Differential Topology and Quantum Field Theory

Chandra Prakash
with Claude (Anthropic), under the author's supervision

chandra.pp@alumni.iitg.ac.in
Latest version: lunar-photon.github.io

30 September 2026

How to read these notes

This book develops the mathematics of differential topology together with the quantum field theory in which it appears, following Charles Nash's *Differential Topology and Quantum Field Theory* chapter by chapter. Nash wrote a fast survey that assumes a great deal. Here every step is carried out. Every object is given explicitly, as a matrix, a transition function or a solution. Every result is traced back to a principle one already trusts: Stokes' theorem, linear algebra, Fourier analysis, gauge invariance, counting zero modes. Over some three thousand pages the aim is to understand each subject at the depth it deserves, rather than to survey it.

Structure

Part I, the Toolkit (Chapters T1–T5), builds what Nash takes for granted: differential forms and the exterior calculus; integration, Stokes' theorem and curvature; Lie groups and gauge fields; Clifford algebras, spinors and Dirac operators, with explicit gamma matrices; and the renormalisation group, from block spins and Polchinski's exact equation to fixed points, conformal invariance and the one-loop β-functions.

Part II, Following Nash (Chapters 1–12), keeps Nash's chapter numbering, so Chapter 4 here is Nash's Chapter IV. Chapter 13, on the mass gap and its topological origins, goes beyond Nash. Each chapter opens with a roadmap and closes with a summary table.

The boxes, and the whole story in seven of them

Each kind of box has one job. Rather than describe them, here is one real box of each kind, taken from the story these notes tell.

Key idea: the whole book in one line

Counting solutions is the same as integrating curvature. For an elliptic operator D (a Dirac operator in a gauge field, say),

$$\underbrace{\dim \operatorname{Ker} D - \dim \operatorname{Ker} D^\dagger}_{\text{an analyst counts zero modes}} = \underbrace{\int_M \hat{A}(M)\ \operatorname{ch}(F)}_{\text{a topologist integrates curvature}} \quad .$$

The left side depends on solving a differential equation; the right side does not change if you deform the metric or the gauge field. Instantons, anomalies, Riemann–Roch and Gauss–Bonnet are all this one equation (Chapter 4).

Key idea boxes hold the one thing to remember.

Working definition 0.1 (fibre bundle, decompressed)

Picture a space M (spacetime, a sphere, a circle). At every point x attach a copy of another space F: a line, a circle, a vector space of internal states. That copy is the **fibre** over x. All the fibres together form the **total space** E, and the projection $\pi : E \to M$ simply remembers which point each fibre hangs over. A **section** is a choice of one point in each fibre, varying smoothly with x. That is exactly what a physicist calls a field. The bundle is twisted when the fibres cannot all be labelled consistently at once: going round the space, the labels come back rotated.

Working definitions unpack every term into a picture you can compute with.

Physical picture: why charge is quantised

A magnetic monopole has no vector potential A defined on the whole sphere around it. Use one potential on the northern cap and another on the southern cap. On the equator they must differ by a gauge transformation $e^{ie\lambda(\phi)}$ of the electron's wavefunction, and going once round the equator that phase must return to itself. *The integer that counts how many times it winds is Dirac's quantisation condition*: topology, not dynamics, makes eg an integer.

Physical pictures say what is really going on.

Worked example 0.1: the monopole flux from two patches

Take $A_N = \frac{g}{2}(1-\cos\theta)\,\mathrm{d}\phi$ and $A_S = -\frac{g}{2}(1+\cos\theta)\,\mathrm{d}\phi$, so that $F = \mathrm{d}A_N = \mathrm{d}A_S = \frac{g}{2}\sin\theta\,\mathrm{d}\theta\wedge\mathrm{d}\phi$. Then

$$\int_{S^2} F \overset{(1)}{=} \int_N \mathrm{d}A_N + \int_S \mathrm{d}A_S \overset{(2)}{=} \oint_{\mathrm{eq}} A_N - \oint_{\mathrm{eq}} A_S \overset{(3)}{=} \oint_{\mathrm{eq}} g\,\mathrm{d}\phi = 2\pi g.$$

(1) Split the sphere into the two caps, and on each use the potential that is smooth there.
(2) Stokes on each cap. The southern cap's boundary is the equator traversed the other way.
(3) $A_N - A_S = g\,\mathrm{d}\phi$ is a pure gauge difference, $\mathrm{d}\lambda$ with $\lambda = g\phi$.

Single-valuedness of $e^{ie\lambda} = e^{ieg\phi}$ forces $eg \in \mathbb{Z}$, so the flux is $2\pi n/e$.

Worked examples are solved to the last line.

Watch out: pure gauge is not trivial

At infinity an instanton is "pure gauge", $A = g^{-1}\mathrm{d}g$, so its field strength vanishes there. Yet the map $g : S^3 \to \mathrm{SU}(2)$ can wrap the sphere k times, and that integer is the instanton number. Zero field strength everywhere you look does not mean nothing is there.

Pitfalls flag the confusions and sign traps.

Reading Nash (p. 32, (2.21)–(2.22))

Nash writes the symbol of $D^\alpha = (-i\partial)^\alpha$ as p^α, but with his own Fourier sign ($\mathcal{F}f = (2\pi)^{-n}\int e^{+ipx} f$) the honest answer is $(-p)^\alpha$. The slip never matters for ellipticity, which only asks where the symbol vanishes. Chapter 2 fixes the convention so his formulas hold as printed.

Reading-Nash boxes map the notes onto Nash's pages, and flag his misprints.

Recap: the dictionary this book builds

connection	gauge potential A
curvature	field strength F
section of a bundle	matter field
Chern number	monopole charge, instanton number
index of the Dirac operator	net number of fermion zero modes, the anomaly
critical points of a Morse function	classical vacua; tunnelling between them

Recaps close every section.

The labelled-equality convention

Every calculation is written as a chain. Each non-trivial equals sign carries a small label, and the labels are explained, in order, right after the chain:

$$\int_M \mathrm{d}\alpha \overset{(1)}{=} \int_{\partial M} \alpha \overset{(2)}{=} 0.$$

(1) Stokes' theorem.
(2) M is closed, so $\partial M = \varnothing$ and the integral is over nothing.

Sources

- C. Nash, *Differential Topology and Quantum Field Theory*, Academic Press 1991, cited as [Nash p. …]. Page numbers are the printed ones.
- T. Dray, *Differential Forms and the Geometry of General Relativity*, CRC Press 2014, cited as [Dray …].
- C. Bär, *Spin Geometry*, lecture notes, Universität Potsdam 2011, cited as [Bär …].
- S. M. Schrans, *Miscellaneous Topics in Quantum Field Theory, with Applications to Particle Physics*, Version 2, 2025, used especially for the explicit anomaly calculations, cited as [QFT notes …].
- A. Gathmann, *Algebraic Geometry*, lecture notes from a course at the University of Kaiserslautern, 2002/2003, used for Chapter 5.
- J. Polchinski, *Renormalization and effective Lagrangians*, Nucl. Phys. B231 (1984) 269, the basis of Chapter T5. Further papers are cited where they are used, and listed at the end of the chapters that use them.

Contents

How to read these notes ii

I The Toolkit 1

T1 Differential Forms: the Working Physicist's Calculus 2

Where this chapter is going . . . 2
T1.1 Manifolds and charts: calculus without global coordinates . . . 3
T1.2 Tangent vectors, covectors and line elements . . . 10
T1.3 The wedge product and the algebra of forms . . . 17
T1.4 Metric, orientation and the Hodge star . . . 25
T1.5 The exterior derivative . . . 34
T1.6 The codifferential and the Laplace–de Rham operator . . . 43
T1.7 Maxwell's equations in the language of forms . . . 50
T1.8 Pullback and naturality . . . 62
T1.9 Interior product, Lie derivative and Cartan's formula . . . 70
T1.10 Complex coordinates and the Dolbeault operator . . . 81
Chapter summary . . . 92

T2 Integration, Stokes' Theorem, Connections and Curvature 99

Where this chapter is going . . . 99
T2.1 Integrating differential forms . . . 101
T2.2 Stokes' theorem . . . 109
T2.3 Closed and exact forms . . . 117
T2.4 A first look at de Rham cohomology . . . 127
T2.5 Connections: how a frame turns from point to point . . . 137
T2.6 Curvature . . . 150
T2.7 Geodesics: straight lines on a curved surface . . . 160
T2.8 The Gauss–Bonnet theorem: integrated curvature is topology . . . 169
T2.9 Applications: equivalence, Lagrangians, spinors and the sphere . . . 182
Chapter summary . . . 198

T3 Lie Groups, Lie Algebras and Gauge Fields 208

Where this chapter is going . . . 208
T3.1 Matrix Lie groups . . . 209
T3.2 Lie algebras: the tangent space at the identity . . . 217
T3.3 The exponential map . . . 224
T3.4 The adjoint representation and the double cover SU(2) → SO(3) . . . 231
T3.5 Representations: spin, weights and roots . . . 238
T3.6 The Maurer–Cartan form $g^{-1}\mathrm{d}g$ . . . 247
T3.7 Group actions, orbits and homogeneous spaces . . . 257
T3.8 Gauge fields: from local symmetry to curvature . . . 264
T3.9 Parallel transport, holonomy and Wilson loops . . . 278
T3.10Pure gauge, flat connections and the Chern–Simons form . . . 291
Chapter summary . . . 311

T4 Clifford Algebras, Spinors and Dirac Operators **320**
Where this chapter is going . . . 320
T4.1 Dirac's square root of the Laplacian . . . 322
T4.2 Clifford algebras: multiplying anticommuting directions . . . 329
T4.3 Gamma matrices in four dimensions . . . 336
T4.4 Gamma matrices in any dimension . . . 346
T4.5 The spin group: rotations built from reflections . . . 354
T4.6 Spin structures: putting spinors on a manifold . . . 364
T4.7 The spin connection and the Dirac operator . . . 373
T4.8 The Lichnerowicz formula, twisted Dirac operators and the index . . . 387
T4.9 Differential forms as Clifford modules . . . 402
Chapter summary . . . 437

T5 The Renormalisation Group **448**
Where this chapter is going . . . 448
T5.1 The renormalisation group: physics at a scale . . . 450
T5.2 The exact renormalisation group . . . 459
T5.3 Fixed points, linearisation and critical exponents . . . 481
T5.4 Running couplings in perturbation theory . . . 493
T5.5 Fixed points are conformal field theories . . . 509
T5.6 Quantum field theories as flows between conformal field theories . . . 521
T5.7 Beta-functions with geometry . . . 525
Reading the literature . . . 538
Chapter summary . . . 540

II Following Nash **541**

1 A Topological Preliminary **542**

Where this chapter is going . . . 542
1.1 From homeomorphism to diffeomorphism . . . 542
1.2 Some algebraic topology: homotopy . . . 554
1.3 Homotopy groups . . . 570
1.4 Cohomology and homology groups . . . 622
1.5 Fibre bundles and fibrations . . . 654
1.6 Differentiable structures for manifolds . . . 684
Chapter summary . . . 712

2 Elliptic Operators **725**
Where this chapter is going . . . 725
2.1 The meaning of ellipticity . . . 725
2.2 Ellipticity and hypo-ellipticity . . . 746
2.3 Ellipticity and vector bundles . . . 766
2.4 Pseudo-differential operators . . . 801
2.5 Pseudo-differential operators and Sobolev spaces . . . 830
2.6 Elliptic operators are Fredholm . . . 853
Chapter summary . . . 874

3 Cohomology of Sheaves and Bundles **886**
Where this chapter is going . . . 886
3.1 Sheaves . . . 886
3.2 Morphisms and exact sequences of sheaves . . . 898
3.3 Čech cohomology . . . 907
3.4 The long exact sequence . . . 919
3.5 Line bundles, $H^1(M;\mathcal{O}^*)$ and the first Chern class . . . 922
3.6 The de Rham theorem . . . 928
3.7 Complex manifolds and the Dolbeault theorem . . . 941
3.8 K-theory: why count bundles, and what we count with . . . 957
3.9 Stable equivalence: ignoring trivial junk . . . 966
3.10 The group $K(M)$: subtracting bundles . . . 973
3.11 Classifying spaces and the stable range . . . 981
3.12 Bundles over spheres: clutching functions and $\widetilde{K}(S^n)$ . . . 990
3.13 Bott periodicity . . . 1002
3.14 Real and quaternionic periodicity: the period eight . . . 1012
3.15 Characteristic classes: what they measure and where they live . . . 1026
3.16 Chern–Weil theory: characteristic classes from curvature . . . 1032
3.17 The splitting principle: pretend every bundle is a sum of lines . . . 1045
3.18 The Chern character and the Todd class . . . 1053
3.19 Real bundles, complexification and Pontryagin classes . . . 1061
3.20 The Euler class, Gauss–Bonnet–Chern, and the genera $\hat{A}$ and L . . . 1070

3.21 Fredholm operators and $K(X)$ . . . 1087
Chapter summary . . . 1110

4 Index Theory for Elliptic Operators 1118
Where this chapter is going . . . 1118
4.1 Counting solutions: the index of a linear map . . . 1118
4.2 Elliptic operators have an index . . . 1125
4.3 K-theory from complexes of bundles . . . 1133
4.4 Elliptic complexes and the rolled-up operator . . . 1141
4.5 The analytical index . . . 1150
4.6 The topological index . . . 1154
4.7 The Atiyah–Singer index theorem . . . 1161
4.8 The heat-kernel view: McKean–Singer . . . 1171
4.9 From K-theory to numbers: the index formula in characteristic classes . . . 1181
4.10 Two vanishing results: odd dimensions and operators on functions . . . 1191
4.11 Orientable manifolds: the index formula on X itself . . . 1198
4.12 The de Rham complex and the Gauss–Bonnet theorem . . . 1206
4.13 The Dolbeault complex and the Riemann–Roch theorem . . . 1214
4.14 The signature complex and the Hirzebruch signature theorem . . . 1224
4.15 The spin complex, the Dirac operator and the $\hat{A}$-genus . . . 1234
4.16 Coupling to gauge fields: counting fermion zero modes . . . 1259
4.17 Twisted complexes . . . 1270
4.18 The index theorem for families of operators . . . 1284
4.19 The index for real families . . . 1298
4.20 Index theory and fixed points . . . 1306
4.21 Index theory for manifolds with boundary . . . 1316
Chapter summary . . . 1368

5 Some Algebraic Geometry 1381
Where this chapter is going . . . 1381
5.1 Algebraic varieties . . . 1383
5.2 Riemann surfaces and divisors . . . 1414
5.3 Serre duality, line bundles and Kähler manifolds . . . 1448
5.4 The Teichmüller space $\mathcal{T}_p$ . . . 1492
5.5 The moduli space $\mathcal{M}_p$ . . . 1506
5.6 The dimension of the moduli space . . . 1518
5.7 Weierstrass gaps and Weierstrass points . . . 1546
Chapter summary . . . 1578

6 Infinite Dimensional Groups 1591
Where this chapter is going . . . 1591

6.1 Some infinite dimensional groups . . . 1593
6.2 Group extensions . . . 1620
6.3 Representations . . . 1660
Chapter summary . . . 1715

7 Morse Theory 1719
Where this chapter is going . . . 1719
7.1 Critical points, the Hessian and the index . . . 1721
7.2 Counting critical points: the Morse inequalities . . . 1730
7.3 How the topology changes at a critical point: sublevel sets, handles and gradient flow . . . 1740
7.4 Three classic calculations: $\mathbb{CP}^n$, surfaces of genus p, and degenerate critical points 1749
7.5 Critical submanifolds: Morse–Bott theory . . . 1759
7.6 Equivariant cohomology: keeping track of a symmetry . . . 1767
7.7 Equivariant Morse theory . . . 1782
7.8 Infinite dimensions: geodesics, loop spaces and Bott periodicity . . . 1789
7.9 Supersymmetric quantum mechanics and Morse theory: the plan . . . 1830
7.10 The supersymmetry algebra . . . 1834
7.11 Supersymmetry from differential forms . . . 1842
7.12 Witten's deformation of the exterior derivative . . . 1850
7.13 The deformed Laplacian: a Schrödinger operator with spin . . . 1855
7.14 Harmonic oscillators at the critical points: the weak Morse inequalities . . . 1862
7.15 Instantons: tunnelling between critical points . . . 1872
7.16 The Morse–Witten complex . . . 1886
Chapter summary . . . 1911

8 Instantons and Monopoles 1920
Where this chapter is going . . . 1920
8.1 The topology of gauge fields . . . 1920
8.2 Secondary characteristic classes . . . 1974
8.3 Instantons: the Yang–Mills action and its critical points . . . 2007
8.4 Self-duality and the Bogomolny bound . . . 2014
8.5 The BPST instanton and the 't Hooft ansatz . . . 2023
8.6 Gauge invariance, gauge fixing and ellipticity . . . 2038
8.7 The instanton moduli space and reducible connections . . . 2045
8.8 The deformation complex . . . 2055
8.9 Counting the moduli: the index theorem . . . 2065
8.10 The $k = 1$ moduli space and the ADHM construction . . . 2079
8.11 The moduli space as a global object . . . 2096
8.12 Monopoles: the Yang–Mills–Higgs system in three dimensions . . . 2108
8.13 The Bogomolny equation and the Prasad–Sommerfield monopole . . . 2119

8.14 Monopoles as symmetric instantons . . . 2129
8.15 Monopole moduli . . . 2141
8.16 Monopole dynamics: forces and the geodesic approximation . . . 2152
8.17 Two monopoles: the Atiyah–Hitchin manifold and scattering . . . 2166
8.18 Critical point theory and gauge theories . . . 2176
Chapter summary . . . 2203

9 The Elliptic Geometry of Strings **2208**
Where this chapter is going . . . 2208
9.1 The bosonic string . . . 2209
9.2 The space of metrics . . . 2236
9.3 The Weil–Petersson metric . . . 2254
Chapter summary . . . 2283

10 Anomalies **2295**
Where this chapter is going . . . 2295
10.1 Introduction: what an anomaly is . . . 2295
10.2 Anomalies and Yang–Mills theories I: chiral fermions, the index and the determinant 2302
10.3 Anomalies and Yang–Mills theories II: the anomalous Ward identity, computed 2313
10.4 Anomalies and Yang–Mills theories III: the measure, the heat kernel and the index 2341
10.5 Anomalies and Yang–Mills theories IV: non-abelian anomalies, consistency and descent . . . 2353
10.6 Anomalies and Yang–Mills theories V: the topology of the anomaly . . . 2372
10.7 Gravitational anomalies . . . 2383
10.8 The critical dimension for strings . . . 2390
10.9 Global anomalies . . . 2409
10.10 Anomalies from a Hamiltonian perspective . . . 2434
Chapter summary . . . 2489

11 Conformal Quantum Field Theories **2495**
Where this chapter is going . . . 2495
11.1 Conformal invariance and quantum field theory . . . 2496
11.2 Weyl rescalings versus conformal transformations . . . 2507
11.3 Conformal field theories in two dimensions . . . 2524
11.4 Relation to the Virasoro algebra . . . 2550
11.5 Critical phenomena and the two-dimensional Ising model . . . 2604
11.6 The Ising model as the $c = \frac{1}{2}$ conformal field theory . . . 2616
11.7 Partition functions on the torus . . . 2630
11.8 Modular forms and the modular invariance of Z . . . 2639
11.9 Operator products and correlation functions . . . 2655
11.10 Fusion rules and the Verlinde formula . . . 2671

11.11Axiomatic approaches I: geometry on moduli space . . . 2690
11.12Axiomatic approaches II: Segal's definition . . . 2706
11.13The c-theorem: the flow runs downhill . . . 2718
Chapter summary . . . 2728

12 Topological Quantum Field Theories 2734
Where this chapter is going . . . 2734
12.1 Introduction: quantum field theory without a metric . . . 2735
12.2 Floer's programme and the Chern–Simons function . . . 2742
12.3 Critical points, the Hessian and the relative Morse index . . . 2753
12.4 Gradient flow lines are instantons . . . 2765
12.5 Floer homology, Casson's invariant and the Chern–Simons theory . . . 2774
12.6 Donaldson's polynomial invariants . . . 2790
12.7 Witten's topological field theory . . . 2821
12.8 Cutting along homology spheres: Floer meets Donaldson . . . 2833
12.9 Knots, links and when two of them are the same . . . 2856
12.10The linking number: Gauss, Ampère and a degree . . . 2868
12.11Knot polynomials: Alexander, Jones and their skein relations . . . 2877
12.12Wilson loops and abelian Chern–Simons theory . . . 2896
12.13Chern–Simons theory at weak coupling: flat connections and torsion . . . 2912
12.14Chern–Simons theory and the Jones polynomial . . . 2926
12.15Surgery and the Jones polynomial . . . 2958
Chapter summary . . . 3008

13 The Mass Gap 3017
Where this chapter is going . . . 3017
13.1 What a mass gap is . . . 3018
13.2 The gap makes correlations decay exponentially . . . 3025
13.3 Elementary ways to make a gap . . . 3036
13.4 A gap from instantons: Polyakov's compact QED in 2+1 dimensions . . . 3045
13.5 A gap from a topological term: the Chern–Simons photon . . . 3052
13.6 θ-terms and the Haldane gap . . . 3058
13.7 Anomalies forbid a trivial gap . . . 3066
13.8 Supersymmetry, the Witten index, and gapped phases as TQFTs . . . 3072
13.9 The Yang–Mills mass gap problem . . . 3075
Chapter summary . . . 3081

Part I

The Toolkit

Chapter T1

Differential Forms: the Working Physicist's Calculus

Where this chapter is going

Nash writes on page 3 that he wants "to make full use of calculus on manifolds to calculate geometric and topological quantities". He then uses that calculus on every page without ever setting it up. This chapter sets it up. It is the calculus a physicist already knows (gradient, curl, divergence, line and surface integrals, Jacobians, Maxwell's equations), rewritten in a language that works in any dimension, in any signature, in any coordinates, and on curved spaces such as spheres and tori where no single coordinate system covers everything.

The language is that of **differential forms**. A differential form is, very concretely, *the thing you put under an integral sign*: $F_x\,\mathrm{d}x + F_y\,\mathrm{d}y + F_z\,\mathrm{d}z$ for a work integral, $B_x\,\mathrm{d}y\,\mathrm{d}z + \dots$ for a flux, $\rho\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z$ for a charge. We will learn to multiply such objects (the wedge product $\wedge$), to convert one kind into another using a metric (the Hodge star $\star$), to differentiate them (the exterior derivative d, which contains gradient, curl and divergence as special cases), and to transport them along maps (the pullback f^*) and along flows (the Lie derivative $\mathcal{L}_X$). Finally we complexify: on a complex manifold d splits into two halves $\partial + \bar\partial$, and $\bar\partial$ is the operator behind the Dolbeault complex that Nash uses in Chapters 3 and 4.

We follow Tevian Dray's Part III (Chapters 11–15) closely in spirit and pace: vector calculus first, then short computations, one idea at a time. Integration and Stokes' theorem are left to the next toolkit chapter (Chapter T2); here we only use them in passing. Figure T1.1 shows how the pieces of this chapter feed each other and the rest of the book.

Key idea: Conventions used in this chapter

- Manifolds are smooth, of dimension n, with local coordinates x^μ ($\mu = 1, \dots, n$). Repeated indices are summed (Einstein).
- A p-form is written $\alpha = \frac{1}{p!}\,\alpha_{\mu_1\dots\mu_p}\,\mathrm{d}x^{\mu_1}\wedge\cdots\wedge\mathrm{d}x^{\mu_p}$ with totally antisymmetric components.
- The wedge product acts on vectors by the determinant rule $(\alpha\wedge\beta)(v,w) = \alpha(v)\beta(w) - \alpha(w)\beta(v)$, so $(\mathrm{d}x\wedge\mathrm{d}y)(\partial_x,\partial_y) = 1$ (Dray's convention).
- The Hodge star is defined by $\alpha\wedge\star\beta = g(\alpha,\beta)\,\mathrm{vol}$.
- Signature s = number of minus signs in the metric. Minkowski space uses $\eta = \mathrm{diag}(-1,+1,+1,+1)$ in coordinates (t,x,y,z), orientation $\mathrm{vol} = \mathrm{d}t\wedge\mathrm{d}x\wedge\mathrm{d}y\wedge\mathrm{d}z$, $c = 1$, Heaviside–Lorentz units.
- The Laplace–de Rham operator $\Delta = \mathrm{d}\delta + \delta\mathrm{d}$ is a *positive* operator: on functions in flat

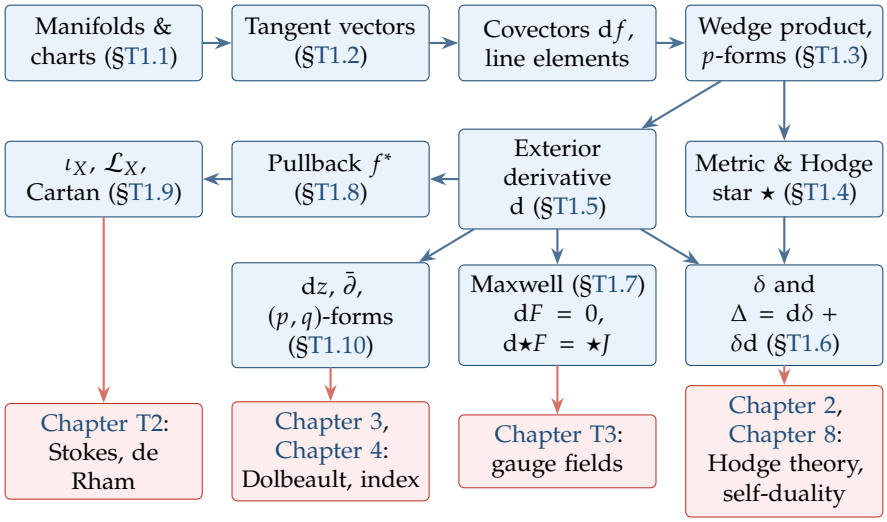


Figure T1.1: The logic of this chapter. Blue boxes are sections; red boxes are the places later in the book that lean on them. Read the top row as "where do vectors and integrands live?", the middle row as "what can we do to integrands?", and the bottom row as "what physics comes out?".

space $\Delta = -\nabla^2$. This is Nash's convention (his (2.42)).

T1.1 Manifolds and charts: calculus without global coordinates

Here is the problem. We want to do calculus (differentiate fields, write field equations, integrate densities) on spaces like the sphere S^2, the circle S^1, the torus T^2, a Euclidean space-time with the point at infinity added (S^4, the home of instantons), or the space of all gauge fields modulo gauge transformations. Calculus needs coordinates. But none of these spaces can be covered by a single coordinate system without something going wrong somewhere: the longitude on a sphere is undefined at the poles, the angle on a circle jumps by 2π somewhere.

The way out is simple, and every physicist already uses it. Cover the space with several overlapping coordinate systems, each of them good where it is defined, and write down how to translate between them on the overlaps. That is all a *manifold* is. Topology enters because the translation rules on the overlaps can carry global information (twists, windings) that no single patch can see.

T1.1.1 Charts: coordinate patches

Working definition T1.1 (chart, or coordinate patch)

A **chart** on a space M is an open region $U \subset M$ [with each of its points, U also contains all sufficiently nearby points of M] together with n functions $x^1, \dots, x^n$ on U that label the points of U one-to-one and continuously by points of an open region of $\mathbb{R}^n$. We write $\varphi : U \to \mathbb{R}^n$, $\varphi(p) = (x^1(p), \dots, x^n(p))$. *Operational test:* nearby points get nearby labels, distinct points get distinct labels, and every label near a used label is also used (the image is open).

The last condition is the one people forget. The usual polar angle $\phi \in [0, 2\pi)$ on the circle labels every point once, but it is *not* a chart on the whole circle: the points just below the positive x-axis get labels near 2π while the point on the axis gets label 0. Nearby points get distant

labels. The fix is to throw away one point, and cover it with a second chart.

Physical picture: A chart is a tracing on a paper cutout

Press a piece of paper onto the sphere, trace the curve drawn there, then lay the paper flat and read the curve off as $x^1(t), x^2(t)$ (Figure T1.2). Differentiating, integrating and parallel transport are all done on this flat tracing. It gives the same answers as embedding the sphere in $\mathbb{R}^3$ and watching how the basis vectors turn from point to point, $\nabla_a e_b = \Gamma^c{}_{ab}\, e_c$ (Example T2.27), without needing any embedding.

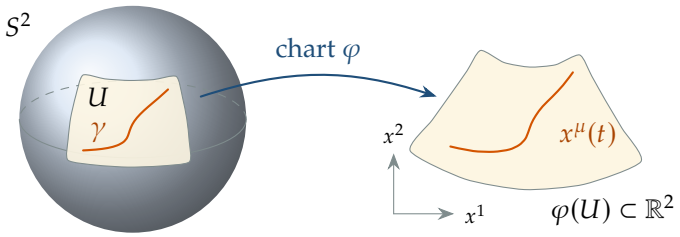


Figure T1.2: A chart as a tracing. Left: a paper cutout pressed onto the sphere covers the region U, and the curve γ is traced onto it. Right: the same cutout laid flat. The tracing is now an ordinary plane curve $x^\mu(t)$, and every derivative and integral is computed there. On the sphere the bottom edge of the cutout is a little longer than the top edge. On the table it is more than twice as long: flattening stretches the paper unevenly, and the metric remembers by how much. (The flattening drawn is the south stereographic chart of Example T1.3, turned so that north is up.)

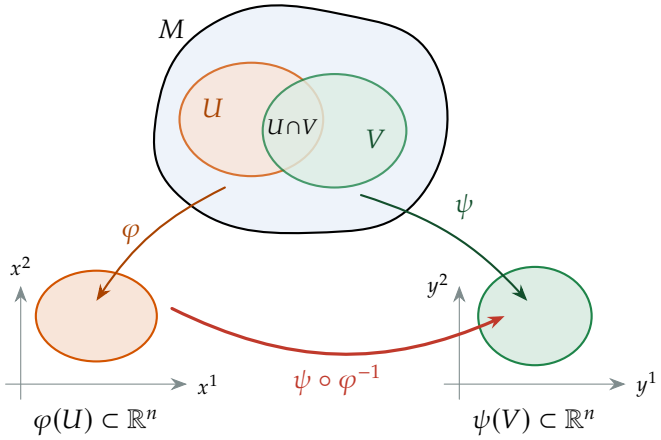


Figure T1.3: Two overlapping charts. A point in the overlap $U \cap V$ has two sets of coordinates, x^μ and y^μ. The red arrow, the *transition map* $y = \psi \circ \varphi^{-1}(x)$, is an ordinary function between regions of $\mathbb{R}^n$, and it is all we ever need to compute with.

T1.1.2 Transition maps and atlases

On the overlap $U \cap V$ of two charts, each point has two coordinate labels, x^μ from φ and y^μ from ψ. The rule converting one into the other,

$$y^\mu = y^\mu(x^1, \ldots, x^n) \qquad \text{i.e.} \qquad \psi \circ \varphi^{-1} : \ \varphi(U \cap V) \to \psi(U \cap V), \tag{T1.1}$$

is called the **transition map**. It is an ordinary map between open sets of $\mathbb{R}^n$, so it makes sense to ask whether it is differentiable.

Working definition T1.2 (smooth manifold)

A **smooth manifold** of dimension n is a space M covered by charts [open regions $U \subset M$, each labelled one-to-one by coordinates $x^1, \dots, x^n$] (a collection of charts covering M is an **atlas**) such that every transition map (T1.1) is infinitely differentiable, with infinitely differentiable inverse. *Operational test:* write down enough charts to cover every point, then compute every transition map on every overlap and check that the formulas involve no division by zero, no square roots of zero, and no absolute values on the overlap.

Formally one also asks that M be Hausdorff and second countable (two technical conditions that exclude pathological examples like "a line with two origins"); every space we meet satisfies them.

Why do we insist on smooth transition maps? Because *everything we define in one chart must mean the same thing in the other*. We will define a smooth function, a tangent vector, a differential form, by giving it in coordinates. The chain rule then converts it into any other chart. The chain rule needs derivatives of $y^\mu(x)$, and higher derivatives of fields need higher derivatives of $y^\mu(x)$. Smoothness of the transition maps is exactly what makes "smooth" a chart-independent notion.

Working definition T1.3 (smooth function on a manifold)

A function $f : M \to \mathbb{R}$ is **smooth** if in every chart the coordinate expression $f(x^1, \dots, x^n)$ is an infinitely differentiable function of the x^μ. *Test:* it suffices to check this in the charts of one atlas [a set of charts that together cover every point of M], since the transition maps [the formulas $y^\mu(x^1, \dots, x^n)$ converting one chart's coordinates into another's] are smooth.

T1.1.3 The circle

Worked example T1.1: S^1 with two angle charts

Let $S^1 = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$. Chart 1: $U_1 = S^1 \setminus \{(1, 0)\}$, with coordinate $\theta \in (0, 2\pi)$ defined by $(x, y) = (\cos\theta, \sin\theta)$. Chart 2: $U_2 = S^1 \setminus \{(-1, 0)\}$, with $\theta' \in (-\pi, \pi)$ defined by the same formula. Each chart misses one point, and together they cover the circle.
The overlap $U_1 \cap U_2$ is the circle minus two points, which has *two* pieces: the upper half ($y > 0$) and the lower half ($y < 0$). On the upper half both angles lie in $(0, \pi)$ and name the same point, so $\theta' = \theta$. On the lower half, $\theta \in (\pi, 2\pi)$ while $\theta' \in (-\pi, 0)$, so

$$\theta' \overset{(1)}{=} \theta - 2\pi \qquad (y < 0),$$
$$\frac{\mathrm{d}\theta'}{\mathrm{d}\theta} \overset{(2)}{=} 1 \qquad \text{on the whole overlap.}$$

(1) Both angles give the same point $(\cos\theta, \sin\theta)$, and the cosine and sine are 2π-periodic; the shift that maps $(\pi, 2\pi)$ onto $(-\pi, 0)$ is -2π.
(2) A constant shift has derivative one; the transition map is smooth (indeed linear) on both components.

So S^1 is a one-dimensional smooth manifold. Notice the constant -2π on one component of the overlap: it is invisible to derivatives, but it will make the 1-form "$\mathrm{d}\theta$" globally defined while θ itself is not. That is the first appearance of topology, and we return to it in Example T1.30 and in Chapter T2.

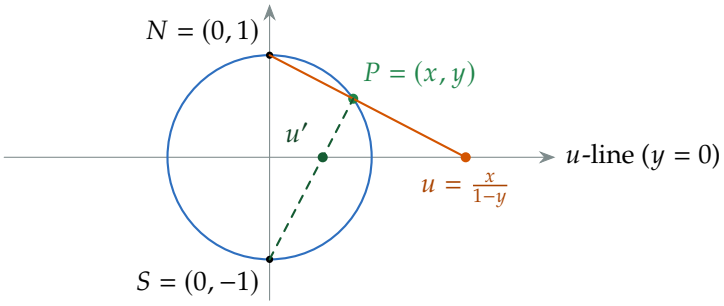


Figure T1.4: Stereographic projection of the circle. The line from the north pole N through P hits the horizontal axis at u; the line from the south pole S through P hits it at u'. Each projection covers every point except its own pole. Points near N are sent far out along the u-line, but near 0 along the u'-line: $u' = 1/u$.

Worked example T1.2: S^1 with two stereographic charts

Project from the north pole $N = (0, 1)$: the straight line through N and $P = (x, y)$ meets the line $y = 0$ at the point u (Figure T1.4). By similar triangles, $u/1 = x/(1 - y)$, so

$$u = \frac{x}{1-y} \quad \text{on } U_N = S^1 \setminus \{N\}, \qquad u' = \frac{x}{1+y} \quad \text{on } U_S = S^1 \setminus \{S\},$$

the second formula coming from projecting from $S = (0, -1)$ instead. On the overlap (the circle minus both poles, where $x \neq 0$),

$$\begin{aligned} u\,u' &\overset{(1)}{=} \frac{x^2}{(1-y)(1+y)} \\ &\overset{(2)}{=} \frac{x^2}{1-y^2} \\ &\overset{(3)}{=} 1, \qquad \text{so} \qquad u' = \frac{1}{u}. \end{aligned}$$

(1) Multiply the two definitions.
(2) Difference of squares.
(3) On the circle, $1 - y^2 = x^2$, and $x \neq 0$ on the overlap.

The transition map $u \mapsto 1/u$ is smooth on the overlap $u \neq 0$ (the two components $u > 0$ and $u < 0$ are the right and left halves of the circle). The inverse map is also useful:

$$x = \frac{2u}{1+u^2}, \qquad y = \frac{u^2-1}{u^2+1},$$

which one checks by substituting into $x^2 + y^2$ and into $x/(1 - y)$. This is the rational parametrisation of the circle behind the "$t = \tan(\theta/2)$" substitution of integral calculus.

Watch out: One chart is never enough for a compact manifold

You cannot cover all of S^1 with one chart. A chart is a continuous one-to-one map onto an *open* interval, and the image of a compact space under a continuous map is compact. A compact open subset of $\mathbb{R}$ is empty. The same argument works for S^2, T^2 and every compact manifold: at least two charts are always needed. Physicists meet this as "coordinate singularities" (the poles of spherical coordinates, $r = 2m$ in Schwarzschild coordinates),

which are artefacts of a bad chart, not of the space.

T1.1.4 The two-sphere, worked out completely

Take $S^2 = \{(X, Y, Z) \in \mathbb{R}^3 : X^2 + Y^2 + Z^2 = 1\}$. Project from the north pole $N = (0, 0, 1)$ onto the equatorial plane $Z = 0$; the image point has coordinates (u, v). Project from the south pole $S = (0, 0, -1)$ instead; call the image (u', v'). The same similar-triangles argument as for the circle gives

$$(u, v) = \left(\frac{X}{1 - Z}, \frac{Y}{1 - Z}\right), \qquad (u', v') = \left(\frac{X}{1 + Z}, \frac{Y}{1 + Z}\right). \tag{T1.2}$$

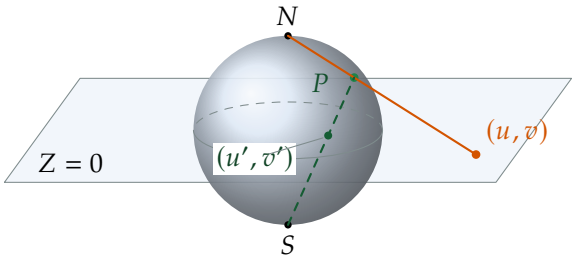


Figure T1.5: Stereographic charts on S^2. A point P in the northern hemisphere is projected from N to a point far from the origin, and from S to a point near the origin. The north chart misses only N (it goes to infinity), the south chart misses only S. The two labels are related by inversion in the unit circle, (T1.3).

Worked example T1.3: Transition map and Jacobian for S^2

First the transition map. Write $\rho^2 = u^2 + v^2$. Then

$$\begin{aligned} \rho^2 &\overset{(1)}{=} \frac{X^2 + Y^2}{(1 - Z)^2} \overset{(2)}{=} \frac{1 - Z^2}{(1 - Z)^2} \overset{(3)}{=} \frac{1 + Z}{1 - Z}, \\ u' &\overset{(4)}{=} \frac{X}{1 + Z} = \frac{X}{1 - Z} \cdot \frac{1 - Z}{1 + Z} \overset{(5)}{=} \frac{u}{\rho^2}, \qquad v' = \frac{v}{\rho^2}. \end{aligned} \tag{T1.3}$$

(1) Definition (T1.2) of u, v.
(2) On the sphere, $X^2 + Y^2 = 1 - Z^2$.
(3) Factor $1 - Z^2 = (1 - Z)(1 + Z)$ and cancel one factor.
(4) Definition of u'; multiply and divide by $1 - Z$.
(5) The first factor is u and the second is $1/\rho^2$ by the first line.

The overlap is the sphere minus both poles, i.e. $\rho \neq 0$ (and $\rho \neq \infty$), where (T1.3) is manifestly smooth: it is inversion in the unit circle. Now the Jacobian matrix:

$$\begin{aligned} \frac{\partial u'}{\partial u} &\overset{(1)}{=} \frac{\rho^2 - 2u^2}{\rho^4} = \frac{v^2 - u^2}{\rho^4}, & \frac{\partial u'}{\partial v} &\overset{(2)}{=} -\frac{2uv}{\rho^4}, \\ \frac{\partial v'}{\partial u} &= -\frac{2uv}{\rho^4}, & \frac{\partial v'}{\partial v} &= \frac{u^2 - v^2}{\rho^4}, \end{aligned}$$

$$\begin{aligned} \det \frac{\partial(u', v')}{\partial(u, v)} &\overset{(3)}{=} \frac{(v^2 - u^2)(u^2 - v^2) - 4u^2v^2}{\rho^8} \\ &\overset{(4)}{=} -\frac{(u^2 - v^2)^2 + 4u^2v^2}{\rho^8} \overset{(5)}{=} -\frac{(u^2 + v^2)^2}{\rho^8} = -\frac{1}{\rho^4}. \end{aligned}$$

(1) Quotient rule on $u/(u^2+v^2)$: $(\rho^2\cdot 1 - u\cdot 2u)/\rho^4$, and $\rho^2 - 2u^2 = v^2 - u^2$.
(2) Quotient rule with the numerator u independent of v.
(3) 2×2 determinant, $ad - bc$.
(4) Pull out the minus sign.
(5) $(a-b)^2 + 4ab = (a+b)^2$ with $a = u^2$, $b = v^2$.

The Jacobian determinant is never zero on the overlap, so the transition map is a diffeomorphism there. It is *negative*. We return to that sign immediately.

For computing, the inverse of (T1.2) is indispensable:

$$X = \frac{2u}{1+\rho^2}, \qquad Y = \frac{2v}{1+\rho^2}, \qquad Z = \frac{\rho^2-1}{\rho^2+1}. \tag{T1.4}$$

Check: $X^2+Y^2+Z^2 = \big(4\rho^2 + (\rho^2-1)^2\big)/(1+\rho^2)^2 = (\rho^2+1)^2/(\rho^2+1)^2 = 1$, and $1 - Z = 2/(1+\rho^2)$, so $X/(1-Z) = u$. The height function Z is a smooth function on S^2 because in the north chart it is the rational function $(\rho^2-1)/(\rho^2+1)$, and in the south chart it is $(1-\rho'^2)/(1+\rho'^2)$, both smooth everywhere they are defined.

T1.1.5 Orientation

A negative Jacobian determinant means the two charts disagree about handedness: a small counter-clockwise loop in the (u,v) plane is a clockwise loop in the (u',v') plane. Integrals of fluxes depend on handedness, so we need to know when a consistent choice exists.

Working definition T1.4 (orientable manifold)

An atlas is **oriented** if every transition map [the formulas $y^\mu(x)$ converting one chart's coordinates into another's on an overlap] has positive Jacobian determinant $\det(\partial y^\mu/\partial x^\nu) > 0$ on every overlap [the two charts agree there on handedness, e.g. in 2D on which small loops run counter-clockwise]. A manifold is **orientable** if it admits an oriented atlas. *Operational test:* compute the Jacobian determinants of your transition maps; if some are negative, try flipping the sign of one coordinate in some charts; if you can make all of them positive, the manifold is orientable.

For S^2 the fix is to flip one coordinate in the south chart: use $(u', -v')$. The new transition map has Jacobian determinant $(-1)\times(-1/\rho^4) = +1/\rho^4 > 0$. So S^2 is orientable.

Physical picture: The Riemann sphere is hiding here

Combine the north coordinates into one complex number $w = u + iv$, and the flipped south coordinates into $w' = u' - iv'$. Then (T1.3) says $u' + iv' = w/|w|^2 = 1/\bar{w}$, so $w' = \overline{1/\bar{w}} = 1/w$. The transition map of the oriented atlas is the holomorphic map $w \mapsto 1/w$: the sphere is the complex projective line $\mathbb{CP}^1$, the Riemann sphere. We use this in Section T1.10.

The standard non-orientable example is the Möbius band: going once around it reverses a small circle's sense of rotation, and any atlas has a transition map with negative determinant. Nash's orientability discussions (the first Stiefel–Whitney class, Chapter 1) are the topological version of this test.

Worked example T1.4: The torus T^2

Let $T^2 = \mathbb{R}^2/\mathbb{Z}^2$: points of the plane with $(x, y) \sim (x + m, y + n)$ for integers m, n. Chart: take any open square of side less than 1 in the plane; its points are pairwise inequivalent, so (x, y) restricted to it is a chart. Four such squares, say centred at $(0, 0)$, $(\frac{1}{2}, 0)$, $(0, \frac{1}{2})$, $(\frac{1}{2}, \frac{1}{2})$ with side 0.9, cover the torus. On an overlap, two labels of the same point differ by an integer vector:

$$(x', y') = (x + m,\ y + n), \qquad \frac{\partial(x', y')}{\partial(x, y)} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

with (m, n) constant on each connected piece of the overlap. All transition maps are translations: smooth, with determinant +1. So T^2 is an orientable smooth manifold, and the differentials $\mathrm{d}x$, $\mathrm{d}y$ make sense globally even though x and y do not (as with $\mathrm{d}\theta$ on S^1). This is the torus on which Nash's string theory (Chapter 9) and many index-theorem examples live.

T1.1.6 Manifolds defined by equations

Most manifolds a physicist meets are defined as the solution set of equations: the sphere $X^2 + Y^2 + Z^2 = 1$, the mass shell $p^2 = -m^2$, a group such as $\mathrm{SU}(2) = \{g : g^\dagger g = 1, \det g = 1\}$. There is a one-line test that such a set is a manifold.

Key idea: Regular level sets are manifolds

Let $f : \mathbb{R}^N \to \mathbb{R}^k$ be smooth and let $M = f^{-1}(c)$. If at every point of M the k differentials $\mathrm{d}f^1, \dots, \mathrm{d}f^k$ are linearly independent (the Jacobian matrix of f has rank k), then M is a smooth manifold of dimension $N - k$.

Why: by the implicit function theorem, near any point of M we can solve the k equations for k of the coordinates in terms of the other $N - k$, and those $N - k$ coordinates form a chart. For the sphere, $f = X^2 + Y^2 + Z^2$ has $\mathrm{d}f = 2(X\,\mathrm{d}X + Y\,\mathrm{d}Y + Z\,\mathrm{d}Z)$, which vanishes only at the origin, not on M; so S^2 is a 2-manifold. Near a point with $Z > 0$ we solve $Z = \sqrt{1 - X^2 - Y^2}$ and use (X, Y) as a chart: six such "hemisphere charts" ($Z > 0$, $Z < 0$, $X > 0$, ...) give yet another atlas of S^2.

Watch out: A manifold is not its coordinates

The same sphere carries the stereographic atlas, the hemisphere atlas and the (θ, ϕ) angles (which fail at the poles and along a line of longitude). Anything physical (a field, an integral, a curvature) must come out the same in every chart. Much of this chapter is about building objects for which that is automatic: $\mathrm{d}f$, p-forms, $\star$, d.

Reading Nash (p. 3 and p. 15)

On p. 3 Nash distinguishes topological, piecewise-linear and differentiable manifolds and then says he will "usually deal with differentiable manifolds" and simply call them manifolds. That is our convention too. On p. 15 he introduces the tangent bundle $TM = \bigcup_p T_pM$ as "one of the most basic examples" of a bundle, taking the tangent space T_pM for granted; the next section constructs it.

Recap

- A manifold is a space covered by overlapping coordinate charts whose transition maps are smooth.
- S^1: angle charts with transition $\theta' = \theta$ or $\theta - 2\pi$; stereographic charts with $u' = 1/u$.
- S^2: stereographic charts with $u' = u/\rho^2$, $v' = v/\rho^2$ and Jacobian $-1/\rho^4$; flipping v' makes the atlas oriented and the transition becomes $w' = 1/w$ (the Riemann sphere).
- Orientable = transition Jacobians can all be made positive.
- Regular level sets $f^{-1}(c)$ of maps with independent $\mathrm{d}f^i$ are manifolds.

T1.2 Tangent vectors, covectors and line elements

What is a vector on a sphere? In $\mathbb{R}^3$ we draw arrows, but an arrow at a point of S^2 sticks out of the sphere unless it happens to be tangent, and a creature living on the sphere has no "outside" to draw it in. We need a notion of vector that uses only the manifold and its charts. There are two equivalent answers, and physicists use both without noticing: a vector is the *velocity of a curve*, and a vector is a *directional derivative*. Once vectors are in place, the object dual to them, the *covector*, turns out to be the familiar differential $\mathrm{d}f$. That is the first differential form.

T1.2.1 Vectors as velocities of curves

Let $\gamma(t)$ be a curve on M passing through a point p at $t = 0$. In a chart it is described by n functions $x^\mu(t)$, and its velocity at p has components

$$v^\mu = \left.\frac{\mathrm{d}x^\mu}{\mathrm{d}t}\right|_{t=0}. \tag{T1.5}$$

In another chart with coordinates y^μ, the same curve is $y^\mu(t) = y^\mu(x(t))$ and the chain rule gives

$$\tilde{v}^\mu = \left.\frac{\mathrm{d}y^\mu}{\mathrm{d}t}\right|_0 = \frac{\partial y^\mu}{\partial x^\nu}\left.\frac{\mathrm{d}x^\nu}{\mathrm{d}t}\right|_0 = \frac{\partial y^\mu}{\partial x^\nu}\, v^\nu. \tag{T1.6}$$

This is the **contravariant** transformation law: components of a vector transform with the Jacobian matrix of the transition map.

Working definition T1.5 (tangent vector)

A **tangent vector** at p is an assignment of n numbers v^μ to each chart around p [each coordinate system x^μ defined on a region containing p], related between charts by (T1.6). Every such assignment is the velocity of some curve (take $x^\mu(t) = x^\mu(p) + tv^\mu$). The set of all tangent vectors at p is the **tangent space** T_pM, an n-dimensional vector space. A **vector field** is a smooth choice of $v \in T_pM$ at every p [in every chart the components $v^\mu(x)$ are infinitely differentiable functions of the x^ν].

T1.2.2 Vectors as directional derivatives

A velocity tells us how fast any function changes as we move along the curve. For a function f on M,

$$\left.\frac{\mathrm{d}}{\mathrm{d}t} f(\gamma(t))\right|_{t=0} \overset{(1)}{=} \frac{\partial f}{\partial x^\mu}\left.\frac{\mathrm{d}x^\mu}{\mathrm{d}t}\right|_0 \overset{(2)}{=} v^\mu \frac{\partial f}{\partial x^\mu}. \tag{T1.7}$$

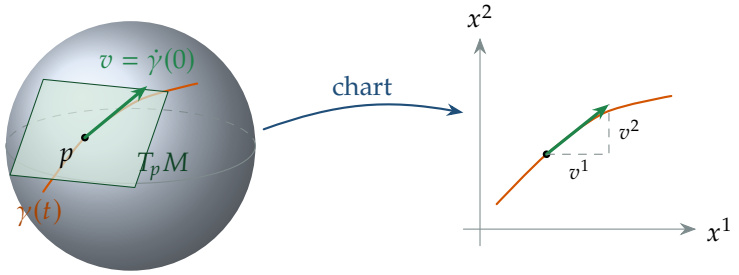


Figure T1.6: A tangent vector is the velocity of a curve through p. On the sphere we can draw it in the tangent plane T_pM; in a chart it is simply the pair of numbers $v^\mu = \dot x^\mu(0)$, read off the flattened tracing of Figure T1.2. Another chart gives other numbers, related by the Jacobian, (T1.6).

(1) Chain rule in the chart.
(2) Definition (T1.5) of the components v^μ.

The right-hand side depends on the curve only through v. So we may *identify* the vector with the differential operator it defines:

$$v = v^\mu \frac{\partial}{\partial x^\mu} \equiv v^\mu \partial_\mu, \qquad v(f) = v^\mu \partial_\mu f. \tag{T1.8}$$

The n operators $\partial_\mu = \partial/\partial x^\mu$ form a basis of T_pM, the **coordinate basis**. This is the notation of every differential geometry book, and it has a pleasant bonus: the transformation law (T1.6) is just the chain rule for partial derivatives,

$$\frac{\partial}{\partial x^\nu} = \frac{\partial y^\mu}{\partial x^\nu}\frac{\partial}{\partial y^\mu}, \qquad \text{so} \qquad v^\nu \frac{\partial}{\partial x^\nu} = \left(\frac{\partial y^\mu}{\partial x^\nu} v^\nu\right)\frac{\partial}{\partial y^\mu} = \tilde v^\mu \frac{\partial}{\partial y^\mu}.$$

The combination $v^\mu \partial_\mu$ is chart-independent: the components and the basis vectors transform oppositely and the changes cancel.

Dictionary: vectors

physicist writes	geometer writes
$\hat x, \hat y, \hat z$	$\partial_x, \partial_y, \partial_z$
$\vec v \cdot \nabla f$	$v(f)$
$\vec v = v_x \hat x + v_y \hat y$	$v = v^x \partial_x + v^y \partial_y$
$\hat r, \hat\phi$ (unit)	$\partial_r, \frac{1}{r}\partial_\phi$

Worked example T1.5: Polar coordinate basis vectors

With $x = r\cos\phi$, $y = r\sin\phi$, the chain rule gives

$$\partial_r \overset{(1)}{=} \frac{\partial x}{\partial r}\partial_x + \frac{\partial y}{\partial r}\partial_y = \cos\phi\,\partial_x + \sin\phi\,\partial_y \overset{(2)}{=} \frac{x\,\partial_x + y\,\partial_y}{r},$$
$$\partial_\phi \overset{(3)}{=} \frac{\partial x}{\partial \phi}\partial_x + \frac{\partial y}{\partial \phi}\partial_y = -r\sin\phi\,\partial_x + r\cos\phi\,\partial_y \overset{(4)}{=} -y\,\partial_x + x\,\partial_y.$$

(1) The rule $\partial/\partial r = (\partial x/\partial r)\,\partial/\partial x + (\partial y/\partial r)\,\partial/\partial y$.
(2) Substitute $\cos\phi = x/r$, $\sin\phi = y/r$.
(3) Same rule for ϕ.
(4) Substitute $r\sin\phi = y$, $r\cos\phi = x$.

So ∂_r is the unit radial vector $\hat r$, but ∂_ϕ has length r: it is $r\hat\phi$. Coordinate basis vectors are

in general *not* unit vectors. The vector field $\partial_\phi = -y\partial_x + x\partial_y$ is the generator of rotations, which we will meet again as a Killing vector and in the Lie derivative (Section T1.9). Acting on functions: $\partial_\phi(x) = -y$, $\partial_\phi(y) = x$, $\partial_\phi(x^2+y^2) = 0$.

Worked example T1.6: A latitude circle on S^2, seen in two charts

Take the circle of constant $Z = Z_0$ on the unit sphere, traversed as $X = a\cos t$, $Y = a\sin t$ with $a = \sqrt{1-Z_0^2}$. In the north stereographic chart (T1.2), $u = a\cos t/(1-Z_0)$, $v = a\sin t/(1-Z_0)$, so the velocity is

$$\dot u = -v, \quad \dot v = u, \qquad \text{i.e.} \qquad V = -v\,\partial_u + u\,\partial_v,$$

a rotation field in the (u,v) plane. Transform it to the south chart with the Jacobian of Example T1.3:

$$\tilde V^{u'} \overset{(1)}{=} \frac{v^2-u^2}{\rho^4}(-v) + \left(-\frac{2uv}{\rho^4}\right)u \overset{(2)}{=} -\frac{v(u^2+v^2)}{\rho^4} \overset{(3)}{=} -v',$$
$$\tilde V^{v'} \overset{(4)}{=} \left(-\frac{2uv}{\rho^4}\right)(-v) + \frac{u^2-v^2}{\rho^4}\,u = \frac{u(u^2+v^2)}{\rho^4} = u'.$$

(1) Contravariant law (T1.6): first row of the Jacobian times $(V^u, V^v) = (-v, u)$.
(2) Expand: $-v^3 + u^2v - 2u^2v = -v(u^2+v^2)$.
(3) $\rho^2 = u^2+v^2$ and $v' = v/\rho^2$ from (T1.3).
(4) Second row of the Jacobian.

So $V = -v'\partial_{u'} + u'\partial_{v'}$: the same rotation field in the south chart. A direct check: ρ is constant along a latitude, so $u'(t) = u(t)/\rho^2$ has velocity $\dot u/\rho^2 = -v/\rho^2 = -v'$. Both routes agree, as they must. (With the orientation-fixed chart $(u', -v')$, the rotation would appear clockwise: the charts disagree about handedness.)

T1.2.3 Covectors and the differential $\mathrm{d}f$

A vector eats a function and returns a number, $v(f)$. Turn this around: fix f and let the vector vary. We get a machine that eats vectors and returns numbers, linearly. That machine is the *differential* of f.

Working definition T1.6 (covector, differential)

A **covector** (or **1-form** at p) is a linear map $\alpha : T_pM \to \mathbb{R}$ [it eats a tangent vector v at p and returns the number $\alpha(v) = \alpha_\mu v^\mu$]. The **differential** of a function f is the covector

$$\mathrm{d}f(v) := v(f) = v^\mu \partial_\mu f. \tag{T1.9}$$

The space of covectors at p is the **cotangent space** T_p^*M. A smooth choice of covector at each point is a **1-form field**, or just a 1-form; the space of them is $\Omega^1(M)$.

Apply (T1.9) to the coordinate functions $f = x^\mu$ themselves:

$$\mathrm{d}x^\mu(\partial_\nu) = \partial_\nu x^\mu = \delta^\mu_\nu. \tag{T1.10}$$

So the $\mathrm{d}x^\mu$ are the *dual basis* to the ∂_ν, and any covector is $\alpha = \alpha_\mu \mathrm{d}x^\mu$ with $\alpha_\mu = \alpha(\partial_\mu)$. In

particular

$$(\mathrm{d}f)_\mu \overset{(1)}{=} \mathrm{d}f(\partial_\mu) \overset{(2)}{=} \partial_\mu f, \qquad \text{so} \qquad \mathrm{d}f = \frac{\partial f}{\partial x^\mu}\,\mathrm{d}x^\mu. \tag{T1.11}$$

(1) Components of a covector are its values on the basis vectors.
(2) Definition (T1.9) with $v = \partial_\mu$, whose only nonzero component is a 1 in slot μ.

This is the formula from first-year calculus, $\mathrm{d}f = f_x\mathrm{d}x + f_y\mathrm{d}y$. Now it has a precise meaning: $\mathrm{d}x$ is not a small number but the linear map "read off the x-component of a vector". Dray calls the vector-calculus version $\mathrm{d}f = \nabla f \cdot \mathrm{d}\vec{r}$ the *master formula* [Dray (13.4)].

Under a change of chart, $\mathrm{d}y^\mu = (\partial y^\mu/\partial x^\nu)\mathrm{d}x^\nu$ (this is (T1.11) applied to $f = y^\mu$), and so the components of $\alpha = \alpha_\mu \mathrm{d}x^\mu = \tilde{\alpha}_\mu \mathrm{d}y^\mu$ obey

$$\alpha_\nu = \frac{\partial y^\mu}{\partial x^\nu}\,\tilde{\alpha}_\mu. \tag{T1.12}$$

This is the **covariant** law: the Jacobian now acts from the other side (equivalently, $\tilde{\alpha}$ is obtained from α with the *inverse* Jacobian). The pairing is chart-independent:

$$\alpha(v) = \alpha_\nu v^\nu \overset{(1)}{=} \tilde{\alpha}_\mu \frac{\partial y^\mu}{\partial x^\nu} v^\nu \overset{(2)}{=} \tilde{\alpha}_\mu \tilde{v}^\mu.$$

(1) Covariant law (T1.12).
(2) Contravariant law (T1.6).

Pictures: a covector is a stack of level surfaces

Dray's picture [Dray §13.8] makes this concrete. Draw the level surfaces $f = c,\ c+1,\ c+2,\ldots$ near p; zoomed in far enough they are parallel, equally spaced planes. Then $\mathrm{d}f(v)$ is *the number of surfaces the arrow v pierces*, counted with sign. Steep functions have closely spaced surfaces and large $\mathrm{d}f$. A vector lying along a level surface pierces none: $\mathrm{d}f(v) = 0$. This is a topographic map, and a 1-form is its linearisation at each point.

Worked example T1.7: $\mathrm{d}f$ in two coordinate systems

Let $f = x^2 + y^2$ on the plane. In Cartesian coordinates $\mathrm{d}f = 2x\,\mathrm{d}x + 2y\,\mathrm{d}y$. In polar coordinates $f = r^2$, so $\mathrm{d}f = 2r\,\mathrm{d}r$. These must be the same covector. Check using $\mathrm{d}x = \cos\phi\,\mathrm{d}r - r\sin\phi\,\mathrm{d}\phi$, $\mathrm{d}y = \sin\phi\,\mathrm{d}r + r\cos\phi\,\mathrm{d}\phi$:

$$\begin{aligned} 2x\,\mathrm{d}x + 2y\,\mathrm{d}y &\overset{(1)}{=} 2r\cos\phi(\cos\phi\,\mathrm{d}r - r\sin\phi\,\mathrm{d}\phi) + 2r\sin\phi(\sin\phi\,\mathrm{d}r + r\cos\phi\,\mathrm{d}\phi) \\ &\overset{(2)}{=} 2r(\cos^2\phi + \sin^2\phi)\,\mathrm{d}r + 2r^2(-\cos\phi\sin\phi + \sin\phi\cos\phi)\,\mathrm{d}\phi \\ &= 2r\,\mathrm{d}r. \end{aligned}$$

(1) Substitute $x = r\cos\phi$, $y = r\sin\phi$ and the expressions for $\mathrm{d}x, \mathrm{d}y$ (themselves instances of (T1.11)).
(2) Collect the $\mathrm{d}r$ and $\mathrm{d}\phi$ terms.

Now evaluate on the vectors of Example T1.5: $\mathrm{d}f(\partial_r) = 2r\,\mathrm{d}r(\partial_r) = 2r$ and $\mathrm{d}f(\partial_\phi) = 0$, exactly as the contour picture (Figure T1.7) predicts.

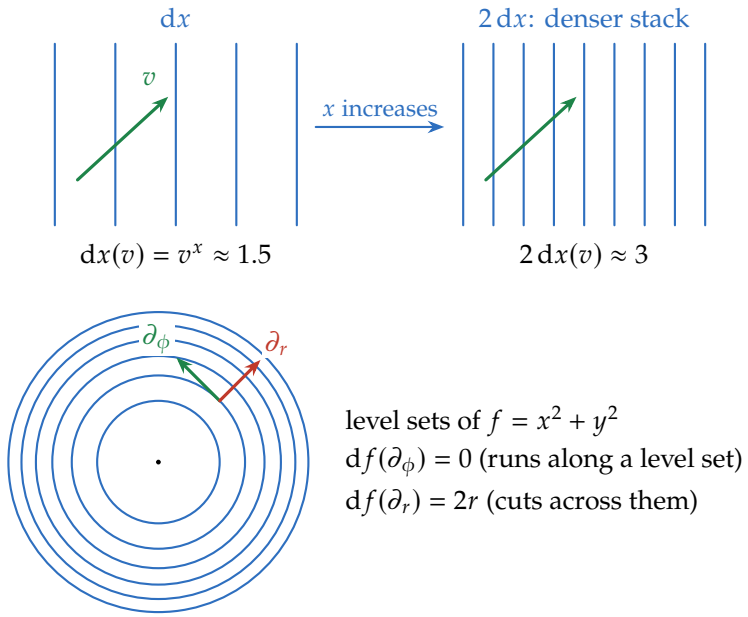


Figure T1.7: Top: the 1-form $\mathrm{d}x$ drawn as a stack of surfaces $x = \text{const}$ at unit spacing; $\mathrm{d}x(v)$ counts how many sheets v crosses. Doubling the form halves the spacing. Bottom: the 1-form $\mathrm{d}f$ for $f = x^2 + y^2$ is the linearised contour map. The rotation vector ∂_ϕ runs along a contour and gets zero; the radial vector cuts across and gets $2r$.

Worked example T1.8: $\mathrm{d}Z$ on the sphere, in both stereographic charts

The height Z is a function on S^2. In the north chart, $Z = (\rho^2-1)/(\rho^2+1) = 1-2/(1+\rho^2)$, so

$$\mathrm{d}Z \overset{(1)}{=} \frac{2\,\mathrm{d}(\rho^2)}{(1+\rho^2)^2} \overset{(2)}{=} \frac{4(u\,\mathrm{d}u + v\,\mathrm{d}v)}{(1+\rho^2)^2}.$$

(1) $\mathrm{d}(-2/w) = 2\,\mathrm{d}w/w^2$ with $w = 1+\rho^2$.
(2) $\mathrm{d}(\rho^2) = \mathrm{d}(u^2+v^2) = 2u\,\mathrm{d}u + 2v\,\mathrm{d}v$.

In the south chart, $Z = (1-\rho'^2)/(1+\rho'^2)$ by the same computation with $Z \to -Z$, so $\mathrm{d}Z = -4\rho'\,\mathrm{d}\rho'/(1+\rho'^2)^2$. On the overlap $\rho' = 1/\rho$ and $\mathrm{d}\rho' = -\mathrm{d}\rho/\rho^2$, hence

$$-\frac{4\rho'\,\mathrm{d}\rho'}{(1+\rho'^2)^2} \overset{(1)}{=} -\frac{4\rho^{-1}(-\rho^{-2}\mathrm{d}\rho)}{(1+\rho^{-2})^2} \overset{(2)}{=} \frac{4\rho^{-3}\,\rho^4\,\mathrm{d}\rho}{(\rho^2+1)^2} = \frac{4\rho\,\mathrm{d}\rho}{(1+\rho^2)^2},$$

(1) Substitute $\rho' = 1/\rho$ and $\mathrm{d}\rho' = -\mathrm{d}\rho/\rho^2$.
(2) Multiply numerator and denominator by ρ^4.

which is the north-chart answer (since $u\,\mathrm{d}u + v\,\mathrm{d}v = \rho\,\mathrm{d}\rho$). On the rotation field $V = -v\partial_u + u\partial_v$ of Example T1.6, $\mathrm{d}Z(V) = 4(u(-v)+vu)/(1+\rho^2)^2 = 0$: moving along a latitude does not change the height.

Physical picture: Wave vectors are covectors

A plane wave $e^{\mathrm{i}\theta(x)}$ has phase $\theta = k_\mu x^\mu$. Its wave "vector" is really $\mathrm{d}\theta = k_\mu \mathrm{d}x^\mu$, a covector: it counts wave crests crossed per unit displacement, exactly like the stack picture. This is why Nash writes the Fourier variable as $k = k_i\,\mathrm{d}x^i$ on p. 40, and why the symbols of differential operators live on the cotangent bundle T^*M (Chapter 2). The rule "$\partial_\mu \to \mathrm{i}k_\mu$" turns a derivative into multiplication by a covector.

T1.2.4 Line elements and the metric

The last ingredient is the one physicists write first: the line element.

$$\mathrm{d}s^2 = g_{\mu\nu}(x)\,\mathrm{d}x^\mu\,\mathrm{d}x^\nu. \tag{T1.13}$$

It is a machine that eats two vectors and returns their dot product, $g(v,w) = g_{\mu\nu}v^\mu w^\nu$, with $g_{\mu\nu} = g_{\nu\mu}$. The products $\mathrm{d}x^\mu\mathrm{d}x^\nu$ in (T1.13) are *symmetric* products ($\mathrm{d}x\,\mathrm{d}y = \mathrm{d}y\,\mathrm{d}x$), not the wedge products of the next section [Dray §6.2]. The length of a curve is $\int\sqrt{g(\dot\gamma,\dot\gamma)}\,\mathrm{d}t$.

Working definition T1.7 (metric)

A (pseudo-)**Riemannian metric** ["pseudo" allows $g(v,v) < 0$ for some nonzero v, as for time-like v in Minkowski space] is a smooth choice of symmetric, non-degenerate ($\det g_{\mu\nu} \neq 0$) bilinear form on each T_pM [the space of tangent vectors v^μ at p]. *In practice:* it is given by a line element $\mathrm{d}s^2$; read off $g_{\mu\nu}$ as the coefficient matrix, remembering that a cross term $2h\,\mathrm{d}x\,\mathrm{d}y$ contributes $g_{xy} = g_{yx} = h$.

Worked example T1.9: The round metric on S^2 in three charts

(a) *Angles.* With $X = \sin\theta\cos\phi$, $Y = \sin\theta\sin\phi$, $Z = \cos\theta$, substituting the differentials into $\mathrm{d}X^2 + \mathrm{d}Y^2 + \mathrm{d}Z^2$ gives the familiar $\mathrm{d}s^2 = \mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2$.
(b) *Stereographic.* Use the complex combination $w = u + \mathrm{i}v$, so that $X + \mathrm{i}Y = 2w/(1+|w|^2)$ and $Z = 1 - 2/(1+|w|^2)$ by (T1.4). Then

$$\begin{aligned}
\mathrm{d}(X+\mathrm{i}Y) &\overset{(1)}{=} \frac{2\big[(1+w\bar w)\mathrm{d}w - w(\bar w\,\mathrm{d}w + w\,\mathrm{d}\bar w)\big]}{(1+|w|^2)^2} = \frac{2(\mathrm{d}w - w^2\mathrm{d}\bar w)}{(1+|w|^2)^2},\\
\mathrm{d}Z &\overset{(2)}{=} \frac{2(\bar w\,\mathrm{d}w + w\,\mathrm{d}\bar w)}{(1+|w|^2)^2},\\
\mathrm{d}X^2+\mathrm{d}Y^2+\mathrm{d}Z^2 &\overset{(3)}{=} \big|\mathrm{d}(X+\mathrm{i}Y)\big|^2 + \mathrm{d}Z^2\\
&\overset{(4)}{=} \frac{4}{(1+|w|^2)^4}\Big[(1+|w|^4)\,\mathrm{d}w\,\mathrm{d}\bar w - \bar w^2\mathrm{d}w^2 - w^2\mathrm{d}\bar w^2\\
&\qquad + \bar w^2\mathrm{d}w^2 + 2|w|^2\mathrm{d}w\,\mathrm{d}\bar w + w^2\mathrm{d}\bar w^2\Big]\\
&= \frac{4(1+|w|^2)^2\,\mathrm{d}w\,\mathrm{d}\bar w}{(1+|w|^2)^4} \overset{(5)}{=} \frac{4(\mathrm{d}u^2+\mathrm{d}v^2)}{(1+u^2+v^2)^2}.
\end{aligned}$$

(1) Quotient rule on $2w/(1+w\bar w)$.
(2) As in Example T1.8, written with $\mathrm{d}(|w|^2) = \bar w\,\mathrm{d}w + w\,\mathrm{d}\bar w$.
(3) $\mathrm{d}X^2+\mathrm{d}Y^2 = (\mathrm{d}X+\mathrm{i}\mathrm{d}Y)(\mathrm{d}X-\mathrm{i}\mathrm{d}Y)$ for symmetric products.
(4) Multiply out $(\mathrm{d}w - w^2\mathrm{d}\bar w)(\mathrm{d}\bar w - \bar w^2\mathrm{d}w)$ and $(\bar w\,\mathrm{d}w + w\,\mathrm{d}\bar w)^2$.
(5) $\mathrm{d}w\,\mathrm{d}\bar w = (\mathrm{d}u+\mathrm{i}\mathrm{d}v)(\mathrm{d}u-\mathrm{i}\mathrm{d}v) = \mathrm{d}u^2+\mathrm{d}v^2$.

The metric is the flat metric times the *conformal factor* $4/(1+\rho^2)^2$: stereographic projection preserves angles.
(c) *Relation between (a) and (b).* In the north chart $\rho^2 = (1+Z)/(1-Z) = \cot^2(\theta/2)$ and the polar angle of (u,v) is ϕ. With $\mathrm{d}\rho = -\tfrac12\csc^2(\theta/2)\,\mathrm{d}\theta$ and $1+\rho^2 = \csc^2(\theta/2)$, $4\,\mathrm{d}\rho^2/(1+\rho^2)^2 = \mathrm{d}\theta^2$ and $4\rho^2\mathrm{d}\phi^2/(1+\rho^2)^2 = 4\cos^2(\theta/2)\sin^2(\theta/2)\,\mathrm{d}\phi^2 = \sin^2\theta\,\mathrm{d}\phi^2$. Same metric, different chart.

The metric on covectors, and raising and lowering

A metric identifies vectors with covectors. Given v, the covector $v^\flat = g(v, \cdot)$ has components $v_\mu = g_{\mu\nu} v^\nu$ ("lowering the index"). Conversely, a covector α gives a vector $\alpha^\sharp$ with components $\alpha^\mu = g^{\mu\nu}\alpha_\nu$, where $g^{\mu\nu}$ is the inverse matrix. The metric on covectors is therefore the inverse matrix:

$$g(\mathrm{d}x^\mu, \mathrm{d}x^\nu) = g^{\mu\nu}. \tag{T1.14}$$

The **gradient vector** is $\nabla f := (\mathrm{d}f)^\sharp$. The gradient is thus a vector only after a metric has been chosen; $\mathrm{d}f$ needs no metric at all.

Worked example T1.10: Why the gradient in polar coordinates has a $1/r$

In polar coordinates $g_{\mu\nu} = \mathrm{diag}(1, r^2)$, so $g^{\mu\nu} = \mathrm{diag}(1, r^{-2})$. Then

$$\nabla f = (\mathrm{d}f)^\sharp \overset{(1)}{=} g^{rr}\partial_r f\, \partial_r + g^{\phi\phi}\partial_\phi f\, \partial_\phi = \partial_r f\, \partial_r + \frac{1}{r^2}\partial_\phi f\, \partial_\phi$$
$$\overset{(2)}{=} \partial_r f\, \hat{r} + \frac{1}{r}\,\partial_\phi f\, \hat{\phi}.$$

(1) Raise the index of $(\mathrm{d}f)_\mu = \partial_\mu f$ with the diagonal inverse metric.

(2) Use the unit vectors $\hat{r} = \partial_r$, $\hat{\phi} = \frac{1}{r}\partial_\phi$ from Example T1.5.

This is the formula in every electrodynamics textbook, and now we see where the $1/r$ comes from: one power from the inverse metric, partly undone by normalising ∂_ϕ. Also, $g(\mathrm{d}r, \mathrm{d}r) = 1$, $g(\mathrm{d}\phi, \mathrm{d}\phi) = 1/r^2$, $g(\mathrm{d}r, \mathrm{d}\phi) = 0$: the *orthonormal coframe* is $\{\mathrm{d}r,\ r\,\mathrm{d}\phi\}$, the two "pieces" of $\mathrm{d}\vec{r} = \mathrm{d}r\,\hat{r} + r\,\mathrm{d}\phi\,\hat{\phi}$ [Dray (13.68)].

Watch out: Is $\mathrm{d}x$ small or not?

In calculus $\mathrm{d}x$ is "an infinitesimal change". Here $\mathrm{d}x$ is a linear map with $\mathrm{d}x(\partial_x) = 1$, not small at all. Both are right [Dray §13.9]: the 1-form lives at a single point (as zoomed-in as one can get), and at that point it behaves like a unit object. A line integral $\int_\gamma \alpha$ feeds the 1-form the velocity $\dot{\gamma}$ and multiplies by $\mathrm{d}t$; the smallness lives in $\mathrm{d}t$.

Reading Nash (p. 15 and p. 40)

Nash's tangent bundle $TM = \bigcup_p T_pM$ (p. 15) is the collection of the spaces T_pM we built above; the cotangent bundle T^*M is the collection of the T_p^*M. On p. 40 he writes the Fourier variable as $k = k_i\,\mathrm{d}x^i$ and uses it as a point of T^*M. Every time Nash says "cotangent", think "stack of level surfaces", "$\mathrm{d}f$", or "wave covector".

Recap

- A tangent vector is a velocity $v^\mu = \dot{x}^\mu$, equivalently the derivation $v = v^\mu\partial_\mu$; components transform with the Jacobian.
- A covector is a linear function on vectors. The differential $\mathrm{d}f(v) = v(f)$ has components $\partial_\mu f$; $\mathrm{d}x^\mu(\partial_\nu) = \delta^\mu_\nu$.
- Picture: a covector is a stack of level surfaces; $\alpha(v)$ counts sheets crossed.
- The line element $\mathrm{d}s^2 = g_{\mu\nu}\mathrm{d}x^\mu\mathrm{d}x^\nu$ is a metric on vectors; its inverse $g^{\mu\nu}$ is the metric on covectors.

- S^2: $\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2 = 4(\mathrm{d}u^2 + \mathrm{d}v^2)/(1+\rho^2)^2$.

T1.3 The wedge product and the algebra of forms

A 1-form is the integrand of a line integral. What is the integrand of a surface integral, or of a volume integral? In calculus we write $\mathrm{d}x\,\mathrm{d}y$ and $\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z$, and treat them as if they were ordinary products of small numbers. That turns out to be almost right. The one correction, which Dray extracts from the change-of-variables formula [Dray §11.3–11.4], is that the product must be *antisymmetric*. Once we accept that, everything else follows: oriented areas, determinants, the cross product, and a clean way to count how many independent integrands of each kind a space has.

T1.3.1 Why the product of differentials is antisymmetric

Consider a change of variables $x = x(u,v)$, $y = y(u,v)$ in the plane. The area element transforms with the Jacobian determinant,

$$\mathrm{d}x\,\mathrm{d}y = \frac{\partial(x,y)}{\partial(u,v)}\,\mathrm{d}u\,\mathrm{d}v, \qquad \frac{\partial(x,y)}{\partial(u,v)} = \det\begin{pmatrix} \partial_u x & \partial_v x \\ \partial_u y & \partial_v y \end{pmatrix}. \tag{T1.15}$$

In a first calculus course we put absolute value bars around the determinant. Let us *not* do that, and keep track of orientation instead: an integral of a flux changes sign when the orientation of the surface is reversed, so the sign carries information. Written the other way round, (T1.15) says $\mathrm{d}u\,\mathrm{d}v = \frac{\partial(u,v)}{\partial(x,y)}\,\mathrm{d}x\,\mathrm{d}y$ for *any* two functions u, v. Now make two silly choices.

$$u = y,\ v = x: \qquad \mathrm{d}y\,\mathrm{d}x \overset{(1)}{=} \frac{\partial(y,x)}{\partial(x,y)}\,\mathrm{d}x\,\mathrm{d}y \overset{(2)}{=} \det\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\mathrm{d}x\,\mathrm{d}y = -\,\mathrm{d}x\,\mathrm{d}y,$$

$$u = v = x: \qquad \mathrm{d}x\,\mathrm{d}x \overset{(3)}{=} \frac{\partial(x,x)}{\partial(x,y)}\,\mathrm{d}x\,\mathrm{d}y \overset{(4)}{=} \det\begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}\mathrm{d}x\,\mathrm{d}y = 0.$$

(1) The transformation rule with $u = y$, $v = x$.
(2) The rows of the Jacobian matrix are the gradients of $u = y$ and $v = x$: $(\partial_x y, \partial_y y) = (0,1)$ and $(1,0)$. The determinant is -1.
(3) The transformation rule with $u = v = x$.
(4) Both rows are the gradient $(1,0)$ of x; equal rows make the determinant vanish.

So if we want the change-of-variables formula to hold *with* signs, products of differentials must obey

$$\mathrm{d}y \wedge \mathrm{d}x = -\,\mathrm{d}x \wedge \mathrm{d}y, \qquad \mathrm{d}x \wedge \mathrm{d}x = 0. \tag{T1.16}$$

We use the symbol $\wedge$ ("wedge") to distinguish this product from the symmetric product $\mathrm{d}x\,\mathrm{d}y = \mathrm{d}y\,\mathrm{d}x$ that appears in line elements.

The two rules in (T1.16) are equivalent. Antisymmetry gives $\mathrm{d}x \wedge \mathrm{d}x = -\mathrm{d}x \wedge \mathrm{d}x$, so it is zero. Conversely, if every 1-form squares to zero then

$$0 \overset{(1)}{=} (\mathrm{d}x + \mathrm{d}y) \wedge (\mathrm{d}x + \mathrm{d}y) \overset{(2)}{=} \mathrm{d}x \wedge \mathrm{d}x + \mathrm{d}x \wedge \mathrm{d}y + \mathrm{d}y \wedge \mathrm{d}x + \mathrm{d}y \wedge \mathrm{d}y \overset{(3)}{=} \mathrm{d}x \wedge \mathrm{d}y + \mathrm{d}y \wedge \mathrm{d}x,$$

(1) $\mathrm{d}x + \mathrm{d}y$ is a 1-form, so its square vanishes.
(2) Distributivity.
(3) The squares $\mathrm{d}x \wedge \mathrm{d}x$ and $\mathrm{d}y \wedge \mathrm{d}y$ vanish.

which is antisymmetry. (This trick is called **polarisation**.)

T1.3.2 What a 2-form does to two vectors: oriented area

A 1-form eats one vector. A product of two 1-forms should eat two vectors, and the natural antisymmetric way to do it is a 2×2 determinant.

Working definition T1.8 (wedge product of two 1-forms)

For 1-forms α, β [covectors: each eats one vector v and returns the number $\alpha(v) = \alpha_\mu v^\mu$] the **wedge product** $\alpha \wedge \beta$ is the machine that eats an ordered pair of vectors and returns

$$(\alpha \wedge \beta)(v, w) = \det \begin{pmatrix} \alpha(v) & \alpha(w) \\ \beta(v) & \beta(w) \end{pmatrix} = \alpha(v)\beta(w) - \alpha(w)\beta(v). \tag{T1.17}$$

It is bilinear [linear in v with w held fixed, and linear in w with v held fixed] and antisymmetric in (v, w) [swapping v and w flips the sign, so $(\alpha \wedge \beta)(v, v) = 0$], and antisymmetric in (α, β).

For the basis forms of the plane, $(dx \wedge dy)(v, w) = v^x w^y - v^y w^x$. That is the signed area of the parallelogram spanned by v and w: positive if w is counter-clockwise from v, negative if clockwise, zero if they are parallel. In three dimensions, $(dx \wedge dy)(v, w)$ is the signed area of the *shadow* of the parallelogram on the xy-plane (Figure T1.8).

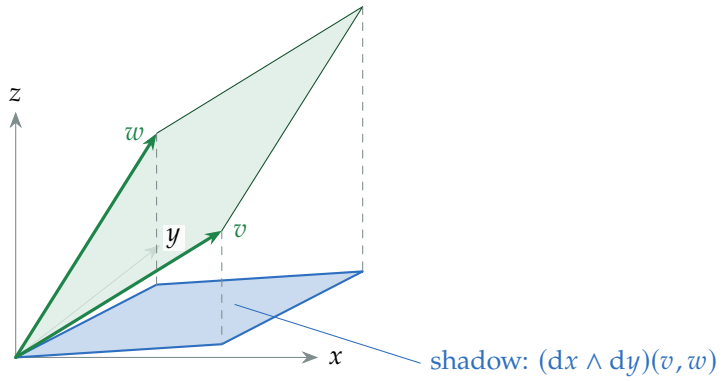


Figure T1.8: The 2-form $dx \wedge dy$ applied to two vectors v, w in $\mathbb{R}^3$ returns the signed area of the shadow (blue) of their parallelogram (green) on the xy-plane. Likewise $dy \wedge dz$ and $dz \wedge dx$ measure the shadows on the other two coordinate planes; the three numbers together are the components of $v \times w$.

Worked example T1.11: The three shadows are the cross product

Take $v = (v^x, v^y, v^z)$ and $w = (w^x, w^y, w^z)$ in $\mathbb{R}^3$. Then by (T1.17) with $dx^\mu(v) = v^\mu$,

$$\begin{aligned} (dy \wedge dz)(v, w) &= v^y w^z - v^z w^y = (v \times w)^x, \\ (dz \wedge dx)(v, w) &= v^z w^x - v^x w^z = (v \times w)^y, \\ (dx \wedge dy)(v, w) &= v^x w^y - v^y w^x = (v \times w)^z. \end{aligned}$$

With numbers: $v = (1, 0, 2)$, $w = (0, 3, 1)$ give shadows $(0 \cdot 1 - 2 \cdot 3,\ 2 \cdot 0 - 1 \cdot 1,\ 1 \cdot 3 - 0 \cdot 0) = (-6, -1, 3)$, which is $v \times w$. The cyclic order (yz, zx, xy) is what makes this match the right-hand rule; this is why Dray [Dray (12.24)] writes the flux 2-form as $F_x\, dy \wedge dz + F_y\, dz \wedge dx + F_z\, dx \wedge dy$.

T1.3.3 Changing coordinates in a 2-form

Wedge products let us change variables by pure algebra: expand each differential by the chain rule, multiply out, and use (T1.16). The Jacobian determinant appears by itself.

Worked example T1.12: $\mathrm{d}x \wedge \mathrm{d}y = r\,\mathrm{d}r \wedge \mathrm{d}\phi$

With $\mathrm{d}x = \cos\phi\,\mathrm{d}r - r\sin\phi\,\mathrm{d}\phi$ and $\mathrm{d}y = \sin\phi\,\mathrm{d}r + r\cos\phi\,\mathrm{d}\phi$,

$$\begin{aligned}\mathrm{d}x \wedge \mathrm{d}y &\overset{(1)}{=} \cos\phi\sin\phi\;\mathrm{d}r\wedge\mathrm{d}r + r\cos^2\phi\;\mathrm{d}r\wedge\mathrm{d}\phi \\ &\quad - r\sin^2\phi\;\mathrm{d}\phi\wedge\mathrm{d}r - r^2\sin\phi\cos\phi\;\mathrm{d}\phi\wedge\mathrm{d}\phi \\ &\overset{(2)}{=} r\cos^2\phi\;\mathrm{d}r\wedge\mathrm{d}\phi - r\sin^2\phi\;\mathrm{d}\phi\wedge\mathrm{d}r \\ &\overset{(3)}{=} r(\cos^2\phi+\sin^2\phi)\,\mathrm{d}r\wedge\mathrm{d}\phi = r\,\mathrm{d}r\wedge\mathrm{d}\phi. \end{aligned} \tag{T1.18}$$

(1) Distribute the product of the two brackets; four terms.
(2) $\mathrm{d}r\wedge\mathrm{d}r = 0 = \mathrm{d}\phi\wedge\mathrm{d}\phi$.
(3) $\mathrm{d}\phi\wedge\mathrm{d}r = -\mathrm{d}r\wedge\mathrm{d}\phi$.

The coefficient r is the Jacobian $\partial(x,y)/\partial(r,\phi)$, obtained without ever writing a determinant. We can also go the other way, following Dray [Dray (13.24)–(13.27)]: from $r^2 = x^2+y^2$ and $\tan\phi = y/x$ we get $r\,\mathrm{d}r = x\,\mathrm{d}x + y\,\mathrm{d}y$ and $r^2\mathrm{d}\phi = x\,\mathrm{d}y - y\,\mathrm{d}x$. Then

$$\begin{aligned} r^3\,\mathrm{d}r\wedge\mathrm{d}\phi &= (x\,\mathrm{d}x + y\,\mathrm{d}y)\wedge(x\,\mathrm{d}y - y\,\mathrm{d}x) \\ &\overset{(1)}{=} x^2\,\mathrm{d}x\wedge\mathrm{d}y - y^2\,\mathrm{d}y\wedge\mathrm{d}x \overset{(2)}{=} (x^2+y^2)\,\mathrm{d}x\wedge\mathrm{d}y = r^2\,\mathrm{d}x\wedge\mathrm{d}y, \end{aligned}$$

(1) The cross terms $-xy\,\mathrm{d}x\wedge\mathrm{d}x$ and $xy\,\mathrm{d}y\wedge\mathrm{d}y$ vanish.
(2) $-\mathrm{d}y\wedge\mathrm{d}x = +\mathrm{d}x\wedge\mathrm{d}y$.

which is (T1.18) again after dividing by r^2.

Key idea: Change of variables is automatic

For any change of coordinates $y^\mu = y^\mu(x)$ in n dimensions,

$$\mathrm{d}y^1\wedge\mathrm{d}y^2\wedge\cdots\wedge\mathrm{d}y^n = \det\!\left(\frac{\partial y^\mu}{\partial x^\nu}\right)\mathrm{d}x^1\wedge\mathrm{d}x^2\wedge\cdots\wedge\mathrm{d}x^n. \tag{T1.19}$$

The Jacobian determinant of multivariable calculus is *defined* by this equation, and its sign records whether the two coordinate systems have the same handedness.

Why: expand each $\mathrm{d}y^\mu = (\partial y^\mu/\partial x^\nu)\mathrm{d}x^\nu$ and multiply out. Every term with a repeated $\mathrm{d}x^\nu$ dies. The surviving terms have each $\mathrm{d}x^\nu$ exactly once, in some order σ; reordering them to $\mathrm{d}x^1\wedge\cdots\wedge\mathrm{d}x^n$ costs the sign $\operatorname{sgn}\sigma$. The coefficient is $\sum_\sigma \operatorname{sgn}\sigma \prod_\mu \partial y^\mu/\partial x^{\sigma(\mu)}$, which is the Leibniz formula for the determinant. For the S^2 stereographic charts, Example T1.3 therefore says $\mathrm{d}u'\wedge\mathrm{d}v' = -\rho^{-4}\,\mathrm{d}u\wedge\mathrm{d}v$: the two charts have opposite orientations.

Worked example T1.13: Spherical coordinates

With $x = r\sin\theta\cos\phi$, $y = r\sin\theta\sin\phi$, $z = r\cos\theta$, first write the three differentials in the orthonormal-looking combinations $\mathrm{d}r$, $r\,\mathrm{d}\theta$, $r\sin\theta\,\mathrm{d}\phi$:

$$\mathrm{d}x = \sin\theta\cos\phi\,\mathrm{d}r + \cos\theta\cos\phi\,(r\,\mathrm{d}\theta) - \sin\phi\,(r\sin\theta\,\mathrm{d}\phi),$$

$$\begin{aligned} \mathrm{d}y &= \sin\theta\sin\phi\,\mathrm{d}r + \cos\theta\sin\phi\,(r\,\mathrm{d}\theta) + \cos\phi\,(r\sin\theta\,\mathrm{d}\phi), \\ \mathrm{d}z &= \cos\theta\,\mathrm{d}r - \sin\theta\,(r\,\mathrm{d}\theta). \end{aligned}$$

The coefficient matrix R (rows: $\mathrm{d}x, \mathrm{d}y, \mathrm{d}z$; columns: $\mathrm{d}r, r\mathrm{d}\theta, r\sin\theta\mathrm{d}\phi$) is a rotation matrix: its columns are the unit vectors $\hat{r}, \hat{\theta}, \hat{\phi}$. Therefore

$$\begin{aligned} \mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z &\overset{(1)}{=} \det R\; \mathrm{d}r \wedge (r\,\mathrm{d}\theta) \wedge (r\sin\theta\,\mathrm{d}\phi) \\ &\overset{(2)}{=} r^2 \sin\theta\; \mathrm{d}r \wedge \mathrm{d}\theta \wedge \mathrm{d}\phi. \end{aligned}$$

(1) (T1.19) applied to the three 1-forms $\mathrm{d}r$, $r\mathrm{d}\theta$, $r\sin\theta\mathrm{d}\phi$ (the argument works for any three 1-forms, not only coordinate differentials).

(2) $\det R = +1$ because $(\hat{r}, \hat{\theta}, \hat{\phi})$ is a right-handed orthonormal triad (check at $\theta = \pi/2, \phi = 0$: $\hat{r} = \hat{x}$, $\hat{\theta} = -\hat{z}$, $\hat{\phi} = \hat{y}$, and $\hat{x} \times (-\hat{z}) = \hat{y}$). Pull the scalar factors r and $r\sin\theta$ out.

T1.3.4 Higher-rank forms and dimension counting

Nothing stops us from multiplying more 1-forms. Products of p of them are p-forms; they are the integrands of p-dimensional integrals.

Working definition T1.9 (p-form)

A **p-form** at a point is a linear combination of products $\mathrm{d}x^{\mu_1} \wedge \cdots \wedge \mathrm{d}x^{\mu_p}$. It eats p vectors, multilinearly and totally antisymmetrically [linear in each vector separately, and the sign flips whenever two of the p vectors are swapped], by the $p \times p$ determinant rule

$$(\alpha^1 \wedge \cdots \wedge \alpha^p)(v_1, \ldots, v_p) = \det\big[\alpha^i(v_j)\big]. \tag{T1.20}$$

A **p-form field** (or just p-form) has function coefficients [$\alpha = \sum_{\mu_1<\cdots<\mu_p} \alpha_{\mu_1\ldots\mu_p}(x)\,\mathrm{d}x^{\mu_1} \wedge \cdots \wedge \mathrm{d}x^{\mu_p}$, with smooth $\alpha_{\mu_1\ldots\mu_p}(x)$]. The space of them is $\Omega^p(M)$. By convention $\Omega^0(M)$ is the space of functions.

A basis of p-forms at a point is given by the products with *increasing* indices, $\{\mathrm{d}x^{\mu_1}\wedge\cdots\wedge\mathrm{d}x^{\mu_p} : \mu_1 < \mu_2 < \cdots < \mu_p\}$. Any other ordering is $\pm$ one of these, and any repetition is zero. So the dimension is the number of ways to choose p distinct indices out of n:

$$\dim \Omega^p_x = \binom{n}{p} = \frac{n!}{p!\,(n-p)!}, \qquad \dim \Omega^p_x = 0 \;\; (p > n). \tag{T1.21}$$

Two facts jump out of Figure T1.9. First, $\binom{n}{p} = \binom{n}{n-p}$: p-forms and $(n-p)$-forms come in equal numbers. The Hodge star of Section T1.4 is the isomorphism between them. Second, the top forms, $p = n$, form a one-dimensional space: every n-form is a function times $\mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^n$. Top forms are the things we integrate over the whole manifold. The total number of independent forms of all ranks is $\sum_p \binom{n}{p} = 2^n$.

Components and the factor $1/p!$

A general p-form can be written in two ways:

$$\alpha = \sum_{\mu_1<\cdots<\mu_p} \alpha_{\mu_1\ldots\mu_p}\,\mathrm{d}x^{\mu_1} \wedge \cdots \wedge \mathrm{d}x^{\mu_p} = \frac{1}{p!}\,\alpha_{\mu_1\ldots\mu_p}\,\mathrm{d}x^{\mu_1} \wedge \cdots \wedge \mathrm{d}x^{\mu_p}, \tag{T1.22}$$

$n\backslash p$	**0**	**1**	**2**	**3**	**4**	**total**
1	1	1	0	0	0	2
2	1	2	1	0	0	4
3	1	3	3	1	0	8
4	1	4	6	4	1	16

$n = 3$: 1-forms and 2-forms both look like "vectors"
$n = 4$: six 2-forms = $\vec{E}$ and $\vec{B}$ together

Figure T1.9: Dimensions $\binom{n}{p}$ of the space of p-forms at a point. Read along a row: the numbers are symmetric under $p \leftrightarrow n-p$ (this is Hodge duality), the last nonzero entry is always 1 (top forms are multiples of a single volume form), and the row sums to 2^n. The coloured cells explain two pieces of physics: why vector calculus in $\mathbb{R}^3$ can get away with one kind of vector, and why the electromagnetic field in space-time has six components.

where in the second expression all indices are summed freely and the components $\alpha_{\mu_1\ldots\mu_p}$ are defined for all index values and taken *totally antisymmetric*. The $p!$ orderings of each index set contribute equal terms, and the $1/p!$ removes the overcounting. With this normalisation the components are the values of α on basis vectors:

$$\alpha(\partial_{\nu_1},\ldots,\partial_{\nu_p}) \overset{(1)}{=} \frac{1}{p!}\,\alpha_{\mu_1\ldots\mu_p}\,\det\!\left[\mathrm{d}x^{\mu_i}(\partial_{\nu_j})\right] \overset{(2)}{=} \frac{1}{p!}\,\alpha_{\mu_1\ldots\mu_p}\,\det\!\left[\delta^{\mu_i}_{\nu_j}\right]$$
$$\overset{(3)}{=} \frac{1}{p!}\sum_{\sigma}\operatorname{sgn}\sigma\;\alpha_{\nu_{\sigma(1)}\ldots\nu_{\sigma(p)}} \overset{(4)}{=} \frac{1}{p!}\,p!\,\alpha_{\nu_1\ldots\nu_p} = \alpha_{\nu_1\ldots\nu_p}.$$

(1) The determinant rule (T1.20) for each term.
(2) Dual basis (T1.10).
(3) The determinant of the matrix $\delta^{\mu_i}_{\nu_j}$ is nonzero only when $(\mu_1,\ldots,\mu_p)$ is a permutation σ of $(\nu_1,\ldots,\nu_p)$, and then equals $\operatorname{sgn}\sigma$.
(4) Antisymmetry of the components: each of the $p!$ terms equals $\alpha_{\nu_1\ldots\nu_p}$.

More generally, by multilinearity, $\alpha(v_1,\ldots,v_p) = \alpha_{\mu_1\ldots\mu_p}v_1^{\mu_1}\cdots v_p^{\mu_p}$. For a 2-form this reads $\alpha(v,w) = \alpha_{\mu\nu}v^\mu w^\nu$, and the antisymmetric matrix $\alpha_{\mu\nu}$ is the *representation* of the 2-form. For example $\mathrm{d}x\wedge\mathrm{d}y$ in the plane has matrix $\left(\begin{smallmatrix}0&1\\-1&0\end{smallmatrix}\right)$, and the electromagnetic field strength $F_{\mu\nu}$ is the 4×4 antisymmetric matrix of a 2-form in space-time (Section T1.7).

Watch out: Two normalisations of components

Many physics texts and Nash himself [Nash p. 39, (2.43)] write $\omega = \omega_{i_1\ldots i_p}\,\mathrm{d}x^{i_1}\wedge\cdots\wedge\mathrm{d}x^{i_p}$ with a free sum and *no* $1/p!$. Their components are then $1/p!$ times ours. Neither is wrong; but formulas with explicit factorials (the Hodge star, the inner product) must be read in the convention they were written in. We always use (T1.22).

T1.3.5 Multiplying forms of any rank

The wedge product of a p-form and a q-form is defined by multiplying out in the basis and using (T1.16). It is associative and distributive, and it is **graded commutative**:

$$\beta\wedge\alpha = (-1)^{pq}\,\alpha\wedge\beta \qquad (\alpha\in\Omega^p,\ \beta\in\Omega^q). \tag{T1.23}$$

Why: move each of the q one-form factors of β past each of the p factors of α; each swap costs a sign; there are pq swaps [Dray (13.50)]. So 1-forms anticommute with each other, but a 2-form commutes with everything.

Worked example T1.14: A 1-form times a 2-form is the dot product

In $\mathbb{R}^3$ let $v = v_x\,dx + v_y\,dy + v_z\,dz$ and $w = w_x\,dy\wedge dz + w_y\,dz\wedge dx + w_z\,dx\wedge dy$. Then

$$\begin{aligned} v\wedge w &\overset{(1)}{=} v_x w_x\,dx\wedge dy\wedge dz + v_y w_y\,dy\wedge dz\wedge dx + v_z w_z\,dz\wedge dx\wedge dy \\ &\overset{(2)}{=} (v_x w_x + v_y w_y + v_z w_z)\,dx\wedge dy\wedge dz. \end{aligned}$$

(1) Of the nine products, only those in which the 1-form supplies the differential missing from the 2-form survive; the other six contain a repeated differential.
(2) $dy\wedge dz\wedge dx$ and $dz\wedge dx\wedge dy$ are cyclic permutations of $dx\wedge dy\wedge dz$; a cyclic permutation of three objects is two swaps, so no sign.

So "vector $\wedge$ flux" is "dot product $\times$ volume" [Dray (13.46)]. And two 1-forms give the cross product as a 2-form: $v\wedge u = (v_y u_z - v_z u_y)\,dy\wedge dz + \ldots$, the content of Example T1.11. The triple product is $u\wedge v\wedge w = \det[u\,v\,w]\,dx\wedge dy\wedge dz$ for three 1-forms.

In components, the product of a p-form and a q-form is the antisymmetrised tensor product:

$$(\alpha\wedge\beta)_{\mu_1\ldots\mu_{p+q}} = \frac{(p+q)!}{p!\,q!}\,\alpha_{[\mu_1\ldots\mu_p}\,\beta_{\mu_{p+1}\ldots\mu_{p+q}]}, \tag{T1.24}$$

where the square bracket means "average over all permutations with signs". For two 1-forms this is $(\alpha\wedge\beta)_{\mu\nu} = \alpha_\mu\beta_\nu - \alpha_\nu\beta_\mu$, which agrees with (T1.17). In the language of tensors, $\alpha\wedge\beta = \alpha\otimes\beta - \beta\otimes\alpha$ for 1-forms, and more generally a p-form is exactly a totally antisymmetric covariant tensor of rank p [Dray §13.9].

Worked example T1.15: A 2-form whose square is not zero

Graded commutativity gives $\alpha\wedge\alpha = -\alpha\wedge\alpha = 0$ for odd p. For even p there is no such argument, and indeed in $\mathbb{R}^4$ with $\omega = dx^1\wedge dx^2 + dx^3\wedge dx^4$,

$$\begin{aligned} \omega\wedge\omega &\overset{(1)}{=} dx^1\wedge dx^2\wedge dx^3\wedge dx^4 + dx^3\wedge dx^4\wedge dx^1\wedge dx^2 \\ &\overset{(2)}{=} 2\,dx^1\wedge dx^2\wedge dx^3\wedge dx^4 \neq 0. \end{aligned}$$

(1) Expand; the terms $(dx^1\wedge dx^2)^{\wedge 2}$ and $(dx^3\wedge dx^4)^{\wedge 2}$ contain repeated factors.
(2) 2-forms commute (T1.23) with $pq = 4$.

A 2-form that can be written as a single product $\alpha\wedge\beta$ of 1-forms is called **decomposable**; it obviously squares to zero. For 2-forms the converse holds too (in any dimension), so "$\omega\wedge\omega = 0$" is the test for a 2-form to be decomposable. For odd forms the test is useless, since they always square to zero: in $\mathbb{R}^6$ the 3-form $dx^1\wedge dx^2\wedge dx^3 + dx^4\wedge dx^5\wedge dx^6$ squares to zero yet is not a single product, which answers the question Dray poses in [Dray §14.15]. This ω fails it. The same computation with the Maxwell field strength gives $F\wedge F \propto \vec{E}\cdot\vec{B}$ (Section T1.7), and with a Yang–Mills field it gives the instanton density $\operatorname{tr} F\wedge F$ of Chapter 8. Squares of 2-forms are where topology lives in four dimensions.

T1.3.6 Top forms and determinants

Here is Dray's elegant use of top forms [Dray §13.4]. A linear map A acting on 1-forms, $A(dx^i) = a^i{}_j\,dx^j$, extends to p-forms by $A(\alpha^1\wedge\cdots\wedge\alpha^p) = A\alpha^1\wedge\cdots\wedge A\alpha^p$. On the one-dimensional space of top forms it can only be multiplication by a number. That number is the determinant:

$$A\,(dx^1\wedge\cdots\wedge dx^n) = \det A\;dx^1\wedge\cdots\wedge dx^n. \tag{T1.25}$$

Worked example T1.16: 2×2 and 3×3 determinants by wedging

In two dimensions,

$$(a\,\mathrm{d}x + b\,\mathrm{d}y)\wedge(c\,\mathrm{d}x + d\,\mathrm{d}y) \overset{(1)}{=} ad\,\mathrm{d}x\wedge\mathrm{d}y + bc\,\mathrm{d}y\wedge\mathrm{d}x \overset{(2)}{=} (ad-bc)\,\mathrm{d}x\wedge\mathrm{d}y.$$

(1) The $ac\,\mathrm{d}x\wedge\mathrm{d}x$ and $bd\,\mathrm{d}y\wedge\mathrm{d}y$ terms vanish.
(2) Antisymmetry.

In three dimensions, with rows $(1,2,0)$, $(0,1,3)$, $(2,0,1)$:

$$\begin{aligned}(\mathrm{d}x+2\mathrm{d}y)\wedge(\mathrm{d}y+3\mathrm{d}z)\wedge(2\mathrm{d}x+\mathrm{d}z) &\overset{(1)}{=} (\mathrm{d}x\wedge\mathrm{d}y + 3\,\mathrm{d}x\wedge\mathrm{d}z + 6\,\mathrm{d}y\wedge\mathrm{d}z)\wedge(2\mathrm{d}x+\mathrm{d}z)\\ &\overset{(2)}{=} \mathrm{d}x\wedge\mathrm{d}y\wedge\mathrm{d}z + 12\,\mathrm{d}y\wedge\mathrm{d}z\wedge\mathrm{d}x\\ &\overset{(3)}{=} 13\,\mathrm{d}x\wedge\mathrm{d}y\wedge\mathrm{d}z.\end{aligned}$$

(1) Multiply the first two factors; $2\,\mathrm{d}y\wedge\mathrm{d}y = 0$.
(2) Only terms with three different differentials survive: $\mathrm{d}x\wedge\mathrm{d}y\wedge\mathrm{d}z$ (coefficient $1\cdot1$) and $6\,\mathrm{d}y\wedge\mathrm{d}z\wedge2\mathrm{d}x$; the term $3\,\mathrm{d}x\wedge\mathrm{d}z\wedge(\dots)$ needs $\mathrm{d}y$, which the last factor does not contain.
(3) Cyclic permutation, no sign.

Indeed $\det\left(\begin{smallmatrix}1&2&0\\0&1&3\\2&0&1\end{smallmatrix}\right) = 1(1) - 2(0-6) + 0 = 13$.

The product rule for determinants is now a one-liner [Dray (13.33)]: $\det(AB)\,\omega = A(B\omega) = A(\det B\,\omega) = \det B\,\det A\,\omega$ for any top form ω. No index gymnastics, and the definition (T1.25) manifestly does not depend on the choice of basis.

T1.3.7 The three-dimensional dictionary

We can now say precisely what the "vectors" of vector calculus are. In $\mathbb{R}^3$ there are four kinds of integrand, and each corresponds to a kind of field [Dray §12.2].

Dictionary: forms in $\mathbb{R}^3$ and vector calculus

rank	form	integrand of	physicist's name
0	f	evaluation at a point	scalar
1	$\vec{F}\cdot\mathrm{d}\vec{r} = F_x\mathrm{d}x + F_y\mathrm{d}y + F_z\mathrm{d}z$	line integral (work)	(polar) vector
2	$\vec{F}\cdot\mathrm{d}\vec{A} = F_x\mathrm{d}y\wedge\mathrm{d}z + \dots$	surface integral (flux)	pseudovector
3	$f\,\mathrm{d}V = f\,\mathrm{d}x\wedge\mathrm{d}y\wedge\mathrm{d}z$	volume integral	density

Products: 1-form $\wedge$ 1-form $\leftrightarrow$ cross product; 1-form $\wedge$ 2-form $\leftrightarrow$ dot product; 1-form $\wedge$ 1-form $\wedge$ 1-form $\leftrightarrow$ triple product.

Why "pseudo"? Under the parity map $P:\vec{x}\mapsto-\vec{x}$ every $\mathrm{d}x^i$ changes sign, so a basis p-form picks up $(-1)^p$. For a field to be parity-even as a form, its components must therefore change by $(-1)^p$. A 1-form's components flip, like the components of the electric field or a velocity: a *polar* vector. A 2-form's components do not flip, like angular momentum $\vec{r}\times\vec{p}$ or the magnetic field: an *axial* or pseudo-vector. The old rule "the cross product of two vectors is a pseudovector" becomes "the wedge product of two 1-forms is a 2-form". Dimensional analysis agrees [Dray §13.5]: $|v\times w|$ is an area, not a length. The wedge product is associative while the

cross product is not, precisely because the wedge keeps track of where each product lives.

T1.3.8 Pictures of forms

Pictures of forms help enormously, and Dray draws them carefully [Dray §13.8]. We met the basic one in Figure T1.7: a 1-form at a point is an oriented *stack* of parallel surfaces, and its value on a vector is the number of sheets pierced. Two further rules make the picture a calculus.

Adding stacks. To draw $\alpha + \beta$, superimpose the stacks of α and β and draw new sheets through the points where a sheet of α meets a sheet of β with labels summing to an integer (Figure T1.10, left). For $a\,dx + b\,dy$ with constants a, b the new sheets are the lines $ax + by =$ const, spaced by $1/\sqrt{a^2+b^2}$: a steeper function has denser contours.

Fields of stacks. A 1-form field has a small stack at each point. When $\alpha = df$ the little stacks fit together into the contour map of f: for $r\,dr = x\,dx + y\,dy = \frac{1}{2}d(r^2)$ they are pieces of circles. But when α is not a differential the stacks need not fit together at all. For $r^2\,d\phi = x\,dy - y\,dx$ the sheets are radial and their spacing shrinks like $1/r^2$ as we move out; following them around the origin one comes back having crossed ever more sheets (Figure T1.10, right). There is no global contour map, because $r^2 d\phi$ is not d of any function. We will make this precise with the exterior derivative in Section T1.5.

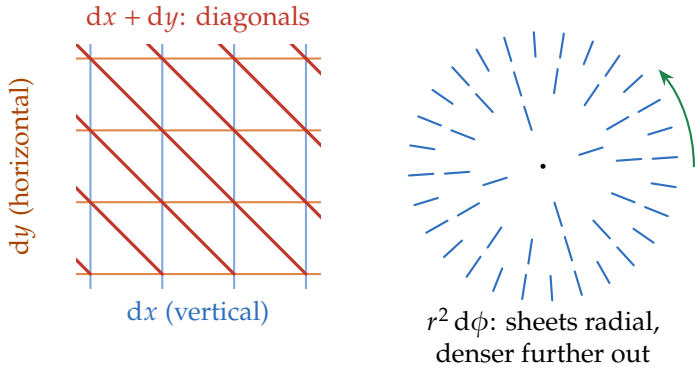


Figure T1.10: Left: adding the stacks dx and dy. The sheets of the sum (red) pass through the corners where the two old stacks cross, and are spaced by $1/\sqrt{2}$. Right: the 1-form field $r^2 d\phi = x\,dy - y\,dx$ drawn as small stacks along three circles. The number of sheets crossed per unit length along a circle grows like r: no contour map produces this pattern.

2-forms and 3-forms. A constant 2-form in the plane, such as $dx \wedge dy$, is drawn as a grid of unit cells made by the stacks of dx and dy. Its value on (v, w) counts the cells covered by the parallelogram spanned by v and w, with a sign from the orientation (Figure T1.11). In three dimensions the two stacks of dx and dy intersect in lines, so $dx \wedge dy$ is a bundle of *tubes* parallel to the z-axis: a flux 2-form literally counts flux tubes piercing a surface. A 3-form is drawn as a honeycomb of boxes, and it counts boxes inside a volume, like a charge density counts charges.

Try it: Stacks and cells

(a) Evaluate $\alpha = 2\,dx - dy$ on $v = (1, 3)$ by counting and by formula. (b) Compute $(dx \wedge dy + dy \wedge dz)(u, w)$ for $u = (1, 1, 0)$, $w = (0, 1, 1)$.
Answers: (a) $2 - 3 = -1$: the vector crosses one sheet against the orientation. (b) $(1 \cdot 1 - 1 \cdot 0) + (1 \cdot 1 - 0 \cdot 1) = 2$.

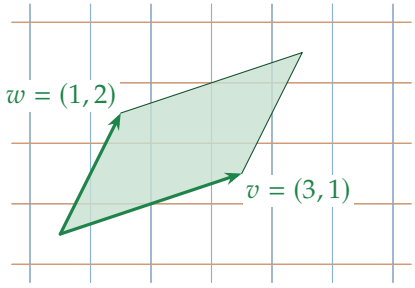


Figure T1.11: The 2-form $\mathrm{d}x \wedge \mathrm{d}y$ as a grid of unit cells. Applied to the pair (v, w) it returns the area of their parallelogram in cell units, here 5; swapping the order of the vectors reverses the orientation and the sign. The 2-form $3\,\mathrm{d}x \wedge \mathrm{d}y$ would be a grid three times finer in area.

Reading Nash (p. 14, (1.45); p. 39, (2.43)–(2.44))

Nash uses the wedge product first on p. 14, to say that the cup product on de Rham cohomology is $[\omega] \cup [\nu] = [\omega \wedge \nu]$, and he notes in (4.76) that the cup product inherits the graded symmetry $u \cup v = (-1)^{pq} v \cup u$ of forms, our (T1.23). On p. 39 he says a p-form "is a section of a bundle over M, the bundle being $\Lambda^p T^*M$, i.e. the pth anti-symmetric power of the cotangent bundle". In our language: at each point, $\Lambda^p T^*_x M$ is the $\binom{n}{p}$-dimensional space spanned by the $\mathrm{d}x^{\mu_1} \wedge \cdots \wedge \mathrm{d}x^{\mu_p}$, and a p-form field picks one element at each point. Recall that his component convention has no $1/p!$.

Recap

- The change-of-variables formula forces $\mathrm{d}y \wedge \mathrm{d}x = -\mathrm{d}x \wedge \mathrm{d}y$ and $\mathrm{d}x \wedge \mathrm{d}x = 0$.
- A p-form eats p vectors by the determinant rule; $\mathrm{d}x \wedge \mathrm{d}y$ measures oriented (shadow) area.
- $\mathrm{d}y^1 \wedge \cdots \wedge \mathrm{d}y^n = \det(\partial y/\partial x)\, \mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^n$; e.g. $\mathrm{d}x \wedge \mathrm{d}y = r\, \mathrm{d}r \wedge \mathrm{d}\phi$.
- $\dim \Omega^p_x = \binom{n}{p}$; top forms are one-dimensional; $\beta \wedge \alpha = (-1)^{pq} \alpha \wedge \beta$.
- In $\mathbb{R}^3$: 0,1,2,3-forms = scalar, vector, pseudovector, density; $\wedge$ gives cross, dot and triple products.
- Pictures: stacks (1-forms), cells or tubes (2-forms), boxes (3-forms).

T1.4 Metric, orientation and the Hodge star

In the three-dimensional dictionary of the last section, a vector field $\vec{F}$ turned into a 1-form $\vec{F} \cdot \mathrm{d}\vec{r}$ *and* into a 2-form $\vec{F} \cdot \mathrm{d}\vec{A}$. The first identification uses only the dot product. The second uses more: to turn $\hat{z}$ into $\mathrm{d}x \wedge \mathrm{d}y$ we needed to know that $\mathrm{d}x$ and $\mathrm{d}y$ are the unit 1-forms perpendicular to $\mathrm{d}z$, *and* which way round to order them. That is, we needed a metric and an orientation. The machine that uses exactly these two ingredients to convert p-forms into $(n-p)$-forms is the *Hodge star* $\star$. It is the most important operator in this chapter after d: it turns d into curl and divergence, it produces the inner product on forms that Nash uses to define adjoints and Laplacians, and in four dimensions it splits 2-forms into self-dual and anti-self-dual halves, which is where instantons live.

T1.4.1 The metric on 1-forms and orthonormal coframes

We saw in (T1.14) that a line element $\mathrm{d}s^2 = g_{\mu\nu}\mathrm{d}x^\mu \mathrm{d}x^\nu$ induces an inner product on 1-forms, $g(\mathrm{d}x^\mu, \mathrm{d}x^\nu) = g^{\mu\nu}$, the inverse matrix. Any symmetric invertible matrix can be diagonalised and the diagonal entries rescaled to ± 1. So at each point we can choose 1-forms $\sigma^1, \dots, \sigma^n$ with

$$g(\sigma^a, \sigma^b) = \pm\delta^{ab}, \qquad \mathrm{d}s^2 = \sum_a \pm(\sigma^a)^2. \tag{T1.26}$$

Such a set is an **orthonormal coframe** [Dray §14.2]. The line element tells us the coframe at a glance: whatever squares appear in $\mathrm{d}s^2$ are the σ^a.

Working definition T1.10 (signature)

The number s of minus signs in (T1.26) does not depend on the coframe (Sylvester's law of inertia [however a symmetric invertible matrix is brought to diagonal form with entries ±1, the number of -1 entries comes out the same]). We call it the **signature** [Dray §14.3]. $s = 0$ is **Riemannian** (Euclidean) geometry; $s = 1$ is **Lorentzian** geometry, the arena of relativity.

Dictionary: standard coframes

space	line element	orthonormal coframe
$\mathbb{R}^2$, polar	$\mathrm{d}r^2 + r^2\mathrm{d}\phi^2$	$\{\mathrm{d}r,\ r\,\mathrm{d}\phi\}$
$\mathbb{R}^3$, spherical	$\mathrm{d}r^2 + r^2\mathrm{d}\theta^2 + r^2\sin^2\theta\,\mathrm{d}\phi^2$	$\{\mathrm{d}r,\ r\,\mathrm{d}\theta,\ r\sin\theta\,\mathrm{d}\phi\}$
S^2, angles	$\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2$	$\{\mathrm{d}\theta,\ \sin\theta\,\mathrm{d}\phi\}$
S^2, stereographic	$\lambda^2(\mathrm{d}u^2 + \mathrm{d}v^2)$, $\lambda = \frac{2}{1+\rho^2}$	$\{\lambda\,\mathrm{d}u,\ \lambda\,\mathrm{d}v\}$
Minkowski $\mathbb{R}^{1,3}$	$-\mathrm{d}t^2 + \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2$	$\{\mathrm{d}t, \mathrm{d}x, \mathrm{d}y, \mathrm{d}z\}$, $s = 1$

T1.4.2 Inner products of p-forms

In $\mathbb{R}^3$ we want $\mathrm{d}x \wedge \mathrm{d}y$, which "is" $\hat{z}$, to have unit length. The natural guess, which Dray arrives at by requiring bilinearity [Dray (14.17)–(14.18)], is a determinant of 1-form inner products.

Working definition T1.11 (inner product of p-forms)

On products of 1-forms,

$$g(\alpha^1 \wedge \cdots \wedge \alpha^p,\ \beta^1 \wedge \cdots \wedge \beta^p) = \det\big[g(\alpha^i, \beta^j)\big], \tag{T1.27}$$

extended bilinearly [a general p-form is a sum of such products, and g is linear in each of its two slots]; on functions $g(f, h) = fh$. In components,

$$g(\alpha, \beta) = \frac{1}{p!}\,\alpha_{\mu_1\dots\mu_p}\,\beta^{\mu_1\dots\mu_p}, \qquad \beta^{\mu_1\dots\mu_p} = g^{\mu_1\nu_1}\cdots g^{\mu_p\nu_p}\beta_{\nu_1\dots\nu_p}. \tag{T1.28}$$

In practice: in an orthonormal coframe [1-forms σ^a with $g(\sigma^a, \sigma^b) = \pm\delta^{ab}$, read off as the squares in $\mathrm{d}s^2$], the basis p-forms $\sigma^{a_1} \wedge \cdots \wedge \sigma^{a_p}$ $(a_1 < \cdots < a_p)$ are orthonormal, and each has squared length equal to the product of the signs of its factors.

The $1/p!$ in (T1.28) compensates the $p!$ equal terms in the free sum, exactly as in (T1.22).

Worked example T1.17: Lengths of area forms

(a) Polar coordinates: $g^{rr} = 1$, $g^{\phi\phi} = r^{-2}$. Then

$$g(\mathrm{d}r \wedge \mathrm{d}\phi, \mathrm{d}r \wedge \mathrm{d}\phi) \overset{(1)}{=} \det \begin{pmatrix} g(\mathrm{d}r,\mathrm{d}r) & g(\mathrm{d}r,\mathrm{d}\phi) \\ g(\mathrm{d}\phi,\mathrm{d}r) & g(\mathrm{d}\phi,\mathrm{d}\phi) \end{pmatrix} \overset{(2)}{=} \det \begin{pmatrix} 1 & 0 \\ 0 & r^{-2} \end{pmatrix} = \frac{1}{r^2}.$$

(1) Definition (T1.27).
(2) Inverse polar metric.

So the unit area form is $r\,\mathrm{d}r \wedge \mathrm{d}\phi = \mathrm{d}x \wedge \mathrm{d}y$, as it should be.
(b) Minkowski space, where $g(\mathrm{d}t, \mathrm{d}t) = -1$: time–space 2-forms have negative squared length and space–space ones positive:

$$g(\mathrm{d}t \wedge \mathrm{d}x, \mathrm{d}t \wedge \mathrm{d}x) = (-1)(+1) = -1, \qquad g(\mathrm{d}y \wedge \mathrm{d}z, \mathrm{d}y \wedge \mathrm{d}z) = (+1)(+1) = +1.$$

For the electromagnetic field $F = E_x\,\mathrm{d}x \wedge \mathrm{d}t + \cdots + B_x\,\mathrm{d}y \wedge \mathrm{d}z + \ldots$ (we set this up carefully in Section T1.7) this gives $g(F,F) = |\vec{B}|^2 - |\vec{E}|^2$, which is -2 times the Maxwell Lagrangian density $\frac{1}{2}(|\vec{E}|^2 - |\vec{B}|^2)$. The famous Lorentz invariant $E^2 - B^2$ is just the squared length of the 2-form F.

Schwarz and reverse Schwarz

Apply (T1.27) to $\alpha \wedge \beta$ with itself:

$$g(\alpha \wedge \beta, \alpha \wedge \beta) = g(\alpha,\alpha)\,g(\beta,\beta) - g(\alpha,\beta)^2. \tag{T1.29}$$

In Euclidean signature every 2-form has non-negative squared length, since in an orthonormal basis of the plane spanned by α, β we have $\alpha \wedge \beta = f\,\sigma^1 \wedge \sigma^2$ with $g(\sigma^1 \wedge \sigma^2, \sigma^1 \wedge \sigma^2) = 1$. So $g(\alpha,\beta)^2 \le g(\alpha,\alpha)g(\beta,\beta)$: the Schwarz inequality. In a Lorentzian plane containing the time direction, $g(\sigma^1 \wedge \sigma^2, \sigma^1 \wedge \sigma^2) = -1$ and the inequality reverses [Dray §14.5].

Worked example T1.18: The reverse Schwarz inequality is $\gamma \ge 1$

In Minkowski 2-space take the two unit timelike 1-forms $\alpha = \mathrm{d}t$ and $\beta = \cosh\eta\,\mathrm{d}t + \sinh\eta\,\mathrm{d}x$ (a boosted time direction with rapidity η). Then

$$g(\alpha,\beta)^2 \overset{(1)}{=} (-\cosh\eta)^2 = \cosh^2\eta,$$
$$g(\alpha,\alpha)\,g(\beta,\beta) \overset{(2)}{=} (-1)(-\cosh^2\eta + \sinh^2\eta) = 1,$$

(1) $g(\mathrm{d}t,\mathrm{d}t) = -1$, $g(\mathrm{d}t,\mathrm{d}x) = 0$.
(2) $g(\beta,\beta) = -\cosh^2\eta + \sinh^2\eta = -1$.

and the reverse inequality $\cosh^2\eta \ge 1$ is the statement that the Lorentz factor $\gamma = \cosh\eta$ is never less than one: time dilation.

T1.4.3 Orientation and the volume form

At each point the top forms make up a one-dimensional space. With an orthonormal coframe we get the unit top form

$$\omega = \sigma^1 \wedge \cdots \wedge \sigma^n, \qquad g(\omega,\omega) = g(\sigma^1,\sigma^1)\cdots g(\sigma^n,\sigma^n) = (-1)^s. \tag{T1.30}$$

There are precisely two unit top forms, $\pm\omega$.

Working definition T1.12 (orientation, volume form)

An **orientation** is a continuous choice of one of the two unit top forms at every point; the chosen one is the **volume form** vol (Dray writes ω). In coordinates x^μ that are positively oriented [$dx^1 \wedge \cdots \wedge dx^n$ is a positive multiple of the chosen unit top form],

$$\mathrm{vol} = \sqrt{|\det g_{\mu\nu}|}\, dx^1 \wedge \cdots \wedge dx^n. \tag{T1.31}$$

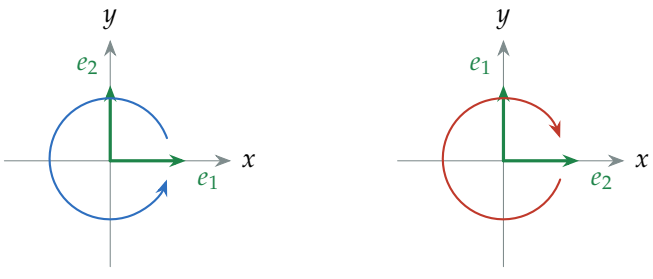
Why (T1.31): write $\sigma^a = e^a{}_\mu dx^\mu$. Then $g_{\mu\nu} = \eta_{ab} e^a{}_\mu e^b{}_\nu$ with $\eta = \mathrm{diag}(\pm 1)$, so $\det g = (\det e)^2(-1)^s$ and $|\det e| = \sqrt{|\det g|}$. By (T1.19), $\sigma^1 \wedge \cdots \wedge \sigma^n = \det e \; dx^1 \wedge \cdots \wedge dx^n$, and "positively oriented" means $\det e > 0$.

This is the same notion of orientation as in Section T1.1: two charts define the same orientation exactly when their coordinate top forms differ by a positive factor, which by (T1.19) is the statement that the transition Jacobian is positive. A manifold is orientable precisely when it admits a nowhere-vanishing top form.

Worked example T1.19: Volume forms we will keep using

$$\begin{aligned}
\mathbb{R}^2 \text{ polar:} &\quad \sqrt{\det \mathrm{diag}(1, r^2)} = r, &&\quad \mathrm{vol} = r\, dr \wedge d\phi, \\
\mathbb{R}^3 \text{ spherical:} &\quad \sqrt{r^4 \sin^2\theta} = r^2 \sin\theta, &&\quad \mathrm{vol} = r^2 \sin\theta\, dr \wedge d\theta \wedge d\phi, \\
S^2 \text{ angles:} &\quad \sqrt{\sin^2\theta} = \sin\theta, &&\quad \mathrm{vol} = \sin\theta\, d\theta \wedge d\phi, \\
S^2 \text{ stereographic:} &\quad \sqrt{\lambda^4} = \lambda^2, &&\quad \mathrm{vol} = \frac{4\, du \wedge dv}{(1+u^2+v^2)^2}, \\
\text{Minkowski:} &\quad \sqrt{|-1|} = 1, &&\quad \mathrm{vol} = dt \wedge dx \wedge dy \wedge dz.
\end{aligned}$$

Two consistency checks. The spherical result agrees with the direct computation of Example T1.13. And the stereographic area of the whole sphere is $\int_{\mathbb{R}^2} 4\, du\, dv/(1+\rho^2)^2 = \int_0^\infty 8\pi\rho\, d\rho/(1+\rho^2)^2 = 4\pi\left[-1/(1+\rho^2)\right]_0^\infty \cdot (-1) \cdot (-1) = 4\pi$, as it must be. With the north chart (u, v) the orientation of vol matches $\sin\theta\, d\theta \wedge d\phi$ only up to a sign: since $\rho = \cot(\theta/2)$ decreases as θ increases, $du \wedge dv = \rho\, d\rho \wedge d\phi$ is a *negative* multiple of $d\theta \wedge d\phi$. Choosing an orientation is a genuine choice.



$+dx \wedge dy$: counter-clockwise $\qquad$ $-dx \wedge dy$: clockwise

Figure T1.12: The two orientations of the plane. An orientation is a choice of unit top form, $+dx \wedge dy$ or $-dx \wedge dy$; equivalently a choice of which ordered bases (e_1, e_2) count as positive (those with $\mathrm{vol}(e_1, e_2) > 0$), equivalently a sense of rotation. On a Möbius band no continuous choice exists.

T1.4.4 The Hodge star

Look again at the dictionary pairs $dx \leftrightarrow dy\wedge dz, dy \leftrightarrow dz\wedge dx, dz \leftrightarrow dx\wedge dy, 1 \leftrightarrow dx\wedge dy\wedge dz$. In each pair, (form)$\wedge$(partner) $= dx \wedge dy \wedge dz$. The partner is the "orthogonal complement, ordered so as to complete the orientation". To make this linear we must let the partner scale with the form, and we arrive at Dray's definition [Dray (14.38)].

Working definition T1.13 (Hodge star)

The **Hodge star** is the linear map $\star : \Omega^p \to \Omega^{n-p}$ defined by

$$\alpha \wedge \star\beta = g(\alpha,\beta)\,\mathrm{vol} \qquad \text{for all } \alpha,\beta \in \Omega^p. \tag{T1.32}$$

Recipe on an orthonormal basis [1-forms σ^a with $g(\sigma^a,\sigma^b) = \pm\delta^{ab}$, e.g. dr, $r\,d\phi$ in polar coordinates]: for a basis p-form $\sigma^I = \sigma^{a_1}\wedge\cdots\wedge\sigma^{a_p}$, let σ^J be the wedge of the remaining σ^b's, in the order (or with the sign) that makes $\sigma^I\wedge\sigma^J = +\mathrm{vol}$ [the chosen unit top form, e.g. $dx\wedge dy\wedge dz$ in $\mathbb{R}^3$]. Then

$$\star\sigma^I = g(\sigma^I,\sigma^I)\,\sigma^J, \tag{T1.33}$$

where $g(\sigma^I,\sigma^I) = \pm 1$ is the product of the signs of the factors. For non-normalised forms, normalise first, star, then undo the normalisation using linearity [Dray §14.10].

Why the recipe works: $\sigma^I\wedge\star\sigma^I$ must equal $g(\sigma^I,\sigma^I)\mathrm{vol}$, so $\star\sigma^I$ contains σ^J with the stated coefficient. It contains nothing else, because every other basis $(n-p)$-form σ^K shares a factor with σ^I and so must satisfy $\sigma^{I'}\wedge\star\sigma^I = g(\sigma^{I'},\sigma^I)\mathrm{vol} = 0$ for the $\sigma^{I'}$ complementary to σ^K. Two immediate consequences in any dimension and signature [Dray (14.40)–(14.42)]:

$$\star 1 = \mathrm{vol}, \qquad \star\mathrm{vol} = (-1)^s. \tag{T1.34}$$

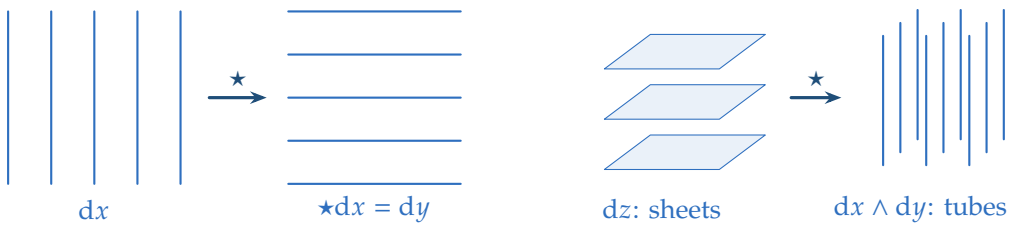


Figure T1.13: What the Hodge star does to pictures. Left: in the Euclidean plane $\star$ rotates a 1-form by 90°; the stack of vertical lines dx becomes the stack of horizontal lines dy. Doing it twice gives $\star\star\,dx = -dx$: on 1-forms in two dimensions $\star$ behaves like multiplication by i. Right: in $\mathbb{R}^3$ the horizontal sheets of dz become the vertical flux tubes of $dx\wedge dy$, which pierce those sheets at right angles.

T1.4.5 Tables of Hodge duals

We now compute $\star$ in every case we will need. Each entry follows from (T1.33); the first few are done in detail.

Worked example T1.20: Euclidean plane and polar coordinates

$\mathbb{R}^2$, $\mathrm{vol} = dx\wedge dy$. For dx the complement is dy since $dx\wedge dy = \mathrm{vol}$; for dy the complement is $-dx$ since $dy\wedge(-dx) = dx\wedge dy$. Both have $g = +1$. So

$$\star 1 = dx\wedge dy, \qquad \star dx = dy, \qquad \star dy = -dx, \qquad \star(dx\wedge dy) = 1.$$

Polar coordinates, coframe $\{\mathrm{d}r, r\,\mathrm{d}\phi\}$, vol $= r\,\mathrm{d}r \wedge \mathrm{d}\phi$: exactly the same pattern gives $\star\mathrm{d}r = r\,\mathrm{d}\phi$ and $\star(r\,\mathrm{d}\phi) = -\mathrm{d}r$. Then linearity handles the unnormalised $\mathrm{d}\phi$:

$$\star\mathrm{d}\phi \overset{(1)}{=} \star\left(\frac{1}{r}\, r\,\mathrm{d}\phi\right) \overset{(2)}{=} \frac{1}{r}\,\star\,(r\,\mathrm{d}\phi) \overset{(3)}{=} -\frac{\mathrm{d}r}{r}, \qquad \star(\mathrm{d}r \wedge \mathrm{d}\phi) \overset{(4)}{=} \frac{1}{r}\star\,(r\,\mathrm{d}r \wedge \mathrm{d}\phi) = \frac{1}{r}.$$

(1) Normalise: $\mathrm{d}\phi = \frac{1}{r}\cdot(r\,\mathrm{d}\phi)$.
(2) $\star$ is linear over functions.
(3) The orthonormal-basis result $\star(r\,\mathrm{d}\phi) = -\mathrm{d}r$.
(4) Same trick, with $\star\mathrm{vol} = 1$.

Check against Cartesian: $r^2\mathrm{d}\phi = x\,\mathrm{d}y - y\,\mathrm{d}x$ should have star $x\star\mathrm{d}y - y\star\mathrm{d}x = -x\,\mathrm{d}x - y\,\mathrm{d}y = -r\,\mathrm{d}r$; and indeed $\star(r^2\mathrm{d}\phi) = r^2(-\mathrm{d}r/r) = -r\,\mathrm{d}r$. ✓

Worked example T1.21: Euclidean $\mathbb{R}^3$ and spherical coordinates

vol $= \mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z$, all signs +. Cyclic permutations of (x, y, z) are even, so [Dray (12.30)–(12.37)]

$$\star 1 = \mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z, \quad \star\mathrm{d}x = \mathrm{d}y \wedge \mathrm{d}z, \quad \star\mathrm{d}y = \mathrm{d}z \wedge \mathrm{d}x, \quad \star\mathrm{d}z = \mathrm{d}x \wedge \mathrm{d}y,$$

$$\star(\mathrm{d}y \wedge \mathrm{d}z) = \mathrm{d}x, \quad \star(\mathrm{d}z \wedge \mathrm{d}x) = \mathrm{d}y, \quad \star(\mathrm{d}x \wedge \mathrm{d}y) = \mathrm{d}z, \quad \star(\mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z) = 1.$$

In spherical coordinates the same pattern holds for the coframe $(\mathrm{d}r,\ r\,\mathrm{d}\theta,\ r\sin\theta\,\mathrm{d}\phi)$, which is right-handed (Example T1.13). Undoing the normalisations:

$$\begin{aligned}
\star\mathrm{d}r &= (r\,\mathrm{d}\theta) \wedge (r\sin\theta\,\mathrm{d}\phi) = r^2\sin\theta\,\mathrm{d}\theta \wedge \mathrm{d}\phi,\\
\star\mathrm{d}\theta &= \tfrac{1}{r}\,(r\sin\theta\,\mathrm{d}\phi) \wedge \mathrm{d}r = \sin\theta\,\mathrm{d}\phi \wedge \mathrm{d}r,\\
\star\mathrm{d}\phi &= \tfrac{1}{r\sin\theta}\,\mathrm{d}r \wedge (r\,\mathrm{d}\theta) = \frac{1}{\sin\theta}\,\mathrm{d}r \wedge \mathrm{d}\theta.
\end{aligned}$$

The first line is the physicist's $\hat{r}\cdot\mathrm{d}\vec{A} = r^2\sin\theta\,\mathrm{d}\theta\,\mathrm{d}\phi$: the flux element through a sphere. So the Coulomb field $\vec{E} = \frac{q}{4\pi r^2}\hat{r}$ as a flux 2-form is $\star\left(\frac{q}{4\pi r^2}\mathrm{d}r\right) = \frac{q}{4\pi}\sin\theta\,\mathrm{d}\theta \wedge \mathrm{d}\phi$, independent of r, and its integral over any sphere is q.

Worked example T1.22: Minkowski space $\mathbb{R}^{1,3}$

$\eta = \mathrm{diag}(-1, 1, 1, 1)$, vol $= \mathrm{d}t \wedge \mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z$, $s = 1$. Every basis form containing $\mathrm{d}t$ has $g = -1$. We do three entries in detail.

- $\star\mathrm{d}t$: the complement with $\mathrm{d}t \wedge (\mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z) = \mathrm{vol}$ is $\mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z$, and $g(\mathrm{d}t, \mathrm{d}t) = -1$, so $\star\mathrm{d}t = -\mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z$.
- $\star\mathrm{d}x$: $\mathrm{d}x \wedge \mathrm{d}t \wedge \mathrm{d}y \wedge \mathrm{d}z = -\mathrm{vol}$ (one swap), so the complement is $-\mathrm{d}t \wedge \mathrm{d}y \wedge \mathrm{d}z$ and $\star\mathrm{d}x = -\mathrm{d}t \wedge \mathrm{d}y \wedge \mathrm{d}z$.
- $\star(\mathrm{d}t \wedge \mathrm{d}x)$: complement $\mathrm{d}y \wedge \mathrm{d}z$ (no reordering needed), $g = -1$, so $\star(\mathrm{d}t \wedge \mathrm{d}x) = -\mathrm{d}y \wedge \mathrm{d}z$.

Proceeding the same way for all sixteen basis forms (with (i, j, k) a cyclic permutation of (x, y, z)):

$\star 1$	$dt \wedge dx \wedge dy \wedge dz$	$\star(dt \wedge dx \wedge dy \wedge dz)$	-1
$\star dt$	$-dx \wedge dy \wedge dz$	$\star(dx \wedge dy \wedge dz)$	$-dt$
$\star dx^i$	$-dt \wedge dx^j \wedge dx^k$	$\star(dt \wedge dx^j \wedge dx^k)$	$-dx^i$
$\star(dt \wedge dx^i)$	$-dx^j \wedge dx^k$	$\star(dx^j \wedge dx^k)$	$dt \wedge dx^i$

The last row is the one electromagnetism uses: $\star$ swaps "electric-type" 2-forms $dt \wedge dx^i$ with "magnetic-type" ones $dx^j \wedge dx^k$, with one minus sign. That minus sign is why $\star\star = -1$ on 2-forms in Lorentzian signature, and why the duality rotation $(\vec{E}, \vec{B}) \to (-\vec{B}, \vec{E})$ of electromagnetism squares to -1.

Worked example T1.23: Euclidean $\mathbb{R}^4$: self-dual and anti-self-dual 2-forms

vol $= dx^1 \wedge dx^2 \wedge dx^3 \wedge dx^4$, all signs $+$. On 2-forms the recipe gives

$$\star(dx^1 \wedge dx^2) = dx^3 \wedge dx^4, \quad \star(dx^1 \wedge dx^3) = dx^4 \wedge dx^2, \quad \star(dx^1 \wedge dx^4) = dx^2 \wedge dx^3,$$

and conversely $\star(dx^3 \wedge dx^4) = dx^1 \wedge dx^2$, etc. (Check the middle one: $dx^1 \wedge dx^3 \wedge dx^4 \wedge dx^2$ is a cyclic permutation of the last three indices of 1234, hence even.) So $\star\star = +1$ on 2-forms, and $\star$ has eigenvalues ± 1. The eigenspaces are spanned by

$$\omega_1^{\pm} = dx^1 \wedge dx^2 \pm dx^3 \wedge dx^4, \quad \omega_2^{\pm} = dx^1 \wedge dx^3 \pm dx^4 \wedge dx^2, \quad \omega_3^{\pm} = dx^1 \wedge dx^4 \pm dx^2 \wedge dx^3,$$

with $\star\omega_i^{\pm} = \pm\omega_i^{\pm}$. The six-dimensional space of 2-forms splits as $3 + 3$: **self-dual** ($\star F = F$) and **anti-self-dual** ($\star F = -F$). A Yang–Mills field that is self-dual automatically solves the field equations; these are the instantons of Chapter 8. Note also $\omega_1^+ \wedge \omega_1^+ = 2\,\mathrm{vol}$ (Example T1.15) while $\omega_1^+ \wedge \omega_1^- = 0$: for a self-dual F, $F \wedge F = F \wedge \star F = |F|^2 \mathrm{vol} \geq 0$. Contrast Minkowski space, where $\star\star = -1$ on 2-forms and the eigenvalues are $\pm i$: real self-dual fields exist only in Euclidean signature, which is one reason Nash (and instanton physics generally) works Euclidean.

T1.4.6 The sign of $\star\star$

Apply the recipe twice. If $\sigma^I \wedge \sigma^J = \mathrm{vol}$ then moving the $(n-p)$ factors of σ^J past the p factors of σ^I gives $\sigma^J \wedge \sigma^I = (-1)^{p(n-p)}\mathrm{vol}$, so the complement of σ^J is $(-1)^{p(n-p)}\sigma^I$. Hence

$$\begin{aligned} \star\star\,\sigma^I &\overset{(1)}{=} g(\sigma^I, \sigma^I)\ \star\sigma^J \overset{(2)}{=} g(\sigma^I, \sigma^I)\, g(\sigma^J, \sigma^J)\,(-1)^{p(n-p)}\sigma^I \\ &\overset{(3)}{=} g(\omega, \omega)\,(-1)^{p(n-p)}\sigma^I \overset{(4)}{=} (-1)^{p(n-p)+s}\,\sigma^I. \end{aligned} \tag{T1.35}$$

(1) Recipe (T1.33) for σ^I.
(2) Recipe for σ^J, whose complement is $(-1)^{p(n-p)}\sigma^I$.
(3) σ^I and σ^J together contain each σ^a once, so the product of their squared lengths is the product of all n signs, $g(\omega, \omega)$.
(4) (T1.30).

Key idea: $\star\star = (-1)^{p(n-p)+s}$ on p-forms

	$p=0$	$p=1$	$p=2$	$p=3$	$p=4$
$\mathbb{R}^2$ $(n=2, s=0)$	$+$	$-$	$+$		
$\mathbb{R}^3$ $(n=3, s=0)$	$+$	$+$	$+$	$+$	
$\mathbb{R}^4$ $(n=4, s=0)$	$+$	$-$	$+$	$-$	$+$
$\mathbb{R}^{1,3}$ $(n=4, s=1)$	$-$	$+$	$-$	$+$	$-$

So $\star^{-1} = \pm\star$, and in odd dimensions with Euclidean signature $\star\star = 1$ always.

T1.4.7 The Hodge star in components

For index computations, and to read Nash, we need the component form. With $\varepsilon_{\mu_1\dots\mu_n}$ the Levi-Civita symbol ($\varepsilon_{12\dots n} = +1$ in positively oriented coordinates, totally antisymmetric),

$$(\star\alpha)_{\nu_1\dots\nu_{n-p}} = \frac{\sqrt{|g|}}{p!}\,\alpha^{\mu_1\dots\mu_p}\,\varepsilon_{\mu_1\dots\mu_p\nu_1\dots\nu_{n-p}}. \tag{T1.36}$$

Worked example T1.24: Checking the component formula

(a) $\mathbb{R}^3$, $\alpha = \mathrm{d}z$: $\alpha^z = 1$, so $(\star\alpha)_{xy} = \varepsilon_{zxy} = 1$, $(\star\alpha)_{yx} = -1$, and $\star\alpha = \frac{1}{2!}\big((\star\alpha)_{xy}\mathrm{d}x\wedge\mathrm{d}y + (\star\alpha)_{yx}\mathrm{d}y\wedge\mathrm{d}x\big) = \mathrm{d}x\wedge\mathrm{d}y$.
(b) Polar, $\alpha = \mathrm{d}\phi$: $\alpha^\phi = g^{\phi\phi}\alpha_\phi = r^{-2}$ and $\sqrt{|g|} = r$, so

$$(\star\alpha)_r \overset{(1)}{=} r\cdot r^{-2}\,\varepsilon_{\phi r} \overset{(2)}{=} -\frac{1}{r}, \qquad \star\mathrm{d}\phi = -\frac{\mathrm{d}r}{r}.$$

(1) (T1.36) with $p=1$, $n=2$, coordinates ordered (r,ϕ).
(2) $\varepsilon_{r\phi} = +1$, so $\varepsilon_{\phi r} = -1$.

This agrees with Example T1.20.
(c) Minkowski, $\alpha = \mathrm{d}t$: $\alpha^t = \eta^{tt} = -1$, so $(\star\alpha)_{xyz} = -\varepsilon_{txyz} = -1$ and $\star\mathrm{d}t = -\mathrm{d}x\wedge\mathrm{d}y\wedge\mathrm{d}z$, as in the table.

T1.4.8 Dot product, cross product, and the L^2 inner product

Since $\star$ encodes g, the vector operations can be written with it [Dray §14.11]:

$$g(\alpha,\beta) = (-1)^s\ \star(\alpha\wedge\star\beta), \qquad \alpha\times\beta = \star(\alpha\wedge\beta) \quad (\mathbb{R}^3 \text{ only}). \tag{T1.37}$$

The first is (T1.32) starred, using $\star\mathrm{vol} = (-1)^s$. The second only makes sense in three dimensions, where the star of a 2-form is again a 1-form.

Integrating the pointwise inner product over a compact oriented manifold gives the inner product on forms that the whole of Hodge theory is built on:

$$\langle\alpha,\beta\rangle = \int_M \alpha\wedge\star\beta = \int_M g(\alpha,\beta)\,\mathrm{vol}. \tag{T1.38}$$

In Riemannian signature it is positive definite: $\langle\alpha,\alpha\rangle = \int|\alpha|^2\mathrm{vol} > 0$ unless $\alpha = 0$.

Worked example T1.25: Two inner products on S^2

Let $Z = \cos\theta$, the height function. Then

$$\langle Z, Z\rangle \overset{(1)}{=} \int_0^{2\pi}\int_0^{\pi} \cos^2\theta\,\sin\theta\,\mathrm{d}\theta\,\mathrm{d}\phi \overset{(2)}{=} 2\pi\cdot\frac{2}{3} = \frac{4\pi}{3},$$

$$\langle \mathrm{d}Z, \mathrm{d}Z\rangle \overset{(3)}{=} \int_{S^2} g(-\sin\theta\,\mathrm{d}\theta, -\sin\theta\,\mathrm{d}\theta)\,\mathrm{vol} \overset{(4)}{=} \int_0^{2\pi}\int_0^{\pi} \sin^3\theta\,\mathrm{d}\theta\,\mathrm{d}\phi \overset{(5)}{=} 2\pi\cdot\frac{4}{3} = \frac{8\pi}{3}.$$

(1) $g(f,h) = fh$ on functions; $\mathrm{vol} = \sin\theta\,\mathrm{d}\theta\wedge\mathrm{d}\phi$.
(2) Substitute $c = \cos\theta$: $\int_{-1}^{1} c^2\,\mathrm{d}c = 2/3$.
(3) $\mathrm{d}Z = -\sin\theta\,\mathrm{d}\theta$.
(4) $g(\mathrm{d}\theta,\mathrm{d}\theta) = 1$, times $\sin^2\theta$ from the coefficient, times $\sin\theta$ from vol.
(5) $\int_0^{\pi}\sin^3\theta\,\mathrm{d}\theta = \int_{-1}^{1}(1-c^2)\,\mathrm{d}c = 4/3$.

The ratio $\langle \mathrm{d}Z,\mathrm{d}Z\rangle/\langle Z,Z\rangle = 2$ is not an accident: it is the eigenvalue $\ell(\ell+1) = 2$ of the Laplacian on the $\ell = 1$ spherical harmonic Z, as we will see in Section T1.6.

Watch out: What the star depends on

$\star$ depends on the metric *and* the orientation; reversing the orientation flips the sign of every $\star$. But on p-forms in dimension $n = 2p$ it is *conformally invariant*: replacing g by $\lambda^2 g$ multiplies $g(\alpha,\beta)$ by λ^{-2p} and vol by λ^n, and these cancel when $n = 2p$. On the sphere in the stereographic chart, with coframe $\{\lambda\mathrm{d}u, \lambda\mathrm{d}v\}$, this gives $\star\mathrm{d}u = \mathrm{d}v$ and $\star\mathrm{d}v = -\mathrm{d}u$, exactly as in the flat plane, although $\star 1 = \lambda^2\mathrm{d}u\wedge\mathrm{d}v$ does feel the conformal factor. The first fact is why the Maxwell and Yang–Mills actions $\int F\wedge\star F$ are conformally invariant in four dimensions, and why a two-dimensional surface's $\star$ on 1-forms only knows its complex structure (Section T1.10).

Reading Nash (p. 39, (2.43); p. 109, (4.77))

Nash's summary list (2.43) contains exactly the objects of this section, in Riemannian signature: the inner product $\langle\omega,\nu\rangle = \int_M \omega\wedge\star\nu$ (our (T1.38)), $\star : \Omega^p \to \Omega^{n-p}$ with $\star^2 = (-1)^{p(n-p)}$ (our (T1.35) with $s = 0$), and the component formula $\star\omega = \frac{\sqrt{g}}{(n-p)!}\,\epsilon^{i_1\dots i_p}{}_{i_{p+1}\dots i_n}\,\omega_{i_1\dots i_p}\,\mathrm{d}x^{i_{p+1}}\wedge\cdots\wedge\mathrm{d}x^{i_n}$. Because his components carry no $1/p!$ (they are ours divided by $p!$), this is the same as (T1.36). He remarks that "some signs would change if the metric was Lorentzian": that is the $(-1)^s$. On p. 109 he uses the rescaled star $\tau = \mathrm{i}^{p(p-1)+n/2}\star$ on complex forms with $n = 4l$, arranged so that $\tau^2 = 1$. For $n = 4$ on 2-forms, $\tau = \mathrm{i}^{2+2}\star = \star$, and its ±1 eigenspaces are the self-dual and anti-self-dual forms of Example T1.23. On 1-forms, $\tau = \mathrm{i}^2\star = -\star$ and on 3-forms $\tau = \mathrm{i}^8\star = \star$, so $\tau^2 = -\star\star = -(-1) = 1$ there too.

Recap

- A metric gives orthonormal coframes σ^a with $g(\sigma^a,\sigma^b) = \pm\delta^{ab}$; s = number of minus signs.
- g extends to p-forms by determinants; $g(F,F) = B^2 - E^2$ in Minkowski space.
- Orientation = choice of unit top form; $\mathrm{vol} = \sqrt{|g|}\,\mathrm{d}x^1\wedge\cdots\wedge\mathrm{d}x^n$.
- $\alpha\wedge\star\beta = g(\alpha,\beta)\mathrm{vol}$; on an orthonormal basis $\star\sigma^I = g(\sigma^I,\sigma^I)\sigma^J$.

- $\star\star = (-1)^{p(n-p)+s}$; in $\mathbb{R}^4$ 2-forms split into self-dual and anti-self-dual.
- $\langle \alpha, \beta \rangle = \int \alpha \wedge \star\beta$ is the inner product used to define adjoints.

T1.5 The exterior derivative

We know how to differentiate a function: $f \mapsto \mathrm{d}f$, the gradient without a metric. Vector calculus has two more derivatives, the curl and the divergence, each with its own formula, its own set of identities, and its own horror-show of expressions in curvilinear coordinates. Is there one derivative, defined for forms of every rank, that contains all of them and that makes sense in any chart of any manifold? There is, and it is almost embarrassingly simple: *differentiate the coefficients and wedge the result on from the left*. This section builds that operator, the *exterior derivative* d, and shows what it measures: circulation per unit area, flux per unit volume, and in general "how much a p-form fails to integrate to zero around the boundary of a tiny $(p+1)$-dimensional box". Its single most important property, $\mathrm{d}^2 = 0$, is the seed of de Rham cohomology and hence of half of Nash's book.

T1.5.1 Why the obvious derivative is wrong

Given a 1-form $\alpha = \alpha_\nu \, \mathrm{d}x^\nu$, the obvious thing to try is the array of partial derivatives $\partial_\mu \alpha_\nu$. The trouble is that it does not mean the same thing in different charts. Change coordinates $x \to y$ and use the covariant law (T1.12), $\alpha_\nu = (\partial y^\rho / \partial x^\nu)\, \tilde{\alpha}_\rho$:

$$\begin{aligned} \frac{\partial \alpha_\nu}{\partial x^\mu} &\overset{(1)}{=} \frac{\partial}{\partial x^\mu}\left(\frac{\partial y^\rho}{\partial x^\nu}\, \tilde{\alpha}_\rho \right) \\ &\overset{(2)}{=} \frac{\partial^2 y^\rho}{\partial x^\mu \partial x^\nu}\, \tilde{\alpha}_\rho + \frac{\partial y^\rho}{\partial x^\nu} \frac{\partial y^\sigma}{\partial x^\mu} \frac{\partial \tilde{\alpha}_\rho}{\partial y^\sigma}. \end{aligned} \tag{T1.39}$$

(1) Covariant transformation law for the components of a 1-form.
(2) Product rule, and the chain rule $\partial/\partial x^\mu = (\partial y^\sigma / \partial x^\mu)\, \partial/\partial y^\sigma$ on $\tilde{\alpha}_\rho$.

The second term is what a well-behaved object with two lower indices would give. The first term, with second derivatives of the transition map, spoils it. But look: the spoiler is *symmetric* in $\mu \leftrightarrow \nu$. So the antisymmetric combination

$$\partial_\mu \alpha_\nu - \partial_\nu \alpha_\mu$$

is immune. Antisymmetric pairs of lower indices are exactly the components of 2-forms. The derivative that makes sense is therefore one that *produces forms*. This is the whole reason differential forms, and not general tensors, are the natural objects to differentiate on a bare manifold (without a metric or connection).

Worked example T1.26: The spoiler term, seen in polar coordinates

Take the constant 1-form $\alpha = \mathrm{d}x$ on the plane. In Cartesian coordinates its components are $(1, 0)$, and all partial derivatives vanish. In polar coordinates, $\alpha = \cos\phi \, \mathrm{d}r - r \sin\phi \, \mathrm{d}\phi$, so $\alpha_r = \cos\phi$ and $\alpha_\phi = -r\sin\phi$. Then

$$\partial_r \alpha_\phi = -\sin\phi \neq 0, \qquad \partial_\phi \alpha_r = -\sin\phi \neq 0,$$

$$\partial_r \alpha_\phi - \partial_\phi \alpha_r \overset{(1)}{=} -\sin\phi - (-\sin\phi) = 0.$$

(1) The two individually nonzero derivatives cancel in the antisymmetric combination.

The individual partial derivatives are artefacts of the curved coordinate lines; the antisymmetric combination agrees with the Cartesian answer 0.

T1.5.2 Definition and first examples

Working definition T1.14 (exterior derivative)

The **exterior derivative** $\mathrm{d}:\Omega^p\to\Omega^{p+1}$ is defined in any chart [every chart gives the same $(p+1)$-form, since the spoiler terms cancel in antisymmetric combinations] by

$$\mathrm{d}\big(f\,\mathrm{d}x^{\mu_1}\wedge\cdots\wedge\mathrm{d}x^{\mu_p}\big)=\mathrm{d}f\wedge\mathrm{d}x^{\mu_1}\wedge\cdots\wedge\mathrm{d}x^{\mu_p},\qquad \mathrm{d}f=\partial_\nu f\,\mathrm{d}x^\nu, \tag{T1.40}$$

extended by linearity [a general p-form is a sum of such terms, and d acts on each term separately] [Dray (15.7)]. *Recipe:* differentiate each coefficient, wedge the differential on the left, simplify with $\mathrm{d}x^\mu\wedge\mathrm{d}x^\mu=0$. In components,

$$(\mathrm{d}\alpha)_{\mu\nu}=\partial_\mu\alpha_\nu-\partial_\nu\alpha_\mu\ \ (p=1),\qquad (\mathrm{d}\alpha)_{\mu_0\mu_1\ldots\mu_p}=\sum_{i=0}^{p}(-1)^i\,\partial_{\mu_i}\alpha_{\mu_0\ldots\widehat{\mu_i}\ldots\mu_p}, \tag{T1.41}$$

where the hat means "omit this index".

Let us check the $p=1$ component formula, since it is the one we use most:

$$\mathrm{d}(\alpha_\nu\,\mathrm{d}x^\nu)\overset{(1)}{=}\partial_\mu\alpha_\nu\,\mathrm{d}x^\mu\wedge\mathrm{d}x^\nu\overset{(2)}{=}\tfrac12\big(\partial_\mu\alpha_\nu-\partial_\nu\alpha_\mu\big)\,\mathrm{d}x^\mu\wedge\mathrm{d}x^\nu,$$

(1) Definition (T1.40), summing over ν, and $\mathrm{d}\alpha_\nu=\partial_\mu\alpha_\nu\,\mathrm{d}x^\mu$.
(2) Split $\partial_\mu\alpha_\nu$ into symmetric and antisymmetric parts; the symmetric part dies against the antisymmetric $\mathrm{d}x^\mu\wedge\mathrm{d}x^\nu$. What is left is in the form $\frac{1}{2!}(\mathrm{d}\alpha)_{\mu\nu}\mathrm{d}x^\mu\wedge\mathrm{d}x^\nu$ of (T1.22).

By the argument around (T1.39), the result is the same whichever chart we compute in. (We give a chart-free formula below.)

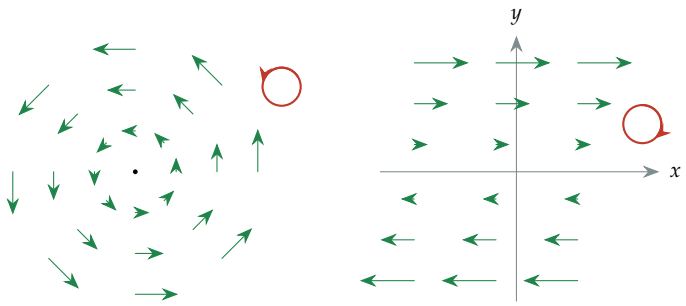
Worked example T1.27: Two flows in the plane

Think of a 1-form as a velocity field $\vec v\cdot\mathrm{d}\vec r$ of a fluid.
Rigid rotation, $\vec v=\Omega(-y,x)$: $v=\Omega(-y\,\mathrm{d}x+x\,\mathrm{d}y)$.

$$\mathrm{d}v\overset{(1)}{=}\Omega\big(-\mathrm{d}y\wedge\mathrm{d}x+\mathrm{d}x\wedge\mathrm{d}y\big)\overset{(2)}{=}2\Omega\,\mathrm{d}x\wedge\mathrm{d}y.$$

(1) (T1.40): $\mathrm{d}(-y)=-\mathrm{d}y$ wedged onto $\mathrm{d}x$, and $\mathrm{d}x$ wedged onto $\mathrm{d}y$.
(2) $\mathrm{d}y\wedge\mathrm{d}x=-\mathrm{d}x\wedge\mathrm{d}y$.

The coefficient 2Ω is the vorticity, twice the angular velocity.
Shear flow, $\vec v=(\kappa y,0)$: $v=\kappa y\,\mathrm{d}x$ and $\mathrm{d}v=\kappa\,\mathrm{d}y\wedge\mathrm{d}x=-\kappa\,\mathrm{d}x\wedge\mathrm{d}y$. Although every fluid particle moves in a straight line, the flow has vorticity: a small paddle wheel placed in it turns clockwise, because the top of the wheel is pushed harder than the bottom (Figure T1.14).
Radial flow, $\vec v=(x,y)$: $v=x\,\mathrm{d}x+y\,\mathrm{d}y$ and $\mathrm{d}v=\mathrm{d}x\wedge\mathrm{d}x+\mathrm{d}y\wedge\mathrm{d}y=0$. No vorticity; and indeed $v=\mathrm{d}(\frac12 r^2)$.



rotation: $\mathrm{d}v = 2\Omega\,\mathrm{d}x \wedge \mathrm{d}y$ shear: $\mathrm{d}v = -\kappa\,\mathrm{d}x \wedge \mathrm{d}y$

Figure T1.14: Two velocity 1-forms and their exterior derivatives. Left: rigid rotation; a small paddle wheel (red) turns counter-clockwise, and $\mathrm{d}v$ is a positive multiple of $\mathrm{d}x \wedge \mathrm{d}y$. Right: shear flow, whose streamlines are straight; the wheel still turns, clockwise here, because the flow above its centre is faster than below. $\mathrm{d}v$ detects local rotation, not curved streamlines.

T1.5.3 What $\mathrm{d}\alpha$ measures: circulation per unit area

The paddle wheel is not just a metaphor. Let us compute the circulation of a 1-form α around a small parallelogram with corner p and sides εv, εw (constant vectors in some chart), keeping terms up to order ε^2 (Figure T1.15). Along a side from q to $q + \varepsilon u$ the line integral is $\int_0^\varepsilon \alpha_\nu(q + tu)u^\nu\,\mathrm{d}t = \varepsilon\,\alpha_\nu(q)u^\nu + \tfrac{1}{2}\varepsilon^2\,\partial_\mu\alpha_\nu(q)u^\mu u^\nu + O(\varepsilon^3)$, by Taylor expansion. The four sides give

$$\begin{aligned}
\oint \alpha &\overset{(1)}{=} \left[\varepsilon\alpha_\nu v^\nu + \tfrac{1}{2}\varepsilon^2\partial_\mu\alpha_\nu v^\mu v^\nu\right] + \left[\varepsilon\alpha_\nu w^\nu + \varepsilon^2\partial_\mu\alpha_\nu v^\mu w^\nu + \tfrac{1}{2}\varepsilon^2\partial_\mu\alpha_\nu w^\mu w^\nu\right] \\
&\quad - \left[\varepsilon\alpha_\nu v^\nu + \varepsilon^2\partial_\mu\alpha_\nu w^\mu v^\nu + \tfrac{1}{2}\varepsilon^2\partial_\mu\alpha_\nu v^\mu v^\nu\right] - \left[\varepsilon\alpha_\nu w^\nu + \tfrac{1}{2}\varepsilon^2\partial_\mu\alpha_\nu w^\mu w^\nu\right] + O(\varepsilon^3) \\
&\overset{(2)}{=} \varepsilon^2\left(\partial_\mu\alpha_\nu v^\mu w^\nu - \partial_\mu\alpha_\nu w^\mu v^\nu\right) + O(\varepsilon^3) \\
&\overset{(3)}{=} \varepsilon^2\left(\partial_\mu\alpha_\nu - \partial_\nu\alpha_\mu\right)v^\mu w^\nu + O(\varepsilon^3) \overset{(4)}{=} \varepsilon^2\,\mathrm{d}\alpha(v,w) + O(\varepsilon^3).
\end{aligned} \tag{T1.42}$$

(1) The four sides, in order: bottom (from p, direction v), right (from $p + \varepsilon v$, direction w, so the Taylor expansion about p picks up an extra $\varepsilon\,v^\mu\partial_\mu$), top traversed backwards (from $p + \varepsilon w$, direction v) and left traversed backwards (from p, direction w). All α_ν and derivatives are at p.
(2) All the order-ε terms and all the "diagonal" $\tfrac{1}{2}\varepsilon^2$ terms cancel in pairs.
(3) Rename the dummy indices $\mu \leftrightarrow \nu$ in the second term.
(4) The components of $\mathrm{d}\alpha$ (T1.41), evaluated on (v, w) as in Section T1.3.

Key idea: Infinitesimal Stokes

$\mathrm{d}\alpha(v, w)$ is the circulation of α around the infinitesimal parallelogram spanned by v and w, per unit ε^2. More generally, for a p-form α, $\mathrm{d}\alpha(v_0, \ldots, v_p)$ is the integral of α over the (oriented) boundary of the small $(p+1)$-dimensional parallelepiped spanned by the v_i. Adding up many small cells, interior edges cancel and we get Stokes' theorem $\int_R \mathrm{d}\alpha = \int_{\partial R}\alpha$ (Chapter T2).

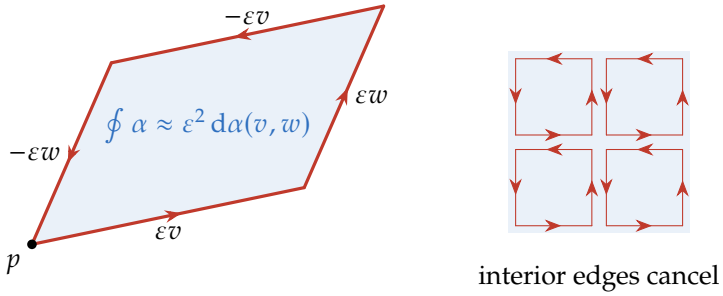


Figure T1.15: Left: the circulation of a 1-form around a small parallelogram is $\varepsilon^2 \mathrm{d}\alpha(v,w)$ to leading order, (T1.42). The order-ε contributions of opposite sides cancel; what survives is the *difference* of α between opposite sides. Right: gluing cells, each interior edge is traversed twice in opposite directions, so the sum of the small circulations is the circulation around the outer edge. That is Stokes' theorem.

T1.5.4 Gradient, curl and divergence are all d

In Euclidean $\mathbb{R}^3$ let $F = F_x\,\mathrm{d}x + F_y\,\mathrm{d}y + F_z\,\mathrm{d}z$ be the 1-form of a vector field $\vec{F}$. Following Dray [Dray (15.9)–(15.10)], first one term:

$$\begin{aligned}\mathrm{d}(F_x\,\mathrm{d}x) &\overset{(1)}{=} \left(\partial_x F_x\,\mathrm{d}x + \partial_y F_x\,\mathrm{d}y + \partial_z F_x\,\mathrm{d}z\right)\wedge \mathrm{d}x \\ &\overset{(2)}{=} \partial_z F_x\,\mathrm{d}z\wedge\mathrm{d}x - \partial_y F_x\,\mathrm{d}x\wedge\mathrm{d}y.\end{aligned}$$

(1) Definition (T1.40).
(2) $\mathrm{d}x\wedge\mathrm{d}x = 0$; $\mathrm{d}y\wedge\mathrm{d}x = -\mathrm{d}x\wedge\mathrm{d}y$.

Adding the two similar terms,

$$\mathrm{d}F = (\partial_y F_z - \partial_z F_y)\,\mathrm{d}y\wedge\mathrm{d}z + (\partial_z F_x - \partial_x F_z)\,\mathrm{d}z\wedge\mathrm{d}x + (\partial_x F_y - \partial_y F_x)\,\mathrm{d}x\wedge\mathrm{d}y. \tag{T1.43}$$

These are the components of $\nabla\times\vec{F}$, arranged as a flux 2-form. Next, a flux 2-form $G = G_x\,\mathrm{d}y\wedge\mathrm{d}z + G_y\,\mathrm{d}z\wedge\mathrm{d}x + G_z\,\mathrm{d}x\wedge\mathrm{d}y$:

$$\begin{aligned}\mathrm{d}G &\overset{(1)}{=} \partial_x G_x\,\mathrm{d}x\wedge\mathrm{d}y\wedge\mathrm{d}z + \partial_y G_y\,\mathrm{d}y\wedge\mathrm{d}z\wedge\mathrm{d}x + \partial_z G_z\,\mathrm{d}z\wedge\mathrm{d}x\wedge\mathrm{d}y \\ &\overset{(2)}{=} (\partial_x G_x + \partial_y G_y + \partial_z G_z)\,\mathrm{d}x\wedge\mathrm{d}y\wedge\mathrm{d}z.\end{aligned} \tag{T1.44}$$

(1) In $\mathrm{d}G_x\wedge\mathrm{d}y\wedge\mathrm{d}z$ only the $\mathrm{d}x$ part of $\mathrm{d}G_x$ survives; similarly for the others.
(2) Cyclic permutations of three 1-forms carry no sign.

This is the divergence, as a density. Using the Hodge star of Example T1.21 to translate back into vectors [Dray (15.12)–(15.19)]:

Dictionary: vector calculus in forms, $\mathbb{R}^3$

vector calculus	forms	ranks
∇f	$\mathrm{d}f$	$0\to 1$
$\nabla\times\vec{F}$	$\star\mathrm{d}F$ (or $\mathrm{d}F$ as a flux)	$1\to 2\,(\to 1)$
$\nabla\cdot\vec{F}$	$\star\mathrm{d}\star F$ (or $\mathrm{d}\star F$ as a density)	$1\to 2\to 3\,(\to 0)$
$\nabla^2 f$	$\star\mathrm{d}\star\mathrm{d}f$	$0\to 1\to 2\to 3\to 0$

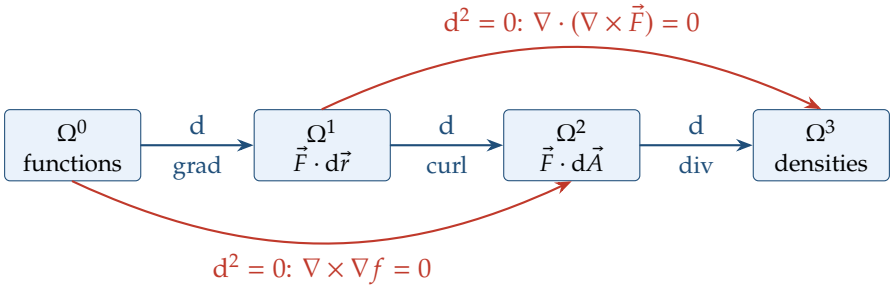


Figure T1.16: The de Rham complex of $\mathbb{R}^3$. One operator d plays the role of gradient, curl and divergence according to the rank it acts on. Two steps in a row always give zero (red arrows); these are the two classic identities of vector calculus. Nash's de Rham cohomology [Nash p. 13, (1.43)] measures how far "$d\beta = 0$" is from "$\beta = d\alpha$" in each box.

Worked example T1.28: Divergence in spherical coordinates, and the Coulomb field

Write a vector field in the orthonormal spherical basis as the 1-form $F = F_r\,dr + F_\theta\, r\,d\theta + F_\phi\, r\sin\theta\,d\phi$. Using the star of the orthonormal coframe (Example T1.21),

$$\begin{aligned} \star F &\overset{(1)}{=} F_r\,(r\,d\theta)\wedge(r\sin\theta\,d\phi) + F_\theta\,(r\sin\theta\,d\phi)\wedge dr + F_\phi\,dr\wedge(r\,d\theta) \\ &= r^2\sin\theta F_r\,d\theta\wedge d\phi + r\sin\theta F_\theta\,d\phi\wedge dr + rF_\phi\,dr\wedge d\theta, \\ d\star F &\overset{(2)}{=} \left[\partial_r(r^2\sin\theta F_r) + \partial_\theta(r\sin\theta F_\theta) + \partial_\phi(rF_\phi)\right] dr\wedge d\theta\wedge d\phi, \\ \nabla\cdot\vec{F} = \star d\star F &\overset{(3)}{=} \frac{1}{r^2}\partial_r(r^2F_r) + \frac{1}{r\sin\theta}\partial_\theta(\sin\theta F_\theta) + \frac{1}{r\sin\theta}\partial_\phi F_\phi. \end{aligned}$$

(1) $\star$ sends each element of the right-handed orthonormal coframe $(dr, r\,d\theta, r\sin\theta\,d\phi)$ to the wedge of the other two in cyclic order.
(2) For each term only the derivative along the missing coordinate survives; $d\phi\wedge dr\wedge d\theta$ and $dr\wedge d\theta\wedge d\phi$ are cyclic, so no signs.
(3) $\star(dr\wedge d\theta\wedge d\phi) = 1/(r^2\sin\theta)$, because $r^2\sin\theta\,dr\wedge d\theta\wedge d\phi = \text{vol}$ has star 1.

This is the textbook formula, obtained with no Christoffel symbols and no scale-factor lore. For the Coulomb field $F_r = q/(4\pi r^2)$, $F_\theta = F_\phi = 0$: $r^2\sin\theta F_r = q\sin\theta/4\pi$ is independent of r, so $d\star F = 0$ for $r \neq 0$. The flux 2-form $\star F = \frac{q}{4\pi}\sin\theta\,d\theta\wedge d\phi$ is closed but, as we shall see, not exact: the charge at the origin is detected by its integral, q, over any sphere.

Try it: Curl in cylindrical coordinates

With coframe $(d\rho, \rho\,d\phi, dz)$ and $F = F_\rho\,d\rho + F_\phi\,\rho\,d\phi + F_z\,dz$, compute the z-component of $\star dF$.
Answer: the $d\rho\wedge d\phi$ part of dF is $\left(\partial_\rho(\rho F_\phi) - \partial_\phi F_\rho\right)d\rho\wedge d\phi$, and $\star(d\rho\wedge d\phi) = \frac{1}{\rho}\star(d\rho\wedge\rho\,d\phi) = \frac{1}{\rho}\,dz$, giving $(\nabla\times\vec{F})_z = \frac{1}{\rho}\left(\partial_\rho(\rho F_\phi) - \partial_\phi F_\rho\right)$.

T1.5.5 $d^2 = 0$

Apply d twice to a function [Dray (15.23)]:

$$d(df) \overset{(1)}{=} d\left(\frac{\partial f}{\partial x^\nu}\right)\wedge dx^\nu \overset{(2)}{=} \frac{\partial^2 f}{\partial x^\mu\partial x^\nu}\,dx^\mu\wedge dx^\nu \overset{(3)}{=} 0. \tag{T1.45}$$

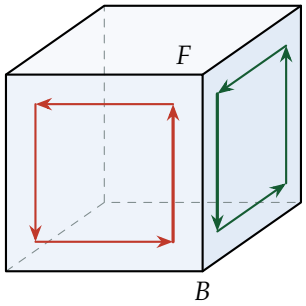
(1) Definition (T1.40) applied to the 1-form $\partial_\nu f\,dx^\nu$.

(2) Differential of the coefficient.
(3) Mixed partial derivatives are symmetric in $\mu\nu$ and $dx^\mu \wedge dx^\nu$ is antisymmetric: swapping the dummy indices shows the sum equals minus itself.

For a general form $\alpha = f\,dx^I$ (with dx^I a product of coordinate differentials), $d\alpha = df \wedge dx^I$ and $dd\alpha = ddf \wedge dx^I = 0$ by the same argument [Dray (15.26)].

Key idea: $d^2 = 0$

For every form α, $d(d\alpha) = 0$. In $\mathbb{R}^3$ this single statement is both $\nabla \times (\nabla f) = 0$ and $\nabla \cdot (\nabla \times \vec{F}) = 0$ [Dray (15.28)–(15.29)]. Geometrically it is dual to "the boundary of a boundary is empty": by Stokes, $\int_R dd\alpha = \int_{\partial R} d\alpha = \int_{\partial\partial R} \alpha = 0$ for every region R, because $\partial\partial R = \varnothing$.



edge BF is traversed upward
by the front face (red) and
downward by the right face (green)

Figure T1.17: Why $\partial\partial = \varnothing$. The boundary of a solid cube is its six faces, each oriented by the outward normal. The boundary of that closed surface is the sum of the faces' edges, but every edge is shared by two faces which traverse it in opposite directions (red and green), so the total is zero. Through Stokes, this is the same fact as $d^2 = 0$.

T1.5.6 The product rule

Working definition T1.15 (graded Leibniz rule)

For $\alpha \in \Omega^p$ [a p-form, a sum of terms $f\,dx^{\mu_1} \wedge \cdots \wedge dx^{\mu_p}$] and any form β,

$$d(\alpha \wedge \beta) = d\alpha \wedge \beta + (-1)^p\, \alpha \wedge d\beta. \tag{T1.46}$$

The sign appears because d has to be "moved past" the p one-form factors of α to act on β.

For monomials $\alpha = f\,dx^I$ (p factors) and $\beta = h\,dx^J$ [Dray (15.31)]:

$$\begin{aligned} d(\alpha \wedge \beta) &\overset{(1)}{=} d(fh) \wedge dx^I \wedge dx^J \overset{(2)}{=} (h\,df + f\,dh) \wedge dx^I \wedge dx^J \\ &\overset{(3)}{=} (df \wedge dx^I) \wedge (h\,dx^J) + (-1)^p\,(f\,dx^I) \wedge (dh \wedge dx^J) = d\alpha \wedge \beta + (-1)^p \alpha \wedge d\beta. \end{aligned}$$

(1) $\alpha \wedge \beta = fh\,dx^I \wedge dx^J$, then definition (T1.40).
(2) Ordinary product rule for functions.
(3) In the second term, move the 1-form dh past the p factors of dx^I; each swap costs -1.

Worked example T1.29: Two vector identities for free

(a) For a function f and a 1-form G in $\mathbb{R}^3$: $\star d(fG) = \star(df \wedge G) + f \star dG$. With $\star(F \wedge G) \leftrightarrow \vec{F} \times \vec{G}$ (T1.37) this is $\nabla \times (f\vec{G}) = \nabla f \times \vec{G} + f\,\nabla \times \vec{G}$.

(b) For two 1-forms F, G, the divergence of the cross product $\star(F \wedge G)$ is [Dray (15.43)]

$$\star \mathrm{d} \star \big(\star(F \wedge G)\big) \overset{(1)}{=} \star \mathrm{d}(F \wedge G) \overset{(2)}{=} \star(\mathrm{d}F \wedge G - F \wedge \mathrm{d}G)$$
$$\overset{(3)}{=} \star\big(G \wedge \mathrm{d}F\big) - \star\big(F \wedge \mathrm{d}G\big) \overset{(4)}{=} \vec{G} \cdot (\nabla \times \vec{F}) - \vec{F} \cdot (\nabla \times \vec{G}).$$

(1) $\star\star = 1$ in $\mathbb{R}^3$ (Section T1.4).
(2) Leibniz rule (T1.46) with $p = 1$.
(3) A 2-form commutes with a 1-form (T1.23).
(4) 1-form $\wedge$ flux = dot product $\times$ volume (Example T1.14), $\mathrm{d}F$ is the flux of $\nabla \times \vec{F}$, and $\star\mathrm{vol} = 1$.

So $\nabla \cdot (\vec{F} \times \vec{G}) = \vec{G} \cdot (\nabla \times \vec{F}) - \vec{F} \cdot (\nabla \times \vec{G})$, an identity that is usually proved by grinding through the ε_{ijk}'s.

T1.5.7 Uniqueness, and a chart-free formula

Key idea: d is unique

There is exactly one operator $\mathrm{d} : \Omega^p \to \Omega^{p+1}$ that (i) is linear, (ii) obeys the graded Leibniz rule (T1.46), (iii) satisfies $\mathrm{d}^2 = 0$, and (iv) gives the differential $\mathrm{d}f$ on functions [Dray §15.12].

Why: by (ii) with $p = 0$, $\mathrm{d}(f\,\mathrm{d}x^I) = \mathrm{d}f \wedge \mathrm{d}x^I + f\,\mathrm{d}(\mathrm{d}x^I)$. By (iv) and (iii), $\mathrm{d}(\mathrm{d}x^\mu) = \mathrm{d}\mathrm{d}x^\mu = 0$, and then by (ii) repeatedly $\mathrm{d}(\mathrm{d}x^{\mu_1} \wedge \cdots \wedge \mathrm{d}x^{\mu_p}) = 0$. So any operator with the four properties is given by (T1.40) in every chart. Since the four properties make no reference to coordinates, the operator defined by (T1.40) in one chart agrees with the one defined in any other chart. This is the cleanest proof that d is coordinate independent.

There is also a formula for d that uses no coordinates at all. For two vector fields X, Y define their **commutator** (or Lie bracket) as the vector field $[X, Y]f = X(Yf) - Y(Xf)$; the second derivatives cancel, so this is again a first-order operator, with components $[X, Y]^\mu = X^\nu \partial_\nu Y^\mu - Y^\nu \partial_\nu X^\mu$ (we study it in Section T1.9). Then for a 1-form α,

$$\mathrm{d}\alpha(X, Y) = X\big(\alpha(Y)\big) - Y\big(\alpha(X)\big) - \alpha([X, Y]). \tag{T1.47}$$

Check in components:

$$X(\alpha_\nu Y^\nu) - Y(\alpha_\nu X^\nu) \overset{(1)}{=} X^\mu \partial_\mu \alpha_\nu\, Y^\nu + \alpha_\nu X^\mu \partial_\mu Y^\nu - Y^\mu \partial_\mu \alpha_\nu\, X^\nu - \alpha_\nu Y^\mu \partial_\mu X^\nu$$
$$\overset{(2)}{=} (\partial_\mu \alpha_\nu - \partial_\nu \alpha_\mu) X^\mu Y^\nu + \alpha_\nu [X, Y]^\nu \overset{(3)}{=} \mathrm{d}\alpha(X, Y) + \alpha([X, Y]).$$

(1) Product rule on each term.
(2) Rename dummies in the third term; collect the terms with derivatives of X, Y into the commutator.
(3) Components of $\mathrm{d}\alpha$.

For coordinate vector fields $[\partial_\mu, \partial_\nu] = 0$ and (T1.47) reduces to the component formula. The formula (T1.47) is the one mathematicians take as the definition, and it is how curvature formulas are usually derived (Chapter T2).

T1.5.8 Closed and exact forms: a first look

Working definition T1.16 (closed, exact)

A form β is **closed** if $\mathrm{d}\beta = 0$ and **exact** if $\beta = \mathrm{d}\alpha$ for some globally defined form α [one smooth α on all of M, not a different one on each chart]. By $\mathrm{d}^2 = 0$ [applying d twice always gives zero, because mixed partial derivatives commute], every exact form is closed. *Test for closedness:* compute $\mathrm{d}\beta$. *Test for non-exactness:* find a closed curve (or surface, ...) over which β has a nonzero integral [by Stokes, $\oint_C \mathrm{d}\alpha = 0$ for every closed curve C, so an exact β integrates to zero there].

Why does a nonzero loop integral rule out exactness? If $\beta = \mathrm{d}f$ with f a smooth function everywhere, then around a closed loop γ the fundamental theorem of calculus gives $\oint_\gamma \mathrm{d}f = f(\text{end}) - f(\text{start}) = 0$.

Worked example T1.30: $\mathrm{d}\phi$ on the punctured plane: closed but not exact

On $\mathbb{R}^2 \setminus \{0\}$ consider

$$\beta = \frac{x\,\mathrm{d}y - y\,\mathrm{d}x}{x^2 + y^2}.$$

Where the angle ϕ is defined, $\beta = \mathrm{d}\phi$ (this is $r^2\mathrm{d}\phi = x\,\mathrm{d}y - y\,\mathrm{d}x$ from Example T1.12, divided by r^2). But β itself is defined on the whole punctured plane, while ϕ is not. Is it closed?

$$\mathrm{d}\beta \overset{(1)}{=} \mathrm{d}\Big(\frac{x}{r^2}\Big) \wedge \mathrm{d}y - \mathrm{d}\Big(\frac{y}{r^2}\Big) \wedge \mathrm{d}x \overset{(2)}{=} \partial_x\Big(\frac{x}{r^2}\Big)\mathrm{d}x \wedge \mathrm{d}y - \partial_y\Big(\frac{y}{r^2}\Big)\mathrm{d}y \wedge \mathrm{d}x$$
$$\overset{(3)}{=} \Big[\frac{r^2 - 2x^2}{r^4} + \frac{r^2 - 2y^2}{r^4}\Big]\mathrm{d}x \wedge \mathrm{d}y \overset{(4)}{=} \frac{2r^2 - 2(x^2+y^2)}{r^4}\,\mathrm{d}x \wedge \mathrm{d}y = 0.$$

(1) Definition (T1.40), with $r^2 = x^2 + y^2$.
(2) Only the $\mathrm{d}x$ part of $\mathrm{d}(x/r^2)$ survives against $\mathrm{d}y$, and only the $\mathrm{d}y$ part of $\mathrm{d}(y/r^2)$ against $\mathrm{d}x$.
(3) Quotient rule, $\partial_x(x/r^2) = (r^2 - 2x^2)/r^4$, and $-\mathrm{d}y \wedge \mathrm{d}x = +\mathrm{d}x \wedge \mathrm{d}y$.
(4) Add the fractions.

So β is closed. Now integrate around the unit circle $(x, y) = (\cos t, \sin t)$, $t \in [0, 2\pi]$, where $\mathrm{d}x = -\sin t\,\mathrm{d}t$, $\mathrm{d}y = \cos t\,\mathrm{d}t$ (this "substitute the curve" step is the pullback of Section T1.8):

$$\oint \beta = \int_0^{2\pi} \big(\cos t \cos t - \sin t(-\sin t)\big)\mathrm{d}t = 2\pi \neq 0.$$

So β is *not* exact on $\mathbb{R}^2 \setminus \{0\}$: there is no smooth single-valued function whose differential it is. The would-be potential ϕ jumps by 2π around the origin, exactly the jump we met in the transition map of Example T1.1. The space $\mathbb{R}^2 \setminus \{0\}$ has a hole, and the closed-but-not-exact form β detects it. This is the prototype of de Rham cohomology: $[\beta]$ is a nonzero element of $H^1_{\mathrm{dR}}(\mathbb{R}^2 \setminus \{0\})$ (Chapter T2).
Contrast $r^2\mathrm{d}\phi = x\,\mathrm{d}y - y\,\mathrm{d}x$ from Figure T1.10: its derivative is $2\,\mathrm{d}x \wedge \mathrm{d}y \neq 0$, so it is not even closed, and it is not d of anything on any region at all. That is why its stacks could not be fitted together into a contour map.

Key idea: Poincaré lemma

On a region that is *star-shaped* (every point can be joined to a fixed centre by a straight segment inside the region, e.g. $\mathbb{R}^n$ or a ball), every closed form is exact. For a closed 1-form α an explicit potential is

$$f(x) = \int_0^1 x^\mu \, \alpha_\mu(tx)\, \mathrm{d}t, \tag{T1.48}$$

the line integral of α along the straight segment from 0 to x.

Why it works:

$$\partial_\nu f(x) \overset{(1)}{=} \int_0^1 \big[\alpha_\nu(tx) + x^\mu\, t\, (\partial_\nu \alpha_\mu)(tx)\big]\mathrm{d}t \overset{(2)}{=} \int_0^1 \big[\alpha_\nu(tx) + t\, x^\mu (\partial_\mu \alpha_\nu)(tx)\big]\mathrm{d}t$$
$$\overset{(3)}{=} \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t}\big[t\, \alpha_\nu(tx)\big]\mathrm{d}t \overset{(4)}{=} \alpha_\nu(x).$$

(1) Differentiate under the integral; the chain rule gives a factor t from the argument tx.
(2) α is closed: $\partial_\nu \alpha_\mu = \partial_\mu \alpha_\nu$.
(3) Product rule: $\frac{\mathrm{d}}{\mathrm{d}t}[t\, \alpha_\nu(tx)] = \alpha_\nu(tx) + t\, x^\mu (\partial_\mu \alpha_\nu)(tx)$.
(4) Fundamental theorem of calculus: $[t\, \alpha_\nu(tx)]_0^1$.

The punctured plane is not star-shaped (the segment to the centre may pass through the hole), and that is exactly where Example T1.30 escapes the lemma. Topology is the study of such escapes.

Worked example T1.31: Finding a potential with the Poincaré formula

Let $\alpha = 2xy\,\mathrm{d}x + x^2\,\mathrm{d}y$ on $\mathbb{R}^2$. It is closed: $\mathrm{d}\alpha = 2x\,\mathrm{d}y \wedge \mathrm{d}x + 2x\,\mathrm{d}x \wedge \mathrm{d}y = 0$. Then

$$f(x,y) \overset{(1)}{=} \int_0^1 \big[x \cdot 2(tx)(ty) + y \cdot (tx)^2\big]\mathrm{d}t \overset{(2)}{=} 3x^2 y \int_0^1 t^2\,\mathrm{d}t = x^2 y.$$

(1) (T1.48) with $\alpha_x(tx,ty) = 2t^2xy$ and $\alpha_y(tx,ty) = t^2x^2$.
(2) Collect: $2t^2x^2y + t^2x^2y = 3t^2x^2y$.

Check: $\mathrm{d}(x^2y) = 2xy\,\mathrm{d}x + x^2\,\mathrm{d}y$. In physics language, a curl-free force on a simply connected region is conservative, and this formula constructs its potential energy (up to sign and a constant).

Reading Nash (p. 13, (1.40)–(1.43))

Nash introduces cochain complexes abstractly, $\cdots \to E^{n-1} \to E^n \to E^{n+1} \to \dots$ with $d_n \circ d_{n-1} = 0$, and then says: take $E^n = \Omega^n(X)$ and d the exterior derivative; "the resulting complex is the de Rham complex" with cohomology $H^n_{\text{de Rham}}(X) = \ker d_n / \operatorname{Im} d_{n-1}$. Everything in that sentence is now concrete: $\Omega^n(X)$ is the space of n-forms of Section T1.3, d is (T1.40), the complex property is $\mathrm{d}^2 = 0$ (T1.45), $\ker d$ is the closed forms, $\operatorname{Im} d$ the exact ones, and Example T1.30 exhibits a nonzero class in $H^1_{\text{de Rham}}(\mathbb{R}^2 \setminus \{0\})$. On p. 14 he adds that the cup product in de Rham cohomology is $[\omega] \cup [\nu] = [\omega \wedge \nu]$; the Leibniz rule (T1.46) is what makes this well defined (closed $\wedge$ closed is closed, and closed $\wedge$ exact is exact: $\omega \wedge \mathrm{d}\alpha = \pm\mathrm{d}(\omega \wedge \alpha)$ when $\mathrm{d}\omega = 0$).

Recap

- $\mathrm{d}(f\,\mathrm{d}x^I) = \mathrm{d}f \wedge \mathrm{d}x^I$; for 1-forms $(\mathrm{d}\alpha)_{\mu\nu} = \partial_\mu\alpha_\nu - \partial_\nu\alpha_\mu$. Antisymmetry is what makes it chart independent.
- $\mathrm{d}\alpha(v,w)$ is the circulation around the small parallelogram on v, w: infinitesimal Stokes.
- In $\mathbb{R}^3$: d is grad, curl, div on 0-, 1-, 2-forms; $\nabla\cdot\vec{F} = \star\mathrm{d}\star F$, $\nabla\times\vec{F} = \star\mathrm{d}F$.
- $\mathrm{d}^2 = 0$ (curl grad = 0, div curl = 0), dual to $\partial\partial = \varnothing$.
- $\mathrm{d}(\alpha\wedge\beta) = \mathrm{d}\alpha\wedge\beta + (-1)^p\alpha\wedge\mathrm{d}\beta$; d is the unique such operator.
- Exact $\Rightarrow$ closed; the converse holds on star-shaped regions (Poincaré) but fails on $\mathbb{R}^2\setminus\{0\}$, where $\mathrm{d}\phi = (x\,\mathrm{d}y - y\,\mathrm{d}x)/r^2$ is closed and not exact.

T1.6 The codifferential and the Laplace–de Rham operator

The exterior derivative raises the rank of a form: grad takes functions to vectors, curl takes vectors to fluxes. Vector calculus also has operators that go the other way. The divergence takes a vector field to a function, and the Laplacian $\nabla^2 = \nabla\cdot\nabla$ is "go up with grad, come back with div". In the language of forms, the rank-lowering operator is the *codifferential* δ, and the Laplacian on forms of any rank is $\Delta = \mathrm{d}\delta + \delta\mathrm{d}$. Nash defines his Laplacian (2.42) exactly this way, using the adjoint d^* of d; the whole of Hodge theory, the de Rham complex as an elliptic complex, and the Euler characteristic as an index (Chapters 2 and 4 of Nash) rest on it. Our job here is to make δ and Δ completely explicit: signs, components, polar and spherical formulas, and the symbol.

T1.6.1 The adjoint of d: integration by parts

In quantum mechanics the adjoint $A^\dagger$ of an operator is defined by $\langle\psi, A\phi\rangle = \langle A^\dagger\psi, \phi\rangle$, and for $A = \mathrm{d}/\mathrm{d}x$ integration by parts gives $A^\dagger = -\mathrm{d}/\mathrm{d}x$ (the boundary terms vanish for decaying or periodic wavefunctions). We do the same with d, using the inner product (T1.38) on forms, $\langle\alpha,\beta\rangle = \int_M \alpha\wedge\star\beta$, on a compact oriented manifold M without boundary.

Working definition T1.17 (codifferential)

The **codifferential** $\delta : \Omega^p \to \Omega^{p-1}$ [it lowers the rank by one, e.g. $\delta = -\mathrm{div}$ on 1-forms in $\mathbb{R}^3$] (Nash's d^* [the adjoint of d, found by integrating by parts as $\mathrm{d}/\mathrm{d}x$ gives $-\mathrm{d}/\mathrm{d}x$]) is the operator with

$$\langle\mathrm{d}\alpha,\beta\rangle = \langle\alpha,\delta\beta\rangle \qquad \text{for all } \alpha\in\Omega^{p-1},\ \beta\in\Omega^p. \tag{T1.49}$$

It is given by

$$\delta = (-1)^{n(p+1)+1+s}\ \star\,\mathrm{d}\,\star \qquad \text{on } p\text{-forms}, \tag{T1.50}$$

with $n = \dim M$ and s the number of minus signs in the metric. By convention $\delta f = 0$ on functions.

Here is the derivation. Take $\alpha \in \Omega^{p-1}$ and $\beta \in \Omega^p$, so that $\alpha \wedge \star\beta$ is an $(n-1)$-form. Then

$$\begin{aligned} 0 &\overset{(1)}{=} \int_M \mathrm{d}(\alpha \wedge \star\beta) \overset{(2)}{=} \int_M \mathrm{d}\alpha \wedge \star\beta + (-1)^{p-1} \int_M \alpha \wedge \mathrm{d}\star\beta \\ &\overset{(3)}{=} \langle \mathrm{d}\alpha, \beta\rangle + (-1)^{p-1}\varepsilon \int_M \alpha \wedge \star\big(\star \mathrm{d}\star\beta\big) \overset{(4)}{=} \langle \mathrm{d}\alpha, \beta\rangle - (-1)^p \varepsilon\, \langle \alpha, \star \mathrm{d}\star\beta\rangle. \end{aligned} \tag{T1.51}$$

(1) Stokes' theorem: $\int_M \mathrm{d}(\cdot) = \int_{\partial M}(\cdot) = 0$ because $\partial M = \varnothing$ (Chapter T2).
(2) Leibniz rule (T1.46) with the $(p-1)$-form α in front.
(3) The first term is the inner product by definition. In the second, $\gamma = \mathrm{d}\star\beta$ is an $(n-p+1)$-form, and by (T1.35) $\gamma = \varepsilon\ \star\star\gamma$ with $\varepsilon = (-1)^{(n-p+1)(p-1)+s}$ (the sign is its own inverse).
(4) Definition (T1.38) of $\langle \alpha, \star\gamma\rangle$; $(-1)^{p-1} = -(-1)^p$.

So $\langle \mathrm{d}\alpha, \beta\rangle = \langle \alpha, (-1)^p \varepsilon\ \star \mathrm{d}\star\beta\rangle$. The sign simplifies modulo 2:

$$\begin{aligned} p + (n-p+1)(p-1) + s &\overset{(1)}{=} p + n(p-1) - (p-1)^2 + s \overset{(2)}{=} p + n(p-1) - (p-1) + s \\ &= n(p-1) + 1 + s \overset{(3)}{=} n(p+1) + 1 + s \pmod 2. \end{aligned}$$

(1) Expand $(n-(p-1))(p-1)$.
(2) $k^2 \equiv k \pmod 2$ for any integer k.
(3) $n(p+1) - n(p-1) = 2n$ is even.

This is (T1.50). For Riemannian signature ($s = 0$) it reads $\delta = (-1)^{np+n+1} \star \mathrm{d} \star$, which is precisely Nash's formula in (2.43).

Dictionary: signs of $\delta = \pm \star \mathrm{d} \star$

space	sign in $\delta = \pm \star \mathrm{d} \star$	comment
n even, Riemannian	$-$ for all p	$\mathbb{R}^2, S^2, T^2, \mathbb{R}^4, S^4$
n odd, Riemannian	$(-1)^p$	$\mathbb{R}^3$: $-$ on 1-forms, $+$ on 2-forms, $-$ on 3-forms
Minkowski $\mathbb{R}^{1,3}$	$+$ for all p	$n = 4$, $s = 1$

Key idea: δ is minus the divergence

On 1-forms in any Riemannian manifold, $\delta\alpha = -\nabla \cdot \alpha^\sharp$:

$$\delta\alpha = -\frac{1}{\sqrt{g}}\, \partial_\mu\big(\sqrt{g}\, g^{\mu\nu}\alpha_\nu\big). \tag{T1.52}$$

On p-forms, $(\delta\alpha)_{\nu_2\ldots\nu_p} = -\nabla^\mu \alpha_{\mu\nu_2\ldots\nu_p}$, "contract a derivative into the first index", where ∇ is the covariant derivative (Chapter T2); in flat Cartesian coordinates it is just $-\partial^\mu \alpha_{\mu\nu_2\ldots}$.

Why (T1.52): for $\alpha = \alpha_\nu \mathrm{d}x^\nu$ write $\alpha^\mu = g^{\mu\nu}\alpha_\nu$. The component formula (T1.36) gives $\star\alpha = \sum_\mu (-1)^{\mu-1}\sqrt{g}\,\alpha^\mu\, \mathrm{d}x^1 \wedge \cdots \wedge \widehat{\mathrm{d}x^\mu} \wedge \cdots \wedge \mathrm{d}x^n$, since $\varepsilon_{\mu 1\ldots\hat{\mu}\ldots n} = (-1)^{\mu-1}$. Then

$$\begin{aligned} \mathrm{d}\star\alpha &\overset{(1)}{=} \sum_\mu (-1)^{\mu-1}\partial_\mu(\sqrt{g}\,\alpha^\mu)\, \mathrm{d}x^\mu \wedge \mathrm{d}x^1 \wedge \cdots \wedge \widehat{\mathrm{d}x^\mu} \wedge \cdots \wedge \mathrm{d}x^n \\ &\overset{(2)}{=} \partial_\mu(\sqrt{g}\,\alpha^\mu)\, \mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^n \overset{(3)}{=} \frac{1}{\sqrt{g}}\partial_\mu(\sqrt{g}\,\alpha^\mu)\ \mathrm{vol}, \end{aligned}$$

(1) Only the derivative along the missing direction x^μ survives the wedge.

(2) Moving dx^μ into its place costs $(\mu-1)$ swaps, which cancels the sign $(-1)^{\mu-1}$.
(3) $\mathrm{vol} = \sqrt{g}\,dx^1\wedge\cdots\wedge dx^n$ (T1.31).

and $\star\mathrm{vol} = 1$, so $\star d\star\alpha = \frac{1}{\sqrt{g}}\partial_\mu(\sqrt{g}\,\alpha^\mu)$, the covariant divergence. With the sign -1 that the table gives on 1-forms (for either parity of n), (T1.52) follows.

Watch out: Why the minus sign?

It is the same minus sign as in $(d/dx)^\dagger = -d/dx$. In one dimension on the circle, $\delta(g\,d\theta) = -g'(\theta)$ and $\langle df, g\,d\theta\rangle = \int f'g\,d\theta = -\int fg'\,d\theta = \langle f, \delta(g\,d\theta)\rangle$. Keeping this sign is what makes $\Delta = d\delta+\delta d$ a *positive* operator, like $-d^2/dx^2 = p^2$ in quantum mechanics.

Two properties follow at once. First, $\delta^2 = 0$: $\delta\delta \propto \star d\star\star d\star = \pm\star dd\star = 0$, using $\star\star = \pm1$. In $\mathbb{R}^3$ this is again $\nabla\cdot(\nabla\times\vec{F}) = 0$ read backwards. Second, a form with $\delta\beta = 0$ is called **co-closed**; a form $\beta = \delta\gamma$ is **co-exact**.

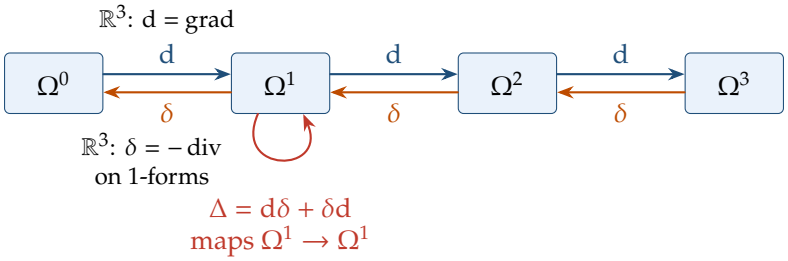


Figure T1.18: The exterior derivative d raises rank by one; its adjoint δ lowers it by one. Neither preserves rank, but the combination $\Delta = d\delta + \delta d$ does: go up and come back, or come down and go back up. On functions only the first route exists and $\Delta = \delta d = -\nabla^2$.

T1.6.2 The Laplace–de Rham operator

Working definition T1.18 (Laplace–de Rham operator)

The **Laplacian** on p-forms is

$$\Delta = d\delta + \delta d : \Omega^p \to \Omega^p \tag{T1.53}$$

(Nash's $\Delta_p = (dd^* + d^*d)_p$, (2.42)). On functions, $\Delta f = \delta df$ [$\delta f = 0$ on functions, so only the route up with d and back down with δ survives], and by (T1.52) this is the **Laplace–Beltrami operator** [the Laplacian built from the metric $g_{\mu\nu}$, with $\sqrt{g} = \sqrt{\det g_{\mu\nu}}$]

$$\Delta f = -\frac{1}{\sqrt{g}}\,\partial_\mu\big(\sqrt{g}\,g^{\mu\nu}\partial_\nu f\big), \qquad \text{flat Cartesian: } \Delta f = -\nabla^2 f. \tag{T1.54}$$

Worked example T1.32: The Laplacian in polar coordinates

Following Dray [Dray (15.22)], but keeping our sign, $\Delta f = -\star\,d\star\,df$ in $\mathbb{R}^2$. Use the orthonormal coframe $\{dr, r\,d\phi\}$ and the stars $\star dr = r\,d\phi$, $\star(r\,d\phi) = -dr$, $\star(dr\wedge d\phi) = 1/r$ from Example T1.20:

$$df \overset{(1)}{=} f_r\,dr + f_\phi\,d\phi = f_r\,dr + \frac{f_\phi}{r}\,(r\,d\phi),$$

$$\star \mathrm{d}f \overset{(2)}{=} f_r\, r\, \mathrm{d}\phi - \frac{f_\phi}{r}\,\mathrm{d}r,$$
$$\mathrm{d}\star\mathrm{d}f \overset{(3)}{=} \partial_r(r f_r)\,\mathrm{d}r\wedge\mathrm{d}\phi - \partial_\phi\Big(\frac{f_\phi}{r}\Big)\mathrm{d}\phi\wedge\mathrm{d}r \overset{(4)}{=} \Big[\partial_r(r f_r) + \frac{f_{\phi\phi}}{r}\Big]\mathrm{d}r\wedge\mathrm{d}\phi,$$
$$\Delta f = -\star\,\mathrm{d}\star\mathrm{d}f \overset{(5)}{=} -\Big[\frac{1}{r}\,\partial_r\big(r\,\partial_r f\big) + \frac{1}{r^2}\,\partial_\phi^2 f\Big].$$

(1) $\mathrm{d}f$ in coordinates; then rewrite with the unit coframe.
(2) Star each coframe element; linearity.
(3) Definition of d; in each term only the derivative along the missing coordinate survives.
(4) $\mathrm{d}\phi\wedge\mathrm{d}r = -\mathrm{d}r\wedge\mathrm{d}\phi$, and $\partial_\phi(f_\phi/r) = f_{\phi\phi}/r$.
(5) $\star(\mathrm{d}r\wedge\mathrm{d}\phi) = 1/r$; overall sign -1 from (T1.50) with $n = 2$.

The bracket is the familiar ∇^2 in polar coordinates. Checks: $\Delta(\ln r) = -\frac{1}{r}\partial_r(r\cdot\frac{1}{r}) = 0$ for $r \neq 0$ ($\ln r$ is the 2D Coulomb potential); $\Delta(r^m\cos m\phi) = -(m^2 r^{m-2} - m^2 r^{m-2})\cos m\phi = 0$ (these are $\mathrm{Re}\, z^m$, harmonic because holomorphic, see Section T1.10).

Worked example T1.33: The Laplacian on the round S^2 and spherical harmonics

On S^2 with $\mathrm{vol} = \sin\theta\,\mathrm{d}\theta\wedge\mathrm{d}\phi$ and coframe $\{\mathrm{d}\theta, \sin\theta\,\mathrm{d}\phi\}$ we have $\star\mathrm{d}\theta = \sin\theta\,\mathrm{d}\phi$, $\star\mathrm{d}\phi = -\mathrm{d}\theta/\sin\theta$ and $\star(\mathrm{d}\theta\wedge\mathrm{d}\phi) = 1/\sin\theta$. Exactly as above,

$$\star\mathrm{d}f \overset{(1)}{=} \sin\theta\, f_\theta\,\mathrm{d}\phi - \frac{f_\phi}{\sin\theta}\,\mathrm{d}\theta,$$
$$\mathrm{d}\star\mathrm{d}f \overset{(2)}{=} \Big[\partial_\theta(\sin\theta\, f_\theta) + \frac{f_{\phi\phi}}{\sin\theta}\Big]\mathrm{d}\theta\wedge\mathrm{d}\phi,$$
$$\Delta f \overset{(3)}{=} -\frac{1}{\sin\theta}\,\partial_\theta\big(\sin\theta\,\partial_\theta f\big) - \frac{1}{\sin^2\theta}\,\partial_\phi^2 f.$$

(1) Star of $f_\theta\mathrm{d}\theta + f_\phi\mathrm{d}\phi$.
(2) Differentiate; $\mathrm{d}\phi\wedge\mathrm{d}\theta = -\mathrm{d}\theta\wedge\mathrm{d}\phi$ turns the minus into a plus.
(3) $\Delta = -\star\,\mathrm{d}\star\mathrm{d}$ (even dimension) and $\star(\mathrm{d}\theta\wedge\mathrm{d}\phi) = 1/\sin\theta$.

This is $\hat{L}^2$, the squared angular momentum of quantum mechanics (in units $\hbar = 1$). Two eigenfunctions:
(a) $Z = \cos\theta$: $\partial_\theta Z = -\sin\theta$, so $\partial_\theta(\sin\theta\cdot(-\sin\theta)) = -2\sin\theta\cos\theta$, and $\Delta Z = 2\cos\theta = 2Z$. Eigenvalue $\ell(\ell+1) = 2$ with $\ell = 1$. This explains the ratio $\langle \mathrm{d}Z, \mathrm{d}Z\rangle/\langle Z, Z\rangle = 2$ of Example T1.25: $\langle \mathrm{d}Z, \mathrm{d}Z\rangle = \langle Z, \delta\mathrm{d}Z\rangle = \langle Z, \Delta Z\rangle = 2\langle Z, Z\rangle$.
(b) $f = \sin^2\theta\cos 2\phi = X^2 - Y^2$:

$$\frac{1}{\sin\theta}\partial_\theta\big(\sin\theta\,\partial_\theta f\big) \overset{(1)}{=} \frac{\cos 2\phi}{\sin\theta}\,\partial_\theta\big(2\sin^2\theta\cos\theta\big) \overset{(2)}{=} 2(2\cos^2\theta - \sin^2\theta)\cos 2\phi,$$
$$\Delta f \overset{(3)}{=} -\big[4\cos^2\theta - 2\sin^2\theta - 4\big]\cos 2\phi \overset{(4)}{=} 6\sin^2\theta\cos 2\phi = 6f.$$

(1) $\partial_\theta f = 2\sin\theta\cos\theta\cos 2\phi$.
(2) $\partial_\theta(\sin^2\theta\cos\theta) = 2\sin\theta\cos^2\theta - \sin^3\theta$; divide by $\sin\theta$.
(3) Add $\frac{1}{\sin^2\theta}\partial_\phi^2 f = -4\cos 2\phi$ and put in the overall minus sign.
(4) $4\cos^2\theta - 4 = -4\sin^2\theta$.

Eigenvalue $6 = \ell(\ell+1)$ with $\ell = 2$. In general the eigenvalues are $\ell(\ell+1)$ with multiplicity $2\ell+1$ (Figure T1.20).

Worked example T1.34: The Laplacian on 1-forms in $\mathbb{R}^3$

For a 1-form $F \leftrightarrow \vec{F}$ in $\mathbb{R}^3$, the table gives $\delta = -\star\mathrm{d}\star$ on 1-forms and $\delta = +\star\mathrm{d}\star$ on 2-forms. Then

$$\begin{aligned} \mathrm{d}\delta F &\overset{(1)}{=} -\mathrm{d}(\star\mathrm{d}\star F) \overset{(2)}{=} -\nabla(\nabla\cdot\vec{F})\cdot\mathrm{d}\vec{r}, \\ \delta\mathrm{d}F &\overset{(3)}{=} \star\mathrm{d}\star(\mathrm{d}F) \overset{(4)}{=} \star\mathrm{d}\big((\nabla\times\vec{F})\cdot\mathrm{d}\vec{r}\big) \overset{(5)}{=} \big(\nabla\times(\nabla\times\vec{F})\big)\cdot\mathrm{d}\vec{r}, \\ \Delta F &\overset{(6)}{=} \big(\nabla\times\nabla\times\vec{F} - \nabla(\nabla\cdot\vec{F})\big)\cdot\mathrm{d}\vec{r} \overset{(7)}{=} -(\nabla^2\vec{F})\cdot\mathrm{d}\vec{r}. \end{aligned}$$

(1) δ on the 1-form F.
(2) $\star\mathrm{d}\star F = \nabla\cdot\vec{F}$ is a function; d of it is the gradient.
(3) δ on the 2-form $\mathrm{d}F$.
(4) $\star\mathrm{d}F$ is the curl as a 1-form (dictionary of Section T1.5).
(5) $\star\mathrm{d}$ of a 1-form is its curl.
(6) Add the two pieces.
(7) The vector identity $\nabla\times(\nabla\times\vec{F}) = \nabla(\nabla\cdot\vec{F}) - \nabla^2\vec{F}$.

So on 1-forms in flat space Δ is minus the componentwise Laplacian, as on functions. In curved space the two routes differ by curvature terms (the Weitzenböck formula), which is where the Laplacian "sees" geometry. Physicists meet $\nabla\times\nabla\times\vec{A} - \nabla(\nabla\cdot\vec{A})$ as the operator of magnetostatics.

T1.6.3 Positivity, harmonic forms and Hodge's theorem

On a compact manifold without boundary, with Riemannian metric,

$$\langle\omega,\Delta\omega\rangle \overset{(1)}{=} \langle\omega,\mathrm{d}\delta\omega\rangle + \langle\omega,\delta\mathrm{d}\omega\rangle \overset{(2)}{=} \langle\delta\omega,\delta\omega\rangle + \langle\mathrm{d}\omega,\mathrm{d}\omega\rangle \overset{(3)}{=} \|\delta\omega\|^2 + \|\mathrm{d}\omega\|^2 \;\geq 0. \tag{T1.55}$$

(1) Definition (T1.53); linearity.
(2) Move d across as δ in the first term and δ across as d in the second, by (T1.49) (the inner product is symmetric).
(3) $\langle\beta,\beta\rangle = \int|\beta|^2\mathrm{vol} \geq 0$ in Riemannian signature.

This is Nash's argument (2.50). A form with $\Delta\omega = 0$ is called **harmonic**, and (T1.55) shows that

$$\Delta\omega = 0 \iff \mathrm{d}\omega = 0 \text{ and } \delta\omega = 0, \tag{T1.56}$$

since a sum of two non-negative numbers vanishes only if both do (Nash (2.51)). Harmonic forms are closed, so they define de Rham classes.

Key idea: Hodge's theorem (Nash p. 41)

On a compact oriented Riemannian manifold without boundary, every de Rham cohomology class contains *exactly one* harmonic form. Hence the Betti numbers are $b_p = \dim\mathrm{Ker}\,\Delta_p$: the number of zero modes of the Laplacian on p-forms. Moreover every p-form splits uniquely and orthogonally as

$$\omega = \mathrm{d}\alpha + \delta\beta + h, \qquad \Delta h = 0 \tag{T1.57}$$

(the **Hodge decomposition**).

Why at most one: if h_1, h_2 are harmonic and in the same class, $h_1 - h_2 = \mathrm{d}\alpha$ is exact and harmonic, so $\|\mathrm{d}\alpha\|^2 = \langle \mathrm{d}\alpha, \mathrm{d}\alpha\rangle = \langle \alpha, \delta(h_1 - h_2)\rangle = 0$. *Why the three pieces of* (T1.57) *are orthogonal:* $\langle \mathrm{d}\alpha, \delta\beta\rangle = \langle \mathrm{dd}\alpha, \beta\rangle = 0$, $\langle \mathrm{d}\alpha, h\rangle = \langle \alpha, \delta h\rangle = 0$ and $\langle \delta\beta, h\rangle = \langle \beta, \mathrm{d}h\rangle = 0$. *Existence* (that the three pieces fill everything) is the hard analytic part: it uses the ellipticity of Δ, the subject of Nash's Chapter 2 (Chapter 2). In $\mathbb{R}^3$ language (T1.57) is the Helmholtz decomposition $\vec{F} = \nabla\phi + \nabla \times \vec{A} + \vec{h}$; on a space with holes the harmonic piece $\vec{h}$ is the part that is neither a gradient nor a curl.

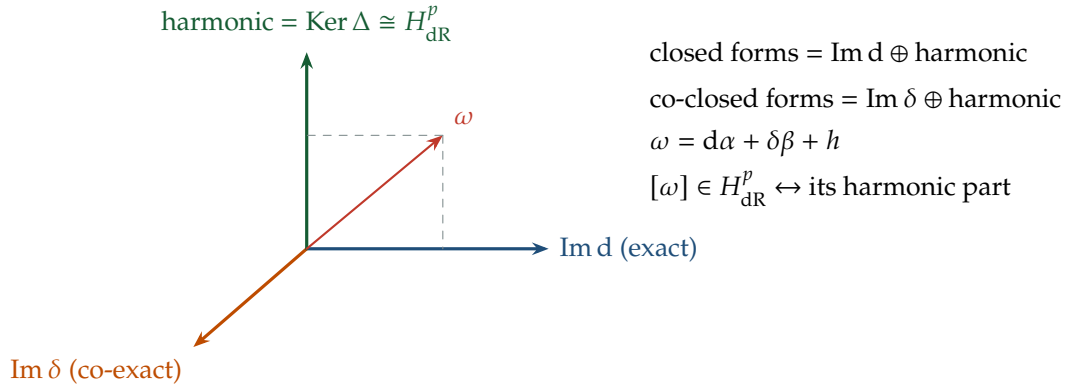


Figure T1.19: The Hodge decomposition drawn as three mutually orthogonal directions in the (infinite-dimensional) space $\Omega^p(M)$. A closed form ω has no co-exact part; its cohomology class is represented by its projection onto the harmonic axis. This is why "counting zero modes of Δ_p" computes the topological invariant b_p.

Worked example T1.35: Harmonic forms on S^1, T^2 and S^2

(a) *Circle.* On functions $\Delta = -\partial_\theta^2$ with eigenfunctions $e^{\mathrm{i}k\theta}$, eigenvalues k^2, $k \in \mathbb{Z}$. Zero modes: constants, so $b_0 = 1$. On 1-forms $f\,\mathrm{d}\theta$, $\Delta(f\,\mathrm{d}\theta) = \mathrm{d}\delta(f\,\mathrm{d}\theta) = \mathrm{d}(-f') = -f''\mathrm{d}\theta$; zero modes $c\,\mathrm{d}\theta$, so $b_1 = 1$. The harmonic representative of the class of the winding form is $\mathrm{d}\theta/2\pi$ (the form with period 1).
(b) *Torus* $T^2 = \mathbb{R}^2/\mathbb{Z}^2$, flat metric. Eigenfunctions $e^{2\pi\mathrm{i}(mx+ny)}$ with eigenvalues $4\pi^2(m^2+n^2)$. Harmonic forms: 1; $\mathrm{d}x, \mathrm{d}y$; $\mathrm{d}x \wedge \mathrm{d}y$ (constant coefficients are the only harmonic ones, since Δ acts componentwise on constant-coefficient forms in flat space). So $(b_0, b_1, b_2) = (1, 2, 1)$ and $\chi = 1 - 2 + 1 = 0$.
(c) *Sphere.* Harmonic functions: constants ($b_0 = 1$). Harmonic 2-forms: c vol, since $\star$ commutes with Δ and maps harmonic functions to harmonic 2-forms ($b_2 = 1$). Harmonic 1-forms: none. Indeed a harmonic ω is closed, and on S^2 every closed 1-form is exact (S^2 is simply connected; Chapter T2), $\omega = \mathrm{d}f$; then $0 = \delta\omega = \delta\mathrm{d}f = \Delta f$ forces f constant and $\omega = 0$. So $(b_0, b_1, b_2) = (1, 0, 1)$ and $\chi = 2$.

T1.6.4 The symbol of d, δ and Δ

How do we see that Δ is "elliptic", like $-\nabla^2$? Replace each derivative by the Fourier variable, $\partial_\mu \to \mathrm{i}k_\mu$, and keep only the highest-order terms. The result is the (principal) **symbol**, a matrix-valued function of the covector $k = k_\mu \mathrm{d}x^\mu$ (Chapter 2). For d, the recipe $\mathrm{d}\alpha = \mathrm{d}x^\mu \wedge \partial_\mu \alpha$ becomes

$$\sigma(\mathrm{d})(k)\,\omega = \mathrm{i}\,k \wedge \omega, \qquad \sigma(\delta)(k)\,\omega = -\mathrm{i}\,\iota_k\,\omega, \tag{T1.58}$$

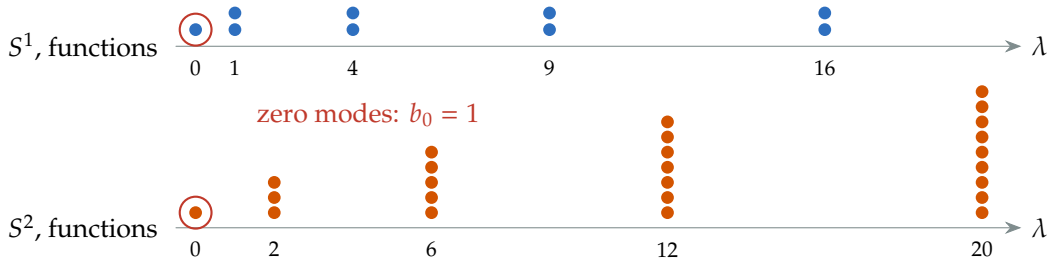


Figure T1.20: Spectra of the Laplacian on functions, each dot one eigenfunction. On S^1: $\lambda = k^2$ with multiplicity 2 (from $e^{\pm ik\theta}$) except for $k = 0$. On S^2: $\lambda = \ell(\ell+1)$ with multiplicity $2\ell+1$ (the spherical harmonics). The eigenvalues depend on the size and shape of the space; the number of zero modes (circled) does not. It is the Betti number b_0, the number of connected components.

where ι_k is the *interior product* with the vector $k^\sharp$ (insert it in the first slot, Section T1.9); the second formula follows from $(\delta\alpha)_{\nu\ldots} = -\partial^\mu \alpha_{\mu\nu\ldots}$. Then

$$\begin{aligned}\sigma(\Delta)(k) &\overset{(1)}{=} \sigma(\mathrm{d})\sigma(\delta) + \sigma(\delta)\sigma(\mathrm{d}) \overset{(2)}{=} (\mathrm{i}k\wedge)(-\mathrm{i}\iota_k) + (-\mathrm{i}\iota_k)(\mathrm{i}k\wedge)\\ &= k\wedge\iota_k + \iota_k\, k\wedge \overset{(3)}{=} |k|^2\,\mathbb{1}. \end{aligned} \tag{T1.59}$$

(1) The symbol of a product of operators is the product of the symbols (Nash (2.45)–(2.46)).
(2) (T1.58).
(3) The interior product is an antiderivation (Section T1.9): $\iota_k(k\wedge\omega) = (\iota_k k)\,\omega - k\wedge\iota_k\omega$, and $\iota_k k = g^{\mu\nu}k_\mu k_\nu = |k|^2$.

The symbol of Δ is $|k|^2$ times the identity, invertible for every $k \neq 0$: that is ellipticity. Nash states this as (2.49), $\sigma_2^{\Delta_p}(x,k) = k^2 I$.

Worked example T1.36: The symbol as explicit 4×4 matrices on $\mathbb{R}^2$

Order the basis of all forms on $\mathbb{R}^2$ as $(1,\ \mathrm{d}x,\ \mathrm{d}y,\ \mathrm{d}x\wedge\mathrm{d}y)$ and let $k = k_1\mathrm{d}x + k_2\mathrm{d}y$. Then $k\wedge$ sends $1 \mapsto k_1\mathrm{d}x + k_2\mathrm{d}y$, $\mathrm{d}x \mapsto k_2\,\mathrm{d}y\wedge\mathrm{d}x = -k_2\,\mathrm{d}x\wedge\mathrm{d}y$, $\mathrm{d}y \mapsto k_1\,\mathrm{d}x\wedge\mathrm{d}y$, $\mathrm{d}x\wedge\mathrm{d}y \mapsto 0$; and ι_k sends $1 \mapsto 0$, $\mathrm{d}x \mapsto k_1$, $\mathrm{d}y \mapsto k_2$, $\mathrm{d}x\wedge\mathrm{d}y \mapsto k_1\mathrm{d}y - k_2\mathrm{d}x$. As matrices (columns are images of basis elements):

$$E = k\wedge = \begin{pmatrix} 0 & 0 & 0 & 0\\ k_1 & 0 & 0 & 0\\ k_2 & 0 & 0 & 0\\ 0 & -k_2 & k_1 & 0 \end{pmatrix}, \qquad I = \iota_k = \begin{pmatrix} 0 & k_1 & k_2 & 0\\ 0 & 0 & 0 & -k_2\\ 0 & 0 & 0 & k_1\\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Multiplying out,

$$EI + IE \overset{(1)}{=} \begin{pmatrix} 0 & 0 & 0 & 0\\ 0 & k_1^2 & k_1k_2 & 0\\ 0 & k_1k_2 & k_2^2 & 0\\ 0 & 0 & 0 & k_1^2+k_2^2 \end{pmatrix} + \begin{pmatrix} k_1^2+k_2^2 & 0 & 0 & 0\\ 0 & k_2^2 & -k_1k_2 & 0\\ 0 & -k_1k_2 & k_1^2 & 0\\ 0 & 0 & 0 & 0 \end{pmatrix} = (k_1^2+k_2^2)\,\mathbb{1}_4.$$

(1) EI: e.g. its $(2,2)$ entry is (row 2 of E)·(column 2 of I) $= k_1\cdot k_1$, and its $(4,4)$ entry is $(-k_2)(-k_2) + k_1k_1$. IE: e.g. $(1,1)$ entry $= k_1k_1 + k_2k_2$ and $(2,2)$ entry $= (-k_2)(-k_2)$.

The off-diagonal terms cancel between the two products, exactly as the antiderivation argument promised. Notice also $E^2 = 0$ and $I^2 = 0$ (the symbols of $\mathrm{d}^2 = 0$ and $\delta^2 = 0$), so the operator $Q = E - I$ satisfies $Q^2 = -(EI + IE) = -|k|^2$. Up to the factor i, Q is the symbol of $\mathrm{d} + \delta$, a "square root" of the Laplacian. Matrices that square to $\pm|k|^2$ are Clifford algebra

generators (Chapter T4): $\mathrm{d}+\delta$ is a Dirac-type operator on forms, and it is the operator of the de Rham complex whose index Nash computes to be the Euler characteristic (Chapter 4).

Reading Nash (pp. 39–41, (2.42)–(2.53))

Nash writes $\Delta_p = (dd^* + d^*d)_p$ (2.42), gives $d^* = (-1)^{np+n+1} \star d \star$ and the inner product $\langle \omega, v \rangle = \int \omega \wedge \star v$ in (2.43), and computes the symbol $\sigma_2^{\Delta_p}(x,k) = k^2 I$ in (2.46)–(2.49) using $\sigma_1^d(x,k) : v \mapsto k \wedge v$ and its adjoint, "interior product with k". He drops the factors of i that we kept in (T1.58): they cancel in Δ, as our chain shows. His (2.50)–(2.51) is our (T1.55)–(T1.56), and his Hodge theorem and (2.53), $\dim \operatorname{Ker} \Delta_p = b_p$, is our key idea above, illustrated in Example T1.35. He also notes that Δ depends on the metric but b_p does not: in Figure T1.20 rescaling the circle moves every nonzero dot but never the zero modes.

Recap

- $\delta = (-1)^{n(p+1)+1+s} \star \mathrm{d} \star$ is the adjoint of d for $\langle \alpha, \beta \rangle = \int \alpha \wedge \star \beta$ (Stokes plus Leibniz).
- On 1-forms $\delta = -\mathrm{div}$; $\delta^2 = 0$.
- $\Delta = \mathrm{d}\delta + \delta\mathrm{d}$; on functions $\Delta f = -\frac{1}{\sqrt{g}} \partial_\mu (\sqrt{g}\, g^{\mu\nu} \partial_\nu f)$, e.g. $-\left(\frac{1}{r}\partial_r r \partial_r + \frac{1}{r^2}\partial_\phi^2\right)$ in polar coordinates.
- $\Delta \geq 0$; harmonic $\iff$ closed and co-closed; Hodge: $b_p = \dim \operatorname{Ker} \Delta_p$.
- $\sigma(\mathrm{d}) = \mathrm{i}k\wedge$, $\sigma(\delta) = -\mathrm{i}\iota_k$, $\sigma(\Delta) = |k|^2$: Δ is elliptic, and $\mathrm{d}+\delta$ is its square root.

T1.7 Maxwell's equations in the language of forms

Here is the question. Maxwell's theory is four vector equations, eight component equations, for two vector fields, with a continuity equation and a gauge freedom tacked on. Why does it have *exactly* this structure? Why are two of the equations "free" (no sources ever appear in them) and two sourced? Why does the charge have to be conserved? Why can we always introduce potentials? And why is $E^2 - B^2$ Lorentz invariant while $E^2 + B^2$ is not?

Differential forms answer all of these at once. In four-dimensional space-time Maxwell's equations become

$$\mathrm{d}F = 0, \qquad \mathrm{d} \star F = \star J.$$

The first is a statement without a metric and without sources: it is pure calculus. The second is where the metric (through $\star$) and the sources enter. Conservation of charge is $\mathrm{d}^2 = 0$. Potentials exist because of the Poincaré lemma. Gauge freedom is $\mathrm{d}^2 = 0$ again. This section is the best possible exercise in everything we have built so far, and it is also our first gauge theory: Yang–Mills theory (Chapter T3, Chapter 8) is this, with matrices.

Conventions (as stated in Chapter T1): Minkowski metric $\eta = \mathrm{diag}(-1,+1,+1,+1)$ in coordinates (t,x,y,z), orientation vol $= \mathrm{d}t \wedge \mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z$, $c = 1$, and Heaviside–Lorentz units (no 4π in the field equations, $\epsilon_0 = 1$). Dray uses Gaussian units and the orientation $\mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z \wedge \mathrm{d}t$ [Dray §15.9]; his signs in intermediate steps therefore differ from ours, but the final equations agree.

T1.7.1 The starting point: vector calculus

In Heaviside–Lorentz units with $c = 1$,

$$\begin{aligned} \nabla\cdot\vec{E} &= \rho, & \nabla\times\vec{E} + \partial_t\vec{B} &= 0, \\ \nabla\cdot\vec{B} &= 0, & \nabla\times\vec{B} - \partial_t\vec{E} &= \vec{J}. \end{aligned} \tag{T1.60}$$

Two consequences are usually derived by hand. Taking the divergence of the Ampère–Maxwell law and using Gauss's law,

$$0 \overset{(1)}{=} \nabla\cdot(\nabla\times\vec{B}) = \nabla\cdot\vec{J} + \partial_t(\nabla\cdot\vec{E}) \overset{(2)}{=} \nabla\cdot\vec{J} + \partial_t\rho,$$

(1) div curl = 0, i.e. $\mathrm{d}^2 = 0$ (Section T1.5).
(2) Gauss's law.

which is the continuity equation. And the ansatz $\vec{B} = \nabla\times\vec{A}$, $\vec{E} = -\nabla\Phi - \partial_t\vec{A}$ solves the two source-free equations identically, again by div curl = 0 and curl grad = 0. Keep an eye on how often "$\mathrm{d}^2 = 0$" is doing the work.

T1.7.2 Which fields are which forms?

Before going to four dimensions, ask the question of Section T1.3: what does each field want to be integrated over?

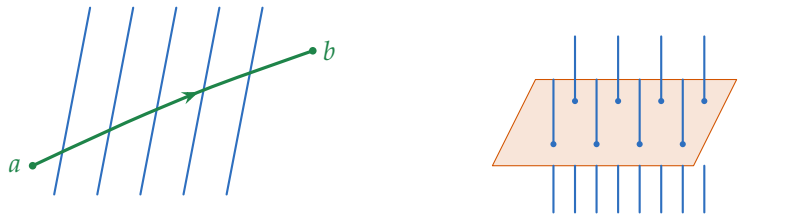
- $\vec{E}$ is integrated along curves: the EMF around a loop is $\oint \vec{E}\cdot\mathrm{d}\vec{r}$, the potential difference is $-\int_a^b \vec{E}\cdot\mathrm{d}\vec{r}$. So $\vec{E}$ wants to be the *1-form* $E = E_x\mathrm{d}x + E_y\mathrm{d}y + E_z\mathrm{d}z$.
- $\vec{B}$ is integrated over surfaces: the magnetic flux is $\int \vec{B}\cdot\mathrm{d}\vec{A}$. So $\vec{B}$ wants to be the *2-form* $\mathcal{B} = B_x\,\mathrm{d}y\wedge\mathrm{d}z + B_y\,\mathrm{d}z\wedge\mathrm{d}x + B_z\,\mathrm{d}x\wedge\mathrm{d}y = \star_3 B$, where $\star_3$ is the three-dimensional Hodge star and $B = B_x\mathrm{d}x + \dots$ is the 1-form made from $\vec{B}$.
- ρ is integrated over volumes: the charge is $\int \rho\,\mathrm{d}^3x$. So ρ wants to be the *3-form* $\rho\,\mathrm{d}x\wedge\mathrm{d}y\wedge\mathrm{d}z$.
- $\vec{J}$ is integrated over surfaces: the current through a surface. So it is the 2-form $\star_3 J$.

This is the "pseudovector" remark of Section T1.3 made physical: $\vec{B}$ is a pseudovector because it is really a 2-form, and it is only the three-dimensional Hodge star, which needs an orientation, that turns it into an arrow.

With this assignment the four equations read, in three dimensions (Dray's "Maxwell II" [Dray §15.8]),

$$\begin{aligned} \mathrm{d}_3\star_3 E &= \rho\,\mathrm{d}^3x, & \mathrm{d}_3 E + \partial_t\mathcal{B} &= 0, \\ \mathrm{d}_3\mathcal{B} &= 0, & \mathrm{d}_3 B - \partial_t\star_3 E &= \star_3 J, \end{aligned} \tag{T1.61}$$

where d_3 is the exterior derivative in x, y, z only and $\mathrm{d}^3x = \mathrm{d}x\wedge\mathrm{d}y\wedge\mathrm{d}z$. Each follows from the dictionary of Section T1.5: $\mathrm{d}_3\star_3 E = (\nabla\cdot\vec{E})\mathrm{d}^3x$, $\mathrm{d}_3 E = \star_3(\nabla\times\vec{E})$, $\mathrm{d}_3\star_3 B = (\nabla\cdot\vec{B})\mathrm{d}^3x$ and $\mathrm{d}_3 B = \star_3(\nabla\times\vec{B})$.



E = 1-form:
stack of equipotentials;
$\int_a^b E$ counts sheets crossed

$\mathcal{B} = \star_3 B$ = 2-form:
flux tubes;
$\int_S \mathcal{B}$ counts tubes piercing S

Figure T1.21: Why $\vec{E}$ is a 1-form and $\vec{B}$ is a 2-form. On the left, the electric field is drawn as a stack of equipotential sheets (Dray's picture of a 1-form, Figure T1.7); the potential difference along the green path is the number of sheets it crosses. On the right, the magnetic field is drawn as a bundle of flux tubes (the picture of a 2-form); the flux through the orange surface is the number of tubes that pierce it. No metric is needed to count sheets or tubes.

Key idea: Where the metric enters

Look at the right-hand column of (T1.61), with $\mathcal{B}$ treated as the basic magnetic variable. The pair $\mathrm{d}_3\mathcal{B} = 0$, $\mathrm{d}_3 E + \partial_t \mathcal{B} = 0$ contains no Hodge star: it makes sense without any metric at all. The other pair needs $\star_3$ (to make the flux $\star_3 E$ from the 1-form E, and the 1-form B from the flux $\mathcal{B}$). In a material these two stars are replaced by the constitutive relations $\vec{D} = \epsilon\vec{E}$ and $\vec{H} = \vec{B}/\mu$. In four dimensions this separation becomes the cleanest statement in the subject: $\mathrm{d}F = 0$ is metric-free; $\mathrm{d} \star F = \star J$ is where the metric lives.

T1.7.3 Splitting space-time forms into space and time

To pass to four dimensions we need a bookkeeping device (Dray's "barred" quantities [Dray §15.9]). Every form on space-time can be written uniquely as

$$\omega = \bar{\omega} + \bar{\eta} \wedge \mathrm{d}t, \tag{T1.62}$$

where $\bar{\omega}$ and $\bar{\eta}$ contain no $\mathrm{d}t$ (they are "spatial" forms whose coefficients may depend on t). On spatial forms we use the spatial exterior derivative d_3 and the spatial star $\star_3$. The rules we need are these.

- *Derivative of a spatial form:* $\mathrm{d}\bar{\omega} = \mathrm{d}_3\bar{\omega} + \mathrm{d}t \wedge \partial_t \bar{\omega}$. (Split $\mathrm{d} = \mathrm{d}x^i \wedge \partial_i + \mathrm{d}t \wedge \partial_t$.)
- *Moving* $\mathrm{d}t$ *through a spatial* p*-form:* $\mathrm{d}t \wedge \bar{\omega} = (-1)^p\, \bar{\omega} \wedge \mathrm{d}t$ (T1.23).
- *The 4D star on spatial forms and on* $(\cdot) \wedge \mathrm{d}t$, read off from the Minkowski table of Example T1.22:

$$\begin{aligned} \star\mathrm{d}t &= -\mathrm{d}^3x, & \star\mathrm{d}^3x &= -\mathrm{d}t, \\ \star\bar{\alpha} &= -(\star_3\bar{\alpha}) \wedge \mathrm{d}t, & \star(\bar{\alpha} \wedge \mathrm{d}t) &= \star_3\bar{\alpha} \quad (\bar{\alpha} \text{ a spatial 1-form}), \\ \star\bar{\beta} &= -(\star_3\bar{\beta}) \wedge \mathrm{d}t & &= (\bar{\beta} \text{ a spatial 2-form}). \end{aligned} \tag{T1.63}$$

Let us check the two rules we use most, on basis elements. From the table, $\star(\mathrm{d}t \wedge \mathrm{d}x) = -\mathrm{d}y \wedge \mathrm{d}z$, so $\star(\mathrm{d}x \wedge \mathrm{d}t) = +\mathrm{d}y \wedge \mathrm{d}z = \star_3\mathrm{d}x$: that is the rule $\star(\bar{\alpha} \wedge \mathrm{d}t) = \star_3\bar{\alpha}$. And $\star(\mathrm{d}y \wedge \mathrm{d}z) = \mathrm{d}t \wedge \mathrm{d}x = -\mathrm{d}x \wedge \mathrm{d}t = -(\star_3(\mathrm{d}y \wedge \mathrm{d}z)) \wedge \mathrm{d}t$, since $\star_3(\mathrm{d}y \wedge \mathrm{d}z) = \mathrm{d}x$: that is the rule for $\bar{\beta}$. Finally $\star\mathrm{d}x = -\mathrm{d}t \wedge \mathrm{d}y \wedge \mathrm{d}z = -(\mathrm{d}y \wedge \mathrm{d}z) \wedge \mathrm{d}t$, the rule for $\bar{\alpha}$.

T1.7.4 The field strength 2-form

We have a 1-form E (3 components) and a 2-form $\mathcal{B}$ (3 components). A 2-form in four dimensions has $\binom{4}{2} = 6$ components. Combine them:

Working definition T1.19 (electromagnetic field strength)

The **field strength** (or Faraday) 2-form is

$$\begin{aligned} F &= E \wedge \mathrm{d}t + \star_3 B \\ &= E_x\,\mathrm{d}x \wedge \mathrm{d}t + E_y\,\mathrm{d}y \wedge \mathrm{d}t + E_z\,\mathrm{d}z \wedge \mathrm{d}t + B_x\,\mathrm{d}y \wedge \mathrm{d}z + B_y\,\mathrm{d}z \wedge \mathrm{d}x + B_z\,\mathrm{d}x \wedge \mathrm{d}y. \end{aligned} \tag{T1.64}$$

Its components $F = \frac{1}{2}F_{\mu\nu}\mathrm{d}x^\mu \wedge \mathrm{d}x^\nu$ [the $\frac{1}{2}$ is there because each pair $\mu \neq \nu$ appears twice in the sum] are $F_{it} = E_i$ and $F_{ij} = \varepsilon_{ijk}B_k$ [$\varepsilon_{ijk} = +1$ if (ijk) is an even permutation of (xyz), -1 if odd, 0 if an index repeats]. As an antisymmetric matrix, rows and columns ordered (t, x, y, z),

$$(F_{\mu\nu}) = \begin{pmatrix} 0 & -E_x & -E_y & -E_z \\ E_x & 0 & B_z & -B_y \\ E_y & -B_z & 0 & B_x \\ E_z & B_y & -B_x & 0 \end{pmatrix}. \tag{T1.65}$$

How to read it: $F(v, w) = F_{\mu\nu}v^\mu w^\nu$. For a charge with 4-velocity $u = \partial_t + \vec{v}\cdot\nabla$, $F(\cdot\,, u)$ is the Lorentz force 1-form (Example T1.37).

Checking a few entries: F_{xt} is the coefficient of $\mathrm{d}x \wedge \mathrm{d}t$, which is E_x, so $F_{tx} = -E_x$ (first row). The coefficient of $\mathrm{d}z \wedge \mathrm{d}x$ is B_y, so $F_{zx} = B_y$ and $F_{xz} = -B_y$ (second row, last entry). This is the 4×4 matrix every physicist has seen; here it is simply the representation of a 2-form in the basis $\mathrm{d}x^\mu \wedge \mathrm{d}x^\nu$ (Section T1.3).

Worked example T1.37: The Lorentz force from the matrix of F

Insert the velocity $u = \partial_t + v^i\partial_i$ into the second slot of F. The result is a 1-form, $f_\mu = F_{\mu\nu}u^\nu$:

$$f_t \overset{(1)}{=} F_{tt}\cdot 1 + F_{ti}v^i = -\vec{E}\cdot\vec{v},$$
$$f_x \overset{(2)}{=} F_{xt}\cdot 1 + F_{xy}v^y + F_{xz}v^z = E_x + B_z v_y - B_y v_z \overset{(3)}{=} \left(\vec{E} + \vec{v}\times\vec{B}\right)_x,$$

(1) First row of (T1.65).
(2) Second row of (T1.65).
(3) $(\vec{v}\times\vec{B})_x = v_y B_z - v_z B_y$.

and similarly for y, z. So $q\,F(\cdot\,, u) = q(\vec{E} + \vec{v}\times\vec{B})\cdot\mathrm{d}\vec{r} - q\,\vec{E}\cdot\vec{v}\,\mathrm{d}t$: the spatial part is the Lorentz force and the time part is minus the power delivered to the charge (raising the index with $\eta^{tt} = -1$ turns it into $+q\vec{E}\cdot\vec{v}$, the rate of change of energy). The whole of charged-particle dynamics is $m\,\mathrm{d}u^\mu/\mathrm{d}\tau = q\,F^\mu{}_\nu u^\nu$.

T1.7.5 The homogeneous equations: $\mathrm{d}F = 0$

Now differentiate (T1.64).

$$\mathrm{d}F \overset{(1)}{=} \mathrm{d}E \wedge \mathrm{d}t + \mathrm{d}(\star_3 B)$$

$$\overset{(2)}{=} \big(\mathrm{d}_3 E + \mathrm{d}t \wedge \partial_t E\big) \wedge \mathrm{d}t + \mathrm{d}_3 \star_3 B + \mathrm{d}t \wedge \partial_t \star_3 B$$
$$\overset{(3)}{=} \mathrm{d}_3 E \wedge \mathrm{d}t + \mathrm{d}_3 \star_3 B + \partial_t(\star_3 B) \wedge \mathrm{d}t$$
$$\overset{(4)}{=} (\nabla \cdot \vec{B})\, \mathrm{d}^3x + \star_3\big(\nabla \times \vec{E} + \partial_t \vec{B}\big) \wedge \mathrm{d}t. \tag{T1.66}$$

(1) Leibniz rule (T1.46) and $\mathrm{d}(\mathrm{d}t) = 0$.
(2) Split d on spatial forms into its space and time parts.
(3) $\mathrm{d}t \wedge \partial_t E \wedge \mathrm{d}t = 0$ because $\mathrm{d}t \wedge \mathrm{d}t = 0$; and $\mathrm{d}t \wedge$ (2-form) = (2-form) $\wedge\, \mathrm{d}t$ by (T1.23).
(4) The 3D dictionary: $\mathrm{d}_3 \star_3 B = (\nabla \cdot \vec{B})\mathrm{d}^3x$ and $\mathrm{d}_3 E = \star_3(\nabla \times \vec{E})$ (Section T1.5).

The two terms are linearly independent (one has no $\mathrm{d}t$, the other has exactly one), so

$$\boxed{\mathrm{d}F = 0 \iff \nabla \cdot \vec{B} = 0 \ \text{ and } \ \nabla \times \vec{E} + \partial_t \vec{B} = 0.} \tag{T1.67}$$

Four component equations in one line. In components, $\mathrm{d}F = \frac{1}{2}\partial_\lambda F_{\mu\nu}\mathrm{d}x^\lambda \wedge \mathrm{d}x^\mu \wedge \mathrm{d}x^\nu$, and $\mathrm{d}F = 0$ is the **Bianchi identity**

$$\partial_\lambda F_{\mu\nu} + \partial_\mu F_{\nu\lambda} + \partial_\nu F_{\lambda\mu} = 0. \tag{T1.68}$$

(For $(\lambda, \mu, \nu) = (x, y, z)$ it reads $\partial_x B_x + \partial_y B_y + \partial_z B_z = 0$; for (t, y, z) it reads $\partial_t B_x + \partial_y F_{zt} + \partial_z F_{ty} = \partial_t B_x + \partial_y E_z - \partial_z E_y = 0$, the x-component of Faraday's law.)

Physical picture: Faraday's law is Stokes' theorem

Integrate $\mathrm{d}F = 0$ over a 3-dimensional "box" in space-time: a surface S in space times a time interval $[t_1, t_2]$. By Stokes (Chapter T2) the integral of F over the boundary of the box is zero. The boundary consists of S at t_2 and at t_1 (where F restricts to $\mathcal{B}$, giving the fluxes $\Phi_B(t_2)$, $\Phi_B(t_1)$) and the "sides" $\partial S \times [t_1, t_2]$ (where F restricts to $E \wedge \mathrm{d}t$, giving $\int \oint_{\partial S} \vec{E} \cdot \mathrm{d}\vec{r}\, \mathrm{d}t$). Keeping track of orientations, $\Phi_B(t_2) - \Phi_B(t_1) = -\int_{t_1}^{t_2} \oint_{\partial S} \vec{E} \cdot \mathrm{d}\vec{r}\, \mathrm{d}t$: the integrated form of Faraday's law. "Magnetic flux changes only by EMF around the edge" is the space-time statement "F has no net flux out of any closed 3-surface".

T1.7.6 The dual field and the inhomogeneous equations

Apply the Minkowski star to F using (T1.63):

$$\star F \overset{(1)}{=} \star(E \wedge \mathrm{d}t) + \star(\star_3 B) \overset{(2)}{=} \star_3 E - \big(\star_3 \star_3 B\big) \wedge \mathrm{d}t \overset{(3)}{=} \star_3 E - B \wedge \mathrm{d}t. \tag{T1.69}$$

(1) Linearity of $\star$.
(2) The rules for $\star(\bar{\alpha} \wedge \mathrm{d}t)$ and $\star\bar{\beta}$ in (T1.63).
(3) In Euclidean $\mathbb{R}^3$, $\star_3\star_3 = +1$ (T1.35).

Compare with $F = E \wedge \mathrm{d}t + \star_3 B$: the star has made the replacement

$$(\vec{E}, \vec{B}) \xrightarrow{\star} (-\vec{B}, \vec{E}), \tag{T1.70}$$

the **electric–magnetic duality** rotation by 90°. Doing it twice gives $(\vec{E}, \vec{B}) \to (-\vec{E}, -\vec{B})$, consistent with $\star\star = -1$ on 2-forms in Lorentzian signature.

Now the current. The 4-current is the vector $J^\mu = (\rho, \vec{J})$; lowering the index with η gives the 1-form

$$J = J_\mu \mathrm{d}x^\mu = -\rho\, \mathrm{d}t + J_x \mathrm{d}x + J_y \mathrm{d}y + J_z \mathrm{d}z, \qquad \star J \overset{(1)}{=} \rho\, \mathrm{d}^3x - (\star_3 \vec{J}) \wedge \mathrm{d}t, \tag{T1.71}$$

(1) $\star \mathrm{d}t = -\mathrm{d}^3x$ and $\star\bar{\alpha} = -(\star_3\bar{\alpha}) \wedge \mathrm{d}t$ from (T1.63), with $\bar{J} = J_i \mathrm{d}x^i$.

The 3-form $\star J$ is the charge–current in its natural home: its integral over a spatial volume at fixed t (where $\mathrm{d}t = 0$) is the charge inside. Next differentiate $\star F$:

$$\begin{aligned} \mathrm{d} \star F &\overset{(1)}{=} \mathrm{d}_3 \star_3 E + \mathrm{d}t \wedge \partial_t \star_3 E - \mathrm{d}_3 B \wedge \mathrm{d}t \\ &\overset{(2)}{=} (\nabla \cdot \vec{E})\, \mathrm{d}^3x + \star_3 \big(\partial_t \vec{E} - \nabla \times \vec{B}\big) \wedge \mathrm{d}t. \end{aligned} \tag{T1.72}$$

(1) Same steps as in (T1.66): split d, drop $\mathrm{d}t \wedge \mathrm{d}t$, move $\mathrm{d}t$ past a 2-form.
(2) 3D dictionary: divergence of $\vec{E}$ and curl of $\vec{B}$.

Setting (T1.72) equal to (T1.71) and matching the parts with and without $\mathrm{d}t$:

$$\boxed{\mathrm{d} \star F = \star J \iff \nabla \cdot \vec{E} = \rho \ \text{ and } \ \nabla \times \vec{B} - \partial_t \vec{E} = \vec{J}.} \tag{T1.73}$$

Key idea: Maxwell's equations

$$\mathrm{d}F = 0, \qquad \mathrm{d} \star F = \star J. \tag{T1.74}$$

Equivalently, applying $\star$ to the second and using the codifferential of Section T1.6 ($\delta = +\star\mathrm{d}\star$ on every degree in Minkowski space, and $\star\star = +1$ on 1-forms there), $\delta F = J$. In components, $\partial_\mu F^{\nu\mu} = J^\nu$ (check $\nu = t$: $\partial_i F^{ti} = \partial_i E_i$, since raising both indices of $F_{ti} = -E_i$ gives $F^{ti} = +E_i$).

Watch out: Signs of $\partial_\mu F^{\mu\nu}$

Textbooks write $\partial_\mu F^{\mu\nu} = \pm J^\nu$ with either sign, depending on the metric signature, the definition $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ versus its negative, and whether $E_i = F_{i0}$ or F_{0i}. The form statement $\mathrm{d} \star F = \star J$ also depends on the orientation (which fixes the sign of $\star$): with Dray's orientation $\mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z \wedge \mathrm{d}t = -\mathrm{vol}$, both sides flip sign and the equation is unchanged. Always check the $\nu = t$ component against Gauss's law $\nabla \cdot \vec{E} = +\rho$.

T1.7.7 Charge conservation is $\mathrm{d}^2 = 0$

$$\begin{aligned} 0 &\overset{(1)}{=} \mathrm{dd} \star F \overset{(2)}{=} \mathrm{d} \star J \overset{(3)}{=} \mathrm{d}\big(\rho\, \mathrm{d}^3x\big) - \mathrm{d}\big((\star_3 \bar{J}) \wedge \mathrm{d}t\big) \\ &\overset{(4)}{=} \partial_t \rho\, \mathrm{d}t \wedge \mathrm{d}^3x - (\nabla \cdot \vec{J})\, \mathrm{d}^3x \wedge \mathrm{d}t \overset{(5)}{=} \big(\partial_t \rho + \nabla \cdot \vec{J}\big)\, \mathrm{vol}. \end{aligned} \tag{T1.75}$$

(1) $\mathrm{d}^2 = 0$ (T1.45).
(2) Maxwell (T1.74).
(3) (T1.71).
(4) In $\mathrm{d}(\rho\, \mathrm{d}^3x)$ only $\partial_t \rho\, \mathrm{d}t$ survives (the spatial derivatives wedge to zero against d^3x); in the second term only the spatial divergence survives.
(5) $\mathrm{d}^3x \wedge \mathrm{d}t = -\mathrm{d}t \wedge \mathrm{d}^3x = -\mathrm{vol}$ (moving $\mathrm{d}t$ through a 3-form), and $\mathrm{d}t \wedge \mathrm{d}^3x = \mathrm{vol}$.

So the inhomogeneous Maxwell equation is *consistent only if* the current is conserved, $\mathrm{d} \star J = 0$. The sources cannot be chosen freely: the operator d on the left has a kernel's worth of built-in constraint, because $\mathrm{d}^2 = 0$. This is the prototype of every "anomaly" argument in Chapter 10: if a current coupled to a gauge field fails to be conserved, the gauge theory is inconsistent.

T1.7.8 Potentials and gauge freedom

The equation $\mathrm{d}F = 0$ says F is closed. On $\mathbb{R}^4$ (star-shaped) the Poincaré lemma (Section T1.5) guarantees a 1-form A with

$$F = \mathrm{d}A, \qquad A = A_\mu \mathrm{d}x^\mu = -\Phi\,\mathrm{d}t + \bar{A}, \quad \bar{A} = A_x\mathrm{d}x + A_y\mathrm{d}y + A_z\mathrm{d}z. \tag{T1.76}$$

Let us see that this is the familiar potential:

$$\begin{aligned}\mathrm{d}A &\overset{(1)}{=} -\mathrm{d}\Phi\wedge\mathrm{d}t + \mathrm{d}_3\bar{A} + \mathrm{d}t\wedge\partial_t\bar{A} \overset{(2)}{=} -\mathrm{d}_3\Phi\wedge\mathrm{d}t + \mathrm{d}_3\bar{A} - \partial_t\bar{A}\wedge\mathrm{d}t\\ &\overset{(3)}{=} \big(-\nabla\Phi - \partial_t\vec{A}\big)\cdot\mathrm{d}\vec{r}\wedge\mathrm{d}t + \star_3\big((\nabla\times\vec{A})\cdot\mathrm{d}\vec{r}\big).\end{aligned}$$

(1) Leibniz rule and the split of d on the spatial form $\bar{A}$.
(2) $\mathrm{d}\Phi\wedge\mathrm{d}t = \mathrm{d}_3\Phi\wedge\mathrm{d}t$ (the $\partial_t\Phi\,\mathrm{d}t$ part dies against $\mathrm{d}t$); $\mathrm{d}t\wedge\partial_t\bar{A} = -\partial_t\bar{A}\wedge\mathrm{d}t$ (moving $\mathrm{d}t$ past a 1-form).
(3) 3D dictionary: d_3 of a function is the gradient, d_3 of a 1-form is the flux form of the curl.

Comparing with $F = E\wedge\mathrm{d}t + \star_3 B$: $\vec{E} = -\nabla\Phi - \partial_t\vec{A}$ and $\vec{B} = \nabla\times\vec{A}$. In components, $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ (T1.41).

Working definition T1.20 (gauge transformation)

For any function λ, $A' = A + \mathrm{d}\lambda$ gives the same field strength, $\mathrm{d}A' = \mathrm{d}A + \mathrm{d}\mathrm{d}\lambda = \mathrm{d}A$. In components, $\Phi' = \Phi - \partial_t\lambda$ and $\vec{A}' = \vec{A} + \nabla\lambda$ (the $\mathrm{d}t$ coefficient of $\mathrm{d}\lambda$ is $\partial_t\lambda$, and it enters A as $-\Phi$). The potential is a *representative*; the physics is in F, or more precisely in A modulo exact forms [A and $A + \mathrm{d}\lambda$ count as the same potential, for every globally defined function λ].

This is a first taste of cohomology: the potentials giving a fixed F form an "affine copy" of the closed 1-forms, and on $\mathbb{R}^4$ all closed 1-forms are exact, so gauge transformations exhaust the ambiguity. On a space with holes (a solenoid removed, a torus) there are closed 1-forms that are not exact, and two potentials with the same F can be physically different: this is the Aharonov–Bohm effect (Chapter T2, Chapter T3).

Worked example T1.38: Lorenz gauge and the wave equation

Insert $F = \mathrm{d}A$ into $\delta F = J$. Since $\delta\mathrm{d}A = (\Delta - \mathrm{d}\delta)A$,

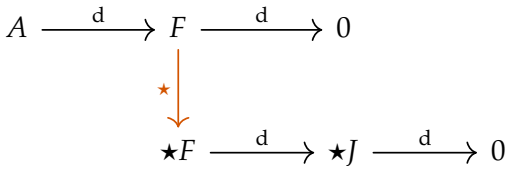
$$\Delta A - \mathrm{d}(\delta A) = J.$$

Use the gauge freedom to impose the **Lorenz gauge** $\delta A = 0$. By (T1.52), $\delta A = -\partial_\mu A^\mu$, and $A^\mu = (\Phi, \vec{A})$, so this is $\partial_t\Phi + \nabla\cdot\vec{A} = 0$. (We can always reach it: if $\delta A \neq 0$, choose λ with $\delta\mathrm{d}\lambda = \Delta\lambda = -\delta A$.) Then Maxwell becomes $\Delta A = J$. On functions in Minkowski space,

$$\Delta f \overset{(1)}{=} \delta\mathrm{d}f \overset{(2)}{=} -\partial_\mu(\eta^{\mu\nu}\partial_\nu f) \overset{(3)}{=} \partial_t^2 f - \nabla^2 f,$$

(1) $\delta f = 0$ on functions.
(2) (T1.54) with $\sqrt{|g|} = 1$.
(3) $\eta^{tt} = -1$, $\eta^{ii} = +1$.

and in Cartesian coordinates Δ acts componentwise on A_μ. So each component obeys the sourced wave equation $(\partial_t^2 - \nabla^2)A_\mu = J_\mu$. In Minkowski signature the "Laplacian" Δ is the wave operator (minus the d'Alembertian): not elliptic. This is exactly why Nash, and instanton physics, Wick-rotate to Euclidean signature, where Δ becomes $-\nabla^2_{\mathbb{R}^4}$, which is elliptic (Section T1.6, Chapter 2).



top row	no metric needed	$\mathrm{d}F = 0$ (Bianchi), $F = \mathrm{d}A$, gauge $A \to A + \mathrm{d}\lambda$
vertical arrow	metric and orientation	$(\vec{E}, \vec{B}) \mapsto (-\vec{B}, \vec{E})$
bottom row	no metric needed again	$\mathrm{d}\star F = \star J$ forces $\mathrm{d}\star J = 0$

Figure T1.22: The logical structure of Maxwell theory. Every horizontal arrow is d, and every composite of two horizontal arrows is zero: this gives the existence of potentials, gauge invariance and charge conservation. The only place the metric enters is the vertical arrow $\star$. Yang–Mills theory has the same diagram with d replaced by a covariant derivative (Chapter T3).

T1.7.9 The action and the two invariants

The two Lorentz-invariant scalars of electromagnetism are the two natural 4-forms one can build from F.

The first invariant. By the defining property of $\star$, $F \wedge \star F = g(F,F)\,\mathrm{vol}$, and $g(F,F) = |\vec{B}|^2 - |\vec{E}|^2$ was computed in Section T1.4 from the determinant formula. It is a good exercise in sign bookkeeping to get the same answer directly from the space–time split.

Worked example T1.39: Computing $F \wedge \star F$ carefully

Only terms containing all four of $\mathrm{d}t, \mathrm{d}x, \mathrm{d}y, \mathrm{d}z$ survive, so only the cross terms between the $\mathrm{d}t$ part of one factor and the spatial part of the other contribute:

$$\begin{aligned} F \wedge \star F &\overset{(1)}{=} (E \wedge \mathrm{d}t) \wedge \star_3 E + \star_3 B \wedge (-B \wedge \mathrm{d}t) \\ &\overset{(2)}{=} E \wedge \star_3 E \wedge \mathrm{d}t - \star_3 B \wedge B \wedge \mathrm{d}t \\ &\overset{(3)}{=} |\vec{E}|^2\, \mathrm{d}^3x \wedge \mathrm{d}t - |\vec{B}|^2\, \mathrm{d}^3x \wedge \mathrm{d}t \overset{(4)}{=} \left(|\vec{B}|^2 - |\vec{E}|^2\right) \mathrm{vol}. \end{aligned}$$

(1) Expand; $(E \wedge \mathrm{d}t) \wedge (B \wedge \mathrm{d}t) = 0$ (two $\mathrm{d}t$'s) and $\star_3 B \wedge \star_3 E = 0$ (a spatial 4-form in 3 dimensions).
(2) $\mathrm{d}t \wedge \star_3 E = \star_3 E \wedge \mathrm{d}t$ ($\mathrm{d}t$ past a 2-form, sign +).
(3) $E \wedge \star_3 E = |\vec{E}|^2 \mathrm{d}^3x$, and $\star_3 B \wedge B = B \wedge \star_3 B = |\vec{B}|^2 \mathrm{d}^3x$ (a 2-form and a 1-form commute).
(4) $\mathrm{d}^3x \wedge \mathrm{d}t = -\mathrm{d}t \wedge \mathrm{d}^3x = -\mathrm{vol}$ ($\mathrm{d}t$ past a 3-form costs $(-1)^3$).

This agrees with $g(F,F) = B^2 - E^2$ from Section T1.4. So the **Maxwell action** with the conventional sign is

$$S[A] = \int \left(-\tfrac{1}{2}\, F \wedge \star F + A \wedge \star J\right) = \int \left(\tfrac{1}{2}(|\vec{E}|^2 - |\vec{B}|^2) - \rho\Phi + \vec{J}\cdot\vec{A}\right) \mathrm{d}t\, \mathrm{d}^3x, \qquad \text{(T1.77)}$$

where $A \wedge \star J = g(A,J)\mathrm{vol} = A_\mu J^\mu \mathrm{vol} = (-\Phi\rho + \vec{A}\cdot\vec{J})\mathrm{vol}$.

Watch out: Moving $\mathrm{d}t$ around

The most common mistake in these computations is a wrong sign from moving $\mathrm{d}t$. The rule is (T1.23): $\mathrm{d}t$ moves past a p-form with sign $(-1)^p$. Past a 2-form it moves freely; past a 1-form or a 3-form it picks up a minus sign. Do one transposition at a time.

Varying the action. Let $A \to A + a$ with a of compact support. Using the symmetry $\alpha \wedge \star\beta = \beta \wedge \star\alpha$,

$$S[A+a] - S[A] \overset{(1)}{=} \int \left(-\mathrm{d}a \wedge \star F + a \wedge \star J\right) + O(a^2)$$
$$\overset{(2)}{=} \int \left(-\mathrm{d}(a \wedge \star F) - a \wedge \mathrm{d}\star F + a \wedge \star J\right) + O(a^2) \overset{(3)}{=} \int a \wedge \left(\star J - \mathrm{d}\star F\right) + O(a^2).$$

(1) $F \to F + \mathrm{d}a$; the cross terms of $-\frac{1}{2}(F + \mathrm{d}a) \wedge \star(F + \mathrm{d}a)$ are equal by the symmetry, giving $-\mathrm{d}a \wedge \star F$.
(2) Leibniz: $\mathrm{d}(a \wedge \star F) = \mathrm{d}a \wedge \star F - a \wedge \mathrm{d}\star F$ (a is a 1-form, sign $(-1)^1$).
(3) Stokes: the total derivative integrates to zero because a vanishes at infinity.

For this to vanish for every a we need $\mathrm{d}\star F = \star J$. The homogeneous equation $\mathrm{d}F = 0$ is not an equation of motion at all: it is automatic once $F = \mathrm{d}A$. And gauge invariance of the coupling term, $\int \mathrm{d}\lambda \wedge \star J = -\int \lambda\, \mathrm{d}\star J$, requires $\mathrm{d}\star J = 0$: again, conservation of the current.

The second invariant. The other 4-form is $F \wedge F$. Only cross terms contribute again:

$$F \wedge F \overset{(1)}{=} 2\,(E \wedge \mathrm{d}t) \wedge \star_3 B \overset{(2)}{=} 2\,E \wedge \star_3 B \wedge \mathrm{d}t \overset{(3)}{=} 2\,(\vec{E}\cdot\vec{B})\,\mathrm{d}^3x \wedge \mathrm{d}t \overset{(4)}{=} -2\,(\vec{E}\cdot\vec{B})\,\mathrm{vol}. \tag{T1.78}$$

(1) $F \wedge F$ with $F = X + Y$, $X = E \wedge \mathrm{d}t$, $Y = \star_3 B$: $X \wedge X = 0$, $Y \wedge Y = 0$ (spatial 4-form), and $X \wedge Y = Y \wedge X$ (2-forms commute).
(2) $\mathrm{d}t$ past the 2-form $\star_3 B$: no sign.
(3) $E \wedge \star_3 B = g_3(E, B)\,\mathrm{d}^3x = (\vec{E}\cdot\vec{B})\,\mathrm{d}^3x$ (definition of $\star_3$).
(4) $\mathrm{d}^3x \wedge \mathrm{d}t = -\mathrm{vol}$.

Cross-check with linear algebra: for a 4×4 antisymmetric matrix, $F\wedge F = 2\,\mathrm{Pf}(F)\,\mathrm{d}t\wedge\mathrm{d}x\wedge\mathrm{d}y\wedge\mathrm{d}z$ with the Pfaffian $\mathrm{Pf} = F_{tx}F_{yz} - F_{ty}F_{xz} + F_{tz}F_{xy}$. From (T1.65), $\mathrm{Pf} = (-E_x)B_x - (-E_y)(-B_y) + (-E_z)B_z = -\vec{E}\cdot\vec{B}$, and $\det(F_{\mu\nu}) = \mathrm{Pf}^2 = (\vec{E}\cdot\vec{B})^2$. The invariant $F \wedge F$ is also a total derivative:

$$F \wedge F \overset{(1)}{=} \mathrm{d}A \wedge \mathrm{d}A \overset{(2)}{=} \mathrm{d}(A \wedge \mathrm{d}A), \tag{T1.79}$$

(1) $F = \mathrm{d}A$.
(2) Leibniz: $\mathrm{d}(A \wedge \mathrm{d}A) = \mathrm{d}A \wedge \mathrm{d}A - A \wedge \mathrm{d}\mathrm{d}A$ and $\mathrm{d}^2 = 0$.

so adding $\theta \int F \wedge F$ to the action changes no equation of motion. Its non-abelian version $\mathrm{tr}\, F \wedge F$ is the instanton number density (Chapter 8), and $A \wedge \mathrm{d}A$ is the abelian Chern–Simons 3-form, the Lagrangian of the topological field theories of Nash's last chapter (Chapter 12).

T1.7.10 Three solutions

Worked example T1.40: The point charge: Gauss's law as a flux of $\star F$

A static charge q at the origin has $\Phi = q/(4\pi r)$ and $\vec{A} = 0$, so $A = -\Phi\,\mathrm{d}t$ and

$$F \overset{(1)}{=} \mathrm{d}A = -\mathrm{d}\Phi \wedge \mathrm{d}t \overset{(2)}{=} \frac{q}{4\pi r^2}\,\mathrm{d}r \wedge \mathrm{d}t, \qquad \star F \overset{(3)}{=} \frac{q}{4\pi r^2}\,\star_3\,\mathrm{d}r \overset{(4)}{=} \frac{q}{4\pi}\,\sin\theta\,\mathrm{d}\theta \wedge \mathrm{d}\phi.$$

(1) (T1.76); $\mathrm{d}\mathrm{d}t = 0$.
(2) $\mathrm{d}(1/r) = -\mathrm{d}r/r^2$. So $E = \frac{q}{4\pi r^2}\mathrm{d}r$, the Coulomb field.
(3) $\star(\bar{\alpha} \wedge \mathrm{d}t) = \star_3\bar{\alpha}$ (T1.63).
(4) In spherical coordinates $\star_3\mathrm{d}r = r^2\sin\theta\,\mathrm{d}\theta \wedge \mathrm{d}\phi$ (Section T1.4): the unit coframe is $\mathrm{d}r$, $r\mathrm{d}\theta$, $r\sin\theta\mathrm{d}\phi$.

Two things to notice. First, $\star F$ does not depend on r at all, and $\mathrm{d}\star F = \frac{q}{4\pi}\cos\theta\,\mathrm{d}\theta\wedge\mathrm{d}\theta\wedge\mathrm{d}\phi = 0$ away from the origin: vacuum Maxwell. Second, its integral over any sphere around the origin is

$$\int_{S^2}\star F = \frac{q}{4\pi}\int_0^{\pi}\int_0^{2\pi}\sin\theta\,\mathrm{d}\phi\,\mathrm{d}\theta = \frac{q}{4\pi}\cdot 2\pi\cdot 2 = q.$$

This is Gauss's law. The 2-form $\star F$ restricted to $\mathbb{R}^3\setminus\{0\}$ is closed but not exact (its integral over the closed surface S^2 is nonzero, and an exact form would integrate to zero by Stokes). It is the 2-dimensional analogue of $\mathrm{d}\phi$ on the punctured plane (Example T1.30), and the charge q is the "period" that measures the hole.

Worked example T1.41: A plane wave is a null field

Take $\vec{E} = E_0\cos(z-t)\,\hat{x}$, $\vec{B} = E_0\cos(z-t)\,\hat{y}$, a wave moving in $+z$. With $u = z-t$, $\mathrm{d}u = \mathrm{d}z - \mathrm{d}t$:

$$F \overset{(1)}{=} E_0\cos u\,(\mathrm{d}x\wedge\mathrm{d}t + \mathrm{d}z\wedge\mathrm{d}x) \overset{(2)}{=} E_0\cos u\,\mathrm{d}x\wedge(\mathrm{d}t-\mathrm{d}z) \overset{(3)}{=} -E_0\cos u\,\mathrm{d}x\wedge\mathrm{d}u,$$
$$\star F \overset{(4)}{=} E_0\cos u\,(\mathrm{d}y\wedge\mathrm{d}z - \mathrm{d}y\wedge\mathrm{d}t) \overset{(5)}{=} E_0\cos u\,\mathrm{d}y\wedge\mathrm{d}u.$$

(1) (T1.64) with $E_x = B_y = E_0\cos u$.
(2) $\mathrm{d}z\wedge\mathrm{d}x = -\mathrm{d}x\wedge\mathrm{d}z$.
(3) $\mathrm{d}t - \mathrm{d}z = -\mathrm{d}u$.
(4) (T1.69): $\star_3 E = E_0\cos u\,\mathrm{d}y\wedge\mathrm{d}z$, and $B\wedge\mathrm{d}t = E_0\cos u\,\mathrm{d}y\wedge\mathrm{d}t$.
(5) $\mathrm{d}z - \mathrm{d}t = \mathrm{d}u$.

Now both Maxwell equations in vacuum are one-liners: $\mathrm{d}F = E_0\sin u\,\mathrm{d}u\wedge\mathrm{d}x\wedge\mathrm{d}u = 0$ and $\mathrm{d}\star F = -E_0\sin u\,\mathrm{d}u\wedge\mathrm{d}y\wedge\mathrm{d}u = 0$, since the only derivative is along $\mathrm{d}u$, which is already present. Both invariants vanish: $F\wedge F \propto \mathrm{d}x\wedge\mathrm{d}u\wedge\mathrm{d}x\wedge\mathrm{d}u = 0$ and $F\wedge\star F \propto \mathrm{d}x\wedge\mathrm{d}u\wedge\mathrm{d}y\wedge\mathrm{d}u = 0$. So $|\vec{E}| = |\vec{B}|$ and $\vec{E}\perp\vec{B}$ in every frame: light is **null**. The whole field is "$\mathrm{d}u$ wedge something transverse", the form-language picture of a transverse wave travelling along the light ray $u = \text{const}$.

Worked example T1.42: The Dirac monopole: a field strength with no global potential

A magnetic monopole of strength g has $\vec{B} = \frac{g}{4\pi r^2}\hat{r}$, i.e. (by the same computation as in Example T1.40)

$$F = \star_3 B = \frac{g}{4\pi}\sin\theta\,\mathrm{d}\theta\wedge\mathrm{d}\phi, \qquad \int_{S^2} F = g. \tag{T1.80}$$

It is closed on $\mathbb{R}^3\setminus\{0\}$ ($\mathrm{d}F$ contains $\mathrm{d}\theta\wedge\mathrm{d}\theta$). Could $F = \mathrm{d}A$ for a global 1-form A on the sphere? No: Stokes would give $\int_{S^2}\mathrm{d}A = \int_{\varnothing} A = 0 \neq g$. But locally there are potentials. Try $A = f(\theta)\,\mathrm{d}\phi$: then $\mathrm{d}A = f'(\theta)\,\mathrm{d}\theta\wedge\mathrm{d}\phi$, so we need $f' = \frac{g}{4\pi}\sin\theta$, $f = -\frac{g}{4\pi}\cos\theta + c$. The constant matters, because $\mathrm{d}\phi$ is singular at the poles (its length is $|\mathrm{d}\phi| = 1/(r\sin\theta)$), so A is smooth at a pole only if f vanishes there:

$$A_N = \frac{g}{4\pi}(1-\cos\theta)\,\mathrm{d}\phi \quad (\text{smooth except at } \theta=\pi),$$
$$A_S = -\frac{g}{4\pi}(1+\cos\theta)\,\mathrm{d}\phi \quad (\text{smooth except at } \theta=0). \tag{T1.81}$$

Near $\theta = 0$, $1-\cos\theta \approx \theta^2/2$, so $|A_N| \approx \frac{g}{4\pi}\frac{\theta^2/2}{r\theta} \to 0$: regular. The singular half-line of each

potential is the **Dirac string**. On the overlap (everything except the poles) the two differ by

$$A_N - A_S \overset{(1)}{=} \frac{g}{4\pi}\big[(1-\cos\theta)+(1+\cos\theta)\big]\mathrm{d}\phi = \frac{g}{2\pi}\,\mathrm{d}\phi \overset{(2)}{=} \mathrm{d}\Big(\frac{g\phi}{2\pi}\Big), \tag{T1.82}$$

(1) Subtract (T1.81).
(2) Locally $\mathrm{d}\phi$ is the differential of the angle.

a gauge transformation with $\lambda = g\phi/2\pi$, which is *not* single-valued. Is that acceptable? A charged wavefunction of charge e transforms as $\psi \to e^{ie\lambda}\psi$ (with $\hbar = 1$), so what must be single-valued on the equator is $e^{ie\lambda} = e^{ieg\phi/2\pi}$. Going once round, $\phi \to \phi + 2\pi$, it changes by e^{ieg}. So

$$eg \in 2\pi\mathbb{Z} \qquad \text{(Dirac quantization)}. \tag{T1.83}$$

The integer $n = eg/2\pi = \frac{e}{2\pi}\int_{S^2} F$ is the first Chern number of a line bundle over S^2, built exactly like the manifold S^2 itself from two charts and a transition function (Section T1.1; bundles in Chapter 1, Chapter T3). For $n = \pm 1$ it is the Hopf bundle.

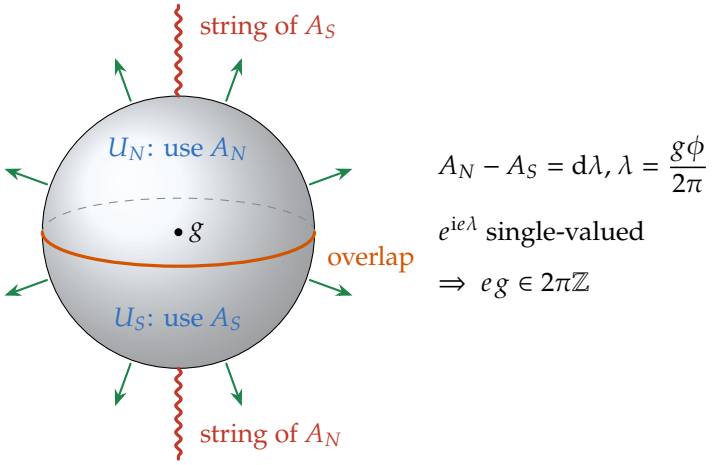


Figure T1.23: The Dirac monopole. The radial field (green) has total flux g through the sphere, so no single potential works everywhere. The potential A_N is good on the northern patch and singular along the south axis (red wavy line, its Dirac string); A_S is the reverse. On the orange overlap they differ by a gauge transformation, and requiring that transformation to be single-valued on charged fields quantizes eg. Compare the two-chart description of S^2 in Figure T1.5.

T1.7.11 Duality, Euclidean signature and self-dual fields

In vacuum ($J = 0$) the equations $\mathrm{d}F = 0$ and $\mathrm{d}\star F = 0$ are symmetric under $F \to \star F$: if F solves both, so does $\star F$, because $\mathrm{d}(\star F) = 0$ and $\mathrm{d}\star(\star F) = -\mathrm{d}F = 0$. This is the duality $(\vec{E},\vec{B}) \to (-\vec{B},\vec{E})$ of (T1.70). Since $\star\star = -1$ on Minkowski 2-forms, its eigenvalues are $\pm\mathrm{i}$: there are no real fields with $\star F = \pm F$.

Now Wick-rotate: $t = -\mathrm{i}\tau$, giving the Euclidean metric $\mathrm{d}\tau^2 + \mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2$ with $\mathrm{vol} = \mathrm{d}x\wedge\mathrm{d}y\wedge\mathrm{d}z\wedge\mathrm{d}\tau$ (the ordering $x^1x^2x^3x^4$ of Example T1.23, with $x^4 = \tau$). Write $F = E\wedge\mathrm{d}\tau + \star_3 B$ as before. The Euclidean table Example T1.23 gives $\star(\mathrm{d}x^1\wedge\mathrm{d}x^4) = \mathrm{d}x^2\wedge\mathrm{d}x^3$ and $\star(\mathrm{d}x^2\wedge\mathrm{d}x^3) = \mathrm{d}x^1\wedge\mathrm{d}x^4$, and cyclically, so now

$$\star F = B\wedge\mathrm{d}\tau + \star_3 E: \qquad (\vec{E},\vec{B}) \xrightarrow{\star} (\vec{B},\vec{E}), \qquad \star F = \pm F \iff \vec{E} = \pm\vec{B}. \tag{T1.84}$$

Real self-dual fields exist, and they solve the vacuum equations automatically: if $F = \mathrm{d}A$ and $\star F = F$, then $\mathrm{d}\star F = \mathrm{d}F = 0$. For a single abelian field on $\mathbb{R}^4$ with finite action the only

such field is $F = 0$, but for SU(2) Yang–Mills there are nontrivial ones, the instantons, and their action is fixed by topology: $\int F \wedge \star F = \pm \int F \wedge F$ (Chapter 8). In Euclidean signature $F \wedge \star F = (|\vec{E}|^2 + |\vec{B}|^2)\mathrm{vol} \geq 0$, so the Euclidean action is positive, which is what makes the Euclidean path integral $\int e^{-S}$ sensible.

Try it: Maxwell in $1+1$ dimensions

In two-dimensional space-time with $\eta = \mathrm{diag}(-1, 1)$ and $\mathrm{vol} = \mathrm{d}t \wedge \mathrm{d}x$, let $F = E\,\mathrm{d}x \wedge \mathrm{d}t$ and $J = -\rho\,\mathrm{d}t + j\,\mathrm{d}x$. Show that $\mathrm{d}F = 0$ is empty, that $\star F = E$ (a function), and that $\mathrm{d} \star F = \star J$ gives $\partial_x E = \rho$ and $\partial_t E = -j$.
Answer: $\mathrm{d}F$ is a 3-form in 2 dimensions, so it vanishes. $F = -E\,\mathrm{vol}$ and $\star\mathrm{vol} = (-1)^s = -1$, so $\star F = E$. From $\mathrm{d}t \wedge \star \mathrm{d}t = -\mathrm{vol}$ we get $\star \mathrm{d}t = -\mathrm{d}x$, and from $\mathrm{d}x \wedge \star \mathrm{d}x = \mathrm{vol}$ we get $\star \mathrm{d}x = -\mathrm{d}t$. So $\star J = \rho\,\mathrm{d}x - j\,\mathrm{d}t$, while $\mathrm{d} \star F = \partial_t E\,\mathrm{d}t + \partial_x E\,\mathrm{d}x$. Matching coefficients gives both equations. This is the setting of the Schwinger model, the simplest anomaly computation (Chapter 10).

Reading Nash (pp. 25, 229 and Ch. 10)

Nash never writes Maxwell's equations, but he uses their structure constantly. On p. 25, (1.74)–(1.75), the Euclidean Yang–Mills action is $S = \|F\|^2 = \int_M \mathrm{tr}(F \wedge \star F)$ and the instanton equation is the self-duality condition $F = \star F$: our (T1.84) with matrices. On p. 229 he proves $\star\star = 1$ on 2-forms in four Euclidean dimensions and splits $F = \frac{1}{2}(F + \star F) + \frac{1}{2}(F - \star F)$, exactly as in Example T1.23. On p. 78 he notes that the Hopf fibration "is also the one that gives the Dirac monopole" (Example T1.42). His anomaly chapter (Chapter 10) rests on the conservation argument (T1.75) and on $F \wedge F$ being a total derivative (T1.79). Remember that Nash's gauge fields are anti-Hermitian: for U(1) his A is i times a real 1-form, $A_{\mathrm{Nash}} = -\mathrm{i}\,e A_{\mathrm{here}}$.

Recap

- $\vec{E}$ is a 1-form (integrated along lines), $\vec{B}$ a 2-form (integrated over surfaces). Together $F = E \wedge \mathrm{d}t + \star_3 B$, with matrix (T1.65).
- $\mathrm{d}F = 0$ is $\nabla \cdot \vec{B} = 0$ and Faraday; it needs no metric.
- $\star F = \star_3 E - B \wedge \mathrm{d}t$, i.e. $(\vec{E}, \vec{B}) \to (-\vec{B}, \vec{E})$; $\mathrm{d} \star F = \star J$ is Gauss and Ampère–Maxwell.
- $\mathrm{d}^2 = 0$ gives charge conservation, potentials $F = \mathrm{d}A$ locally (Poincaré), and gauge freedom $A \to A + \mathrm{d}\lambda$.
- Invariants: $F \wedge \star F = (B^2 - E^2)\mathrm{vol}$ (the action), $F \wedge F = -2\vec{E} \cdot \vec{B}\,\mathrm{vol} = \mathrm{d}(A \wedge \mathrm{d}A)$ (topological).
- The monopole needs two potentials glued by a gauge transformation; consistency gives $eg \in 2\pi\mathbb{Z}$, the first Chern number.
- In Euclidean signature $\star\star = +1$ on 2-forms and self-dual fields ($\vec{E} = \vec{B}$) solve the equations automatically.

T1.8 Pullback and naturality

We have already done the same operation many times without naming it. To get $\mathrm{d}x \wedge \mathrm{d}y = r\,\mathrm{d}r \wedge \mathrm{d}\phi$ we substituted $x = r\cos\phi$, $y = r\sin\phi$ and expanded (Example T1.12). To integrate $\mathrm{d}\phi$ around the unit circle we substituted $x = \cos t$, $y = \sin t$ and got $\mathrm{d}t$ (Example T1.30). To find the metric of the sphere we substituted $X = \sin\theta\cos\phi, \ldots$ into $\mathrm{d}X^2 + \mathrm{d}Y^2 + \mathrm{d}Z^2$ (Example T1.9). Each time a map carried points of one space (θ, ϕ or t) to another ($\mathbb{R}^3$ or $\mathbb{R}^2$), and we used it to drag a form *backwards*, from the target to the source.

That operation is the **pullback**. It is the most important operation on forms after d, for four reasons. (i) Every integral of a form is computed by pulling it back to a parameter domain in $\mathbb{R}^p$. (ii) Restricting a form to a curve or a surface is pulling it back along the inclusion. (iii) It commutes with d ("naturality"), which is why d is coordinate independent and why de Rham cohomology is a topological invariant. (iv) Nash uses it on almost every page of his Chapter 1 to move cohomology classes and bundles between spaces: $f^*\omega$, f^*E, f^*c_i.

T1.8.1 Maps between manifolds, and pushing vectors forward

Let $f : N \to M$ be a smooth map. In coordinates x^μ on N ($\mu = 1, \ldots, n$) and y^a on M ($a = 1, \ldots, m$) it is a list of m functions $y^a = f^a(x^1, \ldots, x^n)$. The dimensions need not agree, and f need not be invertible. Typical examples:

- a curve $\gamma : \mathbb{R} \to M$ ($n = 1$);
- the polar-coordinate map $(r, \phi) \mapsto (r\cos\phi, r\sin\phi)$ ($n = m = 2$);
- the embedding of the sphere $i : S^2 \to \mathbb{R}^3$, $(\theta, \phi) \mapsto (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta)$;
- the n-fold winding map of the circle, $\theta \mapsto n\theta$;
- a projection $\mathbb{R}^3 \to \mathbb{R}^2$, $(x, y, z) \mapsto (x, y)$.

A map carries curves to curves, hence velocities to velocities. If $x(s)$ is a curve in N with velocity $v^\mu = \mathrm{d}x^\mu/\mathrm{d}s$, its image $y^a(s) = f^a(x(s))$ has velocity, by the chain rule,

$$(f_* v)^a = \frac{\partial f^a}{\partial x^\mu}\, v^\mu. \tag{T1.85}$$

Working definition T1.21 (push-forward of a vector)

The **push-forward** (or differential) $f_* v$ of a vector v at $p \in N$ is the vector at $f(p) \in M$ obtained by multiplying the component column v^μ by the $m \times n$ Jacobian matrix $\partial f^a/\partial x^\mu$. As a derivation [a vector acting as a directional derivative, $v(g) = v^\mu \partial_\mu g$], $(f_* v)(g) = v(g \circ f)$: to differentiate a function on M along $f_* v$, pull the function back to N and differentiate it along v.

For the polar map, $f_*\partial_r = \cos\phi\,\partial_x + \sin\phi\,\partial_y$ and $f_*\partial_\phi = -r\sin\phi\,\partial_x + r\cos\phi\,\partial_y = -y\,\partial_x + x\,\partial_y$: the columns of the Jacobian, exactly as in Example T1.5.

T1.8.2 Pulling forms back

Vectors go forward. Forms go backward.

Working definition T1.22 (pullback)

For a function g on M, $f^*g = g \circ f$ [first move the point with f, then evaluate g: $(f^*g)(x) = g(f(x))$]. For a p-form ω on M, $f^*\omega$ is the p-form on N that eats vectors by pushing them forward first:

$$(f^*\omega)(v_1, \dots, v_p) = \omega(f_*v_1, \dots, f_*v_p). \tag{T1.86}$$

Recipe ("substitute and expand"): in ω, replace every coefficient $g(y)$ by $g(f(x))$ and every $\mathrm{d}y^a$ by

$$f^*\mathrm{d}y^a = \mathrm{d}f^a = \frac{\partial f^a}{\partial x^\mu}\,\mathrm{d}x^\mu, \tag{T1.87}$$

then multiply out using the wedge rules [$\mathrm{d}x^\mu \wedge \mathrm{d}x^\nu = -\mathrm{d}x^\nu \wedge \mathrm{d}x^\mu$, so $\mathrm{d}x^\mu \wedge \mathrm{d}x^\mu = 0$].

The recipe works because pullback respects every algebraic operation:

$$f^*(\alpha + \beta) = f^*\alpha + f^*\beta, \qquad f^*(g\,\alpha) = (f^*g)\,f^*\alpha, \qquad f^*(\alpha \wedge \beta) = f^*\alpha \wedge f^*\beta. \tag{T1.88}$$

The first two are clear from (T1.86). For the third, both sides are the determinant-type expression of (T1.20) evaluated on the pushed-forward vectors. And (T1.87) is (T1.86) for $\omega = \mathrm{d}y^a$: $(f^*\mathrm{d}y^a)(\partial_\mu) = \mathrm{d}y^a(f_*\partial_\mu) = \partial f^a/\partial x^\mu$. In components, a 1-form pulls back by the *transpose* of the Jacobian,

$$(f^*\alpha)_\mu = \frac{\partial f^a}{\partial x^\mu}\,\alpha_a(f(x)), \tag{T1.89}$$

which is the covariant transformation law (T1.12) seen from a new angle: a change of coordinates is a pullback along the transition map.

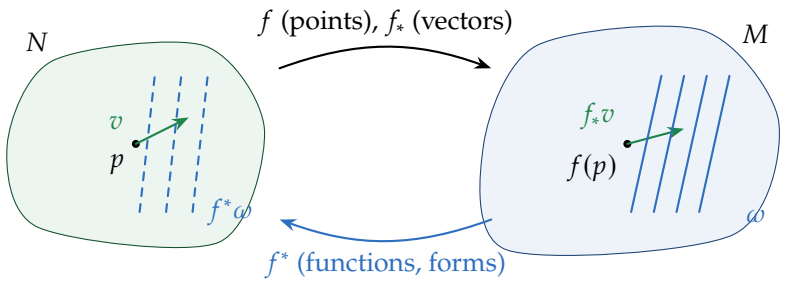


Figure T1.24: Points and vectors travel forward along $f : N \to M$; functions and forms travel backward. The pulled-back stack $f^*\omega$ at p is defined so that the number of its sheets pierced by v equals the number of sheets of ω pierced by f_*v. No inverse of f is needed, which is why we can always pull forms back but cannot, in general, push them forward.

Watch out: Why not push forms forward?

To push a form forward we would need to pull vectors back, i.e. invert the Jacobian. For the projection $(x, y, z) \mapsto (x, y)$ there is no way to say which vector upstairs a vector downstairs came from. Similarly a function on N cannot be pushed to M if two points of N land on the same point of M with different values. Pullback needs no choices; that is why it is the operation that appears in cohomology ($f^* : H^p(M) \to H^p(N)$) and in bundle theory (f^*E).

T1.8.3 Examples: curves, surfaces and top forms

Worked example T1.43: A 1-form pulled back to a curve is the line-integral integrand

Let $\alpha = -y\,dx + x\,dy + z\,dz$ on $\mathbb{R}^3$ and let $\gamma(t) = (\cos t, \sin t, t)$ be a helix. Then

$$\gamma^*\alpha \overset{(1)}{=} -\sin t\,d(\cos t) + \cos t\,d(\sin t) + t\,dt \overset{(2)}{=} \big(\sin^2 t + \cos^2 t + t\big)dt = (1+t)\,dt,$$
$$\int_\gamma \alpha \overset{(3)}{=} \int_0^{2\pi} (1+t)\,dt = 2\pi + 2\pi^2.$$

(1) Substitute $x = \cos t$, $y = \sin t$, $z = t$ in coefficients and differentials (T1.87).
(2) $d(\cos t) = -\sin t\,dt$, $d(\sin t) = \cos t\,dt$.
(3) The integral of a 1-form over an oriented curve is the ordinary integral of its pullback to the parameter interval (Chapter T2).

In vector language this is $\int \vec{F}\cdot d\vec{r} = \int \vec{F}\cdot\dot{\vec{r}}\,dt$ with $\vec{F} = (-y, x, z)$: the familiar rule "$d\vec{r} = \dot{\vec{r}}\,dt$" is the pullback. The coefficient of $\gamma^*\alpha$ is $\alpha(\gamma_*\partial_t) = \alpha(\dot\gamma)$, the form evaluated on the velocity, which is (T1.86). Note also that any 2-form pulls back to *zero* on a curve: there are no 2-forms on a 1-dimensional space.

Worked example T1.44: The flux through a graph: Dray's surface element

Let S be the graph $z = h(x,y)$, parametrised by $\sigma(x,y) = (x, y, h(x,y))$, and let $\star F = F_x\,dy\wedge dz + F_y\,dz\wedge dx + F_z\,dx\wedge dy$ be the flux 2-form of a vector field. Pull back, with $\sigma^*dz = h_x dx + h_y dy$:

$$\sigma^*(dy\wedge dz) \overset{(1)}{=} dy\wedge(h_x dx + h_y dy) = -h_x\,dx\wedge dy,$$
$$\sigma^*(dz\wedge dx) \overset{(2)}{=} (h_x dx + h_y dy)\wedge dx = -h_y\,dx\wedge dy,$$
$$\sigma^*(\star F) \overset{(3)}{=} \big(-F_x h_x - F_y h_y + F_z\big)\big|_{z=h}\,dx\wedge dy.$$

(1) $dy\wedge dy = 0$, $dy\wedge dx = -dx\wedge dy$.
(2) $dx\wedge dx = 0$.
(3) Add the three pieces; coefficients are evaluated on the surface.

This is $\vec{F}\cdot d\vec{A}$ with the upward surface element $d\vec{A} = (-h_x, -h_y, 1)\,dx\,dy$ [Dray (16.15)–(16.18)]. The vector-calculus derivation needs the cross product of two tangent vectors; the pullback does it by antisymmetry alone. For the plane $h = 0$ it gives $F_z\,dx\wedge dy$: Dray's remark "$z = 0$ on S, so $dz = 0$ there" is the statement $\sigma^*dz = 0$.

Key idea: Top forms pull back by the Jacobian determinant

If $\dim N = \dim M = n$, then

$$f^*\big(g\,dy^1\wedge\cdots\wedge dy^n\big) = (g\circ f)\,\det\!\left(\frac{\partial f^a}{\partial x^\mu}\right) dx^1\wedge\cdots\wedge dx^n, \tag{T1.90}$$

by (T1.87), (T1.88) and (T1.19). Hence, for a diffeomorphism f,

$$\int_N f^*\omega = \pm\int_M \omega, \tag{T1.91}$$

with $+$ if f preserves orientation ($\det > 0$) and $-$ if it reverses it. This is the change-of-variables formula of multivariable calculus. The measure-theory version has $|\det|$; forms

keep the sign, because an oriented integral is what flux and Stokes need.

Worked example T1.45: The area form of S^2, three ways

On $\mathbb{R}^3$ consider the 2-form

$$\omega = X\,\mathrm{d}Y\wedge\mathrm{d}Z + Y\,\mathrm{d}Z\wedge\mathrm{d}X + Z\,\mathrm{d}X\wedge\mathrm{d}Y,$$

the flux form of the position vector field $\vec{r}$. (In Section T1.9 we will recognise it as $\iota_{\vec{r}}\,\mathrm{vol}$.)
(a) Pull back along the angle map $i(\theta,\phi) = (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta)$. The differentials are $\mathrm{d}X = \cos\theta\cos\phi\,\mathrm{d}\theta - \sin\theta\sin\phi\,\mathrm{d}\phi$, $\mathrm{d}Y = \cos\theta\sin\phi\,\mathrm{d}\theta + \sin\theta\cos\phi\,\mathrm{d}\phi$, $\mathrm{d}Z = -\sin\theta\,\mathrm{d}\theta$. Then

$$\begin{aligned}
i^*(\mathrm{d}Y\wedge\mathrm{d}Z) &\overset{(1)}{=} \sin\theta\cos\phi\,\mathrm{d}\phi\wedge(-\sin\theta\,\mathrm{d}\theta) = \sin^2\theta\cos\phi\,\mathrm{d}\theta\wedge\mathrm{d}\phi,\\
i^*(\mathrm{d}Z\wedge\mathrm{d}X) &\overset{(2)}{=} (-\sin\theta\,\mathrm{d}\theta)\wedge(-\sin\theta\sin\phi\,\mathrm{d}\phi) = \sin^2\theta\sin\phi\,\mathrm{d}\theta\wedge\mathrm{d}\phi,\\
i^*(\mathrm{d}X\wedge\mathrm{d}Y) &\overset{(3)}{=} \big(\cos\theta\cos\phi\sin\theta\cos\phi + \sin\theta\sin\phi\cos\theta\sin\phi\big)\mathrm{d}\theta\wedge\mathrm{d}\phi\\
&= \sin\theta\cos\theta\,\mathrm{d}\theta\wedge\mathrm{d}\phi,\\
i^*\omega &\overset{(4)}{=} \big(\sin^3\theta\cos^2\phi + \sin^3\theta\sin^2\phi + \cos^2\theta\sin\theta\big)\mathrm{d}\theta\wedge\mathrm{d}\phi\\
&= \sin\theta\,\mathrm{d}\theta\wedge\mathrm{d}\phi.
\end{aligned}$$

(1) Only the $\mathrm{d}\phi$ part of $\mathrm{d}Y$ survives against $\mathrm{d}Z\propto\mathrm{d}\theta$.
(2) Only the $\mathrm{d}\phi$ part of $\mathrm{d}X$ survives.
(3) The $\mathrm{d}\theta\wedge\mathrm{d}\phi$ term and the $\mathrm{d}\phi\wedge\mathrm{d}\theta$ term, the latter with a sign flip: $(-\sin\theta\sin\phi)(\cos\theta\sin\phi)\,\mathrm{d}\phi\wedge\mathrm{d}\theta = +\sin\theta\cos\theta\sin^2\phi\,\mathrm{d}\theta\wedge\mathrm{d}\phi$.
(4) Multiply by the coefficients X, Y, Z pulled back, and use $\cos^2\phi+\sin^2\phi = 1$, $\sin^2\theta + \cos^2\theta = 1$.

So $i^*\omega$ is the area form of the unit sphere, with total $\int_0^{\pi}\int_0^{2\pi}\sin\theta\,\mathrm{d}\phi\,\mathrm{d}\theta = 4\pi$.
(b) Pull back further, to the stereographic chart. In the north chart, $u+\mathrm{i}v = \rho\,e^{\mathrm{i}\phi}$ with $\rho = \cot(\theta/2)$ (Example T1.9(c)). Rather than substituting (T1.4) into ω (possible but long), we use $(f\circ g)^* = g^*f^*$ below and pull back $\sin\theta\,\mathrm{d}\theta\wedge\mathrm{d}\phi$ along $(u,v)\mapsto(\theta,\phi)$. It is easier to go the other way and express $\mathrm{d}u\wedge\mathrm{d}v$ in θ,ϕ:

$$\begin{aligned}
\frac{4\,\mathrm{d}u\wedge\mathrm{d}v}{(1+\rho^2)^2} &\overset{(1)}{=} \frac{4\rho\,\mathrm{d}\rho\wedge\mathrm{d}\phi}{(1+\rho^2)^2} \overset{(2)}{=} 4\sin^4\tfrac{\theta}{2}\,\cot\tfrac{\theta}{2}\left(-\tfrac12\csc^2\tfrac{\theta}{2}\right)\mathrm{d}\theta\wedge\mathrm{d}\phi\\
&\overset{(3)}{=} -2\sin\tfrac{\theta}{2}\cos\tfrac{\theta}{2}\,\mathrm{d}\theta\wedge\mathrm{d}\phi \overset{(4)}{=} -\sin\theta\,\mathrm{d}\theta\wedge\mathrm{d}\phi.
\end{aligned}$$

(1) Polar coordinates in the (u,v) plane: $\mathrm{d}u\wedge\mathrm{d}v = \rho\,\mathrm{d}\rho\wedge\mathrm{d}\phi$ (Example T1.12).
(2) $1+\cot^2 = \csc^2$, so $(1+\rho^2)^{-2} = \sin^4(\theta/2)$; and $\mathrm{d}\rho = \mathrm{d}\cot(\theta/2) = -\tfrac12\csc^2(\theta/2)\,\mathrm{d}\theta$.
(3) Simplify: $\sin^4\cdot\cot\cdot\csc^2 = \sin\cdot\cos$.
(4) Double-angle formula.

So on the sphere $\omega = \sin\theta\,\mathrm{d}\theta\wedge\mathrm{d}\phi = -\dfrac{4\,\mathrm{d}u\wedge\mathrm{d}v}{(1+u^2+v^2)^2}$. The magnitude $4/(1+\rho^2)^2$ is $\sqrt{\det g}$ for the stereographic metric of Example T1.9, as it must be. The *minus sign* says the north stereographic chart (u,v) is orientation-reversing relative to the outward normal. Check at the south pole, where $u=v=0$: there $X\approx 2u$, $Y\approx 2v$, $Z\approx -1$, so $\omega\approx Z\,\mathrm{d}X\wedge\mathrm{d}Y = -4\,\mathrm{d}u\wedge\mathrm{d}v$. Looking at the plane from above, the outward normal at the south pole points *down*. This is the orientation subtlety of Section T1.1, now visible in a single sign.

(c) The Coulomb/monopole form. On $\mathbb{R}^3 \setminus \{0\}$ the 2-form ω/r^3 is closed (it is $4\pi/g$ times the monopole field (T1.80), written in Cartesian coordinates). Pulled back to the sphere of radius R, $(X, Y, Z) = R\,(\sin\theta\cos\phi, \dots)$, it gives $R^3 \sin\theta\, \mathrm{d}\theta \wedge \mathrm{d}\phi/R^3$: independent of R, since each of X, Y, Z and each differential scale by R. The flux through every sphere is 4π, the same number, as a closed form must give on surfaces that can be deformed into each other (Chapter T2).

T1.8.4 Pullback of the metric: induced metrics

The same recipe works for symmetric products of 1-forms. For the inclusion i of a surface, $i^*(\mathrm{d}X^2 + \mathrm{d}Y^2 + \mathrm{d}Z^2)$ is the **induced metric**: the metric that measures lengths of curves drawn on the surface using the ambient ruler. That is what we computed in Example T1.9; in components,

$$(f^*g)_{\mu\nu} = \frac{\partial f^a}{\partial x^\mu}\frac{\partial f^b}{\partial x^\nu}\, g_{ab}(f(x)), \qquad \text{i.e.} \qquad f^*g = J^{\mathsf{T}} g J. \tag{T1.92}$$

Worked example T1.46: The induced metric and area of a graph

For $\sigma(x, y) = (x, y, h)$, with $\mathrm{d}z = h_x \mathrm{d}x + h_y \mathrm{d}y$:

$$\begin{aligned} \sigma^*(\mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2) &\overset{(1)}{=} \mathrm{d}x^2 + \mathrm{d}y^2 + (h_x \mathrm{d}x + h_y \mathrm{d}y)^2 \\ &= (1 + h_x^2)\,\mathrm{d}x^2 + 2h_x h_y\, \mathrm{d}x\, \mathrm{d}y + (1 + h_y^2)\,\mathrm{d}y^2, \\ \det(g_{\mu\nu}) &\overset{(2)}{=} (1 + h_x^2)(1 + h_y^2) - h_x^2 h_y^2 = 1 + h_x^2 + h_y^2. \end{aligned}$$

(1) Substitute and expand; $\mathrm{d}x\,\mathrm{d}y$ here is the *symmetric* product, so the cross term has coefficient $2h_x h_y$ and $g_{xy} = g_{yx} = h_x h_y$.
(2) 2×2 determinant.

The area form is $\sqrt{1 + h_x^2 + h_y^2}\,\mathrm{d}x \wedge \mathrm{d}y$, the familiar formula. And the unit normal flux of Example T1.44, $(-h_x, -h_y, 1)/\sqrt{1 + |\nabla h|^2}$, times this area form gives back $\mathrm{d}\vec{A} = (-h_x, -h_y, 1)\,\mathrm{d}x\,\mathrm{d}y$: the two computations agree.

Watch out: Pullback does not commute with $\star$

The Hodge star uses a metric, and f need not respect it. Take the dilation $f(x, y) = (\lambda x, \lambda y)$ of $\mathbb{R}^2$. Then $f^*(\star 1) = f^*(\mathrm{d}x \wedge \mathrm{d}y) = \lambda^2\, \mathrm{d}x \wedge \mathrm{d}y$, whereas $\star(f^*1) = \star 1 = \mathrm{d}x \wedge \mathrm{d}y$. They agree only for $\lambda = \pm 1$. In general $f^*\star = \pm \star f^*$ only when f is an *isometry* ($f^*g = g$), with the sign given by orientation. This is why $\mathrm{d}F = 0$ is invariant under every diffeomorphism while $\mathrm{d} \star F = \star J$ is only invariant under isometries (Lorentz transformations, rotations).

T1.8.5 Naturality: $\mathrm{d}f^* = f^*\mathrm{d}$

Key idea: Pullback commutes with d

For every smooth map $f : N \to M$ and every form ω on M,

$$\mathrm{d}(f^*\omega) = f^*(\mathrm{d}\omega). \tag{T1.93}$$

"Differentiate, then substitute" equals "substitute, then differentiate".

Why. For a function g this is the chain rule:

$$\mathrm{d}(f^*g) \overset{(1)}{=} \frac{\partial(g\circ f)}{\partial x^\mu}\,\mathrm{d}x^\mu \overset{(2)}{=} \frac{\partial g}{\partial y^a}\frac{\partial f^a}{\partial x^\mu}\,\mathrm{d}x^\mu \overset{(3)}{=} \frac{\partial g}{\partial y^a}\,f^*\mathrm{d}y^a \overset{(4)}{=} f^*(\mathrm{d}g).$$

(1) Definition of d on the function $g\circ f$ on N.
(2) Chain rule.
(3) (T1.87).
(4) $\mathrm{d}g = \partial_a g\,\mathrm{d}y^a$ pulled back by (T1.88).

For a p-form $\omega = g\,\mathrm{d}y^{a_1}\wedge\cdots\wedge\mathrm{d}y^{a_p}$ (and sums of such), $f^*\omega = (g\circ f)\,\mathrm{d}f^{a_1}\wedge\cdots\wedge\mathrm{d}f^{a_p}$. Differentiate with the Leibniz rule: every $\mathrm{d}(\mathrm{d}f^{a_i})$ vanishes by $\mathrm{d}^2=0$ on N, leaving $\mathrm{d}(g\circ f)\wedge\mathrm{d}f^{a_1}\wedge\ldots$, which is $f^*(\mathrm{d}g)\wedge f^*\mathrm{d}y^{a_1}\wedge\cdots = f^*(\mathrm{d}\omega)$ by the function case. Everything reduces to the chain rule plus $\mathrm{d}^2=0$.

Worked example T1.47: Checking naturality on the sphere (Dray §20.5)

Take $\beta = X\,\mathrm{d}Y$ on $\mathbb{R}^3$, so $\mathrm{d}\beta = \mathrm{d}X\wedge\mathrm{d}Y$. Pull back to the unit sphere with the angle map i of Example T1.45.
Differentiate first: $i^*(\mathrm{d}\beta) = i^*(\mathrm{d}X\wedge\mathrm{d}Y) = \sin\theta\cos\theta\,\mathrm{d}\theta\wedge\mathrm{d}\phi$, from step (3) of Example T1.45.
Pull back first:

$$\begin{aligned} i^*\beta &\overset{(1)}{=} \sin\theta\cos\phi\,(\cos\theta\sin\phi\,\mathrm{d}\theta + \sin\theta\cos\phi\,\mathrm{d}\phi) \\ &= \sin\theta\cos\theta\sin\phi\cos\phi\,\mathrm{d}\theta + \sin^2\theta\cos^2\phi\,\mathrm{d}\phi, \\ \mathrm{d}(i^*\beta) &\overset{(2)}{=} \sin\theta\cos\theta\,(\cos^2\phi - \sin^2\phi)\,\mathrm{d}\phi\wedge\mathrm{d}\theta + 2\sin\theta\cos\theta\cos^2\phi\,\mathrm{d}\theta\wedge\mathrm{d}\phi \\ &\overset{(3)}{=} \sin\theta\cos\theta\,(-\cos^2\phi + \sin^2\phi + 2\cos^2\phi)\mathrm{d}\theta\wedge\mathrm{d}\phi = \sin\theta\cos\theta\,\mathrm{d}\theta\wedge\mathrm{d}\phi. \end{aligned}$$

(1) Substitute $X = \sin\theta\cos\phi$ and $\mathrm{d}Y$ from Example T1.45.
(2) d of the $\mathrm{d}\theta$ term: only ∂_ϕ survives; d of the $\mathrm{d}\phi$ term: only ∂_θ survives.
(3) $\mathrm{d}\phi\wedge\mathrm{d}\theta = -\mathrm{d}\theta\wedge\mathrm{d}\phi$; then $\cos^2\phi+\sin^2\phi = 1$.

The two agree, as (T1.93) promised. Now use it. Integrate $\mathrm{d}X\wedge\mathrm{d}Y$ over the upper hemisphere H ($0\le\theta\le\pi/2$):

$$\int_H \sin\theta\cos\theta\,\mathrm{d}\theta\,\mathrm{d}\phi = 2\pi\cdot\tfrac{1}{2} = \pi,$$

while Stokes on the pulled-back form gives the integral of $X\,\mathrm{d}Y$ around the equator, $\int_0^{2\pi}\cos^2\phi\,\mathrm{d}\phi = \pi$. Both are the area of the unit disc: $\mathrm{d}X\wedge\mathrm{d}Y$ measures the *shadow* of the hemisphere on the XY-plane (Figure T1.25). Over the whole sphere the integral is 0: the shadows of the two hemispheres cancel, and the sphere has no boundary.

T1.8.6 Composition, winding and degree

Suppose we have two maps in a row, $P \xrightarrow{g} N \xrightarrow{f} M$. Pulling back along f first and then along g is the same as pulling back along the composite $f\circ g$ in one go:

$$(f\circ g)^* = g^*\circ f^* \qquad \text{for} \qquad P \xrightarrow{g} N \xrightarrow{f} M. \tag{T1.94}$$

The order reverses, just like $(AB)^{\mathsf{T}} = B^{\mathsf{T}}A^{\mathsf{T}}$, and for the same reason: (T1.89) is multiplication by a transposed Jacobian, and the Jacobian of a composite is the product of Jacobians (chain rule). We used this in Example T1.45(b).

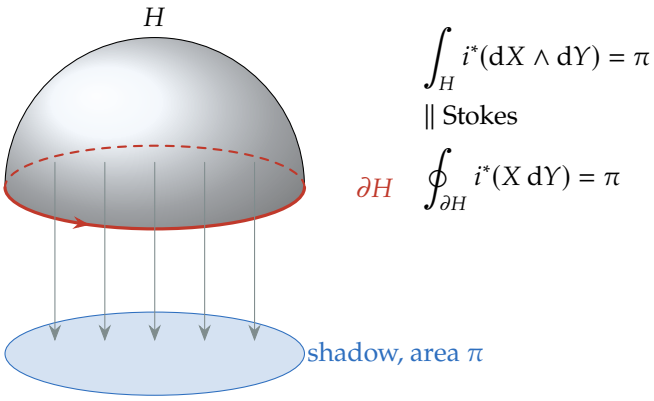


Figure T1.25: Naturality at work. The 2-form $dX \wedge dY$ pulled back to the upper hemisphere measures the area of its shadow on the XY-plane (the blue disc). Because $i^*d = di^*$, Stokes' theorem can be applied on the sphere itself, and the answer is the loop integral of $X\,dY$ around the red equator.

Naturality makes some pullbacks almost free. Here is the cleanest example.

Worked example T1.48: The squaring map pulls $d\phi$ back to $2\,d\phi$

Let $f : \mathbb{R}^2 \setminus \{0\} \to \mathbb{R}^2 \setminus \{0\}$ be $z \mapsto z^2$ in complex notation, i.e. $(x, y) \mapsto (x^2 - y^2,\ 2xy)$, and let $\beta = (x\,dy - y\,dx)/(x^2 + y^2)$ be the closed form of Example T1.30.

The quick way. In polar coordinates $f(r, \phi) = (r^2, 2\phi)$, and locally $\beta = d\phi$. By naturality, $f^*\beta = f^*d\phi = d(f^*\phi) = d(2\phi) = 2\,d\phi = 2\beta$.

The brute-force way, to see that nothing was swept under the rug. With $x' = x^2 - y^2$, $y' = 2xy$: $dx' = 2x\,dx - 2y\,dy$ and $dy' = 2y\,dx + 2x\,dy$. Then

$$\begin{aligned} x'dy' - y'dx' &\overset{(1)}{=} (x^2 - y^2)(2y\,dx + 2x\,dy) - 2xy(2x\,dx - 2y\,dy) \\ &\overset{(2)}{=} \big(2x^2y - 2y^3 - 4x^2y\big)dx + \big(2x^3 - 2xy^2 + 4xy^2\big)dy \\ &\overset{(3)}{=} 2(x^2 + y^2)\,(x\,dy - y\,dx), \\ f^*\beta &\overset{(4)}{=} \frac{2(x^2 + y^2)(x\,dy - y\,dx)}{(x^2 + y^2)^2} = 2\beta. \end{aligned}$$

(1) Substitute.
(2) Collect the dx and dy terms.
(3) $-2x^2y - 2y^3 = -2y(x^2 + y^2)$ and $2x^3 + 2xy^2 = 2x(x^2 + y^2)$.
(4) $x'^2 + y'^2 = (x^2 - y^2)^2 + 4x^2y^2 = (x^2 + y^2)^2$.

Integrating around the unit circle, $\oint f^*\beta = 2\oint \beta = 4\pi$: the image of the circle winds twice around the origin. For $z \mapsto z^n$ the same argument gives $f^*\beta = n\beta$. The number $\frac{1}{2\pi}\oint \gamma^*\beta$ is the **winding number** of a loop γ around the origin, and more generally $\int_N f^*\omega = (\deg f)\int_M \omega$ defines the **degree** of a map between closed oriented manifolds of the same dimension. For $\theta \mapsto n\theta$ on S^1, $f^*d\theta = n\,d\theta$, so $\deg f = n$.

Physical picture: Why pullbacks make cohomology a topological invariant

If $d\omega = 0$ then $d(f^*\omega) = f^*d\omega = 0$, and if $\omega = d\alpha$ then $f^*\omega = d(f^*\alpha)$. So f^* sends closed forms to closed forms and exact forms to exact forms: it gives a well-defined map

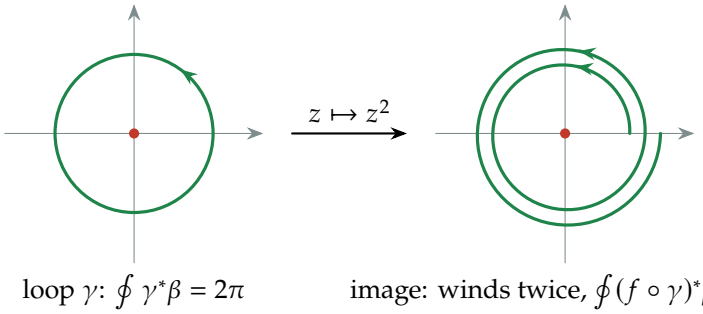


Figure T1.26: The squaring map doubles winding. The image loop (drawn slightly spread out so both turns are visible) goes around the origin twice. Algebraically this is the one-line statement $f^*\beta = 2\beta$, obtained from naturality: $f^*\mathrm{d}\phi = \mathrm{d}(f^*\phi) = \mathrm{d}(2\phi)$.

$f^* : H^p_{\mathrm{dR}}(M) \to H^p_{\mathrm{dR}}(N)$ on cohomology. Moreover, if f_t is a smooth family of maps, then $f_1^*\omega - f_0^*\omega$ is exact for closed ω (this follows from Cartan's formula of Section T1.9, integrated over t; Chapter T2). So homotopic maps induce the same map on cohomology. That is the property Nash quotes in (1.48) for bundles and uses implicitly for cohomology on p. 15.

Try it: Pullback under an orientation-reversing map

Let $r(x, y) = (x, -y)$ be the reflection of $\mathbb{R}^2$. Compute $r^*(\mathrm{d}x \wedge \mathrm{d}y)$ and $r^*\beta$, with β the angle form. What is the winding number of the reflected unit circle?
Answer: $r^*(\mathrm{d}x \wedge \mathrm{d}y) = \mathrm{d}x \wedge (-\mathrm{d}y) = -\mathrm{d}x \wedge \mathrm{d}y$ ($\det = -1$), and $r^*\beta = (x(-\mathrm{d}y) - (-y)\mathrm{d}x)/r^2 = -\beta$. The reflected circle winds -1 times: reflection reverses orientation, and the pullback records it as a sign.

Reading Nash (p. 15, (1.48)–(1.50); p. 27)

Nash introduces pullback on p. 15 in the language of cohomology: "the principal ingredient in any pullback calculation is a map $f : N \to M$, then if $\omega \in H^n(M; G)$ one can pull back ω to $f^*\omega$". For de Rham cohomology this is exactly our pullback of closed forms, which is well defined on classes because of naturality (T1.93). He then pulls back bundles, f^*E, and states that homotopic maps give isomorphic pullbacks (1.48); every bundle with group G is the pullback f^*E_G of a universal bundle, and its characteristic classes are pulled back from the classifying space, $f^* : H^*(BG) \to H^*(M)$ (1.50). All of these rest on the two properties we have established: $(f \circ g)^* = g^*f^*$ and $\mathrm{d}f^* = f^*\mathrm{d}$. His winding numbers and degrees of maps (Chapter 1, and the instanton number in Chapter 8) are the $\int f^*\omega$ of Example T1.48.

Recap

- $f : N \to M$ pushes vectors forward by the Jacobian, $(f_*v)^a = \partial_\mu f^a\, v^\mu$, and pulls forms back: $(f^*\omega)(v, \dots) = \omega(f_*v, \dots)$.
- Recipe: substitute $y = f(x)$ and $\mathrm{d}y^a = \partial_\mu f^a \mathrm{d}x^\mu$, then expand with the wedge rules. f^* respects $+$, $\wedge$ and products with functions.
- Top forms pull back by $\det J$; this is change of variables, with orientation sign. A line integral is $\int \gamma^*\alpha$; a flux is $\int \sigma^*(\star F)$; the induced metric is $f^*g = J^{\mathsf{T}} g J$.

- Naturality: $\mathrm{d}f^* = f^*\mathrm{d}$ (chain rule plus $\mathrm{d}^2 = 0$), and $(f \circ g)^* = g^* f^*$. Hence f^* acts on cohomology.
- f^* does not commute with $\star$ unless f is an isometry.
- Winding numbers and degrees: $\int f^*\omega = \deg f \int \omega$; $z \mapsto z^n$ pulls $\mathrm{d}\phi$ back to $n\,\mathrm{d}\phi$.

T1.9 Interior product, Lie derivative and Cartan's formula

Physics is full of questions of the form "how does this field change as we slide along that flow?". A fluid carries a density and a magnetic field along with it; does the magnetic flux through a comoving loop change? A Hamiltonian flow moves points of phase space; does it squeeze phase-space area? A rotation moves points of a sphere; does it change the metric? In Nash's Chapter 5 the question is: which changes of the metric g of a Riemann surface are "fake", produced merely by sliding the surface along a vector field V (a diffeomorphism), and which change its true shape?

The tool that answers all these questions is the **Lie derivative** $\mathcal{L}_X$: the rate of change of a field as it is dragged along the flow of a vector field X. It sounds complicated, since it involves solving the flow equations, but for forms it collapses to a two-term formula, *Cartan's formula* $\mathcal{L}_X = \mathrm{d}\,\iota_X + \iota_X\mathrm{d}$, which needs only d and one very simple algebraic operation ι_X, "insert X in the first slot". We start with that operation.

T1.9.1 The interior product: feeding a form one vector

A p-form eats p vectors and returns a number. If we feed it only one vector X and leave the other slots open, what remains is a machine that eats $p-1$ vectors: a $(p-1)$-form.

Working definition T1.23 (interior product)

For a vector field X and a p-form α, the **interior product** (or contraction) $\iota_X\alpha$ is the $(p-1)$-form

$$(\iota_X\alpha)(v_2,\dots,v_p) = \alpha(X, v_2,\dots,v_p), \qquad (\iota_X\alpha)_{\nu_2\dots\nu_p} = X^\mu \alpha_{\mu\nu_2\dots\nu_p}. \tag{T1.95}$$

On functions $\iota_X f = 0$. *Recipe:* for a 1-form, $\iota_X \mathrm{d}x^\mu = X^\mu$. For a wedge product, move each factor to the front in turn, with the sign of the move [each move past one 1-form factor gives -1, e.g. $\iota_X(\mathrm{d}x\wedge\mathrm{d}y) = X^x\mathrm{d}y - X^y\mathrm{d}x$], and replace it by its value on X (the antiderivation rule (T1.96) below).

The component formula follows from the $1/p!$ convention (T1.22): in $\alpha = \frac{1}{p!}\alpha_{\mu_1\dots\mu_p}\mathrm{d}x^{\mu_1}\wedge\dots$, inserting X can hit any of the p factors, and after moving each one to the front the p contributions are equal, turning $\frac{1}{p!}$ into $\frac{p}{p!} = \frac{1}{(p-1)!}$, which is the normalisation of a $(p-1)$-form.

Let us see it act. In $\mathbb{R}^2$ with $X = X^x\partial_x + X^y\partial_y$,

$$\iota_X(\mathrm{d}x\wedge\mathrm{d}y) \overset{(1)}{=} (\mathrm{d}x\wedge\mathrm{d}y)(X,\cdot) \overset{(2)}{=} \mathrm{d}x(X)\,\mathrm{d}y - \mathrm{d}y(X)\,\mathrm{d}x \overset{(3)}{=} X^x\,\mathrm{d}y - X^y\,\mathrm{d}x.$$

(1) Definition (T1.95): fill the first slot with X.
(2) The determinant rule for a wedge of two 1-forms: $(a\wedge b)(X,w) = a(X)b(w) - b(X)a(w)$.
(3) $\mathrm{d}x(X) = X^x$ and $\mathrm{d}y(X) = X^y$.

In particular $\iota_{\partial_x}(\mathrm{d}x \wedge \mathrm{d}y) = \mathrm{d}y$ and $\iota_{\partial_y}(\mathrm{d}x \wedge \mathrm{d}y) = -\mathrm{d}x$: to contract with ∂_y we first move $\mathrm{d}y$ to the front, which costs a sign.

Key idea: The interior product is an antiderivation

For a p-form α and any form β,

$$\iota_X(\alpha \wedge \beta) = (\iota_X \alpha) \wedge \beta + (-1)^p\, \alpha \wedge \iota_X \beta, \qquad \iota_X \iota_X = 0. \tag{T1.96}$$

This is the same sign rule as for d (T1.46): an operator of odd degree picks up $(-1)^p$ when it moves past a p-form.

For two 1-forms the rule is exactly step (2) above: $\iota_X(a \wedge b) = a(X)\, b - b(X)\, a = (\iota_X a)\, b - a\, (\iota_X b)$. The general case follows by writing forms as sums of wedges of 1-forms and applying the two-factor case repeatedly. And $\iota_X \iota_X \alpha = \alpha(X, X, \dots) = 0$ because α is antisymmetric; this is the algebraic twin of $\mathrm{d}^2 = 0$.

The three-dimensional dictionary for ι_X

In $\mathbb{R}^3$ with $\mathrm{vol} = \mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z$ and the dictionary of Section T1.3 (a vector $\vec{A}$ corresponds to the 1-form $A = \vec{A} \cdot \mathrm{d}\vec{r}$ and to the flux 2-form $\star A$), the interior product with $X = \vec{X} \cdot \nabla$ is:

$$\begin{aligned} \iota_X A &= \vec{X} \cdot \vec{A}, \\ \iota_X(\star B) &= (\vec{B} \times \vec{X}) \cdot \mathrm{d}\vec{r}, \\ \iota_X(\rho\, \mathrm{vol}) &= \rho \star X = \text{flux form of } \rho \vec{X}. \end{aligned} \tag{T1.97}$$

The first is (T1.95) for 1-forms. For the second, use the recipe term by term:

$$\begin{aligned} &\iota_X\big(B_x\, \mathrm{d}y \wedge \mathrm{d}z + B_y\, \mathrm{d}z \wedge \mathrm{d}x + B_z\, \mathrm{d}x \wedge \mathrm{d}y\big) \\ &\overset{(1)}{=} B_x(X^y \mathrm{d}z - X^z \mathrm{d}y) + B_y(X^z \mathrm{d}x - X^x \mathrm{d}z) + B_z(X^x \mathrm{d}y - X^y \mathrm{d}x) \\ &\overset{(2)}{=} (B_y X^z - B_z X^y)\, \mathrm{d}x + (B_z X^x - B_x X^z)\, \mathrm{d}y + (B_x X^y - B_y X^x)\, \mathrm{d}z \\ &\overset{(3)}{=} (\vec{B} \times \vec{X}) \cdot \mathrm{d}\vec{r}. \end{aligned}$$

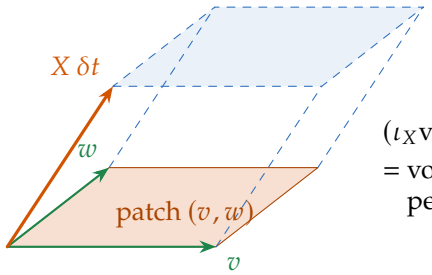
(1) $\iota_X(a \wedge b) = a(X)b - b(X)a$ on each basis 2-form.
(2) Collect the coefficients of $\mathrm{d}x$, $\mathrm{d}y$, $\mathrm{d}z$.
(3) $(\vec{B} \times \vec{X})_x = B_y X^z - B_z X^y$, and cyclically.

For the third, $\iota_X(\mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z) = X^x\, \mathrm{d}y \wedge \mathrm{d}z - X^y\, \mathrm{d}x \wedge \mathrm{d}z + X^z\, \mathrm{d}x \wedge \mathrm{d}y$ (move each factor to the front: $\mathrm{d}y$ costs one sign, $\mathrm{d}z$ two), and $-\mathrm{d}x \wedge \mathrm{d}z = \mathrm{d}z \wedge \mathrm{d}x$, so this is $X^x\, \mathrm{d}y \wedge \mathrm{d}z + X^y\, \mathrm{d}z \wedge \mathrm{d}x + X^z\, \mathrm{d}x \wedge \mathrm{d}y = \star X$.

So the flux 2-form of the position vector, $\omega = X\, \mathrm{d}Y \wedge \mathrm{d}Z + Y\, \mathrm{d}Z \wedge \mathrm{d}X + Z\, \mathrm{d}X \wedge \mathrm{d}Y$ of Example T1.45, is $\iota_{\vec{r}}\, \mathrm{vol}$, as promised there. The same computation in n dimensions (Riemannian) gives

$$\iota_X \mathrm{vol} = \star X^\flat, \tag{T1.98}$$

where $X^\flat = g_{\mu\nu} X^\nu \mathrm{d}x^\mu$ is the 1-form obtained by lowering the index (Section T1.2). The third line of (T1.97) has a physical reading: a charge density ρ moving with velocity $\vec{v}$ gives $\iota_v(\rho\, \mathrm{vol}) = \star J$ with $\vec{J} = \rho \vec{v}$. *Contracting a density with a velocity gives the current.*



$(\iota_X \mathrm{vol})(v,w) = \mathrm{vol}(X,v,w)$
= volume swept by the patch per unit time

Figure T1.27: What $\iota_X \mathrm{vol}$ measures. A fluid moving with velocity X sweeps the orange patch spanned by v, w through the blue parallelepiped in time δt. Its volume, divided by δt, is $\mathrm{vol}(X,v,w)$: the amount of fluid crossing the patch per unit time. That is why $\iota_X \mathrm{vol}$ is the flux 2-form of X, and $\iota_v(\rho\,\mathrm{vol})$ is the current 2-form of a moving charge density.

Worked example T1.49: The Lorentz force as a contraction

In Example T1.37 we inserted the 4-velocity $u = \partial_t + \vec{v}\cdot\nabla$ into the *second* slot of F. With our first-slot convention that is $F(\,\cdot\,, u) = -\iota_u F$. Let us recompute it with the recipe, using $F = E\wedge \mathrm{d}t + \star_3 B$ (T1.64):

$$\iota_u(E\wedge \mathrm{d}t) \overset{(1)}{=} (\iota_u E)\,\mathrm{d}t - E\,(\iota_u \mathrm{d}t) \overset{(2)}{=} (\vec{E}\cdot\vec{v})\,\mathrm{d}t - E,$$
$$\iota_u(\star_3 B) \overset{(3)}{=} \iota_v(\star_3 B) \overset{(4)}{=} (\vec{B}\times\vec{v})\cdot \mathrm{d}\vec{r},$$
$$-q\,\iota_u F \overset{(5)}{=} q\,(\vec{E} + \vec{v}\times\vec{B})\cdot\mathrm{d}\vec{r} - q\,(\vec{E}\cdot\vec{v})\,\mathrm{d}t.$$

(1) Antiderivation rule (T1.96) with the 1-form E.
(2) E has no $\mathrm{d}t$ part, so $\iota_u E = \vec{v}\cdot\vec{E}$; and $\iota_u \mathrm{d}t = u^t = 1$.
(3) $\star_3 B$ contains no $\mathrm{d}t$, so the ∂_t part of u gives nothing.
(4) Second line of (T1.97).
(5) Add, multiply by $-q$, and use $-\vec{B}\times\vec{v} = \vec{v}\times\vec{B}$.

This is the answer of Example T1.37, obtained without writing the 4×4 matrix: the magnetic force $\vec{v}\times\vec{B}$ is literally the flux form $\star_3 B$ with the velocity inserted.

T1.9.2 Flows of vector fields

A vector field X assigns an arrow to every point. Its **integral curves** are the curves that follow the arrows,

$$\frac{\mathrm{d}x^\mu(t)}{\mathrm{d}t} = X^\mu\big(x(t)\big), \tag{T1.99}$$

a system of first-order ODEs that has a unique solution through every starting point (at least for a short time). The **flow** of X is the family of maps φ_t that sends each point to where its integral curve is after time t. Uniqueness of solutions gives the group law

$$\varphi_0 = \mathrm{id}, \qquad \varphi_s\circ\varphi_t = \varphi_{s+t}, \qquad \varphi_t^{-1} = \varphi_{-t}, \tag{T1.100}$$

and for small t,

$$\varphi_t(x)^\mu = x^\mu + t\,X^\mu(x) + O(t^2). \tag{T1.101}$$

Physicists write $\varphi_t = \exp(tX)$, and so does Nash [Nash p. 159, (5.66)]: the vector field is the infinitesimal generator of the one-parameter group of diffeomorphisms, just as the angular momentum operator generates rotations.

Worked example T1.50: Three flows in the plane

(a) Translation, $X = \partial_x$: $\dot{x} = 1$, $\dot{y} = 0$, so $\varphi_t(x, y) = (x + t, y)$.
(b) Rotation, $X = \partial_\phi = -y\,\partial_x + x\,\partial_y$ (Example T1.5):

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} x \\ y \end{pmatrix} \overset{(1)}{=} \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} \quad\Longrightarrow\quad \varphi_t \begin{pmatrix} x \\ y \end{pmatrix} \overset{(2)}{=} \exp\begin{pmatrix} 0 & -t \\ t & 0 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} \overset{(3)}{=} \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}.$$

(1) Read off $\dot{x} = -y$, $\dot{y} = x$ from the components of X.
(2) A linear ODE $\dot{\mathbf{x}} = A\mathbf{x}$ is solved by $\mathbf{x}(t) = e^{tA}\mathbf{x}(0)$.
(3) Sum the exponential series: $A^2 = -\mathbb{1}$, so $e^{tA} = \cos t\,\mathbb{1} + \sin t\,A$.

The flow is rotation by angle t; here "$\varphi_t = \exp(tX)$" is literally a matrix exponential.
(c) Dilation, $X = x\,\partial_x + y\,\partial_y = r\,\partial_r$: $\dot{x} = x$, $\dot{y} = y$, so $\varphi_t(x, y) = (e^t x, e^t y)$, scaling by e^t.

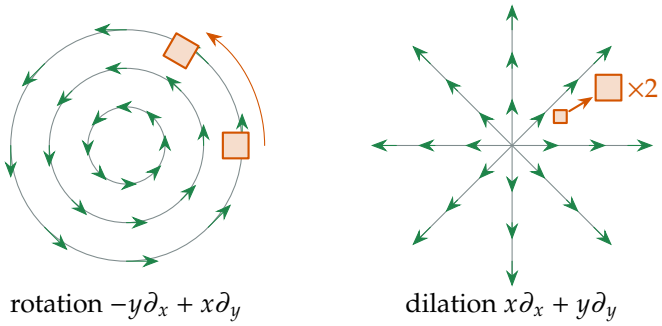


Figure T1.28: The flows of Example T1.50(b) and (c). Green arrows are the vector field; the orange square is carried along by the flow. The rotation moves the square without changing its area or shape; the dilation doubles its side and quadruples its area. The Lie derivatives $\mathcal{L}_X\mathrm{vol} = 0$ and $\mathcal{L}_X\mathrm{vol} = 2\,\mathrm{vol}$ of Example T1.51 are the infinitesimal versions of these two facts.

T1.9.3 The Lie derivative

To compare a field at $\varphi_t(p)$ with the field at p we must bring it back to p, and pullback is exactly the tool that does this without any choices (Section T1.8).

Working definition T1.24 (Lie derivative)

The **Lie derivative** of a function, form or metric T along X [along the flow φ_t of X: $\varphi_t(p)$ is where p arrives after time t following $\dot{x}^\mu = X^\mu$] is

$$\mathcal{L}_X T = \left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} \varphi_t^* T. \tag{T1.102}$$

For a vector field Y, "pull back" means push forward by the inverse, $\varphi_t^* Y = (\varphi_{-t})_* Y$. *Operationally:* use the first-order flow (T1.101), substitute, and keep the terms linear in t. The results are the component formulae (T1.103).

Functions. $\varphi_t^* f(x) = f(x + tX) = f(x) + tX^\mu \partial_\mu f + O(t^2)$, so $\mathcal{L}_X f = X(f)$: the directional derivative, nothing new.

1-forms. By the pullback rule (T1.89),

$$(\varphi_t^*\alpha)_\mu(x) \overset{(1)}{=} \frac{\partial \varphi_t^\nu}{\partial x^\mu}\,\alpha_\nu\big(\varphi_t(x)\big) \overset{(2)}{=} \big(\delta_\mu^\nu + t\,\partial_\mu X^\nu\big)\big(\alpha_\nu + t\,X^\lambda \partial_\lambda \alpha_\nu\big) + O(t^2)$$

$$\overset{(3)}{=} \alpha_\mu + t\left(X^\lambda \partial_\lambda \alpha_\mu + \alpha_\nu\, \partial_\mu X^\nu\right) + O(t^2).$$

(1) A 1-form pulls back by the transposed Jacobian.
(2) Differentiate (T1.101) for the Jacobian, and Taylor expand α_ν about x.
(3) Multiply out and keep the terms of order t.

The first term is the naive "transport the components" derivative. The second is new: it says that even a 1-form with constant components changes if the flow rotates or stretches the coordinate grid.

Vector fields. Now we need $(\varphi_{-t})_*$ applied to Y at $\varphi_t(x)$. The Jacobian of φ_{-t} is $\delta^\mu_\nu - t\,\partial_\nu X^\mu$, so

$$\begin{aligned}\left((\varphi_{-t})_* Y\right)^\mu(x) &\overset{(1)}{=} \left(\delta^\mu_\nu - t\,\partial_\nu X^\mu\right) Y^\nu(x + tX) + O(t^2)\\ &\overset{(2)}{=} Y^\mu + t\left(X^\lambda \partial_\lambda Y^\mu - Y^\nu \partial_\nu X^\mu\right) + O(t^2).\end{aligned}$$

(1) Push forward (T1.85) with the Jacobian of φ_{-t}, applied to the vector sitting at $\varphi_t(x) = x + tX$.
(2) Taylor expand $Y^\nu(x + tX)$ and keep order t.

The bracket is the commutator of Section T1.5, $[X, Y]^\mu = X^\lambda \partial_\lambda Y^\mu - Y^\lambda \partial_\lambda X^\mu$.

Metrics. A metric pulls back with two Jacobians (T1.92), and the same expansion gives one ∂X term per index.

Key idea: Lie derivatives in components

$$\begin{aligned}\mathcal{L}_X f &= X^\lambda \partial_\lambda f,\\ (\mathcal{L}_X Y)^\mu &= [X, Y]^\mu = X^\lambda \partial_\lambda Y^\mu - Y^\lambda \partial_\lambda X^\mu,\\ (\mathcal{L}_X \alpha)_\mu &= X^\lambda \partial_\lambda \alpha_\mu + \alpha_\lambda\, \partial_\mu X^\lambda,\\ (\mathcal{L}_X g)_{\mu\nu} &= X^\lambda \partial_\lambda g_{\mu\nu} + g_{\lambda\nu}\, \partial_\mu X^\lambda + g_{\mu\lambda}\, \partial_\nu X^\lambda.\end{aligned} \tag{T1.103}$$

Pattern: transport term $X^\lambda \partial_\lambda$, then one correction per index, with a + for each lower index and a − for each upper index. Only first derivatives of X appear, and no metric or connection is needed.

Watch out: The Lie derivative is not a directional derivative

$\mathcal{L}_X Y$ depends on the derivatives of X, not only on its value at the point. Example: $X = \partial_x$ and $X' = x\,\partial_x$ agree at $x = 1$, but $\mathcal{L}_X \partial_x = [\partial_x, \partial_x] = 0$ while $\mathcal{L}_{X'} \partial_x = [x\partial_x, \partial_x] = -\partial_x$. The Lie derivative asks how a field changes *relative to the flow*, and the flow near a point depends on the whole vector field nearby. A derivative that depends only on X at the point needs extra structure, a connection (Chapter T2).

The Lie derivative obeys the Leibniz rule with respect to every product, because φ_t^* respects products and $\mathrm{d}/\mathrm{d}t$ obeys the product rule:

$$\mathcal{L}_X(\alpha \wedge \beta) = \mathcal{L}_X \alpha \wedge \beta + \alpha \wedge \mathcal{L}_X \beta, \qquad \mathcal{L}_X\big(\alpha(Y)\big) = (\mathcal{L}_X \alpha)(Y) + \alpha(\mathcal{L}_X Y). \tag{T1.104}$$

Let us check the second in components, since it ties the three formulae (T1.103) together:

$$\begin{aligned}(\mathcal{L}_X \alpha)_\mu Y^\mu + \alpha_\mu [X, Y]^\mu &\overset{(1)}{=} X^\lambda \partial_\lambda \alpha_\mu\, Y^\mu + \alpha_\lambda\, \partial_\mu X^\lambda\, Y^\mu + \alpha_\mu X^\lambda \partial_\lambda Y^\mu - \alpha_\mu Y^\lambda \partial_\lambda X^\mu\\ &\overset{(2)}{=} X^\lambda \partial_\lambda \alpha_\mu\, Y^\mu + \alpha_\mu X^\lambda \partial_\lambda Y^\mu \overset{(3)}{=} X^\lambda \partial_\lambda (\alpha_\mu Y^\mu) \overset{(4)}{=} \mathcal{L}_X\big(\alpha(Y)\big).\end{aligned}$$

(1) Insert (T1.103).
(2) The second and fourth terms cancel (rename $\lambda \leftrightarrow \mu$ in one of them).
(3) Product rule.
(4) $\alpha(Y)$ is a function, and $\mathcal{L}_X$ of a function is X of it.

The correction terms of $\mathcal{L}_X\alpha$ and $\mathcal{L}_X Y$ are designed to cancel in the scalar $\alpha(Y)$: that is the meaning of "+ for lower, − for upper".

Finally, since pullback commutes with d (T1.93), differentiating $\varphi_t^*\mathrm{d}\omega = \mathrm{d}\varphi_t^*\omega$ at $t = 0$ gives

$$\mathcal{L}_X \mathrm{d} = \mathrm{d}\mathcal{L}_X. \tag{T1.105}$$

T1.9.4 Cartan's magic formula

The component formula for $\mathcal{L}_X\alpha$ on a p-form has $p + 1$ terms. There is a much better way.

Key idea: Cartan's formula

On differential forms,

$$\mathcal{L}_X = \mathrm{d}\,\iota_X + \iota_X\,\mathrm{d}. \tag{T1.106}$$

To Lie-differentiate a form: contract and differentiate, differentiate and contract, add.

Check on 1-forms. With $\iota_X\alpha = X^\nu\alpha_\nu$ and $(\mathrm{d}\alpha)_{\nu\mu} = \partial_\nu\alpha_\mu - \partial_\mu\alpha_\nu$:

$$\begin{aligned}\big(\mathrm{d}\,\iota_X\alpha + \iota_X\mathrm{d}\alpha\big)_\mu &\overset{(1)}{=} \partial_\mu(X^\nu\alpha_\nu) + X^\nu(\partial_\nu\alpha_\mu - \partial_\mu\alpha_\nu)\\ &\overset{(2)}{=} \alpha_\nu\partial_\mu X^\nu + X^\nu\partial_\mu\alpha_\nu + X^\nu\partial_\nu\alpha_\mu - X^\nu\partial_\mu\alpha_\nu\\ &\overset{(3)}{=} X^\nu\partial_\nu\alpha_\mu + \alpha_\nu\partial_\mu X^\nu \overset{(4)}{=} (\mathcal{L}_X\alpha)_\mu.\end{aligned}$$

(1) Components of $\mathrm{d}f$ and of ι_X of a 2-form (first slot).
(2) Product rule on the first term.
(3) The $X^\nu\partial_\mu\alpha_\nu$ terms cancel.
(4) (T1.103).

The cancellation in step (3) is the whole trick: $\mathrm{d}\iota_X$ produces the unwanted term $X^\nu\partial_\mu\alpha_\nu$, and $\iota_X\mathrm{d}$ removes it while supplying the transport term.

Why it holds on all forms. Call the right-hand side $K = \mathrm{d}\iota_X + \iota_X\mathrm{d}$. Three facts:

(i) K agrees with $\mathcal{L}_X$ on functions: $Kf = \iota_X\mathrm{d}f = X(f)$ (the term $\mathrm{d}\iota_X f$ vanishes since $\iota_X f = 0$).

(ii) K commutes with d: $K\mathrm{d} = \mathrm{d}\iota_X\mathrm{d} = \mathrm{d}K$, using $\mathrm{d}^2 = 0$ on both sides. So does $\mathcal{L}_X$ (T1.105).

(iii) K obeys the ordinary Leibniz rule $K(\alpha\wedge\beta) = K\alpha\wedge\beta + \alpha\wedge K\beta$. This is because d and ι_X both pick up $(-1)^p$ past a p-form, so in $\mathrm{d}\iota_X + \iota_X\mathrm{d}$ the cross terms carry opposite signs and cancel. $\mathcal{L}_X$ obeys the same rule (T1.104).

Every form is locally a sum of terms $f\,\mathrm{d}x^{\mu_1}\wedge\cdots\wedge\mathrm{d}x^{\mu_p}$, built from functions by d and $\wedge$. Two operators that agree on functions, commute with d and obey the same Leibniz rule therefore agree on everything. (This is the same uniqueness argument that pinned down d in Section T1.5.)

Worked example T1.51: Rotations preserve area, dilations do not

Take $\mathrm{vol} = \mathrm{d}x \wedge \mathrm{d}y$. Since $\mathrm{d}\,\mathrm{vol} = 0$ (a 3-form on $\mathbb{R}^2$), Cartan gives $\mathcal{L}_X \mathrm{vol} = \mathrm{d}(\iota_X \mathrm{vol})$.
Rotation $X = -y\,\partial_x + x\,\partial_y$:

$$\mathcal{L}_X \mathrm{vol} \overset{(1)}{=} \mathrm{d}\big(X^x\,\mathrm{d}y - X^y\,\mathrm{d}x\big) \overset{(2)}{=} \mathrm{d}\big(-y\,\mathrm{d}y - x\,\mathrm{d}x\big) \overset{(3)}{=} -\mathrm{d}y \wedge \mathrm{d}y - \mathrm{d}x \wedge \mathrm{d}x = 0.$$

(1) $\iota_X(\mathrm{d}x \wedge \mathrm{d}y) = X^x \mathrm{d}y - X^y \mathrm{d}x$ (computed above).
(2) $X^x = -y$, $X^y = x$. Note that this 1-form is $-\mathrm{d}(r^2/2)$, exact.
(3) $\mathrm{d}(y\,\mathrm{d}y) = \mathrm{d}y \wedge \mathrm{d}y = 0$; likewise for x.

Dilation $X = x\,\partial_x + y\,\partial_y$:

$$\mathcal{L}_X \mathrm{vol} \overset{(1)}{=} \mathrm{d}\big(x\,\mathrm{d}y - y\,\mathrm{d}x\big) \overset{(2)}{=} \mathrm{d}x \wedge \mathrm{d}y - \mathrm{d}y \wedge \mathrm{d}x \overset{(3)}{=} 2\,\mathrm{d}x \wedge \mathrm{d}y.$$

(1) Same contraction with $X^x = x$, $X^y = y$.
(2) $\mathrm{d}(x\,\mathrm{d}y) = \mathrm{d}x \wedge \mathrm{d}y$, $\mathrm{d}(y\,\mathrm{d}x) = \mathrm{d}y \wedge \mathrm{d}x$.
(3) Antisymmetry.

Check against the flow. $\varphi_t(x, y) = (e^t x, e^t y)$ pulls $\mathrm{d}x \wedge \mathrm{d}y$ back to $e^{2t}\mathrm{d}x \wedge \mathrm{d}y$ (T1.90), whose t-derivative at $t = 0$ is $2\,\mathrm{d}x \wedge \mathrm{d}y$, in agreement with Cartan. For the rotation $\varphi_t^* \mathrm{vol} = \det(\text{rotation})\,\mathrm{vol} = \mathrm{vol}$ for all t.
The numbers 0 and 2 are the divergences of the two fields. In general, on a Riemannian manifold,

$$\mathcal{L}_X \mathrm{vol} \overset{(1)}{=} \mathrm{d}(\iota_X \mathrm{vol}) \overset{(2)}{=} \mathrm{d} \star X^\flat \overset{(3)}{=} (\mathrm{div}\, X)\,\mathrm{vol}, \qquad \mathrm{div}\, X = \frac{1}{\sqrt{g}}\,\partial_\mu\big(\sqrt{g}\,X^\mu\big). \tag{T1.107}$$

(1) Cartan, with $\mathrm{d}\,\mathrm{vol} = 0$ (top form).
(2) (T1.98).
(3) $\mathrm{d}\star$ of a 1-form is the divergence times vol (Section T1.5 and (T1.52)); in coordinates, $\iota_X(\sqrt{g}\,\mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^n)$ has d equal to $\partial_\mu(\sqrt{g}\,X^\mu)\,\mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^n$.

In polar coordinates, $X = r\,\partial_r$ and $\mathrm{vol} = r\,\mathrm{d}r \wedge \mathrm{d}\phi$ give $\iota_X \mathrm{vol} = r^2 \mathrm{d}\phi$ and $\mathrm{d}(r^2 \mathrm{d}\phi) = 2r\,\mathrm{d}r \wedge \mathrm{d}\phi = 2\,\mathrm{vol}$: the same answer in a different chart, as it must be. *The divergence is the fractional rate at which a flow expands volume.*

Worked example T1.52: Lie derivative of a 1-form, three ways

Let $X = -y\,\partial_x + x\,\partial_y$ (rotation) and $\alpha = x\,\mathrm{d}y$.
(1) From the flow. With $\varphi_t(x, y) = (x\cos t - y\sin t,\ x\sin t + y\cos t)$ from Example T1.50,

$$\varphi_t^*(x\,\mathrm{d}y) \overset{(1)}{=} (x\cos t - y\sin t)\,(\sin t\,\mathrm{d}x + \cos t\,\mathrm{d}y),$$
$$\frac{\mathrm{d}}{\mathrm{d}t}\Big|_0 \varphi_t^* \alpha \overset{(2)}{=} (-y)\,\mathrm{d}y + x\,\mathrm{d}x = x\,\mathrm{d}x - y\,\mathrm{d}y.$$

(1) Substitute the new x in the coefficient and d of the new y in the differential (T1.87).
(2) Product rule at $t = 0$: differentiating the first factor gives $(-x\sin t - y\cos t)|_0 = -y$ times $\mathrm{d}y$; differentiating the second gives x times $(\cos t\,\mathrm{d}x - \sin t\,\mathrm{d}y)|_0 = \mathrm{d}x$.

(2) From the components (T1.103), with $\alpha_x = 0$, $\alpha_y = x$: $(\mathcal{L}_X\alpha)_x = X^\lambda \partial_\lambda \alpha_x + \alpha_y \partial_x X^y = 0 + x \cdot 1 = x$, and $(\mathcal{L}_X\alpha)_y = X^\lambda \partial_\lambda \alpha_y + \alpha_y \partial_y X^y = X^x \cdot 1 + 0 = -y$.
(3) From Cartan:

$$\mathcal{L}_X \alpha \overset{(1)}{=} \mathrm{d}(x\,X^y) + \iota_X(\mathrm{d}x \wedge \mathrm{d}y) \overset{(2)}{=} \mathrm{d}(x^2) + (-y\,\mathrm{d}y - x\,\mathrm{d}x) \overset{(3)}{=} x\,\mathrm{d}x - y\,\mathrm{d}y.$$

(1) $\iota_X(x\,\mathrm{d}y) = x\,X^y$ and $\mathrm{d}(x\,\mathrm{d}y) = \mathrm{d}x \wedge \mathrm{d}y$.
(2) $X^y = x$; the contraction of the area form was computed in Example T1.51.
(3) $\mathrm{d}(x^2) = 2x\,\mathrm{d}x$.

Three methods, one answer. The Cartan route needed neither the flow nor index gymnastics.

T1.9.5 Physics with Cartan's formula

Worked example T1.53: Liouville's theorem

Phase space of one degree of freedom has coordinates (q, p) and the **symplectic form** $\omega = \mathrm{d}q \wedge \mathrm{d}p$. A Hamiltonian $H(q,p)$ generates the flow of Hamilton's equations, with vector field $X_H = \frac{\partial H}{\partial p}\,\partial_q - \frac{\partial H}{\partial q}\,\partial_p$. Then

$$\iota_{X_H}\omega \overset{(1)}{=} X_H^q\,\mathrm{d}p - X_H^p\,\mathrm{d}q \overset{(2)}{=} \frac{\partial H}{\partial p}\,\mathrm{d}p + \frac{\partial H}{\partial q}\,\mathrm{d}q \overset{(3)}{=} \mathrm{d}H,$$

$$\mathcal{L}_{X_H}\omega \overset{(4)}{=} \mathrm{d}(\iota_{X_H}\omega) + \iota_{X_H}\mathrm{d}\omega \overset{(5)}{=} \mathrm{d}\mathrm{d}H + 0 = 0.$$

(1) $\iota_X(\mathrm{d}q \wedge \mathrm{d}p) = \mathrm{d}q(X)\,\mathrm{d}p - \mathrm{d}p(X)\,\mathrm{d}q$.
(2) Components of X_H.
(3) Definition of $\mathrm{d}H$.
(4) Cartan's formula.
(5) $\mathrm{d}^2 = 0$, and $\mathrm{d}\omega = 0$ because ω has constant coefficients.

So Hamiltonian flows preserve phase-space area: **Liouville's theorem**, in two lines. Line (3) is worth remembering on its own: Hamilton's equations *are* $\iota_{X_H}\omega = \mathrm{d}H$. Energy conservation is $\mathcal{L}_{X_H}H = \iota_{X_H}\mathrm{d}H = \iota_{X_H}\iota_{X_H}\omega = 0$, by $\iota_X^2 = 0$. For the free particle $H = p^2/2$, $X_H = p\,\partial_q$ is a shear (Figure T1.29); for the oscillator $H = \frac{1}{2}(p^2+q^2)$, $X_H = p\,\partial_q - q\,\partial_p$ is a rotation. With n degrees of freedom $\omega = \sum_i \mathrm{d}q^i \wedge \mathrm{d}p_i$ and the same two lines show $\mathcal{L}_{X_H}\omega = 0$, hence $\mathcal{L}_{X_H}(\omega^n/n!) = 0$: the phase-space volume $\omega^n/n! = \mathrm{d}q^1 \wedge \mathrm{d}p_1 \wedge \ldots$ is conserved.

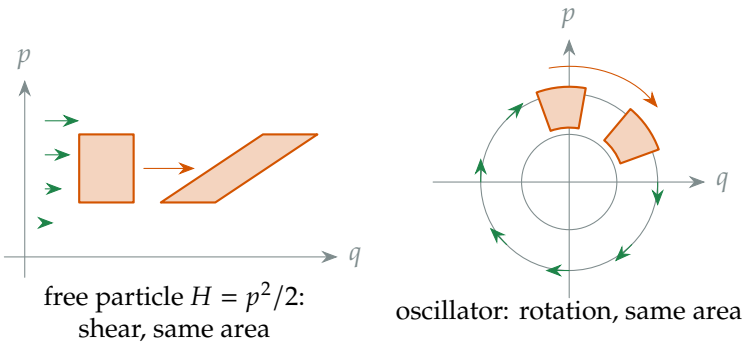


Figure T1.29: Liouville's theorem, $\mathcal{L}_{X_H}(\mathrm{d}q \wedge \mathrm{d}p) = 0$. Left: free motion shears a phase-space cell (fast particles overtake slow ones), but the area of the cell is unchanged. Right: the harmonic oscillator rotates phase space rigidly. In both cases $\iota_{X_H}\omega = \mathrm{d}H$ is exact, so $\mathcal{L}_{X_H}\omega = \mathrm{d}\mathrm{d}H = 0$.

Worked example T1.54: Faraday's flux rule for a moving loop

A wire loop C_t is carried along by a velocity field $\vec{v}$ (with flow φ_t), bounding a moving surface $S_t = \varphi_t(S_0)$. How fast does the magnetic flux $\Phi(t) = \int_{S_t} \mathcal{B}$ change? First, a general *transport rule*: for a time-dependent form ω_t,

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\varphi_t(S)} \omega_t \overset{(1)}{=} \frac{\mathrm{d}}{\mathrm{d}t}\int_S \varphi_t^*\omega_t \overset{(2)}{=} \int_S \varphi_t^*\big(\partial_t\omega_t + \mathcal{L}_v\omega_t\big) \overset{(3)}{=} \int_{\varphi_t(S)} \big(\partial_t\omega_t + \mathcal{L}_v\omega_t\big). \tag{T1.108}$$

(1) Integrals transform by pullback (T1.91).
(2) Product rule in t. The derivative of φ_t^* at time t is $\varphi_t^*\mathcal{L}_v$, because $\varphi_{t+h}^* = \varphi_t^*\varphi_h^*$ by the group law (T1.100) and (T1.94).
(3) Undo step (1).

Now take $\omega_t = \mathcal{B} = \star_3 B$, the magnetic flux 2-form. Then

$$\frac{\mathrm{d}\Phi}{\mathrm{d}t} \overset{(1)}{=} \int_{S_t} \big(\partial_t\mathcal{B} + \mathrm{d}_3\iota_v\mathcal{B} + \iota_v\mathrm{d}_3\mathcal{B}\big) \overset{(2)}{=} \int_{S_t} \big(-\mathrm{d}_3 E + \mathrm{d}_3\iota_v\mathcal{B}\big)$$
$$\overset{(3)}{=} -\oint_{C_t} \big(E - \iota_v\mathcal{B}\big) \overset{(4)}{=} -\oint_{C_t} \big(\vec{E} + \vec{v}\times\vec{B}\big)\cdot\mathrm{d}\vec{r}.$$

(1) Transport rule and Cartan's formula (in space; $\vec{v}$ is spatial).
(2) The metric-free Maxwell equations (T1.61): $\partial_t\mathcal{B} = -\mathrm{d}_3 E$ (Faraday) and $\mathrm{d}_3\mathcal{B} = 0$ (no monopoles).
(3) Stokes' theorem on S_t, whose boundary is C_t (Chapter T2).
(4) $\iota_v\mathcal{B} = (\vec{B}\times\vec{v})\cdot\mathrm{d}\vec{r} = -(\vec{v}\times\vec{B})\cdot\mathrm{d}\vec{r}$ by (T1.97).

The right-hand side is minus the EMF felt by charges riding on the wire: the electric force plus the magnetic Lorentz force. The "flux rule" $\mathcal{E} = -\mathrm{d}\Phi/\mathrm{d}t$, which in textbooks needs a separate argument for the motional part, is one application of Cartan's formula. Notice where each term came from: $\iota_v\mathrm{d}_3\mathcal{B}$ vanished because there are no magnetic monopoles, and $\mathrm{d}_3\iota_v\mathcal{B}$ is the motional EMF.

T1.9.6 Killing vectors: the Lie derivative of the metric

A flow is a *symmetry* of the geometry if it does not change the metric: $\varphi_t^* g = g$. Infinitesimally,

Working definition T1.25 (Killing and conformal Killing vectors)

X is a **Killing vector** if $\mathcal{L}_X g = 0$ [dragging along the flow of X leaves all lengths unchanged, e.g. a rotation of the plane]. It is a **conformal Killing vector** if $\mathcal{L}_X g = C\,g$ for some function C (its flow changes lengths but preserves angles). *Test:* compute $(\mathcal{L}_X g)_{\mu\nu}$ with (T1.103). In Cartesian coordinates on flat space, where $g_{\mu\nu} = \delta_{\mu\nu}$ is constant, this is simply $(\mathcal{L}_X g)_{\mu\nu} = \partial_\mu X_\nu + \partial_\nu X_\mu$.

Worked example T1.55: Killing vectors of the plane and of the sphere

Plane. For the rotation $X = -y\,\partial_x + x\,\partial_y$, $X_x = -y$, $X_y = x$, so $\partial_x X_y + \partial_y X_x = 1 - 1 = 0$ and the diagonal terms $2\partial_x X_x$, $2\partial_y X_y$ vanish: Killing. For the dilation $X = x\,\partial_x + y\,\partial_y$, $(\mathcal{L}_X g)_{\mu\nu} = 2\delta_{\mu\nu}$, so $\mathcal{L}_X g = 2g$: conformal Killing with $C = 2$ (scaling changes lengths uniformly and keeps angles). Translations are Killing trivially. The plane has three Killing vectors (∂_x, ∂_y, rotation) and, as we will see in Section T1.10, infinitely many conformal

Killing vectors.
Sphere. Take $g = \mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2$. For $X = \partial_\phi$ the components are constant and g does not depend on ϕ, so every term in (T1.103) vanishes: Killing. (Dray's slogan: a coordinate the metric does not depend on gives a Killing vector [Dray §A.1].) A less obvious one is the rotation about the X-axis of $\mathbb{R}^3$, the field $K = Z\partial_Y - Y\partial_Z$ (to avoid a clash with the Cartesian coordinate X, we call the vector field K here). On the sphere its components are $K^\theta = \sin\phi$, $K^\phi = \cot\theta\cos\phi$. To see this, differentiate $Z = \cos\theta$ and $\tan\phi = Y/X$ along the flow $\dot X = 0$, $\dot Y = Z$, $\dot Z = -Y$: $-\sin\theta\,\dot\theta = -\sin\theta\sin\phi$ gives $\dot\theta = \sin\phi$, and $\sec^2\phi\,\dot\phi = Z/X = \cos\theta/(\sin\theta\cos\phi)$ gives $\dot\phi = \cot\theta\cos\phi$. Then

$$(\mathcal{L}_K g)_{\theta\theta} \overset{(1)}{=} K^\lambda\partial_\lambda g_{\theta\theta} + 2g_{\theta\theta}\,\partial_\theta K^\theta = 0 + 2\,\partial_\theta(\sin\phi) = 0,$$

$$(\mathcal{L}_K g)_{\theta\phi} \overset{(2)}{=} g_{\phi\phi}\,\partial_\theta K^\phi + g_{\theta\theta}\,\partial_\phi K^\theta = \sin^2\theta\,\big(-\csc^2\theta\cos\phi\big) + \cos\phi = 0,$$

$$(\mathcal{L}_K g)_{\phi\phi} \overset{(3)}{=} K^\theta\partial_\theta g_{\phi\phi} + 2g_{\phi\phi}\,\partial_\phi K^\phi = \sin\phi\cdot 2\sin\theta\cos\theta + 2\sin^2\theta\,(-\cot\theta\sin\phi) = 0.$$

(1) $g_{\theta\theta} = 1$ is constant and $g_{\phi\theta} = 0$.
(2) $g_{\theta\phi} = 0$, so the transport term is absent; one correction term per index; $\partial_\theta\cot\theta = -\csc^2\theta$.
(3) $g_{\phi\phi} = \sin^2\theta$ depends on θ; $\partial_\phi(\cot\theta\cos\phi) = -\cot\theta\sin\phi$.

Each component cancels in a different way: an isometry of the round sphere that no coordinate makes obvious. The sphere has three Killing vectors, the rotations $\mathfrak{so}(3)$.

Physical picture: Killing vectors give conserved quantities

For a free particle moving on a curved space (a geodesic), the quantity $g(X,\dot x) = g_{\mu\nu}X^\mu\dot x^\nu$ is conserved whenever X is Killing. On the sphere with $X = \partial_\phi$ this is $\sin^2\theta\,\dot\phi$: the z-component of angular momentum. This is Noether's theorem in its simplest geometric form, and it is how Dray finds the conserved energy and angular momentum of orbits in Schwarzschild space-time [Dray §2.3, §A.1].

T1.9.7 Flows do not change cohomology

Here is the promise made in Section T1.8. Let ω be a closed form, $\mathrm{d}\omega = 0$, and φ_t the flow of X. Then

$$\varphi_t^*\omega - \omega \overset{(1)}{=} \int_0^t \frac{\mathrm{d}}{\mathrm{d}s}\varphi_s^*\omega\,\mathrm{d}s \overset{(2)}{=} \int_0^t \varphi_s^*\mathcal{L}_X\omega\,\mathrm{d}s \overset{(3)}{=} \int_0^t \varphi_s^*\mathrm{d}\,\iota_X\omega\,\mathrm{d}s \overset{(4)}{=} \mathrm{d}\Big(\int_0^t \varphi_s^*\iota_X\omega\,\mathrm{d}s\Big). \tag{T1.109}$$

(1) Fundamental theorem of calculus, since $\varphi_0^*\omega = \omega$.
(2) Derivative of φ_s^*, as in step (2) of (T1.108).
(3) Cartan's formula with $\mathrm{d}\omega = 0$.
(4) φ_s^* commutes with d (T1.93), and so does integration over the parameter s.

So $\varphi_t^*\omega$ and ω differ by an exact form: *sliding a closed form along any flow does not change its de Rham class*, and integrals of closed forms over closed cycles are unchanged. The same computation, done on the cylinder $[0,1]\times N$ with $X = \partial_t$, shows that any two homotopic maps $f_0, f_1 : N \to M$ satisfy $f_1^*\omega - f_0^*\omega = \mathrm{d}(K\omega)$ with $K\omega = \int_0^1 \iota_{\partial_t}F^*\omega\,\mathrm{d}t$, where F is the homotopy. This **homotopy formula** is behind the homotopy invariance of de Rham cohomology and of characteristic classes (Chapter T2, Chapter 1).

Try it: Is the angle form invariant?

With $\beta = \mathrm{d}\phi = (x\,\mathrm{d}y - y\,\mathrm{d}x)/r^2$, compute $\mathcal{L}_X\beta$ for the rotation $X = \partial_\phi$ and for the dilation $X = r\partial_r$ using Cartan.
Answer: $\mathrm{d}\beta = 0$, so $\mathcal{L}_X\beta = \mathrm{d}(\iota_X\beta)$. For the rotation $\iota_{\partial_\phi}\mathrm{d}\phi = 1$, and $\mathrm{d}1 = 0$. For the dilation $\iota_{r\partial_r}\mathrm{d}\phi = 0$. Both vanish: β is invariant under rotations and scalings, as the winding number (which it computes) should be.

T1.9.8 The Cartan calculus, collected

Dictionary: the six identities of the Cartan calculus

$\mathrm{d}^2 = 0, \quad \iota_X^2 = 0$	both odd operators square to zero
$\mathcal{L}_X = \mathrm{d}\iota_X + \iota_X\mathrm{d}$	Cartan's formula (T1.106)
$\mathcal{L}_X\mathrm{d} = \mathrm{d}\mathcal{L}_X$	flows commute with d (T1.105)
$\mathcal{L}_X\iota_Y - \iota_Y\mathcal{L}_X = \iota_{[X,Y]}$	contraction with a dragged vector
$\mathcal{L}_X\mathcal{L}_Y - \mathcal{L}_Y\mathcal{L}_X = \mathcal{L}_{[X,Y]}$	the flows represent the Lie algebra of vector fields
$(\mathrm{d} + \iota_X)^2 = \mathcal{L}_X$	equivariant d: squares to zero on X-invariant forms

The fourth identity follows from the Leibniz rule (T1.104) applied to $\alpha(Y, \dots)$. The fifth, on functions, is the definition of the commutator: $[\mathcal{L}_X, \mathcal{L}_Y]f = X(Yf) - Y(Xf) = [X, Y]f$. The last follows by squaring: $(\mathrm{d} + \iota_X)^2 = \mathrm{d}^2 + \mathrm{d}\iota_X + \iota_X\mathrm{d} + \iota_X^2 = \mathcal{L}_X$. It is the starting point of *equivariant cohomology* and of the localisation formulae used in topological field theory (Chapter 12): on forms that are invariant under the flow, $\mathrm{d}_X = \mathrm{d} + \iota_X$ is a new "differential" that squares to zero.

Watch out: Sign and slot conventions

We contract in the *first* slot. Some authors contract in the last slot, which changes ι_X by $(-1)^{p-1}$ on p-forms and moves signs around in (T1.96); Cartan's formula is the same in both conventions if the Leibniz rule is adjusted consistently. The commutator sign $[X, Y] = XY - YX$ is universal; but some books define $\mathcal{L}_X Y = -[X, Y]$ by pulling vectors with $(\varphi_t)_*$ instead of $(\varphi_{-t})_*$. Nash's footnote formula (5.71) agrees with ours.

Reading Nash (pp. 159–160, (5.66)–(5.73) and footnote 5; p. 166)

Nash needs the Lie derivative exactly once, but at a crucial point: the count of moduli of Riemann surfaces. A vector field V on the surface Σ exponentiates to a one-parameter family of diffeomorphisms $\alpha_t = \exp[tV]$ (his (5.66), our (T1.101)), which acts on metrics by $\alpha_t \cdot g$, and $g(t) = g(0) + t\mathcal{L}_V g(0) + \dots$ (5.68). So the metric perturbations that are pure diffeomorphisms are $\delta g_{ab} = \mathcal{L}_V g_{ab}$ (5.70), with the component formula (5.71),

$$\mathcal{L}_V g_{ab} = \frac{\partial g_{ab}}{\partial x^i} V^i + g_{ib}\frac{\partial V^i}{\partial x^a} + g_{ai}\frac{\partial V^i}{\partial x^b},$$

which is our (T1.103). In complex coordinates this gives $\delta\mu = 2\partial_{\bar z} V^z$ (5.72), which we derive in Example T1.61. His footnote 5 defines a conformal Killing vector by $\mathcal{L}_K g = Cg$, our working definition above; on p. 166 he uses the fact that the conformal Killing vectors of a torus form a 1-dimensional space, and (footnote 7, p. 165) that the sphere has a three-complex-parameter family, the Möbius transformations. Both counts are done explicitly in Example T1.60.

Recap

- ι_X inserts X in the first slot: $\iota_X \mathrm{d}x^\mu = X^\mu$; it is an antiderivation with $\iota_X^2 = 0$. In 3D: ι_X of a 1-form is a dot product, of a flux form a cross product, of a density a current. $\iota_X \mathrm{vol} = \star X^\flat$.
- The flow $\varphi_t = \exp(tX)$ solves $\dot{x} = X(x)$; the Lie derivative $\mathcal{L}_X T = \frac{\mathrm{d}}{\mathrm{d}t}\varphi_t^* T|_0$ measures change relative to the flow. Components: transport term plus one ∂X correction per index; $\mathcal{L}_X Y = [X, Y]$.
- Cartan: $\mathcal{L}_X = \mathrm{d}\iota_X + \iota_X \mathrm{d}$ on forms. Hence $\mathcal{L}_X \mathrm{vol} = (\operatorname{div} X)\mathrm{vol}$, Liouville's theorem, the flux rule, and invariance of cohomology under flows.
- Killing vectors: $\mathcal{L}_X g = 0$ (symmetries, conserved quantities); conformal Killing: $\mathcal{L}_X g = Cg$. These are the objects Nash counts in the moduli problem of Chapter 5.

T1.10 Complex coordinates and the Dolbeault operator

In two dimensions a physicist very often trades (x, y) for $z = x + \mathrm{i}y$ and $\bar{z} = x - \mathrm{i}y$. String theory and conformal field theory are written this way (Nash's Chapters 9 and 11), and so is the algebraic geometry of Riemann surfaces (Chapter 5). In Chapters 3 and 4 Nash uses the *Dolbeault complex*, the complex-analytic twin of the de Rham complex, as one of his four model elliptic complexes. All of it rests on one observation, which this section develops: in complex coordinates the exterior derivative splits into two halves,

$$\mathrm{d} = \partial + \bar{\partial},$$

and the half $\bar{\partial}$ is a machine for detecting holomorphic functions. The Cauchy–Riemann equations become the single equation $\bar{\partial} f = 0$, Cauchy's theorem becomes Stokes' theorem, and the Dolbeault cohomology groups measure how badly $\bar{\partial} f = g$ can fail to have global solutions.

T1.10.1 Complex-valued forms, $\mathrm{d}z$ and $\mathrm{d}\bar{z}$

Nothing we have done so far requires forms to have *real* coefficients. A **complex-valued form** is $\alpha + \mathrm{i}\beta$ with α, β real forms; $\wedge$, d, pullback and ι_X extend by linearity over $\mathbb{C}$. (Nash does exactly this on p. 103: he takes the de Rham complex "complex valued instead of real valued" so that it can be fed to the Chern character.) On $\mathbb{R}^2 = \mathbb{C}$ define

$$\mathrm{d}z = \mathrm{d}x + \mathrm{i}\,\mathrm{d}y, \qquad \mathrm{d}\bar{z} = \mathrm{d}x - \mathrm{i}\,\mathrm{d}y, \qquad \text{so} \qquad \mathrm{d}x = \tfrac{1}{2}(\mathrm{d}z + \mathrm{d}\bar{z}), \quad \mathrm{d}y = \tfrac{1}{2\mathrm{i}}(\mathrm{d}z - \mathrm{d}\bar{z}). \tag{T1.110}$$

These two complex 1-forms are a basis of complex 1-forms at each point, just as $\mathrm{d}x, \mathrm{d}y$ are.

The dual basis of (complex) vectors is fixed by $\mathrm{d}z(\partial_z) = 1$, $\mathrm{d}\bar{z}(\partial_z) = 0$, $\mathrm{d}z(\partial_{\bar{z}}) = 0$, $\mathrm{d}\bar{z}(\partial_{\bar{z}}) = 1$:

$$\partial_z = \tfrac{1}{2}\big(\partial_x - \mathrm{i}\,\partial_y\big), \qquad \partial_{\bar{z}} = \tfrac{1}{2}\big(\partial_x + \mathrm{i}\,\partial_y\big). \tag{T1.111}$$

Check one of the four conditions:

$$\mathrm{d}z(\partial_z) \overset{(1)}{=} (\mathrm{d}x + \mathrm{i}\,\mathrm{d}y)\big(\tfrac{1}{2}\partial_x - \tfrac{\mathrm{i}}{2}\partial_y\big) \overset{(2)}{=} \tfrac{1}{2} + \mathrm{i}\cdot\big(-\tfrac{\mathrm{i}}{2}\big) \overset{(3)}{=} 1.$$

(1) Definitions (T1.110) and (T1.111).
(2) Linearity, with $\mathrm{d}x(\partial_x) = \mathrm{d}y(\partial_y) = 1$ and the cross terms zero.
(3) $-\mathrm{i}^2 = 1$.

The same arithmetic gives $\mathrm{d}\bar{z}(\partial_z) = \frac{1}{2} + (-\mathrm{i})(-\frac{\mathrm{i}}{2}) = \frac{1}{2} - \frac{1}{2} = 0$. Consequently $\partial_z z = 1$, $\partial_z \bar{z} = 0$, and so on.

Watch out: Are z and $\bar{z}$ independent?

They are not independent as numbers: $\bar{z}$ is determined by z. But $\mathrm{d}z$ and $\mathrm{d}\bar{z}$ are linearly independent 1-forms, and ∂_z, $\partial_{\bar{z}}$ are independent (complex) vectors. Any smooth function of (x, y) can be written as a function of $(z, \bar{z})$ by substituting $x = (z + \bar{z})/2$, $y = (z - \bar{z})/2\mathrm{i}$, and then ∂_z, $\partial_{\bar{z}}$ act on it by the usual rules *as if* $z, \bar{z}$ were independent variables. Example: $\partial_z(z^2\bar{z}) = 2z\bar{z}$, $\partial_{\bar{z}}(z^2\bar{z}) = z^2$. This is a statement about the chain rule, not about the numbers.

Now expand the differential of a function in the new basis:

$$\mathrm{d}f \overset{(1)}{=} \partial_x f\,\mathrm{d}x + \partial_y f\,\mathrm{d}y \overset{(2)}{=} \tfrac{1}{2}(\partial_x - \mathrm{i}\partial_y)f\,\mathrm{d}z + \tfrac{1}{2}(\partial_x + \mathrm{i}\partial_y)f\,\mathrm{d}\bar{z} \overset{(3)}{=} \partial_z f\,\mathrm{d}z + \partial_{\bar{z}} f\,\mathrm{d}\bar{z}. \tag{T1.112}$$

(1) Definition of $\mathrm{d}f$.
(2) Substitute $\mathrm{d}x, \mathrm{d}y$ from (T1.110) and collect the $\mathrm{d}z$ and $\mathrm{d}\bar{z}$ terms: $\frac{1}{2}\partial_x f + \frac{1}{2\mathrm{i}}\partial_y f = \frac{1}{2}(\partial_x - \mathrm{i}\partial_y)f$, and similarly for $\mathrm{d}\bar{z}$.
(3) (T1.111).

We name the two halves:

$$\partial f = \partial_z f\,\mathrm{d}z, \qquad \bar{\partial} f = \partial_{\bar{z}} f\,\mathrm{d}\bar{z}, \qquad \mathrm{d}f = \partial f + \bar{\partial} f. \tag{T1.113}$$

Key idea: Holomorphic means $\bar{\partial} f = 0$

Write $f = u + \mathrm{i}v$. Then

$$2\,\partial_{\bar{z}} f \overset{(1)}{=} (\partial_x + \mathrm{i}\partial_y)(u + \mathrm{i}v) \overset{(2)}{=} (u_x - v_y) + \mathrm{i}\,(v_x + u_y).$$

(1) (T1.111).
(2) Multiply out; $\mathrm{i}^2 = -1$ turns $\mathrm{i}\partial_y(\mathrm{i}v)$ into $-v_y$.

So $\bar{\partial} f = 0$ is exactly the pair of Cauchy–Riemann equations $u_x = v_y$, $u_y = -v_x$: *f is holomorphic iff it does not depend on $\bar{z}$*. For holomorphic f, $\mathrm{d}f = f'(z)\,\mathrm{d}z$, and $\partial_z f$ is the complex derivative f'. This is Nash's remark after (3.25): "the familiar Cauchy–Riemann equations for a holomorphic function f are, of course, simply the statement $\bar{\partial} f = 0$".

Worked example T1.56: Four functions

(a) $f = z^2$: $\partial_{\bar{z}} f = 0$, holomorphic, $\mathrm{d}f = 2z\,\mathrm{d}z$. In real terms $z^2 = (x^2 - y^2) + 2\mathrm{i}xy$ and $\mathrm{d}(x^2 - y^2) + 2\mathrm{i}\,\mathrm{d}(xy) = 2(x + \mathrm{i}y)(\mathrm{d}x + \mathrm{i}\mathrm{d}y)$: check.
(b) $f = \bar{z}$: $\bar{\partial} f = \mathrm{d}\bar{z} \neq 0$. Complex conjugation is the basic non-holomorphic function (it reverses orientation).
(c) $f = |z|^2 = z\bar{z}$: $\mathrm{d}f = \bar{z}\,\mathrm{d}z + z\,\mathrm{d}\bar{z}$. Real, and not holomorphic; $\bar{\partial} f = z\,\mathrm{d}\bar{z}$.
(d) $f = e^z = e^x(\cos y + \mathrm{i}\sin y)$: $u_x = e^x \cos y = v_y$ and $u_y = -e^x \sin y = -v_x$, so it is holomorphic, and $\mathrm{d}f = e^z \mathrm{d}z$.

The Laplacian factorises through ∂_z and $\partial_{\bar{z}}$:

$$4\,\partial_z\partial_{\bar{z}} = (\partial_x - \mathrm{i}\partial_y)(\partial_x + \mathrm{i}\partial_y) = \partial_x^2 + \partial_y^2 = \nabla^2, \qquad \text{so} \qquad \Delta = -4\,\partial_z\partial_{\bar{z}} \tag{T1.114}$$

with our positive Δ of Section T1.6 (the cross terms $\pm\mathrm{i}\partial_x\partial_y$ cancel because partial derivatives commute). Hence *every holomorphic function is harmonic*: $\Delta f = -4\partial_z(\partial_{\bar{z}} f) = 0$, and so are its real and imaginary parts. This is why $\mathrm{Re}\, z^m = r^m \cos m\phi$ solved Laplace's equation in Example T1.32.

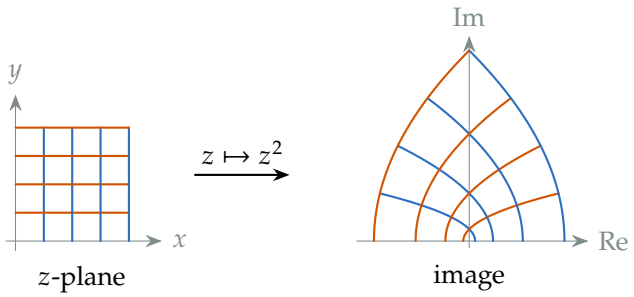


Figure T1.30: A holomorphic map preserves angles. The grid lines $x = \text{const}$ (blue) and $y = \text{const}$ (orange) in the first quadrant are carried by $z \mapsto z^2$ to two families of parabolas, $x' = a^2 - y^2$, $y' = 2ay$ and $x' = x^2 - b^2$, $y' = 2bx$, which still cross at right angles. Infinitesimally a holomorphic map is multiplication by the complex number $f'(z)$: a rotation times a scaling. This is why complex coordinates are the natural language of two-dimensional conformal geometry.

T1.10.2 Area, metric and Hodge star in complex coordinates

The area form and the metric take simple forms:

$$\mathrm{d}z \wedge \mathrm{d}\bar{z} \overset{(1)}{=} (\mathrm{d}x + \mathrm{i}\,\mathrm{d}y) \wedge (\mathrm{d}x - \mathrm{i}\,\mathrm{d}y) \overset{(2)}{=} -\mathrm{i}\,\mathrm{d}x \wedge \mathrm{d}y + \mathrm{i}\,\mathrm{d}y \wedge \mathrm{d}x \overset{(3)}{=} -2\mathrm{i}\,\mathrm{d}x \wedge \mathrm{d}y,$$
$$\mathrm{d}x \wedge \mathrm{d}y = \tfrac{\mathrm{i}}{2}\,\mathrm{d}z \wedge \mathrm{d}\bar{z}, \qquad \mathrm{d}s^2 = \mathrm{d}x^2 + \mathrm{d}y^2 = \mathrm{d}z\,\mathrm{d}\bar{z}. \tag{T1.115}$$

(1) (T1.110).
(2) $\mathrm{d}x \wedge \mathrm{d}x = \mathrm{d}y \wedge \mathrm{d}y = 0$; the cross terms are $\mathrm{d}x \wedge (-\mathrm{i}\mathrm{d}y)$ and $\mathrm{i}\,\mathrm{d}y \wedge \mathrm{d}x$.
(3) $\mathrm{d}y \wedge \mathrm{d}x = -\mathrm{d}x \wedge \mathrm{d}y$.

In the metric, $\mathrm{d}z\,\mathrm{d}\bar{z}$ is the *symmetric* product $\frac{1}{2}(\mathrm{d}z \otimes \mathrm{d}\bar{z} + \mathrm{d}\bar{z} \otimes \mathrm{d}z)$: $(\mathrm{d}x + \mathrm{i}\mathrm{d}y)(\mathrm{d}x - \mathrm{i}\mathrm{d}y) = \mathrm{d}x^2 + \mathrm{d}y^2$, the cross terms cancelling. So in the basis $(\mathrm{d}z, \mathrm{d}\bar{z})$ the metric has components $g_{zz} = g_{\bar{z}\bar{z}} = 0$, $g_{z\bar{z}} = g_{\bar{z}z} = \frac{1}{2}$, and the inverse metric has $g^{z\bar{z}} = 2$. A general metric on a surface can always be brought locally to the **conformal** (or isothermal) form $\mathrm{d}s^2 = \lambda\,\mathrm{d}z\,\mathrm{d}\bar{z}$ (Nash, p. 158).

The Hodge star on 1-forms, from the $\mathbb{R}^2$ table ($\star\mathrm{d}x = \mathrm{d}y$, $\star\mathrm{d}y = -\mathrm{d}x$, Example T1.20):

$$\star\mathrm{d}z \overset{(1)}{=} \star\mathrm{d}x + \mathrm{i} \star \mathrm{d}y \overset{(2)}{=} \mathrm{d}y - \mathrm{i}\,\mathrm{d}x \overset{(3)}{=} -\mathrm{i}\,(\mathrm{d}x + \mathrm{i}\,\mathrm{d}y) = -\mathrm{i}\,\mathrm{d}z, \qquad \star\mathrm{d}\bar{z} = +\mathrm{i}\,\mathrm{d}\bar{z}. \tag{T1.116}$$

(1) $\star$ is linear (over $\mathbb{C}$, for complex-valued forms).
(2) The $\mathbb{R}^2$ table.
(3) Factor out $-\mathrm{i}$: $-\mathrm{i} \cdot \mathrm{i}\,\mathrm{d}y = \mathrm{d}y$.

So $\mathrm{d}z$ and $\mathrm{d}\bar{z}$ are the *eigenvectors* of $\star$ on 1-forms, with eigenvalues $\mp\mathrm{i}$ (consistent with $\star\star = -1$ on 1-forms in two dimensions, (T1.35)). A 1-form $f\,\mathrm{d}z$ is called a $(1,0)$-form, and $f\,\mathrm{d}\bar{z}$ a $(0,1)$-form. Since $\star$ on 1-forms in two dimensions is conformally invariant (Section T1.4), this splitting depends only on the conformal class of the metric, i.e. on the *complex structure*. It is the two-dimensional instance of the decomposition by type that we now generalise.

Worked example T1.57: Cauchy's theorem is Stokes' theorem

Let f be any smooth complex function on a region D. Then

$$\mathrm{d}(f\,\mathrm{d}z) \overset{(1)}{=} \mathrm{d}f \wedge \mathrm{d}z \overset{(2)}{=} \partial_{\bar{z}} f\,\mathrm{d}\bar{z} \wedge \mathrm{d}z \overset{(3)}{=} 2\mathrm{i}\,\partial_{\bar{z}} f\,\mathrm{d}x \wedge \mathrm{d}y.$$

(1) Leibniz rule; $\mathrm{d}(\mathrm{d}z) = \mathrm{d}\mathrm{d}x + \mathrm{i}\,\mathrm{d}\mathrm{d}y = 0$.
(2) (T1.112); the $\mathrm{d}z \wedge \mathrm{d}z$ term vanishes.
(3) $\mathrm{d}\bar{z} \wedge \mathrm{d}z = -\mathrm{d}z \wedge \mathrm{d}\bar{z} = 2\mathrm{i}\,\mathrm{d}x \wedge \mathrm{d}y$ (T1.115).

So $f\,\mathrm{d}z$ is a closed 1-form exactly when f is holomorphic, and Stokes' theorem gives

$$\oint_{\partial D} f\,\mathrm{d}z = 2\mathrm{i}\int_D \partial_{\bar z} f\;\mathrm{d}x\,\mathrm{d}y, \tag{T1.117}$$

which vanishes for holomorphic f: **Cauchy's theorem**. The general formula (T1.117) measures the failure of holomorphy inside the loop, just as $\oint \vec{A}\cdot\mathrm{d}\vec{r}$ measures the curl inside a loop.

Worked example T1.58: $\bar\partial(1/z) = \pi\,\delta^2$ and the Cauchy integral formula

The function $1/z$ is holomorphic on $\mathbb{C}\setminus\{0\}$, so $\partial_{\bar z}(1/z) = 0$ there. What happens at 0? Test it against a smooth function φ of compact support, as one does with distributions: by definition $\int \partial_{\bar z}(1/z)\,\varphi\,\mathrm{d}^2x = -\int (1/z)\,\partial_{\bar z}\varphi\,\mathrm{d}^2x$. Remove a small disc $|z|<\varepsilon$ (its contribution is $O(\varepsilon)$, since $1/|z|$ is integrable in two dimensions):

$$-\int_{|z|>\varepsilon}\frac{\partial_{\bar z}\varphi}{z}\,\mathrm{d}^2x \overset{(1)}{=} -\int_{|z|>\varepsilon}\partial_{\bar z}\Big(\frac{\varphi}{z}\Big)\mathrm{d}^2x \overset{(2)}{=} -\frac{1}{2\mathrm{i}}\int_{|z|>\varepsilon}\mathrm{d}\Big(\frac{\varphi\,\mathrm{d}z}{z}\Big)$$
$$\overset{(3)}{=} +\frac{1}{2\mathrm{i}}\oint_{|z|=\varepsilon}\frac{\varphi\,\mathrm{d}z}{z} \overset{(4)}{=} \frac{1}{2\mathrm{i}}\int_0^{2\pi}\varphi(\varepsilon e^{\mathrm{i}\phi})\,\mathrm{i}\,\mathrm{d}\phi \xrightarrow{\varepsilon\to 0} \pi\,\varphi(0).$$

(1) $1/z$ is holomorphic on the region, so $\partial_{\bar z}$ passes through it.
(2) Example T1.57: $\mathrm{d}(g\,\mathrm{d}z) = 2\mathrm{i}\,\partial_{\bar z} g\,\mathrm{d}^2x$.
(3) Stokes. The region $|z|>\varepsilon$ has as its boundary the small circle traversed *clockwise* (φ vanishes far away), so $\int \mathrm{d}(\cdots) = -\oint_{\mathrm{ccw}}(\cdots)$.
(4) $z=\varepsilon e^{\mathrm{i}\phi}$, $\mathrm{d}z/z = \mathrm{i}\,\mathrm{d}\phi$ (this is the angle form again, Example T1.30).

Hence, as distributions,

$$\partial_{\bar z}\frac{1}{z} = \pi\,\delta^2(x,y), \qquad \partial_z\partial_{\bar z}\log|z|^2 = \pi\,\delta^2(x,y). \tag{T1.118}$$

The second follows from the first because $\partial_z \log|z|^2 = \partial_z(\log z + \log\bar z) = 1/z$. Two consequences. (i) With $\Delta = -4\partial_z\partial_{\bar z}$, the function $G = -\frac{1}{4\pi}\log|z|^2 = -\frac{1}{2\pi}\log r$ solves $\Delta G = \delta^2$: the Green function of the 2D Laplacian (the potential of a line charge). (ii) For f holomorphic on a disc D containing a, apply (T1.117) to $f(z)/(z-a)$: $\oint_{\partial D}\frac{f\,\mathrm{d}z}{z-a} = 2\mathrm{i}\int_D f\,\pi\delta^2(z-a)\,\mathrm{d}^2x = 2\pi\mathrm{i}\,f(a)$, the **Cauchy integral formula**. Similarly $\frac{1}{\pi z}$ is the Green function of $\partial_{\bar z}$: the equation $\partial_{\bar z} f = g$ is solved on $\mathbb{C}$ by

$$f(z) = \frac{1}{\pi}\int_{\mathbb{C}}\frac{g(w)}{z-w}\,\mathrm{d}^2w, \tag{T1.119}$$

for any smooth g of compact support. This is the $\bar\partial$-Poincaré lemma: *locally*, every $(0,1)$-form $g\,\mathrm{d}\bar z$ is $\bar\partial$ of something.

T1.10.3 Several complex variables: (p,q)-forms

On $\mathbb{C}^m$, with complex coordinates $z^j = x^j + \mathrm{i}y^j$ ($j=1,\dots,m$; real dimension $2m$), the complex 1-forms have the basis $\mathrm{d}z^1,\dots,\mathrm{d}z^m,\mathrm{d}\bar z^1,\dots,\mathrm{d}\bar z^m$.

Working definition T1.26 (forms of type (p,q))

A (p,q)**-form** is a sum of terms with p factors $\mathrm{d}z$ and q factors $\mathrm{d}\bar{z}$,

$$\omega = \frac{1}{p!\,q!}\,\omega_{i_1\dots i_p \bar{\jmath}_1\dots\bar{\jmath}_q}\,\mathrm{d}z^{i_1}\wedge\cdots\wedge\mathrm{d}z^{i_p}\wedge\mathrm{d}\bar{z}^{j_1}\wedge\cdots\wedge\mathrm{d}\bar{z}^{j_q} \tag{T1.120}$$

(Nash's (3.24), where as usual he omits the factorials). *Test:* rewrite the form in the basis $\mathrm{d}z^j, \mathrm{d}\bar{z}^j$ and count. Every complex k-form splits uniquely into types, $\Omega^k = \bigoplus_{p+q=k}\Omega^{p,q}$ (Nash's (3.27)) [e.g. on $\mathbb{C}$, $a\,\mathrm{d}x + b\,\mathrm{d}y = \frac{1}{2}(a-\mathrm{i}b)\,\mathrm{d}z + \frac{1}{2}(a+\mathrm{i}b)\,\mathrm{d}\bar{z}$].

Dimension count: at each point, choosing p of the m differentials $\mathrm{d}z^j$ and q of the m differentials $\mathrm{d}\bar{z}^j$ gives

$$\dim\Omega^{p,q}_x = \binom{m}{p}\binom{m}{q}, \qquad \sum_{p+q=k}\binom{m}{p}\binom{m}{q} = \binom{2m}{k}, \tag{T1.121}$$

and the second identity (Vandermonde's) says that the types add up to all k-forms on the real $2m$-dimensional space, as counted in (T1.21). For $\mathbb{C}^2$ ($m=2$) the numbers are arranged in Figure T1.31.

Working definition T1.27 (∂ and $\bar{\partial}$)

On a (p,q)-form $\omega = f\,\mathrm{d}z^I\wedge\mathrm{d}\bar{z}^J$ [$\mathrm{d}z^I = \mathrm{d}z^{i_1}\wedge\cdots\wedge\mathrm{d}z^{i_p}$ for the index list $I=(i_1,\dots,i_p)$, and likewise $\mathrm{d}\bar{z}^J$],

$$\partial\omega = \sum_{j=1}^{m}\frac{\partial f}{\partial z^j}\,\mathrm{d}z^j\wedge\mathrm{d}z^I\wedge\mathrm{d}\bar{z}^J, \qquad \bar{\partial}\omega = \sum_{j=1}^{m}\frac{\partial f}{\partial \bar{z}^j}\,\mathrm{d}\bar{z}^j\wedge\mathrm{d}z^I\wedge\mathrm{d}\bar{z}^J \tag{T1.122}$$

(Nash's (3.25), written as operators $\partial = \sum_j \mathrm{d}z^j\,\partial/\partial z^j$ and $\bar{\partial} = \sum_j \mathrm{d}\bar{z}^j\,\partial/\partial\bar{z}^j$) [$\partial/\partial z^j = \frac{1}{2}(\partial_{x^j} - \mathrm{i}\partial_{y^j})$ and $\partial/\partial\bar{z}^j = \frac{1}{2}(\partial_{x^j}+\mathrm{i}\partial_{y^j})$]. So $\partial:\Omega^{p,q}\to\Omega^{p+1,q}$ and $\bar{\partial}:\Omega^{p,q}\to\Omega^{p,q+1}$.

Key idea: The relations (3.26)

$$\mathrm{d} = \partial + \bar{\partial}, \qquad \partial^2 = 0, \qquad \bar{\partial}^2 = 0, \qquad \partial\bar{\partial} + \bar{\partial}\partial = 0. \tag{T1.123}$$

Why. The first is (T1.112) applied to each coefficient, $\mathrm{d}f = \sum_j(\partial_{z^j}f\,\mathrm{d}z^j + \partial_{\bar{z}^j}f\,\mathrm{d}\bar{z}^j)$, together with $\mathrm{d}(\mathrm{d}z^j) = \mathrm{d}(\mathrm{d}\bar{z}^j) = 0$. For the others, square it:

$$0 \overset{(1)}{=} \mathrm{d}^2 \overset{(2)}{=} \underbrace{\partial^2}_{(p+2,q)} + \underbrace{(\partial\bar{\partial}+\bar{\partial}\partial)}_{(p+1,q+1)} + \underbrace{\bar{\partial}^2}_{(p,q+2)}.$$

(1) $\mathrm{d}^2 = 0$ (T1.45).
(2) Expand $(\partial+\bar{\partial})^2$; each term changes the type by a definite amount, shown underneath.

A form is zero only if each of its type components is zero, so the three pieces vanish separately. Let us see the middle one at work on a function on $\mathbb{C}^2$:

$$\bar{\partial}\partial f + \partial\bar{\partial}f \overset{(1)}{=} \bar{\partial}\big(\partial_j f\,\mathrm{d}z^j\big) + \partial\big(\partial_{\bar{k}}f\,\mathrm{d}\bar{z}^k\big) \overset{(2)}{=} \partial_{\bar{k}}\partial_j f\,\mathrm{d}\bar{z}^k\wedge\mathrm{d}z^j + \partial_j\partial_{\bar{k}}f\,\mathrm{d}z^j\wedge\mathrm{d}\bar{z}^k \overset{(3)}{=} 0.$$

(1) (T1.122) on the function f (sum over j,k).

(2) (T1.122) again on each 1-form.
(3) Mixed partials commute, and $d\bar{z}^k \wedge dz^j = -dz^j \wedge d\bar{z}^k$.

It is the same mechanism as $d^2 = 0$ in Section T1.5: symmetric second derivatives against antisymmetric wedges.

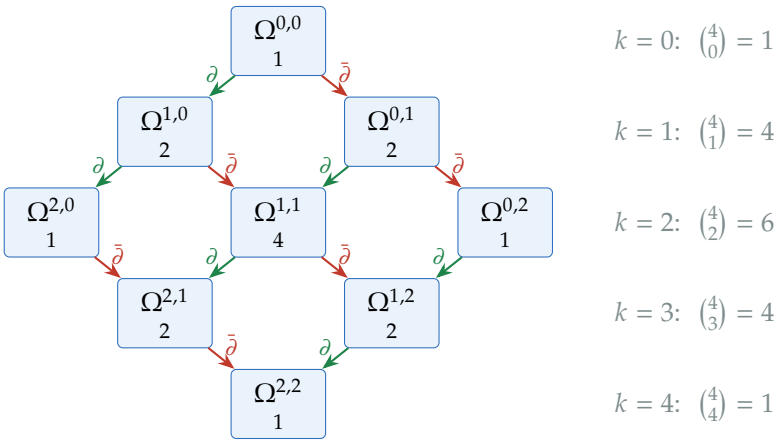


Figure T1.31: The (p, q)-forms on $\mathbb{C}^2$ with the dimension of each space at a point (small numbers). The row $p + q = k$ holds all k-forms, and the dimensions in a row add up to $\binom{4}{k}$. The green arrows are ∂ (down-left), the red arrows are $\bar{\partial}$ (down-right); d is their sum. A chain of red arrows running down to the right (fixed p) is a Dolbeault complex (T1.126).

Worked example T1.59: Computing with ∂ and $\bar{\partial}$ on $\mathbb{C}^2$; the Kähler form

(a) Let $\alpha = \bar{z}^1\, dz^2$, a $(1,0)$-form. Its coefficient depends only on $\bar{z}^1$, so $\partial\alpha = 0$ and $\bar{\partial}\alpha = d\bar{z}^1 \wedge dz^2 = -dz^2 \wedge d\bar{z}^1$, a $(1,1)$-form. Then $d\alpha = \partial\alpha + \bar{\partial}\alpha = d\bar{z}^1 \wedge dz^2$, which you can confirm by writing everything in x, y.
(b) Take the real function $K = |z|^2 = \sum_j z^j \bar{z}^j$. Then

$$\frac{i}{2}\,\partial\bar{\partial}K \overset{(1)}{=} \frac{i}{2}\,\partial\Big(\sum_j z^j\, d\bar{z}^j\Big) \overset{(2)}{=} \frac{i}{2}\sum_j dz^j \wedge d\bar{z}^j \overset{(3)}{=} \sum_j dx^j \wedge dy^j.$$

(1) $\partial_{\bar{z}^j} K = z^j$.
(2) $\partial_{z^k} z^j = \delta^j_k$.
(3) (T1.115) in each complex plane.

This $(1,1)$-form $\omega = \frac{i}{2}\partial\bar{\partial}K$ is the **Kähler form** of flat $\mathbb{C}^m$. It is real, closed ($d\omega = (\partial + \bar{\partial})\partial\bar{\partial}K = 0$ by (T1.123)), and, reading $z^j = q^j + ip_j$, it is the symplectic form $\sum dq^j \wedge dp_j$ of Example T1.53. A function K with $\omega = \frac{i}{2}\partial\bar{\partial}K$ is called a **Kähler potential**; Nash (5.29)–(5.30) defines Kähler manifolds by exactly this kind of closed $(1,1)$-form.

T1.10.4 Complex manifolds, and the Riemann sphere

A **complex manifold** of complex dimension m is a manifold with an atlas of charts valued in $\mathbb{C}^m$ whose transition maps are *holomorphic*. Holomorphic transitions are what make the type of a form chart-independent: if $w = w(z)$ is holomorphic, then $dw^j = \sum_k (\partial w^j/\partial z^k)\, dz^k$ contains no $d\bar{z}$, so a (p, q)-form in one chart is a (p, q)-form in every chart, and ∂, $\bar{\partial}$ are globally defined. (On a surface with a non-holomorphic transition, dw would pick up a $d\bar{z}$ part and the splitting would depend on the chart.)

The basic compact example is the Riemann sphere $\mathbb{CP}^1$, found in Section T1.1: two charts z (north) and w (flipped south), with $w = 1/z$ on the overlap $\mathbb{C} \setminus \{0\}$. There

$$\mathrm{d}w = -\frac{\mathrm{d}z}{z^2}, \qquad \mathrm{d}\bar{w} = -\frac{\mathrm{d}\bar{z}}{\bar{z}^2}, \qquad \partial_z = -w^2\,\partial_w. \tag{T1.124}$$

The last one is the chain rule $\partial_z = (\partial w/\partial z)\,\partial_w = -z^{-2}\partial_w = -w^2\partial_w$.

Worked example T1.60: Holomorphic 1-forms and vector fields on $\mathbb{CP}^1$ and on T^2

Holomorphic 1-forms on $\mathbb{CP}^1$. A global holomorphic $(1,0)$-form is $f(z)\,\mathrm{d}z$ in the north chart with f entire, and must also be holomorphic in the south chart:

$$f(z)\,\mathrm{d}z \overset{(1)}{=} -\frac{f(1/w)}{w^2}\,\mathrm{d}w.$$

(1) $z = 1/w$ and $\mathrm{d}z = -\mathrm{d}w/w^2$ (T1.124).

Regularity at $w = 0$ requires $f(z) = O(z^{-2})$ as $z \to \infty$. An entire function that tends to zero at infinity is bounded, hence constant (Liouville's theorem), hence zero. So $\mathbb{CP}^1$ has *no* holomorphic 1-forms: $\dim H^{1,0}(\mathbb{CP}^1) = 0$. This matches $b_1(S^2) = 0$.

Holomorphic vector fields on $\mathbb{CP}^1$. Now take $v(z)\,\partial_z$ with v entire. In the south chart it is $-w^2 v(1/w)\,\partial_w$, which is regular at $w = 0$ iff $v(z) = O(z^2)$ at infinity, i.e. iff v is a polynomial of degree at most 2:

$$v(z)\,\partial_z \in \mathrm{span}_{\mathbb{C}}\{\partial_z,\ z\,\partial_z,\ z^2\partial_z\}.$$

Their flows are the translations $z \mapsto z + a$, the dilations and rotations $z \mapsto \lambda z$, and the "special conformal" maps $z \mapsto z/(1 - cz)$: together the Möbius group $\mathrm{SL}(2,\mathbb{C})/\{\pm 1\}$, of complex dimension 3. This is Nash's footnote on p. 165: a three-complex-parameter family of conformal Killing vectors on the sphere.

The torus. $T^2 = \mathbb{C}/\Lambda$ with $\Lambda = \mathbb{Z} + \tau\mathbb{Z}$, $\mathrm{Im}\,\tau > 0$. A holomorphic vector field is $v(z)\partial_z$ with v entire and doubly periodic. Such a v is bounded (it is determined by its values on the compact fundamental parallelogram), hence constant. So the torus has exactly one (complex) holomorphic vector field, ∂_z, and by the same argument one holomorphic 1-form, $\mathrm{d}z$: Nash's "the space of conformal Killing vectors is 1-dimensional for genus one" (p. 166).

Why are holomorphic vector fields the same as conformal Killing vectors? Here is the computation Nash does on p. 160.

Worked example T1.61: Lie derivative of a metric in complex coordinates: Nash's $\delta\mu = 2\bar{\partial}V$

Take a conformal metric $g = \lambda\,\mathrm{d}z\,\mathrm{d}\bar{z}$, i.e. $g_{z\bar{z}} = g_{\bar{z}z} = \lambda/2$ and $g_{zz} = g_{\bar{z}\bar{z}} = 0$, and a real vector field $V = V^z\partial_z + V^{\bar{z}}\partial_{\bar{z}}$ with $V^{\bar{z}} = \overline{V^z}$ (for $V = V^x\partial_x + V^y\partial_y$ one has $V^z = V^x + \mathrm{i}V^y$). The component formula (T1.103) works in any basis of coordinates, including $z, \bar{z}$:

$$(\mathcal{L}_V g)_{\bar{z}\bar{z}} \overset{(1)}{=} V^\lambda\partial_\lambda g_{\bar{z}\bar{z}} + g_{\lambda\bar{z}}\,\partial_{\bar{z}}V^\lambda + g_{\bar{z}\lambda}\,\partial_{\bar{z}}V^\lambda \overset{(2)}{=} 2\,g_{z\bar{z}}\,\partial_{\bar{z}}V^z \overset{(3)}{=} \lambda\,\partial_{\bar{z}}V^z,$$

$$(\mathcal{L}_V g)_{z\bar{z}} \overset{(4)}{=} V^\lambda\partial_\lambda g_{z\bar{z}} + g_{z\bar{z}}\big(\partial_z V^z + \partial_{\bar{z}}V^{\bar{z}}\big).$$

(1) (T1.103) with $\mu = \nu = \bar{z}$; λ runs over $z, \bar{z}$.

(2) $g_{\bar{z}\bar{z}} = 0$ kills the first term; in the others only $\lambda = z$ survives, since $g_{\bar{z}\bar{z}} = 0$.

(3) $g_{z\bar{z}} = \lambda/2$ (the conformal factor; the letter λ in the summation index is unrelated).

(4) Same formula with $\mu = z$, $\nu = \bar{z}$: in the first correction term only $g_{\bar{z}\bar{z}}\ldots$ no: $g_{\lambda\bar{z}}\partial_z V^\lambda$ survives for $\lambda = z$, and $g_{z\lambda}\partial_{\bar{z}}V^\lambda$ for $\lambda = \bar{z}$.

The $(z\bar z)$ component is proportional to the metric itself: it only rescales λ. The $(\bar z\bar z)$ component, and its conjugate (zz), change the *shape*: they turn $\mathrm{d}z\,\mathrm{d}\bar z$ into $|\mathrm{d}z+\mu\,\mathrm{d}\bar z|^2$, a metric that is no longer conformal to the old one. Expanding Nash's perturbed line element $(\lambda+\delta\lambda)|\mathrm{d}z+\delta\mu\,\mathrm{d}\bar z|^2$ to first order gives $\delta g_{\bar z\bar z}=\lambda\,\delta\mu$, so a pure diffeomorphism produces

$$\delta\mu=\frac{(\mathcal{L}_V g)_{\bar z\bar z}}{\lambda}=\partial_{\bar z}V^z.$$

Nash normalises $g_{z\bar z}=\lambda$ instead of $\lambda/2$ and gets $\delta\mu=2\partial_{\bar z}V^z$ (5.72); as he says, the factor 2 is not significant. Two conclusions:

- V is a conformal Killing vector ($\mathcal{L}_V g\propto g$) iff $\partial_{\bar z}V^z=0$, i.e. iff $V^z\partial_z$ is a *holomorphic* vector field. On the plane there are infinitely many ($V^z=z^n$ for all $n\geq 0$; with the poles allowed at 0 these generate the Witt algebra of conformal field theory, Chapter 11). On $\mathbb{CP}^1$ there are three and on T^2 one (Example T1.60).
- The shape changes that are *not* fake are the Beltrami differentials $\delta\mu$ modulo those of the form $\bar\partial V$: a Dolbeault-type quotient "ker/Im $\bar\partial$". This is how Nash counts the moduli of Riemann surfaces in Chapter 5.

Worked example T1.62: The Fubini–Study form on $\mathbb{CP}^1$ has total integral 1

Define on the north chart

$$\omega_{FS}=\frac{\mathrm{i}}{2\pi}\,\partial\bar\partial\log\bigl(1+|z|^2\bigr). \tag{T1.125}$$

Compute:

$$\partial\bar\partial\log(1+z\bar z)\overset{(1)}{=}\partial\Bigl(\frac{z\,\mathrm{d}\bar z}{1+z\bar z}\Bigr)\overset{(2)}{=}\partial_z\Bigl(\frac{z}{1+z\bar z}\Bigr)\mathrm{d}z\wedge\mathrm{d}\bar z\overset{(3)}{=}\frac{(1+z\bar z)-z\bar z}{(1+z\bar z)^2}\,\mathrm{d}z\wedge\mathrm{d}\bar z=\frac{\mathrm{d}z\wedge\mathrm{d}\bar z}{(1+|z|^2)^2},$$

$$\omega_{FS}\overset{(4)}{=}\frac{\mathrm{i}}{2\pi}\cdot\frac{-2\mathrm{i}\,\mathrm{d}x\wedge\mathrm{d}y}{(1+r^2)^2}=\frac{1}{\pi}\frac{\mathrm{d}x\wedge\mathrm{d}y}{(1+r^2)^2},$$

$$\int_{\mathbb{CP}^1}\omega_{FS}\overset{(5)}{=}\frac{1}{\pi}\int_0^{2\pi}\!\!\int_0^{\infty}\frac{r\,\mathrm{d}r\,\mathrm{d}\phi}{(1+r^2)^2}\overset{(6)}{=}\frac{1}{\pi}\cdot 2\pi\cdot\frac{1}{2}=1.$$

(1) $\partial_{\bar z}\log(1+z\bar z)=z/(1+z\bar z)$.
(2) ∂ acts on the coefficient and puts $\mathrm{d}z$ in front.
(3) Quotient rule.
(4) (T1.115).
(5) The north chart covers everything except one point, which has zero area; polar coordinates.
(6) $\int_0^\infty r\,\mathrm{d}r/(1+r^2)^2=\bigl[-\frac{1}{2(1+r^2)}\bigr]_0^\infty=\frac12$.

Is ω_{FS} really a global form? In the south chart, the same formula with $w=1/z$ gives $\log(1+|w|^2)=\log(1+|z|^2)-\log|z|^2$, and $\partial\bar\partial\log|z|^2=0$ away from $z=0$ by (T1.118). So the two local definitions agree on the overlap. Comparing with the round metric in stereographic coordinates, $4\,\mathrm{d}u\,\mathrm{d}v/(1+\rho^2)^2$ (Example T1.9), ω_{FS} is the area form of the unit sphere divided by 4π. It is closed and real, and it is the first Chern form of the monopole (hyperplane) line bundle with unit charge: the "1" is the Dirac quantum of Section T1.7. Note that ω_{FS} is *locally* d-exact ($\frac{\mathrm{i}}{2\pi}\partial\bar\partial K=\mathrm{d}\bigl(\frac{\mathrm{i}}{2\pi}\bar\partial K\bigr)$) but not globally, since its integral over the closed surface $\mathbb{CP}^1$ is not zero (the exemplar argument: an exact form would integrate to zero).

Watch out: Normalisations of the Fubini–Study form

Nash (5.30) writes $\omega = \frac{\mathrm{i}}{2}\partial\bar\partial\ln(1+|z|^2)$, without the $1/\pi$; its integral over $\mathbb{CP}^1$ is π. Our ω_{FS} is normalised to integrate to 1 so that it represents the generator of $H^2(\mathbb{CP}^1;\mathbb{Z})$. Also, in (5.30) the factor $\frac{\mathrm{i}}{2}$ belongs with ω, not with the metric: the Hermitian metric coefficients are the real, positive-definite $g_{i\bar\jmath} = \partial_i\partial_{\bar\jmath}\ln(1+|z|^2) = \big[\delta_{ij}(1+|z|^2) - \bar z_i z_j\big]/(1+|z|^2)^2$.

T1.10.5 The Dolbeault complex and Dolbeault cohomology

Fix p. Since $\bar\partial^2 = 0$, the (p,q)-forms form a sequence in which each map composed with the next is zero:

$$0 \longrightarrow \Omega^{p,0} \xrightarrow{\bar\partial} \Omega^{p,1} \xrightarrow{\bar\partial} \Omega^{p,2} \xrightarrow{\bar\partial} \cdots \xrightarrow{\bar\partial} \Omega^{p,m} \longrightarrow 0. \tag{T1.126}$$

This is the **Dolbeault complex** (a chain of red arrows at fixed p in Figure T1.31), and exactly as for d one defines

Working definition T1.28 (Dolbeault cohomology)

$$H^{p,q}_{\bar\partial}(M) = \frac{\ker\big(\bar\partial : \Omega^{p,q} \to \Omega^{p,q+1}\big)}{\mathrm{Im}\big(\bar\partial : \Omega^{p,q-1} \to \Omega^{p,q}\big)}, \qquad h^{p,q} = \dim_{\mathbb{C}} H^{p,q}_{\bar\partial}(M). \tag{T1.127}$$

Operationally: a class is a $\bar\partial$-closed (p,q)-form, and two are equivalent if they differ by $\bar\partial$ of a $(p,q-1)$-form. The group $H^{0,0}$ is the space of global holomorphic functions [functions f on all of M with $\bar\partial f = 0$, i.e. $\partial f/\partial\bar z^j = 0$ for every j], and $H^{p,0}$ the space of global holomorphic p-forms [$(p,0)$-forms on all of M whose coefficient functions are holomorphic] (there is nothing to quotient by).

Locally the cohomology is trivial: (T1.119) solves $\bar\partial f = g\,\mathrm{d}\bar z$ on a disc (multiply g by a cut-off function first). So, as for de Rham cohomology, $H^{p,q}_{\bar\partial}$ is about *global* obstructions. Here is one, computed by hand.

Worked example T1.63: $\mathrm{d}\bar z$ on a torus is $\bar\partial$-closed but not $\bar\partial$-exact

On $T^2 = \mathbb{C}/(\mathbb{Z}+\tau\mathbb{Z})$ the constant form $\mathrm{d}\bar z$ is well defined (translations do not change it) and $\bar\partial$-closed (there are no $(0,2)$-forms in complex dimension 1). Suppose $\mathrm{d}\bar z = \bar\partial f$ for a smooth function f on the torus, i.e. a doubly periodic function on $\mathbb{C}$. Then:

$$\begin{aligned}
\partial_{\bar z}(f-\bar z) = 0 &\overset{(1)}{\Longrightarrow} f = \bar z + h(z),\ h \text{ entire}\\
&\overset{(2)}{\Longrightarrow} h(z+1)-h(z) = -1, \quad h(z+\tau)-h(z) = -\bar\tau\\
&\overset{(3)}{\Longrightarrow} h(z) = cz + d \ \text{ with } \ c=-1,\ c\tau = -\bar\tau\\
&\overset{(4)}{\Longrightarrow} \tau = \bar\tau,
\end{aligned}$$

(1) A function with $\partial_{\bar z} = 0$ is holomorphic.
(2) Periodicity of f under $z \mapsto z+1$ and $z \mapsto z+\tau$, with $\overline{z+\tau} = \bar z + \bar\tau$.
(3) Differentiating (2), h' is doubly periodic and entire, hence bounded, hence constant (Liouville); then read off c from (2).
(4) Combine the two conditions on c.

which contradicts $\mathrm{Im}\,\tau > 0$. So $[\mathrm{d}\bar z] \neq 0$ in $H^{0,1}(T^2)$. With a little Hodge theory (the harmonic $(0,1)$-forms on a flat torus are the constant multiples of $\mathrm{d}\bar z$) one finds $h^{0,1}(T^2) = 1$.

Together with $h^{1,0} = 1$ ($\mathrm{d}z$) this reproduces the de Rham count $b_1(T^2) = 2$: the classes $[\mathrm{d}x], [\mathrm{d}y]$ are recombined as $[\mathrm{d}z], [\mathrm{d}\bar{z}]$. Nash's (5.34)(i), $b_i = \sum_{p+q=i} h_{p,q}$, in the simplest non-trivial case.

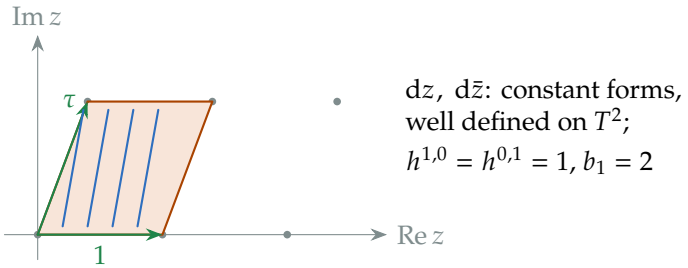


Figure T1.32: The torus $T^2 = \mathbb{C}/(\mathbb{Z} + \tau\mathbb{Z})$: the orange parallelogram with opposite sides glued. Constant forms such as $\mathrm{d}z$ and $\mathrm{d}\bar{z}$ are invariant under the lattice translations and so live on T^2. The function $\bar{z}$ is *not* periodic, which is why $\mathrm{d}\bar{z} = \bar{\partial}\bar{z}$ is closed but not exact on the torus, the Dolbeault analogue of $\mathrm{d}x$ being closed but not exact on a circle.

Key idea: Nash's index of $\bar{\partial}$

For a compact complex manifold, Nash (4.64)–(4.65) defines the index of the Dolbeault complex with $p = 0$ as the alternating sum index $\bar{\partial} = \sum_q (-1)^q h^{0,q}$, the **arithmetic genus**. For a Riemann surface ($m = 1$) only $q = 0, 1$ occur, and $h^{0,0} = 1$ because the only global holomorphic functions on a compact surface are constants. So:

surface	$h^{0,0}$	$h^{1,0}$	$h^{0,1}$	b_1	index $\bar{\partial} = h^{0,0} - h^{0,1}$
$\mathbb{CP}^1 = S^2$	1	0	0	0	1
T^2	1	1	1	2	0
genus g	1	g	g	$2g$	$1 - g$

The entries for $\mathbb{CP}^1$ and T^2 are the hand computations of Examples T1.60 and T1.63, together with Hodge symmetry $h^{0,1} = h^{1,0}$ (Nash's (5.34)(iv)). The index $1 - g$ is what the index theorem computes from the Todd class (Chapter 4), and Riemann–Roch (Chapter 5) is its generalisation to $\bar{\partial}$ acting on sections of a line bundle.

T1.10.6 The symbol and the Laplacian of $\bar{\partial}$

Nash's index theory needs two more facts about $\bar{\partial}$: it is elliptic (as a complex), and its Laplacian is half the ordinary Laplacian on flat space. Both are quick computations on $\mathbb{C}$.

Symbol. Replacing $\partial_\mu \to \mathrm{i}k_\mu$ as in Section T1.6, write the real covector as $k = k_x\mathrm{d}x + k_y\mathrm{d}y = k_z\mathrm{d}z + k_{\bar{z}}\mathrm{d}\bar{z}$ with $k_z = \frac{1}{2}(k_x - \mathrm{i}k_y)$ and $k_{\bar{z}} = \frac{1}{2}(k_x + \mathrm{i}k_y)$. Then

$$\sigma(\bar{\partial})(k)\,\omega = \mathrm{i}\,k_{\bar{z}}\,\mathrm{d}\bar{z} \wedge \omega = \mathrm{i}\,k^{(0,1)} \wedge \omega, \tag{T1.128}$$

and $k_{\bar{z}} \neq 0$ for every real $k \neq 0$. The symbol sequence is therefore exact away from $k = 0$ (the analogue of the de Rham argument with $k\wedge$), which is Nash's statement that the symbol complex (4.66) is exact.

Adjoint and Laplacian. Use the Hermitian L^2 product $\langle \alpha, \beta \rangle = \int \alpha \wedge \star\bar{\beta}$ (the complex version of (T1.38)), for which $\langle f, h \rangle = \int f\bar{h}\,\mathrm{d}^2x$ and $\langle a\,\mathrm{d}\bar{z}, b\,\mathrm{d}\bar{z} \rangle = \int a\bar{b}\,|\mathrm{d}\bar{z}|^2\,\mathrm{d}^2x$ with $|\mathrm{d}\bar{z}|^2 =$

$|\mathrm{d}x|^2 + |\mathrm{d}y|^2 = 2$. For f and b of compact support,

$$\langle \bar\partial f,\, b\,\mathrm{d}\bar z\rangle \overset{(1)}{=} 2\int \partial_{\bar z} f\;\bar b\,\mathrm{d}^2x \overset{(2)}{=} -2\int f\,\partial_{\bar z}\bar b\,\mathrm{d}^2x \overset{(3)}{=} -2\int f\,\overline{\partial_z b}\,\mathrm{d}^2x \overset{(4)}{=} \big\langle f,\,-2\,\partial_z b\big\rangle.$$

(1) $\bar\partial f = \partial_{\bar z} f\,\mathrm{d}\bar z$ and $|\mathrm{d}\bar z|^2 = 2$.
(2) Integration by parts (no boundary terms).
(3) $\overline{\partial_z b} = \partial_{\bar z}\bar b$: conjugating swaps ∂_z and $\partial_{\bar z}$ (T1.111).
(4) Definition of $\langle f, h\rangle$.

So $\bar\partial^*(b\,\mathrm{d}\bar z) = -2\,\partial_z b$, and on functions

$$\Delta_{\bar\partial} f = \bar\partial^*\bar\partial f = -2\,\partial_z\partial_{\bar z} f = \tfrac12\Delta f, \tag{T1.129}$$

by (T1.114). This is Nash's (5.32), $\Delta_d = 2\Delta_{\bar\partial} = 2\Delta_\partial$, checked on functions on flat $\mathbb{C}$ (on a general Kähler manifold it holds on all forms). Its symbol is $(-2\mathrm{i}k_z)(\mathrm{i}k_{\bar z}) = 2k_zk_{\bar z} = \tfrac12|k|^2$, half of the symbol $|k|^2$ of Δ (T1.59). Consequently the $\bar\partial$-harmonic forms and the d-harmonic forms coincide on a compact Kähler manifold, which is what makes the Hodge numbers $h^{p,q}$ computable from harmonic forms.

Reading Nash (pp. 63–65, (3.24)–(3.31); pp. 103, 106–107, (4.49), (4.63)–(4.66); pp. 146–148, (5.29)–(5.34); pp. 158–160)

Chapter 3. Nash introduces (p,q)-forms by the local formula (3.24), the operators $\partial, \bar\partial$ by (3.25) (our (T1.122)), and the relations $\mathrm{d} = \partial + \bar\partial$, $\mathrm{d}^2 = \partial^2 = \partial\bar\partial + \bar\partial\partial = \bar\partial^2 = 0$ (3.26) (our (T1.123)). He then runs the sheaf-theoretic argument that gave de Rham's theorem with d replaced by $\bar\partial$, using the fine sheaves $\mathcal{E}^{p,q}$ and the sheaf Ω^p of holomorphic p-forms (note that Ω^p there means *holomorphic* forms, not all forms), to obtain *Dolbeault's theorem* $H^q(M;\Omega^p) \cong H^{p,q}_{\bar\partial}(M)$ (3.31). The input is the local exactness we saw in (T1.119).
Chapter 4. On p. 103 he complexifies the de Rham complex; on pp. 106–107 he writes the Dolbeault complex and its cohomology (4.63), sets $p = 0$, and defines index $\bar\partial = \sum_q (-1)^q h^{0,q}$ (4.64)–(4.65), the arithmetic genus. The symbol complex (4.66) is our (T1.128); $\Lambda^{0,q}T^*X$ is the bundle whose sections are $(0,q)$-forms.
Chapter 5. Kähler manifolds are defined by a closed $(1,1)$-form $\omega = \frac{\mathrm{i}}{2} g_{i\bar\jmath}\mathrm{d}z^i \wedge \mathrm{d}\bar z^j$ (5.29) (Example T1.59); the Fubini–Study metric is (5.30) (Example T1.62, with the caveats in the pitfall above); the three Laplacians and $\Delta_d = 2\Delta_{\bar\partial}$ are (5.31)–(5.32) (our (T1.129)). On pp. 158–160 the Beltrami differential and $\delta\mu = 2\partial_{\bar z} V^z$ appear (Example T1.61).

Recap

- $\mathrm{d}z = \mathrm{d}x + \mathrm{i}\,\mathrm{d}y$, $\mathrm{d}\bar z = \mathrm{d}x - \mathrm{i}\,\mathrm{d}y$; $\partial_z = \frac12(\partial_x - \mathrm{i}\partial_y)$, $\partial_{\bar z} = \frac12(\partial_x + \mathrm{i}\partial_y)$; $\mathrm{d}f = \partial_z f\,\mathrm{d}z + \partial_{\bar z} f\,\mathrm{d}\bar z$.
- Holomorphic $\Leftrightarrow \bar\partial f = 0$ (Cauchy–Riemann). $\nabla^2 = 4\partial_z\partial_{\bar z}$, so holomorphic functions are harmonic. $\mathrm{d}z\wedge\mathrm{d}\bar z = -2\mathrm{i}\,\mathrm{d}x\wedge\mathrm{d}y$; $\star\mathrm{d}z = -\mathrm{i}\,\mathrm{d}z$, $\star\mathrm{d}\bar z = \mathrm{i}\,\mathrm{d}\bar z$.
- Cauchy's theorem is Stokes for the closed form $f\,\mathrm{d}z$; $\partial_{\bar z}(1/z) = \pi\delta^2$ gives the Cauchy formula and the Green functions of $\bar\partial$ and Δ.
- (p,q)-forms, $\dim = \binom{m}{p}\binom{m}{q}$; $\mathrm{d} = \partial + \bar\partial$ with $\partial^2 = \bar\partial^2 = \partial\bar\partial + \bar\partial\partial = 0$. Holomorphic transition maps make types global.
- Dolbeault complex and $H^{p,q}_{\bar\partial}$. Explicitly: $\mathbb{CP}^1$ has no holomorphic 1-forms and three holomorphic vector fields; T^2 has $h^{1,0} = h^{0,1} = 1$; $\operatorname{index}\bar\partial = 1 - g$.

- $\mathcal{L}_V g$ in complex coordinates: $\delta\mu \propto \partial_{\bar{z}} V^z$; conformal Killing = holomorphic. ω_{FS} integrates to 1. $\Delta_{\bar{\partial}} = \frac{1}{2}\Delta$.

Chapter summary

We have built one calculus that replaces the grad–curl–div toolbox. Its objects are p-forms, the things we integrate over p-dimensional regions. There are six operations on them. Two need nothing but the manifold: the wedge product $\wedge$ and the exterior derivative d. One needs a map, the pullback f^*. Two need a vector field, the interior product ι_X and the Lie derivative $\mathcal{L}_X$. One needs a metric and an orientation: the Hodge star $\star$, and with it the codifferential δ and the Laplacian Δ. On a complex manifold, d splits further into $\partial + \bar{\partial}$.

The summary has three parts. Figure T1.33 is a map of the operators. Table T1.1 says what each operator needs and how it behaves. The concept tables (Tables T1.2 to T1.7) then list every concept of the chapter, with its operational meaning and where the book uses it again.

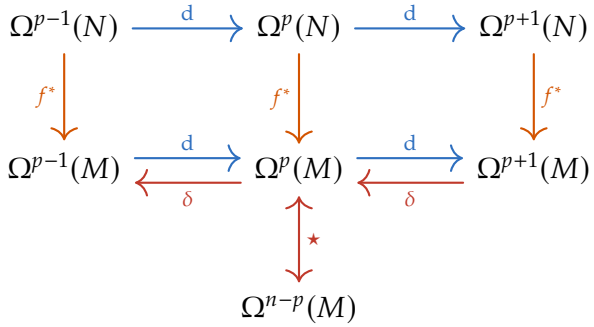


Figure T1.33: The operators of the chapter on one page. Blue arrows (d) and orange arrows (f^*) need no metric, and every square commutes: this is naturality, $\mathrm{d}f^* = f^*\mathrm{d}$ (Section T1.8). Red arrows ($\star$ and $\delta = \pm \star\, \mathrm{d} \star$) need a metric and an orientation. The interior product ι_X goes one step left along the bottom row, and the Lie derivative $\mathcal{L}_X = \mathrm{d}\iota_X + \iota_X \mathrm{d}$ stays at the same degree. See Table T1.1 for the details.

How to read the concept tables. Each row gives a concept, its *operational* meaning (what you actually compute, or how you test for it), and where it comes back later. Section references (§) are to this chapter. Chapter references are to the rest of the book, and page numbers "Nash p. x" are Nash's printed pages.

Operator	Degree	Needs	Key rules
$\alpha \wedge \beta$	$p+q$	nothing	$\alpha \wedge \beta = (-1)^{pq}\beta \wedge \alpha$; forms of degree $> n$ vanish
d	$+1$	nothing	$\mathrm{d}^2 = 0$; $\mathrm{d}(\alpha \wedge \beta) = \mathrm{d}\alpha \wedge \beta + (-1)^p \alpha \wedge \mathrm{d}\beta$
f^*	0	a map $f : M \to N$	$f^*(\alpha \wedge \beta) = f^*\alpha \wedge f^*\beta$; $\mathrm{d}f^* = f^*\mathrm{d}$; $(f \circ g)^* = g^* f^*$
ι_X	-1	a vector field	$\iota_X^2 = 0$; $\iota_X(\alpha \wedge \beta) = \iota_X\alpha \wedge \beta + (-1)^p \alpha \wedge \iota_X\beta$
$\mathcal{L}_X$	0	a vector field	$\mathcal{L}_X = \mathrm{d}\iota_X + \iota_X\mathrm{d}$; $[\mathcal{L}_X, \mathrm{d}] = 0$; $\mathcal{L}_X\mathrm{vol} = (\mathrm{div}\, X)\mathrm{vol}$
$\star$	$p \mapsto n-p$	metric, orientation	$\alpha \wedge \star\beta = g(\alpha,\beta)\mathrm{vol}$; $\star\star = (-1)^{p(n-p)+s}$
δ	-1	metric, orientation	$\delta = (-1)^{n(p+1)+1+s} \star \mathrm{d} \star$; $\langle \mathrm{d}\alpha, \beta\rangle = \langle \alpha, \delta\beta\rangle$; $\delta^2 = 0$
Δ	0	metric	$\Delta = \mathrm{d}\delta + \delta\mathrm{d} \geq 0$; $\Delta f = -\nabla^2 f$ in flat space; symbol $\lvert k\rvert^2$
∂, $\bar\partial$	$(1,0)$, $(0,1)$	complex structure	$\mathrm{d} = \partial + \bar\partial$; $\partial^2 = \bar\partial^2 = 0$; $\partial\bar\partial = -\bar\partial\partial$

Table T1.1: The operators of this chapter. "Degree" is the change in form degree; s is the number of minus signs in the metric.

Concept	Operational meaning	Used later
Chart (§T1.1)	Coordinates x^μ on a patch U, invertible onto an open set of $\mathbb{R}^n$. Write the formula and state which region it covers.	Everywhere; local trivialisations of bundles (Chapter 1)
Transition map	$x' \circ x^{-1}$ on the overlap, e.g. $u' = u/\lvert u\rvert^2$ between the two stereographic charts of S^2. It must be smooth.	Bundle transition functions (Chapters 1 and T3); monopole (§T1.7, Chapter T2)
Atlas, manifold	Enough charts to cover the space, with smooth transitions. Examples: S^1 (2 charts), S^2 (2), T^2, $\mathbb{CP}^1$.	All of the book
Orientation	All Jacobians $\det(\partial x'/\partial x)$ of the transitions are > 0. For S^2 stereographic, fix the sign by flipping one chart.	Integration and Stokes (Chapter T2); $\star$; signature (Chapter 4)
Regular level set	If $\mathrm{d}f \neq 0$ on $f = c$, then $\{f = c\}$ is a manifold of dimension $n-1$ (e.g. $S^2 \subset \mathbb{R}^3$).	Morse theory (Chapter 7)
Tangent vector (§T1.2)	The velocity of a curve, $v^\mu = \dot{x}^\mu$. It acts on functions by $v(f) = v^\mu \partial_\mu f$ and transforms by $v'^\mu = (\partial x'^\mu/\partial x^\nu)v^\nu$.	Vector fields, flows (§T1.9); gradient flows (Chapter 7)
Commutator $[X, Y]$	$[X,Y]^\mu = X^\lambda \partial_\lambda Y^\mu - Y^\lambda \partial_\lambda X^\mu$; it measures whether two flows commute.	Lie algebras (Chapter T3); diffeomorphism groups (Chapter 6)
Covector, $\mathrm{d}f$	A linear machine on vectors: $\mathrm{d}f(v) = v(f)$, $\mathrm{d}x^\mu(\partial_\nu) = \delta^\mu_\nu$. Pictured as a stack of level surfaces.	Every form; symbols of operators (Chapter 2)
Line element, metric	$\mathrm{d}s^2 = g_{\mu\nu}\mathrm{d}x^\mu \mathrm{d}x^\nu$, e.g. $\mathrm{d}r^2 + r^2\mathrm{d}\phi^2$ or $4(\mathrm{d}u^2 + \mathrm{d}v^2)/(1+\lvert u\rvert^2)^2$ on S^2. Raise and lower indices with g.	Hodge theory (Chapter 2); worldsheet metrics (Chapter 9)

Table T1.2: Summary, part 1: manifolds, vectors and covectors.

Concept	Operational meaning	Used later
Wedge product (§T1.3)	An antisymmetric product: $dx \wedge dx = 0$ and $dy \wedge dx = -dx \wedge dy$. On vectors, $(\alpha \wedge \beta)(v, w) = \alpha(v)\beta(w) - \alpha(w)\beta(v)$, the oriented area.	Cup product on cohomology (Nash p. 13–15, Chapter 1)
Change of variables	Substitute and expand; the Jacobian appears by itself: $dx \wedge dy = r\, dr \wedge d\phi$.	Integration (Chapter T2)
p-form	$\alpha = \frac{1}{p!}\alpha_{\mu_1 \ldots \mu_p} dx^{\mu_1} \wedge \cdots \wedge dx^{\mu_p}$ with antisymmetric components; the space has dimension $\binom{n}{p}$.	Throughout; Nash's convention omits the $1/p!$ (Chapter 2)
Graded commutativity	$\alpha \wedge \beta = (-1)^{pq}\beta \wedge \alpha$. So an odd form squares to zero, while $F \wedge F \neq 0$ is possible.	$\operatorname{tr} F \wedge F$ (Chapters 8 and 10)
Top form, determinant	$\alpha^1 \wedge \cdots \wedge \alpha^n = \det(\alpha^i{}_\mu)\, dx^1 \wedge \cdots \wedge dx^n$.	Volume forms; the degree of a map (§T1.8)
Decomposable form	$\alpha = \beta \wedge \gamma$ for 1-forms β, γ. For a 2-form the test is $\alpha \wedge \alpha = 0$ (e.g. $dx^1 \wedge dx^2 + dx^3 \wedge dx^4$ fails it).	Self-dual 2-forms (Chapter 8)
3D dictionary	0-, 1-, 2- and 3-forms ↔ scalar, vector (work), pseudovector (flux) and density (charge). $\wedge$ contains $\times$ and $\cdot$.	Maxwell (§T1.7)
Pictures of forms	A 1-form is a stack of surfaces (count how many you cross); a 2-form is a bundle of tubes (count how many you pierce); a 3-form is a swarm of cells.	Intuition for Stokes (Chapter T2)

Table T1.3: Summary, part 2: the algebra of forms.

Concept	Operational meaning	Used later
Inner product of forms (§T1.4)	Use $g^{\mu\nu}$ on 1-forms. On p-forms, take the determinant of the matrix of 1-form inner products; an orthonormal coframe gives an orthonormal basis of p-forms.	L^2 norms, Hodge theory (Chapter 2)
Signature s	The number of minus signs in g: $s = 0$ Euclidean, $s = 1$ Minkowski. Norms of forms can be negative when $s \neq 0$.	Signs of $\star\star$ and δ
Volume form	$\mathrm{vol} = \sqrt{\lvert g \rvert}\, dx^1 \wedge \cdots \wedge dx^n$ in an oriented chart, e.g. $r^2 \sin\theta\, dr \wedge d\theta \wedge d\phi$.	All integrals; $\mathcal{L}_X \mathrm{vol}$ (§T1.9)
Hodge star	Complete an orthonormal basis form to vol: $\alpha \wedge \star\beta = g(\alpha, \beta)\mathrm{vol}$. Tables for $\mathbb{R}^2$, polar, $\mathbb{R}^3$, spherical, $\mathbb{R}^4$, Minkowski.	Hodge theory (Nash p. 39–41); instantons; Yang–Mills action (Chapters 2, T3 and 8)
$\star\star$ sign	$\star\star = (-1)^{p(n-p)+s}$ on p-forms. It equals +1 on 2-forms in $\mathbb{R}^4$ and −1 on 2-forms in Minkowski space.	Self-duality exists only in Euclidean signature (Chapter 8)
(Anti-)self-dual 2-forms	$\star\omega = \pm\omega$ in $\mathbb{R}^4$, spanned by $dx^1 \wedge dx^2 \pm dx^3 \wedge dx^4$ and its cyclic partners; $3 + 3 = 6$.	BPST instantons (Chapter 8); signature operator (Chapter 4)
L^2 inner product	$\langle \alpha, \beta \rangle = \int_M \alpha \wedge \star\beta$, positive definite for $s = 0$.	Adjoints, δ, actions (Chapter 2)

Table T1.4: Summary, part 3: metric, orientation and the Hodge star.

Concept	Operational meaning	Used later
Exterior derivative (§T1.5)	$\mathrm{d}\alpha = \mathrm{d}x^\mu \wedge \partial_\mu \alpha$: differentiate the components and wedge on a $\mathrm{d}x$. It is chart-independent, and it is the circulation (or flux) per unit area (or volume).	de Rham cohomology (Nash (1.43), Chapter 1); Stokes (Chapter T2)
grad, curl, div	d on 0-, 1- and 2-forms in $\mathbb{R}^3$; with $\star$, it gives curl and div in any coordinates (spherical div worked out).	Maxwell; Laplacians
$\mathrm{d}^2 = 0$	Mixed partials commute. In 3D it is curl grad = 0 and div curl = 0.	Complexes: de Rham, Dolbeault, BRST (Chapters 4 and 12)
Leibniz rule	$\mathrm{d}(\alpha \wedge \beta) = \mathrm{d}\alpha \wedge \beta + (-1)^p \alpha \wedge \mathrm{d}\beta$.	Chern–Simons forms, descent (Chapter 10)
Closed, exact	Closed: $\mathrm{d}\alpha = 0$. Exact: $\alpha = \mathrm{d}\beta$. $\mathrm{d}\phi$ on $\mathbb{R}^2 \setminus 0$ is closed but not exact; its loop integral is 2π.	H^p_{dR} (Chapters 1 and T2)
Poincaré lemma	On a star-shaped region, closed implies exact. Build β by integrating along rays.	Local triviality of cohomology (Chapter 3)
Codifferential (§T1.6)	The adjoint of d: $\delta = (-1)^{n(p+1)+1+s} \star \mathrm{d} \star$. On 1-forms, $\delta = -\operatorname{div}$.	Nash's d^* (Chapter 2)
Laplace–de Rham	$\Delta = \mathrm{d}\delta + \delta\mathrm{d}$. On functions it is $-\frac{1}{\sqrt{g}}\partial_\mu(\sqrt{g}g^{\mu\nu}\partial_\nu)$, e.g. the polar form $-(f_{rr} + f_r/r + f_{\phi\phi}/r^2)$. It is positive.	Heat kernel and index (Chapters 2 and 4)
Harmonic forms, Hodge	$\Delta h = 0 \Leftrightarrow \mathrm{d}h = \delta h = 0$ on a compact M. There is one harmonic form per class, so $b_p = \dim \operatorname{Ker} \Delta_p$, and $\omega = \mathrm{d}\alpha + \delta\beta + h$.	Nash p. 41 (Chapter 2); zero modes (Chapter 4)
Symbols	Replace $\partial_\mu \to \mathrm{i}k_\mu$: $\sigma(\mathrm{d}) = \mathrm{i}k\wedge$, $\sigma(\delta) = -\mathrm{i}\iota_k$, $\sigma(\Delta) = \lvert k\rvert^2$. So Δ is elliptic.	Elliptic complexes (Nash (2.46)–(2.49), Chapter 2)

Table T1.5: Summary, part 4: the exterior derivative, the codifferential and the Laplacian.

Concept	Operational meaning	Used later
Maxwell in forms (§T1.7)	$F = E \wedge \mathrm{d}t + \star_3 B$. $\mathrm{d}F = 0$ gives $\nabla \cdot B = 0$ and Faraday's law; $\mathrm{d} \star F = \star J$ gives Gauss and Ampère.	Yang–Mills $D \star F = \star J$ (Chapters T3 and 8)
Potential, gauge	$F = \mathrm{d}A$ with $A = -\Phi\,\mathrm{d}t + \vec{A} \cdot \mathrm{d}\vec{x}$. Changing $A \to A + \mathrm{d}\lambda$ leaves F fixed because $\mathrm{d}^2 = 0$. Charge conservation is $\mathrm{d} \star J = \mathrm{d}^2 \star F = 0$.	Connections (Chapter T3); gauge orbits (Chapter 6)
Action, invariants	$S = -\frac{1}{2}\int F \wedge \star F$; $F \wedge F = \mathrm{d}(A \wedge \mathrm{d}A)$ is a total derivative ($\propto \vec{E} \cdot \vec{B}$).	θ-term, Chern–Simons, anomalies (Chapters 10 and 12)
Dirac monopole	$F = \frac{g}{4\pi} \sin\theta\,\mathrm{d}\theta \wedge \mathrm{d}\phi$ is closed but not exact on S^2. The two patch potentials differ by $\frac{g}{2\pi}\mathrm{d}\phi$, which forces $eg \in 2\pi\mathbb{Z}$.	Monopole bundle, c_1 (Chapters 1 and T2)
Push-forward, pullback (§T1.8)	f^*: substitute $y = f(x)$ and expand; vectors are pushed forward by the Jacobian. Pulling back to a curve gives the line-integral integrand.	Functoriality of H^*_{dR} (Chapter 1); gauge transformations
Top forms and Jacobians	$f^*(\mathrm{d}y^1 \wedge \cdots \wedge \mathrm{d}y^n) = \det(\partial f/\partial x)\,\mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^n$: this is the change-of-variables formula.	Integration (Chapter T2); degree of a map
Induced metric	$f^*g = J^{\mathsf{T}} g J$, e.g. $S^2 \subset \mathbb{R}^3$ gets $\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2$.	Nambu–Goto and Polyakov (Chapter 9)
Naturality	$\mathrm{d}f^* = f^*\mathrm{d}$: differentiate before or after the substitution, the answer is the same.	Cohomology is a topological invariant (Chapter 1)
Degree, winding	$\deg f = \int f^*\omega \big/ \int \omega$; for $z \mapsto z^k$ on S^1 it is k.	Instanton number (Chapter 8)

Table T1.6: Summary, part 5: Maxwell's equations and the pullback.

Concept	Operational meaning	Used later
Interior product (§T1.9)	Feed X into the first slot: $(\iota_X\alpha)_{\nu\ldots} = X^\mu\alpha_{\mu\nu\ldots}$. In 3D, $\iota_{\vec v}(\star_3 B)$ is $\vec B\times\vec v$ as a 1-form: the magnetic Lorentz force.	Cartan calculus; equivariant forms (Chapter 12)
Flow	Solve $\dot x^\mu = X^\mu(x)$; the solutions form a one-parameter group ϕ_t of diffeomorphisms.	Diffeomorphism groups (Chapter 6); Morse flows (Chapter 7)
Lie derivative	$\mathcal{L}_X\alpha = \frac{\mathrm{d}}{\mathrm{d}t}\phi_t^*\alpha\big\vert_{t=0}$. In components, $(\mathcal{L}_X\alpha)_\mu = X^\lambda\partial_\lambda\alpha_\mu + \alpha_\lambda\partial_\mu X^\lambda$.	Symmetries, Noether currents
Cartan's formula	$\mathcal{L}_X = \mathrm{d}\iota_X + \iota_X\mathrm{d}$ on forms: a two-line calculation instead of components. It gives $\mathcal{L}_X\mathrm{vol} = (\mathrm{div}\,X)\mathrm{vol}$ and Liouville's theorem.	Homotopy invariance of H^*_{dR} (Chapter T2)
Killing, conformal Killing	$\mathcal{L}_X g = 0$ (isometry), or $\mathcal{L}_X g = \lambda g$ (conformal). In 2D the conformal Killing vectors are the holomorphic ones.	Moduli of Riemann surfaces, Nash (5.71) (Chapters 5 and 9)
Complex coordinates (§T1.10)	$\mathrm{d}z = \mathrm{d}x + \mathrm{i}\mathrm{d}y$, $\partial_{\bar z} = \frac12(\partial_x + \mathrm{i}\partial_y)$; holomorphic means $\bar\partial f = 0$; $\nabla^2 = 4\partial_z\partial_{\bar z}$.	Riemann surfaces, CFT (Chapters 5 and 11)
(p,q)-forms	p factors of $\mathrm{d}z$ and q of $\mathrm{d}\bar z$; the space has dimension $\binom{m}{p}\binom{m}{q}$. The type is global because the transitions are holomorphic.	Nash (3.24) (Chapter 3)
∂, $\bar\partial$	$\mathrm{d} = \partial + \bar\partial$ with $\partial^2 = \bar\partial^2 = 0$ and $\partial\bar\partial = -\bar\partial\partial$. Also $\partial_{\bar z}(1/z) = \pi\delta^2$.	Nash (3.25)–(3.26); Green functions (Chapter 11)
Dolbeault cohomology	$H^{p,q}_{\bar\partial} = \mathrm{Ker}\,\bar\partial/\mathrm{Im}\,\bar\partial$. For $\mathbb{CP}^1$, $h^{0,0} = h^{1,1} = 1$ and $h^{1,0} = h^{0,1} = 0$; for T^2 all four are 1.	Dolbeault theorem (Chapter 3); index $\bar\partial = 1 - g$ (Nash (4.63)–(4.65), Chapter 4)
Kähler, Fubini–Study	$\omega = \frac{\mathrm{i}}{2}g_{i\bar\jmath}\mathrm{d}z^i\wedge\mathrm{d}\bar z^j$ is closed; $\omega_{FS} = \frac{\mathrm{i}}{2\pi}\partial\bar\partial\log(1+\vert z\vert^2)$ integrates to 1. $\Delta_{\bar\partial} = \frac12\Delta$.	Nash (5.29)–(5.32) (Chapter 5)

Table T1.7: Summary, part 6: the Cartan calculus and complex forms.

Five calculations to be able to do from memory. If you can do these without looking, you own the chapter.

1. Show that $\mathrm{d}x \wedge \mathrm{d}y = r\,\mathrm{d}r \wedge \mathrm{d}\phi$, and that the pullback of $\mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z$ to spherical coordinates is $r^2 \sin\theta\,\mathrm{d}r \wedge \mathrm{d}\theta \wedge \mathrm{d}\phi$ (§T1.3, §T1.8).
2. Write the Hodge star of $\mathrm{d}r$ and $\mathrm{d}\phi$ in the plane, and from them the polar Laplacian $\Delta f = -\star\,\mathrm{d}\star\,\mathrm{d}f = -(f_{rr} + f_r/r + f_{\phi\phi}/r^2)$ (§T1.4, §T1.6).
3. Show that $\mathrm{d}\phi = (x\,\mathrm{d}y - y\,\mathrm{d}x)/(x^2 + y^2)$ is closed, and that its integral round the unit circle is 2π, so it is not exact (§T1.5).
4. Derive all four Maxwell equations from $\mathrm{d}F = 0$ and $\mathrm{d}\star F = \star J$ with $F = E \wedge \mathrm{d}t + \star_3 B$, and derive charge conservation from $\mathrm{d}^2 = 0$ (§T1.7).
5. Compute $\mathcal{L}_X(\mathrm{d}x \wedge \mathrm{d}y)$ for $X = x\partial_x + y\partial_y$ with Cartan's formula ($= 2\,\mathrm{d}x \wedge \mathrm{d}y$), and show that ω_{FS} on $\mathbb{CP}^1$ has total integral 1 (§T1.9, §T1.10).

Reading Nash

Nash assumes this whole chapter from his first page. The places where he leans on it hardest are these. p. 13: de Rham cohomology (1.43), and the wedge product as the cup product. pp. 39–41: $\star$, d^* and Δ (2.42)–(2.53), the symbols (2.46)–(2.49), and Hodge's theorem. pp. 63–65: (p,q)-forms, ∂ and $\bar\partial$ (3.24)–(3.26), and the Dolbeault theorem. pp. 103–110: the de Rham and Dolbeault complexes in the index theorem (4.63)–(4.66), and the signature operator built from $\star$ (4.77)–(4.80). pp. 155–160: Kähler forms (5.29)–(5.32) and the Lie derivative of the metric (5.71). Chapters 8 and 10: self-dual fields and the density $\mathrm{tr}\,F \wedge F$.

Chapter T2

Integration, Stokes' Theorem, Connections and Curvature

Where this chapter is going

The previous chapter built the machinery of differential forms: the wedge product, the exterior derivative d with $\mathrm{d}^2 = 0$, the Hodge star, and pullbacks. That machinery was all *local*. It told us how to differentiate at a point. This chapter is about what happens when we *add things up* over a whole region or a whole manifold. Two very different stories come out of that, and the punchline of the chapter is that they turn out to be the same story.

The first story is **integration and topology**. Forms are exactly the objects that can be integrated without choosing coordinates. Once we can integrate, Stokes' theorem, $\int_R \mathrm{d}\alpha = \int_{\partial R} \alpha$, becomes the one tool that does everything: it contains the fundamental theorem of calculus, Green's theorem, the curl theorem and Gauss's theorem. Applied to a region with no edge it gives a detector. A form whose integral over a closed surface is non-zero cannot be exact. The forms that are closed but not exact, such as $\mathrm{d}\theta$ around a puncture or the monopole field on a sphere, measure the holes in the space. Organising them gives de Rham cohomology, the Betti numbers and the Euler characteristic.

The second story is **geometry**. To differentiate a *vector* field, rather than a function, we need to know how the basis vectors turn from point to point. That information is a matrix of 1-forms, the *connection* ω. Its failure to be "pure gauge" is the *curvature* $\Omega = \mathrm{d}\omega + \omega \wedge \omega$. This is exactly the structure of a gauge field, with $F = \mathrm{d}A + A \wedge A$; Chapter T3 makes the identification precise.

The two stories meet in the **Gauss–Bonnet theorem**. The integral of the curvature of a closed surface is 2π times its Euler characteristic,

$$\int_\Sigma K \, \mathrm{d}A = 2\pi \, \chi(\Sigma).$$

The left side needs a metric. The right side needs only a count of vertices, edges and faces. This is the first "integral of curvature equals topology" theorem. Chern–Weil theory generalises it to every gauge field, and the Atiyah–Singer index theorem (Chapter 4) turns it into a statement about zero modes of the Dirac operator. Everything later in these notes, from instanton numbers (Chapter 8) to anomalies (Chapter 10), is a descendant of the computation we do at the end of this chapter.

How to read this chapter. Sections T2.1 and T2.2 are vector calculus rewritten in forms; a physicist will recognise every formula. Sections T2.3 and T2.4 are the first genuinely topological ideas, and they are used everywhere in Nash from Chapter 1 onwards. Sections T2.5–T2.7 follow Dray's treatment of frames and connections, done entirely with forms. Section T2.8 puts

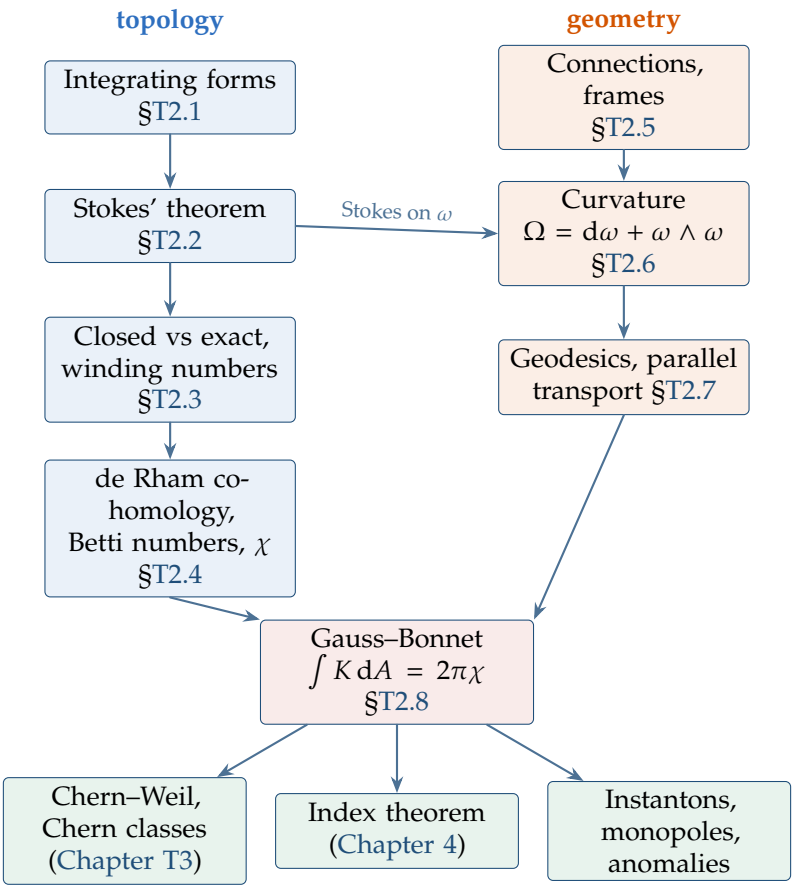


Figure T2.1: The logic of the chapter. The left column is topology: integration leads to Stokes' theorem, which detects closed-but-not-exact forms, which organise into cohomology. The right column is geometry: frames lead to connections, curvature and geodesics. Stokes' theorem applied to the connection is the bridge between them. Both columns meet in Gauss–Bonnet, the prototype of every "curvature integrates to an integer" result used later in physics.

it all together. Section T2.9 collects Dray's closing applications (spinors, the sphere puzzle, the equivalence problem).

Conventions for this chapter.

- Manifolds are smooth, of dimension n, with coordinates x^μ. Surfaces are two-dimensional and Riemannian (Euclidean signature).
- A p-form is $\alpha = \frac{1}{p!}\alpha_{\mu_1\ldots\mu_p}\mathrm{d}x^{\mu_1}\wedge\cdots\wedge\mathrm{d}x^{\mu_p}$.
- The standard orientation of $\mathbb{R}^n$ is $\mathrm{d}x^1\wedge\cdots\wedge\mathrm{d}x^n$. On the sphere we use $\sin\theta\,\mathrm{d}\theta\wedge\mathrm{d}\phi$, which corresponds to the outward normal.
- Boundaries carry the "outward normal first" orientation (Section T2.2). For a planar region this means the boundary is traversed counterclockwise, with the region on the left.
- Orthonormal frames $\{e_i\}$, dual coframes $\{\sigma^i\}$ with $\mathrm{d}\vec{r} = \sigma^i e_i$. Connection 1-forms are defined by $\mathrm{d}e_j = e_i\,\omega^i{}_j$; curvature by $\Omega^i{}_j = \mathrm{d}\omega^i{}_j + \omega^i{}_k\wedge\omega^k{}_j$. These follow Dray.
- Gaussian curvature K is defined by $\Omega^1{}_2 = K\,\sigma^1\wedge\sigma^2$ on an oriented surface; the unit sphere has $K=+1$.

T2.1 Integrating differential forms

In vector calculus we meet three kinds of integral: the line integral $\int_C \vec{F}\cdot\mathrm{d}\vec{r}$, the flux integral $\iint_S \vec{F}\cdot\mathrm{d}\vec{A}$, and the volume integral $\iiint_R f\,\mathrm{d}V$. Each comes with its own little recipe for turning it into an ordinary integral over parameters. Each recipe has a Jacobian or a normal vector hidden in it. We are rarely told what the three have in common.

This section answers one question: *what kind of object is the thing under an integral sign?* The answer is that it is a differential form. A p-form is exactly the object that can be integrated over a p-dimensional oriented surface without choosing coordinates. Once we see this, the three recipes collapse into one: pull the form back to the parameter space and "drop the wedges". Everything later in the chapter, starting with Stokes' theorem, depends on this.

T2.1.1 The integrands you already know

Work in ordinary $\mathbb{R}^3$ with Cartesian coordinates (x,y,z). The **infinitesimal displacement** $\mathrm{d}\vec{r} = \mathrm{d}x\,\hat{x} + \mathrm{d}y\,\hat{y} + \mathrm{d}z\,\hat{z}$ is itself a vector-valued 1-form: feed it a direction and it returns that direction. Dray takes $\mathrm{d}\vec{r}$ as the organising object [Dray §16.1]. With it we can turn any vector field $\vec{F} = F_x\hat{x} + F_y\hat{y} + F_z\hat{z}$ into a 1-form,

$$F \coloneqq \vec{F}\cdot\mathrm{d}\vec{r} = F_x\,\mathrm{d}x + F_y\,\mathrm{d}y + F_z\,\mathrm{d}z. \tag{T2.1}$$

We use the same letter for the vector and the 1-form; the arrow tells them apart. The Hodge star (Chapter T1) turns F into a 2-form,

$$\star F = F_x\,\mathrm{d}y\wedge\mathrm{d}z + F_y\,\mathrm{d}z\wedge\mathrm{d}x + F_z\,\mathrm{d}x\wedge\mathrm{d}y, \tag{T2.2}$$

and a function f into a 3-form, $\star f = f\,\mathrm{d}x\wedge\mathrm{d}y\wedge\mathrm{d}z$.

The claim, which we will check on examples, is the dictionary

$$\int_C \vec{F}\cdot \mathrm{d}\vec{r} = \int_C F, \qquad \iint_S \vec{F}\cdot \mathrm{d}\vec{A} = \int_S \star F, \qquad \iiint_R f\,\mathrm{d}V = \int_R \star f. \tag{T2.3}$$

The circulation of $\vec{F}$ is the integral of the 1-form F. The flux of $\vec{F}$ is the integral of the 2-form $\star F$. The "amount of f" in a region is the integral of the 3-form $\star f = f\,\mathrm{d}V$.

Dictionary: vector calculus integrals as form integrals

vector calculus	form	degree
value $f(P)$ at a point	f	0
circulation $\int_C \vec{F}\cdot \mathrm{d}\vec{r}$	$F = \vec{F}\cdot \mathrm{d}\vec{r}$	1
flux $\iint_S \vec{F}\cdot \mathrm{d}\vec{A}$	$\star F = \vec{F}\cdot \mathrm{d}\vec{A}$	2
total amount $\iiint_R f\,\mathrm{d}V$	$\star f = f\,\mathrm{d}V$	3

Physical picture: a flux form counts field lines through a tile

Think of $\star F$ as a device that is handed a tiny oriented tile of surface and returns the number of field lines of $\vec{F}$ that pierce it, counted with sign. A line integral form F is handed a tiny arrow and returns the work done along it. A volume form is handed a tiny box and returns the amount of stuff inside. Integration is "chop the domain into tiles, feed each tile to the form, add up".

T2.1.2 What a form does to a tiny parallelogram

Here is the mechanism behind the intuition. A p-form at a point is a machine that eats p tangent vectors and returns a number, linearly in each slot and antisymmetrically (Chapter T1). For a 2-form built from two 1-forms,

$$(\alpha\wedge\beta)(\vec{u},\vec{v}) = \alpha(\vec{u})\,\beta(\vec{v}) - \alpha(\vec{v})\,\beta(\vec{u}). \tag{T2.4}$$

Take $\alpha = \mathrm{d}x$, $\beta = \mathrm{d}y$ and two vectors in the plane, $\vec{u} = (u_1, u_2)$ and $\vec{v} = (v_1, v_2)$. Then

$$\begin{aligned}(\mathrm{d}x\wedge\mathrm{d}y)(\vec{u},\vec{v}) &\overset{(1)}{=} \mathrm{d}x(\vec{u})\,\mathrm{d}y(\vec{v}) - \mathrm{d}x(\vec{v})\,\mathrm{d}y(\vec{u})\\ &\overset{(2)}{=} u_1v_2 - v_1u_2\\ &\overset{(3)}{=} \det\begin{pmatrix} u_1 & v_1\\ u_2 & v_2\end{pmatrix}.\end{aligned} \tag{T2.5}$$

(1) The definition (T2.4) of the wedge of two 1-forms acting on two vectors.
(2) $\mathrm{d}x$ picks out the x-component of a vector and $\mathrm{d}y$ the y-component.
(3) This is the determinant whose absolute value is the area of the parallelogram with edges $\vec{u}, \vec{v}$; its sign says whether $\vec{v}$ is counterclockwise from $\vec{u}$.

So $\mathrm{d}x\wedge\mathrm{d}y$ returns the *signed area* of the parallelogram. That is the whole reason forms are integrands: an integral is a sum of (value of the integrand) × (size of a small piece), and a form supplies both factors at once, with a sign that knows about orientation.

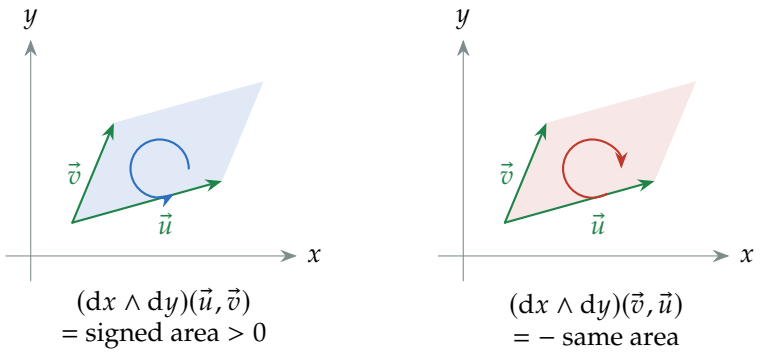


Figure T2.2: A 2-form eats the two edge vectors of a small parallelogram and returns its signed area. Swapping the edges reverses the circulation sense (the small arrow) and flips the sign. The sign is the orientation; this is why integrands must be antisymmetric.

T2.1.3 Why the wedge product does the Jacobian for you

The acid test of an integrand is how it behaves under a change of variables. In ordinary calculus we are told to insert $|\det J|$ by hand:

$$\iint f\,dx\,dy = \iint f \left|\det \frac{\partial(x,y)}{\partial(u,v)}\right| du\,dv.$$

With forms nothing has to be inserted. The Jacobian appears by itself. Write $x = x(u,v)$, $y = y(u,v)$. Then

$$\begin{aligned} dx \wedge dy &\overset{(1)}{=} \left(\frac{\partial x}{\partial u}du + \frac{\partial x}{\partial v}dv\right) \wedge \left(\frac{\partial y}{\partial u}du + \frac{\partial y}{\partial v}dv\right) \\ &\overset{(2)}{=} \frac{\partial x}{\partial u}\frac{\partial y}{\partial v}\,du \wedge dv + \frac{\partial x}{\partial v}\frac{\partial y}{\partial u}\,dv \wedge du \\ &\overset{(3)}{=} \left(\frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u}\right) du \wedge dv \\ &= \det \frac{\partial(x,y)}{\partial(u,v)}\,du \wedge dv. \end{aligned} \tag{T2.6}$$

(1) The chain rule for the differentials of x and y.
(2) Expand; the terms $du \wedge du$ and $dv \wedge dv$ vanish by antisymmetry.
(3) Antisymmetry again: $dv \wedge du = -du \wedge dv$.

The determinant is the Jacobian, and it arrives *with its sign*. The sign is what we want: it records whether the new coordinates have the same orientation as the old ones. For orientation-preserving changes of variables ($\det J > 0$) the form integral and the ordinary integral agree.

Worked example T2.1: Polar coordinates

With $x = r\cos\phi$, $y = r\sin\phi$,

$$\begin{aligned} dx \wedge dy &\overset{(1)}{=} (\cos\phi\,dr - r\sin\phi\,d\phi) \wedge (\sin\phi\,dr + r\cos\phi\,d\phi) \\ &\overset{(2)}{=} r\cos^2\phi\,dr \wedge d\phi - r\sin^2\phi\,d\phi \wedge dr \\ &\overset{(3)}{=} r(\cos^2\phi + \sin^2\phi)\,dr \wedge d\phi = r\,dr \wedge d\phi. \end{aligned}$$

(1) Differentiate x and y.
(2) Only the cross terms survive.

(3) Flip $\mathrm{d}\phi \wedge \mathrm{d}r = -\mathrm{d}r \wedge \mathrm{d}\phi$ and use $\cos^2 + \sin^2 = 1$.

The familiar "$r\,\mathrm{d}r\,\mathrm{d}\phi$" is not a rule to memorise. It is what $\mathrm{d}x \wedge \mathrm{d}y$ *is* in polar coordinates. Notice that (r, ϕ) is positively oriented relative to (x, y) since $r > 0$.

Watch out: the symbol $\mathrm{d}x\,\mathrm{d}y$ is not $\mathrm{d}x \wedge \mathrm{d}y$

In an ordinary double integral, $\mathrm{d}x\,\mathrm{d}y$ means "Lebesgue measure", which is positive and symmetric: $\mathrm{d}x\,\mathrm{d}y = \mathrm{d}y\,\mathrm{d}x$. The form $\mathrm{d}x \wedge \mathrm{d}y$ is antisymmetric. They agree only after we fix an orientation and agree to write the form with its coordinates in positive order. This is the source of every stray minus sign in flux integrals.

T2.1.4 The recipe: pull back and drop the wedges

We now state the general rule. A **parametrised p-surface** in a manifold M is a smooth map $\varphi : U \to M$ from a region $U \subset \mathbb{R}^p$ with coordinates $(u^1, \dots, u^p)$. The **pullback** $\varphi^*\alpha$ of a form (Chapter T1) is obtained by substituting $x^\mu = x^\mu(u)$ and $\mathrm{d}x^\mu = \frac{\partial x^\mu}{\partial u^a}\mathrm{d}u^a$. For a p-form on M the pullback is a top-degree form on U, so it is a single function times $\mathrm{d}u^1 \wedge \cdots \wedge \mathrm{d}u^p$.

Working definition T2.1 (integral of a p-form)

To integrate a p-form α over an oriented p-surface S [a surface on which one order of the coordinate directions has been declared positive]:

1. Choose a parametrisation $\varphi : U \to S$ whose coordinate order $(u^1, \dots, u^p)$ agrees with the orientation of S.
2. Pull back: substitute $x^\mu(u)$ and $\mathrm{d}x^\mu = \partial_a x^\mu\,\mathrm{d}u^a$, and simplify to $\varphi^*\alpha = g(u)\,\mathrm{d}u^1 \wedge \cdots \wedge \mathrm{d}u^p$.
3. Drop the wedges: $\displaystyle\int_S \alpha := \int_U g(u)\,\mathrm{d}u^1 \cdots \mathrm{d}u^p$, an ordinary multiple integral.

If the parametrisation has the wrong orientation [its coordinate order is the one declared negative, e.g. $(u^2, u^1, \dots)$], insert a minus sign. *Formally:* $\int_{\varphi(U)} \alpha = \int_U \varphi^*\alpha$.

Dray puts the whole of this in one sentence: "To integrate a differential form, simply drop the wedge products and do the integral — and check whether the orientation agrees with the one you want, inserting a minus sign if it doesn't" [Dray §20.5].

Why the answer does not depend on the parametrisation. Suppose $\varphi : U \to S$ and $\tilde\varphi : \tilde U \to S$ are two orientation-compatible parametrisations. They are related by a change of parameters $\psi : \tilde U \to U$ with $\tilde\varphi = \varphi \circ \psi$ and $\det(\partial u/\partial \tilde u) > 0$. Write $\varphi^*\alpha = g\,\mathrm{d}^p u$. Then

$$\begin{aligned}\int_{\tilde U} \tilde\varphi^*\alpha &\overset{(1)}{=} \int_{\tilde U} \psi^*\varphi^*\alpha \\ &\overset{(2)}{=} \int_{\tilde U} g(u(\tilde u))\,\det\!\left(\frac{\partial u}{\partial \tilde u}\right)\mathrm{d}^p\tilde u \\ &\overset{(3)}{=} \int_U g(u)\,\mathrm{d}^p u \;=\; \int_U \varphi^*\alpha. \end{aligned} \tag{T2.7}$$

(1) Pullbacks compose backwards: $(\varphi\circ\psi)^* = \psi^*\varphi^*$.

(2) The p-dimensional version of (T2.6): $\mathrm{d}u^1 \wedge \cdots \wedge \mathrm{d}u^p = \det(\partial u/\partial\tilde u)\,\mathrm{d}\tilde u^1 \wedge \cdots \wedge \mathrm{d}\tilde u^p$.

(3) The change-of-variables formula of ordinary calculus, which uses $|\det|$; here $\det > 0$, so $|\det| = \det$.

This is the precise sense in which forms are "the natural integrands". A function alone cannot be integrated over a surface without extra data (a measure). A vector field cannot either (it needs a normal). A p-form over a p-surface needs only an orientation.

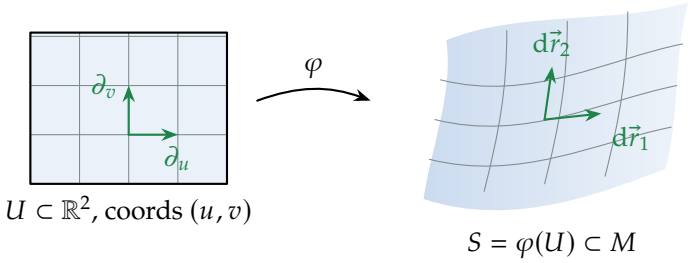


Figure T2.3: Pulling back. The parameter rectangle is chopped into little squares; φ carries each one to a little curved parallelogram on S with edges $d\vec{r}_1 = \partial_u \vec{r}\, du$ and $d\vec{r}_2 = \partial_v \vec{r}\, dv$ (Dray's Figure 16.1). The form is evaluated on those edges. The pullback $\varphi^*\alpha$ is the form on U that gives the same answer on the original little squares.

Worked example T2.2: Circulation around a circle

Let $\alpha = -y\,dx + x\,dy$ and let C be the circle of radius R, traversed counterclockwise. Parametrise by $x = R\cos t$, $y = R\sin t$, $t \in [0, 2\pi]$. Then

$$\begin{aligned}\int_C \alpha &\overset{(1)}{=} \int_0^{2\pi} \big(-R\sin t\,(-R\sin t\,dt) + R\cos t\,(R\cos t\,dt)\big) \\ &\overset{(2)}{=} \int_0^{2\pi} R^2\,dt = 2\pi R^2.\end{aligned}$$

(1) Pull back: $dx = -R\sin t\,dt$ and $dy = R\cos t\,dt$.
(2) $\sin^2 t + \cos^2 t = 1$; then drop the dt and integrate.

The answer is twice the enclosed area. That is no accident: we will see in Example T2.7 that Stokes' theorem turns it into $2\int_D dx \wedge dy$. Traversing the circle clockwise ($t \mapsto -t$) gives $-2\pi R^2$.

Worked example T2.3: Flux through a graph (Dray 16.13–16.18)

Let S be the graph $z = h(x, y)$ over a region D of the xy-plane, oriented by the upward normal. Use (x, y) as parameters. On S, $dz = h_x\,dx + h_y\,dy$ (subscripts are partial derivatives). The pullback of $\star F$ is

$$\begin{aligned}\star F\big|_S &\overset{(1)}{=} F_x\,dy \wedge (h_x dx + h_y dy) + F_y\,(h_x dx + h_y dy) \wedge dx + F_z\,dx \wedge dy \\ &\overset{(2)}{=} F_x h_x\,dy \wedge dx + F_y h_y\,dy \wedge dx + F_z\,dx \wedge dy \\ &\overset{(3)}{=} \big(-F_x h_x - F_y h_y + F_z\big)\,dx \wedge dy.\end{aligned}$$

(1) Substitute dz into (T2.2).
(2) Drop $dy \wedge dy = 0 = dx \wedge dx$.
(3) Flip $dy \wedge dx = -dx \wedge dy$.

This is exactly $\vec{F} \cdot d\vec{A}$ with the textbook area element $d\vec{A} = (-h_x, -h_y, 1)\,dx\,dy$, which one gets from $d\vec{r}_1 \times d\vec{r}_2$ with $d\vec{r}_1 = (\hat{x} + h_x\hat{z})dx$ and $d\vec{r}_2 = (\hat{y} + h_y\hat{z})dy$. The form computation did the cross product for us.

A number. Take $\vec{F} = \vec{r} = (x, y, z)$ and the paraboloid $h = 1 - x^2 - y^2$ over the unit disc. Then $h_x = -2x$, $h_y = -2y$ and

$$\begin{aligned}\int_S \star F &\overset{(1)}{=} \int_D \left(2x^2 + 2y^2 + 1 - x^2 - y^2\right) \mathrm{d}x\,\mathrm{d}y \\ &\overset{(2)}{=} \int_0^{2\pi}\int_0^1 (1+\rho^2)\,\rho\,\mathrm{d}\rho\,\mathrm{d}\phi \\ &= 2\pi\left(\tfrac{1}{2} + \tfrac{1}{4}\right) = \frac{3\pi}{2}.\end{aligned}$$

(1) The formula above with $F_x = x$, $F_y = y$, $F_z = z = 1 - x^2 - y^2$.
(2) Polar coordinates on the disc, $\mathrm{d}x \wedge \mathrm{d}y = \rho\,\mathrm{d}\rho \wedge \mathrm{d}\phi$.

We will recover $3\pi/2$ from Gauss's theorem in Example T2.9.

Worked example T2.4: Area of the sphere and volume of the 3-sphere

On the sphere of radius a with coordinates (θ, ϕ) the orthonormal coframe is $\sigma^\theta = a\,\mathrm{d}\theta$, $\sigma^\phi = a\sin\theta\,\mathrm{d}\phi$ and the area form is $\sigma^\theta \wedge \sigma^\phi$:

$$\begin{aligned}\int_{S^2_a} \sigma^\theta \wedge \sigma^\phi &\overset{(1)}{=} \int_0^{2\pi}\int_0^{\pi} a^2 \sin\theta\,\mathrm{d}\theta\,\mathrm{d}\phi \\ &\overset{(2)}{=} 2\pi a^2\left[-\cos\theta\right]_0^{\pi} = 4\pi a^2.\end{aligned}$$

(1) (θ, ϕ) is positively oriented (outward normal) and covers the sphere once apart from a set of zero area (the poles and one meridian).
(2) The ϕ integral is trivial.

Dray's Problem 16.1 does the same in one dimension more. The 3-sphere of radius r has line element $\mathrm{d}s^2 = r^2\left(\mathrm{d}\psi^2 + \sin^2\psi(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)\right)$, with $\psi, \theta \in [0, \pi]$ and $\phi \in [0, 2\pi)$. The volume form is the product of the orthonormal coframe, $r\,\mathrm{d}\psi \wedge r\sin\psi\,\mathrm{d}\theta \wedge r\sin\psi\sin\theta\,\mathrm{d}\phi$, so

$$\begin{aligned}\mathrm{Vol}(S^3_r) &\overset{(1)}{=} r^3 \int_0^{\pi} \sin^2\psi\,\mathrm{d}\psi \int_0^{\pi} \sin\theta\,\mathrm{d}\theta \int_0^{2\pi} \mathrm{d}\phi \\ &\overset{(2)}{=} r^3 \cdot \frac{\pi}{2} \cdot 2 \cdot 2\pi = 2\pi^2 r^3.\end{aligned}$$

(1) Drop the wedges; the integrand factorises.
(2) $\int_0^{\pi} \sin^2\psi\,\mathrm{d}\psi = \pi/2$ (average of $\sin^2$ is $\tfrac{1}{2}$).

The number $2\pi^2$ will reappear as the volume of $\mathrm{SU}(2) \cong S^3$ when we normalise instanton numbers in Chapter 8.

T2.1.5 Orientation

Every formula above needed an orientation. We now say what that is.

Working definition T2.2 (orientation)

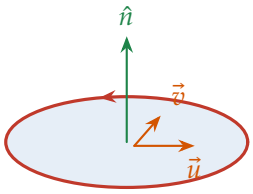
An **orientation** of an n-manifold M is a choice of nowhere-vanishing n-form vol [locally $h\,\mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^n$ with $h \neq 0$ at every point], where two choices that differ by a positive

function count as the same. A chart $(x^1, \dots, x^n)$ is *positively oriented* if $\mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^n = f\,\mathrm{vol}$ with $f > 0$. *Operational test:* M is orientable if you can cover it by charts whose pairwise Jacobians [the matrices $\partial\tilde{x}^\mu/\partial x^\nu$ on the overlap of two charts] all have positive determinant.

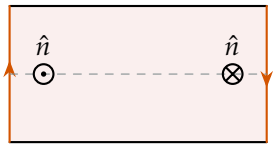
In low dimensions this is all familiar:

- A curve is oriented by a direction of travel.
- A surface in $\mathbb{R}^3$ is oriented by a continuous choice of unit normal $\hat{n}$. The two are related by the right-hand rule: a pair of tangent vectors $(\vec{u}, \vec{v})$ is positive if $\vec{u} \times \vec{v}$ points along $\hat{n}$.
- A region of $\mathbb{R}^3$ is oriented by $\mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z$ (right-handed axes).
- A point is "oriented" by a sign, + or −. Integrating a 0-form over an oriented point means evaluating it and attaching the sign: $\int_{\pm P} f = \pm f(P)$ [Dray (16.23)].

Reversing the orientation of the domain reverses the sign of every integral over it. We write $-S$ for S with the opposite orientation, so $\int_{-S} \alpha = -\int_S \alpha$.



oriented surface: $\vec{u} \times \vec{v} \parallel \hat{n}$



Möbius band: glue the orange edges matching arrows (with a half twist)

Figure T2.4: Left: an orientation of a surface is a consistent choice of normal or, equivalently, of a positive order for pairs of tangent vectors. Right: a Möbius band is a strip whose ends are glued with a half twist (orange arrows). A normal pointing out of the page (⊙) carried along the core line arrives at the glued edge pointing into the page (⊗), so no consistent choice exists. Non-orientable surfaces cannot integrate 2-forms.

Watch out: the poles and the seam do not matter

The coordinates (θ, ϕ) fail at the poles and are discontinuous along the meridian $\phi = 0 \sim 2\pi$. Integrals do not care: sets of lower dimension have zero p-volume. One chart that covers everything except such a set is enough to compute an integral. (It is *not* enough to decide whether a form is globally defined; see Section T2.3.6.)

T2.1.6 Integration on a whole manifold

When no single chart covers M up to a negligible set, we cut the integrand into pieces. A **partition of unity** subordinate to an open cover $\{U_a\}$ is a family of smooth functions $\rho_a \geq 0$ with ρ_a vanishing outside U_a and $\sum_a \rho_a = 1$ everywhere. Then

$$\int_M \alpha := \sum_a \int_{U_a} \rho_a\, \alpha, \tag{T2.8}$$

and each term is computed in the chart of U_a by the recipe above. The result does not depend on the cover or the partition, by the same change-of-variables argument (T2.7).

Worked example T2.5: Two charts on the circle

Cover S^1 by $U_1 = S^1 \setminus \{(1,0)\}$ with angle $\phi_1 \in (0, 2\pi)$, and $U_2 = S^1 \setminus \{(-1,0)\}$ with angle $\phi_2 \in (-\pi, \pi)$. On the overlap, $\phi_1 = \phi_2$ on the upper half and $\phi_1 = \phi_2 + 2\pi$ on the lower half; either way $\mathrm{d}\phi_1 = \mathrm{d}\phi_2$. So there is a single, globally defined 1-form, which we call $\mathrm{d}\phi$, even though there is no global function ϕ. Choose $\rho_1 = \sin^2(\phi/2)$ and $\rho_2 = \cos^2(\phi/2)$; these are smooth on the circle, sum to 1, and ρ_1 vanishes at $(1,0)$, ρ_2 at $(-1,0)$. Then

$$\int_{S^1} \mathrm{d}\phi \overset{(1)}{=} \int_0^{2\pi} \sin^2(\phi_1/2)\,\mathrm{d}\phi_1 + \int_{-\pi}^{\pi} \cos^2(\phi_2/2)\,\mathrm{d}\phi_2 \overset{(2)}{=} \pi + \pi = 2\pi.$$

(1) Definition (T2.8); in each chart the form is $\mathrm{d}\phi_a$.
(2) Each integrand averages to $\frac{1}{2}$ over an interval of length 2π.

Of course one chart would have done. The point of the exercise is the remark in the middle: $\mathrm{d}\phi$ is a perfectly good form that is not d of any function. That observation drives Section T2.3.

Worked example T2.6: The flux of a point charge is independent of the radius

For $\vec{E} = \dfrac{q}{4\pi}\dfrac{\hat{r}}{r^2}$ the flux 2-form is $\star E = \dfrac{q}{4\pi r^2}\,\sigma^\theta \wedge \sigma^\phi$ with $\sigma^\theta \wedge \sigma^\phi = r^2 \sin\theta\,\mathrm{d}\theta \wedge \mathrm{d}\phi$ on the sphere of radius r. Thus

$$\int_{S^2_r} \star E \overset{(1)}{=} \frac{q}{4\pi}\int_0^{2\pi}\int_0^{\pi} \sin\theta\,\mathrm{d}\theta\,\mathrm{d}\phi \overset{(2)}{=} q.$$

(1) The r^2 in the field cancels the r^2 in the area form; in fact $\star E = \frac{q}{4\pi}\sin\theta\,\mathrm{d}\theta \wedge \mathrm{d}\phi$ everywhere on $\mathbb{R}^3 \setminus \{0\}$, with no r in it at all.
(2) The integral of $\sin\theta\,\mathrm{d}\theta \wedge \mathrm{d}\phi$ is 4π (Example T2.4 with $a = 1$).

The radius drops out. Physicists say "Gauss's law". In the language of the next sections: $\star E$ is a closed 2-form on $\mathbb{R}^3 \setminus \{0\}$ that is not exact, and its integral is the charge.

Try it: a cylinder

Integrate $\star F$ for $\vec{F} = (x, y, 0)$ over the curved side of the cylinder $x^2 + y^2 = a^2$, $0 \le z \le L$, oriented outward. *Answer:* with parameters (ϕ, z) one finds $x\,\mathrm{d}y \wedge \mathrm{d}z + y\,\mathrm{d}z \wedge \mathrm{d}x \mapsto a^2(\cos^2\phi + \sin^2\phi)\,\mathrm{d}\phi \wedge \mathrm{d}z$, so the flux is $2\pi a^2 L$. (Check the orientation: $\partial_\phi \times \partial_z$ points outward.)

Reading Nash (integrals everywhere, definitions nowhere)

Nash never explains what it means to integrate a form; he simply writes $\int_X e(X)$ [Nash p. 104, (4.53)], $\int_M \mathrm{tr}(F \wedge \mathrm{d}_A\phi)$ [Nash p. 244, (8.148)] or $\int_{S^2_\infty} \mathrm{tr}(F\phi)$ [Nash p. 245, (8.154)]. Every one of these is computed by the recipe of this section: pull back to parameters, drop the wedges, watch the orientation. When he integrates over "the two sphere at infinity" S^2_∞, he means a large sphere with the outward orientation, the one that makes Stokes' theorem (next section) hold with a plus sign.

Recap

- A p-form is the natural object to integrate over an oriented p-surface; it eats the edge vectors of small parallelograms and returns signed p-volumes.
- Circulation, flux and volume integrals are the integrals of F, $\star F$ and $\star f$ (T2.3).
- Recipe: pull back along a positively oriented parametrisation and drop the wedges. The wedge product supplies the Jacobian with its sign, so the answer is parametrisation-independent.
- Orientation reversal flips the sign. Non-orientable manifolds (Möbius band) cannot integrate top forms.
- $\mathrm{d}\phi$ on S^1 is a global 1-form with $\int_{S^1} \mathrm{d}\phi = 2\pi$, although ϕ is not a global function.

T2.2 Stokes' theorem

Vector calculus has three "big theorems": the gradient theorem, the curl theorem (usually called Stokes' theorem) and the divergence theorem. All three say the same thing: *the integral of a derivative over a region equals the integral of the undifferentiated thing over the edge of the region*. The fundamental theorem of calculus, $\int_a^b f'(x)\,\mathrm{d}x = f(b) - f(a)$, is the one-dimensional case.

With forms the three theorems, and their analogues in every dimension, become a single line. This section states that line, explains why it is true, unpacks every special case, and then draws out the consequences that physicists use daily: integration by parts, boundary terms in action principles, and the fact that some functions and fields are forced to vanish somewhere. It is the most important single formula in these notes.

T2.2.1 Boundaries and their orientation

A $(p+1)$-dimensional region R (a compact manifold with boundary) has an edge ∂R, which is a p-dimensional manifold with no edge of its own. Before Stokes' theorem can be stated we need to know how an orientation of R induces one on ∂R.

Working definition T2.3 (boundary orientation: outward normal first)

Let R be oriented and let $\hat{N}$ be a tangent vector at a point of ∂R pointing *out* of R. A list of tangent vectors $(\vec{v}_1, \ldots, \vec{v}_p)$ to ∂R is positively oriented if and only if $(\hat{N}, \vec{v}_1, \ldots, \vec{v}_p)$ is positively oriented for R [the volume form of R evaluated on this list is > 0].

Unpacking this in the cases we will use:

- $R = [a, b] \subset \mathbb{R}$, oriented left to right. At b the outward direction is $+\hat{x}$, which is positive; at a it is $-\hat{x}$, which is negative. So $\partial[a, b] = \{b\} - \{a\}$: the point b with a plus sign and a with a minus sign.
- R a region in the plane, oriented by $\mathrm{d}x \wedge \mathrm{d}y$. At the right-hand edge of a disc, $\hat{N} = \hat{x}$ and $(\hat{x}, \hat{y})$ is positive, so $\hat{y}$ (upwards) is the positive direction on the boundary. The boundary goes *counterclockwise*, keeping R on the left.
- A region with a hole (an annulus): the outer circle goes counterclockwise, the inner circle *clockwise*. Both keep the region on the left.

- R a surface in $\mathbb{R}^3$ oriented by a normal $\hat{n}$: walk along ∂R with your head along $\hat{n}$; R is on your left. This is the right-hand rule of the curl theorem.
- R a solid region in $\mathbb{R}^3$: ∂R is oriented by the outward normal.

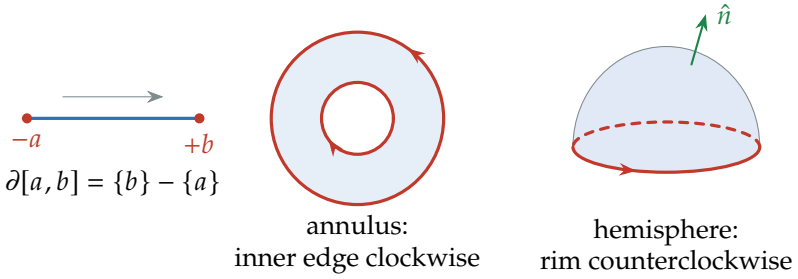


Figure T2.5: Induced boundary orientations. Left: the boundary of an interval is its endpoints with signs. Middle: the boundary of an annulus is two circles, each traversed so that the region lies to the left, so the inner one runs clockwise. Right: for a "butterfly net" surface with upward normal the rim runs counterclockwise seen from above.

T2.2.2 The theorem

Key idea: Stokes' theorem

For every p-form α and every compact oriented $(p+1)$-dimensional region R with boundary ∂R carrying the induced orientation,

$$\int_R d\alpha = \int_{\partial R} \alpha. \tag{T2.9}$$

In words: integrating a derivative over a region only sees the edge.

Dray arrives at (T2.9) by writing the three vector theorems in forms and noticing they have the same shape [Dray (16.28)–(16.33)]. He remarks that "the only real subtlety is keeping track of the relative orientations of R and its boundary", which is why we fixed that first.

T2.2.3 Why it is true: the square, then gluing

The mechanism is the fundamental theorem of calculus, applied one direction at a time. We do the case of a 1-form on a square in full; the general case is the same bookkeeping with more indices.

Let $R = [0,1] \times [0,1]$ in the (x,y)-plane and $\alpha = P\,dx + Q\,dy$. First the left side of (T2.9):

$$\begin{aligned}
\int_R d\alpha &\overset{(1)}{=} \int_R \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dx \wedge dy \\
&\overset{(2)}{=} \int_0^1 \left[\int_0^1 \frac{\partial Q}{\partial x}\,dx\right] dy - \int_0^1 \left[\int_0^1 \frac{\partial P}{\partial y}\,dy\right] dx \\
&\overset{(3)}{=} \int_0^1 \big(Q(1,y) - Q(0,y)\big)\,dy - \int_0^1 \big(P(x,1) - P(x,0)\big)\,dx.
\end{aligned} \tag{T2.10}$$

(1) $d\alpha = \partial_y P\, dy \wedge dx + \partial_x Q\, dx \wedge dy$, and $dy \wedge dx = -dx \wedge dy$.

(2) Drop the wedge (the square is positively oriented by (x, y)) and do each term as an iterated integral, choosing the order so that the derivative is integrated first.
(3) The ordinary fundamental theorem of calculus, once in x and once in y.

Now the right side. The boundary consists of four edges, traversed counterclockwise: bottom ($y = 0$, x from 0 to 1), right ($x = 1$, y from 0 to 1), top ($y = 1$, x from 1 to 0), left ($x = 0$, y from 1 to 0). On a horizontal edge $\mathrm{d}y = 0$ and on a vertical edge $\mathrm{d}x = 0$. So

$$\begin{aligned}\int_{\partial R} \alpha &\overset{(1)}{=} \int_0^1 P(x,0)\,\mathrm{d}x + \int_0^1 Q(1,y)\,\mathrm{d}y + \int_1^0 P(x,1)\,\mathrm{d}x + \int_1^0 Q(0,y)\,\mathrm{d}y \\ &\overset{(2)}{=} \int_0^1 \big(Q(1,y) - Q(0,y)\big)\mathrm{d}y - \int_0^1 \big(P(x,1) - P(x,0)\big)\mathrm{d}x, \end{aligned} \tag{T2.11}$$

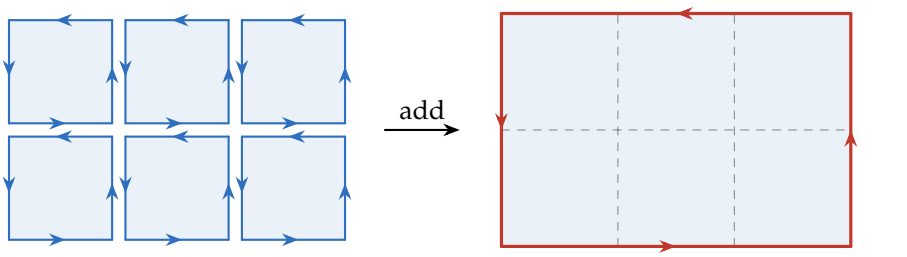
(1) Pull α back to each edge; only one of $\mathrm{d}x$, $\mathrm{d}y$ survives on each.
(2) Reverse the limits on the top and left edges, which costs a sign.

which is (T2.10). Stokes' theorem holds for the square.

Gluing. Any region can be chopped into many small squares (or, on a curved surface, small images of squares under parametrisations). Apply (T2.9) to each and add. On the left we get $\int_R \mathrm{d}\alpha$. On the right, every interior edge is shared by two squares and is traversed in opposite directions by them, so its contributions cancel. Only the outer edges survive, and together they make ∂R with its induced orientation. This is Figure T2.6.

Curved regions. For a region that is the image of a square under φ, we use that pullback commutes with d (Chapter T1), $\varphi^*\mathrm{d}\alpha = \mathrm{d}\varphi^*\alpha$, and that φ maps the boundary of the square to the boundary of the region. Then $\int_{\varphi(R)} \mathrm{d}\alpha = \int_R \mathrm{d}(\varphi^*\alpha) = \int_{\partial R} \varphi^*\alpha = \int_{\partial\varphi(R)} \alpha$.

Higher dimensions. For an $(n-1)$-form on the unit n-cube, each term looks like $f\,\mathrm{d}x^2\wedge\cdots\wedge\mathrm{d}x^n$ (up to relabelling). Its derivative is $\partial_1 f\,\mathrm{d}x^1\wedge\cdots\wedge\mathrm{d}x^n$. Integrating in x^1 first gives $f|_{x^1=1} - f|_{x^1=0}$, the contributions of the two faces perpendicular to x^1, with the signs dictated by "outward normal first". All other faces have some $\mathrm{d}x^j = 0$ for $j \geq 2$ and contribute nothing to this term. The rest is gluing.



each little square: counterclockwise interior edges cancel; only ∂R survives

Figure T2.6: The gluing argument. Each small square satisfies Stokes' theorem by the fundamental theorem of calculus. Adjacent squares traverse their common edge in opposite directions, so interior contributions cancel in pairs. What remains is the circulation around the outer boundary.

T2.2.4 All the special cases

We now read off the classical theorems. In $\mathbb{R}^3$ we need the translation of d into grad, curl and div (Chapter T1); we re-derive the two that matter. For $F = F_x dx + F_y dy + F_z dz$,

$$\begin{aligned} dF &\overset{(1)}{=} (\partial_y F_x\, dy + \partial_z F_x\, dz) \wedge dx + (\partial_x F_y\, dx + \partial_z F_y\, dz) \wedge dy + (\partial_x F_z\, dx + \partial_y F_z\, dy) \wedge dz \\ &\overset{(2)}{=} (\partial_y F_z - \partial_z F_y)\, dy \wedge dz + (\partial_z F_x - \partial_x F_z)\, dz \wedge dx + (\partial_x F_y - \partial_y F_x)\, dx \wedge dy \\ &\overset{(3)}{=} \star(\vec{\nabla} \times \vec{F}) = (\vec{\nabla} \times \vec{F}) \cdot d\vec{A}, \end{aligned} \tag{T2.12}$$

(1) $d(F_x\, dx) = dF_x \wedge dx$ and similarly; the $\partial_x F_x\, dx \wedge dx$ terms vanish.
(2) Collect terms, putting each 2-form in cyclic order $dy \wedge dz$, $dz \wedge dx$, $dx \wedge dy$.
(3) Compare with (T2.2): the coefficients are the Cartesian components of the curl.

and, applying d to $\star F$,

$$\begin{aligned} d\star F &\overset{(1)}{=} \partial_x F_x\, dx \wedge dy \wedge dz + \partial_y F_y\, dy \wedge dz \wedge dx + \partial_z F_z\, dz \wedge dx \wedge dy \\ &\overset{(2)}{=} (\vec{\nabla} \cdot \vec{F})\, dx \wedge dy \wedge dz = (\vec{\nabla} \cdot \vec{F})\, dV. \end{aligned} \tag{T2.13}$$

(1) Only the derivative in the missing direction survives in each term.
(2) Cyclic permutations of three 1-forms are even, so all three 3-forms equal $dx \wedge dy \wedge dz$.

Key idea: the four classical theorems are one theorem

p	α	$\int_R d\alpha = \int_{\partial R} \alpha$ reads	name
0	f	$\int_C \vec{\nabla} f \cdot d\vec{r} = f(B) - f(A)$	FTC / gradient theorem
1	$P dx + Q dy$	$\iint_D (\partial_x Q - \partial_y P)\, dx\, dy = \oint_{\partial D} P dx + Q dy$	Green
1	F	$\iint_S (\vec{\nabla} \times \vec{F}) \cdot d\vec{A} = \oint_{\partial S} \vec{F} \cdot d\vec{r}$	curl (Kelvin–Stokes)
2	$\star F$	$\iiint_V \vec{\nabla} \cdot \vec{F}\, dV = \oiint_{\partial V} \vec{F} \cdot d\vec{A}$	Gauss / divergence

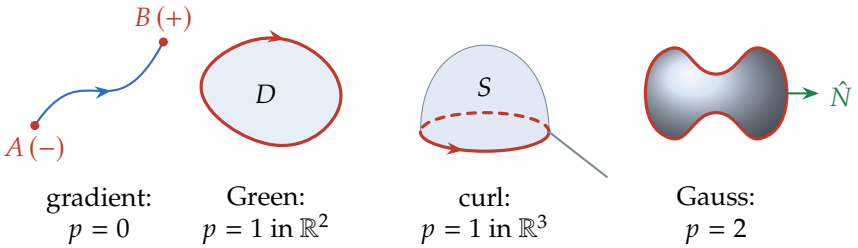


Figure T2.7: The four geometries of Stokes' theorem, in increasing dimension: a curve and its two signed endpoints; a planar region and its counterclockwise rim; Dray's "butterfly net" surface and its rim; a solid "dumbbell" and its outward-oriented surface. The boundary is always drawn in red.

Worked example T2.7: Area from a line integral (Green)

Take $\alpha = \frac{1}{2}(x\, dy - y\, dx)$. Then $d\alpha = \frac{1}{2}(dx \wedge dy - dy \wedge dx) = dx \wedge dy$, so Green's theorem says $\text{Area}(D) = \frac{1}{2}\oint_{\partial D}(x\, dy - y\, dx)$. This is how a planimeter works. For the ellipse

$x = a\cos t$, $y = b\sin t$:

$$\text{Area} \overset{(1)}{=} \tfrac{1}{2}\int_0^{2\pi} \big(a\cos t \cdot b\cos t - b\sin t \cdot (-a\sin t)\big)\,\mathrm{d}t$$
$$\overset{(2)}{=} \tfrac{1}{2}\int_0^{2\pi} ab\,\mathrm{d}t = \pi ab.$$

(1) Pull back: $\mathrm{d}x = -a\sin t\,\mathrm{d}t$, $\mathrm{d}y = b\cos t\,\mathrm{d}t$.
(2) $\cos^2 + \sin^2 = 1$.

For $a = b = R$ this is Example T2.2 divided by 2, as promised.

Worked example T2.8: The curl theorem on a hemisphere, and why the surface does not matter

Let $\alpha = -y\,\mathrm{d}x + x\,\mathrm{d}y$ in $\mathbb{R}^3$ and let S be the upper hemisphere of radius R with the outward (upward) normal. Its rim is the equator traversed counterclockwise, and Example T2.2 gave $\oint_{\partial S}\alpha = 2\pi R^2$. On the other side, $\mathrm{d}\alpha = 2\,\mathrm{d}x\wedge\mathrm{d}y$. On the sphere $x = R\sin\theta\cos\phi$, $y = R\sin\theta\sin\phi$, and the polar-coordinate computation of Section T2.1.3 with $r \to R\sin\theta$ gives $\mathrm{d}x\wedge\mathrm{d}y = R\sin\theta\,\mathrm{d}(R\sin\theta)\wedge\mathrm{d}\phi = R^2\sin\theta\cos\theta\,\mathrm{d}\theta\wedge\mathrm{d}\phi$. So

$$\int_S \mathrm{d}\alpha \overset{(1)}{=} 2R^2\int_0^{2\pi}\int_0^{\pi/2} \sin\theta\cos\theta\,\mathrm{d}\theta\,\mathrm{d}\phi$$
$$\overset{(2)}{=} 2R^2\cdot 2\pi\cdot\tfrac{1}{2} = 2\pi R^2.$$

(1) (θ,ϕ) is positively oriented for the outward normal; restrict to $\theta \le \pi/2$.
(2) $\int_0^{\pi/2}\sin\theta\cos\theta\,\mathrm{d}\theta = \tfrac{1}{2}\sin^2\theta\big|_0^{\pi/2} = \tfrac{1}{2}$.

The flat disc $x^2 + y^2 \le R^2$ with the same rim gives $\int 2\,\mathrm{d}x\wedge\mathrm{d}y = 2\pi R^2$ too. Why must any two surfaces S_1, S_2 with the same rim agree? Glue S_1 to $-S_2$: the result is a closed surface Σ (no edge), and

$$\int_{S_1}\mathrm{d}\alpha - \int_{S_2}\mathrm{d}\alpha \overset{(1)}{=} \int_\Sigma \mathrm{d}\alpha$$
$$\overset{(2)}{=} \int_{\partial\Sigma}\alpha$$
$$\overset{(3)}{=} 0.$$

(1) Reversing orientation flips the sign; the rims cancel when glued.
(2) Stokes' theorem.
(3) A closed surface has empty boundary, and an integral over the empty set is zero.

This is the argument of the model paragraph in our style guide, and we will use it again and again.

Worked example T2.9: Gauss's theorem for the paraboloid

Let V be the solid $0 \le z \le 1 - x^2 - y^2$ and $\vec{F} = \vec{r}$, so $\vec{\nabla}\cdot\vec{F} = 3$. Its boundary is the paraboloid cap S of Example T2.3 together with the unit disc B at $z = 0$ with downward normal. Then

$$\int_S \star F + \int_B \star F \overset{(1)}{=} \int_V \mathrm{d}\star F$$

$$\overset{(2)}{=} 3\int_0^{2\pi}\int_0^1 (1-\rho^2)\,\rho\,\mathrm{d}\rho\,\mathrm{d}\phi$$
$$= 3\cdot 2\pi\left(\tfrac{1}{2}-\tfrac{1}{4}\right) = \frac{3\pi}{2}.$$

(1) Stokes (Gauss) with $\partial V = S \cup B$, both outward.
(2) $\mathrm{d}{\star}F = 3\,\mathrm{d}V$ by (T2.13); the height of the solid above (ρ,ϕ) is $1-\rho^2$.

On B we have $z = 0$ and $\mathrm{d}z = 0$, so ${\star}F|_B = z\,\mathrm{d}x\wedge\mathrm{d}y = 0$. Hence $\int_S {\star}F = 3\pi/2$, agreeing with the direct computation in Example T2.3.

T2.2.5 The boundary of a boundary is empty, and $\mathrm{d}^2 = 0$

The boundary of a disc is a circle, and a circle has no boundary. The boundary of a ball is a sphere, which has no boundary. In general $\partial(\partial R) = \varnothing$. Stokes' theorem says that d and ∂ are *adjoint* under integration, $\int_R \mathrm{d}\alpha = \int_{\partial R}\alpha$. So $\mathrm{d}^2 = 0$ and $\partial^2 = \varnothing$ are two faces of one fact:

$$\int_R \mathrm{d}(\mathrm{d}\beta) \overset{(1)}{=} \int_{\partial R}\mathrm{d}\beta \overset{(2)}{=} \int_{\partial\partial R}\beta \overset{(3)}{=} 0. \tag{T2.14}$$

(1) Stokes on R.
(2) Stokes on the closed manifold ∂R.
(3) $\partial\partial R = \varnothing$.

Since this holds for every R, the form $\mathrm{d}\mathrm{d}\beta$ integrates to zero over every small region, so it vanishes identically. In vector language: $\vec{\nabla}\times\vec{\nabla}f = 0$ and $\vec{\nabla}\cdot(\vec{\nabla}\times\vec{A}) = 0$. This duality between d acting on forms and ∂ acting on regions is exactly the duality between cohomology and homology that Nash sets up in [Nash p. 13, (1.43)–(1.44)].

T2.2.6 Integration by parts

The product rule for forms (Chapter T1) is $\mathrm{d}(\alpha\wedge\beta) = \mathrm{d}\alpha\wedge\beta + (-1)^p\alpha\wedge\mathrm{d}\beta$ for a p-form α. Integrate over R and use Stokes on the total derivative [Dray (16.46)]:

$$\int_R \mathrm{d}\alpha\wedge\beta = \int_{\partial R}\alpha\wedge\beta - (-1)^p\int_R \alpha\wedge\mathrm{d}\beta. \tag{T2.15}$$

Special cases: for functions on a curve, $\int_C f\,\mathrm{d}g = [fg]_{\partial C} - \int_C g\,\mathrm{d}f$, the rule from first-year calculus. For a function f and a form α, $\int_D f\,\mathrm{d}\alpha = \int_{\partial D} f\alpha - \int_D \mathrm{d}f\wedge\alpha$ [Dray (16.48)].

The adjoint of d. On a closed Riemannian manifold (no boundary) define the inner product of two p-forms by $\langle\alpha,\beta\rangle = \int_M \alpha\wedge{\star}\beta$. Then (T2.15) with no boundary term says that moving d from one slot to the other costs only a sign and some Hodge stars. The operator that appears is the adjoint $\mathrm{d}^* = \pm{\star}\mathrm{d}{\star}$ (the sign depends on degree and dimension; Chapter T1), defined by $\langle\mathrm{d}\alpha,\beta\rangle = \langle\alpha,\mathrm{d}^*\beta\rangle$. The Hodge Laplacian $\Delta = \mathrm{d}\mathrm{d}^* + \mathrm{d}^*\mathrm{d}$ is built from it; see Section T2.4.9.

Worked example T2.10: Harmonic functions on a closed manifold are constant

This is Dray's Problems 16.3–16.4. Let M be a closed, connected, Riemannian manifold (for instance the flat torus with $\mathrm{d}s^2 = a^2\mathrm{d}\theta^2 + b^2\mathrm{d}\phi^2$, or the round S^2). Use Dray's Laplacian $\Delta f = \star\mathrm{d}\star\,\mathrm{d}f$ and suppose $\Delta f = 0$. Since $\star$ is invertible, $\mathrm{d}\star\mathrm{d}f = 0$. Then

$$\begin{aligned} 0 &\overset{(1)}{=} \int_M f\,\mathrm{d}\star\mathrm{d}f \\ &\overset{(2)}{=} \int_M \mathrm{d}\big(f\star\mathrm{d}f\big) - \int_M \mathrm{d}f\wedge\star\mathrm{d}f \\ &\overset{(3)}{=} 0 - \int_M |\mathrm{d}f|^2\ \mathrm{vol}. \end{aligned}$$

(1) $\mathrm{d}\star\mathrm{d}f = 0$ by assumption.
(2) The product rule $\mathrm{d}(f\beta) = \mathrm{d}f\wedge\beta + f\,\mathrm{d}\beta$ with $\beta = \star\mathrm{d}f$.
(3) Stokes with $\partial M = \varnothing$ kills the first term; for any 1-form γ, $\gamma\wedge\star\gamma = |\gamma|^2\,\mathrm{vol}$ (the defining property of $\star$).

In Euclidean signature $|\mathrm{d}f|^2 \geq 0$, so the integral of a non-negative function vanishes: $\mathrm{d}f = 0$ everywhere and f is constant. On the torus this says the only solutions of $a^{-2}\partial_\theta^2 f + b^{-2}\partial_\phi^2 f = 0$ are constants, as a Fourier series check confirms: $e^{i(m\theta+n\phi)}$ has eigenvalue $-(m^2/a^2 + n^2/b^2)$, which vanishes only for $m = n = 0$. In Minkowski signature the argument fails because $|\mathrm{d}f|^2$ has no sign: there are plenty of non-constant solutions of the wave equation.

T2.2.7 Corollaries: things that must vanish somewhere

Dray draws some pretty consequences from Stokes' theorem on closed manifolds [Dray §16.7]. Let M be a closed oriented n-manifold with orientation form vol, so $\int_M \mathrm{vol} > 0$.

Key idea: an exact top form must change sign or vanish

If $\mathrm{d}\alpha = h\,\mathrm{vol}$ for some $(n-1)$-form α and function h, then $\int_M h\,\mathrm{vol} = \int_{\partial M}\alpha = 0$. A continuous h with zero integral against a positive measure either vanishes identically or takes both signs; either way *h has a zero*. In particular vol itself is not exact.

- *Every smooth function on a circle has a critical point.* Take $n = 1$, $\alpha = f$, $\mathrm{d}f = f'(\phi)\,\mathrm{d}\phi$. Then f' has a zero. (Obvious from the picture: a periodic function has a maximum.)

- *Every smooth vector field on a sphere has vanishing curl (normal component) somewhere.* Take $\alpha = F = \vec{F}\cdot\mathrm{d}\vec{r}$ restricted to S^2. Then $\mathrm{d}F = (\vec{\nabla}\times\vec{F})\cdot\hat{n}\,\mathrm{vol}$, so $(\vec{\nabla}\times\vec{F})\cdot\hat{n}$ has a zero on S^2.

- *Charge neutrality on a closed universe.* On a closed spatial manifold Gauss's law $\mathrm{d}\star E = \rho\,\mathrm{vol}$ forces $\int\rho\,\mathrm{vol} = 0$: the total charge of a closed universe is zero.

T2.2.8 Boundary terms in action principles

A Lagrangian density is an n-form, and the action is its integral [Dray §20.2]. Varying the action produces a total derivative, and Stokes' theorem turns it into a boundary term. For a scalar

field Φ with $S[\Phi] = \int_R \mathrm{d}\Phi \wedge \star\mathrm{d}\Phi$, vary $\Phi \to \Phi + \delta\Phi$:

$$\begin{aligned}\delta S &\overset{(1)}{=} \int_R \left(\mathrm{d}\delta\Phi \wedge \star\mathrm{d}\Phi + \mathrm{d}\Phi \wedge \star\mathrm{d}\delta\Phi\right)\\ &\overset{(2)}{=} 2\int_R \mathrm{d}\delta\Phi \wedge \star\mathrm{d}\Phi\\ &\overset{(3)}{=} 2\int_R \mathrm{d}\left(\delta\Phi \star\mathrm{d}\Phi\right) - 2\int_R \delta\Phi\, \mathrm{d}\star\mathrm{d}\Phi\\ &\overset{(4)}{=} 2\int_{\partial R} \delta\Phi \star\mathrm{d}\Phi - 2\int_R \delta\Phi\, \mathrm{d}\star\mathrm{d}\Phi. \end{aligned} \tag{T2.16}$$

(1) Keep terms linear in $\delta\Phi$; d commutes with the variation.
(2) The inner product is symmetric: $\alpha \wedge \star\beta = \beta \wedge \star\alpha$ for forms of equal degree.
(3) Product rule, as in Example T2.10.
(4) Stokes' theorem on the total derivative.

Variations supported away from the edge give the field equation $\mathrm{d}\star\mathrm{d}\Phi = 0$, which is Laplace's equation (or the wave equation in Minkowski signature). If $\delta\Phi$ is free on ∂R as well, the boundary term forces $\star\mathrm{d}\Phi|_{\partial R} = 0$: the normal derivative of Φ vanishes (the Neumann condition), because $\star\mathrm{d}\Phi$ restricted to a surface is the flux of $\vec{\nabla}\Phi$ through it. The same computation with $L = F \wedge \star F$, $F = \mathrm{d}A$, gives the vacuum Maxwell equation $\mathrm{d}\star F = 0$; the other half, $\mathrm{d}F = 0$, is $\mathrm{d}^2 = 0$.

Reading Nash (Stokes' theorem in the Bogomolny argument)

Nash uses exactly the reasoning of Section T2.2.7 when he studies monopoles [Nash p. 244, (8.149)–(8.150)]. He shows that the cross term in the energy is a total derivative, $\mathrm{tr}(F \wedge \mathrm{d}_A\phi) = \mathrm{d}\,\mathrm{tr}(F\phi)$, and concludes: "Thus if M is compact and closed Stokes' theorem gives $\langle F, \star\mathrm{d}_A\phi\rangle = -\int_M \mathrm{d}\,\mathrm{tr}(F\phi) = 0$." On $\mathbb{R}^3$ the boundary is the sphere at infinity, and the same integral becomes $\frac{1}{4\pi}\int_{S^2_\infty} \mathrm{tr}(F\phi)$, the magnetic charge [Nash p. 245, (8.154)]. The whole topological content of the monopole lives in a boundary term, and it is Stokes' theorem that puts it there.

Recap

- $\int_R \mathrm{d}\alpha = \int_{\partial R} \alpha$, with ∂R oriented "outward normal first" (counterclockwise for planar regions).
- Mechanism: the fundamental theorem of calculus one direction at a time on a small cube; interior faces cancel when cubes are glued.
- FTC, gradient, Green, curl and divergence theorems are the cases $p = 0, 1, 1, 2$.
- $\mathrm{d}^2 = 0$ is the dual of $\partial^2 = \varnothing$.
- Integration by parts (T2.15); on closed manifolds harmonic functions are constant.
- On a closed manifold an exact top form integrates to zero, so its coefficient must vanish somewhere; the volume form is not exact.
- Varying an action gives field equations plus a boundary term.

T2.3 Closed and exact forms

Every physicist knows the curl test. If $\vec{\nabla}\times\vec{F}=0$, then $\vec{F}$ is the gradient of a potential, $\vec{F}=\vec{\nabla}f$, and the work done along a path depends only on its endpoints. Every physicist also knows that the test has a fine-print clause: "in a simply connected region". The standard counterexample is the magnetic field of a straight wire, $\vec{B}\propto\hat{\phi}/\rho$. Its curl vanishes everywhere away from the wire, yet its circulation around the wire is not zero, so it is *not* a gradient.

This section is about that fine print. In the language of forms the curl test asks: *if* $\mathrm{d}\alpha=0$, *is* $\alpha=\mathrm{d}\beta$*?* The answer is "yes, locally, always" (the Poincaré lemma) and "globally, not necessarily". The failures are not a nuisance. They are the first appearance of topology in these notes: a closed form that is not exact is a detector for a hole in the space. Stokes' theorem is the tool that detects it.

T2.3.1 Two definitions and one implication

Working definition T2.4 (closed and exact)

A p-form α on a manifold M is

- **closed** if $\mathrm{d}\alpha=0$ (test: differentiate it);
- **exact** if $\alpha=\mathrm{d}\beta$ for some $(p-1)$-form β that is smooth *on all of* M (test: find β, or use the Stokes detector below to prove none exists [a closed p-surface S, with no boundary, on which $\int_S\alpha\neq0$]).

The form β is called a **potential** for α. *Formally:* closed forms are $\ker\mathrm{d}$, exact forms are $\mathrm{im}\,\mathrm{d}$.

Exact implies closed, because $\mathrm{d}^2=0$ (Chapter T1):

$$\alpha=\mathrm{d}\beta\quad\Longrightarrow\quad\mathrm{d}\alpha=\mathrm{dd}\beta=0. \tag{T2.17}$$

Dray calls this the Poincaré lemma and its converse "the converse of the Poincaré lemma", and he notes that the converse is often *also* called the Poincaré lemma [Dray §20.4]. We will reserve the name for the converse, Section T2.3.3, as most books do.

In $\mathbb{R}^3$, with the dictionary of Section T2.1, the implication (T2.17) is two familiar identities:

degree	exact	closed	identity
0	—	$\mathrm{d}f=0$: f locally constant	—
1	$F=\mathrm{d}f$: $\vec{F}=\vec{\nabla}f$	$\mathrm{d}F=0$: $\vec{\nabla}\times\vec{F}=0$	$\vec{\nabla}\times\vec{\nabla}f=0$
2	$\star F=\mathrm{d}A$: $\vec{F}=\vec{\nabla}\times\vec{A}$	$\mathrm{d}\star F=0$: $\vec{\nabla}\cdot\vec{F}=0$	$\vec{\nabla}\cdot(\vec{\nabla}\times\vec{A})=0$
3	$\star f=\mathrm{d}\star A$: $f=\vec{\nabla}\cdot\vec{A}$	always (d of a 3-form in $\mathbb{R}^3$ is 0)	—

Physicists meet exact 1-forms as *conservative forces* and exact 2-forms as *magnetic fields with a vector potential*. The table's middle column is what we can test by differentiating. The left column is what we want.

Worked example T2.11: An exact 1-form and a closed 1-form

Let $\alpha=x\,\mathrm{d}x+y\,\mathrm{d}y$ on $\mathbb{R}^2$. It is exact: $\alpha=\mathrm{d}\left(\frac{1}{2}(x^2+y^2)\right)$, as Dray observes [Dray §20.4]. Now

let $\gamma = 2xy\,dx + (x^2 + \cos y)\,dy$. Is it closed?

$$d\gamma \overset{(1)}{=} \partial_y(2xy)\,dy \wedge dx + \partial_x(x^2 + \cos y)\,dx \wedge dy$$
$$\overset{(2)}{=} (-2x + 2x)\,dx \wedge dy = 0.$$

(1) $d(g\,dx) = dg \wedge dx = \partial_y g\,dy \wedge dx$, and similarly for the dy term.
(2) $dy \wedge dx = -dx \wedge dy$.

It is closed, and indeed $\gamma = d(x^2y + \sin y)$. We will find this potential mechanically in Example T2.12.

T2.3.2 The Stokes detector

Here is the owner's favourite argument, stated once and for all. It is the only tool we need to prove that a form is *not* exact.

Key idea: an exact form integrates to zero over anything without boundary

Let $\alpha = d\beta$ be exact and let Z be a closed (compact, boundaryless) oriented p-dimensional submanifold, or more generally any p-dimensional "cycle" with $\partial Z = \varnothing$. Then

$$\int_Z \alpha \overset{(1)}{=} \int_Z d\beta \overset{(2)}{=} \int_{\partial Z} \beta \overset{(3)}{=} \int_{\varnothing} \beta = 0. \tag{T2.18}$$

(1) α is exact, with a potential β defined on all of Z.
(2) Stokes' theorem (T2.9).
(3) Z has no boundary; an integral over the empty set is a sum with no terms.

Read backwards: *one closed cycle Z with $\int_Z \alpha \neq 0$ proves that α is not exact*, however cleverly we search for β.

Two remarks make the detector sharper.

The potential must exist on Z itself. Step (1) needs β to be smooth on all of Z. A potential that blows up somewhere on Z does not count, and this is exactly how the "paradoxes" below are resolved.

For 1-forms the detector is also sufficient. Suppose α is a 1-form on a connected manifold with $\oint_\gamma \alpha = 0$ for *every* closed loop γ. Fix a base point x_0 and define $f(x) = \int_{x_0}^{x} \alpha$ along any path. Two paths from x_0 to x together make a closed loop (go out along one, back along the other), so the answer does not depend on the path. Moving the endpoint a little in the direction $\vec{v}$ changes f by $\alpha(\vec{v})$ times the step, so $df = \alpha$. Hence

$$\alpha \text{ exact} \iff \oint_\gamma \alpha = 0 \text{ for every closed loop } \gamma \iff \text{line integrals of } \alpha \text{ are path independent.} \tag{T2.19}$$

This is the conservative-force theorem of mechanics. In higher degree the analogous statement (a closed p-form is exact if and only if all its integrals over closed p-cycles vanish) is de Rham's theorem, Section T2.4.8; for us it is enough to know that the detector (T2.18) never lies.

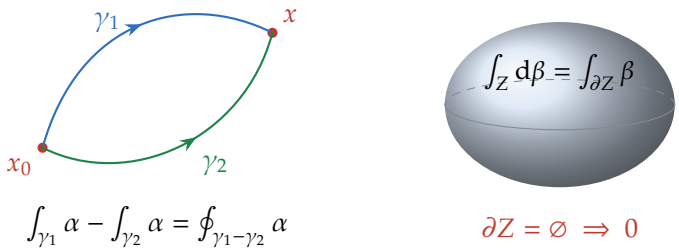


Figure T2.8: Left: two paths with the same endpoints differ by a closed loop, so path independence is the same as "zero circulation around every loop". Right: the Stokes detector. A closed surface has no edge, so an exact form has nothing to be converted into and integrates to zero.

Potentials are not unique. If $\alpha = \mathrm{d}\beta$, then also $\alpha = \mathrm{d}(\beta + \mathrm{d}\gamma)$ for any $(p-2)$-form γ, because $\mathrm{d}\mathrm{d}\gamma = 0$ [Dray (20.29)]. For $p = 1$ this is the familiar "+ constant" of an indefinite integral, $\mathrm{d}(f + c) = \mathrm{d}f$ [Dray (20.30)]. For $p = 2$ it is *gauge freedom*: the vector potential of a magnetic field is determined only up to $A \mapsto A + \mathrm{d}\lambda$, i.e. $\vec{A} \mapsto \vec{A} + \vec{\nabla}\lambda$. Physics lives in α; the potential carries redundant information. We will see in Section T2.5.10 that the connection of a surface has exactly the same redundancy.

T2.3.3 The Poincaré lemma: closed is locally exact

Now the positive result. We show that on a **star-shaped** region $U \subset \mathbb{R}^n$ (one containing a point, which we take as the origin, such that the straight segment from the origin to any point of U lies in U) every closed form is exact. The proof is constructive: it hands us the potential as an explicit integral. Balls, cubes, half-spaces and $\mathbb{R}^n$ itself are star-shaped. Every point of every manifold has a star-shaped coordinate neighbourhood, so every closed form is exact *locally*.

The idea. Shrink the region to the origin by the scaling map $\phi_t(x) = tx$, $0 \le t \le 1$. At $t = 1$ nothing has moved; at $t = 0$ everything sits at the origin, where a p-form with $p \ge 1$ pulled back along a constant map is zero. So $\alpha = \phi_1^*\alpha - \phi_0^*\alpha$ is the integral of a t-derivative. That derivative turns out to be exact when $\mathrm{d}\alpha = 0$.

The 1-form case by hand. Let $\alpha = \alpha_\mu\,\mathrm{d}x^\mu$ be closed, so $\partial_\nu\alpha_\mu = \partial_\mu\alpha_\nu$. Define f by integrating α along the ray from 0 to x, parametrised by $t \mapsto tx$ (so $\mathrm{d}x^\mu \mapsto x^\mu\,\mathrm{d}t$):

$$f(x) := \int_0^1 x^\mu\,\alpha_\mu(tx)\,\mathrm{d}t. \tag{T2.20}$$

Then

$$\begin{aligned}
\partial_\nu f(x) &\overset{(1)}{=} \int_0^1 \Big(\alpha_\nu(tx) + x^\mu\, t\,(\partial_\nu\alpha_\mu)(tx)\Big)\mathrm{d}t \\
&\overset{(2)}{=} \int_0^1 \Big(\alpha_\nu(tx) + t\,x^\mu(\partial_\mu\alpha_\nu)(tx)\Big)\mathrm{d}t \\
&\overset{(3)}{=} \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t}\Big(t\,\alpha_\nu(tx)\Big)\mathrm{d}t \\
&\overset{(4)}{=} \alpha_\nu(x).
\end{aligned} \tag{T2.21}$$

(1) Differentiate under the integral: $\partial_\nu x^\mu = \delta^\mu_\nu$, and the chain rule on $\alpha_\mu(tx)$ gives a factor t.

(2) Closedness, $\partial_\nu \alpha_\mu = \partial_\mu \alpha_\nu$.
(3) Product rule: $\frac{\mathrm{d}}{\mathrm{d}t}\big(t\,\alpha_\nu(tx)\big) = \alpha_\nu(tx) + t\,x^\mu(\partial_\mu\alpha_\nu)(tx)$.
(4) Fundamental theorem of calculus; the lower limit gives $0 \cdot \alpha_\nu(0) = 0$.

So $\mathrm{d}f = \alpha$. Closedness was used exactly once, in step (2). Without it, step (3) fails and the ray integral has the wrong gradient.

All degrees: the homotopy operator. Let $E = x^\mu \partial_\mu$ be the radial (Euler) vector field; its flow is scaling. For a p-form $\alpha = \frac{1}{p!}\alpha_{\mu_1\ldots\mu_p}\mathrm{d}x^{\mu_1}\wedge\cdots\wedge\mathrm{d}x^{\mu_p}$ with $p \geq 1$ define the $(p-1)$-form

$$(K\alpha)(x) := \int_0^1 t^{p-1}\,\big(\iota_x\alpha\big)(tx)\,\mathrm{d}t = \frac{1}{(p-1)!}\int_0^1 t^{p-1}x^\mu\alpha_{\mu\nu_2\ldots\nu_p}(tx)\,\mathrm{d}t\;\mathrm{d}x^{\nu_2}\wedge\cdots\wedge\mathrm{d}x^{\nu_p}, \tag{T2.22}$$

where ι_x means "feed the position vector into the first slot". For $p = 1$ this is (T2.20). The key identity is

$$\mathrm{d}K\alpha + K\mathrm{d}\alpha = \alpha \qquad (p \geq 1), \qquad\qquad K\mathrm{d}f = f - f(0) \qquad (p = 0). \tag{T2.23}$$

Here is why. Write $t = e^s$, so ϕ_t is the flow of E for time s. The derivative of a pulled-back form along a flow is the Lie derivative (Chapter T1), and Cartan's formula gives $\mathcal{L}_E = \mathrm{d}\,\iota_E + \iota_E\,\mathrm{d}$. Then for $p \geq 1$

$$\begin{aligned}
\alpha &\overset{(1)}{=} \phi_1^*\alpha - \phi_0^*\alpha = \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t}\phi_t^*\alpha\;\mathrm{d}t \\
&\overset{(2)}{=} \int_0^1 \frac{1}{t}\,\phi_t^*\big(\mathcal{L}_E\alpha\big)\,\mathrm{d}t \\
&\overset{(3)}{=} \int_0^1 \frac{1}{t}\,\phi_t^*\big(\mathrm{d}\,\iota_E\alpha\big)\,\mathrm{d}t + \int_0^1 \frac{1}{t}\,\phi_t^*\big(\iota_E\,\mathrm{d}\alpha\big)\,\mathrm{d}t \\
&\overset{(4)}{=} \mathrm{d}\Big(\int_0^1 \frac{1}{t}\,\phi_t^*\,\iota_E\alpha\,\mathrm{d}t\Big) + \int_0^1 \frac{1}{t}\,\phi_t^*\,\iota_E(\mathrm{d}\alpha)\,\mathrm{d}t \\
&\overset{(5)}{=} \mathrm{d}K\alpha + K\mathrm{d}\alpha.
\end{aligned} \tag{T2.24}$$

(1) $\phi_1 = \mathrm{id}$, and ϕ_0 maps everything to the origin, so ϕ_0^* kills every form of degree ≥ 1 (all $\mathrm{d}x^\mu$ pull back to 0).
(2) $\frac{\mathrm{d}}{\mathrm{d}t} = \frac{1}{t}\frac{\mathrm{d}}{\mathrm{d}s}$, and $\frac{\mathrm{d}}{\mathrm{d}s}\phi^*\alpha = \phi^*\mathcal{L}_E\alpha$ by the definition of the Lie derivative along a flow.
(3) Cartan's formula.
(4) Pullback commutes with d, and so does integration over the parameter t.
(5) In components, ϕ_t^* replaces x by tx in the coefficients and each $\mathrm{d}x^\nu$ by $t\,\mathrm{d}x^\nu$; $\iota_E\alpha$ has one factor x^μ (giving tx^μ) and $p-1$ differentials, so $\frac{1}{t}\phi_t^*\iota_E\alpha = t^{p-1}x^\mu\alpha_{\mu\ldots}(tx)\,\mathrm{d}x^{\nu_2}\wedge\cdots$, which is the integrand of (T2.22).

(The integrand of (T2.24) is harmless at $t = 0$: after step (5) there is no $1/t$ left.)

Key idea: Poincaré lemma

On a star-shaped region every closed p-form with $p \geq 1$ is exact, with the explicit potential $\beta = K\alpha$ of (T2.22). Every closed 0-form is constant. *Reason:* if $\mathrm{d}\alpha = 0$, the identity $\mathrm{d}K + K\mathrm{d} = 1$ reduces to $\alpha = \mathrm{d}(K\alpha)$.

Worked example T2.12: Finding a potential mechanically

For $\gamma = 2xy\,\mathrm{d}x + (x^2 + \cos y)\,\mathrm{d}y$ from Example T2.11, formula (T2.20) gives

$$\begin{aligned} f(x,y) &\overset{(1)}{=} \int_0^1 \Big(x\cdot 2(tx)(ty) + y\cdot\big((tx)^2 + \cos(ty)\big)\Big)\mathrm{d}t \\ &\overset{(2)}{=} \int_0^1 \big(3t^2x^2y + y\cos(ty)\big)\mathrm{d}t \\ &\overset{(3)}{=} x^2y + \sin y. \end{aligned}$$

(1) $f = \int_0^1 (x\,\gamma_x + y\,\gamma_y)(tx,ty)\,\mathrm{d}t$.
(2) Collect: $2t^2x^2y + t^2x^2y = 3t^2x^2y$.
(3) $\int_0^1 3t^2\mathrm{d}t = 1$ and $\int_0^1 y\cos(ty)\,\mathrm{d}t = \sin(ty)\big|_{t=0}^{t=1} = \sin y$.

Check: $\mathrm{d}(x^2y + \sin y) = 2xy\,\mathrm{d}x + (x^2+\cos y)\,\mathrm{d}y = \gamma$. In three dimensions the same recipe applied to $yz\,\mathrm{d}x + xz\,\mathrm{d}y + xy\,\mathrm{d}z$ gives $\int_0^1 3t^2xyz\,\mathrm{d}t = xyz$.

Worked example T2.13: The vector potential of a uniform magnetic field

A constant field $\vec{B}$ in $\mathbb{R}^3$ is the closed 2-form $\star B = B_x\,\mathrm{d}y\wedge\mathrm{d}z + B_y\,\mathrm{d}z\wedge\mathrm{d}x + B_z\,\mathrm{d}x\wedge\mathrm{d}y$, with components $\beta_{\mu\nu} = \varepsilon_{\mu\nu\lambda}B^\lambda$. The homotopy operator with $p=2$ gives

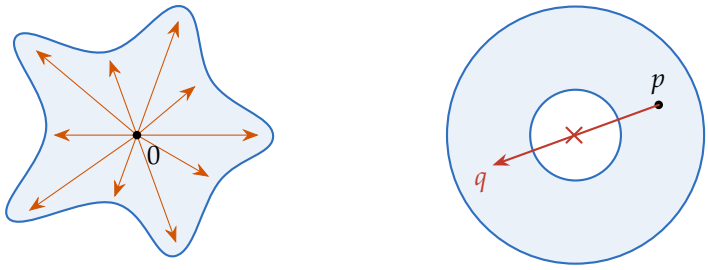
$$\begin{aligned} K(\star B) &\overset{(1)}{=} \int_0^1 t\,x^\mu\,\varepsilon_{\mu\nu\lambda}B^\lambda\,\mathrm{d}t\;\mathrm{d}x^\nu \\ &\overset{(2)}{=} \tfrac{1}{2}\,\varepsilon_{\nu\lambda\mu}B^\lambda x^\mu\,\mathrm{d}x^\nu \\ &\overset{(3)}{=} \tfrac{1}{2}\,(\vec{B}\times\vec{r})\cdot\mathrm{d}\vec{r}. \end{aligned}$$

(1) (T2.22) with $p=2$; the coefficients are constant, so $\beta(tx) = \beta(x)$.
(2) $\int_0^1 t\,\mathrm{d}t = \tfrac{1}{2}$, and $\varepsilon_{\mu\nu\lambda} = \varepsilon_{\nu\lambda\mu}$ (cyclic).
(3) $(\vec{B}\times\vec{r})_\nu = \varepsilon_{\nu\lambda\mu}B^\lambda x^\mu$.

This is the symmetric gauge $\vec{A} = \tfrac{1}{2}\vec{B}\times\vec{r}$ familiar from Landau levels. For $\vec{B} = B\hat{z}$ it reads $A = \frac{B}{2}(x\,\mathrm{d}y - y\,\mathrm{d}x)$, and $\mathrm{d}A = B\,\mathrm{d}x\wedge\mathrm{d}y$ by the computation of Example T2.7. The Landau gauge $A' = Bx\,\mathrm{d}y$ differs from it by $\mathrm{d}\lambda$ with $\lambda = \frac{B}{2}xy$: check $\mathrm{d}\lambda = \frac{B}{2}(y\,\mathrm{d}x + x\,\mathrm{d}y)$ and $A + \mathrm{d}\lambda = Bx\,\mathrm{d}y$.

Watch out: "simply connected" is only enough for 1-forms

Dray states the converse of the Poincaré lemma for simply connected regions of $\mathbb{R}^n$ [Dray §20.4]. For 1-forms that is correct (and follows from (T2.19), since every loop in a simply connected region bounds a disc). For higher forms it is false. The region $\mathbb{R}^3\setminus\{0\}$ is simply connected (every loop can be slid off the missing point), yet the Coulomb flux form $\star E$ of Example T2.6 is closed and has $\int_{S^2}\star E = q \neq 0$, so it is not exact. A 2-form sees 2-dimensional holes, which loops cannot detect. Star-shaped (or, more generally, contractible) is the right hypothesis in every degree.



star-shaped: every ray from 0 stays inside annulus: no centre works

Figure T2.9: Left: in a star-shaped region every point is joined to the centre by a straight segment inside the region; the homotopy operator K integrates the form along these rays. Right: an annulus is not star-shaped. Whatever centre p we choose, some segment pq crosses the hole. This is exactly where closed forms can fail to be exact.

T2.3.4 The angle form on the punctured plane

The punctured plane $\mathbb{R}^2 \setminus \{0\}$ is the simplest space with a hole, and it carries *the* closed-but-not-exact form. The polar angle θ is not a function on the punctured plane (it jumps by 2π somewhere), but its differential is:

$$\mathrm{d}\theta = \frac{x\,\mathrm{d}y - y\,\mathrm{d}x}{x^2 + y^2}. \tag{T2.25}$$

This expression is smooth everywhere except the origin. It is the 1-form of the wire's magnetic field, $\vec{B} \propto \hat{\phi}/\rho$.

It is closed. With $r^2 = x^2 + y^2$,

$$\begin{aligned} \mathrm{d}\left(\frac{x\,\mathrm{d}y - y\,\mathrm{d}x}{r^2}\right) &\overset{(1)}{=} \partial_x\left(\frac{x}{r^2}\right)\mathrm{d}x \wedge \mathrm{d}y - \partial_y\left(\frac{y}{r^2}\right)\mathrm{d}y \wedge \mathrm{d}x \\ &\overset{(2)}{=} \left(\frac{r^2 - 2x^2}{r^4} + \frac{r^2 - 2y^2}{r^4}\right)\mathrm{d}x \wedge \mathrm{d}y \\ &\overset{(3)}{=} \frac{2r^2 - 2(x^2 + y^2)}{r^4}\,\mathrm{d}x \wedge \mathrm{d}y = 0. \end{aligned} \tag{T2.26}$$

(1) $\mathrm{d}(g\,\mathrm{d}y) = \partial_x g\,\mathrm{d}x \wedge \mathrm{d}y$ plus a $\mathrm{d}y \wedge \mathrm{d}y = 0$ term; similarly for $\mathrm{d}x$.
(2) Quotient rule, $\partial_x(x/r^2) = (r^2 - 2x^2)/r^4$; and $\mathrm{d}y \wedge \mathrm{d}x = -\mathrm{d}x \wedge \mathrm{d}y$.
(3) $x^2 + y^2 = r^2$.

It is not exact. On the unit circle, $x = \cos t$, $y = \sin t$:

$$\begin{aligned} \oint_{S^1} \mathrm{d}\theta &\overset{(1)}{=} \int_0^{2\pi} \frac{\cos t\,(\cos t\,\mathrm{d}t) - \sin t\,(-\sin t\,\mathrm{d}t)}{1} \\ &\overset{(2)}{=} \int_0^{2\pi} \mathrm{d}t = 2\pi \neq 0. \end{aligned} \tag{T2.27}$$

(1) Pull back along the parametrisation; $r^2 = 1$ on the circle.
(2) $\cos^2 + \sin^2 = 1$.

By the detector (T2.18) (the circle has no boundary), $\mathrm{d}\theta$ is not d of any smooth function on the punctured plane. The notation is a pun: the "θ" in $\mathrm{d}\theta$ exists only on slit planes such as $\mathbb{R}^2$

minus the negative x-axis, where it is a perfectly good function. On the slit plane $\mathrm{d}\theta$ is exact, which is consistent with the detector because no loop in the slit plane goes around the origin.

Physical picture: the missing point is the source

Apply Stokes' theorem to the unit disc. The left side, $\int_D \mathrm{d}(\mathrm{d}\theta)$, would be zero if $\mathrm{d}\theta$ were smooth on the disc, but the right side is 2π. So the "curl" of $\mathrm{d}\theta$ must be a delta function at the origin: $\mathrm{d}(\mathrm{d}\theta) = 2\pi\,\delta^2(x)\,\mathrm{d}x \wedge \mathrm{d}y$ as a distribution. This is Ampère's law for a line current: $\vec{\nabla} \times \vec{B} = \mu_0 I\,\delta^2\hat{z}$ with the field zero-curl everywhere else. The hole in the space is where the source sits.

T2.3.5 Winding numbers

The period 2π is not an accident of the unit circle. For any closed loop γ in the punctured plane, define the **winding number**

$$w(\gamma) := \frac{1}{2\pi} \oint_\gamma \mathrm{d}\theta. \tag{T2.28}$$

It is an integer. Parametrise γ by $t \in [0,1]$. Along the path we can always choose the angle $\theta(t)$ of $\gamma(t)$ continuously (start with any value and follow it). The pullback of $\mathrm{d}\theta$ is then $\dot{\theta}(t)\,\mathrm{d}t$, so $\oint_\gamma \mathrm{d}\theta = \theta(1) - \theta(0)$. Since $\gamma(1) = \gamma(0)$ is the same point, the two angles differ by a multiple of 2π. So $w(\gamma) \in \mathbb{Z}$ counts the net number of times γ goes counterclockwise around the origin.

It does not change under deformation. Let γ_0 and γ_1 be two loops, and suppose γ_1 can be deformed into γ_0 without crossing the origin. The deformation sweeps out a region A (an annulus, possibly squashed) with $\partial A = \gamma_1 - \gamma_0$. Then

$$\begin{aligned} \oint_{\gamma_1} \mathrm{d}\theta - \oint_{\gamma_0} \mathrm{d}\theta &\overset{(1)}{=} \int_{\partial A} \mathrm{d}\theta \\ &\overset{(2)}{=} \int_A \mathrm{d}(\mathrm{d}\theta) \\ &\overset{(3)}{=} 0. \end{aligned} \tag{T2.29}$$

(1) The boundary of the swept region is γ_1 with its orientation and γ_0 reversed.
(2) Stokes' theorem on A, which does not contain the origin, so $\mathrm{d}\theta$ is smooth on it.
(3) $\mathrm{d}\theta$ is closed, (T2.26).

Notice the division of labour: *closedness* makes the integral a topological invariant (it cannot change under deformation), and *non-exactness* makes the invariant non-trivial.

Worked example T2.14: An ellipse, a small circle, and a double loop

Ellipse around the origin. Take $x = a\cos t$, $y = b\sin t$. Then $x\,\mathrm{d}y - y\,\mathrm{d}x = ab\,\mathrm{d}t$ and

$$\begin{aligned} \oint \mathrm{d}\theta &\overset{(1)}{=} \int_0^{2\pi} \frac{ab\,\mathrm{d}t}{a^2\cos^2 t + b^2 \sin^2 t} \\ &\overset{(2)}{=} 4\int_0^{\infty} \frac{ab\,\mathrm{d}u}{a^2 + b^2u^2} \\ &\overset{(3)}{=} 4 \cdot ab \cdot \frac{1}{ab} \cdot \frac{\pi}{2} = 2\pi. \end{aligned}$$

(1) Pull back (T2.25); as in Example T2.7, $x\,dy - y\,dx = ab(\cos^2 t + \sin^2 t)dt$.
(2) The integrand has period π and is even, so the integral is four times the integral over $[0, \pi/2]$; there substitute $u = \tan t$, $dt = du/(1+u^2)$, and divide numerator and denominator by $\cos^2 t$.
(3) $\int_0^\infty du/(a^2 + b^2u^2) = \frac{1}{ab}\arctan(bu/a)\big|_0^\infty = \frac{\pi}{2ab}$.

The answer is independent of a and b, as (T2.29) predicted: the ellipse deforms to the unit circle without touching the origin.

Circle not around the origin. Take the circle of radius 1 centred at $(3, 0)$. It lies in the half-plane $x > 0$, where $\theta = \arctan(y/x)$ is a genuine smooth function. So $d\theta$ is exact there, and $\oint d\theta = 0$ by (T2.19): $w = 0$.

Double loop. $x = \cos 2t$, $y = \sin 2t$, $t \in [0, 2\pi]$. The pullback is $2\,dt$ and the integral is 4π: $w = 2$. Reversing the direction gives $w = -2$.

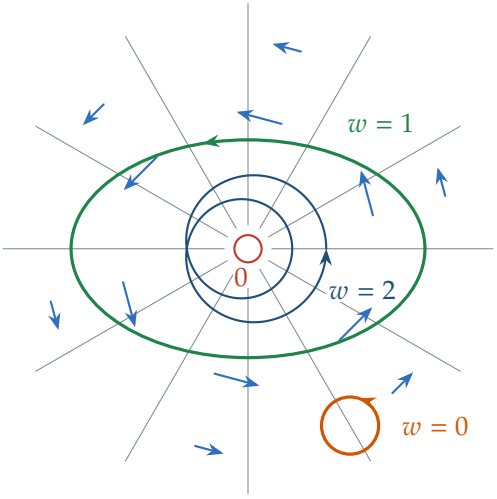
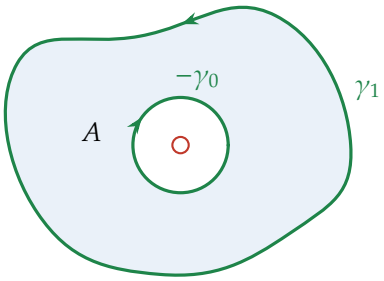


Figure T2.10: The angle form $d\theta$ on the punctured plane. Grey rays are the level sets $\theta = \text{const}$ (on any slit plane); blue arrows show the dual vector field $\hat{\theta}/r$. A loop's winding number counts how many level sets it crosses, with sign, divided by the number of rays per turn: the green ellipse crosses every ray once ($w = 1$), the orange circle crosses each ray it meets twice in opposite directions ($w = 0$), the dark spiral loop goes round twice ($w = 2$).



$$\partial A = \gamma_1 - \gamma_0$$
$$\oint_{\gamma_1} d\theta - \oint_{\gamma_0} d\theta = \int_A dd\theta = 0$$

Figure T2.11: Why winding numbers are deformation invariant. The region A swept out between two loops that surround the puncture is an annulus. Its boundary is the outer loop counterclockwise and the inner loop clockwise. Stokes' theorem and $d(d\theta) = 0$ on A force the two circulations to agree.

Every closed 1-form on the punctured plane. Let α be any closed 1-form on $\mathbb{R}^2 \setminus \{0\}$ and let $c = \oint_{S^1} \alpha$. Then $\alpha - \frac{c}{2\pi} d\theta$ is closed and has zero circulation around the unit circle, hence (by

deformation invariance) around every loop, since every loop is deformable to a multiple of the unit circle. By (T2.19) it is exact:

$$\alpha = \frac{c}{2\pi}\,\mathrm{d}\theta + \mathrm{d}g. \tag{T2.30}$$

So up to exact forms, the closed 1-forms are labelled by one real number. Dray makes the same argument on the circle itself [Dray (20.32)–(20.35)]. This is the statement $H^1_{\mathrm{dR}}(\mathbb{R}^2 \setminus \{0\}) \cong \mathbb{R}$ of the next section.

Worked example T2.15: The Aharonov–Bohm effect

A long thin solenoid along the z-axis carries magnetic flux Φ. Outside it, $\vec{B} = 0$, but the vector potential cannot vanish: by Stokes, its circulation around the solenoid is the enclosed flux. The natural choice outside is

$$A = \frac{\Phi}{2\pi}\,\mathrm{d}\theta, \qquad F = \mathrm{d}A = 0 \text{ outside}, \qquad \oint A = \Phi.$$

Outside the solenoid A is closed but not exact, so it cannot be gauged away: $A \mapsto A + \mathrm{d}\lambda$ never changes $\oint A$. A charged particle going around the solenoid picks up the phase $\exp\left(\frac{iq}{\hbar}\oint A\right) = \exp(iq\Phi/\hbar)$, which is observable in interference even though the particle never meets a magnetic field. The physical content of a closed-but-not-exact A is its period.

T2.3.6 The sphere puzzle and the monopole

Dray poses a puzzle that every student of forms should meet [Dray §20.5]. On the unit sphere take the area form $\omega = \sin\theta\,\mathrm{d}\theta \wedge \mathrm{d}\phi$. We computed $\int_{S^2}\omega = 4\pi \neq 0$ in Example T2.4, so by the detector ω is not exact. Yet

$$\mathrm{d}(-\cos\theta\,\mathrm{d}\phi) = \sin\theta\,\mathrm{d}\theta \wedge \mathrm{d}\phi = \omega. \tag{T2.31}$$

What is going on?

The resolution: $\mathrm{d}\phi$ is singular at the poles. The function $\cos\theta$ is fine everywhere. The trouble is $\mathrm{d}\phi$. Near the north pole use the Cartesian coordinates x, y of the tangent plane. There ϕ is the polar angle, so $\mathrm{d}\phi$ *is* the angle form (T2.25), with its $1/r$ singularity. A smooth 1-form has circulation tending to zero around a shrinking loop; $-\cos\theta\,\mathrm{d}\phi$ has circulation tending to -2π around a tiny circle at the north pole. Dray makes the same point from $\mathbb{R}^3$: $x\,\mathrm{d}y - y\,\mathrm{d}x = r^2\sin^2\theta\,\mathrm{d}\phi$, so the orthonormal 1-form $\sigma^\phi = r\sin\theta\,\mathrm{d}\phi = \cos\phi\,\mathrm{d}y - \sin\phi\,\mathrm{d}x$ has unit length everywhere, while $\cos\theta\,\mathrm{d}\phi = \cot\theta\,\sigma^\phi/r$ blows up at the poles [Dray (20.39)–(20.41)]. So (T2.31) holds only on the sphere minus the two poles, which is a cylinder, not a sphere, and the detector does not apply there.

Two patches. We can repair one pole at a time by adding a constant times $\mathrm{d}\phi$, which does not change d of the potential:

$$A_N = (1-\cos\theta)\,\mathrm{d}\phi \quad (\theta \neq \pi), \qquad A_S = -(1+\cos\theta)\,\mathrm{d}\phi \quad (\theta \neq 0). \tag{T2.32}$$

Near the north pole $1 - \cos\theta \approx \theta^2/2$ and $\theta^2\,\mathrm{d}\phi$ equals $x\,\mathrm{d}y - y\,\mathrm{d}x$ to leading order, which is smooth. So A_N is smooth on the whole sphere except the south pole, and A_S except the north pole. Both satisfy $\mathrm{d}A_N = \mathrm{d}A_S = \omega$. On the overlap they differ by

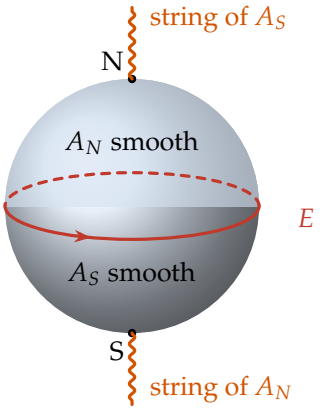
$$A_N - A_S = 2\,\mathrm{d}\phi, \tag{T2.33}$$

which is closed but not exact on the overlap (a band around the equator, topologically a punctured plane). Now compute the total integral with Stokes on each hemisphere, both bounded by the equator E (oriented counterclockwise seen from the north, the boundary orientation of the northern hemisphere N; the southern hemisphere S has boundary $-E$):

$$\begin{aligned}\int_{S^2}\omega &\overset{(1)}{=} \int_N \mathrm{d}A_N + \int_S \mathrm{d}A_S \\ &\overset{(2)}{=} \oint_E A_N - \oint_E A_S \\ &\overset{(3)}{=} \oint_E 2\,\mathrm{d}\phi = 4\pi. \end{aligned} \tag{T2.34}$$

(1) Split the sphere into hemispheres and use the potential that is smooth on each.
(2) Stokes on each hemisphere, with $\partial N = E$ and $\partial S = -E$.
(3) The transition (T2.33), and $\oint_E \mathrm{d}\phi = 2\pi$.

The total integral lives entirely in the mismatch between the two potentials. If a single global potential existed, the mismatch would be zero and so would the integral.



$A_N = (1-\cos\theta)\,\mathrm{d}\phi$
$A_S = -(1+\cos\theta)\,\mathrm{d}\phi$

on the band around E:
$A_N - A_S = 2\,\mathrm{d}\phi$

$\int_{S^2}\omega = \oint_E (A_N - A_S) = 4\pi$

Figure T2.12: The two-patch description of the monopole. A_N is smooth everywhere except the south pole, where it has a "Dirac string" singularity; A_S has its string at the north pole. On the overlap they differ by the closed, non-exact form $2\,\mathrm{d}\phi$. Stokes' theorem on the two hemispheres turns the total flux into the circulation of this mismatch around the equator.

Physical picture: Dirac's monopole

With $F = g\,\omega$ the magnetic field of a monopole of charge g (flux $4\pi g$), the two potentials are gA_N and gA_S, and they differ by the gauge transformation $g(A_N - A_S) = \mathrm{d}(2g\phi)$. A wavefunction of charge e transforms as $\psi_N = e^{2ieg\phi/\hbar}\psi_S$ on the overlap. For this to make sense as ϕ goes once around, $e^{4\pi ieg/\hbar} = 1$:

$$\frac{2eg}{\hbar} \in \mathbb{Z}.$$

That is Dirac's quantisation condition. In the language of Chapter T3 the two patches and the transition function define a nontrivial $U(1)$ bundle over S^2, and the integer is its first Chern number. Everything in this intuition box is Stokes' theorem applied to a closed-but-not-exact form.

Watch out: a formula valid on "almost all" of the space

For *integrals*, a chart missing a set of measure zero is harmless (Section T2.1.5). For *exactness*, it is fatal: a potential that is singular on a single point or a single line is not a potential. Always ask where your β is smooth before invoking Stokes' theorem.

Try it: Dray's experiment in $\mathbb{R}^3$

Take $\beta = x\,\mathrm{d}y$ in $\mathbb{R}^3$, so $\alpha = \mathrm{d}\beta = \mathrm{d}x \wedge \mathrm{d}y$. Pull α back to the unit sphere and integrate. *Answer:* in spherical coordinates the pullback is $\sin\theta\cos\theta\,\mathrm{d}\theta \wedge \mathrm{d}\phi$ (as in Example T2.8), whose integral over $0 \le \theta \le \pi$ is $2\pi \cdot \frac{1}{2}\sin^2\theta\big|_0^\pi = 0$, as the detector demands: β is smooth on all of $\mathbb{R}^3$, hence on the sphere. The northern hemisphere gives $+\pi$ and the southern $-\pi$; the flux of $\hat{z}$ in through the bottom equals the flux out through the top.

Reading Nash (p. 13, (1.41))

Nash's condition "$d_n \circ d_{n-1} = 0$ implies $\operatorname{Im} d_{n-1} \subseteq \ker d_n$" [Nash p. 13, (1.41)] is exactly (T2.17): every exact form is closed. He calls a sequence *exact* when $\operatorname{Im} d_{n-1} = \ker d_n$, i.e. when every closed form is exact, and remarks that "this is false, in general". The Poincaré lemma says the de Rham sequence *is* exact on a star-shaped region; $\mathrm{d}\theta$ and the monopole form show it is not exact on the punctured plane and on S^2. The cohomology groups he defines next measure the failure; that is the subject of the next section.

Recap

- Closed: $\mathrm{d}\alpha = 0$. Exact: $\alpha = \mathrm{d}\beta$ with β smooth everywhere. Exact $\Rightarrow$ closed because $\mathrm{d}^2 = 0$.
- Stokes detector: an exact form integrates to zero over every cycle without boundary. One non-zero integral proves non-exactness. For 1-forms, zero circulation around every loop is also sufficient.
- Poincaré lemma: on a star-shaped region closed forms are exact, with potential $K\alpha = \int_0^1 t^{p-1}\iota_x\alpha(tx)\,\mathrm{d}t$ and the homotopy formula $\mathrm{d}K + K\mathrm{d} = 1$.
- $\mathrm{d}\theta = (x\,\mathrm{d}y - y\,\mathrm{d}x)/r^2$ on $\mathbb{R}^2 \setminus \{0\}$ is closed, has period 2π, and is not exact. Its periods are $2\pi\times$ integer winding numbers, invariant under deformation.
- On S^2, $\omega = \mathrm{d}(-\cos\theta\,\mathrm{d}\phi)$ only away from the poles; two patches with transition $2\,\mathrm{d}\phi$ reproduce $\int \omega = 4\pi$. This is the monopole and Dirac quantisation.

T2.4 A first look at de Rham cohomology

The last section produced a pattern. On a star-shaped region every closed form is exact. On the punctured plane the closed 1-forms are exact *except* for multiples of one form, $\mathrm{d}\theta$. On the sphere the area form is closed and not exact. In each case the "closed modulo exact" forms were labelled by a few real numbers, and those numbers were integrals over cycles that wrap around holes.

De Rham cohomology is the bookkeeping device for this pattern. It is the first topological invariant in these notes that is built out of calculus, and it is the one Nash uses throughout: Betti numbers, the Euler characteristic, Hodge theory, characteristic classes and the index theorem

are all statements about it. The goal of this section is to make it completely concrete: compute it for every space we care about, with explicit representatives, and see how to read off topology from it.

T2.4.1 The definition

Working definition T2.5 (de Rham cohomology)

The p-th **de Rham cohomology** of M is the vector space

$$H^p_{\mathrm{dR}}(M) = \frac{\{\text{closed } p\text{-forms on } M\}}{\{\text{exact } p\text{-forms on } M\}} = \frac{\mathrm{Ker}\big(\mathrm{d} : \Omega^p(M) \to \Omega^{p+1}(M)\big)}{\mathrm{Im}\big(\mathrm{d} : \Omega^{p-1}(M) \to \Omega^p(M)\big)}. \tag{T2.35}$$

An element is a **cohomology class** $[\alpha]$: a closed form α, where two closed forms count as the same class when they differ by an exact form, $[\alpha] = [\alpha + \mathrm{d}\beta]$. Its dimension $b_p = \dim H^p_{\mathrm{dR}}(M)$ is the p-th **Betti number**. *Operational recipe:* (i) find the closed p-forms; (ii) find enough closed p-cycles $Z_1, Z_2, \ldots$ [p-surfaces with no boundary, e.g. loops when $p = 1$] to separate them by their **periods** $\int_{Z_i} \alpha$; (iii) show that a closed form with all periods zero is exact. Then b_p is the number of independent periods.

The quotient in (T2.35) makes sense because exact forms are closed ((T2.17)), so the denominator sits inside the numerator. Two remarks on why this is the right object.

Periods are well defined on classes. If Z is a closed cycle and α is closed, then

$$\int_Z (\alpha + \mathrm{d}\beta) \overset{(1)}{=} \int_Z \alpha + \int_{\partial Z} \beta \overset{(2)}{=} \int_Z \alpha. \tag{T2.36}$$

(1) Linearity of the integral and Stokes' theorem on the second term.
(2) $\partial Z = \varnothing$.

So a class has well-defined periods. Conversely, deforming the cycle does not change the period either: if $Z' - Z = \partial W$ for some $(p+1)$-dimensional region W, then

$$\int_{Z'} \alpha - \int_Z \alpha \overset{(1)}{=} \int_{\partial W} \alpha \overset{(2)}{=} \int_W \mathrm{d}\alpha \overset{(3)}{=} 0. \tag{T2.37}$$

(1) $\partial W = Z' - Z$.
(2) Stokes' theorem on W.
(3) α is closed.

The winding-number argument (T2.29) was the case $p = 1$. Equations (T2.36) and (T2.37) together say that integration is a well-defined pairing between cohomology classes of forms and "homology classes" of cycles, cycles modulo boundaries. This pairing is the whole content of the duality Nash sets up between his (1.43) and (1.44) [Nash p. 13]; we return to it in Section T2.4.8.

Cohomology is a linear-algebra shadow of the space. The spaces $\Omega^p(M)$ are infinite-dimensional, but on a compact manifold the quotients $H^p_{\mathrm{dR}}(M)$ are finite-dimensional, and they vanish for $p > \dim M$ (there are no forms of that degree). All the analysis cancels out; a finite list of numbers survives.

T2.4.2 H^0 counts connected components

A closed 0-form is a function with $\mathrm{d}f = 0$, i.e. with all partial derivatives zero. Such an f is constant on each connected piece of M, but may take different constants on different pieces. There are no exact 0-forms (there are no (-1)-forms to differentiate), so nothing is divided out:

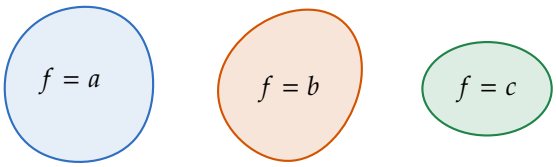
$$H^0_{\mathrm{dR}}(M) = \{\text{locally constant functions}\} \cong \mathbb{R}^{\#\text{components}}, \qquad b_0 = \text{number of connected components}. \tag{T2.38}$$

Worked example T2.16: Components

$\mathbb{R} \setminus \{0\}$. A locally constant function is one constant a on $x < 0$ and another b on $x > 0$; so $H^0_{\mathrm{dR}} \cong \mathbb{R}^2$, $b_0 = 2$. The function $\mathrm{sgn}(x)$ is closed but not constant; this is the lowest-degree cousin of $\mathrm{d}\theta$.

The orthogonal group O(3). It has two components, rotations (det $= +1$) and rotations times parity (det $= -1$). The function det is a closed 0-form taking the values ± 1, and $b_0(\mathrm{O}(3)) = 2$. The same logic applied to the Lorentz group gives $b_0 = 4$ (the signs of det and of $\Lambda^0{}_0$).

A detector for disconnectedness. For 0-forms the "cycles" are points, and the period of f at a point P is $f(P)$. The Stokes detector in degree 0 is the fundamental theorem of calculus: if a path joins P to Q then $f(Q) - f(P) = \int \mathrm{d}f = 0$. A closed f with $f(P) \neq f(Q)$ proves that no path joins them.



closed 0-forms: $\mathrm{d}f = 0$ means constant on each piece; $H^0_{\mathrm{dR}} \cong \mathbb{R}^3$

Figure T2.13: A space with three connected components. A closed function may take a different constant on each, so $H^0_{\mathrm{dR}} \cong \mathbb{R}^3$ and $b_0 = 3$. Nothing is ever divided out in degree zero.

T2.4.3 Contractible spaces have no cohomology

The Poincaré lemma (Section T2.3.3) says precisely that on a star-shaped region every closed p-form with $p \geq 1$ is exact:

$$H^p_{\mathrm{dR}}(\mathbb{R}^n) = \begin{cases} \mathbb{R}, & p = 0, \\ 0, & p \geq 1, \end{cases} \tag{T2.39}$$

and the same for balls, cubes and any star-shaped region. More generally, de Rham cohomology is a **homotopy invariant**: if one space can be continuously deformed into another (a homotopy equivalence, such as squashing a ball to a point or the punctured plane onto the unit circle), they have the same cohomology. The mechanism is the homotopy operator of (T2.23), generalised from the scaling map to any deformation; the integrals of closed forms over cycles cannot change during the deformation, by (T2.37). We shall use this in the form

$$\mathbb{R}^n \simeq \text{point}, \qquad \mathbb{R}^2 \setminus \{0\} \simeq S^1, \qquad \mathbb{R}^3 \setminus \{0\} \simeq S^2, \qquad \mathbb{R}^n \setminus \{0\} \simeq S^{n-1}. \tag{T2.40}$$

T2.4.4 The circle

Dray computes the cohomology of the circle directly [Dray (20.32)–(20.36)]; we fill in the one step he leaves out. Any 1-form on S^1 is $\alpha = f(\phi)\,\mathrm{d}\phi$ with f smooth and 2π-periodic. Every 1-form

on a 1-manifold is closed (there are no 2-forms). Define its period

$$c_\alpha := \oint_{S^1} \alpha = \int_0^{2\pi} f(\phi)\,\mathrm{d}\phi. \tag{T2.41}$$

Claim: $\alpha - \frac{c_\alpha}{2\pi}\mathrm{d}\phi$ is exact. Define $g(\phi) = \int_0^\phi \big(f(s) - \frac{c_\alpha}{2\pi}\big)\mathrm{d}s$. The only question is whether g is a function on the circle, i.e. whether $g(\phi + 2\pi) = g(\phi)$:

$$\begin{aligned} g(\phi+2\pi) - g(\phi) &\overset{(1)}{=} \int_\phi^{\phi+2\pi} \Big(f(s) - \frac{c_\alpha}{2\pi}\Big)\mathrm{d}s \\ &\overset{(2)}{=} \int_0^{2\pi} f(s)\,\mathrm{d}s - c_\alpha \\ &\overset{(3)}{=} 0. \end{aligned} \tag{T2.42}$$

(1) Subtract the two integrals from 0.
(2) f is periodic, so its integral over any interval of length 2π is the same; the constant integrates to c_α.
(3) Definition (T2.41).

So g is a smooth function on S^1 with $\mathrm{d}g = \alpha - \frac{c_\alpha}{2\pi}\mathrm{d}\phi$, i.e. $\alpha = \frac{c_\alpha}{2\pi}\mathrm{d}\phi + \mathrm{d}g$ [Dray (20.35)]. Two 1-forms are in the same class exactly when they have the same period, and the map $[\alpha] \mapsto c_\alpha$ is an isomorphism:

$$H^0_{\mathrm{dR}}(S^1) \cong \mathbb{R}, \qquad H^1_{\mathrm{dR}}(S^1) \cong \mathbb{R} \text{ (generated by } [\mathrm{d}\phi]), \qquad b_0 = b_1 = 1. \tag{T2.43}$$

Dray's summary is exactly right: the fact that this space "is one-dimensional ... turns out to imply that there is one 'hole' ".

By homotopy invariance, the punctured plane has the same answer, $H^1_{\mathrm{dR}}(\mathbb{R}^2 \setminus \{0\}) \cong \mathbb{R}$ generated by $[\mathrm{d}\theta]$, which is what we found directly in (T2.30). A plane with k punctures has $b_1 = k$: one angle form around each puncture, detected by one small loop around each.

Worked example T2.17: Fourier version

Expand $f(\phi) = \sum_{n\in\mathbb{Z}} f_n e^{in\phi}$. Then $\alpha = f_0\,\mathrm{d}\phi + \sum_{n\neq 0} f_n e^{in\phi}\mathrm{d}\phi$, and every $n \neq 0$ term is exact:

$$f_n e^{in\phi}\,\mathrm{d}\phi \overset{(1)}{=} \mathrm{d}\Big(\frac{f_n}{in} e^{in\phi}\Big).$$

(1) $\mathrm{d}e^{in\phi} = in\, e^{in\phi}\mathrm{d}\phi$, and $n \neq 0$ so we may divide.

Only the zero mode $f_0 = c_\alpha/2\pi$ cannot be integrated to a periodic function, because $\int \mathrm{d}\phi = \phi$ is not periodic. *Cohomology is the zero mode that the derivative cannot reach.* This way of thinking, where d is invertible on everything except a finite-dimensional set of modes, is exactly Hodge theory, Section T2.4.9.

T2.4.5 The sphere, and the monopole as a cohomology class

For S^2 we already know $b_0 = 1$ and that the area form ω gives a nonzero class in $H^2_{\mathrm{dR}}(S^2)$. We now show $b_1 = 0$ and $b_2 = 1$, using only two patches and the Poincaré lemma. Cover the sphere by two open caps U_N (everything except a small disc around the south pole) and U_S (everything except a small disc around the north pole). Each cap is a disc, hence star-shaped in a chart. Their overlap $U_N \cap U_S$ is a band around the equator, which deforms to a circle.

$b_1(S^2) = 0$. Let α be a closed 1-form on S^2. By the Poincaré lemma $\alpha = \mathrm{d}f_N$ on U_N and $\alpha = \mathrm{d}f_S$ on U_S. On the band $\mathrm{d}(f_N - f_S) = 0$, so $f_N - f_S$ is a constant k (the band is connected). Replace f_S by $f_S + k$; now $f_N = f_S$ on the overlap, the two glue to a global function f, and $\alpha = \mathrm{d}f$. Every closed 1-form on S^2 is exact. (In words: every loop on the sphere bounds a disc, so every circulation can be computed by Stokes as a flux of $\mathrm{d}\alpha = 0$.)

$b_2(S^2) = 1$. Let β be any 2-form on S^2 (automatically closed) with total integral $c = \int_{S^2} \beta$. Set $\beta' = \beta - \frac{c}{4\pi}\omega$, so $\int_{S^2} \beta' = 0$. We show β' is exact. By the Poincaré lemma $\beta' = \mathrm{d}A_N$ on U_N and $\beta' = \mathrm{d}A_S$ on U_S. On the band, $A_N - A_S$ is closed, hence by (T2.30) equal to $k\,\mathrm{d}\phi + \mathrm{d}g$ for a constant k and a function g on the band. The two-hemisphere computation (T2.34) gives

$$0 = \int_{S^2} \beta' = \oint_E (A_N - A_S) = 2\pi k, \tag{T2.44}$$

so $k = 0$ and $A_N - A_S = \mathrm{d}g$. Now smear g out. Take smooth functions $\rho_N, \rho_S \geq 0$ with $\rho_N + \rho_S = 1$, $\rho_N = 0$ near the south pole and $\rho_S = 0$ near the north pole (a partition of unity as in (T2.8)). Then $\rho_S g$ extends by zero to a smooth function on U_N and $\rho_N g$ to one on U_S. Define

$$A = \begin{cases} A_N - \mathrm{d}(\rho_S g) & \text{on } U_N, \\ A_S + \mathrm{d}(\rho_N g) & \text{on } U_S. \end{cases} \tag{T2.45}$$

On the overlap the two definitions agree:

$$\big(A_N - \mathrm{d}(\rho_S g)\big) - \big(A_S + \mathrm{d}(\rho_N g)\big) \overset{(1)}{=} \mathrm{d}g - \mathrm{d}\big((\rho_S + \rho_N)g\big) \overset{(2)}{=} 0.$$

(1) $A_N - A_S = \mathrm{d}g$ on the band.
(2) $\rho_N + \rho_S = 1$.

So A is a global 1-form, and $\mathrm{d}A = \beta'$ on each patch because $\mathrm{dd} = 0$. Hence $\beta = \frac{c}{4\pi}\omega + \mathrm{d}A$, and the class of a 2-form on S^2 is determined by one number, its total integral:

$$H^0_{\mathrm{dR}}(S^2) \cong \mathbb{R}, \qquad H^1_{\mathrm{dR}}(S^2) = 0, \qquad H^2_{\mathrm{dR}}(S^2) \cong \mathbb{R} \text{ (generated by } [\omega/4\pi]). \tag{T2.46}$$

Physical picture: the monopole field is the generator

Normalise the generator as $[\omega/4\pi]$, so that it integrates to 1. The field of a magnetic monopole of charge g is $F = g\,\omega$, and its class is $4\pi g$ times the generator. Every smooth deformation of the field (moving the monopole off-centre, adding smooth magnetic fields with no net flux) changes F by an exact form and leaves the class alone. The only invariant is the total flux. Dirac quantisation says this real number is actually an integer multiple of $2\pi\hbar/e$; the integer is the first Chern number of Chapter T3. The same argument, with $\mathbb{R}^3 \setminus \{0\} \simeq S^2$, shows that the closed 2-forms on $\mathbb{R}^3 \setminus \{0\}$ are classified by the enclosed charge, which is Gauss's law read as a statement about cohomology.

The pattern generalises: $H^p_{\mathrm{dR}}(S^n) \cong \mathbb{R}$ for $p = 0$ and $p = n$, and zero otherwise. The generator in top degree is the normalised volume form; on S^3 that is $\mathrm{vol}/2\pi^2$ (Example T2.4), the form whose integral counts instanton number in Chapter 8. Nash quotes this as $H^m(S^n;\mathbb{Z}) = \mathbb{Z}$ for $m = 0, n$ [Nash p. 14, (1.46)].

T2.4.6 The torus

Realise the torus as the square $[0,2\pi]\times[0,2\pi]$ with coordinates (θ,ϕ) and opposite edges identified, $T^2=\mathbb{R}^2/(2\pi\mathbb{Z})^2$. The 1-forms $\mathrm{d}\theta$ and $\mathrm{d}\phi$ are globally defined (the coordinates jump by 2π but their differentials do not), closed, and not exact. To see the last point use the two basic loops (Figure T2.14): a, along which θ runs from 0 to 2π at fixed ϕ, and b, along which ϕ runs at fixed θ. The table of periods is

$$\begin{pmatrix}\oint_a \mathrm{d}\theta & \oint_b \mathrm{d}\theta\\ \oint_a \mathrm{d}\phi & \oint_b \mathrm{d}\phi\end{pmatrix}=\begin{pmatrix}2\pi & 0\\ 0 & 2\pi\end{pmatrix}. \tag{T2.47}$$

Since this **period matrix** is invertible, no combination $x\,\mathrm{d}\theta+y\,\mathrm{d}\phi$ with $(x,y)\neq(0,0)$ has all periods zero, so no such combination is exact. Hence $b_1\geq 2$.

Every closed 1-form is $x\,\mathrm{d}\theta+y\,\mathrm{d}\phi+\mathrm{d}h$. Write a 1-form as $\alpha=f\,\mathrm{d}\theta+g\,\mathrm{d}\phi$ with f,g doubly periodic and expand $f=\sum f_{mn}e^{i(m\theta+n\phi)}$, similarly g. The condition $\mathrm{d}\alpha=0$ is $\partial_\theta g=\partial_\phi f$, i.e.

$$m\,g_{mn}=n\,f_{mn}\qquad\text{for all }(m,n)\in\mathbb{Z}^2. \tag{T2.48}$$

For each $(m,n)\neq(0,0)$ define $h_{mn}=f_{mn}/(im)$ if $m\neq 0$ and $h_{mn}=g_{mn}/(in)$ if $m=0$. Then the mode $f_{mn}\mathrm{d}\theta+g_{mn}\mathrm{d}\phi$ (times $e^{i(m\theta+n\phi)}$) equals $\mathrm{d}(h_{mn}e^{i(m\theta+n\phi)})$:

$$\begin{aligned}\mathrm{d}\big(h_{mn}e^{i(m\theta+n\phi)}\big)&\overset{(1)}{=}(im\,h_{mn}\,\mathrm{d}\theta+in\,h_{mn}\,\mathrm{d}\phi)\,e^{i(m\theta+n\phi)}\\&\overset{(2)}{=}(f_{mn}\,\mathrm{d}\theta+g_{mn}\,\mathrm{d}\phi)\,e^{i(m\theta+n\phi)}.\end{aligned}$$

(1) Differentiate the exponential.
(2) If $m\neq 0$: $im\,h_{mn}=f_{mn}$, and $in\,h_{mn}=nf_{mn}/m=g_{mn}$ by (T2.48). If $m=0$ (so $n\neq 0$): (T2.48) gives $f_{0n}=0=im\,h_{0n}$, and $in\,h_{0n}=g_{0n}$.

Summing, $\alpha=f_{00}\,\mathrm{d}\theta+g_{00}\,\mathrm{d}\phi+\mathrm{d}h$. Only the zero modes survive, and they are fixed by the periods, $f_{00}=\frac{1}{2\pi}\oint_a\alpha$ and $g_{00}=\frac{1}{2\pi}\oint_b\alpha$.

Degree 2. Every 2-form $\beta=k(\theta,\phi)\,\mathrm{d}\theta\wedge\mathrm{d}\phi$ is closed. Its non-zero Fourier modes are exact, e.g. $k_{mn}e^{i(m\theta+n\phi)}\mathrm{d}\theta\wedge\mathrm{d}\phi=\mathrm{d}\big(\frac{k_{mn}}{in}e^{i(m\theta+n\phi)}\mathrm{d}\theta\big)\cdot(-1)$ when $n\neq 0$ (check: $\mathrm{d}(e^{\cdots}\mathrm{d}\theta)=in\,e^{\cdots}\mathrm{d}\phi\wedge\mathrm{d}\theta$). The zero mode $k_{00}\,\mathrm{d}\theta\wedge\mathrm{d}\phi$ is not exact, since $\int_{T^2}\mathrm{d}\theta\wedge\mathrm{d}\phi=4\pi^2\neq 0$. So

$$H^0_{\mathrm{dR}}(T^2)\cong\mathbb{R},\qquad H^1_{\mathrm{dR}}(T^2)\cong\mathbb{R}^2\ \big([\mathrm{d}\theta],[\mathrm{d}\phi]\big),\qquad H^2_{\mathrm{dR}}(T^2)\cong\mathbb{R}\ \big([\mathrm{d}\theta\wedge\mathrm{d}\phi]\big). \tag{T2.49}$$

The Betti numbers are $(b_0,b_1,b_2)=(1,2,1)$.

Products of classes. The wedge product of two closed forms is closed, and the wedge of a closed form with an exact one is exact ($\alpha\wedge\mathrm{d}\beta=\pm\mathrm{d}(\alpha\wedge\beta)$ when $\mathrm{d}\alpha=0$). So wedge descends to classes, $[\alpha]\cup[\beta]:=[\alpha\wedge\beta]$. This is the **cup product** Nash writes in [Nash p. 14, (1.45)]. On the torus $[\mathrm{d}\theta]\cup[\mathrm{d}\phi]=[\mathrm{d}\theta\wedge\mathrm{d}\phi]$ is the generator of H^2_{dR}: the two "holes" of the torus multiply to its "volume". On S^2 there are no degree-1 classes to multiply, which is one way to see that the sphere and the torus are different spaces even though both have $b_0=b_2=1$.

T2.4.7 Betti numbers and the Euler characteristic

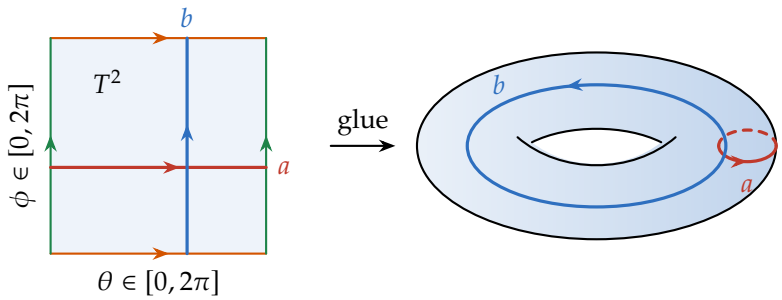


Figure T2.14: The torus as a square with opposite sides glued (orange to orange, green to green, matching arrows). The loop a goes once around in θ, the loop b once around in ϕ. The period matrix of $(\mathrm{d}\theta, \mathrm{d}\phi)$ on (a, b) is 2π times the identity, which proves that $[\mathrm{d}\theta]$ and $[\mathrm{d}\phi]$ are independent classes.

Working definition T2.6 (Euler characteristic)

For a compact manifold M of dimension n,

$$\chi(M) := \sum_{p=0}^{n} (-1)^p b_p(M). \tag{T2.50}$$

[b_p is the number of independent periods of closed p-forms, one for each independent p-dimensional hole] *Operational recipe (for a surface):* chop it into polygons [any two meet only along a whole edge or at a vertex], count
vertices V, edges E and faces F, and compute $\chi = V - E + F$. The two recipes agree (see below).

space	b_0	b_1	b_2	b_3	χ	generators
point, $\mathbb{R}^n$, disc	1	0	0	0	1	1
S^1, $\mathbb{R}^2 \setminus \{0\}$, cylinder	1	1	0	0	0	$1, \mathrm{d}\phi$
plane minus k points	1	k	0	0	$1-k$	$1, \mathrm{d}\theta_1, \ldots, \mathrm{d}\theta_k$
S^2, $\mathbb{R}^3 \setminus \{0\}$	1	0	1	0	2	$1, \omega/4\pi$
T^2	1	2	1	0	0	$1, \mathrm{d}\theta, \mathrm{d}\phi, \mathrm{d}\theta \wedge \mathrm{d}\phi$
genus-g surface Σ_g	1	$2g$	1	0	$2-2g$	$1, 2g$ loop classes, area
$S^3 \cong \mathrm{SU}(2)$	1	0	0	1	0	$1, \mathrm{vol}/2\pi^2$
T^3	1	3	3	1	0	$\mathrm{d}\theta^i, \mathrm{d}\theta^i \wedge \mathrm{d}\theta^j, \ldots$

Two patterns in the table deserve names.

- **Poincaré duality**: for a closed oriented n-manifold, $b_p = b_{n-p}$. The mechanism is that wedging a p-class with an $(n-p)$-class and integrating gives a non-degenerate pairing (on the torus: $\int \mathrm{d}\theta \wedge \mathrm{d}\phi \neq 0$). Nash uses this in Chapter 4.

- Odd-dimensional closed manifolds have $\chi = 0$, by Poincaré duality: the terms b_p and b_{n-p} enter with opposite signs.

Why $V-E+F$ equals $b_0-b_1+b_2$. Here is the linear algebra. Suppose we have any finite sequence of finite-dimensional vector spaces and linear maps with $d_p \circ d_{p-1} = 0$,

$$0 \to C^0 \xrightarrow{d_0} C^1 \xrightarrow{d_1} C^2 \xrightarrow{d_2} \cdots \xrightarrow{d_{n-1}} C^n \to 0,$$

and define its cohomology $H^p = \operatorname{Ker} d_p / \operatorname{Im} d_{p-1}$. Then

$$\begin{aligned} \sum_p (-1)^p \dim H^p &\overset{(1)}{=} \sum_p (-1)^p \left(\dim \operatorname{Ker} d_p - \dim \operatorname{Im} d_{p-1}\right) \\ &\overset{(2)}{=} \sum_p (-1)^p \dim \operatorname{Ker} d_p + \sum_p (-1)^p \dim \operatorname{Im} d_p \\ &\overset{(3)}{=} \sum_p (-1)^p \dim C^p. \end{aligned} \tag{T2.51}$$

(1) Dimension of a quotient.
(2) Shift the summation index in the second sum, $p \to p+1$, which flips the sign $(-1)^{p+1} \to -(-1)^p$.
(3) Rank–nullity for $d_p : C^p \to C^{p+1}$: $\dim \operatorname{Ker} d_p + \dim \operatorname{Im} d_p = \dim C^p$.

The alternating sum of dimensions does not care whether we take cohomology first. Now apply this twice. For a surface chopped into polygons, let C^0, C^1, C^2 be the spaces of real numbers attached to vertices, edges and faces, with d the discrete "difference across the boundary" map (value on an edge = difference of the values at its two ends, and so on). Then $\dim C^0 = V$, $\dim C^1 = E$, $\dim C^2 = F$. The theorem of de Rham (next subsection) says the cohomology of this finite complex equals H^*_{dR} of the surface. So (T2.51) gives $b_0 - b_1 + b_2 = V - E + F$. This also explains the claim, used by Dray, that $V-E+F$ does not depend on how the surface is chopped up [Dray §18.10].

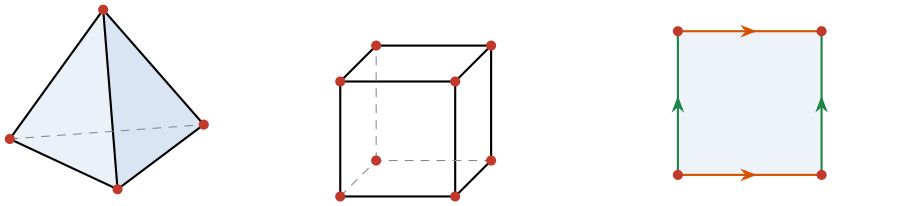
Worked example T2.18: Counting cells

Sphere. A tetrahedron: $V=4$, $E=6$, $F=4$, so $\chi = 2$. A cube: $8-12+6=2$. Dray's decomposition into four "quadrilateral" lunes plus two polar caps: $8-12+6=2$ [Dray (18.77)]. The table predicted $b_0-b_1+b_2 = 1-0+1 = 2$.
Torus. The square with identified sides: all four corners are the same point, so $V=1$; the four sides are two edges (a and b), $E=2$; one face, $F=1$. So $\chi = 1-2+1 = 0 = 1-2+1$. A finer grid of $m \times n$ squares on the torus gives $V=mn$, $E=2mn$, $F=mn$: still 0.
Adding a handle. Cut two discs out of a surface (each removes one face) and join them by a tube made of two quadrilaterals glued along two new edges [Dray Fig. 18.10]. Net change: $\Delta F = -2+2 = 0$, $\Delta E = +2$ (the vertical seams; the circles were already edges), $\Delta V = 0$. So $\Delta\chi = -2$. Starting from the sphere, g handles give $\chi(\Sigma_g) = 2-2g$.

T2.4.8 De Rham's theorem and homology

We have been using cycles freely. Here is the one-paragraph version of the dual theory that Nash writes as (1.44). A **p-chain** is a formal combination of oriented p-dimensional pieces (curves, surfaces, ...) with real coefficients; its boundary ∂ is defined piece by piece with the orientation of Section T2.2.1. Cycles are chains with $\partial Z = 0$; boundaries are chains of the form ∂W. Because $\partial\partial = 0$ (Section T2.2.5), boundaries are cycles, and the **homology** $H_p(M) = \{\text{cycles}\}/\{\text{boundaries}\}$ makes sense. Integration gives a pairing

$$H^p_{\mathrm{dR}}(M) \times H_p(M) \to \mathbb{R}, \qquad \big([\alpha],[Z]\big) \mapsto \int_Z \alpha, \tag{T2.52}$$



tetrahedron
$4 - 6 + 4 = 2$

cube
$8 - 12 + 6 = 2$

torus: $V = 1, E = 2, F = 1$
$1 - 2 + 1 = 0$

Figure T2.15: Three cell decompositions. The tetrahedron and the cube are both spheres and both give $V - E + F = 2$. The torus as a square with glued sides has one vertex (all four red corners are identified), two edges and one face, giving 0. The count never depends on the decomposition, only on the surface.

which is well defined by (T2.36) and (T2.37); Stokes' theorem is exactly the statement that d and ∂ are adjoint under this pairing.

Key idea: de Rham's theorem

The pairing (T2.52) is perfect: $H^p_{\mathrm{dR}}(M)$ is the dual of $H_p(M;\mathbb{R})$, so it is isomorphic to the topologist's real cohomology $H^p(M;\mathbb{R})$ [Nash p. 13, (1.43)]. Concretely: *a closed form is exact if and only if all of its periods vanish*, and any assignment of periods to a basis of cycles is realised by some closed form.

For 1-forms we proved the first half in (T2.19); the computations for S^1, S^2 and T^2 above are the theorem in action. The practical content for physicists: *topological charges are periods*. Winding numbers, magnetic fluxes, instanton numbers and index densities integrated over closed manifolds are all periods of closed forms, and so they cannot change under smooth deformations.

T2.4.9 Hodge theory: a best representative in each class

A cohomology class contains infinitely many forms. On a closed Riemannian manifold there is a best one. Using the inner product $\langle \alpha, \beta \rangle = \int_M \alpha \wedge \star\beta$ and the adjoint d^* of Section T2.2.6, define the **Hodge Laplacian** $\Delta_p = \mathrm{d}\mathrm{d}^* + \mathrm{d}^*\mathrm{d}$ on p-forms. A form is **harmonic** if $\Delta\omega = 0$. Nash gives a two-line argument, which we spell out [Nash p. 40, (2.50)]:

$$\begin{aligned} 0 = \langle \omega, \Delta_p \omega \rangle &\overset{(1)}{=} \langle \omega, \mathrm{d}\mathrm{d}^*\omega \rangle + \langle \omega, \mathrm{d}^*\mathrm{d}\omega \rangle \\ &\overset{(2)}{=} \langle \mathrm{d}^*\omega, \mathrm{d}^*\omega \rangle + \langle \mathrm{d}\omega, \mathrm{d}\omega \rangle \\ &\overset{(3)}{=} \|\mathrm{d}^*\omega\|^2 + \|\mathrm{d}\omega\|^2 . \end{aligned} \tag{T2.53}$$

(1) Definition of Δ_p and linearity.
(2) Move one d or d^* across using the adjoint property (integration by parts (T2.15) with no boundary term, since M is closed).
(3) Each term is the squared norm of a form, which is non-negative in Euclidean signature.

A sum of non-negative terms vanishes only if each does, so $\Delta\omega = 0 \iff \mathrm{d}\omega = 0$ and $\mathrm{d}^*\omega = 0$ [Nash (2.51)]. In particular a harmonic form is closed and defines a class [Nash (2.52)].

Key idea: Hodge's theorem (Nash p. 41)

On a compact closed Riemannian manifold there is exactly one harmonic form in each de Rham class. Hence

$$b_p = \dim H^p_{\mathrm{dR}}(M) = \dim \operatorname{Ker} \Delta_p. \tag{T2.54}$$

The left side is topological; the right side is the number of zero modes of a differential operator built from a metric. The answer does not depend on which metric we pick [Nash p. 41, (2.53)].

Worked example T2.19: Harmonic forms on the flat torus and the round sphere

On the flat torus with $\mathrm{d}s^2 = \mathrm{d}\theta^2 + \mathrm{d}\phi^2$ the Hodge Laplacian acts on the coefficients of a form as $-(\partial_\theta^2 + \partial_\phi^2)$ (the coframe is constant, so nothing mixes). On a Fourier mode,

$$\Delta\big(e^{i(m\theta+n\phi)}\mathrm{d}\theta\big) = (m^2+n^2)\, e^{i(m\theta+n\phi)}\mathrm{d}\theta,$$

which vanishes only for $m = n = 0$. So the harmonic 1-forms are exactly the constant combinations $x\,\mathrm{d}\theta + y\,\mathrm{d}\phi$: $\dim \operatorname{Ker}\Delta_1 = 2 = b_1$. Likewise $\operatorname{Ker}\Delta_0$ is the constants and $\operatorname{Ker}\Delta_2$ is spanned by $\mathrm{d}\theta \wedge \mathrm{d}\phi$, reproducing $(1, 2, 1)$. The harmonic representative of a class is its zero mode, exactly as in Example T2.17.
On the round S^2, $\operatorname{Ker}\Delta_0$ is the constants (Example T2.10); $\operatorname{Ker}\Delta_2$ is spanned by ω (because $\star\omega = 1$ and $\star$ commutes with Δ); and $\operatorname{Ker}\Delta_1 = 0$ because the spherical-harmonic eigenvalues $\ell(\ell+1)$ of $\mathrm{d}Y_\ell$ and $\star\mathrm{d}Y_\ell$ are never zero for $\ell \geq 1$. This matches $(1, 0, 1)$.

Physical picture: zero modes count topology

Equation (T2.54) is the prototype of every "index theorem" statement. Physicists count zero modes of an operator (here Δ, later the Dirac operator) and find a topological number. Combining it with the Euler characteristic gives $\chi(M) = \sum(-1)^p \dim \operatorname{Ker}\Delta_p$, which Nash rewrites as the index of the operator $\mathrm{d} + \mathrm{d}^*$ from even forms to odd forms [Nash p. 104, (4.52)]. The Gauss–Bonnet theorem of Section T2.8 computes the same number as an integral of curvature.

Reading Nash (pp. 13–14 and 40–41)

Nash defines cohomology in general, $H^n = \ker d_n / \operatorname{Im} d_{n-1}$ [Nash p. 13, (1.42)], and then specialises to the de Rham complex $\cdots \to \Omega^{n-1} \to \Omega^n \to \Omega^{n+1} \to \ldots$, asserting the isomorphism with real cohomology "of course" [Nash (1.43)]. The dual homology complex, with $\partial_{n-1} \circ \partial_n = 0$, is his (1.44); the cup product $[\omega] \cup [\nu] = [\omega \wedge \nu]$ is (1.45); the cohomology of spheres is (1.46). In Chapter 2 he returns to the Laplacian $\Delta_p = \mathrm{d}\mathrm{d}^* + \mathrm{d}^*\mathrm{d}$, proves (2.50)–(2.51) exactly as in (T2.53), and quotes Hodge's theorem to obtain $\dim \ker \Delta_p = b_p$ [Nash (2.53)]. Everything in this section is the working physicist's version of those five equations.

Recap

- H^p_{dR} = closed / exact. Classes have well-defined periods over cycles, and periods do not change when cycles are deformed.
- b_0 = number of components; $H^p_{\mathrm{dR}}(\mathbb{R}^n) = 0$ for $p \geq 1$; cohomology is homotopy invariant.

- S^1: $(1,1)$, generator $\mathrm{d}\phi$. S^2: $(1,0,1)$, generator $\omega/4\pi$, the unit monopole. T^2: $(1,2,1)$, generators $\mathrm{d}\theta$, $\mathrm{d}\phi$, $\mathrm{d}\theta\wedge\mathrm{d}\phi$. Fourier: cohomology is the zero mode d cannot reach.
- $\chi=\sum(-1)^p b_p = V-E+F$, by rank–nullity; $\chi(\Sigma_g)=2-2g$.
- de Rham: a closed form is exact iff all periods vanish.
- Hodge: one harmonic form per class; $b_p=\dim\mathrm{Ker}\,\Delta_p$ (zero modes count topology).

T2.5 Connections: how a frame turns from point to point

So far every derivative in these notes has acted on *forms*: we took d of functions, of 1-forms, of 2-forms. Physics also needs derivatives of *vectors*. Acceleration is the rate of change of velocity. The divergence of the electric field, the change of a spin as it is carried around, the equation of motion of a particle in a gravitational field: all of these differentiate a vector field.

In Cartesian coordinates this is trivial. Write $\vec{F}=F^x\hat{x}+F^y\hat{y}$ and differentiate the components, because $\hat{x}$ and $\hat{y}$ are the same at every point. In polar coordinates it is already not trivial: $\hat{r}$ and $\hat{\phi}$ point in different directions at different points, so a vector field with *constant* components $(F^r,F^\phi)=(0,1)$ is not a constant vector field. On a sphere it is worse still: there is no Cartesian frame at all, so there is nothing simple to compare with.

The question of this section is therefore:

> *How do the basis vectors change from point to point, and how do we encode that information so that we can differentiate any vector field?*

The answer is a matrix of 1-forms, the **connection**. We will find that a surface (or any Riemannian manifold) comes with a preferred one, the Levi-Civita connection, and that it can be computed by a few lines of exterior calculus. We will also find that the freedom to rotate the frame is exactly a gauge freedom, and that the connection transforms exactly like a gauge potential. That observation is the bridge from this chapter to Yang–Mills theory (Chapter T3).

We follow Dray, Chapter 17, closely [Dray §§17.1–17.10].

T2.5.1 A motivating example: the polar basis vectors rotate

Take the plane with polar coordinates (r,ϕ) and the unit vectors $\hat{r}$ (pointing away from the origin) and $\hat{\phi}$ (pointing counterclockwise). How do they change as we move?

Geometric answer. If we move radially, nothing happens: at (r,ϕ) and $(r+\mathrm{d}r,\phi)$ the vectors $\hat{r}$ and $\hat{\phi}$ are parallel copies of each other. If we move around a circle through an angle $\mathrm{d}\phi$, both vectors rotate through the same angle $\mathrm{d}\phi$ (Figure T2.16). The difference between two unit vectors separated by a small angle $\mathrm{d}\phi$ is a short chord joining their tips. In the limit the chord becomes tangent to the unit circle, so it is perpendicular to the original vector, and its length is the arc length $1\cdot\mathrm{d}\phi$. For $\hat{r}$ the chord points along $\hat{\phi}$; for $\hat{\phi}$ it points along $-\hat{r}$. So, along any path,

$$\mathrm{d}\hat{r}=\mathrm{d}\phi\;\hat{\phi},\qquad \mathrm{d}\hat{\phi}=-\mathrm{d}\phi\;\hat{r} \tag{T2.55}$$

[Dray (17.2)–(17.3)]. (Radial motion contributes nothing, so the formulae hold along an arbitrary path.)

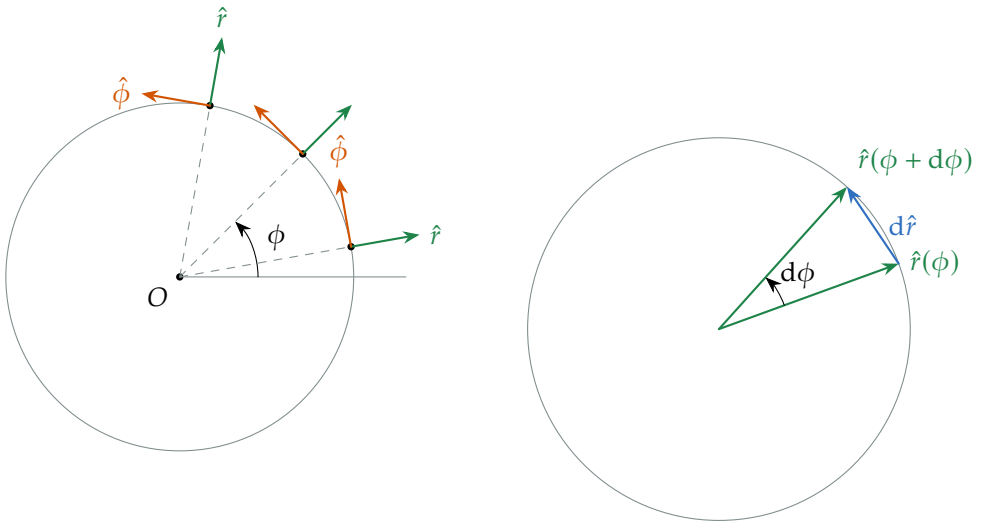


Figure T2.16: Left: the polar frame $(\hat{r}, \hat{\phi})$ at three points on a circle. Moving around the circle by $d\phi$ rotates both vectors by $d\phi$; moving radially changes nothing. Right: the change $d\hat{r}$ is the chord between the tips of two unit vectors, which in the limit is perpendicular to $\hat{r}$ (so along $\hat{\phi}$) with length $d\phi$.

Algebraic answer. Expand in the constant Cartesian frame:

$$\begin{aligned} d\hat{r} &\overset{(1)}{=} d\big(\cos\phi\,\hat{x} + \sin\phi\,\hat{y}\big) \\ &\overset{(2)}{=} -\sin\phi\,d\phi\,\hat{x} + \cos\phi\,d\phi\,\hat{y} \\ &\overset{(3)}{=} d\phi\,\hat{\phi}, \end{aligned}$$

(1) $\hat{r} = \cos\phi\,\hat{x} + \sin\phi\,\hat{y}$ and $\hat{\phi} = -\sin\phi\,\hat{x} + \cos\phi\,\hat{y}$.
(2) $\hat{x}, \hat{y}$ are constant, so only the coefficient functions are differentiated.
(3) Recognise the bracket as $\hat{\phi}$.

and in the same way $d\hat{\phi} = (-\cos\phi\,\hat{x} - \sin\phi\,\hat{y})d\phi = -d\phi\,\hat{r}$.

Watch out: a typo in Dray

Dray's (17.4)–(17.5) read $\hat{r} = r\cos\phi\,\hat{x} + r\sin\phi\,\hat{y}$, with a stray factor of r. That would be the position vector, not the unit vector. The correct expressions are the ones above, without r.

A third answer, which generalises. Differentiate the position vector $\vec{r} = r\hat{r}$:

$$d\vec{r} \overset{(1)}{=} dr\,\hat{r} + r\,d\hat{r},$$

(1) Product rule: $\vec{r}$ is a function r times the vector field $\hat{r}$.

But we also know from geometry that the infinitesimal displacement is $d\vec{r} = dr\,\hat{r} + r\,d\phi\,\hat{\phi}$ (move dr outwards and an arc $r\,d\phi$ around). Comparing gives $d\hat{r} = d\phi\,\hat{\phi}$. For $d\hat{\phi}$ we use only dot products:

$$\begin{aligned} 0 &\overset{(1)}{=} d(\hat{\phi}\cdot\hat{\phi}) \overset{(2)}{=} 2\,\hat{\phi}\cdot d\hat{\phi}, \\ 0 &\overset{(3)}{=} d(\hat{\phi}\cdot\hat{r}) \overset{(4)}{=} d\hat{\phi}\cdot\hat{r} + \hat{\phi}\cdot d\hat{r} \overset{(5)}{=} d\hat{\phi}\cdot\hat{r} + d\phi. \end{aligned}$$

(1) $\hat{\phi}$ has unit length everywhere, so $\hat{\phi}\cdot\hat{\phi}=1$ is constant.
(2) Product rule for the dot product.
(3) $\hat{\phi}\perp\hat{r}$ everywhere.
(4) Product rule again.
(5) $\mathrm{d}\hat{r}=\mathrm{d}\phi\,\hat{\phi}$ and $\hat{\phi}\cdot\hat{\phi}=1$.

The first line says $\mathrm{d}\hat{\phi}$ has no $\hat{\phi}$ component; the second says its $\hat{r}$ component is $-\mathrm{d}\phi$. Hence $\mathrm{d}\hat{\phi}=-\mathrm{d}\phi\,\hat{r}$ [Dray (17.6)–(17.10)]. This third argument never used the Cartesian frame. It used only (i) the formula for $\mathrm{d}\vec{r}$ and (ii) the fact that the frame is orthonormal. Those are exactly the two ingredients that survive on a curved surface, where there is no Cartesian frame and not even a position vector. The rest of this section turns the argument into a machine.

Key idea: derivative of a unit vector

If $\hat{u}$ has constant length, then $\hat{u}\cdot\mathrm{d}\hat{u}=\frac{1}{2}\mathrm{d}(\hat{u}\cdot\hat{u})=0$: *the change of a unit vector is always perpendicular to it.* A unit vector can only rotate. So the change of an orthonormal frame is an infinitesimal rotation, described by an antisymmetric matrix.

T2.5.2 Vector-valued forms and orthonormal coframes (vielbeins)

Vector-valued forms. The infinitesimal displacement $\mathrm{d}\vec{r}$ is both a 1-form and a vector: fed a direction of motion it returns a vector (the displacement in that direction). Dray calls such objects **vector-valued differential forms** [Dray §17.2]. In a basis $\{e_i\}$ of vector fields, a vector-valued p-form is $\alpha^i e_i$ with each α^i an ordinary p-form. A vector field $\vec{v}=v^i e_i$ is a vector-valued 0-form.

Working definition T2.7 (orthonormal frame and coframe)

An **orthonormal frame** on a Riemannian manifold [a manifold with a dot product of tangent vectors at each point, read off from $\mathrm{d}s^2$] is a set of n vector fields $e_1,\dots,e_n$ with $e_i\cdot e_j=\delta_{ij}$ at every point (of some region). The **dual coframe** is the set of 1-forms $\sigma^1,\dots,\sigma^n$ defined by

$$\mathrm{d}\vec{r}=\sigma^i\,e_i,\qquad\text{equivalently}\qquad \sigma^i(e_j)=\delta^i{}_j. \tag{T2.56}$$

How to find them in practice: write the line element as a sum of squares, $\mathrm{d}s^2=\mathrm{d}\vec{r}\cdot\mathrm{d}\vec{r}=(\sigma^1)^2+\cdots+(\sigma^n)^2$; the 1-forms being squared are an orthonormal coframe, and e_i are the unit vectors along the corresponding directions.

The metric is recovered from the coframe, because the frame is orthonormal:

$$\mathrm{d}s^2\overset{(1)}{=}\mathrm{d}\vec{r}\cdot\mathrm{d}\vec{r}\overset{(2)}{=}(\sigma^i e_i)\cdot(\sigma^j e_j)\overset{(3)}{=}\delta_{ij}\,\sigma^i\sigma^j. \tag{T2.57}$$

(1) Definition of the line element [Dray (17.21)].
(2) Expand $\mathrm{d}\vec{r}$ in the frame, (T2.56).
(3) Orthonormality, $e_i\cdot e_j=\delta_{ij}$. (In Lorentzian signature δ_{ij} becomes η_{ij}.)

The coframe as a matrix. Write $\sigma^i=E^i{}_\mu\,\mathrm{d}x^\mu$. The $n\times n$ matrix $E=(E^i{}_\mu)$ is what physicists call the **vielbein** (*zweibein* in 2D, *vierbein* in 4D) and write $e^a{}_\mu$. Its defining property is that it turns the coordinate metric into the identity:

$$g_{\mu\nu}=\delta_{ij}\,E^i{}_\mu E^j{}_\nu,\qquad\text{i.e.}\qquad g=E^{\mathsf{T}}E. \tag{T2.58}$$

The vielbein is a "square root of the metric". The inverse matrix E^{-1} gives the frame vectors in terms of coordinate derivatives, $e_i = (E^{-1})^{\mu}{}_i\,\partial_\mu$. Nash uses exactly this object, $e^a_\mu(x)e^b_\nu(x)\delta_{ab} = g_{\mu\nu}(x)$, to build the Dirac operator [Nash p. 114, (4.100)].

Worked example T2.20: Coframes of the surfaces we will use all chapter

In each case read off the coframe from the line element.
Plane, polar coordinates. $\mathrm{d}s^2 = \mathrm{d}r^2 + r^2\mathrm{d}\phi^2$, so

$$\sigma^1 = \mathrm{d}r, \quad \sigma^2 = r\,\mathrm{d}\phi, \qquad e_1 = \hat{r} = \partial_r, \quad e_2 = \hat{\phi} = \tfrac{1}{r}\partial_\phi, \qquad E = \begin{pmatrix} 1 & 0 \\ 0 & r \end{pmatrix}.$$

Sphere of radius a. $\mathrm{d}s^2 = a^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)$, so

$$\sigma^1 = a\,\mathrm{d}\theta, \quad \sigma^2 = a\sin\theta\,\mathrm{d}\phi, \qquad e_1 = \hat{\theta} = \tfrac{1}{a}\partial_\theta, \quad e_2 = \hat{\phi} = \tfrac{1}{a\sin\theta}\partial_\phi, \qquad E = \begin{pmatrix} a & 0 \\ 0 & a\sin\theta \end{pmatrix}.$$

Torus of revolution. Let the tube (radius a) have its centre on a circle of radius $R > a$ in the xy-plane: $\vec{r} = \big((R + a\cos\theta)\cos\phi,\ (R + a\cos\theta)\sin\phi,\ a\sin\theta\big)$, where θ goes around the tube (starting on the outer equator) and ϕ goes around the z-axis. Moving by $\mathrm{d}\theta$ traces an arc $a\,\mathrm{d}\theta$ of the tube; moving by $\mathrm{d}\phi$ traces an arc of a horizontal circle of radius $R + a\cos\theta$. These are perpendicular, so [Dray (18.81)]

$$\mathrm{d}s^2 = a^2\mathrm{d}\theta^2 + (R + a\cos\theta)^2\mathrm{d}\phi^2, \qquad \sigma^1 = a\,\mathrm{d}\theta, \quad \sigma^2 = (R + a\cos\theta)\,\mathrm{d}\phi.$$

Euclidean $\mathbb{R}^3$, spherical coordinates. $\mathrm{d}s^2 = \mathrm{d}r^2 + r^2\mathrm{d}\theta^2 + r^2\sin^2\theta\,\mathrm{d}\phi^2$, so $(\sigma^1, \sigma^2, \sigma^3) = (\mathrm{d}r,\ r\,\mathrm{d}\theta,\ r\sin\theta\,\mathrm{d}\phi)$ and $(e_1, e_2, e_3) = (\hat{r}, \hat{\theta}, \hat{\phi})$.

The frame is not unique. At each point we may rotate the orthonormal frame by any rotation matrix, and the rotation may depend on the point. In 2D, rotating by an angle $\lambda(x)$ gives another perfectly good orthonormal frame. This freedom will turn out to be a gauge freedom (Section T2.5.10). The Cartesian frame $(\hat{x}, \hat{y})$ and the polar frame $(\hat{r}, \hat{\phi})$ of the plane are related in exactly this way, with $\lambda = \phi$.

T2.5.3 Differentiating vector fields

A vector field is a vector-valued 0-form, so its exterior derivative should be a vector-valued 1-form: fed a direction, it returns how the vector changes in that direction. In a Cartesian frame [Dray (17.11)–(17.13)],

$$\mathrm{d}\vec{F} = \mathrm{d}F^x\,\hat{x} + \mathrm{d}F^y\,\hat{y},$$

because $\mathrm{d}\hat{x} = \mathrm{d}\hat{y} = 0$. In a general frame the product rule forces

$$\mathrm{d}\vec{F} = \mathrm{d}(F^i e_i) = \mathrm{d}F^i\,e_i + F^i\,\mathrm{d}e_i, \tag{T2.59}$$

and we need to know $\mathrm{d}e_i$ [Dray (17.15)]. The properties we demand of d on vector-valued forms are [Dray (17.16)–(17.18)]:

$$\begin{aligned} \mathrm{d}(a\vec{v} + \vec{w}) &= a\,\mathrm{d}\vec{v} + \mathrm{d}\vec{w} \quad (a \text{ constant}), \\ \mathrm{d}(\alpha\,\vec{v}) &= \mathrm{d}\alpha\,\vec{v} + (-1)^p\alpha \wedge \mathrm{d}\vec{v} \quad (\alpha \text{ a } p\text{-form}), \\ \mathrm{d}(\vec{v}\,\alpha) &= \mathrm{d}\vec{v} \wedge \alpha + \vec{v}\,\mathrm{d}\alpha. \end{aligned} \tag{T2.60}$$

The sign in the middle line is the usual graded Leibniz rule: d has to move past the p-form α to reach $\vec{v}$. The last line has no sign because $\vec{v}$ is a 0-form. These rules do not determine $\mathrm{d}e_i$. That extra information is what we now name.

T2.5.4 Connection 1-forms

Working definition T2.8 (connection 1-forms)

Given a frame $\{e_j\}$ [n vector fields forming a basis at each point], the **connection 1-forms** $\omega^i{}_j$ are defined by expanding the change of each frame vector in the frame:

$$\mathrm{d}e_j = e_i\,\omega^i{}_j, \qquad \text{equivalently} \qquad \omega^i{}_j = e_i \cdot \mathrm{d}e_j \quad \text{(orthonormal frame)}. \tag{T2.61}$$

Read it as: $\omega^i{}_j(\vec{v})$ is the i-th component of the rate of change of e_j when we move with velocity $\vec{v}$. Collect them into an $n \times n$ matrix of 1-forms $\omega = (\omega^i{}_j)$ [fed a velocity $\vec{v}$, it gives an ordinary matrix, antisymmetric when the frame is orthonormal] (row i, column j).

Dray writes (T2.61) as $\mathrm{d}e_j = \omega^i{}_j\, e_i$ [Dray (17.19)]; the order does not matter because e_i is a 0-form. A compact matrix notation is useful. Arrange the frame as a *row* $\mathbf{e} = (e_1, \ldots, e_n)$ and the coframe as a *column* $\boldsymbol{\sigma} = (\sigma^1, \ldots, \sigma^n)^{\mathsf{T}}$. Then

$$\mathrm{d}\vec{r} = \mathbf{e}\,\boldsymbol{\sigma}, \qquad \mathrm{d}\mathbf{e} = \mathbf{e}\,\omega. \tag{T2.62}$$

With this, (T2.59) becomes, for $\vec{F} = F^j e_j$,

$$\mathrm{d}\vec{F} \overset{(1)}{=} \mathrm{d}F^j\, e_j + F^j\, e_i\, \omega^i{}_j \overset{(2)}{=} \left(\mathrm{d}F^i + \omega^i{}_j\, F^j\right) e_i. \tag{T2.63}$$

(1) Product rule (T2.59) and the definition (T2.61).
(2) Rename the dummy index $j \to i$ in the first term and collect the coefficient of e_i.

So the components of $\mathrm{d}\vec{F}$ are not $\mathrm{d}F^i$ but $\mathrm{d}F^i + \omega^i{}_j F^j$. The connection is the correction that accounts for the turning of the frame.

Polar coordinates. From (T2.55), with $e_1 = \hat{r}$ and $e_2 = \hat{\phi}$: $\mathrm{d}e_1 = e_2\,\mathrm{d}\phi$, so $\omega^2{}_1 = \mathrm{d}\phi$, $\omega^1{}_1 = 0$; and $\mathrm{d}e_2 = -e_1\,\mathrm{d}\phi$, so $\omega^1{}_2 = -\mathrm{d}\phi$, $\omega^2{}_2 = 0$. As a matrix,

$$\omega_{\text{polar}} = \begin{pmatrix} 0 & -\mathrm{d}\phi \\ \mathrm{d}\phi & 0 \end{pmatrix} = \mathrm{d}\phi \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}. \tag{T2.64}$$

The matrix $\left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ is the generator of counterclockwise rotations in the plane. The connection says: "moving by $\mathrm{d}\phi$ rotates the frame by the angle $\mathrm{d}\phi$".

Worked example T2.21: Constant components, non-constant field; and the reverse

(a) The field $\vec{F} = \hat{\phi}$. Its polar components are $(F^1, F^2) = (0, 1)$, both constant. Yet by (T2.63),

$$\begin{aligned} \mathrm{d}\vec{F} &\overset{(1)}{=} \left(\mathrm{d}F^1 + \omega^1{}_2 F^2\right) e_1 + \left(\mathrm{d}F^2 + \omega^2{}_1 F^1\right) e_2 \\ &\overset{(2)}{=} (0 - \mathrm{d}\phi)\,\hat{r} + (0 + 0)\,\hat{\phi} = -\mathrm{d}\phi\,\hat{r}. \end{aligned}$$

(1) Expand (T2.63); the diagonal $\omega^i{}_i$ vanish.
(2) $F^1 = 0$, $F^2 = 1$ are constants, and $\omega^1{}_2 = -\mathrm{d}\phi$.

Feeding in the unit vector $\hat{\phi}$ (which has $\mathrm{d}\phi(\hat{\phi}) = 1/r$) gives the rate of change of $\hat{\phi}$ as we

move along a circle at unit speed: $-\hat{r}/r$. This is the centripetal acceleration v^2/r of uniform circular motion with $v = 1$, pointing inwards.
(b) The field $\vec{F} = \hat{x}$. In the polar frame, $\hat{x} = \cos\phi\,\hat{r} - \sin\phi\,\hat{\phi}$, so its components vary. But

$$\begin{aligned} \mathrm{d}\hat{x} &\overset{(1)}{=} \big(\mathrm{d}(\cos\phi) + \omega^1{}_2(-\sin\phi)\big)\hat{r} + \big(\mathrm{d}(-\sin\phi) + \omega^2{}_1\cos\phi\big)\hat{\phi} \\ &\overset{(2)}{=} \big(-\sin\phi\,\mathrm{d}\phi + \sin\phi\,\mathrm{d}\phi\big)\hat{r} + \big(-\cos\phi\,\mathrm{d}\phi + \cos\phi\,\mathrm{d}\phi\big)\hat{\phi} = 0. \end{aligned}$$

(1) (T2.63) with $F^1 = \cos\phi$, $F^2 = -\sin\phi$.
(2) $\omega^1{}_2 = -\mathrm{d}\phi$, $\omega^2{}_1 = +\mathrm{d}\phi$.
The change of the components exactly cancels the turning of the frame, as it must for a constant field.

T2.5.5 Metric compatibility: the connection is a rotation generator

We impose two conditions that single out the connection geometry gives us. The first is that differentiation respects the dot product [Dray (17.24)]:

$$\mathrm{d}(\vec{v}\cdot\vec{w}) = \mathrm{d}\vec{v}\cdot\vec{w} + \vec{v}\cdot\mathrm{d}\vec{w}. \tag{T2.65}$$

Apply it to the frame:

$$0 \overset{(1)}{=} \mathrm{d}(e_i\cdot e_j) \overset{(2)}{=} \mathrm{d}e_i\cdot e_j + e_i\cdot\mathrm{d}e_j \overset{(3)}{=} \omega^j{}_i + \omega^i{}_j. \tag{T2.66}$$

(1) $e_i\cdot e_j = \delta_{ij}$ is constant.
(2) The compatibility condition (T2.65).
(3) $\mathrm{d}e_i\cdot e_j = (e_k\,\omega^k{}_i)\cdot e_j = \omega^j{}_i$, and similarly for the second term.

So in an orthonormal frame (Euclidean signature, where indices can be raised and lowered freely) the matrix ω is *antisymmetric*: $\omega_{ij} = -\omega_{ji}$ and the diagonal vanishes [Dray (17.25)]. An antisymmetric $n\times n$ matrix is an element of the Lie algebra $\mathfrak{so}(n)$ of the rotation group. So:

Key idea: the connection is $\mathfrak{so}(n)$-valued

In an orthonormal frame the connection ω is an $\mathfrak{so}(n)$-valued 1-form: for each direction of motion it returns the infinitesimal rotation that the frame undergoes. In 2D, $\mathfrak{so}(2)$ is one-dimensional, so the whole connection is a single 1-form $\omega^1{}_2 = -\omega^2{}_1$. In 3D there are three: $\omega^1{}_2$, $\omega^1{}_3$, $\omega^2{}_3$. In 4D there are six, matching the six generators of SO(4) (or of the Lorentz group SO(3, 1)).

T2.5.6 Torsion-free: Cartan's first structure equation

The second condition concerns d^2. For forms $\mathrm{d}^2 = 0$ always. For vector-valued forms we cannot demand that in general (we will see in Section T2.6 that $\mathrm{d}^2\vec{v} \neq 0$ is precisely curvature). But for the displacement $\mathrm{d}\vec{r}$ there is a strong argument. In Euclidean space $\mathrm{d}\vec{r}$ really is d of the position vector, so $\mathrm{d}(\mathrm{d}\vec{r}) = 0$. On a sphere there is no position vector, but $\mathrm{d}\vec{r}$ still makes sense, and we *demand* [Dray (17.27)]

$$\mathrm{d}(\mathrm{d}\vec{r}) = 0. \tag{T2.67}$$

Expand it using the product rules:

$$\begin{aligned} 0 &\overset{(1)}{=} \mathrm{d}(\sigma^j e_j) \\ &\overset{(2)}{=} \mathrm{d}\sigma^j\, e_j - \sigma^j \wedge \mathrm{d}e_j \\ &\overset{(3)}{=} \mathrm{d}\sigma^i\, e_i - \sigma^j \wedge \omega^i{}_j\, e_i \\ &\overset{(4)}{=} \left(\mathrm{d}\sigma^i + \omega^i{}_j \wedge \sigma^j\right) e_i. \end{aligned} \tag{T2.68}$$

(1) (T2.67) and $\mathrm{d}\vec{r} = \sigma^j e_j$.
(2) The graded product rule (T2.60) with $p = 1$ [Dray (17.28)].
(3) Insert $\mathrm{d}e_j = e_i\, \omega^i{}_j$ and rename $j \to i$ in the first term.
(4) $\sigma^j \wedge \omega^i{}_j = -\omega^i{}_j \wedge \sigma^j$ (two 1-forms anticommute).

Each component must vanish:

$$\boxed{\mathrm{d}\sigma^i + \omega^i{}_j \wedge \sigma^j = 0} \qquad \text{or, in matrices,} \qquad \mathrm{d}\boldsymbol{\sigma} + \omega \wedge \boldsymbol{\sigma} = 0. \tag{T2.69}$$

This is **Cartan's first structure equation**. More generally the left side is the **torsion** 2-form $\Theta^i = \mathrm{d}\sigma^i + \omega^i{}_j \wedge \sigma^j$ [Dray (18.26)], and (T2.69) says the connection is **torsion free**.

Working definition T2.9 (Levi-Civita connection)

The **Levi-Civita connection** of a metric is the unique connection [a choice of the 1-forms $\omega^i{}_j$, i.e. of the rule $\mathrm{d}e_j = e_i\, \omega^i{}_j$] that is

1. metric compatible: $\omega_{ij} = -\omega_{ji}$ in an orthonormal frame [moving in any direction keeps $e_i \cdot e_j = \delta_{ij}$, so the frame only rotates], and
2. torsion free: $\mathrm{d}\sigma^i = -\omega^i{}_j \wedge \sigma^j$.

How to compute it: write down the coframe, compute each $\mathrm{d}\sigma^i$, and guess antisymmetric $\omega^i{}_j$ that reproduce them. If your guess satisfies all n equations, it is *the* answer, by uniqueness.

Why it is unique: counting. In n dimensions an antisymmetric matrix of 1-forms has $\frac{1}{2}n(n-1)$ independent entries, each with n components, so $\frac{1}{2}n^2(n-1)$ unknown functions. The torsion condition is n equations between 2-forms, each 2-form having $\frac{1}{2}n(n-1)$ components, so $\frac{1}{2}n^2(n-1)$ linear equations. The numbers match [Dray §17.6]. The explicit formula of Section T2.5.8 shows that the linear system is indeed invertible, so the solution exists and is unique.

T2.5.7 Computing Levi-Civita connections

The "educated guess" method is almost always fastest [Dray §17.7]. Here are the four cases we need; the first reproduces what we already know, the rest are new.

Worked example T2.22: Polar coordinates, from the structure equations

Coframe $\sigma^1 = \mathrm{d}r$, $\sigma^2 = r\,\mathrm{d}\phi$. The only unknown is $\omega^1{}_2 = -\omega^2{}_1$. The two structure equations are

$$\begin{aligned} 0 &\overset{(1)}{=} \mathrm{d}\sigma^1 + \omega^1{}_2 \wedge \sigma^2 = \omega^1{}_2 \wedge r\,\mathrm{d}\phi, \\ 0 &\overset{(2)}{=} \mathrm{d}\sigma^2 + \omega^2{}_1 \wedge \sigma^1 = \mathrm{d}r \wedge \mathrm{d}\phi - \omega^1{}_2 \wedge \mathrm{d}r. \end{aligned}$$

(1) $\mathrm{d}\sigma^1 = \mathrm{d}\mathrm{d}r = 0$, and $\omega^1{}_1 = 0$.

(2) $\mathrm{d}(r\,\mathrm{d}\phi) = \mathrm{d}r \wedge \mathrm{d}\phi$, and $\omega^2{}_1 = -\omega^1{}_2$, $\omega^2{}_2 = 0$.

The second equation needs $\omega^1{}_2 \wedge \mathrm{d}r = \mathrm{d}r \wedge \mathrm{d}\phi$, so $\omega^1{}_2$ must contain $-\mathrm{d}\phi$. Guess $\omega^1{}_2 = -\mathrm{d}\phi$; then the first equation, $-\mathrm{d}\phi \wedge r\,\mathrm{d}\phi = 0$, also holds. Done: $\omega^1{}_2 = -\mathrm{d}\phi$, in agreement with (T2.64) [Dray (17.32)–(17.39)]. We never used the Cartesian frame.

Worked example T2.23: The sphere of radius a

Coframe $\sigma^1 = a\,\mathrm{d}\theta$, $\sigma^2 = a\sin\theta\,\mathrm{d}\phi$. Then $\mathrm{d}\sigma^1 = 0$ and $\mathrm{d}\sigma^2 = a\cos\theta\,\mathrm{d}\theta \wedge \mathrm{d}\phi$. The structure equations:

$$0 \overset{(1)}{=} \omega^1{}_2 \wedge a\sin\theta\,\mathrm{d}\phi,$$
$$a\cos\theta\,\mathrm{d}\theta \wedge \mathrm{d}\phi \overset{(2)}{=} -\omega^2{}_1 \wedge a\,\mathrm{d}\theta = \omega^1{}_2 \wedge a\,\mathrm{d}\theta.$$

(1) First structure equation for $i = 1$, with $\mathrm{d}\sigma^1 = 0$.
(2) First structure equation for $i = 2$, and antisymmetry.

Try $\omega^1{}_2 = c\,\mathrm{d}\phi$. The first equation holds. The second gives $c\,a\,\mathrm{d}\phi \wedge \mathrm{d}\theta = a\cos\theta\,\mathrm{d}\theta \wedge \mathrm{d}\phi$, so $c = -\cos\theta$:

$$\omega^1{}_2 = -\cos\theta\,\mathrm{d}\phi \qquad \text{(sphere, frame } \hat\theta, \hat\phi\text{)}. \tag{T2.70}$$

Note that the radius a has dropped out. This is the potential $-\cos\theta\,\mathrm{d}\phi$ of the sphere puzzle (T2.31)! We will see why in Section T2.5.10.

Check from outside. Embed the sphere in $\mathbb{R}^3$. In spherical coordinates $\hat\phi = (-\sin\phi, \cos\phi, 0)$, so $\mathrm{d}\hat\phi = -(\cos\phi, \sin\phi, 0)\,\mathrm{d}\phi = -(\sin\theta\,\hat r + \cos\theta\,\hat\theta)\,\mathrm{d}\phi$. Then

$$\omega^1{}_2 = \hat\theta \cdot \mathrm{d}\hat\phi = -\cos\theta\,\mathrm{d}\phi,$$

the same answer. The $\hat r$ part of $\mathrm{d}\hat\phi$ sticks out of the sphere; an ant living on the sphere cannot see it, and it does not enter $\omega^1{}_2$. It will reappear as *extrinsic* curvature in Section T2.6.5.

Worked example T2.24: The torus of revolution

Coframe $\sigma^1 = a\,\mathrm{d}\theta$, $\sigma^2 = (R + a\cos\theta)\,\mathrm{d}\phi$. Then $\mathrm{d}\sigma^1 = 0$ and $\mathrm{d}\sigma^2 = -a\sin\theta\,\mathrm{d}\theta \wedge \mathrm{d}\phi$. The second structure equation reads

$$-a\sin\theta\,\mathrm{d}\theta \wedge \mathrm{d}\phi \overset{(1)}{=} \omega^1{}_2 \wedge a\,\mathrm{d}\theta,$$

(1) As for the sphere: $\mathrm{d}\sigma^2 = -\omega^2{}_1 \wedge \sigma^1 = \omega^1{}_2 \wedge \sigma^1$.

which is solved by

$$\omega^1{}_2 = \sin\theta\,\mathrm{d}\phi \tag{T2.71}$$

(check: $\sin\theta\,\mathrm{d}\phi \wedge a\,\mathrm{d}\theta = -a\sin\theta\,\mathrm{d}\theta \wedge \mathrm{d}\phi$), and the first equation $\omega^1{}_2 \wedge \sigma^2 \propto \mathrm{d}\phi \wedge \mathrm{d}\phi = 0$ holds too [Dray (18.84)–(18.86)]. Neither R nor a appears.

Worked example T2.25: Flat $\mathbb{R}^3$ in spherical coordinates (Dray, Problem 17.1)

Now there are three connection 1-forms. Differentiate the unit vectors (from their Cartesian expressions):

$$\mathrm{d}\hat r = \mathrm{d}\theta\,\hat\theta + \sin\theta\,\mathrm{d}\phi\,\hat\phi,$$
$$\mathrm{d}\hat\theta = -\mathrm{d}\theta\,\hat r + \cos\theta\,\mathrm{d}\phi\,\hat\phi,$$

$$\mathrm{d}\hat{\phi} = -\sin\theta\,\mathrm{d}\phi\,\hat{r} - \cos\theta\,\mathrm{d}\phi\,\hat{\theta}.$$

Reading off $\omega^i{}_j = e_i \cdot \mathrm{d}e_j$ with $(e_1, e_2, e_3) = (\hat{r}, \hat{\theta}, \hat{\phi})$:

$$\omega = \begin{pmatrix} 0 & -\mathrm{d}\theta & -\sin\theta\,\mathrm{d}\phi \\ \mathrm{d}\theta & 0 & -\cos\theta\,\mathrm{d}\phi \\ \sin\theta\,\mathrm{d}\phi & \cos\theta\,\mathrm{d}\phi & 0 \end{pmatrix}. \tag{T2.72}$$

It is antisymmetric, as promised. Check the torsion equation for $i = 3$:

$$\begin{aligned} \mathrm{d}\sigma^3 + \omega^3{}_1 \wedge \sigma^1 + \omega^3{}_2 \wedge \sigma^2 &\overset{(1)}{=} \big(\sin\theta\,\mathrm{d}r \wedge \mathrm{d}\phi + r\cos\theta\,\mathrm{d}\theta \wedge \mathrm{d}\phi\big) \\ &\quad + \sin\theta\,\mathrm{d}\phi \wedge \mathrm{d}r + \cos\theta\,\mathrm{d}\phi \wedge r\,\mathrm{d}\theta \\ &\overset{(2)}{=} 0. \end{aligned}$$

(1) $\sigma^3 = r\sin\theta\,\mathrm{d}\phi$, so $\mathrm{d}\sigma^3 = \sin\theta\,\mathrm{d}r \wedge \mathrm{d}\phi + r\cos\theta\,\mathrm{d}\theta \wedge \mathrm{d}\phi$; insert $\omega^3{}_1$, $\omega^3{}_2$ from (T2.72).
(2) The terms cancel in pairs after swapping the order of the wedges.

The equations for $i = 1, 2$ check the same way. Notice the $(2, 3)$ entry, $\omega^2{}_3 = -\cos\theta\,\mathrm{d}\phi$: it is exactly the sphere's connection (T2.70). Restricting to $r = \text{const}$ (so $\mathrm{d}r = 0$) keeps the tangential rotation and drops the rest [Dray (18.42)].

T2.5.8 An explicit formula: Christoffel symbols and the Koszul formula

Guessing is fast, but it is reassuring to have a formula. Expand the connection 1-forms in the coframe [Dray (17.40)]:

$$\omega^i{}_j = \Gamma^i{}_{jk}\,\sigma^k. \tag{T2.73}$$

The coefficients $\Gamma^i{}_{jk}$ are the **Christoffel symbols** (in an orthonormal frame; physicists also call ω the *spin connection* and its components *Ricci rotation coefficients*). Metric compatibility says $\Gamma_{ijk} = -\Gamma_{jik}$ [Dray (17.42)]. Next expand the derivatives of the coframe:

$$\mathrm{d}\sigma^i = \tfrac{1}{2}\,C^i{}_{jk}\,\sigma^j \wedge \sigma^k, \qquad C^i{}_{jk} = -C^i{}_{kj}. \tag{T2.74}$$

The C's are computed directly from the coframe; no connection is needed. The torsion equation, $\mathrm{d}\sigma^i = -\Gamma^i{}_{jk}\sigma^k \wedge \sigma^j = \Gamma^i{}_{jk}\sigma^j \wedge \sigma^k$, says that $C^i{}_{jk} = \Gamma^i{}_{jk} - \Gamma^i{}_{kj}$ [Dray (17.47)]. We want to invert this, using the antisymmetry in the first two indices:

$$\begin{aligned} C_{ijk} - C_{kij} + C_{jki} &\overset{(1)}{=} (\Gamma_{ijk} - \Gamma_{ikj}) - (\Gamma_{kij} - \Gamma_{kji}) + (\Gamma_{jki} - \Gamma_{jik}) \\ &\overset{(2)}{=} \Gamma_{ijk} + \Gamma_{kij} - \Gamma_{kij} - \Gamma_{jki} + \Gamma_{jki} - \Gamma_{jik} \\ &\overset{(3)}{=} \Gamma_{ijk} - \Gamma_{jik} \overset{(4)}{=} 2\,\Gamma_{ijk}. \end{aligned} \tag{T2.75}$$

(1) Substitute $C_{ijk} = \Gamma_{ijk} - \Gamma_{ikj}$ three times (indices lowered freely, Euclidean signature).
(2) Use $\Gamma_{ikj} = -\Gamma_{kij}$ and $\Gamma_{kji} = -\Gamma_{jki}$ (antisymmetry in the first pair).
(3) Four terms cancel in pairs.
(4) Antisymmetry once more: $-\Gamma_{jik} = \Gamma_{ijk}$.

So

$$\boxed{\Gamma_{ijk} = \tfrac{1}{2}\big(C_{ijk} - C_{kij} + C_{jki}\big)} \tag{T2.76}$$

which is Dray's (17.48), the **Koszul formula** in an orthonormal frame. It proves existence and uniqueness at once: the right side is determined by the coframe alone.

Commutators. The C's have a second meaning. The **commutator** of two vector fields, $[\vec{v},\vec{w}]f := \vec{v}(\vec{w}f) - \vec{w}(\vec{v}f)$, contains no second derivatives of f (they cancel), so it is again a vector field [Dray (17.60)]. For any 1-form the identity $\mathrm{d}\alpha(\vec{v},\vec{w}) = \vec{v}\big(\alpha(\vec{w})\big) - \vec{w}\big(\alpha(\vec{v})\big) - \alpha([\vec{v},\vec{w}])$ holds [Dray (17.61)] (check it on $\alpha = f\,\mathrm{d}g$ in coordinates). Apply it to $\alpha = \sigma^i$ and $\vec{v},\vec{w} = e_j, e_k$: the first two terms are derivatives of the constants $\delta^i{}_k, \delta^i{}_j$, so

$$C^i{}_{jk} = \mathrm{d}\sigma^i(e_j,e_k) = -\sigma^i\big([e_j,e_k]\big), \qquad \text{i.e.} \qquad [e_j,e_k] = -C^i{}_{jk}\,e_i. \tag{T2.77}$$

The frame vectors fail to commute exactly when the coframe fails to be closed. In 2D the Koszul formula collapses to a one-liner [Dray (17.65)–(17.66)]:

$$\omega_{12} = [e_1,e_2]\cdot \mathrm{d}\vec{r}. \tag{T2.78}$$

Worked example T2.26: The one-liner on the plane and on the sphere

Plane. $e_1 = \partial_r$, $e_2 = \frac{1}{r}\partial_\phi$. Acting on a function f,

$$[e_1,e_2]f \overset{(1)}{=} \partial_r\Big(\frac{1}{r}\partial_\phi f\Big) - \frac{1}{r}\partial_\phi\partial_r f \overset{(2)}{=} -\frac{1}{r^2}\partial_\phi f = -\frac{1}{r}\,e_2 f,$$

(1) Definition of the commutator.
(2) Product rule; the mixed second derivatives cancel.

so $[e_1,e_2] = -\frac{1}{r}e_2$ and $\omega_{12} = -\frac{1}{r}\,e_2\cdot\mathrm{d}\vec{r} = -\frac{1}{r}\sigma^2 = -\mathrm{d}\phi$.
Sphere. $e_1 = \frac{1}{a}\partial_\theta$, $e_2 = \frac{1}{a\sin\theta}\partial_\phi$. The same computation gives $[e_1,e_2] = \frac{1}{a}\partial_\theta\big(\frac{1}{a\sin\theta}\big)\partial_\phi = -\frac{\cos\theta}{a\sin\theta}\,e_2$, hence $\omega_{12} = -\frac{\cos\theta}{a\sin\theta}\,a\sin\theta\,\mathrm{d}\phi = -\cos\theta\,\mathrm{d}\phi$, matching (T2.70).

Dictionary: frame connection ↔ coordinate Christoffel symbols

Relativity textbooks use the coordinate basis ∂_μ and write $\nabla_\mu V^\nu = \partial_\mu V^\nu + \Gamma^\nu{}_{\mu\lambda}V^\lambda$ with $\Gamma^\nu{}_{\mu\lambda} = \frac{1}{2}g^{\nu\rho}(\partial_\mu g_{\rho\lambda} + \partial_\lambda g_{\rho\mu} - \partial_\rho g_{\mu\lambda})$. The orthonormal-frame connection (the *spin connection*) is related to it by changing basis with the vielbein:

$$\omega^i{}_j = \sigma^i{}_\nu\big(\partial_\mu E_j{}^\nu + \Gamma^\nu{}_{\mu\lambda}E_j{}^\lambda\big)\mathrm{d}x^\mu, \qquad e_j = E_j{}^\nu\partial_\nu,\ \sigma^i = \sigma^i{}_\nu \mathrm{d}x^\nu.$$

Check on polar coordinates: $\Gamma^r{}_{\phi\phi} = -r$, $\Gamma^\phi{}_{r\phi} = \Gamma^\phi{}_{\phi r} = \frac{1}{r}$, others zero; $e_2 = \frac{1}{r}\partial_\phi$ so $E_2{}^\phi = \frac{1}{r}$. Then $\omega^1{}_2 = \sigma^1{}_r\,\Gamma^r{}_{\phi\phi}E_2{}^\phi\,\mathrm{d}\phi = 1\cdot(-r)\cdot\frac{1}{r}\,\mathrm{d}\phi = -\mathrm{d}\phi$. Same answer, more work. The frame method needs only $n(n-1)/2$ 1-forms; the coordinate method needs $n^2(n+1)/2$ functions.

Worked example T2.27: The sphere's Christoffel symbols, seen from $\mathbb{R}^3$

Sit the unit sphere in $\mathbb{R}^3$, $\vec{r} = (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta)$, with the coordinate basis $e_\theta = \partial_\theta\vec{r}$, $e_\phi = \partial_\phi\vec{r}$, and watch the basis turn:

$$\partial_\theta e_\phi \overset{(1)}{=} (-\cos\theta\sin\phi,\ \cos\theta\cos\phi,\ 0) \overset{(2)}{=} \cot\theta\,e_\phi,$$
$$\partial_\phi e_\phi \overset{(3)}{=} -\sin\theta(\cos\phi,\sin\phi,0) \overset{(4)}{=} -\sin\theta\cos\theta\,e_\theta - \sin^2\theta\,\hat{n}.$$

(1) Differentiate $e_\phi = (-\sin\theta\sin\phi,\ \sin\theta\cos\phi,\ 0)$ in θ.
(2) It is e_ϕ with $\sin\theta$ replaced by $\cos\theta$.
(3) Differentiate e_ϕ in ϕ.
(4) Split $(\cos\phi,\sin\phi,0) = \cos\theta\,e_\theta + \sin\theta\,\hat{n}$ into its part along the sphere and its part along the normal $\hat{n}$.

Keep only the part along the sphere, since $\nabla_a e_b = \Gamma^c{}_{ab}e_c$ is the tangent part of $\partial_a e_b$. This

gives $\Gamma^{\phi}{}_{\theta\phi} = \cot\theta$ and $\Gamma^{\theta}{}_{\phi\phi} = -\sin\theta\cos\theta$: the same numbers the metric formula in the box above gives for $ds^2 = d\theta^2 + \sin^2\theta\, d\phi^2$, and the same connection as $\omega^1{}_2 = -\cos\theta\, d\phi$ of Example T2.23, with no $\mathbb{R}^3$ needed. That is the point of the formalism (Figure T1.2).

T2.5.9 Covariant derivatives

The vector-valued 1-form $d\vec{v}$ answers "how does $\vec{v}$ change?" for every direction at once. Feeding it one direction gives the **covariant derivative**:

Working definition T2.10 (covariant derivative)

For vector fields $\vec{X}$ and $\vec{v} = v^i e_i$,

$$\nabla_{\vec{X}}\vec{v} := d\vec{v}(\vec{X}) = \left(\vec{X}(v^i) + \omega^i{}_j(\vec{X})\, v^j\right) e_i. \tag{T2.79}$$

Recipe: differentiate the components along $\vec{X}$ [$\vec{X}(v^i)$ is the ordinary directional derivative of the function v^i], then add the rotation of the frame, $\omega(\vec{X})$, acting on the column of components. In components $\nabla_k v^i = e_k(v^i) + \Gamma^i{}_{jk} v^j$ [here $\Gamma^i{}_{jk} = \omega^i{}_j(e_k)$, the connection fed the k-th frame vector].

In the physicist's language of gauge theory, $\nabla = d + \omega$ is the *covariant derivative* with gauge potential ω. Nash writes the Dirac operator as $\not{D} = \gamma^a e_a^{\mu}(\partial_\mu + \omega_\mu)$ [Nash p. 114, (4.100)], where ω_μ is this same connection, acting on spinors through the generators of Spin(n) (Chapter T4).

Worked example T2.28: Covariant derivatives on the sphere

On the unit sphere ($a = 1$), take $\vec{v} = \hat{\phi}$ and differentiate it along the latitude circle, $\vec{X} = \hat{\phi}$, and along the meridian, $\vec{X} = \hat{\theta}$. With $e_1 = \hat{\theta}$, $e_2 = \hat{\phi}$, $v^1 = 0$, $v^2 = 1$:

$$\nabla_{\hat{\phi}}\hat{\phi} \overset{(1)}{=} \omega^1{}_2(\hat{\phi})\, e_1 \overset{(2)}{=} -\cos\theta\, d\phi(\hat{\phi})\, \hat{\theta} \overset{(3)}{=} -\cot\theta\, \hat{\theta},$$

$$\nabla_{\hat{\theta}}\hat{\phi} \overset{(4)}{=} \omega^1{}_2(\hat{\theta})\, e_1 = 0.$$

(1) Constant components, so only the ω term of (T2.79) survives; $\omega^2{}_2 = 0$.
(2) (T2.70).
(3) $d\phi(\hat{\phi}) = 1/\sin\theta$, since $\hat{\phi} = \frac{1}{\sin\theta}\partial_\phi$.
(4) $d\phi(\hat{\theta}) = 0$.

So $\hat{\phi}$ does not change as we walk along a meridian, but it turns (towards the pole: $-\hat{\theta}$ points north) as we walk along a latitude circle, at a rate $\cot\theta$. At the equator, $\cot\theta = 0$: walking along the equator, $\hat{\phi}$ does not turn at all. The equator is straight as seen from inside the sphere; it is a geodesic (Section T2.7).

T2.5.10 Rotating the frame is a gauge transformation

Now we come to the observation that connects this chapter with gauge theory. Let $\mathbf{e}' = \mathbf{e}\, g$ be a rotated frame, where $g(x) \in \mathrm{SO}(n)$ is a point-dependent rotation matrix. What is the new connection?

$$d\mathbf{e}' \overset{(1)}{=} d(\mathbf{e}\, g) \overset{(2)}{=} d\mathbf{e}\, g + \mathbf{e}\, dg \overset{(3)}{=} \mathbf{e}\, \omega\, g + \mathbf{e}\, dg$$

$$\overset{(4)}{=} \mathbf{e}'\left(g^{-1}\omega\, g + g^{-1}\mathrm{d}g\right). \tag{T2.80}$$

(1) Definition of the new frame.
(2) Product rule ($\mathbf{e}$ and g are both 0-forms).
(3) $\mathrm{d}\mathbf{e} = \mathbf{e}\,\omega$.
(4) $\mathbf{e} = \mathbf{e}'g^{-1}$.

Hence, together with the coframe (which must satisfy $\mathrm{d}\vec{r} = \mathbf{e}\boldsymbol{\sigma} = \mathbf{e}'\boldsymbol{\sigma}'$),

$$\boxed{\omega' = g^{-1}\omega\, g + g^{-1}\mathrm{d}g, \qquad \boldsymbol{\sigma}' = g^{-1}\boldsymbol{\sigma}} \tag{T2.81}$$

This is *exactly* the gauge transformation law $A \mapsto g^{-1}Ag + g^{-1}\mathrm{d}g$ of a non-abelian gauge potential, in the convention used throughout these notes. The structure group is $\mathrm{SO}(n)$. The "matter fields" are the components v^i of vectors, which transform as $v' = g^{-1}v$.

Two dimensions: an abelian gauge field. Rotate by angle $\lambda(x)$: $g = R(\lambda) = \left(\begin{smallmatrix}\cos\lambda & -\sin\lambda \\ \sin\lambda & \cos\lambda\end{smallmatrix}\right)$, so $e_1' = \cos\lambda\, e_1 + \sin\lambda\, e_2$ and $e_2' = -\sin\lambda\, e_1 + \cos\lambda\, e_2$. Rotations in the plane commute, so $g^{-1}\omega g = \omega$, and

$$g^{-1}\mathrm{d}g \overset{(1)}{=} \begin{pmatrix}\cos\lambda & \sin\lambda \\ -\sin\lambda & \cos\lambda\end{pmatrix}\begin{pmatrix}-\sin\lambda & -\cos\lambda \\ \cos\lambda & -\sin\lambda\end{pmatrix}\mathrm{d}\lambda \overset{(2)}{=} \begin{pmatrix}0 & -1 \\ 1 & 0\end{pmatrix}\mathrm{d}\lambda.$$

(1) $g^{-1} = R(-\lambda)$, and $\mathrm{d}g = R'(\lambda)\,\mathrm{d}\lambda$.
(2) Multiply out; $\cos^2\lambda + \sin^2\lambda = 1$.

Reading off the $(1,2)$ entry:

$$\omega'^1{}_2 = \omega^1{}_2 - \mathrm{d}\lambda. \tag{T2.82}$$

This is the gauge transformation $A \mapsto A - \mathrm{d}\lambda$ of electromagnetism. Two immediate consequences:

- *Polar coordinates are a gauge transformation of Cartesian ones.* The Cartesian frame has $\omega = 0$. The polar frame is the Cartesian frame rotated by $\lambda = \phi$. So $\omega^1{}_2 = 0 - \mathrm{d}\phi$, recovering (T2.64) with no computation at all. The flat plane's connection is *pure gauge*.
- $\mathrm{d}\omega^1{}_2$ *is gauge invariant* in 2D, since $\mathrm{d}\mathrm{d}\lambda = 0$. It is the analogue of the field strength $F = \mathrm{d}A$. It is zero for the plane (pure gauge) and $\mathrm{d}(-\cos\theta\,\mathrm{d}\phi) = \sin\theta\,\mathrm{d}\theta \wedge \mathrm{d}\phi \neq 0$ for the sphere. So no rotation of the frame can make the sphere's connection vanish, even locally. That gauge-invariant obstruction is the curvature of the next section.

Worked example T2.29: The sphere's frame, the monopole, and the hairy ball

The frame $(\hat\theta, \hat\phi)$ is singular at both poles: at the north pole $\hat\theta$ points "away from the pole" in every direction at once (Figure T2.17, left). Let us fix the north pole by a gauge transformation. Near $\theta = 0$ the sphere looks like a plane with polar coordinates (θ, ϕ), and there $\hat\theta \approx \cos\phi\,\hat x + \sin\phi\,\hat y$, $\hat\phi \approx -\sin\phi\,\hat x + \cos\phi\,\hat y$. Rotating by $\lambda = -\phi$ undoes the winding:

$$e_1^N \overset{(1)}{=} \cos\phi\,\hat\theta - \sin\phi\,\hat\phi \overset{(2)}{=} \cos\phi(\cos\phi\,\hat x + \sin\phi\,\hat y) - \sin\phi(-\sin\phi\,\hat x + \cos\phi\,\hat y) \overset{(3)}{=} \hat x,$$

(1) $e_1' = \cos\lambda\, e_1 + \sin\lambda\, e_2$ with $\lambda = -\phi$.
(2) The frame near the north pole.
(3) $\cos^2\phi + \sin^2\phi = 1$ and the $\hat y$ terms cancel.

so the new frame is smooth at the north pole. Its connection is, by (T2.82),

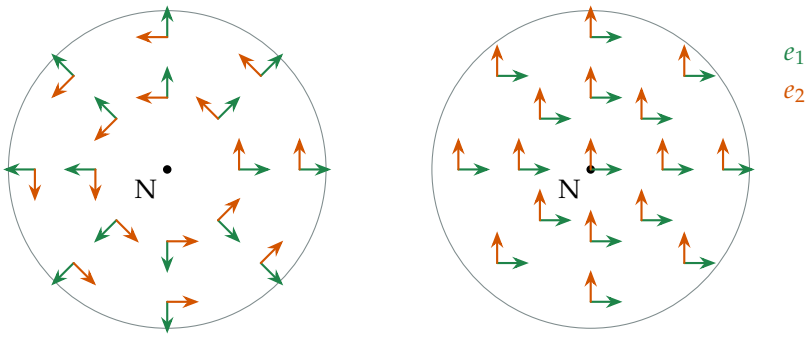
$$\omega_N = \omega^1{}_2 - \mathrm{d}(-\phi) = (1 - \cos\theta)\,\mathrm{d}\phi = A_N,$$

the north-patch monopole potential of (T2.32). Likewise rotating by $\lambda = +\phi$ gives a frame smooth at the south pole ($e_1^S \to -\hat{x}$ there) with $\omega_S = -(1 + \cos\theta)\,\mathrm{d}\phi = A_S$. On the overlap the two good frames differ by a rotation through the angle

$$\lambda_{NS} = 2\phi, \qquad \omega_N - \omega_S = 2\,\mathrm{d}\phi,$$

which winds *twice* as we go once around the equator. So:

- The Levi-Civita connection of the round sphere, written in smooth frames on two patches, *is* a magnetic monopole potential. Its flux is $\int_{S^2} \mathrm{d}\omega = 4\pi$ (Section T2.3.6).
- The transition function winds twice. In the language of Chapter T3, the tangent bundle of S^2, viewed as a complex line bundle, has first Chern number 2. The "2" is the Euler characteristic $\chi(S^2)$; this is the Gauss–Bonnet theorem in disguise (Section T2.8).
- No *single* smooth frame exists on all of S^2: if it did, its connection would be a global potential for the non-exact form $\sin\theta\,\mathrm{d}\theta \wedge \mathrm{d}\phi$. In particular there is no nowhere-vanishing smooth vector field on S^2 (take e_1 to be its normalisation). This is the *hairy ball theorem*, obtained here from the Stokes detector.



$(\hat{\theta}, \hat{\phi})$: winds once, singular at N

(e_1^N, e_2^N), rotated by $\lambda = -\phi$: smooth at N

Figure T2.17: The neighbourhood of the north pole seen from above. Left: the coordinate frame $(\hat{\theta}, \hat{\phi})$ rotates once as we go around the pole, so it has no limit at the pole; its connection $-\cos\theta\,\mathrm{d}\phi$ has a $\mathrm{d}\phi$ singularity there. Right: after the gauge rotation $\lambda = -\phi$ the frame is (to leading order) constant and smooth; its connection is $A_N = (1 - \cos\theta)\,\mathrm{d}\phi$. The analogous frame at the south pole differs from this one by a rotation through 2ϕ on the equator.

Dictionary: frames and connections ↔ gauge theory

Riemannian geometry	Gauge theory
orthonormal frame $\mathbf{e}$	choice of gauge (local trivialisation)
frame rotation $g(x) \in \mathrm{SO}(n)$	gauge transformation
connection ω, $\mathfrak{so}(n)$-valued	gauge potential A, $\mathfrak{g}$-valued
$\omega' = g^{-1}\omega g + g^{-1}\mathrm{d}g$	$A' = g^{-1}Ag + g^{-1}\mathrm{d}g$
$\nabla = \mathrm{d} + \omega$ on components	covariant derivative $D = \mathrm{d} + A$
2D: $\omega^1{}_2 \mapsto \omega^1{}_2 - \mathrm{d}\lambda$	U(1): $A_{\text{phys}} \mapsto A_{\text{phys}} - \mathrm{d}\lambda$
curvature $\Omega = \mathrm{d}\omega + \omega \wedge \omega$ (Section T2.6)	field strength $F = \mathrm{d}A + A \wedge A$

Convention check. In 2D identify the tangent plane with $\mathbb{C}$ by letting i act as rotation by $+90°$ ($i\, e_1 = e_2$). Then e_1 alone is a complex frame, $\mathrm{d}e_1 = e_2\, \omega^2{}_1 = e_1\, (i\, \omega^2{}_1)$, so the anti-Hermitian (mathematicians') U(1) potential is $A = i\, \omega^2{}_1 = -i\, \omega^1{}_2$, and the physicist's Hermitian potential ($A = -iA_{\text{phys}}$) is $A_{\text{phys}} = \omega^1{}_2$. A rotation by λ is $e_1' = e^{i\lambda} e_1$, i.e. $g = e^{i\lambda}$, and $A' = A + g^{-1}\mathrm{d}g = A + i\,\mathrm{d}\lambda$ agrees with (T2.82).

Reading Nash (pp. 86, 114; used from Chapter 3 on)

Nash never computes a Levi-Civita connection. He takes "the usual orthonormal frame (vielbein) of tangent vectors $e_a^\mu(x)$" and "the Riemannian spin connection ω" as known [Nash p. 114, (4.100)], and uses the fact that the structure group of an oriented Riemannian manifold reduces to $\mathrm{SO}(n)$ when he introduces the Euler class [Nash p. 86, (3.120)]. This section supplies all of that: the vielbein is the matrix E of (T2.58); the spin connection is ω, computed from $\mathrm{d}\boldsymbol{\sigma} + \omega \wedge \boldsymbol{\sigma} = 0$; the reduction to $\mathrm{SO}(n)$ is metric compatibility (T2.66); and the identification with gauge fields is (T2.81).

Recap

- A connection records how a frame turns: $\mathrm{d}e_j = e_i\, \omega^i{}_j$. Then $\mathrm{d}\vec{v} = (\mathrm{d}v^i + \omega^i{}_j v^j) e_i$ and $\nabla_{\vec{X}} \vec{v} = \mathrm{d}\vec{v}(\vec{X})$.
- Orthonormal coframe (vielbein): $\mathrm{d}\vec{r} = \sigma^i e_i$, $\mathrm{d}s^2 = \sum (\sigma^i)^2$, $g = E^{\mathsf{T}} E$.
- Levi-Civita: $\omega_{ij} = -\omega_{ji}$ (metric compatible, $\mathfrak{so}(n)$-valued) and $\mathrm{d}\sigma^i + \omega^i{}_j \wedge \sigma^j = 0$ (torsion free, Cartan's first structure equation). Unique, by counting or by the Koszul formula.
- Answers: plane $\omega^1{}_2 = -\mathrm{d}\phi$; sphere $-\cos\theta\, \mathrm{d}\phi$; torus $\sin\theta\, \mathrm{d}\phi$; spherical coordinates (T2.72).
- Rotating the frame, $\mathbf{e}' = \mathbf{e}g$, gives $\omega' = g^{-1}\omega g + g^{-1}\mathrm{d}g$: a gauge transformation. In 2D, $\omega^1{}_2 \mapsto \omega^1{}_2 - \mathrm{d}\lambda$. The sphere's connection in smooth frames is the monopole potential, with transition winding $2 = \chi(S^2)$.

T2.6 Curvature

The previous section ended with a question. On the plane, the connection of the polar frame is $\omega^1{}_2 = -\mathrm{d}\phi$, and a rotation of the frame (back to Cartesian) makes it vanish. On the sphere, the connection is $-\cos\theta\, \mathrm{d}\phi$, and no rotation can make it vanish, because $\mathrm{d}\omega^1{}_2$ is unchanged by rotations and is not zero. So:

What is the gauge-invariant quantity that decides whether a connection can be "gauged away", and what does it measure geometrically?

The answer is the **curvature 2-form** $\Omega = \mathrm{d}\omega + \omega \wedge \omega$. We will derive it from the failure of $\mathrm{d}^2 = 0$ on vectors, see that it measures the rotation of a vector carried around a small loop, compute it for all our examples, relate it to the bending of a surface in $\mathbb{R}^3$ (Gauss's Theorema Egregium), and extract the Riemann, Ricci and scalar curvatures that physicists use. The formula $\Omega = \mathrm{d}\omega + \omega \wedge \omega$ is word for word the Yang–Mills field strength $F = \mathrm{d}A + A \wedge A$; that is the point of this section for the rest of the book. We follow [Dray §§18.1–18.7].

T2.6.1 The second structure equation

In Section T2.5.6 we demanded $\mathrm{d}(\mathrm{d}\vec{r}) = 0$. What about $\mathrm{d}(\mathrm{d}e_j)$? Use the matrix notation $\mathrm{d}\mathbf{e} = \mathbf{e}\,\omega$ of (T2.62):

$$\mathrm{d}(\mathrm{d}\mathbf{e}) \overset{(1)}{=} \mathrm{d}(\mathbf{e}\,\omega) \overset{(2)}{=} \mathrm{d}\mathbf{e} \wedge \omega + \mathbf{e}\,\mathrm{d}\omega \overset{(3)}{=} \mathbf{e}\,\omega \wedge \omega + \mathbf{e}\,\mathrm{d}\omega \overset{(4)}{=} \mathbf{e}\,\big(\mathrm{d}\omega + \omega \wedge \omega\big). \tag{T2.83}$$

(1) Definition of the connection.
(2) Product rule (T2.60) with the vector-valued 0-form on the left: no sign.
(3) Insert $\mathrm{d}\mathbf{e} = \mathbf{e}\,\omega$ again.
(4) Factor out the frame [Dray (18.29)].

Working definition T2.11 (curvature 2-form)

The **curvature** of a connection ω is the matrix of 2-forms

$$\boxed{\Omega = \mathrm{d}\omega + \omega \wedge \omega, \qquad \Omega^i{}_j = \mathrm{d}\omega^i{}_j + \omega^i{}_k \wedge \omega^k{}_j} \tag{T2.84}$$

(**Cartan's second structure equation**, [Dray (18.30)]). Here $(\omega \wedge \omega)^i{}_j = \omega^i{}_k \wedge \omega^k{}_j$ is matrix multiplication with the entries multiplied by $\wedge$. *How to compute:* find ω from the first structure equation [$\mathrm{d}\sigma^i + \omega^i{}_j \wedge \sigma^j = 0$ with ω antisymmetric], then differentiate and add the matrix product. *How to read it:* $\Omega^i{}_j(\vec{u}, \vec{v})$ is the (i, j) entry of the infinitesimal rotation suffered by a vector carried [moved so that its covariant derivative along the path is zero] around the small parallelogram with edges $\vec{u}, \vec{v}$ (see below).

Watch out: $\omega \wedge \omega$ is not zero

For a single 1-form, $\alpha \wedge \alpha = 0$. For a *matrix* of 1-forms, $(\omega \wedge \omega)^i{}_j = \sum_k \omega^i{}_k \wedge \omega^k{}_j$ is a sum of wedges of *different* 1-forms, and it is not zero in general. It vanishes identically when the matrices commute, e.g. in 2D, where ω has only the one independent entry $\omega^1{}_2$. That is why 2D (like electromagnetism) is "abelian".

Curvature is d^2 on vectors. Let $\vec{v} = \mathbf{e}\,v$ with $v = (v^1, \dots, v^n)^{\mathsf{T}}$ a column of component functions. We showed $\mathrm{d}\vec{v} = \mathbf{e}\,(\mathrm{d}v + \omega v)$ in (T2.63). Differentiate once more:

$$\begin{aligned} \mathrm{d}(\mathrm{d}\vec{v}) &\overset{(1)}{=} \mathrm{d}\mathbf{e} \wedge (\mathrm{d}v + \omega v) + \mathbf{e}\,\mathrm{d}(\mathrm{d}v + \omega v) \\ &\overset{(2)}{=} \mathbf{e}\,\big[\omega \wedge \mathrm{d}v + \omega \wedge \omega\, v + \mathrm{d}\omega\, v - \omega \wedge \mathrm{d}v\big] \\ &\overset{(3)}{=} \mathbf{e}\,\Omega\, v. \end{aligned} \tag{T2.85}$$

(1) Product rule on $\mathbf{e}\,(\dots)$.

(2) $\mathrm{d}\mathbf{e} = \mathbf{e}\,\omega$; $\mathrm{d}\mathrm{d}v = 0$; and $\mathrm{d}(\omega v) = \mathrm{d}\omega\, v - \omega \wedge \mathrm{d}v$ (the 1-form ω is passed by d).
(3) The $\omega \wedge \mathrm{d}v$ terms cancel; what is left is (T2.84).

So on vector fields d^2 is not zero; it is *multiplication by the curvature*. There are no derivatives of $\vec{v}$ left: curvature is a tensor, an algebraic operation at each point. In coordinates this is the familiar $[\nabla_\mu, \nabla_\nu]v^i = R^i{}_{j\mu\nu}v^j$.

Physical picture: curvature = rotation around a small loop

Carry a vector around a tiny parallelogram with edges $\epsilon\vec{u}$ and $\epsilon\vec{v}$, keeping it "as constant as possible" (parallel transport, Section T2.6.8). It comes back rotated by the matrix $1 - \epsilon^2\,\Omega(\vec{u},\vec{v}) + O(\epsilon^3)$. This is the geometric content of $\mathrm{d}^2 \neq 0$: the change around a closed loop is Stokes' theorem applied to $\mathrm{d}\vec{v}$, and the integrand of the interior is $\mathrm{d}^2\vec{v} = \Omega\,\vec{v}$. In gauge theory the same statement is that the Wilson loop of a small plaquette is $1 - \epsilon^2 F_{\mu\nu} + \ldots$, which is how lattice gauge theory defines F.

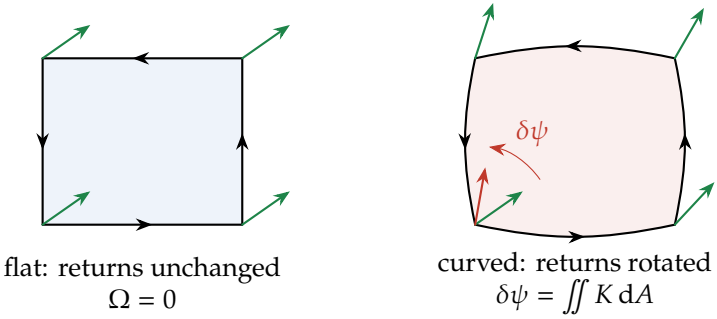


Figure T2.18: Carry a vector around a small closed loop, keeping it parallel to itself at each step. In the flat plane it comes back unchanged. On a curved surface it comes back rotated by an angle $\delta\psi$ equal to the integral of the curvature over the enclosed area. The red arrow at the start is the returned vector. Curvature is this rotation per unit area.

T2.6.2 Symmetries and counting

In an orthonormal frame $\omega_{ij} = -\omega_{ji}$. Then Ω is antisymmetric too [Dray (18.31)–(18.35)]:

$$\Omega_{ji} \overset{(1)}{=} \mathrm{d}\omega_{ji} + \omega_{jk} \wedge \omega_{ki} \overset{(2)}{=} -\mathrm{d}\omega_{ij} + \omega_{kj} \wedge \omega_{ik} \overset{(3)}{=} -\mathrm{d}\omega_{ij} - \omega_{ik} \wedge \omega_{kj} \overset{(4)}{=} -\Omega_{ij}. \tag{T2.86}$$

(1) Definition, with indices lowered (Euclidean signature, orthonormal frame).
(2) $\omega_{ji} = -\omega_{ij}$, and $\omega_{jk} \wedge \omega_{ki} = (-\omega_{kj}) \wedge (-\omega_{ik})$.
(3) Swap the order of the two 1-forms: one minus sign.
(4) Definition of Ω_{ij}.

So Ω is $\mathfrak{so}(n)$-valued, like ω: there are $\frac{1}{2}n(n-1)$ independent curvature 2-forms, each with $\frac{1}{2}n(n-1)$ components. (The first Bianchi identity below cuts the total further, to $n^2(n^2-1)/12$ independent numbers: 1 in 2D, 6 in 3D, 20 in 4D.)

Two dimensions. There is one curvature 2-form,

$$\Omega^1{}_2 \overset{(1)}{=} \mathrm{d}\omega^1{}_2 + \omega^1{}_1 \wedge \omega^1{}_2 + \omega^1{}_2 \wedge \omega^2{}_2 \overset{(2)}{=} \mathrm{d}\omega^1{}_2. \tag{T2.87}$$

(1) Expand the sum over $k = 1, 2$.

(2) The diagonal entries $\omega^1{}_1, \omega^2{}_2$ vanish by antisymmetry.

A 2-form on a surface is a function times the area form $\sigma^1 \wedge \sigma^2$, and that function is the curvature:

Working definition T2.12 (Gaussian curvature)

On a surface with orthonormal coframe σ^1, σ^2,

$$\boxed{\Omega^1{}_2 = \mathrm{d}\omega^1{}_2 = K\,\sigma^1 \wedge \sigma^2} \tag{T2.88}$$

defines the **Gaussian curvature** K [a function on the surface, e.g. $K = 1/a^2$ everywhere on a sphere of radius a] [Dray (18.45)].
Recipe: coframe $\to \omega^1{}_2 \to \mathrm{d}\omega^1{}_2 \to$ divide by the area element $\mathrm{d}A = \sigma^1 \wedge \sigma^2$. Swapping $e_1 \leftrightarrow e_2$ changes the signs of both $\Omega^1{}_2$ and $\sigma^1 \wedge \sigma^2$, so K does not depend on the orientation.

T2.6.3 Gauge covariance, and "flat means pure gauge"

Under a frame rotation $\mathbf{e}' = \mathbf{e}\,g$ we found $\omega' = g^{-1}\omega g + g^{-1}\mathrm{d}g$. The curvature is simpler:

$$\Omega' = g^{-1}\,\Omega\,g. \tag{T2.89}$$

The quickest proof uses (T2.83):

$$\mathbf{e}'\,\Omega' \overset{(1)}{=} \mathrm{d}(\mathrm{d}\mathbf{e}') \overset{(2)}{=} \mathrm{d}\big(\mathrm{d}\mathbf{e}\,g + \mathbf{e}\,\mathrm{d}g\big) \overset{(3)}{=} \mathrm{dd}\mathbf{e}\,g - \mathrm{d}\mathbf{e}\wedge\mathrm{d}g + \mathrm{d}\mathbf{e}\wedge\mathrm{d}g + \mathbf{e}\,\mathrm{dd}g \overset{(4)}{=} \mathbf{e}\,\Omega\,g \overset{(5)}{=} \mathbf{e}'\,g^{-1}\Omega\,g.$$

(1) (T2.83) for the primed frame.
(2) $\mathbf{e}' = \mathbf{e}g$ and the product rule.
(3) Product rule on each term; $\mathrm{d}\mathbf{e}$ is a 1-form so passing it costs a sign.
(4) The middle terms cancel, $\mathrm{dd}g = 0$, and $\mathrm{dd}\mathbf{e} = \mathbf{e}\,\Omega$.
(5) $\mathbf{e} = \mathbf{e}'g^{-1}$.

In 2D the rotations commute, so $\Omega^1{}_2$, and hence K, is *gauge invariant*: it does not depend on the choice of orthonormal frame. In higher dimensions Ω is gauge *covariant*; its invariants are traces of polynomials, such as $\mathrm{tr}\,\Omega\wedge\Omega$. This is the starting point of Chern–Weil theory (Chapter T3).

Key idea: flat ⇔ pure gauge

If $\Omega = 0$ on a simply connected region, there is a frame rotation g that makes $\omega' = 0$: a "Cartesian" frame exists, and the metric is flat Euclidean space in disguise. Conversely, if some frame has $\omega' = 0$ then $\Omega' = 0$ and so $\Omega = g\Omega' g^{-1} = 0$ in every frame.

Why the first direction holds: we want g with $g^{-1}\omega g + g^{-1}\mathrm{d}g = 0$, i.e. $\mathrm{d}g = -\omega g$. This is a first-order system of equations for the matrix g. Applying d to it, the integrability condition $0 = \mathrm{dd}g = -\mathrm{d}\omega\,g + \omega\wedge\mathrm{d}g = -(\mathrm{d}\omega + \omega\wedge\omega)g = -\Omega\,g$ is exactly $\Omega = 0$. In 2D it is just the Poincaré lemma: if $\mathrm{d}\omega^1{}_2 = 0$ on a star-shaped region, then $\omega^1{}_2 = \mathrm{d}\lambda$ for a function λ (Section T2.3.3), and rotating by λ gives $\omega'^1{}_2 = \omega^1{}_2 - \mathrm{d}\lambda = 0$ by (T2.82).

T2.6.4 Examples

Worked example T2.30: Plane, sphere, cylinder

Plane (polar). $\omega^1{}_2 = -\mathrm{d}\phi$, so $\Omega^1{}_2 = -\mathrm{d}\mathrm{d}\phi = 0$. Flat, as it must be.
Sphere of radius a. $\omega^1{}_2 = -\cos\theta\,\mathrm{d}\phi$ from (T2.70), and

$$\Omega^1{}_2 \overset{(1)}{=} \mathrm{d}(-\cos\theta\,\mathrm{d}\phi) \overset{(2)}{=} \sin\theta\,\mathrm{d}\theta\wedge\mathrm{d}\phi \overset{(3)}{=} \frac{1}{a^2}(a\,\mathrm{d}\theta)\wedge(a\sin\theta\,\mathrm{d}\phi) \overset{(4)}{=} \frac{1}{a^2}\,\sigma^1\wedge\sigma^2, \tag{T2.90}$$

(1) (T2.87).
(2) $\mathrm{d}(\cos\theta) = -\sin\theta\,\mathrm{d}\theta$.
(3) Multiply and divide by a^2.
(4) The coframe of the sphere.

so $K = 1/a^2$: constant, positive, and larger for smaller spheres.
Cylinder of radius a. $\mathrm{d}s^2 = a^2\mathrm{d}\phi^2 + \mathrm{d}z^2$, so $\sigma^1 = a\,\mathrm{d}\phi$, $\sigma^2 = \mathrm{d}z$. Both are closed, so $\omega^1{}_2 = 0$ solves the structure equations and $K = 0$. The cylinder is *intrinsically flat*: it is a sheet of paper rolled up, and an ant on it measures triangles with angle sum π. Its bending is visible only from outside (Section T2.6.5).

Worked example T2.31: The torus of revolution

From (T2.71), $\omega^1{}_2 = \sin\theta\,\mathrm{d}\phi$, so

$$\Omega^1{}_2 \overset{(1)}{=} \cos\theta\,\mathrm{d}\theta\wedge\mathrm{d}\phi \overset{(2)}{=} \frac{\cos\theta}{a(R+a\cos\theta)}\,\sigma^1\wedge\sigma^2, \qquad K = \frac{\cos\theta}{a(R+a\cos\theta)} \tag{T2.91}$$

(1) $\mathrm{d}(\sin\theta) = \cos\theta\,\mathrm{d}\theta$.
(2) $\sigma^1\wedge\sigma^2 = a(R+a\cos\theta)\,\mathrm{d}\theta\wedge\mathrm{d}\phi$ [Dray (18.87)–(18.88)].

So $K > 0$ on the outer half ($|\theta| < \pi/2$, sphere-like), $K < 0$ on the inner half (saddle-like), and $K = 0$ on the top and bottom circles $\theta = \pm\pi/2$ (Figure T2.19). The total curvature is

$$\int_{T^2} K\,\mathrm{d}A \overset{(1)}{=} \int_0^{2\pi}\!\!\int_0^{2\pi} \cos\theta\,\mathrm{d}\theta\,\mathrm{d}\phi \overset{(2)}{=} 2\pi\left[\sin\theta\right]_0^{2\pi} = 0. \tag{T2.92}$$

(1) $K\,\mathrm{d}A = \Omega^1{}_2 = \cos\theta\,\mathrm{d}\theta\wedge\mathrm{d}\phi$; the factors $a(R+a\cos\theta)$ cancel.
(2) Nothing depends on ϕ.

The positive and negative regions cancel exactly, whatever R and a are. Gauss–Bonnet will tell us why: $\chi(T^2) = 0$.

Worked example T2.32: Negative curvature: the hyperbolic plane and the pseudosphere

Poincaré half-plane ($Y > 0$) with $\mathrm{d}s^2 = a^2(\mathrm{d}X^2 + \mathrm{d}Y^2)/Y^2$ (Dray, Problem 18.2). Coframe $\sigma^1 = a\,\mathrm{d}X/Y$, $\sigma^2 = a\,\mathrm{d}Y/Y$. Then $\mathrm{d}\sigma^2 = 0$ and

$$\mathrm{d}\sigma^1 \overset{(1)}{=} -\frac{a}{Y^2}\,\mathrm{d}Y\wedge\mathrm{d}X \overset{(2)}{=} \frac{a}{Y^2}\,\mathrm{d}X\wedge\mathrm{d}Y \overset{(3)}{=} \frac{1}{a}\,\sigma^1\wedge\sigma^2.$$

(1) $\mathrm{d}(1/Y) = -\mathrm{d}Y/Y^2$.
(2) Swap the wedge.
(3) $\sigma^1\wedge\sigma^2 = a^2\,\mathrm{d}X\wedge\mathrm{d}Y/Y^2$.

The first structure equation $\mathrm{d}\sigma^1 = -\omega^1{}_2\wedge\sigma^2$ is solved by $\omega^1{}_2 = -\sigma^1/a = -\mathrm{d}X/Y$, and the

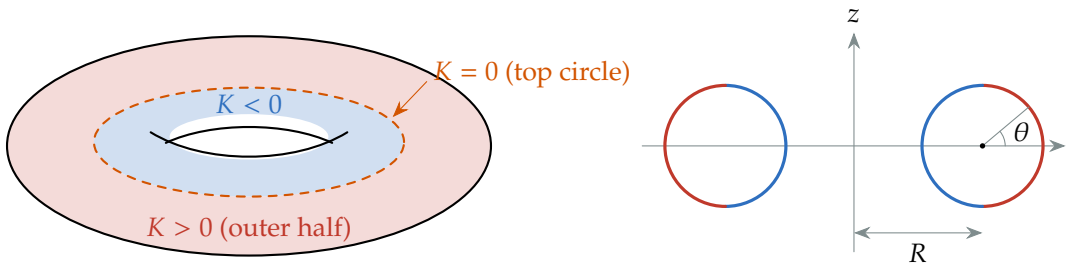


Figure T2.19: The sign of the Gaussian curvature $K = \cos\theta/(a(R + a\cos\theta))$ on a torus of revolution. The outer half (red) curves like a sphere, the inner half (blue, around the hole) like a saddle, and the dashed top circle (like the hidden bottom circle) has $K = 0$. In the cross-section, θ is measured from the outward horizontal. The two contributions to $\int K\,\mathrm{d}A$ cancel exactly, for every shape of torus.

second, $\mathrm{d}\sigma^2 = \omega^1{}_2 \wedge \sigma^1 \propto \sigma^1 \wedge \sigma^1 = 0$, holds. Then

$$\Omega^1{}_2 \overset{(1)}{=} \mathrm{d}\left(-\frac{\mathrm{d}X}{Y}\right) \overset{(2)}{=} \frac{1}{Y^2}\,\mathrm{d}Y \wedge \mathrm{d}X \overset{(3)}{=} -\frac{1}{a^2}\,\sigma^1 \wedge \sigma^2,$$

(1) (T2.87).
(2) $\mathrm{d}(-1/Y) = \mathrm{d}Y/Y^2$.
(3) $\mathrm{d}Y \wedge \mathrm{d}X = -\mathrm{d}X \wedge \mathrm{d}Y = -Y^2\sigma^1 \wedge \sigma^2/a^2$.

so $K = -1/a^2$, constant and negative.
Pseudosphere (Dray, Problem 18.1): the surface of revolution of a tractrix, $\mathrm{d}s^2 = a^2\cot^2\alpha\,\mathrm{d}\alpha^2 + a^2\sin^2\alpha\,\mathrm{d}\phi^2$, $0 < \alpha \le \pi/2$. With $\sigma^1 = a\cot\alpha\,\mathrm{d}\alpha$, $\sigma^2 = a\sin\alpha\,\mathrm{d}\phi$ one finds (try it) $\omega^1{}_2 = -\sin\alpha\,\mathrm{d}\phi$, $\Omega^1{}_2 = -\cos\alpha\,\mathrm{d}\alpha \wedge \mathrm{d}\phi = -a^{-2}\sigma^1 \wedge \sigma^2$: again $K = -1/a^2$. Its area is finite, $\int \sigma^1 \wedge \sigma^2 = \int_0^{2\pi}\int_0^{\pi/2} a^2\cos\alpha\,\mathrm{d}\alpha\,\mathrm{d}\phi = 2\pi a^2$, half that of a sphere of the same radius.

Watch out: Nash's normalisation of the hyperbolic metric

Nash states that the curvature -1 metric on the upper half plane is $\mathrm{d}s^2 = \mathrm{d}z\,\mathrm{d}\bar{z}/(z-\bar{z})^2$ [Nash p. 151, (5.40)]. With $z = X + iY$ we have $(z-\bar{z})^2 = (2iY)^2 = -4Y^2$, so this is $-\frac{1}{4}(\mathrm{d}X^2 + \mathrm{d}Y^2)/Y^2$: negative, and with $a^2 = \frac{1}{4}$ it would give $K = -4$. The curvature -1 metric is $\mathrm{d}s^2 = (\mathrm{d}X^2 + \mathrm{d}Y^2)/Y^2 = -4\,\mathrm{d}z\,\mathrm{d}\bar{z}/(z-\bar{z})^2$. Only the sign of K matters for Nash's argument there.

T2.6.5 Extrinsic curvature and Gauss's Theorema Egregium

The cylinder has $K = 0$, yet it is obviously curved. There are two kinds of curvature, and they must be kept apart: how a surface bends *in* $\mathbb{R}^3$ (extrinsic), and what an inhabitant can measure *without leaving it* (intrinsic). The structure equations make the relation between them transparent [Dray §§18.1–18.5].

Curves. The curvature of a curve is the reciprocal of the radius of its best-fitting circle, $\kappa = 1/r$. With the unit tangent $\hat{T} = \mathrm{d}\vec{r}/\mathrm{d}s$ (s the arc length), a circle has $\mathrm{d}\hat{T}/\mathrm{d}s = \hat{N}/r$ with $\hat{N}$ pointing to the centre; since this involves only first derivatives of $\hat{T}$, the same holds for any curve, $\mathrm{d}\hat{T}/\mathrm{d}s = \kappa\hat{N}$, so [Dray (18.1)–(18.5)]

$$\kappa\,\mathrm{d}s = \mathrm{d}\hat{T}\cdot\hat{N} = -\mathrm{d}\hat{N}\cdot\hat{T}. \tag{T2.93}$$

The curvature of a curve is measured by how its normal turns.

Surfaces: the shape operator. For a surface in $\mathbb{R}^3$, choose an orthonormal frame $\{e_1, e_2, e_3 = \hat{n}\}$ with e_1, e_2 tangent and $\hat{n}$ normal. The normal turns according to the connection of $\mathbb{R}^3$: $-\mathrm{d}\hat{n} \cdot e_i = \hat{n} \cdot \mathrm{d}e_i = \omega^3{}_i$ [Dray (18.6)]. On the surface $\sigma^3 = 0$ (moving within the surface has no normal displacement), so the $i = 3$ torsion equation of $\mathbb{R}^3$ reduces to $\omega^3{}_1 \wedge \sigma^1 + \omega^3{}_2 \wedge \sigma^2 = 0$. Write $\omega^3{}_i = \Gamma^3{}_{ij}\sigma^j$; the equation says $\Gamma^3{}_{12} = \Gamma^3{}_{21}$. The symmetric 2×2 matrix

$$S = -\left(\Gamma^3{}_{ij}\right)_{i,j=1,2} \tag{T2.94}$$

is the **shape operator** (or second fundamental form) [Dray (18.10)]. Being symmetric, it has real eigenvalues κ_1, κ_2, the **principal curvatures**, with orthogonal eigenvectors (the directions of maximal and minimal bending). Its trace $\kappa_1 + \kappa_2$ is the **mean curvature** and its determinant $\kappa_1\kappa_2$ is, as we are about to see, the Gaussian curvature.

Worked example T2.33: Shape operators

Plane: $\hat{n} = \hat{z}$ is constant, $S = 0$. *Cylinder* ($e_1 = \hat{\phi}$, $e_2 = \hat{z}$, $\hat{n} = \hat{r}$): $\mathrm{d}\hat{n} = \mathrm{d}\hat{r} = \mathrm{d}\phi\,\hat{\phi}$, so $\omega^3{}_1 = \hat{r} \cdot \mathrm{d}\hat{\phi} = -\mathrm{d}\phi = -\sigma^1/a$ and $\omega^3{}_2 = 0$:

$$S = \begin{pmatrix} 1/a & 0 \\ 0 & 0 \end{pmatrix}, \qquad \kappa_1\kappa_2 = 0.$$

Sphere ($e_1 = \hat{\theta}$, $e_2 = \hat{\phi}$, $\hat{n} = \hat{r}$): from (T2.72), $\omega^3{}_1 \to \omega^1{}_2$ of $\mathbb{R}^3$ in the ordering $(\hat{r}, \hat{\theta}, \hat{\phi})$, i.e. $\hat{r} \cdot \mathrm{d}\hat{\theta} = -\mathrm{d}\theta = -\sigma^1/a$, and $\hat{r} \cdot \mathrm{d}\hat{\phi} = -\sin\theta\,\mathrm{d}\phi = -\sigma^2/a$:

$$S = \begin{pmatrix} 1/a & 0 \\ 0 & 1/a \end{pmatrix}, \qquad \kappa_1\kappa_2 = \frac{1}{a^2} = K \text{[Dray (18.19)–(18.24)]}.$$

Torus: one principal curvature is $1/a$ (around the tube). The other, along the horizontal circle of radius $R + a\cos\theta$, is $\cos\theta/(R + a\cos\theta)$, not $1/(R + a\cos\theta)$: the circle bends towards the axis, but only the component of that bending along the surface normal counts, and the normal makes angle θ with the horizontal [Dray §18.11]. The product is (T2.91).

The Theorema Egregium. Now use the fact that $\mathbb{R}^3$ is flat: all nine curvature 2-forms of $\mathbb{R}^3$ vanish. Take the $(1,2)$ component and restrict to the surface:

$$\begin{aligned} 0 &\overset{(1)}{=} \Omega^1{}_2\big|_{\mathbb{R}^3} \overset{(2)}{=} \mathrm{d}\omega^1{}_2 + \omega^1{}_3 \wedge \omega^3{}_2 \\ \Longrightarrow \quad \mathrm{d}\omega^1{}_2 &\overset{(3)}{=} \omega^3{}_1 \wedge \omega^3{}_2 \overset{(4)}{=} \left(\Gamma^3{}_{11}\sigma^1 + \Gamma^3{}_{12}\sigma^2\right) \wedge \left(\Gamma^3{}_{21}\sigma^1 + \Gamma^3{}_{22}\sigma^2\right) \\ &\overset{(5)}{=} \left(\Gamma^3{}_{11}\Gamma^3{}_{22} - \Gamma^3{}_{12}\Gamma^3{}_{21}\right)\sigma^1 \wedge \sigma^2 \overset{(6)}{=} \det S\ \sigma^1 \wedge \sigma^2. \end{aligned} \tag{T2.95}$$

(1) Euclidean space is flat.
(2) The second structure equation in 3D; the terms $k = 1, 2$ vanish ($\omega^1{}_1 = 0$, $\omega^2{}_2 = 0$), leaving $k = 3$.
(3) Move to the other side: $-\omega^1{}_3 \wedge \omega^3{}_2 = \omega^3{}_1 \wedge \omega^3{}_2$ by antisymmetry.
(4) Expand in the coframe.
(5) $\sigma^1 \wedge \sigma^1 = \sigma^2 \wedge \sigma^2 = 0$ and $\sigma^2 \wedge \sigma^1 = -\sigma^1 \wedge \sigma^2$.
(6) The bracket is $\det(-S) = \det S$ for a 2×2 matrix.

But the left side, $\mathrm{d}\omega^1{}_2$, is computed from the surface's own coframe: the 3D torsion equations with $\sigma^3 = 0$ are exactly the 2D ones, so $\omega^1{}_2$ of $\mathbb{R}^3$ restricted to the surface is the surface's own Levi-Civita connection [Dray (18.41)–(18.42)]. So

Key idea: Theorema Egregium (Gauss's "remarkable theorem")

$$K = \kappa_1\kappa_2 = \det S.$$

The Gaussian curvature, defined extrinsically as the product of the principal curvatures, depends only on the metric of the surface. An ant that can only measure lengths on the surface can determine K; it cannot determine κ_1 and κ_2 separately.

That is why a sheet of paper can be rolled into a cylinder ($\kappa_1\kappa_2 = 0$ before and after) but not wrapped smoothly around a sphere without crumpling, and why every flat map of the Earth distorts distances. The computation in (T2.95) has a physical reading: the intrinsic curvature $\mathrm{d}\omega^1{}_2$ is exactly cancelled by the product of two extrinsic pieces, so the total curvature of $\mathbb{R}^3$ is zero. For the sphere, $\mathrm{d}\omega^2{}_3 + \omega^2{}_1 \wedge \omega^1{}_3 = \sin\theta\,\mathrm{d}\theta \wedge \mathrm{d}\phi - \sin\theta\,\mathrm{d}\theta \wedge \mathrm{d}\phi = 0$ in the ordering of (T2.72): check it.

T2.6.6 Components: Riemann, Ricci, scalar

Expand the curvature 2-forms in the coframe [Dray (18.47)–(18.51)]:

$$\Omega^i{}_j = \tfrac{1}{2} R^i{}_{jkl}\,\sigma^k \wedge \sigma^l, \qquad R^i{}_{jkl} = -R^i{}_{jlk}. \tag{T2.96}$$

The $R^i{}_{jkl}$ are the components of the **Riemann tensor**. Its traces are the **Ricci tensor** $R_{ij} = R^k{}_{ikj}$ and the **scalar curvature** $R = \delta^{ij}R_{ij}$. In 2D everything is determined by K:

$$\begin{aligned}
&\Omega^1{}_2 \overset{(1)}{=} \tfrac{1}{2}\big(R^1{}_{212}\,\sigma^1 \wedge \sigma^2 + R^1{}_{221}\,\sigma^2 \wedge \sigma^1\big) \overset{(2)}{=} R^1{}_{212}\,\sigma^1 \wedge \sigma^2 \quad\Longrightarrow\quad R^1{}_{212} = K,\\
&R_{11} \overset{(3)}{=} R^2{}_{121} \overset{(4)}{=} K, \qquad R_{22} = R^1{}_{212} = K, \qquad R_{12} = 0,\\
&R \overset{(5)}{=} R_{11} + R_{22} = 2K.
\end{aligned} \tag{T2.97}$$

(1) Only $(k,l) = (1,2), (2,1)$ contribute on a surface.
(2) $R^1{}_{221} = -R^1{}_{212}$ and $\sigma^2 \wedge \sigma^1 = -\sigma^1 \wedge \sigma^2$.
(3) $R_{11} = R^k{}_{1k1}$, and only $k = 2$ survives ($R^1{}_{111} = 0$ by antisymmetry in the last pair).
(4) $R^2{}_{121} = -R^1{}_{221}$ (antisymmetry in the first pair) $= R^1{}_{212} = K$.
(5) Trace.

So in 2D the Ricci tensor is $K\,\delta_{ij}$ and the *scalar curvature is twice the Gaussian curvature*. Keep this factor in mind: in string theory (Chapter 9) the Euler characteristic of a worldsheet is written $\chi = \frac{1}{4\pi}\int R\sqrt{g}\,\mathrm{d}^2x$, which is the same as $\frac{1}{2\pi}\int K\,\mathrm{d}A$.

T2.6.7 The Bianchi identities

Two identities follow from $\mathrm{d}^2 = 0$ on ordinary forms. In matrix notation they are two lines each.

First Bianchi identity (from $\mathrm{d}^2\boldsymbol{\sigma} = 0$):

$$0 \overset{(1)}{=} \mathrm{d}\big(\mathrm{d}\boldsymbol{\sigma} + \omega \wedge \boldsymbol{\sigma}\big) \overset{(2)}{=} \mathrm{d}\omega \wedge \boldsymbol{\sigma} - \omega \wedge \mathrm{d}\boldsymbol{\sigma} \overset{(3)}{=} \mathrm{d}\omega \wedge \boldsymbol{\sigma} + \omega \wedge \omega \wedge \boldsymbol{\sigma} \overset{(4)}{=} \Omega \wedge \boldsymbol{\sigma}. \tag{T2.98}$$

(1) The torsion vanishes identically, so its derivative does too.
(2) $\mathrm{d}\mathrm{d}\boldsymbol{\sigma} = 0$ and the graded product rule on $\omega \wedge \boldsymbol{\sigma}$.
(3) First structure equation: $\mathrm{d}\boldsymbol{\sigma} = -\omega \wedge \boldsymbol{\sigma}$.
(4) Second structure equation [Dray (18.52)].

In components, $\Omega^i{}_j \wedge \sigma^j = 0$ is the cyclic identity $R^i{}_{jkl} + R^i{}_{klj} + R^i{}_{ljk} = 0$ [Dray (18.55)]. Combined with the two antisymmetries it implies that the Ricci tensor is symmetric and that $R_{ijkl} = R_{klij}$ (Dray gives the index gymnastics in (18.58)–(18.60)). On a surface it is empty: $\Omega \wedge \sigma$ is a 3-form.

Second Bianchi identity (from $\mathrm{d}^2\omega = 0$):

$$\mathrm{d}\Omega \overset{(1)}{=} \mathrm{d}\mathrm{d}\omega + \mathrm{d}\omega \wedge \omega - \omega \wedge \mathrm{d}\omega \overset{(2)}{=} (\Omega - \omega \wedge \omega) \wedge \omega - \omega \wedge (\Omega - \omega \wedge \omega) \overset{(3)}{=} \Omega \wedge \omega - \omega \wedge \Omega,$$

(1) Differentiate $\Omega = \mathrm{d}\omega + \omega \wedge \omega$, with the graded product rule.
(2) $\mathrm{d}\mathrm{d}\omega = 0$ and $\mathrm{d}\omega = \Omega - \omega \wedge \omega$.
(3) The cubic terms $\omega \wedge \omega \wedge \omega$ cancel.

that is,

$$D\Omega := \mathrm{d}\Omega + \omega \wedge \Omega - \Omega \wedge \omega = 0 \tag{T2.99}$$

[Dray (18.53)]. In gauge theory the same computation gives $DF = \mathrm{d}F + A \wedge F - F \wedge A = 0$, the non-abelian version of the homogeneous Maxwell equations $\mathrm{d}F = 0$ (no magnetic charges, Faraday's law). In 2D, $\mathrm{d}\Omega^1{}_2$ is a 3-form and vanishes trivially. In 4D the contracted form of (T2.99) is $\nabla^\mu(R_{\mu\nu} - \frac{1}{2}R g_{\mu\nu}) = 0$, the reason Einstein's equations are consistent with energy conservation.

T2.6.8 Parallel transport and holonomy

Working definition T2.13 (parallel transport)

A vector field $\vec{v}$ along a curve $c(t)$ is **parallel transported** if $\nabla_{\dot{c}}\vec{v} = 0$, i.e. $\dot{v}^i + \omega^i{}_j(\dot{c})\, v^j = 0$ [$\dot{v}^i = \mathrm{d}v^i/\mathrm{d}t$, and $\omega(\dot{c})$ is the rate at which the frame rotates along the curve]. This is a linear first-order ODE:
given $\vec{v}$ at the start, it has a unique solution. *Meaning:* the vector is kept "as constant as the geometry allows"; in the plane it is literally constant.

On a surface write a unit vector at angle ψ from e_1, $\vec{v} = \cos\psi\, e_1 + \sin\psi\, e_2$. Then

$$\begin{aligned} \mathrm{d}\vec{v} &\overset{(1)}{=} (-\sin\psi\, e_1 + \cos\psi\, e_2)\,\mathrm{d}\psi + \cos\psi\, e_2\, \omega^2{}_1 + \sin\psi\, e_1\, \omega^1{}_2 \\ &\overset{(2)}{=} (-\sin\psi\, e_1 + \cos\psi\, e_2)\,\big(\mathrm{d}\psi - \omega^1{}_2\big). \end{aligned} \tag{T2.100}$$

(1) Product rule and $\mathrm{d}e_1 = e_2\,\omega^2{}_1$, $\mathrm{d}e_2 = e_1\,\omega^1{}_2$.
(2) $\omega^2{}_1 = -\omega^1{}_2$; collect.

So parallel transport means $\mathrm{d}\psi = \omega^1{}_2$ along the curve: the angle of the vector *relative to the frame* changes by $\omega^1{}_2$ because the frame itself turns. Around a closed curve $C = \partial D$, with a frame that is smooth on D,

$$\Delta\psi \overset{(1)}{=} \oint_C \omega^1{}_2 \overset{(2)}{=} \int_D \mathrm{d}\omega^1{}_2 \overset{(3)}{=} \int_D K\,\mathrm{d}A. \tag{T2.101}$$

(1) Integrate the transport equation $\mathrm{d}\psi = \omega^1{}_2$ around C.
(2) Stokes' theorem; this is where we need the frame (hence $\omega^1{}_2$) smooth on all of D.
(3) (T2.88).

Since the frame returns to itself, $\Delta\psi$ is the actual rotation of the transported vector: the **holonomy** of the loop. *The rotation angle equals the total curvature enclosed.* For an infinitesimal loop this is the intuition box above. It is the precise analogue of the Aharonov–Bohm phase $e^{ie\oint A} = e^{ie\Phi}$: parallel transport is the Wilson line of the connection (Chapter T3).

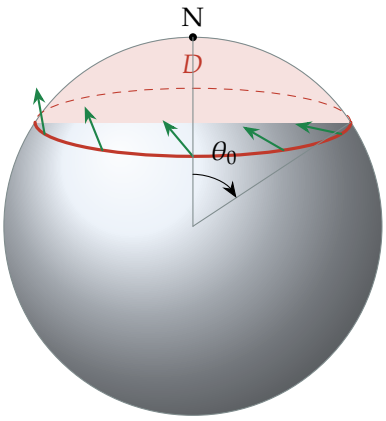
Worked example T2.34: Transport around a line of latitude; the Foucault pendulum

On the unit sphere take the latitude circle $\theta = \theta_0$, bounding the polar cap D ($0 \le \theta \le \theta_0$). The frame $(\hat\theta, \hat\phi)$ is *not* smooth on D (it is singular at the pole), but the rotated frame of Example T2.29 is, with $\omega_N = (1 - \cos\theta)\,\mathrm{d}\phi$:

$$\Delta\psi \overset{(1)}{=} \oint_{\theta=\theta_0} (1-\cos\theta_0)\,\mathrm{d}\phi \overset{(2)}{=} 2\pi(1-\cos\theta_0) \overset{(3)}{=} \int_0^{2\pi}\!\!\int_0^{\theta_0} \sin\theta\,\mathrm{d}\theta\,\mathrm{d}\phi = \mathrm{area}(D).$$

(1) (T2.101) with the smooth frame.
(2) θ is constant on the curve.
(3) The same number from the right side of (T2.101), $K = 1$.

If instead we use the singular frame $(\hat\theta, \hat\phi)$ we get $\oint -\cos\theta_0\,\mathrm{d}\phi = -2\pi\cos\theta_0$, which differs by 2π: the extra turn of the frame around the pole. The rotation of the vector *relative to the local compass directions* is $-2\pi\cos\theta_0$ per circuit. This is the Foucault pendulum: carried around the latitude $90^\circ - \theta_0$ once a (sidereal) day, its swing plane turns by $2\pi\cos\theta_0 = 2\pi\sin(\text{latitude})$ clockwise relative to the ground (in the northern hemisphere). At the equator nothing happens ($\theta_0 = \pi/2$), and at the pole a full turn.



a vector transported around the circle turns relative to $\hat\theta, \hat\phi$; the cap D has area $2\pi(1-\cos\theta_0)$

Figure T2.20: Parallel transport around the latitude circle $\theta = \theta_0$ (red). The vector, kept as constant as possible along the way, returns rotated by the enclosed area $2\pi(1-\cos\theta_0)$ (for the unit sphere). Relative to the local north/east directions it has turned by $-2\pi\cos\theta_0$, the Foucault precession. Along the equator ($\theta_0 = \pi/2$) the enclosed area is a hemisphere, 2π, and the vector comes back unrotated.

Try it: Holonomy on a cone

Cut a wedge of angle β out of a paper disc and glue the edges to make a cone. Away from the tip the cone is flat ($K = 0$). What is the holonomy of a loop around the tip? *Answer:* β (in the direction that makes the total enclosed curvature β). All the curvature is concentrated at the tip, as a delta function of total weight β, the *angle deficit*. A polyhedron is a surface of this kind, with deficits at its vertices; this is Descartes' version of Gauss–Bonnet, $\sum_{\text{vertices}} \beta_v = 4\pi$ for any convex polyhedron (for a cube, $8 \times \frac{\pi}{2} = 4\pi$).

Reading Nash (pp. 129, 151–152, 226–228)

Nash writes the Gauss–Bonnet integrand as "$R/2\pi$ where R is the curvature of X" [Nash p. 129, (4.148)]: his R is our 2-form $\Omega^1{}_2 = K\,\mathrm{d}A$, not the scalar curvature $2K$. In Chapter 5 he uses $R(g)$ for the Gaussian curvature of a metric g and quotes the uniformisation theorem:

every compact Riemann surface carries a metric with $R(g) = +1, 0, -1$ for genus $0, 1, \geq 2$ [Nash pp. 151–152, (5.39)–(5.42)]; our sphere, flat torus and hyperbolic plane are the three model geometries. Parallel transport and holonomy of a general gauge field are his [Nash pp. 226–228]; the latitude circle above is the simplest nontrivial holonomy.

Recap

- $\Omega = \mathrm{d}\omega + \omega \wedge \omega$ (second structure equation) measures $\mathrm{d}^2\vec{v} = \Omega\vec{v}$: the rotation of a vector carried around a small loop. It is $\mathfrak{so}(n)$-valued and transforms covariantly, $\Omega' = g^{-1}\Omega g$.
- Flat ($\Omega = 0$) $\Leftrightarrow$ locally pure gauge $\Leftrightarrow$ a Cartesian frame exists.
- 2D: $\Omega^1{}_2 = \mathrm{d}\omega^1{}_2 = K\,\sigma^1 \wedge \sigma^2$. Sphere $K = 1/a^2$; cylinder 0; torus $\cos\theta/(a(R + a\cos\theta))$ with $\int K\,\mathrm{d}A = 0$; hyperbolic plane and pseudosphere $-1/a^2$.
- Theorema Egregium: $K = \kappa_1\kappa_2 = \det S$ is intrinsic, because $\mathbb{R}^3$ is flat.
- $\Omega^i{}_j = \frac{1}{2}R^i{}_{jkl}\sigma^k \wedge \sigma^l$; in 2D $R_{ij} = K\delta_{ij}$, $R = 2K$.
- Bianchi: $\Omega \wedge \boldsymbol{\sigma} = 0$ and $D\Omega = 0$ (cf. $DF = 0$).
- Holonomy around $\partial D = \int_D K\,\mathrm{d}A$ (Stokes); Foucault pendulum.

T2.7 Geodesics: straight lines on a curved surface

In the plane we know what a straight line is: the curve that does not turn, the path of a free particle, the shortest route between two points. On a sphere there are no straight lines in that naive sense, since every curve on a sphere bends when viewed from $\mathbb{R}^3$. Yet an aeroplane flying from one city to another does follow a preferred route, and a free particle constrained to slide on a frictionless sphere does move along a preferred family of curves. So:

> *What does "straight" mean for a curve on a curved surface, how do we compute it with the connection, and how do we find the straight curves?*

There are three answers, and they agree. A curve is straight if it does not bend to the left or to the right *within the surface* (its geodesic curvature vanishes). A curve is straight if its velocity is parallel transported along itself (no acceleration). A curve is straight if it is a stationary point of length or of energy (the free-particle action). We will see each of these in turn, solve the equations on the plane and the sphere, and then look at what happens to a family of neighbouring geodesics: they spread or focus according to the curvature. That last fact is the seed of the tidal-force equation of general relativity, and the geodesic-curvature term we meet first is exactly the boundary term of the Gauss–Bonnet theorem in the next section. We follow [Dray §18.8 and Ch. 19].

T2.7.1 Geodesic curvature

Take a curve C in an oriented surface Σ with orthonormal frame $\{e_1, e_2\}$, and let s be arc length along C. The unit tangent vector $\hat{T} = \mathrm{d}\vec{r}/\mathrm{d}s$ lies in the tangent plane, so it makes some angle α with e_1 (Figure T2.21):

$$\hat{T} = \cos\alpha\, e_1 + \sin\alpha\, e_2. \tag{T2.102}$$

The angle α is measured *relative to the frame*, so it changes for two reasons: because the curve turns, and because the frame turns. The connection knows how much of the second there is. Differentiate (T2.102) along the curve [Dray (18.61)–(18.62)]:

$$\begin{aligned} \mathrm{d}\hat{T} &\overset{(1)}{=} (-\sin\alpha\, e_1 + \cos\alpha\, e_2)\,\mathrm{d}\alpha + \cos\alpha\,\mathrm{d}e_1 + \sin\alpha\,\mathrm{d}e_2 \\ &\overset{(2)}{=} (-\sin\alpha\, e_1 + \cos\alpha\, e_2)\,\mathrm{d}\alpha + \cos\alpha\, e_2\,\omega^2{}_1 + \sin\alpha\, e_1\,\omega^1{}_2 \\ &\overset{(3)}{=} (-\sin\alpha\, e_1 + \cos\alpha\, e_2)\,(\mathrm{d}\alpha - \omega^1{}_2). \end{aligned} \tag{T2.103}$$

(1) Product rule on each term of (T2.102); $\mathrm{d}\cos\alpha = -\sin\alpha\,\mathrm{d}\alpha$ and $\mathrm{d}\sin\alpha = \cos\alpha\,\mathrm{d}\alpha$.
(2) The connection (T2.61): $\mathrm{d}e_1 = e_2\,\omega^2{}_1$ and $\mathrm{d}e_2 = e_1\,\omega^1{}_2$ (the diagonal entries vanish).
(3) $\omega^2{}_1 = -\omega^1{}_2$; then both connection terms are $-\omega^1{}_2$ times the bracket.

The bracket is the unit vector

$$\hat{N} = -\sin\alpha\, e_1 + \cos\alpha\, e_2, \tag{T2.104}$$

obtained by rotating $\hat{T}$ through $+90°$ in the tangent plane (for a surface in $\mathbb{R}^3$ with unit normal $\hat{n} = e_1 \times e_2$, it is $\hat{n} \times \hat{T}$). It points to the *left* of the direction of travel. So $\mathrm{d}\hat{T}$ is always along $\hat{N}$ (as it must be: $\hat{T}$ has unit length, so its derivative is perpendicular to it), and its size is the rate of turning to the left.

Working definition T2.14 (geodesic curvature)

The **geodesic curvature** κ_g of a curve in a surface is defined by $\mathrm{d}\hat{T} \cdot \hat{N} = \kappa_g\,\mathrm{d}s$ along the curve, i.e.

$$\boxed{\kappa_g\,\mathrm{d}s = \mathrm{d}\alpha - \omega^1{}_2} \tag{T2.105}$$

[Dray (18.64)]. *How to compute it:* write the tangent as $\cos\alpha\, e_1 + \sin\alpha\, e_2$, and subtract the connection 1-form (restricted to the curve [fed only the tangent $\hat{T}$, so it becomes $\omega^1{}_2(\hat{T})\,\mathrm{d}s$]) from the change of α.
Meaning: $\kappa_g > 0$ means the curve bends to the left as seen by an inhabitant of the surface, $\kappa_g < 0$ to the right. A curve with $\kappa_g = 0$ is a **geodesic**.

Two remarks make (T2.105) feel natural.

- *It is frame independent.* Rotate the frame by an angle λ. The tangent's angle relative to the new frame is $\alpha' = \alpha - \lambda$, and $\omega'^1{}_2 = \omega^1{}_2 - \mathrm{d}\lambda$ by (T2.82). So $\mathrm{d}\alpha' - \omega'^1{}_2 = \mathrm{d}\alpha - \mathrm{d}\lambda - \omega^1{}_2 + \mathrm{d}\lambda = \mathrm{d}\alpha - \omega^1{}_2$. The $\mathrm{d}\lambda$'s cancel: $\kappa_g\,\mathrm{d}s$ is a "gauge-invariant combination" of the matter field α and the gauge potential $\omega^1{}_2$, like the covariant derivative $\mathrm{d}\psi - A$ of a phase.
- *It compares with parallel transport.* By (T2.100), a parallel-transported vector has $\mathrm{d}\psi = \omega^1{}_2$. So $\kappa_g\,\mathrm{d}s = \mathrm{d}\alpha - \omega^1{}_2$ is the rate at which the tangent turns *relative to a parallel-transported vector*. A geodesic is a curve whose tangent is parallel transported along it.

Watch out: three different curvatures of a curve

A curve on a surface in $\mathbb{R}^3$ has an ordinary curvature κ (the bending seen in $\mathbb{R}^3$, (T2.93)) and this splits into two perpendicular pieces: the part along the surface normal $\hat{n}$, called the *normal curvature* κ_n, which the surface forces on the curve and which the shape operator measures; and the part along $\hat{N}$ within the surface, the geodesic curvature κ_g. So $\kappa^2 = \kappa_n^2 + \kappa_g^2$. A great circle on a sphere of radius a has $\kappa = 1/a$ (it is a circle of radius a in $\mathbb{R}^3$) but $\kappa_g = 0$: all its bending is forced by the sphere, none of it is chosen by the curve.

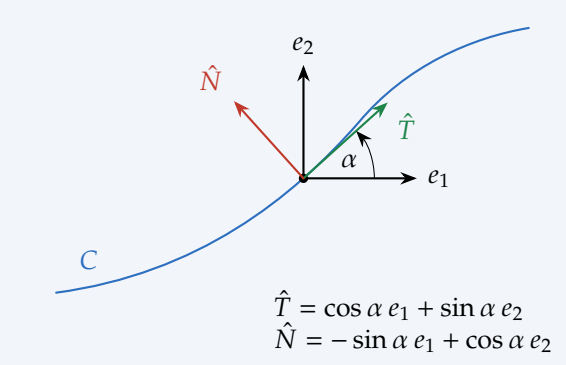


Figure T2.21: The tangent $\hat{T}$ of a curve makes an angle α with the frame vector e_1; the normal $\hat{N}$ (red) is $\hat{T}$ turned through 90° to the left, within the surface. The geodesic curvature measures how fast $\hat{T}$ swings towards $\hat{N}$. The angle α changes both because the curve bends and because the frame turns; subtracting $\omega^1{}_2$ removes the second effect.

Also do not confuse this $\hat{N}$, which depends only on the orientation of the surface, with the principal normal of the curve in $\mathbb{R}^3$ [Dray fn. 5 p. 225].

Worked example T2.35: Circles in the plane and on the sphere

A circle of radius ρ in the plane, traversed counterclockwise, in the polar frame ($e_1 = \hat{r}$, $e_2 = \hat{\phi}$). The tangent is $\hat{\phi} = e_2$, so $\alpha = \pi/2$ is constant, and $\omega^1{}_2 = -\mathrm{d}\phi$:

$$\kappa_g\, \mathrm{d}s \overset{(1)}{=} \mathrm{d}\left(\tfrac{\pi}{2}\right) - (-\mathrm{d}\phi) \overset{(2)}{=} \mathrm{d}\phi \overset{(3)}{=} \frac{\mathrm{d}s}{\rho}.$$

(1) (T2.105) with $\alpha = \pi/2$ and (T2.64).
(2) The derivative of a constant is zero.
(3) Arc length on the circle is $\mathrm{d}s = \rho\, \mathrm{d}\phi$.

So $\kappa_g = 1/\rho$, the familiar curvature of a circle, positive because the circle bends to the left. Notice where the answer came from: entirely from the connection. The tangent does not turn relative to the polar frame at all; it is the frame that turns, and the curve turns with it. *A latitude circle on a sphere of radius a*, $\theta = \theta_0$, traversed eastwards (increasing ϕ), in the frame $e_1 = \hat{\theta}$, $e_2 = \hat{\phi}$. Again $\hat{T} = \hat{\phi} = e_2$ and $\alpha = \pi/2$. With $\omega^1{}_2 = -\cos\theta\, \mathrm{d}\phi$,

$$\kappa_g\, \mathrm{d}s \overset{(1)}{=} 0 - (-\cos\theta_0\, \mathrm{d}\phi) \overset{(2)}{=} \cos\theta_0\, \frac{\mathrm{d}s}{a\sin\theta_0} \qquad \Longrightarrow \qquad \kappa_g = \frac{\cot\theta_0}{a}. \tag{T2.106}$$

(1) (T2.105) with (T2.70) evaluated at $\theta = \theta_0$.
(2) On the circle $\mathrm{d}s = a\sin\theta_0\, \mathrm{d}\phi$ (the circle has radius $a\sin\theta_0$ in $\mathbb{R}^3$).

Read the answer. In the northern hemisphere ($\theta_0 < \pi/2$), $\kappa_g > 0$: walking east, the path curves to the left, around the north pole. Near the pole, $\theta_0 \to 0$ and $\kappa_g \approx 1/(a\theta_0)$, the curvature of a small circle of radius $a\theta_0$ in a nearly flat region. In the southern hemisphere $\kappa_g < 0$. At the equator $\kappa_g = 0$: *the equator is a geodesic*, and it is the only latitude circle that is [Dray (18.65)]. Rotating the sphere, every great circle is a geodesic.

Worked example T2.36: Which circles on the torus are geodesics?

On the torus of revolution, with $\sigma^1 = a\,\mathrm{d}\theta$, $\sigma^2 = (R + a\cos\theta)\,\mathrm{d}\phi$ and $\omega^1{}_2 = \sin\theta\,\mathrm{d}\phi$ (T2.71), there are two families of circles.
Meridians (ϕ = const, going around the tube). The tangent is e_1, so $\alpha = 0$, and $\mathrm{d}\phi = 0$ along the curve:

$$\kappa_g\,\mathrm{d}s \overset{(1)}{=} 0 - \sin\theta\,\mathrm{d}\phi\big|_{\phi=\text{const}} \overset{(2)}{=} 0.$$

(1) (T2.105) with $\alpha = 0$.
(2) $\mathrm{d}\phi$ restricted to a curve of constant ϕ is zero.

Every meridian is a geodesic. (Symmetry says so too: reflection in the plane containing the meridian and the axis maps the torus to itself and fixes the meridian, so the meridian cannot prefer to bend left or right.)
Parallels ($\theta = \theta_0$, going around the axis). The tangent is e_2, so $\alpha = \pi/2$:

$$\kappa_g\,\mathrm{d}s \overset{(1)}{=} -\sin\theta_0\,\mathrm{d}\phi \overset{(2)}{=} -\frac{\sin\theta_0}{R + a\cos\theta_0}\,\mathrm{d}s.$$

(1) (T2.105) with $\alpha = \pi/2$ constant.
(2) $\mathrm{d}s = (R + a\cos\theta_0)\,\mathrm{d}\phi$ along the parallel.

So $\kappa_g = -\sin\theta_0/(R + a\cos\theta_0)$. It vanishes exactly for the *outer equator* $\theta_0 = 0$ and the *inner equator* $\theta_0 = \pi$: these two parallels are geodesics. The top circle $\theta_0 = \pi/2$ has $|\kappa_g| = 1/R$: it is a circle of radius R lying in a horizontal tangent plane, so all its bending is within the surface. Compare with Figure T2.19: the top circle is where $K = 0$, and there the torus looks locally like a plane.

T2.7.2 The geodesic equation

The geodesic-curvature definition is special to surfaces. Here is the definition that works in any dimension [Dray §19.1]. A free particle has zero acceleration. Let $\vec{v} = \mathrm{d}\vec{r}/\mathrm{d}t$ be the velocity of a curve with parameter t. Along the curve, $\mathrm{d}\vec{r} = \vec{v}\,\mathrm{d}t$, so

$$\sigma^i = v^i\,\mathrm{d}t \qquad \text{(restricted to the curve)}, \tag{T2.107}$$

where v^i are the frame components of $\vec{v}$. Differentiate the velocity:

$$\mathrm{d}\vec{v} \overset{(1)}{=} \mathrm{d}(v^j e_j) \overset{(2)}{=} \big(\mathrm{d}v^j + \omega^j{}_i\,v^i\big)e_j \overset{(3)}{=} \big(\mathrm{d}v^j + \Gamma^j{}_{ik}\,v^i\sigma^k\big)e_j \overset{(4)}{=} \big(\dot{v}^j + \Gamma^j{}_{ik}\,v^i v^k\big)\mathrm{d}t\;e_j. \tag{T2.108}$$

(1) Expand in the frame.
(2) The covariant derivative formula (T2.63).
(3) Expand the connection in the coframe, (T2.73).
(4) Along the curve $\mathrm{d}v^j = \dot{v}^j\,\mathrm{d}t$ and (T2.107).

Dividing by $\mathrm{d}t$ gives the acceleration $\dot{\vec{v}} = \nabla_{\vec{v}}\vec{v} = (\dot{v}^j + \Gamma^j{}_{ik}v^i v^k)\,e_j$ [Dray (19.4)–(19.6)].

Working definition T2.15 (geodesic)

A curve is a **geodesic** if its velocity is parallel transported [its frame components change only to undo the rotation of the frame] along it, $\nabla_{\vec{v}}\vec{v} = 0$:

$$\boxed{\dot{v}^j + \Gamma^j{}_{ik}\,v^i v^k = 0} \quad \text{(frame form)}, \qquad \ddot{x}^\mu + \Gamma^\mu{}_{\nu\rho}\,\dot{x}^\nu\dot{x}^\rho = 0 \quad \text{(coordinate form)}. \tag{T2.109}$$

[in the coordinate form, $\Gamma^\mu{}_{\nu\rho}$ are the Christoffel symbols of the basis ∂_μ, built from $g_{\mu\nu}$]

Recipe: write $\vec{v}$ in the frame, compute $\mathrm{d}\vec{v}$ using the derivatives of the frame vectors, divide by $\mathrm{d}t$ and set each component to zero. *Meaning:* the path of a free particle, the "straightest" curve.

How does this compare with $\kappa_g = 0$? Write $\vec{v} = v\,\hat{T}$ with speed $v = |\vec{v}|$. Then $\dot{\vec{v}} = \dot{v}\,\hat{T} + v\,\dot{\hat{T}}$, and $\dot{\hat{T}}$ is perpendicular to $\hat{T}$ (it is along $\hat{N}$, by (T2.103)). The two pieces are perpendicular, so $\dot{\vec{v}} = 0$ exactly when $\dot{v} = 0$ *and* $\dot{\hat{T}} = 0$. On a surface $\dot{\hat{T}} = \kappa_g v\,\hat{N}$, so

$$\nabla_{\vec{v}}\vec{v} = 0 \iff \kappa_g = 0 \text{ and the speed is constant}$$

[Dray (19.9)–(19.15)]. The two definitions pick out the same curves; the acceleration definition also fixes the parametrisation up to $t \mapsto at + b$.

Worked example T2.37: Polar coordinates: the geodesic equation, and why it gives straight lines

Write $\vec{v} = \dot{r}\,\hat{r} + r\dot{\phi}\,\hat{\phi}$ (from $\mathrm{d}\vec{r} = \mathrm{d}r\,\hat{r} + r\,\mathrm{d}\phi\,\hat{\phi}$). Differentiate, using $\mathrm{d}\hat{r} = \mathrm{d}\phi\,\hat{\phi}$ and $\mathrm{d}\hat{\phi} = -\mathrm{d}\phi\,\hat{r}$ (T2.55):

$$\begin{aligned}\mathrm{d}\vec{v} &\overset{(1)}{=} \mathrm{d}\dot{r}\,\hat{r} + \dot{r}\,\mathrm{d}\hat{r} + \mathrm{d}(r\dot{\phi})\,\hat{\phi} + r\dot{\phi}\,\mathrm{d}\hat{\phi}\\ &\overset{(2)}{=} \big(\mathrm{d}\dot{r} - r\dot{\phi}\,\mathrm{d}\phi\big)\hat{r} + \big(\mathrm{d}(r\dot{\phi}) + \dot{r}\,\mathrm{d}\phi\big)\hat{\phi}\\ &\overset{(3)}{=} \Big[\big(\ddot{r} - r\dot{\phi}^2\big)\hat{r} + \big(r\ddot{\phi} + 2\dot{r}\dot{\phi}\big)\hat{\phi}\Big]\mathrm{d}t. \end{aligned} \tag{T2.110}$$

(1) Product rule on each term.
(2) Insert the derivatives of $\hat{r}, \hat{\phi}$ and collect.
(3) Along the curve $\mathrm{d}f = \dot{f}\,\mathrm{d}t$; $\frac{\mathrm{d}}{\mathrm{d}t}(r\dot{\phi}) = \dot{r}\dot{\phi} + r\ddot{\phi}$.

These are the radial and transverse accelerations every mechanics student knows (centripetal $-r\dot{\phi}^2$, Coriolis $2\dot{r}\dot{\phi}$). The geodesic equations are [Dray (19.20)]

$$\ddot{r} - r\dot{\phi}^2 = 0, \qquad r\ddot{\phi} + 2\dot{r}\dot{\phi} = 0.$$

Solving them [Dray §§19.4–19.5]. Multiply the second by r: it says $\frac{\mathrm{d}}{\mathrm{d}t}(r^2\dot{\phi}) = 0$, so $r^2\dot{\phi} = \ell$ is conserved (angular momentum per unit mass). Multiply the first by $\dot{r}$ and use $\dot{\phi} = \ell/r^2$: $\dot{r}\ddot{r} - \ell^2\dot{r}/r^3 = \frac{\mathrm{d}}{\mathrm{d}t}\big(\frac{1}{2}\dot{r}^2 + \frac{1}{2}\ell^2/r^2\big) = 0$, so $\dot{r}^2 + r^2\dot{\phi}^2 = |\vec{v}|^2$ is conserved (energy). Choose unit speed, $t = \lambda$ the arc length. Then

$$\dot{r}^2 = 1 - \frac{\ell^2}{r^2} \quad\Longrightarrow\quad r^2 = \lambda^2 + \ell^2,$$
$$\dot{\phi} = \frac{\ell}{\lambda^2 + \ell^2} \quad\Longrightarrow\quad \phi - \alpha_0 = \arctan\frac{\lambda}{\ell},$$

(check the first: $\dot{r} = \lambda/r$, so $\dot{r}^2 = \lambda^2/r^2 = (r^2 - \ell^2)/r^2$; the origin of λ is chosen at the point of closest approach). Eliminate λ:

$$r\cos(\phi - \alpha_0) \overset{(1)}{=} r\cos\Big(\arctan\frac{\lambda}{\ell}\Big) \overset{(2)}{=} \sqrt{\lambda^2 + \ell^2}\,\frac{\ell}{\sqrt{\lambda^2 + \ell^2}} = \ell,$$

(1) The solution for ϕ.
(2) $\cos(\arctan u) = 1/\sqrt{1 + u^2}$ and $r = \sqrt{\lambda^2 + \ell^2}$.

that is, $x\cos\alpha_0 + y\sin\alpha_0 = \ell$: a straight line at distance ℓ from the origin [Dray (19.53)]. The angular momentum of a unit-speed particle is its impact parameter. We never used Cartesian coordinates, yet the connection $\omega^1{}_2 = -\mathrm{d}\phi$ knew that the plane is flat and produced straight lines.

T2.7.3 Geodesics from an action principle

Physicists usually meet geodesics as the paths of free particles, derived from a Lagrangian. For a curve $x^\mu(t)$, $0 \le t \le 1$, in a manifold with metric $g_{\mu\nu}$, define the **length** and the **energy**

$$L(\gamma) = \int_0^1 \sqrt{g_{\mu\nu}\dot{x}^\mu \dot{x}^\nu}\, \mathrm{d}t, \qquad E(\gamma) = \int_0^1 g_{\mu\nu}\dot{x}^\mu \dot{x}^\nu\, \mathrm{d}t. \tag{T2.111}$$

The energy is the free-particle action (twice the kinetic energy, integrated). Its Euler–Lagrange equations, with $\mathcal{L} = g_{\mu\nu}\dot{x}^\mu\dot{x}^\nu$, are

$$\begin{aligned} 0 &\overset{(1)}{=} \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial \mathcal{L}}{\partial \dot{x}^\rho} - \frac{\partial \mathcal{L}}{\partial x^\rho} \overset{(2)}{=} \frac{\mathrm{d}}{\mathrm{d}t}\left(2g_{\rho\nu}\dot{x}^\nu\right) - \partial_\rho g_{\mu\nu}\dot{x}^\mu\dot{x}^\nu \\ &\overset{(3)}{=} 2g_{\rho\nu}\ddot{x}^\nu + 2\partial_\mu g_{\rho\nu}\dot{x}^\mu\dot{x}^\nu - \partial_\rho g_{\mu\nu}\dot{x}^\mu\dot{x}^\nu \\ &\overset{(4)}{=} 2g_{\rho\nu}\ddot{x}^\nu + \left(\partial_\mu g_{\rho\nu} + \partial_\nu g_{\rho\mu} - \partial_\rho g_{\mu\nu}\right)\dot{x}^\mu\dot{x}^\nu \\ &\overset{(5)}{=} 2g_{\rho\sigma}\left(\ddot{x}^\sigma + \Gamma^\sigma{}_{\mu\nu}\dot{x}^\mu\dot{x}^\nu\right). \end{aligned} \tag{T2.112}$$

(1) Euler–Lagrange equation for the coordinate x^ρ.
(2) $\partial\mathcal{L}/\partial\dot{x}^\rho = 2g_{\rho\nu}\dot{x}^\nu$ by the symmetry of g.
(3) Chain rule: $\frac{\mathrm{d}}{\mathrm{d}t}g_{\rho\nu} = \partial_\mu g_{\rho\nu}\,\dot{x}^\mu$.
(4) Split $2\partial_\mu g_{\rho\nu}\dot{x}^\mu\dot{x}^\nu$ into two equal halves and relabel $\mu \leftrightarrow \nu$ in one of them.
(5) The coordinate Christoffel symbol $\Gamma^\sigma{}_{\mu\nu} = \frac{1}{2}g^{\sigma\rho}(\partial_\mu g_{\rho\nu} + \partial_\nu g_{\rho\mu} - \partial_\rho g_{\mu\nu})$ (the dictionary box of Section T2.5.8).

So the stationary curves of E are exactly the geodesics (T2.109). This is usually the fastest way to *get* the geodesic equations: write down the kinetic energy and turn the crank.

Energy versus length. The length does not depend on how the curve is parametrised; the energy does. They are related by the Cauchy–Schwarz inequality:

$$L(\gamma)^2 \overset{(1)}{=} \left(\int_0^1 1\cdot|\dot{\gamma}|\,\mathrm{d}t\right)^2 \overset{(2)}{\le} \int_0^1 1^2\,\mathrm{d}t \int_0^1 |\dot{\gamma}|^2\,\mathrm{d}t \overset{(3)}{=} E(\gamma), \tag{T2.113}$$

(1) Definition of L, with $|\dot{\gamma}| = \sqrt{g_{\mu\nu}\dot{x}^\mu\dot{x}^\nu}$.
(2) Cauchy–Schwarz for the functions 1 and $|\dot{\gamma}|$; equality holds exactly when $|\dot{\gamma}|$ is constant.
(3) $\int_0^1 \mathrm{d}t = 1$ and the definition of E.

So minimising E does two things at once: it minimises the length, and it forces constant speed. That is why the geodesic equation automatically conserves speed (we saw it in polar coordinates), and why the energy is the better functional for Morse theory (Chapter 7).

Worked example T2.38: The sphere from its Lagrangian; conserved angular momentum

On the sphere of radius a, $\mathcal{L} = a^2(\dot{\theta}^2 + \sin^2\theta\,\dot{\phi}^2)$. The Euler–Lagrange equations are

$$\begin{aligned} \theta: &\quad \frac{\mathrm{d}}{\mathrm{d}t}\left(2a^2\dot{\theta}\right) - 2a^2\sin\theta\cos\theta\,\dot{\phi}^2 = 0 &&\iff \quad \ddot{\theta} - \sin\theta\cos\theta\,\dot{\phi}^2 = 0, \\ \phi: &\quad \frac{\mathrm{d}}{\mathrm{d}t}\left(2a^2\sin^2\theta\,\dot{\phi}\right) = 0 &&\iff \quad \ddot{\phi} + 2\cot\theta\,\dot{\theta}\dot{\phi} = 0, \end{aligned}$$

which are Dray's (19.29), obtained there from the frame computation ($\vec{v} = a\dot{\theta}\,\hat{\theta} + a\sin\theta\,\dot{\phi}\,\hat{\phi}$ and $\mathrm{d}\hat{\theta} = \cos\theta\,\mathrm{d}\phi\,\hat{\phi}$, $\mathrm{d}\hat{\phi} = -\cos\theta\,\mathrm{d}\phi\,\hat{\theta}$ on the sphere; try it). The equator ($\theta = \pi/2$, $\dot{\phi}$

constant) solves both, since $\cos\frac{\pi}{2} = 0$.
The ϕ equation is a conservation law: ϕ does not appear in $\mathcal{L}$, so its conjugate momentum

$$\ell = \sin^2\theta\,\dot\phi \qquad \text{(unit sphere)}$$

is conserved [Dray (19.39)]. It is the z-component of angular momentum, $x\dot y - y\dot x = (\sin\theta)^2\dot\phi$. The θ equation, multiplied by $\dot\theta$, integrates to conservation of $\dot\theta^2 + \sin^2\theta\,\dot\phi^2$, the speed squared [Dray (19.41)].
The quick way to all solutions. A particle sliding freely on a frictionless sphere feels only the normal force, which is radial, so it exerts no torque about the centre. Hence the full angular-momentum vector $\vec L = \vec r \times \vec v$ is constant. Since $\vec r \perp \vec L$ at all times, the particle stays in the plane through the centre perpendicular to $\vec L$: *geodesics on the sphere are great circles*. The z-component $L_z = \ell$ is Dray's constant, and for a unit-speed geodesic on the unit sphere $\left|\vec L\right| = 1$, so $\ell = \cos i$ where i is the angle between the plane of the great circle and the equatorial plane. Dray reaches the same conclusion by integrating $\mathrm{d}\phi/\mathrm{d}\theta$ directly, obtaining $\ell\cot\theta = \sqrt{1-\ell^2}\,\sin(\alpha_0 - \phi)$, which after multiplying by $\sin\theta$ is the equation of a plane through the origin [Dray (19.54)–(19.61)].

Worked example T2.39: Flying between two cities at latitude 45°

Two cities on the unit sphere at colatitude $\theta = \pi/4$, with longitudes 0 and $\pi/2$: $\vec u = \left(\frac{1}{\sqrt2}, 0, \frac{1}{\sqrt2}\right)$, $\vec w = \left(0, \frac{1}{\sqrt2}, \frac{1}{\sqrt2}\right)$. The great-circle distance is the angle between the position vectors:

$$\cos d \overset{(1)}{=} \vec u\cdot\vec w \overset{(2)}{=} 0 + 0 + \tfrac12 \quad\Longrightarrow\quad d = \frac{\pi}{3} \approx 1.047.$$

(1) Arc length on a unit circle equals the angle subtended; the great circle lies in the plane of $\vec u$ and $\vec w$.
(2) Multiply componentwise.

Along the latitude circle the distance is $\sin\frac{\pi}{4}\cdot\frac{\pi}{2} = \frac{\pi}{2\sqrt2} \approx 1.111$. On the Earth ($a = 6371$ km) the two routes are about 6670 km and 7080 km. Where does the great circle go? Its plane has normal $\vec u\times\vec w = \left(-\frac12, -\frac12, \frac12\right)$, of length $\sqrt3/2$, so $\ell = L_z = \pm 1/\sqrt3$. At the highest point $\dot\theta = 0$, so conservation of speed gives $\sin^2\theta\,\dot\phi^2 = 1$, and with $\ell = \sin^2\theta\,\dot\phi$ this means $\sin\theta_{\min} = |\ell| = 1/\sqrt3$: $\theta_{\min} \approx 35.3^\circ$, i.e. latitude 54.7°. The shortest route bulges *towards the pole*, which is why flights from Europe to North America pass over Greenland. The explicit parametrisation is $\vec r(\beta) = \vec u\cos\beta + \hat m\sin\beta$, $0 \le \beta \le d$, with $\hat m$ the unit vector in the plane perpendicular to $\vec u$ (Dray's construction, [Dray Fig. 19.1]).

Try it: Clairaut's relation

On any surface of revolution, $\mathrm{d}s^2 = \mathrm{d}u^2 + \rho(u)^2\mathrm{d}\phi^2$ (with u arc length along a meridian and ρ the distance from the axis), the Lagrangian does not depend on ϕ. Show that along a unit-speed geodesic $\rho\sin\psi = \text{const}$, where ψ is the angle between the geodesic and the meridian. *Answer:* the conserved momentum is $\rho^2\dot\phi$, and the e_2 component of the unit velocity is $\rho\dot\phi = \sin\psi$. On the torus this says that a geodesic starting on the outer equator at a small angle to it can never reach the inner equator (where ρ is smallest) unless $\rho_{\text{out}}\sin\psi \le \rho_{\text{in}}$: some geodesics wind forever around the outside.

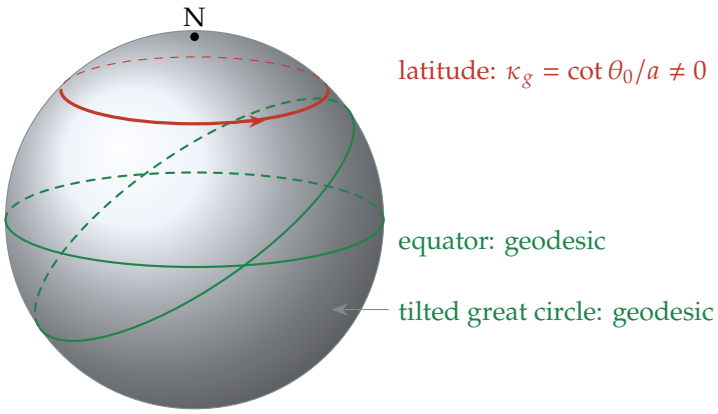


Figure T2.22: Geodesics and non-geodesics on a sphere. Great circles (green), such as the equator and any tilted circle through the centre, do not bend to either side within the sphere. The latitude circle (red) bends to the left, towards the north pole, with geodesic curvature $\cot\theta_0/a$; an inhabitant walking along it must keep turning left.

T2.7.4 Neighbouring geodesics: focusing, spreading and curvature

Two particles leave the north pole along neighbouring meridians. On a plane their separation would grow linearly forever. On a sphere they separate at first, then come together again and meet at the south pole. On a saddle they separate faster than linearly. Curvature is exactly the rate at which neighbouring geodesics accelerate towards or away from each other. The cleanest way to see it uses coordinates adapted to geodesics.

Geodesic polar coordinates. Pick a point p and send out unit speed geodesics in every direction ϕ. Label a point by the direction ϕ of the geodesic that reaches it and the distance r travelled. Near p this is an honest coordinate system (this is Gauss's lemma, which we accept), and the metric takes the form

$$\mathrm{d}s^2 = \mathrm{d}r^2 + G(r,\phi)^2\,\mathrm{d}\phi^2, \qquad G(0,\phi) = 0, \quad \partial_r G(0,\phi) = 1. \tag{T2.114}$$

The conditions at $r = 0$ say that very close to p the metric is the flat polar metric $\mathrm{d}r^2 + r^2\mathrm{d}\phi^2$. The function G has a direct meaning: two radial geodesics at angles ϕ and $\phi + \delta\phi$ are a distance $J = G\,\delta\phi$ apart at radius r. Now compute the curvature with the structure equations. The coframe is $\sigma^1 = \mathrm{d}r$, $\sigma^2 = G\,\mathrm{d}\phi$, and

$$\begin{aligned} \mathrm{d}\sigma^2 &\overset{(1)}{=} \partial_r G\;\mathrm{d}r\wedge\mathrm{d}\phi \overset{(2)}{=} \omega^1{}_2\wedge\sigma^1 \qquad\Longrightarrow\qquad \omega^1{}_2 = -\partial_r G\,\mathrm{d}\phi, \\ \Omega^1{}_2 &\overset{(3)}{=} \mathrm{d}\big(-\partial_r G\,\mathrm{d}\phi\big) \overset{(4)}{=} -\partial_r^2 G\;\mathrm{d}r\wedge\mathrm{d}\phi \overset{(5)}{=} -\frac{\partial_r^2 G}{G}\;\sigma^1\wedge\sigma^2. \end{aligned} \tag{T2.115}$$

(1) $\mathrm{d}(G\,\mathrm{d}\phi) = \partial_r G\,\mathrm{d}r\wedge\mathrm{d}\phi + \partial_\phi G\,\mathrm{d}\phi\wedge\mathrm{d}\phi$, and the second term vanishes.
(2) The first structure equation for $i = 2$, $\mathrm{d}\sigma^2 = -\omega^2{}_1\wedge\sigma^1 = \omega^1{}_2\wedge\sigma^1$. The guess $\omega^1{}_2 = -\partial_r G\,\mathrm{d}\phi$ gives $-\partial_r G\,\mathrm{d}\phi\wedge\mathrm{d}r = \partial_r G\,\mathrm{d}r\wedge\mathrm{d}\phi$; it also satisfies the $i = 1$ equation, $0 = \mathrm{d}\sigma^1 = -\omega^1{}_2\wedge\sigma^2 \propto \mathrm{d}\phi\wedge\mathrm{d}\phi$.
(3) The curvature in 2D is $\mathrm{d}\omega^1{}_2$, (T2.87).
(4) $\mathrm{d}(\partial_r G) = \partial_r^2 G\,\mathrm{d}r + \partial_\phi\partial_r G\,\mathrm{d}\phi$; the $\mathrm{d}\phi$ part dies against $\mathrm{d}\phi$.
(5) $\mathrm{d}r\wedge\mathrm{d}\phi = \sigma^1\wedge\sigma^2/G$.

So $K = -\partial_r^2 G/G$, which we write as the **Jacobi equation**

$$\boxed{\frac{\partial^2 J}{\partial r^2} + K\,J = 0}, \qquad J = G\,\delta\phi. \tag{T2.116}$$

The separation of neighbouring geodesics obeys a harmonic-oscillator equation whose "spring constant" is the Gaussian curvature. (The radial lines really are geodesics: along ϕ = const the tangent is e_1, $\alpha = 0$, and $\omega^1{}_2 \propto \mathrm{d}\phi$ vanishes on the curve, so $\kappa_g = 0$.)

Worked example T2.40: Constant curvature: sin, linear, sinh

With $J(0) = 0$, $J'(0) = \delta\phi$, the Jacobi equation for constant K has the solutions

$$K = +1 : \; G = \sin r, \qquad K = 0 : \; G = r, \qquad K = -1 : \; G = \sinh r.$$

These are the round sphere $\mathrm{d}r^2 + \sin^2 r\,\mathrm{d}\phi^2$ (with $r = \theta$), the flat plane in polar coordinates, and the hyperbolic plane $\mathrm{d}r^2 + \sinh^2 r\,\mathrm{d}\phi^2$. On the sphere the meridians reconverge at $r = \pi$: every geodesic from the north pole passes through the south pole. The south pole is a **conjugate point** of the north pole. Past a conjugate point a geodesic is no longer the shortest path (go from N past S to a point beyond: the other way round is shorter).

An ant's curvature meter. The circumference of the geodesic circle of radius r is $C(r) = \int_0^{2\pi} G\,\mathrm{d}\phi$. Expanding the solution of (T2.116) with K approximately constant near p, $G = r - \frac{1}{6}Kr^3 + O(r^4)$, so

$$C(r) \overset{(1)}{=} 2\pi\Big(r - \frac{K(p)}{6}\,r^3 + O(r^4)\Big) \quad\Longrightarrow\quad K(p) \overset{(2)}{=} \lim_{r\to 0} \frac{3}{\pi}\,\frac{2\pi r - C(r)}{r^3}.$$

(1) Insert the Taylor series of G, which follows from $G(0) = 0$, $G'(0) = 1$, $G''(0) = 0$ and $G'''(0) = -K(p)$ by differentiating (T2.116).

(2) Solve for $K(p)$.

On the unit sphere $C(r) = 2\pi \sin r$ and the limit is $\frac{3}{\pi} \cdot 2\pi \cdot \frac{1}{6} = 1$. An inhabitant who measures that circles are *shorter* than $2\pi r$ knows that $K > 0$, without ever leaving the surface: the Theorema Egregium again.

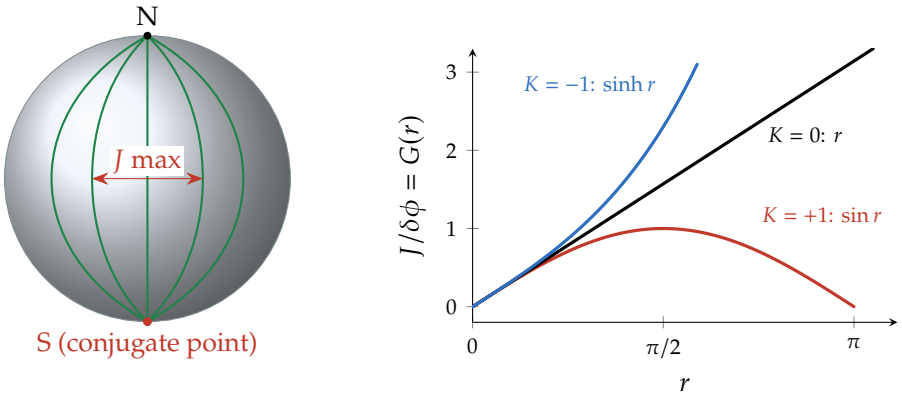


Figure T2.23: Left: geodesics leaving the north pole of a sphere spread apart, reach maximal separation at the equator, and refocus at the south pole, the conjugate point. Right: the separation $J = G(r)\,\delta\phi$ of two neighbouring geodesics solves $J'' + KJ = 0$. Positive curvature focuses ($\sin r$), zero curvature gives linear spreading, negative curvature defocuses exponentially ($\sinh r$).

Physical picture: tidal forces, and why gravity is curvature

In general relativity freely falling bodies follow geodesics of spacetime, and two nearby freely falling bodies drift apart or together according to the *geodesic deviation equation* $\frac{D^2 J^\mu}{\mathrm{d}\tau^2} = -R^\mu{}_{\nu\rho\sigma} u^\nu J^\rho u^\sigma$, the higher-dimensional version of (T2.116). Two apples dropped

side by side above the Earth approach each other: that relative acceleration is a tidal effect, it cannot be removed by choosing a freely falling frame, and it is curvature. The focusing of geodesics by positive curvature is also the mechanism behind the singularity theorems (Raychaudhuri's equation). In Morse theory (Chapter 7) the number of conjugate points along a geodesic is its *index* as a critical point of the energy E: each conjugate point is a direction in which the geodesic can be shortened.

Reading Nash (pp. 205–206, (7.55)–(7.59))

Nash introduces geodesics only in his Morse theory chapter, as the critical points of the energy functional $E(\gamma) = \int_0^1 |\mathrm{d}\gamma/\mathrm{d}t|^2 \,\mathrm{d}t$ on the space of paths from p to q. He writes the geodesic equation $\frac{\mathrm{d}^2\gamma^i}{\mathrm{d}t^2} + \Gamma^i{}_{jk}\frac{\mathrm{d}x^j}{\mathrm{d}t}\frac{\mathrm{d}x^k}{\mathrm{d}t} = 0$ [Nash p. 206, (7.56)]: this is our (T2.109), derived in (T2.112). He states that $E(\gamma) = L^2(\gamma)$ for a geodesic and $E(\gamma) \geq L^2(\gamma)$ in general [Nash (7.57)–(7.58)]: that is the Cauchy–Schwarz computation (T2.113). His remark that "there exist infinitely many geodesics joining a pair of points on a sphere" [Nash p. 192] is visible in Figure T2.22: go round the great circle through the two points once, twice, ..., in either direction. Each winding is a critical point of E with a larger index (more conjugate points passed), and Morse theory uses these to compute the topology of the loop space.

Recap

- Geodesic curvature: $\kappa_g\,\mathrm{d}s = \mathrm{d}\alpha - \omega^1{}_2$, the turning of the tangent relative to a parallel-transported vector. It is gauge invariant. Geodesics have $\kappa_g = 0$.
- Examples: circles in the plane $\kappa_g = 1/\rho$; latitude circles $\kappa_g = \cot\theta_0/a$ (only the equator is a geodesic); on the torus all meridians and the outer and inner equators are geodesics.
- Geodesic equation: $\nabla_{\vec v}\vec v = 0$, i.e. $\dot v^j + \Gamma^j{}_{ik}v^i v^k = 0$. Equivalent to $\kappa_g = 0$ at constant speed. It is the Euler–Lagrange equation of the energy $E = \int g_{\mu\nu}\dot x^\mu \dot x^\nu \,\mathrm{d}t$, and $E \geq L^2$.
- Plane: $r\cos(\phi - \alpha_0) = \ell$, straight lines. Sphere: great circles, by conservation of $\vec L = \vec r \times \vec v$; the shortest route bulges towards the pole.
- Jacobi equation $J'' + KJ = 0$: positive curvature focuses geodesics (conjugate points), negative curvature spreads them; $C(r) = 2\pi(r - Kr^3/6 + \dots)$.

T2.8 The Gauss–Bonnet theorem: integrated curvature is topology

We have computed the total curvature $\int K\,\mathrm{d}A$ of two closed surfaces. For a sphere of radius a it is $\frac{1}{a^2}\cdot 4\pi a^2 = 4\pi$, whatever a is. For a torus of revolution it is 0, whatever the radii R and a are (T2.92). In both cases the metric details dropped out completely. That is suspicious. So:

> *Is the total curvature of a closed surface independent of the metric? If so, what topological number is it?*

The answer is the **Gauss–Bonnet theorem**: $\int_\Sigma K\,\mathrm{d}A = 2\pi\chi(\Sigma)$, where χ is the Euler characteristic of Section T2.4.7. We will prove it in the physicist's way, with nothing but Stokes' theorem applied to the connection 1-form and some careful bookkeeping of angles. Along the way we get the angle sum of a triangle on a curved surface, the version with boundary that Nash quotes,

a second proof of metric independence by the "exact form over a closed surface" argument, the Gauss map and its degree, the Poincaré–Hopf theorem on zeros of vector fields, and finally the reading of Gauss–Bonnet as the first example of Chern–Weil theory and of the index theorem. We follow [Dray §§18.9–18.11].

T2.8.1 The turning tangent

Walk once around the edge of a region, counterclockwise. How much does your heading turn in total? In the plane, for the boundary of a disc, the answer is one full turn, 2π. For a circle this is obvious. For any other simple closed curve, deform it continuously into a circle through simple closed curves: the total turning is always an integer multiple of 2π (the heading returns to where it started), it changes continuously during the deformation, so it cannot change at all. This is Hopf's **turning tangent theorem** (the *Umlaufsatz*).

If the curve has corners, the heading jumps at each corner by the **exterior angle** ε_k (positive for a left turn), and the smooth parts supply the rest (Figure T2.24):

$$\oint_{\text{smooth parts}} \mathrm{d}\alpha + \sum_k \varepsilon_k = 2\pi. \tag{T2.117}$$

The exterior angle is $\varepsilon_k = \pi - \eta_k$, with η_k the interior angle. For a square, $4 \times \frac{\pi}{2} = 2\pi$ comes entirely from the corners.

The same statement holds for a region D on a curved surface, provided D is a topological disc and we measure the heading α relative to a frame that is smooth on all of D. The argument is the same: shrink D continuously to a tiny disc around one of its points. Throughout, the total turning of α is a multiple of 2π and varies continuously, and at the end the region is so small that the surface is flat to any desired accuracy and the frame is constant across it, so we are back in the plane with answer 2π.

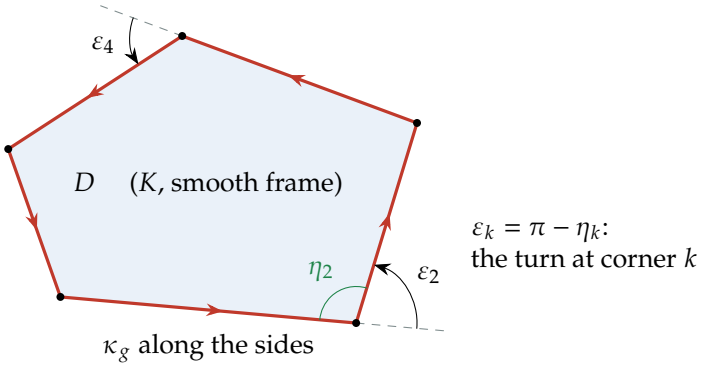


Figure T2.24: A region D with a piecewise smooth boundary, traversed counterclockwise (red). At each corner the heading jumps by the exterior angle ε_k, the angle between the continuation of the incoming side (dashed) and the outgoing side; the interior angle is $\eta_k = \pi - \varepsilon_k$. Along the smooth sides the heading turns at the rate κ_g relative to parallel transport. The sides need not be straight or geodesic.

T2.8.2 Local Gauss–Bonnet

Let D be a topological disc on an oriented surface, with a piecewise smooth boundary $C = \partial D$ traversed in the boundary orientation (region on the left), exterior angles ε_k at the corners, and let $\{e_1, e_2\}$ be an orthonormal frame that is smooth on D. Integrate the geodesic curvature

around the smooth parts of C [Dray (18.66)–(18.70)]:

$$\begin{aligned}\oint_C \kappa_g \,\mathrm{d}s &\overset{(1)}{=} \oint_C \mathrm{d}\alpha - \oint_C \omega^1{}_2 \\ &\overset{(2)}{=} \Big(2\pi - \sum_k \varepsilon_k\Big) - \oint_C \omega^1{}_2 \\ &\overset{(3)}{=} 2\pi - \sum_k \varepsilon_k - \int_D \mathrm{d}\omega^1{}_2 \\ &\overset{(4)}{=} 2\pi - \sum_k \varepsilon_k - \int_D K\,\mathrm{d}A. \end{aligned} \tag{T2.118}$$

(1) The definition (T2.105) of κ_g, integrated along the smooth pieces.
(2) The turning tangent theorem (T2.117); this is where we need D to be a disc and the frame to be smooth on it.
(3) Stokes' theorem for the 1-form $\omega^1{}_2$, which is smooth on D because the frame is. (The corners do not matter: Stokes holds for piecewise smooth boundaries.)
(4) The curvature 2-form, (T2.88): $\mathrm{d}\omega^1{}_2 = K\,\sigma^1 \wedge \sigma^2 = K\,\mathrm{d}A$.

Rearranging:

Key idea: local Gauss–Bonnet

For a disc-shaped region D with piecewise smooth boundary,

$$\boxed{\int_D K\,\mathrm{d}A + \oint_{\partial D} \kappa_g \,\mathrm{d}s + \sum_k \varepsilon_k = 2\pi} \tag{T2.119}$$

[Dray (18.69)–(18.70)]. The three ways a walker's heading can turn (bending of the path, kinks at the corners, and curvature of the surface enclosed) always add up to exactly one full turn.

Notice what happened to the frame: it appears in the middle of the derivation and has disappeared from the result. Every term in (T2.119) is gauge invariant. Compare (T2.101): the curvature term is the holonomy of the boundary loop. So local Gauss–Bonnet also says: *the rotation of a vector parallel transported around ∂D equals 2π minus the total turning of the path.* For a geodesic polygon that is $2\pi - \sum \varepsilon_k$, a purely angular quantity.

Worked example T2.41: Four checks

A flat polygon. $K = 0$, straight sides ($\kappa_g = 0$), so $\sum \varepsilon_k = 2\pi$. For a triangle, $\sum(\pi - \eta_k) = 2\pi$, i.e. $\eta_1 + \eta_2 + \eta_3 = \pi$.
A disc of radius ρ in the plane. $K = 0$, no corners, and $\kappa_g = 1/\rho$ from Example T2.35: $\oint \kappa_g \,\mathrm{d}s = \frac{1}{\rho} \cdot 2\pi\rho = 2\pi$.
A polar cap on the unit sphere, $0 \le \theta \le \theta_0$. The three terms are:

$$\int_D K\,\mathrm{d}A + \oint \kappa_g \,\mathrm{d}s \overset{(1)}{=} 2\pi(1 - \cos\theta_0) + \cot\theta_0 \cdot 2\pi \sin\theta_0 \overset{(2)}{=} 2\pi - 2\pi\cos\theta_0 + 2\pi\cos\theta_0 = 2\pi.$$

(1) $K = 1$ and the area of the cap from Example T2.34; $\kappa_g = \cot\theta_0$ from (T2.106) times the length $2\pi\sin\theta_0$ of the circle.
(2) $\cot\theta_0 \sin\theta_0 = \cos\theta_0$.

As the cap grows, the area term grows and the boundary term shrinks, in exact compensa-

tion. For $\theta_0 > \pi/2$ the cap is more than a hemisphere, $\kappa_g < 0$ (the boundary bends to the right, seen from inside the cap), and the formula still holds.
The hemisphere. $\theta_0 = \pi/2$: the equator is a geodesic, and all 2π comes from the enclosed curvature. (Compare Figure T2.20: a vector carried around the equator returns unrotated relative to the tangent of the equator, but rotated by 2π relative to a frame smooth on the hemisphere; both statements say the same thing.)

T2.8.3 Geodesic triangles and the angle excess

A **geodesic triangle** has three geodesic sides ($\kappa_g = 0$) and interior angles η_1, η_2, η_3. Local Gauss–Bonnet with $\varepsilon_k = \pi - \eta_k$ gives

$$\int_T K \, \mathrm{d}A \overset{(1)}{=} 2\pi - \sum_{k=1}^{3} (\pi - \eta_k) \overset{(2)}{=} \eta_1 + \eta_2 + \eta_3 - \pi. \tag{T2.120}$$

(1) (T2.119) with no κ_g term.
(2) $2\pi - 3\pi = -\pi$.

The angle excess of a geodesic triangle equals the total curvature inside it [Dray (18.71)]. On a sphere of radius a ($K = 1/a^2$) this is Girard's theorem,

$$\mathrm{Area}(T) = a^2 \, (\eta_1 + \eta_2 + \eta_3 - \pi), \tag{T2.121}$$

so spherical triangles always have angle sum larger than π. On the hyperbolic plane ($K = -1/a^2$) the sum is smaller: the *angle deficit* $\pi - \sum \eta_k$ is Area/a^2. Since the deficit cannot exceed π, no hyperbolic triangle has area larger than πa^2, however long its sides; the "ideal triangle" with all three vertices at infinity and all angles zero has exactly that area.

Worked example T2.42: The octant and the lune

The octant. The triangle on the unit sphere with vertices at the north pole and at the equator points of longitude 0 and $\pi/2$ has three right angles (Figure T2.25). Its area is one eighth of the sphere, $4\pi/8 = \pi/2$, and (T2.121) gives $\frac{\pi}{2} + \frac{\pi}{2} + \frac{\pi}{2} - \pi = \frac{\pi}{2}$. It checks.
A thinner triangle. Keep the north pole and put the two equator vertices at longitudes 0 and β. The angles are β (at the pole) and $\frac{\pi}{2}, \frac{\pi}{2}$. Directly:

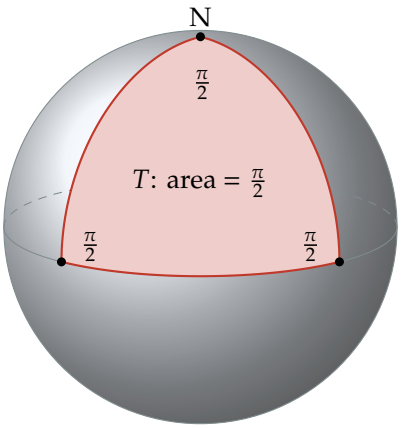
$$\mathrm{Area} \overset{(1)}{=} \int_0^{\beta} \int_0^{\pi/2} \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\phi \overset{(2)}{=} \beta \big[-\cos\theta \big]_0^{\pi/2} = \beta,$$

(1) The area form of the unit sphere over the region $0 \le \theta \le \frac{\pi}{2}, 0 \le \phi \le \beta$.
(2) The ϕ integral gives β.

and (T2.121) gives $\beta + \frac{\pi}{2} + \frac{\pi}{2} - \pi = \beta$.
Girard's own argument, without calculus. Two great circles meeting at angle η cut out a **lune** of area 2η (it is the fraction $\eta/2\pi$ of the total area 4π). The three great circles through the sides of a triangle T make, at each vertex, a lune and its antipodal copy, of total area $2 \cdot 2\eta_k$. The six lunes cover the sphere, covering T and its antipodal copy T' three times each and everything else once. So

$$\sum_{k=1}^{3} 4\eta_k = 4\pi + 2\,\mathrm{Area}(T) + 2\,\mathrm{Area}(T') = 4\pi + 4\,\mathrm{Area}(T),$$

which is (T2.121) with $a = 1$. Two completely different arguments give the same answer, one using only areas and the other only the connection.



angle sum $= \frac{3\pi}{2}$
excess $= \frac{\pi}{2} = \int_T K\,\mathrm{d}A$

Figure T2.25: A geodesic triangle on the unit sphere with three right angles: two meridians from the north pole and a quarter of the equator. Its angle sum exceeds π by $\pi/2$, which is exactly its area, one eighth of 4π. A vector parallel transported around its edge comes back rotated by $\pi/2$.

Physical picture: measuring the curvature of space with triangles

Gauss is said to have measured the angles of the triangle formed by three mountain peaks near Göttingen, looking for an excess. Modern cosmology does the same on a vastly larger scale: the angular size of the hot and cold spots in the cosmic microwave background is a triangle whose apex is at us and whose base is a known physical length at the last scattering surface. The observed size agrees with a Euclidean angle sum to about one percent, which is the statement that the universe is spatially flat.

Try it: carrying a spear around the octant

Stand at the north pole holding a spear pointing along the meridian $\phi = \pi/2$. Walk down the meridian $\phi = 0$ to the equator, keeping the spear parallel (never rotating it relative to your path, which is a geodesic). Walk east along the equator to $\phi = \pi/2$, then back up to the pole. Which way does the spear point? *Answer:* it has turned by $\pi/2$, the enclosed curvature. Along a geodesic the angle between the spear and the path is constant; the path turns by $\varepsilon_k = \pi/2$ at each of the three corners, a total of $3\pi/2$, while a full circuit of the heading would need 2π: the missing $\pi/2$ is the holonomy.

T2.8.4 Global Gauss–Bonnet

Now let Σ be a closed, oriented surface. Chop it into polygons, each a small topological disc with a smooth frame on it, meeting edge to edge: V vertices, E edges, F faces. Apply local Gauss–Bonnet to every face f and add up [Dray (18.72)–(18.76)]. Face f has n_f corners with interior angles $\eta_{f,k}$, and

$$\int_f K\,\mathrm{d}A + \oint_{\partial f} \kappa_g\,\mathrm{d}s + \sum_{k=1}^{n_f} (\pi - \eta_{f,k}) = 2\pi. \tag{T2.122}$$

Sum (T2.122) over all faces, term by term:

$$\begin{aligned} 2\pi F &\overset{(1)}{=} \sum_f \int_f K\,\mathrm{d}A + \sum_f \oint_{\partial f} \kappa_g\,\mathrm{d}s + \pi \sum_f n_f - \sum_{f,k} \eta_{f,k} \\ &\overset{(2)}{=} \int_\Sigma K\,\mathrm{d}A + \sum_f \oint_{\partial f} \kappa_g\,\mathrm{d}s + \pi \sum_f n_f - \sum_{f,k} \eta_{f,k} \end{aligned}$$

$$\overset{(3)}{=} \int_\Sigma K\,\mathrm{d}A + 0 + \pi \sum_f n_f - \sum_{f,k} \eta_{f,k}$$
$$\overset{(4)}{=} \int_\Sigma K\,\mathrm{d}A + 2\pi E - \sum_{f,k} \eta_{f,k}$$
$$\overset{(5)}{=} \int_\Sigma K\,\mathrm{d}A + 2\pi E - 2\pi V. \tag{T2.123}$$

(1) The right side of (T2.122) is 2π for each of the F faces; on the left, split the corner sum.
(2) The faces tile Σ, so the curvature integrals add up to the integral over Σ.
(3) Every edge borders exactly two faces, and their boundary orientations traverse it in opposite directions. Reversing the direction turns a left bend into a right bend, so $\kappa_g\,\mathrm{d}s$ is a 1-form on the edge whose two integrals cancel. All edge terms cancel in pairs (Figure T2.26).
(4) Count corner-edge incidences: each face has as many corners as edges, n_f, and every edge belongs to two faces, so $\sum_f n_f = 2E$.
(5) The interior angles of all faces meeting at a vertex fill the full angle 2π around it, so the sum of all interior angles is $2\pi V$.

Rearranging, we have proved:

Key idea: Gauss–Bonnet theorem

For a closed oriented surface Σ with any Riemannian metric,

$$\boxed{\int_\Sigma K\,\mathrm{d}A = 2\pi\,(V - E + F) = 2\pi\,\chi(\Sigma)} \tag{T2.124}$$

[Dray (18.76)]. The left side needs a metric and two derivatives of it. The right side needs only a count of cells, and by Section T2.4.7 it equals $2\pi(b_0 - b_1 + b_2)$, a statement about closed and exact forms.

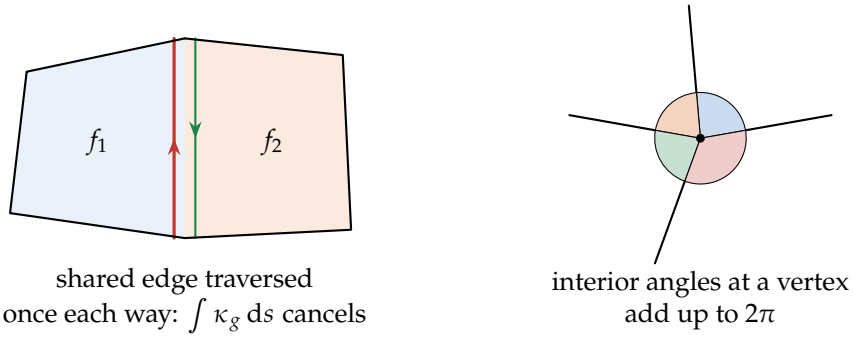


Figure T2.26: The two cancellations behind global Gauss–Bonnet. Left: an edge shared by two faces is traversed in opposite directions by their boundary orientations, so the geodesic-curvature integrals cancel. Right: the interior angles of all the faces meeting at a vertex fill out a full turn. What survives the sum is $2\pi(V - E + F)$.

Two consequences for free.

- *χ does not depend on the decomposition.* Fix a metric. The left side of (T2.124) does not know about the decomposition, so every decomposition gives the same $V - E + F$. (Dray derives

(T2.124) with quadrilaterals only; the derivation above works for arbitrary polygons, which settles his remark that χ does not depend on the shape of the faces.)

- $\int K\,\mathrm{d}A$ *does not depend on the metric.* The right side does not know about the metric. Deform the surface however you like, stretch it, dent it: the total curvature stays fixed.

Worked example T2.43: Sphere, torus, and surfaces of higher genus

Sphere. The cube decomposition gives $\chi = 8 - 12 + 6 = 2$, and the round sphere gives $\int K\,\mathrm{d}A = 4\pi = 2\pi \cdot 2$. Any other metric on S^2 (an ellipsoid, a potato, a dented ping-pong ball) also has total curvature 4π. In particular *no metric on the sphere has $K \le 0$ everywhere*.
Torus. The square with glued sides gives $\chi = 1 - 2 + 1 = 0$, and the torus of revolution gave $\int K\,\mathrm{d}A = 0$ in (T2.92) [Dray (18.89)]. So *no torus has $K > 0$ everywhere*: the positively curved outer half must always be compensated by negative curvature somewhere. The flat torus has $K = 0$ everywhere. It cannot be built in $\mathbb{R}^3$, but it exists in $\mathbb{R}^4$ as $(r\cos\theta, r\sin\theta, s\cos\phi, s\sin\phi)$ [Dray (18.90)–(18.93)]: its line element is $\mathrm{d}s^2 = r^2\mathrm{d}\theta^2 + s^2\mathrm{d}\phi^2$, the coframe $r\,\mathrm{d}\theta$, $s\,\mathrm{d}\phi$ is closed, so $\omega^1{}_2 = 0$ and $K = 0$, consistent with $\chi = 0$. It is the "wrap-around" screen of a video game.
Genus g. Each handle lowers χ by 2 (Example T2.18, [Dray Fig. 18.10]), so $\chi(\Sigma_g) = 2 - 2g$ and

$$\int_{\Sigma_g} K\,\mathrm{d}A = 4\pi(1 - g).$$

For $g \ge 2$ the total curvature is negative. By the uniformisation theorem quoted by Nash (Section T2.6, Reading Nash box), every such surface carries a metric with constant $K = -1$, and then Gauss–Bonnet fixes its area: Area $= -\int K\,\mathrm{d}A = 4\pi(g - 1)$. A hyperbolic surface of genus 2 has area 4π, whatever its shape.

surface	$V - E + F$	χ	$\int K\,\mathrm{d}A$	model metric	K
sphere S^2	$8 - 12 + 6$	2	4π	round	$+1$
torus T^2	$1 - 2 + 1$	0	0	flat (in $\mathbb{R}^4$)	0
genus $g \ge 2$	$1 - 2g + 1$	$2 - 2g$	$4\pi(1 - g)$	hyperbolic	-1
disc (with boundary)	$1 - 1 + 1$	1	$2\pi - \oint \kappa_g\,\mathrm{d}s$	flat or cap	any
annulus, cylinder	$2 - 3 + 1$	0	$-\oint \kappa_g\,\mathrm{d}s$	flat	0

(The genus-g count uses the standard $4g$-gon with its sides glued in pairs: one vertex, $2g$ edges, one face. The disc is one vertex on its boundary circle, one edge, one face. The annulus has a vertex on each boundary circle, the two circles and one radial cut as edges, and one face.)

T2.8.5 Gauss–Bonnet with boundary

If Σ has a boundary, redo the sum (T2.123) with a decomposition in which the boundary is made of edges. The interior edges still cancel in pairs, but the boundary edges belong to only one face and survive. The vertex count changes too: at a boundary vertex the interior angles add up to π (a straight angle) if the boundary is smooth there, or to η at a genuine corner. Redoing the bookkeeping (try it) gives

$$\boxed{\int_\Sigma K\,\mathrm{d}A + \oint_{\partial\Sigma} \kappa_g\,\mathrm{d}s + \sum_{\text{corners}} \varepsilon_k = 2\pi\,\chi(\Sigma)} \tag{T2.125}$$

For a disc, $\chi = 1$, and this is local Gauss–Bonnet (T2.119).

Worked example T2.44: Annulus, cylinder, and a test tube

A flat annulus $\rho_1 \le r \le \rho_2$ in the plane, $\chi = 0$. $K = 0$. The boundary orientation (region on the left) goes counterclockwise around the outer circle and *clockwise* around the inner circle. The outer circle bends left with $\kappa_g = 1/\rho_2$; the inner circle, traversed clockwise, bends right with $\kappa_g = -1/\rho_1$:

$$\oint_{\partial\Sigma} \kappa_g \, \mathrm{d}s = \frac{2\pi\rho_2}{\rho_2} - \frac{2\pi\rho_1}{\rho_1} = 0 = 2\pi\chi.$$

A finite cylinder. $K = 0$, and the boundary circles are geodesics (the cylinder is a rolled-up strip and the circles are the rolled-up straight edges). So both sides of (T2.125) vanish.
A test tube: a hemisphere of radius a with a cylinder of radius a attached along the equator, cut off at some height. $\chi = 1$ (it is a disc). The cylinder and the boundary contribute nothing ($K = 0$, $\kappa_g = 0$), and the hemisphere contributes $\frac{1}{a^2} \cdot 2\pi a^2 = 2\pi$. The boundary term is absent because the metric near the boundary is a product, circle × interval. That is Nash's (4.149).

Reading Nash (p. 129, (4.148)–(4.149))

Nash quotes Gauss–Bonnet for a surface with boundary as

$$\int_X \frac{R}{2\pi} + \int_{\partial X} \frac{R_g}{2\pi} = \chi(X)$$

[Nash p. 129, (4.148)], "where R is the curvature of X and R_g is the geodesic curvature of ∂X in X". In our notation $R = K\,\mathrm{d}A$ (a 2-form) and $R_g = \kappa_g\,\mathrm{d}s$ (a 1-form on the boundary); his boundary is smooth, so there are no corner terms. This is (T2.125) divided by 2π. He then notes that if X is isometric to $\partial X \times \mathbb{R}^+$ near its boundary the boundary integral vanishes and $\int_X R/2\pi = \chi(X)$ [Nash (4.149)]. Why: a product metric near the boundary is $\mathrm{d}s^2 = \mathrm{d}u^2 + \mathrm{d}t^2$ with u the arc length along the boundary and t the distance from it; the coframe $\mathrm{d}u, \mathrm{d}t$ is closed, so $\omega^1{}_2 = 0$ there, the boundary tangent is the constant frame vector e_1, and (T2.105) gives $\kappa_g = 0$. The test tube above is an example. Nash uses this as the warm-up for the Atiyah–Patodi–Singer index theorem, where the boundary correction is not a local integral but the spectral η-invariant.

T2.8.6 Why the metric drops out: an exact form on a closed surface

The polygon argument proves metric independence by comparison with a count. Here is a direct proof that uses nothing but the chapter's favourite fact, $\int_M \mathrm{d}\beta = 0$ when $\partial M = \varnothing$. Change the metric by a **conformal factor**: $g' = e^{2u} g$ for a smooth function u on Σ. (In two dimensions this is no loss of generality: every metric on a closed surface is, after a diffeomorphism, conformal to any other one of the constant curvature metrics listed above, by the uniformisation theorem.) The new coframe is $\sigma'^i = e^u \sigma^i$. Write $\mathrm{d}u = u_1\sigma^1 + u_2\sigma^2$, and recall the Hodge star $\star\sigma^1 = \sigma^2$, $\star\sigma^2 = -\sigma^1$, so $\star\mathrm{d}u = u_1\sigma^2 - u_2\sigma^1$. Guess $\omega'^1{}_2 = \omega^1{}_2 - \star\mathrm{d}u$ and check the first structure equation for $i = 1$:

$$\mathrm{d}\sigma'^1 \overset{(1)}{=} e^u\,\mathrm{d}u \wedge \sigma^1 + e^u\,\mathrm{d}\sigma^1 \overset{(2)}{=} e^u\left(-u_2\,\sigma^1 \wedge \sigma^2 - \omega^1{}_2 \wedge \sigma^2\right)$$
$$\overset{(3)}{=} -e^u\left(\omega^1{}_2 + u_2\sigma^1 - u_1\sigma^2\right) \wedge \sigma^2 \overset{(4)}{=} -\omega'^1{}_2 \wedge \sigma'^2.$$

(1) Product rule.
(2) $\mathrm{d}u \wedge \sigma^1 = u_2\,\sigma^2 \wedge \sigma^1 = -u_2\,\sigma^1 \wedge \sigma^2$, and the old structure equation $\mathrm{d}\sigma^1 = -\omega^1{}_2 \wedge \sigma^2$.
(3) Factor out $\wedge\sigma^2$; the added $-u_1\sigma^2 \wedge \sigma^2$ is zero.
(4) $\omega^1{}_2 + u_2\sigma^1 - u_1\sigma^2 = \omega^1{}_2 - \star\mathrm{d}u$ is the guess, and $e^u\sigma^2 = \sigma'^2$.

The $i = 2$ equation, $\mathrm{d}\sigma'^2 = \omega'^1{}_2 \wedge \sigma'^1$, checks the same way. By uniqueness of the Levi-Civita connection the guess is right:

$$\omega'^1{}_2 = \omega^1{}_2 - \star\mathrm{d}u, \qquad K'\,\mathrm{d}A' = \mathrm{d}\omega'^1{}_2 = K\,\mathrm{d}A - \mathrm{d}\star\mathrm{d}u. \tag{T2.126}$$

Now $\star\mathrm{d}u$ is a *globally defined* 1-form (it is built from the function u and the metric, with no frame choices), so

$$\int_\Sigma K'\,\mathrm{d}A' \overset{(1)}{=} \int_\Sigma K\,\mathrm{d}A - \int_\Sigma \mathrm{d}(\star\mathrm{d}u) \overset{(2)}{=} \int_\Sigma K\,\mathrm{d}A - \int_{\partial\Sigma} \star\mathrm{d}u \overset{(3)}{=} \int_\Sigma K\,\mathrm{d}A. \tag{T2.127}$$

(1) Integrate (T2.126).
(2) Stokes' theorem, legitimate because $\star\mathrm{d}u$ is smooth on all of Σ.
(3) Σ is closed: $\partial\Sigma = \varnothing$, and an integral over the empty set is zero.

The change of total curvature is the integral of an exact form over a closed surface, which vanishes. This is the same mechanism as the monopole computation (T2.34), run in reverse: there the curvature was locally exact but no global potential existed, and the integral was non-zero; here the *change* of curvature has a global potential, and the integral of the change is zero.

Since $\mathrm{d}\star\mathrm{d}u = (\Delta u)\,\mathrm{d}A$ with Δ the Laplacian ($\partial_x^2 + \partial_y^2$ in flat Cartesian coordinates), and $\mathrm{d}A' = e^{2u}\mathrm{d}A$, (T2.126) also gives the useful formula

$$K' = e^{-2u}\big(K - \Delta u\big). \tag{T2.128}$$

Worked example T2.45: The round sphere from the flat plane

Stereographic projection writes the unit sphere as the plane with metric $\mathrm{d}s^2 = \frac{4(\mathrm{d}x^2+\mathrm{d}y^2)}{(1+r^2)^2}$, i.e. $K = 0$ and $u = \ln 2 - \ln(1 + r^2)$. For a radial function $\Delta u = u'' + u'/r$:

$$u' = -\frac{2r}{1+r^2}, \qquad u'' = -\frac{2}{1+r^2} + \frac{4r^2}{(1+r^2)^2},$$
$$\Delta u \overset{(1)}{=} -\frac{4}{1+r^2} + \frac{4r^2}{(1+r^2)^2} \overset{(2)}{=} -\frac{4}{(1+r^2)^2},$$
$$K' \overset{(3)}{=} \frac{(1+r^2)^2}{4}\cdot\frac{4}{(1+r^2)^2} = 1.$$

(1) Add u'' and $u'/r = -2/(1 + r^2)$.
(2) Common denominator: $-4(1 + r^2) + 4r^2 = -4$.
(3) (T2.128) with $K = 0$ and $e^{-2u} = (1 + r^2)^2/4$.

The plane has total curvature 0, the sphere 4π, yet (T2.127) says conformal changes preserve the total. There is no contradiction: the plane is not closed, and $u \to -\infty$ at infinity, so $\star\mathrm{d}u$ is not smooth on the compactified plane (the point at infinity is the north pole). The whole 4π sits in the boundary term at infinity: $-\oint_{r=\Lambda} \star\mathrm{d}u = -2\pi\Lambda\,u'(\Lambda) = 4\pi\Lambda^2/(1 + \Lambda^2) \to 4\pi$.

T2.8.7 The Gauss map: total curvature as a degree

For a surface in $\mathbb{R}^3$ there is a third way to see that total curvature is an integer multiple of something. The **Gauss map** sends each point $p \in \Sigma$ to its unit normal, $\hat{n}(p) \in S^2$. In the adapted frame $\{e_1, e_2, e_3 = \hat{n}\}$ of Section T2.6.5 we had $\mathrm{d}\hat{n} \cdot e_i = -\omega^3{}_i = S_{ij}\sigma^j$, where S is the shape operator. The area form of the unit sphere, evaluated at the point $\hat{n}$, is $\frac{1}{2}\,\hat{n} \cdot (\mathrm{d}\hat{n} \times \mathrm{d}\hat{n})$ (the 2-form that eats two displacements of $\hat{n}$ and returns the area of the parallelogram they span). Pulling it back along the Gauss map:

$$\hat{n}^*\big(\mathrm{d}A_{S^2}\big) \overset{(1)}{=} \tfrac{1}{2}\,\hat{n} \cdot (\mathrm{d}\hat{n} \times \mathrm{d}\hat{n}) \overset{(2)}{=} \tfrac{1}{2}\big[(S_{1j}\sigma^j) \wedge (S_{2k}\sigma^k) - (S_{2k}\sigma^k) \wedge (S_{1j}\sigma^j)\big]$$
$$\overset{(3)}{=} (S_{1j}\sigma^j) \wedge (S_{2k}\sigma^k) \overset{(4)}{=} \det S\ \sigma^1 \wedge \sigma^2 \overset{(5)}{=} K\,\mathrm{d}A. \tag{T2.129}$$

(1) The area form of S^2 written with the embedding; the cross product is taken on the vector parts and the wedge on the form parts.
(2) $\mathrm{d}\hat{n} = (S_{1j}\sigma^j)e_1 + (S_{2k}\sigma^k)e_2$ has no normal part (since $\hat{n} \cdot \hat{n} = 1$), and $\hat{n} \cdot (e_1 \times e_2) = 1 = -\hat{n} \cdot (e_2 \times e_1)$.
(3) Two 1-forms anticommute, so the two terms are equal.
(4) Expand: only $\sigma^1 \wedge \sigma^2$ and $\sigma^2 \wedge \sigma^1$ survive, with coefficient $S_{11}S_{22} - S_{12}S_{21}$.
(5) The Theorema Egregium (T2.95).

The curvature form is the pullback of the area form of the sphere by the Gauss map. (Check on a sphere of radius a: $\hat{n} = \vec{r}/a$ shrinks everything by a, so areas by a^2, and $K = 1/a^2$.) Integrating, $\int_\Sigma K\,\mathrm{d}A$ is the signed area swept out on S^2 by the normal. A map between closed oriented surfaces covers the target an integer number of times, counted with orientation: its **degree**. So

$$\int_\Sigma K\,\mathrm{d}A = 4\pi\,\deg\hat{n}, \qquad \text{and by Gauss–Bonnet} \qquad \deg\hat{n} = \tfrac{1}{2}\chi(\Sigma) = 1 - g. \tag{T2.130}$$

For a convex surface (an ellipsoid, a potato) every direction is the normal at exactly one point, $\deg\hat{n} = 1$, and $\int K\,\mathrm{d}A = 4\pi$ with no calculation at all.

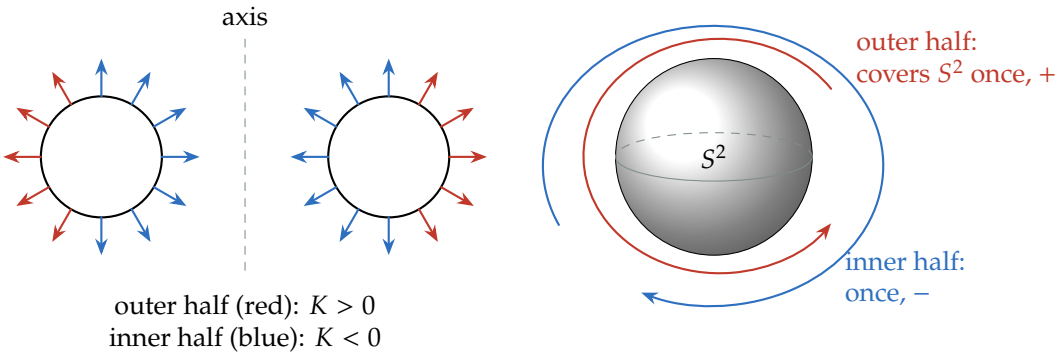


Figure T2.27: The Gauss map of a torus. Every direction in space is the outward normal at exactly one point of the outer half (red) and at exactly one point of the inner half (blue). On the outer half the normal turns the same way as the surface ($K > 0$), on the inner half the opposite way ($K < 0$). The two coverings cancel: $\deg\hat{n} = 0$ and $\int K\,\mathrm{d}A = 0$.

Reading Nash (p. 245, (8.153)–(8.155))

The magnetic charge of a Yang–Mills–Higgs monopole is exactly such a degree. At infinity the Higgs field has unit length, so its direction $\hat{\phi}$ is a map from the sphere at infinity S^2_∞ to the unit sphere in $\mathfrak{su}(2) \cong \mathbb{R}^3$, and Nash writes $k = -\frac{1}{2\pi}\int_{S^2_\infty} \mathrm{tr}\big(\hat{\phi}\,\mathrm{d}\hat{\phi} \wedge \mathrm{d}\hat{\phi}\big)$ [Nash (8.155)].

With the anti-Hermitian basis $T_a = -\frac{i}{2}\sigma_a$ of $\mathfrak{su}(2)$, $\hat\phi = n^a T_a$ with $n \cdot n = 1$:

$$-\frac{1}{2\pi}\operatorname{tr}\big(\hat\phi\,\mathrm{d}\hat\phi \wedge \mathrm{d}\hat\phi\big) \overset{(1)}{=} -\frac{1}{2\pi}\Big(-\frac{i}{2}\Big)^3 \operatorname{tr}(\sigma_a\sigma_b\sigma_c)\, n^a \mathrm{d}n^b \wedge \mathrm{d}n^c \overset{(2)}{=} -\frac{1}{2\pi}\cdot\frac{i}{8}\cdot 2i\,\varepsilon_{abc}\, n^a \mathrm{d}n^b \wedge \mathrm{d}n^c$$
$$\overset{(3)}{=} \frac{1}{8\pi}\,\varepsilon_{abc}\, n^a \mathrm{d}n^b \wedge \mathrm{d}n^c \overset{(4)}{=} \frac{1}{4\pi}\,\hat n^*\big(\mathrm{d}A_{S^2}\big).$$

(1) Insert $\hat\phi = n^a T_a$; the trace is linear.
(2) $(-i/2)^3 = i/8$ and $\operatorname{tr}(\sigma_a\sigma_b\sigma_c) = 2i\,\varepsilon_{abc}$ (from $\sigma_a\sigma_b = \delta_{ab} + i\varepsilon_{abd}\sigma_d$ and $\operatorname{tr}\sigma_d\sigma_c = 2\delta_{dc}$).
(3) $i \cdot i = -1$.
(4) $\frac{1}{2}\varepsilon_{abc} n^a \mathrm{d}n^b \wedge \mathrm{d}n^c = \frac{1}{2}\hat n \cdot (\mathrm{d}\hat n \times \mathrm{d}\hat n)$ is the pulled-back area form, as in (T2.129).

So $k = \frac{1}{4\pi}\int \hat n^*(\mathrm{d}A_{S^2}) = \deg \hat n$: Nash's monopole number is the degree of the Higgs direction field, computed exactly like the Gauss map of a surface. (The unit-length normalisation corresponds to $-2\operatorname{tr}\hat\phi^2 = n \cdot n = 1$.)

T2.8.8 Zeros of vector fields: the Poincaré–Hopf theorem

Why must the wind on the Earth vanish somewhere, while on a doughnut planet it need not? Let $\vec v$ be a vector field on a closed surface with isolated zeros $p_1, \dots, p_m$.

Working definition T2.16 (index of a zero)

Draw a small counterclockwise circle around the zero p [small enough to enclose no other zero] and use a frame that is smooth near p. The **index** of p is the number of times $\vec v$ turns [net number of full counterclockwise turns, a clockwise turn counts -1] as we go once around the circle:

$$\operatorname{ind}_p \vec v = \frac{1}{2\pi}\oint \mathrm{d}\lambda,$$

where λ is the angle of $\vec v$ relative to the frame. It is the winding number (T2.28) of the component map $(v^1, v^2) : \text{circle} \to \mathbb{R}^2 \setminus \{0\}$. *Examples:* a source (x, y), a sink $(-x, -y)$ and a vortex $(-y, x)$ all have index $+1$; a saddle $(x, -y)$ has index -1 (Figure T2.28). In complex notation $v = z^n$ has index n and $v = \bar z^n$ has index $-n$.

Away from its zeros, $\vec v$ defines a frame: $e_1 = \vec v/|\vec v|$, and e_2 the rotation of e_1 by $+90°$. Let ω_v be its connection form. Cut out small discs D_k around the zeros, and let $\Sigma' = \Sigma \setminus \bigcup_k D_k$; its boundary is $\bigcup_k \partial D_k$ with the *opposite* (clockwise) orientation. On each D_k choose a smooth frame $\tilde e$ with connection $\tilde\omega$; on ∂D_k the frame of $\vec v$ is $\tilde e$ rotated by λ, so $\omega_v = \tilde\omega - \mathrm{d}\lambda$ by (T2.82). Then

$$\int_{\Sigma'} K\,\mathrm{d}A \overset{(1)}{=} \int_{\Sigma'} \mathrm{d}\omega_v \overset{(2)}{=} -\sum_k \oint_{\partial D_k} \omega_v \overset{(3)}{=} -\sum_k \oint_{\partial D_k} (\tilde\omega - \mathrm{d}\lambda)$$
$$\overset{(4)}{=} -\sum_k \int_{D_k} K\,\mathrm{d}A + 2\pi \sum_k \operatorname{ind}_{p_k} \vec v. \tag{T2.131}$$

(1) On Σ' the frame of $\vec v$ is smooth, so $K\,\mathrm{d}A = \mathrm{d}\omega_v$ holds on all of Σ'.
(2) Stokes' theorem on Σ', whose boundary circles carry the clockwise orientation (the minus sign).
(3) The change of frame on each circle.
(4) Stokes' theorem for $\tilde\omega$ on the disc D_k (where $\tilde e$ is smooth), and the definition of the index.

Move the D_k integrals to the left: $\int_\Sigma K\,\mathrm{d}A = 2\pi\sum_k \mathrm{ind}_{p_k}\vec{v}$. Combined with Gauss–Bonnet:

Key idea: Poincaré–Hopf theorem

For any vector field with isolated zeros on a closed oriented surface,

$$\sum_{\text{zeros } p} \mathrm{ind}_p\,\vec{v} = \chi(\Sigma). \tag{T2.132}$$

In particular, if $\chi \neq 0$ every vector field has a zero.

Worked example T2.46: Sphere and torus

Sphere, rotation field ∂_ϕ. It vanishes at the two poles. Near the north pole, in local Cartesian coordinates, it is the vortex $(-y, x)$, index +1. Near the south pole, looking from outside (which is the orientation of the sphere there), the rotation appears clockwise, $(y, -x)$: still index +1 (a vortex turned by $-90°$). Total $2 = \chi(S^2)$. *The height function's gradient* on the sphere has a source at the south pole and a sink at the north pole, again $1 + 1 = 2$. This is the hairy ball theorem: you cannot comb a hairy ball flat, and at any moment there is a point on the Earth with no horizontal wind.
Torus. The field ∂_ϕ (around the axis) has no zeros: consistent with $\chi(T^2) = 0$. The gradient of the height of a torus standing on its edge has four zeros: a maximum and a minimum (index +1 each) and two saddles on the rim of the hole (index −1 each). Total 0. Counting critical points of a function with signs is the starting point of Morse theory (Chapter 7).

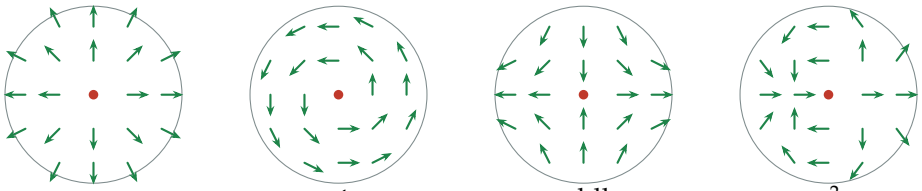


Figure T2.28: Zeros of vector fields and their indices: the number of turns the arrow makes as we go once counterclockwise around the zero. A source and a vortex turn once with the circle (+1), a saddle turns once against it (−1), and the field $z^2 = (x^2 - y^2, 2xy)$ turns twice (+2). On a closed surface the indices of all zeros add up to χ.

T2.8.9 The seed of Chern–Weil theory and of the index theorem

Gauss–Bonnet is the prototype for most of the topology in the rest of these notes. Three readings of it point in three directions.

1. It is a characteristic class (Chern–Weil). Define the **Euler form** of an oriented surface,

$$e := \frac{\Omega^1{}_2}{2\pi} = \frac{K\,\mathrm{d}A}{2\pi}. \tag{T2.133}$$

It is a closed 2-form (every 2-form on a surface is), it is built algebraically from the curvature, it is gauge invariant, and by (T2.127) its cohomology class does not depend on the metric. Gauss–Bonnet says $\int_\Sigma e = \chi(\Sigma)$. Now use the dictionary of Section T2.5.10, which views the

tangent plane as $\mathbb{C}$ and the frame rotation as a U(1) gauge transformation: the anti-Hermitian U(1) potential is $A = -i\,\omega^1{}_2$. Its field strength and first Chern form are

$$c_1 \overset{(1)}{=} \frac{i}{2\pi} F \overset{(2)}{=} \frac{i}{2\pi}\,\mathrm{d}A \overset{(3)}{=} \frac{i}{2\pi}\,(-i)\,\mathrm{d}\omega^1{}_2 \overset{(4)}{=} \frac{\Omega^1{}_2}{2\pi} = e. \tag{T2.134}$$

(1) The convention for the first Chern form used throughout these notes.
(2) For U(1) the term $A \wedge A$ vanishes, so $F = \mathrm{d}A$.
(3) The dictionary $A = -i\,\omega^1{}_2$.
(4) $i \cdot (-i) = 1$, and $\Omega^1{}_2 = \mathrm{d}\omega^1{}_2$ in 2D.

So *the Euler class of a surface is the first Chern class of its tangent bundle*, and Gauss–Bonnet reads $c_1(T\Sigma)[\Sigma] = \chi(\Sigma)$. For the sphere this is the 2 we found as the winding of the transition function between the north and south frames in Example T2.29. Chern–Weil theory (Chapter T3) does the same for any gauge field: invariant polynomials in F (such as $\mathrm{tr}\,F$, $\mathrm{tr}\,F \wedge F$) are closed forms whose integrals are topological, and the monopole charge and the instanton number of Chapter 8 are the next two examples.

2. It generalises to all even dimensions (Chern–Gauss–Bonnet). In dimension $n = 2m$ the Euler form is the **Pfaffian** of the curvature matrix, $e = \mathrm{Pf}(\Omega/2\pi)$, and $\int_M e = \chi(M)$. In four dimensions

$$e = \frac{1}{32\pi^2}\,\varepsilon_{abcd}\,\Omega^{ab} \wedge \Omega^{cd}. \tag{T2.135}$$

Check it on the unit four-sphere, where constant curvature $K = 1$ means $\Omega^{ab} = \sigma^a \wedge \sigma^b$ (the 4D analogue of $\Omega^1{}_2 = \sigma^1 \wedge \sigma^2$):

$$\int_{S^4} e \overset{(1)}{=} \frac{1}{32\pi^2} \int_{S^4} \varepsilon_{abcd}\,\sigma^a \wedge \sigma^b \wedge \sigma^c \wedge \sigma^d \overset{(2)}{=} \frac{24}{32\pi^2} \int_{S^4} \sigma^1 \wedge \sigma^2 \wedge \sigma^3 \wedge \sigma^4 \overset{(3)}{=} \frac{3}{4\pi^2} \cdot \frac{8\pi^2}{3} = 2.$$

(1) Insert the constant-curvature form of Ω into (T2.135).
(2) Each of the $4! = 24$ orderings (a, b, c, d) of $(1, 2, 3, 4)$ contributes $\varepsilon_{abcd} \cdot \varepsilon_{abcd}\,\sigma^1 \wedge \cdots \wedge \sigma^4 = +\sigma^{1234}$.
(3) The volume of the unit S^4 is $8\pi^2/3$.

And $\chi(S^4) = b_0 + b_4 = 2$. In odd dimensions there is no Pfaffian and $\chi = 0$, consistent with Poincaré duality (Section T2.4.7). The same combination $\varepsilon_{abcd} R^{ab}{}_{ef} R^{cd}{}_{gh} \varepsilon^{efgh}$ is the Gauss–Bonnet term of four-dimensional gravity, which is a total derivative and therefore does not change the classical equations of motion.

3. It is an index theorem. By Hodge's theorem (Section T2.4.9) b_p is the number of harmonic p-forms. Group the forms into even ($p = 0, 2$) and odd ($p = 1$) and consider the operator $\mathrm{d} + \mathrm{d}^*$ from even forms to odd forms. Since $(\mathrm{d} + \mathrm{d}^*)^2 = \mathrm{dd}^* + \mathrm{d}^*\mathrm{d} = \Delta$ (the cross terms are $\mathrm{d}^2 = 0$ and $(\mathrm{d}^*)^2 = 0$), its zero modes are exactly the harmonic forms, and

$$\mathrm{index}(\mathrm{d} + \mathrm{d}^*) \overset{(1)}{=} \dim \mathrm{Ker}_{\mathrm{even}} - \dim \mathrm{Ker}_{\mathrm{odd}} \overset{(2)}{=} (b_0 + b_2) - b_1 \overset{(3)}{=} \chi(\Sigma) \overset{(4)}{=} \int_\Sigma e. \tag{T2.136}$$

(1) The index of an operator from even to odd forms: zero modes on the even side minus zero modes of its adjoint on the odd side (the adjoint is again $\mathrm{d} + \mathrm{d}^*$).
(2) Zero modes are harmonic forms, counted by the Betti numbers.
(3) Definition of χ.
(4) Gauss–Bonnet.

The left end is analytic (it counts solutions of a differential equation), the right end is the integral of a local curvature density. This is the Atiyah–Singer index theorem for the de Rham complex. For the Dirac operator the same structure holds with the Euler form replaced by the $\hat{A}$-form (Chapter 4), and in physics that equation computes the chiral anomaly (Chapter 10).

Reading Nash (pp. 104–106, (4.50)–(4.62))

Nash obtains Gauss–Bonnet in the opposite direction. He takes the de Rham complex as an elliptic complex, forms the operator $\mathrm{d}+\mathrm{d}^*$ from even to odd forms, and writes $\operatorname{index}\mathrm{d} = \sum_i (-1)^i \dim H^i_{\text{de Rham}}(X) = \sum (-1)^i b_i = \chi(X)$ [Nash p. 104, (4.52)]. He then evaluates the same index with the general index formula, manipulating Chern characters and the Todd class by the splitting principle, and finds $\operatorname{index}\mathrm{d} = \int_X e(X)$ [Nash (4.53), (4.61)–(4.62)], "the famous Gauss–Bonnet theorem" $\int_X e(X) = \chi(X)$ [Nash (4.54)]. Our derivation used only Stokes' theorem and angle counting, and is special to surfaces; his works in every even dimension but needs the whole machinery of Chapter 4. In two dimensions his $e(X)$ is our $K\,\mathrm{d}A/2\pi$, and (T2.136) is his (4.52)–(4.54) spelled out.

Recap

- Local Gauss–Bonnet: $\int_D K\,\mathrm{d}A + \oint \kappa_g\,\mathrm{d}s + \sum \varepsilon_k = 2\pi$ for a disc; proof: $\kappa_g\,\mathrm{d}s = \mathrm{d}\alpha - \omega^1{}_2$, the turning tangent theorem, and Stokes on $\omega^1{}_2$.
- Geodesic triangles: angle excess $= \int_T K\,\mathrm{d}A$; Girard's area formula on the sphere; angle deficit on the hyperbolic plane.
- Global: $\int_\Sigma K\,\mathrm{d}A = 2\pi\chi$; edges cancel in pairs, angles at vertices sum to 2π. Sphere 4π, torus 0, genus g: $4\pi(1-g)$.
- With boundary (Nash 4.148): $\int K\,\mathrm{d}A + \oint \kappa_g\,\mathrm{d}s = 2\pi\chi$; the boundary term vanishes for a product metric near the boundary (Nash 4.149).
- Metric independence: $K'\mathrm{d}A' = K\,\mathrm{d}A - \mathrm{d}{\star}\mathrm{d}u$; exact form over a closed surface integrates to zero.
- Gauss map: $K\,\mathrm{d}A = \hat{n}^*\mathrm{d}A_{S^2}$, $\deg\hat{n} = 1-g$; the monopole charge is the same kind of degree.
- Poincaré–Hopf: $\sum \mathrm{ind} = \chi$; hairy ball theorem.
- Seed: $e = K\,\mathrm{d}A/2\pi = c_1(T\Sigma)$ (Chern–Weil); Pfaffian in higher dimension ($\int_{S^4} e = 2$); $\operatorname{index}(\mathrm{d}+\mathrm{d}^*) = \chi = \int e$ (index theorem).

T2.9 Applications: equivalence, Lagrangians, spinors and the sphere

We now own a complete toolkit: integration and Stokes' theorem, closed and exact forms, frames, connections, curvature and Gauss–Bonnet. Dray ends his book with a chapter of applications that use the whole kit at once [Dray Ch. 20]. We do the same, with a physicist's eye on what comes next in Nash. Five questions organise the section.

1. Two metrics are written in different coordinates. *Are they the same geometry?* Curvature answers this locally, and holonomy and topology answer it globally (Section T2.9.1).
2. An action is the integral of a top form. *What is the Einstein–Hilbert action in form language,*

and why is gravity in two dimensions pure topology? (Section T2.9.2).

3. Forms of mixed degree can be multiplied with a new product. *Why does this reproduce the Pauli and Dirac matrices, and what is a spinor in this language?* (Sections T2.9.3 and T2.9.4).
4. *Which first-order operator on forms squares to the Laplacian?* The answer, $\mathrm{d}-\delta$, is a Dirac operator in disguise, and its even/odd splitting is the operator whose index is the Euler characteristic (Section T2.9.5).
5. *What does "no holes" really mean, and what happens when we integrate a form from $\mathbb{R}^3$ over a sphere?* Here we correct one statement in Dray and do his closing exercise explicitly (Sections T2.9.6 and T2.9.7).

T2.9.1 The equivalence problem: are two metrics the same?

General relativity is full of rediscovered solutions: a metric is found, published, and later recognised as an old one in new coordinates [Dray §20.1]. The underlying question is:

> *Given two line elements $\mathrm{d}s^2$ and $\mathrm{d}s'^2$, is there a change of coordinates turning one into the other?*

This is the **equivalence problem**. Christoffel introduced connections to attack it.

The flat case is easy. If all the curvature 2-forms vanish, $\Omega^i{}_j = 0$, then by Section T2.6.3 the connection is pure gauge. We can rotate to a frame with $\omega = 0$. In that frame $\mathrm{d}\sigma^i = 0$, so by the Poincaré lemma $\sigma^i = \mathrm{d}X^i$ locally, and $\mathrm{d}s^2 = \sum_i (\mathrm{d}X^i)^2$. The X^i are Cartesian coordinates. So *a metric is locally flat exactly when its curvature vanishes*, and the proof tells us how to find the flat coordinates.

Working definition T2.17 (Flatness test)

Given a line element: (i) choose an orthonormal coframe; (ii) solve $\mathrm{d}\sigma^i = -\omega^i{}_j \wedge \sigma^j$ for the connection; (iii) compute $\Omega = \mathrm{d}\omega + \omega \wedge \omega$. *If* $\Omega = 0$, the metric is locally Euclidean (or Minkowski), and the flat coordinates are found by integrating $\mathrm{d}X^i = \sigma'^i$ in a frame where $\omega' = 0$ [the original frame rotated by a position-dependent angle, which exists locally when $\Omega = 0$]. *If* $\Omega \neq 0$ *anywhere*, no coordinate change can make it flat there.

Worked example T2.47: A flat metric in disguise

Is $\mathrm{d}s^2 = e^{2y}(\mathrm{d}x^2 + \mathrm{d}y^2)$ flat? It does not look it: the coefficients grow exponentially. Take the coframe $\sigma^1 = e^y \mathrm{d}x$, $\sigma^2 = e^y \mathrm{d}y$. Then

$$\mathrm{d}\sigma^1 \overset{(1)}{=} e^y\, \mathrm{d}y \wedge \mathrm{d}x = -e^y\, \mathrm{d}x \wedge \mathrm{d}y, \qquad \mathrm{d}\sigma^2 \overset{(2)}{=} e^y\, \mathrm{d}y \wedge \mathrm{d}y = 0.$$

(1) $\mathrm{d}(e^y \mathrm{d}x) = \mathrm{d}(e^y) \wedge \mathrm{d}x$ and $\mathrm{d}(e^y) = e^y \mathrm{d}y$.
(2) $\mathrm{d}y \wedge \mathrm{d}y = 0$.

Guess $\omega^1{}_2 = c\, \mathrm{d}x$. The structure equations are

$$\mathrm{d}\sigma^1 \overset{(1)}{=} -\omega^1{}_2 \wedge \sigma^2 = -c\, e^y\, \mathrm{d}x \wedge \mathrm{d}y,$$
$$\mathrm{d}\sigma^2 \overset{(2)}{=} -\omega^2{}_1 \wedge \sigma^1 = c\, e^y\, \mathrm{d}x \wedge \mathrm{d}x = 0.$$

(1) First structure equation for $i = 1$.

(2) First structure equation for $i = 2$, with $\omega^2{}_1 = -\omega^1{}_2$.

The second holds for any c, and the first needs $c = 1$. So $\omega^1{}_2 = \mathrm{d}x$ and

$$\Omega^1{}_2 = \mathrm{d}\omega^1{}_2 = \mathrm{d}\mathrm{d}x = 0.$$

The metric is flat. To find the flat coordinates, notice that $\mathrm{d}(e^y) = e^y \mathrm{d}y = \sigma^2$ and $\sigma^1 = e^y \mathrm{d}x$. With $\rho = e^y$ and $\varphi = x$,

$$\mathrm{d}s^2 \overset{(1)}{=} (\sigma^1)^2 + (\sigma^2)^2 \overset{(2)}{=} \rho^2 \,\mathrm{d}\varphi^2 + \mathrm{d}\rho^2.$$

(1) Orthonormal coframe.

(2) $\sigma^1 = \rho \,\mathrm{d}\varphi$ and $\sigma^2 = \mathrm{d}\rho$.

These are polar coordinates, with the frame order swapped: $(e_1, e_2) = (\hat{\varphi}, \hat{\rho})$. (Swapping the frame flips the sign of $\omega^1{}_2$, so $\omega^1{}_2 = +\mathrm{d}\varphi = \mathrm{d}x$ agrees with the polar result $-\mathrm{d}\phi$ of Example T2.22.) The Cartesian coordinates are $X + iY = \rho e^{i\varphi} = e^{y+ix}$. Check: $\left|\mathrm{d}(e^{y+ix})\right|^2 = e^{2y}\,|\mathrm{d}y + i\,\mathrm{d}x|^2 = e^{2y}(\mathrm{d}x^2 + \mathrm{d}y^2)$.

The example has a sting in its tail. The coordinate x runs over all of $\mathbb{R}$, but the angle $\varphi = x$ on the plane is only defined modulo 2π. So the map $(x, y) \mapsto e^{y+ix}$ wraps the (x, y) plane infinitely many times around the punctured plane $\mathbb{R}^2 \setminus \{0\}$. The metric $e^{2y}(\mathrm{d}x^2 + \mathrm{d}y^2)$ is locally flat everywhere, but globally it is the infinite spiral staircase that covers $\mathbb{R}^2 \setminus \{0\}$ (its **universal cover**), not the Euclidean plane. *Curvature decides local equivalence. Global equivalence also needs topology.* The next example shows the difference in its sharpest form.

Worked example T2.48: The cone: flat everywhere except where it isn't

Cut a wedge out of a sheet of paper and glue the edges: you get a cone. In coordinates,

$$\mathrm{d}s^2 = \mathrm{d}r^2 + \beta^2 r^2 \,\mathrm{d}\phi^2, \qquad 0 < \beta \le 1, \quad \phi \sim \phi + 2\pi,$$

where $2\pi\beta$ is the angle of paper that is left. The coframe $\sigma^1 = \mathrm{d}r$, $\sigma^2 = \beta r \,\mathrm{d}\phi$ is the polar one with $\phi \to \beta\phi$, so $\omega^1{}_2 = -\beta \,\mathrm{d}\phi$ and $\Omega^1{}_2 = -\beta \,\mathrm{d}\mathrm{d}\phi = 0$ for $r > 0$. The cone is flat away from the tip. Now apply local Gauss–Bonnet (T2.119) to the disc $D = \{r \le r_0\}$, whose boundary circle has no corners. In the frame $(e_1, e_2) = (\hat{r}, \hat{\phi})$ the tangent of the counterclockwise circle is e_2, so its angle is the constant $\alpha = \pi/2$, and

$$\int_D K \,\mathrm{d}A \overset{(1)}{=} 2\pi - \oint_{r=r_0} \kappa_g \,\mathrm{d}s \overset{(2)}{=} 2\pi - \oint \left(\mathrm{d}\alpha - \omega^1{}_2\right) \overset{(3)}{=} 2\pi - \oint \beta \,\mathrm{d}\phi \overset{(4)}{=} 2\pi(1 - \beta). \quad \text{(T2.137)}$$

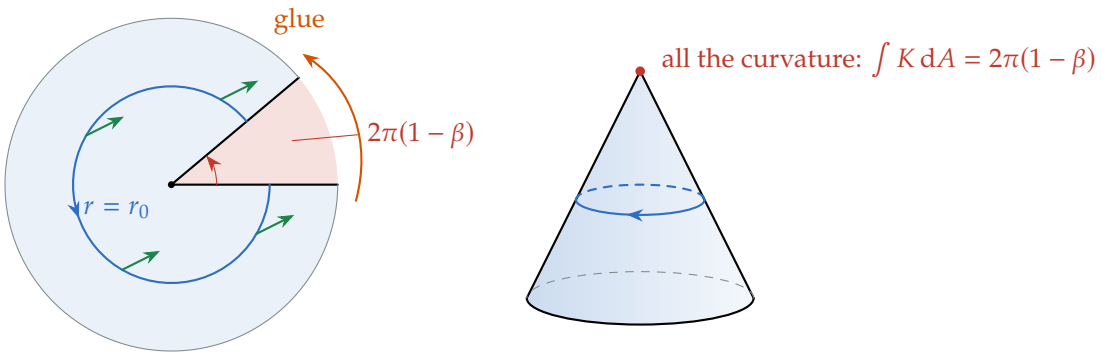
(1) Local Gauss–Bonnet for a disc with smooth boundary: $\sum_k \varepsilon_k = 0$.

(2) The definition (T2.105) of geodesic curvature.

(3) α is constant, so $\mathrm{d}\alpha = 0$; and $-\omega^1{}_2 = \beta \,\mathrm{d}\phi$.

(4) $\oint \mathrm{d}\phi = 2\pi$.

The answer does not depend on r_0. Since $K = 0$ at every $r > 0$, the whole of $2\pi(1 - \beta)$ sits at the tip: $K \,\mathrm{d}A = 2\pi(1 - \beta)\,\delta^2(\text{tip})$. The number $2\pi(1 - \beta)$ is the **deficit angle**, the angle of the wedge we cut out. A vector carried once around the tip by parallel transport comes back rotated by exactly this angle (T2.101). And if we round off the tip by any smooth cap, the cap carries total curvature $2\pi(1 - \beta)$, whatever its shape.



flat paper, wedge removed:
vectors parallel in the plane

$K = 0$ on the sides,
holonomy $2\pi(1-\beta)$ around the tip

Figure T2.29: The cone of Example T2.48. Left: the paper is flat, and a vector field that is parallel in the plane (green) is parallel on the cone. But after gluing the red wedge shut, a vector carried around the circle $r = r_0$ comes back rotated by the missing angle. Right: the cone is flat on its sides, and Gauss–Bonnet puts the whole deficit $2\pi(1-\beta)$ at the tip.

Physical picture: Deficit angles in physics

Descartes' version of Gauss–Bonnet (Section T2.8) is the polyhedral case of (T2.137): each vertex of a polyhedron is a cone point with deficit $2\pi - \sum(\text{face angles})$, and the deficits add up to 4π. The space around a straight cosmic string in general relativity is a cone times a line, with deficit angle $8\pi G\mu$ set by the string tension μ. The orbifold $\mathbb{C}/\mathbb{Z}_N$ of string theory is the cone with $\beta = 1/N$. Regge's lattice gravity builds curved space out of flat simplices and puts all curvature into deficit angles at the hinges.

How hard is the general problem? When the curvature does not vanish, comparing two metrics means comparing their curvature components *and their covariant derivatives*, up to some finite order, after bringing both lists into a canonical form. How many numbers is that? In an orthonormal frame the $\frac{1}{2}n(n-1)$ curvature 2-forms Ω_{ij} each have $N = \frac{1}{2}n(n-1)$ components, giving R_{ijkl} antisymmetric in ij and in kl. Counting,

$$\#R_{ijkl} \overset{(1)}{=} \tfrac{1}{2}N(N+1) - \binom{n}{4} \overset{(2)}{=} \frac{n^2(n^2-1)}{12}, \qquad \text{e.g. } n=4:\ \tfrac{1}{2}\cdot 6\cdot 7 - 1 = 20. \tag{T2.138}$$

(1) The pair symmetry $R_{ijkl} = R_{klij}$ makes R a symmetric $N \times N$ matrix, with $\frac{1}{2}N(N+1)$ entries. The first Bianchi identity (T2.98), $\Omega^i{}_j \wedge \sigma^j = 0$, then removes one more condition for each set of four distinct indices (the totally antisymmetric part $R_{[ijkl]}$ must vanish).

(2) Algebra; the values are $1, 6, 20$ for $n = 2, 3, 4$.

Each derivative adds n more index values. Dray quotes the practical outcome for four-dimensional spacetimes [Dray §20.1], summarised in Table T2.1.

T2.9.2 Lagrangians are top forms: gravity in two dimensions is topology

A Lagrangian density is an n-form, and the action is its integral [Dray §20.2]. We varied the scalar-field and Maxwell actions in Section T2.2.8. Dray lists a third Lagrangian, for gravity [Dray (20.18)]:

$$L_{\mathrm{EH}} = \Omega_{ij} \wedge \star(\sigma^i \wedge \sigma^j), \tag{T2.139}$$

summed over i, j, with the frame index of Ω lowered ($\Omega_{ij} = \Omega^i{}_j$ in Euclidean signature). Let us check that this is the familiar Einstein–Hilbert density $R\,\mathrm{d}V$. Write $\Omega_{ij} = \frac{1}{2}R_{ijkl}\,\sigma^k \wedge \sigma^l$ (T2.96)

Method	highest derivative needed	independent components
naive coordinate components	20	79 310
orthonormal frame (Cartan)	10	8 690
spinor (two-component) formalism	7	3 156
every known exact solution so far	3	430

Table T2.1: The cost of the equivalence problem in four dimensions, as quoted by Dray. Orthonormal frames, the method of this chapter, cut the work by a factor of about ten. Spinors, the "square roots of vectors" of Section T2.9.4, cut it again.

and use $\alpha \wedge \star\beta = \langle \alpha, \beta \rangle \, \mathrm{d}V$ with $\langle \sigma^k \wedge \sigma^l, \sigma^i \wedge \sigma^j \rangle = \delta^{ki}\delta^{lj} - \delta^{kj}\delta^{li}$:

$$\begin{aligned} L_{\mathrm{EH}} &\overset{(1)}{=} \tfrac{1}{2} R_{ijkl}\, \sigma^k \wedge \sigma^l \wedge \star(\sigma^i \wedge \sigma^j) \\ &\overset{(2)}{=} \tfrac{1}{2} R_{ijkl} \big(\delta^{ki}\delta^{lj} - \delta^{kj}\delta^{li}\big)\, \mathrm{d}V \\ &\overset{(3)}{=} \tfrac{1}{2} \big(R_{ijij} - R_{ijji}\big)\, \mathrm{d}V \overset{(4)}{=} R_{ijij}\, \mathrm{d}V \overset{(5)}{=} R\, \mathrm{d}V. \end{aligned} \tag{T2.140}$$

(1) Insert the components of Ω_{ij}.
(2) $\alpha \wedge \star\beta = \langle \alpha, \beta \rangle \, \mathrm{d}V$, and the inner product of basis 2-forms is $+1$ for the same pair in the same order, -1 in the opposite order, 0 otherwise.
(3) Contract the Kronecker deltas.
(4) $R_{ijji} = -R_{ijij}$ (antisymmetry in the last pair).
(5) The scalar curvature is the full contraction $R = R^{ij}{}_{ij}$ (T2.97).

Varying $\int L_{\mathrm{EH}}$ with respect to the coframe gives Einstein's equation in vacuum. That calculation is longer than the scalar one but uses nothing new: the variation of Ω is $\mathrm{d}\,\delta\omega + [\omega, \delta\omega]$, and a Stokes-theorem integration by parts moves the d.

Worked example T2.49: The Einstein–Hilbert action of a closed surface

In two dimensions $\star(\sigma^1 \wedge \sigma^2) = 1$ and $\star(\sigma^2 \wedge \sigma^1) = -1$. Then

$$\int_\Sigma L_{\mathrm{EH}} \overset{(1)}{=} \int_\Sigma \big(\Omega_{12} \cdot 1 + \Omega_{21} \cdot (-1)\big) \overset{(2)}{=} 2 \int_\Sigma \Omega^1{}_2 \overset{(3)}{=} 2 \int_\Sigma K \, \mathrm{d}A \overset{(4)}{=} 4\pi\, \chi(\Sigma). \tag{T2.141}$$

(1) Only $i \neq j$ contributes, since $\Omega_{11} = \Omega_{22} = 0$.
(2) $\Omega_{21} = -\Omega_{12}$ (T2.86).
(3) $\Omega^1{}_2 = K\, \sigma^1 \wedge \sigma^2 = K\, \mathrm{d}A$ (T2.88).
(4) Global Gauss–Bonnet (T2.124).

(Consistently with (T2.140), $R = 2K$ in two dimensions.) The action is the same number for *every* metric on Σ. Its variation with respect to the metric is identically zero, so it produces no field equation at all. *Two-dimensional Einstein gravity has no dynamics; its action is a topological invariant.*

Dictionary: the string coupling counts handles

In string theory the world sheet is a surface Σ and the Polyakov path integral contains exactly this term: $\frac{\lambda}{4\pi} \int_\Sigma R \, \mathrm{d}A = \lambda\, \chi(\Sigma)$. In the sum over world sheets it contributes a weight $e^{-\lambda\chi} = e^{-\lambda(2-2g)} = g_s^{2g-2}$ with $g_s = e^{\lambda}$. So a surface with g handles is suppressed by g_s^{2g-2}: the genus expansion of string perturbation theory is a Gauss–Bonnet expansion (Chapter 9). The four-dimensional analogue, the Pfaffian (T2.135), is likewise a total derivative that can

be added to the gravitational action without changing the field equations.

T2.9.3 The Clifford product of forms

Quantum mechanics hands us objects that the wedge product cannot produce: the Pauli matrices, which square to one, and the Dirac matrices. Dray shows that differential forms already contain them, if we allow ourselves to add forms of different degrees and to multiply them with a new product [Dray §20.3].

Working definition T2.18 (Clifford product)

For functions f and 1-forms α, β define [Dray (20.19)–(20.20)]

$$f \vee \alpha = f\alpha, \qquad \alpha \vee \beta = \alpha \wedge \beta + g(\alpha, \beta), \tag{T2.142}$$

[$g(\alpha,\beta)$ is the metric's dot product of 1-forms, $g(\sigma^i, \sigma^j) = \delta^{ij}$ for an orthonormal Euclidean coframe] and extend to sums of forms of all degrees by requiring the product to be associative and distributive. The symbol $\vee$ is read "vee". The space of mixed-degree forms with this product is the **Clifford algebra** $\mathrm{Cl}(V)$ of the cotangent space V [the 1-forms at one point].

How to compute: in an orthonormal coframe $\{e^i\}$, a product of basis 1-forms is simplified by two rules: distinct ones anticommute and combine into a wedge, $e^i \vee e^j = e^i \wedge e^j = -e^j \vee e^i$ for $i \neq j$; a repeated one cancels in pairs, $e^i \vee e^i = g(e^i, e^i) = \pm 1$ (always +1 in Euclidean signature).

The two rules follow from (T2.142): for $i \neq j$ the metric term $g(e^i, e^j)$ is zero, and for $i = j$ the wedge $e^i \wedge e^i$ is zero. The symmetric and antisymmetric parts of $\vee$ are

$$\alpha \vee \beta + \beta \vee \alpha = 2g(\alpha,\beta), \qquad \alpha \vee \beta - \beta \vee \alpha = 2\,\alpha \wedge \beta. \tag{T2.143}$$

The first of these is the defining relation of the Pauli and Dirac matrices, $\{\gamma^i, \gamma^j\} = 2g^{ij}$. That is the whole secret.

Worked example T2.50: Dray's triple product

Dray states [Dray (20.21)]

$$\alpha \vee \beta \vee \gamma = \alpha \wedge \beta \wedge \gamma + g(\alpha,\beta)\,\gamma - g(\alpha,\gamma)\,\beta + g(\beta,\gamma)\,\alpha. \tag{T2.144}$$

Both sides are linear in each of α, β, γ, so it is enough to check them on orthonormal basis 1-forms in Euclidean signature. There are five cases, up to renaming.

$$e^1 \vee e^2 \vee e^3 \overset{(1)}{=} e^1 \wedge e^2 \wedge e^3 \qquad \text{RHS: } e^{123} + 0 - 0 + 0,$$

$$e^1 \vee e^1 \vee e^2 \overset{(2)}{=} (e^1 \vee e^1) \vee e^2 = e^2 \qquad \text{RHS: } 0 + e^2 - 0 + 0,$$

$$e^1 \vee e^2 \vee e^1 \overset{(3)}{=} -e^1 \vee e^1 \vee e^2 = -e^2 \qquad \text{RHS: } 0 + 0 - e^2 + 0,$$

$$e^2 \vee e^1 \vee e^1 \overset{(4)}{=} e^2 \vee (e^1 \vee e^1) = e^2 \qquad \text{RHS: } 0 + 0 - 0 + e^2,$$

$$e^1 \vee e^1 \vee e^1 \overset{(5)}{=} (e^1 \vee e^1) \vee e^1 = e^1 \qquad \text{RHS: } 0 + e^1 - e^1 + e^1.$$

(1) Distinct basis forms: the product is the wedge. On the right only the wedge survives, since all inner products vanish.
(2) Associativity and $e^1 \vee e^1 = 1$. On the right only $g(\alpha,\beta)\gamma = g(e^1, e^1)e^2$ survives.

(3) Anticommute the middle e^2 past the last e^1. On the right only $-g(\alpha,\gamma)\beta = -e^2$ survives.
(4) Associativity. On the right only $g(\beta,\gamma)\alpha = e^2$ survives.
(5) Associativity. On the right all three metric terms are e^1, with signs $+,-,+$.

All five agree, so (T2.144) holds. Read it as a rule: *the Clifford product is the wedge product plus all possible contractions*, each contraction carrying the sign needed to bring the contracted pair next to each other.

The same bookkeeping gives the product of a 1-form with a p-form, which we need below:

$$\alpha \vee \omega = \alpha \wedge \omega + \iota_{\alpha^\sharp}\omega, \tag{T2.145}$$

where $\alpha^\sharp$ is the vector dual to α and ι is the interior product (Chapter T1). Check it on basis forms in $\mathbb{R}^2$: $e^1 \vee e^{12} = e^1 \vee e^1 \vee e^2 = e^2$, while $e^1 \wedge e^{12} + \iota_{e_1}(e^1 \wedge e^2) = 0 + e^2$; and $e^2 \vee e^{12} = e^2 \vee e^1 \vee e^2 = -e^1 \vee e^2 \vee e^2 = -e^1$, while $e^2 \wedge e^{12} + \iota_{e_2}(e^1 \wedge e^2) = 0 - e^1$. (Here $e^{12} := e^1 \wedge e^2$.)

Worked example T2.51: The Clifford algebra of the plane is the Pauli algebra

On $\mathbb{R}^2$ with coframe $e^1 = \mathrm{d}x$, $e^2 = \mathrm{d}y$, a general element is

$$\Phi = f + a\,e^1 + b\,e^2 + h\,e^{12},$$

so $\mathrm{Cl}(\mathbb{R}^2)$ is four-dimensional. The full multiplication table is shown in Figure T2.30. The only entry needing thought is

$$e^{12} \vee e^{12} \overset{(1)}{=} e^1 \vee e^2 \vee e^1 \vee e^2 \overset{(2)}{=} -e^1 \vee e^1 \vee e^2 \vee e^2 \overset{(3)}{=} -1.$$

(1) $e^{12} = e^1 \wedge e^2 = e^1 \vee e^2$ (distinct factors).
(2) Anticommute the middle pair.
(3) $e^1 \vee e^1 = e^2 \vee e^2 = 1$.

The area element squares to -1: it is a square root of -1 made of geometry. The even part $\{f + h\,e^{12}\}$ is a copy of the complex numbers, with e^{12} playing i. And e^{12} rotates 1-forms by a right angle: $e^1 \vee e^{12} = e^2$ and $e^2 \vee e^{12} = -e^1$.
Now compare the Pauli matrices. The assignment

$$1 \mapsto \mathbb{1}, \qquad e^1 \mapsto \sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad e^2 \mapsto \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \qquad e^{12} \mapsto \sigma_1\sigma_2 = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix} = i\sigma_3 \tag{T2.146}$$

respects every product: $\sigma_1^2 = \sigma_2^2 = \mathbb{1}$ matches $e^i \vee e^i = 1$; $\sigma_1\sigma_2 = -\sigma_2\sigma_1$ matches anticommutation; and $(i\sigma_3)^2 = -\mathbb{1}$ matches $e^{12} \vee e^{12} = -1$. The four matrices $\mathbb{1}, \sigma_1, \sigma_2, i\sigma_3$ are linearly independent and span all 2×2 complex matrices over $\mathbb{C}$, so $\mathrm{Cl}(\mathbb{R}^2) \otimes \mathbb{C}$ is the algebra of 2×2 complex matrices. *Mixed-degree forms on the plane, multiplied with $\vee$, are 2×2 matrices.*

Worked example T2.52: Three dimensions: Pauli matrices and a central i

On $\mathbb{R}^3$ the Clifford algebra has dimension $1+3+3+1 = 8 = 2^3$, one basis element for each subset of $\{1,2,3\}$. The assignment $e^k \mapsto \sigma_k$ works again, since $\sigma_j\sigma_k + \sigma_k\sigma_j = 2\delta_{jk}$. Then the volume element becomes

$$e^{123} = e^1 \vee e^2 \vee e^3 \;\mapsto\; \sigma_1\sigma_2\sigma_3 \overset{(1)}{=} (i\sigma_3)\sigma_3 \overset{(2)}{=} i\,\mathbb{1}.$$

$\vee$	1	e^1	e^2	e^{12}
1	1	e^1	e^2	e^{12}
e^1	e^1	1	e^{12}	e^2
e^2	e^2	$-e^{12}$	1	$-e^1$
e^{12}	e^{12}	$-e^2$	e^1	-1

Pauli image

$\longrightarrow \mathbb{1}$

$\longrightarrow \sigma_1$

$\longrightarrow \sigma_2$

$\longrightarrow \sigma_1\sigma_2 = i\sigma_3$

Figure T2.30: The multiplication table of $\mathrm{Cl}(\mathbb{R}^2)$: the entry in row A, column B is $A \vee B$. Scalars are white, 1-forms blue, the area element orange. The table is not symmetric: distinct 1-forms anticommute. The red entry says the area element squares to -1. On the right, the Pauli matrices reproduce the table exactly.

(1) $\sigma_1\sigma_2 = i\sigma_3$.
(2) $\sigma_3^2 = \mathbb{1}$.

So in three dimensions the volume element commutes with everything and squares to -1: it is "the" i of quantum mechanics. The bivectors become $e^{23} \mapsto i\sigma_1$, $e^{31} \mapsto i\sigma_2$, $e^{12} \mapsto i\sigma_3$, which are (2×) the anti-Hermitian generators of SU(2): the even part of $\mathrm{Cl}(\mathbb{R}^3)$ is the quaternions, and the rotation generators of spin-$\frac{1}{2}$ are 2-forms. This is the start of the Spin(3) = SU(2) story of Chapter T4.

In four-dimensional spacetime with coframe $\{\sigma^\mu\}$, the same construction with $g(\sigma^\mu, \sigma^\nu) = g^{\mu\nu}$ gives $\sigma^\mu \vee \sigma^\nu + \sigma^\nu \vee \sigma^\mu = 2g^{\mu\nu}$. This is the Dirac algebra, under Dray's identification [Dray (20.22)]

$$\gamma^\mu \longleftrightarrow \sigma^\mu.$$

In rectangular coordinates $\sigma^\mu = \mathrm{d}x^\mu$, so *the gamma matrices are the coordinate differentials*, multiplied with $\vee$. The algebra has dimension $2^4 = 16$, the number of independent 4×4 matrices built from $\mathbb{1}, \gamma^\mu, \gamma^{\mu\nu}, \gamma^{\mu\nu\rho}, \gamma^5$: the familiar $1 + 4 + 6 + 4 + 1$ of the Fierz basis is the $1 + 4 + 6 + 4 + 1$ of forms of degree $0, 1, 2, 3, 4$.

T2.9.4 Spinors live inside the Clifford algebra

If the gamma matrices are forms, what does a gamma matrix act *on*? A spinor is a column vector, but the Clifford algebra acts on itself by left multiplication, and that is too big: in 2D it is a four-component space, while a spinor has two. The fix is to keep only a subspace preserved by left multiplication. "Spinors actually live in certain subspaces of Clifford algebras" [Dray §20.3]. Here is that subspace, written down.

Worked example T2.53: Two-component spinors as forms on the plane

In $\mathrm{Cl}(\mathbb{R}^2)$ take

$$P = \tfrac{1}{2}(1 + e^1).$$

It is a projector:

$$P \vee P \overset{(1)}{=} \tfrac{1}{4}(1 + 2e^1 + e^1 \vee e^1) \overset{(2)}{=} \tfrac{1}{4}(2 + 2e^1) = P.$$

(1) Expand; 1 commutes with everything.
(2) $e^1 \vee e^1 = 1$.

The set $S = \mathrm{Cl}(\mathbb{R}^2) \vee P$ of all left multiples of P is closed under left multiplication, since $A \vee (B \vee P) = (A \vee B) \vee P$. Such a subspace is called a **left ideal**. It is two-dimensional, spanned by

$$u_1 = P = \tfrac{1}{2}(1 + e^1), \qquad u_2 = e^2 \vee P = \tfrac{1}{2}(e^2 - e^{12}),$$

because the other two left multiples are not new: $e^1 \vee P = \tfrac{1}{2}(e^1 + 1) = u_1$ and $e^{12} \vee P = \tfrac{1}{2}(e^{12} + e^1 \vee e^2 \vee e^1) = \tfrac{1}{2}(e^{12} - e^2) = -u_2$. Now let the generators act:

$$e^1 \vee u_1 \overset{(1)}{=} u_1, \qquad e^1 \vee u_2 \overset{(2)}{=} -e^2 \vee e^1 \vee P = -e^2 \vee P = -u_2,$$
$$e^2 \vee u_1 \overset{(3)}{=} u_2, \qquad e^2 \vee u_2 \overset{(4)}{=} e^2 \vee e^2 \vee P = u_1.$$

(1) Computed above.
(2) Anticommute e^1 past e^2, then use $e^1 \vee P = P$.
(3) Definition of u_2.
(4) $e^2 \vee e^2 = 1$.

In the basis (u_1, u_2) the left action of the generators is therefore

$$e^1 \mapsto \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = \sigma_3, \qquad e^2 \mapsto \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \sigma_1. \tag{T2.147}$$

These two matrices square to $\mathbb{1}$ and anticommute, so they satisfy the same algebra as e^1, e^2. So a **spinor** on the plane is a form in the ideal S,

$$\psi = \psi_1 u_1 + \psi_2 u_2 = \tfrac{1}{2}\psi_1 + \tfrac{1}{2}\psi_1\,\mathrm{d}x + \tfrac{1}{2}\psi_2\,\mathrm{d}y - \tfrac{1}{2}\psi_2\,\mathrm{d}x \wedge \mathrm{d}y,$$

with components (ψ_1, ψ_2), and the gamma matrices act on it by the Pauli matrices (T2.147). (These are σ_3, σ_1 rather than σ_1, σ_2. The two sets are related by a unitary change of basis; any two irreducible representations of the Clifford algebra are equivalent.) The complementary projector $P' = \tfrac{1}{2}(1 - e^1)$ gives a second ideal S', and since $P + P' = 1$ every form is uniquely the sum of a piece in S and a piece in S': $\mathrm{Cl}(\mathbb{R}^2) = S \oplus S'$, *four form components are two two-component spinors*.

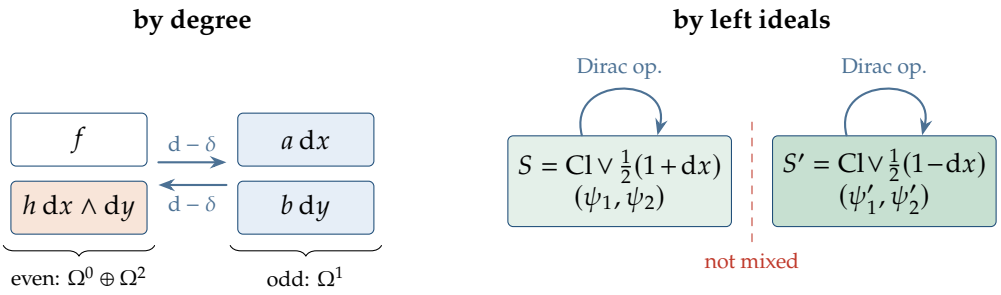


Figure T2.31: Two ways to cut the four components of a form on the plane. Left: by degree parity. The operator $\mathrm{d} - \delta$ always changes degree by one, so it swaps even and odd forms; this is the grading whose index is the Euler characteristic. Right: by left ideals. Left Clifford multiplication preserves each ideal, so $\mathrm{d} - \delta$ acts inside each one as an ordinary two-component Dirac operator.

T2.9.5 The Kähler–Dirac operator: a square root of the Laplacian

Dirac wanted a first-order operator whose square is the wave operator. The gamma matrices were invented for that. We now have gamma matrices made of forms, so we should look for

the corresponding first-order operator on forms. Dray's answer is the **Kähler operator** $\mathrm{d}-\delta$ with [Dray (20.23)]

$$\delta\omega = (-1)^p \star^{-1}\mathrm{d}\star\omega \qquad (\omega \text{ a } p\text{-form}). \tag{T2.148}$$

(This δ has nothing to do with the variational δ of Section T2.9.2.) It lowers the degree by one. In Riemannian signature, using $\star\star = (-1)^{k(n-k)}$ on k-forms, a short sign count shows $\delta = (-1)^{n(p+1)+1}\star\mathrm{d}\star$, which is exactly the adjoint d^* of Sections T2.2.6 and T2.4.9. So $\delta = \mathrm{d}^*$, and $\delta^2 = \pm\star\mathrm{d}\mathrm{d}\star = 0$.

The square. Because $\mathrm{d}^2 = 0$ and $\delta^2 = 0$,

$$(\mathrm{d}-\delta)^2 = \mathrm{d}^2 - \mathrm{d}\delta - \delta\mathrm{d} + \delta^2 = -(\mathrm{d}\delta + \delta\mathrm{d}) \tag{T2.149}$$

[Dray (20.24)]. On a function f we have $\delta f = 0$ (there are no (-1)-forms), so

$$(\mathrm{d}-\delta)^2 f \overset{(1)}{=} -\mathrm{d}\delta f - \delta\mathrm{d}f \overset{(2)}{=} 0 - (-1)^1\star^{-1}\mathrm{d}\star\mathrm{d}f \overset{(3)}{=} \Delta f. \tag{T2.150}$$

(1) (T2.149).
(2) $\delta f = 0$; and $\mathrm{d}f$ is a 1-form, so $p = 1$ in (T2.148).
(3) $\star\mathrm{d}f$ is the flux form of $\vec{\nabla}f$, its d is $(\nabla^2 f)\,\mathrm{d}V$, and $\star^{-1}$ strips off $\mathrm{d}V$: the result is $\nabla^2 f = \Delta f$ [Dray (20.25)].

So $\mathrm{d}-\delta$ is a square root of the Laplacian (and, in Minkowski signature, of the wave operator). That is exactly the job of the Dirac operator.

Watch out: Two Laplacians with opposite signs

Dray's Δ is the physicist's $\nabla^2 = \partial_x^2 + \partial_y^2 + \dots$, whose eigenvalues on a closed manifold are ≤ 0. The Hodge Laplacian $\Delta_{\text{Hodge}} = \mathrm{d}\mathrm{d}^* + \mathrm{d}^*\mathrm{d}$ of Section T2.4.9 and of Nash is $-\nabla^2$ on functions in $\mathbb{R}^n$, with eigenvalues ≥ 0. The two are related by (T2.149): $(\mathrm{d}-\delta)^2 = -\Delta_{\text{Hodge}}$. Always check which sign a book means.

It really is $\gamma^i\partial_i$. On $\mathbb{R}^n$ with Cartesian coordinates, $\mathrm{d}\omega = e^i \wedge \partial_i\omega$ (the derivative acts on the coefficients), and the adjoint is minus the contraction, $\delta\omega = -\iota_{e_i}\partial_i\omega$ (checked below in 2D). Then by (T2.145)

$$e^i \vee \partial_i\omega \overset{(1)}{=} e^i \wedge \partial_i\omega + \iota_{e_i}\partial_i\omega \overset{(2)}{=} \mathrm{d}\omega - \delta\omega. \tag{T2.151}$$

(1) The Clifford product of a 1-form with a p-form (T2.145); e^i is constant, so ∂_i passes through it.
(2) The Cartesian formulas for d and δ.

So $\mathrm{d}-\delta = e^i \vee \partial_i$, which under $e^i \leftrightarrow \gamma^i$ is the Dirac operator $\gamma^i\partial_i$. Its square is $e^i \vee e^j\,\partial_i\partial_j = \frac{1}{2}(e^i\vee e^j + e^j \vee e^i)\,\partial_i\partial_j = \delta^{ij}\partial_i\partial_j = \nabla^2$, by (T2.143), the same computation Dirac did with matrices.

Worked example T2.54: The Kähler–Dirac equation on the plane, component by component

Let $\Phi = f + a\,\mathrm{d}x + b\,\mathrm{d}y + h\,\mathrm{d}x\wedge\mathrm{d}y$ on $\mathbb{R}^2$. With $\star 1 = \mathrm{d}x\wedge\mathrm{d}y$, $\star\mathrm{d}x = \mathrm{d}y$, $\star\mathrm{d}y = -\mathrm{d}x$, $\star(\mathrm{d}x\wedge\mathrm{d}y) = 1$, and $\star^{-1} = -\star$ on 1-forms, $\star^{-1} = \star$ on 0- and 2-forms:

$$\mathrm{d}\Phi \overset{(1)}{=} f_x\,\mathrm{d}x + f_y\,\mathrm{d}y + (b_x - a_y)\,\mathrm{d}x\wedge\mathrm{d}y,$$
$$\delta(a\,\mathrm{d}x + b\,\mathrm{d}y) \overset{(2)}{=} -\star^{-1}\mathrm{d}(a\,\mathrm{d}y - b\,\mathrm{d}x) = -(a_x + b_y),$$

$$\delta(h\,\mathrm{d}x\wedge\mathrm{d}y)\overset{(3)}{=}\star^{-1}\mathrm{d}h=-\star(h_x\,\mathrm{d}x+h_y\,\mathrm{d}y)=h_y\,\mathrm{d}x-h_x\,\mathrm{d}y.$$

(1) $\mathrm{d}f=f_x\mathrm{d}x+f_y\mathrm{d}y$; $\mathrm{d}(a\,\mathrm{d}x+b\,\mathrm{d}y)=(b_x-a_y)\,\mathrm{d}x\wedge\mathrm{d}y$; d of a 2-form on $\mathbb{R}^2$ is zero.
(2) $p=1$: sign -1; $\star$ of the 1-form; $\mathrm{d}(a\,\mathrm{d}y-b\,\mathrm{d}x)=(a_x+b_y)\mathrm{d}x\wedge\mathrm{d}y$; $\star^{-1}$ of that is a_x+b_y. This is minus the divergence, as for d^*.
(3) $p=2$: sign $+1$; $\star(h\,\mathrm{d}x\wedge\mathrm{d}y)=h$; then $\star^{-1}=-\star$ on the 1-form $\mathrm{d}h$.

(Items (2) and (3) confirm $\delta=-\iota_{e_i}\partial_i$: e.g. $-\iota_{\partial_x}\partial_x(h\,\mathrm{d}x\wedge\mathrm{d}y)-\iota_{\partial_y}\partial_y(h\,\mathrm{d}x\wedge\mathrm{d}y)=-h_x\,\mathrm{d}y+h_y\,\mathrm{d}x$.) Collecting by degree,

$$(\mathrm{d}-\delta)\Phi=(a_x+b_y)+(f_x-h_y)\,\mathrm{d}x+(f_y+h_x)\,\mathrm{d}y+(b_x-a_y)\,\mathrm{d}x\wedge\mathrm{d}y. \tag{T2.152}$$

Read off two facts.
(i) Even and odd are swapped. The even data (f,h) appear only in the odd (1-form) part of the result, and the odd data (a,b) only in the even part. Zero modes satisfy two separate systems:

$$f_x=h_y,\quad f_y=-h_x\qquad\text{and}\qquad a_x=-b_y,\quad a_y=b_x.$$

These are the Cauchy–Riemann equations: $f+ih$ and $a-ib$ are holomorphic functions of $z=x+iy$. *The zero modes of the Kähler–Dirac operator on the plane are pairs of holomorphic functions.*
(ii) Each ideal carries a Dirac equation. Insert the spinor of Example T2.53, $f=a=\frac{1}{2}\psi_1$, $b=-h=\frac{1}{2}\psi_2$, into (T2.152):

$$\begin{aligned}(\mathrm{d}-\delta)\psi&\overset{(1)}{=}\tfrac{1}{2}(\partial_x\psi_1+\partial_y\psi_2)(1+\mathrm{d}x)+\tfrac{1}{2}(\partial_y\psi_1-\partial_x\psi_2)(\mathrm{d}y-\mathrm{d}x\wedge\mathrm{d}y)\\&\overset{(2)}{=}(\partial_x\psi_1+\partial_y\psi_2)\,u_1+(\partial_y\psi_1-\partial_x\psi_2)\,u_2.\end{aligned}$$

(1) Substitute: the scalar and $\mathrm{d}x$ coefficients are both $\frac{1}{2}(\partial_x\psi_1+\partial_y\psi_2)$, and the $\mathrm{d}y$ and $\mathrm{d}x\wedge\mathrm{d}y$ coefficients are $\pm\frac{1}{2}(\partial_y\psi_1-\partial_x\psi_2)$.
(2) $u_1=\frac{1}{2}(1+\mathrm{d}x)$ and $u_2=\frac{1}{2}(\mathrm{d}y-\mathrm{d}x\wedge\mathrm{d}y)$.

The result is again in the ideal S, and in components

$$\begin{pmatrix}\psi_1\\\psi_2\end{pmatrix}\longmapsto\begin{pmatrix}\partial_x&\partial_y\\\partial_y&-\partial_x\end{pmatrix}\begin{pmatrix}\psi_1\\\psi_2\end{pmatrix}=\left(\sigma_3\partial_x+\sigma_1\partial_y\right)\begin{pmatrix}\psi_1\\\psi_2\end{pmatrix},$$

the two-dimensional Dirac operator in the representation (T2.147). Its square is $(\partial_x^2+\partial_y^2)\mathbb{1}$, since σ_3 and σ_1 anticommute.

Physical picture: Four tastes on the lattice

In n dimensions forms have 2^n components and a Dirac spinor has $2^{n/2}$ (for n even), so the Kähler–Dirac field is $2^{n/2}$ copies of a Dirac field. In four dimensions that is four copies. Lattice QCD knows this well: the Kogut–Susskind ("staggered") lattice fermion is a discretised Kähler–Dirac field, with one field component per lattice cell of each dimension (sites, links, plaquettes, cubes, hypercubes), and it describes four "tastes" of quark. The forms-to-spinors dictionary of this section is how one untangles them.

The index connection. Nash's operator for the Euler characteristic (Section T2.8.9) is $\mathrm{d}+\mathrm{d}^*$ from even to odd forms. On the plane, (T2.152) shows $\mathrm{d}-\delta$ does the same even→odd swap, and $\mathrm{d}+\delta$ differs from it only by the sign of δ (they have the same kernel on a closed manifold,

the harmonic forms, since $\ker(\mathrm{d}\pm\delta)=\ker\mathrm{d}\cap\ker\delta$ there). So

$$\operatorname{index}\bigl(\mathrm{d}+\mathrm{d}^{*}:\Omega^{\text{even}}\to\Omega^{\text{odd}}\bigr)=\chi(M)=\int_{M}e$$

is an index theorem for the Kähler–Dirac operator with the even/odd grading. Replace the grading by chirality inside one spinor ideal, and the same operator becomes the chiral Dirac operator, whose index is the $\hat{A}$-genus instead of the Euler number (Chapter 4). The Euler characteristic and the Dirac index are two readings of one Clifford module.

Reading Nash (p. 114, (4.99)–(4.101))

Nash introduces the Dirac matrices through their anticommutator, $\gamma_a\gamma_b+\gamma_b\gamma_a=-2\delta_{ab}\mathbb{1}$, and the chirality operator $\gamma_5=i^{n(n+1)/2}\gamma_1\cdots\gamma_n$ with $\gamma_5^2=\mathbb{1}$ [Nash (4.99)]. The minus sign is the anti-Hermitian convention of these notes: Nash's γ_a is i times a Hermitian gamma matrix, and in the language of this section it corresponds to replacing $+g$ by $-g$ in (T2.142). The Dirac operator is $\not{D}=\gamma^a e_a^\mu(x)(\partial_\mu+\omega_\mu)$ with $e_\mu^a e_\nu^b\delta_{ab}=g_{\mu\nu}$ [Nash (4.100)]. Every ingredient is now familiar: e_μ^a is the vielbein of Section T2.5.2, the coframe $\sigma^a=e_\mu^a\mathrm{d}x^\mu$; ω_μ is the connection of Section T2.5 acting on spinors through the rotation generators $\frac{1}{4}\omega_{ab}\gamma^a\gamma^b$ (the 2-forms of Example T2.52); and $\gamma^a e_a^\mu\partial_\mu$ is the flat-space $e^i\vee\partial_i$ of (T2.151). Check the 2D chirality: with $\gamma_a=i\sigma_a$, $\gamma_5=i^3\gamma_1\gamma_2=(-i)(i\sigma_1)(i\sigma_2)=(-i)(-i\sigma_3)=-\sigma_3$, which squares to $\mathbb{1}$ and anticommutes with both γ_a. The splitting $E=E^+\oplus E^-$ by γ_5 [Nash (4.101)] is the chiral analogue of the even/odd splitting of Figure T2.31.

T2.9.6 Topology revisited: what "no holes" means

Dray's topology section [Dray §20.4] restates the facts of Sections T2.3 and T2.4: exact implies closed because $\mathrm{d}^2=0$, which contains $\vec{\nabla}\times\vec{\nabla}f=0$ and $\vec{\nabla}\cdot\vec{\nabla}\times\vec{A}=0$; potentials are unique only up to adding an exact form [Dray (20.29)–(20.30)]; and $\mathrm{d}\phi$ on the circle is closed, has period 2π, and so is not exact, which organises into $H^1_{\mathrm{dR}}(S^1)=\mathbb{R}$ [Dray (20.31)–(20.36)]. We did all of this in detail. Two remarks remain.

Watch out: Naming, and a wrong converse

Naming. "Exact implies closed" is sometimes called the Poincaré lemma (Dray does this), and the converse "closed implies locally exact" is *also* called the Poincaré lemma (we did this in Section T2.3.3). Dray warns about the ambiguity. In these notes *Poincaré lemma* means the converse, on star-shaped regions.
A wrong converse. Dray states that on a simply connected region of $\mathbb{R}^n$ all closed p-forms are exact. That is true for $p=1$ only. Simple connectivity says every loop can be shrunk to a point, so every period of a closed 1-form vanishes and the line-integral construction of the potential works. It says nothing about spheres. The region $\mathbb{R}^3\setminus\{0\}$ is simply connected, since a loop can always slide off the missing point, yet the Coulomb 2-form is closed and not exact (Example T2.6 and Section T2.4.5):

$$E=\frac{x\,\mathrm{d}y\wedge\mathrm{d}z+y\,\mathrm{d}z\wedge\mathrm{d}x+z\,\mathrm{d}x\wedge\mathrm{d}y}{r^3},\qquad \mathrm{d}E=0,\qquad \int_{S^2}E=4\pi\neq0.$$

The correct hypotheses are: contractible (for example star-shaped) regions have *all* closed forms exact; simply connected regions have closed *1-forms* exact.

Which closed-but-not-exact forms a region carries is a fingerprint of the region, and each one is a familiar field. Table T2.2 collects the basic cases in $\mathbb{R}^3$. The most instructive is the last row.

Region	simply conn.?	b_1	b_2	closed, non-exact field
$\mathbb{R}^3$ or a ball	yes	0	0	none: every curl-free field is a gradient, every divergence-free field is a curl
$\mathbb{R}^3 \setminus \{\text{point}\}$	yes	0	1	Coulomb/monopole flux 2-form (flux through an enclosing sphere)
$\mathbb{R}^3 \setminus \{\text{line}\}$	no	1	0	magnetic field of a straight wire, $B \propto \mathrm{d}\phi$ (circulation around the wire)
$\mathbb{R}^3 \setminus \{\text{circle}\}$	no	1	1	field of a current loop (circulation through the ring) and electric field of a charged ring (flux through an enclosing sphere)

Table T2.2: Closed but not exact forms detect holes of different dimensions. A removed point is a hole that a sphere can surround (b_2); a removed line is a hole that a loop can go around (b_1); a removed circle is both.

Worked example T2.55: Ampère's law is the Stokes detector

A steady current I runs along the z-axis. Outside the wire the magnetic field is $\vec{B} = \frac{\mu_0 I}{2\pi\rho}\,\hat{\phi}$ (cylindrical coordinates ρ, ϕ, z). Its 1-form is

$$B \overset{(1)}{=} B_\phi\, \sigma^\phi \overset{(2)}{=} \frac{\mu_0 I}{2\pi\rho}\,\rho\,\mathrm{d}\phi \overset{(3)}{=} \frac{\mu_0 I}{2\pi}\,\mathrm{d}\phi.$$

(1) The 1-form of a vector field is $\vec{B}\cdot\mathrm{d}\vec{r}$, and only the $\hat{\phi}$ component is non-zero.
(2) $\sigma^\phi = \rho\,\mathrm{d}\phi$ is the orthonormal coframe element dual to $\hat{\phi}$.
(3) The factors of ρ cancel.

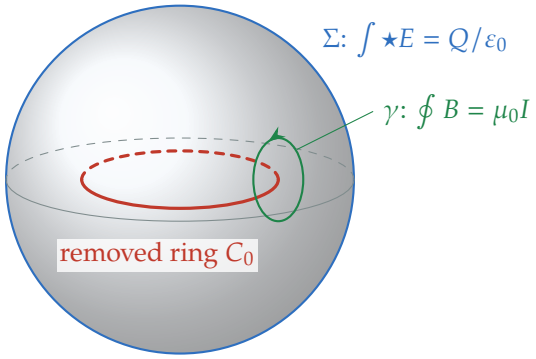
So B is a constant multiple of the angle form: $\mathrm{d}B = 0$ (no current outside the wire, $\vec{\nabla}\times\vec{B} = 0$), and for any loop C

$$\oint_C B = \frac{\mu_0 I}{2\pi}\oint_C \mathrm{d}\phi = \mu_0 I \cdot \mathrm{wind}(C),$$

by (T2.28). That is Ampère's law, with the winding number made explicit. Since $\oint B \neq 0$ around the wire, B is not exact: there is no single-valued magnetic scalar potential Φ_M with $\vec{B} = -\vec{\nabla}\Phi_M$ on $\mathbb{R}^3 \setminus \{\text{wire}\}$. There is a multivalued one, $\Phi_M = -\frac{\mu_0 I}{2\pi}\phi$, which jumps by $\mu_0 I$ across any half-plane $\phi = \text{const}$.

Worked example T2.56: The two holes of a removed circle

Remove the unit circle $C_0 = \{x^2 + y^2 = 1,\ z = 0\}$ from $\mathbb{R}^3$. *Loops:* a small loop γ threading the ring cannot be shrunk to a point in the complement. A current I around C_0 produces a curl-free field outside the wire whose circulation around γ is $\mu_0 I$, by Ampère's law and Stokes' theorem on a small disc pierced by the wire. So $b_1 \geq 1$. *Spheres:* a sphere Σ of radius 2 encloses C_0 and cannot be shrunk to a point either, because it would have to pass through the ring. Spread a charge Q uniformly along C_0. Outside C_0 its electric field is divergence-free, so its flux 2-form $\star E$ is closed on the complement, and Gauss's law gives $\int_\Sigma \star E = Q/\varepsilon_0 \neq 0$. So $b_2 \geq 1$. (In fact $\mathbb{R}^3 \setminus C_0$ can be shrunk onto a sphere with a diameter attached, and $b_1 = b_2 = 1$ exactly.) A thin torus hugging C_0 is homologous to Σ: the region between them has boundary $\Sigma - T$, so by Stokes the flux through the small torus is also Q/ε_0.



$b_1 = 1$: loops that link the ring; $b_2 = 1$: spheres that enclose it

Figure T2.32: The complement of a ring in $\mathbb{R}^3$ has two independent kinds of hole. A loop γ threading the ring (green) detects a closed 1-form, the magnetic field of a current in the ring. A sphere Σ enclosing the ring (blue) detects a closed 2-form, the electric flux of a charge on the ring. Neither loop nor sphere can be shrunk to a point without crossing the ring.

T2.9.7 Integration on the sphere: Dray's exercise, done

Dray closes with an exercise [Dray §20.5]. Pick a 1-form β on $\mathbb{R}^3$ in rectangular coordinates, with $\alpha = \mathrm{d}\beta \neq 0$. Restrict α to the unit sphere and integrate. Then restrict β first and take d of the result. Do the two agree? What must the integral be? We carry this out with two choices of β, and then look at a 2-form that is *not* the restriction of an exact form.

Restricting a form to a surface is the **pullback** ι^* along the inclusion $\iota : S^2 \hookrightarrow \mathbb{R}^3$: substitute $x = \sin\theta\cos\phi$, $y = \sin\theta\sin\phi$, $z = \cos\theta$ and expand, exactly as in Section T2.1.4. The general fact (Chapter T1) is that pullback commutes with d:

$$\iota^*(\mathrm{d}\beta) = \mathrm{d}(\iota^*\beta). \tag{T2.153}$$

$$\begin{array}{ccccc} \beta \in \Omega^1(\mathbb{R}^3) & \xrightarrow{\ \mathrm{d}\ } & \mathrm{d}\beta \in \Omega^2(\mathbb{R}^3) & & \\ \Big\downarrow{\scriptstyle \iota^*} & & \Big\downarrow{\scriptstyle \iota^*} & & \\ \iota^*\beta \in \Omega^1(S^2) & \xrightarrow[\ \mathrm{d}\ ]{} & \iota^*\mathrm{d}\beta = \mathrm{d}\,\iota^*\beta \in \Omega^2(S^2) & \xrightarrow{\int_{S^2}} & 0 \quad (\text{Stokes, } \partial S^2 = \varnothing) \end{array}$$

Figure T2.33: Dray's exercise as a commutative square. Going right then down (differentiate in $\mathbb{R}^3$, then restrict) equals going down then right (restrict, then differentiate on the sphere). Since the bottom-right form is exact on the sphere, its integral over the closed sphere is zero.

Worked example T2.57: $\beta = \frac{1}{2}(x\,\mathrm{d}y - y\,\mathrm{d}x)$: projected area

Here $\mathrm{d}\beta = \frac{1}{2}(\mathrm{d}x\wedge\mathrm{d}y - \mathrm{d}y\wedge\mathrm{d}x) = \mathrm{d}x\wedge\mathrm{d}y$. *Route 1, differentiate then restrict.* On the sphere,

$$\mathrm{d}x = \cos\theta\cos\phi\,\mathrm{d}\theta - \sin\theta\sin\phi\,\mathrm{d}\phi, \qquad \mathrm{d}y = \cos\theta\sin\phi\,\mathrm{d}\theta + \sin\theta\cos\phi\,\mathrm{d}\phi,$$

$$\iota^*(\mathrm{d}x\wedge\mathrm{d}y) \overset{(1)}{=} \big(\cos\theta\cos\phi\cdot\sin\theta\cos\phi + \sin\theta\sin\phi\cdot\cos\theta\sin\phi\big)\,\mathrm{d}\theta\wedge\mathrm{d}\phi$$
$$\overset{(2)}{=} \sin\theta\cos\theta\,\mathrm{d}\theta\wedge\mathrm{d}\phi.$$

(1) Keep the $\mathrm{d}\theta\wedge\mathrm{d}\phi$ terms: (coefficient of $\mathrm{d}\theta$ in $\mathrm{d}x$)(coefficient of $\mathrm{d}\phi$ in $\mathrm{d}y$) minus (coefficient of $\mathrm{d}\phi$ in $\mathrm{d}x$)(coefficient of $\mathrm{d}\theta$ in $\mathrm{d}y$).
(2) $\cos^2\phi + \sin^2\phi = 1$.

Route 2, restrict then differentiate. By Dray's formula $x\,dy - y\,dx = r^2 \sin^2\theta\,d\phi$ [Dray (20.40)] with $r = 1$,

$$d(\iota^*\beta) \overset{(1)}{=} d\left(\tfrac{1}{2}\sin^2\theta\,d\phi\right) \overset{(2)}{=} \sin\theta\cos\theta\,d\theta \wedge d\phi.$$

(1) $\iota^*\beta = \frac{1}{2}\sin^2\theta\,d\phi$.
(2) $d(\frac{1}{2}\sin^2\theta) = \sin\theta\cos\theta\,d\theta$.

The routes agree, as (T2.153) promises. Now integrate over the sphere with the orientation $\sin\theta\,d\theta \wedge d\phi$:

$$\int_{S^2} dx \wedge dy \overset{(1)}{=} \int_0^{2\pi}\int_0^{\pi} \sin\theta\cos\theta\,d\theta\,d\phi \overset{(2)}{=} 2\pi\left[\tfrac{1}{2}\sin^2\theta\right]_0^{\pi} = 0.$$

(1) Drop the wedge; the orientation of $d\theta \wedge d\phi$ agrees with the chosen one.
(2) The ϕ integral gives 2π.

Zero, as Stokes demands. But the hemispheres separately are not zero. Over the northern hemisphere N ($0 \le \theta \le \pi/2$),

$$\int_N dx \wedge dy \overset{(1)}{=} \oint_{\text{equator}} \beta \overset{(2)}{=} \int_0^{2\pi} \tfrac{1}{2}\sin^2\tfrac{\pi}{2}\,d\phi \overset{(3)}{=} \pi,$$

(1) Stokes, with the equator traversed counterclockwise seen from above (the boundary orientation of N).
(2) $\iota^*\beta = \frac{1}{2}\sin^2\theta\,d\phi$ at $\theta = \pi/2$.
(3) Arithmetic.

and the southern hemisphere gives $-\pi$. The geometric meaning: $dx \wedge dy$ restricted to a surface measures the area of its shadow on the xy-plane, with a sign that records whether the outward normal points up or down. Each hemisphere casts a unit-disc shadow of area π, and the two signs cancel (Figure T2.34).

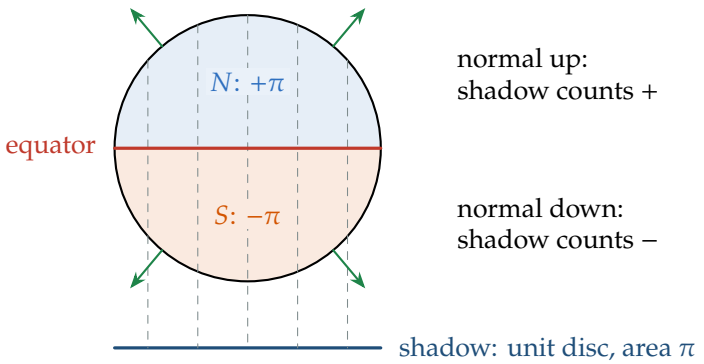


Figure T2.34: The meaning of $\int dx \wedge dy$ over a surface: the signed area of its shadow on the xy-plane. The northern hemisphere (normal up) contributes $+\pi$, the southern (normal down) $-\pi$, and the closed sphere 0. Stokes turns each hemisphere integral into the equator integral of $\beta = \frac{1}{2}(x\,dy - y\,dx)$, which is the area-sweeping form of Kepler's second law.

Worked example T2.58: $\beta = z\,dx$: a less symmetric check

Now $d\beta = dz \wedge dx$. *Route 1.* With $dz = -\sin\theta\,d\theta$,

$$\iota^*(dz \wedge dx) \overset{(1)}{=} (-\sin\theta\,d\theta) \wedge (-\sin\theta\sin\phi\,d\phi) \overset{(2)}{=} \sin^2\theta\sin\phi\,d\theta \wedge d\phi.$$

(1) Only the $d\phi$ part of dx survives the wedge with $d\theta$.
(2) Multiply out.

Route 2. $\iota^*\beta = \cos\theta(\cos\theta\cos\phi\,d\theta - \sin\theta\sin\phi\,d\phi)$, so

$$\begin{aligned} d(\iota^*\beta) &\overset{(1)}{=} d\big(\cos^2\theta\cos\phi\big) \wedge d\theta - d\big(\sin\theta\cos\theta\sin\phi\big) \wedge d\phi \\ &\overset{(2)}{=} -\cos^2\theta\sin\phi\,d\phi \wedge d\theta - \big(\cos^2\theta - \sin^2\theta\big)\sin\phi\,d\theta \wedge d\phi \\ &\overset{(3)}{=} \big(\cos^2\theta - \cos^2\theta + \sin^2\theta\big)\sin\phi\,d\theta \wedge d\phi = \sin^2\theta\sin\phi\,d\theta \wedge d\phi. \end{aligned}$$

(1) $d(f\,d\theta) = df \wedge d\theta$, and similarly for $d\phi$.
(2) In the first term only the ϕ-derivative survives; in the second only the θ-derivative, with $\partial_\theta(\sin\theta\cos\theta) = \cos^2\theta - \sin^2\theta$.
(3) $d\phi \wedge d\theta = -d\theta \wedge d\phi$; collect.

Again the routes agree, and $\int_{S^2}\sin^2\theta\sin\phi\,d\theta\,d\phi = \big(\int_0^\pi \sin^2\theta\,d\theta\big)\big(\int_0^{2\pi}\sin\phi\,d\phi\big) = \frac{\pi}{2}\cdot 0 = 0$. (Geometrically: $dz \wedge dx$ is the signed shadow on the zx-plane, and the halves $y > 0$ and $y < 0$ cancel.)

Worked example T2.59: The area form is a restriction, but not of an exact form

Why does the area form $\omega = \sin\theta\,d\theta \wedge d\phi$ escape this argument? It *is* the restriction of a 2-form on $\mathbb{R}^3$, namely

$$\Omega_0 = x\,dy \wedge dz + y\,dz \wedge dx + z\,dx \wedge dy = \iota_{\vec r}\,(dx \wedge dy \wedge dz),$$

the flux form of the position field $\vec r$ (T2.2). On the unit sphere $\vec r = \hat n$ is the unit normal, so the flux form of $\vec r$ is the area form: $\iota^*\Omega_0 = \omega$. But Ω_0 is not closed, let alone exact:

$$d\Omega_0 \overset{(1)}{=} (\vec\nabla\cdot\vec r)\,dx \wedge dy \wedge dz \overset{(2)}{=} 3\,dV, \qquad \int_{S^2}\omega \overset{(3)}{=} \int_B 3\,dV = 3\cdot\tfrac{4}{3}\pi = 4\pi.$$

(1) d of a flux form is the divergence times the volume form (T2.13).
(2) $\vec\nabla\cdot\vec r = 3$.
(3) Stokes (Gauss) on the unit ball B, whose boundary is S^2.

We can make the extension closed, at a price. Divide by r^3: $E = \Omega_0/r^3$ still restricts to ω on the unit sphere, and

$$dE \overset{(1)}{=} \frac{d\Omega_0}{r^3} + d\big(r^{-3}\big) \wedge \Omega_0 \overset{(2)}{=} \frac{3\,dV}{r^3} - \frac{3}{r^5}(x\,dx + y\,dy + z\,dz) \wedge \Omega_0 \overset{(3)}{=} \frac{3\,dV}{r^3} - \frac{3r^2\,dV}{r^5} = 0.$$

(1) Product rule for d of a function times a form.
(2) $d\Omega_0 = 3\,dV$, and $d(r^{-3}) = -3r^{-4}\,dr = -3r^{-5}(x\,dx + y\,dy + z\,dz)$.
(3) $dx \wedge \Omega_0 = x\,dV$, and similarly for y, z, so the wedge gives $(x^2 + y^2 + z^2)\,dV = r^2\,dV$.

So E is the Coulomb form, closed on $\mathbb{R}^3 \setminus \{0\}$ and singular at the origin. Either way, there is no conflict with Stokes. An extension of ω to all of $\mathbb{R}^3$ cannot be closed (its d must integrate to 4π over the ball), and a closed extension must be singular somewhere inside the sphere. This is the three-dimensional version of the sphere puzzle (T2.31): the non-exactness of ω on S^2 is the charge sitting inside.

Try it: Your own β

Take $\beta = xz\,dy$. Compute $\alpha = d\beta$, restrict it to the unit sphere both ways, and integrate over the sphere and over the northern hemisphere. *Answer:* $\alpha = z\,dx \wedge dy - x\,dy \wedge dz$. On the sphere $\iota^*\beta = \sin\theta\cos\theta\cos\phi\,(\cos\theta\sin\phi\,d\theta + \sin\theta\cos\phi\,d\phi)$. The sphere integral is 0. The hemisphere integral is the equator integral of $\iota^*\beta$ at $\theta = \pi/2$, where $\cos\theta = 0$, so it is also 0. (Check with the shadow picture: $z\,dx \wedge dy$ and $-x\,dy \wedge dz$ each integrate to zero over the northern hemisphere by symmetry in $x \to -x$ and a direct θ integral: $\int_N z\,dx \wedge dy = 2\pi\int_0^{\pi/2}\cos^2\theta\sin\theta\,d\theta = 2\pi/3$ and $\int_N x\,dy \wedge dz = \int_N \sin^3\theta\cos^2\phi\,d\theta\,d\phi = 2\pi/3$, which cancel.)

Reading Nash (where these tools reappear)

Nash never pauses on the material of this section, but he uses all of it. The Clifford algebra and the spin connection are the input to the Dirac operator [Nash p. 114, (4.99)–(4.100)], and the even/odd splitting of forms is his route to the Euler characteristic [Nash p. 104, (4.52)]. The flux of a field through a sphere at infinity is his monopole charge [Nash p. 245, (8.154)], the same Stokes argument as in Example T2.59 with Ω_0/r^3 replaced by $\mathrm{tr}(F\phi)$. Deficit angles and the Euler characteristic of the world sheet return in the string chapter (Chapter 9).

Recap

- *Equivalence:* $\Omega = 0$ iff locally flat; flat coordinates come from integrating a parallel coframe. $e^{2y}(dx^2 + dy^2)$ is flat (polar coordinates with $\rho = e^y$) but globally a covering space. The cone is flat except at its tip, which carries the deficit $2\pi(1-\beta)$. R_{ijkl} has $n^2(n^2-1)/12$ components.
- *Lagrangians:* $\Omega_{ij} \wedge \star(\sigma^i \wedge \sigma^j) = R\,dV$; in 2D its integral is $4\pi\chi$, so 2D gravity is topological and weights string world sheets by g_s^{2g-2}.
- *Clifford product:* $\alpha \vee \beta = \alpha \wedge \beta + g(\alpha,\beta)$; wedge plus all contractions. $\mathrm{Cl}(\mathbb{R}^2)$ is the Pauli algebra with $e^{12} \mapsto i\sigma_3$; in $\mathbb{R}^3$, $e^{123} \mapsto i\mathbb{1}$; in spacetime $\gamma^\mu \leftrightarrow dx^\mu$.
- *Spinors:* a left ideal such as $\mathrm{Cl} \vee \frac{1}{2}(1+e^1)$; in 2D, two components with $e^1 \mapsto \sigma_3$, $e^2 \mapsto \sigma_1$.
- *Kähler–Dirac:* $d - \delta = e^i \vee \partial_i$ squares to ∇^2; on the plane its zero modes are pairs of holomorphic functions; it is $2^{n/2}$ Dirac fields; with the even/odd grading its index is χ.
- *Topology:* simply connected kills H^1, not H^2 ($\mathbb{R}^3 \setminus \{0\}$); Ampère's law is the Stokes detector for $d\phi$; a removed ring has $b_1 = b_2 = 1$.
- *Sphere exercise:* pullback commutes with d; restrictions of exact forms integrate to zero on S^2; $dx \wedge dy$ is signed shadow area; the area form extends to $\mathbb{R}^3$ only as a non-closed form or as the singular Coulomb form.

Chapter summary

The chapter told two stories and then showed they were one. The first story was about *adding up*: forms are the things we can integrate, and Stokes' theorem $\int_R d\alpha = \int_{\partial R}\alpha$ (T2.9) turns the

local statement $d^2 = 0$ into global information. The second story was about *turning*: a frame that is carried across a curved space rotates, the rate of rotation is the connection ω, and the failure of that rotation to be undone around a loop is the curvature $\Omega = d\omega + \omega \wedge \omega$. The stories met in Gauss–Bonnet, $\int_\Sigma K\, dA = 2\pi\chi(\Sigma)$ (T2.124), where a curvature integral turned out to be a count of holes.

If we had to keep a single sentence from the chapter, it would be the owner's favourite: *an exact form integrates to zero over anything without boundary, because the boundary is empty.* That one argument did four different jobs, shown side by side in Figure T2.35. It detected non-exact forms (T2.18). It made periods of cohomology classes well defined (T2.36). It made the total curvature of a closed surface independent of the metric (T2.127). And, as we previewed in §T2.8.9, it is exactly why Chern numbers and instanton numbers are topological in Chapter 4 and Chapter 8.

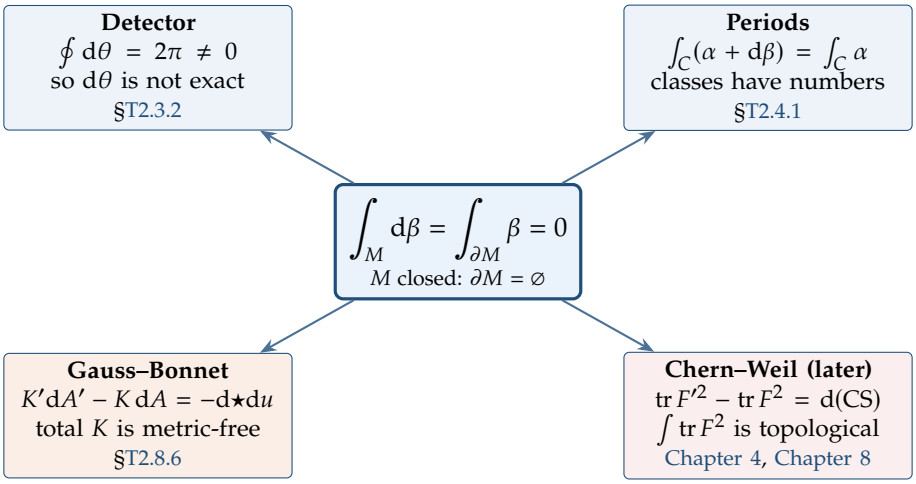


Figure T2.35: One argument, four uses. The hub is Stokes' theorem on a manifold with no boundary. Upper left: a single non-zero loop integral proves a closed form is not exact. Upper right: adding an exact form does not change any period, so a cohomology class carries well-defined numbers. Lower left: a change of metric changes $K\, dA$ by an exact form, so $\int K\, dA$ cannot change. Lower right: the same mechanism, with ω replaced by a gauge potential, makes $\int \mathrm{tr}\, F \wedge F$ topological.

The rest of this summary has five parts: a table of every concept with its operational meaning and where it returns; a catalogue of the spaces we computed; a dictionary between the geometry of this chapter and gauge theory; a list of conventions; and a short self-test.

Every concept, how to use it, and where it returns

The middle column is the *operational* meaning: what you actually do to compute or test the thing. The right column says where in the book the idea comes back. References such as §T2.3.3 point to this chapter.

§T2.1: integrating forms

Concept	Operational meaning	Returns in
p-form as integrand	eats the p edge vectors of a tiny parallelogram and returns its signed p-volume; circulation, flux and volume are $\int F$, $\int \star F$, $\int \star f$ (T2.3)	every action and every topological charge
pull back and drop the wedges	parametrise, substitute $x^\mu(u)$, expand, write $\mathrm{d}u^1 \wedge \cdots \wedge \mathrm{d}u^p \to \mathrm{d}u^1 \cdots \mathrm{d}u^p$; the wedge supplies the Jacobian (T2.6)	all explicit integrals; Chapter 8
orientation	a consistent choice of "positive" ordering of bases; reversing it flips every integral; Möbius band has none	signs of indices, Chapter 4
partition of unity	chop M into chart-sized pieces with weights ρ_a, $\sum \rho_a = 1$, and add (T2.8)	gluing arguments, Chapter 3

§T2.2: Stokes' theorem

Concept	Operational meaning	Returns in
Stokes' theorem	$\int_R \mathrm{d}\alpha = \int_{\partial R} \alpha$; boundary oriented "outward normal first" (T2.9)	everywhere
special cases	FTC ($p = 0$), gradient, Green and curl ($p = 1$), Gauss ($p = 2$ in $\mathbb{R}^3$); Figure T2.7	electromagnetism, Chapter T3
$\partial^2 = \varnothing \leftrightarrow \mathrm{d}^2 = 0$	a boundary has no boundary; dual statement is $\mathrm{dd} = 0$ (T2.14)	chain complexes, Chapter 1
integration by parts	$\int_M \mathrm{d}\alpha \wedge \beta = \int_{\partial M} \alpha \wedge \beta - (-1)^p \int_M \alpha \wedge \mathrm{d}\beta$ (T2.15); gives the adjoint d^*	Hodge theory, Chapter 2
boundary terms	varying $S = \int \mathcal{L}$ gives field equations plus $\int_{\partial M}(\dots)\,\delta\phi$ (T2.16)	anomalies, Chapter 10; TQFT, Chapter 12

§T2.3: closed and exact forms

Concept	Operational meaning	Returns in
closed, exact	closed: compute $\mathrm{d}\alpha$ and get 0; exact: exhibit a global smooth β with $\alpha = \mathrm{d}\beta$; exact $\Rightarrow$ closed (T2.17)	gauge potentials, Chapter T3
Stokes detector	integrate over a cycle without boundary; any non-zero answer proves non-exactness (T2.18)	monopoles and instantons, Chapter 8
Poincaré lemma	on a star-shaped region, $\beta = K\alpha = \int_0^1 t^{p-1} \iota_x \alpha(tx)\,\mathrm{d}t$ solves $\mathrm{d}\beta = \alpha$; $\mathrm{d}K + K\mathrm{d} = 1$ (T2.23)	local gauge potentials; Čech–de Rham, Chapter 3
angle form $\mathrm{d}\theta$	$(x\,\mathrm{d}y - y\,\mathrm{d}x)/r^2$ on $\mathbb{R}^2 \setminus \{0\}$: closed, period 2π, not exact (T2.25)	Aharonov–Bohm; vortices; Chapter 1
winding number	$\frac{1}{2\pi}\oint_C \mathrm{d}\theta \in \mathbb{Z}$; unchanged by deformations that avoid the puncture (T2.28)	degree of maps, π_1, Chapter 1; Chapter 6
sphere puzzle	$\omega = \mathrm{d}(-\cos\theta\,\mathrm{d}\phi)$ fails at the poles; two patches with transition $2\,\mathrm{d}\phi$ give $\int \omega = 4\pi$ (T2.33)	Dirac monopole, Chapter 8

§T2.4: first look at de Rham cohomology

Concept	Operational meaning	Returns in
$H^p_{\mathrm{dR}}(M)$	closed p-forms modulo exact ones (T2.35); label a class by its periods over a basis of cycles	Chapter 1; Chapter 2
H^0 and $H^p(\mathbb{R}^n)$	b_0 = number of components (T2.38); contractible spaces have nothing above degree 0 (T2.39)	homotopy invariance, Chapter 1
Betti numbers	$b_p = \dim H^p_{\mathrm{dR}}$: $S^1\,(1,1)$, $S^2\,(1,0,1)$, $T^2\,(1,2,1)$	Chapter 1; Chapter 7
Euler characteristic	$\chi = \sum_p (-1)^p b_p = V - E + F$ (T2.50); $\chi(\Sigma_g) = 2 - 2g$	index theorem, Chapter 4; strings, Chapter 9
de Rham's theorem	a closed form is exact iff all its periods vanish (T2.52)	Chapter 1
Hodge's theorem	one harmonic representative per class; $b_p = \dim \mathrm{Ker}\,\Delta_p$ (T2.54): zero modes count topology	Chapter 2; Chapter 4

§T2.5: connections

Concept	Operational meaning	Returns in
orthonormal coframe (vielbein)	write $\mathrm{d}\vec{r} = \sigma^i e_i$, i.e. $\mathrm{d}s^2 = \sum_i (\sigma^i)^2$; $g = E^{\mathsf{T}} E$ (T2.58)	spinors on curved space, Chapter T4
connection 1-forms	differentiate the frame: $\mathrm{d}e_j = e_i\,\omega^i{}_j$ (T2.61); then $\mathrm{d}\vec{v} = (\mathrm{d}v^i + \omega^i{}_j v^j) e_i$	gauge potential A, Chapter T3
metric compatibility	$\omega_{ij} = -\omega_{ji}$: ω is $\mathfrak{so}(n)$-valued (T2.66)	structure group of TM, Chapter T3
Cartan's first structure equation	solve $\mathrm{d}\sigma^i + \omega^i{}_j \wedge \sigma^j = 0$ for ω; in 2D read off $\omega^1{}_2$ from $\mathrm{d}\sigma^1, \mathrm{d}\sigma^2$ (T2.69)	spin connection, Chapter T4
Levi-Civita answers	plane (polar) $\omega^1{}_2 = -\mathrm{d}\phi$; sphere $-\cos\theta\,\mathrm{d}\phi$; torus $\sin\theta\,\mathrm{d}\phi$ (T2.70), (T2.71)	examples throughout
Christoffel symbols, Koszul formula	Γ from derivatives of g; frame version via structure constants (T2.75)	geodesics; Chapter 7
frame rotation	$\omega' = g^{-1}\omega g + g^{-1}\mathrm{d}g$; in 2D $\omega^1{}_2 \mapsto \omega^1{}_2 - \mathrm{d}\lambda$ (T2.81)	gauge transformations, Chapter T3

§T2.6: curvature

Concept	Operational meaning	Returns in
curvature 2-form	$\Omega = \mathrm{d}\omega + \omega \wedge \omega$ (T2.84); $\mathrm{d}^2\vec{v} = \Omega\vec{v}$	$F = \mathrm{d}A + A \wedge A$, Chapter T3
Gaussian curvature	in 2D, $\mathrm{d}\omega^1{}_2 = K\,\sigma^1 \wedge \sigma^2$ (T2.87); sphere $1/a^2$, torus (T2.91), hyperbolic $-1/a^2$	Gauss–Bonnet; Chapter 9
flat ⇔ pure gauge	$\Omega = 0$ iff locally $\omega = g^{-1}\mathrm{d}g$ iff a Cartesian frame exists	flat connections, Chapter 12
Theorema Egregium	$K = \kappa_1\kappa_2 = \det S$ (T2.95) yet computable from g alone	intrinsic geometry
Riemann, Ricci, scalar	$\Omega^i{}_j = \frac{1}{2}R^i{}_{jkl}\sigma^k \wedge \sigma^l$ (T2.96); in 2D $R = 2K$	Einstein–Hilbert, Chapter 9
Bianchi identities	$\Omega \wedge \boldsymbol{\sigma} = 0$, $D\Omega = \mathrm{d}\Omega + [\omega, \Omega] = 0$ (T2.98), (T2.99)	$DF = 0$, Chapter T3; Chern–Weil, Chapter 4
holonomy	rotation angle after transport around ∂D equals $\int_D K\,\mathrm{d}A$ (T2.101)	Wilson loops, Chapter T3; Chapter 12

§T2.7: geodesics

Concept	Operational meaning	Returns in
geodesic curvature	$\kappa_g\,\mathrm{d}s = \mathrm{d}\alpha - \omega^1{}_2$ (T2.105): turning of the tangent relative to a transported vector	Gauss–Bonnet with boundary, Chapter 4
geodesic equation	$\ddot{x}^j + \Gamma^j{}_{ik}\dot{x}^i\dot{x}^k = 0$ (T2.109); Euler–Lagrange equation of $E = \int g_{\mu\nu}\dot{x}^\mu\dot{x}^\nu\,\mathrm{d}t$	Morse theory on path space, Chapter 7
Jacobi equation	$J'' + KJ = 0$ (T2.116): $K > 0$ focuses, $K < 0$ spreads	conjugate points and Morse index, Chapter 7

§T2.8: Gauss–Bonnet

Concept	Operational meaning	Returns in
local Gauss–Bonnet	$\int_D K\,\mathrm{d}A + \oint \kappa_g\,\mathrm{d}s + \sum \varepsilon_k = 2\pi$ (T2.119)	Chapter 4
angle excess	geodesic triangle: $\alpha + \beta + \gamma - \pi = \int_T K\,\mathrm{d}A$ (T2.120)	deficit angles, Chapter 9
global Gauss–Bonnet	$\int_\Sigma K\,\mathrm{d}A = 2\pi\chi(\Sigma)$ (T2.124); with boundary add $\oint \kappa_g\,\mathrm{d}s$ (T2.125)	Nash (4.148), Chapter 4
Gauss map degree	$K\,\mathrm{d}A = \hat{n}^*(\mathrm{d}A_{S^2})$, so $\frac{1}{4\pi}\int K\,\mathrm{d}A = \deg\hat{n} = 1 - g$ (T2.130)	monopole charge as degree, Chapter 8
Poincaré–Hopf	sum the indices of the zeros of any vector field; the answer is χ (T2.132)	Morse inequalities, Chapter 7
Euler class	$e = K\,\mathrm{d}A/2\pi = c_1(T\Sigma)$ (T2.133); Pfaffian in 4D (T2.135); $\mathrm{index}(\mathrm{d} + \mathrm{d}^*) = \chi$ (T2.136)	Chern–Weil and the index theorem, Chapter 4

§T2.9: applications

Concept	Operational meaning	Returns in
equivalence problem	compute Ω; if it vanishes, integrate a parallel coframe to find flat coordinates; cones hide curvature at the tip (T2.137)	orbifolds, Chapter 9
2D gravity is topological	$\int R\,\mathrm{d}V = 4\pi\chi$ (T2.141); world sheets weighted by g_s^{2g-2}	string perturbation theory, Chapter 9
Clifford product	$\alpha \vee \beta = \alpha \wedge \beta + g(\alpha,\beta)$ for 1-forms (T2.142); $\mathrm{Cl}(\mathbb{R}^2)$ is the Pauli algebra (T2.146)	gamma matrices, Chapter T4
Kähler–Dirac operator	$\mathrm{d} - \delta = e^i \vee \partial_i$ squares to the Laplacian (T2.149); index with even/odd grading is χ	Dirac operators, Chapter T4; Chapter 4

A catalogue of the spaces we computed

The same handful of spaces appeared in every section. Table T2.3 collects everything we learned about each of them. Read a row from left to right and you replay the chapter: cohomology first (topology only), then a frame and its connection (geometry), then the curvature, and finally the check that total curvature is $2\pi\chi$.

Space	(b_0, b_1, b_2), χ	Generators of H^1_{dR}, H^2_{dR}	Frame data	$\int K\,\mathrm{d}A$
$\mathbb{R}^2$	$(1,0,0)$, 1	none (Poincaré lemma)	polar: $\omega^1{}_2 = -\mathrm{d}\phi$, $K = 0$	–
$\mathbb{R}^2 \setminus \{0\}$	$(1,1,0)$, 0	$\mathrm{d}\theta$, period 2π	polar, $K = 0$	–
S^1	$(1,1)$, 0	$\mathrm{d}\phi$, $\int_{S^1}\mathrm{d}\phi = 2\pi$	–	–
S^2, radius a	$(1,0,1)$, 2	$\omega/4\pi$ (unit monopole)	$\omega^1{}_2 = -\cos\theta\,\mathrm{d}\phi$, $K = 1/a^2$	4π
T^2 in $\mathbb{R}^3$	$(1,2,1)$, 0	$\mathrm{d}\theta$, $\mathrm{d}\phi$; $\mathrm{d}\theta \wedge \mathrm{d}\phi$	$\omega^1{}_2 = \sin\theta\,\mathrm{d}\phi$, $K = \frac{\cos\theta}{a(R + a\cos\theta)}$	0
Σ_g	$(1, 2g, 1)$, $2 - 2g$	$2g$ loop classes; area form	any metric	$4\pi(1-g)$
$\mathbb{R}^3 \setminus \{0\}$	$(1,0,1)$, 2	Coulomb flux $\hat{r}\cdot\mathrm{d}\vec{A}/r^2$	–	–
$\mathbb{R}^3 \setminus$ ring	$(1,1,1)$, 1	linking form (Ampère); flux	–	–

Table T2.3: The chapter's spaces in one table. Betti numbers and χ need only topology; the connection and K need a metric; the last column shows that their integral forgets the metric and remembers χ. For the three-dimensional spaces, χ is the alternating sum of the listed Betti numbers (all higher ones vanish).

The "curvature budget" of Figure T2.36 is the same content for closed surfaces in picture form. Adding a handle costs 4π of total curvature: the metric can move curvature around the surface as it pleases, but it cannot change the total.

Dictionary: the geometry of this chapter as a gauge theory

§T2.5.10 and §T2.6.3 showed that the Levi-Civita connection of a surface is a gauge field whose gauge group is the rotation group $\mathrm{SO}(2) \cong \mathrm{U}(1)$ of the frame. We collect the translation here,

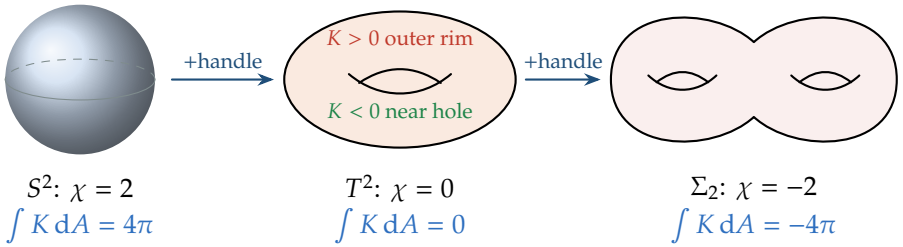


Figure T2.36: The curvature budget. On the sphere the total curvature is 4π however the sphere is dented. On the torus the positive curvature on the outer rim ($K > 0$, red) is cancelled exactly by the negative curvature near the hole ($K < 0$, green), as we checked in (T2.92). Each extra handle adds a saddle-shaped region worth -4π, so $\int_{\Sigma_g} K\,\mathrm{d}A = 4\pi(1-g) = 2\pi\chi$.

because the chapters on gauge theory (Chapter T3) and on instantons (Chapter 8) use it in the other direction.

Dictionary: surface geometry ↔ U(1) gauge theory

Surface geometry	Gauge theory
choice of orthonormal frame (e_1, e_2)	choice of gauge
rotation of frame by angle $\lambda(x)$	gauge transformation $g = e^{i\lambda}$
connection 1-form $\omega^1{}_2$	gauge potential A (with $A_{\text{math}} = -i\,\omega^1{}_2$ in anti-Hermitian form, §T2.5.10)
$\omega^1{}_2 \mapsto \omega^1{}_2 - \mathrm{d}\lambda$	$A \mapsto A + g^{-1}\mathrm{d}g$
$\Omega^1{}_2 = \mathrm{d}\omega^1{}_2 = K\,\mathrm{d}A$	field strength $F = \mathrm{d}A$ (abelian)
$\Omega \mapsto g^{-1}\Omega g$; Bianchi $D\Omega = 0$	$F \mapsto g^{-1}Fg$; Bianchi $DF = 0$
holonomy around a loop $= \int K\,\mathrm{d}A$	Aharonov–Bohm phase, Wilson loop $= \exp \oint A$
sphere: frames smooth on two caps, transition winding 2	monopole of charge 2 in units of Dirac's $\frac{1}{2}$
$\frac{1}{2\pi}\int K\,\mathrm{d}A = \chi$	first Chern number $c_1 = \frac{i}{2\pi}\int F$
geodesic ($\kappa_g = 0$)	path of a charged particle feeling no force
$R\,\mathrm{d}V$ in 2D is a total-derivative density	$\mathrm{tr}\,F \wedge F$ in 4D is locally d(Chern–Simons)
Gauss–Bonnet $\int K\,\mathrm{d}A = 2\pi\chi$	Chern–Weil; index theorem index $= \int$(char. class)

Conventions used in this chapter

- Boundary orientation: *outward normal first*. For a planar region the boundary runs counterclockwise (§T2.2.1).
- A p-form is $\alpha = \frac{1}{p!}\alpha_{\mu_1\ldots\mu_p}\mathrm{d}x^{\mu_1}\wedge\cdots\wedge\mathrm{d}x^{\mu_p}$.
- Frames: $\mathrm{d}e_j = e_i\,\omega^i{}_j$; first structure equation $\mathrm{d}\sigma^i = -\omega^i{}_j\wedge\sigma^j$; curvature $\Omega = \mathrm{d}\omega + \omega\wedge\omega$.
- On the unit sphere in the frame $(\hat\theta, \hat\phi)$ we have $\omega^1{}_2 = -\cos\theta\,\mathrm{d}\phi$ and $\Omega^1{}_2 = \sin\theta\,\mathrm{d}\theta\wedge\mathrm{d}\phi = K\,\sigma^1\wedge\sigma^2$ with $K = +1$. The sign of K is fixed so that spheres are positive.

- Gauss–Bonnet corner terms use the *exterior* (turning) angles $\varepsilon_k = \pi -$ (interior angle).
- Riemannian (Euclidean) signature throughout, except for the remarks on action principles and Maxwell theory, where we stated the signature locally.

Self-test

Each question can be answered in a few lines with the tools of the chapter. The answers follow.

Try it: six quick questions

(a) Without computing any potential, show that $\alpha = \frac{x\,dy\wedge dz + y\,dz\wedge dx + z\,dx\wedge dy}{r^3}$ is not exact on $\mathbb{R}^3 \setminus \{0\}$.

(b) Is $\beta = 2xy\,dx + x^2\,dy$ exact on $\mathbb{R}^2$? If so, find the potential with the homotopy operator K.

(c) A geodesic triangle on the unit sphere has all three angles equal to $\pi/2$. What is its area?

(d) Why can the Gaussian curvature of a closed surface in $\mathbb{R}^3$ of genus 2 not be everywhere positive?

(e) A vector is parallel-transported once around the latitude circle $\theta = \theta_0$ of the unit sphere. By what angle does it come back rotated?

(f) Find the potential for $dx \wedge dy$ on $\mathbb{R}^2$ given by the Poincaré lemma, and check it.

Answers.

(a) It is closed (it is the Coulomb flux form, $d\alpha = 0$ away from the origin), and its integral over the unit sphere is the solid angle:

$$\int_{S^2} \alpha \overset{(1)}{=} \int_{S^2} \sin\theta\, d\theta \wedge d\phi \overset{(2)}{=} 4\pi \neq 0.$$

(1) On the unit sphere $r = 1$, and α restricts to the area form, as in §T2.4.5.

(2) $\int_0^{\pi} \sin\theta\, d\theta = 2$ and $\int_0^{2\pi} d\phi = 2\pi$.

If α were $d\gamma$, Stokes would give $\int_{S^2} \alpha = \int_{\partial S^2} \gamma = 0$, because $\partial S^2 = \varnothing$. The Stokes detector fires.

(b) First test closedness, then build the potential by integrating along rays from the origin:

$$d\beta \overset{(1)}{=} \frac{\partial(2xy)}{\partial y}\,dy\wedge dx + \frac{\partial(x^2)}{\partial x}\,dx\wedge dy \overset{(2)}{=} (-2x + 2x)\,dx\wedge dy = 0,$$

$$K\beta \overset{(3)}{=} \int_0^1 \big(2(tx)(ty)\,x + (tx)^2\,y\big)\,dt \overset{(4)}{=} 3x^2y \int_0^1 t^2\,dt = x^2y.$$

(1) Only the cross derivatives survive in d of a 1-form in the plane.

(2) $dy \wedge dx = -dx \wedge dy$.

(3) The homotopy operator for a 1-form, $K\beta(x) = \int_0^1 \beta_\mu(tx)\,x^\mu\,dt$ (T2.20).

(4) Collect: $2x^2y + x^2y = 3x^2y$, and $\int_0^1 t^2\,dt = \frac{1}{3}$.

Check: $d(x^2y) = 2xy\,dx + x^2\,dy = \beta$. So β is exact, as it must be on a star-shaped region.

(c) By the angle excess formula (T2.120) with $K = 1$, the area is $\frac{\pi}{2} + \frac{\pi}{2} + \frac{\pi}{2} - \pi = \frac{\pi}{2}$. This is the octant of Figure T2.25: one eighth of 4π.

(d) If $K > 0$ everywhere then $\int K \, \mathrm{d}A > 0$. But Gauss–Bonnet gives $\int K \, \mathrm{d}A = 2\pi\chi(\Sigma_2) = 2\pi(2 - 2 \cdot 2) = -4\pi < 0$. So K must be negative somewhere; in fact the negative part must outweigh the positive part by 4π.

(e) The holonomy is the curvature of the enclosed cap (T2.101):

$$\Delta\psi \overset{(1)}{=} \int_{\mathrm{cap}} K \, \mathrm{d}A \overset{(2)}{=} \int_0^{2\pi} \int_0^{\theta_0} \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\phi \overset{(3)}{=} 2\pi(1 - \cos\theta_0).$$

(1) Stokes on $\omega^1{}_2$: the rotation after one circuit is the flux of $\Omega^1{}_2 = K \, \mathrm{d}A$ through the cap.
(2) $K = 1$ and $\mathrm{d}A = \sin\theta \, \mathrm{d}\theta \wedge \mathrm{d}\phi$.
(3) $\int_0^{\theta_0} \sin\theta \, \mathrm{d}\theta = 1 - \cos\theta_0$.

At the equator ($\theta_0 = \pi/2$) this is 2π, the same as no rotation: the equator is a geodesic, so its tangent is itself parallel. This is the Foucault pendulum at co-latitude θ_0.

(f) Apply the homotopy operator $K\alpha = \int_0^1 t^{p-1} \, \iota_x \alpha(tx) \, \mathrm{d}t$ (T2.22), where ι_x inserts the position vector:

$$K(\mathrm{d}x \wedge \mathrm{d}y) \overset{(1)}{=} \int_0^1 t \, \mathrm{d}t \, (x \, \mathrm{d}y - y \, \mathrm{d}x) \overset{(2)}{=} \tfrac{1}{2}(x \, \mathrm{d}y - y \, \mathrm{d}x),$$

$$\mathrm{d}\big(\tfrac{1}{2}(x \, \mathrm{d}y - y \, \mathrm{d}x)\big) \overset{(3)}{=} \tfrac{1}{2}(\mathrm{d}x \wedge \mathrm{d}y - \mathrm{d}y \wedge \mathrm{d}x) = \mathrm{d}x \wedge \mathrm{d}y.$$

(1) Here $p = 2$, so the weight is $t^{p-1} = t$. The coefficient of $\mathrm{d}x \wedge \mathrm{d}y$ is the constant 1, so evaluating it at tx changes nothing. Inserting the position vector $x \, \partial_x + y \, \partial_y$ into the first slot gives $\iota_x(\mathrm{d}x \wedge \mathrm{d}y) = x \, \mathrm{d}y - y \, \mathrm{d}x$.
(2) $\int_0^1 t \, \mathrm{d}t = \frac{1}{2}$.
(3) $\mathrm{d}(x \, \mathrm{d}y) = \mathrm{d}x \wedge \mathrm{d}y$ and $\mathrm{d}(y \, \mathrm{d}x) = \mathrm{d}y \wedge \mathrm{d}x$.

This is the familiar "area $= \frac{1}{2} \oint (x \, \mathrm{d}y - y \, \mathrm{d}x)$" of Green's theorem, now obtained as the Poincaré-lemma potential of the area form.

Where the chapter leads

- Chapter T3 replaces the rotation group $\mathrm{SO}(n)$ of the frame by any Lie group G. Every formula of §T2.5 and §T2.6 survives with $\omega \to A$ and $\Omega \to F$; holonomy becomes the Wilson loop.
- Chapter T4 lifts the frame rotation to spinors. The spin connection is our ω written in the Clifford algebra of §T2.9.3, and the Dirac operator is the honest version of the Kähler–Dirac operator of §T2.9.5.
- Chapter 1 makes cohomology precise, adds integer coefficients and torsion, and proves the invariance we used.
- Chapter 2 develops the Hodge theory previewed in §T2.4.9: harmonic forms as zero modes of the Laplacian.
- Chapter 4 states Gauss–Bonnet as the index theorem for $\mathrm{d} + \mathrm{d}^*$ (Nash (4.52)–(4.54), (4.148)–(4.149)) and generalises the Euler class to all characteristic classes by Chern–Weil.
- Chapter 8 computes monopole and instanton numbers by exactly the two-patch argument of §T2.3.6.

- Chapter 9 sums over surfaces of every genus, weighted by the topological Einstein–Hilbert action of §T2.9.2.

Recap: the whole chapter

- Forms are integrands; Stokes, $\int_R \mathrm{d}\alpha = \int_{\partial R} \alpha$, is the one theorem of integral calculus.
- Exact forms integrate to zero over closed cycles; closed forms that do not are cohomology classes, and they count holes: $b_p = \dim H^p_{\mathrm{dR}}$, $\chi = \sum (-1)^p b_p$.
- A connection is how a frame turns, $\mathrm{d}e = e\,\omega$; the curvature $\Omega = \mathrm{d}\omega + \omega \wedge \omega$ is what does not cancel around a loop. Frame rotation is a gauge transformation.
- Gauss–Bonnet: $\int_\Sigma K\,\mathrm{d}A = 2\pi\chi$. The first "integral of curvature = topology" theorem, and the seed of Chern–Weil theory and the Atiyah–Singer index theorem.

Chapter T3

Lie Groups, Lie Algebras and Gauge Fields

Where this chapter is going

Almost everything in Nash that has a physicist's name attached to it – Yang–Mills fields, instantons, monopoles, anomalies, Chern–Simons theory, Wilson loops, current algebras – is built from two ingredients. The first is a *group of symmetries* that can be described by a finite number of continuous parameters: a Lie group such as U(1), SU(2) or SU(3). The second is the idea of making that symmetry *local*, allowing the symmetry transformation to vary from point to point in spacetime. Once we do that, ordinary derivatives stop making sense and must be replaced by covariant derivatives. The correction term in the covariant derivative is the gauge field A. Geometrically it is a connection, and its curvature $F = \mathrm{d}A + A \wedge A$ is the field strength.

This chapter builds both ingredients from scratch, with explicit matrices at every step. We start with groups of matrices (Section T3.1), look at their infinitesimal versions, the Lie algebras (Section T3.2), and connect the two with the exponential map (Section T3.3). We then meet the most important single example in physics, the two-to-one map $\mathrm{SU}(2) \to \mathrm{SO}(3)$ that is responsible for spin $\frac{1}{2}$ (Section T3.4), and the representation theory that labels particles by spin, weights and roots (Section T3.5). The Maurer–Cartan form $g^{-1}\mathrm{d}g$ (Section T3.6) is the bridge between groups and differential forms: it is the prototype of every gauge field, and its cube counts windings. Group actions and homogeneous spaces (Section T3.7) tell us how spheres and projective spaces arise as quotients of groups. With this toolkit in hand we derive gauge theory (Section T3.8): the transformation law of A, the field strength, the Bianchi identity and the Yang–Mills equations. Finally we look at what a connection does to a particle carried around a loop – parallel transport, holonomy, Wilson loops and the Aharonov–Bohm effect (Section T3.9) – and at connections with zero curvature and the Chern–Simons form, whose derivative is $\operatorname{tr} F \wedge F$ (Section T3.10). That last identity is the doorway to instantons (Chapter 8) and anomalies (Chapter 10).

Conventions for the whole chapter. We use the mathematicians' (and Nash's) conventions. Lie algebra elements of unitary groups are *anti-Hermitian* matrices, a gauge field is a Lie-algebra-valued 1-form $A = A_\mu \, \mathrm{d}x^\mu$, the covariant derivative is $D = \mathrm{d} + A$, and

$$F = \mathrm{d}A + A \wedge A, \qquad A \mapsto g^{-1}Ag + g^{-1}\mathrm{d}g, \qquad F \mapsto g^{-1}Fg.$$

Physicists usually pull out a factor of $-i$ and a coupling constant and use Hermitian generators; the translation is $A_{\text{math}} = -i\, g\, A^a_{\text{phys}} T^a$, and we give a full dictionary in Section T3.8. Spacetime is Euclidean unless we say otherwise. Wedge products of matrix-valued forms multiply the matrices *and* wedge the forms, keeping the order.

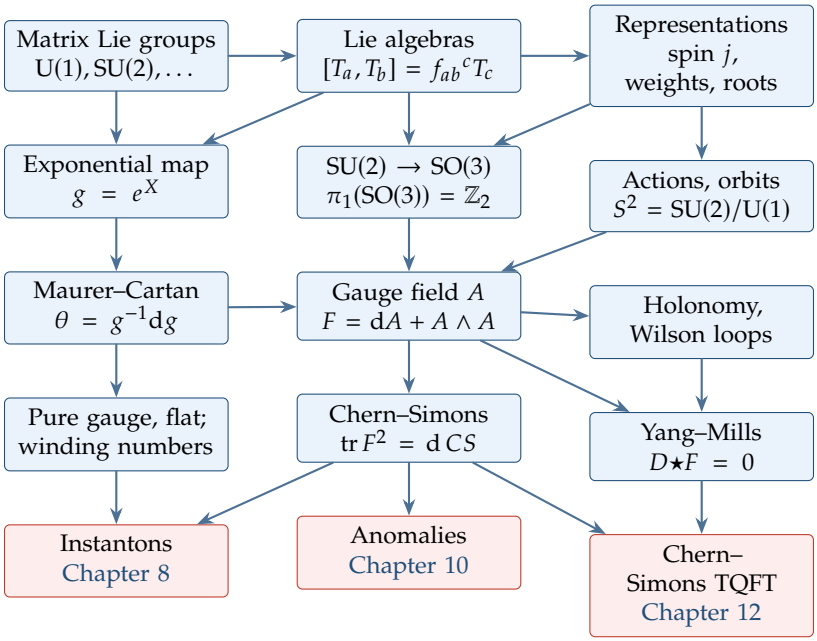


Figure T3.1: The logic of this chapter. Read it top to bottom. The first row is pure algebra: groups, their infinitesimal versions, and the ways they act on vectors. The middle rows turn algebra into geometry: the Maurer–Cartan form $g^{-1}\mathrm{d}g$ is the seed from which the gauge field grows. The red boxes at the bottom are where Nash uses the results.

T3.1 Matrix Lie groups

A symmetry is a transformation that leaves something unchanged: a rotation leaves lengths unchanged, a phase rotation $\psi \mapsto e^{i\alpha}\psi$ leaves $|\psi|^2$ unchanged, an SU(3) colour rotation leaves the strong interactions unchanged. Two symmetries done one after the other give another symmetry, and every symmetry can be undone. That is all a *group* is. When the symmetries can be labelled by a finite number of continuous real parameters (an angle, three Euler angles, eight colour angles) we have a *Lie group*. In physics these groups almost always come to us as groups of matrices acting on a vector space of states or fields. So we start with matrices and never leave them.

T3.1.1 Symmetries as matrices

Working definition T3.1 (group)

A **group** is a set G with a product $g \cdot h$ such that (i) the product of two elements is again in G, (ii) there is an identity e with $e \cdot g = g \cdot e = g$, (iii) every g has an inverse g^{-1} with $g\,g^{-1} = g^{-1}g = e$, and (iv) the product is associative [$(g \cdot h) \cdot k = g \cdot (h \cdot k)$, so a long product needs no brackets]. *Operational test for a set of matrices:* check that it contains $\mathbb{1}$, and that it is closed under matrix multiplication and matrix inversion. Associativity is automatic for matrices.

Working definition T3.2 (matrix Lie group)

A **matrix Lie group** is a set of invertible $n \times n$ real or complex matrices which is a group, and which is cut out of the space of all matrices by smooth equations (such as $R^T R = \mathbb{1}$ or $\det U = 1$), so that it is a smooth manifold [near every element the group looks like a

piece of $\mathbb{R}^d$, so nearby elements are labelled by d real coordinates]. Its **dimension** is the number of independent real parameters needed to describe an element near the identity. *How to compute the dimension:* count the real entries of the matrix, then subtract the number of independent real equations in the defining condition.

Almost every group we need is defined by a sentence of the form *"the matrices that preserve a certain structure."* The structure is an inner product, a volume, an orientation or a symplectic form. The defining equation then says that the structure is unchanged. For example, a real matrix R preserves the dot product if $(Rx)\cdot(Ry) = x\cdot y$ for all x, y. Writing the dot product as $x^T y$,

$$(Rx)\cdot(Ry) \overset{(1)}{=} x^T R^T R\, y \overset{(2)}{=} x^T y \quad \text{for all } x,y \quad \Longleftrightarrow \quad R^T R = \mathbb{1}.$$

(1) The transpose of a product reverses the order: $(Rx)^T = x^T R^T$.
(2) This is the requirement. Taking x, y to be basis vectors picks out each entry of $R^T R$, so the requirement holds for all x, y exactly when $R^T R$ is the identity matrix.

T3.1.2 The catalogue of classical groups

Table T3.1 lists the groups we shall meet, with their defining conditions. We now go through them one at a time and count dimensions, because the counting method is the same every time and it is worth doing once slowly.

Group	Matrices	Condition	Preserves	Real dim.
$\mathrm{GL}(n,\mathbb{R})$	real $n\times n$	$\det M \neq 0$	nothing (all invertible)	n^2
$\mathrm{GL}(n,\mathbb{C})$	complex $n\times n$	$\det M \neq 0$	nothing	$2n^2$
$\mathrm{SL}(n,\mathbb{R})$	real	$\det M = 1$	volume and orientation	n^2-1
$\mathrm{O}(n)$	real	$R^T R = \mathbb{1}$	dot product $x\cdot y$	$\frac{1}{2}n(n-1)$
$\mathrm{SO}(n)$	real	$R^T R = \mathbb{1}$, $\det R = 1$	dot product, orientation	$\frac{1}{2}n(n-1)$
$\mathrm{U}(n)$	complex	$U^\dagger U = \mathbb{1}$	$\langle\xi,\chi\rangle = \xi^\dagger\chi$	n^2
$\mathrm{SU}(n)$	complex	$U^\dagger U = \mathbb{1}$, $\det U = 1$	$\xi^\dagger\chi$ and complex volume	n^2-1
$\mathrm{Sp}(2n,\mathbb{R})$	real $2n\times 2n$	$M^T J M = J$	symplectic form $x^T J y$	$n(2n+1)$
$\mathrm{Sp}(n)$	complex $2n\times 2n$	$U^\dagger U = \mathbb{1}$, $U^T J U = J$	both of the above	$n(2n+1)$

Table T3.1: The classical matrix groups. Here $J = \left(\begin{smallmatrix} 0 & \mathbb{1}_n \\ -\mathbb{1}_n & 0 \end{smallmatrix}\right)$. The last column is the number of real parameters. The compact ones are $\mathrm{O}(n)$, $\mathrm{SO}(n)$, $\mathrm{U}(n)$, $\mathrm{SU}(n)$ and $\mathrm{Sp}(n)$.

$\mathrm{GL}(n)$ and $\mathrm{SL}(n)$. The **general linear group** $\mathrm{GL}(n,\mathbb{R})$ is all invertible real matrices. The condition $\det M \neq 0$ is an inequality, not an equation, so it removes no parameters: $\dim \mathrm{GL}(n,\mathbb{R}) = n^2$. The **special linear group** $\mathrm{SL}(n,\mathbb{R})$ imposes one real equation $\det M = 1$, so its dimension is n^2-1. Geometrically, $\det M$ is the factor by which M scales volumes, so $\mathrm{SL}(n)$ is the group of volume-preserving, orientation-preserving linear maps.

$\mathrm{O}(n)$ and $\mathrm{SO}(n)$. The **orthogonal group** preserves the dot product, $R^T R = \mathbb{1}$. Counting:

$$\begin{aligned}\dim \mathrm{O}(n) &\overset{(1)}{=} n^2 - \#\{\text{independent equations in } R^T R = \mathbb{1}\} \\ &\overset{(2)}{=} n^2 - \tfrac{1}{2}n(n+1) \\ &= \tfrac{1}{2}n(n-1).\end{aligned}$$

(1) Start from the n^2 real entries of R and subtract one parameter per independent real equation.
(2) The matrix $R^T R$ is automatically symmetric, $(R^T R)^T = R^T R$, so the equation $R^T R = \mathbb{1}$ has only as many independent components as a symmetric matrix: n on the diagonal plus $\frac{1}{2}n(n-1)$ above it, total $\frac{1}{2}n(n+1)$.

Taking the determinant of $R^T R = \mathbb{1}$ gives $(\det R)^2 = 1$, so $\det R = \pm 1$. The **special orthogonal group** $\mathrm{SO}(n)$ keeps only $\det R = +1$: the rotations, which also preserve orientation. Because det is continuous and can only take the values ± 1, no continuous path in $\mathrm{O}(n)$ joins a rotation to a reflection. So $\mathrm{O}(n)$ has two *connected components*, and $\mathrm{SO}(n)$ is the one containing $\mathbb{1}$. Adding the discrete condition $\det R = 1$ does not change the dimension.

$\mathrm{U}(n)$ and $\mathrm{SU}(n)$. The **unitary group** preserves the complex inner product $\langle \xi, \chi \rangle = \xi^\dagger \chi = \sum_a \bar{\xi}_a \chi_a$: exactly as for $\mathrm{O}(n)$, this means $U^\dagger U = \mathbb{1}$. It is the group of symmetries of quantum mechanics, because it preserves transition amplitudes.

$$\begin{aligned}\dim \mathrm{U}(n) &\overset{(1)}{=} 2n^2 - \#\{\text{real equations in } U^\dagger U = \mathbb{1}\} \\ &\overset{(2)}{=} 2n^2 - n^2 = n^2.\end{aligned}$$

(1) A complex $n \times n$ matrix has n^2 complex, so $2n^2$ real, entries.
(2) $U^\dagger U$ is automatically Hermitian. A Hermitian matrix has n real diagonal entries and $\frac{1}{2}n(n-1)$ complex entries above the diagonal, so $n + n(n-1) = n^2$ real components.

For a unitary matrix, $|\det U|^2 = \det U^\dagger \det U = \det(U^\dagger U) = 1$, so $\det U = e^{i\varphi}$ is a phase. Requiring $\det U = 1$ is therefore *one* real condition (on φ), and the **special unitary group** has $\dim \mathrm{SU}(n) = n^2 - 1$. So $\dim \mathrm{SU}(2) = 3$ (the three components of isospin or spin) and $\dim \mathrm{SU}(3) = 8$ (the eight gluons).

The symplectic groups. Classical mechanics lives on phase space with coordinates (q, p). The Poisson bracket is built from the antisymmetric form $\omega(x, y) = x^T J y$ with $J = \begin{pmatrix} 0 & \mathbb{1}_n \\ -\mathbb{1}_n & 0 \end{pmatrix}$. Linear canonical transformations preserve it: $M^T J M = J$. They form the real symplectic group $\mathrm{Sp}(2n, \mathbb{R})$.

$$\begin{aligned}\dim \mathrm{Sp}(2n, \mathbb{R}) &\overset{(1)}{=} (2n)^2 - \tfrac{1}{2}(2n)(2n-1) \\ &= n(2n+1).\end{aligned}$$

(1) $M^T J M$ is automatically antisymmetric, because $(M^T J M)^T = M^T J^T M = -M^T J M$. An antisymmetric $2n \times 2n$ matrix has $\frac{1}{2}(2n)(2n-1)$ independent entries, and that is the number of equations.

The *compact* symplectic group $\mathrm{Sp}(n)$ consists of the $2n \times 2n$ complex matrices that are both unitary and symplectic. It has the same dimension $n(2n+1)$. It is also the group of $n \times n$ *quaternionic* matrices preserving the quaternionic inner product; we will see in a moment that $\mathrm{Sp}(1) = \mathrm{SU}(2)$.

Watch out: Two names for the same symplectic group

Some books (including the QFT notes, [QFT notes §31.1]) write $\mathrm{Sp}(2N)$ for the group we call $\mathrm{Sp}(N)$, labelling it by the size of the matrices rather than by the quaternionic dimension. Both conventions give the same Lie algebra, of dimension $N(2N+1)$. Always check which is meant by computing the dimension: Sp(1) in our convention ("Sp(2)" in the other) is 3-dimensional.

U(1). The simplest of all is $\mathrm{U}(1) = \{e^{i\alpha}\}$, the 1×1 unitary matrices. It is the symmetry group of electromagnetism: the phase of a charged field.

T3.1.3 U(1) is a circle, and so is SO(2)

Worked example T3.1: $\mathrm{U}(1) \cong \mathrm{SO}(2) \cong S^1$

A 1×1 complex matrix u is unitary when $\bar{u}u = |u|^2 = 1$. So U(1) is the unit circle in the complex plane, $u = e^{i\alpha}$ with $\alpha \in [0, 2\pi)$. Now look at rotations of the plane. Write $R(\alpha) = \left(\begin{smallmatrix}\cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha\end{smallmatrix}\right)$; every element of SO(2) has this form (the first column is a unit vector $(\cos\alpha, \sin\alpha)$, and the second column must be the unit vector perpendicular to it that gives det = +1). The map $e^{i\alpha} \mapsto R(\alpha)$ respects products:

$$R(\alpha)R(\beta) \overset{(1)}{=} \begin{pmatrix}\cos\alpha\cos\beta - \sin\alpha\sin\beta & -(\sin\alpha\cos\beta + \cos\alpha\sin\beta) \\ \sin\alpha\cos\beta + \cos\alpha\sin\beta & \cos\alpha\cos\beta - \sin\alpha\sin\beta\end{pmatrix}$$
$$\overset{(2)}{=} R(\alpha+\beta),$$

(1) Matrix multiplication, row times column.
(2) The addition formulas for $\cos(\alpha+\beta)$ and $\sin(\alpha+\beta)$.

which is the image of $e^{i\alpha}e^{i\beta} = e^{i(\alpha+\beta)}$. So the two groups are the same group written differently. A map between groups that respects products is called a **homomorphism**; a bijective one is an **isomorphism**. Physically, a charged scalar field $\phi = \phi_1 + i\phi_2$ can be written either as one complex number rotating by $e^{i\alpha}$ or as a real 2-vector (ϕ_1, ϕ_2) rotating by $R(\alpha)$.

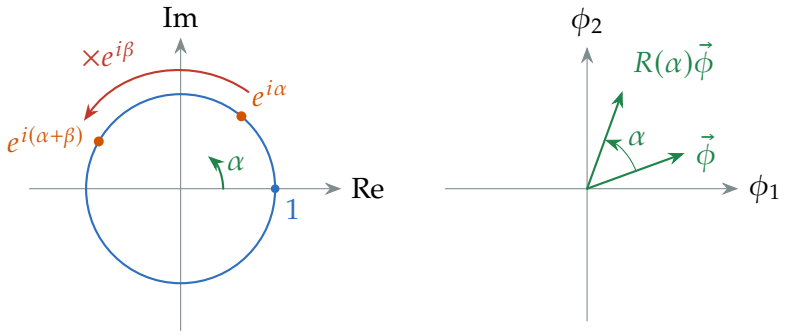


Figure T3.2: Left: U(1) is the unit circle in $\mathbb{C}$; multiplying by $e^{i\beta}$ moves a point a further angle β round the circle, so the group product is addition of angles. Right: the same group acting as SO(2) rotations of a real 2-vector (ϕ_1, ϕ_2).

T3.1.4 SU(2) is the three-sphere

This is the most useful explicit fact in the chapter. Let $U = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ be in SU(2). Because $\det U = 1$, the inverse is given by the familiar 2×2 formula, and because U is unitary the inverse is also $U^\dagger$:

$$\begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \overset{(1)}{=} U^{-1} \overset{(2)}{=} U^\dagger = \begin{pmatrix} \bar a & \bar c \\ \bar b & \bar d \end{pmatrix}.$$

(1) For a 2×2 matrix, $U^{-1} = \frac{1}{\det U}\left(\begin{smallmatrix} d & -b \\ -c & a \end{smallmatrix}\right)$, and $\det U = 1$.
(2) Unitarity: $U^\dagger U = \mathbb{1}$ means $U^{-1} = U^\dagger$.

Comparing entries, $d = \bar a$ and $c = -\bar b$. So

$$U = \begin{pmatrix} a & -\bar b \\ b & \bar a \end{pmatrix}, \qquad \det U = |a|^2 + |b|^2 = 1. \tag{T3.1}$$

Write $a = x^4 + ix^3$ and $b = -x^2 + ix^1$ with four real numbers x^μ. Then

$$\begin{aligned} U &= \begin{pmatrix} x^4 + ix^3 & x^2 + ix^1 \\ -x^2 + ix^1 & x^4 - ix^3 \end{pmatrix} \\ &\overset{(1)}{=} x^4\,\mathbb{1} + i\left(x^1\sigma_1 + x^2\sigma_2 + x^3\sigma_3\right) = x^4\,\mathbb{1} + i\,\vec{x}\cdot\vec{\sigma}, \end{aligned} \tag{T3.2}$$

(1) Using the Pauli matrices

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

we have $ix^1\sigma_1 = \left(\begin{smallmatrix} 0 & ix^1 \\ ix^1 & 0 \end{smallmatrix}\right)$, $ix^2\sigma_2 = \left(\begin{smallmatrix} 0 & x^2 \\ -x^2 & 0 \end{smallmatrix}\right)$, $ix^3\sigma_3 = \mathrm{diag}(ix^3, -ix^3)$; add these to $x^4\mathbb{1}$.

and the condition $\det U = 1$ reads

$$(x^1)^2 + (x^2)^2 + (x^3)^2 + (x^4)^2 = 1. \tag{T3.3}$$

Key idea: SU(2) $\cong S^3$

Every element of SU(2) is $U = x^4\mathbb{1} + i\,\vec{x}\cdot\vec{\sigma}$ for a unique point (x^1, x^2, x^3, x^4) of the unit three-sphere $S^3 \subset \mathbb{R}^4$, and every point of S^3 gives an element of SU(2). The identity is the "north pole" $x^4 = 1$; the element $-\mathbb{1}$ is the "south pole" $x^4 = -1$.

Because S^3 is simply connected (any loop on it can be shrunk to a point), so is SU(2). This fact will be the whole reason spin $\frac{1}{2}$ exists (Section T3.4).

Worked example T3.2: Unit quaternions and Sp(1) = SU(2)

Set $\mathbf{I} = -i\sigma_1$, $\mathbf{J} = -i\sigma_2$, $\mathbf{K} = -i\sigma_3$. Then

$$\begin{aligned} \mathbf{IJ} &\overset{(1)}{=} (-i)^2\sigma_1\sigma_2 \overset{(2)}{=} -i\sigma_3 = \mathbf{K}, \\ \mathbf{I}^2 &\overset{(3)}{=} (-i)^2\sigma_1^2 = -\mathbb{1}, \end{aligned}$$

(1) Definition of $\mathbf{I}, \mathbf{J}$.
(2) The Pauli product rule $\sigma_a\sigma_b = \delta_{ab}\mathbb{1} + i\varepsilon_{abc}\sigma_c$ gives $\sigma_1\sigma_2 = i\sigma_3$, and $(-i)^2\cdot i = -i$.
(3) $\sigma_1^2 = \mathbb{1}$.

and similarly $\mathbf{J}^2 = \mathbf{K}^2 = -\mathbb{1}$, $\mathbf{JK} = \mathbf{I}$, $\mathbf{KI} = \mathbf{J}$. These are Hamilton's quaternion relations. So (T3.2) says $U = x^4 - x^1\mathbf{I} - x^2\mathbf{J} - x^3\mathbf{K}$: SU(2) is the group of *unit quaternions* under multiplication. Unit quaternions are the 1×1 quaternionic unitary matrices, that is Sp(1). The dimension check agrees: $\dim \mathrm{Sp}(1) = 1 \cdot (2+1) = 3 = \dim \mathrm{SU}(2)$.

Worked example T3.3: Conjugacy classes of SU(2) are 2-spheres of latitude

Two group elements U and VUV^{-1} are called **conjugate**; in physics, they are the "same" transformation seen in a rotated frame. Which elements of SU(2) are conjugate? The trace is unchanged by conjugation, $\mathrm{tr}(VUV^{-1}) = \mathrm{tr}(U)$, and from (T3.2) $\mathrm{tr}\, U = 2x^4$ because the Pauli matrices are traceless. So conjugation moves U only within the slice $x^4 = \text{const}$ of S^3. That slice is the 2-sphere $|\vec{x}|^2 = 1 - (x^4)^2$. In fact conjugation moves U over the *whole* slice: writing $\vec{x} = |\vec{x}|\,\hat{n}$, $U = x^4 + i|\vec{x}|\,\hat{n}\cdot\vec{\sigma}$, and we will see in Section T3.4 that $V(\hat{n}\cdot\vec{\sigma})V^{-1} = (R\hat{n})\cdot\vec{\sigma}$ with R an arbitrary rotation. Writing $x^4 = \cos(\theta/2)$ and $|\vec{x}| = \sin(\theta/2)$, the eigenvalues of U are $x^4 \pm i|\vec{x}| = e^{\pm i\theta/2}$, and each conjugacy class is labelled by one angle $\theta \in [0, 2\pi]$. The two poles $\pm\mathbb{1}$ are classes on their own: they commute with everything. They form the **centre** $Z(\mathrm{SU}(2)) = \{\pm\mathbb{1}\} \cong \mathbb{Z}_2$.

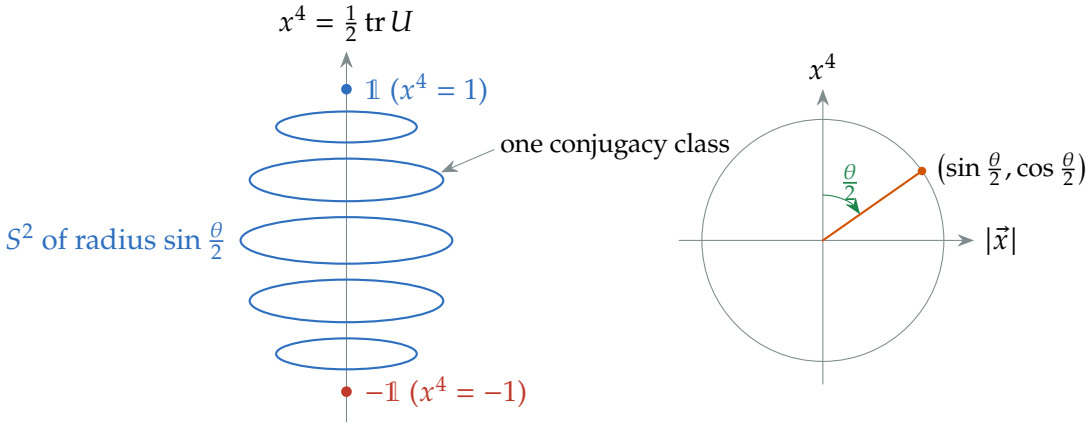


Figure T3.3: SU(2) = S^3 drawn as a stack of 2-spheres, one for each value of $x^4 = \frac{1}{2}\,\mathrm{tr}\, U = \cos\frac{\theta}{2}$. Each 2-sphere is a conjugacy class (all elements with eigenvalues $e^{\pm i\theta/2}$). The stack shrinks to a point at the two poles $\pm\mathbb{1}$, which form the centre. Right: the angle θ that labels the slice.

T3.1.5 SO(3) is a solid ball with its surface glued

Every rotation of $\mathbb{R}^3$ is a rotation by some angle about some axis. This is Euler's theorem, and its proof is a one-line determinant calculation showing that $R \in \mathrm{SO}(3)$ always has eigenvalue 1:

$$\det(R - \mathbb{1}) \overset{(1)}{=} \det(R - RR^T) \overset{(2)}{=} \det R\, \det(\mathbb{1} - R^T) \overset{(3)}{=} \det(\mathbb{1} - R)$$
$$\overset{(4)}{=} (-1)^3 \det(R - \mathbb{1}),$$

(1) $RR^T = \mathbb{1}$ for an orthogonal matrix.
(2) Factor out R on the left; the determinant of a product is the product of determinants.
(3) $\det R = 1$, and a matrix and its transpose have the same determinant.
(4) Pull the factor -1 out of each of the three rows.

so $\det(R-\mathbb{1}) = -\det(R-\mathbb{1}) = 0$. The eigenvector $\hat{n}$ with $R\hat{n} = \hat{n}$ is the axis. In the plane perpendicular to $\hat{n}$, R is an SO(2) rotation by some angle θ. Choosing coordinates with $\hat{n} = \hat{z}$,

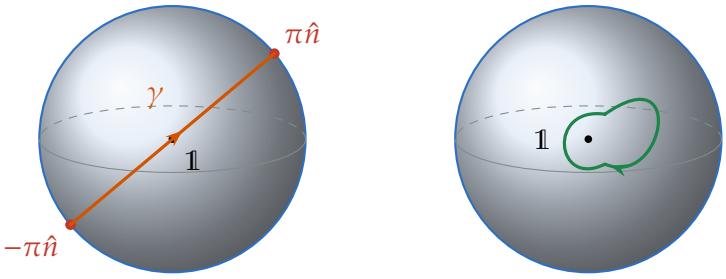
$$R_z(\theta) = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \operatorname{tr} R = 1 + 2\cos\theta. \tag{T3.4}$$

The trace is basis independent, so $\operatorname{tr} R = 1 + 2\cos\theta$ gives the rotation angle of any rotation matrix.

We can now draw SO(3). Represent the rotation by angle $\theta \in [0, \pi]$ about $\hat{n}$ by the vector $\theta\hat{n}$. These vectors fill a solid ball of radius π. Rotations by angles in $(\pi, 2\pi)$ are not new: rotation by θ about $\hat{n}$ equals rotation by $2\pi - \theta$ about $-\hat{n}$. But at the surface there is a double counting: rotation by π about $\hat{n}$ is the same as rotation by π about $-\hat{n}$. So

Key idea: SO(3) as a ball

SO(3) is the solid ball of radius π in which each pair of antipodal points on the surface is identified. This space is the real projective space $\mathbb{RP}^3$.



a closed loop in SO(3) that cannot be shrunk

a loop inside the ball shrinks to $\mathbb{1}$

Figure T3.4: SO(3) as the ball of radius π with antipodal surface points identified: the two red points $\pm\pi\hat{n}$ are the same rotation. Left: the straight path γ from $-\pi\hat{n}$ to $\pi\hat{n}$ (a full 2π rotation about $\hat{n}$) is a *closed* loop, because its endpoints are the same rotation. Moving the endpoints moves both at once (they must stay antipodal), so this loop can never be shrunk. Right: an ordinary loop that stays inside can be shrunk. We show in Section T3.4 that going round γ *twice* can be shrunk.

Worked example T3.4: Axis and angle of a permutation matrix

Take the matrix that cycles the axes, $R\,\hat{x} = \hat{y}$, $R\,\hat{y} = \hat{z}$, $R\,\hat{z} = \hat{x}$:

$$R = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

It is orthogonal (its columns are orthonormal) and $\det R = +1$. Then

$$1 + 2\cos\theta \overset{(1)}{=} \operatorname{tr} R \overset{(2)}{=} 0 \quad \Longrightarrow \quad \cos\theta = -\tfrac{1}{2}, \;\; \theta = \tfrac{2\pi}{3},$$

(1) The trace formula (T3.4), valid in any basis.
(2) All diagonal entries of R vanish.

The axis is the fixed vector: $R(1,1,1)^T = (1,1,1)^T$, so $\hat{n} = (1,1,1)/\sqrt{3}$. Geometrically: a rotation by 120° about the body diagonal of a cube permutes the three axes. Note that $R^3 = \mathbb{1}$, consistent with $3 \times 120° = 360°$.

T3.1.6 Topology at a glance

Several global properties of these groups matter in gauge theory. A group is **compact** if it is closed and bounded as a set of matrices; $\mathrm{U}(n)$ is bounded because each column of a unitary matrix is a unit vector. Compact groups have unitary finite-dimensional representations and positive-definite invariant inner products on their Lie algebras, which is why gauge groups are compact: the kinetic energy must be positive. The **fundamental group** $\pi_1(G)$ classifies loops in G up to deformation, and $\pi_3(G)$ classifies maps $S^3 \to G$ up to deformation. Table T3.2 collects the facts.

G	compact	π_0 (components)	π_1	π_3
$\mathrm{U}(1) = S^1$	yes	0	$\mathbb{Z}$	0
$\mathrm{SU}(2) = S^3$	yes	0	0	$\mathbb{Z}$
$\mathrm{SO}(3) = \mathbb{RP}^3$	yes	0	$\mathbb{Z}_2$	$\mathbb{Z}$
$\mathrm{SU}(n)$, $n \geq 2$	yes	0	0	$\mathbb{Z}$
$\mathrm{U}(n)$	yes	0	$\mathbb{Z}$	$\mathbb{Z}$ $(n \geq 2)$
$\mathrm{SO}(n)$, $n \geq 3$	yes	0	$\mathbb{Z}_2$	$\mathbb{Z}$ $(n = 3, n \geq 5)$, $\mathbb{Z} \oplus \mathbb{Z}$ $(n = 4)$
$\mathrm{O}(n)$	yes	$\mathbb{Z}_2$	as $\mathrm{SO}(n)$	as $\mathrm{SO}(n)$
$\mathrm{Sp}(n)$	yes	0	0	$\mathbb{Z}$
$\mathrm{GL}(n, \mathbb{R})$	no	$\mathbb{Z}_2$ $(\det \gtrless 0)$	as $\mathrm{SO}(n)$	as $\mathrm{SO}(n)$
$\mathrm{SL}(2, \mathbb{R})$	no	0	$\mathbb{Z}$	0

Table T3.2: Topological data of the common Lie groups. Two entries matter most for physics. $\pi_1(\mathrm{U}(1)) = \mathbb{Z}$ is the winding number of a phase, which gives Dirac monopoles and vortices. $\pi_3(G) = \mathbb{Z}$ for every compact simple G is the winding number of a gauge transformation at infinity in four dimensions, which gives instantons.

Three entries can be understood right now. $\pi_1(\mathrm{U}(1)) = \mathbb{Z}$: a loop $\alpha \mapsto e^{in\alpha}$ in the circle winds n times, and loops with different n cannot be deformed into each other. $\pi_1(\mathrm{SU}(2)) = 0$ and $\pi_3(\mathrm{SU}(2)) = \mathbb{Z}$: $\mathrm{SU}(2)$ is S^3, loops on S^3 shrink, and maps $S^3 \to S^3$ have an integer degree (we compute it with an integral in Section T3.6). $\pi_1(\mathrm{U}(n)) = \mathbb{Z}$: the map $\det : \mathrm{U}(n) \to \mathrm{U}(1)$ sends a loop in $\mathrm{U}(n)$ to a loop in the circle, and its winding number is the invariant. The entry $\pi_1(\mathrm{SO}(3)) = \mathbb{Z}_2$ is explained in Section T3.4.

Try it: The Lorentz group has dimension 6

The Lorentz group $\mathrm{O}(1,3)$ consists of real 4×4 matrices with $\Lambda^T \eta \Lambda = \eta$, $\eta = \mathrm{diag}(-1,1,1,1)$. Count its dimension.
Answer: $\Lambda^T \eta \Lambda$ is symmetric, so there are 10 equations and $16 - 10 = 6$ parameters: three rotations and three boosts. It is not compact, since boosts can be arbitrarily large.

Reading Nash (where these groups appear)

Nash takes Lie groups for granted. His structure groups are "G a compact simple Lie group", typically $\mathrm{SU}(2)$ for instantons [Nash p. 25, (1.74)] and $\mathrm{U}(N)$ or $\mathrm{SU}(N)$ when he computes the cohomology of the gauge group [Nash p. 223, (8.38)]. The fact $\pi_3(G) = \mathbb{Z}$ that we quoted in Table T3.2 is what he uses to count the components of the gauge group in three dimensions [Nash p. 227, (8.59)]: $\pi_0(\mathcal{G}) = \pi_3(G) = \mathbb{Z}$. The 3-sphere picture of $\mathrm{SU}(2)$ is used silently throughout Chapter 8.

Recap

- A matrix Lie group is a group of matrices defined by smooth equations; its dimension is the number of entries minus the number of independent equations.
- $\dim \mathrm{SO}(n) = \frac{1}{2}n(n-1)$, $\dim \mathrm{U}(n) = n^2$, $\dim \mathrm{SU}(n) = n^2 - 1$, $\dim \mathrm{Sp}(n) = n(2n+1)$.
- $\mathrm{U}(1) = \mathrm{SO}(2) = S^1$; $\mathrm{SU}(2) = \{x^4 + i\vec{x}\cdot\vec{\sigma}\} = S^3$ = unit quaternions = $\mathrm{Sp}(1)$.
- $\mathrm{SO}(3)$ is a ball of radius π with antipodal surface points glued, $\mathbb{RP}^3$; $\mathrm{tr}\, R = 1 + 2\cos\theta$.
- Topological data: $\pi_1(\mathrm{U}(1)) = \mathbb{Z}$, $\pi_1(\mathrm{SO}(3)) = \mathbb{Z}_2$, $\pi_3(G) = \mathbb{Z}$ for compact simple G.

T3.2 Lie algebras: the tangent space at the identity

A Lie group is a curved space: U(1) is a circle and SU(2) is a 3-sphere. Curved spaces are hard to work with, but near any point they look flat. Near the identity, a group element is "the identity plus something small", and the small somethings form an ordinary vector space. Physicists call its elements *infinitesimal generators*: angular momentum generates rotations, charge generates phase rotations, momentum generates translations. The question of this section is: *what is left of the group structure once we zoom in to the identity?* The answer is a vector space together with one extra operation, the commutator, and that is a Lie algebra.

T3.2.1 Infinitesimal transformations

Take a smooth curve $g(t)$ in the group with $g(0) = \mathbb{1}$. Its velocity at $t = 0$,

$$X = \left.\frac{\mathrm{d}g}{\mathrm{d}t}\right|_{t=0}, \qquad g(\varepsilon) = \mathbb{1} + \varepsilon X + O(\varepsilon^2),$$

is a matrix: a tangent vector to the group at the identity.

Working definition T3.3 (Lie algebra of a matrix group)

The **Lie algebra** $\mathfrak{g}$ of a matrix group G is the set of all velocities $X = \dot{g}(0)$ of curves through the identity. It is a real vector space [if X and Y are velocities, so is every real combination $aX + bY$] of the same dimension as G. *How to find it:* substitute $g = \mathbb{1} + \varepsilon X$ into the defining equations of G and keep the terms of first order in ε. The resulting *linear* equations on X define $\mathfrak{g}$. We write the Lie algebra of a group with the same letters in lower-case Fraktur: $\mathfrak{su}(2)$, $\mathfrak{so}(3)$, $\mathfrak{u}(1)$.

Orthogonal groups. Put $R = \mathbb{1} + \varepsilon X$ in $R^T R = \mathbb{1}$:

$$\mathbb{1} \overset{(1)}{=} (\mathbb{1} + \varepsilon X)^T(\mathbb{1} + \varepsilon X) \overset{(2)}{=} \mathbb{1} + \varepsilon\,(X^T + X) + O(\varepsilon^2) \quad\Longrightarrow\quad X^T = -X.$$

(1) The defining equation of $\mathrm{O}(n)$ with $R = \mathbb{1} + \varepsilon X$.
(2) Multiply out; the $\varepsilon^2 X^T X$ term is second order. The first-order term must vanish on its own.

So $\mathfrak{so}(n) = \mathfrak{o}(n)$ is the space of real *antisymmetric* matrices. An antisymmetric $n \times n$ matrix has $\frac{1}{2}n(n-1)$ free entries, matching $\dim \mathrm{SO}(n)$.

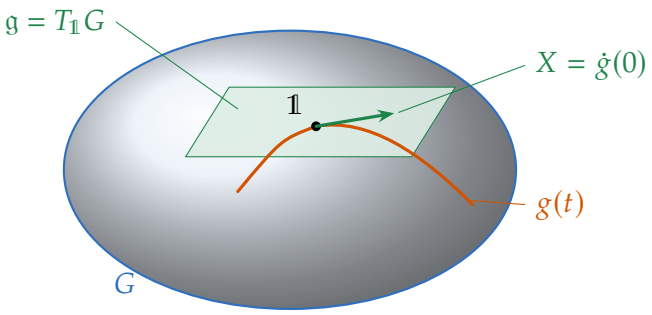


Figure T3.5: A curve $g(t)$ in the group passes through the identity at $t = 0$. Its velocity X lives in the flat tangent plane at $\mathbb{1}$. The set of all such velocities is the Lie algebra $\mathfrak{g}$. Near $\mathbb{1}$, $g(\varepsilon) \approx \mathbb{1} + \varepsilon X$.

Unitary groups. The same calculation with $U^\dagger U = \mathbb{1}$ gives $X^\dagger = -X$: $\mathfrak{u}(n)$ is the space of *anti-Hermitian* matrices, of real dimension n^2. For $\mathrm{SU}(n)$ we also need $\det U = 1$. For small ε,

$$\det(\mathbb{1} + \varepsilon X) \overset{(1)}{=} \prod_{i=1}^{n}(1 + \varepsilon\lambda_i) \overset{(2)}{=} 1 + \varepsilon \sum_i \lambda_i + O(\varepsilon^2) = 1 + \varepsilon\ \mathrm{tr}\, X + O(\varepsilon^2),$$

(1) The determinant is the product of the eigenvalues, and the eigenvalues of $\mathbb{1} + \varepsilon X$ are $1 + \varepsilon\lambda_i$ where λ_i are those of X.
(2) Expand the product to first order; the sum of eigenvalues is the trace.

so $\det U = 1$ linearises to $\mathrm{tr}\, X = 0$. Hence $\mathfrak{su}(n)$ is the space of *traceless anti-Hermitian* matrices, of dimension $n^2 - 1$. The same argument gives $\mathfrak{sl}(n) = \{\mathrm{tr}\, X = 0\}$ and $\mathfrak{gl}(n) =$ all matrices.

Symplectic groups. $M^T J M = J$ linearises to $X^T J + J X = 0$.

Group G	Lie algebra $\mathfrak{g}$	Condition on X	dim
$\mathrm{GL}(n,\mathbb{R})$	$\mathfrak{gl}(n,\mathbb{R})$	none	n^2
$\mathrm{SL}(n,\mathbb{R})$	$\mathfrak{sl}(n,\mathbb{R})$	$\mathrm{tr}\, X = 0$	$n^2 - 1$
$\mathrm{SO}(n)$	$\mathfrak{so}(n)$	$X^T = -X$	$\frac{1}{2}n(n-1)$
$\mathrm{U}(n)$	$\mathfrak{u}(n)$	$X^\dagger = -X$	n^2
$\mathrm{SU}(n)$	$\mathfrak{su}(n)$	$X^\dagger = -X$, $\mathrm{tr}\, X = 0$	$n^2 - 1$
$\mathrm{Sp}(2n,\mathbb{R})$	$\mathfrak{sp}(2n,\mathbb{R})$	$X^T J + J X = 0$	$n(2n+1)$
$\mathrm{Sp}(n)$	$\mathfrak{sp}(n)$	$X^\dagger = -X$, $X^T J + J X = 0$	$n(2n+1)$

Table T3.3: Lie algebras of the classical groups, obtained by linearising the defining equations. The dimensions agree with Table T3.1, as they must: a manifold and its tangent space have the same dimension.

T3.2.2 What survives of the product: the commutator

A vector space cannot know whether the group was commutative. U(1) and $\mathbb{R}$ both have one-dimensional Lie algebras, but so would a hypothetical non-commutative group. The information that survives is *how badly group elements fail to commute*, measured to second order. Take two curves

$$g = \mathbb{1} + \varepsilon X + \varepsilon^2 A + O(\varepsilon^3), \qquad h = \mathbb{1} + \varepsilon Y + \varepsilon^2 B + O(\varepsilon^3).$$

Their inverses are, to the same order, $g^{-1} = \mathbb{1} - \varepsilon X + \varepsilon^2(X^2 - A)$ and $h^{-1} = \mathbb{1} - \varepsilon Y + \varepsilon^2(Y^2 - B)$; multiply out $g\,g^{-1}$ to check. Now compute the **group commutator** $ghg^{-1}h^{-1}$, which is $\mathbb{1}$ exactly when g and h commute:

$$\begin{aligned} gh &\overset{(1)}{=} \mathbb{1} + \varepsilon(X+Y) + \varepsilon^2(A + B + XY), \\ g^{-1}h^{-1} &\overset{(2)}{=} \mathbb{1} - \varepsilon(X+Y) + \varepsilon^2(X^2 + Y^2 - A - B + XY), \\ ghg^{-1}h^{-1} &\overset{(3)}{=} \mathbb{1} + \varepsilon^2\big[(A + B + XY) + (X^2 + Y^2 - A - B + XY) - (X+Y)^2\big] \\ &\overset{(4)}{=} \mathbb{1} + \varepsilon^2\,(XY - YX) + O(\varepsilon^3). \end{aligned} \tag{T3.5}$$

(1) Multiply the two expansions, keeping terms up to ε^2.
(2) The same for the inverses; the cross term is $(-\varepsilon X)(-\varepsilon Y) = \varepsilon^2 XY$.
(3) Multiply $(gh)(g^{-1}h^{-1})$. The ε terms cancel. At order ε^2 we get the two ε^2 coefficients plus the product of the two first-order terms, $(X+Y)(-(X+Y))$.
(4) Expand $(X+Y)^2 = X^2 + XY + YX + Y^2$. Everything cancels except $2XY - XY - YX$. Notice that the arbitrary second-order terms A, B have dropped out.

So the first sign of non-commutativity is the **commutator** $[X,Y] = XY - YX$ of the two generators. It is again in $\mathfrak{g}$: the curve $t \mapsto g(\sqrt{t})h(\sqrt{t})g(\sqrt{t})^{-1}h(\sqrt{t})^{-1}$ lies in G and by (T3.5) has velocity $[X,Y]$ at $t = 0$. You can also check it directly; for example, for X, Y anti-Hermitian, $[X,Y]^\dagger = [Y^\dagger, X^\dagger] = [-Y, -X] = -[X,Y]$, and $\mathrm{tr}[X,Y] = 0$ always.

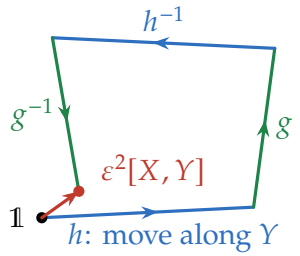


Figure T3.6: Reading right to left, $ghg^{-1}h^{-1}$ is: step along Y, step along X, step back along Y, step back along X. In a commutative group the path would close. In general it misses by $\varepsilon^2[X,Y]$: the commutator is the "gap" left by an infinitesimal rectangle. We will meet the same picture again when the curvature F appears as the failure of a small loop of parallel transports to close (Section T3.9).

Working definition T3.4 (abstract Lie algebra)

A **Lie algebra** is a real (or complex) vector space $\mathfrak{g}$ with a bracket $[\cdot,\cdot] : \mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$ [for matrices, the commutator $[X,Y] = XY - YX$] which is

1. bilinear [linear in each slot, e.g. $[aX + bY, Z] = a[X,Z] + b[Y,Z]$];
2. antisymmetric, $[X,Y] = -[Y,X]$;
3. obeys the **Jacobi identity** $[X,[Y,Z]] + [Y,[Z,X]] + [Z,[X,Y]] = 0$.

In practice: every Lie algebra we meet is a space of matrices closed under the commutator, and (1)–(3) are then automatic.

For matrices, Jacobi is a direct check:

$$\begin{aligned} [X,[Y,Z]] + \text{cyclic} &\overset{(1)}{=} (XYZ - XZY - YZX + ZYX) + (YZX - YXZ - ZXY + XZY) \\ &\quad + (ZXY - ZYX - XYZ + YXZ) \\ &\overset{(2)}{=} 0. \end{aligned}$$

(1) Expand each double commutator, $[X,[Y,Z]] = X(YZ-ZY)-(YZ-ZY)X$, and likewise for the cyclic permutations $X\to Y\to Z\to X$.
(2) Each of the six words $XYZ, XZY, \dots$ appears once with each sign.

T3.2.3 Structure constants

Choose a basis $T_1,\dots,T_N$ of $\mathfrak{g}$ ($N=\dim\mathfrak{g}$). The commutator of two basis elements is again in $\mathfrak{g}$, so it is a combination of basis elements:

$$[T_a,T_b] = f_{ab}{}^{c}\,T_c. \tag{T3.6}$$

The real numbers $f_{ab}{}^{c}$ are the **structure constants**. By bilinearity they determine every commutator: $[X,Y]=X^aY^b f_{ab}{}^{c}T_c$ for $X=X^aT_a$, $Y=Y^bT_b$. They are antisymmetric in a,b. Inserting (T3.6) into the Jacobi identity for T_a,T_b,T_c gives the component form

$$f_{bc}{}^{d}f_{ad}{}^{e}+f_{ca}{}^{d}f_{bd}{}^{e}+f_{ab}{}^{d}f_{cd}{}^{e}=0. \tag{T3.7}$$

For compact groups one can always choose the basis so that f_{abc} (index lowered with δ_{ab}) is totally antisymmetric; we will see why in Section T3.4.

Key idea: The structure constants are the group, to second order

Two groups with the same structure constants look identical near the identity. The Lie algebra captures everything *local* about the group. It misses only global information, such as whether SU(2) or SO(3) is meant (Section T3.4).

T3.2.4 The two most important examples: $\mathfrak{su}(2)$ and $\mathfrak{so}(3)$

Worked example T3.5: $\mathfrak{su}(2)$ from the Pauli matrices

A traceless anti-Hermitian 2×2 matrix is $X=-\tfrac{i}{2}(X^1\sigma_1+X^2\sigma_2+X^3\sigma_3)$ with real X^a (the Pauli matrices span the traceless Hermitian matrices, and $-i$ times Hermitian is anti-Hermitian). So a basis of $\mathfrak{su}(2)$ is

$$T_a=-\tfrac{i}{2}\sigma_a,\qquad a=1,2,3. \tag{T3.8}$$

Their commutators:

$$[T_a,T_b]\overset{(1)}{=}\left(-\tfrac{i}{2}\right)^2[\sigma_a,\sigma_b]\overset{(2)}{=}-\tfrac{1}{4}\cdot 2i\varepsilon_{abc}\sigma_c\overset{(3)}{=}\varepsilon_{abc}\left(-\tfrac{i}{2}\sigma_c\right)=\varepsilon_{abc}T_c.$$

(1) Bilinearity: numerical factors come out of the bracket.
(2) The Pauli matrices obey $\sigma_a\sigma_b=\delta_{ab}\mathbb{1}+i\varepsilon_{abc}\sigma_c$, so $[\sigma_a,\sigma_b]=2i\varepsilon_{abc}\sigma_c$. For instance $\sigma_1\sigma_2=\left(\begin{smallmatrix}0&1\\1&0\end{smallmatrix}\right)\left(\begin{smallmatrix}0&-i\\i&0\end{smallmatrix}\right)=\left(\begin{smallmatrix}i&0\\0&-i\end{smallmatrix}\right)=i\sigma_3$.
(3) Rearrange: $-\tfrac{1}{4}\cdot 2i=-\tfrac{i}{2}$.

So the structure constants of $\mathfrak{su}(2)$ in this basis are $f_{abc}=\varepsilon_{abc}$: $[T_1,T_2]=T_3$ and cyclic permutations.

Worked example T3.6: $\mathfrak{so}(3)$: rotation generators

An antisymmetric real 3×3 matrix has three free entries. A convenient basis is $(L_a)_{bc} = -\varepsilon_{abc}$:

$$L_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}, \quad L_2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}, \quad L_3 = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Each L_a acts on a vector by a cross product with the a-th axis:

$$(L_a v)_b \overset{(1)}{=} -\varepsilon_{abc} v_c \overset{(2)}{=} \varepsilon_{bac} v_c \overset{(3)}{=} (\hat{e}_a \times v)_b,$$

(1) Definition of L_a and matrix multiplication (sum over c).
(2) ε is antisymmetric under swapping two indices.
(3) The cross product in components is $(u \times v)_b = \varepsilon_{bdc} u_d v_c$, and for $u = \hat{e}_a$ only $d = a$ contributes.

so L_3 generates rotations about $\hat{z}$: $L_3(x, y, z)^T = (-y, x, 0)^T$, the velocity of a point rotating anticlockwise about the z-axis. Multiplying out,

$$L_1 L_2 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

$$L_2 L_1 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

$$[L_1, L_2] \overset{(1)}{=} \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = L_3.$$

(1) Subtract the two products and compare with L_3 above.

By cyclic symmetry $[L_a, L_b] = \varepsilon_{abc} L_c$. *These are the same structure constants as* $\mathfrak{su}(2)$. The linear map $T_a \mapsto L_a$ is an isomorphism of Lie algebras, $\mathfrak{su}(2) \cong \mathfrak{so}(3)$. The groups are nevertheless different, as we will see in Section T3.4.

T3.2.5 Mathematicians' and physicists' generators

Physicists prefer Hermitian generators, because Hermitian operators are observables. They write a group element near the identity as $U = \mathbb{1} + i\varepsilon^a J_a$ with $J_a = J_a^\dagger$, and the Lie algebra as

$$[J_a, J_b] = i\, f_{abc}\, J_c. \tag{T3.9}$$

For spin $\frac{1}{2}$, $J_a = \frac{1}{2}\sigma_a$, and (T3.9) with $f_{abc} = \varepsilon_{abc}$ is the angular momentum algebra. The two conventions are related by $T_a = -iJ_a$:

$$[T_a, T_b] \overset{(1)}{=} (-i)^2 [J_a, J_b] \overset{(2)}{=} -i f_{abc} J_c \overset{(3)}{=} f_{abc} T_c,$$

(1) Substitute $T_a = -iJ_a$.
(2) Use (T3.9): $(-1) \cdot i f_{abc} J_c$.
(3) $-iJ_c = T_c$.

so *the real structure constants are the same* in both conventions.

Dictionary: generators

Mathematics (Nash, these notes)	**Physics (QFT notes)**
$X \in \mathfrak{su}(n)$ anti-Hermitian	$J = iX$ Hermitian
$[T_a, T_b] = f_{abc} T_c$	$[J_a, J_b] = i f_{abc} J_c$
$g = e^X$, $X = \alpha^a T_a$	$U = e^{i\alpha^a J_a}$
$T_a = -\frac{i}{2}\sigma_a$	$J_a = \frac{1}{2}\sigma_a$
$\operatorname{tr}(T_a T_b) = -\frac{1}{2}\delta_{ab}$	$\operatorname{tr}(J_a J_b) = \frac{1}{2}\delta_{ab}$

Watch out: Factors of i and signs of f

Some physics texts use $U = e^{-i\alpha^a J_a}$, which flips the sign of α but not of f_{abc}. When comparing formulas, always check one commutator, such as $[J_1, J_2] = iJ_3$ or $[T_1, T_2] = T_3$, in the source you are reading. Nash himself switches to the physicists' form for Kac–Moody algebras, writing if^{abc} [Nash p. 182, (6.9)].

T3.2.6 $\mathfrak{su}(3)$ and the Gell-Mann matrices

$\mathfrak{su}(3)$ has dimension $3^2 - 1 = 8$. The physicists' Hermitian basis is $J_a = \frac{1}{2}\lambda_a$, with the **Gell-Mann matrices**

$$\lambda_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \lambda_2 = \begin{pmatrix} 0 & -i & 0 \\ i & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \lambda_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \lambda_4 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix},$$

$$\lambda_5 = \begin{pmatrix} 0 & 0 & -i \\ 0 & 0 & 0 \\ i & 0 & 0 \end{pmatrix}, \quad \lambda_6 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \quad \lambda_7 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -i \\ 0 & i & 0 \end{pmatrix}, \quad \lambda_8 = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{pmatrix}.$$

They are traceless, Hermitian, and normalised as $\operatorname{tr}(\lambda_a \lambda_b) = 2\delta_{ab}$. The first three are the Pauli matrices sitting in the upper-left 2×2 block, so they generate an SU(2) subgroup ("isospin"). The pairs (λ_4, λ_5) and (λ_6, λ_7) do the same for the $(1,3)$ and $(2,3)$ blocks. The two diagonal ones, λ_3 and λ_8, commute with each other.

The nonzero structure constants $[J_a, J_b] = i f_{abc} J_c$ are, up to antisymmetry,

$$f_{123} = 1, \quad f_{147} = f_{246} = f_{257} = f_{345} = \tfrac{1}{2}, \quad f_{156} = f_{367} = -\tfrac{1}{2}, \quad f_{458} = f_{678} = \tfrac{\sqrt{3}}{2}. \tag{T3.10}$$

Worked example T3.7: Computing two $\mathfrak{su}(3)$ structure constants

Write E_{ij} for the matrix with a 1 in row i, column j and zeros elsewhere, so $E_{ij}E_{kl} = \delta_{jk}E_{il}$. Then $\lambda_4 = E_{13} + E_{31}$ and $\lambda_5 = -iE_{13} + iE_{31}$, and

$$\lambda_4\lambda_5 \overset{(1)}{=} (E_{13} + E_{31})(-iE_{13} + iE_{31}) \overset{(2)}{=} iE_{11} - iE_{33},$$

$$\lambda_5\lambda_4 \overset{(3)}{=} -iE_{11} + iE_{33},$$

$$[\lambda_4, \lambda_5] = 2i(E_{11} - E_{33}) \overset{(4)}{=} 2i\left(\tfrac{1}{2}\lambda_3 + \tfrac{\sqrt{3}}{2}\lambda_8\right).$$

(1) Substitute.
(2) Only $E_{13}E_{31} = E_{11}$ and $E_{31}E_{13} = E_{33}$ survive; the products $E_{13}E_{13}$ and $E_{31}E_{31}$ vanish.
(3) The same computation in the other order.
(4) $\frac{1}{2}\lambda_3 + \frac{\sqrt{3}}{2}\lambda_8 = \operatorname{diag}(\frac{1}{2}, -\frac{1}{2}, 0) + \operatorname{diag}(\frac{1}{2}, \frac{1}{2}, -1) = \operatorname{diag}(1, 0, -1) = E_{11} - E_{33}$.

Dividing by 4 to convert to $J_a = \frac{1}{2}\lambda_a$: $[J_4, J_5] = i\left(\frac{1}{2}J_3 + \frac{\sqrt{3}}{2}J_8\right)$, so $f_{453} = f_{345} = \frac{1}{2}$ and $f_{458} = \frac{\sqrt{3}}{2}$, as in (T3.10). Similarly $[\lambda_1, \lambda_2] = 2i\lambda_3$ is just the Pauli algebra, giving $f_{123} = 1$.

T3.2.7 Abelian, simple and semisimple

A Lie algebra is **abelian** if all brackets vanish ($\mathfrak{u}(1)$, or $\mathbb{R}^n$). A subspace $\mathfrak{h} \subset \mathfrak{g}$ with $[\mathfrak{g}, \mathfrak{h}] \subset \mathfrak{h}$ is an **ideal**; it generates a normal subgroup. A non-abelian $\mathfrak{g}$ with no ideals other than 0 and itself is **simple** ($\mathfrak{su}(n)$, $\mathfrak{so}(n)$ for $n \neq 2, 4$, $\mathfrak{sp}(n)$, and five exceptional algebras); a direct sum of simple algebras is **semisimple**. Two decompositions come up constantly:

- $\mathfrak{u}(n) = \mathfrak{u}(1) \oplus \mathfrak{su}(n)$: split an anti-Hermitian X as $X = \frac{\operatorname{tr} X}{n}\mathbb{1} + \left(X - \frac{\operatorname{tr} X}{n}\mathbb{1}\right)$. The first piece commutes with everything. (The Standard Model gauge algebra $\mathfrak{su}(3) \oplus \mathfrak{su}(2) \oplus \mathfrak{u}(1)$ is of this "semisimple plus abelian" kind.)
- $\mathfrak{so}(4) = \mathfrak{su}(2) \oplus \mathfrak{su}(2)$, which is behind the split of 2-forms in four dimensions into self-dual and anti-self-dual parts (see the next box).

Try it: $\mathfrak{so}(4) \cong \mathfrak{su}(2) \oplus \mathfrak{su}(2)$

Let L_a ($a = 1, 2, 3$) be the rotations of the first three coordinates of $\mathbb{R}^4$ (the matrices L_a above, padded with a zero fourth row and column), and let B_a be the rotation in the $(a, 4)$ plane, $B_a = E_{4a} - E_{a4}$. Multiplying 4×4 matrices one finds $[L_a, L_b] = \varepsilon_{abc}L_c$, $[L_a, B_b] = \varepsilon_{abc}B_c$ and $[B_a, B_b] = \varepsilon_{abc}L_c$. Show that $N_a^{\pm} = \frac{1}{2}(L_a \pm B_a)$ satisfy $[N_a^{\pm}, N_b^{\pm}] = \varepsilon_{abc}N_c^{\pm}$ and $[N_a^{+}, N_b^{-}] = 0$. *Answer:* $[N_a^{+}, N_b^{+}] = \frac{1}{4}\varepsilon_{abc}(L_c + B_c + B_c + L_c) = \varepsilon_{abc}N_c^{+}$, and $[N_a^{+}, N_b^{-}] = \frac{1}{4}([L_a, L_b] - [L_a, B_b] + [B_a, L_b] - [B_a, B_b]) = \frac{1}{4}\varepsilon_{abc}(L_c - B_c + B_c - L_c) = 0$, using $[B_a, L_b] = -[L_b, B_a] = \varepsilon_{abc}B_c$. The N^{+} and N^{-} generate the self-dual and anti-self-dual rotations of $\mathbb{R}^4$ (Chapter 8).

Reading Nash (infinite-dimensional Lie algebras)

Nash's chapter 6 applies the idea "Lie algebra = tangent space at the identity" to infinite-dimensional groups [Nash pp. 173–178]. The recipe is exactly our working definition. For the group $\mathrm{Map}(X, G)$ of G-valued functions, a curve through the identity is $g_t(x)$ with $g_0(x) = \mathbb{1}$, so its velocity is a $\mathfrak{g}$-valued function: the Lie algebra is $\mathrm{Map}(X, \mathfrak{g})$ with the pointwise commutator. For the loop group $LG = \mathrm{Map}(S^1, G)$, expanding in Fourier modes $T_n^a = T^a e^{in\theta}$ gives $[T_m^a, T_n^b] = f_{ab}{}^c T_{m+n}^c$, which Nash then centrally extends to a Kac–Moody algebra. For $\mathrm{Diff}(X)$, a curve of diffeomorphisms through the identity has a velocity vector field, so the Lie algebra is the space of vector fields with the Lie bracket.

Recap

- $\mathfrak{g} = T_{\mathbb{1}}G$: linearise the defining equations. $\mathfrak{so}(n)$ is antisymmetric, $\mathfrak{u}(n)$ anti-Hermitian, $\mathfrak{su}(n)$ traceless anti-Hermitian matrices.
- The group commutator $ghg^{-1}h^{-1} = \mathbb{1} + \varepsilon^2[X, Y] + \ldots$, so $\mathfrak{g}$ is closed under $[X, Y] = XY - YX$, which obeys Jacobi.
- Structure constants: $[T_a, T_b] = f_{ab}{}^c T_c$. For $\mathfrak{su}(2)$ with $T_a = -\frac{i}{2}\sigma_a$ and for $\mathfrak{so}(3)$ with $(L_a)_{bc} = -\varepsilon_{abc}$, $f_{abc} = \varepsilon_{abc}$.
- Physicists use $J_a = iT_a$ Hermitian, $[J_a, J_b] = if_{abc}J_c$; the f's agree.
- $\mathfrak{su}(3)$: Gell-Mann matrices, eight generators, two commuting diagonal ones.

T3.3 The exponential map

The Lie algebra is the set of infinitesimal transformations. But physics needs finite ones: a rotation by 90°, a time evolution over one second, a gauge transformation that winds once around a circle. How do we get from a generator X to a finite group element? The physicist's answer is to compose a huge number of tiny steps. Rotating by θ is the same as rotating N times by θ/N, and for large N each small rotation is $\mathbb{1} + \frac{\theta}{N}X$. So

$$g = \lim_{N\to\infty}\left(\mathbb{1} + \frac{\theta X}{N}\right)^N = e^{\theta X}.$$

This is the same limit that turns compound interest into e^{rt}. The matrix exponential is the bridge from the algebra back to the group. This section computes it explicitly for the groups we care about, finds out how far it reaches, and finds out what happens to products.

T3.3.1 Definition and first properties

Working definition T3.5 (matrix exponential)

For any square matrix X,

$$e^X = \exp X := \sum_{k=0}^{\infty} \frac{X^k}{k!} = \mathbb{1} + X + \tfrac{1}{2}X^2 + \tfrac{1}{6}X^3 + \cdots .$$

The series always converges, because each entry of X^k is bounded by $(n\|X\|)^k$ with $\|X\|$ the largest entry, and $\sum c^k/k! = e^c$ is finite. *How to compute it in practice:* find a polynomial relation that X satisfies (for example $X^2 = -\mathbb{1}$ or $X^3 = -X$), use it to reduce every power X^k to a few low powers, and sum the resulting scalar series, which are usually cos, sin, cosh, sinh. Alternatively, diagonalise [possible when X has n independent eigenvectors, which form the columns of P]: if $X = P\,\mathrm{diag}(\lambda_i)P^{-1}$ then $e^X = P\,\mathrm{diag}(e^{\lambda_i})P^{-1}$.

The diagonalisation recipe works because conjugation passes through every power, $(PDP^{-1})^k = PD^kP^{-1}$, and hence through the whole series. The same argument gives a formula we shall use over and over:

$$g\,e^X g^{-1} = e^{gXg^{-1}} \qquad \text{for any invertible } g. \tag{T3.11}$$

One-parameter subgroups. Fix X and let t vary. Since sX and tX commute, the exponential series multiplies like the ordinary exponential (the binomial theorem needs only commutativity):

$$e^{sX}e^{tX} = e^{(s+t)X}, \qquad e^{0X} = \mathbb{1}, \qquad \left(e^{tX}\right)^{-1} = e^{-tX}. \tag{T3.12}$$

So $t \mapsto e^{tX}$ is a homomorphism from the additive group $\mathbb{R}$ into matrices. It is called the **one-parameter subgroup** generated by X. Differentiating the series term by term,

$$\frac{\mathrm{d}}{\mathrm{d}t}e^{tX} \overset{(1)}{=} \sum_{k=1}^{\infty}\frac{k\,t^{k-1}X^k}{k!} \overset{(2)}{=} X\sum_{k=1}^{\infty}\frac{t^{k-1}X^{k-1}}{(k-1)!} \overset{(3)}{=} X\,e^{tX} = e^{tX}X. \tag{T3.13}$$

(1) Differentiate t^k; the $k = 0$ term is constant and drops out.
(2) Cancel k against $k!$ and pull out one factor of X.
(3) Relabel $k - 1 \to k$; the remaining sum is e^{tX} again. X commutes with every power of itself, so it can stand on either side.

So $g(t) = e^{tX}$ is *the* solution of the linear ODE $\dot{g} = Xg$ with $g(0) = \mathbb{1}$. Its velocity at $t = 0$ is X, as it should be.

Physical picture: Exponentials are everywhere in physics

Time evolution in quantum mechanics is $U(t) = e^{-iHt}$, the solution of $i\dot{U} = HU$. A rotation of a quantum state by θ about $\hat{n}$ is $e^{-i\theta\,\hat{n}\cdot\vec{J}}$. A spatial translation is $e^{-ia\hat{p}}$. In each case a Hermitian generator (an observable) exponentiates to a unitary one-parameter group. In the mathematicians' convention we absorb the $-i$ into an anti-Hermitian generator $X = -iH$ and just write e^{tX}.

The exponential lands in the group. We should check that $e^X \in G$ whenever $X \in \mathfrak{g}$. For the unitary groups:

$$\left(e^X\right)^\dagger e^X \overset{(1)}{=} e^{X^\dagger} e^X \overset{(2)}{=} e^{-X} e^X \overset{(3)}{=} \mathbb{1}.$$

(1) Take the dagger of the series term by term: $(X^k)^\dagger = (X^\dagger)^k$.
(2) $X \in \mathfrak{u}(n)$ is anti-Hermitian.
(3) $-X$ and X commute, so (T3.12) applies with $s = -1, t = 1$.

The orthogonal case is identical with T in place of $\dagger$. For the determinant condition we need one more formula.

T3.3.2 $\det e^X = e^{\operatorname{tr} X}$

Key idea: Determinant of an exponential

For every square matrix X,

$$\det e^X = e^{\operatorname{tr} X}. \tag{T3.14}$$

In particular, traceless generators exponentiate to matrices of determinant one: $e^{\mathfrak{su}(n)} \subset \mathrm{SU}(n)$ and $e^{\mathfrak{sl}(n)} \subset \mathrm{SL}(n)$.

Here is a derivation that works for every X, diagonalisable or not. Let $f(t) = \det e^{tX}$. Then

$$f'(t) \overset{(1)}{=} \frac{\mathrm{d}}{\mathrm{d}s}\Big|_{s=0} \det\left(e^{tX}e^{sX}\right) \overset{(2)}{=} \det\left(e^{tX}\right) \frac{\mathrm{d}}{\mathrm{d}s}\Big|_{s=0} \det e^{sX}$$
$$\overset{(3)}{=} f(t) \frac{\mathrm{d}}{\mathrm{d}s}\Big|_{s=0} \det(\mathbb{1} + sX + O(s^2)) \overset{(4)}{=} f(t)\ \operatorname{tr} X.$$

(1) Write $e^{(t+s)X} = e^{tX}e^{sX}$ using (T3.12), and differentiate in s at $s = 0$.
(2) The determinant of a product is the product of the determinants; the first factor does not depend on s.
(3) Expand e^{sX} to first order.
(4) We showed in Section T3.2 that $\det(\mathbb{1} + sX) = 1 + s \operatorname{tr} X + O(s^2)$.

So $f' = (\operatorname{tr} X) f$ with $f(0) = 1$, whose unique solution is $f(t) = e^{t \operatorname{tr} X}$. Setting $t = 1$ gives (T3.14).

Worked example T3.8: The $\mathrm{U}(1)$ part of a unitary matrix

Any $X \in \mathfrak{u}(n)$ has purely imaginary trace (its eigenvalues are imaginary). Write $\operatorname{tr} X = i\varphi$. Then $\det e^X = e^{i\varphi}$. The overall phase of the unitary matrix is carried entirely by the trace of the generator: this is the $\mathfrak{u}(n) = \mathfrak{u}(1) \oplus \mathfrak{su}(n)$ split of Section T3.2, seen at the level of the

group. For example, with $X = \mathrm{diag}(i\alpha, i\beta)$, $e^X = \mathrm{diag}(e^{i\alpha}, e^{i\beta})$ and $\det e^X = e^{i(\alpha+\beta)} = e^{\mathrm{tr}\, X}$.

T3.3.3 U(1), SO(2) and boosts

Worked example T3.9: Exponentiating the SO(2) generator

The Lie algebra $\mathfrak{so}(2)$ is spanned by $L = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$. Multiplying, $L^2 = -\mathbb{1}$, so $L^{2k} = (-1)^k \mathbb{1}$ and $L^{2k+1} = (-1)^k L$. Then

$$e^{\alpha L} \overset{(1)}{=} \sum_k \frac{\alpha^{2k}(-1)^k}{(2k)!}\,\mathbb{1} + \sum_k \frac{\alpha^{2k+1}(-1)^k}{(2k+1)!}\,L$$
$$\overset{(2)}{=} \cos\alpha\ \mathbb{1} + \sin\alpha\ L = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix} = R(\alpha).$$

(1) Split the series into even and odd powers and reduce them with $L^2 = -\mathbb{1}$.
(2) Recognise the Taylor series of cos and sin.

This is Euler's formula $e^{i\alpha} = \cos\alpha + i\sin\alpha$ in disguise: L plays the role of i, since $L^2 = -\mathbb{1}$. In U(1) itself the generator is the 1×1 matrix i and $e^{i\alpha}$ is literally the exponential.

Notice that $e^{2\pi L} = \mathbb{1}$. The exponential map $\mathfrak{u}(1) = i\mathbb{R} \to \mathrm{U}(1)$ wraps the line around the circle infinitely many times (Figure T3.7). Its *kernel*, the set of generators that map to the identity, is the lattice $2\pi i\mathbb{Z}$. This innocent-looking lattice is behind charge quantisation: a U(1) gauge transformation $e^{i\alpha(x)}$ on a circle must come back to itself, so α can jump by $2\pi n$ but the winding number n must be an integer.

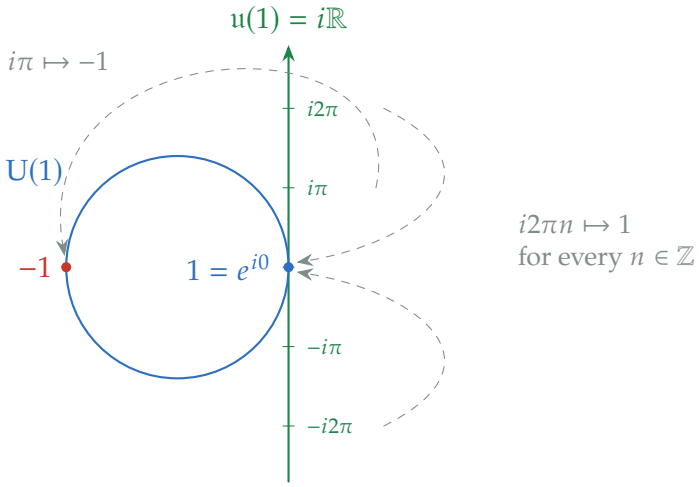


Figure T3.7: The exponential map $i\alpha \mapsto e^{i\alpha}$ rolls the Lie algebra (the green tangent line at 1) around the group circle. Every point of the circle is reached, infinitely often: all the generators $2\pi i n$ land on the identity. The integer n that is lost is exactly the winding number that labels maps $S^1 \to \mathrm{U}(1)$.

Worked example T3.10: Lorentz boosts: rapidities add

A boost along x acts on (t, x) by the generator $K = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, which preserves $-t^2 + x^2$ to first

order. Now $K^2 = +\mathbb{1}$, so

$$e^{\eta K} \overset{(1)}{=} \Big(\sum_k \frac{\eta^{2k}}{(2k)!}\Big)\mathbb{1} + \Big(\sum_k \frac{\eta^{2k+1}}{(2k+1)!}\Big)K \overset{(2)}{=} \begin{pmatrix} \cosh\eta & \sinh\eta \\ \sinh\eta & \cosh\eta \end{pmatrix}.$$

(1) Reduce the powers using $K^2 = \mathbb{1}$ (no alternating signs this time).
(2) The even and odd parts of the exponential series are cosh and sinh.

With $\cosh\eta = \gamma$ and $\sinh\eta = \gamma v$ this is the familiar boost matrix, and (T3.12) says that *rapidities add*: $e^{\eta_1 K} e^{\eta_2 K} = e^{(\eta_1+\eta_2)K}$. The velocity-addition formula $v = \frac{v_1+v_2}{1+v_1 v_2}$ is just $\tanh(\eta_1+\eta_2)$ written out. The one-parameter group of boosts never returns to the identity: this is the non-compact cousin of the rotation group.

T3.3.4 The SU(2) exponential

This is the calculation we use most. Let $\hat{n}$ be a unit vector and θ a real number. With the basis $T_a = -\frac{i}{2}\sigma_a$ of (T3.8), a general element of $\mathfrak{su}(2)$ is

$$X = \theta\, \hat{n}\cdot\vec{T} = -\tfrac{i\theta}{2}\, \hat{n}\cdot\vec{\sigma}.$$

The key fact is that $\hat{n}\cdot\vec{\sigma}$ squares to the identity:

$$(\hat{n}\cdot\vec{\sigma})^2 \overset{(1)}{=} n_a n_b\, \sigma_a\sigma_b \overset{(2)}{=} n_a n_b\big(\delta_{ab}\mathbb{1} + i\varepsilon_{abc}\sigma_c\big) \overset{(3)}{=} |\hat{n}|^2\,\mathbb{1} = \mathbb{1}.$$

(1) Expand the product (sum over a and b).
(2) The Pauli product rule.
(3) $n_a n_b$ is symmetric in a, b and ε_{abc} is antisymmetric, so their contraction vanishes.

Hence $X^2 = -\frac{\theta^2}{4}\mathbb{1}$, and exactly as for SO(2) the series splits into cos and sin:

$$\begin{aligned} e^{-\frac{i\theta}{2}\hat{n}\cdot\vec{\sigma}} &\overset{(1)}{=} \sum_k \frac{(-1)^k(\theta/2)^{2k}}{(2k)!}\mathbb{1} + \sum_k \frac{(-1)^k(\theta/2)^{2k+1}}{(2k+1)!}\big(-i\hat{n}\cdot\vec{\sigma}\big) \\ &\overset{(2)}{=} \cos\tfrac{\theta}{2}\,\mathbb{1} - i\sin\tfrac{\theta}{2}\,\hat{n}\cdot\vec{\sigma} \\ &= \begin{pmatrix} \cos\frac{\theta}{2} - in_3\sin\frac{\theta}{2} & -(in_1+n_2)\sin\frac{\theta}{2} \\ -(in_1-n_2)\sin\frac{\theta}{2} & \cos\frac{\theta}{2} + in_3\sin\frac{\theta}{2} \end{pmatrix}. \end{aligned} \tag{T3.15}$$

(1) Reduce $X^{2k} = (-1)^k(\theta/2)^{2k}\mathbb{1}$ and $X^{2k+1} = (-1)^k(\theta/2)^{2k}X$, and write $X = (\theta/2)(-i\hat{n}\cdot\vec{\sigma})$.
(2) Recognise the cos and sin series in the variable $\theta/2$.

Compare with the parametrisation $U = x^4\mathbb{1} + i\vec{x}\cdot\vec{\sigma}$ of $\mathrm{SU}(2) = S^3$ in (T3.2):

$$x^4 = \cos\tfrac{\theta}{2}, \qquad \vec{x} = -\sin\tfrac{\theta}{2}\,\hat{n}. \tag{T3.16}$$

Indeed $(x^4)^2 + |\vec{x}|^2 = \cos^2 + \sin^2 = 1$. Three consequences follow at once.

1. *The exponential is onto.* Given any $U \in \mathrm{SU}(2)$, i.e. any point $(\vec{x}, x^4)$ of S^3, choose $\theta \in [0, 2\pi]$ with $\cos\frac{\theta}{2} = x^4$, and $\hat{n} = -\vec{x}/|\vec{x}|$ (any $\hat{n}$ if $\vec{x} = 0$). Then $U = e^{\theta\hat{n}\cdot\vec{T}}$.

2. *The angle θ is the conjugacy class.* The eigenvalues are $e^{\mp i\theta/2}$ and $\operatorname{tr} U = 2\cos\frac{\theta}{2}$, in agreement with Example T3.3.

3. *A 2π rotation is $-\mathbb{1}$.* Setting $\theta = 2\pi$ gives $e^{-i\pi\hat{n}\cdot\vec{\sigma}} = \cos\pi\,\mathbb{1} = -\mathbb{1}$ for *every* axis $\hat{n}$. Only $\theta = 4\pi$ gives back $\mathbb{1}$. We will see in Section T3.4 that θ is the rotation angle of the corresponding element of SO(3). So a spin-$\frac{1}{2}$ wavefunction changes sign under a 360° rotation.

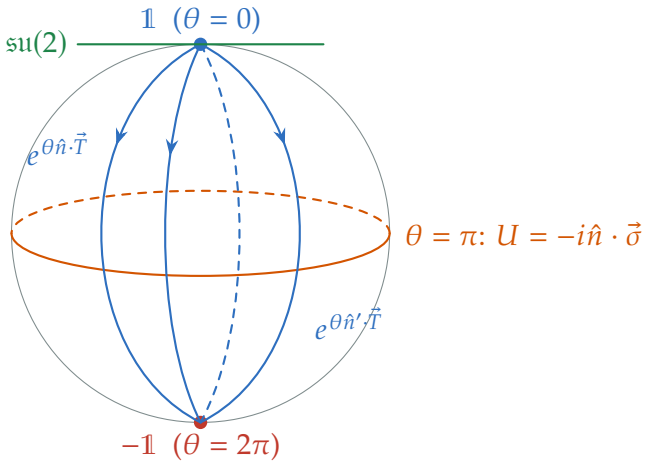


Figure T3.8: SU(2) = S^3 drawn as in Figure T3.3. Each one-parameter subgroup $\theta \mapsto e^{\theta\hat{n}\cdot\vec{T}}$ is a great circle that leaves the identity in the direction $\hat{n}$ (a point of the green tangent space), crosses the equatorial 2-sphere of elements $-i\hat{n}\cdot\vec{\sigma}$ at $\theta = \pi$, and arrives at $-\mathbb{1}$ at $\theta = 2\pi$, *whatever the direction* $\hat{n}$. It then returns to $\mathbb{1}$ at $\theta = 4\pi$ along the opposite half of the same great circle.

Worked example T3.11: A spin-$\frac{1}{2}$ rotation about $\hat{z}$

For $\hat{n} = \hat{z}$, (T3.15) is diagonal: $e^{-\frac{i\theta}{2}\sigma_3} = \mathrm{diag}(e^{-i\theta/2}, e^{i\theta/2})$. Acting on the spin-up state it gives $|\uparrow\rangle \mapsto e^{-i\theta/2}|\uparrow\rangle$. In the physicists' language this is $e^{-i\theta J_3}$ with $J_3 = \frac{1}{2}\sigma_3$ and eigenvalue $m = +\frac{1}{2}$. The phase $e^{-im\theta}$ is the general rule, and $m = \pm\frac{1}{2}$ is what makes $\theta = 2\pi$ give -1. This sign has been measured directly: rotating a beam of neutrons by 2π in a magnetic field and interfering it with an unrotated beam flips the interference pattern (Rauch et al. and Werner et al., 1975).

T3.3.5 Rotations: the Rodrigues formula

For SO(3) we exponentiate θK with $K = \hat{n}\cdot\vec{L}$. From Example T3.6, K acts on a vector by $Kv = \hat{n}\times v$. We need the polynomial relation that K satisfies:

$$K^2 v \overset{(1)}{=} \hat{n}\times(\hat{n}\times v) \overset{(2)}{=} \hat{n}(\hat{n}\cdot v) - v,$$
$$K^3 v \overset{(3)}{=} \hat{n}\times\big(\hat{n}(\hat{n}\cdot v) - v\big) \overset{(4)}{=} -\hat{n}\times v = -Kv.$$

(1) Apply K twice.
(2) The "BAC–CAB" rule $a\times(b\times c) = b(a\cdot c) - c(a\cdot b)$ and $\hat{n}\cdot\hat{n} = 1$.
(3) Apply K to the previous line.
(4) $\hat{n}\times\hat{n} = 0$.

So $K^3 = -K$, hence $K^{2k+1} = (-1)^k K$ and $K^{2k} = (-1)^{k+1}K^2$ for $k \geq 1$. Therefore

$$e^{\theta K} \overset{(1)}{=} \mathbb{1} + \Big(\theta - \frac{\theta^3}{3!} + \cdots\Big)K + \Big(\frac{\theta^2}{2!} - \frac{\theta^4}{4!} + \cdots\Big)K^2$$
$$\overset{(2)}{=} \mathbb{1} + \sin\theta\, K + (1-\cos\theta)\, K^2. \tag{T3.17}$$

(1) Separate the $k = 0$ term, the odd powers and the even powers ≥ 2, and reduce them.
(2) $\theta - \theta^3/3! + \cdots = \sin\theta$ and $\theta^2/2! - \theta^4/4! + \cdots = 1 - \cos\theta$.

Acting on a vector and using the formulas for Kv and $K^2 v$,

$$e^{\theta K} v = \cos\theta\, v + \sin\theta\, (\hat{n} \times v) + (1 - \cos\theta)\, \hat{n}(\hat{n} \cdot v). \tag{T3.18}$$

This is the **Rodrigues rotation formula**: the component of v along $\hat{n}$ is untouched, and the perpendicular part rotates by θ in the plane perpendicular to $\hat{n}$.

Worked example T3.12: Rodrigues reproduces $R_z(\theta)$

Take $\hat{n} = \hat{z}$, $K = L_3$. Then $K^2 = L_3^2 = \mathrm{diag}(-1, -1, 0)$ and

$$e^{\theta L_3} \overset{(1)}{=} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} + \sin\theta \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} + (1 - \cos\theta) \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$
$$= \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \overset{(2)}{=} R_z(\theta).$$

(1) Insert $K = L_3$ into (T3.17).
(2) Compare with (T3.4).

The trace of (T3.17) for general $\hat{n}$ is $3 + 0 + (1 - \cos\theta)\,\mathrm{tr}\, K^2 = 3 - 2(1 - \cos\theta) = 1 + 2\cos\theta$, using $\mathrm{tr}\, K^2 = -2$. This confirms that θ is the rotation angle. Unlike the SU(2) case, $e^{2\pi K} = \mathbb{1}$.

T3.3.6 Products of exponentials: Baker–Campbell–Hausdorff

For numbers $e^x e^y = e^{x+y}$. For matrices this fails as soon as X and Y do not commute, and the correction is controlled by the commutator.

Key idea: BCH formula

For small X, Y,

$$e^X e^Y = \exp\Big(X + Y + \tfrac{1}{2}[X,Y] + \tfrac{1}{12}[X,[X,Y]] - \tfrac{1}{12}[Y,[X,Y]] + \cdots \Big), \tag{T3.19}$$

where every omitted term is a nested commutator. So the group product near the identity is *determined by the bracket*: this is the precise sense in which the Lie algebra knows the group locally.

We check the first two orders. Expand both sides up to terms of total degree two in X, Y:

$$e^X e^Y \overset{(1)}{=} \big(\mathbb{1} + X + \tfrac{1}{2}X^2\big)\big(\mathbb{1} + Y + \tfrac{1}{2}Y^2\big) = \mathbb{1} + X + Y + \tfrac{1}{2}X^2 + XY + \tfrac{1}{2}Y^2,$$
$$e^{X+Y+\frac{1}{2}[X,Y]} \overset{(2)}{=} \mathbb{1} + X + Y + \tfrac{1}{2}[X,Y] + \tfrac{1}{2}(X+Y)^2$$
$$\overset{(3)}{=} \mathbb{1} + X + Y + \tfrac{1}{2}(XY - YX) + \tfrac{1}{2}(X^2 + XY + YX + Y^2)$$
$$= \mathbb{1} + X + Y + \tfrac{1}{2}X^2 + XY + \tfrac{1}{2}Y^2.$$

(1) Truncate each series at second order and multiply.
(2) Truncate $e^Z = \mathbb{1} + Z + \frac{1}{2}Z^2$ with $Z = X + Y + \frac{1}{2}[X,Y]$; in Z^2 only $(X+Y)^2$ is of degree two.
(3) Expand the commutator and the square, keeping the order of factors.

The two agree. Without the $\frac{1}{2}[X,Y]$ term the XY coefficient would be $\frac{1}{2}$ on the right instead of 1. Two consequences:

- If $[X,Y]=0$ then $e^X e^Y = e^{X+Y}$ exactly. (Every nested commutator vanishes.)
- If $[X,Y]$ commutes with both X and Y (as for $\hat{x}$ and $\hat{p}$ in quantum mechanics, $[\hat{x},\hat{p}]=i$), the series stops: $e^X e^Y = e^{X+Y+\frac{1}{2}[X,Y]}$. This is the formula behind coherent states and the Weyl form of the canonical commutation relations.

Try it: Group commutator from BCH

Use (T3.19) to show that $e^{\varepsilon X}e^{\varepsilon Y}e^{-\varepsilon X}e^{-\varepsilon Y} = e^{\varepsilon^2[X,Y]+O(\varepsilon^3)}$, reproducing (T3.5).
Answer: $e^{\varepsilon X}e^{\varepsilon Y} = e^{\varepsilon(X+Y)+\frac{\varepsilon^2}{2}[X,Y]+\cdots}$ and $e^{-\varepsilon X}e^{-\varepsilon Y} = e^{-\varepsilon(X+Y)+\frac{\varepsilon^2}{2}[X,Y]+\cdots}$. Combining these two with BCH once more, the ε terms cancel, the commutator of the two exponents is $O(\varepsilon^3)$, and the ε^2 terms add to $\varepsilon^2[X,Y]$.

T3.3.7 How far does the exponential reach?

For U(1), SU(2) and SO(3) we found that every group element is an exponential. This is true for every *compact connected* Lie group: every element can be conjugated into a maximal torus (a maximal commuting subgroup, such as the diagonal matrices in $\mathrm{U}(n)$), where it is $\mathrm{diag}(e^{i\varphi_1},\dots) = e^{\mathrm{diag}(i\varphi_1,\dots)}$, and conjugation commutes with exp by (T3.11). For non-compact groups it can fail, even when the group is connected.

Worked example T3.13: $\mathrm{SL}(2,\mathbb{R})$: an element that is not an exponential

Let $X \in \mathfrak{sl}(2,\mathbb{R})$, a real traceless 2×2 matrix. For 2×2 matrices the Cayley–Hamilton theorem says $X^2 - (\operatorname{tr}X)X + (\det X)\mathbb{1} = 0$, so $X^2 = \delta\,\mathbb{1}$ with $\delta = -\det X$ real. Three cases:

$$\begin{aligned}
\delta = s^2 > 0: &\quad e^X \overset{(1)}{=} \cosh s\,\mathbb{1} + \tfrac{\sinh s}{s}X, &&\operatorname{tr}e^X \overset{(2)}{=} 2\cosh s \;\geq 2,\\
\delta = 0: &\quad e^X = \mathbb{1} + X, &&\operatorname{tr}e^X = 2,\\
\delta = -s^2 < 0: &\quad e^X \overset{(3)}{=} \cos s\,\mathbb{1} + \tfrac{\sin s}{s}X, &&\operatorname{tr}e^X = 2\cos s \;\geq -2.
\end{aligned}$$

(1) As for boosts: $X^{2k} = s^{2k}\mathbb{1}$, $X^{2k+1} = s^{2k}X$.
(2) $\operatorname{tr}X = 0$, so only the $\mathbb{1}$ term contributes to the trace.
(3) As for rotations: $X^{2k} = (-1)^k s^{2k}\mathbb{1}$.

So *every* exponential has $\operatorname{tr}e^X \geq -2$. But $M = \mathrm{diag}(-2,-\frac{1}{2})$ has $\det M = 1$, so $M \in \mathrm{SL}(2,\mathbb{R})$, and $\operatorname{tr}M = -\frac{5}{2} < -2$. Hence M is not e^X for any X. (It is a product of two exponentials, $M = e^{\pi L}\,e^{\ln 2\,\sigma_3}$, since $e^{\pi L} = -\mathbb{1}$.)

Watch out: The exponential is neither one-to-one nor always onto

Even for compact groups, exp is not injective: $e^{2\pi L} = e^0$ in SO(2), and all of $\{4\pi\hat{n}\cdot\vec{T}\}$ goes to $\mathbb{1}$ in SU(2). It is a diffeomorphism only from a small neighbourhood of 0 in $\mathfrak{g}$ onto a neighbourhood of $\mathbb{1}$ in G. This is why a local calculation (in $\mathfrak{g}$) cannot see global facts such as π_1.

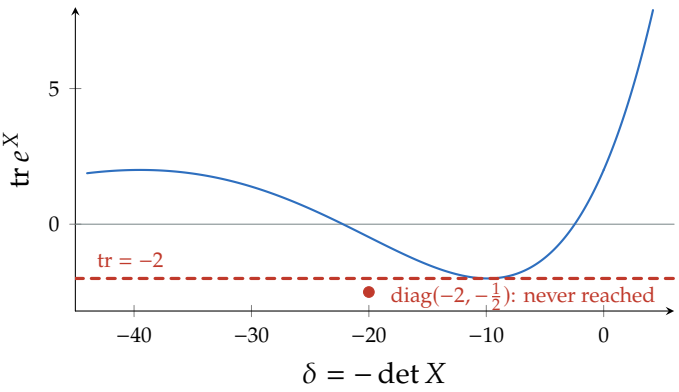


Figure T3.9: The trace of e^X for $X \in \mathfrak{sl}(2,\mathbb{R})$ depends only on $\delta = -\det X$. For $\delta < 0$ (elliptic, rotation-like X) it oscillates between -2 and 2; for $\delta > 0$ (hyperbolic, boost-like X) it grows like $2\cosh\sqrt{\delta}$. It never drops below -2, so group elements with trace below -2 (red dot) are not exponentials.

Reading Nash (where the exponential hides)

Nash rarely writes exp explicitly, but it is behind two of his formulas. The parallel transport equation $\frac{dg}{dt} + A(\dot\gamma)\,g = 0$ [Nash p. 217, (8.4)] is the ODE (T3.13) with the constant generator replaced by a time-dependent one; when $A(\dot\gamma)$ is constant along the path its solution is $g(t) = e^{-tA(\dot\gamma)}g(0)$. We develop this in Section T3.9. And the Fourier modes $T^a e^{in\theta}$ of the loop algebra [Nash p. 182] exponentiate to loops $e^{\alpha T^a e^{in\theta}}$ in LG; for U(1) the loop $\theta \mapsto e^{in\theta}$ is exactly the winding-n loop of Figure T3.7.

Recap

- $e^X = \sum X^k/k!$; $t \mapsto e^{tX}$ is the one-parameter subgroup solving $\dot g = Xg$, $g(0) = \mathbb{1}$.
- $\det e^X = e^{\operatorname{tr} X}$; $ge^Xg^{-1} = e^{gXg^{-1}}$.
- SO(2): $e^{\alpha L} = \cos\alpha + L\sin\alpha$. Boosts: cosh, sinh, rapidities add.
- SU(2): $e^{-\frac{i\theta}{2}\hat n\cdot\vec\sigma} = \cos\frac{\theta}{2} - i\sin\frac{\theta}{2}\,\hat n\cdot\vec\sigma$; onto S^3; $\theta = 2\pi$ gives $-\mathbb{1}$.
- SO(3): Rodrigues $e^{\theta K} = \mathbb{1} + \sin\theta K + (1-\cos\theta)K^2$.
- BCH: $e^Xe^Y = e^{X+Y+\frac12[X,Y]+\cdots}$; the bracket fixes the product locally.
- exp is onto for compact connected groups, but not for $\mathrm{SL}(2,\mathbb{R})$.

T3.4 The adjoint representation and the double cover SU(2) → SO(3)

A rotation acts on vectors. But it also acts on *other rotations*: if h is a rotation about the z-axis and g is some other rotation, then ghg^{-1} is a rotation by the same angle about the rotated axis $g\hat z$. Infinitesimally, a rotation generator is an angular-momentum vector, and a rotation turns one angular-momentum vector into another. So every Lie group acts on its own Lie algebra. This action is the *adjoint representation*. It is the single most important representation in gauge theory, because gauge fields and field strengths transform in it ($F \mapsto g^{-1}Fg$). For SU(2) it produces the famous two-to-one map onto SO(3), which is why electrons have spin $\frac12$ and why a 360° rotation of a spinor gives a minus sign.

T3.4.1 Ad: the group acting on its algebra

Working definition T3.6 (adjoint action)

For $g \in G$ and $X \in \mathfrak{g}$ [a generator, i.e. the velocity $\dot{h}(0)$ of a curve $h(t)$ in G with $h(0) = \mathbb{1}$] define

$$\mathrm{Ad}_g X := g X g^{-1}.$$

It is again in $\mathfrak{g}$: the curve $t \mapsto g\, e^{tX} g^{-1}$ lies in G and by (T3.11) equals $e^{t\, gXg^{-1}}$, whose velocity at $t = 0$ is gXg^{-1}. *How to compute its matrix:* choose a basis T_a, conjugate each basis element, and expand the result in the basis,

$$g\, T_a\, g^{-1} = T_b\, (\mathrm{Ad}_g)_{ba}. \tag{T3.20}$$

The numbers $(\mathrm{Ad}_g)_{ba}$ form a $\dim G \times \dim G$ matrix.

Ad is a **representation**, meaning that it turns the group product into matrix multiplication:

$$\mathrm{Ad}_{gh} X \overset{(1)}{=} (gh)X(gh)^{-1} \overset{(2)}{=} g\big(hXh^{-1}\big)g^{-1} \overset{(3)}{=} \mathrm{Ad}_g\big(\mathrm{Ad}_h X\big).$$

(1) Definition.
(2) $(gh)^{-1} = h^{-1}g^{-1}$ and associativity.
(3) Definition twice.

It also preserves brackets, $\mathrm{Ad}_g[X,Y] = g(XY - YX)g^{-1} = [gXg^{-1}, gYg^{-1}]$, because the $g^{-1}g$ in the middle cancels.

T3.4.2 ad: the algebra acting on itself

Differentiate Ad along a one-parameter subgroup $g = e^{tY}$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} e^{tY} X e^{-tY} \overset{(1)}{=} YX - XY = [Y, X].$$

(1) Product rule with $\frac{\mathrm{d}}{\mathrm{d}t} e^{\pm tY}|_0 = \pm Y$, from (T3.13).

Working definition T3.7 (adjoint representation of the algebra)

$\mathrm{ad}_Y X := [Y, X]$. In a basis, the matrix of ad_{T_a} [a $\dim G \times \dim G$ matrix acting on the components of X] is built from the structure constants:

$$\mathrm{ad}_{T_a} T_b = [T_a, T_b] = f_{ab}{}^c T_c, \qquad \text{so} \qquad (\mathrm{ad}_{T_a})_{cb} = f_{ab}{}^c. \tag{T3.21}$$

That ad respects brackets, $\mathrm{ad}_{[X,Y]} = [\mathrm{ad}_X, \mathrm{ad}_Y]$, is *exactly* the Jacobi identity:

$$[\mathrm{ad}_X, \mathrm{ad}_Y]Z \overset{(1)}{=} [X,[Y,Z]] - [Y,[X,Z]] \overset{(2)}{=} [X,[Y,Z]] + [Y,[Z,X]]$$
$$\overset{(3)}{=} -[Z,[X,Y]] \overset{(4)}{=} [[X,Y],Z] = \mathrm{ad}_{[X,Y]} Z.$$

(1) Definition of the commutator of two linear maps.
(2) Antisymmetry of the inner bracket.
(3) Jacobi: the three cyclic terms sum to zero.
(4) Antisymmetry once more.

The physicists' version of (T3.21) is $(R^b_A)_{ac} = -if^{bac}$ in [QFT notes (31.2.9)], up to their index placement and the factor of i from Hermitian generators.

Key idea: $\mathrm{Ad} = e^{\mathrm{ad}}$

Conjugation by an exponential is the exponential of the commutator:

$$e^Y X e^{-Y} = e^{\mathrm{ad}_Y} X = X + [Y,X] + \tfrac{1}{2!}[Y,[Y,X]] + \tfrac{1}{3!}[Y,[Y,[Y,X]]] + \cdots . \tag{T3.22}$$

Why: let $Z(t) = e^{tY} X e^{-tY}$. Then $\dot{Z} = YZ - ZY = \mathrm{ad}_Y Z$, a linear ODE with constant coefficients and $Z(0) = X$. Its solution is $Z(t) = e^{t\,\mathrm{ad}_Y} X$, by (T3.13) applied to the linear map ad_Y. Set $t = 1$. This is the formula the QFT notes work out order by order in [QFT notes (19.3.16)].

Worked example T3.14: The adjoint representation of $\mathfrak{su}(2)$ is $\mathfrak{so}(3)$

With $f_{ab}{}^c = \varepsilon_{abc}$, (T3.21) gives $(\mathrm{ad}_{T_a})_{cb} = \varepsilon_{abc}$. Compare with the rotation generators of Example T3.6, $(L_a)_{cb} = -\varepsilon_{acb} = \varepsilon_{abc}$. They are equal:

$$\mathrm{ad}_{T_a} = L_a .$$

The adjoint representation of $\mathfrak{su}(2)$ is the vector representation of $\mathfrak{so}(3)$. In physics language: *the generators of isospin themselves carry isospin 1.* The $W^\pm, W^3$ bosons of SU(2) gauge theory form an isospin triplet; the eight gluons of SU(3) form an octet, the adjoint of SU(3).

Worked example T3.15: Using (T3.22) to rotate a Pauli matrix

Take $Y = -\frac{i\theta}{2}\sigma_3 = \theta T_3$ and $X = \sigma_1$. Then $\mathrm{ad}_Y \sigma_1 = -\frac{i\theta}{2}[\sigma_3, \sigma_1] = -\frac{i\theta}{2}(2i\sigma_2) = \theta\sigma_2$ and $\mathrm{ad}_Y \sigma_2 = -\frac{i\theta}{2}[\sigma_3, \sigma_2] = -\frac{i\theta}{2}(-2i\sigma_1) = -\theta\sigma_1$. So ad_Y acts on the pair (σ_1, σ_2) like the SO(2) generator, and

$$e^{-\frac{i\theta}{2}\sigma_3}\,\sigma_1\, e^{\frac{i\theta}{2}\sigma_3} \overset{(1)}{=} \sigma_1 + \theta\sigma_2 - \tfrac{\theta^2}{2!}\sigma_1 - \tfrac{\theta^3}{3!}\sigma_2 + \tfrac{\theta^4}{4!}\sigma_1 + \cdots$$
$$\overset{(2)}{=} \cos\theta\,\sigma_1 + \sin\theta\,\sigma_2 .$$

(1) Apply ad_Y repeatedly: $\sigma_1 \to \theta\sigma_2 \to -\theta^2\sigma_1 \to -\theta^3\sigma_2 \to \cdots$.
(2) Collect the cos and sin series.

The Pauli "vector" $\vec{\sigma}$ is rotated by the angle θ, not $\theta/2$. Keep this in mind for the double cover below.

T3.4.3 The Killing form: a built-in inner product

A Yang–Mills Lagrangian needs a way to square the field strength $F = F^a T_a$ and get a number that does not change under gauge transformations. We need an inner product on $\mathfrak{g}$ that is invariant under Ad. Every Lie algebra carries a canonical candidate.

Working definition T3.8 (Killing form)

The **Killing form** is

$$B(X,Y) := \mathrm{tr}\big(\mathrm{ad}_X\,\mathrm{ad}_Y\big), \qquad B_{ab} = B(T_a, T_b) = f_{ac}{}^d f_{bd}{}^c .$$

In practice: for matrix algebras it is proportional to the trace in the defining representation [the $N \times N$ matrices that define the group], $B(X,Y) = c\ \mathrm{tr}(XY)$, with $c = 2N$ for $\mathfrak{su}(N)$

(derived below). Either one serves as "the" invariant inner product [unchanged when both X and Y are conjugated by the same g]; physicists use tr in the fundamental and normalise $\operatorname{tr}(J_a J_b) = \frac{1}{2}\delta_{ab}$.

The component formula follows from (T3.21): $\operatorname{tr}(\mathrm{ad}_{T_a}\mathrm{ad}_{T_b}) = (\mathrm{ad}_{T_a})_{dc}(\mathrm{ad}_{T_b})_{cd} = f_{ac}{}^{d} f_{bd}{}^{c}$.

Invariance. Both B and $\operatorname{tr}(XY)$ are invariant, because conjugation cancels inside a trace:

$$\operatorname{tr}\big(\mathrm{Ad}_g X\, \mathrm{Ad}_g Y\big) \overset{(1)}{=} \operatorname{tr}\big(gXg^{-1}gYg^{-1}\big) \overset{(2)}{=} \operatorname{tr}\big(XYg^{-1}g\big) = \operatorname{tr}(XY). \tag{T3.23}$$

(1) Definition of Ad.
(2) Cancel $g^{-1}g$ in the middle and move g from the front to the back using cyclicity of the trace, $\operatorname{tr}(AB) = \operatorname{tr}(BA)$.

Differentiating along $g = e^{tZ}$ gives the infinitesimal form, $\operatorname{tr}([Z,X]Y) + \operatorname{tr}(X[Z,Y]) = 0$. In components (with an orthonormal basis, tr or B proportional to δ_{ab}) this says that $f_{abc} \coloneqq f_{ab}{}^{d}\,\delta_{dc}$ is antisymmetric in b and c as well as in a, b: it is *totally antisymmetric*, as promised in Section T3.2.

Worked example T3.16: Killing form of $\mathfrak{su}(N)$ is $2N$ tr

For $X, Y \in \mathfrak{su}(N)$, the map $Z \mapsto \mathrm{ad}_X\mathrm{ad}_Y Z = [X,[Y,Z]]$ can be extended to all $N \times N$ matrices Z. We need one fact: the map $Z \mapsto AZC$ on the N^2-dimensional space of matrices has trace $\operatorname{tr} A\ \operatorname{tr} C$ (check it on the basis E_{ij}: the coefficient of E_{ij} in $AE_{ij}C$ is $A_{ii}C_{jj}$). Then

$$\begin{aligned}\operatorname{tr}_{\text{all } Z}\big(\mathrm{ad}_X\mathrm{ad}_Y\big) &\overset{(1)}{=} \operatorname{tr}_Z\big(Z \mapsto XYZ - XZY - YZX + ZYX\big)\\ &\overset{(2)}{=} N\operatorname{tr}(XY) - \operatorname{tr}X\operatorname{tr}Y - \operatorname{tr}Y\operatorname{tr}X + N\operatorname{tr}(YX)\\ &\overset{(3)}{=} 2N\operatorname{tr}(XY).\end{aligned}$$

(1) Expand the double commutator.
(2) Use the fact for each term: $Z \mapsto (XY)Z\mathbb{1}$ has trace $\operatorname{tr}(XY)\operatorname{tr}\mathbb{1} = N\operatorname{tr}(XY)$, and so on.
(3) X, Y are traceless, and $\operatorname{tr}(YX) = \operatorname{tr}(XY)$.

The space of all matrices is $\mathfrak{sl}(N,\mathbb{C}) \oplus \mathbb{C}\mathbb{1}$, and ad kills the $\mathbb{1}$ direction, so this is also the trace over the Lie algebra. Hence $B(X,Y) = 2N\operatorname{tr}(XY)$. For $\mathfrak{su}(2)$: $B(T_a,T_b) = 4\operatorname{tr}(T_aT_b) = 4\cdot(-\frac{1}{2}\delta_{ab}) = -2\delta_{ab}$. Directly from $\mathrm{ad}_{T_a} = L_a$: $\operatorname{tr}(L_aL_b) = -2\delta_{ab}$ (e.g. $L_3^2 = \mathrm{diag}(-1,-1,0)$). The two computations agree.

Key idea: Compact ⇔ negative-definite Killing form

For anti-Hermitian $X \neq 0$, $\operatorname{tr}(X^2) = -\operatorname{tr}(X^\dagger X) = -\sum_{ij}|X_{ij}|^2 < 0$. So on $\mathfrak{su}(N)$, and on the Lie algebra of any compact semisimple group, $-B$ and $-\operatorname{tr}$ are *positive-definite* invariant inner products. This is why Yang–Mills actions carry a minus sign in the maths convention: $-\operatorname{tr}(F_{\mu\nu}F^{\mu\nu}) = \frac{1}{2}F^a_{\mu\nu}F^{a\,\mu\nu} \geq 0$. For a non-compact group the Killing form is indefinite: in $\mathfrak{sl}(2,\mathbb{R})$, $B(\sigma_3,\sigma_3) = 4\operatorname{tr}\sigma_3^2 = 8 > 0$ while $B(L,L) = 4\operatorname{tr}L^2 = -8 < 0$. A Yang–Mills theory with such a gauge group would have negative-energy (ghost) gluons.

T3.4.4 The double cover SU(2) → SO(3) built explicitly

Now apply Ad to SU(2). Identify a vector $\vec{x} \in \mathbb{R}^3$ with the traceless Hermitian matrix

$$\vec{x}\cdot\vec{\sigma} = \begin{pmatrix} x_3 & x_1 - ix_2 \\ x_1 + ix_2 & -x_3 \end{pmatrix}, \qquad \det(\vec{x}\cdot\vec{\sigma}) = -|\vec{x}|^2.$$

For $U \in \mathrm{SU}(2)$, the matrix $U(\vec{x}\cdot\vec{\sigma})U^\dagger$ is again Hermitian ($(UMU^\dagger)^\dagger = UM^\dagger U^\dagger$) and traceless (cyclicity), so it equals $\vec{x}\,'\cdot\vec{\sigma}$ for some new vector $\vec{x}\,'$, which depends linearly on $\vec{x}$: $\vec{x}\,' = R(U)\vec{x}$. And

$$-|R(U)\vec{x}|^2 \overset{(1)}{=} \det\big(U(\vec{x}\cdot\vec{\sigma})U^\dagger\big) \overset{(2)}{=} \det U\,\det(\vec{x}\cdot\vec{\sigma})\,\det U^\dagger \overset{(3)}{=} -|\vec{x}|^2.$$

(1) The determinant formula above, applied to $\vec{x}\,'$.
(2) Determinant of a product.
(3) $\det U = \det U^\dagger = 1$.

So $R(U)$ preserves lengths: $R(U) \in \mathrm{O}(3)$. To extract its entries, use $\mathrm{tr}(\sigma_i\sigma_k) = 2\delta_{ik}$. From $U\sigma_j U^\dagger = \sigma_k R_{kj}$, multiply by σ_i and take the trace:

Key idea: The covering map

$$R_{ij}(U) = \tfrac{1}{2}\,\mathrm{tr}\big(\sigma_i\, U\sigma_j U^\dagger\big), \qquad U\sigma_j U^\dagger = \sigma_i\, R_{ij}(U). \tag{T3.24}$$

$U \mapsto R(U)$ is a homomorphism SU(2) → SO(3), onto, and exactly two-to-one: $R(U) = R(-U)$.

It is a homomorphism because it is Ad (on the basis $T_a = -\frac{i}{2}\sigma_a$, which is proportional to σ_a). The rest we verify by explicit computation.

Worked example T3.17: The rotation matrix of a general U

Write $U = a\mathbb{1} + i\vec{x}\cdot\vec{\sigma}$ with $a^2 + |\vec{x}|^2 = 1$, as in (T3.2) (here $a = x^4$). We need two Pauli identities: $[\vec{x}\cdot\vec{\sigma}, \sigma_j] = 2i\varepsilon_{kjl}x_k\sigma_l$ and $(\vec{x}\cdot\vec{\sigma})\sigma_j = 2x_j\mathbb{1} - \sigma_j(\vec{x}\cdot\vec{\sigma})$ (from $\sigma_k\sigma_j + \sigma_j\sigma_k = 2\delta_{jk}$). Then

$$\begin{aligned} U\sigma_j U^\dagger &\overset{(1)}{=} (a + i\vec{x}\cdot\vec{\sigma})\,\sigma_j\,(a - i\vec{x}\cdot\vec{\sigma}) \\ &\overset{(2)}{=} a^2\sigma_j + ia\,[\vec{x}\cdot\vec{\sigma}, \sigma_j] + (\vec{x}\cdot\vec{\sigma})\sigma_j(\vec{x}\cdot\vec{\sigma}) \\ &\overset{(3)}{=} a^2\sigma_j - 2a\,\varepsilon_{kjl}x_k\sigma_l + 2x_j(\vec{x}\cdot\vec{\sigma}) - |\vec{x}|^2\sigma_j \\ &\overset{(4)}{=} \sigma_i\Big[(a^2 - |\vec{x}|^2)\delta_{ij} + 2x_ix_j + 2a\,\varepsilon_{ijk}x_k\Big]. \end{aligned}$$

(1) $U^\dagger = a - i\vec{x}\cdot\vec{\sigma}$ since $a, \vec{x}$ are real.
(2) Multiply out; the two cross terms combine into a commutator.
(3) The commutator identity; and for the last term, move σ_j to the right using the anticommutator, then $(\vec{x}\cdot\vec{\sigma})^2 = |\vec{x}|^2$.
(4) Rename the free index $l \to i$ in the second term and use $\varepsilon_{kji} = -\varepsilon_{ijk}$ (an odd permutation), then read off the coefficient of σ_i.

So

$$R_{ij}(U) = (a^2 - |\vec{x}|^2)\,\delta_{ij} + 2x_ix_j + 2a\,\varepsilon_{ijk}x_k. \tag{T3.25}$$

Every term is *quadratic* in $(a, \vec{x})$, so $R(-U) = R(U)$: two antipodal points of S^3 give the

same rotation. Now insert the exponential (T3.16), $a = \cos\frac{\theta}{2}$, $\vec{x} = -\sin\frac{\theta}{2}\,\hat{n}$:

$$R_{ij} \overset{(1)}{=} \left(\cos^2\tfrac{\theta}{2} - \sin^2\tfrac{\theta}{2}\right)\delta_{ij} + 2\sin^2\tfrac{\theta}{2}\, n_i n_j - 2\sin\tfrac{\theta}{2}\cos\tfrac{\theta}{2}\,\varepsilon_{ijk} n_k$$
$$\overset{(2)}{=} \cos\theta\,\delta_{ij} + (1-\cos\theta)\, n_i n_j - \sin\theta\,\varepsilon_{ijk} n_k .$$

(1) Substitute.
(2) Double-angle formulas: $\cos^2 - \sin^2 = \cos\theta$, $2\sin^2\frac{\theta}{2} = 1 - \cos\theta$, $2\sin\frac{\theta}{2}\cos\frac{\theta}{2} = \sin\theta$.

Acting on v, the last term is $-\sin\theta\,\varepsilon_{ijk} n_k v_j = \sin\theta\,(\hat{n}\times v)_i$. This is precisely the Rodrigues formula (T3.18). So

$$R\left(e^{\theta\hat{n}\cdot\vec{T}}\right) = e^{\theta\hat{n}\cdot\vec{L}} :$$

the SU(2) element with parameter θ maps to the rotation by angle θ about $\hat{n}$. Every rotation is hit (onto), and θ and $\theta + 2\pi$, i.e. U and $-U$, hit the same one.

Why exactly two-to-one. $R(U) = \mathbb{1}$ means $U\sigma_j U^\dagger = \sigma_j$ for all j, i.e. U commutes with all three Pauli matrices and hence with every 2×2 matrix. Such a U is a multiple of $\mathbb{1}$, and $\det U = 1$ forces $U = \pm\mathbb{1}$. The kernel is the centre $\{\pm\mathbb{1}\} \cong \mathbb{Z}_2$. And $\det R(U) = +1$ (not -1) because $\det R$ is continuous on the connected space S^3, only takes values ± 1, and equals 1 at $U = \mathbb{1}$. Therefore

$$\mathrm{SO}(3) \cong \mathrm{SU}(2)/\{\pm\mathbb{1}\} = S^3/(\text{antipodal map}) = \mathbb{RP}^3. \tag{T3.26}$$

This matches the ball picture of Section T3.1.5: the upper hemisphere $a \geq 0$ of S^3 (i.e. $\theta \in [0,\pi]$) is a solid ball, and on its boundary $a = 0$ the antipodal identification $\vec{x} \sim -\vec{x}$ is exactly the gluing of $\pi\hat{n}$ to $-\pi\hat{n}$.

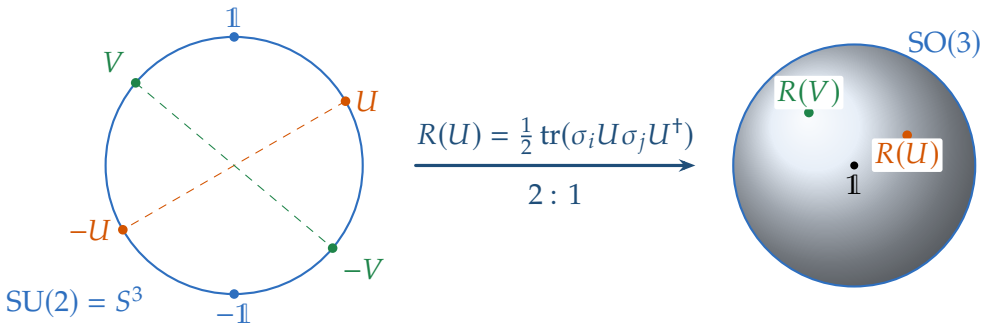


Figure T3.10: The covering map sends each antipodal pair $\{U, -U\}$ of the 3-sphere to a single rotation. Both $\mathbb{1}$ and $-\mathbb{1}$ go to the identity rotation. Locally the map is a diffeomorphism (the Lie algebras agree, Example T3.14); globally it folds S^3 onto $\mathbb{RP}^3$ twice.

T3.4.5 $\pi_1(\mathrm{SO}(3)) = \mathbb{Z}_2$ and the belt trick

The double cover explains the loop we met in Figure T3.4. Consider the loop in SO(3) of rotations about $\hat{z}$ by angle θ running from 0 to 2π: $\gamma(\theta) = R_z(\theta)$. Where does it come from in SU(2)? By Example T3.17, $\gamma(\theta) = R(\tilde{\gamma}(\theta))$ with $\tilde{\gamma}(\theta) = e^{-\frac{i\theta}{2}\sigma_3}$. But $\tilde{\gamma}(0) = \mathbb{1}$ and $\tilde{\gamma}(2\pi) = -\mathbb{1}$: the **lift** of the closed loop γ is an *open* path in S^3, from the north pole to the south pole.

Key idea: Why $\pi_1(\mathrm{SO}(3)) = \mathbb{Z}_2$

Every loop in SO(3) based at $\mathbb{1}$ lifts uniquely (given the starting point $\mathbb{1}$) to a path in SU(2) that ends at $+\mathbb{1}$ or at $-\mathbb{1}$. If we deform the loop continuously, the endpoint of the lift moves continuously in the two-point set $\{\pm\mathbb{1}\}$, so it cannot change. Hence:

- loops whose lift ends at $+\mathbb{1}$ are closed loops in S^3, and every loop in S^3 can be shrunk; so these loops are contractible;
- loops whose lift ends at $-\mathbb{1}$ (such as one full 2π rotation) can never be shrunk;
- going round γ twice lifts to $\theta \in [0, 4\pi]$, which ends at $e^{-2\pi i\sigma_3} = +\mathbb{1}$: a closed loop in S^3, hence contractible.

So there are exactly two classes of loops, and "twice = nothing": $\pi_1(\mathrm{SO}(3)) = \mathbb{Z}_2$.

The same argument works whenever a simply connected group covers another: $\pi_1(G/\Gamma) = \Gamma$ for a discrete central subgroup Γ. Thus $\pi_1(\mathrm{SO}(n)) = \mathbb{Z}_2$ for $n \geq 3$, using the simply connected double cover $\mathrm{Spin}(n) \to \mathrm{SO}(n)$ built from Clifford algebras in Chapter T4, and SU(2) = Spin(3).

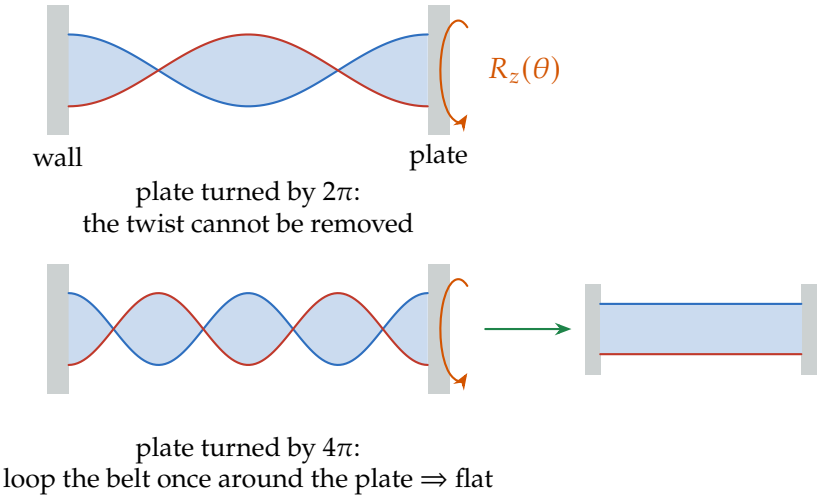


Figure T3.11: The belt trick (Dirac's scissors). A belt joins a fixed wall to a plate. Turning the plate by θ about the belt's axis traces the loop $\theta \mapsto R_z(\theta)$ in SO(3); the orientation of each cross-section of the belt is a path from $\mathbb{1}$ (at the wall) to the plate's rotation. After a 2π turn (top) the belt has a full twist, and no motion of the belt that keeps the plate's orientation fixed can remove it: this is the non-trivial element of $\pi_1(\mathrm{SO}(3))$. After a 4π turn (bottom), passing the belt once around the plate undoes the twist entirely: twice the generator is trivial.

Physical picture: Spin $\frac{1}{2}$

Quantum states are rays: ψ and $e^{i\alpha}\psi$ are the same physical state. So a symmetry need only be represented up to a phase. For rotations, this lets the group SU(2) act instead of SO(3). On a spin-$\frac{1}{2}$ state the 2π rotation is $-\mathbb{1}$: invisible on its own (a phase), but visible in interference, as in the neutron experiment. Topologically, the sign records which class in $\pi_1(\mathrm{SO}(3))$ the rotation path belongs to. Particles whose wavefunction picks up -1 are fermions; that is the origin of the need for spin structures on curved spacetimes (Chapter T4).

Try it: $\mathrm{SO}(4) = (\mathrm{SU}(2) \times \mathrm{SU}(2))/\mathbb{Z}_2$

Identify $\mathbb{R}^4$ with 2×2 matrices $q = x^4\mathbb{1} + i\vec{x}\cdot\vec{\sigma}$ (real multiples of SU(2) matrices, $\det q = |x|^2$). For $(U, V) \in \mathrm{SU}(2) \times \mathrm{SU}(2)$, show that $q \mapsto UqV^\dagger$ preserves $|x|^2$, and that (U, V) and

$(-U,-V)$ give the same map.
Answer: $\det(UqV^\dagger) = \det U \det q \det V^\dagger = \det q$, and the product $UqV^\dagger$ is still of the form $y^4 + i\vec{y}\cdot\vec{\sigma}$ (the quaternions are closed under multiplication). The signs cancel in $(-U)q(-V)^\dagger$. The Lie algebra version is the decomposition $\mathfrak{so}(4) = \mathfrak{su}(2) \oplus \mathfrak{su}(2)$ of Section T3.2.

Reading Nash (the adjoint action in Nash)

Nash uses Ad constantly without naming it. The gauge group $\mathcal{G}$ is the group of sections of the bundle $\mathrm{Ad}\, P = P \times_G G$ built from conjugation, and its Lie algebra is the space of sections of $\mathrm{ad}\, P = P \times_G \mathfrak{g}$ [Nash pp. 176–177, (6.19)–(6.24)]. Both bundles are glued together using Ad from this section, applied point by point in spacetime. His invariant polynomials such as $\det\left(t\mathbb{1} + \frac{iF}{2\pi}\right)$ [Nash p. 87, (3.125)] are the functions on $\mathfrak{g}$ that are unchanged by Ad, exactly like $\mathrm{tr}(XY)$ in (T3.23). The inner product $\langle F, F\rangle = -\int \mathrm{tr}(F \wedge \star F)$ in his Yang–Mills action [Nash p. 227, (8.63)] is the negative trace, which by the key idea above is positive for compact G.

Recap

- $\mathrm{Ad}_g X = gXg^{-1}$ (group on algebra), $\mathrm{ad}_Y X = [Y, X]$ (algebra on algebra), $(\mathrm{ad}_{T_a})_{cb} = f_{ab}{}^c$, $e^Y X e^{-Y} = e^{\mathrm{ad}_Y} X$.
- Killing form $B(X,Y) = \mathrm{tr}(\mathrm{ad}_X \mathrm{ad}_Y) = 2N\,\mathrm{tr}(XY)$ on $\mathfrak{su}(N)$; invariant; negative definite for compact groups.
- $R_{ij}(U) = \frac{1}{2}\mathrm{tr}(\sigma_i U \sigma_j U^\dagger) = (a^2 - |\vec{x}|^2)\delta_{ij} + 2x_i x_j + 2a\varepsilon_{ijk}x_k$ is a $2:1$ homomorphism $\mathrm{SU}(2) \to \mathrm{SO}(3)$ with kernel $\pm\mathbb{1}$; $e^{\theta\hat{n}\cdot\vec{T}} \mapsto$ rotation by θ.
- The 2π rotation lifts to a path from $\mathbb{1}$ to $-\mathbb{1}$; hence $\pi_1(\mathrm{SO}(3)) = \mathbb{Z}_2$ and the belt trick.

T3.5 Representations: spin, weights and roots

In quantum mechanics a symmetry group does not act on points of space; it acts on the Hilbert space of states, by unitary matrices. The same group can act on different spaces in different ways: rotations act on a scalar trivially, on a vector by 3×3 matrices, on an electron's spin by 2×2 matrices. Each such action is a *representation*, and the particles of a theory are sorted into *irreducible* representations, the multiplets: the isospin doublet (p, n), the pion triplet, the quark triplet (u, d, s), the gluon octet. This section answers the question *what are all the ways a given Lie group can act on states, and how do we label them?* The answer for SU(2) is the spin j; for bigger groups it is a *highest weight*, drawn as a point in a lattice decorated by *roots*.

Convention for this section. Representation theory is traditionally done with Hermitian generators, because their eigenvalues (the weights) are real and are the quantum numbers physicists measure. So in this section we use $J_a = iT_a$, with $[J_a, J_b] = if_{abc}J_c$, as in Section T3.2.5. For SU(2), $J_a = \frac{1}{2}\sigma_a$ in the defining representation.

T3.5.1 What a representation is

Working definition T3.9 (representation)

A **representation** of a group G on a vector space V is an assignment $g \mapsto D(g)$ of an invertible matrix to each group element with $D(gh) = D(g)D(h)$ and $D(\mathbb{1}) = \mathbb{1}$. A representation of a Lie algebra is a linear assignment $X \mapsto \rho(X)$ [each generator X gets a matrix $\rho(X)$ acting on the same V, and $\rho(aX + bY) = a\rho(X) + b\rho(Y)$] with $\rho([X, Y]) = [\rho(X), \rho(Y)]$. The two are related by $D(e^X) = e^{\rho(X)}$. *Operational test:* given matrices R_a for each generator, check $[R_a, R_b] = i f_{abc} R_c$ with the *same* structure constants as the group. Then $D(e^{-i\alpha^a J_a}) = e^{-i\alpha^a R_a}$ is (at least near the identity [far from $\mathbb{1}$ one g can equal e^X for several different X, and the formula may then give different matrices]) a representation.

The relation $\rho([X, Y]) = [\rho X, \rho Y]$ follows from $D(gh) = D(g)D(h)$ by the group-commutator computation of (T3.5): the ε^2 term of $D(ghg^{-1}h^{-1})$ is $\varepsilon^2\rho([X, Y])$ on one side and $\varepsilon^2[\rho X, \rho Y]$ on the other.

The representations we already know.

- *Trivial:* $D(g) = 1$ on $V = \mathbb{C}$, $R_a = 0$. (A singlet.)
- *Defining* or *fundamental*: $D(g) = g$ itself, on $\mathbb{C}^N$ for SU(N), on $\mathbb{R}^n$ for SO(n).
- *Adjoint:* $D(g) = \mathrm{Ad}_g$ on $V = \mathfrak{g}$, with $\rho = \mathrm{ad}$ (Section T3.4); dimension $\dim G$.
- *Conjugate:* $D(g) = \overline{g}$ (complex conjugate matrix). If $D(e^{-i\alpha R}) = e^{-i\alpha R}$, then $\overline{D} = e^{+i\alpha\bar{R}} = e^{-i\alpha(-\bar{R})}$, so the generators are $\bar{R}_a \to -R_a^* = -R_a^T$, exactly [QFT notes (31.2.7)]. Antiparticles transform in the conjugate representation.

New representations from old. Given representations D_1 on V_1 and D_2 on V_2, the **direct sum** acts block-diagonally on $V_1 \oplus V_2$, and the **tensor product** acts on $V_1 \otimes V_2$ by $D_1(g) \otimes D_2(g)$. For the generators of the tensor product, put $g = e^{\varepsilon X}$:

$$\begin{aligned} D_1(g) \otimes D_2(g) &\overset{(1)}{=} \big(\mathbb{1} + \varepsilon\rho_1(X)\big) \otimes \big(\mathbb{1} + \varepsilon\rho_2(X)\big) + O(\varepsilon^2) \\ &\overset{(2)}{=} \mathbb{1} \otimes \mathbb{1} + \varepsilon\big(\rho_1(X) \otimes \mathbb{1} + \mathbb{1} \otimes \rho_2(X)\big) + O(\varepsilon^2). \end{aligned} \tag{T3.27}$$

(1) Expand each factor to first order.
(2) The tensor product is bilinear; the ε^2 cross term is dropped.

So generators of a tensor product *add*: $R_a^{(1\otimes 2)} = R_a^{(1)} \otimes \mathbb{1} + \mathbb{1} \otimes R_a^{(2)}$ [QFT notes (31.3.13)]. This is the rule "total angular momentum $\vec{J} = \vec{J}^{(1)} + \vec{J}^{(2)}$" from quantum mechanics.

Working definition T3.10 (irreducible representation)

A representation is **irreducible** if no proper subspace $0 \neq W \subsetneq V$ is mapped into itself by all $D(g)$ [every $D(g)$ sends each vector of W to a vector of W]. *Operational test for unitary representations* [every $D(g)$ satisfies $D(g)^\dagger D(g) = \mathbb{1}$]: it is reducible exactly when some basis change puts all the R_a simultaneously into the same block-diagonal form. For a compact group [a closed and bounded set of matrices, like SU(N) or SO(N)] every representation can be made unitary (average an inner product over the group), and every representation is then a direct sum of irreducible ones.

Key idea: Schur's lemma, physicist's form

If a matrix C commutes with all the generators R_a of an irreducible representation, then $C = c\,\mathbb{1}$.

Why: an eigenspace of C is mapped into itself by every R_a (if $Cv = cv$ then $C(R_a v) = R_a C v = c\,R_a v$). By irreducibility the eigenspace is all of V. In physics: an observable that commutes with all the symmetry generators takes a single value on a whole multiplet. That is why all members of a multiplet have the same mass, and why the Casimir operator below is a number.

T3.5.2 The irreducible representations of SU(2): spin j

We construct every irreducible representation of $[J_a, J_b] = i\varepsilon_{abc}J_c$ by the ladder method. Suppose the J_a are Hermitian operators on a finite-dimensional space V. Define

$$J_\pm = J_1 \pm iJ_2, \qquad J_\pm^\dagger = J_\mp.$$

Their brackets with J_3 and with each other are

$$[J_3, J_\pm] \overset{(1)}{=} [J_3, J_1] \pm i[J_3, J_2] \overset{(2)}{=} iJ_2 \pm i(-iJ_1) = \pm(J_1 \pm iJ_2) = \pm J_\pm,$$
$$[J_+, J_-] \overset{(3)}{=} -i[J_1, J_2] + i[J_2, J_1] \overset{(4)}{=} -i(iJ_3) + i(-iJ_3) = 2J_3.$$

(1) Bilinearity.
(2) $[J_3, J_1] = i\varepsilon_{312}J_2 = iJ_2$ and $[J_3, J_2] = i\varepsilon_{321}J_1 = -iJ_1$.
(3) Expand $[J_1 + iJ_2, J_1 - iJ_2]$; the $[J_1, J_1]$ and $[J_2, J_2]$ terms vanish.
(4) $[J_1, J_2] = iJ_3$.

The first relation says $J_\pm$ are **ladder operators**: if $J_3 v = mv$ then

$$J_3(J_\pm v) \overset{(1)}{=} J_\pm J_3 v + [J_3, J_\pm]v \overset{(2)}{=} (m \pm 1)\,J_\pm v.$$

(1) Write $J_3 J_\pm = J_\pm J_3 + [J_3, J_\pm]$.
(2) $J_3 v = mv$ and $[J_3, J_\pm] = \pm J_\pm$.

So J_+ raises the J_3 eigenvalue by one and J_- lowers it by one. The **Casimir operator** $\vec{J}^2 = J_1^2 + J_2^2 + J_3^2$ commutes with all J_a [QFT notes (31.3.2)], so by Schur it is a number on an irreducible representation. Two rewritings are useful:

$$\vec{J}^2 = J_-J_+ + J_3^2 + J_3 = J_+J_- + J_3^2 - J_3, \tag{T3.28}$$

since $J_\mp J_\pm = J_1^2 + J_2^2 \pm i[J_1, J_2] = J_1^2 + J_2^2 \mp J_3$.

Building the multiplet. V is finite-dimensional, so J_3 has a largest eigenvalue j, with eigenvector $|j, j\rangle$. Then $J_+|j, j\rangle = 0$ (it would have eigenvalue $j + 1$). From (T3.28), $\vec{J}^2|j, j\rangle = (0 + j^2 + j)|j, j\rangle$, so $\vec{J}^2 = j(j + 1)$ on the whole irreducible representation. Now lower repeatedly. The norm of each lowered state is

$$\big\|J_-|j, m\rangle\big\|^2 \overset{(1)}{=} \langle j, m|J_+J_-|j, m\rangle \overset{(2)}{=} \langle j, m|\big(\vec{J}^2 - J_3^2 + J_3\big)|j, m\rangle$$
$$\overset{(3)}{=} j(j + 1) - m^2 + m = (j + m)(j - m + 1). \tag{T3.29}$$

(1) $J_-^\dagger = J_+$.

(2) The second form of (T3.28).
(3) Eigenvalues, with $|j,m\rangle$ normalised.

Starting from $m = j$ and stepping down by one, the right-hand side $(j+m)(j-m+1)$ stays positive until $m = -j$, where it vanishes. If the sequence $j, j-1, j-2, \ldots$ skipped over $-j$, the next norm would be negative, which is impossible. So the ladder must land exactly on $m = -j$: $j-(-j) = 2j$ is a non-negative integer.

Key idea: Irreducible representations of $\mathfrak{su}(2)$

For each $j \in \{0, \frac{1}{2}, 1, \frac{3}{2}, \ldots\}$ there is exactly one irreducible representation, of dimension $2j+1$, with orthonormal basis $|j,m\rangle$, $m = -j, -j+1, \ldots, j$, and

$$J_3|j,m\rangle = m|j,m\rangle, \quad J_\pm|j,m\rangle = \sqrt{(j\mp m)(j\pm m+1)}\,|j,m\pm1\rangle, \quad \vec{J}^{\,2} = j(j+1). \tag{T3.30}$$

The eigenvalues m of J_3 are the **weights**; j is the **highest weight**.

Worked example T3.18: Explicit matrices for $j = \frac{1}{2}$ and $j = 1$

For $j = \frac{1}{2}$, basis $(|\frac{1}{2},\frac{1}{2}\rangle, |\frac{1}{2},-\frac{1}{2}\rangle)$: (T3.30) gives $J_+|\frac{1}{2},-\frac{1}{2}\rangle = \sqrt{1\cdot1}|\frac{1}{2},\frac{1}{2}\rangle$, so

$$J_3 = \tfrac{1}{2}\begin{pmatrix}1&0\\0&-1\end{pmatrix}, \quad J_+ = \begin{pmatrix}0&1\\0&0\end{pmatrix}, \quad J_1 = \tfrac{1}{2}(J_+ + J_-) = \tfrac{1}{2}\sigma_1, \quad J_2 = \tfrac{1}{2i}(J_+ - J_-) = \tfrac{1}{2}\sigma_2.$$

This is the defining representation. For $j = 1$, basis $m = 1, 0, -1$: $J_+|1,0\rangle = \sqrt{1\cdot2}|1,1\rangle$ and $J_+|1,-1\rangle = \sqrt{2\cdot1}|1,0\rangle$, so

$$J_3 = \begin{pmatrix}1&0&0\\0&0&0\\0&0&-1\end{pmatrix}, \quad J_+ = \sqrt{2}\begin{pmatrix}0&1&0\\0&0&1\\0&0&0\end{pmatrix}, \quad J_1 = \tfrac{1}{\sqrt{2}}\begin{pmatrix}0&1&0\\1&0&1\\0&1&0\end{pmatrix}.$$

Check the Casimir on $|1,0\rangle$: $J_3^2|1,0\rangle = 0$ and $J_+J_-|1,0\rangle = J_+(\sqrt{2}|1,-1\rangle) = 2|1,0\rangle$, so $\vec{J}^{\,2} = J_+J_- + J_3^2 - J_3 = 2 = 1\cdot(1+1)$. This representation is the adjoint one, iL_a of Example T3.6, written in a different basis: iL_3 has eigenvalues $+1, 0, -1$ with eigenvectors $(\hat{x}+i\hat{y})/\sqrt{2}$, $\hat{z}$, $(\hat{x}-i\hat{y})/\sqrt{2}$, the "spherical basis" of vectors.

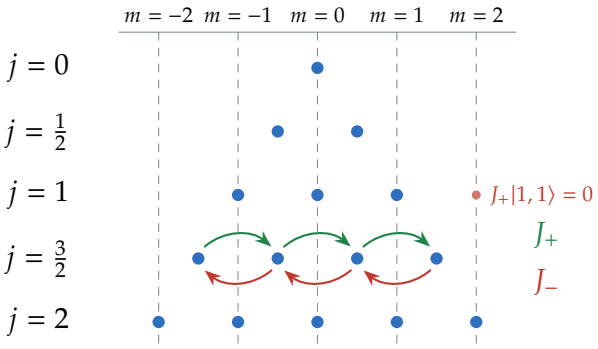


Figure T3.12: Weight diagrams of the spin-j representations of SU(2). Each dot is a state $|j,m\rangle$ at position m (its J_3 eigenvalue). The raising operator J_+ moves one step right, J_- one step left, and the ladder must stop symmetrically at $\pm j$. The difference between neighbouring weights, ±1, is the *root* of $\mathfrak{su}(2)$. Integer j has a weight at $m = 0$; half-integer j does not.

Which representations are representations of SO(3)? A rotation by 2π about $\hat{z}$ is $e^{-2\pi i J_3}$. On $|j,m\rangle$ it is $e^{-2\pi i m} = (-1)^{2m} = (-1)^{2j}$, since m and j differ by an integer. It is the identity for integer j and $-\mathbb{1}$ for half-integer j. So integer spins are genuine representations of SO(3) (orbital angular momentum, $Y_{\ell m}$), while half-integer spins are representations of SU(2) only, in which the element $-\mathbb{1}$ of Section T3.4 acts as -1.

Worked example T3.19: Two spin-$\frac{1}{2}$ particles: $\frac{1}{2}\otimes\frac{1}{2} = 0\oplus 1$

The tensor product has the four states $|\uparrow\uparrow\rangle, |\uparrow\downarrow\rangle, |\downarrow\uparrow\rangle, |\downarrow\downarrow\rangle$ with total weights $m = m_1 + m_2 = 1, 0, 0, -1$ (generators add, so weights add). The highest weight is 1, carried by $|\uparrow\uparrow\rangle$; lower it with the total $J_- = J_-^{(1)} + J_-^{(2)}$:

$$J_-|\uparrow\uparrow\rangle \overset{(1)}{=} (J_-|\uparrow\rangle)\otimes|\uparrow\rangle + |\uparrow\rangle\otimes(J_-|\uparrow\rangle) \overset{(2)}{=} |\downarrow\uparrow\rangle + |\uparrow\downarrow\rangle \overset{(3)}{=} \sqrt{2}\,|1,0\rangle.$$

(1) The tensor-product generator (T3.27).
(2) For $j=\frac{1}{2}$, $J_-|\frac{1}{2},\frac{1}{2}\rangle = \sqrt{1\cdot 1}|\frac{1}{2},-\frac{1}{2}\rangle$.
(3) (T3.30) for $j=1$, $m=1$: $J_-|1,1\rangle = \sqrt{2}|1,0\rangle$.

So the triplet is $|\uparrow\uparrow\rangle$, $\frac{1}{\sqrt{2}}(|\uparrow\downarrow\rangle + |\downarrow\uparrow\rangle)$, $|\downarrow\downarrow\rangle$. The orthogonal combination $\frac{1}{\sqrt{2}}(|\uparrow\downarrow\rangle - |\downarrow\uparrow\rangle)$ has $m=0$ and is annihilated by J_+ (check: J_+ gives $\frac{1}{\sqrt{2}}(|\uparrow\uparrow\rangle - |\uparrow\uparrow\rangle) = 0$), so it is a highest weight with $j=0$: the singlet. Dimensions: $2\times 2 = 1+3$. In general $j_1\otimes j_2 = |j_1-j_2|\oplus\cdots\oplus(j_1+j_2)$; the dimension check is $\sum_{j=|j_1-j_2|}^{j_1+j_2}(2j+1) = (2j_1+1)(2j_2+1)$.

Real, pseudoreal and complex. Is the conjugate of the spin-$\frac{1}{2}$ representation new? The Pauli matrices obey $\sigma_2\sigma_a^*\sigma_2 = -\sigma_a$ (check: $\sigma_1^* = \sigma_1$, $\sigma_3^* = \sigma_3$ anticommute with σ_2; $\sigma_2^* = -\sigma_2$ commutes with it). So $-\sigma_a^* = \sigma_2\sigma_a\sigma_2^{-1}$: the conjugate generators are equivalent to the original ones, $\bar{\mathbf{2}}\cong\mathbf{2}$. The intertwiner $\varepsilon = i\sigma_2$ is antisymmetric, and the invariant $\varepsilon_{\alpha\beta}\xi^\alpha\eta^\beta$ is the spin singlet above. A representation equivalent to its conjugate with an antisymmetric invariant is **pseudoreal**; with a symmetric one (like $\vec{v}\cdot\vec{w}$ for $j=1$) it is **real** [QFT notes §31.2]. For SU(3) the $\mathbf{3}$ and $\bar{\mathbf{3}}$ are genuinely different (*complex*), which is why quarks and antiquarks are distinguishable.

T3.5.3 Cartan subalgebra, weights and roots

For SU(2) we singled out one generator, J_3, diagonalised it, and used the others as ladders. For a bigger group we diagonalise as many generators as possible at once.

Working definition T3.11 (Cartan subalgebra, weights, roots)

A **Cartan subalgebra** is a maximal [no generator outside their span commutes with all of them] set of mutually commuting Hermitian generators $H_1,\dots,H_\ell$. The number ℓ is the **rank**. Commuting Hermitian matrices can be diagonalised together; a common eigenvector $|\mu\rangle$ has an eigenvalue vector

$$H_i|\mu\rangle = \mu_i|\mu\rangle, \qquad \mu = (\mu_1,\dots,\mu_\ell)\in\mathbb{R}^\ell,$$

called its **weight**. The weights of the adjoint representation [the generators themselves form the vector space, and H_i acts by $[H_i,\cdot\,]$] (other than 0) are the **roots**. The corresponding generators E_α satisfy $[H_i,E_\alpha] = \alpha_i E_\alpha$ [QFT notes (31.4.1)]. *How to find them:* for $\mathfrak{su}(N)$ take H_i to be the traceless diagonal matrices; the weights of the fundamental representation are the diagonal entries; the roots are the differences of these weights, carried by the matrix units E_{jk} [the matrix with 1 in row j, column k and zeros elsewhere].

Roots act as ladders on weights, in every representation:

$$H_i\big(E_\alpha|\mu\rangle\big) \overset{(1)}{=} E_\alpha H_i|\mu\rangle + [H_i, E_\alpha]|\mu\rangle \overset{(2)}{=} (\mu_i + \alpha_i)\, E_\alpha|\mu\rangle.$$

(1) Move H_i through E_α.
(2) Eigenvalue equation and the defining property of a root.

So E_α shifts the weight by α. For SU(2): rank 1, $H = J_3$, roots ± 1 with $E_{\pm 1} = J_\pm$. The weight diagrams of Figure T3.12 are one-dimensional because the rank is one.

Weights as characters of the torus. Exponentiating the Cartan generators gives the **maximal torus** $T = \{e^{-i\theta^i H_i}\} \cong \mathrm{U}(1)^\ell$. On a weight vector, $e^{-i\theta\cdot H}$ acts by the phase $e^{-i\theta\cdot\mu}$. So a weight is the same thing as a homomorphism $T \to \mathrm{U}(1)$, $e^{-i\theta\cdot H} \mapsto e^{-i\theta\cdot\mu}$. This is how Nash defines weights [Nash p. 188]. For SU(2), the torus is $\{e^{-i\theta J_3}\}$ with $\theta \in [0, 4\pi)$ and the weight m is the homomorphism $\theta \mapsto e^{-im\theta}$; single-valuedness on the torus forces $2m \in \mathbb{Z}$, the same integrality we found from the ladder.

T3.5.4 SU(3): triangles and a hexagon

In the Gell-Mann basis of Section T3.2.6, the diagonal generators $H_1 = J_3 = \frac{1}{2}\lambda_3$ and $H_2 = J_8 = \frac{1}{2}\lambda_8$ commute, and no other generator commutes with both. So $\mathfrak{su}(3)$ has rank 2 and its weights are points in a plane:

$$H_1 = \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ 0 & -\frac{1}{2} & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad H_2 = \frac{1}{2\sqrt{3}} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{pmatrix}.$$

The fundamental 3. The standard basis vectors e_1, e_2, e_3 of $\mathbb{C}^3$ are common eigenvectors. Reading off the diagonal entries,

$$\mu^{(1)} = \Big(\tfrac{1}{2}, \tfrac{1}{2\sqrt{3}}\Big), \qquad \mu^{(2)} = \Big(-\tfrac{1}{2}, \tfrac{1}{2\sqrt{3}}\Big), \qquad \mu^{(3)} = \Big(0, -\tfrac{1}{\sqrt{3}}\Big).$$

Each has length squared $\frac{1}{4} + \frac{1}{12} = \frac{1}{3}$: the three weights form an equilateral triangle centred at the origin. In the quark model these are u, d, s, with isospin $I_3 = \mu_1$ and hypercharge $Y = \frac{2}{\sqrt{3}}\mu_2$: u, d have $Y = \frac{1}{3}$, s has $Y = -\frac{2}{3}$. The electric charge $Q = I_3 + \frac{Y}{2}$ gives $\frac{2}{3}, -\frac{1}{3}, -\frac{1}{3}$. The conjugate representation $\bar{\mathbf{3}}$ has generators $-R_a^*$, so for the diagonal ones $-H_i$: its weights are $-\mu^{(k)}$, the inverted triangle (the antiquarks).

The roots. For any diagonal matrix $D = \mathrm{diag}(d_1, d_2, d_3)$ and matrix unit E_{jk},

$$[D, E_{jk}] \overset{(1)}{=} \sum_l d_l \big(E_{ll}E_{jk} - E_{jk}E_{ll}\big) \overset{(2)}{=} d_j E_{jk} - d_k E_{jk} = (d_j - d_k)\, E_{jk}.$$

(1) Write $D = \sum_l d_l E_{ll}$.
(2) $E_{ll}E_{jk} = \delta_{lj}E_{lk}$ and $E_{jk}E_{ll} = \delta_{kl}E_{jl}$.

Applying this to H_1, H_2: E_{jk} (for $j \neq k$) is a root vector with root $\alpha = \mu^{(j)} - \mu^{(k)}$. The six roots are

$$\pm(1, 0), \qquad \pm\Big(\tfrac{1}{2}, \tfrac{\sqrt{3}}{2}\Big), \qquad \pm\Big(-\tfrac{1}{2}, \tfrac{\sqrt{3}}{2}\Big),$$

from E_{12}, E_{13}, E_{23} and their transposes. All have length 1 and neighbouring roots are 60° apart: a regular hexagon. The adjoint representation has these six weights plus the weight 0 twice (H_1 and H_2 themselves), $6 + 2 = 8 = \dim \mathrm{SU}(3)$: the gluon (or meson) octet. In terms of Gell-Mann matrices, $E_{12} = \frac{1}{2}(\lambda_1 + i\lambda_2)$ is the isospin raising operator (root $(1, 0)$), $E_{13} = \frac{1}{2}(\lambda_4 + i\lambda_5)$ and $E_{23} = \frac{1}{2}(\lambda_6 + i\lambda_7)$ are the "V-spin" and "U-spin" raising operators.

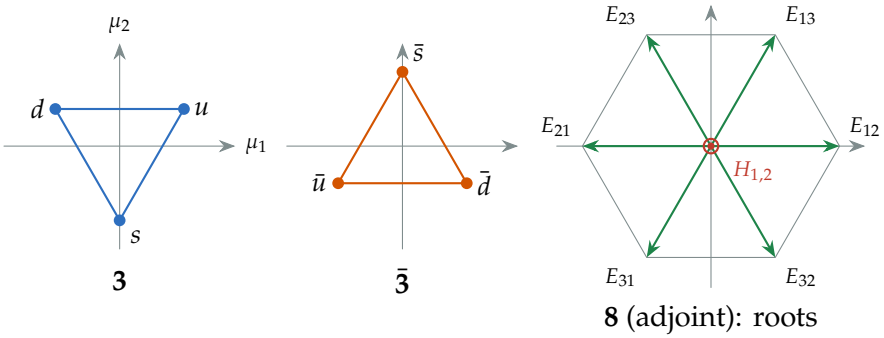


Figure T3.13: Weight diagrams of SU(3) in the (μ_1, μ_2) plane, the eigenvalues of $(H_1, H_2) = (\frac{1}{2}\lambda_3, \frac{1}{2}\lambda_8)$. Left: the quark triplet. Middle: the antiquark triplet, the negative of the first. Right: the adjoint representation: six roots (green arrows) forming a hexagon, each carried by a matrix unit E_{jk}, plus a doubly occupied weight 0 (the two Cartan generators). Each root is the difference of two weights of the **3**: e.g. E_{12} turns d into u.

Worked example T3.20: Checking one root with Gell-Mann matrices

Take $E_{13} = \frac{1}{2}(\lambda_4 + i\lambda_5)$, which is the matrix with a single 1 in position $(1, 3)$. By the formula above, $[H_1, E_{13}] = (\frac{1}{2} - 0)E_{13}$ and $[H_2, E_{13}] = (\frac{1}{2\sqrt{3}} + \frac{2}{2\sqrt{3}})E_{13} = \frac{\sqrt{3}}{2}E_{13}$. So its root is $(\frac{1}{2}, \frac{\sqrt{3}}{2})$. Acting on the s quark e_3: $E_{13}e_3 = e_1$, the u quark, and indeed $\mu^{(3)} + (\frac{1}{2}, \frac{\sqrt{3}}{2}) = (\frac{1}{2}, \frac{\sqrt{3}}{2} - \frac{1}{\sqrt{3}}) = (\frac{1}{2}, \frac{1}{2\sqrt{3}}) = \mu^{(1)}$. Its commutator with its adjoint is $[E_{13}, E_{31}] = E_{11} - E_{33} = \mathrm{diag}(1, 0, -1) = H_1 + \sqrt{3}H_2 = 2\,\alpha \cdot H$, a Cartan element in the direction of the root, as in [QFT notes (31.4.1)].

T3.5.5 Why roots are so rigid: integrality and the rank-two list

Every root α comes with its own copy of $\mathfrak{su}(2)$. Normalise the root vectors so that $[E_\alpha, E_{-\alpha}] = \alpha \cdot H$, and define

$$\mathcal{J}_3^{(\alpha)} = \frac{\alpha \cdot H}{(\alpha, \alpha)}, \qquad (\alpha, \beta) = \sum_i \alpha_i \beta_i .$$

Then $[\mathcal{J}_3^{(\alpha)}, E_{\pm\alpha}] = \pm\frac{(\alpha,\alpha)}{(\alpha,\alpha)}E_{\pm\alpha} = \pm E_{\pm\alpha}$, the $\mathfrak{su}(2)$ relation $[J_3, J_\pm] = \pm J_\pm$. (We use the Euclidean inner product in the orthonormal basis $\mathrm{tr}(H_i H_j) \propto \delta_{ij}$; the QFT notes allow a general Killing metric $g^{ij} = \mathrm{tr}\, H^i H^j$ [QFT notes (31.4.2)].) Any representation of $\mathfrak{g}$ is in particular a representation of this $\mathfrak{su}(2)$, so by Section T3.5.2 the eigenvalues of $\mathcal{J}_3^{(\alpha)}$ are half-integers. Apply this to the adjoint representation and the state E_β of weight β:

$$[\mathcal{J}_3^{(\alpha)}, E_\beta] \overset{(1)}{=} \frac{\alpha_i [H_i, E_\beta]}{(\alpha, \alpha)} \overset{(2)}{=} \frac{(\alpha, \beta)}{(\alpha, \alpha)} E_\beta \quad \Longrightarrow \quad 2\,\frac{(\alpha, \beta)}{(\alpha, \alpha)} \in \mathbb{Z}.$$

(1) Definition of $\mathcal{J}_3^{(\alpha)}$.
(2) $[H_i, E_\beta] = \beta_i E_\beta$; the eigenvalue must be in $\frac{1}{2}\mathbb{Z}$.

The same holds with α and β swapped. Multiplying the two integers $n = 2(\alpha,\beta)/(\alpha,\alpha)$ and $m = 2(\alpha,\beta)/(\beta,\beta)$,

$$\cos^2\theta_{\alpha\beta} = \frac{(\alpha,\beta)^2}{(\alpha,\alpha)(\beta,\beta)} = \frac{mn}{4} \in \{0, \tfrac14, \tfrac12, \tfrac34, 1\}$$

[QFT notes (31.4.6)]. So the angle between two roots can only be 90°, 60° (or 120°), 45° (or 135°), 30° (or 150°), with length ratios $\sqrt{m/n}$ equal to 1, 1, $\sqrt2$, $\sqrt3$ respectively. The **Weyl reflection** $s_\alpha(\beta) = \beta - 2\frac{(\alpha,\beta)}{(\alpha,\alpha)}\alpha$, the mirror image of β in the plane perpendicular to α, maps roots to roots (it is the $\mathrm{SU}(2)^{(\alpha)}$ rotation by π, which exchanges m and $-m$). For SU(3) the Weyl reflections permute the three weights of the triangle: the Weyl group is S_3.



$A_1 \times A_1$: 90° $\mathfrak{so}(4)$, dim 2 + 4 | A_2: 60° $\mathfrak{su}(3)$, dim 2 + 6 | $B_2 = C_2$: 45° $\mathfrak{so}(5) \cong \mathfrak{sp}(2)$, dim 2 + 8 | G_2: 30° dim 2 + 12

Figure T3.14: The four root systems of rank two, one for each allowed angle between simple roots. Green: short roots; orange: long roots (length ratio $\sqrt2$ for B_2, $\sqrt3$ for G_2). The dimension of the Lie algebra is the rank plus the number of roots. Nothing else is possible: the integrality of $2(\alpha,\beta)/(\alpha,\alpha)$ forbids every other angle.

Watch out: A misprint in the QFT notes

In the list of rank-two algebras the QFT notes call the 45° case "$\mathfrak{so}(8) \equiv \mathfrak{sp}(4)$" [QFT notes §31.4, item 3]. The correct statement is $\mathfrak{so}(5) \cong \mathfrak{sp}(4)$ in their convention (our $\mathfrak{sp}(2)$): dimension 10, as the count 2 + 8 shows. ($\mathfrak{so}(8)$ has dimension 28 and rank 4.)

Simple roots, the Cartan matrix and Dynkin diagrams. Call a root **positive** if its first nonzero component is positive. A positive root that is not a sum of two positive roots is **simple**; there are exactly ℓ of them, and every positive root is a combination of simple roots with non-negative integer coefficients. The **Cartan matrix** $A_{ij} = 2(\alpha_i,\alpha_j)/(\alpha_j,\alpha_j)$ of the simple roots determines the whole algebra.

Worked example T3.21: Simple roots and Cartan matrix of $\mathfrak{su}(3)$

By our rule the positive roots are $(1,0)$, $(\frac12, \frac{\sqrt3}{2})$ and $(\frac12, -\frac{\sqrt3}{2})$; the root $(-\frac12, \frac{\sqrt3}{2})$ counts as negative because its first component is negative. The first is the sum of the other two, $(1,0) = (\frac12, \frac{\sqrt3}{2}) + (\frac12, -\frac{\sqrt3}{2})$, so it is not simple, and the simple roots are $\alpha_1 = (\frac12, \frac{\sqrt3}{2})$, $\alpha_2 = (\frac12, -\frac{\sqrt3}{2})$. Then

$$(\alpha_1,\alpha_1) = (\alpha_2,\alpha_2) = 1, \quad (\alpha_1,\alpha_2) \overset{(1)}{=} \tfrac14 - \tfrac34 = -\tfrac12, \quad A \overset{(2)}{=} \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}.$$

(1) Dot product of the two vectors.
(2) $A_{12} = 2(-\frac12)/1 = -1 = A_{21}$, and $A_{ii} = 2$ always.

The angle between the simple roots is 120°. The Weyl reflection of α_2 in α_1 is $s_{\alpha_1}(\alpha_2) = \alpha_2 - (-1)\alpha_1 = \alpha_1 + \alpha_2 = (1,0)$, the third positive root.

The Dynkin diagram encodes the Cartan matrix: one node per simple root, and $A_{ij}A_{ji} \in$

$\{0,1,2,3\}$ lines between nodes i and j, with an arrow pointing to the shorter root when the lengths differ. The classification of possible diagrams gives the complete list of compact simple Lie algebras (Figure T3.15).

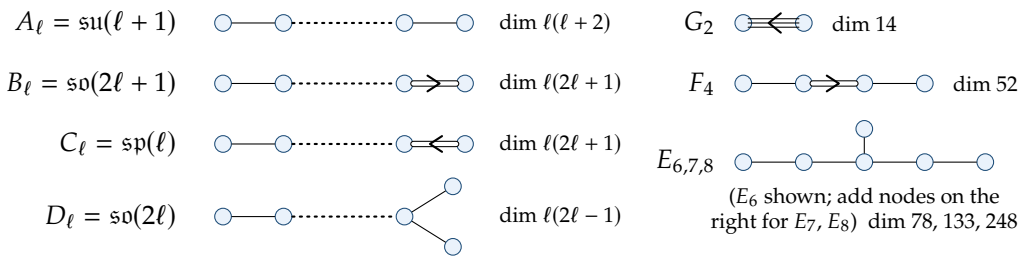


Figure T3.15: Dynkin diagrams of all compact simple Lie algebras. Nodes are simple roots; no line means 90°, one line 120°, two lines 135°, three lines 150°; arrows point to the shorter root. The four infinite families are the classical algebras of Table T3.3; five exceptional algebras complete the list.

T3.5.6 Casimirs and the index of a representation

Two numbers attached to each irreducible representation R appear in every loop calculation in gauge theory. With Hermitian generators R_a normalised so that $\mathrm{tr}_{\text{fund}}(J_a J_b) = \frac{1}{2}\delta_{ab}$:

$$\sum_a R_a R_a = C_2(R)\,\mathbb{1}_{\dim R}, \qquad \mathrm{tr}(R_a R_b) = C(R)\,\delta_{ab}. \tag{T3.31}$$

The first is Schur's lemma applied to the Casimir; the second holds because $\mathrm{tr}(R_a R_b)$ is an invariant symmetric form on a simple algebra and hence proportional to the Killing form. They are related by taking the trace of the first and contracting the second with δ_{ab}:

$$C_2(R)\dim R \overset{(1)}{=} \mathrm{tr}\sum_a R_a R_a \overset{(2)}{=} \sum_a C(R)\,\delta_{aa} = C(R)\dim G. \tag{T3.32}$$

(1) Trace of the first equation in (T3.31).
(2) Set $a = b$ in the second and sum over a; $\delta_{aa} = \dim G$.

Worked example T3.22: Casimirs of $\mathrm{SU}(N)$

Fundamental: $C(\mathbf{N}) = \frac{1}{2}$ by normalisation, so (T3.32) gives $C_2(\mathbf{N}) = \frac{N^2-1}{2N}$. For SU(2) that is $\frac{3}{4} = \frac{1}{2}(\frac{1}{2}+1)$, the spin-$\frac{1}{2}$ value of $\vec{J}^2$. For SU(3) it is $\frac{4}{3}$, the "colour factor" C_F of a quark. *Adjoint:* the adjoint generators are $J^A_a = i\,\mathrm{ad}_{T_a}$, so

$$\mathrm{tr}\big(J^A_a J^A_b\big) \overset{(1)}{=} -\mathrm{tr}\big(\mathrm{ad}_{T_a}\mathrm{ad}_{T_b}\big) \overset{(2)}{=} -B(T_a,T_b) \overset{(3)}{=} -2N\,\mathrm{tr}(T_a T_b) \overset{(4)}{=} 2N\,\mathrm{tr}(J_a J_b) = N\,\delta_{ab}.$$

(1) $i^2 = -1$.
(2) Definition of the Killing form.
(3) Example T3.16.
(4) $T_a = -iJ_a$, so $\mathrm{tr}(T_a T_b) = -\mathrm{tr}(J_a J_b) = -\frac{1}{2}\delta_{ab}$.

So $C(A) = N$, and since $\dim A = \dim G$, also $C_2(A) = N$ [QFT notes (31.3.22)]. For SU(2), $C_2(A) = 2 = 1\cdot(1+1)$, the spin-1 value. For QCD, $C_A = 3$. The one-loop beta function coefficient $b_0 = \frac{11}{3}C_A - \frac{4}{3}T_F n_f$ (with $T_F = C(\mathbf{N}) = \frac{1}{2}$) is built from exactly these numbers; $b_0 > 0$ is asymptotic freedom.

Try it: $\mathbf{N}\otimes\bar{\mathbf{N}} = \mathbf{1}\oplus$ adjoint

A quark–antiquark pair $\psi^i\bar{\chi}_j$ is an $N\times N$ matrix $M^i{}_j$ transforming as $M\mapsto UMU^\dagger$. Split $M = \frac{\operatorname{tr}M}{N}\mathbb{1} + (M - \frac{\operatorname{tr}M}{N}\mathbb{1})$ and identify the two pieces.
Answer: the trace is invariant ($\operatorname{tr}UMU^\dagger = \operatorname{tr}M$): a singlet, the colour-neutral meson. The traceless part transforms by conjugation, i.e. in the adjoint representation, of dimension N^2-1. So $N^2 = 1 + (N^2-1)$ [QFT notes (31.3.20)].

Reading Nash (weights, roots and highest weights in Nash)

Nash's account of the Borel–Weil theorem [Nash pp. 188–190, (6.63)–(6.70)] uses exactly the objects of this section. The weights of a representation are homomorphisms $\lambda_i : T\to \mathrm{U}(1)$ from the maximal torus; an irreducible representation is labelled by its highest weight λ and written V_λ. His decomposition $\mathfrak{g} = \mathfrak{t}\oplus\mathfrak{h}$ [Nash (6.66)] splits the Lie algebra into the Cartan subalgebra $\mathfrak{t}$ and the sum $\mathfrak{h}$ of root spaces; after complexifying, $\mathfrak{h}_{\mathbb{C}} = \mathfrak{h}^+_{\mathbb{C}}\oplus\mathfrak{h}^-_{\mathbb{C}}$ [Nash (6.68)] is our split into positive and negative root vectors $E_{\pm\alpha}$, the "step operators". For SU(2) these are $J_\pm$. The Casimir operator he uses for Kac–Moody algebras [Nash p. 182, (6.39)] is the $\sum_a R_aR_a$ of (T3.31).

Recap

- A representation turns group products into matrix products; its generators satisfy the same commutation relations. Tensor products add generators; conjugates use $-R^*$.
- Schur: anything commuting with an irrep is a number. Hence Casimirs are constants on multiplets.
- SU(2) irreps: spin $j\in\frac12\mathbb{Z}_{\geq0}$, weights $m = -j,\dots,j$, $J_\pm$ ladders with norms $(j\mp m)(j\pm m+1)$. Integer j only for SO(3).
- Cartan subalgebra = maximal commuting set; weights = simultaneous eigenvalues; roots = weights of the adjoint, acting as ladders.
- SU(3): rank 2, $\mathbf{3}$ is a triangle, $\bar{\mathbf{3}}$ the inverted triangle, roots form a hexagon.
- $2(\alpha,\beta)/(\alpha,\alpha)\in\mathbb{Z}$ fixes the allowed angles and leads to the Dynkin classification.
- $C_2(R)\dim R = C(R)\dim G$; SU($N$): $C(\mathbf{N}) = \frac12$, $C_2(\mathbf{N}) = \frac{N^2-1}{2N}$, $C_2(A) = N$.

T3.6 The Maurer–Cartan form $g^{-1}\mathrm{d}g$

So far a group element has been a fixed matrix. In gauge theory it becomes a *field*: a matrix $g(x)$ that depends on the point x of spacetime. The first thing a physicist wants to do with a field is differentiate it. Here we hit a small but important snag. The derivative $\mathrm{d}g$ of a group-valued function is not in the Lie algebra: for $g(x)\in\mathrm{SU}(2)$, the matrix $\partial_\mu g$ is neither anti-Hermitian nor traceless in general. It is a tangent vector to the group at the point g, not at the identity. The question of this section is: *how do we turn the derivative of a group-valued field into something that lives in the Lie algebra, so that we can expand it in generators T_a?* The answer, $g^{-1}\mathrm{d}g$, is the single most important formula connecting Lie groups to differential forms. It is the prototype of every gauge field (it is "pure gauge"), its structure equation is the prototype of $F = \mathrm{d}A + A\wedge A$, and the trace of its cube counts how many times a map wraps around the group.

T3.6.1 Pulling a velocity back to the identity

Take a curve $g(t)$ in a matrix group, passing through $g_0 = g(0)$. Its velocity $\dot{g}(0)$ is a tangent vector at g_0. Multiply it on the left by g_0^{-1}. The curve $t \mapsto g_0^{-1} g(t)$ lies in the group (it is a product of group elements) and passes through $\mathbb{1}$ at $t = 0$. So its velocity $g_0^{-1}\dot{g}(0)$ is in $\mathfrak{g}$, by the working definition of Section T3.2. Left multiplication by g_0^{-1} has carried the tangent vector from g_0 back to the identity.

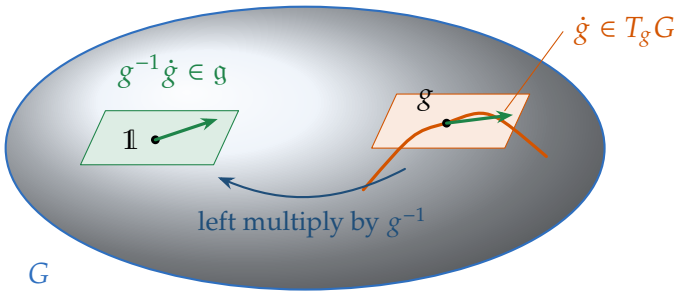


Figure T3.16: The velocity $\dot{g}$ of a curve through g is a tangent vector at g (orange plane), not in the Lie algebra. Left multiplication by g^{-1} maps a neighbourhood of g to a neighbourhood of $\mathbb{1}$ and carries $\dot{g}$ to $g^{-1}\dot{g}$, which lies in the tangent plane at the identity: the Lie algebra (green plane). Doing this at every point at once is the Maurer–Cartan form.

Now let g depend on several variables x^μ, the coordinates of a manifold M (spacetime, or the group itself). Doing the same thing for each partial derivative gives a Lie-algebra-valued 1-form.

Working definition T3.12 (Maurer–Cartan form)

For a group-valued function $g : M \to G$, its **Maurer–Cartan form** is the matrix of 1-forms [it turns a small step δx^μ in M into the generator $g^{-1}\delta g = g^{-1}\partial_\mu g\, \delta x^\mu$]

$$\theta \coloneqq g^{-1}\mathrm{d}g = g^{-1}\partial_\mu g\, \mathrm{d}x^\mu .$$

Each component $g^{-1}\partial_\mu g$ lies in $\mathfrak{g}$, so we may expand $\theta = \theta^a T_a$ with ordinary real 1-forms θ^a. *How to compute it:* differentiate the matrix entry by entry, multiply on the left by g^{-1}, and read off the coefficient of each generator T_a (for $\mathfrak{su}(N)$, $\theta^a = -2\,\mathrm{tr}(T_a\theta)$ when $\mathrm{tr}(T_aT_b) = -\frac{1}{2}\delta_{ab}$). When $M = G$ and g is the identity map, θ is called *the* Maurer–Cartan form of the group.

Let us check that the components really are in $\mathfrak{u}(n)$ when g is unitary:

$$(g^{-1}\partial_\mu g)^\dagger \overset{(1)}{=} (\partial_\mu g)^\dagger\, g \overset{(2)}{=} -g^\dagger\,\partial_\mu g\, g^\dagger g \overset{(3)}{=} -g^{-1}\partial_\mu g .$$

(1) The dagger of a product reverses the order, and $(g^{-1})^\dagger = g$ for unitary g.
(2) Differentiate $g^\dagger g = \mathbb{1}$: $(\partial_\mu g^\dagger)g = -g^\dagger\partial_\mu g$; multiply on the right by $g^\dagger$ and use $gg^\dagger = \mathbb{1}$, so $(\partial_\mu g)^\dagger = -g^\dagger(\partial_\mu g)g^\dagger$.
(3) $g^\dagger g = \mathbb{1}$ and $g^\dagger = g^{-1}$.

For $\mathrm{SU}(n)$ it is also traceless, because $\mathrm{tr}(g^{-1}\partial_\mu g) = \partial_\mu \log\det g = 0$ when $\det g = 1$ (a consequence of $\det e^X = e^{\mathrm{tr}\,X}$, Section T3.3).

Worked example T3.23: U(1): the Maurer–Cartan form is $i\,\mathrm{d}\alpha$

Let $g = e^{i\alpha(x)}$ be a phase field. Then

$$g^{-1}\mathrm{d}g \overset{(1)}{=} e^{-i\alpha}\,\mathrm{d}\big(e^{i\alpha}\big) \overset{(2)}{=} e^{-i\alpha}\,e^{i\alpha}\,i\,\mathrm{d}\alpha = i\,\mathrm{d}\alpha.$$

(1) Definition.
(2) Chain rule: $\mathrm{d}(e^{i\alpha}) = e^{i\alpha}\,i\,\mathrm{d}\alpha$.

The answer is in $\mathfrak{u}(1) = i\mathbb{R}$, as it must be. On the group $\mathrm{U}(1) = S^1$ itself, with the angle α as coordinate, $\theta = i\,\mathrm{d}\alpha$. Note that $\mathrm{d}\alpha$ is a perfectly good 1-form on the circle even though α is not a function there (it jumps by 2π): this is the standard closed-but-not-exact form of Chapter T2. The Maurer–Cartan form of U(1) is i times the angle form.

Worked example T3.24: SO(2) written as matrices

With $R(\alpha) = \left(\begin{smallmatrix}\cos\alpha & -\sin\alpha\\ \sin\alpha & \cos\alpha\end{smallmatrix}\right)$,

$$R^{-1}\mathrm{d}R \overset{(1)}{=} \begin{pmatrix}\cos\alpha & \sin\alpha\\ -\sin\alpha & \cos\alpha\end{pmatrix}\begin{pmatrix}-\sin\alpha & -\cos\alpha\\ \cos\alpha & -\sin\alpha\end{pmatrix}\mathrm{d}\alpha$$
$$\overset{(2)}{=} \begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix}\mathrm{d}\alpha = L\,\mathrm{d}\alpha.$$

(1) $R^{-1} = R^T$ for a rotation; differentiate each entry of R.
(2) Multiply out: the $(1,1)$ entry is $-\cos\alpha\sin\alpha + \sin\alpha\cos\alpha = 0$, the $(1,2)$ entry is $-\cos^2\alpha - \sin^2\alpha = -1$, and so on.

The α-dependence has disappeared: θ is the constant generator L times $\mathrm{d}\alpha$. This is the matrix version of Example T3.23, and it is the connection 1-form of polar coordinates in disguise. Dray computes $\mathrm{d}\hat{r} = \mathrm{d}\phi\,\hat{\phi}$, $\mathrm{d}\hat{\phi} = -\mathrm{d}\phi\,\hat{r}$ [Dray (17.2)–(17.3)]. Put the frame in a row, $(\hat{r},\hat{\phi}) = (\hat{x},\hat{y})R(\phi)$: then $\mathrm{d}(\hat{r},\hat{\phi}) = (\hat{x},\hat{y})\mathrm{d}R = (\hat{r},\hat{\phi})\,R^{-1}\mathrm{d}R$, and $R^{-1}\mathrm{d}R = L\,\mathrm{d}\phi$ reproduces exactly Dray's two formulas. The connection 1-forms of a rotating frame *are* a Maurer–Cartan form. We return to this in Section T3.8.

T3.6.2 Left invariance

Why multiply on the left and not on the right? Either works, and each gives a form with a symmetry. For a constant group element h,

$$(hg)^{-1}\mathrm{d}(hg) \overset{(1)}{=} g^{-1}h^{-1}\,h\,\mathrm{d}g \overset{(2)}{=} g^{-1}\mathrm{d}g, \tag{T3.33}$$
$$(gh)^{-1}\mathrm{d}(gh) \overset{(3)}{=} h^{-1}g^{-1}\,\mathrm{d}g\,h \overset{(4)}{=} h^{-1}\,\theta\,h = \mathrm{Ad}_{h^{-1}}\theta. \tag{T3.34}$$

(1) $(hg)^{-1} = g^{-1}h^{-1}$, and $\mathrm{d}(hg) = h\,\mathrm{d}g$ because h is constant.
(2) $h^{-1}h = \mathbb{1}$.
(3) Similarly, $\mathrm{d}(gh) = \mathrm{d}g\,h$.
(4) Definition of θ and of the adjoint action (Section T3.4).

So θ is **left-invariant**: it does not change if we rotate the field from the left by a constant. Under a constant rotation from the right it rotates in the adjoint representation. The other choice, $\Theta \coloneqq \mathrm{d}g\,g^{-1}$, is **right-invariant** and rotates by Ad_h under $g \mapsto hg$. The two are related by $\Theta = g\,\theta\,g^{-1} = \mathrm{Ad}_g\theta$.

Physical picture: Body frame and space frame

Think of $g(t) \in \mathrm{SO}(3)$ as the orientation of a spinning top. Then $g^{-1}\dot{g}$ is the angular velocity measured in the *body* frame (what an observer riding on the top sees), and $\dot{g}\,g^{-1}$ is the angular velocity in the *space* frame. Rotating the laboratory (left multiplication) does not change what the rider sees: that is left invariance. Euler's equations for the rigid body are written in the body frame precisely because $g^{-1}\dot{g}$ does not care where the laboratory is.

T3.6.3 The structure equation $\mathrm{d}\theta + \theta \wedge \theta = 0$

The Maurer–Cartan form is not an arbitrary Lie-algebra-valued 1-form. It is the "derivative" of something, so its own derivative is constrained, just as a gradient field has zero curl. First we need the derivative of an inverse. Differentiating $g^{-1}g = \mathbb{1}$ gives $\mathrm{d}(g^{-1})\,g + g^{-1}\mathrm{d}g = 0$, so

$$\mathrm{d}(g^{-1}) = -g^{-1}\,\mathrm{d}g\,g^{-1}. \tag{T3.35}$$

This is the matrix version of $\mathrm{d}(1/x) = -\mathrm{d}x/x^2$. Now

$$\begin{aligned} \mathrm{d}\theta &\overset{(1)}{=} \mathrm{d}(g^{-1}) \wedge \mathrm{d}g + g^{-1}\,\mathrm{d}\mathrm{d}g \\ &\overset{(2)}{=} -g^{-1}\mathrm{d}g\,g^{-1} \wedge \mathrm{d}g \\ &\overset{(3)}{=} -(g^{-1}\mathrm{d}g) \wedge (g^{-1}\mathrm{d}g) = -\theta \wedge \theta. \end{aligned} \tag{T3.36}$$

(1) Product rule for the exterior derivative. The first factor g^{-1} is a (matrix-valued) 0-form, so no sign appears.
(2) Equation (T3.35), and $\mathrm{d}\mathrm{d} = 0$ applied to each matrix entry of g.
(3) Regroup the matrix product; the order of the matrices is kept, and the only wedge is between the two $\mathrm{d}g$'s.

Key idea: Maurer–Cartan structure equation

$$\mathrm{d}\theta + \theta \wedge \theta = 0, \qquad \text{in components} \qquad \mathrm{d}\theta^a + \tfrac{1}{2} f_{bc}{}^a\, \theta^b \wedge \theta^c = 0. \tag{T3.37}$$

It says that the "curvature" $\mathrm{d}A + A \wedge A$ of the special gauge field $A = g^{-1}\mathrm{d}g$ vanishes. We will see in Section T3.10 that this is exactly what "pure gauge" means.

The component form needs one small calculation, because $\theta \wedge \theta$ looks as if it should vanish (a 1-form wedged with itself). It does not, because the matrix coefficients do not commute:

$$\begin{aligned} \theta \wedge \theta &\overset{(1)}{=} \theta^b \wedge \theta^c\, T_b T_c \\ &\overset{(2)}{=} \tfrac{1}{2}\,\theta^b \wedge \theta^c\, \left(T_b T_c - T_c T_b\right) \\ &\overset{(3)}{=} \tfrac{1}{2}\,\theta^b \wedge \theta^c\, f_{bc}{}^a\, T_a. \end{aligned} \tag{T3.38}$$

(1) Expand $\theta = \theta^b T_b$ in both factors. The θ^b are ordinary 1-forms, so they can be moved past the constant matrices.
(2) $\theta^b \wedge \theta^c$ is antisymmetric in b, c, so only the antisymmetric part of $T_b T_c$ survives: write $T_b T_c = \tfrac{1}{2}[T_b, T_c] + \tfrac{1}{2}\{T_b, T_c\}$ and drop the symmetric anticommutator.
(3) Structure constants, (T3.6).

We often write this as $\theta \wedge \theta = \tfrac{1}{2}[\theta, \theta]$, where the bracket of two Lie-algebra-valued 1-forms means $[\alpha, \beta] = \alpha^b \wedge \beta^c\,[T_b, T_c]$. For an abelian group $\theta \wedge \theta = 0$ and the structure equation is just $\mathrm{d}\theta = 0$: for $\mathrm{U}(1)$, $\mathrm{d}(i\,\mathrm{d}\alpha) = 0$.

Watch out: Signs for matrix-valued forms

For matrix-valued forms, $\alpha \wedge \beta \neq \pm\beta \wedge \alpha$ in general, because the matrices do not commute. Two rules replace graded commutativity. (i) Under a trace, the *graded* cyclic rule holds: $\mathrm{tr}(\alpha \wedge \beta) = (-1)^{pq}\,\mathrm{tr}(\beta \wedge \alpha)$ for a p-form α and a q-form β. (ii) The product rule is $\mathrm{d}(\alpha \wedge \beta) = \mathrm{d}\alpha \wedge \beta + (-1)^p \alpha \wedge \mathrm{d}\beta$, with the matrices kept in order. Every calculation in the rest of this chapter uses only these two rules. For the right-invariant form $\Theta = \mathrm{d}g\, g^{-1}$ the same steps give $\mathrm{d}\Theta - \Theta \wedge \Theta = 0$: note the opposite sign.

The dual picture: left-invariant vector fields. For each generator T_a there is a vector field e_a on the group whose value at g is gT_a (the velocity of $t \mapsto g\, e^{tT_a}$). These are the vector fields dual to the θ^a: $\theta(e_a) = g^{-1}gT_a = T_a$, i.e. $\theta^b(e_a) = \delta^b_a$. Their Lie bracket is $[e_a, e_b] = f_{ab}{}^c e_c$, and the component equation (T3.37) is the same statement written with forms. (Use Cartan's formula $\mathrm{d}\theta^c(e_a, e_b) = e_a\big(\theta^c(e_b)\big) - e_b\big(\theta^c(e_a)\big) - \theta^c([e_a, e_b])$: the first two terms vanish because $\theta^c(e_b)$ is constant, leaving $\mathrm{d}\theta^c(e_a, e_b) = -f_{ab}{}^c$, which is what $-\frac{1}{2} f_{de}{}^c \theta^d \wedge \theta^e$ gives on (e_a, e_b).) So a Lie algebra is "the algebra of left-invariant vector fields", and this is how Nash thinks of $\mathfrak{g}$ for infinite-dimensional groups like $\mathrm{Diff}(X)$ [Nash p. 178].

T3.6.4 SU(2) in Euler angles: the forms written out

Let us compute the three 1-forms θ^a on $\mathrm{SU}(2) = S^3$ explicitly, and check the structure equation by hand. Parametrise

$$g(\phi, \vartheta, \psi) = e^{\phi T_3}\, e^{\vartheta T_2}\, e^{\psi T_3}, \qquad 0 \le \vartheta \le \pi, \quad 0 \le \phi < 2\pi, \quad 0 \le \psi < 4\pi, \tag{T3.39}$$

with $T_a = -\frac{i}{2}\sigma_a$. (The range 4π for ψ is needed because $e^{2\pi T_3} = -\mathbb{1}$: we are covering SU(2), not SO(3).) We need how conjugation by an exponential rotates the generators. From $[T_a, T_b] = \varepsilon_{abc} T_c$, ad_{T_a} acts on the components of $X = X^b T_b$ exactly as the rotation generator L_a acts on a 3-vector (Example T3.14), so $\mathrm{Ad}_{e^{\beta T_a}} = e^{\beta\,\mathrm{ad}_{T_a}}$ is the rotation by β about the a-th axis ((T3.22)):

$$e^{-\psi T_3}\, T_1\, e^{\psi T_3} = \cos\psi\, T_1 - \sin\psi\, T_2, \qquad e^{-\psi T_3}\, T_2\, e^{\psi T_3} = \sin\psi\, T_1 + \cos\psi\, T_2,$$
$$e^{-\vartheta T_2}\, T_3\, e^{\vartheta T_2} = \cos\vartheta\, T_3 - \sin\vartheta\, T_1.$$

(These are rotations by $-\psi$ about the 3-axis and $-\vartheta$ about the 2-axis applied to the unit vectors $\hat{e}_1, \hat{e}_2, \hat{e}_3$. Check one: to first order in ψ, $e^{-\psi T_3} T_1 e^{\psi T_3} = T_1 - \psi[T_3, T_1] = T_1 - \psi T_2$.) Now differentiate the product (T3.39) with the product rule:

$$\begin{aligned} g^{-1}\mathrm{d}g &\overset{(1)}{=} e^{-\psi T_3} e^{-\vartheta T_2}\,(T_3\,\mathrm{d}\phi)\, e^{\vartheta T_2} e^{\psi T_3} + e^{-\psi T_3}\,(T_2\,\mathrm{d}\vartheta)\, e^{\psi T_3} + T_3\,\mathrm{d}\psi \\ &\overset{(2)}{=} \mathrm{d}\phi\; e^{-\psi T_3}\big(\cos\vartheta\, T_3 - \sin\vartheta\, T_1\big) e^{\psi T_3} + \mathrm{d}\vartheta\,\big(\sin\psi\, T_1 + \cos\psi\, T_2\big) + \mathrm{d}\psi\, T_3 \\ &\overset{(3)}{=} \big(\sin\psi\,\mathrm{d}\vartheta - \sin\vartheta\cos\psi\,\mathrm{d}\phi\big) T_1 + \big(\cos\psi\,\mathrm{d}\vartheta + \sin\vartheta\sin\psi\,\mathrm{d}\phi\big) T_2 \\ &\qquad + \big(\mathrm{d}\psi + \cos\vartheta\,\mathrm{d}\phi\big) T_3. \end{aligned}$$

(1) $\mathrm{d}(e^{\phi T_3}) = e^{\phi T_3} T_3\,\mathrm{d}\phi$ and similarly for the other factors; after multiplying by $g^{-1} = e^{-\psi T_3} e^{-\vartheta T_2} e^{-\phi T_3}$ on the left, each derivative is sandwiched between the factors to its right and their inverses.

(2) The conjugation formulas for T_3 by $e^{\vartheta T_2}$ and for T_2 by $e^{\psi T_3}$; T_3 commutes with $e^{\psi T_3}$.

(3) Conjugate T_1 by $e^{\psi T_3}$ with the first formula and collect the coefficients of T_1, T_2, T_3.

So the left-invariant forms of SU(2) are

$$\begin{aligned}\theta^1 &= \sin\psi\,\mathrm{d}\vartheta - \sin\vartheta\cos\psi\,\mathrm{d}\phi,\\ \theta^2 &= \cos\psi\,\mathrm{d}\vartheta + \sin\vartheta\sin\psi\,\mathrm{d}\phi,\\ \theta^3 &= \mathrm{d}\psi + \cos\vartheta\,\mathrm{d}\phi.\end{aligned} \tag{T3.40}$$

Worked example T3.25: Checking the structure equation by hand

With $f_{bc}{}^{a} = \varepsilon_{abc}$, (T3.37) reads $\mathrm{d}\theta^3 + \theta^1\wedge\theta^2 = 0$ and cyclic permutations. For the third component:

$$\begin{aligned}\theta^1\wedge\theta^2 &\overset{(1)}{=} \sin^2\psi\,\sin\vartheta\,\mathrm{d}\vartheta\wedge\mathrm{d}\phi - \sin\vartheta\cos^2\psi\,\mathrm{d}\phi\wedge\mathrm{d}\vartheta\\ &\overset{(2)}{=} \sin\vartheta\left(\sin^2\psi+\cos^2\psi\right)\mathrm{d}\vartheta\wedge\mathrm{d}\phi = \sin\vartheta\,\mathrm{d}\vartheta\wedge\mathrm{d}\phi,\\ \mathrm{d}\theta^3 &\overset{(3)}{=} \mathrm{d}(\cos\vartheta)\wedge\mathrm{d}\phi = -\sin\vartheta\,\mathrm{d}\vartheta\wedge\mathrm{d}\phi.\end{aligned}$$

(1) Multiply out; the terms $\mathrm{d}\vartheta\wedge\mathrm{d}\vartheta$ and $\mathrm{d}\phi\wedge\mathrm{d}\phi$ vanish, leaving the two cross terms.
(2) $\mathrm{d}\phi\wedge\mathrm{d}\vartheta = -\mathrm{d}\vartheta\wedge\mathrm{d}\phi$.
(3) $\mathrm{dd}\psi = 0$ and $\mathrm{dd}\phi = 0$.

The sum is zero. For the first component, $\mathrm{d}\theta^1 = \cos\psi\,\mathrm{d}\psi\wedge\mathrm{d}\vartheta - \cos\vartheta\cos\psi\,\mathrm{d}\vartheta\wedge\mathrm{d}\phi + \sin\vartheta\sin\psi\,\mathrm{d}\psi\wedge\mathrm{d}\phi$ and $\theta^2\wedge\theta^3 = \cos\psi\,\mathrm{d}\vartheta\wedge\mathrm{d}\psi + \cos\psi\cos\vartheta\,\mathrm{d}\vartheta\wedge\mathrm{d}\phi + \sin\vartheta\sin\psi\,\mathrm{d}\phi\wedge\mathrm{d}\psi$; the three pairs cancel term by term. The Maurer–Cartan equation is not an abstract identity: it is a statement about these trigonometric functions.

The metric and the volume. Because the θ^a are left-invariant, anything built from them with constant coefficients is left-invariant too. Two things are especially useful.

- The metric $\mathrm{d}s^2 = (\theta^1)^2 + (\theta^2)^2 + (\theta^3)^2 = -2\,\mathrm{tr}(\theta\otimes\theta)$ is invariant under both left and right multiplication (right multiplication rotates θ^a by Ad, an orthogonal matrix). From (T3.40),
$$\mathrm{d}s^2 = \mathrm{d}\vartheta^2 + \mathrm{d}\phi^2 + \mathrm{d}\psi^2 + 2\cos\vartheta\,\mathrm{d}\phi\,\mathrm{d}\psi,$$
which is the round metric on the 3-sphere of radius 2. (The radius is 2 rather than 1 because $T_a = -\frac{i}{2}\sigma_a$ carries a $\frac{1}{2}$: going once round the group takes ψ from 0 to 4π.)

- The volume form $\theta^1\wedge\theta^2\wedge\theta^3$. Using the first part of Example T3.25,
$$\begin{aligned}\theta^1\wedge\theta^2\wedge\theta^3 &\overset{(1)}{=} \sin\vartheta\,\mathrm{d}\vartheta\wedge\mathrm{d}\phi\wedge(\mathrm{d}\psi+\cos\vartheta\,\mathrm{d}\phi) \overset{(2)}{=} \sin\vartheta\,\mathrm{d}\vartheta\wedge\mathrm{d}\phi\wedge\mathrm{d}\psi,\\ \int_{\mathrm{SU}(2)}\theta^1\wedge\theta^2\wedge\theta^3 &\overset{(3)}{=} \int_0^{\pi}\sin\vartheta\,\mathrm{d}\vartheta\int_0^{2\pi}\mathrm{d}\phi\int_0^{4\pi}\mathrm{d}\psi = 2\cdot 2\pi\cdot 4\pi = 16\pi^2.\end{aligned}$$
(1) $\theta^1\wedge\theta^2 = \sin\vartheta\,\mathrm{d}\vartheta\wedge\mathrm{d}\phi$.
(2) $\mathrm{d}\phi\wedge\mathrm{d}\phi = 0$.
(3) The ranges (T3.39), with the orientation $\mathrm{d}\vartheta\wedge\mathrm{d}\phi\wedge\mathrm{d}\psi$.

This is $2^3\times 2\pi^2$, the volume of a 3-sphere of radius 2 (the unit S^3 has volume $2\pi^2$), as the metric predicted.

Physical picture: Haar measure

The normalised volume $\mathrm{d}\mu = \theta^1 \wedge \theta^2 \wedge \theta^3/16\pi^2$ is the **Haar measure**: the unique way of "averaging over the group" that is unchanged by left and right multiplication, $\int f(hg)\,\mathrm{d}\mu(g) = \int f(g)\,\mathrm{d}\mu(g)$. It is the measure used for the link variables of lattice gauge theory, and for the "average over the group" that makes every representation of a compact group unitary (Section T3.5). For a function of the conjugacy class only (a function of the rotation angle θ, like a character), the Haar measure reduces to $\frac{1}{\pi}\sin^2\frac{\theta}{2}\,\mathrm{d}\theta$ on $[0, 2\pi]$: the classes near $\pm\mathbb{1}$ are small 2-spheres (Figure T3.3) and carry little weight.

T3.6.5 The cube of θ counts windings

Here is the promised topological use. Consider first a map from a circle into U(1), $g(\varphi) = e^{i\alpha(\varphi)}$ with $\alpha(2\pi) = \alpha(0) + 2\pi n$. By Example T3.23,

$$\frac{1}{2\pi i}\oint_{S^1} g^{-1}\mathrm{d}g = \frac{1}{2\pi}\oint \mathrm{d}\alpha = \frac{\alpha(2\pi) - \alpha(0)}{2\pi} = n. \tag{T3.41}$$

The integral of the Maurer–Cartan form counts how many times the loop winds around the group: it is the element of $\pi_1(\mathrm{U}(1)) = \mathbb{Z}$ of Table T3.2.

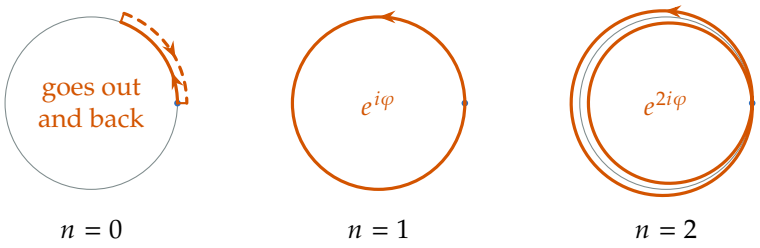


Figure T3.17: Loops $\varphi \mapsto g(\varphi)$ in U(1) (drawn slightly off the circle when they overlap). The left loop can be pulled back to a point; the others wind once and twice and cannot. The number n is computed by the integral $\frac{1}{2\pi i}\oint g^{-1}\mathrm{d}g$, which is unchanged by any continuous deformation of the loop.

For maps from a 3-sphere into SU(2) (or any compact simple group) the analogue is built from the cube of θ. We need the trace of three generators. For $\mathfrak{su}(2)$, only the part antisymmetric in a, b, c survives when contracted with $\theta^a \wedge \theta^b \wedge \theta^c$, and

$$\mathrm{tr}\big(T_a\,[T_b, T_c]\big) \overset{(1)}{=} \varepsilon_{bcd}\ \mathrm{tr}(T_a T_d) \overset{(2)}{=} -\tfrac{1}{2}\,\varepsilon_{bca} = -\tfrac{1}{2}\,\varepsilon_{abc}.$$

(1) $[T_b, T_c] = \varepsilon_{bcd} T_d$.
(2) $\mathrm{tr}(T_a T_d) = -\frac{1}{2}\delta_{ad}$; and $\varepsilon_{bca} = \varepsilon_{abc}$ (cyclic).

Therefore

$$\begin{aligned}\mathrm{tr}(\theta \wedge \theta \wedge \theta) &\overset{(1)}{=} \theta^a \wedge \theta^b \wedge \theta^c\ \mathrm{tr}(T_a T_b T_c)\\ &\overset{(2)}{=} \tfrac{1}{2}\,\theta^a \wedge \theta^b \wedge \theta^c\ \mathrm{tr}\big(T_a [T_b, T_c]\big)\\ &\overset{(3)}{=} -\tfrac{1}{4}\,\varepsilon_{abc}\,\theta^a \wedge \theta^b \wedge \theta^c \overset{(4)}{=} -\tfrac{3}{2}\,\theta^1 \wedge \theta^2 \wedge \theta^3.\end{aligned} \tag{T3.42}$$

(1) Expand each θ in the basis; the scalar 1-forms come out of the trace.
(2) Antisymmetry of $\theta^b \wedge \theta^c$ in b, c keeps only $\frac{1}{2}[T_b, T_c]$, as in (T3.38).
(3) The trace just computed.

(4) The sum over a, b, c has $3! = 6$ nonzero terms, each equal to $+\theta^1 \wedge \theta^2 \wedge \theta^3$ (both ε and the wedge change sign under a swap).

Integrating over the group with the volume computed above,

$$\int_{\mathrm{SU}(2)} \operatorname{tr}(\theta \wedge \theta \wedge \theta) = -\tfrac{3}{2} \cdot 16\pi^2 = -24\pi^2. \tag{T3.43}$$

Working definition T3.13 (winding number of a map $S^3 \to \mathrm{SU}(2)$)

For a map $g : S^3 \to \mathrm{SU}(2)$ (or into $\mathrm{SU}(N)$), the **winding number** (degree) [how many times g covers SU(2), each covering counted $+1$ or -1 according to orientation] is

$$w(g) = -\frac{1}{24\pi^2} \int_{S^3} \operatorname{tr}\big(g^{-1}\mathrm{d}g\big)^3, \tag{T3.44}$$

where $(g^{-1}\mathrm{d}g)^3$ means the triple wedge product. *How to compute it:* pull back, i.e. substitute $g(x)$, compute the 3×3 determinant-like combination $\varepsilon^{ijk} \operatorname{tr}(g^{-1}\partial_i g\, g^{-1}\partial_j g\, g^{-1}\partial_k g)$ in coordinates on S^3 and integrate. It is an integer, it is unchanged by continuous deformations of g, and it labels $\pi_3(\mathrm{SU}(2)) = \mathbb{Z}$ [maps that can be continuously deformed into each other count as one class, and the classes are labelled by one integer].

The normalisation is fixed by (T3.43): for the identity map $S^3 \to \mathrm{SU}(2)$, with S^3 oriented by $\theta^1 \wedge \theta^2 \wedge \theta^3$ (which, at the north pole $\mathbb{1}$, is the standard orientation of S^3 as the boundary of the unit ball in $\mathbb{R}^4$), $w = 1$. For a map that covers the group k times, the integral is k times as big.

Watch out: The sign in front

Many books write $w = +\frac{1}{24\pi^2}\int \operatorname{tr}(g^{-1}\mathrm{d}g)^3$. The difference is a choice of orientation of S^3 or of Hermitian versus anti-Hermitian generators. With anti-Hermitian θ and the orientation above, the minus sign is needed for the identity map to have $w = +1$. The advantage of our sign is that the instanton number $k = \frac{1}{8\pi^2}\int \operatorname{tr} F \wedge F$ then equals w of the gauge transformation at infinity, with no extra sign (Section T3.10).

Why is w unchanged by deformations? Deform g to $g\, e^{\varepsilon X}$ with $X(x) \in \mathfrak{g}$ small. To first order the Maurer–Cartan form changes by

$$\delta\theta \overset{(1)}{=} \big(g(1+\varepsilon X)\big)^{-1}\mathrm{d}\big(g(1+\varepsilon X)\big) - g^{-1}\mathrm{d}g \overset{(2)}{=} \varepsilon\,\big(\mathrm{d}X + \theta X - X\theta\big) + O(\varepsilon^2),$$

(1) Definition; $e^{\varepsilon X} = 1 + \varepsilon X + O(\varepsilon^2)$.
(2) $(g(1+\varepsilon X))^{-1} = (1-\varepsilon X)g^{-1}$; expand $(1-\varepsilon X)g^{-1}(\mathrm{d}g + \varepsilon\,\mathrm{d}g\,X + \varepsilon\, g\,\mathrm{d}X)$ to first order.

and the change of the integrand is exact:

$$\begin{aligned}\delta \operatorname{tr}\theta^3 &\overset{(1)}{=} 3\operatorname{tr}\big(\theta^2\,\delta\theta\big) \overset{(2)}{=} 3\varepsilon\,\operatorname{tr}\big(\theta^2\mathrm{d}X\big) + 3\varepsilon\,\operatorname{tr}\big(\theta^3 X - \theta^2 X\theta\big)\\ &\overset{(3)}{=} 3\varepsilon\,\operatorname{tr}\big(\theta^2\mathrm{d}X\big) \overset{(4)}{=} 3\varepsilon\,\mathrm{d}\operatorname{tr}\big(\theta^2 X\big).\end{aligned}$$

(1) Vary each of the three factors; graded cyclicity of the trace brings each varied factor to the end with sign $+1$ (a 1-form moved past a 2-form).
(2) Substitute $\delta\theta$.
(3) $\operatorname{tr}(\theta^2 X\theta) = \operatorname{tr}(\theta\,\theta^2 X) = \operatorname{tr}(\theta^3 X)$ by graded cyclicity (moving the 1-form θ past the 2-form $\theta^2 X$ costs $(-1)^2 = +1$).

(4) $\mathrm{d}\,\mathrm{tr}(\theta^2 X) = \mathrm{tr}(\mathrm{d}(\theta^2)X) + \mathrm{tr}(\theta^2 \mathrm{d}X)$, and $\mathrm{d}(\theta^2) = \mathrm{d}\theta\,\theta - \theta\,\mathrm{d}\theta = -\theta^3 + \theta^3 = 0$ by the structure equation.

By Stokes' theorem an exact 3-form integrates to zero on the closed manifold S^3, so $\delta w = 0$. This is the exemplar argument of the style guide in a new setting: *topological invariance is Stokes' theorem applied to an exact variation.*

Worked example T3.26: Windings add: $w(gh) = w(g) + w(h)$

Let $g, h : S^3 \to \mathrm{SU}(2)$ and multiply them pointwise. Write $a = g^{-1}\mathrm{d}g$ and $b = \mathrm{d}h\,h^{-1}$ (the right-invariant form of h). Then

$$(gh)^{-1}\mathrm{d}(gh) \overset{(1)}{=} h^{-1}g^{-1}(\mathrm{d}g\,h + g\,\mathrm{d}h) = h^{-1}a\,h + h^{-1}\mathrm{d}h \overset{(2)}{=} h^{-1}(a+b)\,h,$$
$$\mathrm{tr}\big((gh)^{-1}\mathrm{d}(gh)\big)^3 \overset{(3)}{=} \mathrm{tr}(a+b)^3 \overset{(4)}{=} \mathrm{tr}\,a^3 + \mathrm{tr}\,b^3 + 3\,\mathrm{tr}(a^2b) + 3\,\mathrm{tr}(ab^2).$$

(1) Product rule.
(2) $h^{-1}\mathrm{d}h = h^{-1}(\mathrm{d}h\,h^{-1})h = h^{-1}bh$.
(3) Conjugation by h cancels inside the trace.
(4) Expand the cube keeping the order of factors; graded cyclicity for three 1-forms has no sign, so e.g. $\mathrm{tr}(aba) = \mathrm{tr}(a^2b)$.

The cross terms are exact. Using $\mathrm{d}a = -a^2$ and $\mathrm{d}b = +b^2$ (the two structure equations), $\mathrm{d}\,\mathrm{tr}(ab) = \mathrm{tr}(\mathrm{d}a\,b) - \mathrm{tr}(a\,\mathrm{d}b) = -\mathrm{tr}(a^2b) - \mathrm{tr}(ab^2)$. So $3\,\mathrm{tr}(a^2b) + 3\,\mathrm{tr}(ab^2) = -3\,\mathrm{d}\,\mathrm{tr}(ab)$, which integrates to zero. And $\mathrm{tr}\,b^3 = \mathrm{tr}(h\,h^{-1}\mathrm{d}h\,h^{-1}\mathrm{d}h\,h^{-1}\mathrm{d}h\,h^{-1}) = \mathrm{tr}(h^{-1}\mathrm{d}h)^3$. Integrating,

$$w(gh) = w(g) + w(h), \qquad \text{and in particular} \qquad w(g^{-1}) = -w(g), \quad w(g^k) = k\,w(g).$$

The winding number turns pointwise multiplication of maps into addition of integers: it is the group isomorphism $\pi_3(\mathrm{SU}(2)) \cong \mathbb{Z}$.

Worked example T3.27: The basic winding-one map on $\mathbb{R}^3$ and on S^3_∞

Two versions of the identity map $S^3 \to \mathrm{SU}(2)$ appear constantly. *On the 3-sphere at infinity of* $\mathbb{R}^4$: for $x = (x^1, x^2, x^3, x^4)$ with $|x| = r$,

$$g(x) = \frac{x^4 \mathbb{1} + i\,\vec{x}\cdot\vec{\sigma}}{r}.$$

On the sphere $|x| = r$ this is (T3.2) with a rescaled radius, so it covers SU(2) exactly once: $w = 1$. This is the gauge transformation at infinity of the BPST instanton (Chapter 8). *On* $\mathbb{R}^3$ *compactified to* S^3: stereographic projection sends $\vec{y} \in \mathbb{R}^3$ to the point $\big(\frac{2\vec{y}}{1+|\vec{y}|^2}, \frac{|\vec{y}|^2-1}{|\vec{y}|^2+1}\big)$ of the unit S^3, and

$$g(\vec{y}) = \frac{(|\vec{y}|^2 - 1)\mathbb{1} + 2i\,\vec{y}\cdot\vec{\sigma}}{|\vec{y}|^2 + 1}$$

tends to $\mathbb{1}$ as $|\vec{y}| \to \infty$ and equals $-\mathbb{1}$ at the origin. As $\vec{y}$ ranges over $\mathbb{R}^3$ it covers S^3 once, so $w = \pm 1$ (the sign depends on the orientation of $\mathbb{R}^3$). Such "large" gauge transformations, which cannot be deformed to the identity, are what split the gauge group in three dimensions into components labelled by $\pi_3(\mathrm{SU}(2)) = \mathbb{Z}$ [Nash p. 227, (8.59)].

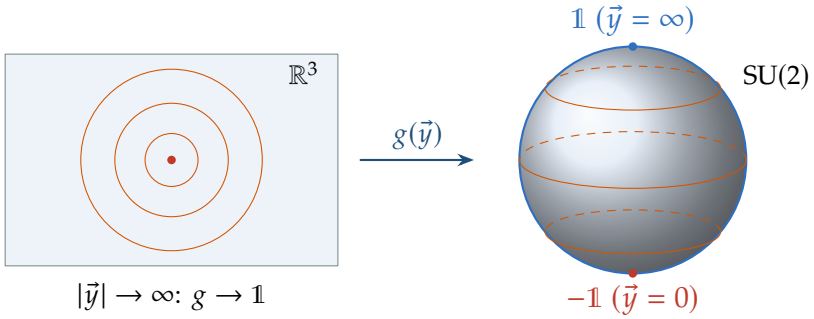


Figure T3.18: The winding-one map of Example T3.27. Spheres $|\vec{y}| = \text{const}$ in $\mathbb{R}^3$ (orange rings) go to the 2-spheres of latitude of SU(2) = S^3 (conjugacy classes, Figure T3.3); the origin goes to the south pole $-\mathbb{1}$ and infinity to the north pole $\mathbb{1}$. Every point of the group is hit exactly once, so $\frac{1}{24\pi^2}\left|\int \mathrm{tr}(g^{-1}\mathrm{d}g)^3\right| = 1$.

Try it: Winding number of a map that misses a point

Suppose $g : S^3 \to \mathrm{SU}(2)$ never takes the value $-\mathbb{1}$. Show that $w(g) = 0$.
Answer: SU(2) minus a point is a 3-sphere minus a point, which can be contracted to $\mathbb{1}$ (slide everything away from $-\mathbb{1}$ along the great circles through it). So g can be deformed to the constant map, whose Maurer–Cartan form vanishes, and w is deformation invariant. A nonzero winding number means the map is *onto*.

Reading Nash (Maurer–Cartan forms in Nash)

Nash never names $g^{-1}\mathrm{d}g$, but it runs through chapter 8. His gauge transformation law $g^{-1}(A+\mathrm{d})g = A$ for a fixed point [Nash p. 232, (8.88)] reads, once expanded, $g^{-1}Ag + g^{-1}\mathrm{d}g = A$: the Maurer–Cartan form is the inhomogeneous term. The "transgression" formulas [Nash p. 224, (8.41)–(8.42)] applied to the pure-gauge connection $A = g^{-1}\mathrm{d}g$ produce, up to normalisation, $\mathrm{tr}(g^{-1}\mathrm{d}g)^3$; his forms μ_{2j-1} on the gauge group [Nash pp. 225–226, (8.48)–(8.58)] are generalisations of our winding number density to families of maps. The statement $\pi_0(\mathcal{G}) = \pi_3(G) = \mathbb{Z}$ [Nash p. 227, (8.59)] is the fact that w of (T3.44) labels the components of the group of gauge transformations of S^3.

Recap

- $\theta = g^{-1}\mathrm{d}g$ carries the derivative of a group-valued field back to the identity; it is a $\mathfrak{g}$-valued 1-form, $\theta = \theta^a T_a$.
- Left-invariant: $(hg)^{-1}\mathrm{d}(hg) = \theta$; under $g \mapsto gh$ it rotates by $\mathrm{Ad}_{h^{-1}}$. U(1): $\theta = i\,\mathrm{d}\alpha$.
- Structure equation $\mathrm{d}\theta + \theta \wedge \theta = 0$, i.e. $\mathrm{d}\theta^a + \frac{1}{2} f_{bc}{}^a \theta^b \wedge \theta^c = 0$ (from $\mathrm{d}(g^{-1}) = -g^{-1}\mathrm{d}g\, g^{-1}$).
- SU(2) in Euler angles: (T3.40); the metric $\sum(\theta^a)^2$ is the round S^3 of radius 2, volume $16\pi^2$.
- $\mathrm{tr}\,\theta^3 = -\frac{3}{2}\theta^1\theta^2\theta^3$, $\int_{\mathrm{SU}(2)} \mathrm{tr}\,\theta^3 = -24\pi^2$; winding number $w = -\frac{1}{24\pi^2}\int \mathrm{tr}(g^{-1}\mathrm{d}g)^3 \in \mathbb{Z}$, invariant under deformation (the variation is exact) and additive.

T3.7 Group actions, orbits and homogeneous spaces

A group is interesting because of what it *does*. Rotations move points of space; SU(2) moves spin states; gauge transformations move gauge fields. Very often the thing being moved is itself a curved space: the direction of a spin is a point on a 2-sphere, the vacuum of a Higgs field is a point on a sphere, a quantum state up to phase is a point of a complex projective space. The question of this section is: *how do we describe a space on which a group acts, and how do we build such spaces out of groups?* The answer is the orbit–stabiliser principle, which says that a space on which G acts transitively is a quotient G/H. This one idea produces spheres, projective spaces and flag manifolds from matrices, gives the Hopf fibration $S^3 \to S^2$ that underlies the Dirac monopole, and prepares the language of principal bundles and of the gauge group acting on the space of connections.

T3.7.1 Actions

Working definition T3.14 (group action)

An **action** of a group G on a space X is a rule that assigns to each $g \in G$ a map $x \mapsto g \cdot x$ of X to itself, such that

$$\mathbb{1} \cdot x = x, \qquad g \cdot (h \cdot x) = (gh) \cdot x.$$

Operational test: check that the identity does nothing and that doing h then g is the same as doing gh. A representation (Section T3.5) is the special case in which X is a vector space and each $x \mapsto g \cdot x$ is linear [$g \cdot (av + bw) = a\, g \cdot v + b\, g \cdot w$, so each g acts by a matrix].

Examples we already know, and some new ones:

- SO(3) acts on $\mathbb{R}^3$ by $R \cdot \vec{x} = R\vec{x}$; SU(2) acts on $\mathbb{C}^2$ by $U \cdot \psi = U\psi$.
- G acts on itself by *left multiplication* $g \cdot h = gh$ and by *conjugation* $g \cdot h = ghg^{-1}$, and on its Lie algebra by Ad (Section T3.4).
- SU(2) acts on the unit 2-sphere through SO(3): $U \cdot \hat{n}$ is defined by $(U \cdot \hat{n}) \cdot \vec{\sigma} = U(\hat{n} \cdot \vec{\sigma})U^\dagger$.
- U(1) acts on $\mathbb{C}^{n+1}$ by an overall phase, $e^{i\alpha} \cdot z = e^{i\alpha} z$.
- The group $\mathcal{G}$ of gauge transformations $g(x)$ acts on gauge fields by $g \cdot A = g^{-1}Ag + g^{-1}\mathrm{d}g$ (Section T3.8). (Strictly this is a right action: $(gh) \cdot A = h \cdot (g \cdot A)$. Nothing below depends on the difference.)

T3.7.2 Orbits and stabilisers

Working definition T3.15 (orbit and stabiliser)

The **orbit** of a point x_0 is the set of all points it can be moved to, $G \cdot x_0 = \{g \cdot x_0 : g \in G\}$. The **stabiliser** (or isotropy group, or little group [the physicist's name when x_0 is a momentum p and G is the Lorentz group]) of x_0 is the set of group elements that do not move it, $H = G_{x_0} = \{g : g \cdot x_0 = x_0\}$. It is a subgroup: if g and h fix x_0, so do gh and g^{-1}. The action is **transitive** if there is only one orbit (every point can be reached from every other), and **free** if every stabiliser is trivial ($g \cdot x = x$ only for $g = \mathbb{1}$).

Worked example T3.28: Rotations of $\mathbb{R}^3$: spheres and axes

Take SO(3) acting on $\mathbb{R}^3$ and a point $\vec{x}_0 \neq 0$. A rotation preserves length, so the orbit lies in the sphere $|\vec{x}| = |\vec{x}_0|$; and any point of that sphere can be reached by rotating about the axis $\vec{x}_0 \times \vec{x}$. So the orbit is the whole 2-sphere of radius $|\vec{x}_0|$. The stabiliser of $\vec{x}_0 = r\hat{z}$ consists of the rotations that fix $\hat{z}$, which are the rotations $R_z(\alpha)$ about the z-axis: $H \cong \mathrm{SO}(2)$. Dimensions:

$$\dim(\text{orbit}) = 2 = 3 - 1 = \dim \mathrm{SO}(3) - \dim \mathrm{SO}(2).$$

The origin is a special orbit on its own, a single point, with stabiliser the whole group. So $\mathbb{R}^3$ is sliced into orbits: concentric spheres plus one point (Figure T3.19). The orbits are labelled by one number, the radius, which is the *invariant* of the action.

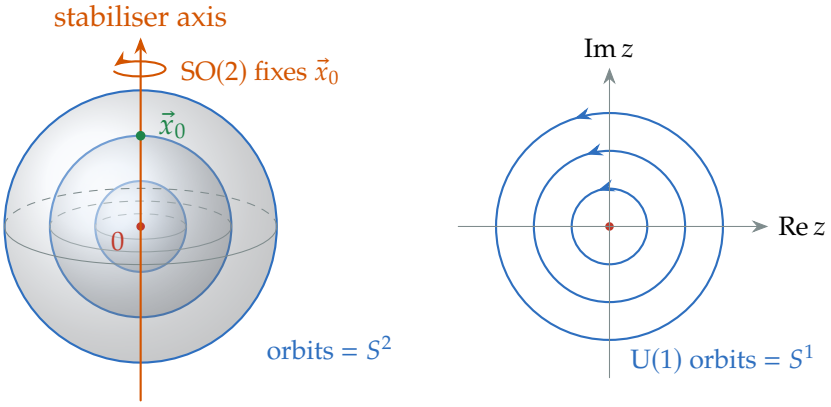


Figure T3.19: Left: SO(3) acting on $\mathbb{R}^3$. The orbits are concentric spheres (blue) and the single point 0 (red). A point $\vec{x}_0$ on the z-axis is fixed by the SO(2) of rotations about that axis. Right: U(1) acting on $\mathbb{C}$ by phases: orbits are circles, and the action is free except at the origin. In both pictures the orbit through x_0 is a copy of G/H.

T3.7.3 Cosets and the orbit–stabiliser principle

Two group elements g and gh with $h \in H$ move x_0 to the same place, because $gh \cdot x_0 = g \cdot (h \cdot x_0) = g \cdot x_0$. So points of the orbit correspond not to single group elements but to whole families $gH = \{gh : h \in H\}$, called **left cosets** of H. The set of left cosets is written G/H.

Key idea: Orbit–stabiliser

The map $G/H \to G \cdot x_0$, $gH \mapsto g \cdot x_0$, is well defined and a one-to-one correspondence. So every orbit is a quotient G/H of the group by the stabiliser, and

$$\dim(G \cdot x_0) = \dim G - \dim H.$$

A space on which G acts transitively is called a **homogeneous space**; it is G/H for the stabiliser H of any of its points.

Why one-to-one: *well defined*, because different representatives g, gh of the same coset give the same point (just shown); *onto*, by the definition of the orbit; *injective*, because

$$g_1 \cdot x_0 = g_2 \cdot x_0 \overset{(1)}{\Longleftrightarrow} g_2^{-1} g_1 \cdot x_0 = x_0 \overset{(2)}{\Longleftrightarrow} g_2^{-1} g_1 \in H \overset{(3)}{\Longleftrightarrow} g_1 H = g_2 H.$$

(1) Act on both sides with g_2^{-1} and use the action rule $g_2^{-1} \cdot (g_1 \cdot x_0) = (g_2^{-1} g_1) \cdot x_0$.

(2) Definition of the stabiliser.
(3) $g_1 = g_2 h$ with $h \in H$ means exactly that g_1 lies in the coset of g_2.

The dimension formula follows by differentiating: near x_0 the orbit is swept out by $e^{tX} \cdot x_0$, $X \in \mathfrak{g}$, and the directions $X \in \mathfrak{h}$ do not move x_0, so the tangent space of the orbit is $\mathfrak{g}/\mathfrak{h}$.

Watch out: G/H is usually not a group

G/H is a space, not a group, unless H is a *normal* subgroup ($gHg^{-1} = H$ for all g). $\mathrm{SO}(3) = \mathrm{SU}(2)/\{\pm\mathbb{1}\}$ is a group because $\{\pm\mathbb{1}\}$ is central. $S^2 = \mathrm{SU}(2)/\mathrm{U}(1)$ is not: the $\mathrm{U}(1)$ of diagonal matrices is not normal, and the 2-sphere has no continuous group structure.

T3.7.4 Spheres as homogeneous spaces

Worked example T3.29: $S^{n-1} = \mathrm{SO}(n)/\mathrm{SO}(n-1)$

$\mathrm{SO}(n)$ acts on the unit sphere $S^{n-1} \subset \mathbb{R}^n$ transitively (the first column of a rotation can be any unit vector: complete it to an oriented orthonormal basis). The stabiliser of the north pole e_n is the set of R with $Re_n = e_n$. Such an R has last column e_n, and because its rows are orthonormal its last row is e_n^T as well:

$$R = \begin{pmatrix} R' & 0 \\ 0 & 1 \end{pmatrix}, \qquad R' \in \mathrm{SO}(n-1).$$

So $S^{n-1} = \mathrm{SO}(n)/\mathrm{SO}(n-1)$. Dimension check: $\frac{1}{2}n(n-1) - \frac{1}{2}(n-1)(n-2) = n-1$. ✓ In the same way $\mathrm{U}(n)$ and $\mathrm{SU}(n)$ act transitively on the unit sphere $S^{2n-1} \subset \mathbb{C}^n$ with stabiliser $\mathrm{U}(n-1)$ and $\mathrm{SU}(n-1)$: $S^{2n-1} = \mathrm{SU}(n)/\mathrm{SU}(n-1)$. For $n = 2$ the stabiliser $\mathrm{SU}(1)$ is trivial and we recover $S^3 = \mathrm{SU}(2)$.

T3.7.5 The 2-sphere as $\mathrm{SU}(2)/\mathrm{U}(1)$, and the Hopf map

This is the example to know by heart. Let $\mathrm{SU}(2)$ act on traceless Hermitian matrices of unit "length" $\hat{n} \cdot \vec{\sigma}$ by conjugation. Every such matrix has eigenvalues ± 1 and so is conjugate to σ_3: the action is transitive on the 2-sphere of unit vectors $\hat{n}$. Define

$$\pi : \mathrm{SU}(2) \to S^2, \qquad \pi(U) \cdot \vec{\sigma} = U \sigma_3 U^\dagger. \tag{T3.45}$$

The stabiliser of σ_3 is the set of U commuting with σ_3, which is the diagonal subgroup $\{e^{i\chi\sigma_3} = \mathrm{diag}(e^{i\chi}, e^{-i\chi})\} \cong \mathrm{U}(1)$. By orbit–stabiliser, $S^2 = \mathrm{SU}(2)/\mathrm{U}(1)$, and the dimensions agree: $3 - 1 = 2$.

Let us write π out. With $U = \left(\begin{smallmatrix} a & -\bar{b} \\ b & \bar{a} \end{smallmatrix}\right)$, $|a|^2 + |b|^2 = 1$, as in (T3.1), call the columns $u_1 = (a, b)^T$ and $u_2 = (-\bar{b}, \bar{a})^T$. Then

$$\begin{aligned} U\sigma_3 U^\dagger &\overset{(1)}{=} u_1 u_1^\dagger - u_2 u_2^\dagger \\ &\overset{(2)}{=} \begin{pmatrix} |a|^2 & a\bar{b} \\ \bar{a}b & |b|^2 \end{pmatrix} - \begin{pmatrix} |b|^2 & -a\bar{b} \\ -\bar{a}b & |a|^2 \end{pmatrix} = \begin{pmatrix} |a|^2 - |b|^2 & 2a\bar{b} \\ 2\bar{a}b & |b|^2 - |a|^2 \end{pmatrix}. \end{aligned}$$

(1) $U\sigma_3 U^\dagger = U\,\mathrm{diag}(1,-1)\,U^\dagger = \sum_k (\sigma_3)_{kk} u_k u_k^\dagger$, the spectral decomposition with eigenvectors the columns of U.
(2) Outer products: $(u_1 u_1^\dagger)_{ij} = (u_1)_i \overline{(u_1)_j}$, and likewise for u_2.

Comparing with $\hat{n}\cdot\vec{\sigma} = \begin{pmatrix} n_3 & n_1 - in_2 \\ n_1 + in_2 & -n_3 \end{pmatrix}$:

$$n_1 + in_2 = 2\bar{a}\,b, \qquad n_3 = |a|^2 - |b|^2. \tag{T3.46}$$

Check that it lands on the sphere:

$$n_1^2 + n_2^2 + n_3^2 \overset{(1)}{=} 4|a|^2|b|^2 + \left(|a|^2 - |b|^2\right)^2 \overset{(2)}{=} \left(|a|^2 + |b|^2\right)^2 \overset{(3)}{=} 1.$$

(1) $|n_1 + in_2|^2 = |2\bar{a}b|^2$.
(2) $(p-q)^2 + 4pq = (p+q)^2$ with $p = |a|^2$, $q = |b|^2$.
(3) $\det U = 1$.

The map (T3.46) depends only on the first column $(a, b) \in S^3 \subset \mathbb{C}^2$. It is the **Hopf map** $S^3 \to S^2$. Its fibres are the cosets $U\,\mathrm{U}(1)$: multiplying on the right by $\mathrm{diag}(e^{i\chi}, e^{-i\chi})$ sends $(a, b) \mapsto e^{i\chi}(a, b)$, which leaves $\bar{a}b$ and $|a|^2 - |b|^2$ unchanged. So above each point of S^2 sits a circle in S^3.

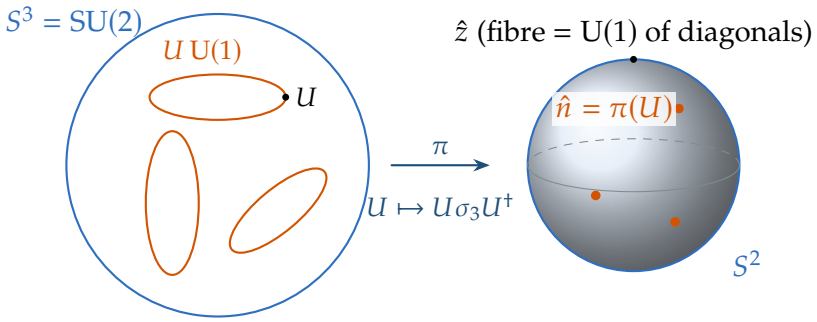


Figure T3.20: The Hopf fibration $S^3 \to S^2$ as the quotient $\mathrm{SU}(2) \to \mathrm{SU}(2)/\mathrm{U}(1)$. Each orange circle in S^3 is a coset $U\,\mathrm{U}(1) = \{Ue^{i\chi\sigma_3}\}$ and is collapsed to a single point of S^2. Any two fibres are linked circles in S^3 (the drawing only suggests this). The fibre over the north pole is the subgroup of diagonal matrices itself.

Worked example T3.30: The Bloch sphere

A spin-$\frac{1}{2}$ state is a unit vector $\psi = (a, b)^T \in \mathbb{C}^2$, a point of S^3. Physically, ψ and $e^{i\chi}\psi$ are the same state. The expectation value of spin is

$$\langle\sigma_1\rangle + i\langle\sigma_2\rangle \overset{(1)}{=} \psi^\dagger(\sigma_1 + i\sigma_2)\psi \overset{(2)}{=} 2\bar{a}\,b,$$
$$\langle\sigma_3\rangle \overset{(3)}{=} |a|^2 - |b|^2,$$

(1) Linearity of the expectation value.
(2) $\sigma_1 + i\sigma_2 = \left(\begin{smallmatrix} 0 & 0 \\ 2 & 0 \end{smallmatrix}\right)$, so $\psi^\dagger(\sigma_1 + i\sigma_2)\psi = \bar{a}\cdot 2b$.
(3) $\sigma_3 = \mathrm{diag}(1, -1)$.

This is (T3.46): the Hopf map *is* the map from a normalised spinor to its spin direction, and the fibre is the unobservable overall phase. The Bloch sphere of physical states is $S^3/\mathrm{U}(1) = S^2 = \mathbb{CP}^1$. The state with $\hat{n} = \hat{z}$ is spin-up $(1, 0)$; spin-down $(0, 1)$ goes to $\hat{n} = -\hat{z}$; and $(1, 1)/\sqrt{2}$ goes to $\hat{x}$.

Physical picture: Why the Hopf fibration is the monopole

A section of the Hopf map would be a smooth choice of phase for the spinor $\psi(\hat{n})$ over the whole sphere. There is none: the phase must wind once around the equator when we compare the choices made near the north and south poles. This winding is the Dirac monopole of charge one, and the natural connection on the Hopf fibration (Example T3.37) is the monopole vector potential. Berry's phase for a spin-$\frac{1}{2}$ in a slowly rotating magnetic field is the holonomy of the same connection (Section T3.9).

T3.7.6 Complex projective space $\mathbb{CP}^n = \mathrm{SU}(n+1)/\mathrm{U}(n)$

$\mathbb{CP}^n$ is the set of complex lines through the origin of $\mathbb{C}^{n+1}$, or equivalently of nonzero vectors z modulo $z \sim \lambda z$, $\lambda \in \mathbb{C}^\times$. Normalising $|z| = 1$ leaves the phase freedom $z \sim e^{i\alpha} z$, so

$$\mathbb{CP}^n = S^{2n+1}/\mathrm{U}(1), \qquad \dim_{\mathbb{R}} \mathbb{CP}^n = 2n.$$

In quantum mechanics, $\mathbb{CP}^n$ is the space of pure states of an $(n+1)$-level system. $\mathrm{SU}(n+1)$ acts on it by $[z] \mapsto [Uz]$, and the action is transitive because it is already transitive on S^{2n+1}.

Worked example T3.31: The stabiliser of a line

Take the line through $e_0 = (1, 0, \ldots, 0)$. A matrix $U \in \mathrm{U}(n+1)$ fixes this line if $Ue_0 = e^{i\alpha} e_0$. Then its first column is $e^{i\alpha} e_0$, and unitarity (orthonormal rows) forces its first row to be $(e^{i\alpha}, 0, \ldots, 0)$:

$$U = \begin{pmatrix} e^{i\alpha} & 0 \\ 0 & V \end{pmatrix}, \qquad V \in \mathrm{U}(n).$$

The stabiliser in $\mathrm{U}(n+1)$ is $\mathrm{U}(1) \times \mathrm{U}(n)$. Inside $\mathrm{SU}(n+1)$ we need in addition $\det U = e^{i\alpha} \det V = 1$, which fixes $e^{i\alpha} = (\det V)^{-1}$. So the stabiliser in $\mathrm{SU}(n+1)$ is $\{\mathrm{diag}((\det V)^{-1}, V) : V \in \mathrm{U}(n)\} \cong \mathrm{U}(n)$, and

$$\mathbb{CP}^n = \mathrm{SU}(n+1)/\mathrm{U}(n) = \mathrm{U}(n+1)/\big(\mathrm{U}(1) \times \mathrm{U}(n)\big),$$

$$\dim \mathbb{CP}^n \overset{(1)}{=} \big((n+1)^2 - 1\big) - n^2 = 2n.$$

(1) $\dim \mathrm{SU}(n+1) - \dim \mathrm{U}(n)$, from Table T3.1.

For $n = 1$: $\mathbb{CP}^1 = \mathrm{SU}(2)/\mathrm{U}(1) = S^2$, the Hopf case again. For $n = 2$: $\mathbb{CP}^2 = \mathrm{SU}(3)/\mathrm{U}(2)$ is four-dimensional; it is a favourite compact 4-manifold for instantons (Chapter 8) and algebraic geometry (Chapter 5).

T3.7.7 Flag manifolds G/T

Take the stabiliser to be as small as a commutative subgroup can reasonably be: a maximal torus T, the group $e^{\mathfrak{t}}$ generated by a Cartan subalgebra (Section T3.5.3). For $\mathrm{SU}(N)$, T is the diagonal matrices $\mathrm{diag}(e^{i\phi_1}, \ldots, e^{i\phi_N})$ with $\sum \phi_k = 0$, of dimension $N - 1$. The homogeneous space G/T is the orbit of a *generic* diagonal Hermitian matrix $D = \mathrm{diag}(d_1, \ldots, d_N)$ (all d_k different) under conjugation:

$$\mathrm{SU}(N)/T \cong \{UDU^\dagger\}, \qquad \dim = (N^2 - 1) - (N - 1) = N(N-1).$$

Such a matrix is the same thing as an ordered set of N orthogonal eigenlines in $\mathbb{C}^N$, a "flag". For SU(2), $G/T = S^2$ (again); for SU(3), G/T is six-dimensional. The dimension $N(N-1)$ is the number of roots (Section T3.5): the tangent space of G/T at the identity coset is $\mathfrak{g}/\mathfrak{t}$, which is spanned by the root vectors. This is the space on which Nash's Borel–Weil construction lives.

T3.7.8 Free actions and a first look at bundles

When an action is *free*, every orbit is a copy of G itself ($H = \{\mathbb{1}\}$). The space X is then sliced into copies of G, one for each point of the quotient X/G:

$$G \longrightarrow X \longrightarrow X/G.$$

This is the shape of a **principal G-bundle**: a total space X, a base X/G, and fibres that are copies of the group. Examples:

- U(1) acts freely on $S^3 \subset \mathbb{C}^2$ by phases, with quotient S^2: the Hopf bundle $\mathrm{U}(1) \to S^3 \to S^2$ of Section T3.7.5.
- U(1) acts freely on S^{2n+1} with quotient $\mathbb{CP}^n$.
- Any subgroup H acts freely on G by right multiplication, with quotient G/H: $H \to G \to G/H$. For example $\mathrm{SO}(n-1) \to \mathrm{SO}(n) \to S^{n-1}$ and $T \to G \to G/T$ ("G forms a principal T-bundle over the homogeneous space G/T", [Nash p. 189]).
- $\{\pm\mathbb{1}\}$ acts freely on SU(2) with quotient SO(3): a principal $\mathbb{Z}_2$-bundle, which is exactly the double cover.

When the action is not free, the quotient is singular at points with larger stabilisers: the orbit space of SO(3) on $\mathbb{R}^3$ is the half-line $r \geq 0$, with an "edge" at $r = 0$ where the stabiliser jumps from SO(2) to SO(3).

Worked example T3.32: Vacuum manifolds of Higgs fields

A Higgs potential $V(\phi) = \lambda(|\phi|^2 - v^2)^2$ is minimised on an orbit of the gauge group, the **vacuum manifold** $\mathcal{V} = G/H$, where H is the unbroken subgroup (the stabiliser of one vacuum).

- SU(2) *doublet* (the electroweak Higgs without hypercharge): $\phi \in \mathbb{C}^2$ with $|\phi| = v$. SU(2) acts freely and transitively on this S^3, so $H = \{\mathbb{1}\}$, $\mathcal{V} = S^3$: all three generators are broken, giving three massive vector bosons.
- SU(2) *adjoint* (the Georgi–Glashow model): $\phi = \phi^a \sigma_a$ with $|\vec{\phi}| = v$, acted on by conjugation. By Section T3.7.5, $\mathcal{V} = S^2 = \mathrm{SU}(2)/\mathrm{U}(1)$: one generator survives (the photon) and two are broken ($W^\pm$).

Counting: the number of broken generators is $\dim G - \dim H = \dim \mathcal{V}$, the number of Goldstone directions. Topology of $\mathcal{V}$ then classifies defects: a field configuration that tends to the vacuum at spatial infinity S^2_∞ defines a map $S^2_\infty \to \mathcal{V}$, and for $\mathcal{V} = S^2$ these are classified by $\pi_2(S^2) = \mathbb{Z}$: the charge of the 't Hooft–Polyakov monopole.

T3.7.9 The gauge group acting on connections

The most important action in Nash is infinite-dimensional: the gauge group $\mathcal{G}$ acting on the space $\mathcal{A}$ of all connections. Two gauge fields on the same orbit, $A' = g^{-1}Ag + g^{-1}\mathrm{d}g$, describe the same physics, so the physical configuration space is the orbit space $\mathcal{A}/\mathcal{G}$. Three facts from this section carry over directly.

- *Stabilisers.* The stabiliser of a connection A is the set of gauge transformations with $g^{-1}Ag + g^{-1}\mathrm{d}g = A$, i.e. with $\mathrm{d}g + [A, g] = 0$: covariantly constant g's. Constant elements

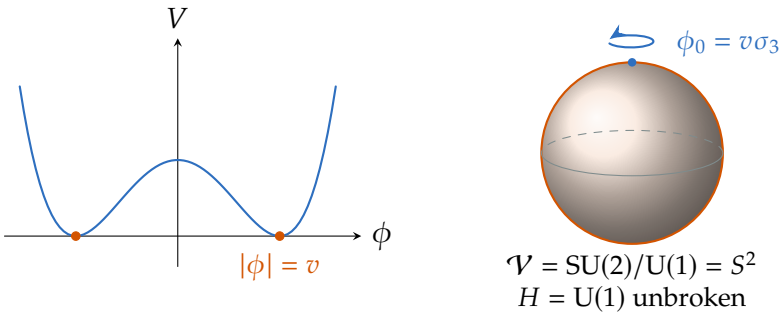


Figure T3.21: Left: a slice of the Higgs potential; its minima form one orbit of the gauge group. Right: for an adjoint SU(2) Higgs the orbit is a 2-sphere, $\mathcal{V} = G/H$. The vacuum $v\sigma_3$ is fixed by the diagonal U(1) (the electromagnetic gauge group), and moving along $\mathcal{V}$ costs no energy: these are the two Goldstone directions, eaten by $W^{\pm}$.

of the centre $Z(G)$ always qualify. A connection with a bigger stabiliser is called *reducible*; its holonomy (Section T3.9) is smaller than G. The orbit space $\mathcal{A}/\mathcal{G}$ is singular at reducible connections, exactly as the orbit space $r \geq 0$ of SO(3) on $\mathbb{R}^3$ is singular at $r = 0$.

- *Freeness.* If we use the *based* gauge group (gauge transformations equal to $\mathbb{1}$ at a chosen point x_0), the action is free, so $\mathcal{G} \to \mathcal{A} \to \mathcal{A}/\mathcal{G}$ is a principal bundle.

- *Dimension counting.* Orbit dimensions subtract stabiliser dimensions; this is how the dimension of the instanton moduli space is computed in Chapter 8.

Reading Nash (actions in Nash)

Nash proves that the gauge group $\mathcal{G}$ acts freely on $\mathcal{A}$ [Nash pp. 216–218, (8.7)–(8.11)] by showing that gauge transformations commute with parallel transport, $\alpha_b \circ PT_{\gamma_{ab}}(A) = PT_{\gamma_{ab}}(\alpha \cdot A) \circ \alpha_a$: a gauge transformation fixing A and equal to $\mathbb{1}$ at one point is then $\mathbb{1}$ everywhere. The quotient bundle $\mathcal{G} \to \mathcal{A} \to \mathcal{A}/\mathcal{G}$ [Nash (8.11)] is the "free action gives a principal bundle" of this section. The stabiliser of a connection under constant gauge transformations is the centraliser of its holonomy group [Nash p. 232, (8.88)–(8.90)], which is why he cuts down to irreducible connections and to $\mathrm{Ad}\, G = G/Z(G)$. The homogeneous space G/T and the bundle $T \to G \to G/T$ are the setting of the Borel–Weil theorem [Nash pp. 188–190].

Recap

- An action is a rule $g \cdot x$ with $\mathbb{1} \cdot x = x$ and $g \cdot (h \cdot x) = (gh) \cdot x$. Orbit $G \cdot x_0$; stabiliser $H = G_{x_0}$.
- Orbit–stabiliser: $G \cdot x_0 \cong G/H$, dim = dim G − dim H; homogeneous space = single orbit = G/H.
- $S^{n-1} = \mathrm{SO}(n)/\mathrm{SO}(n-1)$, $S^{2n-1} = \mathrm{SU}(n)/\mathrm{SU}(n-1)$, $S^2 = \mathrm{SU}(2)/\mathrm{U}(1) = \mathbb{CP}^1$, $\mathbb{CP}^n = \mathrm{SU}(n+1)/\mathrm{U}(n)$, flags G/T of dimension #roots.
- Hopf map $(a, b) \mapsto (2\bar{a}b, |a|^2 - |b|^2)$: spinor ↦ spin direction; fibre = phase. Free actions give principal bundles $G \to X \to X/G$.
- Vacuum manifolds G/H, broken generators = dim G/H; $\mathcal{A}/\mathcal{G}$ is singular at reducible connections.

T3.8 Gauge fields: from local symmetry to curvature

Everything so far has been about groups on their own. Now we let a group act on *fields*, and we let the group element depend on the point of spacetime. The question of this section is: *if a theory is invariant under $\psi(x) \mapsto g^{-1}(x)\psi(x)$ with $g(x)$ an arbitrary G-valued function, what must its derivatives look like?* The answer is forced on us. Ordinary derivatives fail, and the repair is a new field A, a Lie-algebra-valued 1-form: the gauge field. Its transformation law $A \mapsto g^{-1}Ag + g^{-1}\mathrm{d}g$, its curvature $F = \mathrm{d}A + A \wedge A$, the Bianchi identity and the Yang–Mills equations all follow from this one demand, and we derive each of them line by line. Geometrically, A is a *connection*, the same object as the connection 1-forms of Dray's Chapter 17 (reviewed in Chapter T2), now with an internal symmetry group in place of rotations of tangent vectors.

T3.8.1 Global versus local symmetry

Take a field $\psi(x)$ with n complex components, a column vector in $\mathbb{C}^n$ at every point x of spacetime M (coordinates x^μ, $\mu = 1, \dots, \dim M$). A group $G \subset \mathrm{U}(n)$ acts on $\mathbb{C}^n$ by matrix multiplication: SU(2) on an isospin doublet $(u, d)^T$, SU(3) on a colour triplet, U(1) on a single charged field by a phase. A typical kinetic term,

$$\mathcal{L} = \partial_\mu \psi^\dagger \, \partial_\mu \psi \qquad \text{or} \qquad \mathcal{L} = \bar{\psi} \gamma^\mu \partial_\mu \psi,$$

is invariant under $\psi \mapsto g^{-1}\psi$ for a *constant* unitary g, because $\partial_\mu(g^{-1}\psi) = g^{-1}\partial_\mu\psi$ and $(g^{-1})^\dagger g^{-1} = \mathbb{1}$. This is a **global symmetry**.

Why g^{-1}? We write the transformation as $\psi \mapsto g^{-1}\psi$ rather than $\psi \mapsto g\psi$ to match Nash and the mathematicians. The reason is the frame picture of Section T3.8.4: if we rotate the *basis* of $\mathbb{C}^n$ by g, the *components* of a fixed vector change by g^{-1}, just as rotating your axes by $+30^\circ$ rotates the coordinates of a fixed arrow by -30°. Physicists usually write $\psi \mapsto V\psi$; then simply $V = g^{-1}$ (see the dictionary in Section T3.8.9).

Now let g depend on x. The mass term $\psi^\dagger\psi$ is still invariant, but the derivative is not:

$$\begin{aligned} \mathrm{d}\big(g^{-1}\psi\big) &\overset{(1)}{=} \mathrm{d}(g^{-1})\,\psi + g^{-1}\mathrm{d}\psi \\ &\overset{(2)}{=} -g^{-1}\mathrm{d}g\, g^{-1}\psi + g^{-1}\mathrm{d}\psi \\ &\overset{(3)}{=} g^{-1}\mathrm{d}\psi - \theta\,\big(g^{-1}\psi\big), \qquad \theta := g^{-1}\mathrm{d}g. \end{aligned} \tag{T3.47}$$

(1) Product rule; d acts on each matrix entry, and the order of the matrix factors is kept.
(2) The derivative of an inverse, (T3.35).
(3) Name the Maurer–Cartan form $\theta = g^{-1}\mathrm{d}g$ of the map $x \mapsto g(x)$ (Section T3.6) and regroup.

If $\mathrm{d}\psi$ transformed like ψ we would have only the first term. The unwanted second term is the Maurer–Cartan form of the gauge transformation. It vanishes when g is constant and only then.

Physical picture: A derivative compares two different vector spaces

$\partial_\mu\psi = \lim_{\varepsilon\to 0} \frac{1}{\varepsilon}\big[\psi(x + \varepsilon\hat{\mu}) - \psi(x)\big]$ subtracts a vector in the copy of $\mathbb{C}^n$ at $x + \varepsilon\hat{\mu}$ from a vector in the copy at x. With a local symmetry, the basis of each copy can be rotated independently, so "the same components" at two different points means nothing. It is like subtracting the velocity of a ship measured in one harbour's coordinates from one

measured in another harbour's coordinates, without being told how the two sets of axes are related. We need an extra piece of data that tells us how to carry a vector from one copy to the next. That piece of data is the gauge field (Figure T3.22).

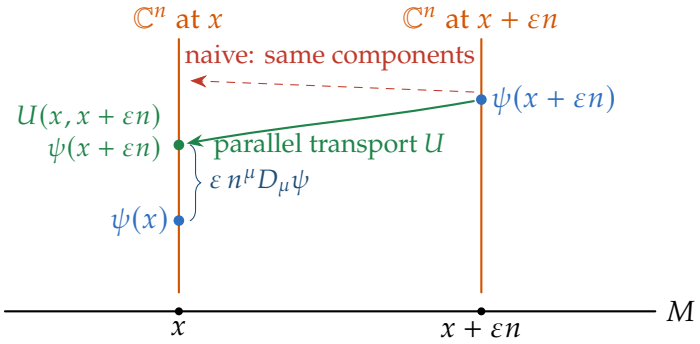


Figure T3.22: The covariant derivative. Above each point of spacetime M sits a copy of $\mathbb{C}^n$ (orange). Comparing $\psi(x + \varepsilon n)$ with $\psi(x)$ by "keeping the components fixed" (red dashed) depends on arbitrary choices of basis at the two points. Instead we carry $\psi(x+\varepsilon n)$ back to x with the comparator $U(x, x + \varepsilon n)$ (green) and subtract there. The difference, divided by ε, is $n^\mu D_\mu \psi$.

T3.8.2 The comparator and the covariant derivative

Following the QFT notes [QFT notes §19.2–19.3], introduce a matrix $U(y, x)$ that carries a vector in the copy of $\mathbb{C}^n$ at x to the copy at y. We want $U(y, x)\psi(x)$ to transform like a field at y. Since $\psi(x) \mapsto g(x)^{-1}\psi(x)$ and $\psi(y) \mapsto g(y)^{-1}\psi(y)$, this requires

$$U(y, x) \;\longmapsto\; g(y)^{-1}\, U(y, x)\, g(x). \tag{T3.48}$$

Check: $U(y, x)\psi(x) \mapsto g(y)^{-1}U(y, x)g(x)\, g(x)^{-1}\psi(x) = g(y)^{-1}\big(U(y, x)\psi(x)\big)$.

Working definition T3.16 (comparator and connection components)

A **comparator** (or parallel transporter) assigns to each pair of nearby points a matrix $U(y, x) \in G$ with $U(x, x) = \mathbb{1}$ and the transformation law (T3.48). Its first-order Taylor coefficients define the **connection components** $A_\mu(x)$:

$$U(x, x + \varepsilon n) = \mathbb{1} + \varepsilon\, n^\mu A_\mu(x) + O(\varepsilon^2). \tag{T3.49}$$

How to read it: A_μ is "the rotation of internal space [the copy of $\mathbb{C}^n$ at x in which $\psi(x)$ lives] you must apply per unit length in the μ direction to bring a vector back to x".

Now define the derivative by carrying the neighbouring value back before subtracting (Figure T3.22):

$$\begin{aligned} n^\mu D_\mu \psi(x) &\overset{(1)}{=} \lim_{\varepsilon \to 0} \frac{1}{\varepsilon}\Big[U(x, x + \varepsilon n)\, \psi(x + \varepsilon n) - \psi(x)\Big] \\ &\overset{(2)}{=} \lim_{\varepsilon \to 0} \frac{1}{\varepsilon}\Big[\big(\mathbb{1} + \varepsilon n^\mu A_\mu\big)\big(\psi + \varepsilon n^\mu \partial_\mu \psi\big) - \psi\Big] \\ &\overset{(3)}{=} n^\mu \big(\partial_\mu + A_\mu\big)\psi. \end{aligned} \tag{T3.50}$$

(1) Definition: transport the value at $x + \varepsilon n$ back to x, then subtract. Both terms now live in the copy of $\mathbb{C}^n$ at x.

(2) Expand U by (T3.49) and ψ by Taylor's theorem, everything evaluated at x.
(3) The zeroth-order terms $\psi - \psi$ cancel; the $O(\varepsilon)$ terms are $\varepsilon n^\mu(\partial_\mu\psi + A_\mu\psi)$; the $O(\varepsilon^2)$ terms vanish in the limit.

So the **covariant derivative** is $D_\mu = \partial_\mu + A_\mu$, or, as a 1-form,

$$D\psi = \mathrm{d}\psi + A\psi, \qquad A = A_\mu\,\mathrm{d}x^\mu. \tag{T3.51}$$

What kind of matrix is A_μ? U lies in G and $U(x,x) = \mathbb{1}$, so $\varepsilon \mapsto U(x, x+\varepsilon n)$ is a curve in G through the identity, and its velocity $n^\mu A_\mu$ lies in the Lie algebra $\mathfrak{g} = T_{\mathbb{1}}G$ (Section T3.2). For $G = \mathrm{U}(n)$: differentiating $U^\dagger U = \mathbb{1}$ gives $A_\mu^\dagger + A_\mu = 0$, anti-Hermitian. For $G = \mathrm{SU}(n)$: differentiating $\det U = 1$ gives $\operatorname{tr} A_\mu = 0$ as well. For $G = \mathrm{U}(1)$: $A_\mu = ia_\mu$ with a_μ real.

Working definition T3.17 (gauge field = connection)

A **gauge field** (or **connection**) with gauge group G on a region of M is a $\mathfrak{g}$-valued 1-form [at each x, one generator $A_\mu(x) \in \mathfrak{g}$ for each direction μ]

$$A = A_\mu(x)\,\mathrm{d}x^\mu = A_\mu^a(x)\,T_a\,\mathrm{d}x^\mu,$$

where T_a is a basis of $\mathfrak{g}$ (for SU(2), $T_a = -\frac{i}{2}\sigma_a$). *How to count it:* $\dim G \times \dim M$ real functions $A_\mu^a(x)$; for SU(3) in four dimensions, $8 \times 4 = 32$ fields (eight gluons, four components each). *How to use it:* $D\psi = \mathrm{d}\psi + A\psi$, where A acts on ψ in whatever representation ψ carries, i.e. T_a is replaced by its representation matrix.

T3.8.3 The transformation law $A \mapsto g^{-1}Ag + g^{-1}\mathrm{d}g$

Derivation 1: from the comparator. Put the expansion (T3.49) into the law (T3.48) with $y = x$ and the second point $x + \varepsilon n$:

$$\begin{aligned}
\mathbb{1} + \varepsilon n^\mu A'_\mu &\overset{(1)}{=} g(x)^{-1}\big(\mathbb{1} + \varepsilon n^\mu A_\mu\big)\, g(x+\varepsilon n) \\
&\overset{(2)}{=} g^{-1}\big(\mathbb{1} + \varepsilon n^\mu A_\mu\big)\big(g + \varepsilon n^\mu \partial_\mu g\big) + O(\varepsilon^2) \\
&\overset{(3)}{=} \mathbb{1} + \varepsilon n^\mu\big(g^{-1}A_\mu g + g^{-1}\partial_\mu g\big) + O(\varepsilon^2).
\end{aligned}$$

(1) The transformed comparator is $g(x)^{-1}U(x, x+\varepsilon n)g(x+\varepsilon n)$ by (T3.48), with the roles $y \to x$, $x \to x + \varepsilon n$.
(2) Taylor-expand $g(x+\varepsilon n)$; now everything is evaluated at x.
(3) Multiply out; $g^{-1}g = \mathbb{1}$, and keep the terms of order ε.

Comparing the coefficients of εn^μ:

$$\boxed{A \longmapsto A' = g^{-1}Ag + g^{-1}\mathrm{d}g} \tag{T3.52}$$

The first term rotates A like an element of the Lie algebra (it is $\mathrm{Ad}_{g^{-1}}A$, Section T3.4). The second, inhomogeneous term is the Maurer–Cartan form of g.

Derivation 2: demand that $D\psi$ transforms like ψ. We want $D'\psi' = g^{-1}D\psi$, where $\psi' = g^{-1}\psi$ and $D' = \mathrm{d} + A'$. Using (T3.47),

$$\begin{aligned} D'\psi' &\overset{(1)}{=} \mathrm{d}(g^{-1}\psi) + A'g^{-1}\psi \overset{(2)}{=} g^{-1}\mathrm{d}\psi - g^{-1}\mathrm{d}g\, g^{-1}\psi + A'g^{-1}\psi \\ &\overset{(3)}{=} g^{-1}\mathrm{d}\psi - g^{-1}\mathrm{d}g\, g^{-1}\psi + \big(g^{-1}Ag + g^{-1}\mathrm{d}g\big)g^{-1}\psi \\ &\overset{(4)}{=} g^{-1}\big(\mathrm{d}\psi + A\psi\big) = g^{-1}D\psi. \end{aligned}$$

(1) Definition of D' acting on ψ'.
(2) Equation (T3.47).
(3) Insert the proposed law (T3.52).
(4) $g\,g^{-1} = \mathbb{1}$ in both brackets: the two $g^{-1}\mathrm{d}g\,g^{-1}\psi$ terms cancel, and $g^{-1}Agg^{-1}\psi = g^{-1}A\psi$.

So (T3.52) is exactly what is needed, and it is the *only* choice: if A' had any other value, the difference $(A' - g^{-1}Ag - g^{-1}\mathrm{d}g)g^{-1}\psi$ would spoil the result for some ψ. With $D\psi$ covariant, the Lagrangian $(D_\mu\psi)^\dagger D_\mu\psi$ is gauge invariant, because $(g^{-1}D\psi)^\dagger(g^{-1}D\psi) = (D\psi)^\dagger D\psi$ for unitary g.

Key idea: Gauge principle

Promoting a global symmetry $\psi \mapsto g^{-1}\psi$ to a local one forces the replacement $\mathrm{d} \to D = \mathrm{d} + A$, with A a $\mathfrak{g}$-valued 1-form transforming as $A \mapsto g^{-1}Ag + g^{-1}\mathrm{d}g$. Then $D\psi \mapsto g^{-1}D\psi$, and any globally invariant Lagrangian built with D in place of d is locally invariant.

Try it: The law is a right action

Show that doing g and then h is the same as doing gh: $(A^g)^h = A^{gh}$, where $A^g := g^{-1}Ag + g^{-1}\mathrm{d}g$.
Answer: $(A^g)^h = h^{-1}(g^{-1}Ag + g^{-1}\mathrm{d}g)h + h^{-1}\mathrm{d}h = (gh)^{-1}A(gh) + (gh)^{-1}(\mathrm{d}g\,h + g\,\mathrm{d}h) = (gh)^{-1}A(gh) + (gh)^{-1}\mathrm{d}(gh)$. This is the "right action" remarked on in Section T3.7.

T3.8.4 The frame picture: a connection is a rotating basis

Dray's Chapter 17 [Dray §17.1–17.5] gives a third derivation, and the most geometric one. In the plane with polar coordinates, the basis vectors $\hat{r}, \hat{\phi}$ change from point to point, and Dray records how they change with *connection 1-forms*: $\mathrm{d}\hat{e}_j = \omega^i{}_j\,\hat{e}_i$. For polar coordinates he finds $\mathrm{d}\hat{r} = \mathrm{d}\phi\,\hat{\phi}$ and $\mathrm{d}\hat{\phi} = -\mathrm{d}\phi\,\hat{r}$, i.e. $\omega^r{}_\phi = -\mathrm{d}\phi$ [Dray (17.2)–(17.3), (17.38)].

Do the same in the internal space. Choose at every point x a basis ("frame") $e_1(x), \dots, e_n(x)$ of the copy of $\mathbb{C}^n$, and write a field as $\Psi = e_i\,\psi^i$, where $\psi = (\psi^1, \dots, \psi^n)^T$ are its components. Record how the frame changes by a matrix of 1-forms A:

$$\mathrm{d}e_j = e_i\,A^i{}_j.$$

Then the derivative of the (frame-independent) field Ψ is

$$\mathrm{d}\Psi \overset{(1)}{=} \mathrm{d}e_j\,\psi^j + e_i\,\mathrm{d}\psi^i \overset{(2)}{=} e_i\big(\mathrm{d}\psi^i + A^i{}_j\psi^j\big) = e_i\,(D\psi)^i,$$

(1) Product rule (Dray's (17.15)).
(2) Substitute $\mathrm{d}e_j = e_iA^i{}_j$ and rename the dummy index.

so the components of the derivative are exactly $D\psi = \mathrm{d}\psi + A\psi$. Now change frame, $e'_j = e_i\,g^i{}_j$, with $g(x)$ an invertible matrix (unitary if both frames are orthonormal). A fixed field has

$\Psi = e'\psi' = e\, g\, \psi'$, so $\psi' = g^{-1}\psi$: *components transform by g^{-1} when the frame transforms by g*. The new connection matrix is read off from

$$\mathrm{d}e'_j \overset{(1)}{=} \mathrm{d}e_i\, g^i{}_j + e_i\, \mathrm{d}g^i{}_j \overset{(2)}{=} e_k \big(A^k{}_i\, g^i{}_j + \mathrm{d}g^k{}_j\big) \overset{(3)}{=} e'_l\, \big(g^{-1}\big)^l{}_k \big(Ag + \mathrm{d}g\big)^k{}_j,$$

(1) Product rule applied to $e'_j = e_i g^i{}_j$.

(2) $\mathrm{d}e_i = e_k A^k{}_i$; rename $i \to k$ in the second term.

(3) Express the old frame in terms of the new one, $e_k = e'_l (g^{-1})^l{}_k$.

so $A' = g^{-1}Ag + g^{-1}\mathrm{d}g$ once more. A gauge transformation is nothing but a change of frame in the internal space, and the gauge field records how the chosen frames twist from point to point.

Worked example T3.33: Polar coordinates as a pure-gauge connection

In the plane, the Cartesian frame $(\hat{x}, \hat{y})$ is constant, so its connection is $A = 0$. The polar frame is the Cartesian frame rotated by the angle ϕ (Figure T3.23):

$$(\hat{r}, \hat{\phi}) = (\hat{x}, \hat{y})\, g, \qquad g = \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix} \in \mathrm{SO}(2).$$

(Column j of g holds the Cartesian components of the j-th new basis vector: $\hat{r} = \cos\phi\, \hat{x} + \sin\phi\, \hat{y}$ and $\hat{\phi} = -\sin\phi\, \hat{x} + \cos\phi\, \hat{y}$.) By (T3.52) with $A = 0$, the polar connection is

$$A' = g^{-1}\mathrm{d}g \overset{(1)}{=} \begin{pmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix} \begin{pmatrix} -\sin\phi & -\cos\phi \\ \cos\phi & -\sin\phi \end{pmatrix} \mathrm{d}\phi \overset{(2)}{=} \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \mathrm{d}\phi.$$

(1) $g^{-1} = g^T$ for a rotation; differentiate each entry of g.

(2) Multiply: the $(1,1)$ entry is $-\cos\phi\sin\phi + \sin\phi\cos\phi = 0$, the $(1,2)$ entry is $-\cos^2\phi - \sin^2\phi = -1$, and so on.

So $A'^r{}_\phi = -\mathrm{d}\phi$ and $A'^\phi{}_r = +\mathrm{d}\phi$, exactly Dray's $\omega^r{}_\phi = -\mathrm{d}\phi$. The polar connection is not zero, but it is *pure gauge*: it comes from the trivial connection by a change of frame. That is why the plane is flat even though $\omega \neq 0$ in polar coordinates. Its curvature, computed below, vanishes (Section T3.10 develops this).

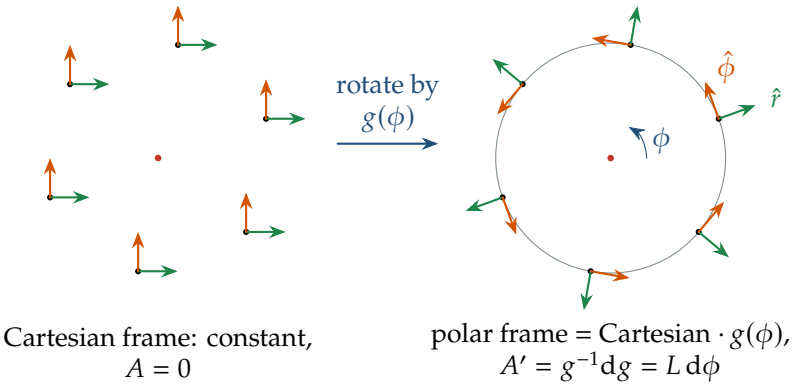


Figure T3.23: Dray's polar-coordinate example as a gauge transformation. Left: the Cartesian frame is the same at every point, so it has zero connection. Right: the polar frame $(\hat{r}, \hat{\phi})$ is the Cartesian one rotated by $g(\phi) \in \mathrm{SO}(2)$. Its connection $A' = g^{-1}\mathrm{d}g$ is nonzero, with $L = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$, but it is pure gauge. The polar frame is ill defined at the origin (red), where ϕ has no value.

T3.8.5 Infinitesimal gauge transformations

Take $g = e^{\varepsilon X}$ with $X(x) \in \mathfrak{g}$ and ε small. To first order,

$$\begin{aligned} A' &\overset{(1)}{=} (\mathbb{1} - \varepsilon X)\, A\, (\mathbb{1} + \varepsilon X) + (\mathbb{1} - \varepsilon X)\, \mathrm{d}(\mathbb{1} + \varepsilon X) + O(\varepsilon^2) \\ &\overset{(2)}{=} A + \varepsilon\big(\mathrm{d}X + AX - XA\big) + O(\varepsilon^2) = A + \varepsilon\big(\mathrm{d}X + [A, X]\big) + O(\varepsilon^2). \end{aligned} \tag{T3.53}$$

(1) Insert $g = \mathbb{1} + \varepsilon X + O(\varepsilon^2)$ and $g^{-1} = \mathbb{1} - \varepsilon X + O(\varepsilon^2)$ into (T3.52).
(2) Multiply out and keep the terms of order ε; $\mathrm{d}\mathbb{1} = 0$.

The combination $\mathrm{d}X + [A, X]$ is the **covariant derivative in the adjoint representation**, written $D_A X$ or DX. So

$$\delta A = DX = \mathrm{d}X + [A, X].$$

Why this combination? A Lie-algebra-valued field X transforms by conjugation, $X \mapsto g^{-1} X g$, because it acts on ψ like a matrix: $X\psi \mapsto g^{-1} X g\, g^{-1}\psi$. The covariant derivative of X must obey the Leibniz rule $D(X\psi) = (DX)\psi + X\, D\psi$. Solving for DX,

$$(DX)\psi \overset{(1)}{=} D(X\psi) - X\, D\psi \overset{(2)}{=} \mathrm{d}X\, \psi + X\mathrm{d}\psi + AX\psi - X\mathrm{d}\psi - XA\psi = \big(\mathrm{d}X + [A, X]\big)\psi.$$

(1) The Leibniz rule we demand.
(2) Expand $D(X\psi) = \mathrm{d}(X\psi) + AX\psi$ and $D\psi = \mathrm{d}\psi + A\psi$.

In components, with $X = X^a T_a$, $A = A^b_\mu T_b\, \mathrm{d}x^\mu$ and $[T_b, T_c] = f_{bc}{}^a T_a$:

$$\delta A^a_\mu = \partial_\mu X^a + f_{bc}{}^a\, A^b_\mu X^c. \tag{T3.54}$$

The first term is the abelian gauge transformation; the second, new for non-abelian G, rotates A in the adjoint representation (the ad matrices of (T3.21)). This is the QFT notes' [QFT notes (19.3.19)–(19.3.20)] in our conventions.

Worked example T3.34: QED as the $\mathrm{U}(1)$ case

For $G = \mathrm{U}(1)$ acting on a charged field, write $g = e^{-i\alpha(x)}$, so that $\psi \mapsto g^{-1}\psi = e^{i\alpha}\psi$, the physicists' convention of [QFT notes (19.2.1)]. A $\mathfrak{u}(1)$-valued 1-form is $A = i\, e\, a_\mu \mathrm{d}x^\mu$ with a_μ real (e is a constant, the charge, pulled out for convenience). Then

$$A' \overset{(1)}{=} A + g^{-1}\mathrm{d}g \overset{(2)}{=} A - i\, \mathrm{d}\alpha, \qquad \text{i.e.} \qquad a'_\mu = a_\mu - \tfrac{1}{e}\partial_\mu \alpha,$$
$$D_\mu \psi = (\partial_\mu + ie\, a_\mu)\psi.$$

(1) Everything commutes in U(1), so $g^{-1} A g = A$.
(2) $g^{-1}\mathrm{d}g = e^{i\alpha}\mathrm{d}(e^{-i\alpha}) = -i\, \mathrm{d}\alpha$.

These are the QED formulas [QFT notes (19.2.9), (19.2.11)]. Nothing rotates: the abelian gauge field only shifts by a gradient.

Worked example T3.35: Gauge transformations of a constant $\mathrm{SU}(2)$ potential

Take $G = \mathrm{SU}(2)$, $T_a = -\frac{i}{2}\sigma_a$, and the constant potential $A = a\, T_1\, \mathrm{d}x$ on $\mathbb{R}^2$ with coordinates (x, y).
(i) A constant gauge transformation $g = e^{\beta T_3}$. From the Pauli-matrix rotation of Section T3.4,

$e^{\beta T_3} T_1 e^{-\beta T_3} = \cos\beta\, T_1 + \sin\beta\, T_2$, and replacing $\beta \to -\beta$,

$$A' \overset{(1)}{=} g^{-1} A g \overset{(2)}{=} a\big(\cos\beta\, T_1 - \sin\beta\, T_2\big)\mathrm{d}x.$$

(1) g is constant, so $g^{-1}\mathrm{d}g = 0$.
(2) $g^{-1} T_1 g = e^{-\beta T_3} T_1 e^{\beta T_3}$, the rotation of the "colour vector" $(1,0,0)$ by $-\beta$ about the 3-axis.

A constant gauge transformation just rotates the colour components A^a as a vector in the adjoint (isospin-1) representation.
(ii) A position-dependent one, $g = e^{\beta x\, T_3}$. Now $g^{-1}\mathrm{d}g = \beta\, T_3\, \mathrm{d}x$, and

$$A' = a\big(\cos\beta x\, T_1 - \sin\beta x\, T_2\big)\mathrm{d}x + \beta\, T_3\, \mathrm{d}x.$$

The colour direction of A' now twists as we move along x, and a new component along T_3 has appeared. Both A and A' describe the same physics. Since A is constant and has only an x-component, its field strength (Section T3.8.6) is $F = \mathrm{d}A + A \wedge A = a^2 T_1^2\, \mathrm{d}x \wedge \mathrm{d}x = 0$; so A' must have $F' = 0$ too. Check it yourself with (T3.57): $F'_{xy} = 0$ trivially because nothing depends on y and there is no $\mathrm{d}y$ component.

T3.8.6 Curvature from the commutator of covariant derivatives

Ordinary derivatives commute: $\mathrm{dd} = 0$. Covariant derivatives need not. To see what replaces $\mathrm{dd} = 0$, extend D to vector-valued *forms*: for a $\mathbb{C}^n$-valued p-form ϕ (such as $D\psi$, a $\mathbb{C}^n$-valued 1-form) set $D\phi = \mathrm{d}\phi + A \wedge \phi$. Then

$$\begin{aligned} D(D\psi) &\overset{(1)}{=} \mathrm{d}\big(\mathrm{d}\psi + A\psi\big) + A \wedge \big(\mathrm{d}\psi + A\psi\big) \\ &\overset{(2)}{=} \mathrm{d}A\,\psi - A \wedge \mathrm{d}\psi + A \wedge \mathrm{d}\psi + A \wedge A\,\psi \\ &\overset{(3)}{=} \big(\mathrm{d}A + A \wedge A\big)\psi. \end{aligned} \tag{T3.55}$$

(1) Apply $D = \mathrm{d} + A\wedge$ to the 1-form $D\psi = \mathrm{d}\psi + A\psi$.
(2) $\mathrm{dd}\psi = 0$; the product rule for the 1-form A times the 0-form ψ gives $\mathrm{d}(A\psi) = \mathrm{d}A\,\psi - A \wedge \mathrm{d}\psi$ (a minus sign for moving d past a 1-form).
(3) The two $A \wedge \mathrm{d}\psi$ terms cancel.

All derivatives of ψ have dropped out: D^2 is not a differential operator at all, but multiplication by a matrix-valued 2-form.

Working definition T3.18 (curvature / field strength)

The **curvature** (field strength) of A is the $\mathfrak{g}$-valued 2-form [for each pair of directions μ, ν, a generator $F_{\mu\nu} = -F_{\nu\mu}$ in $\mathfrak{g}$]

$$F := \mathrm{d}A + A \wedge A, \qquad D^2\psi = F\psi. \tag{T3.56}$$

How to compute it in components: write $F = \frac{1}{2} F_{\mu\nu} \mathrm{d}x^\mu \wedge \mathrm{d}x^\nu$; then

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu + [A_\mu, A_\nu], \qquad F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu + f_{bc}{}^a A^b_\mu A^c_\nu. \tag{T3.57}$$

How to recognise it: $F_{\mu\nu} = [D_\mu, D_\nu]$, the failure of covariant derivatives to commute.

The component formula comes from

$$\begin{aligned} \mathrm{d}A + A \wedge A &\overset{(1)}{=} \partial_\mu A_\nu \, \mathrm{d}x^\mu \wedge \mathrm{d}x^\nu + A_\mu A_\nu \, \mathrm{d}x^\mu \wedge \mathrm{d}x^\nu \\ &\overset{(2)}{=} \tfrac{1}{2}\big(\partial_\mu A_\nu - \partial_\nu A_\mu\big)\mathrm{d}x^\mu \wedge \mathrm{d}x^\nu + \tfrac{1}{2}\big(A_\mu A_\nu - A_\nu A_\mu\big)\mathrm{d}x^\mu \wedge \mathrm{d}x^\nu, \end{aligned}$$

(1) $\mathrm{d}(A_\nu \mathrm{d}x^\nu) = \partial_\mu A_\nu \, \mathrm{d}x^\mu \wedge \mathrm{d}x^\nu$ and $A \wedge A = A_\mu \mathrm{d}x^\mu \wedge A_\nu \mathrm{d}x^\nu$ (the matrices stay in order; the scalar forms $\mathrm{d}x^\mu$ pass through them).
(2) $\mathrm{d}x^\mu \wedge \mathrm{d}x^\nu$ is antisymmetric, so only the antisymmetric part of each coefficient survives; write it with the factor $\frac{1}{2}$.

and from $[A_\mu, A_\nu] = A_\mu^b A_\nu^c [T_b, T_c] = f_{bc}{}^a A_\mu^b A_\nu^c T_a$. The commutator version: acting on ψ,

$$\begin{aligned} [D_\mu, D_\nu]\psi &\overset{(1)}{=} (\partial_\mu + A_\mu)(\partial_\nu \psi + A_\nu \psi) - (\mu \leftrightarrow \nu) \\ &\overset{(2)}{=} \partial_\mu \partial_\nu \psi + (\partial_\mu A_\nu)\psi + A_\nu \partial_\mu \psi + A_\mu \partial_\nu \psi + A_\mu A_\nu \psi - (\mu \leftrightarrow \nu) \\ &\overset{(3)}{=} \big(\partial_\mu A_\nu - \partial_\nu A_\mu + A_\mu A_\nu - A_\nu A_\mu\big)\psi = F_{\mu\nu}\psi. \end{aligned}$$

(1) Definition of the commutator of two operators.
(2) Expand, using the product rule on $\partial_\mu(A_\nu \psi)$.
(3) Subtracting the $\mu \leftrightarrow \nu$ terms cancels $\partial_\mu \partial_\nu \psi$ (partial derivatives commute) and the symmetric pair $A_\nu \partial_\mu \psi + A_\mu \partial_\nu \psi$.

For electromagnetism this is the QFT notes' $[D_\mu, D_\nu] = ieF_{\mu\nu}$ [QFT notes (19.2.26)]; for Yang–Mills it is [QFT notes (19.3.22)–(19.3.25)].

How F transforms. Since $D\psi \mapsto g^{-1}D\psi$ for every ψ, applying the rule twice gives $D'^2\psi' = g^{-1}D^2\psi$, i.e. $F' g^{-1}\psi = g^{-1}F\psi$ for every ψ, so

$$\boxed{F \longmapsto F' = g^{-1} F g} \tag{T3.58}$$

It is instructive to check this directly, because it shows the inhomogeneous terms cancelling. Write $\tilde{A} = g^{-1}Ag$ and $\theta = g^{-1}\mathrm{d}g$, so $A' = \tilde{A} + \theta$. We need two derivatives:

$$\begin{aligned} \mathrm{d}\tilde{A} &\overset{(1)}{=} \mathrm{d}(g^{-1}) \wedge Ag + g^{-1}\mathrm{d}A\, g - g^{-1}A \wedge \mathrm{d}g \\ &\overset{(2)}{=} -\theta \wedge \tilde{A} + g^{-1}\mathrm{d}A\, g - \tilde{A} \wedge \theta, \\ \mathrm{d}\theta &\overset{(3)}{=} -\theta \wedge \theta. \end{aligned} \tag{T3.59}$$

(1) Product rule on the triple product $g^{-1} \cdot A \cdot g$; the last term has a minus sign because d moves past the 1-form A.
(2) $\mathrm{d}(g^{-1}) = -g^{-1}\mathrm{d}g\, g^{-1}$ gives $-g^{-1}\mathrm{d}g\, g^{-1}Ag = -\theta \wedge \tilde{A}$; and $g^{-1}A \wedge \mathrm{d}g = (g^{-1}Ag) \wedge (g^{-1}\mathrm{d}g)$ after inserting gg^{-1}.
(3) The Maurer–Cartan structure equation (T3.36).

Then

$$\begin{aligned} F' &\overset{(1)}{=} \mathrm{d}\tilde{A} + \mathrm{d}\theta + (\tilde{A} + \theta) \wedge (\tilde{A} + \theta) \\ &\overset{(2)}{=} \big(-\theta\tilde{A} + g^{-1}\mathrm{d}A\, g - \tilde{A}\theta\big) - \theta\theta + \tilde{A}\tilde{A} + \tilde{A}\theta + \theta\tilde{A} + \theta\theta \\ &\overset{(3)}{=} g^{-1}\mathrm{d}A\, g + \tilde{A} \wedge \tilde{A} \overset{(4)}{=} g^{-1}\big(\mathrm{d}A + A \wedge A\big)g = g^{-1}Fg. \end{aligned}$$

(1) Definition of F' with $A' = \tilde{A} + \theta$.
(2) Insert (T3.59) and expand the square, keeping the order (wedges suppressed).
(3) Every term containing θ cancels against a partner.
(4) $\tilde{A} \wedge \tilde{A} = g^{-1}Ag\,g^{-1} \wedge Ag = g^{-1}A \wedge A\,g$.

Key idea: Field strength is covariant, not invariant

$F = \mathrm{d}A + A \wedge A$ transforms by conjugation, $F \mapsto g^{-1}Fg$: it lives in the adjoint representation and carries colour. For U(1) conjugation does nothing, so the electromagnetic field is gauge *invariant*. For non-abelian G only conjugation-invariant combinations are physical: $\operatorname{tr} F \wedge F$, $\operatorname{tr} F \wedge \star F$, $\operatorname{tr} F_{\mu\nu}F_{\mu\nu}$, $\det(\mathbb{1} + \frac{i}{2\pi}F)$. These are the gauge-invariant polynomials that build the Yang–Mills action and the characteristic classes (Chapter 1, [Nash p. 87, (3.125)]).

Worked example T3.36: A constant potential with nonzero field strength

On $\mathbb{R}^2$ take the SU(2) potential $A = a\,T_1\,\mathrm{d}x + b\,T_2\,\mathrm{d}y$ with constants a, b. In electromagnetism a constant potential has zero field. Here

$$F \overset{(1)}{=} A \wedge A \overset{(2)}{=} ab\,(T_1T_2\,\mathrm{d}x \wedge \mathrm{d}y + T_2T_1\,\mathrm{d}y \wedge \mathrm{d}x) \overset{(3)}{=} ab\,[T_1, T_2]\,\mathrm{d}x \wedge \mathrm{d}y \overset{(4)}{=} ab\,T_3\,\mathrm{d}x \wedge \mathrm{d}y.$$

(1) A is constant, so $\mathrm{d}A = 0$.
(2) Multiply out; the $\mathrm{d}x \wedge \mathrm{d}x$ and $\mathrm{d}y \wedge \mathrm{d}y$ terms vanish.
(3) $\mathrm{d}y \wedge \mathrm{d}x = -\mathrm{d}x \wedge \mathrm{d}y$.
(4) $[T_1, T_2] = T_3$.

So $F_{xy} = ab\,T_3 \neq 0$. It is not a gauge artefact: $\operatorname{tr}(F_{xy}F_{xy}) = a^2b^2\operatorname{tr}(T_3^2) = -\frac{1}{2}a^2b^2 \neq 0$ is invariant. A non-abelian field can be "magnetic" purely through the self-interaction $[A_\mu, A_\nu]$. Its holonomy is computed in Example T3.41.

Worked example T3.37: The Dirac monopole: one connection, two patches

Here is a gauge field that *cannot* be written with a single potential on its whole domain, the prototype of every topologically nontrivial gauge field in Nash. Recall the Hopf map of Section T3.7.5: the unit spinor $\psi = (a, b)^T$ has spin direction $\hat{n}$ with $n_1 + in_2 = 2\bar{a}b$, $n_3 = |a|^2 - |b|^2$. A *choice* of spinor for each direction $\hat{n}$, in spherical angles (ϑ, ϕ), is

$$\psi_N(\vartheta,\phi) = \begin{pmatrix} \cos\frac{\vartheta}{2} \\ e^{i\phi}\sin\frac{\vartheta}{2} \end{pmatrix}, \qquad \psi_S(\vartheta,\phi) = e^{-i\phi}\psi_N = \begin{pmatrix} e^{-i\phi}\cos\frac{\vartheta}{2} \\ \sin\frac{\vartheta}{2} \end{pmatrix}.$$

Check: $2\bar{a}b = 2\cos\frac{\vartheta}{2}\sin\frac{\vartheta}{2}e^{i\phi} = \sin\vartheta\,e^{i\phi}$ and $|a|^2 - |b|^2 = \cos\vartheta$, the point $\hat{n}(\vartheta,\phi)$. At the north pole $\vartheta = 0$ we get $\psi_N = (1, 0)$ whatever ϕ is, so ψ_N is smooth there; at the south pole $\psi_N = (0, e^{i\phi})$ depends on the direction of approach, so ψ_N is singular. For ψ_S it is the other way round. Each is a good choice of phase on one patch: $U_N = S^2 \setminus \{\text{south pole}\}$, $U_S = S^2 \setminus \{\text{north pole}\}$.
The **Berry connection** of a family of states is $A = \psi^\dagger \mathrm{d}\psi$. It is imaginary, i.e. $\mathfrak{u}(1)$-valued, because $0 = \mathrm{d}(\psi^\dagger\psi) = \mathrm{d}\psi^\dagger\,\psi + \psi^\dagger\mathrm{d}\psi = 2\operatorname{Re}(\psi^\dagger\mathrm{d}\psi)$. Changing the phase, $\psi \mapsto \psi\,h$ with $h = e^{i\chi}$, gives $A \mapsto A + h^{-1}\mathrm{d}h$, the gauge law. On the two patches:

$$\begin{aligned} A_N = \psi_N^\dagger \mathrm{d}\psi_N &\overset{(1)}{=} \cos\tfrac{\vartheta}{2}\,\mathrm{d}\big(\cos\tfrac{\vartheta}{2}\big) + \sin\tfrac{\vartheta}{2}\,e^{-i\phi}\,\mathrm{d}\big(e^{i\phi}\sin\tfrac{\vartheta}{2}\big) \\ &\overset{(2)}{=} \tfrac{1}{2}\mathrm{d}\big(\cos^2\tfrac{\vartheta}{2} + \sin^2\tfrac{\vartheta}{2}\big) + i\sin^2\tfrac{\vartheta}{2}\,\mathrm{d}\phi \overset{(3)}{=} \tfrac{i}{2}\,(1 - \cos\vartheta)\,\mathrm{d}\phi, \\ A_S = \psi_S^\dagger \mathrm{d}\psi_S &\overset{(4)}{=} A_N + e^{i\phi}\mathrm{d}e^{-i\phi} = A_N - i\,\mathrm{d}\phi \overset{(5)}{=} -\tfrac{i}{2}\,(1 + \cos\vartheta)\,\mathrm{d}\phi. \end{aligned}$$

(1) Write out $\bar{a}\,\mathrm{d}a + \bar{b}\,\mathrm{d}b$ with the components of ψ_N.
(2) $\mathrm{d}(e^{i\phi}\sin\frac{\vartheta}{2}) = e^{i\phi}(\mathrm{d}\sin\frac{\vartheta}{2} + i\sin\frac{\vartheta}{2}\mathrm{d}\phi)$; the real terms combine to half the derivative of $\cos^2 + \sin^2$.
(3) That derivative is $\mathrm{d}(1) = 0$, and $\sin^2\frac{\vartheta}{2} = \frac{1}{2}(1-\cos\vartheta)$.
(4) $\psi_S = \psi_N h$ with $h = e^{-i\phi}$: the gauge law with $h^{-1}\mathrm{d}h = -i\,\mathrm{d}\phi$.
(5) Substitute A_N: $\frac{i}{2}(1-\cos\vartheta) - i = -\frac{i}{2}(1+\cos\vartheta)$.

A_N is smooth on U_N: near the north pole $1-\cos\vartheta \approx \vartheta^2/2$ kills the bad behaviour of $\mathrm{d}\phi$. At the south pole it is singular (a Dirac string). A_S is the mirror image. Both give the *same* field strength,

$$F = \mathrm{d}A_N = \mathrm{d}A_S = \tfrac{i}{2}\sin\vartheta\,\mathrm{d}\vartheta\wedge\mathrm{d}\phi,$$

(abelian, so $A\wedge A = 0$) which is smooth on the whole sphere: it is $\frac{i}{2}$ times the area form. The total flux is

$$\int_{S^2} F = \tfrac{i}{2}\cdot 4\pi = 2\pi i, \qquad c_1 = \frac{i}{2\pi}\int_{S^2} F = -1.$$

The transition function $h = e^{-i\phi}$ on the overlap (a band around the equator) winds once around U(1) as we go once around the equator (Figure T3.24). This is not an accident. If one smooth A existed on all of S^2, Stokes' theorem would give $\int_{S^2} F = \int_{S^2}\mathrm{d}A = 0$. The nonzero flux *forces* two patches, and the winding of the transition function measures the flux:

$$\int_{S^2} F \overset{(1)}{=} \int_{\text{north cap}} \mathrm{d}A_N + \int_{\text{south cap}} \mathrm{d}A_S \overset{(2)}{=} \oint_{\text{equator}} (A_N - A_S) \overset{(3)}{=} \oint i\,\mathrm{d}\phi = 2\pi i.$$

(1) Split the sphere along the equator and use the potential that is smooth on each half.
(2) Stokes on each cap. The equator bounds the north cap with the orientation of increasing ϕ and the south cap with the opposite orientation.
(3) $A_N - A_S = i\,\mathrm{d}\phi$ from the computation above.

In physics units, a monopole of magnetic charge g_m has $A_N^{\text{phys}} = g_m(1-\cos\vartheta)\mathrm{d}\phi$, flux $4\pi g_m$, and a particle of charge e sees $A = -ie\,A^{\text{phys}}$ (Section T3.8.9). The transition function is then $h = e^{-2ieg_m\phi}$, which is single-valued on the equator only if $2eg_m \in \mathbb{Z}$. That is the **Dirac quantisation condition**. The Hopf connection above is the case $2eg_m = -1$: the magnetic field seen by the phase of a spin-$\frac{1}{2}$ state.

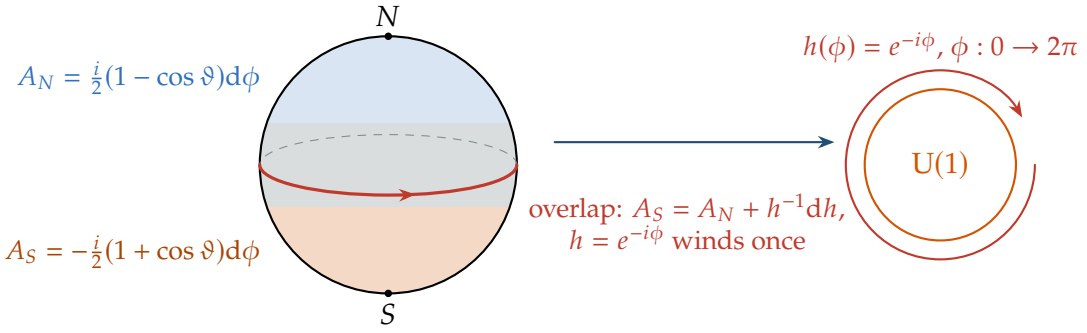


Figure T3.24: The monopole connection of Example T3.37. The sphere is covered by two patches, each missing one pole (blue north, orange south), overlapping in a band around the equator (grey). On each patch the potential is smooth; on the overlap they differ by the gauge transformation $h = e^{-i\phi}$, which goes once around U(1) as ϕ goes once around the equator. That winding number is the magnetic charge, $c_1 = -1$.

Watch out: A singular potential is not a singular field

A_N blows up at the south pole, but F does not. Singularities of a potential that disappear under a gauge transformation (Dirac strings) are artefacts of a bad choice of frame, like the singularity of the polar frame at the origin in Example T3.33. Only F and gauge-invariant quantities built from it are physical. The topological statement is that the underlying bundle is nontrivial: Chapter 1 and Chapter 8 treat the monopole as the Hopf bundle with $c_1 = \pm 1$.

T3.8.7 The Bianchi identity

The covariant derivative extends to Lie-algebra-valued p-forms ω (which transform by conjugation, like F) by

$$D\omega = \mathrm{d}\omega + A \wedge \omega - (-1)^p\, \omega \wedge A \;=:\; \mathrm{d}\omega + [A, \omega]. \tag{T3.60}$$

For $p = 0$ this is $\mathrm{d}X + [A, X]$ from above; the sign $(-1)^p$ makes the graded Leibniz rule $D(\omega\,\phi) = (D\omega)\phi + (-1)^p \omega \wedge D\phi$ hold for a $\mathbb{C}^n$-valued form ϕ. For a 2-form such as F the sign is +: $DF = \mathrm{d}F + A \wedge F - F \wedge A$.

Key idea: Bianchi identity

Every curvature satisfies $DF = \mathrm{d}F + [A, F] = 0$ identically.

Direct check:

$$\begin{aligned} \mathrm{d}F &\overset{(1)}{=} \mathrm{d}\mathrm{d}A + \mathrm{d}A \wedge A - A \wedge \mathrm{d}A \overset{(2)}{=} \mathrm{d}A \wedge A - A \wedge \mathrm{d}A, \\ A \wedge F - F \wedge A &\overset{(3)}{=} A \wedge \mathrm{d}A + A \wedge A \wedge A - \mathrm{d}A \wedge A - A \wedge A \wedge A = A \wedge \mathrm{d}A - \mathrm{d}A \wedge A, \\ DF &\overset{(4)}{=} 0. \end{aligned} \tag{T3.61}$$

(1) $\mathrm{d}(A \wedge A) = \mathrm{d}A \wedge A - A \wedge \mathrm{d}A$ (the product rule for a 1-form in front).
(2) $\mathrm{d}\mathrm{d} = 0$.
(3) Substitute $F = \mathrm{d}A + A \wedge A$ in both products; the cubic terms cancel.
(4) Add the two lines.

There is a slicker reason. Associativity says $D(D^2\psi) = D^2(D\psi)$, i.e. $D(F\psi) = F \wedge D\psi$. By the Leibniz rule the left side is $(DF)\psi + F \wedge D\psi$ (sign + for the 2-form F). So $(DF)\psi = 0$ for every ψ. In components, $DF = 0$ reads

$$D_\lambda F_{\mu\nu} + D_\mu F_{\nu\lambda} + D_\nu F_{\lambda\mu} = 0,$$

which is the Jacobi identity $[D_\lambda, [D_\mu, D_\nu]] + \text{cyclic} = 0$ for the operators D_μ.

Physical picture: Half of Maxwell's equations for free

For U(1) the commutator term vanishes and the Bianchi identity is $\mathrm{d}F = 0$. With F the electromagnetic field 2-form, $\mathrm{d}F = 0$ is $\nabla \cdot \vec{B} = 0$ together with Faraday's law $\nabla \times \vec{E} + \partial_t \vec{B} = 0$ (Chapter T1). These two equations are not dynamics: they say only that F comes from a potential. The other half, which contains the sources, comes from the action.

T3.8.8 The Yang–Mills action and field equations

We want a Lagrangian for A itself: gauge invariant, Lorentz (or rotation) invariant, and quadratic in first derivatives so that A propagates. The natural candidate uses the Hodge star

$\star$ of Chapter T1. In d Euclidean dimensions, $\alpha \wedge \star\beta = \langle \alpha, \beta \rangle$ vol for p-forms, with $\langle \cdot, \cdot \rangle$ the pointwise inner product; for 2-forms $\langle F, G \rangle = \frac{1}{2} F_{\mu\nu} G_{\mu\nu}$ (sum over all μ, ν). In four Euclidean dimensions, for example, $\star(\mathrm{d}x^1 \wedge \mathrm{d}x^2) = \mathrm{d}x^3 \wedge \mathrm{d}x^4$.

Working definition T3.19 (Yang–Mills action)

$$S_{\mathrm{YM}}[A] = -\frac{1}{g_{\mathrm{YM}}^2} \int_M \mathrm{tr}\big(F \wedge \star F\big) = -\frac{1}{2g_{\mathrm{YM}}^2} \int_M \mathrm{tr}\big(F_{\mu\nu} F_{\mu\nu}\big)\, \mathrm{d}^d x. \tag{T3.62}$$

Nash absorbs the coupling [the constant g_{YM} that sets the strength of the gluon self-interaction] ($g_{\mathrm{YM}} = 1$) and writes $S = \|F\|^2 = -\int \mathrm{tr}(F \wedge \star F)$ [Nash p. 227, (8.63)]. *Why the minus sign:* for anti-Hermitian X [$X^\dagger = -X$, as every $F_{\mu\nu}$ is for $G = \mathrm{U}(n)$ or $\mathrm{SU}(n)$], $\mathrm{tr}(X^2) = -\mathrm{tr}(X^\dagger X) = -\sum_{ij} |X_{ij}|^2 \le 0$, so the minus makes the action positive. With $\mathrm{tr}(T_a T_b) = -\frac{1}{2}\delta_{ab}$,

$$-\mathrm{tr}(F_{\mu\nu} F_{\mu\nu}) = \tfrac{1}{2} F^a_{\mu\nu} F^a_{\mu\nu}, \qquad S_{\mathrm{YM}} = \frac{1}{4g_{\mathrm{YM}}^2} \int F^a_{\mu\nu} F^a_{\mu\nu}\, \mathrm{d}^d x,$$

a sum of squares, the non-abelian version of $\frac{1}{2}(\vec{E}^2 + \vec{B}^2)$.

It is gauge invariant: $\mathrm{tr}(g^{-1} F g \wedge \star g^{-1} F g) = \mathrm{tr}(g^{-1} F \wedge \star F\, g) = \mathrm{tr}(F \wedge \star F)$, since $\star$ acts on the form part only and the trace is cyclic.

The field equations. Vary $A \to A_t = A + t\,a$ with a an arbitrary $\mathfrak{g}$-valued 1-form of compact support (Nash's (8.64)). First the curvature:

$$\begin{aligned} F(A + ta) &\overset{(1)}{=} \mathrm{d}A + t\,\mathrm{d}a + (A + ta) \wedge (A + ta) \\ &\overset{(2)}{=} F + t\big(\mathrm{d}a + A \wedge a + a \wedge A\big) + t^2 a \wedge a \overset{(3)}{=} F + t\,Da + t^2\, a \wedge a. \end{aligned} \tag{T3.63}$$

(1) Definition of F for the connection $A + ta$.
(2) Expand the product keeping the order of factors.
(3) For the 1-form a, (T3.60) with $p = 1$ gives $Da = \mathrm{d}a + A \wedge a + a \wedge A$.

We also need integration by parts under a trace. For $\mathfrak{g}$-valued forms ω (degree p) and η (degree q),

$$\mathrm{d}\ \mathrm{tr}(\omega \wedge \eta) = \mathrm{tr}(D\omega \wedge \eta) + (-1)^p\, \mathrm{tr}(\omega \wedge D\eta). \tag{T3.64}$$

Proof: the d terms are the ordinary product rule, so we must show the A terms cancel:

$$\begin{aligned} &\mathrm{tr}\big([A, \omega] \wedge \eta\big) + (-1)^p\, \mathrm{tr}\big(\omega \wedge [A, \eta]\big) \\ &\overset{(1)}{=} \mathrm{tr}(A\omega\eta) - (-1)^p\, \mathrm{tr}(\omega A \eta) + (-1)^p\, \mathrm{tr}(\omega A \eta) - (-1)^{p+q}\, \mathrm{tr}(\omega \eta A) \\ &\overset{(2)}{=} \mathrm{tr}(A\omega\eta) - (-1)^{p+q}(-1)^{p+q}\, \mathrm{tr}(A\omega\eta) = 0. \end{aligned}$$

(1) Expand $[A, \omega] = A\omega - (-1)^p \omega A$ and $[A, \eta] = A\eta - (-1)^q \eta A$ (wedges suppressed).
(2) The middle terms cancel. Graded cyclicity moves the 1-form A from the back past the $(p+q)$-form $\omega\eta$: $\mathrm{tr}(\omega\eta A) = (-1)^{p+q}\, \mathrm{tr}(A\omega\eta)$.

Now vary the action (with $g_{\mathrm{YM}} = 1$; the constant just multiplies everything):

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} S[A + ta] &\overset{(1)}{=} -\int \mathrm{tr}\big(Da \wedge \star F\big) - \int \mathrm{tr}\big(F \wedge \star Da\big) \overset{(2)}{=} -2 \int \mathrm{tr}\big(Da \wedge \star F\big) \\ &\overset{(3)}{=} -2 \int \mathrm{d}\,\mathrm{tr}\big(a \wedge \star F\big) - 2 \int \mathrm{tr}\big(a \wedge D \star F\big) \overset{(4)}{=} -2 \int \mathrm{tr}\big(a \wedge D \star F\big). \end{aligned}$$

(1) Insert (T3.63); the t^2 term does not contribute at $t = 0$.
(2) $\alpha \wedge \star\beta = \beta \wedge \star\alpha$ for two forms of the same degree (both equal $\langle\alpha, \beta\rangle \mathrm{vol}$), applied entry by entry inside the trace.
(3) Equation (T3.64) with $\omega = a$ ($p = 1$) and $\eta = \star F$: $\mathrm{tr}(Da \wedge \star F) = \mathrm{d}\,\mathrm{tr}(a \wedge \star F) + \mathrm{tr}(a \wedge D \star F)$.
(4) Stokes: a has compact support (or M is closed), so the exact term integrates to zero.

This vanishes for every a exactly when

$$\boxed{D \star F = \mathrm{d} \star F + A \wedge \star F - (-1)^{d-2} \star F \wedge A = 0} \qquad \Longleftrightarrow \qquad D_\mu F_{\mu\nu} = \partial_\mu F_{\mu\nu} + [A_\mu, F_{\mu\nu}] = 0. \tag{T3.65}$$

These are the **Yang–Mills equations**. Nash writes them as $d_A^* F = 0$ with $d_A^* = - \star d_A \star$ [Nash pp. 228–229, (8.69)–(8.75)]; the extra $\star$'s do not change the content. Together with the Bianchi identity,

$$DF = 0, \qquad D \star F = 0,$$

they are the non-abelian Maxwell equations. With matter the right-hand side of the second equation becomes the colour current, $D_\mu F^a_{\mu\nu} \propto j^a_\nu$ [QFT notes (19.3.36)].

Physical picture: Non-linearity: gluons carry colour

For U(1), $\mathrm{d} \star F = 0$ is linear in A: light passes through light. For non-abelian G, $[A_\mu, F_{\mu\nu}]$ contains A^2 and A^3 terms, and the action $\mathrm{tr}\, F^2$ contains cubic and quartic self-couplings:

$$\tfrac{1}{2} F^a_{\mu\nu} F^a_{\mu\nu} = \underbrace{(\partial_\mu A^a_\nu - \partial_\nu A^a_\mu)\partial_\mu A^a_\nu}_{\text{free}} + \underbrace{2 f_{bc}{}^a\, \partial_\mu A^a_\nu\, A^b_\mu A^c_\nu}_{\text{3-gluon vertex}} + \underbrace{\tfrac{1}{2} f_{bc}{}^a f_{de}{}^a\, A^b_\mu A^c_\nu A^d_\mu A^e_\nu}_{\text{4-gluon vertex}}.$$

(Square (T3.57); the cross term doubles because $\partial_\mu A_\nu - \partial_\nu A_\mu$ and $f A_\mu A_\nu$ are both anti-symmetric in $\mu\nu$.) The gauge field is itself charged under the group, which is why gluons interact with each other, and why Yang–Mills theory is asymptotically free.

Worked example T3.38: Self-duality solves the Yang–Mills equations

In four Euclidean dimensions $\star\star = +1$ on 2-forms, so a 2-form splits into self-dual and anti-self-dual parts, $F = F^+ + F^-$ with $\star F^\pm = \pm F^\pm$. If $F = \pm \star F$ then

$$D \star F = \pm DF = 0$$

by the Bianchi identity: a first-order equation implies the second-order Yang–Mills equation (Nash's (8.75)–(8.76)). Moreover such fields minimise the action in their topological class. Using $\star F \wedge \star\star F = \star F \wedge F$ and $\alpha \wedge \star\beta = \beta \wedge \star\alpha$,

$$\begin{aligned} 0 &\overset{(1)}{\leq} -\tfrac{1}{2} \int \mathrm{tr}\big((F \mp \star F) \wedge \star (F \mp \star F)\big) \\ &\overset{(2)}{=} -\tfrac{1}{2} \int \mathrm{tr}\big(F \wedge \star F \mp F \wedge F \mp \star F \wedge \star F + \star F \wedge F\big) \overset{(3)}{=} -\int \mathrm{tr}(F \wedge \star F) \pm \int \mathrm{tr}(F \wedge F), \end{aligned}$$

(1) $-\mathrm{tr}(X \wedge \star X) \geq 0$ pointwise for any $\mathfrak{su}(N)$-valued 2-form X (the positivity above).
(2) Expand; $\star\star F = F$.
(3) $\star F \wedge \star\star F = \star F \wedge F = F \wedge \star F$.

so $S \geq \mp \int \mathrm{tr}\, F \wedge F$ for either sign, i.e. $S \geq \left| \int \mathrm{tr}\, F \wedge F \right|$, with equality exactly when $F = \pm \star F$. We will see in Section T3.10 that $\int \mathrm{tr}\, F \wedge F = 8\pi^2 k$ with k an integer; so $S_{\mathrm{YM}} \geq 8\pi^2 |k| / g^2_{\mathrm{YM}}$, the instanton bound, and $e^{-S} \leq e^{-8\pi^2|k|/g^2_{\mathrm{YM}}}$ is the non-perturbative factor of the QFT notes [QFT notes (4.13.12)–(4.13.14)]. Instantons are the subject of Chapter 8.

T3.8.9 Mathematicians' and physicists' conventions

The QFT notes (like most physics texts) write the covariant derivative as $D_\mu = \partial_\mu - ig\,\mathcal{A}^a_\mu R_a$ with Hermitian generators R_a of the representation of ψ, $[R_a, R_b] = i f_{abc} R_c$ and $\mathrm{tr}(R_a R_b) = C(R)\delta_{ab}$ ($= \frac{1}{2}\delta_{ab}$ in the fundamental), and $\psi \mapsto V\psi$ with $V = e^{i\alpha^a R_a}$ [QFT notes (19.3.1)–(19.3.10)]. To avoid a clash with the group element g, we write their coupling as g_{YM} and their Hermitian gauge field as $\mathcal{A}$. Matching $D = \mathrm{d} + A$:

$$A_{\mathrm{math}} = -i\, g_{\mathrm{YM}}\, \mathcal{A}^a_\mu R_a\, \mathrm{d}x^\mu, \qquad V = g^{-1}. \tag{T3.66}$$

Then:

$$\begin{aligned} F_{\mathrm{math},\mu\nu} &\overset{(1)}{=} -ig_{\mathrm{YM}}\big(\partial_\mu \mathcal{A}^a_\nu - \partial_\nu \mathcal{A}^a_\mu\big)R_a + (-ig_{\mathrm{YM}})^2 \mathcal{A}^b_\mu \mathcal{A}^c_\nu [R_b, R_c] \\ &\overset{(2)}{=} -ig_{\mathrm{YM}}\Big(\partial_\mu \mathcal{A}^a_\nu - \partial_\nu \mathcal{A}^a_\mu + g_{\mathrm{YM}} f_{bca}\mathcal{A}^b_\mu \mathcal{A}^c_\nu\Big)R_a = -ig_{\mathrm{YM}}\,\mathcal{F}^a_{\mu\nu} R_a, \end{aligned}$$

(1) Insert (T3.66) into (T3.57).
(2) $(-ig)^2[R_b, R_c] = -g^2\, i f_{bca} R_a = -ig\,(g f_{bca})R_a$.

which is the QFT notes' field strength [QFT notes (19.3.25)]. The gauge law with $g = V^{-1}$ becomes $A' = VAV^{-1} - (\mathrm{d}V)V^{-1}$, i.e. $\mathcal{A}' = V\mathcal{A}V^{-1} - \frac{i}{g_{\mathrm{YM}}}(\mathrm{d}V)V^{-1}$, which is [QFT notes (19.3.14)].

Dictionary: gauge fields

Mathematics (Nash, these notes)	**Physics (QFT notes)**
$A = A_\mu \mathrm{d}x^\mu$, anti-Hermitian	$\mathcal{A}_\mu = \mathcal{A}^a_\mu R_a$, Hermitian
$A = -ig_{\mathrm{YM}}\mathcal{A}$	$\mathcal{A} = \frac{i}{g_{\mathrm{YM}}}A$
$D = \mathrm{d} + A$	$D_\mu = \partial_\mu - ig_{\mathrm{YM}}\mathcal{A}^a_\mu R_a$
$\psi \mapsto g^{-1}\psi$	$\psi \mapsto V\psi$, $V = e^{i\alpha^a R_a}$
$A \mapsto g^{-1}Ag + g^{-1}\mathrm{d}g$	$\mathcal{A} \mapsto V\mathcal{A}V^\dagger - \frac{i}{g_{\mathrm{YM}}}(\partial V)V^\dagger$
$\delta A = \mathrm{d}X + [A, X]$	$\delta\mathcal{A}^a_\mu = \frac{1}{g_{\mathrm{YM}}}\partial_\mu \alpha^a + f_{abc}\mathcal{A}^b_\mu \alpha^c$
$F = \mathrm{d}A + A \wedge A$	$\mathcal{F}^a = \partial\mathcal{A}^a - \partial\mathcal{A}^a + g_{\mathrm{YM}} f_{abc}\mathcal{A}^b\mathcal{A}^c$
$F_{\mu\nu} = [D_\mu, D_\nu]$	$[D_\mu, D_\nu] = -ig_{\mathrm{YM}}\mathcal{F}^a_{\mu\nu}R_a$
$F \mapsto g^{-1}Fg$	$\mathcal{F} \mapsto V\mathcal{F}V^\dagger$
$-\frac{1}{g^2_{\mathrm{YM}}}\int \mathrm{tr}\, F \wedge \star F$	$\frac{1}{4}\int \mathcal{F}^a_{\mu\nu}\mathcal{F}^a_{\mu\nu}$ (Euclidean)
$D \star F = 0$	$\partial_\mu \mathcal{F}^a_{\mu\nu} + g_{\mathrm{YM}} f_{abc}\mathcal{A}^b_\mu \mathcal{F}^c_{\mu\nu} = 0$
QED: $A = ie\,a$, $g = e^{-i\alpha}$	$D = \partial + iea$, $\psi \mapsto e^{i\alpha}\psi$

(For the $\delta\mathcal{A}$ line, take $X = -i\alpha^a R_a$, which is $g = V^{-1} = e^X$ to first order in the sign convention $V = e^{i\alpha R}$, so the table's α is minus the α of (T3.53); always check one line against the source you are reading.) The QED line uses the QFT notes' convention $D = \partial + iea$ [QFT notes (19.2.9)], which is the U(1) case with $R = -1$ and $g_{\mathrm{YM}} = e$.

Watch out: Minkowski signature

The dictionary above is Euclidean. In Minkowski signature the physics Lagrangian is $-\frac{1}{4}\mathcal{F}^a_{\mu\nu}\mathcal{F}^{a\,\mu\nu}$ and $\star\star = -1$ on 2-forms in four dimensions, so there are no real self-dual solutions: $F = \star F$ would give $F = \star\star F = -F$ [QFT notes §4.13.4]. This is why instantons live in Euclidean space.

Reading Nash (gauge fields in Nash)

Nash uses the mathematicians' conventions throughout: A is a $\mathfrak{g}$-valued 1-form (anti-Hermitian for unitary G), $F = \mathrm{d}A + A\wedge A$, and gauge transformations act by $A \mapsto g^{-1}(A+\mathrm{d})g$ [Nash p. 232, (8.88)]. The action appears first as $S = -\int \mathrm{tr}(F \wedge \star F)$ with self-dual solutions $F = \star F$ [Nash p. 25, (1.74)–(1.75)], and in detail in [Nash pp. 227–229, (8.63)–(8.76)]: the family $A_t = A + ta$, the expansion $F(A_t) = F(A) + t\, d_A a + t^2 a \wedge a$ (our (T3.63)), the critical-point equation $d_A^* F = 0$, the Bianchi identity $d_A F = 0$, and the split $S = \|F^+\|^2 + \|F^-\|^2$. Nash's gauge group is the group of gauge transformations $g(x)$ (sections of $\mathrm{Ad}\, P$) and its Lie algebra is the space of $X(x)$ (sections of $\mathrm{ad}\, P$) [Nash pp. 176–177, (6.19)–(6.24)]; our infinitesimal law $\delta A = DX$ is the action of that Lie algebra on $\mathcal{A}$.

Recap

- Local symmetry $\psi \mapsto g^{-1}(x)\psi$ spoils $\mathrm{d}\psi$ by the Maurer–Cartan term $-g^{-1}\mathrm{d}g$; the cure is $D = \mathrm{d} + A$.
- $A = A_\mu^a T_a \mathrm{d}x^\mu$ is a $\mathfrak{g}$-valued 1-form (the linearised comparator); $A \mapsto g^{-1}Ag + g^{-1}\mathrm{d}g$, derived three ways (comparator, covariance of $D\psi$, change of frame). Infinitesimally $\delta A = \mathrm{d}X + [A, X]$.
- $D^2 = F = \mathrm{d}A + A\wedge A$, $F_{\mu\nu} = [D_\mu, D_\nu] = \partial_\mu A_\nu - \partial_\nu A_\mu + [A_\mu, A_\nu]$, $F \mapsto g^{-1}Fg$; constant non-abelian potentials can have $F \neq 0$.
- Monopole: $A_N = \frac{i}{2}(1-\cos\vartheta)\mathrm{d}\phi$, $A_S = -\frac{i}{2}(1+\cos\vartheta)\mathrm{d}\phi$, glued by $e^{-i\phi}$; flux $2\pi i$, $c_1 = -1$; Dirac quantisation.
- Bianchi $DF = 0$; Yang–Mills $S = -\frac{1}{g^2}\int \mathrm{tr}\, F \wedge \star F$, field equation $D \star F = 0$; self-dual fields solve it and saturate $S \geq |\int \mathrm{tr}\, F \wedge F|$.
- Physics dictionary: $A_{\text{math}} = -ig_{\text{YM}}\mathcal{A}^a R_a$, $V = g^{-1}$.

T3.9 Parallel transport, holonomy and Wilson loops

In Section T3.8 the gauge field entered as the first Taylor coefficient of a comparator $U(x, x+\varepsilon n)$ that carries a vector between two *infinitely close* copies of $\mathbb{C}^n$. The question of this section is: *how do we carry a vector along a finite path, and what do we learn when the path is closed?* The answer is an ordinary differential equation along the path, whose solution is a path-ordered exponential of A. For a closed loop the result is a group element, the **holonomy**, and its trace, the **Wilson loop**, is the simplest gauge-invariant observable there is. Small loops measure the curvature F. Large loops can see gauge fields that have no curvature at all where the particle travels, which is the Aharonov–Bohm effect. The whole of lattice gauge theory, and much of Nash's Chapter 12, is written in terms of these objects.

T3.9.1 Parallel transport along a curve

Let $\gamma : [0,1] \to M$, $t \mapsto x^\mu(t)$, be a smooth path from $a = \gamma(0)$ to $b = \gamma(1)$, with velocity $\dot\gamma^\mu = \mathrm{d}x^\mu/\mathrm{d}t$. Break the path into many short steps of duration Δt. Over each step the comparator of (T3.49) tells us how to move a vector, and we simply chain the steps together.

A field ψ along the path is **parallel transported** if it is "covariantly constant along γ": at every step, the value at the new point is what the comparator produces from the value at the

old point. By the derivation of (T3.50), that means the covariant derivative in the direction of motion vanishes:

$$0 = \dot\gamma^\mu D_\mu\psi \overset{(1)}{=} \dot\gamma^\mu\partial_\mu\psi + \dot\gamma^\mu A_\mu\psi \overset{(2)}{=} \frac{\mathrm{d}\psi}{\mathrm{d}t} + A(\dot\gamma)\,\psi, \qquad A(\dot\gamma) \coloneqq A_\mu\big(x(t)\big)\,\dot\gamma^\mu(t). \tag{T3.67}$$

(1) $D_\mu = \partial_\mu + A_\mu$, (T3.51).
(2) Chain rule: $\dot\gamma^\mu\partial_\mu\psi = \frac{\mathrm{d}}{\mathrm{d}t}\psi(x(t))$. $A(\dot\gamma)$ is the 1-form A evaluated on the velocity vector: a matrix-valued function of t alone.

This is a linear, first-order ODE for the column vector $\psi(t)$, with a time-dependent matrix coefficient $-A(\dot\gamma)$. It has a unique solution for every initial value $\psi(0)$, and the solution depends linearly on $\psi(0)$. So there is a matrix $W_\gamma(t)$ with

$$\psi(t) = W_\gamma(t)\,\psi(0), \qquad \frac{\mathrm{d}W_\gamma}{\mathrm{d}t} + A(\dot\gamma)\,W_\gamma = 0, \qquad W_\gamma(0) = \mathbb{1}.$$

The matrix $W_\gamma \coloneqq W_\gamma(1)$ is the **parallel transport** (or **Wilson line**) along γ.

Working definition T3.20 (parallel transport)

Given a gauge field A and a path γ from a to b, solve

$$\frac{\mathrm{d}W(t)}{\mathrm{d}t} = -A\big(\dot\gamma(t)\big)\,W(t), \qquad W(0) = \mathbb{1}, \tag{T3.68}$$

and set $W_\gamma = W(1)$. It maps the copy of $\mathbb{C}^n$ at a to the copy at b. *Operational check:* $W_\gamma \in G$ (for A valued in $\mathfrak{g}$), it does not depend on how the path is parametrised [how fast γ is traversed, i.e. replacing t by an increasing $s(t)$ with $s(0) = 0$, $s(1) = 1$], and it is the identity if $A(\dot\gamma) = 0$ along the path.

Two of the claims in the box deserve a line each. $W_\gamma \in G$: for $G = \mathrm{U}(n)$,

$$\frac{\mathrm{d}}{\mathrm{d}t}(W^\dagger W) \overset{(1)}{=} \dot W^\dagger W + W^\dagger\dot W \overset{(2)}{=} -W^\dagger A(\dot\gamma)^\dagger W - W^\dagger A(\dot\gamma)W \overset{(3)}{=} 0,$$

(1) Product rule.
(2) The ODE (T3.68) and its Hermitian conjugate $\dot W^\dagger = -W^\dagger A(\dot\gamma)^\dagger$.
(3) A is anti-Hermitian: $A^\dagger = -A$.

so $W^\dagger W$ stays equal to its initial value $\mathbb{1}$. (For $\mathrm{SU}(n)$, $\frac{\mathrm{d}}{\mathrm{d}t}\det W = -\operatorname{tr}A(\dot\gamma)\det W = 0$ as well.) The ODE is a curve in G whose velocity, translated back to the identity, lies in $\mathfrak{g}$: it is the ODE $\dot g = Xg$ of the one-parameter subgroups (T3.13) with a generator that changes along the way. *Reparametrisation:* if $t = t(s)$, then $A(\mathrm{d}\gamma/\mathrm{d}s) = A(\dot\gamma)\,\mathrm{d}t/\mathrm{d}s$ and $\mathrm{d}W/\mathrm{d}s = (\mathrm{d}W/\mathrm{d}t)(\mathrm{d}t/\mathrm{d}s)$, so the equation keeps its form; W depends only on the geometric path and its direction.

The abelian case. For $G = \mathrm{U}(1)$ everything commutes, and the ODE is solved by an ordinary exponential:

$$W_\gamma = \exp\Big(-\int_0^1 A(\dot\gamma)\,\mathrm{d}t\Big) = \exp\Big(-\int_\gamma A\Big). \tag{T3.69}$$

Check: differentiating $\exp(-\int_0^t A(\dot\gamma)\mathrm{d}t')$ with respect to t brings down $-A(\dot\gamma(t))$, which commutes with the exponential. With $A = ie\,a$ (the QED convention of Example T3.34) this is the phase $e^{-ie\int_\gamma a}$, the QFT notes' Wilson line [QFT notes (19.4.1)].

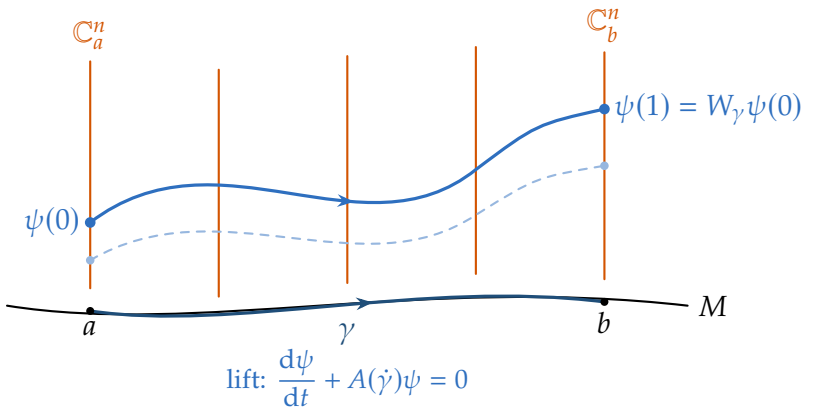


Figure T3.25: Parallel transport. Over each point of the path γ sits a copy of $\mathbb{C}^n$ (orange). Starting from any vector $\psi(0)$ in the copy over a, the ODE (T3.67) produces a unique lifted curve (blue) ending over b. A different starting vector (faint) gives a different lift. The map $\psi(0) \mapsto \psi(1)$ is linear: it is the matrix W_γ, Nash's $PT_{\gamma_{ab}}(A) : P_a \to P_b$.

T3.9.2 Path ordering

For non-abelian G the matrices $A(\dot\gamma(t))$ at different times do not commute and (T3.69) is wrong. The cure is the same one used for time-dependent Hamiltonians in quantum mechanics. Write $\mathcal{A}(t) := A(\dot\gamma(t))$ and turn the ODE into an integral equation, then iterate:

$$\begin{aligned} W(t) &\overset{(1)}{=} \mathbb{1} - \int_0^t dt_1\, \mathcal{A}(t_1)\, W(t_1) \\ &\overset{(2)}{=} \mathbb{1} - \int_0^t dt_1\, \mathcal{A}(t_1) + \int_0^t dt_1 \int_0^{t_1} dt_2\, \mathcal{A}(t_1)\mathcal{A}(t_2)\, W(t_2) \\ &\overset{(3)}{=} \sum_{k=0}^{\infty} (-1)^k \int_{t>t_1>t_2>\cdots>t_k>0} \mathcal{A}(t_1)\mathcal{A}(t_2)\cdots\mathcal{A}(t_k)\, dt_1 \cdots dt_k. \end{aligned} \tag{T3.70}$$

(1) Integrate (T3.68) from 0 to t and use $W(0) = \mathbb{1}$.
(2) Substitute the same equation for $W(t_1)$ inside the integral.
(3) Keep going. The k-th term has k factors of $-\mathcal{A}$, ordered with the *latest time on the left*.

In the k-th term the region $t > t_1 > \cdots > t_k > 0$ is one of the $k!$ orderings of the cube $[0,t]^k$. If we define the **path-ordering** symbol $\mathcal{P}$ to rearrange any product so that later times stand to the left, then each of the $k!$ orderings contributes the same, and

$$\boxed{W_\gamma = \mathcal{P}\exp\Big(-\int_\gamma A\Big) := \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} \int_0^1 \cdots \int_0^1 \mathcal{P}\big[\mathcal{A}(t_1)\cdots\mathcal{A}(t_k)\big]\, dt_1 \cdots dt_k} \tag{T3.71}$$

This is the QFT notes' [QFT notes (19.4.6)–(19.4.8)], where the ordering is written with step functions $\theta(\sigma_1 - \sigma_2)$.

Physical picture: It is the Dyson series

In the interaction picture of quantum mechanics, $i\,dU/dt = H_I(t)U$ is solved by $U = \mathcal{T}\exp(-i\int H_I\,dt)$, with the time-ordering symbol $\mathcal{T}$ putting later Hamiltonians to the left. Parallel transport is the same mathematics with $iH_I(t) \to A(\dot\gamma(t))$ and "time" $\to$ "position along the path". The Wilson line is the time-evolution operator of the internal degrees of freedom (colour, isospin) of a very heavy particle dragged along γ.

The most useful way to *compute* a path-ordered exponential is to chop the path into pieces on which $\mathcal{A}$ is (nearly) constant. The ODE is then solved exactly on each piece, $W_{\text{piece}} = e^{-\Delta t\,\mathcal{A}}$, and the pieces multiply in order:

$$W_\gamma = \lim_{N\to\infty} e^{-\Delta t\,\mathcal{A}(t_N)} \cdots e^{-\Delta t\,\mathcal{A}(t_2)}\, e^{-\Delta t\,\mathcal{A}(t_1)}, \qquad \Delta t = 1/N. \tag{T3.72}$$

This is exactly how lattice gauge theory defines things, with one group element per link (Section T3.9.8).

Composition and reversal. Two facts follow directly from the ODE. If γ_1 runs from a to b and γ_2 from b to c, then $\gamma_2 * \gamma_1$ (first γ_1, then γ_2) has

$$W_{\gamma_2*\gamma_1} = W_{\gamma_2}\, W_{\gamma_1}, \qquad W_{\gamma^{-1}} = W_\gamma^{-1}, \tag{T3.73}$$

where γ^{-1} is γ run backwards. The first holds because the solution on the second leg starts from the value reached at the end of the first. The second holds because running backwards, the transported vector retraces its steps: the composite $\gamma^{-1} * \gamma$ transports $\psi(0)$ to $\psi(1)$ and back to $\psi(0)$. Note the order: *the first path taken stands on the right*, just as in (T3.72).

Worked example T3.39: A path with two straight legs

Take $G = \mathrm{SU}(2)$ with $T_a = -\frac{i}{2}\sigma_a$, and the constant potential $A = \alpha\, T_1\, \mathrm{d}x + \beta\, T_3\, \mathrm{d}y$ on $\mathbb{R}^2$. Go from $(0,0)$ to $(1,0)$ along x (γ_1), then from $(1,0)$ to $(1,1)$ along y (γ_2). On γ_1, $\dot\gamma = \partial_x$ and $A(\dot\gamma) = \alpha T_1$; on γ_2, $A(\dot\gamma) = \beta T_3$. Both are constant, so each leg is a single exponential and

$$\begin{aligned} W_{\gamma_2*\gamma_1} &\overset{(1)}{=} e^{-\beta T_3}\, e^{-\alpha T_1} \overset{(2)}{=} \Big(\cos\tfrac{\beta}{2} + i\sin\tfrac{\beta}{2}\,\sigma_3\Big)\Big(\cos\tfrac{\alpha}{2} + i\sin\tfrac{\alpha}{2}\,\sigma_1\Big) \\ &\overset{(3)}{=} \cos\tfrac{\alpha}{2}\cos\tfrac{\beta}{2}\,\mathbb{1} + i\sin\tfrac{\alpha}{2}\cos\tfrac{\beta}{2}\,\sigma_1 - i\sin\tfrac{\alpha}{2}\sin\tfrac{\beta}{2}\,\sigma_2 + i\cos\tfrac{\alpha}{2}\sin\tfrac{\beta}{2}\,\sigma_3. \end{aligned}$$

(1) Composition (T3.73): the first leg on the right; on a leg with constant $\mathcal{A}$ the ODE gives $e^{-\mathcal{A}}$.
(2) The SU(2) exponential (T3.15): $e^{-\alpha T_1} = e^{\frac{i\alpha}{2}\sigma_1} = \cos\frac{\alpha}{2} + i\sin\frac{\alpha}{2}\sigma_1$.
(3) Multiply out, using $\sigma_3\sigma_1 = i\sigma_2$, so $(i\sigma_3)(i\sigma_1) = -i\sigma_2$.

Going the other way round, first along y and then along x via $(0,1)$, gives $e^{-\alpha T_1}e^{-\beta T_3}$, which differs only in the sign of the σ_2 term (because $\sigma_1\sigma_3 = -i\sigma_2$). *Parallel transport depends on the path.* The two results agree only if $\sin\frac{\alpha}{2}\sin\frac{\beta}{2} = 0$. For comparison, this constant potential has $F = \alpha\beta\,[T_1, T_3]\,\mathrm{d}x\wedge\mathrm{d}y = -\alpha\beta\, T_2\,\mathrm{d}x\wedge\mathrm{d}y \neq 0$, as in Example T3.36. We return to the precise relation between path dependence and F below.

T3.9.3 Gauge covariance of parallel transport

Under a gauge transformation, $\psi \mapsto g^{-1}\psi$ and $A \mapsto A^g = g^{-1}Ag + g^{-1}\mathrm{d}g$. A vector transported along γ should still be transported after we change all the frames, so we expect

$$W_\gamma(A^g) = g(b)^{-1}\, W_\gamma(A)\, g(a), \tag{T3.74}$$

the finite version of the comparator law (T3.48). Here is the check, following the QFT notes' argument [QFT notes (19.4.9)–(19.4.24)]. Write $g_t = g(\gamma(t))$ and $V(t) \coloneqq g_t^{-1}W(t)g_0$, with $W(t)$ the

solution for A. Then

$$\begin{aligned}\frac{\mathrm{d}V}{\mathrm{d}t} &\overset{(1)}{=} -g_t^{-1}\dot g_t\, g_t^{-1}\, W(t)g_0 + g_t^{-1}\dot W(t)g_0\\ &\overset{(2)}{=} -g_t^{-1}\dot g_t\, g_t^{-1} W(t)g_0 - g_t^{-1}\mathcal{A}(t)W(t)g_0\\ &\overset{(3)}{=} -\big(g_t^{-1}\mathcal{A}(t)g_t + g_t^{-1}\dot g_t\big)\, g_t^{-1}W(t)g_0 \overset{(4)}{=} -A^g(\dot\gamma)\,V(t).\end{aligned}$$

(1) Product rule and (T3.35); $\dot g_t = \frac{\mathrm{d}}{\mathrm{d}t}g(\gamma(t)) = (\mathrm{d}g)(\dot\gamma)$.
(2) The ODE for W.
(3) Insert $\mathbb{1} = g_t g_t^{-1}$ between $\mathcal{A}(t)$ and $W(t)$ and factor out $g_t^{-1}Wg_0 = V$.
(4) The bracket is A^g evaluated on $\dot\gamma$.

So V solves the transport equation for A^g, and $V(0) = g_0^{-1}g_0 = \mathbb{1}$. By uniqueness of solutions, $V(t)$ *is* the transport for A^g, which is (T3.74).

Reading Nash (p. 217, (8.2)–(8.9))

Nash sets up parallel transport on the principal bundle P: in local coordinates (x, g) a path γ_{ab} is lifted to $\hat\gamma_{ab}(t) = (\gamma_{ab}(t), g(t))$ with $\frac{\mathrm{d}g}{\mathrm{d}t} + A\dot\gamma_{ab}\, g = 0$ [Nash (8.4)] – our (T3.68) with W renamed g. The map $PT_{\gamma_{ab}}(A) : P_a \to P_b$, $g(0) \mapsto g(1)$ [Nash (8.7)] is our W_γ. His commutation relation $\alpha_b \circ PT_{\gamma_{ab}}(A) = PT_{\gamma_{ab}}(\alpha\cdot A)\circ\alpha_a$ [Nash (8.8)] is (T3.74), written as $g(b)^{-1}W_\gamma(A) = W_\gamma(A^g)g(a)^{-1}$ with his $\alpha = g^{-1}$ acting on fibres. From it he proves that the based gauge group acts freely on $\mathcal{A}$: if $A^g = A$ and $g(a) = \mathbb{1}$, then $g(b)^{-1}W_\gamma = W_\gamma$ for every path from a, so $g(b) = \mathbb{1}$ for every b [Nash (8.9)].

T3.9.4 Holonomy and the Wilson loop

Now close the path: γ is a loop based at a, $\gamma(0) = \gamma(1) = a$. Parallel transport maps the copy of $\mathbb{C}^n$ at a to *itself*, so W_γ is simply an element of G, acting on one vector space.

Working definition T3.21 (holonomy and Wilson loop)

The **holonomy** of A around a loop γ based at a is

$$\mathrm{Hol}_\gamma(A) = W_\gamma = \mathcal{P}\exp\Big(-\oint_\gamma A\Big) \in G$$

[$\mathcal{P}$ is path ordering: the product of the short-step transports $\mathbb{1} - A(\dot\gamma)\Delta t$, with later steps standing to the left]. Under a gauge transformation it is conjugated, $\mathrm{Hol}_\gamma(A^g) = g(a)^{-1}\,\mathrm{Hol}_\gamma(A)\,g(a)$ (put $b = a$ in (T3.74)). Its trace in a representation R,

$$W_R(\gamma) := \mathrm{tr}_R\ \mathcal{P}\exp\Big(-\oint_\gamma A\Big), \tag{T3.75}$$

is therefore *gauge invariant*: the **Wilson loop**. *How to compute it:* chop γ into legs, multiply the leg transports right to left in the order travelled, and take the trace.

The gauge invariance of the trace is one line, as in [QFT notes (19.4.25)]: $\mathrm{tr}(g^{-1}Wg) = \mathrm{tr}(Wgg^{-1}) = \mathrm{tr}\,W$. More generally, any conjugation-invariant function of the holonomy (its eigenvalues, its conjugacy class) is gauge invariant, and for SU(2) the trace in the fundamental already determines the conjugacy class (Example T3.3).

Changing the base point. If we move the base point from a to c along a path σ from a to c, the loop $\sigma * \gamma * \sigma^{-1}$ based at c has holonomy $W_\sigma \operatorname{Hol}_\gamma W_\sigma^{-1}$ by (T3.73). Again only the conjugacy class matters: the trace does not care where on the loop we start.

Working definition T3.22 (holonomy group; reducible and irreducible)

As γ ranges over all loops based at a, the holonomies form a subgroup $\operatorname{Hol}_a(A) \subset G$: loops compose by (T3.73), and the reversed loop gives the inverse. This is the **holonomy group**. For a connected [any two points are joined by a path] M the groups at different base points are conjugate. If $\operatorname{Hol}_a(A) = G$, the connection is **irreducible**; if it is a proper subgroup, A is **reducible**. *Test:* a connection is reducible if there is a gauge in which all its components $A_\mu(x)$ lie in a proper subalgebra of $\mathfrak{g}$ [a smaller space of generators closed under commutators, e.g. the multiples of T_3 inside $\mathfrak{su}(2)$] (then every transport lies in the corresponding subgroup).

Worked example T3.40: A reducible SU(2) connection and its stabiliser

Let $A = a_\mu(x)\, T_3\, \mathrm{d}x^\mu$ with arbitrary real functions a_μ. Every $A(\dot\gamma)$ is a multiple of T_3, and multiples of T_3 commute, so

$$\operatorname{Hol}_\gamma(A) = \exp\Big(-T_3 \oint_\gamma a\Big) = \begin{pmatrix} e^{\frac{i}{2}\Phi} & 0 \\ 0 & e^{-\frac{i}{2}\Phi} \end{pmatrix}, \qquad \Phi = \oint_\gamma a .$$

The holonomy group lies in the maximal torus $\mathrm{U}(1) = \{e^{sT_3}\} \subset \mathrm{SU}(2)$: A is reducible. Which constant gauge transformations fix A? We need $g^{-1}T_3 g = T_3$, i.e. g commutes with T_3: $g = e^{sT_3}$. So the stabiliser is the whole U(1), not just the centre $\{\pm\mathbb{1}\}$. This is the general rule quoted in Section T3.7: the stabiliser of a connection is the centraliser of its holonomy group [Nash p. 232, (8.88)–(8.90)]. An irreducible connection has holonomy group SU(2), whose centraliser is the centre $\{\pm\mathbb{1}\}$.

Reading Nash (p. 218, (8.10))

Nash closes the path to a loop γ_a and obtains $PT_{\gamma_a}(A) : P_a \to P_a$ [Nash (8.10)]. "As γ_a varies over all the loops based at a the set of all the $PT_{\gamma_a}(A)$ form a group known as the holonomy group of A at a." If it is all of G, A is irreducible; otherwise reducible. This is exactly the working definition above. Reducible connections are where the space of gauge orbits $\mathcal{A}/\mathcal{G}$ is singular, which matters for instanton moduli spaces (Chapter 8) and for Floer theory (Chapter 12).

T3.9.5 Small loops measure the curvature

We now make precise the statement, made pictorially in Figure T3.6, that curvature is the failure of a small loop of transports to close. Put the base point at x and go around a small parallelogram spanned by $\varepsilon\, e_1$ and $\delta\, e_2$ (two coordinate directions $\mu = 1, 2$), counterclockwise: along e_1, then e_2, then back along $-e_1$, then $-e_2$ (Figure T3.26).

Write $a = A_1(x)$, $b = A_2(x)$. With each potential evaluated at the midpoint of its leg (which is accurate to the order we need) the four leg exponents are

$$X_1 = -\varepsilon\big(a + \tfrac{\varepsilon}{2}\partial_1 a\big), \qquad X_2 = -\delta\big(b + \varepsilon\,\partial_1 b + \tfrac{\delta}{2}\partial_2 b\big),$$
$$X_3 = +\varepsilon\big(a + \tfrac{\varepsilon}{2}\partial_1 a + \delta\,\partial_2 a\big), \qquad X_4 = +\delta\big(b + \tfrac{\delta}{2}\partial_2 b\big),$$

up to third-order terms, and the holonomy is $e^{X_4}e^{X_3}e^{X_2}e^{X_1}$. The Baker–Campbell–Hausdorff

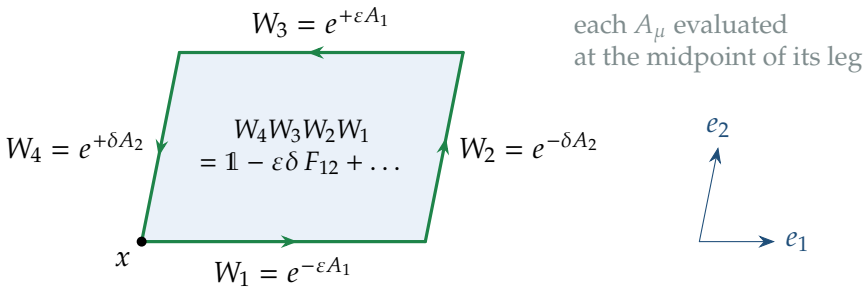


Figure T3.26: A small loop, travelled counterclockwise from x. Each leg contributes the exponential of $\mp$(length)$\times A_\mu$ at its midpoint. The four factors would cancel if the A_μ were the same on opposite legs and commuted with each other. The first failure is the change of A across the loop ($\partial_1 A_2 - \partial_2 A_1$); the second is non-commutativity ($[A_1, A_2]$). Together they are F_{12}. In lattice gauge theory this loop is a plaquette.

formula (T3.19), to second order, gives $\log(e^{X_4}e^{X_3}e^{X_2}e^{X_1}) = \sum_i X_i + \frac{1}{2}\sum_{i>j}[X_i, X_j] + O(3)$. First the linear terms:

$$X_1 + X_2 + X_3 + X_4 \overset{(1)}{=} \varepsilon\delta\,\partial_2 a - \varepsilon\delta\,\partial_1 b = -\varepsilon\delta\left(\partial_1 A_2 - \partial_2 A_1\right).$$

(1) In $X_1 + X_3$ everything cancels except $\varepsilon\delta\,\partial_2 a$: the potential on the top leg exceeds the one on the bottom leg by the change of A_1 over the height δ. In $X_2 + X_4$ everything cancels except $-\varepsilon\delta\,\partial_1 b$ from the change of A_2 over the width ε.

Then the commutators, where only the leading parts $X_1 \approx -\varepsilon a$, $X_2 \approx -\delta b$, $X_3 \approx \varepsilon a$, $X_4 \approx \delta b$ matter:

$$\tfrac{1}{2}\sum_{i>j}[X_i, X_j] \overset{(1)}{=} \tfrac{1}{2}\Big([X_2, X_1] + [X_3, X_2] + [X_4, X_1] + [X_4, X_3]\Big)$$
$$\overset{(2)}{=} \tfrac{1}{2}\,\varepsilon\delta\Big([b, a] - [a, b] + [a, b] + [b, a]\Big) \overset{(3)}{=} -\varepsilon\delta\,[A_1, A_2].$$

(1) $[X_3, X_1]$ and $[X_4, X_2]$ vanish at this order (opposite legs have proportional leading exponents).
(2) $[X_2, X_1] = \varepsilon\delta[b, a]$, $[X_3, X_2] = -\varepsilon\delta[a, b]$, $[X_4, X_1] = -\varepsilon\delta[b, a] = \varepsilon\delta[a, b]$, $[X_4, X_3] = \varepsilon\delta[b, a]$.
(3) $[b, a] = -[a, b]$, so the bracket is $-2[a, b]$.

Adding, and exponentiating,

$$\boxed{\mathrm{Hol}_\square(A) = \mathbb{1} - \varepsilon\delta\, F_{12}(x) + O(3), \qquad F_{12} = \partial_1 A_2 - \partial_2 A_1 + [A_1, A_2]} \tag{T3.76}$$

which is the component formula (T3.57). More invariantly: for a small loop spanning the tangent 2-plane of vectors u, v at x, $\mathrm{Hol} = \mathbb{1} - F(u, v) + \ldots$, where $F(u, v) = F_{\mu\nu}u^\mu v^\nu$ is the curvature 2-form evaluated on the two vectors.

Key idea: Curvature is holonomy per unit area

$F_{\mu\nu}(x)$ is (minus) the rotation of internal space picked up by going around an infinitesimal loop in the $\mu\nu$ plane at x, per unit area. Zero curvature means small loops have trivial holonomy; nonzero curvature means parallel transport is path dependent. The Wilson loop of a small loop is

$$\mathrm{tr}_R\ \mathrm{Hol}_\square = \dim R - \varepsilon\delta\ \mathrm{tr}_R\, F_{12} + \tfrac{1}{2}\varepsilon^2\delta^2\ \mathrm{tr}_R(F_{12}F_{12}) + \ldots,$$

and for $\mathfrak{su}(N)$ the linear term is zero ($\operatorname{tr} F = 0$), so the first gauge-invariant information is $\operatorname{tr} F_{12}^2$: exactly the Yang–Mills Lagrangian density.

Abelian Stokes, and why there is no simple non-abelian Stokes. For U(1) the path ordering is irrelevant, and Stokes' theorem turns the line integral into a flux: if $\gamma = \partial S$,

$$\mathrm{Hol}_\gamma(A) = \exp\Big(-\oint_{\partial S} A\Big) = \exp\Big(-\int_S F\Big), \tag{T3.77}$$

the QFT notes' [QFT notes (19.4.4)–(19.4.5)]. In QED this is the phase $e^{-ie\Phi_B}$ with Φ_B the magnetic flux through S. For non-abelian G there is no formula this simple: the local contributions $\mathbb{1} - F\,\mathrm{d}\sigma$ from the little plaquettes tiling S must each be transported back to the base point before they are multiplied, and they do not commute. The small-loop formula (T3.76) is the only piece that survives unchanged. The next example shows what happens for a finite loop.

Worked example T3.41: Wilson loop of a constant non-abelian field

Return to Example T3.36: $G = \mathrm{SU}(2)$, $A = a\,T_1\,\mathrm{d}x + b\,T_2\,\mathrm{d}y$ with constants a, b, so $F = ab\,T_3\,\mathrm{d}x \wedge \mathrm{d}y$ is constant too. Go counterclockwise around the rectangle $[0, L_1] \times [0, L_2]$ from the origin, and write $\alpha = aL_1$, $\beta = bL_2$. Each leg has constant $A(\dot\gamma)$, so

$$\mathrm{Hol} \overset{(1)}{=} e^{+\beta T_2}\, e^{+\alpha T_1}\, e^{-\beta T_2}\, e^{-\alpha T_1} \overset{(2)}{=} \big(PQP^{-1}\big)\, Q^{-1}, \qquad P = e^{\beta T_2},\ Q = e^{\alpha T_1}.$$

(1) Legs in order: $+x$ ($e^{-\alpha T_1}$), $+y$ ($e^{-\beta T_2}$), $-x$ ($e^{+\alpha T_1}$), $-y$ ($e^{+\beta T_2}$); the first leg stands on the right.

(2) Group the factors: the holonomy is a *group commutator* (T3.5).

Now use the SU(2) exponential. $Q = \cos\frac{\alpha}{2} - i \sin\frac{\alpha}{2}\,\sigma_1$, and $PQP^{-1} = \cos\frac{\alpha}{2} - i \sin\frac{\alpha}{2}\,P\sigma_1 P^{-1}$, where conjugating by $P = e^{-\frac{i\beta}{2}\sigma_2}$ rotates σ_1 about the 2-axis: $P\sigma_1 P^{-1} = \cos\beta\,\sigma_1 - \sin\beta\,\sigma_3$ (Section T3.4). The trace of a product of two SU(2) elements $c - i\,s\,\hat m \cdot \vec\sigma$ and $c' + i\,s'\,\hat n \cdot \vec\sigma$ is $2(cc' + ss'\,\hat m \cdot \hat n)$, since $\operatorname{tr}(\sigma_i\sigma_j) = 2\delta_{ij}$. Hence

$$\operatorname{tr}\mathrm{Hol} \overset{(1)}{=} 2\Big(\cos^2\tfrac{\alpha}{2} + \sin^2\tfrac{\alpha}{2}\,\cos\beta\Big) \overset{(2)}{=} 2 - 2\sin^2\tfrac{\alpha}{2}\,(1 - \cos\beta) \overset{(3)}{=} 2 - 4\sin^2\tfrac{aL_1}{2}\,\sin^2\tfrac{bL_2}{2}.$$

(1) $Q^{-1} = \cos\frac{\alpha}{2} + i\sin\frac{\alpha}{2}\,\sigma_1$, so $\hat n = \hat e_1$; and $\hat m = (\cos\beta, 0, -\sin\beta)$, so $\hat m \cdot \hat n = \cos\beta$.

(2) $\cos^2 + \sin^2 = 1$.

(3) $1 - \cos\beta = 2\sin^2\frac{\beta}{2}$.

Small loop check. For small L_1, L_2, $\operatorname{tr}\mathrm{Hol} \approx 2 - 4\cdot\frac{\alpha^2}{4}\cdot\frac{\beta^2}{4} = 2 - \frac{1}{4}(ab\,L_1L_2)^2$. The small-loop formula predicts $\operatorname{tr}\big(\mathbb{1} - L_1L_2F_{12} + \frac{1}{2}L_1^2L_2^2F_{12}^2\big) = 2 + \frac{1}{2}(abL_1L_2)^2\operatorname{tr}T_3^2 = 2 - \frac{1}{4}(abL_1L_2)^2$, using $\operatorname{tr}T_3 = 0$ and $\operatorname{tr}T_3^2 = -\frac{1}{2}$. They agree. *Large loop.* The naive guess "exp(−area × F)" would give $\operatorname{tr} e^{-L_1L_2\,ab\,T_3} = 2\cos\frac{abL_1L_2}{2}$, which depends only on the area L_1L_2. The true answer does not: at fixed area it depends on the shape, and it is periodic in L_1 and L_2 separately. The non-abelian field is not "flux through the loop".

T3.9.6 The Aharonov–Bohm effect: holonomy without field

Everything so far suggests that holonomy is a way of measuring F. The Aharonov–Bohm effect shows it is more: a gauge field can have *zero* field strength everywhere a particle goes and still produce a measurable holonomy. The physical set-up (Figure T3.28): an infinitely long thin

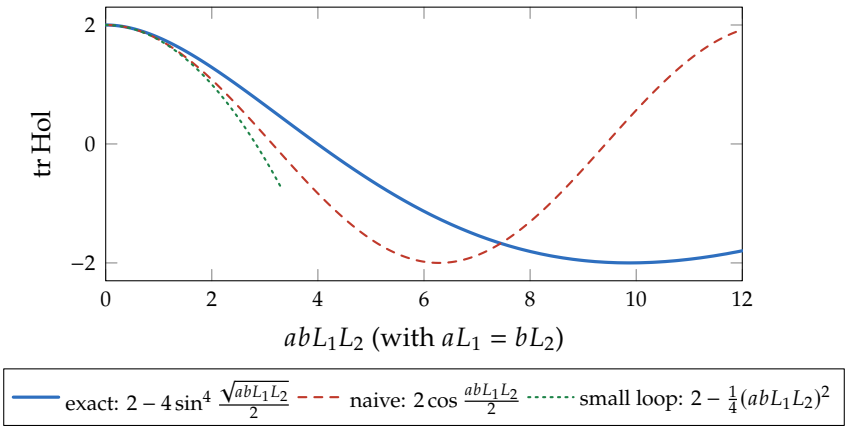


Figure T3.27: The Wilson loop of Example T3.41 for square loops ($aL_1 = bL_2$), as a function of the "flux" abL_1L_2. All three curves agree for small loops (the green small-loop formula (T3.76)). For larger loops the exact answer (blue) never goes below -2 and oscillates at a different rate from the naive abelian guess (red): non-abelian holonomy is not the exponential of a flux.

solenoid along the z-axis carries magnetic flux Φ, confined inside it. Electrons are kept out of the solenoid by a wall, so they live in the plane with a disc removed, $M = \mathbb{R}^2 \setminus \{r < r_0\}$. Outside the solenoid $\vec{B} = 0$, yet the vector potential cannot vanish: Stokes' theorem around any circle enclosing the solenoid gives $\oint \vec{a} \cdot \mathrm{d}\vec{l} = \Phi$. The symmetric choice is

$$\vec{a} = \frac{\Phi}{2\pi r}\,\hat{\phi}, \qquad a = \frac{\Phi}{2\pi}\,\mathrm{d}\phi \quad (r > r_0),$$

since the 1-form dual to $\hat{\phi}/r$ is $\mathrm{d}\phi$ (Chapter T1). Its field strength is $\mathrm{d}a = \frac{\Phi}{2\pi}\mathrm{d}\mathrm{d}\phi = 0$ on M. (The "$\mathrm{d}\phi$" is closed but not exact on M: ϕ is not a single-valued function there.)

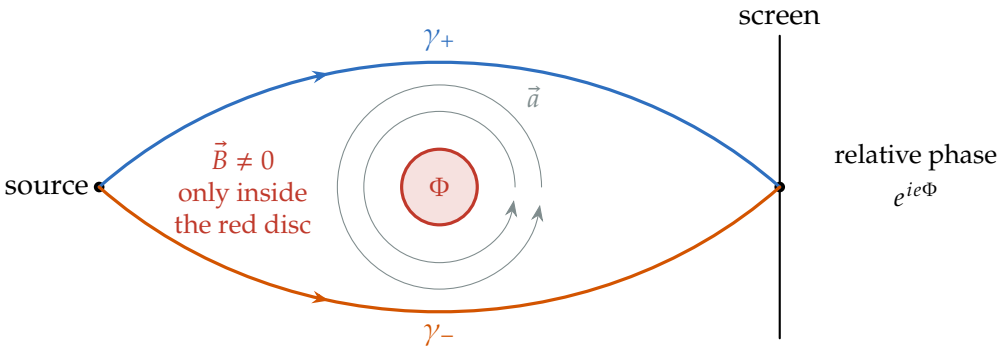


Figure T3.28: The Aharonov–Bohm experiment. The flux Φ is confined inside the solenoid (red), so electrons on either path $\gamma_\pm$ see $\vec{B} = 0$. But the vector potential (grey circles) circulates around the solenoid, and the two paths together form a loop $\gamma_-^{-1} * \gamma_+$ enclosing it. The relative phase at the screen is the holonomy $e^{ie\Phi}$ of that (clockwise) loop.

In the language of this chapter the electron's covariant derivative is $D = \mathrm{d} + A$ with $A = ie\,a$ (Example T3.34). The amplitude to reach a point on the screen along $\gamma_\pm$ is the field-free amplitude multiplied by the parallel transport, the phase $W_{\gamma_\pm} = e^{-ie\int_{\gamma_\pm} a}$ (in the semiclassical, or path-integral, picture each path carries its Wilson line). The interference depends on the ratio:

$$W_{\gamma_-}^{-1} W_{\gamma_+} \overset{(1)}{=} W_{\gamma_-^{-1}*\gamma_+} \overset{(2)}{=} \exp\Big(-ie \oint_{\gamma_-^{-1}*\gamma_+} a\Big) \overset{(3)}{=} \exp\big(+ie\Phi\big). \tag{T3.78}$$

(1) Composition and reversal (T3.73): go out along γ_+ and come back along γ_- reversed. For U(1) the order of the factors does not matter.
(2) Abelian transport (T3.69) around the closed loop.
(3) The loop goes once around the solenoid *clockwise* as drawn (over the top from left to right, back underneath), so $\oint \frac{\Phi}{2\pi} \mathrm{d}\phi = -\Phi$. Any loop winding once around the solenoid gives the same, however it is deformed: a is closed, so by Stokes the integral depends only on the winding number.

The interference pattern shifts by $e\Phi$ in phase: the fringe intensity goes like $\cos(\delta_0 + e\Phi)$, where δ_0 is the field-free path difference. The shift is periodic: it vanishes when $e\Phi \in 2\pi\mathbb{Z}$, i.e. when the flux is a multiple of the flux quantum $2\pi/e$ (which is h/e in SI units).

Key idea: Aharonov–Bohm = holonomy of a flat connection

On $M = \mathbb{R}^2 \setminus \text{disc}$ the connection $A = ie\frac{\Phi}{2\pi}\mathrm{d}\phi$ is *flat*, $F = 0$, yet its holonomy once counterclockwise around the hole is $e^{-ie\Phi} \neq 1$ (unless $e\Phi \in 2\pi\mathbb{Z}$). Locally it is pure gauge, $A = g^{-1}\mathrm{d}g$ with $g = e^{ie\Phi\phi/2\pi}$, but this g is not single-valued around the hole. Holonomy detects gauge fields that curvature cannot see; it needs a loop that cannot be shrunk, i.e. $\pi_1(M) \neq 0$. This is Nash's remark that a flat connection's "only departure from triviality lies in its holonomy, which relies on $\pi_1(M)$ to be non-vanishing" [Nash p. 244].

Watch out: Is the potential physical?

The Aharonov–Bohm effect does not make A itself observable: A can still be changed by any single-valued gauge transformation. What is observable is the gauge-invariant holonomy $e^{-ie\oint a}$, a non-local quantity. The lesson is that the complete gauge-invariant content of a connection is its set of Wilson loops, not its field strength alone. (Knowing all Wilson loops determines the connection up to gauge; knowing F does not, as this example shows.)

T3.9.7 Holonomy on the 2-sphere and Berry's phase

Two more classic holonomies are both computed on a circle of latitude $\vartheta = \vartheta_0$ of the unit sphere, traversed once in the direction of increasing ϕ. They are two faces of the same fact: on a curved space the holonomy of a small-ish loop is the enclosed curvature.

Worked example T3.42: A tangent vector carried around a latitude

Parallel transport of tangent vectors on S^2 (the Levi-Civita connection of Dray's Chapters 17–18, Chapter T2) is an SO(2) gauge field. In the orthonormal frame $(\hat{e}_\vartheta, \hat{e}_\phi)$, the tangential parts of the frame derivatives are [Dray Ch. 17–18]

$$\mathrm{d}\hat{e}_\vartheta = \cos\vartheta\,\mathrm{d}\phi\;\hat{e}_\phi, \qquad \mathrm{d}\hat{e}_\phi = -\cos\vartheta\,\mathrm{d}\phi\;\hat{e}_\vartheta,$$

so, in the convention $\mathrm{d}e_j = e_i A^i{}_j$ of Section T3.8.4,

$$A = \begin{pmatrix} 0 & -\cos\vartheta \\ \cos\vartheta & 0 \end{pmatrix}\mathrm{d}\phi = \cos\vartheta\, L\,\mathrm{d}\phi, \qquad L = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},$$

where e^{sL} is the rotation by angle s (Example T3.9). Along the latitude $\phi \mapsto (\vartheta_0, \phi)$ we have $A(\dot\gamma) = \cos\vartheta_0\, L$, constant, so

$$\mathrm{Hol} \overset{(1)}{=} \exp\!\big(-2\pi\cos\vartheta_0\, L\big) \overset{(2)}{=} \text{rotation by } -2\pi\cos\vartheta_0 \overset{(3)}{=} \text{rotation by } 2\pi(1-\cos\vartheta_0).$$

(1) Constant $A(\dot\gamma)$ on a path of parameter length 2π.
(2) e^{sL} is the rotation by s.
(3) Add 2π: a full turn is the identity in SO(2).

The angle $2\pi(1-\cos\vartheta_0)$ is the area of the polar cap enclosed by the latitude. A vector carried around a small circle near the north pole comes back almost unrotated; carried around the equator it comes back rotated by 2π, i.e. unrotated, and the equator is a geodesic, along which the tangent vector is transported without turning relative to the curve. The curvature is $F = \mathrm{d}A = -\sin\vartheta\,\mathrm{d}\vartheta\wedge\mathrm{d}\phi\;L$, so $-F$ is L times the area form, and the small-loop formula gives rotation by (area) × (Gaussian curvature = 1), in agreement. For a cap, abelian Stokes (T3.77) gives the same answer exactly, because SO(2) is abelian: this is the local Gauss–Bonnet theorem.

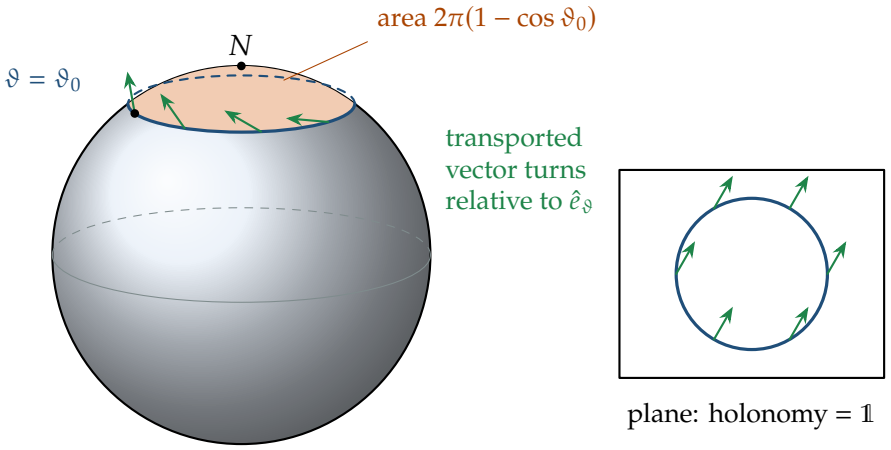


Figure T3.29: Left: on the sphere a tangent vector carried around the latitude $\vartheta=\vartheta_0$ comes back rotated by the enclosed area $2\pi(1-\cos\vartheta_0)$ (orange cap), because the Levi-Civita connection has curvature. Right: in the flat plane the same experiment returns every vector unchanged; the polar-frame connection of Example T3.33 is nonzero but pure gauge, so its holonomy is trivial.

Worked example T3.43: Berry's phase of a spin-$\frac12$

A spin-$\frac12$ in a magnetic field $\vec B = B\hat n$ has ground state $\psi_N(\hat n)$ of Example T3.37, up to a phase. If the direction $\hat n$ is turned slowly around a closed loop γ on S^2, the adiabatic theorem says the state follows $\psi_N(\hat n(t))$ up to a phase $e^{i\chi(t)}$ (apart from the dynamical phase $e^{-i\int E\,\mathrm{d}t}$, which we drop), and the phase is fixed by the condition that no component of $\frac{\mathrm{d}}{\mathrm{d}t}\psi$ lies along ψ: $\psi^\dagger\frac{\mathrm{d}\psi}{\mathrm{d}t}=0$. With $\psi = e^{i\chi}\psi_N$,

$$0 = \psi^\dagger\frac{\mathrm{d}\psi}{\mathrm{d}t} \overset{(1)}{=} i\dot\chi + \psi_N^\dagger\frac{\mathrm{d}\psi_N}{\mathrm{d}t} \overset{(2)}{=} i\dot\chi + A_N(\dot\gamma),$$

(1) Product rule and $\psi_N^\dagger\psi_N = 1$.
(2) The Berry connection $A_N = \psi_N^\dagger\mathrm{d}\psi_N$.

so $\frac{\mathrm{d}}{\mathrm{d}t}e^{i\chi} = -A_N(\dot\gamma)e^{i\chi}$: the phase is parallel transported with the monopole connection. For the latitude ϑ_0,

$$e^{i\chi(1)} \overset{(1)}{=} \exp\Big(-\oint A_N\Big) \overset{(2)}{=} \exp\Big(-\tfrac{i}{2}(1-\cos\vartheta_0)\cdot 2\pi\Big) = e^{-\frac{i}{2}\Omega}, \qquad \Omega = 2\pi(1-\cos\vartheta_0),$$

(1) Abelian parallel transport (T3.69) with $\chi(0)=0$.

(2) $A_N = \frac{i}{2}(1-\cos\vartheta)\mathrm{d}\phi$, constant along the latitude.

The **Berry phase** is $-\frac{1}{2}\Omega$, minus half the solid angle swept out by the spin. It is the holonomy of the monopole (Hopf) connection, and by abelian Stokes it equals $\exp(-\int_{\text{cap}} F)$ with $F = \frac{i}{2}\sin\vartheta\,\mathrm{d}\vartheta\wedge\mathrm{d}\phi$, the monopole field of strength $\frac{1}{2}$ sitting at the degeneracy point $\vec{B}=0$. Using A_S instead (i.e. the other patch) gives $e^{-\frac{i}{2}\Omega}e^{2\pi i}$, the same number: the Dirac quantisation of Example T3.37 is what makes the Berry phase independent of the patch. For $\vartheta_0 = \pi/2$ (spin rotated once around the equator) the phase is $e^{-i\pi} = -1$: the same sign as a 2π rotation of a spinor (Section T3.4.5).

T3.9.8 Lattice gauge theory: Wilson's action

Wilson's formulation of gauge theory on a lattice takes holonomy as the *fundamental* variable. Replace spacetime by a hypercubic lattice of spacing ℓ (we write ℓ rather than the usual a to avoid a clash). A matter field lives on the sites x. The gauge field lives on the links: to each link from x to $x+\ell\hat{\mu}$ we attach the group element

$$U_{x,\mu} := W_{\text{link}} \approx e^{-\ell A_\mu(x+\frac{\ell}{2}\hat{\mu})} \in G,$$

which carries vectors from x to $x+\ell\hat{\mu}$. Gauge transformations act on sites, $U_{x,\mu} \mapsto g(x+\ell\hat{\mu})^{-1}\,U_{x,\mu}\,g(x)$, which is (T3.74) for one link. The smallest gauge-invariant object is the Wilson loop of an elementary square, a **plaquette** (Figure T3.30):

$$U_p = U_{x,\nu}^{-1}\,U_{x+\ell\hat{\nu},\mu}^{-1}\,U_{x+\ell\hat{\mu},\nu}\,U_{x,\mu}, \qquad \operatorname{tr} U_p \text{ gauge invariant.}$$

(Right to left: along μ, then along ν, then back along μ, then back along ν.)

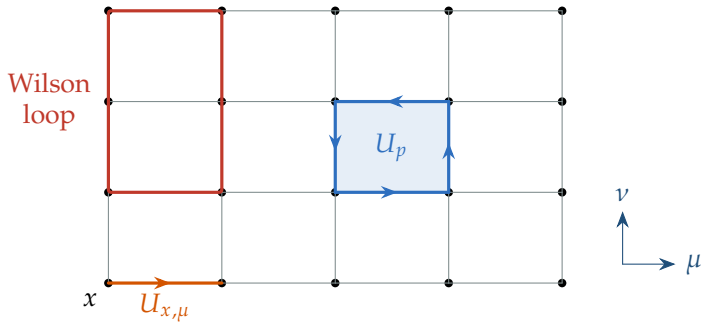


Figure T3.30: Lattice gauge theory. Group elements $U_{x,\mu}$ live on the links (orange); gauge transformations act at the sites (dots). The trace of the product around an elementary square is the plaquette variable $\operatorname{tr} U_p$ (blue), whose continuum limit is $\operatorname{tr} F_{\mu\nu}^2$. Larger Wilson loops (red) are the order parameters of confinement.

Worked example T3.44: Wilson's action becomes Yang–Mills

Wilson's action for $G=\mathrm{SU}(N)$ is $S_W = \beta\sum_p\big(1-\frac{1}{N}\operatorname{Re}\operatorname{tr}U_p\big)$, a sum over all plaquettes. By (T3.76) with $\varepsilon=\delta=\ell$, $U_p = \exp\big(-\ell^2 F_{\mu\nu} + O(\ell^3)\big)$, and for an $\mathfrak{su}(N)$ matrix X (traceless, anti-Hermitian) $\operatorname{tr}e^X = N + \frac{1}{2}\operatorname{tr}X^2 + O(X^3)$. For the symmetric plaquette the $O(\ell^3)$ terms drop out of the real part, and

$$1-\tfrac{1}{N}\operatorname{Re}\operatorname{tr}U_p \overset{(1)}{=} -\tfrac{\ell^4}{2N}\operatorname{tr}\big(F_{\mu\nu}F_{\mu\nu}\big) + O(\ell^6) \qquad \text{(no sum)},$$

$$S_W \overset{(2)}{=} \beta \sum_x \sum_{\mu<\nu} \left(-\tfrac{\ell^4}{2N}\right) \operatorname{tr} F_{\mu\nu}^2 \overset{(3)}{=} -\frac{\beta}{2N} \int \mathrm{d}^4x \sum_{\mu<\nu} \operatorname{tr} F_{\mu\nu}^2 \overset{(4)}{=} -\frac{\beta}{4N} \int \mathrm{d}^4x \, \operatorname{tr}\big(F_{\mu\nu}F_{\mu\nu}\big).$$

(1) $\operatorname{tr} F_{\mu\nu} = 0$, and the quadratic term of the exponential.
(2) Each plaquette is labelled by its corner x and a pair $\mu < \nu$.
(3) $\sum_x \ell^4 \to \int \mathrm{d}^4x$ as $\ell \to 0$.
(4) $\sum_{\mu<\nu}$ is half the sum over all μ, ν (diagonal terms vanish).

Comparing with $S_{\mathrm{YM}} = -\frac{1}{2g_{\mathrm{YM}}^2} \int \operatorname{tr}(F_{\mu\nu}F_{\mu\nu})$ of (T3.62), the lattice coupling is

$$\beta = \frac{2N}{g_{\mathrm{YM}}^2}.$$

Both terms are positive: $-\operatorname{tr} F^2 \geq 0$ for anti-Hermitian F, and $\operatorname{Re} \operatorname{tr} U \leq N$ for unitary U. The lattice theory is gauge invariant exactly, at every spacing, with no gauge fixing: only holonomies appear.

Physical picture: Wilson loops and confinement

A large rectangular Wilson loop of time extent T and spatial width R is the amplitude for creating a static quark–antiquark pair at separation R, holding it for time T, and annihilating it. Hence $\langle W(\gamma)\rangle \sim e^{-V(R)\,T}$ with $V(R)$ the static potential. If $\langle W\rangle$ falls like $e^{-\sigma RT}$, the *area law*, the potential is linear, $V = \sigma R$: quarks are confined by a string of tension σ. If it falls only like the perimeter, $e^{-c(R+T)}$, they are not. The Wilson loop is therefore the order parameter of confinement, and it is a holonomy.

Watch out: Signs and orders in the literature

Our transport solves $\dot{W} = -A(\dot{\gamma})W$, so $W = \mathcal{P}\exp(-\int A)$ with later points on the left. The QFT notes write $\mathcal{P}\exp(+ig\int \mathcal{A}^a R_a)$ [QFT notes (19.4.26)], which is the same thing under the dictionary $A = -ig_{\mathrm{YM}}\mathcal{A}$ of Section T3.8.9. Nash writes the Wilson line of Chern–Simons theory as $W(R, C) = \operatorname{tr} \mathcal{P}\exp\big[\oint_C A\big]$ [Nash p. 343, (12.84)], without the minus sign. That is the holonomy of the loop run the other way (or with the opposite ordering convention), and for a trace in a unitary representation it is the complex conjugate of ours. When comparing formulas, fix which end of the path is on the left and the sign in the exponent; nothing physical depends on the choice.

Reading Nash (holonomy and Wilson lines in Nash)

Holonomy appears three times in Nash. (i) In the proof that the based gauge group acts freely and in the definition of irreducible connections [Nash pp. 216–218, (8.2)–(8.10)]; (ii) in the statement that a flat connection is detected only by its holonomy, which requires $\pi_1(M) \neq 0$ [Nash p. 244, after (8.151)], and its use in Floer theory, where flat SU(2) connections on a homology 3-sphere are identified with representations $\operatorname{Hom}(\pi_1(M), \mathrm{SU}(2))/\operatorname{Ad}\mathrm{SU}(2)$ [Nash p. 325, (12.9)–(12.10)]; (iii) in Witten's Chern–Simons theory, where the observables are products of Wilson lines $W(R_i, C_i)$ along the components of a knot or link, and their expectation values give the Jones polynomial [Nash pp. 343–344, (12.84)–(12.85)]. The global anomaly discussion also computes a holonomy, of a determinant line bundle [Nash pp. 289–290] (Chapter 10).

Recap

- Parallel transport along γ: $\dot{W} = -A(\dot{\gamma})W$, $W(0) = \mathbb{1}$; solution $W_\gamma = \mathcal{P}\exp(-\int_\gamma A)$ (Dyson series, later points on the left), $W_\gamma \in G$. Abelian: $e^{-\int_\gamma A}$.
- $W_{\gamma_2 * \gamma_1} = W_{\gamma_2} W_{\gamma_1}$, $W_{\gamma^{-1}} = W_\gamma^{-1}$; gauge law $W_\gamma \mapsto g(b)^{-1} W_\gamma\, g(a)$.
- Holonomy of a loop: conjugated by $g(a)$; Wilson loop $W_R(\gamma) = \operatorname{tr}_R \operatorname{Hol}_\gamma$ is gauge invariant. Holonomy group; irreducible if it is all of G.
- Small loop: $\operatorname{Hol} = \mathbb{1} - \varepsilon\delta F_{12} + \ldots$; curvature is holonomy per unit area. Abelian Stokes $e^{-\oint A} = e^{-\int F}$; no simple non-abelian version (constant SU(2) field example).
- Aharonov–Bohm: flat connection, holonomy $e^{\mp ie\Phi}$; needs $\pi_1 \neq 0$. Sphere: rotation by enclosed area; Berry phase $-\Omega/2$ is the monopole holonomy.
- Lattice: $U_{x,\mu}$ on links, $S_W = \beta \sum_p (1 - \frac{1}{N}\operatorname{Re}\operatorname{tr} U_p) \to S_{\mathrm{YM}}$ with $\beta = 2N/g_{\mathrm{YM}}^2$; area law = confinement.

T3.10 Pure gauge, flat connections and the Chern–Simons form

Two questions close the chapter. The first: *when is a gauge field "nothing at all"?* Clearly $A = 0$ describes no field. But so does anything gauge equivalent to it, $A = g^{-1}\mathrm{d}g$, and such an A can look very complicated. Is $F = 0$ the test? The Aharonov–Bohm effect of Section T3.9.6 already warns us that the answer is "locally yes, globally no": a curvature-free field can have holonomy. We shall see that the gap between "flat" and "pure gauge" is measured exactly by the holonomy around non-contractible loops, i.e. by the topology of spacetime.

The second question: *the density* $\operatorname{tr} F \wedge F$ *is gauge invariant and closed; can we write it as* d *of something?* The answer is the **Chern–Simons form**

$$CS(A) = \operatorname{tr}\Big(A \wedge \mathrm{d}A + \tfrac{2}{3}\, A \wedge A \wedge A\Big), \qquad \mathrm{d}\, CS(A) = \operatorname{tr}(F \wedge F).$$

This one identity has three consequences that run through the rest of Nash. By Stokes' theorem the four-dimensional integral $\int \operatorname{tr} F \wedge F$ becomes a three-dimensional integral at infinity, where the field is pure gauge; there CS reduces to the winding-number density $\operatorname{tr}(g^{-1}\mathrm{d}g)^3$ of Section T3.6.5, so the instanton number is an integer. The form $CS(A)$ is itself a Lagrangian in three dimensions, the Chern–Simons action of Witten's knot theory (Chapter 12), whose critical points are flat connections. And the failure of CS to be gauge invariant, computed below, is the first step of the descent equations that produce anomalies (Chapter 10).

Throughout, $\theta = g^{-1}\mathrm{d}g$ is the Maurer–Cartan form of a map g into G, satisfying $\mathrm{d}\theta = -\theta \wedge \theta$ (T3.36). We write θ^2 for $\theta \wedge \theta$, A^3 for $A \wedge A \wedge A$, and so on: powers of matrix-valued forms always mean wedge products in order.

Graded cyclicity. We will use one trace rule over and over, so we state it once. If α is a matrix-valued p-form and β a matrix-valued q-form, then

$$\operatorname{tr}(\alpha \wedge \beta) = (-1)^{pq}\, \operatorname{tr}(\beta \wedge \alpha). \tag{T3.79}$$

It is ordinary cyclicity of the trace, $\operatorname{tr}(MN) = \operatorname{tr}(NM)$, combined with the sign $(-1)^{pq}$ from moving the q one-forms of β past the p one-forms of α. Three consequences we need:

- $\operatorname{tr}(\theta^2) = 0$ for any matrix-valued 1-form θ (take $p = q = 1$: $\operatorname{tr}(\theta \wedge \theta) = -\operatorname{tr}(\theta \wedge \theta)$). More generally $\operatorname{tr}(\alpha^{2m}) = 0$ for an odd form α; in particular $\operatorname{tr}(A^4) = 0$.

- A 1-form may be moved cyclically past a 2-form without a sign: $\operatorname{tr}(A \wedge \omega) = \operatorname{tr}(\omega \wedge A)$ for a 2-form ω.

- A product of three 1-forms is cyclic without a sign: $\operatorname{tr}(abc) = \operatorname{tr}(bca)$ (move the 1-form a past the 2-form bc).

T3.10.1 Pure gauge connections

Start from the trivial connection $A = 0$ and gauge transform it with an arbitrary $g(x)$. By the law $A \mapsto g^{-1}Ag + g^{-1}\mathrm{d}g$ (T3.52),

$$0 \;\longmapsto\; g^{-1}\,0\,g + g^{-1}\mathrm{d}g = \theta.$$

Working definition T3.23 (pure gauge)

A connection is **pure gauge** on a region U if there is a map $g : U \to G$ with

$$A = g^{-1}\mathrm{d}g \qquad \text{on } U. \tag{T3.80}$$

It is then gauge equivalent to $A = 0$ on U, and has no physical effect there [every gauge-invariant quantity on U, such as F or a Wilson loop inside U, takes its $A = 0$ value]. *Test:* try to solve $\mathrm{d}g = g\,A$ for g (a linear first-order PDE for the matrix g); pure gauge means a solution exists on all of U.

The field strength of a pure-gauge connection vanishes. This had better be so, because F transforms covariantly and $F(0) = 0$; but it is worth seeing the cancellation directly, because it is exactly the Maurer–Cartan structure equation:

$$\begin{aligned} F(g^{-1}\mathrm{d}g) &\overset{(1)}{=} \mathrm{d}\theta + \theta \wedge \theta \\ &\overset{(2)}{=} -g^{-1}\mathrm{d}g \wedge g^{-1}\mathrm{d}g + g^{-1}\mathrm{d}g \wedge g^{-1}\mathrm{d}g \\ &= 0. \end{aligned} \tag{T3.81}$$

(1) $F = \mathrm{d}A + A \wedge A$ (T3.56) with $A = \theta$.
(2) $\mathrm{d}\theta = \mathrm{d}(g^{-1}) \wedge \mathrm{d}g = -g^{-1}\mathrm{d}g\, g^{-1} \wedge \mathrm{d}g$, by $\mathrm{d}(g^{-1}) = -g^{-1}\mathrm{d}g\, g^{-1}$ (T3.35).

So the Maurer–Cartan equation $\mathrm{d}\theta + \theta^2 = 0$ *is* the statement "pure gauge fields are flat". Its content is not new; its interpretation is.

What does parallel transport do in a pure-gauge field? Transport for $A = 0$ is $W_\gamma(0) = \mathbb{1}$, and the gauge law (T3.74) gives

$$W_\gamma\big(g^{-1}\mathrm{d}g\big) = g(b)^{-1}\,\mathbb{1}\,g(a) = g(b)^{-1}g(a), \tag{T3.82}$$

for a path γ from a to b. The answer depends only on the end points, not on the path. For a closed loop, $b = a$, the holonomy is $g(a)^{-1}g(a) = \mathbb{1}$. So: *pure gauge* $\Rightarrow$ *path-independent transport* $\Rightarrow$ *trivial holonomy around every loop in* U.

Worked example T3.45: Checking (T3.82) by solving the ODE

Let $W(t) = g(\gamma(t))^{-1} g(a)$ and write $g_t = g(\gamma(t))$. Then

$$\frac{\mathrm{d}W}{\mathrm{d}t} \overset{(1)}{=} -g_t^{-1}\dot{g}_t\, g_t^{-1}\, g(a) \overset{(2)}{=} -\theta(\dot\gamma)\, W(t),$$

(1) Derivative of an inverse, $\frac{\mathrm{d}}{\mathrm{d}t} g_t^{-1} = -g_t^{-1}\dot{g}_t\, g_t^{-1}$.
(2) $g_t^{-1}\dot{g}_t = (g^{-1}\mathrm{d}g)(\dot\gamma) = \theta(\dot\gamma)$, and $g_t^{-1} g(a) = W(t)$.

and $W(0) = g(a)^{-1} g(a) = \mathbb{1}$. So W solves the transport equation (T3.68) for $A = \theta$, and by uniqueness it is the transport. Concretely, for U(1) with $g = e^{i\alpha(x)}$ we get $A = i\,\mathrm{d}\alpha$ and $W_\gamma = e^{-i(\alpha(b)-\alpha(a))}$, which is $e^{-\int_\gamma A}$ by the fundamental theorem of calculus.

Worked example T3.46: A pure-gauge SU(2) field that looks complicated

Take $g(x) = e^{\phi(x)T_3}\, e^{\chi(x)T_1}$ with $T_a = -\frac{i}{2}\sigma_a$ and two arbitrary functions ϕ, χ on $\mathbb{R}^2$. Then

$$\begin{aligned} A = g^{-1}\mathrm{d}g &\overset{(1)}{=} e^{-\chi T_1} e^{-\phi T_3}\left(\mathrm{d}\phi\, T_3\, e^{\phi T_3} e^{\chi T_1} + e^{\phi T_3}\,\mathrm{d}\chi\, T_1\, e^{\chi T_1}\right) \\ &\overset{(2)}{=} \mathrm{d}\phi\; e^{-\chi T_1} T_3\, e^{\chi T_1} + \mathrm{d}\chi\, T_1 \\ &\overset{(3)}{=} \mathrm{d}\phi\,\left(\cos\chi\, T_3 + \sin\chi\, T_2\right) + \mathrm{d}\chi\, T_1. \end{aligned}$$

(1) Product rule on g; $\mathrm{d}e^{\phi T_3} = \mathrm{d}\phi\, T_3 e^{\phi T_3}$ because T_3 commutes with its own exponential.
(2) $e^{-\phi T_3} T_3 e^{\phi T_3} = T_3$ and $e^{-\phi T_3} e^{\phi T_3} = \mathbb{1}$.
(3) $e^{-\chi T_1} T_3 e^{\chi T_1} = \mathrm{Ad}_{e^{-\chi T_1}} T_3$ is the adjoint action of $e^{-\chi T_1}$ on T_3 (Example T3.14), a rotation of the 2–3 plane by the angle χ; with $[T_1, T_3] = -T_2$ the Hadamard series (T3.22) gives $T_3 - \chi[T_1, T_3] + \cdots = \cos\chi\, T_3 + \sin\chi\, T_2$.

So $A^1 = \mathrm{d}\chi$, $A^2 = \sin\chi\,\mathrm{d}\phi$, $A^3 = \cos\chi\,\mathrm{d}\phi$. Check flatness in components. With $F^a = \mathrm{d}A^a + \frac{1}{2}\varepsilon_{abc} A^b \wedge A^c$ (the $\mathfrak{su}(2)$ form of $F = \mathrm{d}A + A \wedge A$):

$$\begin{aligned} F^1 &= 0 + A^2 \wedge A^3 = \sin\chi\cos\chi\,\mathrm{d}\phi \wedge \mathrm{d}\phi = 0, \\ F^2 &= \cos\chi\,\mathrm{d}\chi \wedge \mathrm{d}\phi + A^3 \wedge A^1 = \cos\chi\,\mathrm{d}\chi \wedge \mathrm{d}\phi + \cos\chi\,\mathrm{d}\phi \wedge \mathrm{d}\chi = 0, \\ F^3 &= -\sin\chi\,\mathrm{d}\chi \wedge \mathrm{d}\phi + A^1 \wedge A^2 = -\sin\chi\,\mathrm{d}\chi \wedge \mathrm{d}\phi + \sin\chi\,\mathrm{d}\chi \wedge \mathrm{d}\phi = 0. \end{aligned}$$

Each component of $\mathrm{d}A$ is non-zero but is cancelled exactly by the commutator term. A single component taken alone, say A^3, is *not* a flat abelian field ($\mathrm{d}A^3 \neq 0$): non-abelian flatness is a collective property of all the components together.

T3.10.2 Flat means locally pure gauge

We showed pure gauge ⇒ flat. Now the converse, which is the real content of this subsection.

Working definition T3.24 (flat connection)

A connection is **flat** if its curvature vanishes identically, $F = \mathrm{d}A + A \wedge A = 0$. *Test:* compute $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu + [A_\mu, A_\nu]$ for every pair $\mu < \nu$. *Physical meaning:* parallel transport around every small loop is trivial [it returns every vector unchanged, $W_\gamma = \mathbb{1}$], by (T3.76).

Key idea: flat = locally pure gauge

If $F = 0$ on a simply connected region U (every loop in U can be shrunk to a point inside U), then $A = g^{-1}\mathrm{d}g$ on U for some $g : U \to G$, unique up to $g \mapsto h\,g$ with h a constant. On a region that is *not* simply connected this can fail, and the failure is measured by the holonomies around the non-contractible loops.

Here is why, in three steps.

Step 1: the equation for g and its integrability condition. We want g with $g^{-1}\mathrm{d}g = A$, i.e. $\mathrm{d}g = g\,A$, i.e. $\partial_\mu g = g\,A_\mu$ for each μ. That is n first-order equations for one unknown matrix, so it is overdetermined: the mixed partials must agree, $\partial_\nu\partial_\mu g = \partial_\mu\partial_\nu g$. Compute:

$$\begin{aligned}\partial_\nu(\partial_\mu g) - \partial_\mu(\partial_\nu g) &\overset{(1)}{=} \partial_\nu(gA_\mu) - \partial_\mu(gA_\nu)\\ &\overset{(2)}{=} (\partial_\nu g)A_\mu + g\,\partial_\nu A_\mu - (\partial_\mu g)A_\nu - g\,\partial_\mu A_\nu\\ &\overset{(3)}{=} g\,A_\nu A_\mu + g\,\partial_\nu A_\mu - g\,A_\mu A_\nu - g\,\partial_\mu A_\nu\\ &\overset{(4)}{=} -g\left(\partial_\mu A_\nu - \partial_\nu A_\mu + [A_\mu, A_\nu]\right) = -g\,F_{\mu\nu}.\end{aligned}$$

(1) Insert the equation $\partial_\mu g = gA_\mu$.
(2) Product rule.
(3) Use the equation again for $\partial_\nu g$ and $\partial_\mu g$.
(4) Collect terms; the bracket is the component formula (T3.57).

So the equations are consistent only if $F_{\mu\nu} = 0$. In form language: if $\mathrm{d}g = gA$ then $0 = \mathrm{d}\mathrm{d}g = \mathrm{d}g \wedge A + g\,\mathrm{d}A = g(A \wedge A + \mathrm{d}A) = gF$. This is the non-abelian version of "a vector field with zero curl is locally a gradient".

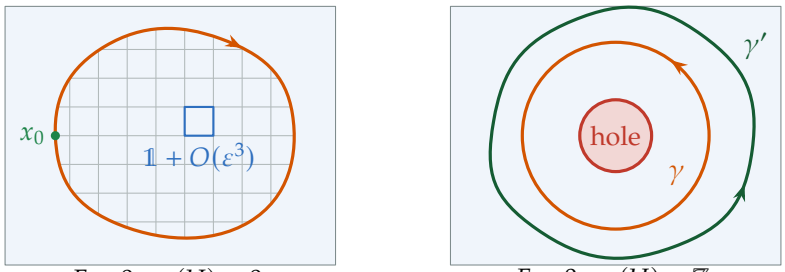
Step 2: construct g by parallel transport. Pick a base point $x_0 \in U$. For each $x \in U$ choose a path γ_x from x_0 to x and let $h(x) = W_{\gamma_x}$ be the parallel transport along it. *If* $h(x)$ does not depend on the choice of path, then extending the path by a small step εn at the end gives, by (T3.49) and (T3.73), $h(x + \varepsilon n) = (\mathbb{1} - \varepsilon\, n^\mu A_\mu)h(x) + O(\varepsilon^2)$, i.e. $\mathrm{d}h = -A\,h$. Then $g := h^{-1}$ does the job:

$$\mathrm{d}g \overset{(1)}{=} -h^{-1}\,\mathrm{d}h\,h^{-1} \overset{(2)}{=} -h^{-1}(-A\,h)h^{-1} = h^{-1}A = g\,A.$$

(1) Derivative of an inverse.
(2) $\mathrm{d}h = -Ah$.

Step 3: path independence is flatness plus simple connectivity. Two paths γ, γ' from x_0 to x give the same transport if and only if the loop $\gamma'^{-1} * \gamma$ has trivial holonomy (by (T3.73)). If U is simply connected, this loop can be shrunk to a point. Shrinking it sweeps out a surface, which we can tile by a fine grid of small plaquettes. The holonomy of the big loop is a product of conjugated plaquette holonomies (each plaquette is visited by a "lollipop": out along a tail, around, and back along the same tail, and the tails cancel in pairs), and each plaquette holonomy is $\mathbb{1} - \varepsilon^2 F_{12} + \cdots = \mathbb{1} + O(\varepsilon^3)$ when $F = 0$, by (T3.76). There are about $1/\varepsilon^2$ plaquettes, so the total error is $O(\varepsilon) \to 0$: the holonomy is exactly $\mathbb{1}$. (Figure T3.31 shows the picture.)

The same plaquette argument, applied to the strip between two loops that can be deformed into one another, gives a result we will use constantly.



$F = 0$, $\pi_1(U) = 0$: every holonomy $= \mathbb{1}$

$F = 0$, $\pi_1(U) = \mathbb{Z}$: $\mathrm{Hol}_\gamma = \mathrm{Hol}_{\gamma'}$ may be $\neq \mathbb{1}$

Figure T3.31: Left: in a flat connection on a simply connected region, a loop can be filled in by small plaquettes, each with holonomy $\mathbb{1} + O(\varepsilon^3)$; so the loop's holonomy is $\mathbb{1}$ and transport is path independent. Right: around a hole, the orange and green loops, both running counterclockwise, can be deformed into each other (so they have the same holonomy, up to conjugation when G is non-abelian), but neither can be filled in. Their common holonomy is the only information a flat connection carries.

Key idea: holonomy of a flat connection is topological

For a flat connection, the holonomy around a loop based at x_0 depends only on the *homotopy class* of the loop. It therefore defines a map

$$\rho_A : \pi_1(M, x_0) \longrightarrow G, \qquad [\gamma] \longmapsto \mathrm{Hol}_\gamma(A),$$

which is a group homomorphism (a **representation of the fundamental group**) because holonomies compose, (T3.73). A gauge transformation conjugates it, $\rho_{A^g} = g(x_0)^{-1}\rho_A g(x_0)$. Conversely every homomorphism $\pi_1(M) \to G$ arises from some flat connection (possibly on a non-trivial bundle). So

$$\{\text{flat connections on } M\}/\{\text{gauge}\} \;\cong\; \mathrm{Hom}\big(\pi_1(M), G\big)/\text{conjugation}. \tag{T3.83}$$

A small bookkeeping remark about homomorphism order: with the convention "first path on the right" of (T3.73), the holonomy of "γ_1 then γ_2" is $\mathrm{Hol}_{\gamma_2}\mathrm{Hol}_{\gamma_1}$, so ρ_A respects products if we also write loop composition right to left. Nothing below depends on this.

Watch out: flat is not trivial

On $\mathbb{R}^n$, or on any simply connected M such as S^2, S^3 or S^4, (T3.83) says that the only flat connection up to gauge is $A = 0$: flat $\Leftrightarrow$ pure gauge. On a circle, a torus, a punctured plane, a knot complement or a lens space, there are many inequivalent flat connections. They carry no field strength and no energy, yet they are physically distinct: they are detected by Wilson loops, as in Aharonov–Bohm.

Worked example T3.47: Flat connections on a circle

On $M = S^1$ with coordinate $\phi \in [0, 2\pi)$ there are no 2-forms, so every connection is flat. Any connection can be gauge transformed to a constant one, $A = X\,\mathrm{d}\phi$ with $X \in \mathfrak{g}$ constant (transport around the circle gives a group element W; write $W = e^{-2\pi X}$ and gauge away the rest). Its holonomy is found by solving $\mathrm{d}W/\mathrm{d}\phi = -XW$:

$$\mathrm{Hol}(A) = e^{-2\pi X}.$$

U(1): $X = ia$ with $a \in \mathbb{R}$; the holonomy is $e^{-2\pi i a}$. The gauge transformation $g = e^{in\phi}$ ($n \in \mathbb{Z}$,

single valued) gives $g^{-1}\mathrm{d}g = in\,\mathrm{d}\phi$, so

$$A^g \overset{(1)}{=} g^{-1}(ia\,\mathrm{d}\phi)g + g^{-1}\mathrm{d}g \overset{(2)}{=} i(a+n)\,\mathrm{d}\phi.$$

(1) Gauge law (T3.52).
(2) U(1) is abelian so $g^{-1}(\cdot)g$ does nothing; $g^{-1}\mathrm{d}g = e^{-in\phi}\,in\,e^{in\phi}\,\mathrm{d}\phi$.

So a and $a+n$ are gauge equivalent: the space of flat U(1) connections on S^1 is $\mathbb{R}/\mathbb{Z}$, a circle, labelled faithfully by the holonomy $e^{-2\pi ia} \in \mathrm{U}(1)$. The transformations $e^{in\phi}$ with $n \neq 0$ cannot be deformed to the identity (they wind n times, $\pi_1(\mathrm{U}(1)) = \mathbb{Z}$); they are the simplest **large gauge transformations**.
SU(2): conjugate X into the Cartan subalgebra, $X = ia\sigma_3$, holonomy $\mathrm{diag}(e^{-2\pi ia}, e^{2\pi ia})$. Large gauge transformations $e^{in\phi\sigma_3}$ shift $a \mapsto a+n$, and the Weyl reflection (conjugation by $i\sigma_2$) sends $a \mapsto -a$. A fundamental domain is $a \in [0, \frac{1}{2}]$: the flat SU(2) connections on S^1 form an *interval*, the space of conjugacy classes of SU(2) (the latitudes of Figure T3.3). At the endpoints the holonomy is central ($\pm\mathbb{1}$) and the stabiliser is all of SU(2); in the interior it is the U(1) of diagonal matrices (compare Example T3.40).

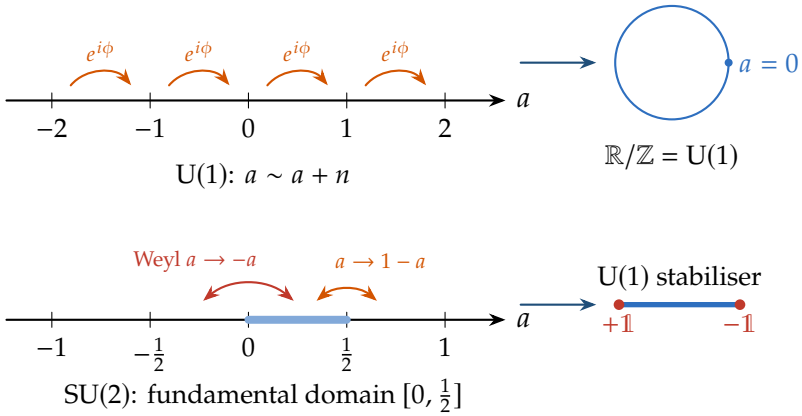


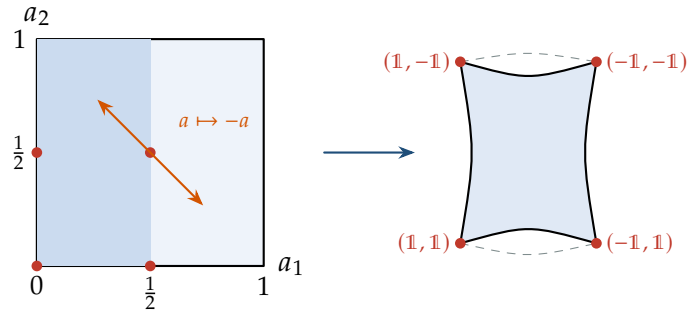
Figure T3.32: Flat connections on a circle, $A = ia\sigma\,\mathrm{d}\phi$. Top: for U(1) the large gauge transformations $e^{in\phi}$ shift a by integers, and the moduli space is a circle. Bottom: for SU(2) we also quotient by the Weyl reflection; the moduli space is the interval $[0, \frac{1}{2}]$, whose endpoints (holonomy $\pm\mathbb{1}$, red) have the largest stabiliser.

Worked example T3.48: Flat connections on a torus: the pillowcase

On T^2 with angles (ϕ_1, ϕ_2), $\pi_1(T^2) = \mathbb{Z}^2$ is generated by the two cycles γ_1, γ_2, with the single relation $\gamma_1\gamma_2\gamma_1^{-1}\gamma_2^{-1} = 1$ (the boundary of the square in the usual picture of the torus). So a flat connection is a pair of *commuting* holonomies (U_1, U_2), up to simultaneous conjugation.

- U(1): $A = i(a_1\mathrm{d}\phi_1 + a_2\mathrm{d}\phi_2)$ with constants a_1, a_2; $F = \mathrm{d}A = 0$. Large gauge transformations $e^{i(n_1\phi_1 + n_2\phi_2)}$ shift (a_1, a_2) by integers. The moduli space is the torus $\mathbb{R}^2/\mathbb{Z}^2$ (the "dual torus", or Jacobian, of Chapter 5).
- SU(2): two commuting elements of SU(2) can be simultaneously diagonalised, so $A = i(a_1\mathrm{d}\phi_1 + a_2\mathrm{d}\phi_2)\sigma_3$, again with $(a_1, a_2) \in \mathbb{R}^2/\mathbb{Z}^2$, but now also identified under the Weyl reflection $(a_1, a_2) \mapsto (-a_1, -a_2)$. The quotient of a torus by this reflection is a sphere with four corner points, the **pillowcase** (Figure T3.33). The corners are the four pairs $(\pm\mathbb{1}, \pm\mathbb{1})$ of central holonomies.

Check that $F = 0$ in the SU(2) case: $\mathrm{d}A = 0$ (constant coefficients) and $A \wedge A = -(a_1 \mathrm{d}\phi_1 + a_2 \mathrm{d}\phi_2)^2 \sigma_3^2 = 0$ since the scalar 1-form squares to zero. The flatness is "abelian" because the components commute.



$T^2 = \mathbb{R}^2/\mathbb{Z}^2$, shaded half is a fundamental domain

pillowcase: flat SU(2) connections on T^2

Figure T3.33: The moduli space of flat SU(2) connections on a torus. Commuting holonomies are conjugated to $(e^{-2\pi i a_1 \sigma_3}, e^{-2\pi i a_2 \sigma_3})$ with (a_1, a_2) on a torus (left). The Weyl reflection $a \mapsto -a$ folds the torus in half and sews the edges, giving a sphere with four singular corners (right). The corners, marked in red, are the four flat connections with central holonomy, where the stabiliser jumps to all of SU(2).

Try it: Flat connections on $\mathbb{RP}^3$ and on S^3

(a) $\pi_1(S^3) = 0$. What are the flat SU(2) connections on S^3? (b) $\pi_1(\mathbb{RP}^3) = \mathbb{Z}_2$. Classify flat SU(2) connections on $\mathbb{RP}^3$.
Answer: (a) Only $A = 0$ up to gauge: the trivial representation. (b) A homomorphism $\mathbb{Z}_2 \to \mathrm{SU}(2)$ sends the generator to U with $U^2 = \mathbb{1}$. In SU(2) the only such elements are $\pm\mathbb{1}$ (an element $\cos\alpha + i \sin\alpha\, \hat{n} \cdot \vec{\sigma}$ squares to $\mathbb{1}$ only for $\alpha \in \{0, \pi\}$). So there are exactly two flat connections, both with central holonomy. The non-trivial one is the "$-\mathbb{1}$ around the non-contractible loop" connection, the gauge-theory cousin of the spin-$\frac{1}{2}$ sign of Section T3.4.5.

Reading Nash (flat connections in Nash)

Nash meets flat connections three times. In the monopole chapter, energy $E = \frac{1}{2}(\|F\|^2 + \|d_A\phi\|^2) = 0$ forces $F = 0$, and he remarks that the connection's "only departure from triviality lies in its holonomy, which relies on $\pi_1(M)$ to be non-vanishing" [Nash p. 244, (8.150)–(8.151)]: that is our (T3.83). In Floer theory the critical points of the Chern–Simons function are the flat connections, identified with $\mathrm{Hom}(\pi_1(M), \mathrm{SU}(2))/\mathrm{Ad}\,\mathrm{SU}(2)$ [Nash p. 325, (12.8)–(12.10)]; on a homology 3-sphere they are automatically irreducible, because a reducible one would give an abelian representation of π_1, which factors through $H_1(M;\mathbb{Z}) = 0$. And in Chern–Simons theory the semiclassical expansion of the partition function is a sum over flat connections [Nash pp. 346–348].

T3.10.3 The Chern–Simons form: $\mathrm{tr}\, F \wedge F = \mathrm{d}\, CS$

Consider the 4-form $\mathrm{tr}(F \wedge F)$. It has two properties that make it special.

- *Gauge invariant:* $\mathrm{tr}(g^{-1}Fg \wedge g^{-1}Fg) = \mathrm{tr}(g^{-1}F \wedge F\, g) = \mathrm{tr}(F \wedge F)$.

- *Closed:* $\mathrm{d}\,\mathrm{tr}(F\wedge F) = \mathrm{tr}(\mathrm{d}F\wedge F)+\mathrm{tr}(F\wedge \mathrm{d}F) = 2\,\mathrm{tr}(\mathrm{d}F\wedge F)$, and by Bianchi $\mathrm{d}F = F\wedge A - A\wedge F$, so $\mathrm{d}\,\mathrm{tr}(F^2) = 2\,\mathrm{tr}(FAF - AF^2) = 0$, since $\mathrm{tr}(FAF) = \mathrm{tr}(AF^2)$ by (T3.79) (move the 2-form F at the front to the back, past the 3-form AF: sign $(-1)^{2\cdot 3} = +1$).

Closed forms are locally exact (Poincaré lemma, Chapter T2). So there should be a 3-form whose derivative is $\mathrm{tr}\,F\wedge F$. In the abelian case it is easy to guess, since $F = \mathrm{d}A$:

$$F\wedge F = \mathrm{d}A\wedge \mathrm{d}A = \mathrm{d}(A\wedge \mathrm{d}A).$$

For non-abelian A we need a correction cubic in A.

Key idea: the Chern–Simons 3-form

For any $\mathfrak{g}$-valued 1-form A with curvature $F = \mathrm{d}A + A\wedge A$,

$$CS(A) := \mathrm{tr}\Big(A\wedge \mathrm{d}A + \tfrac{2}{3}\,A\wedge A\wedge A\Big) \qquad \text{satisfies} \qquad \mathrm{d}\,CS(A) = \mathrm{tr}(F\wedge F). \tag{T3.84}$$

Equivalently $CS(A) = \mathrm{tr}\big(A\wedge F - \tfrac{1}{3}A\wedge A\wedge A\big)$.

Derivation, line by line. First expand the right-hand side:

$$\begin{aligned}\mathrm{tr}(F\wedge F) &\overset{(1)}{=} \mathrm{tr}\big((\mathrm{d}A + A^2)\wedge(\mathrm{d}A + A^2)\big)\\ &\overset{(2)}{=} \mathrm{tr}(\mathrm{d}A\,\mathrm{d}A) + \mathrm{tr}(\mathrm{d}A\,A^2) + \mathrm{tr}(A^2\,\mathrm{d}A) + \mathrm{tr}(A^4)\\ &\overset{(3)}{=} \mathrm{tr}(\mathrm{d}A\,\mathrm{d}A) + 2\,\mathrm{tr}(\mathrm{d}A\,A^2).\end{aligned} \tag{T3.85}$$

(1) Definition of F.
(2) Multiply out, keeping the order of the factors.
(3) $\mathrm{tr}(A^2\mathrm{d}A) = \mathrm{tr}(\mathrm{d}A\,A^2)$ by (T3.79) with $p = q = 2$; and $\mathrm{tr}(A^4) = 0$ since $A^4 = A\wedge A^3$ and $\mathrm{tr}(A\wedge A^3) = (-1)^{1\cdot 3}\,\mathrm{tr}(A^3\wedge A) = -\,\mathrm{tr}(A^4)$.

Now differentiate the two terms of CS separately:

$$\mathrm{d}\,\mathrm{tr}(A\,\mathrm{d}A) \overset{(1)}{=} \mathrm{tr}(\mathrm{d}A\,\mathrm{d}A) - \mathrm{tr}(A\,\mathrm{d}\mathrm{d}A) \overset{(2)}{=} \mathrm{tr}(\mathrm{d}A\,\mathrm{d}A), \tag{T3.86}$$

$$\begin{aligned}\mathrm{d}\,\mathrm{tr}(A^3) &\overset{(3)}{=} \mathrm{tr}(\mathrm{d}A\,A^2) - \mathrm{tr}(A\,\mathrm{d}A\,A) + \mathrm{tr}(A^2\,\mathrm{d}A)\\ &\overset{(4)}{=} \mathrm{tr}(\mathrm{d}A\,A^2) + \mathrm{tr}(\mathrm{d}A\,A^2) + \mathrm{tr}(\mathrm{d}A\,A^2) = 3\,\mathrm{tr}(\mathrm{d}A\,A^2).\end{aligned} \tag{T3.87}$$

(1) Leibniz rule $\mathrm{d}(\alpha\wedge\beta) = \mathrm{d}\alpha\wedge\beta + (-1)^p\alpha\wedge\mathrm{d}\beta$ with $\alpha = A$ a 1-form; d commutes with the trace.
(2) $\mathrm{dd} = 0$.
(3) Leibniz on the triple product: the sign alternates each time d passes a 1-form.
(4) Middle term: $\mathrm{tr}(A\wedge(\mathrm{d}A\,A)) = (-1)^{1\cdot 3}\,\mathrm{tr}(\mathrm{d}A\,A\wedge A)$, so $-\,\mathrm{tr}(A\,\mathrm{d}A\,A) = +\,\mathrm{tr}(\mathrm{d}A\,A^2)$. Last term: move the 2-form A^2 past the 2-form $\mathrm{d}A$, no sign.

Putting them together,

$$\mathrm{d}\,CS(A) \overset{(1)}{=} \mathrm{tr}(\mathrm{d}A\,\mathrm{d}A) + \tfrac{2}{3}\cdot 3\,\mathrm{tr}(\mathrm{d}A\,A^2) \overset{(2)}{=} \mathrm{tr}(F\wedge F).$$

(1) (T3.86) and (T3.87).
(2) Compare with (T3.85).

The coefficient $\frac{2}{3}$ is forced: it is the only number that makes the $\mathrm{tr}(\mathrm{d}A\,A^2)$ terms match. The alternative form follows from $\mathrm{tr}(A\,\mathrm{d}A) = \mathrm{tr}(A(F - A^2)) = \mathrm{tr}(AF) - \mathrm{tr}(A^3)$, so $CS = \mathrm{tr}(AF) - \mathrm{tr}\,A^3 + \frac{2}{3}\,\mathrm{tr}\,A^3 = \mathrm{tr}(AF) - \frac{1}{3}\,\mathrm{tr}\,A^3$.

In components. On a 3-manifold with coordinates x^1, x^2, x^3, write $A = A_\mu \mathrm{d}x^\mu$ and use $\mathrm{d}x^\mu \wedge \mathrm{d}x^\nu \wedge \mathrm{d}x^\rho = \varepsilon^{\mu\nu\rho}\,\mathrm{d}^3x$:

$$CS(A) \overset{(1)}{=} \operatorname{tr}\big(A_\mu \partial_\nu A_\rho + \tfrac{2}{3} A_\mu A_\nu A_\rho\big)\,\mathrm{d}x^\mu \wedge \mathrm{d}x^\nu \wedge \mathrm{d}x^\rho$$
$$\overset{(2)}{=} \varepsilon^{\mu\nu\rho}\,\operatorname{tr}\Big(A_\mu \partial_\nu A_\rho + \tfrac{2}{3} A_\mu A_\nu A_\rho\Big)\,\mathrm{d}^3x. \tag{T3.88}$$

(1) $\mathrm{d}A = \partial_\nu A_\rho\,\mathrm{d}x^\nu \wedge \mathrm{d}x^\rho$; the matrix coefficients stay in order.
(2) The wedge of three coordinate differentials is $\varepsilon^{\mu\nu\rho}$ times the coordinate volume form.

This is the form found in the physics literature, e.g. as the Lagrangian of 2+1-dimensional Chern–Simons theory. Note what is *missing*: there is no metric anywhere, no $\star$, no $g_{\mu\nu}$. This is why Chern–Simons theory is topological.

Worked example T3.49: CS for SU(2) in components

Write $A = A^a T_a$ with $T_a = -\frac{i}{2}\sigma_a$, so $\operatorname{tr}(T_a T_b) = -\frac{1}{2}\delta_{ab}$ and $\operatorname{tr}(T_a[T_b, T_c]) = -\frac{1}{2}\varepsilon_{abc}$ (computed before (T3.42)). Then

$$\operatorname{tr}(A \wedge \mathrm{d}A) \overset{(1)}{=} A^a \wedge \mathrm{d}A^b\ \operatorname{tr}(T_a T_b) = -\tfrac{1}{2}\,A^a \wedge \mathrm{d}A^a,$$
$$\operatorname{tr}(A \wedge A \wedge A) \overset{(2)}{=} -\tfrac{1}{4}\,\varepsilon_{abc}\,A^a \wedge A^b \wedge A^c = -\tfrac{3}{2}\,A^1 \wedge A^2 \wedge A^3,$$
$$CS(A) \overset{(3)}{=} -\tfrac{1}{2}\Big(A^a \wedge \mathrm{d}A^a + \tfrac{1}{3}\varepsilon_{abc}\,A^a \wedge A^b \wedge A^c\Big) = -\tfrac{1}{2}A^a \wedge \mathrm{d}A^a - A^1 \wedge A^2 \wedge A^3.$$

(1) Scalar 1-forms come out of the trace.
(2) Exactly the computation (T3.42) with θ replaced by A.
(3) $\frac{2}{3}\cdot(-\frac{1}{4}) = -\frac{1}{6} = -\frac{1}{2}\cdot\frac{1}{3}$; and $\varepsilon_{abc}A^aA^bA^c = 6A^1A^2A^3$.

In physicists' Hermitian variables ($A^a = -g_{\mathrm{YM}}\mathcal{A}^a$, see Section T3.8.9) this is, up to normalisation, $\mathcal{A}^a \wedge \mathrm{d}\mathcal{A}^a + \frac{g}{3}\varepsilon_{abc}\mathcal{A}^a\mathcal{A}^b\mathcal{A}^c$, the form in which it is usually quoted.

Worked example T3.50: CS of a pure-gauge field is the winding density

Put $A = \theta = g^{-1}\mathrm{d}g$. Then $F = 0$ by (T3.81), and

$$CS(\theta) \overset{(1)}{=} \operatorname{tr}\big(\theta \wedge F - \tfrac{1}{3}\theta^3\big) \overset{(2)}{=} -\tfrac{1}{3}\operatorname{tr}\big(g^{-1}\mathrm{d}g\big)^3. \tag{T3.89}$$

(1) The alternative form of (T3.84).
(2) $F(\theta) = 0$.

So on pure-gauge configurations the Chern–Simons form *is* the winding-number density of (T3.44), up to a constant. Integrating over a 3-sphere and using $\int \operatorname{tr}(g^{-1}\mathrm{d}g)^3 = -24\pi^2 w(g)$:

$$\frac{1}{8\pi^2}\int_{S^3} CS\big(g^{-1}\mathrm{d}g\big) = \frac{1}{8\pi^2}\cdot\Big(-\frac{1}{3}\Big)\cdot\big(-24\pi^2\,w(g)\big) = w(g). \tag{T3.90}$$

For the identity map of SU(2) (the Maurer–Cartan form on the group itself), $\frac{1}{8\pi^2}\int CS = 1$. The normalisation $\frac{1}{8\pi^2}$ is the one that makes Chern–Simons integrals integers on pure-gauge fields.

T3.10.4 Transgression: where the Chern–Simons form comes from

Guessing the $\frac{2}{3}$ worked, but it does not generalise. Nash uses a systematic construction, **transgression**, that produces a Chern–Simons form for any invariant polynomial [Nash p. 224,

(8.41)–(8.42)]. The idea is to switch the connection on gradually.

Let $A_t = tA$ for $t \in [0,1]$, interpolating between $A_0 = 0$ and $A_1 = A$. Its curvature is

$$F_t = \mathrm{d}(tA) + tA \wedge tA = t\,\mathrm{d}A + t^2 A \wedge A, \tag{T3.91}$$

Nash's (8.41). At $t = 0$, $F_0 = 0$; at $t = 1$, $F_1 = F$. So

$$\operatorname{tr}(F \wedge F) = \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t} \operatorname{tr}(F_t \wedge F_t)\,\mathrm{d}t.$$

We compute the integrand. First, the t-derivative of F_t is the covariant derivative of A with respect to the connection A_t:

$$\frac{\mathrm{d}F_t}{\mathrm{d}t} \overset{(1)}{=} \mathrm{d}A + 2t\,A \wedge A \overset{(2)}{=} \mathrm{d}A + A_t \wedge A + A \wedge A_t \overset{(3)}{=} D_t A, \tag{T3.92}$$

(1) Differentiate (T3.91).
(2) $A_t = tA$, so $A_t \wedge A + A \wedge A_t = 2tA \wedge A$.
(3) $D_t\omega = \mathrm{d}\omega + A_t \wedge \omega - (-1)^p \omega \wedge A_t$ is the covariant derivative (T3.60) built from A_t; for the 1-form $\omega = A$ the sign is $+$.

Second, the trace of a covariant derivative against F_t is an ordinary derivative, because of Bianchi. For any $\mathfrak{g}$-valued 1-form a and any connection B with curvature F_B,

$$\begin{aligned}
\mathrm{d}\operatorname{tr}(a \wedge F_B) &\overset{(1)}{=} \operatorname{tr}(\mathrm{d}a \wedge F_B) - \operatorname{tr}(a \wedge \mathrm{d}F_B) \\
&\overset{(2)}{=} \operatorname{tr}(\mathrm{d}a \wedge F_B) - \operatorname{tr}(a \wedge F_B \wedge B) + \operatorname{tr}(a \wedge B \wedge F_B) \\
&\overset{(3)}{=} \operatorname{tr}(\mathrm{d}a \wedge F_B) + \operatorname{tr}(B \wedge a \wedge F_B) + \operatorname{tr}(a \wedge B \wedge F_B) \\
&\overset{(4)}{=} \operatorname{tr}\big(D_B a \wedge F_B\big).
\end{aligned} \tag{T3.93}$$

(1) Leibniz rule; a is a 1-form.
(2) Bianchi identity $\mathrm{d}F_B = F_B \wedge B - B \wedge F_B$ (T3.61).
(3) Move the 1-form B from the back to the front past the 3-form $a \wedge F_B$: sign $(-1)^3 = -1$.
(4) $D_B a = \mathrm{d}a + B \wedge a + a \wedge B$.

Now put the pieces together:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} \operatorname{tr}(F_t \wedge F_t) &\overset{(1)}{=} 2 \operatorname{tr}\Big(\frac{\mathrm{d}F_t}{\mathrm{d}t} \wedge F_t\Big) \overset{(2)}{=} 2 \operatorname{tr}\big(D_t A \wedge F_t\big) \overset{(3)}{=} 2\,\mathrm{d}\operatorname{tr}\big(A \wedge F_t\big), \\
\operatorname{tr}(F \wedge F) &\overset{(4)}{=} \mathrm{d}\Big[\,2\int_0^1 \operatorname{tr}(A \wedge F_t)\,\mathrm{d}t\,\Big].
\end{aligned} \tag{T3.94}$$

(1) Product rule; the two terms are equal because 2-forms commute inside the trace, (T3.79) with $p = q = 2$.
(2) (T3.92).
(3) (T3.93) with $a = A$, $B = A_t$.
(4) Integrate over t from 0 to 1 (with $F_0 = 0$) and pull d, which acts on x only, outside the t-integral.

The bracket is the Chern–Simons form. Evaluating the t-integral:

$$2\int_0^1 \operatorname{tr}(A \wedge F_t)\,\mathrm{d}t \overset{(1)}{=} 2\int_0^1 \big(t\,\operatorname{tr}(A\,\mathrm{d}A) + t^2 \operatorname{tr}(A^3)\big)\mathrm{d}t \overset{(2)}{=} \operatorname{tr}(A\,\mathrm{d}A) + \tfrac{2}{3}\operatorname{tr}(A^3) = CS(A).$$

(1) (T3.91).
(2) $\int_0^1 2t\,\mathrm{d}t = 1$, $\int_0^1 2t^2\,\mathrm{d}t = \frac{2}{3}$.

The mysterious $\frac{2}{3}$ is $\int_0^1 2t^2\,\mathrm{d}t$.

Working definition T3.25 (transgression formula)

For an invariant polynomial P_j of degree j (symmetric, multilinear in j matrix arguments, and invariant under simultaneous conjugation [replacing every argument X by gXg^{-1} with the same g], such as $P_2(X,Y) = \mathrm{tr}(XY)$ or $P_3(X,Y,Z) = \frac{1}{2}\,\mathrm{tr}(X\{Y,Z\})$) set

$$TP_j(A) = j\int_0^1 P_j\big(A, F_t, \dots, F_t\big)\,\mathrm{d}t, \qquad F_t = t\,\mathrm{d}A + t^2 A\wedge A \tag{T3.95}$$

[F_t is the curvature of the scaled connection tA]. Then $\mathrm{d}\,TP_j(A) = P_j(F,\dots,F)$ [Nash (8.42)]. *How to use it:* substitute F_t, expand in powers of t, integrate each power. TP_2 with $P_2 = \mathrm{tr}$ is $CS(A)$.

The proof is the one just given: $\frac{\mathrm{d}}{\mathrm{d}t}P_j(F_t,\dots,F_t) = j\,P_j(D_tA, F_t,\dots,F_t)$ by symmetry, and invariance plus Bianchi turn $P_j(D_tA,F_t,\dots)$ into $\mathrm{d}P_j(A,F_t,\dots)$.

Worked example T3.51: The five-dimensional Chern–Simons form

In six dimensions the relevant invariant is $\mathrm{tr}(F\wedge F\wedge F)$; its Chern–Simons form governs anomalies of four-dimensional chiral fermions (Chapter 10). By (T3.95) with $j=3$:

$$\begin{aligned} CS_5(A) &\overset{(1)}{=} 3\int_0^1 \mathrm{tr}\big(A\wedge F_t\wedge F_t\big)\mathrm{d}t \\ &\overset{(2)}{=} 3\int_0^1 \Big[t^2\,\mathrm{tr}\big(A(\mathrm{d}A)^2\big) + t^3\,\mathrm{tr}\big(A\,\mathrm{d}A\,A^2 + A^3\,\mathrm{d}A\big) + t^4\,\mathrm{tr}(A^5)\Big]\mathrm{d}t \\ &\overset{(3)}{=} \mathrm{tr}\big(A(\mathrm{d}A)^2\big) + \tfrac{3}{4}\cdot 2\,\mathrm{tr}\big(A^3\,\mathrm{d}A\big) + \tfrac{3}{5}\,\mathrm{tr}(A^5) \\ &= \mathrm{tr}\Big(A(\mathrm{d}A)^2 + \tfrac{3}{2}A^3\,\mathrm{d}A + \tfrac{3}{5}A^5\Big). \end{aligned}$$

(1) Transgression with $P_3(X,Y,Z)$ symmetrised; for $\mathrm{tr}(AF_tF_t)$ the symmetrisation is automatic because the two F_t are equal.
(2) $F_t^2 = t^2(\mathrm{d}A)^2 + t^3(\mathrm{d}A\,A^2 + A^2\,\mathrm{d}A) + t^4A^4$.
(3) $\int_0^1 3t^2 = 1$, $\int_0^1 3t^3 = \frac{3}{4}$, $\int_0^1 3t^4 = \frac{3}{5}$; and $\mathrm{tr}(A\,\mathrm{d}A\,A^2) = \mathrm{tr}(A^2\,A\,\mathrm{d}A) = \mathrm{tr}(A^3\mathrm{d}A)$ by moving the 2-form A^2 from the back to the front past a 4-form (no sign).

So $\mathrm{d}\,CS_5 = \mathrm{tr}\,F^3$. Checking this directly takes a page of algebra; the transgression formula does it for us.

Chern–Weil: characteristic classes do not depend on the connection. The same computation with $A_t = A_0 + t(A_1 - A_0)$ interpolating between two arbitrary connections (each A_t is again a connection, because the difference $A_1 - A_0$ transforms homogeneously, $a \mapsto g^{-1}ag$: the inhomogeneous terms $g^{-1}\mathrm{d}g$ cancel) gives, with $a = A_1 - A_0$ and $\frac{\mathrm{d}F_t}{\mathrm{d}t} = D_t a$,

$$\mathrm{tr}(F_1\wedge F_1) - \mathrm{tr}(F_0\wedge F_0) = \mathrm{d}\Big[2\int_0^1 \mathrm{tr}\big(a\wedge F_t\big)\mathrm{d}t\Big]. \tag{T3.96}$$

The difference of the two 4-forms is exact. On a closed 4-manifold its integral vanishes by Stokes, so $\int_M \operatorname{tr} F\wedge F$ is the same for *every* connection on a given bundle: it is a topological invariant of the bundle, a characteristic number (Chapter 1). The Chern–Simons form is the special case $A_0 = 0$, which is available only where the bundle is trivial, i.e. locally, or on all of $\mathbb{R}^4$.

Reading Nash (secondary characteristic classes, pp. 223–224)

Nash's §8.2 is this subsection in general form. He writes a characteristic class as an invariant polynomial, $c_j(F) = P_j(F,\dots,F)$ [Nash (8.40)]; he pulls it back to the total space P of the bundle, where it becomes exact, $\pi^* c_j(F) = \mathrm{d}\,TP_j(A)$, with TP_j given by (T3.95) using $A_t = tA$, $F_t = t\,\mathrm{d}A + t^2 A\wedge A$ [Nash (8.41)–(8.42)]. (Working on P is the mathematicians' way of saying "in a local trivialisation, where $A_0 = 0$ makes sense".) If $c_j(F)$ happens to vanish as a form, then $\mathrm{d}\,TP_j(A) = 0$ and TP_j defines a cohomology class of odd degree, a **secondary characteristic class** [Nash (8.43)–(8.45)]; it is only defined modulo integers on the base, because changing gauge shifts it by an integral class, as we compute next. With $P_2(F) = -\frac{1}{8\pi^2}\operatorname{tr}(F\wedge F)$ he obtains $TP_2(A) = -\frac{1}{8\pi^2}\operatorname{tr}(A\wedge \mathrm{d}A + \frac{2}{3}A\wedge A\wedge A)$ [Nash (8.61)], which is $-\frac{1}{8\pi^2}CS(A)$ in our notation. The same transgression with $j=3$, pulled back to the gauge group, gives his formula for the generator μ_1 of $H^1(\mathcal{G})$ [Nash (8.56)–(8.58)], which is the anomaly (Chapter 10).

Watch out: the sign of P_2

With the Chern class convention $c(F) = \det\!\left(\mathbb{1} + \frac{i}{2\pi}F\right)$ of this book, $c_2(F) = \frac{1}{8\pi^2}\left[\operatorname{tr}(F\wedge F) - \operatorname{tr}F\wedge \operatorname{tr}F\right]$, which for $\mathfrak{su}(N)$ ($\operatorname{tr}F = 0$) is $+\frac{1}{8\pi^2}\operatorname{tr}(F\wedge F)$. Nash's (8.61) writes $P_2(F) = -\frac{1}{8\pi^2}\operatorname{tr}(F\wedge F)$ for $\mathrm{SU}(N)$. The overall sign is a convention (it flips with the orientation, or with the sign of i in $c(F)$) and does not affect anything in this section except the sign of the integer. The structure $\operatorname{tr}(A\,\mathrm{d}A + \frac{2}{3}A^3)$ is the same in every convention.

T3.10.5 How the Chern–Simons form changes under a gauge transformation

$\operatorname{tr}F\wedge F$ is gauge invariant, but $CS(A)$ is built from A itself and is not. Since $\mathrm{d}\,CS(A^g) = \operatorname{tr}F^g\wedge F^g = \operatorname{tr}F\wedge F = \mathrm{d}\,CS(A)$, the difference $CS(A^g) - CS(A)$ is a *closed* 3-form. We now compute it exactly. Write

$$A^g = \tilde{A} + \theta, \qquad \tilde{A} := g^{-1}Ag, \qquad \theta := g^{-1}\mathrm{d}g, \qquad \rho := \mathrm{d}g\,g^{-1},$$

so that $F^g = g^{-1}Fg =: \tilde{F}$, and note two facts about the right-invariant form ρ: it is θ conjugated, $\theta = g^{-1}\rho\,g$, and it obeys the structure equation with the opposite sign,

$$\mathrm{d}\rho = -\mathrm{d}g\wedge \mathrm{d}(g^{-1}) = \mathrm{d}g\wedge g^{-1}\mathrm{d}g\,g^{-1} = \rho\wedge\rho. \tag{T3.97}$$

Then, using the alternative form $CS = \operatorname{tr}(AF) - \frac{1}{3}\operatorname{tr}A^3$,

$$\begin{aligned}
CS(A^g) &\overset{(1)}{=} \operatorname{tr}\big((\tilde{A}+\theta)\tilde{F}\big) - \tfrac{1}{3}\operatorname{tr}(\tilde{A}+\theta)^3 \\
&\overset{(2)}{=} \operatorname{tr}(AF) + \operatorname{tr}(\rho F) - \tfrac{1}{3}\Big[\operatorname{tr}A^3 + 3\operatorname{tr}(A^2\rho) + 3\operatorname{tr}(A\rho^2) + \operatorname{tr}\theta^3\Big] \\
&\overset{(3)}{=} CS(A) + \operatorname{tr}(\rho F) - \operatorname{tr}(A^2\rho) - \operatorname{tr}(A\rho^2) - \tfrac{1}{3}\operatorname{tr}\theta^3 \\
&\overset{(4)}{=} CS(A) + \operatorname{tr}(\rho\,\mathrm{d}A) - \operatorname{tr}(A\rho^2) - \tfrac{1}{3}\operatorname{tr}\theta^3 \\
&\overset{(5)}{=} CS(A) - \mathrm{d}\operatorname{tr}(\rho\wedge A) - \tfrac{1}{3}\operatorname{tr}\theta^3.
\end{aligned} \tag{T3.98}$$

(1) Definition, with $A^g = \tilde{A} + \theta$ and $F^g = \tilde{F}$.
(2) Conjugation drops out of traces: $\operatorname{tr}(\tilde{A}\tilde{F}) = \operatorname{tr}(AF)$. Also $\operatorname{tr}(\theta\tilde{F}) = \operatorname{tr}(g^{-1}\mathrm{d}g\, g^{-1}Fg) = \operatorname{tr}(\mathrm{d}g\, g^{-1}F) = \operatorname{tr}(\rho F)$. In the cube, the three terms with two $\tilde{A}$'s and one θ are equal by cyclicity of three 1-forms, and $\operatorname{tr}(\tilde{A}^2\theta) = \operatorname{tr}(g^{-1}A^2g\, g^{-1}\mathrm{d}g) = \operatorname{tr}(A^2\rho)$; likewise the three terms with one $\tilde{A}$ give $3\operatorname{tr}(\tilde{A}\theta^2) = 3\operatorname{tr}(A\rho^2)$.
(3) Collect $\operatorname{tr}(AF) - \frac{1}{3}\operatorname{tr}A^3 = CS(A)$.
(4) $\operatorname{tr}(\rho F) = \operatorname{tr}(\rho\,\mathrm{d}A) + \operatorname{tr}(\rho A^2)$ and $\operatorname{tr}(\rho A^2) = \operatorname{tr}(A^2\rho)$ (a 1-form past a 2-form, no sign).
(5) $\mathrm{d}\operatorname{tr}(\rho\wedge A) = \operatorname{tr}(\mathrm{d}\rho\wedge A) - \operatorname{tr}(\rho\wedge \mathrm{d}A) = \operatorname{tr}(\rho^2A) - \operatorname{tr}(\rho\,\mathrm{d}A)$ by (T3.97), and $\operatorname{tr}(\rho^2A) = \operatorname{tr}(A\rho^2)$.

Key idea: gauge variation of the Chern–Simons form

$$CS(A^g) = CS(A) - \mathrm{d}\operatorname{tr}\big(\mathrm{d}g\, g^{-1}\wedge A\big) - \tfrac{1}{3}\operatorname{tr}\big(g^{-1}\mathrm{d}g\big)^3. \tag{T3.99}$$

The change is an exact term plus the winding density. On a closed oriented 3-manifold M the exact term integrates to zero, and by (T3.90)

$$\frac{1}{8\pi^2}\int_M CS(A^g) = \frac{1}{8\pi^2}\int_M CS(A) + w(g), \qquad w(g)\in\mathbb{Z}. \tag{T3.100}$$

So $\frac{1}{8\pi^2}\int_M CS(A)$ is gauge invariant *modulo integers*: it is a well-defined element of $\mathbb{R}/\mathbb{Z}$ on the space of gauge orbits.

Two special cases are worth isolating.

- $A = 0$. Then $A^g = \theta$ and (T3.99) gives $CS(\theta) = -\frac{1}{3}\operatorname{tr}\theta^3$, which is (T3.89) again.
- *Small gauge transformations.* Put $g = e^{\varepsilon X}$ with $X(x)\in\mathfrak{g}$. Then $\rho = \varepsilon\,\mathrm{d}X + O(\varepsilon^2)$ and $\operatorname{tr}\theta^3 = O(\varepsilon^3)$, so to first order

$$\delta_X CS(A) = -\varepsilon\,\mathrm{d}\operatorname{tr}(\mathrm{d}X\wedge A) = \varepsilon\,\mathrm{d}\operatorname{tr}(X\,\mathrm{d}A), \tag{T3.101}$$

where the last step uses $\operatorname{tr}(\mathrm{d}X\wedge A) = \mathrm{d}\operatorname{tr}(XA) - \operatorname{tr}(X\mathrm{d}A)$ and $\mathrm{dd} = 0$. The change is exact, and $\int_M CS$ is invariant under every gauge transformation that can be reached from the identity: $w(e^{\varepsilon X}) = 0$.

The general variation. For the equations of motion we need the change of CS under an arbitrary small change $A\to A+\delta A$:

$$\begin{aligned}\delta\, CS(A) &\overset{(1)}{=} \operatorname{tr}(\delta A\,\mathrm{d}A) + \operatorname{tr}(A\,\mathrm{d}\delta A) + \tfrac{2}{3}\cdot 3\operatorname{tr}(\delta A\, A^2)\\ &\overset{(2)}{=} \operatorname{tr}(\delta A\,\mathrm{d}A) + \operatorname{tr}(\delta A\,\mathrm{d}A) - \mathrm{d}\operatorname{tr}(A\wedge\delta A) + 2\operatorname{tr}(\delta A\, A^2)\\ &\overset{(3)}{=} 2\operatorname{tr}(\delta A\wedge F) - \mathrm{d}\operatorname{tr}(A\wedge\delta A).\end{aligned} \tag{T3.102}$$

(1) Vary each factor; the three variations of $\operatorname{tr}A^3$ are equal by cyclicity of three 1-forms.
(2) $\mathrm{d}\operatorname{tr}(A\wedge\delta A) = \operatorname{tr}(\mathrm{d}A\,\delta A) - \operatorname{tr}(A\,\mathrm{d}\delta A)$. So $\operatorname{tr}(A\,\mathrm{d}\delta A) = \operatorname{tr}(\mathrm{d}A\,\delta A) - \mathrm{d}\operatorname{tr}(A\,\delta A)$, and $\operatorname{tr}(\mathrm{d}A\,\delta A) = \operatorname{tr}(\delta A\,\mathrm{d}A)$.
(3) $\mathrm{d}A + A^2 = F$.

(For $\delta A = DX$ one can check that this reproduces (T3.101), using $\operatorname{tr}(DX\wedge F) = \mathrm{d}\operatorname{tr}(XF)$, which is Bianchi again.)

T3.10.6 The instanton number is a winding number

Here is the payoff. Consider a gauge field on $\mathbb{R}^4$ with finite action. Finite action needs $F \to 0$ at infinity faster than $1/r^2$, so on a large 3-sphere S^3_R the field is *flat*, and since S^3 is simply connected it is *pure gauge* there:

$$A \xrightarrow{r\to\infty} g^{-1}\mathrm{d}g, \qquad g : S^3_\infty \to G.$$

The map g is a map from a 3-sphere to the group, and it has a winding number $w(g) \in \pi_3(G) = \mathbb{Z}$.

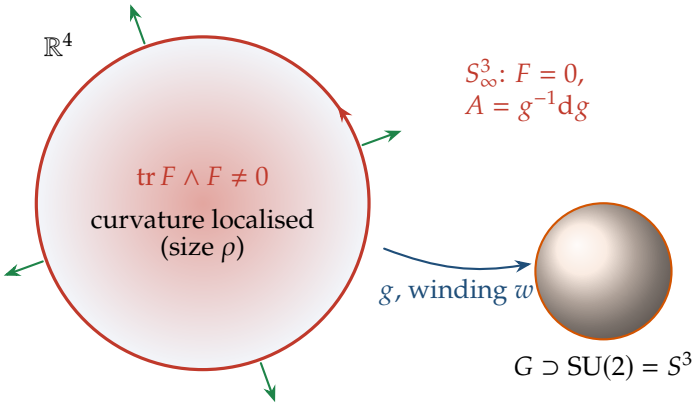


Figure T3.34: An instanton on $\mathbb{R}^4$. The action density (red) is concentrated in a region of size ρ. Far away the field is pure gauge, $A = g^{-1}\mathrm{d}g$, so it defines a map g from the 3-sphere at infinity to the group. Stokes' theorem converts $\int \mathrm{tr}\, F \wedge F$ over the whole of $\mathbb{R}^4$ into $\int CS$ over this sphere, where CS is the winding density of g.

Key idea: instanton number = winding at infinity

With $\mathbb{R}^4$ oriented by $\mathrm{d}x^1 \wedge \mathrm{d}x^2 \wedge \mathrm{d}x^3 \wedge \mathrm{d}x^4$ and S^3_∞ oriented as its boundary (outward normal first),

$$k := \frac{1}{8\pi^2}\int_{\mathbb{R}^4} \mathrm{tr}(F \wedge F) = w(g) \in \mathbb{Z}. \tag{T3.103}$$

The derivation is four lines:

$$\int_{\mathbb{R}^4} \mathrm{tr}(F\wedge F) \overset{(1)}{=} \lim_{R\to\infty}\int_{B_R} \mathrm{d}\,CS(A) \overset{(2)}{=} \lim_{R\to\infty}\int_{S^3_R} CS(A)$$
$$\overset{(3)}{=} \int_{S^3_\infty} CS\big(g^{-1}\mathrm{d}g\big) \overset{(4)}{=} 8\pi^2\, w(g).$$

(1) $\mathrm{tr}\, F \wedge F = \mathrm{d}\, CS(A)$ (T3.84); on $\mathbb{R}^4$ the gauge field is a single globally defined 1-form, so $CS(A)$ is defined everywhere. B_R is the ball of radius R.
(2) Stokes' theorem.
(3) On S^3_R, $A = g^{-1}\mathrm{d}g + o(1/R)$; CS is cubic in A and its derivatives, which are $O(1/R)$ and $O(1/R^2)$, while the area of S^3_R grows like R^3, so the corrections vanish in the limit.
(4) (T3.90).

This is the exemplar argument of the style guide again, one level up: the integral of an exact form sees only the boundary, and on the boundary the field is pure gauge and carries only topological information. The integrality $\int \mathrm{tr}\, F \wedge F \in 8\pi^2\mathbb{Z}$ quoted in the QFT notes [QFT notes (4.13.13)] is exactly this.

Worked example T3.52: Every radial ansatz has $k = 1$

Let $g(x) = (x^4\mathbb{1} + i\vec{x}\cdot\vec{\sigma})/r$, the winding-one map of Example T3.27, and $\theta = g^{-1}\mathrm{d}g$. Since g depends only on the direction x/r, θ has no $\mathrm{d}r$ component. Take

$$A = f(r)\,\theta, \qquad f(0) = 0, \quad f(\infty) = 1,$$

with f smooth and $f \sim r^2$ near 0 (so that A, which behaves like f/r times a smooth angular form, is smooth at the origin). Then:

$$\begin{aligned} F &\overset{(1)}{=} \mathrm{d}f\wedge\theta + f\,\mathrm{d}\theta + f^2\theta\wedge\theta \overset{(2)}{=} f'\,\mathrm{d}r\wedge\theta + (f^2 - f)\,\theta^2, \\ \operatorname{tr}(F\wedge F) &\overset{(3)}{=} (f^2-f)f'\left[\operatorname{tr}(\mathrm{d}r\wedge\theta\wedge\theta^2) + \operatorname{tr}(\theta^2\wedge\mathrm{d}r\wedge\theta)\right] + (f^2-f)^2\operatorname{tr}\theta^4 \\ &\overset{(4)}{=} 2(f^2-f)f'\,\mathrm{d}r\wedge\operatorname{tr}\theta^3. \end{aligned}$$

(1) $\mathrm{d}(f\theta) = \mathrm{d}f\wedge\theta + f\,\mathrm{d}\theta$, and $A\wedge A = f^2\theta^2$.
(2) $\mathrm{d}f = f'\mathrm{d}r$ and the structure equation $\mathrm{d}\theta = -\theta^2$.
(3) Expand the square; the term $(f'\mathrm{d}r\wedge\theta)^2$ vanishes because it contains $\mathrm{d}r\wedge\mathrm{d}r = 0$.
(4) $\operatorname{tr}\theta^4 = 0$ (graded cyclicity); $\theta^2\wedge\mathrm{d}r = \mathrm{d}r\wedge\theta^2$ (a 2-form commutes with a 1-form), so both cross terms equal $\mathrm{d}r\wedge\operatorname{tr}\theta^3$.

Now integrate, using $\mathbb{R}^4\setminus\{0\} = (0,\infty)\times S^3$ with orientation $\mathrm{d}r\wedge$ (boundary orientation of S^3):

$$\begin{aligned} \int_{\mathbb{R}^4}\operatorname{tr}(F\wedge F) &\overset{(1)}{=} \int_0^\infty 2(f^2-f)f'\,\mathrm{d}r\int_{S^3}\operatorname{tr}\theta^3 \overset{(2)}{=} \left[\tfrac{2}{3}f^3 - f^2\right]_{f=0}^{f=1}\cdot(-24\pi^2) \\ &= \left(-\tfrac{1}{3}\right)(-24\pi^2) = 8\pi^2. \end{aligned}$$

(1) Fubini; the S^3 factor is the same on every sphere since θ depends only on the direction.
(2) Substitute $u = f(r)$: $\int(2u^2-2u)\,\mathrm{d}u$; and $\int_{S^3}\operatorname{tr}\theta^3 = -24\pi^2 w(g) = -24\pi^2$ (T3.43).

So $k = 1$, for *every* profile f. The shape of f decides where the topological charge density sits (Figure T3.35), but not how much there is. The choice $f = r^2/(r^2+\rho^2)$ is the one for which F is (anti-)self-dual and the action is minimal, $S = 8\pi^2/g_{\mathrm{YM}}^2$: the BPST instanton of Chapter 8 in "regular gauge". Check with (T3.103): at infinity $A\to\theta$ with $w = 1$. Also check the CS form directly: $CS(f\theta) = \operatorname{tr}\big(f\theta(f'\mathrm{d}r\wedge\theta - f\theta^2)\big) + \frac{2}{3}f^3\operatorname{tr}\theta^3 = (\frac{2}{3}f^3 - f^2)\operatorname{tr}\theta^3$ (the $\mathrm{d}r$ term is $-ff'\mathrm{d}r\wedge\operatorname{tr}\theta^2 = 0$), which at $f = 1$ is $-\frac{1}{3}\operatorname{tr}\theta^3$, as it must be.

Watch out: which sign of duality has $k > 0$?

In Example T3.38 we found $S_{\mathrm{YM}} \geq \mp\int\operatorname{tr}F\wedge F$ with equality iff $F = \pm\star F$. With our conventions ($\operatorname{tr}(X^2)\leq 0$ for anti-Hermitian X, and $k = \frac{1}{8\pi^2}\int\operatorname{tr}F\wedge F$), a self-dual field has $\int\operatorname{tr}F\wedge F = \int\operatorname{tr}F\wedge\star F = -g_{\mathrm{YM}}^2 S \leq 0$, i.e. $k\leq 0$; the fields with $k > 0$ that saturate the bound are anti-self-dual. Many books define $k = -\frac{1}{8\pi^2}\int\operatorname{tr}F\wedge F$ (or use the opposite orientation) so that instantons are self-dual with $k > 0$. Nothing physical depends on this, but it is the most common source of stray minus signs. Always check one example, such as Example T3.52.

The same argument on S^4: transition functions. On the compact S^4 a gauge field with $k\neq 0$ cannot be written as a single globally defined 1-form (if it could, $k = \frac{1}{8\pi^2}\int_{S^4}\mathrm{d}\,CS = 0$

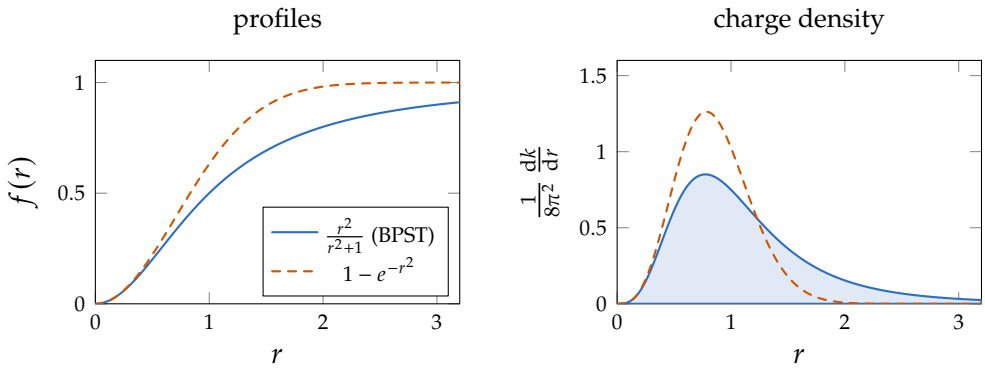


Figure T3.35: Two radial profiles $f(r)$ (left) and the corresponding radial density of topological charge $6(f-f^2)f'$ (right; the $\frac{1}{8\pi^2}$ and the S^3 integral are included). The curves have different shapes but both enclose area exactly 1: the charge is fixed by $f(0)=0$, $f(\infty)=1$ and the winding of θ, not by the details.

by Stokes). Instead, as for the monopole of Example T3.37, we use two patches: A_N on the northern hemisphere D_N and A_S on the southern D_S, glued on the equator S^3 by a transition function, $A_N = A_S^g = g^{-1}A_S g + g^{-1}\mathrm{d}g$. Then

$$\int_{S^4} \operatorname{tr} F\wedge F \overset{(1)}{=} \int_{D_N} \mathrm{d}\,CS(A_N) + \int_{D_S} \mathrm{d}\,CS(A_S) \overset{(2)}{=} \int_{S^3}\big[CS(A_N) - CS(A_S)\big]$$
$$\overset{(3)}{=} \int_{S^3}\Big[-\mathrm{d}\operatorname{tr}(\mathrm{d}g\,g^{-1}\wedge A_S) - \tfrac{1}{3}\operatorname{tr}(g^{-1}\mathrm{d}g)^3\Big] \overset{(4)}{=} 8\pi^2\, w(g).$$

(1) Split the sphere; on each hemisphere A is a genuine 1-form, so $\operatorname{tr}F\wedge F = \mathrm{d}\,CS$ there.
(2) Stokes on each hemisphere; the equator is the boundary of both, with opposite orientations. We orient S^3 as ∂D_N.
(3) (T3.99) with $A = A_S$.
(4) The exact term integrates to zero on the closed S^3; (T3.90) for the rest.

So *the instanton number is the winding number of the transition function on the equator*, exactly as the monopole charge was the winding of the U(1) transition function $e^{-i\phi}$ on the equator of S^2. On $\mathbb{R}^4$ the "southern patch" is the point at infinity, and g is the gauge transformation at infinity. This is the statement, used throughout Chapter 8, that SU(2) bundles over S^4 are classified by $\pi_3(\mathrm{SU}(2)) = \mathbb{Z}$, with $k = c_2$ up to sign.

T3.10.7 Chern–Simons number and the vacua of Yang–Mills theory

The same identity has a Hamiltonian reading. Split Euclidean $\mathbb{R}^4$ into time and space, $\mathbb{R}_t\times\mathbb{R}^3$, oriented by $\mathrm{d}t\wedge\mathrm{d}^3x$, and consider the slab $[-T,T]\times\mathbb{R}^3$. Suppose $A\to 0$ fast enough at spatial infinity that the sides of the slab contribute nothing. Define the **Chern–Simons number** of a spatial configuration

$$N_{\mathrm{CS}}[A] \coloneqq \frac{1}{8\pi^2}\int_{\mathbb{R}^3} CS(A). \tag{T3.104}$$

Then Stokes' theorem on the slab gives (the top face has the orientation d^3x, the bottom face the opposite)

$$k = \frac{1}{8\pi^2}\int_{[-T,T]\times\mathbb{R}^3} \operatorname{tr}F\wedge F = N_{\mathrm{CS}}[A(T)] - N_{\mathrm{CS}}[A(-T)]. \tag{T3.105}$$

The instanton number is the change in Chern–Simons number.

Now look at the classical vacua, the spatial configurations with zero energy: $F = 0$ on $\mathbb{R}^3$, hence $A = h^{-1}\mathrm{d}h$ with $h : \mathbb{R}^3 \to G$. If we require $h \to \mathbb{1}$ at spatial infinity, h is a map from $\mathbb{R}^3 \cup \{\infty\} = S^3$ to G, and by (T3.90)

$$N_{\mathrm{CS}}\left[h^{-1}\mathrm{d}h\right] = w(h) \in \mathbb{Z}.$$

So there is not one classical vacuum but a whole $\mathbb{Z}$'s worth, $|n\rangle$ with $N_{\mathrm{CS}} = n$, related by the large gauge transformations of Example T3.27. They cannot be connected by a path of zero-energy configurations, because N_{CS} is an integer on vacua and changes continuously along a path. A path from $|n\rangle$ to $|n+1\rangle$ must climb over an energy barrier, and in Euclidean time the tunnelling event is an instanton with $k = 1$ (Figure T3.36). Its amplitude is suppressed by $e^{-S} = e^{-8\pi^2/g_{\mathrm{YM}}^2}$ [QFT notes (4.13.14)]: invisible in perturbation theory, but not zero.

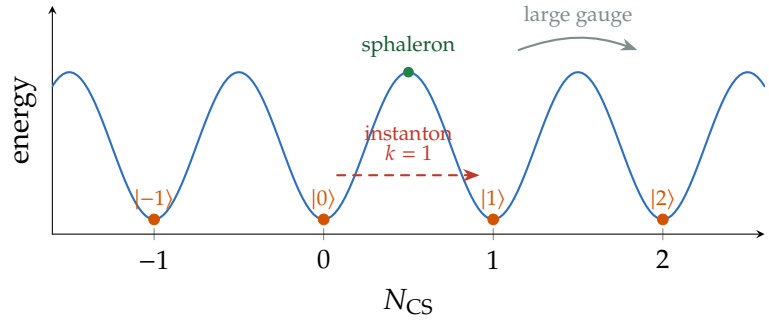


Figure T3.36: Schematic energy of static gauge fields as a function of the Chern–Simons number. The classical vacua (orange) sit at the integers and are pure gauge, $A = h^{-1}\mathrm{d}h$ with winding $w(h) = N_{\mathrm{CS}}$; large gauge transformations shift between them. An instanton is the Euclidean tunnelling path from $|n\rangle$ to $|n+k\rangle$ under the barrier; the lowest point on the barrier (green) is the sphaleron. (Energy is not literally a function of N_{CS} alone: the picture is a one-dimensional slice.)

Physical picture: θ-vacua in one paragraph

Because the large gauge transformation U maps $|n\rangle \to |n+1\rangle$ and commutes with the Hamiltonian, the true energy eigenstates are its eigenstates, the Bloch-wave combinations $|\vartheta\rangle = \sum_n e^{in\vartheta}|n\rangle$, exactly as for an electron in a periodic potential. The angle ϑ is a new parameter of the theory. In the Euclidean path integral it appears as a weight $e^{i\vartheta k}$ for each sector of instanton number k, i.e. as an extra term $\frac{i\vartheta}{8\pi^2}\int \mathrm{tr}\, F \wedge F$ in the exponent, which is a total derivative and so invisible in the classical equations of motion. The analogous statement in electroweak theory, combined with the anomaly (Chapter 10), is that baryon number changes by the number of generations times ΔN_{CS}.

T3.10.8 The Chern–Simons action in three dimensions

In three dimensions $CS(A)$ is a 3-form, so it can be integrated over the spacetime itself. Witten's Chern–Simons theory, the subject of the last part of Nash (Chapter 12), takes it as the entire action. In Nash's normalisation the path-integral weight is [Nash p. 343, (12.85)]

$$e^{iS_k[A]}, \qquad S_k[A] = -\frac{k}{4\pi}\int_M \mathrm{tr}\left(A \wedge \mathrm{d}A + \tfrac{2}{3}A \wedge A \wedge A\right) = -\frac{k}{4\pi}\int_M CS(A), \tag{T3.106}$$

with M a closed oriented 3-manifold and k a constant, the **level**.

The level is quantised. Under a gauge transformation g, (T3.100) gives

$$S_k[A^g] \overset{(1)}{=} -\frac{k}{4\pi}\Big(\int_M CS(A) + 8\pi^2 w(g)\Big) = S_k[A] - 2\pi k\, w(g),$$
$$e^{iS_k[A^g]} \overset{(2)}{=} e^{iS_k[A]}\, e^{-2\pi i\, k\, w(g)}.$$

(1) (T3.100).
(2) Exponentiate.

The action itself is not gauge invariant, but the weight is, for every g, if and only if

$$k \in \mathbb{Z}.$$

(On S^3 there are maps g with $w(g) = 1$, so the condition is necessary.) This is the first example of a phenomenon we will meet again: a quantum theory can be consistent even though its classical action is not invariant, provided the non-invariance is an integer multiple of 2π. Nash's version for the topologically massive theory of Deser–Jackiw–Templeton and Schonfeld is the action

$$S = \|F\|^2 + \alpha\, TP_2(A) = -\operatorname{tr}\int_M \Big\{F\wedge\star F + \frac{\alpha}{8\pi^2}\Big(\mathrm{d}A\wedge A + \tfrac{2}{3}A\wedge A\wedge A\Big)\Big\}$$

[Nash p. 227, (8.60)–(8.62)], with $TP_2(A) = -\frac{1}{8\pi^2}CS(A)$. Under g, $\alpha\int TP_2$ shifts by $-\alpha\, w(g)$, so e^{iS} is gauge invariant iff $\alpha \in 2\pi\mathbb{Z}$, his quantisation condition. (Note $\operatorname{tr}(\mathrm{d}A\wedge A) = \operatorname{tr}(A\wedge \mathrm{d}A)$.) The Chern–Simons term gives the gauge field a mass without a Higgs field, and the mass is quantised in units set by the Yang–Mills coupling; this is why the $\pi_0(\mathcal{G}) = \pi_3(G) = \mathbb{Z}$ of [Nash (8.59)] has physical consequences.

Physicists' normalisation. With the dictionary $A = -i\mathcal{A}$ of Section T3.8.9 (coupling absorbed into $\mathcal{A}$),

$$CS(A) \overset{(1)}{=} \operatorname{tr}\big((-i)^2\mathcal{A}\,\mathrm{d}\mathcal{A} + \tfrac{2}{3}(-i)^3\mathcal{A}^3\big) \overset{(2)}{=} -\operatorname{tr}\Big(\mathcal{A}\,\mathrm{d}\mathcal{A} - \tfrac{2i}{3}\mathcal{A}^3\Big),$$

(1) Substitute; the factors of $-i$ are scalars.
(2) $(-i)^2 = -1$ and $(-i)^3 = i$.

so $S_k = \frac{k}{4\pi}\int \operatorname{tr}(\mathcal{A}\,\mathrm{d}\mathcal{A} - \frac{2i}{3}\mathcal{A}^3)$, the form in which Witten writes it, with Hermitian $\mathcal{A}$ and tr in the fundamental.

The equations of motion are flatness. Vary $A \to A + \delta A$ in (T3.106). By (T3.102) the exact term integrates to zero on the closed M, and

$$\delta S_k = -\frac{k}{4\pi}\int_M 2\operatorname{tr}(\delta A\wedge F) = -\frac{k}{2\pi}\int_M \operatorname{tr}(\delta A\wedge F).$$

This vanishes for every δA iff $F = 0$. *The classical solutions of Chern–Simons theory are the flat connections*, which by (T3.83) are the representations of $\pi_1(M)$ in G. There are no propagating degrees of freedom; everything is topological. The same computation is Nash's derivation that the critical points of the Chern–Simons function $f(A) = -\frac{1}{8\pi^2}\int_M CS(A)$ used in Floer theory are the flat connections: $\delta f = -\frac{1}{4\pi^2}\int_M \operatorname{tr}(\delta A\wedge F)$, i.e. "$\mathrm{d}f(A) = -F(A)/4\pi^2$" [Nash p. 324–325, (12.4)–(12.8)], and $f(A^g) = f(A) - w(g)$ is his (12.5).

Key idea: why Chern–Simons theory is topological

The action (T3.106) is built from d, $\wedge$ and tr only. None of these uses a metric. So the classical theory, and (with care about the gauge fixing, which does need a metric and leaves a framing dependence) the quantum theory, depend only on the topology of M and of any Wilson loops inserted in it. The Wilson loop expectation values are knot invariants.

Worked example T3.53: Abelian Chern–Simons: magnetic helicity and linking

For $G = \mathrm{U}(1)$ write $A = ia$ with a a real 1-form (the electromagnetic potential, $b = \mathrm{d}a$ the magnetic field 2-form). Then $A \wedge A = 0$ and

$$CS(A) = (i)^2\, a \wedge \mathrm{d}a = -a \wedge \mathrm{d}a, \qquad \int_{\mathbb{R}^3} a \wedge \mathrm{d}a = \int \vec{a} \cdot \vec{b}\, \mathrm{d}^3x =: H,$$

the **magnetic helicity** of plasma physics. Its value for two thin, closed, untwisted flux tubes along curves C_1, C_2 carrying fluxes Φ_1, Φ_2 (Figure T3.37) is

$$H \overset{(1)}{=} \int_{\mathbb{R}^3} a \wedge b \overset{(2)}{=} \Phi_1 \oint_{C_1} a + \Phi_2 \oint_{C_2} a \overset{(3)}{=} \Phi_1 \int_{\Sigma_1} b + \Phi_2 \int_{\Sigma_2} b$$
$$\overset{(4)}{=} \Phi_1\, \Phi_2\, \mathrm{Lk}(C_1, C_2) + \Phi_2\, \Phi_1\, \mathrm{Lk}(C_2, C_1) \overset{(5)}{=} 2\, \Phi_1 \Phi_2\, \mathrm{Lk}(C_1, C_2).$$

(1) $b = \mathrm{d}a$.
(2) For a thin tube along C with flux Φ (field pointing along C), $\int \alpha \wedge b = \Phi \oint_C \alpha$ for any 1-form α: in local coordinates with C along z, $b = \Phi\, \delta^2(x, y)\, \mathrm{d}x \wedge \mathrm{d}y$ and $\alpha \wedge b = \Phi\, \alpha_z\, \delta^2(x, y)\, \mathrm{d}^3x$.
(3) Stokes, with Σ_i a surface bounded by C_i (a Seifert surface).
(4) The flux through Σ_1 comes only from tube 2, which pierces Σ_1 a net $\mathrm{Lk}(C_1, C_2)$ times; tube 1 contributes nothing because it is untwisted (its own field lines do not cross Σ_1).
(5) The linking number is symmetric.

So the Chern–Simons integral of a pair of flux tubes counts how many times they link. The linking number has Gauss's integral formula [Nash p. 344, (12.90)]

$$\mathrm{Lk}(C_1, C_2) = \frac{1}{4\pi} \oint_{C_1} \mathrm{d}x^i \oint_{C_2} \mathrm{d}y^j\, \varepsilon_{ijk} \frac{(x - y)^k}{|x - y|^3},$$

which is the Biot–Savart law: $\oint_{C_1} a$ with a the potential of a unit current along C_2. In the U(1) Chern–Simons path integral with Wilson lines $W = e^{in \oint_C a}$, the Gaussian integral over a produces exactly this double integral, and Nash obtains $\langle W(n_1, C_1) \cdots \rangle = \exp\left[\frac{i\pi}{k} \sum n_l n_m \mathrm{Lk}(C_l, C_m)\right]$ [Nash (12.92)]: a knot invariant computed by a quantum field theory. The self-linking terms $l = m$ need a framing ("untwisted" above), which is the framing dependence mentioned in the key idea.

T3.10.9 Descent: a preview of anomalies

Look again at the chain of identities we have derived:

$$\mathrm{tr}(F \wedge F) = \mathrm{d}\, CS(A),$$
$$\delta_X CS(A) = \mathrm{d}\ \mathrm{tr}(X\, \mathrm{d}A) \qquad (\text{with } \varepsilon = 1). \tag{T3.107}$$

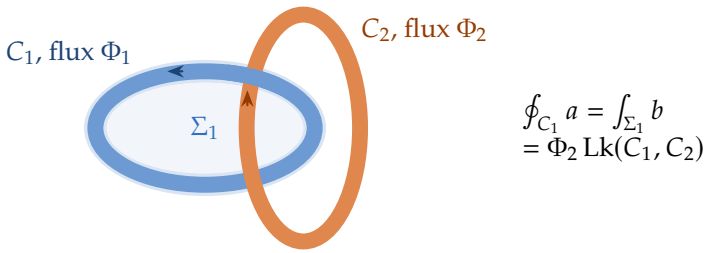


Figure T3.37: Two linked flux tubes (a Hopf link). Tube 2 (orange) threads the surface Σ_1 spanning tube 1 (blue) once, so the circulation of the vector potential around C_1 is Φ_2. The magnetic helicity $\int a \wedge \mathrm{d}a$, the abelian Chern–Simons integral, is $2\Phi_1\Phi_2\,\mathrm{Lk}(C_1, C_2)$.

A gauge-invariant closed 4-form determines a 3-form, which is not gauge invariant; its gauge variation is exact and determines a 2-form linear in the gauge parameter X. Writing $\omega_4 = \operatorname{tr} F^2$, $\omega_3 = CS$, $\omega_2^1 = \operatorname{tr}(X\mathrm{d}A)$,

$$\omega_4 = \mathrm{d}\omega_3, \qquad \delta_X \omega_3 = \mathrm{d}\omega_2^1.$$

These are the first two of the **descent equations**. Their importance is this. In Chapter 10 we will find that a chiral fermion in two dimensions, coupled to A, has an effective action $\Gamma[A]$ that fails to be gauge invariant, and that the failure is

$$\delta_X \Gamma[A] \;\propto\; \int_{M^2} \operatorname{tr}(X\,\mathrm{d}A) = \int_{M^2} \omega_2^1,$$

the bottom of the descent from $\operatorname{tr} F^2$ in four dimensions. Because ω_2^1 comes from descent, it automatically satisfies the Wess–Zumino consistency condition (two successive gauge variations commute in the right way). In four dimensions the same construction starts from $\operatorname{tr} F^3$ in six dimensions: $\operatorname{tr} F^3 = \mathrm{d}\,CS_5$ (Example T3.51), and the variation of CS_5 is $\mathrm{d}\omega_4^1$ with $\omega_4^1 = \operatorname{tr}\big(X\,\mathrm{d}(A\,\mathrm{d}A + \tfrac{1}{2}A^3)\big)$ (a standard result which we quote here and derive in Chapter 10). That 4-form is the non-abelian anomaly, and Nash's μ_1 of [Nash (8.58)] is the same object seen as a cohomology class on the gauge group.

Try it: Abelian descent

For $G = \mathrm{U}(1)$, check (T3.107) directly: $CS = A \wedge \mathrm{d}A$, and under $A \to A + \mathrm{d}X$ (with X imaginary), $\delta CS = \mathrm{d}X \wedge \mathrm{d}A = \mathrm{d}(X\,\mathrm{d}A)$.
Answer: $\delta(A \wedge \mathrm{d}A) = \mathrm{d}X \wedge \mathrm{d}A + A \wedge \mathrm{d}\mathrm{d}X = \mathrm{d}X \wedge \mathrm{d}A$, and $\mathrm{d}(X\mathrm{d}A) = \mathrm{d}X \wedge \mathrm{d}A + X\,\mathrm{d}\mathrm{d}A = \mathrm{d}X \wedge \mathrm{d}A$. In two dimensions this gives the abelian anomaly $\propto \int X\,F$, the $\epsilon^{\mu\nu}F_{\mu\nu}$ of the Schwinger model [QFT notes Ch. 22].

Reading Nash (the Chern–Simons form throughout Nash)

The Chern–Simons form appears in Nash in five guises, all covered in this section. (i) As the transgression $TP_2(A)$ of the second Chern class, $\operatorname{tr} F \wedge F = \mathrm{d}(\cdots)$ [Nash pp. 223–224, (8.39)–(8.45)], with $A_t = tA$, $F_t = t\,\mathrm{d}A + t^2 A \wedge A$ [Nash (8.41)]; our (T3.94). (ii) As an $\mathbb{R}/\mathbb{Z}$-valued secondary class, because a gauge transformation shifts it by an integer [Nash (8.45)]; our (T3.100). (iii) As a term in the 3D gauge theory action with quantised coefficient [Nash p. 227, (8.59)–(8.62)]. (iv) As the Morse function of Floer theory, whose critical points are flat connections [Nash pp. 324–325, (12.4)–(12.10)]. (v) As the action of Witten's topological field theory, whose Wilson-loop correlators give linking numbers and the Jones polynomial [Nash pp. 342–348, (12.84)–(12.94)]. Nash also writes the instanton number as $\frac{1}{8\pi^2}\int \operatorname{tr} F \wedge F$ and uses $\pi_3(\mathrm{SU}(2)) = \mathbb{Z}$ to classify SU(2) bundles on S^4 and the components of the 3D gauge group; our (T3.103) and its two-patch version are the reason these are the same

integer.

Recap

- Pure gauge $A = g^{-1}\mathrm{d}g$: $F = 0$ is the Maurer–Cartan equation; transport $W_\gamma = g(b)^{-1}g(a)$, trivial holonomy.
- Flat $\Rightarrow$ locally pure gauge (integrability of $\mathrm{d}g = gA$ is $F = 0$; construct g by parallel transport). Globally, flat connections up to gauge = $\mathrm{Hom}(\pi_1(M), G)$/conjugation. Circle: conjugacy classes ($\mathbb{R}/\mathbb{Z}$ for U(1), $[0, \frac{1}{2}]$ for SU(2)); torus: commuting pairs, the pillowcase.
- $CS(A) = \mathrm{tr}(A\,\mathrm{d}A + \frac{2}{3}A^3) = \mathrm{tr}(AF - \frac{1}{3}A^3)$, $\mathrm{d}\,CS = \mathrm{tr}\,F \wedge F$; transgression $CS = 2\int_0^1 \mathrm{tr}(A \wedge F_t)\,\mathrm{d}t$, $\frac{2}{3} = \int_0^1 2t^2$; $CS_5 = \mathrm{tr}(A(\mathrm{d}A)^2 + \frac{3}{2}A^3\mathrm{d}A + \frac{3}{5}A^5)$; Chern–Weil.
- $CS(A^g) = CS(A) - \mathrm{d}\,\mathrm{tr}(\mathrm{d}g\,g^{-1}A) - \frac{1}{3}\,\mathrm{tr}(g^{-1}\mathrm{d}g)^3$; $\frac{1}{8\pi^2}\int_{M^3} CS$ shifts by $w(g) \in \mathbb{Z}$; infinitesimally $\delta CS = \mathrm{d}\,\mathrm{tr}(X\mathrm{d}A)$; general variation $2\,\mathrm{tr}(\delta A\,F)$+exact.
- $k = \frac{1}{8\pi^2}\int_{\mathbb{R}^4} \mathrm{tr}\,F \wedge F = w(g_\infty)$; every radial ansatz $f\,\theta$ has $k = 1$; on S^4, $k = w$(transition function); $k = \Delta N_{\mathrm{CS}}$, vacua $|n\rangle$, ϑ-vacua.
- 3D Chern–Simons action $-\frac{k}{4\pi}\int CS$: level $k \in \mathbb{Z}$, equations of motion $F = 0$, metric independent; abelian case = helicity = $2\Phi_1\Phi_2\mathrm{Lk}$.
- Descent $\mathrm{tr}\,F^2 \to CS \to \mathrm{tr}(X\mathrm{d}A)$: the seed of the anomaly.

Chapter summary

The chapter had one job: to take the physicist's two words "symmetry" and "gauge" and turn them into objects we can compute with. The first half built the symmetry side. A Lie group is a set of matrices closed under multiplication and cut out by smooth equations; its Lie algebra is the set of velocities at the identity, closed under the commutator; the exponential map goes back from the algebra to the group; and a representation lets the group act on the fields. The second half made the symmetry local. A field $g : M \to G$ has a derivative $g^{-1}\mathrm{d}g$ that lives in the Lie algebra. A covariant derivative needs a Lie-algebra-valued 1-form A that absorbs such terms, and the curvature of A is the field strength. Once there is a connection we can ask what it does around loops (holonomy, Wilson loops), when it is trivial (pure gauge versus flat), and how its topology is measured (the Chern–Simons form and the winding number).

The summary has four parts. Figures T3.38 and T3.39 put the two halves of the chapter on one page each. The concept tables (Tables T3.4 to T3.10) list every concept, what it means operationally, and where the rest of the book uses it. Table T3.11 collects the numbers worth knowing by heart. The chapter closes with the calculations a reader should be able to do without looking, and a map from Nash's pages back to the sections of this chapter.

The two halves on one page

Concept tables

In each table the last column says where the idea is used again. References to chapters are to the rest of the book; references to sections are to this chapter.

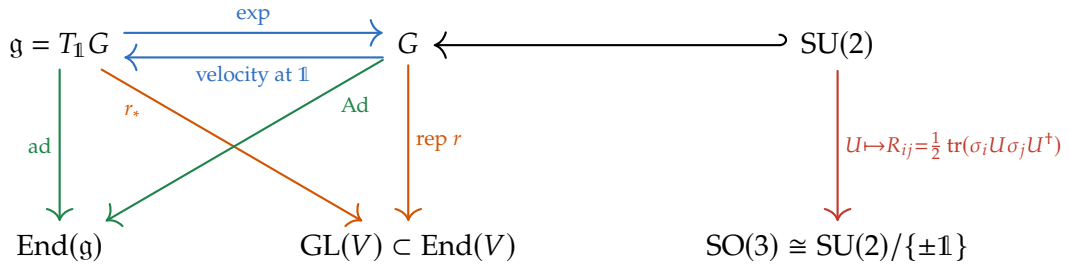


Figure T3.38: The algebraic half of the chapter. The exponential map (blue) goes from the Lie algebra to the group, and "velocity at the identity" comes back. A representation r of the group (orange) differentiates to a representation r_* of the algebra with the same structure constants. The group acts on its own algebra by $\mathrm{Ad}_g X = gXg^{-1}$ and the algebra by $\mathrm{ad}_Y X = [Y, X]$ (green). On the right is the most important example in physics: the adjoint representation of SU(2) is the double cover of SO(3) (red), with kernel $\pm\mathbb{1}$, which is why spin $\frac{1}{2}$ exists (Section T3.4).

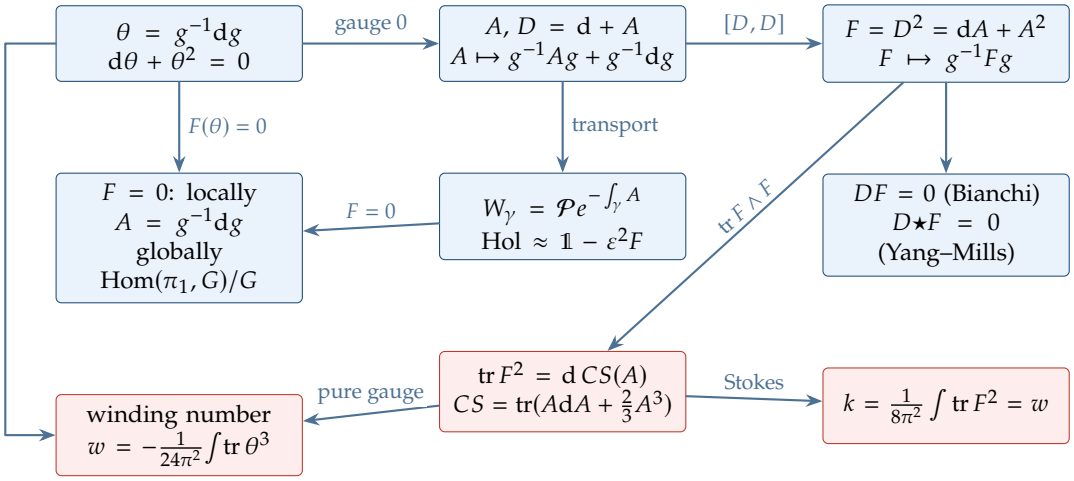


Figure T3.39: The geometric half of the chapter. The Maurer–Cartan form (top left) is what you get by gauge transforming $A = 0$; the gauge field A is its generalisation, and its curvature F is the commutator of covariant derivatives. Parallel transport measures F around small loops. When $F = 0$ only the topology of the loops is left. The red boxes are the topological numbers: the Chern–Simons form turns $\int \mathrm{tr}\,F \wedge F$ into a boundary term, which for a pure-gauge boundary is the winding number of g.

Concept	Operational meaning	Used later in
Matrix Lie group	A group of $n \times n$ matrices cut out by smooth equations ($R^T R = \mathbb{1}$, $U^\dagger U = \mathbb{1}$, $\det = 1, \dots$). Dimension = number of entries minus number of independent equations (Section T3.1).	Every gauge group; Chapters 6 and 8
Classical groups	$\mathrm{GL}(n)$, $\mathrm{SL}(n)$, $\mathrm{O}(n)$, $\mathrm{SO}(n)$, $\mathrm{U}(n)$, $\mathrm{SU}(n)$, $\mathrm{Sp}(n)$; test which equations a matrix satisfies (Table T3.1).	Structure groups of bundles, Chapters 1 and 4
$\mathrm{U}(1) = S^1$	$e^{i\alpha}$, $\alpha \sim \alpha + 2\pi$; the same as SO(2).	Electromagnetism, monopoles, Chapter 8
$\mathrm{SU}(2) = S^3$	$U = x^4\mathbb{1} + i\vec{x}\cdot\vec{\sigma}$ with $\sum x_\mu^2 = 1$ (T3.1); unit quaternions = Sp(1).	Instantons, $\pi_3 = \mathbb{Z}$, Chapter 8
$\mathrm{SO}(3) = \mathbb{RP}^3$	Ball of radius π (axis × angle) with antipodal surface points identified; $\mathrm{tr}\, R = 1 + 2\cos\theta$.	Spin, Chapter T4
Topological data	π_0 (components), π_1 (non-contractible loops), π_3 (winding of $S^3 \to G$) of each group (Table T3.2).	Instanton sectors, θ-vacua, anomalies

Table T3.4: Groups (Section T3.1).

Numbers worth knowing by heart

Group	dim	rank	π_1	π_3	$\mathfrak{g}$
U(1)	1	1	$\mathbb{Z}$	0	$i\mathbb{R}$
$\mathrm{SU}(2) = S^3$	3	1	0	$\mathbb{Z}$	traceless anti-Hermitian 2×2
$\mathrm{SO}(3) = \mathbb{RP}^3$	3	1	$\mathbb{Z}_2$	$\mathbb{Z}$	antisymmetric real 3×3
$\mathrm{SU}(N)$	N^2-1	$N-1$	0	$\mathbb{Z}$	traceless anti-Hermitian
$\mathrm{U}(N)$	N^2	N	$\mathbb{Z}$	$\mathbb{Z}$ $(N \geq 2)$	anti-Hermitian
$\mathrm{SO}(n)$, $n \geq 3$	$\frac{1}{2}n(n-1)$	$\lfloor n/2 \rfloor$	$\mathbb{Z}_2$	$\mathbb{Z}$ $(n \neq 4)$, $\mathbb{Z}^2$ $(n = 4)$	antisymmetric
$\mathrm{Sp}(n)$	$n(2n+1)$	n	0	$\mathbb{Z}$	quaternionic anti-Hermitian

Quantity	Value
$\mathfrak{su}(2)$ structure constants ($T_a = -\frac{i}{2}\sigma_a$)	$f_{abc} = \varepsilon_{abc}$
Killing form on $\mathfrak{su}(N)$	$B(X,Y) = 2N\,\mathrm{tr}(XY)$
Index and Casimirs of $\mathrm{SU}(N)$	$C(\mathbf{N}) = \frac{1}{2}$, $C_2(\mathbf{N}) = \frac{N^2-1}{2N}$, $C_2(\mathrm{adj}) = N$
Spin j	$\dim = 2j+1$, $C_2 = j(j+1)$
Volume of SU(2) in the metric $\sum(\theta^a)^2$	$16\pi^2$ (round S^3 of radius 2)
$\int_{\mathrm{SU}(2)} \mathrm{tr}(g^{-1}\mathrm{d}g)^3$	$-24\pi^2$
Instanton number normalisation	$k = \frac{1}{8\pi^2}\int \mathrm{tr}\, F \wedge F \in \mathbb{Z}$
Minimal instanton action	$S = 8\pi^2\lvert k\rvert/g_{\mathrm{YM}}^2$
Chern–Simons level	e^{iS_k} gauge invariant iff $k \in \mathbb{Z}$
Lattice coupling	$\beta = 2N/g_{\mathrm{YM}}^2$

Table T3.11: Dimensions, ranks and homotopy groups of the groups used in the book (top), and the constants that recur (bottom). The π_3 column is where instantons live: a finite-action field on $\mathbb{R}^4$ defines a map from S^3_∞ into G, and $\pi_3(G) = \mathbb{Z}$ is its winding number.

Concept	Operational meaning	Used later in
Lie algebra $\mathfrak{g}$	Linearise the defining equations at $\mathbb{1}$: write $g = \mathbb{1} + \varepsilon X$ and keep $O(\varepsilon)$. $\mathfrak{su}(n)$: traceless anti-Hermitian; $\mathfrak{so}(n)$: antisymmetric.	Values of gauge fields; Chapter 6
Commutator	$[X, Y] = XY - YX$, the $O(\varepsilon^2)$ part of the group commutator (T3.5); obeys Jacobi.	F contains $[A_\mu, A_\nu]$
Structure constants	$[T_a, T_b] = f_{ab}{}^c T_c$ (T3.6); for $\mathfrak{su}(2)$, $f_{abc} = \varepsilon_{abc}$.	Kac–Moody algebras, Chapter 6; $F^a_{\mu\nu}$
Maths vs physics generators	T_a anti-Hermitian, $J_a = iT_a$ Hermitian, $[J_a, J_b] = i f_{abc} J_c$ (Section T3.2.5).	Every comparison with QFT texts
Gell-Mann matrices	Eight Hermitian traceless 3×3 matrices, λ_3, λ_8 diagonal (Section T3.2.6).	QCD examples
Exponential map	$e^X = \sum X^k/k!$; e^{tX} solves $\dot{g} = Xg$; $\det e^X = e^{\operatorname{tr} X}$ (T3.14).	Gauge transformations $g = e^X$, path ordering
SU(2) exponential	$e^{-\frac{i\theta}{2}\hat{n}\cdot\vec{\sigma}} = \cos\frac{\theta}{2} - i \sin\frac{\theta}{2}\, \hat{n}\cdot\vec{\sigma}$ (T3.15); $\theta = 2\pi$ gives $-\mathbb{1}$.	Spinor sign under 2π, Chapter T4
Rodrigues formula	$e^{\theta K} = \mathbb{1} + \sin\theta\, K + (1 - \cos\theta) K^2$ for rotations (T3.17).	Explicit rotations
BCH	$e^X e^Y = e^{X+Y+\frac{1}{2}[X,Y]+\cdots}$ (T3.19): the bracket determines the product near $\mathbb{1}$.	Non-abelian Stokes, lattice
Ad, ad	$\mathrm{Ad}_g X = gXg^{-1}$; $\mathrm{ad}_Y X = [Y, X]$, matrix $(\mathrm{ad}_{T_a})_{cb} = f_{ab}{}^c$ (T3.21).	$F \mapsto g^{-1} F g$; adjoint bundle, Chapter 6
Killing form	$B(X, Y) = \operatorname{tr}(\mathrm{ad}_X \mathrm{ad}_Y) = 2N \operatorname{tr}(XY)$ on $\mathfrak{su}(N)$; invariant; negative definite $\Leftrightarrow$ compact semisimple (Section T3.4.3).	The metric in $\operatorname{tr} F \wedge \star F$
Double cover	$R_{ij} = \frac{1}{2}\operatorname{tr}(\sigma_i U \sigma_j U^\dagger)$ (T3.24); $2:1$, kernel $\pm\mathbb{1}$.	Spin groups, Chapter T4
$\pi_1(\mathrm{SO}(3)) = \mathbb{Z}_2$	The 2π rotation lifts to a path from $\mathbb{1}$ to $-\mathbb{1}$; 4π is contractible (belt trick, Section T3.4.5).	Spin structures, Chapter T4

Table T3.5: Lie algebras, the exponential map and the adjoint representation (Sections T3.2 to T3.4).

Concept	Operational meaning	Used later in
Representation	Matrices $r(g)$ with $r(gh) = r(g)r(h)$; generators R_a satisfy the same commutation relations as T_a.	Matter fields, twisted Dirac operators, Chapter 4
Irreducible, Schur	No invariant subspace; anything commuting with every $r(g)$ is a multiple of $\mathbb{1}$. Hence Casimirs are numbers on a multiplet.	Peter–Weyl, Borel–Weil, Chapter 6
Tensor product, conjugate	Generators add, $R \otimes \mathbb{1} + \mathbb{1} \otimes R'$ (T3.27); conjugate generators $-R^*$ (or R^* in anti-Hermitian form).	Coupling of multiplets
Spin j	J_3 eigenvalues $m = -j, \dots, j$; ladders $J_\pm$ with $\|J_\pm \|j, m\rangle\|^2 = (j \mp m)(j \pm m + 1)$ (T3.29); explicit matrices (T3.30).	Angular momentum, index of Dirac on S^2
Cartan subalgebra, rank	A maximal set of commuting generators (diagonal ones); rank = how many (Section T3.5.3).	Maximal torus T, G/T
Weights	Simultaneous eigenvalues of the Cartan generators on a representation; plot them.	Borel–Weil, character formulae
Roots	Weights of the adjoint; root vectors are ladder operators $[H_i, E_\alpha] = \alpha_i E_\alpha$.	Kac–Moody, Chapter 6
SU(3) diagrams	**3** triangle, $\bar{\mathbf{3}}$ inverted triangle, adjoint hexagon plus two zeros (Figure T3.13).	Quark model; anomaly cancellation
Simple roots, Dynkin	$2(\alpha, \beta)/(\alpha, \alpha) \in \mathbb{Z}$ allows only $90°, 120°, 135°, 150°$; classify by diagrams (Figure T3.15).	Choice of gauge group
Casimir and index	$C_2(R)\mathbb{1} = \sum_a R_a R_a$, $\operatorname{tr} R_a R_b = C(R)\delta_{ab}$, $C_2(R) \dim R = C(R) \dim G$ (T3.32).	Anomaly coefficients, β-functions, Chapter 10

Table T3.6: Representations (Section T3.5).

Concept	Operational meaning	Used later in
Maurer–Cartan form	$\theta = g^{-1}\mathrm{d}g$: differentiate the group-valued field and carry the result back to $\mathbb{1}$; a $\mathfrak{g}$-valued 1-form.	Pure gauge fields; WZW models, Chapter 11
Left invariance	$(hg)^{-1}\mathrm{d}(hg) = g^{-1}\mathrm{d}g$ for constant h (T3.33); right translation rotates it by $\mathrm{Ad}_{h^{-1}}$.	Global vs local symmetry
Structure equation	$\mathrm{d}\theta + \theta \wedge \theta = 0$ (T3.36), from $\mathrm{d}(g^{-1}) = -g^{-1}\mathrm{d}g\, g^{-1}$; in components $\mathrm{d}\theta^a = -\frac{1}{2} f_{bc}{}^a \theta^b \wedge \theta^c$.	"Pure gauge is flat"
Euler-angle forms	Explicit θ^a on SU(2) (T3.40); metric $\sum(\theta^a)^2$ is the round S^3 of radius 2.	Instanton computations
Winding number	$w(g) = -\frac{1}{24\pi^2}\int_{S^3} \mathrm{tr}(g^{-1}\mathrm{d}g)^3 \in \mathbb{Z}$ (T3.44); U(1): $\frac{1}{2\pi i}\oint g^{-1}\mathrm{d}g$.	Instanton number, CS level, Chapters 8 and 12
Group action	A rule $x \mapsto g \cdot x$ with $\mathbb{1} \cdot x = x$, $g \cdot (h \cdot x) = (gh) \cdot x$.	Gauge group on $\mathcal{A}$, Chapter 6
Orbit, stabiliser	Everything x_0 can reach; everything fixing it. $G \cdot x_0 \cong G/H$, $\dim = \dim G - \dim H$.	Reducible connections, moduli singularities
Homogeneous spaces	$S^{n-1} = \mathrm{SO}(n)/\mathrm{SO}(n-1)$, $S^2 = \mathrm{SU}(2)/\mathrm{U}(1) = \mathbb{CP}^1$ (Hopf map (T3.45)), $\mathbb{CP}^n = \mathrm{SU}(n+1)/\mathrm{U}(n)$, flags G/T.	Borel–Weil, σ-models, Chapters 5 and 6
Free actions	Every stabiliser trivial; then $G \to X \to X/G$ is a principal bundle.	Principal bundles, Chapter 1

Table T3.7: The Maurer–Cartan form, group actions and homogeneous spaces (Sections T3.6 and T3.7).

Concept	Operational meaning	Used later in
Local symmetry	$\psi \mapsto g^{-1}(x)\psi$; then $\mathrm{d}\psi$ picks up $-g^{-1}\mathrm{d}g\, g^{-1}\psi$ (T3.47) and is no longer covariant.	Motivation for all gauge theory
Comparator	$U(y,x)$ transports the internal frame from x to y; expanding it gives A (T3.49).	Lattice links; Wilson lines
Connection A	$A = A^a_\mu T_a\,\mathrm{d}x^\mu$, a $\mathfrak{g}$-valued 1-form; $D = \mathrm{d} + A$ (T3.51).	Every Nash chapter from 6 on
Gauge law	$A \mapsto g^{-1}Ag + g^{-1}\mathrm{d}g$ (T3.52); infinitesimally $\delta A = \mathrm{d}X + [A, X] = DX$ (T3.53).	Gauge orbits, Faddeev–Popov, BRST
Frame picture	A is the rate of rotation of a chosen basis; change of basis gives the same law (Section T3.8.4, after Dray).	Levi-Civita and spin connections, Chapters T2 and T4
Curvature F	$F = D^2 = \mathrm{d}A + A \wedge A$ (T3.56); $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu + [A_\mu, A_\nu]$; $F \mapsto g^{-1}Fg$.	Characteristic classes, Chapters 4 and 8
Monopole connection	$A_N = \frac{i}{2}(1-\cos\vartheta)\mathrm{d}\phi$, $A_S = -\frac{i}{2}(1+\cos\vartheta)\mathrm{d}\phi$, glued by $e^{-i\phi}$; $c_1 = -1$.	Hopf bundle, index on S^2
Bianchi identity	$DF = \mathrm{d}F + [A, F] = 0$ (T3.61), automatic.	Closedness of $\operatorname{tr} F^k$
Yang–Mills	$S = -\frac{1}{g^2}\int \operatorname{tr} F \wedge \star F$ (T3.62); $D \star F = 0$ (T3.65); self-dual fields solve it.	Instantons, monopoles, Chapter 8
Dictionary	$A_{\text{math}} = -ig_{\text{YM}}\mathcal{A}^a R_a$, $V = g^{-1}$ (T3.66) (Section T3.8.9).	Comparing with the QFT notes

Table T3.8: Gauge fields (Section T3.8).

Concept	Operational meaning	Used later in
Parallel transport	Solve $\dot{W} = -A(\dot\gamma)W$, $W(0) = \mathbb{1}$ (T3.67); $W_\gamma = \mathcal{P}\exp(-\int_\gamma A)$ (T3.71), later points on the left.	Nash §8.1; holonomy of moduli
Gauge covariance	$W_\gamma \mapsto g(b)^{-1}W_\gamma\, g(a)$ (T3.74): only the endpoints feel the gauge.	Gauge-invariant observables
Holonomy, Wilson loop	Transport round a closed loop; $W_R(\gamma) = \operatorname{tr}_R \mathrm{Hol}_\gamma$ (T3.75) is gauge invariant.	Chern–Simons TQFT, knot invariants, Chapter 12
Holonomy group	All holonomies at a point; the connection is irreducible if it is all of G.	Smooth points of the moduli space
Small loops	$\mathrm{Hol} = \mathbb{1} - \varepsilon\delta F_{12} + \cdots$ (T3.76): curvature is holonomy per unit area.	Lattice action; flatness test
Aharonov–Bohm	$F = 0$ on the electron's region, holonomy $e^{\mp ie\Phi}$ (T3.78); needs a non-contractible loop.	Flat connections, Chapter 12
Berry phase	Holonomy of the monopole connection round a loop on S^2: $-\Omega/2$.	Geometric phases, anomalies
Lattice gauge theory	Links $U_{x,\mu} \in G$, $S_W = \beta\sum_p(1 - \frac{1}{N}\operatorname{Re}\operatorname{tr} U_p) \to S_{\text{YM}}$, $\beta = 2N/g^2_{\text{YM}}$ (Section T3.9.8).	Non-perturbative intuition

Table T3.9: Parallel transport, holonomy and Wilson loops (Section T3.9).

Concept	Operational meaning	Used later in
Pure gauge	$A = g^{-1}\mathrm{d}g$ (T3.80); $F = 0$ is the Maurer–Cartan equation; $W_\gamma = g(b)^{-1}g(a)$.	Vacua, boundary conditions at ∞
Flat connection	$F = 0$. Locally pure gauge (construct g by transport); globally classified by $\mathrm{Hom}(\pi_1(M), G)/G$ (T3.83).	Floer theory, CS theory, Chapter 12
Flat on S^1, T^2	Conjugacy classes of one holonomy (circle; interval $[0, \frac{1}{2}]$ for SU(2)); commuting pairs on T^2 (the pillowcase, Figure T3.33).	Moduli spaces, Chapter 12
Chern–Simons form	$CS(A) = \mathrm{tr}(A\,\mathrm{d}A + \frac{2}{3}A^3)$ with $\mathrm{d}\,CS = \mathrm{tr}\,F \wedge F$ (T3.84); check by expanding both sides.	Instantons, anomalies, TQFT
Transgression	$CS = 2\int_0^1 \mathrm{tr}(A \wedge F_t)\,\mathrm{d}t$ with $A_t = tA$ (T3.94); same recipe gives CS_5; Chern–Weil.	Nash (8.39)–(8.45); Chapter 10
Gauge variation of CS	$CS(A^g) = CS(A) - \mathrm{d}\,\mathrm{tr}(\mathrm{d}g\,g^{-1}A) - \frac{1}{3}\mathrm{tr}(g^{-1}\mathrm{d}g)^3$ (T3.99); $\frac{1}{8\pi^2}\int CS$ shifts by $w(g)$.	Level quantisation
Instanton number	$k = \frac{1}{8\pi^2}\int_{\mathbb{R}^4} \mathrm{tr}\,F \wedge F = w(g_\infty)$ (T3.103) (Stokes plus pure gauge at infinity).	Chapters 4 and 8
CS number, θ-vacua	N_{CS} labels classical vacua; $k = \Delta N_{\mathrm{CS}}$ (T3.105); $\lvert\vartheta\rangle = \sum e^{in\vartheta}\lvert n\rangle$.	Strong CP, baryon number violation
3D CS action	$S_k = -\frac{k}{4\pi}\int CS$ (T3.106); $k \in \mathbb{Z}$; field equations $F = 0$; no metric; abelian case = linking.	Witten's knot theory, Chapter 12
Descent	$\mathrm{tr}\,F^2 = \mathrm{d}\,CS$, $\delta_X CS = \mathrm{d}\,\mathrm{tr}(X\mathrm{d}A)$ (T3.107): the bottom form is the anomaly.	Chapter 10

Table T3.10: Pure gauge, flat connections and the Chern–Simons form (Section T3.10).

Calculations to be able to do without looking

A toolkit chapter has done its job when its core calculations become reflexes. These are the ones the rest of the book assumes.

1. *Exponentiate a Pauli matrix.* Use $(\hat{n}\cdot\vec{\sigma})^2 = \mathbb{1}$ to split the series into even and odd parts: $e^{-\frac{i\theta}{2}\hat{n}\cdot\vec{\sigma}} = \cos\frac{\theta}{2}\,\mathbb{1} - i\sin\frac{\theta}{2}\,\hat{n}\cdot\vec{\sigma}$ (Section T3.3.4).

2. *Build the rotation of a spinor matrix.* $R_{ij} = \frac{1}{2}\mathrm{tr}(\sigma_i U \sigma_j U^\dagger)$; for $U = e^{-\frac{i\theta}{2}\sigma_3}$ this is the rotation by θ about z, and U and $-U$ give the same R (Section T3.4.4).

3. *Build spin j.* Start from the highest weight, lower with J_-, and use $\|J_-|j,m\rangle\|^2 = (j+m)(j-m+1)$ to fix the matrix entries (Section T3.5.2).

4. *Derive the gauge law.* Demand $D(g^{-1}\psi) = g^{-1}D\psi$ and solve for the new A: $A \mapsto g^{-1}Ag + g^{-1}\mathrm{d}g$ (Section T3.8).

5. *Compute F from D^2.* $D^2\psi = (\mathrm{d}A)\psi + A\wedge A\,\psi$, the $\mathrm{d}\psi$ terms cancelling by the graded Leibniz rule (Section T3.8.6).

6. *Check that pure gauge is flat.* $\mathrm{d}(g^{-1}\mathrm{d}g) = -g^{-1}\mathrm{d}g \wedge g^{-1}\mathrm{d}g$, which cancels $A\wedge A$ exactly (Section T3.10.1).

7. *Derive* $\mathrm{tr}\,F\wedge F = \mathrm{d}\,CS$. Expand both sides; use $\mathrm{dd} = 0$, graded cyclicity (T3.79) and $\mathrm{tr}\,A^4 = 0$; the coefficient $\frac{2}{3}$ is forced (Section T3.10.3).

8. *Turn $\int \mathrm{tr}\,F\wedge F$ into a winding number.* Stokes' theorem on a large ball, $A \to g^{-1}\mathrm{d}g$ on the boundary, and $CS(g^{-1}\mathrm{d}g) = -\frac{1}{3}\mathrm{tr}(g^{-1}\mathrm{d}g)^3$ (Section T3.10.6).

9. *Compute an Aharonov–Bohm phase.* Holonomy of a flat abelian connection round the solenoid is $e^{-\oint A}$, fixed by the enclosed flux (Section T3.9.6).

Where Nash uses this chapter

Nash	What he uses	Here
p. 25, (1.74)–(1.75)	Yang–Mills action $-\int \operatorname{tr} F \wedge \star F$, self-duality	Section T3.8
p. 87, (3.125)	Invariant polynomials $\det(t\mathbb{1} + \frac{iF}{2\pi})$; gauge invariance of $\operatorname{tr} F^k$	Sections T3.8 and T3.10.4
pp. 173–178	Lie algebras of $\mathrm{Map}(X, G)$, loop groups, $\mathrm{Diff}(X)$; gauge group = sections of $\mathrm{Ad}\, P$	Sections T3.2 and T3.7
p. 182, (6.39); (6.9)	Casimir operator; Kac–Moody brackets with if^{abc}	Sections T3.2.5 and T3.5.6
pp. 189–190, (6.66)–(6.68)	Borel–Weil: G/T, root spaces, highest weights	Sections T3.5 and T3.7
pp. 216–218, (8.1)–(8.10)	Space of connections, parallel transport, holonomy group, reducible connections	Sections T3.7 and T3.9
p. 219, (8.15)–(8.19); p. 227, (8.59)	$\pi_3(G) = \mathbb{Z}$, winding numbers, instanton sectors	Sections T3.6.5 and T3.10.6
pp. 223–224, (8.39)–(8.45)	Transgression $A_t = tA$, Chern–Simons form as a secondary class	Sections T3.10.4 and T3.10.5
p. 227, (8.60)–(8.62)	Chern–Simons level quantisation	Section T3.10.8
p. 228, (8.63)–(8.70)	Yang–Mills variation, $d_A^* F = 0$, Bianchi	Section T3.8
p. 244, (8.150)–(8.151)	Flat connections and π_1	Section T3.10.2
pp. 324–325, (12.4)–(12.10)	Floer theory: CS function, flat SU(2) connections	Section T3.10
pp. 342–348, (12.84)–(12.94)	Chern–Simons TQFT, Wilson loops, linking numbers	Sections T3.9 and T3.10.8

Table T3.12: Where Nash relies on this toolkit chapter, with the section of this chapter that supplies the background.

Recap: the chapter in five sentences

- A Lie group is a smooth family of matrices; its Lie algebra is the set of its velocities at $\mathbb{1}$, and exp goes back.
- Representations are the ways the group acts on fields; for SU(2) they are labelled by spin j, and in general by weights, whose differences are roots.
- $g^{-1}\mathrm{d}g$ is what a local symmetry does to a derivative; the gauge field A is introduced to cancel it, and transforms as $A \mapsto g^{-1}Ag + g^{-1}\mathrm{d}g$.
- The curvature $F = \mathrm{d}A + A \wedge A$ is the commutator of covariant derivatives and the holonomy per unit area; a flat connection sees only π_1.
- $\operatorname{tr} F \wedge F = \mathrm{d}\, CS(A)$ makes the instanton number a winding number, the Chern–Simons action a topological field theory, and its gauge variation the seed of anomalies.

Chapter T4

Clifford Algebras, Spinors and Dirac Operators

Where this chapter is going

Every fermion in the Standard Model is described by a field ψ that obeys a *first-order* wave equation, the Dirac equation. Dirac found it by asking a naive question: can we take the square root of the wave operator $\partial_t^2 - \nabla^2$? The answer is yes, but only if the coefficients are matrices that *anticommute*. That single requirement is the seed of everything in this chapter.

Here is the chain of ideas we will follow.

- The anticommutation rule $\{\gamma^\mu, \gamma^\nu\} = 2\eta^{\mu\nu}$ defines an algebra, the **Clifford algebra**. We never leave it abstract. In every dimension we write down the matrices (Pauli matrices, quaternions, the 4×4 Dirac and Weyl matrices, and tensor products of Pauli matrices in general) and check the relations by multiplying them out.
- Inside the Clifford algebra lives a group, $\mathrm{Spin}(n)$, which is a double cover of the rotation group. Rotating a spinor by 2π multiplies it by -1. This sign is the reason spinors are subtle on curved manifolds.
- Transporting spinors around a manifold requires lifting the rotation matrices that glue coordinate patches to $\mathrm{Spin}(n)$ consistently. When this is possible we have a **spin structure**. On a circle there are two (the Ramond and Neveu–Schwarz sectors of string theory); on $\mathbb{CP}^2$ there are none.
- With a spin structure and a metric we build the **Dirac operator** D. We compute its spectrum on $\mathbb{R}^n$, S^1, T^2 and S^2 by hand. The Lichnerowicz formula $D^2 = \nabla^*\nabla + R/4$ ties its zero modes to curvature.
- Splitting spinors by chirality turns D into a map $D^+ : S^+ \to S^-$. The number $\dim \ker D^+ - \dim \ker D^-$ is the *index*. It is the central object of Nash's Chapter 4 (Chapter 4) and the source of the anomalies of Chapter 10 (Chapter 10).
- Finally, differential forms themselves form a Clifford module, and $\mathrm{d} + \mathrm{d}^*$ is a Dirac-type operator. This is how de Rham theory, the signature and the Dolbeault operator all fit into one pattern.

Conventions used in this chapter

Signs and factors of i in the spinor literature are a notorious mess. Every source in the bibliography uses a different convention. We fix ours once and for all here, and give a

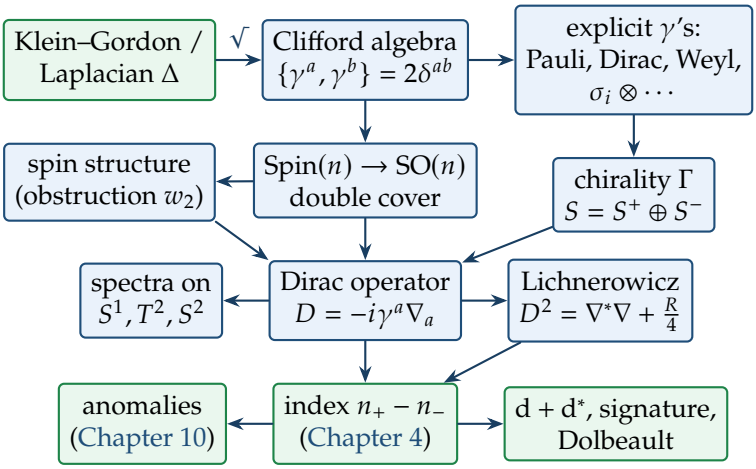


Figure T4.1: The logic of this chapter. Green boxes are the physics that goes in and comes out; blue boxes are the tools we build. Read it top to bottom: square-rooting the Laplacian forces the Clifford algebra; the algebra contains the spin group and chirality; spin structures let us transport spinors on a manifold; the Dirac operator and its index are the payoff.

translation table.

Dictionary: our conventions and the sources'

- **Euclidean gamma matrices (our default).** $\gamma_1, \dots, \gamma_n$ are *Hermitian* matrices with

$$\gamma_a \gamma_b + \gamma_b \gamma_a = 2\delta_{ab}\,\mathbf{1}, \qquad \text{so } \gamma_a^2 = \mathbf{1}.$$

  Indices $a, b, c, \dots$ refer to an orthonormal frame; they are raised and lowered with δ_{ab}, so $\gamma^a = \gamma_a$.

- **Dirac operator.** $D = -i\gamma^a \nabla_a$ (sum over a). It is formally self-adjoint and $D^2 = -\nabla^a \nabla_a + \dots$ is a *positive* Laplacian.

- **Mathematicians' Clifford multiplication** (Bär). Bär writes $e_a \cdot e_a = -1$ and $D = \sum_a e_a \cdot \nabla_{e_a}$. The translation is $c(e_a) = -i\gamma_a$: indeed $(-i\gamma_a)^2 = -\mathbf{1}$. With this dictionary Bär's D is exactly our D.

- **Nash** [Nash p. 114, (4.99)] uses anti-Hermitian matrices with $\{\gamma_a, \gamma_b\} = -2\delta_{ab}$, i.e. $\gamma_a^{\text{Nash}} = i\gamma_a^{\text{ours}}$, and $\not{D} = \gamma^a_{\text{Nash}}(\partial_a + \omega_a)$, which is $-D$ in our notation. Spectra are symmetric in all our examples, so nothing physical changes.

- **Minkowski space** (following the QFT notes [QFT notes §3.5]). $\eta^{\mu\nu} = \text{diag}(+1, -1, -1, -1)$, $\{\gamma^\mu, \gamma^\nu\} = 2\eta^{\mu\nu}$, γ^0 Hermitian, γ^i anti-Hermitian, $\gamma^5 = i\gamma^0\gamma^1\gamma^2\gamma^3$, Dirac equation $(i\gamma^\mu \partial_\mu - m)\psi = 0$.

- **Clifford algebras.** $\text{Cl}_{p,q}$ is generated by p vectors squaring to $+1$ and q vectors squaring to -1. Bär's Cl_n is our $\text{Cl}_{0,n}$. (Warning: Lawson–Michelsohn use the opposite order.)

- **Gauge fields** follow the house convention: A is anti-Hermitian, $\nabla = \mathrm{d} + A$, $F = \mathrm{d}A + A \wedge A$. The physicists' Hermitian field is a with $A = -ia$ for U(1) (and $A = -ig\, A^a_{\text{phys}} T^a$ in general).

T4.1 Dirac's square root of the Laplacian

In 1928 Dirac wanted a relativistic wave equation that was *first order* in time, like the Schrödinger equation. The obvious relativistic equation, the Klein–Gordon equation $(\partial_t^2 - \nabla^2 + m^2)\phi = 0$, is second order. So he asked: is there a first-order operator whose square is the Klein–Gordon operator? This section answers that question. The answer forces on us a new kind of algebra, and it is worth seeing it appear with no input except "take a square root".

T4.1.1 Trying to factorise a second-order operator

We start in Euclidean space $\mathbb{R}^n$ with coordinates $x^1, \dots, x^n$, because there the signs are simplest. The (positive) Laplacian is

$$\Delta := -\sum_{a=1}^{n} \partial_a \partial_a = -\partial_1^2 - \cdots - \partial_n^2 .$$

On a plane wave $e^{ik\cdot x}$ it acts as multiplication by $|k|^2 \geq 0$. We look for a first-order operator with constant coefficients,

$$D = -i \sum_a \gamma^a \partial_a \qquad \text{(coefficients } \gamma^a \text{ to be found)},$$

such that $D^2 = \Delta$. (The factor $-i$ is only there to make D Hermitian later; it plays no role in the algebra.) Square it:

$$\begin{aligned} D^2 &\overset{(1)}{=} (-i)^2 \sum_{a,b} \gamma^a \gamma^b \, \partial_a \partial_b \\ &\overset{(2)}{=} -\frac{1}{2} \sum_{a,b} \left(\gamma^a \gamma^b + \gamma^b \gamma^a \right) \partial_a \partial_b \\ &\overset{(3)}{=} -\sum_a (\gamma^a)^2 \partial_a^2 - \sum_{a<b} \left(\gamma^a \gamma^b + \gamma^b \gamma^a \right) \partial_a \partial_b . \end{aligned} \tag{T4.1}$$

(1) The coefficients are constants, so they pass through the derivatives; we just multiply out the two sums.

(2) Partial derivatives commute, $\partial_a \partial_b = \partial_b \partial_a$, so only the symmetric part of $\gamma^a \gamma^b$ survives the double sum.

(3) Split the double sum into diagonal terms $a = b$ and off-diagonal pairs $a < b$ (each unordered pair appears twice in the sum, which cancels the $\frac{1}{2}$).

For this to equal $-\sum_a \partial_a^2$ *for every function*, the coefficients of ∂_a^2 and of $\partial_a \partial_b$ must match:

$$(\gamma^a)^2 = 1, \qquad \gamma^a \gamma^b + \gamma^b \gamma^a = 0 \ (a \neq b), \qquad \text{i.e.} \qquad \boxed{\{\gamma^a, \gamma^b\} = 2\delta^{ab}} \tag{T4.2}$$

where $\{A, B\} := AB + BA$ is the **anticommutator**.

Key idea: the square root forces anticommutation

A first-order operator $-i\gamma^a \partial_a$ squares to the Laplacian if and only if the coefficients satisfy $\{\gamma^a, \gamma^b\} = 2\delta^{ab}$. For $n \geq 2$ ordinary numbers cannot do this: if $\gamma^1 \gamma^2 + \gamma^2 \gamma^1 = 2\gamma^1 \gamma^2 = 0$ then one of them vanishes, contradicting $(\gamma^a)^2 = 1$. The coefficients must be *matrices*, and the wave function they act on must have several components.

Now repeat the computation in Minkowski space. We use the metric of the QFT notes, $\eta^{\mu\nu} = \mathrm{diag}(+1, -1, -1, -1)$, and the d'Alembertian

$$\partial^2 = \eta^{\mu\nu}\partial_\mu\partial_\nu = \partial_t^2 - \nabla^2,$$

gives the Minkowski version. Suppose $\{\gamma^\mu, \gamma^\nu\} = 2\eta^{\mu\nu}$. Then

$$\begin{aligned}(-i\gamma^\mu\partial_\mu - m)(i\gamma^\nu\partial_\nu - m)\psi &\overset{(1)}{=} \left(\gamma^\mu\gamma^\nu\partial_\mu\partial_\nu + im\gamma^\mu\partial_\mu - im\gamma^\nu\partial_\nu + m^2\right)\psi \\ &\overset{(2)}{=} \left(\tfrac{1}{2}\{\gamma^\mu, \gamma^\nu\}\partial_\mu\partial_\nu + m^2\right)\psi \\ &\overset{(3)}{=} (\partial^2 + m^2)\psi. \end{aligned} \tag{T4.3}$$

(1) Multiply out; $(-i)(i) = 1$ and the two cross terms carry $+im$ and $-im$.
(2) The two cross terms are the same operator with a renamed dummy index, so they cancel; and only the symmetric part of $\gamma^\mu\gamma^\nu$ survives against $\partial_\mu\partial_\nu$.
(3) Use $\{\gamma^\mu, \gamma^\nu\} = 2\eta^{\mu\nu}$.

So every solution of the **Dirac equation** $(i\gamma^\mu\partial_\mu - m)\psi = 0$ solves the Klein–Gordon equation component by component [QFT notes (3.5.30)]. The Dirac equation is a first-order "square root" of Klein–Gordon.

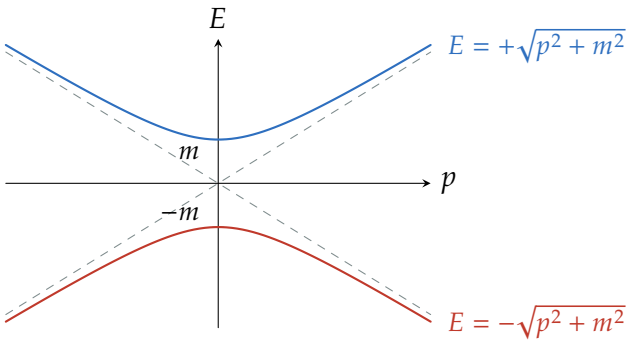


Figure T4.2: The Klein–Gordon dispersion relation $E^2 = p^2 + m^2$ has two branches. A first-order equation $E\psi = (\alpha \cdot p + \beta m)\psi$ reproduces both only if $(\alpha \cdot p + \beta m)^2 = p^2 + m^2$ as a *matrix* identity; its eigenvalues are then exactly $\pm\sqrt{p^2 + m^2}$. The negative branch (red) is the one Dirac later reinterpreted as antiparticles.

T4.1.2 Solving the anticommutation relations in low dimensions

Now we find the matrices. The goal is the smallest matrices that work.

Dimension 1. Take $\gamma^1 = 1$. Then $D = -i\,\mathrm{d}/\mathrm{d}x$ and $D^2 = -\mathrm{d}^2/\mathrm{d}x^2$. Nothing new: a first derivative is already the square root of a second derivative.

Dimension 2. We need two anticommuting matrices, each squaring to $\mathbf{1}$. The **Pauli matrices**

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \qquad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{T4.4}$$

will do. Let us multiply them out once and for all.

$$\sigma_1\sigma_2 \overset{(1)}{=} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix} \overset{(2)}{=} i\sigma_3,$$
$$\sigma_2\sigma_1 \overset{(3)}{=} \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} -i & 0 \\ 0 & i \end{pmatrix} = -i\sigma_3. \tag{T4.5}$$

(1) Row times column: the $(1,1)$ entry is $0\cdot 0 + 1\cdot i = i$, the $(2,2)$ entry is $1\cdot(-i) + 0\cdot 0 = -i$, and the off-diagonal entries vanish.
(2) Compare with $\sigma_3 = \operatorname{diag}(1,-1)$.
(3) Same computation in the other order.

So $\sigma_1\sigma_2 + \sigma_2\sigma_1 = 0$. The same kind of multiplication gives the complete set of Pauli relations,

$$\sigma_j\sigma_k = \delta_{jk}\mathbf{1} + i\,\varepsilon_{jkl}\,\sigma_l, \qquad \{\sigma_j,\sigma_k\} = 2\delta_{jk}, \qquad [\sigma_j,\sigma_k] = 2i\varepsilon_{jkl}\sigma_l. \tag{T4.6}$$

In two dimensions we take $\gamma^1 = \sigma_1$, $\gamma^2 = \sigma_2$.

Worked example T4.1: the 2D Dirac operator is the Cauchy–Riemann operator

With $\gamma^1 = \sigma_1$, $\gamma^2 = \sigma_2$ and the complex coordinate $z = x^1 + ix^2$, put $\partial_z = \tfrac{1}{2}(\partial_1 - i\partial_2)$ and $\partial_{\bar z} = \tfrac{1}{2}(\partial_1 + i\partial_2)$. Then

$$D \overset{(1)}{=} -i\begin{pmatrix} 0 & \partial_1 \\ \partial_1 & 0 \end{pmatrix} - i\begin{pmatrix} 0 & -i\partial_2 \\ i\partial_2 & 0 \end{pmatrix}$$
$$\overset{(2)}{=} -i\begin{pmatrix} 0 & \partial_1 - i\partial_2 \\ \partial_1 + i\partial_2 & 0 \end{pmatrix}$$
$$\overset{(3)}{=} \begin{pmatrix} 0 & -2i\,\partial_z \\ -2i\,\partial_{\bar z} & 0 \end{pmatrix}.$$

(1) Insert the Pauli matrices into $D = -i(\sigma_1\partial_1 + \sigma_2\partial_2)$.
(2) Add the two matrices entry by entry.
(3) Recognise $\partial_1 \mp i\partial_2 = 2\partial_z,\ 2\partial_{\bar z}$.

Squaring,

$$D^2 \overset{(1)}{=} \begin{pmatrix} (-2i)^2\partial_z\partial_{\bar z} & 0 \\ 0 & (-2i)^2\partial_{\bar z}\partial_z \end{pmatrix} \overset{(2)}{=} -4\partial_z\partial_{\bar z}\,\mathbf{1} \overset{(3)}{=} -(\partial_1^2 + \partial_2^2)\,\mathbf{1}.$$

(1) The product of two off-diagonal matrices is diagonal: $(1,1)$ entry is $(-2i\partial_z)(-2i\partial_{\bar z})$ and so on.
(2) $(-2i)^2 = -4$, and $\partial_z, \partial_{\bar z}$ commute.
(3) $4\partial_z\partial_{\bar z} = (\partial_1 - i\partial_2)(\partial_1 + i\partial_2) = \partial_1^2 + \partial_2^2$.

So $D^2 = \Delta$ as promised. And the equation $D\psi = 0$ for $\psi = (\psi_+, \psi_-)^{\mathsf T}$ says $\partial_{\bar z}\psi_+ = 0$ and $\partial_z\psi_- = 0$: the upper component is *holomorphic* and the lower one *antiholomorphic*. Zero modes of the 2D Dirac operator are solutions of the Cauchy–Riemann equations. We will meet this again when we count zero modes on surfaces.

Dimension 3. Now all three Pauli matrices anticommute, so $\gamma^a = \sigma_a$, $a = 1,2,3$, works. This is the familiar operator $\sigma\cdot p$ of the Pauli equation.

Worked example T4.2: the Pauli identity and $(\sigma \cdot p)^2 = p^2$

For any two vectors $\vec{a}, \vec{b}$ (whose components commute with the σ's),

$$\begin{aligned}(\sigma \cdot \vec{a})(\sigma \cdot \vec{b}) &\overset{(1)}{=} \sum_{j,k} a_j b_k \, \sigma_j \sigma_k \\ &\overset{(2)}{=} \sum_{j,k} a_j b_k \left(\delta_{jk} + i \varepsilon_{jkl} \sigma_l\right) \\ &\overset{(3)}{=} \vec{a} \cdot \vec{b} + i \, \sigma \cdot (\vec{a} \times \vec{b}).\end{aligned}$$

(1) Expand both sums.
(2) Use the product rule (T4.6).
(3) $\sum_{j,k} \varepsilon_{jkl} a_j b_k = (\vec{a} \times \vec{b})_l$.

With $\vec{a} = \vec{b} = \vec{p}$ the cross product vanishes, so $(\sigma \cdot \vec{p})^2 = |\vec{p}|^2$: the operator $-i\sigma \cdot \nabla$ is a square root of the three-dimensional Laplacian. With $\vec{p} \to \vec{p} - e\vec{A}$ the cross product no longer vanishes, because the components of $\vec{p} - e\vec{A}$ do not commute; one finds $(\sigma \cdot \vec{\pi})^2 = \vec{\pi}^2 - e\hbar \, \sigma \cdot \vec{B}$, the Pauli Hamiltonian with $g = 2$. This is our first "twisted" Dirac operator; we return to it in Section T4.8.

Dimension 4: the 2×2 matrices run out. In four dimensions (Euclidean or Minkowski) we need *four* mutually anticommuting matrices. Can they still be 2×2? No. Any 2×2 matrix can be written $M = m_0 \mathbf{1} + \vec{m} \cdot \sigma$. If M anticommutes with all three Pauli matrices, then

$$0 = \{\sigma_j, M\} \overset{(1)}{=} 2m_0 \sigma_j + \sum_k m_k \{\sigma_j, \sigma_k\} \overset{(2)}{=} 2m_0 \sigma_j + 2m_j \mathbf{1},$$

(1) σ_j commutes with $m_0 \mathbf{1}$, so $\{\sigma_j, m_0 \mathbf{1}\} = 2m_0 \sigma_j$.
(2) $\{\sigma_j, \sigma_k\} = 2\delta_{jk}$.

and since $\mathbf{1}, \sigma_j$ are linearly independent, $m_0 = m_j = 0$. So at most three 2×2 matrices can anticommute pairwise. There is also a size argument valid in any dimension $n \geq 2$: each γ^a is traceless, because

$$\operatorname{tr} \gamma^1 \overset{(1)}{=} \operatorname{tr}(\gamma^1 \gamma^2 \gamma^2) \overset{(2)}{=} \operatorname{tr}(\gamma^2 \gamma^1 \gamma^2) \overset{(3)}{=} -\operatorname{tr}(\gamma^1 \gamma^2 \gamma^2) = -\operatorname{tr} \gamma^1,$$

(1) Insert $(\gamma^2)^2 = \mathbf{1}$.
(2) Cyclicity of the trace moves the last γ^2 to the front.
(3) Anticommute $\gamma^2 \gamma^1 = -\gamma^1 \gamma^2$.

and $(\gamma^a)^2 = \mathbf{1}$ means its eigenvalues are ± 1. Traceless with eigenvalues ± 1 forces equally many $+1$'s and -1's, so the matrix size is *even*. In four dimensions the smallest size turns out to be 4×4; Dirac found the matrices, and in Section T4.3 we write them down in two standard forms. The general pattern (Section T4.4) is that in n dimensions the smallest size is $2^{\lfloor n/2 \rfloor}$.

Watch out: why not just use $\sqrt{\Delta}$?

One can define $\sqrt{\Delta}$ by Fourier transform: it multiplies $e^{ik \cdot x}$ by $|k|$. But $|k| = \sqrt{k_1^2 + \cdots + k_n^2}$ is not a polynomial in k, so $\sqrt{\Delta}$ is not a differential operator; it is non-local (its kernel spreads over all of space). Dirac's trick makes $|k|$ polynomial at the price of matrices:

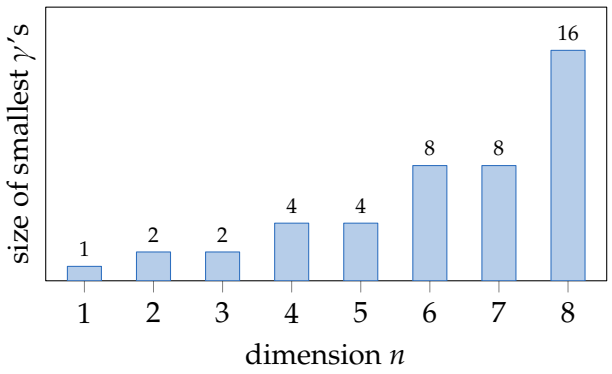


Figure T4.3: How big the gamma matrices must be. Each time we add *two* dimensions the size doubles: $2^{\lfloor n/2 \rfloor}$. In odd dimension the extra matrix is "free": it is the product of all the others (for $n = 3$, $\sigma_3 \propto \sigma_1\sigma_2$). This staircase is why Dirac spinors in 4D have four components.

$(\gamma^a k_a)^2 = |k|^2$, and $\gamma^a k_a$ is linear in k.

T4.1.3 Symbols: Laplace-type and Dirac-type operators

On a curved manifold, or with a gauge field, the operators we care about have variable coefficients and lower-order terms. What survives of Dirac's computation is the *leading* part. We need a way to talk about it.

Working definition T4.1 (principal symbol)

Let $P = \sum_{|\alpha| \le k} A^\alpha(x)\,\partial_\alpha$ [α lists which partial derivatives are taken, and $|\alpha|$ counts them, so $\partial_\alpha = \partial_1\partial_1\partial_3$ has $|\alpha| = 3$] be a differential operator of order k acting on vector-valued functions, with matrix coefficients A^α. Its **principal symbol** at the covector $\xi = \xi_\mu \mathrm{d}x^\mu$ [one number ξ_μ per coordinate, playing the role of the wave vector k_μ of $e^{ik\cdot x}$] is obtained by keeping only the k-th order terms and replacing each ∂_μ by $i\xi_\mu$:

$$\sigma_P(x,\xi) = \sum_{|\alpha|=k} A^\alpha(x)\,(i\xi)^\alpha .$$

Operationally: it is the matrix you get by letting P act on a rapidly oscillating wave $e^{i\lambda\,\xi\cdot x}u$ and keeping the leading power λ^k. It is a matrix-valued polynomial in ξ, and it changes under coordinate changes like a function of a covector.

Worked example T4.3: symbols of the basic operators

- Laplacian on a Riemannian manifold, $\Delta = -g^{\mu\nu}\partial_\mu\partial_\nu$ + (first order):

$$\sigma_\Delta(\xi) \overset{(1)}{=} -g^{\mu\nu}(i\xi_\mu)(i\xi_\nu) \overset{(2)}{=} g^{\mu\nu}\xi_\mu\xi_\nu = |\xi|^2 .$$

  (1) Replace each ∂ by $i\xi$; the first-order terms drop.
  (2) $i^2 = -1$ cancels the minus sign.

- Our Dirac operator $D = -i\gamma^a\partial_a$ on $\mathbb{R}^n$: $\sigma_D(\xi) = -i\gamma^a(i\xi_a) = \gamma^a\xi_a =: \not{\xi}$.
- The exterior derivative on functions, $\mathrm{d}f = \partial_\mu f\,\mathrm{d}x^\mu$: $\sigma_{\mathrm{d}}(\xi)f = i\xi\, f$, i.e. "multiply by $i\xi$". On p-forms, $\sigma_{\mathrm{d}}(\xi)\alpha = i\,\xi\wedge\alpha$.

The whole content of (T4.1) is then a statement about symbols: $\sigma_{D^2} = \sigma_D^2$ (symbols of compositions multiply), so D^2 has the symbol of the Laplacian exactly when $\not{\xi}^2 = |\xi|^2$ for all ξ. This suggests the following definitions, which are Bär's [Bär Def. 1.3.1, 1.3.7] (Bär omits the factors of i in the symbol, so his Laplace-type condition reads $\sigma_2 = -|\xi|^2$).

Working definition T4.2 (Laplace-type and Dirac-type)

A second-order operator P on sections of a vector bundle E [fields with several components at each point, such as spinor fields or forms] is of **Laplace type** if $\sigma_P(\xi) = |\xi|^2\,\mathbf{1}_E$. In local coordinates: $P = -g^{\mu\nu}\partial_\mu\partial_\nu\,\mathbf{1} +$ (first and zeroth order).
A first-order operator $D : \Gamma(E) \to \Gamma(F)$ is of **Dirac type** if both D^*D and DD^* are of Laplace type [D^* is the adjoint, fixed by $\langle D\psi, \chi\rangle = \langle \psi, D^*\chi\rangle$ for all ψ, χ]. If $E = F$ and $D = D^*$ this just says D^2 is of Laplace type [Bär Rem. 1.3.8].
Test: compute the symbol $\sigma_D(\xi)$, a matrix linear in ξ, and check whether $\sigma_D(\xi)^\dagger\sigma_D(\xi) = |\xi|^2$.

Now comes the key observation [Bär Rem. 1.3.14]. Let D be a formally self-adjoint Dirac-type operator. Then $\sigma_D(\xi)^2 = |\xi|^2$ for every ξ. Apply this to $\xi + \eta$:

$$\begin{aligned} |\xi|^2 + 2\langle\xi,\eta\rangle + |\eta|^2 &\overset{(1)}{=} |\xi+\eta|^2 \\ &\overset{(2)}{=} \big(\sigma_D(\xi) + \sigma_D(\eta)\big)^2 \\ &\overset{(3)}{=} |\xi|^2 + \sigma_D(\xi)\sigma_D(\eta) + \sigma_D(\eta)\sigma_D(\xi) + |\eta|^2. \end{aligned} \tag{T4.7}$$

(1) Expand the squared norm of a sum.
(2) D is of Dirac type and σ_D is linear in ξ.
(3) Expand the square of a sum of (non-commuting) matrices and use the Dirac-type condition on each of the two squares.

Comparing both sides:

$$\sigma_D(\xi)\sigma_D(\eta) + \sigma_D(\eta)\sigma_D(\xi) = 2\langle\xi,\eta\rangle\,\mathbf{1}. \tag{T4.8}$$

Key idea: every Dirac-type operator carries a Clifford algebra

At each point x, the symbol matrices $\sigma_D(\xi)$ of a self-adjoint Dirac-type operator satisfy the same anticommutation relations as Dirac's γ's. So the fibre E_x is automatically a representation space of the Clifford algebra of T^*_xM. Clifford algebras are not an optional decoration: they are forced on us the moment we ask for a first-order square root of a Laplacian. This is Bär's (1.18), and it is why "it is much more difficult to construct Dirac-type operators than Laplace-type operators".

Worked example T4.4: the gradient is *not* of Dirac type

Take $D = \mathrm{grad}$ from functions to vector fields on $\mathbb{R}^n$. Its adjoint is $D^* = -\mathrm{div}$ (integrate by parts). Then $D^*D = -\mathrm{div}\,\mathrm{grad} = \Delta$, which is of Laplace type. But $DD^* = -\mathrm{grad}\,\mathrm{div}$ acts on a vector field v as

$$(-\mathrm{grad}\,\mathrm{div}\,v)_j \overset{(1)}{=} -\partial_j \sum_k \partial_k v_k, \qquad \sigma_{DD^*}(\xi)_{jk} \overset{(2)}{=} -(i\xi_j)(i\xi_k) = \xi_j\xi_k.$$

(1) Definitions of divergence and gradient.
(2) Replace $\partial \to i\xi$.

The matrix $\xi\xi^{\mathsf T}$ is a projector onto ξ times $|\xi|^2$; it is not $|\xi|^2\mathbf{1}$ (it kills vectors perpendicular to ξ). So grad fails the test [Bär Ex. 1.3.9]. The fix, as we will see in Section T4.9, is to enlarge

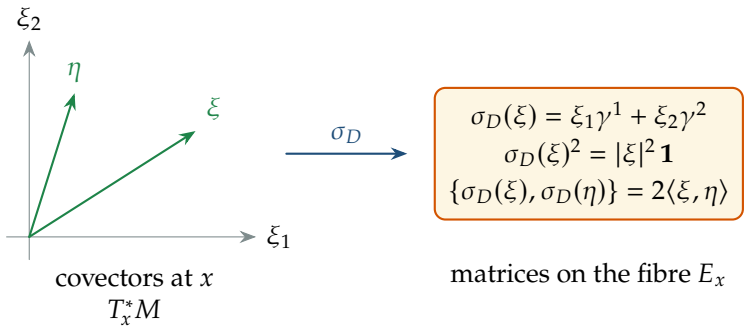


Figure T4.4: The principal symbol of a Dirac-type operator sends each covector ξ to a matrix whose square is $|\xi|^2$. Polarising (feeding in $\xi + \eta$) turns this into the Clifford relation, so the fibre is a Clifford module.

the bundles: $\mathrm{d} + \mathrm{d}^*$ acting on *all* forms is of Dirac type.

Worked example T4.5: Dray's Kähler operator

Dray [Dray §20.3, (20.23)–(20.25)] considers $\mathrm{d} - \delta$ acting on formal sums of forms, where $\delta = (-1)^p \star^{-1} \mathrm{d} \star$ is (minus) the codifferential. Since $\mathrm{d}^2 = 0$ and $\delta^2 = 0$,

$$(\mathrm{d} - \delta)^2 \overset{(1)}{=} \mathrm{d}^2 - \mathrm{d}\delta - \delta\mathrm{d} + \delta^2 \overset{(2)}{=} -(\mathrm{d}\delta + \delta\mathrm{d}),$$

(1) Expand the square keeping the order of the operators.
(2) $\mathrm{d}^2 = 0 = \delta^2$.

and on a function f this is $-\delta\mathrm{d}f$, which is the Laplacian of f (Dray's sign conventions make it the positive or negative Laplacian depending on signature). So Dray's $\mathrm{d} - \delta$ is "a sort of square root of the Laplacian" built without any matrices, because the forms themselves provide the anticommuting objects: $\mathrm{d}x^1 \wedge \mathrm{d}x^2 = -\mathrm{d}x^2 \wedge \mathrm{d}x^1$. We make this precise in Section T4.9.

Try it: a first-order root of the harmonic oscillator

Show that $Q = \sigma_1 p + \sigma_2 x$, with $p = -i\,\mathrm{d}/\mathrm{d}x$, satisfies $Q^2 = p^2 + x^2 + \sigma_3$. *Answer:* $Q^2 = \sigma_1^2 p^2 + \sigma_2^2 x^2 + \sigma_1\sigma_2\, px + \sigma_2\sigma_1\, xp = p^2 + x^2 + i\sigma_3[p, x] = p^2 + x^2 + \sigma_3$, using $[p, x] = -i$. The extra term is the "spin" part; this is the simplest supersymmetric quantum mechanics, and we meet it again with Witten's approach to Morse theory (Chapter 7).

Reading Nash (p. 114–115, (4.99)–(4.100))

Nash introduces the Dirac matrices in one sentence: "a set of n anti-commuting $2^{n/2} \times 2^{n/2}$ matrices which generate the Clifford algebra". His relation is $\gamma_a\gamma_b + \gamma_b\gamma_a = -2\delta_{ab}$ (anti-Hermitian γ's; ours times i). On p. 115 he notes that the Dirac operator is elliptic "because its symbol is given locally by the matrix $\sigma_1(x, p) = i\gamma^a e_a^\mu p_\mu$ and the inverse of this matrix is $i\gamma^a e_a^\mu p_\mu / p^2$". That is exactly the statement $\sigma_D(\xi)^2 = |\xi|^2$ above (with his signs), and it is the reason for the whole construction.

Recap

- Demanding $(-i\gamma^a\partial_a)^2 = \Delta$ forces $\{\gamma^a, \gamma^b\} = 2\delta^{ab}$; in Minkowski space $\{\gamma^\mu, \gamma^\nu\} = 2\eta^{\mu\nu}$ and Dirac $\Rightarrow$ Klein–Gordon.
- Solutions: 1 in 1D; σ_1, σ_2 in 2D (the Cauchy–Riemann operator); $\sigma_1, \sigma_2, \sigma_3$ in 3D (the Pauli operator); 4×4 matrices in 4D. Sizes grow like $2^{\lfloor n/2 \rfloor}$.
- The principal symbol of a Dirac-type operator satisfies the Clifford relation, so every Dirac-type operator acts on a Clifford module. The gradient alone is not of Dirac type; $\mathrm{d} + \mathrm{d}^*$ is.

T4.2 Clifford algebras: multiplying anticommuting directions

Once we have matrices $\gamma^1, \dots, \gamma^n$ obeying $\{\gamma^a, \gamma^b\} = 2\delta^{ab}$, we can multiply them together: $\gamma^1\gamma^2$, $\gamma^1\gamma^2\gamma^3$, and so on. These products are not decorations. $\gamma^1\gamma^2$ will turn out to generate rotations of spinors in the 12-plane, and the product of all the γ's will be the chirality operator γ^5. The set of all linear combinations of all products is an algebra, the Clifford algebra. In this section we find out what it looks like, dimension by dimension, and we always identify it with a concrete algebra of matrices.

T4.2.1 The working definition

Working definition T4.3 (Clifford algebra $\mathrm{Cl}_{p,q}$)

Let $n = p + q$. The **Clifford algebra** $\mathrm{Cl}_{p,q}$ is the real algebra generated by a unit 1 and n symbols $e_1, \dots, e_n$ [its elements are real linear combinations of products $e_{i_1}e_{i_2}\cdots$, multiplied associatively] subject only to

$$e_ie_j + e_je_i = 2\eta_{ij}, \qquad \eta = \mathrm{diag}(\underbrace{+1, \dots, +1}_{p}, \underbrace{-1, \dots, -1}_{q}). \tag{T4.9}$$

So $e_i^2 = +1$ for the first p generators, $e_i^2 = -1$ for the last q, and distinct generators anticommute. *Operationally*: an element of $\mathrm{Cl}_{p,q}$ is a polynomial in the e_i; to simplify a product, reorder the factors using $e_je_i = -e_ie_j$ ($i \neq j$) and replace e_i^2 by ± 1. A **representation** (or Clifford module) is a choice of matrices γ_i [square matrices, all of the same size, real or complex] with $\gamma_i\gamma_j + \gamma_j\gamma_i = 2\eta_{ij}\mathbf{1}$.

Two remarks make the definition less arbitrary.

(1) It is really a statement about all vectors. Put $v = v^ie_i$ for any real numbers v^i. Then

$$v\,v \overset{(1)}{=} \sum_{i,j} v^iv^je_ie_j \overset{(2)}{=} \frac{1}{2}\sum_{i,j} v^iv^j(e_ie_j + e_je_i) \overset{(3)}{=} \sum_{i,j}\eta_{ij}v^iv^j =: Q(v). \tag{T4.10}$$

(1) Multiply out; the v^i are numbers and commute with everything.
(2) v^iv^j is symmetric in i, j, so only the symmetric part of e_ie_j contributes.
(3) The relation (T4.9).

So "the square of a vector is its length squared" (with signature η). Conversely, if $v\,v = Q(v)$ for all v, then applying it to $v + w$ and subtracting the squares gives $vw + wv = 2\langle v, w\rangle_\eta$. This

is the same polarisation trick as (T4.7). Bär's definition [Bär Def. 2.1.1] is exactly this, with the opposite sign: $v \cdot v = -\beta(v,v)$.

(2) "Only" these relations. The phrase "subject only to" means we impose no further relations: the algebra is as big as it can be. (Formally, $\mathrm{Cl}(V,Q)$ is the tensor algebra $\bigoplus_k V^{\otimes k}$ divided by the ideal generated by $v \otimes v - Q(v)1$; equivalently it is characterised by a universal property: any linear map $f : V \to A$ into an algebra with $f(v)^2 = Q(v)$ extends uniquely to an algebra map $\mathrm{Cl}(V,Q) \to A$ [Bär Prop. 2.1.5].) A representation may be smaller than the algebra, as we will see with $\mathrm{Cl}_{0,3}$ below.

Watch out: which sign?

Physicists write $\{\gamma^a, \gamma^b\} = 2\delta^{ab}$ (vectors square to $+1$, that is $\mathrm{Cl}_{n,0}$); many mathematicians, including Bär and Nash, use $\{e_a, e_b\} = -2\delta_{ab}$ (vectors square to -1, $\mathrm{Cl}_{0,n}$). Over the real numbers these are *different algebras* (we will see $\mathrm{Cl}_{1,0} \neq \mathrm{Cl}_{0,1}$ and $\mathrm{Cl}_{2,0} \neq \mathrm{Cl}_{0,2}$). Over the complex numbers they are the same: if $\gamma_a^2 = 1$ then $(i\gamma_a)^2 = -1$. Since spinors are complex in almost all of physics, the sign is usually a harmless convention, but it matters for real (Majorana) spinors.

T4.2.2 Basis and dimension: 2^n

Any product of generators can be reordered, at the cost of signs, so that the indices increase, and repeated indices cancel in pairs ($e_i^2 = \pm 1$). So every element is a combination of the **basis monomials**

$$e_I \coloneqq e_{i_1} e_{i_2} \cdots e_{i_k}, \qquad I = \{i_1 < i_2 < \cdots < i_k\} \subseteq \{1, \dots, n\}, \tag{T4.11}$$

with $e_\varnothing = 1$. There is one monomial for each subset I, so 2^n of them, and $\binom{n}{k}$ of degree k. That they are linearly independent (so $\dim \mathrm{Cl}_{p,q} = 2^n$ exactly) follows from the matrix representations below, where we will simply see 2^n independent matrices [Bär Cor. 2.1.13]. The number $k = |I|$ is the **degree** or grade of e_I.

Worked example T4.6: multiplying two monomials

In $\mathrm{Cl}_{3,0}$ compute $(e_1e_2)(e_2e_3)$ and $(e_1e_2)(e_1e_3)$:

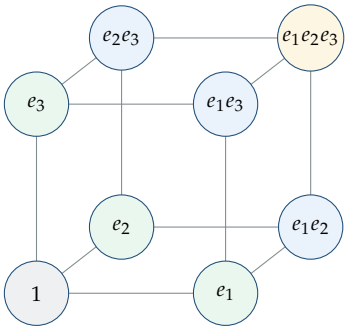
$$(e_1e_2)(e_2e_3) \overset{(1)}{=} e_1(e_2e_2)e_3 \overset{(2)}{=} e_1e_3,$$
$$(e_1e_2)(e_1e_3) \overset{(3)}{=} -e_2e_1e_1e_3 \overset{(4)}{=} -e_2e_3.$$

(1) Associativity.
(2) $e_2^2 = +1$.
(3) Anticommute $e_1e_2 = -e_2e_1$.
(4) $e_1^2 = +1$.

General rule: $e_I e_J = \pm e_{I \triangle J}$ (times ± 1 from the squares), where $I \triangle J$ is the symmetric difference: indices in both sets cancel. The product of monomials of degrees k and l has degree $k + l - 2|I \cap J|$, which has the same *parity* as $k + l$. The degree itself is not preserved, only its parity.

T4.2.3 One dimension: $\mathbb{C}$ and $\mathbb{R} \oplus \mathbb{R}$

With a single generator e the algebra has basis $\{1, e\}$.



degree k: number $\binom{3}{k}$

$k = 0$: 1 scalar

$k = 1$: 3 vectors

$k = 2$: 3 bivectors

$k = 3$: 1 trivector

$1 + 3 + 3 + 1 = 2^3$

Figure T4.5: The basis of Cl_3 drawn as the vertices of a cube: a monomial e_I is a vertex, and I records which of the three directions you have moved along. Multiplying by e_i moves you along edge i (with a sign). The number of monomials of degree k is a row of Pascal's triangle, and the total is $2^3 = 8$. Even monomials (grey, blue) and odd ones (green, yellow) alternate around the cube.

Worked example T4.7: $\mathrm{Cl}_{0,1} \cong \mathbb{C}$

If $e^2 = -1$, then e behaves exactly like i: the map $a + be \mapsto a + bi$ is an algebra isomorphism [Bär Ex. 2.1.9]. As a real 2×2 matrix, $e \mapsto \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$, the rotation by $90°$, whose square is $-\mathbf{1}$.

Worked example T4.8: $\mathrm{Cl}_{1,0} \cong \mathbb{R} \oplus \mathbb{R}$

If $e^2 = +1$, the elements $P_\pm := \frac{1}{2}(1 \pm e)$ satisfy

$$P_\pm^2 \overset{(1)}{=} \tfrac{1}{4}(1 \pm 2e + e^2) \overset{(2)}{=} \tfrac{1}{4}(2 \pm 2e) = P_\pm,$$
$$P_+P_- \overset{(3)}{=} \tfrac{1}{4}(1 - e^2) \overset{(4)}{=} 0, \qquad P_+ + P_- = 1.$$

(1) Expand the square.
(2) $e^2 = 1$.
(3) $(1 + e)(1 - e) = 1 - e + e - e^2$.
(4) $e^2 = 1$ again.

So $P_\pm$ are complementary projectors, and $a + be = (a + b)P_+ + (a - b)P_-$. Multiplication is componentwise in the $P_\pm$ basis: the algebra is two copies of $\mathbb{R}$ side by side [Bär Ex. 2.1.10]. Concretely, $e \mapsto \sigma_3 = \mathrm{diag}(1, -1)$ and $a + be \mapsto \mathrm{diag}(a + b,\ a - b)$. Note the difference from $\mathbb{C}$: $\mathbb{R} \oplus \mathbb{R}$ has zero divisors ($P_+P_- = 0$), while $\mathbb{C}$ is a field. This is the simplest instance of the "$\oplus$" that appears in odd dimensions throughout the classification.

T4.2.4 Two dimensions: real matrices and quaternions

Worked example T4.9: $\mathrm{Cl}_{2,0} \cong \mathbb{R}(2)$, the real 2×2 matrices

We need two anticommuting matrices squaring to $+1$. We could use σ_1, σ_2, but σ_2 is imaginary; the *real* choice $e_1 \mapsto \sigma_1$, $e_2 \mapsto \sigma_3$ also works because $\sigma_1\sigma_3 = -\sigma_3\sigma_1$. The fourth basis element is

$$e_1e_2 \mapsto \sigma_1\sigma_3 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},$$

the 90° rotation, which squares to $-\mathbf{1}$ (check: $(e_1e_2)^2 = e_1e_2e_1e_2 = -e_1e_1e_2e_2 = -1$). The four matrices

$$\mathbf{1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad \sigma_1\sigma_3 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

are linearly independent and span all real 2×2 matrices: $\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \frac{a+d}{2}\mathbf{1} + \frac{a-d}{2}\sigma_3 + \frac{b+c}{2}\sigma_1 + \frac{c-b}{2}\sigma_1\sigma_3$. So $\mathrm{Cl}_{2,0} \cong \mathbb{R}(2)$, where $\mathbb{R}(m)$ denotes the algebra of real $m \times m$ matrices.

Worked example T4.10: $\mathrm{Cl}_{0,2} \cong \mathbb{H}$, the quaternions

Now $e_1^2 = e_2^2 = -1$. The monomial e_1e_2 also squares to -1: $(e_1e_2)^2 = -e_1^2e_2^2 = -1$. So all three of e_1, e_2, e_1e_2 square to -1, and they anticommute pairwise; for instance $e_1(e_1e_2) = e_1^2e_2 = -e_2$ while $(e_1e_2)e_1 = -e_1^2e_2 = e_2$. Writing $\mathbf{i} = e_1$, $\mathbf{j} = e_2$, $\mathbf{k} = e_1e_2$ we get Hamilton's rules

$$\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = \mathbf{ijk} = -1, \qquad \mathbf{ij} = -\mathbf{ji} = \mathbf{k}, \quad \mathbf{jk} = -\mathbf{kj} = \mathbf{i}, \quad \mathbf{ki} = -\mathbf{ik} = \mathbf{j}. \tag{T4.12}$$

So $\mathrm{Cl}_{0,2} \cong \mathbb{H}$ [Bär Ex. 2.1.14]. A matrix representation uses the Pauli matrices times $-i$:

$$\mathbf{i} \mapsto -i\sigma_1, \qquad \mathbf{j} \mapsto -i\sigma_2, \qquad \mathbf{k} \mapsto -i\sigma_3. \tag{T4.13}$$

Check one relation: $\mathbf{ij} \mapsto (-i\sigma_1)(-i\sigma_2) \overset{(1)}{=} -\sigma_1\sigma_2 \overset{(2)}{=} -i\sigma_3 \leftrightarrow \mathbf{k}$.

(1) $(-i)^2 = -1$.

(2) $\sigma_1\sigma_2 = i\sigma_3$ from (T4.5).

A general quaternion $q = a + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}$ becomes

$$q \mapsto \begin{pmatrix} a - id & -c - ib \\ c - ib & a + id \end{pmatrix} = \begin{pmatrix} \alpha & -\bar{\beta} \\ \beta & \bar{\alpha} \end{pmatrix}, \qquad \alpha = a - id,\ \beta = c - ib, \tag{T4.14}$$

and $\det = |\alpha|^2 + |\beta|^2 = a^2 + b^2 + c^2 + d^2 = |q|^2$. The unit quaternions are exactly the matrices of this form with determinant 1, i.e. the group SU(2). This is the first appearance of the spin group: Spin(3) = SU(2) (Section T4.5).

$\times$	1	$\mathbf{i} = e_1$	$\mathbf{j} = e_2$	$\mathbf{k} = e_1e_2$
1	1	$\mathbf{i}$	$\mathbf{j}$	$\mathbf{k}$
$\mathbf{i}$	$\mathbf{i}$	-1	$\mathbf{k}$	$-\mathbf{j}$
$\mathbf{j}$	$\mathbf{j}$	$-\mathbf{k}$	-1	$\mathbf{i}$
$\mathbf{k}$	$\mathbf{k}$	$\mathbf{j}$	$-\mathbf{i}$	-1

$\times$	1	e_1	e_2	e_1e_2
1	1	e_1	e_2	e_1e_2
e_1	e_1	$+1$	e_1e_2	e_2
e_2	e_2	$-e_1e_2$	$+1$	$-e_1$
e_1e_2	e_1e_2	$-e_2$	e_1	-1

Table T4.1: Multiplication tables (row $\times$ column) of $\mathrm{Cl}_{0,2} = \mathbb{H}$ (left) and $\mathrm{Cl}_{2,0} = \mathbb{R}(2)$ (right). The only difference is the sign of the squares of the generators, yet the left algebra is a division algebra and the right one is not (it contains the non-invertible matrix $\frac{1}{2}(1 + e_1)$). In both, the bivector e_1e_2 squares to -1: that is the rotation generator.

T4.2.5 Three dimensions: the Pauli algebra

Worked example T4.11: $\mathrm{Cl}_{3,0} \cong \mathbb{C}(2)$: the algebra of Pauli matrices

Take $e_a \mapsto \sigma_a$. The eight basis monomials become

$$1 \mapsto \mathbf{1}; \qquad e_1, e_2, e_3 \mapsto \sigma_1, \sigma_2, \sigma_3;$$

$$e_1e_2 \mapsto i\sigma_3, \quad e_2e_3 \mapsto i\sigma_1, \quad e_3e_1 \mapsto i\sigma_2; \qquad e_1e_2e_3 \mapsto \sigma_1\sigma_2\sigma_3 \overset{(1)}{=} i\sigma_3\sigma_3 \overset{(2)}{=} i\mathbf{1}.$$

(1) $\sigma_1\sigma_2 = i\sigma_3$.
(2) $\sigma_3^2 = \mathbf{1}$.

As a *real* vector space, the span of $\{\mathbf{1}, \sigma_a, i\sigma_a, i\mathbf{1}\}$ is all of $\mathbb{C}(2)$, the complex 2×2 matrices (real dimension 8). So the map is onto an 8-dimensional algebra, hence injective: $\mathrm{Cl}_{3,0} \cong \mathbb{C}(2)$ and the eight monomials are linearly independent, confirming dim $= 2^3$.
The **volume element** $\omega := e_1e_2e_3$ is special:

- it commutes with every e_a (moving e_a through ω passes two other generators and one copy of itself: $(-1)^2 = 1$), hence it is central;
- $\omega^2 = e_1e_2e_3e_1e_2e_3 = (e_1e_2)^2e_3^2 \cdot (+1) = -1$ (moving e_3 past e_1e_2 costs $(-1)^2$).

A central element squaring to -1 acts as $\pm i$; in the Pauli representation it is exactly $i\mathbf{1}$. This is why the "real" Clifford algebra of three-space secretly contains the complex numbers.

Physical picture: geometric algebra of space

In $\mathrm{Cl}_{3,0}$ the four grades have a direct geometric meaning. Vectors $v = v^a e_a$ are arrows. Bivectors e_1e_2, e_2e_3, e_3e_1 are oriented plane elements; since $e_1e_2 = i\sigma_3$, a bivector is "i times the normal vector", which is the Clifford version of the fact that an axial vector (angular momentum, magnetic field) is really an oriented plane. The trivector $e_1e_2e_3$ is an oriented volume, the pseudoscalar. The product of two vectors is scalar plus bivector: $ab = a \cdot b + i\,\sigma \cdot (a \times b)$, which is the Pauli identity of Example T4.2 read geometrically.

Worked example T4.12: $\mathrm{Cl}_{0,3} \cong \mathbb{H} \oplus \mathbb{H}$: a representation can be smaller than the algebra

With $e_a^2 = -1$ try $e_a \mapsto -i\sigma_a$ (the quaternions of (T4.13)). The relations hold, but

$$e_1e_2e_3 \mapsto (-i)^3\sigma_1\sigma_2\sigma_3 \overset{(1)}{=} i \cdot i\mathbf{1} = -\mathbf{1}.$$

(1) $(-i)^3 = i$ and $\sigma_1\sigma_2\sigma_3 = i\mathbf{1}$.

So the degree-3 monomial is represented by a multiple of the identity: the eight monomials are *not* independent in this representation, which has only real dimension 4. What happened? Now the volume element $\omega = e_1e_2e_3$ is central with $\omega^2 = +1$ (same computation as before with $e_3^2 = -1$ and $(e_1e_2)^2 = -1$). Exactly as in $\mathrm{Cl}_{1,0}$, the elements $\frac{1}{2}(1 \pm \omega)$ are central complementary projectors, and they split the algebra into two pieces, each a copy of $\mathbb{H}$. The representation $e_a \mapsto -i\sigma_a$ has $\omega = -1$, so it sees only one piece; the other representation $e_a \mapsto +i\sigma_a$ has $\omega = +1$ and sees the other. The faithful representation is block diagonal, $e_a \mapsto \mathrm{diag}(-i\sigma_a, i\sigma_a)$, and $\mathrm{Cl}_{0,3} \cong \mathbb{H} \oplus \mathbb{H}$. The lesson: in odd dimensions a central volume element with square $+1$ splits the algebra, and there are two inequivalent irreducible representations.

T4.2.6 Clifford algebra versus exterior algebra

The basis monomials e_I look exactly like the basis $e_{i_1} \wedge \cdots \wedge e_{i_k}$ of the exterior algebra $\Lambda^\bullet V$. The difference is only in what happens when an index repeats: in Λ, $e_i \wedge e_i = 0$; in Cl, $e_i e_i = \pm 1$. Dray [Dray §20.3] makes this precise with his "vee" product on forms: for 1-forms α, β,

$$\alpha \vee \beta = \alpha \wedge \beta + g(\alpha, \beta) \tag{T4.15}$$

(Dray's (20.20); we use the + convention). The Clifford product of two vectors is their wedge product plus their dot product. Let us derive Dray's formula (20.21) for three 1-forms. We need the product of a 2-form with a 1-form, $(\alpha \wedge \beta) \vee \gamma$. Write $\alpha \wedge \beta = \alpha \vee \beta - g(\alpha, \beta)$ and use associativity; for orthonormal e_1, e_2, e_3 we check directly:

$$e_1 \vee e_2 \vee e_1 \overset{(1)}{=} -e_1 \vee e_1 \vee e_2 \overset{(2)}{=} -e_2,$$

$$\text{formula (20.21):} \quad e_1 \wedge e_2 \wedge e_1 + g(e_1, e_2)e_1 - g(e_1, e_1)e_2 + g(e_2, e_1)e_1 \overset{(3)}{=} 0 + 0 - e_2 + 0 = -e_2.$$

(1) Anticommute the first two factors ($e_1 \perp e_2$).
(2) $e_1 \vee e_1 = g(e_1, e_1) = 1$.
(3) A wedge with a repeated factor vanishes, and $g(e_1, e_2) = 0$.

Checking all orderings of orthonormal triples (with repeats) proves

$$\alpha \vee \beta \vee \gamma = \alpha \wedge \beta \wedge \gamma + g(\alpha, \beta)\gamma - g(\alpha, \gamma)\beta + g(\beta, \gamma)\alpha \tag{T4.16}$$

by linearity. We will use exactly this identity (with γ matrices) in the Lichnerowicz computation of Section T4.8.

Key idea: Clifford = exterior, deformed by the metric

The linear map $\Lambda^\bullet V \to \mathrm{Cl}(V)$, $e_{i_1} \wedge \cdots \wedge e_{i_k} \mapsto e_{i_1} \cdots e_{i_k}$ (for an orthonormal basis) is an isomorphism of *vector spaces* but not of algebras [Bär Rem. 2.1.15]. Under it, the Clifford product of a vector with anything is

$$v \cdot \phi = v \wedge \phi + \iota_v \phi \qquad (\text{with } v^2 = +|v|^2),$$

wedge plus contraction. Every Clifford algebra is the exterior algebra with its product "corrected" by the inner product. This is the bridge that lets differential forms carry a Clifford action (Section T4.9).

T4.2.7 Even and odd parts

Because every relation (T4.9) involves only even-degree terms ($e_i e_j + e_j e_i$ and 1), the parity of the degree is well defined, even though the degree itself is not (Example T4.6). So the algebra splits

$$\mathrm{Cl}_{p,q} = \mathrm{Cl}^0_{p,q} \oplus \mathrm{Cl}^1_{p,q}, \qquad \mathrm{Cl}^i \cdot \mathrm{Cl}^j \subseteq \mathrm{Cl}^{i+j \bmod 2}, \tag{T4.17}$$

where Cl^0 (the **even part**) is spanned by monomials of even degree and Cl^1 (the odd part) by those of odd degree. Each has dimension 2^{n-1}. The even part is a subalgebra; the odd part is not. A slick way to say this [Bär p. 71]: the map $e_i \mapsto -e_i$ preserves the relations, so it extends to an automorphism α with $\alpha^2 = 1$; Cl^0 and Cl^1 are its ± 1 eigenspaces.

Worked example T4.13: the even part of $\mathrm{Cl}_{3,0}$ is the quaternions

$\mathrm{Cl}^0_{3,0}$ is spanned by $1, e_2e_3, e_3e_1, e_1e_2 \mapsto \mathbf{1}, i\sigma_1, i\sigma_2, i\sigma_3$. Each bivector squares to -1, and $(e_2e_3)(e_3e_1) = e_2e_1 = -e_1e_2$. So with $\mathbf{i} = -e_2e_3$, $\mathbf{j} = -e_3e_1$, $\mathbf{k} = -e_1e_2$ (matrices $-i\sigma_a$) we recover Hamilton's rules: $\mathrm{Cl}^0_{3,0} \cong \mathbb{H} \cong \mathrm{Cl}_{0,2}$. The even part of the 3D algebra is the 2D algebra with the opposite sign. This is a general phenomenon [Bär Lemma 2.3.12]:

$$\mathrm{Cl}^0_{p,q+1} \cong \mathrm{Cl}_{p,q}, \qquad \text{via } e_i \mapsto e_ie_{n+1}\ (i \le n), \tag{T4.18}$$

where e_{n+1} is the extra generator with $e_{n+1}^2 = -1$. The check: $(e_ie_{n+1})^2 = -e_i^2e_{n+1}^2 = e_i^2$, and for $i \neq j$, $(e_ie_{n+1})(e_je_{n+1}) = -e_ie_je_{n+1}^2 = e_ie_j$ anticommutes with $(e_je_{n+1})(e_ie_{n+1}) = e_je_i$. So the elements e_ie_{n+1} satisfy the relations of $\mathrm{Cl}_{p,q}$, and a dimension count (2^n on both sides) gives the isomorphism.

Physical picture: why physicists care about the even part

Rotations are built from products of an *even* number of vectors (Section T4.5), so the spin group and its Lie algebra live in Cl^0. The Lie algebra is spanned by the bivectors e_ie_j: in 3D these are $i\sigma_a$, the generators $-\frac{i}{2}\sigma_a$ of SU(2) up to a factor. The isomorphism (T4.18) is the algebraic reason why the spinors of $n+1$ dimensions, restricted to rotations, look like the spinors of n dimensions (e.g. a 4D Weyl spinor restricted to spatial rotations is a 3D Pauli spinor).

T4.2.8 Building big algebras from small ones

If $V = V_1 \oplus V_2$ is an orthogonal splitting, then generators from V_1 anticommute with generators from V_2. So $\mathrm{Cl}(V)$ is built from $\mathrm{Cl}(V_1)$ and $\mathrm{Cl}(V_2)$, but with a twist: we must remember that odd elements of the two factors anticommute. Bär [Bär Def. 2.1.11, Prop. 2.1.12] calls this the **graded tensor product**:

$$\mathrm{Cl}(V_1 \oplus V_2) \cong \mathrm{Cl}(V_1) \hat{\otimes} \mathrm{Cl}(V_2), \qquad (a \otimes b)(a' \otimes b') = (-1)^{|b|\,|a'|}\, aa' \otimes bb', \tag{T4.19}$$

with $v_1 + v_2 \mapsto v_1 \otimes 1 + 1 \otimes v_2$. The sign is exactly what is needed: $(v_1 \otimes 1)(1 \otimes v_2) = v_1 \otimes v_2$ while $(1 \otimes v_2)(v_1 \otimes 1) = -v_1 \otimes v_2$. Taking dimensions gives $\dim \mathrm{Cl}(V) = \dim \mathrm{Cl}(V_1) \cdot \dim \mathrm{Cl}(V_2)$, and induction from the one-dimensional case gives $\dim \mathrm{Cl}(V) = 2^{\dim V}$ [Bär Cor. 2.1.13]. For computing *which* matrix algebra we get, the ungraded tensor product is more useful; we use it in Section T4.4.5. In physics this sign rule is the familiar one: fermionic operators from independent subsystems anticommute.

Watch out: "the Clifford algebra of $\mathrm{Spin}(n)$"

Nash [Nash p. 114] says the γ-matrices "generate the Clifford algebra of $\mathrm{Spin}(n)$". Strictly, the Clifford algebra belongs to the vector space $\mathbb{R}^n$ with its inner product; $\mathrm{Spin}(n)$ is a group sitting *inside* it (Section T4.5). The phrase is a harmless shorthand, but do not look for a Clifford algebra "of a group" in general.

Recap

- $\mathrm{Cl}_{p,q}$: generators with $e_ie_j + e_je_i = 2\eta_{ij}$; equivalently $v^2 = Q(v)$. Basis e_I indexed by subsets; $\dim = 2^n$.
- Low dimensions: $\mathrm{Cl}_{0,1} = \mathbb{C}$, $\mathrm{Cl}_{1,0} = \mathbb{R} \oplus \mathbb{R}$, $\mathrm{Cl}_{0,2} = \mathbb{H}$ (matrices $-i\sigma_a$), $\mathrm{Cl}_{2,0} = \mathbb{R}(2)$

(matrices σ_1, σ_3), $\mathrm{Cl}_{3,0} = \mathbb{C}(2)$ (Pauli), $\mathrm{Cl}_{0,3} = \mathbb{H} \oplus \mathbb{H}$.

- $\mathrm{Cl} \cong \Lambda^\bullet$ as vector spaces; Clifford product of a vector = wedge + contraction (Dray's vee product).
- $\mathbb{Z}_2$-grading $\mathrm{Cl} = \mathrm{Cl}^0 \oplus \mathrm{Cl}^1$; $\mathrm{Cl}^0_{p,q+1} \cong \mathrm{Cl}_{p,q}$; rotations live in the even part.
- In odd dimensions a central volume element can split the algebra into two pieces, giving two inequivalent irreducible representations.

T4.3 Gamma matrices in four dimensions

We saw in Section T4.1.2 that three anticommuting 2×2 matrices exist but four do not. Four dimensions is where physics lives, so we need four anticommuting matrices, and they must be at least 4×4. This section writes them down in the representations that appear in the literature, checks every relation by block multiplication, and then uses them: chirality γ^5, Weyl spinors, the Lorentz generators, and the sixteen bilinears $\bar\psi \Gamma \psi$. Everything later in the book (the index of the Dirac operator in Chapter 4, the chiral anomaly in Chapter 10) is written in terms of these matrices.

The one trick we use over and over is **block multiplication**. Write a 4×4 matrix as a 2×2 array of 2×2 blocks. Then

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} A' & B' \\ C' & D' \end{pmatrix} = \begin{pmatrix} AA' + BC' & AB' + BD' \\ CA' + DC' & CB' + DD' \end{pmatrix},$$

exactly as for numbers, except that the order of the blocks in each product must be kept, because 2×2 matrices do not commute. With Pauli matrices in the blocks, every computation reduces to the Pauli product rule (T4.6).

T4.3.1 Euclidean signature: the chiral representation

In Euclidean $\mathbb{R}^4$ we want Hermitian $\gamma_1, \dots, \gamma_4$ with $\{\gamma_a, \gamma_b\} = 2\delta_{ab}$. The idea is to put 2×2 blocks off the diagonal. Define the four 2×2 matrices

$$\tau_a = (\tau_1, \tau_2, \tau_3, \tau_4) \coloneqq (i\sigma_1,\ i\sigma_2,\ i\sigma_3,\ \mathbf{1}), \qquad \tau_a^\dagger = (-i\sigma_1,\ -i\sigma_2,\ -i\sigma_3,\ \mathbf{1}), \tag{T4.20}$$

and set

$$\gamma_a \coloneqq \begin{pmatrix} 0 & \tau_a \\ \tau_a^\dagger & 0 \end{pmatrix}, \qquad a = 1, 2, 3, 4. \tag{T4.21}$$

Written out, $\gamma_4 = \begin{pmatrix} 0 & \mathbf{1} \\ \mathbf{1} & 0 \end{pmatrix}$ and $\gamma_j = \begin{pmatrix} 0 & i\sigma_j \\ -i\sigma_j & 0 \end{pmatrix}$ for $j = 1, 2, 3$. Each γ_a is Hermitian, because the lower-left block is the Hermitian conjugate of the upper-right block.

The four τ_a are the *quaternion units* (up to sign) from (T4.13), and the relation we need is

$$\tau_a \tau_b^\dagger + \tau_b \tau_a^\dagger = 2\delta_{ab}\mathbf{1}, \qquad \tau_a^\dagger \tau_b + \tau_b^\dagger \tau_a = 2\delta_{ab}\mathbf{1}. \tag{T4.22}$$

Check it case by case. For $a = b = 4$: $2 \cdot \mathbf{1} \cdot \mathbf{1} = 2$. For $a = j$, $b = 4$: $\tau_j \tau_4^\dagger + \tau_4 \tau_j^\dagger = i\sigma_j - i\sigma_j = 0$. For $a = j, b = k$ spatial:

$$\tau_j \tau_k^\dagger + \tau_k \tau_j^\dagger \overset{(1)}{=} (i\sigma_j)(-i\sigma_k) + (i\sigma_k)(-i\sigma_j) \overset{(2)}{=} \sigma_j\sigma_k + \sigma_k\sigma_j \overset{(3)}{=} 2\delta_{jk}.$$

(1) Insert (T4.20).
(2) $(i)(-i) = 1$.
(3) The Pauli anticommutator.

The second relation in (T4.22) is checked the same way.

Worked example T4.14: the Euclidean Clifford relations by block multiplication

Using the block rule,

$$\gamma_a\gamma_b + \gamma_b\gamma_a \overset{(1)}{=} \begin{pmatrix} \tau_a\tau_b^\dagger & 0 \\ 0 & \tau_a^\dagger\tau_b \end{pmatrix} + \begin{pmatrix} \tau_b\tau_a^\dagger & 0 \\ 0 & \tau_b^\dagger\tau_a \end{pmatrix} \overset{(2)}{=} \begin{pmatrix} 2\delta_{ab}\mathbf{1} & 0 \\ 0 & 2\delta_{ab}\mathbf{1} \end{pmatrix} = 2\delta_{ab}\mathbf{1}_4.$$

(1) Product of two block-off-diagonal matrices: the upper-left block is (upper right)×(lower left) $= \tau_a\tau_b^\dagger$, and the lower-right block is (lower left)×(upper right) $= \tau_a^\dagger\tau_b$.
(2) The relations (T4.22).

So (T4.21) is a representation of the Euclidean Clifford algebra on $\mathbb{C}^4$. Note what the computation shows in passing: the product of *two* gamma matrices is block *diagonal*. This fact is behind everything that follows about chirality.

The chirality matrix. Multiply all four:

$$\begin{aligned} \gamma_1\gamma_2 &\overset{(1)}{=} \begin{pmatrix} \tau_1\tau_2^\dagger & 0 \\ 0 & \tau_1^\dagger\tau_2 \end{pmatrix} \overset{(2)}{=} \begin{pmatrix} \sigma_1\sigma_2 & 0 \\ 0 & \sigma_1\sigma_2 \end{pmatrix} = \begin{pmatrix} i\sigma_3 & 0 \\ 0 & i\sigma_3 \end{pmatrix}, \\ \gamma_3\gamma_4 &\overset{(3)}{=} \begin{pmatrix} \tau_3 & 0 \\ 0 & \tau_3^\dagger \end{pmatrix} = \begin{pmatrix} i\sigma_3 & 0 \\ 0 & -i\sigma_3 \end{pmatrix}, \\ \gamma_1\gamma_2\gamma_3\gamma_4 &\overset{(4)}{=} \begin{pmatrix} (i\sigma_3)(i\sigma_3) & 0 \\ 0 & (i\sigma_3)(-i\sigma_3) \end{pmatrix} = \begin{pmatrix} -\mathbf{1} & 0 \\ 0 & \mathbf{1} \end{pmatrix}. \end{aligned} \tag{T4.23}$$

(1) Block rule for two off-diagonal matrices.
(2) $(i\sigma_1)(-i\sigma_2) = \sigma_1\sigma_2$ and $(-i\sigma_1)(i\sigma_2) = \sigma_1\sigma_2$.
(3) Same rule with $\tau_4 = \mathbf{1}$.
(4) Multiply the two diagonal matrices block by block; $\sigma_3^2 = \mathbf{1}$.

We define the Euclidean **chirality operator** in four dimensions by

$$\Gamma := (-i)^2\gamma_1\gamma_2\gamma_3\gamma_4 = -\gamma_1\gamma_2\gamma_3\gamma_4 = \begin{pmatrix} \mathbf{1} & 0 \\ 0 & -\mathbf{1} \end{pmatrix}. \tag{T4.24}$$

The phase $(-i)^2$ is part of a general rule, $\Gamma = (-i)^k\gamma_1\cdots\gamma_{2k}$, that we justify in Section T4.4; it is chosen so that Γ is Hermitian with $\Gamma^2 = 1$. In this representation Γ is diagonal: the upper two components of a spinor have chirality $+1$ and the lower two have chirality -1. That is why this is called the **chiral representation**.

T4.3.2 Minkowski signature: Dirac and Weyl representations

We now follow the QFT notes [QFT notes §3.5]. The metric is $\eta^{\mu\nu} = \mathrm{diag}(+1,-1,-1,-1)$, so we need

$$(\gamma^0)^2 = \mathbf{1}, \qquad (\gamma^i)^2 = -\mathbf{1}\ (i=1,2,3), \qquad \gamma^\mu\gamma^\nu = -\gamma^\nu\gamma^\mu\ (\mu \neq \nu).$$

Since $(\gamma^i)^2 = -1$, the γ^i cannot be Hermitian (a Hermitian matrix has real eigenvalues, so its square is positive). The standard choice is γ^0 Hermitian and γ^i anti-Hermitian. All the representations below share the spatial matrices

$$\gamma^i = \begin{pmatrix} 0 & \sigma^i \\ -\sigma^i & 0 \end{pmatrix}, \qquad i = 1,2,3, \tag{T4.25}$$

and differ in γ^0.

Working definition T4.4 (Dirac and Weyl (chiral) representations)

The **Dirac representation** has $\gamma^0_{\mathrm{D}} = \begin{pmatrix} \mathbf{1} & 0 \\ 0 & -\mathbf{1} \end{pmatrix}$ and (T4.25). The **Weyl** or **chiral representation** [the basis of $\mathbb{C}^4$ in which γ^5 is diagonal, so the upper two and lower two components have opposite chirality] [QFT notes (3.5.13)] has $\gamma^0_{\mathrm{W}} = \begin{pmatrix} 0 & \mathbf{1} \\ \mathbf{1} & 0 \end{pmatrix}$ and (T4.25). With

$$\sigma^\mu := (\mathbf{1}, \sigma^1, \sigma^2, \sigma^3), \qquad \bar\sigma^\mu := (\mathbf{1}, -\sigma^1, -\sigma^2, -\sigma^3), \tag{T4.26}$$

the Weyl representation is compactly $\gamma^\mu = \begin{pmatrix} 0 & \sigma^\mu \\ \bar\sigma^\mu & 0 \end{pmatrix}$ [QFT notes (3.5.51)]. *Test for any proposed set:* check $\{\gamma^\mu, \gamma^\nu\} = 2\eta^{\mu\nu}$ on all ten pairs, e.g. by block multiplication as below.

Worked example T4.15: verifying the Weyl representation

Exactly as in Example T4.14,

$$\{\gamma^\mu, \gamma^\nu\} \overset{(1)}{=} \begin{pmatrix} \sigma^\mu\bar\sigma^\nu + \sigma^\nu\bar\sigma^\mu & 0 \\ 0 & \bar\sigma^\mu\sigma^\nu + \bar\sigma^\nu\sigma^\mu \end{pmatrix}.$$

(1) Block rule for off-diagonal matrices.

Now evaluate the upper-left block on each type of pair:

$$\begin{aligned} \mu = \nu = 0: &\quad \sigma^0\bar\sigma^0 + \sigma^0\bar\sigma^0 \overset{(1)}{=} 2\mathbf{1} = 2\eta^{00}, \\ \mu = 0, \nu = i: &\quad \sigma^0\bar\sigma^i + \sigma^i\bar\sigma^0 \overset{(2)}{=} -\sigma^i + \sigma^i = 0 = 2\eta^{0i}, \\ \mu = i, \nu = j: &\quad \sigma^i\bar\sigma^j + \sigma^j\bar\sigma^i \overset{(3)}{=} -(\sigma^i\sigma^j + \sigma^j\sigma^i) \overset{(4)}{=} -2\delta^{ij} = 2\eta^{ij}. \end{aligned}$$

(1) $\sigma^0 = \bar\sigma^0 = \mathbf{1}$.
(2) $\bar\sigma^i = -\sigma^i$ and $\bar\sigma^0 = \mathbf{1}$.
(3) Again $\bar\sigma^j = -\sigma^j$.
(4) Pauli anticommutator; and $\eta^{ij} = -\delta^{ij}$.

The lower-right block is the same computation with $\sigma \leftrightarrow \bar\sigma$ and gives the same answers. So $\{\gamma^\mu, \gamma^\nu\} = 2\eta^{\mu\nu}\mathbf{1}_4$. The Dirac representation is checked the same way; the only new products are $\gamma^0_{\mathrm{D}}\gamma^i + \gamma^i\gamma^0_{\mathrm{D}} = \begin{pmatrix} 0 & \sigma^i \\ \sigma^i & 0 \end{pmatrix} + \begin{pmatrix} 0 & -\sigma^i \\ -\sigma^i & 0 \end{pmatrix} = 0$.

Why two representations, and how they are related. Both sets of matrices satisfy the same algebra. In Section T4.4.3 we will see that any two irreducible representations of the complex Clifford algebra in even dimensions are *equivalent*: there is an invertible U with $\gamma^\mu_{\mathrm{W}} = U\gamma^\mu_{\mathrm{D}}U^{-1}$ for all μ. Here it is explicitly.

Worked example T4.16: the change of basis from Dirac to Weyl

Let $U = \frac{1}{\sqrt{2}}\begin{pmatrix} \mathbf{1} & -\mathbf{1} \\ \mathbf{1} & \mathbf{1} \end{pmatrix}$, which is unitary with $U^\dagger = \frac{1}{\sqrt{2}}\begin{pmatrix} \mathbf{1} & \mathbf{1} \\ -\mathbf{1} & \mathbf{1} \end{pmatrix}$. Then

$$U\gamma^0_{\mathrm{D}}U^\dagger \overset{(1)}{=} \tfrac{1}{2}\begin{pmatrix} \mathbf{1} & \mathbf{1} \\ \mathbf{1} & -\mathbf{1} \end{pmatrix}\begin{pmatrix} \mathbf{1} & \mathbf{1} \\ -\mathbf{1} & \mathbf{1} \end{pmatrix} \overset{(2)}{=} \tfrac{1}{2}\begin{pmatrix} 0 & 2\cdot\mathbf{1} \\ 2\cdot\mathbf{1} & 0 \end{pmatrix} = \gamma^0_{\mathrm{W}},$$

$$U\gamma^i U^\dagger \overset{(3)}{=} \tfrac{1}{2}\begin{pmatrix} \sigma^i & \sigma^i \\ -\sigma^i & \sigma^i \end{pmatrix}\begin{pmatrix} \mathbf{1} & \mathbf{1} \\ -\mathbf{1} & \mathbf{1} \end{pmatrix} \overset{(4)}{=} \begin{pmatrix} 0 & \sigma^i \\ -\sigma^i & 0 \end{pmatrix} = \gamma^i.$$

(1) $U\gamma^0_{\mathrm{D}}$: multiplying by $\mathrm{diag}(\mathbf{1}, -\mathbf{1})$ on the right flips the sign of the second block column of U.

(2) Block multiplication: upper-left $\mathbf{1} - \mathbf{1} = 0$, upper-right $\mathbf{1} + \mathbf{1}$, and so on.

(3) $U\gamma^i = \frac{1}{\sqrt{2}}\left(\begin{smallmatrix} \sigma^i & \sigma^i \\ -\sigma^i & \sigma^i \end{smallmatrix}\right)$ by block multiplication.

(4) Upper-left $\sigma^i - \sigma^i = 0$, upper-right $\sigma^i + \sigma^i$, lower-left $-\sigma^i - \sigma^i$, lower-right $-\sigma^i + \sigma^i$; then divide by 2.

So a Dirac-representation spinor ψ_{D} corresponds to the Weyl-representation spinor $\psi_{\mathrm{W}} = U\psi_{\mathrm{D}}$. The physics is identical; only the basis of $\mathbb{C}^4$ has changed. The Dirac representation is convenient near rest (then γ^0 is diagonal and separates particle from antiparticle components); the Weyl representation is convenient for massless particles and for chirality.

Hermiticity in one line. In both representations

$$(\gamma^\mu)^\dagger = \gamma^0\gamma^\mu\gamma^0. \tag{T4.27}$$

For $\mu = 0$ this says γ^0 is Hermitian (using $(\gamma^0)^2 = 1$). For $\mu = i$: $\gamma^0\gamma^i\gamma^0 = -\gamma^i(\gamma^0)^2 = -\gamma^i = (\gamma^i)^\dagger$. Since U is unitary, (T4.27) survives any unitary change of basis. We will use it for the Dirac conjugate $\bar\psi = \psi^\dagger\gamma^0$.

T4.3.3 The chirality matrix γ^5

Working definition T4.5 (γ^5 in Minkowski space)

$\gamma^5 := i\gamma^0\gamma^1\gamma^2\gamma^3$ [QFT notes (3.5.122)]. It satisfies

$$(\gamma^5)^\dagger = \gamma^5, \qquad (\gamma^5)^2 = \mathbf{1}, \qquad \{\gamma^5, \gamma^\mu\} = 0. \tag{T4.28}$$

Its eigenvalues are ± 1, each twice. The **chiral projectors** [$P_L\psi$ keeps only the part of ψ with $\gamma^5 = -1$, and $P_R\psi$ only the part with $\gamma^5 = +1$] are $P_L = \frac{1}{2}(1 - \gamma^5)$ and $P_R = \frac{1}{2}(1 + \gamma^5)$.

Let us derive the three properties. The factor i is there to make the square positive:

$$\begin{aligned}(\gamma^5)^2 &\overset{(1)}{=} i^2\,\gamma^0\gamma^1\gamma^2\gamma^3\gamma^0\gamma^1\gamma^2\gamma^3 \\ &\overset{(2)}{=} -(-1)^3\,\gamma^0\gamma^0\,\gamma^1\gamma^2\gamma^3\gamma^1\gamma^2\gamma^3 \\ &\overset{(3)}{=} \gamma^1\gamma^2\gamma^3\gamma^1\gamma^2\gamma^3 \overset{(4)}{=} (-1)^2\gamma^1\gamma^1\,\gamma^2\gamma^3\gamma^2\gamma^3 \\ &\overset{(5)}{=} -\gamma^2\gamma^3\gamma^2\gamma^3 \overset{(6)}{=} \gamma^2\gamma^2\gamma^3\gamma^3 \overset{(7)}{=} \mathbf{1}.\end{aligned}$$

(1) Write out the product.

(2) Move the second γ^0 to the left past $\gamma^3, \gamma^2, \gamma^1$: three anticommutations.
(3) $(\gamma^0)^2 = 1$ and $-(-1)^3 = 1$.
(4) Move the second γ^1 left past γ^3, γ^2: two signs.
(5) $(\gamma^1)^2 = -1$.
(6) Anticommute $\gamma^3\gamma^2 = -\gamma^2\gamma^3$ in the middle.
(7) $(\gamma^2)^2(\gamma^3)^2 = (-1)(-1) = 1$.

For the anticommutation: γ^μ anticommutes with the three factors of $\gamma^0\gamma^1\gamma^2\gamma^3$ different from it and commutes with the one equal to it, so moving it through the product costs $(-1)^3 = -1$. For Hermiticity:

$$(\gamma^5)^\dagger \overset{(1)}{=} -i(\gamma^3)^\dagger(\gamma^2)^\dagger(\gamma^1)^\dagger(\gamma^0)^\dagger \overset{(2)}{=} -i(-1)^3\gamma^3\gamma^2\gamma^1\gamma^0 \overset{(3)}{=} i\,(-1)^6\gamma^0\gamma^1\gamma^2\gamma^3 = \gamma^5.$$

(1) The dagger reverses the order and conjugates i.
(2) γ^0 is Hermitian, the three γ^i are anti-Hermitian.
(3) Reversing four objects takes $3 + 2 + 1 = 6$ transpositions.

Worked example T4.17: γ^5 in the two representations

In the Weyl representation,

$$\gamma^0\gamma^1 \overset{(1)}{=} \begin{pmatrix} 0 & \mathbf{1} \\ \mathbf{1} & 0 \end{pmatrix}\begin{pmatrix} 0 & \sigma^1 \\ -\sigma^1 & 0 \end{pmatrix} = \begin{pmatrix} -\sigma^1 & 0 \\ 0 & \sigma^1 \end{pmatrix},$$

$$\gamma^2\gamma^3 \overset{(2)}{=} \begin{pmatrix} -\sigma^2\sigma^3 & 0 \\ 0 & -\sigma^2\sigma^3 \end{pmatrix} \overset{(3)}{=} \begin{pmatrix} -i\sigma^1 & 0 \\ 0 & -i\sigma^1 \end{pmatrix},$$

$$\gamma^5 = i\gamma^0\gamma^1\gamma^2\gamma^3 \overset{(4)}{=} i\begin{pmatrix} (-\sigma^1)(-i\sigma^1) & 0 \\ 0 & \sigma^1(-i\sigma^1) \end{pmatrix} = \begin{pmatrix} -\mathbf{1} & 0 \\ 0 & \mathbf{1} \end{pmatrix}.$$

(1) Block multiplication.
(2) Two off-diagonal blocks: upper-left is $\sigma^2(-\sigma^3)$, lower-right is $(-\sigma^2)\sigma^3$.
(3) $\sigma^2\sigma^3 = i\sigma^1$.
(4) Multiply the diagonal blocks and use $(\sigma^1)^2 = \mathbf{1}$; then $i \cdot i = -1$ and $i \cdot (-i) = 1$.

So $\gamma^5_{\mathrm{W}} = \mathrm{diag}(-\mathbf{1}, \mathbf{1})$ [QFT notes (3.5.128)], and $P_L = \mathrm{diag}(\mathbf{1}, 0)$, $P_R = \mathrm{diag}(0, \mathbf{1})$. In the Dirac representation $\gamma^5_{\mathrm{D}} = U^\dagger\gamma^5_{\mathrm{W}}U = \begin{pmatrix} 0 & \mathbf{1} \\ \mathbf{1} & 0 \end{pmatrix}$: chirality is *not* diagonal there, which is why the Weyl basis is used for chiral physics.

Watch out: the sign of γ^5 and a typo in the QFT notes

With $\gamma^5_{\mathrm{W}} = \mathrm{diag}(-\mathbf{1}, \mathbf{1})$, the left-handed spinor $\psi = (\psi_L, 0)$ has $\gamma^5\psi = -\psi$, and $P_L = \frac{1}{2}(1-\gamma^5)$ picks it out. The QFT notes say (just after (3.5.128)) that a left-handed spinor has eigenvalue $+1$; that is a slip, as their own projector $P_L = \frac{1}{2}(1 - \gamma^5)$ shows. Other books (Srednicki, some older texts) use $\gamma^5 = -i\gamma^0\gamma^1\gamma^2\gamma^3$ or put ψ_L in the lower components. Always check which components P_L projects onto.

T4.3.4 Lorentz transformations of spinors

Why do anticommuting matrices have anything to do with relativity? Because their commutators generate Lorentz transformations. The QFT notes define

$$S^{\mu\nu} := \frac{i}{4}[\gamma^\mu, \gamma^\nu] \tag{T4.29}$$

[QFT notes (3.5.2)] and show, using only $\{\gamma^\mu, \gamma^\nu\} = 2\eta^{\mu\nu}$, that these obey the commutation relations of the Lorentz algebra [QFT notes (3.5.5)]. The spinor transformation with parameters $\omega_{\mu\nu} = -\omega_{\nu\mu}$ is

$$\psi \longmapsto \Lambda_{1/2}\psi, \qquad \Lambda_{1/2} = \exp\Big(-\frac{i}{2}\omega_{\mu\nu}S^{\mu\nu}\Big). \tag{T4.30}$$

Let us compute the generators in the Weyl representation.

Worked example T4.18: boosts and rotations as block-diagonal matrices

For $\mu \neq \nu$, $[\gamma^\mu, \gamma^\nu] = 2\gamma^\mu\gamma^\nu$, so $S^{\mu\nu} = \frac{i}{2}\gamma^\mu\gamma^\nu$. Then

$$S^{0i} \overset{(1)}{=} \frac{i}{2}\begin{pmatrix}0 & \mathbf{1}\\ \mathbf{1} & 0\end{pmatrix}\begin{pmatrix}0 & \sigma^i\\ -\sigma^i & 0\end{pmatrix} \overset{(2)}{=} \frac{i}{2}\begin{pmatrix}-\sigma^i & 0\\ 0 & \sigma^i\end{pmatrix} = -\frac{i}{2}\begin{pmatrix}\sigma^i & 0\\ 0 & -\sigma^i\end{pmatrix},$$
$$S^{ij} \overset{(3)}{=} \frac{i}{2}\begin{pmatrix}-\sigma^i\sigma^j & 0\\ 0 & -\sigma^i\sigma^j\end{pmatrix} \overset{(4)}{=} \frac{1}{2}\varepsilon^{ijk}\begin{pmatrix}\sigma^k & 0\\ 0 & \sigma^k\end{pmatrix} \quad (i \neq j).$$

(1) $S^{0i} = \frac{i}{2}\gamma^0\gamma^i$ with the Weyl matrices.
(2) Block multiplication (as in Example T4.17).
(3) Two off-diagonal matrices multiply to a diagonal one with blocks $\sigma^i(-\sigma^j)$ and $(-\sigma^i)\sigma^j$.
(4) For $i \neq j$, $\sigma^i\sigma^j = i\varepsilon^{ijk}\sigma^k$, and $\frac{i}{2}\cdot(-i) = \frac{1}{2}$.

These agree with [QFT notes (3.5.15), (3.5.17)]. All six generators are block diagonal, so Lorentz transformations never mix the upper and lower pairs of components. The rotation generators S^{ij} are Hermitian; the boost generators S^{0i} are anti-Hermitian, so boosts act by non-unitary matrices. This is unavoidable: the Lorentz group is non-compact, and a non-compact simple group has no faithful finite-dimensional unitary representations.

Worked example T4.19: finite rotations and the minus sign at 2π

A rotation by angle θ about the z-axis has $\omega_{12} = -\omega_{21} = \theta$ and all other parameters zero, so $-\frac{i}{2}\omega_{\mu\nu}S^{\mu\nu} = -i\theta S^{12}$. Thus

$$\Lambda_{1/2}(\theta) \overset{(1)}{=} \exp\Big(-\frac{i\theta}{2}\begin{pmatrix}\sigma^3 & 0\\ 0 & \sigma^3\end{pmatrix}\Big) \overset{(2)}{=} \operatorname{diag}\big(e^{-i\theta/2}, e^{i\theta/2}, e^{-i\theta/2}, e^{i\theta/2}\big).$$

(1) The two terms $\omega_{12}S^{12} + \omega_{21}S^{21} = 2\theta S^{12}$, and $S^{12} = \frac{1}{2}\operatorname{diag}(\sigma^3, \sigma^3)$ from the previous example.
(2) The exponent is diagonal, so exponentiate entry by entry.

At $\theta = 2\pi$ every entry is $e^{\pm i\pi} = -1$:

$$\Lambda_{1/2}(2\pi) = -\mathbf{1}_4.$$

A rotation by 2π, which is the identity on vectors, multiplies every Dirac spinor by -1. Only at $\theta = 4\pi$ do we return to $+\mathbf{1}$. This is the first appearance of the double cover Spin $\to$ SO (Section T4.5).

Worked example T4.20: a finite boost stretches spinors

A boost with rapidity β along z has $\omega_{03} = -\omega_{30} = \beta$, so the exponent is $-i\beta S^{03} = -i\beta\cdot(-\frac{i}{2})\operatorname{diag}(\sigma^3, -\sigma^3) = -\frac{\beta}{2}\operatorname{diag}(\sigma^3, -\sigma^3)$ and

$$\Lambda_{1/2} = \operatorname{diag}\big(e^{-\beta/2},\, e^{\beta/2},\, e^{\beta/2},\, e^{-\beta/2}\big).$$

The entries are real and not of modulus one: boosts rescale the components. On the upper (left-handed) block the boost is $e^{-\beta\sigma^3/2}$, on the lower (right-handed) block it is $e^{+\beta\sigma^3/2}$. The two blocks transform *inequivalently*; this is the content of [QFT notes (3.5.43)],

$$\psi_L \mapsto \big(1 - i\vec{\theta}\cdot\tfrac{\vec{\sigma}}{2} - \vec{\beta}\cdot\tfrac{\vec{\sigma}}{2}\big)\psi_L, \qquad \psi_R \mapsto \big(1 - i\vec{\theta}\cdot\tfrac{\vec{\sigma}}{2} + \vec{\beta}\cdot\tfrac{\vec{\sigma}}{2}\big)\psi_R.$$

Gammas are invariant tensors. The key identity that makes the Dirac equation Lorentz covariant is

$$[\gamma^\mu, S^{\rho\sigma}] = i\big(\eta^{\mu\rho}\gamma^\sigma - \eta^{\mu\sigma}\gamma^\rho\big). \tag{T4.31}$$

It follows from the identity $[A, BC] = \{A, B\}C - B\{A, C\}$:

$$\begin{aligned}
[\gamma^\mu, S^{\rho\sigma}] &\overset{(1)}{=} \frac{i}{4}\big([\gamma^\mu, \gamma^\rho\gamma^\sigma] - [\gamma^\mu, \gamma^\sigma\gamma^\rho]\big) \\
&\overset{(2)}{=} \frac{i}{4}\big(\{\gamma^\mu, \gamma^\rho\}\gamma^\sigma - \gamma^\rho\{\gamma^\mu, \gamma^\sigma\} - \{\gamma^\mu, \gamma^\sigma\}\gamma^\rho + \gamma^\sigma\{\gamma^\mu, \gamma^\rho\}\big) \\
&\overset{(3)}{=} \frac{i}{4}\big(4\eta^{\mu\rho}\gamma^\sigma - 4\eta^{\mu\sigma}\gamma^\rho\big) = i\big(\eta^{\mu\rho}\gamma^\sigma - \eta^{\mu\sigma}\gamma^\rho\big).
\end{aligned}$$

(1) Definition (T4.29), expanding the commutator.
(2) $[A, BC] = ABC - BCA = (AB + BA)C - B(AC + CA)$.
(3) $\{\gamma^\mu, \gamma^\nu\} = 2\eta^{\mu\nu}$; the numbers commute with the matrices, so the four terms combine in pairs.

The right-hand side of (T4.31) is exactly an infinitesimal Lorentz transformation acting on the vector index of γ^μ. Exponentiating gives [QFT notes (3.5.26)]

$$\Lambda_{1/2}^{-1}\gamma^\mu\Lambda_{1/2} = \Lambda^\mu{}_\nu\gamma^\nu, \tag{T4.32}$$

where $\Lambda^\mu{}_\nu$ is the vector Lorentz matrix with the same parameters. In words: rotating the spinor indices of γ^μ is the same as rotating its vector index. With this, if ψ solves the Dirac equation then so does $\psi'(x) = \Lambda_{1/2}\psi(\Lambda^{-1}x)$ [QFT notes (3.5.29)]: the operator $i\gamma^\mu\partial_\mu$ transforms into $\Lambda_{1/2}\, i\gamma^\nu\partial_\nu\Lambda_{1/2}^{-1}$ because the change of γ^μ cancels the change of ∂_μ. The same identity in Euclidean signature is the statement that $v \mapsto uvu^{-1}$ is a rotation, which we make the definition of the spin group in Section T4.5.

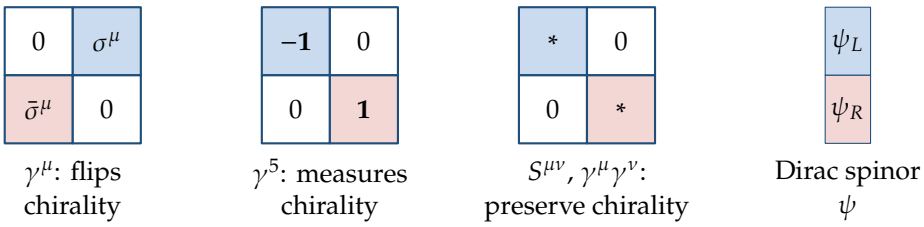


Figure T4.6: Block structure in the Weyl representation. A single γ^μ is off-diagonal, so it exchanges the left-handed (blue) and right-handed (red) halves of a Dirac spinor. Products of an even number of γ's, in particular the Lorentz generators, are block diagonal: Lorentz transformations never mix ψ_L and ψ_R. γ^5 is diagonal with entries ∓ 1 and simply reads off which half we are in.

T4.3.5 Weyl spinors

Because every Lorentz generator is block diagonal, the four-component Dirac spinor splits into two two-component pieces that transform separately:

$$\psi = \begin{pmatrix} \psi_L \\ \psi_R \end{pmatrix}, \qquad P_L\psi = \begin{pmatrix} \psi_L \\ 0 \end{pmatrix}, \quad P_R\psi = \begin{pmatrix} 0 \\ \psi_R \end{pmatrix}.$$

ψ_L and ψ_R are **Weyl spinors**. They are the two inequivalent two-dimensional representations of the Lorentz group, often written $(\frac{1}{2}, 0)$ and $(0, \frac{1}{2})$.

Worked example T4.21: the Dirac equation in chiral components

With the Weyl matrices,

$$(i\gamma^\mu\partial_\mu - m)\psi \overset{(1)}{=} \begin{pmatrix} -m & i\sigma^\mu\partial_\mu \\ i\bar\sigma^\mu\partial_\mu & -m \end{pmatrix}\begin{pmatrix} \psi_L \\ \psi_R \end{pmatrix} \overset{(2)}{=} \begin{pmatrix} i\sigma^\mu\partial_\mu\psi_R - m\psi_L \\ i\bar\sigma^\mu\partial_\mu\psi_L - m\psi_R \end{pmatrix}.$$

(1) Insert $\gamma^\mu = \left(\begin{smallmatrix} 0 & \sigma^\mu \\ \bar\sigma^\mu & 0 \end{smallmatrix}\right)$; the mass term is $m\mathbf{1}_4$.
(2) Multiply the block matrix into the column.

The mass couples ψ_L to ψ_R. For $m = 0$ the equations decouple into the **Weyl equations** [QFT notes (3.5.52)]

$$\bar\sigma^\mu\partial_\mu\psi_L = (\partial_0 - \vec\sigma\cdot\nabla)\psi_L = 0, \qquad \sigma^\mu\partial_\mu\psi_R = (\partial_0 + \vec\sigma\cdot\nabla)\psi_R = 0.$$

Try a plane wave $\psi_L = \xi\, e^{-iEt + i\vec p\cdot\vec x}$ with a constant two-spinor ξ:

$$0 = (\partial_0 - \vec\sigma\cdot\nabla)\psi_L \overset{(1)}{=} (-iE - i\vec\sigma\cdot\vec p)\psi_L \quad\Longrightarrow\quad (\vec\sigma\cdot\hat p)\,\xi \overset{(2)}{=} -\frac{E}{|\vec p|}\,\xi.$$

(1) $\partial_0 \to -iE$ and $\nabla \to i\vec p$ on the plane wave.
(2) Divide by $-i|\vec p|$; here $\hat p = \vec p/|\vec p|$.

Since $(\vec\sigma\cdot\hat p)^2 = 1$, its eigenvalues are ±1, so $E = \pm|\vec p|$ (massless dispersion). For positive energy $\vec\sigma\cdot\hat p\,\xi = -\xi$: the spin points *opposite* to the momentum. The **helicity** $\frac{1}{2}\vec\sigma\cdot\hat p$ is $-\frac{1}{2}$, and the particle is left-handed. The same computation gives helicity $+\frac{1}{2}$ for ψ_R. This is where the names come from.

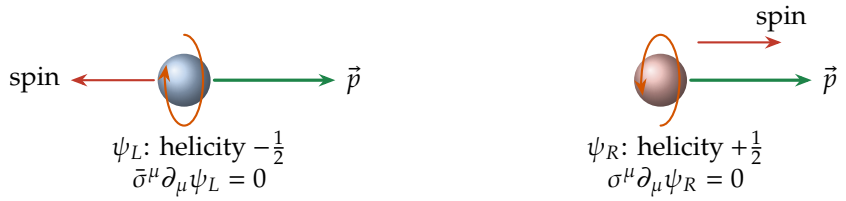


Figure T4.7: Massless Weyl particles. A positive-energy solution of the left-handed Weyl equation has its spin anti-parallel to its momentum; a right-handed one has it parallel. For a massless particle this is Lorentz invariant, because no boost can overtake a particle moving at the speed of light and reverse its momentum.

Worked example T4.22: $\sigma^2\psi_L^*$ is right-handed

We need the identity $\sigma^2(\sigma^i)^* = -\sigma^i\sigma^2$. For $i = 2$: $(\sigma^2)^* = -\sigma^2$, so $\sigma^2(\sigma^2)^* = -(\sigma^2)^2 = -\sigma^2\sigma^2$. For $i = 1, 3$: σ^1, σ^3 are real, and they anticommute with σ^2. Now take the complex conjugate

of the left-handed transformation law:

$$\sigma^2\psi_L^* \overset{(1)}{=} \sigma^2\Big(1 + i\vec{\theta}\cdot\frac{\vec{\sigma}^*}{2} - \vec{\beta}\cdot\frac{\vec{\sigma}^*}{2}\Big)\psi_L^* \overset{(2)}{=} \Big(1 - i\vec{\theta}\cdot\frac{\vec{\sigma}}{2} + \vec{\beta}\cdot\frac{\vec{\sigma}}{2}\Big)\sigma^2\psi_L^*.$$

(1) Complex conjugate of $\big(1 - i\vec{\theta}\cdot\frac{\vec{\sigma}}{2} - \vec{\beta}\cdot\frac{\vec{\sigma}}{2}\big)\psi_L$; θ, β are real.
(2) Move σ^2 to the right using $\sigma^2\sigma^{i*} = -\sigma^i\sigma^2$; this flips the sign of both the rotation and boost terms.

This is exactly the right-handed law [QFT notes (3.5.46)]. So the complex conjugate of a left-handed spinor, rotated by σ^2 = "$i\times$ the SU(2)-invariant antisymmetric tensor", is right-handed. One Weyl spinor therefore contains the same information as the other: a Majorana fermion can be built from ψ_L alone, and the charge conjugation matrix of the Dirac theory is built from σ^2.

T4.3.6 Bilinears, currents and traces

Observables are built from pairs of spinors. Because boosts are not unitary, $\psi^\dagger\psi$ is not Lorentz invariant (it is the time component of a current). The fix is the **Dirac conjugate** $\bar\psi := \psi^\dagger\gamma^0$ [QFT notes (3.5.32)], which transforms as $\bar\psi \mapsto \bar\psi\Lambda_{1/2}^{-1}$. Then $\bar\psi\psi$ is a scalar, and $\bar\psi\gamma^\mu\psi$ is a vector by (T4.32).

A general bilinear is $\bar\psi\Gamma\psi$ with Γ any 4×4 matrix. The 4×4 matrices form a 16-dimensional space, and the 16 basis monomials of the Clifford algebra (Section T4.2.2) are a basis. It is convenient to trade the three- and four-index monomials for $\gamma^\mu\gamma^5$ and γ^5, via [QFT notes (3.5.131)] $\gamma^{[\mu}\gamma^\nu\gamma^{\rho]} = i\varepsilon^{\mu\nu\rho\sigma}\gamma_\sigma\gamma^5$ (with $\varepsilon^{0123} = +1$).

basis matrix Γ	grade in Cl	number	$\bar\psi\Gamma\psi$ transforms as
$\mathbf{1}$	0	1	scalar
γ^μ	1	4	vector
$\sigma^{\mu\nu} = \frac{i}{2}[\gamma^\mu,\gamma^\nu] = 2S^{\mu\nu}$	2	6	antisymmetric tensor
$\gamma^\mu\gamma^5$	3	4	axial vector
γ^5	4	1	pseudoscalar
total		$16 = 2^4$	

Table T4.2: The sixteen Dirac bilinears [QFT notes Table 3.5.1]. The grade is the number of γ's in the corresponding Clifford monomial; the numbers are the binomial coefficients $\binom{4}{k}$. "Pseudo" and "axial" mean an extra sign under parity, which comes from the γ^5.

Worked example T4.23: vector and axial currents

Let ψ solve the Dirac equation, so $\gamma^\mu\partial_\mu\psi = -im\psi$ and (taking the Dirac conjugate) $\partial_\mu\bar\psi\gamma^\mu = im\bar\psi$. Then for $j^\mu = \bar\psi\gamma^\mu\psi$ and $j^{\mu5} = \bar\psi\gamma^\mu\gamma^5\psi$:

$$\partial_\mu j^\mu \overset{(1)}{=} (\partial_\mu\bar\psi)\gamma^\mu\psi + \bar\psi\gamma^\mu\partial_\mu\psi \overset{(2)}{=} im\bar\psi\psi - im\bar\psi\psi = 0,$$
$$\partial_\mu j^{\mu5} \overset{(3)}{=} (\partial_\mu\bar\psi)\gamma^\mu\gamma^5\psi - \bar\psi\gamma^5\gamma^\mu\partial_\mu\psi \overset{(4)}{=} im\bar\psi\gamma^5\psi + im\bar\psi\gamma^5\psi = 2im\,\bar\psi\gamma^5\psi.$$

(1) Product rule.
(2) The Dirac equation and its conjugate.

(3) Product rule, and move γ^5 to the left past γ^μ with a sign.
(4) The Dirac equation and its conjugate again.

So the vector current is always conserved, and the axial current is conserved when $m = 0$ [QFT notes (3.5.136)–(3.5.137)]. They are the Noether currents of $\psi \mapsto e^{i\alpha}\psi$ and $\psi \mapsto e^{i\alpha\gamma^5}\psi$. The second conservation law is classical only: in the quantum theory it fails by a term proportional to $\varepsilon^{\mu\nu\rho\sigma}F_{\mu\nu}F_{\rho\sigma}$, the chiral anomaly, and Nash's Chapter 10 (Chapter 10) explains it as the index of a Dirac operator.

Worked example T4.24: trace identities from the Clifford relation alone

We already know $\operatorname{tr}\gamma^\mu = 0$ (Section T4.1.2). The same argument with γ^5 inserted shows that the trace of any *odd* number of γ's vanishes: $\operatorname{tr}(\gamma^{\mu_1}\cdots\gamma^{\mu_{2j+1}}) = \operatorname{tr}(\gamma^5\gamma^5\gamma^{\mu_1}\cdots) = -\operatorname{tr}(\gamma^5\gamma^{\mu_1}\cdots\gamma^5) = -\operatorname{tr}(\cdots)$ (move one γ^5 through $2j+1$ gammas, then use cyclicity). For two:

$$\operatorname{tr}(\gamma^\mu\gamma^\nu) \overset{(1)}{=} \tfrac{1}{2}\operatorname{tr}\{\gamma^\mu,\gamma^\nu\} \overset{(2)}{=} \tfrac{1}{2}\cdot 2\eta^{\mu\nu}\operatorname{tr}\mathbf{1}_4 = 4\eta^{\mu\nu}.$$

(1) Cyclicity: $\operatorname{tr}(\gamma^\mu\gamma^\nu) = \operatorname{tr}(\gamma^\nu\gamma^\mu)$.
(2) The Clifford relation.

For four, anticommute γ^μ all the way to the right, then use cyclicity:

$$\begin{aligned}\operatorname{tr}(\gamma^\mu\gamma^\nu\gamma^\rho\gamma^\sigma) &\overset{(1)}{=} 2\eta^{\mu\nu}\operatorname{tr}(\gamma^\rho\gamma^\sigma) - 2\eta^{\mu\rho}\operatorname{tr}(\gamma^\nu\gamma^\sigma) + 2\eta^{\mu\sigma}\operatorname{tr}(\gamma^\nu\gamma^\rho) - \operatorname{tr}(\gamma^\nu\gamma^\rho\gamma^\sigma\gamma^\mu)\\ &\overset{(2)}{=} 4\big(\eta^{\mu\nu}\eta^{\rho\sigma} - \eta^{\mu\rho}\eta^{\nu\sigma} + \eta^{\mu\sigma}\eta^{\nu\rho}\big).\end{aligned}$$

(1) Each step $\gamma^\mu\gamma^\alpha = 2\eta^{\mu\alpha} - \gamma^\alpha\gamma^\mu$ moves γ^μ one place right, leaving a term with the pair removed.
(2) The last trace equals the left-hand side by cyclicity; bring it over, divide by 2, and use the two-gamma trace.

Finally the trace with γ^5: $\operatorname{tr}(\gamma^5\gamma^\mu\gamma^\nu\gamma^\rho\gamma^\sigma)$ is totally antisymmetric (repeated indices reduce it to $\operatorname{tr}(\gamma^5\gamma^\alpha\gamma^\beta)$, which vanishes by the same kind of argument), so it is a multiple of $\varepsilon^{\mu\nu\rho\sigma}$. The multiple is fixed by one component: $\gamma^0\gamma^1\gamma^2\gamma^3 = -i\gamma^5$, so $\operatorname{tr}(\gamma^5\gamma^0\gamma^1\gamma^2\gamma^3) = -i\operatorname{tr}(\gamma^5)^2 = -4i$, i.e.

$$\operatorname{tr}(\gamma^5\gamma^\mu\gamma^\nu\gamma^\rho\gamma^\sigma) = -4i\,\varepsilon^{\mu\nu\rho\sigma} \qquad (\varepsilon^{0123} = +1).$$

This is the trace that produces the $\varepsilon^{\mu\nu\rho\sigma}F_{\mu\nu}F_{\rho\sigma}$ of the anomaly (and the $\hat{A}$ and ch forms of the index theorem) in the heat-kernel calculation of Chapter 10.

Physical picture: why the trace with γ^5 needs four gammas

In $n = 2k$ dimensions the chirality operator is proportional to the product of all $2k$ gammas. A trace of Γ times fewer than $2k$ gammas vanishes, because such a product is a Clifford monomial of the wrong grade and all non-trivial monomials are traceless. So the first non-zero trace $\operatorname{tr}(\Gamma\gamma^{a_1}\cdots\gamma^{a_{2k}}) \propto \varepsilon^{a_1\ldots a_{2k}}$ needs exactly one of each direction. In the index theorem this is why, among all the terms of a heat-kernel expansion, only the top-degree form (the one you can integrate over M^{2k}) survives the supertrace.

T4.3.7 The $\mathrm{SL}(2,\mathbb{C})$ picture

There is a neat way to see all of this at once. To a four-vector x^μ associate the Hermitian 2×2 matrix

$$X := x^0\mathbf{1} + \vec{x}\cdot\vec{\sigma} = \begin{pmatrix} x^0+x^3 & x^1 - ix^2 \\ x^1+ix^2 & x^0-x^3 \end{pmatrix}. \tag{T4.33}$$

Its determinant is the Minkowski norm: $\det X = (x^0)^2 - (x^3)^2 - (x^1-ix^2)(x^1+ix^2) = (x^0)^2 - |\vec{x}|^2$. Any $A \in \mathrm{SL}(2,\mathbb{C})$ acts by $X \mapsto AXA^\dagger$. The result is again Hermitian, and $\det(AXA^\dagger) = |\det A|^2 \det X = \det X$. So each A gives a Lorentz transformation, and A and $-A$ give the same one. This is the Lorentzian analogue of $\mathrm{Spin}(3) = \mathrm{SU}(2) \to \mathrm{SO}(3)$, and it says $\mathrm{Spin}(1,3) \cong \mathrm{SL}(2,\mathbb{C})$; we return to it in Section T4.5.7. One Weyl spinor transforms by A and the other by $(A^\dagger)^{-1}$; for A unitary (a rotation, $A = e^{-i\vec{\theta}\cdot\vec{\sigma}/2}$) these agree, while for A Hermitian (a boost, $A = e^{\mp\vec{\beta}\cdot\vec{\sigma}/2}$) they are inverse to each other, matching Example T4.20.

Reading Nash (p. 114, (4.99)–(4.101))

Nash works in Euclidean signature throughout, and uses anti-Hermitian $\gamma_a^{\text{Nash}} = i\gamma_a$ with $\{\gamma_a,\gamma_b\} = -2\delta_{ab}$. His chirality operator is $\gamma_5 = i^{n(n+1)/2}\gamma_1\cdots\gamma_n$. For $n=4$, $i^{10} = -1$ and $\gamma_1^{\text{Nash}}\cdots\gamma_4^{\text{Nash}} = i^4\gamma_1\cdots\gamma_4$, so $\gamma_5^{\text{Nash}} = -\gamma_1\gamma_2\gamma_3\gamma_4 = \Gamma$, exactly our (T4.24); and his (4.101), $\gamma_5 = \mathrm{diag}(I,-I)$, is our Euclidean chiral representation. Note that this Euclidean Γ is not the Minkowski $\gamma^5 = i\gamma^0\gamma^1\gamma^2\gamma^3$: the two are related by the Wick rotation $\gamma^0 \leftrightarrow \gamma_4$, $\gamma^j \leftrightarrow i\gamma_j$ (up to convention-dependent signs). Nash's remark on p. 115 that "in a space of Lorentzian signature the Dirac operator is not elliptic" is visible here: the Minkowski symbol $\gamma^\mu p_\mu$ squares to $p^\mu p_\mu$, which vanishes on the light cone.

Recap

- Euclidean 4D: $\gamma_a = \left(\begin{smallmatrix} 0 & \tau_a \\ \tau_a^\dagger & 0 \end{smallmatrix}\right)$ with $\tau_a = (i\vec{\sigma},\mathbf{1})$; $\Gamma = -\gamma_1\gamma_2\gamma_3\gamma_4 = \mathrm{diag}(\mathbf{1},-\mathbf{1})$.
- Minkowski 4D: Dirac rep (γ^0 diagonal) and Weyl rep ($\gamma^\mu = \left(\begin{smallmatrix} 0 & \sigma^\mu \\ \bar{\sigma}^\mu & 0 \end{smallmatrix}\right)$), related by an explicit unitary U.
- $\gamma^5 = i\gamma^0\gamma^1\gamma^2\gamma^3$ is Hermitian, squares to 1, anticommutes with all γ^μ; in the Weyl rep it is $\mathrm{diag}(-\mathbf{1},\mathbf{1})$ and $\psi = (\psi_L,\psi_R)$.
- $S^{\mu\nu} = \frac{i}{4}[\gamma^\mu,\gamma^\nu]$ generate Lorentz transformations; they are block diagonal, so ψ_L,ψ_R do not mix; a 2π rotation gives -1.
- Sixteen bilinears $= 2^4$ Clifford monomials; the axial current is classically conserved only for $m=0$; $\mathrm{tr}(\gamma^5\gamma^\mu\gamma^\nu\gamma^\rho\gamma^\sigma) = -4i\varepsilon^{\mu\nu\rho\sigma}$.

T4.4 Gamma matrices in any dimension

Nash's index theorem for the Dirac operator is stated on a manifold of any even dimension $n = 2k$; the anomalies of Chapter 10 appear in QED in $2k$ dimensions; superstrings live in ten. So we need gamma matrices in every dimension, and we need to know how big they are, what the chirality operator is, and whether they can be chosen real. The answer is a construction so simple that it fits on one line (tensor products of Pauli matrices), plus a classification with a surprising period of eight.

T4.4.1 The tensor product construction

Recall that the tensor (Kronecker) product of matrices is $A \otimes B = (a_{ij}B)$, a block matrix whose (i,j) block is $a_{ij}B$. Its two properties we need are

$$(A \otimes B)(C \otimes D) = AC \otimes BD, \qquad (A \otimes B)^\dagger = A^\dagger \otimes B^\dagger. \tag{T4.34}$$

So two tensor products anticommute exactly when an *odd* number of their factors anticommute (and the rest commute).

Working definition T4.6 (gamma matrices in dimension $2k$)

For $j = 1, \dots, k$ put

$$\begin{aligned} \gamma_{2j-1} &:= \underbrace{\sigma_3 \otimes \cdots \otimes \sigma_3}_{j-1} \otimes \sigma_1 \otimes \underbrace{\mathbf{1} \otimes \cdots \otimes \mathbf{1}}_{k-j}, \\ \gamma_{2j} &:= \underbrace{\sigma_3 \otimes \cdots \otimes \sigma_3}_{j-1} \otimes \sigma_2 \otimes \underbrace{\mathbf{1} \otimes \cdots \otimes \mathbf{1}}_{k-j}. \end{aligned} \tag{T4.35}$$

[each is a tensor product of k Pauli-type factors, and the j-th factor is where σ_1 or σ_2 sits]. These are $2^k \times 2^k$ Hermitian matrices with $\{\gamma_a, \gamma_b\} = 2\delta_{ab}$. The **chirality operator** [a Hermitian matrix with $\Gamma^2 = \mathbf{1}$ that anticommutes with every γ_a, so its eigenvalues ± 1 split spinors into two halves] is

$$\Gamma := (-i)^k \gamma_1 \gamma_2 \cdots \gamma_{2k} = \sigma_3 \otimes \sigma_3 \otimes \cdots \otimes \sigma_3. \tag{T4.36}$$

Let us verify the claims.

Squares. Each factor in (T4.35) squares to $\mathbf{1}$, so by (T4.34) $\gamma_a^2 = \mathbf{1} \otimes \cdots \otimes \mathbf{1}$.

Anticommutation. Compare γ_{2j-1} with γ_{2j}: they differ only in slot j, where σ_1 and σ_2 anticommute; all other slots are equal and commute. One anticommuting slot: they anticommute. Now compare γ_a in "pair" j with γ_b in a later pair $l > j$. In slot j, γ_a has σ_1 or σ_2 while γ_b has σ_3: these anticommute. In every other slot at least one of the two has $\mathbf{1}$ or both have σ_3: these commute. Again exactly one anticommuting slot. The string of σ_3's is there precisely to produce this one sign.

Chirality. Group the product in pairs:

$$-i\gamma_{2j-1}\gamma_{2j} \overset{(1)}{=} -i\,(\sigma_3^2)^{\otimes(j-1)} \otimes \sigma_1\sigma_2 \otimes \mathbf{1}^{\otimes(k-j)} \overset{(2)}{=} \mathbf{1}^{\otimes(j-1)} \otimes \sigma_3 \otimes \mathbf{1}^{\otimes(k-j)},$$

$$\Gamma \overset{(3)}{=} \prod_{j=1}^{k} (-i\gamma_{2j-1}\gamma_{2j}) \overset{(4)}{=} \sigma_3 \otimes \sigma_3 \otimes \cdots \otimes \sigma_3.$$

(1) Multiply slot by slot using (T4.34).
(2) $\sigma_3^2 = \mathbf{1}$ and $-i\sigma_1\sigma_2 = -i \cdot i\sigma_3 = \sigma_3$.
(3) $\gamma_1\gamma_2\cdots\gamma_{2k}$ is already ordered in consecutive pairs; the k factors of $-i$ are distributed one per pair.
(4) Each pair contributes a σ_3 in its own slot.

In general (for *any* Hermitian representation) Γ is Hermitian and $\Gamma^2 = 1$. Indeed, reversing $2k$ anticommuting factors takes $\binom{2k}{2} = k(2k-1)$ transpositions, so

$$(\gamma_1 \cdots \gamma_{2k})^2 \overset{(1)}{=} (-1)^{k(2k-1)} \gamma_1 \cdots \gamma_{2k}\gamma_{2k} \cdots \gamma_1 \overset{(2)}{=} (-1)^k,$$

$$\Gamma^2 \overset{(3)}{=} (-i)^{2k}(-1)^k = (-1)^k(-1)^k = 1,$$

(1) Reverse the order of the second copy.
(2) The middle pairs $\gamma_{2k}\gamma_{2k}$, then $\gamma_{2k-1}\gamma_{2k-1}, \dots$ cancel one by one; and $k(2k-1) \equiv k \pmod 2$.
(3) The phase $(-i)^k$ was chosen for exactly this.

and $\Gamma^\dagger = i^k\gamma_{2k}\cdots\gamma_1 = i^k(-1)^k\gamma_1\cdots\gamma_{2k} = \Gamma$. Finally, Γ anticommutes with each γ_a: moving γ_a through $\gamma_1\cdots\gamma_{2k}$ passes $2k-1$ other gammas.

Worked example T4.25: dimensions 2, 4 and 6 explicitly

$d = 2$ ($k = 1$): $\gamma_1 = \sigma_1$, $\gamma_2 = \sigma_2$, $\Gamma = -i\sigma_1\sigma_2 = \sigma_3$. This is Example T4.1: chirality ± 1 are the holomorphic and antiholomorphic components.
$d = 4$ ($k = 2$):

$$\gamma_1 = \sigma_1 \otimes \mathbf{1} = \begin{pmatrix} 0 & \mathbf{1} \\ \mathbf{1} & 0 \end{pmatrix}, \quad \gamma_2 = \sigma_2 \otimes \mathbf{1} = \begin{pmatrix} 0 & -i\mathbf{1} \\ i\mathbf{1} & 0 \end{pmatrix},$$
$$\gamma_3 = \sigma_3 \otimes \sigma_1 = \begin{pmatrix} \sigma_1 & 0 \\ 0 & -\sigma_1 \end{pmatrix}, \quad \gamma_4 = \sigma_3 \otimes \sigma_2 = \begin{pmatrix} \sigma_2 & 0 \\ 0 & -\sigma_2 \end{pmatrix},$$
$$\Gamma = \sigma_3 \otimes \sigma_3 = \operatorname{diag}(1, -1, -1, 1).$$

Check one mixed pair by blocks: $\gamma_1\gamma_3 = \left(\begin{smallmatrix} 0 & -\sigma_1 \\ \sigma_1 & 0 \end{smallmatrix}\right)$ and $\gamma_3\gamma_1 = \left(\begin{smallmatrix} 0 & \sigma_1 \\ -\sigma_1 & 0 \end{smallmatrix}\right)$, so they anticommute. Here Γ is diagonal but its $+1$ eigenvectors are the first and fourth basis vectors, not the first two. Permuting the basis to $(1, 4, 2, 3)$ brings it to $\operatorname{diag}(1, 1, -1, -1)$ and the gammas to block-off-diagonal form, like (T4.21); by the uniqueness result below, the two sets are unitarily equivalent.
$d = 6$ ($k = 3$): 8×8 matrices

$$\gamma_1 = \sigma_1 \otimes \mathbf{1} \otimes \mathbf{1}, \qquad \gamma_2 = \sigma_2 \otimes \mathbf{1} \otimes \mathbf{1},$$
$$\gamma_3 = \sigma_3 \otimes \sigma_1 \otimes \mathbf{1}, \qquad \gamma_4 = \sigma_3 \otimes \sigma_2 \otimes \mathbf{1},$$
$$\gamma_5 = \sigma_3 \otimes \sigma_3 \otimes \sigma_1, \qquad \gamma_6 = \sigma_3 \otimes \sigma_3 \otimes \sigma_2,$$

with $\Gamma = (-i)^3\gamma_1\cdots\gamma_6 = \sigma_3 \otimes \sigma_3 \otimes \sigma_3 = \operatorname{diag}(1, -1, -1, 1, -1, 1, 1, -1)$. Four $+1$'s and four -1's: $S^\pm$ are each four-dimensional.

Watch out: the recursion in the QFT notes

The QFT notes [QFT notes (3.5.9)–(3.5.10)] give the same construction recursively: from $2n$-dimensional gammas $\gamma_i^{(n)}$ build $\gamma_i^{(n+1)} = \gamma_i^{(n)} \otimes \tau_3$ ($i \le 2n$), $\gamma_{2n+1}^{(n+1)} = \mathbf{1} \otimes \tau_1$, $\gamma_{2n+2}^{(n+1)} = \mathbf{1} \otimes \tau_2$. Two things to watch. First, they write $\tau_i = \frac{1}{2}\sigma_i$, but then $\tau_i^2 = \frac{1}{4}$ and the Clifford relation fails; read $\tau_i = \sigma_i$. Second, their explicit block matrices, e.g. $\gamma_i^{(n)} \otimes \tau_3 = \operatorname{diag}(\gamma_i^{(n)}, -\gamma_i^{(n)})$, use the opposite ordering of the Kronecker product from ours ($A \otimes B$ with B as the *outer* block matrix). Either ordering works; just be consistent.

T4.4.2 Spinors as a fermionic Fock space

The tensor-product structure has a physical meaning that makes the whole theory transparent. Pair the gammas into complex combinations

$$b_j := \tfrac{1}{2}(\gamma_{2j-1} + i\gamma_{2j}), \qquad b_j^\dagger = \tfrac{1}{2}(\gamma_{2j-1} - i\gamma_{2j}), \qquad j = 1, \dots, k. \tag{T4.37}$$

Then the Clifford relations become the canonical anticommutation relations of k fermionic modes:

$$\{b_j, b_l^\dagger\} \overset{(1)}{=} \tfrac{1}{4}\big(\{\gamma_{2j-1}, \gamma_{2l-1}\} + \{\gamma_{2j}, \gamma_{2l}\} - i\{\gamma_{2j-1}, \gamma_{2l}\} + i\{\gamma_{2j}, \gamma_{2l-1}\}\big) \overset{(2)}{=} \tfrac{1}{4}(2\delta_{jl} + 2\delta_{jl}) = \delta_{jl},$$

$$\{b_j, b_l\} \overset{(3)}{=} \tfrac{1}{4}\big(\{\gamma_{2j-1},\gamma_{2l-1}\} - \{\gamma_{2j},\gamma_{2l}\} + i\{\gamma_{2j-1},\gamma_{2l}\} + i\{\gamma_{2j},\gamma_{2l-1}\}\big) \overset{(4)}{=} 0. \tag{T4.38}$$

(1) Expand the bilinear anticommutator.
(2) $\{\gamma_a,\gamma_b\} = 2\delta_{ab}$; the mixed terms involve an odd and an even index, which are never equal.
(3) Same expansion without the conjugation.
(4) The two diagonal terms cancel and the mixed ones vanish.

In the representation (T4.35), $b_j = \sigma_3^{\otimes(j-1)} \otimes \sigma^+ \otimes \mathbf{1}^{\otimes(k-j)}$ with $\sigma^+ = \tfrac{1}{2}(\sigma_1 + i\sigma_2) = \left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right)$. This is the **Jordan–Wigner** representation of fermions on a chain of k spins. Let $|\uparrow\rangle = (1,0)^{\mathsf{T}}$ and $|\downarrow\rangle = (0,1)^{\mathsf{T}}$; then $\sigma^+|\uparrow\rangle = 0$. So:

- The *vacuum* $|0\rangle = |\uparrow\uparrow\cdots\uparrow\rangle$ is killed by every b_j.
- $b_j^\dagger$ flips spin j down (with a sign from the σ_3 string): it creates a fermion in mode j.
- The 2^k states $b_{j_1}^\dagger \cdots b_{j_r}^\dagger|0\rangle$ ($j_1 < \cdots < j_r$) form a basis: spinor space is the fermionic Fock space $S \cong \Lambda^\bullet \mathbb{C}^k$, of dimension 2^k.
- $\Gamma = \sigma_3^{\otimes k}$ is $+1$ on $|\uparrow\rangle$ and -1 on $|\downarrow\rangle$ in each slot, so $\Gamma = (-1)^F$ where F is the number of fermions (down spins).

Thus the chirality splitting is the split into even and odd fermion number:

$$S = S^+ \oplus S^-, \qquad S^+ = \Lambda^{\text{even}}\mathbb{C}^k, \quad S^- = \Lambda^{\text{odd}}\mathbb{C}^k, \qquad \dim S^\pm = 2^{k-1}. \tag{T4.39}$$

A single γ_a is linear in b's and $b^\dagger$'s, so it changes the fermion number by one and maps $S^\pm \to S^\mp$; a product of two gammas preserves the parity. Bär's construction [Bär pp. 85–88] is exactly this: his $z_j = \tfrac{1}{2}(e_{2j-1} - ie_{2j})$ are the b's (his e_a square to -1; with $e_a = -i\gamma_a$ one gets $z_j = -\tfrac{i}{2}(\gamma_{2j-1} - i\gamma_{2j}) = -ib_j^\dagger$), his relations $z_j^2 = \bar z_j^2 = 0$ are (T4.38), his basis $z(j_1,\dots,j_r)$ is our Fock basis, and his $\Sigma_n^\pm = \{i^m\omega\cdot z = \pm z\}$ [Bär (2.9)] is our $\Gamma = \pm 1$ (since $i^m\omega = i^m(-i)^{2m}\gamma_1\cdots\gamma_{2m} = \Gamma$).

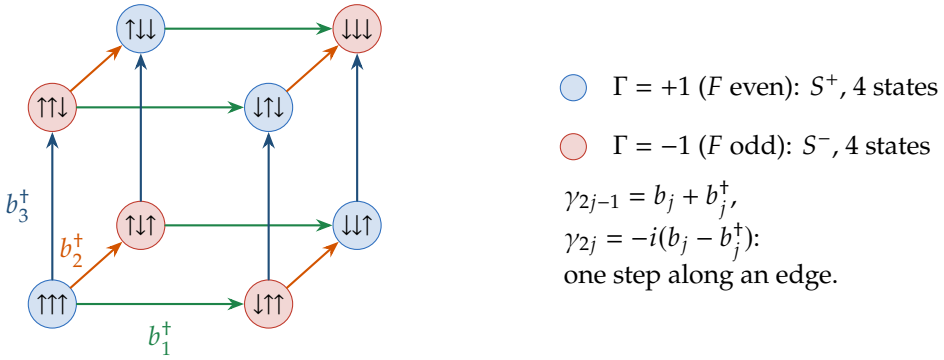


Figure T4.8: The 8-dimensional spinor space in $d = 6$ as the Fock space of three fermionic modes, drawn as a cube. Each vertex is a basis state; a gamma matrix moves along one edge (creating or destroying one fermion), so it always connects a blue vertex (S^+) to a red one (S^-). Chirality is fermion parity.

T4.4.3 Uniqueness: all irreducible representations are equivalent

We have used the word "the" spinor representation. Here is why that is justified in even dimensions.

Key idea: uniqueness of the spinor representation, $n = 2k$

Any set of Hermitian matrices $\gamma'_1, \dots, \gamma'_{2k}$ on $\mathbb{C}^N$ with $\{\gamma'_a, \gamma'_b\} = 2\delta_{ab}$ is unitarily equivalent to a direct sum of copies of (T4.35). In particular N is a multiple of 2^k, and an irreducible set ($N = 2^k$) is unique up to a change of basis.

Why (a Fock-space argument). Form $b'_j = \frac{1}{2}(\gamma'_{2j-1} + i\gamma'_{2j})$; they obey (T4.38). The operators $n_j = b'^\dagger_j b'_j$ are commuting projectors (from the CAR, $n_j^2 = b'^\dagger_j(1 - b'^\dagger_j b'_j)b'_j = n_j$), so they can be diagonalised together. Take a joint eigenvector and apply b'_j to it wherever $n_j = 1$: since $b'_j n_j = b'_j$ and $b'^2_j = 0$, this produces a non-zero vector $|0'\rangle$ with $b'_j|0'\rangle = 0$ for all j, a "vacuum". Then the 2^k vectors $b'^\dagger_{j_1} \cdots b'^\dagger_{j_r}|0'\rangle$ are orthonormal (compute their inner products with the CAR, exactly as for a harmonic oscillator) and span a subspace on which the γ''s act by the *same* matrices as in (T4.35). The orthogonal complement is again invariant, so repeat. The map sending one Fock basis to the other is the unitary equivalence.

A second, more algebraic way to say it: the 2^{2k} matrices γ_I (products over subsets I) of (T4.35) are orthogonal for the trace inner product, $\mathrm{tr}(\gamma_I^\dagger \gamma_J) = 2^k \delta_{IJ}$ (for $I \neq J$ the product is $\pm\gamma_{I\triangle J}$, which is traceless). So they are linearly independent and span all $2^k \times 2^k$ matrices: the complexified Clifford algebra is a full matrix algebra,

$$\mathrm{Cl}_{2k} \otimes \mathbb{C} \cong \mathbb{C}(2^k), \tag{T4.40}$$

which is Bär's Prop. 2.3.11. A full matrix algebra has only one irreducible module, the column vectors. This is the fact we used in Example T4.16 to relate the Dirac and Weyl representations.

T4.4.4 Odd dimensions

In dimension $2k + 1$ we need one more matrix anticommuting with $\gamma_1, \dots, \gamma_{2k}$ and squaring to one. We already have it: Γ. So

$$\gamma_{2k+1} := \pm\Gamma \tag{T4.41}$$

gives a representation of the $(2k + 1)$-dimensional Clifford algebra on the same 2^k-dimensional space. For $k = 1$ this is $\sigma_1, \sigma_2, \pm\sigma_3$: the Pauli matrices in 3D.

The two signs give two representations that are *not* equivalent. The reason is the volume element $\omega = \gamma_1 \cdots \gamma_{2k+1}$. In odd dimensions ω commutes with every generator (moving γ_a through it passes $2k$ other gammas), so it is central, and it is represented by a multiple of the identity: with $\gamma_{2k+1} = \pm\Gamma$,

$$\gamma_1 \cdots \gamma_{2k}\gamma_{2k+1} \overset{(1)}{=} \pm i^k \Gamma \cdot \Gamma \overset{(2)}{=} \pm i^k \mathbf{1}.$$

(1) $\gamma_1 \cdots \gamma_{2k} = i^k \Gamma$ by (T4.36).
(2) $\Gamma^2 = 1$.

A change of basis cannot change a multiple of the identity, so the sign is an invariant. This is the $\mathrm{Cl}_{0,3} = \mathbb{H} \oplus \mathbb{H}$ phenomenon of Example T4.12 in general: in odd dimensions the complex Clifford algebra is $\mathbb{C}(2^k) \oplus \mathbb{C}(2^k)$ and has two irreducible representations, one for each value of ω. Either one restricts to the *same* representation of the spin group, because the spin group lies in the even part, which does not see the sign of γ_{2k+1}. Bär realises the odd-dimensional spinor space as $\Sigma_{2k+1} := \Sigma^+_{2k+2}$, using $\mathrm{Cl}_{2k+1} \cong \mathrm{Cl}^0_{2k+2}$ [Bär Lemma 2.3.12].

Key idea: size of spinors

In dimension n the irreducible complex spinor space has dimension $2^{\lfloor n/2 \rfloor}$ [Nash (4.97)]: $1, 2, 2, 4, 4, 8, 8, 16, \ldots$ for $n = 1, 2, 3, \ldots$ In even dimensions it splits into $S^+ \oplus S^-$, each of dimension $2^{n/2-1}$; in odd dimensions it does not split.

T4.4.5 Real Clifford algebras and the period eight

Over $\mathbb{C}$ the story is now complete: period two, with a full matrix algebra in even dimension and two copies in odd dimension. Over $\mathbb{R}$, and for mixed signatures, the algebra can be a matrix algebra over $\mathbb{R}$, $\mathbb{C}$ or $\mathbb{H}$, and the answer repeats with period eight. This matters in physics for one reason: *real* (Majorana) spinors exist exactly when the relevant Clifford algebra is a real matrix algebra.

We have computed the first few cases by hand: $\mathrm{Cl}_{1,0} = \mathbb{R} \oplus \mathbb{R}$, $\mathrm{Cl}_{0,1} = \mathbb{C}$, $\mathrm{Cl}_{2,0} = \mathbb{R}(2)$, $\mathrm{Cl}_{0,2} = \mathbb{H}$, $\mathrm{Cl}_{3,0} = \mathbb{C}(2)$, $\mathrm{Cl}_{0,3} = \mathbb{H} \oplus \mathbb{H}$ (Section T4.2). The rest follows from two shifting rules.

Key idea: shifting by two dimensions

$$\mathrm{Cl}_{0,n+2} \cong \mathrm{Cl}_{n,0} \otimes \mathrm{Cl}_{0,2} = \mathrm{Cl}_{n,0} \otimes \mathbb{H}, \qquad \mathrm{Cl}_{n+2,0} \cong \mathrm{Cl}_{0,n} \otimes \mathrm{Cl}_{2,0} = \mathrm{Cl}_{0,n} \otimes \mathbb{R}(2). \tag{T4.42}$$

Adding two dimensions tensors with $\mathbb{H}$ or $\mathbb{R}(2)$, *and flips the sign* of the original generators.

The generators, explicitly. Let $e_1, \ldots, e_n$ generate $\mathrm{Cl}_{n,0}$ ($e_i^2 = +1$) and f_1, f_2 generate $\mathrm{Cl}_{0,2}$ ($f^2 = -1$). Inside $\mathrm{Cl}_{n,0} \otimes \mathrm{Cl}_{0,2}$ (ordinary, ungraded tensor product: the two factors commute) define

$$E_i \coloneqq e_i \otimes f_1 f_2 \ (i \le n), \qquad E_{n+1} \coloneqq 1 \otimes f_1, \qquad E_{n+2} \coloneqq 1 \otimes f_2.$$

Then

$$E_i^2 \overset{(1)}{=} e_i^2 \otimes (f_1 f_2)^2 \overset{(2)}{=} (+1)(-1) = -1, \qquad E_{n+1}^2 = E_{n+2}^2 = -1,$$
$$E_i E_{n+1} \overset{(3)}{=} e_i \otimes f_1 f_2 f_1 \overset{(4)}{=} -e_i \otimes f_1 f_1 f_2 = -E_{n+1} E_i,$$
$$E_i E_j \overset{(5)}{=} e_i e_j \otimes (f_1 f_2)^2 = -e_i e_j \otimes 1, \text{ which anticommutes with } E_j E_i \ (i \neq j).$$

(1) Rule (T4.34) for tensor products.
(2) $e_i^2 = 1$ and $(f_1 f_2)^2 = -f_1^2 f_2^2 = -1$.
(3) Multiply in each factor.
(4) $f_2 f_1 = -f_1 f_2$.
(5) Same rule; then $e_i e_j = -e_j e_i$.

The case $E_i E_{n+2}$ is identical. So the E's are $n+2$ anticommuting elements squaring to -1: they generate a copy of $\mathrm{Cl}_{0,n+2}$, and comparing dimensions ($2^{n+2} = 2^n \cdot 4$) gives the isomorphism. The second rule is the same computation with all signs reversed. (This is why the ungraded tensor product is the useful one here: the "twist" by $f_1 f_2$ supplies the anticommutation by hand.)

Now iterate, using $\mathbb{H} \otimes \mathbb{R}(2) = \mathbb{H}(2)$, $\mathbb{C} \otimes \mathbb{H} = \mathbb{C}(2)$, $\mathbb{H} \otimes \mathbb{H} = \mathbb{R}(4)$ (the last two can be checked by writing down the obvious actions on $\mathbb{C}^2$ and on $\mathbb{H} \cong \mathbb{R}^4$, $x \mapsto p x \bar{q}$). For example

$$\mathrm{Cl}_{0,4} \overset{(1)}{=} \mathrm{Cl}_{2,0} \otimes \mathbb{H} = \mathbb{R}(2) \otimes \mathbb{H} \overset{(2)}{=} \mathbb{H}(2),$$
$$\mathrm{Cl}_{0,6} \overset{(3)}{=} \mathrm{Cl}_{4,0} \otimes \mathbb{H} = \big(\mathrm{Cl}_{0,2} \otimes \mathbb{R}(2)\big) \otimes \mathbb{H} = \mathbb{H}(2) \otimes \mathbb{H} \overset{(4)}{=} \mathbb{R}(8),$$

$$\mathrm{Cl}_{0,n+8} \overset{(5)}{=} \mathrm{Cl}_{0,n} \otimes \big(\mathbb{R}(2) \otimes \mathbb{H} \otimes \mathbb{R}(2) \otimes \mathbb{H}\big) \overset{(6)}{=} \mathrm{Cl}_{0,n} \otimes \mathbb{R}(16).$$

(1) First rule of (T4.42) with $n = 2$.
(2) A 2×2 matrix with quaternion entries.
(3) First rule with $n = 4$, then the second rule for $\mathrm{Cl}_{4,0}$.
(4) $\mathbb{H} \otimes \mathbb{H} \cong \mathbb{R}(4)$, and $\mathbb{R}(2) \otimes \mathbb{R}(4) = \mathbb{R}(8)$.
(5) Apply the two rules alternately four times; the signs flip back after an even number of steps.
(6) $\mathbb{H} \otimes \mathbb{H} = \mathbb{R}(4)$ and $\mathbb{R}(2) \otimes \mathbb{R}(2) \otimes \mathbb{R}(4) = \mathbb{R}(16)$.

The last line is **Bott periodicity** for Clifford algebras: $\mathrm{Cl}_{0,n+8} \cong \mathrm{Cl}_{0,n} \otimes \mathbb{R}(16)$, so the *type* of the algebra depends only on n mod 8. Complexifying kills the distinction between $\mathbb{R}$, $\mathbb{C}$ and $\mathbb{H}$ and leaves period two.

n	$\mathrm{Cl}_{0,n}$ ($e^2 = -1$, Bär)	$\mathrm{Cl}_{n,0}$ ($\gamma^2 = +1$)	$\mathrm{Cl}_n \otimes \mathbb{C}$	spinor dim $2^{\lfloor n/2 \rfloor}$
0	$\mathbb{R}$	$\mathbb{R}$	$\mathbb{C}$	1
1	$\mathbb{C}$	$\mathbb{R} \oplus \mathbb{R}$	$\mathbb{C} \oplus \mathbb{C}$	1
2	$\mathbb{H}$	$\mathbb{R}(2)$	$\mathbb{C}(2)$	2
3	$\mathbb{H} \oplus \mathbb{H}$	$\mathbb{C}(2)$	$\mathbb{C}(2) \oplus \mathbb{C}(2)$	2
4	$\mathbb{H}(2)$	$\mathbb{H}(2)$	$\mathbb{C}(4)$	4
5	$\mathbb{C}(4)$	$\mathbb{H}(2) \oplus \mathbb{H}(2)$	$\mathbb{C}(4) \oplus \mathbb{C}(4)$	4
6	$\mathbb{R}(8)$	$\mathbb{H}(4)$	$\mathbb{C}(8)$	8
7	$\mathbb{R}(8) \oplus \mathbb{R}(8)$	$\mathbb{C}(8)$	$\mathbb{C}(8) \oplus \mathbb{C}(8)$	8
8	$\mathbb{R}(16)$	$\mathbb{R}(16)$	$\mathbb{C}(16)$	16

Table T4.3: The classification of Clifford algebras. $\mathbb{R}(m), \mathbb{C}(m), \mathbb{H}(m)$ are $m \times m$ matrices over the reals, complexes and quaternions. The pattern of the real columns repeats with period 8 (with matrix sizes multiplied by 16); the complex column repeats with period 2. A "$\oplus$" appears exactly when the volume element is central with square $+1$ (odd n of the right sign), as in $\mathrm{Cl}_{1,0}$ and $\mathrm{Cl}_{0,3}$.

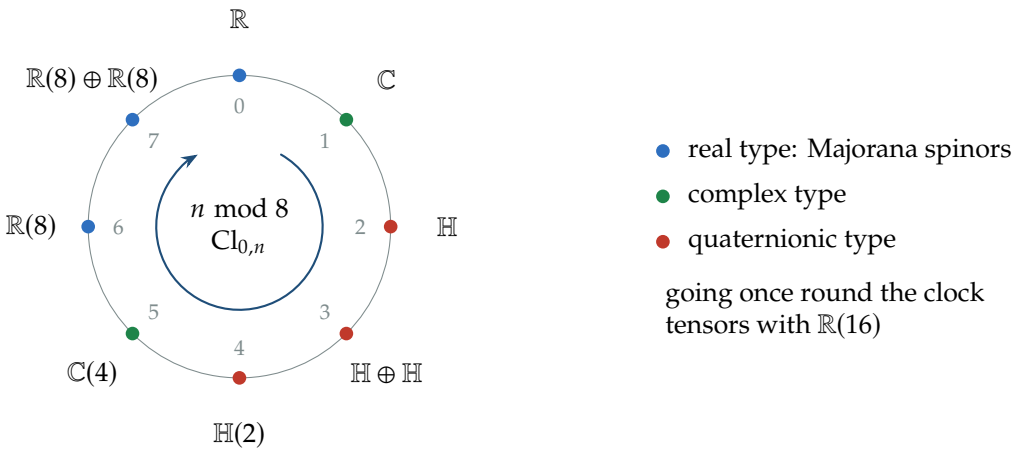


Figure T4.9: The Bott clock for $\mathrm{Cl}_{0,n}$ (vectors square to -1). Reading clockwise: real, complex, quaternionic, quaternionic, quaternionic, complex, real, real, and back to real at $n = 8$. The same eightfold pattern reappears in the real K-theory of spheres, $KO^{-n}(\mathrm{pt}) = \mathbb{Z}, \mathbb{Z}_2, \mathbb{Z}_2, 0, \mathbb{Z}, 0, 0, 0$, and it is why dimensions $8m + 1$ and $8m + 2$ carry mod-2 indices.

Mixed signature and Majorana spinors. One more rule, $\mathrm{Cl}_{p+1,q+1} \cong \mathrm{Cl}_{p,q} \otimes \mathbb{R}(2)$ (same proof, with the extra pair $f_1^2 = +1$, $f_2^2 = -1$, so $(f_1 f_2)^2 = +1$ and no sign flip), handles any signature.

In four-dimensional spacetime:

$$\mathrm{Cl}_{1,3} \cong \mathrm{Cl}_{0,2} \otimes \mathbb{R}(2) = \mathbb{H}(2), \qquad \mathrm{Cl}_{3,1} \cong \mathrm{Cl}_{2,0} \otimes \mathbb{R}(2) = \mathbb{R}(4).$$

With the QFT notes' mostly-minus metric, $\{\gamma^\mu, \gamma^\nu\} = 2\eta^{\mu\nu}$ is $\mathrm{Cl}_{1,3} = \mathbb{H}(2)$: there are no *real* 4×4 gammas. But $\tilde\gamma^\mu := i\gamma^\mu$ satisfy $\{\tilde\gamma^\mu, \tilde\gamma^\nu\} = -2\eta^{\mu\nu}$, which is $\mathrm{Cl}_{3,1} = \mathbb{R}(4)$: *real* 4×4 matrices exist. The next example finds them.

Worked example T4.26: a Majorana representation in four dimensions

We want real 4×4 matrices with $(\tilde\gamma^0)^2 = -1$, $(\tilde\gamma^i)^2 = +1$, all anticommuting. Real symmetric matrices square to positive matrices, and $\epsilon := i\sigma_2 = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$ is real antisymmetric with $\epsilon^2 = -1$. Try

$$\tilde\gamma^0 = \epsilon \otimes \sigma_3, \quad \tilde\gamma^1 = \sigma_1 \otimes \mathbf{1}, \quad \tilde\gamma^2 = \sigma_3 \otimes \mathbf{1}, \quad \tilde\gamma^3 = \epsilon \otimes \epsilon.$$

All are real. Squares: $(\epsilon \otimes \sigma_3)^2 = \epsilon^2 \otimes \sigma_3^2 = -1$; $(\epsilon \otimes \epsilon)^2 = (-1)(-1) = +1$; the middle two square to $+1$. Anticommutation, by counting anticommuting slots (Section T4.4.1): ϵ anticommutes with σ_1 and σ_3, and σ_1 anticommutes with σ_3, so every pair among $\tilde\gamma^0, \tilde\gamma^1, \tilde\gamma^2$ and every pair $(\tilde\gamma^{1,2}, \tilde\gamma^3)$ has exactly one anticommuting slot (the first). For $(\tilde\gamma^0, \tilde\gamma^3)$ the first slots ϵ, ϵ commute and the second slots σ_3, ϵ anticommute: again one. So these are real gammas for $\mathrm{Cl}_{3,1}$, and $\gamma^\mu = -i\tilde\gamma^\mu$ are purely imaginary gammas for the QFT notes' convention. In this basis $i\gamma^\mu\partial_\mu = \tilde\gamma^\mu\partial_\mu$ is a real operator and $\Lambda_{1/2} = \exp\left(-\frac{1}{8}\omega_{\mu\nu}[\tilde\gamma^\mu, \tilde\gamma^\nu]\right)$ is a real matrix, so the condition "ψ is real" is Lorentz invariant and compatible with the massless and massive Dirac equations. That is a **Majorana** spinor: four real components instead of four complex ones.

Dictionary: which dimensions have which spinors

- Weyl (chiral) spinors: even dimension n (Euclidean or Lorentzian).
- Majorana spinors in Lorentzian signature $(1, n-1)$ exist for $n \equiv 2, 3, 4 \pmod 8$ (with the appropriate sign convention; e.g. $n = 4$ above); Majorana–Weyl for $n \equiv 2 \pmod 8$, e.g. $n = 2$ and $n = 10$. This is why the heterotic and type I superstrings live in ten dimensions with sixteen real supercharges (Chapter 9).
- Euclidean signature (Nash's setting) typically has no real structure compatible with chirality in $n = 4$: $\mathrm{Cl}_{0,4} = \mathrm{Cl}_{4,0} = \mathbb{H}(2)$ is quaternionic, and $S^\pm$ are quaternionic ($\cong \mathbb{H}$) rather than real. This is the origin of the "SU(2) = Sp(1)" structures of instanton theory.

Reading Nash (p. 113–114, (4.97), (4.99))

Nash quotes: "S has dimensions $2^{n/2}$ and $2^{(n-1)/2}$ for n even and odd respectively. For odd n the representation S is irreducible but for n even S is reducible into two irreducible components of equal dimension" (4.97). Here we have seen why: S is a fermionic Fock space of $\lfloor n/2 \rfloor$ modes, and in even dimension chirality is fermion parity. His chirality operator $\gamma_5 = i^{n(n+1)/2}\gamma_1 \cdots \gamma_n$ (in his anti-Hermitian gammas) equals $i^{n/2}\gamma_1 \cdots \gamma_n$ in ours, which is our $\Gamma = (-i)^{n/2}\gamma_1 \cdots \gamma_n$ when $n \equiv 0 \pmod 4$ and $-\Gamma$ when $n \equiv 2 \pmod 4$. So in two dimensions Nash's E^+ is our S^-; the index changes sign, which is harmless but worth knowing when comparing with Chapter 4.

Recap

- In dimension $2k$: $\gamma_{2j-1} = \sigma_3^{\otimes(j-1)} \otimes \sigma_1 \otimes \mathbf{1}^{\otimes(k-j)}$, $\gamma_{2j} = \sigma_3^{\otimes(j-1)} \otimes \sigma_2 \otimes \mathbf{1}^{\otimes(k-j)}$; $\Gamma = (-i)^k \gamma_1 \cdots \gamma_{2k} = \sigma_3^{\otimes k}$.
- $b_j = \frac{1}{2}(\gamma_{2j-1} + i\gamma_{2j})$ are fermion annihilators; spinors = Fock space $\Lambda^\bullet \mathbb{C}^k$; $\Gamma = (-1)^F$, $S^\pm$ = even/odd.
- The irreducible representation is unique in even dimensions ($\mathrm{Cl}_{2k} \otimes \mathbb{C} = \mathbb{C}(2^k)$); in odd dimensions add $\gamma_{2k+1} = \pm\Gamma$, giving two inequivalent representations. Dimension $2^{\lfloor n/2 \rfloor}$.
- Real Clifford algebras: explicit shift rules $\mathrm{Cl}_{0,n+2} = \mathrm{Cl}_{n,0} \otimes \mathbb{H}$, $\mathrm{Cl}_{n+2,0} = \mathrm{Cl}_{0,n} \otimes \mathbb{R}(2)$ give period 8 (Bott); Majorana spinors exist when the algebra is real, e.g. $\mathrm{Cl}_{3,1} = \mathbb{R}(4)$.

T4.5 The spin group: rotations built from reflections

We have met the minus sign twice already. In Example T4.19 a rotation by 2π multiplied a Dirac spinor by -1, and in Example T4.10 the unit quaternions turned out to be SU(2), which is "twice as big" as SO(3). This section explains both facts with one construction. The question it answers is:

Which matrices act on spinors when we rotate the coordinate axes?

For vectors the answer is $\mathrm{SO}(n)$. For spinors the answer is a group $\mathrm{Spin}(n)$ that lives *inside the Clifford algebra* and covers $\mathrm{SO}(n)$ twice. The recipe is surprisingly geometric: a unit vector u, viewed as the Clifford element $u = u^a \gamma_a$, acts on vectors by a reflection; two reflections make a rotation; so products of an *even* number of unit vectors are "square roots" of rotations. Everything in this section is a consequence of the one identity $u^2 = |u|^2$.

Throughout, $\gamma_1, \dots, \gamma_n$ are Hermitian with $\{\gamma_a, \gamma_b\} = 2\delta_{ab}$ (our convention, Chapter T4), and we identify a vector $x \in \mathbb{R}^n$ with the matrix

$$\not{x} := x^a \gamma_a, \qquad \not{x}\not{y} + \not{y}\not{x} = 2\langle x, y\rangle \mathbf{1}, \qquad \not{x}^2 = |x|^2 \mathbf{1}. \tag{T4.43}$$

When no confusion is possible we drop the slash and write x for $\not{x}$.

T4.5.1 A unit vector acts by a reflection

Let u be a unit vector, so $u^2 = 1$ and $u^{-1} = u$. Conjugate an arbitrary vector x by u:

$$\begin{aligned} u\, x\, u^{-1} &\overset{(1)}{=} u\, x\, u \\ &\overset{(2)}{=} u\big(-u\, x + 2\langle x, u\rangle\big) \\ &\overset{(3)}{=} -x + 2\langle x, u\rangle\, u \\ &\overset{(4)}{=} -\big(x - 2\langle x, u\rangle u\big) \;=\; -R_u(x). \end{aligned} \tag{T4.44}$$

(1) $u^2 = |u|^2 = 1$, so u is its own inverse.
(2) The Clifford relation (T4.43) in the form $xu = -ux + 2\langle x, u\rangle$.
(3) $u^2 = 1$ again, and the number $\langle x, u\rangle$ commutes with everything.

(4) $R_u(x) \coloneqq x - 2\langle x, u\rangle u$ is the **reflection** in the hyperplane $u^\perp$ perpendicular to u: it flips the component of x along u and leaves the perpendicular part alone.

So conjugation by a unit vector is a reflection followed by the total inversion $x \mapsto -x$. The result is again a vector (a linear combination of the γ_a), which is not obvious in advance: conjugating a vector by a general element of the Clifford algebra produces higher-degree junk.

Worked example T4.27: reflections with Pauli matrices

In three dimensions $\gamma_a = \sigma_a$, and a vector is the Hermitian traceless matrix $X = \vec{x}\cdot\vec{\sigma}$. Take $u = \sigma_3$ (the unit vector along z). Then

$$\sigma_3\,(x^1\sigma_1 + x^2\sigma_2 + x^3\sigma_3)\,\sigma_3 \overset{(1)}{=} -x^1\sigma_1 - x^2\sigma_2 + x^3\sigma_3,$$

(1) $\sigma_3\sigma_1\sigma_3 = -\sigma_1\sigma_3\sigma_3 = -\sigma_1$, likewise for σ_2, and $\sigma_3\sigma_3\sigma_3 = \sigma_3$.

which is $-(x^1, x^2, -x^3)$: minus the reflection in the xy-plane, exactly as (T4.44) says. In three dimensions $-R_u$ is a rotation by π about the axis u, so it has determinant $+1$; but that is an accident of odd dimension. The honest statement is that $x \mapsto uxu^{-1}$ has determinant $(-1)^n \cdot (-1) = (-1)^{n+1}$, and we will only ever use it in pairs, where the signs cancel.

T4.5.2 Two reflections make a rotation

Composing two reflections in hyperplanes that meet at angle α gives a rotation by 2α in the plane spanned by the two normals. This is the kaleidoscope principle (Figure T4.10). In the Clifford algebra it becomes a computation. Let u_1, u_2 be unit vectors. Then

$$(u_2u_1)\,x\,(u_2u_1)^{-1} \overset{(1)}{=} u_2\,(u_1xu_1^{-1})\,u_2^{-1} \overset{(2)}{=} u_2\big(-R_{u_1}x\big)u_2^{-1} \overset{(3)}{=} R_{u_2}R_{u_1}(x). \tag{T4.45}$$

(1) $(u_2u_1)^{-1} = u_1^{-1}u_2^{-1}$.
(2) (T4.44) for u_1.
(3) (T4.44) for u_2, applied to the vector $-R_{u_1}x$; the two minus signs cancel.

So the *even* product u_2u_1 acts on vectors by the rotation $R_{u_2}R_{u_1} \in \mathrm{SO}(n)$.

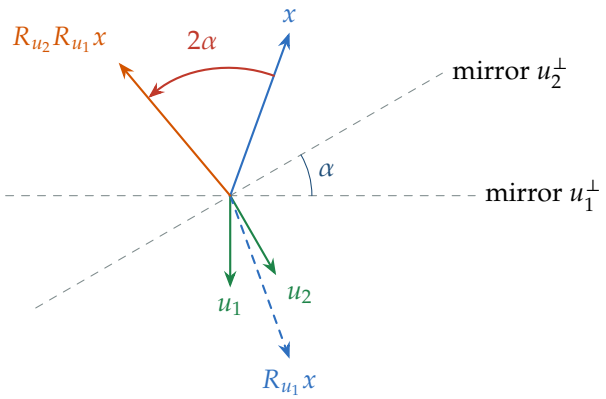


Figure T4.10: Two reflections make a rotation. Reflect x in the mirror $u_1^\perp$ (dashed image), then in the mirror $u_2^\perp$. The result is x rotated by *twice* the angle α between the mirrors. In the Clifford algebra the rotation is implemented by the product u_2u_1, which "knows" only the angle α: this factor of two is the origin of the half-angles in spinor rotations.

Worked example T4.28: the lift of a plane rotation

Take $u_1 = \gamma_1$ and $u_2 = \cos\frac{\theta}{2}\,\gamma_1 + \sin\frac{\theta}{2}\,\gamma_2$, two unit vectors at angle $\theta/2$. Their product is

$$a(\theta) := u_2 u_1 \overset{(1)}{=} \cos\tfrac{\theta}{2}\,\gamma_1\gamma_1 + \sin\tfrac{\theta}{2}\,\gamma_2\gamma_1 \overset{(2)}{=} \cos\tfrac{\theta}{2} - \sin\tfrac{\theta}{2}\,\gamma_1\gamma_2 \overset{(3)}{=} \exp\!\Big(-\frac{\theta}{2}\,\gamma_1\gamma_2\Big). \tag{T4.46}$$

(1) Multiply out.
(2) $\gamma_1^2 = 1$ and $\gamma_2\gamma_1 = -\gamma_1\gamma_2$.
(3) $(\gamma_1\gamma_2)^2 = -\gamma_1^2\gamma_2^2 = -1$, so $\gamma_1\gamma_2$ behaves like the imaginary unit and $e^{-\phi\gamma_1\gamma_2} = \cos\phi - \gamma_1\gamma_2\sin\phi$ (Euler's formula; expand the exponential series and separate even and odd powers).

Let us check directly that it rotates the 12-plane by θ. The element $\gamma_1\gamma_2$ anticommutes with γ_1 and γ_2 and commutes with $\gamma_3, \dots, \gamma_n$. Therefore $a(\theta)\gamma_1 = \gamma_1 a(-\theta)$, and

$$a(\theta)\,\gamma_1\,a(\theta)^{-1} \overset{(1)}{=} \gamma_1\,a(-\theta)\,a(-\theta) \overset{(2)}{=} \gamma_1\,e^{\theta\gamma_1\gamma_2} \overset{(3)}{=} \gamma_1(\cos\theta + \sin\theta\,\gamma_1\gamma_2) = \cos\theta\,\gamma_1 + \sin\theta\,\gamma_2,$$

$$a(\theta)\,\gamma_2\,a(\theta)^{-1} \overset{(4)}{=} \gamma_2\,e^{\theta\gamma_1\gamma_2} = \cos\theta\,\gamma_2 - \sin\theta\,\gamma_1, \qquad a(\theta)\,\gamma_j\,a(\theta)^{-1} = \gamma_j\ (j \geq 3).$$

(1) Move γ_1 to the left through $a(\theta)$; $a(\theta)^{-1} = a(-\theta)$ because the exponents add.
(2) $a(-\theta)^2 = e^{\theta\gamma_1\gamma_2}$.
(3) Euler's formula again, and $\gamma_1\gamma_1\gamma_2 = \gamma_2$.
(4) Same steps, with $\gamma_2\gamma_1\gamma_2 = -\gamma_1$.

So $\rho(a(\theta))$ is the rotation by θ taking e_1 towards e_2. Now the punchline:

$$a(2\pi) = \cos\pi - \sin\pi\,\gamma_1\gamma_2 = -\mathbf{1}, \qquad a(4\pi) = +\mathbf{1}. \tag{T4.47}$$

The rotation by 2π is the identity on vectors, but its lift through the Clifford algebra is $-\mathbf{1}$. This is the same sign as Example T4.19, now derived without any Lorentz group theory.

T4.5.3 The groups $\mathrm{Pin}(n)$ and $\mathrm{Spin}(n)$

Working definition T4.7 (spin group)

Inside the Clifford algebra (concretely: inside the algebra of $2^{\lfloor n/2\rfloor}$ dimensional gamma-matrix products) let

$$\mathrm{Pin}(n) := \{u_1 u_2 \cdots u_m \ :\ u_j \in S^{n-1} \text{ unit vectors},\ m \geq 0\},$$
$$\mathrm{Spin}(n) := \{u_1 u_2 \cdots u_m \ :\ u_j \in S^{n-1},\ m \text{ even}\}.$$

[a unit vector u is used as the matrix $u^a\gamma_a$, which squares to $\mathbf{1}$]. The **spin group** acts on vectors by

$$\rho(a)x := a\,x\,a^{-1}, \qquad a \in \mathrm{Spin}(n), \tag{T4.48}$$

and on spinors simply by matrix multiplication, $\psi \mapsto a\psi$. *To test whether a matrix a is in* $\mathrm{Spin}(n)$*:* it must be even [a sum of products of an even number of γ's], it must satisfy $a\,x\,a^{-1} \in \mathbb{R}^n$ for all vectors x (i.e. a combination of γ's), and it must have "unit norm", $a\,a^{\dagger} = 1$, where † reverses the order of the factors. For Hermitian gammas and real u_j this says a is unitary.

Bär's definitions [Bär Def. 2.2.3, 2.2.5] are identical, with his convention $v^2 = -|v|^2$; since u and $-u$ are both unit vectors, the sign convention changes nothing in the group. Three facts

make Spin(n) useful.

1. It is a group. The product of two even products of unit vectors is again one. The inverse of $u_1 \cdots u_m$ is $u_m \cdots u_1$, because the factors cancel from the middle out ($u_m u_m = 1$, then $u_{m-1}u_{m-1} = 1, \dots$). With Hermitian γ's every unit vector $u = u^a \gamma_a$ is Hermitian and unitary, so $(u_1 \cdots u_m)^\dagger = u_m \cdots u_1 = (u_1 \cdots u_m)^{-1}$: *the spinor representation of* Spin(n) *is unitary*.

2. ρ lands in SO(n) and hits all of it. By (T4.45), $\rho(u_1 \cdots u_{2k})$ is a product of $2k$ reflections, hence orthogonal with determinant $(-1)^{2k} = 1$. Conversely, the theorem of Cartan and Dieudonné says every element of O(n) is a product of at most n reflections; an element of SO(n) needs an even number. So every rotation is $\rho(a)$ for some a [Bär Prop. 2.2.10]. For a physicist the proof in the case that matters is Example T4.28: any rotation is a product of rotations in commuting planes (the normal form of an orthogonal matrix), and each plane rotation lifts as $a(\theta)$.

3. The kernel is $\{\pm 1\}$. If $\rho(a) = \mathrm{id}$ then a commutes with every γ_a, hence with the whole algebra; an even element with this property is a multiple of $\mathbf{1}$ (this is the "exercise" in Bär's proof: an even monomial $\gamma_I \neq 1$ contains some γ_a and anticommutes with a suitable γ_b). So $a = \alpha\mathbf{1}$. Reversing the order of the factors does nothing to a multiple of $\mathbf{1}$, so $a^\dagger = \alpha\mathbf{1}$ too, while $a\,a^\dagger = u_1 \cdots u_m\, u_m \cdots u_1 = 1$. Hence $\alpha^2 = 1$ and $a = \pm\mathbf{1}$. Both are in Spin(n): $-1 = \gamma_1(-\gamma_1)$.

Key idea: the spin group is a double cover

The sequence

$$1 \longrightarrow \{\pm 1\} \longrightarrow \mathrm{Spin}(n) \xrightarrow{\rho} \mathrm{SO}(n) \longrightarrow 1 \tag{T4.49}$$

is exact [Bär Prop. 2.2.10]: every rotation R has exactly two lifts $\pm a$, and they act on spinors with opposite signs. A spinor "knows" not just the rotation R but also which of the two lifts was used.

Watch out: spinors are not a representation of SO(n)

If a representation of Spin(n) came from a representation λ' of SO(n), then $-1 \in \mathrm{Spin}(n)$ would act as $\lambda'(\rho(-1)) = \lambda'(\mathrm{id}) = \mathrm{id}$ [Bär Rem. 2.3.4]. On spinors -1 acts as $-\mathrm{id}$. So the spinor representation is *not* a representation of the rotation group, only of its double cover. Tensors (vectors, forms, metrics) are the representations on which -1 acts trivially; spinors are the ones on which it acts as -1. Physicists say "spinors are double valued representations of SO(n)".

T4.5.4 Small cases, completely explicitly

Worked example T4.29: Spin(1) = $\mathbb{Z}_2$ and Spin(2) = U(1)

$n = 1$. The unit vectors are $\pm\gamma_1$, so even products are ± 1: Spin(1) = $\{\pm 1\} = \mathbb{Z}_2$, covering the trivial group SO(1) [Bär Ex. 2.2.7]. This tiny group is what distinguishes the two spin structures on a circle (Section T4.6).
$n = 2$. Every unit vector is $u(\phi) = \cos\phi\,\gamma_1 + \sin\phi\,\gamma_2$, and

$$u(\phi_2)\,u(\phi_1) \overset{(1)}{=} \cos\phi_2\cos\phi_1 + \sin\phi_2\sin\phi_1 + (\cos\phi_2\sin\phi_1 - \sin\phi_2\cos\phi_1)\gamma_1\gamma_2$$
$$\overset{(2)}{=} \cos(\phi_2 - \phi_1) - \sin(\phi_2 - \phi_1)\gamma_1\gamma_2 \overset{(3)}{=} a\big(2(\phi_2 - \phi_1)\big).$$

(1) Expand; use $\gamma_1^2 = \gamma_2^2 = 1$ and $\gamma_2\gamma_1 = -\gamma_1\gamma_2$.
(2) Angle-difference formulas.
(3) Compare with (T4.46).

Products of such pairs multiply like $a(\theta)a(\theta') = a(\theta+\theta')$, so $\mathrm{Spin}(2) = \{a(\theta)\}$ is a circle, parametrised by $\theta \in [0, 4\pi)$ [Bär Ex. 2.2.8]. In the representation $\gamma_1 = \sigma_1$, $\gamma_2 = \sigma_2$ of Example T4.1, $\gamma_1\gamma_2 = i\sigma_3$ and

$$a(\theta) = e^{-i\theta\sigma_3/2} = \begin{pmatrix} e^{-i\theta/2} & 0 \\ 0 & e^{i\theta/2} \end{pmatrix}. \tag{T4.50}$$

The two chiral components of a 2D spinor carry "spin" $\mp\frac{1}{2}$: they pick up the phases $e^{\mp i\theta/2}$ under a rotation by θ, while a vector $v^1 + iv^2$ picks up $e^{i\theta}$. The lower entry $w = e^{i\theta/2}$ of $a(\theta)$ squares to the rotation $z = e^{i\theta}$, so ρ is the squaring map $\mathrm{U}(1) \to \mathrm{U}(1)$ (Bär writes $z \mapsto z^2$).

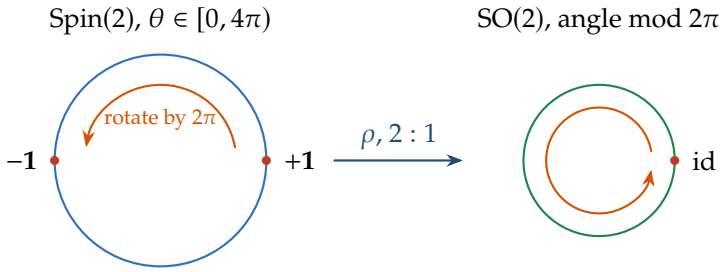


Figure T4.11: $\mathrm{Spin}(2) \to \mathrm{SO}(2)$ is the double cover of a circle by a circle. Going once around SO(2) (a full 2π rotation, orange loop on the right) lifts to going *half* way around Spin(2), from $+\mathbf{1}$ to $-\mathbf{1}$. Only a 4π rotation closes the loop upstairs.

Worked example T4.30: Spin(3) = SU(2)

Now $\gamma_a = \sigma_a$ and a unit vector is $\hat{u}\cdot\vec{\sigma}$. By the Pauli product rule (T4.6),

$$(\hat{u}\cdot\vec{\sigma})(\hat{w}\cdot\vec{\sigma}) = (\hat{u}\cdot\hat{w})\mathbf{1} + i(\hat{u}\times\hat{w})\cdot\vec{\sigma}, \tag{T4.51}$$

and $(\hat{u}\cdot\hat{w})^2 + |\hat{u}\times\hat{w}|^2 = 1$. So a product of two unit vectors has the form $a_0\mathbf{1} + i\vec{a}\cdot\vec{\sigma}$ with $a_0^2 + |\vec{a}|^2 = 1$. These matrices are exactly

$$\begin{pmatrix} a_0 + ia_3 & ia_1 + a_2 \\ ia_1 - a_2 & a_0 - ia_3 \end{pmatrix} = \begin{pmatrix} \alpha & -\bar{\beta} \\ \beta & \bar{\alpha} \end{pmatrix}, \qquad |\alpha|^2 + |\beta|^2 = 1,$$

the group SU(2), topologically the 3-sphere S^3 (the unit quaternions, (T4.14)). Every element can be written

$$U(\theta, \hat{n}) = \exp\!\left(-\frac{i\theta}{2}\,\hat{n}\cdot\vec{\sigma}\right) = \cos\frac{\theta}{2}\,\mathbf{1} - i\sin\frac{\theta}{2}\,\hat{n}\cdot\vec{\sigma}, \tag{T4.52}$$

and $\rho(U)$ is the rotation by θ about $\hat{n}$. Let us verify this for $\hat{n} = \hat{z}$ by explicit multiplication. With $X = \vec{x}\cdot\vec{\sigma}$ and $U = \mathrm{diag}(e^{-i\theta/2}, e^{i\theta/2})$:

$$UXU^\dagger \overset{(1)}{=} \begin{pmatrix} e^{-i\theta/2} & 0 \\ 0 & e^{i\theta/2} \end{pmatrix} \begin{pmatrix} x^3 & x^1 - ix^2 \\ x^1 + ix^2 & -x^3 \end{pmatrix} \begin{pmatrix} e^{i\theta/2} & 0 \\ 0 & e^{-i\theta/2} \end{pmatrix}$$
$$\overset{(2)}{=} \begin{pmatrix} x^3 & e^{-i\theta}(x^1 - ix^2) \\ e^{i\theta}(x^1 + ix^2) & -x^3 \end{pmatrix}.$$

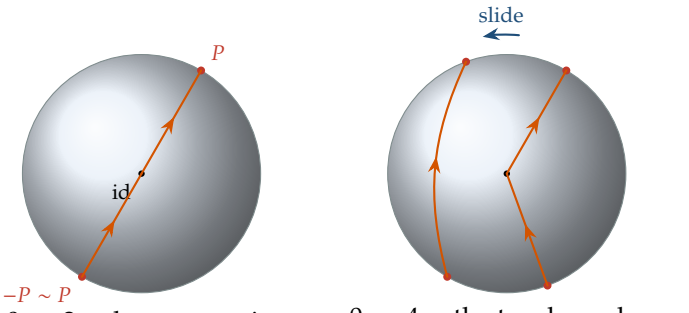
(1) Write out X and $U^\dagger = U^{-1}$.
(2) Multiplying by diagonal matrices on both sides scales entry (j,k) by $U_{jj}\bar{U}_{kk}$: the diagonal is unchanged and the off-diagonal entries acquire $e^{\mp i\theta}$.

So $x^1 + ix^2 \mapsto e^{i\theta}(x^1 + ix^2)$ and $x^3 \mapsto x^3$: a rotation by θ about z. The rotation matrix in general is read off as $R_{ab} = \frac{1}{2}\operatorname{tr}(\sigma_a U \sigma_b U^\dagger)$. The kernel is visible too: $U(2\pi, \hat{n}) = -\mathbf{1}$ for every axis, and $\pm U$ give the same R.

The topology of SO(3)**.** Since $\mathrm{SU}(2) = S^3$ and ρ identifies U with $-U$, the rotation group is the 3-sphere with antipodal points identified: $\mathrm{SO}(3) \cong \mathbb{RP}^3$. A more visual model [Bär proof of Prop. 2.2.13]: represent the rotation by angle $\theta \in [0, \pi]$ about the axis $\hat{n}$ by the point $\theta\hat{n}/\pi$ of the closed unit ball. Rotations by π about $\hat{n}$ and $-\hat{n}$ coincide, so antipodal points of the boundary sphere are identified. A path that starts at the centre, runs out to the boundary, reappears at the antipodal point and returns to the centre is a closed loop (the family of rotations by $0 \to 2\pi$ about a fixed axis). It cannot be shrunk to a point: shrinking would require pulling the two ends of the chord off the boundary together, but they are stuck at antipodal points. Traversing it twice gives a loop that *can* be shrunk (Figure T4.12). Hence

$$\pi_1(\mathrm{SO}(3)) = \mathbb{Z}_2, \qquad \pi_1(\mathrm{SU}(2)) = \pi_1(S^3) = 0.$$

The 2π loop is exactly the loop whose lift to SU(2) ends at $-\mathbf{1}$ instead of $+\mathbf{1}$: a lift of a loop in SO(3) closes in SU(2) precisely when the loop is contractible.



$0 \to 2\pi$ about one axis: not contractible

$0 \to 4\pi$: the two boundary pairs slide together and cancel

Figure T4.12: SO(3) as a ball of radius π with antipodal boundary points identified. Left: the 2π rotation loop leaves through P and re-enters at $-P$; it cannot be contracted. Right: the 4π loop crosses the boundary twice; moving the second crossing point around the sphere until it meets the first lets the two crossings annihilate, and the loop shrinks to the identity. This is the geometry behind Dirac's belt trick: a belt twisted by 4π can be untwisted without rotating its ends, one twisted by 2π cannot.

Physical picture: the belt trick

Hold one end of a belt fixed and rotate the buckle by 2π. The belt is twisted and you cannot untwist it without rotating the buckle back. Rotate by 4π instead: now, keeping the buckle's orientation fixed, you can loop the belt around the buckle and the twist disappears. The state of the belt records an element of SU(2), not just of SO(3). Neutron interferometry measured the -1 directly in 1975: a neutron beam rotated by 2π in a magnetic field interferes destructively with an unrotated one.

Connectedness in general. For $n \geq 2$, $\mathrm{Spin}(n)$ is connected: the path $c(t) = (\cos t\, \gamma_1 + \sin t\, \gamma_2)\gamma_1 = \cos t - \sin t\, \gamma_1\gamma_2$ joins $c(0) = 1$ to $c(\pi) = -1$, and every other element is joined to

one of them by lifting a path in the connected group $\mathrm{SO}(n)$. For $n \geq 3$ it is simply connected: $\pi_1(\mathrm{SO}(n)) = \mathbb{Z}_2$ for all $n \geq 3$ (adding dimensions does not change π_1, since $\mathrm{SO}(n+1)/\mathrm{SO}(n) = S^n$ is simply connected), and the nontrivial loop is the 2π rotation in one plane, whose lift is the *open* path from 1 to -1. So $\mathrm{Spin}(n)$ is the **universal cover** of $\mathrm{SO}(n)$ for $n \geq 3$ [Bär Prop. 2.2.13]; this is how Nash introduces it.

T4.5.5 The Lie algebra: $\gamma_a\gamma_b/4$

Infinitesimal rotations are antisymmetric matrices $A \in \mathfrak{so}(n)$, $A_{ab} = -A_{ba}$, acting as $x \mapsto Ax$. Which infinitesimal element of the Clifford algebra implements them? Differentiating (T4.46) at $\theta = 0$ gives $-\frac{1}{2}\gamma_1\gamma_2$ for the rotation generator of the 12-plane. So we expect bivectors. The key commutator is

$$\begin{aligned}[\gamma_a\gamma_b, \gamma_c] &\overset{(1)}{=} \gamma_a\gamma_b\gamma_c - \gamma_c\gamma_a\gamma_b \overset{(2)}{=} \gamma_a(2\delta_{bc} - \gamma_c\gamma_b) - \gamma_c\gamma_a\gamma_b \\ &\overset{(3)}{=} 2\delta_{bc}\gamma_a - \{\gamma_a, \gamma_c\}\gamma_b \overset{(4)}{=} 2\delta_{bc}\gamma_a - 2\delta_{ac}\gamma_b. \end{aligned} \tag{T4.53}$$

(1) Definition of the commutator.
(2) $\gamma_b\gamma_c = 2\delta_{bc} - \gamma_c\gamma_b$.
(3) Collect the two terms ending in γ_b.
(4) Clifford relation.

Now contract with $\frac{1}{4}A_{ab}$ and apply to a vector $x = x^c\gamma_c$:

$$\begin{aligned}\left[\tfrac{1}{4}A_{ab}\gamma_a\gamma_b,\ x^c\gamma_c\right] &\overset{(1)}{=} \tfrac{1}{4}A_{ab}x^c(2\delta_{bc}\gamma_a - 2\delta_{ac}\gamma_b) \overset{(2)}{=} \tfrac{1}{2}(A_{ab}x^b\gamma_a - A_{ab}x^a\gamma_b) \\ &\overset{(3)}{=} \tfrac{1}{2}(A_{ab}x^b\gamma_a + A_{ba}x^a\gamma_b) \overset{(4)}{=} (Ax)^a\gamma_a. \end{aligned} \tag{T4.54}$$

(1) (T4.53).
(2) Do the c sums with the deltas.
(3) A is antisymmetric.
(4) Rename $a \leftrightarrow b$ in the second term; the two terms are equal.

Since the derivative of $x \mapsto axa^{-1}$ at $a = 1$ in the direction X is $x \mapsto [X, x]$, this says:

Key idea: spin generators

The Lie algebra $\mathfrak{spin}(n)$ is spanned by the $\binom{n}{2}$ bivectors $\gamma_a\gamma_b$ $(a < b)$, and the infinitesimal rotation $A \in \mathfrak{so}(n)$ lifts to

$$\rho_*^{-1}(A) = \frac{1}{4}A_{ab}\gamma_a\gamma_b = \frac{1}{2}A_{ab}\Sigma_{ab}, \qquad \Sigma_{ab} := \frac{1}{4}[\gamma_a, \gamma_b]. \tag{T4.55}$$

A finite rotation e^A acts on spinors by $\exp\left(\frac{1}{4}A_{ab}\gamma_a\gamma_b\right)$.

The Σ_{ab} satisfy the commutation relations of $\mathfrak{so}(n)$. From (T4.53), $[\Sigma_{ab}, \gamma_c] = \delta_{bc}\gamma_a - \delta_{ac}\gamma_b$ (for $a \neq b$, $\Sigma_{ab} = \frac{1}{2}\gamma_a\gamma_b$), and then for $c \neq d$

$$\begin{aligned}[\Sigma_{ab}, \Sigma_{cd}] &\overset{(1)}{=} \tfrac{1}{2}[\Sigma_{ab}, \gamma_c]\gamma_d + \tfrac{1}{2}\gamma_c[\Sigma_{ab}, \gamma_d] \\ &\overset{(2)}{=} \delta_{bc}\Sigma_{ad} - \delta_{ac}\Sigma_{bd} + \delta_{bd}\Sigma_{ca} - \delta_{ad}\Sigma_{cb}, \end{aligned}$$

(1) Leibniz rule for commutators, $[X, YZ] = [X, Y]Z + Y[X, Z]$.

(2) Insert $[\Sigma_{ab}, \gamma_c]$ and use $\frac{1}{2}\gamma_x\gamma_y = \Sigma_{xy}$ for $x \neq y$ (when an index coincidence produces $\frac{1}{2}\gamma_x\gamma_x = \frac{1}{2}$, the two such constants appear with opposite signs and cancel).

which is the familiar $\mathfrak{so}(n)$ algebra of the generators $L_{ab} = E_{ab} - E_{ba}$. Bär computes the same thing as $\rho_*(e_ie_j) = 2(E_{ji} - E_{ij})$ [Bär p. 82]; with $e_a = -i\gamma_a$, $e_ie_j = -\gamma_i\gamma_j$, and our (T4.54) gives $\rho_*(-\gamma_i\gamma_j) = -4(E_{ij} - E_{ji})$, in agreement.

Dictionary: Euclidean Σ_{ab} versus Lorentz $S^{\mu\nu}$

- Mathematicians use the anti-Hermitian $\Sigma_{ab} = \frac{1}{4}[\gamma_a, \gamma_b]$; group elements are $\exp(\frac{1}{2}\omega_{ab}\Sigma_{ab})$ with real ω.
- Physicists use Hermitian generators $J_{ab} = i\Sigma_{ab}$ and write $\exp(-\frac{i}{2}\omega_{ab}J_{ab})$. In Minkowski space the QFT notes' $S^{\mu\nu} = \frac{i}{4}[\gamma^\mu, \gamma^\nu]$ (T4.29) is this J.
- In 3D: $\Sigma_{ab} = \frac{1}{2}\sigma_a\sigma_b = \frac{i}{2}\varepsilon_{abc}\sigma_c$ for $a \neq b$, so $J_{12} = i\Sigma_{12} = -\frac{1}{2}\sigma_3$, and cyclically: the Hermitian spin operators are $\pm\frac{1}{2}\vec{\sigma}$, the overall sign being a convention. "Spin $\frac{1}{2}$" is the statement that the rotation generators are *half* the Pauli matrices, which is the infinitesimal version of the half-angle in $a(\theta)$.

T4.5.6 $\mathrm{Spin}(4) = \mathrm{SU}(2) \times \mathrm{SU}(2)$ and self-duality

Four dimensions is special: the rotation group splits into two commuting pieces. The chiral representation (T4.21) shows this at once. Write a vector $x \in \mathbb{R}^4$ as the 2×2 matrix

$$X := x^a\tau_a = x^4\mathbf{1} + i\,\vec{x}\cdot\vec{\sigma} = \begin{pmatrix} x^4 + ix^3 & ix^1 + x^2 \\ ix^1 - x^2 & x^4 - ix^3 \end{pmatrix}, \qquad \det X = |x|^2, \tag{T4.56}$$

a quaternion (Example T4.10). Then $\not{x} = x^a\gamma_a = \left(\begin{smallmatrix} 0 & X \\ X^\dagger & 0 \end{smallmatrix}\right)$. Since even products of γ's are block diagonal, a spin element has the form $a = \mathrm{diag}(p, q)$, and

$$a\not{x}a^{-1} \overset{(1)}{=} \begin{pmatrix} p & 0 \\ 0 & q \end{pmatrix}\begin{pmatrix} 0 & X \\ X^\dagger & 0 \end{pmatrix}\begin{pmatrix} p^\dagger & 0 \\ 0 & q^\dagger \end{pmatrix} \overset{(2)}{=} \begin{pmatrix} 0 & pXq^\dagger \\ qX^\dagger p^\dagger & 0 \end{pmatrix}. \tag{T4.57}$$

(1) a is unitary (fact 1 above), so $a^{-1} = a^\dagger$.
(2) Block multiplication.

The result is again of the form $\not{x}'$, with $X' = pXq^\dagger$, because $(pXq^\dagger)^\dagger = qX^\dagger p^\dagger$. For this to be a vector for every X, $pXq^\dagger$ must again be a real combination of $\mathbf{1}, i\sigma_j$, which holds for $p, q \in \mathrm{SU}(2)$; and $\det(pXq^\dagger) = \det X$, so the map is a rotation.

Worked example T4.31: $\mathrm{Spin}(4) \to \mathrm{SO}(4)$ via quaternions

The map $(p, q) \mapsto \big(X \mapsto pXq^\dagger\big)$ sends $\mathrm{SU}(2) \times \mathrm{SU}(2)$ onto $\mathrm{SO}(4)$. Its kernel: if $pXq^\dagger = X$ for all X, then $X = \mathbf{1}$ gives $p = q$, and then p commutes with every quaternion, so $p = \pm\mathbf{1}$. The kernel is $\{(\mathbf{1}, \mathbf{1}), (-\mathbf{1}, -\mathbf{1})\}$, the $\{\pm 1\}$ of (T4.49). So

$$\mathrm{Spin}(4) = \mathrm{SU}(2)_+ \times \mathrm{SU}(2)_-, \qquad \mathrm{SO}(4) = \big(\mathrm{SU}(2) \times \mathrm{SU}(2)\big)/\mathbb{Z}_2.$$

The positive-chirality spinors S^+ (upper block) transform under p only, the negative ones S^- under q only. In particle-physics language: a Euclidean Weyl spinor is a doublet of one SU(2) and a singlet of the other, the Euclidean shadow of the $(\frac{1}{2}, 0)$ and $(0, \frac{1}{2})$ representations of the Lorentz group.

Self-dual and anti-self-dual generators. Which rotations does S^+ see? Compute the blocks of $\Sigma_{ab} = \frac{1}{2}\gamma_a\gamma_b$ ($a \neq b$). The upper block is $\frac{1}{2}\tau_a\tau_b^\dagger$ and the lower block is $\frac{1}{2}\tau_a^\dagger\tau_b$ (Example T4.14). For the two "opposite" planes 12 and 34:

$$\Sigma_{12} \overset{(1)}{=} \tfrac{1}{2}\,\mathrm{diag}\big((i\sigma_1)(-i\sigma_2),\ (-i\sigma_1)(i\sigma_2)\big) \overset{(2)}{=} \tfrac{i}{2}\,\mathrm{diag}(\sigma_3,\ \sigma_3),$$
$$\Sigma_{34} \overset{(3)}{=} \tfrac{1}{2}\,\mathrm{diag}\big((i\sigma_3)\mathbf{1},\ (-i\sigma_3)\mathbf{1}\big) = \tfrac{i}{2}\,\mathrm{diag}(\sigma_3,\ -\sigma_3).$$

(1) Insert $\tau_1 = i\sigma_1$, $\tau_2 = i\sigma_2$ from (T4.20).
(2) $\sigma_1\sigma_2 = i\sigma_3$.
(3) $\tau_3 = i\sigma_3$, $\tau_4 = \mathbf{1}$.

So on S^+, $\Sigma_{12} = \Sigma_{34}$, while on S^-, $\Sigma_{12} = -\Sigma_{34}$. The same holds for the pairs $(13, 42)$ and $(14, 23)$. The combinations $\Sigma_{12} - \Sigma_{34}$ etc. vanish on S^+. Recall that the Hodge star in $\mathbb{R}^4$ exchanges $\mathrm{d}x^1 \wedge \mathrm{d}x^2 \leftrightarrow \mathrm{d}x^3 \wedge \mathrm{d}x^4$. A two-form $F = \frac{1}{2}F_{ab}\mathrm{d}x^a \wedge \mathrm{d}x^b$ acts on spinors through $\frac{1}{2}F_{ab}\Sigma_{ab}$, and the computation shows

$$\tfrac{1}{2}F_{ab}\Sigma_{ab}\big|_{S^+} \text{ depends only on } F^+ = \tfrac{1}{2}(F + \star F), \qquad \tfrac{1}{2}F_{ab}\Sigma_{ab}\big|_{S^-} \text{ only on } F^- = \tfrac{1}{2}(F - \star F). \quad \text{(T4.58)}$$

This is the algebra behind the splitting $\Lambda^2\mathbb{R}^4 = \Lambda^2_+ \oplus \Lambda^2_-$, $\mathfrak{so}(4) = \mathfrak{su}(2)_+ \oplus \mathfrak{su}(2)_-$, and it is why instantons (self-dual fields, Chapter 8) couple to one chirality in a special way.

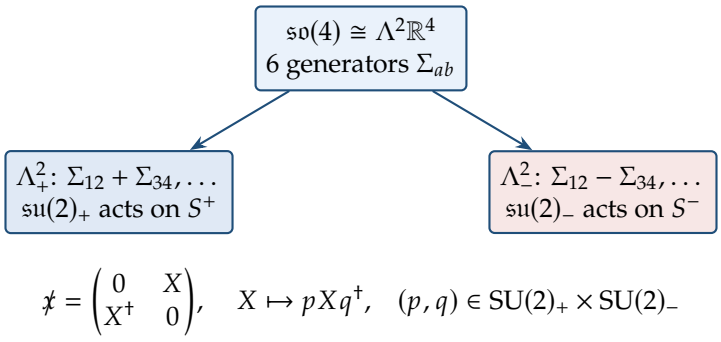


Figure T4.13: The four-dimensional miracle. The six rotation generators split into three self-dual and three anti-self-dual combinations that commute with each other. The first three act only on positive-chirality spinors, the last three only on negative ones. At the group level, a vector is a quaternion X and a rotation is $X \mapsto pXq^\dagger$.

T4.5.7 Lorentzian signature

The same construction works for any signature [Bär §6.1]. A vector v with $\langle v, v\rangle = \pm 1$ is invertible in the Clifford algebra, $v\,x\,v^{-1}$ is (minus) the reflection in $v^\perp$, and even products of such v form $\mathrm{Spin}(p, q)$, a double cover of $\mathrm{SO}(p, q)$. Two new features appear.

Worked example T4.32: $\mathrm{Spin}(1, 1)$: boosts in two dimensions

Take a timelike γ^0 with $(\gamma^0)^2 = 1$ and a spacelike γ^1 with $(\gamma^1)^2 = -1$, e.g. $\gamma^0 = \sigma_1$, $\gamma^1 = i\sigma_2$. The bivector $\gamma^0\gamma^1 = \sigma_1\, i\sigma_2 = -\sigma_3$ now squares to $+1$, so exponentials give hyperbolic functions:

$$b(\beta) := \exp\Big(\frac{\beta}{2}\gamma^0\gamma^1\Big) = \cosh\frac{\beta}{2} + \sinh\frac{\beta}{2}\,\gamma^0\gamma^1 = \mathrm{diag}\big(e^{-\beta/2}, e^{\beta/2}\big).$$

Since $\gamma^0\gamma^1$ anticommutes with γ^0, $b(\beta)\gamma^0 = \gamma^0 b(-\beta)$, and

$$b\gamma^0 b^{-1} \overset{(1)}{=} \gamma^0\, b(-\beta)^2 \overset{(2)}{=} \gamma^0\big(\cosh\beta - \sinh\beta\,\gamma^0\gamma^1\big) \overset{(3)}{=} \cosh\beta\,\gamma^0 - \sinh\beta\,\gamma^1,$$

(1) Move γ^0 through $b(\beta)$; $b(\beta)^{-1} = b(-\beta)$.
(2) $b(-\beta)^2 = \exp(-\beta\gamma^0\gamma^1)$, and $(\gamma^0\gamma^1)^2 = +1$ gives cosh and sinh instead of cos and sin.
(3) $\gamma^0\gamma^0\gamma^1 = \gamma^1$.

a boost with rapidity β (the sign depends on which way we call "forward"). The two components of the spinor are rescaled by $e^{\mp\beta/2}$: they are the left- and right-movers, and boosts, unlike rotations, are not unitary. SO(1, 1) has two components and Spin(1, 1) has four [Bär Ex. 6.1.11], because the non-compact direction does not close up.

Spin(1, 3) = SL(2, $\mathbb{C}$). In four-dimensional Minkowski space the identity component $\mathrm{Spin}_0(1,3)$ is $\mathrm{SL}(2,\mathbb{C})$ acting by $X \mapsto AXA^\dagger$ on Hermitian matrices $X = x^0\mathbf{1} + \vec{x}\cdot\vec{\sigma}$ (Section T4.3.7). The proof that this is a double cover of the proper orthochronous Lorentz group mirrors Example T4.30: $\det(AXA^\dagger) = \det X$ shows it is a Lorentz transformation; $AXA^\dagger = X$ for all X forces $A = \pm\mathbf{1}$; and the Lie algebra $\mathfrak{sl}(2,\mathbb{C})$ is spanned by $\frac{i}{2}\sigma_j$ (rotations) and $\frac{1}{2}\sigma_j$ (boosts), matching the block-diagonal $S^{\mu\nu}$ of Example T4.18. The Weyl spinors are the defining representation A and its conjugate-inverse $(A^\dagger)^{-1}$. As in Euclidean signature, $\mathrm{Spin}_0(1,3)$ is simply connected while $\mathrm{SO}_0(1,3)$ has $\pi_1 = \mathbb{Z}_2$ [Bär Prop. 6.1.17]; but $\mathrm{SL}(2,\mathbb{C})$ is not compact, so these spinor representations are not unitary.

Try it: the rotation matrix from SU(2)

Compute $R_{ab} = \frac{1}{2}\operatorname{tr}(\sigma_a U \sigma_b U^\dagger)$ for $U = e^{-i\theta\sigma_1/2}$ and check that it is the rotation by θ about the x-axis. *Answer:* $U\sigma_1U^\dagger = \sigma_1$, $U\sigma_2U^\dagger = \cos\theta\,\sigma_2 + \sin\theta\,\sigma_3$, $U\sigma_3U^\dagger = \cos\theta\,\sigma_3 - \sin\theta\,\sigma_2$, so R has columns $(1,0,0)$, $(0,\cos\theta,\sin\theta)$, $(0,-\sin\theta,\cos\theta)$.

Reading Nash (p. 113)

Nash's whole treatment of the spin group is one sentence: "Spinors first arise when one encounters the universal cover Spin(n) of the orthogonal group SO(n)." We have now seen what that cover is concretely: even products of unit vectors $u = u^a\gamma_a$, acting on vectors by $x \mapsto axa^{-1}$ and on spinors by matrix multiplication; it is universal (simply connected) for $n \geq 3$, and the two lifts $\pm a$ of each rotation differ in their action on spinors. The "spin representation" Nash quotes on the same page is the restriction of the Clifford representation of Section T4.4 to Spin(n); since $\mathrm{Spin}(n) \subset \mathrm{Cl}^0$ and even elements commute with Γ, it splits as $S^+ \oplus S^-$ in even dimensions (his (4.97)).

Recap

- A unit vector u acts by $x \mapsto uxu^{-1} = -R_u(x)$; two of them give a rotation by twice the angle between them.
- $\mathrm{Spin}(n) = \{\text{even products of unit vectors}\}$, $\rho(a)x = axa^{-1}$; $1 \to \mathbb{Z}_2 \to \mathrm{Spin}(n) \to \mathrm{SO}(n) \to 1$; spinors are acted on by a itself.
- Plane rotation lift $a(\theta) = \exp(-\frac{\theta}{2}\gamma_1\gamma_2)$, and $a(2\pi) = -1$. Spin(2) = U(1) (squaring map), Spin(3) = SU(2), Spin(4) = SU(2) × SU(2), $\mathrm{Spin}_0(1,3) = \mathrm{SL}(2,\mathbb{C})$.
- $\mathfrak{spin}(n)$ is spanned by $\Sigma_{ab} = \frac{1}{4}[\gamma_a,\gamma_b]$; $A \in \mathfrak{so}(n)$ lifts to $\frac{1}{4}A_{ab}\gamma_a\gamma_b$ (T4.55).

- In 4D, S^+ sees only self-dual rotations/field strengths, S^- only anti-self-dual ones.

T4.6 Spin structures: putting spinors on a manifold

On $\mathbb{R}^n$ a spinor field is simply a function $\psi : \mathbb{R}^n \to \mathbb{C}^{2^{\lfloor n/2 \rfloor}}$. On a curved manifold we must say what the components of ψ *mean*. The components of a vector field are meaningful only relative to a frame; change the frame by a rotation R and the components change by R. The components of a spinor are also meaningful only relative to a frame, but a change of frame acts on them by a *lift* $\pm a$ of R to $\mathrm{Spin}(n)$, and there are two lifts. The question of this section is:

Can we choose the signs of the lifts consistently over the whole manifold?

When the answer is yes, and we have made a choice, we have a **spin structure**. The answer is sometimes "yes, in several inequivalent ways" (the circle has two, the torus four) and sometimes "no" ($\mathbb{CP}^2$). Both answers have physical consequences. On a circle, the two spin structures are periodic and antiperiodic fermions: the Ramond and Neveu–Schwarz sectors of string theory, or the $(-1)^F$-twisted and untwisted thermal traces. On $\mathbb{CP}^2$, the impossibility means there is no Dirac equation at all unless we couple the fermion to an extra gauge field.

T4.6.1 Frames, transition functions and associated bundles

We first recall how vectors and forms are glued together on a manifold, written so that the spinor case becomes a small modification.

Let M be an oriented Riemannian manifold of dimension n. Cover M by open sets U_α on each of which we have chosen an oriented orthonormal frame $h_\alpha = (e_1^{(\alpha)}, \dots, e_n^{(\alpha)})$ (for instance by applying Gram–Schmidt to the coordinate vectors ∂_μ). On an overlap $U_\alpha \cap U_\beta$ two frames are available, and since both are oriented and orthonormal they differ by a rotation:

$$e_b^{(\beta)} = e_a^{(\alpha)} \left(g_{\alpha\beta}\right)_{ab}, \qquad g_{\alpha\beta} : U_\alpha \cap U_\beta \to \mathrm{SO}(n). \tag{T4.59}$$

We write this compactly as $h_\beta = h_\alpha\, g_{\alpha\beta}$. The functions $g_{\alpha\beta}$ are the **transition functions** of the frame bundle.

Components of a vector. A vector v has components $v_\alpha \in \mathbb{R}^n$ in the frame h_α, meaning $v = e_a^{(\alpha)} v_\alpha^a$. Comparing the two expansions on an overlap:

$$e_a^{(\alpha)} v_\alpha^a \overset{(1)}{=} v \overset{(2)}{=} e_b^{(\beta)} v_\beta^b \overset{(3)}{=} e_a^{(\alpha)} (g_{\alpha\beta})_{ab} v_\beta^b,$$

(1) Expansion of v in the frame h_α.
(2) Expansion of the same vector in the frame h_β.
(3) Insert (T4.59).

so $v_\alpha = g_{\alpha\beta}\, v_\beta$. On a triple overlap $U_\alpha \cap U_\beta \cap U_\gamma$ we can go from γ to α directly or through β, and consistency requires the **cocycle condition**

$$g_{\alpha\beta}\, g_{\beta\gamma} = g_{\alpha\gamma}, \qquad \text{equivalently} \qquad g_{\alpha\beta}\, g_{\beta\gamma}\, g_{\gamma\alpha} = \mathbf{1}, \qquad g_{\beta\alpha} = g_{\alpha\beta}^{-1}. \tag{T4.60}$$

For frames this is automatic, because it is literally the statement $h_\gamma = h_\alpha g_{\alpha\gamma} = h_\beta g_{\beta\gamma} = h_\alpha g_{\alpha\beta} g_{\beta\gamma}$.

Associated bundles. Every other tensor field is glued with the same $g_{\alpha\beta}$, fed through a representation. A k-form has components $\omega_\alpha \in \Lambda^k\mathbb{R}^n$ with $\omega_\alpha = \Lambda^k(g_{\alpha\beta})\,\omega_\beta$; a metric-type tensor has $T_\alpha = g_{\alpha\beta}T_\beta g_{\alpha\beta}^{\mathsf{T}}$; and so on. In general, a representation $\lambda : \mathrm{SO}(n) \to \mathrm{GL}(V)$ gives a vector bundle whose sections are collections $\phi_\alpha : U_\alpha \to V$ with $\phi_\alpha = \lambda(g_{\alpha\beta})\phi_\beta$. The cocycle condition for the $\lambda(g_{\alpha\beta})$ follows from (T4.60) because λ is a homomorphism. Bär calls this the **associated vector bundle** $P_{\mathrm{SO}}(M) \times_\lambda V$ [Bär Def. 2.4.2]; his table is:

fibre V	representation of $\mathrm{SO}(n)$	bundle
$\mathbb{R}$	trivial	functions, $M \times \mathbb{R}$
$\mathbb{R}^n$	defining, $v \mapsto gv$	TM
$(\mathbb{R}^n)^*$	dual, $\xi \mapsto \xi g^{-1}$	T^*M
$\Lambda^k(\mathbb{R}^n)^*$	Λ^k of the dual	$\Lambda^k T^*M$ (k-forms)
$\otimes^k\mathbb{R}^n \otimes \otimes^l(\mathbb{R}^n)^*$	tensor products	(k,l) tensors

Formally, $P_{\mathrm{SO}}(M)$ is the **oriented orthonormal frame bundle**: the space of all pairs (point, oriented orthonormal frame at that point), on which $\mathrm{SO}(n)$ acts by rotating frames [Bär Def. 2.4.1]. Everything we need is captured by the transition functions $g_{\alpha\beta}$.

Watch out: orientability comes first

If M is not orientable we can only arrange $g_{\alpha\beta} \in \mathrm{O}(n)$, and $\det g_{\alpha\beta} = \pm 1$ is a $\mathbb{Z}_2$-valued cocycle. Its class $w_1(M) \in H^1(M;\mathbb{Z}_2)$ is the **first Stiefel–Whitney class**; it vanishes exactly when the signs can be removed by flipping some frames, i.e. when M is orientable. This is the first of Nash's two conditions (4.96). The Möbius band and $\mathbb{RP}^2$ fail it.

T4.6.2 Lifting the transition functions

A spinor field should be a collection of local spinor-valued functions $\psi_\alpha : U_\alpha \to \mathbb{C}^N$ ($N = 2^{\lfloor n/2 \rfloor}$) glued by matrices that rotate spinors the way $g_{\alpha\beta}$ rotates vectors. From Section T4.5 we know which matrices those are: elements $\hat g_{\alpha\beta} \in \mathrm{Spin}(n)$ with $\rho(\hat g_{\alpha\beta}) = g_{\alpha\beta}$. So we try

$$\psi_\alpha = \hat g_{\alpha\beta}\,\psi_\beta, \qquad \rho(\hat g_{\alpha\beta}) = g_{\alpha\beta}, \qquad \hat g_{\beta\alpha} = \hat g_{\alpha\beta}^{-1}. \tag{T4.61}$$

Two things must be checked.

(i) Local lifts exist. On a contractible overlap (we may always refine the cover so that all overlaps are contractible), $g_{\alpha\beta}$ is a map from a contractible set into $\mathrm{SO}(n)$, and such a map can always be lifted continuously through the double cover ρ; the only freedom is an overall sign $\pm\hat g_{\alpha\beta}$ on each overlap, because a continuous choice of sign is constant on a connected set.

(ii) The cocycle condition. On a triple overlap,

$$\rho\big(\hat g_{\alpha\beta}\hat g_{\beta\gamma}\hat g_{\gamma\alpha}\big) \overset{(1)}{=} \rho(\hat g_{\alpha\beta})\,\rho(\hat g_{\beta\gamma})\,\rho(\hat g_{\gamma\alpha}) \overset{(2)}{=} g_{\alpha\beta}g_{\beta\gamma}g_{\gamma\alpha} \overset{(3)}{=} \mathbf{1},$$

$$\text{hence} \quad \hat g_{\alpha\beta}\,\hat g_{\beta\gamma}\,\hat g_{\gamma\alpha} \overset{(4)}{=} \varepsilon_{\alpha\beta\gamma}\,\mathbf{1}, \qquad \varepsilon_{\alpha\beta\gamma} \in \{+1,-1\}. \tag{T4.62}$$

(1) ρ is a group homomorphism.
(2) The $\hat g$'s are lifts.
(3) The cocycle condition (T4.60) for frames.
(4) The kernel of ρ is $\{\pm\mathbf{1}\}$ (T4.49); by continuity the sign is constant on the (connected) triple overlap.

If some $\varepsilon_{\alpha\beta\gamma} = -1$, the rule (T4.61) is inconsistent: going around the three patches, a spinor comes back as $-\psi$. We are allowed to flip the sign of any $\hat g_{\alpha\beta}$ (with $\hat g_{\beta\alpha}$ flipped too), which flips

the three signs ε of the triple overlaps containing $U_\alpha \cap U_\beta$. The question is whether all signs can be made $+1$.

Working definition T4.8 (spin structure)

A **spin structure** on an oriented Riemannian manifold is a choice of lifts $\hat{g}_{\alpha\beta} \in \mathrm{Spin}(n)$ of the frame transition functions [at each point, one of the two elements $\pm\hat{g}$ with $\rho(\hat{g}) = g_{\alpha\beta}$, the rotation converting frame components on U_β into those on U_α], continuous on each overlap, such that

$$\hat{g}_{\alpha\beta}\,\hat{g}_{\beta\gamma}\,\hat{g}_{\gamma\alpha} = +\mathbf{1} \qquad \text{on every triple overlap.}$$

Two choices $\hat{g}$ and $\hat{g}'$ are **equivalent** if $\hat{g}'_{\alpha\beta} = s_\alpha\,\hat{g}_{\alpha\beta}\,s_\beta$ for some signs $s_\alpha = \pm 1$, one per patch (this is just flipping the sign convention for the spinor components on U_α, $\psi_\alpha \to s_\alpha\psi_\alpha$).
How to test existence: pick any lifts, compute the signs $\varepsilon_{\alpha\beta\gamma}$, and ask whether they can all be removed by sign flips. *How to count:* if one spin structure exists, the others are obtained by multiplying the lifts by signs $\eta_{\alpha\beta}$ with $\eta_{\alpha\beta}\eta_{\beta\gamma}\eta_{\gamma\alpha} = 1$; in practice, by choosing independently the sign acquired around each non-contractible loop.
Formal definition [Bär Def. 2.4.3]: a principal $\mathrm{Spin}(n)$-bundle $P_{\mathrm{Spin}}(M)$ [a copy of $\mathrm{Spin}(n)$ over each point, glued between patches by multiplying with $\hat{g}_{\alpha\beta}$] with a two-to-one map $\bar{\rho} : P_{\mathrm{Spin}}(M) \to P_{\mathrm{SO}}(M)$ compatible with ρ. Its transition functions are our $\hat{g}_{\alpha\beta}$.

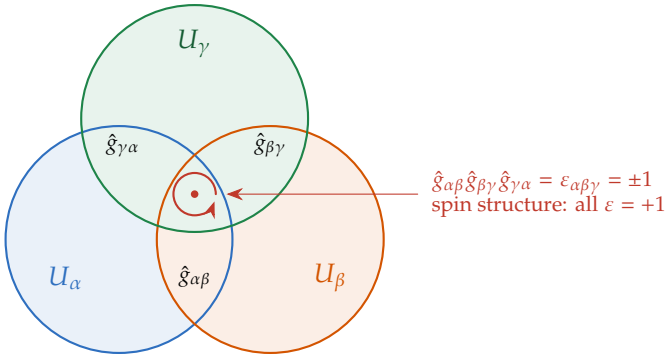


Figure T4.14: Three patches and their overlaps. On each pairwise overlap the frame rotation $g_{\alpha\beta}$ has two lifts $\pm\hat{g}_{\alpha\beta}$. Going once around the triple overlap (red loop) the product of the chosen lifts is a sign $\varepsilon_{\alpha\beta\gamma}$. A spin structure is a choice of lifts for which every such sign is $+1$; the collection of signs, modulo flips, is the obstruction w_2.

The obstruction w_2 and the counting by H^1. The pattern of signs $\varepsilon_{\alpha\beta\gamma}$ is a $\mathbb{Z}_2$-valued Čech 2-cocycle, and changing the lifts changes it by a coboundary. Its class

$$w_2(M) := [\varepsilon] \in H^2(M;\mathbb{Z}_2)$$

is the **second Stiefel–Whitney class**. It is independent of every choice we made, and

$$M \text{ admits a spin structure} \iff w_1(M) = 0 \text{ and } w_2(M) = 0. \tag{T4.63}$$

When a spin structure exists, any other differs from it by signs $\eta_{\alpha\beta}$ forming a $\mathbb{Z}_2$ Čech 1-cocycle, and equivalent ones differ by coboundaries. So spin structures are counted by $H^1(M;\mathbb{Z}_2)$. For a connected M, $H^1(M;\mathbb{Z}_2) = \mathrm{Hom}(\pi_1(M),\mathbb{Z}_2)$: a spin structure difference is a rule assigning a sign to every loop, multiplicative under composition of loops. For the rest of this chapter the operational content is:

Key idea: spin structures, operationally

- Existence: all triple-overlap signs can be made $+1$ ($w_2 = 0$). On a 2-sphere inside M this reduces to a parity check on a single winding number (below).
- Choice: once one exists, we may still choose independently whether spinors pick up $+1$ or -1 around each independent non-contractible loop. The number of spin structures is $|H^1(M;\mathbb{Z}_2)|$; if M is simply connected the spin structure is unique.

Watch out: a torsor, not a group

Nash says "each choice of a spin structure determines an element of $H^1(X;\mathbb{Z}_2)$, and vice versa" [Nash p. 113]. More precisely, the *difference* of two spin structures is an element of $H^1(X;\mathbb{Z}_2)$. The set of spin structures is in bijection with H^1, but there is in general no preferred "zero" spin structure. (On the circle, calling the periodic one "trivial" is natural because it is the product cover, but it is the other one that extends over the disc, as we see next.)

T4.6.3 The circle: periodic and antiperiodic spinors

Worked example T4.33: the two spin structures on S^1

On the circle, $n = 1$, $\mathrm{SO}(1) = \{1\}$ and $\mathrm{Spin}(1) = \{\pm 1\}$ (Example T4.29). The unit tangent vector $e_1 = \partial_\theta/|\partial_\theta|$ is a global frame, so we may take all $g_{\alpha\beta} = 1$. Its lifts are $\hat{g}_{\alpha\beta} = \pm 1$, and the cocycle condition just says the product of the signs is consistent. Cover S^1 by two arcs U_1, U_2 whose overlap consists of *two* disjoint intervals I and I'. On each interval we choose a sign. Flipping the sign convention on U_2 flips both, so only the product

$$\varepsilon = \hat{g}_{12}\big|_I \cdot \hat{g}_{12}\big|_{I'} = \pm 1$$

matters. This is the sign a spinor acquires when transported once around the circle. Unwrapping the circle to the real line, a spinor field is a function $\psi(\theta)$ with

$$\psi(\theta + 2\pi) = +\psi(\theta) \quad \text{(periodic)}, \qquad \text{or} \qquad \psi(\theta + 2\pi) = -\psi(\theta) \quad \text{(antiperiodic)}. \tag{T4.64}$$

These are Bär's trivial and non-trivial spin structures [Bär Ex. 2.4.6, 2.4.8]. As spaces, the two double covers $P_{\mathrm{Spin}}(S^1) \to P_{\mathrm{SO}}(S^1) = S^1$ are the disconnected cover $S^1 \sqcup S^1$ and the connected cover $z \mapsto z^2$ (Figure T4.15). They are inequivalent because one total space is connected and the other is not. And $H^1(S^1;\mathbb{Z}_2) = \mathbb{Z}_2$: two spin structures.

Which of the two is "natural"? A circle usually appears as the boundary of a disc, and then the answer is forced.

Worked example T4.34: the circle bounding a disc carries the antiperiodic structure

The plane $\mathbb{R}^2$ has the global frame (∂_x, ∂_y) and is simply connected, so it has exactly one spin structure: spinors are ordinary functions $\psi_{\mathrm{cart}}(x, y)$. Restrict to the unit circle, but now use the frame adapted to the circle, $e_r = \cos\theta\,\partial_x + \sin\theta\,\partial_y$, $e_\theta = -\sin\theta\,\partial_x + \cos\theta\,\partial_y$. This polar frame is the Cartesian one rotated by the angle θ:

$$h_{\mathrm{polar}} = h_{\mathrm{cart}}\,R(\theta), \qquad R(\theta) = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

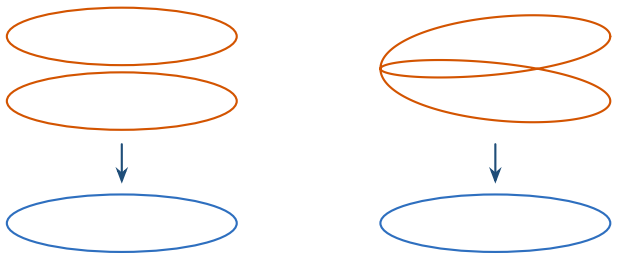


Figure T4.15: The two spin structures on a circle, drawn as double covers of the circle of frames (blue). Left: two separate copies; following a lift once around brings you back to the same sheet, so spinors are periodic. Right: a single curve that winds twice; once around the base takes you to the *other* sheet, which is the sign flip -1: spinors are antiperiodic.

By (T4.61) the spinor components in the two frames are related by a lift of $R(\theta)$, and the continuous lift is $a(\theta)$ of (T4.50):

$$\psi_{\text{polar}}(\theta) \overset{(1)}{=} a(\theta)^{-1}\,\psi_{\text{cart}} \overset{(2)}{=} \begin{pmatrix} e^{i\theta/2} & 0 \\ 0 & e^{-i\theta/2} \end{pmatrix} \psi_{\text{cart}},$$

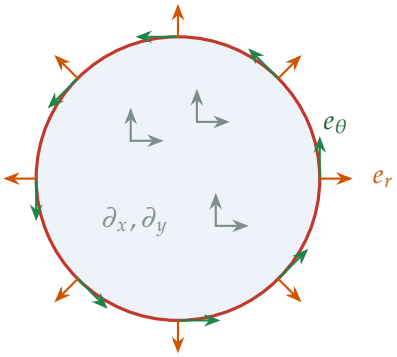
$$\psi_{\text{polar}}(\theta + 2\pi) \overset{(3)}{=} \begin{pmatrix} e^{i\theta/2}e^{i\pi} & 0 \\ 0 & e^{-i\theta/2}e^{-i\pi} \end{pmatrix} \psi_{\text{cart}} = -\psi_{\text{polar}}(\theta).$$

(1) $\psi_{\text{cart}} = \hat{g}\,\psi_{\text{polar}}$ with $\hat{g} = a(\theta)$ lifting $R(\theta)$; $a(\theta)$ must vary continuously with θ because the frames do.

(2) $a(\theta) = \operatorname{diag}(e^{-i\theta/2}, e^{i\theta/2})$ in the representation $\gamma_1 = \sigma_1$, $\gamma_2 = \sigma_2$.

(3) ψ_{cart} is a genuine single-valued function of the point, so only the half-angle phases change, each by $e^{\pm i\pi} = -1$.

So a perfectly smooth spinor on the disc, written in the frame tangent to its boundary, is *antiperiodic* around the boundary. The Neveu–Schwarz structure is the one that "bounds"; the Ramond structure does not extend over a disc. (Bär's general statement: a spin structure on M induces one on every oriented hypersurface [Bär §2.6, (2.45)]; the frame of the hypersurface is completed by the normal ν, here e_r.)



polar frame turns by 2π
relative to (∂_x, ∂_y);
its lift $a(\theta)$ ends at -1:
boundary spinors antiperiodic

Figure T4.16: Why the boundary of a disc carries the antiperiodic spin structure. The grey Cartesian frame is global on the disc. The frame adapted to the boundary (green tangent e_θ, orange normal e_r) rotates once as we go around. Spinor components in the adapted frame pick up the half-angle phase $e^{\pm i\theta/2}$, which is -1 after one circuit.

Dictionary: spin structures on a circle in physics

- **String theory.** A closed-string worldsheet fermion on the spatial circle is either periodic (**Ramond**, R) or antiperiodic (**Neveu–Schwarz**, NS). R-sector fermions have zero modes (Example T4.44), which is why the Ramond ground state is a spinor of spacetime; the NS sector contains the graviton. Under the conformal map from the cylinder to the plane (radial quantisation, Chapter 11), the NS sector is the one that extends smoothly over the origin: Example T4.34.
- **Thermal field theory.** The partition function $\operatorname{Tr} e^{-\beta H}$ is a path integral over Euclidean time $\tau \in [0, \beta]$ with fermions *antiperiodic* in τ (the Matsubara frequencies $(2k+1)\pi/\beta$). Inserting $(-1)^F$ flips the boundary condition: $\operatorname{Tr}(-1)^F e^{-\beta H}$ has *periodic* fermions. This twisted trace is the Witten index, a first physical appearance of the index of a Dirac operator (Chapter 4).

T4.6.4 Tori and surfaces

Worked example T4.35: the four spin structures on T^2

Let $T^2 = \mathbb{R}^2/(2\pi\mathbb{Z})^2$ with the flat metric. The coordinate frame (∂_x, ∂_y) is global and orthonormal, so the frame bundle is trivial and all $g_{\alpha\beta} = \mathbf{1}$. Every choice of lifts is a choice of sign around each of the two independent loops ($x \to x + 2\pi$ and $y \to y + 2\pi$). A spinor field is a function on $\mathbb{R}^2$ with

$$\psi(x + 2\pi, y) = \varepsilon_1\, \psi(x, y), \qquad \psi(x, y + 2\pi) = \varepsilon_2\, \psi(x, y), \qquad \varepsilon_1, \varepsilon_2 \in \{\pm 1\}. \tag{T4.65}$$

The loop $x \to x + 2\pi$ followed by $y \to y + 2\pi$ is homotopic to the opposite order, and the signs commute, so there is no further condition. We get

$$(\varepsilon_1, \varepsilon_2) \in \{(+,+), (+,-), (-,+), (-,-)\}, \qquad |H^1(T^2; \mathbb{Z}_2)| = |\mathbb{Z}_2 \oplus \mathbb{Z}_2| = 4.$$

In string language: (R,R), (R,NS), (NS,R), (NS,NS). The three structures with at least one antiperiodic direction are called **even** and the fully periodic one **odd**; the modular group of the torus permutes the three even ones among themselves and fixes the odd one (the parity is the Arf invariant). Only the odd structure $(+,+)$ will have zero modes of the Dirac operator (Example T4.45).

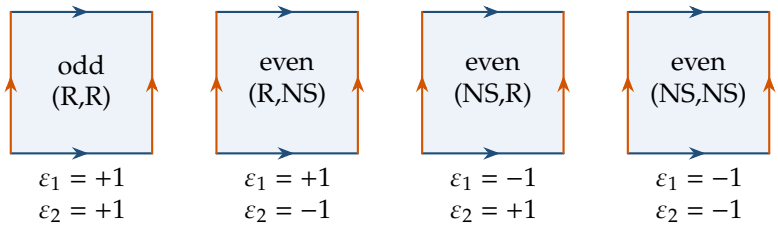


Figure T4.17: The four spin structures on the torus. Opposite edges of each square are glued. Crossing an orange edge (going once around in x) multiplies a spinor by ε_1; crossing a blue edge (going once around in y) multiplies it by ε_2. Only the fully periodic choice (far left) admits constant spinors.

The same reasoning gives 2^n spin structures on the flat n-torus T^n (one sign per circle), and more generally 2^{b_1} on any parallelisable manifold with $H^1(M; \mathbb{Z}_2) = \mathbb{Z}_2^{b_1}$ (for instance every Lie group, whose tangent bundle is trivialised by left-invariant vector fields). A closed oriented

surface of genus g has $H^1 = \mathbb{Z}_2^{2g}$, hence 2^{2g} spin structures; it always has at least one, because w_2 of a surface is the Euler characteristic mod 2, and $\chi = 2 - 2g$ is even. (This parity statement is exactly the computation we now do for the sphere.)

T4.6.5 The sphere S^2: one spin structure, found by a winding number

The sphere is simply connected, so $H^1(S^2;\mathbb{Z}_2) = 0$ and there is at most one spin structure. Does it exist? We compute.

Worked example T4.36: the clutching function of TS^2 and its lift

Cover $S^2 = \mathbb{C} \cup \{\infty\}$ by $U_0 = S^2 \setminus \{\infty\}$ with complex coordinate z and $U_\infty = S^2 \setminus \{0\}$ with $w = 1/z$. The round metric is $ds^2 = 4|dz|^2/(1+|z|^2)^2$, conformal to the flat one, so on U_0 the rescaled coordinate vectors $e_1^{(0)} \propto \partial_x$, $e_2^{(0)} \propto \partial_y$ are an oriented orthonormal frame; similarly on U_∞ using w. Identify a tangent vector at a point of U_0 with the complex number δz describing its direction. The frame vector $e_1^{(\infty)}$ points in the direction $\delta w = 1$. Since $\delta w = -\delta z/z^2$ (the derivative of $w = 1/z$), in the z-picture

$$\delta z\big|_{e_1^{(\infty)}} \overset{(1)}{=} -z^2\,\delta w \overset{(2)}{=} -z^2 \overset{(3)}{=} |z|^2 e^{i(\pi+2\phi)} \qquad (z = |z|e^{i\phi}).$$

(1) Invert $\delta w = -\delta z/z^2$.
(2) $e_1^{(\infty)}$ has δw real and positive; only the direction matters.
(3) $-1 = e^{i\pi}$ and $z^2 = |z|^2 e^{2i\phi}$.

So on the overlap the frame h_∞ is h_0 rotated by the angle $\pi + 2\phi$:

$$g_{0\infty}(z) = R(\pi + 2\phi), \qquad \phi = \arg z.$$

Going once around the equator ($\phi : 0 \to 2\pi$) the rotation angle increases by 4π: the **clutching function** has winding number 2. (This is the Euler number $\chi(S^2) = 2$: the frames h_0 and h_∞ each extend over one hemisphere, and their relative winding counts the zeros of a vector field.) Now lift. With two patches there are no triple overlaps, so a spin structure is just a continuous lift of $g_{0\infty}$ on the overlap $\mathbb{C}^\times$. Try

$$\hat{g}_{0\infty}(z) = a(\pi + 2\phi) = \exp\!\left(-\frac{\pi + 2\phi}{2}\,\gamma_1\gamma_2\right).$$

As $\phi \to \phi + 2\pi$ the argument of a increases by 4π, and $a(\theta + 4\pi) = a(\theta)$ by (T4.47). So $\hat{g}_{0\infty}$ is single valued: the lift exists and S^2 is spin, with a unique spin structure. In the representation $\gamma_j = \sigma_j$, $\hat{g}_{0\infty} = \mathrm{diag}(e^{-i(\pi+2\phi)/2}, e^{i(\pi+2\phi)/2}) = \mathrm{diag}(-i\,e^{-i\phi},\ i\,e^{i\phi})$, visibly single valued.

Key idea: the parity test on a sphere

For an oriented rank-2 bundle over S^2, glued by a clutching function of winding number d, the lift of the clutching function to Spin(2) closes up exactly when d is even: the lift of a rotation by total angle $2\pi d$ is $a(2\pi d) = (-1)^d$. The same holds for higher-rank bundles over S^2, with d the class of the clutching loop in $\pi_1(\mathrm{SO}(r)) = \mathbb{Z}_2$ ($r \geq 3$). So

$$\langle w_2(E), [S^2] \rangle = d \bmod 2.$$

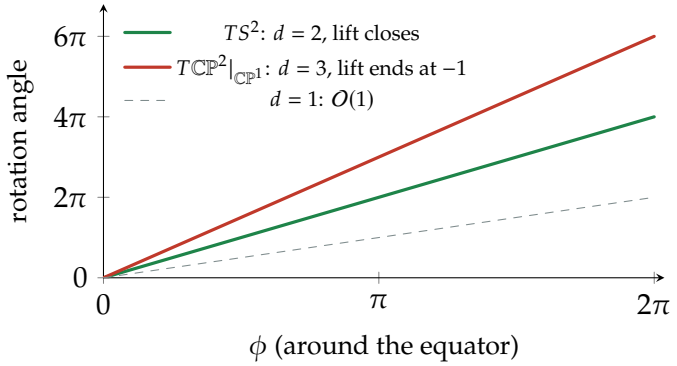


Figure T4.18: Clutching functions over a 2-sphere, drawn as the total rotation angle of the frame as we go once around the equator. A rotation by total angle $2\pi d$ lifts to $a(2\pi d) = (-1)^d$. For the tangent bundle of S^2 the angle grows by 4π ($d = 2$) and the lifted transition function is single valued. For $\mathbb{CP}^2$ restricted to a line the total is $d = 3$, odd, and no consistent sign exists.

T4.6.6 $\mathbb{CP}^2$: a manifold with no spin structure

Nash states that $\mathbb{CP}^n$ is spin exactly when n is odd [Nash p. 113]. The first failure is $\mathbb{CP}^2$, and the parity test shows why.

Worked example T4.37: $\mathbb{CP}^2$ is not spin

Inside $\mathbb{CP}^2$ sits the line $\mathbb{CP}^1 = \{[z_0 : z_1 : 0]\} \cong S^2$. Along it the tangent bundle of $\mathbb{CP}^2$ splits into directions along the line and directions normal to it:

$$T\mathbb{CP}^2\big|_{\mathbb{CP}^1} = T\mathbb{CP}^1 \oplus N = O(2) \oplus O(1).$$

Here $O(k)$ denotes the complex line bundle over S^2 whose clutching function winds k times. We found $T\mathbb{CP}^1 = TS^2$ has $k = 2$. The normal bundle of a line in $\mathbb{CP}^2$ has $k = 1$: in the chart $z_0 = 1$ the normal direction is the coordinate $u = z_2/z_0$, in the chart $z_1 = 1$ it is $u' = z_2/z_1$, and on the overlap $u = (z_1/z_0)\,u' = z\,u'$ with $z = z_1/z_0$: one winding. So the rank-4 real bundle $T\mathbb{CP}^2|_{\mathbb{CP}^1}$ is glued by

$$g(\phi) = \mathrm{diag}\big(R(2\phi + \mathrm{const}),\, R(\phi)\big) \in \mathrm{SO}(2) \times \mathrm{SO}(2) \subset \mathrm{SO}(4).$$

Its would-be lift is a product of lifts in the two commuting planes,

$$\hat{g}(\phi) \overset{(1)}{=} a_{12}(2\phi + c)\, a_{34}(\phi),$$
$$\hat{g}(\phi + 2\pi) \overset{(2)}{=} a_{12}(2\phi + c + 4\pi)\, a_{34}(\phi + 2\pi) \overset{(3)}{=} (+1)(-1)\, \hat{g}(\phi) = -\hat{g}(\phi),$$

(1) $a_{12}(\theta) = \exp(-\frac{\theta}{2}\gamma_1\gamma_2)$ and $a_{34}(\theta) = \exp(-\frac{\theta}{2}\gamma_3\gamma_4)$ commute and lift the two plane rotations; any other continuous lift differs by an overall sign.

(2) Go once around the equator.

(3) $a(\theta + 4\pi) = a(\theta)$ but $a(\theta + 2\pi) = -a(\theta)$ (T4.47).

No continuous single-valued lift exists, so $w_2(\mathbb{CP}^2)$ is nonzero on the line: $\mathbb{CP}^2$ has no spin structure. (Restricting a spin structure of $\mathbb{CP}^2$ to the line would produce such a lift.) For $\mathbb{CP}^n$ the same computation gives $T\mathbb{CP}^n|_{\mathbb{CP}^1} = O(2) \oplus O(1)^{\oplus(n-1)}$, total winding $n + 1$: spin iff n is odd, in agreement with Nash. In characteristic-class language, w_2 of a complex manifold is c_1 mod 2, and $c_1(\mathbb{CP}^n) = (n + 1)h$.

Physical picture: what goes wrong physically, and the way out

On $\mathbb{CP}^2$ one cannot write the Dirac equation for an uncharged fermion. But the obstruction is "half a unit of monopole flux" through the line: the sign -1 that spoils the lift is the same sign a charged particle picks up going around half a Dirac string. So if the fermion carries charge under a U(1) gauge field whose flux through the line is *half-integral*, the two signs cancel and the combined object is well defined. This is a **spin^c structure**; every complex manifold (indeed every oriented 4-manifold) has one, and it is how Seiberg–Witten theory and the Dolbeault operator (Section T4.9) handle $\mathbb{CP}^2$.

Watch out: a useful test in four dimensions

For a closed oriented 4-manifold, $w_2(M)$ evaluated on a surface Σ equals the self-intersection $\Sigma \cdot \Sigma \bmod 2$ (the Wu formula). So a spin 4-manifold has an *even* intersection form. $\mathbb{CP}^2$ has intersection form (1) (the line meets a displaced copy once), so it is not spin. S^4, $S^2 \times S^2$ (form $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$) and the K3 surface are spin. We use this in the instanton and index chapters.

T4.6.7 The spinor bundle and Clifford multiplication

With a spin structure in hand, spinor fields are defined by (T4.61): a spinor field is a collection $\psi_\alpha : U_\alpha \to \mathbb{C}^N$ with $\psi_\alpha = \hat{g}_{\alpha\beta}\psi_\beta$. The bundle they are sections of is the **spinor bundle** $S = P_{\mathrm{Spin}}(M) \times_\sigma \mathbb{C}^N$ [Bär Def. 2.4.7]; Nash calls it E [Nash (4.98)]. Three structures carry over from a single fibre to the whole bundle, each for a one-line reason.

Hermitian metric. Define $\langle \psi, \chi \rangle := \psi_\alpha^\dagger \chi_\alpha$ in any patch. It does not depend on the patch because $\hat{g}_{\alpha\beta}$ is unitary (Section T4.5.3, fact 1): $\psi_\alpha^\dagger \chi_\alpha = \psi_\beta^\dagger \hat{g}_{\alpha\beta}^\dagger \hat{g}_{\alpha\beta} \chi_\beta = \psi_\beta^\dagger \chi_\beta$.

Clifford multiplication. A tangent vector v acts on spinors by $v \cdot \psi := \not{v}_\alpha \psi_\alpha = v_\alpha^a \gamma_a \psi_\alpha$ in the patch U_α [Bär Def. 2.4.9]. This is patch independent:

$$\not{v}_\alpha \, \psi_\alpha \overset{(1)}{=} \gamma(g_{\alpha\beta} v_\beta)\, \hat{g}_{\alpha\beta} \psi_\beta \overset{(2)}{=} \hat{g}_{\alpha\beta}\, \gamma(v_\beta)\, \hat{g}_{\alpha\beta}^{-1}\, \hat{g}_{\alpha\beta} \psi_\beta \overset{(3)}{=} \hat{g}_{\alpha\beta} \big(\not{v}_\beta \psi_\beta \big),$$

(1) Vector components glue with g, spinor components with $\hat{g}$; write $\gamma(v) := v^a \gamma_a$.
(2) $\rho(\hat{g})v = \hat{g}\, \gamma(v)\, \hat{g}^{-1}$ is the definition of ρ (T4.48).
(3) Cancel $\hat{g}^{-1}\hat{g}$.

so $v \cdot \psi$ glues like a spinor. The Clifford relation $v \cdot w \cdot \psi + w \cdot v \cdot \psi = 2\langle v, w \rangle \psi$ holds pointwise because it holds for the γ's, and $v\cdot$ is Hermitian (Bär's skew-symmetry, with his $v = -i\gamma(v)$).

Chirality. In even dimension $n = 2k$, $\hat{g}_{\alpha\beta}$ is even, so it commutes with the chirality matrix Γ (T4.36). Hence the eigenspaces of Γ are glued separately and $S = S^+ \oplus S^-$, with rank $S^\pm = 2^{k-1}$: Nash's (4.97)–(4.98). Clifford multiplication by a vector anticommutes with Γ and so swaps $S^+ \leftrightarrow S^-$.

Worked example T4.38: the spinor bundle of S^2 in the two charts

With $\hat{g}_{0\infty} = \mathrm{diag}(-ie^{-i\phi},\ ie^{i\phi})$ from Example T4.36, the positive-chirality component (upper entry, $\Gamma = \sigma_3$) is glued by $\psi_0^+ = -ie^{-i\phi}\psi_\infty^+$ and the negative one by $\psi_0^- = ie^{i\phi}\psi_\infty^-$. Each is a complex line bundle over S^2 with clutching winding ∓ 1. Compare TS^2, a complex line bundle with winding $+2$ (Example T4.36): the spinor bundles are its "square roots" $S^\pm \cong (TS^2)^{\mp 1/2}$, in holomorphic language $S^+ \cong \mathcal{O}(-1)$ and $S^- \cong \mathcal{O}(1)$ (up to the orientation

convention). This is the concrete meaning of "spinors are square roots of vectors", and it only makes sense because the winding 2 is even.

Reading Nash (p. 113, (4.96)–(4.98))

Nash compresses this section into one paragraph: a spin structure is a lift of the structure group from $\mathrm{SO}(n)$ to $\mathrm{Spin}(n)$; it exists iff $w_1(X) = 0$ (orientability) and $w_2(X) = 0$ ("spinors are globally defined"), (4.96); spin structures are parametrised by $H^1(X;\mathbb{Z}_2)$ (read: form a torsor over it); simply connected manifolds have at most one; and $\mathbb{CP}^n$ is spin iff n is odd. His (4.98), $E^\pm = \tilde{P} \times_{\mathrm{Spin}(n)} S^\pm$, is our spinor bundle glued by $\hat{g}_{\alpha\beta}$ acting on the two chiral halves. We have made each claim concrete: (T4.62) is the w_2 cocycle, Example T4.33 and Example T4.35 are the H^1 counts, Example T4.36 and Example T4.37 are the parity test.

Recap

- Frame transition functions $g_{\alpha\beta} \in \mathrm{SO}(n)$ glue vectors and tensors. Spinors need lifts $\hat{g}_{\alpha\beta} \in \mathrm{Spin}(n)$ satisfying the cocycle condition; the obstruction is the sign pattern $\varepsilon_{\alpha\beta\gamma}$, whose class is $w_2(M)$.
- When spin structures exist, they are counted by $H^1(M;\mathbb{Z}_2)$: a free choice of sign around each independent loop. S^1: two (R and NS); T^2: four; genus g: 2^{2g}; S^n ($n \geq 2$): one.
- Over a 2-sphere the test is the parity of a winding number: TS^2 winds twice (spin), $T\mathbb{CP}^2|_{\mathbb{CP}^1}$ three times (not spin). $\mathbb{CP}^n$ is spin iff n is odd.
- The circle bounding a disc inherits the antiperiodic (NS) structure.
- The spinor bundle carries a Hermitian metric, Clifford multiplication by vectors, and (in even dimensions) a chiral splitting $S = S^+ \oplus S^-$.

T4.7 The spin connection and the Dirac operator

We now have spinor fields on a spin manifold: local functions ψ_α glued by the lifted transition functions $\hat{g}_{\alpha\beta}$. To write a wave equation for them we must *differentiate* them. On $\mathbb{R}^n$ we would write $-i\gamma^a\partial_a\psi$. On a curved manifold $\partial_a\psi_\alpha$ is not a spinor: the gluing matrices $\hat{g}_{\alpha\beta}$ depend on the point, so $\partial(\hat{g}\psi) \neq \hat{g}\,\partial\psi$. This is the same problem as for vectors, where the cure is the Levi-Civita connection, and the same problem as for charged fields, where the cure is a gauge field. For spinors the cure is a connection built out of the Levi-Civita one, and the recipe is fixed completely by the spin generators $\frac{1}{4}\gamma_a\gamma_b$ of Section T4.5.5.

The questions of this section are:

- How do we differentiate a spinor field? (Answer: the **spin connection** $\nabla = \mathrm{d} + \frac{1}{4}\omega_{ab}\gamma^a\gamma^b$.)
- What is the Dirac operator on a curved manifold, and what are its properties? (Answer: $D = -i\gamma^a\nabla_a$; it is self-adjoint and swaps chiralities.)
- What does its spectrum look like? We compute it by hand on $\mathbb{R}^n$, S^1, T^2 and S^2, and watch the spin structure decide whether there are zero modes.

T4.7.1 Reminder: the Levi-Civita connection in an orthonormal frame

Everything in this section is done in a local oriented orthonormal frame $e_1, \dots, e_n$ with dual coframe $e^1, \dots, e^n$ (so $e^a(e_b) = \delta^a_b$ and the metric is $g = \sum_a e^a \otimes e^a$). The Levi-Civita connection is encoded in the matrix of **connection 1-forms** $\omega^a{}_b$, defined by how the frame vectors turn as we move:

$$\nabla_X e_b = e_a\, \omega^a{}_b(X), \qquad \omega_{ab} \coloneqq \omega^a{}_b = -\omega_{ba}. \tag{T4.66}$$

Indices are raised and lowered with δ_{ab}, so the position of frame indices does not matter. Antisymmetry is metric compatibility: differentiating $\langle e_a, e_b\rangle = \delta_{ab}$ gives $0 = \langle \nabla e_a, e_b\rangle + \langle e_a, \nabla e_b\rangle = \omega_{ba} + \omega_{ab}$. So ω is an $\mathfrak{so}(n)$-valued 1-form, an infinitesimal rotation per unit displacement. The fastest way to compute it is **Cartan's first structure equation** (torsion-freeness written in the frame; see Chapter T2):

$$\mathrm{d}e^a + \omega^a{}_b \wedge e^b = 0. \tag{T4.67}$$

Its curvature is the matrix of 2-forms

$$R^a{}_b = \mathrm{d}\omega^a{}_b + \omega^a{}_c \wedge \omega^c{}_b, \qquad R_{ab} = \tfrac{1}{2} R_{abcd}\, e^c \wedge e^d, \qquad R(X,Y)e_b = e_a\, R^a{}_b(X,Y), \tag{T4.68}$$

so that $R_{abcd} = \langle R(e_c, e_d)e_b, e_a\rangle$. With this convention the sectional curvature of the plane $e_a \wedge e_b$ is R_{abab} (no sum), the Ricci tensor is $\mathrm{Ric}_{bd} = \sum_a R_{abad}$ and the scalar curvature is $R = \sum_{a,b} R_{abab}$. The unit sphere has $R_{abab} = 1$ for $a \neq b$ and $R = n(n-1)$.

In coordinates, with vielbein components $e_b = e_b{}^\nu \partial_\nu$, the connection forms are $\omega^a{}_{b\mu} = e^a{}_\nu \big(\partial_\mu e_b{}^\nu + \Gamma^\nu_{\mu\lambda} e_b{}^\lambda\big)$, which is just (T4.66) with $X = \partial_\mu$. Bär writes $\nabla_{b_i} b_j = \sum_k \Gamma^k_{ij} b_k$ for the same thing, so his "Christoffel symbols of the frame" are $\Gamma^k_{ij} = \omega_{kj}(b_i)$ [Bär (2.19)].

Worked example T4.39: the round sphere S^2: frame, connection and curvature

Use polar coordinates, $g = \mathrm{d}\theta^2 + \sin^2\theta\, \mathrm{d}\phi^2$, and the frame

$$e_1 = \partial_\theta, \qquad e_2 = \frac{1}{\sin\theta}\,\partial_\phi, \qquad e^1 = \mathrm{d}\theta, \qquad e^2 = \sin\theta\, \mathrm{d}\phi.$$

It is oriented ($e^1 \wedge e^2 = \sin\theta\, \mathrm{d}\theta \wedge \mathrm{d}\phi$ is the area form) and orthonormal, but singular at the two poles, where ∂_ϕ vanishes. Apply (T4.67):

$$\mathrm{d}e^2 \overset{(1)}{=} \cos\theta\, \mathrm{d}\theta \wedge \mathrm{d}\phi \overset{(2)}{=} -\omega^2{}_1 \wedge e^1 = \mathrm{d}\theta \wedge \omega^2{}_1 \quad \Longrightarrow \quad \omega^2{}_1 = \cos\theta\, \mathrm{d}\phi,$$
$$\mathrm{d}e^1 \overset{(3)}{=} 0 \overset{(4)}{=} -\omega^1{}_2 \wedge e^2 = \cos\theta\, \mathrm{d}\phi \wedge \sin\theta\, \mathrm{d}\phi = 0.$$

(1) Differentiate $\sin\theta\, \mathrm{d}\phi$.
(2) Cartan's equation for $a = 2$; $\omega^2{}_2 = 0$ by antisymmetry, and $\alpha \wedge \beta = -\beta \wedge \alpha$ for 1-forms.
(3) $e^1 = \mathrm{d}\theta$ is exact.
(4) Consistency check of the $a = 1$ equation with $\omega^1{}_2 = -\cos\theta\, \mathrm{d}\phi$: $\mathrm{d}\phi \wedge \mathrm{d}\phi = 0$.

So $\omega_{12} = -\cos\theta\, \mathrm{d}\phi$. The curvature:

$$R_{12} = \mathrm{d}\omega_{12} + \underbrace{\omega_{1c} \wedge \omega_{c2}}_{=0\ (c\in\{1,2\})} = \sin\theta\, \mathrm{d}\theta \wedge \mathrm{d}\phi = e^1 \wedge e^2,$$

so $R_{1212} = 1$: Gauss curvature 1, scalar curvature $R = R_{1212} + R_{2121} = 2$. (The quadratic term vanishes because $\omega_{11} = \omega_{22} = 0$.)

T4.7.2 The spin connection

Working definition T4.9 (spin connection)

In a local orthonormal frame [vector fields $e_1, \dots, e_n$ on a patch U with $g(e_a, e_b) = \delta_{ab}$], with spinor components $\psi : U \to \mathbb{C}^N$ and constant gamma matrices γ_a, the **spin connection** is

$$\nabla\psi := \mathrm{d}\psi + \Omega\,\psi, \qquad \Omega := \frac{1}{4}\,\omega_{ab}\,\gamma^a\gamma^b = \frac{1}{2}\sum_{a<b}\omega_{ab}\,\gamma^a\gamma^b, \tag{T4.69}$$

where ω_{ab} are the connection 1-forms of the frame [fixed by $\mathrm{d}e^a = -\omega^a{}_b \wedge e^b$ and $\omega_{ab} = -\omega_{ba}$], i.e. $\nabla_X\psi = \partial_X\psi + \frac{1}{4}\omega_{ab}(X)\gamma^a\gamma^b\psi$. In coordinates, $\nabla_\mu\psi = \partial_\mu\psi + \frac{1}{4}\omega_{ab\mu}\gamma^a\gamma^b\psi$. *Recipe:* (1) choose a frame; (2) find ω_{ab} from (T4.67); (3) replace each infinitesimal rotation ω_{ab} by its spin generator: $\Omega = \rho_*^{-1}(\omega)$ from (T4.55). Bär's formula [Bär (2.21)] is identical after $e_a \to -i\gamma_a$.

Where does the $\frac{1}{4}$ come from? There are three ways to see that (T4.69) is forced, and each teaches something.

1. The Leibniz rule for Clifford multiplication. We want the spin connection to be compatible with everything we already have, and in particular with Clifford multiplication by vectors: for a vector field $Y = Y^c e_c$ we demand

$$\nabla_X\big(\gamma(Y)\psi\big) = \gamma(\nabla_X Y)\,\psi + \gamma(Y)\,\nabla_X\psi, \qquad \gamma(Y) := Y^c\gamma_c. \tag{T4.70}$$

Try $\nabla = \mathrm{d} + \Omega$ with an unknown matrix-valued 1-form Ω. Then

$$\begin{aligned}\nabla\big(\gamma(Y)\psi\big) - \gamma(\nabla Y)\psi - \gamma(Y)\nabla\psi &\overset{(1)}{=} (\mathrm{d}Y^c)\gamma_c\psi + Y^c\gamma_c\,\mathrm{d}\psi + \Omega\,Y^c\gamma_c\psi \\ &\quad - (\mathrm{d}Y^a + \omega_{ac}Y^c)\gamma_a\psi - Y^c\gamma_c(\mathrm{d}\psi + \Omega\psi) \\ &\overset{(2)}{=} Y^c\big([\Omega, \gamma_c] - \omega_{ac}\gamma_a\big)\psi.\end{aligned}$$

(1) The γ_c are constant matrices in our trivialisation; the covariant derivative of Y is $\nabla Y = e_a(\mathrm{d}Y^a + \omega_{ac}Y^c)$ by (T4.66).

(2) The $\mathrm{d}Y$ and $\mathrm{d}\psi$ terms cancel; what is left is a commutator.

So (T4.70) holds iff $[\Omega, \gamma_c] = \omega_{ac}\gamma_a$ for every c: *conjugation by Ω must rotate the gammas by the infinitesimal rotation ω.* That is precisely the property (T4.54) of $\frac{1}{4}A_{ab}\gamma_a\gamma_b$, with $A = \omega$ and $x = e_c$:

$$\Big[\tfrac{1}{4}\omega_{ab}\gamma_a\gamma_b,\ \gamma_c\Big] \overset{(1)}{=} (\omega\,e_c)^a\gamma_a \overset{(2)}{=} \omega_{ac}\gamma_a.$$

(1) (T4.54), which holds for any antisymmetric A, here the (form-valued) matrix ω.

(2) The a-th component of the column c of the matrix ω.

Is the answer unique? If Ω' also works, then $\Omega' - \Omega$ commutes with every γ_c, hence with the whole Clifford algebra, and so is a multiple of $\mathbf{1}$ (in even dimensions the γ's generate all matrices). The spin connection is therefore unique *up to adding a* U(1) *gauge field* $A\,\mathbf{1}$. Keeping that freedom is exactly what "twisting by a line bundle" means (Section T4.8.4); forbidding it gives (T4.69).

2. The half-angle. In two dimensions, $\Omega = \frac{1}{2}\omega_{12}\gamma^1\gamma^2$. If the frame turns by an angle $\vartheta(x)$ relative to a fixed one, then $\omega_{12} = -\mathrm{d}\vartheta$ (see the next example), and $\Omega = -\frac{1}{2}\mathrm{d}\vartheta\,\gamma_1\gamma_2 = a^{-1}\mathrm{d}a$ with $a = \exp(-\frac{\vartheta}{2}\gamma_1\gamma_2)$ the lift (T4.46). One factor $\frac{1}{2}$ counts ω_{12} and ω_{21} once; the other is the *half-angle* of the spinor lift. Spinors turn at half the rate of the frame.

3. Gluing. On an overlap, frames are related by $h' = h\,g$ with $g : U \cap U' \to \mathrm{SO}(n)$, the connection matrix transforms like a gauge field, $\omega' = g^{-1}\omega g + g^{-1}\mathrm{d}g$ (Chapter T3), and spinor components are related by $\psi = \hat{g}\,\psi'$ with $\rho(\hat{g}) = g$. For $\nabla\psi$ to glue like a spinor we need

$$\nabla\psi = \mathrm{d}(\hat{g}\psi') + \Omega\,\hat{g}\psi' \overset{(1)}{=} \hat{g}\big(\mathrm{d}\psi' + (\hat{g}^{-1}\mathrm{d}\hat{g} + \hat{g}^{-1}\Omega\hat{g})\psi'\big) \overset{(2)}{=} \hat{g}\,(\mathrm{d} + \Omega')\psi',$$

(1) Leibniz rule for $\mathrm{d}(\hat{g}\psi')$ and factor out $\hat{g}$.
(2) This is the requirement: the bracket must be the spin connection $\Omega' = \rho_*^{-1}(\omega')$ built from the primed frame.

It holds because ρ_*^{-1} converts each piece correctly: $\rho_*^{-1}(g^{-1}\omega g) = \hat{g}^{-1}\rho_*^{-1}(\omega)\hat{g}$ (conjugating a rotation generator by g corresponds to conjugating its spin generator by the lift $\hat{g}$, by the definition $\rho(\hat{g})x = \hat{g}x\hat{g}^{-1}$), and $\rho_*^{-1}(g^{-1}\mathrm{d}g) = \hat{g}^{-1}\mathrm{d}\hat{g}$ (differentiate $\rho(\hat{g}) = g$). Note that only the *lift* $\hat{g}$ enters: this is where the spin structure is used. Nash's one line "ω denotes the Riemannian spin connection on X" [Nash p. 114] hides all of this.

Worked example T4.40: flat space in polar coordinates: the spin connection is pure gauge

On $\mathbb{R}^2$ take the polar frame $e_1 = \partial_r$, $e_2 = \frac{1}{r}\partial_\phi$, $e^1 = \mathrm{d}r$, $e^2 = r\,\mathrm{d}\phi$. Then

$$\mathrm{d}e^2 \overset{(1)}{=} \mathrm{d}r \wedge \mathrm{d}\phi \overset{(2)}{=} \mathrm{d}r \wedge \omega^2{}_1 \implies \omega_{21} = \mathrm{d}\phi,\ \ \omega_{12} = -\mathrm{d}\phi,$$
$$\Omega \overset{(3)}{=} \tfrac{1}{2}\omega_{12}\gamma_1\gamma_2 \overset{(4)}{=} -\tfrac{i}{2}\,\sigma_3\,\mathrm{d}\phi \overset{(5)}{=} a(\phi)^{-1}\mathrm{d}a(\phi), \qquad a(\phi) = \mathrm{diag}(e^{-i\phi/2}, e^{i\phi/2}).$$

(1) Differentiate $r\,\mathrm{d}\phi$.
(2) Cartan's equation, as in Example T4.39.
(3) (T4.69); the two terms $ab = 12, 21$ are equal.
(4) $\gamma_1\gamma_2 = \sigma_1\sigma_2 = i\sigma_3$ in the representation $\gamma_j = \sigma_j$.
(5) $a^{-1}\mathrm{d}a = \mathrm{diag}(-\frac{i}{2}, \frac{i}{2})\mathrm{d}\phi$; $a(\phi)$ is the lift (T4.50) of the rotation $R(\phi)$ relating the polar frame to the Cartesian one.

The curvature $\mathrm{d}\Omega + \Omega \wedge \Omega = 0$, as it must be in flat space, but Ω is not zero: it is the Maurer–Cartan form of the frame rotation. Its holonomy around the origin is $\exp(-\oint \Omega) = a(2\pi)^{-1}a(0) = -\mathbf{1}$. This is the antiperiodicity of polar-frame spinors from Example T4.34 seen by the connection: a covariantly constant spinor, written in the rotating frame, comes back with a minus sign.

Three properties. The spin connection inherits three properties, each in one line.

1. *Leibniz rule for Clifford multiplication*, (T4.70) [Bär (2.23)]: shown above.

2. *Metric compatibility* [Bär (2.22)]. For $a \neq b$, $(\gamma_a\gamma_b)^\dagger = \gamma_b\gamma_a = -\gamma_a\gamma_b$, so Ω is anti-Hermitian, and

$$\mathrm{d}(\psi^\dagger\chi) \overset{(1)}{=} (\mathrm{d}\psi)^\dagger\chi + \psi^\dagger\mathrm{d}\chi \overset{(2)}{=} (\mathrm{d}\psi + \Omega\psi)^\dagger\chi + \psi^\dagger(\mathrm{d}\chi + \Omega\chi) = \langle\nabla\psi, \chi\rangle + \langle\psi, \nabla\chi\rangle.$$

   (1) Product rule.
   (2) Add $(\Omega\psi)^\dagger\chi + \psi^\dagger\Omega\chi = \psi^\dagger(\Omega^\dagger + \Omega)\chi = 0$.

3. *Chirality is preserved* [Bär Rem. 2.4.11]. In even dimension Ω is a sum of products of *two* gammas, which commute with Γ. So ∇ maps S^+-valued fields to S^+-valued forms, and likewise for S^-.

Curvature of the spin connection. Since $\Omega = \rho_*^{-1}(\omega)$ and ρ_*^{-1} is a Lie algebra homomorphism (it preserves commutators, as we checked for the Σ_{ab} in Section T4.5.5), the curvature of ∇ is the image of the Riemann curvature:

$$\begin{aligned} R^S := \mathrm{d}\Omega + \Omega\wedge\Omega &\overset{(1)}{=} \rho_*^{-1}(\mathrm{d}\omega) + \tfrac{1}{2}[\Omega\wedge\Omega] \overset{(2)}{=} \rho_*^{-1}(\mathrm{d}\omega) + \rho_*^{-1}\big(\tfrac{1}{2}[\omega\wedge\omega]\big) \\ &\overset{(3)}{=} \rho_*^{-1}(\mathrm{d}\omega + \omega\wedge\omega) \overset{(4)}{=} \frac{1}{4}R_{ab}\,\gamma^a\gamma^b = \frac{1}{8}R_{abcd}\,\gamma^a\gamma^b\,e^c\wedge e^d. \end{aligned} \tag{T4.71}$$

(1) ρ_*^{-1} is linear and x-independent, so it commutes with d. For a matrix-valued 1-form, $\Omega\wedge\Omega = \frac{1}{2}[\Omega\wedge\Omega]$, where $[\alpha\wedge\beta] := \alpha\wedge\beta + \beta\wedge\alpha$ is the graded commutator (write $\Omega = \Omega_\mu \mathrm{d}x^\mu$: $\Omega\wedge\Omega = \frac{1}{2}[\Omega_\mu,\Omega_\nu]\mathrm{d}x^\mu\wedge\mathrm{d}x^\nu$).
(2) ρ_*^{-1} preserves commutators.
(3) Undo step 1 for ω.
(4) Definition (T4.68) of the curvature matrix, and $\rho_*^{-1}(A) = \frac{1}{4}A_{ab}\gamma^a\gamma^b$ once more.

Evaluated on two vectors, $R^S(e_c,e_d) = \frac{1}{4}R_{abcd}\gamma^a\gamma^b$; this is Bär's (2.24),

$$R^\Sigma(X,Y)\phi = -\tfrac{1}{4}\textstyle\sum_i (R(X,Y)b_i)\cdot b_i\cdot\phi,$$

after the translation $b\to -i\gamma$.

Worked example T4.41: spin connection and spinor curvature on S^2

From Example T4.39, $\omega_{12} = -\cos\theta\,\mathrm{d}\phi$ and $R_{12} = e^1\wedge e^2$. With $\gamma_1=\sigma_1$, $\gamma_2=\sigma_2$, $\gamma_1\gamma_2 = i\sigma_3$:

$$\begin{aligned} \Omega &\overset{(1)}{=} \tfrac{1}{2}\omega_{12}\gamma_1\gamma_2 = -\tfrac{i}{2}\cos\theta\,\sigma_3\,\mathrm{d}\phi, \\ R^S &\overset{(2)}{=} \tfrac{1}{2}R_{12}\gamma_1\gamma_2 = \tfrac{i}{2}\,\sigma_3\,e^1\wedge e^2 = \begin{pmatrix} \frac{i}{2} & 0 \\ 0 & -\frac{i}{2}\end{pmatrix}\sin\theta\,\mathrm{d}\theta\wedge\mathrm{d}\phi. \end{aligned}$$

(1) (T4.69) in two dimensions.
(2) (T4.71); check directly: $\mathrm{d}\Omega = \frac{i}{2}\sin\theta\,\sigma_3\,\mathrm{d}\theta\wedge\mathrm{d}\phi$ and $\Omega\wedge\Omega = 0$ because Ω is a multiple of the single form $\mathrm{d}\phi$.

The chirality $\Gamma = \sigma_3$ is diagonal, so S^+ and S^- are two complex line bundles, each carrying a U(1) connection with curvature $F^\pm = \pm\frac{i}{2}\,\mathrm{vol}$. Their first Chern numbers are

$$c_1(S^\pm) \overset{(1)}{=} \frac{i}{2\pi}\int_{S^2} F^\pm \overset{(2)}{=} \frac{i}{2\pi}\Big(\pm\frac{i}{2}\Big)4\pi = \mp 1.$$

(1) The house convention $c_1 = \frac{i}{2\pi}\operatorname{tr}F$ (Chapter T3).
(2) The area of the unit sphere is 4π.

This agrees with the clutching computation of Example T4.38 (windings ∓1). So *on S^2 the chiral spinors are monopoles of charge $\mp\frac{1}{2}$ relative to the tangent bundle, which has $c_1 = \chi(S^2) = 2$.* We will see the charge again in the spectrum.

T4.7.3 The Dirac operator on a spin manifold

Working definition T4.10 (Dirac operator)

On a Riemannian spin manifold [one on which a spin structure has been chosen, so that spinor fields are defined], in a local orthonormal frame [$e_a = e_a{}^\mu \partial_\mu$ with $g(e_a, e_b) = \delta_{ab}$],

$$D\psi := -i\gamma^a \nabla_{e_a}\psi = -i\gamma^a e_a{}^\mu \Big(\partial_\mu + \frac{1}{4}\omega_{bc\mu}\gamma^b\gamma^c\Big)\psi. \tag{T4.72}$$

Recipe: compute Ω, then contract the covariant derivative with the gammas. The result does not depend on the frame: under $h' = hg$ both γ^a and ∇_{e_a} rotate, and $\gamma^a\nabla_{e_a}$ is a contraction. (In Clifford-module language, D is "Clifford multiplication composed with the connection", $D = -i\,c \circ \nabla$, Bär's (2.25) with $c(e_a) = -i\gamma_a$.)

Reading Nash (p. 114, (4.99)–(4.101))

Nash writes $\not{D} = \gamma^a e_a^\mu(x)(\partial_\mu + \omega_\mu)$ with anti-Hermitian gammas $\{\gamma_a, \gamma_b\} = -2\delta_{ab}$ and the vielbein relation $e_a^\mu e_b^\nu\, g_{\mu\nu} = \delta_{ab}$ written in (4.100) with the indices placed the other way. His ω_μ is our $\Omega_\mu = \frac{1}{4}\omega_{ab\mu}\gamma^a\gamma^b$ (the product of two of his gammas is minus the product of two of ours, and his ω_{ab} is our $-\omega_{ab}$ or ours depending on the index order; only the combination Ω_μ matters). His $\gamma_5 = i^{n(n+1)/2}\gamma_1\cdots\gamma_n$ is our Γ up to the change $\gamma^{\text{Nash}} = i\gamma$: in $n = 4$, $i^{10}(i)^4\gamma_1\gamma_2\gamma_3\gamma_4 = -\gamma_1\gamma_2\gamma_3\gamma_4 = \Gamma$. Since $\gamma_a^{\text{Nash}} = i\gamma_a$, his $\not{D}$ is $i\gamma^a\nabla_a = -D$. Nothing depends on the overall sign.

The symbol, and ellipticity. Replacing ∂_μ by $i\xi_\mu$ in the top-order part:

$$\sigma_D(x,\xi) = -i\gamma^a e_a{}^\mu(i\xi_\mu) = \gamma^a\xi_a = \not{\xi}, \qquad \not{\xi}^2 = |\xi|^2\,\mathbf{1}.$$

So D is of Dirac type (Section T4.1.3); the connection term is zeroth order and does not appear. The symbol is invertible for every $\xi \neq 0$, with $\not{\xi}^{-1} = \not{\xi}/|\xi|^2$: D is **elliptic**. Nash makes the same point [Nash p. 115]: the symbol of the chiral operator is $i\gamma^a e_a^\mu p_\mu$ with inverse $-i\gamma^a e_a^\mu p_\mu/p^2$, and this needs Euclidean signature: in Lorentzian signature $\not{p}^2 = p^2$ vanishes on the light cone and the Dirac operator is *not* elliptic (it is hyperbolic, which is why it propagates waves instead of having a discrete spectrum).

Formal self-adjointness. On a closed manifold (compact, no boundary)

$$(D\psi, \chi) := \int_M \langle D\psi, \chi\rangle \,\mathrm{vol} = (\psi, D\chi)$$

[Bär Lemma 2.5.5]. The proof is Stokes' theorem applied to a current. Define the complex vector field V by $V(Y) := \psi^\dagger\gamma(Y)\chi$ for all vectors Y, i.e. $V = V^a e_a$ with $V^a = \psi^\dagger\gamma^a\chi$. (In Minkowski physics this is the current $\bar\psi\gamma^\mu\chi$.) Compute its divergence in a frame with $(\nabla e_a)(x) = 0$ at the point x under consideration (such a **synchronous frame** always exists: parallel transport a frame out radially from x):

$$\begin{aligned}\operatorname{div} V &\overset{(1)}{=} \sum_a e_a\big(V^a\big) \overset{(2)}{=} \sum_a \Big[(\nabla_a\psi)^\dagger\gamma^a\chi + \psi^\dagger\gamma^a\nabla_a\chi\Big] \\ &\overset{(3)}{=} \big(i\,D\psi\big)^\dagger\chi + \psi^\dagger\big(i\,D\chi\big) \overset{(4)}{=} -i\big(\langle D\psi,\chi\rangle - \langle\psi, D\chi\rangle\big).\end{aligned} \tag{T4.73}$$

(1) $\operatorname{div} V = \sum_a \langle \nabla_{e_a} V, e_a\rangle = \sum_a \big[e_a(V^a) - \langle V, \nabla_{e_a} e_a\rangle\big]$, and the last term vanishes at x.

(2) Metric compatibility of ∇ and the Leibniz rule (T4.70); the extra term $\gamma(\nabla_{e_a} e_a)$ vanishes at x.
(3) $\gamma^a \nabla_a \psi = iD\psi$ by definition (T4.72), and $(\nabla_a \psi)^\dagger \gamma^a = (\gamma^a \nabla_a \psi)^\dagger$ since γ^a is Hermitian.
(4) $(iD\psi)^\dagger = -i(D\psi)^\dagger$.

Both sides of (T4.73) are frame independent, so it holds everywhere. Integrating over a closed M and using $\int_M \operatorname{div} V \operatorname{vol} = 0$ (the divergence theorem, i.e. Stokes for the $(n-1)$-form $\iota_V \operatorname{vol}$ on a manifold with empty boundary) gives $(D\psi, \chi) = (\psi, D\chi)$. On a manifold with boundary the same computation leaves a flux,

$$(D\psi, \chi) - (\psi, D\chi) = i \int_{\partial M} \psi^\dagger \, \gamma(\nu) \, \chi \, \mathrm{d}A, \tag{T4.74}$$

with ν the outward unit normal (Bär's (2.27) in his conventions). The boundary term is the flux of the fermion-number current through the boundary; choosing boundary conditions that kill it is the subject of the Atiyah–Patodi–Singer problem in Chapter 4.

Key idea: the Dirac operator and chirality

In even dimension $n = 2k$, γ^a anticommutes with Γ and ∇ commutes with it, so $D\Gamma = -\Gamma D$. In the basis $S = S^+ \oplus S^-$ in which $\Gamma = \operatorname{diag}(\mathbf{1}, -\mathbf{1})$,

$$D = \begin{pmatrix} 0 & D^- \\ D^+ & 0 \end{pmatrix}, \qquad D^\pm : \Gamma(S^\pm) \to \Gamma(S^\mp), \qquad D^- = (D^+)^*. \tag{T4.75}$$

The last statement is self-adjointness of D written in blocks [Bär Rem. 2.5.3]. Nash's chiral operator $\not{D}(1+\gamma_5)/2$ is D^+ [Nash (4.102)]. Its **index** $\dim \operatorname{Ker} D^+ - \dim \operatorname{Ker} D^-$ is the topic of Section T4.8.5.

Watch out: chirality is not a symmetry of D

Because D *anticommutes* with Γ, an eigenspinor of D with $\lambda \neq 0$ never has definite chirality: if $D\psi = \lambda\psi$ then $D(\Gamma\psi) = -\lambda\,\Gamma\psi$. Only zero modes can be chosen chiral. In Minkowski language: the massless Dirac equation preserves chirality, but a mass term (a nonzero eigenvalue in the Euclidean problem) mixes the two.

T4.7.4 Flat space: plane waves and holomorphic functions

Worked example T4.42: the Dirac operator on $\mathbb{R}^n$

With the Cartesian frame $\omega = 0$, so $D = -i\gamma^a \partial_a$. Try a plane wave $\psi = u\, e^{ik\cdot x}$ with a constant spinor u:

$$D\psi \overset{(1)}{=} -i\gamma^a (ik_a) u\, e^{ik\cdot x} = \not{k}\, u\, e^{ik\cdot x},$$
$$\not{k}^2 \overset{(2)}{=} \tfrac{1}{2} k_a k_b \{\gamma^a, \gamma^b\} = |k|^2.$$

(1) $\partial_a e^{ik\cdot x} = ik_a e^{ik\cdot x}$.
(2) Only the symmetric part of $\gamma^a \gamma^b$ survives the symmetric sum; Clifford relation.

So $\not{k}$ is a Hermitian matrix squaring to $|k|^2$: its eigenvalues are $\pm|k|$. For $n \geq 2$ it is traceless ($\operatorname{tr} \gamma^a = 0$, Example T4.24), so the two eigenvalues have equal multiplicity $\frac{1}{2} \cdot 2^{\lfloor n/2 \rfloor}$. The spectrum of D on $\mathbb{R}^n$ is all of $\mathbb{R}$, continuous, and $D^2 = -\partial_a \partial_a$ is the Laplacian. This is Dirac's square root again (Section T4.1), now reinterpreted: *each Fourier mode splits into a positive and a negative "energy" branch*, $\lambda = \pm|k|$.

In two dimensions with $\gamma_1 = \sigma_1$, $\gamma_2 = \sigma_2$ and $k = |k|(\cos\beta, \sin\beta)$,

$$\not{k} = \begin{pmatrix} 0 & k_1 - ik_2 \\ k_1 + ik_2 & 0 \end{pmatrix} = |k| \begin{pmatrix} 0 & e^{-i\beta} \\ e^{i\beta} & 0 \end{pmatrix}, \qquad u_\pm = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ \pm e^{i\beta} \end{pmatrix}, \quad \not{k}\, u_\pm = \pm|k|\, u_\pm.$$

Rotating k by 2π ($\beta \to \beta + 2\pi$) returns $u_\pm$ to itself, but the *relative* phase between the components winds once: this is the "Berry phase π" of massless Dirac fermions in graphene.

Worked example T4.43: in two dimensions D is the Cauchy–Riemann operator

Let $z = x + iy$, $\partial_z = \tfrac{1}{2}(\partial_x - i\partial_y)$, $\partial_{\bar z} = \tfrac{1}{2}(\partial_x + i\partial_y)$. Then

$$D \overset{(1)}{=} -i \begin{pmatrix} 0 & \partial_x - i\partial_y \\ \partial_x + i\partial_y & 0 \end{pmatrix} \overset{(2)}{=} -2i \begin{pmatrix} 0 & \partial_z \\ \partial_{\bar z} & 0 \end{pmatrix}, \qquad D^+ = -2i\,\partial_{\bar z}, \quad D^- = -2i\,\partial_z. \tag{T4.76}$$

(1) $\sigma_1\partial_x + \sigma_2\partial_y$ written out.
(2) Definitions of $\partial_z, \partial_{\bar z}$.

So a positive-chirality zero mode, $D^+\psi_+ = 0$, is exactly a *holomorphic* function, and a negative-chirality zero mode is an antiholomorphic one. This is the first place where the Dirac operator meets complex geometry; it is the seed of the statement that on a Riemann surface D^+ is the Dolbeault operator $\bar\partial$ acting on $S^+ = K^{1/2}$ (Section T4.9.8). On all of $\mathbb{R}^2$ there are no normalisable holomorphic functions (a bounded entire function is constant, and a nonzero constant is not square integrable), so D has no L^2 zero modes. The same equation with a magnetic field will have many (Example T4.52).

T4.7.5 The circle

Worked example T4.44: the Dirac operator on S^1 for both spin structures

Take the circle of length L, coordinate $t \in [0, L)$. In one dimension the Clifford algebra has a single generator with $\gamma_1^2 = 1$; we take $\gamma_1 = 1$ on spinors $\mathbb{C}^1$. The frame $e_1 = \partial_t$ is global and $\omega = 0$, so

$$D = -i\frac{\mathrm{d}}{\mathrm{d}t}.$$

It acts on functions with the boundary condition (T4.64) of the spin structure.

- **Periodic (Ramond, Bär's "trivial").** $\psi(t + L) = \psi(t)$. Fourier modes $\psi_k = e^{2\pi i k t/L}$, $k \in \mathbb{Z}$:
$$D\psi_k \overset{(1)}{=} -i \cdot \frac{2\pi i k}{L}\psi_k = \frac{2\pi k}{L}\psi_k, \qquad \operatorname{spec} D = \frac{2\pi}{L}\mathbb{Z}.$$
- **Antiperiodic (Neveu–Schwarz).** $\psi(t+L) = -\psi(t)$. The function $e^{i\pi t/L}\psi(t)$ is periodic (the factor contributes the missing -1), so $\psi_k = e^{2\pi i(k-\frac{1}{2})t/L}$, and
$$D\psi_k = \frac{2\pi}{L}\Big(k - \frac{1}{2}\Big)\psi_k, \qquad \operatorname{spec} D = \frac{2\pi}{L}\Big(\mathbb{Z} + \frac{1}{2}\Big).$$

(1) $\dfrac{\mathrm{d}}{\mathrm{d}t}e^{2\pi i k t/L} = \dfrac{2\pi i k}{L}e^{2\pi i k t/L}$ and $(-i)(i) = 1$.

All multiplicities are 1 [Bär Ex. 2.5.8]. The periodic structure has exactly one zero mode, the constant spinor; the antiperiodic one has none. Both spectra are symmetric under $\lambda \to -\lambda$.

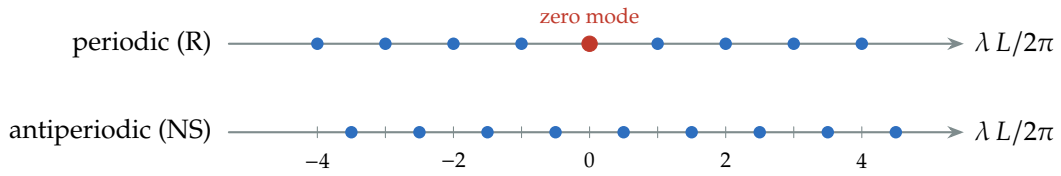


Figure T4.19: Spectrum of $D = -i\,\mathrm{d}/\mathrm{d}t$ on a circle of length L. Top: the periodic spin structure gives the integers, including a single zero mode (red, the constant spinor). Bottom: the antiperiodic structure shifts every eigenvalue by $\frac{1}{2}$ and there is no zero mode. The spin structure changes the spectrum, not just the bookkeeping.

The two spin structures as a flat gauge field. The two spin structures differ by a sign around the circle, and a sign is the holonomy of a flat U(1) connection. Let us make this precise, since it is the simplest example of *spectral flow*. Couple the periodic spinor to a constant gauge field $A = -i\,a\,\mathrm{d}t$ (anti-Hermitian, house convention; the physicist's field is a, with charge 1). The twisted operator is

$$D_a = -i\Big(\frac{\mathrm{d}}{\mathrm{d}t} - ia\Big) = -i\frac{\mathrm{d}}{\mathrm{d}t} - a, \qquad D_a e^{2\pi i k t/L} = \Big(\frac{2\pi k}{L} - a\Big)e^{2\pi i k t/L}.$$

The holonomy of the gauge field is $e^{-\oint A} = e^{iaL}$. At $a = 0$ we have the Ramond spectrum; at $a = \pi/L$ (holonomy -1) the spectrum is $\frac{2\pi}{L}(\mathbb{Z} - \frac{1}{2})$, the Neveu–Schwarz spectrum; indeed $e^{-i\pi t/L}$ times a periodic function is antiperiodic, and $D_{\pi/L}$ acting on periodic functions is unitarily equivalent to D_0 acting on antiperiodic ones. At $a = 2\pi/L$ the spectrum is back to itself, but *every eigenvalue has moved down by one step* (Figure T4.20). The net number of eigenvalues crossing zero from above, here 1, is the **spectral flow**. It is a topological number (it only depends on the winding of the holonomy e^{iaL} around U(1)) and it is the one-dimensional origin of the chiral anomaly: in Chapter 10 the adiabatic change of a background field pumps one unit of charge out of the Dirac sea.

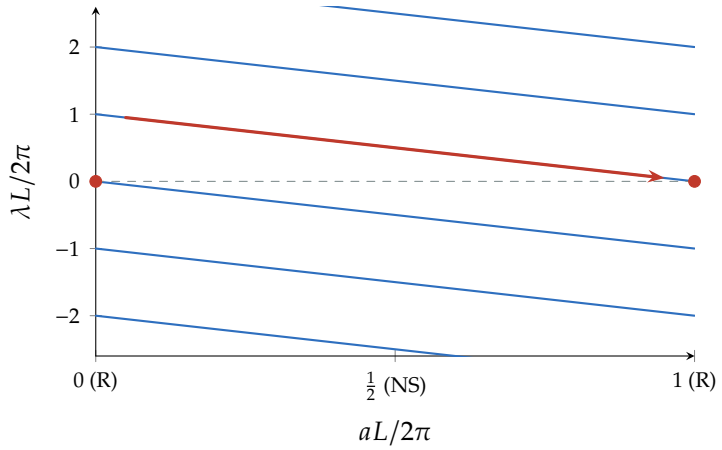


Figure T4.20: Spectral flow on the circle. As the flat gauge field a increases from 0 to $2\pi/L$, every eigenvalue $\frac{2\pi k}{L} - a$ slides down one step. Half way, at holonomy -1, we pass through the Neveu–Schwarz spectrum. One eigenvalue (red) crosses zero: the spectral flow is 1. The spectrum returns to itself as a set, but the eigenvalues have been relabelled.

Dictionary: the circle spectrum in physics

- **Ramond zero mode.** A fermion zero mode on the spatial circle is a Grassmann variable ψ_0 with no energy cost; quantising it gives a two-fold degenerate ground

state (the Clifford algebra of one fermionic zero mode). With many worldsheet fermions the zero modes form a Clifford algebra and the Ramond ground state is a *spacetime spinor* (Chapter 9).

- **Zero-point energies.** Formally the vacuum energy is a sum of eigenvalues. With ζ-function regularisation, $\sum_{k\geq 1} k = \zeta(-1) = -\frac{1}{12}$ and $\sum_{k\geq 1}(k-\frac{1}{2}) = \zeta(-1,\frac{1}{2}) = +\frac{1}{24}$ (Hurwitz zeta). These are the numbers behind the different normal-ordering constants of the R and NS sectors.
- **Thermal fermions.** With $L=\beta$ the antiperiodic spectrum is the set of fermionic Matsubara frequencies $(2k+1)\pi/\beta$.

T4.7.6 The torus

Worked example T4.45: the Dirac operator on T^2 for the four spin structures

Let $T^2 = \mathbb{R}^2/(2\pi\mathbb{Z})^2$, flat, with the global frame (∂_x, ∂_y), so $\omega = 0$ and D is (T4.76). A spinor field obeys the boundary conditions (T4.65) with signs $(\varepsilon_1, \varepsilon_2)$. Plane waves $u\, e^{i(px+qy)}$ satisfy them iff

$$p \in \mathbb{Z} \text{ if } \varepsilon_1 = +1, \quad p \in \mathbb{Z} + \tfrac{1}{2} \text{ if } \varepsilon_1 = -1; \qquad \text{and the same for } q \text{ and } \varepsilon_2.$$

For each allowed (p,q), Example T4.42 gives two eigenvalues

$$\lambda = \pm\sqrt{p^2+q^2},$$

each with multiplicity one, except that when $(p,q)=(0,0)$ the matrix $\not{k}=0$ and both spinors $u=(1,0)^{\mathsf{T}}$ and $(0,1)^{\mathsf{T}}$ are zero modes. The momentum $(0,0)$ is allowed only for $(\varepsilon_1,\varepsilon_2)=(+,+)$. So:

spin structure $(\varepsilon_1,\varepsilon_2)$	$(+,+)$	$(+,-)$	$(-,+)$	$(-,-)$
allowed (p,q)	$\mathbb{Z}\times\mathbb{Z}$	$\mathbb{Z}\times(\mathbb{Z}+\frac{1}{2})$	$(\mathbb{Z}+\frac{1}{2})\times\mathbb{Z}$	$(\mathbb{Z}+\frac{1}{2})^2$
zero modes (n_+,n_-)	$(1,1)$	$(0,0)$	$(0,0)$	$(0,0)$
smallest $\lvert\lambda\rvert>0$	1 (mult. 4 each sign)	$\frac{1}{2}$ (mult. 2)	$\frac{1}{2}$ (mult. 2)	$\frac{1}{\sqrt{2}}$ (mult. 4)
index n_+-n_-	0	0	0	0

In the $(+,+)$ case the zero modes are the constants: $D^+\psi_+ = -2i\partial_{\bar z}\psi_+ = 0$ means ψ_+ is a holomorphic doubly periodic function, hence constant (Liouville on the compact torus), and similarly for ψ_-. In the other cases a constant does not satisfy the antiperiodic condition, and neither does any other holomorphic function. The index vanishes in every case, as it must by the index theorem in two dimensions without a gauge field ($\hat A = 1$ has no 2-form part; Section T4.8.5).

Try it: a torus with a flat twist

Couple the $(+,+)$ spinor on T^2 to a constant gauge field $A = -i(a_1\mathrm{d}x + a_2\mathrm{d}y)$. Show that the eigenvalues become $\pm\sqrt{(p-a_1)^2+(q-a_2)^2}$, $p,q\in\mathbb{Z}$, so that zero modes exist only when $(a_1,a_2)\in\mathbb{Z}^2$, and that $(a_1,a_2)=(\frac{1}{2},0)$ reproduces the $(-,+)$ spectrum. *Answer:* $D_A = -i\gamma^a(\partial_a - ia_a)$ acts on $u\,e^{i(px+qy)}$ as $\gamma^a(k_a-a_a)$ with $k=(p,q)$; square it. The shift by

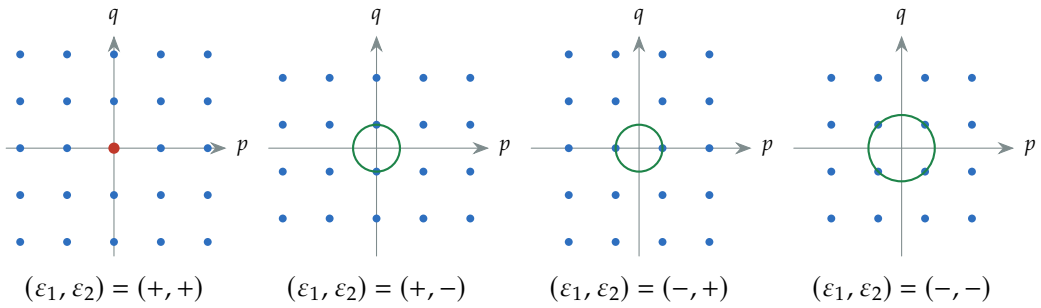


Figure T4.21: Allowed momenta (p, q) on the square torus for the four spin structures. Each dot gives the eigenvalues $\pm\sqrt{p^2+q^2}$. Only the periodic–periodic lattice contains the origin (red), which gives the two constant zero modes. For the other three, the lattice is shifted by half a step and the green circle shows the resulting gap $|\lambda| \geq \frac{1}{2}$ or $\frac{1}{\sqrt{2}}$.

$\frac{1}{2}$ in p is the same as the shift of the allowed momenta from $\mathbb{Z}$ to $\mathbb{Z}+\frac{1}{2}$ (the minus sign is irrelevant since the lattice is symmetric).

T4.7.7 The round two-sphere

The sphere is the first example where the spin connection is not removable, and the answer is beautiful: the eigenvalues are $\pm1, \pm2, \pm3, \ldots$ with multiplicities $2, 4, 6, \ldots$. We derive the operator, find the lowest eigenspinors explicitly, check that they are smooth, and then obtain the whole spectrum by recognising an old friend from the hydrogen atom.

Worked example T4.46: the Dirac operator on S^2 in polar coordinates

With the frame of Example T4.39 and Ω of Example T4.41:

$$D \overset{(1)}{=} -i\Big[\sigma_1\nabla_{e_1} + \sigma_2\nabla_{e_2}\Big] \overset{(2)}{=} -i\Big[\sigma_1\partial_\theta + \sigma_2\frac{1}{\sin\theta}\Big(\partial_\phi - \frac{i}{2}\cos\theta\,\sigma_3\Big)\Big]$$
$$\overset{(3)}{=} -i\Big[\sigma_1\Big(\partial_\theta + \frac{1}{2}\cot\theta\Big) + \frac{\sigma_2}{\sin\theta}\,\partial_\phi\Big]. \tag{T4.77}$$

(1) (T4.72) with $\gamma_1 = \sigma_1$, $\gamma_2 = \sigma_2$.
(2) $\Omega(e_1) = \Omega(\partial_\theta) = 0$ and $\Omega(e_2) = \frac{1}{\sin\theta}\Omega(\partial_\phi) = -\frac{i}{2}\frac{\cos\theta}{\sin\theta}\sigma_3$.
(3) $\sigma_2\sigma_3 = i\sigma_1$, so $\sigma_2\big(-\frac{i}{2}\cot\theta\,\sigma_3\big) = -\frac{i}{2}\cot\theta\,(i\sigma_1) = \frac{1}{2}\cot\theta\,\sigma_1$.

The spin connection has become the term $\frac{1}{2}\cot\theta$. It has a simple meaning: $\partial_\theta + \frac{1}{2}\cot\theta = (\sin\theta)^{-1/2}\,\partial_\theta\,(\sin\theta)^{1/2}$, i.e. it makes ∂_θ antisymmetric with respect to the measure $\sin\theta\,\mathrm{d}\theta$ (the one-dimensional version of formal self-adjointness).
Because the frame rotates once around each pole, spinor components in this frame are *antiperiodic* in ϕ (Example T4.34), so we separate variables with half-integer m:

$$\psi = \begin{pmatrix} f(\theta) \\ g(\theta) \end{pmatrix} e^{im\phi}, \qquad m \in \mathbb{Z} + \tfrac{1}{2}.$$

Using $\sigma_1 = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, $\sigma_2 = \left(\begin{smallmatrix} 0 & -i \\ i & 0 \end{smallmatrix}\right)$ and $\partial_\phi \to im$,

$$D\begin{pmatrix} f \\ g \end{pmatrix} e^{im\phi} = -i\begin{pmatrix} \big(\partial_\theta + \frac{1}{2}\cot\theta + \frac{m}{\sin\theta}\big)g \\ \big(\partial_\theta + \frac{1}{2}\cot\theta - \frac{m}{\sin\theta}\big)f \end{pmatrix} e^{im\phi}. \tag{T4.78}$$

(The upper entry: σ_2 contributes $-i\cdot im/\sin\theta = m/\sin\theta$ acting on g; the lower entry: $+i\cdot im/\sin\theta = -m/\sin\theta$ acting on f.) The lower-left block is D^+, the upper-right is D^-, as in (T4.75).

Worked example T4.47: the lowest eigenspinors on S^2: $\lambda = \pm 1$

Take $m = \frac{1}{2}$ and try $f = \cos\frac{\theta}{2}$, $g = c\,\sin\frac{\theta}{2}$ with a constant c. We need two derivatives. Writing $\sin\theta = 2\sin\frac{\theta}{2}\cos\frac{\theta}{2}$:

$$\Big(\partial_\theta + \tfrac{1}{2}\cot\theta + \tfrac{1}{2\sin\theta}\Big)\sin\tfrac{\theta}{2} \overset{(1)}{=} \tfrac{1}{2}\cos\tfrac{\theta}{2} + \frac{\cos\theta+1}{2\sin\theta}\sin\tfrac{\theta}{2} \overset{(2)}{=} \tfrac{1}{2}\cos\tfrac{\theta}{2} + \frac{2\cos^2\frac{\theta}{2}}{4\cos\frac{\theta}{2}} = \cos\tfrac{\theta}{2},$$

$$\Big(\partial_\theta + \tfrac{1}{2}\cot\theta - \tfrac{1}{2\sin\theta}\Big)\cos\tfrac{\theta}{2} \overset{(3)}{=} -\tfrac{1}{2}\sin\tfrac{\theta}{2} + \frac{\cos\theta-1}{2\sin\theta}\cos\tfrac{\theta}{2} \overset{(4)}{=} -\tfrac{1}{2}\sin\tfrac{\theta}{2} - \frac{2\sin^2\frac{\theta}{2}}{4\sin\frac{\theta}{2}} = -\sin\tfrac{\theta}{2}.$$

(1) Differentiate, and combine $\frac{1}{2}\cot\theta + \frac{1}{2\sin\theta} = \frac{\cos\theta+1}{2\sin\theta}$.
(2) $1+\cos\theta = 2\cos^2\frac{\theta}{2}$ and $\sin\theta = 2\sin\frac{\theta}{2}\cos\frac{\theta}{2}$.
(3) Same steps with the opposite sign of the m term.
(4) $\cos\theta - 1 = -2\sin^2\frac{\theta}{2}$.

Insert into (T4.78): $D\psi = -i\big(c\cos\frac{\theta}{2},\ -\sin\frac{\theta}{2}\big)^{\mathsf T}e^{i\phi/2}$. For this to be $\lambda\psi = \lambda(\cos\frac{\theta}{2},\ c\sin\frac{\theta}{2})^{\mathsf T}e^{i\phi/2}$ we need $-ic = \lambda$ (upper components) and $i = \lambda c$ (lower components). The first gives $c = i\lambda$; inserting it into the second gives $i = i\lambda^2$, so $\lambda^2 = 1$ and $c = \pm i$:

$$\psi_\pm^{(1/2)} = \begin{pmatrix}\cos\frac{\theta}{2}\\ \pm i\sin\frac{\theta}{2}\end{pmatrix}e^{i\phi/2}, \qquad D\psi_\pm^{(1/2)} = \pm\psi_\pm^{(1/2)}. \tag{T4.79}$$

The same computation with $m = -\frac{1}{2}$ gives $\psi_\pm^{(-1/2)} = \big(\sin\frac{\theta}{2},\ \mp i\cos\frac{\theta}{2}\big)^{\mathsf T}e^{-i\phi/2}$, also with $D = \pm 1$. So $\lambda = +1$ and $\lambda = -1$ each occur (at least) twice.

Are these genuine smooth spinors? The polar frame is singular at the poles, so we must re-express ψ in a frame that is smooth there. Near the north pole, (e_1, e_2) is the polar frame of a small disc with polar angle ϕ, and by Example T4.34 components in a smooth (Cartesian-like) frame are $\psi_{\text{cart}} = a(\phi)\psi = \operatorname{diag}(e^{-i\phi/2}, e^{i\phi/2})\psi$:

$$\psi_{\text{cart}} = \begin{pmatrix}\cos\frac{\theta}{2}\\ \pm i\sin\frac{\theta}{2}\,e^{i\phi}\end{pmatrix} \approx \begin{pmatrix}1\\ \pm\frac{i}{2}\,\theta e^{i\phi}\end{pmatrix} = \begin{pmatrix}1\\ \pm\frac{i}{2}(x+iy)\end{pmatrix} \quad (\theta\to 0),$$

with local Cartesian coordinates $x + iy = \theta e^{i\phi}$: smooth. Near the south pole the oriented polar angle is $-\phi$ and $e_1 = \partial_\theta$ points *towards* the pole, so (e_1, e_2) is the smooth frame rotated by $\pi - \phi$; multiplying by $a(\pi-\phi) = \operatorname{diag}(-ie^{i\phi/2}, ie^{-i\phi/2})$ gives components $\big(-ie^{i\phi}\cos\frac{\theta}{2},\ \mp\sin\frac{\theta}{2}\big)$, which with $\theta' = \pi - \theta$ and $x' - iy' = \theta' e^{i\phi}$ behave as $\big(-\frac{i}{2}(x'-iy'),\ \mp 1\big)$: smooth again. The half-integer m was essential: an integer m would leave a factor $e^{\pm i\phi/2}$ that is discontinuous at the poles.

The whole spectrum: the spin–orbit operator of the hydrogen atom. The eigenvalues ± 1 look like the start of a pattern. To see it, embed $S^2 \subset \mathbb{R}^3$ and use the flat Dirac operator of $\mathbb{R}^3$, $-i\vec\sigma\cdot\nabla$. Recall the Pauli identity $(\vec\sigma\cdot\vec a)(\vec\sigma\cdot\vec b) = \vec a\cdot\vec b + i\vec\sigma\cdot(\vec a\times\vec b)$ (Example T4.2), valid as long as $\vec a$ stands to the left of the derivatives in $\vec b$. With $\hat r = \vec r/r$ and the orbital angular momentum

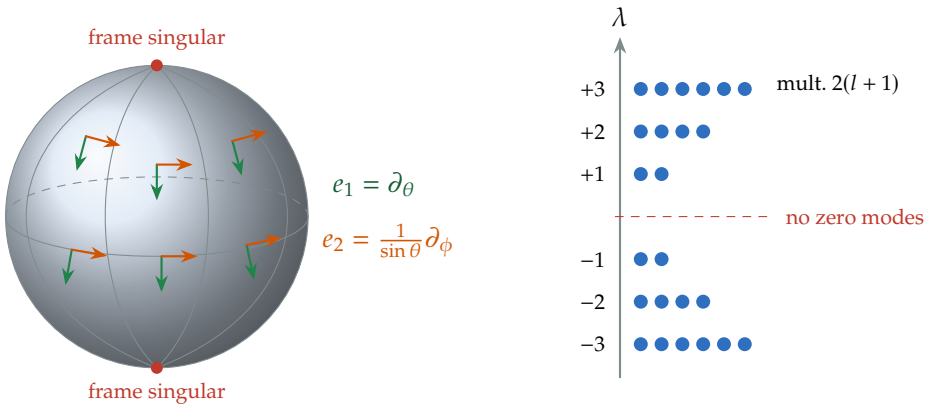


Figure T4.22: Left: the polar frame on S^2 (green along meridians, orange along parallels). It is smooth except at the poles, where it rotates once around each pole; hence spinor components in this frame have half-integer angular momentum m. Right: the spectrum of the Dirac operator on the unit sphere, $\lambda = \pm(l+1)$ with multiplicity $2(l+1)$ for each sign (dots count eigenspinors). It is symmetric and has a gap: $|\lambda| \geq 1$.

$\vec{L} = -i\,\vec{r}\times\nabla$:

$$(\vec{\sigma}\cdot\hat{r})(\vec{\sigma}\cdot\nabla) \overset{(1)}{=} \hat{r}\cdot\nabla + i\vec{\sigma}\cdot(\hat{r}\times\nabla) \overset{(2)}{=} \partial_r + \frac{i}{r}\,\vec{\sigma}\cdot(i\vec{L}) = \partial_r - \frac{\vec{\sigma}\cdot\vec{L}}{r},$$
$$-i\vec{\sigma}\cdot\nabla \overset{(3)}{=} -i(\vec{\sigma}\cdot\hat{r})\Big(\partial_r - \frac{\vec{\sigma}\cdot\vec{L}}{r}\Big) = -i(\vec{\sigma}\cdot\hat{r})\Big(\partial_r + \frac{1}{r}\Big) + \frac{i(\vec{\sigma}\cdot\hat{r})}{r}\,K, \qquad K \coloneqq \vec{\sigma}\cdot\vec{L} + 1. \tag{T4.80}$$

(1) Pauli identity with $\vec{a} = \hat{r}$, $\vec{b} = \nabla$.
(2) $\hat{r}\cdot\nabla = \partial_r$, and $\vec{r}\times\nabla = i\vec{L}$.
(3) Multiply on the left by $\vec{\sigma}\cdot\hat{r}$, which squares to 1; add and subtract $1/r$.

The first term differentiates in the normal direction (the $1/r$ is the mean curvature term that makes ∂_r antisymmetric with respect to $r^2\mathrm{d}r$). The second term involves only derivatives along the sphere. The general hypersurface formula of Bär's §2.6 identifies it: *the intrinsic Dirac operator of the unit S^2, with its unique spin structure induced from $\mathbb{R}^3$, is unitarily equivalent to* $K = \vec{\sigma}\cdot\vec{L} + 1$, acting on two-component functions on the sphere. (Clifford multiplication by the normal, $\vec{\sigma}\cdot\hat{r}$, anticommutes with K and exchanges its $\pm$ eigenspaces, which is the same pairing as Γ in Section T4.7.3.) We accept this identification and check it against what we computed. The eigenvalues of K follow from angular momentum addition. With $\vec{J} = \vec{L} + \frac{1}{2}\vec{\sigma}$,

$$\vec{J}^2 \overset{(1)}{=} \vec{L}^2 + \vec{\sigma}\cdot\vec{L} + \tfrac{3}{4} \quad\Longrightarrow\quad K = j(j+1) - l(l+1) + \tfrac{1}{4},$$
$$j = l + \tfrac{1}{2}:\ K \overset{(2)}{=} l + 1 = j + \tfrac{1}{2}, \qquad j = l - \tfrac{1}{2}:\ K \overset{(3)}{=} -l = -(j + \tfrac{1}{2}).$$

(1) $(\vec{L} + \frac{1}{2}\vec{\sigma})^2 = \vec{L}^2 + \vec{L}\cdot\vec{\sigma} + \frac{1}{4}\vec{\sigma}^2$ (the components of $\vec{L}$ and $\vec{\sigma}$ commute) and $\vec{\sigma}^2 = 3$.
(2) $(l+\frac{1}{2})(l+\frac{3}{2}) - l(l+1) + \frac{1}{4} = l + 1$.
(3) $(l-\frac{1}{2})(l+\frac{1}{2}) - l(l+1) + \frac{1}{4} = -l$.

For each $j = \frac{1}{2}, \frac{3}{2}, \ldots$ the multiplet has $2j+1$ states, and $K = +(j+\frac{1}{2})$ (from $l = j - \frac{1}{2}$) and $K = -(j+\frac{1}{2})$ (from $l = j + \frac{1}{2}$) each occur once per state. Writing $j + \frac{1}{2} = l + 1$:

$$\operatorname{spec} D_{S^2} = \{\pm(l+1) :\ l = 0, 1, 2, \ldots\}, \qquad \text{multiplicity } 2(l+1) \text{ for each sign.} \tag{T4.81}$$

For a sphere of radius r every eigenvalue is divided by r (D has dimensions of inverse length).

Checks. (i) $l = 0$: $\lambda = \pm 1$ with multiplicity 2 each: the four eigenspinors (T4.79). (In the $\mathbb{R}^3$ picture, $K = +1$ is the constant spinors χ, which have $\vec{L}\chi = 0$, and $K = -1$ is $(\vec{\sigma}\cdot\hat{r})\chi$.) (ii) The m values: for the multiplet j the angular momentum J_3 takes the $2j+1$ half-integer values $-j, \dots, j$, which matches the half-integer m we were forced to use in the polar frame; for $j = \frac{1}{2}$ these are the two values $m = \pm\frac{1}{2}$ we found. (iii) There is no zero mode, consistent with the Lichnerowicz argument of Section T4.8 ($R = 2 > 0$). (iv) The spectrum is symmetric, as it must be in even dimension: Γ anticommutes with D and maps the λ-eigenspace onto the $(-\lambda)$-eigenspace (in the $\mathbb{R}^3$ picture this role is played by $\vec{\sigma}\cdot\hat{r}$).

Dictionary: the sphere spectrum in physics

- **Relativistic hydrogen.** K is Dirac's spin–orbit operator; its eigenvalue $-\kappa = \pm(j+\frac{1}{2})$ is the quantum number that labels the $s_{1/2}, p_{1/2}, p_{3/2}, \dots$ levels. The Dirac operator on the unit sphere *is* the angular part of the Dirac Hamiltonian.
- **Radial quantisation.** On $\mathbb{R}\times S^2$ (or on S^2 as the celestial sphere) free massless fermions have energies $\pm(l+1)/r$: the conformal dimensions $l+1 = \frac{3}{2} - \frac{1}{2} + l$ of the free-fermion operators $\psi, \partial\psi, \dots$ shifted by the curvature of the sphere.
- **The gap** $|\lambda| \geq 1$ is a curvature effect: on a flat torus of the same area there can be zero modes. Positive curvature pushes spinor eigenvalues away from zero. The next section explains why.

Reading Nash (p. 113–115)

Nash defines the Dirac operator in one line, (4.100), and moves directly to its index. Everything in this section is background he assumes: the spin connection ω ("the Riemannian spin connection on X") and its curvature, ellipticity of the symbol $i\gamma^a e_a^\mu p_\mu$, self-adjointness, and the chiral splitting (4.101)–(4.102). Nash never computes a Dirac spectrum; the examples S^1, T^2 and S^2 are there so that the index $n_+ - n_-$ of (4.106) has concrete meaning: on all of them it is 0, for different reasons (no zero modes at all, or zero modes that come in chiral pairs).

Recap

- The spin connection is $\nabla = \mathrm{d} + \frac{1}{4}\omega_{ab}\gamma^a\gamma^b$: each infinitesimal frame rotation ω is replaced by its spin generator. It is forced by the Leibniz rule for Clifford multiplication, it encodes the half-angle, and it glues correctly only because the transition functions are lifted to $\mathrm{Spin}(n)$.
- Its curvature is $R^S = \frac{1}{4}R_{ab}\gamma^a\gamma^b$; on S^2 the chiral halves are line bundles with $c_1 = \mp 1$.
- The Dirac operator $D = -i\gamma^a\nabla_a$ is elliptic (symbol $\not{\xi}$), formally self-adjoint (divergence of the current $\psi^\dagger\gamma^a\chi$), and anticommutes with Γ, so $D = \left(\begin{smallmatrix} 0 & D^- \\ D^+ & 0 \end{smallmatrix}\right)$.
- Spectra: $\mathbb{R}^n$: $\pm|k|$. S^1: $\frac{2\pi}{L}\mathbb{Z}$ (R, one zero mode) or $\frac{2\pi}{L}(\mathbb{Z}+\frac{1}{2})$ (NS, none). T^2: $\pm\sqrt{p^2+q^2}$ with zero modes only for $(+,+)$. S^2: $\pm(l+1)$ with multiplicity $2(l+1)$, no zero modes.
- In two dimensions $D^+ = -2i\partial_{\bar{z}}$: chiral zero modes are holomorphic.

T4.8 The Lichnerowicz formula, twisted Dirac operators and the index

In Section T4.7 we computed three spectra by hand and noticed a pattern. On the flat torus with periodic boundary conditions there were zero modes: the constant spinors. On the round sphere there were none, and the eigenvalues kept a respectful distance $|\lambda| \geq 1$ from zero. Is this an accident of the examples, or is curvature *forcing* the spinors away from zero?

It is forced, and the mechanism is one identity: the square of the Dirac operator is a positive "kinetic" term plus a quarter of the scalar curvature,

$$D^2 = \nabla^*\nabla + \frac{R}{4}.$$

This is the **Lichnerowicz formula**. It is the spinor analogue of the statement that the Pauli Hamiltonian $(\sigma \cdot \vec{\pi})^2$ is the kinetic energy $\vec{\pi}^2$ plus the magnetic moment term $-\sigma \cdot \vec{B}$: squaring a first-order operator produces a second-order kinetic term *plus a curvature term*, and the curvature term is where all the physics sits. With a gauge field in addition, the curvature term acquires exactly the Pauli coupling $-\frac{1}{2}\gamma^a\gamma^b F_{ab}$.

The questions of this section are:

- What is $\nabla^*\nabla$, and why can every Laplace-type operator be written as $\nabla^*\nabla$ plus a potential?
- Why is the potential for D^2 exactly $R/4$? (A two-line computation with the Clifford relation and the Bianchi identity.)
- What does it imply? (No harmonic spinors when $R > 0$; lower bounds on $|\lambda|$; the bound is attained on round spheres.)
- What happens when spinors are also charged? (Twisted Dirac operators, Landau levels, the Aharonov–Casher zero modes, monopoles on S^2.)
- Why is $\dim \operatorname{Ker} D^+ - \dim \operatorname{Ker} D^-$ so robust, and what formula computes it? (A preview of the index theorem of Nash's Chapter 4.)

Throughout, M is a Riemannian spin manifold of dimension n, γ_a are our Hermitian gamma matrices with $\{\gamma_a, \gamma_b\} = 2\delta_{ab}$, and $D = -i\gamma^a\nabla_a$ with the spin connection of Section T4.7.2.

T4.8.1 The connection Laplacian

On functions on $\mathbb{R}^n$ the Laplacian is $-\sum_a \partial_a^2$, and integrating by parts gives $\int \bar{f}(-\partial_a^2 f) = \int |\partial f|^2 \geq 0$. For sections of any bundle with a connection there is a direct analogue.

Working definition T4.11 (connection Laplacian)

Let $E \to M$ be a bundle with a Hermitian metric [an inner product $\langle\psi, \chi\rangle$ on the components at each point] and a metric connection ∇ [one that obeys $\partial_X\langle\psi, \chi\rangle = \langle\nabla_X\psi, \chi\rangle + \langle\psi, \nabla_X\chi\rangle$] (for us: spinors with the spin connection, or spinors tensored with a gauge bundle). The **connection Laplacian** (or **rough Laplacian**) is the operator $\nabla^*\nabla$ defined by

$$(\nabla^*\nabla\psi, \chi) = (\nabla\psi, \nabla\chi) := \int_M \sum_a \langle\nabla_a\psi, \nabla_a\chi\rangle \,\mathrm{vol}$$

for all compactly supported χ [χ vanishes outside a bounded region, so boundary terms drop]. *Recipe* [Bär Lemma 2.5.10, (2.33)]: in a local orthonormal frame

$$\nabla^*\nabla = -\sum_{a=1}^{n}\Big(\nabla_{e_a}\nabla_{e_a} - \nabla_{\nabla_{e_a}e_a}\Big), \tag{T4.82}$$

and at a point where the frame is synchronous ($\nabla e_a = 0$ there) it is simply $-\sum_a \nabla_a\nabla_a$.

The second term in (T4.82) is there because the frame vectors themselves turn. Let us see why the recipe is right. The argument is the same divergence trick we used for self-adjointness of D in (T4.73). Define the complex vector field X by $\langle X, Y\rangle = \langle\psi, \nabla_Y\chi\rangle$ for all vectors Y, i.e. $X^a = \langle\psi, \nabla_a\chi\rangle$. At a point x with a synchronous frame,

$$\begin{aligned}\operatorname{div} X &\overset{(1)}{=} \sum_a e_a\big(\langle\psi, \nabla_a\chi\rangle\big) \overset{(2)}{=} \sum_a\Big(\langle\nabla_a\psi, \nabla_a\chi\rangle + \langle\psi, \nabla_a\nabla_a\chi\rangle\Big)\\ &\overset{(3)}{=} \langle\nabla\psi, \nabla\chi\rangle - \big\langle\psi, \nabla^*\nabla\chi\big\rangle.\end{aligned}$$

(1) Divergence in an orthonormal frame; the term $\langle X, \nabla_{e_a}e_a\rangle$ vanishes at x.
(2) The connection is metric: $e_a\langle\psi,\phi\rangle = \langle\nabla_a\psi,\phi\rangle + \langle\psi,\nabla_a\phi\rangle$.
(3) At x, (T4.82) reduces to $-\sum_a \nabla_a\nabla_a$.

Both ends of the chain are frame-independent, so the identity holds everywhere. Integrating over a closed M kills the divergence (Stokes), and what remains is $(\nabla\psi, \nabla\chi) = (\psi, \nabla^*\nabla\chi)$: the recipe computes the adjoint. Setting $\chi = \psi$ gives the one property we will use again and again:

$$(\nabla^*\nabla\psi, \psi) = \|\nabla\psi\|^2 \geq 0, \qquad \text{with equality iff } \nabla\psi = 0. \tag{T4.83}$$

So $\nabla^*\nabla$ is a *non-negative* operator whose kernel is the parallel sections. It is the "kinetic energy" of the field.

Worked example T4.48: the rough Laplacian of functions in polar coordinates

On $\mathbb{R}^2$ with the polar frame $e_1 = \partial_r$, $e_2 = \frac{1}{r}\partial_\phi$ of Example T4.40, $\omega_{12} = -\mathrm{d}\phi$. By (T4.66), $\nabla_X e_2 = e_1\,\omega_{12}(X) = -\mathrm{d}\phi(X)\,e_1$ and $\nabla_X e_1 = e_2\,\omega_{21}(X) = \mathrm{d}\phi(X)\,e_2$. Hence

$$\nabla_{e_1}e_1 = \mathrm{d}\phi(\partial_r)\,e_2 = 0, \qquad \nabla_{e_2}e_2 = -\mathrm{d}\phi\big(\tfrac{1}{r}\partial_\phi\big)\,e_1 = -\tfrac{1}{r}\,e_1.$$

For a function f (trivial bundle, $\nabla = \mathrm{d}$):

$$\begin{aligned}\nabla^*\nabla f &\overset{(1)}{=} -\Big(e_1e_1f - (\nabla_{e_1}e_1)f + e_2e_2f - (\nabla_{e_2}e_2)f\Big)\\ &\overset{(2)}{=} -\Big(\partial_r^2 f + \frac{1}{r^2}\partial_\phi^2 f + \frac{1}{r}\partial_r f\Big).\end{aligned}$$

(1) (T4.82) with $n = 2$.
(2) $e_1e_1f = \partial_r^2 f$, $e_2e_2f = r^{-2}\partial_\phi^2 f$, $-(\nabla_{e_2}e_2)f = +\frac{1}{r}\partial_r f$.

This is the familiar polar Laplacian. The $\frac{1}{r}\partial_r$ term, which students usually memorise, is the correction for the turning frame.

Every Laplace-type operator is a rough Laplacian plus a potential. Bär proves a small but important structural fact [Bär Lemma 1.3.5]: if P is a formally self-adjoint operator of Laplace

type on E (symbol $|\xi|^2\mathbf{1}$), there is *exactly one* metric connection ∇ on E and one Hermitian endomorphism field K with

$$P = \nabla^*\nabla + K. \tag{T4.84}$$

The mechanism is simple. The second-order part of P is fixed by the symbol. The first-order part $B^a\partial_a$ can be absorbed by choosing the connection $\nabla_a = \partial_a - \frac{1}{2}B^a + \dots$, exactly as one completes the square $x^2 + bx = (x + \frac{b}{2})^2 - \frac{b^2}{4}$. What is left over has order zero: it is K. Any formula of the type (T4.84) with an explicit K is called a **Weitzenböck formula**. The Lichnerowicz formula identifies the connection and the potential for $P = D^2$: the connection is the spin connection and the potential is $R/4$.

Physical picture: completing the square

In quantum mechanics $H = (\vec{p} - \vec{a})^2 + V$ is "kinetic energy with a gauge field, plus a potential". (T4.84) says every Laplace-type operator can be put in this form in exactly one way. The gauge field is the connection ∇, and the potential K is what the operator "feels" at each point beyond kinetic energy. For D^2, the potential is curvature.

T4.8.2 Deriving the Lichnerowicz formula

Key idea: Lichnerowicz formula [Bär Thm. 2.5.11]

On any Riemannian spin manifold, with R the scalar curvature,

$$D^2 = \nabla^*\nabla + \frac{R}{4}. \tag{T4.85}$$

The derivation has three steps. Each is short, and each uses one idea.

Step 1: square D and split into symmetric and antisymmetric parts. Work at a point x in a synchronous frame, so that $\nabla e_a = 0$ at x. Then $\gamma(e_b) = \gamma_b$ is covariantly constant at x by the Leibniz rule (T4.70), and $[e_a, e_b] = \nabla_{e_a}e_b - \nabla_{e_b}e_a = 0$ at x, so the commutator of covariant derivatives is pure curvature: $[\nabla_a, \nabla_b] = R^S(e_a, e_b) =: R^S_{ab}$. Now

$$\begin{aligned} D^2\psi &\overset{(1)}{=} (-i\gamma^a\nabla_a)(-i\gamma^b\nabla_b\psi) \overset{(2)}{=} -\gamma^a\gamma^b\,\nabla_a\nabla_b\psi \\ &\overset{(3)}{=} -\sum_a \nabla_a\nabla_a\psi - \sum_{a\neq b}\gamma^a\gamma^b\,\nabla_a\nabla_b\psi \\ &\overset{(4)}{=} -\sum_a \nabla_a\nabla_a\psi - \frac{1}{2}\sum_{a\neq b}\gamma^a\gamma^b\,[\nabla_a, \nabla_b]\psi \\ &\overset{(5)}{=} \nabla^*\nabla\psi - \frac{1}{2}\,\gamma^a\gamma^b\,R^S_{ab}\,\psi. \end{aligned} \tag{T4.86}$$

(1) Definition (T4.72).
(2) $(-i)^2 = -1$, and ∇_a passes through γ^b because γ^b is parallel at x.
(3) Separate $a = b$ (where $\gamma^a\gamma^a = \mathbf{1}$) from $a \neq b$.
(4) For $a \neq b$, $\gamma^b\gamma^a = -\gamma^a\gamma^b$, so the sum over ordered pairs only sees the antisymmetric part of $\nabla_a\nabla_b$: $\sum_{a\neq b}\gamma^a\gamma^b\nabla_a\nabla_b = \frac{1}{2}\sum_{a\neq b}\gamma^a\gamma^b(\nabla_a\nabla_b - \nabla_b\nabla_a)$.
(5) (T4.82) at x; $[\nabla_a, \nabla_b] = R^S_{ab}$ at x, and the $a = b$ terms may be added back since $R^S_{aa} = 0$.

This step is pure Clifford algebra and holds for *any* connection compatible with Clifford multiplication. It is the curved-space version of the Pauli computation $(\sigma\cdot\vec{\pi})^2 = \vec{\pi}^{\,2} + i\sigma_3[\pi_1, \pi_2]$.

Step 2: insert the spinor curvature. By (T4.71), $R^S_{ab} = R^S(e_a, e_b) = \frac{1}{4} R_{cdab}\gamma^c\gamma^d$. So

$$D^2 = \nabla^*\nabla - \frac{1}{8}\, R_{cdab}\, \gamma^a\gamma^b\gamma^c\gamma^d = \nabla^*\nabla - \frac{1}{8}\, R_{abcd}\, \gamma^a\gamma^b\gamma^c\gamma^d, \tag{T4.87}$$

using the pair symmetry $R_{cdab} = R_{abcd}$ of the Riemann tensor.

Step 3: a four-gamma identity. We need $X := R_{abcd}\gamma^a\gamma^b\gamma^c\gamma^d$. Hold a fixed and look at the last three gammas. The gamma-matrix version of Dray's triple product (T4.16) is

$$\gamma^b\gamma^c\gamma^d = \gamma^{[bcd]} + \delta^{bc}\gamma^d - \delta^{bd}\gamma^c + \delta^{cd}\gamma^b, \tag{T4.88}$$

where $\gamma^{[bcd]}$ is the totally antisymmetric part (equal to $\gamma^b\gamma^c\gamma^d$ when b, c, d are distinct and zero otherwise). (Check one case: $b = d = 1$, $c = 2$ gives $\gamma^1\gamma^2\gamma^1 = -\gamma^2$ on the left and $0 + 0 - \gamma^2 + 0$ on the right.) Contract with R_{abcd}, using the symmetries $R_{abcd} = -R_{bacd} = -R_{abdc} = R_{cdab}$, the first Bianchi identity $R_{abcd} + R_{acdb} + R_{adbc} = 0$, and $\mathrm{Ric}_{bd} = \sum_a R_{abad}$ (Section T4.7.1):

$$\begin{aligned} R_{abcd}\gamma^b\gamma^c\gamma^d &\overset{(1)}{=} R_{abcd}\gamma^{[bcd]} + R_{abbd}\gamma^d - R_{abcb}\gamma^c + R_{abcc}\gamma^b \\ &\overset{(2)}{=} 0 + R_{abbd}\gamma^d - R_{abcb}\gamma^c + 0 \\ &\overset{(3)}{=} -\mathrm{Ric}_{ad}\gamma^d - \mathrm{Ric}_{ac}\gamma^c = -2\,\mathrm{Ric}_{ad}\gamma^d. \end{aligned} \tag{T4.89}$$

(1) Insert (T4.88) and do the δ contractions (repeated indices summed).
(2) The antisymmetrisation of R_{abcd} over b, c, d is one third of the cyclic sum in the Bianchi identity, so it vanishes. $R_{abcc} = 0$ by antisymmetry in the last pair.
(3) $\sum_b R_{abbd} = -\sum_b R_{babd} = -\mathrm{Ric}_{ad}$, and $\sum_b R_{abcb} = \sum_b R_{babc} = \mathrm{Ric}_{ac}$ (swap both pairs' entries).

Equation (T4.89) is Bär's Lemma 2.5.9 in our language ($\sum_j \gamma^j R^S(e_j, X) = -\frac{1}{2}\gamma(\mathrm{Ric}\, X)$ up to his sign conventions). Now multiply by γ^a and sum:

$$X = \gamma^a\left(R_{abcd}\gamma^b\gamma^c\gamma^d\right) \overset{(1)}{=} -2\,\mathrm{Ric}_{ad}\,\gamma^a\gamma^d \overset{(2)}{=} -2\,\mathrm{Ric}_{ad}\,\tfrac{1}{2}\{\gamma^a, \gamma^d\} \overset{(3)}{=} -2\,\mathrm{Ric}_{aa} = -2R.$$

(1) (T4.89).
(2) Ric is symmetric, so only the symmetric part of $\gamma^a\gamma^d$ survives.
(3) Clifford relation; the trace of the Ricci tensor is the scalar curvature.

Insert into (T4.87): $D^2 = \nabla^*\nabla - \frac{1}{8}(-2R) = \nabla^*\nabla + \frac{R}{4}$. That is (T4.85).

Worked example T4.49: checking the four-gamma identity on a round sphere

The unit sphere S^n has $R_{abcd} = \delta_{ac}\delta_{bd} - \delta_{ad}\delta_{bc}$ (sectional curvature 1). Then, with $\gamma^b\gamma^b = n$ and $\gamma^a\gamma^b\gamma^a = (2-n)\gamma^b$ (anticommute γ^b past γ^a for the $n-1$ values $a \neq b$),

$$X \overset{(1)}{=} \gamma^a\gamma^b\gamma^a\gamma^b - \gamma^a\gamma^b\gamma^b\gamma^a \overset{(2)}{=} (2-n)\gamma^b\gamma^b - n\,\gamma^a\gamma^a \overset{(3)}{=} (2-n)n - n^2 = -2n(n-1).$$

(1) Insert R_{abcd}; the two δ terms give the two orderings.
(2) $\gamma^a\gamma^b\gamma^a = (2-n)\gamma^b$ in the first term, $\gamma^b\gamma^b = n$ in the second.
(3) $\gamma^b\gamma^b = n$ again.

Since $R = n(n-1)$ for the unit sphere, $X = -2R$ as claimed. So on S^n, $D^2 = \nabla^*\nabla + \frac{n(n-1)}{4}$; on the unit S^2, $D^2 = \nabla^*\nabla + \frac{1}{2}$.

Worked example T4.50: the S^2 ground state: splitting $\lambda^2 = 1$ into kinetic and curvature parts

Take the eigenspinor $\psi = \left(\cos\frac{\theta}{2},\ i\sin\frac{\theta}{2}\right)^{\mathsf{T}} e^{i\phi/2}$ of (T4.79) with $\lambda = +1$, and compute its covariant derivatives with the spin connection $\Omega = -\frac{i}{2}\cos\theta\,\sigma_3\,\mathrm{d}\phi$ of Example T4.41. Along $e_1 = \partial_\theta$, $\Omega(e_1) = 0$:

$$\nabla_{e_1}\psi \overset{(1)}{=} \left(-\tfrac{1}{2}\sin\tfrac{\theta}{2},\ \tfrac{i}{2}\cos\tfrac{\theta}{2}\right)^{\mathsf{T}} e^{i\phi/2} \overset{(2)}{=} \tfrac{i}{2}\,\gamma_1\psi.$$

(1) Differentiate the components.

(2) $\gamma_1 = \sigma_1$ swaps the components: $\frac{i}{2}\sigma_1\psi = \frac{i}{2}\left(i\sin\frac{\theta}{2}, \cos\frac{\theta}{2}\right)^{\mathsf{T}} e^{i\phi/2}$.

Along $e_2 = \frac{1}{\sin\theta}\partial_\phi$, $\partial_\phi \to \frac{i}{2}$ and $\Omega(\partial_\phi) = -\frac{i}{2}\cos\theta\,\sigma_3$:

$$\nabla_{e_2}\psi \overset{(1)}{=} \frac{1}{\sin\theta}\left[\frac{i}{2}\begin{pmatrix}\cos\frac{\theta}{2}\\ i\sin\frac{\theta}{2}\end{pmatrix} - \frac{i}{2}\cos\theta\begin{pmatrix}\cos\frac{\theta}{2}\\ -i\sin\frac{\theta}{2}\end{pmatrix}\right]e^{i\phi/2} \overset{(2)}{=} \frac{1}{\sin\theta}\begin{pmatrix}\frac{i}{2}(1-\cos\theta)\cos\frac{\theta}{2}\\ -\frac{1}{2}(1+\cos\theta)\sin\frac{\theta}{2}\end{pmatrix}e^{i\phi/2}$$
$$\overset{(3)}{=} \begin{pmatrix}\frac{i}{2}\sin\frac{\theta}{2}\\ -\frac{1}{2}\cos\frac{\theta}{2}\end{pmatrix}e^{i\phi/2} \overset{(4)}{=} \tfrac{i}{2}\,\gamma_2\psi.$$

(1) $\nabla_{e_2} = \frac{1}{\sin\theta}(\partial_\phi + \Omega(\partial_\phi))$ and $\sigma_3 = \mathrm{diag}(1,-1)$.

(2) Collect terms; $\frac{i}{2}\cdot i = -\frac{1}{2}$.

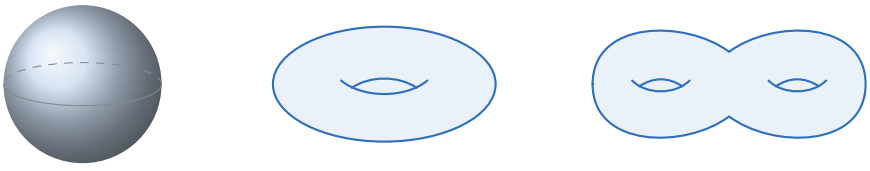
(3) $1-\cos\theta = 2\sin^2\frac{\theta}{2}$, $1+\cos\theta = 2\cos^2\frac{\theta}{2}$, $\sin\theta = 2\sin\frac{\theta}{2}\cos\frac{\theta}{2}$.

(4) $\sigma_2(a,b)^{\mathsf{T}} = (-ib, ia)^{\mathsf{T}}$, so $\frac{i}{2}\sigma_2\psi = \frac{i}{2}\left(\sin\frac{\theta}{2},\ i\cos\frac{\theta}{2}\right)^{\mathsf{T}} e^{i\phi/2}$.

So $\nabla_X\psi = \frac{i}{2}\gamma(X)\psi$ for every X: ψ is a **Killing spinor**. Consistency: $D\psi = -i\gamma^a\cdot\frac{i}{2}\gamma_a\psi = \frac{1}{2}\gamma^a\gamma_a\psi = \frac{n}{2}\psi = \psi$. ✓ Its kinetic energy density is $|\nabla\psi|^2 = \sum_a |\frac{i}{2}\gamma_a\psi|^2 = 2\cdot\frac{1}{4}|\psi|^2 = \frac{1}{2}|\psi|^2$ (each γ_a is unitary). Hence

$$\lambda^2\|\psi\|^2 = \|D\psi\|^2 = \underbrace{\|\nabla\psi\|^2}_{\frac{1}{2}\|\psi\|^2} + \underbrace{\int \tfrac{R}{4}|\psi|^2}_{\frac{1}{2}\|\psi\|^2} = \|\psi\|^2,$$

and the Lichnerowicz formula is verified on a nontrivial eigenspinor: half of $\lambda^2 = 1$ is kinetic, half is curvature.



S^2: $R = 2 > 0$
$D^2 \geq \frac{R}{4} > 0$
no zero modes

T^2: $R = 0$
$D^2 = \nabla^*\nabla$
zero modes = parallel spinors

genus 2: $R = -2 < 0$
no constraint from R
zero modes may exist

Figure T4.23: What the Lichnerowicz formula $D^2 = \nabla^*\nabla + R/4$ says for surfaces of constant curvature. For positive R both terms are non-negative and the curvature term strictly positive, so D cannot vanish on a nonzero spinor. For $R = 0$ a zero mode must be parallel ($\nabla\psi = 0$): the constant spinors of the $(+,+)$ torus. For negative R the curvature term pulls the other way and the formula gives no obstruction; whether harmonic spinors exist then depends on the conformal structure and the spin structure.

Watch out: which Laplacian?

D^2 is *not* the rough Laplacian $\nabla^*\nabla$, and neither of them is the Hodge Laplacian of forms. Three different "Laplacians" act on three different bundles, and each Weitzenböck formula relates two of them: $D^2 = \nabla^*\nabla + \frac{R}{4}$ on spinors, $\Delta = \nabla^*\nabla + \mathrm{Ric}$ on 1-forms (Section T4.9.5), and $D_A^2 = \nabla^*\nabla + \frac{R}{4} - \gamma(F)$ on charged spinors (Section T4.8.4). On flat space with no gauge field all of them reduce to $-\sum \partial_a^2$ acting componentwise.

T4.8.3 Consequences: harmonic spinors and positive curvature

A **harmonic spinor** is a solution of $D\psi = 0$, a zero mode. The Lichnerowicz formula gives two quick results, both obtained by pairing (T4.85) with ψ and integrating.

Key idea: no harmonic spinors in positive scalar curvature [Bär Cor. 2.5.12–2.5.13]

Let M be closed and connected.

1. If $R \geq 0$, every harmonic spinor is parallel, $\nabla\psi = 0$. Hence $\dim \ker D \leq 2^{\lfloor n/2 \rfloor}$, the rank of the spinor bundle.
2. If moreover $R > 0$ somewhere, there are no harmonic spinors at all.

Here is the whole argument. Let $D\psi = 0$. Then

$$0 \overset{(1)}{=} \|D\psi\|^2 \overset{(2)}{=} (D^2\psi, \psi) \overset{(3)}{=} (\nabla^*\nabla\psi, \psi) + \int_M \frac{R}{4}|\psi|^2 \,\mathrm{vol} \overset{(4)}{=} \|\nabla\psi\|^2 + \int_M \frac{R}{4}|\psi|^2 \,\mathrm{vol}. \tag{T4.90}$$

(1) $D\psi = 0$.
(2) D is formally self-adjoint (Section T4.7.3).
(3) Lichnerowicz (T4.85).
(4) (T4.83).

If $R \geq 0$ both terms on the right are ≥ 0 and they add to zero, so each vanishes: $\nabla\psi = 0$. A parallel section is determined by its value at one point (solve the transport equation along paths from that point), which gives the dimension bound. For part 2: $\nabla\psi = 0$ implies $\mathrm{d}|\psi|^2 = 2\,\mathrm{Re}\langle\nabla\psi, \psi\rangle = 0$, so $|\psi|$ is a constant c, and the second term is $\frac{c^2}{4}\int_M R \,\mathrm{vol}$. It vanishes only if $c = 0$, because $\int_M R > 0$ when $R \geq 0$ everywhere and $R > 0$ on some open set.

Checks. The $(+,+)$ torus has $R = 0$ and two harmonic spinors, the constants: parallel, and $2 = 2^{\lfloor 2/2 \rfloor}$ saturates the bound. The other three torus spin structures have no parallel spinors (a constant is not antiperiodic), hence no zero modes, as found in Example T4.45. The round S^2 has $R = 2$ and no zero modes, as found in (T4.81).

A spectral gap. The same computation with $D\psi = \lambda\psi$ and $R \geq s_0 > 0$ gives [Bär (2.36)]

$$\lambda^2\|\psi\|^2 = \|\nabla\psi\|^2 + \int_M \tfrac{R}{4}|\psi|^2 \;\geq\; \tfrac{s_0}{4}\|\psi\|^2 \quad\Longrightarrow\quad |\lambda| \geq \tfrac{1}{2}\sqrt{s_0}.$$

On the unit S^2 this says $|\lambda| \geq 1/\sqrt{2} \approx 0.71$, but the true answer is $|\lambda| \geq 1$. The estimate throws away $\|\nabla\psi\|^2$ entirely, and we saw in Example T4.50 that half of λ^2 is kinetic. A better estimate keeps part of it.

Friedrich's inequality. The trick is to bound the kinetic term from below *by D itself*. At each point, with $\psi_a := \nabla_a \psi$,

$$|D\psi|^2 = \Big|\sum_a \gamma^a \psi_a\Big|^2 \overset{(1)}{\leq} \Big(\sum_a |\gamma^a \psi_a|\Big)^2 \overset{(2)}{=} \Big(\sum_a |\psi_a|\Big)^2 \overset{(3)}{\leq} n \sum_a |\psi_a|^2 = n\,|\nabla\psi|^2. \tag{T4.91}$$

(1) Triangle inequality (the $-i$ in D does not change norms).
(2) Each γ^a is Hermitian with square $\mathbf{1}$, hence unitary: $|\gamma^a \phi| = |\phi|$ [Bär (2.38)].
(3) Cauchy–Schwarz for the vectors $(1,\dots,1)$ and $(|\psi_1|,\dots,|\psi_n|)$ in $\mathbb{R}^n$.

Now let $D\psi = \lambda\psi$ and $R \geq s_0 > 0$:

$$\lambda^2 \|\psi\|^2 = \|D\psi\|^2 \overset{(1)}{=} \|\nabla\psi\|^2 + \int_M \tfrac{R}{4}|\psi|^2 \overset{(2)}{\geq} \tfrac{1}{n}\|D\psi\|^2 + \tfrac{s_0}{4}\|\psi\|^2 = \tfrac{\lambda^2}{n}\|\psi\|^2 + \tfrac{s_0}{4}\|\psi\|^2.$$

(1) (T4.90) without setting $D\psi = 0$.
(2) Integrate (T4.91), and $R \geq s_0$.

Rearranging gives **Friedrich's inequality** [Bär Thm. 2.5.15, (2.37)]:

$$\lambda^2 \;\geq\; \frac{n}{n-1}\,\frac{s_0}{4}. \tag{T4.92}$$

When is it sharp? Equality needs equality in both steps of (T4.91): all $\gamma^a\psi_a$ must be parallel vectors of the same length. The solutions are exactly the Killing spinors, $\nabla_X\psi = \mu\,\gamma(X)\psi$ with constant μ; then $D\psi = -i\mu\gamma^a\gamma_a\psi = -in\mu\,\psi$, so $\mu = i\lambda/n$. On the unit S^2 we found such a spinor in Example T4.50 with $\mu = \frac{i}{2}$, $\lambda = 1$, and indeed (T4.92) with $n = 2$, $s_0 = 2$ gives $\lambda^2 \geq 1$: *equality*. On the unit S^n, $s_0 = n(n-1)$ and Friedrich gives $|\lambda| \geq n/2$, which is exactly the smallest eigenvalue of the round sphere (the full spectrum is $\pm(\frac{n}{2}+k)$, $k = 0,1,2,\dots$, generalising $\pm(l+1)$ for $n = 2$) [Bär Rem. 2.5.16].

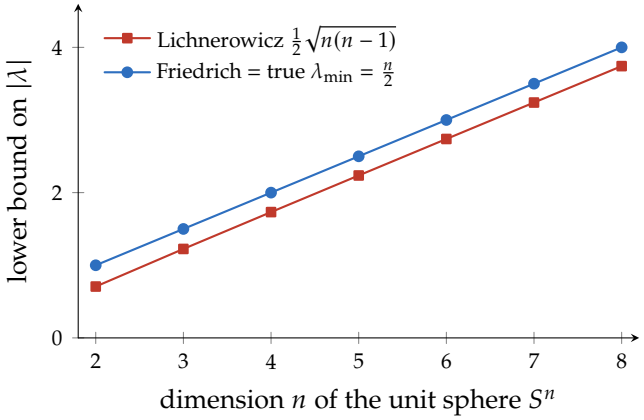


Figure T4.24: Two lower bounds for the smallest Dirac eigenvalue on the unit sphere S^n. The Lichnerowicz gap $\frac{1}{2}\sqrt{s_0}$ (red) discards all kinetic energy and is never attained. Friedrich's bound (blue) keeps a fraction $\frac{1}{n}$ of it and coincides with the true lowest eigenvalue $n/2$, attained by Killing spinors.

A bound with no curvature assumption. On a two-sphere something even stronger holds. Bär proved [Bär Thm. 2.5.17, (2.40)] that for *every* Riemannian metric on S^2,

$$\lambda^2 \;\geq\; \frac{4\pi}{\operatorname{area}(S^2, g)}. \tag{T4.93}$$

Check on the round sphere of radius r: $\lambda_{\min} = 1/r$ and the area is $4\pi r^2$, so $\lambda_{\min}^2 = 1/r^2 = 4\pi/(4\pi r^2)$, equality. Bär also shows that equality characterises constant curvature. Unlike Friedrich's bound this needs no sign of R; it reflects the topology (a two-sphere has a unique spin structure, and its spinor bundle is "twisted" by $c_1(S^\pm) = \mp 1$). By contrast, every closed orientable surface of genus ≥ 1, and S^3, admit metrics with harmonic spinors [Bär Rem. 2.5.18], so no such estimate can hold there. (Hitchin showed that in dimensions $n \equiv 0, 1, 7 \pmod 8$ there are metrics with harmonic spinors on many spin manifolds; this is where the mod-8 periodicity of Section T4.4.5 reappears in geometry.)

Dictionary: the Lichnerowicz term in physics

- **Fermion masses from curvature.** On $\mathbb{R} \times M$ a massless Dirac field has frequencies $\omega = \pm\lambda$, $\lambda \in \operatorname{spec} D_M$. Positive curvature gives a gap: fermions on a sphere of radius r cannot have energy below $1/r$. In a Kaluza–Klein compactification on a positively curved internal space there are therefore no massless fermions, a classic no-go argument (Witten, 1981) that pushes model builders towards Ricci-flat or gauge-twisted internal spaces.
- **Positive energy.** Witten's spinorial proof of the positive energy theorem in general relativity uses the Lichnerowicz formula on a spacelike slice: the ADM energy is written as $\|\nabla\psi\|^2$ plus $\frac{R}{4}|\psi|^2$, and the dominant energy condition makes the curvature term non-negative.

T4.8.4 Twisted Dirac operators

Physical fermions carry charges. An electron is a spinor *and* a section of a U(1) bundle; a quark is a spinor and a colour triplet. Mathematically the field lives in the tensor product $S \otimes E$ of the spinor bundle with a gauge bundle E (rank r, with a connection A). Locally ψ is a $2^{\lfloor n/2 \rfloor} \times r$ array: a spinor index and a gauge index.

Working definition T4.12 (twisted Dirac operator)

Let E carry a Hermitian metric and a unitary connection $\nabla^E = \mathrm{d} + A$ [parallel transport along any path preserves the inner product on the gauge index] with anti-Hermitian A_μ (house convention; for U(1) with charge 1, $A = -ia$ with the physicists' real potential a). On $S \otimes E$ use the product connection $\nabla_a = \nabla_a^S \otimes \mathbf{1} + \mathbf{1} \otimes \nabla_a^E$, i.e. in components

$$\nabla_\mu \psi = \partial_\mu \psi + \tfrac{1}{4}\omega_{ab\mu}\gamma^a\gamma^b\psi + A_\mu\psi,$$

with the gammas acting on the spinor index and A_μ on the gauge index. The **twisted Dirac operator** is

$$D_A\psi := -i\gamma^a\nabla_a\psi = -i\gamma^a e_a{}^\mu\big(\partial_\mu + \Omega_\mu + A_\mu\big)\psi. \tag{T4.94}$$

This is Bär's D^{∇^C} [Bär Def. 2.5.4, (2.26)] with $C = E$, and Nash's $\not{D}_F = \gamma^a e_a^\mu(\partial_\mu + \omega_\mu + A_\mu)$ [Nash p. 118].

Three properties carry over from D with no work.

1. *Same symbol.* A has no derivatives, so $\sigma_{D_A}(\xi) = \not{\xi} \otimes \mathbf{1}_E$ [Bär (1.34)]. Hence D_A is elliptic and of Dirac type [Bär Cor. 1.3.22]. The gauge field is "invisible to the symbol", which is why the index can depend on E only through topology.
2. *Self-adjoint.* If A is anti-Hermitian (a unitary connection), the proof of (T4.73) goes through

with $\langle\cdot,\cdot\rangle$ the metric of $S\otimes E$ [Bär Cor. 2.5.6].

3. *Chirality.* A commutes with Γ (it acts on the other index), so D_A still anticommutes with Γ and splits as $D_A^{\pm}:\Gamma(S^{\pm}\otimes E)\to\Gamma(S^{\mp}\otimes E)$. Nash's $\not{D}_F(1+\gamma_5)/2$ is D_A^+.

The square. Step 1 of the Lichnerowicz derivation used only Clifford algebra and $[\nabla_a,\nabla_b]=$ curvature. For the product connection the curvature is the sum, $[\nabla_a,\nabla_b]=R^S_{ab}\otimes\mathbf{1}+\mathbf{1}\otimes F_{ab}$, with $F=\mathrm{d}A+A\wedge A$ and $F_{ab}=F(e_a,e_b)$. So (T4.86) and Steps 2–3 give at once

$$D_A^2=\nabla^*\nabla+\frac{R}{4}-\gamma(F),\qquad \gamma(F):=\frac{1}{2}F_{ab}\,\gamma^a\gamma^b=\sum_{a<b}F_{ab}\,\gamma^a\gamma^b. \tag{T4.95}$$

Here $\gamma(F)$ is the Clifford action of the curvature 2-form: the gammas act on the spinor index, F_{ab} on the gauge index. This is Bär's formula $D^2=\nabla^*\nabla+\frac{\mathrm{scal}}{4}+\mathfrak{F}^E$ [Bär §5], with $\mathfrak{F}^E=\frac{1}{2}\sum e_ie_jF_{ij}=-\gamma(F)$ after $e_i\to -i\gamma_i$.

Dictionary: the Pauli term

With $A=-ia$, $F=-if$ ($f=\mathrm{d}a$ the physicists' field strength),

$$-\gamma(F)=\tfrac{i}{2}f_{ab}\gamma^a\gamma^b=-\tfrac{1}{2}f_{ab}\,\Sigma^{ab}_{\text{phys}},\qquad \Sigma^{ab}_{\text{phys}}:=-\tfrac{i}{2}[\gamma^a,\gamma^b]\ (a\neq b).$$

In three dimensions with $\gamma_j=\sigma_j$, $\gamma^1\gamma^2=i\sigma_3$ etc., this is $-\vec\sigma\cdot\vec B$: (T4.95) on flat $\mathbb{R}^3$ *is* the Pauli Hamiltonian $(\vec\sigma\cdot\vec\pi)^2=\vec\pi^{\,2}-\vec\sigma\cdot\vec B$ with gyromagnetic ratio $g=2$ (Example T4.2). In Minkowski QFT the same term is $\frac{e}{2}F_{\mu\nu}\sigma^{\mu\nu}$, obtained by squaring $i\not{D}$ [QFT notes §3.5]. The value $g=2$ is not an input: it is forced by the Clifford algebra.

Worked example T4.51: Landau levels of a Dirac fermion

Flat $\mathbb{R}^2$, $\gamma_1=\sigma_1$, $\gamma_2=\sigma_2$, $\Gamma=\sigma_3$, and a constant magnetic field $f_{12}=B>0$. Write $\pi_a=p_a-a_a$ with $p_a=-i\partial_a$, so that $D_A=\sigma_1\pi_1+\sigma_2\pi_2$. First the commutator:

$$[\pi_1,\pi_2]\overset{(1)}{=}-[p_1,a_2]-[a_1,p_2]\overset{(2)}{=}i\,\partial_1a_2-i\,\partial_2a_1=iB.$$

(1) $[p_1,p_2]=0=[a_1,a_2]$.
(2) $[p_a,g]=-i\,\partial_a g$ for a function g.

So π_1,π_2 are like x,p of a harmonic oscillator. With $b:=(\pi_1+i\pi_2)/\sqrt{2B}$ one checks $[b,b^\dagger]=\frac{1}{2B}\big(-i[\pi_1,\pi_2]+i[\pi_2,\pi_1]\big)=1$ and $\pi_1^2+\pi_2^2=B(2b^\dagger b+1)$. Now square D_A:

$$D_A^2\overset{(1)}{=}\pi_1^2+\pi_2^2+\sigma_1\sigma_2\pi_1\pi_2+\sigma_2\sigma_1\pi_2\pi_1\overset{(2)}{=}\pi_1^2+\pi_2^2+i\sigma_3[\pi_1,\pi_2]\overset{(3)}{=}B(2b^\dagger b+1)-B\sigma_3.$$

(1) Expand; $\sigma_1^2=\sigma_2^2=\mathbf{1}$.
(2) $\sigma_1\sigma_2=i\sigma_3=-\sigma_2\sigma_1$.
(3) $i\cdot iB=-B$, and the oscillator form of π^2.

This is (T4.95) with $R=0$ and $-\gamma(F)=-B\sigma_3$. With $N=b^\dagger b=0,1,2,\ldots$:

$$\operatorname{spec}D_A^2=\begin{cases}2NB & \text{on } S^+\ (\sigma_3=+1),\\ 2(N+1)B & \text{on } S^-\ (\sigma_3=-1),\end{cases}\qquad \operatorname{spec}D_A=\{0\}\cup\{\pm\sqrt{2mB}:m=1,2,\ldots\}.$$

The zero modes are the lowest Landau level of the spin-up component: $D_A^-D_A^+$ has the Landau spectrum shifted *down* by B, so its bottom lands exactly on zero. Every nonzero

level $2mB$ appears once in S^+ (with $N = m$) and once in S^- (with $N = m-1$): they pair up. Each Landau level has degeneracy $B/2\pi$ per unit area, i.e. one state per flux quantum. This is the relativistic Landau problem of graphene, where the zero-energy level is responsible for the anomalous half-integer quantum Hall effect.

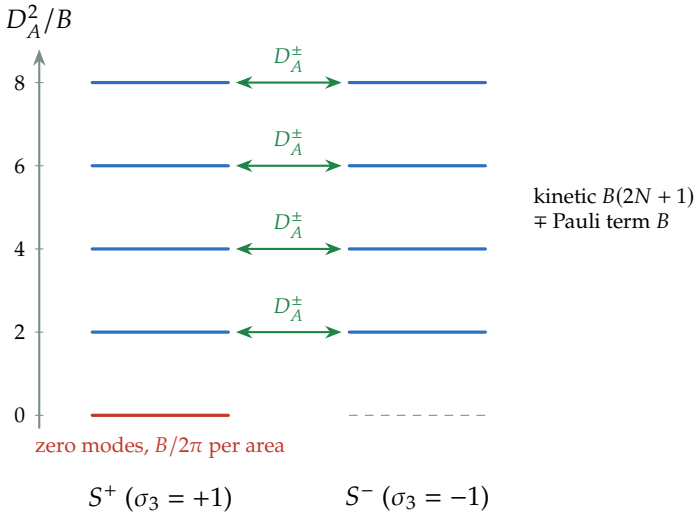


Figure T4.25: Landau levels of $D_A^2 = B(2b^\dagger b + 1) - B\sigma_3$ on $\mathbb{R}^2$. The Pauli term shifts the spin-up tower down and the spin-down tower up by B. Nonzero levels pair up between S^+ and S^- (green arrows: $D_A^\pm$ maps one onto the other). The bottom of the S^+ tower sits exactly at zero and has no partner: these unpaired zero modes are what the index counts.

Worked example T4.52: Aharonov–Casher: the number of zero modes is the flux

Now let $B(x,y) = f_{12}$ be any smooth field on $\mathbb{R}^2$ with compact support and total flux $\Phi = \int B\,\mathrm{d}x\,\mathrm{d}y > 0$. Choose the potential in Coulomb gauge through a scalar φ:

$$a_1 = -\partial_2\varphi, \quad a_2 = \partial_1\varphi \quad\Longrightarrow\quad f_{12} = \partial_1 a_2 - \partial_2 a_1 = \Delta_0\varphi := (\partial_1^2 + \partial_2^2)\varphi = B.$$

The solution is $\varphi(x) = \frac{1}{2\pi}\int \ln|x-y|\,B(y)\,\mathrm{d}^2y$ (the 2D Green's function), so far away $\varphi \approx \frac{\Phi}{2\pi}\ln r$. In complex notation (Example T4.43), $D_A^+ = -2i(\partial_{\bar z} - i a_{\bar z})$ with $a_{\bar z} = \frac{1}{2}(a_1 + i a_2)$. Then

$$-i a_{\bar z} \overset{(1)}{=} -\tfrac{i}{2}\big(-\partial_2\varphi + i\partial_1\varphi\big) = \tfrac{1}{2}(\partial_1 + i\partial_2)\varphi \overset{(2)}{=} \partial_{\bar z}\varphi,$$
$$D_A^+ \overset{(3)}{=} -2i\,(\partial_{\bar z} + \partial_{\bar z}\varphi) \overset{(4)}{=} -2i\,e^{-\varphi}\,\partial_{\bar z}\,e^{\varphi}.$$

(1) Insert the Coulomb-gauge potential.
(2) $\partial_{\bar z} = \frac{1}{2}(\partial_1 + i\partial_2)$.
(3) Substitute.
(4) $e^{-\varphi}\partial_{\bar z}(e^{\varphi}\psi) = \partial_{\bar z}\psi + (\partial_{\bar z}\varphi)\psi$.

So $D_A^+\psi_+ = 0$ iff $e^{\varphi}\psi_+$ is holomorphic: $\psi_+ = e^{-\varphi}h(z)$. The same steps give $D_A^- = -2i\,e^{\varphi}\partial_z e^{-\varphi}$, whose zero modes $e^{\varphi}\bar h(\bar z)$ grow like $r^{\Phi/2\pi}$ and are never normalisable. For ψ_+ normalisability decides everything. At large r, $|\psi_+| \approx r^{-\Phi/2\pi}|h|$, and for $h = z^k$

$$\int^\infty |\psi_+|^2\, r\,\mathrm{d}r \sim \int^\infty r^{2k-\Phi/\pi+1}\,\mathrm{d}r < \infty \iff 2k - \frac{\Phi}{\pi} + 1 < -1 \iff k < \frac{\Phi}{2\pi} - 1.$$

(Any h growing faster than a polynomial is excluded, and polynomials of degree k behave like z^k.) So the zero modes are $e^{-\varphi}z^k$ with $0 \le k < \Phi/2\pi - 1$, all of positive chirality:

$$n_+ = \left\lceil \frac{\Phi}{2\pi} \right\rceil - 1, \qquad n_- = 0. \tag{T4.96}$$

Up to the rounding, *the number of zero modes is the number of flux quanta*. For a uniform field this reproduces the lowest Landau level: $\varphi = \frac{B}{4}r^2$, $\psi_+ = z^k e^{-Br^2/4}$. For $\Phi < 0$ the roles of $S^\pm$ are exchanged. This is the Aharonov–Casher theorem (1979); it is the simplest non-trivial instance of the index theorem, whose answer here is $\frac{1}{2\pi}\int f = \Phi/2\pi$ (the rounding comes from $\mathbb{R}^2$ not being compact).

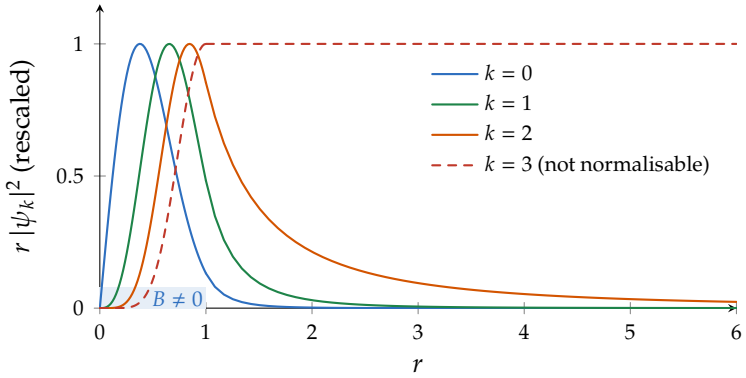


Figure T4.26: Aharonov–Casher zero modes $\psi_k = z^k e^{-\varphi}$ for a uniform flux tube of radius 1 carrying $\Phi/2\pi = 3.5$ flux quanta. Outside the tube $r|\psi_k|^2 \propto r^{2k+1-\Phi/\pi} = r^{2k-6}$. For $k = 0, 1, 2$ the density decays and the mode is normalisable; for $k = 3$ it tends to a constant (red, dashed) and the norm diverges. Three flux quanta (rounded down), three zero modes.

Worked example T4.53: a monopole on S^2: twisting the sphere

Let $L_k \to S^2$ be the line bundle of Chern number k (the Dirac monopole of charge $k/2$ in units where the minimal Dirac charge is $\frac{1}{2}$; see Chapter T3), with the rotationally invariant connection. Its curvature is a constant multiple of the area form, fixed by the Chern number:

$$\frac{i}{2\pi}\int_{S^2} F = k \implies F = -\frac{ik}{2}\, e^1 \wedge e^2, \qquad F_{12} = -\frac{ik}{2}.$$

(Check: $\frac{i}{2\pi}\cdot(-\frac{ik}{2})\cdot 4\pi = k$.) The Clifford action is $\gamma(F) = F_{12}\gamma^1\gamma^2 = (-\frac{ik}{2})(i\sigma_3) = \frac{k}{2}\sigma_3$, and $R/4 = \frac{1}{2}$ on the unit sphere, so (T4.95) reads

$$D_A^2 = \nabla^*\nabla + \frac{1}{2} - \frac{k}{2}\,\sigma_3 = \begin{cases} \nabla^*\nabla + \frac{1-k}{2} & \text{on } S^+ \otimes L_k, \\ \nabla^*\nabla + \frac{1+k}{2} & \text{on } S^- \otimes L_k. \end{cases} \tag{T4.97}$$

Read off, for $k \ge 0$:

- On $S^- \otimes L_k$ the potential is $\ge \frac{1}{2} > 0$, so there are *no* negative-chirality zero modes, by the argument (T4.90).
- $k = 0$: both potentials are $\frac{1}{2}$; no zero modes at all, as in (T4.81).
- $k = 1$: the potential vanishes on $S^+ \otimes L_1$. In fact $S^+ \otimes L_1$ is *flat*: its curvature is $\frac{i}{2}\mathrm{vol}$ (from S^+, Example T4.41) plus $-\frac{i}{2}\mathrm{vol}$ (from L_1). A flat line bundle on the simply connected S^2 has a parallel section, and that section is the unique zero mode: $n_+ = 1$.

- $k \geq 2$: the potential on $S^+ \otimes L_k$ is negative and zero modes are allowed. Their number is $n_+ = k$. The cleanest way to see it is holomorphic: on $S^2 = \mathbb{CP}^1$, D_A^+ is (a multiple of) $\bar\partial$ acting on sections of $S^+ \otimes L_k$, a line bundle of degree $-1 + k$ (Section T4.9.8). Holomorphic sections of the degree-$(k-1)$ bundle $\mathcal{O}(k-1)$ are the polynomials of degree $\leq k-1$ in the affine coordinate z: $1, z, \dots, z^{k-1}$, which is k of them.

The complete spectrum is known in closed form (Wu–Yang monopole harmonics): the eigenvalues are $\pm\sqrt{(j+\frac{1}{2})^2 - \frac{k^2}{4}}$ with $j = \frac{|k|-1}{2}, \frac{|k|+1}{2}, \dots$, each with multiplicity $2j+1$, and the value $j = \frac{|k|-1}{2}$ gives $|k|$ zero modes. For $k = 0$ this is $\pm(j+\frac{1}{2}) = \pm 1, \pm 2, \dots$ with multiplicities $2, 4, \dots$, our (T4.81).

Try it: the Pauli term on S^2

Show that for $k < 0$ the roles are reversed: $n_+ = 0$ and $n_- = |k|$. *Answer:* in (T4.97) the S^+ potential $\frac{1-k}{2} = \frac{1+|k|}{2}$ is now positive, and the S^- potential $\frac{1-|k|}{2}$ can be negative; the holomorphic count applies to the antiholomorphic side. In all cases $n_+ - n_- = k$.

T4.8.5 Chirality and the index: a preview

Every example so far had the same structure. The nonzero eigenvalues of D came in pairs, one in each chirality sector of D^2, and only the zero modes could be lopsided. Nash's Chapter 4 is about the number that measures the lopsidedness.

Working definition T4.13 (index of the Dirac operator)

In even dimension n, with $D^+ : \Gamma(S^+ \otimes E) \to \Gamma(S^- \otimes E)$ [D restricted to fields of chirality $\Gamma = +1$, which it maps to chirality -1] and its adjoint D^- (T4.75), let $n_\pm := \dim \operatorname{Ker} D^\pm$ be the numbers of positive and negative chirality harmonic spinors [independent solutions of $D\psi = 0$, i.e. zero modes]. The **index** is

$$\operatorname{index} D^+ := \dim \operatorname{Ker} D^+ - \dim \operatorname{Ker} D^- = n_+ - n_-. \tag{T4.98}$$

How to compute it in examples: find the zero modes and sort them by chirality. *How to compute it in general:* the Atiyah–Singer formula (T4.103) below.

Nash packages this as the **spin complex**, a two-term sequence $0 \to \Gamma(X, E^+) \xrightarrow{\not{D}} \Gamma(X, E^-) \to 0$ together with its symbol sequence [Nash (4.102)]; "elliptic complex" just means that the symbol $\sigma(\xi) = \not{\xi}$ is an isomorphism $S^+ \to S^-$ for $\xi \neq 0$, which is our $\not{\xi}^2 = |\xi|^2$. Since $D^- = (D^+)^*$ and $\operatorname{Ker} D = \operatorname{Ker} D^+ \oplus \operatorname{Ker} D^-$, the index is also $\dim \operatorname{Ker} D^+ - \dim \operatorname{Coker} D^+$.

Why the index is robust. The key fact is that nonzero eigenvalues pair. Write $D^2 = \operatorname{diag}(D^-D^+, D^+D^-)$ on $S^+ \oplus S^-$. Let $D^-D^+\psi = \mu\psi$ with $\psi \in \Gamma(S^+)$ and $\mu \neq 0$, and put $\chi := D^+\psi$. Then

$$D^+D^-\chi \overset{(1)}{=} D^+(D^-D^+\psi) = \mu\, D^+\psi = \mu\,\chi,$$
$$\|\chi\|^2 \overset{(2)}{=} (D^+\psi, D^+\psi) \overset{(3)}{=} (D^-D^+\psi, \psi) = \mu\|\psi\|^2 \neq 0.$$

(1) Associativity of composition and the eigenvalue equation.
(2) Definition of χ.

(3) D^- is the adjoint of D^+.

So D^+ maps the μ-eigenspace of D^-D^+ injectively into the μ-eigenspace of D^+D^-; by the same argument D^- maps back, and $\frac{1}{\mu}D^-D^+ = \mathbf{1}$ on the first space. The two eigenspaces are isomorphic for every $\mu \neq 0$. (This is the spectral pairing $\lambda \leftrightarrow -\lambda$ of Section T4.7.3, seen through D^2.) Only $\mu = 0$ can be unbalanced. Consequences:

- *Deformation invariance.* If we change the metric or the gauge field continuously, eigenvalues of D^2 move continuously. An eigenvalue can reach zero only together with its partner, so n_+ and n_- can jump only by the same amount. The difference is constant.
- *McKean–Singer formula.* For every $t > 0$

$$\operatorname{Tr}\big(\Gamma\, e^{-tD^2}\big) \overset{(1)}{=} \sum_{\mu} e^{-t\mu}\big(\dim E^+_\mu - \dim E^-_\mu\big) \overset{(2)}{=} n_+ - n_-. \tag{T4.99}$$

  (1) Diagonalise D^2 on S^+ and S^-; $\Gamma = \pm 1$ on the two sectors.
  (2) All terms with $\mu \neq 0$ cancel by the pairing.

  The left side can be computed as $t \to 0$ from purely local data (the short-time heat kernel), which is how the index theorem is proved in Chapter 4. In physics, $\operatorname{Tr}(-1)^F e^{-\beta H}$ with $H = D^2$ is the **Witten index** of supersymmetric quantum mechanics.

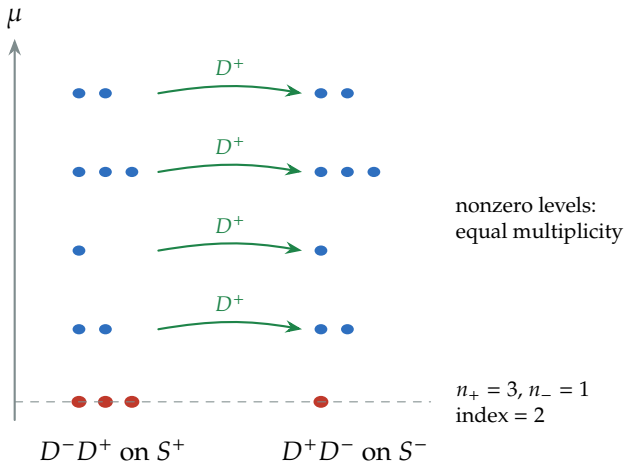


Figure T4.27: Why only zero modes count. For $\mu \neq 0$, D^+ is an isomorphism from the μ-eigenspace of D^-D^+ to that of D^+D^-, so the blue levels always have equal multiplicities on the two sides. Under deformations, levels may descend to zero, but only in pairs (one from each side), so $n_+ - n_-$ (red) cannot change.

The indices we have already computed.

manifold	twisting	n_+	n_-	index
T^2, $(+,+)$	none	1	1	0
T^2, other three	none	0	0	0
S^2	none	0	0	0
S^2	monopole L_k, $k > 0$	k	0	k
$\mathbb{R}^2$	flux Φ	$\lceil \Phi/2\pi \rceil - 1$	0	$\approx \Phi/2\pi$

The pattern is striking: in two dimensions the index is zero without a gauge field, and equal to the number of flux quanta $\frac{i}{2\pi}\int F = c_1(E)$ with one. The index theorem explains this and gives the answer in every dimension.

The index formula. Nash states the result in three steps [Nash (4.107)–(4.111)]. We follow them and fill in the algebra. Write the Pontryagin (or Chern) roots of TM as $\pm x_1, \dots, \pm x_k$ ($n = 2k$): formally, $TM \otimes \mathbb{C}$ splits into line bundles with first Chern classes $\pm x_i$, and $p_1 = \sum_i x_i^2$, $p_2 = \sum_{i<j} x_i^2 x_j^2$, $e(TM) = \prod_i x_i$.

Step 1: the Chern character of the spinor symbol. The weights of the spin representation are $\frac{1}{2}(\pm x_1 \pm \cdots \pm x_k)$: this is the Fock-space description of Section T4.4.2, where each of the k fermion modes is empty or full and a rotation by θ_i in the i-th plane acts with phases $e^{\pm i\theta_i/2}$. The even states make S^+, the odd ones S^-. So

$$\operatorname{ch} S^+ - \operatorname{ch} S^- = \prod_{i=1}^{k} \left(e^{x_i/2} - e^{-x_i/2}\right) \tag{T4.100}$$

up to an overall sign $(-1)^k$ fixed by which half is called S^+ (Nash's (4.107)). Check in $n = 2$: $S^\pm$ are line bundles with $c_1 = \mp\frac{x}{2}$, $x = e(TS^2)$. On S^2, $\int x = \chi = 2$ and $c_1(S^\pm) = \mp 1$, matching Example T4.41.

Step 2: divide by the Euler class and multiply by the Todd class. The general Atiyah–Singer recipe [Nash (4.108)] is index $= (-1)^k \{\operatorname{ch}(\sigma)\, e(TM)^{-1} \operatorname{td}(TM\otimes\mathbb{C})\}[M]$, with $\operatorname{td}(TM\otimes\mathbb{C}) = \prod_i \frac{x_i}{1-e^{-x_i}}\cdot\frac{-x_i}{1-e^{x_i}}$ (roots $\pm x_i$). For one root:

$$\begin{aligned} \frac{e^{x/2}-e^{-x/2}}{x}\cdot\frac{x}{1-e^{-x}}\cdot\frac{-x}{1-e^{x}} &\overset{(1)}{=} \frac{-x\left(e^{x/2}-e^{-x/2}\right)}{2-e^{x}-e^{-x}} \overset{(2)}{=} \frac{-x\left(e^{x/2}-e^{-x/2}\right)}{-\left(e^{x/2}-e^{-x/2}\right)^2} \\ &\overset{(3)}{=} \frac{x}{e^{x/2}-e^{-x/2}} = \frac{x/2}{\sinh(x/2)}. \end{aligned} \tag{T4.101}$$

(1) Cancel x against x, and multiply out $(1-e^{-x})(1-e^{x}) = 2 - e^{x} - e^{-x}$.
(2) $(e^{x/2}-e^{-x/2})^2 = e^{x} - 2 + e^{-x}$.
(3) Cancel one factor; $e^{y} - e^{-y} = 2\sinh y$.

Step 3: the $\hat{A}$ genus. So the index is the integral of

$$\hat{A}(M) := \prod_{i=1}^{k} \frac{x_i/2}{\sinh(x_i/2)} = 1 - \frac{p_1}{24} + \frac{7p_1^2 - 4p_2}{5760} + \cdots \tag{T4.102}$$

[Nash (4.109)–(4.110)]. The expansion is a two-line exercise. From $\frac{\sinh y}{y} = 1 + \frac{y^2}{6} + \frac{y^4}{120} + \dots$ with $y = x/2$:

$$\begin{aligned} \frac{x/2}{\sinh(x/2)} &\overset{(1)}{=} \left(1 + \frac{x^2}{24} + \frac{x^4}{1920}\right)^{-1} \overset{(2)}{=} 1 - \frac{x^2}{24} + \left(\frac{1}{576} - \frac{1}{1920}\right)x^4 = 1 - \frac{x^2}{24} + \frac{7x^4}{5760}, \\ \prod_i \left(1 - \frac{x_i^2}{24} + \frac{7x_i^4}{5760}\right) &\overset{(3)}{=} 1 - \frac{\sum x_i^2}{24} + \frac{7\sum x_i^4}{5760} + \frac{\sum_{i<j} x_i^2 x_j^2}{576} \overset{(4)}{=} 1 - \frac{p_1}{24} + \frac{7(p_1^2 - 2p_2) + 10p_2}{5760}. \end{aligned}$$

(1) $y^2/6 = x^2/24$ and $y^4/120 = x^4/1920$.
(2) $(1+u)^{-1} = 1 - u + u^2 - \dots$ with $u = \frac{x^2}{24} + \frac{x^4}{1920}$; the x^4 terms are $\frac{x^4}{576} - \frac{x^4}{1920}$, and $\frac{1}{576} - \frac{1}{1920} = \frac{10-3}{5760}$.
(3) Multiply out to fourth order.
(4) $\sum x_i^4 = p_1^2 - 2p_2$, $\sum_{i<j} x_i^2 x_j^2 = p_2$, and $\frac{1}{576} = \frac{10}{5760}$.

The last numerator is $7p_1^2 - 4p_2$, which is Nash's (4.110).

Key idea: the index theorem for Dirac operators [Nash (4.111), (4.122)]

On a closed even-dimensional spin manifold, with E a gauge bundle with connection,

$$\operatorname{index} D_A^+ = \int_M \hat{A}(M)\, \operatorname{ch}(E)\Big|_{\text{top degree}} \tag{T4.103}$$

(up to the orientation sign $(-1)^{n/2}$ that Nash keeps explicitly). With no twisting ($\operatorname{ch} E = 1$) only the degree-n part of $\hat{A}$ contributes, and since $\hat{A}$ has components only in degrees $0, 4, 8, \ldots$, *the untwisted index vanishes unless* $4 \mid n$.

Worked example T4.54: the index formula in two and four dimensions

$n = 2$. $\hat{A} = 1 + (\text{degree} \geq 4)$, so $\hat{A}$ contributes only its 1, and $\operatorname{ch} E = \operatorname{rk} E + c_1(E) + \ldots$ contributes its 2-form part:

$$\operatorname{index} D_A^+ \overset{(1)}{=} \int_M c_1(E) \overset{(2)}{=} \frac{i}{2\pi} \int_M \operatorname{tr} F.$$

(1) The top-degree part of $\hat{A} \operatorname{ch} E$ in two dimensions.
(2) House convention for c_1.

Untwisted: 0 on every surface (torus and sphere, as found). Monopole L_k on S^2: k. Flux on $\mathbb{R}^2$: $\Phi/2\pi$, which Aharonov–Casher confirmed up to the boundary rounding.
$n = 4$, **untwisted.** $\operatorname{index} D^+ = -\frac{1}{24}\int_M p_1$. The Hirzebruch signature theorem (Section T4.9.7) gives $\int p_1 = 3\sigma(M)$, so

$$\operatorname{index} D^+ = -\frac{\sigma(M)}{8}. \tag{T4.104}$$

The K3 surface has $\sigma = -16$, so $\hat{A} = 2$: two harmonic spinors (in fact parallel ones, since K3 carries a Ricci-flat metric). For $\mathbb{CP}^2$, $\sigma = 1$ and the formula would give $-\frac{1}{8}$, *not an integer*. There is no contradiction: $\mathbb{CP}^2$ is not spin (Example T4.37), so there is no Dirac operator whose index this could be. The fractional answer is the index theorem's own way of detecting $w_2 \neq 0$.
$n = 4$, **twisted by an** SU(2) **bundle.** $\operatorname{ch} E = 2 + \operatorname{ch}_2(E)$ with $\operatorname{ch}_2 = \frac{1}{2}\operatorname{tr}\left(\frac{iF}{2\pi}\right)^2 = -\frac{1}{8\pi^2}\operatorname{tr} F \wedge F$ (and $c_1 = 0$). So

$$\operatorname{index} D_A^+ \overset{(1)}{=} \int_M \operatorname{ch}_2(E) + 2\int_M \left(-\frac{p_1}{24}\right) \overset{(2)}{=} -\frac{1}{8\pi^2}\int_M \operatorname{tr} F \wedge F - \frac{\sigma(M)}{4}.$$

(1) Degree-4 part of $(1 - \frac{p_1}{24})(2 + \operatorname{ch}_2)$.
(2) Definitions, and $\int p_1 = 3\sigma$.

On S^4 ($\sigma = 0$) the index is minus the instanton number $k = \frac{1}{8\pi^2}\int \operatorname{tr} F \wedge F$ of the house convention: an instanton of charge k has $|k|$ fermion zero modes of one chirality in the fundamental representation. These are 't Hooft's zero modes, the source of instanton-induced fermion interactions and of the anomaly (Chapter 8, Chapter 10).

Key idea: Lichnerowicz meets Atiyah–Singer

Combine Section T4.8.3 with (T4.103). If a closed spin manifold carries *some* metric with $R > 0$, then $\operatorname{Ker} D = 0$ for that metric, hence $\operatorname{index} D^+ = 0$; but the index does not depend

on the metric. So

$$\hat{A}(M) \neq 0 \quad \Longrightarrow \quad M \text{ admits no metric of positive scalar curvature.}$$

Example: the K3 surface ($\hat{A} = 2$). A differential-geometric question (can the curvature be positive?) has been answered by a topological invariant computed from zero modes. This is the prototype of every argument in Nash's book: *count zero modes analytically, compute them topologically, and compare*.

Reading Nash (pp. 113–118, (4.96)–(4.122))

Nash's treatment of the Dirac operator as an index problem occupies pp. 113–116, and the twisted version is on p. 118. The correspondences are: spinor bundles $E^{\pm}$ (4.98) = our $S^{\pm}$; $\not{D}$ (4.100) = $-D$; chiral operator $\not{D}(1+\gamma_5)/2$ (4.102) = our D^+ (up to sign); harmonic spinors $\Psi^{\pm}$ (4.104) = our $\operatorname{Ker} D^{\pm}$; $\operatorname{index}\not{D} = n^+ - n^-$ (4.106) = (T4.98); the character formula (4.107) = (T4.100); the passage (4.108) from $\operatorname{ch}/e \cdot \operatorname{td}$ to $\prod \frac{x_i/2}{\sinh(x_i/2)}$ = our chain (T4.101); $\hat{A}$ and its expansion (4.109)–(4.110) = (T4.102); (4.111) and the twisted (4.122) = (T4.103). The sign $(-1)^{n/2}$ in Nash depends on his choice of γ_5 and orientation; in our conventions (with $\Gamma = \sigma_3$ in two dimensions and positive flux) (T4.103) holds with a plus sign, as the Aharonov–Casher and monopole examples confirm. Nash does not discuss the Lichnerowicz formula; it is the standard route to the vanishing results he uses implicitly.

Recap

- The rough Laplacian $\nabla^*\nabla = -\sum(\nabla_a\nabla_a - \nabla_{\nabla_a e_a})$ is non-negative with kernel the parallel sections. Every Laplace-type operator is $\nabla^*\nabla + K$.
- Lichnerowicz: $D^2 = \nabla^*\nabla + \frac{R}{4}$. Proof: split $\gamma^a\gamma^b\nabla_a\nabla_b$ into symmetric part (rough Laplacian) and antisymmetric part (curvature), then $R_{abcd}\gamma^a\gamma^b\gamma^c\gamma^d = -2R$ by the three-gamma identity and Bianchi.
- $R \geq 0$: harmonic spinors are parallel; $R > 0$ somewhere: none. Gap $|\lambda| \geq \frac{1}{2}\sqrt{s_0}$, improved by Friedrich to $\lambda^2 \geq \frac{n}{n-1}\frac{s_0}{4}$, sharp on round spheres (Killing spinors). On S^2, $\lambda^2 \geq 4\pi/\text{area}$ for every metric.
- Twisted: $D_A = -i\gamma^a(\nabla^S_a + A_a)$, $D_A^2 = \nabla^*\nabla + \frac{R}{4} - \gamma(F)$: the Pauli term with $g = 2$. Landau levels, Aharonov–Casher ($n_+ \approx \Phi/2\pi$) and the monopole on S^2 ($n_+ - n_- = k$).
- Nonzero modes pair between chiralities, so $\operatorname{index} D^+ = n_+ - n_-$ is deformation invariant and equals $\operatorname{Tr}\Gamma e^{-tD^2}$. Formula: $\int \hat{A}\,\operatorname{ch} E$, $\hat{A} = 1 - \frac{p_1}{24} + \dots$. Positive scalar curvature forces $\hat{A} = 0$.

T4.9 Differential forms as Clifford modules

Everything so far needed a spin structure. On $\mathbb{CP}^2$ there is none (Example T4.37), so on $\mathbb{CP}^2$ there is no Dirac operator D acting on spinors. And yet $\mathbb{CP}^2$ has a perfectly good Laplacian on differential forms, a Hodge theory, an Euler characteristic ($\chi = 3$), a signature ($\sigma = 1$) and Nash computes all of these as indices of elliptic operators [Nash pp. 103–112]. Where do those operators come from, if there are no spinors?

The answer is the observation we made at the end of Section T4.2.6: the exterior algebra

already carries an action of the Clifford algebra. Wedging with a 1-form and contracting with it are a creation and an annihilation operator, and their sum and difference obey the Clifford relations. So the bundle of all forms $\Lambda^\bullet T^*M$ is a Clifford module on *every* Riemannian manifold, and the operator $\mathrm{d} + \mathrm{d}^*$ is a Dirac-type operator on it. The questions of this section are:

- What exactly is the Clifford action on forms? Can we write it as explicit matrices and check the relations?
- How are forms related to spinors? (Answer: a form is a spinor with an extra spinor index; in two dimensions, a 2×2 matrix whose columns are spinors.)
- Why is $\mathrm{d} + \mathrm{d}^*$ of Dirac type, and what are its zero modes? (Answer: harmonic forms, which are de Rham cohomology.)
- Which gradings of $\Lambda^\bullet$ anticommute with $\mathrm{d} + \mathrm{d}^*$, and what indices do they produce? (Answer: the Euler characteristic, the signature and, on complex manifolds, the arithmetic genus.)

These are the three "classical elliptic complexes" that Nash treats as examples of the index theorem before he reaches the spin complex. Here we build the operators; Chapter 4 computes their indices.

Notation. V is an oriented Euclidean space of dimension n (later $V = T^*_x M$) with orthonormal basis $e^1, \dots, e^n$. We identify vectors and covectors with the metric, so we write e^a and e_a interchangeably. For an increasing multi-index $I = (i_1 < \cdots < i_k)$ we write $e^I = e^{i_1} \wedge \cdots \wedge e^{i_k}$; these 2^n forms are an orthonormal basis of $\Lambda^\bullet V$. The degree operator N multiplies a k-form by k. Forms are complex-valued when we need eigenvalues, with the Hermitian inner product in which the e^I are orthonormal.

T4.9.1 Wedge and contraction: creation and annihilation operators

There are two natural linear maps on $\Lambda^\bullet V$ attached to a vector v.

Working definition T4.14 (exterior and interior multiplication)

For $v \in V$ define, on every form α,

$$\varepsilon(v)\alpha := v \wedge \alpha, \qquad \iota(v)\alpha := \iota_v \alpha,$$

[$\varepsilon(v)$ raises the degree of a form by one and $\iota(v)$ lowers it by one, like fermion creation and annihilation operators], where ι_v is the interior product ("feed v into the first slot", Chapter T1). *Operationally*, on a basis form: $\varepsilon(e^j)e^I$ is 0 if $j \in I$ and otherwise $\pm e^{I \cup \{j\}}$; $\iota(e^j)e^I$ is 0 if $j \notin I$ and otherwise $\pm e^{I \setminus \{j\}}$. The sign is $(-1)^s$ where s is the number of indices in I smaller than j (the number of swaps needed to move e^j to the front).

For example, in $\mathbb{R}^3$: $\varepsilon(e^2)e^{13} = e^2 \wedge e^1 \wedge e^3 = -e^{123}$ and $\iota(e^2)e^{123} = \iota_{e_2}(e^1 \wedge e^2 \wedge e^3) = -e^{13}$: in both cases one swap past e^1.

Physical picture: forms are a fermionic Fock space

Think of the n basis directions as n fermionic modes. The form $e^{i_1} \wedge \cdots \wedge e^{i_k}$ is the state with modes $i_1, \dots, i_k$ occupied; the function 1 is the vacuum. Wedging with e^j *creates* a fermion in mode j (and gives zero if it is already there: Pauli exclusion is $e^j \wedge e^j = 0$).

Contracting with e_j *annihilates* it. The degree N is the fermion number. This is the same Fock space we met for spinors in Section T4.4.2, but with n modes instead of $n/2$, which is why forms have 2^n components and spinors only $2^{n/2}$.

The three anticommutators are the whole algebra.

Key idea: canonical anticommutation relations on forms

For all $u, v \in V$,

$$\{\varepsilon(u), \varepsilon(v)\} = 0, \qquad \{\iota(u), \iota(v)\} = 0, \qquad \{\varepsilon(u), \iota(v)\} = \langle u, v\rangle\, \mathbf{1}. \tag{T4.105}$$

Moreover $\iota(v)$ is the adjoint of $\varepsilon(v)$: $\varepsilon(v)^\dagger = \iota(v)$.

The first relation is the antisymmetry of the wedge product: $u \wedge v \wedge \alpha = -v \wedge u \wedge \alpha$. The second is the same statement for contractions, $\iota_u \iota_v \alpha = \alpha(v, u, \dots) = -\alpha(u, v, \dots) = -\iota_v \iota_u \alpha$. The third is the only one with content. We need the graded Leibniz rule for the interior product (Chapter T1): for a 1-form u and any form α, $\iota_v(u \wedge \alpha) = (\iota_v u)\,\alpha - u \wedge \iota_v \alpha$. Then

$$\begin{aligned} \{\varepsilon(u), \iota(v)\}\alpha &\overset{(1)}{=} u \wedge \iota_v \alpha + \iota_v(u \wedge \alpha) \\ &\overset{(2)}{=} u \wedge \iota_v \alpha + \langle u, v\rangle\, \alpha - u \wedge \iota_v \alpha \\ &\overset{(3)}{=} \langle u, v\rangle\, \alpha. \end{aligned} \tag{T4.106}$$

(1) Definition of the anticommutator: $\varepsilon(u)\iota(v) + \iota(v)\varepsilon(u)$.
(2) Graded Leibniz rule; $\iota_v u = u(v) = \langle u, v\rangle$ for a 1-form u.
(3) The two wedge terms cancel.

For the adjoint statement take $j \notin I$. Moving e^j to its place in $e^j \wedge e^I$ and moving it back out of $e^{I\cup\{j\}}$ to contract it take the same number s of swaps, so

$$\langle e^{I\cup\{j\}}, \varepsilon(e^j) e^I\rangle = (-1)^s = \langle \iota(e^j) e^{I\cup\{j\}}, e^I\rangle,$$

and all other matrix elements of both operators vanish. So in the orthonormal basis e^I the matrix of $\iota(e^j)$ is the transpose of the (real) matrix of $\varepsilon(e^j)$. Bär gives a basis-free version of the same argument [Bär Ex. 1.3.10]: split every form as $\alpha = \xi \wedge \alpha_\xi + \alpha_{\xi^\perp}$ with no ξ in $\alpha_\xi, \alpha_{\xi^\perp}$, and compare $\langle \xi \wedge \alpha, \beta\rangle$ with $\langle \alpha, \iota_\xi \beta\rangle$.

Now the Clifford algebra appears. From one creation and one annihilation operator we can form two Hermitian-or-anti-Hermitian combinations.

Working definition T4.15 (the two Clifford actions on forms)

For $v \in V$ put

$$c(v) := \varepsilon(v) - \iota(v), \qquad \hat{c}(v) := \varepsilon(v) + \iota(v). \tag{T4.107}$$

Then

$$\{c(u), c(v)\} = -2\langle u, v\rangle, \qquad \{\hat{c}(u), \hat{c}(v)\} = +2\langle u, v\rangle, \qquad \{c(u), \hat{c}(v)\} = 0. \tag{T4.108}$$

$c(v)$ is anti-Hermitian [$c(v)^\dagger = -c(v)$, because $\iota(v)$ is the adjoint of $\varepsilon(v)$] and squares to $-|v|^2$: it is Bär's Clifford multiplication (Chapter T4, conventions box). $\hat{c}(v)$ is Hermitian and squares to $+|v|^2$: it is a gamma matrix in our convention. The Hermitian gamma matrices that go with c are

$$\gamma(v) := i\, c(v) = i\big(\varepsilon(v) - \iota(v)\big), \qquad \gamma(v)^2 = |v|^2. \tag{T4.109}$$

Test: to check that a family of operators is a Clifford action, compute the three anticommutators in (T4.108) using only (T4.105) [ε's anticommute among themselves, ι's likewise, and $\{\varepsilon(u), \iota(v)\} = \langle u, v\rangle$].

Here is the computation of the first relation; the other two are identical in structure.

$$\{c(u), c(v)\} \overset{(1)}{=} \{\varepsilon(u), \varepsilon(v)\} - \{\varepsilon(u), \iota(v)\} - \{\iota(u), \varepsilon(v)\} + \{\iota(u), \iota(v)\}$$
$$\overset{(2)}{=} 0 - \langle u, v\rangle - \langle v, u\rangle + 0 = -2\langle u, v\rangle.$$

(1) The anticommutator is bilinear; expand $c = \varepsilon - \iota$ in both slots.
(2) (T4.105), and the inner product is symmetric.

For $\hat{c}$ the two middle terms come with + signs, giving $+2\langle u, v\rangle$. For the mixed one, $\{c(u), \hat{c}(v)\} = \{\varepsilon, \varepsilon\} + \{\varepsilon(u), \iota(v)\} - \{\iota(u), \varepsilon(v)\} - \{\iota, \iota\} = \langle u, v\rangle - \langle v, u\rangle = 0$. Anti-Hermiticity of c: $c(v)^\dagger = \iota(v) - \varepsilon(v) = -c(v)$ by the adjoint relation.

Watch out: two actions, not one

$\Lambda^\bullet V$ carries *two* anticommuting Clifford actions, c and $\hat{c}$. Together they give $2n$ anticommuting generators acting on a space of dimension 2^n, which is exactly the size of the irreducible spinor representation in $2n$ dimensions (Section T4.4). So $\Lambda^\bullet V$ is *the spinor space of* $V \oplus V$. As a module over one copy of $\mathrm{Cl}(V)$ it is far from irreducible: it is $2^{n/2}$ copies of the spinor space (for even n), as we will see in Section T4.9.2. When a formula says "Clifford multiplication on forms", check which of the two is meant.

Worked example T4.55: the Clifford action on $\Lambda^\bullet\mathbb{R}^2$ as 4×4 matrices

Order the basis of $\Lambda^\bullet\mathbb{R}^2$ as $(1,\ e^1,\ e^2,\ e^{12})$ and write a form as the column $(f, a_1, a_2, b)^{\mathsf{T}}$ for $\alpha = f + a_1 e^1 + a_2 e^2 + b\, e^{12}$. Read off the action of the four basic operators on the basis:

	1	e^1	e^2	e^{12}
$\varepsilon(e^1)$	e^1	0	e^{12}	0
$\iota(e^1)$	0	1	0	e^2
$\varepsilon(e^2)$	e^2	$-e^{12}$	0	0
$\iota(e^2)$	0	0	1	$-e^1$

(for instance $\varepsilon(e^2)e^1 = e^2 \wedge e^1 = -e^{12}$ and $\iota(e^2)e^{12} = \iota_{e_2}(e^1 \wedge e^2) = -e^1$). The columns of the matrices are the images of the basis vectors, so

$$c(e^1) = \begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \quad c(e^2) = \begin{pmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix},$$

$$\hat{c}(e^1) = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \quad \hat{c}(e^2) = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix}.$$

These are 2×2 block matrices, so they are Kronecker products. With $A \otimes B = \left(\begin{smallmatrix} a_{11}B & a_{12}B \\ a_{21}B & a_{22}B \end{smallmatrix}\right)$ and $J := -i\sigma_2 = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$,

$$c(e^1) = \mathbf{1} \otimes J, \qquad c(e^2) = J \otimes \sigma_3, \qquad \hat{c}(e^1) = \mathbf{1} \otimes \sigma_1, \qquad \hat{c}(e^2) = \sigma_1 \otimes \sigma_3. \tag{T4.110}$$

Now every relation is a one-line check with the rule $(A\otimes B)(C\otimes D)=AC\otimes BD$ of (T4.34). Using $J^2=-\mathbf{1}$, $\sigma_1 J=-J\sigma_1=\sigma_3$ and $\sigma_3 J=-J\sigma_3$:

$$c(e^2)c(e^1)+c(e^1)c(e^2)\overset{(1)}{=}J\otimes\sigma_3 J+J\otimes J\sigma_3\overset{(2)}{=}J\otimes(\sigma_3 J+J\sigma_3)=0,$$

$$c(e^1)^2\overset{(3)}{=}\mathbf{1}\otimes J^2=-\mathbf{1},\qquad c(e^2)^2\overset{(4)}{=}J^2\otimes\sigma_3^2=-\mathbf{1},$$

$$c(e^1)\hat{c}(e^1)+\hat{c}(e^1)c(e^1)\overset{(5)}{=}\mathbf{1}\otimes(J\sigma_1+\sigma_1 J)=0.$$

(1) Multiply slot by slot.
(2) Collect the common first factor J.
(3) $J^2=-\mathbf{1}$.
(4) $J^2=-\mathbf{1}$, $\sigma_3^2=\mathbf{1}$.
(5) J and σ_1 anticommute ($J=-i\sigma_2$).

The Hermitian gammas (T4.109) are $\gamma(e^1)=ic(e^1)=\mathbf{1}\otimes\sigma_2$ and $\gamma(e^2)=\sigma_2\otimes\sigma_3$, since $iJ=\sigma_2$. The parity of the degree is $(-1)^N=\mathrm{diag}(1,-1,-1,1)=\sigma_3\otimes\sigma_3$, and it anticommutes with all four generators. This is a Jordan–Wigner representation of the kind in (T4.35), with the σ_3 string on the other side (the sign in $e^2\wedge e^1=-e^{12}$ is what puts it there).

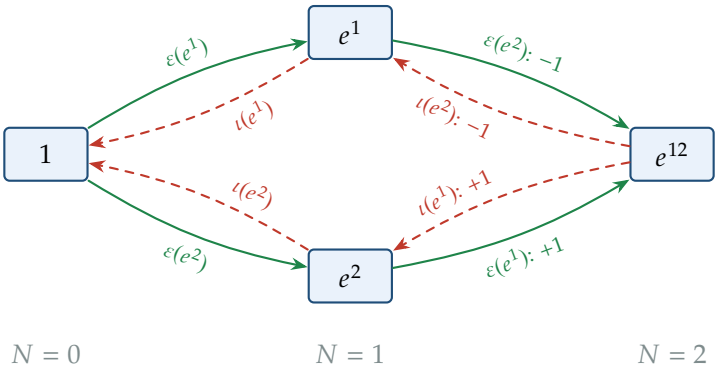


Figure T4.28: The exterior algebra of $\mathbb{R}^2$ as a two-mode fermionic Fock space. Green arrows are wedge products (creation operators), red dashed arrows are contractions (annihilation operators); the labels on the right-hand arrows are the signs. Following $\varepsilon(e^1)$ then $\varepsilon(e^2)$ from 1 gives $-e^{12}$ (top route), following them in the other order gives $+e^{12}$ (bottom route): that is $\{\varepsilon(e^1),\varepsilon(e^2)\}=0$. Going out and back along one mode, in either order, returns the state once: that is $\{\varepsilon(e^j),\iota(e^j)\}=1$.

T4.9.2 Forms are spinors with an extra spinor index

Dray says it in words [Dray §20.3]: "the space of formal sums of differential forms under the Clifford product is a Clifford algebra … spinors actually live in certain subspaces of Clifford algebras". Let us make the statement concrete.

Working definition T4.16 (the quantisation map)

Let $\gamma_1,\dots,\gamma_n$ be Hermitian gamma matrices, $\{\gamma_a,\gamma_b\}=2\delta_{ab}$. The **quantisation map** (or **symbol map** read backwards) sends a form to a matrix by replacing wedge products of basis 1-forms by products of gammas:

$$Q\big(e^{i_1}\wedge\cdots\wedge e^{i_k}\big):=\gamma_{i_1}\gamma_{i_2}\cdots\gamma_{i_k}\qquad(i_1<\cdots<i_k),\tag{T4.111}$$

extended linearly. It is a linear isomorphism $\Lambda^\bullet V \otimes \mathbb{C} \to \mathrm{Cl}(V) \otimes \mathbb{C}$ [$\mathrm{Cl}(V) \otimes \mathbb{C}$ is all complex linear combinations of products of the γ's] [Bär Rem. 2.1.15], but not an algebra map [it does not satisfy $Q(\alpha \wedge \beta) = Q(\alpha)Q(\beta)$ in general]: $Q(e^1 \wedge e^1) = 0 \neq \gamma_1\gamma_1 = \mathbf{1}$.

The point of Q is that it turns the action $\hat{c}$ into ordinary matrix multiplication from the left, and c into multiplication from the right:

$$Q\big(\hat{c}(v)\alpha\big) = \gamma(v)\,Q(\alpha), \qquad Q\big(c(v)\alpha\big) = (-1)^k\,Q(\alpha)\,\gamma(v) \quad (\alpha \in \Lambda^k), \tag{T4.112}$$

where on the right $\gamma(v) = v^a\gamma_a$. To check the first, take $v = e^j$ and a basis form e^I, with j in position p of $I \cup \{j\}$. If $j \notin I$, then $\gamma_j\gamma_{i_1}\cdots\gamma_{i_k}$ can be reordered into increasing order by moving γ_j past the $p-1$ smaller indices, each move costing a sign, which is exactly the sign in $e^j \wedge e^I$. If $j \in I$ (in position p), then

$$\gamma_j\,Q(e^I) \overset{(1)}{=} (-1)^{p-1}\,\gamma_{i_1}\cdots\gamma_{i_{p-1}}\,\gamma_j\gamma_j\,\gamma_{i_{p+1}}\cdots\gamma_{i_k} \overset{(2)}{=} (-1)^{p-1}\,Q\big(e^{I\setminus\{j\}}\big) \overset{(3)}{=} Q\big(\iota(e^j)e^I\big).$$

(1) Move γ_j to the right past the $p-1$ smaller gammas; they anticommute.
(2) $\gamma_j^2 = \mathbf{1}$.
(3) The contraction $\iota_{e_j}e^I$ has the same sign $(-1)^{p-1}$ (move e^j to the front).

In both cases $\gamma_j Q(e^I) = Q(\varepsilon(e^j)e^I + \iota(e^j)e^I) = Q(\hat{c}(e^j)e^I)$. The right multiplication rule is checked the same way; the extra $(-1)^k$ comes from moving γ_j past all k factors instead of the ones on its left.

Worked example T4.56: a form on $\mathbb{R}^2$ is a 2×2 matrix whose columns are spinors

Take $\gamma_1 = \sigma_1$, $\gamma_2 = \sigma_2$, so $\gamma_1\gamma_2 = \sigma_1\sigma_2 = i\sigma_3$. For $\alpha = f + a_1e^1 + a_2e^2 + b\,e^{12}$,

$$Q(\alpha) \overset{(1)}{=} f\,\mathbf{1} + a_1\sigma_1 + a_2\sigma_2 + b\,i\sigma_3 \overset{(2)}{=} \begin{pmatrix} f + ib & a_1 - ia_2 \\ a_1 + ia_2 & f - ib \end{pmatrix}. \tag{T4.113}$$

(1) Definition (T4.111).
(2) Write out the Pauli matrices.

The four real (or complex) components of α fill the four entries of a 2×2 matrix, and Q is invertible: $f = \frac{1}{2}\operatorname{tr} Q$, $b = -\frac{i}{2}\operatorname{tr}(\sigma_3 Q)$, and so on. Left multiplication by $\gamma(v)$ acts on each *column* separately, as a gamma matrix acts on a two-component spinor. So

$$\Lambda^\bullet\mathbb{R}^2 \otimes \mathbb{C} \;\cong\; \{2\times 2 \text{ matrices}\} \;=\; S \oplus S \;=\; S \otimes \mathbb{C}^2,$$

two copies of the spinor space $S = \mathbb{C}^2$, one per column. Check with $v = e^1$ and $\alpha = 1$: $\hat{c}(e^1)1 = e^1$ and indeed $\sigma_1 \cdot \mathbf{1} = \sigma_1 = Q(e^1)$. Check with $\alpha = e^{12}$: $\hat{c}(e^1)e^{12} = \iota_{e_1}e^{12} = e^2$, and $\sigma_1\,(i\sigma_3) = i\sigma_1\sigma_3 = i(-i\sigma_2) = \sigma_2 = Q(e^2)$.

The general statement follows from the classification of Section T4.4.3: for even n, $\mathrm{Cl}(V)\otimes\mathbb{C}$ is the full matrix algebra $\mathrm{End}(S)$ with $\dim S = 2^{n/2}$.

Key idea: forms are bispinors

For n even,

$$\Lambda^\bullet V \otimes \mathbb{C} \;\cong\; \mathrm{Cl}(V)\otimes\mathbb{C} \;\cong\; \mathrm{End}(S) \;=\; S \otimes S^*, \qquad 2^n = 2^{n/2}\cdot 2^{n/2}, \tag{T4.114}$$

with $\hat{c}$ acting on the S factor and c on the S^* factor. On a spin manifold this holds bundle-wise: $\Lambda^\bullet T^*M \otimes \mathbb{C} \cong S \otimes S^*$.

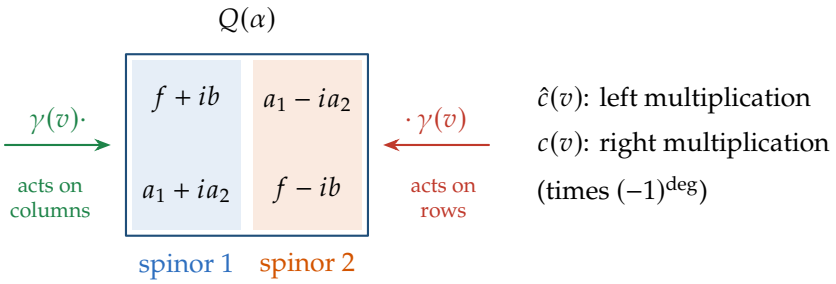


Figure T4.29: The quantisation map turns a form on $\mathbb{R}^2$ into a 2×2 matrix. The Clifford action $\hat{c}$ multiplies from the left and so acts on each column, which is therefore a spinor. The other action c multiplies from the right and acts on the row index. A form is a spinor with an extra (dual) spinor index: $\Lambda^\bullet \otimes \mathbb{C} \cong S \otimes S^*$.

Physical picture: why forms never need a spin structure

On a spin manifold the spinor bundle is glued with lifted transition functions $\tilde{g}_{\alpha\beta} \in \mathrm{Spin}(n)$, each defined only up to a sign, and the signs must be chosen consistently (Section T4.6.2). A bispinor $S \otimes S^*$ transforms with $\tilde{g} \otimes (\tilde{g}^{-1})^{\mathsf{T}}$, in which the sign appears *twice* and cancels. So the bundle $S \otimes S^*$ can be glued on any manifold, spin or not, and it is just $\Lambda^\bullet T^*M \otimes \mathbb{C}$. The same cancellation is why a vector ($\cong$ a traceless bispinor, cf. $X = x^\mu \sigma_\mu$ in Section T4.3.7) never notices the 2π sign. Every operator built on forms, in particular $\mathrm{d} + \mathrm{d}^*$, exists on every Riemannian manifold.

Dictionary: forms and fermions

- *Kähler–Dirac fermions.* The equation $(\mathrm{d} + \mathrm{d}^*)\Phi = 0$ for an inhomogeneous form Φ is equivalent to $2^{n/2}$ Dirac equations, one per column of $Q(\Phi)$ (Example T4.57 below). In four dimensions a form has 16 components: four Dirac spinors.
- *Staggered lattice fermions.* Lattice QCD exploits exactly this: placing the 2^4 components of a form on the corners of a hypercube gives a lattice Dirac operator with four "tastes" of quark, because a form is four spinors.
- *Supersymmetric quantum mechanics.* On the Fock space of forms, ε and ι are fermion creation and annihilation operators, and d and d^* are supercharges (Section T4.9.6, Chapter 7).

T4.9.3 The operator $\mathrm{d} + \mathrm{d}^*$ is of Dirac type

Now differentiate. Let M be an oriented Riemannian manifold, d the exterior derivative on $\Omega^\bullet(M)$ and d^* its formal adjoint, the **codifferential**, defined by $(\mathrm{d}\alpha, \beta) = (\alpha, \mathrm{d}^*\beta)$ for compactly supported forms, with $(\alpha, \beta) = \int_M \langle \alpha, \beta \rangle$ vol. In terms of the Hodge star (defined by $\alpha \wedge \star\beta = \langle \alpha, \beta \rangle$ vol, Chapter T1) it is [Bär Lemma 1.3.17]

$$\mathrm{d}^* = (-1)^{n(k+1)+1} \star \mathrm{d} \star \quad \text{on } k\text{-forms}, \qquad \text{so } \mathrm{d}^* = - \star \mathrm{d} \star \quad \text{for } n \text{ even}. \tag{T4.115}$$

For our purposes the more useful form is in terms of ε and ι.

Key idea: d and d^* in a frame

In any local orthonormal frame, with ∇ the Levi-Civita connection on forms,

$$\mathrm{d} = \sum_a \varepsilon(e^a)\,\nabla_{e_a}, \qquad \mathrm{d}^* = -\sum_a \iota(e^a)\,\nabla_{e_a}, \qquad \mathrm{d}+\mathrm{d}^* = \sum_a c(e^a)\,\nabla_{e_a} = -i\,\gamma^a\nabla_a. \tag{T4.116}$$

So $\mathrm{d}+\mathrm{d}^*$ has *exactly* the form of our Dirac operator $D = -i\gamma^a\nabla_a$ of (T4.72), with the spinor bundle replaced by $\Lambda^\bullet T^*M$ and the gammas by $\gamma(e^a) = ic(e^a)$ of (T4.109).

Why (T4.116) is true. In flat space with Cartesian coordinates, $\alpha = \frac{1}{k!}\alpha_{a_1\ldots a_k}e^{a_1}\wedge\cdots\wedge e^{a_k}$ with constant $e^a = \mathrm{d}x^a$, and $\mathrm{d}\alpha = \mathrm{d}x^a\wedge\partial_a\alpha$: this is the definition of d in components. The adjoint of a product is the product of adjoints in reverse order, so

$$\mathrm{d}^* \overset{(1)}{=} \Big(\sum_a \varepsilon(e^a)\,\partial_a\Big)^\dagger \overset{(2)}{=} \sum_a (\partial_a)^\dagger\,\varepsilon(e^a)^\dagger \overset{(3)}{=} \sum_a(-\partial_a)\,\iota(e^a) \overset{(4)}{=} -\sum_a \iota(e^a)\,\partial_a.$$

(1) The flat-space formula for d; d^* is its adjoint.
(2) $(AB)^\dagger = B^\dagger A^\dagger$.
(3) Integration by parts, $(\partial_a)^\dagger = -\partial_a$; and $\varepsilon^\dagger = \iota$ from (T4.105).
(4) $\iota(e^a)$ has constant coefficients, so it commutes with ∂_a.

On a curved manifold replace ∂_a by ∇_a. For d this is allowed because the Levi-Civita connection is torsion-free: the Christoffel symbols $\Gamma^\lambda_{\mu\nu}$ are symmetric in $\mu\nu$ and drop out when antisymmetrised by the wedge. For d^* the same integration by parts works with the divergence theorem, exactly as in the self-adjointness argument for D (T4.73). On a 1-form $\alpha = \alpha_a e^a$ the formula gives $\mathrm{d}^*\alpha = -\nabla_a\alpha_a = -\operatorname{div}\alpha$, which is the familiar statement that the adjoint of grad is $-\operatorname{div}$ (Example T4.4).

The symbol. Replacing ∇_a by $i\xi_a$ (Section T4.1.3) in (T4.116),

$$\sigma_{\mathrm{d}}(\xi) = i\,\varepsilon(\xi), \qquad \sigma_{\mathrm{d}^*}(\xi) = -i\,\iota(\xi), \qquad \sigma_{\mathrm{d}+\mathrm{d}^*}(\xi) = i\,c(\xi) = \gamma(\xi). \tag{T4.117}$$

The Dirac-type test of Section T4.1.3 is Bär's computation (1.15) in our language:

$$\begin{aligned}\sigma_{\mathrm{dd}^*+\mathrm{d}^*\mathrm{d}}(\xi) &\overset{(1)}{=} \sigma_{\mathrm{d}}(\xi)\sigma_{\mathrm{d}^*}(\xi) + \sigma_{\mathrm{d}^*}(\xi)\sigma_{\mathrm{d}}(\xi)\\ &\overset{(2)}{=} (i\varepsilon(\xi))(-i\iota(\xi)) + (-i\iota(\xi))(i\varepsilon(\xi))\\ &\overset{(3)}{=} \{\varepsilon(\xi),\iota(\xi)\} \overset{(4)}{=} |\xi|^2\,\mathbf{1}.\end{aligned} \tag{T4.118}$$

(1) Symbols of compositions multiply.
(2) (T4.117).
(3) $i\cdot(-i) = 1$.
(4) (T4.105) with $u = v = \xi$.

So $\Delta := \mathrm{dd}^* + \mathrm{d}^*\mathrm{d}$, the **Hodge Laplacian**, is of Laplace type, and since $(\mathrm{d}+\mathrm{d}^*)^2 = \mathrm{dd}^* + \mathrm{d}^*\mathrm{d}$ (because $\mathrm{d}^2 = 0 = (\mathrm{d}^*)^2$), the operator $\mathrm{d}+\mathrm{d}^*$ on $\Lambda^\bullet T^*M$ is a formally self-adjoint operator of Dirac type [Bär Ex. 1.3.10, Rem. 1.3.13]. This is how the gradient's failure (Example T4.4) is repaired: add the curl and the divergence and act on all degrees at once.

Worked example T4.57: $\mathrm{d}+\mathrm{d}^*$ on $\mathbb{R}^2$: the Cauchy–Riemann operator twice

Use the basis and matrices of Example T4.55. Then $\mathrm{d}+\mathrm{d}^* = c(e^1)\partial_1 + c(e^2)\partial_2$ acts on (f, a_1, a_2, b) as

$$(\mathrm{d}+\mathrm{d}^*)\begin{pmatrix} f \\ a_1 \\ a_2 \\ b \end{pmatrix} \overset{(1)}{=} \begin{pmatrix} -\partial_1 a_1 \\ \partial_1 f \\ -\partial_1 b \\ \partial_1 a_2 \end{pmatrix} + \begin{pmatrix} -\partial_2 a_2 \\ \partial_2 b \\ \partial_2 f \\ -\partial_2 a_1 \end{pmatrix} \overset{(2)}{=} \begin{pmatrix} -(\partial_1 a_1 + \partial_2 a_2) \\ \partial_1 f + \partial_2 b \\ \partial_2 f - \partial_1 b \\ \partial_1 a_2 - \partial_2 a_1 \end{pmatrix}. \tag{T4.119}$$

(1) Multiply the columns by the matrices $c(e^1)$ and $c(e^2)$.
(2) Add.

Read the rows: the 0-form part is $-\operatorname{div} a$ (d^* of the 1-form), the 1-form part is grad f plus $\mathrm{d}^*(b\,e^{12}) = \partial_2 b\,e^1 - \partial_1 b\,e^2$, and the 2-form part is $\mathrm{d}a = (\partial_1 a_2 - \partial_2 a_1)e^{12}$, the curl. Applying the operator twice gives $-(\partial_1^2 + \partial_2^2)$ on every component (try it: the first row becomes $-\partial_1(\partial_1 f + \partial_2 b) - \partial_2(\partial_2 f - \partial_1 b) = -\nabla^2 f$).
The zero modes: $(\mathrm{d}+\mathrm{d}^*)\alpha = 0$ splits into an even part (rows 2 and 3 involve only f, b) and an odd part (rows 1 and 4 involve only a_1, a_2):

$$\partial_1 f = -\partial_2 b, \quad \partial_2 f = \partial_1 b \qquad \text{and} \qquad \partial_1 a_1 = -\partial_2 a_2, \quad \partial_1 a_2 = \partial_2 a_1.$$

The first pair is the Cauchy–Riemann system for $f - ib$; the second for $a_1 - ia_2$. So harmonic forms on $\mathbb{R}^2$ are *two* holomorphic functions, one for each column of $Q(\alpha)$: compare Example T4.43, where one harmonic spinor was one holomorphic function.
Finally, the columns. Under Q the operator $\sum_a \hat{c}(e^a)\partial_a = \mathrm{d} - \mathrm{d}^*$ becomes left multiplication $\sigma_1\partial_1 Q + \sigma_2\partial_2 Q = i\,D_{\mathrm{col}}Q$, where $D_{\mathrm{col}} = -i(\sigma_1\partial_1 + \sigma_2\partial_2)$ is our Dirac operator acting on each column. The two operators $\mathrm{d} \pm \mathrm{d}^*$ are unitarily equivalent: with $U := i^N$ (i^k on k-forms),

$$U\mathrm{d}U^{-1}\alpha \overset{(1)}{=} i^{k+1}\,\mathrm{d}(i^{-k}\alpha) = i\,\mathrm{d}\alpha, \qquad U\mathrm{d}^*U^{-1}\alpha \overset{(2)}{=} i^{k-1}\,\mathrm{d}^*(i^{-k}\alpha) = -i\,\mathrm{d}^*\alpha,$$

(1) d raises the degree by one.
(2) d^* lowers the degree by one; $i^{-1} = -i$.

so $U(\mathrm{d}+\mathrm{d}^*)U^{-1} = i(\mathrm{d}-\mathrm{d}^*)$, which under Q is $-D_{\mathrm{col}}$. Hence $\mathrm{d}+\mathrm{d}^*$ *on forms is unitarily equivalent to (minus) the Dirac operator acting on two spinors side by side.* This is the precise version of Dray's "the Dirac equation can be rewritten in terms of differential forms".

Worked example T4.58: $\mathrm{d}+\mathrm{d}^*$ on $\mathbb{R}^3$: grad, curl and div are one operator

On $\mathbb{R}^3$ identify a form with a quadruple of scalar and vector fields,

$$\alpha = f + A\cdot\mathrm{d}x + B_1\,e^{23} + B_2\,e^{31} + B_3\,e^{12} + g\,e^{123} \;\longleftrightarrow\; (f, \vec{A}, \vec{B}, g).$$

Then $\mathrm{d}f = \nabla f\cdot\mathrm{d}x$, $\mathrm{d}(A\cdot\mathrm{d}x) \leftrightarrow \nabla\times\vec{A}$, $\mathrm{d}(B) = (\nabla\cdot\vec{B})\,e^{123}$, and from $\mathrm{d}^* = -\sum_a \iota(e^a)\partial_a$:

$$\mathrm{d}^*(A\cdot\mathrm{d}x) \overset{(1)}{=} -\partial_a A_a = -\nabla\cdot\vec{A},$$
$$\mathrm{d}^*B \overset{(2)}{=} -\big[\partial_1(B_3e^2 - B_2e^3) + \partial_2(B_1e^3 - B_3e^1) + \partial_3(B_2e^1 - B_1e^2)\big] \overset{(3)}{=} (\nabla\times\vec{B})\cdot\mathrm{d}x,$$
$$\mathrm{d}^*(g\,e^{123}) \overset{(4)}{=} -\big(\partial_1 g\,e^{23} + \partial_2 g\,e^{31} + \partial_3 g\,e^{12}\big) \;\leftrightarrow\; -\nabla g.$$

(1) $\iota(e^a)e^b = \delta^{ab}$.
(2) $\iota_{e_1}B = B_3e^2 - B_2e^3$ (from $\iota_{e_1}e^{31} = -e^3$, $\iota_{e_1}e^{12} = e^2$), and cyclically.

(3) Collect the coefficient of each e^a: e.g. of e^1, $\partial_2 B_3 - \partial_3 B_2$.
(4) $\iota_{e_1} e^{123} = e^{23}$, $\iota_{e_2} e^{123} = -e^{13} = e^{31}$, $\iota_{e_3} e^{123} = e^{12}$.

So

$$(\mathrm{d} + \mathrm{d}^*)(f, \vec{A}, \vec{B}, g) = \big(-\nabla \cdot \vec{A},\ \ \nabla f + \nabla \times \vec{B},\ \ \nabla \times \vec{A} - \nabla g,\ \ \nabla \cdot \vec{B}\big). \tag{T4.120}$$

All of first-order vector calculus is one Dirac-type operator. Now square it. The 1-form component of $(\mathrm{d} + \mathrm{d}^*)^2$ is

$$\nabla(-\nabla \cdot \vec{A}) + \nabla \times (\nabla \times \vec{A} - \nabla g) \overset{(1)}{=} -\nabla(\nabla \cdot \vec{A}) + \nabla \times \nabla \times \vec{A} \overset{(2)}{=} -\nabla^2 \vec{A},$$

(1) $\nabla \times \nabla g = 0$ (this is $\mathrm{d}^2 = 0$).
(2) The identity $\nabla \times \nabla \times \vec{A} = \nabla(\nabla \cdot \vec{A}) - \nabla^2 \vec{A}$.

Read backwards, *the vector identity "curl curl = grad div − Laplacian" is the statement* $(\mathrm{d} + \mathrm{d}^*)^2 = \mathrm{d}^*\mathrm{d} + \mathrm{d}\mathrm{d}^* = -\nabla^2$ *on 1-forms*. The other components work the same way ($\nabla \cdot \nabla \times = 0$ kills the cross terms).

Physical picture: Maxwell's equations are a Dirac equation

Let F be the electromagnetic field strength 2-form. Maxwell's equations are $\mathrm{d}F = 0$ (no monopoles, Faraday) and $\mathrm{d}^*F = J$ (Gauss and Ampère, with the current 1-form J; the sign depends on conventions and, in Lorentzian signature, on the metric used to define d^*). Because $\mathrm{d}F$ is a 3-form and d^*F a 1-form, the two equations are the components of the single equation

$$(\mathrm{d} + \mathrm{d}^*)F = J.$$

In Lorenz gauge $\mathrm{d}^*A = 0$ we have $F = \mathrm{d}A = (\mathrm{d} + \mathrm{d}^*)A$, and then $(\mathrm{d} + \mathrm{d}^*)^2 A = \Delta A = J$: the wave equation for the potential is the square of the Dirac-type equation, exactly as Klein–Gordon is the square of Dirac (Section T4.1).

Watch out: $\mathrm{d} + \mathrm{d}^*$ mixes degrees

$\mathrm{d} + \mathrm{d}^*$ does not map k-forms to k-forms; it maps Λ^k to $\Lambda^{k+1} \oplus \Lambda^{k-1}$. Its square $\Delta = \mathrm{d}\mathrm{d}^* + \mathrm{d}^*\mathrm{d}$ does preserve the degree. So "the Dirac operator on forms" only makes sense on all of $\Lambda^\bullet$, just as the spinor Dirac operator only makes sense on all of $S = S^+ \oplus S^-$ and exchanges the two halves.

T4.9.4 Hodge theory: harmonic forms are the zero modes

For the Dirac operator the interesting objects were the zero modes. For $\mathrm{d} + \mathrm{d}^*$ the zero modes turn out to be cohomology classes.

Working definition T4.17 (harmonic forms)

A form h on a closed manifold [compact and without boundary, so integration by parts produces no boundary terms] is **harmonic** if $\Delta h = 0$ [$\Delta = \mathrm{d}\mathrm{d}^* + \mathrm{d}^*\mathrm{d}$, the Hodge Laplacian]. *Test:* h is harmonic iff it is both closed and coclosed,

$$\Delta h = 0 \iff (\mathrm{d} + \mathrm{d}^*)h = 0 \iff \mathrm{d}h = 0 \text{ and } \mathrm{d}^* h = 0.$$

The test follows from positivity, as for the rough Laplacian (T4.83):

$$(\Delta\alpha, \alpha) \overset{(1)}{=} (\mathrm{d}\mathrm{d}^*\alpha, \alpha) + (\mathrm{d}^*\mathrm{d}\alpha, \alpha) \overset{(2)}{=} (\mathrm{d}^*\alpha, \mathrm{d}^*\alpha) + (\mathrm{d}\alpha, \mathrm{d}\alpha) = \|\mathrm{d}^*\alpha\|^2 + \|\mathrm{d}\alpha\|^2,$$

$$\|(\mathrm{d}+\mathrm{d}^*)\alpha\|^2 \overset{(3)}{=} \|\mathrm{d}\alpha\|^2 + \|\mathrm{d}^*\alpha\|^2 + 2\,\mathrm{Re}\,(\mathrm{d}\alpha, \mathrm{d}^*\alpha) \overset{(4)}{=} \|\mathrm{d}\alpha\|^2 + \|\mathrm{d}^*\alpha\|^2. \tag{T4.121}$$

(1) Definition of Δ.
(2) Move one factor to the other side of the inner product; d and d^* are adjoint.
(3) Expand the norm of a sum.
(4) $(\mathrm{d}\alpha, \mathrm{d}^*\alpha) = (\mathrm{dd}\alpha, \alpha) = 0$ since $\mathrm{d}^2 = 0$.

All three quantities vanish together, which is the test. The kernel of $\mathrm{d}+\mathrm{d}^*$ on $\Omega^\bullet$ is therefore the direct sum over degrees of the harmonic k-forms $\mathcal{H}^k$, since Δ preserves degree.

Key idea: Hodge's theorem

On a closed oriented Riemannian manifold every de Rham class contains exactly one harmonic form:

$$H^k_{\mathrm{dR}}(M) \;\cong\; \mathcal{H}^k = \operatorname{Ker}\Delta\big|_{\Omega^k}, \qquad \operatorname{Ker}(\mathrm{d}+\mathrm{d}^*) = \bigoplus_k \mathcal{H}^k, \qquad \dim \mathcal{H}^k = b_k. \tag{T4.122}$$

Equivalently, every form splits orthogonally as $\Omega^k = \mathcal{H}^k \oplus \mathrm{d}\Omega^{k-1} \oplus \mathrm{d}^*\Omega^{k+1}$ (the **Hodge decomposition**).

Two parts of this are elementary and worth seeing. *Orthogonality* of the three pieces: for h harmonic, $(\mathrm{d}\beta, \mathrm{d}^*\gamma) = (\mathrm{dd}\beta, \gamma) = 0$, $(h, \mathrm{d}\beta) = (\mathrm{d}^*h, \beta) = 0$ and $(h, \mathrm{d}^*\gamma) = (\mathrm{d}h, \gamma) = 0$. *Uniqueness and meaning* of the harmonic representative: if α is closed and coclosed, then for every β

$$\|\alpha + \mathrm{d}\beta\|^2 \overset{(1)}{=} \|\alpha\|^2 + 2\,\mathrm{Re}(\alpha, \mathrm{d}\beta) + \|\mathrm{d}\beta\|^2 \overset{(2)}{=} \|\alpha\|^2 + 2\,\mathrm{Re}(\mathrm{d}^*\alpha, \beta) + \|\mathrm{d}\beta\|^2 \overset{(3)}{=} \|\alpha\|^2 + \|\mathrm{d}\beta\|^2.$$

(1) Expand the norm.
(2) Adjointness.
(3) $\mathrm{d}^*\alpha = 0$.

So the harmonic form is the *unique shortest element* of its cohomology class $\alpha + \mathrm{d}\Omega^{k-1}$, and two harmonic forms in the same class differ by an exact form $\mathrm{d}\beta$ that is itself harmonic, so $\|\mathrm{d}\beta\|^2 = (\mathrm{d}^*\mathrm{d}\beta, \beta) = 0$. In physics language, choosing the harmonic representative is choosing the *Coulomb gauge* $\mathrm{d}^*\alpha = 0$ in the class of α under gauge transformations $\alpha \to \alpha + \mathrm{d}\beta$: it minimises the energy $\|\alpha\|^2$. Existence (that the minimum is attained) is analysis: it is the statement that Δ is elliptic, proved in Chapter 2.

Worked example T4.59: harmonic forms on the flat torus and on the round sphere

Flat $T^2 = \mathbb{R}^2/(2\pi\mathbb{Z})^2$**.** By (T4.119) and Fourier series, a form is harmonic iff each component function is harmonic, $-\nabla^2 f = 0$, and on a torus that means constant (expand $f = \sum f_{pq} e^{i(px+qy)}$: $-\nabla^2$ multiplies by p^2+q^2). So

$$\mathcal{H}^0 = \langle 1 \rangle, \quad \mathcal{H}^1 = \langle \mathrm{d}x, \mathrm{d}y \rangle, \quad \mathcal{H}^2 = \langle \mathrm{d}x \wedge \mathrm{d}y \rangle, \qquad (b_0, b_1, b_2) = (1, 2, 1).$$

Under Q these four constant forms become the constant matrices $\mathbf{1}, \sigma_1, \sigma_2, i\sigma_3$, whose columns are constant spinors. On the torus with the periodic spin structure there are 2 constant (harmonic) spinors (Example T4.45); two columns of two harmonic spinors each give $4 = 2 \times 2$ harmonic forms, as (T4.114) predicts. (For the other spin structures there are no harmonic spinors, but the forms do not care: the two sign choices in $S \otimes S^*$ cancel.)
Round S^2**.** $b_0 = b_2 = 1$, $b_1 = 0$: the harmonic forms are the constants and multiples of the area form $\mathrm{vol} = \sin\theta\, \mathrm{d}\theta \wedge \mathrm{d}\phi$. The constant is harmonic trivially. For the area form, $\mathrm{d}\,\mathrm{vol} = 0$ (top degree) and $\mathrm{d}^*\mathrm{vol} = -\star\,\mathrm{d}\star\,\mathrm{vol} = -\star\,\mathrm{d}1 = 0$. There are no harmonic 1-forms:

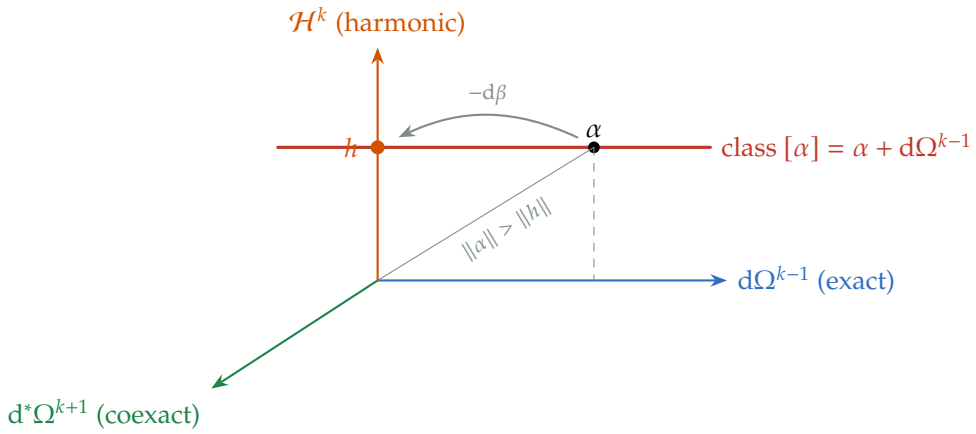


Figure T4.30: The Hodge decomposition. The three subspaces are mutually orthogonal. A closed form has no coexact part, so its cohomology class is a line parallel to the exact axis (red). The harmonic representative h is the point of the line closest to the origin: subtracting the exact part $\mathrm{d}\beta$ of α lowers the squared norm by $\|\mathrm{d}\beta\|^2$.

a harmonic 1-form h would be closed, hence $h = \mathrm{d}f$ on the simply connected S^2, and then $0 = \mathrm{d}^*\mathrm{d}f = \Delta f$ forces f constant and $h = 0$.

Watch out: harmonic forms exist where harmonic spinors cannot

The round S^2 has no harmonic spinors at all (positive scalar curvature, Section T4.8.3), yet it has the two harmonic forms 1 and vol. No contradiction: forms are $S \otimes S^*$, i.e. spinors *twisted* by the bundle S^*, and the twisted Lichnerowicz formula of Section T4.8.4 has an extra curvature term from the twisting bundle that can be negative. Positivity of R alone does not kill harmonic forms. What *does* kill some of them is positivity of Ricci, as we now show.

T4.9.5 The Bochner formula

The first step of the Lichnerowicz derivation (T4.86) used nothing but the Clifford relations and a parallel Clifford action. So it applies verbatim to $\mathrm{d} + \mathrm{d}^* = \sum_a c(e^a)\nabla_a$, with the only change that $c(e^a)^2 = -1$ where $\gamma_a^2 = +1$.

Key idea: Weitzenböck formula for a Dirac-type operator

If $D = \sum_a c(e^a)\nabla_a$ with $\{c(e^a), c(e^b)\} = -2\delta^{ab}$ and $\nabla c = 0$, then

$$D^2 = \nabla^*\nabla + \frac{1}{2}\sum_{a,b} c(e^a)c(e^b)\, R^{\nabla}(e_a, e_b), \tag{T4.123}$$

where $R^{\nabla}(e_a, e_b) = [\nabla_a, \nabla_b]$ (in a synchronous frame) is the curvature of the bundle.

The derivation at a point with a synchronous frame:

$$D^2 \overset{(1)}{=} \sum_{a,b} c(e^a)c(e^b)\nabla_a\nabla_b \overset{(2)}{=} \sum_{a,b} \tfrac{1}{2}\{c(e^a), c(e^b)\}\nabla_a\nabla_b + \sum_{a,b} \tfrac{1}{2}[c(e^a), c(e^b)]\nabla_a\nabla_b$$
$$\overset{(3)}{=} -\sum_a \nabla_a\nabla_a + \tfrac{1}{2}\sum_{a,b} c(e^a)c(e^b)\,[\nabla_a, \nabla_b] \overset{(4)}{=} \nabla^*\nabla + \tfrac{1}{2}\sum_{a,b} c(e^a)c(e^b)R^{\nabla}(e_a, e_b).$$

(1) $c(e^b)$ is parallel at the point, so ∇_a passes through it.
(2) Split $c^a c^b$ into its symmetric and antisymmetric parts.
(3) Clifford relation $\{c^a, c^b\} = -2\delta^{ab}$; the antisymmetric part $\frac{1}{2}[c^a, c^b]$ pairs with the antisymmetric part of $\nabla_a \nabla_b$, and for $a \neq b$ it equals $c^a c^b$ (for $a = b$ the commutator $[\nabla_a, \nabla_a] = 0$ anyway).
(4) (T4.82) at the point, and $[\nabla_a, \nabla_b] = R^{\nabla}(e_a, e_b)$ since $[e_a, e_b] = 0$ there.

For spinors ($c \to -i\gamma$) the curvature term is $\frac{R}{4}$, the Lichnerowicz formula. For forms the curvature R^{∇} is the Riemann tensor acting on forms as a derivation. On 1-forms it acts by $R^{\nabla}(e_a, e_b)e^d = \sum_c R_{cdab}\, e^c$ in the convention $R_{abcd} = \langle R(e_c, e_d)e_b, e_a\rangle$ of (T4.68).

Key idea: Bochner formula [Bär Prop. 1.3.12, (1.17)]

On 1-forms,

$$\Delta = \mathrm{d}\mathrm{d}^* + \mathrm{d}^*\mathrm{d} = \nabla^*\nabla + \mathrm{Ric}, \tag{T4.124}$$

where Ric acts on a 1-form $\theta = \theta_d e^d$ by $(\mathrm{Ric}\,\theta)_b = \mathrm{Ric}_{bd}\theta_d$.

To derive it we evaluate the curvature term of (T4.123) on $\theta = \theta_d e^d$. First the Clifford algebra: using $c = \varepsilon - \iota$ twice,

$$c(e^b)e^c \overset{(1)}{=} e^b \wedge e^c - \delta^{bc},$$
$$c(e^a)c(e^b)e^c \overset{(2)}{=} e^a \wedge e^b \wedge e^c - \delta^{ab}e^c + \delta^{ac}e^b - \delta^{bc}e^a.$$

(1) $\varepsilon(e^b)e^c = e^b \wedge e^c$ and $\iota(e^b)e^c = \delta^{bc}$.
(2) Apply $c(e^a)$: $\iota_{e_a}(e^b \wedge e^c) = \delta^{ab}e^c - \delta^{ac}e^b$, and $c(e^a)1 = e^a$.

Now contract with the curvature (all indices summed):

$$\begin{aligned}\tfrac{1}{2}\, c(e^a)c(e^b)R^{\nabla}(e_a, e_b)\theta &\overset{(1)}{=} \tfrac{1}{2}\,\theta_d R_{cdab}\, c(e^a)c(e^b)e^c \\ &\overset{(2)}{=} \tfrac{1}{2}\,\theta_d\big(R_{cdab}\, e^a \wedge e^b \wedge e^c - R_{cdaa}e^c + R_{adab}e^b - R_{bdab}e^a\big) \\ &\overset{(3)}{=} \tfrac{1}{2}\,\theta_d\big(0 - 0 + R_{adab}e^b + R_{adab}e^b\big) \overset{(4)}{=} \mathrm{Ric}_{bd}\,\theta_d\, e^b.\end{aligned}$$

(1) $R^{\nabla}(e_a, e_b)e^d = R_{cdab}e^c$.
(2) Insert the Clifford computation.
(3) The first term is the antisymmetrisation of R_{cdab} over a, b, c, which vanishes by the first Bianchi identity; $R_{cdaa} = 0$; in the last term rename $a \leftrightarrow b$ and use $R_{adba} = -R_{adab}$.
(4) $\mathrm{Ric}_{bd} = \sum_a R_{abad} = \sum_a R_{adab}$ (pair symmetry).

With (T4.123) and $D^2 = \Delta$ this is (T4.124). The calculation is the same three-gamma identity (T4.88) that produced $-2\,\mathrm{Ric}$ in the Lichnerowicz derivation; there it was contracted once more to give the scalar curvature, here the free index is the form index.

Key idea: Bochner's vanishing theorem

On a closed manifold, integrate $(\Delta h, h) = 0$ for a harmonic 1-form h:

$$0 = \|\nabla h\|^2 + \int_M \mathrm{Ric}(h, h)\,\mathrm{vol}. \tag{T4.125}$$

If $\mathrm{Ric} \geq 0$, both terms vanish, so every harmonic 1-form is parallel and $b_1 \leq n$. If moreover $\mathrm{Ric} > 0$ somewhere, then $h = 0$ there, hence everywhere (it is parallel), and $b_1(M) = 0$.

Worked example T4.60: Bochner on spheres, tori and $S^1 \times S^2$

- *Round* S^n ($n \geq 2$): Ric $= (n-1)g > 0$, so $b_1 = 0$, as we found directly for S^2. Quantitatively $\Delta \geq n-1$ on 1-forms. On S^2 the lowest 1-form eigenvalue is 2 (the forms $\mathrm{d}Y_{1m}$, since $\Delta \mathrm{d}f = \mathrm{d}\Delta f$ and $\Delta Y_{1m} = 2Y_{1m}$), which respects the bound ≥ 1.
- *Flat* T^n: Ric $= 0$, harmonic 1-forms are parallel, and indeed they are the constant forms $\mathrm{d}x^1, \ldots, \mathrm{d}x^n$: $b_1 = n$, saturating the bound.
- $S^1 \times S^2$ with the product metric: Ric ≥ 0 (zero along S^1, positive along S^2). The parallel 1-form $\mathrm{d}\theta$ along the circle is harmonic, and $b_1 = 1$.
- *Surfaces of genus* $g \geq 2$ with constant curvature -1: Ric < 0, no obstruction, and indeed $b_1 = 2g$.

bundle	Dirac-type operator	Weitzenböck potential K	vanishing result
spinors S	$D = -i\gamma^a \nabla_a$	$\frac{R}{4}$	$R > 0 \Rightarrow$ no harmonic spinors
$S \otimes E$	D_A	$\frac{R}{4} - \gamma(F)$	needs F small
1-forms	$\mathrm{d} + \mathrm{d}^*$	Ric	Ric $> 0 \Rightarrow b_1 = 0$
functions	$\mathrm{d} + \mathrm{d}^*$	0	harmonic $\Rightarrow$ constant

T4.9.6 The Euler operator: the index is the Euler characteristic

We now have a Dirac-type operator $\mathrm{d} + \mathrm{d}^*$ on $\Lambda^\bullet T^*M$ on every Riemannian manifold. For the spinor Dirac operator the interesting number was not the kernel itself but the *index*, the difference of zero modes in two sectors exchanged by D (Section T4.8.5). To imitate that we need a grading: a splitting $\Lambda^\bullet = E^+ \oplus E^-$ such that $\mathrm{d} + \mathrm{d}^*$ maps E^+ to E^- and back. In the language of Section T4.3.3, we need an operator that squares to $\mathbf{1}$ and *anticommutes* with every Clifford generator $c(e^a)$, and that is parallel. Differential forms carry two such operators. The first one is obvious.

Working definition T4.18 (the Euler grading and the Euler operator)

Let $(-1)^N$ be $+1$ on even forms and -1 on odd forms. The **Euler operator** (Nash's $\mathbf{d}$, [Nash (4.51)]) is the restriction of $\mathrm{d} + \mathrm{d}^*$ to even forms [sums of forms of degree $0, 2, 4, \ldots$],

$$D_{\mathrm{E}} := (\mathrm{d} + \mathrm{d}^*)\big|_{\Omega^{\mathrm{even}}} : \Omega^{\mathrm{even}}(M) \longrightarrow \Omega^{\mathrm{odd}}(M), \tag{T4.126}$$

with adjoint $D_{\mathrm{E}}^* = (\mathrm{d} + \mathrm{d}^*)|_{\Omega^{\mathrm{odd}}}$. *Test* that it is well defined: $(-1)^N$ must anticommute with $\mathrm{d} + \mathrm{d}^*$.

The test is one line per operator. Since $\varepsilon(v)$ raises the degree by one and $\iota(v)$ lowers it by one, on a k-form α

$$(-1)^N c(v)\,\alpha \overset{(1)}{=} (-1)^{k+1}\varepsilon(v)\alpha - (-1)^{k-1}\iota(v)\alpha \overset{(2)}{=} -(-1)^k\big(\varepsilon(v) - \iota(v)\big)\alpha \overset{(3)}{=} -c(v)\,(-1)^N\alpha,$$

(1) $\varepsilon(v)\alpha$ has degree $k+1$ and $\iota(v)\alpha$ has degree $k-1$.
(2) $(-1)^{k+1} = (-1)^{k-1} = -(-1)^k$.
(3) $(-1)^N\alpha = (-1)^k\alpha$.

so $\{(-1)^N, c(v)\} = 0$, and the same computation gives $\{(-1)^N, \hat{c}(v)\} = 0$. The Levi-Civita

connection preserves degrees, so ∇_a commutes with $(-1)^N$, and therefore, from (T4.116),

$$(-1)^N(\mathrm{d}+\mathrm{d}^*) \overset{(1)}{=} \sum_a (-1)^N c(e^a)\nabla_a \overset{(2)}{=} -\sum_a c(e^a)(-1)^N\nabla_a \overset{(3)}{=} -(\mathrm{d}+\mathrm{d}^*)(-1)^N.$$

(1) (T4.116).
(2) $(-1)^N$ anticommutes with $c(e^a)$.
(3) $(-1)^N$ commutes with ∇_a.

Of course one sees this directly: d and d^* each change the degree by one. The Clifford version is worth having because it is the template for the second grading below, where "directly" is much less obvious.

Key idea: the index of the Euler operator [Nash (4.52)]

On a closed oriented Riemannian manifold,

$$\operatorname{index} D_{\mathrm{E}} = \sum_k (-1)^k b_k(M) = \chi(M). \tag{T4.127}$$

The derivation is just Hodge theory:

$$\operatorname{index} D_{\mathrm{E}} \overset{(1)}{=} \dim\operatorname{Ker}\big((\mathrm{d}+\mathrm{d}^*)|_{\text{even}}\big) - \dim\operatorname{Ker}\big((\mathrm{d}+\mathrm{d}^*)|_{\text{odd}}\big) \overset{(2)}{=} \sum_{k\ \text{even}} \dim\mathcal{H}^k - \sum_{k\ \text{odd}} \dim\mathcal{H}^k$$
$$\overset{(3)}{=} \sum_k (-1)^k b_k \overset{(4)}{=} \chi(M).$$

(1) Definition of the index, with $\operatorname{Coker} D_{\mathrm{E}} = \operatorname{Ker} D_{\mathrm{E}}^*$ (as for D^+ in Section T4.8.5).
(2) The kernel of $\mathrm{d}+\mathrm{d}^*$ is the direct sum of the harmonic forms of each degree, by the test (T4.121); sort them by parity.
(3) Hodge's theorem (T4.122): $\dim\mathcal{H}^k = b_k$.
(4) Euler–Poincaré: the alternating sum of Betti numbers is the Euler characteristic (the alternating count of cells, $V - E + F$ for a surface, Chapter T2).

The index theorem computes the same number as the integral of a curvature expression, the Euler form: that is the Gauss–Bonnet–Chern theorem $\int_M e(TM) = \chi(M)$ [Nash (4.53)–(4.54)]. In two dimensions $e = \frac{1}{2\pi}K\,\mathrm{vol}$ and this is the Gauss–Bonnet theorem of Chapter T2. We will see in Section T4.9.9 how the Euler form arises from the $\hat{A}$ genus.

Worked example T4.61: the Euler operator on the circle: every level pairs, including zero

On S^1 with coordinate $\theta \in [0, 2\pi)$ and metric $\mathrm{d}\theta^2$, an even form is a function f and an odd form is $g\,\mathrm{d}\theta$. From $\mathrm{d}f = f'\,\mathrm{d}\theta$ and $\mathrm{d}^*(g\,\mathrm{d}\theta) = -\iota(\mathrm{d}\theta)\partial_\theta(g\,\mathrm{d}\theta) = -g'$,

$$(\mathrm{d}+\mathrm{d}^*)\begin{pmatrix} f \\ g \end{pmatrix} = \begin{pmatrix} -g' \\ f' \end{pmatrix} = \begin{pmatrix} 0 & -\partial_\theta \\ \partial_\theta & 0 \end{pmatrix}\begin{pmatrix} f \\ g \end{pmatrix}. \tag{T4.128}$$

On a Fourier mode $(f, g) = (a, b)\,e^{in\theta}$ the operator is the 2×2 matrix

$$\begin{pmatrix} 0 & -in \\ in & 0 \end{pmatrix} = n\,\sigma_2, \qquad \text{eigenvalues } \pm n, \quad n \in \mathbb{Z}.$$

So the spectrum of $\mathrm{d}+\mathrm{d}^*$ is $\{\pm n\}$, the spectrum of the *periodic* (Ramond) Dirac operator of Example T4.44, doubled. It is never the half-integer spectrum of the antiperiodic spin

structure, because $S \otimes S^*$ does not see the sign of the spin structure (Section T4.9.2). The Laplacian $\Delta = (\mathrm{d} + \mathrm{d}^*)^2 = -\partial_\theta^2$ has eigenvalue n^2 on both sides:

$\lambda = n^2$	even (functions)	odd (1-forms)	D_{E} maps
0	1	$\mathrm{d}\theta$	$1 \mapsto 0$
$n^2,\ n \geq 1$	$e^{\pm in\theta}$	$e^{\pm in\theta}\mathrm{d}\theta$	$e^{in\theta} \mapsto in\, e^{in\theta}\mathrm{d}\theta$

Every nonzero level has the same multiplicity (2) on both sides, as the pairing argument of Section T4.8.5 requires, and at $\lambda = 0$ there is one zero mode on each side:

$$\operatorname{index} D_{\mathrm{E}} = \dim \mathcal{H}^0 - \dim \mathcal{H}^1 = 1 - 1 = 0 = \chi(S^1).$$

Worked example T4.62: the Euler operator on the round S^2

The eigenfunctions of $\Delta = -\nabla^2$ on functions are the spherical harmonics Y_{lm}, with eigenvalue $l(l+1)$ and multiplicity $2l+1$. We build all other eigenforms from them.

- *2-forms.* $\star$ commutes with Δ (because $\mathrm{d}^* = -\star\,\mathrm{d}\star$ in even dimensions), so $Y_{lm}\,\mathrm{vol} = \star Y_{lm}$ has the same eigenvalue.
- *1-forms.* Δ commutes with d ($\Delta\mathrm{d} = \mathrm{d}\mathrm{d}^*\mathrm{d} = \mathrm{d}\Delta$), so $\mathrm{d}Y_{lm}$ is an eigenform with eigenvalue $l(l+1)$, and so is $\star\mathrm{d}Y_{lm}$. They are nonzero for $l \geq 1$: $\|\mathrm{d}Y_{lm}\|^2 = (\mathrm{d}^*\mathrm{d}Y_{lm}, Y_{lm}) = l(l+1)$. By the Hodge decomposition, with $\mathcal{H}^1 = 0$, every 1-form is exact plus coexact, so these exhaust the 1-forms.

Sorting by parity:

λ	even: Y_{lm} and $Y_{lm}\mathrm{vol}$	odd: $\mathrm{d}Y_{lm}$ and $\star\mathrm{d}Y_{lm}$	difference
0	$1 + 1 = 2$	0	2
$l(l+1),\ l \geq 1$	$(2l+1) + (2l+1)$	$(2l+1) + (2l+1)$	0

The pairing map is explicit: for $l \geq 1$,

$$D_{\mathrm{E}}\big(Y_{lm}\big) \overset{(1)}{=} \mathrm{d}Y_{lm}, \qquad D_{\mathrm{E}}\big(Y_{lm}\mathrm{vol}\big) \overset{(2)}{=} \mathrm{d}^\star \star Y_{lm} \overset{(3)}{=} -\star\,\mathrm{d}Y_{lm},$$

(1) d^* of a function is zero.
(2) d of a top form is zero, and $Y\mathrm{vol} = \star Y$.
(3) $\mathrm{d}^* = -\star\,\mathrm{d}\star$ and $\star\star = 1$ on functions.

so the even eigenspace at $l(l+1)$ is mapped isomorphically onto the odd one. Only the two harmonic even forms 1 and vol are unpaired, and index $D_{\mathrm{E}} = 2 = \chi(S^2)$.
Compare with the spinor Dirac operator on the same sphere (Example T4.47): there the spectrum was $\pm(l+1)$ with *no* zero modes. The forms have D_{E}^2-eigenvalues $l(l+1)$ and two zero modes. Positive curvature kills harmonic spinors; it does not kill the Euler characteristic.

Worked example T4.63: Euler characteristics as indices

With the Betti numbers found by Hodge theory in Example T4.59 and in Chapter T2:

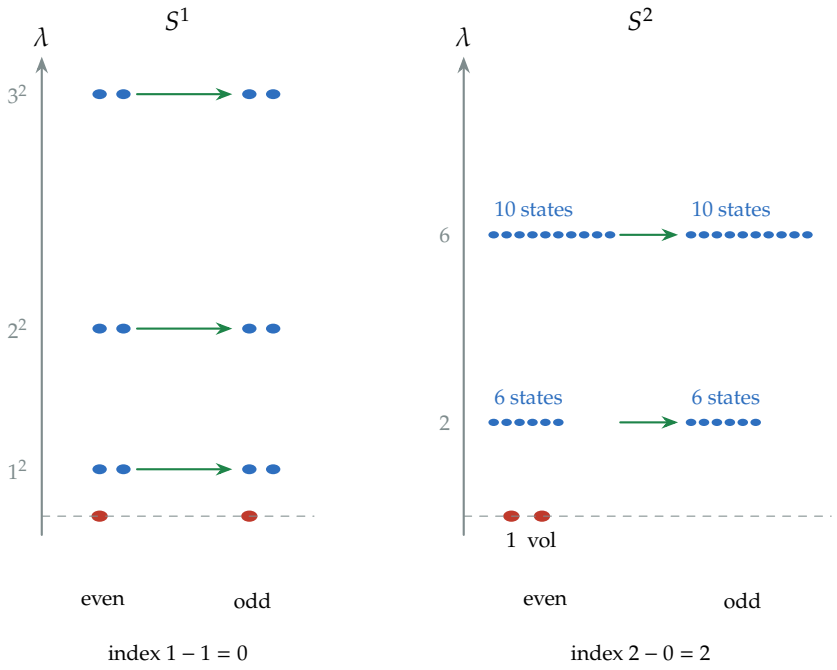


Figure T4.31: Spectrum of the Hodge Laplacian $\Delta = D_{\mathrm{E}}^* D_{\mathrm{E}} \oplus D_{\mathrm{E}} D_{\mathrm{E}}^*$ on even and odd forms. Left: the circle, levels n^2; right: the round sphere, levels $l(l+1)$ with $2(2l+1)$ states on each side. The green arrows are D_{E}, an isomorphism between the two sides at every nonzero level. The red dots are the harmonic forms. Only they can be unbalanced, and their imbalance is χ.

M	$(b_0, b_1, \ldots, b_n)$	$\chi = \operatorname{index} D_{\mathrm{E}}$	spin?
S^1	$(1,1)$	0	yes (two)
S^2	$(1,0,1)$	2	yes (one)
T^2	$(1,2,1)$	0	yes (four)
Σ_g	$(1,2g,1)$	$2-2g$	yes (2^{2g})
S^4	$(1,0,0,0,1)$	2	yes
$S^2 \times S^2$	$(1,0,2,0,1)$	4	yes
$\mathbb{CP}^2$	$(1,0,1,0,1)$	3	**no**

The last line is the reason for this whole section: $\mathbb{CP}^2$ has no Dirac operator on spinors, but its Euler operator is perfectly well defined and has index 3. In odd dimensions $\chi = 0$ always (Poincaré duality pairs b_k with b_{n-k}, which enter with opposite signs), consistent with the vanishing of the index of every differential operator in odd dimensions [Nash p. 103].

Physical picture: supersymmetric quantum mechanics on forms

The Fock-space picture of Section T4.9.1 turns the Euler operator into a quantum system. The Hilbert space is $\Omega^\bullet(M)$, the fermion number is the degree N, and

$$Q := \mathrm{d}, \qquad Q^\dagger = \mathrm{d}^*, \qquad H := \{Q, Q^\dagger\} = \mathrm{d}\mathrm{d}^* + \mathrm{d}^*\mathrm{d} = \Delta, \qquad Q^2 = 0.$$

This is **supersymmetric quantum mechanics**: Q changes fermion number by one, commutes with H, and squares to zero. The Hamiltonian is the Laplacian, a particle moving on M with fermionic partners $\varepsilon(e^a)$, $\iota(e^a)$. Every state with $E > 0$ has a partner of opposite fermion parity (namely Q or $Q^\dagger$ of it), exactly the pairing of Figure T4.31. Supersymmetric

ground states ($E = 0$) are harmonic forms, i.e. cohomology. The **Witten index** is

$$\operatorname{Tr}\big((-1)^N e^{-tH}\big) \overset{(1)}{=} \sum_{\lambda} e^{-t\lambda}\big(\dim \Omega^{\text{even}}_{\lambda} - \dim \Omega^{\text{odd}}_{\lambda}\big) \overset{(2)}{=} \chi(M), \tag{T4.129}$$

(1) Diagonalise Δ in each parity sector.
(2) Only $\lambda = 0$ survives the pairing, and there the count is (T4.127).

independent of t and of the metric: the McKean–Singer formula (T4.99) for forms. Witten's deformation $Q_s = e^{-sf}\,\mathrm{d}\,e^{sf}$ by a function f still squares to zero, so it leaves the index unchanged, and for large s the ground states localise at the critical points of f. That is the physicist's proof of the Morse inequalities (Chapter 7).

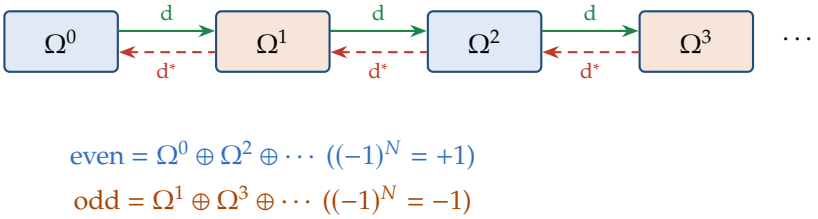


Figure T4.32: The de Rham complex as a ladder of fermion numbers. Both d (green) and d* (red, dashed) move one rung, so each maps even forms to odd forms and vice versa. That is why $(-1)^N$ anticommutes with $\mathrm{d} + \mathrm{d}^*$, and why the Euler operator D_{E} (all arrows, starting from the blue boxes) is a map from even to odd forms.

T4.9.7 The signature operator

The Euler grading uses only the degree. A four-manifold has a finer invariant that the degree cannot see. On H^2 the wedge product gives a quadratic form, $q(u,v) = \int_M u \wedge v$, and a symmetric real form has a signature: the number of positive minus the number of negative eigenvalues. For $\mathbb{CP}^2$ it is $+1$, for $S^2 \times S^2$ it is 0, for the K3 surface it is -16. Is this number also the index of a Dirac-type operator? It is, and the grading that does it is the *chirality operator of the Clifford action* c, which turns out to be a multiple of the Hodge star.

The intersection form and the signature. Let $n = 2m$. The middle cohomology $H^m(M)$ carries the pairing [Nash (4.74)]

$$q(u,v) := \int_M u \wedge v \qquad (u, v \in H^m(M;\mathbb{R})), \tag{T4.130}$$

which is well defined on classes by Stokes ($\int (u + \mathrm{d}\alpha)\wedge v = \int u\wedge v + \int \mathrm{d}(\alpha \wedge v)$ for closed v) and nondegenerate by Poincaré duality. Since $u \wedge v = (-1)^{m^2} v \wedge u$ [Nash (4.76)], q is antisymmetric when m is odd, and a real antisymmetric form has no real eigenvalues. So the **signature**

$$\sigma(M) := b^+ - b^- \qquad (\text{Nash's Sign}(X), \text{[Nash (4.75)]}) \tag{T4.131}$$

is interesting only when m is even, i.e. $4 \mid n$. Reversing the orientation flips the sign of $\int$, hence of σ.

Working definition T4.19 (the involution τ [Nash (4.77)], [Bär Ex. 1.3.19])

On complex-valued k-forms on an oriented Riemannian manifold of dimension $n = 2m$, put

$$\tau\alpha := i^{\,k(k-1)+m} \star \alpha \qquad (\alpha \in \Omega^k). \tag{T4.132}$$

[$\star$ is the Hodge star, sending a k-form to an $(n-k)$-form, e.g. $\star e^1 = e^2$ on $\mathbb{R}^2$] Then $\tau^2 = 1$, so $\Lambda^\bullet T^*M \otimes \mathbb{C} = \Lambda^+ \oplus \Lambda^-$ splits into the ± 1 eigenbundles [at each point, the forms with $\tau\alpha = +\alpha$ and those with $\tau\alpha = -\alpha$] [Nash (4.79)]. *How to compute:* apply $\star$ and multiply by the phase; the table for $n = 4$ is below.

The phase is chosen precisely to make τ an involution. Using $\star^2 = (-1)^{k(n-k)} = (-1)^k$ on k-forms in even dimension ([Bär Lemma 1.3.16], and $k(n-k) = 2mk - k^2 \equiv k \bmod 2$):

$$\tau^2\alpha \overset{(1)}{=} i^{k(k-1)+m}\,\tau(\star\alpha) \overset{(2)}{=} i^{k(k-1)+m}\,i^{(n-k)(n-k-1)+m}\,\star^2\alpha \overset{(3)}{=} i^{2k^2}\,\star^2\alpha \overset{(4)}{=} (-1)^k(-1)^k\alpha = \alpha. \tag{T4.133}$$

(1) Definition (T4.132) on the k-form α.
(2) $\star\alpha$ is an $(n-k)$-form, so τ acts on it with the phase for degree $n-k$.
(3) Exponent: $k^2 - k + (n-k)^2 - (n-k) + 2m = 2k^2 + n^2 - 2nk$ with $n = 2m$; the terms $n^2 = 4m^2$ and $2nk = 4mk$ are multiples of 4 and drop out of a power of i.
(4) $i^{2k^2} = (-1)^{k^2} = (-1)^k$, and $\star^2 = (-1)^k$.

This is Nash's (4.78), with every exponent spelled out. For $n = 4$ ($m = 2$), the phases $i^{k(k-1)+2}$ for $k = 0, \dots, 4$ are $-1, -1, +1, +1, -1$:

degree k	0	1	2	3	4
τ	$-\star$	$-\star$	$\star$	$\star$	$-\star$
$+1$ eigenforms	$f - f\,\mathrm{vol}$	$\theta - \star\theta$	self-dual	(paired with degrees 1, 0)	

In the middle degree $\tau = \star$ *exactly* whenever $4 \mid n$: the phase is $i^{m(m-1)+m} = i^{m^2} = 1$ for m even. So on 2-forms in four dimensions Λ^+ is the bundle of *self-dual* forms and Λ^- the anti-self-dual ones: the same splitting $\Lambda^2 = \Lambda^2_+ \oplus \Lambda^2_-$ that underlies instantons and $\mathrm{Spin}(4) = \mathrm{SU}(2) \times \mathrm{SU}(2)$ (Section T4.5.6, Chapter 8).

τ is the chirality operator. Why this odd phase? Because τ is the Clifford volume element. Recall that for Hermitian gammas the chirality operator is $\Gamma = (-i)^m \gamma_1 \cdots \gamma_n$ (Section T4.4.1). With the gammas $\gamma(e^a) = i\,c(e^a)$ of (T4.109) this is

$$\Gamma_\Lambda \overset{(1)}{=} (-i)^m\, i^n\, c(e^1)\cdots c(e^n) \overset{(2)}{=} i^m\, c(e^1)c(e^2)\cdots c(e^n) \;=\; \tau. \tag{T4.134}$$

(1) $\gamma(e^a) = i\,c(e^a)$, and there are n factors.
(2) $(-i)^m i^{2m} = (-i)^m(-1)^m = i^m$.

The last equality, $\tau = i^m c(e^1)\cdots c(e^n)$, is the content; it is a sign count on basis forms, and it is quickest to check it on examples.

- *On the function* 1, *any* n: each $c(e^a) = \varepsilon(e^a) - \iota(e^a)$ acting on a form not containing e^a just wedges, so $c(e^1)\cdots c(e^n)1 = e^1 \wedge \cdots \wedge e^n = \mathrm{vol}$, and $i^m \mathrm{vol} = i^{0+m} \star 1 = \tau 1$.
- *On* $\Lambda^\bullet\mathbb{R}^2$, *all four basis forms.* With the matrices of Example T4.55, $c(e^1)c(e^2) = (\mathbf{1} \otimes J)(J \otimes \sigma_3) = J \otimes J\sigma_3$ and $J\sigma_3 = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right) = \sigma_1$, so

$$i\,c(e^1)c(e^2) = i\,J \otimes \sigma_1 = \begin{pmatrix} 0 & 0 & 0 & -i \\ 0 & 0 & -i & 0 \\ 0 & i & 0 & 0 \\ i & 0 & 0 & 0 \end{pmatrix}.$$

Its columns say $1 \mapsto i\,e^{12}$, $e^1 \mapsto i\,e^2$, $e^2 \mapsto -i\,e^1$, $e^{12} \mapsto -i$. Directly from (T4.132) with $m = 1$: $\tau 1 = i \star 1 = i\,e^{12}$; $\tau e^1 = i \star e^1 = ie^2$; $\tau e^2 = i \star e^2 = -ie^1$; $\tau e^{12} = i^{2+1} \star e^{12} = -i$. They agree.

- *On e^{12} in $\mathbb{R}^4$.* Working from the right, $c(e^4)e^{12} = e^{412} = e^{124}$, $c(e^3)e^{124} = e^{3124} = e^{1234}$, $c(e^2)e^{1234} = -\iota_{e_2}e^{1234} = e^{134}$, $c(e^1)e^{134} = -e^{34}$. So $i^2 c(e^1)\cdots c(e^4)e^{12} = e^{34} = \star e^{12} = \tau e^{12}$.

Key idea: τ anticommutes with $\mathrm{d} + \mathrm{d}^*$ [Nash p. 110]

Since τ is the chirality operator of the Clifford module $(\Lambda^\bullet, c)$, it anticommutes with every $c(e^a)$; being built from the metric and orientation it is parallel. Hence

$$\tau(\mathrm{d} + \mathrm{d}^*) = -(\mathrm{d} + \mathrm{d}^*)\tau, \qquad D_S := (\mathrm{d} + \mathrm{d}^*)\big|_{\Omega^+} : \Omega^+(M) \longrightarrow \Omega^-(M), \tag{T4.135}$$

where $\Omega^\pm = \Gamma(\Lambda^\pm)$. D_S is the **signature operator** (Nash's d_+, [Nash (4.80)–(4.81)]); its adjoint is the restriction of $\mathrm{d} + \mathrm{d}^*$ to Ω^-.

It is reassuring to see the anticommutation without Clifford algebra. Separately, τ turns d into $-\mathrm{d}^*$. For $\alpha \in \Omega^k$ (so $\mathrm{d}\alpha \in \Omega^{k+1}$ and $\star\alpha \in \Omega^{n-k}$), with $\mathrm{d}^* = -\star\,\mathrm{d}\star$ in even dimension:

$$\mathrm{d}^*\tau\alpha \overset{(1)}{=} i^{k(k-1)+m}\,\mathrm{d}^* \star \alpha \overset{(2)}{=} -i^{k(k-1)+m} \star \mathrm{d} \star \star\alpha \overset{(3)}{=} -(-1)^k\, i^{k(k-1)+m} \star \mathrm{d}\alpha$$
$$\overset{(4)}{=} -i^{2k}\, i^{k(k-1)+m} \star \mathrm{d}\alpha \overset{(5)}{=} -i^{(k+1)k+m} \star \mathrm{d}\alpha \overset{(6)}{=} -\tau\mathrm{d}\alpha.$$

(1) Definition of τ on the k-form α.
(2) $\mathrm{d}^* = -\star\,\mathrm{d}\star$ on forms of any degree when n is even (T4.115).
(3) $\star\star\alpha = (-1)^k\alpha$.
(4) $(-1)^k = i^{2k}$.
(5) $2k + k(k-1) = (k+1)k$.
(6) $\mathrm{d}\alpha$ has degree $k+1$, and the phase of τ in degree $k+1$ is $i^{(k+1)k+m}$.

So $\tau\mathrm{d} = -\mathrm{d}^*\tau$; conjugating by τ (using $\tau^2 = 1$) gives $\tau\mathrm{d}^* = -\mathrm{d}\tau$, and adding, $\tau(\mathrm{d}+\mathrm{d}^*) = -(\mathrm{d}+\mathrm{d}^*)\tau$. Note that τ does *not* anticommute with d alone: it exchanges d and $-\mathrm{d}^*$. Only the Dirac-type combination is graded.

Key idea: the index of the signature operator [Nash (4.84)–(4.89)]

If $4 \mid n$,

$$\operatorname{index} D_S = h_+ - h_- = \sigma(M). \tag{T4.136}$$

Here $h_\pm$ is the dimension of the space of harmonic forms in $\Omega^\pm$.

The argument is linear algebra on harmonic forms. Since τ anticommutes with $\mathrm{d} + \mathrm{d}^*$, it commutes with $\Delta = (\mathrm{d} + \mathrm{d}^*)^2$ and maps harmonic forms to harmonic forms, taking $\mathcal{H}^p$ to $\mathcal{H}^{n-p}$. The kernel of D_S is $\mathcal{H}^\bullet \cap \Omega^+$ and the kernel of its adjoint is $\mathcal{H}^\bullet \cap \Omega^-$. Group the harmonic forms as Nash does [Nash (4.85)]:

$$\mathcal{H}^\bullet = \underbrace{\left(\mathcal{H}^0 \oplus \mathcal{H}^n\right)}_{I^0} \oplus \underbrace{\left(\mathcal{H}^1 \oplus \mathcal{H}^{n-1}\right)}_{I^1} \oplus \cdots \oplus \underbrace{\left(\mathcal{H}^{m-1} \oplus \mathcal{H}^{m+1}\right)}_{I^{m-1}} \oplus \mathcal{H}^m.$$

- *The pairs I^p ($p < m$) contribute nothing.* τ swaps the two summands. For every $h \in \mathcal{H}^p$ the forms $h \pm \tau h$ are eigenforms with eigenvalue ± 1 (as $\tau^2 = 1$), and every element of I^p is a combination of these. So $I^p = I^p_+ \oplus I^p_-$ with $\dim I^p_+ = \dim I^p_- = b_p$, and the contributions to $h_+ - h_-$ cancel.

- *The middle degree contributes $b^+ - b^-$.* On $\mathcal{H}^m$, $\tau = \star$ (the phase is 1 when m is even), which is real. Its ± 1 eigenspaces $\mathcal{H}^m_\pm$ are complexifications of real spaces of self-dual and anti-self-dual harmonic forms. For a real h with $\star h = \pm h$,

$$q(h,h) \overset{(1)}{=} \int_M h \wedge h \overset{(2)}{=} \pm \int_M h \wedge \star h \overset{(3)}{=} \pm \|h\|^2. \tag{T4.137}$$

  (1) Definition (T4.130).
  (2) $h = \pm \star h$.
  (3) Defining property of the Hodge star, $\alpha \wedge \star\beta = \langle \alpha, \beta \rangle \mathrm{vol}$.

  And $q(h_+, h_-) = \int h_+ \wedge \star h_- \cdot (-1) = -(h_+, h_-)$ while also $q(h_+, h_-) = \int \star h_+ \wedge h_- = (h_-, h_+)$, so $(h_-, h_+) = -(h_+, h_-)$; the real inner product is symmetric, so both vanish and $q(h_+, h_-) = 0$. Thus q is positive definite on $\mathcal{H}^m_+$, negative definite on $\mathcal{H}^m_-$, and the two are q-orthogonal: $\dim \mathcal{H}^m_\pm = b^\pm$.

Adding up,

$$h_+ - h_- \overset{(1)}{=} \sum_{p<m} \left(\dim I^p_+ - \dim I^p_-\right) + \dim \mathcal{H}^m_+ - \dim \mathcal{H}^m_- \overset{(2)}{=} 0 + b^+ - b^- \overset{(3)}{=} \sigma(M).$$

(1) Decompose $\mathcal{H}^\bullet$ as above; each piece is τ-invariant.
(2) The pairs cancel; the middle degree gives $b^\pm$.
(3) Definition (T4.131).

When m is odd the same bookkeeping applies, but on $\mathcal{H}^m$ we have $\tau = \pm i\star$ with $\star^2 = -1$, and complex conjugation exchanges the two eigenspaces, so $h_+ = h_-$ and the index is 0, as it should be for an antisymmetric q.

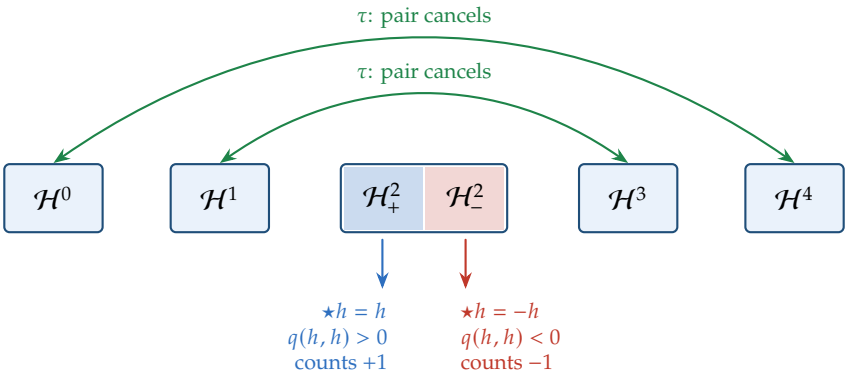


Figure T4.33: Why the signature operator counts $b^+ - b^-$ ($n = 4$). τ exchanges $\mathcal{H}^p$ and $\mathcal{H}^{4-p}$, so each such pair contains as many $\tau = +1$ as $\tau = -1$ forms and drops out of $h_+ - h_-$. In the middle degree $\tau = \star$, and a self-dual harmonic form has positive intersection with itself, an anti-self-dual one negative. The index sees only the middle.

Worked example T4.64: signatures of four-manifolds from harmonic forms

- S^4: $b_2 = 0$, so $\sigma = 0$.
- T^4 with flat coordinates $x^1, \dots, x^4$: $\mathcal{H}^2$ is spanned by the six constant forms e^{ij}. Since $\star e^{12} = e^{34}$, $\star e^{13} = -e^{24}$, $\star e^{14} = e^{23}$, the combinations

$$e^{12} \pm e^{34}, \qquad e^{13} \mp e^{24}, \qquad e^{14} \pm e^{23}$$

are self-dual (upper signs) and anti-self-dual (lower signs): $b^+ = b^- = 3$ and $\sigma(T^4) = 0$. Check with q: $\int (e^{12} + e^{34}) \wedge (e^{12} + e^{34}) = 2 \int e^{1234} > 0$.

- $S^2 \times S^2$ with the product of round metrics: $\mathcal{H}^2$ is spanned by the two area forms ω_1, ω_2 (normalised to $\int_{S^2} \omega_i = 1$). Here $\star\omega_1 = \omega_2$ (the Hodge star of the area form of one factor is the area form of the other, for equal radii), so $\omega_1 + \omega_2$ is self-dual and $\omega_1 - \omega_2$ anti-self-dual. In the basis ω_1, ω_2, $q = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ (since $\omega_i \wedge \omega_i = 0$ and $\int \omega_1 \wedge \omega_2 = 1$), with eigenvalues ± 1: $\sigma = 0$.
- $\mathbb{CP}^2$ with the Fubini–Study metric: $b_2 = 1$, and $\mathcal{H}^2$ is spanned by the Kähler form ω. On a Kähler surface ω is self-dual (in local unitary coordinates $\omega = e^{12} + e^{34}$), so $b^+ = 1$, $b^- = 0$, $\sigma(\mathbb{CP}^2) = 1$. With the opposite orientation, $\overline{\mathbb{CP}}^2$ has $\sigma = -1$.
- **K3** (quoted): $b_2 = 22$ with $b^+ = 3$, $b^- = 19$, so $\sigma = -16$.

The index formula: Hirzebruch's theorem. The Atiyah–Singer theorem computes index D_S as the integral of a characteristic class. Nash does it with the general recipe [Nash (4.90)–(4.93)]; in Section T4.9.9 we will see a shorter route through the spinor index formula. The answer, in terms of the Pontryagin roots $\pm x_i$ of Section T4.8.5, is

$$\sigma(M) = \int_M L(M), \qquad L(M) = \prod_{i=1}^{m} \frac{x_i}{\tanh x_i}\Big|_{\text{degree } n} = \left(1 + \frac{p_1}{3} + \frac{7p_2 - p_1^2}{45} + \cdots\right)_{\text{degree } n}. \tag{T4.138}$$

This is the **Hirzebruch signature theorem** [Nash (4.94)–(4.95)]. The expansion follows from $\frac{y}{\tanh y} = 1 + \frac{y^2}{3} - \frac{y^4}{45} + \cdots$ and the same symmetric-function manipulation as for $\hat{A}$ in Section T4.8.5. In four dimensions

$$\sigma(M) = \frac{1}{3} \int_M p_1(M).$$

Check on $\mathbb{CP}^2$: $p_1(\mathbb{CP}^2) = 3h^2$ with $\int h^2 = 1$ (h the hyperplane class), so $\sigma = 1$, matching Example T4.64. *Check on $S^2 \times S^2$:* the tangent bundle is a sum of two bundles pulled back from S^2, whose Pontryagin classes live in $H^4(S^2) = 0$, so $p_1 = 0$ and $\sigma = 0$. Since index $D^+ = -\frac{1}{24} \int p_1$ for spinors, this is what gave index $D^+ = -\sigma/8$ in (T4.104).

Watch out: two normalisations of the L-polynomial

Nash writes the result as index $d_+ = 2^{n/2} \mathfrak{L}(X)[X]$ with $\mathfrak{L} = \prod_i \frac{x_i/2}{\tanh(x_i/2)}$ [Nash (4.93)], and remarks that Hirzebruch's $\prod_i x_i/\tanh x_i$ gives the same top-degree term. The reason: a homogeneous polynomial of form-degree n in the x_i has polynomial degree $n/2$, so substituting $x_i \to x_i/2$ multiplies it by $2^{-n/2}$, which the prefactor $2^{n/2}$ undoes. Check in $n = 4$: $\frac{x/2}{\tanh(x/2)} = 1 + \frac{x^2}{12} + \cdots$, so $2^2 \cdot \frac{x_1^2 + x_2^2}{12} = \frac{p_1}{3}$. The lower-degree terms of the two polynomials differ, so they are not interchangeable in formulas (such as the twisted index, or the η-invariant corrections of Chapter 10) that use them.

Physical picture: why a physicist meets the signature

In four dimensions Λ^+ contains the self-dual 2-forms. The deformation theory of instantons (Chapter 8) uses exactly the complex $\Omega^0 \xrightarrow{\mathrm{d}_A} \Omega^1 \xrightarrow{\mathrm{d}_A^+} \Omega^2_+$. Untwisted, its alternating sum of cohomology dimensions is $b_0 - b_1 + b^+ = \frac{1}{2}(\chi + \sigma)$, built from the Euler and signature indices of this section; this is the term $-3(1 - b_1 + b^+)$ in the instanton moduli space dimension

$8k - 3(1 - b_1 + b^+)$. The gravitational contribution to the axial anomaly is proportional to $\int p_1$, i.e. to the signature (Chapter 10).

T4.9.8 The Dolbeault operator: holomorphic sections as zero modes

In Example T4.43 we found that a harmonic spinor on $\mathbb{R}^2$ is a pair of a holomorphic and an antiholomorphic function. On a complex manifold the natural first-order operator is $\bar\partial$, whose kernel on functions is the holomorphic functions. Is $\bar\partial$ (suitably completed) of Dirac type, and how is it related to the spinor Dirac operator? It is, and on a Kähler manifold the two are the same operator in disguise: spinors *are* $(0,q)$-forms, up to a square root of a line bundle. This explains why spin structures on a Riemann surface are counted by square roots of the canonical bundle, and why the monopole zero modes of Example T4.53 are holomorphic polynomials.

Complex forms in one line each. On a complex manifold with local holomorphic coordinates $z^j = x^j + iy^j$ ($j = 1, \dots, m$, real dimension $n = 2m$) put

$$\mathrm{d}z^j = \mathrm{d}x^j + i\,\mathrm{d}y^j, \quad \mathrm{d}\bar z^j = \mathrm{d}x^j - i\,\mathrm{d}y^j, \quad \frac{\partial}{\partial z^j} = \tfrac12\Big(\frac{\partial}{\partial x^j} - i\frac{\partial}{\partial y^j}\Big), \quad \frac{\partial}{\partial \bar z^j} = \tfrac12\Big(\frac{\partial}{\partial x^j} + i\frac{\partial}{\partial y^j}\Big).$$

Then $\mathrm{d}f = \sum_j \partial_{z^j} f\,\mathrm{d}z^j + \sum_j \partial_{\bar z^j} f\,\mathrm{d}\bar z^j =: \partial f + \bar\partial f$ [Bär (1.23)], and $\bar\partial f = 0$ is the Cauchy–Riemann system: f is holomorphic. A form with p factors $\mathrm{d}z$ and q factors $\mathrm{d}\bar z$ is a **(p,q)-form**; $\Lambda^k T^*M \otimes \mathbb{C} = \bigoplus_{p+q=k} \Lambda^{p,q}$. The exterior derivative splits as $\mathrm{d} = \partial + \bar\partial$ with $\partial : \Lambda^{p,q} \to \Lambda^{p+1,q}$ and $\bar\partial : \Lambda^{p,q} \to \Lambda^{p,q+1}$ (differentiate the coefficients; $\mathrm{d}(\mathrm{d}z^j) = \mathrm{d}(\mathrm{d}\bar z^j) = 0$). Sorting $0 = \mathrm{d}^2$ by bidegree,

$$0 = \mathrm{d}^2 = \underbrace{\partial^2}_{(p+2,q)} + \underbrace{(\partial\bar\partial + \bar\partial\partial)}_{(p+1,q+1)} + \underbrace{\bar\partial^2}_{(p,q+2)} \quad\Longrightarrow\quad \partial^2 = 0,\ \ \partial\bar\partial = -\bar\partial\partial,\ \ \bar\partial^2 = 0 \tag{T4.139}$$

[Bär (1.24)–(1.26)], because the three pieces lie in different bidegrees and must vanish separately. So each $(\Omega^{p,\bullet}, \bar\partial)$ is a complex, the **Dolbeault complex** [Nash (4.63)], with cohomology $H^{p,q}_{\bar\partial}(M)$ and **Hodge numbers** $h^{p,q} = \dim H^{p,q}_{\bar\partial}$. We take $p = 0$ as Nash does.

Metric conventions: with the Euclidean metric, $|\mathrm{d}z|^2 = |\mathrm{d}\bar z|^2 = 2$ and $\mathrm{d}z \perp \mathrm{d}\bar z$. For a real covector v write $v = v^{1,0} + v^{0,1}$; then $v^{1,0} = \overline{v^{0,1}}$ and, for real u, v,

$$\langle u, v\rangle = \langle u^{1,0}, v^{1,0}\rangle + \langle u^{0,1}, v^{0,1}\rangle = 2\,\mathrm{Re}\,\langle u^{0,1}, v^{0,1}\rangle. \tag{T4.140}$$

For example $\mathrm{d}x^{0,1} = \tfrac12 \mathrm{d}\bar z$ has $|\tfrac12 \mathrm{d}\bar z|^2 = \tfrac12 = \tfrac12 |\mathrm{d}x|^2$.

Working definition T4.20 (Clifford action on $(0,q)$-forms and the Dolbeault–Dirac operator)

On $\Lambda^{0,\bullet} = \bigoplus_q \Lambda^{0,q}$ let a real covector v act by

$$c(v) := \sqrt{2}\,\big(\varepsilon(v^{0,1}) - \iota(v^{0,1})\big), \tag{T4.141}$$

where $\iota(w) := \varepsilon(w)^\dagger$ is contraction with the metric dual of w. The **Dolbeault–Dirac operator** [Bär Ex. 1.3.20] is

$$D_{\bar\partial} := \sqrt{2}\,\big(\bar\partial + \bar\partial^*\big) : \ \Omega^{0,\mathrm{even}} \longrightarrow \Omega^{0,\mathrm{odd}}. \tag{T4.142}$$

[$\bar\partial^*$ is the adjoint of $\bar\partial$ and lowers the number of $\mathrm{d}\bar z$'s by one. $\Omega^{0,\mathrm{even}}$ holds forms built

from an even number of $\mathrm{d}\bar z$'s and no $\mathrm{d}z$] *Test:* check $\{c(u),c(v)\} = -2\langle u,v\rangle$, and that $D_{\bar\partial} = \sum_a c(e^a)\nabla_a$.

The Clifford relation follows from the canonical anticommutators, now with complex vectors. For $a,b \in \Lambda^{0,1}$ we have $\{\varepsilon(a),\varepsilon(b)\} = 0$, $\{\iota(a),\iota(b)\} = 0$ and $\{\iota(b),\varepsilon(a)\} = \langle a,b\rangle$ (the Hermitian inner product, linear in a), exactly as in (T4.105). So, with $a = u^{0,1}$ and $b = v^{0,1}$,

$$\{c(u),c(v)\} \overset{(1)}{=} 2\big(-\{\varepsilon(a),\iota(b)\} - \{\iota(a),\varepsilon(b)\}\big) \overset{(2)}{=} -2\big(\langle a,b\rangle + \langle b,a\rangle\big)$$
$$\overset{(3)}{=} -4\,\mathrm{Re}\langle u^{0,1},v^{0,1}\rangle \overset{(4)}{=} -2\langle u,v\rangle.$$

(1) Expand; the $\varepsilon\varepsilon$ and $\iota\iota$ anticommutators vanish.
(2) The mixed anticommutators are inner products.
(3) $\langle b,a\rangle = \overline{\langle a,b\rangle}$.
(4) (T4.140).

And the operator: on flat $\mathbb{C}^m$, $\mathrm{d} = \sum_a \varepsilon(e^a)\partial_a$ (T4.116); its $(0,1)$ part is $\bar\partial = \sum_a \varepsilon(e^{a\,(0,1)})\partial_a$, the adjoint is $\bar\partial^* = -\sum_a \iota(e^{a\,(0,1)})\partial_a$ (as for d^*), and so $\sqrt{2}(\bar\partial + \bar\partial^*) = \sum_a c(e^a)\partial_a$. On a curved *Kähler* manifold the same holds with $\partial_a \to \nabla_a$, because there the Levi-Civita connection preserves the splitting into (p,q) types. Squaring, using $\bar\partial^2 = 0 = (\bar\partial^*)^2$,

$$D_{\bar\partial}^2 = 2\big(\bar\partial\bar\partial^* + \bar\partial^*\bar\partial\big) = 2\Delta_{\bar\partial}, \tag{T4.143}$$

so $2\Delta_{\bar\partial}$ is of Laplace type [Bär (1.27)]; on a Kähler manifold $2\Delta_{\bar\partial} = \Delta$, the Hodge Laplacian.

Worked example T4.65: the Dolbeault–Dirac operator on $\mathbb{C}$ is the Dirac operator

On $\mathbb{C}$ with $z = x+iy$, $\Lambda^{0,\bullet}$ has orthonormal basis $(1,\ \mathrm{d}\bar z/\sqrt{2})$. Write a section as $f + h\,\mathrm{d}\bar z/\sqrt{2}$, i.e. as the column (f,h). First the adjoint of $\bar\partial$: for $g\,\mathrm{d}\bar z$,

$$(\bar\partial f, g\,\mathrm{d}\bar z) \overset{(1)}{=} \int \partial_{\bar z} f\ \bar g\ \langle \mathrm{d}\bar z, \mathrm{d}\bar z\rangle\,\mathrm{d}x\,\mathrm{d}y \overset{(2)}{=} 2\int \partial_{\bar z} f\ \bar g \overset{(3)}{=} -2\int f\ \partial_{\bar z}\bar g \overset{(4)}{=} \int f\ \overline{(-2\partial_z g)},$$

(1) $\bar\partial f = \partial_{\bar z} f\,\mathrm{d}\bar z$ and the L^2 inner product.
(2) $|\mathrm{d}\bar z|^2 = 2$.
(3) Integrate by parts (compact support).
(4) $\overline{\partial_z g} = \partial_{\bar z}\bar g$.

so $\bar\partial^*(g\,\mathrm{d}\bar z) = -2\partial_z g$. In the orthonormal basis $(g = h/\sqrt{2})$:

$$D_{\bar\partial}\begin{pmatrix} f \\ h\end{pmatrix} \overset{(1)}{=} \sqrt{2}\begin{pmatrix} -2\partial_z(h/\sqrt{2}) \\ \sqrt{2}\,\partial_{\bar z} f\end{pmatrix} \overset{(2)}{=} \begin{pmatrix} 0 & -2\partial_z \\ 2\partial_{\bar z} & 0\end{pmatrix}\begin{pmatrix} f \\ h\end{pmatrix} = \begin{pmatrix} 0 & -(\partial_x - i\partial_y) \\ \partial_x + i\partial_y & 0\end{pmatrix}\begin{pmatrix} f \\ h\end{pmatrix}. \tag{T4.144}$$

(1) $\bar\partial f = \partial_{\bar z} f\,\mathrm{d}\bar z = \sqrt{2}\,\partial_{\bar z} f\cdot(\mathrm{d}\bar z/\sqrt{2})$, and the $\bar\partial^*$ just computed.
(2) Collect.

Compare with our Dirac operator with $\gamma_1 = \sigma_1$, $\gamma_2 = \sigma_2$: $D = -i(\sigma_1\partial_x + \sigma_2\partial_y) = \left(\begin{smallmatrix} 0 & -2i\partial_z \\ -2i\partial_{\bar z} & 0\end{smallmatrix}\right)$. With the constant unitary $U = \mathrm{diag}(1,i)$,

$$UDU^{-1} = \begin{pmatrix} 0 & -2i\partial_z\cdot\frac{1}{i} \\ i\cdot(-2i)\partial_{\bar z} & 0\end{pmatrix} = \begin{pmatrix} 0 & -2\partial_z \\ 2\partial_{\bar z} & 0\end{pmatrix} = D_{\bar\partial}.$$

So on $\mathbb{C}$ the Dolbeault–Dirac operator *is* the Dirac operator, with $S^+ = \Lambda^{0,0}$ (functions) and $S^- = \Lambda^{0,1}$. Its symbol is $\gamma(\xi)$ with $\gamma(e^1) = \sigma_2$, $\gamma(e^2) = -\sigma_1$, which one checks from (T4.141)

($\gamma = ic$): $c(\mathrm{d}x) = \sqrt{2}(\varepsilon - \iota)(\frac{1}{2}\mathrm{d}\bar{z})$ sends $1 \mapsto \mathrm{d}\bar{z}/\sqrt{2}$ and $\mathrm{d}\bar{z}/\sqrt{2} \mapsto -1$, i.e. $c(\mathrm{d}x) = J$ and $ic(\mathrm{d}x) = \sigma_2$. The zero modes: $D_{\bar\partial}(f, h) = 0$ says $\partial_{\bar z} f = 0$ and $\partial_z h = 0$: f holomorphic and h antiholomorphic, the result of Example T4.43.

Worked example T4.66: $\Lambda^{0,\bullet}\mathbb{C}^2$ is the four-dimensional spinor space

On $\mathbb{C}^2$ the bundle $\Lambda^{0,\bullet}$ has basis 1, $\mathrm{d}\bar{z}^1$, $\mathrm{d}\bar{z}^2$, $\mathrm{d}\bar{z}^1 \wedge \mathrm{d}\bar{z}^2$ (normalised by powers of $\sqrt{2}$): dimension $4 = 2^{4/2}$, the size of Dirac spinors in four dimensions. The operators $\frac{1}{\sqrt{2}}\varepsilon(\mathrm{d}\bar{z}^j)$ and their adjoints obey the canonical anticommutation relations of two fermionic modes; they are the $b_j^\dagger$, b_j of Section T4.4.2, and $\Lambda^{0,\bullet}\mathbb{C}^2$ is the Fock space $\Lambda^\bullet\mathbb{C}^2$ of (T4.39). The chirality $\Gamma = (-1)^F$ is the parity of q:

$$S^+ \cong \Lambda^{0,0} \oplus \Lambda^{0,2} = \langle 1,\ \mathrm{d}\bar{z}^1 \wedge \mathrm{d}\bar{z}^2 \rangle, \qquad S^- \cong \Lambda^{0,1} = \langle \mathrm{d}\bar{z}^1,\ \mathrm{d}\bar{z}^2 \rangle.$$

In general $\dim \Lambda^{0,\bullet}\mathbb{C}^m = 2^m = 2^{n/2}$: *the Fock-space construction of spinors is the statement that spinors on $\mathbb{C}^m$ are $(0,q)$-forms.* Compare the forms of Section T4.9.1: all of $\Lambda^\bullet\mathbb{R}^{2m}$ is the Fock space of $2m$ modes; the $(0,q)$-forms use only the m antiholomorphic ones.

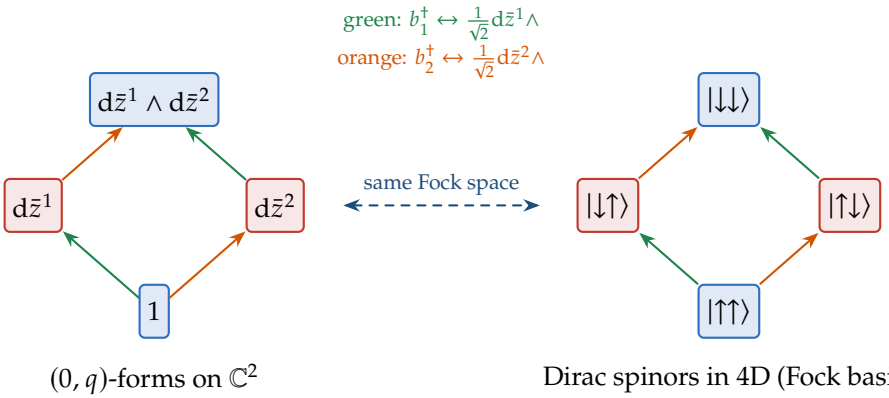


Figure T4.34: Spinors on $\mathbb{C}^2$ are $(0,q)$-forms. Left: the four basic $(0,q)$-forms, with wedging by $\mathrm{d}\bar{z}^1$ and $\mathrm{d}\bar{z}^2$ as the arrows. Right: the Fock basis of four-dimensional spinors of Section T4.4.2, with the creation operators $b_1^\dagger, b_2^\dagger$. The two squares are the same. Blue boxes are even (S^+), red boxes odd (S^-).

The global statement: spinors are $(0,q)$-forms twisted by $K^{1/2}$. Flat $\mathbb{C}^m$ hides a subtlety: under rotations, spinors and $(0,q)$-forms do *not* transform in the same way. We compute on $\mathbb{C}$. Let R_θ rotate the plane by θ (from e_1 towards e_2).

- *Spinors.* By Example T4.28 the lift is $a(\theta) = \exp(-\frac{\theta}{2}\gamma_1\gamma_2)$, and with $\gamma_1\gamma_2 = \sigma_1\sigma_2 = i\sigma_3$ this is $\mathrm{diag}(e^{-i\theta/2}, e^{i\theta/2})$: S^+ has weight $-\frac{1}{2}$, S^- weight $+\frac{1}{2}$.
- *Forms.* A rotation moves covectors like vectors, $e^1 \mapsto \cos\theta\, e^1 + \sin\theta\, e^2$, $e^2 \mapsto -\sin\theta\, e^1 + \cos\theta\, e^2$. Hence

$$\mathrm{d}z = e^1 + ie^2 \overset{(1)}{=} (\cos\theta - i\sin\theta)\, e^1 + (\sin\theta + i\cos\theta)\, e^2 \overset{(2)}{=} e^{-i\theta}\big(e^1 + ie^2\big) = e^{-i\theta}\mathrm{d}z,$$

  (1) Substitute the rotated covectors and collect.
  (2) $\sin\theta + i\cos\theta = i(\cos\theta - i\sin\theta) = ie^{-i\theta}$.

  and $\mathrm{d}\bar{z} \mapsto e^{i\theta}\mathrm{d}\bar{z}$. So $\Lambda^{0,0}$ has weight 0 and $\Lambda^{0,1}$ weight +1.

The weights of the forms are those of the spinors shifted by $+\frac{1}{2}$. The line bundle $K \coloneqq \Lambda^{m,0}$ (the **canonical bundle**; on $\mathbb{C}$ it is spanned by $\mathrm{d}z$, weight -1) has a square root of weight $-\frac{1}{2}$, and $(0, \frac{1}{2}-\frac{1}{2},\ 1-\frac{1}{2}) = (-\frac{1}{2}, \frac{1}{2})$. Globally:

Key idea: spinors on a Kähler manifold

On a Kähler manifold of complex dimension m,

$$S \ \cong\ \Lambda^{0,\bullet}T^*M \otimes K^{1/2}, \qquad S^+ \cong \Lambda^{0,\mathrm{even}} \otimes K^{1/2}, \quad S^- \cong \Lambda^{0,\mathrm{odd}} \otimes K^{1/2}, \tag{T4.145}$$

and under this isomorphism D is the Dolbeault–Dirac operator $\sqrt{2}(\bar\partial + \bar\partial^*)$ acting on $(0,q)$-forms with values in $K^{1/2}$. *Spin structures are the same thing as holomorphic square roots $K^{1/2}$ of K*; one exists iff $c_1(K)$ is even, which is $w_2 = 0$ for a complex manifold. Harmonic spinors of positive chirality are holomorphic sections of $K^{1/2}$ (in real dimension two; in general, $\bar\partial$-cohomology of $K^{1/2}$).

Worked example T4.67: Riemann surfaces: spin structures are square roots of K

On a compact Riemann surface Σ_g, $K = T^*\Sigma$ (the $(1,0)$-forms) has degree

$$\deg K \overset{(1)}{=} -\deg T\Sigma \overset{(2)}{=} -\chi(\Sigma_g) = 2g-2,$$

(1) K is dual to the holomorphic tangent bundle, and dualising flips the degree.
(2) The degree (first Chern number) of the tangent bundle is the Euler number (Gauss–Bonnet, or the Poincaré–Hopf count of zeros of a vector field).

which is even, so square roots exist: every orientable surface is spin, as we found by hand.

- $S^2 = \mathbb{CP}^1$: $K = \mathcal{O}(-2)$ (spanned by $\mathrm{d}z$, which has a double pole at $z=\infty$: $\mathrm{d}z = -w^{-2}\mathrm{d}w$ with $w = 1/z$). The square root $K^{1/2} = \mathcal{O}(-1)$ is unique, and so is the spin structure (Example T4.36). $S^+ = \mathcal{O}(-1)$ has no holomorphic sections (a negative-degree bundle has none). The negative-chirality zero modes are the Dolbeault cohomology $H^{0,1}(K^{1/2})$, and Serre duality ($H^1(L) \cong H^0(K\otimes L^{-1})^*$) turns this into $H^0(K^{1/2})^*$, which is zero again: *no harmonic spinors*, as Lichnerowicz said. Twisting by the monopole bundle $L_k = \mathcal{O}(k)$ gives $S^+ \otimes L_k = \mathcal{O}(k-1)$, with the k holomorphic sections $1, z, \ldots, z^{k-1}$ (Example T4.53).

- $T^2 = \mathbb{C}/\Lambda$: K is trivial ($\mathrm{d}z$ is a global nowhere-zero section). Its square roots are the flat line bundles whose holonomy around each of the two cycles is ±1: $2^2 = 4$ of them, the four spin structures of Example T4.35. Only the trivial root (periodic–periodic) has a holomorphic section (the constant), matching the two zero modes, one of each chirality, of Example T4.45.

- **Genus** g: $K^{1/2}$ has degree $g-1$, and the square roots differ by the 2^{2g} flat bundles with holonomies ±1: $2^{2g} = |H^1(\Sigma_g;\mathbb{Z}_2)|$ spin structures, Nash's count [Nash p. 113]. Algebraic geometers call them **theta characteristics**. The number of harmonic spinors $h^0(K^{1/2})$ depends on the choice (and the complex structure), but its parity does not: $2^{g-1}(2^g+1)$ spin structures are "even" and $2^{g-1}(2^g-1)$ "odd". String theorists sum over all of them (Chapter 9).

- $\mathbb{CP}^2$: $K = \mathcal{O}(-3)$, of odd degree on a line, has no square root: $\mathbb{CP}^2$ is not spin (Example T4.37). But $\Lambda^{0,\bullet}$ is still a perfectly good Clifford module and $D_{\bar\partial}$ exists. Formally $\Lambda^{0,\bullet} = S \otimes K^{-1/2}$: neither factor exists alone, but the sign ambiguities of the two cancel in the product, as they did for $S\otimes S^*$. Such a pair is a **spin^c structure**; every complex manifold has one.

Watch out: the index of $\bar\partial$ on $K^{1/2}$ is always zero on a surface

Riemann–Roch for a holomorphic line bundle L on Σ_g says $h^0(L) - h^1(L) = \deg L + 1 - g$. For $L = K^{1/2}$ this is $(g-1) + 1 - g = 0$, whatever the spin structure. This is the statement index $D^+ = \int \hat A = 0$ in two dimensions (Example T4.54): zero modes can occur, but always in chirality pairs.

Key idea: the index of the Dolbeault operator: Riemann–Roch [Nash (4.64)–(4.72)]

On a compact complex manifold,

$$\operatorname{index} D_{\bar\partial} = \sum_q (-1)^q\, h^{0,q} =: \chi(M,\mathcal{O}) = \int_M \operatorname{td}(M), \tag{T4.146}$$

$$\operatorname{td} = \prod_i \frac{x_i}{1-e^{-x_i}} = 1 + \frac{c_1}{2} + \frac{c_1^2 + c_2}{12} + \cdots$$

where now $x_1, \dots, x_m$ are the Chern roots of the holomorphic tangent bundle. The integer $\chi(M,\mathcal{O})$ is the **arithmetic genus** [Nash (4.65)]; the equality with $\int \operatorname{td}$ is the Hirzebruch–Riemann–Roch theorem [Nash (4.72)].

The first equality is again Hodge theory: $\operatorname{Ker} D_{\bar\partial}$ on $\Omega^{0,q}$ is the space of $\Delta_{\bar\partial}$-harmonic $(0,q)$-forms, which represents $H^{0,q}_{\bar\partial}$ (the Dolbeault version of (T4.122)), and the even and odd q are separated by the grading.

Worked example T4.68: arithmetic genera

- Σ_g. $h^{0,0} = 1$ (holomorphic functions on a compact surface are constant) and $h^{0,1} = g$. So
$$\operatorname{index} D_{\bar\partial} \overset{(1)}{=} 1 - g \overset{(2)}{=} \int_\Sigma \frac{c_1}{2} \overset{(3)}{=} \frac{\chi(\Sigma_g)}{2} = \frac{2-2g}{2}.$$
  (1) Hodge numbers.
  (2) The degree-2 part of td.
  (3) $c_1(T\Sigma)$ is the Euler class; its integral is χ.

- $\mathbb{CP}^2$. The total Chern class is $c(T\mathbb{CP}^2) = (1+h)^3$, so $c_1 = 3h$, $c_2 = 3h^2$, with $\int h^2 = 1$:
$$\int_{\mathbb{CP}^2} \operatorname{td} \overset{(1)}{=} \int \frac{c_1^2 + c_2}{12} \overset{(2)}{=} \frac{9+3}{12} = 1 \overset{(3)}{=} h^{0,0} - h^{0,1} + h^{0,2} = 1 - 0 + 0.$$
  (1) Degree-4 part of td.
  (2) $c_1^2 = 9h^2$, $c_2 = 3h^2$.
  (3) Hodge numbers of $\mathbb{CP}^2$: only $h^{0,0} = h^{1,1} = h^{2,2} = 1$ are nonzero.

  A non-spin manifold with a perfectly integral Dirac-type index.

- **K3.** K is trivial, so $K^{1/2}$ is trivial and $S = \Lambda^{0,\bullet}$ exactly. The Hodge numbers $h^{0,0} = 1$, $h^{0,1} = 0$, $h^{0,2} = 1$ give index $D_{\bar\partial} = 2$, which must equal index $D^+ = \hat A(\mathrm{K3}) = -\sigma/8 = 2$ (Example T4.54). The two harmonic spinors are the constant function 1 and the holomorphic 2-form (the conjugate of the harmonic $(0,2)$-form $\overline{\Omega}$).

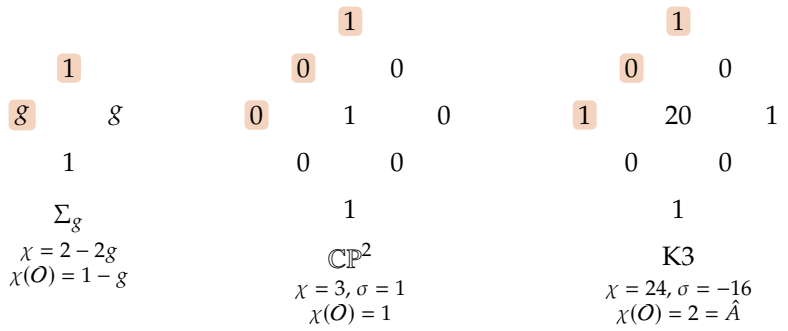


Figure T4.35: Hodge diamonds: $h^{p,q}$ sits in row $p+q$ (top row $p+q=0$), with $h^{0,q}$ on the upper-left edge (shaded). Row sums are Betti numbers, so the alternating sum of all entries is χ [Nash (4.73)]. The alternating sum along the shaded edge is the arithmetic genus $\chi(\mathcal{O}) = \operatorname{index} D_{\bar\partial}$. For K3 the shaded edge is $1, 0, 1$, which is the number $\hat A = 2$ of harmonic spinors.

T4.9.9 One Dirac operator, four indices

Nash computes the indices of the Euler, Dolbeault and signature operators by three separate manipulations of characteristic classes [Nash (4.55)–(4.62), (4.67)–(4.72), (4.90)–(4.93)], and the spinor index by a fourth [Nash (4.107)–(4.111)]. Each of them needs a "standard linear-algebraic fact" about exterior powers, and for the signature he quotes the key identity (4.91) with the words "this computation requires the combining of the splitting principle with a little group theory for SO(n)". The bispinor picture of Section T4.9.2 supplies that group theory in a few lines, and does more: it shows that, on a spin manifold, *all four operators are the one spin Dirac operator, twisted by different bundles*. Then the single formula $\operatorname{index} D^+_W = \int \hat A \operatorname{ch} W$ of (T4.103) produces the Euler class, the L-polynomial and the Todd class by one-line algebra. The questions are: which twisting bundle W, and which grading?

Step 1: a second Clifford action that commutes with the first. The two actions c and $\hat c$ on forms *anticommute*. A tensor product $S \otimes W$ has two actions that *commute*, one on each factor. To see $\Lambda^\bullet$ as a tensor product we must turn one anticommuting family into a commuting one. The chirality τ does it.

Working definition T4.21 (the twisting Clifford action)

Let $n = 2m$ and let $\tau = i^m\, c(e^1)\cdots c(e^n)$ be the chirality operator (T4.134) [$\tau^2 = 1$, and τ anticommutes with every $c(u)$ but commutes with every $\hat c(u)$]. For $v \in V$ put

$$w(v) := \hat c(v)\,\tau. \tag{T4.147}$$

[$\hat c(v) = \varepsilon(v) + \iota(v)$, the Hermitian Clifford action on forms] *Test:* $w(v)$ must commute with every $c(u)$, be Hermitian, and obey $\{w(u), w(v)\} = 2\langle u, v\rangle$.

First, τ commutes with every $\hat c(v)$: it is a product of an even number n of operators $c(e^a)$, each of which anticommutes with $\hat c(v)$ by (T4.108), so moving $\hat c(v)$ through it costs $(-1)^n = 1$. It anticommutes with every $c(u)$: that is what a chirality operator does (Section T4.9.7). With these two facts,

$$\begin{aligned} c(u)\,w(v) &\overset{(1)}{=} c(u)\,\hat c(v)\,\tau \overset{(2)}{=} -\hat c(v)\,c(u)\,\tau \overset{(3)}{=} \hat c(v)\,\tau\,c(u) = w(v)\,c(u), \\ w(u)w(v) + w(v)w(u) &\overset{(4)}{=} \hat c(u)\hat c(v)\,\tau^2 + \hat c(v)\hat c(u)\,\tau^2 \overset{(5)}{=} \{\hat c(u), \hat c(v)\} \overset{(6)}{=} 2\langle u, v\rangle. \end{aligned} \tag{T4.148}$$

(1) Definition (T4.147).

(2) $\{c(u), \hat{c}(v)\} = 0$.
(3) τ anticommutes with $c(u)$: the second sign cancels the first.
(4) Move the middle τ to the right through $\hat{c}(v)$ (they commute).
(5) $\tau^2 = 1$ (T4.133).
(6) (T4.108).

Hermiticity: $c(e^a)$ is anti-Hermitian with $c(e^a)^2 = -1$, so it is unitary, and so is τ; a unitary involution is Hermitian. Then $w(v)^\dagger = \tau^\dagger \hat{c}(v)^\dagger = \tau\, \hat{c}(v) = \hat{c}(v)\, \tau = w(v)$.

So on forms we now have *two commuting sets of Hermitian gamma matrices*,

$$\gamma(v) = i\, c(v) \quad \text{and} \quad w(v) = \hat{c}(v)\, \tau, \qquad [\gamma(u), w(v)] = 0,$$

each generating a copy of $\mathrm{Cl}(V) \otimes \mathbb{C} \cong \mathrm{End}(\mathbb{C}^{2^m})$ (Section T4.4.3). Their chirality operators are

$$\Gamma_S := (-i)^m \gamma(e^1) \cdots \gamma(e^n) = \tau, \qquad \Gamma_W := (-i)^m\, w(e^1) \cdots w(e^n) = (-i)^m\, \hat{c}(e^1) \cdots \hat{c}(e^n). \tag{T4.149}$$

The first is (T4.134). For the second, move all the τ's to the right through the $\hat{c}$'s (they commute) and use $\tau^n = 1$.

Key idea: forms are spinors tensor spinors

For $n = 2m$ there is an isomorphism

$$\Lambda^\bullet V \otimes \mathbb{C} \;\cong\; S \otimes W, \qquad \gamma(v) \leftrightarrow \gamma_S(v) \otimes \mathbf{1}, \qquad w(v) \leftrightarrow \mathbf{1} \otimes \gamma_W(v), \tag{T4.150}$$

where S and W are two copies of the 2^m-dimensional spinor space. The chirality of the first factor is τ and that of the second is Γ_W.

Why it is true. Pick a joint eigenvector-type decomposition: as a module for the γ's alone, $\Lambda^\bullet$ is a sum of copies of the unique irreducible module S (Section T4.4.3); there are $2^n/2^m = 2^m$ copies. The operators that commute with all the γ's are then exactly the $2^m \times 2^m$ matrices that mix the copies (Schur's lemma), an algebra of dimension 4^m. The w's commute with the γ's and generate an algebra of that dimension, so they generate *all* of it. Labelling the copies by a vector space W, the w's act on W, and a 2^m-dimensional module of $\mathrm{Cl}(V) \otimes \mathbb{C}$ is the spinor module. This is the precise form of "a form is a spinor with an extra spinor index".

Step 2: the Euler grading in the new language. The degree parity $(-1)^N$ anticommutes with all $2n$ generators $c, \hat{c}$. So does the product $\Gamma_S \Gamma_W$, since Γ_S anticommutes with the γ's and commutes with the w's, and vice versa. They are proportional, and the constant is found from a one-mode identity. Write $n_a := \varepsilon(e^a)\iota(e^a)$ for the occupation number of mode a (1 on forms containing e^a, 0 otherwise). Then

$$c(e^a)\, \hat{c}(e^a) \overset{(1)}{=} \big(\varepsilon_a - \iota_a\big)\big(\varepsilon_a + \iota_a\big) \overset{(2)}{=} \varepsilon_a \iota_a - \iota_a \varepsilon_a \overset{(3)}{=} n_a - (1 - n_a) = -(-1)^{n_a}, \tag{T4.151}$$

(1) Abbreviate $\varepsilon_a = \varepsilon(e^a)$, $\iota_a = \iota(e^a)$; use (T4.107).
(2) $\varepsilon_a^2 = 0$ and $\iota_a^2 = 0$ (Pauli exclusion).
(3) $\iota_a \varepsilon_a = 1 - \varepsilon_a \iota_a$ by the anticommutation relation (T4.105).

since $2n_a - 1$ is $+1$ on an occupied mode and -1 on an empty one. Now multiply over all modes:

$$\begin{aligned} \Gamma_S \Gamma_W &\overset{(1)}{=} i^m (-i)^m\, c_1 c_2 \cdots c_n\, \hat{c}_1 \hat{c}_2 \cdots \hat{c}_n \overset{(2)}{=} (-1)^{n(n-1)/2}\, (c_1 \hat{c}_1)(c_2 \hat{c}_2) \cdots (c_n \hat{c}_n) \\ &\overset{(3)}{=} (-1)^m \prod_{a=1}^{n} \big(-(-1)^{n_a}\big) \overset{(4)}{=} (-1)^m\, (-1)^N. \end{aligned} \tag{T4.152}$$

(1) (T4.149), with $c_a = c(e^a)$, $\hat{c}_a = \hat{c}(e^a)$.
(2) Move $\hat{c}_1$ left past $c_n, \dots, c_2$ ($n-1$ swaps), then $\hat{c}_2$ past $c_n, \dots, c_3$ ($n-2$ swaps), and so on: $n(n-1)/2$ swaps, each a sign.
(3) $n(n-1)/2 = m(2m-1) \equiv m \pmod 2$, and (T4.151) for each pair.
(4) There are n (an even number of) minus signs, and $\sum_a n_a = N$.

So

$$(-1)^N = (-1)^m \, \Gamma_S \, \Gamma_W, \qquad \tau = \Gamma_S. \tag{T4.153}$$

The signature grading is the chirality of the *first* factor alone; the Euler grading is the product of the chiralities of *both* factors, up to the sign $(-1)^m$. This single line contains the whole difference between the two complexes.

Worked example T4.69: $\Lambda^\bullet \mathbb{R}^2 = S \otimes W$ with explicit matrices

Use the Kronecker forms (T4.110): $c(e^1) = \mathbf{1} \otimes J$, $c(e^2) = J \otimes \sigma_3$, $\hat{c}(e^1) = \mathbf{1} \otimes \sigma_1$, $\hat{c}(e^2) = \sigma_1 \otimes \sigma_3$, with $J = -i\sigma_2$, and recall $\tau = ic(e^1)c(e^2) = \sigma_2 \otimes \sigma_1$ (Section T4.9.7). The twisting gammas are

$$w(e^1) \overset{(1)}{=} (\mathbf{1} \otimes \sigma_1)(\sigma_2 \otimes \sigma_1) \overset{(2)}{=} \sigma_2 \otimes \mathbf{1},$$
$$w(e^2) \overset{(3)}{=} (\sigma_1 \otimes \sigma_3)(\sigma_2 \otimes \sigma_1) \overset{(4)}{=} \sigma_1\sigma_2 \otimes \sigma_3\sigma_1 \overset{(5)}{=} (i\sigma_3) \otimes (i\sigma_2) = -\sigma_3 \otimes \sigma_2.$$

(1) Definition (T4.147) with $v = e^1$.
(2) $(A \otimes B)(C \otimes D) = AC \otimes BD$ and $\sigma_1^2 = \mathbf{1}$.
(3) Definition with $v = e^2$.
(4) Multiply slot by slot.
(5) $\sigma_1\sigma_2 = i\sigma_3$ and $\sigma_3\sigma_1 = i\sigma_2$.

Check against the gammas $\gamma(e^1) = \mathbf{1} \otimes \sigma_2$, $\gamma(e^2) = \sigma_2 \otimes \sigma_3$ of Example T4.55. Two Kronecker products commute when the number of slots in which the factors anticommute is even. $w(e^1) = \sigma_2 \otimes \mathbf{1}$ against $\mathbf{1} \otimes \sigma_2$: zero slots; against $\sigma_2 \otimes \sigma_3$: zero slots. $w(e^2) = -\sigma_3 \otimes \sigma_2$ against $\mathbf{1} \otimes \sigma_2$: zero slots; against $\sigma_2 \otimes \sigma_3$: two slots. All four commute. And $w(e^1)$, $w(e^2)$ anticommute with each other (one slot, σ_2 against σ_3) and square to $\mathbf{1}$.
The two chiralities are $\Gamma_S = \tau = \sigma_2 \otimes \sigma_1$ and

$$\Gamma_W \overset{(1)}{=} -i \, w(e^1)w(e^2) \overset{(2)}{=} i \, \sigma_2\sigma_3 \otimes \sigma_2 \overset{(3)}{=} i(i\sigma_1) \otimes \sigma_2 = -\sigma_1 \otimes \sigma_2,$$

(1) (T4.149) with $m = 1$.
(2) $w(e^1)w(e^2) = -(\sigma_2 \otimes \mathbf{1})(\sigma_3 \otimes \sigma_2) = -\sigma_2\sigma_3 \otimes \sigma_2$.
(3) $\sigma_2\sigma_3 = i\sigma_1$.

and $\Gamma_S\Gamma_W = -(\sigma_2\sigma_1) \otimes (\sigma_1\sigma_2) = -(-i\sigma_3) \otimes (i\sigma_3) = -\sigma_3 \otimes \sigma_3$. Since $(-1)^N = \mathrm{diag}(1,-1,-1,1) = \sigma_3 \otimes \sigma_3$, this is $(-1)^N = -\Gamma_S\Gamma_W$: (T4.153) with $m = 1$.
Now find the four joint eigenvectors. On even forms $\Gamma_W = -\Gamma_S(-1)^N = -\tau$, and on odd forms $\Gamma_W = +\tau$. The eigenvectors of τ were found in Section T4.9.7: $\tau(1 \pm i\,e^{12}) = \pm(1 \pm i\,e^{12})$ and $\tau(e^1 \pm i\,e^2) = \pm(e^1 \pm i\,e^2)$. With $\mathrm{d}z = e^1 + ie^2$ and $\mathrm{d}\bar{z} = e^1 - ie^2$:

form	$\mathrm{d}z$	$1 + i\,e^{12}$	$1 - i\,e^{12}$	$\mathrm{d}\bar{z}$
$\Gamma_S = \tau$	$+1$	$+1$	-1	-1
Γ_W	$+1$	-1	$+1$	-1
sits in	$S^+ \otimes W^+$	$S^+ \otimes W^-$	$S^- \otimes W^+$	$S^- \otimes W^-$
$(-1)^N$	-1	$+1$	$+1$	-1

Every one of the four lines $S^\pm \otimes W^\pm$ is spanned by one explicit form. The Euler grading picks the anti-diagonal ($S^+ \otimes W^- \oplus S^- \otimes W^+$ are the even forms); the signature grading picks the top row ($S^+ \otimes W$). See Figure T4.36.

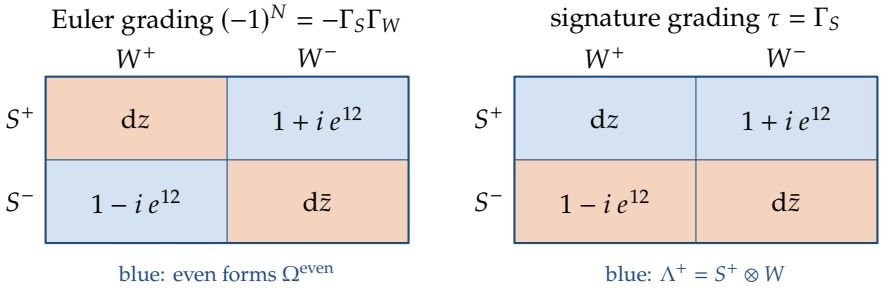


Figure T4.36: The four forms on $\mathbb{R}^2$ that diagonalise both chiralities, placed in the grid $S^\pm \otimes W^\pm$. Left: the degree parity is a *checkerboard*, because it is the product of the row and column chiralities (with a sign). Right: the signature grading τ sees only the row. So the Euler operator is the Dirac operator twisted by the *graded* bundle $W^+ \ominus W^-$, and the signature operator is the Dirac operator twisted by the whole of W.

Step 3: the connection splits too. To identify $\mathrm{d} + \mathrm{d}^*$ with a twisted Dirac operator we need more than algebra at a point: the Levi-Civita connection on forms must be the product of spin connections on S and W. An infinitesimal rotation $A_{ab} = -A_{ba}$ acts on forms by the derivation

$$\mathcal{D}_A := \sum_{a,b} A_{ab}\, \varepsilon(e^a)\, \iota(e^b),$$

which rotates each 1-form, $\mathcal{D}_A e^c = \sum_a A_{ac} e^a$, and obeys the Leibniz rule on wedge products (because ι is an antiderivation and $\varepsilon(e^a)$ moves past a k-form with the sign $(-1)^k$). The Levi-Civita connection on forms is $\nabla = \mathrm{d} + \mathcal{D}_\omega$ in an orthonormal coframe, since $\nabla e^c = -\omega^c{}_b e^b = \omega_{bc} e^b$ by (T4.66). Now rewrite $\mathcal{D}_A$ with $\varepsilon = \tfrac{1}{2}(\hat{c} + c)$ and $\iota = \tfrac{1}{2}(\hat{c} - c)$:

$$\begin{aligned}\mathcal{D}_A &\overset{(1)}{=} \frac{1}{4}\sum_{a,b} A_{ab}\big(\hat{c}^a + c^a\big)\big(\hat{c}^b - c^b\big) \overset{(2)}{=} \frac{1}{4}\sum_{a,b} A_{ab}\big(\hat{c}^a\hat{c}^b - c^a c^b\big) + \frac{1}{4}\sum_{a,b} A_{ab}\big(c^a\hat{c}^b - \hat{c}^a c^b\big)\\ &\overset{(3)}{=} \frac{1}{4}\sum_{a,b} A_{ab}\big(\hat{c}^a\hat{c}^b - c^a c^b\big) \overset{(4)}{=} \frac{1}{4}\sum_{a,b} A_{ab}\,\gamma^a\gamma^b + \frac{1}{4}\sum_{a,b} A_{ab}\,w^a w^b. \end{aligned} \tag{T4.154}$$

(1) Write $c^a = c(e^a)$, $\hat{c}^a = \hat{c}(e^a)$ and substitute.
(2) Multiply out and group.
(3) $\hat{c}^a c^b = -c^b\hat{c}^a$, so $c^a\hat{c}^b - \hat{c}^a c^b = c^a\hat{c}^b + c^b\hat{c}^a$ is symmetric in a, b, and its contraction with the antisymmetric A_{ab} vanishes.
(4) $\gamma = ic$ gives $-c^a c^b = \gamma^a\gamma^b$; and $w^a w^b = \hat{c}^a\tau\hat{c}^b\tau = \hat{c}^a\hat{c}^b\tau^2 = \hat{c}^a\hat{c}^b$.

With $A = \omega$ the first term is exactly the spin connection $\Omega = \frac{1}{4}\omega_{ab}\gamma^a\gamma^b$ of (T4.69), acting on S; the second is the same expression in the w's, the spin connection of W. So

$$\nabla^\Lambda = \nabla^S \otimes \mathbf{1} + \mathbf{1} \otimes \nabla^W, \qquad \mathrm{d} + \mathrm{d}^* = \sum_a c(e^a)\nabla_a = -i\gamma^a\nabla_a = D_W,$$

the Dirac operator of S twisted by W with its own spin connection (T4.94). On a spin manifold W is glued with the same lifted transition functions as S, so as bundles with connection and grading $W \cong S$, and

$$\Lambda^\bullet T^*M \otimes \mathbb{C} \;\cong\; S \otimes S, \qquad \mathrm{d} + \mathrm{d}^* = D_S \;\text{(Dirac twisted by spinors)}. \tag{T4.155}$$

Step 4: read off the indices. Everything now follows from $\operatorname{index} D_W^+ = \int \hat{A}\,\operatorname{ch} W$ and the spinor character $\operatorname{ch} S^+ - \operatorname{ch} S^- = (-1)^m \prod_i (e^{x_i/2} - e^{-x_i/2})$, where the sign is the one left open in (T4.100): in our conventions $c_1(S^\pm) = \mp\frac{x}{2}$ in two dimensions, so the sign is -1 per plane. Also $\operatorname{ch} S = \operatorname{ch} S^+ + \operatorname{ch} S^- = \prod_i (e^{x_i/2} + e^{-x_i/2})$ (all 2^m weights $\frac{1}{2}(\pm x_1 \pm \cdots \pm x_m)$ with a plus sign).

The Euler operator. By (T4.153) the even forms are $S^+ \otimes W^\epsilon \oplus S^- \otimes W^{-\epsilon}$ with $\epsilon = (-1)^m$, and D_W acts only on the S factor. So

$$\begin{aligned}\operatorname{index} D_{\mathrm{E}} &\overset{(1)}{=} \big(n_+(W^\epsilon) - n_-(W^{-\epsilon})\big) - \big(n_-(W^\epsilon) - n_+(W^{-\epsilon})\big) \overset{(2)}{=} \operatorname{index} D^+_{W^\epsilon} - \operatorname{index} D^+_{W^{-\epsilon}} \\ &\overset{(3)}{=} (-1)^m \int_M \hat{A}(M)\,\big(\operatorname{ch} W^+ - \operatorname{ch} W^-\big) \overset{(4)}{=} \int_M \prod_{i=1}^m \frac{x_i/2}{\sinh(x_i/2)}\big(e^{x_i/2} - e^{-x_i/2}\big) \\ &\overset{(5)}{=} \int_M \prod_{i=1}^m x_i \overset{(6)}{=} \int_M e(TM). \end{aligned} \tag{T4.156}$$

(1) The index counts zero modes on even forms minus zero modes on odd forms; $n_\pm(W^{\pm\epsilon})$ is the number of zero modes of D_W on $S^\pm \otimes W^{\pm\epsilon}$, and the odd forms are $S^- \otimes W^\epsilon \oplus S^+ \otimes W^{-\epsilon}$.
(2) Regroup by twisting bundle: $n_+(W^\epsilon) - n_-(W^\epsilon) = \operatorname{index} D^+_{W^\epsilon}$, and likewise for $W^{-\epsilon}$.
(3) The twisted index formula (T4.103) for each; the difference is ϵ times $(\operatorname{ch} W^+ - \operatorname{ch} W^-)$.
(4) $W \cong S$ as graded bundles, so $\operatorname{ch} W^+ - \operatorname{ch} W^- = (-1)^m \prod (e^{x_i/2} - e^{-x_i/2})$, and $(-1)^m(-1)^m = 1$; $\hat{A}$ from (T4.102).
(5) $\frac{x/2}{\sinh(x/2)} \cdot 2\sinh(x/2) = x$, one plane at a time.
(6) $\prod_i x_i$ is the Euler class in terms of the Pontryagin roots (Section T4.8.5).

This is the Gauss–Bonnet–Chern theorem $\chi(M) = \int_M e(TM)$ [Nash (4.53)–(4.54)], obtained without Nash's detour through $c_n(TX_{\mathbb{C}}) = (-1)^{n/2} e^2$ [Nash (4.61)–(4.62)]. The mechanism is visible: the sinh in the denominator of $\hat{A}$, which comes from the spinors, is cancelled by the sinh in the character of the *graded* twisting spinors. Every trace of the spin structure disappears from the answer, as it must, because χ is defined on every manifold.

The signature operator. Now the grading is $\tau = \Gamma_S \otimes \mathbf{1}$, so $\Omega^+ = \Gamma(S^+ \otimes W)$ and $D_{\mathrm{S}} = D^+_W$ with the *whole* of W:

$$\begin{aligned}\operatorname{index} D_{\mathrm{S}} &\overset{(1)}{=} \int_M \hat{A}(M)\,\operatorname{ch} W \overset{(2)}{=} \int_M \prod_{i=1}^m \frac{x_i/2}{\sinh(x_i/2)}\, 2\cosh(x_i/2) \overset{(3)}{=} \int_M \prod_{i=1}^m \frac{x_i}{\tanh(x_i/2)} \\ &\overset{(4)}{=} 2^m \int_M \prod_{i=1}^m \frac{x_i/2}{\tanh(x_i/2)} \overset{(5)}{=} \int_M \prod_{i=1}^m \frac{x_i}{\tanh x_i} = \sigma(M). \end{aligned} \tag{T4.157}$$

(1) The twisted index formula (T4.103) with the ungraded twist W.
(2) $\operatorname{ch} W = \operatorname{ch} S = \prod_i (e^{x_i/2} + e^{-x_i/2})$ and $e^y + e^{-y} = 2\cosh y$.
(3) $\cosh/\sinh = 1/\tanh$, and $\frac{x}{2} \cdot 2 = x$.
(4) Pull a factor 2 out of each of the m numerators.
(5) In top degree $n = 2m$ (polynomial degree m in the x_i) the substitution $x_i \to x_i/2$ multiplies by 2^{-m}, which cancels the prefactor (the pitfall after (T4.138)); then Hirzebruch's theorem.

The penultimate expression is Nash's $2^{n/2}\mathfrak{L}(X)[X]$ [Nash (4.92)–(4.93)] exactly. Two bonuses come for free. First, $x/\tanh(x/2)$ is an *even* function of x, so only degrees divisible by 4 survive, and the index vanishes when $n \equiv 2 \pmod 4$: the analytic shadow of the antisymmetry of the intersection form [Nash (4.76)]. Second, in four dimensions, one plane gives $\frac{x}{\tanh(x/2)} = 2 + \frac{x^2}{6} + \cdots$, so the degree-4 part of the product of two is $2 \cdot \frac{x_1^2 + x_2^2}{6} = \frac{p_1}{3}$: $\sigma = \frac{1}{3}\int p_1$, as in Section T4.9.7.

The Dolbeault operator. On a Kähler manifold, (T4.145) says $\Lambda^{0,\bullet} = S \otimes K^{-1/2}$ with the grading $\Gamma_S \otimes \mathbf{1}$, and $D_{\bar\partial}$ is D twisted by the line bundle $K^{-1/2}$. Its first Chern class is $-\frac12 c_1(K) = \frac12 c_1(M) = \frac12 \sum_i x_i$, where the x_i are the Chern roots of the holomorphic tangent bundle (and also the Pontryagin roots, since $TM \otimes \mathbb{C} = T^{1,0} \oplus T^{0,1}$ has roots $\pm x_i$). So

$$\operatorname{index} D_{\bar\partial} \overset{(1)}{=} \int_M \hat A(M)\, e^{c_1(M)/2} \overset{(2)}{=} \int_M \prod_{i=1}^m \frac{x_i/2}{\sinh(x_i/2)}\, e^{x_i/2} \overset{(3)}{=} \int_M \prod_{i=1}^m \frac{x_i\, e^{x_i/2}}{e^{x_i/2} - e^{-x_i/2}}$$
$$\overset{(4)}{=} \int_M \prod_{i=1}^m \frac{x_i}{1 - e^{-x_i}} \overset{(5)}{=} \int_M \operatorname{td}(M). \tag{T4.158}$$

(1) The twisted index formula with $\operatorname{ch} K^{-1/2} = e^{c_1(K^{-1/2})}$.
(2) $e^{\sum x_i/2} = \prod e^{x_i/2}$, and (T4.102).
(3) $2\sinh(x/2) = e^{x/2} - e^{-x/2}$.
(4) Divide numerator and denominator by $e^{x/2}$.
(5) The Todd class (T4.146).

This is Hirzebruch–Riemann–Roch [Nash (4.71)–(4.72)], again as a consequence of the spinor formula. The identity behind it, $\operatorname{td} = \hat A\, e^{c_1/2}$, holds as an identity of characteristic classes on *every* complex manifold, spin or not.

Worked example T4.70: non-integers that add up: $\mathbb{CP}^2$ and $\operatorname{td} = \hat A\, e^{c_1/2}$

On $\mathbb{CP}^2$, with h the hyperplane class, $c_1 = 3h$ and $p_1 = 3h^2$, and $\int h^2 = 1$. Then

$$\int_{\mathbb{CP}^2} \hat A\, e^{c_1/2} \overset{(1)}{=} \int \Big(1 - \frac{3h^2}{24}\Big)\Big(1 + \frac{3h}{2} + \frac{9h^2}{8}\Big) \overset{(2)}{=} \frac98 - \frac18 = 1 \overset{(3)}{=} \int_{\mathbb{CP}^2} \operatorname{td}.$$

(1) $\hat A = 1 - \frac{p_1}{24}$ and $e^{3h/2} = 1 + \frac{3h}{2} + \frac12\big(\frac{3h}{2}\big)^2 + \cdots$.
(2) Keep the h^2 terms: $\frac98 h^2$ from the second factor alone, $-\frac18 h^2$ from the first alone.
(3) Example T4.68.

Each factor on its own gives a non-integer: $\int \hat A = -\frac18$ (there is no spinor bundle, Example T4.54) and "$\int e^{c_1/2} = \frac98$" (there is no line bundle $K^{-1/2}$, since $c_1 = 3h$ is odd). The product bundle $S \otimes K^{-1/2} = \Lambda^{0,\bullet}$ exists, and its index is the integer 1. This is the spin^c mechanism of Example T4.67 seen in the characteristic classes.

Worked example T4.71: Nash's character identities from bispinors

Nash needs two characters of exterior powers. Both are one line from $\Lambda^\bullet \otimes \mathbb{C} \cong S \otimes W$ and (T4.153).

The signature character [Nash (4.91)]. With $\Lambda^\pm = S^\pm \otimes W$:

$$\operatorname{ch}\Lambda^+ - \operatorname{ch}\Lambda^- \overset{(1)}{=} \big(\operatorname{ch} S^+ - \operatorname{ch} S^-\big)\operatorname{ch} W \overset{(2)}{=} (-1)^m \prod_i \big(e^{x_i/2} - e^{-x_i/2}\big)\big(e^{x_i/2} + e^{-x_i/2}\big)$$
$$\overset{(3)}{=} (-1)^m \prod_i \big(e^{x_i} - e^{-x_i}\big) = \prod_{i=1}^m \big(e^{-x_i} - e^{x_i}\big),$$

(1) The Chern character is multiplicative on tensor products and additive on sums.
(2) The spinor characters above, with $W \cong S$.
(3) $(a-b)(a+b) = a^2 - b^2$ with $a = e^{x/2}$, $b = e^{-x/2}$.

which is (4.91) including Nash's sign. *Direct check on a surface:* from Example T4.69, Λ^+ is

spanned by $1 + i\,\mathrm{vol}$ (a trivial line) and dz (the canonical line $K = \Lambda^{1,0}$, with $c_1(K) = -x$), and Λ^- by $1 - i\,\mathrm{vol}$ and $d\bar{z}$ (the line $\bar{K} \cong K^{-1}$). So $\mathrm{ch}\,\Lambda^+ - \mathrm{ch}\,\Lambda^- = e^{-x} - e^{x}$.
The de Rham character [Nash (4.59)], for the real bundle $TX \otimes \mathbb{C}$ with roots $\pm x_i$:

$$\sum_p (-1)^p \,\mathrm{ch}\,\Lambda^p \overset{(1)}{=} (-1)^m \big(\mathrm{ch}\,S^+ - \mathrm{ch}\,S^-\big)\big(\mathrm{ch}\,W^+ - \mathrm{ch}\,W^-\big) \overset{(2)}{=} (-1)^m \prod_i \big(e^{x_i/2} - e^{-x_i/2}\big)^2$$
$$\overset{(3)}{=} (-1)^m \prod_i \big(e^{x_i} - 2 + e^{-x_i}\big) \overset{(4)}{=} \prod_i (1 - e^{x_i})(1 - e^{-x_i}),$$

(1) (T4.153): the even-minus-odd forms are $(-1)^m(S^+ - S^-)\otimes(W^+ - W^-)$ as virtual bundles.
(2) Both factors equal $(-1)^m \prod(e^{x_i/2} - e^{-x_i/2})$; $(-1)^{3m} = (-1)^m$.
(3) Square out.
(4) $(1 - e^{x})(1 - e^{-x}) = 2 - e^{x} - e^{-x} = -(e^{x} - 2 + e^{-x})$, one minus sign per plane.

which is Nash's (4.59) with the $2m$ roots $x_1, \ldots, x_m, -x_1, \ldots, -x_m$. The "standard linear-algebraic fact" (4.56) and the $\mathrm{SO}(n)$ group theory behind (4.91) are the same statement: a form is a product of two spinors.

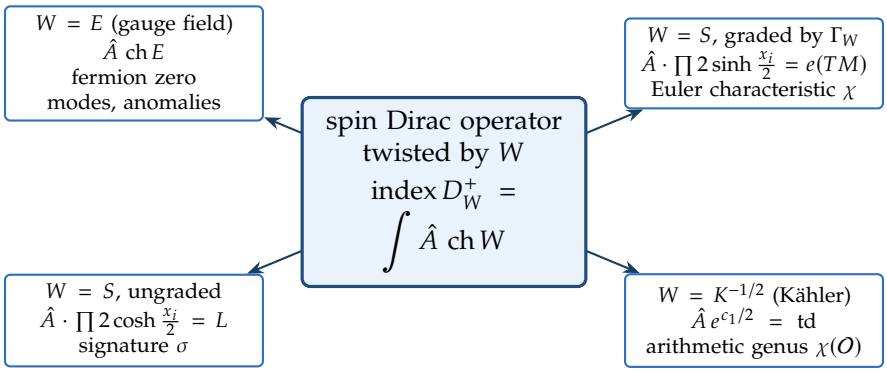


Figure T4.37: One index formula, four classical theorems. Each classical elliptic operator is the spin Dirac operator twisted by a bundle W; the integrand is $\hat{A}$ times the character of W. The twisting by spinors turns $\hat{A}$ into the Euler class (graded twist) or the L-class (ungraded twist); the twisting by $K^{-1/2}$ turns it into the Todd class. The three right-hand and lower answers make sense on non-spin manifolds, because $S \otimes W$ exists even when S does not.

operator	bundle	as D_W	integrand	index
Dirac D^+	S^+	$W = \mathbb{C}$	$\hat{A}$	$\hat{A}(M)$
twisted Dirac D^+_A	$S^+ \otimes E$	$W = E$	$\hat{A}\,\mathrm{ch}\,E$	$n_+ - n_-$ of D_A
Euler $d + d^*$	Λ^{even}	$W = S$, graded	$e(TM)$	$\chi(M)$
signature $d + d^*$	Λ^+ (via τ)	$W = S$	$L(M)$	$\sigma(M)$
Dolbeault $\sqrt{2}(\bar{\partial} + \bar{\partial}^*)$	$\Lambda^{0,\mathrm{even}}$	$W = K^{-1/2}$	$\mathrm{td}(M)$	$\chi(M, \mathcal{O})$

Physical picture: the supersymmetric quantum mechanics picture

In the path-integral proof of the index theorem (Chapter 4), $\mathrm{index}\,D = \mathrm{Tr}\,\Gamma e^{-tD^2}$ is the Witten index of a particle moving on M with fermions: one real fermion ψ^a per direction for the Dirac operator (these are the γ^a), *two* per direction for $d + d^*$ (these are c^a and $\hat{c}^a$). In the limit $t \to 0$ each rotation plane contributes a bosonic factor $\frac{x/2}{\sinh(x/2)}$ (this is $\hat{A}$) and each extra fermion pair contributes the fermionic factor $2\sinh(x/2)$ with periodic boundary

conditions (graded) or $2\cosh(x/2)$ with antiperiodic ones (ungraded). The Euler class is what remains when the extra fermions exactly cancel the bosons: $e(TM) = \mathrm{Pf}(R/2\pi)$ is a pure Berezin integral. The periodic/antiperiodic choice is the Ramond/Neveu–Schwarz choice of Section T4.6.3 in time.

Watch out: signs and orientations in the four formulas

The sign $(-1)^m$ appears twice in the Euler computation (in (T4.153) and in $\mathrm{ch}\, S^+ - \mathrm{ch}\, S^-$) and cancels. Nash's version has the same cancellation in the form $(-1)^{n/2}(-1)^{n/2}$ [Nash (4.62)]. If a different convention for Γ or for the orientation flips one of them, it flips the other too: χ does not depend on the orientation. The signature does: reversing the orientation flips τ and hence σ. The check that fixes all conventions at once is S^2: $\chi = 2 = \int e$, with $\int x = 2$.

Try it

Twist the Euler operator by a gauge bundle E, i.e. let $\mathrm{d}_A + \mathrm{d}_A^*$ act on $\Lambda^\bullet \otimes E$. Show that its index is $\mathrm{rk}(E)\,\chi(M)$ for every connection. *Answer:* the integrand is $\hat{A}\,(\mathrm{ch}\, W^+ - \mathrm{ch}\, W^-)(-1)^m\ \mathrm{ch}\, E = e(TM)\ \mathrm{ch}\, E$; since $e(TM)$ already has top degree, only $\mathrm{ch}_0 E = \mathrm{rk}\, E$ contributes. The twisted *signature* operator is different: its index $\int \prod_i \frac{x_i}{\tanh(x_i/2)}\ \mathrm{ch}\, E$ picks up $\mathrm{ch}_2 E$ in four dimensions, and it is the operator Atiyah and Singer used in their original proof.

Reading Nash (pp. 103–112, (4.47)–(4.95))

Nash treats the classical complexes as examples of his second index formula (4.48), $\mathrm{index}\, D = (-1)^{n/2}\{\mathrm{ch}(\sum(-1)^p[E^p])\, e(X)^{-1}\, \mathrm{td}(TX_{\mathbb{C}})\}[X]$, valid on even-dimensional oriented manifolds. The correspondences with this section are:

- *De Rham complex* (4.49)–(4.51): our $\mathrm{d} + \mathrm{d}^*$ on Ω^{even}, the Euler operator D_{E} (T4.126). Its index (4.52) is χ by Hodge theory (T4.127). Nash's characteristic class manipulations (4.55)–(4.62) giving Gauss–Bonnet (4.53)–(4.54) are our (T4.156); the identity (4.59) is derived in Example T4.71. His footnote that $b_i = \dim_{\mathbb{C}} H^i(X;\mathbb{C})$ is why we complexify forms throughout.
- *Dolbeault complex* (4.63)–(4.72): our $D_{\bar\partial}$ of Section T4.9.8 with index $\sum(-1)^q h^{0,q}$ (4.64)–(4.65), the arithmetic genus. Riemann–Roch (4.71)–(4.72) is (T4.158). The relation (4.73) between Betti and Hodge numbers is Figure T4.35. Nash uses only the $p = 0$ row, as we do; the general p is the twist by $\Lambda^{p,0}$.
- *Signature complex* (4.74)–(4.95): intersection form (4.74), signature (4.75), its vanishing for $n \equiv 2 \pmod 4$ (4.76), $\tau = i^{p(p-1)+n/2}\star$ (4.77) and $\tau^2 = 1$ (4.78) are our (T4.130)–(T4.133). The splitting (4.79), d_+ (4.80)–(4.82) and the Hodge-theoretic count (4.83)–(4.89) are (T4.135)–(T4.136). The character (4.90)–(4.91) is in Example T4.71, and (4.92)–(4.93) is (T4.157). The two normalisations of the L-polynomial (p. 112) are in the pitfall after (T4.138); (4.94)–(4.95) is Hirzebruch's theorem.

Nash writes the eigenvalues of τ as "∓ 1" in (4.79) but uses Ω^+ for the $+1$ eigenspace, as we do. He does not mention the Clifford structure on forms: in his treatment the three complexes are unrelated to the Dirac operator of pp. 113–118. The bispinor identification (T4.155) is what links them, and is the route taken in [Bär §1.3] and in Atiyah–Singer's later papers.

Recap

- Wedge $\varepsilon(v)$ and contraction $\iota(v)$ are fermion creation and annihilation operators on $\Lambda^\bullet$. The combinations $c = \varepsilon - \iota$ and $\hat{c} = \varepsilon + \iota$ are two anticommuting Clifford actions ($c^2 = -|v|^2$, $\hat{c}^2 = +|v|^2$). On $\Lambda^\bullet\mathbb{R}^2$ they are explicit 4×4 Kronecker products.
- The quantisation map Q turns a form into a matrix: forms are bispinors, $\Lambda^\bullet \otimes \mathbb{C} \cong S \otimes S^*$ (even n), which is why forms never need a spin structure.
- $\mathrm{d} + \mathrm{d}^* = \sum_a c(e^a)\nabla_a = -i\gamma^a\nabla_a$ is of Dirac type; its square is the Hodge Laplacian, and its zero modes are the harmonic forms, which represent de Rham cohomology (Hodge). Bochner: $\Delta = \nabla^*\nabla + \mathrm{Ric}$ on 1-forms, so $\mathrm{Ric} > 0 \Rightarrow b_1 = 0$.
- Gradings that anticommute with $\mathrm{d} + \mathrm{d}^*$: the degree parity (Euler operator, index χ) and the Clifford chirality $\tau = i^{k(k-1)+m}\star$ (signature operator, index σ, $\tau = \star$ on middle forms when $4 \mid n$). On Kähler manifolds $\sqrt{2}(\bar{\partial} + \bar{\partial}^*)$ on $\Lambda^{0,\bullet}$ is Dirac type too (index $\chi(M, \mathcal{O})$), and $S = \Lambda^{0,\bullet} \otimes K^{1/2}$: spin structures are square roots of K.
- Unification: $w(v) = \hat{c}(v)\tau$ commutes with c, so $\Lambda^\bullet \otimes \mathbb{C} = S \otimes W$ with $W \cong S$; the Levi-Civita connection splits; $\mathrm{d} + \mathrm{d}^* = D_W$. The Euler grading is $(-1)^m\Gamma_S\Gamma_W$ and the signature grading is Γ_S. From $\int \hat{A}\,\mathrm{ch}\,W$: $e(TM)$, $L(M)$ and $\mathrm{td}(M) = \hat{A}\,e^{c_1/2}$.

Chapter summary

The chapter started from one question: can the wave operator be written as the square of a first-order operator? Dirac's answer was yes, provided the coefficients anticommute, $\{\gamma^a, \gamma^b\} = 2\delta^{ab}$. Everything else followed from taking that relation seriously. The relation defines the Clifford algebra. Its matrix solutions are the gamma matrices, which we built in every dimension from tensor products of Pauli matrices, and whose size $2^{\lfloor n/2 \rfloor}$ is forced. Inside the algebra sit products of unit vectors. These act on vectors by reflections, and pairs of them act by rotations, so they form the spin group, a double cover of $\mathrm{SO}(n)$ in which a 2π rotation is -1.

The second half of the chapter moved the algebra onto a curved manifold. To do this we had to lift the frame transition functions from $\mathrm{SO}(n)$ to $\mathrm{Spin}(n)$ with consistent signs. When this is possible we have a spin structure, and there may be several of them: two on S^1, four on T^2, one on S^2 and none on $\mathbb{CP}^2$. On a spin manifold the Levi-Civita connection lifts to the spin connection $\nabla = \mathrm{d} + \frac{1}{4}\omega_{ab}\gamma^a\gamma^b$, and $D = -i\gamma^a\nabla_a$ is the Dirac operator. We computed its spectrum by hand on $\mathbb{R}^n$, S^1, T^2 and S^2. The Lichnerowicz formula $D^2 = \nabla^*\nabla + R/4$ explained the gaps we found, and the chirality splitting $D^+ : S^+ \to S^-$ produced the index $n_+ - n_-$. Finally the exterior algebra turned out to be a Clifford module too, so $\mathrm{d} + \mathrm{d}^*$ is a Dirac operator in disguise and the Euler characteristic, the signature and the arithmetic genus all come out of the same index formula.

The summary has five parts. Figures T4.38 and T4.39 put the algebra and the spectra on one page each. The concept tables (Tables T4.4 to T4.10) list every concept, what it means operationally, and where the rest of the book uses it. Tables T4.11 and T4.12 collect the numbers worth knowing by heart. The chapter closes with the calculations a reader should be able to do without looking, and a map from Nash's pages back to the sections of this chapter.

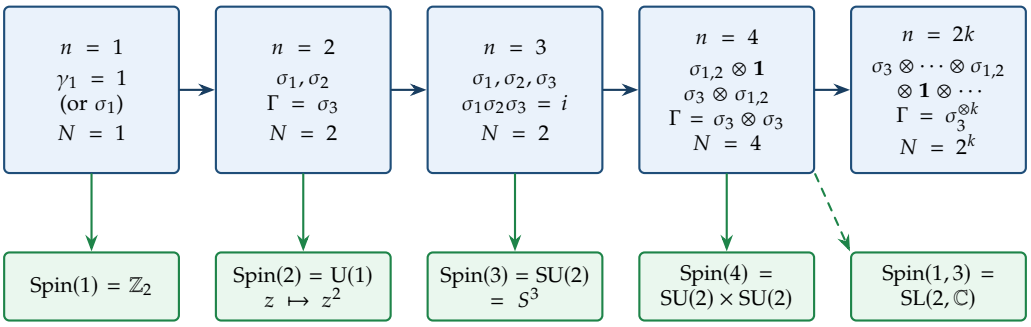


Figure T4.38: The algebraic half of the chapter in one picture. Top row: explicit Hermitian gamma matrices with $\{\gamma_a, \gamma_b\} = 2\delta_{ab}$, and the spinor dimension $N = 2^{\lfloor n/2 \rfloor}$. Each arrow is one step of the construction. Going from $n = 2k$ to $2k+1$ adds the chirality matrix Γ as the last gamma. Going from $2k+1$ to $2k+2$ doubles the size: the first $2k$ gammas become $\gamma_a \otimes \mathbf{1}$, and the last one ($\Gamma$) becomes $\Gamma \otimes \sigma_1$ and $\Gamma \otimes \sigma_2$ (Section T4.4.1). Starting from $\gamma_1 = 1$ at $n = 1$, this reproduces every box. Bottom row: the spin groups, generated by the even products of unit vectors, $\exp(-\frac{\theta}{2}\gamma_1\gamma_2)$ in each plane (Section T4.5). Each is a double cover of the rotation group, with kernel ± 1. The dashed arrow is the change to Minkowski signature, where the same four matrices generate Spin(1, 3) = SL(2, $\mathbb{C}$).

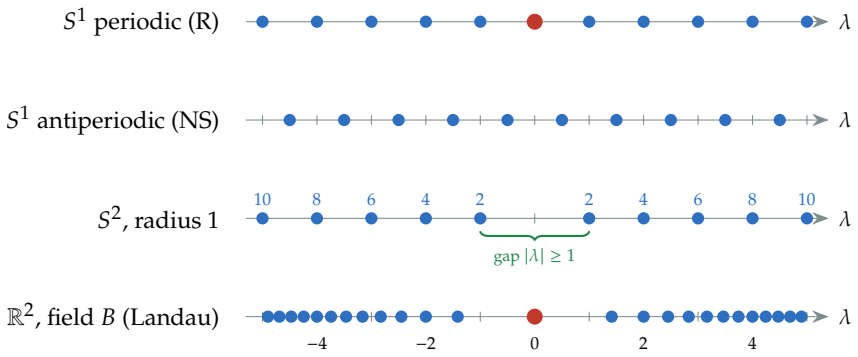


Figure T4.39: The Dirac spectra computed in this chapter, as dots on the λ axis (red: zero modes). The two circle spectra (length 2π) differ only by the spin structure: integers with one zero mode (Ramond), half-integers with none (Neveu–Schwarz) (Section T4.7.5). On the unit sphere the eigenvalues are $\pm(l+1)$ with multiplicity $2(l+1)$ (Section T4.7.7); the gap $|\lambda| \geq 1$ is Friedrich's bound $\lambda^2 \geq \frac{n}{n-1}\frac{R}{4}$ with $R = 2$, saturated (Section T4.8.3). In a magnetic field on $\mathbb{R}^2$ (here $B = 1$) the levels are $\pm\sqrt{2mB}$ plus a zero-energy level with one state per flux quantum, all of positive chirality (Section T4.8.4). Every spectrum is symmetric under $\lambda \to -\lambda$, because Γ anticommutes with D.

The algebra and the spectra on one page

Concept tables

In each table the last column says where the idea is used again. References to chapters are to the rest of the book; references to sections are to this chapter.

Concept	Operational meaning	Used later in
Square root of Δ	Demand $(-i\gamma^a\partial_a)^2 = -\partial^a\partial_a$; the cross terms cancel iff $\{\gamma^a, \gamma^b\} = 2\delta^{ab}$ (T4.2); no numbers do it, so γ^a are matrices (Section T4.1.1).	Every Dirac operator; $\sqrt{\text{Laplacian}}$ in Chapter 2
Klein–Gordon from Dirac	$(i\gamma^\mu\partial_\mu + m)(i\gamma^\nu\partial_\nu - m) = -(\partial^2 + m^2)$ with $\{\gamma^\mu, \gamma^\nu\} = 2\eta^{\mu\nu}$ (T4.3).	Fermions in Chapter 10
Symbol	Replace $\partial_a \to i\xi_a$ and keep the top order. Laplace type: $\sigma(\xi) = \lvert\xi\rvert^2$; Dirac type: $\sigma(\xi)^2 = \lvert\xi\rvert^2$, i.e. $\sigma(\xi)$ is Clifford multiplication by ξ (T4.7) (Section T4.1.3).	Ellipticity, Chapter 2; symbol class in $K(T^*X)$, Chapter 4
Clifford algebra $\mathrm{Cl}_{p,q}$	Polynomials in anticommuting generators with $e_i^2 = +1$ (p of them) or -1 (q of them); reduce any word using $vw + wv = 2\langle v, w\rangle$ (T4.9).	Everything in this chapter
Basis, dimension 2^n	Ordered monomials $e_{i_1}\cdots e_{i_k}$, $i_1 < \cdots < i_k$: one per subset, so 2^n (T4.11); picture as the vertices of a cube (Figure T4.5).	Counting spinor components
Low dimensions	$\mathrm{Cl}_{0,1} = \mathbb{C}$, $\mathrm{Cl}_{1,0} = \mathbb{R}\oplus\mathbb{R}$, $\mathrm{Cl}_{0,2} = \mathbb{H}$ (quaternions, $ij = k$), $\mathrm{Cl}_{2,0} = \mathbb{R}(2)$, $\mathrm{Cl}_{3,0} = \mathbb{C}(2)$ (Pauli), $\mathrm{Cl}_{0,3} = \mathbb{H}\oplus\mathbb{H}$ (Sections T4.2.3 to T4.2.5).	Spin groups, instantons ($\mathbb{H}$), Chapter 8
Clifford vs exterior	$vw = v\wedge w + \langle v, w\rangle$: the same vector space $\Lambda^\bullet V$ with a deformed product (Dray's $\vee$ product, (T4.15)).	Forms as Clifford modules, Section T4.9
Even and odd parts	$\mathrm{Cl} = \mathrm{Cl}^0 \oplus \mathrm{Cl}^1$ by word length mod 2 (T4.17); $\mathrm{Cl}^0_{p,q} \cong \mathrm{Cl}_{p,q-1}$ (T4.18), e.g. $\mathrm{Cl}^0_{3,0} = \mathbb{H}$.	$\mathrm{Spin}(n) \subset \mathrm{Cl}^0$; supersymmetry gradings
Graded tensor product	$\mathrm{Cl}(V\oplus W) = \mathrm{Cl}(V)\hat{\otimes}\mathrm{Cl}(W)$, with a sign when two odd elements pass (T4.19).	Building gammas in higher n

Table T4.4: The Dirac square root and Clifford algebras (Sections T4.1 and T4.2).

Numbers worth knowing by heart

The first table is the dimension count behind every spinor calculation. Read across a row: in dimension n the gamma matrices are $N \times N$ with $N = 2^{\lfloor n/2\rfloor}$, a chirality matrix exists only for even n, and then the Weyl spinors have half the components.

Calculations to be able to do without looking

A toolkit chapter has done its job when its core calculations become reflexes. These are the ones the rest of the book assumes.

1. *Square the Dirac operator on $\mathbb{R}^n$.* Split $\gamma^a\gamma^b\partial_a\partial_b$ into its symmetric part, which is $\delta^{ab}\partial_a\partial_b$, and its antisymmetric part, which vanishes because partial derivatives commute (Section T4.1.1).

Concept	Operational meaning	Used later in
Euclidean chiral gammas ($n = 4$)	$\gamma_a = \left(\begin{smallmatrix} 0 & \tau_a \\ \tau_a^\dagger & 0 \end{smallmatrix}\right)$ with $\tau = (\vec{\sigma}, \pm i\mathbf{1})$ (T4.21); $\Gamma = \mathrm{diag}(\mathbf{1}, -\mathbf{1})$ (T4.24) (Section T4.3.1).	Instantons, Chapter 8; Nash (4.99)
Dirac and Weyl representations	Minkowski γ^0 Hermitian, γ^i anti-Hermitian; Weyl: $\gamma^\mu = \left(\begin{smallmatrix} 0 & \sigma^\mu \\ \bar{\sigma}^\mu & 0 \end{smallmatrix}\right)$ (T4.26); related by a unitary change of basis (Example T4.16).	QFT notes; Chapter 10
Chirality γ^5, Γ	$\gamma^5 = i\gamma^0\gamma^1\gamma^2\gamma^3$; squares to $\mathbf{1}$, anticommutes with every γ^μ (T4.28); projectors $\frac{1}{2}(1 \pm \gamma^5)$.	Chiral fermions, Nash (10.1); index
Lorentz generators	$S^{\mu\nu} = \frac{i}{4}[\gamma^\mu, \gamma^\nu]$ (T4.29) obey the $\mathfrak{so}(1,3)$ algebra; $\Lambda_{1/2} = \exp(-\frac{i}{2}\omega_{\mu\nu}S^{\mu\nu})$ (T4.30).	Spin connection, Section T4.7
2π rotation	$\exp(-i\theta S^{12})$ at $\theta = 2\pi$ is $-\mathbf{1}$ (Example T4.19); boosts are not unitary (Example T4.20).	Spin structures
Weyl spinors	$\psi = (\psi_L, \psi_R)$, $\gamma^5 = \mp 1$; massless Dirac equation decouples; helicity (Section T4.3.5 and Figure T4.7).	Chiral anomalies, Chapter 10
Bilinears, traces	$\bar{\psi}\Gamma\psi$ for $\Gamma \in \{\mathbf{1}, \gamma^\mu, \sigma^{\mu\nu}, \gamma^\mu\gamma^5, \gamma^5\}$ (Table T4.2); $\mathrm{tr}\,\gamma^\mu\gamma^\nu = 4\eta^{\mu\nu}$, $\mathrm{tr}\,\gamma^5\gamma^\mu\gamma^\nu\gamma^\rho\gamma^\sigma = -4i\varepsilon^{\mu\nu\rho\sigma}$ (Example T4.24).	Triangle diagram, Chapter 10
SL$(2,\mathbb{C})$ picture	$X = x^\mu\sigma_\mu$, $\det X = x^\mu x_\mu$; $X \mapsto AXA^\dagger$ is a Lorentz transformation (T4.33) (Section T4.3.7).	Spin$(1,3)$
Gammas in $n = 2k$	Jordan–Wigner tensor products (T4.35); $\Gamma = (-i)^k\gamma_1\cdots\gamma_{2k} = \sigma_3^{\otimes k}$ (T4.36); explicit $n = 2, 4, 6$ (Example T4.25).	D in any dimension; $d = 6, 10, 12$ anomalies
Fock space	$b_j = \frac{1}{2}(\gamma_{2j-1} - i\gamma_{2j})$ obey the CAR (T4.38); spinors = states of k fermion modes, $\Gamma = (-1)^F$ (Section T4.4.2).	Forms as fermions, Chapter 7; Dolbeault
Uniqueness	Every irreducible representation in even n is equivalent to the one built here; odd n has two, $\Gamma \to \pm$ (Sections T4.4.3 and T4.4.4).	Well-defined spinor bundle
Classification	$\mathrm{Cl}_{0,n}$, $\mathrm{Cl}_{n,0}$, $\mathrm{Cl}_n \otimes \mathbb{C}$ as matrix algebras over $\mathbb{R}, \mathbb{C}, \mathbb{H}$ (Table T4.3); period 8 (real) and 2 (complex) from $\mathrm{Cl}_{0,n+2} = \mathrm{Cl}_{n,0} \otimes \mathbb{H}$ (T4.42).	Bott periodicity, KO, Chapter 3; real families, Chapter 4
Majorana spinors	Real representations exist when the algebra is of real type; $\mathrm{Cl}_{3,1} = \mathbb{R}(4)$ (Example T4.26).	Free Majorana fermion, Chapter 11

Table T4.5: Gamma matrices in four and in general dimensions (Sections T4.3 and T4.4).

Concept	Operational meaning	Used later in
Reflection	For a unit vector u, $x \mapsto uxu^{-1} = -R_u(x)$ (T4.44): conjugation by u is (minus) the mirror image in the plane $\perp u$.	Parity; Pin groups
Two reflections	Mirrors at angle $\theta/2$ compose to a rotation by θ (T4.45); the lift is $\exp(-\frac{\theta}{2}\gamma_1\gamma_2)$ (T4.46) (Figure T4.10).	Every spinor rotation
$\mathrm{Pin}(n)$, $\mathrm{Spin}(n)$	Products of unit vectors; Spin: an even number of them (Section T4.5.3).	Structure group of S
Double cover	$1 \to \{\pm 1\} \to \mathrm{Spin}(n) \to \mathrm{SO}(n) \to 1$ (T4.49); $a(2\pi) = -1$ (T4.47); the spinor representation is not a representation of $\mathrm{SO}(n)$.	Spin structures, w_2
Spin(2), Spin(3)	$\mathrm{Spin}(2) = \mathrm{U}(1)$ covering $\mathrm{SO}(2)$ by $z \mapsto z^2$ (T4.50); $\mathrm{Spin}(3) = \mathrm{SU}(2)$ with $R_{ij} = \frac{1}{2}\operatorname{tr}(\sigma_i U \sigma_j U^\dagger)$ (T4.52); $\mathrm{SO}(3) = \mathbb{RP}^3$, belt trick (Figure T4.12).	Monopoles, Chapter 8
$\mathfrak{spin}(n)$	Spanned by $\gamma_a\gamma_b$ $(a < b)$; an infinitesimal rotation A_{ab} lifts to $\frac{1}{4}A_{ab}\gamma_a\gamma_b = \frac{1}{2}A_{ab}\Sigma_{ab}$, $\Sigma_{ab} = \frac{1}{4}[\gamma_a, \gamma_b]$ (T4.55).	Spin connection $\frac{1}{4}\omega_{ab}\gamma^a\gamma^b$
Spin(4)	Unit quaternions (p, q) acting on $x \in \mathbb{H}$ by $x \mapsto pxq$ (T4.57); $\mathfrak{so}(4) = \mathfrak{su}(2) \oplus \mathfrak{su}(2)$, and $S^\pm$ carry the two factors; self-dual 2-forms act on S^+ only (T4.58).	Instantons, ADHM, Chapters 8 and 12
Lorentzian signature	$\mathrm{Spin}(1,3) = \mathrm{SL}(2,\mathbb{C})$; $\mathrm{Spin}(1,1)$ as the boosts (Example T4.32); Lorentz group has two components beyond Spin (Section T4.5.7).	Worldsheet fermions, Chapter 11

Table T4.6: The spin group (Section T4.5).

Concept	Operational meaning	Used later in
Frame bundle	Orthonormal frames glued by $g_{\alpha\beta} : U_\alpha \cap U_\beta \to \mathrm{SO}(n)$ with the cocycle rule (T4.60) (Section T4.6.1).	Every associated bundle, Chapter 1
Spin structure	Choose lifts $\hat{g}_{\alpha\beta} \in \mathrm{Spin}(n)$; the triple-overlap signs $\varepsilon_{\alpha\beta\gamma} = \hat{g}_{\alpha\beta}\hat{g}_{\beta\gamma}\hat{g}_{\gamma\alpha}$ (T4.62) must all be made $+1$ (Section T4.6.2).	All fermions on curved space
w_1, w_2	w_1: orientability; $w_2 = [\varepsilon_{\alpha\beta\gamma}] \in H^2(M;\mathbb{Z}_2)$ is the obstruction (T4.63); spin iff $w_1 = w_2 = 0$.	Nash (10.1) requirement; Chapters 1 and 10
Counting	If spin, the structures are counted by $H^1(M;\mathbb{Z}_2) = \mathrm{Hom}(\pi_1 M, \mathbb{Z}_2)$: a sign for each loop.	Nash p. 113; string sectors
S^1	Two: periodic (Ramond) and antiperiodic (Neveu–Schwarz) (T4.64); only the NS one bounds the disc (Example T4.34).	Superstrings, Chapter 11
T^2, T^n	Signs $(\varepsilon_1, \varepsilon_2)$ around the two cycles (T4.65): four; T^n has 2^n.	Fermions on the torus, Chapter 11
S^2	One; the clutching map $S^1 \to \mathrm{SO}(2)$ has degree 2, which is even, so it lifts (Figure T4.18).	Monopoles, S^{2m} in Nash (10.3)
$\mathbb{CP}^2$	None: $T\mathbb{CP}^2\big\|_{\mathbb{CP}^1} = \mathcal{O}(2) \oplus \mathcal{O}(1)$ has odd total degree 3 (Example T4.37); $\mathbb{CP}^n$ spin iff n odd.	index would be $-\frac{1}{8}$; Spin^c
Spinor bundle	$S = P_{\mathrm{Spin}} \times_{\mathrm{Spin}} \mathbb{C}^N$: components glued by $\hat{g}_{\alpha\beta}$ (T4.61); Clifford multiplication $TM \otimes S \to S$ (Section T4.6.7).	$E^\pm$ in Nash (4.100)

Table T4.7: Spin structures (Section T4.6).

Concept	Operational meaning	Used later in
Connection 1-forms	Solve $de^a + \omega^a{}_b \wedge e^b = 0$, $\omega_{ab} = -\omega_{ba}$ (T4.67); e.g. $\omega_{12} = -\cos\theta\, d\phi$ on S^2 (Example T4.39).	Gravitational anomalies, Chapter 10
Spin connection	$\nabla\psi = d\psi + \frac{1}{4}\omega_{ab}\gamma^a\gamma^b\psi$ (T4.69); Leibniz for Clifford multiplication (T4.70); preserves $S^{\pm}$.	Γ^μ in Nash (10.1)
Spin curvature	$R^S = \frac{1}{4}R_{ab}\gamma^a\gamma^b$ (T4.71).	Lichnerowicz; $\hat{A}$
Dirac operator	$D = -i\gamma^a\nabla_{e_a}$ (T4.72); formally self-adjoint by a divergence (T4.73); $D = \left(\begin{smallmatrix} 0 & D^- \\ D^+ & 0 \end{smallmatrix}\right)$ (T4.75).	Chapters 4 and 10
$\mathbb{R}^n$	Plane waves $u\, e^{ik\cdot x}$, $\lambda = \pm\lvert k\rvert$; in $n = 2$, $D^+ = -2i\partial_{\bar z}$ so zero modes are holomorphic (T4.76) (Example T4.43).	Riemann surfaces, Chapter 5
S^1	spec $D = \frac{2\pi}{L}\mathbb{Z}$ (R, one zero mode) or $\frac{2\pi}{L}(\mathbb{Z} + \frac{1}{2})$ (NS, none); spectral flow under a flat U(1) twist (Figure T4.20).	η-invariant, global anomalies, Chapter 10
T^2	$\lambda = \pm\sqrt{p^2 + q^2}$ on the shifted lattices of Figure T4.21; only $(+,+)$ has zero modes, $(n_+, n_-) = (1, 1)$, index 0.	Torus partition functions, Chapter 11
S^2	$\pm(l+1)$ with multiplicity $2(l+1)$ (T4.81), via $K = \vec{\sigma}\cdot\vec{L} + 1$ (T4.80); no zero modes.	Monopole harmonics, Chapter 8

Table T4.8: The spin connection and the Dirac operator (Section T4.7).

Concept	Operational meaning	Used later in
Connection Laplacian	$\nabla^*\nabla = -\sum_a(\nabla_a\nabla_a - \nabla_{\nabla_a e_a})$ (T4.82); $\langle\psi, \nabla^*\nabla\psi\rangle = \lVert\nabla\psi\rVert^2 \geq 0$ (T4.83).	Heat kernels, Chapter 4
Lichnerowicz	$D^2 = \nabla^*\nabla + \frac{R}{4}$ (T4.85): expand $\gamma^a\gamma^b\nabla_a\nabla_b$ into symmetric and antisymmetric parts; $\gamma^a\gamma^b\gamma^c R_{abcd}$ collapses to $-2\,\mathrm{Ric}$ by the Bianchi identity (Section T4.8.2).	Vanishing theorems, Chapters 4 and 12
Positive curvature	$R > 0 \Rightarrow \ker D = 0$ (T4.90), hence $\hat{A}(M) = 0$; Friedrich $\lambda^2 \geq \frac{n}{n-1}\frac{\min R}{4}$ (T4.92); on S^2, $\lambda^2 \geq 4\pi/\mathrm{area}$ (T4.93).	Obstructions to psc metrics
Twisted Dirac	$D_A = -i\gamma^a(\nabla_a + A_a)$ on $S \otimes E$ (T4.94); $D_A^2 = \nabla^*\nabla + \frac{R}{4} - \frac{1}{2}F_{ab}\gamma^a\gamma^b$ (T4.95).	Nash (4.122), (10.2)
Landau levels	On $\mathbb{R}^2$ with field B: $D_A^2 = B(2b^\dagger b + 1) - B\sigma_3$, spectrum $0, \pm\sqrt{2mB}$ (Example T4.51).	Graphene; spectral asymmetry
Aharonov–Casher	Zero modes $e^{-\varphi}z^k$ with $\nabla^2\varphi = B$: $n_+ = \lceil\Phi/2\pi\rceil - 1$, $n_- = 0$ (T4.96).	"Zero modes = flux"
Monopole on S^2	Twist by L_k: $n_+ - n_- = k$ (Example T4.53), holomorphic sections of $\mathcal{O}(k-1)$.	Nash (10.3) with $m = 1$
Index	$\mathrm{index}\, D^+ = n_+ - n_-$ (T4.98); nonzero modes pair up; $\mathrm{Tr}\,\Gamma e^{-tD^2}$ is independent of t (McKean–Singer, (T4.99)).	Central object of Chapter 4; anomalies
$\hat{A}$ and ch	$\mathrm{index}\, D_A^+ = \int \hat{A}(M)\,\mathrm{ch}(E)$ (T4.103), $\hat{A} = \prod \frac{x_i/2}{\sinh(x_i/2)} = 1 - \frac{p_1}{24} + \cdots$ (T4.102); $n = 4$: $\mathrm{index}\, D^+ = -\sigma/8$ (T4.104).	Nash (4.107)–(4.111), Chapters 4 and 10

Table T4.9: The Lichnerowicz formula, twisted Dirac operators and the index (Section T4.8).

Concept	Operational meaning	Used later in
Wedge and contraction	$\varepsilon(v) = v\wedge$, $\iota(v)$; $\{\varepsilon(v), \iota(w)\} = \langle v, w\rangle$ (T4.105): fermion creation and annihilation on $\Lambda^\bullet$.	SUSY QM, Nash (7.64)–(7.69), Chapter 7
Clifford actions on forms	$c = \varepsilon - \iota$, $\hat{c} = \varepsilon + \iota$ (T4.107): $c^2 = -\lvert v\rvert^2$, $\hat{c}^2 = +\lvert v\rvert^2$, $\{c, \hat{c}\} = 0$; explicit 4×4 on $\Lambda^\bullet\mathbb{R}^2$ (T4.110).	Witten's Q_1, Q_2
Forms = bispinors	$Q(e^{a_1}\wedge\cdots) = \gamma^{a_1}\cdots$ (T4.111); $\Lambda^\bullet\otimes\mathbb{C} \cong S\otimes S^*$ (T4.114); forms need no spin structure.	Twisting in TQFT, Chapter 12
$\mathrm{d} + \mathrm{d}^*$	$= \sum_a c(e^a)\nabla_a$ (T4.116); on $\mathbb{R}^2$ it is Cauchy–Riemann, on $\mathbb{R}^3$ grad, curl, div (T4.120); its square is the Hodge Laplacian.	Nash (4.49)–(4.52); Morse, Chapter 7
Hodge theory	Harmonic forms = zero modes of $\mathrm{d} + \mathrm{d}^*$ = de Rham classes (T4.122) (Figure T4.30).	H^p computations throughout
Bochner	On 1-forms $\Delta = \nabla^*\nabla + \mathrm{Ric}$ (T4.124); $\mathrm{Ric} > 0 \Rightarrow b_1 = 0$.	Vanishing theorems
Euler operator	Grade by degree parity: index $= \sum(-1)^p b_p = \chi(M)$ (T4.127); Witten index $\mathrm{Tr}(-1)^F e^{-tH}$ (T4.129).	Gauss–Bonnet, Morse inequalities, Chapter 7
Signature operator	Grade by $\tau = i^{k(k-1)+m}\star$ (T4.132); index $= \sigma(M) = b^+ - b^-$ (T4.136); Hirzebruch $\sigma = \int L = \frac{1}{3}\int p_1$ (T4.138).	Nash (4.74)–(4.95); gravitational anomalies
Dolbeault operator	$\sqrt{2}(\bar\partial + \bar\partial^*)$ on $\Lambda^{0,\bullet}$ (T4.142); index $= \sum(-1)^q h^{0,q}$; $S = \Lambda^{0,\bullet}\otimes K^{1/2}$ (T4.145).	Riemann–Roch, theta characteristics, Chapter 5
Unification	$\Lambda^\bullet\otimes\mathbb{C} = S\otimes W$, $\mathrm{d} + \mathrm{d}^* = D_W$ (T4.155); e, L, td all from $\int \hat{A}\,\mathrm{ch}\,W$ (Figure T4.37).	Nash (4.48)

Table T4.10: Differential forms as Clifford modules (Section T4.9).

n	$\dim \mathrm{Cl}$	$N = 2^{\lfloor n/2 \rfloor}$	chirality Γ	$\dim S^{\pm}$	$\mathrm{Spin}(n)$
1	2	1	no	–	$\mathbb{Z}_2$
2	4	2	σ_3	1	U(1)
3	8	2	no	–	SU(2)
4	16	4	$\sigma_3 \otimes \sigma_3$	2	$\mathrm{SU}(2) \times \mathrm{SU}(2)$
5	32	4	no	–	(double cover of SO(5))
6	64	8	$\sigma_3^{\otimes 3}$	4	(double cover of SO(6))
10	2^{10}	32	$\sigma_3^{\otimes 5}$	16	(superstring spacetime)

Quantity	Value
Clifford relation (ours)	$\{\gamma_a, \gamma_b\} = 2\delta_{ab}$, $\gamma_a^\dagger = \gamma_a$
Translation to Bär and Nash	$c(e_a) = -i\gamma_a$; $\gamma_a^{\mathrm{Nash}} = i\gamma_a$, $\not{D}_{\mathrm{Nash}} = -D$
Spinor rotation in the 12-plane	$\exp(-\frac{\theta}{2}\gamma_1\gamma_2) = \cos\frac{\theta}{2} - \sin\frac{\theta}{2}\,\gamma_1\gamma_2$
Spin generator	$\Sigma_{ab} = \frac{1}{4}[\gamma_a, \gamma_b]$; Minkowski $S^{\mu\nu} = \frac{i}{4}[\gamma^\mu, \gamma^\nu]$
Spin connection	$\Omega = \frac{1}{4}\omega_{ab}\gamma^a\gamma^b$
Lichnerowicz	$D^2 = \nabla^*\nabla + \frac{R}{4}$; twisted: $-\frac{1}{2}F_{ab}\gamma^a\gamma^b$ extra
Unit S^n	$R = n(n-1)$; unit S^2: $R = 2$, $\lambda_{\min}^2 = 1$
$\hat{A}$ expansion	$1 - \frac{p_1}{24} + \frac{7p_1^2 - 4p_2}{5760} + \cdots$
Four-dimensional index	$\mathrm{index}\, D^+ = -\frac{\sigma}{8}$; twisted by SU(2): $-k - \frac{\sigma}{4}$
Spin structures	$\lvert H^1(M;\mathbb{Z}_2)\rvert$: S^1: 2, T^2: 4, Σ_g: 2^{2g}, simply connected: ≤ 1

Table T4.11: Dimensions of Clifford algebras, spinors and spin groups (top), and the constants that recur (bottom). The $n = 10$ row is the superstring's: the Weyl condition halves 32 complex components to 16, and in signature $(1, 9)$ a Majorana condition can be imposed as well, leaving 16 real components.

Manifold	spin str.	spectrum of D	zero modes (n_+, n_-)	index
S^1, length 2π	2	R: $\mathbb{Z}$; NS: $\mathbb{Z} + \frac{1}{2}$	R: 1; NS: 0 (no chirality)	–
$T^2 = \mathbb{R}^2/(2\pi\mathbb{Z})^2$	4	$\pm\sqrt{p^2+q^2}$, (p, q) in $\mathbb{Z}$ or $\mathbb{Z} + \frac{1}{2}$ per cycle	$(+,+)$: $(1,1)$; else $(0,0)$	0
S^2, radius 1	1	$\pm(l+1)$, mult. $2(l+1)$	$(0,0)$	0
S^2 with L_k	1	k zero modes, the rest in $\pm$ pairs	$(k,0)$ for $k \ge 0$	k
$\mathbb{R}^2$, flux Φ	1	$0, \pm\sqrt{2mB}$ (uniform B)	$(\lceil \frac{\Phi}{2\pi} \rceil - 1, 0)$	$\approx \frac{\Phi}{2\pi}$
K3	1	(not computed)	$(2,0)$	$2 = -\frac{\sigma}{8}$
$\mathbb{CP}^2$	0	no Dirac operator	–	$-\frac{1}{8} \notin \mathbb{Z}$

Table T4.12: The manifolds of this chapter and their Dirac data. The index column vanishes on every untwisted surface because $\hat{A}$ has no 2-form part; it becomes nonzero only when a gauge field (the monopole, the flux) or four-dimensional topology (K3) supplies a top-degree class.

2. *Check a Clifford relation by multiplying matrices.* For example $(\sigma_3 \otimes \sigma_1)(\sigma_1 \otimes \mathbf{1}) = \sigma_3\sigma_1 \otimes \sigma_1 = -\sigma_1\sigma_3 \otimes \sigma_1$: the σ_3 prefix is what makes different slots anticommute (T4.34).

3. *Build the gammas in dimension $2k$ and the chirality matrix.* Use (T4.35), then $\gamma_{2j-1}\gamma_{2j} = \mathbf{1} \otimes \cdots \otimes i\sigma_3 \otimes \cdots$, so $\Gamma = (-i)^k \prod \gamma_a = \sigma_3^{\otimes k}$ (Example T4.25).

4. *Lift a rotation.* Two reflections in unit vectors at angle $\theta/2$ give $\exp(-\frac{\theta}{2}\gamma_1\gamma_2)$; at $\theta = 2\pi$ this is -1 (T4.46).

5. *Compute the rotation of a spin matrix.* $R_{ij} = \frac{1}{2}\operatorname{tr}(\sigma_i U \sigma_j U^\dagger)$ for $U \in \mathrm{SU}(2)$, and see that $\pm U$ give the same R (Example T4.30).

6. *Count spin structures.* Pick a sign for each generator of π_1 (on S^1, T^2, Σ_g); on simply connected manifolds decide whether the clutching winding number is even (Sections T4.6.5 and T4.6.6).

7. *Find the spin connection.* Solve Cartan's equation $\mathrm{d}e^a = -\omega^a{}_b \wedge e^b$ in an orthonormal frame, then $\Omega = \frac{1}{4}\omega_{ab}\gamma^a\gamma^b$ (Example T4.41).

8. *Compute a spectrum on a flat manifold.* Plane waves with momenta on the lattice fixed by the spin structure; $\lambda = \pm|k|$ (Examples T4.44 and T4.45).

9. *Derive the Lichnerowicz formula.* Split $\gamma^a\gamma^b\nabla_a\nabla_b$ as in item 1, but now the antisymmetric part is $\frac{1}{2}\gamma^a\gamma^b R^S_{ab} = \frac{R}{4}$ by the three-gamma identity (Section T4.8.2); add F for a twisted operator.

10. *Run the vanishing argument.* $\|D\psi\|^2 = \|\nabla\psi\|^2 + \int \frac{R}{4}|\psi|^2$, so $R > 0$ and $D\psi = 0$ force $\psi = 0$ (T4.90).

11. *Count zero modes by holomorphy.* In two dimensions $D^+ \propto \partial_{\bar z}$ (twisted: $\partial_{\bar z} + A_{\bar z}$), so zero modes are holomorphic sections; count them (Examples T4.52 and T4.53).

12. *Write $\mathrm{d} + \mathrm{d}^*$ as a Dirac operator.* $\mathrm{d} = \sum \varepsilon(e^a)\nabla_a$, $\mathrm{d}^* = -\sum \iota(e^a)\nabla_a$, add them (Section T4.9.3).

Where Nash uses this chapter

Nash has no chapter on Clifford algebras; he uses the material in passing, mostly in his Chapters 4, 7 and 10. The table lists the places, with the section of this chapter that supplies the background. Page numbers are Nash's printed pages.

Nash	What he uses	Here
p. 16	Stiefel–Whitney classes as $\mathbb{Z}_2$ characteristic classes	Section T4.6.2
pp. 28–35	Symbols of differential operators; ellipticity of $\partial_{\bar z}$ and the Laplacian	Section T4.1.3
pp. 75–78	Bott periodicity; the real and quaternionic periodicities "can be traced" to Clifford algebras	Section T4.4.5
pp. 98–112, (4.47)–(4.95)	de Rham, Dolbeault and signature complexes and their indices	Section T4.9
p. 113, (4.96)–(4.98)	Spin structures, $w_1 = w_2 = 0$, counting by $H^1(X;\mathbb{Z}_2)$, $\mathbb{CP}^n$ spin iff n odd	Section T4.6
pp. 113–115, (4.99)–(4.106)	Clifford algebra, $\gamma_5 = i^{n(n+1)/2}\gamma_1\cdots\gamma_n$, $E = E^+ \oplus E^-$, the Dirac operator, harmonic spinors, index $\not{D} = n_+ - n_-$	Sections T4.4, T4.7 and T4.8.5
pp. 115–118, (4.107)–(4.122)	$\hat A$ genus, index of the (twisted) Dirac operator	Sections T4.8.4 and T4.8.5
pp. 121–122	Real families, skew-adjoint operators, mod 2 indices	Section T4.4.5
pp. 207–210, (7.60)–(7.75)	SUSY QM: $Q_1 = \mathrm{d} + \mathrm{d}^*$, $Q_2 = i(\mathrm{d} - \mathrm{d}^*)$, $(-1)^F$, forms as fermions	Sections T4.9.1 and T4.9.6
p. 241, (8.137)	Dimensions of instanton moduli spaces for Spin(N) and other groups (twisted Dirac index)	Sections T4.5.6 and T4.8.4
pp. 270–278, (10.1)–(10.3)	Chiral fermions, $\gamma_\mu(\partial^\mu + \Gamma^\mu + A^\mu)(1+\gamma_5)$, spin manifold requirement, index $\not{D}_A = \int \hat A\,\mathrm{ch}\,F$	Sections T4.3.3, T4.7.2 and T4.8.5
pp. 279–281, (10.35)	Gravitational anomalies: metric dependence of the Dirac operator	Section T4.7
pp. 283–290, (10.46)–(10.59)	Global anomalies: five-dimensional Dirac operators, mod 2 index, 11- and 12-dimensional Dirac operators	Sections T4.4 and T4.4.4
p. 295, (10.80)–(10.81)	Self-adjoint Dirac operator on S^1 and its spectral flow	Section T4.7.5
p. 303, (11.5)–(11.10)	Free Majorana fermion on the plane; spin $h - \bar h$	Sections T4.4.5 and T4.7.4
p. 332, (12.38)–(12.41)	Fermions of integral spin in Donaldson–Witten theory (twisting)	Sections T4.5.6 and T4.9.2

Table T4.13: Where Nash relies on this toolkit chapter.

Recap: the chapter in five sentences

- Square-rooting the Laplacian forces $\{\gamma^a, \gamma^b\} = 2\delta^{ab}$, and the solutions are explicit matrices: Pauli matrices, the 4×4 Dirac and Weyl matrices, and tensor products $\sigma_3 \otimes \cdots \otimes \sigma_{1,2} \otimes \mathbf{1}\cdots$ of size $2^{\lfloor n/2\rfloor}$, with chirality $\Gamma = \sigma_3^{\otimes k}$ in even dimensions.
- Even products of unit vectors form Spin(n), which double covers SO(n) through reflections; its Lie algebra is spanned by $\frac{1}{4}[\gamma_a, \gamma_b]$, and a 2π rotation is -1 on spinors.
- Spinors live on a manifold only after the frame transition functions are lifted with consistent signs; the obstruction is w_2 and the choices are counted by $H^1(M;\mathbb{Z}_2)$ (Ramond and Neveu–Schwarz on S^1).
- The Dirac operator $D = -i\gamma^a\nabla_a = -i\gamma^a e_a{}^\mu(\partial_\mu + \frac{1}{4}\omega_{bc\mu}\gamma^b\gamma^c)$ has computable spectra, obeys $D^2 = \nabla^*\nabla + \frac{R}{4} - \gamma(F)$, and has no zero modes when $R > 0$ and $F = 0$.

- Chirality splits D into $D^{\pm}$ and defines the index $n_+ - n_- = \int \hat{A}\ \mathrm{ch}\, E$; with forms as a Clifford module, $\mathrm{d} + \mathrm{d}^*$ is a twisted Dirac operator whose indices are χ, σ and the arithmetic genus.

Chapter T5

The Renormalisation Group

Where this chapter is going

Quantum field theories come with a scale. A lattice has a spacing, a regularised path integral has a cutoff, a scattering experiment has an energy. The **renormalisation group** (RG) is the systematic answer to the question "how does the description of a system change when we change the scale at which we look at it?". The answer is a flow on the space of all possible couplings. Its fixed points are theories without any scale, which turn out to be conformal field theories, and every quantum field theory can be viewed as a trajectory of this flow running from one fixed point at short distances to another at long distances.

The logic has seven steps. Coarse graining on a lattice and in momentum space defines the flow (Section T5.1). Polchinski's exact flow equation makes it a differential equation, which can be solved in the local potential approximation (Section T5.2). Near a fixed point the flow is linear, its eigenvalues are the critical exponents, and the Wilson–Fisher fixed point in $4-\varepsilon$ dimensions gives them explicitly (Section T5.3). The same running can be computed with ordinary Feynman diagrams. The Callan–Symanzik equation turns the freedom to choose the renormalisation scale into power laws, and one-loop diagrams give the β-functions of φ^4 theory and of Yang–Mills theory, whose sign is asymptotic freedom (Section T5.4). A fixed point is conformally invariant under a condition on the stress tensor that always holds for unitary theories in two dimensions. Away from a fixed point the trace of the stress tensor is the β-function, which is the scale anomaly (Section T5.5). Put together, these facts give a dictionary between quantum field theories and conformal field theories (Section T5.6). Finally, two β-functions carry geometry. For a sigma model with a curved target it is the Ricci tensor, so fixed points solve Einstein's vacuum equations and the flow is Ricci flow. For Chern–Simons theory the one-loop correction shifts the level k to $k + h^\vee$ (Section T5.7). Figure T5.1 shows the logic and where each piece is used later.

Dictionary: the words of this chapter

RG language	what it means physically
coupling g_i, theory space	the numbers in the action; the space of all actions
RG flow, $\beta_i(g)$	how the effective couplings change as we zoom out
fixed point	a theory with no length scale (critical point, CFT)
relevant / irrelevant	grows / dies away at long distances
RG eigenvalue y	d minus the scaling dimension of the operator
critical surface	theories that flow into the fixed point
Wilsonian action S_Λ	the action that reproduces all physics below Λ

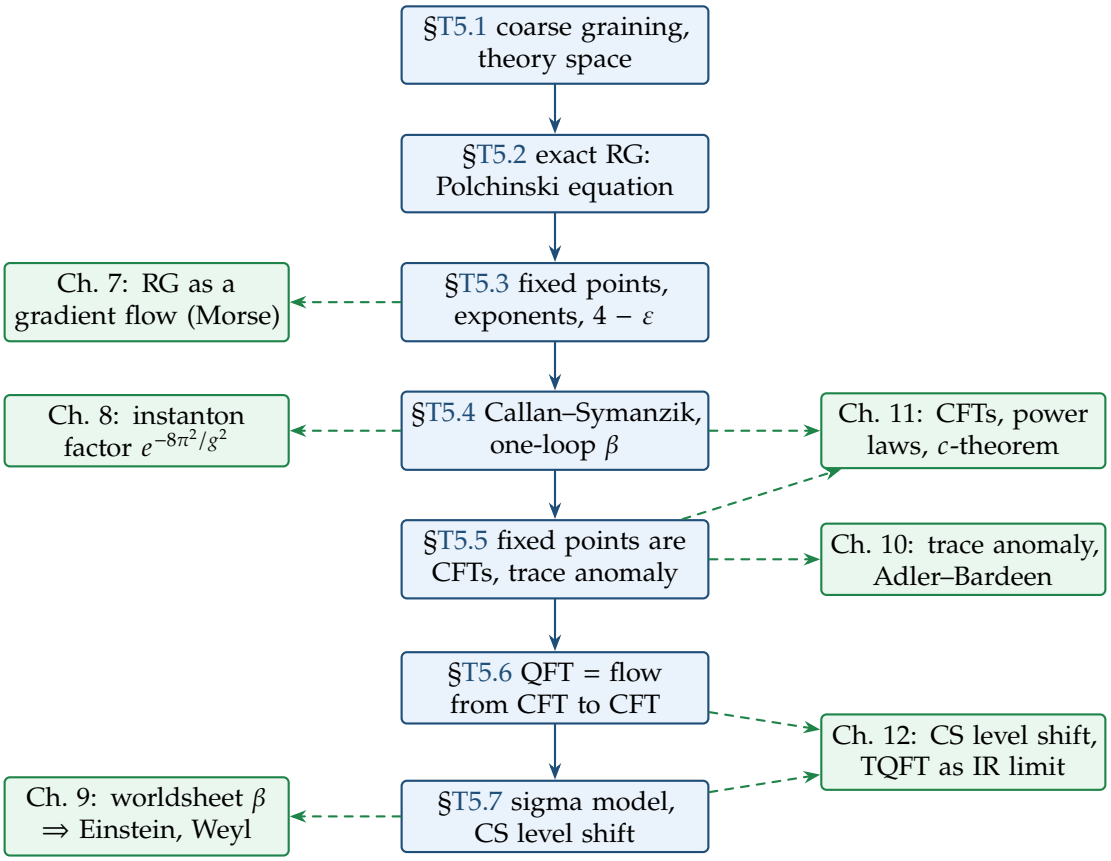


Figure T5.1: The logic of this chapter (blue) and the later chapters that use it (green). The main line runs from coarse graining to the statement that a quantum field theory is an RG trajectory between two conformal field theories, and ends with two β-functions that carry geometry. What to see: every Nash chapter from 7 to 12 uses one piece of this line.

T5.1 The renormalisation group: physics at a scale

Two puzzles sit behind everything in this chapter. The first comes from particle physics. Nobody knows what happens at distances of 10^{-35} m, yet quantum electrodynamics predicts the magnetic moment of the electron to twelve digits using only two measured numbers, e and m_e. Why does ignorance of the shortest distances not spoil predictions at long ones? The second comes from statistical mechanics. The critical exponents of a liquid–gas transition, of a uniaxial magnet and of a binary alloy are the same numbers (in three dimensions $\nu \approx 0.630$, $\eta \approx 0.036$), although the microscopic physics could hardly be more different. Why?

Both puzzles have the same answer, the **renormalisation group** (RG). The same machinery gives the Callan–Symanzik equation and anomalous dimensions (Chapter 11), the β-functions of the string worldsheet (Chapter 9) and the scale anomaly (Chapter 10). Its end point is the statement that makes conformal field theory relevant to ordinary quantum field theory: *a QFT is an RG flow from one CFT to another.*

Physical picture: a camera with finite resolution

Photograph a magnet near its critical point, then zoom out by a factor of two and photograph it again with the same camera. Each pixel of the new picture averages four pixels of the old one. Fine detail is lost, but the picture is still a picture of a magnet, described by the same kind of rules with slightly different numbers (a different effective temperature and field). The RG is the map "old numbers $\mapsto$ new numbers". Iterating it is zooming out further and further. What survives many iterations is what we see in the laboratory.

T5.1.1 Coarse graining a lattice: block spins

Kadanoff's version of the idea is the easiest to draw. Take the Ising model, the simplest model of a uniaxial magnet (its two-dimensional version is solved exactly by conformal field theory in Chapter 11): spins $s_i = \pm 1$ on a square lattice with spacing a, weight e^{-H} with

$$-H = K \sum_{\langle ij \rangle} s_i s_j + h \sum_i s_i$$

($K = J/k_B T$ and h is the magnetic field in units of $k_B T$). Group the spins into $b \times b$ blocks and give each block a single **block spin** $s'_I = \pm 1$, for example by majority rule (Figure T5.2). Then sum the Boltzmann weight over all configurations of the original spins that are compatible with a given configuration of block spins. What is left is a Boltzmann weight $e^{-H'}$ for the block spins alone.

Working definition T5.1 (RG transformation)

The **couplings** $\{K_1, K_2, \dots\}$ [the numbers in the Boltzmann weight: nearest-neighbour coupling, four-spin coupling, field] specify a statistical system. An **RG transformation** R_b [one step of zooming out by a factor b] is the rule $\{K_i\} \mapsto \{K'_i\}$ obtained by

1. summing the Boltzmann weight over the degrees of freedom on length scales between a and ba [the spins inside each block of Figure T5.2],
2. rescaling lengths by $1/b$ [so that the new lattice spacing is again called a],
3. reading off the new couplings K'_i from the Boltzmann weight of the block spins that is left.

Operational test: the partition function is unchanged, $Z(K') = Z(K)$ up to a spin-independent factor, and every long-distance quantity is the same function of the new couplings, measured in the new units. In particular the correlation length [the size of a typical patch of aligned

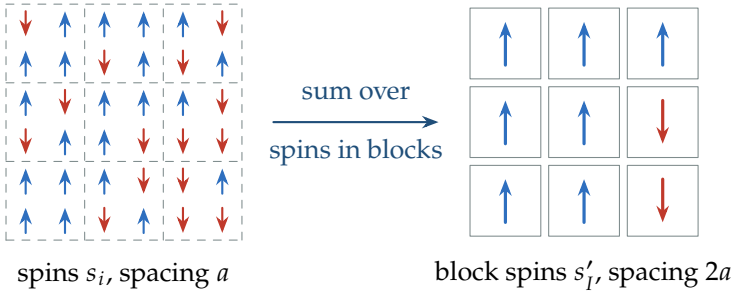


Figure T5.2: One block-spin step with $b = 2$. Left: a configuration of 36 Ising spins; the dashed lines cut the lattice into 2×2 blocks. Right: each block is replaced by one spin, here by majority rule (a tie is broken upwards). The new lattice has spacing $2a$ and 9 spins. What to see: the new picture has the same kind of variables as the old one, so it can be described by the same kind of Hamiltonian with new couplings (K', h').

spins, in units of the lattice spacing] obeys

$$\xi(K') = \frac{\xi(K)}{b}. \tag{T5.1}$$

Equation (T5.1) is worth pausing on. The physical correlation length (in metres) does not change, because nothing physical has been done. But we now measure it in units of the new lattice spacing ba, so the number ξ drops by b. Two consequences follow at once.

- A **fixed point** $K^* = R_b(K^*)$ has $\xi(K^*) = \xi(K^*)/b$, so $\xi = 0$ or $\xi = \infty$. The first kind are trivial (zero or infinite temperature). The second kind are *critical points*. A fixed point with $\xi = \infty$ has no length scale at all: it is a scale-invariant theory, the subject of Chapter 11.

- Couplings that approach a fixed point under repeated RG steps describe systems that look, at long distances, exactly like the fixed point. All such systems share the same long-distance physics. This is **universality**.

Working definition T5.2 (correlation length, scaling dimension, critical exponents)

Take the connected correlation $G(r) = \langle s(x)s(0)\rangle - \langle s\rangle^2$ [how much a spin at x still "knows" about the spin at 0, a distance r away]. Away from a critical point it dies exponentially, $G \sim e^{-r/\xi}$, and ξ is the **correlation length** [the size of a typical correlated patch]. As the reduced temperature $t = (T - T_c)/T_c$ goes to zero, ξ diverges as $\xi \propto |t|^{-\nu}$. At $t = 0$ no length is left, so G is a pure power,

$$G(r) \propto \frac{1}{r^{2\Delta}},$$

and Δ is the **scaling dimension** of the local operator [the power of length it carries once all scales are gone, read off from how fast its correlations fall]. For the order parameter one writes $2\Delta = d - 2 + \eta$. The numbers ν and η are the first two **critical exponents** [the powers that describe the singular behaviour near T_c]. *Operational test:* on a log–log plot of $G(r)$ at T_c, the slope is -2Δ; on a log–log plot of ξ against $|t|$, the slope is $-\nu$. (For the two-dimensional Ising model $\nu = 1$ and $\eta = \frac{1}{4}$ exactly; the full list of exponents is derived from conformal field theory in Chapter 11.)

T5.1.2 An exact RG step: the one-dimensional Ising chain

In one dimension we can do a block-spin step exactly. Instead of majority rule we use **decimation**: sum over every second spin and keep the others.

Worked example T5.1: Decimating the Ising chain

Take $h = 0$ and the Boltzmann weight $\prod_i e^{K s_i s_{i+1}}$. A spin s_2 sitting between s_1 and s_3 appears in exactly two factors. Summing over it,

$$\sum_{s_2=\pm 1} e^{K s_1 s_2} e^{K s_2 s_3} \overset{(1)}{=} \sum_{s_2=\pm 1} e^{K s_2 (s_1+s_3)} \overset{(2)}{=} 2\cosh\bigl(K(s_1+s_3)\bigr).$$

(1) Combine the exponents.
(2) $e^{x} + e^{-x} = 2\cosh x$.

We want this to have the old form, $A\, e^{K' s_1 s_3}$, with a new coupling K' and a spin-independent factor A. There are only two cases:

$$\begin{aligned} s_1 = s_3 &: \quad 2\cosh 2K = A\, e^{K'}, \\ s_1 = -s_3 &: \quad 2 = A\, e^{-K'}. \end{aligned}$$

Dividing and multiplying the two lines,

$$e^{2K'} \overset{(1)}{=} \cosh 2K, \qquad A^2 \overset{(2)}{=} 4\cosh 2K. \tag{T5.2}$$

(1) Divide the first line by the second.
(2) Multiply the two lines.

An equivalent and nicer form uses tanh:

$$\tanh K' \overset{(1)}{=} \frac{e^{2K'}-1}{e^{2K'}+1} \overset{(2)}{=} \frac{\cosh 2K - 1}{\cosh 2K + 1} \overset{(3)}{=} \frac{2\sinh^2 K}{2\cosh^2 K} = \tanh^2 K.$$

(1) Definition of tanh, multiplied top and bottom by $e^{K'}$.
(2) (T5.2).
(3) $\cosh 2K = 1 + 2\sinh^2 K = 2\cosh^2 K - 1$.

So one RG step with $b = 2$ is the map $\tanh K \mapsto \tanh^2 K$.

The map has two fixed points, $\tanh K = 0$ and $\tanh K = 1$. Between them, $\tanh^2 K < \tanh K$, so every finite coupling decreases under coarse graining (Figure T5.3). The $K = \infty$ fixed point is *unstable*: start near it and you are pushed away. The $K = 0$ fixed point is *stable*: everything ends there. Let us check (T5.1) with the exact correlation length.

Worked example T5.2: The correlation length of the chain obeys $\xi' = \xi/2$

First rewrite each bond factor. Since $s s' = \pm 1$,

$$e^{K s s'} \overset{(1)}{=} \cosh K + s s' \sinh K \overset{(2)}{=} \cosh K \,\bigl(1 + s s' \tanh K\bigr).$$

(1) For $s s' = +1$ both sides are $e^{K} = \cosh K + \sinh K$; for $s s' = -1$ both are $e^{-K} = \cosh K - \sinh K$.
(2) Take out the factor $\cosh K$ and use $\tanh K = \sinh K / \cosh K$.

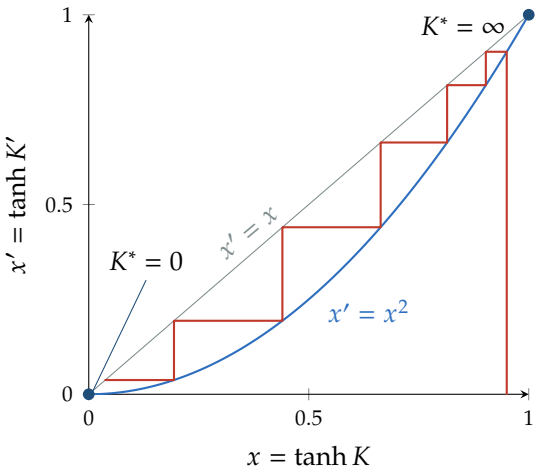


Figure T5.3: The decimation map of the 1D Ising chain, $\tanh K' = \tanh^2 K$ (blue), and a "cobweb" of repeated RG steps (red) starting from $\tanh K = 0.95$. Each step goes up to the curve and across to the diagonal. What to see: the only fixed points are the ends, $K = 0$ (infinite temperature) and $K = \infty$ (zero temperature). Every finite K flows to $K = 0$: at long distances a 1D magnet at any $T > 0$ looks disordered, so there is no phase transition at finite temperature.

For an open chain of N spins $s_0, \dots, s_{N-1}$ with $0 < r < N$,

$$\langle s_0 s_r \rangle \overset{(1)}{=} \frac{\sum_{\{s\}} s_0 s_r \prod_i (1 + s_i s_{i+1} \tanh K)}{\sum_{\{s\}} \prod_i (1 + s_i s_{i+1} \tanh K)} \overset{(2)}{=} \frac{2^N \tanh^r K}{2^N} = \tanh^r K.$$

(1) The factors $\cosh K$ cancel between numerator and denominator.

(2) Expand the product. Summing $s_j = \pm 1$ kills every term in which some spin appears an odd number of times, since $\sum_{s=\pm 1} s = 0$. In the denominator only the term 1 survives. In the numerator, s_0 and s_r must be paired by the bonds $(0,1), (1,2), \dots, (r-1, r)$, each contributing $\tanh K$; this chain of bonds is the only surviving term. Each surviving term gives 2^N (N spins, each summed over two values).

Since $\tanh^r K = e^{r \ln \tanh K}$ and $\ln \tanh K < 0$ for $0 < K < \infty$, this is $\langle s_0 s_r \rangle = e^{-r/\xi}$ with $\xi(K) = -1/\ln \tanh K$, measured in lattice spacings. After one decimation step,

$$\xi(K') \overset{(1)}{=} \frac{-1}{\ln \tanh K'} \overset{(2)}{=} \frac{-1}{\ln \tanh^2 K} = \frac{1}{2} \frac{-1}{\ln \tanh K} = \frac{\xi(K)}{2},$$

(1) The formula for ξ applied to the new coupling.

(2) $\tanh K' = \tanh^2 K$ from Example T5.1.

exactly (T5.1) with $b = 2$. How large is ξ at low temperature, where $K = J/k_B T$ is large? Write $q := e^{-2K} \ll 1$:

$$\tanh K \overset{(1)}{=} \frac{1 - e^{-2K}}{1 + e^{-2K}} \overset{(2)}{=} (1 - q)(1 - q + q^2 - \cdots) \overset{(3)}{=} 1 - 2q + O(q^2),$$
$$\xi \overset{(4)}{=} \frac{-1}{\ln(1 - 2q + O(q^2))} \overset{(5)}{=} \frac{1}{2q + O(q^2)} \approx \tfrac{1}{2} e^{2K} = \tfrac{1}{2} e^{2J/k_B T}.$$

(1) Multiply numerator and denominator of $\sinh K / \cosh K$ by e^{-K}.

(2) Geometric series $1/(1+q) = 1 - q + q^2 - \cdots$.

(3) Multiply out and keep the first order in q.

(4) $\xi = -1/\ln \tanh K$.

(5) $\ln(1-x) = -x + O(x^2)$.

The correlation length is finite at every $T > 0$ and diverges only as $T \to 0$.

Watch out: the RG is a semigroup, not a group

Decimation can be iterated but not undone: from K' we cannot recover the configurations of the spins we summed over. So R_b has no inverse in general, and the "group" in renormalisation group is a historical misnomer (only the composition law $R_b R_{b'} = R_{bb'}$ holds). In continuum language we will still run the flow both ways, $\Lambda \to \Lambda e^{\pm \delta t}$, but only the direction towards long distances has the meaning "integrate out".

T5.1.3 Integrating out modes in a field theory

In a continuum field theory the analogue of "summing over the spins inside a block" is Wilson's **integrating out**. Work in Euclidean signature: after the Wick rotation $t = -i\tau$ the weight e^{iS} becomes a Boltzmann weight e^{-S_E}, so a Euclidean field theory in d dimensions is a statistical system in d dimensions. Fourier transform the field, $\varphi(x) = \int \frac{\mathrm{d}^d p}{(2\pi)^d}\, e^{ip\cdot x} \varphi(p)$. A theory with an ultraviolet cutoff Λ contains only the modes with $|p| < \Lambda$; the cutoff plays the role of the inverse lattice spacing, $\Lambda \sim 1/a$. Split

$$\varphi = \varphi_< + \varphi_>, \qquad \varphi_< : \ |p| < \Lambda/b, \qquad \varphi_> : \ \Lambda/b < |p| < \Lambda, \tag{T5.3}$$

and define the new action of the slow modes by doing the path integral over the fast ones:

$$e^{-S_{\Lambda/b}[\varphi_<]} := \int \mathcal{D}\varphi_>\, e^{-S_\Lambda[\varphi_< + \varphi_>]}. \tag{T5.4}$$

By construction $\int \mathcal{D}\varphi_<\, e^{-S_{\Lambda/b}[\varphi_<]} = \int \mathcal{D}\varphi\, e^{-S_\Lambda[\varphi]}$: the partition function, and every correlation function of slow modes, is unchanged. The functional S_Λ is the **Wilsonian effective action** at scale Λ. Figure T5.4 shows the shell of modes that is removed.

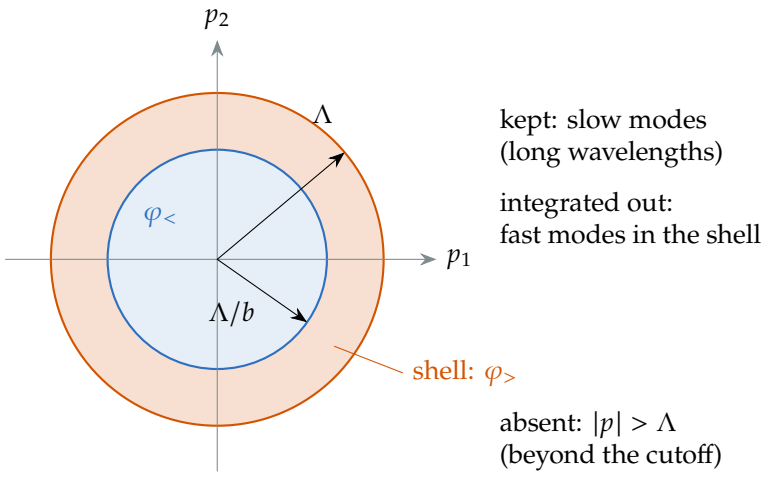


Figure T5.4: Wilson's momentum shell in $d = 2$. The theory with cutoff Λ contains the modes inside the orange circle. One RG step integrates out the orange annulus $\Lambda/b < |p| < \Lambda$ and keeps the blue disc. Then momenta are rescaled by b so that the blue disc becomes a disc of radius Λ again. What to see: this is the block-spin step of Figure T5.2 in Fourier space; short wavelengths are thrown away and their effect is absorbed into new couplings.

What does the integral over fast modes do to the couplings? The simplest possible model shows it: a "field theory in zero dimensions", in which the path integral is an ordinary integral over two numbers.

Worked example T5.3: Integrating out a heavy variable generates a new coupling

Let x be a light variable and y a heavy one, with action

$$S(x,y) = \tfrac{1}{2}x^2 + \tfrac{1}{2}M^2y^2 + g\,x^2y, \qquad M \gg 1.$$

Integrate out y. The y dependence is quadratic, so complete the square:

$$\tfrac{1}{2}M^2y^2 + gx^2y \overset{(1)}{=} \tfrac{1}{2}M^2\Big(y + \frac{gx^2}{M^2}\Big)^2 - \frac{g^2x^4}{2M^2},$$

$$\int_{-\infty}^{\infty} \mathrm{d}y\, e^{-S(x,y)} \overset{(2)}{=} e^{-\frac{1}{2}x^2 + \frac{g^2x^4}{2M^2}} \int_{-\infty}^{\infty} \mathrm{d}y'\, e^{-\frac{1}{2}M^2y'^2} \overset{(3)}{=} \sqrt{\frac{2\pi}{M^2}}\; e^{-\frac{1}{2}x^2 + \frac{g^2}{2M^2}x^4}.$$

(1) Expand the square: $\frac{1}{2}M^2(y^2 + 2gx^2y/M^2 + g^2x^4/M^4)$; the last term is subtracted again.
(2) Shift $y' = y + gx^2/M^2$; the shift does not change $\int \mathrm{d}y$ over the whole line.
(3) Gaussian integral $\int \mathrm{d}y\, e^{-\frac{1}{2}Ay^2} = \sqrt{2\pi/A}$.

The effective action of the light variable is defined, as in (T5.4), by $e^{-S_{\text{eff}}(x)} = \int \mathrm{d}y\, e^{-S(x,y)}$:

$$S_{\text{eff}}(x) \overset{(1)}{=} -\ln \int \mathrm{d}y\, e^{-S(x,y)} \overset{(2)}{=} \tfrac{1}{2}x^2 - \frac{g^2}{2M^2}x^4 - \ln\sqrt{\frac{2\pi}{M^2}}$$

$$\overset{(3)}{=} \tfrac{1}{2}x^2 + \frac{\lambda_{\text{eff}}}{4!}x^4 + \text{const}, \qquad \lambda_{\text{eff}} \overset{(4)}{=} -\frac{12\,g^2}{M^2}.$$

(1) Take $-\ln$ of the definition.
(2) The result of the previous chain; $-\ln e^X = -X$.
(3) Name the x^4 coefficient $\lambda_{\text{eff}}/4!$, the usual normalisation of a quartic coupling. The constant does not depend on x, so it drops out of every average.
(4) $\lambda_{\text{eff}}/24 = -g^2/(2M^2)$, so $\lambda_{\text{eff}} = -24g^2/(2M^2)$.

Three lessons. (i) Integrating out *generates* a coupling (x^4) that was not in the original action. (ii) The new coupling is suppressed by $1/M^2$, a power of the heavy scale. (iii) Its value is fixed by the old couplings, g and M.

Fermi's four-fermion interaction arises in exactly this way from W-boson exchange. At momenta far below M_W the kinetic term of the W is negligible, and the terms of the (Minkowski) Lagrangian that contain it are $\mathcal{L}_W = M_W^2 W_\mu^+ W^{-\mu} - \frac{g_W}{2\sqrt{2}}\big(W_\mu^+ J^{-\mu} + W_\mu^- J^{+\mu}\big)$, where g_W is the weak coupling and $J^{\pm\mu}$ are the charged currents built from the lepton and quark fields. With $a := g_W/(2\sqrt{2}M_W^2)$,

$$\mathcal{L}_W \overset{(1)}{=} M_W^2\big(W_\mu^+ - aJ_\mu^+\big)\big(W^{-\mu} - aJ^{-\mu}\big) - M_W^2 a^2 J_\mu^+ J^{-\mu} \;\longrightarrow\; \mathcal{L}_{\text{eff}} \overset{(2)}{=} -\frac{g_W^2}{8M_W^2}\, J_\mu^+ J^{-\mu}.$$

(1) Complete the square, as for y above: the cross terms of the product are $-M_W^2 a(W_\mu^+ J^{-\mu} + J_\mu^+ W^{-\mu})$ and $M_W^2 a = g_W/(2\sqrt{2})$; the a^2 term of the product is subtracted again.
(2) Shift W and integrate it out; the shifted Gaussian integral is a constant. What is left is $-M_W^2 a^2 = -g_W^2/(8M_W^2)$.

Fermi wrote this interaction as $-\frac{G_F}{\sqrt{2}} J_\mu^+ J^{-\mu}$, so $G_F/\sqrt{2} = g_W^2/(8M_W^2)$: a coupling of negative mass dimension, suppressed by the heavy mass, exactly as λ_{eff}.

In a real field theory the fast modes are not a single heavy variable but a whole shell, and the integral cannot be done exactly unless the theory is free. Two cases can nevertheless be handled. A shell of infinitesimal thickness can be integrated out exactly (Section T5.2), and a thin shell at weak coupling can be integrated out in perturbation theory (Section T5.3). The three lessons of Example T5.3 hold in both.

T5.1.4 Theory space, running couplings and power counting

The heavy-variable example shows that the RG does not stay inside any finite family of actions: it generates every coupling that the symmetries allow. So the arena of the RG is large.

Working definition T5.3 (theory space and RG flow)

List all local operators $O_i(x)$ allowed by the symmetries [for a $\mathbb{Z}_2$-symmetric scalar: φ^2, φ^4, $(\partial\varphi)^2$, φ^6, $\varphi^2(\partial\varphi)^2$, …]. A general Wilsonian action [the action that reproduces all physics below the cutoff Λ] is

$$S_\Lambda[\varphi] = \sum_i g_i(\Lambda) \int \mathrm{d}^d x \, O_i(x).$$

The **theory space** [one axis for each coupling g_i, so one point for each possible action] is the space of coupling vectors $\{g_i\}$. Lowering the cutoff moves the point. The **RG flow** [an arrow at every point saying where it moves as we zoom out] is this vector field. We parametrise it by the **RG time** [the number of e-folds by which we have zoomed out]

$$t = \ln \frac{\Lambda_0}{\Lambda}, \qquad \frac{\mathrm{d}\hat{g}_i}{\mathrm{d}t} = \beta_i(\hat{g}), \tag{T5.5}$$

where Λ_0 is the cutoff we started with and $\hat{g}_i = g_i \Lambda^{-[g_i]}$ are the **dimensionless couplings** [each coupling measured in units of the current cutoff]. Only these can have fixed points, since a dimensionful coupling changes under any change of units.

Rewriting in dimensionless couplings has an immediate consequence even without any interaction. Suppose g_i has mass dimension $[g_i]$ and does not change at all when modes are integrated out. Then

$$\frac{\mathrm{d}\hat{g}_i}{\mathrm{d}t} \overset{(1)}{=} -\Lambda \frac{\mathrm{d}}{\mathrm{d}\Lambda}\left(g_i \Lambda^{-[g_i]}\right) \overset{(2)}{=} [g_i]\, g_i \Lambda^{-[g_i]} = [g_i]\, \hat{g}_i. \tag{T5.6}$$

(1) $\mathrm{d}t = -\mathrm{d}\Lambda/\Lambda$ from (T5.5).

(2) g_i is constant by assumption; differentiate the power of Λ.

So $\hat{g}_i(t) = \hat{g}_i(0)\, e^{[g_i]t}$: a coupling of positive mass dimension grows towards the infrared, one of negative dimension dies away. This is the **classical** (or **canonical**) part of every β-function. The number $[g_i]$ is the RG eigenvalue y_i (defined in general in Section T5.3.1) at the free (Gaussian) fixed point.

Working definition T5.4 (relevant, marginal, irrelevant)

Near a fixed point a coupling grows or shrinks like e^{yt}, with y its RG eigenvalue [the growth rate per e-fold of zooming out]. It is **relevant** if $y > 0$ [it matters more and more at long distances, like the temperature deviation of a magnet], **irrelevant** if $y < 0$ [its effect dies off as a power of the length scale, like the x^4 coupling suppressed by $1/M^2$ in Example T5.3], and **marginal** if $y = 0$ [whether it grows or shrinks is decided by the β_i

beyond linear order]. *Test:* $y_i = d - \Delta_i$, where Δ_i is the scaling dimension of the operator $\mathcal{O}_i$ that the coupling multiplies (derived in Section T5.3.1, (T5.48)).

Worked example T5.4: Power counting at the Gaussian fixed point

Count mass dimensions, with $\hbar = c = 1$: a length has dimension -1, so $[\mathrm{d}^d x] = -d$ and $[\partial_\mu] = +1$. At the free massless point the action $\frac{1}{2}\int \mathrm{d}^d x (\partial\varphi)^2$ is a pure number, so

$$0 \overset{(1)}{=} [\mathrm{d}^d x] + 2[\partial] + 2[\varphi] = -d + 2 + 2[\varphi] \quad \Longrightarrow \quad [\varphi] = \tfrac{d-2}{2}.$$

(1) Dimensions of a product add; the integrand has two derivatives and two fields.

An operator with n fields and k derivatives therefore has $\Delta = n\frac{d-2}{2} + k$. The action $g\int \mathrm{d}^d x\, \mathcal{O}$ is dimensionless, so

$$y = [g] \overset{(1)}{=} d - \Delta \overset{(2)}{=} d - n\,\frac{d-2}{2} - k.$$

(1) $[\mathrm{d}^d x] = -d$, and the three dimensions must add to zero.
(2) Count the fields and derivatives.

For $\mathcal{O} = \varphi^n$ ($k = 0$):

	φ^2	φ^4	φ^6	φ^8
$d = 4$	2	0	-2	-4
$d = 3$	2	1	0	-1
$d = 2$	2	2	2	2

In four dimensions the mass is relevant, φ^4 is marginal, and everything else is irrelevant: this is why $\lambda\varphi^4$ is the most general renormalisable scalar theory. In three dimensions φ^4 is relevant (and φ^6 marginal). In two dimensions $[\varphi] = 0$, every φ^n is relevant, and the Gaussian counting is useless; this is one reason two-dimensional critical points need the exact methods of conformal field theory (Chapter 11). The derivative operator $(\partial\varphi)^2$ has $y = 0$ in every d: it is the kinetic term, which is marginal by construction (we use it to define the normalisation of φ).

Figure T5.5 is the picture to keep in mind. It explains both puzzles at the start of this section. Universality: magnets with very different microscopic couplings sit at different points of theory space, but all are pushed onto the same flow line near the fixed point, so they share its exponents. Predictivity: a QFT with an unknown ultraviolet completion starts somewhere unknown, but by the time the flow reaches accessible energies it has been squeezed onto a low-dimensional surface labelled by the few relevant and marginal couplings (here, e and m_e). We make the squeezing precise in Section T5.2.6.

Watch out: two sign conventions for β

The Callan–Symanzik treatment of Chapter 11 (and [QFT notes §15.4]) writes $\beta(g) = M\,\mathrm{d}g/\mathrm{d}M$, with M a renormalisation scale that *increases* towards the ultraviolet. Here the RG time $t = \ln(\Lambda_0/\Lambda)$ increases towards the *infrared*, and $\mathrm{d}\hat{g}/\mathrm{d}t = \beta(\hat{g})$. Identify the two

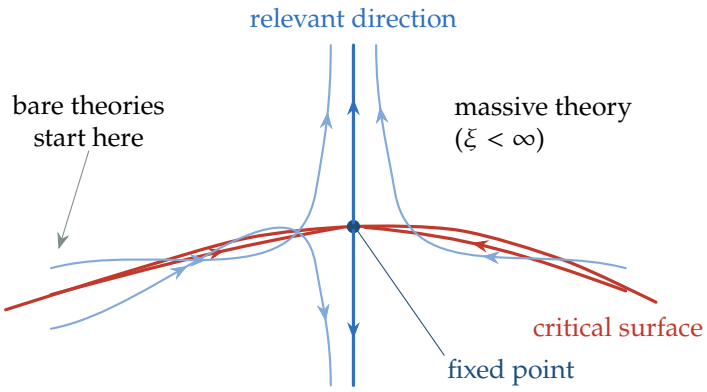


Figure T5.5: A cartoon of theory space. The red curve stands for the **critical surface**: all theories that flow into the fixed point. Its directions are the irrelevant couplings. The blue arrow leaving the fixed point is a relevant coupling (for a magnet: the temperature deviation). A theory that starts slightly off the critical surface (thin blue curves) first approaches the fixed point, lingers near it, and is then expelled along the relevant direction. What to see: the long-distance physics of all such theories is controlled by the fixed point and by one number, how far off the surface they started.

scales, $M = \Lambda$ (this is exact at the leading order where the two schemes agree). Then

$$\frac{\mathrm{d}\hat{g}}{\mathrm{d}t} \overset{(1)}{=} \frac{\mathrm{d}\Lambda}{\mathrm{d}t}\,\frac{\mathrm{d}\hat{g}}{\mathrm{d}\Lambda} \overset{(2)}{=} -\Lambda\,\frac{\mathrm{d}\hat{g}}{\mathrm{d}\Lambda} \overset{(3)}{=} -M\,\frac{\mathrm{d}\hat{g}}{\mathrm{d}M}.$$

(1) Chain rule.
(2) $\Lambda = \Lambda_0 e^{-t}$, so $\mathrm{d}\Lambda/\mathrm{d}t = -\Lambda$.
(3) $M = \Lambda$.

So the two β's differ by a sign. Relevant couplings have $\mathrm{d}\hat{g}/\mathrm{d}t > 0$ here, and $M\,\mathrm{d}g/\mathrm{d}M < 0$ in the other convention. For example, φ^4 in $d = 4 - \varepsilon$ has $M\,\mathrm{d}g/\mathrm{d}M = -\varepsilon g + 3g^2/16\pi^2$ (the form used in Chapter 11), which is $\mathrm{d}g/\mathrm{d}t = \varepsilon g - 3g^2/16\pi^2$ here; this β-function is derived by a shell integration in Section T5.3.3, (T5.54).

Reading Nash (p. 301, Ch. XI §1)

Nash devotes one sentence to the RG, in his chapter on conformal field theory: "the renormalisation properties of correlation functions are constrained by the need to obey the Callan–Symanzik renormalisation group equations", and then states that a zero of $\beta(g)$ is "necessary, but insufficient" for conformal invariance. The Callan–Symanzik equation and its consequence at a fixed point are unpacked in Chapter 11. Behind Nash's sentence lie three facts he takes for granted: the flow comes from integrating out short distances (Sections T5.1 and T5.2), its fixed points carry the critical exponents (Section T5.3), and the "insufficient" condition becomes sufficient for unitary two-dimensional theories (Section T5.5).

Recap

- An RG step removes the degrees of freedom between two length scales and rescales lengths; the couplings move, $K \mapsto R_b(K)$, and $\xi \mapsto \xi/b$.
- Fixed points have $\xi = 0$ or $\xi = \infty$; the latter are the scale-invariant critical theories.
- Exact example: decimating the 1D Ising chain gives $\tanh K' = \tanh^2 K$; every finite

coupling flows to $K = 0$, and $\xi(K') = \xi(K)/2$ exactly.

- Integrating out generates new couplings, suppressed by powers of the heavy scale and fixed by the old couplings ($\lambda_{\text{eff}} = -12g^2/M^2$).
- Near the Gaussian point a coupling of O scales with $y = d - \Delta_O$; $y > 0$ relevant, $y < 0$ irrelevant, $y = 0$ marginal.

T5.2 The exact renormalisation group

The definition (T5.4) integrates out a shell of finite thickness. For an interacting theory that integral cannot be done in closed form. But a shell of *infinitesimal* thickness can be integrated out exactly, because in a very thin shell the fast modes are almost free: their propagator is tiny, so only the first order in it matters. Doing this step after step turns (T5.4) into a differential equation for the Wilsonian action, the **exact renormalisation group equation**. Several versions exist (Wilson 1974, Wegner and Houghton 1973, Wetterich 1993). The one due to Polchinski has the simplest form, and it gives a short proof that $\lambda\varphi^4$ theory in four dimensions is renormalisable. Two questions organise this section. What equation does the Wilsonian action obey as the cutoff is lowered? And why does the answer at low energies forget almost everything about the theory we started with?

T5.2.1 A smooth cutoff

A sharp cutoff (keep $|p| < \Lambda$, drop the rest) produces awkward non-local effects in position space, because the Fourier transform of a step function decays only slowly. It is cleaner to suppress high momenta smoothly.

Working definition T5.5 (regularised propagator and shell propagator)

Choose a **cutoff function** $K(x)$ [a dial that is 1 for modes we keep and 0 for modes we have removed, with $x = p^2/\Lambda^2$] with $K(0) = 1$, decreasing, and falling to zero faster than any power as $x \to \infty$. The standard example is $K(x) = e^{-x}$. The **regularised propagator** at scale Λ [the propagator of only those modes that are still present at cutoff Λ] is

$$C_\Lambda(p) = \frac{K(p^2/\Lambda^2)}{p^2}. \tag{T5.7}$$

Picture it as the ordinary massless propagator $1/p^2$ for $|p| \ll \Lambda$, killed off for $|p| \gg \Lambda$. (The massive version $K/(p^2 + m^2)$ works equally well; nothing below depends on the choice.) The **shell propagator** [the propagator of the thin layer of modes removed when Λ is lowered a little] is the logarithmic derivative with respect to the cutoff,

$$\Delta_\Lambda(p) := \Lambda\frac{\partial C_\Lambda(p)}{\partial\Lambda} = -\frac{2}{\Lambda^2}\,K'\!\left(\frac{p^2}{\Lambda^2}\right) \geq 0. \tag{T5.8}$$

Operational meaning: lowering the cutoff from Λ to $\Lambda - \delta\Lambda$ removes from C_Λ exactly the propagator of the modes near $|p| \sim \Lambda$,

$$C_\Lambda - C_{\Lambda-\delta\Lambda} = \Delta_\Lambda\,\frac{\delta\Lambda}{\Lambda} + O(\delta\Lambda^2). \tag{T5.9}$$

Let us check (T5.8). With $x = p^2/\Lambda^2$,

$$\Lambda\frac{\partial}{\partial\Lambda}\frac{K(p^2/\Lambda^2)}{p^2} \overset{(1)}{=} \frac{1}{p^2}K'(x)\,\Lambda\frac{\partial x}{\partial\Lambda} \overset{(2)}{=} \frac{1}{p^2}K'(x)\left(-\frac{2p^2}{\Lambda^2}\right) = -\frac{2}{\Lambda^2}K'(x).$$

(1) Chain rule; $1/p^2$ does not depend on Λ.
(2) $\Lambda\,\partial_\Lambda(p^2\Lambda^{-2}) = -2p^2\Lambda^{-2}$.

Because K is decreasing, $K' \le 0$ and $\Delta_\Lambda \ge 0$: a larger cutoff contains more modes. For $K(x) = e^{-x}$, $K'(x) = -e^{-x}$ and so $\Delta_\Lambda(p) = 2e^{-p^2/\Lambda^2}/\Lambda^2$. Which modes does it describe? In d dimensions the modes with $|p|$ between p and $p + \mathrm{d}p$ fill a spherical shell of volume $S_{d-1}p^{d-1}\mathrm{d}p$ (with S_{d-1} the area of the unit sphere), so their weight is $p^{d-1}\Delta_\Lambda(p)$. Its maximum is at

$$0 = \frac{\mathrm{d}}{\mathrm{d}p}\left(p^{d-1}e^{-p^2/\Lambda^2}\right) \overset{(1)}{=} \left(\frac{d-1}{p} - \frac{2p}{\Lambda^2}\right)p^{d-1}e^{-p^2/\Lambda^2} \quad\Longrightarrow\quad p \overset{(2)}{=} \sqrt{\tfrac{d-1}{2}}\,\Lambda,$$

(1) Product rule.
(2) Set the bracket to zero and solve for $p > 0$.

which is $1.22\,\Lambda$ in $d = 4$: the removed modes sit near $|p| \approx \Lambda$ (Figure T5.6).

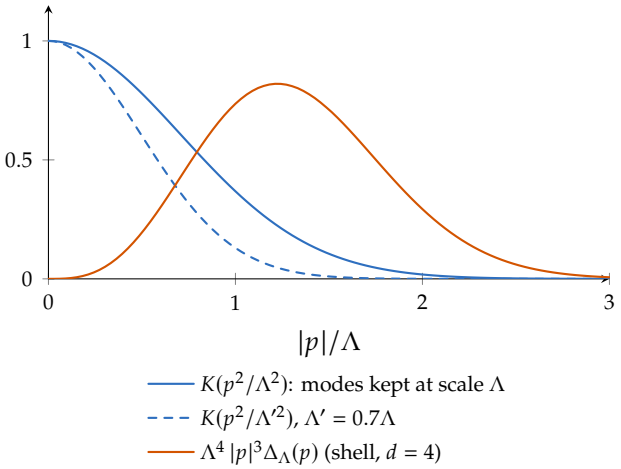


Figure T5.6: The smooth cutoff $K(x) = e^{-x}$. Solid blue: the fraction of each mode kept at scale Λ. Dashed blue: the same at a lower scale Λ'; the modes between the two curves have been integrated out. Orange: the radial weight of the shell propagator in four dimensions, $\Lambda^4|p|^3\Delta_\Lambda = 2(|p|/\Lambda)^3e^{-p^2/\Lambda^2}$. What to see: a thin RG step removes a smeared shell of modes centred near $|p| \approx \Lambda$, the smooth version of the annulus in Figure T5.4.

We write the Wilsonian action at scale Λ as a regularised kinetic term plus everything else,

$$S_\Lambda[\varphi] = \tfrac{1}{2}\,\varphi\cdot C_\Lambda^{-1}\cdot\varphi + S_\Lambda^{\mathrm{I}}[\varphi], \qquad \varphi\cdot A\cdot\varphi \coloneqq \int\frac{\mathrm{d}^dp}{(2\pi)^d}\,\varphi(-p)A(p)\varphi(p). \tag{T5.10}$$

The **interaction action** S_Λ^{I} contains all interactions, and may also contain mass terms and corrections to the kinetic term. In position space A becomes a kernel $A(x-y)$ and $\varphi\cdot A\cdot\varphi = \int \mathrm{d}^dx\,\mathrm{d}^dy\,\varphi(x)A(x-y)\varphi(y)$. We use the same dot notation for functional derivatives, for example

$$\frac{\delta}{\delta\varphi}\cdot A\cdot\frac{\delta}{\delta\varphi} \coloneqq \int \mathrm{d}^dx\,\mathrm{d}^dy\,A(x-y)\,\frac{\delta}{\delta\varphi(x)}\frac{\delta}{\delta\varphi(y)}.$$

T5.2.2 Two facts about Gaussian integrals

Everything in the derivation rests on two properties of Gaussian integrals. We state them for N real variables $\chi = (\chi_1, \dots, \chi_N)$; a field is the case $N \to \infty$ with the index i replaced by x (or p). For a positive symmetric matrix D write

$$\langle F \rangle_D := \frac{\int \mathrm{d}^N\chi \; e^{-\frac{1}{2}\chi\cdot D^{-1}\cdot\chi}\, F(\chi)}{\int \mathrm{d}^N\chi \; e^{-\frac{1}{2}\chi\cdot D^{-1}\cdot\chi}} \tag{T5.11}$$

for the Gaussian average with covariance D (the name is justified by (T5.13) below: $\langle \chi_i\chi_j\rangle_D = D_{ij}$).

Fact 1 (the generating function). For any vector J,

$$\left\langle e^{J\cdot\chi}\right\rangle_D \overset{(1)}{=} \frac{\int \mathrm{d}^N\chi \; e^{-\frac{1}{2}(\chi-DJ)\cdot D^{-1}\cdot(\chi-DJ)+\frac{1}{2}J\cdot D\cdot J}}{\int \mathrm{d}^N\chi \; e^{-\frac{1}{2}\chi\cdot D^{-1}\cdot\chi}} \overset{(2)}{=} e^{\frac{1}{2}J\cdot D\cdot J}. \tag{T5.12}$$

(1) Complete the square: $-\frac{1}{2}\chi D^{-1}\chi + J\chi = -\frac{1}{2}(\chi - DJ)D^{-1}(\chi - DJ) + \frac{1}{2}JDJ$. To check it, expand the right side: $-\frac{1}{2}\chi D^{-1}\chi + \frac{1}{2}\chi D^{-1}DJ + \frac{1}{2}JDD^{-1}\chi - \frac{1}{2}JDD^{-1}DJ + \frac{1}{2}JDJ$. With $D^{-1}D = \mathbb{1}$ and $D = D^{\mathsf{T}}$ the second and third terms add up to $J\cdot\chi$, and the last two cancel.

(2) Shift the integration variable $\chi \to \chi + DJ$; the numerator becomes $e^{\frac{1}{2}JDJ}$ times the denominator.

Expanding both sides of (T5.12) to second order in J identifies D as the covariance:

$$\left\langle \chi_i\chi_j\right\rangle_D \overset{(1)}{=} \frac{\partial^2}{\partial J_i \partial J_j}\left\langle e^{J\cdot\chi}\right\rangle_D\Big|_{J=0} \overset{(2)}{=} \frac{\partial^2}{\partial J_i \partial J_j} e^{\frac{1}{2}J\cdot D\cdot J}\Big|_{J=0} \overset{(3)}{=} D_{ij}. \tag{T5.13}$$

(1) Each $\partial/\partial J_i$ brings down one factor χ_i; then set $J = 0$.

(2) (T5.12).

(3) $\partial_{J_j} e^{\frac{1}{2}JDJ} = (DJ)_j e^{\frac{1}{2}JDJ}$, and the second derivative at $J = 0$ keeps only $\partial_{J_i}(DJ)_j = D_{ji} = D_{ij}$.

A second consequence: if χ_1 and χ_2 are independent Gaussians with covariances A and B, then $\langle e^{J\cdot(\chi_1+\chi_2)}\rangle = e^{\frac{1}{2}J\cdot A\cdot J} e^{\frac{1}{2}J\cdot B\cdot J} = e^{\frac{1}{2}J\cdot(A+B)\cdot J}$, so $\chi_1 + \chi_2$ *is Gaussian with covariance* $A + B$*: covariances add.* (A distribution is fixed by its generating function.) This is the statement "the propagator of all modes is the propagator of the slow modes plus the propagator of the fast modes".

Fact 2 (Gaussian smearing is a heat operator). For any polynomial (or analytic) function F and any fixed vector φ,

$$\left\langle F(\varphi+\chi)\right\rangle_D \overset{(1)}{=} \left\langle e^{\chi\cdot\partial_\varphi}\right\rangle_D F(\varphi) \overset{(2)}{=} \exp\!\left(\tfrac{1}{2}\,\partial_\varphi\cdot D\cdot\partial_\varphi\right)F(\varphi). \tag{T5.14}$$

(1) Taylor's theorem in N variables: $F(\varphi+\chi) = \sum_n \frac{1}{n!}(\chi\cdot\partial_\varphi)^n F(\varphi) = e^{\chi\cdot\partial_\varphi}F(\varphi)$ (in one variable this is the familiar $e^{a\,\mathrm{d}/\mathrm{d}x}f(x) = f(x+a)$).

(2) Fact 1 with $J = \partial_\varphi$. This is legitimate because the components of ∂_φ commute with each other, so every algebraic step in (T5.12) goes through term by term in the power series.

In words: averaging a function over Gaussian jitter of size D is the same as running the heat equation $\partial_s F = \frac{1}{2}\partial\cdot D\cdot\partial F$ for unit time. Heat flow smooths; so does integrating out.

Worked example T5.5: Checking the heat-operator formula on φ^4

Take $N = 1$, D a number, $F(\varphi) = \varphi^4$. Directly, using $\langle\chi\rangle = \langle\chi^3\rangle = 0$, $\langle\chi^2\rangle = D$ and $\langle\chi^4\rangle = 3D^2$ (Wick's theorem: three ways to pair four χ's),

$$\langle(\varphi+\chi)^4\rangle_D \overset{(1)}{=} \varphi^4 + 6\varphi^2\langle\chi^2\rangle + \langle\chi^4\rangle = \varphi^4 + 6D\varphi^2 + 3D^2.$$

(1) Binomial expansion; the odd powers of χ average to zero.

With the heat operator,

$$e^{\frac{D}{2}\partial^2}\varphi^4 \overset{(1)}{=} \Big(1 + \frac{D}{2}\partial^2 + \frac{D^2}{8}\partial^4\Big)\varphi^4 \overset{(2)}{=} \varphi^4 + \frac{D}{2}\cdot 12\varphi^2 + \frac{D^2}{8}\cdot 24 = \varphi^4 + 6D\varphi^2 + 3D^2.$$

(1) Exponential series; $\partial^6\varphi^4 = 0$, so it stops.
(2) $\partial^2\varphi^4 = 12\varphi^2$ and $\partial^4\varphi^4 = 24$.

The two agree. The $6D\varphi^2$ term is the "tadpole": two of the four legs of φ^4 are joined by the propagator D of the jitter.

T5.2.3 Deriving the Polchinski equation

Now the derivation. The partition function at cutoff Λ is

$$Z = \int \mathcal{D}\varphi\; e^{-\frac{1}{2}\varphi\cdot C_\Lambda^{-1}\cdot\varphi}\, e^{-S^{\mathrm{I}}_\Lambda[\varphi]}. \tag{T5.15}$$

Read the first factor as a Gaussian measure with covariance C_Λ: Z is (up to a constant) the average $\langle e^{-S^{\mathrm{I}}_\Lambda}\rangle_{C_\Lambda}$. Let $\Lambda' = \Lambda - \delta\Lambda$ and write $C_\Lambda = C_{\Lambda'} + \delta C$, with $\delta C = \Delta_\Lambda\,\delta\Lambda/\Lambda$ by (T5.9). Then

$$Z \overset{(1)}{=} \mathcal{N}\big\langle e^{-S^{\mathrm{I}}_\Lambda[\varphi]}\big\rangle_{C_\Lambda} \overset{(2)}{=} \mathcal{N}\Big\langle\big\langle e^{-S^{\mathrm{I}}_\Lambda[\varphi'+\chi]}\big\rangle_{\delta C}\Big\rangle_{C_{\Lambda'}} \overset{(3)}{=} \mathcal{N}\big\langle e^{-S^{\mathrm{I}}_{\Lambda'}[\varphi']}\big\rangle_{C_{\Lambda'}}, \tag{T5.16}$$

where we have *defined*

$$e^{-S^{\mathrm{I}}_{\Lambda'}[\varphi]} := \big\langle e^{-S^{\mathrm{I}}_\Lambda[\varphi+\chi]}\big\rangle_{\delta C} = \exp\Big(\tfrac{1}{2}\frac{\delta}{\delta\varphi}\cdot\delta C\cdot\frac{\delta}{\delta\varphi}\Big)e^{-S^{\mathrm{I}}_\Lambda[\varphi]}. \tag{T5.17}$$

(1) (T5.15) with the normalisation $\mathcal{N} = \int\mathcal{D}\varphi\, e^{-\frac{1}{2}\varphi C_\Lambda^{-1}\varphi}$ pulled out; it does not depend on the fields, so it never affects correlation functions.
(2) Fact 1: a Gaussian with covariance $C_{\Lambda'} + \delta C$ is the sum $\varphi' + \chi$ of independent Gaussians with covariances $C_{\Lambda'}$ (the slow modes) and δC (the thin shell). The average over φ becomes an average over both.
(3) Do the inner (shell) average first, and call the result $e^{-S^{\mathrm{I}}_{\Lambda'}}$. The second form in (T5.17) is Fact 2.

The last expression in (T5.16) is again of the form (T5.15), with the lower cutoff Λ' and the new interaction $S^{\mathrm{I}}_{\Lambda'}$. So (T5.17) is one exact RG step. It is Wilson's (T5.4) for a thin shell, with the kinetic term of the slow modes automatically correct.

Now expand to first order in δC. Write $S_x := \delta S^{\mathrm{I}}_\Lambda/\delta\varphi(x)$ and $S_{xy} := \delta^2 S^{\mathrm{I}}_\Lambda/\delta\varphi(x)\delta\varphi(y)$.

$$\begin{aligned} e^{-S^{\mathrm{I}}_{\Lambda'}} &\overset{(1)}{=} \Big(1 + \tfrac{1}{2}\int_{x,y}\delta C(x-y)\,\frac{\delta}{\delta\varphi(x)}\frac{\delta}{\delta\varphi(y)}\Big)e^{-S^{\mathrm{I}}_\Lambda} + O(\delta C^2) \\ &\overset{(2)}{=} e^{-S^{\mathrm{I}}_\Lambda}\Big(1 + \tfrac{1}{2}\int_{x,y}\delta C(x-y)\big[S_xS_y - S_{xy}\big]\Big) + O(\delta C^2). \end{aligned} \tag{T5.18}$$

(1) First two terms of the exponential series in (T5.17).
(2) Chain rule twice: $\frac{\delta}{\delta\varphi(y)}e^{-S} = -S_y\, e^{-S}$, and then $\frac{\delta}{\delta\varphi(x)}\big(-S_y e^{-S}\big) = (-S_{xy} + S_x S_y)\, e^{-S}$.

Now take $-\ln$ of both sides and compare with a Taylor expansion of the left side in $\delta\Lambda$:

$$\begin{aligned} S^{\mathrm{I}}_\Lambda - \delta\Lambda\,\partial_\Lambda S^{\mathrm{I}}_\Lambda + O(\delta\Lambda^2) &\overset{(1)}{=} S^{\mathrm{I}}_{\Lambda-\delta\Lambda} \overset{(2)}{=} S^{\mathrm{I}}_\Lambda - \ln\Big(1 + \tfrac{1}{2}\int_{x,y} \delta C(x-y)\big[S_xS_y - S_{xy}\big]\Big) \\ &\overset{(3)}{=} S^{\mathrm{I}}_\Lambda - \tfrac{1}{2}\int_{x,y} \delta C(x-y)\big[S_xS_y - S_{xy}\big] + O(\delta\Lambda^2) \\ &\overset{(4)}{=} S^{\mathrm{I}}_\Lambda - \frac{\delta\Lambda}{\Lambda}\,\tfrac{1}{2}\int_{x,y} \Delta_\Lambda(x-y)\big[S_xS_y - S_{xy}\big] + O(\delta\Lambda^2). \end{aligned}$$

(1) Taylor expansion of $S^{\mathrm{I}}_{\Lambda'}$ in its label Λ' about Λ.
(2) $-\ln$ of (T5.18); $-\ln e^{-S} = S$ and the log of a product is the sum of the logs.
(3) $\ln(1+\epsilon) = \epsilon + O(\epsilon^2)$, and $\epsilon \propto \delta C = O(\delta\Lambda)$.
(4) $\delta C = \Delta_\Lambda\, \delta\Lambda/\Lambda$ by (T5.9).

Cancel S^{I}_Λ, divide by $-\delta\Lambda/\Lambda$ and let $\delta\Lambda \to 0$.

Key idea: the Polchinski equation

The interaction action obeys

$$\Lambda\frac{\partial S^{\mathrm{I}}_\Lambda}{\partial\Lambda} = \frac{1}{2}\int \mathrm{d}^dx\,\mathrm{d}^dy\;\Delta_\Lambda(x-y)\left[\frac{\delta S^{\mathrm{I}}_\Lambda}{\delta\varphi(x)}\frac{\delta S^{\mathrm{I}}_\Lambda}{\delta\varphi(y)} - \frac{\delta^2 S^{\mathrm{I}}_\Lambda}{\delta\varphi(x)\,\delta\varphi(y)}\right], \tag{T5.19}$$

or, in the dot notation and with the RG time $t = \ln(\Lambda_0/\Lambda)$,

$$\partial_t S^{\mathrm{I}} = -\tfrac{1}{2}\frac{\delta S^{\mathrm{I}}}{\delta\varphi}\cdot\Delta\cdot\frac{\delta S^{\mathrm{I}}}{\delta\varphi} + \tfrac{1}{2}\frac{\delta}{\delta\varphi}\cdot\Delta\cdot\frac{\delta S^{\mathrm{I}}}{\delta\varphi}, \qquad \text{equivalently} \qquad \partial_t\, e^{-S^{\mathrm{I}}} = \tfrac{1}{2}\frac{\delta}{\delta\varphi}\cdot\Delta\cdot\frac{\delta}{\delta\varphi}\, e^{-S^{\mathrm{I}}}. \tag{T5.20}$$

The equation is exact: no expansion in a coupling was made. It says that $e^{-S^{\mathrm{I}}}$ diffuses in field space, with the shell propagator as the diffusion constant.

Both forms of (T5.20) follow from (T5.19). With $\partial_t = -\Lambda\partial_\Lambda$ (because $t = \ln(\Lambda_0/\Lambda)$),

$$\begin{aligned} \partial_t S^{\mathrm{I}} &\overset{(1)}{=} -\tfrac{1}{2}\int_{x,y}\Delta(x-y)\big[S_xS_y - S_{xy}\big] \overset{(2)}{=} -\tfrac{1}{2}\frac{\delta S^{\mathrm{I}}}{\delta\varphi}\cdot\Delta\cdot\frac{\delta S^{\mathrm{I}}}{\delta\varphi} + \tfrac{1}{2}\frac{\delta}{\delta\varphi}\cdot\Delta\cdot\frac{\delta S^{\mathrm{I}}}{\delta\varphi}, \\ \partial_t\, e^{-S^{\mathrm{I}}} &\overset{(3)}{=} -(\partial_t S^{\mathrm{I}})\, e^{-S^{\mathrm{I}}} \overset{(4)}{=} \tfrac{1}{2}\int_{x,y}\Delta(x-y)\big[S_xS_y - S_{xy}\big]e^{-S^{\mathrm{I}}} \overset{(5)}{=} \tfrac{1}{2}\frac{\delta}{\delta\varphi}\cdot\Delta\cdot\frac{\delta}{\delta\varphi}\, e^{-S^{\mathrm{I}}}. \end{aligned}$$

(1) Multiply (T5.19) by -1.
(2) The dot notation of (T5.10): $\int_{x,y}\Delta(x-y)S_xS_y$ is $\frac{\delta S}{\delta\varphi}\cdot\Delta\cdot\frac{\delta S}{\delta\varphi}$, and $\int_{x,y}\Delta(x-y)S_{xy}$ is $\frac{\delta}{\delta\varphi}\cdot\Delta\cdot\frac{\delta S}{\delta\varphi}$.
(3) Chain rule for the exponential.
(4) The first line.
(5) Step (2) of (T5.18) read backwards: $\frac{\delta}{\delta\varphi(x)}\frac{\delta}{\delta\varphi(y)}e^{-S} = (S_xS_y - S_{xy})e^{-S}$.

The exponential form shows the structure at a glance: a linear equation (heat flow) for $\psi = e^{-S^{\mathrm{I}}}$, which becomes the non-linear equation (T5.19) for $S^{\mathrm{I}} = -\ln\psi$. The same substitution turns the Schrödinger-like heat equation into the Hamilton–Jacobi-like equation of a classical action, and that is exactly what the two terms mean.

T5.2.4 Polchinski's derivation: the partition function does not move

The same equation can be reached from the opposite direction (Polchinski 1984, eqs. (16)–(19)). Instead of integrating out a shell, ask: if we lower the cutoff in the propagator, how must the interaction change so that the generating functional $Z[J]$, and hence every correlation function, stays exactly the same? Put a source J into (T5.15) that only couples to low momenta, $J(p) = 0$ wherever $\Delta_\Lambda(p) \neq 0$. To keep the bookkeeping transparent, treat the momentum modes as a discrete list φ_i (so $\varphi \cdot C^{-1} \cdot \varphi = \sum_i C_i^{-1}\varphi_i^2$, the propagator being diagonal in momentum), and write $\partial_i = \partial/\partial\varphi_i$, $S_i = \partial_i S^{\mathrm{I}}$, $S_{ii} = \partial_i^2 S^{\mathrm{I}}$ and

$$E \coloneqq \exp\Big(-\tfrac{1}{2}\sum_i C_i^{-1}\varphi_i^2 - S^{\mathrm{I}} + \sum_i J_i\varphi_i\Big), \qquad Z = \int \prod_i \mathrm{d}\varphi_i \; E.$$

Differentiating under the integral,

$$\Lambda\frac{\mathrm{d}Z}{\mathrm{d}\Lambda} \overset{(1)}{=} \int \prod_i \mathrm{d}\varphi_i \Big[-\tfrac{1}{2}\sum_i \varphi_i^2\, \Lambda\partial_\Lambda C_i^{-1} - \Lambda\partial_\Lambda S^{\mathrm{I}}\Big] E \overset{(2)}{=} \int \prod_i \mathrm{d}\varphi_i \Big[\tfrac{1}{2}\sum_i \Delta_i C_i^{-2}\varphi_i^2 - \Lambda\partial_\Lambda S^{\mathrm{I}}\Big] E. \tag{T5.21}$$

(1) Only C_Λ and S^{I}_Λ depend on Λ; J does not.
(2) $\Lambda\partial_\Lambda C^{-1} = -C^{-2}\Lambda\partial_\Lambda C = -C^{-2}\Delta$.

The integral of a total derivative $\partial_i(\cdots)$ vanishes (the integrand falls off at large φ_i). So we look for a total derivative that reproduces the bracket. Try $Y \coloneqq \sum_i \partial_i\big[(a_i\varphi_i + b_i\partial_i)E\big]$ with numbers a_i, b_i to be chosen. Using $\partial_i E = w_i E$ with $w_i = -C_i^{-1}\varphi_i - S_i$ on the shell modes (where $J_i = 0$),

$$\begin{aligned} Y &\overset{(1)}{=} \sum_i \big[a_i + a_i\varphi_i w_i + b_i(\partial_i w_i + w_i^2)\big]E \\ &\overset{(2)}{=} \sum_i \Big[a_i - b_i C_i^{-1} + (b_i C_i^{-2} - a_i C_i^{-1})\varphi_i^2 + (2b_i C_i^{-1} - a_i)\varphi_i S_i + b_i\big(S_i^2 - S_{ii}\big)\Big]E. \end{aligned}$$

(1) Product rule: $\partial_i(\varphi_i E) = E + \varphi_i w_i E$ and $\partial_i^2 E = \partial_i(w_i E) = (\partial_i w_i + w_i^2)E$.
(2) $\partial_i w_i = -C_i^{-1} - S_{ii}$ and $w_i^2 = C_i^{-2}\varphi_i^2 + 2C_i^{-1}\varphi_i S_i + S_i^2$; collect powers of φ_i.

Choose $a_i = 2b_i C_i^{-1}$ to kill the mixed term $\varphi_i S_i$, which has no counterpart in (T5.21). The φ_i^2 coefficient becomes $b_i C_i^{-2} - 2b_i C_i^{-2} = -b_i C_i^{-2}$ and the constant becomes $2b_i C_i^{-1} - b_i C_i^{-1} = b_i C_i^{-1}$ (field independent; its effect is worked out below). Matching the φ^2 term of (T5.21) requires $b_i = -\tfrac{1}{2}\Delta_i$. Then

$$Y = \Big[\tfrac{1}{2}\sum_i \Delta_i C_i^{-2}\varphi_i^2 - \tfrac{1}{2}\sum_i \Delta_i\big(S_i^2 - S_{ii}\big)\Big]E - \tfrac{1}{2}\sum_i \Delta_i C_i^{-1}E, \qquad Y = -\sum_i \Delta_i\,\partial_i\Big[\Big(C_i^{-1}\varphi_i + \tfrac{1}{2}\partial_i\Big)E\Big]. \tag{T5.22}$$

(The sums run only over shell modes, since a_i and b_i are proportional to Δ_i.) Now subtract Y from the integrand of (T5.21) and write $c \coloneqq \tfrac{1}{2}\sum_i \Delta_i C_i^{-1}$ for the field-independent constant:

$$\begin{aligned} \Lambda\frac{\mathrm{d}Z}{\mathrm{d}\Lambda} &\overset{(1)}{=} \int \prod_i \mathrm{d}\varphi_i \; Y + \int \prod_i \mathrm{d}\varphi_i \Big[\tfrac{1}{2}\sum_i \Delta_i\big(S_i^2 - S_{ii}\big) - \Lambda\partial_\Lambda S^{\mathrm{I}} + c\Big]E \\ &\overset{(2)}{=} \int \prod_i \mathrm{d}\varphi_i \Big[\tfrac{1}{2}\sum_i \Delta_i\big(S_i^2 - S_{ii}\big) - \Lambda\partial_\Lambda S^{\mathrm{I}}\Big]E + c\,Z. \end{aligned}$$

(1) Add and subtract Y, using the first form of (T5.22) with the constant written out: the φ_i^2 terms cancel.

(2) The integral of the total derivative Y vanishes (second form of (T5.22)); c does not depend on the fields, so $\int cE = cZ$.

The bracket vanishes exactly when $\Lambda\partial_\Lambda S^{\mathrm{I}} = \frac{1}{2}\sum_i \Delta_i(S_i^2 - S_{ii})$, which is (T5.19) written for discrete modes. So if S^{I} obeys the Polchinski equation, $\Lambda\,\mathrm{d}Z/\mathrm{d}\Lambda = cZ$: Z changes only by a field-independent and J-independent factor, which cancels in every normalised correlation function $\langle\varphi\cdots\varphi\rangle = Z^{-1}\partial_J\cdots\partial_J Z$. All n-point functions at low momenta are cutoff independent.

Watch out: why the source must avoid the shell

The step "$J_i = 0$ on the shell" is not a technicality. If J reached into the modes being removed, the correlators of those modes would change, since they are no longer in the theory. In practice one fixes a low scale Λ_R [the lowest cutoff we will ever use, where measurements are made], takes $J(p) = 0$ for $p^2 > \Lambda_R^2$, and lowers the cutoff only down to Λ_R. With a cutoff function that is exactly 1 for $p^2 < \Lambda^2$, as in Section T5.2.7, $\Delta_\Lambda(p)$ vanishes for $p^2 < \Lambda^2$, so the shell never reaches the source. Then we only ask questions at momenta below Λ_R.

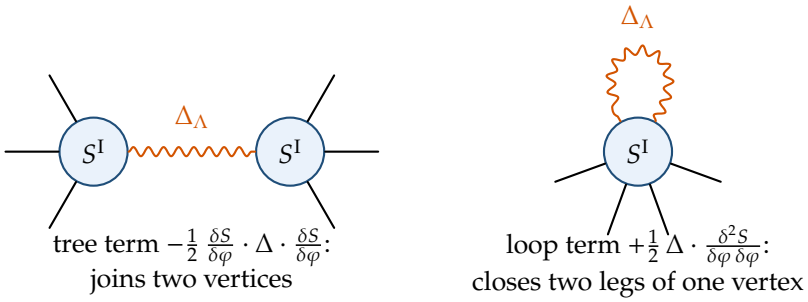


Figure T5.7: The two terms of the Polchinski equation as Feynman diagrams. A blob is a vertex of S^{I}_Λ; black lines are external legs (slow fields φ); the wavy orange line is the shell propagator Δ_Λ of the modes being integrated out. Left: two vertices are joined by one shell line, which generates a new vertex with the legs of both, as in Example T5.3. Right: two legs of one vertex are joined, which lowers the number of legs by two, as the tadpole $6D\varphi^2$ of Example T5.5. What to see: iterating these two moves builds every Feynman diagram, with internal lines made of the modes already integrated out.

Watch out: what is and is not approximated

The Polchinski equation is exact, but to use it we must truncate the infinite set of couplings in S^{I}. The truncation is the approximation, not the equation. Two standard truncations are perturbation theory (expand in the couplings, Example T5.6 below and Section T5.3) and the derivative expansion, whose lowest order is the local potential approximation of Section T5.2.5.

Since (T5.20) is a heat equation, its solution is the heat kernel: integrating from the bare scale Λ_0 down to Λ,

$$e^{-S^{\mathrm{I}}_\Lambda[\varphi]} = \exp\Big(\tfrac{1}{2}\frac{\delta}{\delta\varphi}\cdot(C_{\Lambda_0} - C_\Lambda)\cdot\frac{\delta}{\delta\varphi}\Big)e^{-S^{\mathrm{I}}_{\Lambda_0}[\varphi]} = \big\langle e^{-S^{\mathrm{I}}_{\Lambda_0}[\varphi+\chi]}\big\rangle_{C_{\Lambda_0}-C_\Lambda}. \tag{T5.23}$$

Why is this the solution? Write $\mathsf{H}_t := \frac{1}{2}\frac{\delta}{\delta\varphi}\cdot\Delta_{\Lambda(t)}\cdot\frac{\delta}{\delta\varphi}$, so that (T5.20) reads $\partial_t\psi = \mathsf{H}_t\psi$ for $\psi = e^{-S^{\mathrm{I}}}$. The operators H_t at different times commute (each is a second derivative in φ with a field-independent kernel, and partial derivatives commute), so they can be integrated like

numbers:

$$\psi(t) \overset{(1)}{=} \exp\Big(\int_0^t \mathrm{d}s\ \mathsf{H}_s\Big)\psi(0), \qquad \int_0^t \mathrm{d}s\ \Delta_{\Lambda(s)} \overset{(2)}{=} \int_\Lambda^{\Lambda_0} \frac{\mathrm{d}\Lambda'}{\Lambda'}\,\Lambda' \frac{\partial C_{\Lambda'}}{\partial \Lambda'} \overset{(3)}{=} C_{\Lambda_0} - C_\Lambda .$$

(1) Differentiate in t: since the H_s commute, the derivative of the exponential is H_t times the exponential, as for $e^{\int_0^t a(s)\mathrm{d}s}$ with a number $a(s)$.
(2) $s = \ln(\Lambda_0/\Lambda')$, so $\mathrm{d}s = -\mathrm{d}\Lambda'/\Lambda'$, and s from 0 to t means Λ' from Λ_0 down to Λ; the minus sign flips the limits. Then (T5.8).
(3) The integral of a derivative.

This is the first form of (T5.23): the "diffusion time" is the propagator $C_{\Lambda_0} - C_\Lambda$. The second form is Fact 2 (T5.14) with $D = C_{\Lambda_0} - C_\Lambda$. So *the Wilsonian interaction at scale Λ is the sum of all connected Feynman diagrams built from the bare vertices, with internal lines $C_{\Lambda_0} - C_\Lambda$* (the propagator of the modes between Λ and Λ_0) and with the slow field φ on the external legs. The minus sign and the logarithm pick out the connected diagrams, as always. Here is the simplest case, solved exactly.

Worked example T5.6: A mass term as an exact solution

Try $S^{\mathrm{I}}_\Lambda = \frac{1}{2}\varphi\cdot M_\Lambda\cdot\varphi + E_\Lambda$, a quadratic interaction with kernel $M_\Lambda(p)$ and a field-independent constant E_Λ. In the dot notation its derivatives are $\frac{\delta S^{\mathrm{I}}}{\delta\varphi} = M_\Lambda\cdot\varphi$ and $\frac{\delta^2 S^{\mathrm{I}}}{\delta\varphi\,\delta\varphi} = M_\Lambda$ (the kernel itself). Insert them into the first form of (T5.20):

$$\begin{aligned}\tfrac{1}{2}\,\varphi\cdot\partial_t M_\Lambda\cdot\varphi + \partial_t E_\Lambda &\overset{(1)}{=} -\tfrac{1}{2}\,(M_\Lambda\cdot\varphi)\cdot\Delta_\Lambda\cdot(M_\Lambda\cdot\varphi) + \tfrac{1}{2}\,\mathrm{Tr}\big(\Delta_\Lambda M_\Lambda\big)\\ &\overset{(2)}{=} -\tfrac{1}{2}\,\varphi\cdot M_\Lambda\Delta_\Lambda M_\Lambda\cdot\varphi + \tfrac{1}{2}\,\mathrm{Tr}\big(\Delta_\Lambda M_\Lambda\big).\end{aligned}$$

(1) Left: only M and E depend on t. Right: the tree term, and the loop term $\frac{1}{2}\int_{x,y}\Delta(x-y)M(x-y)$, which is a number (the trace of the product of kernels).
(2) In momentum space all kernels are diagonal (functions of p), so $M\cdot\varphi$ is $M(p)\varphi(p)$ and the product of kernels is the product of functions, $M(p)\Delta(p)M(p)$.

The φ^2 terms give $\partial_t M = -M\Delta M$, and the field-independent terms only fix E_Λ. Since $\partial_t = -\Lambda\partial_\Lambda$,

$$\Lambda\partial_\Lambda M_\Lambda = M_\Lambda\,\Delta_\Lambda\,M_\Lambda \quad\Longrightarrow\quad \Lambda\partial_\Lambda\Big(\frac{1}{M_\Lambda}\Big) \overset{(1)}{=} -\frac{\Lambda\partial_\Lambda M_\Lambda}{M_\Lambda^2} \overset{(2)}{=} -\Delta_\Lambda \overset{(3)}{=} -\Lambda\partial_\Lambda C_\Lambda .$$

(1) Derivative of $1/M$.
(2) The flow equation just derived.
(3) Definition (T5.8).

So $1/M_\Lambda + C_\Lambda$ is independent of Λ. Start at Λ_0 with a plain mass, $M_{\Lambda_0} = m^2$, and write $D = C_{\Lambda_0} - C_\Lambda$:

$$M_\Lambda(p) \overset{(1)}{=} \frac{m^2}{1 + m^2 D(p)} \overset{(2)}{=} m^2 - m^2 D m^2 + m^2 D m^2 D m^2 - \cdots$$

(1) $1/M_\Lambda = 1/m^2 + C_{\Lambda_0} - C_\Lambda$.
(2) Geometric series.

This is the diagrammatic statement: a chain of mass insertions (crosses) joined by propagators D of the modes already integrated out, with the sign (-1) per extra insertion from e^{-S}. It is the tree term of Figure T5.7 iterated.

As a check, let $\Lambda \to 0$ with Λ_0 finite, so that all modes are integrated out ($C_\Lambda = K(\infty)/p^2 = 0$). Write $C \coloneqq C_{\Lambda_0} = K_0/p^2$ with $K_0 \coloneqq K(p^2/\Lambda_0^2)$. Then

$$M_{\Lambda\to 0} \overset{(1)}{=} \frac{m^2}{1+m^2 C} \overset{(2)}{=} \frac{m^2 p^2}{p^2 + m^2 K_0}.$$

(1) $D = C_{\Lambda_0} - 0 = C$.
(2) Multiply top and bottom by p^2 and use $p^2 C = K_0$.

What should it be? Put $\Lambda = 0$ in the second form of (T5.23) and change variables to $\chi' = \varphi + \chi$, the full field:

$$e^{-S_0^{\mathrm{I}}[\varphi]} \overset{(1)}{=} \frac{\int \mathcal{D}\chi'\, e^{-\frac{1}{2}(\chi'-\varphi)\cdot C^{-1}\cdot(\chi'-\varphi)} e^{-S^{\mathrm{I}}_{\Lambda_0}[\chi']}}{\int \mathcal{D}\chi\, e^{-\frac{1}{2}\chi\cdot C^{-1}\cdot\chi}} \overset{(2)}{=} e^{-\frac{1}{2}\varphi\cdot C^{-1}\cdot\varphi}\, \frac{Z[J]}{Z_{\mathrm{free}}}, \qquad J \coloneqq C^{-1}\cdot\varphi,$$

(1) The definition (T5.11) of the average; the shift does not change the measure.
(2) Expand the square: $-\frac{1}{2}\chi' C^{-1}\chi' + J\cdot\chi' - \frac{1}{2}\varphi C^{-1}\varphi$. What remains is the full partition function of the bare theory with source J, $Z[J] = \int \mathcal{D}\chi'\, e^{-\frac{1}{2}\chi' C^{-1}\chi' - S^{\mathrm{I}}_{\Lambda_0}[\chi'] + J\cdot\chi'}$, divided by its free value at $J = 0$.

So $-S_0^{\mathrm{I}}[\varphi] = W[J] - \frac{1}{2}J\cdot C\cdot J$ with $W = \ln(Z/Z_{\mathrm{free}})$ the connected generating functional (we used $\varphi\cdot C^{-1}\cdot\varphi = J\cdot C\cdot J$). For our bare theory, a free field with propagator $G \coloneqq (C^{-1}+m^2)^{-1} = K_0/(p^2+m^2K_0)$, the Gaussian integral gives $W = \frac{1}{2}J\cdot G\cdot J$. Compare the two sides, working at each momentum p:

$$W - \tfrac{1}{2}J\cdot C\cdot J \overset{(1)}{=} \tfrac{1}{2}J\cdot(G-C)\cdot J \overset{(2)}{=} -\tfrac{1}{2}J\cdot\frac{m^2C^2}{1+m^2C}\cdot J \overset{(3)}{=} -\tfrac{1}{2}\varphi\cdot\frac{m^2}{1+m^2C}\cdot\varphi \overset{(4)}{=} -S_0^{\mathrm{I}}[\varphi].$$

(1) Combine the two quadratic forms.
(2) $G = C/(1+m^2C)$, so $G - C = \big(C - C - m^2C^2\big)/(1+m^2C)$.
(3) $J = C^{-1}\varphi$ cancels the two factors of C.
(4) $S_0^{\mathrm{I}} = \frac{1}{2}\varphi\cdot M_{\Lambda\to 0}\cdot\varphi$ with $M_{\Lambda\to 0}$ found above (up to the constant E).

The flow reproduces the full massive propagator $K_0/(p^2+m^2K_0)$.

T5.2.5 The local potential approximation

The most useful truncation keeps an arbitrary function of the field but throws away all momentum dependence of the vertices:

$$S_\Lambda^{\mathrm{I}}[\varphi] \approx \int \mathrm{d}^d x\; V_\Lambda\big(\varphi(x)\big). \tag{T5.24}$$

This is the **local potential approximation** (LPA). It keeps every φ^n coupling at once (all the couplings in the Taylor series of V_Λ) and drops the derivative couplings such as $\varphi^2(\partial\varphi)^2$. It is the lowest order of an expansion in derivatives, and it is good when the anomalous dimension is small, since the anomalous dimension lives in the $(\partial\varphi)^2$ term.

To get the flow of V_Λ, insert (T5.24) into (T5.19) and evaluate both sides on a *constant* field $\varphi(x) = \varphi$ in a large volume $\mathcal{V}$. (On a non-constant field the tree term generates momentum-dependent vertices, so the ansatz is not exactly preserved. The LPA keeps only what a constant

field sees.) The ingredients are

$$\frac{\delta S^{\mathrm{I}}}{\delta\varphi(x)} \overset{(1)}{=} V'(\varphi), \qquad \frac{\delta^2 S^{\mathrm{I}}}{\delta\varphi(x)\delta\varphi(y)} \overset{(2)}{=} V''(\varphi)\,\delta^d(x-y),$$

$$\int_{x,y} \Delta(x-y)\,V'V' \overset{(3)}{=} \mathcal{V}\,\Delta_\Lambda(p{=}0)\,V'^2, \qquad \int_{x,y} \Delta(x-y)\,V''\delta^d(x-y) \overset{(4)}{=} \mathcal{V}\,V''\int \frac{\mathrm{d}^d p}{(2\pi)^d}\Delta_\Lambda(p).$$

(1) Changing φ at the single point x changes $\int \mathrm{d}^d z\, V(\varphi(z))$ only through the integrand at $z = x$, by $V'(\varphi(x))$.
(2) Differentiate $V'(\varphi(x))$ once more; it depends on $\varphi(y)$ only if $y = x$, which gives $\delta^d(x-y)$.
(3) V' is constant, $\int \mathrm{d}^d x = \mathcal{V}$, and $\int \mathrm{d}^d y\,\Delta(x-y) = \int \mathrm{d}^d z\,\Delta(z)e^{-i0\cdot z}$ is the Fourier transform at zero momentum.
(4) The delta function sets $y = x$, leaving $\mathcal{V}\,V''\,\Delta(z{=}0)$, and $\Delta(z{=}0) = \int \frac{\mathrm{d}^d p}{(2\pi)^d} e^{ip\cdot 0}\Delta(p)$ is the integral over all momenta.

The left side of (T5.19) is $\Lambda\partial_\Lambda \int \mathrm{d}^d x\, V = \mathcal{V}\,\Lambda\partial_\Lambda V$. Insert the four ingredients and divide by $\mathcal{V}$:

$$\Lambda\frac{\partial V_\Lambda}{\partial\Lambda} = \tfrac{1}{2}\Delta_\Lambda(0)\,V'^2 - \tfrac{1}{2}V''\int \frac{\mathrm{d}^d p}{(2\pi)^d}\,\Delta_\Lambda(p). \tag{T5.25}$$

For the exponential cutoff both numbers are elementary. Setting $p = 0$ in $\Delta_\Lambda(p) = 2e^{-p^2/\Lambda^2}/\Lambda^2$ gives $\Delta_\Lambda(0) = 2/\Lambda^2$, and

$$\int \frac{\mathrm{d}^d p}{(2\pi)^d}\,\frac{2}{\Lambda^2}e^{-p^2/\Lambda^2} \overset{(1)}{=} \frac{2}{\Lambda^2}\,\frac{(\pi\Lambda^2)^{d/2}}{(2\pi)^d} \overset{(2)}{=} \frac{2\,\Lambda^{d-2}}{(4\pi)^{d/2}}.$$

(1) The Gaussian integral $\int \mathrm{d}^d p\, e^{-p^2/\Lambda^2} = (\pi\Lambda^2)^{d/2}$: since $p^2 = p_1^2 + \cdots + p_d^2$, it is the product of d one-dimensional integrals $\int \mathrm{d}p_1\, e^{-p_1^2/\Lambda^2} = \sqrt{\pi}\,\Lambda$.
(2) $(2\pi)^d = (4\pi^2)^{d/2}$, so $\pi^{d/2}/(4\pi^2)^{d/2} = (4\pi)^{-d/2}$.

So, with $\alpha_d := (4\pi)^{-d/2}$, the two terms of (T5.25) are $\frac{1}{2}\cdot\frac{2}{\Lambda^2}V'^2$ and $\frac{1}{2}V''\cdot 2\alpha_d\Lambda^{d-2}$:

$$\Lambda\frac{\partial V_\Lambda}{\partial\Lambda} = \frac{V'^2}{\Lambda^2} - \alpha_d\,\Lambda^{d-2}\,V''. \tag{T5.26}$$

Dimensionless variables. Fixed points need dimensionless quantities. Measure the field and the potential in units of the cutoff,

$$\varphi = \Lambda^{\frac{d-2}{2}}\,\hat{\varphi}, \qquad V_\Lambda(\varphi) = \Lambda^d\,v_t(\hat{\varphi}), \qquad t = \ln\frac{\Lambda_0}{\Lambda}. \tag{T5.27}$$

Primes on v mean $\mathrm{d}/\mathrm{d}\hat{\varphi}$. Since $\hat{\varphi} = \Lambda^{-\frac{d-2}{2}}\varphi$, each $\mathrm{d}/\mathrm{d}\varphi$ acting on $V = \Lambda^d v(\hat{\varphi})$ brings a factor $\Lambda^{-\frac{d-2}{2}}$:

$$V' = \Lambda^{d-\frac{d-2}{2}}v' = \Lambda^{\frac{d+2}{2}}v', \qquad V'' = \Lambda^{d-(d-2)}v'' = \Lambda^2 v''.$$

Differentiating at fixed $\hat{\varphi}$,

$$\begin{aligned}
\partial_t v &\overset{(1)}{=} -\Lambda\partial_\Lambda\big(\Lambda^{-d}V_\Lambda(\Lambda^{\frac{d-2}{2}}\hat{\varphi})\big) \\
&\overset{(2)}{=} d\,v - \Lambda^{-d}\Big[\Lambda\partial_\Lambda V\big|_\varphi + V'\cdot\tfrac{d-2}{2}\Lambda^{\frac{d-2}{2}}\hat{\varphi}\Big] \\
&\overset{(3)}{=} d\,v - \tfrac{d-2}{2}\,\hat{\varphi}\,v' - v'^2 + \alpha_d\,v''.
\end{aligned} \tag{T5.28}$$

(1) $\partial_t = -\Lambda\partial_\Lambda$ and (T5.27).
(2) Product rule: the Λ^{-d} gives $d\,v$; the Λ inside the argument gives the chain-rule term.
(3) Insert (T5.26): $\Lambda^{-d}V'^2/\Lambda^2 = v'^2$ and $\Lambda^{-d}\alpha_d\Lambda^{d-2}V'' = \alpha_d v''$; and $\Lambda^{-d}V'\Lambda^{(d-2)/2}\hat\varphi = \hat\varphi v'$.

The first two terms are pure dimensional analysis (they are there even with no interactions, as in (T5.6)); the last two are the tree and loop terms.

Worked example T5.7: The LPA flow of the mass, the quartic and the sextic coupling

Expand the dimensionless potential of a $\mathbb{Z}_2$-symmetric theory,

$$v = \frac{r}{2}\hat\varphi^2 + \frac{u}{4!}\hat\varphi^4 + \frac{w}{6!}\hat\varphi^6 + \cdots,$$

so that $r = m^2/\Lambda^2$ is the dimensionless mass, u the quartic and w the sextic coupling. We need

$$v' = r\hat\varphi + \tfrac{u}{6}\hat\varphi^3 + \tfrac{w}{120}\hat\varphi^5 + \cdots, \qquad v'' = r + \tfrac{u}{2}\hat\varphi^2 + \tfrac{w}{24}\hat\varphi^4 + \cdots,$$
$$\hat\varphi v' = r\hat\varphi^2 + \tfrac{u}{6}\hat\varphi^4 + \tfrac{w}{120}\hat\varphi^6 + \cdots, \qquad v'^2 = r^2\hat\varphi^2 + \tfrac{ru}{3}\hat\varphi^4 + \left(\tfrac{u^2}{36} + \tfrac{rw}{60}\right)\hat\varphi^6 + \cdots.$$

The first three lines differentiate the series term by term. For v'^2, square v' and collect: the $\hat\varphi^4$ term is $2\cdot r\cdot\frac{u}{6} = \frac{ru}{3}$, and the $\hat\varphi^6$ term is $(\frac{u}{6})^2 + 2\cdot r\cdot\frac{w}{120} = \frac{u^2}{36} + \frac{rw}{60}$. Now match powers of $\hat\varphi$ in (T5.28). On the left, $\partial_t v = \frac{\partial_t r}{2}\hat\varphi^2 + \frac{\partial_t u}{4!}\hat\varphi^4 + \frac{\partial_t w}{6!}\hat\varphi^6 + \cdots$, so we multiply the coefficient of $\hat\varphi^2$ by 2, of $\hat\varphi^4$ by 4! and of $\hat\varphi^6$ by 6!. On the right each coefficient collects four pieces: from $d\,v$, from $-\frac{d-2}{2}\hat\varphi v'$, from $-v'^2$ and from $\alpha_d v''$:

$$\begin{aligned}
\partial_t r &\overset{(1)}{=} d\,r - (d-2)r - 2r^2 + \alpha_d u = 2r - 2r^2 + \alpha_d u,\\
\partial_t u &\overset{(2)}{=} d\,u - 2(d-2)u - 8ru + \alpha_d w = (4-d)u - 8ru + \alpha_d w,\\
\partial_t w &\overset{(3)}{=} d\,w - 3(d-2)w - 20u^2 - 12rw + \alpha_d w_8 = (6-2d)w - 20u^2 - 12rw + \alpha_d w_8.
\end{aligned} \tag{T5.29}$$

(1) $2\cdot\left[\frac{d}{2}r - \frac{d-2}{2}r - r^2 + \alpha_d\frac{u}{2}\right]$.
(2) $24\cdot\left[\frac{d}{24}u - \frac{d-2}{2}\frac{u}{6} - \frac{ru}{3} + \alpha_d\frac{w}{24}\right]$.
(3) $720\cdot\left[\frac{d}{720}w - \frac{d-2}{2}\frac{w}{120} - \frac{u^2}{36} - \frac{rw}{60} + \alpha_d\frac{w_8}{720}\right]$, where w_8 is the $\hat\varphi^8$ coupling ($v \supset w_8\hat\varphi^8/8!$, so $v'' \supset w_8\hat\varphi^6/6!$).

The linear terms $2, 4-d, 6-2d$ are the Gaussian eigenvalues of Example T5.4, as they must be. The constant α_d can be absorbed into the couplings. Define $\tilde u := \alpha_d u$, $\tilde w := \alpha_d^2 w$ and $\tilde w_8 := \alpha_d^3 w_8$. Multiply the second line by α_d and the third by α_d^2:

$$\partial_t r = 2r - 2r^2 + \tilde u, \qquad \partial_t\tilde u \overset{(1)}{=} (4-d)\tilde u - 8r\tilde u + \tilde w,$$
$$\partial_t\tilde w \overset{(2)}{=} (6-2d)\tilde w - 20\tilde u^2 - 12r\tilde w + \tilde w_8.$$

(1) $\alpha_d\cdot\alpha_d w = \alpha_d^2 w = \tilde w$.
(2) $\alpha_d^2 u^2 = \tilde u^2$ and $\alpha_d^2\cdot\alpha_d w_8 = \tilde w_8$.

This is (T5.29) with $\alpha_d = 1$. We use these variables from now on and drop the tildes.

Three features of (T5.29) deserve comment.

- The loop term (α_d) feeds each coupling from the next higher one: u into r, w into u. The tree term (v'^2) feeds from lower ones: u^2 into w, as in Example T5.3. So the tower does

not close; in practice one truncates it.

- In $d = 4 - \varepsilon$ the sextic coupling has eigenvalue $6 - 2d = -2 + 2\varepsilon$: strongly irrelevant. It relaxes quickly to the value where $\partial_t w \approx 0$. Dropping $12rw$ and w_8, which are of higher order in the small couplings,

$$0 \overset{(1)}{=} (-2 + 2\varepsilon)w - 20u^2 \implies w \overset{(2)}{=} -\frac{20u^2}{2(1-\varepsilon)} \overset{(3)}{=} -10u^2\big(1 + O(\varepsilon)\big).$$

  (1) The third line of (T5.29) with $d = 4 - \varepsilon$ and $\alpha_d = 1$.
  (2) Solve for w.
  (3) $1/(1-\varepsilon) = 1 + \varepsilon + \cdots$.

  Section T5.2.6 shows why the flow really lands there. Substituting $w = -10u^2$ into the second line, where $(4-d)u = \varepsilon u$, gives a closed system for (r, u):

$$\partial_t r = 2r - 2r^2 + u, \qquad \partial_t u = \varepsilon u - 8ru - 10u^2. \tag{T5.30}$$

- Fixed points of (T5.30): the **Gaussian** point $r = u = 0$, and, for $\varepsilon > 0$, a second one, the **Wilson–Fisher** point, computed in the next example.

Worked example T5.8: The Wilson–Fisher fixed point in the LPA

Set both right sides of (T5.30) to zero with $u \neq 0$ and keep the leading order in ε (we will see $r^*, u^* = O(\varepsilon)$, so r^2 and ru corrections are $O(\varepsilon^2)$):

$$0 = 2r^* - 2r^{*2} + u^* \implies r^* \overset{(1)}{=} -\tfrac{1}{2}u^* + O(\varepsilon^2),$$
$$0 = \varepsilon - 8r^* - 10u^* \implies \varepsilon + 4u^* - 10u^* \overset{(2)}{=} 0 \implies u^* = \frac{\varepsilon}{6}, \quad r^* = -\frac{\varepsilon}{12}.$$

(1) Drop $r^{*2} = O(\varepsilon^2)$.
(2) Divide the second equation by $u^* \neq 0$ and insert r^*.

The **stability matrix** $M_{ij} = \partial(\partial_t g_i)/\partial g_j$ at the fixed point, with $g = (r, u)$, is

$$M \overset{(1)}{=} \begin{pmatrix} 2 - 4r^* & 1 \\ -8u^* & \varepsilon - 8r^* - 20u^* \end{pmatrix} \overset{(2)}{=} \begin{pmatrix} 2 + \frac{\varepsilon}{3} & 1 \\ -\frac{4\varepsilon}{3} & -\frac{5\varepsilon}{3} \end{pmatrix} + O(\varepsilon^2).$$

(1) Differentiate (T5.30).
(2) Insert $r^* = -\varepsilon/12$, $u^* = \varepsilon/6$: $\varepsilon + \frac{2\varepsilon}{3} - \frac{10\varepsilon}{3} = -\frac{5\varepsilon}{3}$.

Its eigenvalues y solve the characteristic equation $\det(M - y\mathbb{1}) = y^2 - (\operatorname{tr} M)\, y + \det M = 0$. First the two coefficients:

$$\operatorname{tr} M \overset{(1)}{=} 2 + \tfrac{\varepsilon}{3} - \tfrac{5\varepsilon}{3} = 2 - \tfrac{4\varepsilon}{3},$$
$$\det M \overset{(2)}{=} \big(2 + \tfrac{\varepsilon}{3}\big)\big(-\tfrac{5\varepsilon}{3}\big) - 1 \cdot \big(-\tfrac{4\varepsilon}{3}\big) = -\tfrac{10\varepsilon}{3} + \tfrac{4\varepsilon}{3} + O(\varepsilon^2) = -2\varepsilon + O(\varepsilon^2).$$

(1) Sum of the diagonal entries.
(2) $\det \begin{pmatrix} a & b \\ c & e \end{pmatrix} = ae - bc$; the product $\frac{\varepsilon}{3} \cdot \frac{5\varepsilon}{3}$ is $O(\varepsilon^2)$.

Then the roots, using that $\det M$ is small compared with $(\operatorname{tr} M)^2 \approx 4$:

$$y_{1,2} \overset{(1)}{=} \tfrac{1}{2}\Big(\operatorname{tr} M \pm \sqrt{(\operatorname{tr} M)^2 - 4\det M}\Big) \overset{(2)}{=} \tfrac{1}{2}\operatorname{tr} M\Big(1 \pm \Big(1 - \frac{2\det M}{(\operatorname{tr} M)^2}\Big)\Big) + O(\varepsilon^2),$$
$$y_1 \overset{(3)}{=} \operatorname{tr} M - \frac{\det M}{\operatorname{tr} M} = 2 - \tfrac{4\varepsilon}{3} + \varepsilon + O(\varepsilon^2), \qquad y_2 \overset{(4)}{=} \frac{\det M}{\operatorname{tr} M} = \frac{-2\varepsilon}{2} + O(\varepsilon^2).$$

(1) Quadratic formula.
(2) Take $\operatorname{tr} M$ out of the root and use $\sqrt{1-x} = 1 - \frac{x}{2} + O(x^2)$ with $x = 4 \det M/(\operatorname{tr} M)^2 = O(\varepsilon)$.
(3) Upper sign; $\det M/\operatorname{tr} M = -2\varepsilon/2 + O(\varepsilon^2) = -\varepsilon + O(\varepsilon^2)$.
(4) Lower sign; only $\operatorname{tr} M = 2$ is needed at this order.

So

$$y_1 = 2 - \frac{\varepsilon}{3} + O(\varepsilon^2), \qquad y_2 = -\varepsilon + O(\varepsilon^2). \tag{T5.31}$$

One relevant direction (the temperature) and one irrelevant one. By the relation $\nu = 1/y_t$ derived in Section T5.3.1, $\nu = 1/y_1 = \frac{1}{2}(1 - \frac{\varepsilon}{6})^{-1} = \frac{1}{2} + \frac{\varepsilon}{12} + O(\varepsilon^2)$. (A numerical solution of the untruncated three-coupling system (T5.29) at $\varepsilon = 10^{-3}$ gives $y_1 = 1.99967$ and $y_2 = -0.00100$, as predicted.)

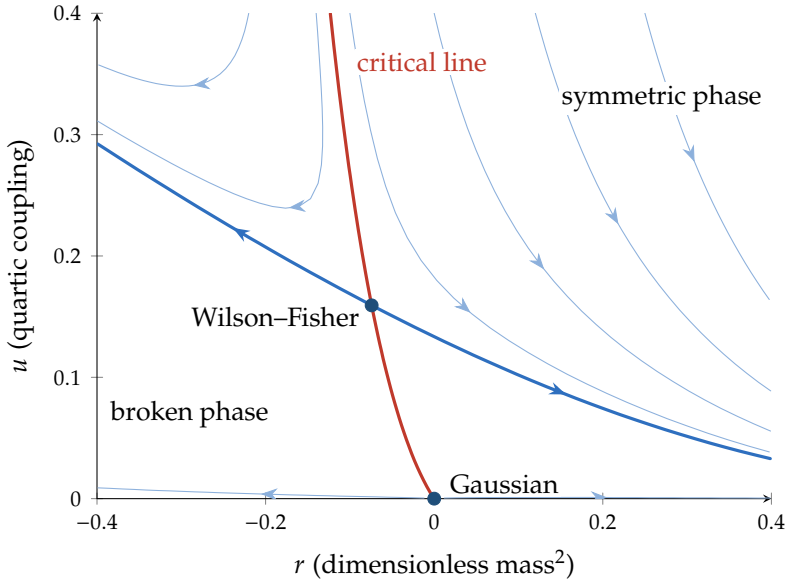


Figure T5.8: The LPA flow (T5.30) in the (r, u) plane for $\varepsilon = 1$ ($d = 3$), computed numerically; arrows point towards the infrared. Red: the critical line, the set of theories that flow into the Wilson–Fisher point (from the Gaussian point below and from large u above). Thick blue: the relevant direction leaving the Wilson–Fisher point. Theories to the right of the red line flow to large positive r (a massive, disordered theory); theories to the left flow to negative r (the ordered phase). What to see: the geometry of Figure T5.5 in an honest calculation. (For $\varepsilon = 1$ the $O(\varepsilon)$ formulas are only rough: the numerical fixed point is at $(r^*, u^*) = (-0.074, 0.159)$ with eigenvalues 1.94 and −1.23.)

Watch out: fixed-point coordinates are not physical; eigenvalues are

The location $u^* = \varepsilon/6$ depends on the cutoff function K and on the truncation. In Section T5.3 a different scheme (a sharp momentum shell, keeping only one-loop diagrams) gives $u^* = \varepsilon/3$ in the same units, twice the LPA value, and dimensional regularisation gives yet another number. What does *not* depend on the scheme are the eigenvalues of the stability matrix: a change of scheme is a smooth change of coordinates $g \mapsto \tilde{g}(g)$ on theory space, with Jacobian $J_{ij} = \partial \tilde{g}_i/\partial g_j$. Near the fixed point write $\delta\tilde{g} = J\,\delta g + O(\delta g^2)$. Then

$$\partial_t\, \delta\tilde{g} \overset{(1)}{=} J\, \partial_t\, \delta g \overset{(2)}{=} JM\, \delta g \overset{(3)}{=} JMJ^{-1}\, \delta\tilde{g},$$

(1) J is evaluated at the fixed point, so it is constant along the linearised flow.

(2) Linearised flow $\partial_t \delta g = M \delta g$.
(3) $\delta g = J^{-1} \delta \tilde{g}$.

so $\tilde{M} = JMJ^{-1}$. If $Me = ye$ then $\tilde{M}(Je) = JMe = y\,(Je)$: the eigenvalues are unchanged. This is why critical exponents are universal while coupling constants at fixed points are not.

T5.2.6 Polchinski's toy flow: one relevant and one irrelevant coupling

We can now make the "squeezing" of Figure T5.5 quantitative. The mechanism is clearest in a model with just two couplings (Polchinski 1984, §2), chosen to mimic the φ^4 and φ^6 couplings of a four-dimensional scalar: g_4, dimensionless, and g_6, of mass dimension -2. Their flow is

$$\Lambda \frac{\mathrm{d}g_4}{\mathrm{d}\Lambda} = \beta_4\big(g_4, \Lambda^2 g_6\big), \qquad \Lambda \frac{\mathrm{d}g_6}{\mathrm{d}\Lambda} = \Lambda^{-2}\beta_6\big(g_4, \Lambda^2 g_6\big),$$

with β_4, β_6 dimensionless functions. The powers of Λ are fixed by dimensional analysis: the only scale is Λ, so g_6 can enter only through the dimensionless $\Lambda^2 g_6$, and the left side of the second equation has dimension -2, which the factor Λ^{-2} supplies. (In this subsection we use $\Lambda\,\mathrm{d}/\mathrm{d}\Lambda$, which points towards the ultraviolet, the opposite of $\partial_t = -\Lambda\,\mathrm{d}/\mathrm{d}\Lambda$.) In the dimensionless variables $\lambda_4 = g_4$, $\lambda_6 = \Lambda^2 g_6$,

$$\Lambda \frac{\mathrm{d}\lambda_6}{\mathrm{d}\Lambda} \overset{(1)}{=} 2\Lambda^2 g_6 + \Lambda^2 \Lambda \frac{\mathrm{d}g_6}{\mathrm{d}\Lambda} \overset{(2)}{=} 2\lambda_6 + \beta_6(\lambda_4, \lambda_6) \quad \Longrightarrow \quad \Lambda \frac{\mathrm{d}\lambda_4}{\mathrm{d}\Lambda} = \beta_4, \qquad \Lambda \frac{\mathrm{d}\lambda_6}{\mathrm{d}\Lambda} - 2\lambda_6 = \beta_6. \tag{T5.32}$$

(1) Product rule on $\Lambda^2 g_6$.
(2) The flow of g_6: $\Lambda^2 \cdot \Lambda^{-2}\beta_6 = \beta_6$.

The $-2\lambda_6$ says that λ_6 is irrelevant: with β_6 switched off, $\lambda_6 \propto \Lambda^2$, so it dies like $(\Lambda/\Lambda_0)^2$ on the way from Λ_0 down to Λ.

Two neighbouring trajectories. Take one solution $\bar{\lambda}_i(\Lambda)$ and a nearby one, $\lambda_i = \bar{\lambda}_i + \epsilon_i$. Taylor expand the right sides of (T5.32) about $\bar{\lambda}$ and subtract the equations for $\bar{\lambda}$; to first order in ϵ,

$$\Lambda \frac{\mathrm{d}\epsilon}{\mathrm{d}\Lambda} = J\epsilon, \qquad J = \begin{pmatrix} \bar{\partial}_4\beta_4 & \bar{\partial}_6\beta_4 \\ \bar{\partial}_4\beta_6 & 2 + \bar{\partial}_6\beta_6 \end{pmatrix}, \tag{T5.33}$$

where $\bar{\partial}_a\beta_b = \partial\beta_b/\partial\lambda_a$ at $\bar{\lambda}(\Lambda)$. Part of ϵ is uninteresting: the second trajectory may be the *same* curve in the (λ_4, λ_6) plane, just run a little ahead of the first. Such a displacement points along the tangent vector $\tau_a := \Lambda\,\mathrm{d}\bar{\lambda}_a/\mathrm{d}\Lambda$, with $\tau_4 = \bar{\beta}_4$ and $\tau_6 = 2\bar{\lambda}_6 + \bar{\beta}_6$ by (T5.32). The tangent itself solves (T5.33):

$$\Lambda \frac{\mathrm{d}\tau_a}{\mathrm{d}\Lambda} \overset{(1)}{=} \Lambda \frac{\mathrm{d}}{\mathrm{d}\Lambda} F_a\big(\bar{\lambda}(\Lambda)\big) \overset{(2)}{=} \sum_b \frac{\partial F_a}{\partial \lambda_b}\Big|_{\bar{\lambda}} \Lambda \frac{\mathrm{d}\bar{\lambda}_b}{\mathrm{d}\Lambda} \overset{(3)}{=} \sum_b J_{ab}\tau_b,$$

(1) Write the right sides of (T5.32) as $F_4 = \beta_4$ and $F_6 = 2\lambda_6 + \beta_6$, so $\tau_a = F_a(\bar{\lambda})$.
(2) Chain rule along the trajectory.
(3) $\partial F_a/\partial\lambda_b = J_{ab}$ is exactly the matrix of (T5.33), and $\Lambda\,\mathrm{d}\bar{\lambda}_b/\mathrm{d}\Lambda = \tau_b$.

Split ϵ into a piece along the tangent and a purely vertical piece:

$$\epsilon = a\,\tau + \begin{pmatrix} 0 \\ \xi_6 \end{pmatrix}, \qquad a = \frac{\epsilon_4}{\tau_4}. \tag{T5.34}$$

The number ξ_6 is the genuine vertical distance between the two trajectories, measured at the same λ_4. Insert (T5.34) into (T5.33), with $e_6 := (0,1)$ the vertical unit vector:

$$\Big(\Lambda\frac{\mathrm{d}a}{\mathrm{d}\Lambda}\Big)\tau + a\,J\tau + \Lambda\frac{\mathrm{d}\xi_6}{\mathrm{d}\Lambda}\,e_6 \overset{(1)}{=} \Lambda\frac{\mathrm{d}\epsilon}{\mathrm{d}\Lambda} \overset{(2)}{=} J\epsilon \overset{(3)}{=} a\,J\tau + \xi_6\,Je_6.$$

(1) Product rule on $a\tau + \xi_6 e_6$, and $\Lambda\,\mathrm{d}\tau/\mathrm{d}\Lambda = J\tau$.
(2) (T5.33).
(3) Linearity of J.

The terms $aJ\tau$ cancel. With $Je_6 = (\bar\partial_6\beta_4,\ 2+\bar\partial_6\beta_6)$, the second column of J, the two components read

$$\begin{aligned}\text{row 4:}\quad & (\Lambda\,\mathrm{d}a/\mathrm{d}\Lambda)\,\tau_4 = \bar\partial_6\beta_4\,\xi_6,\\ \text{row 6:}\quad & (\Lambda\,\mathrm{d}a/\mathrm{d}\Lambda)\,\tau_6 + \Lambda\frac{\mathrm{d}\xi_6}{\mathrm{d}\Lambda} = (2+\bar\partial_6\beta_6)\,\xi_6.\end{aligned}$$

Eliminate $\Lambda\,\mathrm{d}a/\mathrm{d}\Lambda$ between them:

$$\Lambda\frac{\mathrm{d}\xi_6}{\mathrm{d}\Lambda} - 2\xi_6 \overset{(1)}{=} \Big(\bar\partial_6\beta_6 - \frac{\tau_6}{\tau_4}\bar\partial_6\beta_4\Big)\xi_6 \overset{(2)}{=} \Big(\bar\partial_6\beta_6 + \bar\partial_4\beta_4 - \Lambda\frac{\mathrm{d}}{\mathrm{d}\Lambda}\ln\bar\beta_4\Big)\xi_6. \tag{T5.35}$$

(1) Row 4 gives $\Lambda\,\mathrm{d}a/\mathrm{d}\Lambda = \bar\partial_6\beta_4\,\xi_6/\tau_4$; put it in row 6.
(2) Chain rule along the trajectory: $\Lambda\,\mathrm{d}\bar\beta_4/\mathrm{d}\Lambda = \bar\partial_4\beta_4\,\tau_4 + \bar\partial_6\beta_4\,\tau_6$; divide by $\tau_4 = \bar\beta_4$, so $-\frac{\tau_6}{\tau_4}\bar\partial_6\beta_4 = \bar\partial_4\beta_4 - \Lambda\frac{\mathrm{d}}{\mathrm{d}\Lambda}\ln\bar\beta_4$.

This is a linear first-order equation for one function, so it integrates directly:

$$\begin{aligned}\ln\frac{\xi_6(\Lambda)}{\xi_6(\Lambda_0)} &\overset{(1)}{=} \int_{\Lambda_0}^{\Lambda}\frac{\mathrm{d}\Lambda'}{\Lambda'}\Big(2+\bar\partial_6\beta_6+\bar\partial_4\beta_4\Big) - \ln\frac{\bar\beta_4(\Lambda)}{\bar\beta_4(\Lambda_0)}\\ \xi_6(\Lambda) &\overset{(2)}{=} \xi_6(\Lambda_0)\,\frac{\Lambda^2}{\Lambda_0^2}\,\frac{\bar\beta_4(\Lambda_0)}{\bar\beta_4(\Lambda)}\,\exp\int_{\Lambda_0}^{\Lambda}\frac{\mathrm{d}\Lambda'}{\Lambda'}\Big(\bar\partial_6\beta_6+\bar\partial_4\beta_4\Big)(\Lambda').\end{aligned} \tag{T5.36}$$

(1) Divide (T5.35) by ξ_6 and integrate $\mathrm{d}\ln\xi_6 = \frac{\mathrm{d}\Lambda}{\Lambda}(\cdots)$ from Λ_0 to Λ; the total-derivative term integrates to a logarithm.
(2) Exponentiate; $\int_{\Lambda_0}^{\Lambda} 2\,\mathrm{d}\Lambda'/\Lambda' = \ln(\Lambda^2/\Lambda_0^2)$.

At weak coupling the integrand of the exponential is small (these are anomalous dimensions, of order the couplings), and the ratio of β_4's varies slowly. So ξ_6 is dominated by Λ^2/Λ_0^2: *two neighbouring trajectories approach each other like* $(\Lambda/\Lambda_0)^2$ *in the infrared.* The infrared theory lies very close to a one-dimensional curve in the (λ_4,λ_6) plane, and the one remaining parameter only says where on that curve we are: it is "transmuted into a scale".

Why this is renormalisability. Start at the bare cutoff Λ_0 with a bare theory $\lambda_4 = \lambda_4^0$, $\lambda_6 = 0$, and flow down to a low scale Λ_R, where λ_4 is *defined* to have the measured value λ_4^R. For a larger bare cutoff Λ_0' choose the bare coupling so that the flow again reaches λ_4^R at Λ_R. This defines a function $\lambda_4^0(\lambda_4^R,\Lambda_R,\Lambda_0)$. The two flows end at the same λ_4, so their difference at Λ_R is purely vertical, and its size is the ξ_6 of (T5.36):

$$\big|\lambda_6(\Lambda_R;\Lambda_0') - \lambda_6(\Lambda_R;\Lambda_0)\big| \overset{(1)}{=} |\xi_6(\Lambda_R)| \overset{(2)}{=} |\xi_6(\Lambda_0)|\,\frac{\Lambda_R^2}{\Lambda_0^2}\,\big(1+O(\lambda)\big) \overset{(3)}{=} O\Big(\frac{\Lambda_R^2}{\Lambda_0^2}\Big).$$

(1) Both flows are trajectories of (T5.32) passing through the same λ_4^R at Λ_R; for Λ_0' close to Λ_0 the difference is small and the linear analysis applies (for a large change of cutoff, add up small steps).

(2) (T5.36) with $\Lambda = \Lambda_R$; at weak coupling λ the exponential and the ratio of β_4's are $1 + O(\lambda)$ up to slowly varying logarithms.
(3) At Λ_0 the first flow has $\lambda_6 = 0$, while the second has already run from Λ_0' to Λ_0, so it has some $\lambda_6 = O(1)$ in cutoff units. So $|\xi_6(\Lambda_0)|$ is a pure number of order one at most.

So as $\Lambda_0 \to \infty$ with Λ_R and λ_4^R held fixed, $\lambda_6(\Lambda_R)$ approaches a limit, with corrections that are inverse powers of the cutoff. That is exactly the definition of a renormalisable theory: *holding the renormalised couplings fixed, everything else has a limit as the cutoff goes to infinity*. The same statement can be packaged in a vector that generalises to infinitely many couplings:

$$v_i(\Lambda) := \Lambda_0 \frac{\partial \lambda_i}{\partial \Lambda_0} - \frac{\partial \lambda_i}{\partial \lambda_4^0}\Big(\frac{\partial \lambda_4}{\partial \lambda_4^0}\Big)^{-1} \Lambda_0 \frac{\partial \lambda_4}{\partial \Lambda_0} = \Lambda_0 \frac{\mathrm{d}}{\mathrm{d}\Lambda_0} \lambda_i\big(\Lambda, \Lambda_0, \lambda_4^0(\Lambda, \lambda_4, \Lambda_0)\big), \tag{T5.37}$$

the change of the couplings at Λ when the bare cutoff moves while $\lambda_4(\Lambda)$ is held fixed. (The two forms agree by the chain rule: the derivative $\Lambda_0\,\mathrm{d}/\mathrm{d}\Lambda_0$ acts both on the explicit Λ_0 and, through λ_4^0, on the bare coupling, and $\Lambda_0\,\partial\lambda_4^0/\partial\Lambda_0$ is fixed by demanding $v_4 = 0$.) By construction $v_4 = 0$, so v is purely vertical, and $v_6(\Lambda)$ is the vertical distance computed above per unit change of $\ln \Lambda_0$: $v_6(\Lambda) = O(\Lambda^2/\Lambda_0^2)$. Now integrate over the cutoff, from Λ_0 to ∞, at $\Lambda = \Lambda_R$:

$$\big|\lambda_6(\Lambda_R;\infty) - \lambda_6(\Lambda_R;\Lambda_0)\big| \overset{(1)}{=} \Big|\int_{\Lambda_0}^{\infty} \frac{\mathrm{d}\Lambda_0'}{\Lambda_0'}\, v_6(\Lambda_R;\Lambda_0')\Big| \overset{(2)}{\le} c \int_{\Lambda_0}^{\infty} \frac{\mathrm{d}\Lambda_0'}{\Lambda_0'} \frac{\Lambda_R^2}{\Lambda_0'^2} \overset{(3)}{=} \frac{c}{2}\frac{\Lambda_R^2}{\Lambda_0^2},$$

(1) $\Lambda_0\,\mathrm{d}\lambda_6/\mathrm{d}\Lambda_0 = v_6$, integrated in $\ln\Lambda_0$.
(2) The bound on v_6, with c a pure number.
(3) $\int_{\Lambda_0}^{\infty} x^{-3}\mathrm{d}x = 1/(2\Lambda_0^2)$.

so the limit $\lambda_6(\Lambda_R;\infty)$ exists and is approached like Λ_R^2/Λ_0^2.

The LPA flow provides a concrete toy of exactly this kind. Stay in $d = 4$ and look at the sextic coupling w of (T5.29), irrelevant with eigenvalue -2 (with α_d absorbed, and in the infrared time t):

$$\partial_t w = -2w - 20u^2 + R_w, \qquad \partial_t u = -8ru + w =: \beta_u, \tag{T5.38}$$

where $R_w = -12rw + w_8$ collects terms of higher order in the small couplings. Instead of the vertical distance to a neighbouring trajectory, we measure w relative to the value it would have if it had relaxed completely.

Worked example T5.9: A concrete toy: the bare value of w is forgotten

Define the deviation $\delta := w + 10u^2$. Then

$$\partial_t \delta \overset{(1)}{=} \partial_t w + 20u\,\partial_t u \overset{(2)}{=} -2w - 20u^2 + R_w + 20u\beta_u \overset{(3)}{=} -2\delta + \underbrace{R_w + 20u\beta_u}_{=:R}.$$

(1) Chain rule.
(2) (T5.38).
(3) $-2w - 20u^2 = -2(w + 10u^2) = -2\delta$.

This linear equation is solved by an integrating factor:

$$\delta(t) \overset{(1)}{=} e^{-2t}\delta(0) + \int_0^t \mathrm{d}s\; e^{-2(t-s)} R(s),$$

$$|\delta(t)| \overset{(2)}{\le} e^{-2t}|\delta(0)| + \max_{s\le t}|R(s)| \int_0^t \mathrm{d}s\; e^{-2(t-s)} \overset{(3)}{\le} \Big(\frac{\Lambda}{\Lambda_0}\Big)^2 |\delta(0)| + \tfrac{1}{2}\max|R|.$$

(1) $\partial_t(e^{2t}\delta) = e^{2t}R$; integrate from 0 to t.

(2) Triangle inequality, and bound $|R|$ by its maximum.
(3) $e^{-2t} = (\Lambda/\Lambda_0)^2$ by the definition of t, and $\int_0^t e^{-2(t-s)}\mathrm{d}s = \frac{1}{2}(1-e^{-2t}) \le \frac{1}{2}$.

Now R is cubic in the small couplings. Near the critical surface the mass is tied to u ($\partial_t r = 0$ in (T5.29) gives $r \approx -u/2$), so $r = O(u)$, $w = O(u^2)$ and $w_8 = O(u^3)$, and term by term

$$R \overset{(1)}{=} -12rw + w_8 + 20u(-8ru + w) \overset{(2)}{=} O(u\cdot u^2) + O(u^3) + O(u\cdot u\cdot u) + O(u\cdot u^2) = O(u^3).$$

(1) $R = R_w + 20u\beta_u$ with R_w and β_u from (T5.38).
(2) Insert the orders of magnitude of r, w, w_8.

Since $\delta = w + 10u^2$, the bound gives, at the low scale Λ_R,

$$w(\Lambda_R) = -10\,u(\Lambda_R)^2 + O(u^3) + O\Big(\frac{\Lambda_R^2}{\Lambda_0^2}\Big) \times (\text{bare data}).$$

The first two terms are fixed by the marginal coupling u at the low scale. All memory of the bare value $w(\Lambda_0)$ sits in the last term, which vanishes as $\Lambda_0 \to \infty$.

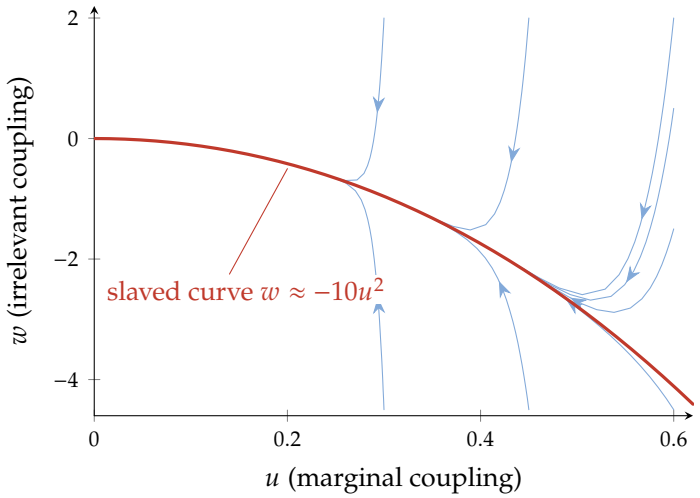


Figure T5.9: Slaving of an irrelevant coupling, computed for the model flow $\partial_t u = -u^2/5$, $\partial_t w = -2w - 20u^2$ (blue curves; arrows towards the infrared). Trajectories start (top and bottom edges) at a few values of the marginal coupling u and with very different irrelevant couplings w; within a few units of RG time they all land on the red curve $w = -10u^2 - 2u^3 - 0.6u^4$ (the attracting solution, computed as a power series), and then move slowly along it. What to see: at low energies the irrelevant coupling is a function of the marginal one. The theory space is squeezed onto a surface whose dimension is the number of relevant and marginal couplings.

T5.2.7 The exact equation in components, and the bound

For $\lambda\varphi^4$ theory in four dimensions the same logic has to control infinitely many couplings with full momentum dependence. The signs are tidier for the **effective interaction** $L(\varphi,\Lambda) := -S^{\mathrm{I}}_{\Lambda}$ [the interaction as it appears in the exponent of e^{+L}], and we use it in this subsection. The bare interaction at Λ_0 is

$$L_{\mathrm{int}} = \int \mathrm{d}^4x \Big(-\tfrac{1}{2}\rho_1^0\varphi^2 - \tfrac{1}{2}\rho_2^0(\partial_\mu\varphi)^2 - \tfrac{1}{4!}\rho_3^0\varphi^4\Big),$$

with the bare mass counterterm $\rho_1^0 = \delta m^2$, the wave-function counterterm $\rho_2^0 = Z - 1$ and the bare quartic coupling $\rho_3^0 = \lambda^0$. The propagator is the massive one, $K(p^2/\Lambda^2)/(p^2 + m^2)$, so the shell propagator is $\Delta_\Lambda = \Lambda\partial_\Lambda K/(p^2 + m^2)$. Put $S^{\mathrm{I}} = -L$ into (T5.19) and go to momentum space:

$$-\Lambda\frac{\partial L}{\partial\Lambda} \overset{(1)}{=} \frac{1}{2}\int_{x,y}\Delta_\Lambda(x-y)\Big[\frac{\delta L}{\delta\varphi(x)}\frac{\delta L}{\delta\varphi(y)} + \frac{\delta^2 L}{\delta\varphi(x)\,\delta\varphi(y)}\Big]$$
$$\overset{(2)}{=} \frac{1}{2}\int\frac{\mathrm{d}^4p}{(2\pi)^4}\,\Delta_\Lambda(p)\Big[\frac{\delta L}{\delta\varphi(-p)}\frac{\delta L}{\delta\varphi(p)} + \frac{\delta^2 L}{\delta\varphi(p)\,\delta\varphi(-p)}\Big].$$

(1) The product of two first derivatives does not change sign under $S^{\mathrm{I}} \to -L$; the left side and the second derivative do.

(2) Insert $\Delta(x-y) = \int_p e^{ip\cdot(x-y)}\Delta(p)$ and $\frac{\delta}{\delta\varphi(x)} = \int_p e^{-ip\cdot x}\frac{\delta}{\delta\varphi(p)}$. This is normalised so that $\delta\varphi(q)/\delta\varphi(p) = (2\pi)^4\delta^4(p-q)$. The x and y integrals give delta functions that set the two momenta to p and $-p$.

Multiplying by -1:

$$\Lambda\frac{\partial L}{\partial\Lambda} = -\frac{1}{2}\int\frac{\mathrm{d}^4p}{(2\pi)^4}\,\Delta_\Lambda(p)\Big[\frac{\delta L}{\delta\varphi(-p)}\frac{\delta L}{\delta\varphi(p)} + \frac{\delta^2 L}{\delta\varphi(p)\,\delta\varphi(-p)}\Big] \tag{T5.39}$$

Its two terms are the tree and loop diagrams of Figure T5.7.

Vertices and relevant coordinates. Expand the effective interaction in vertex functions,

$$L(\varphi,\Lambda) = \sum_{m\geq 1}\frac{1}{(2m)!}\int\prod_{i=1}^{2m}\frac{\mathrm{d}^4p_i}{(2\pi)^4}\,(2\pi)^4\delta^4\Big(\sum_i p_i\Big)\,L_{2m}(p_1,\dots,p_{2m};\Lambda)\,\varphi(p_1)\cdots\varphi(p_{2m}), \tag{T5.40}$$

with each L_{2m} symmetric in its momenta (only even m-point functions, since we keep $\varphi \to -\varphi$). The three **relevant parameters** are

$$\rho_1(\Lambda) = -L_2(0,0;\Lambda), \qquad \rho_2(\Lambda) = -\frac{1}{8}\,\frac{\partial^2}{\partial p_1^\mu\partial p_1^\mu}L_2(p_1,-p_1;\Lambda)\Big|_{p_1=0}, \qquad \rho_3(\Lambda) = -L_4(0,0,0,0;\Lambda). \tag{T5.41}$$

To see where the $\frac{1}{8}$ comes from, read off L_2 for the bare interaction:

$$-\tfrac{1}{2}\int\mathrm{d}^4x\,(\rho_1^0\varphi^2 + \rho_2^0(\partial_\mu\varphi)^2) \overset{(1)}{=} -\tfrac{1}{2}\int\frac{\mathrm{d}^4p}{(2\pi)^4}(\rho_1^0 + \rho_2^0p^2)\varphi(p)\varphi(-p),$$
$$\Longrightarrow \quad L_2(p,-p) \overset{(2)}{=} -\rho_1^0 - \rho_2^0p^2.$$

(1) Fourier transform: $\int\mathrm{d}^4x\,\varphi^2 = \int_p\varphi(p)\varphi(-p)$, and each ∂_μ becomes ip_μ, so $(\partial\varphi)^2 \to (ip)\cdot(-ip) = p^2$.

(2) Compare with the $m = 1$ term of (T5.40),

$$\frac{1}{2!}\int_{p_1,p_2}(2\pi)^4\delta^4(p_1+p_2)\,L_2\,\varphi(p_1)\varphi(p_2) = \tfrac{1}{2}\int_p L_2(p,-p)\,\varphi(p)\varphi(-p).$$

The four-dimensional Laplacian of p^2 is $\partial_\mu\partial_\mu(p_\nu p_\nu) = 2\delta_{\mu\mu} = 8$, so the $\frac{1}{8}$ in ρ_2 returns ρ_2^0. These are the mass, the wave-function normalisation and the quartic coupling, read off at zero momentum: the coordinates on the three-dimensional surface of Figure T5.5. Every other piece of L (the L_{2m} with $m \geq 3$, and the momentum dependence of L_2 and L_4 beyond these three numbers) is irrelevant.

Worked example T5.10: The component equation

Insert (T5.40) into (T5.39) and read off the coefficient of $\varphi(p_1)\cdots\varphi(p_{2m})/(2m)!$. We use $\delta\varphi(q)/\delta\varphi(p) = (2\pi)^4\delta^4(p-q)$, so a derivative removes one field and its momentum integral.

- *Loop term.* $\delta^2/\delta\varphi(p)\delta\varphi(-p)$ acting on the $(2m+2)$-point term removes two fields in $(2m+2)(2m+1)$ ways, which cancels $(2m+2)!$ down to $(2m)!$. It leaves $-\frac{1}{2}\int_p \Delta_\Lambda(p)\, L_{2m+2}(p_1,\dots,p_{2m},p,-p)$.
- *Tree term.* $\delta L/\delta\varphi(p)$ from the $2l$-point term is $\frac{1}{(2l-1)!}\int L_{2l}(q_1,\dots,q_{2l-1},p)\varphi^{2l-1}$. The product of the $2l$ and the $2l'$ terms has $2l-1+2l'-1$ fields, which is $2m$ for $l' = m+1-l$, with $l = 1,\dots,m$. Symmetrising over which $2l-1$ of the $2m$ external legs sit on the first vertex gives $\frac{(2m)!}{(2l-1)!(2m+1-2l)!} = \binom{2m}{2l-1}$ distinct terms, and the factorials cancel against $1/(2m)!$. The shell line carries the momentum P equal to the sum of the momenta on the first vertex, by momentum conservation at that vertex. Both terms carry the prefactor $-\frac{1}{2}$ of (T5.39).

Now measure everything in units of Λ. The vertex L_{2m} has mass dimension $4-2m$: the term (T5.40) is dimensionless, $\varphi(p)$ has dimension $[\varphi(x)] - 4 = 1-4 = -3$, each d^4p has dimension 4 and δ^4 has -4, so

$$0 \overset{(1)}{=} 2m\cdot 4 - 4 + [L_{2m}] + 2m\cdot(-3) \quad\Longrightarrow\quad [L_{2m}] = 4-2m.$$

(1) Add the dimensions of the $2m$ measures, the delta function, the vertex and the $2m$ fields.

So define the dimensionless vertices $A_{2m} := \Lambda^{2m-4}L_{2m}$ and the dimensionless shell factor $Q(p,\Lambda,m^2) := \Lambda^2\Delta_\Lambda(p)$ (recall $[\Delta] = -2$). By the product rule $\Lambda\partial_\Lambda A_{2m} = (2m-4)A_{2m} + \Lambda^{2m-4}\Lambda\partial_\Lambda L_{2m}$, so the left side below is $\Lambda^{2m-4}\Lambda\partial_\Lambda L_{2m}$:

$$\Big(\Lambda\frac{\partial}{\partial\Lambda} + 4 - 2m\Big)A_{2m}(p_1,\dots,p_{2m}) \overset{(1)}{=} -\frac{1}{2}\sum_{l=1}^{m}\sum_{\binom{2m}{2l-1}\text{ splits}} Q(P)\,A_{2l}(\dots,P)\,A_{2m+2-2l}(\dots,-P) - \frac{1}{2}\int\frac{\mathrm{d}^4p}{(2\pi\Lambda)^4}\,A_{2m+2}(p_1,\dots,p_{2m},p,-p)\,Q(p). \tag{T5.42}$$

(1) The two bullets, multiplied by Λ^{2m-4}. The powers of Λ work out: in the tree term $\Delta L_{2l}L_{2m+2-2l} = (Q/\Lambda^2)\Lambda^{4-2l}\Lambda^{2-2m+2l}AA = Q\Lambda^{4-2m}AA$; in the loop term $\Delta L_{2m+2} = (Q/\Lambda^2)\Lambda^{2-2m}A$, and $\Lambda^{2m-4}\cdot\Lambda^{-2m}\int\mathrm{d}^4p/(2\pi)^4 = \int\mathrm{d}^4p/(2\pi\Lambda)^4$.

This is the component form of the exact RG equation (Polchinski 1984, eq. (34)).

The left side of (T5.42) is pure dimensional analysis: with the right side switched off, $A_{2m}\propto\Lambda^{2m-4}$, so the $2m$-point vertices with $m\geq 3$ die like $(\Lambda/\Lambda_0)^{2m-4}$ as Λ decreases. The right side is where the interactions feed one vertex into another.

The bound. Choose K to be exactly 1 for $p^2\leq\Lambda^2$ and exactly 0 for $p^2\geq 4\Lambda^2$, smooth in between. With $x = p^2/\Lambda^2$, $\Lambda\partial_\Lambda K(x) = -2xK'(x)$ (as in (T5.8)), which is non-zero only on the shell $1<x<4$. Let $k_1 := \max_{1\leq x\leq 4}|2xK'(x)|$, a pure number fixed by the choice of K, and assume $|m^2|\leq\frac{1}{2}\Lambda_R^2$, with $\Lambda\geq\Lambda_R$ throughout. Two elementary estimates follow. On the shell,

$$|Q(p,\Lambda,m^2)| \overset{(1)}{=} \frac{\Lambda^2\,|\Lambda\partial_\Lambda K|}{p^2+m^2} \overset{(2)}{\leq} \frac{\Lambda^2\,k_1}{\Lambda^2-\frac{1}{2}\Lambda^2} \overset{(3)}{=} 2k_1 =: D_0,$$

$$\int \frac{\mathrm{d}^4p}{(2\pi\Lambda)^4}\,|Q| \overset{(4)}{\le} D_0\,\frac{\mathrm{vol}\{\Lambda^2<p^2<4\Lambda^2\}}{(2\pi\Lambda)^4} \overset{(5)}{=} D_0\,\frac{\frac{\pi^2}{2}(16-1)\Lambda^4}{16\pi^4\Lambda^4} = \frac{15}{32\pi^2}\,D_0 =: C.$$

(1) Definition $Q=\Lambda^2\Delta_\Lambda$ with $\Delta_\Lambda=\Lambda\partial_\Lambda K/(p^2+m^2)$.
(2) On the shell $p^2>\Lambda^2$ and $|m^2|\le\frac12\Lambda_R^2\le\frac12\Lambda^2$, so $p^2+m^2\ge\frac12\Lambda^2$; and $|\Lambda\partial_\Lambda K|\le k_1$.
(3) Arithmetic.
(4) Q vanishes off the shell, and $|Q|\le D_0$ on it.
(5) The volume of a ball of radius R in four dimensions is $\frac{\pi^2}{2}R^4$; subtract the ball of radius Λ from the ball of radius 2Λ.

Write $\|f\|:=\max_{p_i^2<4\Lambda^2}|f(p_1,\dots)|$, the largest value on the momenta that are still present. Taking absolute values in (T5.42) and replacing every factor by its maximum,

$$\Big\|\Big(\Lambda\frac{\partial}{\partial\Lambda}+4-2m\Big)A_{2m}\Big\| \;\le\; \sum_{l=1}^{m}\frac12\binom{2m}{2l-1}D_0\,\|A_{2l}\|\,\|A_{2m+2-2l}\|+\frac12 C\,\|A_{2m+2}\|. \tag{T5.43}$$

(Each tree term is bounded by $\frac12 D_0\|A_{2l}\|\|A_{2m+2-2l}\|$ and there are $\binom{2m}{2l-1}$ of them; the loop term is bounded by $\|A_{2m+2}\|$ times the integral of $|Q|$.) This bound is the heart of the proof: because every momentum integral runs over a bounded range, the momentum dependence can be estimated naively, and the infinite system becomes as simple as the toy flow.

Integrating the bound. For an irrelevant vertex ($m\ge3$), write the right side of (T5.42) as $R_{2m}(\Lambda)$ and solve the linear equation by an integrating factor, from the bare scale down:

$$\begin{aligned}
&\Lambda\frac{\partial}{\partial\Lambda}\big(\Lambda^{4-2m}A_{2m}\big)=\Lambda^{4-2m}R_{2m} \quad\Longrightarrow\\
&A_{2m}(\Lambda)\overset{(1)}{=}\Big(\frac{\Lambda}{\Lambda_0}\Big)^{2m-4}A_{2m}(\Lambda_0)-\int_\Lambda^{\Lambda_0}\frac{\mathrm{d}\Lambda'}{\Lambda'}\Big(\frac{\Lambda}{\Lambda'}\Big)^{2m-4}R_{2m}(\Lambda'),\\
&\|A_{2m}(\Lambda)\|\overset{(2)}{\le}\max_{\Lambda\le\Lambda'\le\Lambda_0}\|R_{2m}(\Lambda')\|\int_\Lambda^{\Lambda_0}\frac{\mathrm{d}\Lambda'}{\Lambda'}\Big(\frac{\Lambda}{\Lambda'}\Big)^{2m-4}\\
&\qquad\overset{(3)}{\le}\frac{\max\|R_{2m}\|}{2m-4}.
\end{aligned} \tag{T5.44}$$

(1) The first equation holds because $\Lambda\partial_\Lambda\Lambda^{4-2m}=(4-2m)\Lambda^{4-2m}$. Integrate it in $\ln\Lambda'$ from Λ to Λ_0: $\Lambda_0^{4-2m}A_{2m}(\Lambda_0)-\Lambda^{4-2m}A_{2m}(\Lambda)=\int_\Lambda^{\Lambda_0}\frac{\mathrm{d}\Lambda'}{\Lambda'}\Lambda'^{4-2m}R_{2m}(\Lambda')$, then solve for $A_{2m}(\Lambda)$ by multiplying by Λ^{2m-4}.
(2) The bare vertices with $m\ge3$ vanish, $A_{2m}(\Lambda_0)=0$, since L_{int} has only φ^2, $(\partial\varphi)^2$ and φ^4.
(3) $\int_\Lambda^{\Lambda_0}(\Lambda/\Lambda')^{2m-4}\mathrm{d}\Lambda'/\Lambda'=\frac{1-(\Lambda/\Lambda_0)^{2m-4}}{2m-4}\le\frac{1}{2m-4}$.

So an irrelevant vertex is bounded by its sources, uniformly in Λ_0. The relevant parts ($m=1$ and $m=2$ at zero momentum) must *not* be integrated down from Λ_0. For $m=1$ the first term of (T5.44) is $(\Lambda/\Lambda_0)^{-2}A_2(\Lambda_0)=(\Lambda_0/\Lambda)^2A_2(\Lambda_0)$: a mass integrated down grows like $(\Lambda_0/\Lambda)^2$, and for $m=2$ the factor $(\Lambda/\Lambda_0)^0=1$ gives no suppression either. Instead the relevant parts are fixed at the low scale by **renormalisation conditions** [the measured mass, normalisation and coupling at the scale Λ_R],

$$\rho_1(\Lambda_R)=0,\qquad \rho_2(\Lambda_R)=0,\qquad \rho_3(\Lambda_R)=\lambda^R, \tag{T5.45}$$

and integrates *up* from Λ_R. The fact that the mass must be fixed at the bottom, by a finely tuned bare ρ_1^0, is the naturalness problem of light scalars.

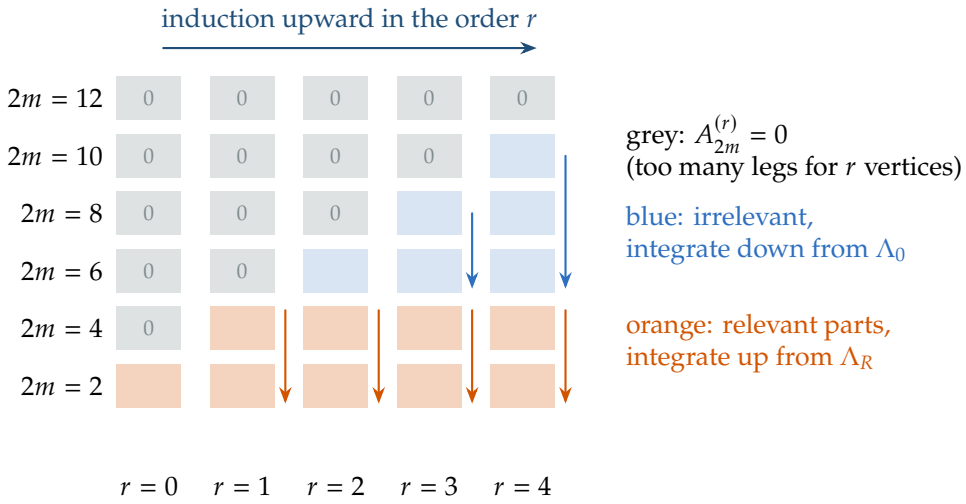


Figure T5.10: The order in which Polchinski's bounds are proved. Each cell is one vertex A_{2m} at order r in the renormalised coupling λ^R. At order r a tree of r vertices has at most $2r+2$ legs, so the grey cells vanish. Within a column the proof goes down in m (arrows), because the bound (T5.43) on A_{2m} only needs A_{2m+2} at the same order and products of lower orders. The two lowest rows hold the relevant parameters; they come last in each column, and they are anchored at Λ_R and integrated upwards in Λ (orange) instead of down from Λ_0. What to see: at every step only quantities already bounded appear on the right side.

Key idea: Polchinski's theorem (1984)

Fix $\rho_a(\Lambda_R)$ by (T5.45) and let $\tilde{L}(\varphi, \Lambda_R, \lambda^R, \Lambda_0)$ be the effective interaction at Λ_R. Then, order by order in λ^R, the limit $\Lambda_0 \to \infty$ of $\tilde{L}$ exists, and at order r (Polchinski 1984, eq. (47))

$$\big\|\tilde{L}^{(r)}_{2m}(\Lambda_R, \Lambda_0) - \tilde{L}^{(r)}_{2m}(\Lambda_R, \infty)\big\| \le \Lambda_R^{4-2m}\Big(\frac{\Lambda_R}{\Lambda_0}\Big)^2 P\big(\ln(\Lambda_0/\Lambda_R)\big),$$

with P a polynomial with pure-number coefficients (and zero if $m > r+1$). The correlation functions at momenta below Λ_R therefore depend on the cutoff only through inverse powers: $\lambda\varphi^4$ theory in four dimensions is perturbatively renormalisable.

The proof is the induction of Figure T5.10, run twice. Write $A^{(r)}_{2m}$ for the part of A_{2m} of order r in λ^R, and $\ell := \ln(\Lambda_0/\Lambda_R)$.

Step 1: the vertices are bounded by polynomials in ℓ. At order r a tree of r quartic vertices has at most $2r+2$ external legs, so $A^{(r)}_{2m} = 0$ for $m > r+1$. Assume the bound for all orders below r, and at order r for all $m' > m$. On the right side of (T5.43) for $A^{(r)}_{2m}$ the tree terms are products $A^{(r_1)}A^{(r_2)}$ with $r_1 + r_2 = r$ and $r_1, r_2 \ge 1$ (lower orders), and the loop term is $A^{(r)}_{2m+2}$ (same order, more legs). All are already bounded, so $\|R^{(r)}_{2m}\| \le P(\ell)$ for some polynomial P. Now integrate:

- $m \ge 3$ (irrelevant): (T5.44) gives $\|A^{(r)}_{2m}\| \le \max \|R^{(r)}_{2m}\|/(2m-4)$.
- $m = 2$ at zero momentum (ρ_3, marginal): integrate $\Lambda\partial_\Lambda A_4 = R_4$ *up* from Λ_R, where $A_4 = -\lambda^R$: $|A_4(\Lambda)| \le \lambda^R + \max|R_4|\,\ln(\Lambda/\Lambda_R)$. This is where the powers of ℓ come from.
- $m = 1$ at zero momentum (ρ_1, relevant): integrate $\Lambda\partial_\Lambda(\Lambda^2 A_2) = \Lambda^2 R_2$ up from Λ_R, where $A_2 = 0$: $|A_2(\Lambda)| \le \Lambda^{-2}\int_{\Lambda_R}^{\Lambda} \frac{\mathrm{d}\Lambda'}{\Lambda'}\Lambda'^2 \max|R_2| \le \frac{1}{2}\max|R_2|$.

(The momentum dependence of A_2 and A_4, and ρ_2, are handled in the same way, by Taylor

expanding in the external momenta.) So every $\|A^{(r)}_{2m}\|$ is bounded by a polynomial in ℓ, uniformly in Λ_0.

Step 2: the cutoff dependence is $(\Lambda_R/\Lambda_0)^2$. Define the analogue of (T5.37),

$$V(\Lambda) = \Lambda_0 \frac{\partial L}{\partial \Lambda_0} - \sum_{a,b} \frac{\partial L}{\partial \rho^0_a} \frac{\partial \rho^0_a}{\partial \rho_b(\Lambda)} \Lambda_0 \frac{\partial \rho_b(\Lambda)}{\partial \Lambda_0},$$

the change of the effective interaction when Λ_0 moves with the three relevant parameters held fixed. By construction V has no relevant component. Differentiating (T5.39) with respect to Λ_0 gives an equation *linear* in V: its left side is the operator $\Lambda\partial_\Lambda + 4 - 2m$ of (T5.42) acting on V_{2m}, and its right side has one factor V in each term (the other factor is an A, already bounded in step 1). The dimensionless V_{2m} start at Λ_0 with values of order one (times powers of ℓ). Running the same induction, the integrating factor $(\Lambda/\Lambda_0)^{2m-4} \le (\Lambda/\Lambda_0)^2$ for $m \ge 3$ makes every component of V at Λ_R at most $\Lambda_R^{4-2m}(\Lambda_R/\Lambda_0)^2 P(\ell)$ in dimensionful units (Polchinski 1984 carries out this estimate in detail). Finally integrate $\Lambda_0\,\partial\tilde{L}/\partial\Lambda_0 = V$ over the cutoff. For a term ℓ^k of P,

$$\int_{\Lambda_0}^{\infty} \frac{\mathrm{d}\Lambda_0'}{\Lambda_0'} \Big(\frac{\Lambda_R}{\Lambda_0'}\Big)^2 \Big(\ln\frac{\Lambda_0'}{\Lambda_R}\Big)^k \overset{(1)}{=} \Big(\frac{\Lambda_R}{\Lambda_0}\Big)^2 \int_0^{\infty} \mathrm{d}s\; e^{-2s}(s+\ell)^k \overset{(2)}{=} \Big(\frac{\Lambda_R}{\Lambda_0}\Big)^2 \sum_{j=0}^{k} \binom{k}{j} \frac{j!}{2^{j+1}}\, \ell^{k-j}.$$

(1) Substitute $s = \ln(\Lambda_0'/\Lambda_0)$, so $\Lambda_0' = \Lambda_0 e^s$ and $\ln(\Lambda_0'/\Lambda_R) = s + \ell$.

(2) Binomial expansion and $\int_0^\infty s^j e^{-2s}\mathrm{d}s = j!/2^{j+1}$.

The result is again $(\Lambda_R/\Lambda_0)^2$ times a polynomial in ℓ, which is the bound of the theorem. The proof never mentions divergent integrals, counterterms or overlapping subdivergences; it is a statement about the attraction of a three-dimensional surface in theory space.

Two remarks complete the picture. First, the relevant coupling behaves the opposite way. By (T5.29) the dimensionless mass has eigenvalue 2: near the critical surface, at the critical value r_c of the bare mass, the linearised flow is $\partial_t(r - r_c) = 2(r - r_c)$. Then

$$r(t_R) - r_c \overset{(1)}{=} \big(r(0) - r_c\big) e^{2t_R} \overset{(2)}{=} \big(r(0) - r_c\big) \frac{\Lambda_0^2}{\Lambda_R^2} \quad\Longrightarrow\quad r(0) \overset{(3)}{=} r_c + \big(r(t_R) - r_c\big) \frac{\Lambda_R^2}{\Lambda_0^2}.$$

(1) Solve the linear equation from $t = 0$ (the scale Λ_0) to $t_R = \ln(\Lambda_0/\Lambda_R)$.

(2) $e^{t_R} = \Lambda_0/\Lambda_R$.

(3) Solve for the bare value.

A finite mass at Λ_R, $r(t_R) - r_c = O(1)$, needs a bare value tuned to the critical surface with ever greater precision, $r(\Lambda_0) = r_c + O(\Lambda_R^2/\Lambda_0^2)$. This is the fine-tuning (or hierarchy) problem of scalar masses ([QFT notes §15.9.4]). Second, a theory with an irrelevant coupling of its own (Fermi theory, gravity) is not thereby useless: at energies $E \ll \Lambda_0$ it is an **effective field theory**, predictive order by order in E/Λ_0 ([QFT notes §15.9.5]).

Recap

- A smooth cutoff $C_\Lambda = K(p^2/\Lambda^2)/p^2$; its derivative $\Delta_\Lambda = \Lambda\partial_\Lambda C_\Lambda \ge 0$ is the propagator of a thin shell.
- Covariances of independent Gaussians add, and Gaussian smearing is the heat operator $e^{\frac{1}{2}\partial\cdot D\cdot\partial}$.
- Together they give the exact Polchinski equation $\partial_t e^{-S^{\mathrm{I}}} = \frac{1}{2}\frac{\delta}{\delta\varphi}\cdot\Delta\cdot\frac{\delta}{\delta\varphi}e^{-S^{\mathrm{I}}}$: a tree term

joining vertices and a loop term closing legs.

- Its solution is the sum of Feynman diagrams with internal lines $C_{\Lambda_0} - C_\Lambda$; a mass insertion is resummed exactly.
- The LPA gives $\partial_t v = dv - \frac{d-2}{2}\hat{\varphi}v' - v'^2 + v''$; in $d = 4 - \varepsilon$ it has a Wilson–Fisher fixed point with eigenvalues $2 - \varepsilon/3$ and $-\varepsilon$.
- Polchinski's toy flow: neighbouring trajectories converge like $(\Lambda/\Lambda_0)^2$, so irrelevant couplings become functions of the relevant ones.
- For $\lambda\varphi^4$ the dimensionless vertices obey $(\Lambda\partial_\Lambda + 4 - 2m)A_{2m} = (\text{tree}) + (\text{loop})$, bounded as in (T5.43); induction in the order and in m gives cutoff dependence $(\Lambda_R/\Lambda_0)^2$: renormalisability.

T5.3 Fixed points, linearisation and critical exponents

A fixed point is a theory without a scale. Near it the flow is almost linear, and a linear flow is characterised by the eigenvalues of a matrix. Those eigenvalues are the critical exponents, and they are also the scaling dimensions of the operators of the fixed-point theory. How do we compute them? The most important example in statistical physics shows the whole procedure: the Wilson–Fisher fixed point of φ^4 theory in $d = 4 - \varepsilon$ dimensions, which describes the Ising universality class (magnets with one preferred axis, the liquid–gas critical point, binary alloys).

T5.3.1 The linearised flow and scaling fields

Let g^* be a fixed point of (T5.5), $\beta_i(g^*) = 0$, and write $g_i = g_i^* + \delta g_i$. Taylor expanding the β-functions to first order,

$$\frac{\mathrm{d}\,\delta g_i}{\mathrm{d}t} \overset{(1)}{=} \beta_i(g^* + \delta g) \overset{(2)}{=} \sum_j M_{ij}\,\delta g_j + O(\delta g^2), \qquad M_{ij} := \left.\frac{\partial \beta_i}{\partial g_j}\right|_{g^*}. \tag{T5.46}$$

(1) g^* is constant, so $\mathrm{d}g/\mathrm{d}t = \mathrm{d}\,\delta g/\mathrm{d}t$.
(2) First-order Taylor expansion; $\beta_i(g^*) = 0$.

Suppose the **stability matrix** M has a basis of eigenvectors, $Me_a = y_a e_a$ (this is the generic case; the eigenvalues are real in every unitary example we meet). Expand $\delta g = \sum_a u_a\, e_a$. Then

$$\sum_a \frac{\mathrm{d}u_a}{\mathrm{d}t} e_a \overset{(1)}{=} M \sum_a u_a e_a \overset{(2)}{=} \sum_a y_a u_a e_a \quad\Longrightarrow\quad u_a(t) = u_a(0)\, e^{y_a t} = u_a(0)\left(\frac{\Lambda_0}{\Lambda}\right)^{y_a}. \tag{T5.47}$$

(1) Insert the expansion into (T5.46).
(2) $Me_a = y_a e_a$; then compare coefficients of the independent e_a.

Working definition T5.6 (scaling fields and RG eigenvalues)

Near a fixed point, replace the couplings g_i [the numbers we write in a Lagrangian] by the **scaling fields** u_a [the components of δg along the eigenvectors e_a of the stability matrix M]. In these coordinates the flow is diagonal: each u_a is multiplied by e^{y_a} per unit of RG time, and the y_a are the **RG eigenvalues** [the eigenvalues of M]. Picture the fixed point as a saddle in a landscape: downhill directions ($y_a > 0$, relevant, the flow runs away), uphill directions ($y_a < 0$, irrelevant, the flow comes in) and possibly flat ones ($y_a = 0$, marginal).

Operational recipe: compute β_i, find g^*, differentiate to get M, diagonalise.

Each scaling field multiplies a definite operator. If the Wilsonian action near the fixed point is $S = S^* + \sum_a u_a \int \mathrm{d}^d x\, \mathcal{O}_a(x)$, the operator $\mathcal{O}_a$ has a definite scaling dimension Δ_a at the fixed point: measured in the rescaled units $x' = x/b$ it becomes $\mathcal{O}_a(x) = b^{-\Delta_a}\mathcal{O}'_a(x')$, with $\mathcal{O}'_a$ normalised in the new units exactly as $\mathcal{O}_a$ was in the old. Follow one RG step with factor b, which is RG time $\ln b$:

$$u_a \int \mathrm{d}^d x\, \mathcal{O}_a(x) \overset{(1)}{=} u_a \int b^d\, \mathrm{d}^d x'\; b^{-\Delta_a}\mathcal{O}'_a(x') \overset{(2)}{=} \left(b^{d-\Delta_a} u_a\right) \int \mathrm{d}^d x'\, \mathcal{O}'_a(x'),$$
$$u'_a \overset{(3)}{=} b^{y_a} u_a \quad \Longrightarrow \quad y_a = d - \Delta_a. \tag{T5.48}$$

(1) Substitute $x = bx'$, so $\mathrm{d}^d x = b^d \mathrm{d}^d x'$, and the scaling of the operator.
(2) Collect the powers of b: in the new units the coupling of $\mathcal{O}'_a$ is $u'_a = b^{d-\Delta_a} u_a$.
(3) (T5.47) with $t = \ln b$: $e^{y_a \ln b} = b^{y_a}$. Comparing the two expressions for u'_a gives the result.

The *spectrum of the linearised RG* and the *spectrum of scaling dimensions of the fixed-point theory* are the same list of numbers.

The correlation length follows from (T5.1). Let there be one relevant scaling field u_t (the temperature) with eigenvalue y_t. After RG time t, lengths in the new units are e^t times shorter, so

$$\xi\big(u_t(0)\big) \overset{(1)}{=} e^t\, \xi\big(u_t(0) e^{y_t t}\big) \overset{(2)}{=} u_t(0)^{-1/y_t}\, \xi(1),$$

(1) (T5.1) with $b = e^t$ and (T5.47).
(2) Stop the flow when $u_t = 1$, that is at $e^t = u_t(0)^{-1/y_t}$; from there on the system is far from criticality and $\xi(1)$ is a number of order one.

So $\xi \propto u_t^{-1/y_t}$, and since $u_t \propto T - T_c$ close to the critical point, $\nu = 1/y_t$. (Chapter 11 applies the same argument to the two-dimensional Ising model.)

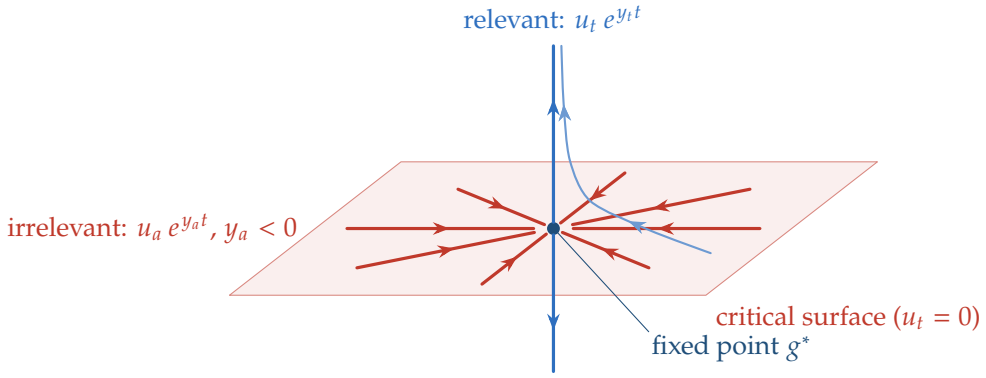


Figure T5.11: The linearised flow near a fixed point with one relevant and two irrelevant directions. Inside the critical surface (red plane) all flow lines converge on the fixed point, each component shrinking like $e^{y_a t}$. The relevant scaling field (blue axis) grows like $e^{y_t t}$. A theory starting slightly above the surface (thin blue curve) first slides in towards g^*, then leaves along the relevant axis. What to see: the number of relevant directions is the number of parameters one must tune to reach the critical point (one, the temperature, for the Ising class at zero field; two if the field is counted).

Worked example T5.11: The scaling operators of the Gaussian fixed point

Linearise the LPA (T5.28) (with $\alpha_d = 1$) about the Gaussian point $v = 0$. The quadratic term v'^2 drops, leaving the linear operator

$$\partial_t\, \delta v = \mathsf{L}\, \delta v, \qquad \mathsf{L} \coloneqq d - \tfrac{d-2}{2}\, \hat\varphi\, \frac{\mathrm{d}}{\mathrm{d}\hat\varphi} + \frac{\mathrm{d}^2}{\mathrm{d}\hat\varphi^2}.$$

On a monomial,

$$\mathsf{L}\, \hat\varphi^n \overset{(1)}{=} \Big(d - n\, \frac{d-2}{2}\Big)\hat\varphi^n + n(n-1)\hat\varphi^{n-2}. \tag{T5.49}$$

(1) $\hat\varphi\, \mathrm{d}\hat\varphi^n/\mathrm{d}\hat\varphi = n\hat\varphi^n$ (the degree counter of a monomial) and $\mathrm{d}^2\hat\varphi^n/\mathrm{d}\hat\varphi^2 = n(n-1)\hat\varphi^{n-2}$.

So L lowers the degree or keeps it: in the basis $1, \hat\varphi^2, \hat\varphi^4, \ldots$ it is an upper-triangular matrix, and its eigenvalues are the diagonal entries $y_n = d - n\frac{d-2}{2}$, the power-counting values of Example T5.4. The eigenfunctions are polynomials $\hat\varphi^n$ + (lower). For $n = 2$ try $\hat\varphi^2 + c$:

$$\mathsf{L}(\hat\varphi^2 + c) \overset{(1)}{=} 2\hat\varphi^2 + 2 + dc \overset{(2)}{=} 2(\hat\varphi^2 + c) \quad\Longleftrightarrow\quad c = -\frac{2}{d-2}.$$

(1) (T5.49) with $n = 2$ and $n = 0$.

(2) Demand the eigenvalue $y_2 = 2$ and solve $2 + dc = 2c$.

For $n = 4$ try $\hat\varphi^4 + a\hat\varphi^2 + b$, with eigenvalue $y_4 = d - 2(d-2) = 4 - d$:

$$\begin{aligned}\mathsf{L}(\hat\varphi^4 + a\hat\varphi^2 + b) &\overset{(1)}{=} (4-d)\hat\varphi^4 + 12\hat\varphi^2 + a\big(2\hat\varphi^2 + 2\big) + d\,b \\ &\overset{(2)}{=} (4-d)\hat\varphi^4 + (12 + 2a)\hat\varphi^2 + (2a + d\,b).\end{aligned}$$

(1) (T5.49) with $n = 4$ (diagonal entry $d - 2(d-2) = 4 - d$, lower term $4 \cdot 3\,\hat\varphi^2$), with $n = 2$ and with $n = 0$.

(2) Collect powers of $\hat\varphi$.

Demand that this equal $(4-d)(\hat\varphi^4 + a\hat\varphi^2 + b)$ and match the two lower powers:

$$\begin{aligned}\hat\varphi^2 &: \; 12 + 2a = (4-d)a \;\Longrightarrow\; a \overset{(1)}{=} \frac{12}{2-d}, \\ \hat\varphi^0 &: \; 2a + d\,b = (4-d)b \;\Longrightarrow\; b \overset{(2)}{=} \frac{2a}{4-2d} = \frac{a}{2-d} = \frac{12}{(2-d)^2}.\end{aligned}$$

(1) $a(4 - d - 2) = 12$.

(2) $b(4 - d - d) = 2a$, then insert a.

In $d = 4$: $c = -1$, $a = -6$, $b = 3$, and the two eigenfunctions are

$$\hat\varphi^2 - 1, \qquad \hat\varphi^4 - 6\hat\varphi^2 + 3,$$

the Hermite polynomials He_2 and He_4. Physically they are the normal-ordered operators $:\varphi^2:$ and $:\varphi^4:$: subtracting the self-contractions (the tadpoles of Example T5.5) is exactly what makes φ^n an eigenoperator of the RG. The "lower" terms are the mixing of φ^n with operators of lower dimension.

Worked example T5.12: A marginal coupling runs logarithmically

At $d = 4$ the quartic coupling has $y = 0$; its flow starts at second order, $\mathrm{d}u/\mathrm{d}t = -bu^2$ with

$b > 0$ (we find b in Example T5.13). Separate variables:

$$-\frac{\mathrm{d}u}{u^2} = b\,\mathrm{d}t \implies \frac{1}{u(t)} - \frac{1}{u(0)} \overset{(1)}{=} bt \implies u(t) \overset{(2)}{=} \frac{u(0)}{1 + b\,u(0)\ln(\Lambda_0/\Lambda)}.$$

(1) Integrate from 0 to t; $\int u^{-2}\mathrm{d}u = -u^{-1}$.
(2) Solve for $u(t)$ and use $t = \ln(\Lambda_0/\Lambda)$.

The coupling dies off towards the infrared, but only logarithmically: it is **marginally irrelevant**. Run backwards (towards the ultraviolet), it grows and blows up at $\ln(\Lambda_0/\Lambda) = -1/(bu(0))$, the Landau pole. QCD is the opposite case, **marginally relevant** ($b < 0$): weak at short distances and strong at long ones.

T5.3.2 From two eigenvalues to all six exponents

A magnet near its critical point has two relevant directions: the reduced temperature t (eigenvalue y_t, operator the energy density ε) and the magnetic field h (eigenvalue y_h, operator the spin σ). The **singular free energy density** $f(t,h)$ (the non-analytic part of $-\ln Z$ per unit volume) has dimension (length)$^{-d}$. One RG step with factor b leaves Z unchanged but shrinks the volume in new units by b^{-d}, and moves the couplings to $(b^{y_t}t, b^{y_h}h)$, so

$$f(t,h) = b^{-d}\,f\left(b^{y_t}t,\; b^{y_h}h\right). \tag{T5.50}$$

The standard thermodynamic quantities are derivatives of f: the specific heat $C \propto \partial^2 f/\partial t^2$, the magnetisation $M = -\partial f/\partial h$ and the susceptibility $\chi = \partial^2 f/\partial h^2$. Their exponents are defined by $C \sim |t|^{-\alpha}$, $M \sim (-t)^{\beta}$ at $h = 0$, $\chi \sim |t|^{-\gamma}$, and $M \sim h^{1/\delta}$ at $t = 0$. Choosing b cleverly gives each of them.

$$\begin{aligned}
f(t,0) &\overset{(1)}{=} |t|^{d/y_t} f(\pm 1, 0) && \implies C \propto |t|^{d/y_t - 2}, \quad \alpha = 2 - \frac{d}{y_t},\\
M(t) &\overset{(2)}{=} b^{y_h - d} M(b^{y_t}t, 0) = |t|^{(d-y_h)/y_t} M(-1,0) && \implies \beta = \frac{d - y_h}{y_t},\\
\chi(t) &\overset{(3)}{=} b^{2y_h - d}\chi(b^{y_t}t, 0) && \implies \gamma = \frac{2y_h - d}{y_t},\\
M(0,h) &\overset{(4)}{=} b^{y_h - d} M(0, b^{y_h}h) = h^{(d-y_h)/y_h} M(0,1) && \implies \delta = \frac{y_h}{d - y_h}.
\end{aligned}$$

(1) Put $h = 0$ and $b = |t|^{-1/y_t}$ in (T5.50), so that the rescaled temperature is ± 1 and $b^{-d} = |t|^{d/y_t}$; differentiate twice in t.
(2) Differentiate (T5.50) once in h (chain rule gives b^{y_h}), then choose $b = |t|^{-1/y_t}$ with $t < 0$.
(3) Differentiate twice in h (factor b^{2y_h}), same choice of b.
(4) Put $t = 0$ and choose $b = h^{-1/y_h}$.

Together with $\nu = 1/y_t$ and, from the power law of the spin correlation at T_c, $2\Delta_\sigma = d - 2 + \eta$ with $\Delta_\sigma = d - y_h$ by (T5.48):

Key idea: critical exponents from two scaling dimensions

With $\Delta_\varepsilon = d - y_t$ and $\Delta_\sigma = d - y_h$,

$$\nu = \frac{1}{d - \Delta_\varepsilon}, \qquad \eta = 2\Delta_\sigma + 2 - d, \qquad \alpha = 2 - \frac{d}{d - \Delta_\varepsilon},$$

$$\beta = \frac{\Delta_\sigma}{d-\Delta_\varepsilon}, \qquad \gamma = \frac{d-2\Delta_\sigma}{d-\Delta_\varepsilon}, \qquad \delta = \frac{d-\Delta_\sigma}{\Delta_\sigma}.$$

Six exponents from two numbers imply four **scaling relations**, for example $\alpha + 2\beta + \gamma = 2$ and $\gamma = \nu(2-\eta)$, which hold for every universality class. Chapter 11 uses this table with the exact two-dimensional values $\Delta_\sigma = \frac{1}{8}$, $\Delta_\varepsilon = 1$.

The four relations follow by inserting the table. Write $y_t = d - \Delta_\varepsilon$:

$$\alpha + 2\beta + \gamma \overset{(1)}{=} \frac{2y_t - d + 2\Delta_\sigma + d - 2\Delta_\sigma}{y_t} = 2, \qquad \nu(2-\eta) \overset{(2)}{=} \frac{2 - 2\Delta_\sigma - 2 + d}{y_t} = \gamma,$$

$$\beta(\delta - 1) \overset{(3)}{=} \frac{\Delta_\sigma}{y_t} \cdot \frac{d - 2\Delta_\sigma}{\Delta_\sigma} = \gamma, \qquad d\,\nu \overset{(4)}{=} \frac{d}{y_t} = 2 - \alpha.$$

(1) Put everything over the common denominator y_t: $\alpha = (2y_t - d)/y_t$.
(2) $\nu = 1/y_t$ and $2 - \eta = 2 - (2\Delta_\sigma + 2 - d) = d - 2\Delta_\sigma$.
(3) $\delta - 1 = (d - 2\Delta_\sigma)/\Delta_\sigma$ (Widom's relation).
(4) Hyperscaling (Josephson's relation), which involves d explicitly.

As a check, the Gaussian fixed point in $d = 4$ has $\Delta_\sigma = \Delta_\varphi = 1$ and $\Delta_\varepsilon = \Delta_{\varphi^2} = 2$, so $y_t = 2$, which gives the mean-field values $\nu = \frac{1}{2}$, $\eta = 2 + 2 - 4 = 0$, $\alpha = 2 - \frac{4}{2} = 0$, $\beta = \frac{1}{2}$, $\gamma = \frac{4-2}{2} = 1$, $\delta = \frac{4-1}{1} = 3$.

T5.3.3 The Wilson–Fisher fixed point at one loop

We now compute the flow of φ^4 theory perturbatively, in the way Wilson did it: integrate out a thin shell with a sharp cutoff, keeping all diagrams with one loop of shell modes. The action is

$$S = \int \mathrm{d}^d x \left[\tfrac{1}{2}(\partial\varphi)^2 + \tfrac{1}{2} m^2 \varphi^2 + \frac{\lambda}{4!}\varphi^4 \right], \qquad |p| < \Lambda. \tag{T5.51}$$

Split $\varphi = \varphi_< + \varphi_>$ with $\varphi_>$ in the shell $\Lambda e^{-\delta t} < |p| < \Lambda$. The shell modes are Gaussian with propagator $G_>(k) = 1/(k^2 + m^2)$ restricted to the shell (the mass is kept in the propagator; only the interaction is treated as a perturbation).

The shell integral. Every loop over a thin shell is the same kind of integral. Let S_{d-1} be the area of the unit sphere in $\mathbb{R}^d$. For any function f of $|k|$, use spherical coordinates, $\mathrm{d}^d k = k^{d-1}\mathrm{d}k\,\mathrm{d}\Omega$ with $\int \mathrm{d}\Omega = S_{d-1}$:

$$\int_{\text{shell}} \frac{\mathrm{d}^d k}{(2\pi)^d} f(|k|) \overset{(1)}{=} \frac{S_{d-1}}{(2\pi)^d} \int_{\Lambda e^{-\delta t}}^{\Lambda} \mathrm{d}k\; k^{d-1} f(k) \overset{(2)}{=} \frac{S_{d-1}}{(2\pi)^d}\, \Lambda^{d-1}\, (\Lambda\,\delta t)\, f(\Lambda)$$
$$= K_d\, \Lambda^d\, f(\Lambda)\, \delta t, \qquad K_d := \frac{S_{d-1}}{(2\pi)^d}. \tag{T5.52}$$

(1) The integrand does not depend on the direction, so the angular integral gives S_{d-1}.
(2) The shell is thin: the integrand is $\Lambda^{d-1} f(\Lambda)$ up to $O(\delta t)$, and its thickness is $\Lambda - \Lambda e^{-\delta t} = \Lambda\,\delta t + O(\delta t^2)$.

To find S_{d-1}, compute the Gaussian integral $\int \mathrm{d}^d k\, e^{-k^2}$ in two ways:

$$\pi^{d/2} \overset{(1)}{=} \int \mathrm{d}^d k\, e^{-k^2} \overset{(2)}{=} S_{d-1} \int_0^\infty \mathrm{d}k\, k^{d-1} e^{-k^2} \overset{(3)}{=} \tfrac{1}{2} S_{d-1}\, \Gamma(d/2),$$

$$\Longrightarrow \quad S_{d-1} = \frac{2\pi^{d/2}}{\Gamma(d/2)}, \qquad K_d \overset{(4)}{=} \frac{2}{(4\pi)^{d/2}\Gamma(d/2)}.$$

(1) Product of d one-dimensional integrals $\int \mathrm{d}k_i\, e^{-k_i^2} = \sqrt{\pi}$.
(2) Spherical coordinates.
(3) Substitute $s = k^2$, so $k^{d-1}\mathrm{d}k = \frac{1}{2}s^{d/2-1}\mathrm{d}s$, and use $\int_0^\infty s^{d/2-1}e^{-s}\mathrm{d}s = \Gamma(d/2)$.
(4) Divide S_{d-1} by $(2\pi)^d = 2^d\pi^d$, and $2^d\pi^{d/2} = (4\pi)^{d/2}$.

So $S_1 = 2\pi$, $S_2 = 4\pi$, $S_3 = 2\pi^2$ (with $\Gamma(1) = 1$, $\Gamma(\frac{3}{2}) = \frac{\sqrt{\pi}}{2}$, $\Gamma(2) = 1$), and in $d = 4$, $K_4 = 2\pi^2/(2\pi)^4 = 1/(8\pi^2)$.

The effective action to second order. Write $S = S_0[\varphi_<] + S_0[\varphi_>] + S_{\mathrm{int}}[\varphi_< + \varphi_>]$, where S_0 is the Gaussian part (it does not couple slow and fast modes, since they live at different momenta). Then

$$e^{-S_{\mathrm{eff}}[\varphi_<]} \overset{(1)}{=} e^{-S_0[\varphi_<]}\big\langle e^{-S_{\mathrm{int}}[\varphi_< + \varphi_>]}\big\rangle_>$$
$$S_{\mathrm{eff}} \overset{(2)}{=} S_0[\varphi_<] + \langle S_{\mathrm{int}}\rangle_> - \tfrac{1}{2}\Big(\langle S_{\mathrm{int}}^2\rangle_> - \langle S_{\mathrm{int}}\rangle_>^2\Big) + \cdots \tag{T5.53}$$

(1) Definition (T5.4), with the (normalised) Gaussian average $\langle\cdot\rangle_>$ over the shell modes.
(2) The **cumulant expansion**: expand $\ln\langle e^{-X}\rangle = \ln\big(1 - \langle X\rangle + \frac{1}{2}\langle X^2\rangle - \cdots\big) = -\langle X\rangle + \frac{1}{2}(\langle X^2\rangle - \langle X\rangle^2) + O(X^3)$ using $\ln(1+\epsilon) = \epsilon - \frac{1}{2}\epsilon^2 + \cdots$.

The variance $\langle S^2\rangle - \langle S\rangle^2$ keeps only the terms in which the two copies of S_{int} are connected by at least one shell propagator: the *connected* diagrams, as in (T5.23).

Worked example T5.13: One-loop shell integration of φ^4

Mass. In $\langle S_{\mathrm{int}}\rangle_>$ expand $(\varphi_< + \varphi_>)^4$ and keep the term with two slow and two fast fields (odd numbers of fast fields average to zero; the term with four fast fields is a constant):

$$\frac{\lambda}{4!}\int \mathrm{d}^dx \binom{4}{2}\varphi_<^2\langle\varphi_>^2\rangle \overset{(1)}{=} \frac{\lambda}{4}\int \mathrm{d}^dx\,\varphi_<^2 \int_{\mathrm{shell}} \frac{\mathrm{d}^dk}{(2\pi)^d}\,\frac{1}{k^2+m^2} \overset{(2)}{=} \frac{1}{2}\Big[\frac{\lambda}{2}\,\frac{K_d\Lambda^d}{\Lambda^2+m^2}\,\delta t\Big]\int \mathrm{d}^dx\,\varphi_<^2.$$

(1) $\binom{4}{2} = 6$ ways to choose the two fast fields; $\langle\varphi_>(x)^2\rangle$ is the shell propagator at coincident points, the integral of $G_>(k)$ over the shell.
(2) (T5.52); written as $\frac{1}{2}\,\delta m^2\int\varphi^2$.

So $\delta m^2 = \frac{\lambda}{2}K_d\Lambda^d\delta t/(\Lambda^2+m^2)$: the tadpole of Example T5.5. Integrating out adds positive mass.

Quartic coupling. In the variance, take from each copy of S_{int} the term with two slow and two fast fields, and contract the two fast fields of one vertex with the two of the other (a loop of two shell propagators). The number of ways is $\binom{4}{2}\cdot\binom{4}{2}\cdot 2 = 72$ (choose the fast pair at each vertex, then two ways to pair them across). So

$$-\tfrac{1}{2}\big(\langle S_{\mathrm{int}}^2\rangle - \langle S_{\mathrm{int}}\rangle^2\big) \overset{(1)}{=} -\frac{1}{2}\Big(\frac{\lambda}{4!}\Big)^2\cdot 72\int_{x,y}\varphi_<^2(x)\,G_>(x-y)^2\,\varphi_<^2(y)$$
$$\overset{(2)}{=} -\frac{\lambda^2}{16}\Big(\int_{\mathrm{shell}}\frac{\mathrm{d}^dk}{(2\pi)^d}\,\frac{1}{(k^2+m^2)^2}\Big)\int\mathrm{d}^dx\,\varphi_<^4 + \cdots$$
$$\overset{(3)}{=} \frac{1}{4!}\Big[-\frac{3\lambda^2}{2}\,\frac{K_d\Lambda^d}{(\Lambda^2+m^2)^2}\,\delta t\Big]\int\mathrm{d}^dx\,\varphi_<^4.$$

(1) Wick's theorem for the shell modes; each contraction of a fast field at x with one at y gives $G_>(x-y)$.
(2) $\frac{1}{2}\cdot\frac{72}{576} = \frac{1}{16}$. The slow fields vary slowly on the scale $1/\Lambda$ over which $G_>(x-$

$y)^2$ is concentrated, so set $\varphi_<^2(y) \approx \varphi_<^2(x)$ (the dots are derivative terms). Then $\int \mathrm{d}^d z\, G_>(z)^2 = \int_k G_>(k)G_>(-k)$ by Parseval; both k and $-k$ lie in the shell.

(3) (T5.52), and $\frac{24}{16} = \frac{3}{2}$.

So $\delta\lambda = -\frac{3}{2}\lambda^2 K_d \Lambda^d \delta t/(\Lambda^2 + m^2)^2$. There are three more second-order terms, and none of them contributes: (i) the term with one fast field at each vertex (a single shell line between two vertices) needs the shell momentum to equal the sum of the three slow momenta at one vertex; for small external momenta this sum is far below Λ, not in the shell, so the term vanishes (at larger momenta it builds a six-leg vertex, which is irrelevant); (ii) the terms with three fast fields at a vertex give higher-order corrections to the mass; (iii) the kinetic term $(\partial\varphi)^2$ is not corrected at one loop, because the tadpole does not depend on the external momentum. Point (iii) means *the anomalous dimension vanishes at one loop*: $\eta = O(\lambda^2)$.

Rescaling. After the integration the slow field lives on $|p| < \Lambda' := \Lambda e^{-\delta t}$, with couplings $m^2 + \delta m^2$ and $\lambda + \delta\lambda$, and the kinetic term still has coefficient $\frac{1}{2}$ (point (iii)). Rescaling momenta by $e^{\delta t}$ so that the cutoff is again called Λ is the same as measuring every coupling in units of the new cutoff Λ'. So use the dimensionless couplings

$$r := \frac{m^2}{\Lambda^2}, \qquad \bar{u} := K_d\, \lambda\, \Lambda^{d-4},$$

and compare them before and after the step. With $\Lambda'^{-2} = \Lambda^{-2}(1+2\delta t)$ and $\Lambda'^{d-4} = \Lambda^{d-4}(1+\varepsilon\,\delta t)$ to first order, where $\varepsilon := 4 - d$,

$$\begin{aligned} r' - r &\overset{(1)}{=} \frac{m^2 + \delta m^2}{\Lambda^2}(1 + 2\delta t) - \frac{m^2}{\Lambda^2} \overset{(2)}{=} 2r\,\delta t + \frac{\lambda}{2}\,\frac{K_d \Lambda^{d-4}}{1+r}\,\delta t \\ &\overset{(3)}{=} \Big(2r + \frac{\bar{u}}{2(1+r)}\Big)\delta t, \\ \bar{u}' - \bar{u} &\overset{(4)}{=} K_d(\lambda + \delta\lambda)\Lambda^{d-4}(1 + \varepsilon\,\delta t) - K_d\lambda\Lambda^{d-4} \overset{(5)}{=} \varepsilon\,\bar{u}\,\delta t - \frac{3}{2}\,\frac{K_d^2\lambda^2\Lambda^{2d-8}}{(1+r)^2}\,\delta t \\ &\overset{(6)}{=} \Big(\varepsilon\,\bar{u} - \frac{3}{2}\,\frac{\bar{u}^2}{(1+r)^2}\Big)\delta t. \end{aligned}$$

(1) Definition of r at the new cutoff, $r' = (m^2 + \delta m^2)/\Lambda'^2$.
(2) Keep first order in δt; insert δm^2 and write $\Lambda^d/(\Lambda^2(\Lambda^2 + m^2)) = \Lambda^{d-4}/(1 + r)$.
(3) $K_d\lambda\Lambda^{d-4} = \bar{u}$.
(4) Definition of $\bar{u}$ at the new cutoff.
(5) First order in δt; insert $\delta\lambda$ and write $K_d\Lambda^{d-4} \cdot K_d\Lambda^d/(\Lambda^2 + m^2)^2 = K_d^2\Lambda^{2d-8}/(1 + r)^2$.
(6) $K_d^2\lambda^2\Lambda^{2d-8} = \bar{u}^2$.

The first terms, $2r$ and $\varepsilon\bar{u}$, are the canonical factors $e^{[g]\delta t}$ of (T5.6). Dividing by δt,

$$\frac{\mathrm{d}r}{\mathrm{d}t} = 2r + \frac{\bar{u}}{2(1+r)}, \qquad \frac{\mathrm{d}\bar{u}}{\mathrm{d}t} = \varepsilon\,\bar{u} - \frac{3}{2}\,\frac{\bar{u}^2}{(1+r)^2}, \qquad \varepsilon = 4 - d. \tag{T5.54}$$

In $d = 4$ ($\varepsilon = 0$) and at $r = 0$, $\bar{u} = K_4\lambda = \lambda/8\pi^2$, and the second equation becomes

$$\frac{1}{8\pi^2}\,\frac{\mathrm{d}\lambda}{\mathrm{d}t} \overset{(1)}{=} -\frac{3}{2}\,\frac{\lambda^2}{(8\pi^2)^2} \quad\Longrightarrow\quad \frac{\mathrm{d}\lambda}{\mathrm{d}t} \overset{(2)}{=} -\frac{3\lambda^2}{16\pi^2},$$

(1) Insert $\bar{u} = \lambda/8\pi^2$.

(2) Multiply by $8\pi^2$: $\frac{3}{2}\cdot\frac{1}{8\pi^2}=\frac{3}{16\pi^2}$.

the familiar one-loop β-function of φ^4 theory ([QFT notes (15.7.56)]), with the sign convention of the pitfall in Section T5.1.4. In Example T5.12, $b=3/16\pi^2$. For $d=4-\varepsilon$ the same substitution, with $K_d=K_4+O(\varepsilon)$, gives $\mathrm{d}\hat\lambda/\mathrm{d}t=\varepsilon\hat\lambda-3\hat\lambda^2/16\pi^2$ at leading order for the dimensionless coupling $\hat\lambda:=\lambda\Lambda^{d-4}$.

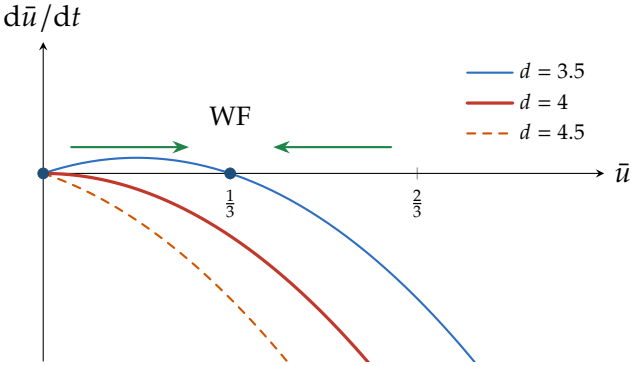


Figure T5.12: The one-loop flow of the quartic coupling at $r=0$, $\mathrm{d}\bar u/\mathrm{d}t=\varepsilon\bar u-\frac{3}{2}\bar u^2$, for $d=3.5$ ($\varepsilon=\frac12$, blue), $d=4$ (red) and $d=4.5$ (orange, dashed). Green arrows: the direction of flow towards the infrared for $d=3.5$. What to see: below four dimensions a second zero appears at $\bar u^*=\frac{2}{3}\varepsilon$ (here $\frac13$), and it attracts the flow from both sides; this is the Wilson–Fisher point. At $d=4$ it merges with the Gaussian point, and the coupling decays logarithmically (Example T5.12). Chapter 11 draws the same curve with the opposite sign convention.

Worked example T5.14: Fixed point and exponents to first order in ε

Setting the right sides of (T5.54) to zero with $\bar u\neq0$, at leading order in ε:

$$\bar u^*\overset{(1)}{=}\tfrac{2}{3}\varepsilon,\qquad r^*\overset{(2)}{=}-\tfrac14\bar u^*=-\tfrac{\varepsilon}{6}.$$

(1) $\varepsilon=\frac{3}{2}\bar u^*/(1+r^*)^2$ and $(1+r^*)^2=1+O(\varepsilon)$.
(2) $2r^*+\frac12\bar u^*=0$ at leading order.

The stability matrix at this point is

$$M\overset{(1)}{=}\begin{pmatrix}2-\dfrac{\bar u^*}{2(1+r^*)^2} & \dfrac{1}{2(1+r^*)}\\ \dfrac{3\bar u^{*2}}{(1+r^*)^3} & \varepsilon-\dfrac{3\bar u^*}{(1+r^*)^2}\end{pmatrix}\overset{(2)}{=}\begin{pmatrix}2-\dfrac{\varepsilon}{3} & \dfrac12\\ O(\varepsilon^2) & -\varepsilon\end{pmatrix}+O(\varepsilon^2).$$

(1) Differentiate (T5.54) with respect to r and $\bar u$.
(2) Insert $\bar u^*=\frac23\varepsilon$: $\frac12\bar u^*=\frac{\varepsilon}{3}$ and $\varepsilon-3\bar u^*=\varepsilon-2\varepsilon=-\varepsilon$.

The matrix is triangular up to $O(\varepsilon^2)$. Write it as $\left(\begin{smallmatrix}a & b\\ c & e\end{smallmatrix}\right)$ with $a=2-\varepsilon/3$, $b=\frac12$, $c=O(\varepsilon^2)$, $e=-\varepsilon$. Its characteristic polynomial is

$$(a-y)(e-y)-bc\overset{(1)}{=}(a-y)(e-y)+O(\varepsilon^2),$$

(1) $bc=\frac12\cdot O(\varepsilon^2)$.

so its roots are a and e up to corrections $bc/(a-e)=O(\varepsilon^2)$:

$$y_t=2-\frac{\varepsilon}{3},\qquad y_{\mathrm{irr}}=-\varepsilon.$$

These agree with the LPA result (T5.31), although the fixed point sits elsewhere ($\bar{u}^* = \frac{2}{3}\varepsilon$ here, which is $\tilde{u}^* = \frac{\varepsilon}{3}$ in the LPA units since $\bar{u} = K_4\lambda = 2\alpha_4\lambda$; the LPA had $\tilde{u}^* = \frac{\varepsilon}{6}$). Now use Section T5.3.2:

$$\nu \overset{(1)}{=} \frac{1}{y_t} = \frac{1}{2-\varepsilon/3} \overset{(2)}{=} \frac{1}{2} + \frac{\varepsilon}{12} + O(\varepsilon^2),$$

$$\eta \overset{(3)}{=} 0 + O(\varepsilon^2), \qquad \Delta_\varphi = \frac{d-2+\eta}{2} = 1 - \frac{\varepsilon}{2} + O(\varepsilon^2), \qquad \Delta_{\varphi^2} = d - y_t = 2 - \frac{2\varepsilon}{3} + O(\varepsilon^2).$$

(1) The relation $\nu = 1/y_t$ derived above.
(2) $\frac{1}{2}(1-\frac{\varepsilon}{6})^{-1} = \frac{1}{2}(1+\frac{\varepsilon}{6}+\cdots)$.
(3) No field renormalisation at one loop (Example T5.13, point (iii)).

The first non-zero anomalous dimension comes from a two-loop diagram, $\eta = \varepsilon^2/54$; Example T5.15 below computes it. The other exponents follow from the key idea of Section T5.3.2 with $\Delta_\sigma = \Delta_\varphi = 1 - \frac{\varepsilon}{2}$, $\Delta_\varepsilon = \Delta_{\varphi^2}$ and $y_t = d - \Delta_\varepsilon = 2 - \frac{\varepsilon}{3}$, expanding each ratio to first order with $(2-\frac{\varepsilon}{3})^{-1} = \frac{1}{2}(1+\frac{\varepsilon}{6}) + O(\varepsilon^2)$:

$$\alpha \overset{(1)}{=} 2 - \frac{4-\varepsilon}{2-\varepsilon/3} = 2 - \left(2-\tfrac{\varepsilon}{2}\right)\left(1+\tfrac{\varepsilon}{6}\right) = \frac{\varepsilon}{6}, \qquad \beta \overset{(2)}{=} \frac{1-\varepsilon/2}{2-\varepsilon/3} = \tfrac{1}{2}\left(1-\tfrac{\varepsilon}{2}\right)\left(1+\tfrac{\varepsilon}{6}\right) = \frac{1}{2} - \frac{\varepsilon}{6},$$

$$\gamma \overset{(3)}{=} \frac{4-\varepsilon-2+\varepsilon}{2-\varepsilon/3} = \frac{2}{2-\varepsilon/3} = 1 + \frac{\varepsilon}{6}, \qquad \delta \overset{(4)}{=} \frac{3-\varepsilon/2}{1-\varepsilon/2} = \left(3-\tfrac{\varepsilon}{2}\right)\left(1+\tfrac{\varepsilon}{2}\right) = 3 + \varepsilon,$$

(1) $\alpha = 2 - d/y_t$; drop ε^2: $2 - \frac{\varepsilon}{2} + \frac{\varepsilon}{3} = 2 - \frac{\varepsilon}{6}$.
(2) $\beta = \Delta_\sigma/y_t$; $\frac{1}{2}(1 - \frac{\varepsilon}{2} + \frac{\varepsilon}{6}) = \frac{1}{2} - \frac{\varepsilon}{6}$.
(3) $\gamma = (d - 2\Delta_\sigma)/y_t$ with $2\Delta_\sigma = 2 - \varepsilon$.
(4) $\delta = (d - \Delta_\sigma)/\Delta_\sigma$ and $(1 - \frac{\varepsilon}{2})^{-1} = 1 + \frac{\varepsilon}{2} + \cdots$; then $\frac{3\varepsilon}{2} - \frac{\varepsilon}{2} = \varepsilon$.

all to $O(\varepsilon^2)$. At $\varepsilon = 0$ these are the mean-field values.

Worked example T5.15: The sunset diagram gives $\eta = \varepsilon^2/54$

The field is not renormalised at one loop, so η starts at order λ^2. We compute it at the massless critical point, in **dimensional regularisation** [do the loop integrals in $d = 4 - \varepsilon$ dimensions, where they are finite, and read off the divergences as poles in $1/\varepsilon$]. Exponents are scheme independent (Section T5.2.5, last pitfall), so any scheme will do. To keep the coupling dimensionless write $\lambda = \mu^\varepsilon g$, with μ an arbitrary mass scale. The one-loop flow derived above, rewritten in the convention of the pitfall in Section T5.1.4, is $\mu\,\mathrm{d}g/\mathrm{d}\mu = -\varepsilon g + 3g^2/16\pi^2$, with fixed point $g^* = 16\pi^2\varepsilon/3$.
(a) The diagram. At order λ^2 the correction to $\langle\varphi\varphi\rangle$ that depends on the external momentum p is the **sunset** diagram: the two external legs attach to two vertices, which are joined by three propagators. Its weight is

$$\frac{1}{2!}\left(\frac{\lambda}{4!}\right)^2 \cdot 2\cdot 4\cdot 4\cdot 3! \overset{(1)}{=} \frac{\lambda^2\cdot 192}{2\cdot 576} \overset{(2)}{=} \frac{\lambda^2}{6}.$$

(1) Second order of $e^{-S_{\text{int}}}$ gives $\frac{1}{2!}\langle S_{\text{int}}^2\rangle$ (the two minus signs cancel). Choose which vertex the first external leg joins (2), which of its four legs (4), which leg of the other vertex the second external leg joins (4), and pair the remaining three legs at one vertex with the three at the other (3!).
(2) Arithmetic.

So the self-energy is $\Sigma(p) = \frac{\lambda^2}{6} I(p)$ with

$$I(p) \coloneqq \int \frac{\mathrm{d}^d k}{(2\pi)^d} \frac{\mathrm{d}^d q}{(2\pi)^d} \frac{1}{k^2\, q^2\, (p-k-q)^2},$$

and, since the correction adds to the free propagator as $G_0 + G_0 \Sigma G_0 + \cdots = 1/(p^2 - \Sigma)$, the inverse propagator is $\Gamma^{(2)}(p) = p^2 - \Sigma(p)$.
(b) The key integral. Every one-loop integral we need is of the form

$$J(a,b) \coloneqq \int \frac{\mathrm{d}^d k}{(2\pi)^d} \frac{1}{(k^2)^a \left((p-k)^2\right)^b}.$$

With Schwinger parameters,

$$\begin{aligned}
J(a,b) &\overset{(1)}{=} \frac{1}{\Gamma(a)\Gamma(b)} \int_0^\infty \mathrm{d}s\, \mathrm{d}t\; s^{a-1} t^{b-1} \int \frac{\mathrm{d}^d k}{(2\pi)^d}\, e^{-sk^2 - t(p-k)^2} \\
&\overset{(2)}{=} \frac{1}{\Gamma(a)\Gamma(b)} \int_0^\infty \mathrm{d}s\, \mathrm{d}t\; \frac{s^{a-1} t^{b-1}}{(4\pi(s+t))^{d/2}}\, e^{-\frac{st}{s+t} p^2} \\
&\overset{(3)}{=} \frac{1}{\Gamma(a)\Gamma(b)(4\pi)^{d/2}} \int_0^1 \mathrm{d}x\, x^{a-1}(1-x)^{b-1} \int_0^\infty \mathrm{d}\rho\; \rho^{a+b-1-d/2} e^{-\rho x(1-x)p^2} \\
&\overset{(4)}{=} \frac{\Gamma(a+b-\frac{d}{2})}{\Gamma(a)\Gamma(b)(4\pi)^{d/2}}\, (p^2)^{\frac{d}{2}-a-b} \int_0^1 \mathrm{d}x\, x^{\frac{d}{2}-b-1}(1-x)^{\frac{d}{2}-a-1} \\
&\overset{(5)}{=} \frac{(p^2)^{\frac{d}{2}-a-b}}{(4\pi)^{d/2}}\, \frac{\Gamma(a+b-\frac{d}{2})\,\Gamma(\frac{d}{2}-a)\,\Gamma(\frac{d}{2}-b)}{\Gamma(a)\,\Gamma(b)\,\Gamma(d-a-b)}.
\end{aligned}$$

(1) $A^{-a} = \frac{1}{\Gamma(a)} \int_0^\infty \mathrm{d}s\, s^{a-1} e^{-sA}$ for $A > 0$, once for each denominator.
(2) Complete the square, $sk^2 + t(p-k)^2 = (s+t)\big(k - \frac{t}{s+t} p\big)^2 + \frac{st}{s+t} p^2$, and use $\int \frac{\mathrm{d}^d k}{(2\pi)^d} e^{-ck^2} = (4\pi c)^{-d/2}$ (Section T5.2.5).
(3) $s = \rho x$, $t = \rho(1-x)$ with $\mathrm{d}s\, \mathrm{d}t = \rho\, \mathrm{d}\rho\, \mathrm{d}x$.
(4) $\int_0^\infty \mathrm{d}\rho\, \rho^{z-1} e^{-\rho c} = \Gamma(z) c^{-z}$ with $z = a + b - \frac{d}{2}$, $c = x(1-x)p^2$; collect the powers of x and $1-x$.
(5) The Beta integral $\int_0^1 x^{\alpha-1}(1-x)^{\beta-1} \mathrm{d}x = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$, with $\alpha + \beta = d - a - b$.

(c) The sunset as two nested loops. Do the q integral first, at fixed $P \coloneqq p - k$: it is $J(1,1)$ with external momentum P. The result is a power of P^2, so the k integral is again of the form $J(a,b)$:

$$I(p) \overset{(1)}{=} \frac{\Gamma(2-\frac{d}{2})\Gamma(\frac{d}{2}-1)^2}{(4\pi)^{d/2}\Gamma(d-2)} \int \frac{\mathrm{d}^d k}{(2\pi)^d} \frac{1}{k^2\, ((p-k)^2)^{2-d/2}} \overset{(2)}{=} \frac{(p^2)^{d-3}}{(4\pi)^d} \frac{\Gamma(\frac{d}{2}-1)^3\, \Gamma(3-d)}{\Gamma(\frac{3d}{2}-3)}.$$

(1) $J(1,1) = \frac{(P^2)^{d/2-2}}{(4\pi)^{d/2}} \frac{\Gamma(2-\frac{d}{2})\Gamma(\frac{d}{2}-1)^2}{\Gamma(d-2)}$.
(2) $J(1, 2-\frac{d}{2})$: here $a + b - \frac{d}{2} = 3 - d$, $\frac{d}{2} - a = \frac{d}{2} - 1$, $\frac{d}{2} - b = d - 2$, $d - a - b = \frac{3d}{2} - 3$, and the power is $(p^2)^{d-3}$; the factors $\Gamma(2 - \frac{d}{2})$ and $\Gamma(d-2)$ cancel.

(d) The pole. Put $d = 4 - \varepsilon$. Then $\Gamma(\frac{d}{2}-1) = \Gamma(1-\frac{\varepsilon}{2}) = 1 + O(\varepsilon)$, $\Gamma(\frac{3d}{2}-3) = \Gamma(3-\frac{3\varepsilon}{2}) = 2 + O(\varepsilon)$, and $\Gamma(3-d) = \Gamma(-1+\varepsilon) = \frac{\Gamma(\varepsilon)}{-1+\varepsilon} = -\frac{1}{\varepsilon} + O(1)$ (from $\Gamma(z+1) = z\Gamma(z)$ and $\Gamma(\varepsilon) = \frac{1}{\varepsilon} + O(1)$). So

$$I(p) = -\frac{p^2}{2\varepsilon\,(4\pi)^4} + (\text{finite}) \cdot p^2.$$

The divergence multiplies p^2: the sunset renormalises the kinetic term, which the one-loop tadpole could not do.
(e) The anomalous dimension. Insert into $\Gamma^{(2)}$, with $\lambda^2 = \mu^{2\varepsilon} g^2$ (the factor $\mu^{2\varepsilon}(p^2)^{-\varepsilon}$ is finite and only changes the finite part):

$$\Gamma^{(2)}(p) \overset{(1)}{=} p^2 - \frac{g^2}{6} I(p) \overset{(2)}{=} p^2\left(1 + \frac{g^2}{12\varepsilon(4\pi)^4}\right) + (\text{finite}) \cdot p^2.$$

(1) $\Gamma^{(2)} = p^2 - \Sigma$ with $\Sigma = \frac{\lambda^2}{6} I$.
(2) $-\frac{1}{6} \cdot \left(-\frac{1}{2\varepsilon}\right) = \frac{1}{12\varepsilon}$.

A rescaled field $\varphi = Z^{1/2}\varphi_R$ has $\Gamma^{(2)}_R = Z\,\Gamma^{(2)}$, which is finite if $Z = 1 - g^2/(12\varepsilon(4\pi)^4)$. The **anomalous dimension** of the field is $\gamma_\varphi := \frac{1}{2}\mu\,\mathrm{d}\ln Z/\mathrm{d}\mu$ at fixed bare coupling. At a fixed point the Callan–Symanzik equation (Chapter 11) makes the propagator a power, $\langle\varphi(x)\varphi(0)\rangle \propto |x|^{-(d-2)-2\gamma_\varphi}$, so $\Delta_\varphi = \frac{d-2}{2} + \gamma_\varphi$ and, comparing with $2\Delta_\varphi = d - 2 + \eta$, $\eta = 2\gamma_\varphi(g^*)$. Now

$$\gamma_\varphi \overset{(1)}{=} \frac{1}{2} \cdot \left(-\frac{2g}{12\varepsilon(4\pi)^4}\right)\mu\frac{\mathrm{d}g}{\mathrm{d}\mu} \overset{(2)}{=} \frac{1}{2} \cdot \left(-\frac{2g}{12\varepsilon(4\pi)^4}\right)(-\varepsilon g) = \frac{g^2}{12(4\pi)^4},$$
$$\eta \overset{(3)}{=} 2\gamma_\varphi(g^*) \overset{(4)}{=} \frac{2}{12(4\pi)^4}\left(\frac{(4\pi)^2\varepsilon}{3}\right)^2 = \frac{2\varepsilon^2}{108} = \frac{\varepsilon^2}{54}.$$

(1) $\ln Z = -g^2/(12\varepsilon(4\pi)^4) + O(g^4)$; differentiate with the chain rule.
(2) $\mu\,\mathrm{d}g/\mathrm{d}\mu = -\varepsilon g + O(g^2)$; the $O(g^2)$ term gives $O(g^3)$ in γ_φ. The $1/\varepsilon$ cancels: the anomalous dimension is finite.
(3) The relation just derived.
(4) $g^* = 16\pi^2\varepsilon/3 = (4\pi)^2\varepsilon/3$.

For the N-component model the same steps give $\eta = (N+2)\varepsilon^2/(2(N+8)^2)$ ([QFT notes (16.2.81)–(16.2.82)]), which is $3\varepsilon^2/(2\cdot 81) = \varepsilon^2/54$ at $N = 1$.

Watch out: η comes from the field, not from φ^2

Two anomalous dimensions appear at the Wilson–Fisher point and are easily confused. The field anomalous dimension γ_φ gives $\eta = 2\gamma_\varphi(g^*)$, which starts at $O(\varepsilon^2)$. The anomalous dimension γ_m of the operator φ^2 is already $O(\varepsilon)$: from $\Delta_{\varphi^2} = d - 2 + \gamma_m$ and $\Delta_{\varphi^2} = 2 - \frac{2\varepsilon}{3}$ we get $\gamma_m = \frac{\varepsilon}{3}$, and it controls $\nu = 1/y_t = (2 - \gamma_m)^{-1}$, not η.

T5.3.4 How good is the ε-expansion, and where the CFTs come in

Set $\varepsilon = 1$ and $\varepsilon = 2$ in the first-order formulas and compare with what is known in $d = 3$ (from the conformal bootstrap and Monte Carlo) and in $d = 2$ (Onsager's solution, and the $c = \frac{1}{2}$ CFT of Chapter 11).

	ν	η	α	β	γ	δ
mean field ($d \geq 4$)	0.5	0	0	0.5	1	3
$O(\varepsilon)$ at $\varepsilon = 1$	0.583	0	0.167	0.333	1.167	4
3D Ising (bootstrap)	0.630	0.036	0.110	0.326	1.237	4.79
$O(\varepsilon)$ at $\varepsilon = 2$	0.667	0	0.333	0.167	1.333	5
2D Ising (exact)	1	0.25	0	0.125	1.75	15

In three dimensions first order is already within about 10%; higher orders, resummed, give three or four digits. In two dimensions the expansion is hopeless ($\delta = 5$ against 15). That is not a failure of the RG but of perturbation theory about the Gaussian point: at $\varepsilon = 2$ the fixed point is far from free. The exact answer comes from the other end: the fixed point is a CFT, and in two dimensions CFTs can be classified. In Chapter 11 we will find that the two-dimensional Ising critical point is the "minimal model" $\mathcal{M}_3$, the CFT with central charge $c = \frac{1}{2}$ whose two non-trivial primary operators σ and ε have conformal weights $h = \bar{h} = \frac{1}{16}$ and $\frac{1}{2}$, so $\Delta_\sigma = \frac{1}{8}$ and $\Delta_\varepsilon = 1$. The RG eigenvalues $y_t = 2 - 1 = 1$ and $y_h = 2 - \frac{1}{8} = \frac{15}{8}$ then reproduce the last line of the table through Section T5.3.2. More generally, Zamolodchikov identified the multicritical fixed point of $\varphi^{2(m-1)}$ theory in two dimensions with the minimal model $\mathcal{M}_m$ of central charge $c_m = 1 - 6/(m(m+1))$: $\varphi^4 \leftrightarrow$ Ising ($m = 3$), $\varphi^6 \leftrightarrow$ tricritical Ising ($m = 4$), and so on. This identification drives the flows between minimal models in the c-theorem section of Chapter 11 (Section 11.13.4).

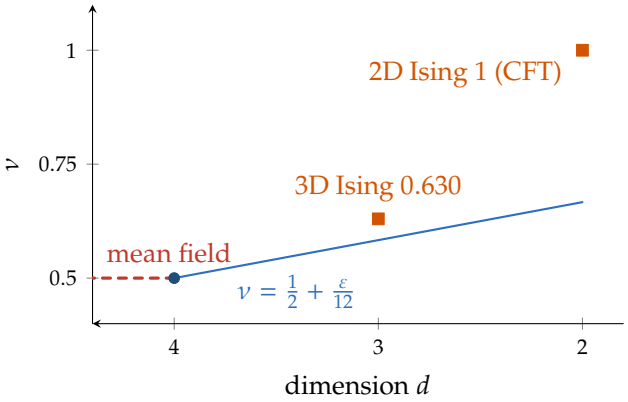


Figure T5.13: The correlation-length exponent against dimension (read from right to left, starting at $d = 4$). Blue: the first-order ε-expansion. Orange squares: the known values in $d = 3$ and $d = 2$. What to see: the expansion about the Gaussian point is good near $d = 4$ and poor at $d = 2$, where the exact value comes instead from the Ising CFT of Chapter 11.

Recap

- Linearising $\mathrm{d}g/\mathrm{d}t = \beta(g)$ at a fixed point gives a stability matrix M; its eigenvectors define scaling fields $u_a \propto e^{y_a t}$.
- RG eigenvalues and scaling dimensions are the same data: $y_a = d - \Delta_a$; $\nu = 1/y_t$.
- At the Gaussian point the eigenoperators are the normal-ordered $:\varphi^n:$ (Hermite polynomials) with $y_n = d - n\frac{d-2}{2}$.
- One-loop shell integration of φ^4 gives $\mathrm{d}r/\mathrm{d}t = 2r + \bar{u}/2(1+r)$, $\mathrm{d}\bar{u}/\mathrm{d}t = \varepsilon\bar{u} - \frac{3}{2}\bar{u}^2/(1+r)^2$; the Wilson–Fisher point $\bar{u}^* = \frac{2}{3}\varepsilon$ has $y_t = 2 - \varepsilon/3$, so $\nu = \frac{1}{2} + \frac{\varepsilon}{12}$, $\eta = \varepsilon^2/54$.

- Fixed-point locations are scheme dependent; eigenvalues are not.
- Scaling of the free energy turns (y_t, y_h) into all six exponents.
- In $d = 2$ the Wilson–Fisher point is the Ising CFT $\mathcal{M}_3$, whose exact weights give the exact exponents (Chapter 11).

T5.4 Running couplings in perturbation theory

Wilson's RG removes shells of momenta below a cutoff Λ and follows the couplings as Λ is lowered. A particle physicist usually works differently. The cutoff is removed (or replaced by dimensional regularisation), and the coupling is *defined* by a measurement at some chosen momentum scale μ. Nothing physical can depend on that choice. This section asks what that requirement says about correlation functions, and it computes the first coefficient of the β-function in the two theories that matter most in this book: φ^4 theory and Yang–Mills theory. The answers are the Callan–Symanzik equation (Section T5.4.1), $\beta = 3\lambda^2/16\pi^2$ (Section T5.4.2) and $\beta = -(g^3/16\pi^2)(\frac{11N}{3} - \frac{2n_f}{3})$ (Section T5.4.3).

Throughout, $\beta(g) \coloneqq \mu\, \mathrm{d}g/\mathrm{d}\mu$ with μ increasing towards the ultraviolet. This is minus the $\mathrm{d}g/\mathrm{d}t$ of (T5.5) (the pitfall in Section T5.1.4).

T5.4.1 The Callan–Symanzik equation

[Callan 1970, Symanzik 1970] Take massless φ^4 theory in $d = 4 - \varepsilon$ Euclidean dimensions, with the **bare** action

$$S = \int \mathrm{d}^d x \left[\tfrac{1}{2}(\partial\varphi_0)^2 + \frac{\lambda_0}{4!}\varphi_0^4 \right].$$

The field φ_0 and coupling λ_0 are the ones that appear in the path integral. The **bare correlators** $G_0^{(n)}(x_1, \dots, x_n) = \langle \varphi_0(x_1) \cdots \varphi_0(x_n) \rangle$ depend on λ_0 and on ε (which plays the role of the cutoff), and on nothing else. Their poles at $\varepsilon \to 0$ are removed by rescaling the field and the coupling.

Working definition T5.7 (renormalised coupling and anomalous dimension)

Pick a **renormalisation scale** μ [the momentum at which we agree to call the strength of the interaction λ] and write

$$\varphi_0 = Z^{1/2}\varphi, \qquad \lambda_0 = \mu^{\varepsilon} Z_\lambda\, \lambda, \tag{T5.55}$$

with Z and Z_λ chosen so that the **renormalised correlators** $G^{(n)} \coloneqq Z^{-n/2} G_0^{(n)}$ are finite as $\varepsilon \to 0$ [Z absorbs the part of the field's normalisation that the loops make infinite]. The factor μ^ε keeps λ dimensionless, since $[\lambda_0] = 4 - d = \varepsilon$. Holding λ_0 fixed and changing μ changes λ and Z. The two rates are the **β-function** and the **anomalous dimension**

$$\beta(\lambda) \coloneqq \mu \frac{\mathrm{d}\lambda}{\mathrm{d}\mu}\bigg|_{\lambda_0}, \qquad \gamma(\lambda) \coloneqq \frac{1}{2}\,\mu \frac{\mathrm{d}\ln Z}{\mathrm{d}\mu}\bigg|_{\lambda_0}. \tag{T5.56}$$

Here γ [how fast the field's normalisation must be readjusted as μ moves] is the anomalous dimension. *Operational test:* compute a finite quantity in terms of $\lambda(\mu)$, demand that it does not change with μ, and read off β (this is done in Example T5.18).

The renormalised correlators depend on μ and λ, but the bare ones do not know about μ at

all. That single fact gives a differential equation.

$$
\begin{aligned}
0 &\overset{(1)}{=} \mu \frac{\mathrm{d}}{\mathrm{d}\mu}\Big|_{\lambda_0} G_0^{(n)} \overset{(2)}{=} \mu \frac{\mathrm{d}}{\mathrm{d}\mu}\Big|_{\lambda_0} \left[Z^{n/2}\, G^{(n)}\big(x_i; \mu, \lambda(\mu)\big) \right] \\
&\overset{(3)}{=} Z^{n/2} \left[\frac{n}{2}\, \mu \frac{\mathrm{d} \ln Z}{\mathrm{d}\mu} + \mu \frac{\partial}{\partial \mu} + \mu \frac{\mathrm{d}\lambda}{\mathrm{d}\mu} \frac{\partial}{\partial \lambda} \right] G^{(n)} \\
&\overset{(4)}{=} Z^{n/2} \left[\mu \frac{\partial}{\partial \mu} + \beta(\lambda) \frac{\partial}{\partial \lambda} + n\, \gamma(\lambda) \right] G^{(n)}.
\end{aligned}
\tag{T5.57}
$$

(1) $G_0^{(n)}$ is a function of x_i, λ_0 and ε only.
(2) $G_0^{(n)} = Z^{n/2} G^{(n)}$, the definition of $G^{(n)}$.
(3) Product rule, and the chain rule for the μ-dependence that enters through $\lambda(\mu)$. The factor $Z^{n/2}$ gives $\mu\, \mathrm{d}Z^{n/2}/\mathrm{d}\mu = Z^{n/2}\, \frac{n}{2}\, \mu\, \mathrm{d} \ln Z/\mathrm{d}\mu$.
(4) The definitions (T5.56).

Dividing by $Z^{n/2}$ gives the result.

Key idea: the Callan–Symanzik equation

Renormalised correlators obey

$$
\left[\mu \frac{\partial}{\partial \mu} + \beta(\lambda) \frac{\partial}{\partial \lambda} + n\, \gamma(\lambda) \right] G^{(n)}(x_1, \ldots, x_n; \mu, \lambda) = 0. \tag{T5.58}
$$

A change of the arbitrary scale μ is compensated by a change of the coupling (rate β) and of the normalisation of each field (rate γ).

What does (T5.58) say physically? Combine it with dimensional analysis. The field has classical dimension $\Delta_{\mathrm{cl}} = \frac{d-2}{2}$ (Example T5.4), so the two-point function must have the form

$$
G^{(2)}(x; \mu, \lambda) = |x|^{-2\Delta_{\mathrm{cl}}}\, F(\mu\, |x|\,, \lambda), \tag{T5.59}
$$

with F a dimensionless function of the only two dimensionless numbers available. Write $r = |x|$ and $u = \mu r$. Then

$$
r \frac{\partial G^{(2)}}{\partial r} \overset{(1)}{=} -2\Delta_{\mathrm{cl}}\, G^{(2)} + r^{-2\Delta_{\mathrm{cl}}}\, u\, \partial_u F \overset{(2)}{=} -2\Delta_{\mathrm{cl}}\, G^{(2)} + \mu \frac{\partial G^{(2)}}{\partial \mu},
$$

(1) Product rule; $r\, \partial_r F(\mu r) = u\, \partial_u F$.
(2) $\mu\, \partial_\mu F(\mu r) = u\, \partial_u F$ as well.

so $\mu \partial_\mu = r \partial_r + 2\Delta_{\mathrm{cl}}$ on $G^{(2)}$, and (T5.58) becomes

$$
\left[r \frac{\partial}{\partial r} + 2\Delta_{\mathrm{cl}} + \beta(\lambda) \frac{\partial}{\partial \lambda} + 2\gamma(\lambda) \right] G^{(2)} = 0. \tag{T5.60}
$$

Read it as a statement about zooming out: moving the two points apart ($r\partial_r$) has the same effect as changing the coupling at the rate β and the normalisation at the rate γ. This is Wilson's picture again, now written for correlators.

Solving it: the method of characteristics. Equation (T5.58) is a first-order linear partial differential equation in the two variables (μ, λ). Such an equation is solved along the curves

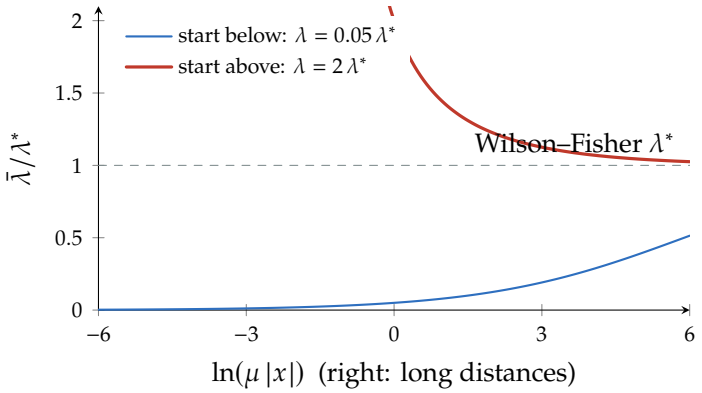


Figure T5.14: The running coupling $\bar\lambda$ of Example T5.17 in $d = 4 - \varepsilon$ ($\varepsilon = \frac{1}{2}$), as a function of the distance at which the two-point function is probed. Both curves are fixed by $\bar\lambda = \lambda$ at $\mu\,|x| = 1$. What to see: at long distances (right) every starting value is drawn to the Wilson–Fisher value λ^*, so the correlator becomes the pure power of the fixed point; at short distances (left) the coupling from below dies away (the theory is free in the ultraviolet), while the one from above grows and runs out of the region of validity of one loop.

on which its derivative part is an ordinary derivative. Define the **running coupling** $\bar\lambda(\ell)$ [the coupling we would have to use if we renormalised at the scale μe^{ℓ} instead of μ] by

$$\frac{\mathrm{d}\bar\lambda}{\mathrm{d}\ell} = \beta(\bar\lambda), \qquad \bar\lambda(0) = \lambda, \tag{T5.61}$$

and follow the curve $\ell \mapsto (\mu e^{\ell}, \bar\lambda(\ell))$ in the (μ, λ) plane. Along it,

$$\frac{\mathrm{d}}{\mathrm{d}\ell}\, G^{(2)}\big(x; \mu e^{\ell}, \bar\lambda(\ell)\big) \overset{(1)}{=} \Big[\mu' \frac{\partial}{\partial \mu'} + \frac{\mathrm{d}\bar\lambda}{\mathrm{d}\ell}\,\frac{\partial}{\partial \lambda}\Big] G^{(2)} \overset{(2)}{=} -2\gamma\big(\bar\lambda(\ell)\big)\, G^{(2)},$$

(1) Chain rule with $\mu' = \mu e^{\ell}$, so $\mathrm{d}\mu'/\mathrm{d}\ell = \mu'$.
(2) (T5.61), then (T5.58) with $n = 2$, evaluated at $(\mu', \bar\lambda)$.

This is an ordinary linear equation in ℓ, solved by an exponential: $G^{(2)}(x; \mu e^{\ell}, \bar\lambda(\ell)) = G^{(2)}(x; \mu, \lambda) \exp\big(-2\int_0^{\ell} \gamma(\bar\lambda(\ell'))\mathrm{d}\ell'\big)$. Now choose ℓ so that the new scale is the inverse distance, $\mu e^{\ell_x} = 1/r$, that is $\ell_x = -\ln(\mu r)$. At that scale the dimensionless function in (T5.59) is evaluated at $\mu' r = 1$, and

$$G^{(2)}(x; \mu, \lambda) = \frac{F\big(1, \bar\lambda(\ell_x)\big)}{r^{2\Delta_{\mathrm{cl}}}}\, \exp\Big(2\int_0^{\ell_x} \gamma\big(\bar\lambda(\ell')\big)\,\mathrm{d}\ell'\Big), \qquad \ell_x = -\ln(\mu r). \tag{T5.62}$$

Everything about the distance dependence is now carried by the running coupling at the scale $1/r$ and by the accumulated anomalous dimension. If $\bar\lambda(\ell_x)$ is small, $F(1, \bar\lambda)$ can be computed in perturbation theory even when $\ln(\mu r)$ is large: the solution *resums* the large logarithms that spoil a naive expansion in λ.

At a fixed point: a pure power law. Suppose $\beta(\lambda^*) = 0$ and start at $\lambda = \lambda^*$. Then $\bar\lambda(\ell) = \lambda^*$ for all ℓ (the right side of (T5.61) vanishes), $\gamma(\bar\lambda) = \gamma^* := \gamma(\lambda^*)$ is constant, and

$$G^{(2)}(x) \overset{(1)}{=} \frac{F(1, \lambda^*)}{r^{2\Delta_{\mathrm{cl}}}}\, e^{2\gamma^* \ell_x} \overset{(2)}{=} \frac{F(1, \lambda^*)}{r^{2\Delta_{\mathrm{cl}}}}\, (\mu r)^{-2\gamma^*} = \frac{F(1, \lambda^*)\, \mu^{-2\gamma^*}}{r^{2(\Delta_{\mathrm{cl}} + \gamma^*)}}. \tag{T5.63}$$

(1) (T5.62) with a constant integrand.

(2) $\ell_x = -\ln(\mu r)$, so $e^{2\gamma^*\ell_x} = (\mu r)^{-2\gamma^*}$.

The correlator is a pure power, $G \propto r^{-2\Delta}$, with

$$\Delta = \Delta_{\mathrm{cl}} + \gamma(\lambda^*). \tag{T5.64}$$

The anomalous dimension at the fixed point is exactly the shift of the scaling dimension away from its classical value. This is the scaling dimension of the workdef in Section T5.1.1, the Δ of $y = d - \Delta$ ((T5.48)), and, in two dimensions, the conformal weight $\Delta = h + \bar{h}$ of the primary field that the conformal field theory of Chapter 11 assigns to φ. The same steps for any local operator $\mathcal{O}$ with its own anomalous dimension $\gamma_{\mathcal{O}}$ give $\Delta_{\mathcal{O}} = \Delta_{\mathcal{O},\mathrm{cl}} + \gamma_{\mathcal{O}}(\lambda^*)$.

Near, rather than at, the fixed point, expand $\beta(\lambda) \approx \beta'(\lambda^*)(\lambda - \lambda^*)$. Then (T5.61) gives $\bar{\lambda}(\ell) - \lambda^* \propto e^{\beta'(\lambda^*)\ell} = (\mu r)^{-\beta'(\lambda^*)}$. If $\beta'(\lambda^*) > 0$ this dies away at long distances: the coupling is irrelevant at λ^*, and its RG eigenvalue in the language of Section T5.3.1 is $y = -\beta'(\lambda^*)$ (the minus sign is the change from t to $\ln \mu$).

Worked example T5.16: Massless φ^4 in $d = 4$: logarithms instead of powers

At one loop in $d = 4$, $\beta = b\lambda^2$ with $b = 3/16\pi^2$ (derived in Section T5.4.2), and at two loops $\gamma = c\lambda^2$ with $c = 1/(12(4\pi)^4)$ (Example T5.15). The running coupling and the exponent in (T5.62) are

$$\bar{\lambda}(\ell) \overset{(1)}{=} \frac{\lambda}{1 - b\lambda\ell},$$
$$\int_0^\ell \gamma\big(\bar{\lambda}(\ell')\big)\mathrm{d}\ell' \overset{(2)}{=} \int_0^\ell \frac{c\,\bar{\lambda}^2}{b\,\bar{\lambda}^2}\frac{\mathrm{d}\bar{\lambda}}{\mathrm{d}\ell'}\,\mathrm{d}\ell' \overset{(3)}{=} \frac{c}{b}\big(\bar{\lambda}(\ell) - \lambda\big).$$

(1) Separate variables in $\mathrm{d}\bar{\lambda}/\mathrm{d}\ell = b\bar{\lambda}^2$: $1/\lambda - 1/\bar{\lambda} = b\ell$.
(2) Insert $1 = \beta(\bar{\lambda})^{-1}\mathrm{d}\bar{\lambda}/\mathrm{d}\ell'$ with $\beta = b\bar{\lambda}^2$.
(3) The integrand is $\frac{c}{b}\,\mathrm{d}\bar{\lambda}/\mathrm{d}\ell'$, a total derivative.

With $\ell_x = -\ln(\mu r)$ the two-point function is therefore

$$G^{(2)}(x) = \frac{F(1, \bar{\lambda}_x)}{r^2}\,\exp\!\Big[\frac{2c}{b}\big(\bar{\lambda}_x - \lambda\big)\Big], \qquad \bar{\lambda}_x = \frac{\lambda}{1 + b\lambda\ln(\mu r)}.$$

At long distances $\ln(\mu r) \to \infty$, so $\bar{\lambda}_x \approx 1/(b\ln(\mu r)) \to 0$ and $F(1, \bar{\lambda}_x) \to F(1,0) = 1/(4\pi^2)$, the free massless propagator. The correlator tends to $\frac{1}{4\pi^2 r^2}e^{-2c\lambda/b}$: a *free* power law with no anomalous exponent, approached only logarithmically. (Numerically $2c/b = 1/(18\cdot 16\pi^2) \approx 3.5 \times 10^{-4}$, so the constant is very close to 1.) This is the statement that four-dimensional φ^4 theory is trivial at long distances, the marginally irrelevant flow of Example T5.12. At short distances $\bar{\lambda}_x$ grows and reaches the Landau pole at $r = \mu^{-1}e^{-1/(b\lambda)}$, where one loop can no longer be trusted.

Worked example T5.17: $d = 4 - \varepsilon$: crossover to the Wilson–Fisher power law

Now $\beta = -\varepsilon\lambda + b\lambda^2$, with fixed point $\lambda^* = \varepsilon/b = 16\pi^2\varepsilon/3$. For $u := 1/\bar{\lambda}$ the flow equation becomes linear:

$$\frac{\mathrm{d}u}{\mathrm{d}\ell} \overset{(1)}{=} -\frac{1}{\bar{\lambda}^2}\frac{\mathrm{d}\bar{\lambda}}{\mathrm{d}\ell} \overset{(2)}{=} \frac{\varepsilon}{\bar{\lambda}} - b = \varepsilon\,u - b \quad\Longrightarrow\quad u(\ell) \overset{(3)}{=} \frac{b}{\varepsilon} + \Big(\frac{1}{\lambda} - \frac{b}{\varepsilon}\Big)e^{\varepsilon\ell},$$

(1) Chain rule for $u = \bar{\lambda}^{-1}$.
(2) Insert $\beta(\bar{\lambda}) = -\varepsilon\bar{\lambda} + b\bar{\lambda}^2$.

(3) The general solution of $u' = \varepsilon u - b$ is the constant solution b/ε plus $Ce^{\varepsilon\ell}$; $u(0) = 1/\lambda$ fixes C.

so, with $e^{\varepsilon \ell_x} = (\mu r)^{-\varepsilon}$ and $b/\varepsilon = 1/\lambda^*$,

$$\bar\lambda_x = \frac{\lambda^*}{1 + \left(\frac{\lambda^*}{\lambda} - 1\right)(\mu r)^{-\varepsilon}}.$$

Figure T5.14 plots it. As $r \to \infty$, $(\mu r)^{-\varepsilon} \to 0$ and $\bar\lambda_x \to \lambda^*$ from either side, and (T5.63) applies: $G \propto r^{-2\Delta_\varphi}$ with $\Delta_\varphi = 1 - \frac{\varepsilon}{2} + \gamma(\lambda^*) = 1 - \frac{\varepsilon}{2} + \frac{\varepsilon^2}{108}$, which is $\eta = 2\gamma(\lambda^*) = \varepsilon^2/54$ of Example T5.15. As $r \to 0$, $\bar\lambda_x \approx \lambda(\mu r)^\varepsilon \to 0$: below four dimensions the theory is free at short distances. The crossover happens where the two terms in the denominator are equal; for $\lambda \ll \lambda^*$ this is at $\mu r_\times \approx (\lambda^*/\lambda)^{1/\varepsilon}$. The approach to the fixed point is $\bar\lambda_x - \lambda^* \propto (\mu r)^{-\varepsilon}$, the power $\beta'(\lambda^*) = -\varepsilon + 2b\lambda^* = \varepsilon$, which is the irrelevant eigenvalue $y_{\rm irr} = -\varepsilon$ of Example T5.14.

Watch out: which γ, which sign

Three conventions coexist. (i) Some books define γ with $\mu\,\mathrm{d}/\mathrm{d}\mu$ acting on $\ln Z^{-1}$, or write the equation with $-n\gamma$; then $\Delta = \Delta_{\rm cl} - \gamma^*$. Check with (T5.64): in a unitary theory the field's dimension is raised by interactions, so $\gamma^* > 0$ in the convention used here. (ii) The exponent η is $2\gamma^*$, not γ^*, because the two-point function contains two fields. (iii) With the RG time t of (T5.5) the β-function changes sign but γ does not.

Reading Nash (p. 301, Ch. XI §1)

"The renormalisation properties of correlation functions are constrained by the need to obey the Callan–Symanzik renormalisation group equations." Equation (T5.58) is that constraint, and (T5.63) is Nash's remark that at a zero of β the scale dimension differs from the canonical one by the anomalous dimension.

T5.4.2 The one-loop β-function of φ^4 theory

Section T5.3.3 found $\mathrm{d}\lambda/\mathrm{d}t = -3\lambda^2/16\pi^2$ in $d = 4$ by integrating out a shell. We now get the same number from the scattering amplitude, with dimensional regularisation, as it is done in particle physics. The calculation is short, but it uses four standard tools (Feynman parameters, the Wick rotation, the d-dimensional loop integral and the expansion of Γ near its pole), and we derive each of them.

For this subsection only we work in Minkowski space, with metric $\eta = \mathrm{diag}(+1,-1,-1,-1)$, $k^2 = k_0^2 - \vec{k}^2$, Lagrangian $\mathcal{L} = \frac12 \partial_\mu\varphi\,\partial^\mu\varphi - \frac12 m^2\varphi^2 - \frac{\lambda_0}{4!}\varphi^4$, and the Feynman rules: vertex $-i\lambda_0$, propagator $i/(k^2 - m^2 + i0)$, a factor $\int \mathrm{d}^dk/(2\pi)^d$ per loop. For $2 \to 2$ scattering with incoming momenta p_1, p_2 and outgoing p_3, p_4, the Mandelstam variables are $s = (p_1+p_2)^2$, $t = (p_1 - p_3)^2$, $u = (p_1 - p_4)^2$. The tree amplitude is $i\mathcal{M} = -i\lambda_0$.

Step 1: the diagrams and their weight. At order λ_0^2 the amplitude comes from the three diagrams of Figure T5.15. For the s-channel, expand $e^{i\int\mathcal{L}}$ to second order and count the Wick contractions that attach p_1, p_2 to one vertex and p_3, p_4 to the other:

$$\frac{1}{2!}\left(\frac{-i\lambda_0}{4!}\right)^2 \times \underbrace{2\cdot(4\cdot3)\cdot(4\cdot3)\cdot2}_{576} \overset{(1)}{=} \frac{(-i\lambda_0)^2}{2}.$$

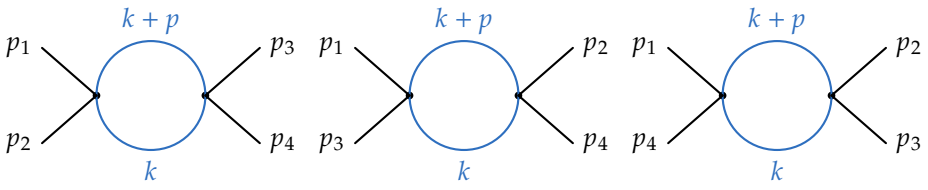


Figure T5.15: The three one-loop corrections to $2 \to 2$ scattering in φ^4 theory. In each, two external lines meet at one vertex and two at the other, and the two vertices are joined by a loop of two propagators; p is the momentum flowing through the loop ($p = p_1 + p_2$, $p_1 - p_3$ or $p_1 - p_4$). What to see: the three diagrams are the same integral evaluated at $p^2 = s, t$ and u, and each has the same logarithmic divergence, which is why the β-function has a factor 3.

(1) Choose the vertex for p_1, p_2 (2 ways), the legs there for p_1 and p_2 ($4 \cdot 3$), the legs at the other vertex for p_3, p_4 ($4 \cdot 3$), and pair the two leftover legs of one vertex with those of the other (2); $576/(2 \cdot 576) = \frac{1}{2}$.

So, with $p = p_1 + p_2$,

$$i\mathcal{M}_s = \frac{(-i\lambda_0)^2}{2} \int \frac{\mathrm{d}^d k}{(2\pi)^d} \frac{i}{k^2 - m^2 + i0} \frac{i}{(k+p)^2 - m^2 + i0} = \frac{\lambda_0^2}{2} I(p^2), \tag{T5.65}$$

where $I(p^2) := \int \frac{\mathrm{d}^d k}{(2\pi)^d} \frac{1}{(k^2-m^2+i0)((k+p)^2-m^2+i0)}$ and $(-i)^2 \cdot i^2 = +1$. The factor $\frac{1}{2}$ is the **symmetry factor** of the bubble: exchanging its two internal lines gives the same diagram.

Step 2: Feynman parameters. Two denominators are combined into one with

$$\frac{1}{AB} = \int_0^1 \frac{\mathrm{d}x}{\big[xA + (1-x)B\big]^2}. \tag{T5.66}$$

To check it, write the denominator as $B + x(A - B)$:

$$\int_0^1 \frac{\mathrm{d}x}{\big[B + x(A-B)\big]^2} \overset{(1)}{=} \Big[-\frac{1}{(A-B)\big(B + x(A-B)\big)}\Big]_0^1 \overset{(2)}{=} \frac{1}{A-B}\Big(\frac{1}{B} - \frac{1}{A}\Big) = \frac{1}{AB}.$$

(1) $\int \mathrm{d}x\,(B + cx)^{-2} = -1/(c(B + cx))$ with $c = A - B$.
(2) Upper limit gives A, lower limit gives B; $\frac{1}{B} - \frac{1}{A} = \frac{A-B}{AB}$.

Take $A = (k+p)^2 - m^2$ and $B = k^2 - m^2$:

$$xA + (1-x)B \overset{(1)}{=} k^2 + 2x\,k \cdot p + x\,p^2 - m^2 \overset{(2)}{=} (k + xp)^2 - \Delta, \qquad \Delta := m^2 - x(1-x)p^2.$$

(1) Expand $(k+p)^2 = k^2 + 2k \cdot p + p^2$ and collect.
(2) $(k + xp)^2 = k^2 + 2xk \cdot p + x^2 p^2$, and $xp^2 - x^2 p^2 = x(1-x)p^2$.

Shift the loop momentum to $\ell = k + xp$ (a shift does not change $\int \mathrm{d}^d k$ over all of momentum space):

$$I(p^2) = \int_0^1 \mathrm{d}x \int \frac{\mathrm{d}^d \ell}{(2\pi)^d} \frac{1}{(\ell^2 - \Delta + i0)^2}. \tag{T5.67}$$

Step 3: the Wick rotation. For fixed $\vec{\ell}$ the integrand, as a function of ℓ_0, has double poles where $\ell_0^2 = \vec{\ell}^2 + \Delta - i0$, that is at $\ell_0 = +E - i0$ and $\ell_0 = -E + i0$ with $E = \sqrt{\vec{\ell}^2 + \Delta}$ (take $\Delta > 0$ first). The first lies just below the positive real axis, the second just above the negative real axis (Figure T5.16). So the first and third quadrants contain no poles, and the integration contour along the real axis can be rotated anticlockwise onto the imaginary axis, $\ell_0 = i\ell_E^0$ with ℓ_E^0 real. The arcs at infinity contribute nothing, because the integrand falls like $|\ell_0|^{-4}$. With $\mathrm{d}\ell_0 = i\,\mathrm{d}\ell_E^0$ and $\ell^2 = -(\ell_E^0)^2 - \vec{\ell}^2 := -\ell_E^2$,

$$\int \frac{\mathrm{d}^d\ell}{(2\pi)^d}\,\frac{1}{(\ell^2-\Delta)^2} \overset{(1)}{=} i\int \frac{\mathrm{d}^d\ell_E}{(2\pi)^d}\,\frac{1}{(-\ell_E^2-\Delta)^2} = i\int \frac{\mathrm{d}^d\ell_E}{(2\pi)^d}\,\frac{1}{(\ell_E^2+\Delta)^2}. \tag{T5.68}$$

(1) Rotate the ℓ_0 contour; the other $d-1$ components are untouched, and together with ℓ_E^0 they form a Euclidean vector ℓ_E.

For $\Delta < 0$ (above threshold) the result is defined by analytic continuation in Δ, which is what the $+i0$ prescription instructs.

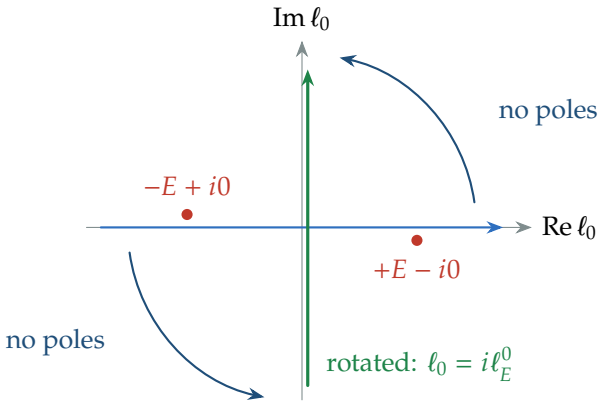


Figure T5.16: The Wick rotation. The poles of the propagators (red) sit just below the positive real axis and just above the negative one. What to see: turning the real contour (blue) anticlockwise by 90° sweeps only through the first and third quadrants, where there are no poles, so the integral along the imaginary axis (green) equals the original one.

Step 4: the d-dimensional integral. For $n > d/2$ and $\Delta > 0$,

$$\int \frac{\mathrm{d}^d\ell_E}{(2\pi)^d}\,\frac{1}{(\ell_E^2+\Delta)^n} \overset{(1)}{=} \frac{1}{\Gamma(n)}\int_0^\infty \mathrm{d}s\; s^{n-1}e^{-s\Delta}\int \frac{\mathrm{d}^d\ell_E}{(2\pi)^d}\,e^{-s\ell_E^2} \overset{(2)}{=} \frac{1}{\Gamma(n)}\int_0^\infty \mathrm{d}s\;\frac{s^{n-1}e^{-s\Delta}}{(4\pi s)^{d/2}}$$
$$\overset{(3)}{=} \frac{\Gamma(n-\frac{d}{2})}{(4\pi)^{d/2}\,\Gamma(n)}\,\Delta^{\frac{d}{2}-n}. \tag{T5.69}$$

(1) $X^{-n} = \frac{1}{\Gamma(n)}\int_0^\infty \mathrm{d}s\; s^{n-1}e^{-sX}$ for $X > 0$ (substitute $s' = sX$ to see it), with $X = \ell_E^2 + \Delta$; then exchange the order of integration.

(2) The Gaussian $\int \frac{\mathrm{d}^d\ell}{(2\pi)^d}e^{-s\ell^2} = (4\pi s)^{-d/2}$ (Section T5.2.5).

(3) $\int_0^\infty \mathrm{d}s\; s^{z-1}e^{-s\Delta} = \Gamma(z)\Delta^{-z}$ with $z = n - \frac{d}{2}$.

The right side makes sense for any complex d with $n - d/2$ not a non-positive integer. **Dimensional regularisation** [evaluate the loop in a dimension where it converges, then continue in d] *defines* the loop integral by this formula. For $n = 2$ and $d = 4$ it has a pole: the logarithmic divergence of the bubble.

Step 5: the pole. Put $d = 4 - \varepsilon$ and $n = 2$. Three expansions are needed:

$$\Gamma\!\left(\tfrac{\varepsilon}{2}\right) \overset{(1)}{=} \frac{2}{\varepsilon} - \gamma_E + O(\varepsilon), \qquad (4\pi)^{-d/2} \overset{(2)}{=} \frac{1}{16\pi^2}\Big(1 + \frac{\varepsilon}{2}\ln 4\pi + O(\varepsilon^2)\Big),$$
$$\left(\frac{\Delta}{\mu^2}\right)^{-\varepsilon/2} \overset{(3)}{=} 1 - \frac{\varepsilon}{2}\ln\frac{\Delta}{\mu^2} + O(\varepsilon^2).$$

(1) $\Gamma(z) = \Gamma(z+1)/z$ and $\Gamma(1+z) = 1 - \gamma_E z + O(z^2)$, with $\gamma_E = 0.5772\ldots$ the Euler constant.
(2) $(4\pi)^{-2+\varepsilon/2} = (4\pi)^{-2} e^{\frac{\varepsilon}{2}\ln 4\pi}$.
(3) $a^{-\varepsilon/2} = e^{-\frac{\varepsilon}{2}\ln a}$. Here μ is an arbitrary mass, inserted to make the argument of the logarithm dimensionless; it will be the renormalisation scale.

Combining (T5.67), (T5.68) and (T5.69), and multiplying and dividing by $\mu^{-\varepsilon}$,

$$I(p^2) \overset{(1)}{=} i\,\mu^{-\varepsilon}\int_0^1 \mathrm{d}x\;\frac{\Gamma(\frac{\varepsilon}{2})}{(4\pi)^{d/2}}\left(\frac{\Delta}{\mu^2}\right)^{-\varepsilon/2}$$
$$\overset{(2)}{=} \frac{i\,\mu^{-\varepsilon}}{16\pi^2}\left[\frac{2}{\varepsilon} - \gamma_E + \ln 4\pi - \int_0^1 \mathrm{d}x\;\ln\frac{m^2 - x(1-x)p^2}{\mu^2}\right] + O(\varepsilon). \tag{T5.70}$$

(1) $\Delta^{d/2-2} = \Delta^{-\varepsilon/2} = \mu^{-\varepsilon}(\Delta/\mu^2)^{-\varepsilon/2}$, and $\Gamma(2) = 1$.
(2) Multiply the three expansions and keep terms up to $O(\varepsilon^0)$: the $\frac{2}{\varepsilon}$ times each $O(\varepsilon)$ correction gives the finite terms.

Step 6: renormalise and read off β. Add the three channels. The pole is the same in each, and the amplitude is

$$\mathcal{M} \overset{(1)}{=} -\lambda_0 + \frac{\lambda_0^2}{2i}\big[I(s) + I(t) + I(u)\big] \overset{(2)}{=} -\lambda_0 + \frac{3\lambda_0^2\mu^{-\varepsilon}}{16\pi^2\,\varepsilon} + (\text{finite}). \tag{T5.71}$$

(1) Divide $i\mathcal{M}$ by i; tree plus (T5.65) for each channel.
(2) Each I contributes $\frac{i\mu^{-\varepsilon}}{16\pi^2}\frac{2}{\varepsilon}$; three channels and the prefactor $\frac{1}{2}$ give $3\cdot\frac{1}{2}\cdot\frac{2}{\varepsilon} = \frac{3}{\varepsilon}$.

In terms of the renormalised coupling of (T5.55), written as $\lambda_0 = \mu^\varepsilon\big(\lambda + a\lambda^2/\varepsilon + O(\lambda^3)\big)$ (the field is not renormalised at one loop, Example T5.13), (T5.71) is

$$\mu^{-\varepsilon}\mathcal{M} = -\lambda - \frac{a\lambda^2}{\varepsilon} + \frac{3\lambda^2}{16\pi^2\varepsilon} + (\text{finite}) + O(\lambda^3),$$

which is finite exactly when $a = 3/16\pi^2$. Now hold λ_0 fixed:

$$0 \overset{(1)}{=} \mu\frac{\mathrm{d}\lambda_0}{\mathrm{d}\mu} \overset{(2)}{=} \mu^\varepsilon\left[\varepsilon\left(\lambda + \frac{a\lambda^2}{\varepsilon}\right) + \left(1 + \frac{2a\lambda}{\varepsilon}\right)\beta(\lambda)\right]$$
$$\Longrightarrow \quad \beta(\lambda) \overset{(3)}{=} -\varepsilon\left(\lambda + \frac{a\lambda^2}{\varepsilon}\right)\left(1 - \frac{2a\lambda}{\varepsilon} + O(\lambda^2)\right) \overset{(4)}{=} -\varepsilon\lambda + a\lambda^2 + O(\lambda^3). \tag{T5.72}$$

(1) The bare coupling does not know about μ.
(2) Differentiate μ^ε (giving $\varepsilon\mu^\varepsilon$) and the bracket (giving $(1 + 2a\lambda/\varepsilon)\,\mu\,\mathrm{d}\lambda/\mathrm{d}\mu$).
(3) Solve for β and expand $(1 + 2a\lambda/\varepsilon)^{-1}$ to first order in λ.
(4) Multiply out: $-\varepsilon\lambda - a\lambda^2 + 2a\lambda^2$. The poles in $1/\varepsilon$ cancel, as they must for a finite β.

Key idea: one-loop β-function of φ^4

$$\beta(\lambda) = \mu\frac{\mathrm{d}\lambda}{\mathrm{d}\mu} = -\varepsilon\lambda + \frac{3\lambda^2}{16\pi^2} + O(\lambda^3), \qquad \text{so in } d=4: \quad \beta = \frac{3\lambda^2}{16\pi^2}. \tag{T5.73}$$

Check against Wilson's shell integration. In the convention of (T5.5) this is $\mathrm{d}\lambda/\mathrm{d}t = \varepsilon\lambda - 3\lambda^2/16\pi^2$, the same coefficient that the sharp-cutoff shell calculation gave in (T5.54). The fixed point $\lambda^* = 16\pi^2\varepsilon/3$ is the Wilson–Fisher point: in Wilson's variable $\bar{u} = K_4\lambda = \lambda/8\pi^2$ it is $\bar{u}^* = \frac{2}{3}\varepsilon$, as in Example T5.14, and its slope $\beta'(\lambda^*) = -\varepsilon + 2\cdot\frac{3}{16\pi^2}\lambda^* = \varepsilon$ is the eigenvalue $y_{\mathrm{irr}} = -\varepsilon$ with the sign flipped. Why do two quite different schemes agree? For a marginal coupling the one-loop coefficient is scheme independent. If a second scheme uses $\lambda' = \lambda + c\lambda^2 + \cdots$, then

$$\beta'(\lambda') \overset{(1)}{=} \mu\frac{\mathrm{d}\lambda'}{\mathrm{d}\mu} \overset{(2)}{=} (1+2c\lambda)\,b\lambda^2 + O(\lambda^4) \overset{(3)}{=} b\lambda'^2 + O(\lambda'^3),$$

(1) Definition of β in the new scheme.
(2) Chain rule, $\mathrm{d}\lambda'/\mathrm{d}\lambda = 1 + 2c\lambda$, and $\beta(\lambda) = b\lambda^2 + O(\lambda^3)$ in $d = 4$.
(3) $\lambda^2 = \lambda'^2 + O(\lambda'^3)$; the change starts at order λ^3.

so b is the same in every scheme (the location of a fixed point in $d < 4$ is not, as the LPA showed, but its eigenvalues are).

Worked example T5.18: The β-function from μ-independence of the amplitude

Take $m = 0$ and Euclidean momenta, so $s, t, u < 0$ and every logarithm is real. From (T5.70), $\int_0^1 \mathrm{d}x\,\ln\!\big(x(1-x)(-s)/\mu^2\big) = \ln(-s/\mu^2) - 2$ (use $\int_0^1 \ln x\,\mathrm{d}x = -1$ twice). After the pole is removed, at $\varepsilon \to 0$,

$$\mathcal{M} = -\lambda(\mu) + \frac{\lambda^2}{32\pi^2}\sum_{q^2=s,t,u}\left[2 - \gamma_E + \ln 4\pi - \ln\frac{-q^2}{\mu^2}\right] + O(\lambda^3).$$

A physical amplitude cannot depend on the bookkeeping scale μ:

$$0 = \mu\frac{\mathrm{d}\mathcal{M}}{\mathrm{d}\mu} \overset{(1)}{=} -\beta(\lambda) + \frac{\lambda^2}{32\pi^2}\cdot 3\cdot 2 + O(\lambda^3) \quad\Longrightarrow\quad \beta(\lambda) \overset{(2)}{=} \frac{3\lambda^2}{16\pi^2}.$$

(1) $\mu\,\partial_\mu$ acting on $-\ln(-q^2/\mu^2)$ gives $+2$ for each of the three channels; $\mu\,\mathrm{d}\lambda/\mathrm{d}\mu = \beta$ acts on the tree term. The derivative of the λ^2 prefactor gives $O(\lambda^3)$.
(2) $6/32\pi^2 = 3/16\pi^2$.

The same logarithms say what "running" means: at large momentum transfer $Q^2 = -s$ the one-loop amplitude grows like $\lambda + \frac{3\lambda^2}{16\pi^2}\ln\frac{Q}{\mu}$, which is the first term of $\bar{\lambda}(\ln\frac{Q}{\mu}) = \lambda/(1 - \frac{3\lambda}{16\pi^2}\ln\frac{Q}{\mu})$. For $\lambda = 1$ at μ the Landau pole sits at $\ln(Q/\mu) = 16\pi^2/3 \approx 52.6$, some 10^{22} times higher: harmless in practice, but a sign that φ^4 theory in $d = 4$ needs new physics at some short distance.

Worked example T5.19: The $O(N)$ model: $\beta = (N+8)\lambda^2/48\pi^2$

Take N scalars φ_a with potential $\frac{\lambda}{4!}(\varphi_a\varphi_a)^2$. Its fourth derivative is the vertex $-i\lambda\,S_{abcd}/3$ with $S_{abcd} := \delta_{ab}\delta_{cd} + \delta_{ac}\delta_{bd} + \delta_{ad}\delta_{bc}$ (check at $N = 1$: $S = 3$, vertex $-i\lambda$). The s-channel

bubble carries the internal sum

$$\sum_{e,f} S_{abef} S_{efcd} \overset{(1)}{=} N\,\delta_{ab}\delta_{cd} + 4\,\delta_{ab}\delta_{cd} + 2\,\delta_{ac}\delta_{bd} + 2\,\delta_{ad}\delta_{bc}$$
$$\overset{(2)}{=} (N+4)\,\delta_{ab}\delta_{cd} + 2\,\delta_{ac}\delta_{bd} + 2\,\delta_{ad}\delta_{bc}.$$

(1) Multiply out the nine products. The product $\delta_{ab}\delta_{ef}\cdot\delta_{ef}\delta_{cd}$ has a free sum $\delta_{ee}=N$. The four products that contain exactly one of $\delta_{ab}\delta_{ef}$ and $\delta_{ef}\delta_{cd}$ each give $\delta_{ab}\delta_{cd}$. Of the four remaining products, two give $\delta_{ac}\delta_{bd}$ and two give $\delta_{ad}\delta_{bc}$.
(2) Collect the $\delta_{ab}\delta_{cd}$ terms.

The t- and u-channels are the same with (b,c) and (b,d) exchanged. Adding the three channels,

$$\sum_{\text{channels}} \overset{(1)}{=} (N+4+2+2)\left(\delta_{ab}\delta_{cd}+\delta_{ac}\delta_{bd}+\delta_{ad}\delta_{bc}\right) \overset{(2)}{=} (N+8)\,S_{abcd}.$$

(1) Each structure gets $N+4$ from the channel in which it is the "$\delta_{ab}\delta_{cd}$" structure and 2 from each of the other two.
(2) The definition of S_{abcd}.

At $N=1$ the same sum is $9\,S_{abcd}$, so the one-loop pole of the four-point vertex is $\frac{N+8}{9}$ times its $N=1$ value. Hence

$$\beta(\lambda) = -\varepsilon\lambda + \frac{N+8}{9}\cdot\frac{3\lambda^2}{16\pi^2} = -\varepsilon\lambda + \frac{(N+8)\lambda^2}{48\pi^2}, \qquad \lambda^* = \frac{48\pi^2\varepsilon}{N+8}.$$

At $N=1$ this is (T5.73). The factor $N+8$ is the one that appears in the $O(N)$ exponents, for example $\eta=(N+2)\varepsilon^2/(2(N+8)^2)$ quoted after Example T5.15.

T5.4.3 Yang–Mills theory at one loop: the background-field method

The β-function of Yang–Mills theory is the most consequential number in this chapter. Its sign decides that quarks are almost free at short distances and confined at long ones, it produces Λ_{QCD} by dimensional transmutation (Example T5.28), and it sets the size of the trace anomaly (Section T5.5.3). We compute it with the **background-field method**, which reduces the problem to a single two-point function [Abbott 1981].

Conventions. Euclidean signature. Hermitian generators T^a in a representation r, with $[T^a,T^b]=if^{abc}T^c$ and $\mathrm{tr}(T^aT^b)=T(r)\,\delta^{ab}$. For SU($N$) the fundamental has $T(\mathrm{fund})=\frac12$, and the adjoint, with matrices $(T^a_{\mathrm{adj}})_{bc}=-if^{abc}$, has $T(\mathrm{adj})=N$, which is the identity $\sum_{c,d} f^{acd}f^{bcd}=N\delta^{ab}$. The gauge field is $A_\mu=A^a_\mu T^a$ with the coupling absorbed, so that

$$S_{\mathrm{YM}} = \frac{1}{4g^2}\int \mathrm{d}^4x\; F^a_{\mu\nu}F^a_{\mu\nu}, \qquad F^a_{\mu\nu}=\partial_\mu A^a_\nu - \partial_\nu A^a_\mu + f^{abc}A^b_\mu A^c_\nu, \tag{T5.74}$$

matter in r has $D_\mu=\partial_\mu - iA_\mu$ and $[D_\mu,D_\nu]=-iF_{\mu\nu}$ with $F_{\mu\nu}=F^a_{\mu\nu}T^a$, and an adjoint field X^a has $(D_\mu X)^a=\partial_\mu X^a + f^{abc}A^b_\mu X^c$. (In the anti-Hermitian convention of Chapter T3, $A_{\mathrm{math}}=-iA$ and $F_{\mathrm{math}}=-iF$. Rescaling $A=gA_c$ gives the canonical normalisation $\frac14 F_cF_c$, with g in the commutator.)

Worked example T5.20: SU(2): the adjoint is the vector representation

For SU(2), $f^{abc} = \epsilon^{abc}$ and $T^a = \sigma^a/2$ in the fundamental, so $\mathrm{tr}(T^3T^3) = \frac{1}{4}\,\mathrm{tr}(\sigma^3)^2 = \frac{1}{2}$. The adjoint matrices $(T^a)_{bc} = -i\epsilon^{abc}$ are

$$T^3_{\text{adj}} = \begin{pmatrix} 0 & -i & 0 \\ i & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \mathrm{tr}\big(T^3_{\text{adj}} T^3_{\text{adj}}\big) \overset{(1)}{=} (-i)(i) + (i)(-i) = 2 = N.$$

(1) The only non-zero entries are $(T^3)_{12} = -i\epsilon^{312} = -i$ and $(T^3)_{21} = -i\epsilon^{321} = i$; the trace of the square is $\sum_{b,c}(T^3)_{bc}(T^3)_{cb} = (T^3)_{12}(T^3)_{21} + (T^3)_{21}(T^3)_{12}$.

T^3_{adj} has eigenvalues $+1, 0, -1$: the three gluons of SU(2) carry charges ± 1 and 0 under the T^3 direction. The ratio $T(\text{adj})/T(\text{fund}) = 2N = 4$ is used again in Section T5.7.2.

Step 1: the background field and its gauge invariance. Split the gauge field into a fixed classical background and a fluctuation,

$$A_\mu = \bar{A}_\mu + a_\mu,$$

and integrate over a only. Write $\bar{D}$ and $\bar{F}$ for the covariant derivative and curvature of $\bar{A}$. Expanding (T5.74),

$$F^a_{\mu\nu} \overset{(1)}{=} \bar{F}^a_{\mu\nu} + (\bar{D}_\mu a_\nu)^a - (\bar{D}_\nu a_\mu)^a + f^{abc} a^b_\mu a^c_\nu,$$

(1) The terms linear in a are $\partial_\mu a^a_\nu - \partial_\nu a^a_\mu + f^{abc}(\bar{A}^b_\mu a^c_\nu + a^b_\mu \bar{A}^c_\nu)$; in the last one relabel $b \leftrightarrow c$ and use $f^{acb} = -f^{abc}$ to get $-f^{abc}\bar{A}^b_\nu a^c_\mu$. Grouping gives $(\bar{D}_\mu a_\nu)^a - (\bar{D}_\nu a_\mu)^a$.

We fix the gauge of the fluctuation with the **background-field gauge** $(\bar{D}_\mu a_\mu)^a = 0$, imposed by adding $\frac{1}{2g^2}\int (\bar{D}_\mu a_\mu)^a (\bar{D}_\nu a_\nu)^a$ to the action (Feynman gauge). Here is the point of the method. Under a gauge transformation of the *background*, $\bar{A} \mapsto U\bar{A}U^{-1} + iU\partial U^{-1}$ together with $a \mapsto UaU^{-1}$, the total A transforms as a gauge field, the action is invariant, and $\bar{D}_\mu a_\mu$ rotates covariantly, so the gauge-fixing term is invariant too. Hence the effective action $\Gamma[\bar{A}]$ obtained by integrating out a is a gauge-invariant functional of $\bar{A}$. Its divergent part must be a local, gauge-invariant functional of dimension four, and there is only one:

$$\Gamma[\bar{A}] = \frac{1}{4}\Big(\frac{1}{g_0^2} + \delta\Big)\int \bar{F}^a_{\mu\nu}\bar{F}^a_{\mu\nu} + (\text{finite}).$$

So the renormalisation of g is read off from the coefficient of $\bar{F}^2$, and that coefficient can be found from the part of Γ that is quadratic in $\bar{A}$, where $\bar{F}^2$ reduces to $(\partial_\mu \bar{A}^a_\nu - \partial_\nu \bar{A}^a_\mu)^2$. The quadratic part is a two-point function of the background: the **vacuum polarisation**.

Step 2: the quadratic action of the fluctuations. Keep the terms of $\frac{1}{4}F^aF^a$ quadratic in a, add the gauge-fixing term, and integrate by parts.

$$\tfrac{1}{4}F^aF^a\big|_{a^2} \overset{(1)}{=} \tfrac{1}{2}(\bar{D}_\mu a_\nu)^a(\bar{D}_\mu a_\nu)^a - \tfrac{1}{2}(\bar{D}_\mu a_\nu)^a(\bar{D}_\nu a_\mu)^a + \tfrac{1}{2}\bar{F}^a_{\mu\nu} f^{abc} a^b_\mu a^c_\nu,$$

$$-\tfrac{1}{2}\int (\bar{D}_\mu a_\nu)^a(\bar{D}_\nu a_\mu)^a \overset{(2)}{=} \tfrac{1}{2}\int a^a_\nu\big(\bar{D}_\mu\bar{D}_\nu a_\mu\big)^a \overset{(3)}{=} \tfrac{1}{2}\int a^a_\nu\big(\bar{D}_\nu\bar{D}_\mu a_\mu + [\bar{D}_\mu,\bar{D}_\nu]a_\mu\big)^a$$

$$\overset{(4)}{=} -\tfrac{1}{2}\int (\bar{D}_\mu a_\mu)^a(\bar{D}_\nu a_\nu)^a + \tfrac{1}{2}\int a^a_\nu f^{abc}\bar{F}^b_{\mu\nu}a^c_\mu,$$

$$\tfrac{1}{2}\,a^a_\nu f^{abc}\bar{F}^b_{\mu\nu}a^c_\mu \overset{(5)}{=} \tfrac{1}{2}\,\bar{F}^a_{\mu\nu}f^{abc}a^b_\mu a^c_\nu.$$

(1) Square the linear term, $(\bar D_\mu a_\nu - \bar D_\nu a_\mu)^2 = 2(\bar D_\mu a_\nu)^2 - 2(\bar D_\mu a_\nu)(\bar D_\nu a_\mu)$ summed over μ,ν, and keep the cross term $2\bar F^a_{\mu\nu} f^{abc} a^b_\mu a^c_\nu$ between $\bar F$ and the quadratic part of F.
(2) Integrate by parts; for adjoint fields $\int X^a (\bar D_\mu Y)^a = -\int (\bar D_\mu X)^a Y^a$ because f^{abc} is antisymmetric.
(3) Commute the two derivatives.
(4) Integrate the first term by parts again. For adjoint fields $([\bar D_\mu, \bar D_\nu]X)^a = f^{abc}\bar F^b_{\mu\nu} X^c$ (the adjoint case of $[D_\mu, D_\nu] = -iF_{\mu\nu}$, with $-iT^b_{\text{adj}}$ having entries $-f^{bac} = f^{abc}$).
(5) Rename $a \to b$, $b \to a$ (using $f^{bac} = -f^{abc}$), then swap $\mu \leftrightarrow \nu$ (using $\bar F_{\nu\mu} = -\bar F_{\mu\nu}$).

The $-\frac12(\bar D\cdot a)^2$ cancels the gauge-fixing term exactly, the two $\frac12 \bar F f a a$ terms add, and one more integration by parts turns $\frac12(\bar D a)^2$ into $\frac12 a(-\bar D^2 a)$:

$$S_2 = \frac{1}{2g^2}\int a^b_\mu \Big[(-\bar D^2)^{bc}\,\delta_{\mu\nu} + 2f^{abc}\bar F^a_{\mu\nu}\Big] a^c_\nu \coloneqq \frac{1}{2g^2}\int a\,\big(-\bar D^2 + X\big)a, \qquad X^{bc}_{\mu\nu} = 2f^{abc}\bar F^a_{\mu\nu}. \tag{T5.75}$$

The operator has two parts. $-\bar D^2$ is the kinetic operator of four adjoint scalars $a_1,\dots,a_4$: it describes how a charged particle's *orbit* responds to the field. X mixes the Lorentz index μ with ν in proportion to $\bar F_{\mu\nu}$: it is the coupling of the gluon's *spin* to the field, a magnetic-moment term. (With the spin-1 generators $(J_{\rho\sigma})_{\mu\nu} = \delta_{\rho\mu}\delta_{\sigma\nu} - \delta_{\sigma\mu}\delta_{\rho\nu}$ one has $X = \bar F^a_{\rho\sigma}(J_{\rho\sigma})\otimes(\cdots)$ with gyromagnetic factor 2.)

Step 3: the one-loop effective action. The Faddeev–Popov operator of the gauge $\bar D_\mu a_\mu = 0$ is $\bar D_\mu D_\mu$, which is $\bar D^2$ at one loop, so the ghosts are two anticommuting adjoint scalars with action $\int \bar c^a(-\bar D^2)^{ab}c^b$. Add n_f Dirac fermions in the fundamental, with action $\int \bar\psi\gamma_\mu D_\mu\psi$. Doing the three Gaussian integrals (a real bosonic Gaussian gives $\det^{-1/2}$, a pair of Grassmann variables gives $\det^{+1}$),

$$\Gamma_1[\bar A] = \tfrac12 \ln\det\big(-\bar D^2 + X\big)_{\text{adj, vector}} - \ln\det\big(-\bar D^2\big)_{\text{adj}} - n_f \ln\det\big(\gamma_\mu \bar D_\mu\big)_{\text{fund}}. \tag{T5.76}$$

For the fermions use $\det \vec{\not D} = \det(-\vec{\not D}^2)^{1/2}$ (the eigenvalues of the anti-Hermitian $\vec{\not D}$ come in pairs $\pm i\lambda$) and

$$\vec{\not D}^2 \overset{(1)}{=} \tfrac12\{\gamma_\mu,\gamma_\nu\}\bar D_\mu\bar D_\nu + \tfrac14[\gamma_\mu,\gamma_\nu][\bar D_\mu,\bar D_\nu] \overset{(2)}{=} \bar D^2 - \tfrac{i}{2}\gamma_{\mu\nu}\bar F_{\mu\nu}, \qquad \gamma_{\mu\nu} \coloneqq \tfrac12[\gamma_\mu,\gamma_\nu], \tag{T5.77}$$

(1) Write $\gamma_\mu\gamma_\nu$ as its symmetric plus antisymmetric part, and do the same for $\bar D_\mu\bar D_\nu$; the mixed products vanish when summed.
(2) $\{\gamma_\mu,\gamma_\nu\} = 2\delta_{\mu\nu}$ (Euclidean, Chapter T4) and $[\bar D_\mu,\bar D_\nu] = -i\bar F_{\mu\nu}$.

so $-\vec{\not D}^2 = -\bar D^2 + X_f$ with $X_f = \frac{i}{2}\gamma_{\mu\nu}\bar F_{\mu\nu}$, the spin coupling of the fermion. Every term in (T5.76) is now of the form $\ln\det(-\bar D^2 + \text{spin term})$, and we need only two master calculations.

Step 4 (master calculation A): the orbital part. Let $\Gamma_s[\bar A] \coloneqq \ln\det(-\bar D^2)$ for one *complex* scalar in a representation r (this is minus the log of $\int \mathcal{D}\phi^\dagger\mathcal{D}\phi\, e^{-\int\phi^\dagger(-\bar D^2)\phi}$). Expand $-\bar D^2 = -\partial^2 + V_1 + V_2$ in powers of $\bar A$, with $V_1 = i(\partial_\mu\bar A_\mu + \bar A_\mu\partial_\mu)$ (an operator: first multiply by $\bar A_\mu$, then differentiate, plus the other order) and $V_2 = \bar A_\mu\bar A_\mu$. With $G = (-\partial^2)^{-1}$ and $\ln\det = \operatorname{Tr}\ln$,

$$\Gamma_s\big|_{\bar A^2} \overset{(1)}{=} \operatorname{Tr}\big(GV_2\big) - \tfrac12\operatorname{Tr}\big(GV_1GV_1\big) \overset{(2)}{=} \frac12\int_p \bar A^a_\mu(-p)\bar A^b_\nu(p)\,T(r)\,\delta^{ab}\,\Pi_{\mu\nu}(p),$$

$$\Pi_{\mu\nu}(p) \overset{(3)}{=} 2\delta_{\mu\nu}\int_k \frac{1}{k^2} - \int_k \frac{(2k+p)_\mu(2k+p)_\nu}{k^2\,(k+p)^2},$$

(1) $\operatorname{Tr}\ln(G^{-1}+V)=\operatorname{Tr}\ln G^{-1}+\operatorname{Tr}(GV)-\frac12\operatorname{Tr}(GVGV)+\cdots$, the matrix identity $\ln(1+Y)=Y-\frac12Y^2+\cdots$ with $Y=GV$; keep the terms with two powers of $\bar A$.
(2) In momentum space V_1 takes a plane wave e^{ikx} to one of momentum $k+p$ with amplitude $i\cdot i(2k+p)_\mu\bar A_\mu(p)=-(2k+p)_\mu\bar A_\mu(p)$; the colour trace gives $\operatorname{tr}(T^aT^b)=T(r)\delta^{ab}$; we write $\int_k$ for $\int\frac{\mathrm{d}^dk}{(2\pi)^d}$.
(3) The first term is the "seagull" V_2, the second the loop with two V_1 vertices.

In dimensional regularisation $\int_k k^{-2}=0$ (it has no scale to depend on). The second integral is done exactly as the bubble: Feynman parameter x, shift $\ell=k+xp$ so that $2k+p=2\ell+(1-2x)p$ and the denominator is $(\ell^2+\Delta)^2$ with $\Delta=x(1-x)p^2$, then (T5.69). The term linear in ℓ vanishes by symmetry and $\ell_\mu\ell_\nu\to\delta_{\mu\nu}\ell^2/d$:

$$\begin{aligned}\Pi_{\mu\nu}&\overset{(1)}{=}-\int_0^1\mathrm{d}x\Big[\frac{4\delta_{\mu\nu}}{d}\int_\ell\frac{\ell^2}{(\ell^2+\Delta)^2}+(1-2x)^2p_\mu p_\nu\int_\ell\frac{1}{(\ell^2+\Delta)^2}\Big]\\&\overset{(2)}{=}-\frac{\Gamma(2-\frac d2)}{(4\pi)^{d/2}}\int_0^1\mathrm{d}x\,\Delta^{\frac d2-2}\Big[\frac{2\Delta}{1-\frac d2}\,\delta_{\mu\nu}+(1-2x)^2p_\mu p_\nu\Big]\\&\overset{(3)}{=}\frac{\Gamma(2-\frac d2)}{(4\pi)^{d/2}}\,(p^2)^{\frac d2-2}\,C_d\,\big(p^2\delta_{\mu\nu}-p_\mu p_\nu\big),\qquad C_d:=\int_0^1\mathrm{d}x\,(1-2x)^2\big(x(1-x)\big)^{\frac d2-2}.\end{aligned}$$

(1) Insert the numerator $4\ell_\mu\ell_\nu+(1-2x)^2p_\mu p_\nu$ and average $\ell_\mu\ell_\nu$ over directions.
(2) Write $\ell^2=(\ell^2+\Delta)-\Delta$ and use (T5.69) with $n=1,2$: $\frac4d\big[\frac{\Gamma(1-\frac d2)}{(4\pi)^{d/2}}\Delta^{\frac d2-1}-\frac{\Gamma(2-\frac d2)}{(4\pi)^{d/2}}\Delta^{\frac d2-1}\big]$, and $\Gamma(1-\frac d2)=\Gamma(2-\frac d2)/(1-\frac d2)$, so the bracket is $\frac{\Gamma(2-\frac d2)}{(4\pi)^{d/2}}\Delta^{\frac d2-1}\frac{d/2}{1-d/2}$; times $\frac4d$ this is $\frac{2\Delta}{1-d/2}$.
(3) Integration by parts in x: $0=\int_0^1\mathrm{d}x\,\frac{\mathrm{d}}{\mathrm{d}x}\big[(x(1-x))^{\frac d2-1}(1-2x)\big]=(\frac d2-1)\int(1-2x)^2(x(1-x))^{\frac d2-2}-2\int(x(1-x))^{\frac d2-1}$, so $\int\frac{2\Delta^{d/2-1}}{1-d/2}\mathrm{d}x=-(p^2)^{\frac d2-1}C_d$.

The result is exactly *transverse*, $p_\mu\Pi_{\mu\nu}=0$, as background gauge invariance demands. At $d=4-\varepsilon$: $\Gamma(\frac\varepsilon2)=\frac2\varepsilon+\cdots$ and $C_4=\int_0^1(1-2x)^2\mathrm{d}x=\frac13$. Since $\frac12\int_p\bar A^a_\mu(-p)\big(p^2\delta_{\mu\nu}-p_\mu p_\nu\big)\bar A^a_\nu(p)=\frac14\int(\partial_\mu\bar A^a_\nu-\partial_\nu\bar A^a_\mu)^2$, we get

$$\Gamma_s\big|_{\bar F^2}=\frac14\cdot\frac{T(r)}{3}\,L(p)\int\bar F^a_{\mu\nu}\bar F^a_{\mu\nu},\qquad L(p):=\frac{1}{16\pi^2}\Big(\frac2\varepsilon-\ln\frac{p^2}{\mu^2}+\text{const}\Big).\tag{T5.78}$$

(With a momentum cutoff instead, $L(p)=\frac{1}{16\pi^2}\ln\frac{\Lambda^2}{p^2}$.) A charged scalar *increases* the coefficient of $\bar F^2$ at long distances: it screens the charge, like the electron–positron pairs of QED. This is the magnetic response of the orbital motion of charges: Landau diamagnetism.

Step 5 (master calculation B): the spin part. For an operator $-\bar D^2+X$ with X linear in $\bar F$, the new terms quadratic in the background come from two insertions of X (a cross term of X with the orbital vertex V_1 has a vanishing Lorentz trace, because $X_{\mu\nu}$ is antisymmetric in the gluon case and $\operatorname{tr}\gamma_{\mu\nu}=0$ in the fermion case):

$$\begin{aligned}\ln\det\big(-\bar D^2+X\big)-\ln\det\big(-\bar D^2\big)&\overset{(1)}{=}-\tfrac12\operatorname{Tr}\big(GXGX\big)+\cdots\overset{(2)}{=}-\frac12\int_p\operatorname{tr}X(-p)X(p)\int_k\frac{1}{k^2(k+p)^2}\\&\overset{(3)}{=}-\frac12\,L(p)\int\operatorname{tr}X^2.\end{aligned}\tag{T5.79}$$

(1) The second-order term of $\operatorname{Tr}\ln(G^{-1}+X)$, as in step 4.

(2) X has no derivatives, so each insertion just transfers momentum; the trace tr is over colour and spin (or Lorentz) indices.

(3) The loop is (T5.70) at $m = 0$ in Euclidean form, $\int_k \frac{1}{k^2(k+p)^2} = \frac{1}{16\pi^2}\big(\frac{2}{\varepsilon} - \ln\frac{p^2}{\mu^2} + \cdots\big) = L(p)$; in position space $\int_p \operatorname{tr} X(-p)X(p) = \int \mathrm{d}^4x\ \operatorname{tr} X^2$.

The two traces we need are:

$$\text{gluon:}\quad \operatorname{tr} X^2 \overset{(1)}{=} 4 f^{abc} f^{dcb} \bar F^a_{\mu\nu} \bar F^d_{\nu\mu} \overset{(2)}{=} 4 f^{abc} f^{dbc} \bar F^a_{\mu\nu} \bar F^d_{\mu\nu} \overset{(3)}{=} 4N\, \bar F^a_{\mu\nu} \bar F^a_{\mu\nu},$$

$$\text{fermion:}\quad \operatorname{tr} X_f^2 \overset{(4)}{=} -\tfrac14 \operatorname{tr}\big(\gamma_{\mu\nu}\gamma_{\rho\sigma}\big)\ \operatorname{tr}\big(\bar F_{\mu\nu}\bar F_{\rho\sigma}\big) \overset{(5)}{=} -T(r)\big(\delta_{\mu\sigma}\delta_{\nu\rho} - \delta_{\mu\rho}\delta_{\nu\sigma}\big)\bar F^a_{\mu\nu}\bar F^a_{\rho\sigma} \overset{(6)}{=} 2T(r)\, \bar F^a_{\mu\nu}\bar F^a_{\mu\nu}.$$

(1) $\operatorname{tr} X^2 = \sum X^{bc}_{\mu\nu} X^{cb}_{\nu\mu}$ with $X^{bc}_{\mu\nu} = 2 f^{abc} \bar F^a_{\mu\nu}$ from (T5.75).

(2) $f^{dcb} = -f^{dbc}$ and $\bar F^d_{\nu\mu} = -\bar F^d_{\mu\nu}$.

(3) $\sum_{b,c} f^{abc} f^{dbc} = N\delta^{ad}$.

(4) $X_f = \frac{i}{2}\gamma_{\mu\nu}\bar F_{\mu\nu}$, so $X_f^2 = -\frac14 \gamma_{\mu\nu}\gamma_{\rho\sigma}\bar F_{\mu\nu}\bar F_{\rho\sigma}$; the trace factorises into a spinor trace and a colour trace.

(5) $\operatorname{tr}(\gamma_{\mu\nu}\gamma_{\rho\sigma}) = 4(\delta_{\mu\sigma}\delta_{\nu\rho} - \delta_{\mu\rho}\delta_{\nu\sigma})$ for 4×4 Euclidean gamma matrices (from $\operatorname{tr}\gamma_\mu\gamma_\nu\gamma_\rho\gamma_\sigma = 4(\delta_{\mu\nu}\delta_{\rho\sigma} - \delta_{\mu\rho}\delta_{\nu\sigma} + \delta_{\mu\sigma}\delta_{\nu\rho})$ with $\mu\neq\nu$, $\rho\neq\sigma$), and $\operatorname{tr}(\bar F_{\mu\nu}\bar F_{\rho\sigma}) = T(r)\bar F^a_{\mu\nu}\bar F^a_{\rho\sigma}$.

(6) The first term gives $\bar F^a_{\mu\nu}\bar F^a_{\nu\mu} = -\bar F^a_{\mu\nu}\bar F^a_{\mu\nu}$.

Both are positive. Inserted in (T5.79) with the minus sign, a spin coupling *decreases* the coefficient of $\bar F^2$ for bosons: the spins line up with the field and lower its energy, which is Pauli paramagnetism. For fermions the overall sign of the determinant is reversed.

Step 6: add up. Write each contribution to Γ_1 as $\frac14 c\, L(p) \int \bar F^a \bar F^a$. A real field in a real representation gives half of a complex one; the vector a_μ has four components and the Dirac spinor four.

term of (T5.76)	origin	c	how
gluon, orbital	$\frac12 \cdot 4\,\Gamma_s(\text{adj})$	$+\frac{2N}{3}$	(T5.78), four real adjoint scalars
gluon, spin	$\frac12\cdot\big(-\frac12\big)\operatorname{tr} X^2$	$-4N$	(T5.79): $\frac12\cdot\frac12\cdot 4N\cdot 4 = 4N$
ghosts	$-\Gamma_s(\text{adj})$	$-\frac{N}{3}$	(T5.78), Grassmann sign
quarks, orbital	$-n_f\cdot\frac12\cdot 4\,\Gamma_s(\text{fund})$	$-\frac{n_f}{3}$	$2n_f\cdot\frac{T_F}{3}$ with $T_F=\frac12$
quarks, spin	$-n_f\cdot\frac12\cdot\big(-\frac12\big)\operatorname{tr} X_f^2$	$+n_f$	$\frac14\cdot 2T_F\cdot 4\,n_f$
total		$\frac{2N}{3} - 4N - \frac{N}{3} - \frac{n_f}{3} + n_f = -\big(\frac{11N}{3} - \frac{2n_f}{3}\big)$	

(In the "how" column the spin entries convert $-\frac12 L\int \operatorname{tr} X^2$ into the form $\frac14 cL\int \bar F^2$: for the gluon, $\frac12\cdot(-\frac12)\cdot 4N = -N = \frac14\cdot(-4N)$.) So

$$\Gamma[\bar A] = \frac14\Big[\frac{1}{g_0^2} - b_0\, L(p)\Big]\int \bar F^a_{\mu\nu}\bar F^a_{\mu\nu}, \qquad b_0 := \frac{11N}{3} - \frac{2n_f}{3}. \tag{T5.80}$$

In minimal subtraction the bare coupling absorbs the pole, $\frac{1}{g_0^2} = \mu^{-\varepsilon}\big[\frac{1}{g^2(\mu)} + \frac{b_0}{16\pi^2}\frac{2}{\varepsilon}\big]$, and what is left is the finite *effective coupling at momentum* p,

$$\frac{1}{g^2_{\text{eff}}(p)} = \frac{1}{g^2(\mu)} + \frac{b_0}{16\pi^2}\ln\frac{p^2}{\mu^2} + \text{const.} \tag{T5.81}$$

It cannot depend on μ, so

$$0 = \mu\frac{\mathrm{d}}{\mathrm{d}\mu}\frac{1}{g_{\text{eff}}^2} \overset{(1)}{=} -\frac{2}{g^3}\beta(g) - \frac{2b_0}{16\pi^2} \quad\Longrightarrow\quad \beta(g) \overset{(2)}{=} -\frac{b_0\, g^3}{16\pi^2}. \tag{T5.82}$$

(1) $\mu\,\mathrm{d}(g^{-2})/\mathrm{d}\mu = -2g^{-3}\beta$, and $\mu\,\partial_\mu \ln(p^2/\mu^2) = -2$.
(2) Solve for β.

Key idea: the one-loop Yang–Mills β-function

For gauge group SU(N) with n_f Dirac fermions in the fundamental,

$$\beta(g) = \mu\frac{\mathrm{d}g}{\mathrm{d}\mu} = -\frac{g^3}{16\pi^2}\left(\frac{11N}{3} - \frac{2n_f}{3}\right) + O(g^5). \tag{T5.83}$$

[Gross and Wilczek 1973, Politzer 1973] The $\frac{11}{3}$ is $4 - \frac{1}{3}$: the gluon's spin (paramagnetic, antiscreening) gives 4, its orbital motion (diamagnetic, screening, with the ghosts removing the two unphysical polarisations) gives $-\frac{1}{3}$. For a quark the signs flip and $(2s)^2 = 1$ replaces 4: $-(1 - \frac{1}{3})\cdot 2T_F = -\frac{2}{3}$ per flavour.

Physical picture: the vacuum as a paramagnet (Nielsen and Hughes)

[Nielsen 1981, Hughes 1981] In a medium, a static charge is screened if the dielectric constant $\epsilon > 1$. A relativistic vacuum must have $\epsilon\,\mu_{\mathrm{m}} = 1$ (the product sets the speed of light), so a vacuum that is *paramagnetic*, $\mu_{\mathrm{m}} > 1$, has $\epsilon < 1$: it antiscreens, and the effective charge grows with distance. The magnetic susceptibility of free charged particles has two parts: Pauli paramagnetism of the spin magnetic moment, $\propto (g_L s_z)^2$, and Landau diamagnetism of the orbits, $\propto -\frac{1}{3}$ relative to it. For a particle of spin s with $g_L = 2$, the combination is $(2s)^2 - \frac{1}{3}$, times $(-1)^{2s}$ because fermion loops have the opposite sign. Spin-1 gluons give $4 - \frac{1}{3} > 0$: the Yang–Mills vacuum is a paramagnet, and that is asymptotic freedom. Charged scalars give $-\frac{1}{3}$ and quarks $-(1 - \frac{1}{3})$: both screen. QED has no charged spin-1 particles, which is why its coupling grows at short distances.

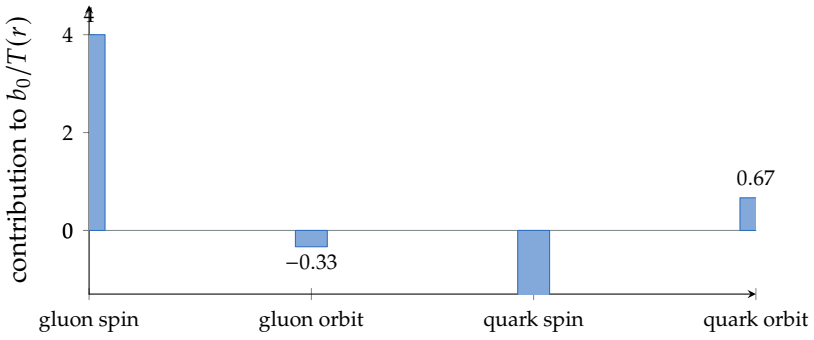


Figure T5.17: The four pieces of the one-loop coefficient b_0, each divided by the group factor $T(r)$ of the particle in the loop (N for gluons, $\frac{1}{2}$ per quark flavour). Positive bars antiscreen (drive asymptotic freedom), negative bars screen. What to see: the gluon spin term 4 dominates everything; the orbital terms are the small $\mp\frac{1}{3}$ corrections. For quarks the bars are the gluon-like values $(2s)^2 = 1$ and $-\frac{1}{3}$, reversed in sign by Fermi statistics and doubled for the four components of a Dirac spinor: -2 and $+\frac{2}{3}$, so $T_F(-2 + \frac{2}{3}) = -\frac{2}{3}$ per flavour.

Worked example T5.21: QCD numbers: α_s from M_Z to 1 TeV, and $\Lambda_{\rm QCD}$

For $N = 3$: $b_0 = 11 - \frac{2}{3}n_f$, positive for $n_f \le 16$; with all six quarks $b_0 = 7$. In terms of $\alpha_s = g^2/4\pi$,

$$\mu\frac{\mathrm{d}}{\mathrm{d}\mu}\frac{1}{\alpha_s} \overset{(1)}{=} 4\pi\,\mu\frac{\mathrm{d}}{\mathrm{d}\mu}\frac{1}{g^2} \overset{(2)}{=} 4\pi\cdot\frac{b_0}{8\pi^2} = \frac{b_0}{2\pi} \quad\Longrightarrow\quad \frac{1}{\alpha_s(Q)} = \frac{1}{\alpha_s(\mu)} + \frac{b_0}{2\pi}\ln\frac{Q}{\mu}.$$

(1) $1/\alpha_s = 4\pi/g^2$.
(2) From (T5.82), $\mu\,\mathrm{d}(g^{-2})/\mathrm{d}\mu = -2\beta/g^3 = 2b_0/16\pi^2$.

Between m_b and m_t five quarks are light, $b_0 = \frac{23}{3}$ and $b_0/2\pi = 1.220$. Starting from the measured $\alpha_s(M_Z) = 0.118$ at $M_Z = 91.2\,\mathrm{GeV}$, so $1/\alpha_s = 8.47$:

Q	$\ln(Q/M_Z)$	$1/\alpha_s(Q)$	$\alpha_s(Q)$
10 GeV	−2.21	8.47 − 2.70 = 5.77	0.173
1 TeV	+2.39	8.47 + 2.92 = 11.39	0.088

close to the measured 0.18 and 0.088. The coupling blows up where $1/\alpha_s = 0$, at $\Lambda = M_Z\exp(-2\pi/(b_0\alpha_s(M_Z))) = 91.2\,\mathrm{GeV}\times e^{-6.95} \approx 0.09\,\mathrm{GeV}$. Two-loop running and quark thresholds move this to the familiar $\Lambda_{\rm QCD} \approx 0.2\,\mathrm{GeV}$, but the order of magnitude is already right.

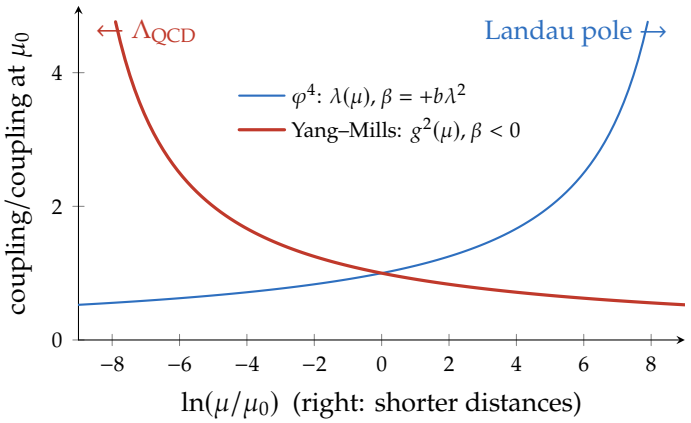


Figure T5.18: One-loop running in four dimensions, with the same strength of running 0.1 per e-fold for both. Blue: φ^4 theory, $\lambda(\mu) = \lambda_0/(1 - 0.1\ln(\mu/\mu_0))$, grows towards short distances and has a Landau pole in the ultraviolet. Red: Yang–Mills, $g^2(\mu) = g_0^2/(1 + 0.1\ln(\mu/\mu_0))$, shrinks towards short distances (asymptotic freedom) and blows up in the infrared, at the scale Λ of dimensional transmutation. What to see: the two curves are mirror images, and the sign of β alone decides at which end the theory is weakly coupled.

Worked example T5.22: Dimensional transmutation and the instanton factor

For g^2 the flow is $\mathrm{d}g^2/\mathrm{d}t = -\mu\,\mathrm{d}g^2/\mathrm{d}\mu = -2g\beta = \frac{b_0}{8\pi^2}g^4$, which is Example T5.28 with $b = b_0/8\pi^2$. So the RG-invariant mass is

$$\Lambda = \mu\, e^{-8\pi^2/(b_0\, g^2(\mu))}. \tag{T5.84}$$

Check that it does not depend on μ:

$$\mu\frac{\mathrm{d}\ln\Lambda}{\mathrm{d}\mu} \overset{(1)}{=} 1 - \frac{8\pi^2}{b_0}\,\mu\frac{\mathrm{d}}{\mathrm{d}\mu}\frac{1}{g^2} \overset{(2)}{=} 1 - \frac{8\pi^2}{b_0}\cdot\frac{b_0}{8\pi^2} = 0.$$

(1) Take the logarithm of (T5.84) and differentiate.
(2) $\mu\,\mathrm{d}(g^{-2})/\mathrm{d}\mu = b_0/8\pi^2$ from (T5.82).

Now raise (T5.84) to the power b_0:

$$e^{-8\pi^2/g^2(\mu)} \overset{(1)}{=} \left(\frac{\Lambda}{\mu}\right)^{b_0}.$$

(1) $(\Lambda/\mu)^{b_0} = e^{-b_0\cdot 8\pi^2/(b_0 g^2)}$.

The left side is the weight e^{-S} of a Yang–Mills instanton, whose action is $8\pi^2/g^2$ in the normalisation (T5.74) (Chapter 8). The coupling in it should be the one at the instanton's own scale, $\mu = 1/\rho$ for an instanton of size ρ. So the tunnelling amplitude of a size-ρ instanton is $(\rho\Lambda)^{b_0}$, which is $(\rho\Lambda)^{22/3}$ for pure SU(2): small instantons are exponentially rare, large ones are not suppressed at all, and the dilute instanton gas breaks down at $\rho \sim 1/\Lambda$, the same scale at which the coupling becomes strong.

Watch out: the background-field shortcut and the Ward identity

In a covariant gauge without a background, g appears in the three- and four-gluon vertices and in the quark–gluon vertex, and each renormalises differently; the β-function comes out only after combining the gluon, ghost and vertex renormalisations ($Z_g = Z_1/Z_3^{3/2}$). The background field gauge avoids this because background gauge invariance forces $Z_g Z_{\bar{A}}^{1/2} = 1$: the coupling renormalises only through the normalisation of $\bar{A}$, which is the single coefficient b_0 computed above. The answer (T5.83) is the same in both methods, as it must be for a one-loop coefficient of a marginal coupling.

Recap

- The Callan–Symanzik equation $[\mu\partial_\mu + \beta\partial_\lambda + n\gamma]G^{(n)} = 0$ says that bare correlators do not know the renormalisation scale.
- Its solution by characteristics expresses $G^{(2)}(x)$ through the running coupling at $\mu = 1/|x|$ and $\exp(2\int\gamma)$; at a zero of β it is a pure power with $\Delta = \Delta_{\mathrm{cl}} + \gamma(\lambda^*)$.
- φ^4: the s, t, u bubbles give $\beta = -\varepsilon\lambda + 3\lambda^2/16\pi^2$, the same as Wilson's shell integration; $O(N)$: $(N+8)\lambda^2/48\pi^2$.
- Yang–Mills in the background-field gauge: orbital (diamagnetic) and spin (paramagnetic) parts of $\ln\det(-\bar{D}^2 + X)$ give $\beta = -\frac{g^3}{16\pi^2}(\frac{11N}{3} - \frac{2n_f}{3})$, with $\frac{11}{3} = 4 - \frac{1}{3}$.
- $\Lambda = \mu e^{-8\pi^2/(b_0 g^2)}$ and $e^{-8\pi^2/g^2(1/\rho)} = (\rho\Lambda)^{b_0}$: the instanton factor is a power of the transmuted scale.

T5.5 Fixed points are conformal field theories

A fixed point of the RG has no length scale, so it is scale invariant. Nash calls this "necessary, but insufficient" for conformal invariance [Nash p. 301]. Conformal invariance is invariance under maps that rescale lengths by a *position-dependent* factor while preserving angles. In two dimensions these maps are all holomorphic functions $z \mapsto f(z)$, infinitely many, and that infinite symmetry is what makes two-dimensional CFTs solvable (Chapter 11). So it matters whether a fixed point is merely scale invariant or fully conformal. Which extra condition

upgrades scale invariance to conformal invariance? The answer is a condition on the trace of the stress tensor. In two dimensions it always holds for unitary theories with a discrete spectrum (the Zamolodchikov–Polchinski theorem), while in higher dimensions a free unitary theory shows how it can fail. The same trace turns out to *be* the list of β-functions, the fact behind the scale anomaly of Chapter 10.

T5.5.1 Conformal maps and the stress tensor

Working definition T5.8 (conformal transformation)

A change of coordinates $x \mapsto x'$ is **conformal** if it changes the metric only by an overall factor, $\mathrm{d}x'^{\mu}\mathrm{d}x'^{\mu} = \Omega(x)^2\,\mathrm{d}x^{\mu}\mathrm{d}x^{\mu}$ [near each point it is a rotation times a stretch, and the stretch Ω may differ from point to point]. Such maps preserve angles but not lengths. Infinitesimally, $x'^{\mu} = x^{\mu} + v^{\mu}(x)$ with v a **conformal Killing vector** [a vector field whose flow stretches every small square into a bigger or smaller square, never a rectangle or a rhombus]. *Test:* v is conformal Killing exactly when

$$\partial_{\mu}v_{\nu} + \partial_{\nu}v_{\mu} = \frac{2}{d}\,(\partial\cdot v)\,\delta_{\mu\nu}. \tag{T5.85}$$

Here is where (T5.85) comes from. Under $x' = x + v$,

$$\mathrm{d}x'^{\mu}\mathrm{d}x'^{\mu} \overset{(1)}{=} (\mathrm{d}x^{\mu} + \partial_{\nu}v^{\mu}\,\mathrm{d}x^{\nu})(\mathrm{d}x^{\mu} + \partial_{\rho}v^{\mu}\,\mathrm{d}x^{\rho}) \overset{(2)}{=} \mathrm{d}x^{\mu}\mathrm{d}x^{\mu} + (\partial_{\mu}v_{\nu} + \partial_{\nu}v_{\mu})\,\mathrm{d}x^{\mu}\mathrm{d}x^{\nu} + O(v^2).$$

(1) $\mathrm{d}x'^{\mu} = \mathrm{d}x^{\mu} + \partial_{\nu}v^{\mu}\mathrm{d}x^{\nu}$ by the chain rule.
(2) Expand to first order in v and rename dummy indices.

This is $\Omega^2\mathrm{d}x\cdot\mathrm{d}x$ only if the symmetric matrix $\partial_{\mu}v_{\nu} + \partial_{\nu}v_{\mu}$ is a multiple of $\delta_{\mu\nu}$, say $c\,\delta_{\mu\nu}$. Taking the trace of both sides, $2\,\partial\cdot v = c\,d$, so $c = \frac{2}{d}\partial\cdot v$, which is (T5.85). The solutions in d dimensions are: translations $v = a$; rotations $v_{\mu} = \omega_{\mu\nu}x_{\nu}$ with ω antisymmetric (for both the left side of (T5.85) vanishes, and so does $\partial\cdot v$); the dilatation $v = x$, with $\partial_{\mu}x_{\nu} + \partial_{\nu}x_{\mu} = 2\delta_{\mu\nu} = \frac{2}{d}(d)\delta_{\mu\nu}$; and the special conformal fields $v_{\nu} = 2(b\cdot x)x_{\nu} - x^2 b_{\nu}$ for a constant vector b:

$$\begin{aligned}\partial_{\mu}v_{\nu} + \partial_{\nu}v_{\mu} &\overset{(1)}{=} \big(2b_{\mu}x_{\nu} + 2(b\cdot x)\delta_{\mu\nu} - 2x_{\mu}b_{\nu}\big) + \big(2b_{\nu}x_{\mu} + 2(b\cdot x)\delta_{\mu\nu} - 2x_{\nu}b_{\mu}\big)\\ &\overset{(2)}{=} 4(b\cdot x)\,\delta_{\mu\nu} \overset{(3)}{=} \frac{2}{d}\,(\partial\cdot v)\,\delta_{\mu\nu}.\end{aligned}$$

(1) $\partial_{\mu}(b\cdot x) = b_{\mu}$, $\partial_{\mu}x_{\nu} = \delta_{\mu\nu}$ and $\partial_{\mu}x^2 = 2x_{\mu}$; the second bracket is the first with $\mu \leftrightarrow \nu$.
(2) The terms $2b_{\mu}x_{\nu}$ and $2x_{\mu}b_{\nu}$ cancel against their partners.
(3) $\partial\cdot v = 2d\,(b\cdot x)$, computed in (T5.87) below.

In $d = 2$ write $v = v^1 + iv^2$ and $z = x^1 + ix^2$. The (11) component of (T5.85) is $2\partial_1 v^1 = \partial_1 v^1 + \partial_2 v^2$, that is $\partial_1 v^1 = \partial_2 v^2$, and the (12) component is $\partial_1 v^2 + \partial_2 v^1 = 0$. These are the Cauchy–Riemann equations for $v^1 + iv^2$ as a function of $x^1 + ix^2$. So in two dimensions *every* holomorphic $v(z)$ is conformal Killing, which is the infinite-dimensional symmetry of Chapter 11.

Working definition T5.9 (stress tensor)

The **stress tensor** $T^{\mu\nu}$ [the local density and flux of energy and momentum] is the response of the action to a position-dependent translation $\delta\varphi = v^{\mu}(x)\partial_{\mu}\varphi$ [moving the fields by an amount that varies from place to place]:

$$\delta S = \int \mathrm{d}^d x\; T^{\mu\nu}\,\partial_{\mu}v_{\nu}.$$

A constant v is a symmetry, so only ∂v can appear. For a rotation-invariant theory $T^{\mu\nu}$ can be chosen symmetric. *Recipe:* vary the action with $\delta\varphi = v^\mu \partial_\mu \varphi$, integrate by parts until only ∂v appears, read off the coefficient.

Two consequences. First, conservation: integrating by parts, $\delta S = -\int v_\nu \, \partial_\mu T^{\mu\nu}$ for v vanishing at infinity. On shell the action is stationary under *every* small change of the fields, including this one, so $\delta S = 0$ for every v, and therefore $\partial_\mu T^{\mu\nu} = 0$ on shell. Second, the recipe at work on the free scalar $S = \frac{1}{2}\int (\partial\varphi)^2$:

$$\delta S \overset{(1)}{=} \int \partial_\nu\varphi \, \partial_\nu\big(v^\mu \partial_\mu \varphi\big) \overset{(2)}{=} \int (\partial_\nu v^\mu)\, \partial_\nu\varphi \, \partial_\mu\varphi + \int v^\mu \, \partial_\mu\big(\tfrac{1}{2}(\partial\varphi)^2\big)$$
$$\overset{(3)}{=} \int (\partial_\nu v_\mu)\Big[\partial_\nu\varphi \, \partial_\mu\varphi - \tfrac{1}{2}\delta_{\mu\nu}(\partial\varphi)^2\Big].$$

(1) $\delta\frac{1}{2}(\partial\varphi)^2 = \partial_\nu\varphi \, \partial_\nu\delta\varphi$.
(2) Product rule on $\partial_\nu(v^\mu\partial_\mu\varphi)$; in the second term $\partial_\nu\varphi \, \partial_\mu\partial_\nu\varphi = \partial_\mu\big(\frac{1}{2}(\partial\varphi)^2\big)$.
(3) Integrate the second term by parts, which puts the derivative on v^μ: $-\int(\partial_\mu v^\mu)\frac{1}{2}(\partial\varphi)^2 = -\int(\partial_\nu v_\mu)\frac{1}{2}\delta_{\mu\nu}(\partial\varphi)^2$.

So $T_{\mu\nu} = \partial_\mu\varphi\partial_\nu\varphi - \frac{1}{2}\delta_{\mu\nu}(\partial\varphi)^2$, which is symmetric.

For a conformal v the change of the action involves only the trace:

$$\delta S \overset{(1)}{=} \int T^{\mu\nu}\,\tfrac{1}{2}\big(\partial_\mu v_\nu + \partial_\nu v_\mu\big) \overset{(2)}{=} \int T^{\mu\nu}\,\tfrac{1}{d}(\partial\cdot v)\,\delta_{\mu\nu} = \frac{1}{d}\int (\partial\cdot v)\, T^\mu{}_\mu.$$

(1) $T^{\mu\nu}$ is symmetric, so only the symmetric part of $\partial_\mu v_\nu$ contributes.
(2) (T5.85).

So a theory with $T^\mu{}_\mu = 0$ is invariant under every conformal map. When does scale invariance, which only needs the *integrated* trace to vanish, already imply that?

T5.5.2 Currents from vector fields

Let $T^{\mu\nu}$ be the symmetric conserved stress tensor of a Euclidean field theory in d dimensions, $\partial_\mu T^{\mu\nu} = 0$ (on shell, away from other operator insertions). For any vector field $v^\nu(x)$ build the current $J_v^\mu := T^{\mu\nu}v_\nu$. Its divergence is

$$\partial_\mu J_v^\mu \overset{(1)}{=} (\partial_\mu T^{\mu\nu})v_\nu + T^{\mu\nu}\partial_\mu v_\nu \overset{(2)}{=} \tfrac{1}{2}T^{\mu\nu}\big(\partial_\mu v_\nu + \partial_\nu v_\mu\big). \tag{T5.86}$$

(1) Product rule.
(2) Conservation kills the first term; $T^{\mu\nu}$ is symmetric, so only the symmetric part of $\partial_\mu v_\nu$ survives.

So J_v is conserved whenever $\partial_\mu v_\nu + \partial_\nu v_\mu = 0$, that is, for translations and rotations (the Killing vectors of flat space). For the other two kinds of vector field in Figure T5.19 the right side is not zero:

- *Dilatation*, $v_\nu = x_\nu$: $\partial_\mu x_\nu = \delta_{\mu\nu}$, so $\partial_\mu(T^{\mu\nu}x_\nu) = T^\mu{}_\mu$.
- *Conformal Killing vectors*, $\partial_\mu v_\nu + \partial_\nu v_\mu = \frac{2}{d}(\partial\cdot v)\delta_{\mu\nu}$ by (T5.85): then $\partial_\mu J_v^\mu = \frac{1}{d}(\partial\cdot v)\,T^\mu{}_\mu$. The typical example is the **special conformal** field $v_\nu = 2(b\cdot x)x_\nu - x^2 b_\nu$ for a constant vector b, for which

$$\partial\cdot v \overset{(1)}{=} 2(b\cdot x) + 2d\,(b\cdot x) - 2(b\cdot x) = 2d\,(b\cdot x). \tag{T5.87}$$

(1) $\partial_\mu[(b\cdot x)x_\mu] = b_\mu x_\mu + (b\cdot x)\,d$ and $\partial_\mu(x^2 b_\mu) = 2x_\mu b_\mu$.

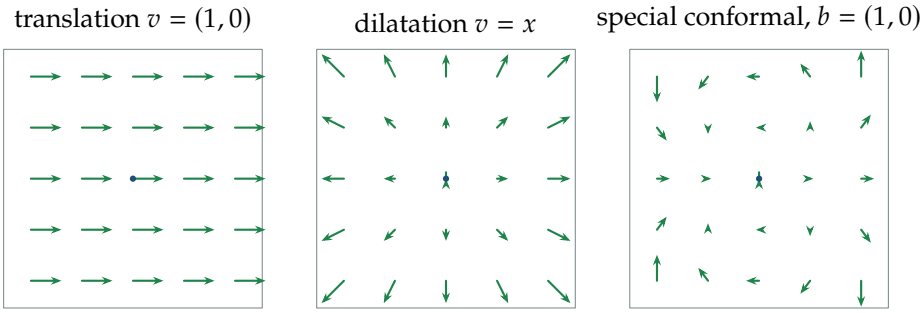


Figure T5.19: Three vector fields on the plane. Left: a translation, a Killing vector; its current $T^{\mu\nu}v_\nu$ is conserved in every theory. Middle: the dilatation $v = x$; its current has divergence $T^\mu{}_\mu$. Right: the special conformal field $v = 2(b\cdot x)x - x^2 b$ with $b = (1,0)$, whose components are $(x^2 - y^2,\, 2xy)$: the holomorphic field $v(z) = z^2$. It stretches by the position-dependent amount $\partial\cdot v = 4(b\cdot x)$, so its current has divergence $2(b\cdot x)T^\mu{}_\mu$. What to see: the middle and right fields rescale lengths, uniformly and non-uniformly; both are symmetries only if the trace of T is under control.

Working definition T5.10 (virial current)

A theory is **scale invariant** if it has a conserved dilatation current [a current whose charge generates $x \mapsto \lambda x$ and does not change in time]. The natural candidate $x_\nu T^{\mu\nu}$ has divergence $T^\mu{}_\mu$ [how the energy responds to a uniform stretch of all lengths]. It may be corrected by subtracting a local operator $K^\mu(x)$ [built from the fields at x, with no explicit x in it]:

$$D^\mu := x_\nu T^{\mu\nu} - K^\mu, \qquad \partial_\mu D^\mu = T^\mu{}_\mu - \partial_\mu K^\mu. \tag{T5.88}$$

So scale invariance means precisely $T^\mu{}_\mu = \partial_\mu K^\mu$ for some local K^μ, the **virial current** [the flux whose divergence is the trace, so that the trace integrates to zero over all space]. (The name is from Clausius' virial $\sum \vec{r}\cdot\vec{F}$, which plays the same role for a gas.) *Test:* compute $T^\mu{}_\mu$ on shell and try to write it as a divergence of a local current.

Now try to build a conserved current for a special conformal transformation. With $T^\mu{}_\mu = \partial_\mu K^\mu$,

$$\partial_\mu\big(v_\nu T^{\mu\nu}\big) \overset{(1)}{=} \tfrac{1}{d}(\partial\cdot v)\,T^\mu{}_\mu \overset{(2)}{=} 2(b\cdot x)\,\partial_\mu K^\mu \overset{(3)}{=} \partial_\mu\big(2(b\cdot x)K^\mu\big) - 2\,b_\mu K^\mu. \tag{T5.89}$$

(1) (T5.86) with the conformal Killing equation.
(2) (T5.87) and the virial current.
(3) Product rule backwards: $\partial_\mu(2(b\cdot x)K^\mu) = 2b_\mu K^\mu + 2(b\cdot x)\partial_\mu K^\mu$.

The first term can be moved into the current. The second, $-2b\cdot K$, can also be written as a divergence for every b exactly when $K_\mu = \partial_\mu L$ for some local scalar L (then $b\cdot K = \partial_\mu(b^\mu L)$). In that case $C^\mu_b = v_\nu T^{\mu\nu} - 2(b\cdot x)K^\mu + 2b^\mu L$ is conserved. So:

Key idea: when scale invariance becomes conformal invariance

A scale-invariant theory with $T^\mu{}_\mu = \partial_\mu K^\mu$ is conformally invariant if and only if the virial current is a gradient, $K_\mu = \partial_\mu L$ (up to a separately conserved current, which does not contribute to the trace). In that case the **improved** stress tensor

$$\Theta_{\mu\nu} := T_{\mu\nu} + \frac{1}{d-1}\big(\partial_\mu\partial_\nu - \delta_{\mu\nu}\partial^2\big)L \tag{T5.90}$$

is symmetric, conserved, has the same charges as $T_{\mu\nu}$, and is traceless.

The claims about (T5.90) are each one line:

$$\partial^\mu\big(\partial_\mu\partial_\nu - \delta_{\mu\nu}\partial^2\big)L \overset{(1)}{=} \partial^2\partial_\nu L - \partial_\nu\partial^2 L = 0,$$
$$\Theta^\mu{}_\mu \overset{(2)}{=} T^\mu{}_\mu + \frac{1}{d-1}\big(\partial^2 - d\,\partial^2\big)L \overset{(3)}{=} \partial_\mu K^\mu - \partial^2 L \overset{(4)}{=} 0.$$

(1) Partial derivatives commute: the added term is conserved identically, without using any equation of motion.
(2) $\delta^\mu{}_\mu = d$.
(3) $(1-d)/(d-1) = -1$ and $T^\mu{}_\mu = \partial_\mu K^\mu$.
(4) $K_\mu = \partial_\mu L$.

The charges do not change because the added term in $\Theta_{0\nu}$ is $(\partial_0\partial_\nu - \delta_{0\nu}\partial^2)L$; for $\nu = 0$ this is $-\partial_i\partial_i L$, and for $\nu = j$ it is $\partial_j(\partial_0 L)$. Both are spatial divergences, which integrate to zero over a time slice.

Worked example T5.23: The free scalar: scale invariant and conformal in every d

For $S = \frac{1}{2}\int(\partial\varphi)^2$ the canonical tensor is $T_{\mu\nu} = \partial_\mu\varphi\partial_\nu\varphi - \frac{1}{2}\delta_{\mu\nu}(\partial\varphi)^2$ (Section T5.5.1). Its trace, on shell ($\partial^2\varphi = 0$), is

$$T^\mu{}_\mu \overset{(1)}{=} \Big(1 - \frac{d}{2}\Big)(\partial\varphi)^2 \overset{(2)}{=} \frac{2-d}{4}\,\partial^2(\varphi^2) = \partial_\mu\Big[\frac{2-d}{4}\,\partial^\mu(\varphi^2)\Big],$$

(1) $\delta^\mu{}_\mu = d$.
(2) $\partial^2(\varphi^2) = 2(\partial\varphi)^2 + 2\varphi\,\partial^2\varphi = 2(\partial\varphi)^2$ on shell.

so the virial current is $K^\mu = \frac{2-d}{4}\partial^\mu(\varphi^2)$, a gradient with $L = \frac{2-d}{4}\varphi^2$. The theory is conformal in every dimension, and (T5.90) gives the improvement term $-\frac{d-2}{4(d-1)}(\partial_\mu\partial_\nu - \delta_{\mu\nu}\partial^2)\varphi^2$. In $d = 4$ the coefficient is $-\frac{1}{6}$: the classic "improved" stress tensor of Callan, Coleman and Jackiw. In $d = 2$ the trace vanishes already, without any improvement.

Worked example T5.24: Free Maxwell theory in $d \neq 4$: scale but not conformal

Take $S = \frac{1}{4}\int F_{\mu\nu}F^{\mu\nu}$ in d dimensions. The recipe of Section T5.5.1 applied to A_μ gives a tensor that is not gauge invariant, so use instead the metric definition of the stress tensor, $\delta S = \frac{1}{2}\int\sqrt{g}\,T_{\mu\nu}\,\delta g^{\mu\nu}$ (introduced in Section T5.5.3; for the scalar it reproduces the same tensor). With $S = \frac{1}{4}\int\sqrt{g}\,g^{\mu\alpha}g^{\nu\beta}F_{\mu\nu}F_{\alpha\beta}$,

$$\delta S \overset{(1)}{=} \frac{1}{4}\int\sqrt{g}\Big[2\,\delta g^{\mu\alpha}\,F_{\mu\nu}F_\alpha{}^\nu - \tfrac{1}{2}g_{\mu\alpha}\,\delta g^{\mu\alpha}\,F^2\Big] \overset{(2)}{=} \frac{1}{2}\int\sqrt{g}\Big[F_{\mu\rho}F_\nu{}^\rho - \tfrac{1}{4}g_{\mu\nu}F^2\Big]\delta g^{\mu\nu},$$

(1) $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ does not involve the metric. The two inverse metrics give two equal terms, and $\delta\sqrt{g} = -\frac{1}{2}\sqrt{g}\,g_{\mu\nu}\delta g^{\mu\nu}$.
(2) Take out $\frac{1}{2}$ and rename indices.

so in flat space $T_{\mu\nu} = F_{\mu\rho}F_\nu{}^\rho - \frac{1}{4}\delta_{\mu\nu}F^2$, which is gauge invariant. On shell ($\partial_\mu F^{\mu\nu} = 0$),

$$T^\mu{}_\mu \overset{(1)}{=} \Big(1 - \frac{d}{4}\Big)F^2 \overset{(2)}{=} \frac{4-d}{2}\,\partial_\mu\big(A_\nu F^{\mu\nu}\big),$$

(1) $F_{\mu\rho}F^{\mu\rho} = F^2$ and $\delta^\mu{}_\mu = d$.

(2) $\partial_\mu(A_\nu F^{\mu\nu}) = (\partial_\mu A_\nu)F^{\mu\nu} + A_\nu \partial_\mu F^{\mu\nu} = \frac{1}{2}(\partial_\mu A_\nu - \partial_\nu A_\mu)F^{\mu\nu} = \frac{1}{2}F^2$, using the equation of motion and the antisymmetry of $F^{\mu\nu}$.

So the theory is scale invariant with virial current $K^\mu = \frac{4-d}{2}A_\nu F^{\mu\nu}$. Is it a gradient $\partial^\mu L$ of a local *gauge-invariant* scalar? $T^\mu{}_\mu$ has dimension d, so $\partial_\mu K^\mu$ does too, $[K] = d-1$, and L would need dimension $d-2$. But $[A] = \frac{d-2}{2}$ and $[F] = \frac{d}{2}$, and every gauge-invariant local scalar is built from F and its derivatives, the lowest being F^2 of dimension d. There is none of dimension $d-2$. So for $d \neq 4$ the virial current cannot be removed and the free Maxwell field is scale invariant but *not* conformally invariant. For $d > 4$ the representation theory of the conformal group gives an independent check: in a unitary CFT a primary 2-form must have $\Delta \geq d-2$ (the unitarity bound, Mack 1977), and $\Delta_F = \frac{d}{2} < d-2$ when $d > 4$. (In $d = 3$ the photon is dual to a free scalar, $F_{\mu\nu} = \epsilon_{\mu\nu\rho}\partial^\rho\phi$, and the theory sits inside the conformal free-scalar theory; the subtlety there is the gauge non-invariance of K^μ, not unitarity.)

Watch out: what the Maxwell example does and does not show

Maxwell theory in $d \neq 4$ is unitary, so it shows that "unitary + scale invariant $\Rightarrow$ conformal" is false as stated in general dimension. The loophole is that its virial current is not gauge invariant, so it is not a genuine local operator of the theory. Theories that are scale but not conformally invariant with a genuine virial current are known only when unitarity (reflection positivity) fails; the standard example is the Riva–Cardy theory of elasticity, $\int\left[\frac{1}{4}(\partial_\mu v_\nu - \partial_\nu v_\mu)^2 + \frac{\alpha}{2}(\partial\cdot v)^2\right]$ for a displacement vector v_μ with no gauge symmetry, which is conformal only for $\alpha = (d-4)/d$ (Riva and Cardy 2005). Scale invariance for every α is quick to see. Any Lagrangian $\mathcal{L}$ quadratic in the first derivatives of free fields v_ρ has the conserved canonical tensor $T_c^{\mu\nu} = \pi^{\mu\rho}\partial^\nu v_\rho - \delta^{\mu\nu}\mathcal{L}$, with $\pi^{\mu\rho} := \partial\mathcal{L}/\partial(\partial_\mu v_\rho)$, and on shell

$$T_c{}^\mu{}_\mu \overset{(1)}{=} \pi^{\mu\rho}\partial_\mu v_\rho - d\,\mathcal{L} \overset{(2)}{=} (2-d)\,\mathcal{L} \overset{(3)}{=} \frac{2-d}{2}\,\partial_\mu\big(\pi^{\mu\rho}v_\rho\big).$$

(1) $\delta^\mu{}_\mu = d$.
(2) Euler's theorem for a function of degree 2 in ∂v: $\pi^{\mu\rho}\partial_\mu v_\rho = 2\mathcal{L}$.
(3) $\partial_\mu(\pi^{\mu\rho}v_\rho) = (\partial_\mu\pi^{\mu\rho})v_\rho + \pi^{\mu\rho}\partial_\mu v_\rho = 0 + 2\mathcal{L}$, using the equation of motion $\partial_\mu\pi^{\mu\rho} = 0$.

So $\partial_\mu(x_\nu T_c^{\mu\nu}) = T_c{}^\mu{}_\mu$ is a total derivative, and a conserved dilatation current exists. (For the free scalar, $\pi^\mu = \partial^\mu\varphi$, this gives $\frac{2-d}{4}\partial^2\varphi^2$, as in Example T5.23.) Whether the virial current can be made a gradient is decided by the symmetric stress tensor, and that is where α matters.

T5.5.3 The trace of the stress tensor measures the β-functions

Away from a fixed point the trace is not a total derivative, and it is worth knowing exactly what it is. Near a fixed point write $S = S^* + \sum_i g^i \int \mathrm{d}^d x\, \mathcal{O}_i$, with $\mathcal{O}_i$ of scaling dimension Δ_i at the fixed point, so that g^i has mass dimension $y_i = d - \Delta_i$. Define the stress tensor equivalently as the response to the metric, $\delta S = \frac{1}{2}\int \sqrt{g}\, T_{\mu\nu}\,\delta g^{\mu\nu}$ (for the free scalar this reproduces the tensor of Section T5.5.1). A uniform stretch of all lengths, $g_{\mu\nu} \to (1+2\sigma)g_{\mu\nu}$, has $\delta g^{\mu\nu} = -2\sigma g^{\mu\nu}$, so $\delta S = \frac{1}{2}\int T_{\mu\nu}(-2\sigma\delta^{\mu\nu}) = -\sigma\int T^\mu{}_\mu$. On the other hand the fixed-point action does not respond (it has no scale), the volume element grows, $\sqrt{g} = \sqrt{\det g} \to (1+2\sigma)^{d/2}\sqrt{g} = (1+d\sigma)\sqrt{g}$, and each operator shrinks by its dimension, $\mathcal{O}_i \to (1-\Delta_i\sigma)\mathcal{O}_i$:

$$-\sigma\int T^\mu{}_\mu \overset{(1)}{=} \sigma\sum_i (d-\Delta_i)\,g^i\int \mathcal{O}_i \quad\Longrightarrow\quad T^\mu{}_\mu \overset{(2)}{=} -\sum_i y_i\,g^i\,\mathcal{O}_i \overset{(3)}{=} -\sum_i \beta^i(g)\,\mathcal{O}_i + \partial_\mu K^\mu. \tag{T5.91}$$

(1) Collect the factors of $1+\sigma$ from the measure and from each operator, to first order in σ.
(2) The coefficient of each $\int \mathcal{O}_i$ must agree; $d-\Delta_i = y_i$.
(3) $y_i g^i$ is the classical β-function (T5.6). In the quantum theory, stretching all lengths by $1+\sigma$ is one RG step of RG time σ: the theory with couplings g in the stretched metric is the theory with couplings $g+\sigma\beta(g)$ in the old one, by the definition (T5.5) of β. So the change of the action is $\sigma\sum_i \beta^i \int \mathcal{O}_i$, and the full β^i replaces $y_i g^i$ (this is the operator form of the Callan–Symanzik equation). Terms whose integral vanishes, a divergence $\partial_\mu K^\mu$ and contact terms, are not fixed by this argument and may be added.

At a fixed point all $\beta^i = 0$ and only the virial term is left, which is the starting point of Section T5.5.2.

Worked example T5.25: The massive scalar and the scale anomaly of Yang–Mills

Massive scalar. For $S = \int \sqrt{g}\,[\frac{1}{2}g^{\mu\nu}\partial_\mu\varphi\partial_\nu\varphi + \frac{1}{2}m^2\varphi^2]$,

$$\delta S \overset{(1)}{=} \int \sqrt{g}\,\tfrac{1}{2}\partial_\mu\varphi\partial_\nu\varphi\,\delta g^{\mu\nu} - \int \tfrac{1}{2}\sqrt{g}\,g_{\mu\nu}\delta g^{\mu\nu}\left[\tfrac{1}{2}(\partial\varphi)^2 + \tfrac{1}{2}m^2\varphi^2\right]$$
$$\overset{(2)}{=} \frac{1}{2}\int \sqrt{g}\left[\partial_\mu\varphi\partial_\nu\varphi - g_{\mu\nu}\left(\tfrac{1}{2}(\partial\varphi)^2 + \tfrac{1}{2}m^2\varphi^2\right)\right]\delta g^{\mu\nu},$$

(1) Vary the inverse metric in the kinetic term, and use $\delta\sqrt{g} = -\frac{1}{2}\sqrt{g}\,g_{\mu\nu}\delta g^{\mu\nu}$.
(2) Take out $\frac{1}{2}$.

so in flat space $T_{\mu\nu} = \partial_\mu\varphi\partial_\nu\varphi - \delta_{\mu\nu}\left(\frac{1}{2}(\partial\varphi)^2 + \frac{1}{2}m^2\varphi^2\right)$ (at $m=0$ this is the tensor found in Section T5.5.1). On shell, $\partial^2\varphi = m^2\varphi$, so $\partial^2(\varphi^2) = 2(\partial\varphi)^2 + 2m^2\varphi^2$, and

$$T^\mu{}_\mu \overset{(1)}{=} \left(1-\frac{d}{2}\right)(\partial\varphi)^2 - \frac{d}{2}m^2\varphi^2 \overset{(2)}{=} \frac{2-d}{4}\,\partial^2(\varphi^2) - m^2\varphi^2.$$

(1) $\delta^\mu{}_\mu = d$.
(2) Replace $(\partial\varphi)^2 = \frac{1}{2}\partial^2(\varphi^2) - m^2\varphi^2$; then $-\frac{2-d}{2}m^2 - \frac{d}{2}m^2 = -m^2$.

The last term is $-y\,g\,\mathcal{O}$ with $\mathcal{O}=\varphi^2$, $g=\frac{1}{2}m^2$, $y=2$, exactly as (T5.91) predicts; the first is a virial term.

Yang–Mills. Write $S = \frac{1}{4g^2}\int F^a_{\mu\nu}F^a_{\mu\nu}$ in $d=4$. Classically F^2 has dimension 4, so the theory is scale invariant ($y=0$). Quantum mechanically the coupling $\kappa = 1/g^2$ of $\mathcal{O} = \frac{1}{4}F^aF^a$ runs, with $\beta_\kappa = \mathrm{d}(1/g^2)/\mathrm{d}t = -\frac{2}{g^3}\frac{\mathrm{d}g}{\mathrm{d}t}$. So (T5.91) gives

$$T^\mu{}_\mu \overset{(1)}{=} -\beta_\kappa\,\tfrac{1}{4}F^aF^a \overset{(2)}{=} \frac{1}{2g^3}\frac{\mathrm{d}g}{\mathrm{d}t}\,F^a_{\mu\nu}F^a_{\mu\nu}$$

(1) (T5.91) with the single coupling κ; there is no classical term since $y=0$.
(2) $-\left(-\frac{2}{g^3}\right)\cdot\frac{1}{4} = \frac{1}{2g^3}$.

(the overall sign depends on the sign convention for T). A classically scale invariant theory has a non-zero trace because the coupling runs: the **scale (trace) anomaly**. Insert the one-loop running of SU(N) Yang–Mills, $\mathrm{d}g/\mathrm{d}t = \frac{11N}{3}\frac{g^3}{16\pi^2}$ (asymptotic freedom: g grows towards the infrared):

$$T^\mu{}_\mu \overset{(1)}{=} \frac{1}{2g^3}\cdot\frac{11N}{3}\frac{g^3}{16\pi^2}\,F^aF^a \overset{(2)}{=} \frac{11N}{96\pi^2}\,F^aF^a.$$

(1) The previous line with the one-loop $\mathrm{d}g/\mathrm{d}t$.
(2) g^3 cancels and $2\cdot3\cdot16 = 96$.

Unlike the chiral anomaly, whose coefficient is fixed at one loop (the Adler–Bardeen theorem), the trace anomaly follows the full β-function; Chapter 10 compares the two.

The Yang–Mills trace anomaly in particle-physics form. The example used the infrared-ward flow $\mathrm{d}g/\mathrm{d}t$ and the normalisation with $1/g^2$ in front of the action. The formula usually quoted uses $\beta(g) = \mu\,\mathrm{d}g/\mathrm{d}\mu$ of (T5.83) and canonically normalised fields $A = gA_c$ (Section T5.4.3). Then $F = gF_c$ exactly, because the commutator term of F_c carries one factor of g, and

$$T^{\mu}{}_{\mu} \overset{(1)}{=} \frac{1}{2g^3}\frac{\mathrm{d}g}{\mathrm{d}t}\,F^aF^a \overset{(2)}{=} -\frac{\beta(g)}{2g^3}\,F^aF^a \overset{(3)}{=} -\frac{\beta(g)}{2g}\,F^a_cF^a_c \overset{(4)}{=} \frac{b_0\,g^2}{32\pi^2}\,F^a_cF^a_c. \tag{T5.92}$$

(1) The Yang–Mills part of Example T5.25.
(2) t increases towards the infrared, so $\mathrm{d}g/\mathrm{d}t = -\mu\,\mathrm{d}g/\mathrm{d}\mu = -\beta$ (the pitfall in Section T5.1.4).
(3) $F^a = gF^a_c$, so $F^aF^a = g^2F^a_cF^a_c$.
(4) $\beta = -b_0g^3/16\pi^2$ from (T5.83), so $-\beta/2g = b_0g^2/32\pi^2$. At $n_f = 0$, $b_0g^2F_cF_c/32\pi^2 = \frac{11N}{96\pi^2}F^aF^a$, the value found in the example.

Quark masses. A quark mass m multiplies $\bar\psi\psi$, of dimension 3, so $y = 1$ and the classical part of (T5.91) is $-m\bar\psi\psi$, exactly as for $\frac12 m^2\varphi^2$ above. The mass also runs. Write $\mu\,\mathrm{d}m/\mathrm{d}\mu = -\gamma_m\,m$, which defines the mass anomalous dimension γ_m. The coupling that enters (T5.91) is the dimensionless $\hat m = m/\mu$, and

$$\frac{\mathrm{d}\ln\hat m}{\mathrm{d}t} \overset{(1)}{=} -\mu\frac{\mathrm{d}}{\mathrm{d}\mu}\big(\ln m - \ln\mu\big) \overset{(2)}{=} \gamma_m + 1.$$

(1) $t = -\ln\mu$ and $\ln\hat m = \ln m - \ln\mu$.
(2) $\mu\,\mathrm{d}\ln m/\mathrm{d}\mu = -\gamma_m$ and $\mu\,\mathrm{d}\ln\mu/\mathrm{d}\mu = 1$.

So each quark adds $-(1+\gamma_m)\,m\,\bar\psi\psi$ to $T^{\mu}{}_{\mu}$.

Minkowski signature. The tensor of this chapter is Euclidean, with $\delta S = \frac12\int\sqrt{g}\,T_{\mu\nu}\delta g^{\mu\nu}$. The physical tensor $\theta_{\mu\nu}$ in signature $(+,-,-,-)$ has θ_{00} = energy density. For the scalar of the example, with Euclidean time $\tau = it$ and $V = \frac12 m^2\varphi^2$,

$$T_{\tau\tau} \overset{(1)}{=} (\partial_\tau\varphi)^2 - \tfrac12(\partial_\tau\varphi)^2 - \tfrac12(\nabla\varphi)^2 - V \overset{(2)}{=} -\tfrac12(\partial_t\varphi)^2 - \tfrac12(\nabla\varphi)^2 - V \overset{(3)}{=} -\theta_{00}.$$

(1) The flat-space $T_{\mu\nu}$ of the example at $\mu = \nu = \tau$.
(2) $\partial_\tau = -i\partial_t$, so $(\partial_\tau\varphi)^2 = -(\partial_t\varphi)^2$.
(3) $\theta_{00} = \frac12(\partial_t\varphi)^2 + \frac12(\nabla\varphi)^2 + V$ is the energy density.

The same substitution gives $T_{ij} = \theta_{ij}$ for the spatial components. The Euclidean trace is $T_{\tau\tau} + \sum_i T_{ii} = -\theta_{00} + \sum_i \theta_{ii}$, and the Minkowski trace is $\theta^{\mu}{}_{\mu} = \theta_{00} - \sum_i \theta_{ii}$. So $\theta^{\mu}{}_{\mu} = -T^{\mu}{}_{\mu}$. Both tensors come from the same metric variation, so this holds for every field. Finally $F^a_{c\,\mu\nu}F^{a\,\mu\nu}_c$ (Euclidean) equals $F^a_{c\,\mu\nu}F^{a\,\mu\nu}_c = 2(\vec B^2 - \vec E^2)$ (Minkowski), since $A_\tau = -iA_0$ makes $F_{\tau i}F_{\tau i} = -F_{0i}F_{0i}$.

Key idea: the QCD trace anomaly

In Minkowski signature, with θ^{00} the energy density and canonically normalised fields,

$$\theta^{\mu}{}_{\mu} = \frac{\beta(g)}{2g}\,F^a_{c\,\mu\nu}F^{a\,\mu\nu}_c + \sum_q(1+\gamma_m)\,m_q\,\bar\psi_q\psi_q, \qquad \frac{\beta(g)}{2g} = -\frac{b_0\,\alpha_s}{8\pi} + O(\alpha_s^2), \tag{T5.93}$$

with $\alpha_s = g^2/4\pi$ and $b_0 = \frac{11N}{3} - \frac{2n_f}{3}$ [Collins, Duncan, Joglekar 1977]. A theory with no scale in its classical action acquires one, and the operator that measures it is F^2 times the β-function. Chapter 10 compares it with the chiral anomaly.

Worked example T5.26: where the proton mass and the vacuum energy come from

Proton mass. Normalise one-particle states as $\langle p'|p\rangle = 2E(2\pi)^3\delta^3(\vec{p}-\vec{p}')$. The four-momentum is $P^\nu = \int \mathrm{d}^3x\, \theta^{0\nu}$, and $\theta(x) = e^{iPx}\theta(0)e^{-iPx}$. Then

$$p^\nu\, 2E(2\pi)^3\delta^3(\vec{p}-\vec{p}') \overset{(1)}{=} \langle p'|P^\nu|p\rangle \overset{(2)}{=} \int \mathrm{d}^3x\; e^{i(\vec{p}-\vec{p}')\cdot\vec{x}}\,\langle p'|\theta^{0\nu}(0)|p\rangle$$
$$\overset{(3)}{=} (2\pi)^3\delta^3(\vec{p}-\vec{p}')\,\langle p|\theta^{0\nu}(0)|p\rangle.$$

(1) $P^\nu|p\rangle = p^\nu|p\rangle$ and the normalisation.
(2) Translate $\theta^{0\nu}(x)$ to the origin. At equal times the phase is $e^{i(\vec{p}-\vec{p}')\cdot\vec{x}}$.
(3) $\int \mathrm{d}^3x\, e^{i\vec{k}\cdot\vec{x}} = (2\pi)^3\delta^3(\vec{k})$.

So $\langle p|\theta^{0\nu}|p\rangle = 2Ep^\nu$. The only symmetric tensors built from p are $p^\mu p^\nu$ and $M^2 g^{\mu\nu}$, so $\langle p|\theta^{\mu\nu}|p\rangle = a\,p^\mu p^\nu + b\,M^2 g^{\mu\nu}$. The component $\nu = i$ gives $2Ep^i = aEp^i$, so $a = 2$. The component $\nu = 0$ gives $2E^2 = 2E^2 + bM^2$, so $b = 0$. Taking the trace with $p^\mu p_\mu = M^2$,

$$2M^2 \overset{(1)}{=} \langle p|\theta^\mu{}_\mu|p\rangle \overset{(2)}{=} \frac{\beta(g)}{2g}\,\langle p|F^a_c F^a_c|p\rangle \qquad (m_q = 0).$$

(1) $\langle p|\theta^{\mu\nu}|p\rangle = 2p^\mu p^\nu$ contracted with $g_{\mu\nu}$.
(2) (T5.93) with massless quarks.

With massless quarks the entire mass of the proton is the matrix element of the anomaly. Classically the theory has no scale, and M must be a pure number times the transmuted scale Λ of (T5.84). The physical u, d masses add the quark term, the pion–nucleon sigma term, which is of order 50 MeV, about 5% of M [Gasser, Leutwyler, Sainio 1991].
Vacuum energy. Lorentz invariance of the vacuum forces $\langle 0|\theta^{\mu\nu}|0\rangle = \epsilon\, g^{\mu\nu}$, and θ^{00} gives ϵ as the energy density. With $g^\mu{}_\mu = 4$ and massless quarks,

$$\epsilon \overset{(1)}{=} \tfrac{1}{4}\langle\theta^\mu{}_\mu\rangle \overset{(2)}{=} \frac{\beta(g)}{8g}\,\langle F^a_c F^a_c\rangle \overset{(3)}{=} -\frac{b_0}{32}\left\langle\frac{\alpha_s}{\pi}F^a_c F^a_c\right\rangle.$$

(1) Trace of $\epsilon\, g^{\mu\nu}$.
(2) (T5.93).
(3) $\beta/8g = -b_0 g^2/128\pi^2$ and $g^2/4\pi^2 = \alpha_s/\pi$.

The gluon condensate, the non-perturbative part of $\langle\frac{\alpha_s}{\pi}F_cF_c\rangle$, is about $0.012\ \mathrm{GeV}^4$ [Shifman, Vainshtein, Zakharov 1979]. With $n_f = 3$ light flavours, $b_0 = 9$ and $\epsilon \approx -\frac{9}{32}\cdot 0.012\ \mathrm{GeV}^4 \approx -(0.24\,\mathrm{GeV})^4$. The confining vacuum lies below the perturbative one. This energy difference is the bag constant of bag models of hadrons.

T5.5.4 Two dimensions: the Zamolodchikov–Polchinski argument

In two dimensions the upgrade is a theorem. Use complex coordinates $z = x^1 + ix^2$, $\partial = \frac{1}{2}(\partial_1 - i\partial_2)$, $\bar\partial = \frac{1}{2}(\partial_1 + i\partial_2)$, metric $\mathrm{d}s^2 = \mathrm{d}z\,\mathrm{d}\bar z$, so $g_{z\bar z} = \frac{1}{2}$ and $g^{z\bar z} = 2$. Define

$$T := -2\pi T_{zz}, \qquad \Theta := -2\pi T_{z\bar z} = -\frac{\pi}{2}\,T^\mu{}_\mu. \tag{T5.94}$$

The factor -2π is the standard normalisation of two-dimensional CFT (used throughout Chapter 11): at a conformal fixed point $T = T(z)$ is holomorphic, and its operator product with itself is $T(w)T(z) \sim \frac{c/2}{(w-z)^4} + \cdots$, where the number c, the **central charge**, counts degrees of freedom ($c = 1$ for a free boson, $c = \frac{1}{2}$ for a free Majorana fermion). The second form of

Θ uses $T^{\mu}{}_{\mu} = g^{z\bar z}T_{z\bar z} + g^{\bar z z}T_{\bar z z} = 4T_{z\bar z}$. Away from a fixed point T is no longer holomorphic; conservation now reads

$$0 = \partial^{\mu}T_{\mu z} \overset{(1)}{=} g^{z\bar z}\partial_{\bar z}T_{zz} + g^{\bar z z}\partial_z T_{\bar z z} \overset{(2)}{=} 2\big(\bar\partial T_{zz} + \partial T_{z\bar z}\big) \quad\Longrightarrow\quad \bar\partial T + \partial\Theta \overset{(3)}{=} 0. \tag{T5.95}$$

(1) Raise the index with the inverse metric, whose only entries are $g^{z\bar z} = g^{\bar z z}$.
(2) $g^{z\bar z} = 2$, and T is symmetric.
(3) Multiply by $-\pi$ and use (T5.94).

At a conformal point $\Theta = 0$ and this says $\bar\partial T = 0$: T is holomorphic (Nash's (11.13)).

The three two-point functions. Rotation $z \mapsto e^{i\alpha}z$ acts on T as a spin-2 object ($T \mapsto e^{-2i\alpha}T$, since T_{zz} carries two lower z indices) and leaves the scalar Θ unchanged. Translation and rotation invariance therefore force

$$\langle T(z,\bar z)T(0)\rangle = \frac{F(z\bar z)}{z^4}, \qquad \langle T(z,\bar z)\Theta(0)\rangle = \frac{G(z\bar z)}{z^3\bar z}, \qquad \langle \Theta(z,\bar z)\Theta(0)\rangle = \frac{H(z\bar z)}{z^2\bar z^2}, \tag{T5.96}$$

with F, G, H functions of $r^2 = z\bar z$ only: under the rotation each prefactor z^{-4}, $z^{-3}\bar z^{-1}$, $z^{-2}\bar z^{-2}$ picks up exactly the phase of the operators ($e^{-4i\alpha}$, $e^{-2i\alpha}$, 1), so F, G, H must be invariant. (Also $\langle\Theta(z)T(0)\rangle = \langle T(z)\Theta(0)\rangle$: translate by $-z$ and note $(-z)^3(-\bar z) = z^3\bar z$.) At a conformal fixed point $F = \frac{c}{2}$ and $G = H = 0$.

What conservation says about F, G, H. Put (T5.95) inside correlators with $T(0)$ and with $\Theta(0)$, at $z \neq 0$. Write $\dot X := \mathrm{d}X/\mathrm{d}\ln(z\bar z) = u\,\mathrm{d}X/\mathrm{d}u$ with $u = z\bar z$, and note $\bar\partial X(u) = zX'(u)$, $\partial X(u) = \bar z X'(u)$.

$$\bar\partial\frac{F}{z^4} = -\partial\frac{G}{z^3\bar z} \overset{(1)}{=} \frac{F'}{z^3} = -\frac{G'}{z^3} + \frac{3G}{z^4\bar z} \overset{(2)}{=} \dot F = -\dot G + 3G, \tag{T5.97}$$

$$\bar\partial\frac{G}{z^3\bar z} = -\partial\frac{H}{z^2\bar z^2} \overset{(3)}{=} \frac{G'}{z^2\bar z} - \frac{G}{z^3\bar z^2} = -\frac{H'}{z^2\bar z} + \frac{2H}{z^3\bar z^2} \overset{(4)}{=} \dot G - G = -\dot H + 2H. \tag{T5.98}$$

(1) $\bar\partial(Fz^{-4}) = zF'z^{-4}$; $\partial(Gz^{-3}\bar z^{-1}) = \bar z G' z^{-3}\bar z^{-1} - 3Gz^{-4}\bar z^{-1}$.
(2) Multiply by z^3, which gives $F' = -G' + 3G/u$, then by $u = z\bar z$, using $uF' = \dot F$.
(3) $\bar\partial(Gz^{-3}\bar z^{-1}) = zG'z^{-3}\bar z^{-1} - Gz^{-3}\bar z^{-2}$; $\partial(Hz^{-2}\bar z^{-2}) = \bar z H' z^{-2}\bar z^{-2} - 2Hz^{-3}\bar z^{-2}$.
(4) Multiply by $z^3\bar z^2$, which turns $G'z\bar z$ into $\dot G$.

These two equations are the whole content of energy–momentum conservation for the two-point functions.

Key idea: Zamolodchikov and Polchinski: scale invariance implies conformal invariance in 2D

Assume a two-dimensional Euclidean field theory that is (i) local, with a symmetric conserved $T_{\mu\nu}$; (ii) reflection positive (unitary after continuation); (iii) scale invariant; (iv) with a discrete spectrum of scaling dimensions, in particular no operator of dimension zero other than the identity. Then $\Theta = 0$ as an operator, and the theory is conformally invariant.

Here is the argument, in three steps.

Step 1: F, G, H are constants. In a scale-invariant theory the dilatations act on local operators, and a conserved $T_{\mu\nu}$ has engineering dimension $d = 2$. It could fail to be a scaling operator only by mixing with another symmetric, identically conserved operator of the same dimension, and the only such operators are the improvement terms $(\partial_\mu\partial_\nu - \delta_{\mu\nu}\partial^2)\mathcal{O}$ of (T5.90). These have dimension $2 + \Delta_{\mathcal O}$, so they need $\Delta_{\mathcal O} = 0$, which (iv) excludes (Polchinski 1988). So T and Θ have

dimension exactly 2, and each two-point function in (T5.96) obeys $\langle X(\lambda x)Y(0)\rangle = \lambda^{-4}\langle X(x)Y(0)\rangle$. For the first one,

$$\frac{F(\lambda^2 z\bar z)}{\lambda^4 z^4} \overset{(1)}{=} \lambda^{-4}\,\frac{F(z\bar z)}{z^4} \quad\Longrightarrow\quad F(\lambda^2 u) \overset{(2)}{=} F(u) \text{ for all } \lambda > 0,$$

(1) Replace $z \to \lambda z$, $\bar z \to \lambda \bar z$ on the left, and use the scaling law on the right.
(2) Cancel $\lambda^{-4} z^{-4}$, with $u = z\bar z$.

so F is a constant; the same argument applies to G and H.

Step 2: $G = H = 0$. With $\dot F = \dot G = \dot H = 0$, (T5.97) gives $0 = 3G$, and then (T5.98) gives $0 = -G + 0 = 2H$. So

$$\langle \Theta(x)\Theta(0)\rangle = 0 \qquad \text{for all } x \neq 0.$$

Step 3: $\Theta = 0$. Take x along the Euclidean time direction, $x = (\tau, 0)$ with $\tau > 0$, and insert a complete set of energy eigenstates $|n\rangle$ (with $H|0\rangle = 0$). Since $\Theta(\tau) = e^{H\tau}\Theta(0)e^{-H\tau}$ and Θ is Hermitian,

$$0 = \langle \Theta(\tau)\Theta(0)\rangle \overset{(1)}{=} \sum_n \langle 0|\Theta(0)|n\rangle\langle n|\Theta(0)|0\rangle\, e^{-E_n\tau} \overset{(2)}{=} \sum_n \big|\langle n|\Theta(0)|0\rangle\big|^2 e^{-E_n\tau}. \tag{T5.99}$$

(1) Insert $\sum_n |n\rangle\langle n| = 1$; $e^{-H\tau}|0\rangle = |0\rangle$ and $\langle 0|e^{H\tau} = \langle 0|$, while $e^{-H\tau}$ acting on $|n\rangle$ gives $e^{-E_n\tau}$.
(2) $\langle 0|\Theta|n\rangle = \overline{\langle n|\Theta^\dagger|0\rangle} = \overline{\langle n|\Theta|0\rangle}$.

Every term is non-negative, so every term vanishes: $\langle n|\Theta|0\rangle = 0$ for all n, that is, $\Theta|0\rangle = 0$. A local operator that annihilates the vacuum is zero (the Reeh–Schlieder theorem; physically, a local operator acting on the vacuum always creates *something* unless it is the zero operator). So $\Theta = T^\mu{}_\mu \cdot (-\pi/2) = 0$, $\bar\partial T = 0$, and the full infinite-dimensional (Virasoro) symmetry of two-dimensional CFT follows (Chapter 11).

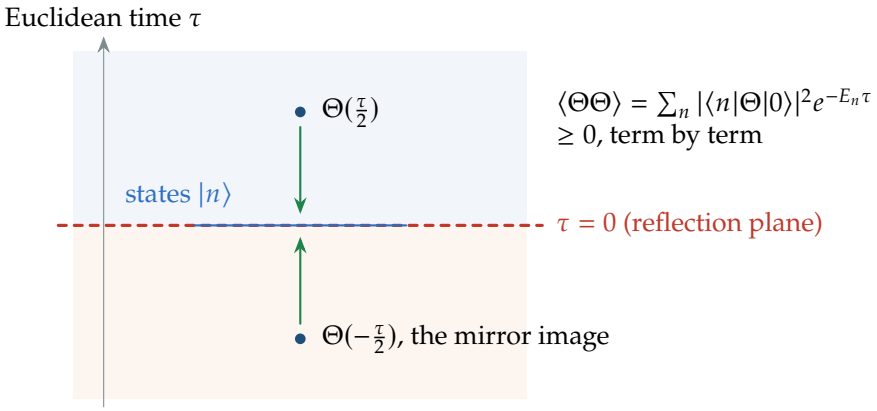


Figure T5.20: Reflection positivity. A two-point function of a Hermitian operator at Euclidean times $\pm\tau/2$ is the norm of the state $e^{-H\tau/2}\Theta|0\rangle$ that the upper operator creates on the slice $\tau = 0$; the lower operator is its mirror image. What to see: such a correlator is a sum of squares, so it can vanish only if the operator annihilates the vacuum. This is the step that needs unitarity, and it fails for the non-unitary Riva–Cardy theory.

The role of each assumption. Unitarity enters only in step 3. The discrete spectrum enters only in step 1. When a dimension-zero operator $\mathcal{O}$ is allowed, T can mix with $(\partial\partial - \delta\partial^2)\mathcal{O}$, the correlator of $\mathcal{O}$ with itself grows like $\ln r$ instead of being a power, and F, G, H acquire powers of $\ln r$. Then $\dot F, \dot G, \dot H \neq 0$, the conservation equations no longer force $H = 0$, and the theorem fails (Nakayama 2013, §5.3.1). This is not academic: the string worldsheet theory for a

non-compact target space has such operators (and is not unitary either), which is why string theory must check conformal, not just scale, invariance [Nash Ch. 9] (Chapter 9).

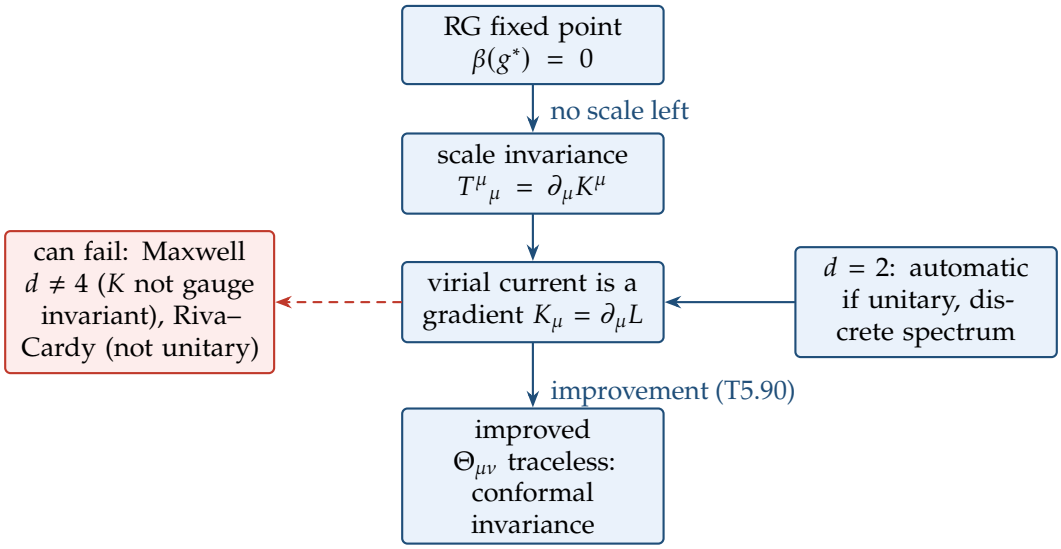


Figure T5.21: The logic of this section. Each downward arrow is one step of Section T5.5.2. The box on the right is the Zamolodchikov–Polchinski theorem; the red box on the left lists the ways the middle step can fail. What to see: the only non-trivial condition between a fixed point and a CFT is whether the virial current is a gradient.

Higher dimensions. In $d = 4$, Luty, Polchinski and Rattazzi (2012) and Fortin, Grinstein and Stergiou showed that scale invariance implies conformal invariance in perturbation theory, using the four-dimensional a-function of Section 11.13.6; non-perturbatively the question is open, and no unitary interacting counterexample is known in any d. For the fixed points met in practice (critical lattice models, Wilson–Fisher points, minimal models) conformal invariance is established.

Reading Nash (p. 301, Ch. XI §1)

"A necessary, but insufficient, condition for a theory to possess conformal invariance is that the renormalisation group flow has a fixed point." The fixed point gives scale invariance. The extra condition is that the virial current be a gradient (Section T5.5.2), and in two dimensions, for the unitary theories with discrete spectra that Nash has in mind (critical statistical models, unitary minimal models), it is automatic (Section T5.5.4). So for Nash's purposes the "insufficient" condition is in fact sufficient; in string theory, where the worldsheet theory is not unitary, it is not.

Recap

- For a vector field v, $\partial_\mu(T^{\mu\nu}v_\nu) = \frac{1}{2}T^{\mu\nu}(\partial_\mu v_\nu + \partial_\nu v_\mu)$: dilatations and conformal maps see only the trace $T^\mu{}_\mu$.
- Scale invariance $\Leftrightarrow T^\mu{}_\mu = \partial_\mu K^\mu$ (virial current). Conformal invariance $\Leftrightarrow$ in addition $K_\mu = \partial_\mu L$; then $\Theta_{\mu\nu} = T_{\mu\nu} + \frac{1}{d-1}(\partial_\mu\partial_\nu - \delta_{\mu\nu}\partial^2)L$ is traceless.
- Free scalar: $L = \frac{2-d}{4}\varphi^2$, conformal in all d. Free Maxwell in $d \neq 4$: $K^\mu = \frac{4-d}{2}A_\nu F^{\mu\nu}$, scale but not conformal.

- In 2D, conservation gives $\dot{F} = -\dot{G} + 3G$, $\dot{G} - G = -\dot{H} + 2H$; scale invariance makes F, G, H constant, so $G = H = 0$; reflection positivity then forces $\Theta = 0$.

T5.6 Quantum field theories as flows between conformal field theories

Fixed points of the RG are scale invariant, and (under the conditions of Section T5.5) conformal. Near a fixed point the flow is controlled by a finite list of relevant operators. Far from it, irrelevant couplings are slaved to relevant ones (Section T5.2.6). Now turn the picture around. Instead of asking what the RG does to a given theory, ask what a quantum field theory *is* from the RG point of view. The answer is short ([QFT notes §15.9.5]): *a quantum field theory is an RG trajectory, starting at a conformal field theory in the ultraviolet and ending at another one (possibly empty) in the infrared*.

T5.6.1 A QFT is a trajectory

Working definition T5.11 (relevant deformation of a CFT)

Start from a conformal field theory $\mathrm{CFT}_{\mathrm{UV}}$ [a theory with no length scale, such as a free massless field or a critical point] and add relevant operators [operators $\mathcal{O}_i$ of scaling dimension $\Delta_i < d$, so that their couplings grow towards long distances]:

$$S = S_{\mathrm{CFT}_{\mathrm{UV}}} + \sum_i \lambda_i \int \mathrm{d}^d x \, \mathcal{O}_i(x), \qquad [\lambda_i] = y_i = d - \Delta_i > 0. \tag{T5.100}$$

The result is a **relevant deformation** [a QFT whose short-distance limit is the CFT we started from]. Each λ_i carries a positive mass dimension y_i, so λ_i^{1/y_i} is a mass and it defines a length $\ell_i = \lambda_i^{-1/y_i}$. At distances $r \ll \ell_i$ the deformation is invisible; at $r \gg \ell_i$ it dominates. *Test:* the correlation functions at $r \to 0$ are those of $\mathrm{CFT}_{\mathrm{UV}}$; the RG trajectory leaves the UV fixed point along the relevant directions of Figure T5.11.

Where does the trajectory end? There are only two possibilities for a unitary theory at long distances.

- *A gapped phase.* All correlations die exponentially beyond some length ξ. The infrared fixed point is "empty" (a trivial theory, $c = 0$ in two dimensions) or a **topological field theory** [a theory with no local degrees of freedom whose correlators depend only on topology]. The latter is how Chern–Simons theory arises in Chapter 12.
- *A massless phase.* The trajectory ends at a non-trivial $\mathrm{CFT}_{\mathrm{IR}}$, which it approaches along its *irrelevant* directions.

Limit cycles are excluded in two dimensions by the c-theorem, and in four by the a-theorem (Section 11.13 in Chapter 11).

Key idea: the RG definition of a quantum field theory

A renormalisable quantum field theory is a relevant (or marginally relevant) deformation of a UV conformal field theory, with the couplings of (T5.100) as its parameters. Its number of free parameters is the number of relevant operators allowed by the symmetries. An effective field theory is the reverse: an infrared CFT (often free) deformed by *irrelevant*

operators, valid below the scale where they become strong.

Two examples make this concrete. Massive $\lambda\varphi^4$ theory in $d = 4$ is the free massless scalar (a CFT) deformed by φ^2 ($y = 2$) and the marginally irrelevant φ^4; it flows to the empty theory beyond the Compton wavelength. In $d = 3$ the same Lagrangian flows from the free scalar in the UV, past the Wilson–Fisher point (Figure T5.8), to a gapped phase: the theory is "super-renormalisable", its UV end is free and its long-distance physics near the critical line is the 3D Ising CFT.

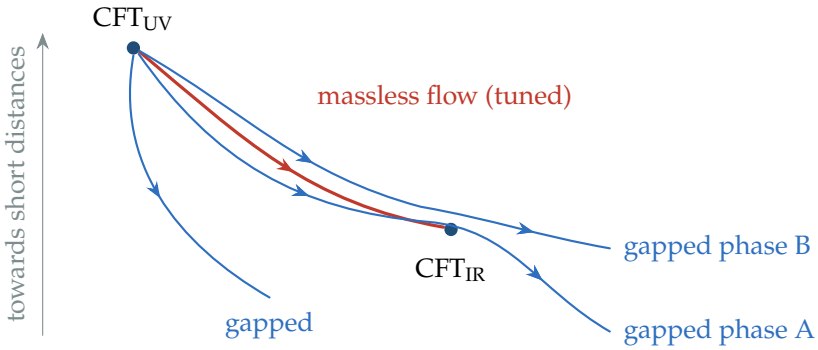


Figure T5.22: Quantum field theories as RG trajectories. Every trajectory leaves the UV fixed point along a relevant direction. The red one, tuned to lie on the critical surface of the IR fixed point, ends there: a massless theory interpolating between two CFTs. The blue ones pass close to CFT$_{\mathrm{IR}}$ (their physics at intermediate scales is governed by it) and are then pushed off into gapped phases. What to see: a QFT is the whole curve, its two ends are CFTs, and its parameters say which curve it is.

T5.6.2 Masses, correlation lengths and dimensional transmutation

With a single relevant coupling λ of dimension y, dimensional analysis alone fixes every mass in the theory up to pure numbers:

$$m = (\text{number}) \times \lambda^{1/y}, \qquad \xi \propto \lambda^{-1/y} = \lambda^{-\nu}, \qquad \nu = \frac{1}{y}. \tag{T5.101}$$

There is nothing else with the dimension of a mass. This is the continuum form of $\nu = 1/y_t$ (Section T5.3.1).

Worked example T5.27: The two-dimensional Ising field theory

Near its critical point the 2D Ising model is the $c = \frac{1}{2}$ CFT perturbed by the energy operator ε ($\Delta_\varepsilon = 1$) with coupling $\propto t$, and by the spin σ ($\Delta_\sigma = \frac{1}{8}$) with coupling $\propto h$ (these dimensions are derived in Chapter 11). By (T5.101),

$$\begin{aligned} h = 0: &\quad y_t = 2 - 1 = 1 \implies m \propto |t|, \\ t = 0: &\quad y_h = 2 - \tfrac{1}{8} = \tfrac{15}{8} \implies m \propto h^{8/15}. \end{aligned}$$

The first is Onsager's $\nu = 1$: the thermal perturbation is a free Majorana fermion of mass $\propto |t|$. The second is the celebrated $h^{8/15}$ law of the Ising model in a field at T_c, whose particle spectrum Zamolodchikov found to be organised by the exceptional Lie algebra E_8.

A *marginally* relevant coupling has $y = 0$ and defines no length by dimensional analysis. It still produces one, through its running.

Worked example T5.28: Dimensional transmutation

Let $\mathrm{d}g/\mathrm{d}t = +b\,g^2$ with $b > 0$ (towards the infrared the coupling grows; for Yang–Mills g^2 plays this role, cf. Example T5.25). Then

$$\frac{1}{g(t)} \overset{(1)}{=} \frac{1}{g_0} - bt, \qquad \Lambda_* := \Lambda\, e^{-1/(b\,g(\Lambda))},$$
$$\frac{\mathrm{d}\ln\Lambda_*}{\mathrm{d}t} \overset{(2)}{=} \frac{\mathrm{d}\ln\Lambda}{\mathrm{d}t} + \frac{1}{b\,g^2}\frac{\mathrm{d}g}{\mathrm{d}t} \overset{(3)}{=} -1 + 1 = 0.$$

(1) Separate variables as in Example T5.12, with the opposite sign.
(2) Differentiate $\ln\Lambda_* = \ln\Lambda - 1/(bg)$.
(3) $t = \ln(\Lambda_0/\Lambda)$ gives $\mathrm{d}\ln\Lambda/\mathrm{d}t = -1$; and $\mathrm{d}g/\mathrm{d}t = bg^2$.

So Λ_* is the same at every scale: an RG-invariant mass built from a dimensionless coupling. Where does the coupling become strong? Start at Λ_0 with g_0:

$$\frac{1}{g(\Lambda)} \overset{(1)}{=} \frac{1}{g_0} - b\ln\frac{\Lambda_0}{\Lambda} = 0 \quad\Longleftrightarrow\quad \Lambda \overset{(2)}{=} \Lambda_0\, e^{-1/(bg_0)} \overset{(3)}{=} \Lambda_*.$$

(1) The first line of the chain above with $t = \ln(\Lambda_0/\Lambda)$; $1/g = 0$ means $g = \infty$.
(2) Solve for $\ln(\Lambda_0/\Lambda) = 1/(bg_0)$.
(3) The definition of Λ_* evaluated at Λ_0 (it has the same value at every scale).

So the coupling blows up at $\Lambda = \Lambda_*$ (at one loop; in practice it becomes of order one near there), and every mass of the theory is a pure number times Λ_*. In QCD this is $\Lambda_{\mathrm{QCD}} \approx 200\,\mathrm{MeV}$, which sets the proton mass. A dimensionless parameter has been traded for a scale: this is **dimensional transmutation** ([QFT notes §15.9.4]).

T5.6.3 Crossover

Between the two ends of a trajectory a correlation function interpolates. Dimensional analysis gives it the **scaling form**

$$G(r) = \frac{1}{r^{2\Delta_{\mathrm{UV}}}}\,\Phi\!\left(\frac{r}{\xi}\right), \qquad \Phi(0) = \text{const}, \tag{T5.102}$$

with Δ_{UV} the dimension at the UV fixed point and Φ a universal function. For $r \ll \xi$ we see the UV CFT; for $r \gg \xi$ we see the IR end (Φ decays exponentially in a gapped phase, or as a power $(r/\xi)^{2\Delta_{\mathrm{UV}}-2\Delta_{\mathrm{IR}}}$ if the flow ends at a CFT). A good diagnostic is the **effective dimension**

$$\Delta_{\mathrm{eff}}(r) := -\frac{1}{2}\frac{\mathrm{d}\ln G}{\mathrm{d}\ln r}, \tag{T5.103}$$

which equals Δ on any stretch where G is a pure power.

Worked example T5.29: Crossover of a free massive field in three dimensions

The propagator of a free field of mass m in $d = 3$ is the Fourier transform of $1/(p^2+m^2)$. With $r = |x|$ and spherical coordinates around the direction of x,

$$G(r) \overset{(1)}{=} \int \frac{\mathrm{d}^3p}{(2\pi)^3}\,\frac{e^{ip\cdot x}}{p^2+m^2} \overset{(2)}{=} \frac{1}{4\pi^2}\int_0^\infty \mathrm{d}p\,\frac{p^2}{p^2+m^2}\int_{-1}^{1}\mathrm{d}c\; e^{iprc}$$
$$\overset{(3)}{=} \frac{1}{2\pi^2 r}\int_0^\infty \mathrm{d}p\,\frac{p\sin pr}{p^2+m^2} \overset{(4)}{=} \frac{e^{-mr}}{4\pi r}.$$

(1) Definition of the propagator.
(2) $\mathrm{d}^3p = p^2\mathrm{d}p\,\mathrm{d}c\,\mathrm{d}\phi$ with $c = \cos\theta$; the ϕ integral gives 2π, and $2\pi/(2\pi)^3 = 1/(4\pi^2)$.
(3) $\int_{-1}^{1} e^{iprc}\mathrm{d}c = 2\sin(pr)/(pr)$.
(4) The integrand is even in p, so the integral is $\frac{1}{2}\int_{-\infty}^{\infty}$ of $\mathrm{Im}\; pe^{ipr}/(p^2+m^2)$. Close the contour in the upper half plane, where e^{ipr} decays; the pole at $p = im$ has residue $\frac{im\,e^{-mr}}{2im} = \frac{1}{2}e^{-mr}$, so $\int_{-\infty}^{\infty}\frac{pe^{ipr}}{p^2+m^2}\mathrm{d}p = 2\pi i\cdot\frac{1}{2}e^{-mr}$, and the original integral is $\frac{\pi}{2}e^{-mr}$.

It has the scaling form with $\Delta_{\mathrm{UV}} = \frac{1}{2}$ (at $m = 0$, $G = 1/(4\pi r)$), $\xi = 1/m$ and $\Phi(x) = e^{-x}/4\pi$. Its effective dimension is

$$\Delta_{\mathrm{eff}}(r) \overset{(1)}{=} -\frac{1}{2}\,r\frac{\mathrm{d}}{\mathrm{d}r}\left(-mr - \ln r - \ln 4\pi\right) \overset{(2)}{=} \frac{1}{2} + \frac{mr}{2}.$$

(1) $\ln G = -mr - \ln r - \ln 4\pi$ and $\mathrm{d}/\mathrm{d}\ln r = r\,\mathrm{d}/\mathrm{d}r$.
(2) $r\,\mathrm{d}(mr)/\mathrm{d}r = mr$ and $r\,\mathrm{d}\ln r/\mathrm{d}r = 1$.

It starts at the free-field value $\frac{1}{2}$ and grows without bound once $r \gtrsim 1/m$: the signature of a gapped phase.

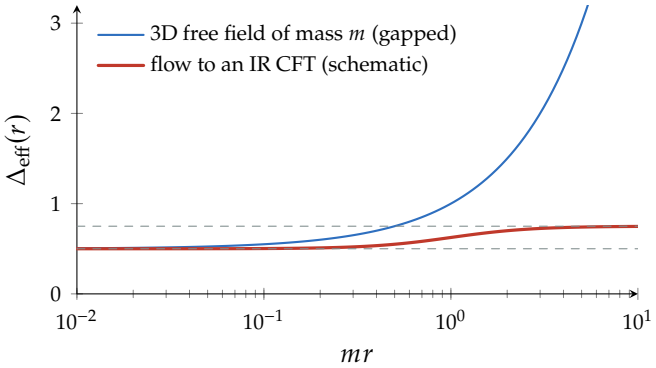


Figure T5.23: The effective dimension (T5.103) along two RG trajectories. Blue: the free massive field of Example T5.29, which starts at the UV value $\frac{1}{2}$ and then shoots up as correlations die exponentially. Red (a schematic interpolation, not a computed flow): a massless flow from $\Delta_{\mathrm{UV}} = \frac{1}{2}$ to a different IR dimension, here $\frac{3}{4}$, with a plateau at each end. What to see: the crossover happens around $r \sim \xi$, and the plateaus are the two CFTs at the ends of the trajectory.

Theory space can be a space of functions. Consider a two-dimensional sigma model

$$S = \frac{1}{4\pi\alpha'}\int \mathrm{d}^2x\, G_{\mu\nu}(X)\,\partial_a X^\mu \partial_a X^\nu.$$

Here the "coupling" is a whole metric $G_{\mu\nu}$ on the target space, one coupling for every point, and $R_{\mu\nu}$ below is its Ricci tensor. Its β-function is, at one loop and in the convention $\beta = \mu\,\mathrm{d}G/\mathrm{d}\mu$, $\beta^G_{\mu\nu} = \alpha' R_{\mu\nu}$ (Friedan 1980). It comes from the one-loop tadpole of the fluctuations of X around a point, whose coupling to the background is the curvature of G in Riemann normal coordinates; Chapter 9 takes this up. By (T5.91), with the operator $\mathcal{O} = \frac{1}{4\pi\alpha'}\partial_a X^\mu \partial_a X^\nu$ multiplying the coupling $G_{\mu\nu}$, the trace of the worldsheet stress tensor is proportional to $\beta^G_{\mu\nu}\,\partial_a X^\mu \partial_a X^\nu$. So conformal invariance of the worldsheet, which string theory requires, is $R_{\mu\nu} = 0$ at this order: *Einstein's equations in vacuum are an RG fixed-point condition*. The flow itself is $\partial_t G_{\mu\nu} = -\mu\,\mathrm{d}G_{\mu\nu}/\mathrm{d}\mu = -\alpha' R_{\mu\nu}$ (the sign flip of the pitfall in Section T5.1.4): Ricci flow.

T5.6.4 The dictionary

quantum field theory	RG / conformal field theory
renormalisable QFT	relevant deformation of a UV CFT, (T5.100)
effective field theory	IR CFT deformed by irrelevant operators
number of free parameters	number of relevant (and marginal) operators
mass m, correlation length ξ	$\lambda^{1/y}$, $\lambda^{-1/y}$, (T5.101)
running coupling $g(\mu)$	position along the trajectory
β-function	the RG vector field, (T5.5)
anomalous dimension γ	$\Delta - \Delta_{\text{canonical}}$ at a fixed point
asymptotic freedom	UV end is the Gaussian (free) fixed point
Λ_{QCD}	dimensional transmutation of a marginally relevant coupling
trace anomaly $T^{\mu}{}_{\mu} \neq 0$	$T^{\mu}{}_{\mu} = -\beta^i \mathcal{O}_i$, (T5.91)
critical point, universality class	fixed point, its CFT
critical exponents	scaling dimensions, $y = d - \Delta$
phase of matter	basin of attraction of an IR fixed point
gapped phase	trivial or topological IR theory
irreversibility of coarse graining	c- and a-theorems (Chapter 11)

Recap

- A QFT is an RG trajectory: a UV CFT deformed by relevant operators, ending in a gapped phase or at an IR CFT.
- Masses follow from dimensional analysis, $m \propto \lambda^{1/y}$; for the 2D Ising model $m \propto |t|$ and $m \propto h^{8/15}$.
- A marginally relevant coupling generates the RG-invariant scale $\Lambda_* = \Lambda e^{-1/bg(\Lambda)}$.
- Correlators cross over between the UV and IR dimensions; the effective dimension shows the plateaus.
- For a sigma model the fixed-point condition is $R_{\mu\nu} = 0$.

T5.7 Beta-functions with geometry

In Section T5.4 the couplings were numbers, λ and g. Two theories that matter later in the book have couplings of a different kind. In the two-dimensional sigma model of string theory the coupling is a whole Riemannian metric, and its β-function turns out to be the Ricci tensor (Section T5.7.1). In three-dimensional Chern–Simons theory the coupling is an integer, the level k, which cannot run at all, and yet one loop shifts it by a fixed integer (Section T5.7.2). Both answers are one-loop calculations of the kind done in Section T5.4. What is new is that the answer is a geometric object: a curvature in the first case, a spectral invariant in the second.

We keep the convention of Section T5.4: $\beta := \mu\, \mathrm{d}(\text{coupling})/\mathrm{d}\mu$, with the scale μ increasing towards the ultraviolet.

T5.7.1 The sigma model: the coupling is a metric

The string worldsheet of Chapter 9 is a surface Σ with D scalar fields $X^\mu(\sigma)$ on it, the position of the string in space-time. In flat space-time they are free fields. In a curved space-time the flat metric is replaced by a metric $G_{\mu\nu}(X)$, and the theory is no longer free. Is it still scale invariant, as string theory requires? This subsection computes the answer: only if G is Ricci flat, to first order in the curvature.

Working definition T5.12 (nonlinear sigma model)

A **nonlinear sigma model** is a field theory whose field is a map $X : \Sigma \to M$ [each point σ of the surface is sent to a point $X(\sigma)$ of a space M] from a surface Σ with metric g_{ab} into a **target** M with metric $G_{\mu\nu}(x)$ [a matrix at each point x of M that measures lengths there, and changes from point to point]. In Euclidean signature, with weight e^{-S}, the action is

$$S[X] = \frac{1}{4\pi\alpha'} \int_\Sigma \mathrm{d}^2\sigma \, \sqrt{g}\, g^{ab}\, G_{\mu\nu}(X)\, \partial_a X^\mu \partial_b X^\nu . \tag{T5.104}$$

It is the energy of the map [how much X stretches the surface, with lengths measured by G] divided by $2\pi\alpha'$, where α' is a constant of dimension length^2 (the string length squared). *How to use it:* expand X about a background map, and read the Taylor coefficients of G as coupling constants.

For $G_{\mu\nu} = \delta_{\mu\nu}$ and $T = 1/2\pi\alpha'$ this is the Polyakov action of Chapter 9. On a flat worldsheet ($g_{ab} = \delta_{ab}$) we write $\partial_a X^\mu \partial_a X^\nu$ with a summed.

Why the metric is a coupling. Choose coordinates x on M centred at a point x_0 in which $G_{\mu\nu}(x_0) = \delta_{\mu\nu}$ and $\partial_\rho G_{\mu\nu}(x_0) = 0$ (Riemann normal coordinates, Chapter T2). Put $X = x_0 + y$ and Taylor expand. The first non-trivial coefficient is the second derivative of G, whose size is $1/r_c^2$ for a curvature radius r_c. Rescale $y = \sqrt{\alpha'}\,\tilde{y}$ so that the kinetic term is $\frac{1}{4\pi}(\partial\tilde{y})^2$:

$$\begin{aligned} S &\overset{(1)}{=} \frac{1}{4\pi\alpha'} \int \mathrm{d}^2\sigma \Big[\delta_{\mu\nu} + \tfrac{1}{2}\,\partial_\alpha\partial_\beta G_{\mu\nu}(x_0)\, y^\alpha y^\beta + \dots \Big] \partial_a y^\mu \partial_a y^\nu \\ &\overset{(2)}{=} \frac{1}{4\pi} \int \mathrm{d}^2\sigma \Big[\partial_a \tilde{y}^\mu \partial_a \tilde{y}^\mu + \tfrac{1}{2}\,\alpha'\,\partial_\alpha\partial_\beta G_{\mu\nu}(x_0)\, \tilde{y}^\alpha \tilde{y}^\beta\, \partial_a \tilde{y}^\mu \partial_a \tilde{y}^\nu + \dots \Big] . \end{aligned}$$

(1) Taylor expansion of $G_{\mu\nu}(x_0 + y)$; the linear term is absent because $\partial G(x_0) = 0$, and $\partial_a X = \partial_a y$ because x_0 is constant.

(2) $y^\alpha y^\beta \partial y^\mu \partial y^\nu = \alpha'^2\, \tilde{y}^\alpha \tilde{y}^\beta \partial\tilde{y}^\mu \partial\tilde{y}^\nu$, and $\alpha'^2/\alpha' = \alpha'$.

The quartic vertex has the dimensionless strength $\alpha'\,\partial\partial G \sim \alpha'/r_c^2$. Perturbation theory in the sigma model is an expansion in (string length/curvature radius)^2. The Taylor coefficients at every point are couplings, so "the coupling" is the whole function $G_{\mu\nu}(x)$, and its β-function is a symmetric tensor field $\beta^G_{\mu\nu}(x)$ on M.

Worked example T5.30: The sphere: the $O(n)$ model

Take $M = S^{n-1}$ of radius r, written as $X = r\,\mathbf{n}$ with a unit vector $\mathbf{n} \in \mathbb{R}^n$, $\mathbf{n}\cdot\mathbf{n} = 1$. The induced metric is $G = r^2 \hat{G}$, where $\hat{G}$ is the metric of the unit sphere, and

$$S \overset{(1)}{=} \frac{r^2}{4\pi\alpha'} \int \mathrm{d}^2\sigma\, \partial_a\mathbf{n}\cdot\partial_a\mathbf{n} \overset{(2)}{=} \frac{1}{2g^2} \int \mathrm{d}^2\sigma\, \partial_a\mathbf{n}\cdot\partial_a\mathbf{n}, \qquad g^2 := \frac{2\pi\alpha'}{r^2}.$$

(1) $G_{\mu\nu}\partial_a X^\mu \partial_a X^\nu$ is the squared length of $\partial_a X = r\,\partial_a\mathbf{n}$ in $\mathbb{R}^n$.

(2) Definition of g^2: $\frac{r^2}{4\pi\alpha'} = \frac{1}{2g^2}$.

This is the $O(n)$ model of statistical mechanics (for $n = 3$, the continuum Heisenberg ferromagnet). A large sphere is weak coupling. Its Ricci tensor (sign conventions in (T5.105) below) follows from the Riemann tensor of a sphere of radius r, $R_{\rho\sigma\mu\nu} = r^{-2}(G_{\rho\mu}G_{\sigma\nu} - G_{\rho\nu}G_{\sigma\mu})$, which for $n = 3$ is the statement that the Gauss curvature is $1/r^2$ (Chapter T2):

$$R_{\sigma\nu} = G^{\rho\mu}R_{\rho\sigma\mu\nu} \overset{(1)}{=} \frac{1}{r^2}\big((n-1)G_{\sigma\nu} - G_{\sigma\nu}\big) = \frac{n-2}{r^2}\,G_{\sigma\nu} = (n-2)\,\hat{G}_{\sigma\nu}.$$

(1) $G^{\rho\mu}G_{\rho\mu} = \dim M = n - 1$ and $G^{\rho\mu}G_{\rho\nu}G_{\sigma\mu} = G_{\sigma\nu}$.

The Ricci tensor of a round sphere does not depend on its radius.

Step 1: expand along geodesics. To integrate out short-wavelength fluctuations we split X into a slowly varying background $x(\sigma)$ and a fluctuation. The naive split $X = x + y$ is coordinate dependent: y is a difference of coordinates, not a vector, and the coefficients ∂G, $\partial\partial G$ of its expansion are not tensors. The cure is to describe the fluctuation by a tangent vector $\xi(\sigma) \in T_{x(\sigma)}M$ and to reach $X(\sigma)$ by following the geodesic that leaves $x(\sigma)$ with velocity $\xi(\sigma)$ for unit time (Figure T5.24). Then every coefficient of the expansion is a tensor at x [Alvarez-Gaumé, Freedman and Mukhi 1981].

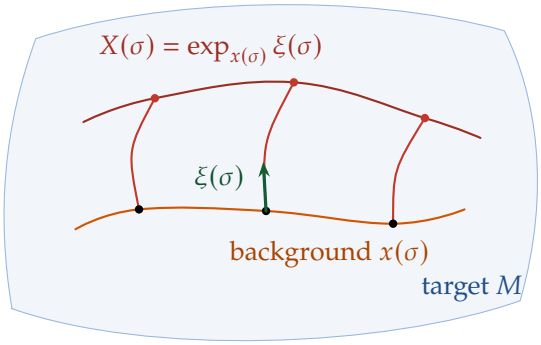


Figure T5.24: The geodesic split of the field. The orange curve is the background map $x(\sigma)$ (one line of the worldsheet drawn in the target). From each of its points a geodesic (red) leaves with initial velocity $\xi(\sigma)$ (green), a tangent vector at $x(\sigma)$; after unit time it arrives at the point $X(\sigma)$ on the full field (upper red curve). Because ξ is a vector at x, the expansion of the action in powers of ξ has tensors as coefficients, and the first one that matters is the Riemann tensor.

Write $X_s(\sigma)$ for the point reached at time s, so $X_0 = x$, $X_1 = X$, $\partial_s X_s|_{s=0} = \xi$, and $\nabla_s\partial_s X_s = 0$ (each curve $s \mapsto X_s(\sigma)$ is a geodesic). Let $E(s) = \frac{1}{2}\int \mathrm{d}^2\sigma\, G(\partial_a X_s, \partial_a X_s)$, so that $S[X_s] = E(s)/2\pi\alpha'$ on a flat worldsheet. Our curvature convention is

$$[\nabla_\mu, \nabla_\nu]W^\rho = R^\rho{}_{\sigma\mu\nu}W^\sigma, \qquad R_{\sigma\nu} = R^\rho{}_{\sigma\rho\nu}, \tag{T5.105}$$

for which a round sphere has positive Ricci tensor (Example T5.30). Along the two-parameter family $X_s(\sigma)$ the covariant derivatives ∇_s and ∇_a of vector fields along the map obey $\nabla_s\partial_a X = \nabla_a\partial_s X$ (the Levi-Civita connection has no torsion and $[\partial_s, \partial_a] = 0$) and $\nabla_s\nabla_a W - \nabla_a\nabla_s W = R(\partial_s X, \partial_a X)W$, which is (T5.105) with W^ρ carried along the map. Then

$$E'(s) \overset{(1)}{=} \int \mathrm{d}^2\sigma\, G\big(\partial_a X_s, \nabla_s\partial_a X_s\big) \overset{(2)}{=} \int \mathrm{d}^2\sigma\, G\big(\partial_a X_s, \nabla_a\partial_s X_s\big),$$

$$
\begin{aligned}
E''(s) &\overset{(3)}{=} \int \mathrm{d}^2\sigma \Big[G\big(\nabla_s \partial_a X_s, \nabla_a \partial_s X_s\big) + G\big(\partial_a X_s, \nabla_s \nabla_a \partial_s X_s\big) \Big] \\
&\overset{(4)}{=} \int \mathrm{d}^2\sigma \Big[G\big(\nabla_a \partial_s X_s, \nabla_a \partial_s X_s\big) + G\big(\partial_a X_s, R(\partial_s X_s, \partial_a X_s)\partial_s X_s\big) \Big], \\
E''(0) &\overset{(5)}{=} \int \mathrm{d}^2\sigma \Big[G_{\mu\nu} \nabla_a \xi^\mu \nabla_a \xi^\nu + R_{\mu\alpha\beta\nu}\, \xi^\alpha \xi^\beta\, \partial_a x^\mu \partial_a x^\nu \Big].
\end{aligned}
$$

(1) The metric is covariantly constant, so $\partial_s G(V,V) = 2G(V, \nabla_s V)$; the $\frac{1}{2}$ cancels the 2.
(2) No torsion: $\nabla_s \partial_a X = \nabla_a \partial_s X$.
(3) Differentiate the first line of E' again with the same product rule.
(4) In the first term use (2) again. In the second, swap ∇_s and ∇_a at the price of a curvature term; the swapped term $\nabla_a \nabla_s \partial_s X$ vanishes because the curves $s \mapsto X_s$ are geodesics.
(5) At $s = 0$: $\partial_s X = \xi$ and $X = x$. In components $G(\partial_a x, R(\xi, \partial_a x)\xi) = G_{\mu\rho} \partial_a x^\mu R^\rho{}_{\alpha\beta\nu} \xi^\alpha \xi^\beta \partial_a x^\nu$, with $R(U,V)W^\rho = R^\rho{}_{\sigma\mu\nu} W^\sigma U^\mu V^\nu$; lowering ρ gives $R_{\mu\alpha\beta\nu}$.

Here $\nabla_a \xi^\mu = \partial_a \xi^\mu + \Gamma^\mu_{\nu\rho}(x)\, \partial_a x^\nu\, \xi^\rho$ is the covariant derivative of ξ along the background. Taylor's theorem, $S[X_1] = \frac{1}{2\pi\alpha'}\big(E(0) + E'(0) + \frac{1}{2}E''(0) + \dots\big)$, gives

$$
\begin{aligned}
S[X] = S[x] &+ \frac{1}{2\pi\alpha'} \int \mathrm{d}^2\sigma\, G_{\mu\nu}\, \partial_a x^\mu \nabla_a \xi^\nu \\
&+ \frac{1}{4\pi\alpha'} \int \mathrm{d}^2\sigma \Big[G_{\mu\nu} \nabla_a \xi^\mu \nabla_a \xi^\nu + R_{\mu\alpha\beta\nu}\, \xi^\alpha \xi^\beta\, \partial_a x^\mu \partial_a x^\nu \Big] + O(\xi^3).
\end{aligned}
\tag{T5.106}
$$

On the round sphere the curvature term is $r^{-2}\big[(\xi \cdot \partial_a x)^2 - |\xi|^2\, |\partial_a x|^2\big] \le 0$, by the Riemann tensor of Example T5.30. Pushing a stretched map sideways on a sphere lowers its energy, because neighbouring geodesics converge. This focusing is what will make the sphere shrink under the RG.

Step 2: integrate out a shell. Take the background $x(\sigma)$ to contain only wavelengths longer than $1/\Lambda'$, and the fluctuation ξ only wave numbers in the shell $\Lambda' < |k| < \Lambda$, as in Wilson's calculation of Section T5.3.3. Three simplifications make the one-loop calculation short.

- The term linear in ξ is $E'(0)/2\pi\alpha'$. Integrating by parts, $E'(0) = -\int G(\nabla_a \partial_a x, \xi)$, and $\nabla_a \partial_a x^\mu = \partial^2 x^\mu + \Gamma^\mu_{\nu\rho} \partial_a x^\nu \partial_a x^\rho = 0$ is the classical equation of motion. We expand about a classical solution, so this term is absent.
- Only terms quadratic in ξ enter at one loop, because the one-loop integral is the Gaussian integral over ξ. Each extra power of ξ costs a factor $\sqrt{\alpha'}$, by the rescaling $y = \sqrt{\alpha'}\,\tilde{y}$ above.
- Use an orthonormal frame $e^i{}_\mu(x)$ [D perpendicular unit vectors at each point of M] and write $\xi^i = e^i{}_\mu \xi^\mu$. Then $G_{\mu\nu} \nabla_a \xi^\mu \nabla_a \xi^\nu = (\partial_a \xi^i + \omega_a{}^i{}_j \xi^j)^2$, where $\omega_a = \omega_\mu(x) \partial_a x^\mu$ is the spin connection pulled back to the worldsheet. It acts on ξ^i like a two-dimensional $\mathrm{SO}(D)$ gauge field, and rotating the frame is a gauge transformation of it. As in the background-field gauge of Section T5.4.3, the divergences it produces must be gauge invariant. In two dimensions the only gauge-invariant local term of dimension two is $\mathrm{tr}(\epsilon^{ab} \mathcal{F}_{ab})$, with $\mathcal{F}_{ab}$ the curvature of ω_a. It is the trace of an antisymmetric $\mathfrak{so}(D)$ matrix, which is zero. So ω produces no divergence, and the free kinetic term $\frac{1}{4\pi\alpha'}(\partial \xi^i)^2$ is all we need.

The shell average of two fluctuations at the same point is

$$
\big\langle \xi^i(\sigma) \xi^j(\sigma) \big\rangle_{\text{shell}} \overset{(1)}{=} \delta^{ij} \int_{\Lambda' < |k| < \Lambda} \frac{\mathrm{d}^2 k}{(2\pi)^2} \frac{2\pi\alpha'}{k^2} \overset{(2)}{=} \delta^{ij} \frac{2\pi\alpha'}{(2\pi)^2} 2\pi \int_{\Lambda'}^{\Lambda} \frac{\mathrm{d}k}{k} = \delta^{ij}\, \alpha' \ln \frac{\Lambda}{\Lambda'}. \tag{T5.107}
$$

(1) In Fourier space $\frac{1}{4\pi\alpha'}\int(\partial\xi)^2 = \frac{1}{2}\int\frac{\mathrm{d}^2k}{(2\pi)^2}\,\tilde\xi(-k)\,\frac{k^2}{2\pi\alpha'}\,\tilde\xi(k)$. A Gaussian $e^{-\frac{1}{2}Av^2}$ has $\langle v^2\rangle = 1/A$, so the propagator is $2\pi\alpha'/k^2$.
(2) Polar coordinates, $\mathrm{d}^2k = 2\pi k\,\mathrm{d}k$.

To first order in the curvature vertex, integrating out the shell replaces $\xi^\alpha\xi^\beta$ in (T5.106) by its average, so the action of the slow field becomes

$$\begin{aligned} S_{\Lambda'}[x] &\overset{(1)}{=} \frac{1}{4\pi\alpha'}\int \mathrm{d}^2\sigma\Big[G_{\mu\nu} + R_{\mu\alpha\beta\nu}\big\langle\xi^\alpha\xi^\beta\big\rangle_{\text{shell}}\Big]\partial_a x^\mu\partial_a x^\nu \\ &\overset{(2)}{=} \frac{1}{4\pi\alpha'}\int \mathrm{d}^2\sigma\Big[G_{\mu\nu} + \alpha'\ln\frac{\Lambda}{\Lambda'}\,G^{\alpha\beta}R_{\mu\alpha\beta\nu}\Big]\partial_a x^\mu\partial_a x^\nu \\ &\overset{(3)}{=} \frac{1}{4\pi\alpha'}\int \mathrm{d}^2\sigma\Big[G_{\mu\nu} - \alpha'\ln\frac{\Lambda}{\Lambda'}\,R_{\mu\nu}\Big]\partial_a x^\mu\partial_a x^\nu. \end{aligned} \tag{T5.108}$$

(1) $\int\mathcal{D}\xi\,e^{-S_0-V} = e^{-\langle V\rangle+\cdots}$ times a constant, with V the curvature vertex (the cumulant expansion used in Section T5.3.3).
(2) (T5.107) with $\xi^\alpha = e_i{}^\alpha\xi^i$ and $\sum_i e_i{}^\alpha e_i{}^\beta = G^{\alpha\beta}$.
(3) $G^{\alpha\beta}R_{\mu\alpha\beta\nu} = -G^{\alpha\beta}R_{\alpha\mu\beta\nu} = -R^\beta{}_{\mu\beta\nu} = -R_{\mu\nu}$: antisymmetry in the first pair of indices, then (T5.105).

So the metric seen at scale Λ' is $G_{\mu\nu}(\Lambda') = G_{\mu\nu}(\Lambda) - \alpha'\ln(\Lambda/\Lambda')R_{\mu\nu}$. Calling the lower scale μ,

$$\mu\frac{\mathrm{d}G_{\mu\nu}}{\mathrm{d}\mu} \overset{(1)}{=} \frac{\mathrm{d}}{\mathrm{d}\ln\Lambda'}\Big[G_{\mu\nu}(\Lambda) - \alpha'(\ln\Lambda - \ln\Lambda')R_{\mu\nu}\Big] = \alpha' R_{\mu\nu}. \tag{T5.109}$$

(1) $\mu = \Lambda'$, with Λ and the bare metric held fixed; the change of $R_{\mu\nu}$ is of order α' and contributes at two loops.

Key idea: the sigma-model β-function is the Ricci tensor

For the sigma model (T5.104), at one loop,

$$\beta^G_{\mu\nu} := \mu\frac{\mathrm{d}G_{\mu\nu}}{\mathrm{d}\mu} = \alpha' R_{\mu\nu} + O(\alpha'^2). \tag{T5.110}$$

Positive curvature shrinks the target towards the infrared; negative curvature expands it. The next term is $\frac{1}{2}\alpha'^2 R_{\mu\kappa\lambda\tau}R_\nu{}^{\kappa\lambda\tau}$ [Friedan 1980].

Worked example T5.31: Asymptotic freedom of the $O(n)$ model

For the sphere of Example T5.30, $G = r^2\hat G$ and $R_{\mu\nu} = (n-2)\hat G_{\mu\nu}$. So (T5.110) is a flow of the radius alone, and of $g^2 = 2\pi\alpha'/r^2$:

$$\mu\frac{\mathrm{d}r^2}{\mathrm{d}\mu}\,\hat G_{\mu\nu} \overset{(1)}{=} \alpha'(n-2)\,\hat G_{\mu\nu} \quad\Longrightarrow\quad \mu\frac{\mathrm{d}r^2}{\mathrm{d}\mu} = \alpha'(n-2),$$
$$\mu\frac{\mathrm{d}g^2}{\mathrm{d}\mu} \overset{(2)}{=} -\frac{2\pi\alpha'}{r^4}\,\mu\frac{\mathrm{d}r^2}{\mathrm{d}\mu} \overset{(3)}{=} -\frac{(n-2)}{2\pi}\,g^4.$$

(1) $\beta^G = \mu\,\mathrm{d}(r^2\hat G)/\mathrm{d}\mu$, and $\hat G$ does not change.
(2) Differentiate $g^2 = 2\pi\alpha'/r^2$.
(3) First line, and $(2\pi\alpha'/r^2)^2 = g^4$.

For $n \geq 3$ the coupling decreases at short distances: the $O(n)$ model is asymptotically free [Polyakov 1975], like Yang–Mills theory (compare $\beta(g) = -b_0 g^3/16\pi^2$ in (T5.83)). As there,

the coupling transmutes into a scale (Example T5.28): $\mu\,\mathrm{d}(g^{-2})/\mathrm{d}\mu = (n-2)/2\pi$ integrates to

$$\Lambda_{O(n)} \overset{(1)}{=} \mu\,\exp\!\Big(-\frac{2\pi}{(n-2)\,g^2(\mu)}\Big),$$

(1) $\ln(\mu/\Lambda_{O(n)}) = \frac{2\pi}{n-2}g^{-2}(\mu)$ makes $\mu\,\mathrm{d}(g^{-2})/\mathrm{d}\mu = (n-2)/2\pi$ hold, and $\Lambda_{O(n)}$ is then μ-independent, exactly as for (T5.84).

and the model has a mass gap of order $\Lambda_{O(n)}$ (Chapter 13). For $n = 2$ the target is a circle, $R_{\mu\nu} = 0$, and the radius does not run: the free compact boson.

Worked example T5.32: The linearised check: the graviton is massless

Near flat space, $G_{\mu\nu} = \delta_{\mu\nu} + \chi_{\mu\nu}$ with χ small, the sigma model is the free string perturbed by a graviton vertex operator. Its Weyl anomaly can be computed directly from the operator product of two ∂X's, and Polchinski finds, to first order in χ, $\beta^G_{\mu\nu} = -\frac{\alpha'}{2}\big(\partial^2\chi_{\mu\nu} - \partial_\nu\partial^\omega\chi_{\mu\omega} - \partial_\mu\partial^\omega\chi_{\omega\nu} + \partial_\mu\partial_\nu\chi^\omega{}_\omega\big)$ [Polchinski, *String Theory* I, (3.7.13)]. Linearising the Ricci tensor reproduces this. To first order, $\Gamma^\rho_{\mu\nu} = \frac{1}{2}(\partial_\mu\chi^\rho{}_\nu + \partial_\nu\chi^\rho{}_\mu - \partial^\rho\chi_{\mu\nu})$, and

$$\begin{aligned} R_{\mu\nu} &\overset{(1)}{=} \partial_\rho\Gamma^\rho_{\nu\mu} - \partial_\nu\Gamma^\rho_{\rho\mu} \\ &\overset{(2)}{=} \tfrac{1}{2}\big(\partial_\rho\partial_\nu\chi^\rho{}_\mu + \partial_\rho\partial_\mu\chi^\rho{}_\nu - \partial^2\chi_{\mu\nu}\big) - \tfrac{1}{2}\partial_\nu\partial_\mu\chi^\rho{}_\rho \\ &\overset{(3)}{=} -\tfrac{1}{2}\big(\partial^2\chi_{\mu\nu} - \partial_\nu\partial^\omega\chi_{\mu\omega} - \partial_\mu\partial^\omega\chi_{\omega\nu} + \partial_\mu\partial_\nu\chi^\omega{}_\omega\big). \end{aligned}$$

(1) $R^\rho{}_{\mu\rho\nu}$ from (T5.105), $R^\rho{}_{\sigma\mu\nu} = \partial_\mu\Gamma^\rho_{\nu\sigma} - \partial_\nu\Gamma^\rho_{\mu\sigma} + \Gamma\Gamma$, dropping $\Gamma\Gamma = O(\chi^2)$.
(2) Insert Γ; in $\Gamma^\rho_{\rho\mu}$ the first and third terms cancel, leaving $\frac{1}{2}\partial_\mu\chi^\rho{}_\rho$.
(3) Rename $\rho \to \omega$ and use the symmetry of χ.

So $\alpha' R_{\mu\nu}$ is exactly the vertex-operator answer. In the gauge $\partial^\omega\chi_{\omega\mu} = \frac{1}{2}\partial_\mu\chi^\omega{}_\omega$ it is $-\frac{\alpha'}{2}\partial^2\chi_{\mu\nu}$, and $\beta^G = 0$ becomes the massless wave equation for a plane wave $\chi \propto e^{ik\cdot X}$: $k^2 = 0$, the graviton is massless.

Step 3: the Weyl anomaly and the string. String theory needs the worldsheet theory to be invariant under local rescalings $g_{ab} \to e^{2\omega(\sigma)}g_{ab}$ (Chapter 9). The general statement (T5.91) says that the trace of the stress tensor measures the β-functions. We translate it into the string normalisation, in which $T^{ab} := (4\pi/\sqrt{g})\,\delta S/\delta g_{ab}$ [Polchinski, §2.3], so that $\delta S = \frac{1}{2\pi}\int\sqrt{g}\,\delta\omega\,T^a{}_a$ for $\delta g_{ab} = 2\delta\omega\,g_{ab}$:

$$T^a{}_a \overset{(1)}{=} -2\pi\,T^\mu{}_\mu\big|_{(\mathrm{T5.91})} \overset{(2)}{=} 2\pi\sum_i \frac{\mathrm{d}g^i}{\mathrm{d}t}\,\mathcal{O}_i \overset{(3)}{=} -2\pi\sum_i \mu\frac{\mathrm{d}g^i}{\mathrm{d}\mu}\,\mathcal{O}_i \overset{(4)}{=} -\frac{1}{2\alpha'}\,\beta^G_{\mu\nu}\,g^{ab}\partial_a X^\mu\partial_b X^\nu. \quad \text{(T5.111)}$$

(1) (T5.91) uses $\delta S = \frac{1}{2}\int\sqrt{g}\,T_{ab}\,\delta g^{ab}$. Since $\delta g^{ab} = -g^{ac}g^{bd}\delta g_{cd}$, the string definition gives $\delta S = -\frac{1}{4\pi}\int\sqrt{g}\,T_{ab}\,\delta g^{ab}$; comparing, $T_{\text{string}} = -2\pi\,T$.
(2) (T5.91), $T^\mu{}_\mu = -\sum_i\beta^i\mathcal{O}_i$, whose β^i is the flow $\mathrm{d}g^i/\mathrm{d}t$ towards the infrared.
(3) $\mathrm{d}/\mathrm{d}t = -\mu\,\mathrm{d}/\mathrm{d}\mu$ (the pitfall in Section T5.1.4).
(4) The couplings are the values $G_{\mu\nu}(x)$ at every target point, and the operator multiplying $G_{\mu\nu}(X(\sigma))$ in (T5.104) is $\frac{1}{4\pi\alpha'}g^{ab}\partial_a X^\mu\partial_b X^\nu$.

So the worldsheet is Weyl invariant exactly when $\beta^G_{\mu\nu} = 0$. The string can move consistently only through a space-time whose metric is a fixed point of the sigma-model RG.

Strings also couple to the antisymmetric tensor $B_{\mu\nu}$ and to the dilaton Φ, the other two

massless closed-string fields:

$$S = \frac{1}{4\pi\alpha'}\int \mathrm{d}^2\sigma\,\sqrt{g}\Big[\big(g^{ab}G_{\mu\nu} + i\epsilon^{ab}B_{\mu\nu}\big)\partial_a X^\mu \partial_b X^\nu + \alpha' R^{(2)}\Phi(X)\Big], \tag{T5.112}$$

with $R^{(2)}$ the worldsheet curvature and $\sqrt{g}\,\epsilon^{12} = 1$. The field strength $H_{\omega\mu\nu} = \partial_\omega B_{\mu\nu} + \partial_\mu B_{\nu\omega} + \partial_\nu B_{\omega\mu}$ is invariant under $\delta B_{\mu\nu} = \partial_\mu\zeta_\nu - \partial_\nu\zeta_\mu$, which changes the action by a total derivative. The dilaton term is *classically* not Weyl invariant, and its failure is of the same order in α' as the one-loop effect of G. Under $g \to e^{2\delta\omega}g$ in two dimensions $\sqrt{g}\,R^{(2)} \to \sqrt{g}\,(R^{(2)} - 2\nabla^2\delta\omega)$ (Chapter 9), and

$$\delta S_\Phi \overset{(1)}{=} \frac{1}{4\pi}\int \mathrm{d}^2\sigma\,\sqrt{g}\,\big(-2\nabla^2\delta\omega\big)\,\Phi(X) \overset{(2)}{=} \frac{1}{2\pi}\int \mathrm{d}^2\sigma\,\sqrt{g}\,\delta\omega\,\big(-\nabla^2\Phi(X)\big),$$

$$\nabla^2\Phi(X) \overset{(3)}{=} \partial_\mu\partial_\nu\Phi\,\partial_a X^\mu\partial^a X^\nu + \partial_\rho\Phi\,\nabla^2 X^\rho \overset{(4)}{=} \nabla_\mu\nabla_\nu\Phi\,\partial_a X^\mu\partial^a X^\nu + \partial_\rho\Phi\big(\nabla^2 X^\rho + \Gamma^\rho_{\mu\nu}\partial_a X^\mu\partial^a X^\nu\big)$$
$$\overset{(5)}{=} \nabla_\mu\nabla_\nu\Phi\,\partial_a X^\mu\partial^a X^\nu + O(\alpha').$$

(1) The $\frac{1}{4\pi\alpha'}\cdot\alpha' = \frac{1}{4\pi}$ term of (T5.112) and the variation of $\sqrt{g}R^{(2)}$.
(2) Integrate ∇^2 by parts twice (the worldsheet is closed).
(3) Chain rule twice: $\partial_a\Phi(X) = \partial_\mu\Phi\,\partial_a X^\mu$, then ∇^a of that.
(4) Target covariant derivative, $\partial_\mu\partial_\nu\Phi = \nabla_\mu\nabla_\nu\Phi + \Gamma^\rho_{\mu\nu}\partial_\rho\Phi$.
(5) The bracket is the classical equation of motion of X at $B = \Phi = 0$ (step 2); the B and Φ corrections to it are of higher order.

So the dilaton adds $-\nabla_\mu\nabla_\nu\Phi\,\partial X^\mu\partial X^\nu$ to $T^a{}_a$, which in the form (T5.111) is a shift $\beta^G_{\mu\nu} \to \beta^G_{\mu\nu} + 2\alpha'\nabla_\mu\nabla_\nu\Phi$. The complete one-loop result has the same structure for all three fields.

Key idea: the string's β-functions

With the conventions of (T5.112) and $T^{ab} = (4\pi/\sqrt{g})\,\delta S/\delta g_{ab}$,

$$T^a{}_a = -\frac{1}{2\alpha'}\beta^G_{\mu\nu}g^{ab}\partial_a X^\mu\partial_b X^\nu - \frac{i}{2\alpha'}\beta^B_{\mu\nu}\epsilon^{ab}\partial_a X^\mu\partial_b X^\nu - \frac{1}{2}\beta^\Phi R^{(2)},$$

and [Polchinski, (3.7.14)]

$$\beta^G_{\mu\nu} = \alpha' R_{\mu\nu} + 2\alpha'\nabla_\mu\nabla_\nu\Phi - \tfrac{\alpha'}{4}H_{\mu\lambda\omega}H_\nu{}^{\lambda\omega} + O(\alpha'^2),$$
$$\beta^B_{\mu\nu} = -\tfrac{\alpha'}{2}\nabla^\omega H_{\omega\mu\nu} + \alpha'\nabla^\omega\Phi\,H_{\omega\mu\nu} + O(\alpha'^2),$$
$$\beta^\Phi = \frac{D-26}{6} - \tfrac{\alpha'}{2}\nabla^2\Phi + \alpha'\nabla_\omega\Phi\nabla^\omega\Phi - \tfrac{\alpha'}{24}H_{\mu\nu\lambda}H^{\mu\nu\lambda} + O(\alpha'^2).$$

Weyl invariance, $\beta^G = \beta^B = \beta^\Phi = 0$, is Einstein's equation with sources, a Maxwell equation for B, and a condition that fixes the number of dimensions. The $R_{\mu\nu}$ and $\nabla\nabla\Phi$ terms are derived above; the $\frac{D-26}{6}$ is the flat-space conformal anomaly of D scalars and the ghosts (Chapters 9 and 10); the H terms come from the same tadpole with the connection $\Gamma^\mu_{\nu\rho} \pm \frac{1}{2}H^\mu{}_{\nu\rho}$.

Worked example T5.33: The linear dilaton: $D \neq 26$ with a tilted string coupling

Take $G_{\mu\nu} = \delta_{\mu\nu}$, $B = 0$ and $\Phi(X) = V_\mu X^\mu$ with a constant vector V. Then $\nabla\nabla\Phi = 0$, so

$\beta^G = \beta^B = 0$, and

$$\beta^\Phi \overset{(1)}{=} \frac{D-26}{6} + \alpha' V_\mu V^\mu = 0 \quad \Longleftrightarrow \quad V_\mu V^\mu = \frac{26-D}{6\alpha'}.$$

(1) $\nabla^2\Phi = 0$, $\nabla_\omega \Phi = V_\omega$, $H = 0$ in the keyidea.

The background is exact (the action is still Gaussian in X), and it is a consistent string background in any D: the dilaton gradient supplies the missing central charge, $c = D + 6\alpha' V^2 = 26$ [Polchinski, (3.7.18)].

Step 4: fixed points and Ricci flow. At a fixed point with $B = 0$ and constant Φ, (T5.110) says $R_{\mu\nu} = 0$: the target is Ricci flat. Flat tori are fixed points for every choice of radii, a whole manifold of fixed points (the radii are exactly marginal couplings). Curved Ricci-flat examples are the K3 surface and the Calabi–Yau spaces of string compactification. Away from a fixed point, in the RG time $t = \ln(\mu_0/\mu)$ of (T5.5) that runs towards the infrared,

$$\frac{\partial G_{\mu\nu}}{\partial t} \overset{(1)}{=} -\mu \frac{\mathrm{d}G_{\mu\nu}}{\mathrm{d}\mu} = -\alpha' R_{\mu\nu} \qquad \Longleftrightarrow \qquad \frac{\partial G_{\mu\nu}}{\partial \tau} \overset{(2)}{=} -2R_{\mu\nu}, \quad \tau = \tfrac{1}{2}\alpha' t. \tag{T5.113}$$

(1) The pitfall in Section T5.1.4, and (T5.110).
(2) Chain rule, $\partial_t = \frac{1}{2}\alpha' \partial_\tau$.

The second form is Hamilton's **Ricci flow** [Hamilton 1982], the equation Perelman used to prove the Poincaré conjecture. The RG of the sigma model coarse-grains the geometry of its target: bumps of positive curvature shrink, regions of negative curvature grow, and flat directions stay put (Figure T5.25).

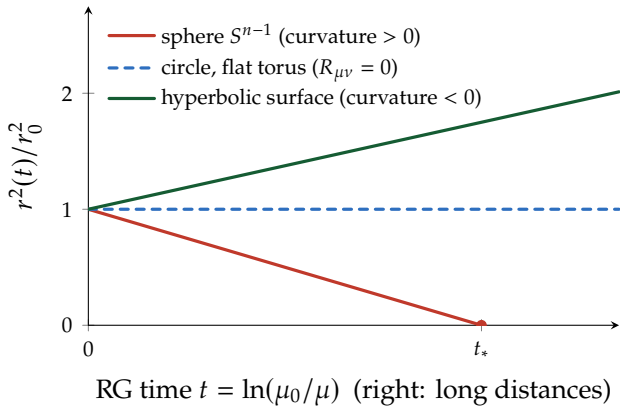


Figure T5.25: Ricci flow as the sigma-model RG, drawn for targets whose shape does not change, so only the size r flows. From Example T5.31, r^2 of a sphere decreases linearly and reaches zero at the finite RG time $t_* = r_0^2/\alpha'(n-2)$: there the coupling $g^2 = 2\pi\alpha'/r^2$ has become strong and perturbation theory stops, at the scale $\mu_0 e^{-t_*} = \Lambda_{O(n)}$. A flat target does not flow. A target of negative curvature grows, so its sigma model is weakly coupled at long distances and strongly coupled at short ones.

Worked example T5.34: The collapse time is the transmuted scale

For S^{n-1}, $\partial_t r^2 = -\mu\, \mathrm{d}r^2/\mathrm{d}\mu = -\alpha'(n-2)$ by Example T5.31, so $r^2(t) = r_0^2 - \alpha'(n-2)t$ reaches

zero at

$$t_* \overset{(1)}{=} \frac{r_0^2}{\alpha'(n-2)} \overset{(2)}{=} \frac{2\pi}{(n-2)\,g_0^2}, \qquad \mu_0 e^{-t_*} \overset{(3)}{=} \mu_0 \exp\!\left(-\frac{2\pi}{(n-2)g_0^2}\right) = \Lambda_{O(n)}.$$

(1) Solve $r^2(t_*) = 0$.
(2) $r_0^2 = 2\pi\alpha'/g_0^2$ with $g_0 = g(\mu_0)$.
(3) $t = \ln(\mu_0/\mu)$, and the formula for $\Lambda_{O(n)}$ in Example T5.31.

The finite-time singularity of Ricci flow on a round sphere and the dynamically generated mass of the $O(n)$ model are the same event. For a two-dimensional target, $R_{\mu\nu} = KG_{\mu\nu}$ with K the Gauss curvature, and the same computation with $n = 3$ shows that the area $4\pi r^2$ decreases at the constant rate $4\pi\alpha'$: the rate is $\alpha' \int K\,\mathrm{d}A = 2\pi\alpha'\chi(S^2)$ by Gauss–Bonnet (Chapter T2).

Watch out: what is and is not fixed about β^G

- β^G is defined only up to a change of coordinates on the target. A field redefinition $X^\mu \to X^\mu + \epsilon v^\mu(X)$ changes G by $\nabla_\mu v_\nu + \nabla_\nu v_\mu$ and describes the same theory, so $\beta^G_{\mu\nu}$ and $\beta^G_{\mu\nu} + \nabla_\mu v_\nu + \nabla_\nu v_\mu$ are equivalent flows. The dilaton term is of this form, $2\alpha'\nabla_\mu\nabla_\nu\Phi = \nabla_\mu v_\nu + \nabla_\nu v_\mu$ with $v = \alpha'\nabla\Phi$, which is why Perelman's version of Ricci flow carries an extra function that plays the role of the dilaton.
- Two signs are easy to lose when deriving (T5.111) in dimensional regularisation instead. In $d = 2-\varepsilon$, $\sqrt{g}\,g^{ab} \to e^{(d-2)\delta\omega}\sqrt{g}\,g^{ab}$, so the classical Weyl variation is $-\varepsilon\,\delta\omega$ times the action. And $\int \mathrm{d}^d\sigma\,\partial X\partial X$ has mass dimension $-\varepsilon$, so the bare coupling $\lambda_0 = G_0/\alpha'_0$ carries $\mu^{-\varepsilon}$, $\lambda_0 = \mu^{-\varepsilon}(\lambda + \text{poles})$, which gives $\varepsilon\lambda_0 = \beta\,\partial\lambda_0/\partial\lambda$. Together, $T^a{}_a = -2\pi\varepsilon\lambda_0 O = -2\pi\beta\,[O]$, in agreement with (T5.111). A derivative with respect to g^{ab} instead of g_{ab} in the definition of T flips the overall sign.

Recap

- The sigma-model coupling is the target metric; the dimensionless expansion parameter is α'/r_c^2.
- Expanding along geodesics, $S_2 = \frac{1}{4\pi\alpha'}\int[(\nabla\xi)^2 + R_{\mu\alpha\beta\nu}\xi^\alpha\xi^\beta\partial x^\mu\partial x^\nu]$; the shell tadpole $\langle\xi\xi\rangle = \alpha'\ln(\Lambda/\Lambda')$ gives $\beta^G_{\mu\nu} = \alpha' R_{\mu\nu}$.
- The worldsheet trace is $T^a{}_a = -\frac{1}{2\alpha'}\beta^G\partial X\partial X + \ldots$, so Weyl invariance is $R_{\mu\nu} = 0$ (plus dilaton and H terms), Einstein's equation.
- Towards the infrared the flow is Ricci flow; the sphere collapses at the transmuted scale of the asymptotically free $O(n)$ model.

T5.7.2 The Chern–Simons level shift

Chern–Simons theory on a closed oriented three-manifold M with gauge group $G = \mathrm{SU}(N)$ has the partition function

$$Z(M) = \int \mathcal{D}A\; e^{2\pi i k\,\mathrm{CS}(A)}, \qquad \mathrm{CS}(A) := -\frac{1}{8\pi^2}\int_M \mathrm{tr}\!\left(A\wedge\mathrm{d}A + \tfrac{2}{3}A\wedge A\wedge A\right), \qquad \text{(T5.114)}$$

with A an anti-Hermitian matrix-valued 1-form and tr the trace in the N-dimensional representation. This is the weight $e^{-ikS(A)/4\pi}$ of Chapter 12, since $2\pi/8\pi^2 = 1/4\pi$. A large gauge transformation changes $\mathrm{CS}(A)$ by an integer (Chapter 12), so $e^{2\pi i k\,\mathrm{CS}}$ is gauge invariant only for integer k. The coupling $1/k$ is *quantised*. It cannot run: a β-function would move it continuously. Yet every semiclassical formula of Chapter 12 for SU(2) contains $k+2$, not k. This subsection derives the shift $k \to k + h^\vee\,\mathrm{sgn}(k)$ from the phase of the one-loop determinant.

Working definition T5.13 (dual Coxeter number)

The **dual Coxeter number** $h^\vee$ of a simple Lie group is the number with

$$\mathrm{tr}_{\mathrm{adj}}(XY) = 2h^\vee\,\mathrm{tr}(XY) \qquad \text{for all } X, Y \text{ in the Lie algebra,}$$

where $\mathrm{tr}_{\mathrm{adj}}$ is the trace in the adjoint representation [the matrices by which X acts on the Lie algebra itself, $Y \mapsto [X,Y]$, the representation of the gluons] and tr the trace in the fundamental one [the $N \times N$ matrices, the representation of the quarks]. It measures how strongly the gauge field acts on itself, compared with how strongly it acts on a quark. *Recipe:* compute both traces for one generator and divide. For SU(N), $h^\vee = N$. The values for the other groups are in the table of Chapter 8.

For SU(2), Example T5.20 found $\mathrm{tr}(T^3_{\mathrm{adj}}T^3_{\mathrm{adj}}) = 2$ and $\mathrm{tr}(T^3T^3) = \frac{1}{2}$, a ratio $4 = 2h^\vee$, so $h^\vee = 2$.

Step 1: the one-loop phase. Around a flat connection A_f the gauge-fixed quadratic action of Chapter 12 packages the fluctuation a and the Lagrange multiplier ϕ into $v = (a, \phi)$, with exponent $\frac{ik}{4\pi}\langle v, D_{A_f} v\rangle$. Here

$$D_A = \begin{pmatrix} \star\mathrm{d}_A & \mathrm{d}_A \\ \mathrm{d}_A^\dagger & 0 \end{pmatrix} \qquad \text{on } \Omega^1(M,\mathfrak{g}) \oplus \Omega^0(M,\mathfrak{g}) \tag{T5.115}$$

is the self-adjoint **odd signature operator** [curl on the Lie-algebra-valued 1-forms, glued to gradient and divergence]. A single Fresnel integral has the phase $\int \mathrm{d}x\, e^{i\lambda x^2/2} = \sqrt{2\pi/|\lambda|}\, e^{i\pi\,\mathrm{sgn}(\lambda)/4}$. So

$$\text{phase of } \int \mathcal{D}v\, e^{\frac{ik}{4\pi}\langle v, Dv\rangle} \overset{(1)}{=} \text{phase of } \int \mathcal{D}v\, e^{\frac{i}{2}\langle v, \frac{k}{2\pi} Dv\rangle}$$
$$\overset{(2)}{=} \prod_j e^{\frac{i\pi}{4}\,\mathrm{sgn}(k\lambda_j)} \overset{(3)}{=} \exp\!\left(\frac{i\pi}{4}\,\mathrm{sgn}(k)\,\eta(A_f)\right). \tag{T5.116}$$

(1) $\frac{k}{4\pi} = \frac{1}{2}\cdot\frac{k}{2\pi}$.
(2) Expand v in eigenvectors of D, with eigenvalues λ_j; each mode is a Fresnel integral with $\lambda = k\lambda_j/2\pi$.
(3) $\sum_j \mathrm{sgn}(k\lambda_j) = \mathrm{sgn}(k)\sum_j \mathrm{sgn}(\lambda_j)$, regularised as the eta invariant $\eta(A) := \eta_{D_A}(0)$, $\eta_D(s) = \sum_j \mathrm{sgn}(\lambda_j)\,|\lambda_j|^{-s}$ (Chapter 12).

The ghosts contribute the determinant of the positive operator $\mathrm{d}_A^\dagger \mathrm{d}_A$, which has no phase. So each flat connection enters Z as $e^{2\pi i k\,\mathrm{CS}(A_f)}\, e^{i\pi\,\mathrm{sgn}(k)\eta(A_f)/4}$ times a positive modulus. Everything now depends on how $\eta(A)$ depends on A.

Worked example T5.35: The circle: η is linear in the connection, and jumps by 2

Chapter 12 computed the eta invariant of $-i\,\mathrm{d}/\mathrm{d}x$ on a circle with holonomy $e^{2\pi i\alpha}$. The eigenvalues are $n + \alpha$, $n \in \mathbb{Z}$, and $\eta = 1 - 2\alpha$ for $0 < \alpha < 1$. The number $\alpha = \frac{1}{2\pi i}\oint A$ is the

one-dimensional Chern–Simons functional of the U(1) connection, and like CS it is defined only modulo integers. So

$$\eta(\alpha+\delta)-\eta(\alpha)\overset{(1)}{=}-2\delta \quad (0<\alpha<\alpha+\delta<1), \qquad \lim_{\alpha\to 0^+}\eta-\lim_{\alpha\to 1^-}\eta\overset{(2)}{=}1-(-1)=2.$$

(1) $\eta=1-2\alpha$ on the interval.
(2) $\alpha=0^+$ and $\alpha=1^-$ are the same holonomy approached from the two sides. In between, the eigenvalue $\alpha-1$ has crossed zero from below.

Two features (Figure T5.26): away from zero crossings η changes linearly with the Chern–Simons functional, with a slope fixed by the local index density; at a crossing it jumps by 2. The three-dimensional statement below has the same shape, with slope $8h^\vee$.

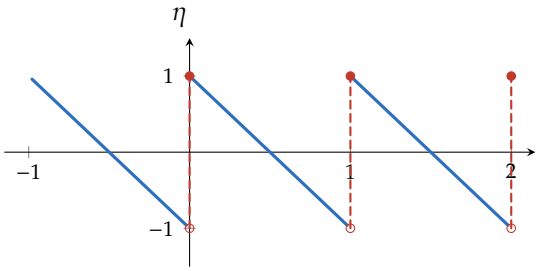


Figure T5.26: Spectral asymmetry of $-i\,\mathrm{d}/\mathrm{d}x$ on a circle as the connection is changed. Between integers, $\eta=1-2\alpha$ falls linearly (blue): this is the part fixed by the local index density. At each integer one eigenvalue crosses zero and η jumps up by 2 (red dashes), which multiplies the phase $e^{i\pi\eta/4}$ by i. The Chern–Simons case has the same shape: a linear rise $8h^\vee\,\mathrm{CS}(A)$ interrupted by jumps of ±2.

Step 2: $\eta(A)$ from the Atiyah–Patodi–Singer theorem. Recall the APS theorem (Chapter 4): if a first-order elliptic operator on a compact X has the form $\sigma\,(\partial_n+L)$ near ∂X, with n the inward normal, σ an invertible bundle map and L self-adjoint on ∂X, then with APS boundary conditions

$$\text{index}=\int_X P-\sum_{\text{boundary pieces}}\tfrac{1}{2}\big(h(L)+\eta_L(0)\big), \tag{T5.117}$$

with P the same local index density as on a closed manifold and $h(L)=\dim\ker L$. Apply it to the cylinder $X=[0,1]\times M$ (Figure T5.27), with the product metric, orientation $\mathrm{d}t\wedge\mathrm{vol}_M$, and the connection $\mathbb{A}=A_t$ that has no $\mathrm{d}t$ component and interpolates from $A_0=0$ to $A_1=A$ (constant in t near both ends). The operator is the one whose index counts instantons (Chapter 8),

$$\mathcal{D}=(-\mathrm{d}_{\mathbb{A}}^\dagger)\oplus\mathrm{d}_{\mathbb{A}}^+:\ \Omega^1(X,\mathfrak{g})\to\Omega^0(X,\mathfrak{g})\oplus\Omega^2_+(X,\mathfrak{g}),$$

where d^+ is the self-dual part of d. On a closed four-manifold its index is $4h^\vee c_2-\dim G\,(1-b_1+b_+)$, with $c_2=\frac{1}{8\pi^2}\int\mathrm{tr}(F\wedge F)$ (positive for anti-self-dual instantons); for SU(2) on S^4 this is $8k-3$. So P is $4h^\vee\cdot\frac{1}{8\pi^2}\,\mathrm{tr}(\mathbb{F}\wedge\mathbb{F})$ plus densities built from the curvature of X.

(a) The boundary operator is $\pm D$. Write a 1-form on X as $a=a_0\,\mathrm{d}t+b$, with $b(t)$ a 1-form on M. We show the computation for $\mathbb{A}=0$; with $\mathbb{A}\neq0$ every d becomes d_{A_t}. A self-dual 2-form is $\mathrm{d}t\wedge\beta+\star_M\beta$ for a 1-form β on M, because $\star(\mathrm{d}t\wedge\beta)=\star_M\beta$ in this orientation; so

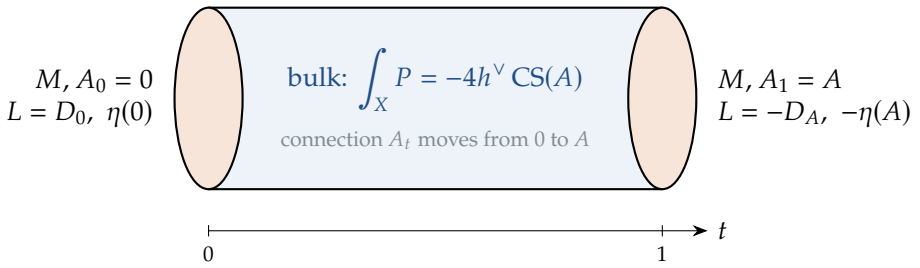


Figure T5.27: The cylinder that compares $\eta(A)$ with $\eta(0)$. The two ends are copies of M carrying the trivial connection and A. Seen from inside, the normal at the right end points backwards in t, so the boundary operator there is $-D_A$ and contributes $-\eta(A)$. The index density in between is local and integrates to a Chern–Simons functional. The index is an integer, so the difference of the two eta invariants is that local integral up to integers.

$\Omega^2_+(X) \cong \Omega^1(M)$ at each t. Then

$$
\begin{aligned}
-\mathrm{d}^\dagger a &\overset{(1)}{=} \partial_t a_0 - \mathrm{d}^\dagger_M b, \\
\mathrm{d}a &\overset{(2)}{=} \mathrm{d}t \wedge (\partial_t b - \mathrm{d}_M a_0) + \mathrm{d}_M b, \\
\mathrm{d}^+ a &\overset{(3)}{=} \tfrac{1}{2}\big[\mathrm{d}t \wedge \beta + \star_M \beta\big], \qquad \beta = \partial_t b - \mathrm{d}_M a_0 + \star_M \mathrm{d}_M b.
\end{aligned}
$$

(1) $\mathrm{d}^\dagger a = -\nabla^\mu a_\mu = -\partial_t a_0 - \nabla^i b_i$, and $-\nabla^i b_i = \mathrm{d}^\dagger_M b$.
(2) $\mathrm{d}(a_0 \mathrm{d}t) = \mathrm{d}_M a_0 \wedge \mathrm{d}t = -\mathrm{d}t \wedge \mathrm{d}_M a_0$ and $\mathrm{d}b = \mathrm{d}t \wedge \partial_t b + \mathrm{d}_M b$.
(3) For $F = \mathrm{d}t \wedge E + B$: $\star F = \star_M E + \mathrm{d}t \wedge \star_M B$, so $\frac{1}{2}(F + \star F) = \frac{1}{2}[\mathrm{d}t \wedge \beta + \star_M \beta]$ with $\beta = E + \star_M B$ (using $\star_M \star_M = 1$).

Identifying $\mathrm{d}^+ a$ with β (an invertible bundle map, the σ of (T5.117)),

$$
\mathcal{D}\begin{pmatrix} a_0 \\ b \end{pmatrix} \cong \partial_t \begin{pmatrix} a_0 \\ b \end{pmatrix} + L \begin{pmatrix} a_0 \\ b \end{pmatrix}, \qquad L = \begin{pmatrix} 0 & -\mathrm{d}^\dagger_M \\ -\mathrm{d}_M & \star_M \mathrm{d}_M \end{pmatrix}.
$$

Changing the sign of a_0 turns L into $\begin{pmatrix} 0 & \mathrm{d}^\dagger_M \\ \mathrm{d}_M & \star_M \mathrm{d}_M \end{pmatrix}$, which is D of (T5.115) with the two components of v listed in the other order. A change of basis does not change eigenvalues, so $\eta_L = \eta_D$.

(b) APS on the cylinder. At $t = 0$ the inward normal is ∂_t and $L = D_0$. At $t = 1$ it is $-\partial_t$, and $\partial_t + D_A = -(\partial_n - D_A)$, so the boundary operator is $-D_A$, whose eta invariant is $-\eta(A)$. With $h_0 = \dim \mathrm{Ker}\, D_0$ and $h_1 = \dim \mathrm{Ker}\, D_A$,

$$
\begin{aligned}
\operatorname{index} \mathcal{D} &\overset{(1)}{=} \int_X P - \tfrac{1}{2}\big(h_0 + \eta(0)\big) - \tfrac{1}{2}\big(h_1 - \eta(A)\big), \\
\tfrac{1}{2}\big(\eta(A) - \eta(0)\big) &\overset{(2)}{=} \operatorname{index} \mathcal{D} - \int_X P + \tfrac{1}{2}(h_0 + h_1).
\end{aligned}
$$

(1) (T5.117) with the two boundary operators; $\eta_{-D} = -\eta_D$.
(2) Rearrange.

(c) The bulk integral. For the product metric every curvature 2-form of X has legs only along M, and a 4-form built from them would need four independent directions in the three-dimensional M; so the gravitational part of P vanishes. The gauge part is a total derivative, $\mathrm{tr}(\mathbb{F} \wedge \mathbb{F}) = \mathrm{d}\,\mathrm{tr}(\mathbb{A} \wedge \mathrm{d}\mathbb{A} + \frac{2}{3}\mathbb{A} \wedge \mathbb{A} \wedge \mathbb{A})$ (Chapter 12), and

$$
\int_X P \overset{(1)}{=} \frac{4h^\vee}{8\pi^2} \int_X \mathrm{tr}(\mathbb{F} \wedge \mathbb{F}) \overset{(2)}{=} \frac{4h^\vee}{8\pi^2}\Big[\int_M \mathrm{tr}\big(A \wedge \mathrm{d}A + \tfrac{2}{3}A^3\big) - 0\Big] \overset{(3)}{=} -4h^\vee \mathrm{CS}(A).
$$

(1) The gravitational part vanishes.
(2) Stokes. With outward normal first, ∂X is M at $t = 1$ with its own orientation and M at $t = 0$ with the opposite one; the Chern–Simons form of $A_0 = 0$ is zero.
(3) Definition (T5.114).

(d) Put together. Steps (b) and (c) give

$$\eta(A) = \eta(0) + 8h^\vee \,\mathrm{CS}(A) + \big(2\,\mathrm{index}\,\mathcal{D} + h_0 + h_1\big). \tag{T5.118}$$

The bracket is an integer. It depends on the path from 0 to A only through integers, and $\eta(A)$ does not depend on it at all: a large gauge transformation shifts $\mathrm{CS}(A)$ by an integer m and the index by $-4h^\vee m$ (the instanton number of the path changes by $-m$), and $8h^\vee m - 2\cdot 4h^\vee m = 0$.

Key idea: the level shift

Inserting (T5.118) into the one-loop phase (T5.116),

$$e^{2\pi i k\,\mathrm{CS}(A_f)}\, e^{\frac{i\pi}{4}\,\mathrm{sgn}(k)\eta(A_f)} \overset{(1)}{=} e^{\frac{i\pi}{4}\,\mathrm{sgn}(k)\eta(0)}\; e^{\frac{i\pi}{4}\,\mathrm{sgn}(k)(2\,\mathrm{index}\,\mathcal{D}+h_0+h_1)}\; e^{2\pi i\,(k+h^\vee\,\mathrm{sgn}\,k)\,\mathrm{CS}(A_f)}.$$

(1) $\frac{i\pi}{4}\cdot 8h^\vee\,\mathrm{CS} = 2\pi i h^\vee\,\mathrm{CS}$, combined with $e^{2\pi i k\,\mathrm{CS}}$.

At one loop the level multiplying the Chern–Simons invariant of every flat connection is

$$k \;\longrightarrow\; k + h^\vee\,\mathrm{sgn}(k), \qquad k \to k+2 \;\text{ for SU(2)},\; k > 0.$$

The shift is an integer, so the renormalised weight is still gauge invariant. The other two factors do not depend on the level. The integer bracket gives an eighth root of unity for each flat connection (Witten's spectral-flow phase). $\eta(0)$ does not depend on A but depends on the metric of M; APS applied to a family of metrics shows that its variation is the gravitational Chern–Simons form, and cancelling it with a counterterm is the framing dependence of Chapter 12 [Witten 1989, §2].

Worked example T5.36: SU(2): the exact answers know about $k+2$

With $h^\vee = 2$ the one-loop answer for S^3 is a function of $k+2$. The exact result of Chapter 12 is $Z(S^3) = \sqrt{2/(k+2)}\,\sin(\pi/(k+2))$. Expanding,

$$Z(S^3) \overset{(1)}{=} \sqrt{\frac{2}{k+2}}\left(\frac{\pi}{k+2} + O\big((k+2)^{-3}\big)\right) = \sqrt{2}\,\pi\,(k+2)^{-3/2}\big(1 + O(k^{-2})\big).$$

(1) $\sin x = x - x^3/6 + \ldots$ with $x = \pi/(k+2)$.

The power $k^{-3/2}$ is the one-loop count of the three zero modes of the trivial connection, and the +2 is the level shift: at $k = 10$ the exact value is 0.1057, while $\sqrt{2}\pi(k+2)^{-3/2} = 0.1069$ and $\sqrt{2}\pi k^{-3/2} = 0.1405$. The same denominator appears in the $\mathrm{SU}(2)_k$ WZW model of Chapter 11: the central charge $c = 3k/(k+2)$ and the conformal weights $j(j+1)/(k+2)$ come from the Sugawara construction, whose normalisation is $k + h^\vee$.

Worked example T5.37: U(1): no shift

For U(1) the structure constants vanish, so the adjoint representation is trivial: $\mathrm{tr}_{\mathrm{adj}}(XY) = 0$ and $h^\vee = 0$. In (T5.118) the gauge field drops out of the index density, $\eta(A) = \eta(0)+$ integer, and the level is not shifted. This fits the fact that the abelian theory is Gaussian: CS is quadratic in A, the stationary-phase evaluation is exact, and there is nothing for a loop to renormalise except the metric-dependent phase. For SU(N) the shift is $k \to k + N$.

Physical picture: a quantised coupling can only jump

A coupling that must be an integer cannot run logarithmically, since a β-function would move it through non-integer values. What it can do is change by a finite integer when heavy modes are integrated out. If Chern–Simons theory is regularised by adding a Yang–Mills term, the gluons become massive, and integrating them out at their mass scale shifts the Chern–Simons coefficient by $h^\vee \operatorname{sgn}(k)$ [Pisarski and Rao 1985]. In the language of Section T5.6, three-dimensional Yang–Mills theory with a Chern–Simons term is a trajectory whose infrared end is a topological theory (Chapter 12), and the level shift is the one-loop threshold effect along it. Here the same number came out of a gauge-invariant regularisation (the eta invariant) with no heavy fields at all.

Watch out: which η, which power

Books write the one-loop phase as $e^{i\pi\eta/4}$, $e^{i\pi\eta/2}$ or $e^{-i\pi\eta/4}$, depending on whether η is taken for D on $\Omega^1 \oplus \Omega^0$ (as here), for the operator on all odd forms, or for $\star\mathrm{d}$ on 1-forms alone, and on the sign of the exponent. The statement that does not depend on these choices is the coefficient of $\mathrm{CS}(A)$ in the logarithm of the phase, $2\pi i h^\vee \operatorname{sgn}(k)$. A check of any convention: it must reproduce $\sqrt{2}\pi(k+2)^{-3/2}$ for $Z(S^3)$ at $G = \mathrm{SU}(2)$.

Recap

- The Chern–Simons level is quantised, so it cannot run; the one-loop Gaussian contributes the phase $e^{i\pi \operatorname{sgn}(k)\eta(A)/4}$.
- APS on $[0,1]\times M$ with the instanton-counting operator gives $\eta(A) = \eta(0) + 8h^\vee \mathrm{CS}(A) +$ integer.
- Hence $e^{2\pi i k\,\mathrm{CS}} \to e^{2\pi i(k+h^\vee \operatorname{sgn} k)\mathrm{CS}}$: $k \to k+2$ for SU(2), $k \to k+N$ for SU(N), no shift for U(1).
- The leftover $\eta(0)$ is metric dependent; removing it is the framing choice of Chapter 12.

Reading the literature

Sources. The exact RG equation, the two-coupling toy flow and the renormalisability bound follow J. Polchinski, "Renormalization and effective Lagrangians", Nucl. Phys. B231 (1984) 269. The Gaussian-smearing derivation of the same equation is in O. Rosten, "Fundamentals of the exact renormalization group", arXiv:1003.1366, §III; the exponential form also appears in [QFT notes §15.9.1, (15.9.19)]. Effective field theory from this point of view is developed in J. Polchinski, "Effective field theory and the Fermi surface", hep-th/9210046. The relation between scale and conformal invariance, including the Maxwell and Riva–Cardy examples and the two-dimensional theorem (J. Polchinski, Nucl. Phys. B303 (1988) 226), is reviewed in Y. Nakayama, "Scale invariance vs conformal invariance", arXiv:1302.0884, §§2.5, 4.1, 5.1–5.3 and 5–6.

Dictionary: notation in the original papers

Polchinski (1984)	here
weight e^S, interaction $L(\varphi,\Lambda)$	weight e^{-S}, $S^{\mathrm{I}}_\Lambda = -L$
propagator $K(p^2/\Lambda^2)/(p^2+m^2)$	$C_\Lambda(p)$, massless in most examples
$\Lambda\,\partial_\Lambda K/(p^2+m^2)$	shell propagator $\Delta_\Lambda(p)$
$Q = \Lambda^3\partial_\Lambda K/(p^2+m^2)$	$\Lambda^2\Delta_\Lambda(p)$
$\Lambda\,\mathrm{d}/\mathrm{d}\Lambda$ (towards the UV)	$\partial_t = -\Lambda\,\partial_\Lambda$ (towards the IR)
eqs. (1)–(10), (18), (20)–(21), (34), (35), (45), (47)	(T5.32)–(T5.37), (T5.39), (T5.40)–(T5.41), (T5.42), (T5.43), (T5.45), theorem

Rosten writes the flow in terms of $\dot{C} = -\Delta$. In [QFT notes Ch. 16], eqs. (16.2.24) and (16.2.84) state $\eta = 2\gamma_m(\lambda^*)$ with γ_m the anomalous dimension of φ^2; the relation (16.2.23) there shows that the correct statement is $\eta = 2\gamma(\lambda^*)$ with the field anomalous dimension (Example T5.15), while $\nu = [2-\gamma_m(\lambda^*)]^{-1}$ in (16.2.84) is correct.

Sources for Section T5.4. The Callan–Symanzik equation is C. G. Callan, Phys. Rev. D2 (1970) 1541, and K. Symanzik, Commun. Math. Phys. 18 (1970) 227. The Yang–Mills β-function is D. J. Gross and F. Wilczek, Phys. Rev. Lett. 30 (1973) 1343, and H. D. Politzer, *ibid.* 1346. The background-field method used here is L. F. Abbott, Nucl. Phys. B185 (1981) 189. The spin-versus-orbital split of $\frac{11}{3}$ is N. K. Nielsen, Am. J. Phys. 49 (1981) 1171, and R. J. Hughes, Nucl. Phys. B186 (1981) 376.

Sources for the trace anomaly (Section T5.5.3). The operator equation for $T^\mu{}_\mu$ in gauge theories is J. C. Collins, A. Duncan and S. D. Joglekar, Phys. Rev. D16 (1977) 438. The pion–nucleon σ-term is from J. Gasser, H. Leutwyler and M. E. Sainio, Phys. Lett. B253 (1991) 252. The gluon condensate is from M. A. Shifman, A. I. Vainshtein and V. I. Zakharov, Nucl. Phys. B147 (1979) 385.

Sources for Section T5.7. The sigma-model β-function is D. Friedan, Phys. Rev. Lett. 45 (1980) 1057, with the full treatment in Ann. Phys. 163 (1985) 318. The $O(N)$ case is A. M. Polyakov, Phys. Lett. 59B (1975) 79. The covariant (Riemann-normal-coordinate) expansion is L. Alvarez-Gaumé, D. Z. Freedman and S. Mukhi, Ann. Phys. 134 (1981) 85. The Weyl-anomaly route and the string β-functions follow J. Polchinski, *String Theory* I, §3.7, eqs. (3.7.13)–(3.7.18). Ricci flow is R. S. Hamilton, J. Diff. Geom. 17 (1982) 255, and G. Perelman, arXiv:math/0211159. The level shift is in E. Witten, Commun. Math. Phys. 121 (1989) 351, §2. It uses the η-invariant theorem of M. F. Atiyah, V. K. Patodi and I. M. Singer, Math. Proc. Camb. Phil. Soc. 77 (1975) 43. The perturbative shift is R. D. Pisarski and S. Rao, Phys. Rev. D32 (1985) 2081.

Chapter summary

concept	operational meaning	used later in
RG step, $\xi' = \xi/b$	integrate out short scales, rescale; fixed points have $\xi = 0, \infty$	Chapter 11 (critical points)
Wilsonian action S_Λ	action reproducing all physics below Λ	Chapters 9 and 10
Polchinski equation	$\partial_t e^{-S^{\mathrm{I}}} = \frac{1}{2}\frac{\delta}{\delta\varphi}\cdot\Delta\cdot\frac{\delta}{\delta\varphi}e^{-S^{\mathrm{I}}}$	Chapter 9 (worldsheet β)
relevant / irrelevant	$y = d - \Delta > 0 \,/\, < 0$	every chapter with a cutoff
Polchinski's theorem	irrelevant couplings slaved, cutoff effects $(\Lambda_R/\Lambda_0)^2$	Chapter 10 (regulator independence)
stability matrix, y_a	linearised flow at g^*; $\nu = 1/y_t$	Chapter 11 (Ising exponents)
six exponents from (y_t, y_h)	scaling of the free energy	Chapter 11
Wilson–Fisher point	$\bar{u}^* = \frac{2}{3}\varepsilon$, $\nu = \frac{1}{2} + \frac{\varepsilon}{12}$	Chapter 11
Callan–Symanzik equation	$[\mu\partial_\mu + \beta\partial_\lambda + n\gamma]G^{(n)} = 0$; at $\beta(\lambda^*) = 0$, $\Delta = \Delta_{\mathrm{cl}} + \gamma(\lambda^*)$	Chapter 11 (power laws)
one-loop φ^4	$\beta = 3\lambda^2/16\pi^2$ in $d = 4$; $-\varepsilon\lambda$ added gives Wilson–Fisher	Chapter 11 (Ising exponents)
one-loop Yang–Mills	$\beta = -\frac{g^3}{16\pi^2}(\frac{11N}{3} - \frac{2n_f}{3})$, $\frac{11}{3} = 4 - \frac{1}{3}$ (spin, orbit)	Chapter 8 ($e^{-8\pi^2/g^2}$), Chapter 13
virial current, improvement	scale $\Rightarrow$ conformal iff $K_\mu = \partial_\mu L$	Chapters 9 and 11
$T^\mu{}_\mu = -\beta^i O_i$	trace measures running	Chapter 9 (Weyl anomaly, Einstein eqs), Chapter 10 (scale anomaly, Adler–Bardeen)
QCD trace anomaly	$\theta^\mu{}_\mu = \frac{\beta}{2g}F_c F_c + \sum_q (1+\gamma_m) m_q \bar{\psi}_q \psi_q$; the proton mass	Chapter 10 (scale versus chiral anomaly)
2D scale $\Rightarrow$ conformal	$G = H = 0$, reflection positivity	Chapter 11
C-function, gradient flow	C decreases, Morse index = number of relevant operators	Chapter 11 (Section 11.13), Chapter 7
QFT = trajectory	UV CFT + relevant operators $\to$ IR CFT or gapped	Chapter 12 (TQFT as IR limit, CS level shift)
dimensional transmutation	$\Lambda_* = \Lambda e^{-1/bg}$	Chapters 8 and 10
sigma-model β	$\beta^G_{\mu\nu} = \alpha' R_{\mu\nu}$ (+ dilaton, H); fixed point $R_{\mu\nu} = 0$; flow is Ricci flow	Chapter 9 (Einstein equations from Weyl invariance)
Chern–Simons level shift	$\eta(A) = \eta(0) + 8h^\vee \mathrm{CS}(A)+$ integer, so $k \to k + h^\vee$ ($k+2$ for SU(2))	Chapter 12 ($Z(S^3)$, $q = e^{2\pi i/(k+2)}$)

Two uses deserve a sentence each. In Chapter 7 the RG reappears as a gradient flow: in two dimensions $\beta^i = -\mathsf{G}^{ij}\partial_j(C/6\pi^2)$, and the Morse index of C at a fixed point is its number of relevant operators (Section 11.13.5). In Chapter 12 a gapped phase is described at long distances by a topological theory; Chern–Simons theory is the infrared end of three-dimensional Yang–Mills with a Chern–Simons term (the Yang–Mills coupling $1/g^2$ has negative mass dimension there, so it is irrelevant), and the famous shift of the level $k \to k+2$ for SU(2) is a one-loop renormalisation of a quantised coupling (Section T5.7.2).

Part II

Following Nash

Chapter 1

A Topological Preliminary

Where this chapter is going

Nash's first chapter is a guided tour of the topologist's toolbox. Nothing in it is quantum field theory yet. But every later chapter leans on it. When a physicist says "the instanton number is an integer", "this anomaly is topological", "the vortex cannot unwind" or "the Dirac monopole needs two patches", a piece of this chapter is doing the work. The chapter answers one question in a sequence of increasingly clever ways: *how can we tell two spaces apart, or two field configurations apart, without solving any equations?*

The answer comes in layers. First (§1) we decide what "the same space" means: homeomorphic, piecewise-linearly equivalent, or diffeomorphic. Then (§2) we relax "the same" to "deformable into each other": homotopy. Homotopy classes of loops and spheres give the *homotopy groups* (§3), which classify vortices, monopoles and instantons. They are hard to compute, so (§4) we trade them for the more computable *cohomology groups*, which de Rham's theorem ties to differential forms and hence to field strengths. We then (§5) study spaces that are locally products, the *fibre bundles*: these are the arenas of gauge theory. Finally (§6) we return to the surprising fact from §1 that one topological manifold may carry several inequivalent calculus structures. That fact reappears in Donaldson theory and in global gravitational anomalies. This part of the notes covers §1–§3. Figure 1.1 shows how the pieces feed into each other and into physics.

Throughout the chapter we keep one physical question in mind. A field configuration is a map from space (or spacetime) into some target space of field values. Two configurations with finite energy can often be deformed into one another continuously, and then they are "in the same sector". If they cannot, some integer (a winding number, a magnetic charge, an instanton number) is conserved for purely topological reasons. The mathematics of this chapter is the systematic theory of such integers.

1.1 From homeomorphism to diffeomorphism

Physicists usually meet "space" already equipped with coordinates, a metric and derivatives. Topology asks what survives if we throw almost all of that away and keep only the notion of *nearness*. When are two spaces the same in this weak sense? When are they the same in the stronger sense needed for calculus? And, surprisingly, can these two notions of sameness disagree? This section answers these questions informally. It also introduces the three kinds of manifold (topological, piecewise-linear, differentiable) that Nash distinguishes on his first two pages.

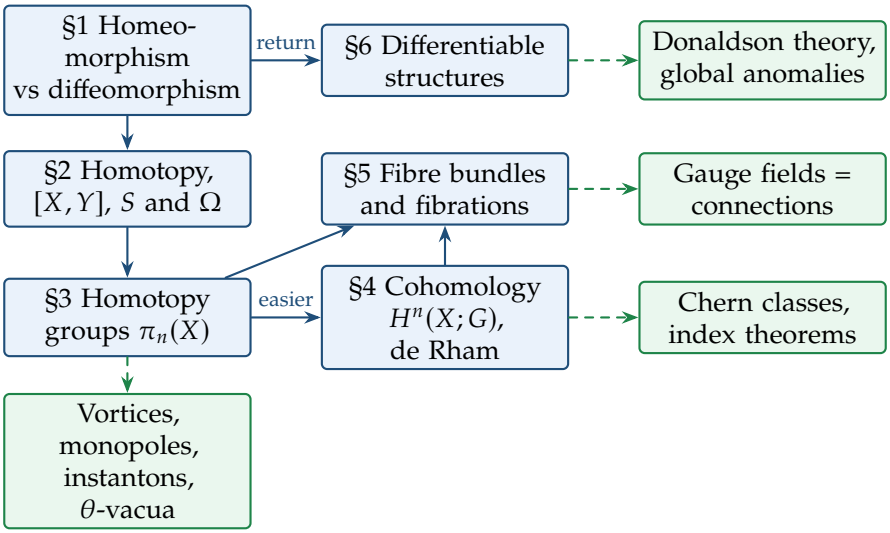


Figure 1.1: The logic of Chapter I. Blue boxes are Nash's sections; green boxes are the physics they feed. Homotopy groups (§3) classify topological defects directly; cohomology (§4) is the computable shadow of homotopy and becomes Chern classes; bundles (§5) are where gauge fields live. This part covers the left column.

1.1.1 Continuity and homeomorphism

Start with the notion of nearness. In $\mathbb{R}^n$ we know what it means for a function to be continuous: points that are close go to points that are close. In ϵ–δ language, f is continuous at x if for every $\epsilon > 0$ there is a $\delta > 0$ with $|y - x| < \delta \Rightarrow |f(y) - f(x)| < \epsilon$. Topologists rephrase this without distances, using **open sets**: sets U such that every point of U has a small ball around it that is still inside U. Then:

Working definition 1.1 (continuous map)

A map $f : X \to Y$ is **continuous** if the preimage $f^{-1}(V)$ [the set of all points of X that f sends into V] of every open set $V \subset Y$ is open in X. *Operational test in practice:* if X and Y sit inside Euclidean spaces and f is given by formulas, check that small changes in the input produce small changes in the output (no jumps, no tearing). Compositions of continuous maps are continuous; polynomials, sin, exp, $\sqrt{\ }$ on its domain, and quotients with nonvanishing denominators are continuous.

A **topological space** is a set together with a declared collection of "open" subsets (closed under arbitrary unions and finite intersections, and containing $\varnothing$ and the whole set). For everything in this book, the spaces are subsets of some $\mathbb{R}^N$, or spaces glued out of such pieces, and "open" means what it means in $\mathbb{R}^N$. So the abstract definition will never bite us.

Working definition 1.2 (homeomorphism)

Two spaces X and Y are **homeomorphic**, written $X \cong Y$, if there are maps

$$\alpha : X \to Y, \qquad \alpha^{-1} : Y \to X, \qquad \alpha \circ \alpha^{-1} = \mathrm{id}_Y, \quad \alpha^{-1} \circ \alpha = \mathrm{id}_X, \tag{1.1}$$

with *both* α and α^{-1} continuous. *Operational test:* exhibit an explicit formula for α, check it is a bijection [each point of Y is hit by exactly one point of X] by writing down the inverse formula, and check that both formulas are continuous.

A homeomorphism is a "rubber-sheet" equivalence: stretch, bend, squash, but do not cut

and do not glue. The coffee cup and the doughnut are the standard cartoon. Here are honest calculations.

Worked example 1.1: An open interval is homeomorphic to the whole line

Let $X = (-1, 1)$ and $Y = \mathbb{R}$. Define $\alpha(x) = \tan(\pi x/2)$. We claim $\alpha^{-1}(y) = \frac{2}{\pi}\arctan y$. Check both compositions:

$$\alpha\big(\alpha^{-1}(y)\big) \overset{(1)}{=} \tan\Big(\frac{\pi}{2}\cdot\frac{2}{\pi}\arctan y\Big) \overset{(2)}{=} y,$$
$$\alpha^{-1}\big(\alpha(x)\big) \overset{(3)}{=} \frac{2}{\pi}\arctan\big(\tan(\pi x/2)\big) \overset{(4)}{=} x.$$

(1) Substitute the candidate inverse into α.
(2) $\tan(\arctan y) = y$ for every real y.
(3) Substitute α into the candidate inverse.
(4) For $x \in (-1, 1)$ the angle $\pi x/2$ lies in $(-\pi/2, \pi/2)$, which is exactly the range of arctan, so $\arctan(\tan\theta) = \theta$ there.

Both tan (away from $\pm\pi/2$) and arctan are continuous, so $(-1, 1) \cong \mathbb{R}$. Length is *not* a topological notion: a finite interval and an infinite line are the same space.

Worked example 1.2: The boundary of a square is homeomorphic to a circle

Let X be the boundary of the square $\max(|x|, |y|) = 1$ and Y the unit circle $x^2 + y^2 = 1$. Radial projection $\alpha(\vec{r}) = \vec{r}/|\vec{r}|$ sends X to Y; its inverse is $\alpha^{-1}(\vec{u}) = \vec{u}/\max(|u_x|, |u_y|)$. Check:

$$\alpha^{-1}\big(\alpha(\vec{r})\big) \overset{(1)}{=} \frac{\vec{r}/|\vec{r}|}{\max(|x|, |y|)/|\vec{r}|} \overset{(2)}{=} \frac{\vec{r}}{\max(|x|, |y|)} \overset{(3)}{=} \vec{r}.$$

(1) The components of $\vec{u} = \vec{r}/|\vec{r}|$ are $x/|\vec{r}|$ and $y/|\vec{r}|$, so $\max(|u_x|, |u_y|) = \max(|x|, |y|)/|\vec{r}|$.
(2) Cancel the positive number $|\vec{r}|$.
(3) On the square, $\max(|x|, |y|) = 1$.

The other composition is checked the same way. Both maps are continuous (neither denominator vanishes on its domain), so the square's corners are invisible to topology (Figure 1.2). A physicist's version: a Wilson loop around a square path and around a circular path detect the same enclosed flux.

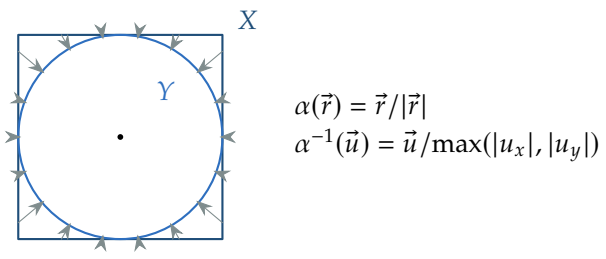


Figure 1.2: Radial projection is a homeomorphism from the square to the circle. Each grey arrow moves a point of the square along its ray from the origin; no two points collide and nothing is torn. Corners are not topological features.

Watch out: A continuous bijection need not be a homeomorphism

Wrap the half-open interval $X = [0, 1)$ onto the circle by $\alpha(t) = (\cos 2\pi t, \sin 2\pi t)$. This is continuous and bijective. But the inverse is *not* continuous: points on the circle just below $(1, 0)$ come from $t \approx 1$, points just above come from $t \approx 0$, so α^{-1} jumps at $(1, 0)$. That is why (1.1) demands continuity of α^{-1} separately. And indeed $[0, 1) \not\cong S^1$: removing the point $t = \frac{1}{2}$ disconnects $[0, 1)$, while removing any single point leaves the circle connected.

That last argument is typical: to prove two spaces are *not* homeomorphic, we find a property preserved by homeomorphisms that one space has and the other lacks. Such a property is a **topological invariant**. The simplest ones are:

- *Connectedness* and the number of connected pieces.
- *Compactness*: for subsets of $\mathbb{R}^N$, being closed and bounded. S^1 is compact, $\mathbb{R}$ is not, so $S^1 \not\cong \mathbb{R}$.
- *What happens when you remove a point.* $\mathbb{R}$ minus a point has two pieces; $\mathbb{R}^2$ minus a point has one. So $\mathbb{R} \not\cong \mathbb{R}^2$.
- *Dimension.* $\mathbb{R}^n \cong \mathbb{R}^m$ only if $n = m$ (Brouwer's "invariance of domain"; easy for $n = 1$ by the previous item, hard in general).
- *Numbers built by counting*, such as the Euler characteristic below, and *algebraic objects* such as the homotopy groups of §1.3.

Reading Nash (p. 1, (1.1))

Nash's (1.1) is exactly our (1.1); his I is the identity map. He announces the goal of topology as "classifying all topological spaces up to homeomorphism" and immediately says that this goal is not reached except in special cases. The rest of the chapter is about retreating to more modest but achievable goals.

1.1.2 The one great success: closed surfaces

The classification problem *is* solved in dimension two, for surfaces that are compact, have no boundary ("closed") and are orientable (they have two sides; no Möbius strips inside). The answer:

Key idea: Classification of closed orientable surfaces

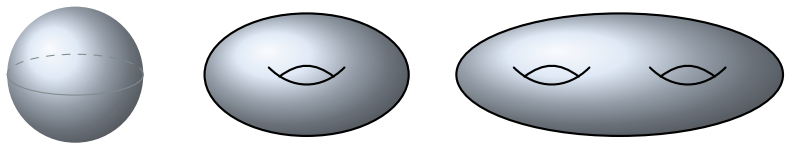
Every closed orientable surface is homeomorphic to a sphere with g handles attached, for exactly one $g \in \{0, 1, 2, \dots\}$, called the **genus**. Two such surfaces are homeomorphic if and only if they have the same genus. The genus is read off from the Euler characteristic by

$$\chi = 2 - 2g. \tag{1.2}$$

Working definition 1.3 (Euler characteristic)

Cut the surface into polygons (faces) meeting along edges and vertices, in any way you like, as long as the pieces are glued edge to edge. Count V vertices, E edges and F faces. Then

$$\chi := V - E + F.$$



$g = 0, \chi = 2$ $g = 1, \chi = 0$ $g = 2, \chi = -2$

Figure 1.3: The first three closed orientable surfaces. Each handle lowers the Euler characteristic by 2. As Riemann surfaces these are the Riemann sphere $\mathbb{CP}^1$, the complex tori $\mathbb{C}/\Lambda$, and the genus-2 curves; string perturbation theory sums over exactly this list (Chapter 9).

The number does not depend on how you cut, so it is a topological invariant [homeomorphic surfaces always give the same χ].

Why is χ independent of the cutting? Any two cuttings have a common refinement, and refining changes nothing: adding a vertex in the middle of an edge adds one vertex and one edge ($\Delta\chi = 1 - 1 = 0$); adding an edge across a face that joins two existing vertices adds one edge and one face ($\Delta\chi = -1 + 1 = 0$). Dray gives the physicist's reason [Dray §18.10]: the Gauss–Bonnet theorem

$$\int_{\Sigma} K \, \mathrm{d}A = 2\pi\chi(\Sigma) \tag{1.3}$$

expresses χ as an integral of the Gaussian curvature K. Deforming the surface changes K locally, but the total is pinned to 2π times an integer, so it cannot change continuously. (We return to Gauss–Bonnet in Chapter T2.) This "an integral that must be an integer is deformation-invariant" argument will recur all through the book.

Worked example 1.3: Euler characteristics of the sphere and the torus

Sphere. Inflate a cube into a sphere. It has $V = 8$ corners, $E = 12$ edges and $F = 6$ faces:

$$\chi(S^2) = 8 - 12 + 6 = 2, \qquad \text{so } g = \tfrac{1}{2}(2 - \chi) = 0.$$

Torus. A torus is a square with opposite edges glued (Figure 1.4). Cut the square into a 3×3 grid of small squares, and then each small square into two triangles by a diagonal. After gluing, the grid points form a 3×3 array (the points on the right and top edges are identified with those on the left and bottom), so

$$V \overset{(1)}{=} 3 \cdot 3 = 9,$$
$$F \overset{(2)}{=} 2 \cdot 9 = 18,$$
$$E \overset{(3)}{=} \tfrac{3}{2}F = 27,$$
$$\chi(T^2) = 9 - 27 + 18 = 0, \qquad g = 1.$$

(1) Gluing identifies the 4×4 lattice of grid points to 3×3.
(2) Nine small squares, two triangles each.
(3) Each triangle has three edges and each edge is shared by exactly two triangles after gluing.

The crudest possible cutting (one square, with its four edges glued in pairs and all four corners glued to one point) gives $V = 1$, $E = 2$, $F = 1$ and the same $\chi = 0$, as it must.

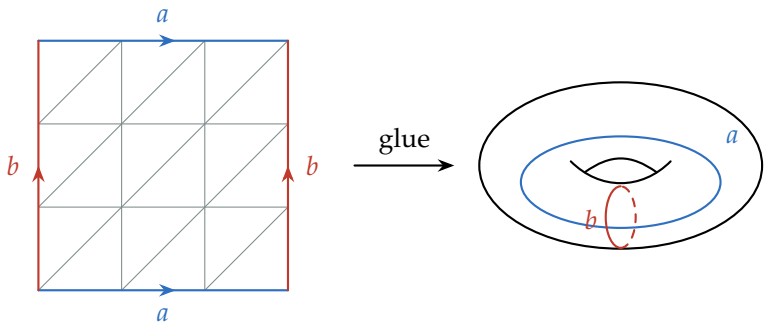


Figure 1.4: The torus as a square with opposite edges glued: a to a and b to b, respecting the arrows. The 3×3 triangulated grid on the left counts $V = 9$, $E = 27$, $F = 18$ after gluing, so $\chi = 0$. The two edge classes become the two basic loops a and b on the torus; they return in §1.3 as generators of $\pi_1(T^2) = \mathbb{Z} \times \mathbb{Z}$.

Try it: A genus-g surface

A closed genus-g surface can be made from a $4g$-gon by gluing its edges in the pattern $a_1 b_1 a_1^{-1} b_1^{-1} \cdots a_g b_g a_g^{-1} b_g^{-1}$ (read around the boundary; a^{-1} means edge a traversed backwards). All $4g$ corners end up glued to a single point. Compute χ. *Answer:* $V = 1$, $E = 2g$ (edges are glued in pairs), $F = 1$, so $\chi = 2 - 2g$, in agreement with (1.2).

Nash adds that these surfaces are "the celebrated Riemann surfaces". The reason: every closed orientable surface can be given a **complex structure**, i.e. an atlas of charts to $\mathbb{C}$ whose transition functions are holomorphic. Then it becomes a one-dimensional complex manifold, a **Riemann surface**. The genus is then the number of holes, and also the number of independent holomorphic 1-forms (we will meet that in Chapter 5). What topology does *not* fix is which complex structure: a torus can be $\mathbb{C}/(\mathbb{Z} + \tau\mathbb{Z})$ for any τ in the upper half plane. The space of such choices is a moduli space, central in string theory (Chapter 9).

Watch out: Orientability matters

If we drop orientability, there are more closed surfaces: the projective plane $\mathbb{RP}^2$ (a disc with antipodal boundary points glued, $\chi = 1$), the Klein bottle ($\chi = 0$), and connected sums of these. The Klein bottle and the torus both have $\chi = 0$ but are not homeomorphic; the full classification needs χ *and* orientability. This is why Nash restricts to "compact closed orientable surfaces".

1.1.3 Why dimension two is special: room to move

In higher dimensions, no list like "sphere with g handles" exists. Nash gives a heuristic for why low and high dimensions behave differently. The heuristic rests on one counting rule, which is worth making explicit.

Key idea: General position: a dimension count

In an n-dimensional space, a p-dimensional and a q-dimensional object in "general position" (wiggled slightly at random) meet in a set of dimension

$$p + q - n. \tag{1.4}$$

If $p + q - n < 0$, a small wiggle makes them miss each other entirely.

The reason is linear algebra. Locally, a p-dimensional object is cut out by $n-p$ equations and a q-dimensional one by $n-q$ equations. Their intersection solves all $(n-p)+(n-q)$ equations in n unknowns. Generically the solution set has dimension $n-(n-p)-(n-q) = p+q-n$, and if that number is negative, generic equations have no common solution. Two random lines in a plane ($1+1-2=0$) meet in a point; two random lines in space ($1+1-3=-1$) miss.

Now Nash's point. Suppose we want to deform something inside a 2-dimensional manifold M, and a submanifold N is in the way. N has dimension 1 or 2. A path of the deformation is 1-dimensional, and $1+1-2=0$: a curve generically *hits* a curve in a surface, and there is no room to go around. In a manifold of large dimension n, a 1- or 2-dimensional obstacle is hit by a 2-dimensional family of paths only if $2+2-n \geq 0$, i.e. only if $n \leq 4$. For $n \geq 5$ there is room to push the deformation off the obstacle (Figure 1.5).

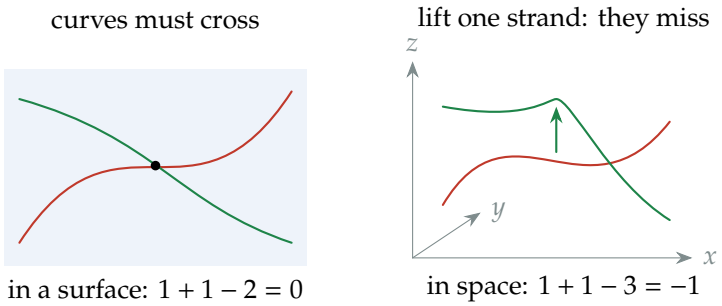


Figure 1.5: General position. In a surface, two curves that cross cannot be pulled apart by a small wiggle: the crossing point is stable because $1+1-2=0$. With one extra dimension, one strand lifts over the other, and the count $1+1-3<0$ says they generically miss. High dimensions are "roomy"; low dimensions are "cramped".

1.1.4 Knots, and the dimension at which things become easy

Nash's second illustration is knot theory.

Working definition 1.4 (knot)

A **knot** is a smooth [K can be differentiated any number of times] one-to-one map $K : S^1 \to \mathbb{R}^3$ (a closed loop of string with no self-intersections). Two knots are **equivalent** if one can be deformed into the other through knots, i.e. without ever passing the string through itself. A knot is **trivial** (the **unknot**) if it is equivalent to a round circle.

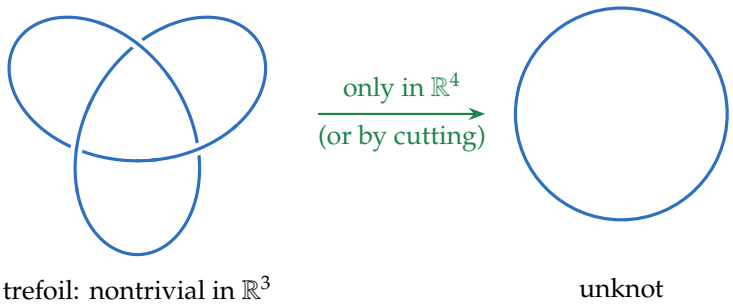


Figure 1.6: The trefoil cannot be deformed into the round circle inside $\mathbb{R}^3$ without the string passing through itself. In $\mathbb{R}^4$ it can: at each crossing one strand is pushed into the fourth direction, slid past, and brought back.

Many knots are non-trivial: the trefoil (Figure 1.6) is the simplest. Proving it is non-trivial

needs an invariant (for example the number of ways of colouring its three arcs with three colours so that at each crossing either all colours agree or all differ: the trefoil has 9 such colourings, the unknot only 3). But, as Nash says, *every* knot unties in $\mathbb{R}^4$.

Worked example 1.4: Why every knot unties in four dimensions

Think of a deformation of a knot as a movie: at each moment $s \in [0,1]$ of deformation time we have a closed curve in $\mathbb{R}^n$. A crossing change happens when two points of the string collide at some moment. Count the dimension of the set of collisions with (1.4): at fixed s, take two separate arcs of the string, then let s vary:

$$\dim\{\text{collisions}\} \overset{(1)}{=} \underbrace{(1+1-n)}_{\text{two arcs at fixed time}} + \underbrace{1}_{\text{time}}$$

$$= 3-n \begin{cases} = 0 & n=3: \text{ isolated collisions, cannot be avoided,} \\ = -1 & n=4: \text{ generically no collisions at all.} \end{cases}$$

(1) At a fixed time, two arcs in $\mathbb{R}^n$ meet in dimension $1+1-n$ by (1.4); letting time vary adds one dimension to the family.

So in $\mathbb{R}^4$ any homotopy of a loop (any deformation allowed to pass the string through itself) can be wiggled into one that never self-intersects. Every loop in $\mathbb{R}^4$ can be shrunk to a round circle by a homotopy, hence by an isotopy: every knot in $\mathbb{R}^4$ is trivial.

The general lesson: interesting knotting happens in codimension two (S^1 in $\mathbb{R}^3$, S^2 in $\mathbb{R}^4$, S^n in $\mathbb{R}^{n+2}$). Higher codimension gives too much room.

Physical picture: Knots in physics

Knots are not a mathematical curiosity for physicists. A Wilson loop $\operatorname{tr} P \exp \oint_K A$ is a function of a closed curve K; in Chern–Simons theory its expectation value depends only on the knot type of K, and it computes the Jones polynomial. This is one of the founding examples of a topological quantum field theory (Chapter 12). Knotted vortex lines in fluids and superfluids, and knotted flux tubes, are also real objects.

Nash then asks: if $\dim M = n$, from what n onwards does M behave like a "space of higher dimension", with enough room for general-position arguments? The answer is about $n = 5$. The technical reason is the **Whitney trick**: to cancel two intersection points of two submanifolds of complementary dimensions, one pushes across an embedded 2-disc. The 2-disc can be made embedded by general position when $2+2-n < 0$, i.e. $n \geq 5$. Dimensions 3 and 4 are the hard ones: in dimension 3 the geometry of 3-manifolds is special (Thurston, Perelman), and dimension 4 is where the gauge theory of this book (instantons, Seiberg–Witten) discovered completely unexpected phenomena. Nash returns to this in §6 of this chapter and in Chapter 8.

1.1.5 Topological manifolds

Since classifying all topological spaces is hopeless, Nash "lowers his sights" to spaces that look locally like $\mathbb{R}^n$.

Working definition 1.5 (topological manifold of dimension n)

A space M is a **topological manifold** of dimension n if

(i) it is covered by open sets (**patches** or **charts**) U_α, each equipped with a homeomorphism [a continuous bijection whose inverse is also continuous] $\phi_\alpha : U_\alpha \to \phi_\alpha(U_\alpha) \subset \mathbb{R}^n$ onto an open subset of $\mathbb{R}^n$ (the **coordinates** on U_α);

(ii) it is **Hausdorff**: any two distinct points have disjoint neighbourhoods [open sets containing them, with no point in common].

On an overlap $U_\alpha \cap U_\beta$ the two coordinate systems are related by the **transition function**

$$g_{\alpha\beta} := \phi_\alpha \circ \phi_\beta^{-1} : \phi_\beta(U_\alpha \cap U_\beta) \to \phi_\alpha(U_\alpha \cap U_\beta), \tag{1.5}$$

a map between open subsets of $\mathbb{R}^n$. For a topological manifold, the $g_{\alpha\beta}$ are homeomorphisms (automatically, being composites of homeomorphisms). A collection of charts covering M is an **atlas**.

The transition functions are the whole story. They say how to glue flat pieces of $\mathbb{R}^n$ together. The gluing idea returns in §5 of this chapter, where the transition functions take values in a gauge group and describe how to glue a fibre bundle: a gauge field on a non-trivial bundle is exactly a collection of local gauge potentials related by gauge transformations $g_{\alpha\beta}$ on overlaps.

Watch out: Why "Hausdorff"?

Without condition (ii) strange things can happen. Take two copies of the real line and glue them along every point except 0. The result, the "line with two origins", is covered by two charts (each copy of $\mathbb{R}$), so it is locally Euclidean. But the two origins cannot be separated: every neighbourhood of one meets every neighbourhood of the other. A sequence like $x_k = 1/k$ converges to both origins. Physics needs limits to be unique, so we exclude this. (Usually one also asks that M be covered by countably many charts; this is automatic for everything we meet.)

Worked example 1.5: An atlas for the circle

Let $S^1 = \{(x, y) : x^2 + y^2 = 1\}$. Use **stereographic projection**: from the north pole $N = (0, 1)$ draw the line through a point $P = (x, y) \neq N$ and record where it hits the x-axis. Similarly from the south pole $S = (0, -1)$. This gives two charts,

$$U_N = S^1 \setminus \{N\}, \quad u = \phi_N(x, y) = \frac{x}{1 - y}; \qquad U_S = S^1 \setminus \{S\}, \quad v = \phi_S(x, y) = \frac{x}{1 + y}.$$

On the overlap (the circle minus both poles), $u \neq 0$ and

$$u\, v \overset{(1)}{=} \frac{x^2}{1 - y^2} \overset{(2)}{=} 1, \qquad \text{so} \qquad g_{SN}(u) = \phi_S \circ \phi_N^{-1}(u) = \frac{1}{u}.$$

(1) Multiply the two formulas: $(1 - y)(1 + y) = 1 - y^2$.
(2) On the circle, $x^2 = 1 - y^2$.

The transition function $u \mapsto 1/u$ is a homeomorphism of $\mathbb{R} \setminus \{0\}$ to itself, so this is an atlas of a topological manifold of dimension 1. It is even infinitely differentiable, which will matter shortly. The same construction for $S^2 \subset \mathbb{R}^3$ gives charts $\vec{u}_N = (x, y)/(1 - z)$ and $\vec{u}_S = (x, y)/(1 + z)$ with $\vec{u}_S = \vec{u}_N/|\vec{u}_N|^2$; in complex form $\zeta_N = (x + iy)/(1 - z)$, $\zeta_S = (x - iy)/(1 + z)$ and $\zeta_S = 1/\zeta_N$, which is holomorphic: S^2 is the Riemann sphere $\mathbb{CP}^1$.

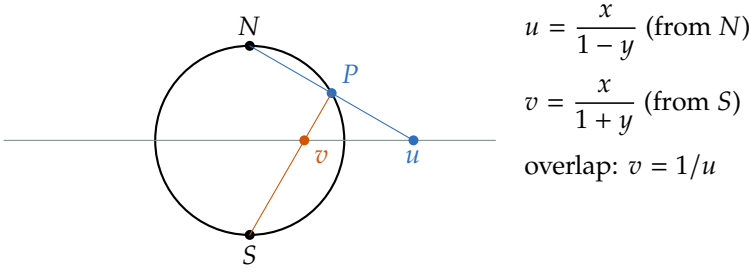


Figure 1.7: Two stereographic charts cover the circle. A point P gets the coordinate u (blue line from N) and v (orange line from S). Neither chart alone covers the whole circle, since each misses its own pole. On the overlap the coordinates are related by the smooth transition function $v = 1/u$.

1.1.6 Piecewise-linear manifolds

Topological manifolds are very general and "sparsely provided with concrete mathematical structures", as Nash puts it: we cannot differentiate on them, and we cannot count anything on them directly. The first way to add structure is combinatorial.

A **triangulation** of a space cuts it into simplices (points, edges, triangles, tetrahedra and their higher analogues) glued face to face. The sphere is triangulated by the surface of a tetrahedron or an octahedron. A map is **piecewise-linear** (PL) if, after subdividing into small enough simplices, it is affine (linear plus constant) on each simplex.

Working definition 1.6 (PL manifold)

A **piecewise-linear manifold** is a topological manifold with an atlas whose transition functions $g_{\alpha\beta}$ [the maps $\phi_\alpha \circ \phi_\beta^{-1}$ converting one chart's coordinates into another's] and their inverses are piecewise-linear. Equivalently, it is a manifold built from a triangulation in which the simplices around every vertex fit together like the simplices of a triangulated ball in $\mathbb{R}^n$. *Operational meaning:* you can *count*. Invariants such as $\chi = \sum_k (-1)^k (\#k\text{-simplices})$ and the simplicial homology groups [groups counting independent holes of each dimension] are computed by finite linear algebra on the triangulation.

Worked example 1.6: Euler characteristic of S^2 from two triangulations

Tetrahedron: $V = 4$, $E = 6$, $F = 4$, so $\chi = 4 - 6 + 4 = 2$. Octahedron: $V = 6$, $E = 12$, $F = 8$, so $\chi = 6 - 12 + 8 = 2$. The octahedron triangulation is a subdivision-equivalent refinement of the same PL structure: a PL homeomorphism (radial projection followed by straightening edges) carries one to the other. Lattice field theorists meet PL manifolds constantly: a lattice discretisation of spacetime *is* a triangulation (or cubulation), and lattice gauge theory lives on its edges (link variables) and faces (plaquettes).

1.1.7 Differentiable manifolds

The physicist, and the differential topologist, want calculus. Calculus needs the notion of a differentiable function, and on a manifold "differentiable" only makes sense if it means the same thing in every chart. That forces the transition functions to be differentiable.

Working definition 1.7 (differentiable manifold, diffeomorphism)

A **differentiable** (or **smooth**) manifold is a topological manifold with an atlas whose transition functions $g_{\alpha\beta}$ (and hence their inverses $g_{\beta\alpha} = g_{\alpha\beta}^{-1}$) are infinitely differentiable (C^∞). A function $f : M \to \mathbb{R}$ is smooth if $f \circ \phi_\alpha^{-1}$ is smooth on $\phi_\alpha(U_\alpha) \subset \mathbb{R}^n$ for every α. A map $h : M \to N$ between smooth manifolds is a **diffeomorphism** if it is a homeomorphism and both h and h^{-1} are smooth when written in charts. *Operational test:* write h in local coordinates, check that all partial derivatives of all orders exist and are continuous, and check the Jacobian $\det(\partial h^i/\partial x^j)$ never vanishes (then the inverse is smooth too, by the inverse function theorem [a smooth map with nonzero Jacobian at a point has a smooth inverse near that point]).

Why does "smooth in one chart" imply "smooth in every chart"? Because on an overlap

$$f \circ \phi_\beta^{-1} \overset{(1)}{=} f \circ \phi_\alpha^{-1} \circ \phi_\alpha \circ \phi_\beta^{-1} \overset{(2)}{=} \left(f \circ \phi_\alpha^{-1}\right) \circ g_{\alpha\beta},$$

(1) Insert $\phi_\alpha^{-1} \circ \phi_\alpha = \mathrm{id}$ on $U_\alpha \cap U_\beta$.
(2) Definition (1.5) of the transition function.

and a composite of smooth maps is smooth (chain rule). This is exactly why the $g_{\alpha\beta}$ must be C^∞. The circle atlas of Example 1.5 has $g_{SN}(u) = 1/u$, smooth on $u \neq 0$, so S^1 is a smooth manifold; likewise S^2, and every sphere S^n. Tori $T^n = S^1 \times \cdots \times S^1$, the projective spaces $\mathbb{RP}^n$ and $\mathbb{CP}^n$, and all the matrix Lie groups (U(1), SU(2), SO(3), ...) are smooth manifolds. The toolkit chapter Chapter T1 builds forms and derivatives on such manifolds.

Worked example 1.7: Homeomorphic but not a diffeomorphism; different atlases but the same manifold

Consider $h : \mathbb{R} \to \mathbb{R}$, $h(x) = x^3$. It is a homeomorphism: its inverse $h^{-1}(y) = y^{1/3}$ (real cube root) is continuous. It is smooth. But h^{-1} is not differentiable at 0:

$$\frac{\mathrm{d}}{\mathrm{d}y}\, y^{1/3} = \frac{1}{3}\, y^{-2/3} \xrightarrow{y \to 0} \infty.$$

So h is a homeomorphism but not a diffeomorphism. (In the Jacobian language: $h'(0) = 0$.) Now use h to define a new atlas on the set $\mathbb{R}$: the single chart $\psi(x) = x^3$. Call the resulting smooth manifold $\mathbb{R}'$. The standard atlas $\{\mathrm{id}\}$ and the new atlas $\{\psi\}$ are *not compatible*: the transition function $\mathrm{id} \circ \psi^{-1}(y) = y^{1/3}$ is not smooth. So they are different smooth structures on the same set. Are $\mathbb{R}$ and $\mathbb{R}'$ different smooth manifolds? No. The map $f : \mathbb{R} \to \mathbb{R}'$, $f(x) = x^{1/3}$, is a diffeomorphism. To see this, write it in the charts (id on the source, ψ on the target):

$$\psi \circ f \circ \mathrm{id}^{-1}(x) \overset{(1)}{=} \left(x^{1/3}\right)^3 \overset{(2)}{=} x,$$

(1) Apply f, then the chart $\psi(t) = t^3$ of $\mathbb{R}'$.
(2) Real cube roots cube back to the original number.

which is the identity, smooth with smooth inverse. So different atlases can give diffeomorphic manifolds. When mathematicians count "smooth structures" on a topological manifold, they count them *up to diffeomorphism*, and this example counts as one structure, not two.

1.1.8 Three kinds of manifold, and why the distinction matters

We now have three levels of structure: M_{TOP} (transition functions homeomorphisms), M_{PL} (piecewise-linear) and M_{DIFF} (smooth). Every smooth manifold can be triangulated in an essentially unique way compatible with its smooth structure (Whitehead), and every PL manifold is a topological manifold. So there are forgetful arrows

$$\mathrm{DIFF} \longrightarrow \mathrm{PL} \longrightarrow \mathrm{TOP}.$$

One might guess that these arrows can always be reversed: that any topological manifold has, uniquely, a PL structure and a smooth structure. Nash stresses that this is false in both respects.

Key idea: Existence and uniqueness both fail

- *Existence can fail.* There are topological manifolds that admit no PL structure, and topological or PL manifolds that admit no smooth structure.
- *Uniqueness can fail.* A topological manifold can carry several smooth structures that are homeomorphic but *not* diffeomorphic to each other.

This is the *raison d'être* of **differential topology** as a separate subject: smoothness is genuinely extra information.

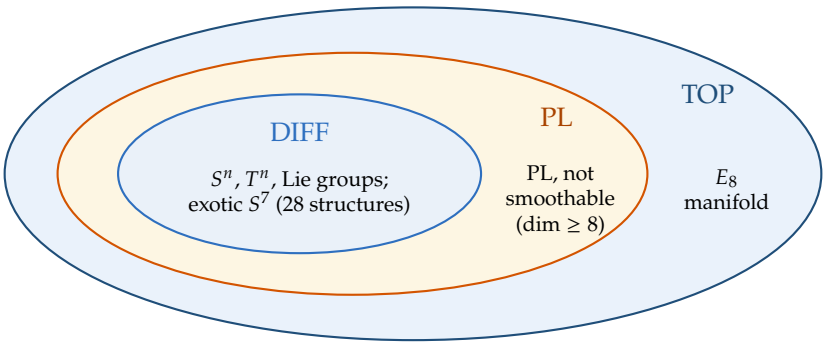


Figure 1.8: Three nested worlds. Every smooth manifold is PL, and every PL manifold is topological, but not conversely. Freedman's E_8 manifold is a closed topological 4-manifold with no PL and no smooth structure. Inside DIFF, one topological sphere S^7 carries 28 inequivalent smooth structures (counting orientation).

Some landmark facts, to make these statements concrete (we only name them here; Nash develops some of them in §6 of this chapter):

dimension	what happens	name
$n \le 3$	TOP = PL = DIFF: every topological manifold has a unique smooth structure	Radó, Moise
$n = 4$	topological 4-manifolds with no smooth structure	Freedman, Donaldson
	$\mathbb{R}^4$ has uncountably many smooth structures	Taubes, Gompf
$n \ge 5$	TOP manifolds with no PL structure	Kirby–Siebenmann
$n = 7$	S^7 has 28 smooth structures (oriented)	Milnor, Kervaire–Milnor
$n = 11$	S^{11} has 992 smooth structures, i.e. 991 exotic ones	Kervaire–Milnor

Physical picture: Why a physicist should care

Nash anticipates the objection that "the particular properties of manifolds described here might be thought unlikely to have any relevance for quantum field theory", and says the opposite is true. Two examples we will meet:

- *Donaldson theory* (Chapter 8): moduli spaces of instantons, solutions of the Yang–Mills self-duality equations, are sensitive to the smooth structure of a 4-manifold. Gauge theory is how exotic 4-dimensional smooth structures were discovered.
- *Global gravitational anomalies* (Chapter 10): the 991 exotic 11-spheres enter the check that ten-dimensional superstring theories are free of global anomalies, because a large diffeomorphism of S^{10} can be used to glue an exotic S^{11}.

More simply: a path integral over metrics or fields uses derivatives, so it needs a smooth structure. If a spacetime topology admits several, the "sum over geometries" has a choice to make.

Reading Nash (pp. 1–3)

Nash's order is: goal of topology (1.1); surfaces solved by genus; dimension 2 restricted by obstacles; knots untie in $\mathbb{R}^4$; "higher dimension" starts around $n = 5$ (pointer to his §6); topological manifolds with transition functions $g_{\alpha\beta}$; PL manifolds ($g_{\alpha\beta}$, $g_{\alpha\beta}^{-1}$ piecewise-linear); differentiable manifolds ($g_{\alpha\beta}$, $g_{\alpha\beta}^{-1}$ of class C^∞); M_{TOP}, M_{PL}, M_{DIFF} may fail to exist or to be unique. He closes by adopting the convention that "manifold" means differentiable manifold unless stated otherwise. We adopt the same convention from now on.

Recap

- A homeomorphism is a continuous bijection with continuous inverse (1.1); homeomorphic spaces are topologically identical. Length, corners and curvature are not topological.
- Closed orientable surfaces are classified by the genus g, read off from $\chi = V - E + F = 2 - 2g$.
- General position (1.4): objects of dimensions p, q in $\mathbb{R}^n$ generically miss if $p + q < n$. Low dimensions are cramped, dimension ≥ 5 is roomy; knots untie in $\mathbb{R}^4$.
- A manifold is glued from patches of $\mathbb{R}^n$ by transition functions $g_{\alpha\beta} = \phi_\alpha \circ \phi_\beta^{-1}$. Their regularity (homeomorphism, PL, C^∞) defines TOP, PL, DIFF manifolds.
- Smooth structures may fail to exist or to be unique: exotic spheres, exotic $\mathbb{R}^4$. This matters for instantons and anomalies.

1.2 Some algebraic topology: homotopy

Homeomorphism is too fine a relation to be useful in general: we cannot classify spaces up to homeomorphism. Algebraic topology therefore changes the question. Instead of deciding whether two spaces are *the same*, it attaches to each space some **invariants**, quantities that do not change under homeomorphism, and compares those. If the invariants differ, the spaces differ. Nash lists the two kinds:

- *numbers*: the dimension, the Euler–Poincaré characteristic χ we met in §1.1.2;
- *algebraic objects*: groups such as the homotopy groups (this section and the next) and the homology and cohomology groups (§4 of this chapter).

The most powerful single idea for building invariants is *homotopy*: continuous deformation of maps. A physicist already thinks this way. A field configuration is a map $\phi : X \to Y$ from space X to a space Y of field values. Time evolution with finite energy deforms ϕ continuously. So anything that cannot change under continuous deformation is a *conserved topological charge*. This section builds the language for such charges.

1.2.1 Maps and homotopies

Fix two spaces X and Y. Write

$$\mathrm{Map}(X, Y) := \{\alpha : X \to Y \text{ continuous}\} \tag{1.6}$$

for the set of all continuous maps from X to Y. For a physicist, $\mathrm{Map}(\mathbb{R}^3, \mathbb{C})$ is the space of (continuous) complex scalar field configurations, and $\mathrm{Map}(S^1, Y)$ is the space of closed strings moving in a target space Y.

Working definition 1.8 (homotopy)

Two maps $\alpha, \beta \in \mathrm{Map}(X, Y)$ are **homotopic**, written $\alpha \simeq \beta$, if there is a continuous map [small changes in the pair (t, x) give small changes in $F(t, x)$]

$$F : [0,1] \times X \to Y \qquad \text{with} \qquad F(0, x) = \alpha(x), \quad F(1, x) = \beta(x) \quad \text{for all } x \in X. \tag{1.7}$$

F is called a **homotopy** from α to β. Think of the first variable $t \in [0,1]$ as time: $F_t := F(t, \cdot)$ is a one-parameter family of maps (a "movie") that starts at α and ends at β. *Operational test:* to show $\alpha \simeq \beta$, write down a formula $F(t, x)$, check it is continuous in both variables jointly, check it lands in Y for all t, and check the endpoints.

The condition "lands in Y for all t" is where all the content is. Writing down *some* interpolation between α and β is easy; keeping it inside Y may be impossible.

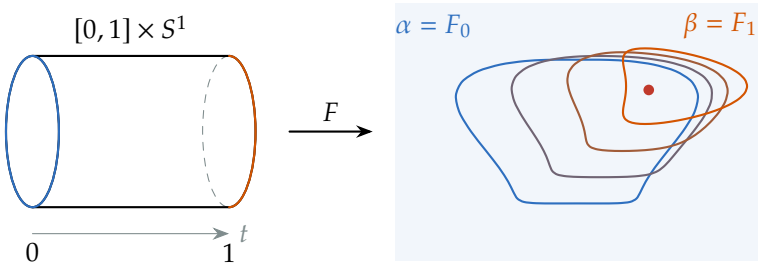


Figure 1.9: A homotopy is a map from the cylinder $[0,1] \times X$ (here $X = S^1$) into Y. Its left end is the loop α and its right end the loop β; the intermediate slices are a movie of loops (colours interpolating). In the plane with a hole (the red dot), the movie is legal only if no intermediate loop passes through the hole. Here every slice keeps the hole inside, so this is a genuine homotopy in $\mathbb{R}^2 \setminus \{\text{hole}\}$. Sliding the loop off the hole would require crossing it, and we will see in §1.3.2 that no homotopy can do that.

Worked example 1.8: Everything is homotopic in $\mathbb{R}^n$: the straight-line homotopy

Let $Y = \mathbb{R}^n$ and let $\alpha, \beta \in \mathrm{Map}(X, \mathbb{R}^n)$ be *any* two maps. Define $F(t,x) = (1-t)\,\alpha(x) + t\,\beta(x)$. Then

$$F(0,x) \overset{(1)}{=} \alpha(x), \qquad F(1,x) \overset{(1)}{=} \beta(x),$$

(1) Set $t = 0$ or $t = 1$ in the formula.

and F is continuous because it is built from continuous functions by multiplication and addition. It lands in $\mathbb{R}^n$ trivially. So any two maps into $\mathbb{R}^n$ are homotopic. The same formula works for maps into any **convex** set (one containing the straight segment between any two of its points), such as a ball. It fails in $\mathbb{R}^2 \setminus \{0\}$, where the straight segment from $\alpha(x)$ to $\beta(x)$ may pass through the origin.

Worked example 1.9: The antipodal map of an odd-dimensional sphere

Let $X = Y = S^{2k-1}$, viewed as the unit sphere in $\mathbb{C}^k$: $S^{2k-1} = \{z = (z_1, \dots, z_k) : |z_1|^2 + \cdots + |z_k|^2 = 1\}$. The **antipodal map** is $a(z) = -z$. Claim: $a \simeq \mathrm{id}$. Take $F(t,z) = e^{i\pi t} z$. Then

$$|F(t,z)|^2 \overset{(1)}{=} \sum_{j=1}^{k} |e^{i\pi t} z_j|^2 \overset{(2)}{=} \sum_{j=1}^{k} |z_j|^2 \overset{(3)}{=} 1,$$

$$F(0,z) = z = \mathrm{id}(z), \qquad F(1,z) \overset{(4)}{=} -z = a(z).$$

(1) Definition of the length in $\mathbb{C}^k$.
(2) $|e^{i\pi t}| = 1$, so each term is unchanged.
(3) z was on the unit sphere.
(4) $e^{i\pi} = -1$.

So F stays on the sphere for all t: it is the rotation by angle πt simultaneously in each complex plane. For $k = 1$ this is just the rotation of a circle by 180°. On even-dimensional spheres, such as S^2, the antipodal map is *not* homotopic to the identity (it has degree -1, §1.3.8); this is the hairy-ball phenomenon.

Worked example 1.10: Homotopy is an equivalence relation

We check the three properties, since everything below depends on them.

- *Reflexive:* $\alpha \simeq \alpha$ via the "do nothing" homotopy $F(t,x) = \alpha(x)$.
- *Symmetric:* if F goes from α to β, then $\bar{F}(t,x) := F(1-t,x)$ goes from β to α (run the movie backwards).
- *Transitive:* if F goes from α to β and G from β to γ, play the two movies in succession at double speed:

$$H(t,x) = \begin{cases} F(2t,x) & 0 \le t \le \tfrac{1}{2}, \\ G(2t-1,x) & \tfrac{1}{2} \le t \le 1. \end{cases}$$

The only thing to check is that H is well defined and continuous at $t = \frac{1}{2}$, where both formulas apply:

$$F(2\cdot\tfrac{1}{2}, x) = F(1,x) \overset{(1)}{=} \beta(x) \overset{(2)}{=} G(0,x) = G(2\cdot\tfrac{1}{2} - 1, x).$$

(1) F ends at β.
(2) G starts at β.

Two continuous functions on the closed pieces $[0, \frac{1}{2}] \times X$ and $[\frac{1}{2}, 1] \times X$ that agree on the common edge glue to a continuous function (the "gluing lemma"). This double-speed concatenation reappears as the product of loops in §1.3.

Because homotopy is an equivalence relation, it divides $\mathrm{Map}(X, Y)$ into equivalence classes.

Working definition 1.9 (homotopy classes)

The set of homotopy classes [one class for each collection of maps that can all be deformed into one another] of maps from X to Y is written

$$[X, Y] := \mathrm{Map}(X, Y)/\simeq . \tag{1.8}$$

The class of a map α is written $[\alpha]$. *Physics reading:* $[X, Y]$ is the set of *topological sectors* of field configurations $X \to Y$: two configurations are in the same sector iff one can be continuously deformed into the other.

Some first examples, all of which we justify in this part:

$[X, Y]$	answer	reason
$[X, \mathbb{R}^n]$	one element	straight-line homotopy, Example 1.8
$[\{\mathrm{pt}\}, Y]$	path components of Y, written $\pi_0(Y)$	a map from a point is a point of Y; a homotopy is a path between points
$[S^1, S^1]$	$\mathbb{Z}$	winding number, §1.3.2
$[S^1, S^2]$	one element	loops slide off a point of S^2, §1.3.7
$[S^2, S^2]$	$\mathbb{Z}$	degree, §1.3.8
$[S^3, \mathrm{SU}(2)]$	$\mathbb{Z}$	instanton number, §1.3.14

Dictionary: homotopy language

mathematics	physics
map $\alpha : X \to Y$	field configuration on space X with values in Y
homotopy $F : [0, 1] \times X \to Y$	continuous evolution / deformation of the field
class $[\alpha] \in [X, Y]$	topological sector, "charge" of the configuration
homotopy invariant	topologically conserved quantity
$[X, Y]$ has one element	no topological sectors (everything deforms to vacuum)

$[X, Y]$ is a topological invariant of Y

Nash observes that if X is fixed and Y is replaced by a homeomorphic space Y', the set $[X, Y]$ does not change. Let $h : Y \to Y'$ be a homeomorphism. Send each map $\alpha : X \to Y$ to $h \circ \alpha : X \to Y'$. If F is a homotopy from α to β, then $h \circ F$ is a homotopy from $h \circ \alpha$ to $h \circ \beta$:

$$(h \circ F)(0, x) \overset{(1)}{=} h\big(F(0, x)\big) \overset{(2)}{=} h\big(\alpha(x)\big), \qquad (h \circ F)(1, x) = h\big(\beta(x)\big),$$

(1) Definition of composition.
(2) F starts at α; the same argument at $t = 1$ gives β.

and $h \circ F$ is continuous as a composite of continuous maps. So $[\alpha] \mapsto [h \circ \alpha]$ is a well-defined map $[X, Y] \to [X, Y']$. Doing the same with h^{-1} gives the inverse map. Hence $[X, Y]$ and

$[X, Y']$ are canonically identified. The same argument, with h acting on X by precomposition, shows $[X, Y]$ is also a topological invariant of X.

The argument used much less than a homeomorphism. It only used that composition with a fixed continuous map carries homotopies to homotopies:

$$\alpha \simeq \beta \implies h \circ \alpha \simeq h \circ \beta \quad \text{and} \quad \alpha \circ k \simeq \beta \circ k \tag{1.9}$$

for any continuous $h : Y \to Y'$ and $k : X' \to X$ (use $h \circ F$ and $F(t, k(x'))$ respectively). That observation leads to the next notion.

1.2.2 Homotopy type

Homeomorphism asks for maps $\alpha : X \to Y$ and $\beta : Y \to X$ that are exact inverses. Homotopy theory asks only that they be inverses up to homotopy.

Working definition 1.10 (homotopy equivalence)

Spaces X and Y have the same **homotopy type** (are **homotopy equivalent**), written $X \simeq Y$, if there are continuous maps

$$\alpha : X \to Y, \qquad \beta : Y \to X, \qquad \alpha \circ \beta \simeq \mathrm{id}_Y, \qquad \beta \circ \alpha \simeq \mathrm{id}_X. \tag{1.10}$$

[here $\simeq$ between maps means homotopic, i.e. deformable into each other, not equal]. β is a **homotopy inverse** of α. A space homotopy equivalent to a single point is **contractible**. *Operational test:* guess a "shrinking" map α onto a simpler space and the inclusion β back [each point of the simpler space sent to itself inside the original one]; one composite is usually the identity on the nose; write down an explicit homotopy for the other.

Homotopy equivalence is an equivalence relation. Transitivity uses (1.9): if $X \simeq Y$ via (α, β) and $Y \simeq Z$ via (γ, δ), then

$$(\gamma \circ \alpha) \circ (\beta \circ \delta) \overset{(1)}{=} \gamma \circ (\alpha \circ \beta) \circ \delta \overset{(2)}{\simeq} \gamma \circ \mathrm{id}_Y \circ \delta = \gamma \circ \delta \overset{(3)}{\simeq} \mathrm{id}_Z,$$

(1) Composition of maps is associative.
(2) $\alpha \circ \beta \simeq \mathrm{id}_Y$, and composing on both sides with fixed maps preserves homotopy, (1.9).
(3) $\gamma \circ \delta \simeq \mathrm{id}_Z$ by assumption.

and similarly $(\beta \circ \delta) \circ (\gamma \circ \alpha) \simeq \mathrm{id}_X$.

Worked example 1.11: $\mathbb{R}^n$ is contractible

Let $\alpha : \mathbb{R}^n \to \{0\}$ send everything to the origin and let $\beta : \{0\} \to \mathbb{R}^n$ be the inclusion. Then $\alpha \circ \beta = \mathrm{id}_{\{0\}}$ exactly. The other composite $\beta \circ \alpha$ is the constant map $x \mapsto 0$. Use $F(t, x) = t\,x$:

$$F(0, x) = 0 = (\beta \circ \alpha)(x), \qquad F(1, x) = x = \mathrm{id}_{\mathbb{R}^n}(x).$$

So $\beta \circ \alpha \simeq \mathrm{id}$ and $\mathbb{R}^n \simeq \{\mathrm{pt}\}$. The homotopy is the movie in which space is uniformly shrunk to the origin (Figure 1.10, left).
Nash draws the moral. $\mathbb{R}^n$ $(n > 0)$ is *not* homeomorphic to a point (it is not even the same size as a set), and $\mathbb{R}^n \cong \mathbb{R}^m$ only if $n = m$. But all $\mathbb{R}^n$ have the homotopy type of a point. So dimension is a topological invariant which is *not* a homotopy invariant, and every homotopy invariant (homotopy groups, homology groups) of $\mathbb{R}^n$ equals that of a point, i.e. is trivial.

Worked example 1.12: The punctured plane has the homotopy type of a circle

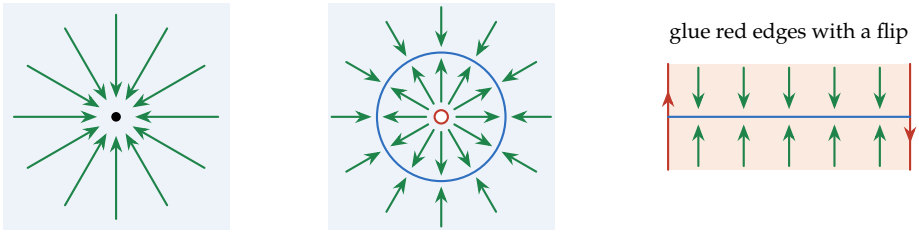
Let $X = \mathbb{R}^2 \setminus \{0\}$ and $Y = S^1$. Take $\alpha(x) = x/|x|$ (radial projection) and $\beta : S^1 \hookrightarrow X$ the inclusion. Then $\alpha \circ \beta = \mathrm{id}_{S^1}$ since unit vectors are already unit. For the other composite use the straight-line homotopy $F(t, x) = (1 - t)\,\frac{x}{|x|} + t\,x$. The danger is that F might hit the origin, which is not in X. It does not:

$$F(t,x) \overset{(1)}{=} \Big(\frac{1-t}{|x|} + t\Big)x, \qquad \text{and} \qquad \frac{1-t}{|x|} + t \overset{(2)}{>} 0,$$

(1) Factor out the vector x.

(2) For $t \in [0, 1]$ both terms are ≥ 0, and they cannot both vanish ($t = 0$ makes the first term $1/|x| > 0$).

So $F(t, x)$ is a positive multiple of $x \neq 0$: it slides each point along its own ray, never through the origin. $F(0, \cdot) = \beta \circ \alpha$ and $F(1, \cdot) = \mathrm{id}_X$. Hence $\mathbb{R}^2 \setminus \{0\} \simeq S^1$. In the same way $\mathbb{R}^n \setminus \{0\} \simeq S^{n-1}$, which is why the field of a point charge in $\mathbb{R}^3$ is "measured" by a sphere S^2 around it.



$\mathbb{R}^2 \simeq \{0\}$: $F(t, x) = tx$ $\quad$ $\mathbb{R}^2 \setminus \{0\} \simeq S^1$ $\quad$ Möbius band $\simeq S^1$

Figure 1.10: Three deformation retractions (green arrows show where points move as t runs from 1 down to 0). Left: the plane shrinks to a point. Middle: the punctured plane retracts onto the unit circle, points outside moving in and points inside moving out, never crossing the hole. Right: a Möbius band (a strip whose ends are glued with a flip) retracts onto its central circle, shown in blue.

Worked example 1.13: More spaces with the homotopy type of a circle

Each of the following retracts onto a circle by squashing the "thick" directions, exactly as in Figure 1.10:

- the cylinder $S^1 \times [0, 1]$ (squash $[0, 1]$ to a point) and $S^1 \times \mathbb{R}^k$ (use $F(t, (z, y)) = (z, ty)$);
- the Möbius band (squash each vertical segment to its midpoint);
- the solid torus $S^1 \times D^2$ (squash each disc to its centre);
- $\mathbb{R}^3$ minus a straight line, for instance the space around an infinitely long solenoid or a straight vortex line (project onto a plane perpendicular to the line and use Example 1.12);
- the thickened letters O, D, P, A, Q, R (one hole each). The letter B is homotopy equivalent to the figure eight $S^1 \vee S^1$ (two holes), and C, E, T, ... are contractible.

The cylinder and the Möbius band have the same homotopy type but are not homeomorphic (the Möbius band has only one boundary circle). Homotopy type forgets orientability.

Key idea: Invariants of homotopy type

If $Y \simeq Y'$ then $[X, Y] = [X, Y']$ for every X; if $X \simeq X'$ then $[X, Y] = [X', Y]$ for every Y. In Nash's form,

$$[X, Y] = [X, Y'] \quad \text{for } Y \text{ homeomorphic to } Y', \tag{1.11}$$
$$[X, Y] = [X, Y'] \quad \text{for } Y \text{ of the same homotopy type as } Y'. \tag{1.12}$$

The second is stronger. Most invariants of algebraic topology (homotopy groups, homology and cohomology groups, the Euler characteristic) are invariants of homotopy type.

Why does (1.12) hold? Let $\alpha : Y \to Y'$, $\beta : Y' \to Y$ be a homotopy equivalence. Define $\alpha_* : [X, Y] \to [X, Y']$ by $[f] \mapsto [\alpha \circ f]$ and β_* similarly. By (1.9) these are well defined, and

$$\beta_*\alpha_*[f] \overset{(1)}{=} [\beta \circ \alpha \circ f] \overset{(2)}{=} [\mathrm{id}_Y \circ f] = [f],$$

(1) Apply the definitions of α_* and then β_*.
(2) $\beta \circ \alpha \simeq \mathrm{id}_Y$; compose on the right with f using (1.9).

and likewise $\alpha_*\beta_* = \mathrm{id}$. So α_* is a bijection.

Watch out: Asking for a homotopy invariant is asking for less

Nash warns: an invariant of homotopy type is a *coarser* invariant than an invariant of the homeomorphism class. Homotopy invariants cannot tell $\mathbb{R}^3$ from a point, a Möbius band from a cylinder, or a solid torus from a circle. That is the price of computability. When we need finer information (dimension, smooth structure) we must use other tools.

1.2.3 Retracts and deformation retracts

All the examples above have the same shape: a big space X is squashed onto a subspace A that it contains, without moving the points of A. This pattern has a name.

Working definition 1.11 (retract, deformation retract)

Let $A \subset X$ and let $i : A \hookrightarrow X$ be the inclusion [each point of A sent to itself, now regarded as a point of X].

- A is a **retract** of X if there is a continuous $r : X \to A$ with $r(a) = a$ for all $a \in A$ (i.e. $r \circ i = \mathrm{id}_A$).
- A is a **deformation retract** of X if moreover $i \circ r \simeq \mathrm{id}_X$ [the squashing map $X \to A \subset X$ can be deformed continuously into doing nothing] by a homotopy that keeps A fixed at all times.

A deformation retract is automatically a homotopy equivalence ($\alpha = r$, $\beta = i$). *Operational test:* find a "squashing" formula $F(t, x)$, equal to x at $t = 1$, landing in A at $t = 0$, and fixing points of A.

Worked example 1.14: The circle is not a retract of the disc; Brouwer's fixed-point theorem

The punctured plane retracts onto S^1. The full disc D^2 does not. Suppose $r : D^2 \to S^1$ were a retraction. Consider the loop $\gamma(s) = (\cos 2\pi s, \sin 2\pi s)$ on the boundary circle, and

the homotopy $\Gamma(t,s) = r\big(t\,\gamma(s)\big)$ (shrink the loop into the disc, then apply r). Then

$$\Gamma(1,s) \overset{(1)}{=} r(\gamma(s)) \overset{(2)}{=} \gamma(s), \qquad \Gamma(0,s) = r(0) = \text{a constant point of } S^1.$$

(1) Set $t = 1$.
(2) r fixes the points of S^1, and $\gamma(s) \in S^1$.

So Γ would be a homotopy *in* S^1 from the loop that winds once around the circle to a constant loop. We will prove in §1.3.2 that the winding number is a homotopy invariant, and these two loops have winding numbers 1 and 0. Contradiction: no retraction exists.
A famous consequence is **Brouwer's fixed-point theorem**: every continuous $f : D^2 \to D^2$ has a fixed point. If not, $f(x) \neq x$ for all x, and we could define $r(x)$ as the point where the ray from $f(x)$ through x leaves the disc. This r is continuous and fixes the boundary, so it is a retraction $D^2 \to S^1$, which does not exist. Stir a cup of coffee (continuously, without splashing) and some point of the surface ends where it started.

Try it: Retracts of the torus

Show that the circle $S^1 \times \{1\}$ is a retract of the torus $T^2 = S^1 \times S^1$ but not a deformation retract. *Answer:* $r(z,w) = (z,1)$ is a retraction. It is not a deformation retract, since then $T^2 \simeq S^1$; but $\pi_1(T^2) = \mathbb{Z}\times\mathbb{Z}$ while $\pi_1(S^1) = \mathbb{Z}$ (§1.3.3), and π_1 is a homotopy invariant.

1.2.4 Base points: why loops need a common point

The most important choice of X in $[X,Y]$ is the circle: $[S^1,Y]$ is the set of deformation classes of closed loops in Y. We would like these classes to form a *group*, with the product "go around one loop, then the other". But two arbitrary loops in Y need not meet, so there is no obvious way to follow one by the other. The cure is simple and, as Nash says, "powerful": make all loops start and end at one agreed point.

Working definition 1.12 (pointed spaces and based maps)

A **pointed space** $\{X,*\}$ is a space X with a chosen **base point** $* = x_0 \in X$. A **based map** $\alpha : \{X,x_0\} \to \{Y,y_0\}$ is a continuous map with $\alpha(x_0) = y_0$. We write

$$\mathrm{Map}_0(X,Y) := \{\alpha \in \mathrm{Map}(X,Y) : \alpha(x_0) = y_0\}. \tag{1.13}$$

A **based homotopy** is a homotopy F [a continuous $F : [0,1]\times X \to Y$ deforming one map, $F(0,\cdot)$, into another, $F(1,\cdot)$] with $F(t,x_0) = y_0$ for all t: the base point is pinned throughout the movie. The set of based homotopy classes is

$$[X,Y]_0 := \mathrm{Map}_0(X,Y)/\!\simeq_{\text{based}}. \tag{1.14}$$

By contrast $[X,Y]$ is the set of **free** homotopy classes: neither the maps nor the homotopies need respect base points.

Every based class determines a free class (forget that the base point was pinned), giving a map $[X,Y]_0 \to [X,Y]$. It need not be one-to-one: two based maps can be freely homotopic without being based homotopic. We work out exactly how the two differ in §1.3.5.

Worked example 1.15: π_0: path components as based classes

Let $S^0 = \{-1, +1\}$, the "zero-sphere" (the solutions of $x^2 = 1$), with base point $+1$. A based map $\alpha : S^0 \to Y$ is determined by the single point $\alpha(-1) \in Y$, since $\alpha(+1) = y_0$ is forced. A based homotopy is a path $t \mapsto F(t, -1)$ moving that point around. So

$$[S^0, Y]_0 = \{\text{path components of } Y\} =: \pi_0(Y).$$

For example $\pi_0(\mathrm{O}(3))$ has two elements (rotations and reflections, $\det R = \pm 1$), and $\pi_0(\text{Lorentz group})$ has four. A space is **path-connected** if $\pi_0(Y)$ has one element: any two points can be joined by a path. π_0 has no group structure in general (it does if Y is a group: $\pi_0(\mathrm{O}(3)) \cong \mathbb{Z}_2$). In physics, π_0 of the vacuum manifold classifies domain walls (§1.3.13).

1.2.5 Maps into mapping spaces: the exponential law

We now drop base points for a moment. Nash records a simple but very useful bijection: for spaces X, Y, Z,

$$i : \mathrm{Map}\big(X, \mathrm{Map}(Y, Z)\big) \longrightarrow \mathrm{Map}(X \times Y, Z). \tag{1.15}$$

A map that assigns to each x a map $Y \to Z$ is the same thing as a map of two variables $(x, y) \mapsto Z$. Programmers call this *currying*. The formula is

$$i(g)(x, y) \coloneqq g(x)(y), \qquad \text{with inverse} \qquad i^{-1}(f)(x) \coloneqq f(x, \cdot).$$

Where does the name "exponential law" come from? Write Y^X for the set of all functions (continuous or not) from X to Y. If X and Y are finite sets, there are exactly $|Y|^{|X|}$ such functions (for each of the $|X|$ inputs choose one of $|Y|$ outputs), so the notation is honest:

$$|Y^X| = |Y|^{|X|}. \tag{1.16}$$

Then the set-theoretic version of (1.15) is the familiar law of exponents,

$$Z^{(Y \times X)} \cong Z^{(X \times Y)} \cong \left(Z^X\right)^Y, \tag{1.17}$$

and for finite sets we can check it by counting:

$$\left|(Z^X)^Y\right| \overset{(1)}{=} |Z^X|^{|Y|} \overset{(2)}{=} \left(|Z|^{|X|}\right)^{|Y|} \overset{(3)}{=} |Z|^{|X|\,|Y|} \overset{(4)}{=} \left|Z^{(X \times Y)}\right|.$$

(1) (1.16) with Z^X in place of Y and Y in place of X.
(2) (1.16) again.
(3) The ordinary law of exponents for numbers.
(4) $|X \times Y| = |X|\,|Y|$ and (1.16).

Nash's remark is that to establish (1.15) one only needs to check that the set-theoretic bijection (1.17) respects continuity. It does, provided we give $\mathrm{Map}(Y, Z)$ the natural topology in which "$g(x)$ close to $g(x')$" means the two maps are uniformly close on compact pieces of Y (the *compact-open topology*), and Y is a reasonable space (locally compact, as all manifolds are).

Worked example 1.16: A homotopy is a path in the space of maps

Apply (1.15) with $X = [0, 1]$: a homotopy $F : [0, 1] \times Y \to Z$ is the same as a path

$$t \longmapsto F_t = F(t, \cdot) \in \mathrm{Map}(Y, Z)$$

in the space of maps. Two maps are homotopic exactly when they lie in the same path

component of $\mathrm{Map}(Y,Z)$. Therefore

$$[Y,Z] = \pi_0\big(\mathrm{Map}(Y,Z)\big). \tag{1.18}$$

This is precisely the physicist's picture: $\mathrm{Map}(Y,Z)$ is the configuration space of a field theory on space Y with target Z; a finite-energy time evolution is a path in configuration space; and topological sectors are the disconnected pieces of configuration space. For instance, the configuration space of a U(1)-valued field on a circle, $\mathrm{Map}(S^1,\mathrm{U}(1))$, falls into pieces labelled by the winding number $n \in \mathbb{Z}$.

1.2.6 Wedge and smash products

We want a based version of (1.15). In it, X, Y and Z are pointed, and all maps are based. $\mathrm{Map}_0(Y,Z)$ has a natural base point: the constant map $y \mapsto z_0$. The one space in (1.15) with no natural base point is the product $X \times Y$: the obvious candidate is (x_0, y_0), but that turns out to be the wrong answer. Nash describes the right construction in two steps.

Working definition 1.13 (wedge and smash)

- The **wedge** (one-point union) $X \vee Y$ is the disjoint union [a copy of X and a copy of Y side by side, sharing no point] of X and Y with the two base points identified:

$$X \vee Y := (X \sqcup Y)/(x_0 \sim y_0). \tag{1.19}$$

  Inside $X \times Y$ it sits as the "axes" $(X \times \{y_0\}) \cup (\{x_0\} \times Y)$.

- The **smash product** (reduced product) collapses those axes to a single point, which becomes the base point:

$$X \wedge Y := \frac{X \times Y}{X \vee Y}. \tag{1.20}$$

Here A/B (for $B \subset A$) means: identify all points of B with each other, leave the rest alone. *Operational picture:* draw $X \times Y$, draw the two axes through the base point, and shrink both axes to one point.

Worked example 1.17: The figure eight, and $S^1 \wedge S^1 = S^2$

Take $X = Y = S^1$. Then $S^1 \vee S^1$ is the figure eight: two circles touching at one point. In the torus $S^1 \times S^1$ drawn as a square with opposite edges glued (Figure 1.4), the wedge $S^1 \vee S^1$ is exactly the union of the edges a and b, i.e. the whole boundary of the square. Collapsing it to a point leaves the *open* square plus one extra point. The open square is homeomorphic to $\mathbb{R}^2$ (use the map of Example 1.1 in each coordinate), and adding one point "at infinity" to $\mathbb{R}^2$ gives the sphere S^2 (inverse stereographic projection, Example 1.5). So

$$S^1 \wedge S^1 \cong S^2.$$

Think of pulling a drawstring bag closed around the boundary of the square (Figure 1.11).

Now the based exponential law. A based map $g : X \to \mathrm{Map}_0(Y,Z)$ gives, by currying, $f(x,y) = g(x)(y)$ on $X \times Y$. Which points of $X \times Y$ are automatically sent to z_0?

$$f(x,y_0) \overset{(1)}{=} g(x)(y_0) \overset{(2)}{=} z_0,$$

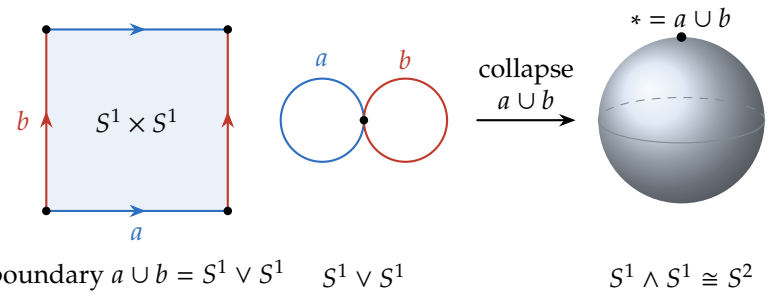


Figure 1.11: The smash product of two circles. The torus is a square with glued edges; the glued edges form the wedge $S^1 \vee S^1$ (a figure eight, middle). Collapsing the wedge to a single point closes the square up like a drawstring bag into a sphere, with the collapsed boundary as the base point.

$$f(x_0, y) \overset{(3)}{=} g(x_0)(y) \overset{(4)}{=} z_0.$$

(1) Definition of currying.
(2) Each $g(x)$ lies in $\mathrm{Map}_0(Y, Z)$, so it is a based map.
(3) Definition of currying.
(4) g is based, so $g(x_0)$ is the base point of $\mathrm{Map}_0(Y, Z)$, the constant map at z_0.

So f sends the whole wedge $X \vee Y = (X \times \{y_0\}) \cup (\{x_0\} \times Y)$ to z_0, and therefore descends to a based map on the quotient $X \wedge Y$. Conversely, any based map on $X \wedge Y$ uncurries to a based g. This is Nash's based exponential law:

$$i_* : \mathrm{Map}_0\big(X, \mathrm{Map}_0(Y, Z)\big) \xrightarrow{\cong} \mathrm{Map}_0(X \wedge Y, Z). \tag{1.21}$$

That is why the smash product, and not $X \times Y$ with base point (x_0, y_0), is the correct "product" of pointed spaces.

1.2.7 Suspension and loop space

Two special cases of the smash product and of based mapping spaces are used constantly.

Working definition 1.14 (suspension and loop space)

- The (reduced) **suspension** of a pointed space X is

$$SX \coloneqq S^1 \wedge X$$

[the smash product: $S^1 \times X$ with the axes $S^1 \vee X$ collapsed to one point]. Concretely: take the cylinder $[0, 1] \times X$, collapse the top $\{1\} \times X$, the bottom $\{0\} \times X$ and the vertical line $[0, 1] \times \{x_0\}$ all to one point.

- The **loop space** of X is the space of based loops,

$$\Omega X \coloneqq \mathrm{Map}_0(S^1, X) \tag{1.22}$$

[continuous maps $S^1 \to X$ sending the base point of S^1 to x_0]. We parametrise S^1 as $[0, 1]$ with $0 \sim 1$ the base point, so a point of ΩX is a path $\gamma : [0, 1] \to X$ with $\gamma(0) = \gamma(1) = x_0$.

The suspension is a "double cone" on X (Figure 1.12): each point x of X becomes a meridian running from the south pole ($t = 0$) to the north pole ($t = 1$). Suspending raises dimension by

one. The key example (Nash's (1.20)):

$$S^1 \wedge S^n = SS^n \cong S^{n+1}. \tag{1.23}$$

Worked example 1.18: Why $S^m \wedge S^n \cong S^{m+n}$

Use the fact (stereographic projection) that S^n is $\mathbb{R}^n$ plus one point at infinity: $S^n \cong (\mathbb{R}^n)^+$, where W^+ means "W with one point ∞ added", and we choose ∞ as the base point. Then

$$\begin{aligned} S^m \wedge S^n &\overset{(1)}{=} (\mathbb{R}^m)^+ \wedge (\mathbb{R}^n)^+ \\ &\overset{(2)}{=} \frac{(\mathbb{R}^m \cup \infty) \times (\mathbb{R}^n \cup \infty)}{(\mathbb{R}^m \cup \infty) \times \{\infty\} \ \cup \ \{\infty\} \times (\mathbb{R}^n \cup \infty)} \\ &\overset{(3)}{=} (\mathbb{R}^m \times \mathbb{R}^n) \cup \{\text{one point}\} \\ &\overset{(4)}{=} (\mathbb{R}^{m+n})^+ \overset{(5)}{=} S^{m+n}. \end{aligned}$$

(1) Stereographic projection, with the base point at ∞.
(2) Definition (1.20): the wedge consists of all pairs in which at least one entry is the base point ∞.
(3) The product of the two sets consists of the pairs (x, y) with both finite, which form $\mathbb{R}^m \times \mathbb{R}^n$, plus the pairs with at least one entry infinite, which is the wedge; the wedge is collapsed to one point.
(4) $\mathbb{R}^m \times \mathbb{R}^n = \mathbb{R}^{m+n}$, and the collapsed point is the new point at infinity (a sequence leaves every compact set of $\mathbb{R}^{m+n}$ exactly when one of its two components runs off to infinity).
(5) Stereographic projection in dimension $m + n$.

With $m = 1$ this is (1.23); with $m = n = 1$ it is Example 1.17 again. Iterating, $S^n \cong S^1 \wedge S^1 \wedge \cdots \wedge S^1$ (n factors) is the n-cube with its whole boundary collapsed to a point: $S^n \cong [0,1]^n/\partial[0,1]^n$. We use this cube picture of the sphere for the higher homotopy groups in §1.3.6.

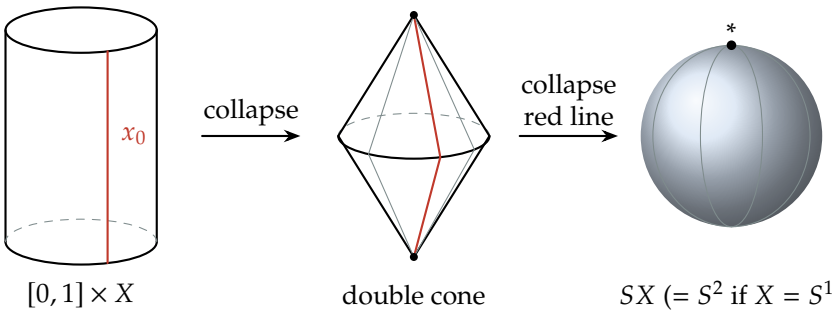


Figure 1.12: Building the reduced suspension $SX = S^1 \wedge X$. Start with the cylinder over X, collapse top and bottom (the double cone, or unreduced suspension), then collapse the red meridian through the base point x_0. For $X = S^1$ the result is the sphere S^2, drawn with meridians (grey), one for each point of X, all starting and ending at the base point.

The loop space ΩX is the space of closed strings pinned at one point of X. It is an infinite-dimensional space even when X is a finite manifold. Loop spaces of Lie groups, ΩG, are the loop groups of Chapter 6, which appear in two-dimensional conformal field theory and string theory.

S and Ω are adjoint

Take $Y = S^1$ in the based exponential law (1.21), and rename:

$$\mathrm{Map}_0(X, \Omega Z) \overset{(1)}{=} \mathrm{Map}_0\big(X, \mathrm{Map}_0(S^1, Z)\big) \overset{(2)}{=} \mathrm{Map}_0(X \wedge S^1, Z)$$
$$\overset{(3)}{=} \mathrm{Map}_0(S^1 \wedge X, Z) = \mathrm{Map}_0(SX, Z).$$

(1) Definition (1.22) of ΩZ.
(2) The based exponential law (1.21) with $Y = S^1$.
(3) $X \wedge S^1 \cong S^1 \wedge X$ by swapping the two factors.

Concretely, a based map $f : SX \to Z$, $f(s, x)$, corresponds to the map $g : X \to \Omega Z$ that sends x to the loop $s \mapsto f(s, x)$. This bijection (Nash's (1.21)) is compatible with homotopies (a homotopy is just one more parameter), so it passes to classes:

$$\mathrm{Map}_0(SX, Y) \cong \mathrm{Map}_0(X, \Omega Y), \tag{1.24}$$
$$[SX, Y]_0 \cong [X, \Omega Y]_0. \tag{1.25}$$

Physical picture: Adjoint, as in Hilbert space

Nash calls S and Ω *functors*: operations that act on spaces, and on maps between spaces, compatibly with composition. They are **adjoint**: (1.25) has exactly the shape of the definition of the adjoint of an operator,

$$\langle Sx, y\rangle = \langle x, \Omega\, y\rangle,$$

with the "inner product" $\langle A, B\rangle$ replaced by the set $[A, B]_0$. You can move S from the first slot to the second slot, where it becomes Ω. Suspension raises the dimension of the source; looping lowers the dimension of the target. The two effects cancel in $[\cdot, \cdot]_0$.

Worked example 1.19: A sphere of maps is a loop of loops

Take $X = S^1$ in (1.25): since $SS^1 = S^2$,

$$[S^2, Y]_0 \cong [S^1, \Omega Y]_0.$$

A based map of the sphere into Y is a based loop in the space of based loops. To see it, use the meridian picture of Figure 1.12: each meridian of S^2 (labelled by a point x of the equator S^1) is mapped to a loop in Y through y_0. As x goes once around the equator, we get a closed family of loops, starting and ending at the constant loop (the meridian through the base point is collapsed). Physically: a static configuration of a field on a sphere is the same as a periodic "movie" of field configurations on a line segment with fixed ends. This is how Nash will compute (1.44) below: $\pi_n(Y) = \pi_{n-1}(\Omega Y)$.

1.2.8 Group structures on $[X, Y]_0$: H-spaces and coH-spaces

Groups are extremely valuable, so we want the sets $[X, Y]_0$ to be groups. There are two ways this can happen, dual to each other: a structure on the target Y (Nash's "P"), or a structure on the source X (Nash's "P'").

Structure on the target: H-spaces

Suppose first that $Y = G$ is a group (a topological group: multiplication and inversion are continuous), with base point the identity e. Then maps into G can be multiplied pointwise.

Working definition 1.15 (pointwise product)

For $\alpha, \beta \in \mathrm{Map}_0(X, G)$ [continuous maps $X \to G$ sending the base point of X to the identity e] define

$$\alpha \bullet \beta : X \to G, \qquad x \longmapsto \alpha(x) * \beta(x), \tag{1.26}$$

where $*$ is the product in G. *Physics reading:* if X is space and G a gauge group, $\mathrm{Map}(X, G)$ is the group of gauge transformations $g(x)$, and $\bullet$ is how two gauge transformations compose. This is the gauge group $\mathcal{G}$.

This makes $\mathrm{Map}_0(X, G)$ itself a group: $\alpha \bullet \beta$ is based since $e * e = e$, the identity element is the constant map e, and the inverse of α is $x \mapsto \alpha(x)^{-1}$. It passes to homotopy classes: if F is a based homotopy from α to α' and H one from β to β', then $(t, x) \mapsto F(t, x) * H(t, x)$ is a based homotopy from $\alpha \bullet \beta$ to $\alpha' \bullet \beta'$. So

$$[\alpha] \bullet [\beta] \coloneqq [\alpha \bullet \beta]$$

is well defined, and $[X, G]_0$ is a group, for *any* X.

Worked example 1.20: $[S^1, \mathrm{U}(1)]_0$ with the pointwise product

Let $G = \mathrm{U}(1) = \{e^{i\theta}\}$ and $X = S^1$, parametrised by $s \in [0, 1]$. The maps $\alpha_n(s) = e^{2\pi i n s}$ ($n \in \mathbb{Z}$) are based ($\alpha_n(0) = 1$) and wind n times around $\mathrm{U}(1)$. Their pointwise product is

$$(\alpha_m \bullet \alpha_n)(s) \overset{(1)}{=} e^{2\pi i m s}\, e^{2\pi i n s} \overset{(2)}{=} e^{2\pi i (m+n) s} = \alpha_{m+n}(s).$$

(1) Definition (1.26), with $*$ ordinary multiplication of complex numbers.
(2) $e^a e^b = e^{a+b}$.

So under $\bullet$, winding numbers *add*. We will see in §1.3.2 that every based loop in $\mathrm{U}(1)$ is homotopic to exactly one α_n, so $([S^1, \mathrm{U}(1)]_0, \bullet) \cong (\mathbb{Z}, +)$. In gauge theory on a circle (or on the S^1 at the edge of a vortex), α_n is a *large* gauge transformation when $n \neq 0$: it cannot be deformed to the identity.

Nash's next observation: nothing in this argument used that G is a group on the nose. We only needed a continuous multiplication $* : Y \times Y \to Y$ whose group axioms hold *up to homotopy*, because we only ever look at homotopy classes.

Working definition 1.16 (H-space)

A pointed space Y is a (homotopy associative) **H-space** if it has a continuous multiplication $* : Y \times Y \to Y$ such that, up to based homotopy [the two sides need not be equal, only deformable into each other with the base point held fixed],

- the base point is a two-sided identity: $y * y_0 \simeq y \simeq y_0 * y$;
- $*$ is associative: $(y * y') * y'' \simeq y * (y' * y'')$ (as maps $Y \times Y \times Y \to Y$);
- there is an inverse map $\nu : Y \to Y$ with $y * \nu(y) \simeq y_0$.

Then $[X, Y]_0$ is a group for every X, under the pointwise product. *Examples:* every

topological group; every loop space ΩX [based loops in X, multiplied by running one loop and then the other] (we check this in §1.3.1).

The phrase "homotopy associative" is the whole point. On $\mathrm{Map}_0(X,Y)$ itself the product need not be associative; only on the classes $[X,Y]_0$ is it associative.

Structure on the source: coH-spaces

Now the dual idea. Instead of a multiplication $Y \times Y \to Y$ on the target, put a *comultiplication* on the source.

Working definition 1.17 (coH-space)

A pointed space X is a (homotopy associative) **coH-space** if it has a continuous based map

$$*' : X \longrightarrow X \vee X \tag{1.27}$$

[$X \vee X$ is two copies of X joined at their base points] (a "comultiplication") satisfying the duals of the H-space axioms [for example, $*'$ followed by squashing the second copy to its base point is homotopic to id_X] up to homotopy. Given $\alpha, \beta \in \mathrm{Map}_0(X,Y)$, define $\alpha \vee \beta : X \vee X \to Y$ to be α on the first copy and β on the second (consistent at the shared base point since both send it to y_0), and

$$\bullet' : [X,Y]_0 \times [X,Y]_0 \to [X,Y]_0, \qquad [\alpha] \bullet' [\beta] := \big[(\alpha \vee \beta) \circ *'\big]. \tag{1.28}$$

Then $[X,Y]_0$ is a group for every Y.

Watch out: Why $X \vee X$ and not $X \times X$?

Nash notes that $X \vee X$ appears "where one would expect $X \times X$". The reason is concrete. To combine α and β on the target of $*'$ we need a map out of it built from α and β alone. On $X \vee X$ this is $\alpha \vee \beta$: each point lies in one copy or the other. On $X \times X$ we would have to assign a single value to a pair (x, x') using both $\alpha(x)$ and $\beta(x')$, which requires multiplying points of Y: exactly the H-space structure we are trying not to assume. (Also, $X \vee X$ is where base points behave: $\alpha \vee \beta$ is automatically based.)

Worked example 1.21: Suspensions are coH-spaces: the pinch map

Let $X = SZ$, with points written (t,z), $t \in [0,1]$ (Figure 1.12; all of $t=0$, $t=1$ and $z = z_0$ is the base point). Define the **pinch map**: squeeze the "equator" $t = \frac{1}{2}$ to the base point, which splits SZ into two copies of itself glued at the base point:

$$*'(t,z) = \begin{cases} (2t,z) \text{ in the first copy of } SZ, & 0 \le t \le \tfrac{1}{2}, \\ (2t-1,z) \text{ in the second copy of } SZ, & \tfrac{1}{2} \le t \le 1. \end{cases}$$

At $t = \frac{1}{2}$ the two formulas give $(1,z)$ and $(0,z)$, both the base point, so $*'$ is continuous. The induced product (1.28) is

$$(\alpha \bullet' \beta)(t,z) = \begin{cases} \alpha(2t,z), & 0 \le t \le \tfrac{1}{2}, \\ \beta(2t-1,z), & \tfrac{1}{2} \le t \le 1, \end{cases}$$

"do α on the southern half, β on the northern half". For $Z = S^0$ we have $SZ = S^1$ and this is the concatenation of loops, Nash's (1.27), the product in the fundamental group. For

$Z = S^{n-1}$ we get $SZ = S^n$ and the product of the higher homotopy groups (Figure 1.13).

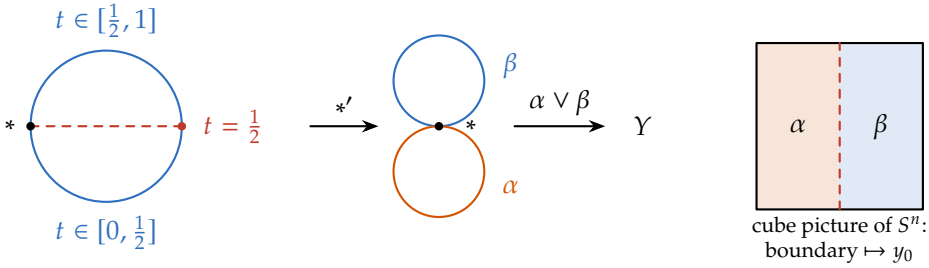


Figure 1.13: The pinch map. Squeezing the point $t = \frac{1}{2}$ of the loop parameter to the base point (the red point meets $*$) gives a figure eight $S^1 \vee S^1$; then $\alpha \vee \beta$ runs α on one lobe and β on the other. This is loop concatenation. In the cube picture of S^n (right), the pinch splits the cube in half along the first coordinate: α fills one half, β the other, and the whole boundary goes to y_0.

When both structures are present

Nash ends §2 with an "elementary calculation": if Y is an H-space *and* X is a coH-space, then the two group structures $\bullet$ and $\bullet'$ on $[X, Y]_0$ coincide, and the resulting group is *commutative*. Since the calculation is short and is the reason why higher homotopy groups are abelian, we do it. It is known as the **Eckmann–Hilton argument**.

The one fact we need is the *interchange law*:

$$(\alpha \bullet' \beta) \bullet (\gamma \bullet' \delta) = (\alpha \bullet \gamma) \bullet' (\beta \bullet \delta). \tag{1.29}$$

Both sides are the same map: take $X = SZ$ with the pinch map for concreteness. On the southern half ($t \le \frac{1}{2}$) the left side is $\alpha(2t, z) * \gamma(2t, z)$ and on the northern half $\beta(2t-1, z) * \delta(2t-1, z)$; the right side is literally the same. Write e for the constant map at the base point; it is the identity for both products (up to homotopy). Then for any α, δ:

$$\alpha \bullet \delta \overset{(1)}{=} (\alpha \bullet' e) \bullet (e \bullet' \delta) \overset{(2)}{=} (\alpha \bullet e) \bullet' (e \bullet \delta) \overset{(3)}{=} \alpha \bullet' \delta,$$
$$\alpha \bullet \delta \overset{(4)}{=} (e \bullet' \alpha) \bullet (\delta \bullet' e) \overset{(5)}{=} (e \bullet \delta) \bullet' (\alpha \bullet e) \overset{(6)}{=} \delta \bullet' \alpha \overset{(7)}{=} \delta \bullet \alpha.$$

(1) e is the identity for $\bullet'$ (in homotopy classes).
(2) The interchange law (1.29) with $\beta = \gamma = e$.
(3) e is the identity for $\bullet$.
(4) e is the identity for $\bullet'$, now inserted on the other side.
(5) The interchange law with $(\alpha, \beta, \gamma, \delta) \to (e, \alpha, \delta, e)$.
(6) e is the identity for $\bullet$.
(7) The first line, with α and δ swapped: $\bullet' = \bullet$.

The first line says the two products agree; the second says the product is commutative. All equalities are of homotopy classes.

Key idea: Consequence

If Y is an H-space and X a coH-space (for example $Y = G$ a group and $X = SZ$ a suspension), then $[X, Y]_0$ has a single natural group structure, and it is abelian. In particular $[SZ, G]_0$ is always an abelian group.

Reading Nash (pp. 3–9, (1.2)–(1.25))

Nash's §2 runs: invariants (numbers, groups); $\mathrm{Map}(X,Y)$ (1.2), homotopy (1.3)–(1.5); classes $[X,Y]$ (1.6) and their invariance under homeomorphism of Y; homotopy type (1.7), (1.8); invariants of homotopy type (1.9), (1.10) (our (1.11), (1.12)); $\mathbb{R}^n$ versus a point; base points, Map_0 (1.11), $[X,Y]_0$ (1.12); exponential law (1.13)–(1.15); wedge (1.16), smash (1.17), based exponential law (1.18); $SX = S^1 \wedge X$, ΩX (1.19), $SS^n \cong S^{n+1}$ (1.20); adjunction (1.21), (1.22); pointwise product (1.23), H-spaces; comultiplication (1.24) and product (1.25), coH-spaces; coincidence and commutativity when both are present. He cites Whitehead for (1.20), Husemoller for coH-spaces and Spanier for the "elementary calculation", which is our Eckmann–Hilton chain. Nash writes $\cong$ for bijections and $\simeq$ for homotopy, as we do. Note that his (1.9) contains the misprint "$[X,Y]$".

Recap

- A homotopy is a continuous movie of maps; $[X,Y]$ is the set of topological sectors, $[X,Y] = \pi_0(\mathrm{Map}(X,Y))$.
- $X \simeq Y$ (homotopy equivalence) means maps both ways whose composites are homotopic to identities. $\mathbb{R}^n \simeq$ point, $\mathbb{R}^n \setminus 0 \simeq S^{n-1}$, Möbius band $\simeq S^1$. $[X,Y]$ depends only on homotopy types.
- A deformation retract squashes X onto $A \subset X$ keeping A fixed. S^1 is not a retract of D^2, which gives Brouwer's theorem.
- Based maps and based homotopies give $[X,Y]_0$. The correct product of pointed spaces is the smash $X \wedge Y = X \times Y / X \vee Y$, with $S^m \wedge S^n = S^{m+n}$.
- $SX = S^1 \wedge X$ and $\Omega X = \mathrm{Map}_0(S^1, X)$ are adjoint: $[SX,Y]_0 = [X,\Omega Y]_0$.
- A group (H-space) target or a suspension (coH-space) source makes $[X,Y]_0$ a group; with both, the group is abelian (Eckmann–Hilton).

1.3 Homotopy groups

We now have all the machinery of §1.2: based maps, homotopy classes $[X,Y]_0$, suspensions, loop spaces, and two ways of making $[X,Y]_0$ a group. Nash's §3 applies this machinery to the simplest possible sources, the spheres. The questions are concrete:

- How many essentially different ways are there to draw a closed loop in a space X? (The fundamental group $\pi_1(X)$.)
- How many essentially different ways are there to wrap a sphere S^n around X? (The higher homotopy groups $\pi_n(X)$.)

For a physicist these are the questions "which vortices can exist?", "which monopoles?", "which instantons?". A vortex line in three-space is surrounded by a circle, so its charge is a class of loops in the space of field values. A point defect is surrounded by a sphere S^2. A finite-action gauge field on $\mathbb{R}^4$ is controlled by what happens on the sphere S^3 at infinity. We will compute the basic examples by hand, using nothing more than the winding-number integral $\oint \mathrm{d}\theta/2\pi$ and some 2×2 matrices. At the end of the section we collect the physics.

1.3.1 The fundamental group

Recall that we parametrise the circle by $t \in [0,1]$ with the ends glued, $0 \sim 1$, and take the glued point as the base point (the circle has unit circumference, as in Nash). A based map $S^1 \to X$ is therefore the same thing as a path that starts and ends at the base point.

Working definition 1.18 (fundamental group)

Let $\{X, x_0\}$ be a pointed space. A **loop at** x_0 is a continuous $\alpha : [0,1] \to X$ with $\alpha(0) = \alpha(1) = x_0$. The **fundamental group** of X is the set of based homotopy classes of such loops,

$$\pi_1(X) \coloneqq [S^1, X]_0. \tag{1.30}$$

Two loops define the same element iff one can be deformed into the other through loops at x_0, keeping the base point pinned at all times. *Operational test:* to show $[\alpha] = [\beta]$, write down $F(s,t)$ with $F(0,t) = \alpha(t)$, $F(1,t) = \beta(t)$ and $F(s,0) = F(s,1) = x_0$; to show $[\alpha] \neq [\beta]$, find an invariant (for example a winding number [the net number of times the loop goes around a hole, counted with sign]) that takes different values on them.

Because $\pi_1(X)$ consists of classes, we never distinguish between two homotopic loops. What makes π_1 a *group* is the product: first run around α, then around β, each at double speed.

Working definition 1.19 (product of loops)

For loops α, β at x_0, the product $\alpha \cdot \beta$ is the loop

$$(\alpha \cdot \beta)(t) \coloneqq \begin{cases} \alpha(2t), & 0 \le t \le \tfrac{1}{2}, \\ \beta(2t-1), & \tfrac{1}{2} \le t \le 1, \end{cases} \qquad [\alpha][\beta] \coloneqq [\alpha \cdot \beta]. \tag{1.31}$$

[$[\alpha]$ is the class of all loops at x_0 that can be deformed into α with x_0 held fixed]

This is Nash's (1.27). Let us check that $\alpha \cdot \beta$ really is a loop, i.e. continuous at the junction and based:

$$\alpha(2 \cdot \tfrac{1}{2}) \overset{(1)}{=} \alpha(1) \overset{(2)}{=} x_0 \overset{(3)}{=} \beta(0) = \beta(2 \cdot \tfrac{1}{2} - 1),$$
$$(\alpha \cdot \beta)(0) = \alpha(0) \overset{(4)}{=} x_0, \qquad (\alpha \cdot \beta)(1) = \beta(1) \overset{(4)}{=} x_0.$$

(1) Substitute $t = \frac{1}{2}$ into the first branch.
(2) α is a loop at x_0, so it ends at x_0.
(3) β is a loop at x_0, so it starts at x_0; the second branch at $t = \frac{1}{2}$ gives the same point, so the two halves join up (gluing lemma).
(4) Both loops start and end at the base point.

We saw in Example 1.21 that (1.31) is precisely the coH-space product $\bullet'$ coming from the pinch map of $S^1 = SS^0$ (Figure 1.13). The product is also well defined on classes: if F deforms α to α' and G deforms β to β', then running $F(s,2t)$ on the first half and $G(s,2t-1)$ on the second deforms $\alpha \cdot \beta$ to $\alpha' \cdot \beta'$.

The group axioms hold up to homotopy

On the nose, the product (1.31) is *not* associative: $(\alpha \cdot \beta) \cdot \gamma$ spends a quarter of the time on α, a quarter on β and a half on γ, while $\alpha \cdot (\beta \cdot \gamma)$ spends a half on α. The two loops trace the same

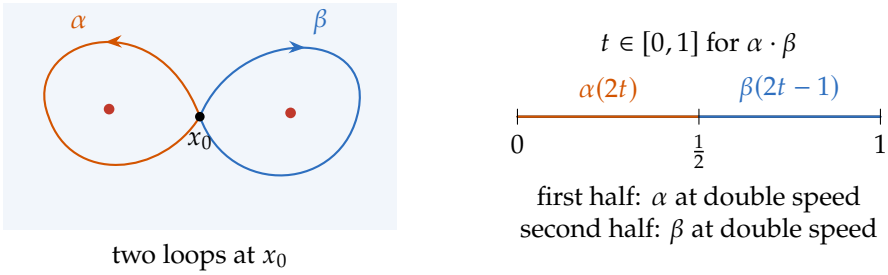


Figure 1.14: The product of loops in the plane with two holes (red). Both loops start and end at x_0, so we can run one after the other. The product $\alpha \cdot \beta$ spends the first half of its time on α and the second half on β. Neither loop can be shrunk to x_0 without crossing a hole.

route at different speeds. Nash calls checking the axioms up to homotopy "the well known process". It rests on a single fact, which we state and prove.

Key idea: Reparametrisation does not change the class

Let $\phi : [0,1] \to [0,1]$ be continuous with $\phi(0) = 0$ and $\phi(1) = 1$. Then for every loop α, $\alpha \circ \phi \simeq \alpha$ as based loops.

The homotopy slides the time variable linearly from $\phi(t)$ to t: $F(s,t) = \alpha\big((1-s)\phi(t) + st\big)$. The argument $(1-s)\phi(t) + st$ is a weighted average of two numbers in $[0,1]$, so it lies in $[0,1]$ and F makes sense. Check the ends:

$$F(0,t) = \alpha(\phi(t)), \qquad F(1,t) = \alpha(t),$$

$$F(s,0) \overset{(1)}{=} \alpha\big((1-s)\cdot 0 + 0\big) = \alpha(0) = x_0, \qquad F(s,1) \overset{(2)}{=} \alpha\big((1-s)+s\big) = \alpha(1) = x_0.$$

(1) $\phi(0) = 0$.
(2) $\phi(1) = 1$.

So the base point stays pinned. Now the axioms follow.

- *Associativity.* $(\alpha \cdot \beta) \cdot \gamma = \big(\alpha \cdot (\beta \cdot \gamma)\big) \circ \phi$, where ϕ is the piecewise-linear map with $\phi(\frac{1}{4}) = \frac{1}{2}$, $\phi(\frac{1}{2}) = \frac{3}{4}$:
$$\phi(t) = \begin{cases} 2t, & 0 \le t \le \frac{1}{4}, \\ t + \frac{1}{4}, & \frac{1}{4} \le t \le \frac{1}{2}, \\ \frac{t+1}{2}, & \frac{1}{2} \le t \le 1. \end{cases}$$
Check on the first quarter: $\big(\alpha \cdot (\beta \cdot \gamma)\big)(\phi(t)) = \big(\alpha \cdot (\beta \cdot \gamma)\big)(2t) = \alpha(4t)$, which is what $(\alpha \cdot \beta) \cdot \gamma$ does for $t \in [0, \frac{1}{4}]$. The other two pieces are checked the same way. The keyidea gives $[(\alpha \cdot \beta) \cdot \gamma] = [\alpha \cdot (\beta \cdot \gamma)]$.
- *Identity.* Let e be the constant loop, $e(t) = x_0$. Then $e \cdot \alpha = \alpha \circ \phi$ with $\phi(t) = 0$ for $t \le \frac{1}{2}$ and $\phi(t) = 2t - 1$ for $t \ge \frac{1}{2}$. So $[e][\alpha] = [\alpha]$, and similarly $[\alpha][e] = [\alpha]$.
- *Inverse.* Let $\bar{\alpha}(t) := \alpha(1-t)$ be α run backwards. Then $\alpha \cdot \bar{\alpha} \simeq e$ via "go out along α only as far as $1-s$, then come back":
$$H(s,t) = \begin{cases} \alpha\big(2t(1-s)\big), & 0 \le t \le \frac{1}{2}, \\ \alpha\big((2-2t)(1-s)\big), & \frac{1}{2} \le t \le 1. \end{cases}$$
At $s = 0$ this is $\alpha(2t)$ followed by $\alpha(2-2t) = \bar{\alpha}(2t-1)$, i.e. $\alpha \cdot \bar{\alpha}$. At $s = 1$ it is $\alpha(0) = x_0$ throughout, i.e. e. At $t = \frac{1}{2}$ both branches give $\alpha(1-s)$, so H is continuous, and $H(s,0) = H(s,1) = \alpha(0) = x_0$.

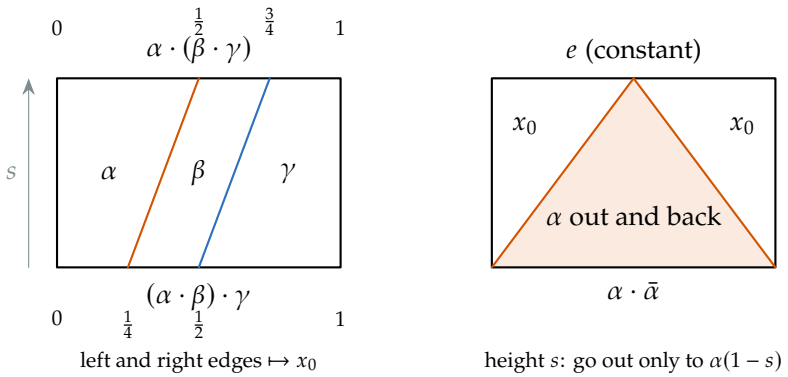


Figure 1.15: The group axioms as pictures of homotopies. Each square is the (t, s) domain of a based homotopy: the bottom edge is the starting loop, the top edge the final loop, and the vertical edges go to x_0. Left: the associativity homotopy just slides the break points $\frac{1}{4}, \frac{1}{2}$ to $\frac{1}{2}, \frac{3}{4}$. Right: the loop $\alpha \cdot \bar{\alpha}$ is reeled in; at height s it goes out along α only as far as $\alpha(1-s)$ and retraces its steps. The white regions map to x_0.

Key idea: $\pi_1(X)$ is a group, and ΩX is an H-space

With the product (1.31), identity $[e]$ and inverse $[\alpha]^{-1} = [\bar{\alpha}]$, the set $\pi_1(X)$ is a group. In Nash's words, checking this "is precisely the verification that ΩX is (another example of) a H-space": the loop product is a multiplication $\Omega X \times \Omega X \to \Omega X$ that is associative, unital and invertible up to homotopy.

A space whose fundamental group is trivial (every loop contracts to the base point) is called **simply connected**.

Worked example 1.22: Convex sets are simply connected

Let $X \subset \mathbb{R}^n$ be convex (for example $\mathbb{R}^n$ itself or a ball) and α a loop at x_0. The straight-line homotopy to the constant loop is $F(s,t) = (1-s)\,\alpha(t) + s\,x_0$. It stays in X by convexity, and it is based:

$$F(s,0) \overset{(1)}{=} (1-s)\,x_0 + s\,x_0 = x_0, \qquad F(s,1) \overset{(1)}{=} x_0.$$

(1) $\alpha(0) = \alpha(1) = x_0$.

So $\pi_1(X) = 0$. (We write 0 or 1 for the trivial group.) The same argument with $x_0 = 0$ works for any *star-shaped* set, one containing the segment from the origin to each of its points.

Maps induce homomorphisms

A based map $f : \{X, x_0\} \to \{Y, y_0\}$ sends loops in X to loops in Y, and by (1.9) homotopic loops to homotopic loops. So it defines $f_* : \pi_1(X) \to \pi_1(Y)$, $f_*[\alpha] = [f \circ \alpha]$. It respects products, because applying f commutes with running two loops in succession:

$$f \circ (\alpha \cdot \beta)(t) \overset{(1)}{=} \begin{cases} f(\alpha(2t)) & t \le \tfrac{1}{2} \\ f(\beta(2t-1)) & t \ge \tfrac{1}{2} \end{cases} \overset{(2)}{=} \big((f \circ \alpha) \cdot (f \circ \beta)\big)(t).$$

(1) Apply f to each branch of (1.31).
(2) This is the definition of the product of the loops $f \circ \alpha$ and $f \circ \beta$.

So f_* is a group homomorphism. Clearly $(g \circ f)_* = g_* \circ f_*$ and $(\mathrm{id})_* = \mathrm{id}$. If f is a based

homotopy equivalence, f_* is therefore an isomorphism, by the same argument we gave for (1.12). The fundamental group is an invariant of homotopy type.

1.3.2 The circle: $\pi_1(S^1) = \mathbb{Z}$ and the winding number

Everything else in this section rests on one computation: the fundamental group of the circle. Physicists know the answer in their bones. The phase $e^{i\theta(\vec{x})}$ of a superfluid order parameter, or of a single-valued wavefunction, can wind around the circle an integer number of times as $\vec{x}$ goes around a closed path, and that integer cannot change under small deformations. We make this precise.

Unwrapping the circle

Write $S^1 = \mathrm{U}(1) = \{z \in \mathbb{C} : |z| = 1\}$ with base point 1, and let

$$p : \mathbb{R} \to S^1, \qquad p(\theta) = e^{i\theta}$$

be the map that wraps the real line around the circle, once every 2π (Figure 1.16). Draw $\mathbb{R}$ as a helix sitting above the circle: p is the vertical projection. Above each point $z \in S^1$ sits a whole ladder of points $\theta, \theta \pm 2\pi, \theta \pm 4\pi, \dots$, all the possible values of "arg z".

Key idea: Path lifting

Every continuous path $\gamma : [0,1] \to S^1$ has a **lift**: a continuous $\tilde{\theta} : [0,1] \to \mathbb{R}$ with $e^{i\tilde{\theta}(t)} = \gamma(t)$. Once $\tilde{\theta}(0)$ is chosen from the ladder above $\gamma(0)$, the lift is unique.

Why: over any arc of the circle shorter than a half-turn, the argument function has a continuous branch (for instance $\arg z = \arctan(y/x)$ plus a constant). Chop $[0,1]$ into many short pieces on each of which γ moves by less than a half-turn (possible since γ is uniformly continuous on the compact interval). On the first piece use the branch that matches $\tilde{\theta}(0)$; on each next piece use the branch that agrees with the previous one at the junction. Uniqueness holds because two lifts differ by a continuous function with values in $2\pi\mathbb{Z}$, hence by a constant, and that constant is zero at $t = 0$. Physically: to follow the phase of a wavefunction continuously, you are never allowed to jump by 2π.

Working definition 1.20 (winding number)

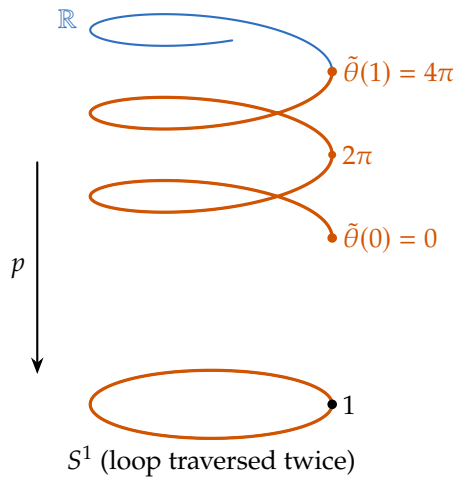
For a loop γ in S^1 based at 1, let $\tilde{\theta}$ be its lift [a continuous real-valued angle $\tilde{\theta}(t)$ with $\gamma(t) = e^{i\tilde{\theta}(t)}$, not reduced mod 2π] with $\tilde{\theta}(0) = 0$. Since $e^{i\tilde{\theta}(1)} = \gamma(1) = 1$, the endpoint $\tilde{\theta}(1)$ is a multiple of 2π. The **winding number** (or **degree**) is

$$W[\gamma] := \frac{\tilde{\theta}(1) - \tilde{\theta}(0)}{2\pi} \in \mathbb{Z}. \tag{1.32}$$

For a smooth loop it is an integral. Writing $\gamma(t) = e^{i\tilde{\theta}(t)}$,

$$W[\gamma] = \frac{1}{2\pi}\oint_\gamma \mathrm{d}\theta = \frac{1}{2\pi i}\oint_\gamma \frac{\mathrm{d}z}{z} = \frac{1}{2\pi i}\int_0^1 \frac{\dot{\gamma}(t)}{\gamma(t)}\,\mathrm{d}t. \tag{1.33}$$

For a loop in $\mathbb{C} \setminus \{0\}$ that is not on the unit circle, use the same integral: it measures the winding of $\gamma/|\gamma|$.



A loop γ in S^1 (orange, below) lifts to a path $\tilde\theta$ on the helix (orange, above). The lift need not close up: it ends at $\tilde\theta(1) = 2\pi W$, and

$$W = \frac{\tilde\theta(1) - \tilde\theta(0)}{2\pi} = \frac{1}{2\pi}\oint_\gamma d\theta$$

counts the turns. Here $\gamma(t) = e^{4\pi i t}$ runs twice around the circle and its lift climbs two storeys: $W = 2$.

Figure 1.16: Unwrapping the circle. The real line (blue helix) covers the circle by vertical projection p. A path on the circle lifts uniquely to the helix once a starting storey is chosen. A closed loop on the circle lifts to a path that ends an integer number of storeys (2π each) above where it started. That integer is the winding number.

The equalities in (1.33) are a two-line computation:

$$\frac{1}{2\pi i}\int_0^1 \frac{\dot\gamma}{\gamma}\,dt \overset{(1)}{=} \frac{1}{2\pi i}\int_0^1 \frac{i\dot{\tilde\theta}\,e^{i\tilde\theta}}{e^{i\tilde\theta}}\,dt \overset{(2)}{=} \frac{1}{2\pi}\int_0^1 \dot{\tilde\theta}\,dt \overset{(3)}{=} \frac{\tilde\theta(1)-\tilde\theta(0)}{2\pi}.$$

(1) Chain rule: $\dot\gamma = \frac{d}{dt}e^{i\tilde\theta} = i\dot{\tilde\theta}\,e^{i\tilde\theta}$.
(2) Cancel $e^{i\tilde\theta} \neq 0$ and the factor i.
(3) The fundamental theorem of calculus.

Here "$d\theta$" is the familiar 1-form $d\theta = (x\,dy - y\,dx)/(x^2+y^2)$ on the punctured plane. It is closed but not exact: there is no global function θ (the angle jumps by 2π somewhere), which is exactly why its integral around a loop can be nonzero. We return to this in Chapter T2 and in §4 of this chapter (de Rham cohomology).

Worked example 1.23: The standard loops α_n

Let $\alpha_n(t) = e^{2\pi i n t}$, $n \in \mathbb{Z}$. Then

$$W[\alpha_n] \overset{(1)}{=} \frac{1}{2\pi i}\int_0^1 \frac{2\pi i n\, e^{2\pi i n t}}{e^{2\pi i n t}}\,dt \overset{(2)}{=} n\int_0^1 dt = n.$$

(1) Formula (1.33) with $\dot\alpha_n = 2\pi i n\, e^{2\pi i n t}$.
(2) Cancel.

Equivalently, the lift is $\tilde\theta(t) = 2\pi n t$, which climbs n storeys. For $n < 0$ the loop runs clockwise.

Worked example 1.24: An off-centre circle: winding by Fourier analysis

Let $\gamma_a(\phi) = a + e^{i\phi}$, $\phi \in [0, 2\pi]$, with real $a \neq \pm 1$: the unit circle centred at a. It avoids 0, and we ask how often it winds around 0. Here $dz = ie^{i\phi}d\phi$.

Case $|a| < 1$. Expand $1/z$ in a geometric series in $ae^{-i\phi}$:

$$W \overset{(1)}{=} \frac{1}{2\pi i}\int_0^{2\pi} \frac{ie^{i\phi}}{e^{i\phi}(1+ae^{-i\phi})}\,\mathrm{d}\phi \overset{(2)}{=} \frac{1}{2\pi}\int_0^{2\pi}\sum_{k\geq 0}(-a)^k e^{-ik\phi}\,\mathrm{d}\phi \overset{(3)}{=} \frac{1}{2\pi}\cdot 2\pi = 1.$$

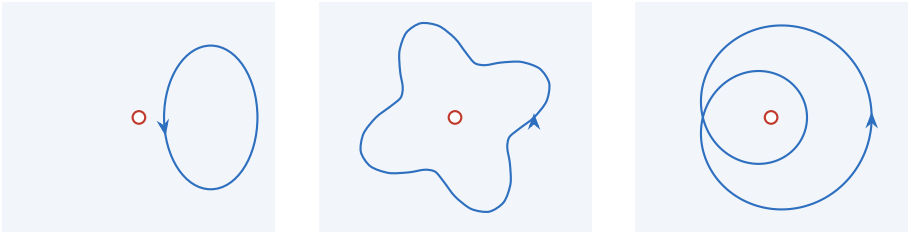
(1) Formula (1.33); factor $z = e^{i\phi}(1+ae^{-i\phi})$.
(2) $\frac{1}{1+w} = \sum_k (-w)^k$ with $|w| = |a| < 1$; the series converges uniformly, so we may integrate term by term.
(3) $\int_0^{2\pi} e^{-ik\phi}\mathrm{d}\phi = 2\pi\delta_{k0}$: only the constant Fourier mode survives.

Case $|a| > 1$. Now expand in $e^{i\phi}/a$ instead:

$$W \overset{(1)}{=} \frac{1}{2\pi i}\int_0^{2\pi} \frac{ie^{i\phi}}{a(1+e^{i\phi}/a)}\,\mathrm{d}\phi \overset{(2)}{=} \frac{1}{2\pi a}\int_0^{2\pi}\sum_{k\geq 0}(-a)^{-k} e^{i(k+1)\phi}\,\mathrm{d}\phi \overset{(3)}{=} 0.$$

(1) Factor $z = a(1+e^{i\phi}/a)$.
(2) Geometric series with $|e^{i\phi}/a| < 1$.
(3) Every term has frequency $k+1 \geq 1$, so it integrates to zero.

So the circle winds once around the origin when it encloses it and zero times when it does not, as the picture suggests (Figure 1.17). The jump from 1 to 0 happens only at $|a| = 1$, where the loop passes through the origin and leaves the space $\mathbb{C}\setminus\{0\}$.



$W = 0$ $W = 1$ $W = 2$

Figure 1.17: Loops in the punctured plane $\mathbb{C}\setminus\{0\}$ and their winding numbers. Wiggles do not matter (middle); only the net number of times the loop goes around the hole counts. The right-hand loop, $z(t) = 0.95e^{2it} + 0.45e^{it}$, goes around the origin twice, crossing itself once. In a superfluid these are the possible circulations $\oint \nabla\theta\cdot\mathrm{d}\vec{l} = 2\pi W$ around a vortex at the hole.

The winding number cannot change under homotopy

This is the heart of the matter. Here is the physicist's proof, for a smooth based homotopy $F(s,t)$ of loops in $\mathbb{C}\setminus\{0\}$. Write $W(s) = \frac{1}{2\pi i}\int_0^1 \frac{\partial_t F}{F}\,\mathrm{d}t$ for the winding number of the loop $F(s,\cdot)$. Then

$$\frac{\mathrm{d}W}{\mathrm{d}s} \overset{(1)}{=} \frac{1}{2\pi i}\int_0^1 \partial_s\Big(\frac{\partial_t F}{F}\Big)\mathrm{d}t \overset{(2)}{=} \frac{1}{2\pi i}\int_0^1 \frac{F\,\partial_s\partial_t F - \partial_s F\,\partial_t F}{F^2}\,\mathrm{d}t$$
$$\overset{(3)}{=} \frac{1}{2\pi i}\int_0^1 \partial_t\Big(\frac{\partial_s F}{F}\Big)\mathrm{d}t \overset{(4)}{=} \frac{1}{2\pi i}\Big[\frac{\partial_s F}{F}\Big]_{t=0}^{t=1} \overset{(5)}{=} 0.$$

(1) Differentiate under the integral sign (everything is smooth and $F \neq 0$).
(2) Quotient rule.
(3) Quotient rule in the other order: $\partial_t(\partial_s F/F) = (F\partial_t\partial_s F - \partial_t F\partial_s F)/F^2$, and mixed partials commute.

(4) Fundamental theorem of calculus in t.
(5) The base point is pinned, $F(s,0) = F(s,1) = x_0$ for all s, so $\partial_s F = 0$ at both ends. (For a free homotopy of closed loops, $F(s,0) = F(s,1)$ and the two boundary terms are equal and cancel.)

So $W(s)$ is constant. Even without this computation, $W(s)$ is a continuous function of s with values in the integers, and a continuous integer-valued function on an interval is constant. (For merely continuous homotopies one lifts the whole homotopy to $\mathbb{R}$, square by small square, by the same patching argument as for paths.)

Key idea: Nash's (1.32), first line: $\pi_1(S^1) \cong \mathbb{Z}$

The winding number is a group isomorphism

$$W : \pi_1(S^1) \xrightarrow{\cong} \mathbb{Z}, \qquad [\gamma] \mapsto W[\gamma].$$

Every loop in the circle is homotopic to exactly one standard loop α_n, and $n = W[\gamma]$.

We check the four things this says.

(a) *Well defined:* homotopic loops have equal W, by the computation above.

(b) *Homomorphism:* the lift of $\alpha \cdot \beta$ is the lift of α (at double speed) followed by the lift of β shifted up to start where the first one ended, so

$$W[\alpha \cdot \beta] \overset{(1)}{=} \frac{1}{2\pi i}\int_0^{1/2} \frac{2\dot\alpha(2t)}{\alpha(2t)}\mathrm{d}t + \frac{1}{2\pi i}\int_{1/2}^{1} \frac{2\dot\beta(2t-1)}{\beta(2t-1)}\mathrm{d}t \overset{(2)}{=} W[\alpha] + W[\beta].$$

(1) Formula (1.33) on the two halves of (1.31), with the chain rule giving the factor 2.
(2) Substitute $u = 2t$ in the first integral and $u = 2t-1$ in the second.

(c) *Onto:* $W[\alpha_n] = n$ (Example 1.23).

(d) *One-to-one:* suppose $W[\alpha] = W[\beta] = n$, with lifts $\tilde\theta_\alpha$, $\tilde\theta_\beta$ from 0 to $2\pi n$. Average the *lifts*, not the loops:

$$F(s,t) = \exp\big(i\,[(1-s)\tilde\theta_\alpha(t) + s\,\tilde\theta_\beta(t)]\big).$$

This is on the circle for all s,t; at $s = 0,1$ it is α,β; and at $t = 0$ and $t = 1$ the exponent is 0 and $2\pi i n$, so $F(s,0) = F(s,1) = 1$. Hence $\alpha \simeq \beta$ as based loops.

Step (d) is the whole trick: on the helix $\mathbb{R}$ we may interpolate linearly (it is contractible), and projecting down gives a homotopy on the circle.

Worked example 1.25: The fundamental theorem of algebra

Winding numbers prove that every polynomial $P(z) = z^n + c_{n-1}z^{n-1} + \cdots + c_0$ with $n \geq 1$ has a complex root. Suppose not: $P(z) \neq 0$ for all z. For $r \geq 0$ consider the loop $\gamma_r(t) = P(re^{2\pi i t})$ in $\mathbb{C} \setminus \{0\}$.

- As r goes from 0 to R, the loops γ_r form a homotopy in $\mathbb{C} \setminus \{0\}$ from the constant loop $\gamma_0 = P(0)$ (winding 0) to γ_R. (It is a free homotopy of loops, so the free version of the invariance argument applies.) So $W[\gamma_R] = 0$.
- For large R the leading term dominates. Put $Q(z) = c_{n-1}z^{n-1} + \cdots + c_0$ and deform

z^n into $P(z)$ by $H_s(z) = z^n + sQ(z)$, $s \in [0,1]$, on the circle $|z| = R$. It never vanishes:

$$|H_s(z)| \overset{(1)}{\geq} |z|^n - s|Q(z)| \overset{(2)}{\geq} R^n - \left(|c_{n-1}|R^{n-1} + \cdots + |c_0|\right) \overset{(3)}{>} 0.$$

(1) Reverse triangle inequality $|a+b| \geq |a| - |b|$.
(2) $s \leq 1$ and the triangle inequality on Q.
(3) True for $R > 1 + \sum_k |c_k|$, since then the bracket is at most $R^{n-1}\sum_k |c_k| < R^n$.

So γ_R is homotopic in $\mathbb{C} \setminus \{0\}$ to $t \mapsto R^n e^{2\pi i n t}$, which has winding number n (Example 1.23).

Thus $0 = W[\gamma_R] = n \geq 1$, a contradiction. So P has a root.

Physical picture: Flux quantisation and the Aharonov–Bohm effect

In a superconducting ring the Cooper-pair condensate is $\psi = |\psi|e^{i\theta}$, with charge $q = 2e$. Deep inside the ring the current vanishes, which forces $\hbar\nabla\theta = q\vec{A}$. Integrating around a loop C inside the ring,

$$\Phi = \oint_C \vec{A}\cdot \mathrm{d}\vec{l} \overset{(1)}{=} \frac{\hbar}{q}\oint_C \nabla\theta \cdot \mathrm{d}\vec{l} \overset{(2)}{=} \frac{\hbar}{q} 2\pi W \overset{(3)}{=} W\,\frac{h}{2e}.$$

(1) Zero current in the bulk: $\vec{J} \propto \hbar\nabla\theta - q\vec{A} = 0$; the first equality is Stokes' theorem for the enclosed flux.
(2) ψ is single valued, so θ changes by $2\pi W$ around C, where W is the winding number of the loop $\psi/|\psi| : C \to S^1$.
(3) $2\pi\hbar = h$ and $q = 2e$.

Flux comes in integer multiples of $h/2e$ because $\pi_1(S^1) = \mathbb{Z}$. Similarly, an electron moving in $\mathbb{R}^3$ minus an infinitely long solenoid lives in a space homotopy equivalent to S^1 (Example 1.13). Its paths fall into classes labelled by $\pi_1 = \mathbb{Z}$, and the Aharonov–Bohm phase $e^{iqW\Phi/\hbar}$ weights them in the path integral.

Now we can finish the argument of Example 1.14: the identity loop of S^1 has $W = 1$ and a constant loop has $W = 0$, so they are not homotopic, and S^1 is not a retract of the disc. Brouwer's fixed-point theorem is proved.

1.3.3 Computing more fundamental groups

With $\pi_1(S^1) = \mathbb{Z}$ in hand, homotopy invariance and a few simple rules give the fundamental groups of most spaces a physicist meets.

Homotopy invariance

Since π_1 depends only on the homotopy type, Examples 1.12 and 1.13 give at once

$$\pi_1(\mathbb{R}^2 \setminus \{0\}) = \pi_1(\text{Möbius band}) = \pi_1(S^1 \times D^2) = \pi_1(\mathbb{R}^3 \setminus \text{line}) = \mathbb{Z},$$

and $\pi_1(\mathbb{R}^n) = 0$. In each case a generator is "go once around the hole". The winding number around the z-axis of a loop in $\mathbb{R}^3 \setminus \{z\text{-axis}\}$ is computed by the same integral $\frac{1}{2\pi}\oint \mathrm{d}\phi$, with ϕ the azimuthal angle.

Products

A loop in $X \times Y$ is a pair of loops, one in each factor, run simultaneously, and a based homotopy is a pair of based homotopies. So

$$\pi_1(X \times Y) \cong \pi_1(X) \times \pi_1(Y), \qquad [(\alpha, \beta)] \mapsto ([\alpha], [\beta]). \tag{1.34}$$

The product of loops is computed componentwise:

$$(\alpha, \beta) \cdot (\alpha', \beta') \overset{(1)}{=} (\alpha \cdot \alpha',\ \beta \cdot \beta'),$$

(1) Both sides run (α, β) at double speed on $[0, \frac{1}{2}]$ and (α', β') on $[\frac{1}{2}, 1]$, component by component.

so (1.34) is a group isomorphism. (The same argument works for all the higher π_n.)

Worked example 1.26: The torus: $\pi_1(T^2) = \mathbb{Z} \times \mathbb{Z}$

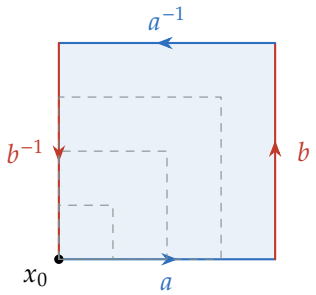
Let $T^2 = S^1 \times S^1$ with angles (θ, ϕ). By (1.34) and $\pi_1(S^1) = \mathbb{Z}$, a loop on the torus is classified by two winding numbers,

$$(m, n) = \left(\frac{1}{2\pi} \oint \mathrm{d}\theta,\ \frac{1}{2\pi} \oint \mathrm{d}\phi \right) \in \mathbb{Z} \times \mathbb{Z}.$$

The loop $t \mapsto (e^{2\pi i p t}, e^{2\pi i q t})$ has class (p, q); for coprime p, q it is a (p, q) torus knot when the torus is drawn in $\mathbb{R}^3$ (the trefoil of Figure 1.6 is $(2, 3)$). The group is abelian. In terms of the generators $a = (1, 0)$ and $b = (0, 1)$ (the edges of the square in Figure 1.4),

$$aba^{-1}b^{-1} \simeq e.$$

Geometrically: the loop $aba^{-1}b^{-1}$ runs around the boundary of the square, and the square itself provides the contraction: shrink the boundary of the square through the interior to a corner (Figure 1.18).



$aba^{-1}b^{-1}$ bounds the square

Figure 1.18: On the torus (a square with opposite sides glued), the loop $aba^{-1}b^{-1}$ is the boundary of the square traversed once. The square itself is a disc filling it in, so the loop shrinks (grey dashed squares) to the corner x_0. Hence $ab = ba$ in $\pi_1(T^2)$.

Spheres of dimension at least two are simply connected

Worked example 1.27: $\pi_1(S^n) = 0$ for $n \geq 2$

Take a loop α in S^n. First deform it slightly to a smooth loop (any continuous loop is homotopic to a smooth one by smoothing, e.g. by convolution followed by projection back onto the sphere). A smooth loop is 1-dimensional, and by the general-position count (1.4) it meets a chosen point $P \neq x_0$ of S^n in a set of dimension $1 + 0 - n < 0$, i.e. generically not at all. (If it happens to pass through P, choose another point: the image of a smooth curve has zero area in S^n when $n \geq 2$.) So α lies in $S^n \setminus \{P\}$, which is homeomorphic to $\mathbb{R}^n$ by stereographic projection from P, and there α contracts by Example 1.22. Hence every loop contracts: $\pi_1(S^n) = 0$ for $n \geq 2$.
The smoothing step is not pedantry: there are continuous surjections $[0,1] \to S^2$ (space-filling curves), which do not miss any point. They are still contractible, but only because they can first be smoothed.

Wedges: the figure eight and the plane with two holes

For wedges, the rule is different: loops in the two pieces do not commute. The general statement is the Seifert–van Kampen theorem, which Nash assumes silently (it is behind his remark on Riemann surfaces).

Key idea: Seifert–van Kampen, in the form we need

- $\pi_1(X \vee Y)$ is the **free product** $\pi_1(X) * \pi_1(Y)$: its elements are words $g_1 h_1 g_2 h_2 \cdots$ alternating between elements of $\pi_1(X)$ and $\pi_1(Y)$, multiplied by juxtaposition, with no relations between the two factors.
- Gluing a disc onto a space along a loop λ adds the relation $\lambda = 1$.

A consequence: any assignment of the generators to elements of a group G that respects the relations defines a homomorphism $\pi_1 \to G$.

So $\pi_1(S^1 \vee S^1) = \mathbb{Z} * \mathbb{Z} = F_2$, the free group on two generators a, b. The plane with two punctures deformation retracts onto a figure eight (Figure 1.14), so it has the same π_1.

Worked example 1.28: The figure eight has a non-abelian π_1

We show $ab \neq ba$ in $F_2 = \pi_1(S^1 \vee S^1)$. It is enough to find a homomorphism $\rho : F_2 \to G$ into some group with $\rho(a)\rho(b) \neq \rho(b)\rho(a)$. Take $G = S_3$, the permutations of $\{1,2,3\}$, and $\rho(a) = \sigma = (12)$, $\rho(b) = \tau = (23)$; since F_2 has no relations this is allowed. (Geometrically, ρ is the monodromy of a three-sheeted covering of the figure eight: going around a swaps sheets $1,2$, going around b swaps sheets $2,3$.) Compose permutations right to left and follow the element 1 through the commutator:

$$\sigma\tau\sigma^{-1}\tau^{-1}(1) \overset{(1)}{=} \sigma\tau\sigma\tau(1) \overset{(2)}{=} \sigma\tau\sigma(1) \overset{(3)}{=} \sigma\tau(2) \overset{(4)}{=} \sigma(3) = 3 \neq 1.$$

(1) Transpositions are their own inverses.
(2) $\tau = (23)$ fixes 1.
(3) $\sigma = (12)$ sends 1 to 2.
(4) τ sends 2 to 3; then σ fixes 3.

So $\rho(aba^{-1}b^{-1}) \neq 1$, hence $aba^{-1}b^{-1} \neq e$ and π_1 is non-abelian. Physically: an electron

encircling first flux tube a then flux tube b, versus the reverse order, follows paths in different homotopy classes. For abelian (ordinary) fluxes the Aharonov–Bohm phases still commute; for non-abelian fluxes (holonomies in a non-abelian gauge group) the order is observable.

Worked example 1.29: Riemann surfaces of genus $g \geq 2$: Nash's non-abelian π_1

A genus-g surface Σ_g is a $4g$-gon with edges glued as in the Try-it of §1.1.2 (Figure 1.19). After gluing, the boundary of the polygon becomes a wedge of $2g$ circles $a_1, b_1, \ldots, a_g, b_g$, and the interior is a disc glued along the loop $a_1 b_1 a_1^{-1} b_1^{-1} \cdots a_g b_g a_g^{-1} b_g^{-1}$. By van Kampen,

$$\pi_1(\Sigma_g) = \langle a_1, b_1, \ldots, a_g, b_g \mid [a_1, b_1][a_2, b_2] \cdots [a_g, b_g] = 1 \rangle, \qquad [a, b] := aba^{-1}b^{-1}. \tag{1.35}$$

For $g = 1$ this is $\langle a, b \mid aba^{-1}b^{-1} = 1 \rangle = \mathbb{Z} \times \mathbb{Z}$, the torus again. For $g = 2$, Nash says π_1 is non-commutative. To see it, map to S_3 again: $a_1 \mapsto \sigma = (12)$, $b_1 \mapsto \tau = (23)$, $a_2 \mapsto \tau$, $b_2 \mapsto \sigma$. The relation is respected: with $c := \sigma\tau$ (a 3-cycle, so $c^3 = 1$) and $\tau\sigma = c^{-1}$,

$$[\sigma, \tau][\tau, \sigma] \overset{(1)}{=} (\sigma\tau\sigma\tau)(\tau\sigma\tau\sigma) \overset{(2)}{=} c^2 c^{-2} = 1.$$

(1) Transpositions are self-inverse, so $[\sigma, \tau] = \sigma\tau\sigma\tau$ and $[\tau, \sigma] = \tau\sigma\tau\sigma$.

(2) $\sigma\tau\sigma\tau = (\sigma\tau)^2 = c^2$ and $\tau\sigma\tau\sigma = (\tau\sigma)^2 = c^{-2}$, where $\tau\sigma = (\sigma\tau)^{-1}$ because both are self-inverse.

So this defines a homomorphism $\pi_1(\Sigma_2) \to S_3$, and a_1, b_1 map to non-commuting permutations (Example 1.28). Hence $a_1 b_1 \neq b_1 a_1$ in $\pi_1(\Sigma_2)$. The same trick (with the extra generators sent to 1) works for all $g \geq 2$.

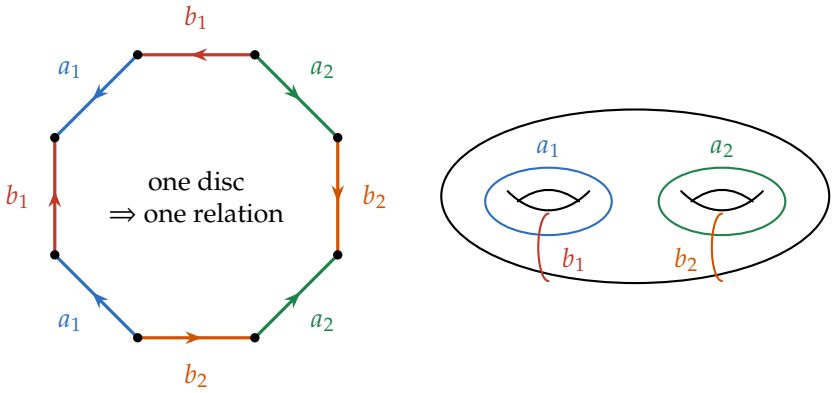


Figure 1.19: A genus-2 surface as an octagon with edges glued in the pattern $a_1 b_1 a_1^{-1} b_1^{-1} a_2 b_2 a_2^{-1} b_2^{-1}$ (reading clockwise from the bottom left). All eight corners become one point. The glued edges are four loops on the surface (right). The interior of the octagon is a single disc glued along the boundary word, which gives the single relation $[a_1, b_1][a_2, b_2] = 1$ in (1.35).

Worked example 1.30: The projective plane: $\pi_1(\mathbb{RP}^2) = \mathbb{Z}_2$

$\mathbb{RP}^2$ is the sphere S^2 with antipodal points identified, $\vec{n} \sim -\vec{n}$ (the space of lines through the origin in $\mathbb{R}^3$, or of unoriented directions: the order-parameter space of a nematic liquid crystal). Let γ be a path on S^2 from the north pole N to the south pole $S = -N$ along a meridian. Since $N \sim S$ in $\mathbb{RP}^2$, γ becomes a closed loop there. It is not contractible: a homotopy of loops in $\mathbb{RP}^2$ lifts to a homotopy of paths in S^2 starting at N, and the lifted

endpoint can only be N or $-N$; by continuity it stays $-N$, whereas the constant loop lifts to a path ending at N. But $\gamma\cdot\gamma$ lifts to a path from N to S and back to N along the antipodal meridian: a closed loop on S^2, which contracts because $\pi_1(S^2)=0$. So $\pi_1(\mathbb{RP}^2)=\{e,[\gamma]\}\cong\mathbb{Z}_2$. The same argument gives $\pi_1(\mathbb{RP}^n)=\mathbb{Z}_2$ for all $n\geq 2$. In a nematic this is why there are line defects of "half-integer" strength: the director rotates by π, not 2π, around the defect.

1.3.4 Topological groups, $\mathrm{SO}(3)$ and the belt trick

Nash promised at the end of §2 that "the fundamental group of a topological group G (e.g. G is some Lie group) is always commutative". In §3 he proves it by writing

$$\pi_1(G)=[S^1,G]_0=[SS^0,G]_0, \tag{1.36}$$

where $S^0=\{x_1 : x_1^2=1\}=\{\pm1\}$ is the zero-sphere (Nash's footnote) and $S^1=SS^0$ by (1.23) with $n=0$. So $\pi_1(G)$ has the form $[X,Y]_0$ with $Y=G$ an H-space and $X=SS^0$ a coH-space, and the Eckmann–Hilton argument of §1.2.8 applies.

It is instructive to see the argument with actual loops. Let α,β be loops in G at the identity e, write e also for the constant loop, and let $\bullet$ be the pointwise product. Then

$$\alpha\cdot\beta \overset{(1)}{=} (\alpha\cdot e)\bullet(e\cdot\beta) \overset{(2)}{\simeq} (e\cdot\alpha)\bullet(\beta\cdot e) \overset{(3)}{=} \beta\cdot\alpha.$$

(1) For $t\leq\frac12$ the right side is $\alpha(2t)\,e=\alpha(2t)$; for $t\geq\frac12$ it is $e\,\beta(2t-1)=\beta(2t-1)$. This is literally $\alpha\cdot\beta$.

(2) $\alpha\cdot e\simeq e\cdot\alpha$ and $e\cdot\beta\simeq\beta\cdot e$ (both sides are reparametrisations of the same loop), and pointwise products of homotopies are homotopies.

(3) For $t\leq\frac12$ the middle expression is $e\,\beta(2t)=\beta(2t)$, for $t\geq\frac12$ it is $\alpha(2t-1)\,e$; this is $\beta\cdot\alpha$.

The picture: slide the loop α "through" β, using the group multiplication to let them pass each other. So $\pi_1(G)$ is abelian. For instance, the figure eight is not a topological group, and no group structure can be put on it.

$\mathrm{SU}(2)$ is the three-sphere

Every 2×2 matrix of the form

$$U=a_0\,\mathbf{1}+i\,\vec a\cdot\vec\sigma=\begin{pmatrix}a_0+ia_3 & ia_1+a_2\\ ia_1-a_2 & a_0-ia_3\end{pmatrix},\qquad (a_0,\vec a)\in\mathbb{R}^4, \tag{1.37}$$

with the Pauli matrices $\sigma_1=\left(\begin{smallmatrix}0&1\\1&0\end{smallmatrix}\right)$, $\sigma_2=\left(\begin{smallmatrix}0&-i\\i&0\end{smallmatrix}\right)$, $\sigma_3=\left(\begin{smallmatrix}1&0\\0&-1\end{smallmatrix}\right)$, satisfies

$$U^\dagger U \overset{(1)}{=} (a_0-i\,\vec a\cdot\vec\sigma)(a_0+i\,\vec a\cdot\vec\sigma) \overset{(2)}{=} a_0^2+(\vec a\cdot\vec\sigma)^2 \overset{(3)}{=} (a_0^2+|\vec a|^2)\,\mathbf{1},$$

$$\det U \overset{(4)}{=} (a_0+ia_3)(a_0-ia_3)-(ia_1+a_2)(ia_1-a_2)=a_0^2+a_3^2+a_1^2+a_2^2.$$

(1) The Pauli matrices are Hermitian, so $U^\dagger=a_0-i\,\vec a\cdot\vec\sigma$.

(2) The cross terms $\pm ia_0\,\vec a\cdot\vec\sigma$ cancel.

(3) $\sigma_j\sigma_k+\sigma_k\sigma_j=2\delta_{jk}$, so $(\vec a\cdot\vec\sigma)^2=|\vec a|^2\,\mathbf{1}$.

(4) Determinant of the explicit matrix; $(ia_1)^2=-a_1^2$.

So $U\in\mathrm{SU}(2)$ iff $a_0^2+|\vec a|^2=1$, and conversely every element of $\mathrm{SU}(2)$ has this form. Hence

$$\mathrm{SU}(2)\cong S^3\subset\mathbb{R}^4,$$

and in particular $\pi_1(\mathrm{SU}(2))=0$ by Example 1.27.

The double cover $\mathrm{SU}(2) \to \mathrm{SO}(3)$

SU(2) acts on traceless Hermitian matrices $\vec{x}\cdot\vec{\sigma}$ by conjugation, $\vec{x}\cdot\vec{\sigma} \mapsto U(\vec{x}\cdot\vec{\sigma})U^\dagger$. The result is again traceless and Hermitian, so equals $\vec{x}'\cdot\vec{\sigma}$ for a unique $\vec{x}'$, and det is preserved, so $|\vec{x}'|^2 = -\det(\vec{x}'\cdot\vec{\sigma}) = |\vec{x}|^2$. Thus $\vec{x}' = R(U)\vec{x}$ with $R(U) \in \mathrm{SO}(3)$; explicitly

$$U(\vec{x}\cdot\vec{\sigma})U^\dagger = \big(R(U)\vec{x}\big)\cdot\vec{\sigma}, \qquad R(U)_{jk} = \tfrac{1}{2}\operatorname{tr}\big(\sigma_j U \sigma_k U^\dagger\big). \tag{1.38}$$

Since U and $-U$ give the same conjugation, $R(-U) = R(U)$: the map $U \mapsto R(U)$ is two-to-one. (One can check it is onto and that $\pm U$ are the only elements with the same image.)

Worked example 1.31: Rotating by 2π gives -1 in SU(2)

Let $U_\theta = e^{-i\theta\sigma_3/2} = \operatorname{diag}(e^{-i\theta/2}, e^{i\theta/2})$. Its action on σ_1:

$$U_\theta\sigma_1 U_\theta^\dagger \overset{(1)}{=} \begin{pmatrix} e^{-i\theta/2} & 0 \\ 0 & e^{i\theta/2}\end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & 0\end{pmatrix}\begin{pmatrix} e^{i\theta/2} & 0 \\ 0 & e^{-i\theta/2}\end{pmatrix} \overset{(2)}{=} \begin{pmatrix} 0 & e^{-i\theta} \\ e^{i\theta} & 0\end{pmatrix}$$
$$\overset{(3)}{=} \begin{pmatrix} 0 & \cos\theta - i\sin\theta \\ \cos\theta + i\sin\theta & 0\end{pmatrix} \overset{(4)}{=} \cos\theta\,\sigma_1 + \sin\theta\,\sigma_2.$$

(1) $U_\theta^\dagger = \operatorname{diag}(e^{i\theta/2}, e^{-i\theta/2})$.
(2) Multiply out: the $(1,2)$ entry is $e^{-i\theta/2}\cdot 1\cdot e^{-i\theta/2}$ and the $(2,1)$ entry is $e^{i\theta/2}\cdot 1\cdot e^{i\theta/2}$.
(3) Euler's formula.
(4) Read off the coefficients of $\sigma_1 = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0\end{smallmatrix}\right)$ and $\sigma_2 = \left(\begin{smallmatrix} 0 & -i \\ i & 0\end{smallmatrix}\right)$.

The same computation gives $U_\theta\sigma_2 U_\theta^\dagger = -\sin\theta\,\sigma_1 + \cos\theta\,\sigma_2$, and σ_3 commutes with U_θ. So

$$R(U_\theta) = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1\end{pmatrix},$$

the rotation by angle θ about the z-axis. As θ goes from 0 to 2π, the rotation goes once around a closed loop in SO(3), from $\mathbb{1}$ back to $\mathbf{1}$. But the matrix U_θ goes from $U_0 = \mathbf{1}$ to

$$U_{2\pi} = \operatorname{diag}(e^{-i\pi}, e^{i\pi}) = -\mathbf{1}.$$

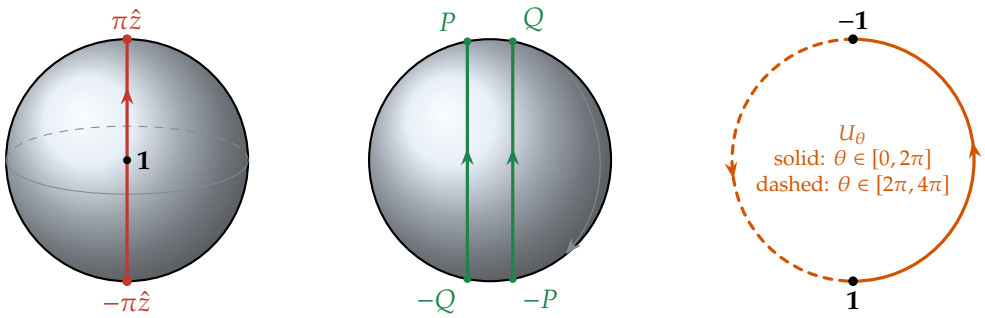
The lift of the closed loop is *not* closed. Only at $\theta = 4\pi$ does U_θ return to $\mathbf{1}$.

Key idea: $\pi_1(\mathrm{SO}(3)) = \mathbb{Z}_2$

Let γ be the loop $\theta \mapsto R(U_\theta)$, $\theta \in [0, 2\pi]$.

- γ is *not* contractible. Any loop in SO(3) at $\mathbf{1}$ lifts uniquely to a path in SU(2) starting at $\mathbf{1}$ (the same patching argument as for $\mathbb{R} \to S^1$, since the map is locally a homeomorphism). Its endpoint is $+\mathbf{1}$ or $-\mathbf{1}$. Under a based homotopy of the loop, the endpoint of the lift moves continuously in the two-point set $\{\pm\mathbf{1}\}$, so it cannot change. The constant loop lifts to a path ending at $+\mathbf{1}$; γ lifts to one ending at $-\mathbf{1}$.
- $\gamma\cdot\gamma$ (rotation by 4π) *is* contractible. Its lift U_θ, $\theta \in [0, 4\pi]$, is a closed loop in $\mathrm{SU}(2) \cong S^3$, which contracts since $\pi_1(S^3) = 0$. Projecting the contraction by R contracts $\gamma\cdot\gamma$.

Every loop lifts to a path ending at $\pm\mathbf{1}$, and loops with the same endpoint are homotopic (project a homotopy of lifts in the simply connected S^3). So $\pi_1(\mathrm{SO}(3)) = \{[e], [\gamma]\} \cong \mathbb{Z}_2$.



SO(3): ball of radius π, antipodes on the surface glued

4π: move the two surface crossings around and cancel

lift to SU(2) $\cong S^3$

Figure 1.20: Three views of $\pi_1(\mathrm{SO}(3)) = \mathbb{Z}_2$. Left: a rotation is a vector $\theta\hat{n}$ with $|\theta| \le \pi$, so SO(3) is a solid ball whose antipodal surface points $\pm\pi\hat{n}$ are the same rotation. The 2π loop runs up the z-axis, leaves at $\pi\hat{z}$ and re-enters at $-\pi\hat{z}$; it cannot be shrunk. Middle: the 4π loop crosses the surface twice; the pair of crossings can be moved round the sphere until they meet and cancel. Right: in SU(2) the 2π rotation lifts to a path from $\mathbf{1}$ to $-\mathbf{1}$ (solid), and only the 4π rotation closes up (solid plus dashed).

Physical picture: The belt trick, and why electrons have spin $\frac{1}{2}$

Hold one end of a belt fixed and rotate the buckle by 2π about the belt's axis. The belt is twisted, and no amount of moving the buckle around (without rotating it) will untwist it. Rotate by another 2π: now the belt can be untwisted by looping it around the buckle. This is Dirac's demonstration of $\pi_1(\mathrm{SO}(3)) = \mathbb{Z}_2$: the family of orientations along the belt is a path in SO(3), and the 2π loop is non-contractible while the 4π loop is contractible.

Quantum mechanics represents rotations on states up to phase, and it is really the simply connected group SU(2) = Spin(3) that acts. A spin-$\frac{1}{2}$ state picks up $U_{2\pi} = -\mathbf{1}$ under a 2π rotation. The sign was measured directly in neutron interferometry (1975): a neutron beam passed through a magnetic field that precesses its spin by 2π interferes destructively with the unrotated beam. The generalisation, $\pi_1(\mathrm{SO}(n)) = \mathbb{Z}_2$ for $n \ge 3$ with universal cover $\mathrm{Spin}(n)$, is the subject of Chapter T4.

Worked example 1.32: $\pi_1(\mathrm{U}(n)) = \mathbb{Z}$ via the determinant

$\mathrm{SU}(n)$ is simply connected for all n (for $n = 2$ we just showed it; in general it follows from the exact sequences of §1.3.11). The map

$$\mathrm{SU}(n) \times \mathrm{U}(1) \to \mathrm{U}(n), \qquad (V, z) \mapsto V\,\mathrm{diag}(z, 1, \dots, 1)$$

is a homeomorphism (inverse: $z = \det M$, $V = M\,\mathrm{diag}(z^{-1}, 1, \dots, 1)$), though not a group isomorphism. So $\pi_1(\mathrm{U}(n)) \cong \pi_1(\mathrm{SU}(n)) \times \pi_1(\mathrm{U}(1)) = \mathbb{Z}$, and the isomorphism is the winding number of the determinant. For example, in U(2):

$$W\left[\det \mathrm{diag}(e^{2\pi i t}, 1)\right] = W[e^{2\pi i t}] \overset{(1)}{=} 1, \qquad W\left[\det(e^{2\pi i t}\mathbf{1}_2)\right] \overset{(2)}{=} W[e^{4\pi i t}] \overset{(1)}{=} 2.$$

(1) Example 1.23.
(2) $\det(\lambda \mathbf{1}_2) = \lambda^2$.

So the loop of overall phases $e^{2\pi i t}\mathbf{1}_2$ is twice the generator. Summary for the groups we use most:

G	U(1)	U(n)	SU(n)	SO(2)	SO(n), $n \ge 3$	Sp(n), Spin(n) ($n \ge 3$)
$\pi_1(G)$	$\mathbb{Z}$	$\mathbb{Z}$	0	$\mathbb{Z}$	$\mathbb{Z}_2$	0

All are abelian, as Nash's argument requires.

1.3.5 Free versus based homotopy classes

In §1.2.4 we introduced base points for convenience. Now we pay the bill: how much does the answer depend on them? Nash summarises the answer in one line,

$$[X, Y]_0/\pi_1(Y) \cong [X, Y], \tag{1.39}$$

which says: $\pi_1(Y)$ acts on the based classes, and the free classes are the orbits of this action. We unpack it.

Changing the base point

Temporarily write $[X, x_0; Y, y_0]_0$ to show the base points. Keep x_0 fixed and move y_0 to y_1. Choose a path γ from y_0 to y_1 (here we need Y path-connected, or at least y_0, y_1 in the same path component). For loops ($X = S^1$) the translation is explicit:

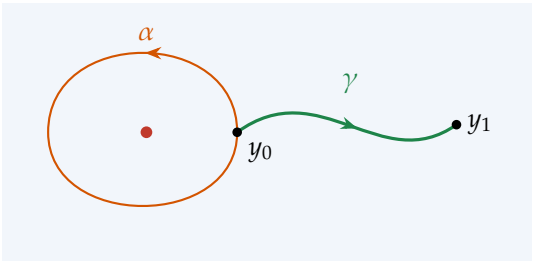
$$i_\gamma : \pi_1(Y, y_0) \to \pi_1(Y, y_1), \qquad i_\gamma[\alpha] := [\bar\gamma \cdot \alpha \cdot \gamma], \tag{1.40}$$

"walk back from y_1 to y_0 along γ, do α, return along γ" (Figure 1.21). (The products of paths are defined as for loops, whenever the end of one is the start of the next.) This is a group isomorphism: its inverse is $i_{\bar\gamma}$, and

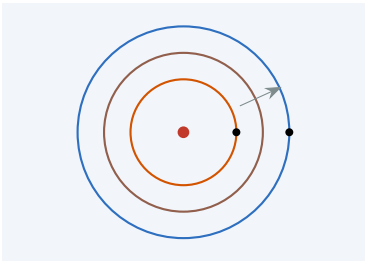
$$i_\gamma[\alpha]\, i_\gamma[\beta] \overset{(1)}{=} [\bar\gamma \cdot \alpha \cdot \gamma \cdot \bar\gamma \cdot \beta \cdot \gamma] \overset{(2)}{=} [\bar\gamma \cdot \alpha \cdot \beta \cdot \gamma] = i_\gamma[\alpha\beta].$$

(1) Definition (1.40) and associativity up to homotopy.
(2) $\gamma \cdot \bar\gamma$ is homotopic (with endpoints fixed) to the constant path at y_0, by the "reel it in" homotopy of §1.3.1.

For general X one uses the same idea: a based map α is dragged along γ by a homotopy that moves only a neighbourhood of the base point.



$i_\gamma[\alpha] = [\bar\gamma \cdot \alpha \cdot \gamma]$: a loop at y_1



free homotopy: base point may move

Figure 1.21: Left: translating a loop from base point y_0 to y_1 along a path γ. Right: a free homotopy of loops is allowed to move the base point; if the base point travels around a closed loop δ during the homotopy, the based class changes by conjugation, $[\alpha] \mapsto [\delta]^{-1}[\alpha][\delta]$.

The isomorphism depends only on the class of the path

If $\gamma \simeq \gamma'$ with endpoints fixed, then $\bar\gamma \cdot \alpha \cdot \gamma \simeq \bar\gamma' \cdot \alpha \cdot \gamma'$, so $i_\gamma = i_{\gamma'}$. If γ and γ' are *not* homotopic, they differ by a loop: $\mu := \bar\gamma \cdot \gamma'$ (back from y_1 to y_0 along γ, then out again along γ') is a loop at y_1, with inverse $\bar\mu = \bar\gamma' \cdot \gamma$. Then

$$i_{\gamma'}[\alpha] \overset{(1)}{=} [\bar\gamma' \cdot \alpha \cdot \gamma'] \overset{(2)}{=} [\bar\gamma' \cdot \gamma \cdot \bar\gamma \cdot \alpha \cdot \gamma \cdot \bar\gamma \cdot \gamma'] \overset{(3)}{=} [\bar\mu]\; i_\gamma[\alpha]\; [\mu] = [\mu]^{-1}\, i_\gamma[\alpha]\, [\mu].$$

(1) Definition (1.40).
(2) Insert $\gamma \cdot \bar{\gamma}$ (out to y_1 and back), which is homotopic to the constant path at y_0, on both sides of α.
(3) Group the six factors as $(\bar{\gamma}' \cdot \gamma) \cdot (\bar{\gamma} \cdot \alpha \cdot \gamma) \cdot (\bar{\gamma} \cdot \gamma')$: the outer pieces are $\bar{\mu}$ and μ, the middle is $i_\gamma[\alpha]$.

So two different paths give identifications that differ by conjugation by a loop, i.e. by the action of $\pi_1(Y)$ on itself. In words: the ambiguity in identifying based classes at different base points is exactly the action of $\pi_1(Y)$, by conjugation for π_1 itself. If $\pi_1(Y) = 0$ there is no ambiguity. This is Nash's argument.

Free loops are conjugacy classes

Now take $X = S^1$. A free homotopy of loops may drag the base point around. If the base point travels around a loop δ at y_0 during the homotopy, then the starting and final loops α, α' are related by $[\alpha'] = [\delta]^{-1}[\alpha][\delta]$. Conversely, α and $\bar{\delta} \cdot \alpha \cdot \delta$ are always freely homotopic (slide the base point back along δ). Hence (1.39) for $X = S^1$ reads

$$[S^1, Y] \cong \{\text{conjugacy classes of } \pi_1(Y)\}. \tag{1.41}$$

Worked example 1.33: Free and based loops in the figure eight

In $F_2 = \pi_1(S^1 \vee S^1)$, the loops a and bab^{-1} are conjugate, so they are freely homotopic. They are not based homotopic: under the homomorphism ρ of Example 1.28,

$$\rho(bab^{-1})(1) \overset{(1)}{=} \tau\sigma\tau(1) \overset{(2)}{=} \tau\sigma(1) \overset{(3)}{=} \tau(2) = 3, \qquad \text{while} \qquad \rho(a)(1) = \sigma(1) = 2.$$

(1) $\rho(b) = \tau = (23) = \tau^{-1}$, $\rho(a) = \sigma = (12)$, applied right to left.
(2) τ fixes 1.
(3) $\sigma(1) = 2$.

So $\rho(bab^{-1}) = (13) \neq (12) = \rho(a)$, and $a \neq bab^{-1}$ in π_1. In the abelian case (circle, torus, any Lie group) conjugation does nothing, and free and based classes of loops coincide: the winding number of a loop in S^1 needs no base point.

Worked example 1.34: When Y is simply connected, base points can be ignored

If $\pi_1(Y) = 0$, (1.39) says $[X, Y]_0 = [X, Y]$. Two important cases: $[S^2, S^2] = \pi_2(S^2)$, the degree of a map of spheres (§1.3.8), and $[S^3, \mathrm{SU}(2)] = \pi_3(\mathrm{SU}(2))$, the instanton number (§1.3.14). Physically this is a relief. A field configuration on space has no preferred base point, so it is the *free* class that is physical; when the target is simply connected the free class equals the based class, which is what we can compute. For $Y = \mathbb{RP}^2$ (nematics, $\pi_1 = \mathbb{Z}_2$), π_1 acts on $\pi_2(\mathbb{RP}^2) = \mathbb{Z}$ by $n \mapsto -n$, and free classes of point defects are labelled by $|n|$ only: a hedgehog and an anti-hedgehog in a nematic are the same defect.

Watch out: Path-connectedness

Nash notes that we "must assume that Y is path-connected, otherwise such γ's need not exist". For example $Y = \mathrm{O}(3)$ has two components (det $= \pm 1$). Based classes at the identity only see SO(3); maps into the other component are not related to them by any change of base point.

Physical picture: Non-abelian vortices

In a biaxial nematic liquid crystal the order-parameter space has $\pi_1 = Q_8 = \{\pm 1, \pm i, \pm j, \pm k\}$, the quaternion group. Line defects are labelled by *conjugacy classes*, $\{1\}, \{-1\}, \{\pm i\}, \{\pm j\}, \{\pm k\}$, by (1.41). Moreover, two defect lines labelled by non-commuting elements cannot cross each other without leaving a connecting line behind, labelled by their commutator. The algebra of π_1 becomes the mechanics of defect lines.

1.3.6 Higher homotopy groups

A vortex line in three-dimensional space is detected by a loop around it. A *point* defect in three-dimensional space, such as a magnetic monopole or the hedgehog in a ferromagnet, is not: any loop around a point in $\mathbb{R}^3$ can be slipped off the point. To detect a point defect we must surround it by a sphere S^2 and ask how the field on that sphere wraps around the space of field values. In four Euclidean dimensions the object that surrounds everything is the three-sphere at infinity. So we need to replace the circle in $\pi_1(X) = [S^1, X]_0$ by higher spheres. That is Nash's next step.

Working definition 1.21 (higher homotopy groups)

For $n \geq 1$ the **n-th homotopy group** of a pointed space $\{X, x_0\}$ is

$$\pi_n(X) \coloneqq [S^n, X]_0, \qquad n = 1, 2, \ldots \tag{1.42}$$

the set of based homotopy classes of based maps $S^n \to X$ [maps sending the base point of S^n to x_0, counted as the same when deformable into each other with the base point held fixed]. *Cube picture:* since $S^n \cong [0,1]^n/\partial[0,1]^n$ [the cube with its whole boundary squashed to a single point] (Example 1.18), an element is represented by a map $f(t_1, \ldots, t_n)$ of the unit cube I^n into X that sends the *whole boundary* of the cube to x_0. A homotopy is a family $F(s, t_1, \ldots, t_n)$ of such maps, with the boundary kept at x_0 for every s. The product is concatenation in the first coordinate:

$$(f+g)(t_1, t_2, \ldots, t_n) \coloneqq \begin{cases} f(2t_1, t_2, \ldots, t_n), & 0 \leq t_1 \leq \tfrac{1}{2}, \\ g(2t_1 - 1, t_2, \ldots, t_n), & \tfrac{1}{2} \leq t_1 \leq 1. \end{cases} \tag{1.43}$$

For $n = 1$ this is the loop product (1.31). We write the product as "+" because we shall shortly see that for $n \geq 2$ it is commutative. As for loops, $f + g$ is continuous on the wall $t_1 = \frac{1}{2}$: there the first branch gives $f(1, t_2, \ldots)$ and the second $g(0, t_2, \ldots)$, and both are x_0 because $t_1 = 1$ and $t_1 = 0$ are faces of the cube. Also $f + g$ sends every face of the cube to x_0, so it is again an element of π_n. The identity is the constant map $e \equiv x_0$, and the inverse of f is f run backwards in t_1, $(-f)(t_1, \ldots) = f(1 - t_1, \ldots)$. The associativity, identity and inverse homotopies are those of §1.3.1, acting on the variable t_1 and carried along trivially in $t_2, \ldots, t_n$.

$\pi_n(X)$ is π_{n-1} of the loop space

The right-hand panel of Figure 1.22 is Nash's (1.29). Read a map of the cube one row at a time: for each fixed $(t_2, \ldots, t_n)$ the function $t_1 \mapsto f(t_1, t_2, \ldots, t_n)$ is a loop at x_0, because the faces $t_1 = 0$ and $t_1 = 1$ go to x_0. So f is a map from the $(n-1)$-cube into the loop space ΩX. When $(t_2, \ldots, t_n)$ is on the boundary of the $(n-1)$-cube, the whole row lies in a face of I^n, so the loop is constant: the boundary of I^{n-1} goes to the base point of ΩX (the constant loop). In terms of

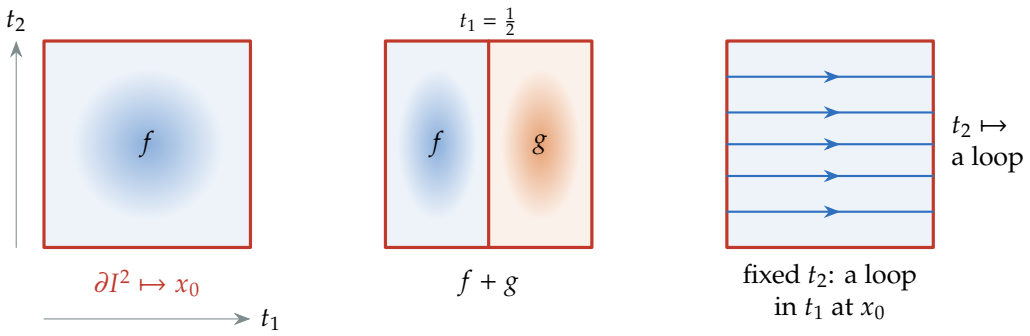


Figure 1.22: The cube picture of π_2. Left: an element is a map of the square whose whole boundary (red) goes to the base point x_0; only the interior (shaded) is doing anything. Middle: the product $f + g$ squeezes f into the left half and g into the right half. Right: read row by row, the same map is a one-parameter family of loops (horizontal lines), beginning and ending with the constant loop. This is (1.44): $\pi_2(X) = \pi_1(\Omega X)$.

the adjunction of §1.2.7:

$$\pi_n(X) \overset{(1)}{=} [S^n, X]_0 \overset{(2)}{=} [SS^{n-1}, X]_0 \overset{(3)}{=} [S^{n-1}, \Omega X]_0 \overset{(4)}{=} \pi_{n-1}(\Omega X). \tag{1.44}$$

(1) Definition (1.42).
(2) $S^n \cong S^1 \wedge S^{n-1} = SS^{n-1}$, (1.23).
(3) The adjunction $[SZ, X]_0 \cong [Z, \Omega X]_0$, (1.25), with $Z = S^{n-1}$; concretely the row-by-row reading above.
(4) Definition (1.42) again, now for the pointed space $\{\Omega X, \text{constant loop}\}$.

Iterating, $\pi_n(X) = \pi_{n-2}(\Omega^2 X) = \cdots = \pi_1(\Omega^{n-1} X) = \pi_0(\Omega^n X)$: an element of $\pi_n(X)$ is a path component of the n-fold loop space. The last form is how physicists often meet it. For instance, $\pi_3(G) = \pi_0(\Omega^3 G)$, and $\Omega^3 G$ is the space of gauge transformations on $\mathbb{R}^3$ that go to 1 at infinity. Its path components are the "large gauge transformations" of §1.3.14.

Why π_n is a group, and abelian for $n \geq 2$

By (1.44), $\pi_n(X) = [S^{n-1}, \Omega X]_0$. Since ΩX is an H-space (loops can be multiplied), $[Z, \Omega X]_0$ is a group for any Z (§1.2.8). So $\pi_n(X)$ is a group, and the group law coming from ΩX is exactly the concatenation (1.43) in t_1. For $n \geq 2$ we may peel off one more suspension, which is Nash's (1.30):

$$\pi_n(X) \overset{(1)}{=} [S^{n-1}, \Omega X]_0 \overset{(2)}{=} [SS^{n-2}, \Omega X]_0, \qquad n \geq 2. \tag{1.45}$$

(1) Equation (1.44).
(2) $S^{n-1} \cong SS^{n-2}$ by (1.23); for $n = 2$ this is $S^1 = SS^0$ with $S^0 = \{\pm 1\}$.

Now the target ΩX is an H-space *and* the source SS^{n-2} is a coH-space. By the Eckmann–Hilton argument (§1.2.8) the two group structures coincide and are commutative. So:

Key idea: Higher homotopy groups are abelian

For $n \geq 2$, $\pi_n(X)$ is an abelian group, and its product can be computed by concatenating in *any* of the n cube coordinates: all these products agree. The fundamental group $\pi_1(X)$, by contrast, can be non-abelian (the figure eight, genus $g \geq 2$ surfaces: Examples 1.28 and 1.29).

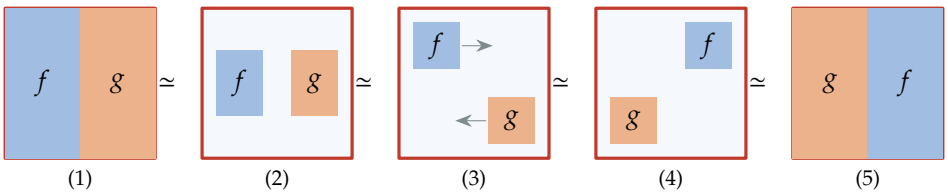


Figure 1.23: Why π_2 is abelian: $f + g \simeq g + f$. (1) The product $f + g$. (2) Shrink each map into a smaller box, filling the rest of the square with the base point (this is a reparametrisation, hence a homotopy). (3) Slide f up and g down, using the second direction. (4) Move them past each other horizontally. (5) Expand again: this is $g + f$. Every intermediate map still sends the boundary to x_0. In π_1 there is no second direction, the two blocks meet head-on, and the argument fails: like two particles on a line, which cannot exchange without colliding, while in the plane they can.

Figure 1.23 is the Eckmann–Hilton argument drawn in the cube. Each step is an explicit homotopy. For example, the step (1)→(2) is $F(s, t_1, t_2) = (f + g)\big(\phi_s(t_1), \psi_s(t_2)\big)$, where ϕ_s and ψ_s are piecewise-linear maps of $[0, 1]$ that squeeze the middle part of the interval onto the whole interval and send the edges to the ends; the parts of the square that are sent to the boundary get the value x_0.

Watch out: Abelian does not mean "free of structure"

For $n \geq 2$ the group $\pi_n(X)$ is abelian, but $\pi_1(X)$ still acts on it (by dragging the base point around a loop, §1.3.5). For $\mathbb{RP}^2$ the action on $\pi_2(\mathbb{RP}^2) = \mathbb{Z}$ is $n \mapsto -n$. For simply connected spaces, and for topological groups, the action is trivial.

First properties

The following facts are proved exactly as for π_1; we use them freely.

- *Functoriality.* A based map $h : X \to Y$ induces a homomorphism $h_* : \pi_n(X) \to \pi_n(Y)$, $h_*[f] = [h \circ f]$, with $(h \circ k)_* = h_* \circ k_*$.
- *Homotopy invariance.* A homotopy equivalence induces isomorphisms on all π_n. In particular a contractible space has $\pi_n = 0$ for all n: $\pi_n(\mathbb{R}^k) = 0$, $\pi_n(\text{ball}) = 0$.
- *Products.* $\pi_n(X \times Y) \cong \pi_n(X) \times \pi_n(Y)$, as in (1.34).
- *Topological groups.* For a topological group G the pointwise product also gives $\pi_n(G)$ a group law, and by Eckmann–Hilton it agrees with (1.43). This is often the easiest way to add two elements of $\pi_3(\mathrm{SU}(2))$: multiply the matrices.

Coverings do not see higher homotopy

We found $\pi_1(S^1)$ by unwrapping the circle onto the line $\mathbb{R}$. The same unwrapping shows that the circle has *no* higher homotopy: Nash's (1.32), second line. The mechanism is general.

Key idea: Lifting spheres through a covering

Let $p : \tilde{X} \to X$ be a **covering map**: every point of X has a neighbourhood U such that $p^{-1}(U)$ is a disjoint union of copies of U, each mapped homeomorphically onto U by p (examples: $\mathbb{R} \to S^1$, $S^2 \to \mathbb{RP}^2$, $\mathrm{SU}(2) \to \mathrm{SO}(3)$, $\mathbb{R}^2 \to T^2$). Then for $n \geq 2$

$$p_* : \pi_n(\tilde{X}) \xrightarrow{\cong} \pi_n(X).$$

Reason: a based map $f : S^n \to X$ lifts to $\tilde f : S^n \to \tilde X$ because S^n is simply connected (below), and so does every homotopy of such maps (the cylinder $[0,1] \times S^n$ is simply connected too). Lifts of homotopic maps are homotopic, and p of a lift gives back the map.

Worked example 1.35: $\pi_n(S^1) = 0$ for $n \geq 2$, by an explicit lift

Let $f : S^n \to S^1$ be based, $f(x_0) = 1$, with $n \geq 2$. We build a lift $\tilde f : S^n \to \mathbb{R}$ with $e^{i\tilde f} = f$ and $\tilde f(x_0) = 0$. For any point $x \in S^n$ choose a path c from x_0 to x, lift the path $f \circ c$ in S^1 to $\mathbb{R}$ starting at 0 (the path lifting of §1.3.2), and let $\tilde f(x)$ be its endpoint.

- *The answer does not depend on the path.* If c' is another path from x_0 to x, then $c \cdot \bar c'$ is a loop in S^n at x_0. The difference of the two lifted endpoints is

$$\widetilde{f \circ c}(1) - \widetilde{f \circ c'}(1) \overset{(1)}{=} 2\pi\, W\big[f \circ (c \cdot \bar c')\big] \overset{(2)}{=} 2\pi\, W[\text{const}] \overset{(3)}{=} 0.$$

  (1) Lifting $f \circ c$ and then $f \circ \bar c'$ (which walks the second lift backwards) gives the lift of the loop $f \circ (c \cdot \bar c')$, whose total climb is 2π times its winding number (1.32).
  (2) $\pi_1(S^n) = 0$ (Example 1.27), so $c \cdot \bar c'$ contracts in S^n, hence $f \circ (c \cdot \bar c')$ contracts in S^1, and W is a homotopy invariant.
  (3) The constant loop has winding number 0.

- *The lift is null-homotopic, hence so is f.* On $\mathbb{R}$ we may interpolate linearly. Define

$$F(s,x) = \exp\big(i(1-s)\tilde f(x)\big), \qquad s \in [0,1].$$

  Then $F(0,x) = f(x)$, $F(1,x) = 1$, and $F(s,x_0) = e^0 = 1$ for all s: a based homotopy from f to the constant map.

So $\pi_n(S^1) = 0$ for $n \geq 2$. Together with $\pi_1(S^1) = \mathbb{Z}$ this is Nash's

$$\pi_m(S^1) = \begin{cases} \mathbb{Z}, & m = 1, \\ 0, & m > 1. \end{cases} \tag{1.46}$$

The same lift works for any simply connected source: every map from a simply connected space into S^1 is null-homotopic. For $n = 1$ the argument fails precisely at step (2), and the failure is measured by the winding number.

Worked example 1.36: Consequences of the lifting principle

- $T^k = \mathbb{R}^k/\mathbb{Z}^k$ is covered by $\mathbb{R}^k$, so $\pi_n(T^k) = \pi_n(\mathbb{R}^k) = 0$ for $n \geq 2$. A torus has only $\pi_1 = \mathbb{Z}^k$.
- Every closed orientable surface of genus $g \geq 1$ is covered by the plane (flat for $g = 1$, hyperbolic for $g \geq 2$), so $\pi_n(\Sigma_g) = 0$ for $n \geq 2$. Such spaces, with only one nonzero homotopy group, are the Eilenberg–MacLane spaces of Nash §4: S^1 is a $K(\mathbb{Z},1)$ and Σ_g is a $K(\pi_1(\Sigma_g),1)$.
- $S^2 \to \mathbb{RP}^2$ is a two-sheeted covering, so $\pi_n(\mathbb{RP}^2) = \pi_n(S^2)$ for $n \geq 2$; in particular $\pi_2(\mathbb{RP}^2) = \mathbb{Z}$ (point defects in nematics).
- $\mathrm{SU}(2) \to \mathrm{SO}(3)$ is a two-sheeted covering (§1.3.4), so $\pi_n(\mathrm{SO}(3)) = \pi_n(S^3)$ for $n \geq 2$. In particular $\pi_3(\mathrm{SO}(3)) = \pi_3(S^3) = \mathbb{Z}$.

1.3.7 Spheres in spheres, I: $\pi_m(S^n) = 0$ for $m < n$

Nash now turns to the most basic question: what are the homotopy groups of the spheres themselves? He quotes (from Bott and Tu) the answer for $m \le n$,

$$\pi_m(S^n) = \begin{cases} 0, & m < n, \\ \mathbb{Z}, & m = n, \end{cases} \tag{1.47}$$

and calls it "relatively easy". We now give the reasons. The first line is the argument of Example 1.27 in higher dimension.

Key idea: A lower-dimensional sphere slides off

If $m < n$, every map $f : S^m \to S^n$ misses a point after a small deformation, and a map into S^n minus a point is null-homotopic because S^n minus a point is $\mathbb{R}^n$.

The steps are these.

(a) *Smooth it.* Any continuous f is homotopic to a smooth map, through maps close to f (smooth the components of f in $\mathbb{R}^{n+1}$ by averaging over small balls, then divide by the length to project back onto S^n). Close maps are homotopic by the straight-line homotopy followed by projection, which never passes through 0 if the maps are less than a distance 1 apart.

(b) *Count dimensions.* The image of a smooth map from an m-dimensional manifold has zero n-dimensional volume when $m < n$ (locally it is the image of a small m-cube, which fits in an ε-neighbourhood of an m-dimensional patch: volume $\sim \varepsilon^{n-m} \to 0$). So there is a point $P \in S^n$ not in the image, and we can choose $P \neq x_0$.

(c) *Contract through the complement.* Let $\sigma : S^n \setminus \{P\} \to \mathbb{R}^n$ be stereographic projection from P, and $c = \sigma(x_0)$. Set

$$F(s, x) = \sigma^{-1}\Big((1 - s)\,\sigma\big(f(x)\big) + s\,c\Big). \tag{1.48}$$

We check that (1.48) is a based null-homotopy:

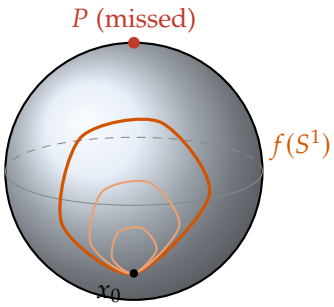
$$F(0, x) \overset{(1)}{=} \sigma^{-1}\big(\sigma(f(x))\big) = f(x), \qquad F(1, x) \overset{(2)}{=} \sigma^{-1}(c) = x_0,$$
$$F(s, x_0) \overset{(3)}{=} \sigma^{-1}\big((1 - s)\,c + s\,c\big) = \sigma^{-1}(c) = x_0.$$

(1) $f(x) \neq P$, so $\sigma(f(x))$ is defined; $\sigma^{-1} \circ \sigma = \mathrm{id}$.
(2) At $s = 1$ only the constant c survives.
(3) $f(x_0) = x_0$, so $\sigma(f(x_0)) = c$.

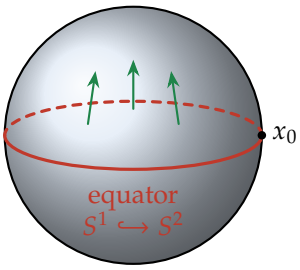
The straight line in $\mathbb{R}^n$ never reaches the "point at infinity" P, so F is continuous. Hence $[f] = 0$ and $\pi_m(S^n) = 0$ for $m < n$.

Watch out: Why the smoothing step is needed

Without step (a), step (b) fails: there are continuous surjections $S^1 \to S^2$ (Peano curves). The conclusion is still true, but only because every continuous map is homotopic to a smooth one. The whole of differential topology is full of this move: approximate a continuous object by a smooth one, then count dimensions.



a loop on S^2 misses P
and shrinks through $S^2 \setminus P \cong \mathbb{R}^2$



$m = n$ is different: the identity
$S^2 \to S^2$ hits every point

Figure 1.24: Left: for $m < n$ a (smoothed) map $S^m \to S^n$ leaves a point P uncovered. Stereographic projection from P turns the rest of the sphere into the plane, where the loop contracts to the base point (fainter copies). Right: the equator can be slid up (green arrows) over the north pole and shrunk, while the identity map of S^2, which covers every point, has no uncovered point to project from. That is why $\pi_2(S^2)$ can be nonzero.

1.3.8 Spheres in spheres, II: the degree and $\pi_n(S^n) = \mathbb{Z}$

For $m = n$ the dimension count gives nothing: an n-sphere can cover an n-sphere. The question becomes: *how many times* does it cover it? For $n = 1$ the answer was the winding number. For general n it is the **degree**. We give three equivalent working definitions. They are the same number seen by a counter, by an integrator, and by a topologist.

Working definition 1.22 (degree of a map $f : S^n \to S^n$)

(i) *Counting.* Pick a point y of the target that is a **regular value**: at every preimage $x \in f^{-1}(y)$ the Jacobian $\det(\mathrm{d}f_x)$ (computed in oriented local coordinates) is nonzero. Then there are finitely many preimages, and

$$\deg f = \sum_{x \in f^{-1}(y)} \operatorname{sgn} \det(\mathrm{d}f_x). \tag{1.49}$$

[each preimage counts $+1$ if f keeps orientation near it, -1 if it reverses it]

(ii) *Integrating.* Let ω be any n-form [an integrand $\rho(y)\,\mathrm{d}y^1 \cdots \mathrm{d}y^n$ in local coordinates] on the target with $\int_{S^n} \omega = 1$ (for instance the normalised volume form). Then

$$\deg f = \int_{S^n} f^*\omega. \tag{1.50}$$

[$f^*\omega$ is ω with $y = f(x)$ substituted, so $\mathrm{d}y$ becomes the Jacobian times $\mathrm{d}x$]

(iii) *For fields* $\vec{n} : S^2 \to S^2$ (unit vectors, $\vec{n} \cdot \vec{n} = 1$) in local coordinates (u^1, u^2) on the source:

$$\deg \vec{n} = \frac{1}{4\pi} \int \vec{n} \cdot \left(\partial_1 \vec{n} \times \partial_2 \vec{n}\right) \mathrm{d}u^1\,\mathrm{d}u^2. \tag{1.51}$$

The first definition makes it obvious that $\deg f$ is an integer; the second and third make it obvious that it is a homotopy invariant [unchanged when f is deformed continuously]; it is a theorem (below) that they agree, and that the answer does not depend on the choices of y or ω.

The integrand of (1.51) has a simple meaning. The small coordinate rectangle $\mathrm{d}u^1\,\mathrm{d}u^2$ in the source is mapped to a small parallelogram on the target sphere, spanned by the tangent vectors $\partial_1\vec{n}\,\mathrm{d}u^1$ and $\partial_2\vec{n}\,\mathrm{d}u^2$. Its area, with a sign telling whether its orientation agrees with the outward normal $\vec{n}$, is the triple product $\vec{n}\cdot(\partial_1\vec{n}\times\partial_2\vec{n})\,\mathrm{d}u^1\mathrm{d}u^2$. So (1.51) is "total signed area swept out on the target, divided by the area 4π of the unit sphere". It is (1.50) with ω the area form divided by 4π. For $n=1$ the analogue is the winding number (1.33): total signed angle swept, divided by 2π.

Worked example 1.37: The identity and the k-fold wrapping of S^2

Use spherical coordinates (θ,ϕ) on the source, oriented so that $\mathrm{d}\theta\,\mathrm{d}\phi$ is positive (this is the outward orientation, since $\hat\theta\times\hat\phi=\hat r$). For an integer k let

$$\vec{n}_k(\theta,\phi)=\big(\sin\theta\cos k\phi,\ \sin\theta\sin k\phi,\ \cos\theta\big),$$

which wraps each circle of latitude k times around itself; $\vec{n}_1$ is the identity. Then

$$\begin{aligned}\partial_\theta\vec{n}_k\times\partial_\phi\vec{n}_k &\overset{(1)}{=}(\cos\theta\cos k\phi,\ \cos\theta\sin k\phi,\ -\sin\theta)\times(-k\sin\theta\sin k\phi,\ k\sin\theta\cos k\phi,\ 0)\\ &\overset{(2)}{=}k\sin\theta\,\big(\sin\theta\cos k\phi,\ \sin\theta\sin k\phi,\ \cos\theta\big)=k\sin\theta\,\vec{n}_k,\end{aligned}$$

(1) Differentiate each component.

(2) Call the two vectors $\vec{a}$ and $\vec{b}$. The components of $\vec{a}\times\vec{b}$ are x: $\cos\theta\sin k\phi\cdot 0-(-\sin\theta)(k\sin\theta\cos k\phi)=k\sin^2\theta\cos k\phi$; y: $(-\sin\theta)(-k\sin\theta\sin k\phi)-\cos\theta\cos k\phi\cdot 0=k\sin^2\theta\sin k\phi$; z: $\cos\theta\cos k\phi\cdot k\sin\theta\cos k\phi+\cos\theta\sin k\phi\cdot k\sin\theta\sin k\phi=k\sin\theta\cos\theta$.

and therefore

$$\deg\vec{n}_k\overset{(1)}{=}\frac{1}{4\pi}\int_0^{2\pi}\!\!\int_0^{\pi}k\sin\theta\,(\vec{n}_k\cdot\vec{n}_k)\,\mathrm{d}\theta\,\mathrm{d}\phi\overset{(2)}{=}\frac{k}{4\pi}\cdot 2\cdot 2\pi=k.$$

(1) Formula (1.51) with the cross product just computed.

(2) $\vec{n}_k\cdot\vec{n}_k=1$, $\int_0^\pi\sin\theta\,\mathrm{d}\theta=2$, $\int_0^{2\pi}\mathrm{d}\phi=2\pi$.

Check by counting: take the regular value $y=(1,0,0)$. Its preimages have $\theta=\pi/2$ and $\cos k\phi=1$, i.e. $\phi=2\pi j/k$, $j=0,\dots,|k|-1$: exactly $|k|$ points. At each, the Jacobian is $\partial(\theta,k\phi)/\partial(\theta,\phi)=k$, with sign $\operatorname{sgn}k$. So (1.49) gives $|k|\operatorname{sgn}k=k$ as well. Every integer is the degree of some map.

In complex language: stereographic projection identifies S^2 with the Riemann sphere $\mathbb{C}\cup\{\infty\}$, $z=e^{i\phi}\tan(\theta/2)$, and $\vec{n}_k$ becomes (after a harmless reparametrisation of θ) $z\mapsto z^k$. More generally a rational map $z\mapsto P(z)/Q(z)$ with no common factors has degree $\max(\deg P,\deg Q)$: a generic point has that many preimages, all counted $+1$ because holomorphic maps preserve orientation.

Worked example 1.38: Reflections and the antipodal map

The reflection $r(x_1,x_2,\dots,x_{n+1})=(-x_1,x_2,\dots,x_{n+1})$ of S^n has degree -1. By counting: the point $y=(0,\dots,0,1)$ has the single preimage y itself, and near it r in the local coordinates $(x_1,\dots,x_n)$ is $\operatorname{diag}(-1,1,\dots,1)$, with determinant -1.

The antipodal map $a(x)=-x$ is the composition of $n+1$ reflections, one for each coordinate. Degrees multiply under composition (the preimage count of a composition is the product

of counts, signs included), so

$$\deg a = (-1)^{n+1}. \tag{1.52}$$

On S^2 we can confirm with (1.51):

$$\deg(-\vec{n}) \overset{(1)}{=} \frac{1}{4\pi}\int (-\vec{n})\cdot\big((-\partial_1\vec{n})\times(-\partial_2\vec{n})\big)\,\mathrm{d}^2u \overset{(2)}{=} -\frac{1}{4\pi}\int \vec{n}\cdot(\partial_1\vec{n}\times\partial_2\vec{n})\,\mathrm{d}^2u = -\deg\vec{n}.$$

(1) Replace $\vec{n}$ by $-\vec{n}$ in (1.51).
(2) Three minus signs make one.

With $\vec{n}$ the identity, the antipodal map of S^2 has degree -1, so it is *not* homotopic to the identity. On odd spheres it is: on S^1 it is the rotation by π, and on $S^{2k-1}\subset\mathbb{C}^k$ it is the end point of the rotation $x\mapsto e^{i\pi s}x$, $s\in[0,1]$.

Worked example 1.39: The hairy ball theorem

Claim: every continuous tangent vector field on S^{2k} vanishes somewhere. ("You cannot comb a hairy ball flat"; there is always a point on Earth where the horizontal wind speed is zero.)
Suppose $\vec{v}(x)$ is tangent, $\vec{v}(x)\cdot x = 0$, and never zero. Normalise, $\hat{v} = \vec{v}/|\vec{v}|$, and rotate each point x towards $\hat{v}(x)$ along a great circle:

$$F(s,x) = \cos(\pi s)\,x + \sin(\pi s)\,\hat{v}(x), \qquad s\in[0,1].$$

This stays on the sphere:

$$|F(s,x)|^2 \overset{(1)}{=} \cos^2(\pi s)\,|x|^2 + 2\cos(\pi s)\sin(\pi s)\,x\cdot\hat{v} + \sin^2(\pi s)\,|\hat{v}|^2 \overset{(2)}{=} \cos^2(\pi s)+\sin^2(\pi s) = 1.$$

(1) Expand the square.
(2) $|x| = |\hat{v}| = 1$ and $x\cdot\hat{v} = 0$ (tangency).

At $s=0$, F is the identity; at $s=1$, $F(1,x) = -x$ is the antipodal map. So the identity would be homotopic to the antipodal map. But their degrees are 1 and $(-1)^{2k+1} = -1$ by (1.52), and the degree is a homotopy invariant (below). Contradiction. On odd spheres there is no contradiction, and indeed $\vec{v}(x_1,x_2,\dots) = (-x_2,x_1,-x_4,x_3,\dots)$ is a nowhere-zero tangent field on S^{2k-1}.
This is the promised explanation of the remark in Example 1.9 that the antipodal map of S^2 is not homotopic to the identity. It is also our first "index theorem": the zeros of a vector field on S^2, counted with their indices, add up to the Euler characteristic $\chi(S^2) = 2$ (the Poincaré–Hopf theorem; the index of a zero is the degree of $\hat{v}$ on a small circle around it).

Why the degree is an integer and a homotopy invariant

We give the argument for fields $\vec{n}: S^2\to S^2$, where everything is explicit. Two facts carry it.

Fact 1: the degree integral does not change under deformations. Let $\vec{n}$ depend smoothly on a deformation parameter, and write $\delta\vec{n}$ for its derivative. Since $\vec{n}\cdot\vec{n} = 1$, the three vectors $\delta\vec{n}$, $\partial_1\vec{n}$, $\partial_2\vec{n}$ are all orthogonal to $\vec{n}$, so they lie in one plane and any triple product of them vanishes. With the density $q := \vec{n}\cdot(\partial_1\vec{n}\times\partial_2\vec{n})$,

$$\begin{aligned}\delta q &\overset{(1)}{=} \delta\vec{n}\cdot(\partial_1\vec{n}\times\partial_2\vec{n}) + \vec{n}\cdot(\partial_1\delta\vec{n}\times\partial_2\vec{n}) + \vec{n}\cdot(\partial_1\vec{n}\times\partial_2\delta\vec{n})\\ &\overset{(2)}{=} \vec{n}\cdot(\partial_1\delta\vec{n}\times\partial_2\vec{n}) + \vec{n}\cdot(\partial_1\vec{n}\times\partial_2\delta\vec{n})\end{aligned}$$

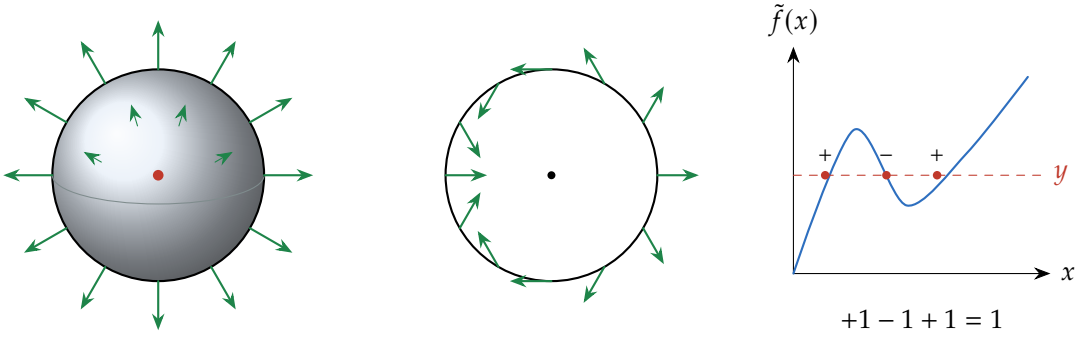


Figure 1.25: Pictures of degree. Left: the hedgehog field $\vec n(\vec x) = \hat r$ on a sphere surrounding a point defect (red) is the identity map $S^2 \to S^2$, of degree 1. Middle: the equator of $\vec n_2$ of Example 1.37: as we go once around, the arrow turns twice. Right: counting preimages, drawn for S^1 via the lift $\tilde f$. The horizontal line y is crossed three times, twice going up (+) and once going down (−). The net count 1 is the same for every regular value y: moving the line creates or destroys crossings in ± pairs.

$$\overset{(3)}{=} \partial_1\big[\vec n \cdot (\delta\vec n \times \partial_2\vec n)\big] + \partial_2\big[\vec n \cdot (\partial_1\vec n \times \delta\vec n)\big].$$

(1) Product rule on the triple product; δ commutes with ∂_i.
(2) The first term is a triple product of three vectors tangent to the sphere at $\vec n$: zero.
(3) Expand the right side by the product rule. The terms where ∂_i hits $\vec n$ outside the bracket are $\partial_1\vec n \cdot (\delta\vec n \times \partial_2\vec n)$ and $\partial_2\vec n \cdot (\partial_1\vec n \times \delta\vec n)$: triple products of tangent vectors, zero. The terms with second derivatives are $\vec n \cdot (\delta\vec n \times \partial_1\partial_2\vec n) + \vec n \cdot (\partial_2\partial_1\vec n \times \delta\vec n) = 0$ by antisymmetry of $\times$. What is left is the line above.

So δq is a total divergence, and its integral over the closed sphere vanishes (Stokes, or the divergence theorem, with no boundary). Hence $\delta(\deg \vec n) = 0$: the integral (1.51) is constant along any smooth homotopy. The same holds for (1.50) in general, in one line: a homotopy changes $f^*\omega$ by an exact form $\mathrm{d}\eta$, and $\int_{S^n} \mathrm{d}\eta = \int_\varnothing \eta = 0$. (This is the fact "the integral of an exact form over a closed manifold vanishes" of Chapter T2; de Rham cohomology in Nash §4 turns it into a theory.)

Fact 2: the integral equals the signed count. By Fact 1 applied to the *target* form, $\int f^*\omega$ depends on ω only through $\int \omega = 1$: if ω' is another n-form with total integral 1, then $\omega' - \omega = \mathrm{d}\eta$ (on S^n a top form with zero integral is exact; this is $H^n_{\mathrm{dR}}(S^n) = \mathbb{R}$), and $\int f^*\mathrm{d}\eta = \int \mathrm{d}f^*\eta = 0$. Now choose ω to be a bump form concentrated in a tiny disc D around a regular value y. Then $f^*\omega$ lives only on $f^{-1}(D)$, which (by the inverse function theorem at each regular preimage) is a disjoint union of small discs $D_1, \dots, D_N$ around the preimages $x_1, \dots, x_N$, each mapped diffeomorphically onto D. So

$$\int_{S^n} f^*\omega \overset{(1)}{=} \sum_{j=1}^{N} \int_{D_j} f^*\omega \overset{(2)}{=} \sum_{j=1}^{N} \operatorname{sgn}\det(\mathrm{d}f_{x_j}) \int_D \omega \overset{(3)}{=} \sum_{j=1}^{N} \operatorname{sgn}\det(\mathrm{d}f_{x_j}).$$

(1) $f^*\omega = 0$ outside $f^{-1}(D) = D_1 \cup \cdots \cup D_N$.
(2) Change of variables $x \mapsto f(x)$ on each D_j: the integral of a pulled-back form is the integral of the form, times +1 if f preserves orientation there and -1 if it reverses it.
(3) $\int_D \omega = \int_{S^n} \omega = 1$.

This proves that (1.49) and (1.50) agree, that the count is independent of the regular value y,

and that the integral is an integer. (Regular values exist, in fact almost every y is one: this is Sard's theorem.)

Key idea: Nash's (1.33), second line: $\pi_n(S^n) \cong \mathbb{Z}$ (Hopf)

The degree defines a group isomorphism

$$\deg : \pi_n(S^n) \xrightarrow{\cong} \mathbb{Z}.$$

It is well defined by Fact 1; it is onto because the maps $\vec{n}_k$ (and their analogues, suspensions of $z \mapsto z^k$) realise every integer; it is a homomorphism because the preimages of y under $f + g$ are those of f (squeezed into one half of the cube) together with those of g. That it is one-to-one, i.e. *maps of equal degree are homotopic*, is Hopf's theorem: the one genuinely hard step, which we take on trust. Since S^n is simply connected for $n \geq 2$, base points do not matter and $[S^n, S^n] = \pi_n(S^n) = \mathbb{Z}$ (Example 1.34).

Physical picture: Baby skyrmions and the Belavin–Polyakov bound

In a two-dimensional ferromagnet the magnetisation direction is a unit vector field $\vec{n}(x, y)$. Finite energy requires $\vec{n} \to \hat{z}$ at infinity, so we may add the point at infinity and regard $\vec{n}$ as a map from $\mathbb{R}^2 \cup \{\infty\} \cong S^2$ to S^2. Its degree $Q = \frac{1}{4\pi} \int q \, \mathrm{d}x \, \mathrm{d}y$ is the **skyrmion number**; it is conserved by any smooth time evolution, by Fact 1. It also bounds the exchange energy $E = \frac{1}{2} \int |\partial_i \vec{n}|^2 \, \mathrm{d}^2x$ from below. Since a square is non-negative,

$$\begin{aligned} 0 &\overset{(1)}{\leq} \int \left|\partial_i \vec{n} \pm \varepsilon_{ij}\, \vec{n} \times \partial_j \vec{n}\right|^2 \mathrm{d}^2x \\ &\overset{(2)}{=} \int \Big(|\partial_i \vec{n}|^2 + |\vec{n} \times \partial_j \vec{n}|^2 \pm 2\varepsilon_{ij}\, \partial_i \vec{n} \cdot (\vec{n} \times \partial_j \vec{n}) \Big) \mathrm{d}^2x \\ &\overset{(3)}{=} \int \left(2|\partial_i \vec{n}|^2 \mp 4\, q\right) \mathrm{d}^2x \overset{(4)}{=} 4E \mp 16\pi Q, \end{aligned}$$

(1) The integrand is a sum of squares (sum over $i = 1, 2$; $\varepsilon_{12} = -\varepsilon_{21} = 1$).
(2) Expand; $\varepsilon_{ij}\varepsilon_{ik} = \delta_{jk}$ in two dimensions turns the second square into $\sum_j |\vec{n} \times \partial_j \vec{n}|^2$.
(3) $|\vec{n} \times \partial_j \vec{n}| = |\partial_j \vec{n}|$ because $\partial_j \vec{n} \perp \vec{n}$ and $|\vec{n}| = 1$; and the cross term is $\varepsilon_{ij}\, \vec{n} \cdot (\partial_j \vec{n} \times \partial_i \vec{n})$ (cyclic triple product), which for $(i, j) = (1, 2)$ and $(2, 1)$ gives $-q$ twice.
(4) $\int |\partial_i \vec{n}|^2 = 2E$ and $\int q = 4\pi Q$.

So $E \geq 4\pi|Q|$, with equality exactly when $\partial_i \vec{n} = \mp \varepsilon_{ij}\, \vec{n} \times \partial_j \vec{n}$, a first-order "self-duality" equation. Topology forces a minimum energy in each sector. This is the two-dimensional model of the instanton bound $S_{\mathrm{YM}} \geq 8\pi^2|k|/g^2$ of Chapter 8.

Try it: Degree of a map given by a formula

Compute the degree of $f : \mathbb{C} \cup \{\infty\} \to \mathbb{C} \cup \{\infty\}$, $f(z) = \bar{z}^3$, by counting the preimages of $y = 1$ with signs.
Answer: the preimages are the three cube roots of unity. Near each, f is the composition of the holomorphic map $z \mapsto z^3$ (orientation preserving, Jacobian $|3z^2|^2 > 0$) with complex conjugation (a reflection, Jacobian -1). So each counts -1 and $\deg f = -3$.

1.3.9 Spheres in spheres, III: the Hopf map and $\pi_3(S^2) = \mathbb{Z}$

So far the pattern is tidy: $\pi_m(S^n) = 0$ below the dimension and $\mathbb{Z}$ at the dimension. One might guess that $\pi_m(S^n) = 0$ for $m > n$ as well, as it is for S^1. It is not. The first counterexample, found by Hopf in 1931, is a map from S^3 onto S^2 that cannot be deformed to a constant. Every physicist already knows it: it is the map from a spin-$\frac{1}{2}$ state to its Bloch vector.

Working definition 1.23 (the Hopf map)

Write $S^3 = \{(z_1, z_2) \in \mathbb{C}^2 : |z_1|^2 + |z_2|^2 = 1\}$, i.e. normalised spinors $\psi = \binom{z_1}{z_2}$. The **Hopf map** $h : S^3 \to S^2$ sends a spinor to the expectation value of the Pauli matrices:

$$h(\psi) = \vec{n}, \qquad n_a = \psi^\dagger \sigma_a \psi, \qquad \vec{n} = \big(2\,\mathrm{Re}(\bar{z}_1 z_2),\ 2\,\mathrm{Im}(\bar{z}_1 z_2),\ |z_1|^2 - |z_2|^2\big). \tag{1.53}$$

[$\vec{n}$ is the Bloch vector. It has unit length, since $|\vec{n}| = |z_1|^2 + |z_2|^2 = 1$]

The components follow from the Pauli matrices of (1.37): $\psi^\dagger\sigma_1\psi = \bar{z}_1 z_2 + \bar{z}_2 z_1 = 2\,\mathrm{Re}(\bar{z}_1 z_2)$, $\psi^\dagger\sigma_2\psi = -i\bar{z}_1 z_2 + i\bar{z}_2 z_1 = 2\,\mathrm{Im}(\bar{z}_1 z_2)$ and $\psi^\dagger\sigma_3\psi = |z_1|^2 - |z_2|^2$. The image really is on the unit sphere:

$$|\vec{n}|^2 \overset{(1)}{=} 4\,|\bar{z}_1 z_2|^2 + \big(|z_1|^2 - |z_2|^2\big)^2 \overset{(2)}{=} \big(|z_1|^2 + |z_2|^2\big)^2 \overset{(3)}{=} 1.$$

(1) $(2\,\mathrm{Re}\,w)^2 + (2\,\mathrm{Im}\,w)^2 = 4|w|^2$ with $w = \bar{z}_1 z_2$.
(2) $4ab + (a-b)^2 = (a+b)^2$ with $a = |z_1|^2$, $b = |z_2|^2$.
(3) ψ is normalised.

The fibres are circles: the global phase

Multiplying ψ by a phase does not change $\vec{n}$, because $(e^{i\alpha}\psi)^\dagger\sigma_a(e^{i\alpha}\psi) = e^{-i\alpha}e^{i\alpha}\psi^\dagger\sigma_a\psi$. Conversely, if $h(\psi) = h(\psi')$ then $\psi' = e^{i\alpha}\psi$ for some α. Indeed $\vec{n}$ fixes $|z_1|^2 = \frac{1}{2}(1 + n_3)$ and $|z_2|^2 = \frac{1}{2}(1 - n_3)$ and the product $\bar{z}_1 z_2 = \frac{1}{2}(n_1 + i n_2)$. If $z_1 \neq 0$, write $z_1' = e^{i\alpha} z_1$ (same modulus); then $\bar{z}_1' z_2' = \bar{z}_1 z_2$ gives $z_2' = e^{i\alpha} z_2$. If $z_1 = 0$ then $z_1' = 0$, and z_2, z_2' are two unit complex numbers, so again $z_2' = e^{i\alpha} z_2$. So

$$h^{-1}(\vec{n}) = \{e^{i\alpha}\psi : \alpha \in [0, 2\pi)\} \cong S^1.$$

The three-sphere is a union of disjoint circles, one over each point of S^2:

$$S^1 \longrightarrow S^3 \xrightarrow{\ h\ } S^2. \tag{1.54}$$

This is the **Hopf fibration**, our first nontrivial fibre bundle (Nash §5, next part). In quantum-mechanical language: the space of normalised spinors is S^3, the space of physical states (rays) is the Bloch sphere $S^2 = \mathbb{CP}^1$, and the fibre is the unobservable global phase U(1). The same bundle, viewed over S^2 as a U(1) bundle, is the Dirac monopole of charge one.

Dictionary: the Hopf fibration

normalised spinor $\psi \in \mathbb{C}^2$	point of S^3
Bloch vector $\langle\vec{\sigma}\rangle$, or the ray $[z_1 : z_2] \in \mathbb{CP}^1$	point of S^2
global phase $e^{i\alpha}$	fibre $S^1 = \mathrm{U}(1)$
Berry connection $i\psi^\dagger \mathrm{d}\psi$	connection on the bundle
Berry phase of a loop on the Bloch sphere	holonomy; $-\frac{1}{2}\times$ solid angle

Why h is not null-homotopic: linked fibres

Pick two different points $\vec{n}, \vec{n}'$ of S^2. Their preimages are two disjoint circles in S^3. The claim is that they are *linked*, and that the linking number cannot change under homotopy.

Take the north and south poles. With $z_1 = x_1 + ix_2$, $z_2 = x_3 + ix_4$:

- $h^{-1}(0,0,1)$: $n_3 = 1$ forces $|z_1| = 1$, $z_2 = 0$: the circle $C_N = \{x_1^2 + x_2^2 = 1,\ x_3 = x_4 = 0\}$.
- $h^{-1}(0,0,-1)$: $z_1 = 0$, $|z_2| = 1$: the circle $C_S = \{x_3^2 + x_4^2 = 1,\ x_1 = x_2 = 0\}$.

To see them, project S^3 stereographically from the point $(0,0,0,1)$ onto $\mathbb{R}^3$: $(x_1, x_2, x_3, x_4) \mapsto (x_1, x_2, x_3)/(1 - x_4)$. Then C_N (where $x_4 = 0$) becomes the unit circle in the (x_1, x_2) plane, and C_S becomes $\{(0, 0, x_3/(1 - x_4))\}$, which runs over the entire x_3-axis as the point moves around C_S (through ∞ at $x_4 = 1$). The axis pierces the disc bounded by the unit circle exactly once: the linking number is 1 (Figure 1.26). Any two fibres of h are linked once in the same way.

Key idea: The Hopf invariant

For a smooth map $f : S^3 \to S^2$ and two regular values $y \neq y'$, the **Hopf invariant** $H(f)$ is the linking number of the curves $f^{-1}(y)$ and $f^{-1}(y')$ in S^3. It is a homotopy invariant, and

$$H : \pi_3(S^2) \xrightarrow{\cong} \mathbb{Z}, \qquad H(h) = 1.$$

So $\pi_3(S^2) = \mathbb{Z}$, generated by the Hopf map. A constant map has $H = 0$ (the preimage of $y' \neq$ the constant value is empty), so h is not null-homotopic.

There is also an integral formula, due to Whitehead, which makes the Hopf invariant look like physics. Let ω be the area form of S^2 divided by 4π, and $F := f^*\omega$, a closed 2-form on S^3. Since every closed 2-form on S^3 is exact (Nash §4: $H^2_{\mathrm{dR}}(S^3) = 0$), $F = \mathrm{d}A$ for some 1-form A, and

$$H(f) = \int_{S^3} A \wedge F. \tag{1.55}$$

Read F as a magnetic field $\vec{B}$ on (compactified) $\mathbb{R}^3$ and A as its vector potential: (1.55) is the **magnetic helicity** $\int \vec{A} \cdot \vec{B}\, \mathrm{d}^3x$ of plasma physics, which measures how much the lines of $\vec{B}$ are linked. Here the field lines of $\vec{B}$ are exactly the fibres $f^{-1}(y)$, and each pair links H times. Knotted solitons in the Faddeev–Skyrme model ("hopfions"), and knotted vortex configurations in some magnets, are classified by exactly this integer.

Worked example 1.40: The Hopf map in the language of spinors: its degree on a slice

A consistency check that ties $\pi_3(S^2)$ to $\pi_2(S^2)$. Restrict h to the set $\Sigma = \{(z_1, z_2) \in S^3 : z_2 \in \mathbb{R},\ z_2 \geq 0\}$ where z_2 is real and non-negative. Every fibre $\{e^{i\alpha}\psi\}$ meets Σ exactly once if $z_2 \neq 0$ (choose α to make z_2 real positive), and the fibre over the north pole ($z_2 = 0$) lies entirely in the boundary circle $z_2 = 0$ of Σ. So Σ is a disc $\{|z_1| \leq 1\}$ (with $z_2 = \sqrt{1 - |z_1|^2}$), and h maps its interior one-to-one onto S^2 minus the north pole, and its boundary circle to the north pole. Explicitly, with $z_1 = re^{i\phi}$, $z_2 = \sqrt{1 - r^2}$:

$$\vec{n} \overset{(1)}{=} \left(2r\sqrt{1 - r^2}\cos\phi,\ -2r\sqrt{1 - r^2}\sin\phi,\ 2r^2 - 1\right) \overset{(2)}{=} \left(\sin\theta\cos\phi,\ -\sin\theta\sin\phi,\ -\cos\theta\right),$$

(1) (1.53): $\bar{z}_1 z_2 = r\sqrt{1 - r^2}\, e^{-i\phi}$ and $|z_1|^2 - |z_2|^2 = r^2 - (1 - r^2)$.

(2) Put $r = \sin(\theta/2)$, so $2r\sqrt{1 - r^2} = \sin\theta$ and $2r^2 - 1 = -\cos\theta$, with $\theta \in [0, \pi]$ as $r \in [0, 1]$.

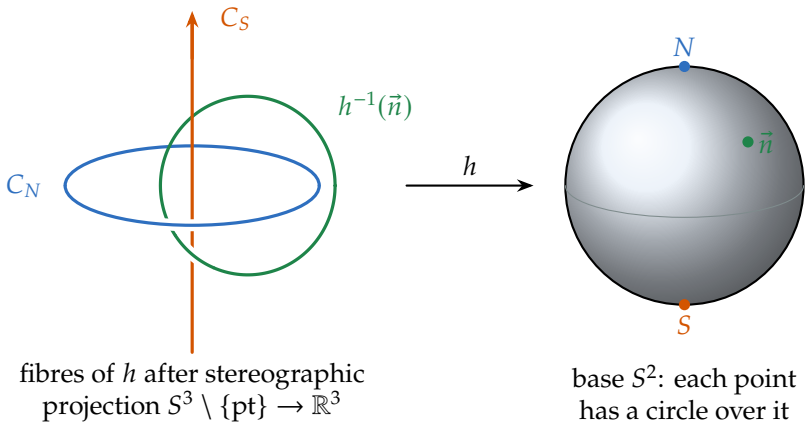


Figure 1.26: The Hopf fibration. Left: in S^3 (drawn in $\mathbb{R}^3$ by stereographic projection) the fibre over the north pole is the unit circle C_N (blue), the fibre over the south pole is the vertical axis C_S closed up through infinity (orange), and the fibre over any other point is a circle (green) that links both. Every pair of fibres links exactly once: the Hopf invariant of h is 1. Right: the base.

This is the disc, with its boundary collapsed, wrapped once over the sphere: a degree ± 1 map $D^2/\partial D^2 \cong S^2 \to S^2$. So the Hopf map "contains" the generator of $\pi_2(S^2)$. Read in reverse, this is the boundary map $\partial : \pi_2(S^2) \to \pi_1(S^1)$ of the exact sequence of §1.3.11: the boundary circle of Σ, which is a whole fibre, is the loop of phases that the generator of $\pi_2(S^2)$ is sent to.

1.3.10 The table of $\pi_m(S^n)$

Nash states the general situation for $m > n$, quoting Serre's theorem (1951):

$$\pi_m(S^n) = \begin{cases} \text{a finite group}, & m > n,\ n \text{ odd}, \\ \text{a finite group}, & m > n,\ n \text{ even},\ m \neq 2n-1, \\ \mathbb{Z} \times F, & m > n,\ n \text{ even},\ m = 2n-1, \end{cases} \tag{1.56}$$

with F some finite group. In words: apart from $\pi_n(S^n) = \mathbb{Z}$, the only infinite homotopy groups of spheres are $\pi_{2n-1}(S^n)$ for even n, and each of these contains exactly one copy of $\mathbb{Z}$. The extra $\mathbb{Z}$ is detected by the Hopf invariant, defined for maps $S^{2n-1} \to S^n$ as the linking number of two $(n-1)$-dimensional preimages in S^{2n-1}. (For odd n the linking number of two preimages is antisymmetric under exchanging them, so the Hopf invariant vanishes.) Maps of Hopf invariant 1 exist only for $n = 2, 4, 8$; they are built exactly like (1.53) from the complex numbers, the quaternions and the octonions ($S^3 \to S^2$, $S^7 \to S^4$, $S^{15} \to S^8$). This is Adams' theorem, and it is why there are only four normed division algebras $\mathbb{R}, \mathbb{C}, \mathbb{H}, \mathbb{O}$.

The actual groups are known only in a range. Nash remarks that "$\pi_m(S^2)$ is not known for all m"; that is still true. Figure 1.27 tabulates the beginning, arranged so that its structure is visible: row k lists $\pi_{n+k}(S^n)$, the maps that go up in dimension by k.

Stability: the Freudenthal suspension theorem

Look along a row of Figure 1.27. The entries change for small n and then settle down. Nash's (1.35)–(1.36) record this:

$$\pi_{m+n}(S^n) = \pi_{m+n+1}(S^{n+1}) = \cdots =: G_m \quad \text{for } n \geq m+2, \qquad G_m = \pi_{2m+2}(S^{m+2}). \tag{1.57}$$

	S^1	S^2	S^3	S^4	S^5	S^6	S^7	S^8	G_k
$k=0$	$\mathbb{Z}$	$\mathbb{Z}$	$\mathbb{Z}$	$\mathbb{Z}$	$\mathbb{Z}$	$\mathbb{Z}$	$\mathbb{Z}$	$\mathbb{Z}$	$\mathbb{Z}$
$k=1$	0	$\mathbb{Z}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_2$
$k=2$	0	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_2$
$k=3$	0	$\mathbb{Z}_2$	$\mathbb{Z}_{12}$	$\mathbb{Z}\times\mathbb{Z}_{12}$	$\mathbb{Z}_{24}$	$\mathbb{Z}_{24}$	$\mathbb{Z}_{24}$	$\mathbb{Z}_{24}$	$\mathbb{Z}_{24}$
$k=4$	0	$\mathbb{Z}_{12}$	$\mathbb{Z}_2$	$\mathbb{Z}_2^2$	$\mathbb{Z}_2$	0	0	0	0
$k=5$	0	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_2^2$	$\mathbb{Z}_2$	$\mathbb{Z}$	0	0	0
$k=6$	0	$\mathbb{Z}_2$	$\mathbb{Z}_3$	$\mathbb{Z}_{24}\times\mathbb{Z}_3$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_2$
$k=7$	0	$\mathbb{Z}_3$	$\mathbb{Z}_{15}$	$\mathbb{Z}_{15}$	$\mathbb{Z}_{30}$	$\mathbb{Z}_{60}$	$\mathbb{Z}_{120}$	$\mathbb{Z}\times\mathbb{Z}_{120}$	$\mathbb{Z}_{240}$

Figure 1.27: Homotopy groups of spheres: row k, column S^n contains $\pi_{n+k}(S^n)$. The diagonal $k=0$ is $\pi_n(S^n)=\mathbb{Z}$ (the degree). The red boxes are the infinite groups $\pi_{2n-1}(S^n)$, n even, predicted by Serre (1.56) and detected by the Hopf invariant. In the shaded region $n \geq k+2$ each row is constant: this is the stability (1.57) of Nash's (1.35), and the constant value is the stable group G_k of (1.36), listed on the right. Column S^1 is (1.46); column S^2 is column S^3 shifted down by one row (Example 1.45).

(Nash's m is our row label k.) The mechanism is suspension. A map $f : S^p \to S^n$ gives a map of suspensions

$$Sf : S^{p+1} = S^1 \wedge S^p \to S^1 \wedge S^n = S^{n+1}, \qquad (t,x) \mapsto (t, f(x)),$$

which "does f on every meridian" (Figure 1.12). A homotopy of f gives a homotopy of Sf, so this defines the **suspension homomorphism** $E : \pi_p(S^n) \to \pi_{p+1}(S^{n+1})$.

Key idea: Freudenthal (1937)

$E : \pi_p(S^n) \to \pi_{p+1}(S^{n+1})$ is an isomorphism for $p < 2n-1$ and onto for $p = 2n-1$. With $p = n+k$ the condition $p < 2n-1$ is $n \geq k+2$, which gives (1.57).

Why a bound like $p < 2n-1$? By the adjunction (1.25), $\pi_{p+1}(S^{n+1}) = [S^1 \wedge S^p, S^{n+1}]_0 = [S^p, \Omega S^{n+1}]_0 = \pi_p(\Omega S^{n+1})$, and under this identification E is induced by the inclusion

$$j : S^n \hookrightarrow \Omega S^{n+1}, \qquad x \mapsto \big(\text{the meridian loop of } S^{n+1} \text{ through } x\big).$$

So E is an isomorphism exactly when the loop space ΩS^{n+1} "looks like" S^n as far as p-spheres can tell. James showed that ΩS^{n+1} is built from S^n by attaching further cells of dimensions $2n, 3n, 4n, \ldots$ (roughly: loops that pass over the north pole twice, three times, ...). Now the general-position count (1.4) does the rest. A p-sphere in a space can be pushed off the interior of a $2n$-cell if $p < 2n$, so every element of $\pi_p(\Omega S^{n+1})$ comes from S^n (this is "onto" for $p \leq 2n-1$); a homotopy between two p-spheres is a $(p+1)$-dimensional object, which can be pushed off if $p+1 < 2n$ (this is "one-to-one" for $p < 2n-1$). Beyond that range the $2n$-cell makes itself felt, and it is precisely there that the Hopf invariant lives.

Worked example 1.41: Stability at work: from $\pi_3(S^2)=\mathbb{Z}$ to $\pi_4(S^3)=\mathbb{Z}_2$

Take $n=2$, $p=3=2n-1$. Freudenthal says $E : \pi_3(S^2) \to \pi_4(S^3)$ is onto but need not be one-to-one. Indeed $\pi_3(S^2) = \mathbb{Z}$ (generated by the Hopf map h) while $\pi_4(S^3) = \mathbb{Z}_2$

(generated by the suspension $\eta = Sh$), so

$$E(k\,[h]) = k\,[\eta] = \begin{cases} 0, & k \text{ even}, \\ [\eta], & k \text{ odd}. \end{cases}$$

The linking-number invariant dies on suspension: in S^4 two circles can be unlinked, and only a mod-2 remnant survives. From $n = 3$ on, E is an isomorphism, and the row $k = 1$ is constant: $G_1 = \mathbb{Z}_2$. Physics: since $\mathrm{SU}(2) \cong S^3$, this says $\pi_4(\mathrm{SU}(2)) = \mathbb{Z}_2$, and that is the origin of Witten's global SU(2) anomaly: a theory with an odd number of Weyl fermion doublets is inconsistent (Chapter 10).

Watch out: Homotopy groups do not stop at the dimension

S^2 is two-dimensional, yet $\pi_m(S^2) \neq 0$ for every $m \geq 2$ that has been computed. Maps of big spheres into small spheres can wrap in intricate, knotted ways (the fibres of the Hopf map are linked circles). By contrast, the cohomology groups of Nash §4 vanish above the dimension of the space. This is the main reason Nash calls homotopy groups "very difficult to calculate" and turns to cohomology.

1.3.11 Exact sequences

How were the entries of Figure 1.27 computed, and how do we know that $\pi_1(\mathrm{SU}(n)) = 0$ or $\pi_3(\mathrm{SU}(3)) = \mathbb{Z}$? The main tool is the exact sequence. Nash introduces exact sequences in §4 (for complexes) and uses the exact homotopy sequence of a fibre bundle in §5; since these tools are behind much of what we have just stated, we give a working introduction now.

Working definition 1.24 (exact sequence)

A sequence of groups and homomorphisms [maps respecting the group law, $f(aa') = f(a)f(a')$]

$$\cdots \longrightarrow A \xrightarrow{f} B \xrightarrow{g} C \longrightarrow \cdots$$

is **exact at** B if $\mathrm{Im}\, f = \mathrm{Ker}\, g$: what f produces is exactly what g kills. It is **exact** if it is exact at every place. Operational consequences (with 0 the trivial group):

- $0 \to A \xrightarrow{f} B$ exact $\iff$ f is one-to-one;
- $B \xrightarrow{g} C \to 0$ exact $\iff$ g is onto;
- $0 \to A \xrightarrow{f} B \to 0$ exact $\iff$ f is an isomorphism;
- $0 \to A \xrightarrow{f} B \xrightarrow{g} C \to 0$ exact (a **short exact sequence**) $\iff$ A is a subgroup of B and $C \cong B/A$ [B with any two elements differing by an element of A counted as the same].

Worked example 1.42: Short exact sequences a physicist knows

- $0 \to \mathbb{Z} \xrightarrow{\times 2} \mathbb{Z} \to \mathbb{Z}_2 \to 0$: doubling is one-to-one, its image is the even integers, and the quotient is "even or odd".

- $0 \to \mathbb{Z} \to \mathbb{R} \xrightarrow{\theta \mapsto e^{i\theta}} \mathrm{U}(1) \to 1$ with the first map $n \mapsto 2\pi n$: the kernel of $\theta \mapsto e^{i\theta}$ is $2\pi\mathbb{Z}$. This is the helix of Figure 1.16 as algebra.
- $1 \to \mathbb{Z}_2 \to \mathrm{SU}(2) \xrightarrow{R} \mathrm{SO}(3) \to 1$, with $\mathbb{Z}_2 = \{\pm\mathbf{1}\}$: the kernel of $U \mapsto R(U)$ in (1.38) is $\pm\mathbf{1}$, and R is onto.

The exact homotopy sequence of a fibration

Let $p : E \to B$ be a fibre bundle with fibre $F = p^{-1}(b_0)$ over the base point b_0, with base point $e_0 \in F$. Examples: the covering $\mathbb{R} \to S^1$ (fibre $\mathbb{Z}$, a discrete set); $\mathrm{SU}(2) \to \mathrm{SO}(3)$ (fibre $\{\pm\mathbf{1}\}$); the Hopf fibration (1.54) (fibre S^1). The essential property, which Nash takes as the definition of a *fibration* in §5, is *homotopy lifting*: a homotopy in B can be lifted to a homotopy in E, once the lift of its starting map is given. For coverings this is the path lifting of §1.3.2.

Key idea: The long exact homotopy sequence

For a fibration $F \xrightarrow{i} E \xrightarrow{p} B$ there is an exact sequence

$$\cdots \to \pi_n(F) \xrightarrow{i_*} \pi_n(E) \xrightarrow{p_*} \pi_n(B) \xrightarrow{\partial} \pi_{n-1}(F) \xrightarrow{i_*} \pi_{n-1}(E) \to \cdots \to \pi_0(E) \to \pi_0(B). \quad (1.58)$$

Here i_* and p_* are induced by inclusion and projection, and ∂ is the **boundary map**: lift a sphere in the base to a disc in the total space and look at the boundary of the disc, which lies in the fibre (Figure 1.28).

Let us see ∂ concretely in the cube picture. An element of $\pi_n(B)$ is a map $f : I^n \to B$ with $f(\partial I^n) = b_0$. Regard it as a homotopy, in the last variable t_n, of maps of the $(n-1)$-cube, starting at the constant map. Lift it, starting from the constant map at e_0: we get $\tilde{f} : I^n \to E$ with $p \circ \tilde{f} = f$. All faces of the cube except the top one ($t_n = 1$) still go to e_0 (they are lifts of constant maps that start at e_0, and such a lift can be chosen constant). The top face goes into $p^{-1}(b_0) = F$, and its boundary goes to e_0. So the top face is an element of $\pi_{n-1}(F)$: that is $\partial[f]$.

For $n = 1$ this is exactly the lift of a loop: a loop γ in B at b_0 lifts to a path in E from e_0 to some point $\tilde{\gamma}(1)$ of the fibre, and $\partial[\gamma]$ is the component of F containing that endpoint.

Why exactness holds, in two representative places.

- *At $\pi_n(E)$:* $\mathrm{Im}\, i_* = \mathrm{Ker}\, p_*$. If g is a sphere in the fibre, then $p \circ g$ is the constant map b_0, so $p_* i_* = 0$. Conversely, if $g : S^n \to E$ has $p \circ g$ null-homotopic, lift the null-homotopy of $p \circ g$ to a homotopy of g; at its end g has been deformed to a map whose projection is constant, i.e. a sphere in F.
- *At $\pi_n(B)$:* $\mathrm{Im}\, p_* = \mathrm{Ker}\, \partial$. If $f = p \circ g$ comes from E, then g itself is a lift with all faces at e_0, so the top face is constant: $\partial[f] = 0$. Conversely, if the top face is null-homotopic in F, use that null-homotopy to push the top face of $\tilde{f}$ to e_0 as well; the result is a sphere in E lying over f.

The remaining places are similar. We now use (1.58) as a machine. The recurring pattern: if the two neighbours of a group are zero, the map between them is an isomorphism.

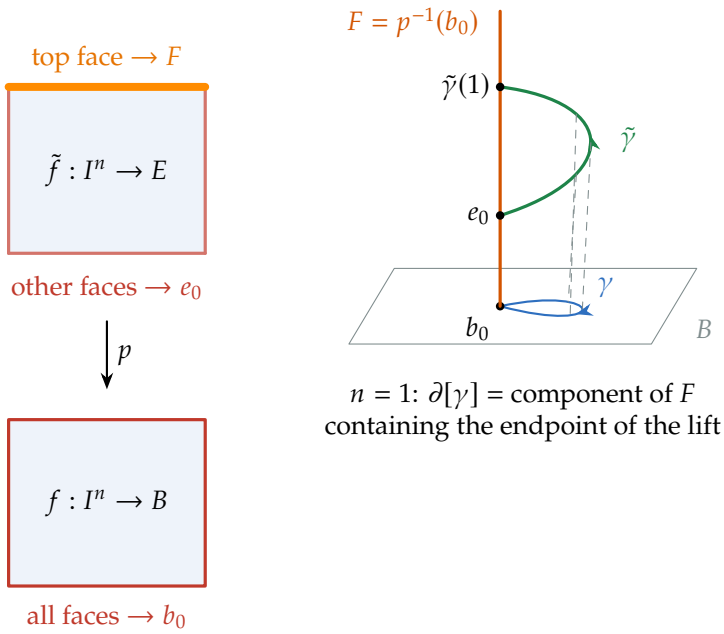


Figure 1.28: The boundary map $\partial : \pi_n(B) \to \pi_{n-1}(F)$. Left: an element of $\pi_n(B)$ (all faces to b_0) lifts to a map of the cube into E with all faces but the top one at e_0. The top face (orange) lies in the fibre over b_0 and defines $\partial[f] \in \pi_{n-1}(F)$. Right: for $n = 1$, a loop in the base lifts to a path that ends somewhere in the fibre. For $\mathbb{R} \to S^1$ the endpoint is $2\pi W$: the boundary map is the winding number.

Worked example 1.43: The circle again: $\mathbb{Z} \to \mathbb{R} \to S^1$

Here $F = 2\pi\mathbb{Z}$ is discrete, so $\pi_k(F) = 0$ for $k \geq 1$ and $\pi_0(F) = \mathbb{Z}$ (one component per point), and $E = \mathbb{R}$ is contractible, so $\pi_k(\mathbb{R}) = 0$. The segments of (1.58) read

$$\underbrace{\pi_n(\mathbb{R})}_{0} \to \pi_n(S^1) \xrightarrow{\partial} \underbrace{\pi_{n-1}(\mathbb{Z})}_{0\ (n\geq 2)} \qquad \Longrightarrow \quad \pi_n(S^1) = 0 \ \ (n \geq 2),$$

$$\underbrace{\pi_1(\mathbb{R})}_{0} \to \pi_1(S^1) \xrightarrow{\partial} \pi_0(\mathbb{Z}) \to \underbrace{\pi_0(\mathbb{R})}_{\text{one point}} \qquad \Longrightarrow \quad \pi_1(S^1) \cong \pi_0(\mathbb{Z}) = \mathbb{Z}.$$

The boundary map is the endpoint of the lift divided by 2π: the winding number (1.32). So (1.46) drops out of the machine in one line.

Worked example 1.44: $\pi_1(\mathrm{SO}(3))$ again: $\mathbb{Z}_2 \to \mathrm{SU}(2) \to \mathrm{SO}(3)$

Fibre $\{\pm\mathbf{1}\}$, total space $\mathrm{SU}(2) \cong S^3$ with $\pi_1 = \pi_2 = 0$:

$$\underbrace{\pi_1(\mathrm{SU}(2))}_{0} \to \pi_1(\mathrm{SO}(3)) \xrightarrow{\partial} \underbrace{\pi_0(\{\pm\mathbf{1}\})}_{\mathbb{Z}_2} \to \underbrace{\pi_0(\mathrm{SU}(2))}_{\text{one point}} \quad \Longrightarrow \quad \pi_1(\mathrm{SO}(3)) \cong \mathbb{Z}_2,$$

and $\partial[\gamma]$ is the endpoint $\pm\mathbf{1}$ of the lift, exactly as in §1.3.4. For $n \geq 2$ the fibre has no homotopy and $\pi_n(\mathrm{SO}(3)) \cong \pi_n(\mathrm{SU}(2)) = \pi_n(S^3)$.

Worked example 1.45: The Hopf fibration: $\pi_2(S^2)$ and $\pi_3(S^2)$

For $S^1 \to S^3 \to S^2$ we know $\pi_k(S^1)$ ((1.46)) and $\pi_k(S^3) = 0$ for $k = 1, 2$ (§1.3.7). Then:

$$\underbrace{\pi_2(S^3)}_{0} \to \pi_2(S^2) \xrightarrow{\partial} \underbrace{\pi_1(S^1)}_{\mathbb{Z}} \to \underbrace{\pi_1(S^3)}_{0} \qquad \Longrightarrow \quad \pi_2(S^2) \cong \mathbb{Z},$$

$$\underbrace{\pi_n(S^1)}_{0} \to \pi_n(S^3) \xrightarrow{h_*} \pi_n(S^2) \xrightarrow{\partial} \underbrace{\pi_{n-1}(S^1)}_{0\ (n\geq 3)} \qquad \Longrightarrow \quad \pi_n(S^2) \cong \pi_n(S^3) \quad (n \geq 3).$$

The first line recovers $\pi_2(S^2) = \mathbb{Z}$ without the degree: the boundary map sends the generator of $\pi_2(S^2)$ (the identity) to the loop of phases in the fibre, which is Example 1.40 read backwards. The second line with $n = 3$ gives

$$\pi_3(S^2) \cong \pi_3(S^3) = \mathbb{Z}, \qquad \text{generated by } h_*[\mathrm{id}_{S^3}] = [h].$$

For larger n it explains why the column S^2 of Figure 1.27 is the column S^3 shifted by one row: $\pi_4(S^2) = \pi_4(S^3) = \mathbb{Z}_2$, $\pi_5(S^2) = \pi_5(S^3) = \mathbb{Z}_2$, $\pi_6(S^2) = \pi_6(S^3) = \mathbb{Z}_{12}$, and so on.

$$\cdots \to \underset{0}{\pi_3(S^1)} \to \underset{\mathbb{Z}}{\pi_3(S^3)} \xrightarrow{h_*} \underset{\mathbb{Z}}{\pi_3(S^2)} \xrightarrow{\partial} \underset{0}{\pi_2(S^1)} \to \underset{0}{\pi_2(S^3)} \to \underset{\mathbb{Z}}{\pi_2(S^2)} \xrightarrow{\partial} \underset{\mathbb{Z}}{\pi_1(S^1)} \to \underset{0}{\pi_1(S^3)}$$

Figure 1.29: The long exact sequence of the Hopf fibration $S^1 \to S^3 \to S^2$, with the value of each group written underneath. Between two zeros (red) the map is an isomorphism, so the unknowns (blue) are forced: $\pi_3(S^2) \cong \pi_3(S^3) = \mathbb{Z}$ and $\pi_2(S^2) \cong \pi_1(S^1) = \mathbb{Z}$.

Worked example 1.46: The unitary and orthogonal groups, by induction

Let $\mathrm{SU}(n)$ act on unit vectors $v \in \mathbb{C}^n$, i.e. on S^{2n-1}. The map

$$p : \mathrm{SU}(n) \to S^{2n-1}, \qquad U \mapsto Ue_1 \ \text{ (the first column of } U)$$

is onto (any unit vector can be completed to a unitary matrix with determinant 1 by Gram–Schmidt). The fibre over e_1 consists of the U with first column e_1; unitarity forces the first row to be e_1^T as well, so

$$p^{-1}(e_1) = \left\{ \begin{pmatrix} 1 & 0 \\ 0 & V \end{pmatrix} : V \in \mathrm{SU}(n-1) \right\} \cong \mathrm{SU}(n-1).$$

So there is a fibration $\mathrm{SU}(n-1) \to \mathrm{SU}(n) \to S^{2n-1}$. (For $n = 2$ the fibre is trivial and p is the homeomorphism $\mathrm{SU}(2) \cong S^3$ of (1.37).) The segment

$$\underbrace{\pi_{k+1}(S^{2n-1})}_{0\ (k+1<2n-1)} \xrightarrow{\partial} \pi_k(\mathrm{SU}(n-1)) \xrightarrow{i_*} \pi_k(\mathrm{SU}(n)) \xrightarrow{p_*} \underbrace{\pi_k(S^{2n-1})}_{0\ (k<2n-1)}$$

shows $\pi_k(\mathrm{SU}(n-1)) \cong \pi_k(\mathrm{SU}(n))$ for $k \leq 2n-3$. Running up from $\mathrm{SU}(2) \cong S^3$:

- $k = 1$: $\pi_1(\mathrm{SU}(n)) = \pi_1(\mathrm{SU}(2)) = 0$ for all n (this was used in Example 1.32);
- $k = 2$: $\pi_2(\mathrm{SU}(n)) = \pi_2(S^3) = 0$;

- $k = 3$ (needs $n \geq 3$): $\pi_3(\mathrm{SU}(n)) = \pi_3(\mathrm{SU}(2)) = \pi_3(S^3) = \mathbb{Z}$. The generator is the SU(2) winding embedded in the top-left corner.

The same argument with $\mathrm{SO}(n)$ acting on $S^{n-1} \subset \mathbb{R}^n$ gives a fibration $\mathrm{SO}(n-1) \to \mathrm{SO}(n) \to S^{n-1}$ and $\pi_k(\mathrm{SO}(n-1)) \cong \pi_k(\mathrm{SO}(n))$ for $k \leq n-3$. With $k = 1$: $\pi_1(\mathrm{SO}(n)) = \pi_1(\mathrm{SO}(3)) = \mathbb{Z}_2$ for all $n \geq 3$, the fact behind spinors in every dimension (Chapter T4). With $\mathrm{U}(n)$: $\pi_1(\mathrm{U}(n)) = \pi_1(\mathrm{U}(1)) = \mathbb{Z}$.

Worked example 1.47: Homogeneous spaces: $\pi_2(G/H) = \pi_1(H)$

If a group G acts transitively on a space M and H is the stabiliser of a point, then $M = G/H$ and $H \to G \to G/H$ is a fibration. If G is a compact simply connected Lie group, then $\pi_1(G) = \pi_2(G) = 0$ (the vanishing of π_2 holds for every compact Lie group, a theorem of É. Cartan), and

$$\underbrace{\pi_2(G)}_{0} \to \pi_2(G/H) \xrightarrow{\partial} \pi_1(H) \to \underbrace{\pi_1(G)}_{0} \quad \Longrightarrow \quad \pi_2(G/H) \cong \pi_1(H).$$

For $G = \mathrm{SU}(2)$ acting on the Bloch sphere by rotations, the stabiliser of the north pole is the diagonal U(1), $G/H = S^2$, and $\pi_2(S^2) = \pi_1(\mathrm{U}(1)) = \mathbb{Z}$: this is the Hopf fibration once more. This formula is the backbone of the classification of magnetic monopoles in grand unified theories (§1.3.13): a symmetry breaking $G \to H$ admits monopoles labelled by $\pi_1(H)$, which is nonzero whenever H contains an unbroken U(1) factor, such as electromagnetism.

1.3.12 Homotopy groups of Lie groups and Bott periodicity

Nash remarks that the homotopy groups of the classical compact Lie groups are "somewhat better known" than those of spheres. The fibrations of Example 1.46 explain why. Including $\mathrm{U}(n) \subset \mathrm{U}(n+1)$ as the top-left block, we found $\pi_k(\mathrm{U}(n)) \cong \pi_k(\mathrm{U}(n+1))$ as soon as $k \leq 2n-1$. So for fixed k the groups $\pi_k(\mathrm{U}(n))$ stop changing once n is large, and we may define the **stable** groups

$$\pi_k(\mathrm{U}) := \pi_k(\mathrm{U}(n)) \ \ (2n > k), \qquad \pi_k(\mathrm{O}) := \pi_k(\mathrm{O}(n)) \ \ (n \geq k+2),$$
$$\pi_k(\mathrm{Sp}) := \pi_k(\mathrm{Sp}(n)) \ \ (4n \geq k-1).$$

The ranges come from the fibrations $\mathrm{O}(n) \to \mathrm{O}(n+1) \to S^n$ and $\mathrm{Sp}(n) \to \mathrm{Sp}(n+1) \to S^{4n+3}$ in the same way.

Key idea: Bott periodicity (1959)

The stable homotopy groups of the classical groups are periodic:

$$\pi_{k+2}(\mathrm{U}) \cong \pi_k(\mathrm{U}), \qquad \pi_{k+8}(\mathrm{O}) \cong \pi_k(\mathrm{O}), \qquad \pi_{k+4}(\mathrm{Sp}) \cong \pi_k(\mathrm{O}),$$

with the values

$k \bmod 8$	0	1	2	3	4	5	6	7
$\pi_k(\mathrm{U})$	0	$\mathbb{Z}$	0	$\mathbb{Z}$	0	$\mathbb{Z}$	0	$\mathbb{Z}$
$\pi_k(\mathrm{O})$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	0	$\mathbb{Z}$	0	0	0	$\mathbb{Z}$
$\pi_k(\mathrm{Sp})$	0	0	0	$\mathbb{Z}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	0	$\mathbb{Z}$

We can check the first few entries against what we already computed: $\pi_1(\mathrm{U}) = \pi_1(\mathrm{U}(1)) = \mathbb{Z}$ (winding of the determinant), $\pi_0(\mathrm{O}) = \mathbb{Z}_2$ (the two components $\det = \pm 1$), $\pi_1(\mathrm{O}) = \pi_1(\mathrm{SO}(3)) = \mathbb{Z}_2$, $\pi_3(\mathrm{U}) = \pi_3(\mathrm{SU}(n)) = \mathbb{Z}$ and $\pi_3(\mathrm{Sp}) = \pi_3(\mathrm{Sp}(1)) = \pi_3(\mathrm{SU}(2)) = \mathbb{Z}$ (since Sp(1), the unit quaternions, is SU(2)). The pattern is astonishing: an infinite list of hard computations reduced to period two and period eight. Nash returns to it in Chapter 3, where it becomes the periodicity of K-theory, and the period eight is the same period eight as that of the real Clifford algebras (Chapter T4).

Outside the stable range the groups are not periodic, but the first few are known for every group we need.

G	π_1	π_2	π_3	π_4	π_5	π_6
U(1)	$\mathbb{Z}$	0	0	0	0	0
SU(2) $\cong S^3$	0	0	$\mathbb{Z}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_{12}$
SO(3)	$\mathbb{Z}_2$	0	$\mathbb{Z}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_{12}$
SU(3)	0	0	$\mathbb{Z}$	0	$\mathbb{Z}$	$\mathbb{Z}_6$
SU(n), $n \geq 4$	0	0	$\mathbb{Z}$	0	$\mathbb{Z}$	0

Physical picture: Which homotopy group does which physics

- $\pi_2(G) = 0$ for every compact Lie group: there are no "monopoles" valued in a group; monopoles need a coset G/H (Example 1.47).
- $\pi_3(G) = \mathbb{Z}$ for every compact simple Lie group: Yang–Mills instantons exist for every simple gauge group, labelled by one integer (§1.3.14).
- $\pi_4(\mathrm{SU}(2)) = \mathbb{Z}_2$: Witten's global anomaly (Example 1.41).
- $\pi_5(\mathrm{SU}(3)) = \mathbb{Z}$: the coefficient of the Wess–Zumino–Witten term of the chiral Lagrangian must be an integer, which turns out to be the number of colours N_c.
- $\pi_{2n+1}(\mathrm{U}) = \mathbb{Z}$ in general: the non-abelian anomalies in $2n$ dimensions (Chapter 10).

1.3.13 Physics I: topological defects

We now have enough homotopy theory to answer a question that runs through condensed matter, particle physics and cosmology: *which stable "lumps" can a field form?* A domain wall in a magnet, a vortex in a superfluid, a flux tube in a superconductor, a magnetic monopole in a grand unified theory, a skyrmion in nuclear physics: each is a configuration that cannot be removed by any smooth, finite-energy change of the field. In every case the reason is that a certain map of a sphere into a certain space is not null-homotopic. Nash does not discuss this; it is the most direct physical use of §3, and we give it in full.

The space of vacua

Let ϕ be a field (real, complex, vector or matrix valued) with a potential energy $V(\phi) \geq 0$. The set of field values that minimise V is the **vacuum manifold** (or **order-parameter space**)

$$\mathcal{M}_0 := \{\phi : V(\phi) = \min V\}.$$

Finite energy forces the field to lie in $\mathcal{M}_0$ wherever it is far from the "core" of a defect, because a region of volume $\mathcal{V}$ outside $\mathcal{M}_0$ costs energy of order $\mathcal{V} \cdot (V - \min V)$. When a symmetry group G acts on the fields, leaves V invariant and acts transitively on the minima, and $H \subset G$ is

the stabiliser (the **unbroken subgroup**) of one minimum ϕ_0, then every minimum is $g\phi_0$ for some g, and $g\phi_0 = g'\phi_0$ exactly when $g^{-1}g' \in H$. So

$$\mathcal{M}_0 \cong G/H.$$

Dictionary: vacuum manifolds

system	order parameter	$\mathcal{M}_0$
Ising magnet, ϕ^4 kink	real ϕ, $V = \frac{\lambda}{4}(\phi^2 - v^2)^2$	$\{+v, -v\} = S^0$
superfluid, superconductor, cosmic string	complex ψ, $V = \frac{\lambda}{4}(\lvert\psi\rvert^2 - v^2)^2$	$\mathrm{U}(1)/1 = S^1$
Heisenberg ferromagnet	unit magnetisation $\vec{n}$	$\mathrm{SO}(3)/\mathrm{SO}(2) = S^2$
nematic liquid crystal	director $\pm\vec{n}$	$S^2/\mathbb{Z}_2 = \mathbb{RP}^2$
Georgi–Glashow SO(3) model	real triplet $\vec{\phi}$, $\lvert\vec{\phi}\rvert = v$	$\mathrm{SO}(3)/\mathrm{SO}(2) = S^2$
chiral pions	$U \in \mathrm{SU}(2)$	$\frac{\mathrm{SU}(2)_L \times \mathrm{SU}(2)_R}{\mathrm{SU}(2)_V} = S^3$
electroweak Higgs	doublet Φ, $\Phi^\dagger\Phi = v^2/2$	$\frac{\mathrm{SU}(2)\times\mathrm{U}(1)}{\mathrm{U}(1)} = S^3$

The codimension rule

Suppose the field is allowed to leave $\mathcal{M}_0$ only in a small region of space, the core. What matters is the **codimension** d of the core: the number of directions in which one can move away from it. In D-dimensional space, a wall has $d = 1$, a line has $d = 2$, a point has $d = 3$ (for $D = 3$). The set of points at a fixed small distance from the core, in the d transverse directions, is a sphere S^{d-1}: two points for a wall, a circle for a line, a two-sphere for a point (Figure 1.30). On that sphere the field lies in $\mathcal{M}_0$, so it defines a map

$$\phi\big|_{S^{d-1}} : S^{d-1} \longrightarrow \mathcal{M}_0.$$

Working definition 1.25 (topological charge of a defect)

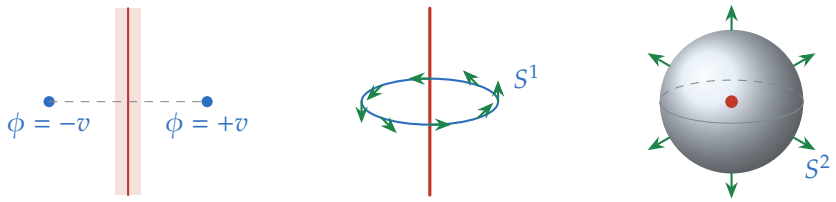
Surround the core by a sphere S^{d-1} in the transverse directions. The **topological charge** of the defect is the homotopy class [all maps from the sphere into the vacuum manifold $\mathcal{M}_0$ that can be deformed into this one]

$$[\phi|_{S^{d-1}}] \in [S^{d-1}, \mathcal{M}_0] \;=\; \pi_{d-1}(\mathcal{M}_0)\big/\pi_1(\mathcal{M}_0).$$

The quotient by π_1 is Nash's (1.31), (1.39): space has no preferred base point, so the physical charge is a *free* homotopy class. When $\mathcal{M}_0$ is simply connected [every loop in $\mathcal{M}_0$ can be shrunk to a point, $\pi_1(\mathcal{M}_0) = 0$] the charge is simply an element of $\pi_{d-1}(\mathcal{M}_0)$. *Operational test:* compute the relevant winding number ((1.33) for $d = 2$, the degree integral (1.51) for $d = 3$) of the field on a sphere far from the core. If it is nonzero, the defect is topologically stable.

Key idea: Why the charge is conserved

Any smooth time evolution of the field that keeps the energy finite keeps the field in $\mathcal{M}_0$ on the surrounding sphere (as long as the core does not cross it). So the map $\phi_t|_{S^{d-1}}$, $t \in [0, 1]$, is a homotopy, and its class cannot change. *Topological charges are conserved because homotopy classes are discrete and time evolution is continuous.* No equation of motion is needed. The only ways to change the charge are discontinuous: pay an infinite energy, or bring in another defect through the sphere.



wall, $d = 1$
S^0 = two points
charge in $\pi_0(\mathcal{M}_0)$

line, $d = 2$
loop around it
charge in $\pi_1(\mathcal{M}_0)$

point, $d = 3$
sphere around it
charge in $\pi_2(\mathcal{M}_0)$

Figure 1.30: The codimension rule. The core of a defect (red) is surrounded, in the directions transverse to it, by a sphere S^{d-1} (blue): two points on either side of a wall, a loop around a line, a two-sphere around a point. On that sphere the field sits in the vacuum manifold, and its homotopy class in $\pi_{d-1}(\mathcal{M}_0)$ is the charge. The green arrows show the field on the sphere: winding once around a vortex, pointing outward on a hedgehog.

Charges also add, in the sense of the group law. Two vortices of windings n_1 and n_2, enclosed together by a big loop, have total winding $n_1 + n_2$: deform the big loop into the product of two small loops, one around each (Figure 1.14). For point defects the same squeezing of a big sphere into two small spheres touching at a point is the cube-picture sum (1.43). A defect and an anti-defect can annihilate; two defects of charge +1 cannot.

Dictionary: defects and homotopy groups

defect	codim. d	classified by	examples
domain wall	1	$\pi_0(\mathcal{M}_0)$	Ising walls, ϕ^4 kink, axion walls
vortex, string	2	$\pi_1(\mathcal{M}_0)$	superfluid vortex, Abrikosov flux tube, cosmic string
monopole, hedgehog	3	$\pi_2(\mathcal{M}_0)$	't Hooft–Polyakov monopole, hedgehog in a magnet
texture, skyrmion	(whole space)	$\pi_D(\mathcal{M}_0)$	skyrmion (π_3), baby skyrmion (π_2 in $D = 2$)
instanton	(all of spacetime)	$\pi_3(G)$	Yang–Mills instanton (§1.3.14)

The last two rows are a different kind of object: **textures** have no core at all. The field stays in $\mathcal{M}_0$ everywhere, and tends to one fixed value ϕ_∞ at spatial infinity. Adding the point at infinity turns $\mathbb{R}^D$ into S^D (stereographic projection, Figure 1.7), so the whole field is a map $S^D \to \mathcal{M}_0$ and has a class in $\pi_D(\mathcal{M}_0)$. We now work through one example of each type.

Worked example 1.48: The ϕ^4 kink: $\pi_0(S^0)$ and a Bogomolny bound

Take a real scalar in one space dimension with energy

$$E[\phi] = \int_{-\infty}^{\infty} \Big(\tfrac{1}{2}\phi'^2 + V(\phi) \Big) \mathrm{d}x, \qquad V(\phi) = \tfrac{\lambda}{4}(\phi^2 - v^2)^2.$$

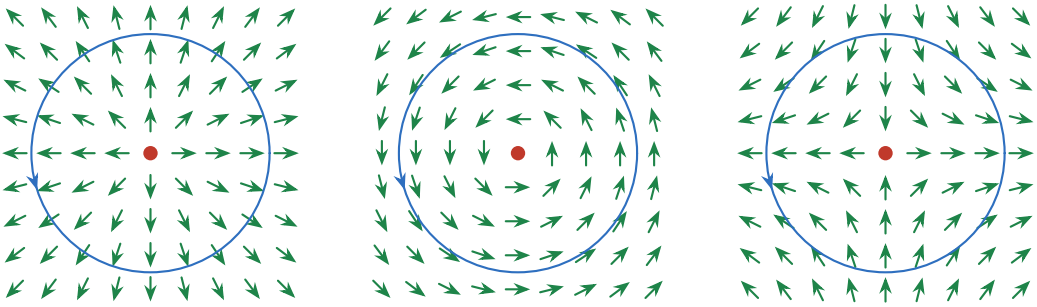
Here $\mathcal{M}_0 = \{\pm v\} = S^0$, and $\pi_0(\mathcal{M}_0)$ is the set of its two points (Example 1.15). Finite energy forces $\phi(\pm\infty) \in \{\pm v\}$, so there are four sectors labelled by $(\phi(-\infty), \phi(+\infty))$. The sectors with $\phi(-\infty) \neq \phi(+\infty)$ contain a wall. To find the lowest energy in the sector $(-v, +v)$, set $W(\phi) := \sqrt{\lambda/2}\,(v^2\phi - \phi^3/3)$, so that $W'(\phi) = \sqrt{\lambda/2}\,(v^2 - \phi^2)$ and $\frac{1}{2}W'^2 = V$. Then

$$E[\phi] \overset{(1)}{=} \int \Big(\tfrac{1}{2}\phi'^2 + \tfrac{1}{2}W'(\phi)^2 \Big) \mathrm{d}x$$

$$\overset{(2)}{=} \int \tfrac{1}{2}\big(\phi' - W'(\phi)\big)^2 \mathrm{d}x + \int W'(\phi)\,\phi'\,\mathrm{d}x$$
$$\overset{(3)}{=} \int \tfrac{1}{2}\big(\phi' - W'(\phi)\big)^2 \mathrm{d}x + W(v) - W(-v)$$
$$\overset{(4)}{\geq} W(v) - W(-v) = \tfrac{4}{3}\sqrt{\tfrac{\lambda}{2}}\,v^3.$$

(1) $V = \frac{1}{2}W'^2$ by construction.
(2) Complete the square: $\frac{1}{2}a^2 + \frac{1}{2}b^2 = \frac{1}{2}(a-b)^2 + ab$.
(3) Chain rule, $W'(\phi)\phi' = \mathrm{d}W(\phi(x))/\mathrm{d}x$, and the fundamental theorem of calculus with $\phi(\pm\infty) = \pm v$. The result depends only on the boundary values: it is topological.
(4) A square is non-negative; $W(v) - W(-v) = 2\sqrt{\lambda/2}\,(v^3 - v^3/3)$.

Equality holds when $\phi' = W'(\phi) = \sqrt{\lambda/2}\,(v^2 - \phi^2)$, a first-order equation. Its solution is the **kink** $\phi(x) = v\tanh(mx)$ with $m = v\sqrt{\lambda/2}$: indeed $\phi' = vm\,\mathrm{sech}^2(mx)$ and $\sqrt{\lambda/2}\,(v^2 - v^2\tanh^2 mx) = \sqrt{\lambda/2}\,v^2\,\mathrm{sech}^2(mx) = vm\,\mathrm{sech}^2(mx)$. The kink cannot decay into the vacuum $\phi \equiv v$: that would change $\phi(-\infty)$ from $-v$ to $+v$ through values outside $\mathcal{M}_0$, at infinite cost. Its energy is fixed by the boundary data. This is the pattern of every **Bogomolny bound**: energy $\geq$ constant $\times$|topological charge|, saturated by a first-order equation. We met it for the baby skyrmion (the Belavin–Polyakov bound of §1.3.8), and will meet it again for instantons.



$\psi \propto e^{i\theta}$: $W = +1$ (hedgehog in the plane)
$\psi \propto ie^{i\theta}$: $W = +1$ (circulating flow)
$\psi \propto e^{-i\theta}$: $W = -1$ (anti-vortex)

Figure 1.31: The phase of ψ drawn as an arrow at angle $\arg\psi$. Walk around the blue loop and count how many times the arrow turns: that is the winding number $W \in \pi_1(S^1) = \mathbb{Z}$. The left and middle fields differ by a constant rotation of every arrow, a homotopy, and have the same $W = +1$; the right one has $W = -1$ (the arrow turns backwards as we walk forwards). In each case the phase is undefined at the red core, where $|\psi|$ must vanish.

Worked example 1.49: Vortices and quantised flux: $\pi_1(S^1) = \mathbb{Z}$

Global vortex. A complex order parameter in the plane, far from the core, has the form $\psi = v\,e^{i\chi(x,y)}$, with χ defined modulo 2π. On a circle C_R of radius R around the core, ψ/v is a loop in S^1, with winding number $W = \frac{1}{2\pi}\oint_{C_R}\mathrm{d}\chi$ ((1.33)). The simplest representative is $\chi = W\theta$ (Figure 1.31). Its gradient energy between the core radius a and R is

$$\int_{a<r<R} |\nabla\psi|^2\,\mathrm{d}^2x \overset{(1)}{=} \int_a^R\int_0^{2\pi} v^2\left|\frac{1}{r}\frac{\partial\chi}{\partial\theta}\right|^2 r\,\mathrm{d}\theta\,\mathrm{d}r \overset{(2)}{=} 2\pi W^2v^2\int_a^R\frac{\mathrm{d}r}{r} \overset{(3)}{=} 2\pi W^2v^2\ln\frac{R}{a}.$$

(1) $|\nabla\psi|^2 = v^2|\nabla\chi|^2$ and, for $\chi = W\theta$, only the angular component $\frac{1}{r}\partial_\theta\chi$ of the gradient is nonzero.

(2) $\partial_\theta \chi = W$; the θ integral gives 2π.
(3) Elementary integral.

The energy grows logarithmically with the size of the system. The vortex cannot unwind (that would change W), but it costs a lot: this is why superfluid vortices appear only when the container rotates, which makes the winding energetically favourable, and why the $W = 2$ vortex, with energy $\propto W^2 = 4$, prefers to split into two $W = 1$ vortices ($1 + 1 = 2 < 4$).

Gauged vortex. In a superconductor ψ has charge q and couples to $\vec{A}$ through $\vec{D}\psi = (\nabla - iq\vec{A})\psi$ (physicists' convention, $\hbar = 1$). Finite energy now needs $\vec{D}\psi \to 0$ at large r, not $\nabla\psi \to 0$. For $\psi \to v e^{iW\theta}$ this fixes the gauge field at large r, and the magnetic flux through the plane follows:

$$\Phi \overset{(1)}{=} \oint_{C_R} \vec{A} \cdot \mathrm{d}\vec{l} \overset{(2)}{=} \frac{1}{q} \oint_{C_R} \nabla\chi \cdot \mathrm{d}\vec{l} \overset{(3)}{=} \frac{2\pi W}{q}.$$

(1) Stokes' theorem: $\int \vec{B} \cdot \mathrm{d}\vec{S} = \oint \vec{A} \cdot \mathrm{d}\vec{l}$ for C_R large enough to enclose all the flux.
(2) $\vec{D}\psi = iv\, e^{i\chi}(\nabla\chi - q\vec{A}) = 0$ at large r, so $\vec{A} = \frac{1}{q}\nabla\chi$ there.
(3) $\oint \nabla\chi \cdot \mathrm{d}\vec{l} = \oint \mathrm{d}\chi = 2\pi W$.

So each Abrikosov flux tube carries an integer number of flux quanta $\Phi_0 = 2\pi\hbar/q = h/2e$ (Cooper pairs have $q = 2e$), the same quantisation we found for a ring (§1.3.2), now for a tube of flux in the bulk. The gradient energy at large r is gone, since $\vec{D}\psi \to 0$ there, so the gauged vortex has finite energy per unit length. In a relativistic theory the same object is a **cosmic string** (Nielsen–Olesen vortex), with $\mathcal{M}_0 = S^1$ and charge in $\pi_1(S^1) = \mathbb{Z}$.

Watch out: Topology gives stability, not existence

A nonzero homotopy class says that the field *cannot be deformed to the vacuum*. It does not say that a static solution of minimum energy exists in that class. The energy may be lowered forever by shrinking or spreading the configuration. Derrick's theorem makes this precise: for a scalar field with only two-derivative terms in $D \geq 3$ space dimensions, rescaling $\phi(x) \mapsto \phi(\lambda x)$ always lowers the energy of a static configuration, so no static texture exists. Stabilisation requires something extra: gauge fields (vortices, monopoles), four-derivative terms (the Skyrme term), or $D = 2$ where the gradient energy is scale invariant (baby skyrmions).

Worked example 1.50: The hedgehog: $\pi_2(S^2) = \mathbb{Z}$

In the Georgi–Glashow model a real triplet $\vec{\phi} = (\phi^1, \phi^2, \phi^3)$ has $V = \frac{\lambda}{4}(\vec{\phi}\cdot\vec{\phi} - v^2)^2$ and $\mathcal{M}_0$ is the sphere $|\vec{\phi}| = v$. Point defects in $\mathbb{R}^3$ are labelled by $\pi_2(S^2) = \mathbb{Z}$ (S^2 is simply connected, so free and based classes agree, Example 1.34). The **hedgehog**

$$\vec{\phi}(\vec{x}) = v\, h(r)\, \hat{\vec{x}}, \qquad h(0) = 0, \quad h(\infty) = 1,$$

points radially outward (Figure 1.30, right). On a large sphere of radius R with coordinates (θ, φ) its direction is $\hat{\vec{\phi}} = \hat{\vec{x}} = (\sin\theta\cos\varphi, \sin\theta\sin\varphi, \cos\theta)$, which is the map $\vec{n}_1$ of Example 1.37. So

$$N \overset{(1)}{=} \frac{1}{4\pi} \int \hat{\vec{\phi}} \cdot \left(\partial_\theta \hat{\vec{\phi}} \times \partial_\varphi \hat{\vec{\phi}}\right) \mathrm{d}\theta\, \mathrm{d}\varphi \overset{(2)}{=} \frac{1}{4\pi} \int_0^{2\pi} \int_0^{\pi} \sin\theta\, \mathrm{d}\theta\, \mathrm{d}\varphi \overset{(3)}{=} 1.$$

(1) The degree integral (1.51) on the sphere $r = R$.

(2) The computation of Example 1.37 with $k = 1$: $\partial_\theta \hat{\vec{x}} \times \partial_\varphi \hat{\vec{x}} = \sin\theta\, \hat{\vec{x}}$.
(3) $\int_0^\pi \sin\theta \, \mathrm{d}\theta = 2$ and $\int_0^{2\pi} \mathrm{d}\varphi = 2\pi$.

Since $N \neq 0$ the field cannot be combed flat: $\vec{\phi}$ must vanish somewhere inside every sphere, which is the core. (If $\vec{\phi}$ had no zero inside the ball, $\hat{\vec{\phi}}$ would be defined on the whole ball, and shrinking the sphere to a point would be a null-homotopy.) With the SO(3) gauge field included, the hedgehog is the **'t Hooft–Polyakov monopole**. Far from the core the unbroken gauge group is the U(1) of rotations about $\hat{\vec{\phi}}$, which plays the role of electromagnetism, and its magnetic flux through the large sphere is

$$\oint \vec{B} \cdot \mathrm{d}\vec{S} = \frac{4\pi}{e} N,$$

with e the gauge coupling: the magnetic charge is the degree. We do not derive this (it needs 't Hooft's gauge-invariant field strength); Chapter 8 returns to monopoles.

The monopole charge can also be read in a second way, which prepares for fibre bundles (Nash §5, next part) and makes contact with the exact sequence of Example 1.47. For SU(2) → U(1) that sequence says $\pi_2(S^2) \cong \pi_1(\mathrm{U}(1))$: the degree of the Higgs field on the sphere equals the winding of a U(1) gauge transformation around the equator. The Dirac monopole shows that winding directly.

Worked example 1.51: The Dirac monopole: π_2 of the sphere becomes $\pi_1(\mathrm{U}(1))$

A point magnetic charge g has $\vec{B} = \frac{g}{4\pi}\frac{\hat{\vec{r}}}{r^2}$, i.e. the 2-form $F = \frac{g}{4\pi}\sin\theta\,\mathrm{d}\theta \wedge \mathrm{d}\varphi$ on each sphere around the origin. It is closed but not exact on S^2 (its integral is $g \neq 0$; see the exemplar argument in Chapter T2). So no single vector potential works. Two do:

$$A_N = \frac{g}{4\pi}(1-\cos\theta)\,\mathrm{d}\varphi \ \ (\theta < \pi), \qquad A_S = -\frac{g}{4\pi}(1+\cos\theta)\,\mathrm{d}\varphi \ \ (\theta > 0).$$

A_N is smooth at the north pole because its coefficient vanishes there ($1-\cos\theta \to 0$ as $\theta \to 0$, where $\mathrm{d}\varphi$ is singular); likewise A_S at the south pole. Both have $\mathrm{d}A = \frac{g}{4\pi}\sin\theta\,\mathrm{d}\theta\wedge\mathrm{d}\varphi = F$. On the overlap they differ by

$$A_N - A_S \overset{(1)}{=} \frac{g}{4\pi}\big[(1-\cos\theta)+(1+\cos\theta)\big]\mathrm{d}\varphi \overset{(2)}{=} \frac{g}{2\pi}\,\mathrm{d}\varphi \overset{(3)}{=} \frac{1}{q}\,\mathrm{d}\chi, \qquad \chi \coloneqq \frac{qg}{2\pi}\varphi.$$

(1) Subtract the two formulas.
(2) The $\cos\theta$ terms cancel.
(3) Define χ; this has the form of a gauge transformation $\vec{A} \mapsto \vec{A} + \frac{1}{q}\nabla\chi$, under which a charge-$q$ wavefunction changes as $\psi \mapsto e^{i\chi}\psi$.

A charged particle's wavefunctions in the two patches must be related by $\psi_N = e^{i\chi}\psi_S$. For this to make sense on the whole overlap the **transition function** $e^{i\chi} = e^{iqg\varphi/2\pi}$ must be single valued around the equator, which is **Dirac's quantisation condition**

$$\frac{qg}{2\pi} = N \in \mathbb{Z}.$$

Now look at what N is: the transition function is a loop $\varphi \mapsto e^{iN\varphi}$ in U(1), and N is its winding number, an element of $\pi_1(\mathrm{U}(1)) = \mathbb{Z}$. And the total flux is (by Stokes on each

hemisphere, whose common boundary is the equator, traversed in opposite directions)

$$\int_{S^2} F \overset{(1)}{=} \int_{\text{north}} \mathrm{d}A_N + \int_{\text{south}} \mathrm{d}A_S \overset{(2)}{=} \oint_{\text{eq}} A_N - \oint_{\text{eq}} A_S \overset{(3)}{=} \oint_{\text{eq}} \frac{1}{q}\mathrm{d}\chi \overset{(4)}{=} \frac{2\pi N}{q}.$$

(1) Split the sphere into hemispheres and use the potential that is smooth on each.
(2) Stokes. The equator is the boundary of the northern hemisphere with the direction of increasing φ and of the southern hemisphere with the opposite direction.
(3) The difference of the potentials, computed above.
(4) χ increases by $2\pi N$ once around the equator.

Magnetic charge = winding of the transition function. The degree of a map $S^2 \to S^2$ has become the winding of a map $S^1 \to \mathrm{U}(1)$ on the equator. This is our first glimpse of how bundles over S^2 are classified by π_1 of their structure group: the "clutching" construction of Nash §5.

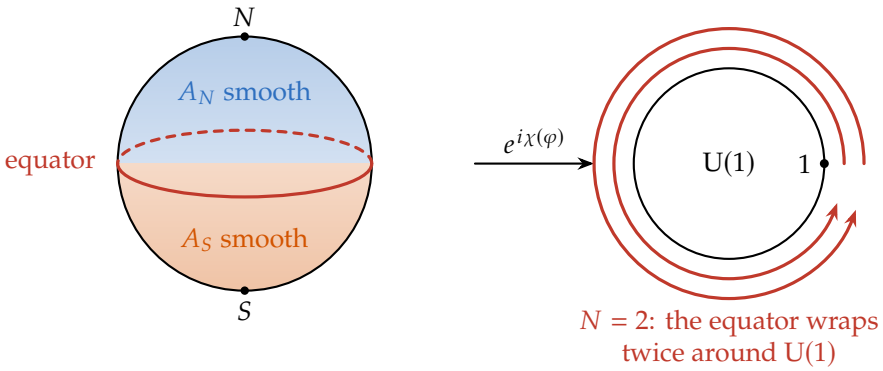


Figure 1.32: The Dirac monopole. Each hemisphere carries its own smooth vector potential. On the equator (red) they are glued by a gauge transformation $e^{i\chi}$, $\chi = N\varphi$, a loop in U(1). Its winding number $N \in \pi_1(\mathrm{U}(1)) = \mathbb{Z}$ is the magnetic charge in units of $2\pi/q$. The picture on the right shows $N = 2$.

Worked example 1.52: The skyrmion: a texture in $\pi_3(\mathrm{SU}(2)) = \mathbb{Z}$

In the low-energy theory of pions the field is a matrix $U(\vec{x}) \in \mathrm{SU}(2)$, with $U \to \mathbf{1}$ at spatial infinity. So U is a map $\mathbb{R}^3 \cup \{\infty\} \cong S^3 \to \mathrm{SU}(2) \cong S^3$, with a class in $\pi_3(\mathrm{SU}(2)) = \mathbb{Z}$. Skyrme's insight (1961) was that this integer is the *baryon number*: nucleons are textures of the pion field. The integer is computed by the winding integral that we set up in the next subsection, (1.60); here we evaluate it for the **hedgehog skyrmion**

$$U(\vec{x}) = \cos f(r)\,\mathbf{1} + i \sin f(r)\,\hat{\vec{x}} \cdot \vec{\sigma}, \qquad f(0) = \pi, \quad f(\infty) = 0.$$

Under the identification $\mathrm{SU}(2) \cong S^3$ of (1.37), $U = x^0\mathbf{1} + i\,x^a\sigma_a$ with $(x^0, \vec{x}') = (\cos f,\ \sin f\,\hat{\vec{x}})$. In hyperspherical coordinates (χ, θ, φ) on S^3, $x^0 = \cos\chi$, $\vec{x}' = \sin\chi\,(\sin\theta\cos\varphi, \sin\theta\sin\varphi, \cos\theta)$, the map U is simply

$$(r, \theta, \varphi) \longmapsto (\chi, \theta, \varphi) = (f(r), \theta, \varphi).$$

The volume form of the unit S^3 in these coordinates is $\sin^2\chi\,\sin\theta\,\mathrm{d}\chi \wedge \mathrm{d}\theta \wedge \mathrm{d}\varphi$ (the round metric is $\mathrm{d}\chi^2 + \sin^2\chi\,(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\varphi^2)$), and its total is $2\pi^2$. The degree of U is the

pulled-back volume divided by the total volume:

$$\deg U \overset{(1)}{=} \frac{1}{2\pi^2}\int_{\mathbb{R}^3} \sin^2 f(r)\, f'(r)\, \sin\theta \, \mathrm{d}r\, \mathrm{d}\theta\, \mathrm{d}\varphi$$
$$\overset{(2)}{=} \frac{4\pi}{2\pi^2}\int_0^\infty \sin^2 f \,\frac{\mathrm{d}f}{\mathrm{d}r}\,\mathrm{d}r \overset{(3)}{=} \frac{2}{\pi}\int_{f(0)}^{f(\infty)} \sin^2 f \,\mathrm{d}f$$
$$\overset{(4)}{=} \frac{2}{\pi}\left[\frac{f}{2} - \frac{\sin 2f}{4}\right]_\pi^0 \overset{(5)}{=} -1.$$

(1) (1.50) with $\omega = \mathrm{vol}_{S^3}/2\pi^2$: substitute $\chi = f(r)$, so $\mathrm{d}\chi = f'\,\mathrm{d}r$, and keep θ, φ.
(2) The angular integral is 4π.
(3) Change variables from r to f; the limits are $f(0) = \pi$ and $f(\infty) = 0$.
(4) $\sin^2 f = \frac{1}{2}(1 - \cos 2f)$.
(5) $\frac{2}{\pi}\left(0 - \frac{\pi}{2}\right) = -1$.

The answer depends only on the boundary values $f(0) = \pi$, $f(\infty) = 0$, not on the profile, as a homotopy invariant must. Geometrically: as r runs from 0 to ∞, $\chi = f(r)$ runs from the south pole $-\mathbf{1}$ of S^3 to the north pole $\mathbf{1}$, and each sphere $r = \text{const}$ is mapped to the latitude sphere $\chi = f(r)$ of S^3: U covers S^3 exactly once. The sign is an orientation convention: here $\mathbb{R}^3$ has its standard orientation and S^3 the one fixed in (1.60) below. With these conventions the skyrmion has $\deg U = -1$, and the Skyrme-model literature defines the baryon number as $B = -\deg U$ so that it is $+1$. (Equivalently, $f(0) = -\pi$ gives $\deg U = +1$.) Derrick's theorem (the pitfall above) forbids a static skyrmion in the two-derivative pion Lagrangian; Skyrme added a four-derivative term that stabilises it.

Recap: defects

- Finite energy puts the field in the vacuum manifold $\mathcal{M}_0 = G/H$ away from the core.
- A defect of codimension d is surrounded by S^{d-1}; its charge is the (free) class in $\pi_{d-1}(\mathcal{M}_0)$. Walls: π_0; vortices: π_1; monopoles: π_2. Textures with a fixed value at infinity: $\pi_D(\mathcal{M}_0)$.
- Charges are conserved because homotopy classes are discrete; they add by the group law; topology gives stability, not existence.
- Examples: kink ($\pi_0(S^0)$, Bogomolny bound), vortex and flux quantum $2\pi/q$ ($\pi_1(S^1)$), hedgehog and Dirac monopole ($\pi_2(S^2) = \pi_1(\mathrm{U}(1))$), skyrmion ($\pi_3(S^3)$).

1.3.14 Physics II: instantons and $\pi_3(\mathrm{SU}(2)) = \mathbb{Z}$

The Euclidean path integral of a gauge theory is a sum over gauge fields on $\mathbb{R}^4$ weighted by e^{-S}. Semiclassically it is dominated by configurations of finite action. The question of this subsection is: *do the finite-action gauge fields on $\mathbb{R}^4$ fall into disconnected classes?* If so, the path integral splits into a sum over sectors, and the sectors cannot be reached from each other by perturbation theory. The answer is yes, the classes are labelled by $\pi_3(\mathrm{SU}(2)) = \mathbb{Z}$, and the label is the **instanton number**. This is the physics that the whole of Nash's book is building towards (Chapter 8), and we can already see its topological core.

Dictionary: conventions for gauge fields in this part

We use Nash's (mathematicians') convention. The gauge field $A = A_\mu \mathrm{d}x^\mu$ is a 1-form with values in anti-Hermitian 2×2 matrices ($\mathfrak{su}(2)$),

$$F = \mathrm{d}A + A\wedge A, \qquad A \mapsto g^{-1}Ag + g^{-1}\mathrm{d}g, \qquad F \mapsto g^{-1}Fg.$$

The physicists' Hermitian fields are related by $A = -ig\,A^a_{\text{phys}}T^a$, $F = -ig\,F^a_{\text{phys}}T^a$, with $T^a = \sigma^a/2$ and $\operatorname{tr}(T^aT^b) = \frac{1}{2}\delta^{ab}$, where g is the coupling constant (context distinguishes it from a gauge transformation $g(x)$). Euclidean $\mathbb{R}^4$ has coordinates (x^0,x^1,x^2,x^3), orientation $\mathrm{d}x^0\wedge\mathrm{d}x^1\wedge\mathrm{d}x^2\wedge\mathrm{d}x^3$, and the three-sphere S^3_R of radius R carries the boundary orientation (outward normal first). See Chapter T3 for connections and curvature.

Finite action forces a map $S^3 \to \mathrm{SU}(2)$

The Yang–Mills action is

$$S = \frac{1}{4}\int F^a_{\mu\nu}F^a_{\mu\nu}\,\mathrm{d}^4x = -\frac{1}{2g^2}\int \operatorname{tr}(F_{\mu\nu}F_{\mu\nu})\,\mathrm{d}^4x.$$

For it to be finite, F must fall off faster than $1/r^2$ as $r=|x|\to\infty$. So on a large sphere S^3_R the field is *flat*, $F\approx 0$. A flat connection on a simply connected region is *pure gauge*: parallel transport is path independent (its holonomy around small loops is $1+O(F)$, and every loop can be shrunk), so transporting from a fixed point defines a gauge g in which $A=0$, i.e. $A = g^{-1}\mathrm{d}g$ in the original gauge. On the region $r>R$, which is homotopic to S^3 and hence simply connected (Example 1.27), we therefore have

$$A \xrightarrow{\;r\to\infty\;} g^{-1}\mathrm{d}g, \qquad g : S^3_\infty \to \mathrm{SU}(2). \tag{1.59}$$

The finite-action field thus defines a class $[g]\in[S^3,\mathrm{SU}(2)]$. Since $\mathrm{SU}(2)\cong S^3$ is simply connected, free and based classes coincide (Example 1.34), and

$$[S^3,\mathrm{SU}(2)] = \pi_3(\mathrm{SU}(2)) = \pi_3(S^3) = \mathbb{Z}$$

by the degree (§1.3.8). Figure 1.33 shows the picture.

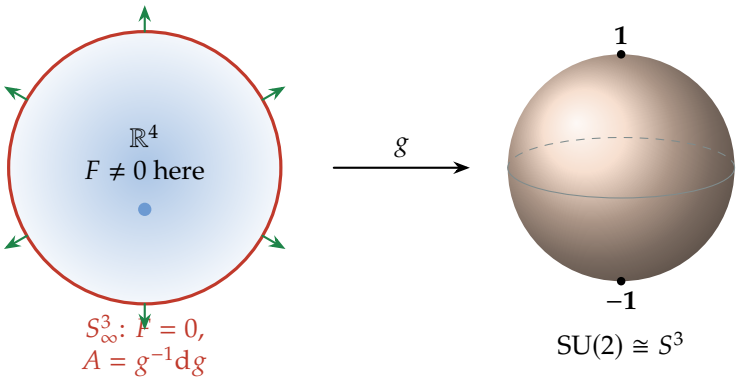


Figure 1.33: Why instantons are classified by π_3. Finite action forces the field strength to vanish on the three-sphere at infinity (red), so there the gauge field is a pure gauge $g^{-1}\mathrm{d}g$. The gauge transformation g maps S^3_∞ to the group SU(2), which is itself a three-sphere. How many times it wraps around, $n(g)\in\pi_3(\mathrm{SU}(2))=\mathbb{Z}$, is the instanton number. The energy density (blue) sits in a lump inside.

Why can we not simply undo the winding by a gauge transformation? A gauge transformation h that is smooth on the whole of $\mathbb{R}^4$ restricts on S^3_R to a null-homotopic map: $h_t(x) = h(tx)$, $t \in [0,1]$, is a homotopy from the constant map $h(0)$ to $h|_{S^3_R}$. So it changes g to gh with $[gh] = [g]$ (this uses the additivity (1.62) proved below). The winding at infinity is gauge invariant.

The winding number as an integral

For $\pi_1(S^1)$ the winding number was $\frac{1}{2\pi}\oint \mathrm{d}\theta = \frac{1}{2\pi i}\oint g^{-1}\mathrm{d}g$ for $g = e^{i\theta}$. The three-dimensional analogue uses the matrix-valued 1-form $g^{-1}\mathrm{d}g$ cubed.

Working definition 1.26 (winding number of $g : S^3 \to \mathrm{SU}(2)$)

$$n(g) = \frac{1}{24\pi^2}\int_{S^3} \mathrm{tr}\big[(g^{-1}\mathrm{d}g)^3\big] = \frac{1}{24\pi^2}\int \varepsilon^{ijk}\,\mathrm{tr}\big(L_iL_jL_k\big)\,\mathrm{d}^3u, \qquad L_i := g^{-1}\partial_i g, \tag{1.60}$$

where u^i are oriented local coordinates on S^3, $(g^{-1}\mathrm{d}g)^3$ means $g^{-1}\mathrm{d}g \wedge g^{-1}\mathrm{d}g \wedge g^{-1}\mathrm{d}g$ with matrix multiplication [$\wedge$ antisymmetrises in the three directions, which is where ε^{ijk} comes from], and $\varepsilon^{123} = 1$. *Operational meaning:* $n(g)$ is the degree of g [the number of times g covers S^3, counted with orientation signs] as a map $S^3 \to S^3$ (shown below), so it is an integer, and it can be computed either by the integral or by counting signed preimages of a point.

Worked example 1.53: The identity map has winding number one

Let $g_1 : S^3 \to \mathrm{SU}(2)$ be the identification (1.37), $g_1(a_0, \vec{a}) = a_0\mathbf{1} + i\,\vec{a}\cdot\vec{\sigma}$. We evaluate the integrand in two steps: first at the point $\mathbf{1} = (1,0,0,0)$, then everywhere by symmetry.

At $\mathbf{1}$. Near $(1,\vec{0})$ we may use a^1, a^2, a^3 as coordinates on S^3; there $g_1 = \mathbf{1}$ and $\mathrm{d}g_1 = i\sigma_a\,\mathrm{d}a^a$. Then

$$\mathrm{tr}\big[(g_1^{-1}\mathrm{d}g_1)^3\big]\Big|_{\mathbf{1}} \overset{(1)}{=} i^3\,\mathrm{tr}(\sigma_a\sigma_b\sigma_c)\,\mathrm{d}a^a\wedge\mathrm{d}a^b\wedge\mathrm{d}a^c \overset{(2)}{=} -i\cdot 2i\,\varepsilon_{abc}\,\mathrm{d}a^a\wedge\mathrm{d}a^b\wedge\mathrm{d}a^c$$
$$\overset{(3)}{=} 2\cdot 6\,\mathrm{d}a^1\wedge\mathrm{d}a^2\wedge\mathrm{d}a^3 = 12\,\mathrm{d}a^1\wedge\mathrm{d}a^2\wedge\mathrm{d}a^3.$$

(1) $g_1^{-1}\mathrm{d}g_1 = i\sigma_a\mathrm{d}a^a$ at $\mathbf{1}$; the wedge product of three copies, with the matrices multiplied in order.

(2) $\mathrm{tr}(\sigma_a\sigma_b\sigma_c) = 2i\,\varepsilon_{abc}$: from $\sigma_a\sigma_b = \delta_{ab}\mathbf{1} + i\varepsilon_{abd}\sigma_d$ and $\mathrm{tr}\,\sigma_c = 0$, $\mathrm{tr}(\sigma_d\sigma_c) = 2\delta_{dc}$.

(3) $\varepsilon_{abc}\,\mathrm{d}a^a\wedge\mathrm{d}a^b\wedge\mathrm{d}a^c$ has $3! = 6$ nonzero terms, each equal to $+\mathrm{d}a^1\wedge\mathrm{d}a^2\wedge\mathrm{d}a^3$ after reordering.

At $\mathbf{1}$ the outward normal of $S^3 \subset \mathbb{R}^4$ is e_0, and (e_0, e_1, e_2, e_3) is the standard frame, so (e_1, e_2, e_3) is positively oriented on S^3 there. Hence $\mathrm{d}a^1\wedge\mathrm{d}a^2\wedge\mathrm{d}a^3 = \mathrm{vol}_{S^3}$ at $\mathbf{1}$, and the integrand equals $12\,\mathrm{vol}_{S^3}$ at $\mathbf{1}$.

Everywhere. The 3-form $\mathrm{tr}[(g^{-1}\mathrm{d}g)^3]$ on the group is *bi-invariant*. Under $g \mapsto hg$ with h constant, $(hg)^{-1}\mathrm{d}(hg) = g^{-1}h^{-1}h\,\mathrm{d}g = g^{-1}\mathrm{d}g$; under $g \mapsto gh$, $g^{-1}\mathrm{d}g \mapsto h^{-1}(g^{-1}\mathrm{d}g)h$, and the trace of a product is unchanged by conjugation. Left multiplication by a unit quaternion h is a rotation of $\mathbb{R}^4$ (it preserves $a_0^2 + |\vec{a}|^2 = \det$, and it is connected to the identity), so it also preserves vol_{S^3}. Two left-invariant 3-forms that agree at $\mathbf{1}$ agree everywhere, so

$\operatorname{tr}[(g_1^{-1}\mathrm{d}g_1)^3] = 12\,\mathrm{vol}_{S^3}$ on all of S^3, and

$$n(g_1) \overset{(1)}{=} \frac{12}{24\pi^2}\operatorname{Vol}(S^3) \overset{(2)}{=} \frac{12}{24\pi^2}\int_0^{\pi}\!\!\int_0^{\pi}\!\!\int_0^{2\pi} \sin^2\chi\,\sin\theta\,\mathrm{d}\varphi\,\mathrm{d}\theta\,\mathrm{d}\chi \overset{(3)}{=} \frac{12}{24\pi^2}\cdot\frac{\pi}{2}\cdot 2\cdot 2\pi \overset{(4)}{=} 1.$$

(1) The integrand is $12\,\mathrm{vol}_{S^3}$.
(2) Hyperspherical coordinates, as in Example 1.52.
(3) $\int_0^{\pi}\sin^2\chi\,\mathrm{d}\chi = \frac{\pi}{2}$, $\int_0^{\pi}\sin\theta\,\mathrm{d}\theta = 2$.
(4) So $\operatorname{Vol}(S^3) = 2\pi^2$, and $12\cdot 2\pi^2 = 24\pi^2$.

This explains the strange normalisation $1/24\pi^2$: it is $1/(12\cdot\operatorname{Vol}(S^3))$. For a general map g, $\operatorname{tr}[(g^{-1}\mathrm{d}g)^3]$ is the pullback by g of the same form on the group, i.e. $12\,g^*\mathrm{vol}_{S^3}$, so $n(g) = \int g^*(\mathrm{vol}/2\pi^2) = \deg g$ by (1.50). In particular $n(g)\in\mathbb{Z}$ and it is a homotopy invariant. (The skyrmion of Example 1.52 has $n(U) = \deg U = -1$.)

Worked example 1.54: Winding numbers add under pointwise multiplication

Let $g = g_1g_2$ (pointwise matrix product) and write $a := g_1^{-1}\mathrm{d}g_1$, $b := \mathrm{d}g_2\,g_2^{-1}$. Then $g^{-1}\mathrm{d}g = g_2^{-1}g_1^{-1}(\mathrm{d}g_1\,g_2 + g_1\mathrm{d}g_2) = g_2^{-1}(a+b)g_2$, and

$$\begin{aligned}\operatorname{tr}\big[(g^{-1}\mathrm{d}g)^3\big] &\overset{(1)}{=} \operatorname{tr}\big[(a+b)^3\big]\\ &\overset{(2)}{=} \operatorname{tr}a^3 + \operatorname{tr}b^3 + 3\operatorname{tr}(a^2b) + 3\operatorname{tr}(ab^2)\\ &\overset{(3)}{=} \operatorname{tr}\big[(g_1^{-1}\mathrm{d}g_1)^3\big] + \operatorname{tr}\big[(g_2^{-1}\mathrm{d}g_2)^3\big] - 3\,\mathrm{d}\,\operatorname{tr}(ab).\end{aligned} \tag{1.61}$$

(1) The conjugation by g_2 drops out of the trace.
(2) Expand the eight terms. For 1-forms, $\operatorname{tr}(xyz) = \operatorname{tr}(yzx)$ (moving a 1-form past a 2-form costs no sign), so $\operatorname{tr}(aab) = \operatorname{tr}(aba) = \operatorname{tr}(baa)$, and similarly for the terms with two b's.
(3) $\operatorname{tr}b^3 = \operatorname{tr}[g_2(g_2^{-1}\mathrm{d}g_2)^3g_2^{-1}] = \operatorname{tr}[(g_2^{-1}\mathrm{d}g_2)^3]$. For the cross terms: $\mathrm{d}a = \mathrm{d}(g_1^{-1})\wedge\mathrm{d}g_1 = -g_1^{-1}\mathrm{d}g_1\,g_1^{-1}\wedge\mathrm{d}g_1 = -a^2$ and similarly $\mathrm{d}b = b^2$, so $\mathrm{d}\operatorname{tr}(ab) = \operatorname{tr}(\mathrm{d}a\,b) - \operatorname{tr}(a\,\mathrm{d}b) = -\operatorname{tr}(a^2b) - \operatorname{tr}(ab^2)$.

The last term is exact, and an exact form integrates to zero over the closed manifold S^3 (Stokes, $\partial S^3 = \varnothing$). Therefore

$$n(g_1g_2) = n(g_1) + n(g_2), \qquad n(g^{-1}) = -n(g), \qquad n(g^k) = k\,n(g). \tag{1.62}$$

The second follows from the first with $g_1g_2 = \mathbf{1}$ and $n(\mathbf{1}) = 0$. This is the Eckmann–Hilton statement of §1.2.8 made concrete: the pointwise product of maps into a group gives the same group law on π_3 as the cube-picture concatenation. An explicit map of winding k is g_1^k. In the polar form $g_1 = \cos\chi + i\sin\chi\,\hat{\vec{n}}\cdot\vec{\sigma} = \exp(i\chi\,\hat{\vec{n}}\cdot\vec{\sigma})$ it is $g_1^k = \exp(ik\chi\,\hat{\vec{n}}\cdot\vec{\sigma})$: the same hedgehog shape, turning k times as fast.

From the field strength to the winding: Chern–Simons and Stokes

The winding at infinity can be written as a volume integral of the field strength, exactly as the magnetic charge was an integral of $\vec{B}$.

Key idea: $\operatorname{tr} F \wedge F$ is a total derivative

With the **Chern–Simons form** $\mathrm{CS}(A) := \operatorname{tr}\big(A \wedge \mathrm{d}A + \frac{2}{3} A \wedge A \wedge A\big)$,

$$\operatorname{tr}(F \wedge F) = \mathrm{d}\,\mathrm{CS}(A). \tag{1.63}$$

Here is the verification, using only graded cyclicity $\operatorname{tr}(\alpha \wedge \beta) = (-1)^{pq} \operatorname{tr}(\beta \wedge \alpha)$ for matrix-valued p- and q-forms.

$$\mathrm{d}\,\mathrm{CS} \overset{(1)}{=} \operatorname{tr}(\mathrm{d}A \wedge \mathrm{d}A) + \tfrac{2}{3} \operatorname{tr}\big(\mathrm{d}A\, A^2 - A\, \mathrm{d}A\, A + A^2 \mathrm{d}A\big) \overset{(2)}{=} \operatorname{tr}(\mathrm{d}A \wedge \mathrm{d}A) + 2 \operatorname{tr}(\mathrm{d}A \wedge A^2),$$

$$\operatorname{tr}(F \wedge F) \overset{(3)}{=} \operatorname{tr}(\mathrm{d}A\, \mathrm{d}A) + \operatorname{tr}(\mathrm{d}A\, A^2) + \operatorname{tr}(A^2 \mathrm{d}A) + \operatorname{tr}(A^4) \overset{(4)}{=} \operatorname{tr}(\mathrm{d}A \wedge \mathrm{d}A) + 2 \operatorname{tr}(\mathrm{d}A \wedge A^2).$$

(1) Leibniz rule; $\mathrm{d}(\mathrm{d}A) = 0$; a minus sign each time d passes a 1-form.
(2) $\operatorname{tr}(A\, \mathrm{d}A\, A) = (-1)^{1\cdot 3} \operatorname{tr}(\mathrm{d}A\, A\, A) = -\operatorname{tr}(\mathrm{d}A\, A^2)$ and $\operatorname{tr}(A^2 \mathrm{d}A) = (-1)^{2\cdot 2} \operatorname{tr}(\mathrm{d}A\, A^2)$; so the bracket is $3 \operatorname{tr}(\mathrm{d}A\, A^2)$.
(3) Expand $F \wedge F = (\mathrm{d}A + A^2) \wedge (\mathrm{d}A + A^2)$.
(4) $\operatorname{tr}(A^2 \mathrm{d}A) = \operatorname{tr}(\mathrm{d}A\, A^2)$ as before, and $\operatorname{tr}(A^4) = \operatorname{tr}(A \cdot A^3) = (-1)^{1\cdot 3} \operatorname{tr}(A^3 \cdot A) = -\operatorname{tr}(A^4)$, so it vanishes.

Now integrate over a large ball B_R and use Stokes' theorem. Assume A is smooth on all of $\mathbb{R}^4$ and becomes the pure gauge (1.59) at large r.

$$\begin{aligned} \int_{\mathbb{R}^4} \operatorname{tr}(F \wedge F) &\overset{(1)}{=} \lim_{R\to\infty} \int_{S^3_R} \mathrm{CS}(A) \overset{(2)}{=} \int_{S^3_\infty} \operatorname{tr}\big(-A^3 + \tfrac{2}{3} A^3\big) \\ &\overset{(3)}{=} -\frac{1}{3} \int_{S^3_\infty} \operatorname{tr}\big[(g^{-1}\mathrm{d}g)^3\big] \overset{(4)}{=} -8\pi^2\, n(g). \end{aligned} \tag{1.64}$$

(1) (1.63) and Stokes on the ball, whose boundary is S^3_R with the outward orientation.
(2) On S^3_∞, $F = 0$, so $\mathrm{d}A = -A \wedge A$.
(3) Collect terms; $A = g^{-1}\mathrm{d}g$ there.
(4) The definition (1.60): $\frac{1}{3} \cdot 24\pi^2 = 8\pi^2$.

Working definition 1.27 (instanton number)

$$k := -\frac{1}{8\pi^2} \int_{\mathbb{R}^4} \operatorname{tr}(F \wedge F) = \frac{g^2}{32\pi^2} \int F^a_{\mu\nu} \tilde{F}^a_{\mu\nu}\, \mathrm{d}^4x = n(g) \in \mathbb{Z}, \qquad \tilde{F}^a_{\mu\nu} := \tfrac{1}{2} \varepsilon_{\mu\nu\rho\sigma} F^a_{\rho\sigma}$$

[$n(g)$ is the winding number of the gauge transformation g for which $A = g^{-1}\mathrm{d}g$ on the sphere at infinity]. The minus sign in the first expression is the price of anti-Hermitian fields [$A^\dagger = -A$, as for $A = g^{-1}\mathrm{d}g$ with g unitary]: in Hermitian conventions it is $+\frac{1}{8\pi^2} \int \operatorname{tr}(F_{\text{phys}} \wedge F_{\text{phys}})$ (with g absorbed). *Operational meaning:* either integrate the "topological charge density" $\frac{g^2}{32\pi^2} F\tilde{F}$ over spacetime, or read off the winding of the gauge transformation at infinity.

The middle expression follows by substituting the dictionary:

$$-\frac{1}{8\pi^2} \operatorname{tr}(F \wedge F) \overset{(1)}{=} -\frac{1}{8\pi^2} (-ig)^2 \operatorname{tr}(T^a T^b)\, F^a \wedge F^b \overset{(2)}{=} \frac{g^2}{16\pi^2} F^a \wedge F^a \overset{(3)}{=} \frac{g^2}{32\pi^2} F^a_{\mu\nu} \tilde{F}^a_{\mu\nu}\, \mathrm{d}^4x.$$

(1) $F = -igF^aT^a$, where $F^a = \frac{1}{2} F^a_{\mu\nu} \mathrm{d}x^\mu \wedge \mathrm{d}x^\nu$.
(2) $(-ig)^2 = -g^2$ and $\operatorname{tr}(T^a T^b) = \frac{1}{2}\delta^{ab}$.
(3) $F^a \wedge F^a = \frac{1}{4} F^a_{\mu\nu} F^a_{\rho\sigma}\, \mathrm{d}x^\mu \wedge \mathrm{d}x^\nu \wedge \mathrm{d}x^\rho \wedge \mathrm{d}x^\sigma = \frac{1}{4} \varepsilon_{\mu\nu\rho\sigma} F^a_{\mu\nu} F^a_{\rho\sigma}\, \mathrm{d}^4x = \frac{1}{2} F^a_{\mu\nu} \tilde{F}^a_{\mu\nu}\, \mathrm{d}^4x$.

Physical picture: Integral of a local density = integer

$F\tilde{F}$ is a local density built from the fields, and in general it has no reason to integrate to anything special. Yet its integral is always $32\pi^2/g^2$ times an integer. The mechanism is the same as for the monopole's flux: the density is a total derivative, so the integral only sees the boundary, where the field is a map between spheres, and maps between spheres are counted by integers. The same mechanism will produce the index theorem's integrality (Chapter 4).

The action bound and the tunnelling amplitude

Topology also bounds the action from below, exactly as in the kink and the baby skyrmion. In four Euclidean dimensions $\tilde{\tilde{F}} = F$ and $\tilde{F}^a_{\mu\nu}\tilde{F}^a_{\mu\nu} = F^a_{\mu\nu}F^a_{\mu\nu}$. So

$$S \overset{(1)}{=} \frac{1}{4}\int F^a_{\mu\nu}F^a_{\mu\nu}\,\mathrm{d}^4x \overset{(2)}{=} \frac{1}{8}\int \left(F^a_{\mu\nu} \mp \tilde{F}^a_{\mu\nu}\right)^2 \mathrm{d}^4x \pm \frac{1}{4}\int F^a_{\mu\nu}\tilde{F}^a_{\mu\nu}\,\mathrm{d}^4x$$
$$\overset{(3)}{=} \frac{1}{8}\int \left(F \mp \tilde{F}\right)^2 \mathrm{d}^4x \pm \frac{8\pi^2}{g^2}\,k \overset{(4)}{\geq} \frac{8\pi^2}{g^2}\,|k|.$$

(1) The Yang–Mills action in Hermitian components.
(2) $(F \mp \tilde{F})^2 = FF + \tilde{F}\tilde{F} \mp 2F\tilde{F} = 2FF \mp 2F\tilde{F}$, so $\frac{1}{8}(F \mp \tilde{F})^2 = \frac{1}{4}FF \mp \frac{1}{4}F\tilde{F}$.
(3) The instanton number: $\int F\tilde{F} = 32\pi^2 k/g^2$.
(4) Choose the sign equal to the sign of k; a square is non-negative.

Equality holds exactly for **self-dual** ($F = \tilde{F}$, $k > 0$) or **anti-self-dual** ($F = -\tilde{F}$, $k < 0$) fields: first-order equations, whose solutions are automatically solutions of the second-order Yang–Mills equations (by the Bianchi identity). These are the **instantons**. In the path integral each contributes a weight $e^{-S} = e^{-8\pi^2|k|/g^2}$. As a function of g this has a vanishing Taylor series at $g = 0$: instanton effects are invisible to every order of perturbation theory [QFT notes §4.13.4, p. 403].

Worked example 1.55: The BPST instanton has $k = 1$ without computing F

Belavin, Polyakov, Schwarz and Tyupkin (1975) found the solution, in the **regular gauge**,

$$A = \frac{r^2}{r^2+\rho^2}\,g_1^{-1}\mathrm{d}g_1, \qquad g_1(x) = \frac{x^0\mathbf{1} + i\,x^a\sigma_a}{r}, \qquad r^2 = \sum_\mu (x^\mu)^2,$$

with an arbitrary size $\rho > 0$ (and an arbitrary centre, here the origin). Two properties decide its instanton number.

- *It is smooth at the origin.* $g_1^{-1}\mathrm{d}g_1$ behaves like $1/r$ near $r = 0$ (it is homogeneous of degree -1), but the prefactor behaves like r^2/ρ^2. In fact $g_1^{-1}\mathrm{d}g_1 = r^{-2}(x^0 - ix^a\sigma_a)(\mathrm{d}x^0 + i\sigma_b\mathrm{d}x^b) + (\text{a term} \propto \mathrm{d}r)$, and $A = (r^2+\rho^2)^{-1} \times (\text{polynomial in } x)$ is smooth. So Stokes on $\mathbb{R}^4$ is legitimate.
- *At infinity it is pure gauge with $g = g_1$.* As $r \to \infty$, $r^2/(r^2+\rho^2) \to 1$. And g_1 restricted to S^3_R is the map of Example 1.53 after rescaling by $1/R$, so $n(g_1) = 1$.

By (1.64), $k = n(g_1) = 1$. One checks by direct computation (done in Chapter 8) that F is (anti-)self-dual, the sign being fixed by the orientation conventions, so $S = 8\pi^2/g^2$: the instanton saturates the bound. Its field strength is concentrated in a region of size ρ around the centre, as in Figure 1.33. With g_1^{-1} in place of g_1 the same construction gives $k = -1$, the anti-instanton, by (1.62).

The vacuum is not unique: winding vacua and θ

There is a Hamiltonian version of the same topology. Work in the temporal gauge $A_0 = 0$ and consider static field configurations of zero energy on space $\mathbb{R}^3$. They have $F = 0$, so $A_i = g^{-1}\partial_i g$ with $g : \mathbb{R}^3 \to \mathrm{SU}(2)$. Require $g \to \mathbf{1}$ at spatial infinity (gauge transformations that do not change the state at infinity). Adding the point at infinity, $\mathbb{R}^3 \cup \{\infty\} \cong S^3$, so each classical vacuum has a winding number $n(g) \in \pi_3(\mathrm{SU}(2)) = \mathbb{Z}$, computed by (1.60) over $\mathbb{R}^3$. We get infinitely many classical vacua $|n\rangle$, $n \in \mathbb{Z}$. Their winding is the **Chern–Simons number** $N_{\mathrm{CS}}[A] := -\frac{1}{8\pi^2}\int_{\mathbb{R}^3} \mathrm{CS}(A)$, which on a pure gauge equals $n(g)$ by steps (2)–(4) of (1.64).

A gauge transformation of winding one maps $|n\rangle$ to $|n+1\rangle$: it is a symmetry of the Hamiltonian, but it cannot be deformed to the identity (it is a **large gauge transformation**). So the vacua $|n\rangle$ are physically equivalent but distinct as states. Between them the energy functional has barriers (Figure 1.34): to go from $|n\rangle$ to $|n+1\rangle$ continuously the field must pass through configurations with $F \neq 0$. Quantum mechanically it can tunnel, and the tunnelling path in Euclidean time is an instanton. Apply Stokes to the slab $\mathbb{R}^3 \times [-T, T]$:

$$k \overset{(1)}{=} -\frac{1}{8\pi^2}\int_{\mathbb{R}^3\times[-T,T]} \mathrm{d}\,\mathrm{CS} \overset{(2)}{=} -\frac{1}{8\pi^2}\left(\int_{x^0=T} \mathrm{CS} - \int_{x^0=-T} \mathrm{CS}\right) \overset{(3)}{=} N_{\mathrm{CS}}(T) - N_{\mathrm{CS}}(-T) \xrightarrow{T\to\infty} n_+ - n_-.$$

(1) (1.63), with T large enough to contain the instanton.
(2) Stokes. The outward normal is $+e_0$ on the top face and $-e_0$ on the bottom face, hence the relative sign; the sides at spatial infinity contribute nothing because $A \to 0$ there fast enough (we have required $g \to \mathbf{1}$).
(3) Definition of N_{CS} with the standard orientation of $\mathbb{R}^3$.

An instanton of charge k is a tunnelling event from $|n\rangle$ to $|n+k\rangle$.

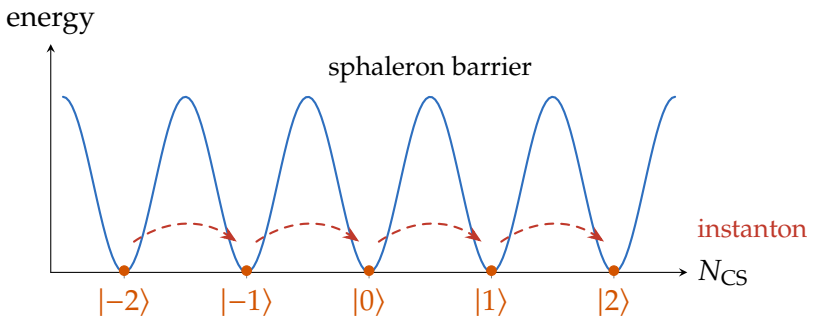


Figure 1.34: Schematic energy of gauge-field configurations as a function of the Chern–Simons number. The classical vacua $|n\rangle$ (orange) sit at the integers and are related by large gauge transformations. Classically they are separated by barriers; quantum mechanically an instanton (red) tunnels from $|n\rangle$ to $|n+1\rangle$. The true vacuum is a Bloch wave over this "crystal": the θ-vacuum.

This is the quantum mechanics of a particle in a periodic potential, and the answer is the same: the energy eigenstates are Bloch waves. Let $\mathcal{T}$ be the operator of the winding-one large gauge transformation, $\mathcal{T}|n\rangle = |n+1\rangle$. It commutes with the Hamiltonian, so we may diagonalise them together. Its eigenstates are the θ-vacua

$$|\theta\rangle := \sum_{n\in\mathbb{Z}} e^{-in\theta}|n\rangle, \qquad \theta \in [0, 2\pi),$$

because

$$\mathcal{T}|\theta\rangle \overset{(1)}{=} \sum_n e^{-in\theta}|n+1\rangle \overset{(2)}{=} \sum_m e^{-i(m-1)\theta}|m\rangle \overset{(3)}{=} e^{i\theta}|\theta\rangle.$$

(1) Linearity and $\mathcal{T}|n\rangle = |n+1\rangle$.
(2) Relabel $m = n + 1$; the sum over all integers is unchanged.
(3) Pull out $e^{i\theta}$.

Since $\mathcal{T}$ commutes with every gauge-invariant observable, different θ's never mix: θ is a new constant of nature, like a superselection label. What does it do to the path integral? The amplitude $\langle m|e^{-HT}|n\rangle$ is the path integral over fields that interpolate from $|n\rangle$ to $|m\rangle$, i.e. over the sector $k = m - n$; call it I_{m-n}. Then

$$\langle\theta'|e^{-HT}|\theta\rangle \overset{(1)}{=} \sum_{m,n} e^{im\theta'} e^{-in\theta}\, I_{m-n} \overset{(2)}{=} \sum_{n} e^{in(\theta'-\theta)} \sum_{k} e^{ik\theta'} I_k$$
$$\overset{(3)}{=} 2\pi\,\delta(\theta'-\theta) \sum_{k\in\mathbb{Z}} e^{ik\theta} \int_{[k]} \mathcal{D}A\; e^{-S[A]}.$$

(1) Expand both θ-states in the $|n\rangle$; $\langle\theta'| = \sum_m e^{im\theta'}\langle m|$.
(2) Put $m = n + k$ and sum over n and k independently.
(3) $\sum_n e^{in\alpha} = 2\pi\delta(\alpha)$ for $\alpha \in (-\pi, \pi]$; and I_k is the path integral over the sector of instanton number k.

So the θ-vacuum is the same as adding to the Euclidean action the term $-i\theta k = -i\theta\frac{g^2}{32\pi^2}\int F\tilde{F}\,\mathrm{d}^4x$. It is a total derivative, so it changes neither the classical equations nor perturbation theory, yet it is physical. In QCD it violates CP, and the bound on the neutron electric dipole moment forces $|\theta| \lesssim 10^{-10}$: the **strong CP problem**. We return to all of this, and to the counting of instanton moduli, in Chapter 8.

Dictionary: the θ-vacuum as a crystal

particle in a periodic potential	Yang–Mills vacuum
lattice site n	classical vacuum $\lvert n\rangle$, winding $n \in \pi_3(\mathrm{SU}(2))$
translation by one lattice spacing	large gauge transformation $\mathcal{T}$
tunnelling through a barrier	instanton, weight $e^{-8\pi^2/g^2}$
Bloch momentum × lattice spacing	θ
Bloch wave $\sum_n e^{-in\theta}\lvert n\rangle$	θ-vacuum $\lvert\theta\rangle$

1.3.15 Why Nash now turns to cohomology

Nash closes §3 with a warning: homotopy groups, "though simple and intuitive to describe, can be very difficult to calculate", and "a much more calculable set of invariants … are the cohomology groups". After this section we can say precisely what goes wrong and what cohomology will fix.

	homotopy groups $\pi_m(X)$	cohomology groups $H^m(X)$ (Nash §4)
definition	maps *into* X from spheres	(for de Rham) closed forms *on* X modulo exact ones
above dim X	usually nonzero ($\pi_3(S^2) = \mathbb{Z}$, Figure 1.27)	always zero
cutting and gluing	no simple rule (van Kampen only for π_1)	Mayer–Vietoris exact sequence
computed by	exact sequences, much ingenuity; unknown in general	linear algebra; integrals of forms
algebra	non-abelian π_1, abelian $\pi_{m\geq 2}$	abelian groups, with a ring structure (wedge product)
physics	defect and instanton charges	fluxes, characteristic classes, anomalies

The two are closely related, and Nash's first definition of cohomology uses homotopy. The circle has only one nonzero homotopy group, $\pi_1(S^1) = \mathbb{Z}$ (Nash's (1.32)): it is an **Eilenberg–MacLane space** $K(\mathbb{Z}, 1)$. Nash defines $H^n(X; G) = [X, K(G, n)]_0$, so in particular $H^1(X; \mathbb{Z}) = [X, S^1]_0$: the first integral cohomology of X is the set of ways of mapping X to a circle, and these are counted by winding numbers.

Worked example 1.56: $H^1(T^2;\mathbb{Z}) = [T^2, S^1] = \mathbb{Z}^2$ by winding numbers

A map $f : T^2 \to S^1$, restricted to the two basic circles of the torus (Example 1.26 picture), has two winding numbers

$$(m, n) = \left(\frac{1}{2\pi} \oint_a f^* \mathrm{d}\theta,\ \frac{1}{2\pi} \oint_b f^* \mathrm{d}\theta \right) \in \mathbb{Z}^2.$$

Every pair occurs: $f_{m,n}(\theta, \phi) = e^{i(m\theta + n\phi)}$ has windings (m, n), since $f_{m,n}^* \mathrm{d}\theta = m\,\mathrm{d}\theta + n\,\mathrm{d}\phi$. Conversely, if f has windings (m, n), then $f \cdot f_{m,n}^{-1}$ has windings $(0, 0)$, so it lifts to a real function $\tilde{h}$ on T^2 with $f = f_{m,n}\, e^{i\tilde{h}}$, and $f_{m,n} e^{it\tilde{h}}$, $t \in [0, 1]$, is a homotopy from $f_{m,n}$ to f. So $[T^2, S^1] = \mathbb{Z}^2$ (and base points do not matter since S^1 is abelian). The answer is the pair of periods of the closed 1-form $f^*\mathrm{d}\theta/2\pi$: an element of $H^1(T^2;\mathbb{Z})$. The de Rham version of this, $H^1_{\mathrm{dR}}(T^2) = \mathbb{R}^2$, is where the next part starts.

Reading Nash (pp. 9–12, (1.26)–(1.36))

- *(1.26)–(1.27)*: $\pi_1(X) = [S^1, X]_0$ with the product "first α at double speed, then β"; the group axioms hold up to homotopy, i.e. ΩX is an H-space. Our §1.3.1, with the homotopies written out.
- *(1.28)–(1.30)*: $\pi_n(X) = [S^n, X]_0 = [S^{n-1}, \Omega X]_0 = [SS^{n-2}, \Omega X]_0$; the last form has an H-space target and a coH-space source, so by the remark at the end of Nash §2 (our Eckmann–Hilton argument) π_n is abelian for $n \geq 2$. Our §1.3.6, with the cube picture of Figure 1.23.
- $\pi_1(G)$ *abelian*: $\pi_1(G) = [SS^0, G]_0$, G an H-space; the footnote explains $S^0 = \{x_1^2 = 1\}$, two points. Our §1.3.4; the example $\pi_1(\mathrm{SO}(3)) = \mathbb{Z}_2$ and the belt trick are our additions.
- *Genus $g \geq 2$ Riemann surfaces* have non-abelian π_1: our Figure 1.19 and the presentation (1.35).
- *(1.31)*: $[X, Y]_0/\pi_1(Y) \cong [X, Y]$ via change of base point along paths γ, depending

only on the homotopy class of γ, and requiring Y path-connected. Our §1.3.5. Nash's notation $[X, x_0; Y, y_0]_0$ is our $[X, Y]_0$ with the base points shown.

- *(1.32)–(1.34)*: $\pi_m(S^1)$, $\pi_m(S^n)$ for $m \leq n$, and Serre's result for $m > n$. Nash cites Bott and Tu; we computed (1.32) by lifting, (1.33) by general position and the degree, and illustrated (1.34) by the Hopf map and Figure 1.27.
- *(1.35)–(1.36)*: stability, $G_m = \pi_{2m+2}(S^{m+2})$. Our (1.57) and the Freudenthal theorem; note Nash's m is the row index k of our table.
- *Lie groups and Bott periodicity*, deferred by Nash to his chapter 3 (K-theory): our §1.3.12.
- *Cohomology is more calculable*: our §1.3.15.

Everything in §§1.3.11, 1.3.13 and 1.3.14 (exact homotopy sequences, defects, instantons, θ) goes beyond Nash's text in this section; it anticipates his §5 and chapters 8 and 10.

Recap: homotopy groups (Nash §3)

- $\pi_1(X) = [S^1, X]_0$: based loops up to homotopy, multiplied by concatenation. $\pi_1(S^1) = \mathbb{Z}$ by the winding number $\frac{1}{2\pi}\oint \mathrm{d}\theta$; $\pi_1(T^2) = \mathbb{Z}^2$; π_1 of the figure eight and of genus $g \geq 2$ surfaces is non-abelian; $\pi_1(\mathbb{RP}^2) = \pi_1(\mathrm{SO}(3)) = \mathbb{Z}_2$ (the belt trick, spin $\frac{1}{2}$).
- $\pi_n(X) = [S^n, X]_0$: maps of cubes with boundary at the base point; abelian for $n \geq 2$ (Eckmann–Hilton); coverings do not change $\pi_{n\geq 2}$.
- Spheres: $\pi_m(S^n) = 0$ for $m < n$ (general position), $\pi_n(S^n) = \mathbb{Z}$ (degree $= \frac{1}{\mathrm{Vol}}\int f^*\mathrm{vol}$), $\pi_3(S^2) = \mathbb{Z}$ (Hopf map, linking number), finite otherwise except $\pi_{2n-1}(S^n)$ for even n (Serre); stable along diagonals (Freudenthal).
- Free classes are based classes modulo the action of π_1: conjugacy classes for loops.
- The long exact sequence of a fibration computes π_*: $\pi_2(S^2) = \pi_1(S^1)$, $\pi_3(S^2) = \pi_3(S^3)$, $\pi_k(\mathrm{SU}(n))$ stable, $\pi_2(G/H) = \pi_1(H)$. Stable groups of classical groups are periodic (Bott).
- Physics: defects of codimension d carry charges in π_{d-1} of the vacuum manifold; textures in π_D. Instantons carry $k = -\frac{1}{8\pi^2}\int \mathrm{tr}\, F \wedge F = n(g) \in \pi_3(\mathrm{SU}(2)) = \mathbb{Z}$, the winding of the gauge transformation at infinity; $S \geq 8\pi^2|k|/g^2$; the vacuum is a θ-vacuum.
- Homotopy groups are hard to compute; cohomology (next) is easier.

1.4 Cohomology and homology groups

Homotopy groups are the most natural invariants we have met so far. A loop that cannot be shrunk, a sphere that cannot be collapsed: these are things we can picture. Their weakness is that they are *hard to compute*. Nobody knows all the groups $\pi_m(S^2)$. Even the sphere, the simplest non-trivial space there is, defeats us.

This section is about a second family of invariants, the **cohomology groups** $H^n(X; G)$ and their duals, the **homology groups** $H_n(X; G)$. They are coarser than homotopy groups: they see

less. In exchange they are *computable*, in the most literal sense. For a space built out of finitely many triangles, computing homology is a matter of writing down a few integer matrices and finding their ranks. For a manifold, computing real cohomology is a matter of asking which closed differential forms fail to be exact.

For a physicist there are three reasons to care.

- Cohomology classes are what *topological charges* are. Magnetic charge, instanton number and winding number are all integrals of a closed form over a closed surface. Those integrals only depend on the cohomology class of the form.

- Cohomology counts *zero modes*. By Hodge theory (Chapter 2) the dimension of $H^p_{\mathrm{dR}}(M)$ equals the number of harmonic p-forms. These are the zero modes of the Laplacian on p-forms.

- The alternating sum of these dimensions, the Euler characteristic, is the simplest *index*. The whole of Chapter 4 generalises it.

Nash gives two definitions of cohomology in quick succession. One uses maps into special spaces (Eilenberg–MacLane spaces). The other uses complexes, meaning sequences of linear maps whose consecutive composites vanish. We follow him, and then spend most of our time on the second definition, because that is the one we can compute with. Figure 1.35 shows how the pieces fit together.

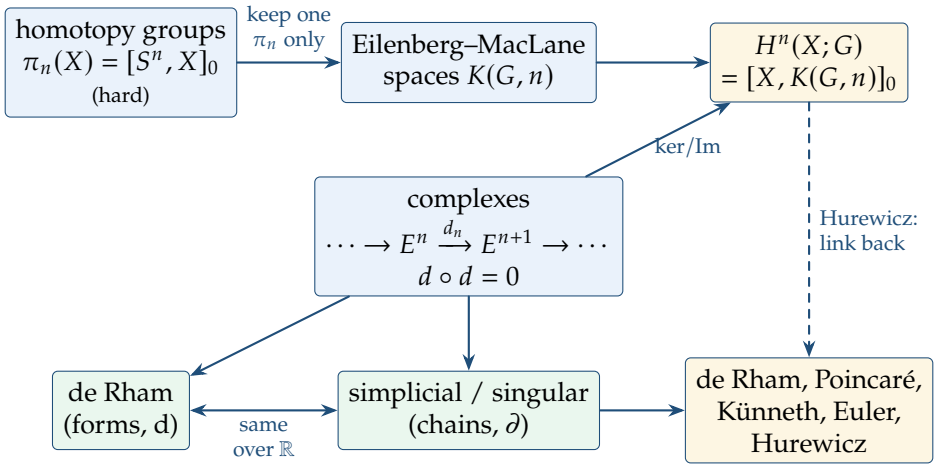


Figure 1.35: Roadmap of Section 1.4. Nash's first definition (top row) makes it obvious that cohomology depends only on homotopy type. The second (middle) turns cohomology into linear algebra, with two concrete realisations (bottom row): differential forms and triangulations. The theorems at the bottom right let us compute.

1.4.1 Eilenberg–MacLane spaces: one homotopy group and nothing else

Recall from earlier in the chapter that for a pointed space X the m-th homotopy group is $\pi_m(X) = [S^m, X]_0$. This is the set of based homotopy classes of maps from the m-sphere into X. A general space has many non-zero homotopy groups, all tangled together. Eilenberg and MacLane asked a simple question: what if a space had *only one*?

Working definition 1.28 (Eilenberg–MacLane space)

Let G be a group and $n \geq 1$ an integer. A $K(G,n)$ is a path-connected [any two points can be joined by a path in K] space K with

$$\pi_m\big(K(G,n)\big) = \begin{cases} G & m = n, \\ 0 & m \neq n. \end{cases} \tag{1.65}$$

Operational test: list the homotopy groups of your candidate space. Exactly one of them must be non-zero. For $n \geq 2$, G must be abelian [$gh = hg$ for all $g, h \in G$], because π_n is always abelian for $n \geq 2$.

Such spaces exist for every G (abelian when $n \geq 2$). Up to homotopy equivalence there is only one of each. So we can speak of *the* $K(G,n)$, just as we speak of *the* circle without fixing a radius.

Here is the idea behind existence. Start with a bunch of n-spheres glued at a point. Their π_n is a free group on the spheres. Glue on $(n+1)$-dimensional discs to impose the relations of G. Now π_n is exactly G, but the gluing has created unwanted higher homotopy groups. Kill each class of π_{n+1} by gluing in an $(n+2)$-disc along it. Then kill π_{n+2} by $(n+3)$-discs, and so on for ever. The result is usually infinite-dimensional, and we will see that this cannot be avoided.

Worked example 1.57: The circle is a $K(\mathbb{Z},1)$

We already know $\pi_1(S^1) = \mathbb{Z}$: a loop is classified by how many times it winds. The higher groups vanish. To see why, unroll the circle onto its universal cover $\mathbb{R} \to S^1$, $t \mapsto e^{2\pi i t}$. A map $S^m \to S^1$ with $m \geq 2$ lifts to a map $S^m \to \mathbb{R}$, because S^m is simply connected. The lift can be shrunk to a point, since $\mathbb{R}$ is contractible. Push the shrinking back down to S^1 and the original map becomes null-homotopic. Hence

$$\begin{aligned} \pi_m(S^1) &\overset{(1)}{=} \pi_m(\mathbb{R}) \\ &\overset{(2)}{=} 0 \qquad (m \geq 2), \end{aligned}$$

(1) Every map (and every homotopy) from a simply connected S^m lifts through the covering, and lifts of homotopic maps are homotopic.

(2) $\mathbb{R}$ is contractible, so all its homotopy groups vanish.

Therefore $S^1 = K(\mathbb{Z},1)$, which is the example Nash gives. In Section 1.5.13 we will get the same result from the exact homotopy sequence of the fibration $\mathbb{Z} \to \mathbb{R} \to S^1$.

Worked example 1.58: More Eilenberg–MacLane spaces

- *Torus.* $T^k = S^1 \times \cdots \times S^1$ has universal cover $\mathbb{R}^k$, which is contractible. Therefore $T^k = K(\mathbb{Z}^k, 1)$.
- *Surfaces.* A closed orientable surface Σ_g of genus $g \geq 1$ has the contractible plane (for $g = 1$) or the hyperbolic disc (for $g \geq 2$) as universal cover. So $\Sigma_g = K(\pi_1(\Sigma_g), 1)$. The sphere $\Sigma_0 = S^2$ is *not* a $K(G,n)$, since π_2 and π_3 are both $\mathbb{Z}$.
- *Infinite real projective space.* Let $S^\infty = \bigcup_N S^N$ be the unit sphere in $\mathbb{R}^\infty$. It is contractible. The antipodal map $x \mapsto -x$ acts freely, and the quotient is $\mathbb{RP}^\infty$. It has $\pi_1 = \mathbb{Z}_2$ and nothing else: $\mathbb{RP}^\infty = K(\mathbb{Z}_2, 1)$.

- *Infinite complex projective space.* $U(1)$ acts freely on the unit sphere $S^\infty \subset \mathbb{C}^\infty$ by $z \mapsto e^{i\alpha}z$. The quotient $\mathbb{CP}^\infty$ has $\pi_2 = \mathbb{Z}$ and nothing else: $\mathbb{CP}^\infty = K(\mathbb{Z}, 2)$. We prove this in Example 1.101 with the exact homotopy sequence.

Notice that the finite-dimensional versions fail. $\mathbb{RP}^2$ has $\pi_2(\mathbb{RP}^2) = \pi_2(S^2) = \mathbb{Z}$, and $\mathbb{CP}^1 = S^2$ has $\pi_3 = \mathbb{Z}$. Only in the infinite-dimensional limit do the unwanted groups die.

Watch out: $K(G, n)$ is almost never a manifold

Apart from $K(\mathbb{Z}, 1) = S^1$, the tori, the surfaces and similar aspherical manifolds, Eilenberg–MacLane spaces are infinite-dimensional. You should not try to picture $K(\mathbb{Z}, 2)$ as something smooth and finite. Picture it as the receptacle for a certain kind of map. It is a tool, not an object of study.

1.4.2 Nash's first definition: cohomology as maps into $K(G, n)$

We can now state Nash's first definition. The n-th cohomology group of a pointed space X with coefficients in G is

$$H^n(X; G) \coloneqq [X, K(G, n)]_0. \tag{1.66}$$

This is the set of based homotopy classes of maps from X into the Eilenberg–MacLane space. Compare it with homotopy groups. There we map a *fixed* simple space (a sphere) *into* X. Here we map X *into* a fixed simple space. Cohomology is "homotopy the other way round".

Key idea: Cohomology is an invariant of homotopy type

If $X \simeq X'$ are homotopy equivalent, then $[X, Y]_0 = [X', Y]_0$ for any Y (this was shown for $[X, Y]$ earlier in the chapter). Taking $Y = K(G, n)$ gives

$$X \simeq X' \quad \Longrightarrow \quad H^n(X; G) \cong H^n(X'; G).$$

So a contractible space, such as $\mathbb{R}^k$ or a disc, has the cohomology of a point.

The definition becomes concrete when $n = 1$ and $G = \mathbb{Z}$, because then $K(\mathbb{Z}, 1) = S^1$ and

$$H^1(X; \mathbb{Z}) = [X, S^1]_0. \tag{1.67}$$

A degree-one integral cohomology class *is* a homotopy class of maps into the circle. Equivalently, since $S^1 = U(1)$, it is a homotopy class of $U(1)$-valued functions. Physicists call these *gauge transformations up to deformation*, and label them by winding numbers.

Worked example 1.59: $H^1(S^1; \mathbb{Z}) = \mathbb{Z}$ and the winding number

A based map $f : S^1 \to S^1 \subset \mathbb{C}$ is a phase $f(e^{i\theta})$ with $f(1) = 1$. Its homotopy class is fixed by its **winding number**, or **degree**,

$$\deg f = \frac{1}{2\pi i} \oint_{S^1} f^{-1} \mathrm{d}f. \tag{1.68}$$

For $f_k(e^{i\theta}) = e^{ik\theta}$ we get

$$\deg f_k \overset{(1)}{=} \frac{1}{2\pi i} \int_0^{2\pi} e^{-ik\theta} \frac{\mathrm{d}}{\mathrm{d}\theta}\left(e^{ik\theta}\right) \mathrm{d}\theta$$

$$\overset{(2)}{=} \frac{1}{2\pi i}\int_0^{2\pi} e^{-ik\theta}\,(ik)\,e^{ik\theta}\,\mathrm{d}\theta$$
$$= \frac{ik\cdot 2\pi}{2\pi i} = k.$$

(1) Pull the 1-form $f^{-1}\mathrm{d}f$ back to the parameter $\theta \in [0, 2\pi]$.
(2) Chain rule.

The integrand $f^{-1}\mathrm{d}f = i\,\mathrm{d}(\arg f)$ changes by an exact form under a small deformation of f. So $\deg f$ is a homotopy invariant, and it is always an integer, because $\arg f$ must come back to itself modulo 2π. Conversely, two maps with the same degree are homotopic: interpolate their lifts $\arg f_0$ and $\arg f_1$ linearly. The group law on $[S^1, U(1)]_0$ is pointwise multiplication of phases, and

$$\deg(fg) \overset{(1)}{=} \frac{1}{2\pi i}\oint (fg)^{-1}(\mathrm{d}f\,g + f\,\mathrm{d}g)$$
$$\overset{(2)}{=} \frac{1}{2\pi i}\oint f^{-1}\mathrm{d}f + \frac{1}{2\pi i}\oint g^{-1}\mathrm{d}g = \deg f + \deg g.$$

(1) Leibniz rule for $\mathrm{d}(fg)$.
(2) $U(1)$ is abelian, so $(fg)^{-1} = f^{-1}g^{-1}$ and the factors cancel term by term.

So $H^1(S^1;\mathbb{Z}) \cong \mathbb{Z}$ *as a group*, with pointwise multiplication of maps becoming addition of integers.

Worked example 1.60: $H^1(T^2;\mathbb{Z}) = \mathbb{Z}^2$

Put angles (θ_1, θ_2) on the torus. For every pair of integers (m, n) there is a map

$$f_{m,n} : T^2 \to S^1, \qquad f_{m,n}(\theta_1, \theta_2) = e^{i(m\theta_1 + n\theta_2)}.$$

Restricted to the loop $a = \{\theta_2 = 0\}$ it has degree m. Restricted to the loop $b = \{\theta_1 = 0\}$ it has degree n. Both degrees are homotopy invariants, so different pairs give non-homotopic maps. Every map $T^2 \to S^1$ is homotopic to exactly one $f_{m,n}$: lift the phase to $\mathbb{R}$ after subtracting $m\theta_1 + n\theta_2$, and the remainder is a genuine periodic function, which can be shrunk to zero. Hence

$$H^1(T^2;\mathbb{Z}) = [T^2, S^1]_0 \cong \mathbb{Z}^2, \qquad [f_{m,n}] \mapsto (m, n).$$

Figure 1.36 shows the picture.

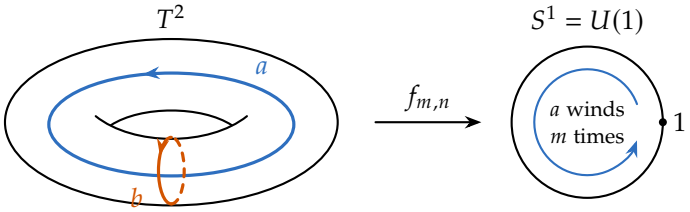


Figure 1.36: A map from the torus to the circle is fixed, up to homotopy, by two integers: how many times the image of the loop a (blue) winds around S^1, and how many times the image of b (orange) does. This is $H^1(T^2;\mathbb{Z}) = \mathbb{Z}^2$, read off as "one integer per independent loop".

Worked example 1.61: $H^1(S^2;\mathbb{Z}) = 0$

Any map $f : S^2 \to S^1$ lifts to $\tilde{f} : S^2 \to \mathbb{R}$, because S^2 is simply connected. Now $t\tilde{f}$, $t \in [0,1]$, shrinks the lift to 0, and $e^{2\pi i t\tilde{f}}$ shrinks f to the constant map. So $[S^2, S^1]_0$ has one element. The sphere has no "winding-number sectors" for $U(1)$-valued fields.

Physical picture: Large gauge transformations

In a $U(1)$ gauge theory on a spatial manifold X, a gauge transformation is a map $g : X \to U(1)$. The ones that can be deformed to the identity are "small". The rest are "large", and they fall into sectors labelled by $[X, U(1)] = H^1(X;\mathbb{Z})$. On a spatial circle these are the winding numbers $k \in \mathbb{Z}$ of $g = e^{ik\theta}$. On a spatial torus they are pairs (m, n), as above. On a sphere there are none.

What about $n = 2$? We will see in Section 1.5.11 that $K(\mathbb{Z}, 2) = \mathbb{CP}^\infty$ is also the *classifying space* of the group $U(1)$. Then (1.66) says

$$H^2(X;\mathbb{Z}) = [X, \mathbb{CP}^\infty]_0 = \{U(1) \text{ bundles over } X\}/\text{iso}.$$

This is the mathematical statement of Dirac quantisation: the magnetic flux through a closed surface is an integer.

1.4.3 Why $[X, K(G,n)]_0$ is a group: loop spaces of $K(G,n)$

A set of homotopy classes is only a group if something supplies a multiplication. Earlier in the chapter we saw the two ways this can happen. Either the *target* is an H-space (it has a product $Y \times Y \to Y$ that is associative and invertible up to homotopy), or the *source* is a coH-space (for example a suspension). For cohomology the source X is arbitrary, so the product must come from the target. Nash's argument is that $K(G,n)$ is itself a loop space, and loop spaces are H-spaces.

The key fact, recalled from the discussion of (1.29), is that looping shifts homotopy groups down by one:

$$\pi_m(\Omega Y) = [S^m, \Omega Y]_0 = [SS^m, Y]_0 = [S^{m+1}, Y]_0 = \pi_{m+1}(Y). \tag{1.69}$$

Apply this to $Y = K(G, n+1)$:

$$\pi_m\big(\Omega K(G,n+1)\big) \overset{(1)}{=} \pi_{m+1}\big(K(G,n+1)\big) \overset{(2)}{=} \begin{cases} G & m+1 = n+1, \\ 0 & \text{otherwise} \end{cases} = \begin{cases} G & m = n, \\ 0 & \text{otherwise.} \end{cases} \tag{1.70}$$

(1) Looping lowers the index by one, (1.69).
(2) Definition (1.65) of $K(G, n+1)$.

So $\Omega K(G, n+1)$ has exactly the homotopy groups of a $K(G,n)$. By uniqueness up to homotopy equivalence,

$$\Omega K(G,n+1) \simeq K(G,n), \qquad \text{and more generally} \qquad \Omega^m K(G,n+m) \simeq K(G,n). \tag{1.71}$$

Now $K(G,n)$ is (up to homotopy) the loop space ΩY with $Y = K(G,n+1)$. Loops can be concatenated, and concatenation is an H-space product. This makes $[X, K(G,n)]_0 = [X, \Omega K(G,n+1)]_0$ a group: two maps $f, g : X \to \Omega Y$ multiply to the map sending x to the loop "first $f(x)$, then $g(x)$". Because $K(G,n)$ is even a *double* loop space, $\Omega^2 K(G,n+2)$, this group is abelian. This is the same argument that made π_n abelian for $n \geq 2$.

Worked example 1.62: Checking (1.71) in the two simplest cases

- $\Omega S^1 = \Omega K(\mathbb{Z},1)$ should be a "$K(\mathbb{Z},0)$", meaning a space whose path components are labelled by $\mathbb{Z}$ and are each contractible. Indeed, a based loop in S^1 lies, up to homotopy, in exactly one winding sector $k \in \mathbb{Z}$, and each sector deforms to the single loop $e^{ik\theta}$.
- $\Omega\mathbb{CP}^\infty = \Omega K(\mathbb{Z},2) \simeq K(\mathbb{Z},1) = S^1$. A loop in $\mathbb{CP}^\infty$ is "the same as" an element of $U(1)$. This reflects the fibration $U(1) \to S^\infty \to \mathbb{CP}^\infty$ of Section 1.5.11: the fibre of the path fibration over a classifying space BG is $\Omega BG \simeq G$.

Physical picture: Spectra and generalised cohomology

A sequence of spaces K_n with $\Omega K_{n+1} \simeq K_n$ is called a **spectrum**. Ordinary cohomology comes from the spectrum $K_n = K(G,n)$. Replace it by the Bott spectrum $\mathbb{Z}\times BU,\ U,\ \mathbb{Z}\times BU,\ U,\dots$ (which obeys $\Omega^2 BU \simeq \mathbb{Z}\times BU$, Bott periodicity) and you get *K-theory*, the cohomology theory of vector bundles that drives the index theorem (Chapters 3 and 4). This is why Nash insists on the $[X,K(G,n)]_0$ viewpoint: it generalises.

1.4.4 Complexes: cohomology by linear algebra

The definition $[X,K(G,n)]_0$ is conceptually clean but hard to compute with: the target is an infinite-dimensional monster. The second definition is built from linear algebra, and it is the one we actually use.

Working definition 1.29 (cochain complex and its cohomology)

A **cochain complex** is a sequence of abelian groups (or vector spaces) E^n and homomorphisms (or linear maps) $d_n : E^n \to E^{n+1}$,

$$\cdots \xrightarrow{d_{n-2}} E^{n-1} \xrightarrow{d_{n-1}} E^n \xrightarrow{d_n} E^{n+1} \xrightarrow{d_{n+1}} \cdots \tag{1.72}$$

with the one crucial property

$$d_n \circ d_{n-1} = 0 \qquad \text{for all } n. \tag{1.73}$$

Elements of $\operatorname{Ker} d_n$ [the $e \in E^n$ with $d_n e = 0$] are **cocycles** ("closed"). Elements of $\operatorname{Im} d_{n-1}$ [those of the form $d_{n-1}f$ with $f \in E^{n-1}$] are **coboundaries** ("exact"). The **n-th cohomology** of the complex is

$$H^n_E = \frac{\operatorname{Ker} d_n}{\operatorname{Im} d_{n-1}} = \frac{\{\text{closed elements of } E^n\}}{\{\text{exact elements of } E^n\}}. \tag{1.74}$$

[two closed elements count as the same class when they differ by an exact one]. *Operational recipe* (finite dimensions, real coefficients): write each d_n as a matrix D_n and compute its rank r_n. Then $\dim H^n_E = \dim E^n - r_n - r_{n-1}$.

The quotient in (1.74) makes sense because every exact element is closed. If $e = d_{n-1}f$ then

$$\begin{aligned} d_n e &\overset{(1)}{=} d_n(d_{n-1}f) \\ &\overset{(2)}{=} 0, \qquad \text{so} \qquad \operatorname{Im} d_{n-1} \subseteq \operatorname{Ker} d_n. \end{aligned} \tag{1.75}$$

(1) e is exact: $e = d_{n-1}f$ for some $f \in E^{n-1}$.

(2) The defining property $d \circ d = 0$, (1.73).

If equality holds, $\operatorname{Im} d_{n-1} = \operatorname{Ker} d_n$, at every n, the sequence is called **exact**. In general it is not, and H^n_E measures by how much exactness fails at the n-th spot. "Cohomology is the obstruction to exactness" is the slogan.

The rank formula in the box follows from rank–nullity:

$$\begin{aligned} \dim H^n_E &\overset{(1)}{=} \dim \operatorname{Ker} d_n - \dim \operatorname{Im} d_{n-1} \\ &\overset{(2)}{=} (\dim E^n - r_n) - r_{n-1}. \end{aligned}$$

(1) The dimension of a quotient of vector spaces is the difference of dimensions.
(2) Rank–nullity for $d_n : E^n \to E^{n+1}$ gives $\dim \operatorname{Ker} d_n = \dim E^n - r_n$. The image of d_{n-1} has dimension r_{n-1} by definition of rank.

Worked example 1.63: A discrete circle: cohomology of a triangle

Take three vertices v_0, v_1, v_2 joined in a loop by edges e_{01}, e_{12}, e_{20}. Let $E^0 = \mathbb{R}^3$ be the functions on vertices and $E^1 = \mathbb{R}^3$ the functions on edges. Define the "discrete gradient" $(d_0 f)(e_{ij}) = f(v_j) - f(v_i)$. All other E^n are zero. In the bases of vertices and edges,

$$D_0 = \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 1 & 0 & -1 \end{pmatrix} \begin{matrix} \leftarrow e_{01} \\ \leftarrow e_{12} \\ \leftarrow e_{20} \end{matrix}$$

The rows add up to zero, so $r_0 \le 2$; the first two rows are independent, so $r_0 = 2$. Then

$$\begin{aligned} \dim H^0 &\overset{(1)}{=} \dim E^0 - r_0 - r_{-1} = 3 - 2 - 0 = 1, \\ \dim H^1 &\overset{(2)}{=} \dim E^1 - r_1 - r_0 = 3 - 0 - 2 = 1. \end{aligned}$$

(1) Rank formula. There is no map into E^0, so $r_{-1} = 0$.
(2) $d_1 : E^1 \to E^2 = 0$ is the zero map, so $r_1 = 0$.

Interpretation: H^0 is spanned by the constant function (gradient zero). H^1 is detected by the "discrete line integral" $a \mapsto a(e_{01}) + a(e_{12}) + a(e_{20})$. Every gradient has zero total, because the sum telescopes around the loop. An edge function with non-zero total is *not* a gradient: it is a discrete $\mathrm{d}\theta$. This is the lattice-gauge-theory picture of a Wilson loop.

Key idea: Euler characteristic of a complex

For a finite complex of finite-dimensional spaces,

$$\sum_n (-1)^n \dim E^n = \sum_n (-1)^n \dim H^n_E. \tag{1.76}$$

The alternating sum is the same before and after taking cohomology.

This follows from the rank formula:

$$\begin{aligned} \sum_n (-1)^n \dim H^n_E &\overset{(1)}{=} \sum_n (-1)^n \dim E^n - \sum_n (-1)^n r_n - \sum_n (-1)^n r_{n-1} \\ &\overset{(2)}{=} \sum_n (-1)^n \dim E^n - \sum_n (-1)^n r_n + \sum_m (-1)^m r_m \end{aligned}$$

$$= \sum_n (-1)^n \dim E^n .$$

(1) Insert $\dim H^n_E = \dim E^n - r_n - r_{n-1}$.
(2) Rename $m = n - 1$ in the last sum, so $(-1)^n = -(-1)^m$.

For the triangle, $3 - 3 = 0 = 1 - 1$. This small identity is the seed of the index theorem. The left side is "combinatorial" (count cells). The right side is "analytic" (count solutions of $\mathrm{d}f = 0$ modulo gauge). In Chapter 4 the complex will be a complex of differential operators, and the right side becomes an index.

1.4.5 The de Rham complex

The most important example for us takes $E^n = \Omega^n(X)$, the smooth n-forms on a manifold X, and $d_n = \mathrm{d}$, the exterior derivative (Chapter T1). We recall the two facts we need.

- In coordinates, d of a p-form $\alpha = \frac{1}{p!}\alpha_{\mu_1\dots\mu_p}\mathrm{d}x^{\mu_1}\wedge\cdots\wedge\mathrm{d}x^{\mu_p}$ is $\mathrm{d}\alpha = \frac{1}{p!}\partial_\nu\alpha_{\mu_1\dots\mu_p}\,\mathrm{d}x^\nu\wedge\mathrm{d}x^{\mu_1}\wedge\cdots$.
- $\mathrm{d}\circ\mathrm{d} = 0$, because mixed partial derivatives are symmetric while $\mathrm{d}x^\nu\wedge\mathrm{d}x^\lambda$ is antisymmetric.

So the de Rham complex

$$0 \to \Omega^0(X) \xrightarrow{\mathrm{d}} \Omega^1(X) \xrightarrow{\mathrm{d}} \Omega^2(X) \xrightarrow{\mathrm{d}} \cdots \xrightarrow{\mathrm{d}} \Omega^{\dim X}(X) \to 0 \tag{1.77}$$

is a cochain complex. Its cohomology

$$H^p_{\mathrm{dR}}(X) = \frac{\{\alpha \in \Omega^p(X) : \mathrm{d}\alpha = 0\}}{\{\mathrm{d}\beta : \beta \in \Omega^{p-1}(X)\}} \tag{1.78}$$

is the **de Rham cohomology** of X. In $\mathbb{R}^3$ the complex is vector calculus in disguise (Figure 1.37).

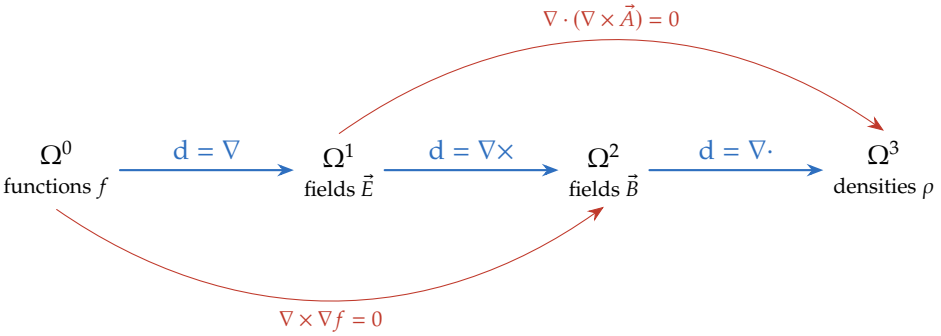


Figure 1.37: The de Rham complex of $\mathbb{R}^3$ is grad, curl, div. The two identities of vector calculus (red) are $\mathrm{d}\circ\mathrm{d} = 0$. On all of $\mathbb{R}^3$ the complex is exact: curl-free fields are gradients and divergence-free fields are curls. On a region with holes it is not, and the failure is the cohomology.

Worked example 1.64: H^0_{dR} counts connected components

There is nothing before Ω^0, so nothing is exact in degree 0, and $H^0_{\mathrm{dR}}(X) = \{f : \mathrm{d}f = 0\}$. A function with zero gradient is constant on each connected piece. So $H^0_{\mathrm{dR}}(X) = \mathbb{R}^c$, where c is the number of connected components. For $X = \mathbb{R}\setminus\{0\}$, $c = 2$ and $H^0_{\mathrm{dR}} = \mathbb{R}^2$.

Worked example 1.65: The Poincaré lemma: $\mathbb{R}^n$ has no cohomology in degree ≥ 1

Let $\alpha = \alpha_\mu \mathrm{d}x^\mu$ be a closed 1-form on $\mathbb{R}^n$: $\partial_\mu \alpha_\nu = \partial_\nu \alpha_\mu$. We construct a potential explicitly by integrating along the straight ray from the origin:

$$f(x) = \int_0^1 x^\mu \, \alpha_\mu(tx) \, \mathrm{d}t.$$

Then

$$\begin{aligned}
\partial_\nu f(x) &\overset{(1)}{=} \int_0^1 \Big[\alpha_\nu(tx) + x^\mu \, t \, (\partial_\nu \alpha_\mu)(tx) \Big] \mathrm{d}t \\
&\overset{(2)}{=} \int_0^1 \Big[\alpha_\nu(tx) + t \, x^\mu (\partial_\mu \alpha_\nu)(tx) \Big] \mathrm{d}t \\
&\overset{(3)}{=} \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t} \Big[t \, \alpha_\nu(tx) \Big] \mathrm{d}t \\
&\overset{(4)}{=} \alpha_\nu(x).
\end{aligned}$$

(1) Differentiate under the integral. The product rule acts on x^μ and on $\alpha_\mu(tx)$, and the chain rule supplies the factor t.
(2) α is closed: $\partial_\nu \alpha_\mu = \partial_\mu \alpha_\nu$.
(3) Product rule: $\frac{\mathrm{d}}{\mathrm{d}t}[t \alpha_\nu(tx)] = \alpha_\nu(tx) + t \, x^\mu (\partial_\mu \alpha_\nu)(tx)$.
(4) Fundamental theorem of calculus: the bracket is $\alpha_\nu(x)$ at $t = 1$ and 0 at $t = 0$.

So $\alpha = \mathrm{d}f$ and $H^1_{\mathrm{dR}}(\mathbb{R}^n) = 0$. The same "integrate along rays" construction works in every degree (and on any star-shaped region), giving $H^p_{\mathrm{dR}}(\mathbb{R}^n) = 0$ for $p \geq 1$. This is the curl test for conservative fields, made precise. It fits the keyidea above: $\mathbb{R}^n$ is contractible, so it has the cohomology of a point.

Worked example 1.66: $H^1_{\mathrm{dR}}(S^1) = \mathbb{R}$

This follows [Dray §20.4]. Every 1-form on S^1 is $\alpha = f(\phi)\,\mathrm{d}\phi$ with f 2π-periodic, and every such form is closed (there are no 2-forms on S^1). Set $c_\alpha = \oint \alpha = \int_0^{2\pi} f \, \mathrm{d}\phi$. If $c_\alpha = 0$, the function

$$g(\phi) = \int_0^\phi f(\phi') \, \mathrm{d}\phi'$$

is periodic, because $g(2\pi) = c_\alpha = 0 = g(0)$. It satisfies $\mathrm{d}g = \alpha$, so α is exact. In general $\alpha - \frac{c_\alpha}{2\pi}\mathrm{d}\phi$ has zero integral, so

$$\alpha = \frac{c_\alpha}{2\pi} \, \mathrm{d}\phi + \mathrm{d}g,$$

and the class of α is fixed by the single number c_α. The form $\mathrm{d}\phi$ is closed but not exact, even though it is written as "d of ϕ". There is no single-valued smooth function ϕ on the whole circle. So $H^1_{\mathrm{dR}}(S^1) = \mathbb{R}$, spanned by $[\mathrm{d}\phi]$.

Worked example 1.67: The punctured plane and Aharonov–Bohm

On $X = \mathbb{R}^2 \setminus \{0\}$ consider

$$\alpha = \frac{x \, \mathrm{d}y - y \, \mathrm{d}x}{x^2 + y^2} \quad (= \mathrm{d}\theta \text{ locally}).$$

With $\alpha_x = -y/r^2$, $\alpha_y = x/r^2$ and $r^2 = x^2 + y^2$,

$$\begin{aligned} \mathrm{d}\alpha &\overset{(1)}{=} (\partial_x \alpha_y - \partial_y \alpha_x)\,\mathrm{d}x \wedge \mathrm{d}y \\ &\overset{(2)}{=} \left(\frac{r^2 - 2x^2}{r^4} + \frac{r^2 - 2y^2}{r^4}\right)\mathrm{d}x \wedge \mathrm{d}y \\ &\overset{(3)}{=} \frac{2r^2 - 2(x^2+y^2)}{r^4}\,\mathrm{d}x \wedge \mathrm{d}y = 0. \end{aligned}$$

(1) d of a 1-form in two dimensions; $\mathrm{d}y \wedge \mathrm{d}x = -\mathrm{d}x \wedge \mathrm{d}y$.
(2) Quotient rule: $\partial_x(x/r^2) = (r^2 - 2x^2)/r^4$ and $\partial_y(-y/r^2) = -(r^2 - 2y^2)/r^4$.
(3) $x^2 + y^2 = r^2$.

So α is closed. On the unit circle, $x = \cos\theta$, $y = \sin\theta$, it pulls back to $\mathrm{d}\theta$, so $\oint \alpha = 2\pi \neq 0$ and α is not exact (exemplar logic: an exact form integrates to zero on a closed curve). Therefore $H^1_{\mathrm{dR}}(\mathbb{R}^2 \setminus \{0\}) \neq 0$. In fact it is $\mathbb{R}$: the punctured plane deformation-retracts onto the circle, and cohomology is a homotopy invariant.

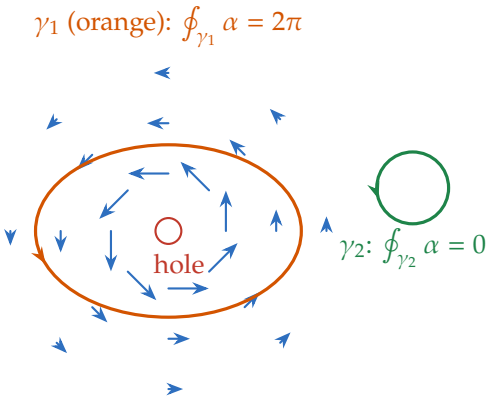


Figure 1.38: The closed form $\alpha = \mathrm{d}\theta$ on the punctured plane (blue arrows show the dual vector field, of strength $1/r$). The loop γ_1 encircles the hole and picks up 2π. The loop γ_2 does not, and picks up 0. The loop integral depends only on the *homology class* of the loop, which is "how many times it encircles the hole". For a solenoid carrying flux Φ, the vector potential outside is $A = \frac{\Phi}{2\pi}\alpha$: closed ($B = 0$) but not exact, and a charge going around γ_1 picks up the Aharonov–Bohm phase $e^{iq\Phi/\hbar}$.

Worked example 1.68: $H^2_{\mathrm{dR}}(S^2) \neq 0$

The area form $\omega = \sin\theta\,\mathrm{d}\theta \wedge \mathrm{d}\phi$ is closed (it is a top form) and $\int_{S^2} \omega = 4\pi \neq 0$, so it is not exact. This is the magnetic field of a monopole. We will show in Example 1.79 that $H^2_{\mathrm{dR}}(S^2) = \mathbb{R}$, spanned by $[\omega]$: *every* closed 2-form on S^2 is $\frac{c}{4\pi}\omega + \mathrm{d}\beta$, where c is its total flux.

Watch out: Closed does not mean exact, and "$\mathrm{d}\phi$" is a lie

The notation $\mathrm{d}\phi$, $\mathrm{d}\theta$ suggests exactness. On the circle, the punctured plane or the sphere, these forms are closed but not exact, because the would-be potential (ϕ, or the Dirac potential) is multivalued or singular somewhere. Cohomology is precisely the bookkeeping of such "potentials that exist locally but not globally".

Reading Nash (pp. 12–13, (1.37)–(1.43))

Nash's §4 opens with Eilenberg–MacLane spaces (1.37), the example $S^1 = K(\mathbb{Z}, 1)$ and the definition $H^n(X; G) = [X, K(G, n)]_0$ (1.38). He notes that $\Omega K(G, n+1)$ is a $K(G, n)$ (1.39), which makes $K(G, n)$ an H-space and $[X, K(G, n)]_0$ a group. He then switches to complexes $\{E^n(X), d_n\}$ with $d_n \circ d_{n-1} = 0$ (1.40), $\operatorname{Im} d_{n-1} \subseteq \operatorname{Ker} d_n$ (1.41), $H_E^n = \operatorname{Ker} d_n / \operatorname{Im} d_{n-1}$ (1.42), and the de Rham complex (1.43), whose cohomology is "of course" isomorphic to $H^n(X; \mathbb{R})$. That "of course" is de Rham's theorem, which we explain in Section 1.4.8. Everything in our examples above (winding numbers, discrete circle, Poincaré lemma, $\mathrm{d}\theta$) is background that Nash assumes.

1.4.6 Homology: the things we integrate over

So far cohomology has been built from things we *integrate*: forms, or functions on edges. The dual objects are the things we integrate *over*: points, curves, surfaces, and formal combinations of them. Stokes' theorem already tells us how the two sides talk to each other:

$$\int_c \mathrm{d}\alpha = \int_{\partial c} \alpha. \tag{1.79}$$

The exterior derivative d on forms is "adjoint" to the boundary operator ∂ on regions. Nash says this in one line: "passing to the dual will generate the corresponding homology groups". In this subsection we build the region side properly and compute with it.

Simplices and their orientation

A **simplex** is the simplest convex shape in each dimension:

- a 0-simplex is a point $[v_0]$;
- a 1-simplex is a segment $[v_0, v_1]$;
- a 2-simplex is a filled triangle $[v_0, v_1, v_2]$;
- a 3-simplex is a solid tetrahedron $[v_0, v_1, v_2, v_3]$.

In general $[v_0, \dots, v_n]$ is the convex hull of $n+1$ points in general position. The **standard n-simplex** is $\Delta^n = \{(t_0, \dots, t_n) \in \mathbb{R}^{n+1} : t_i \geq 0,\ \sum t_i = 1\}$.

The order in which we list the vertices carries an *orientation*. Swapping two vertices reverses it:

$$[v_0, v_1] = -[v_1, v_0], \qquad [v_0, v_1, v_2] = -[v_1, v_0, v_2] = [v_1, v_2, v_0].$$

An even permutation of the vertices gives back the same oriented simplex. An odd one gives its negative. For a triangle, "$v_0 \to v_1 \to v_2$ counterclockwise" is one orientation and clockwise is the other. This is exactly the sign rule of $\mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^n$, and it is no accident: we will integrate n-forms over oriented n-simplices.

The boundary operator

Working definition 1.30 (boundary of a simplex)

$$\partial[v_0, \dots, v_n] = \sum_{i=0}^{n} (-1)^i \, [v_0, \dots, \widehat{v_i}, \dots, v_n], \tag{1.80}$$

where the hat means "omit this vertex". *In words:* the boundary is the sum of the faces [the $(n-1)$-simplices obtained by dropping one vertex], each taken with the orientation it inherits from the simplex [for a triangle, the three edges then run round in the same sense as $v_0 \to v_1 \to v_2$].

The first cases are exactly what our intuition wants:

$$\begin{aligned} \partial[v_0, v_1] &= [v_1] - [v_0] \qquad \text{(end minus start)}, \\ \partial[v_0, v_1, v_2] &= [v_1, v_2] - [v_0, v_2] + [v_0, v_1] \\ &= [v_0, v_1] + [v_1, v_2] + [v_2, v_0] \qquad \text{(walk around the triangle)}. \end{aligned}$$

The second line uses $-[v_0, v_2] = [v_2, v_0]$. Figure 1.39 shows these.

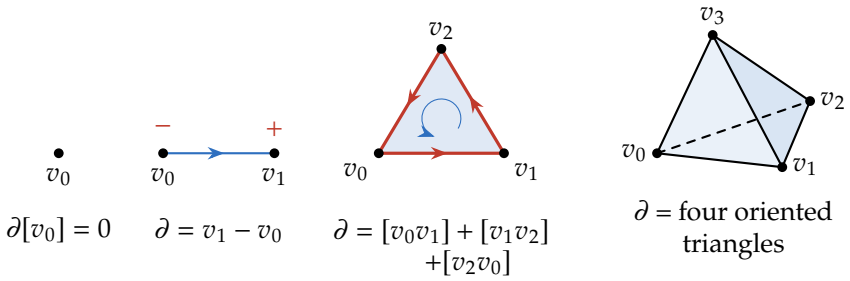


Figure 1.39: Oriented simplices in dimensions 0 to 3 and their boundaries. The boundary of a segment is "end minus start". The boundary of a triangle is the loop around its edge (red), traversed in the direction the orientation (blue circular arrow) dictates. The boundary of a tetrahedron is its four faces, all oriented outward.

Key idea: The boundary of a boundary is zero: $\partial \circ \partial = 0$

Apply ∂ twice to the triangle:

$$\begin{aligned} \partial\partial[v_0, v_1, v_2] &\overset{(1)}{=} \partial[v_1, v_2] - \partial[v_0, v_2] + \partial[v_0, v_1] \\ &\overset{(2)}{=} (v_2 - v_1) - (v_2 - v_0) + (v_1 - v_0) \\ &= 0. \end{aligned}$$

(1) (1.80) with $n = 2$, and linearity of ∂.
(2) (1.80) with $n = 1$ on each edge.

In general, $\partial\partial[v_0, \dots, v_n]$ is a sum over pairs $i < j$ of the face with v_i and v_j both removed. Each such face arises twice: remove v_i first, or remove v_j first. The two terms carry signs $(-1)^i(-1)^{j-1}$ and $(-1)^j(-1)^i$, which are opposite, so they cancel. Geometrically, a closed loop has no endpoints, and the surface of a solid has no edge.

Chains, cycles and boundaries

Let K be a space cut into simplices (a **simplicial complex**): the simplices meet only along whole common faces. An **n-chain** is a formal integer combination $c = \sum_i m_i \sigma_i$ of oriented n-simplices of K. Negative coefficients mean "reversed orientation". Coefficient 2 means "counted twice". The n-chains form an abelian group $C_n(K) \cong \mathbb{Z}^{N_n}$, where N_n is the number of n-simplices. The boundary operator extends linearly to maps $\partial_n : C_n \to C_{n-1}$, and we get a **chain complex**, with arrows pointing *down* in degree:

$$\cdots \xleftarrow{\partial_{n-1}} C_{n-1} \xleftarrow{\partial_n} C_n \xleftarrow{\partial_{n+1}} C_{n+1} \xleftarrow{\partial_{n+2}} \cdots, \qquad \partial_{n-1} \circ \partial_n = 0. \tag{1.81}$$

Working definition 1.31 (simplicial homology)

Cycles are chains with no boundary, $Z_n = \operatorname{Ker} \partial_n$. **Boundaries** are chains that bound something, $B_n = \operatorname{Im} \partial_{n+1}$. Since $\partial\partial = 0$, $B_n \subseteq Z_n$, and the **homology** is

$$H_n(K;\mathbb{Z}) = \frac{Z_n}{B_n} = \frac{\operatorname{Ker} \partial_n}{\operatorname{Im} \partial_{n+1}}. \tag{1.82}$$

[two cycles count as the same class when their difference is a boundary]. *Recipe:* write each ∂_n as an $N_{n-1} \times N_n$ integer matrix. Then

- the **Betti number** is $b_n = \operatorname{rank} H_n$ [the number of independent copies of $\mathbb{Z}$ in H_n], $b_n = N_n - \operatorname{rank} \partial_n - \operatorname{rank} \partial_{n+1}$;
- the **torsion** in H_n is $\mathbb{Z}_{d_1} \oplus \mathbb{Z}_{d_2} \oplus \cdots$, where the $d_i > 1$ are the non-unit diagonal entries (**elementary divisors**) in the Smith normal form of ∂_{n+1}. That form is a diagonalisation by integer row and column operations.

Physically, a cycle is a closed curve or surface, possibly with multiplicities. It is homologous to zero when it is the edge of something. H_1 asks: "which closed loops are not the edge of any surface?" H_2 asks the same for closed surfaces and solids.

Worked example 1.69: Homology of the circle (a hollow triangle)

Let K be the three vertices and three edges of a triangle, with no face. Use the edges $e_{01} = [v_0, v_1]$, $e_{12} = [v_1, v_2]$, $e_{20} = [v_2, v_0]$. Then $\partial e_{ij} = v_j - v_i$, and in the bases (e_{01}, e_{12}, e_{20}) and (v_0, v_1, v_2)

$$\partial_1 = \begin{pmatrix} -1 & 0 & 1 \\ 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix}.$$

This is the *transpose* of the matrix D_0 of Example 1.63. The boundary is the adjoint of the discrete gradient, as the Stokes pairing (1.79) predicts. The rank is 2, and there are no 2-simplices, so

$$b_0 = N_0 - \operatorname{rank} \partial_0 - \operatorname{rank} \partial_1 = 3 - 0 - 2 = 1,$$
$$b_1 = N_1 - \operatorname{rank} \partial_1 - \operatorname{rank} \partial_2 = 3 - 2 - 0 = 1.$$

The kernel of ∂_1 is spanned by $z = e_{01} + e_{12} + e_{20}$: the loop once around. Nothing bounds it, since there are no triangles. Hence $H_1(S^1;\mathbb{Z}) = \mathbb{Z}$, generated by the fundamental loop, and $H_0 = \mathbb{Z}$.

Worked example 1.70: Homology of S^2 as the surface of a tetrahedron

The hollow tetrahedron on vertices $0,1,2,3$ is a triangulated sphere. It has $N_0 = 4$ vertices, $N_1 = 6$ edges $e_{01}, e_{02}, e_{03}, e_{12}, e_{13}, e_{23}$, and $N_2 = 4$ triangles $f_{012}, f_{013}, f_{023}, f_{123}$ (vertices always listed in increasing order). From (1.80),

$$\partial f_{012} = e_{12} - e_{02} + e_{01}, \qquad \partial f_{013} = e_{13} - e_{03} + e_{01},$$
$$\partial f_{023} = e_{23} - e_{03} + e_{02}, \qquad \partial f_{123} = e_{23} - e_{13} + e_{12}.$$

Reading off columns, the boundary matrices are

$$\partial_2 = \begin{pmatrix} 1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & -1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & 1 \end{pmatrix} \begin{matrix} e_{01} \\ e_{02} \\ e_{03} \\ e_{12}, \\ e_{13} \\ e_{23} \end{matrix} \qquad \partial_1 = \begin{pmatrix} -1 & -1 & -1 & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & -1 & 0 \\ 0 & 1 & 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix} \begin{matrix} v_0 \\ v_1 \\ v_2 \\ v_3 \end{matrix}.$$

Kernel of ∂_2. Put $c = af_{012} + bf_{013} + c'f_{023} + df_{123}$. The rows give $a + b = 0$, $-a + c' = 0$, $-b - c' = 0$, $a + d = 0$, $b - d = 0$, $c' + d = 0$. Hence $b = -a$, $c' = a$, $d = -a$, and all six equations hold. So

$$\operatorname{Ker} \partial_2 = \mathbb{Z} \cdot \big(f_{012} - f_{013} + f_{023} - f_{123}\big), \qquad \operatorname{rank} \partial_2 = 3.$$

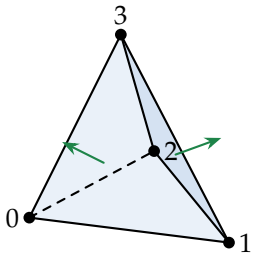
This cycle is (minus) the boundary of the solid tetrahedron $[0,1,2,3]$: all four faces, coherently oriented. That is the sphere itself, the **fundamental class**.
Rank of ∂_1. The graph of edges is connected, with 4 vertices, so $\operatorname{rank} \partial_1 = 4 - 1 = 3$. (The rows sum to zero, and any three are independent.)
Betti numbers.

$$b_0 = 4 - 0 - 3 = 1, \qquad b_1 = 6 - 3 - 3 = 0, \qquad b_2 = 4 - 3 - 0 = 1.$$

Torsion. The 3×3 minor of ∂_2 in rows e_{01}, e_{02}, e_{13} and columns $f_{012}, f_{013}, f_{023}$ is $\det\left(\begin{smallmatrix} 1 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & 1 & 0 \end{smallmatrix}\right) = -1$. So the gcd of the 3×3 minors is 1, all elementary divisors are 1, and there is no torsion. Therefore

$$H_0(S^2;\mathbb{Z}) = \mathbb{Z}, \qquad H_1(S^2;\mathbb{Z}) = 0, \qquad H_2(S^2;\mathbb{Z}) = \mathbb{Z}.$$

Every closed loop on the sphere bounds (it is the edge of a cap), but the sphere as a whole bounds nothing inside itself. Check: $N_0 - N_1 + N_2 = 4 - 6 + 4 = 2 = b_0 - b_1 + b_2$.



$[S^2] = \pm\big(f_{012} - f_{013} + f_{023} - f_{123}\big)$
all faces oriented outward: a 2-cycle bounding nothing

Figure 1.40: The surface of a tetrahedron is a triangulated sphere. The alternating sum of its four faces, oriented so that the normals (green) all point outward, is the unique (up to sign) 2-cycle. It generates $H_2(S^2;\mathbb{Z}) = \mathbb{Z}$. Integrating a 2-form over this cycle is integrating over the sphere.

The torus by boundary matrices

For the torus it pays to relax the rules slightly. We allow a triangle's vertices to be glued to each other, as happens when we build the torus from a square by identifying opposite sides. Such a decomposition is called a **Δ-complex**. The recipe (1.80) and all the linear algebra are unchanged.

Worked example 1.71: Homology of T^2 from two triangles

Take the unit square with bottom and top edges identified (both called a, pointing right) and left and right edges identified (both called b, pointing up). Draw the diagonal c from the bottom-left to the top-right corner. All four corners become *one* vertex v. There are three edges a, b, c and two triangles

$$L = [(0,0),(1,0),(1,1)] \text{ (lower right)}, \qquad U = [(0,0),(0,1),(1,1)] \text{ (upper left)}.$$

(Figure 1.41.) Apply (1.80), reading off which edge each face is:

$$\partial L \overset{(1)}{=} [(1,0),(1,1)] - [(0,0),(1,1)] + [(0,0),(1,0)] = b - c + a,$$
$$\partial U \overset{(2)}{=} [(0,1),(1,1)] - [(0,0),(1,1)] + [(0,0),(0,1)] = a - c + b,$$
$$\partial a \overset{(3)}{=} v - v = 0, \qquad \partial b = \partial c = 0.$$

(1) The right side of the square is b, the diagonal is c, the bottom is a.
(2) The top side is a (identified with the bottom), the diagonal is c, the left side is b.
(3) Every edge starts and ends at the single vertex v.

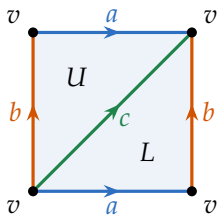
So, in the bases (U, L), (a, b, c) and (v),

$$\partial_2 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ -1 & -1 \end{pmatrix}, \qquad \partial_1 = \begin{pmatrix} 0 & 0 & 0 \end{pmatrix}.$$

Now read off the homology.

- H_0: $\operatorname{Ker}\partial_0 = \mathbb{Z}v$ and $\operatorname{Im}\partial_1 = 0$, so $H_0 = \mathbb{Z}$.
- H_1: $\operatorname{Ker}\partial_1 = \mathbb{Z}^3$ (everything) and $\operatorname{Im}\partial_2 = \mathbb{Z}\cdot(a+b-c)$. Change basis of $\mathbb{Z}^3$ to $\{a,\ b,\ a+b-c\}$. This is allowed, since the matrix taking (a,b,c) to it has determinant -1. Then
$$H_1(T^2;\mathbb{Z}) = \frac{\mathbb{Z}a \oplus \mathbb{Z}b \oplus \mathbb{Z}(a+b-c)}{\mathbb{Z}(a+b-c)} \cong \mathbb{Z}a \oplus \mathbb{Z}b = \mathbb{Z}^2.$$
The diagonal c is homologous to $a+b$: going diagonally is the same, up to a boundary, as going along a and then b.
- H_2: $\operatorname{Ker}\partial_2 = \{xU + yL : x + y = 0\} = \mathbb{Z}(L-U)$, so $H_2(T^2;\mathbb{Z}) = \mathbb{Z}$, generated by the fundamental class $[T^2] = L - U$. Checking orientations directly from the vertex order (the cross product of consecutive edge vectors), $L = [(0,0),(1,0),(1,1)]$ is counterclockwise but $U = [(0,0),(0,1),(1,1)]$ is clockwise; to combine the two triangles into one consistently counterclockwise-oriented cycle we must flip U, giving $L + (-U) = L - U$ rather than $U - L$. (Either sign generates the same $\mathbb{Z}$; we fix this one because it is the sign used consistently in Example 1.81, where $(\alpha \cup \beta)[T^2] = +1$ matches $\int_{T^2} \omega_1 \wedge \omega_2 = +1$.)

Euler characteristic check: $N_0 - N_1 + N_2 = 1 - 3 + 2 = 0 = 1 - 2 + 1$.



$$\partial_2 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ -1 & -1 \end{pmatrix}$$

$$\partial_1 = (0\ 0\ 0)$$

$$H_* = \mathbb{Z},\ \mathbb{Z}^2,\ \mathbb{Z}$$

Figure 1.41: The torus as a square with opposite sides glued (same letter, same arrow). The diagonal c cuts it into two triangles L and U. All four corners are one vertex. The two small boundary matrices give the whole homology of the torus. The generators of H_1 are the two loops a and b of Figure 1.36.

Watch out: A real triangulation needs more simplices

In a genuine simplicial complex, two different simplices cannot have the same vertex set. The two-triangle torus violates this badly. The smallest honest triangulation of T^2 (Möbius, Császár) has 7 vertices, 21 edges and 14 triangles. The obvious one, a 3×3 grid of squares each cut by a diagonal, has 9 vertices, 27 edges and 18 triangles. Its boundary matrices are 9×27 and 27×18. That is fine for a computer, and it gives the same homology: $9 - 27 + 18 = 0$, $\operatorname{rank} \partial_1 = 8$, $\operatorname{rank} \partial_2 = 17$, so $b_0 = 1$, $b_1 = 27 - 8 - 17 = 2$, $b_2 = 18 - 17 = 1$. The answer does not depend on the triangulation. That is the content of the statement that homology is a topological invariant.

Cells instead of simplices: T^2, $\mathbb{RP}^2$ and the Klein bottle in three lines

The torus computation suggests a further shortcut. Why cut the square into triangles at all? A **CW** or **cell complex** is built by gluing discs ("cells") of increasing dimension. The **cellular chain complex** has one generator per cell. The boundary of a 2-cell is the word you read walking around its edge, written additively: each edge counts +1 if you traverse it along its arrow and -1 against it. Cellular homology agrees with simplicial homology, and it is usually far quicker.

Worked example 1.72: Three surfaces from a square

Each surface below has one 2-cell D (the square) and 1-cells a, b.

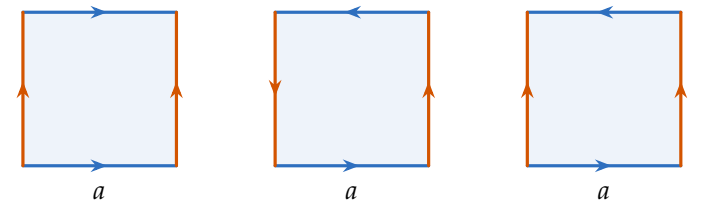
- *Torus.* Boundary word $aba^{-1}b^{-1}$, one vertex. Then $\partial D = a + b - a - b = 0$ and $\partial a = \partial b = 0$. So $H_0 = \mathbb{Z}$, $H_1 = \mathbb{Z}^2$, $H_2 = \mathbb{Z}$, as before.
- *Real projective plane* $\mathbb{RP}^2$ (the disc with antipodal boundary points identified). Boundary word aa (the square with a along the bottom and a reversed along the top, b likewise, collapses to one 1-cell). One vertex, one edge a, one face D, with $\partial D = 2a$ and $\partial a = 0$. Then

$$H_2 \overset{(1)}{=} \operatorname{Ker}(\partial_2 : \mathbb{Z} \to \mathbb{Z},\ 1 \mapsto 2) = 0,$$
$$H_1 \overset{(2)}{=} \mathbb{Z}a/2\mathbb{Z}a = \mathbb{Z}_2,$$
$$H_0 = \mathbb{Z}.$$

  (1) Multiplication by 2 on $\mathbb{Z}$ is injective. No 2-cycle exists, because $\mathbb{RP}^2$ is not orientable.
  (2) Every 1-chain is a cycle. The boundaries are the even multiples of a.

  The loop a is not a boundary, but *twice* it is. This is **torsion**.

- *Klein bottle* K. Boundary word $aba^{-1}b$, one vertex. $\partial D = a + b - a + b = 2b$, so $H_2 = 0$ and $H_1 = (\mathbb{Z}a \oplus \mathbb{Z}b)/\langle 2b \rangle = \mathbb{Z} \oplus \mathbb{Z}_2$.



torus T^2 $H_1 = \mathbb{Z}^2$, $H_2 = \mathbb{Z}$ | $\mathbb{RP}^2$ $H_1 = \mathbb{Z}_2$, $H_2 = 0$ | Klein bottle $H_1 = \mathbb{Z} \oplus \mathbb{Z}_2$, $H_2 = 0$

Figure 1.42: Three surfaces from one square, differing only in the arrows (blue sides are identified with each other, and so are orange sides). The torus keeps both pairs parallel. $\mathbb{RP}^2$ reverses both. The Klein bottle reverses one. A reversed pair makes the boundary of the square count an edge twice instead of zero times. That produces $\mathbb{Z}_2$ torsion in H_1 and kills H_2: the surface is non-orientable.

Coefficients and torsion

We can use coefficients other than $\mathbb{Z}$. Allow chains $\sum g_i \sigma_i$ with g_i in any abelian group G, such as $\mathbb{R}$, $\mathbb{Q}$ or $\mathbb{Z}_2$. The same matrices now act on G^{N_n}, and we get $H_n(K; G)$.

- Over $\mathbb{R}$ (or any field of characteristic zero) torsion disappears, because 2 is invertible, and $H_n(K; \mathbb{R}) = \mathbb{R}^{b_n}$. For $\mathbb{RP}^2$: $H_1(\mathbb{RP}^2; \mathbb{R}) = \mathbb{R}/2\mathbb{R} = 0$.
- Over $\mathbb{Z}_2$, signs do not matter and orientation is irrelevant. For $\mathbb{RP}^2$ the map $\partial D = 2a = 0$ vanishes, so $H_1(\mathbb{RP}^2; \mathbb{Z}_2) = \mathbb{Z}_2$ and $H_2(\mathbb{RP}^2; \mathbb{Z}_2) = \mathbb{Z}_2$. Non-orientable surfaces still have a $\mathbb{Z}_2$ fundamental class.

The general rule is the **universal coefficient theorem**: integral homology determines homology with any coefficients. In particular $b_n = \dim_{\mathbb{R}} H_n(K; \mathbb{R})$ is the rank of $H_n(K; \mathbb{Z})$, and torsion is invisible to real numbers, hence to differential forms. This is why de Rham cohomology cannot see $\mathbb{Z}_2$ phenomena such as spin structures or the Kirby–Siebenmann class of Section 1.6.

Singular homology

Simplicial homology needs a triangulation, and it is not obvious that the answer is independent of the triangulation chosen. **Singular homology** removes the choice by brute force.

Working definition 1.32 (singular homology)

A **singular n-simplex** in a space X is *any* continuous map $\sigma : \Delta^n \to X$ [Δ^n is the standard n-simplex, a point, segment, triangle, ...]. It can be crumpled, degenerate or self-intersecting. Singular n-chains are finite integer combinations $\sum m_i \sigma_i$. The boundary is $\partial\sigma = \sum_i (-1)^i \, \sigma|_{i\text{-th face}}$, with the faces parametrised as in (1.80). Then $\partial\partial = 0$ as before, and $H_n(X; \mathbb{Z}) = \operatorname{Ker} \partial_n / \operatorname{Im} \partial_{n+1}$. *Use:* never compute with this directly. Use it to know that homology is defined for every space, is a topological invariant, and is homotopy invariant [homotopy equivalent spaces have the same H_n]. Then compute simplicially or cellularly [from a decomposition into balls glued along their boundaries].

Three facts make singular homology the right foundation:

1. *Functoriality.* A continuous map $f : X \to Y$ pushes singular simplices forward, $\sigma \mapsto f \circ \sigma$. This commutes with ∂, so it induces $f_* : H_n(X) \to H_n(Y)$. A homeomorphism therefore induces an isomorphism.
2. *Homotopy invariance.* If $f \simeq g$ via $F : [0,1] \times X \to Y$, then the "prism" $F \circ (\mathrm{id} \times \sigma)$ swept out by a simplex has boundary $g \circ \sigma - f \circ \sigma$ plus the prism over $\partial\sigma$. So a cycle is moved to a homologous cycle, and $f_* = g_*$. Hence homology, like cohomology, is an invariant of homotopy type.
3. *Agreement.* For a simplicial (or cell) complex, singular, simplicial and cellular homology coincide.

Worked example 1.73: H_0 counts path components

A singular 1-simplex is a path $\sigma : [0,1] \to X$ with $\partial\sigma = \sigma(1) - \sigma(0)$. So two points joined by a path are homologous 0-chains. Conversely, the "total coefficient in each path component" is unchanged by adding boundaries. Hence $H_0(X;\mathbb{Z}) = \mathbb{Z}^c$, where c is the number of path components. This is the homology twin of Example 1.64.

Worked example 1.74: A loop is a 1-cycle

A based loop $\gamma : [0,1] \to X$ with $\gamma(0) = \gamma(1)$ is a singular 1-simplex with $\partial\gamma = \gamma(1) - \gamma(0) = 0$, so it is a cycle and defines a class $[\gamma] \in H_1(X;\mathbb{Z})$. Homotopic loops give the same class. Concatenation gives the sum, because the triangle "γ_1, then γ_2, back along $\gamma_1 \cdot \gamma_2$" is a singular 2-simplex with boundary $\gamma_2 - \gamma_1 \cdot \gamma_2 + \gamma_1$. This defines the **Hurewicz map** $h : \pi_1(X) \to H_1(X;\mathbb{Z})$, which we study in Section 1.4.10.

Physical picture: Integration is the pairing between chains and forms

For a smooth p-form α and a (smooth) singular p-chain $c = \sum m_i \sigma_i$, define

$$\langle \alpha, c \rangle = \int_c \alpha = \sum_i m_i \int_{\Delta^p} \sigma_i^* \alpha.$$

Stokes' theorem for each simplex gives $\langle \mathrm{d}\alpha, c \rangle = \langle \alpha, \partial c \rangle$. In other words, d is the transpose of ∂. Consequences:

- closed form, boundary chain: $\langle \alpha, \partial b \rangle = \langle \mathrm{d}\alpha, b \rangle = 0$;
- exact form, cycle: $\langle \mathrm{d}\beta, z \rangle = \langle \beta, \partial z \rangle = 0$.

So integration descends to a well-defined pairing $H^p_{\mathrm{dR}}(X) \times H_p(X;\mathbb{R}) \to \mathbb{R}$. De Rham's theorem says this pairing is perfect.

Watch out: Homology is not cohomology

Nash writes that singular homology "is isomorphic to the integral cohomology group $H_n(X;\mathbb{Z})$". Read this as "is the integral *homology* group $H_n(X;\mathbb{Z})$". It is a naming, not a theorem. Homology and cohomology are genuinely different in general. For $\mathbb{RP}^2$:

$$H_0, H_1, H_2 = \mathbb{Z},\ \mathbb{Z}_2,\ 0, \qquad \text{but} \qquad H^0, H^1, H^2 = \mathbb{Z},\ 0,\ \mathbb{Z}_2.$$

The torsion moves up one degree. (This is the universal coefficient theorem for cohomology: $H^n \cong \mathrm{free}(H_n) \oplus \mathrm{torsion}(H_{n-1})$.) The two are related by Poincaré duality on closed oriented manifolds, $H^n(M;\mathbb{Z}) \cong H_{\dim M - n}(M;\mathbb{Z})$. For real coefficients the distinction is just vector-

space duality, $H^n(X;\mathbb{R}) = H_n(X;\mathbb{R})^*$.

1.4.7 Cochains: cohomology from a triangulation

We now have two complexes that look like mirror images. Forms with d raise degree. Chains with ∂ lower it. Stokes' theorem (1.79) says d is the transpose of ∂. We can turn this around and *define* a cohomology from a triangulation by transposing the boundary matrices. This is what Nash means when he says the two complexes are dual.

Working definition 1.33 (simplicial cochains and coboundary)

Let K be a simplicial (or Δ-, or cell) complex [a space glued from simplices, or from balls, along their faces] and G an abelian group. An **n-cochain** with values in G is a function that assigns an element of G to each oriented n-simplex, extended linearly to chains: $C^n(K;G) = \mathrm{Hom}(C_n(K), G) \cong G^{N_n}$ [one value in G for each of the N_n simplices, and $\varphi(c + c') = \varphi(c) + \varphi(c')$ on chains]. The **coboundary** $\delta : C^n \to C^{n+1}$ is defined by

$$(\delta\varphi)(\sigma) = \varphi(\partial\sigma) \qquad \text{for every } (n+1)\text{-simplex } \sigma. \tag{1.83}$$

In matrix form, *the matrix of δ_n is the transpose of the matrix of ∂_{n+1}*. The **simplicial cohomology** is $H^n(K;G) = \mathrm{Ker}\,\delta_n / \mathrm{Im}\,\delta_{n-1}$ [cochains with $\delta\varphi = 0$, counted as the same when they differ by some $\delta\psi$]. *Recipe:* transpose the boundary matrices you already have, and compute kernels and images again.

Because δ is a transpose, $\delta \circ \delta = 0$ is automatic:

$$\begin{aligned}(\delta\delta\varphi)(\tau) &\overset{(1)}{=} (\delta\varphi)(\partial\tau)\\ &\overset{(2)}{=} \varphi(\partial\partial\tau)\\ &\overset{(3)}{=} 0.\end{aligned}$$

(1) Definition (1.83), applied to $\delta\varphi$.
(2) Definition (1.83) again, applied to each simplex in the chain $\partial\tau$ (linearity).
(3) The boundary of a boundary is zero.

We already met the simplest case without saying so: the discrete gradient of Example 1.63 is exactly δ_0 on the hollow triangle, and its matrix D_0 is the transpose of ∂_1 in Example 1.69.

Worked example 1.75: Integral cohomology of the torus by transposing

Use the two-triangle torus of Example 1.71, with bases (v), (a,b,c), (U,L). Transposing $\partial_1 = (0\ 0\ 0)$ and ∂_2 gives

$$\delta_0 = \begin{pmatrix}0\\0\\0\end{pmatrix}, \qquad \delta_1 = \partial_2^T = \begin{pmatrix}1 & 1 & -1\\ 1 & 1 & -1\end{pmatrix}.$$

A 1-cochain is a triple $\varphi = (\varphi(a), \varphi(b), \varphi(c)) = (x,y,z)$. Then

$$\begin{aligned}H^0(T^2;\mathbb{Z}) &\overset{(1)}{=} \mathrm{Ker}\,\delta_0 = \mathbb{Z},\\ H^1(T^2;\mathbb{Z}) &\overset{(2)}{=} \{(x,y,z) \in \mathbb{Z}^3 : x + y - z = 0\}/0 \cong \mathbb{Z}^2,\\ H^2(T^2;\mathbb{Z}) &\overset{(3)}{=} \mathbb{Z}^2/\mathrm{Im}\,\delta_1 = \mathbb{Z}^2/\mathbb{Z}(1,1) \cong \mathbb{Z}.\end{aligned}$$

(1) $\delta_0 = 0$, so every 0-cochain (a single integer on the single vertex) is a cocycle.
(2) The cocycle condition is $(\delta_1\varphi)(U) = (\delta_1\varphi)(L) = x + y - z = 0$, so $z = x + y$ and (x, y) are free integers. Nothing is exact, because $\delta_0 = 0$.
(3) Every 2-cochain is a cocycle (there are no 3-simplices). The image of δ_1 is spanned by $(1, 1)$, and $(p, q) \mapsto p + q$ identifies the quotient with $\mathbb{Z}$.

A 1-cocycle is a rule "how much phase do I pick up along a, along b, and along the diagonal c". The cocycle condition says the phase around the edge of each triangle vanishes: *zero flux through each plaquette*, in lattice-gauge language. The two free integers are the two holonomies around the two non-contractible loops. They match $H^1(T^2;\mathbb{Z}) = [T^2, S^1]_0 = \mathbb{Z}^2$ of Example 1.60, where the pair (m, n) were the windings along a and b.

Worked example 1.76: Torsion moves up: the cohomology of $\mathbb{RP}^2$

The cell complex of $\mathbb{RP}^2$ in Example 1.72 has one cell in each dimension, with $\partial_2 = (2)$ and $\partial_1 = (0)$. Transposing, $\delta_0 = (0)$ and $\delta_1 = (2)$. Hence

$$H^0 = \mathbb{Z}, \qquad H^1 = \mathrm{Ker}(\times 2) = 0, \qquad H^2 = \mathbb{Z}/2\mathbb{Z} = \mathbb{Z}_2.$$

Compare $H_1 = \mathbb{Z}_2$, $H_2 = 0$. This is the "torsion moves up one degree" of the pitfall at the end of the homology subsection, now seen as nothing more than the difference between the kernel and the cokernel of the 1×1 matrix (2).

1.4.8 The de Rham theorem: forms know the real topology

Nash asserts that de Rham cohomology is "of course" isomorphic to $H^n(X;\mathbb{R})$. This is **de Rham's theorem**. It is the reason a physicist can do topology with differential forms. Here is what it says and why it is true.

The de Rham map

Triangulate the manifold X by smooth simplices. Given a p-form α, build a real p-cochain by integrating:

$$I(\alpha)(\sigma) = \int_\sigma \alpha \qquad \text{for each oriented } p\text{-simplex } \sigma. \tag{1.84}$$

The map $I : \Omega^p(X) \to C^p(K;\mathbb{R})$ is the **de Rham map**. It intertwines the two differentials:

$$\begin{aligned} I(\mathrm{d}\alpha)(\tau) &\overset{(1)}{=} \int_\tau \mathrm{d}\alpha \\ &\overset{(2)}{=} \int_{\partial\tau} \alpha \\ &\overset{(3)}{=} I(\alpha)(\partial\tau) \\ &\overset{(4)}{=} (\delta I(\alpha))(\tau). \end{aligned} \tag{1.85}$$

(1) Definition (1.84), for a $(p+1)$-simplex τ.
(2) Stokes' theorem on the simplex τ.
(3) The integral over the chain $\partial\tau$ is the sum of integrals over its faces with their signs, which is the linear extension of $I(\alpha)$.
(4) Definition (1.83) of the coboundary.

So $I \circ \mathrm{d} = \delta \circ I$. A map between complexes that commutes with the differentials is called a **chain map**. It sends closed to closed and exact to exact, so it induces a map on cohomology.

Key idea: de Rham's theorem

For a smooth manifold X the de Rham map induces an isomorphism

$$I_* : H^p_{\mathrm{dR}}(X) \xrightarrow{\cong} H^p(X;\mathbb{R}) \qquad \text{for all } p. \tag{1.86}$$

Equivalently, in terms of the integration pairing with cycles:

1. a closed p-form is exact *if and only if* all its **periods** $\oint_z \alpha$ over p-cycles z vanish;
2. any assignment of real periods to a basis of $H_p(X;\mathbb{Z})$ is realised by some closed p-form.

Half of statement 1 is the exemplar argument: exact forms have zero periods by Stokes. The other half is the real content. Here is why it is true, in the style of the Poincaré lemma.

- *Locally* there is nothing to prove. On a small ball every closed form is exact (Example 1.65), and the ball's cohomology is that of a point. So both theories agree on each simplex's star.
- *Gluing* works the same way in both theories: both satisfy the Mayer–Vietoris sequence of Section 1.4.9 below.
- Build X up one open set at a time. At each stage the two long exact sequences are compared by I. The "five lemma" of linear algebra then transfers the isomorphism from the pieces to the union.

So de Rham's theorem is "the Poincaré lemma plus gluing". Weil's proof via Čech cohomology makes this into an honest argument, and we will meet that machinery in Chapter 3.

$$\begin{array}{ccccccc}
\cdots & \xrightarrow{\mathrm{d}} & \Omega^p(X) & \xrightarrow{\mathrm{d}} & \Omega^{p+1}(X) & \xrightarrow{\mathrm{d}} & \cdots \\
 & & \downarrow{\scriptstyle I=\int} & & \downarrow{\scriptstyle I=\int} & & \\
\cdots & \xrightarrow[\delta]{} & C^p(K;\mathbb{R}) & \xrightarrow[\delta=\partial^T]{} & C^{p+1}(K;\mathbb{R}) & \xrightarrow[\delta]{} & \cdots
\end{array}$$

Figure 1.43: The de Rham map is a ladder between two complexes. Going across then down (differentiate, then integrate over a simplex) equals going down then across (integrate over the faces of the simplex). That square is Stokes' theorem, (1.85). De Rham's theorem says the ladder induces an isomorphism on cohomology: the smooth and the combinatorial pictures of real topology are the same.

Worked example 1.77: The period matrix of the torus

On T^2 with angles $\theta_1, \theta_2 \in [0, 2\pi)$, take the closed forms $\omega_1 = \mathrm{d}\theta_1/2\pi$ and $\omega_2 = \mathrm{d}\theta_2/2\pi$, and the cycles a ($\theta_2 = 0$, θ_1 running from 0 to 2π) and b ($\theta_1 = 0$). The periods are

$$\oint_a \omega_1 \overset{(1)}{=} \frac{1}{2\pi}\int_0^{2\pi} \mathrm{d}\theta_1 = 1, \qquad \oint_b \omega_1 \overset{(2)}{=} 0,$$
$$\oint_a \omega_2 = 0, \qquad \oint_b \omega_2 = 1.$$

(1) Parametrise a by θ_1.

(2) Along b, θ_1 is constant, so $\mathrm{d}\theta_1$ pulls back to zero.

The **period matrix** $\left(\oint_{z_j} \omega_i\right)$ is the identity. By statement 1 of de Rham's theorem, ω_1 and ω_2 are linearly independent in cohomology (a combination $x\omega_1 + y\omega_2$ with x, y not both zero has a non-zero period). By statement 2 they span: any closed 1-form α with periods $A = \oint_a \alpha$, $B = \oint_b \alpha$ satisfies

$$\alpha = A\,\omega_1 + B\,\omega_2 + \mathrm{d}f$$

for a periodic function f. So $H^1_{\mathrm{dR}}(T^2) = \mathbb{R}^2$. Its *integral lattice*, the forms with integer periods, is $\mathbb{Z}\omega_1 \oplus \mathbb{Z}\omega_2$. This is $H^1(T^2;\mathbb{Z})$ sitting inside $H^1(T^2;\mathbb{R})$, and it is exactly the set of forms $\frac{1}{2\pi i} f^{-1}\mathrm{d}f$ coming from maps $f_{m,n} : T^2 \to U(1)$ in Example 1.60.

Worked example 1.78: A form on the punctured plane is fixed by one number

On $\mathbb{R}^2 \setminus \{0\}$, $H_1 = \mathbb{Z}$ is generated by the unit circle γ. De Rham's theorem says: a closed 1-form β is exact exactly when $\oint_\gamma \beta = 0$, and in general $\beta - \frac{1}{2\pi}\left(\oint_\gamma \beta\right)\alpha$ is exact, with α the angle form of Example 1.67. For a static magnetic field confined to a solenoid, β is the vector potential outside, and the only gauge-invariant information it carries in the field-free region is the single period $\oint_\gamma A = \Phi$, the enclosed flux.

Watch out: What de Rham cannot see

De Rham cohomology is cohomology with *real* coefficients. It sees the Betti numbers and nothing else. The $\mathbb{Z}_2$ in $H^2(\mathbb{RP}^2;\mathbb{Z})$ is invisible to forms, since $H^2_{\mathrm{dR}}(\mathbb{RP}^2) = 0$. So are all the $\mathbb{Z}_2$-valued invariants of Section 1.6 (Stiefel–Whitney classes, the Kirby–Siebenmann class). Integrality, too, is extra information. De Rham tells you that $[F/2\pi] \in H^2_{\mathrm{dR}}(S^2)$, but the fact that its period is an integer (Dirac quantisation) comes from $H^2(S^2;\mathbb{Z})$ and the bundle structure, not from the forms.

1.4.9 Gluing: the Mayer–Vietoris sequence

How do we compute cohomology of a space we cannot triangulate by hand, like S^n for large n? We cut it into simple pieces and glue. The bookkeeping of gluing is an exact sequence.

Let $X = U \cup V$ with U, V open. There are two natural maps of forms:

$$0 \to \Omega^p(X) \xrightarrow{\ r\ } \Omega^p(U) \oplus \Omega^p(V) \xrightarrow{\ s\ } \Omega^p(U \cap V) \to 0, \tag{1.87}$$
$$r(\omega) = (\omega|_U, \omega|_V), \qquad s(\alpha,\beta) = \beta|_{U\cap V} - \alpha|_{U\cap V}.$$

This short sequence is exact. r is injective, since a form that vanishes on U and on V vanishes. $\operatorname{Im} r = \operatorname{Ker} s$, since two forms that agree on the overlap glue to a global form. s is surjective: this is where a **partition of unity** enters. Choose smooth functions $\rho_U, \rho_V \geq 0$ with $\rho_U + \rho_V = 1$, $\operatorname{supp}\rho_U \subset U$ and $\operatorname{supp}\rho_V \subset V$. Given ω on $U \cap V$, set

$$\alpha = -\rho_V\,\omega \ \text{ on } U, \qquad \beta = \rho_U\,\omega \ \text{ on } V. \tag{1.88}$$

(Here $\rho_V\omega$ extends by zero to all of U, because ρ_V vanishes near the edge of V.) Then $s(\alpha,\beta) = \rho_U\omega + \rho_V\omega = \omega$.

Key idea: Mayer–Vietoris

A short exact sequence of complexes gives a long exact sequence of cohomology:

$$\begin{aligned} 0 \to H^0(X) &\xrightarrow{r^*} H^0(U)\oplus H^0(V) \xrightarrow{s^*} H^0(U\cap V) \\ &\xrightarrow{\mathrm{d}^*} H^1(X) \xrightarrow{r^*} H^1(U)\oplus H^1(V) \xrightarrow{s^*} H^1(U\cap V) \\ &\xrightarrow{\mathrm{d}^*} H^2(X) \xrightarrow{r^*} H^2(U)\oplus H^2(V) \xrightarrow{s^*} \cdots \end{aligned} \tag{1.89}$$

The **connecting map** d^* sends a closed p-form ω on $U\cap V$ to the closed $(p+1)$-form on X

$$\mathrm{d}^*\omega = \mathrm{d}\rho_U \wedge \omega, \tag{1.90}$$

which is supported inside $U\cap V$ and extended by zero.

Let us check that (1.90) is the right recipe. Lift ω to (α,β) by (1.88) and differentiate:

$$\begin{aligned} \mathrm{d}\beta &\overset{(1)}{=} \mathrm{d}\rho_U\wedge\omega + \rho_U\,\mathrm{d}\omega = \mathrm{d}\rho_U\wedge\omega \quad \text{on } V, \\ \mathrm{d}\alpha &\overset{(2)}{=} -\mathrm{d}\rho_V\wedge\omega \overset{(3)}{=} \mathrm{d}\rho_U\wedge\omega \quad \text{on } U. \end{aligned}$$

(1) Leibniz rule, and $\mathrm{d}\omega = 0$.
(2) The same, for $\alpha = -\rho_V\omega$.
(3) $\rho_U + \rho_V = 1$, so $\mathrm{d}\rho_V = -\mathrm{d}\rho_U$.

The two pieces agree on $U\cap V$, so they glue to one global closed form. It is not in general exact on X, although it is exact on U (as $\mathrm{d}\alpha$) and on V (as $\mathrm{d}\beta$). This is exactly the monopole situation: a field strength that has a potential on each patch but not globally. Exactness of the long sequence is then a (tedious but routine) exercise in chasing definitions.

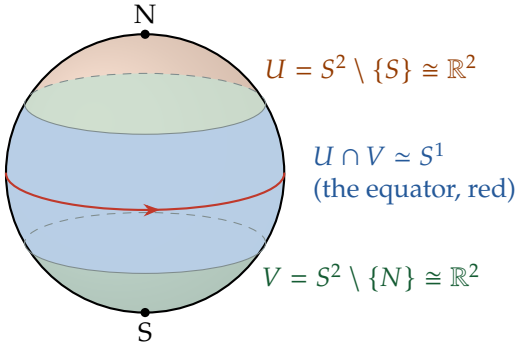


Figure 1.44: The Mayer–Vietoris cover of the sphere. Each cap is contractible, so all the topology of S^2 lives in how the caps overlap, along a band that retracts onto the equator. The class $[\mathrm{d}\phi/2\pi]$ on the band is sent by the connecting map to the generator of $H^2_{\mathrm{dR}}(S^2)$. In physics the same cover carries the Wu–Yang monopole: A_N on U, A_S on V, and a gauge transformation $e^{in\phi}$ on the band.

Worked example 1.79: Mayer–Vietoris computes the cohomology of S^2

Take U and V as in Figure 1.44. Both are contractible, so $H^0=\mathbb{R}$ and $H^p=0$ for $p\geq 1$ on each (Poincaré lemma). The overlap retracts onto the equator, so $H^0(U\cap V)=H^1(U\cap V)=\mathbb{R}$ (Example 1.66). The sequence (1.89) becomes

$$0\to H^0(S^2)\xrightarrow{r^*}\mathbb{R}\oplus\mathbb{R}\xrightarrow{s^*}\mathbb{R}\xrightarrow{\mathrm{d}^*}H^1(S^2)\to 0\oplus 0\to\mathbb{R}\xrightarrow{\mathrm{d}^*}H^2(S^2)\to 0\oplus 0.$$

Read it from the left.

$$\begin{aligned} H^0(S^2) &\overset{(1)}{=} \mathbb{R}, \\ \operatorname{Im} s^* &\overset{(2)}{=} \{b-a : (a,b) \in \mathbb{R}^2\} = \mathbb{R}, \\ H^1(S^2) &\overset{(3)}{=} 0, \\ H^2(S^2) &\overset{(4)}{=} \mathrm{d}^*\big(H^1(U \cap V)\big) \cong \mathbb{R}. \end{aligned}$$

(1) S^2 is connected, so its closed 0-forms are the constants.
(2) On H^0, s^* takes the constant a on U and b on V to the constant $b-a$ on the overlap. Every constant arises.
(3) Exactness at $H^0(U \cap V)$: since s^* is onto, $\operatorname{Ker} \mathrm{d}^* = \mathbb{R}$ and $\mathrm{d}^* = 0$. Then exactness at $H^1(S^2)$: the next map lands in 0, so its kernel is all of $H^1(S^2)$, which must equal $\operatorname{Im} \mathrm{d}^* = 0$.
(4) Exactness at $H^1(U \cap V)$ and at $H^2(S^2)$: the map into $H^1(U \cap V)$ comes from 0, so d^* is injective, and the map out of $H^2(S^2)$ goes to 0, so d^* is surjective.

The explicit generator. Let $\rho_U = \rho(\theta)$ depend only on the polar angle, with $\rho = 1$ for $\theta < \pi/3$ and $\rho = 0$ for $\theta > 2\pi/3$. Using the orientation $\mathrm{d}\theta \wedge \mathrm{d}\phi$ (the one in which $\sin\theta\, \mathrm{d}\theta \wedge \mathrm{d}\phi$ is positive),

$$\begin{aligned} \int_{S^2} \mathrm{d}^*\Big(\frac{\mathrm{d}\phi}{2\pi}\Big) &\overset{(1)}{=} \int_{S^2} \rho'(\theta)\, \mathrm{d}\theta \wedge \frac{\mathrm{d}\phi}{2\pi} \\ &\overset{(2)}{=} \int_0^{\pi} \rho'(\theta)\, \mathrm{d}\theta \\ &\overset{(3)}{=} \rho(\pi) - \rho(0) = -1. \end{aligned}$$

(1) (1.90) with $\mathrm{d}\rho_U = \rho'(\theta)\, \mathrm{d}\theta$.
(2) The ϕ integral gives 2π.
(3) Fundamental theorem of calculus.

So $-\mathrm{d}\rho_U \wedge \mathrm{d}\phi/2\pi$ is a closed 2-form, a "band of flux" concentrated between latitudes, with total flux 1. Every closed 2-form on S^2 is a multiple of it plus an exact form. In particular, the monopole's $\frac{1}{4\pi} \sin\theta\, \mathrm{d}\theta \wedge \mathrm{d}\phi$ and this band are cohomologous: they have the same total flux.

Worked example 1.80: All spheres at once: Nash's (1.46)

Cover S^n by $U = S^n \setminus \{S\}$ and $V = S^n \setminus \{N\}$, both $\cong \mathbb{R}^n$, with $U \cap V \simeq S^{n-1}$. For $p \geq 1$ the groups $H^p(U) \oplus H^p(V)$ vanish, so the segment

$$0 = H^{p-1}(U) \oplus H^{p-1}(V) \to H^{p-1}(U \cap V) \xrightarrow{\mathrm{d}^*} H^p(S^n) \to H^p(U) \oplus H^p(V) = 0 \qquad (p \geq 2)$$

shows $H^p(S^n) \cong H^{p-1}(S^{n-1})$ for $p \geq 2$. The low degrees are handled as in Example 1.79: $H^0(S^n) = \mathbb{R}$, and $H^1(S^n) = 0$ for $n \geq 2$. Iterating the shift,

$$\begin{aligned} H^n(S^n) &\overset{(1)}{=} H^{n-1}(S^{n-1}) \overset{(1)}{=} \cdots \overset{(1)}{=} H^1(S^1) \overset{(2)}{=} \mathbb{R}, \\ H^p(S^n) &\overset{(3)}{=} H^{p-1}(S^{n-1}) \overset{(3)}{=} \cdots \overset{(3)}{=} H^1(S^{n-p+1}) \overset{(4)}{=} 0 \qquad (2 \leq p < n), \end{aligned}$$

(1) The shift isomorphism, applied $n-1$ times; each step lowers a degree that is still ≥ 2.

(2) Example 1.66.
(3) The shift applied $p-1$ times. Degree and dimension drop together, so we end in degree 1 on a sphere of dimension $n-p+1 \geq 2$.
(4) The low-degree result: $H^1(S^m) = 0$ for $m \geq 2$.

So $H^m(S^n) = \mathbb{R}$ for $m = 0, n$ and 0 otherwise. The same argument with integer coefficients (Mayer–Vietoris holds for singular cohomology too) gives Nash's

$$H^m(S^n;\mathbb{Z}) \cong H_m(S^n;\mathbb{Z}) = \begin{cases} \mathbb{Z} & m = 0 \text{ or } m = n, \\ 0 & \text{otherwise.} \end{cases} \tag{1.91}$$

The generator of $H^n(S^n)$ is "a bump of unit total integral", for example the normalised volume form. Compare this three-line answer with the homotopy groups of spheres, which fill tables and are still not all known.

Try it: The torus by Mayer–Vietoris

Cut T^2 into two cylinders U and V (each $\simeq S^1$) whose intersection is two disjoint cylinders ($\simeq S^1 \sqcup S^1$). Show that the sequence forces $b_0 = 1$, $b_1 = 2$, $b_2 = 1$. *Answer sketch:* H^0: $\mathbb{R}^2 \to \mathbb{R}^2$, $(a,b) \mapsto (b-a, b-a)$ has rank 1, so the cokernel $\mathbb{R}$ injects into $H^1(T^2)$. H^1: $\mathbb{R} \oplus \mathbb{R} \to \mathbb{R} \oplus \mathbb{R}$ sends $(a,b) \mapsto (b-a, b-a)$, kernel $\mathbb{R}$ and cokernel $\mathbb{R}$. So $b_1 = 1+1 = 2$ and $b_2 = \dim \operatorname{coker} = 1$.

1.4.10 Four facts Nash lists

Nash closes §4 by recalling "a few well known and useful facts". Each deserves its own explanation.

Finiteness: cohomology stops at the dimension

Key idea: Finiteness

For a manifold X of dimension n, $H^p(X) = 0$ for $p > n$ (and trivially for $p < 0$). If X is compact, each $H^p(X;\mathbb{Z})$ is finitely generated, so each $H^p_{\mathrm{dR}}(X)$ is finite-dimensional.

The first statement is obvious for forms: a p-form with $p > n$ is a totally antisymmetric tensor with more indices than there are directions, so it is zero. For triangulations: an n-manifold needs no simplices of dimension above n. The second statement follows from the existence of a finite triangulation for a compact smooth manifold: then all the chain groups are $\mathbb{Z}^{N_n}$ with finite N_n. Contrast homotopy: $\pi_3(S^2) = \mathbb{Z} \neq 0$, even though $3 > \dim S^2$. Homotopy groups do not stop at the dimension.

The cohomology ring and the cup product

Cohomology classes can be multiplied. For forms the product is the wedge product. If $\mathrm{d}\omega = 0$ and $\mathrm{d}\nu = 0$, then

$$\mathrm{d}(\omega \wedge \nu) \overset{(1)}{=} \mathrm{d}\omega \wedge \nu + (-1)^p\, \omega \wedge \mathrm{d}\nu \overset{(2)}{=} 0,$$
$$(\omega + \mathrm{d}\lambda) \wedge \nu \overset{(3)}{=} \omega \wedge \nu + \mathrm{d}(\lambda \wedge \nu). \tag{1.92}$$

(1) Graded Leibniz rule for a p-form ω.

(2) Both forms are closed.
(3) $\mathrm{d}(\lambda \wedge \nu) = \mathrm{d}\lambda \wedge \nu \pm \lambda \wedge \mathrm{d}\nu = \mathrm{d}\lambda \wedge \nu$, because ν is closed.

So closed $\wedge$ closed is closed, and changing ω by an exact form changes $\omega \wedge \nu$ by an exact form. The product of classes, Nash's (1.45),

$$\cup : H^m(X;\mathbb{R}) \times H^n(X;\mathbb{R}) \to H^{m+n}(X;\mathbb{R}), \qquad [\omega] \cup [\nu] = [\omega \wedge \nu], \tag{1.93}$$

is therefore well defined. It makes $H^*(X;\mathbb{R}) = \bigoplus_{n\geq 0} H^n(X;\mathbb{R})$ into a **graded-commutative ring**: $[\omega] \cup [\nu] = (-1)^{mn}[\nu] \cup [\omega]$. The **cup product** exists for any coefficient ring (for example $\mathbb{Z}$ or $\mathbb{Z}_2$), where forms are not available. On simplicial cochains it is the "front face times back face" formula

$$(\varphi \cup \psi)[v_0, \dots, v_{m+n}] = \varphi[v_0, \dots, v_m]\, \psi[v_m, \dots, v_{m+n}]. \tag{1.94}$$

Worked example 1.81: The cohomology ring of the torus, two ways

With forms. $x = [\mathrm{d}\theta_1/2\pi]$ and $y = [\mathrm{d}\theta_2/2\pi]$ span $H^1(T^2)$. Then $x \cup x = 0$ (since $\mathrm{d}\theta_1 \wedge \mathrm{d}\theta_1 = 0$), $y \cup y = 0$, $y \cup x = -x \cup y$, and

$$\int_{T^2} x \cup y = \frac{1}{4\pi^2} \int_0^{2\pi} \int_0^{2\pi} \mathrm{d}\theta_1\, \mathrm{d}\theta_2 = 1.$$

So $x \cup y$ generates $H^2(T^2;\mathbb{Z})$, and $H^*(T^2) = \Lambda[x, y]$ is the exterior algebra on two generators.
With cochains. On the two-triangle torus, take the cocycles of Example 1.75

$$\alpha : (a,b,c) \mapsto (1,0,1), \qquad \beta : (a,b,c) \mapsto (0,1,1)$$

(both satisfy $x + y - z = 0$). The triangle $L = [(0,0),(1,0),(1,1)]$ has front edge a and back edge b. The triangle $U = [(0,0),(0,1),(1,1)]$ has front edge b and back edge a. With the standard (counterclockwise) orientation the fundamental class is $[T^2] = L - U$. By (1.94),

$$(\alpha \cup \beta)[T^2] \overset{(1)}{=} \alpha(a)\beta(b) - \alpha(b)\beta(a) = 1 \cdot 1 - 0 \cdot 0 = 1,$$
$$(\beta \cup \alpha)[T^2] \overset{(1)}{=} \beta(a)\alpha(b) - \beta(b)\alpha(a) = 0 \cdot 0 - 1 \cdot 1 = -1.$$

(1) Front edge times back edge on L, minus the same on U.

The combinatorial formula reproduces $\int \omega_1 \wedge \omega_2 = 1$ and the antisymmetry $y \cup x = -x \cup y$. (At the cochain level $\cup$ is not graded-commutative; only on cohomology is it.)

Worked example 1.82: Same numbers, different rings

T^2 and the wedge $S^1 \vee S^1 \vee S^2$ (two circles and a sphere glued at a point) both have Betti numbers $1, 2, 1$. In T^2 the product of the two degree-one generators is the degree-two generator. In the wedge it is zero: the circles and the sphere touch at a single point, and a 1-form supported on one circle wedged with one on the other circle vanishes. So the cohomology *ring* tells them apart when the groups cannot. The ring structure is also the origin of the intersection form of 4-manifolds, (1.162) below.

Poincaré duality

On a closed oriented n-manifold M we can pair a p-form with an $(n-p)$-form by wedging and integrating.

Key idea: Poincaré duality

If M is closed (compact, no boundary) and oriented, the pairing

$$H^p_{\mathrm{dR}}(M) \times H^{n-p}_{\mathrm{dR}}(M) \to \mathbb{R}, \qquad ([\omega],[\nu]) \mapsto \int_M \omega \wedge \nu \tag{1.95}$$

is well defined and non-degenerate. Hence $b_p = b_{n-p}$. With integer coefficients, $H^p(M;\mathbb{Z}) \cong H_{n-p}(M;\mathbb{Z})$.

Well defined: shifting ω by $\mathrm{d}\lambda$ changes the integral by $\int_M \mathrm{d}(\lambda \wedge \nu) = \int_{\partial M} \lambda \wedge \nu = 0$, by (1.92) and Stokes. Non-degenerate is the real content. It is most transparent with Hodge theory (Chapter 2): each class has a harmonic representative ω, and $\star\omega$ is again harmonic, so

$$\int_M \omega \wedge \star\omega = \|\omega\|^2_{L^2} > 0 \quad \text{for } \omega \neq 0.$$

The class $[\star\omega] \in H^{n-p}_{\mathrm{dR}}$ therefore pairs non-trivially with $[\omega]$.

The geometric meaning is *intersection*. A closed oriented $(n-p)$-dimensional submanifold $S \subset M$ has a **Poincaré dual** form η_S, a closed p-form concentrated near S, with $\int_S \nu = \int_M \nu \wedge \eta_S$ for every closed $(n-p)$-form ν. It is a smeared delta function transverse to S. Two such submanifolds of complementary dimension then have

$$\int_M \eta_S \wedge \eta_{S'} = \#(S \cap S') \quad \text{(counted with signs)}.$$

Worked example 1.83: Duality on the torus

On T^2 the loops a ($\theta_2 = 0$) and b ($\theta_1 = 0$) cross once. Their Poincaré duals are bump forms across them: $\eta_a = \delta_\epsilon(\theta_2)\,\mathrm{d}\theta_2$ and $\eta_b = -\delta_\epsilon(\theta_1)\,\mathrm{d}\theta_1$, where δ_ϵ is a narrow bump of unit area. (Check: $\int_a \nu = \int_{T^2} \nu \wedge \eta_a$ for $\nu = \mathrm{d}\theta_1/2\pi$ gives $1 = 1$. The sign in η_b is fixed the same way, since $\mathrm{d}\theta_2 \wedge \mathrm{d}\theta_1 = -\mathrm{d}\theta_1 \wedge \mathrm{d}\theta_2$.) In cohomology $\eta_a \sim \mathrm{d}\theta_2/2\pi = y$ and $\eta_b \sim -\mathrm{d}\theta_1/2\pi = -x$. The pairing matrix on H^1 in the basis (x, y) is

$$Q = \begin{pmatrix} \int x\wedge x & \int x\wedge y \\ \int y\wedge x & \int y\wedge y \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \qquad \det Q = 1 \neq 0.$$

It is antisymmetric because $p = n - p = 1$ is odd, and unimodular (determinant ± 1) because Poincaré duality holds over $\mathbb{Z}$. In dimension 4 with $p = 2$ the same construction gives a *symmetric* unimodular matrix: the intersection form of Section 1.6.8.

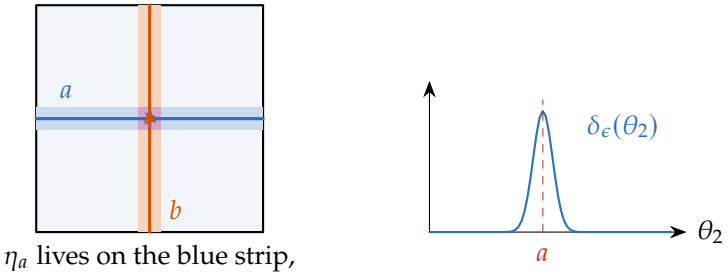
Künneth: cohomology of products

Key idea: Künneth formula

With field coefficients (for example $\mathbb{R}$),

$$H^k(X \times Y) \cong \bigoplus_{p+q=k} H^p(X) \otimes H^q(Y), \tag{1.96}$$

where $[\omega] \otimes [\nu] \mapsto [\pi_X^*\omega \wedge \pi_Y^*\nu]$. In terms of the **Poincaré polynomial** $P_X(t) = \sum_p b_p(X)\, t^p$, this says $P_{X\times Y} = P_X\, P_Y$.



η_a lives on the blue strip, η_b on the orange strip

profile of η_a across a: unit area

Figure 1.45: Poincaré duality as smeared delta functions. The dual of the loop a is a 1-form supported in a thin strip around a, with unit integral across it (right). The wedge of the two duals is supported only where the strips cross (purple square), so $\int \eta_a \wedge \eta_b = \pm 1$ counts the single crossing point of a and b.

The mechanism is separation of variables. A closed form on $X \times Y$ is cohomologous to a sum of products of closed forms on the factors, just as a function on a product is expanded in products of eigenfunctions. (With Hodge theory, harmonic forms on $X \times Y$ are exactly the sums of products of harmonic forms on X and Y.)

Worked example 1.84: Tori, and products of spheres

$$P_{T^n}(t) \overset{(1)}{=} (1+t)^n = \sum_{p=0}^{n} \binom{n}{p} t^p,$$

$$P_{S^2\times S^2}(t) \overset{(2)}{=} (1+t^2)^2 = 1 + 2t^2 + t^4,$$

$$P_{S^1\times S^3}(t) \overset{(3)}{=} (1+t)(1+t^3) = 1 + t + t^3 + t^4.$$

(1) $P_{S^1} = 1 + t$. A basis of $H^p(T^n)$ is $\mathrm{d}\theta_{i_1} \wedge \cdots \wedge \mathrm{d}\theta_{i_p}$ with $i_1 < \cdots < i_p$, and there are $\binom{n}{p}$ of these: the exterior algebra on n generators.

(2) $P_{S^2} = 1 + t^2$. $H^2(S^2 \times S^2)$ is spanned by the area forms ω_1, ω_2 of the two factors.

(3) $S^1 \times S^3$ has $b_1 = 1$ but $b_2 = 0$. It is a compact complex manifold (a Hopf surface) that cannot be Kähler, because Kähler manifolds have even b_1.

For $S^2 \times S^2$ the ring gives the intersection matrix on H^2 in the basis $(\omega_1/4\pi, \omega_2/4\pi)$: $\int (\omega_1/4\pi)^2 = 0$, $\int \omega_1 \wedge \omega_2/16\pi^2 = 1$, so $Q = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$. It is even, of signature zero. We will need it in Section 1.6.8.

The Euler characteristic

Working definition 1.34 (Euler characteristic)

$$\chi(X) = \sum_p (-1)^p\, b_p(X) = \sum_p (-1)^p\, N_p,$$

where b_p is the Betti number [$\dim H^p(X;\mathbb{R})$, the number of independent p-dimensional holes] and N_p is the number of p-cells [pieces that are p-dimensional balls, e.g. vertices, edges, faces] in any cell decomposition of X. *Recipe:* count vertices minus edges plus faces minus ..., or take $P_X(-1)$ [the Poincaré polynomial $\sum_p b_p t^p$ at $t = -1$].

The equality of the two sums is (1.76) applied to the cellular chain complex. So the alternating count of cells does not depend on how you cut the space up. That is the content of Euler's polyhedron formula.

Worked example 1.85: Euler characteristics you should know

$$\chi(\text{cube surface}) \overset{(1)}{=} 8 - 12 + 6 = 2 = \chi(S^2),$$
$$\chi(\Sigma_g) \overset{(2)}{=} 1 - 2g + 1 = 2 - 2g,$$
$$\chi(S^n) \overset{(3)}{=} 1 + (-1)^n,$$
$$\chi(X \times Y) \overset{(4)}{=} \chi(X)\,\chi(Y).$$

(1) Vertices, edges, faces of a cube; any convex polyhedron gives 2.
(2) The genus-g surface is a $4g$-gon with edges identified in pairs $(a_1 b_1 a_1^{-1} b_1^{-1} \cdots a_g b_g a_g^{-1} b_g^{-1})$: one vertex, $2g$ edges, one face.
(3) From (1.91): $b_0 = b_n = 1$.
(4) Künneth: $\chi = P(-1)$ and $P_{X\times Y} = P_X P_Y$.

So $\chi(T^2) = 0$ and $\chi(S^2 \times S^2) = 4$. Every closed odd-dimensional manifold has $\chi = 0$, because Poincaré duality pairs b_p with b_{n-p} and they enter with opposite signs.

Physical picture: Three faces of one number

The Euler characteristic of a closed surface appears in three guises that look unrelated.

- *Combinatorics:* $V - E + F$.
- *Curvature (Gauss–Bonnet, Chapter T2):* $\frac{1}{2\pi}\int_\Sigma K\,\mathrm{d}A = \chi(\Sigma)$.
- *Zeros of vector fields (Poincaré–Hopf):* the sum of indices of the zeros of any vector field is χ. This forces a zero on S^2: the hairy ball theorem of Section 1.5.8.

And in the analytic guise $\chi = \sum(-1)^p \dim\{\text{harmonic } p\text{-forms}\}$ it is the index of the operator $\mathrm{d} + \mathrm{d}^\dagger$ from even to odd forms. The Atiyah–Singer theorem (Chapter 4) is the grand generalisation of "combinatorics = curvature = zero modes".

Hurewicz: where homotopy and homology meet

Homotopy groups and homology groups are both built from maps of spheres or simplices into X. There is a natural map between them. A based map $f : S^n \to X$ pushes forward the fundamental class of the sphere:

$$h : \pi_n(X) \to H_n(X;\mathbb{Z}), \qquad [f] \mapsto f_*[S^n]. \tag{1.97}$$

This is the **Hurewicz map**. For $n = 1$ it is Example 1.74: a loop, viewed as a 1-cycle.

Key idea: The Hurewicz theorem

Let X be path connected.

1. If $\pi_k(X) = 0$ for all $1 \le k < n$, with $n \ge 2$, then $H_k(X;\mathbb{Z}) = 0$ for $1 \le k < n$ and $h : \pi_n(X) \to H_n(X;\mathbb{Z})$ is an isomorphism. The first non-vanishing homotopy and homology groups occur in the same dimension and agree.

2. For $n = 1$: h is onto, and its kernel is the commutator subgroup, so

$$H_1(X;\mathbb{Z}) \cong \pi_1(X)/[\pi_1(X), \pi_1(X)]. \tag{1.98}$$

Why is H_1 the *abelianised* π_1? Homology adds chains, and addition is commutative: $a+b = b+a$. So whatever π_1 records about the *order* of loops is lost. Concretely, the commutator loop $aba^{-1}b^{-1}$ is the boundary of a surface: it is the edge of the square that becomes a punctured torus when a and b are glued (Figure 1.46). As a 1-chain it is $a + b - a - b = 0$ anyway. The theorem says that this is *all* that is lost.

Worked example 1.86: Hurewicz at work

- *Spheres.* S^n is $(n-1)$-connected, so $\pi_n(S^n) \cong H_n(S^n;\mathbb{Z}) = \mathbb{Z}$. The degree of a map $S^n \to S^n$ can therefore be computed homologically: $f_*[S^n] = (\deg f)[S^n]$, or with forms, $\deg f = \int_{S^n} f^*\mathrm{vol}/\int_{S^n} \mathrm{vol}$.
- *Only the first group.* $\pi_3(S^2) = \mathbb{Z}$ but $H_3(S^2) = 0$. Beyond the first non-vanishing group, homotopy and homology part company. The Hopf map (Section 1.5.7) generates a homotopy class that homology cannot see at all.
- *The figure eight* $S^1 \vee S^1$: one vertex and two edges a, b with $\partial = 0$, so $H_1 = \mathbb{Z}^2$. But π_1 is the free group on a, b, in which $ab \neq ba$. Abelianising the free group on two letters gives $\mathbb{Z}^2$.
- *Genus two.* $\pi_1(\Sigma_2) = \langle a_1, b_1, a_2, b_2 \mid a_1 b_1 a_1^{-1} b_1^{-1} a_2 b_2 a_2^{-1} b_2^{-1} = 1\rangle$, a non-abelian group (Nash's example). Abelianising turns the relation into $a_1+b_1-a_1-b_1+a_2+b_2-a_2-b_2 = 0$, which is empty. So $H_1(\Sigma_2;\mathbb{Z}) = \mathbb{Z}^4$.
- $\mathbb{RP}^2$: $\pi_1 = \mathbb{Z}_2$ is already abelian, and $H_1 = \mathbb{Z}_2$, as computed cellularly.

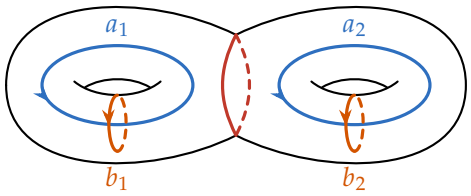


Figure 1.46: The genus-2 surface. The red loop cuts it into two punctured tori. As an element of π_1 it equals the commutator $a_1 b_1 a_1^{-1} b_1^{-1}$, which is not the identity: the loop cannot be shrunk. As a 1-cycle it is the boundary of the left half, so it is zero in H_1. This is the content of $H_1 = \pi_1/[\pi_1, \pi_1]$: homology forgets exactly the commutators.

Dictionary: cohomology in physics

mathematics	physics
closed but not exact 1-form	vector potential of a flux tube (Aharonov–Bohm)
period $\oint_\gamma A$	enclosed magnetic flux, holonomy phase
$H^1(X;\mathbb{Z}) = [X, U(1)]$	sectors of large $U(1)$ gauge transformations
$H^2(X;\mathbb{Z})$	magnetic charges; $U(1)$ bundles (Section 1.5.11)
Mayer–Vietoris connecting map	Wu–Yang patching of the monopole
Poincaré dual form η_S	delta-function source (current) on S
$b_p = \dim H^p_{\mathrm{dR}}$	number of harmonic p-form zero modes
$\chi = \sum(-1)^p b_p$	index of $\mathrm{d} + \mathrm{d}^\dagger$; Witten index of SUSY QM

Reading Nash (pp. 13–14, (1.44)–(1.46) and Hurewicz)

Nash dualises the complex (1.44) to get homology, $\partial_{n-1} \circ \partial_n = 0$ and $H^E_n = \operatorname{Ker} \partial_{n-1} / \operatorname{Im} \partial_n$. (His indexing names each map by its target; ours names it by its source, $\partial_n : C_n \to C_{n-1}$; the content is identical.) He takes the singular chain complex and says its homology "is isomorphic to the integral cohomology group $H_n(X;\mathbb{Z})$"; read "homology" (see the pitfall in the homology subsection). He then lists: finiteness ($0 \le n \le \dim X$); the ring $H^*_E(X)$ with cup product equal to the wedge product for de Rham (1.45); the sphere result (1.46), which we derived by Mayer–Vietoris in Example 1.80; the Hurewicz isomorphism; and $H_1 = \pi_1/[\pi_1, \pi_1]$ with the genus-2 example. Poincaré duality, Künneth and the Euler characteristic are not in Nash's §4. We added them because Nash uses them freely later: Poincaré duality in the definition of the intersection form (1.72), and the Euler characteristic throughout the index theorem chapters.

Recap: §1.4 Cohomology and homology

- $K(G, n)$ has one homotopy group, G in degree n. Cohomology is $H^n(X;G) = [X, K(G,n)]_0$, a group because $K(G,n) = \Omega K(G, n+1)$. Examples: $H^1(X;\mathbb{Z}) = [X, S^1]_0$ (winding numbers).
- A complex is a sequence with $d \circ d = 0$. Its cohomology $\operatorname{Ker} d / \operatorname{Im} d$ measures failure of exactness. The Euler characteristic of the complex survives passage to cohomology.
- de Rham: forms and d. Homology: chains and ∂, computed by integer boundary matrices (torus: $\partial_2 = (1, 1, -1)^T$ twice, $\partial_1 = 0$). Cochains: transpose. Torsion ($\mathbb{RP}^2$) sits in H_1 but in H^2.
- de Rham's theorem: $H^p_{\mathrm{dR}}(X) \cong H^p(X;\mathbb{R})$; a closed form is exact iff all periods vanish.
- Mayer–Vietoris glues: it gives $H^*(S^n)$ in three lines, and its connecting map is the monopole's patching.
- Cup product = wedge; Poincaré duality $b_p = b_{n-p}$ and intersection numbers; Künneth $P_{X\times Y} = P_X P_Y$; $\chi = \sum(-1)^p b_p = V - E + F$.
- Hurewicz: first non-zero π_n equals first non-zero H_n; $H_1 = \pi_1^{\mathrm{ab}}$.

1.5 Fibre bundles and fibrations

A field in physics is a rule that attaches a value to each point of space. For a scalar field the value is a number. For a vector field it is a vector in the tangent space at that point. For a charged field it is a complex number defined *only up to a phase convention*, and the convention may be changed from point to point by a gauge transformation. In every case there is a *space of possible values* sitting over each point, and a field picks one value from each.

When all these value spaces can be identified with one fixed space F in a single global way, the field is just a function $M \to F$. Often they cannot be. The tangent planes of a sphere cannot be identified with $\mathbb{R}^2$ smoothly everywhere at once. The phases of a charged field around a magnetic monopole cannot be fixed smoothly on the whole sphere surrounding it. A **fibre bundle** is the structure that describes value spaces that are *locally* a product but *globally* twisted. This section builds bundles concretely, from transition functions, then turns to the homotopy theory Nash needs: pullbacks, classifying spaces, the exact homotopy sequence and fibrations.

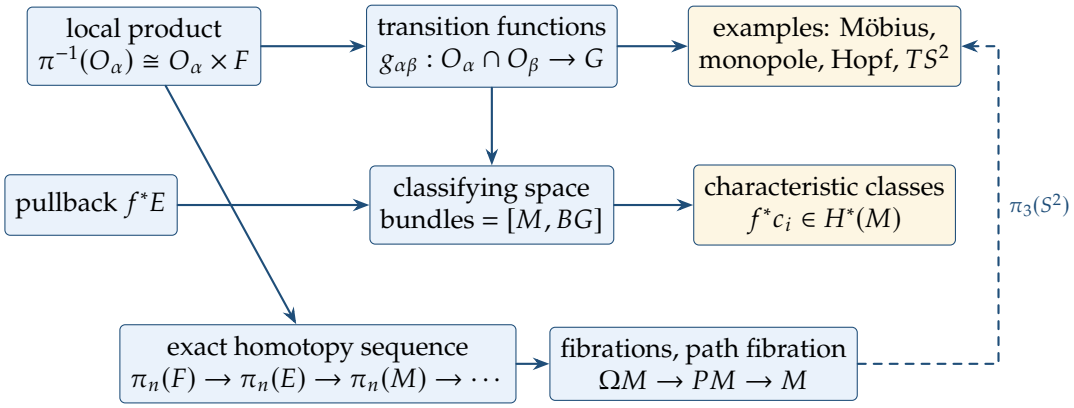


Figure 1.47: Roadmap of Section 1.5. The top row is geometry: bundles are glued from products by transition functions. The middle row is classification: every bundle is pulled back from a universal one, and its characteristic classes are pulled back cohomology. The bottom row is homotopy theory: the exact sequence of a bundle or fibration, which (among other things) computes $\pi_3(S^2) = \mathbb{Z}$ from the Hopf fibration.

1.5.1 The first example: the tangent bundle

Nash starts with the tangent bundle of a manifold M,

$$TM = \bigcup_{p \in M} T_pM, \tag{1.99}$$

the collection of all tangent spaces, one for each point. It comes with a **projection** $\pi : TM \to M$ that sends a tangent vector to the point where it is attached. The vocabulary for every bundle is already here.

- The **base space** is M.
- The **total space** is TM.
- The **fibre** over p is $\pi^{-1}(p) = T_pM$, a copy of $\mathbb{R}^n$, $n = \dim M$.

When the fibres are vector spaces, and the gluing respects the linear structure, we speak of a **vector bundle**. Its **rank** is the dimension of the fibre. TM is a vector bundle of rank n.

Concretely. On $S^2 \subset \mathbb{R}^3$ the tangent plane at p is $T_pS^2 = \{v \in \mathbb{R}^3 : v \cdot p = 0\}$. So

$$TS^2 = \{(p, v) \in \mathbb{R}^3 \times \mathbb{R}^3 : \ |p| = 1, \ v \cdot p = 0\}, \qquad \pi(p, v) = p.$$

It is a 4-dimensional manifold: 2 dimensions of base plus 2 of fibre.

In coordinates. On a chart with coordinates x^μ, a tangent vector is $v = v^\mu \partial_\mu$, and the chart identifies $\pi^{-1}(\text{chart})$ with $(\text{chart}) \times \mathbb{R}^n$ via (x^μ, v^μ). On an overlap with a second chart x'^μ, the chain rule gives

$$v'^\mu = \frac{\partial x'^\mu}{\partial x^\nu} v^\nu. \tag{1.100}$$

The *Jacobian matrix* $\partial x'/\partial x$, an element of $\mathrm{GL}(n, \mathbb{R})$ at each point of the overlap, tells us how to glue the local pieces $(\text{chart}) \times \mathbb{R}^n$ together. That is the entire content of a bundle, as we now explain.

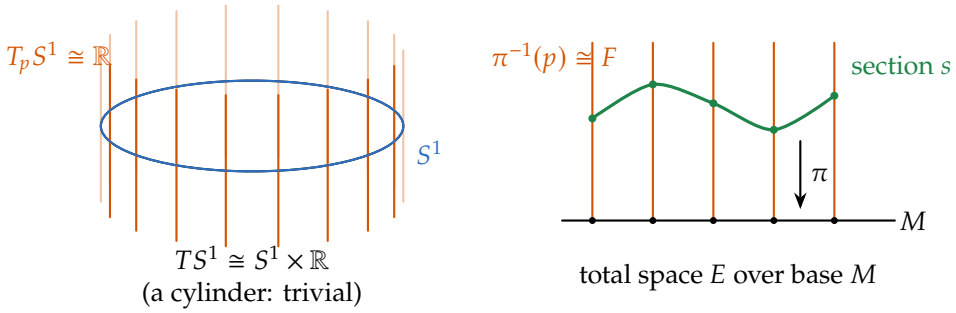


Figure 1.48: Left: the tangent bundle of the circle is a cylinder, one tangent line (orange) over each point of S^1. Right: the general picture of a bundle. Each point p of the base M carries a fibre $\pi^{-1}(p)$, a copy of F. The projection π collapses each fibre to its point. A section (green) picks one point in each fibre: it is a field.

1.5.2 Local triviality and transition functions

Working definition 1.35 (fibre bundle)

A **fibre bundle** with fibre F over M is a space E with a map $\pi : E \to M$ such that M has an open cover $\{O_\alpha\}$ [open sets whose union is all of M] with **local trivialisations**

$$\phi_\alpha : \ \pi^{-1}(O_\alpha) \xrightarrow{\cong} O_\alpha \times F, \qquad \phi_\alpha(e) = (\pi(e), \ f_\alpha(e)). \tag{1.101}$$

[$\pi^{-1}(O_\alpha)$ is the part of E lying over O_α, and $\cong$ means homeomorphic]. Over each patch the bundle is a product. On an overlap $O_\alpha \cap O_\beta$ the two identifications of the fibre differ by a map

$$f_\alpha(e) = g_{\alpha\beta}(\pi(e)) \cdot f_\beta(e), \qquad g_{\alpha\beta} : O_\alpha \cap O_\beta \to G, \tag{1.102}$$

where G is a group acting on F [each $g \in G$ moves the points of F, with $(gh)\cdot f = g\cdot(h\cdot f)$], the **structure group**, and the $g_{\alpha\beta}$ are the **transition functions**. *Operational form:* a bundle *is* a cover, a group G acting on F, and a set of G-valued functions on the overlaps obeying the cocycle conditions (1.104) [$g_{\alpha\beta}\, g_{\beta\gamma} = g_{\alpha\gamma}$ wherever three patches overlap].

The transition functions are not arbitrary. On any point of a triple overlap $O_\alpha \cap O_\beta \cap O_\gamma$ we

can change coordinates $\gamma \to \beta \to \alpha$ or go directly $\gamma \to \alpha$, and the answers must agree:

$$\begin{aligned} g_{\alpha\gamma}\, f_\gamma &\overset{(1)}{=} f_\alpha \\ &\overset{(2)}{=} g_{\alpha\beta}\, f_\beta \\ &\overset{(3)}{=} g_{\alpha\beta}\, g_{\beta\gamma}\, f_\gamma . \end{aligned} \tag{1.103}$$

(1) (1.102) for the pair (α, γ).
(2) (1.102) for the pair (α, β).
(3) (1.102) for the pair (β, γ).

Since this holds for every point of the fibre (and we take G to act faithfully), we get the **cocycle conditions**

$$g_{\alpha\alpha} = 1, \qquad g_{\beta\alpha} = g_{\alpha\beta}^{-1}, \qquad g_{\alpha\beta}\, g_{\beta\gamma}\, g_{\gamma\alpha} = 1 \ \text{ on } O_\alpha \cap O_\beta \cap O_\gamma . \tag{1.104}$$

Key idea: A bundle is its transition functions

Conversely, given a cover $\{O_\alpha\}$ and functions $g_{\alpha\beta}$ obeying (1.104), we can *build* the bundle:

$$E = \Big(\bigsqcup_\alpha O_\alpha \times F \Big) \Big/ \sim, \qquad (x, f)_\beta \sim (x,\ g_{\alpha\beta}(x) f)_\alpha \ \text{ for } x \in O_\alpha \cap O_\beta . \tag{1.105}$$

The cocycle conditions are exactly what makes $\sim$ an equivalence relation (reflexive, symmetric, transitive). Changing the trivialisations by G-valued functions h_α on each patch replaces

$$g_{\alpha\beta} \longmapsto h_\alpha\, g_{\alpha\beta}\, h_\beta^{-1}, \tag{1.106}$$

and gives an isomorphic bundle. A bundle is **trivial**, $E \cong M \times F$, exactly when its transition functions can be written as $g_{\alpha\beta} = h_\alpha h_\beta^{-1}$.

Notice the resemblance with cohomology. "Cocycle" modulo "coboundary" $h_\alpha h_\beta^{-1}$ is a cohomology in which the cochains are functions on overlaps: Čech cohomology, the subject of Chapter 3. For $G = U(1)$ this is literal. Write $g_{\alpha\beta} = e^{i\lambda_{\alpha\beta}}$. The cocycle condition says $\lambda_{\alpha\beta} + \lambda_{\beta\gamma} + \lambda_{\gamma\alpha} \in 2\pi\mathbb{Z}$, and the integers $n_{\alpha\beta\gamma} = (\lambda_{\alpha\beta} + \lambda_{\beta\gamma} + \lambda_{\gamma\alpha})/2\pi$ form a class in $H^2(M;\mathbb{Z})$: the first Chern class of the bundle. We will compute it for the monopole below.

Physical picture: Transition functions are gauge transformations on the overlaps

In a gauge theory with gauge group G, a choice of trivialisation over a patch is a choice of gauge. On the overlap of two patches two gauges are in use, and they are related by a gauge transformation: that is $g_{\alpha\beta}$. The replacement (1.106) is a change of gauge on each patch. A bundle is *non-trivial* when no choice of gauges makes all the patch-to-patch transformations disappear. The magnetic monopole is the prototype.

Worked example 1.87: Cylinder and Möbius band: the simplest twist

Take the base S^1, with angle ϕ, covered by two arcs $O_1 = \{-\frac{\pi}{2} - \epsilon < \phi < \frac{\pi}{2} + \epsilon\}$ and $O_2 = \{\frac{\pi}{2} - \epsilon < \phi < \frac{3\pi}{2} + \epsilon\}$. Their overlap has two pieces: A near $\phi = \pi/2$ and B near $\phi = -\pi/2$. Take fibre $F = \mathbb{R}$ and structure group $G = \mathbb{Z}_2 = \{\pm 1\} = \mathrm{O}(1)$.

- $g_{12} = +1$ on A and $+1$ on B: the cylinder $S^1 \times \mathbb{R}$.

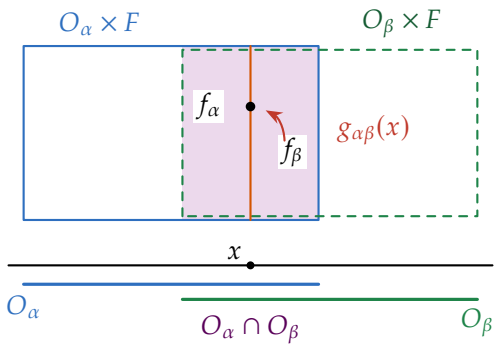


Figure 1.49: Gluing a bundle. Over O_α the bundle is the product $O_\alpha \times F$ (blue box), over O_β it is $O_\beta \times F$ (green). Over the overlap (purple) a single point of the fibre above x has two names, f_α and f_β, related by the transition function: $f_\alpha = g_{\alpha\beta}(x) f_\beta$. A twisted bundle arises when these relabellings cannot all be undone.

- $g_{12} = +1$ on A and -1 on B: the **Möbius band**.

There are no triple overlaps, so the cocycle conditions hold trivially. Can we untwist the Möbius band by (1.106)? The new transition function is $h_1 g_{12} h_2^{-1}$ with $h_i : O_i \to \{\pm 1\}$ continuous, hence constant on each (connected) arc. So

$$h_1 g_{12} h_2^{-1}\big|_A = h_1 h_2^{-1}, \qquad h_1 g_{12} h_2^{-1}\big|_B = -h_1 h_2^{-1}.$$

The product of the values on A and B is -1 for every choice. It is an invariant, so the band cannot be made trivial. Allowing $G = \mathrm{GL}(1,\mathbb{R}) = \mathbb{R}^\times$ does not help: h_i must be nowhere zero, hence of constant sign, and the same argument applies.

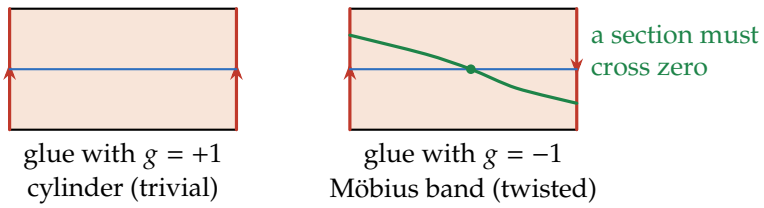


Figure 1.50: The two line bundles over the circle. Both are a strip $[0, 2\pi] \times \mathbb{R}$ with the ends glued. Gluing with $+1$ gives the cylinder. Gluing with -1 (arrows reversed) gives the Möbius band. The blue line is the zero section. A section of the Möbius band (green) must take value $s(2\pi) = -s(0)$, so by the intermediate value theorem it vanishes somewhere.

1.5.3 Sections: fields as bundle maps

Working definition 1.36 (section)

A **section** of $\pi : E \to M$ is a map $s : M \to E$ with $\pi \circ s = \mathrm{id}_M$: it picks one point in each fibre [the copy $\pi^{-1}(x)$ of F sitting over the point x]. *Operationally*, it is a collection of local functions $s_\alpha : O_\alpha \to F$ that obey

$$s_\alpha(x) = g_{\alpha\beta}(x)\, s_\beta(x) \qquad \text{on } O_\alpha \cap O_\beta. \tag{1.107}$$

[$g_{\alpha\beta}$ are the transition functions, converting the β description of a point of F into the α one]

A physical field of a given type is a section of the appropriate bundle. On a trivial bundle sections are just functions $M \to F$. On a twisted bundle the local functions must be patched with the transition functions, and this can force zeros.

Worked example 1.88: Every section of the Möbius band has a zero

Cut the circle at $\phi = 0$ and describe a section by a continuous function $s(\phi)$, $\phi \in [0, 2\pi]$. Going around once, the fibre coordinate comes back multiplied by the product of the transition functions encountered, which is -1. So $s(2\pi) = -s(0)$. If $s(0) > 0$ then $s(2\pi) < 0$, and by the intermediate value theorem s vanishes somewhere. The zero section is a section, but there is no nowhere-vanishing one. Consequently the Möbius band has no global frame: it is not trivial. (A nowhere-zero section s of a line bundle would give a trivialisation $\lambda s(x) \mapsto (x, \lambda)$.)

Key idea: Sections detect triviality

A vector bundle of rank k is trivial exactly when it has k sections that are linearly independent at every point (a global frame). A principal bundle (below) is trivial exactly when it has a single global section.

1.5.4 New vector bundles from old

Nash notes that we can build new bundles "by carrying out the standard operations of linear algebra on the fibres". In terms of transition functions this is mechanical. Let E and E' have transition functions $g_{\alpha\beta}$ and $g'_{\alpha\beta}$ (matrices, on a common cover).

bundle	fibre over x	transition functions
direct sum $E \oplus E'$	$E_x \oplus E'_x$	$\begin{pmatrix} g_{\alpha\beta} & 0 \\ 0 & g'_{\alpha\beta} \end{pmatrix}$
tensor product $E \otimes E'$	$E_x \otimes E'_x$	$g_{\alpha\beta} \otimes g'_{\alpha\beta}$ (Kronecker product)
dual E^*	$(E_x)^*$	$(g_{\alpha\beta}^T)^{-1}$
determinant $\Lambda^k E$	$\Lambda^k E_x$	$\det g_{\alpha\beta}$
exterior power $\Lambda^p E$	$\Lambda^p E_x$	$\Lambda^p g_{\alpha\beta}$ ($p \times p$ minors)

Each line is checked in the same way: apply the rule to the fibres, and verify that the new transition functions still satisfy the cocycle condition. For the dual, a covector ξ pairs with vectors to give numbers, and this pairing must be gauge invariant. If $v_\alpha = g_{\alpha\beta} v_\beta$, then

$$\xi_\alpha^T v_\alpha \overset{(1)}{=} \xi_\alpha^T g_{\alpha\beta} v_\beta$$

$$\overset{(2)}{=} \xi_\beta^T v_\beta \qquad \Longleftrightarrow \qquad \xi_\alpha = (g_{\alpha\beta}^T)^{-1}\xi_\beta.$$

(1) Transform the vector.
(2) Require invariance of the pairing for every v_β; this forces $g_{\alpha\beta}^T \xi_\alpha = \xi_\beta$.

Worked example 1.89: Forms are sections of bundles built from TM

The cotangent bundle T^*M is the dual of TM. Its transition functions are the inverse transpose of the Jacobian (1.100), which is the familiar covector rule $\xi'_\mu = \frac{\partial x^\nu}{\partial x'^\mu}\xi_\nu$. The p-forms of Chapter T1 are sections of $\Lambda^p T^*M$. The top power $\Lambda^n T^*M$ has transition functions $\det(\partial x/\partial x') = 1/\det(\partial x'/\partial x)$, which is the Jacobian factor in the change of variables of an integral. M is **orientable** exactly when we can choose charts with all Jacobian determinants positive. Then $\Lambda^n T^*M$ has transition functions in $\mathbb{R}_{>0}$, which can be gauged to 1, so it is trivial and has a nowhere-zero section: a volume form. For the Möbius band the Jacobians of the obvious charts have determinants of both signs, and no choice fixes it.

1.5.5 Principal bundles and associated bundles

Underneath every vector bundle Nash places a **principal bundle**, one whose fibre is the structure group itself.

Working definition 1.37 (principal bundle)

A **principal G-bundle** $P \to M$ is a bundle with fibre $F = G$, transition functions acting on the fibre by *left* multiplication, $f_\alpha = g_{\alpha\beta} f_\beta$. Because left and right multiplications commute [$(gf)h = g(fh)$ in G], G also acts on P from the *right*, $p \mapsto p \cdot h$, fibre by fibre, without reference to any trivialisation [a choice of patch identification $\pi^{-1}(O_\alpha) \cong O_\alpha \times G$]. *Operationally*: a principal bundle is just the set of transition functions $g_{\alpha\beta}$, regarded as data on their own, with no particular fibre in mind.

Given P and a representation $\rho : G \to \mathrm{GL}(V)$, the **associated vector bundle** $E = P \times_\rho V$ is the bundle with fibre V and transition functions $\rho(g_{\alpha\beta})$. The same principal bundle gives many vector bundles, one for each representation.

Worked example 1.90: The frame bundle of a manifold

A **frame** at $p \in M$ is an ordered basis $(e_1, \dots, e_n)$ of T_pM. The set of all frames at p is a copy of $\mathrm{GL}(n,\mathbb{R})$: fix one frame, and every other is obtained from it by a unique invertible matrix. The **frame bundle** $F(M)$ is the principal $\mathrm{GL}(n,\mathbb{R})$ bundle of all frames. Its transition functions are the Jacobians (1.100). The tangent bundle is associated to it via the defining representation of $\mathrm{GL}(n,\mathbb{R})$ on $\mathbb{R}^n$, the cotangent bundle via the dual representation, and tensor bundles via tensor representations. This is Nash's "underlying principal bundle with fibre $G = \mathrm{GL}(n,\mathbb{R})$".
A metric lets us restrict to orthonormal frames. Then the transition functions can be chosen in $\mathrm{O}(n)$, and with an orientation in $\mathrm{SO}(n)$. This **reduction of the structure group** will be important in Section 1.6 and Chapter T4, where lifting from $\mathrm{SO}(n)$ to $\mathrm{Spin}(n)$ defines spinors.

Dictionary: bundles and gauge theory

bundle language	physics language
principal G-bundle P	gauge group G, with topological sector fixed
local trivialisation over O_α	choice of gauge on O_α
transition function $g_{\alpha\beta}$	gauge transformation relating two gauges
associated bundle $P \times_\rho V$	matter field in the representation ρ
section of $P \times_\rho V$	matter field configuration $\psi(x)$
connection on P	gauge potential A_μ
curvature	field strength $F_{\mu\nu}$
non-trivial bundle	topologically non-trivial sector (monopole, instanton)

Gauge potentials on a patched bundle. We use the convention stated in Chapter T3: A is a Lie-algebra-valued 1-form with anti-Hermitian values, the covariant derivative on sections is $D = \mathrm{d} + A$, and $F = \mathrm{d}A + A \wedge A$. (The physicists' Hermitian field is related by $A_{\text{math}} = -ig\, A^a_{\text{phys}} T^a$; for electromagnetism with charge e, $A_{\text{math}} = -ie\, A_{\text{phys}}$.) Each patch carries its own local potential A_α. For D to be a well-defined operation on sections (1.107), we need $D_\alpha s_\alpha = g_{\alpha\beta} D_\beta s_\beta$. Writing $g = g_{\alpha\beta}$,

$$\begin{aligned} D_\alpha s_\alpha &\overset{(1)}{=} \mathrm{d}(g\, s_\beta) + A_\alpha\, g\, s_\beta \\ &\overset{(2)}{=} g\, \mathrm{d}s_\beta + (\mathrm{d}g)\, s_\beta + A_\alpha\, g\, s_\beta \\ &\overset{(3)}{=} g\big(\mathrm{d}s_\beta + (g^{-1} A_\alpha g + g^{-1}\mathrm{d}g)\, s_\beta\big). \end{aligned} \tag{1.108}$$

(1) Definition of D_α and the patching rule $s_\alpha = g s_\beta$.
(2) Leibniz rule.
(3) Factor out g on the left.

This equals $g\, D_\beta s_\beta$ exactly when

$$A_\beta = g_{\alpha\beta}^{-1} A_\alpha\, g_{\alpha\beta} + g_{\alpha\beta}^{-1} \mathrm{d} g_{\alpha\beta} \qquad \text{on } O_\alpha \cap O_\beta, \tag{1.109}$$

which is the gauge transformation law. The potentials are not global objects; they are local representatives glued by (1.109).

1.5.6 Clutching: bundles over spheres

For a sphere S^n there is a very economical description. Cover S^n by two caps O_N and O_S, each contractible, overlapping in a band around the equator S^{n-1}. Over a contractible set every bundle is trivial (we will see why in Section 1.5.10), so a bundle over S^n is completely described by one transition function on the band, and up to homotopy that is a map from the equator,

$$g_{NS} : S^{n-1} \longrightarrow G. \tag{1.110}$$

Deforming g_{NS} continuously deforms the bundle, and homotopic transition functions give isomorphic bundles. Changing trivialisations on the caps multiplies g_{NS} by maps that extend over the caps, which are null-homotopic. The result is the **clutching construction**: for connected G,

$$\{\text{principal } G\text{-bundles over } S^n\}/\text{iso} \;\cong\; \pi_{n-1}(G). \tag{1.111}$$

base	group	classified by	physics
S^2	$U(1)$	$\pi_1(U(1)) = \mathbb{Z}$	Dirac monopole of charge n
S^2	SO(3)	$\pi_1(\mathrm{SO}(3)) = \mathbb{Z}_2$	't Hooft–Polyakov sectors mod 2
S^4	SU(2)	$\pi_3(\mathrm{SU}(2)) = \mathbb{Z}$	BPST instantons of charge k (Chapter 8)
S^3	any Lie G	$\pi_2(G) = 0$	every bundle over S^3 is trivial

The monopole bundle

Worked example 1.91: The Dirac monopole as a $U(1)$ bundle over S^2

Cover S^2 by $O_N = \{0 \le \theta < \frac{\pi}{2} + \epsilon\}$ and $O_S = \{\frac{\pi}{2} - \epsilon < \theta \le \pi\}$. Take $G = U(1)$ and the transition function on the band

$$g_{NS}(\phi) = e^{in\phi}. \tag{1.112}$$

For this to be a continuous function on the band, $e^{in\cdot 2\pi} = 1$: *n must be an integer*. On the two patches choose the potentials

$$A_N = -\frac{in}{2}(1 - \cos\theta)\,\mathrm{d}\phi, \qquad A_S = +\frac{in}{2}(1 + \cos\theta)\,\mathrm{d}\phi. \tag{1.113}$$

Each is smooth on its own patch. A_N has a coefficient that vanishes at $\theta = 0$, where $\mathrm{d}\phi$ is singular, like $(1 - \cos\theta) \sim \theta^2/2$; in Cartesian coordinates near the north pole $\theta^2\mathrm{d}\phi \approx x\,\mathrm{d}y - y\,\mathrm{d}x$ is smooth. Likewise A_S at $\theta = \pi$.
They are glued correctly. For abelian G, (1.109) reads $A_S = A_N + g^{-1}\mathrm{d}g$:

$$\begin{aligned} A_N + g_{NS}^{-1}\mathrm{d}g_{NS} &\overset{(1)}{=} -\frac{in}{2}(1 - \cos\theta)\,\mathrm{d}\phi + in\,\mathrm{d}\phi \\ &= \frac{in}{2}(1 + \cos\theta)\,\mathrm{d}\phi = A_S. \end{aligned}$$

(1) $g^{-1}\mathrm{d}g = e^{-in\phi}\,(in)\,e^{in\phi}\mathrm{d}\phi = in\,\mathrm{d}\phi$.

The curvature is global. On both patches

$$F = \mathrm{d}A_N \overset{(1)}{=} -\frac{in}{2}\sin\theta\,\mathrm{d}\theta \wedge \mathrm{d}\phi, \qquad \mathrm{d}A_S \overset{(2)}{=} -\frac{in}{2}\sin\theta\,\mathrm{d}\theta \wedge \mathrm{d}\phi,$$

(1) $\mathrm{d}\big((1 - \cos\theta)\mathrm{d}\phi\big) = \sin\theta\,\mathrm{d}\theta \wedge \mathrm{d}\phi$.
(2) $\mathrm{d}\big((1 + \cos\theta)\mathrm{d}\phi\big) = -\sin\theta\,\mathrm{d}\theta \wedge \mathrm{d}\phi$.

so F is a single, smooth, closed 2-form on S^2: a radial magnetic field of constant strength.
The first Chern number. With $c_1 = \frac{i}{2\pi}F$ (orientation $\mathrm{d}\theta \wedge \mathrm{d}\phi$),

$$\begin{aligned} \int_{S^2} c_1 &\overset{(1)}{=} \frac{i}{2\pi}\Big(-\frac{in}{2}\Big)\int_0^{2\pi}\!\!\int_0^{\pi} \sin\theta\,\mathrm{d}\theta\,\mathrm{d}\phi \\ &\overset{(2)}{=} \frac{n}{4\pi}\cdot 4\pi = n. \end{aligned} \tag{1.114}$$

(1) Insert F.
(2) $i\cdot(-i) = 1$, and the area of the unit sphere is 4π.

We can also get this without any curvature, by Stokes on the two hemispheres H_N, H_S with common boundary the equator γ (oriented as the boundary of H_N):

$$\begin{aligned}\int_{S^2} F &\overset{(1)}{=} \int_{H_N} \mathrm{d}A_N + \int_{H_S} \mathrm{d}A_S \\ &\overset{(2)}{=} \oint_\gamma A_N - \oint_\gamma A_S \\ &\overset{(3)}{=} -\oint_\gamma g_{NS}^{-1}\mathrm{d}g_{NS} = -2\pi i\, n.\end{aligned}$$

(1) Split the sphere; on each hemisphere use the potential that is smooth there.
(2) Stokes. The equator is $+\gamma$ as the boundary of H_N and $-\gamma$ as the boundary of H_S.
(3) $A_S - A_N = g^{-1}\mathrm{d}g$, and $\oint in\,\mathrm{d}\phi = 2\pi i n$.

So $\frac{i}{2\pi}\int F = n$ again: *the total flux is the winding number of the transition function*.

Physical picture: Dirac quantisation is topology

In physical units, $A_{\text{math}} = -ie\,A_{\text{phys}}$ (with $\hbar = c = 1$), so the magnetic flux is $\Phi = \int B = \frac{i}{e}\int F = \frac{2\pi n}{e}$. That is Dirac's condition $e\Phi \in 2\pi\mathbb{Z}$, or $eg = n/2$ for a monopole of flux $4\pi g$. In Dirac's original picture, the "string" is where A_N would be singular if we insisted on using it on the whole sphere ($\theta = \pi$). The string is invisible exactly when $e^{ie\Phi} = 1$. Wu and Yang's patching removes the string altogether. The integer n is the clutching class in $\pi_1(U(1)) = \mathbb{Z}$, it is the first Chern number, and it is the class $[F/2\pi]$ in $H^2(S^2;\mathbb{Z}) = \mathbb{Z}$.

1.5.7 The Hopf fibration

The most famous non-trivial bundle is a principal $U(1)$ bundle over S^2 whose total space is the 3-sphere. Nash uses it (via the exact sequence) to compute $\pi_3(S^2)$. Let us build it by hand.

The map

Write $S^3 = \{(z_1, z_2) \in \mathbb{C}^2 : |z_1|^2 + |z_2|^2 = 1\}$ and think of $z = (z_1, z_2)^T$ as a normalised two-component spinor. Define

$$\pi(z_1, z_2) = z^\dagger \vec{\sigma} z = \big(2\,\mathrm{Re}(\bar z_1 z_2),\ 2\,\mathrm{Im}(\bar z_1 z_2),\ |z_1|^2 - |z_2|^2\big), \tag{1.115}$$

with the Pauli matrices $\sigma_1 = \left(\begin{smallmatrix}0&1\\1&0\end{smallmatrix}\right)$, $\sigma_2 = \left(\begin{smallmatrix}0&-i\\i&0\end{smallmatrix}\right)$, $\sigma_3 = \left(\begin{smallmatrix}1&0\\0&-1\end{smallmatrix}\right)$. The image lies on the unit sphere:

$$\begin{aligned}|\pi(z)|^2 &\overset{(1)}{=} 4|\bar z_1 z_2|^2 + \big(|z_1|^2 - |z_2|^2\big)^2 \\ &\overset{(2)}{=} \big(|z_1|^2 + |z_2|^2\big)^2 = 1.\end{aligned}$$

(1) $(2\,\mathrm{Re}\,w)^2 + (2\,\mathrm{Im}\,w)^2 = 4|w|^2$ with $w = \bar z_1 z_2$.
(2) $4ab + (a-b)^2 = (a+b)^2$ with $a = |z_1|^2$, $b = |z_2|^2$.

Physically, $\pi(z)$ is the *Bloch vector* $\langle\vec\sigma\rangle$ of the spin-$\frac12$ state z.

The fibres are circles

Multiplying the spinor by a phase does not change the Bloch vector: $\pi(e^{i\alpha}z) = (e^{i\alpha}z)^\dagger\vec\sigma(e^{i\alpha}z) = z^\dagger\vec\sigma z$. Conversely, if $\pi(z) = \pi(z')$ then $z' = e^{i\alpha}z$ (a spin state is fixed by its Bloch vector up to

phase). So

$$\pi^{-1}(\text{point}) = \{e^{i\alpha}z : \alpha \in [0, 2\pi)\} \cong U(1),$$

and $U(1)$ acts on S^3 from the right by $z \mapsto ze^{i\alpha}$, freely and preserving fibres. This is a principal $U(1)$ bundle

$$U(1) = S^1 \longrightarrow S^3 \xrightarrow{\pi} S^2, \tag{1.116}$$

the **Hopf fibration**. Equivalently $S^2 = \mathbb{CP}^1$, the space of complex lines in $\mathbb{C}^2$, and $\pi(z_1, z_2) = [z_1 : z_2]$.

Coordinates, trivialisations and the transition function

Parametrise S^3 by **Hopf coordinates**

$$z_1 = \cos\tfrac{\theta}{2}\, e^{i\xi_1}, \qquad z_2 = \sin\tfrac{\theta}{2}\, e^{i\xi_2}, \qquad 0 \le \theta \le \pi. \tag{1.117}$$

Then

$$\begin{aligned}\pi(z) &\overset{(1)}{=} \Big(2\cos\tfrac{\theta}{2}\sin\tfrac{\theta}{2}\cos(\xi_2-\xi_1),\ 2\cos\tfrac{\theta}{2}\sin\tfrac{\theta}{2}\sin(\xi_2-\xi_1),\ \cos^2\tfrac{\theta}{2}-\sin^2\tfrac{\theta}{2}\Big)\\ &\overset{(2)}{=} \big(\sin\theta\cos\phi,\ \sin\theta\sin\phi,\ \cos\theta\big), \qquad \phi = \xi_2 - \xi_1.\end{aligned}$$

(1) $\bar z_1 z_2 = \cos\frac{\theta}{2}\sin\frac{\theta}{2}\, e^{i(\xi_2-\xi_1)}$.
(2) Double-angle formulas.

So θ is the polar angle on the base, the azimuth is $\phi = \xi_2 - \xi_1$, and the overall phase $(\xi_1 + \xi_2)/2$ moves along the fibre.

Cover S^2 by $O_N = \{\theta < \pi\}$ (where $z_1 \neq 0$) and $O_S = \{\theta > 0\}$ (where $z_2 \neq 0$), and trivialise

$$\phi_N(z) = \Big(\pi(z),\ \frac{z_1}{|z_1|}\Big), \qquad \phi_S(z) = \Big(\pi(z),\ \frac{z_2}{|z_2|}\Big) \in S^2 \times U(1). \tag{1.118}$$

Each is a bijection from $\pi^{-1}(O_{N,S})$ to $O_{N,S} \times U(1)$, compatible with the $U(1)$ action. On the overlap the fibre coordinates are related by

$$\frac{z_2/|z_2|}{z_1/|z_1|} = e^{i(\xi_2-\xi_1)} = e^{i\phi} \qquad \Longrightarrow \qquad g_{SN}(\phi) = e^{i\phi}. \tag{1.119}$$

The transition function winds once around $U(1)$ as we go once around the equator. Comparing with (1.112): *the Hopf bundle is the monopole bundle with* $|n| = 1$. (Whether one calls it $n = +1$ or $n = -1$ depends on orientation conventions; the magnitude is what matters here.)

The natural connection is the monopole potential

The spinor z carries a natural $U(1)$ connection, the *Berry connection* $\mathcal{A} = z^\dagger \mathrm{d}z$. It is anti-Hermitian (purely imaginary), since $0 = \mathrm{d}(z^\dagger z) = z^\dagger \mathrm{d}z + (z^\dagger \mathrm{d}z)^*$. Pull it back to the base with the local sections

$$s_N(\theta,\phi) = \big(\cos\tfrac{\theta}{2},\ \sin\tfrac{\theta}{2}\, e^{i\phi}\big), \qquad s_S(\theta,\phi) = \big(\cos\tfrac{\theta}{2}\, e^{-i\phi},\ \sin\tfrac{\theta}{2}\big), \tag{1.120}$$

($\xi_1 = 0$ and $\xi_2 = 0$ respectively). s_N is smooth at $\theta = 0$ because $\sin\frac{\theta}{2}$ kills the ill-defined phase there, and s_S is smooth at $\theta = \pi$. Then

$$\begin{aligned}s_N^*\mathcal{A} &\overset{(1)}{=} \cos\tfrac{\theta}{2}\,\mathrm{d}\cos\tfrac{\theta}{2} + \sin\tfrac{\theta}{2}\, e^{-i\phi}\,\mathrm{d}\big(\sin\tfrac{\theta}{2}\, e^{i\phi}\big)\\ &\overset{(2)}{=} \tfrac{1}{2}\mathrm{d}\big(\cos^2\tfrac{\theta}{2} + \sin^2\tfrac{\theta}{2}\big) + i\sin^2\tfrac{\theta}{2}\,\mathrm{d}\phi\\ &\overset{(3)}{=} \frac{i}{2}(1-\cos\theta)\,\mathrm{d}\phi,\end{aligned}$$

(1) $\mathcal{A} = \bar{z}_1 \mathrm{d}z_1 + \bar{z}_2 \mathrm{d}z_2$ evaluated on s_N.
(2) Leibniz rule on the second term; the real parts combine into a total derivative.
(3) $\cos^2 + \sin^2 = 1$ has zero derivative, and $\sin^2 \frac{\theta}{2} = \frac{1}{2}(1 - \cos\theta)$.

and similarly $s_S^* \mathcal{A} = -\frac{i}{2}(1 + \cos\theta)\,\mathrm{d}\phi$. These are the monopole potentials (1.113) with $n = -1$. The curvature $\mathrm{d}\mathcal{A} = \frac{i}{2}\sin\theta\,\mathrm{d}\theta \wedge \mathrm{d}\phi$ integrates to $\int \frac{i}{2\pi} F = -1$. So in our conventions the Hopf bundle, with the Berry connection, has first Chern number -1. (Using $\bar{z}$ instead of z gives the conjugate bundle, with $+1$.)

Physical picture: Berry phase of a spin

Carry a spin-$\frac{1}{2}$ state adiabatically around a loop C of Bloch vectors, always choosing the state aligned with the instantaneous direction. The state returns with an extra phase $e^{-\oint_C s^* \mathcal{A}}$, which by Stokes is $e^{-\int_\Sigma F}$ over a cap Σ bounded by C: minus one half of the solid angle, $e^{-i\Omega/2}$ (the sign depends on the direction of travel and on the convention for $\mathcal{A}$). This is Berry's phase. The "monopole in parameter space" Berry found is the curvature of the Hopf bundle, and the fact that its total flux is an integer is the fact that the Hopf bundle is a genuine (non-trivial) $U(1)$ bundle.

Every two fibres link

Stereographic projection from $(0, 1) \in S^3$ sends $S^3 \setminus \{\mathrm{pt}\}$ to $\mathbb{R}^3$. The fibre over the south pole ($z_1 = 0$) passes through the projection point and becomes the vertical axis. The fibre over the north pole ($z_2 = 0$) becomes the unit circle in the horizontal plane. Every other fibre becomes a circle that winds once around the axis and once through the unit circle. *Any two fibres are linked exactly once* (Figure 1.51). The linking number is the **Hopf invariant**. It cannot change under deformation, which is the geometric reason the Hopf map is not homotopic to a constant: $\pi_3(S^2) \neq 0$.

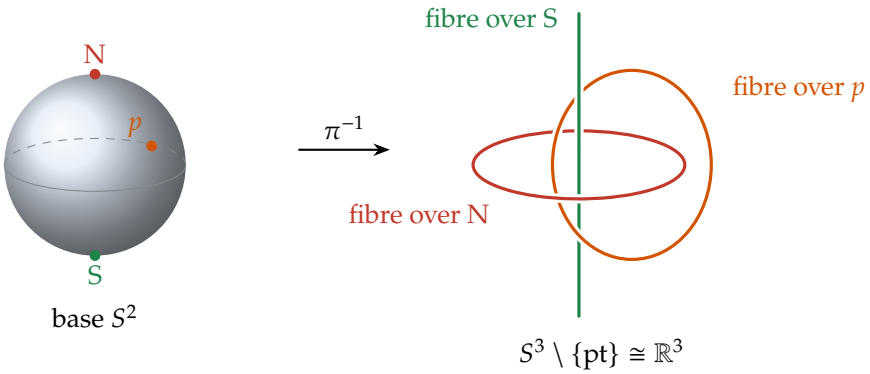


Figure 1.51: The Hopf fibration seen in $\mathbb{R}^3$ after stereographic projection of S^3. Each point of the base S^2 (left) has a circle as its fibre (right). The fibre over the south pole is the vertical axis (closed up through the point at infinity), the fibre over the north pole is the horizontal unit circle, and the fibre over a generic point p is a circle threading both. Any two fibres are linked once. The whole of S^3 is filled by such circles, nested on tori.

Watch out: Locally a product, globally not

Locally S^3 looks like $S^2 \times S^1$. Globally it is not: S^3 is simply connected, while $\pi_1(S^2 \times S^1) = \mathbb{Z}$. If the Hopf bundle were trivial, $\pi_1(S^3)$ would be $\mathbb{Z}$. So the Hopf bundle is non-trivial, and it has no global section. (A global section would be a global choice of phase for spin states:

impossible, as every physicist who has tried to write a spin coherent state without a Dirac string knows.)

Try it: The quaternionic Hopf fibration

Replace $\mathbb{C}$ by the quaternions $\mathbb{H}$. The unit sphere $S^7 \subset \mathbb{H}^2$ maps to $\mathbb{H}P^1 = S^4$ with fibre the unit quaternions $S^3 = \mathrm{SU}(2)$. The clutching function on the equator S^3 of S^4 is $q \mapsto q$ (or its inverse), which has degree one. What bundle over S^4 is this? *Answer:* the principal SU(2) bundle of instanton number ± 1, generating $\pi_3(\mathrm{SU}(2)) = \mathbb{Z}$ in the clutching table above. Its natural connection is the BPST instanton of Chapter 8.

1.5.8 The tangent bundle of S^2 and the hairy ball

We return to Nash's first example and show that TS^2 is not trivial. Use the complex stereographic coordinates of the sphere: z on $O_0 = S^2 \setminus \{N\}$ and w on $O_1 = S^2 \setminus \{S\}$, related by $w = 1/z$ on the overlap. The real 2-dimensional tangent plane at each point is a complex line, spanned by ∂_z (we identify $v = v^x \partial_x + v^y \partial_y$ with the complex number $v^z = v^x + i v^y$).

The transition function. A tangent vector $v = v^z \partial_z = v^w \partial_w$ has components related by the chain rule (1.100):

$$\begin{aligned} v^w &\overset{(1)}{=} \frac{\mathrm{d}w}{\mathrm{d}z}\, v^z \\ &\overset{(2)}{=} -\frac{1}{z^2}\, v^z. \end{aligned} \tag{1.121}$$

(1) The 1×1 complex Jacobian of the holomorphic change of coordinates.
(2) $w = 1/z$, so $\mathrm{d}w/\mathrm{d}z = -1/z^2$.

On the equator $z = e^{i\phi}$ this is $g_{10}(\phi) = -e^{-2i\phi}$: the transition function winds *twice* around $U(1)$ (the sign -1 is a constant and can be deformed away). By the clutching construction, TS^2 is the $U(1)$ bundle, or complex line bundle, of charge $|n| = 2$. It is not trivial, and $2 = \chi(S^2)$.

Key idea: The hairy ball theorem

Every continuous tangent vector field on S^2 vanishes somewhere. "You cannot comb a hairy ball flat."

Suppose v were a nowhere-vanishing vector field. On the closed disc $|z| \le 1$ (the southern hemisphere in the z chart) its component v^z is a nowhere-zero complex function. The winding number of $v^z(e^{i\phi})$ around 0, as ϕ goes from 0 to 2π, is therefore zero: the loop $\phi \mapsto v^z(re^{i\phi})$ can be shrunk to the constant $v^z(0)$ by letting $r \to 0$, without passing through zero. By the same argument on the disc $|w| \le 1$, the winding of v^w around the equator is zero. (On the equator $w = e^{-i\phi}$, so ϕ runs the opposite way around the w disc; this reverses the sign of a winding number but does not change zero.) But (1.121) gives

$$\begin{aligned} \mathrm{wind}_\phi\big(v^w\big) &\overset{(1)}{=} \mathrm{wind}_\phi\big(-e^{-2i\phi}\big) + \mathrm{wind}_\phi\big(v^z\big) \\ &\overset{(2)}{=} -2 + 0 = -2 \neq 0, \end{aligned}$$

(1) Winding numbers add under multiplication of non-vanishing functions (Example 1.59: $\deg(fg) = \deg f + \deg g$).
(2) The factor $-e^{-2i\phi}$ winds -2 times; v^z winds 0 times.

a contradiction. So v must vanish somewhere. The same argument shows TS^2 has no nowhere-zero section, hence is not trivial.

Worked example 1.92: Two combings of the sphere and their zeros

Rotation field $v = \partial_\phi$ (the velocity field of a rigid rotation about the z axis). It vanishes at the two poles. Near each pole it circulates, like $-y\partial_x + x\partial_y$, which has index +1 (the vector turns once as we go once around the zero). Total index $1 + 1 = 2$.
Dipole field v = the projection onto TS^2 of the constant vector $\hat{e}_3$ *rotated* so that it flows out of the north pole, around, and back in. The standard example is the pushforward of ∂_x on the plane under inverse stereographic projection from the north pole: on the z chart it is $v^z = 1$ (no zeros there), and in the w chart (1.121) gives $v^w = -1/z^2 = -w^2$, a double zero at $w = 0$, of index 2. Total index 2 again.
In both cases the indices add to $\chi(S^2) = 2$. This is the Poincaré–Hopf theorem: for any vector field with isolated zeros on a closed manifold, the sum of the indices is χ. For the torus $\chi = 0$, and indeed ∂_{θ_1} combs the torus with no zeros.

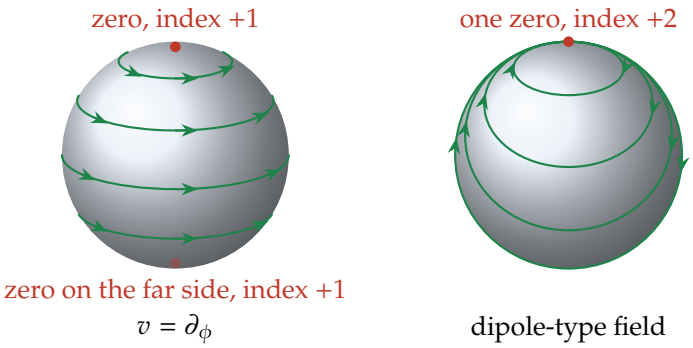


Figure 1.52: Two attempts to comb the sphere. Left: circulation around the axis leaves a zero at each pole, each of index +1. Right: flowing from a single point around the sphere and back into it gives one zero of index +2. However the vector field is chosen, the indices of its zeros sum to $\chi(S^2) = 2$, the winding of the transition function of TS^2.

Physical picture: Where hairy balls appear in physics

- There is always a point on Earth with no horizontal wind.
- The transverse polarisation vectors $\vec{\epsilon}(\hat{k})$ of a photon, as $\hat{k}$ ranges over the sphere of directions, are tangent vectors to S^2. They cannot be chosen continuously for all directions. The helicity basis $\vec{\epsilon}_\pm$ lives in two line bundles with Chern numbers ± 2, the photon's "monopole charge" in momentum space.
- Only S^1, S^3 and S^7 among spheres have trivial tangent bundles (Adams). S^1 and S^3 are the Lie groups $U(1)$ and SU(2), and a Lie group always has a global frame of left-invariant vector fields.

1.5.9 Pulling back cohomology

So far we have built bundles by hand. Nash now turns to the question every physicist asks of a new object: *how do we tell two of them apart?* The tool is a very simple operation, the pullback, applied first to cohomology and then to bundles.

Let $f : N \to M$ be a smooth map. A differential form on M can be *pulled back* to N: evaluate

it on the pushed-forward vectors. In coordinates y^a on N and x^μ on M, with $x^\mu = f^\mu(y)$, we substitute and use the chain rule,

$$f^*\big(\omega_\mu(x)\,\mathrm{d}x^\mu\big) = \omega_\mu(f(y))\,\frac{\partial f^\mu}{\partial y^a}\,\mathrm{d}y^a. \tag{1.122}$$

(This is the "change of variables" of Chapter T1.) Two facts make f^* pass to cohomology.

- f^* commutes with d: $\mathrm{d}(f^*\omega) = f^*(\mathrm{d}\omega)$. So closed forms go to closed forms and exact forms to exact forms, $f^*(\mathrm{d}\alpha) = \mathrm{d}(f^*\alpha)$.
- f^* respects the wedge product: $f^*(\omega \wedge \nu) = f^*\omega \wedge f^*\nu$.

Hence we get a well-defined map

$$f^* : H^p(M;G) \longrightarrow H^p(N;G), \qquad f^*(a \cup b) = f^*a \cup f^*b, \tag{1.123}$$

a **ring homomorphism** of cohomology rings. Notice the direction: f goes from N to M, but f^* goes from M to N. Cohomology is *contravariant*. (With the definition $H^n(X;G) = [X, K(G,n)]_0$ of Section 1.4.2 this is even simpler: a class on M is a map $c : M \to K(G,n)$, and f^*c is the composite $c \circ f : N \to K(G,n)$.)

Worked example 1.93: Pulling back the angle form along $z \mapsto z^k$

Let $f_k : S^1 \to S^1$, $e^{i\theta} \mapsto e^{ik\theta}$, so in the angle coordinate $\theta \mapsto k\theta$. The generator of $H^1(S^1;\mathbb{Z})$ is represented by the form $x = \mathrm{d}\theta/2\pi$, whose integral is 1. Then

$$\int_{S^1} f_k^* x \overset{(1)}{=} \int_0^{2\pi} \frac{1}{2\pi}\,\frac{\mathrm{d}(k\theta)}{\mathrm{d}\theta}\,\mathrm{d}\theta \overset{(2)}{=} k.$$

(1) Apply (1.122): the coefficient of $\mathrm{d}\theta$ picks up the derivative of $\theta \mapsto k\theta$.
(2) The integrand is the constant $k/2\pi$.

So $f_k^* x = k\,x$: on $H^1(S^1;\mathbb{Z}) = \mathbb{Z}$ the pullback is multiplication by the winding number k. In the same way a map $f : S^2 \to S^2$ of degree k multiplies the class of the area form by k. Pullback turns "how many times does the map wrap" into linear algebra.

Key idea: Homotopic maps pull back equally

If $f_0 \simeq f_1 : N \to M$, then $f_0^* = f_1^*$ on cohomology. So f^* depends only on the homotopy class $[f]$, and cohomology is an invariant of homotopy type (as Nash said in §4).

Here is the de Rham reason. Let $F : N \times [0,1] \to M$ be the homotopy, with $F(y,0) = f_0(y)$ and $F(y,1) = f_1(y)$, and let ω be a closed form on M. Split the pulled-back form on the cylinder $N \times [0,1]$ into the part without $\mathrm{d}t$ and the part with it, $F^*\omega = \alpha(t) + \mathrm{d}t \wedge \beta(t)$, where $\alpha(t), \beta(t)$ are forms on N depending on t. Write d_N for the exterior derivative in the N directions. Then

$$\begin{aligned} 0 &\overset{(1)}{=} F^*(\mathrm{d}\omega) = \mathrm{d}(F^*\omega) \\ &\overset{(2)}{=} \mathrm{d}_N\alpha + \mathrm{d}t \wedge \partial_t\alpha - \mathrm{d}t \wedge \mathrm{d}_N\beta = \mathrm{d}_N\alpha + \mathrm{d}t \wedge \big(\partial_t\alpha - \mathrm{d}_N\beta\big). \end{aligned}$$

(1) ω is closed, and d commutes with pullback.
(2) The full d on $N \times [0,1]$ is $\mathrm{d}_N + \mathrm{d}t \wedge \partial_t$; moving d_N past $\mathrm{d}t$ costs a sign.

The part with $\mathrm{d}t$ must vanish on its own, so $\partial_t \alpha = \mathrm{d}_N \beta$. Now

$$\begin{aligned} f_1^*\omega - f_0^*\omega &\overset{(1)}{=} \alpha(1) - \alpha(0) \\ &\overset{(2)}{=} \int_0^1 \partial_t \alpha \, \mathrm{d}t \\ &\overset{(3)}{=} \mathrm{d}_N \Big(\int_0^1 \beta(t) \, \mathrm{d}t \Big). \end{aligned} \tag{1.124}$$

(1) At $t = 0$ and $t = 1$ the $\mathrm{d}t$ part of $F^*\omega$ does not contribute to the restriction to $N \times \{t\}$, which is $f_t^*\omega$.
(2) Fundamental theorem of calculus in t.
(3) $\partial_t \alpha = \mathrm{d}_N \beta$, and d_N commutes with the t integral.

The difference is exact, so $[f_1^*\omega] = [f_0^*\omega]$. This single identity is behind Nash's remark that "two homotopic maps produce identical pullbacks".

1.5.10 Pullback bundles

Bundles can be pulled back too, and this is the operation that will classify them.

Working definition 1.38 (pullback bundle)

Given a bundle $\pi : E \to M$ and a map $f : N \to M$, the **pullback bundle** $f^*E \to N$ has as its fibre over $y \in N$ the fibre of E over $f(y)$:

$$f^*E = \{(y, e) \in N \times E : \ f(y) = \pi(e)\}, \qquad (f^*E)_y = E_{f(y)}. \tag{1.125}$$

Operationally: if E has transition functions $g_{\alpha\beta}$ on the cover $\{O_\alpha\}$ of M, then f^*E has transition functions $g_{\alpha\beta} \circ f$ [at $y \in N$, use the gluing $g_{\alpha\beta}(f(y))$ of E at the point $f(y)$] on the cover $\{f^{-1}(O_\alpha)\}$ [the points of N that f sends into O_α] of N.

The cocycle conditions (1.104) for $g_{\alpha\beta} \circ f$ are those of $g_{\alpha\beta}$, evaluated at $f(y)$, so they hold automatically. The pullback of a section s of E is the section $y \mapsto s(f(y))$ of f^*E.

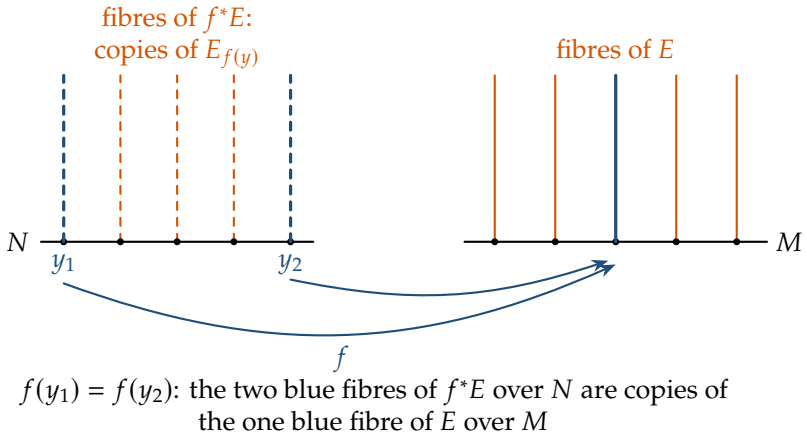


Figure 1.53: The pullback bundle. Each point y of N is assigned a copy of the fibre of E at its image $f(y)$. If f is not injective, several fibres of f^*E are copies of one fibre of E. If f is constant, every fibre of f^*E is the same fibre, and f^*E is the trivial bundle $N \times F$.

Worked example 1.94: Three pullbacks you can do in your head

1. *Restriction.* If $i : S \hookrightarrow M$ is a submanifold, $i^*E = E|_S$. For example TS^2 restricted to the equator S^1 is $TS^1 \oplus$ (normal line), a trivial rank-2 bundle.

2. *Constant maps.* If $f(y) = x_0$ for all y, then $g_{\alpha\beta} \circ f$ is constant on each connected overlap, and it can be gauged away by constant h_α in (1.106). So $f^*E \cong N \times F$ is trivial.

3. *The Möbius band along f_k.* Pull back the Möbius band Mb along $f_k : S^1 \to S^1$ of Example 1.93. Going once around the source circle, f_k goes k times around the target, and the fibre coordinate comes back multiplied by the product of the transition functions met on the way:

$$\begin{aligned} s(2\pi) &\overset{(1)}{=} \big((-1)\cdot\ldots\cdot(-1)\big)\, s(0) \quad (k \text{ factors}) \\ &\overset{(2)}{=} (-1)^k\, s(0). \end{aligned}$$

   (1) Each passage through the twisted overlap B of Example 1.87 contributes -1.
   (2) There are k such passages.

   So f_k^*Mb is the Möbius band for k odd and the cylinder for k even. (The double cover of a Möbius band's core circle untwists it: cut a paper Möbius band down the middle and see.)

Worked example 1.95: Pulling back the monopole bundle

Let $P_n \to S^2$ be the monopole bundle of Example 1.91, with clutching function $g_{NS}(\phi) = e^{in\phi}$. Let $f : S^2 \to S^2$ be the map that, in polar angles, sends $(\theta, \phi) \mapsto (\theta, k\phi)$. It has degree k and preserves the two caps. Then f^*P_n has clutching function

$$\begin{aligned} (g_{NS} \circ f)(\phi) &\overset{(1)}{=} g_{NS}(k\phi) \\ &\overset{(2)}{=} e^{ink\phi}, \end{aligned}$$

(1) Rule for pullback transition functions, restricted to the equator.
(2) Insert (1.112).

so $f^*P_n = P_{nk}$. The charge multiplies by the degree. In terms of characteristic classes, $c_1(f^*P) = f^*c_1(P)$, and f^* on $H^2(S^2;\mathbb{Z}) = \mathbb{Z}$ is multiplication by k (Example 1.93). The bundle and its cohomology class pull back together. This compatibility is the whole point of (1.129) below.

Key idea: Nash's (1.48): homotopic maps give isomorphic pullbacks

If $f_0 \simeq f_1 : N \to M$, then $f_0^*E \cong f_1^*E$.

There is a transparent physicist's argument. Let $F : N \times [0,1] \to M$ be the homotopy, and consider the single bundle F^*E over the cylinder $N \times [0,1]$. Its restrictions to the two ends are f_0^*E and f_1^*E. Choose any connection on F^*E (a partition of unity always provides one) and *parallel transport* along the segments $\{y\} \times [0,1]$. Parallel transport is a linear (or G-equivariant) isomorphism from the fibre at $(y,0)$ to the fibre at $(y,1)$, it depends smoothly on y, and it has an inverse (transport back). So it is a bundle isomorphism $f_0^*E \to f_1^*E$ (Figure 1.54).

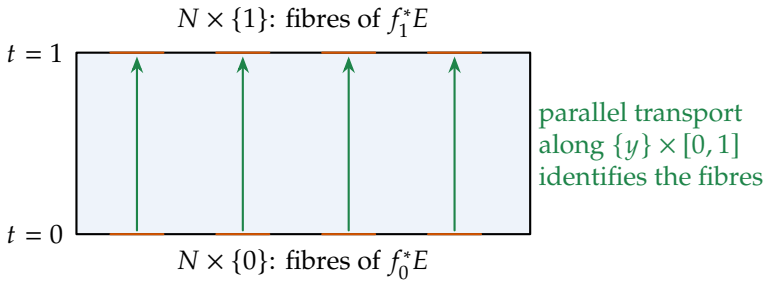


Figure 1.54: Why homotopic maps pull back isomorphic bundles. The homotopy F gives one bundle over the cylinder $N \times [0,1]$. Parallel transport along the vertical segments (green) carries each fibre at the bottom isomorphically to the fibre above it at the top, smoothly in y. That is an isomorphism $f_0^*E \cong f_1^*E$.

Two consequences are used constantly.

- *Every bundle over a contractible space is trivial.* If M is contractible, $\mathrm{id}_M \simeq c$, a constant map. Then $E = \mathrm{id}_M^* E \cong c^*E$, which is trivial by Example 1.94. This is what allowed us to trivialise the bundle over each cap in the clutching construction (Section 1.5.6), and it is why local potentials A_α always exist on contractible patches.

- *Topological sectors are discrete.* A continuous deformation of a gauge field configuration cannot change the bundle it lives on. The bundle can only jump, and the possible jumps are counted by homotopy classes of maps, as we now see.

1.5.11 Universal bundles and classifying spaces

Nash's next point is striking: for each group G there is *one* bundle from which all others are pulled back.

Working definition 1.39 (universal bundle, classifying space)

A principal G-bundle [a bundle whose fibres are copies of G, with G acting on them by right multiplication] $E_G \to BG$ is **universal** if its total space E_G is contractible [it can be shrunk continuously to one of its points] (and G acts freely on it [only the identity of G fixes any point]). The base $BG = E_G/G$ is the **classifying space** of G. *Operationally*: find a contractible space with a free G action; its quotient is BG.

Key idea: Nash's (1.49): bundles are maps into BG

Every principal G-bundle over M is isomorphic to f^*E_G for some map $f : M \to BG$, the **classifying map**, and $f^*E_G \cong f'^*E_G$ exactly when $f \simeq f'$. So

$$\{\text{principal } G\text{-bundles over } M\}/\text{iso} \;\cong\; [M, BG]. \tag{1.126}$$

Why contractibility of E_G? A classifying map for $P \to M$ is the same thing as a G-equivariant map $P \to E_G$ (it then descends to $M = P/G \to E_G/G = BG$). Such maps are sections of the bundle $P \times_G E_G$ over M with fibre E_G. When the fibre is contractible, the obstruction theory of Section 1.6.4 says a section always exists and any two sections are homotopic: all the obstruction groups $H^{k+1}(M; \pi_k(E_G))$ vanish. Existence gives the map f; uniqueness up to homotopy gives (1.126).

How to build a classifying map: the Grassmannian

For vector bundles the construction is completely concrete. Let $E \to M$ be a complex vector bundle of rank k over a compact M, with local trivialisations $\phi_\alpha(e) = (\pi(e), f_\alpha(e))$, $f_\alpha(e) \in \mathbb{C}^k$, on finitely many patches $\alpha = 1, \dots, m$, and a partition of unity ρ_α (smooth functions with $\operatorname{supp}\rho_\alpha \subset O_\alpha$ and $\sum_\alpha \rho_\alpha = 1$). Define

$$\Phi : E \to \mathbb{C}^{km}, \qquad \Phi(e) = \big(\rho_1(\pi e) f_1(e),\ \dots,\ \rho_m(\pi e) f_m(e)\big). \tag{1.127}$$

On each fibre Φ is linear, and it is injective because at least one ρ_α is positive there. So each fibre E_x becomes a k-dimensional subspace $\Phi(E_x) \subset \mathbb{C}^{km}$, and

$$f : M \to \mathrm{Gr}_k(\mathbb{C}^{km}), \qquad x \mapsto \Phi(E_x),$$

is a map into the **Grassmannian** of k-planes. By construction $E = f^*\gamma$, where γ is the **tautological bundle** whose fibre over a plane is the plane itself. Letting $m \to \infty$, the universal case is $BU(k) = \mathrm{Gr}_k(\mathbb{C}^\infty)$.

Worked example 1.96: The tautological line bundle over $\mathbb{CP}^1$ is the Hopf bundle

Over $[z_1 : z_2] \in \mathbb{CP}^1 = S^2$ the tautological line is $\ell = \mathbb{C}\cdot(z_1, z_2) \subset \mathbb{C}^2$. On $U_1 = \{z_1 \neq 0\}$ use the inhomogeneous coordinate $w = z_2/z_1$ and the frame $e_1 = (1, w)$; on $U_2 = \{z_2 \neq 0\}$ use $e_2 = (1/w, 1)$. On the overlap

$$\begin{aligned} e_1 &\overset{(1)}{=} (1, w) = w\,(1/w,\ 1) \\ &\overset{(2)}{=} w\, e_2, \end{aligned}$$

(1) Factor w out of both components.
(2) Definition of e_2.

so the transition function is $w = z_2/z_1$. On the equator $|z_1| = |z_2|$ it is the phase $e^{i(\xi_2 - \xi_1)} = e^{i\phi}$ of (1.119): it winds once. The unit vectors in the tautological lines form $\{(\ell, v) : v \in \ell,\ |v| = 1\} = S^3$, and the projection $v \mapsto \ell$ is exactly the Hopf map. So:

- the Hopf fibration is the unit circle bundle of the tautological line bundle over $\mathbb{CP}^1$;
- the Berry connection of Section 1.5.7 is the connection the tautological bundle inherits from the flat $\mathbb{C}^2$ (project the derivative of a section back onto the line), and its Chern number -1 is the Chern number of the tautological bundle, often written $\mathcal{O}(-1)$;
- its classifying map is the inclusion $\mathbb{CP}^1 \hookrightarrow \mathbb{CP}^\infty = BU(1)$, covered by $S^3 \hookrightarrow S^\infty = EU(1)$.

A table of classifying spaces

G	E_G (contractible)	BG	bundles classified
$\mathbb{Z}$	$\mathbb{R}$	S^1	$\mathbb{Z}$-covers; $[M, S^1] = H^1(M;\mathbb{Z})$
$\mathbb{Z}_2$	$S^\infty \subset \mathbb{R}^\infty$	$\mathbb{RP}^\infty = K(\mathbb{Z}_2, 1)$	double covers, real line bundles; $H^1(M;\mathbb{Z}_2)$
$U(1)$	$S^\infty \subset \mathbb{C}^\infty$	$\mathbb{CP}^\infty = K(\mathbb{Z}, 2)$	complex line bundles; $H^2(M;\mathbb{Z})$
SU(2)	$S^\infty \subset \mathbb{H}^\infty$	$\mathbb{H}P^\infty$	SU(2) gauge sectors; c_2 on 4-manifolds
$U(k)$	k-frames in $\mathbb{C}^\infty$	$\mathrm{Gr}_k(\mathbb{C}^\infty)$	rank-k complex vector bundles
O(k)	k-frames in $\mathbb{R}^\infty$	$\mathrm{Gr}_k(\mathbb{R}^\infty)$	rank-k real vector bundles

Here $S^\infty = \bigcup_N S^N$ is contractible: any sphere mapped into it lies in some S^N of higher dimension, where it can be shrunk (see Example 1.101). The $U(1)$ row is Nash's link between §4 and §5: $\mathbb{CP}^\infty$ is at the same time $K(\mathbb{Z},2)$ and $BU(1)$, so

$$\{\text{complex line bundles over } M\}/\text{iso} = [M, \mathbb{CP}^\infty] = [M, K(\mathbb{Z},2)] = H^2(M;\mathbb{Z}), \tag{1.128}$$

and the isomorphism is $L \mapsto c_1(L)$. Every magnetic charge is realised by a bundle, and the charge determines the bundle.

Physical picture: Topological sectors of a gauge theory

In a gauge theory with group G on a compact Euclidean spacetime M, the field configurations split into sectors labelled by $[M, BG]$. For $U(1)$ on a closed surface Σ the sectors are $H^2(\Sigma;\mathbb{Z}) = \mathbb{Z}$, the total flux. For SU(2) on a closed simply connected 4-manifold they are labelled by the instanton number $c_2 \in \mathbb{Z}$. The path integral is a sum over sectors, each an integral over connections on one fixed bundle.

1.5.12 Characteristic classes

We now have a way to detect non-triviality. If $E = f^*E_G$, pull back the cohomology of BG along the same map:

$$f^* : H^*(BG) \longrightarrow H^*(M), \qquad c_i \longmapsto c_i(E) := f^* c_i. \tag{1.129}$$

The generators c_i of the ring $H^*(BG)$ are the **universal characteristic classes**; their pullbacks are the **characteristic classes** of E. Because f is determined up to homotopy, and homotopic maps pull back equally, $c_i(E)$ depends only on the isomorphism class of E.

Working definition 1.40 (characteristic class)

A characteristic class assigns to each bundle $E \to M$ a class $c(E) \in H^*(M)$ [a cohomology class, e.g. a closed form up to exact ones] that is *natural*: $c(f^*E) = f^*c(E)$ [computing c for the pulled-back bundle gives the pullback of $c(E)$ along f]. *Operational test for non-triviality:* a trivial bundle [one isomorphic to $M \times F$] is pulled back from a point, so all its characteristic classes of positive degree vanish. One non-zero characteristic class proves the bundle is twisted.

Which groups, which classes

For vector bundles over a manifold we can always simplify the structure group. A real vector bundle has structure group $\mathrm{GL}(n,\mathbb{R})$, but a fibre metric (which exists by a partition of unity) lets us use orthonormal frames, with transition functions in $\mathrm{O}(n)$. Nothing topological is lost, because $\mathrm{GL}(n,\mathbb{R})$ deformation retracts onto $\mathrm{O}(n)$. Write $A \in \mathrm{GL}(n,\mathbb{R})$ in polar form $A = OP$ with $O \in \mathrm{O}(n)$ and $P = (A^TA)^{1/2}$ symmetric positive definite, and set $A_t = OP^t$. Then

$$\det A_t \overset{(1)}{=} \det O\,(\det P)^t \overset{(2)}{=} \pm(\det P)^t \neq 0,$$

(1) det is multiplicative, and P^t has eigenvalues λ_i^t where $\lambda_i > 0$ are those of P.
(2) $\det O = \pm 1$ and $\det P > 0$.

so A_t is a path in $\mathrm{GL}(n,\mathbb{R})$ from $A_1 = A$ to $A_0 = O \in \mathrm{O}(n)$, fixing $\mathrm{O}(n)$ throughout. Likewise $\mathrm{GL}(n,\mathbb{C})$ retracts onto $U(n)$. This is Nash's remark that the structure group is "naturally" $\mathrm{O}(n)$ or $U(n)$.

group	$H^*(BG)$	classes	degree	Chern–Weil formula
$U(n)$	$\mathbb{Z}[c_1,\dots,c_n]$	Chern c_i	$2i$	$\det(1+\frac{i}{2\pi}F)=1+c_1+c_2+\cdots$
$\mathrm{O}(n)$	$\mathbb{Z}_2[w_1,\dots,w_n]$ (mod 2)	Stiefel–Whitney w_i	i	none (mod 2)
$\mathrm{O}(n)$	from $H^*(BO(n);\mathbb{Z})$	Pontryagin p_i	$4i$	$\det(1-\frac{F}{2\pi})=1+p_1+\cdots$
$\mathrm{SO}(n)$, n even	extra generator	Euler e	n	$\mathrm{Pf}(F/2\pi)$, $e^2=p_{n/2}$

The Pontryagin classes are defined by complexifying, $p_i(E)=(-1)^i c_{2i}(E\otimes\mathbb{C})$. The Chern–Weil formulas express each class by the curvature F of any connection. That these forms are closed and that their classes do not depend on the connection is the subject of Chapter T2 and Chapter 4.

Expanding the determinant with $X=\frac{i}{2\pi}F$ gives the first two Chern classes, which are the ones a physicist meets most:

$$\begin{aligned}\det(1+X) &\overset{(1)}{=} 1+\operatorname{tr}X+\tfrac{1}{2}\big[(\operatorname{tr}X)^2-\operatorname{tr}(X^2)\big]+\cdots\\ &\overset{(2)}{=} 1+\frac{i}{2\pi}\operatorname{tr}F+\frac{1}{8\pi^2}\big[\operatorname{tr}(F\wedge F)-\operatorname{tr}F\wedge\operatorname{tr}F\big]+\cdots .\end{aligned}\tag{1.130}$$

(1) For a diagonalisable X with eigenvalues x_i, $\det(1+X)=\prod(1+x_i)=1+\sum x_i+\sum_{i<j}x_ix_j+\cdots$, and $\sum_{i<j}x_ix_j=\frac{1}{2}[(\sum x_i)^2-\sum x_i^2]$.

(2) Insert $X=\frac{i}{2\pi}F$; $(\frac{i}{2\pi})^2=-\frac{1}{4\pi^2}$, and the $\frac{1}{2}$ makes $\frac{1}{8\pi^2}$.

So $c_1=\frac{i}{2\pi}\operatorname{tr}F$ and, for $\mathrm{SU}(n)$ where $\operatorname{tr}F=0$, $c_2=\frac{1}{8\pi^2}\operatorname{tr}(F\wedge F)$. These are the monopole charge and the instanton number.

Worked example 1.97: Characteristic classes of the bundles we have met

1. *Monopole.* $c_1(P_n)[S^2]=n$, computed in (1.114).
2. *Möbius band.* $w_1(\mathrm{Mb})\in H^1(S^1;\mathbb{Z}_2)=\mathbb{Z}_2$ is the product of the signs of the transition functions around the circle, the invariant -1 found in Example 1.87. It is the non-zero element. For a manifold, $w_1(TM)=0$ exactly when M is orientable (Example 1.89).
3. *Tangent bundle of S^2.* As a complex line bundle, its transition function winds twice (1.121), so $c_1(TS^2)[S^2]=2$. As an oriented real rank-2 bundle, its Euler class is the same number, $e(TS^2)[S^2]=2=\chi(S^2)$. The Gauss–Bonnet theorem is the Chern–Weil form of this: $\frac{1}{2\pi}\int_{S^2}K\,\mathrm{d}A=2$.
4. *Tautological bundle.* $c_1(\mathcal{O}(-1))[\mathbb{CP}^1]=-1$, the Berry-phase computation of Section 1.5.7.

Worked example 1.98: The universal class for line bundles

The cohomology ring of $\mathbb{CP}^\infty$ is a polynomial ring, $H^*(\mathbb{CP}^\infty;\mathbb{Z})=\mathbb{Z}[c]$ with c in degree 2. Here is why. $\mathbb{CP}^N$ is built from one cell in each even dimension $0,2,\dots,2N$: $\mathbb{CP}^N=\mathbb{CP}^{N-1}\cup(\text{a }2N\text{-cell }\{[z_0:\cdots:z_{N-1}:1]\}\cong\mathbb{C}^N)$. With cells only in even dimensions, every cellular boundary map goes from an even degree to an odd one, where there are no cells, so all boundary maps vanish and

$$H^{2j}(\mathbb{CP}^N;\mathbb{Z})=\mathbb{Z}\quad(0\le j\le N),\qquad H^{\mathrm{odd}}=0.$$

The ring structure: the class $c\in H^2$ is Poincaré dual to a hyperplane $\mathbb{CP}^{N-1}$, and c^j is dual to the intersection of j generic hyperplanes, a $\mathbb{CP}^{N-j}$. For $j=N$ that is a single point, so c^N

generates H^{2N} and $H^*(\mathbb{CP}^N) = \mathbb{Z}[c]/(c^{N+1})$. Letting $N \to \infty$ removes the relation. For $\mathbb{CP}^2$ the pairing on H^2 is $\int c \cup c = 1$: two lines in the projective plane meet in exactly one point. This 1×1 matrix (1) is the intersection form of $\mathbb{CP}^2$ that we meet again in Section 1.6.8. The first Chern class of a line bundle L with classifying map f is $c_1(L) = f^*c$.

Watch out: Characteristic classes detect, but need not classify

A non-zero class proves non-triviality. The converse can fail: there are non-trivial bundles with all characteristic classes zero (for example non-trivial SU(2) bundles over S^5, classified by $\pi_4(\mathrm{SU}(2)) = \mathbb{Z}_2$, which no cohomology class of S^5 in degree 2 or 4 can see). For $U(1)$ bundles, however, c_1 is a complete invariant, by (1.128).

1.5.13 The exact homotopy sequence of a bundle

We now turn from cohomology to homotopy. A bundle is written

$$F \xrightarrow{i} E \xrightarrow{\pi} M, \tag{1.131}$$

where i is the inclusion of the fibre $F = \pi^{-1}(*)$ over the base point. Its homotopy groups are tied together by Nash's "valuable" long exact sequence

$$\cdots \to \pi_n(F) \xrightarrow{i_*} \pi_n(E) \xrightarrow{\pi_*} \pi_n(M) \xrightarrow{\partial} \pi_{n-1}(F) \xrightarrow{i_*} \pi_{n-1}(E) \to \cdots \to \pi_0(E) \to \pi_0(M) \to 0. \tag{1.132}$$

Recall from Section 1.4.4 that *exact* means: at every spot the image of the incoming map equals the kernel of the outgoing one. The maps i_* and π_* are just composition with i and π. The interesting map is ∂.

The boundary map: lift a disc and look at its edge

An element of $\pi_n(M)$ can be represented by a map $\sigma : D^n \to M$ from the n-disc which sends the whole boundary sphere $\partial D^n = S^{n-1}$ to the base point $*$ (collapsing the boundary of a disc gives a sphere). The disc is contractible, so σ *lifts*: there is $\tilde{\sigma} : D^n \to E$ with $\pi \circ \tilde{\sigma} = \sigma$. (Over a contractible domain the pulled-back bundle σ^*E is trivial, by Section 1.5.10, and a trivial bundle has sections.) On the boundary, $\pi(\tilde{\sigma}(x)) = *$, so $\tilde{\sigma}$ maps S^{n-1} into the fibre F. Define

$$\partial[\sigma] = \left[\tilde{\sigma}|_{S^{n-1}} : S^{n-1} \to F\right] \in \pi_{n-1}(F). \tag{1.133}$$

With the three maps defined, exactness can be checked by the same lifting idea at each spot.

- *At* $\pi_n(E)$. A sphere in F projects to the point $*$, so $\pi_* i_* = 0$. Conversely, if a sphere τ in E projects to a contractible sphere in M, lift the contraction (homotopy lifting, Section 1.5.15); it deforms τ into a sphere lying over $*$, that is, in F.
- *At* $\pi_n(M)$. If $\sigma = \pi \circ \tau$ comes from E, we may lift with $\tilde{\sigma} = \tau$, whose edge is a point: $\partial \pi_* = 0$. Conversely, if the edge of the lift can be contracted in F, we can use the contraction to close up the lifted disc into a sphere in E projecting to σ.
- *At* $\pi_{n-1}(F)$. An edge $\partial[\sigma]$ bounds the disc $\tilde{\sigma}$ in E, so it is contractible in E: $i_* \partial = 0$. Conversely, a sphere in F that bounds a disc in E is the edge of that disc, whose projection is a sphere in M.

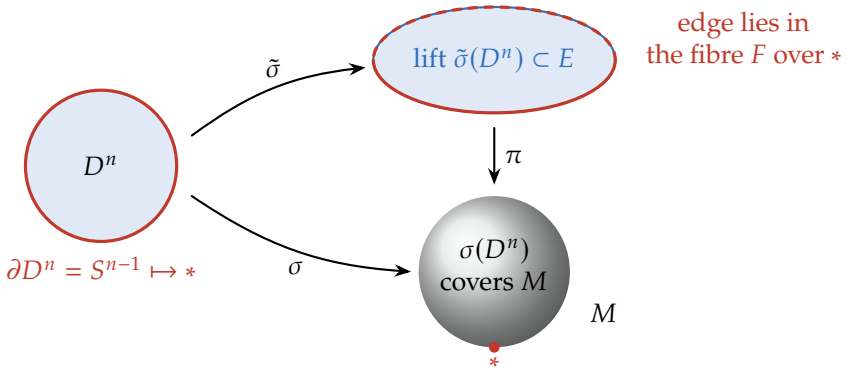


Figure 1.55: The boundary map $\partial : \pi_n(M) \to \pi_{n-1}(F)$. A sphere in the base is a disc whose edge is sent to the base point. Lift the disc into the total space. The lifted disc's edge must lie over $*$, that is, in the fibre F (red). That edge is an $(n-1)$-sphere in F, and its class is $\partial[\sigma]$. If the bundle were trivial we could lift with a constant fibre coordinate and the edge would be a point.

Key idea: The boundary map of a bundle over a sphere is its clutching function

For a principal G-bundle over S^n with clutching function $g : S^{n-1} \to G$ (1.110), the boundary map sends the generator of $\pi_n(S^n) = \mathbb{Z}$ to the class of the clutching function:

$$\partial\,[\mathrm{id}_{S^n}] = [g] \in \pi_{n-1}(G). \tag{1.134}$$

To see it, put the base point at the south pole and view S^n as a disc D^n whose boundary is squashed to the south pole: the meridians from the north pole are the radii of the disc, and all their end points go to S. Lift as follows. Over the northern cap use the constant section $f_N = 1$ of the northern trivialisation. At a point x of the equator this reads, in the southern trivialisation, $f_S = g_{SN}(x)$. Continue down each meridian to the south pole keeping f_S constant, equal to $g_{SN}(x)$. The lift is continuous, and at the south pole the meridian through x arrives at the point $g_{SN}(x)$ of the fibre G. So the edge of the lifted disc is the map $x \mapsto g_{SN}(x)$ from S^{n-1} to the fibre: that is the clutching function (up to the inverse, which only changes an overall sign in $\pi_{n-1}(G)$). A trivial bundle has null-homotopic clutching function and $\partial = 0$; the more the bundle twists, the more the edge winds.

Computations

Worked example 1.99: The trivial bundle, and the covering $\mathbb{Z} \to \mathbb{R} \to S^1$

Trivial bundle. For $E = M \times F$ the section $x \mapsto (x, f_0)$ lifts every sphere in M with a point as edge, so $\partial = 0$. The sequence breaks into short exact pieces $0 \to \pi_n(F) \to \pi_n(M \times F) \to \pi_n(M) \to 0$, and in fact $\pi_n(M \times F) = \pi_n(M) \oplus \pi_n(F)$ (a sphere in a product is a pair of spheres). This is Nash's remark; for a non-trivial bundle the sequence does not split like this, and it carries information.

The covering $\exp : \mathbb{R} \to S^1$, $t \mapsto e^{2\pi i t}$, with fibre $\mathbb{Z}$. The total space $\mathbb{R}$ is contractible, so $\pi_n(\mathbb{R}) = 0$ and the sequence reads

$$\underbrace{\pi_n(\mathbb{R})}_{0} \to \pi_n(S^1) \xrightarrow{\partial} \pi_{n-1}(\mathbb{Z}) \to \underbrace{\pi_{n-1}(\mathbb{R})}_{0}.$$

Hence $\pi_n(S^1) \cong \pi_{n-1}(\mathbb{Z})$. A discrete space has $\pi_0 =$ its set of points and no higher homotopy,

so

$$\pi_1(S^1) \overset{(1)}{=} \pi_0(\mathbb{Z}) = \mathbb{Z},$$
$$\pi_n(S^1) \overset{(2)}{=} \pi_{n-1}(\mathbb{Z}) = 0 \qquad (n \geq 2).$$

(1) Exactness with zeros on both sides. The boundary map is completely explicit: the loop $t \mapsto e^{2\pi i k t}$, $t \in [0,1]$, lifts to the path $t \mapsto kt$ in $\mathbb{R}$ starting at 0, whose end point $k \in \mathbb{Z}$ lies in the fibre. ∂ is the winding number.

(2) For $n \geq 2$, π_{n-1} of a discrete space is trivial.

This recovers $S^1 = K(\mathbb{Z}, 1)$ (Section 1.4.1) from the exact sequence, as promised.

Worked example 1.100: The Hopf fibration gives $\pi_3(S^2) = \mathbb{Z}$

Apply (1.132) to $S^1 \to S^3 \to S^2$:

$$\cdots \to \pi_n(S^1) \to \pi_n(S^3) \xrightarrow{\pi_*} \pi_n(S^2) \xrightarrow{\partial} \pi_{n-1}(S^1) \to \pi_{n-1}(S^3) \to \cdots$$

We know $\pi_k(S^1) = 0$ for $k \geq 2$ (Example 1.99) and $\pi_k(S^3) = 0$ for $k = 1, 2$ (Chapter 1, §3).

- $n \geq 3$. Both neighbours $\pi_n(S^1)$ and $\pi_{n-1}(S^1)$ vanish, since $n - 1 \geq 2$:

$$\pi_3(S^2) \overset{(1)}{=} \pi_3(S^3) \overset{(2)}{=} \mathbb{Z}. \tag{1.135}$$

  (1) $0 \to \pi_3(S^3) \xrightarrow{\pi_*} \pi_3(S^2) \to 0$ is exact, so π_* is an isomorphism.

  (2) $\pi_n(S^n) = \mathbb{Z}$, generated by the identity (the degree).

  The generator of $\pi_3(S^2)$ is $\pi_*[\mathrm{id}_{S^3}] = [\pi]$: *the Hopf map itself*. Its multiples are detected by the Hopf invariant, the linking number of two fibres (Figure 1.51). The same step gives $\pi_n(S^2) = \pi_n(S^3)$ for every $n \geq 3$: for example $\pi_4(S^2) = \pi_4(S^3) = \mathbb{Z}_2$ and $\pi_6(S^2) = \pi_6(S^3) = \mathbb{Z}_{12}$, entries of the table in Chapter 1.

- $n = 2$. $\underbrace{\pi_2(S^3)}_{0} \to \pi_2(S^2) \xrightarrow{\partial} \pi_1(S^1) \to \underbrace{\pi_1(S^3)}_{0}$, so $\pi_2(S^2) \cong \pi_1(S^1) = \mathbb{Z}$. By (1.134), ∂ sends the identity of S^2 to the clutching function $e^{i\phi}$ of (1.119), which winds once: ∂ is an isomorphism, as it must be.

Physical picture: Hopf solitons and skyrmions

Maps $\mathbb{R}^3 \cup \{\infty\} = S^3 \to S^2$ are the field configurations of a unit 3-vector field $\vec{n}(x)$ (a Heisenberg ferromagnet, or the Faddeev–Skyrme model) that tends to a constant at infinity. The sectors are $\pi_3(S^2) = \mathbb{Z}$. The integer is the linking number of the preimages of two target points, the Hopf invariant. Knotted solitons ("hopfions") carry it. The formula $H = \frac{1}{16\pi^2}\int_{S^3} a \wedge \mathrm{d}a$, where a is any real 1-form with $\mathrm{d}a = f^*\omega$ the pullback of the area form ω of S^2 (total area 4π), is an abelian Chern–Simons integral: the first appearance of Chern–Simons theory, Chapter 12.

Worked example 1.101: $\mathbb{CP}^\infty = K(\mathbb{Z},2)$ and $\pi_n(BG) = \pi_{n-1}(G)$

First, S^∞ is contractible. A map $S^k \to S^\infty$ has compact image, so it lies in some S^N; enlarge N so that $N > k$; then $\pi_k(S^N) = 0$ and the map is null-homotopic. So $\pi_k(S^\infty) = 0$ for all k (and with a little more work S^∞ contracts to a point). Apply (1.132) to $U(1) \to S^\infty \to \mathbb{CP}^\infty$:

$$\pi_n(\mathbb{CP}^\infty) \overset{(1)}{=} \pi_{n-1}(U(1)) \overset{(2)}{=} \begin{cases} \mathbb{Z} & n = 2, \\ 0 & n \neq 2. \end{cases}$$

(1) $\pi_n(S^\infty) = \pi_{n-1}(S^\infty) = 0$ on both sides of ∂.
(2) $U(1) = S^1$, whose only non-zero homotopy group is $\pi_1 = \mathbb{Z}$ (Example 1.99).

So $\mathbb{CP}^\infty$ is a $K(\mathbb{Z},2)$, as claimed in Example 1.58. The finite stages show how the higher groups die: $S^1 \to S^{2N+1} \to \mathbb{CP}^N$ gives $\pi_2(\mathbb{CP}^N) = \mathbb{Z}$ and $\pi_k(\mathbb{CP}^N) = \pi_k(S^{2N+1})$ for $k \geq 3$, which is zero for $k \leq 2N$. As $N \to \infty$ nothing survives above degree 2.
The same argument for any universal bundle $G \to E_G \to BG$, with E_G contractible, gives

$$\pi_n(BG) \cong \pi_{n-1}(G). \tag{1.136}$$

For connected G this matches the clutching construction: bundles over S^n are $[S^n, BG] = \pi_n(BG) = \pi_{n-1}(G)$, as in (1.111). It also matches $\Omega BG \simeq G$ from Section 1.4.3.

Worked example 1.102: Two fibrations of $\mathrm{SO}(3)$ and the belt trick

Route 1: the double cover. $\mathrm{SU}(2) = S^3$ double covers $\mathrm{SO}(3)$, with fibre $\{\pm 1\} = \mathbb{Z}_2$ (Chapter 1, belt trick). The sequence $\underbrace{\pi_1(S^3)}_{0} \to \pi_1(\mathrm{SO}(3)) \xrightarrow{\partial} \pi_0(\mathbb{Z}_2) \to \underbrace{\pi_0(S^3)}_{\mathrm{pt}}$ gives $\pi_1(\mathrm{SO}(3)) = \mathbb{Z}_2$.
For $n \geq 2$, $\pi_n(\mathrm{SO}(3)) = \pi_n(S^3)$: $\pi_2(\mathrm{SO}(3)) = 0$, $\pi_3(\mathrm{SO}(3)) = \mathbb{Z}$.
Route 2: the unit tangent bundle of S^2. A rotation R is fixed by where it sends the frame $(\hat{e}_3, \hat{e}_1)$: the point $p = R\hat{e}_3 \in S^2$ and the unit tangent vector $R\hat{e}_1 \perp p$. So $\mathrm{SO}(3)$ is the bundle of unit tangent vectors of S^2, a principal $\mathrm{SO}(2)$ bundle $\mathrm{SO}(2) \to \mathrm{SO}(3) \to S^2$ whose clutching function is that of TS^2, winding twice (1.121). The sequence reads

$$\underbrace{\pi_2(\mathrm{SO}(3))}_{0} \to \underbrace{\pi_2(S^2)}_{\mathbb{Z}} \xrightarrow{\partial} \underbrace{\pi_1(\mathrm{SO}(2))}_{\mathbb{Z}} \to \pi_1(\mathrm{SO}(3)) \to \underbrace{\pi_1(S^2)}_{0}.$$

By (1.134), ∂ is multiplication by the winding of the clutching function, 2. Exactness then gives

$$\pi_1(\mathrm{SO}(3)) \overset{(1)}{=} \pi_1(\mathrm{SO}(2))/\operatorname{Im}\partial \overset{(2)}{=} \mathbb{Z}/2\mathbb{Z} = \mathbb{Z}_2.$$

(1) The map $\pi_1(\mathrm{SO}(2)) \to \pi_1(\mathrm{SO}(3))$ is onto (the next group $\pi_1(S^2)$ is zero) with kernel $\operatorname{Im}\partial$.
(2) ∂ is multiplication by 2.

The two routes agree. The second shows something remarkable: *the belt trick and the hairy ball theorem are the same number* $2 = \chi(S^2)$. A 4π rotation is contractible because the tangent bundle of the sphere twists twice.

Try it: Homotopy groups of $\mathrm{SU}(n)$ by induction

$\mathrm{SU}(n)$ acts transitively on the unit sphere $S^{2n-1} \subset \mathbb{C}^n$, and the stabiliser of e_n is $\mathrm{SU}(n-1)$. So $\mathrm{SU}(n-1) \to \mathrm{SU}(n) \to S^{2n-1}$ is a bundle. Show that $\pi_1(\mathrm{SU}(n)) = 0$ and $\pi_3(\mathrm{SU}(n)) = \mathbb{Z}$ for all $n \geq 2$. *Answer:* for $k \leq 2n-3$ the groups $\pi_k(S^{2n-1})$ and $\pi_{k+1}(S^{2n-1})$ vanish, so $\pi_k(\mathrm{SU}(n-1)) \cong \pi_k(\mathrm{SU}(n))$. Starting from $\mathrm{SU}(2) = S^3$: $\pi_1 = 0$ and $\pi_3 = \mathbb{Z}$ are passed up for $n \geq 3$ (for $k = 3$ we need $3 \leq 2n-3$, true for $n \geq 3$). Physically: every $\mathrm{SU}(n)$ instanton sector has an integer charge, and the minimal instanton is an SU(2) instanton sitting in an SU(2) subgroup.

1.5.14 Cohomology of a bundle: no simple rule

Nash warns that "there is no simple relation between the cohomology groups of E, F and M". A trivial bundle obeys Künneth, $P_{M\times F} = P_M P_F$ (Example 1.84). A twisted one does not. The cleanest demonstration is the family of all circle bundles over S^2.

Worked example 1.103: The homology of the charge-n circle bundle over S^2

Let $E_n \to S^2$ be the circle bundle with clutching function $e^{in\phi}$, so that $f_N = e^{in\phi} f_S$ on the equator (Example 1.91). Cover E_n by $E_N = \pi^{-1}(O_N) \cong D^2 \times S^1$ and $E_S = \pi^{-1}(O_S) \cong D^2 \times S^1$. Their intersection is the circle bundle over the equatorial band, $\simeq S^1 \times S^1 = T^2$. On $H_1(T^2) = \mathbb{Z}^2$ use the basis

- b: the fibre circle over a point of the equator;
- a: the equator traced at fixed northern fibre coordinate, $\phi \mapsto (\phi, f_N = 1)$.

In $E_N \cong D^2 \times S^1$, $H_1 = \mathbb{Z}\beta_N$ (the fibre); $b \mapsto \beta_N$, and $a \mapsto 0$ because a bounds the disc $D^2 \times \{1\}$. In E_S, $H_1 = \mathbb{Z}\beta_S$; $b \mapsto \beta_S$, but a has southern fibre coordinate $f_S = e^{-in\phi}$, which winds $-n$ times around the fibre, so $a \mapsto -n\beta_S$. The Mayer–Vietoris sequence (Section 1.4.9, in its homology form)

$$H_1(T^2) \xrightarrow{\Phi} H_1(E_N) \oplus H_1(E_S) \to H_1(E_n) \to \tilde{H}_0(T^2) = 0$$

gives $H_1(E_n) = \operatorname{Coker}\Phi$. In the bases (a, b) and (β_N, β_S),

$$\begin{aligned} H_1(E_n) &\overset{(1)}{=} \mathbb{Z}^2 \Big/ \operatorname{Im}\begin{pmatrix} 0 & 1 \\ -n & 1 \end{pmatrix} \\ &\overset{(2)}{=} \mathbb{Z}^2/\langle (0,-n), (1,1)\rangle \\ &\overset{(3)}{=} \mathbb{Z}/n\mathbb{Z}. \end{aligned}$$

(1) The columns are the images of a and b.
(2) The image is spanned by the columns.
(3) The relation $(1,1)$ sets $\beta_N = -\beta_S$, leaving one generator β_S subject to $n\beta_S = 0$.

Also $H_2(E_n) = \operatorname{Ker}\Phi$ (since $H_2(E_N) = H_2(E_S) = 0$), which is 0 for $n \neq 0$ because $\det\Phi = n \neq 0$. So:

n	E_n	$H_1(E_n)$	$H_2(E_n)$
0	$S^2 \times S^1$	$\mathbb{Z}$	$\mathbb{Z}$
± 1	S^3 (Hopf)	0	0
± 2	$\mathrm{SO}(3) = \mathbb{RP}^3$ (unit tangent bundle of S^2)	$\mathbb{Z}_2$	0
n	lens space $L(n,1) = S^3/\mathbb{Z}_n$	$\mathbb{Z}_n$	0

Same fibre, same base, different total spaces with different homology. The twist n enters as a matrix entry.

Physical picture: The spectral sequence in one sentence

The tool Nash names is the **Serre spectral sequence**. One starts from the "Künneth guess" $E_2^{p,q} = H^p(M; H^q(F))$, laid out on a grid, and corrects it by differentials $d_r : E_r^{p,q} \to E_r^{p+r,q-r+1}$ that encode the twisting. For a circle bundle over S^2 the grid has four entries, at $(p,q) = (0,0), (2,0), (0,1), (2,1)$, and the only differential is $d_2 : E_2^{0,1} = \mathbb{Z} \to E_2^{2,0} = \mathbb{Z}$, which is multiplication by the Euler class n. Its kernel is $H^1(E_n)$ and its cokernel is $H^2(E_n) = \mathbb{Z}_n$, in agreement with the example (torsion in H_1 reappears in H^2, Example 1.76).

One number is insensitive to the twist: the Euler characteristic.

Key idea: χ is multiplicative in bundles

For a bundle $F \to E \to M$ with compact fibre and base, $\chi(E) = \chi(F)\,\chi(M)$.

Triangulate M so finely that E is trivial over each simplex. Each k-simplex σ of M, together with a cell decomposition of F, contributes cells $\sigma \times (\text{cell of } F)$ to E. Counting with signs,

$$\chi(E) \overset{(1)}{=} \sum_{\sigma} \sum_{c} (-1)^{\dim \sigma + \dim c} \overset{(2)}{=} \sum_{\sigma} (-1)^{\dim \sigma}\, \chi(F) = \chi(M)\,\chi(F).$$

(1) χ is the alternating count of cells (Section 1.4.10), and the cells of E are products.
(2) The inner sum over cells c of F is $\chi(F)$; the outer sum is $\chi(M)$.

Checks: the Hopf fibration, $\chi(S^3) = 0 = 0 \cdot 2$; the unit tangent bundle $\mathbb{RP}^3$ of S^2, $0 = 0 \cdot 2$; the Möbius band, which retracts to its core circle, $0 = 1 \cdot 0$.

1.5.15 Fibrations and the homotopy lifting property

Every argument in the last two subsections used one property of a bundle: *maps into the base can be lifted, and so can their deformations*. Nash promotes this property to a definition.

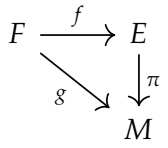
Working definition 1.41 (fibration)

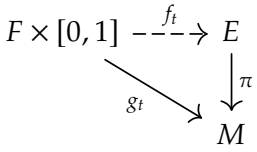
A map $\pi : E \to M$ is a **fibration** if it has the **homotopy lifting property**: for every space F, every map $f : F \to E$ and every homotopy $g_t : F \to M$ [a family of maps depending continuously on $t \in [0,1]$] of its projection $g_0 = \pi \circ f$ [f followed by π, i.e. f pushed down to M], there is a homotopy $f_t : F \to E$ with

$$f_0 = f \qquad \text{and} \qquad \pi \circ f_t = g_t \ \text{ for all } t.$$

Operationally: "whenever you move the shadow, you can move the object so that it casts that shadow."

In diagrams, Nash's (1.53)–(1.55) are





$$g = \pi \circ f, \quad f_0 = f, \quad g_t = \pi \circ f_t. \tag{1.137}$$

The dashed arrow is what the property guarantees.

Worked example 1.104: Bundles and coverings are fibrations

Fibre bundles. Over a single patch $O_\alpha \times F'$ lifting is trivial: keep the fibre coordinate fixed, $f_t(y) = (g_t(y),\ f_\alpha(f(y)))$. For a general homotopy, cut $[0,1]$ into short time steps during which each point stays within one patch, and lift step by step, re-expressing the fibre coordinate with $g_{\alpha\beta}$ when crossing to a new patch. (With a connection there is an even simpler lift: parallel transport along the paths $t \mapsto g_t(y)$.) Nash calls these "locally trivial" fibrations.
Coverings. For $\mathbb{R} \to S^1$ the lift of a path is unique once the starting point is chosen: this is the lifting that computed $\pi_1(S^1)$. Coverings are the fibrations with discrete fibre.

The fibre of a fibration is $F = \pi^{-1}(*)$. The fibres over different points of a path-connected base need not be homeomorphic, but they all have the same homotopy type. The reason is transport: lifting a path γ from x to y for every point of $\pi^{-1}(x)$ at once gives a map $\pi^{-1}(x) \to \pi^{-1}(y)$, and the reversed path gives a map back, whose composites are homotopic to the identities.

Worked example 1.105: A fibration whose fibres are not all homeomorphic

Let $E = \{(x,t) \in \mathbb{R}^2 : 0 \le t \le 1,\ 0 \le x \le t\}$, a closed triangle, and $\pi(x,t) = t$. The fibre over $t > 0$ is the interval $[0,t]$; over $t = 0$ it is a single point. These are not homeomorphic, but all are contractible. Homotopy lifting holds. Given $f(y) = (x_0(y), t_0(y))$ and a homotopy $t_s(y)$ of t_0, define

$$f_s(y) = \big(\min\{x_0(y), t_s(y)\},\ t_s(y)\big),$$

$$\text{check:}\quad f_0(y) \overset{(1)}{=} \big(x_0(y), t_0(y)\big) = f(y), \qquad \pi f_s(y) \overset{(2)}{=} t_s(y).$$

(1) $x_0 \le t_0$ because $f(y) \in E$, so the minimum is x_0.
(2) By construction; and $0 \le \min\{x_0, t_s\} \le t_s$, so $f_s(y) \in E$.

The minimum of continuous functions is continuous, so f_s is a lift. A fibre bundle could never behave like this; a fibration can. This extra flexibility is exactly what makes fibrations useful: every map can be replaced, up to homotopy, by a fibration, and the most important example arises this way.

Nash's next statement is that fibrations have the *same* exact homotopy sequence (1.132). Indeed, our derivation of ∂ and of exactness used only lifting of discs and of homotopies. The spectral sequence relating the cohomology of F, E and M also carries over.

1.5.16 The path fibration

Take as total space the based paths in M:

$$PM = \mathrm{Map}_0([0,1], M) = \{p : [0,1] \to M \text{ continuous},\ p(0) = *\}, \tag{1.138}$$

and let $\pi : PM \to M$ send a path to its end point, $\pi(p) = p(1)$. The fibre over $*$ is the set of paths starting and ending at $*$: the loop space ΩM of Chapter 1. Nash's picture is

$$\Omega M \longrightarrow PM \xrightarrow{\pi} M, \qquad \pi(p) = p(1). \tag{1.139}$$

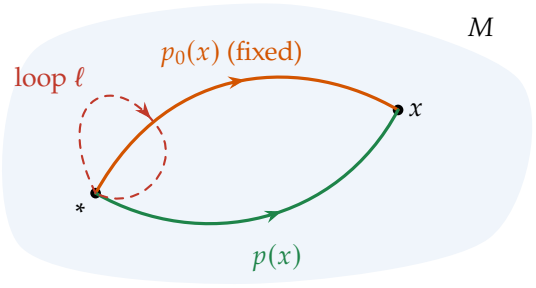


Figure 1.56: The path fibration. The fibre over x consists of all paths from $*$ to x. Fix one of them, $p_0(x)$ (orange). Any other path $p(x)$ (green) differs from it by the loop $p(x) \cdot \overline{p_0(x)}$ at $*$ (go out along p, come back along p_0). Conversely a loop ℓ (red) followed by $p_0(x)$ is a path to x. This identifies the fibre over x with ΩM up to homotopy.

Every fibre is a loop space. Nash's argument, made explicit: fix a path $p_0(x)$ from $*$ to x. Define maps

$$\Omega M \to \pi^{-1}(x), \ \ \ell \mapsto \ell \cdot p_0(x), \qquad \pi^{-1}(x) \to \Omega M, \ \ p \mapsto p \cdot \overline{p_0(x)},$$

where $\cdot$ is concatenation and the bar reverses a path. The composites send $\ell \mapsto \ell \cdot p_0 \cdot \overline{p_0}$ and $p \mapsto p \cdot \overline{p_0} \cdot p_0$, which are homotopic to ℓ and p by retracting the back-and-forth excursion (the same homotopies that make π_1 a group, Chapter 1). So $\pi^{-1}(x) \simeq \Omega M$, and in particular the fibre over $*$ is ΩM itself.

The total space is contractible. Shrink every path back along itself:

$$H_u(p)(t) = p(ut), \qquad u \in [0,1],$$
$$H_1(p) \overset{(1)}{=} p, \qquad H_0(p) \overset{(2)}{=} \text{the constant path at } *.$$

(1) $p(1 \cdot t) = p(t)$.
(2) $p(0 \cdot t) = p(0) = *$ for every t.

Each $H_u(p)$ still starts at $*$, so it lies in PM. Hence PM has the homotopy type of a point.

Homotopy lifting. Suppose p is a path from $*$ to x_0 and the end point moves along a path $s \mapsto x_s$. Lift by *extending* the path: follow p, then follow the motion of the end point.

$$p_s(t) = \begin{cases} p\big((1+s)t\big), & 0 \le t \le \frac{1}{1+s}, \\ x_{(1+s)t-1}, & \frac{1}{1+s} \le t \le 1. \end{cases} \tag{1.140}$$

At the junction $t = \frac{1}{1+s}$ both lines give $p(1) = x_0$; at $t = 1$ the end point is x_s; at $s = 0$ the second line is empty and $p_0 = p$. Doing this for a whole family of paths at once (the $y \in F$ of the definition) is the same formula. So π is a fibration.

Key idea: Loop space shifts homotopy down by one

The exact sequence of $\Omega M \to PM \to M$, with $\pi_n(PM) = 0$ for all n, gives

$$\pi_n(M) \cong \pi_{n-1}(\Omega M). \tag{1.141}$$

This is the isomorphism $[S^n, M]_0 = [S^{n-1}, \Omega M]_0$ of Chapter 1 (Nash's (1.29)), now as a

consequence of a fibration. With $M = K(G, n+1)$ it says $\Omega K(G, n+1)$ has a single homotopy group G in degree n, which is Nash's (1.39). With $M = BG$ and (1.136) it says $\Omega BG \simeq G$.

Why Nash cares: computing the cohomology of ΩM

Loop spaces are infinite dimensional, and there is no direct way to triangulate them. But in the path fibration the total space is a point up to homotopy, so its cohomology is known: $\mathbb{Z}$ in degree 0 and nothing else. The spectral sequence must turn the grid $E_2^{p,q} = H^p(M) \otimes H^q(\Omega M)$ into that, and this is a very strong constraint.

Worked example 1.106: The cohomology of ΩS^n

Take $M = S^n$, $n \geq 2$. The grid has only two non-zero columns, $p = 0$ and $p = n$, each a copy of $H^q(\Omega S^n)$. A differential d_r moves r steps right and $r-1$ steps down, so the only one that can connect the two columns is $d_n : E^{0,q} \to E^{n,q-n+1}$. Everything except $E^{0,0} = \mathbb{Z}$ must be killed, so each d_n must be an isomorphism:

$$H^q(\Omega S^n) \overset{(1)}{=} E_2^{0,q} \overset{(2)}{=} E_2^{n,q-n+1} = H^{q-n+1}(\Omega S^n) \qquad (q \geq 1).$$

(1) The column $p = 0$ is $H^0(S^n) \otimes H^q(\Omega S^n) = H^q(\Omega S^n)$.

(2) d_n must be an isomorphism for everything to cancel, and the column $p = n$ is another copy of $H^*(\Omega S^n)$.

Starting from $H^0 = \mathbb{Z}$ and $H^q = 0$ for $0 < q < n-1$, the recursion gives

$$H^q(\Omega S^n; \mathbb{Z}) = \begin{cases} \mathbb{Z}, & q = 0,\ n-1,\ 2(n-1),\ 3(n-1), \dots \\ 0, & \text{otherwise.} \end{cases}$$

For $n = 2$, ΩS^2 has $\mathbb{Z}$ in every degree. The infinite-dimensional loop space has cohomology in infinitely many degrees, which is characteristic of loop spaces and of the loop groups of Chapter 6.

Reading Nash (pp. 14–19, (1.47)–(1.58))

Nash's §5 is a rapid tour; we have filled in each stop.

- (1.47) is local triviality, our (1.101). Nash does not write transition functions at this point; we made them the working definition because every computation runs through them.
- The pullback of cohomology, $f^* : H^n(M; G) \to H^n(N; G)$, and of bundles, f^*E, and (1.48) $f_1^*E \cong f_2^*E$ for $f_1 \simeq f_2$: our Sections 1.5.9 and 1.5.10, with the homotopy operator and the parallel-transport argument added.
- (1.49) bundles $\leftrightarrow [M, BG]$ and (1.50) $c_i \mapsto f^*c_i$: Sections 1.5.11 and 1.5.12. Nash names Chern, Pontryagin, Stiefel–Whitney and Euler classes without formulas; we added the Chern–Weil expressions and the explicit classes of our examples.
- (1.51)–(1.52): the fibre sequence and the exact homotopy sequence, with "$\pi_n(E) \cong \pi_n(M) + \pi_n(F)$" for a trivial bundle (read + as direct sum). Nash gives no examples

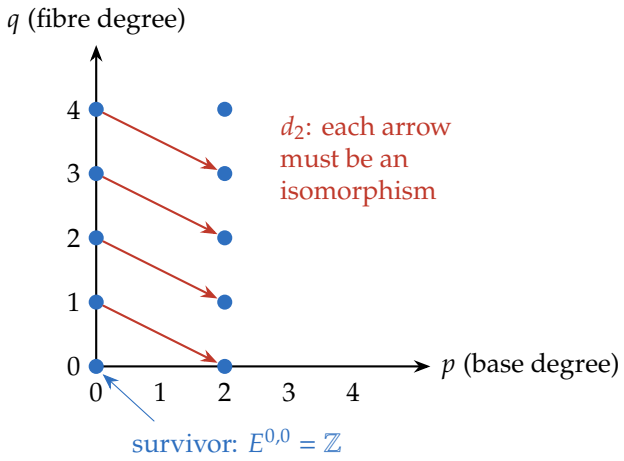


Figure 1.57: The spectral sequence of $\Omega S^2 \to PS^2 \to S^2$. Dots are the groups $H^p(S^2) \otimes H^q(\Omega S^2)$, in the columns $p = 0$ and $p = 2$. The differential d_2 (red) goes two steps right and one step down. The total space is contractible, so every dot except the bottom-left one must be cancelled by an arrow. This forces $H^q(\Omega S^2) \cong H^{q-1}(\Omega S^2)$, so $\mathbb{Z}$ appears in every degree.

of the sequence in use; we computed $\pi_n(S^1)$, $\pi_3(S^2)$, $\pi_n(\mathbb{CP}^\infty)$, $\pi_1(\mathrm{SO}(3))$ (two ways) and $\pi_3(\mathrm{SU}(n))$.

- (1.53)–(1.55): fibrations and the homotopy lifting property, Nash's diagrams being our (1.137). His remarks that fibres need only share a homotopy type and that the same exact sequence holds are Example 1.105 and the discussion after it.
- (1.56)–(1.58): the path fibration $\Omega M \to PM \to M$, with Nash's fixed-path argument that $\pi^{-1}(x) \simeq \Omega M$. We added the explicit contraction of PM, the lifting formula (1.140), and the computation of $H^*(\Omega S^n)$ that shows why "this encoding of ΩM as the fibre of a fibration" is a "crucial step".

Recap: §1.5 Fibre bundles and fibrations

- A bundle is local products $O_\alpha \times F$ glued by transition functions $g_{\alpha\beta} \in G$ obeying the cocycle condition; changing trivialisations is a gauge transformation $h_\alpha g_{\alpha\beta} h_\beta^{-1}$.
- Fields are sections. Twisting can force zeros: Möbius band, hairy ball (TS^2 has clutching winding 2).
- Principal bundles carry the transition functions; associated bundles $P \times_\rho V$ carry matter in representation ρ; gauge potentials glue as $A_\beta = g^{-1} A_\alpha g + g^{-1} \mathrm{d} g$.
- Bundles over S^n are classified by $\pi_{n-1}(G)$ (clutching). The monopole is $e^{in\phi}$; its charge is $c_1 = n$. The Hopf fibration $S^1 \to S^3 \to S^2$ is the $|n| = 1$ case, and it is the unit circle bundle of the tautological line bundle.
- Pullback f^* is contravariant and homotopy invariant. Bundles are pullbacks of a universal bundle, $\{\text{bundles}\} = [M, BG]$; characteristic classes are pullbacks of $H^*(BG)$. Line bundles are $H^2(M; \mathbb{Z})$.
- Exact sequence $\cdots \to \pi_n(F) \to \pi_n(E) \to \pi_n(M) \xrightarrow{\partial} \pi_{n-1}(F) \to \cdots$; ∂ is "lift a disc, take its edge", and over a sphere it is the clutching class. Hopf gives $\pi_3(S^2) = \mathbb{Z}$.

- Fibrations need only homotopy lifting. The path fibration $\Omega M \to PM \to M$ has contractible total space, so $\pi_n(M) = \pi_{n-1}(\Omega M)$, and it computes $H^*(\Omega S^n)$.

1.6 Differentiable structures for manifolds

In §1.1 we met three kinds of manifold: topological (M_{TOP}), piecewise-linear (M_{PL}) and smooth (M_{DIFF}). They differ only in what we demand of the transition functions $g_{\alpha\beta}$ that glue the coordinate patches: homeomorphisms, piecewise-linear homeomorphisms, or diffeomorphisms. Two questions follow at once.

- *Existence.* Given a topological manifold, can we choose the atlas so that all the $g_{\alpha\beta}$ are piecewise-linear ("straighten" it)? Given a PL manifold, can we make them smooth ("smooth" it)?
- *Uniqueness.* If we can, in how many essentially different ways?

Why should a physicist care? Every action we write, $\int \left(\frac{1}{2}\partial_\mu\phi\,\partial^\mu\phi - V(\phi)\right)$ or $-\int \mathrm{tr}(F \wedge \star F)$, needs derivatives, and derivatives need a smooth structure. If one topological space carried two inequivalent smooth structures, "the same" spacetime would carry two inequivalent sets of smooth fields and two different field theories. This really happens: on S^7 there are 28 smooth structures, and on $\mathbb{R}^4$, the most important space in physics, there are uncountably many. And the most spectacular result in the subject, Donaldson's theorem, was proved by solving the Yang–Mills equations. Physics was used to answer a question about smoothness.

The pleasant surprise is that most of the answer can be computed with the tools of the previous two sections. The plan is:

1. turn the three kinds of transition function into three groups $TOP_n \supset PL_n \supset DIFF_n$;
2. use their classifying spaces (Section 1.5.11) to say what a "topological tangent bundle" is;
3. recast "can M be straightened or smoothed?" as a *lifting problem* through a fibration (Section 1.5.15);
4. solve lifting problems cell by cell with *obstruction theory*, whose answers are cohomology classes (Section 1.4);
5. feed in the homotopy groups of the fibres TOP/PL and $PL/DIFF$, which are known.

The result works beautifully in dimension ≥ 5 and fails spectacularly in dimension 4. Figure 1.58 shows the whole section on one page.

1.6.1 Three groups of transition functions

A transition function $g_{\alpha\beta}$ is a map from an open subset of $\mathbb{R}^n$ to another open subset of $\mathbb{R}^n$. To turn these into a *group*, we look at invertible maps of $\mathbb{R}^n$ to itself, and (as for vector bundles) only care about what happens near one point, which we may put at the origin. So we define:

- TOP_n: homeomorphisms $f : \mathbb{R}^n \to \mathbb{R}^n$ with $f(0) = 0$;
- PL_n: piecewise-linear homeomorphisms with $f(0) = 0$ ("piecewise-linear": there is a triangulation of $\mathbb{R}^n$ on each simplex of which f is affine; the inverse is then automatically PL too);

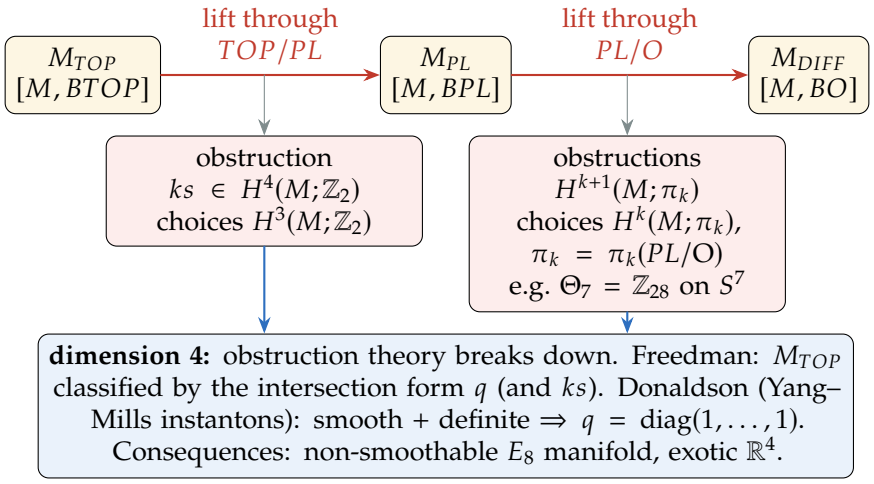


Figure 1.58: Road map of Section 1.6. Going from left to right adds structure. Each arrow is a lifting problem through a fibration, answered by obstruction theory: the red boxes list the cohomology group that contains the obstruction and the group that counts the answers. In dimension ≥ 5 that is the whole story. In dimension 4 (blue box) it is replaced by Freedman's and Donaldson's theorems.

- $DIFF_n$: diffeomorphisms with $f(0) = 0$.

Each is a group under composition, and each is a topological group (we can say when two homeomorphisms are close). The **structure group** of a manifold's tangent bundle is then TOP_n, PL_n or $DIFF_n$ according to its type.

Worked example 1.107: Elements of one group that are not in the next

Work in $n = 1$.

- $f(x) = x^3$ is a homeomorphism of $\mathbb{R}$ (continuous, bijective, with continuous inverse $x^{1/3}$). It is smooth, but its inverse is not differentiable at 0. So $f \in TOP_1$, $f \notin DIFF_1$. It is not PL either: no finite collection of straight pieces reproduces a cubic.
- $f(x) = x$ for $x \leq 0$ and $f(x) = 2x$ for $x \geq 0$ is PL (two linear pieces), a homeomorphism, but has a kink at 0. So $f \in PL_1$, $f \notin DIFF_1$.
- $f(x) = x + x^3$ is a diffeomorphism ($f' = 1 + 3x^2 > 0$), but it is not affine on any interval. So $f \in DIFF_1$, $f \notin PL_1$.

The last example shows that $DIFF_n \subset PL_n$ is *not* a literal inclusion of sets; see the pitfall below.

These groups are huge (infinite-dimensional), but for the smooth one there is an enormous simplification: $DIFF_n$ can be squashed down, without changing its homotopy type, onto a finite-dimensional Lie group.

Key idea: A diffeomorphism is homotopic to its derivative

Every $f \in DIFF_n$ is joined to the linear map $\mathrm{d}f_0 \in \mathrm{GL}(n,\mathbb{R})$ by the path

$$f_t(x) = \frac{f(tx)}{t} \quad (0 < t \leq 1), \qquad f_0(x) = \mathrm{d}f_0(x). \tag{1.142}$$

Consequently $DIFF_n \simeq \mathrm{GL}(n,\mathbb{R}) \simeq \mathrm{O}(n)$, and $BDIFF_n \simeq BO(n)$: a smooth tangent bundle is, up to homotopy, just an ordinary vector bundle.

Let us check that the path really ends at the derivative. Expand f around 0:

$$\begin{aligned} f_t(x) &\overset{(1)}{=} \frac{1}{t}\Big(f(0) + t\,\mathrm{d}f_0(x) + \tfrac{t^2}{2}\,\mathrm{d}^2 f_0(x,x) + O(t^3)\Big) \\ &\overset{(2)}{=} \mathrm{d}f_0(x) + \tfrac{t}{2}\,\mathrm{d}^2 f_0(x,x) + O(t^2) \\ &\xrightarrow{t\to 0} \mathrm{d}f_0(x). \end{aligned}$$

(1) Taylor's theorem for $f(tx)$ in powers of t; $\mathrm{d}f_0$ is the Jacobian matrix at the origin and $\mathrm{d}^2 f_0$ the Hessian.

(2) $f(0) = 0$ because elements of $DIFF_n$ fix the origin; divide through by t.

Each f_t is a diffeomorphism (a rescaled copy of f), so this is a path inside $DIFF_n$ that shrinks the whole group onto its linear part. The second step, $\mathrm{GL}(n,\mathbb{R}) \simeq \mathrm{O}(n)$, is Gram–Schmidt: write $A = O\,P$ with O orthogonal and $P = (A^T A)^{1/2}$ positive definite (the polar decomposition), and slide P to the identity along P^s, $s \in [0,1]$. Physicists know this from the vielbein: any frame can be deformed to an orthonormal one.

The trick fails for TOP_n and PL_n, because a homeomorphism has no derivative to shrink onto. That is the whole difference between the three theories: TOP_n and PL_n are genuinely bigger than $\mathrm{O}(n)$, and the "extra" homotopy is measured by the quotients TOP_n/PL_n and $PL_n/\mathrm{O}(n)$.

Watch out: "$DIFF \subset PL$" is an inclusion only up to homotopy

Nash writes the inclusions $DIFF_n \subset PL_n \subset TOP_n$ [Nash p. 19, (1.59)]. The second is literal. The first is not, by Example 1.107. What is true is J. H. C. Whitehead's theorem: every smooth manifold has a triangulation compatible with its smooth structure, unique up to PL homeomorphism. So each smooth structure determines a PL structure, and on the level of groups one gets a map $\mathrm{O}(n) \simeq DIFF_n \to PL_n$ defined up to homotopy. That is all the theory uses, and it is why topologists often write O instead of $DIFF$, as in PL/O.

Letting $n \to \infty$: direct limits

Each group sits inside the next one: a map f of $\mathbb{R}^n$ extends to $\mathbb{R}^{n+1} = \mathbb{R}^n \times \mathbb{R}$ by doing nothing to the last coordinate, $f \mapsto f \times \mathrm{id}_{\mathbb{R}}$. This gives Nash's three towers

$$\begin{array}{ccccccccc} TOP_1 & \subset & TOP_2 & \subset & \cdots & \subset & TOP_n & \subset & \cdots \\ PL_1 & \subset & PL_2 & \subset & \cdots & \subset & PL_n & \subset & \cdots \\ DIFF_1 & \subset & DIFF_2 & \subset & \cdots & \subset & DIFF_n & \subset & \cdots \end{array} \tag{1.143}$$

Working definition 1.42 (direct limit)

Given a sequence $O_1 \xrightarrow{\phi_1} O_2 \xrightarrow{\phi_2} O_3 \to \cdots$, the **direct limit** $\varinjlim O_i$ is "the union of all of them, with each O_i identified with its image in the next" [$\phi_i(O_i) \subset O_{i+1}$]. An element is an element of some O_i; two elements are equal if they become equal further along [$x \in O_i$ and $x' \in O_j$ are equal when their images in some later O_k coincide]. For groups and spaces the operations and the topology are inherited from the finite stages. *Operationally*: a statement about the limit is a statement about O_i for all sufficiently large i. For example $\pi_k(\varinjlim O_i) = \varinjlim \pi_k(O_i)$, which is $\pi_k(O_i)$ for i large once these groups stabilise [the maps $\pi_k(O_i) \to \pi_k(O_{i+1})$ are isomorphisms from some i on]. The dual notion, for maps going the other way $O_1 \leftarrow O_2 \leftarrow \cdots$, is the **inverse limit** $\varprojlim O_i$: sequences $(x_1, x_2, \dots)$ with

$x_i \in O_i$ compatible under the maps [the map $O_{i+1} \to O_i$ sends each x_{i+1} to x_i].

Worked example 1.108: Three direct limits

- $\mathbb{R}^\infty = \varinjlim \mathbb{R}^n$, with $\mathbb{R}^n \subset \mathbb{R}^{n+1}$ as $(x_1, \dots, x_n, 0)$: the space of real sequences with finitely many nonzero entries.
- $S^\infty = \varinjlim S^n$, the unit sphere in $\mathbb{R}^\infty$. It is contractible (every sphere S^n is contractible inside S^{n+1}), and $\mathbb{Z}_2 = \{\pm 1\}$ acts freely on it, so it is $E_{\mathbb{Z}_2}$ and $S^\infty/\mathbb{Z}_2 = \mathbb{RP}^\infty = B\mathbb{Z}_2$ (Section 1.5.11).
- $\mathrm{O} = \varinjlim \mathrm{O}(n)$, with $A \mapsto \left(\begin{smallmatrix} A & 0 \\ 0 & 1 \end{smallmatrix}\right)$: infinite orthogonal matrices that differ from the identity in a finite block. Its homotopy groups are the stable ones, $\pi_k(\mathrm{O}) = \pi_k(\mathrm{O}(n))$ for $n > k+1$; Bott periodicity (§1.3, Chapter 3) says they repeat with period 8: $\mathbb{Z}_2, \mathbb{Z}_2, 0, \mathbb{Z}, 0, 0, 0, \mathbb{Z}$ for $k = 0, 1, \dots, 7$.

The limits of (1.143) are the three groups

$$TOP = \varinjlim TOP_n, \qquad PL = \varinjlim PL_n, \qquad DIFF = \varinjlim DIFF_n \simeq \mathrm{O}. \tag{1.144}$$

Why go to the limit? Because the questions we ask are *stable*. Replacing M by $M \times \mathbb{R}^k$ adds a trivial bundle to its tangent bundle, $T(M \times \mathbb{R}^k) = TM \oplus \underline{\mathbb{R}}^k$, and does not change whether M can be smoothed (in the range of dimensions where the theory works). Stable questions are governed by the limit groups, whose homotopy groups are simpler and are known.

1.6.2 Classifying spaces and the three tangent bundles

Each of the groups TOP, PL, $DIFF$ has a classifying space, $BTOP$, BPL, $BDIFF \simeq BO$, built exactly as in Section 1.5.11: a contractible space with a free action of the group, divided by that action. By (1.126), a bundle with structure group G over M is the same thing as a homotopy class of maps $M \to BG$.

What bundle? The tangent bundle. For a smooth manifold this is the ordinary TM, glued by the Jacobians. For a PL or a topological manifold there are no Jacobians, but there is still a "tangent bundle" whose transition functions are the $g_{\alpha\beta}$ themselves, germs of homeomorphisms near a point.

Working definition 1.43 (topological tangent bundle)

For a topological manifold M, take a small neighbourhood U of the diagonal $\{(x,x)\}$ in $M \times M$ and project to the first factor, $(x,y) \mapsto x$. The fibre over x is a neighbourhood of x in M, a copy of $\mathbb{R}^n$ with a marked point x (the "origin"). Changing coordinates near x changes this $\mathbb{R}^n$ by an element of TOP_n [a homeomorphism of $\mathbb{R}^n$ fixing the origin]. This bundle with fibre $\mathbb{R}^n$ and structure group TOP_n is the **tangent microbundle** TM_{TOP} (Milnor). For a smooth M, the exponential map [a tangent vector v at x goes to the point reached by the geodesic from x with velocity v after unit time] identifies it with TM. *Operationally*: it is recorded by its classifying map $\tau_M : M \to BTOP_n$ [the map, unique up to homotopy, along which this bundle is pulled back from the universal one], stabilised to $M \to BTOP$.

Physical picture: The tangent space of a manifold without derivatives

Zoom in on a point of a smooth manifold and it looks like a vector space: that is T_xM. Zoom in on a topological manifold and you still see $\mathbb{R}^n$, but you have lost the linear structure, so "tangent vectors" can be added only up to homeomorphism. The microbundle keeps exactly the part of the tangent bundle that survives: the local copy of $\mathbb{R}^n$ and how these copies are glued.

So each type of manifold comes with a map, and the type of M is encoded in which space the map lands in:

$$TM_{TOP} \leftrightarrow [M, BTOP], \qquad TM_{PL} \leftrightarrow [M, BPL], \qquad TM_{DIFF} \leftrightarrow [M, BDIFF]. \tag{1.145}$$

A PL manifold is in particular topological, so its tangent map $M \to BPL$ can be composed with $BPL \to BTOP$ to give the topological tangent map. The question "does the topological manifold M admit a PL structure?" therefore has a *necessary condition* that is pure homotopy theory: $\tau_M : M \to BTOP$ must factor through BPL.

The fibre of $BH \to BG$ is G/H

Nash states that $BPL \to BTOP$ can be made into a fibration with fibre TOP/PL. This is an instance of a general and very useful fact. Let $H \subset G$ be a subgroup, and let E_G be the contractible space with free G action. Then H also acts freely on E_G, so E_G serves as E_H too, and

$$BH \overset{(1)}{=} E_G/H \xrightarrow{\ \pi\ } E_G/G \overset{(2)}{=} BG, \qquad \pi^{-1}\big([e]\big) \overset{(3)}{=} \{[eg]_H : g \in G\} \overset{(4)}{=} G/H. \tag{1.146}$$

(1) E_G is contractible with a free H action, which is all a universal H-bundle needs.
(2) The definition of BG.
(3) π sends an H-orbit to the G-orbit containing it. The H-orbits inside the G-orbit of a point e are the classes $[eg]_H$.
(4) eg and eg' lie in the same H-orbit iff $g' \in Hg$, since the action is free. So the fibre is the coset space (right or left cosets, depending on the side on which the group acts).

So for every inclusion of groups there is a fibration

$$G/H \longrightarrow BH \xrightarrow{\ \pi\ } BG, \tag{1.147}$$

and Nash's (1.148) below is the case $H = PL$, $G = TOP$.

Worked example 1.109: Fibres we already know

- $H = \mathrm{O}(n-1) \subset G = \mathrm{O}(n)$. The group $\mathrm{O}(n)$ moves the unit vector e_n to any unit vector, and the matrices fixing e_n form $\mathrm{O}(n-1)$. So $\mathrm{O}(n)/\mathrm{O}(n-1) = S^{n-1}$, and $S^{n-1} \to BO(n-1) \to BO(n)$. Lifting the tangent map of M from $BO(n)$ to $BO(n-1)$ means splitting $TM = L \oplus \xi$ with L a trivial line: it means finding a *nowhere-vanishing vector field*.
- $H = \mathrm{SO}(n) \subset G = \mathrm{O}(n)$. Here $\mathrm{O}(n)/\mathrm{SO}(n) = \{\pm 1\}$ via the determinant. Lifting from $BO(n)$ to $BSO(n)$ means choosing an *orientation*.
- $\mathrm{Spin}(n) \to \mathrm{SO}(n)$ is not an inclusion but a double cover with kernel $\mathbb{Z}_2$ (Chapter T4). The same argument with G/H replaced by B(kernel) gives the fibration $B\mathbb{Z}_2 \to$

$B\mathrm{Spin}(n) \to BSO(n)$, with fibre $B\mathbb{Z}_2 = \mathbb{RP}^\infty = K(\mathbb{Z}_2, 1)$. Lifting means choosing a *spin structure*.

Vector fields, orientations and spinors are all lifting problems of exactly the same type as straightening and smoothing. We shall solve these three first, because we can check the answers by hand.

1.6.3 Straightening and smoothing as lifting problems

Apply (1.147) to $PL \subset TOP$:

$$\begin{array}{ccc} TOP/PL & \longrightarrow & BPL \\ & & \downarrow{\scriptstyle \pi} \\ & & BTOP \end{array} \tag{1.148}$$

A topological manifold M gives its tangent map $\alpha = \tau_M : M \to BTOP$. If M carries a PL structure, its PL tangent bundle gives $\beta : M \to BPL$ with $\pi \circ \beta = \alpha$. Nash's proposal is to run this backwards: *try to factorise α through BPL; if you succeed, M acquires a PL structure.*

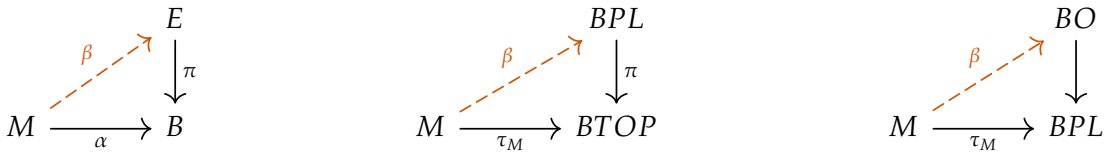


Figure 1.59: Three lifting problems. Left: the general problem (1.149) for a fibration $F \to E \to B$. Middle: "can the topological manifold M be straightened?" Right: "can the PL manifold M be smoothed?" In each case we are given the solid arrow and ask for the dashed (orange) one making the triangle commute, $\pi \circ \beta = \alpha$.

The general form of the question is [Nash p. 20, (1.61)]: given a fibration $F \to E \xrightarrow{\pi} B$ and a map $\alpha : M \to B$, find

$$\beta : M \to E \qquad \text{with} \qquad \alpha = \pi \circ \beta. \tag{1.149}$$

There is an equivalent way to say this that will make the solution transparent. Pull the fibration back along α (Section 1.5.10): $\alpha^*E = \{(m, e) \in M \times E : \alpha(m) = \pi(e)\}$ is a fibration over M with the same fibre F. A lift β is the same thing as a *section* $m \mapsto (m, \beta(m))$ of α^*E. So

lifting α through E = finding a section of a bundle over M with fibre F.

We know from the Möbius band and the hairy ball (Sections 1.5.3 and 1.5.8) that sections of twisted bundles can fail to exist. Obstruction theory measures exactly how.

Watch out: A lift is not yet a structure

Every PL structure gives a lift, but the converse "every lift comes from a PL structure" is a theorem, not a tautology. It is the content of the work of Kirby and Siebenmann, and it holds when $\dim M \geq 5$ ($\dim M \geq 6$ if M has a boundary) [Nash p. 20, footnote 5]. The dimension restriction is the same "room to move" we met in §1.1: the proof needs the Whitney trick, which separates two intersecting discs by sliding one over a third disc, and that disc can be pushed off everything else only when $\dim M \geq 5$. In dimension 4 the lifting problem is still well posed but its answer no longer decides the geometric question; see Section 1.6.7. The smoothing theory of Hirsch and Mazur plays the same role for

$PL \to DIFF$.

1.6.4 Obstruction theory: extending a section cell by cell

We want a section of a bundle $F \to \alpha^*E \to M$. Build M out of cells: points (0-cells), then segments glued along their end points (1-cells), then discs glued along their boundary circles (2-cells), and so on. This is a triangulation, or more economically a cell decomposition like those of Section 1.4.6. The union of the cells of dimension $\leq k$ is the **k-skeleton** $M^{(k)}$. We build the section one skeleton at a time.

- *0-cells.* Over each vertex choose any point of the fibre. Possible as long as $F \neq \varnothing$.
- *1-cells.* Over each edge, the section is already given at its two end points, two points of F. We need a path in F joining them. Possible for every edge if F is path-connected, $\pi_0(F) = 0$.
- *k-cells.* Over each cell e^k the bundle is trivial (e^k is contractible), so a section over e^k is a map $e^k \to F$. The section is already defined on the boundary $\partial e^k \cong S^{k-1}$. That boundary map $S^{k-1} \to F$ extends over the disc iff it is null-homotopic, that is, iff its class in $\pi_{k-1}(F)$ vanishes.

Working definition 1.44 (obstruction cocycle and obstruction class)

Suppose a section s has been built on the $(k-1)$-skeleton. For each k-cell e^k let

$$c(s)(e^k) = \left[\, s|_{\partial e^k} : S^{k-1} \to F \,\right] \in \pi_{k-1}(F). \tag{1.150}$$

This assigns an element of the group $\pi_{k-1}(F)$ to every k-cell: it is a k-cochain with coefficients in $\pi_{k-1}(F)$ (Section 1.4.7). It is a cocycle [$\delta c(s) = 0$: summed over the boundary of any $(k+1)$-cell, $c(s)$ gives 0], and changing s on the $(k-1)$-cells changes $c(s)$ by a coboundary [a cochain of the form δd]. Its class

$$\mathfrak{o}_k \in H^k\big(M; \pi_{k-1}(F)\big) \tag{1.151}$$

is the **obstruction class**. *Operationally*: s extends over the k-skeleton (after possibly changing it on the $(k-1)$-skeleton) iff $\mathfrak{o}_k = 0$. Writing $k = n+1$ gives Nash's form $H^{n+1}(M; \pi_n(F))$ [Nash p. 21, (1.62)].

Two facts make this practical.

- *Only the homotopy groups of F enter.* If $\pi_j(F) = 0$ for all $j < n$, there are no obstructions until $k = n+1$: the first possibly non-zero one is the **primary obstruction** in $H^{n+1}(M; \pi_n(F))$. If F is contractible there are no obstructions at all, which is why every bundle has a classifying map (Section 1.5.11).
- *Counting answers.* Two sections that agree on the $(n-1)$-skeleton differ on each n-cell by an element of $\pi_n(F)$ (glue the two discs along their common boundary to get an n-sphere in F). This **difference cochain** is an n-cochain with values in $\pi_n(F)$, and the two sections are homotopic iff its class in $H^n(M; \pi_n(F))$ vanishes. When F has a single homotopy group $\pi_n(F)$, that is when $F = K(\pi, n)$ (Section 1.4.1), homotopy classes of lifts, if there are any, are counted exactly by $H^n(M; \pi_n(F))$ [Nash p. 21, (1.66)].

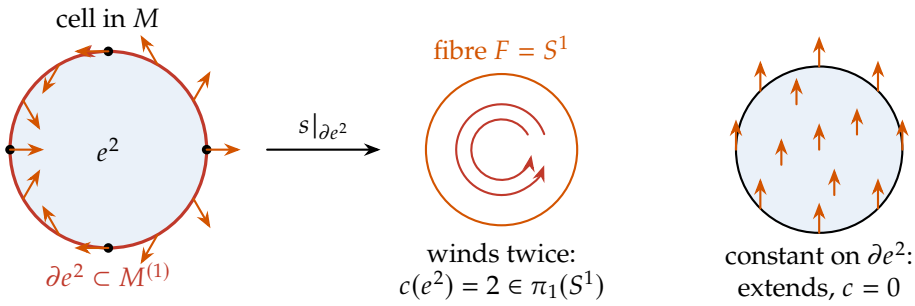


Figure 1.60: Obstruction theory for a section with fibre S^1 (a unit vector field on a surface). The section is known on the boundary of a 2-cell. Left: the arrows turn twice as we go once round the boundary, so the boundary map $S^1 \to F = S^1$ has degree 2, and no continuous choice of arrows fills the disc: the obstruction cochain takes the value 2 on this cell. Right: if the boundary arrows do not turn, we just continue them inwards, and the section extends.

Why is the obstruction cochain a cocycle? Look at the case $k = 2$ with fibre S^1 of the figure, and at a 3-cell e^3. Its boundary is a 2-sphere tiled by 2-cells (faces), and over the whole of e^3 the bundle is trivial, so the section on the edges of the tiling is just an angle ϑ attached to each point of each edge. The obstruction on a face is its winding number, $\frac{1}{2\pi}$ times the total change of ϑ as we run once round the edges of that face. Now add up over all faces of the 2-sphere. Each edge lies on exactly two faces, and is traversed in opposite directions by them, so its change of ϑ enters once with each sign:

$$\begin{aligned}(\delta c)(e^3) &\overset{(1)}{=} \sum_{\text{faces } f \subset \partial e^3} \pm c(f) \\ &\overset{(2)}{=} \frac{1}{2\pi} \sum_{f} \sum_{\text{edges } \ell \subset \partial f} \pm \Delta_\ell \vartheta \\ &\overset{(3)}{=} 0.\end{aligned}$$

(1) The coboundary of a cochain evaluates it on the boundary of the cell, with incidence signs (Section 1.4.7).
(2) The winding number of a face is the sum of the angle changes $\Delta_\ell \vartheta$ along its edges; the signs record the direction in which the face runs along each edge.
(3) Every edge appears twice, with opposite signs: this is $\partial\partial = 0$ for the tiled sphere.

So $\delta c = 0$. In higher degree the argument is the same with "angle changes" replaced by homotopy classes: it is the discrete Stokes theorem of Section 1.4.7, which is why obstructions are *cohomology* classes.

Worked examples of lifting problems

Worked example 1.110: Nowhere-vanishing vector fields on surfaces: the obstruction is χ

A nowhere-zero vector field on an oriented surface M is (after normalising it) a section of the unit tangent bundle $S^1 \to UTM \to M$. The fibre S^1 is connected, and $\pi_1(S^1) = \mathbb{Z}$. So the first obstruction lies in

$$H^2\big(M; \pi_1(S^1)\big) = H^2(M; \mathbb{Z}) \cong \mathbb{Z},$$

the last step for closed oriented M by Poincaré duality (1.95).
The sphere. Use the two patches of Section 1.5.8: $O_0 = S^2 \setminus \{N\}$ with coordinate z and

$O_1 = S^2 \setminus \{S\}$ with coordinate $w = 1/z$. Take the section $v = \partial_z$ on O_0. It is constant in the z-trivialisation, so it is a perfectly good nowhere-zero section over everything except the north pole. What is left is to extend it over one 2-cell, the northern cap $D_N = \{|w| \le 1\}$. In the w-trivialisation its component is, by (1.121),

$$v^w = \frac{\mathrm{d}w}{\mathrm{d}z} \cdot 1 = -\frac{1}{z^2} = -w^2.$$

The obstruction on D_N is the winding number of this map from the boundary circle $w = e^{i\psi}$ to the fibre (directions in the tangent plane):

$$\begin{aligned} c(D_N) &\overset{(1)}{=} \frac{1}{2\pi i} \oint_{|w|=1} \frac{\mathrm{d}v^w}{v^w} \\ &\overset{(2)}{=} \frac{1}{2\pi i} \int_0^{2\pi} 2i \, \mathrm{d}\psi \\ &\overset{(3)}{=} 2. \end{aligned}$$

(1) The degree of a map $S^1 \to S^1$, $\psi \mapsto u(\psi)$, is $\frac{1}{2\pi i} \oint \mathrm{d} \log u$ (§1.3). Only the direction of v matters, and d log ignores the modulus.
(2) $\mathrm{d}\log(-w^2) = 2\,\mathrm{d}w/w = 2i\,\mathrm{d}\psi$ on $w = e^{i\psi}$.
(3) The integrand is constant.

So $\mathfrak{v}_2 = 2 \neq 0$ and v cannot be extended without a zero: the hairy ball theorem again. Indeed the obvious extension $v^w = -w^2$ has a double zero, of index 2, at the north pole, exactly as in Figure 1.60 (left). The number is $2 = \chi(S^2)$.
The torus. $T^2 = \mathbb{R}^2/\mathbb{Z}^2$ has the global field ∂_{θ_1}, so the obstruction must vanish, and $\chi(T^2) = 0$.
In general the obstruction class, evaluated on $[M]$, is the sum of the indices of the zeros of any vector field (each zero is a 2-cell over which the field failed to extend), and by the Poincaré–Hopf theorem that sum is $\chi(M) = 2 - 2g$. The obstruction class is the **Euler class** of TM (Section 1.5.12). We will derive Poincaré–Hopf from Morse theory in Chapter 7.

Worked example 1.111: Orientation and spin structures

Let M be a Riemannian manifold with tangent map $\tau : M \to BO(n)$.
Orientation. The fibration $\mathrm{O}(n)/\mathrm{SO}(n) = \mathbb{Z}_2 \to BSO(n) \to BO(n)$ has discrete fibre, $\pi_0(\mathbb{Z}_2) = \mathbb{Z}_2$ (two points, thought of as the group $\mathbb{Z}_2$) and no higher homotopy. The only obstruction is to extending over 1-cells:

$$w_1(M) \in H^1(M; \mathbb{Z}_2).$$

On a closed loop γ, $w_1[\gamma]$ records whether an orthonormal frame carried round γ comes back with the same or the opposite orientation, i.e. the sign of det of the holonomy of the frame bundle. For the Möbius band and for $\mathbb{RP}^2$ it is 1 on the core loop, so they are non-orientable. If $w_1 = 0$ the choices are counted by $H^0(M; \mathbb{Z}_2) = \mathbb{Z}_2$ for connected M: the two orientations.
Spin structure. Now let M be oriented, and use $B\mathbb{Z}_2 \to B\mathrm{Spin}(n) \to BSO(n)$. The fibre is $K(\mathbb{Z}_2, 1)$, whose only homotopy group is $\pi_1 = \mathbb{Z}_2$. So the obstruction is the second Stiefel–Whitney class and the choices form a group:

$$\text{obstruction } w_2(M) \in H^2(M; \mathbb{Z}_2), \qquad \text{spin structures} \;\leftrightarrow\; H^1(M; \mathbb{Z}_2) \;\; (\text{if } w_2 = 0). \quad (1.152)$$

For the torus T^2 (which has a global frame, so $w_2 = 0$),

$$\#\{\text{spin structures on } T^2\} \overset{(1)}{=} \left|H^1(T^2;\mathbb{Z}_2)\right| \overset{(2)}{=} |\mathbb{Z}_2 \oplus \mathbb{Z}_2| = 4.$$

(1) (1.152); the fibre is a $K(\mathbb{Z}_2,1)$, so lifts are counted by H^1.
(2) $H^1(T^2;\mathbb{Z}) = \mathbb{Z}^2$ and there is no torsion, so reducing mod 2 gives $\mathbb{Z}_2^2$ (universal coefficients, Section 1.4.7).

These four are the four choices of periodic or antiperiodic boundary conditions for a fermion around the two cycles of the torus: $(P,P),(P,A),(A,P),(A,A)$, the Ramond and Neveu–Schwarz sectors of string theory (Chapter 9). A class in $H^1(M;\mathbb{Z}_2)$ is precisely a choice of sign ± 1 for each loop: the sign a spinor picks up when transported around it.

structure	fibration	fibre	obstruction	choices
vector field $\neq 0$	$BO(n-1) \to BO(n)$	S^{n-1}	$e(M) \in H^n(M;\mathbb{Z})$	(harder)
orientation	$BSO(n) \to BO(n)$	$\mathbb{Z}_2 = K(\mathbb{Z}_2,0)$	$w_1 \in H^1(M;\mathbb{Z}_2)$	$H^0(M;\mathbb{Z}_2)$
spin structure	$B\mathrm{Spin}(n) \to BSO(n)$	$K(\mathbb{Z}_2,1)$	$w_2 \in H^2(M;\mathbb{Z}_2)$	$H^1(M;\mathbb{Z}_2)$
PL structure	$BPL \to BTOP$	$K(\mathbb{Z}_2,3)$	$ks \in H^4(M;\mathbb{Z}_2)$	$H^3(M;\mathbb{Z}_2)$
smooth structure	$BO \to BPL$	PL/O	$H^{k+1}(M;\pi_k)$	$H^k(M;\pi_k)$, $\pi_k = \pi_k(PL/\mathrm{O})$

The table is the point of this subsection: *straightening and smoothing are the same kind of problem as choosing an orientation or a spin structure.* The only new input is the homotopy of the fibres TOP/PL and PL/O, to which we now turn.

Watch out: Local coefficients

In $H^{n+1}(M;\pi_n(F))$ the coefficient group is really a family of groups $\pi_n(F_b)$, one for each fibre F_b over $b \in B$. They are all isomorphic, since all fibres have the same homotopy type, but not canonically: carrying F_b around a loop in B can act on $\pi_n(F_b)$ by a non-trivial automorphism. When π_1 acts non-trivially, one must use **cohomology with local coefficients**, a twisted version in which the coefficients themselves form a bundle of groups [Nash p. 21, footnote 6]. Example: the Euler class of a non-orientable surface lives in H^2 with $\mathbb{Z}$ twisted by the orientation character, because going round an orientation-reversing loop sends the winding number $w \in \pi_1(S^1)$ to $-w$. For TOP/PL the coefficients are $\mathbb{Z}_2$, which has no non-trivial automorphism, so no twisting can occur, and Nash may safely write $\pi_n(TOP/PL)$.

1.6.5 From topological to PL: the Kirby–Siebenmann class

For the lifting (1.148) the obstructions lie in [Nash p. 21, (1.63)]

$$H^{n+1}\big(M;\pi_n(TOP/PL)\big), \tag{1.153}$$

and Kirby and Siebenmann computed the coefficients [Nash p. 21, (1.64)]:

$$\pi_n(TOP/PL) = \begin{cases} \mathbb{Z}_2 & n = 3, \\ 0 & \text{otherwise.} \end{cases} \tag{1.154}$$

A space with exactly one non-zero homotopy group is an Eilenberg–MacLane space (Section 1.4.1), so $TOP/PL \simeq K(\mathbb{Z}_2, 3)$. Everything now follows from the table above.

Key idea: Straightening a topological manifold, $\dim M \geq 5$

- There is exactly one obstruction, the **Kirby–Siebenmann class**

$$ks(M) \in H^4(M; \mathbb{Z}_2). \tag{1.155}$$

  M admits a PL structure iff $ks(M) = 0$.

- If $ks(M) = 0$, the PL structures (up to concordance, i.e. up to deformation) are in one-to-one correspondence with

$$H^3(M; \mathbb{Z}_2). \tag{1.156}$$

Both lines come from (1.151) and the difference-cochain count, with the single non-zero coefficient group:

$$\text{obstructions} \overset{(1)}{=} \bigoplus_n H^{n+1}\big(M; \pi_n(TOP/PL)\big) \overset{(2)}{=} H^{3+1}(M; \mathbb{Z}_2),$$

$$\text{choices} \overset{(3)}{=} H^n\big(M; \pi_n(TOP/PL)\big)\big|_{n=3} = H^3(M; \mathbb{Z}_2).$$

(1) The possible obstructions, one for each homotopy group of the fibre.
(2) By (1.154) only $n = 3$ survives.
(3) The fibre is $K(\mathbb{Z}_2, 3)$, so lifts are counted by H^3 with coefficients in its one homotopy group.

Nash calls the class $e(M)$. We write $ks(M)$ to avoid a clash with the Euler class. As Nash notes [Nash p. 21, footnote 7], it is a characteristic class: there is a *universal* class $d \in H^4(BTOP; \mathbb{Z}_2)$, the obstruction for the universal bundle, and $ks(M) = \tau_M^* d$ is its pullback by the tangent map. So ks is to topological manifolds what w_1 is to orientability and w_2 to spin.

Worked example 1.112: How many PL structures?

Spheres. For $n \geq 5$, $H^4(S^n; \mathbb{Z}_2) = 0$ and $H^3(S^n; \mathbb{Z}_2) = 0$ by (1.91) (with $\mathbb{Z}_2$ in place of $\mathbb{Z}$; the same Mayer–Vietoris argument works). So S^n has a PL structure and it is unique.
Tori. For the torus T^n, $n \geq 5$, the obstruction vanishes because T^n is smooth. The count is

$$\#\{\text{PL structures on } T^n\} \overset{(1)}{=} \left|H^3(T^n; \mathbb{Z}_2)\right| \overset{(2)}{=} \left|\mathbb{Z}_2^{\binom{n}{3}}\right| = 2^{\binom{n}{3}},$$

(1) (1.156).
(2) The Künneth theorem (Section 1.4.10): $H^*(T^n)$ is the exterior algebra on n generators of degree 1, so H^3 has a basis $\mathrm{d}\theta_i \wedge \mathrm{d}\theta_j \wedge \mathrm{d}\theta_k$, $i < j < k$, one for each of the $\binom{n}{3}$ triples. There is no torsion, so the same holds mod 2.

For T^5 that is $2^{10} = 1024$ PL structures up to concordance. Some of them give PL manifolds that are homeomorphic to T^5 but not PL-homeomorphic to it: the "fake tori" of Hsiang–Shaneson and Wall, which played a central role in Kirby's proof.

Physical picture: The Hauptvermutung

The *Hauptvermutung* ("main conjecture", Steinitz and Tietze, 1908) said that two triangulations of the same space always have a common subdivision: triangulations are unique up to refinement, so "PL" adds nothing to "TOP". For polyhedra Milnor disproved it in 1961. For manifolds, (1.156) disproves it whenever $H^3(M;\mathbb{Z}_2)$ produces genuinely different PL manifolds, as for the fake tori. Nash's two consequences [Nash p. 22]:

- there are PL manifolds that are homeomorphic but admit no PL homeomorphism between them;
- if $ks(M) \neq 0$, the topological manifold M has no PL structure at all, i.e. it cannot be triangulated by a triangulation that is a PL manifold.

In low dimension the question is empty: Radó (1925) proved that every surface has a unique PL structure, and Moise (1952) did the same for 3-manifolds. So $TOP = PL$ for $\dim M \leq 3$. Dimension 4 is the interesting case; Nash postpones it and so do we, to Section 1.6.7.

1.6.6 From PL to smooth, and the exotic 7-spheres

Smoothing a PL manifold M with $\dim M \geq 5$ is the same kind of problem, with the fibration [Nash p. 22, (1.68)]

$$\begin{array}{ccc} PL/DIFF & \longrightarrow & BDIFF \\ & & \downarrow{\scriptstyle \pi} \\ & & BPL \end{array} \tag{1.157}$$

The difference is in the fibre. Its homotopy groups are much richer [Nash p. 22, (1.69)]. We complete Nash's table:

n	≤ 6	7	8	9	10	11	12	13	14	15	16
$\pi_n(PL/\mathrm{O})$	0	$\mathbb{Z}_{28}$	$\mathbb{Z}_2$	$\mathbb{Z}_2^3$	$\mathbb{Z}_6$	$\mathbb{Z}_{992}$	0	$\mathbb{Z}_3$	$\mathbb{Z}_2$	$\mathbb{Z}_{8128} \oplus \mathbb{Z}_2$	$\mathbb{Z}_2$

(1.158)

For $n \geq 7$ (in fact for $n \geq 5$) these groups are the groups Θ_n of Kervaire and Milnor: the set of smooth structures on the topological n-sphere, up to orientation-preserving diffeomorphism, made into an abelian group by connected sum.

Key idea: Smoothing, $\dim M \geq 5$

Obstructions to smoothing a PL manifold lie in $H^{k+1}(M;\pi_k(PL/\mathrm{O}))$, and when they vanish the smooth structures (up to concordance) are counted by the groups $H^k(M;\pi_k(PL/\mathrm{O}))$, in the sense of obstruction theory. Since $\pi_k(PL/\mathrm{O}) = 0$ for $k \leq 6$, *every PL manifold of dimension* ≤ 7 *is smoothable, and in dimension* ≤ 6 *the smoothing is unique.*

Worked example 1.113: The 28 smooth structures on S^7

Take $M = S^7$ with its PL structure. The first non-zero coefficient group is $\pi_7(PL/\mathrm{O}) = \mathbb{Z}_{28}$.

$$\begin{aligned} \text{obstruction} &\overset{(1)}{=} H^8\big(S^7;\pi_7(PL/\mathrm{O})\big) \overset{(2)}{=} 0, \\ \text{smoothings} &\overset{(3)}{=} H^7\big(S^7;\pi_7(PL/\mathrm{O})\big) \overset{(4)}{=} \mathbb{Z}_{28}. \end{aligned} \tag{1.159}$$

(1) The first possibly non-zero obstruction: $k = 7$ in $H^{k+1}(M;\pi_k)$ [Nash (1.70)].

(2) A 7-dimensional space has no cohomology in degree 8: there are no 8-cells.
(3) The first possibly non-zero choice group [Nash (1.71)]. The groups $H^k(S^7;\pi_k)$ for other k vanish, either because $\pi_k = 0$ ($k \le 6$) or because $H^k(S^7) = 0$ ($k \ge 8$).
(4) $H^7(S^7;G) = G$ for any coefficient group G, by (1.91).

So S^7 can be smoothed (we knew that: the round sphere is smooth), and there are 28 smoothings. One is the standard one; the other 27 are **exotic spheres**: smooth manifolds homeomorphic, but not diffeomorphic, to S^7 (Milnor 1956, Kervaire–Milnor 1963).

Watch out: 28 or 15?

The 28 counts oriented smooth structures up to *orientation-preserving* diffeomorphism. Reversing orientation sends the class $x \in \Theta_7 = \mathbb{Z}_{28}$ to $-x$. If we allow orientation-reversing diffeomorphisms too, x and $-x$ are identified, leaving 15 classes (0, 14, and thirteen pairs $\pm x$). Nash's "27 exotic spheres" is the oriented count.

Obstruction theory tells us there are 28 smoothings, but it does not show us one. Milnor's original construction is explicit, and it is a beautiful application of everything in Section 1.5: clutching, characteristic classes and the signature. We give it in full, because it is the prototype of every exotic structure.

Worked example 1.114: Milnor's exotic spheres as S^3 bundles over S^4

The bundles. By clutching (1.111), vector bundles of rank 4 over S^4 are classified by $\pi_3(\mathrm{SO}(4)) \cong \mathbb{Z} \oplus \mathbb{Z}$. Explicit representatives: identify $\mathbb{R}^4$ with the quaternions $\mathbb{H}$ and S^3 with the unit quaternions, and for integers h, j let

$$g_{h,j} : S^3 \to \mathrm{SO}(4), \qquad g_{h,j}(u)\,v = u^h\, v\, u^j \qquad (u \in S^3,\ v \in \mathbb{H}). \tag{1.160}$$

(Left and right multiplication by unit quaternions preserve $|v|$, so $g_{h,j}(u) \in \mathrm{SO}(4)$.) Let $\xi_{h,j}$ be the rank-4 bundle with this clutching function, $W_{h,j}$ its unit disc bundle (an 8-manifold with boundary) and $M_{h,j} = \partial W_{h,j}$ its unit sphere bundle, a 7-manifold fibred as $S^3 \to M_{h,j} \to S^4$. The two characteristic numbers of $\xi_{h,j}$ are (Milnor)

$$e(\xi_{h,j})[S^4] = h + j, \qquad p_1(\xi_{h,j})[S^4] = \pm 2(h - j). \tag{1.161}$$

For $h = 1, j = 0$ this is the quaternionic Hopf bundle, whose sphere bundle is $S^7 \to S^4$ (the quaternionic analogue of $S^3 \to S^2$, Section 1.5.7).

They are topological spheres. Take $h + j = 1$ and put $k = h - j$ (an odd integer). Milnor wrote down a smooth function on $M_{h,j}$ with exactly two critical points, a maximum and a minimum. A closed manifold with such a function is homeomorphic to a sphere (Reeb's theorem: it is two discs glued along their boundary; Chapter 7). So each $M_k = M_{h,j}$ is homeomorphic to S^7.

Some are not diffeomorphic to S^7. For a closed smooth oriented 8-manifold X, the Hirzebruch signature theorem (Chapter 4) gives

$$\sigma(X) = \frac{7p_2 - p_1^2}{45}[X], \qquad \text{so} \qquad 45\,\sigma(X) + p_1^2[X] = 7\,p_2[X] \equiv 0 \pmod 7.$$

Since $45 \equiv 3$, multiplying by 2 gives $6\sigma + 2p_1^2 \equiv -\sigma + 2p_1^2 \equiv 0 \pmod 7$. Now suppose M_k were diffeomorphic to S^7. Then we could glue an 8-disc onto $W = W_{h,j}$ along M_k and get a closed smooth 8-manifold X with the same H^4, the same p_1 and the same signature as W.

Compute:

$$\begin{aligned} 0 &\overset{(1)}{=} 2p_1^2[X] - \sigma(X) \pmod 7 \\ &\overset{(2)}{=} 2(2k)^2 - 1 \\ &\overset{(3)}{=} 8k^2 - 1 \;\equiv\; k^2 - 1 \pmod 7. \end{aligned}$$

(1) The mod-7 consequence of the signature theorem derived above, applied to the closed manifold X.

(2) $H^4(X) = H^4(W) = H^4(S^4) = \mathbb{Z}$ (the disc bundle retracts onto its zero section, and the 8-disc adds nothing in degree 4), generated by a class x. The tangent bundle of W is $\pi^*(TS^4 \oplus \xi)$ and TS^4 is stably trivial, so $p_1(X) = p_1(\xi) = \pm 2k\,x$ by (1.161). The self-intersection of the zero section $S^4 \subset W$ is the Euler number $h + j = 1$, so $x^2[X] = 1$: the intersection form is the 1×1 matrix (1). Hence $p_1^2[X] = 4k^2$ and $\sigma(X) = 1$.

(3) $8 \equiv 1 \pmod 7$.

So if $k^2 \not\equiv 1 \pmod 7$, M_k is *not* diffeomorphic to S^7. For $k = 1$ ($h = 1, j = 0$) we get 0: consistent, it is the Hopf S^7. For $k = 3$ ($h = 2, j = -1$) we get $k^2 - 1 = 8 \equiv 1 \neq 0$: M_3 *is an exotic 7-sphere.* Milnor's invariant $\lambda(M_k) = k^2 - 1 \bmod 7$ takes the values $0, 1, 3, 6$ for $k = 1, 3, 5, \ldots$ (as $k^2 \bmod 7 \in \{1, 2, 4, 0\}$), so it already detects several distinct exotic spheres.

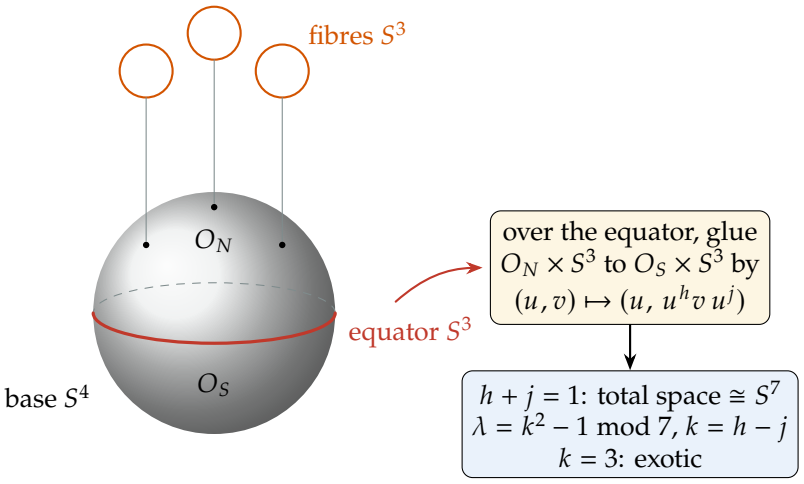


Figure 1.61: Milnor's construction. A 7-manifold is built as a bundle of 3-spheres over S^4, glued over the equatorial S^3 by the quaternionic clutching function (1.160). When $h + j = 1$ the result is always a topological 7-sphere, but the mod-7 invariant computed from p_1 and the signature shows that for $h - j = \pm 3$ it cannot be the standard smooth sphere.

Physical picture: Where are the exotic spheres "hidden"?

Every exotic 7-sphere is obtained from two standard 7-discs by gluing their boundary 6-spheres with a diffeomorphism $\phi : S^6 \to S^6$ that is not isotopic to the identity (Smale). Topologically every such ϕ can be straightened (the Alexander trick: cone it off, $x \mapsto |x|\ \phi(x/|x|)$, which is a homeomorphism of the disc but not smooth at the centre). Smoothly it cannot. So an exotic sphere is a "twisted double" whose twist is invisible to continuity and visible to derivatives: the smooth analogue of a non-trivial transition function. Another

model, due to Brieskorn, sits inside $\mathbb{C}^5$:

$$\Sigma_k = \{z \in \mathbb{C}^5 : \ z_1^{6k-1} + z_2^3 + z_3^2 + z_4^2 + z_5^2 = 0, \ |z| = 1\}, \qquad k = 1, \dots, 28,$$

and Σ_k represents k times a generator of $\Theta_7 = \mathbb{Z}_{28}$, so these 28 algebraic links run through all the smooth structures on S^7, with Σ_{28} standard.

n	1	2	3	4	5	6	7	8	9	10	11
smooth structures on S^n	1	1	1	?	1	1	28	2	8	6	992
smooth structures on $\mathbb{R}^n$	1	1	1	$2^{\aleph_0}$	1	1	1	1	1	1	1

The table (oriented counts) shows how odd dimension 4 is. For S^4 nobody knows whether there is an exotic structure: that is the *smooth 4-dimensional Poincaré conjecture*, still open. For $\mathbb{R}^4$ there are uncountably many (Section 1.6.12). In every other dimension $\mathbb{R}^n$ has exactly one.

Finally, PL and DIFF agree in low dimension: by the work of Kirby–Siebenmann and Kuiper cited by Nash [Nash p. 23], and in fact because $\pi_k(PL/\mathrm{O}) = 0$ for all $k \leq 6$, a PL manifold of dimension ≤ 6 has a unique smooth structure. (Nash states the bijection for dim $M \leq 3$; it holds in dimension 4 too, by Cerf's theorem, which is why Nash says a PL 4-manifold has a unique smooth structure.) So in dimension 4 the only gap is between TOP and PL=DIFF, and that is where we now go.

Reading Nash (pp. 19–23, (1.59)–(1.71))

Nash introduces TOP_n, PL_n, $DIFF_n$ and their direct limits in half a page (1.59), and the classifying spaces and $[M, BTOP]$, $[M, BPL]$, $[M, BDIFF]$ in a paragraph. We added: the linearisation (1.142) showing $DIFF_n \simeq \mathrm{O}(n)$; the tangent microbundle; the general fibre G/H of $BH \to BG$ (1.146), of which his (1.60) and (1.68) are instances; and the cell-by-cell construction behind the obstruction groups (1.62). His "one can check that the fibre is TOP/PL" is (1.146). His $e(M)$ is our $ks(M)$. His (1.66)–(1.67) ("the answer is just the whole cohomology group") is the difference-cochain count, valid because TOP/PL is a $K(\mathbb{Z}_2, 3)$. His (1.69) is our (1.158); his (1.70)–(1.71) are (1.159). The explicit Milnor spheres, the worked examples of vector fields, orientations and spin structures, and the fake tori are additions.

1.6.7 Why dimension four is different

In dimension 4, Nash says [Nash p. 23], "there are phenomena which have no counterpart in any other dimension":

- there are topological 4-manifolds with no smooth structure (although a PL 4-manifold has a unique smooth structure, so the whole difficulty is between TOP and PL=DIFF);
- the obstruction to smoothing is not of the type met in Section 1.6.4: it is not the pullback of a class from a classifying space, i.e. not a characteristic class. The Kirby–Siebenmann class $ks(M) \in H^4(M; \mathbb{Z}_2)$ is still defined, and $ks \neq 0$ still forbids smoothing, but $ks = 0$ is no longer enough.

Why does the machine break exactly here? The homotopy theory of the lifting problem is the same in every dimension. What fails is the step "a lift gives a structure", which rests on the geometric **Whitney trick**.

Physical picture: The Whitney trick needs five dimensions

In an n-manifold, a p-dimensional and a q-dimensional submanifold in general position meet in a set of dimension $p + q - n$ (count equations: each surface is cut out by $n - p$ and $n - q$ equations). When $p + q = n$ they meet in isolated points, each with a sign $\pm$ from orientations. The algebraic count is a homological invariant, the intersection number. Suppose two intersection points have opposite signs. Join them by an arc in each submanifold; the two arcs form a loop, and if the manifold is simply connected the loop bounds a disc D, the **Whitney disc**. Pushing one submanifold across D removes both intersection points (Figure 1.62), so the geometric count becomes the algebraic count.
The catch is that D must be embedded and must not meet the submanifolds in its interior. A 2-disc meets a p-dimensional submanifold in dimension $2 + p - n$, and meets itself in dimension $2 + 2 - n$. For $n \geq 5$ (with $p, q \geq 3$, or after the standard manipulations) these are negative, so generic position gives a clean disc. For $n = 4$ and $p = q = 2$ they are 0: *the Whitney disc itself generically has intersection points*, and the trick needs another Whitney disc to remove those, and so on forever. Casson made that infinite tower of discs precise (Casson handles), and Freedman proved, by an extraordinary piece of point-set topology, that the tower always contains a *topological* disc. It does not contain a smooth one in general. That is the gap between TOP and DIFF in dimension 4.

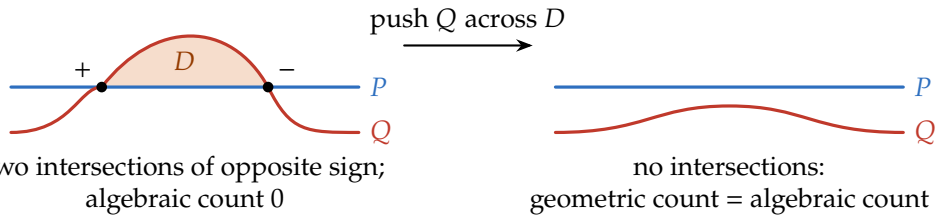


Figure 1.62: The Whitney trick, drawn one dimension down. Two submanifolds P and Q of complementary dimension cross twice with opposite signs. The arcs between the crossings bound the Whitney disc D (orange). Pushing Q across D removes both crossings. In dimension $n \geq 5$ the disc can be chosen embedded and disjoint from everything else; in dimension 4 it generically hits P, Q and itself in points, and the trick fails smoothly.

So in dimension 4 we need a new invariant, one that captures "how 2-dimensional surfaces intersect" directly. That is the intersection form.

1.6.8 The intersection form of a 4-manifold

Let M be a closed, oriented, simply connected 4-manifold. Its interesting cohomology is $H^2(M;\mathbb{Z})$: we have $H^0 = H^4 = \mathbb{Z}$, and $H^1 = H^3 = 0$ because $\pi_1 = 0$ (Hurewicz and Poincaré duality, Section 1.4.10). Two classes of degree 2 multiply to a class of degree 4, which we can evaluate on the fundamental class. Nash's definition [Nash p. 23, (1.72)]:

$$q(\alpha, \beta) = (\alpha \cup \beta)[M] \in \mathbb{Z}, \qquad \alpha, \beta \in H^2(M;\mathbb{Z}). \tag{1.162}$$

Working definition 1.45 (intersection form)

Three equivalent ways to compute q:

- *Forms.* Represent α, β by closed 2-forms with integer periods [$\int_\Sigma \alpha \in \mathbb{Z}$ for every closed surface Σ in M]; then $q(\alpha, \beta) = \int_M \alpha \wedge \beta$.
- *Surfaces.* By Poincaré duality (1.95) each class is dual to an oriented surface Σ_α.

Make $\Sigma_\alpha, \Sigma_\beta$ transverse [meeting at isolated points, where their two tangent planes together span all four directions]; then $q(\alpha,\beta) = \#(\Sigma_\alpha \cap \Sigma_\beta)$, counted with signs. For $q(\alpha,\alpha)$ push Σ_α off itself along a normal vector field and count the zeros of that field: the **self-intersection** number.

- *Matrix.* Choose a basis $e_1, \dots, e_n$ of $H^2(M;\mathbb{Z}) \cong \mathbb{Z}^n$ and write $Q_{ij} = q(e_i, e_j)$. A change of basis $P \in \mathrm{GL}(n,\mathbb{Z})$ [an integer matrix whose inverse is also an integer matrix, so $\det P = \pm 1$] gives $Q \mapsto P^T Q P$.

Its basic properties, each with a one-line reason:

- *Symmetric*: $\alpha \wedge \beta = (-1)^{2\cdot 2}\beta \wedge \alpha = \beta \wedge \alpha$ for 2-forms.
- *Unimodular*: $\det Q = \pm 1$. Poincaré duality over $\mathbb{Z}$ says every homomorphism $H^2(M;\mathbb{Z}) \to \mathbb{Z}$ is $q(\alpha,\cdot)$ for exactly one α, i.e. Q is invertible over the integers. (Simple connectivity ensures $H^2(M;\mathbb{Z})$ is free, with no torsion to spoil this.) Note $\det(P^T Q P) = \det Q$ since $\det P = \pm 1$.
- *Invariants*: the **rank** $n = b_2$; the **signature** $\sigma(q) = b_+ - b_-$, the number of positive minus the number of negative eigenvalues of Q [Nash p. 24, footnote 8]; the **type**, *even* if $q(\alpha,\alpha)$ is even for all α (equivalently all diagonal entries Q_{ii} are even) and *odd* otherwise; and *definite* ($b_+ = 0$ or $b_- = 0$) or *indefinite*.

Why the diagonal entries decide the type: for $x = \sum x_i e_i$,

$$q(x,x) \overset{(1)}{=} \sum_i Q_{ii}x_i^2 + \sum_{i\neq j} Q_{ij}x_i x_j$$
$$\overset{(2)}{=} \sum_i Q_{ii}x_i^2 + 2\sum_{i<j} Q_{ij}x_i x_j \;\equiv\; \sum_i Q_{ii}x_i^2 \pmod 2.$$

(1) Bilinearity.
(2) Symmetry $Q_{ij} = Q_{ji}$ pairs the off-diagonal terms, which contribute an even number.

Worked example 1.115: The basic intersection forms

- S^4: $H^2 = 0$, so q is the empty form, $q = \varnothing$. Rank 0, even.
- $\mathbb{CP}^2$: $H^2 = \mathbb{Z}c$, with c dual to a line $\mathbb{CP}^1$; two lines meet in one point, so $Q = (1)$ (Section 1.5.12). Odd, positive definite, $\sigma = 1$. Reversing the orientation, $\overline{\mathbb{CP}^2}$ has $Q = (-1)$.
- $S^2 \times S^2$: in the basis dual to $S^2 \times \{pt\}$ and $\{pt\} \times S^2$, $Q = H := \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ (Section 1.4.10). Each sphere can be pushed off itself (to a parallel copy), so its self-intersection is 0; the two spheres meet once. Even, indefinite, and since the eigenvalues of H are ± 1, $\sigma = 0$.
- Connected sums: removing a ball from each of M_1, M_2 and gluing along the boundary spheres gives $M_1 \# M_2$ with $H^2(M_1\#M_2) = H^2(M_1) \oplus H^2(M_2)$ and $q = q_1 \oplus q_2$. For example $\mathbb{CP}^2 \# \overline{\mathbb{CP}^2}$ has $Q = \mathrm{diag}(1,-1)$, and $\#^k \mathbb{CP}^2$ has $Q = I_k$.
- The K3 surface (a smooth quartic $\{z_0^4 + z_1^4 + z_2^4 + z_3^4 = 0\} \subset \mathbb{CP}^3$) has $b_2 = 22$ and $q = (-E_8) \oplus (-E_8) \oplus H \oplus H \oplus H$, of signature -16; it is even. We meet E_8 in (1.164).

Worked example 1.116: $S^2 \times S^2$ and $\mathbb{CP}^2\#\overline{\mathbb{CP}^2}$ are different

Both forms have rank 2 and signature 0. Over $\mathbb{R}$ they are equivalent (both are $\mathrm{diag}(1,-1)$ after diagonalising H by $(e_1 \pm e_2)/\sqrt{2}$). Over $\mathbb{Z}$ they are not. For $x = (a, b)$:

$$H(x,x) \overset{(1)}{=} 2ab \ \in 2\mathbb{Z},$$
$$\mathrm{diag}(1,-1)(x,x) \overset{(2)}{=} a^2 - b^2, \qquad \text{which is 1 for } x = (1,0).$$

(1) $x^T H x = a \cdot b + b \cdot a$.
(2) $x^T \mathrm{diag}(1,-1) x = a^2 - b^2$.

Parity of $q(x,x)$ does not depend on the basis, so H is even and $\mathrm{diag}(1,-1)$ is odd, and no $P \in \mathrm{GL}(2,\mathbb{Z})$ relates them. (The real change of basis needed the non-integer $1/\sqrt{2}$.) The two manifolds have the same Betti numbers and the same signature but are not homotopy equivalent. The integers matter.

Dictionary: the intersection form in physics

- A U(1) gauge field on M is a connection on a line bundle L, and line bundles are classified by $m = c_1(L) = [F/2\pi] \in H^2(M;\mathbb{Z})$ (1.128): the magnetic fluxes through the 2-cycles, quantised by Dirac's argument. (Here $F = \mathrm{d}A$ is the real field strength.)
- The θ-term is then $\frac{\theta}{8\pi^2}\int_M F \wedge F = \frac{\theta}{2}\, q(m,m)$. If q is even, $e^{i\frac{\theta}{2}q(m,m)}$ is invariant under $\theta \to \theta + 2\pi$; if q is odd, only under $\theta \to \theta + 4\pi$. The parity of the intersection form is a physical property of the spacetime.
- The lattice $(H^2(M;\mathbb{Z}), q)$ is the lattice of electric and magnetic fluxes in the Euclidean path integral of Maxwell theory; electromagnetic duality acts on it.

Unimodular forms: what exists

Because they are so central, it is worth knowing the classification of unimodular symmetric forms over $\mathbb{Z}$ (Serre, *A Course in Arithmetic*).

- *Indefinite, odd*: always diagonalisable, $q \cong m(1) \oplus n(-1)$. Determined by rank and signature.
- *Indefinite, even*: $q \cong a\,E_8 \oplus b\,H$ (with $a \in \mathbb{Z}$, and $-E_8$ meaning E_8 with the sign reversed). Again determined by rank and signature, and $\sigma = 8a$.
- *Definite*: a zoo. In each rank there are finitely many, but their number grows fast. Even positive definite forms exist only in ranks divisible by 8: E_8 in rank 8; $E_8 \oplus E_8$ and D_{16}^+ in rank 16; the 24 Niemeier lattices in rank 24 (one of them the Leech lattice). Odd ones include I_n and, for example, $E_8 \oplus (1)$ in rank 9.

The signature of an even form is always divisible by 8. The reason is a neat mod-8 identity. Call w **characteristic** if $q(w,x) \equiv q(x,x) \pmod 2$ for all x. Characteristic vectors always exist (reduce mod 2: $x \mapsto q(x,x)$ is linear mod 2, so by unimodularity it is $q(w,\cdot)$ for some w), and van der Blij's lemma says

$$q(w,w) \equiv \sigma(q) \pmod 8 \qquad \text{for every characteristic } w. \tag{1.163}$$

Checks: for (1), $w = 1$ is characteristic and $q(w,w) = 1 = \sigma$. For $\mathrm{diag}(1,-1)$, $w = (1,1)$ gives $1 - 1 = 0 = \sigma$. For an *even* form, $w = 0$ is characteristic, so $\sigma \equiv 0 \pmod 8$. The E_8 form, with $\sigma = 8$, is the smallest possibility.

1.6.9 Freedman's theorem

That the intersection form is important is old news: Whitehead and Milnor showed (1949–58) that q determines the *homotopy type* of a closed simply connected 4-manifold. Freedman (1982) showed that it almost determines the *homeomorphism type*.

Key idea: Freedman's theorem [Nash p. 23–24]

Let M be a closed, simply connected topological 4-manifold with intersection form q.

- If q is even, M is unique up to homeomorphism.
- If q is odd, there are exactly two non-homeomorphic M with this q. They are distinguished by the Kirby–Siebenmann class $ks(M) \in H^4(M;\mathbb{Z}_2) = \mathbb{Z}_2$, and the one with $ks \neq 0$ has no smooth structure.
- Every unimodular form is the intersection form of some such M.

Compare with dimension ≥ 5, where the classification of simply connected manifolds needs far more data: here a single integer matrix, up to change of basis, together with one bit, is everything.

Worked example 1.117: The topological Poincaré conjecture in dimension 4

The Poincaré conjecture in dimension 4 says: a closed 4-manifold S^4_h with the homotopy type of S^4 is homeomorphic to S^4. Nash's proof from Freedman's theorem [Nash p. 24] is a chain of implications:

$$\begin{aligned} S^4_h \simeq S^4 &\overset{(1)}{\Longrightarrow} \pi_1(S^4_h) = 0 \text{ and } H^2(S^4_h;\mathbb{Z}) = H^2(S^4;\mathbb{Z}) = 0 \\ &\overset{(2)}{\Longrightarrow} q_{S^4_h} = \varnothing = q_{S^4}, \text{ which is even} \\ &\overset{(3)}{\Longrightarrow} S^4_h \cong S^4. \end{aligned}$$

(1) Homotopy groups and cohomology are homotopy invariants (§1.2 and Section 1.4).
(2) With no classes of degree 2, the intersection form is the empty form, and it has no odd diagonal entry.
(3) Freedman: an even form determines a unique homeomorphism class, and S^4 is in it.

The Poincaré conjecture had been proved for $n \geq 5$ by Smale (1961). Nash wrote that it remained open only for $n = 3$, the case Poincaré himself proposed; Perelman settled that in 2003, using Ricci flow. The *smooth* Poincaré conjecture in dimension 4 (is every smooth homotopy 4-sphere *diffeomorphic* to S^4?) is still open.

Worked example 1.118: Two manifolds with $q = (1)$

The form (1) is odd, so Freedman gives two homeomorphism classes. One is $\mathbb{CP}^2$, with $ks = 0$. The other is the **Chern manifold** (or "fake $\mathbb{CP}^2$") $*\mathbb{CP}^2$, with $ks = 1$. It is homotopy equivalent to $\mathbb{CP}^2$ (same q) but not homeomorphic to it, and it has no smooth or PL structure

at all. Its existence shows that in dimension 4 the Kirby–Siebenmann class is a genuine obstruction, as in higher dimensions.

The E_8 form and the E_8 manifold

The most famous unimodular form is [Nash p. 24, (1.73)]

$$q = \begin{pmatrix} 2 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 2 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 2 & -1 & 0 & -1 \\ 0 & 0 & 0 & 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 2 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 2 \end{pmatrix} = E_8. \tag{1.164}$$

It is the Cartan matrix of the exceptional Lie algebra $\mathfrak{e}_8$ (Chapter T3): the diagonal entries are 2 and the entry -1 appears exactly when two nodes of the Dynkin diagram are joined. Reading off the -1's, node i is joined to node $i+1$ for $i = 1, \dots, 6$, and node 8 is joined to node 5 (Figure 1.63).

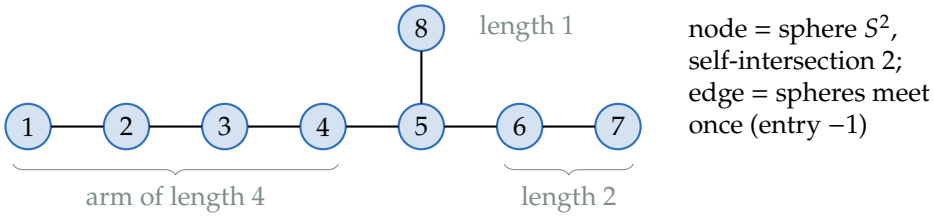


Figure 1.63: The E_8 Dynkin diagram with Nash's numbering of (1.164). Node 5 is the branch point; the three arms have lengths 4, 2 and 1. To build a 4-manifold with this intersection form, take one copy of the tangent disc bundle of S^2 (self-intersection $\chi(S^2) = 2$) for each node and *plumb* two of them (glue them so that their zero sections cross once) for each edge.

Worked example 1.119: E_8 is even, unimodular, positive definite, of signature 8

Even. All diagonal entries are 2. Explicitly $q(x,x) = 2\sum_i x_i^2 - 2\sum_{\text{edges } ij} x_i x_j \in 2\mathbb{Z}$.
Determinant and definiteness by Gaussian elimination. Eliminate the nodes of the arms from the tips inwards. Removing a node i with current diagonal entry (pivot) d_i, joined by an entry -1 to a neighbour j, replaces Q_{jj} by $Q_{jj} - (-1)^2/d_i$ (the Schur complement). Along an arm starting at the tip,

$$d^{(1)} = 2, \quad d^{(2)} = 2 - \tfrac{1}{2} = \tfrac{3}{2}, \quad d^{(3)} = 2 - \tfrac{2}{3} = \tfrac{4}{3}, \quad d^{(4)} = 2 - \tfrac{3}{4} = \tfrac{5}{4}, \qquad d^{(m)} = \tfrac{m+1}{m}.$$

So the arm $1,2,3,4$ gives pivots $2, \frac{3}{2}, \frac{4}{3}, \frac{5}{4}$; the arm $7,6$ gives $2, \frac{3}{2}$; the arm 8 gives 2. The branch node 5 is joined to the last node of each arm, so its pivot is

$$\begin{aligned} d_5 &\overset{(1)}{=} 2 - \frac{1}{d_4} - \frac{1}{d_6} - \frac{1}{d_8} \\ &\overset{(2)}{=} 2 - \frac{4}{5} - \frac{2}{3} - \frac{1}{2} \\ &\overset{(3)}{=} \frac{60 - 24 - 20 - 15}{30} = \frac{1}{30}. \end{aligned}$$

(1) Three Schur-complement corrections, one from each arm.

(2) The last pivots of the arms: $d_4 = \frac{5}{4}$ (arm of length 4), $d_6 = \frac{3}{2}$ (length 2), $d_8 = 2$ (length 1).
(3) Common denominator 30.

The determinant is the product of the pivots:

$$\det E_8 \overset{(1)}{=} \left(2 \cdot \tfrac{3}{2} \cdot \tfrac{4}{3} \cdot \tfrac{5}{4}\right)\left(2 \cdot \tfrac{3}{2}\right)(2)\left(\tfrac{1}{30}\right)$$
$$\overset{(2)}{=} 5 \cdot 3 \cdot 2 \cdot \tfrac{1}{30} = 1.$$

(1) Gaussian elimination $Q = LDL^T$ with L unit lower triangular, so $\det Q = \det D$.
(2) Each arm's product telescopes: $\prod_{m=1}^{\ell} \frac{m+1}{m} = \ell + 1$.

So E_8 is unimodular. All eight pivots are positive, so (Sylvester's criterion, in its LDL^T form) E_8 is positive definite: $b_+ = 8$, $b_- = 0$, $\sigma = 8$. Being even, it has no vectors with $q(x, x) = 1$; its 240 vectors with $q(x, x) = 2$ are the roots of $\mathfrak{e}_8$.

By Freedman's theorem there is a closed, simply connected topological 4-manifold M_{E_8} with $q = E_8$, unique up to homeomorphism. Concretely: the plumbing of Figure 1.63 is a smooth 4-manifold with boundary whose intersection form is E_8; its boundary is the Poincaré homology sphere $\Sigma(2, 3, 5)$ (a 3-manifold with the homology of S^3 but π_1 the binary icosahedral group of order 120). Freedman proved that every homology 3-sphere bounds a contractible *topological* 4-manifold. Capping off with it gives M_{E_8}. The cap cannot be chosen smooth, and the next result shows why.

1.6.10 Rohlin's theorem

Key idea: Rohlin's theorem [Nash p. 24]

If M is a closed, smooth, simply connected 4-manifold with even intersection form, then $\sigma(q)$ is divisible by 16.

For simply connected M, "q even" is the same as "M is spin" ($w_2 = 0$, Example 1.111). Indeed Wu's formula says $w_2 \cup x = x \cup x \bmod 2$ for every $x \in H^2(M; \mathbb{Z}_2)$, so w_2 is the mod-2 reduction of a characteristic vector, and it vanishes iff 0 is characteristic, i.e. iff q is even. The purely algebraic bound (1.163) gives $8 \mid \sigma$; smoothness buys one more factor of 2. There is a beautiful physicist's explanation, through the index of the Dirac operator.

Physical picture: Rohlin from fermion zero modes

On a smooth spin 4-manifold the Dirac operator $\not{D}$ exists (Chapter T4), and its chiral part $\not{D}^+ : S^+ \to S^-$ has an index, index $\not{D}^+ = n_+ - n_-$, the number of positive-chirality minus negative-chirality zero modes. The Atiyah–Singer index theorem (Chapter 4) computes it:

$$\operatorname{index} \not{D}^+ \overset{(1)}{=} \hat{A}(M) = -\frac{1}{24}\, p_1[M]$$
$$\overset{(2)}{=} -\frac{\sigma(M)}{8}.$$

(1) The index theorem for the Dirac operator in 4 dimensions: the $\hat{A}$ genus is $1 - \frac{p_1}{24} + \cdots$.
(2) The Hirzebruch signature theorem in 4 dimensions, $\sigma = \frac{1}{3} p_1[M]$.

In 4 Euclidean dimensions $\mathrm{Spin}(4) = \mathrm{SU}(2) \times \mathrm{SU}(2)$, and each chiral spinor is a doublet of one SU(2): a *pseudo-real* (quaternionic) representation. The antilinear map $\psi \mapsto \epsilon \psi^*$

(with $\epsilon = i\sigma_2$, $\epsilon^2 = -1$) commutes with $\not{D}$ and squares to -1, like time reversal for a spin-$\frac{1}{2}$ particle. So zero modes come in Kramers pairs, $n_\pm$ are even, and index $\not{D}^+$ is even. Then $\sigma/8$ is even: $16 \mid \sigma$.

Worked example 1.120: The E_8 manifold cannot be smoothed

M_{E_8} is simply connected with even form and $\sigma = 8$ (Example 1.119). If it were smooth, Rohlin would force $16 \mid 8$. So M_{E_8} has no smooth structure. In fact it has no PL structure either: for simply connected topological spin 4-manifolds $ks(M) \equiv \sigma(M)/8 \pmod 2$, so $ks(M_{E_8}) = 1$.
For comparison: $E_8 \oplus E_8$ has $\sigma = 16$ and *passes* Rohlin's test; the K3 surface, with $q = -2E_8 \oplus 3H$ and $\sigma = -16$, passes it and is indeed smooth, with index $\not{D}^+ = -\sigma/8 = 2$. Rohlin cannot decide whether the manifold $M_{E_8 \oplus E_8}$ is smoothable. Donaldson can.

1.6.11 Donaldson's theorem

Key idea: Donaldson's theorem [Nash p. 25]

If M is a closed, smooth, simply connected 4-manifold with positive definite intersection form (always achievable for a definite form by choosing the orientation), then q is diagonalisable over the integers:

$$q \cong \mathrm{diag}(1, 1, \ldots, 1) = I_n.$$

The consequences are immediate and dramatic.

- *No smooth manifold has an even definite form* (other than the empty one): I_n has $q(e_1, e_1) = 1$, so it is odd. So M_{E_8} (already excluded by Rohlin) and $M_{E_8 \oplus E_8}$ (not excluded by Rohlin) are not smoothable. Before Donaldson, surgery theorists had tried hard to construct a smooth $M_{E_8 \oplus E_8}$; the attempts were doomed.
- *Odd forms are not safe either*: $E_8 \oplus (1)$ is odd and positive definite of rank 9 but not diagonalisable (it has only two vectors of square 1, $\pm e_9$, while I_9 has eighteen), so it is not smooth either.
- Freedman: *every* unimodular form is realised topologically. Donaldson: among the positive definite ones, only *one* per rank is realised smoothly. The gap between topological and smooth 4-manifolds is enormous.

What makes the theorem so striking is its proof. It uses the Yang–Mills equations. We now sketch it, following Nash.

Yang–Mills and instantons

Let $P \to M$ be a principal SU(2) bundle (Section 1.5.5) and A a connection on it, locally a 1-form with values in anti-Hermitian traceless 2×2 matrices, with curvature $F = \mathrm{d}A + A \wedge A$ (Chapter T3). Its size is measured by the Euclidean Yang–Mills action [Nash p. 25, (1.74)]

$$S(A) = \|F\|^2 = -\int_M \mathrm{tr}(F \wedge \star F). \tag{1.165}$$

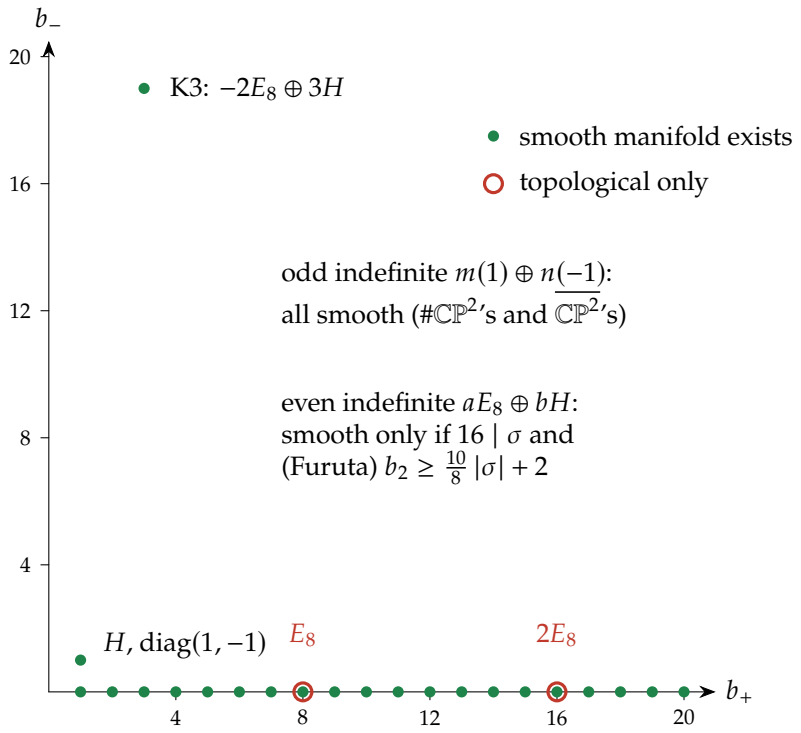


Figure 1.64: The landscape of closed simply connected 4-manifolds, by (b_+, b_-). Freedman realises every unimodular form topologically. Along the positive-definite axis $b_- = 0$, Donaldson allows only I_n (green dots); the even definite forms E_8, $E_8 \oplus E_8$ (red circles) and every other non-diagonal definite form exist only as topological manifolds. The same holds on the negative-definite axis by reversing orientation.

The minus sign makes $S \geq 0$. With $F = \frac{1}{2}F_{\mu\nu}\mathrm{d}x^\mu \wedge \mathrm{d}x^\nu$ in an orthonormal frame,

$$-\operatorname{tr}(F \wedge \star F) \overset{(1)}{=} -\tfrac{1}{2}\sum_{\mu,\nu} \operatorname{tr}\big(F_{\mu\nu}F_{\mu\nu}\big)\,\mathrm{vol}$$
$$\overset{(2)}{=} \tfrac{1}{2}\sum_{\mu,\nu} \operatorname{tr}\big(F^\dagger_{\mu\nu}F_{\mu\nu}\big)\,\mathrm{vol} \overset{(3)}{\geq} 0.$$

(1) For 2-forms $\alpha \wedge \star\beta = \frac{1}{2}\alpha_{\mu\nu}\beta_{\mu\nu}\,\mathrm{vol}$ (Chapter T1); the trace acts on the matrix values.
(2) Each $F_{\mu\nu}$ is anti-Hermitian, $F^\dagger_{\mu\nu} = -F_{\mu\nu}$.
(3) $\operatorname{tr}(X^\dagger X) = \sum_{ab}|X_{ab}|^2 \geq 0$.

Dictionary: mathematicians' and physicists' Yang–Mills

With $A_{\text{math}} = -ig\,A^a_\mu T^a \mathrm{d}x^\mu$, $T^a = \sigma^a/2$ Hermitian with $\operatorname{tr}(T^aT^b) = \frac{1}{2}\delta^{ab}$, the same computation gives $S_{\text{math}} = \frac{g^2}{4}\int F^a_{\mu\nu}F^a_{\mu\nu}\,\mathrm{d}^4x = g^2 S_{\text{phys}}$. So the Euclidean path-integral weight is $e^{-S_{\text{phys}}} = e^{-S_{\text{math}}/g^2}$, and the bound below becomes the famous instanton action $8\pi^2|k|/g^2$.

The Hodge star on 2-forms in 4 dimensions squares to +1, so every F splits into self-dual and anti-self-dual parts, $F = F_+ + F_-$ with $\star F_\pm = \pm F_\pm$. The topological charge is the second Chern class (Section 1.5.12), and with Nash's choice of self-dual instantons we fix the sign of the instanton number as

$$k := -c_2(P)[M] = -\frac{1}{8\pi^2}\int_M \operatorname{tr}(F \wedge F). \tag{1.166}$$

(It is an integer independent of A, by Chern–Weil.) Then

$$\begin{aligned} S &\overset{(1)}{=} -\int \operatorname{tr}\big((F_+ + F_-) \wedge (F_+ - F_-)\big) \\ &\overset{(2)}{=} -\int \operatorname{tr}(F_+ \wedge F_+) + \int \operatorname{tr}(F_- \wedge F_-) = \|F_+\|^2 + \|F_-\|^2 , \\ 8\pi^2 k &\overset{(3)}{=} -\int \operatorname{tr}(F_+ \wedge F_+) - \int \operatorname{tr}(F_- \wedge F_-) = \|F_+\|^2 - \|F_-\|^2 , \end{aligned}$$

(1) $\star F = F_+ - F_-$.
(2) The cross terms $F_+ \wedge F_- = -F_+ \wedge \star F_-$ vanish, because self-dual and anti-self-dual forms are orthogonal ($\alpha \wedge \star\beta = \langle \alpha, \beta \rangle \mathrm{vol}$ and $\langle F_+, F_- \rangle = \langle \star F_+, \star F_- \rangle = -\langle F_+, F_- \rangle$). Then use $F_\pm \wedge F_\pm = \pm F_\pm \wedge \star F_\pm$ and $\|F_\pm\|^2 = -\int \operatorname{tr}(F_\pm \wedge \star F_\pm)$.
(3) Definition (1.166) and the same splitting.

Subtracting, $S - 8\pi^2 k = 2\|F_-\|^2 \geq 0$. So in the sector with instanton number $k \geq 0$,

$$S \geq 8\pi^2 k, \qquad \text{with equality iff} \qquad F = \star F. \tag{1.167}$$

The absolute minima of the action in each sector are the **instantons**, the solutions of the self-duality equations [Nash p. 25, (1.75)] $F = \star F$. (A self-dual connection automatically solves the Yang–Mills equations $\mathrm{d}_A \star F = 0$, by the Bianchi identity $\mathrm{d}_A F = 0$.)

The moduli space

Two connections related by a gauge transformation are physically the same. The **moduli space** is

$$\mathcal{M} = \{A : F_A = \star F_A\} / \mathcal{G},$$

self-dual connections modulo gauge. Perturbing an instanton, $A \to A + a$, and linearising $F = \star F$ modulo gauge gives an elliptic system of equations for a (Chapter 2), so $\mathcal{M}$ is finite-dimensional, and its dimension is computed by an index theorem (Chapters 4 and 8). For SU(2), instanton number k and M simply connected with positive definite form,

$$\dim \mathcal{M} = 8k - 3, \qquad \text{so for } k = 1 : \quad \dim \mathcal{M} = 5.$$

On S^4 the 5 parameters of the minimal ($k = 1$) instanton are visible in the BPST solution: 4 for its centre $x_0 \in S^4$ and 1 for its size $\lambda > 0$. So $\mathcal{M}_{S^4}$ is an open 5-ball, with the centre of the ball the round, maximally spread-out instanton and the boundary sphere reached as $\lambda \to 0$: an instanton shrunk to a point of S^4. *The boundary of the moduli space is a copy of the manifold itself.*

Donaldson's analysis of $\mathcal{M}$ for $k = 1$ on a general M (smooth, simply connected, positive definite, with a generic metric) gives four facts.

1. *The collar.* Very concentrated instantons are labelled by a point $x \in M$ and a small size λ (Taubes constructed them by gluing a tiny BPST instanton into M). So $\mathcal{M}$ contains a collar $M \times (0, \lambda_0)$ and is non-compact, but compactness fails *only* there (Uhlenbeck). This shows $\mathcal{M} \neq \varnothing$.

2. *Smoothness.* Away from finitely many special points $\mathcal{M}$ is a smooth, orientable 5-manifold (Freed and Uhlenbeck for generic metrics; Donaldson for the orientation).

3. *Reducible connections.* The special points are connections whose structure group reduces from SU(2) to U(1): $P = L \oplus L^{-1}$ for a line bundle L with $e = c_1(L)$, with a self-dual U(1) connection. For these

$$c(L \oplus L^{-1}) \overset{(1)}{=} (1 + e)(1 - e) \overset{(2)}{=} 1 - e^2 \quad \Rightarrow \quad k = -c_2[M] \overset{(3)}{=} q(e, e).$$

(1) The total Chern class is multiplicative on direct sums, and $c_1(L^{-1}) = -e$.
(2) Expand; the degree-4 part is $c_2 = -e^2$.
(3) Definition (1.166) and $e^2[M] = q(e,e)$ (1.162).

So for $k = 1$ the reducibles correspond to classes with $q(e,e) = 1$, with e and $-e$ giving the same connection (swap L and L^{-1}). (A self-dual U(1) connection exists for each such e: its curvature is the harmonic representative of $-2\pi i e$, which is self-dual because q is positive definite, so all harmonic 2-forms are self-dual.) Let

$$p = \#\{\pm e \in H^2(M;\mathbb{Z}) : \; q(e,e) = 1\} \tag{1.168}$$

be the number of such pairs.

4. *Cones.* A reducible connection is fixed by a U(1) of gauge transformations, not just by the centre ±1. Near it, $\mathcal{M}$ is modelled on (deformations)/U(1). The deformation space has real dimension $\dim \mathcal{M} + \dim \mathrm{U}(1) = 6$, i.e. it is $\mathbb{C}^3$, and U(1) acts on it by phases. So $\mathcal{M}$ looks like $\mathbb{C}^3/\mathrm{U}(1)$ near the point: a cone on $(\mathbb{C}^3 \setminus 0)/\mathrm{U}(1)$ restricted to the unit sphere, $S^5/\mathrm{U}(1) = \mathbb{CP}^2$. *Each reducible is a cone point whose link is* $\mathbb{CP}^2$ [Nash p. 26].

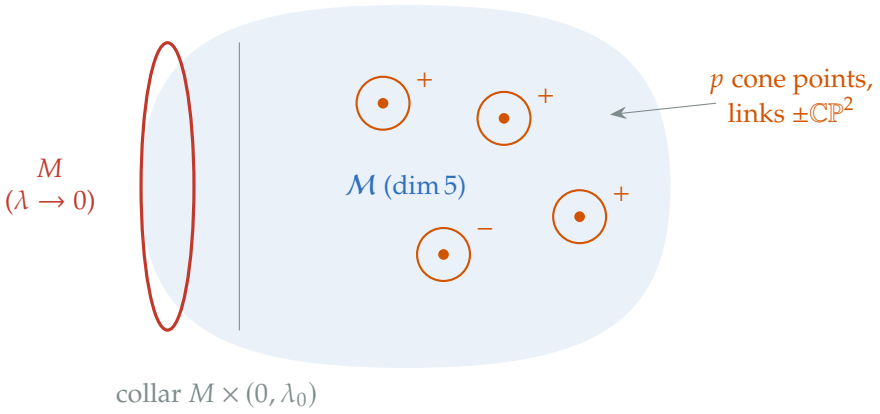


Figure 1.65: Donaldson's moduli space of $k = 1$ instantons on a smooth, simply connected, positive definite M. At one end (red) the instantons shrink to points of M, and the moduli space looks like $M \times (0, \lambda_0)$. Inside there are p singular points (orange), the reducible connections, each a cone on $\mathbb{CP}^2$. Removing small neighbourhoods of both gives a cobordism between M and p copies of $\mathbb{CP}^2$ with various orientations.

Signature is a cobordism invariant

The last ingredient is topological. If a closed oriented 4-manifold X is the boundary of a compact oriented 5-manifold W, then $\sigma(X) = 0$. Here is the mechanism. Let $L \subset H^2(X;\mathbb{R})$ be the classes that extend to closed forms on W. For $a = [\alpha|_X]$, $b = [\beta|_X]$ in L,

$$\begin{aligned} q_X(a,b) &\overset{(1)}{=} \int_X \alpha \wedge \beta \\ &\overset{(2)}{=} \int_W \mathrm{d}(\alpha \wedge \beta) \\ &\overset{(3)}{=} \int_W (\mathrm{d}\alpha \wedge \beta + \alpha \wedge \mathrm{d}\beta) = 0. \end{aligned}$$

(1) The de Rham form of the intersection form, using the restrictions of the extensions.
(2) Stokes' theorem, $\partial W = X$.
(3) Leibniz rule for d (the sign in front of the second term is $(-1)^2 = +1$); α, β are closed on W.

So q vanishes identically on L. Poincaré–Lefschetz duality shows $\dim L = \frac{1}{2}b_2(X)$ ("half the cycles of the boundary bound in the interior"). Now a non-degenerate form that vanishes on a half-dimensional subspace has signature zero: a subspace V_+ on which q is positive definite meets L only in 0, so $b_+ = \dim V_+ \le b_2 - \dim L = \frac{1}{2}b_2$; likewise $b_- \le \frac{1}{2}b_2$; and $b_+ + b_- = b_2$ forces $b_+ = b_-$, i.e. $\sigma(X) = 0$.

Donaldson's argument

Apply this to $W = \mathcal{M}_0$. Its boundary is M together with the p links, a of them oriented like $\mathbb{CP}^2$ and b like $\overline{\mathbb{CP}^2}$, with $a + b = p$. As inner boundary components they enter with the orientation opposite to M's, so $0 = \sigma(\partial\mathcal{M}_0) = \sigma(M) - (a - b)$. Let $n = b_2(M)$. The chain of Nash [Nash p. 26, (1.76)–(1.79)]:

$$\begin{aligned} n &\overset{(1)}{=} \sigma(M) \\ &\overset{(2)}{=} a - b \\ &\overset{(3)}{\le} a + b \\ &\overset{(4)}{=} p \\ &\overset{(5)}{\le} n. \end{aligned}$$

(1) q is positive definite of rank n, so all n eigenvalues are positive [Nash (1.77)].
(2) Cobordism invariance of the signature applied to $\mathcal{M}_0$, with $\sigma(\mathbb{CP}^2) = 1$ and $\sigma(\overline{\mathbb{CP}^2}) = -1$ [Nash (1.76)].
(3) $b \ge 0$.
(4) Each reducible gives exactly one cone point [Nash (1.78)].
(5) The lattice argument below: a positive definite unimodular form of rank n has at most n pairs $\pm e$ with $q(e, e) = 1$.

So $n \le p \le n$, hence $p = n$ [Nash (1.79)]: *there are exactly n pairs of classes with square* 1. It remains to see that this forces $q \cong I_n$.

Worked example 1.121: Vectors of square 1 in a positive definite lattice

Let e, f satisfy $q(e, e) = q(f, f) = 1$ with $f \ne \pm e$. Then

$$\begin{aligned} |q(e,f)| &\overset{(1)}{<} \sqrt{q(e,e)\,q(f,f)} \\ &\overset{(2)}{=} 1, \qquad \text{so} \qquad q(e,f) \overset{(3)}{=} 0. \end{aligned}$$

(1) Cauchy–Schwarz for the positive definite inner product q, strict because e and f are not parallel (the only multiples of e with square 1 are $\pm e$).
(2) Both have square 1.
(3) $q(e, f)$ is an integer of absolute value less than 1.

So one representative from each of the p pairs gives an *orthonormal* set $e_1, \dots, e_p$. Orthonormal vectors are linearly independent, so $p \le n$: this is step (5) above. If $p = n$ they span a sublattice $\Lambda' \subset \Lambda = H^2(M; \mathbb{Z})$ of full rank with Gram matrix I_n. For a sublattice of index d

the Gram determinants satisfy $\det(I_n) = d^2 \det(q)$, i.e. $1 = d^2 \cdot 1$, so $d = 1$ and $\Lambda' = \Lambda$: in the basis $e_1, \dots, e_n$, $q = I_n$. That is Donaldson's theorem.
Test cases for the count p: I_n has $p = n$ (the pairs $\pm e_i$); E_8 has $p = 0 < 8$ (it is even); $E_8 \oplus (1)$ has $p = 1 < 9$. The last two cannot be the forms of smooth manifolds.

Watch out: Nash's "p unit eigenvalues"

Nash phrases step (5) as "q regarded as an $n \times n$ matrix with precisely p unit eigenvalues (clearly $p \leq n$)" [Nash p. 26]. Read literally this is not right: the eigenvalues of a matrix change under $Q \mapsto P^T Q P$, and E_8 has no eigenvalue equal to 1 in any sense relevant here. The correct statement is the lattice one: p is the number of pairs of *integral* classes of square 1, which is basis-independent, is at most n by Example 1.121, and equals n exactly when $q \cong I_n$.

Physical picture: Why a physicist should be amazed

The moduli space of solutions of a classical field equation, the self-duality equation, is a 5-dimensional space that remembers the 4-manifold: one of its ends *is* M, and its singularities count the square-one classes. Field theory turned into a cobordism. Donaldson went on to extract numerical invariants of smooth 4-manifolds from these moduli spaces (Donaldson invariants), and Witten showed they are the correlation functions of a topological quantum field theory, a twisted version of $N = 2$ super Yang–Mills (Chapter 12). The Seiberg–Witten equations later reproduced all of this with a U(1) monopole equation and compact moduli spaces. We return to instantons in Chapter 8.

1.6.12 Non-compact manifolds and exotic $\mathbb{R}^4$

For non-compact manifolds, Nash reports [Nash p. 26]:

- *Quinn*: every non-compact connected 4-manifold is smoothable. (No closed 4-manifold phenomenon like M_{E_8} occurs.)
- *Uniqueness fails badly.* For $n \neq 4$, $\mathbb{R}^n$ has exactly one smooth structure up to diffeomorphism (Stallings, for $n \geq 5$; Moise for $n \leq 3$). For $n = 4$ there is an **exotic** (or *fake*) $\mathbb{R}^4$: a smooth manifold homeomorphic to $\mathbb{R}^4$ but not diffeomorphic to it. This follows from Freedman's and Donaldson's work.
- *Gompf* (and Taubes): there is a continuum, i.e. uncountably many, of pairwise non-diffeomorphic smooth structures on $\mathbb{R}^4$.

This is in sharp contrast with the compact case, where obstruction theory produced a finite, discrete set of structures (28 on S^7).

Physical picture: How an exotic $\mathbb{R}^4$ arises

The logic, stripped of its (formidable) technicalities, is a proof by contradiction that uses both theorems.

1. Take a smooth 4-manifold whose intersection form splits algebraically as a definite non-diagonal piece plus hyperbolic pieces; the K3 surface, $q = (-E_8 \oplus -E_8) \oplus 3H$, is the standard choice.
2. Freedman's topological methods split the manifold *topologically* along this decom-

position, into a piece carrying $-E_8 \oplus -E_8$ and a piece carrying $3H$. The splitting is controlled by a certain open subset that is *homeomorphic* to $\mathbb{R}^4$. Being an open subset of a smooth manifold, it inherits a smooth structure.

3. If that smooth $\mathbb{R}^4$ were the standard one, a smooth ball could be found where the topological theory needs it, and the splitting could be carried out *smoothly*, producing a closed smooth manifold with definite form $-E_8 \oplus -E_8$.

4. Donaldson forbids that. So the inherited smooth structure on this $\mathbb{R}^4$ is exotic.

Exotic $\mathbb{R}^4$'s of this kind cannot even be smoothly embedded in the standard $\mathbb{R}^4$. Others ("small" ones) arise from Donaldson's discovery that the smooth h-cobordism theorem fails in dimension 4.

Watch out: What an exotic $\mathbb{R}^4$ is not

An exotic $\mathbb{R}^4$ is not a curved metric on $\mathbb{R}^4$, nor a non-trivial topology. It is homeomorphic to $\mathbb{R}^4$ and, locally, every smooth manifold looks like standard $\mathbb{R}^4$ in a small chart. The difference is global and lies in *which functions count as smooth*: there is no homeomorphism $\mathbb{R}^4_{\text{exotic}} \to \mathbb{R}^4$ that is smooth with smooth inverse. Physicists have speculated (Brans, Asselmeyer-Maluga and others) that exotic smoothness of spacetime might have physical consequences, e.g. in the gravitational path integral, but this remains speculation.

Reading Nash (pp. 23–26, (1.72)–(1.79))

Nash's treatment of dimension 4 is a guided tour. He *defines* the intersection form (1.72), states Freedman's theorem, deduces the 4-dimensional Poincaré conjecture (our Example 1.117), states Rohlin, writes down E_8 (1.73) "by inspection even and, by calculation, of signature 8" (our Example 1.119 does the calculation), states Donaldson, and sketches the proof through (1.74)–(1.79). We added: the Whitney-trick explanation of why dimension 4 is special; the three ways to compute q and its invariants; the list of basic forms and the H versus $\mathrm{diag}(1,-1)$ example; the classification of unimodular forms and the mod-8 theorem; the fake $\mathbb{CP}^2$; the E_8 plumbing; the Dirac-operator explanation of Rohlin; the positivity of S and the bound $S \geq 8\pi^2 k$; the dimension count, the reducibles $L \oplus L^{-1}$ and the cone structure of $\mathcal{M}$; the Stokes proof that signature is a cobordism invariant; and the lattice argument that turns $p = n$ into $q \cong I_n$, correcting Nash's "unit eigenvalues". His last paragraph (Quinn, exotic $\mathbb{R}^4$, Gompf) is our Section 1.6.12. Note that Nash's book predates Perelman (2003) and Seiberg–Witten theory (1994).

Recap: §1.6 Differentiable structures

- Transition functions come in three kinds, forming groups $TOP_n \supset PL_n \supset DIFF_n \simeq \mathrm{O}(n)$, with direct limits TOP, PL, O. The type of a manifold is the lift of its tangent map $M \to BTOP$ to BPL or BO.
- Lifting through a fibration with fibre F is finding a section; obstructions lie in $H^{k+1}(M;\pi_k(F))$ and choices in $H^k(M;\pi_k(F))$. The same pattern gives the Euler class (vector fields), w_1 (orientation) and w_2 (spin structures; four on T^2).
- $TOP/PL = K(\mathbb{Z}_2,3)$: for $\dim M \geq 5$ the only obstruction to a PL structure is $ks(M) \in H^4(M;\mathbb{Z}_2)$, and PL structures are counted by $H^3(M;\mathbb{Z}_2)$. The Hauptvermutung is false.

- $\pi_k(PL/\mathrm{O}) = \Theta_k$ is 0 for $k \le 6$ and $\mathbb{Z}_{28}$ for $k = 7$: S^7 has 28 smooth structures. Milnor's S^3-bundles over S^4 realise them, detected by $2p_1^2 - \sigma \bmod 7$.
- In dimension 4 the Whitney trick fails. The invariant is the intersection form $q(\alpha, \beta) = (\alpha \cup \beta)[M]$, symmetric and unimodular. Freedman: q (plus ks if odd) classifies simply connected topological 4-manifolds, and all forms occur; hence the 4-dimensional topological Poincaré conjecture.
- Rohlin: smooth + even $\Rightarrow 16 \mid \sigma$, so the E_8 manifold is not smooth. Donaldson: smooth + positive definite $\Rightarrow q \cong I_n$, proved by the moduli space of self-dual SU(2) instantons, a 5-dimensional cobordism from M to p copies of $\pm\mathbb{CP}^2$.
- $\mathbb{R}^n$ has a unique smooth structure except for $n = 4$, where there are uncountably many.

Chapter summary

The chapter asked one question: *how can we tell two spaces, or two field configurations, apart without solving any equations?* We answered it with a ladder of invariants. At the top of the ladder sit the finest notions of "the same" (diffeomorphic, PL homeomorphic, homeomorphic). They are what we would really like to decide, and they are almost never decidable by hand. Lower down sit coarser notions (homotopy equivalent, same cohomology ring, same Betti numbers, same Euler characteristic). Each rung sees less than the one above it, but is easier to compute. The art of topology is to climb down only as far as necessary. The whole of Figure 1.66 is the chapter in one picture: each downward arrow is an implication, and each red example on the right shows that the arrow cannot be reversed.

The rungs "isomorphic groups" and "same Betti numbers" are separated by *torsion*. Real coefficients, and hence differential forms, see only the ranks b_k of the homology groups and forget every finite cyclic summand $\mathbb{Z}_k$. The red box between them ($\mathbb{RP}^3$ against S^3) is the reason the chapter needed integer (simplicial and singular) homology as well as de Rham cohomology.

The rest of this summary has four parts: a table of every concept of the chapter with its operational meaning and where it returns; a catalogue of the spaces we computed; a dictionary to physics; and a self-test.

Every concept, how to use it, and where it returns

The tables follow Nash's sections. The middle column is the *operational* meaning: what you actually do to compute or test the thing. The right column says where in the book the idea comes back. References of the form §1.4 point to our own sections; Nash §1–§3 are Sections 1.1–1.3 of these notes.

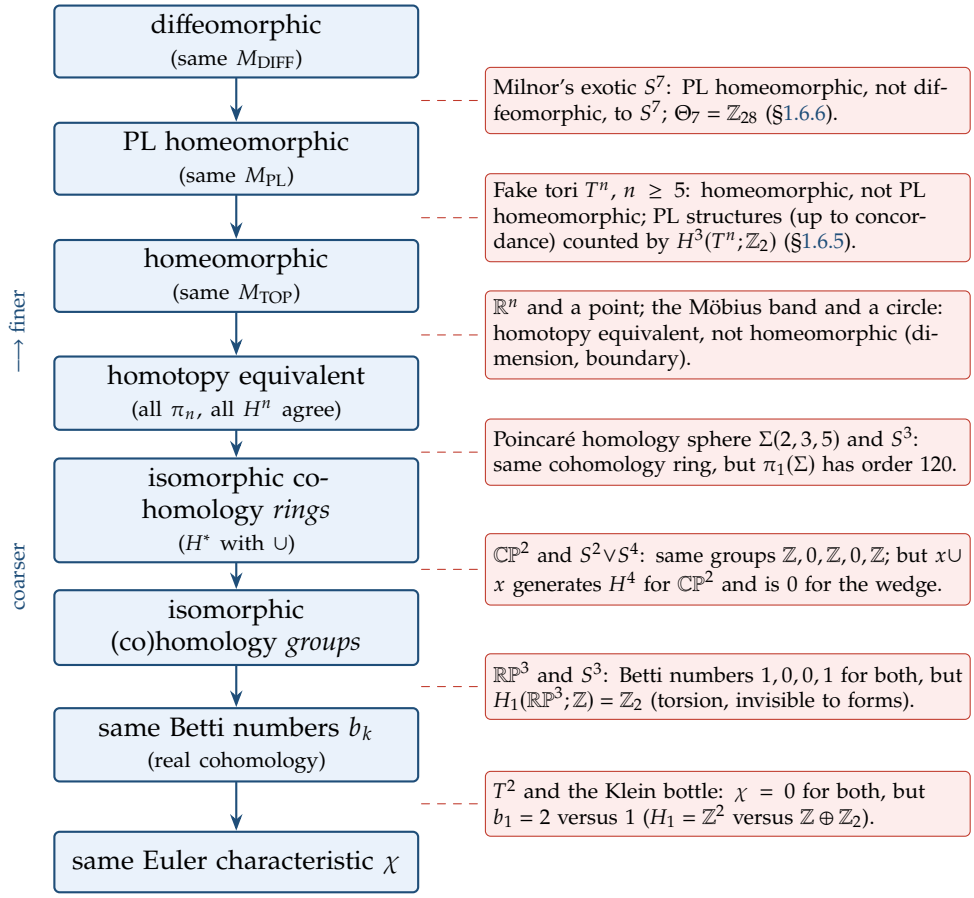


Figure 1.66: The ladder of "sameness" built in this chapter. Read downwards: each notion implies the next. Read the red boxes: each is a pair of spaces that agree on the lower rung but differ on the upper one, so no arrow can be reversed. Nash §1 and §6 live on the top three rungs; §2–§4 build the lower ones.

Nash §1: from homeomorphism to diffeomorphism

Concept	Operational meaning	Returns in
homeomorphism	a continuous bijection with continuous inverse; to *show* two spaces differ, find an invariant (dimension, compactness, connectedness, χ, π_n, H^n) that differs	everywhere
classification of surfaces	closed orientable surfaces are classified by the genus g; compute g from $\chi = 2 - 2g = V - E + F$ of any triangulation	Chapter 5, Chapter 9, Chapter 11
general position	submanifolds of dimensions p, q in M^n generically miss each other if $p + q < n$ and meet in a $(p + q - n)$-manifold otherwise; the reason low dimensions are hard	§1.6.7, Chapter 12
knots	embeddings $S^1 \hookrightarrow \mathbb{R}^3$ up to isotopy; all knots untie in $\mathbb{R}^4$ $(1 + 1 < 4)$	Chapter 12 (Jones polynomial)
topological manifold	cover by charts $U_\alpha \to \mathbb{R}^n$; transition functions $g_{\alpha\beta}$ are homeomorphisms	§1.6.1
PL manifold	transition functions are piecewise linear; equivalently a nice triangulation	§1.4.6, §1.6.5
smooth manifold	transition functions are C^∞ with C^∞ inverse (check the Jacobian is invertible)	all of Chapter T1–Chapter T4
M_{TOP}, M_{PL}, M_{DIFF}	the same set with three kinds of atlas; existence and uniqueness of the finer ones can fail	§1.6, Chapter 10, Chapter 12

Nash §2: some algebraic topology: homotopy

Concept	Operational meaning	Returns in
homotopy $F : [0,1] \times X \to Y$	a continuous one-parameter family of maps; write one down explicitly (straight line $F_t = (1-t)f + tg$ in a convex target, or a radial shrink)	everywhere
$[X, Y]$	the set of maps modulo homotopy; a field configuration's *topological sector*	Chapter 8, Chapter 10
homotopy equivalence, contractible	maps $f : X \to Y$, $g : Y \to X$ with $gf \simeq \mathrm{id}$, $fg \simeq \mathrm{id}$; contractible = homotopy equivalent to a point (e.g. $\mathbb{R}^n$, a disc, a vector space)	Poincaré lemma (Example 1.65), Chapter 2
retract, deformation retract	a continuous projection $r : X \to A$ with $r\|_A = \mathrm{id}$; a deformation retract is a homotopy from id_X to r (punctured plane → circle)	Example 1.67, §1.4.9
base point, $[X, Y]_0$	maps and homotopies that keep $* \mapsto *$; needed to make loops composable	§1.4.2
exponential law	$\mathrm{Map}(X \times Y, Z) = \mathrm{Map}(X, \mathrm{Map}(Y, Z))$: a family of maps is a map into a space of maps	path fibration (§1.5.16), Chapter 6
wedge $X \vee Y$, smash $X \wedge Y$	glue at the base point; collapse $X \vee Y$ inside $X \times Y$; $S^m \wedge S^n = S^{m+n}$	Example 1.82
suspension SX, loop space ΩX	$SX = S^1 \wedge X$; ΩX = based loops; adjoint: $[SX, Y]_0 = [X, \Omega Y]_0$	§1.4.3, §1.5.16, Chapter 7, Chapter 6
H-space, coH-space	a target with a product up to homotopy (ΩY, a group G) or a source with a coproduct (SX) makes $[X, Y]_0$ a group	§1.4.3
Eckmann–Hilton	when both structures exist they agree and are commutative (slide squares past each other)	$\pi_{n \geq 2}$ abelian

Nash §3: homotopy groups

Concept	Operational meaning	Returns in
$\pi_1(X) = [S^1, X]_0$	based loops up to homotopy, composed by "first one, then the other"; compute by van Kampen (glue along simply connected overlaps) or by coverings	§1.4.10, Chapter 8
winding number	$\frac{1}{2\pi}\oint f^*\mathrm{d}\theta = \frac{1}{2\pi i}\oint \frac{\mathrm{d}f}{f} \in \mathbb{Z}$: the isomorphism $\pi_1(S^1) \to \mathbb{Z}$	§1.4.2, §1.5.6
π_1 examples	$\pi_1(T^2) = \mathbb{Z}^2$; genus g: $\langle a_i, b_i \mid \prod[a_i, b_i]\rangle$; $\pi_1(\mathbb{RP}^2) = \mathbb{Z}_2$; $\pi_1(\mathrm{SO}(3)) = \mathbb{Z}_2$ (belt trick, spin $\frac{1}{2}$)	Example 1.86, Chapter T4
$\pi_n(X) = [S^n, X]_0$	based maps of spheres; $= [S^{n-1}, \Omega X]_0$; abelian for $n \geq 2$	§1.5.13
free vs based classes	$[X, Y] = [X, Y]_0/\pi_1(Y)$; for simply connected Y base points do not matter	§1.4.2
degree	$\deg f = \int_{S^n} f^*\omega$ with $\int \omega = 1$, or count preimages with Jacobian signs; gives $\pi_n(S^n) = \mathbb{Z}$	§1.5.8, Chapter 4
$\pi_m(S^n) = 0$ for $m < n$	a (smoothed) map from a lower-dimensional sphere misses a point, and $S^n \setminus \{\mathrm{pt}\} = \mathbb{R}^n$ is contractible	§1.4.1
Hopf map, $\pi_3(S^2) = \mathbb{Z}$	$S^3 \to S^2$, $(z_1, z_2) \mapsto z_1/z_2$; Hopf invariant = linking number of two fibres	§1.5.7, Example 1.100
table, stability	$\pi_{n+k}(S^n)$ depends only on k once $n > k + 1$ (Freudenthal)	§1.6.6
exact sequences	"image = kernel" at every step; a zero on each side of a term makes a map an isomorphism	§1.5.13, §1.4.9
Lie groups, Bott	$\pi_1(\mathrm{U}(n)) = \mathbb{Z}$, $\pi_3(\mathrm{SU}(n)) = \mathbb{Z}$, $\pi_1(\mathrm{SO}(n \geq 3)) = \mathbb{Z}_2$; Bott periodicity: $\pi_k(\mathrm{U})$ has period 2, $\pi_k(\mathrm{O})$ period 8	Chapter 3, Chapter 10
defects	a defect of codimension $k + 1$ with vacuum manifold $\mathcal{V}$ is classified by $\pi_k(\mathcal{V})$: walls (π_0), vortices (π_1), monopoles (π_2), textures (π_3)	Chapter 8
instantons, θ	$k = \frac{1}{8\pi^2}\int \mathrm{tr}\, F \wedge F \in \pi_3(\mathrm{SU}(2)) = \mathbb{Z}$; vacua labelled by k, superposed into θ-vacua	§1.6.11, Chapter 8

Nash §4: cohomology and homology groups (§1.4)

Concept	Operational meaning	Returns in
$K(G,n)$	a space with $\pi_n = G$ and all other $\pi_m = 0$; $S^1 = K(\mathbb{Z},1)$, $\mathbb{RP}^\infty = K(\mathbb{Z}_2,1)$, $\mathbb{CP}^\infty = K(\mathbb{Z},2)$	§1.5.11
$H^n(X;G) = [X, K(G,n)]_0$	e.g. $H^1(X;\mathbb{Z})$ = maps to a circle up to homotopy, counted by winding numbers around the 1-cycles	Example 1.60, §1.5.12
$\Omega K(G,n+1) = K(G,n)$	shifting degree by looping; makes $[X, K(G,n)]_0$ an abelian group	Chapter 3
complex, $H = \mathrm{Ker}/\mathrm{Im}$	a sequence of linear maps with $d \circ d = 0$; its cohomology is *rank arithmetic*: $\dim H^n = \dim \mathrm{Ker}\, d_n - \mathrm{rank}\, d_{n-1}$	every complex in Chapter 2, Chapter 4
de Rham cohomology	closed p-forms modulo exact ones; to find a class, integrate over cycles; to kill one, find a primitive	Chapter T2, Chapter 4
simplicial homology	chains = integer combinations of oriented simplices; write the boundary matrices ∂_k, then $H_k = \mathrm{Ker}\,\partial_k/\mathrm{Im}\,\partial_{k+1}$ (Smith normal form gives torsion)	§1.4.6, Chapter 7
singular homology	the same with all continuous simplices; manifestly a topological invariant; agrees with simplicial	§1.4.6
cochains, coboundary	functions on simplices; $\delta = \partial^{\mathsf{T}}$; Stokes $\langle \delta c, \sigma\rangle = \langle c, \partial\sigma\rangle$	§1.4.7
universal coefficients	$H^n(X;\mathbb{Z}) \cong \mathrm{free}(H_n) \oplus \mathrm{tors}(H_{n-1})$; with $\mathbb{R}$ coefficients torsion disappears	Example 1.76
de Rham theorem	$H^p_{\mathrm{dR}}(M) \cong H^p(M;\mathbb{R})$, the map being "integrate the form over chains"; periods determine the class	§1.4.8, Chapter 3
Mayer–Vietoris	a long exact sequence for $M = U \cup V$; compute $H^*(S^n)$ from two discs	§1.4.9, Chapter 3
finiteness	$H^n = 0$ for $n > \dim M$ and $n < 0$ (unlike π_n)	
cup product, ring	wedge of representatives, $[\alpha] \cup [\beta] = [\alpha \wedge \beta]$; graded commutative	§1.6.8
Poincaré duality	closed oriented M^n: $H^p \cong H_{n-p}$, $b_p = b_{n-p}$; the pairing $\int_M \alpha \wedge \beta$ is non-degenerate	§1.6.8, Chapter 4
Künneth	$H^*(X \times Y) = H^*(X) \otimes H^*(Y)$ (field coefficients); Poincaré polynomials multiply	Example 1.84
Euler characteristic	$\chi = \sum(-1)^k b_k = \sum(-1)^k N_k$ (count simplices) = index of $\mathrm{d} + \mathrm{d}^\dagger$	Chapter 4, Chapter 7
Hurewicz	for simply connected X the first non-zero π_n equals the first non-zero H_n; $H_1 = \pi_1/[\pi_1, \pi_1]$	§1.4.10

Nash §5: fibre bundles and fibrations (§1.5)

Concept	Operational meaning	Returns in
fibre bundle $F \to E \xrightarrow{\pi} M$	locally $\pi^{-1}(O_\alpha) \cong O_\alpha \times F$; to specify one, give a cover and transition functions	Chapter 2 onwards
transition functions, cocycle	$g_{\alpha\beta} : O_\alpha \cap O_\beta \to G$ with $g_{\alpha\beta} g_{\beta\gamma} = g_{\alpha\gamma}$; the structure group G is where they take values	§1.5.2, Chapter 3
change of trivialisation	$g_{\alpha\beta} \mapsto h_\alpha^{-1} g_{\alpha\beta} h_\beta$: a gauge transformation; the bundle is trivial iff one can reach $g_{\alpha\beta} = 1$	Chapter T3, Chapter 8
section	a map $s : M \to E$ with $\pi \circ s = \mathrm{id}$; locally a F-valued function, patched by $s_\alpha = g_{\alpha\beta} s_\beta$; a matter field	§1.5.3, Chapter 2
vector bundle, rank	fibre $\mathbb{R}^k$ or $\mathbb{C}^k$, $g_{\alpha\beta} \in \mathrm{GL}(k)$; TM has $g_{\alpha\beta} =$ Jacobian	§1.5.1
$\oplus$, $\otimes$, dual, Λ^p	do the linear algebra fibrewise, and apply it to the transition matrices ($g \oplus h$, $g \otimes h$, $(g^{\mathsf{T}})^{-1}$)	Chapter 2, Chapter 3
principal bundle P, associated bundle $P \times_\rho V$	fibre G acting on itself; associated bundles use the same $g_{\alpha\beta}$ acting through a representation ρ; frame bundle of TM	§1.5.5, Chapter T3, Chapter T4
Möbius band	the simplest non-trivial bundle, $g = -1$ on one overlap; every section vanishes	Example 1.87
clutching	bundles over $S^n \leftrightarrow [S^{n-1}, G] = \pi_{n-1}(G)$ (connected G)	§1.5.6, Chapter 3, Chapter 8
monopole bundle	$g_{NS} = e^{in\phi}$ on the equator; $\int_{S^2} \frac{i}{2\pi} F = n$: Dirac quantisation	§1.5.6, Chapter 8
Hopf fibration	$S^1 \to S^3 \to S^2$, the $n = 1$ monopole bundle; fibres are linked circles	§1.5.7
TS^2, hairy ball	transition $-1/z^2$ winds twice; $c_1 = \chi(S^2) = 2$; every vector field on S^2 vanishes	§1.5.8, Chapter 7
pullback f^*	pull back forms (substitute), classes, and bundles ($g_{\alpha\beta} \circ f$); homotopic maps pull back isomorphic bundles	§1.5.10, Chapter 4
$EG \to BG$	bundles over $M \leftrightarrow [M, BG]$; $B\mathrm{U}(1) = \mathbb{CP}^\infty$, $B\mathrm{O}(1) = \mathbb{RP}^\infty$; $\pi_n(BG) = \pi_{n-1}(G)$	§1.5.11, Chapter 3
characteristic classes	pull back universal classes: c_i ($\mathrm{U}(n)$), p_i ($\mathrm{O}(n)$), w_i (mod 2), e ($\mathrm{SO}(2m)$); in practice $c_1 = [\frac{i}{2\pi} \operatorname{tr} F]$, $c_2 = \frac{1}{8\pi^2}[\operatorname{tr} F \wedge F]$	§1.5.12, Chapter 3, Chapter 4
homotopy exact sequence	$\cdots \to \pi_n(F) \to \pi_n(E) \to \pi_n(M) \xrightarrow{\partial} \pi_{n-1}(F) \to \cdots$; computes $\pi_3(S^2)$, $\pi_2(G/H)$	§1.5.13, Chapter 8
cohomology of a bundle	no simple rule; the spectral sequence; $\chi(E) = \chi(F)\chi(M)$ survives	§1.5.14, Chapter 7
fibration, homotopy lifting	a map along which homotopies lift; all fibres homotopy equivalent, not necessarily homeomorphic	§1.5.15
path fibration	$\Omega M \to PM \to M$ with PM contractible; gives $\pi_n(\Omega M) = \pi_{n+1}(M)$	§1.5.16, Chapter 7, Chapter 6

Nash §6: differentiable structures for manifolds (§1.6)

Concept	Operational meaning	Returns in
$TOP_n \supset PL_n \supset DIFF_n$	the three groups of local transition functions; $DIFF_n \simeq \mathrm{O}(n)$; direct limits TOP, PL, O	§1.6.1
$BTOP$, BPL, $B\mathrm{O}$	the tangent bundle is a map $M \to BTOP$; a PL or smooth structure is a lift to BPL or $B\mathrm{O}$	§1.6.2
lifting problem	a lift through $F \to E \to B$ is a section of a pulled-back bundle with fibre F	§1.6.3
obstruction theory	extend a section cell by cell; the obstruction to the next step lies in $H^{k+1}(M;\pi_k(F))$, choices in $H^k(M;\pi_k(F))$	§1.6.4, Chapter 10
Euler class, w_1, w_2	obstructions to a non-vanishing vector field, to orientation, to a spin structure; spin structures counted by $H^1(M;\mathbb{Z}_2)$	Example 1.110, Example 1.111, Chapter T4
Kirby–Siebenmann class	$TOP/PL = K(\mathbb{Z}_2,3)$; for $\dim M \geq 5$ a PL structure exists iff $ks(M) = 0 \in H^4(M;\mathbb{Z}_2)$; count by $H^3(M;\mathbb{Z}_2)$	§1.6.5
Hauptvermutung	"homeomorphic triangulated spaces are PL homeomorphic"; false	§1.6.5
PL/O, Θ_n	$\pi_k(PL/\mathrm{O}) = \Theta_k$: 0 for $k \leq 6$, $\mathbb{Z}_{28}$ for $k = 7$; smooth structures on S^7	§1.6.6, Chapter 10
Milnor spheres	S^3-bundles over S^4 with Euler number 1, told apart by $2p_1^2 - \sigma \bmod 7$	Example 1.114
Whitney trick	cancel intersection points across an embedded disc; works for $n \geq 5$, fails for $n = 4$	§1.6.7
intersection form q	$q(\alpha,\beta) = \int_M \alpha \wedge \beta$ on $H^2(M;\mathbb{Z})$; symmetric unimodular matrix; rank, signature, type	§1.6.8, Chapter 4, Chapter 12
Freedman	q (and ks if odd) classifies simply connected closed topological 4-manifolds; all forms occur	§1.6.9
E_8	the Cartan matrix of $\mathfrak{e}_8$: even, positive definite, $\sigma = 8$; the E_8 manifold is not smoothable	Example 1.119, Example 1.120
Rohlin	smooth + spin $\Rightarrow 16 \mid \sigma$; from the quaternionic zero modes of the Dirac operator	§1.6.10, Chapter 4
Donaldson	smooth + simply connected + definite $\Rightarrow q \cong \pm I_n$; proof via the 5-dimensional moduli space of $k = 1$ instantons	§1.6.11, Chapter 8, Chapter 12
exotic $\mathbb{R}^4$	$\mathbb{R}^n$ has one smooth structure for $n \neq 4$ and uncountably many for $n = 4$	§1.6.12

A catalogue of the spaces we met

Most statements in this chapter were tested on the same dozen spaces. Table 1.1 collects what we know about them. Knowing this table by heart is the best single preparation for the rest of the book. The integral homology column lists $H_0, H_1, H_2, \ldots$ up to the dimension; cohomology follows from the universal coefficient rule (torsion moves up one degree).

A few entries deserve a line of justification, since not all were computed in the text.

- *Aspherical surfaces.* T^2, $\Sigma_{g\geq 2}$ and K have contractible universal cover $\mathbb{R}^2$ (or the hyperbolic plane). A covering map induces isomorphisms on π_n for $n \geq 2$, so all their higher homotopy groups vanish.
- $\mathbb{RP}^2$. Its universal cover is S^2, so $\pi_n(\mathbb{RP}^2) = \pi_n(S^2)$ for $n \geq 2$: in particular $\pi_3(\mathbb{RP}^2) = \mathbb{Z}$, although $\mathbb{RP}^2$ is only 2-dimensional.
- $\mathbb{CP}^2$. The fibration $S^1 \to S^5 \to \mathbb{CP}^2$ (the Hopf construction one dimension up) gives

space X	π_1	π_2	π_3	$H_0, H_1, H_2, \dots$	χ
point, $\mathbb{R}^n$, disc	0	0	0	$\mathbb{Z}$	1
$S^1 = \mathrm{U}(1)$	$\mathbb{Z}$	0	0	$\mathbb{Z}, \mathbb{Z}$	0
$S^2 = \mathbb{CP}^1$	0	$\mathbb{Z}$	$\mathbb{Z}$	$\mathbb{Z}, 0, \mathbb{Z}$	2
$S^3 = \mathrm{SU}(2)$	0	0	$\mathbb{Z}$	$\mathbb{Z}, 0, 0, \mathbb{Z}$	0
S^n, $n \geq 4$	0	0	0	$\mathbb{Z}$ in degrees $0, n$	$1 + (-1)^n$
T^2	$\mathbb{Z}^2$	0	0	$\mathbb{Z}, \mathbb{Z}^2, \mathbb{Z}$	0
Σ_g, $g \geq 2$	$\langle a_i, b_i \mid \prod_i [a_i, b_i] \rangle$	0	0	$\mathbb{Z}, \mathbb{Z}^{2g}, \mathbb{Z}$	$2 - 2g$
Klein bottle K	$\langle a, b \mid abab^{-1} \rangle$	0	0	$\mathbb{Z}, \mathbb{Z} \oplus \mathbb{Z}_2, 0$	0
$\mathbb{RP}^2$	$\mathbb{Z}_2$	$\mathbb{Z}$	$\mathbb{Z}$	$\mathbb{Z}, \mathbb{Z}_2, 0$	1
$\mathbb{RP}^3 = \mathrm{SO}(3)$	$\mathbb{Z}_2$	0	$\mathbb{Z}$	$\mathbb{Z}, \mathbb{Z}_2, 0, \mathbb{Z}$	0
$\mathbb{CP}^2$	0	$\mathbb{Z}$	0	$\mathbb{Z}, 0, \mathbb{Z}, 0, \mathbb{Z}$	3
$S^2 \times S^2$	0	$\mathbb{Z}^2$	$\mathbb{Z}^2$	$\mathbb{Z}, 0, \mathbb{Z}^2, 0, \mathbb{Z}$	4
$\mathrm{SU}(3)$	0	0	$\mathbb{Z}$	$\mathbb{Z}$ in degrees $0, 3, 5, 8$	0
$\mathbb{CP}^\infty = B\mathrm{U}(1) = K(\mathbb{Z}, 2)$	0	$\mathbb{Z}$	0	$\mathbb{Z}$ in every even degree	–
$\mathbb{RP}^\infty = B\mathrm{O}(1) = K(\mathbb{Z}_2, 1)$	$\mathbb{Z}_2$	0	0	$\mathbb{Z}, \mathbb{Z}_2, 0, \mathbb{Z}_2, 0, \dots$	–

Table 1.1: The spaces of the chapter. Read the columns against each other: π_1 abelianises to H_1 (Hurewicz); the first non-zero π_n of a simply connected space equals H_n; the higher π_n carry information that homology throws away (compare $\pi_3(S^2) = \mathbb{Z}$ with $H_3(S^2) = 0$); χ is the alternating sum of the ranks.

$\pi_n(\mathbb{CP}^2) = \pi_n(S^5) = 0$ for $3 \leq n \leq 4$ and $\pi_2(\mathbb{CP}^2) = \pi_1(S^1) = \mathbb{Z}$, by the same argument as Example 1.100.

- $S^2 \times S^2$. Homotopy groups of products are products; homology by Künneth (Example 1.84).
- $\mathrm{SU}(3)$. The fibration $\mathrm{SU}(2) \to \mathrm{SU}(3) \to S^5$ gives $\pi_3(\mathrm{SU}(3)) = \pi_3(\mathrm{SU}(2)) = \mathbb{Z}$; its cohomology is an exterior algebra on generators of degrees 3 and 5, so the Poincaré polynomial is $(1 + t^3)(1 + t^5)$ and $\chi = 0$, as for every compact Lie group of positive dimension.

One space, every tool: the chapter on S^2

The best test of a toolbox is to use every tool on the same object. We do it for the two-sphere, the space on which the chapter's ideas were most often illustrated.

Worked example 1.122: Everything about S^2, in one place

Homotopy. $\pi_1(S^2) = 0$ because a loop can be moved off a point and then shrunk in $S^2 \setminus \{\mathrm{pt}\} \cong \mathbb{R}^2$. $\pi_2(S^2) = \mathbb{Z}$ by the degree. $\pi_3(S^2) = \mathbb{Z}$ by the Hopf fibration:

$$\pi_3(S^2) \overset{(1)}{=} \pi_3(S^3) \overset{(2)}{=} \mathbb{Z}.$$

(1) The exact sequence of $S^1 \to S^3 \to S^2$ reads $\pi_3(S^1) \to \pi_3(S^3) \to \pi_3(S^2) \to \pi_2(S^1)$, and both ends vanish because the universal cover of S^1 is the contractible line $\mathbb{R}$ (Example 1.100).

(2) The degree of maps $S^3 \to S^3$.

Homology from a triangulation. The hollow tetrahedron (Example 1.70) has 4 vertices, 6 edges and 4 faces, and its boundary matrices give $H_0 = \mathbb{Z}$, $H_1 = 0$, $H_2 = \mathbb{Z}$. Its Euler characteristic

is

$$\begin{aligned}\chi(S^2) &\overset{(1)}{=} N_0 - N_1 + N_2 \\ &= 4 - 6 + 4 = 2 \\ &\overset{(2)}{=} b_0 - b_1 + b_2 = 1 - 0 + 1.\end{aligned}$$

(1) The Euler characteristic of the simplicial chain complex is the alternating count of simplices (Example 1.85).
(2) The Euler–Poincaré formula: the same alternating sum computed from the homology, since rank arithmetic cancels in pairs.

De Rham. The area form $\omega = \sin\theta\,\mathrm{d}\theta \wedge \mathrm{d}\phi$ is closed (it is a top form) and is not exact, because

$$\begin{aligned}\int_{S^2} \omega &\overset{(1)}{=} \int_0^{2\pi}\int_0^{\pi} \sin\theta\,\mathrm{d}\theta\,\mathrm{d}\phi = 4\pi \\ &\overset{(2)}{\neq} 0 = \int_{S^2} \mathrm{d}\alpha.\end{aligned}$$

(1) Direct integration in spherical coordinates.
(2) Stokes: $\int_{S^2} \mathrm{d}\alpha = \int_{\partial S^2} \alpha = 0$ since $\partial S^2 = \varnothing$.

So $H^2_{\mathrm{dR}}(S^2) = \mathbb{R}$, spanned by $[\omega/4\pi]$; Mayer–Vietoris with two discs gives the same answer (Example 1.79), as the de Rham theorem demands.

Bundles. By clutching, U(1) bundles on S^2 are classified by $\pi_1(\mathrm{U}(1)) = \mathbb{Z}$: the transition function on the equator is $e^{in\phi}$ and the Chern number is $\int_{S^2} \frac{i}{2\pi} F = n$ (Example 1.91). The case $n = 1$ is the Hopf bundle. The tangent bundle, viewed as a complex line bundle, has transition function $-1/z^2$, which winds twice, and its Chern number equals the Euler characteristic, as Gauss–Bonnet (Chapter T2) confirms for the round sphere of radius r:

$$\begin{aligned}\frac{1}{2\pi}\int_{S^2} K\,\mathrm{d}A &\overset{(1)}{=} \frac{1}{2\pi}\cdot\frac{1}{r^2}\cdot 4\pi r^2 \\ &= 2 = \chi(S^2).\end{aligned}$$

(1) The Gaussian curvature of the round sphere is the constant $1/r^2$ and its area is $4\pi r^2$.

Since $\chi \neq 0$, TS^2 has no nowhere-zero section: the hairy ball theorem (§1.5.8).

Smooth structures. In dimension 2 every topological manifold has exactly one smooth structure (Radó; §1.6). So for S^2 the top three rungs of Figure 1.66 coincide, and the coarse invariants above say everything: a closed simply connected surface is S^2.

The formulas of the chapter

Key idea: Formulas to know by heart

winding number / degree: $\deg f = \dfrac{1}{2\pi i}\oint \dfrac{\mathrm{d}f}{f}$, $\deg f = \displaystyle\int_{S^n} f^*\omega \ \left(\int_{S^n}\omega = 1\right)$

homology of a complex: $H^n = \dfrac{\operatorname{Ker} d_n}{\operatorname{Im} d_{n-1}}$, $d_n \circ d_{n-1} = 0$

Euler characteristic:	$\chi = \sum_k (-1)^k N_k = \sum_k (-1)^k b_k, \qquad \chi(\Sigma_g) = 2 - 2g$
products:	$P_{X\times Y}(t) = P_X(t)\,P_Y(t), \qquad \chi(X \times Y) = \chi(X)\chi(Y)$
Poincaré duality:	$b_p(M) = b_{n-p}(M) \quad (M^n$ closed, oriented$)$
Hurewicz:	$H_1(X;\mathbb{Z}) = \pi_1(X)/[\pi_1(X), \pi_1(X)]$
cocycle condition:	$g_{\alpha\beta}\, g_{\beta\gamma} = g_{\alpha\gamma}$ on $O_\alpha \cap O_\beta \cap O_\gamma$
gauge change:	$g_{\alpha\beta} \mapsto h_\alpha^{-1} g_{\alpha\beta} h_\beta, \qquad A \mapsto g^{-1}Ag + g^{-1}\mathrm{d}g$
clutching:	$\{G$-bundles on $S^n\} \leftrightarrow \pi_{n-1}(G) \quad (G$ connected$)$
classifying spaces:	$\{G$-bundles on $M\} \leftrightarrow [M, BG], \qquad \pi_n(BG) = \pi_{n-1}(G)$
Chern numbers:	$c_1 = \left[\frac{i}{2\pi}\operatorname{tr} F\right], \qquad k = \frac{1}{8\pi^2}\int_M \operatorname{tr} F \wedge F$
fibration sequence:	$\cdots \to \pi_n(F) \to \pi_n(E) \to \pi_n(M) \to \pi_{n-1}(F) \to \cdots$
obstructions:	$\mathfrak{o} \in H^{k+1}(M;\pi_k(F)), \qquad$ choices $\in H^k(M;\pi_k(F))$
intersection form:	$q(\alpha,\beta) = \int_M \alpha \wedge \beta, \qquad \sigma = b_+ - b_-$
Rohlin:	smooth, closed, spin 4-manifold $\Rightarrow 16 \mid \sigma$
Donaldson:	smooth, $\pi_1 = 0$, definite $\Rightarrow q \cong \pm I$
Yang–Mills bound:	$S = \|F\|^2 \geq 8\pi^2\,\lvert k\rvert$, equality iff $F = \pm \star F$

And the numbers: $\pi_3(S^2) = \mathbb{Z}$, $\pi_1(\mathrm{SO}(3)) = \mathbb{Z}_2$, $\pi_3(\mathrm{SU}(2)) = \mathbb{Z}$, $\pi_3(TOP/PL) = \mathbb{Z}_2$, $\Theta_7 = \mathbb{Z}_{28}$, $\dim \mathcal{M}_k = 8k-3$ for SU(2) instantons on S^4 or in Donaldson's set-up (simply connected, definite form; $= 5$ for $k = 1$), and $\sigma(E_8) = 8$.

Dictionary: topology and physics

The dictionary below translates each tool of the chapter into the physics it describes. It is the table to consult when a physics paper uses a topological word without explaining it.

Dictionary: The chapter's mathematics and its physical avatars

$[X, Y]$, **homotopy class:** topological sector of a field $\phi : X \to Y$; conserved because finite action forbids jumps

$\pi_0, \pi_1, \pi_2, \pi_3$ **of the vacuum manifold:** domain walls, vortices / flux tubes, monopoles, textures and skyrmions

$\pi_1(\mathrm{SO}(3)) = \mathbb{Z}_2$**:** spin $\frac{1}{2}$: a 2π rotation is not homotopic to the identity, a 4π rotation is

$\pi_3(\mathrm{SU}(2)) = \mathbb{Z}$**:** instanton number, large gauge transformations, θ-vacua, baryon number of skyrmions

$\pi_3(S^2) = \mathbb{Z}$ **(Hopf invariant):** Hopf solitons; linking of flux lines; the Hopf term in $2+1$ dimensions

closed form modulo exact, $H^2(M;\mathbb{Z})$**:** quantised flux: magnetic charge, $\oint F/2\pi \in \mathbb{Z}$

$H^1(M;\mathbb{R})$**:** Aharonov–Bohm phases, flat connections, Wilson lines around non-contractible loops

de Rham theorem, periods: a field strength is determined, up to gauge, by its fluxes through cycles

Betti numbers, harmonic forms: zero modes of the Laplacian on forms

$\chi(M)$**:** index of $\mathrm{d} + \mathrm{d}^\dagger$; Gauss–Bonnet; number of zeros of a vector field, with signs

fibre bundle, transition functions: gauge field configuration defined patch by patch; $g_{\alpha\beta}$ are gauge transformations on overlaps

principal bundle P: the gauge structure itself; G the gauge group

section of an associated bundle: a matter field in the representation ρ

monopole bundle, Hopf fibration: Dirac monopole, Wu–Yang patches, Dirac quantisation

c_1, c_2: magnetic charge, instanton number

w_2: obstruction to having fermions (a spin structure) at all

obstruction theory: "can this field configuration be extended inside?"; the answer is a topological charge

loop space ΩG, path fibration: loop groups, Kac–Moody symmetry, Wilson loops, strings

intersection form, E_8: lattice of electric and magnetic fluxes; θ periodicity; even lattices in string theory

instanton moduli space $\mathcal{M}_k$: collective coordinates of instantons; Donaldson invariants as correlation functions of a topological field theory

smooth structure: a choice hidden in every path integral over metrics; exotic spheres enter global gravitational anomalies

Where Chapter 1 goes next

Figure 1.67 shows how the chapter's five main tools feed the rest of the book. Nothing in Chapter 1 was analysis: there were no differential equations, only their topological shadows. The next chapter supplies the analysis. An elliptic operator acts on sections of vector bundles (§1.5); its kernel is the analytic incarnation of a cohomology group (§1.4); and its index turns out to be a characteristic number (§1.5.12). The simplest case is already in hand: the index of the de Rham operator $\mathrm{d} + \mathrm{d}^\dagger$ is $\chi(M)$.

Watch out: Five confusions to leave behind in Chapter 1

- *Homotopy groups are not homology groups.* They agree in the lowest non-trivial degree (Hurewicz) and then diverge: $\pi_3(S^2) = \mathbb{Z}$ but $H_3(S^2) = 0$. Homology vanishes above the dimension; homotopy does not.
- *H_n and H^n are not simply "isomorphic".* With integer coefficients torsion shifts up a degree: $H_1(\mathbb{RP}^2) = \mathbb{Z}_2$ but $H^1(\mathbb{RP}^2;\mathbb{Z}) = 0$ and $H^2(\mathbb{RP}^2;\mathbb{Z}) = \mathbb{Z}_2$ (Example 1.76). Only with field coefficients are they dual vector spaces.
- *De Rham cohomology is blind to torsion.* It computes $H^*(M;\mathbb{R})$, which forgets every $\mathbb{Z}_k$. Spin structures, w_2 and the Kirby–Siebenmann class all live in torsion, so they are invisible to forms.
- *The fibres of a fibration need not be homeomorphic*, only homotopy equivalent; those of a bundle are homeomorphic to F.
- *Homeomorphic does not mean diffeomorphic.* In dimension 7 there are 28 smooth spheres; on $\mathbb{R}^4$ uncountably many smooth structures. Physics done "on a manifold" silently picks one.

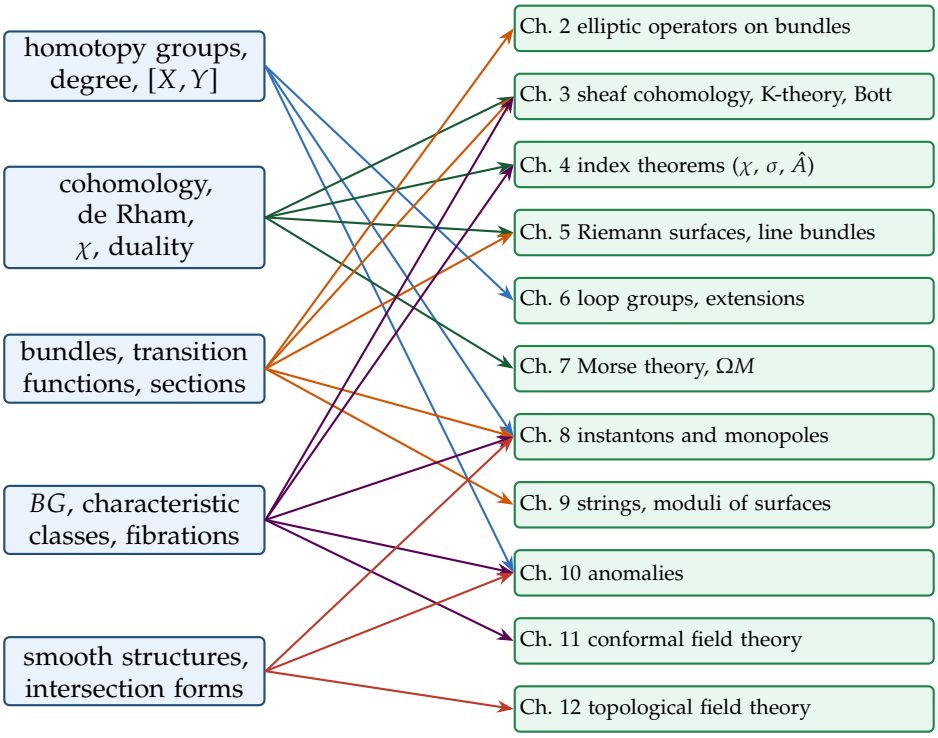


Figure 1.67: How Chapter 1 feeds the book. Each tool (left) is drawn with its own colour of arrow, so one can follow, for example, the red arrows of smooth structures to instantons (Donaldson), anomalies (exotic spheres and global gravitational anomalies) and topological field theory (Donaldson invariants). Cohomology and bundles are used almost everywhere; the arrows show only where they are the main actors.

Self-test

Try it: Ten questions on Chapter 1

(1) Compute χ of the octahedron and of a triangulated torus with 9 vertices.

(2) Why is every map $S^2 \to T^2$ null-homotopic?

(3) What is $[T^2, S^1]$, and what are the classes' representatives?

(4) Use Mayer–Vietoris to show $H^1_{\mathrm{dR}}(S^1 \vee S^1) = \mathbb{R}^2$ (cover by two thickened circles).

(5) How many U(1) bundles are there on S^2? On S^3? On S^4 with structure group SU(2)?

(6) Does T^2 admit a nowhere-vanishing vector field? Does Σ_2?

(7) How many spin structures does Σ_g have?

(8) Check that the exact sequence of $\mathrm{SO}(2) \to \mathrm{SO}(3) \to S^2$ is consistent with $\pi_2(S^2) = \mathbb{Z}$, and find the image of the generator of $\pi_2(S^2)$ in $\pi_1(\mathrm{SO}(2))$.

(9) Is $S^2 \times S^2$ spin? What does Rohlin say about it?

(10) Which of the forms I_3, E_8, $E_8 \oplus E_8$, $H \oplus H$ are intersection forms of smooth closed simply connected 4-manifolds?

Answers.

(1) Octahedron: $6-12+8=2$. Torus: 9 vertices, 27 edges, 18 triangles, $\chi = 9-27+18 = 0$.

(2) The universal cover of T^2 is the contractible $\mathbb{R}^2$, and S^2 is simply connected, so every map lifts to $\mathbb{R}^2$ and contracts there.

(3) $\mathbb{Z}^2$, represented by $e^{i(m\theta+n\phi)}$, detected by the two winding numbers (Example 1.60).

(4) Each piece retracts to a circle and the overlap (a neighbourhood of the wedge point) is contractible, so the sequence is $0 \to H^0 \to \mathbb{R}^2 \to \mathbb{R} \xrightarrow{\mathrm{d}^*} H^1 \to \mathbb{R}^2 \to 0$. The difference map $\mathbb{R}^2 \to \mathbb{R}$ is onto, so $\mathrm{d}^* = 0$ and $H^1 \cong \mathbb{R}^2$.

(5) On S^2: $\pi_1(\mathrm{U}(1)) = \mathbb{Z}$, one for each monopole charge. On S^3: $\pi_2(\mathrm{U}(1)) = 0$, only the trivial one. On S^4: $\pi_3(\mathrm{SU}(2)) = \mathbb{Z}$, one for each instanton number.

(6) Yes: $\chi(T^2) = 0$ and ∂_θ works. No: $\chi(\Sigma_2) = -2 \neq 0$.

(7) $|H^1(\Sigma_g; \mathbb{Z}_2)| = 2^{2g}$ (orientable surfaces have $w_2 = 0$).

(8) $\pi_2(\mathrm{SO}(3)) = 0 \to \pi_2(S^2) \to \pi_1(\mathrm{SO}(2)) = \mathbb{Z} \to \pi_1(\mathrm{SO}(3)) = \mathbb{Z}_2 \to \pi_1(S^2) = 0$. The map $\mathbb{Z} \to \mathbb{Z}_2$ is onto, so its kernel $2\mathbb{Z} \cong \mathbb{Z}$ is the image of $\pi_2(S^2)$; and ∂ is injective because $\pi_2(\mathrm{SO}(3)) = 0$, so $\pi_2(S^2) \cong 2\mathbb{Z} \cong \mathbb{Z}$ (a degree-one class maps to twice the generator: the frame of a sphere turns twice around the equator, as for TS^2).

(9) Yes: $q = H$ is even and $S^2 \times S^2$ is simply connected, so $w_2 = 0$. Rohlin requires $16 \mid \sigma$, and indeed $\sigma = 0$.

(10) I_3 ($\#^3\mathbb{CP}^2$) and $H \oplus H$ ($\#^2 S^2 \times S^2$) yes. E_8 no (Rohlin). $E_8 \oplus E_8$ no (Donaldson: it is definite but not diagonal).

With these tools in hand we can turn from topology to analysis. The next chapter asks what makes a differential operator "elliptic", and why elliptic operators on compact manifolds have finite-dimensional kernels whose difference of dimensions, the index, is a topological invariant of exactly the kind this chapter taught us to compute.

Chapter 2

Elliptic Operators

Where this chapter is going

Physics is full of differential operators: the Laplacian of electrostatics, the d'Alembertian of wave propagation, the heat operator of diffusion, the Schrödinger operator, the Dirac operator. They do not all behave alike. The Laplacian is "kind": feed it a smooth source and every solution is smooth, and on a compact space it has finitely many zero modes. The wave operator is "unkind": a kink in the initial data travels along light rays forever and never smooths out. This chapter isolates the property that makes the Laplacian kind. It is called **ellipticity**, and it can be read off from a single polynomial, the *principal symbol*, obtained by Fourier transforming the highest derivatives. Once we know how to recognise ellipticity, we can carry it to operators acting on sections of vector bundles, such as the Hodge Laplacian on p-forms and the Dirac operator. That leads to Hodge's theorem: the zero modes of the Laplacian on p-forms count the p-dimensional holes of the manifold. The chapter then widens the class of operators to pseudo-differential operators and Sobolev spaces (part 2b), which is the analytic engine behind the index theorem of Chapter 4.

In this part (Nash §1–§3) we build the core ideas: the symbol, ellipticity, hypo-ellipticity, elliptic regularity, the bundle version of ellipticity, the symbol calculus of d, d^* and Δ_p, and the Hodge theorem. Figure 2.1 shows how these ideas feed into each other and into the physics used later in the book.

Conventions for this part. We work on $\mathbb{R}^n$ with coordinates $x = (x_1, \dots, x_n)$ and on smooth manifolds M of dimension n. Unless we say otherwise, everything is Euclidean (signature $+\cdots+$). When we need Minkowski space we write $(t, x_1, \dots, x_n)$ and use the wave operator $\partial_t^2 - \nabla^2$, so the metric is $(+, -, \dots, -)$ in the "∂_t^2 first" sense. We write p (Nash's notation in §1–§2) or k (Nash's notation in §3) for the Fourier variable. Our Fourier convention is fixed in Section 2.1.2; it differs from Nash's footnote by a sign, and we explain why in the "Watch out" box on page 728.

2.1 The meaning of ellipticity

Here is the question this section answers. We have three famous second-order operators in two variables:

$$-\Big(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\Big) \text{ (Laplace)}, \qquad \frac{\partial^2}{\partial t^2} - \frac{\partial^2}{\partial x^2} \text{ (wave)}, \qquad \frac{\partial}{\partial t} - \frac{\partial^2}{\partial x^2} \text{ (heat)}.$$

They look almost the same: each is a combination of second derivatives, the first two differ only by a sign. Yet their solutions behave completely differently. Solutions of Laplace's equation are

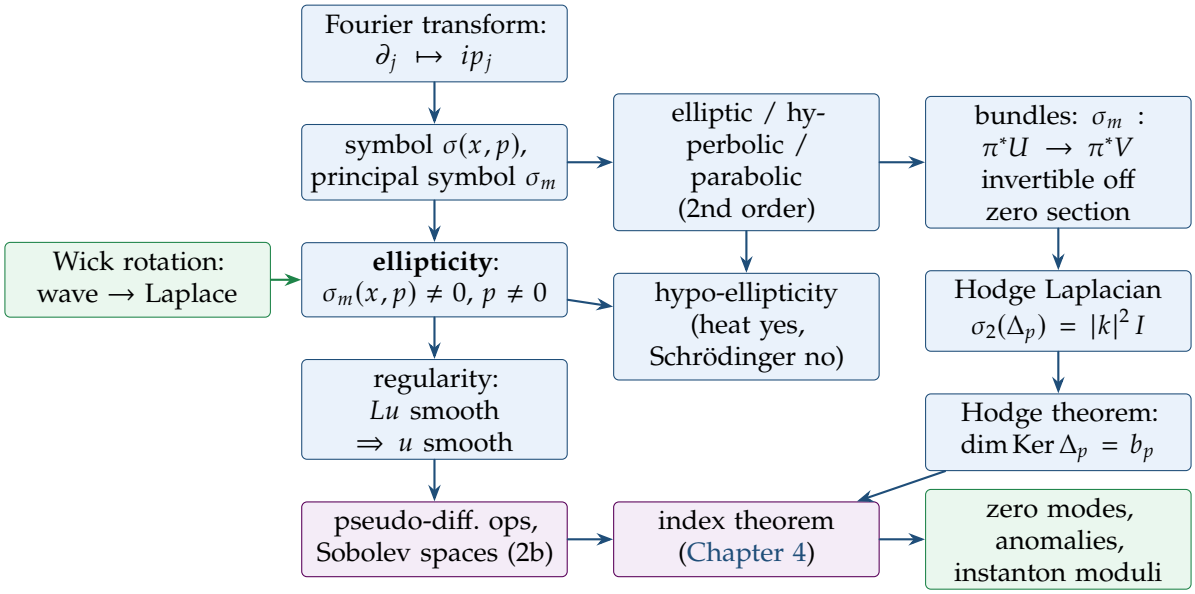


Figure 2.1: The logic of Chapter 2. Read it top to bottom. The Fourier transform turns derivatives into polynomials. The top-degree part of the polynomial (the principal symbol) decides whether the operator is elliptic. Ellipticity gives smoothness of solutions, survives the passage to vector bundles, and powers Hodge theory. Hodge theory and the pseudo-differential machinery (purple, part 2b) together feed the index theorem, which is what physics uses (green) to count zero modes.

as smooth as can be. Solutions of the wave equation can carry a kink forever. Solutions of the heat equation become smooth instantly, but only forwards in time. *What feature of the operator decides this?* Nash's answer, which we develop here, is: *Fourier transform the operator, keep only the highest-order terms, and look at where the resulting polynomial vanishes.*

Before we start, it helps to know why a book on topology and physics cares.

2.1.1 Why ellipticity matters for topology and physics

Take a compact Riemannian manifold M (a sphere, a torus, a Calabi–Yau space, Euclidean space-time compactified to S^4). On it there is a Laplacian Δ_p acting on p-forms. We will define it carefully in Section 2.3.7; for now think of it as the operator $-\nabla^2$ applied component by component. The celebrated **Hodge–de Rham theorem**, which we reach in Section 2.3.12, says

$$\dim \operatorname{Ker} \Delta_p \;=\; \dim H^p(M;\mathbb{R}) \;=\; b_p, \tag{2.1}$$

where $H^p(M;\mathbb{R})$ is the p-th de Rham cohomology group (closed p-forms modulo exact ones, see Chapter T2) and b_p is the p-th Betti number, the number of independent p-dimensional "holes" of M.

Read (2.1) as a physicist. The left side counts **zero modes**: solutions of $\Delta_p \omega = 0$, the static, source-free field configurations. The right side is topology, so it does not change when we deform the metric. Something analytic (a count of solutions of a PDE) equals something topological (a count of holes). Three things are needed to make sense of the left side:

1. the kernel must be *finite-dimensional*, or the count means nothing;
2. the solutions must be *smooth*, so that they really are differential forms and represent de Rham classes;
3. the count must be *stable* under deformations.

All three are consequences of ellipticity. The index theorem of Chapter 4 generalises (2.1) from Δ_p to any elliptic operator, and in physics it counts fermion zero modes in an instanton background, computes anomalies, and gives the dimension of instanton moduli spaces. So ellipticity is the entry ticket to everything that follows.

Physical picture: Elliptic = "no preferred direction of propagation"

An elliptic equation describes *equilibrium*: a soap film, an electrostatic potential, a static field. Every point feels every other point, and there is no direction along which disturbances run away. A hyperbolic equation describes *propagation*: signals move along special directions (light rays), and a sharp front stays sharp. A parabolic equation describes *diffusion*: there is a preferred time direction, and sharp features melt away as time goes forwards.

2.1.2 The Fourier transform: derivatives become multiplication

The one tool behind everything in this chapter is the Fourier transform. We recall it now and fix conventions.

Working definition 2.1 (Fourier transform)

For a function $u(x)$ on $\mathbb{R}^n$ that decays fast enough [fast enough for the integral below to converge, e.g. $\int |u| \, \mathrm{d}^n x < \infty$], its **Fourier transform** is

$$\hat{u}(p) \;=\; \mathcal{F}(u)(p) \;=\; \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} e^{-ip\cdot x}\, u(x)\, \mathrm{d}^n x, \tag{2.2}$$

and u is recovered by the **inverse transform**

$$u(x) \;=\; \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} e^{ip\cdot x}\, \hat{u}(p)\, \mathrm{d}^n p. \tag{2.3}$$

Here $p \cdot x = p_1 x_1 + \cdots + p_n x_n$. *Operationally*: u is a superposition of plane waves $e^{ip\cdot x}$, and $\hat{u}(p)$ is the amplitude of the wave with wave-vector p.

The key property is what a derivative does. Differentiate the inverse transform (2.3) under the integral sign:

$$\begin{aligned} \frac{\partial u}{\partial x_j}(x) &\overset{(1)}{=} \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} \frac{\partial}{\partial x_j}\left(e^{ip\cdot x}\right) \hat{u}(p)\, \mathrm{d}^n p \\ &\overset{(2)}{=} \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} e^{ip\cdot x}\, \big(i p_j\, \hat{u}(p)\big)\, \mathrm{d}^n p. \end{aligned} \tag{2.4}$$

(1) Only the plane wave depends on x; $\hat{u}(p)$ does not.
(2) $\partial_{x_j} e^{ip\cdot x} = i p_j\, e^{ip\cdot x}$, since $p \cdot x$ contains x_j only in the term $p_j x_j$.

Comparing with (2.3), the function $\partial_j u$ has Fourier transform $i p_j \hat{u}(p)$. So

$$\boxed{\frac{\partial}{\partial x_j} \;\longleftrightarrow\; \text{multiplication by } i p_j} \qquad \text{and} \qquad -i\frac{\partial}{\partial x_j} \;\longleftrightarrow\; \text{multiplication by } p_j. \tag{2.5}$$

The second form is why Nash (and quantum mechanics) like the operator $D_j := -i\,\partial/\partial x_j$: it is the momentum operator (with $\hbar = 1$), and on a plane wave it just reads off the momentum, $D_j e^{ip\cdot x} = p_j e^{ip\cdot x}$.

Applying (2.5) twice, a second derivative becomes $\partial_j \partial_k \leftrightarrow (ip_j)(ip_k) = -p_j p_k$. Any linear differential operator with constant coefficients therefore becomes *multiplication by a polynomial*. Solving a constant-coefficient PDE becomes division by that polynomial. Everything difficult about the PDE is hidden in *where the polynomial vanishes*, because that is where we cannot divide.

Watch out: Nash's Fourier convention

Nash's footnote on p. 28 defines $\mathcal{F}(f) = (2\pi)^{-n} \int e^{+ipx} f \, \mathrm{d}^n x$ (note: the full $(2\pi)^{-n}$, not our symmetric $(2\pi)^{-n/2}$), and in (2.21) he writes the inverse transform with e^{-ipx} and *no* extra 2π factor. So Nash's pair is asymmetric: all of $(2\pi)^{-n}$ sits on the forward transform, none on the inverse. Ours (2.2)–(2.3) splits it symmetrically, $(2\pi)^{-n/2}$ each way, and uses the opposite sign convention for which map is "forward". With Nash's sign, $-i\partial_j$ becomes multiplication by $-p_j$, not p_j, so his formula (2.21) is off by a factor $(-1)^{|\alpha|}$ in each term compared with ours. None of this matters for ellipticity, which only asks whether the top-order polynomial *vanishes*: a sign or a real constant prefactor never changes a zero set. It also never matters for second-order operators, where $(-1)^2 = 1$. We use (2.2)–(2.3) throughout, which makes $D_j \leftrightarrow p_j$ exactly and Nash's symbol formulas correct as written.

Two more facts we will use repeatedly:

- **The delta function is the flat spectrum.** Formally $\hat{\delta}(p) = (2\pi)^{-n/2}$, so
$$\delta(x) \;=\; \frac{1}{(2\pi)^n} \int_{\mathbb{R}^n} e^{ip\cdot x} \, \mathrm{d}^n p. \tag{2.6}$$
A point source contains every wavelength with equal weight.
- **Smoothness is decay in p.** If $|\hat{u}(p)| \le C(1+|p|)^{-N}$ with $N > n + k$, then we may differentiate (2.3) k times under the integral: each derivative brings down a factor of size $|p|$, and the integral of $|p|^k (1+|p|)^{-N}$ still converges. So *fast decay of $\hat{u}$ means many derivatives of u*; decay faster than every power means u is C^∞. Conversely a jump or a kink in u produces a slowly decaying tail in $\hat{u}$ (a step gives $\hat{u} \sim 1/p$, a kink gives $\sim 1/p^2$).

Worked example 2.1: Fourier transforming the Laplacian

Nash multiplies the Laplacian by -1 so that it is a *positive* operator:
$$\Delta \;=\; -\frac{\partial^2}{\partial x_1^2} - \cdots - \frac{\partial^2}{\partial x_n^2}. \tag{2.7}$$
Apply it to a plane wave:
$$\Delta\, e^{ip\cdot x} \overset{(1)}{=} -\sum_{j=1}^{n} \frac{\partial^2}{\partial x_j^2} e^{ip\cdot x} \overset{(2)}{=} -\sum_{j=1}^{n} (ip_j)^2 \, e^{ip\cdot x} \overset{(3)}{=} \left(p_1^2 + \cdots + p_n^2\right) e^{ip\cdot x}.$$
(1) Definition (2.7).
(2) Each ∂_j brings down ip_j, by (2.4).
(3) $(ip_j)^2 = -p_j^2$, and the minus signs cancel.

So under Fourier transformation Δ is multiplication by
$$\sigma(\Delta)(p) \;=\; p_1^2 + \cdots + p_n^2 \;=\; |p|^2, \tag{2.8}$$
Nash's (2.3). This is a quadratic form in p. It is *positive*, which is the point of Nash's sign:

for a real function u that decays at infinity,

$$\int_{\mathbb{R}^n} u\,\Delta u\,\mathrm{d}^n x \overset{(1)}{=} -\sum_j \int_{\mathbb{R}^n} u\,\partial_j^2 u\,\mathrm{d}^n x \overset{(2)}{=} \sum_j \int_{\mathbb{R}^n} (\partial_j u)^2\,\mathrm{d}^n x \;\geq\; 0.$$

(1) Definition (2.7).
(2) Integrate by parts in x_j; the boundary term vanishes because u decays.

In quantum mechanics Δ is (twice the mass times) the kinetic energy operator $\hat{\vec{p}}^{\,2}$, and positivity is the statement that kinetic energy is never negative.

2.1.3 Quadratic forms, level sets and the word "elliptic"

Where does the word *elliptic* come from? Nash's answer is geometric. Set the symbol (2.8) equal to a positive constant c: the set $|p|^2 = c$ is a sphere of radius $\sqrt{c}$ in p-space. Now stretch the coordinates. Replace the Laplacian by

$$L \;=\; -\sum_{i=1}^{n} a_i \frac{\partial^2}{\partial x_i^2}, \qquad a_i \in \mathbb{R} \text{ constants}, \tag{2.9}$$

Nash's (2.4). Exactly as in Example 2.1, Fourier transformation turns L into multiplication by

$$\sigma(L)(p) \;=\; \sum_{i=1}^{n} a_i\, p_i^2, \tag{2.10}$$

Nash's (2.5), and the level set $\sigma(L) = c$ is

$$a_1 p_1^2 + \cdots + a_n p_n^2 \;=\; c, \tag{2.11}$$

Nash's (2.6). If every $a_i > 0$ this is an *ellipsoid* with semi-axes $\sqrt{c/a_i}$. That is the origin of the name.

A word about **signature**. A real quadratic form $\sigma(p) = \sum_{ij} a_{ij} p_i p_j$ (with a_{ij} symmetric) can be diagonalised by a rotation of p-space; the diagonal entries are the eigenvalues of the matrix (a_{ij}). If n_+ of them are positive, n_- negative and n_0 zero, the signature is often quoted as the single number $n_+ - n_-$, and that is how Nash uses the word. "Signature $+n$" means all n eigenvalues are positive: the form is **positive definite**. "Signature $n-2$" with $\det \neq 0$ means $n_+ = n-1$, $n_- = 1$: the form is Lorentzian.

Working definition 2.2 (elliptic, constant-coefficient second-order case)

The operator $L = -\sum_i a_i \partial_i^2$ is **elliptic** if its Fourier polynomial [what L becomes after Fourier transform, with each ∂_j replaced by ip_j] $\sigma(L) = \sum_i a_i p_i^2$ is positive definite (signature $+n$), or negative definite (signature $-n$; then $-L$ is positive definite). *Test:* all a_i nonzero and of the same sign. *Equivalent test:* $\sigma(L)(p) \neq 0$ for every real $p \neq 0$; the level sets $\sigma = c$ are ellipsoids, which are bounded.

The last reformulation is the one that generalises. An ellipsoid is *bounded*: if $\sigma(L)(p) = c$ then $|p|$ is at most $\sqrt{c/\min a_i}$. Put differently, $|\sigma(L)(p)|$ grows like $|p|^2$ in *every* direction. For the wave or heat operators the level sets are hyperbolae or parabolae, which run off to infinity: there are directions in p-space along which the symbol does *not* grow. Figure 2.2 shows the three pictures side by side.

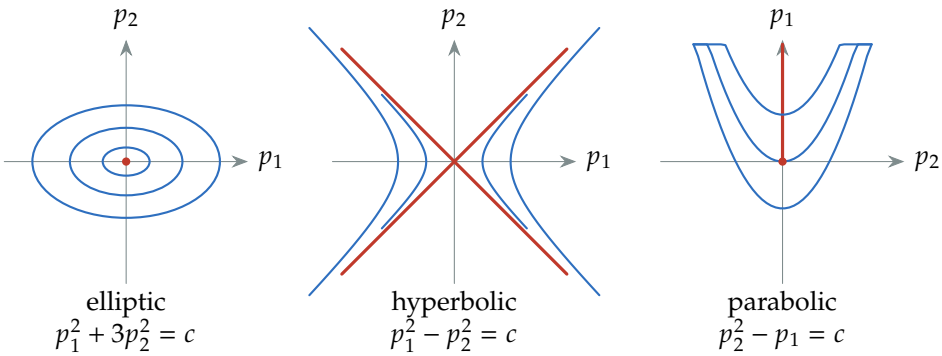


Figure 2.2: Level sets $\sigma(p) = c$ of the Fourier polynomial for the three archetypes (blue), and the directions where the *top-order* part of σ vanishes (red). Elliptic: closed ellipses, the top-order part vanishes only at $p = 0$. Hyperbolic (wave operator): the curves run to infinity along the red light cone $p_1 = \pm p_2$, where σ does not grow. Parabolic (heat or Schrödinger operator): the parabolas open along the p_1-axis. There the second-order part p_2^2 vanishes, and σ grows only linearly.

2.1.4 The payoff: smooth sources give smooth solutions

Why is the ellipsoid picture so important? Because of the following property, which is the main theme of the chapter.

Key idea: Elliptic regularity (Nash's first statement, (2.7))

If L is elliptic and f is a C^∞ function, then every solution u of

$$Lu = f \tag{2.12}$$

is automatically C^∞. More precisely, and this is the useful form, if f is smooth on an open set Ω, then u is smooth on Ω, whatever f does elsewhere.

This is surprising at first. The equation $Lu = f$ only constrains a particular *combination* of second derivatives of u. Why should it control *all* derivatives of all orders? For example, $-\partial_x^2 u - \partial_y^2 u = 0$ says nothing directly about $\partial_x \partial_y u$. The answer is that for an elliptic operator the combination is "fat enough" in Fourier space: it controls $|p|^2\,\hat{u}$, hence every $p_i p_j \hat{u}$. Compare the wave operator: the combination $p_1^2 - p_2^2$ vanishes on the light cone, and $\hat{u}$ can be as large as it likes there without Lu noticing. We now make this precise with Green's functions, following Nash.

Green's functions

Working definition 2.3 (Green's function / fundamental solution)

A **Green's function** (mathematicians say **fundamental solution**) of a constant-coefficient operator L on $\mathbb{R}^n$ is a function or distribution [a generalised function such as δ, known only through its integrals against smooth functions] G with

$$L\,G(x) = \delta(x). \tag{2.13}$$

Operationally: G is the field produced by a unit point source at the origin. *How to compute it*: Fourier transform. Since L becomes multiplication by $\sigma(L)(p)$ and δ has flat spectrum

[$\hat{\delta}(p) = (2\pi)^{-n/2}$, the same for every p] (2.6),

$$G(x) \;=\; \frac{1}{(2\pi)^n}\int_{\mathbb{R}^n} \frac{e^{ip\cdot x}}{\sigma(L)(p)}\,\mathrm{d}^n p, \tag{2.14}$$

which is Nash's (2.9) (Nash drops the $(2\pi)^{-n}$ and uses the opposite sign in the exponent; neither affects the argument).

Once we have G, we can solve $Lu = f$ for any source by superposing point sources:

$$u(x) \;=\; (G * f)(x) \;=\; \int_{\mathbb{R}^n} G(x-y)\,f(y)\,\mathrm{d}^n y, \tag{2.15}$$

Nash's (2.8). Let us check that this works:

$$Lu(x) \overset{(1)}{=} \int_{\mathbb{R}^n} \big(L_x G(x-y)\big)\,f(y)\,\mathrm{d}^n y \overset{(2)}{=} \int_{\mathbb{R}^n} \delta(x-y)\,f(y)\,\mathrm{d}^n y \overset{(3)}{=} f(x).$$

(1) L acts on x; only $G(x-y)$ depends on x.
(2) L has constant coefficients, so $L_x G(x-y) = (LG)(x-y) = \delta(x-y)$ by (2.13).
(3) The defining property of the delta function.

This is the principle of superposition familiar from electrostatics: the potential of a charge distribution ρ is the sum of Coulomb potentials of its point charges.

Worked example 2.2: The Coulomb potential as the Green's function of Δ on $\mathbb{R}^3$

We compute (2.14) for $L = \Delta$ in $\mathbb{R}^3$, where $\sigma = |p|^2$. Put $r = |x|$ and use spherical polar coordinates in p-space with the polar axis along x, so $p\cdot x = pr\cos\theta$ with $p = |p|$.

$$\begin{aligned} G(x) &\overset{(1)}{=} \frac{1}{(2\pi)^3}\int_0^\infty\!\!\int_0^\pi\!\!\int_0^{2\pi} \frac{e^{ipr\cos\theta}}{p^2}\,p^2\sin\theta\,\mathrm{d}\phi\,\mathrm{d}\theta\,\mathrm{d}p \\ &\overset{(2)}{=} \frac{1}{(2\pi)^2}\int_0^\infty\!\!\int_{-1}^1 e^{iprs}\,\mathrm{d}s\,\mathrm{d}p \overset{(3)}{=} \frac{1}{(2\pi)^2}\int_0^\infty \frac{2\sin(pr)}{pr}\,\mathrm{d}p \\ &\overset{(4)}{=} \frac{1}{(2\pi)^2}\,\frac{2}{r}\,\frac{\pi}{2} \;=\; \frac{1}{4\pi r}. \end{aligned}$$

(1) Formula (2.14) in spherical coordinates; the volume element is $p^2\sin\theta\,\mathrm{d}p\,\mathrm{d}\theta\,\mathrm{d}\phi$.
(2) The p^2 cancels, the ϕ integral gives 2π, and we substitute $s = \cos\theta$, $\mathrm{d}s = -\sin\theta\,\mathrm{d}\theta$.
(3) $\int_{-1}^1 e^{iprs}\mathrm{d}s = (e^{ipr} - e^{-ipr})/(ipr) = 2\sin(pr)/(pr)$.
(4) Substitute $q = pr$ and use the Dirichlet integral $\int_0^\infty (\sin q/q)\,\mathrm{d}q = \pi/2$.

So the Green's function of the positive Laplacian in $\mathbb{R}^3$ is the Coulomb potential $G = 1/(4\pi r)$. Notice two things. It is singular only at $r = 0$, where the source sits. Away from $r = 0$ it is C^∞, indeed real analytic.

Worked example 2.3: The same result from Gauss's law

We can confirm $\Delta(1/4\pi r) = \delta$ without Fourier transforms, using only Stokes' theorem in its Gauss form (Chapter T2). Away from the origin, with $G = 1/(4\pi r)$,

$$\Delta G \overset{(1)}{=} -\frac{1}{r^2}\frac{\partial}{\partial r}\Big(r^2\frac{\partial G}{\partial r}\Big) \overset{(2)}{=} -\frac{1}{r^2}\frac{\partial}{\partial r}\Big(r^2\cdot\frac{-1}{4\pi r^2}\Big) \overset{(3)}{=} 0 \qquad (r \neq 0).$$

(1) For a radial function, $\nabla^2 G = r^{-2}\partial_r(r^2\partial_r G)$, and $\Delta = -\nabla^2$.
(2) $\partial_r(1/4\pi r) = -1/(4\pi r^2)$.
(3) $r^2 \cdot (-1/4\pi r^2) = -1/4\pi$ is constant, so its r-derivative is zero.

So ΔG is concentrated at the origin. Its total strength is the integral over a ball B_R of radius R:

$$\int_{B_R} \Delta G \, \mathrm{d}^3x \overset{(1)}{=} -\int_{B_R} \nabla\cdot\nabla G \, \mathrm{d}^3x \overset{(2)}{=} -\oint_{\partial B_R} \nabla G \cdot \mathrm{d}\vec{A} \overset{(3)}{=} -\Big(\frac{-1}{4\pi R^2}\Big) 4\pi R^2 \;=\; 1.$$

(1) $\Delta = -\nabla\cdot\nabla$.
(2) Gauss's divergence theorem (Stokes for the 2-form $\star \mathrm{d}G$).
(3) $\nabla G = -\hat{r}/(4\pi r^2)$ is radial and of constant size on the sphere of area $4\pi R^2$.

A distribution concentrated at a point with total integral 1 is δ. Physically: the electric field of a unit charge has unit flux through any sphere around it.

For later use, here are the Green's functions of Δ in all dimensions. Write $\omega_{n-1} = 2\pi^{n/2}/\Gamma(n/2)$ for the area of the unit sphere S^{n-1} (so $\omega_1 = 2\pi$, $\omega_2 = 4\pi$). The same Gauss's-law argument, with flux $-\partial_r G \cdot \omega_{n-1} r^{n-1} = 1$, gives

$$G_n(x) \;=\; \begin{cases} -\dfrac{1}{2\pi}\ln|x|\,, & n = 2, \\[2ex] \dfrac{1}{(n-2)\,\omega_{n-1}\,|x|^{n-2}}, & n \geq 3. \end{cases} \tag{2.16}$$

Every one of them is singular only at $x = 0$ and real analytic elsewhere.

Why the Green's function of an elliptic operator is smooth away from the source

Now we can see the mechanism behind elliptic regularity. It has two steps.

Step 1: G is smooth for $x \neq 0$. This is where ellipticity enters. We want to show that $G(x)$ in (2.14) has derivatives of all orders away from the origin. Recall the principle "smoothness is decay in p". The function $1/\sigma(p)$ only decays like $|p|^{-2}$, which is not enough to differentiate (2.14) even once in $n \geq 2$ dimensions. That is fine: G really is singular at $x = 0$. The trick is to multiply by powers of x, which kill the singularity at 0, and to see that this improves the decay in p. Since $\partial_{p_j} e^{ip\cdot x} = i x_j e^{ip\cdot x}$, integration by parts gives

$$\begin{aligned} x_j\, G(x) &\overset{(1)}{=} \frac{1}{(2\pi)^n}\int_{\mathbb{R}^n}\Big(-i\frac{\partial}{\partial p_j}e^{ip\cdot x}\Big)\frac{1}{\sigma(p)}\,\mathrm{d}^n p \\ &\overset{(2)}{=} \frac{1}{(2\pi)^n}\int_{\mathbb{R}^n} e^{ip\cdot x}\, i\frac{\partial}{\partial p_j}\Big(\frac{1}{\sigma(p)}\Big)\,\mathrm{d}^n p. \end{aligned} \tag{2.17}$$

(1) $x_j e^{ip\cdot x} = -i\,\partial_{p_j} e^{ip\cdot x}$.
(2) Integrate by parts in p_j (ignoring for the moment the point $p = 0$; see below). The boundary terms at $|p| \to \infty$ vanish because $1/\sigma \to 0$.

So *multiplying G by x_j is the same as differentiating $1/\sigma$ with respect to p_j*. Now use ellipticity. Because σ is a positive definite quadratic form, it is bounded below on the unit sphere, $\sigma(p) \geq \lambda\,|p|^2$ with $\lambda = \min_i a_i > 0$. Differentiating $1/\sigma$ produces terms like $\partial_j\sigma/\sigma^2$, each of which is homogeneous of degree -3, and in general

$$\Big|\partial_p^\gamma\Big(\frac{1}{\sigma(p)}\Big)\Big| \;\leq\; C_\gamma\,|p|^{-2-|\gamma|}\,, \tag{2.18}$$

where ∂_p^γ is any combination of $|\gamma|$ derivatives. *Each p-derivative improves the decay at infinity by one power.* Therefore, for any number k of x-derivatives, the function $|x|^{2N}\,\partial_x^\beta G$ with $|\beta| = k$ has a Fourier transform bounded by a constant times $|p|^{k-2-2N}$ at large $|p|$, which is integrable once $2N > k - 2 + n$. So $|x|^{2N}\,\partial_x^\beta G$ is a continuous function, and dividing by $|x|^{2N}$ (allowed for $x \neq 0$) shows that $\partial_x^\beta G$ exists and is continuous for $x \neq 0$. Since k was arbitrary, G is C^∞ away from the origin.

There is a technical point we skipped: the extra p-derivatives make the singularity of $1/\sigma$ at $p = 0$ worse. It is handled by splitting $1/\sigma = \chi(p)/\sigma + (1 - \chi(p))/\sigma$, where χ is a smooth bump equal to 1 near $p = 0$. The first piece has compact support in p, so its inverse Fourier transform is automatically smooth in x (for $n \geq 3$, where $1/|p|^2$ is integrable near 0; for $n = 2$ one subtracts a constant, which is where the logarithm comes from). The second piece is smooth everywhere in p, so the decay argument applies to it without any trouble at $p = 0$.

Key idea: Where exactly ellipticity was used

The whole argument rests on (2.18), and that estimate holds only because σ *has no zeros except* $p = 0$. If σ vanished on some ray $p = s\,\hat{p}_0$ going to infinity, then $1/\sigma$ would be singular all along that ray. Differentiating would make it more singular, not more decaying, and the argument would collapse. Nash puts it this way: "had the quadratic form $\sigma(L)$ vanished away from the origin the whole argument would have failed."

Nash phrases Step 1 slightly differently. He rescales $p_i \to \tilde{p}_i = \sqrt{a_i}\,p_i$ so that σ becomes $|\tilde{p}|^2$, passes to spherical polar coordinates in $\tilde{p}$-space so that the only nontrivial integral is over the radius $|\tilde{p}|$, and deforms that radial contour in the complex plane (Cauchy's theorem) to make the integral converge absolutely when $x \neq 0$. That is the same mechanism in different clothing: the integrand has no poles for real $\tilde{p} \neq 0$ precisely because σ vanishes only at the origin.

Step 2: from smooth G to smooth u. Suppose $Lu = f$ and f is smooth on an open set Ω. Fix a point $x_0 \in \Omega$ and a small ball B around it with $\bar{B} \subset \Omega$. The following narrative argument shows that u is smooth near x_0.

1. Choose a smooth cutoff χ equal to 1 on a smaller ball B' around x_0 and 0 outside B. The function χu has compact support, so it equals $G * L(\chi u)$ (a compactly supported solution is determined by its source, because $G*$ inverts L on such functions).

2. Compute the source using the product rule: $L(\chi u) = \chi\,Lu + [L, \chi]u = \chi f + [L, \chi]u$. The commutator $[L, \chi]u$ involves derivatives of χ, so it vanishes on B' and outside B: it lives in the shell $B \setminus B'$.

3. Hence near x_0, $\chi u = G*(\chi f) + G*([L, \chi]u)$. The first term is smooth because χf is smooth with compact support: we may move all derivatives onto χf, $\partial^\beta(G * \chi f) = G * \partial^\beta(\chi f)$. The second term is smooth near x_0 because it only uses $G(x - y)$ with x near x_0 and y in the shell. Then $x - y \neq 0$, and G is smooth there by Step 1.

So $u = \chi u$ is smooth near x_0. Only *two* properties of G were used: it inverts L, and it is smooth off the origin. In the language of distributions this is the statement

$$\operatorname{sing\,supp}(u) \;\subseteq\; \operatorname{sing\,supp}(Lu) \qquad \text{whenever} \qquad \operatorname{sing\,supp}(G) = \{0\}, \tag{2.19}$$

where the **singular support** of a function is the closure of the set of points where it fails to be C^∞.

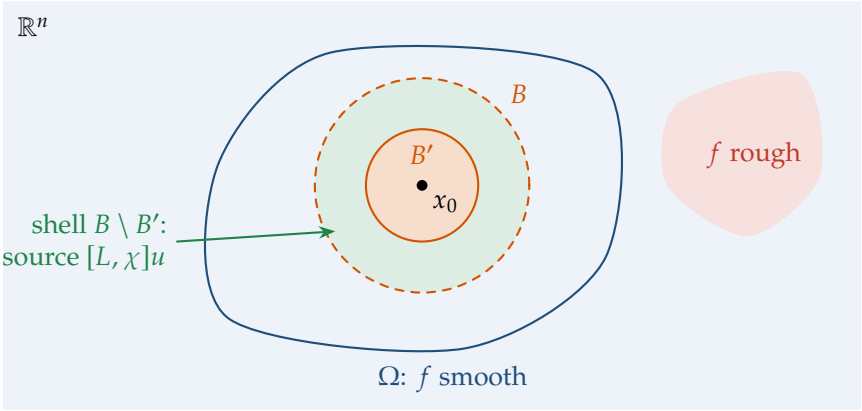


Figure 2.3: The localisation argument of Step 2. The source f is smooth on Ω but may be rough elsewhere (red). Near x_0 we cut u off with χ ($\chi = 1$ on the inner ball B', $\chi = 0$ outside B). The price is a new source $[L, \chi]u$ living only in the shell $B \setminus B'$. Its effect on points near x_0 is felt through G at nonzero separation, where G is smooth. Roughness far away cannot spoil smoothness near x_0.

Weyl's lemma: harmonic means smooth

The simplest instance of elliptic regularity is the case $f = 0$: solutions of Laplace's equation $\Delta u = 0$ (**harmonic functions**) are C^∞, and in fact real analytic. Historically this is **Weyl's lemma**. It says more: even a *distribution* u, a priori just a "generalised function" that need not have any derivatives at all, is a smooth function if $\Delta u = 0$. There is a beautifully physical proof in two dimensions, through the mean value property.

Worked example 2.4: Mean value property of harmonic functions

Let u be harmonic on a region of $\mathbb{R}^2$ containing the closed disc of radius R about x_0. Define the circle average

$$A(r) \;=\; \frac{1}{2\pi}\int_0^{2\pi} u(x_0 + r\cos\phi,\; y_0 + r\sin\phi)\,\mathrm{d}\phi.$$

Then, writing ∂_r for the radial derivative and D_r for the disc of radius r,

$$\frac{\mathrm{d}A}{\mathrm{d}r} \overset{(1)}{=} \frac{1}{2\pi}\int_0^{2\pi} \partial_r u\,\mathrm{d}\phi \overset{(2)}{=} \frac{1}{2\pi r}\oint_{\partial D_r} \nabla u\cdot\hat{n}\,\mathrm{d}s \overset{(3)}{=} \frac{1}{2\pi r}\int_{D_r} \nabla^2 u\,\mathrm{d}^2x \overset{(4)}{=} 0.$$

(1) Differentiate under the integral; ∂_r at fixed ϕ is the radial derivative.
(2) On the circle, $\partial_r u = \nabla u\cdot\hat{n}$ with $\hat{n}$ the outward normal, and arc length is $\mathrm{d}s = r\,\mathrm{d}\phi$.
(3) Divergence theorem in the plane (Stokes).
(4) u is harmonic: $\nabla^2 u = -\Delta u = 0$.

So $A(r)$ is constant, and letting $r \to 0$ gives $A(r) = u(x_0)$: *a harmonic function equals its average over every circle around a point.* The same argument in $\mathbb{R}^n$ uses spheres.

Here is why this gives smoothness. Let $\phi_\varepsilon(x)$ be a smooth radial bump of total integral 1 supported in the disc of radius ε. Averaging the mean-value identity over radii with weight ϕ_ε gives

$$u \;=\; u * \phi_\varepsilon \qquad \text{near any point at distance} > \varepsilon \text{ from the edge of the region.}$$

But a convolution with a smooth function is smooth, since all derivatives can be moved onto ϕ_ε: $\partial^\beta(u * \phi_\varepsilon) = u * \partial^\beta\phi_\varepsilon$. So u is C^∞. The argument used *only* integrals of u, so it works even

for merely integrable u, which is Weyl's lemma. Harmonic functions are self-smoothing: each is its own local average.

Worked example 2.5: Harmonic polynomials are smooth, and every harmonic function in the plane is a real part

In $\mathbb{R}^2$ write $z = x + iy$. For any integer $k \geq 0$ the functions $u_k = \operatorname{Re} z^k$ and $v_k = \operatorname{Im} z^k$ are harmonic. For $k = 2$: $z^2 = (x^2 - y^2) + 2ixy$, and

$$\Delta(x^2 - y^2) \overset{(1)}{=} -(2) - (-2) = 0, \qquad \Delta(2xy) \overset{(2)}{=} -0 - 0 = 0.$$

(1) $\partial_x^2(x^2 - y^2) = 2$ and $\partial_y^2(x^2 - y^2) = -2$.
(2) $2xy$ is linear in each variable separately.

More generally, if $F(z)$ is complex analytic then $\operatorname{Re} F$ and $\operatorname{Im} F$ are harmonic, by the Cauchy–Riemann equations. Conversely, every harmonic function on a disc is the real part of an analytic function. So in two dimensions *elliptic regularity for Δ is the smoothness of analytic functions*. We will see this again when we meet the Cauchy–Riemann operator in Section 2.2.10.

Watch out: Elliptic regularity is local and needs smooth coefficients

Elliptic regularity does not say solutions exist, and it does not say solutions are bounded or decay: $u = e^x \cos y$ is harmonic on all of $\mathbb{R}^2$ and grows exponentially. It says only that solutions have as many derivatives as the source allows. It is also local: $\Delta u = f$ with f rough on a region K gives a u that is smooth outside K but may be rough on K. For example, the potential of a uniformly charged ball has a discontinuous second derivative on the surface of the ball, where f jumps.

2.1.5 Hyperbolic operators: the wave equation

Now we look at what goes wrong when the quadratic form is *not* definite. Nash takes the quadratic form $\sum_i a_i p_i^2$ with signature $n - 2$: $n - 1$ positive eigenvalues and one negative. The associated operator is called **hyperbolic**. The simplest case is $n = 2$ with

$$L = \frac{\partial^2}{\partial x_1^2} - \frac{\partial^2}{\partial x_2^2}, \qquad \sigma(L)(p) = -p_1^2 + p_2^2. \tag{2.20}$$

(Each ∂_j^2 becomes $(ip_j)^2 = -p_j^2$. Nash writes $p_1^2 - p_2^2$; the overall sign is irrelevant for everything that follows.) The level curves $\sigma = c$ are now *hyperbolae* (Figure 2.2), and the symbol vanishes on the two lines $p_1 = \pm p_2$, which extend to infinity. Our Step 1 argument now fails, and we expect that solutions of $Lu = f$ with smooth f may be singular. We shall see that they are, and that the singularities can be traced to the Green's function.

Rename the variables $(x_1, x_2) \to (x, t)$, so $L = \partial_x^2 - \partial_t^2$. The key to the whole analysis is a change of variables to **light-cone coordinates**

$$\xi = x + t, \qquad \eta = x - t. \tag{2.21}$$

By the chain rule $\partial_x = \partial_\xi + \partial_\eta$ and $\partial_t = \partial_\xi - \partial_\eta$, so

$$\begin{aligned} L &\overset{(1)}{=} (\partial_\xi + \partial_\eta)^2 - (\partial_\xi - \partial_\eta)^2 \\ &\overset{(2)}{=} \left(\partial_\xi^2 + 2\partial_\xi\partial_\eta + \partial_\eta^2\right) - \left(\partial_\xi^2 - 2\partial_\xi\partial_\eta + \partial_\eta^2\right) = 4\,\partial_\xi\partial_\eta. \end{aligned} \tag{2.22}$$

(1) Substitute the chain-rule expressions for ∂_x and ∂_t.
(2) Expand the squares; partial derivatives commute, so the cross terms are $2\partial_\xi\partial_\eta$.

So the wave operator is just a mixed second derivative. Its solutions with $f = 0$ are therefore

$$u(x,t) \;=\; F(x+t) + H(x-t) \tag{2.23}$$

for *arbitrary* functions F and H of one variable (d'Alembert). Here is the problem in one line: F and H can be as rough as we like, and u still solves $Lu = 0$. Nothing smooths them out. Roughness travels along the lines $x + t = \text{const}$ and $x - t = \text{const}$, the **characteristics** or light rays.

Worked example 2.6: Nash's Green's function for the wave operator, (2.10)

Nash claims that

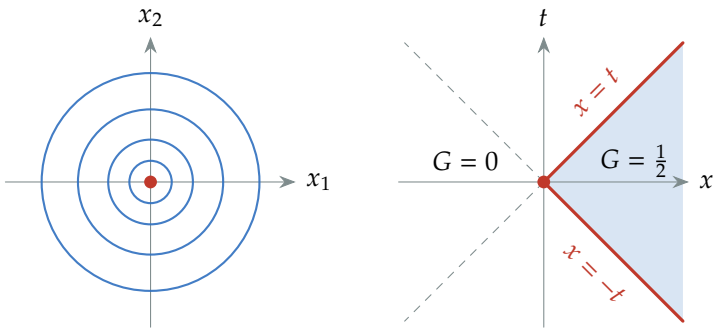
$$G(x,t) \;=\; \tfrac{1}{2}\,\theta(x+t)\,\theta(x-t), \qquad \theta(s) = \begin{cases} 1, & s \geq 0, \\ 0, & s < 0, \end{cases} \tag{2.24}$$

satisfies $LG = \delta(x)\delta(t)$ for $L = \partial_x^2 - \partial_t^2$. We check this in light-cone coordinates, where $G = \frac{1}{2}\theta(\xi)\theta(\eta)$:

$$LG \overset{(1)}{=} 4\,\partial_\xi\partial_\eta\big(\tfrac{1}{2}\theta(\xi)\theta(\eta)\big) \overset{(2)}{=} 2\,\delta(\xi)\,\delta(\eta) \overset{(3)}{=} \delta(x)\,\delta(t).$$

(1) The factorisation (2.22).
(2) The derivative of a step function is a delta function, $\theta' = \delta$, and the two factors depend on different variables.
(3) The Jacobian of $(x,t) \mapsto (\xi,\eta)$ is $|\partial(\xi,\eta)/\partial(x,t)| = \left|\det\left(\begin{smallmatrix} 1 & 1 \\ 1 & -1 \end{smallmatrix}\right)\right| = 2$, so $\mathrm{d}\xi\,\mathrm{d}\eta = 2\,\mathrm{d}x\,\mathrm{d}t$. Hence $\int \delta(\xi)\delta(\eta)\,\mathrm{d}x\,\mathrm{d}t = \frac{1}{2}$, that is, $\delta(\xi)\delta(\eta) = \frac{1}{2}\delta(x)\delta(t)$.

So G is a Green's function. It equals $\frac{1}{2}$ on the wedge $x \geq |t|$ and 0 elsewhere. It is *discontinuous across the whole of the two half-lines* $x = \pm t$, $x \geq 0$, not just at the origin (Figure 2.4). Compare the Laplacian, whose Green's function is singular only at the source. This is the Green's-function face of non-ellipticity.



Laplace: $G = -\frac{1}{2\pi}\ln r$
singular only at 0

wave: $G = \frac{1}{2}\theta(x+t)\theta(x-t)$
singular on two whole half-lines

Figure 2.4: Where the Green's function is singular (red). Left: for the (elliptic) Laplacian in the plane, only at the source; the blue level curves of G are smooth circles. Right: for the (hyperbolic) wave operator $\partial_x^2 - \partial_t^2$, Nash's Green's function (2.10) jumps from 0 to $\frac{1}{2}$ across the two characteristic half-lines $x = \pm t$. Every point of these lines is in the singular support.

Worked example 2.7: A solution of $Lu = 0$ that is only C^2, Nash's (2.11)

Take $F(s) = s^{5/2}$ and $H = 0$ in (2.23):

$$u(x,t) \;=\; (x+t)^{5/2}. \tag{2.25}$$

For $s = x + t < 0$ we need to say what $s^{5/2}$ means. Nash takes the principal branch: $s^{5/2} = |s|^{5/2}$ for $s > 0$ and $s^{5/2} = i\,|s|^{5/2}$ for $s < 0$ (the positive-imaginary choice). Either way u is a function of $x + t$ alone, so $Lu = 4\partial_\xi\partial_\eta u = 0$ wherever u is twice differentiable. Now differentiate along s for $s > 0$:

$$\frac{\mathrm{d}}{\mathrm{d}s}s^{5/2} = \tfrac{5}{2}s^{3/2}, \qquad \frac{\mathrm{d}^2}{\mathrm{d}s^2}s^{5/2} = \tfrac{15}{4}s^{1/2}, \qquad \frac{\mathrm{d}^3}{\mathrm{d}s^3}s^{5/2} = \tfrac{15}{8}s^{-1/2} \xrightarrow{\;s\to 0^+\;} \infty.$$

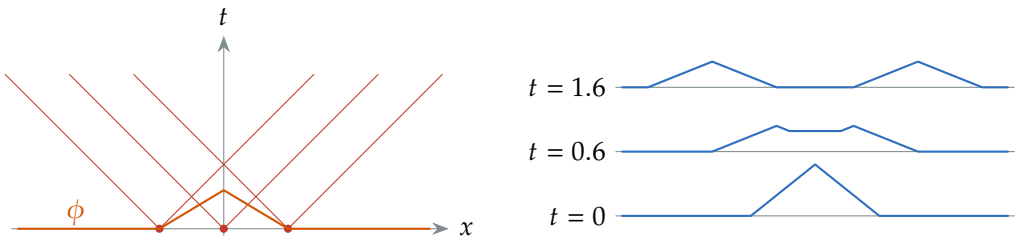
The same happens for $s < 0$ with an extra factor i. So u has two continuous derivatives across the line $x = -t$ (the first and second derivatives tend to 0 from both sides), but its third derivative blows up there. The data were smooth (zero source), yet the solution is only C^2. This cannot happen for an elliptic operator.

Worked example 2.8: A kink that never heals

Solve $(\partial_t^2 - \partial_x^2)u = 0$ with initial data $u(x,0) = \phi(x)$, $\partial_t u(x,0) = 0$. From (2.23), $u(x,0) = F(x) + H(x) = \phi(x)$ and $\partial_t u(x,0) = F'(x) - H'(x) = 0$, so $F = H = \frac{1}{2}\phi$ (up to a constant that cancels), and

$$u(x,t) \;=\; \tfrac{1}{2}\big[\phi(x+t) + \phi(x-t)\big]. \tag{2.26}$$

Take a "tent" $\phi(x) = \max(0, 1 - |x|)$, which has kinks at $x = 0, \pm 1$. Then $u(\cdot, t)$ consists of two half-height tents moving left and right at unit speed. The kinks never smooth out: at time t they sit at $x = \pm t, \pm t \pm 1$. The singular support of u is the union of the characteristic lines through the kinks of the initial data (Figure 2.5). Contrast the heat equation, where the same tent would become C^∞ at any $t > 0$ (Section 2.2.8).



kinks of the data travel on characteristics — snapshots $u(\cdot, t)$: kinks stay sharp

Figure 2.5: Hyperbolic equations propagate singularities. Left: the tent-shaped initial data ϕ has kinks at $x = -1, 0, 1$; each sends a kink along both light rays $x = \text{const} \pm t$ (red). Right: d'Alembert's solution $u = \frac{1}{2}[\phi(x+t) + \phi(x-t)]$ at three times; the profile splits into two half-height tents with permanently sharp corners. At $t = 0.6$ the two tents still overlap, and between their peaks they add up to a flat plateau.

2.1.6 The Cauchy problem, light cones and Huygens' principle

The natural question for a hyperbolic equation is not "solve $Lu = f$ on all of space" but an **initial value problem**, also called a **Cauchy problem**: given u and $\partial_t u$ at an initial time t_0 (the

Cauchy data), find u at later times. Nash states it for the wave operator in n space dimensions,

$$Lu = f, \qquad L = \frac{\partial^2}{\partial t^2} - \frac{\partial^2}{\partial x_1^2} - \cdots - \frac{\partial^2}{\partial x_n^2}, \tag{2.27}$$

Nash's (2.12). This is the d'Alembertian $\Box$ of $(n+1)$-dimensional Minkowski space, and the characteristic lines $x = \pm t$ of the $1+1$ case become **light cones** $|x - x_0| = |t - t_0|$.

Recall that the **support** of a function is the closure of the set where it is non-zero (Nash's footnote 2). The **singular support** is the closure of the set where it fails to be C^∞. Nash's description of how singularities propagate is:

Key idea: Propagation of singularities for the wave equation

Suppose the Cauchy data at time t_0 is singular on a set S (its singular support), and f is smooth. Then the solution $u(x,t)$ at a later time can be singular at a point (x,t) only if the backward light cone from (x,t) meets S at time t_0. In other words: singularities move at the speed of light, along light rays, and nowhere else.

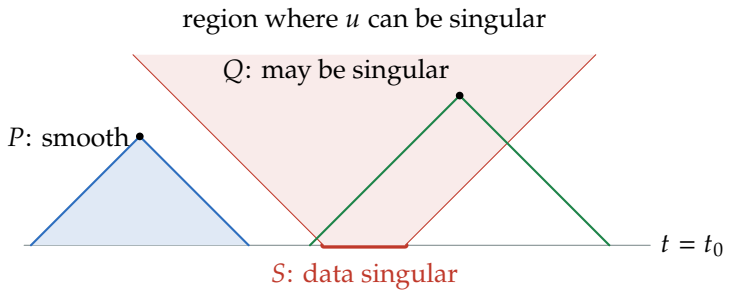


Figure 2.6: The Cauchy problem for the wave equation. The Cauchy data at $t = t_0$ is singular on S (thick red). At a later point P whose backward light cone (blue) misses S, the solution is smooth. At Q the backward cone (green) meets S, so u may be singular there. In odd space dimensions $n \geq 3$ only the *edges* of the shaded region (the light cones through ∂S) carry the singularities that started on ∂S; this is Huygens' principle.

How do we see this? Once again, through the Green's function. The solution of the Cauchy problem is built from the **retarded Green's function** G_{ret}, the field of a flash of light emitted at the origin at $t = 0$ ($LG_{\text{ret}} = \delta$, $G_{\text{ret}} = 0$ for $t < 0$). The data at time t_0 is propagated to time t by convolution with $G_{\text{ret}}(\cdot, t - t_0)$ and its time derivative. Singularities of u come from singularities of the data, carried along by the singularities of G_{ret}, exactly as in (2.19). So the whole question is where G_{ret} is singular.

Worked example 2.9: The retarded Green's function in three space dimensions

Fourier transform in x only. For each wave-vector p, the equation $(\partial_t^2 + |p|^2)g = \delta(t)$ with $g = 0$ for $t < 0$ is solved by $g(t) = \theta(t)\sin(|p|\,t)/|p|$. (Check: g is continuous at $t = 0$ and g' jumps by 1 there, which produces the $\delta(t)$ when we differentiate twice; for $t > 0$, g solves the harmonic-oscillator equation.) So, with $r = |x|$, $k = |p|$ and $t > 0$,

$$\begin{aligned} G_{\text{ret}}(x,t) &\overset{(1)}{=} \frac{1}{(2\pi)^3}\int_{\mathbb{R}^3} e^{ip\cdot x}\,\frac{\sin(kt)}{k}\,\mathrm{d}^3p \overset{(2)}{=} \frac{1}{(2\pi)^2}\int_0^\infty k^2\,\frac{2\sin(kr)}{kr}\,\frac{\sin(kt)}{k}\,\mathrm{d}k \\ &\overset{(3)}{=} \frac{1}{2\pi^2 r}\int_0^\infty \sin(kr)\sin(kt)\,\mathrm{d}k \overset{(4)}{=} \frac{1}{2\pi^2 r}\cdot\frac{\pi}{2}\big[\delta(r-t) - \delta(r+t)\big] \\ &\overset{(5)}{=} \frac{\delta(t-r)}{4\pi r}. \end{aligned}$$

(1) Inverse Fourier transform in the three space variables.
(2) Angular integration exactly as in Example 2.2.
(3) Simplify: $k^2 \cdot 2/(kr) \cdot 1/k = 2/r$, and $2/(2\pi)^2 = 1/(2\pi^2)$.
(4) $\sin a \sin b = \frac{1}{2}[\cos(a-b) - \cos(a+b)]$ and $\int_0^\infty \cos(ks)\,\mathrm{d}k = \pi\delta(s)$.
(5) For $t > 0$ and $r \geq 0$, $r + t > 0$, so $\delta(r + t) = 0$.

The retarded Green's function in three dimensions is supported *exactly on the forward light cone* $r = t$: a flash of light is seen once, as a sharp shell, and then it is dark again.

Doing the same calculation in other dimensions gives (with $r = |x|$)

$$G_{\mathrm{ret}}(x,t) \;=\; \begin{cases} \frac{1}{2}\,\theta(t-|x|), & n=1, \\ \dfrac{\theta(t-r)}{2\pi\sqrt{t^2-r^2}}, & n=2, \\ \dfrac{\delta(t-r)}{4\pi r}, & n=3. \end{cases} \tag{2.28}$$

In every case the *singular support* is the light cone $r = t$. But the *support* differs. For $n = 3$ (and, it turns out, all odd $n \geq 3$) the support is the cone itself. For $n = 1, 2$ (and all even n) the support is the whole solid interior of the cone. This is Nash's remark that "for $n = 3, 5, \ldots$ the support and the singular support of the Green's function actually coincide" (he cites Treves). The physical name is **Huygens' principle**: in three space dimensions a sharp signal stays sharp and leaves no wake. In two dimensions it does leave a wake: drop a stone in a pond and the water keeps rippling inside the expanding circle.

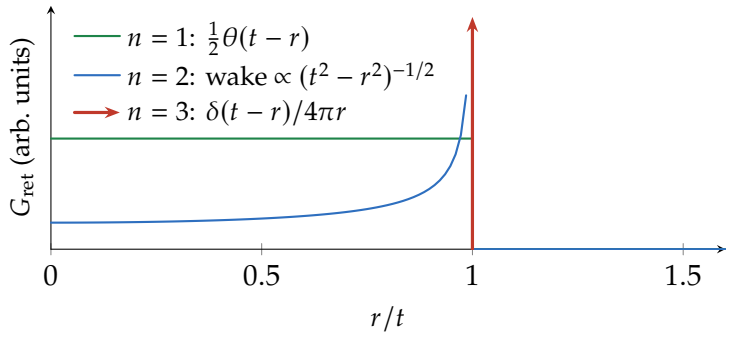


Figure 2.7: The retarded Green's function at a fixed time t, as a function of distance r. In $n = 1$ (green) the signal fills the interior of the light cone with a constant. In $n = 2$ (blue) there is a wake inside the cone, singular at the front $r = t$. In $n = 3$ (red arrow) everything is concentrated on the front: Huygens' principle. In all three cases the singularities sit on the light cone $r = t$.

2.1.7 Wick rotation: turning the wave operator into the Laplacian

The wave operator $\partial_x^2 - \partial_t^2$ and the Laplacian $-\partial_x^2 - \partial_y^2$ differ only in a sign. Substituting $t = iy$ gives $\partial_t = -i\partial_y$, hence $\partial_t^2 = -\partial_y^2$, and

$$\frac{\partial^2}{\partial x^2} - \frac{\partial^2}{\partial t^2} \;\xrightarrow{\;t=iy\;}\; \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \;=\; -\Delta. \tag{2.29}$$

Physicists call this kind of substitution a **Wick rotation**, and it is the bridge between Minkowski and Euclidean field theory. Nash uses it to make a subtle point.

Nash's puzzle: is $z^{5/2}$ a non-smooth harmonic function?

Apply the substitution to the C^2 solution (2.25):

$$u(x,t) = (x+t)^{5/2} \xrightarrow{\;t=iy\;} u(x,y) = (x+iy)^{5/2} = z^{5/2}.$$

One might claim that $z^{5/2}$ solves Laplace's equation and is singular at the origin (its third derivatives blow up there), contradicting elliptic regularity. It does not, and the reason is instructive.

In the real (x,t)-plane, $x+t$ is a real number, and Nash's rule for $(x+t)^{5/2}$ (principal branch) gives a single-valued function on the whole plane. In the complex z-plane, $z^{5/2}$ is *multi-valued*. To make it a function we must cut the plane along a ray from 0 to ∞, say the negative real axis. Write $z = re^{i\theta}$ with $-\pi < \theta \le \pi$, so $z^{5/2} = r^{5/2}e^{5i\theta/2}$. Approaching the negative real axis from above and from below,

$$\lim_{\theta\to\pi^-} z^{5/2} \;=\; r^{5/2}e^{5i\pi/2} \;=\; i\,r^{5/2}, \qquad \lim_{\theta\to-\pi^+} z^{5/2} \;=\; r^{5/2}e^{-5i\pi/2} \;=\; -i\,r^{5/2}.$$

So $z^{5/2}$ *jumps by* $2ir^{5/2}$ *across the cut*. It is not continuous there, let alone differentiable, so it does not satisfy Laplace's equation on any open set that meets the cut. Every neighbourhood of the origin meets the cut, so $z^{5/2}$ is not a solution of $\Delta u = 0$ near the origin at all, and elliptic regularity has nothing to say about it. On any open set that avoids the cut, $z^{5/2}$ is complex analytic, hence harmonic *and* C^∞, exactly as elliptic regularity predicts (Figure 2.8).

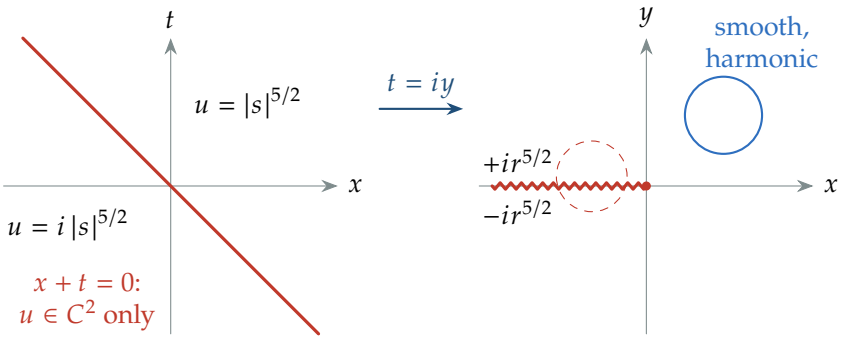


Figure 2.8: Nash's warning about Wick rotation. Left: $(x+t)^{5/2}$ is single valued on the real (x,t)-plane and solves the wave equation, but it is only C^2 across the characteristic $x+t=0$. Right: after $t \to iy$ it becomes $z^{5/2}$, which needs a branch cut (zigzag). On any disc away from the cut (blue) it is analytic and smooth. On a disc crossing the cut (dashed red) it is discontinuous, so it does not solve Laplace's equation there. Elliptic regularity survives.

Wick rotation in quantum field theory

The same mechanism is why quantum field theorists love to Wick rotate. Take the Klein–Gordon operator in $3+1$ dimensions with metric $(+,-,-,-)$:

$$L_{\mathrm{M}} \;=\; \frac{\partial^2}{\partial t^2} - \nabla^2 + m^2. \tag{2.30}$$

Using (2.5) with Fourier variables $(p_0, \vec{p})$ for $(t, \vec{x})$, its symbol is

$$\sigma(L_{\mathrm{M}}) \;=\; -p_0^2 + |\vec{p}|^2 + m^2, \tag{2.31}$$

which vanishes on the **mass shell** $p_0 = \pm\sqrt{|\vec{p}|^2 + m^2}$: a two-sheeted hyperboloid running off to infinity. The Green's function $\int e^{ip\cdot x}/\sigma \, \mathrm{d}^4p$ is ill-defined until we say how to go around the poles. That is the $i\varepsilon$ prescription, and each choice (retarded, advanced, Feynman) gives a different Green's function, all singular on the light cone.

Now set $t = -i\tau$ ("imaginary time"). Then $\partial_t = i\partial_\tau$, $\partial_t^2 = -\partial_\tau^2$, and

$$L_{\mathrm{M}} \overset{(1)}{=} -\frac{\partial^2}{\partial\tau^2} - \nabla^2 + m^2 \overset{(2)}{=} \Delta_{\mathbb{R}^4} + m^2 \;=:\; L_{\mathrm{E}}, \tag{2.32}$$

(1) $\partial_t^2 = (i\partial_\tau)^2 = -\partial_\tau^2$.
(2) The positive Laplacian of $\mathbb{R}^4$ with coordinates $(\tau, \vec{x})$.

The Euclidean operator has symbol $p_4^2 + |\vec{p}|^2 + m^2 \geq m^2 > 0$. It never vanishes: L_{E} is elliptic, its Green's function

$$G_{\mathrm{E}}(x) \;=\; \frac{1}{(2\pi)^4} \int_{\mathbb{R}^4} \frac{e^{ip\cdot x}}{|p|^2 + m^2} \, \mathrm{d}^4p$$

is unique (no pole prescription needed), decays like $e^{-m|x|}$, and is smooth everywhere except at $x = 0$. In momentum space the Wick rotation is the rotation of the p_0 integration contour onto the imaginary axis, $p_0 = ip_4$ (Figure 2.9). The poles of the Feynman propagator sit at $p_0 = \pm(\omega_{\vec{p}} - i\varepsilon)$ with $\omega_{\vec{p}} = \sqrt{|\vec{p}|^2 + m^2}$, in the fourth and second quadrants, so the rotation through the first and third quadrants crosses no poles:

$$\int_{-\infty}^{\infty} \frac{\mathrm{d}p_0}{-p_0^2 + \omega_{\vec{p}}^2 - i\varepsilon} \overset{(1)}{=} \int_{-i\infty}^{i\infty} \frac{\mathrm{d}p_0}{-p_0^2 + \omega_{\vec{p}}^2} \overset{(2)}{=} i \int_{-\infty}^{\infty} \frac{\mathrm{d}p_4}{p_4^2 + \omega_{\vec{p}}^2} \;=\; \frac{i\pi}{\omega_{\vec{p}}}.$$

(1) Rotate the contour anticlockwise by 90°. The arcs at infinity contribute nothing because the integrand falls off like p_0^{-2}, and no poles are crossed. The $i\varepsilon$ may then be dropped.
(2) Substitute $p_0 = ip_4$, so $\mathrm{d}p_0 = i\,\mathrm{d}p_4$ and $-p_0^2 = p_4^2$.

(The last integral is $\pi/\omega_{\vec{p}}$ by $\int \mathrm{d}s/(s^2 + a^2) = \pi/a$.) A Minkowski integral with delicate poles has become a Euclidean integral with a strictly positive, elliptic denominator.

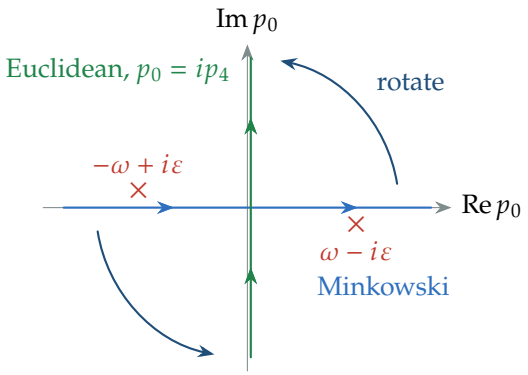


Figure 2.9: Wick rotation in the complex p_0-plane. The Feynman poles (red crosses) sit just below the positive real axis and just above the negative real axis. The real-axis contour (blue) can be swung anticlockwise onto the imaginary axis (green) without crossing them. On the imaginary axis the denominator $p_4^2 + \omega^2$ never vanishes: the operator has become elliptic.

Dictionary: Minkowski versus Euclidean

Minkowski (real time t)	Euclidean (imaginary time $\tau = it$)
wave / Klein–Gordon operator: hyperbolic	Laplacian $+ m^2$: elliptic
symbol vanishes on the mass shell	symbol $\geq m^2 > 0$
Green's function needs $i\varepsilon$; singular on light cone	unique Green's function; singular only at $x = 0$
propagation of signals	correlations decaying like $e^{-m\lvert x\rvert}$
e^{iS}, oscillatory path integral	$e^{-S_{\mathrm{E}}}$, damped path integral
Schrödinger evolution e^{-iHt}	heat / diffusion evolution $e^{-H\tau}$

This is why Nash, and the index theory that powers Chapters 8 and 10, work in Euclidean signature: only there are the relevant operators elliptic.

2.1.8 Parabolic operators: Schrödinger and heat

There is a third class. Nash's example is the free Schrödinger equation in one space dimension (with $\hbar = 2m = 1$), written as

$$Lu \;=\; -i\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} \;=\; 0, \tag{2.33}$$

Nash's (2.13). Let (p_1, p_2) be the Fourier variables for (t, x). Then $-i\partial_t \to -i(ip_1) = p_1$ and $-\partial_x^2 \to p_2^2$, so

$$\sigma(L) \;=\; p_1 + p_2^2, \tag{2.34}$$

and the level curves $\sigma = c$ are the parabolas $p_2^2 + p_1 = c$. (Nash's (2.14), $\sigma = -p_1 + p_2^2$, comes from his opposite Fourier sign; the parabola simply opens the other way.) So this is a **parabolic** equation. The heat operator $\partial_t - \partial_x^2$ is the same up to replacing $-i\partial_t$ by ∂_t. Its symbol is $ip_1 + p_2^2$.

Nash then makes a point that matters for everything that follows. Although the symbol (2.34) happens to describe a parabola, the classification is *not* really made from the whole symbol. It is made from the *highest-order* part only, here p_2^2, which comes from ∂_x^2. In the elliptic and hyperbolic examples every term had the same order, 2, so the distinction never arose. For the Schrödinger operator the second-order part p_2^2 is a *degenerate* quadratic form: its matrix $\left(\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ has a zero eigenvalue in the p_1 (time) direction. That degenerate direction is what makes it parabolic.

2.1.9 The classification of second-order operators

Nash now gives one definition covering all three types. Take the general second-order operator with constant coefficients on $\mathbb{R}^n$,

$$L \;=\; -\sum_{i,j=1}^{n} a_{ij}\frac{\partial^2}{\partial x_i \partial x_j} + \sum_{i=1}^{n} b_i \frac{\partial}{\partial x_i} + c, \qquad a_{ij} = a_{ji}, \tag{2.35}$$

Nash's (2.16). (We may take a_{ij} symmetric because $\partial_i\partial_j = \partial_j\partial_i$; only the symmetric part of a_{ij} contributes.) Fourier transform only the second-order part:

$$-\sum_{i,j} a_{ij}\,\partial_i\partial_j \;\longrightarrow\; -\sum_{i,j} a_{ij}(ip_i)(ip_j) \;=\; \sum_{i,j} a_{ij}\,p_i p_j \;=:\; \sigma_2(L)(p), \tag{2.36}$$

Nash's (2.17). This is a quadratic form, with symmetric matrix $A = (a_{ij})$.

Working definition 2.4 (elliptic / hyperbolic / parabolic, Nash's (2.18))

Diagonalise the symmetric matrix $A = (a_{ij})$ of the second-order part and count the signs of its eigenvalues. Then L is

- **elliptic** if all n eigenvalues have the same sign (signature [number of positive eigenvalues minus number of negative ones] $\pm n$);
- **hyperbolic** if $\det A \neq 0$ and exactly one eigenvalue has the opposite sign to the others (signature $\pm(n-2)$);
- **parabolic** if $\det A = 0$, i.e. at least one eigenvalue is zero.

In two variables, with $L = -(A\,\partial_x^2 + 2B\,\partial_x\partial_y + C\,\partial_y^2) + \dots$, the test is the sign of $\det A = AC - B^2$: positive means elliptic, negative hyperbolic, zero parabolic.

Watch out: The classification is not exhaustive, and "parabolic" needs more

For $n \geq 4$ there are forms with two or more eigenvalues of each sign, such as $p_1^2 + p_2^2 - p_3^2 - p_4^2$. They are called **ultrahyperbolic** and fit none of Nash's three boxes. Also, $\det A = 0$ alone does not make a good parabolic equation. The heat and Schrödinger equations also have a *first-order* derivative in the degenerate direction (∂_t), which is what makes the time evolution well posed. The operator ∂_x^2 on $\mathbb{R}^2$ has $\det A = 0$ but no ∂_y term, and it is really a one-dimensional operator with y as a spectator.

Worked example 2.10: Classifying three operators by diagonalising

Classify, on $\mathbb{R}^2$, the operators $L_k = -(\partial_x^2 + 2\beta_k\,\partial_x\partial_y + \partial_y^2)$ with $\beta_1 = 2$, $\beta_2 = \tfrac{1}{2}$, $\beta_3 = 1$. The matrix of the second-order part is $A = \left(\begin{smallmatrix} 1 & \beta \\ \beta & 1 \end{smallmatrix}\right)$. Its eigenvectors are $(1,1)/\sqrt{2}$ and $(1,-1)/\sqrt{2}$, with eigenvalues

$$A\begin{pmatrix}1\\1\end{pmatrix} \overset{(1)}{=} \begin{pmatrix}1+\beta\\ \beta+1\end{pmatrix} = (1+\beta)\begin{pmatrix}1\\1\end{pmatrix}, \qquad A\begin{pmatrix}1\\-1\end{pmatrix} \overset{(2)}{=} \begin{pmatrix}1-\beta\\ \beta-1\end{pmatrix} = (1-\beta)\begin{pmatrix}1\\-1\end{pmatrix}.$$

(1) Multiply out: first row $1\cdot 1 + \beta\cdot 1$, second row $\beta\cdot 1 + 1\cdot 1$.
(2) First row $1-\beta$, second row $\beta - 1$.

So the eigenvalues are $1 \pm \beta$:

β	eigenvalues	$\det A = 1-\beta^2$	type
2	$3, -1$	-3	hyperbolic
$\frac{1}{2}$	$\frac{3}{2}, \frac{1}{2}$	$\frac{3}{4}$	elliptic
1	$2, 0$	0	parabolic (degenerate)

In the rotated coordinates $s = (x+y)/\sqrt{2}$, $w = (x-y)/\sqrt{2}$ the operator becomes $L = -(1+\beta)\partial_s^2 - (1-\beta)\partial_w^2$. For $\beta = 2$ this is $-3\partial_s^2 + \partial_w^2$, a wave operator in which w plays the role of time and the "speed of light" is $\sqrt{3}$. For $\beta = 1$ it is $-2\partial_s^2$: the operator does not differentiate in the w direction at all.

Worked example 2.11: An equation that changes type: the Tricomi equation

Coefficients need not be constant, and then the type can change from point to point. The Tricomi operator on $\mathbb{R}^2$,

$$L \;=\; -\big(y\,\partial_x^2 + \partial_y^2\big), \qquad \sigma_2(L)(x,y;p) \;=\; y\,p_1^2 + p_2^2,$$

is elliptic in the upper half-plane $y > 0$ (both coefficients positive), hyperbolic in the lower half-plane $y < 0$, and parabolic on the line $y = 0$. It is the model for *transonic flow*: $y > 0$ is the subsonic region (the flow "feels" disturbances in every direction), $y < 0$ the supersonic region (disturbances propagate only inside a Mach cone), and $y = 0$ is the sonic line. This example is the reason Nash moves on, in Section 2.2, to operators with *variable* coefficients, where ellipticity becomes a property checked point by point.

2.1.10 Characteristics: the directions an operator cannot see

The zero set of the principal symbol has a direct meaning in position space, and it explains the word "characteristic" that we have been using.

Working definition 2.5 (characteristic covector and hypersurface)

A nonzero covector [a list of n numbers, like a momentum p or a gradient $\partial_i\varphi$] $\xi = (\xi_1,\dots,\xi_n)$ at a point x is **characteristic** for L if the principal symbol [the second-order part only, $\sigma_2(L)(x,\xi) = \sum_{ij} a_{ij}(x)\xi_i\xi_j$] vanishes on it, $\sigma_2(L)(x,\xi) = 0$. A hypersurface [a surface of dimension $n-1$ in $\mathbb{R}^n$] $S = \{\varphi(x) = 0\}$ is a **characteristic surface** if its normal covector $\mathrm{d}\varphi = (\partial_1\varphi,\dots,\partial_n\varphi)$ is characteristic at every point of S. *Test for ellipticity:* L is elliptic exactly when it has *no real characteristic covectors*, hence no characteristic surfaces.

Why do characteristic surfaces matter? Try to solve $Lu = 0$ near a surface S on which we know u and its normal derivative (Cauchy data). Choose coordinates with $S = \{x_1 = 0\}$, so the normal covector is $\xi = (1,0,\dots,0)$. On S we know u, hence all its tangential derivatives $\partial_j u$, $\partial_j\partial_k u$ for $j,k \geq 2$. We know $\partial_1 u$, hence its tangential derivatives $\partial_1\partial_j u$. The only second derivative we do not know is $\partial_1^2 u$. The equation (2.35) reads

$$0 \;=\; Lu \overset{(1)}{=} -a_{11}\,\partial_1^2 u \;-\; \Big(\text{terms known on } S\Big) \overset{(2)}{=} -\sigma_2(L)(\xi)\,\partial_1^2 u \;-\; \Big(\text{terms known on } S\Big).$$

(1) Separate the one term containing $\partial_1^2 u$; every other term involves at most one normal derivative and is known on S.

(2) For $\xi = (1,0,\dots,0)$ we have $\sigma_2(\xi) = \sum a_{ij}\xi_i\xi_j = a_{11}$.

If $\sigma_2(\xi) \neq 0$ we can solve for $\partial_1^2 u$ on S. Differentiating the equation, we then get all higher normal derivatives: the Cauchy data determines u (formally, as a Taylor series) near S. If $\sigma_2(\xi) = 0$, the equation gives no information about $\partial_1^2 u$. Then two solutions can agree on one side of S and differ on the other, with a jump in the second derivative across S. That is exactly what we saw in Example 2.7 and Example 2.8: the kinks sit on the characteristic lines $x \pm t = \text{const}$, whose normals $(1,\pm 1)$ satisfy $-\xi_1^2 + \xi_2^2 = 0$.

Physical picture: Why elliptic solutions cannot have kinks

A kink or jump in a solution must sit on a surface across which the equation does not see the jump, i.e. on a characteristic surface. An elliptic operator has no characteristic surfaces, so there is nowhere for a singularity to hide. For the wave operator the characteristic surfaces are the null (light-like) surfaces, and singularities ride on them at the speed of

light.

	Laplace $-\nabla^2$	wave $\partial_t^2 - \nabla^2$	heat $\partial_t - \nabla^2$
principal symbol	$\lvert p\rvert^2$	$-p_0^2 + \lvert\vec{p}\rvert^2$	$\lvert\vec{p}\rvert^2$ (degenerate)
type	elliptic	hyperbolic	parabolic
characteristics	none	light cones	$t = \text{const}$
Green's function	$\propto r^{2-n}$, smooth off 0	singular on light cone	smooth off origin
smooth $f \Rightarrow$ smooth u?	yes	no (kinks propagate)	yes (Section 2.2.8)
physics	statics, Euclidean QFT	radiation, Minkowski QFT	diffusion, imaginary time

Table 2.1: The three archetypes of second-order operator compared. The heat equation is not elliptic but still smooths; this is the phenomenon of hypo-ellipticity studied in Section 2.2.

Nash closes §1 by announcing the two generalisations that the connection between the smoothness of f and the smoothness of u calls for: operators of *arbitrary order with variable coefficients*, and the property of *hypo-ellipticity*. These are the subject of the next section.

Reading Nash (§1, pp. 27–32, (2.1)–(2.18))

- p. 27, (2.1): the Hodge result $\dim \operatorname{Ker} \Delta_p = \dim H^p(M;\mathbb{R})$, our (2.1), proved in Section 2.3.12.
- p. 28, (2.2)–(2.6): positive Laplacian, its Fourier symbol $\lvert p\rvert^2$, stretched Laplacian $L = -\sum a_i \partial_i^2$ with symbol $\sum a_i p_i^2$, and the ellipsoid picture; our Sections 2.1.2 and 2.1.3. Beware the Fourier sign in his footnote 1 (page 728).
- pp. 28–29, (2.7)–(2.9): regularity via the Green's function. Nash's "it is easy to check" and "appropriate use of Cauchy's theorem" are our Steps 1 and 2 in Section 2.1.4.
- pp. 29–30, (2.10)–(2.12): the wave operator, its Green's function, the C^2 solution $(x+t)^{5/2}$, the Cauchy problem and Huygens' principle; our Sections 2.1.5 and 2.1.6.
- pp. 30–31: the $z^{5/2}$ puzzle; our Section 2.1.7, where we also add the QFT Wick rotation.
- p. 31, (2.13)–(2.18): Schrödinger as parabolic, the general second-order classification; our Sections 2.1.8 and 2.1.9. Nash's signs in (2.14) differ from ours by the Fourier convention.

Recap: The meaning of ellipticity

- Fourier transformation turns ∂_j into ip_j and a constant-coefficient operator into a polynomial $\sigma(p)$. Solving $Lu = f$ means dividing by σ.
- Elliptic (second order): the top-order quadratic form is definite; level sets are ellipsoids; $\sigma(p) \neq 0$ for $p \neq 0$; no real characteristics.
- Elliptic regularity: the Green's function is smooth away from the source, because $1/\sigma$ gains decay with every p-derivative. Hence smooth f on Ω gives smooth u on Ω.

Weyl: harmonic functions are smooth, being their own averages.

- Hyperbolic: the symbol vanishes on a cone; the Green's function is singular on the light cone; kinks propagate along characteristics; Huygens' principle holds in odd space dimensions ≥ 3.
- Wick rotation $t \to -i\tau$ turns the hyperbolic wave operator into the elliptic Laplacian. Nash's $z^{5/2}$ is not a counterexample, because of its branch cut.
- Parabolic: the top-order form is degenerate (heat, Schrödinger). The classification only looks at the highest-order terms.

2.2 Ellipticity and hypo-ellipticity

So far every operator we met had *constant* coefficients and was of *second* order. That is far too narrow for what follows. The operators of geometry and physics come in all orders and nearly always have coefficients that depend on position:

- The Cauchy–Riemann operator $\bar\partial$ and the Dirac operator $\not{D}$ are of *first* order.
- The biharmonic operator Δ^2 of elasticity (bending of plates) is of *fourth* order.
- The Laplacian on a curved space, written in coordinates, is $-g^{ij}(x)\partial_i\partial_j + \dots$, with coefficients given by the inverse metric. They depend on x unless the space is flat and the coordinates Cartesian. Even the flat Laplacian in polar coordinates has variable coefficients.
- The Tricomi operator of Example 2.11 changes type from point to point.

This section answers two questions. *How do we define ellipticity for an operator of any order with variable coefficients?* And, since the heat equation smooths just as well as Laplace's equation without being elliptic, *what is the right name for "smooth source gives smooth solution", and which operators have it?* The answers are the *leading symbol* and *hypo-ellipticity*.

2.2.1 Multi-index notation

A general operator of order m on $\mathbb{R}^n$ is a sum of terms, each a coefficient times some mixed partial derivative of total order at most m. Writing this out with explicit indices is painful, so everybody uses multi-indices.

Working definition 2.6 (multi-index notation)

A **multi-index** is an n-tuple of non-negative integers $\alpha = (\alpha_1, \dots, \alpha_n)$. Some of the α_i may be zero. With it we write

$$
\begin{aligned}
|\alpha| &= \alpha_1 + \cdots + \alpha_n && \text{(the \textbf{length}, or order),}\\
p^\alpha &= p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_n^{\alpha_n} && \text{(a monomial of degree } |\alpha|\text{),}\\
\partial^\alpha &= \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \cdots \frac{\partial^{\alpha_n}}{\partial x_n^{\alpha_n}} && \text{(a mixed partial derivative of order } |\alpha|\text{),}\\
D^\alpha &= (-i)^{|\alpha|}\,\partial^\alpha = D_1^{\alpha_1}\cdots D_n^{\alpha_n}, \quad D_j = -i\frac{\partial}{\partial x_j} && \text{(the same, times a convenient phase).}
\end{aligned}
$$

Operationally: α_j counts how many times we differentiate with respect to x_j, and $|\alpha|$ is the total number of derivatives.

With this notation the general linear differential operator of order m is

$$L \;=\; \sum_{|\alpha|\le m} a_\alpha(x)\, D^\alpha, \tag{2.37}$$

Nash's (2.19). Written out in full, this is

$$L \;=\; \sum_{|\alpha|\le m} a_{(\alpha_1,\dots,\alpha_n)}(x)\,(-i)^{|\alpha|}\,\frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}}\cdots\frac{\partial^{\alpha_n}}{\partial x_n^{\alpha_n}}, \tag{2.38}$$

Nash's (2.20). The sum runs over all multi-indices of length $0, 1, \dots, m$. For example, in $n = 2$ variables and order $m = 2$ there are six of them:

$$(0,0);\quad (1,0),\ (0,1);\quad (2,0),\ (1,1),\ (0,2),$$

corresponding to 1; D_1, D_2; D_1^2, D_1D_2, D_2^2. In general the number of multi-indices of length exactly k in n variables is $\binom{n+k-1}{k}$ (the number of ways of distributing k derivatives among n variables).

Why the factor $(-i)^{|\alpha|}$? Because $D_j = -i\partial_j$ becomes multiplication by p_j under our Fourier transform (2.5), so D^α becomes multiplication by p^α with no stray factors of i. Nash says the factor "removes it from expressions obtained later by Fourier transformation." The price is that the coefficients a_α of a real operator may be imaginary, as the next example shows.

Worked example 2.12: Translating an operator into multi-index form

Consider on $\mathbb{R}^2$, with coordinates (x, y),

$$L \;=\; \frac{\partial^3}{\partial x^2\,\partial y} \;-\; 3x\,\frac{\partial}{\partial x} \;+\; 5.$$

Each partial derivative is $\partial_j = iD_j$, so $\partial^\alpha = i^{|\alpha|}D^\alpha$. Hence

$$\frac{\partial^3}{\partial x^2 \partial y} \overset{(1)}{=} \partial^{(2,1)} \overset{(2)}{=} i^3\, D^{(2,1)} \overset{(3)}{=} -i\, D^{(2,1)},$$
$$-3x\,\frac{\partial}{\partial x} \overset{(4)}{=} -3x \cdot i\, D^{(1,0)} \;=\; -3ix\, D^{(1,0)}.$$

(1) Two derivatives in x and one in y: $\alpha = (2,1)$, $|\alpha| = 3$.
(2) $\partial^\alpha = i^{|\alpha|}D^\alpha$, the inverse of $D^\alpha = (-i)^{|\alpha|}\partial^\alpha$.
(3) $i^3 = -i$.
(4) One derivative in x: $\alpha = (1,0)$, and $\partial_x = iD_x$.

So the non-zero coefficients are $a_{(2,1)} = -i$, $a_{(1,0)}(x,y) = -3ix$ and $a_{(0,0)} = 5$, and $m = 3$. The coefficients are imaginary even though L is a real operator. That is the bookkeeping cost of using D instead of ∂.

2.2.2 The symbol: what L does to a plane wave

We now Fourier transform the general operator (2.37), as Nash does in (2.21). Write u as a superposition of plane waves (2.3) and apply L term by term:

$$\begin{aligned} Lu(x) &\overset{(1)}{=} \sum_{|\alpha|\le m} a_\alpha(x)\, D^\alpha \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} e^{ip\cdot x}\, \hat{u}(p)\, \mathrm{d}^n p \\ &\overset{(2)}{=} \sum_{|\alpha|\le m} a_\alpha(x) \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} p^\alpha\, e^{ip\cdot x}\, \hat{u}(p)\, \mathrm{d}^n p \\ &\overset{(3)}{=} \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} \sigma(x,p)\, \hat{u}(p)\, e^{ip\cdot x}\, \mathrm{d}^n p, \end{aligned} \tag{2.39}$$

(1) Definition (2.37) and the inverse Fourier transform (2.3).
(2) Differentiate under the integral. Each D_j acting on $e^{ip\cdot x}$ brings down p_j, so $D^\alpha e^{ip\cdot x} = p^\alpha e^{ip\cdot x}$. The coefficient $a_\alpha(x)$ multiplies *after* differentiation, so it is not differentiated.
(3) Pull the finite sum inside the integral and collect it into one function.

The function

$$\sigma(x,p) \;=\; \sum_{|\alpha|\le m} a_\alpha(x)\, p^\alpha \tag{2.40}$$

is the **symbol** of L, Nash's (2.22). It is a polynomial of degree m in p whose coefficients depend on x. Formula (2.39) says: *to apply L, multiply the Fourier amplitudes by $\sigma(x,p)$ and transform back.* For constant coefficients this is exactly the "derivatives become multiplication" of Section 2.1.2. For variable coefficients it is still true, but σ now depends on the point x where we evaluate Lu.

The most useful way to remember (2.40) is as a test you apply to one plane wave.

Working definition 2.7 (symbol, by the plane-wave test)

To find the symbol of a differential operator [a finite sum $L = \sum_{|\alpha|\le m} a_\alpha(x) D^\alpha$] L, apply L to the plane wave $e^{ip\cdot x}$ and divide by the plane wave:

$$\sigma(x,p) \;=\; e^{-ip\cdot x}\, L\big(e^{ip\cdot x}\big). \tag{2.41}$$

How to compute: replace every ∂_j by ip_j (equivalently every D_j by p_j), keeping each coefficient where it is.

Formula (2.41) follows from $D^\alpha e^{ip\cdot x} = p^\alpha e^{ip\cdot x}$, which is the step (2) of (2.39). It is the version we will generalise to manifolds and vector bundles.

Watch out: Variable coefficients: the symbol is not a multiplier

For constant coefficients, $\widehat{Lu}(p) = \sigma(p)\hat{u}(p)$. For variable coefficients this is *false*: $\sigma(x,p)$ depends on both x and p, so (2.39) is not the inverse Fourier transform of anything simple. It is a mixed position-momentum representation, like a classical phase-space function $H(x,p)$ in quantum mechanics. The plane-wave test (2.41) is exact, but the product of two operators does *not* simply have the product of the symbols as its symbol, because D also differentiates the coefficients of the operator to its right. We come back to this in Section 2.3.8, and in part 2b it becomes the calculus of pseudo-differential operators.

2.2.3 The leading symbol

In Section 2.1.8 we saw that the classification of second-order operators used only the second-order terms. Nash now makes this the general rule.

Working definition 2.8 (leading symbol)

The **leading symbol** of $L = \sum_{|\alpha|\le m} a_\alpha(x) D^\alpha$ is the top-degree part of its symbol [$\sigma(x,p) = \sum_{|\alpha|\le m} a_\alpha(x) p^\alpha$, the factor by which L multiplies the plane wave $e^{ip\cdot x}$]:

$$\sigma_m(x,p) \;=\; \sum_{|\alpha|=m} a_\alpha(x)\, p^\alpha, \tag{2.42}$$

Nash's (2.23). It is also called the **principal symbol** or **highest-order symbol**. *How to compute*: keep only the terms with the maximal number m of derivatives, and replace each ∂_j by ip_j. It is a homogeneous polynomial of degree m in p:

$$\sigma_m(x,\lambda p) \;=\; \lambda^m\, \sigma_m(x,p) \qquad (\lambda \in \mathbb{R}). \tag{2.43}$$

Nash stresses the obvious but important point that different operators of order m with the same top-order terms, differing only in lower orders, have the *same* leading symbol and *different* full symbols.

Worked example 2.13: Same leading symbol, different full symbols

Here are four second-order operators on $\mathbb{R}^3$ from physics, with their full symbols obtained by the plane-wave test (replace ∂_j by ip_j):

operator	full symbol $\sigma(x,p)$	physics
$-\nabla^2$	$\lvert p\rvert^2$	electrostatics
$-\nabla^2 + m^2$	$\lvert p\rvert^2 + m^2$	static Yukawa field
$-\nabla^2 + \vec{b}\cdot\nabla$	$\lvert p\rvert^2 + i\,\vec{b}\cdot\vec{p}$	steady advection–diffusion
$-\nabla^2 + V(x)$	$\lvert p\rvert^2 + V(x)$	time-independent Schrödinger

For the third, $\vec{b}\cdot\nabla e^{ip\cdot x} = i\,\vec{b}\cdot\vec{p}\, e^{ip\cdot x}$. All four have the same leading symbol $\sigma_2 = |p|^2$, so (as we will define in a moment) all four are elliptic, and all four have smooth solutions for smooth sources and potentials. Their full symbols differ, and so do their Green's functions: $1/(4\pi r)$ for the first, the Yukawa potential $e^{-mr}/(4\pi r)$ for the second. But *near the source* all the Green's functions have the same leading singularity $1/(4\pi r)$, as the next paragraph explains.

Why only the top order matters: high frequencies

Why should the lower-order terms be irrelevant for questions of smoothness? Because smoothness is about *high frequencies* ("smoothness is decay in p", Section 2.1.2), and at high frequencies the top-order terms win. Fix a direction $\hat{p}$ with $|\hat{p}| = 1$ and follow the symbol out along the ray $p = \lambda\hat{p}$:

$$\sigma(x,\lambda\hat{p}) \overset{(1)}{=} \sum_{k=0}^{m} \sum_{|\alpha|=k} a_\alpha(x)\, \lambda^k\, \hat{p}^{\,\alpha}$$

$$\overset{(2)}{=} \lambda^m \left[\sigma_m(x,\hat{p}) + \frac{1}{\lambda}\,\sigma_{m-1}(x,\hat{p}) + \cdots + \frac{1}{\lambda^m}\,\sigma_0(x) \right]. \tag{2.44}$$

(1) Group the terms of (2.40) by their degree $k = |\alpha|$, and use $(\lambda\hat{p})^\alpha = \lambda^{|\alpha|}\hat{p}^{\,\alpha}$.
(2) Factor out λ^m; here $\sigma_k = \sum_{|\alpha|=k} a_\alpha p^\alpha$ is the degree-k part of the symbol.

As $\lambda \to \infty$ the bracket tends to $\sigma_m(x,\hat{p})$. So for a wave of very short wavelength in the direction $\hat{p}$, the operator acts like multiplication by $\lambda^m \sigma_m(x,\hat{p})$, whatever the lower-order terms are. *The leading symbol is the operator as seen by short waves.* There is one exception: if $\sigma_m(x,\hat{p}) = 0$ for some direction, the top order does not dominate in that direction, and the lower-order terms decide what happens. That is exactly the parabolic case, where the first-order term ∂_t took over in the direction where p_2^2 vanished (Figure 2.10).

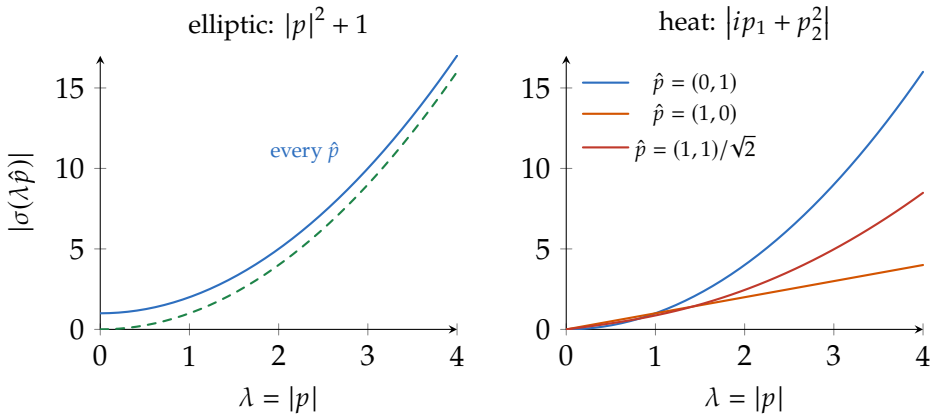


Figure 2.10: How the symbol grows along rays $p = \lambda\hat{p}$. Left: for the elliptic operator $-\nabla^2 + 1$ the symbol grows like λ^2 in every direction; the lower order term (+1, the gap between the blue curve and the dashed green λ^2) is irrelevant at large λ. Right: for the heat operator $\partial_t - \partial_x^2$ the growth is quadratic in most directions but only *linear* in the direction $\hat{p} = (1,0)$ (orange), where the leading symbol p_2^2 vanishes and the lower-order term ip_1 takes over.

The same scaling explains why the Green's functions in Example 2.13 all have the same singularity. The singularity of G at $x = 0$ comes from the large-$|p|$ part of $\int e^{ip\cdot x}/\sigma(p)\,\mathrm{d}^n p$, and there $1/\sigma \approx 1/\sigma_m$. For instance, for the Yukawa operator,

$$\frac{1}{|p|^2 + m^2} \overset{(1)}{=} \frac{1}{|p|^2} \cdot \frac{1}{1 + m^2/|p|^2} \overset{(2)}{=} \frac{1}{|p|^2} - \frac{m^2}{|p|^4} + \frac{m^4}{|p|^6} - \cdots \qquad (|p| > m),$$

(1) Factor out the leading symbol.
(2) Geometric series in $m^2/|p|^2 < 1$.

The first term gives the Coulomb singularity $1/(4\pi r)$. The second decays faster by two powers of p, so it produces a milder singularity (in $\mathbb{R}^3$, the inverse Fourier transform of $1/|p|^4$ is $-r/(8\pi)$, which is continuous). Indeed $e^{-mr}/(4\pi r) = 1/(4\pi r) - m/(4\pi) + m^2 r/(8\pi) - \cdots$: the Coulomb term, then a constant, then the r term. Up to the smooth constant term, this matches the series. Each lower-order correction is smoother than the one before. This "expansion in powers of $1/|p|$" is the germ of the symbol calculus of part 2b.

2.2.4 Ellipticity for general operators

We are ready for Nash's definition.

Working definition 2.9 (elliptic operator (Nash, p. 33))

The operator $L = \sum_{|\alpha|\le m} a_\alpha(x) D^\alpha$ of order m is **elliptic** if its leading symbol [$\sigma_m(x,p) = \sum_{|\alpha|=m} a_\alpha(x) p^\alpha$, from the terms with exactly m derivatives] never vanishes for non-zero p:

$$\sigma_m(x,p) \neq 0 \qquad \text{for every } x \text{ and every real } p \neq 0. \tag{2.45}$$

It is **elliptic at the point** x_0 if this holds at $x = x_0$, and elliptic on an open set Ω if it holds at every point of Ω. *How to test*: compute σ_m by replacing $\partial_j \to i p_j$ in the top-order terms only. Check that the resulting homogeneous polynomial has no real zeros other than $p = 0$. By homogeneity (2.43) it is enough to check the unit sphere $|p| = 1$.

For second-order operators with real coefficients this reproduces Section 2.1.9: $\sigma_2(x,p) = \sum a_{ij}(x) p_i p_j$ vanishes only at $p = 0$ exactly when the matrix $(a_{ij}(x))$ is definite. The new definition applies to every order, and to complex coefficients, and it is checked point by point.

Key idea: Ellipticity means "the symbol is large in every direction"

Because σ_m is continuous and non-zero on the unit sphere $|p| = 1$, which is compact, it has a positive minimum there, $|\sigma_m(x,\hat{p})| \ge c(x) > 0$. By homogeneity, $|\sigma_m(x,p)| \ge c(x)\,|p|^m$ for all p. So an elliptic operator of order m "costs" m powers of $|p|$ in *every* direction of momentum space. Inverting it gains back m powers in every direction. That is the source of elliptic regularity: *if Lu has k derivatives, then u has $k+m$.*

The lower-order terms cannot spoil this at high frequencies. At a fixed point x, bound the lower-order part of the symbol by $|\sigma(x,p) - \sigma_m(x,p)| \le C(1+|p|)^{m-1}$. Then

$$|\sigma(x,p)| \overset{(1)}{\ge} |\sigma_m(x,p)| - |\sigma(x,p) - \sigma_m(x,p)| \overset{(2)}{\ge} c\,|p|^m - C(1+|p|)^{m-1} \overset{(3)}{\ge} \tfrac{c}{2}\,|p|^m \qquad (|p| \ge R), \tag{2.46}$$

(1) The triangle inequality $|a| \ge |b| - |a-b|$.
(2) Ellipticity (the key idea above) for the first term; the lower-order bound for the second.
(3) $C(1+|p|)^{m-1}/|p|^m \to 0$ as $|p| \to \infty$, so it is below $c/2$ once $|p|$ exceeds some radius R.

So the *full* symbol of an elliptic operator can vanish only inside a bounded ball $|p| < R$. That does not matter for smoothness, which is about large $|p|$. The Yukawa example shows the full symbol need not vanish at all. For $-\nabla^2 - k^2$ (the Helmholtz operator), the full symbol $|p|^2 - k^2$ vanishes on the sphere $|p| = k$, which is bounded, so Helmholtz is elliptic. It gives the wave-number-k solutions of scattering theory, all smooth.

Ellipticity as invertibility

Nash remarks that one can think of ellipticity "as being characterised by the invertibility of the leading symbol of L, or the invertibility of L itself modulo lower order terms." For scalar operators "invertible" just means "non-zero". The point of the phrase becomes clear for matrix-valued symbols in Section 2.3. But the second half deserves an explanation now, because it is the idea behind the *parametrix* of part 2b.

Take L with constant coefficients for simplicity, and let Q be the operator that, in Fourier space, multiplies by

$$q(p) \;=\; \frac{1 - \chi(p)}{\sigma_m(p)},$$

where χ is a smooth bump equal to 1 near $p = 0$, inserted only to avoid dividing by zero at the

origin. Then LQ multiplies by

$$\sigma(p)\,q(p) \overset{(1)}{=} \frac{\sigma_m(p) + \big(\sigma(p) - \sigma_m(p)\big)}{\sigma_m(p)}\,\big(1 - \chi(p)\big) \overset{(2)}{=} 1 \;-\; \chi(p) \;+\; \frac{\sigma(p) - \sigma_m(p)}{\sigma_m(p)}\big(1 - \chi(p)\big).$$

(1) Split the full symbol into its top-order and lower-order parts.
(2) Multiply out.

The term $\chi(p)$ has compact support in p, so it corresponds to an operator that turns any function into a smooth one (a "smoothing" operator). The last term is of size $|p|^{m-1}/|p|^m = |p|^{-1}$ at large $|p|$: it is "of order -1", one order better than the identity. So $LQ = 1 + (\text{order} \le -1)$, and Q inverts L *up to lower-order terms*. This is possible precisely because σ_m can be divided by, which is ellipticity. Iterating the construction improves the error to order $-2, -3, \ldots$. The result, a **parametrix**, is an approximate inverse whose error is smoothing, and it is the rigorous replacement for the Green's function when the coefficients vary.

2.2.5 The leading symbol lives on the cotangent space

For the definition (2.45) to make sense on a manifold, ellipticity must not depend on the coordinates we choose. The full symbol does depend on the coordinates. The leading symbol does not, *provided we transform p as a covector*. This is the key to the bundle picture of Section 2.3, and Nash alludes to it in his footnote 4, so we work it out.

The oscillatory test. Replace the plane wave in (2.41) by a wave with a general phase, $e^{i\lambda\psi(x)}$, where ψ is a smooth real function and λ is a large parameter. Locally this is a plane wave with wave-vector $\lambda\,\nabla\psi(x)$. Differentiate once: $\partial_j e^{i\lambda\psi} = i\lambda(\partial_j\psi)e^{i\lambda\psi}$. Differentiate again:

$$\partial_j\partial_k e^{i\lambda\psi} \overset{(1)}{=} \partial_j\big(i\lambda\,\partial_k\psi\,e^{i\lambda\psi}\big) \overset{(2)}{=} \big[(i\lambda)^2\,\partial_j\psi\,\partial_k\psi + i\lambda\,\partial_j\partial_k\psi\big]e^{i\lambda\psi}.$$

(1) The first derivative, computed above.
(2) Product rule: the derivative hits either the exponential (giving another $i\lambda\,\partial_j\psi$) or the factor $\partial_k\psi$ (giving no new λ).

The pattern is general: each ∂_j gives $i\lambda\,\partial_j\psi$ plus terms with fewer powers of λ. For an operator of order m,

$$e^{-i\lambda\psi}\,L\big(e^{i\lambda\psi}\big) \;=\; \lambda^m\,\sigma_m\big(x, \mathrm{d}\psi(x)\big) + O(\lambda^{m-1}), \qquad \text{so} \qquad \sigma_m(x, \mathrm{d}\psi(x)) \;=\; \lim_{\lambda\to\infty} \lambda^{-m} e^{-i\lambda\psi} L\, e^{i\lambda\psi}, \tag{2.47}$$

where $\mathrm{d}\psi(x) = (\partial_1\psi, \ldots, \partial_n\psi)$ is the gradient covector. The right-hand side mentions no coordinates at all: just a function ψ, the operator, and a limit. So σ_m *is a well-defined function of the point x and of the covector $\xi = \mathrm{d}\psi(x) \in T^*_xM$.* Every covector at x is $\mathrm{d}\psi(x)$ for some ψ (take ψ linear in local coordinates), so this defines σ_m on the whole cotangent space.

The transformation law. In practice: if we change coordinates $x \mapsto y(x)$, the momentum variable must change like the components of a gradient. By the chain rule, $\partial\psi/\partial x_i = \sum_k (\partial y_k/\partial x_i)\,\partial\psi/\partial y_k$, so

$$p_i^{(x)} \;=\; \sum_{k=1}^{n} \frac{\partial y_k}{\partial x_i}\, p_k^{(y)}, \tag{2.48}$$

and $\sigma_m^{(x)}(x, p^{(x)}) = \sigma_m^{(y)}(y(x), p^{(y)})$. The "$p$" of the symbol is a momentum, and momenta are covectors, just as in Hamiltonian mechanics, where (x, p) are coordinates on phase space T^*M.

Worked example 2.14: The Laplacian in polar coordinates

In the plane the positive Laplacian in polar coordinates (r, θ) is

$$\Delta \;=\; -\frac{\partial^2}{\partial r^2} - \frac{1}{r}\frac{\partial}{\partial r} - \frac{1}{r^2}\frac{\partial^2}{\partial \theta^2}.$$

Its coefficients depend on r. Let (p_r, p_θ) be the momenta conjugate to (r, θ). The plane-wave test (replace $\partial_r \to i p_r$, $\partial_\theta \to i p_\theta$) gives the full and leading symbols

$$\sigma(r, \theta; p) \;=\; p_r^2 - \frac{i}{r}\, p_r + \frac{p_\theta^2}{r^2}, \qquad \sigma_2(r, \theta; p) \;=\; p_r^2 + \frac{p_\theta^2}{r^2}. \tag{2.49}$$

Now check that σ_2 is the Cartesian symbol $p_x^2 + p_y^2$, transformed as a covector. With $x = r\cos\theta$, $y = r\sin\theta$, the law (2.48) (with the roles of old and new coordinates chosen so that (r, θ) are the "x") gives

$$\begin{aligned} p_r &\overset{(1)}{=} \frac{\partial x}{\partial r} p_x + \frac{\partial y}{\partial r} p_y \;=\; \cos\theta\, p_x + \sin\theta\, p_y, \\ p_\theta &\overset{(2)}{=} \frac{\partial x}{\partial \theta} p_x + \frac{\partial y}{\partial \theta} p_y \;=\; -r\sin\theta\, p_x + r\cos\theta\, p_y. \end{aligned}$$

(1) (2.48) with $i = r$; $\partial_r(r\cos\theta) = \cos\theta$, $\partial_r(r\sin\theta) = \sin\theta$.
(2) (2.48) with $i = \theta$; $\partial_\theta(r\cos\theta) = -r\sin\theta$, $\partial_\theta(r\sin\theta) = r\cos\theta$.

Therefore

$$\begin{aligned} p_r^2 + \frac{p_\theta^2}{r^2} &\overset{(1)}{=} (\cos\theta\, p_x + \sin\theta\, p_y)^2 + (-\sin\theta\, p_x + \cos\theta\, p_y)^2 \\ &\overset{(2)}{=} (\cos^2\theta + \sin^2\theta)\, p_x^2 + (\sin^2\theta + \cos^2\theta)\, p_y^2 \\ &\qquad + 2(\cos\theta\sin\theta - \sin\theta\cos\theta)\, p_x p_y \;=\; p_x^2 + p_y^2. \end{aligned}$$

(1) Substitute the two formulas above; the factor r in p_θ cancels the $1/r^2$.
(2) Expand both squares; the cross terms cancel.

So the leading symbol is coordinate independent, as promised. The full symbol is not. In Cartesian coordinates it has no first-order part. In polar coordinates it has the first-order part $-ip_r/r$, produced by the change of coordinates. Notice also that σ_2 in (2.49) is positive for $(p_r, p_\theta) \neq 0$ only when $r > 0$. At $r = 0$ it degenerates to p_r^2. This is not a failure of ellipticity: polar coordinates are singular at the origin, and the "operator in coordinates" there is not the Laplacian in any honest chart.

Worked example 2.15: The Laplace–Beltrami operator of a curved metric

On a Riemannian manifold with metric $g_{ij}(x)$, inverse metric g^{ij} and $\sqrt{g} = \sqrt{\det g_{ij}}$, the positive Laplacian on functions is

$$\Delta_g u \;=\; -\frac{1}{\sqrt{g}}\,\partial_i\big(\sqrt{g}\, g^{ij}\, \partial_j u\big) \;=\; -g^{ij}\,\partial_i\partial_j u \;-\; \frac{1}{\sqrt{g}}\,\partial_i\big(\sqrt{g}\, g^{ij}\big)\,\partial_j u, \tag{2.50}$$

with summation over repeated indices (the second form is the product rule). Only the first term has two derivatives, so

$$\sigma_2(\Delta_g)(x, p) \;=\; g^{ij}(x)\, p_i p_j \;=\; |p|_g^2, \tag{2.51}$$

the squared length of the covector p in the metric. A Riemannian metric is positive definite, so $g^{ij}p_ip_j > 0$ for $p \neq 0$: *the Laplacian of any Riemannian metric is elliptic.* For the round unit sphere, $g = \mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2$, so $g^{\theta\theta} = 1$, $g^{\phi\phi} = 1/\sin^2\theta$ and $\sigma_2 = p_\theta^2 + p_\phi^2/\sin^2\theta$. For a Lorentzian metric the same formula gives $g^{\mu\nu}p_\mu p_\nu$, which vanishes on the null cone: the wave operator of general relativity is hyperbolic, as it must be.

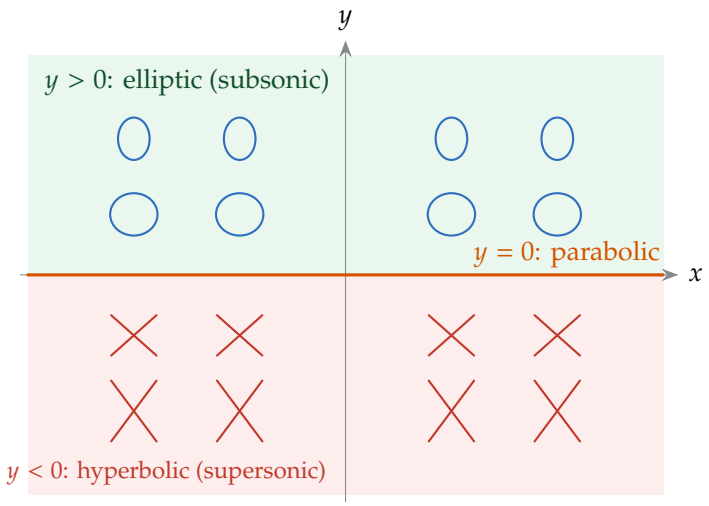


Figure 2.11: Ellipticity is checked point by point. For the Tricomi operator $-(y\partial_x^2 + \partial_y^2)$ we draw, at sample points, the level set $\sigma_2(x, y; p) = yp_1^2 + p_2^2 = c$ in the momentum plane attached to that point (a small copy of p-space at each point). In the upper half-plane these are ellipses, which get wider as $y \to 0$. In the lower half-plane the leading symbol vanishes on a pair of null directions (red crosses). They open up as y becomes more negative, and they are the directions normal to the Mach lines. On the sonic line $y = 0$ the form is degenerate.

2.2.6 More examples of the definition at work

Worked example 2.16: The biharmonic operator: elliptic of order four

The deflection $u(x, y)$ of a thin elastic plate under a load f obeys $\Delta^2 u = f$, with

$$\Delta^2 = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)^2 = \frac{\partial^4}{\partial x^4} + 2\frac{\partial^4}{\partial x^2 \partial y^2} + \frac{\partial^4}{\partial y^4}.$$

In multi-index form, $\partial^\alpha = i^{|\alpha|}D^\alpha$ with $|\alpha| = 4$ gives $i^4 = 1$, so $\Delta^2 = D^{(4,0)} + 2D^{(2,2)} + D^{(0,4)}$ and

$$\sigma_4(p) \overset{(1)}{=} p_1^4 + 2p_1^2p_2^2 + p_2^4 \overset{(2)}{=} (p_1^2 + p_2^2)^2 = |p|^4.$$

(1) Replace each D^α by p^α.
(2) Recognise a perfect square.

This vanishes only at $p = 0$, so Δ^2 is elliptic. It is also the square of an elliptic operator, and indeed the leading symbol of a composition is the product of leading symbols (we derive this in Section 2.3.8). By contrast, the fourth-order operator $\partial_x^4 - \partial_y^4$ has $\sigma_4 = p_1^4 - p_2^4 = (p_1^2 - p_2^2)(p_1^2 + p_2^2)$, which vanishes on $p_1 = \pm p_2$. It is the product of an elliptic and a hyperbolic operator, and it is not elliptic.

Worked example 2.17: Why a scalar first-order operator is almost never elliptic

Take a first-order scalar operator $L = \sum_j a_j(x)\partial_j + b(x)$ on $\mathbb{R}^n$ with complex coefficients. Its leading symbol is

$$\sigma_1(x,p) \;=\; i\sum_{j=1}^{n} a_j(x)\,p_j \;=\; i\,(\vec{u}\cdot p + i\,\vec{v}\cdot p),$$

where we split the complex vector $a = \vec{u} + i\vec{v}$ into real and imaginary parts $\vec{u}, \vec{v} \in \mathbb{R}^n$. For real p, $\vec{u}\cdot p$ and $\vec{v}\cdot p$ are real, so $\sigma_1 = 0$ exactly when $\vec{u}\cdot p = 0$ *and* $\vec{v}\cdot p = 0$. That is, p is orthogonal to both $\vec{u}$ and $\vec{v}$. Now count:

- $n = 1$: $\sigma_1 = ia\,p$ is non-zero for $p \neq 0$ iff $a \neq 0$. Every non-degenerate first-order ODE operator is elliptic.
- $n = 2$: two vectors $\vec{u}, \vec{v}$ in the plane have no common non-zero orthogonal vector iff they are linearly independent. For example $a = (1, i)$, i.e. $\vec{u} = (1, 0)$, $\vec{v} = (0, 1)$, which is $L = \partial_x + i\partial_y = 2\bar{\partial}$. So *the Cauchy–Riemann operator is elliptic*, and up to a change of coordinates it is the only type of elliptic scalar first-order operator in the plane.
- $n \geq 3$: the vectors orthogonal to both $\vec{u}$ and $\vec{v}$ form a subspace of dimension at least $n - 2 \geq 1$. There is always a non-zero characteristic p. *No scalar first-order operator in three or more dimensions is elliptic.*

A real first-order operator (such as the transport operator $\partial_t + c\,\partial_x$) is never elliptic for $n \geq 2$: its symbol $i(p_t + cp_x)$ vanishes on a whole line. Transport is "pure propagation". This counting is the reason the Dirac operator is a *matrix* operator: to take a first-order "square root" of the Laplacian in $n \geq 3$ dimensions we are forced to use matrices (gamma matrices). The next section builds the framework, vector bundles, in which such operators live.

Try it: Ellipticity of a few operators

Decide whether each operator on $\mathbb{R}^2$ is elliptic, and where. (a) $-\partial_x^2 - 4\partial_x\partial_y - 5\partial_y^2 + 7\partial_x$. (b) $-\partial_x^2 - (1 + x^2)\partial_y^2$. (c) $\partial_x^2\partial_y^2$. (d) $\partial_x + ix\,\partial_y$.

Answers. (a) $\sigma_2 = p_1^2 + 4p_1p_2 + 5p_2^2 = (p_1 + 2p_2)^2 + p_2^2 > 0$ for $p \neq 0$: elliptic (the $7\partial_x$ is irrelevant). (b) $\sigma_2 = p_1^2 + (1 + x^2)p_2^2 > 0$: elliptic everywhere. (c) $\sigma_4 = p_1^2p_2^2$ vanishes on both axes: not elliptic. (d) $\sigma_1 = i(p_1 + ixp_2)$, which vanishes for $p = (0, 1)$ exactly when $x = 0$: elliptic for $x \neq 0$, not elliptic on the line $x = 0$. (This is the Mizohata operator, a famous source of counterexamples in PDE theory.)

2.2.7 Hypo-ellipticity: smoothing without ellipticity

In Section 2.1 we saw that the benefit of ellipticity is elliptic regularity: $Lu = f$ with f smooth forces u to be smooth. Is ellipticity *necessary* for this? No. Some parabolic operators, the heat operator above all, smooth just as well. So Nash gives the regularity property its own name.

Working definition 2.10 (hypo-elliptic operator (Nash, p. 33))

A differential operator L is **hypo-elliptic** if, for every open set Ω and every (generalised) function [possibly as rough as a point source δ] u,

$$Lu \text{ is } C^\infty \text{ on } \Omega \quad\Longrightarrow\quad u \text{ is } C^\infty \text{ on } \Omega. \tag{2.52}$$

Operationally: solutions are never rougher than the source, locally. In the language of (2.19), $\operatorname{sing\,supp} u = \operatorname{sing\,supp} Lu$ [the singular support is the closed set of points near which the function is not smooth]. *How to test* (constant coefficients): find the Green's function G and check whether it is C^∞ away from the origin. If it is, the localisation argument of Section 2.1.4 (Step 2) goes through word for word, and L is hypo-elliptic.

The inclusion one way is what Section 2.1.4 showed.

Key idea: Elliptic ⇒ hypo-elliptic, not conversely

Every elliptic operator is hypo-elliptic: this is elliptic regularity. For variable coefficients the proof replaces the Green's function by the parametrix of Section 2.2.4, an approximate inverse whose error is smoothing. But there are hypo-elliptic operators that are not elliptic. Nash's example is the heat operator.

2.2.8 The heat operator: hypo-elliptic but not elliptic

The heat (diffusion) operator on $\mathbb{R}^{n+1}$, with coordinates $(x,t) = (x_1,\dots,x_n,t)$, is

$$L \;=\; \frac{\partial}{\partial t} - \frac{\partial^2}{\partial x_1^2} - \cdots - \frac{\partial^2}{\partial x_n^2} \;=\; \frac{\partial}{\partial t} + \Delta_x, \tag{2.53}$$

Nash's (2.24). Its leading symbol, in the variables (p,p_0) dual to (x,t), is $\sigma_2 = |p|^2$. It does not involve p_0 at all, so it vanishes at $(p,p_0) = (0,1) \neq 0$. The heat operator is *not* elliptic, which is what we called parabolic in Section 2.1.9. Yet it smooths. We see this by computing its Green's function.

Worked example 2.18: The heat kernel, Nash's (2.25)

We want G with $LG = \delta(x)\delta(t)$ and $G = 0$ for $t < 0$ (the heat released by a unit burst at the origin at time 0). Fourier transform in x only: $G(x,t) = (2\pi)^{-n}\int e^{ip\cdot x} g(p,t)\,\mathrm{d}^n p$. Then $-\partial_{x_j}^2 \to p_j^2$ and $\delta(x) \to 1$ in these units, so for each p

$$\frac{\partial g}{\partial t} + |p|^2\, g \;=\; \delta(t), \qquad g = 0 \text{ for } t<0, \qquad \text{hence} \qquad g(p,t) = \theta(t)\, e^{-|p|^2 t}.$$

(Check: for $t>0$ this solves $g' = -|p|^2 g$, and the jump of g from 0 to 1 at $t=0$ produces the $\delta(t)$.) Each Fourier mode simply decays, the short wavelengths fastest. Transforming back, the integral factorises into n one-dimensional Gaussian integrals. For $t>0$ and one coordinate,

$$\int_{-\infty}^{\infty} e^{ipx - p^2 t}\,\mathrm{d}p \overset{(1)}{=} \int_{-\infty}^{\infty} e^{-t\,(p - \frac{ix}{2t})^2 - \frac{x^2}{4t}}\,\mathrm{d}p \overset{(2)}{=} e^{-x^2/4t}\int_{-\infty}^{\infty} e^{-tq^2}\,\mathrm{d}q \overset{(3)}{=} \sqrt{\frac{\pi}{t}}\; e^{-x^2/4t}.$$

(1) Complete the square: $-t\big(p - \frac{ix}{2t}\big)^2 = -tp^2 + ipx + \frac{x^2}{4t}$, so $ipx - tp^2 = -t\big(p - \frac{ix}{2t}\big)^2 - \frac{x^2}{4t}$.
(2) Shift the contour, $q = p - ix/2t$. This is allowed by Cauchy's theorem because the Gaussian decays at both ends of the strip between $\operatorname{Im} q = 0$ and $\operatorname{Im} q = -x/2t$.
(3) The Gaussian integral $\int e^{-tq^2}\mathrm{d}q = \sqrt{\pi/t}$.

Multiplying n of these and the prefactor $(2\pi)^{-n}$,

$$G(x,t) \;=\; \theta(t)\,\frac{1}{(2\pi)^n}\Big(\frac{\pi}{t}\Big)^{n/2} e^{-|x|^2/4t} \;=\; \frac{\theta(t)}{(4\pi t)^{n/2}}\,\exp\Big[-\frac{\sum_i x_i^2}{4t}\Big], \tag{2.54}$$

which is Nash's (2.25), since $(4\pi t)^{n/2} = (2\sqrt{\pi t})^n$. It is a Gaussian of width $\sqrt{2t}$ and total heat $\int G\,\mathrm{d}^n x = 1$ for every $t > 0$.
As a direct check that $LG = 0$ for $t > 0$, write $r^2 = |x|^2$:

$$\frac{\partial G}{\partial t} \overset{(1)}{=} \Big(-\frac{n}{2t} + \frac{r^2}{4t^2}\Big)G, \qquad \frac{\partial^2 G}{\partial x_j^2} \overset{(2)}{=} \Big(-\frac{1}{2t} + \frac{x_j^2}{4t^2}\Big)G,$$

(1) Differentiate $(4\pi t)^{-n/2}$ (giving $-n/2t$ times itself) and the exponent $-r^2/4t$ (giving $+r^2/4t^2$).
(2) $\partial_j G = -\frac{x_j}{2t}G$; differentiate again with the product rule.

Summing the second formula over j gives $\nabla^2 G = (-\frac{n}{2t} + \frac{r^2}{4t^2})G = \partial_t G$, so $(\partial_t - \nabla^2)G = 0$ for $t > 0$.

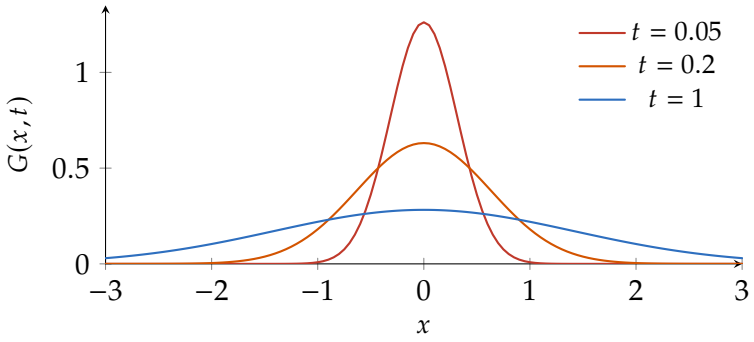


Figure 2.12: The one-dimensional heat kernel (2.54) at three times. A unit burst of heat at the origin spreads into Gaussians of width $\sqrt{2t}$ and constant area 1. For every $t > 0$ the profile is C^∞ (indeed analytic) in x. As $t \to 0^+$ it concentrates into $\delta(x)$.

Why the heat kernel is smooth away from the origin

Nash's key claim is that G in (2.54) is singular *only* at the origin $x = t = 0$ of $\mathbb{R}^{n+1}$. For $t > 0$ it is obviously smooth. For $t < 0$ it is identically zero. The only question is the hyperplane $t = 0$ with $x \neq 0$, where the two formulas meet. Every partial derivative of G for $t > 0$ has the form

$$\partial_x^\beta \partial_t^k G(x,t) \;=\; P\big(x, t^{-1}\big)\, t^{-n/2}\, e^{-|x|^2/4t}, \tag{2.55}$$

where P is a polynomial. (Each derivative either hits the prefactor, producing a power of $1/t$, or hits the exponential, producing a factor like x_j/t or $|x|^2/t^2$.) Fix $x \neq 0$ and let $t \to 0^+$. With $s = 1/t \to \infty$ and $a = |x|^2/4 > 0$, every term is a power of s times e^{-as}, and

$$\lim_{t\to 0^+} t^{-N} e^{-a/t} \overset{(1)}{=} \lim_{s\to\infty} s^N e^{-as} \overset{(2)}{=} \lim_{s\to\infty} \frac{s^N}{1 + as + \cdots + \frac{(as)^{N+1}}{(N+1)!} + \cdots} \;=\; 0.$$

(1) Substitute $s = 1/t$.
(2) e^{as} is its Taylor series, every term of which is positive; keeping only the term of degree $N+1$ in the denominator shows the ratio is at most $(N+1)!/(a^{N+1}s) \to 0$.

So every derivative of G tends to 0 as $t \to 0^+$ at fixed $x \neq 0$, matching the derivatives of the zero function on the other side. Hence G is C^∞ at every point with $x \neq 0$ (Nash: "all other partial derivatives of $G(x,t)$ converge (strongly) for $x \neq 0$ and $t > 0$, and vanish (because of the

$\theta(t)$) for $x \neq 0$ and $t < 0$"). At $x = 0$, however, $G(0,t) = (4\pi t)^{-n/2} \to \infty$: the origin is the only singular point. The same proof as in Section 2.1.4 then shows the heat operator is hypo-elliptic.

Physical picture: Diffusion kills high frequencies

In Fourier space the heat evolution is $\hat{u}(p,t) = e^{-|p|^2 t}\hat{u}(p,0)$. For any $t > 0$ the factor $e^{-|p|^2 t}$ decays faster than every power of $|p|$. By the principle "smoothness is decay in p", *any* initial temperature profile, however rough, is C^∞ in x an instant later. A step function becomes the error function $\frac{1}{2}[1 + \operatorname{erf}(x/2\sqrt{t})]$. A kink rounds off at once. This is exactly the opposite of Example 2.8, where the wave equation kept kinks forever. The price is irreversibility: running the heat equation backwards multiplies by $e^{+|p|^2 t}$, which blows up the high frequencies, and the backward problem is ill-posed.

Worked example 2.19: A function that is smooth but not analytic: $e^{-1/t}$

The one-variable function behind the heat kernel is

$$h(t) \;=\; \begin{cases} e^{-1/t}, & t > 0, \\ 0, & t \le 0. \end{cases}$$

The argument above (with $a = 1$) shows that every derivative $h^{(k)}(t)$ tends to 0 as $t \to 0^+$. So h is C^∞ on all of $\mathbb{R}$, with $h^{(k)}(0) = 0$ for every k. Its Taylor series at 0 is therefore

$$\sum_{k=0}^{\infty} \frac{h^{(k)}(0)}{k!}\, t^k \;=\; 0 + 0 + 0 + \cdots \;=\; 0,$$

which converges (to zero) but does not equal $h(t) = e^{-1/t} > 0$ for any $t > 0$. A function is **real analytic** at a point if it equals its convergent Taylor series in a neighbourhood of that point. So h is smooth but *not* analytic at $t = 0$. For each fixed $x \neq 0$ the heat kernel, as a function of t, is a copy of this function: $G(x,t) = (4\pi t)^{-n/2} h(4t/|x|^2)$. This is the content of Nash's footnote 3: G is C^∞ but not analytic on the hyperplane $t = 0$, and analytic off it.

2.2.9 The Schrödinger operator: not hypo-elliptic

The free Schrödinger operator (with $\hbar = 2m = 1$) in n space dimensions,

$$L \;=\; -i\frac{\partial}{\partial t} - \nabla^2, \tag{2.56}$$

looks like the heat operator with t replaced by it. Both are parabolic with the same leading symbol $|p|^2$. Nash shows that, despite this similarity, the Schrödinger operator is *not* hypo-elliptic. Again the Green's function tells the story.

Worked example 2.20: The Schrödinger Green's function, Nash's (2.26)

Fourier transform in x as before. The mode equation is $-i\,\partial_t g + |p|^2 g = \delta(t)$ with $g = 0$ for $t < 0$. Its solution is $g(p,t) = i\,\theta(t)\,e^{-i|p|^2 t}$. To check it,

$$\begin{aligned} -i\frac{\partial g}{\partial t} + |p|^2 g &\overset{(1)}{=} -i \cdot i\Big[\delta(t)\,e^{-i|p|^2 t} - i\,|p|^2\;\theta(t)\,e^{-i|p|^2 t}\Big] + i\,|p|^2\;\theta(t)\,e^{-i|p|^2 t} \\ &\overset{(2)}{=} \delta(t) - i\,|p|^2\;\theta(t)\,e^{-i|p|^2 t} + i\,|p|^2\;\theta(t)\,e^{-i|p|^2 t} \;=\; \delta(t). \end{aligned}$$

(1) Product rule on $\theta(t)e^{-i|p|^2 t}$, using $\theta' = \delta$.

(2) $-i \cdot i = 1$, and $\delta(t)e^{-i|p|^2 t} = \delta(t)$ because the exponential equals 1 at $t = 0$.

Transforming back uses the same Gaussian integral as Example 2.18 with t replaced by it (justified by analytic continuation, or by inserting a damping factor $e^{-\varepsilon|p|^2}$ and letting $\varepsilon \to 0$):

$$G(x,t) \;=\; i\,\frac{\theta(t)}{(4\pi i t)^{n/2}}\,\exp\!\Big[-\frac{\sum_j x_j^2}{4it}\Big] \;=\; i\,\frac{\theta(t)}{(4\pi i t)^{n/2}}\,e^{\,i|x|^2/4t}, \tag{2.57}$$

Nash's (2.26). This is the free-particle propagator of quantum mechanics.

Now compare the two kernels on the hyperplane $t = 0$, $x \neq 0$. For the heat kernel the exponential factor is $e^{-|x|^2/4t}$. It is real and tends to zero faster than any power of t, which is what made everything smooth. For the Schrödinger kernel the exponential factor is $e^{i|x|^2/4t}$, a pure phase of modulus one. Then

$$|G(x,t)| \;=\; \frac{1}{(4\pi t)^{n/2}} \quad (t > 0),$$

which blows up as $t \to 0^+$ at *every* x, while the phase $|x|^2/4t$ spins round infinitely often. Nash says it "behaves monstrously" on $t = 0$. G is not even continuous there, let alone C^∞ (Figure 2.13).

The failure of hypo-ellipticity follows at once. Take any point $(x_0, 0)$ with $x_0 \neq 0$ and a small ball Ω around it not containing the origin. On Ω, $LG = \delta(x)\delta(t) = 0$, which is certainly smooth. But G is not smooth on Ω. So (2.52) fails: *the Schrödinger operator is not hypo-elliptic.*

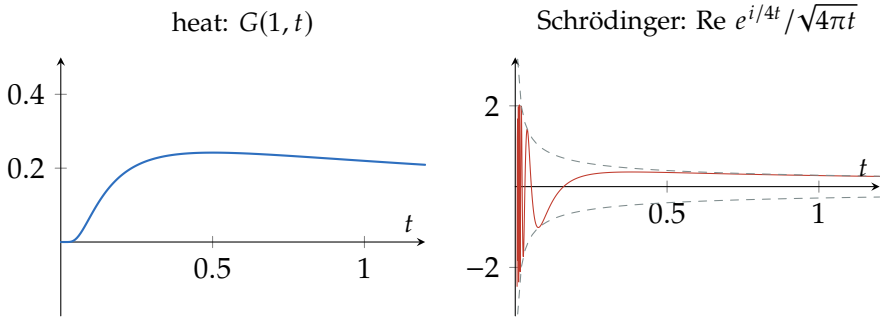


Figure 2.13: The two Green's functions at the fixed position $x = 1$ (one space dimension), as functions of time. Left: the heat kernel $e^{-1/4t}/\sqrt{4\pi t}$ leaves $t = 0$ so flatly that every derivative vanishes there: it is C^∞ (though not analytic). Right: (the real part of) the Schrödinger kernel oscillates infinitely fast inside an envelope $\pm 1/\sqrt{4\pi t}$ (dashed) that blows up as $t \to 0^+$: it is not even continuous at $t = 0$. This is why the heat operator is hypo-elliptic and the Schrödinger operator is not.

Physical picture: Dispersion moves high frequencies, it does not kill them

In Fourier space, Schrödinger evolution is $\hat{u}(p,t) = e^{-i|p|^2 t}\hat{u}(p,0)$. The multiplier has modulus one, so $|\hat{u}(p,t)| = |\hat{u}(p,0)|$: the high-frequency content is never damped. The phase makes a wave-packet of momentum p move with group velocity $2p$. So short waves are not destroyed, only sent away very fast. Time reversal is a symmetry ($t \to -t$ with complex conjugation), so short waves can equally well arrive from far away and focus to a singularity at a chosen point and time. The Green's function is exactly such a focused wave. The heat equation has no such symmetry: it damps short waves, and nothing can

reassemble them.

Watch out: Non-hypo-elliptic does not mean "never smooths"

Free Schrödinger evolution of nice, rapidly decaying initial data is in fact smooth for $t \neq 0$: the roughness disperses off to infinity. (Physicists call this dispersive smoothing.) This uses information about the solution *at infinity*. Hypo-ellipticity is a *local* property: it must hold on every small open set, whatever the solution does elsewhere. The Green's function shows that locally, near $(x_0, 0)$, a solution of $Lu = 0$ can be as singular as we like. By contrast, for the heat equation every local solution is smooth.

Hörmander's test for constant coefficients

There is a neat criterion, due to Hörmander, that explains all these examples at once in terms of the *full* symbol. We state it because it makes the difference between heat and Schrödinger visible in one line. A constant-coefficient operator L with symbol $\sigma(p)$ is hypo-elliptic if and only if

$$\frac{\left|\partial_p^\gamma \sigma(p)\right|}{|\sigma(p)|} \longrightarrow 0 \qquad \text{as } |p| \to \infty,\ p \text{ real, for every } \gamma \neq 0. \tag{2.58}$$

In words: at high frequency, the symbol is much larger than all its derivatives. This is the property (2.18) that made the Green's function smooth: each p-derivative of $1/\sigma$ gains decay. In particular σ must be non-zero for large $|p|$.

- **Laplace**, $\sigma = |p|^2$: $\partial_j \sigma / \sigma = 2p_j/|p|^2 = O(1/|p|) \to 0$. Hypo-elliptic.
- **Heat**, $\sigma = ip_0 + |p|^2$ (with p_0 dual to t), so $|\sigma| = \sqrt{p_0^2 + |p|^4}$. Then $\left|\partial_{p_0}\sigma\right|/|\sigma| = 1/\sqrt{p_0^2 + |p|^4} \to 0$. Also $\left|\partial_{p_j}\sigma\right|/|\sigma| = 2\left|p_j\right|/\sqrt{p_0^2 + |p|^4}$. If $|p|^2 \geq |p_0|$ this is at most $2/|p|$, and if $|p|^2 < |p_0|$ it is at most $2\,|p_0|^{1/2}/|p_0|$; either way it tends to 0. Hypo-elliptic, even though the leading symbol vanishes at $(p, p_0) = (0, 1)$.
- **Schrödinger**, $\sigma = p_0 + |p|^2$: it vanishes on the whole paraboloid $p_0 = -|p|^2$, which goes off to infinity. The ratio is infinite there. Not hypo-elliptic.
- **Wave**, $\sigma = -p_0^2 + |p|^2$: vanishes on the light cone. Not hypo-elliptic.

The factor i in front of p_0 is the whole difference between heat and Schrödinger. It moves the zero set of the full symbol off the real p-space.

2.2.10 The Cauchy–Riemann operator

Nash's last example of the section is the most important one for later chapters: the operator whose solutions are the holomorphic functions. On $\mathbb{C} = \mathbb{R}^2$ with $z = x + iy$ set

$$L = \bar\partial = \frac{\partial}{\partial \bar z} = \frac{1}{2}\Big(\frac{\partial}{\partial x} + i\frac{\partial}{\partial y}\Big), \qquad \partial = \frac{\partial}{\partial z} = \frac{1}{2}\Big(\frac{\partial}{\partial x} - i\frac{\partial}{\partial y}\Big), \tag{2.59}$$

Nash's (2.27). These are defined so that $\partial z = 1$, $\bar\partial z = 0$, $\partial \bar z = 0$, $\bar\partial \bar z = 1$. Check one: $\bar\partial z = \frac{1}{2}(\partial_x + i\partial_y)(x + iy) = \frac{1}{2}(1 + i \cdot i) = 0$.

Holomorphic means $\bar\partial f = 0$. Write $f = u + iv$ with u, v real. Then

$$2\bar\partial f \overset{(1)}{=} (\partial_x + i\partial_y)(u + iv) \overset{(2)}{=} (\partial_x u - \partial_y v) + i\,(\partial_y u + \partial_x v).$$

(1) Definition (2.59).
(2) Multiply out and collect real and imaginary parts, using $i \cdot i = -1$.

So $\bar\partial f = 0$ is the pair of Cauchy–Riemann equations $\partial_x u = \partial_y v$, $\partial_y u = -\partial_x v$: f is holomorphic.

$\bar\partial$ is a square root of the Laplacian. Compose the two operators:

$$4\,\partial\bar\partial \overset{(1)}{=} (\partial_x - i\partial_y)(\partial_x + i\partial_y) \overset{(2)}{=} \partial_x^2 + i\partial_x\partial_y - i\partial_y\partial_x + \partial_y^2 \overset{(3)}{=} \partial_x^2 + \partial_y^2 \;=\; -\Delta. \tag{2.60}$$

(1) Definitions (2.59); the factors $\frac12 \cdot \frac12 \cdot 4 = 1$.
(2) Multiply out; $(-i)(i) = 1$.
(3) Mixed partials commute, so the middle terms cancel.

This is the two-dimensional version of what the Dirac operator does in any dimension: a first-order operator whose "square" is the Laplacian.

Worked example 2.21: The Green's function of $\bar\partial$ is $1/\pi z$, Nash's (2.28)–(2.29)

Nash states that

$$\bar\partial\, G \;=\; \delta(x,y) \qquad \text{is solved by} \qquad G(x,y) \;=\; \frac{1}{\pi(x+iy)} \;=\; \frac{1}{\pi z}. \tag{2.61}$$

We give two derivations.
From the Laplacian. The Green's function of Δ in the plane is $G_2 = -\frac{1}{2\pi}\ln|z|$ (2.16). By (2.60), $\Delta G_2 = \delta$ means $\bar\partial(-4\,\partial G_2) = \delta$. So $G = -4\partial G_2$ works, and

$$-4\,\partial G_2 \overset{(1)}{=} \frac{4}{2\pi}\,\partial\Big(\tfrac12 \ln(z\bar z)\Big) \overset{(2)}{=} \frac{1}{\pi}\cdot\frac{\bar z}{z\bar z} \;=\; \frac{1}{\pi z}.$$

(1) $\ln|z| = \frac12 \ln|z|^2 = \frac12 \ln(z\bar z)$.
(2) ∂ treats $\bar z$ as a constant, so $\partial \ln(z\bar z) = \bar z/(z\bar z)$. With the prefactors: $\frac{4}{2\pi}\cdot\frac12 = \frac{1}{\pi}$.

From Stokes' theorem. Test against a smooth compactly supported function φ. By definition of the derivative of a distribution, $\langle \bar\partial G, \varphi\rangle = -\int G\,\bar\partial\varphi\,\mathrm{d}x\,\mathrm{d}y$. Remove a small disc D_ε around the singularity. On the rest, $1/z$ is holomorphic, so $\frac{1}{z}\bar\partial\varphi = \bar\partial(\varphi/z)$. We also need $\mathrm{d}z \wedge \mathrm{d}\bar z$. Since $\mathrm{d}\bar z \wedge \mathrm{d}z = (\mathrm{d}x - i\mathrm{d}y)\wedge(\mathrm{d}x + i\mathrm{d}y) = 2i\,\mathrm{d}x\wedge\mathrm{d}y$, we have $\mathrm{d}(\psi\,\mathrm{d}z) = \bar\partial\psi\,\mathrm{d}\bar z\wedge\mathrm{d}z = 2i\,\bar\partial\psi\,\mathrm{d}x\wedge\mathrm{d}y$ for any function ψ. Hence

$$\begin{aligned} -\int_{|z|>\varepsilon} \frac{1}{\pi z}\,\bar\partial\varphi\,\mathrm{d}x\,\mathrm{d}y &\overset{(1)}{=} -\frac{1}{\pi}\int_{|z|>\varepsilon} \bar\partial\Big(\frac{\varphi}{z}\Big)\mathrm{d}x\,\mathrm{d}y \overset{(2)}{=} -\frac{1}{2\pi i}\int_{|z|>\varepsilon} \mathrm{d}\Big(\frac{\varphi}{z}\,\mathrm{d}z\Big) \\ &\overset{(3)}{=} +\frac{1}{2\pi i}\oint_{|z|=\varepsilon} \frac{\varphi\,\mathrm{d}z}{z} \;\xrightarrow{\varepsilon\to 0}\; \frac{1}{2\pi i}\,\varphi(0)\cdot 2\pi i \;=\; \varphi(0). \end{aligned}$$

(1) $1/z$ is holomorphic away from 0, so $\bar\partial$ passes through it.
(2) The identity $\mathrm{d}(\psi\,\mathrm{d}z) = 2i\,\bar\partial\psi\,\mathrm{d}x\wedge\mathrm{d}y$ with $\psi = \varphi/z$.
(3) Stokes' theorem. The region $|z| > \varepsilon$ has as its boundary the small circle traversed *clockwise*, which gives the sign flip. (φ vanishes far away.) Then $\oint \mathrm{d}z/z = 2\pi i$ counter-clockwise and $\varphi \to \varphi(0)$ on the shrinking circle.

So $\langle\bar\partial G, \varphi\rangle = \varphi(0)$, i.e. $\bar\partial G = \delta$.

The leading symbol. Using $\partial_x \to ip_1$, $\partial_y \to ip_2$,

$$\sigma_1(\bar\partial)(z,p) \;=\; \frac{1}{2}\big(ip_1 + i\cdot ip_2\big) \;=\; \frac{i}{2}\,(p_1+ip_2). \tag{2.62}$$

Its modulus is $\frac{1}{2}\,|p|$, which vanishes only at $p=0$. So $\bar\partial$ is elliptic, as Example 2.17 predicted. (Nash's (2.30) reads $\frac{1}{2}(p_1+ip_2)$, without the factor i. The factor is a matter of convention and does not affect ellipticity.) Consistently with (2.60), $4\,\sigma_1(\partial)\sigma_1(\bar\partial) = 4\cdot\frac{i}{2}(p_1-ip_2)\cdot\frac{i}{2}(p_1+ip_2) = -\,|p|^2$, which is the symbol of $-\Delta$.

Elliptic regularity for $\bar\partial$ is Cauchy's integral formula. The Green's function $1/\pi z$ is smooth, indeed real analytic, away from $z=0$. So by the argument of Section 2.1.4, $\bar\partial u = f$ with f smooth forces u to be smooth. For $f=0$ this is the classical fact that holomorphic functions are infinitely differentiable, and the localisation argument of Step 2 becomes Cauchy's integral formula. If $\bar\partial u = 0$ on a neighbourhood of the closed disc $|w|\le R$, then for $|z|<R$,

$$u(z) \overset{(1)}{=} \frac{1}{2\pi i}\oint_{|w|=R}\frac{u(w)}{w-z}\,\mathrm{d}w \overset{(2)}{=} \frac{1}{2\pi i}\oint_{|w|=R}\frac{u(w)}{w}\sum_{k=0}^{\infty}\Big(\frac{z}{w}\Big)^k\mathrm{d}w \overset{(3)}{=} \sum_{k=0}^{\infty} c_k\,z^k. \tag{2.63}$$

(1) Cauchy's integral formula. It is the Stokes computation of Example 2.21, run with the Green's function $1/\pi(w-z)$ centred at z and with the boundary term on $|w|=R$ kept.

(2) $\frac{1}{w-z} = \frac{1}{w}\,\frac{1}{1-z/w}$, a geometric series that converges uniformly on the circle because $|z/w| = |z|\,/R < 1$.

(3) Integrate term by term, with $c_k = \frac{1}{2\pi i}\oint u(w)\,w^{-k-1}\,\mathrm{d}w$.

The right-hand side is a convergent power series. So u is not merely smooth but real analytic. The values of u on a circle (the "shell" of Figure 2.3, shrunk to a curve) determine u inside through the smooth kernel $1/(w-z)$, $w\neq z$.

Analytic hypo-ellipticity

This suggests a stronger property than hypo-ellipticity.

Working definition 2.11 (analytic-hypo-elliptic operator)

L is **analytic-hypo-elliptic** if, whenever Lu is real analytic [equal to its convergent Taylor series near every point] on an open set Ω, u is real analytic on Ω. *How to test* (constant coefficients): check whether the Green's function is real analytic away from the origin.

- $\bar\partial$ is analytic-hypo-elliptic: $G = 1/\pi z$ is real analytic for $z\neq 0$, and (2.63) shows solutions of $\bar\partial u = 0$ are analytic.
- The Laplacian is analytic-hypo-elliptic (Nash: "not difficult to check"). Its Green's functions (2.16) are real analytic off the origin. In the plane, a harmonic function is locally the real part of a holomorphic function (Example 2.5), so it is analytic.
- The heat operator is hypo-elliptic but *not* analytic-hypo-elliptic (Nash's footnote 3). Near a point $(x_0,0)$ with $x_0\neq 0$ the heat kernel solves $LG=0$ (analytic source), but as a function of t it is a copy of $e^{-1/t}$, smooth and not analytic (Example 2.19). Physically: a solution of the heat equation can be exactly zero up to time 0 and non-zero afterwards, which an analytic function of t could never do.

For constant coefficients the picture is completely clean (a theorem of Petrowsky): *analytic-hypo-elliptic is the same as elliptic*. So we have the nested classes of Figure 2.14.

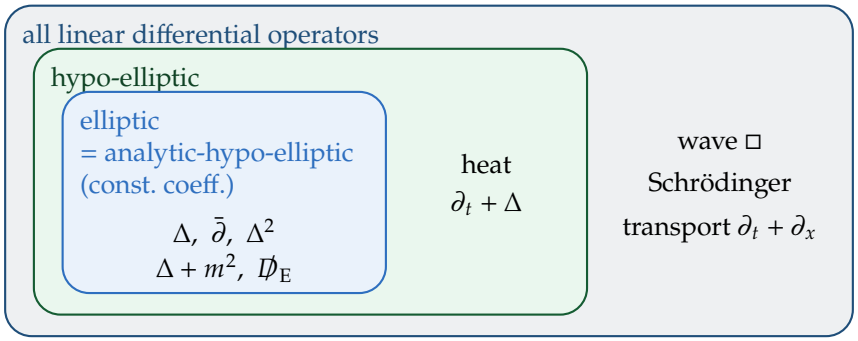


Figure 2.14: Nested classes of operators. Elliptic operators (blue) are hypo-elliptic (green); for constant coefficients they are exactly the analytic-hypo-elliptic ones. The heat operator smooths but is not elliptic. The wave, Schrödinger and transport operators propagate singularities and are not hypo-elliptic. ($\not{D}_{\mathrm{E}}$ is the Euclidean Dirac operator of Section 2.3.)

2.2.11 Ordinary differential equations

Nash ends the section with ODEs, where everything can be seen by hand. The general linear ODE of order n with smooth coefficients is

$$a_n(x)\frac{\mathrm{d}^n u}{\mathrm{d}x^n} + \cdots + a_1(x)\frac{\mathrm{d}u}{\mathrm{d}x} + a_0(x)\,u \;=\; f(x), \tag{2.64}$$

Nash's (2.31). Only the first term has n derivatives, so by the plane-wave test

$$\sigma_n(x,p) \;=\; a_n(x)\,(ip)^n, \tag{2.65}$$

Nash's (2.32), up to the harmless factor i^n which Nash absorbs into the coefficient. In one dimension $p \neq 0$ means $p^n \neq 0$, so:

Key idea: Ellipticity of an ODE

The ODE operator (2.64) is elliptic exactly where the top coefficient $a_n(x)$ is non-zero. Points where $a_n(x_0) = 0$ are the **singular points** of the ODE. Constant-coefficient ODEs with $a_n \neq 0$ are elliptic everywhere, and their solutions (combinations of $x^k e^{rx}$) are indeed smooth.

Regularity for ODEs, by induction (Nash's (2.37)). If a_n never vanishes we may divide by it:

$$\frac{\mathrm{d}^n u}{\mathrm{d}x^n} \;=\; \frac{f(x)}{a_n(x)} - \frac{a_{n-1}(x)}{a_n(x)}\frac{\mathrm{d}^{n-1}u}{\mathrm{d}x^{n-1}} - \cdots - \frac{a_0(x)}{a_n(x)}\,u. \tag{2.66}$$

Suppose u is a solution, so u has at least n continuous derivatives: $u \in C^n$. Suppose we know $u \in C^{n+j}$ for some $j \geq 0$ (true for $j = 0$). Then

$$\frac{\mathrm{d}^n u}{\mathrm{d}x^n} \overset{(1)}{\in} C^{j+1} \quad\Longrightarrow\quad u \overset{(2)}{\in} C^{n+j+1}.$$

(1) Look at the right side of (2.66). The functions f/a_n and a_k/a_n are C^∞, because $a_n \neq 0$ and everything is smooth. The derivatives $u^{(k)}$ with $k \leq n-1$ are in $C^{n+j-k} \subseteq C^{j+1}$. So the whole right side is in C^{j+1}.

(2) If the n-th derivative of u has $j+1$ continuous derivatives, then u has $n+j+1$.

By induction on j, $u \in C^\infty$. Every step used $a_n \neq 0$: *dividing by the leading symbol is what gains derivatives.* This is elliptic regularity in its most elementary form: the order-n operator gains back n derivatives.

Where a_n vanishes, this breaks down, and solutions can indeed be singular. Nash's two examples are the classic equations of mathematical physics with regular singular points.

Worked example 2.22: Bessel's equation, Nash's (2.33)–(2.34)

Bessel's equation of order λ,

$$x^2\frac{\mathrm{d}^2u}{\mathrm{d}x^2} + x\frac{\mathrm{d}u}{\mathrm{d}x} + (x^2-\lambda^2)\,u \;=\; 0, \tag{2.67}$$

arises from separating the Laplacian in cylindrical coordinates. Its leading symbol is $\sigma_2 = x^2(ip)^2 = -x^2p^2$ (Nash writes x^2p^2), which vanishes at $x=0$. So the operator is elliptic on $x>0$ and on $x<0$, but not at the origin. How do solutions behave there? Try a power $u = x^r(1+O(x))$ (the method of Frobenius). The terms of lowest order in x are

$$x^2\,r(r-1)x^{r-2} + x\,r\,x^{r-1} - \lambda^2x^r \overset{(1)}{=} \left[r(r-1)+r-\lambda^2\right]x^r \overset{(2)}{=} (r^2-\lambda^2)\,x^r.$$

(1) $(x^r)' = rx^{r-1}$, $(x^r)'' = r(r-1)x^{r-2}$; the x^2u term is of higher order, x^{r+2}, and is dropped.
(2) $r^2 - r + r = r^2$.

The lowest-order terms cancel only if $r^2=\lambda^2$ (the **indicial equation**), so $r=\pm\lambda$. The two solutions behave like x^λ and $x^{-\lambda}$ near 0 (with a logarithm, $x^\lambda\ln x$, when the two roots differ by an integer). Unless λ is a non-negative integer and we pick the right solution, u is not C^∞ at $x=0$.

Worked example 2.23: Bessel of order $\frac12$: explicit non-smooth solutions

For $\lambda=\frac12$ the solutions are elementary. We check that $u = x^{-1/2}\cos x$ solves (2.67). First,

$$u' \overset{(1)}{=} -\tfrac12 x^{-3/2}\cos x - x^{-1/2}\sin x,$$
$$u'' \overset{(2)}{=} \tfrac34 x^{-5/2}\cos x + x^{-3/2}\sin x - x^{-1/2}\cos x.$$

(1) Product rule.
(2) Differentiate each term of u' with the product rule; the two $\sin x$ terms, $\frac12x^{-3/2}\sin x$ each, add up.

Now assemble the three terms of (2.67), sorted by the type of term:

$$x^2u'' + xu' + \left(x^2-\tfrac14\right)u \overset{(1)}{=} \left(\tfrac34-\tfrac12-\tfrac14\right)x^{-1/2}\cos x + (1-1)\,x^{1/2}\sin x$$
$$+ (-1+1)\,x^{3/2}\cos x \;=\; 0.$$

(1) $x^2u'' = \frac34x^{-1/2}\cos x + x^{1/2}\sin x - x^{3/2}\cos x$; $xu' = -\frac12x^{-1/2}\cos x - x^{1/2}\sin x$; $(x^2-\frac14)u = x^{3/2}\cos x - \frac14x^{-1/2}\cos x$. Collect coefficients.

The same calculation works for $x^{-1/2}\sin x$. Both solutions are C^∞ on $x>0$, where the operator is elliptic. Neither is smooth at 0. The first blows up like $x^{-1/2}$. The second behaves like $x^{1/2}(1-x^2/6+\cdots)$, which is continuous but has infinite slope at 0 (Figure 2.15). The source is zero, hence smooth. The singularity of the solutions sits exactly at the zero of the leading symbol.

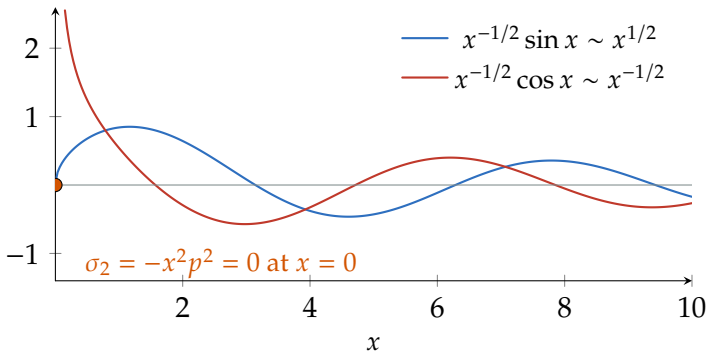


Figure 2.15: The two solutions of Bessel's equation of order $\frac{1}{2}$. On $x > 0$ the operator is elliptic and both are smooth (oscillating like sin and cos with a slowly decaying amplitude). At $x = 0$, where the leading symbol vanishes (orange), the red solution blows up and the blue one has a vertical tangent.

Worked example 2.24: Gauss's hypergeometric equation, Nash's (2.35)–(2.36)

The hypergeometric equation

$$x(1-x)\frac{\mathrm{d}^2u}{\mathrm{d}x^2} + \{\gamma - (\alpha+\beta+1)x\}\frac{\mathrm{d}u}{\mathrm{d}x} - \alpha\beta\, u \;=\; 0 \tag{2.68}$$

has leading symbol $\sigma_2 = -x(1-x)p^2$ (Nash: $x(1-x)p^2$), vanishing at $x = 0$ and $x = 1$. So it is non-elliptic at two points. One solution is the hypergeometric series

$$F(\alpha,\beta;\gamma;x) \;=\; \sum_{k=0}^{\infty} c_k x^k, \qquad \frac{c_{k+1}}{c_k} \;=\; \frac{(k+\alpha)(k+\beta)}{(k+\gamma)(k+1)}, \quad c_0 = 1,$$

whose ratio of coefficients tends to 1, so its radius of convergence is exactly 1: the distance to the next singular point. Take $\alpha = \beta = 1$, $\gamma = 2$. Then

$$\frac{c_{k+1}}{c_k} \overset{(1)}{=} \frac{(k+1)^2}{(k+2)(k+1)} \;=\; \frac{k+1}{k+2} \quad\Longrightarrow\quad c_k \;=\; \frac{1}{k+1},$$

$$F(1,1;2;x) \overset{(2)}{=} \sum_{k=0}^{\infty} \frac{x^k}{k+1} \overset{(3)}{=} \frac{1}{x}\sum_{k=1}^{\infty} \frac{x^k}{k} \overset{(4)}{=} -\frac{\ln(1-x)}{x}.$$

(1) The ratio formula with $\alpha = \beta = 1$, $\gamma = 2$; then $c_k = \frac{1}{2}\cdot\frac{2}{3}\cdots\frac{k}{k+1} = \frac{1}{k+1}$ (telescoping product).
(2) Insert the coefficients.
(3) Shift the index $k \to k-1$ and pull out $1/x$.
(4) The Mercator series $-\ln(1-x) = \sum_{k\geq 1} x^k/k$ for $|x| < 1$.

So this solution has a logarithmic branch point at $x = 1$, the zero of the leading symbol, exactly as Nash says. It is smooth everywhere else on $x < 1$, including at $x = 0$ ($-\ln(1-x)/x \to 1$). A second, independent solution near 0 behaves like $x^{1-\gamma}$ (with logarithms when γ is an integer) and is singular at $x = 0$ instead.

Watch out: Singular points of ODEs are not "bad" operators

The Bessel and hypergeometric equations are among the most useful equations in physics, and their singular points carry the physics: $x = 0$ is the axis of a cylinder, or the position

of a Coulomb centre. Non-ellipticity at a point just warns us that solutions need not be smooth there, so boundary conditions ("regular at the origin") must be imposed by hand. That is exactly how quantisation conditions arise in the hydrogen atom.

Reading Nash (§2, pp. 32–36, (2.19)–(2.37))

- p. 32, (2.19)–(2.23): multi-indices, $D^\alpha = (-i)^{|\alpha|}\partial^\alpha$, the symbol $\sigma(x,p) = \sum a_\alpha p^\alpha$ and the leading symbol σ_m; our Sections 2.2.1 to 2.2.3. His (2.21) has the opposite Fourier sign to ours (Page 728).
- p. 33: the definition of ellipticity ($\sigma_m(x,p) \neq 0$ for $p \neq 0$), the remark on invertibility modulo lower order, and the definition of hypo-ellipticity; our Sections 2.2.4 and 2.2.7. We added the covector transformation law (Section 2.2.5), which Nash only mentions in his footnote 4 on p. 37.
- pp. 33–34, (2.24)–(2.26): heat kernel and Schrödinger kernel; our Sections 2.2.8 and 2.2.9, with Hörmander's criterion added.
- pp. 34–35, (2.27)–(2.30) and footnote 3: Cauchy–Riemann operator, $G = 1/\pi z$, analytic hypo-ellipticity; our Section 2.2.10. Nash's symbol (2.30) lacks our factor i.
- pp. 35–36, (2.31)–(2.37): ODEs, Bessel, hypergeometric, and regularity by induction; our Section 2.2.11.

Recap: Ellipticity and hypo-ellipticity

- A general operator is $L = \sum_{|\alpha|\le m} a_\alpha(x) D^\alpha$. Its symbol $\sigma(x,p) = e^{-ip\cdot x} L e^{ip\cdot x} = \sum a_\alpha(x) p^\alpha$; its leading symbol σ_m keeps only $|\alpha| = m$.
- The leading symbol is what L looks like to short waves. It is homogeneous of degree m and transforms as a function of a *covector* $p \in T^*_x M$, so it is coordinate independent. The full symbol is not.
- Elliptic: $\sigma_m(x,p) \neq 0$ for $p \neq 0$. Then $|\sigma| \ge \frac{c}{2}|p|^m$ at large $|p|$, L is invertible modulo lower order (parametrix), and L gains m derivatives.
- Hypo-elliptic: Lu smooth on $\Omega \Rightarrow u$ smooth on Ω. Elliptic $\Rightarrow$ hypo-elliptic. The heat operator is hypo-elliptic but not elliptic (its kernel is smooth off the origin, like $e^{-1/t}$). The Schrödinger operator is not hypo-elliptic (its kernel oscillates wildly on $t = 0$).
- $\bar\partial$ is elliptic with Green's function $1/\pi z$. Its regularity is Cauchy's integral formula, which gives analyticity: $\bar\partial$ and Δ are analytic-hypo-elliptic, the heat operator is not.
- ODEs: elliptic where $a_n(x) \neq 0$. Singular points (Bessel at 0, hypergeometric at $0, 1$) are zeros of the leading symbol, and there solutions can be singular.

2.3 Ellipticity and vector bundles

Almost no field in physics is a single real function. The electric field has three components. A gauge potential A_μ has four, and each is a matrix in the Lie algebra. A Dirac spinor has four complex components, and on a curved space-time those components are not even functions: they are sections of a spinor bundle, and they mix when we change frames. The operators that

act on such fields, like $\nabla\times$, $\gamma^\mu\partial_\mu$ or d acting on p-forms, are *matrices of differential operators*.

So here is the question this section answers: *what does it mean for a matrix of differential operators, acting on sections of a vector bundle, to be elliptic?* The answer is a direct upgrade of what we have: the leading symbol becomes a *matrix*, and "non-zero" becomes "invertible". Once we have it, we will compute the symbol of the most important operator in the book, the Laplacian on p-forms, and arrive at Hodge's theorem.

2.3.1 Vector bundles and their sections: a working reminder

Nash uses vector bundles freely. We recall what we need; bundles are treated at length in Chapter 1 and Chapter T3.

Working definition 2.12 (vector bundle, in practice)

A (real or complex) **vector bundle** U of **rank** q over a manifold [a space covered by patches, each with coordinates $x_1,\dots,x_n$] M is a family of q-dimensional vector spaces U_x, the **fibres**, one for each point $x\in M$, glued together smoothly. *Operationally*: cover M by open sets Ω_a. Over each Ω_a choose a **local frame**, that is, q smoothly varying bases $e_1(x),\dots,e_q(x)$ of the fibres. Then a point of the bundle over Ω_a is labelled by (x,u) with $x\in\Omega_a$ and $u\in\mathbb{R}^q$ (or $\mathbb{C}^q$), the components in the frame: this is a **local trivialisation** $U|_{\Omega_a}\cong\Omega_a\times\mathbb{R}^q$. On an overlap $\Omega_a\cap\Omega_b$ the two sets of components are related by an invertible matrix,

$$u_{(a)} \;=\; h_{ab}(x)\,u_{(b)}, \qquad h_{ab}(x)\in \mathrm{GL}(q)\,[\text{ the invertible } q\times q \text{ matrices }], \tag{2.69}$$

the **transition functions**. The bundle is *trivial* if one frame works on all of M.

Working definition 2.13 (section)

A **section** s of U picks one vector $s(x)\in U_x$ for each x, smoothly. *Operationally*: over each Ω_a it is a column vector of q smooth functions $u_{(a)}(x)$, and on overlaps these obey (2.69). The space of smooth sections is written $\Gamma(M,U)$, or $\Gamma(U)$ for short. It is an infinite-dimensional vector space: the space of field configurations.

The examples we need are all built from the tangent and cotangent bundles.

- The **trivial bundle** $M\times\mathbb{R}^q$. Its sections are just $\mathbb{R}^q$-valued functions. On $\mathbb{R}^n$ every bundle is trivial, so on $\mathbb{R}^n$ "operators between bundles" are simply matrices of operators acting on vector-valued functions.
- The **tangent bundle** TM, of rank n. Its fibre T_xM is the space of velocity vectors at x, with local frame $\partial/\partial x^i$. A section is a vector field.
- The **cotangent bundle** T^*M, of rank n. Its fibre T^*_xM is the dual space of covectors (gradients) at x, with local frame $\mathrm{d}x^i$. A section is a 1-form. Under a change of coordinates the components of a covector transform as in (2.48): h_{ab} is the Jacobian matrix, transposed.
- The **bundle of p-forms** Λ^pT^*M, with local frame $\mathrm{d}x^{i_1}\wedge\cdots\wedge\mathrm{d}x^{i_p}$ $(i_1<\cdots<i_p)$, of rank $\binom{n}{p}$. A section is a p-form, and $\Gamma(\Lambda^pT^*M)=\Omega^p(M)$.
- Twisted bundles such as $\Lambda^pT^*M\otimes E$, where E carries the gauge indices: Lie-algebra-valued forms, like a Yang–Mills potential.

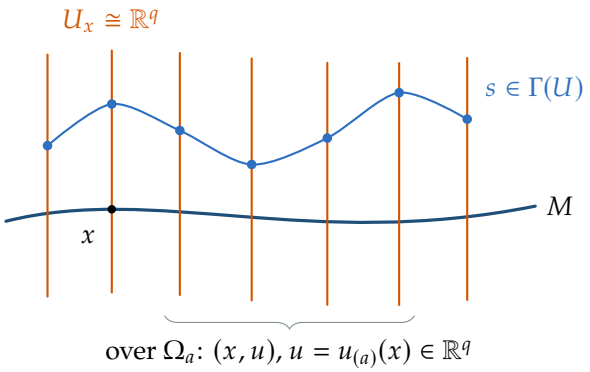


Figure 2.16: A vector bundle drawn as fibres (orange) standing over the base M (blue-black curve). Each fibre U_x is a copy of $\mathbb{R}^q$, drawn as a line. A section (blue) chooses one point in each fibre, smoothly: it is a "vector-valued function whose values live in different spaces at different points". Over a patch Ω_a a frame identifies every fibre with $\mathbb{R}^q$, and then the section is an honest column of q functions.

Worked example 2.25: The tangent bundle of S^2 is not trivial, but has transition functions

Cover S^2 by the two stereographic charts, from the north and south poles, with complex coordinates z and $w = 1/z$ on the overlap $\mathbb{C} \setminus \{0\}$. A tangent vector with component $v_{(z)}$ in the frame ∂_z has component $v_{(w)}$ in the frame ∂_w given by the chain rule:

$$v_{(w)} \overset{(1)}{=} \frac{\mathrm{d}w}{\mathrm{d}z}\, v_{(z)} \overset{(2)}{=} -\frac{1}{z^2}\, v_{(z)},$$

(1) A vector's components transform with the Jacobian (here a 1×1 complex matrix, since we treat TS^2 as a complex line bundle).

(2) $w = 1/z$, so $\mathrm{d}w/\mathrm{d}z = -1/z^2$.

So $h(z) = -z^{-2}$. On the equator $z = e^{i\phi}$ this is $-e^{-2i\phi}$, which winds -2 times around the origin as ϕ goes once around. The winding number is (up to sign conventions) the Euler number $\chi(S^2) = 2$, and it is the reason the bundle cannot be trivial: "you cannot comb a hairy ball". We mention this because the same kind of winding reappears below inside the *symbol* of an elliptic operator.

2.3.2 Why smooth sections?

Nash insists (p. 36) that the operator

$$L : \Gamma(M, U) \longrightarrow \Gamma(M, V) \tag{2.70}$$

act between spaces of *smooth* sections. This is his (2.38). His reason is elliptic regularity. We are doing differential topology, so smooth objects are the natural ones. For an elliptic L the smooth world is closed under "solving": if $Lu = f$ with f smooth, then u is automatically smooth. So restricting to smooth sections loses no solutions of elliptic equations. For a non-elliptic operator this is false, and the equation $Lu = f$ with smooth f may have non-smooth solutions that the smooth setting cannot see (the kinks of the wave equation, Example 2.8).

Later (in part 2b) we shall enlarge $\Gamma(M, U)$ to Sobolev spaces, where the analysis is easier, and then use elliptic regularity to come back.

2.3.3 Operators between bundles are matrices of operators

Take a patch $\Omega \subset M$ over which both bundles are trivial and on which we have coordinates $x = (x_1, \dots, x_n)$. Write U with frame components $u = (u_1, \dots, u_q)$, and V with components $(v_1, \dots, v_p)$. Then a linear differential operator of order m from sections of U to sections of V takes the form

$$(Lu)_i(x) \;=\; \sum_{j=1}^{q} \sum_{|\alpha| \le m} a_\alpha^{ij}(x)\, D^\alpha u_j(x), \qquad i = 1, \dots, p, \tag{2.71}$$

Nash's (2.39). Read it as follows: the i-th component of the output is a sum, over all input components j, of a scalar differential operator L_{ij} acting on u_j. In matrix form, $Lu = \sum_\alpha a_\alpha(x) D^\alpha u$, where each coefficient $a_\alpha(x)$ is now a $p \times q$ matrix: p rows because the output has p components, q columns because the input has q.

The symbol is found by the same plane-wave test as before (Definition 2.7), now applied to a plane wave with a constant *vector* amplitude.

Working definition 2.14 (matrix symbol and leading symbol)

For L as in (2.71), the **symbol** is the $p \times q$ matrix

$$\sigma(x, k) \;=\; \sum_{|\alpha| \le m} a_\alpha(x)\, k^\alpha, \tag{2.72}$$

Nash's (2.40), and the **leading symbol** is its top-degree part,

$$\sigma_m(x, k) \;=\; \sum_{|\alpha| = m} a_\alpha(x)\, k^\alpha, \tag{2.73}$$

Nash's (2.41). *How to compute*: for each fixed vector $v \in \mathbb{R}^q$,

$$\sigma(x, k)\, v \;=\; e^{-ik\cdot x}\, L\big(e^{ik\cdot x}\, v\big), \tag{2.74}$$

that is, in each entry L_{ij} replace $\partial_l \to i k_l$. Keep only the terms with m derivatives for σ_m. At each point x and for each covector [the momentum $k = (k_1, \dots, k_n)$ of the plane wave $e^{ik\cdot x}$] k, $\sigma_m(x, k)$ is a linear map

$$\sigma_m(x, k) : U_x \longrightarrow V_x.$$

Nash now writes k instead of p for the Fourier variable, and we follow him in this section. It is the same momentum variable, and (Section 2.2.5) it is a covector: $k = k_i\, \mathrm{d}x^i \in T_x^* M$.

Worked example 2.26: Gradient, curl and divergence as bundle operators on $\mathbb{R}^3$

On $\mathbb{R}^3$ take U and V trivial, of the appropriate ranks. We compute the three symbols with (2.74).

Gradient, $\nabla : \Gamma(\mathbb{R}^3 \times \mathbb{R}) \to \Gamma(\mathbb{R}^3 \times \mathbb{R}^3)$, a 3×1 matrix of operators. Apply it to $e^{ik\cdot x} v$ with $v \in \mathbb{R}$:

$$\nabla\big(e^{ik\cdot x} v\big) \overset{(1)}{=} \begin{pmatrix} \partial_1 \\ \partial_2 \\ \partial_3 \end{pmatrix} e^{ik\cdot x} v \overset{(2)}{=} e^{ik\cdot x}\, i \begin{pmatrix} k_1 \\ k_2 \\ k_3 \end{pmatrix} v, \qquad \text{so} \qquad \sigma_1(\nabla)(k) = i\,\vec{k}.$$

(1) The gradient as a column of three derivatives.
(2) Each ∂_l brings down $i k_l$.

Divergence, $\nabla\cdot : \Gamma(\mathbb{R}^3\times\mathbb{R}^3)\to\Gamma(\mathbb{R}^3\times\mathbb{R})$, the 1×3 row $(\partial_1,\partial_2,\partial_3)$. The same computation gives $\sigma_1(\nabla\cdot)(k)\,\vec v = i\,\vec k\cdot\vec v$: the row matrix $i\vec k^{\,T}$.
Curl, a 3×3 matrix of operators,

$$\nabla\times \;=\; \begin{pmatrix} 0 & -\partial_3 & \partial_2 \\ \partial_3 & 0 & -\partial_1 \\ -\partial_2 & \partial_1 & 0 \end{pmatrix}, \qquad \sigma_1(\nabla\times)(k) \;=\; i\begin{pmatrix} 0 & -k_3 & k_2 \\ k_3 & 0 & -k_1 \\ -k_2 & k_1 & 0 \end{pmatrix} =:\; i\,[\vec k]_\times,$$

because $\nabla\times(e^{ik\cdot x}\vec v) = e^{ik\cdot x}\,i\,\vec k\times\vec v$. The matrix $[\vec k]_\times$ is "cross product with $\vec k$".
All three are first order and have constant coefficients, so the symbol equals the leading symbol. Now look at the shapes: 3×1, 1×3 and 3×3. Only the last one can possibly be invertible. We shall see in a moment that it is not.

Worked example 2.27: The Cauchy–Riemann equations as a 2×2 real system

Write a complex function as $f = u + iv$. The Cauchy–Riemann equations $u_x = v_y$, $u_y = -v_x$ say that the real operator

$$L\begin{pmatrix} u \\ v \end{pmatrix} \;=\; \begin{pmatrix} \partial_x & -\partial_y \\ \partial_y & \partial_x \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix} \;=\; \begin{pmatrix} u_x - v_y \\ u_y + v_x \end{pmatrix}$$

vanishes. (It is $2\bar\partial f = (\partial_x + i\partial_y)(u+iv)$, split into real and imaginary parts.) Its leading symbol and determinant are

$$\sigma_1(L)(k) \overset{(1)}{=} i\begin{pmatrix} k_1 & -k_2 \\ k_2 & k_1 \end{pmatrix}, \qquad \det\sigma_1(L)(k) \overset{(2)}{=} i^2\big(k_1^2+k_2^2\big) \;=\; -\,|k|^2\,.$$

(1) Replace $\partial_x\to ik_1$, $\partial_y\to ik_2$ in each entry.
(2) The determinant of a 2×2 matrix, with the factor i pulled out of both rows.

So $\sigma_1(L)(k)$ is invertible for every $k\neq 0$: L is elliptic in the matrix sense. Geometrically, $\begin{pmatrix} k_1 & -k_2 \\ k_2 & k_1 \end{pmatrix}$ is $|k|$ times a rotation, the real form of multiplication by the complex number $k_1 + ik_2$. This matches the scalar symbol $\sigma_1(2\bar\partial) = i(k_1+ik_2)$ of Section 2.2.10. The same operator is elliptic whether we view it as one complex equation or two real ones.

2.3.4 How the matrix symbol transforms

Nash's footnote 4 warns that the coefficients $a_\alpha^{ij}(x)$ "will change by a Jacobian-type transformation" when we change local coordinates, and he suppresses this from the notation. Let us see exactly what changes, because the result is the reason the leading symbol is a *geometric* object while the rest of the symbol is not.

There are two kinds of change: of coordinates on M, and of frame in the fibres. The first we have already done: k transforms as a covector (2.48). The second is new. Change frames in U and V by invertible matrix functions $h_U(x)$ and $h_V(x)$, so the new components are $u' = h_U u$ and $v' = h_V v$. The operator in the new frames is $L' = h_V\,L\,h_U^{-1}$. For a first-order operator $L = \sum_j A^j(x)\partial_j + B(x)$ this is

$$L'u' \overset{(1)}{=} h_V\Big(\sum_j A^j\partial_j + B\Big)\big(h_U^{-1}u'\big)$$

$$\overset{(2)}{=} \sum_j h_V A^j h_U^{-1}\, \partial_j u' \;+\; \Big(\sum_j h_V A^j\, \partial_j\big(h_U^{-1}\big) + h_V B h_U^{-1}\Big) u'. \tag{2.75}$$

(1) Express $u = h_U^{-1} u'$, apply L, and convert the output to the new frame with h_V.
(2) Product rule: ∂_j hits either u' or the matrix h_U^{-1}.

So the top-order coefficients transform *tensorially*, $A'^j = h_V A^j h_U^{-1}$, and therefore

$$\sigma_1(L')(x,k) \;=\; h_V(x)\, \sigma_1(L)(x,k)\, h_U(x)^{-1}. \tag{2.76}$$

This is exactly how a linear map $U_x \to V_x$ must transform when we change bases in both spaces. The zeroth-order coefficient, by contrast, picks up the extra term $h_V A^j \partial_j h_U^{-1}$, which depends on *derivatives* of the frame change. The same happens in every order: the top-order part transforms as a linear map between fibres, and the lower-order parts mix with derivatives of the transition functions.

Dictionary: Frame changes are gauge transformations

Physicists have met (2.75) before. Take $U = V$, $h_U = h_V = g(x)^{-1}$ and $L = \gamma^\mu(\partial_\mu + A_\mu)$, a Dirac operator coupled to a gauge field. Then the zeroth-order part $\gamma^\mu A_\mu$ transforms into $\gamma^\mu(g^{-1}A_\mu g + g^{-1}\partial_\mu g)$: the inhomogeneous term is the $g^{-1}\mathrm{d}g$ of a gauge transformation (the convention of Chapter T3). The leading symbol $i\gamma^\mu k_\mu$ does not see A_μ at all. *Gauge fields are lower order.* So coupling any elliptic operator to a gauge field, or to gravity through a spin connection, never changes whether it is elliptic.

2.3.5 The leading symbol as a bundle map over T^*M

At a point x, for each covector $k \in T_x^*M$, we have a linear map $\sigma_m(x,k) : U_x \to V_x$. Nash now packages this family globally, and the packaging is worth understanding because it is exactly the input of the index theorem in Chapter 4.

Step 1: the base becomes phase space. The pair (x,k) is a point of the cotangent bundle T^*M, the classical phase space of a particle on M. Let

$$\pi : T^*M \to M, \qquad \pi(x,k) = x,$$

be the projection that forgets the momentum. T^*M is a manifold of dimension $2n$.

Step 2: pull the bundles back. We want σ_m to be a map between bundles *over* T^*M. So we build a bundle over T^*M whose fibre over (x,k) is U_x.

Working definition 2.15 (pullback bundle)

Given a map $\pi : T^*M \to M$ and a bundle U over M, the **pullback bundle** π^*U is the bundle over T^*M whose fibre at the point (x,k) is

$$(\pi^*U)_{(x,k)} \;=\; U_{\pi(x,k)} \;=\; U_x.$$

Operationally: same fibres, but each fibre U_x is now repeated once for every momentum k at x. Its transition functions [the matrices h_{ab} relating components in two overlapping frames] are those of U, composed with π: they depend on x only, not on k.

Step 3: the symbol is a bundle homomorphism. A **bundle homomorphism** $\phi : E \to F$ between two bundles over the same base assigns to each base point b a linear map $\phi_b : E_b \to F_b$, depending smoothly on b. Over the base point $b = (x, k) \in T^*M$, the fibres of π^*U and π^*V are U_x and V_x, and $\sigma_m(x, k)$ is a linear map between them. So

$$\sigma_m : \pi^*U \longrightarrow \pi^*V \tag{2.77}$$

is a bundle homomorphism over T^*M. This is Nash's "elucidation" on p. 38. The transformation law (2.76), together with the covector law for k, is precisely the statement that the local matrices glue into a global bundle map.

Step 4: delete the zero section. Ellipticity is a condition for $k \neq 0$ only. At $k = 0$ the leading symbol, being homogeneous of degree $m \geq 1$, always vanishes: $\sigma_m(x, 0) = 0$. So we remove the set $\{(x, 0)\} \subset T^*M$ of zero covectors. This set is the image of the **zero section** $x \mapsto (x, 0)$, a copy of M sitting inside T^*M. What remains, $T^*M \setminus \{\text{zero section}\}$, is the bundle of non-zero covectors.

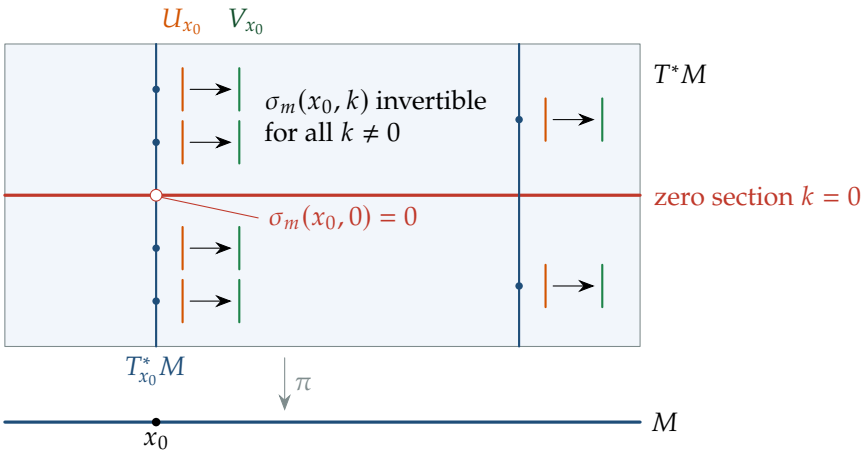


Figure 2.17: The global picture of the leading symbol. The cotangent bundle T^*M (shaded) sits over M: above each point x_0 stands the whole momentum space $T^*_{x_0}M$ (vertical blue line). At every point (x_0, k) of T^*M the pulled-back bundles supply a copy of U_{x_0} (orange) and of V_{x_0} (green), and the symbol is a linear map between them. Ellipticity asks that every one of these maps be invertible, except on the zero section (red line), where the symbol always vanishes.

2.3.6 The definition of ellipticity for bundles

With all this in place Nash gives two versions of the definition. They say the same thing.

Working definition 2.16 (elliptic operator between vector bundles (Nash, p. 38))

The m-th order linear differential operator $L : \Gamma(M, U) \to \Gamma(M, V)$ is **elliptic** if, in every local chart and trivialisation [a choice of local frame, so that sections become columns of functions], its leading symbol $\sigma_m(x, k)$ is an *invertible matrix* for every x and every $k \neq 0$. *Equivalently* (Nash's "more abstract definition"), the bundle homomorphism [one linear map $U_x \to V_x$ at each point (x, k) of T^*M, varying smoothly] $\sigma_m : \pi^*U \to \pi^*V$ is a **bundle isomorphism off the zero section** [the points $(x, 0)$ with $k = 0$] of T^*M. *How to test*: compute the $p \times q$ matrix $\sigma_m(x, k)$; check that it is square; check that $\det \sigma_m(x, k) \neq 0$ for $k \neq 0$. By homogeneity it suffices to check $|k| = 1$.

Three remarks make this concrete.

(a) It is the old definition when $p = q = 1$. A 1×1 matrix is invertible exactly when it is non-zero, so we recover Definition 2.9. This is Nash's remark that the new definition is "the natural linear algebraic generalisation".

(b) An elliptic operator connects bundles of the same rank. An invertible matrix is square. So if L is elliptic, $p = q$: the bundles U and V have equal rank. The gradient ($p = 3$, $q = 1$) and the divergence ($p = 1$, $q = 3$) fail at once. Nash notes that "in the coordinate free method of dealing with symbols" this is the abstract definition: an isomorphism between π^*U and π^*V forces equal rank.

(c) It is coordinate- and frame-independent. By (2.76), changing frames replaces σ_m by $h_V \sigma_m h_U^{-1}$, which is invertible exactly when σ_m is. Changing coordinates replaces k by the transformed covector, which is non-zero exactly when k is. So it makes sense to say "L is elliptic" on a manifold.

Key idea: Ellipticity for systems

An operator between bundles is elliptic when, for every non-zero momentum k, the matrix "what L does to a short wave of momentum k" can be inverted. Physically: no polarisation of a short wave is invisible to the operator (injectivity), and every polarisation of the output can be produced (surjectivity).

Worked example 2.28: The curl is not elliptic: the longitudinal mode

The curl on $\mathbb{R}^3$ is a square (3×3) operator, so the rank test passes. But $\sigma_1(\nabla\times)(k) = i[\vec{k}]_\times$ and

$$
\begin{aligned}
[\vec{k}]_\times \vec{k} &\overset{(1)}{=} \vec{k} \times \vec{k} \overset{(2)}{=} 0, \\
\det[\vec{k}]_\times &\overset{(3)}{=} 0 \cdot \big(0 \cdot 0 + k_1^2\big) - (-k_3)\big(k_3 \cdot 0 - k_1 k_2\big) + k_2\big(k_3 k_1 - 0 \cdot k_2\big) \\
&= -k_1 k_2 k_3 + k_1 k_2 k_3 \;=\; 0.
\end{aligned}
$$

(1) The matrix $[\vec{k}]_\times$ is, by construction, cross product with $\vec{k}$.

(2) Any vector crossed with itself is zero. So $\vec{k} \neq 0$ lies in the kernel.

(3) Expand the determinant along the first row $(0, -k_3, k_2)$. (Faster: an antisymmetric 3×3 matrix A has $\det A = \det A^T = \det(-A) = -\det A$, so $\det A = 0$.)

The kernel of the symbol is the *longitudinal* polarisation $\vec{v} \parallel \vec{k}$. The transverse polarisations $\vec{v} \perp \vec{k}$ are rotated by 90° about $\vec{k}$ and stretched by $|k|$; the longitudinal one is annihilated (Figure 2.18).

The failure of ellipticity has a concrete price: there is no regularity. Take any function f that is continuous but nowhere differentiable in some region, and let $\vec{B} = \nabla f$ (in the sense of distributions). Then $\nabla \times \vec{B} = 0$, which is as smooth as can be, but $\vec{B}$ is very rough. The rough part lives entirely in the longitudinal modes $\hat{\vec{B}}(k) \parallel \vec{k}$, which the curl cannot see. This is also the physicist's statement that $\nabla \times \vec{B} = \mu_0 \vec{J}$ determines $\vec{B}$ only up to a gradient: we need $\nabla \cdot \vec{B} = 0$ as well.

Worked example 2.29: Repairing the curl: the div–grad–curl operator is elliptic

The cure is to add the missing information. Act on a pair (scalar f, vector field $\vec{A}$) and

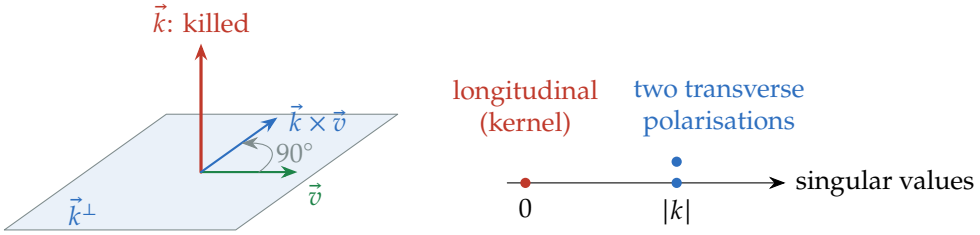


Figure 2.18: Why the curl fails to be elliptic. Left: its symbol $i\,\vec{k}\times(\cdot)$ rotates each transverse vector $\vec{v}\perp\vec{k}$ by 90° within the plane $\vec{k}^{\perp}$ and scales it by $|k|$, but it annihilates the longitudinal direction $\vec{k}$ (red). Right: the singular values of $[\vec{k}]_\times$ are $|k|, |k|, 0$. One zero singular value, at every $k\neq 0$, is enough to destroy ellipticity.

output a pair (scalar, vector):

$$L\begin{pmatrix} f \\ \vec{A}\end{pmatrix} = \begin{pmatrix} \nabla\cdot\vec{A} \\ \nabla f + \nabla\times\vec{A}\end{pmatrix}, \qquad L:\Gamma(\mathbb{R}^3\times\mathbb{R}^4)\to\Gamma(\mathbb{R}^3\times\mathbb{R}^4).$$

Using Example 2.26, its leading symbol is the 4×4 matrix

$$\sigma_1(L)(k) = i\,M(k), \qquad M(k) = \begin{pmatrix} 0 & \vec{k}^{\,T} \\ \vec{k} & [\vec{k}]_\times\end{pmatrix} = \begin{pmatrix} 0 & k_1 & k_2 & k_3 \\ k_1 & 0 & -k_3 & k_2 \\ k_2 & k_3 & 0 & -k_1 \\ k_3 & -k_2 & k_1 & 0\end{pmatrix}.$$

Is it invertible? Compute M^TM block by block, using $[\vec{k}]_\times^T = -[\vec{k}]_\times$:

$$\begin{aligned} M^TM &\overset{(1)}{=} \begin{pmatrix} 0 & \vec{k}^{\,T} \\ \vec{k} & -[\vec{k}]_\times\end{pmatrix}\begin{pmatrix} 0 & \vec{k}^{\,T} \\ \vec{k} & [\vec{k}]_\times\end{pmatrix} \\ &\overset{(2)}{=} \begin{pmatrix} \vec{k}^{\,T}\vec{k} & \vec{k}^{\,T}[\vec{k}]_\times \\ -[\vec{k}]_\times\vec{k} & \vec{k}\vec{k}^{\,T} - [\vec{k}]_\times^2\end{pmatrix} \\ &\overset{(3)}{=} \begin{pmatrix} |k|^2 & 0 \\ 0 & \vec{k}\vec{k}^{\,T} - \left(\vec{k}\vec{k}^{\,T} - |k|^2\,I_3\right)\end{pmatrix} = |k|^2\,I_4. \end{aligned}$$

(1) The transpose of a block matrix transposes the blocks and swaps the off-diagonal ones; $(\vec{k})^T = \vec{k}^{\,T}$ and $[\vec{k}]_\times^T = -[\vec{k}]_\times$.

(2) Multiply the blocks.

(3) $[\vec{k}]_\times\vec{k} = \vec{k}\times\vec{k} = 0$, and its transpose $\vec{k}^{\,T}[\vec{k}]_\times = -([\vec{k}]_\times\vec{k})^T = 0$. For the corner use $\vec{k}\times(\vec{k}\times\vec{v}) = \vec{k}(\vec{k}\cdot\vec{v}) - |k|^2\,\vec{v}$ ("BAC–CAB"), i.e. $[\vec{k}]_\times^2 = \vec{k}\vec{k}^{\,T} - |k|^2\,I_3$.

So $M(k)/|k|$ is an orthogonal matrix, $\det M(k) = \pm|k|^4 \neq 0$, and L is elliptic. The longitudinal mode killed by the curl is now caught by the gradient term, and the divergence adds the equation that the curl lacked. Moreover $\sigma_1(L)^\dagger\sigma_1(L) = |k|^2\,I_4$ is the symbol of the Laplacian: L is a *square root of the Laplacian* on four-component fields. In Section 2.3.11 we will recognise L as $\mathrm{d}+\mathrm{d}^*$ acting on (0-forms $\oplus$ 2-forms) in $\mathbb{R}^3$, and it is also the Dirac operator of $\mathbb{R}^3$ written in quaternionic form.

Worked example 2.30: The Dirac operator is elliptic in Euclidean signature

The Euclidean Dirac operator on $\mathbb{R}^n$ is $\not{\partial} = \gamma^\mu \partial_\mu$, where the γ^μ are Hermitian matrices with

$$\gamma^\mu \gamma^\nu + \gamma^\nu \gamma^\mu \;=\; 2\,\delta^{\mu\nu}\, I \tag{2.78}$$

(Chapter T4). The leading symbol and its square are

$$\begin{aligned} \sigma_1(\not{\partial})(k) &\overset{(1)}{=} i\,\gamma^\mu k_\mu, \\ \big(\sigma_1(\not{\partial})(k)\big)^2 &\overset{(2)}{=} -\gamma^\mu\gamma^\nu k_\mu k_\nu \overset{(3)}{=} -\tfrac{1}{2}\big(\gamma^\mu\gamma^\nu + \gamma^\nu\gamma^\mu\big)k_\mu k_\nu \overset{(4)}{=} -\,|k|^2\; I. \end{aligned} \tag{2.79}$$

(1) Plane-wave test: $\partial_\mu \to i k_\mu$; the γ's are constant matrices.
(2) $i^2 = -1$.
(3) $k_\mu k_\nu$ is symmetric in $\mu\nu$, so only the symmetric part of $\gamma^\mu\gamma^\nu$ contributes.
(4) The Clifford relation (2.78).

So $\sigma_1(\not{\partial})(k)$ is invertible for $k \neq 0$, with inverse $\sigma_1(\not{\partial})(k)^{-1} = -i\gamma^\mu k_\mu/|k|^2$. This is the Euclidean fermion propagator, $1/(i\not{k})$. The Dirac operator is elliptic.

Explicitly in two dimensions. Take $\gamma^1 = \sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ and $\gamma^2 = \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}$. Then

$$\not{\partial} \;=\; \begin{pmatrix} 0 & \partial_1 - i\partial_2 \\ \partial_1 + i\partial_2 & 0 \end{pmatrix} \;=\; \begin{pmatrix} 0 & 2\partial \\ 2\bar\partial & 0 \end{pmatrix}, \qquad \sigma_1(\not{\partial})(k) \;=\; i\begin{pmatrix} 0 & k_1 - ik_2 \\ k_1 + ik_2 & 0 \end{pmatrix},$$

with $\partial = \tfrac{1}{2}(\partial_1 - i\partial_2)$ and $\bar\partial = \tfrac{1}{2}(\partial_1 + i\partial_2)$. The determinant is $-i^2(k_1 - ik_2)(k_1 + ik_2) = |k|^2$. The two-dimensional Dirac operator is the Cauchy–Riemann operator and its conjugate, stacked. The chiral half $2\bar\partial : \Gamma(S^+) \to \Gamma(S^-)$ is itself elliptic, between rank-one bundles.

Explicitly in four dimensions. In the chiral representation

$$\gamma^j = \begin{pmatrix} 0 & -i\sigma_j \\ i\sigma_j & 0 \end{pmatrix}\ (j = 1,2,3), \qquad \gamma^4 = \begin{pmatrix} 0 & I_2 \\ I_2 & 0 \end{pmatrix},$$

(check: $\gamma^j\gamma^l = \mathrm{diag}(\sigma_j\sigma_l, \sigma_j\sigma_l)$, whose symmetric part is $\delta_{jl}I$, and $\gamma^4\gamma^j = -\gamma^j\gamma^4 = \mathrm{diag}(i\sigma_j, -i\sigma_j)$), the symbol is

$$\sigma_1(\not{\partial})(k) \;=\; i\begin{pmatrix} 0 & q(k)^\dagger \\ q(k) & 0 \end{pmatrix}, \qquad q(k) \;=\; k_4\, I_2 + i\,\vec{k}\cdot\vec{\sigma}.$$

The block $q(k)$ is the symbol (up to the factor i) of the chiral Dirac operator $\Gamma(S^+) \to \Gamma(S^-)$, and

$$q(k)^\dagger q(k) \overset{(1)}{=} (k_4 - i\vec{k}\cdot\vec{\sigma})(k_4 + i\vec{k}\cdot\vec{\sigma}) \overset{(2)}{=} k_4^2 + (\vec{k}\cdot\vec{\sigma})^2 \overset{(3)}{=} k_4^2 + \left|\vec{k}\right|^2 \;=\; |k|^2\,,$$

(1) The Pauli matrices are Hermitian, so $(i\vec{k}\cdot\vec{\sigma})^\dagger = -i\vec{k}\cdot\vec{\sigma}$.
(2) Multiply out; the cross terms $\pm ik_4\,\vec{k}\cdot\vec{\sigma}$ cancel.
(3) $(\vec{k}\cdot\vec{\sigma})^2 = \left|\vec{k}\right|^2 I$, from $\sigma_j\sigma_l + \sigma_l\sigma_j = 2\delta_{jl}$.

So $q(k)/|k| \in \mathrm{SU}(2)$ for $k \neq 0$: it is a unit quaternion.

In Minkowski signature the same computation, with $\{\gamma^\mu, \gamma^\nu\} = 2\eta^{\mu\nu}$, gives $(i\gamma^\mu k_\mu)^2 = -\eta^{\mu\nu}k_\mu k_\nu = -(k_0^2 - \left|\vec{k}\right|^2)$, which vanishes on the light cone. The Minkowski Dirac operator is not elliptic: it is hyperbolic, like the wave operator whose square root it is. Wick rotation $k_0 \to ik_4$ turns $k_0^2 - \left|\vec{k}\right|^2$ into $-\,|k|^2$ and restores ellipticity, exactly as in Section 2.1.7.

A topological shadow: the symbol on the unit sphere

Because σ_m is homogeneous, $\sigma_m(x,\lambda k)=\lambda^m\sigma_m(x,k)$, all its information is in its values on unit covectors $|k|=1$. At each x these form a sphere S^{n-1}, and together they form the **unit cosphere bundle** $S(T^*M)$. An elliptic symbol therefore gives, at each x, a map

$$S^{n-1}\ni\hat{k}\;\longmapsto\;\sigma_m(x,\hat{k})\in \mathrm{GL}(q)\quad\text{(invertible matrices)}.$$

Maps from a sphere into a group of invertible matrices can wind around, like the transition function of Example 2.25. The winding cannot be undone without making the symbol singular somewhere, that is, without destroying ellipticity. So it is a *topological* invariant of the elliptic operator, and it survives any continuous deformation through elliptic operators.

Two examples. For $2\bar\partial$ on $\mathbb{R}^2$, on the unit circle $\hat{k}=(\cos\phi,\sin\phi)$ the symbol is $i(\cos\phi+i\sin\phi)=ie^{i\phi}$. As ϕ goes once around, this goes once around the origin of $\mathbb{C}\setminus\{0\}=\mathrm{GL}(1,\mathbb{C})$: winding number 1 (Figure 2.19). For the chiral Dirac operator on $\mathbb{R}^4$, $\hat{k}\mapsto q(\hat{k})$ maps S^3 to the unit quaternions $\mathrm{SU}(2)\cong S^3$, and it is the identity map, of degree 1. The Atiyah–Singer index theorem of Chapter 4 says that this kind of winding of the symbol, combined with the twisting of the bundles, determines the number of zero modes. We only note here that the topological data of an elliptic operator *is* its leading symbol, and it lives on T^*M minus the zero section, which is why Nash took such care to put it there.

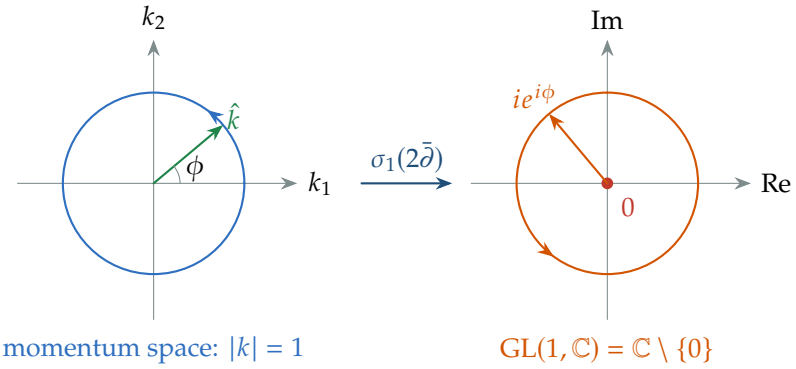


Figure 2.19: The leading symbol of the Cauchy–Riemann operator, restricted to unit momenta, is a map from the circle $|k|=1$ (left) to the non-zero complex numbers (right): $\phi\mapsto ie^{i\phi}$. It winds once around the forbidden point 0 (red), where the symbol would fail to be invertible. No deformation through elliptic symbols can unwind it. This winding number is the simplest example of the topological information carried by an elliptic symbol, which the index theorem converts into a count of zero modes.

Watch out: Operators of mixed order

Nash's definition assumes one order m for the whole matrix. Some important systems mix orders: the Stokes system of slow viscous flow, $-\nu\nabla^2\vec{u}+\nabla P=\vec{f}$, $\nabla\cdot\vec{u}=0$, has second derivatives of $\vec{u}$ but first derivatives of P. If we kept only the second-order terms we would drop ∇P and $\nabla\cdot\vec{u}$ completely and wrongly conclude that the system is not elliptic. The fix (due to Douglis and Nirenberg) assigns separate weights to rows and columns. We never need it in this book: all our operators ($\mathrm{d}+\mathrm{d}^*$, Δ_p, $\not{D}$, $\bar\partial$) have a single order.

Try it: Which systems are elliptic?

On $\mathbb{R}^2$ decide whether each operator is elliptic. (a) $L(u,v)=(u_x+v_y,\ u_y-v_x)$. (b) $L(u,v)=(u_x,\ v_y)$. (c) The Laplacian acting on $\mathbb{R}^q$-valued functions, $u\mapsto-\nabla^2u$ (each

component separately). (d) $L(u,v) = (u_x - v_y,\ u_x + v_y)$.

Answers. (a) $\sigma_1 = i\begin{pmatrix} k_1 & k_2 \\ k_2 & -k_1 \end{pmatrix}$, $\det = i^2(-k_1^2 - k_2^2) = |k|^2 \neq 0$: elliptic (it is the Cauchy–Riemann system for $u - iv$). (b) $\sigma_1 = i\,\mathrm{diag}(k_1, k_2)$, $\det = -k_1 k_2$, which vanishes on both axes: not elliptic (two uncoupled transport operators). (c) $\sigma_2 = |k|^2\, I_q$, invertible: elliptic. (d) $\sigma_1 = i\begin{pmatrix} k_1 & -k_2 \\ k_1 & k_2 \end{pmatrix}$, $\det = i^2(2k_1k_2) = -2k_1k_2$: not elliptic, although it looks similar to Cauchy–Riemann. The determinant, not the appearance, decides.

2.3.7 The Laplacian on p-forms

Nash illustrates the bundle definition with the most important example in the subject. Let M be a compact Riemannian manifold of dimension n with metric g_{ij}. On p-forms we want a Laplacian Δ_p, a second-order operator

$$\Delta_p : \Gamma(M, \Lambda^p T^*M) \longrightarrow \Gamma(M, \Lambda^p T^*M), \tag{2.80}$$

Nash's (2.44). The bundle is $U = V = \Lambda^p T^*M$, of rank $\binom{n}{p}$, and Δ_p maps p-forms to p-forms. Nash defines it by

$$\Delta_p \;=\; \big(\mathrm{d}\mathrm{d}^* + \mathrm{d}^*\mathrm{d}\big)_p, \tag{2.81}$$

his (2.42), where the subscript means "restricted to p-forms" and d^* is the *adjoint* of the exterior derivative. To make sense of d^* we need an inner product on forms, and that is where the metric and the Hodge star come in. We recall them now, with all the formulas of Nash's (2.43). For more on forms and $\star$ see Chapter T1.

Components and the Hodge star

Convention for components. We write a p-form as

$$\omega \;=\; \frac{1}{p!}\,\omega_{i_1\ldots i_p}\,\mathrm{d}x^{i_1}\wedge\cdots\wedge\mathrm{d}x^{i_p} \;=\; \sum_{i_1<\cdots<i_p} \omega_{i_1\ldots i_p}\,\mathrm{d}x^{i_1}\wedge\cdots\wedge\mathrm{d}x^{i_p},$$

with $\omega_{i_1\ldots i_p}$ totally antisymmetric. The first form sums over all index values (with the $1/p!$ correcting the overcounting); the second sums over increasing indices only. Nash writes $\omega = \omega_{i_1\ldots i_p}\mathrm{d}x^{i_1}\wedge\cdots\wedge\mathrm{d}x^{i_p}$ and means the second.

Working definition 2.17 (Hodge star)

The **Hodge star** $\star : \Lambda^p T^*_x M \to \Lambda^{n-p} T^*_x M$ is the linear map defined, in an *oriented orthonormal* coframe [n one-forms at x with $\langle e^i, e^j\rangle = \delta^{ij}$ in the metric, listed in right-handed order] $e^1, \ldots, e^n$ at x, by

$$\star\big(e^{i_1}\wedge\cdots\wedge e^{i_p}\big) \;=\; \pm\, e^{j_1}\wedge\cdots\wedge e^{j_{n-p}}, \tag{2.82}$$

where $\{j_1,\ldots,j_{n-p}\}$ is the complement of $\{i_1,\ldots,i_p\}$ and the sign is chosen so that $e^{i_1}\wedge\cdots\wedge e^{i_p}\wedge\star(e^{i_1}\wedge\cdots\wedge e^{i_p}) = e^1\wedge\cdots\wedge e^n$, the volume form. *In general coordinates* this becomes Nash's formula

$$\star\omega \;=\; \frac{\sqrt{g}}{(n-p)!}\,\epsilon^{i_1\ldots i_p}{}_{i_{p+1}\ldots i_n}\,\omega_{i_1\ldots i_p}\,\mathrm{d}x^{i_{p+1}}\wedge\cdots\wedge\mathrm{d}x^{i_n} \qquad (\text{sum over } i_1<\cdots<i_p), \tag{2.83}$$

where $\sqrt{g} = \sqrt{\det g_{ij}}$, ϵ is the Levi-Civita symbol ($\epsilon_{12\ldots n} = +1$), and its first p indices are raised with g^{ij}.

Geometrically, $\star$ turns a p-dimensional "plane element" into the complementary $(n-p)$-dimensional plane element orthogonal to it, with the same size (Figure 2.20). Two facts from (2.43) are all we will need:

$$\star\star \;=\; (-1)^{p(n-p)} \quad \text{on } p\text{-forms}, \qquad \langle \omega, \nu \rangle \;=\; \int_M \omega \wedge \star\nu \quad (\omega, \nu \in \Omega^p(M)). \tag{2.84}$$

The first follows from (2.82): applying $\star$ twice returns the original basis form, and the sign is the sign of moving p one-forms past $n-p$ others. The second defines the L^2 **inner product** of forms. In an orthonormal coframe $\omega \wedge \star\nu = \left(\sum_I \omega_I \nu_I\right)\mathrm{vol}$, where I runs over increasing multi-indices. So $\langle \omega, \omega \rangle = \int_M \sum_I \omega_I^2 \,\mathrm{vol} \geq 0$, with equality only for $\omega = 0$. This positivity uses the *Riemannian* signature. Nash remarks that some signs would change for a Lorentzian metric; worse, positivity fails.

Worked example 2.31: The Hodge star in $\mathbb{R}^2$ and $\mathbb{R}^3$

In $\mathbb{R}^3$ with the Euclidean metric, $e^i = \mathrm{d}x^i$ and the rule (2.82) gives

$$\star 1 = \mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z, \qquad \star \mathrm{d}x = \mathrm{d}y \wedge \mathrm{d}z, \quad \star \mathrm{d}y = \mathrm{d}z \wedge \mathrm{d}x, \quad \star \mathrm{d}z = \mathrm{d}x \wedge \mathrm{d}y,$$

and $\star(\mathrm{d}y \wedge \mathrm{d}z) = \mathrm{d}x$ and cyclically, $\star(\mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z) = 1$. Check one sign: $\mathrm{d}x \wedge \star \mathrm{d}x = \mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z$, the volume form, as required. Here $p(3-p)$ is always even, so $\star\star = +1$. This is the dictionary of vector calculus: a vector $\vec{A}$ can be stored as the 1-form $A_i \mathrm{d}x^i$ or as the 2-form $\star(A_i \mathrm{d}x^i) = A_1 \mathrm{d}y \wedge \mathrm{d}z + A_2 \mathrm{d}z \wedge \mathrm{d}x + A_3 \mathrm{d}x \wedge \mathrm{d}y$ (its "flux form").
In $\mathbb{R}^2$: $\star 1 = \mathrm{d}x \wedge \mathrm{d}y$, $\star \mathrm{d}x = \mathrm{d}y$, $\star \mathrm{d}y = -\mathrm{d}x$, $\star(\mathrm{d}x \wedge \mathrm{d}y) = 1$. On 1-forms $\star\star = -1$, in agreement with $(-1)^{1\cdot 1}$: $\star$ is rotation by 90°, and doing it twice is rotation by 180°.

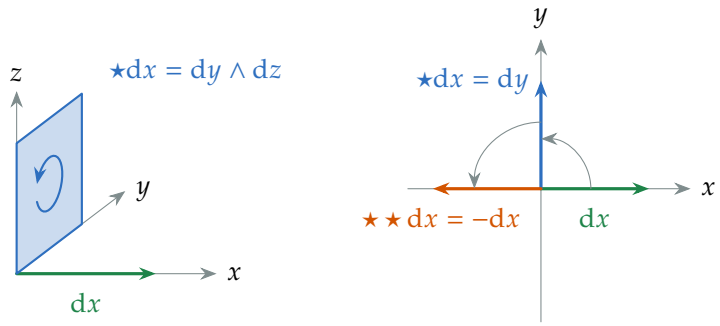


Figure 2.20: The Hodge star turns a p-plane element into the orthogonal complementary one. Left, in $\mathbb{R}^3$: the 1-form $\mathrm{d}x$ (green arrow, pointing along the normal) becomes the 2-form $\mathrm{d}y \wedge \mathrm{d}z$ (blue square in the yz-plane, oriented so that x followed by y, z is right-handed). Right, in $\mathbb{R}^2$: $\star$ on 1-forms is rotation by 90°, so $\star\star = -1$ there, in agreement with $\star\star = (-1)^{p(n-p)}$.

The adjoint d^*

Working definition 2.18 (co-differential)

The **co-differential** (or adjoint of d) $\mathrm{d}^* : \Omega^p(M) \to \Omega^{p-1}(M)$ is the operator with

$$\langle \mathrm{d}\mu, \omega \rangle \;=\; \langle \mu, \mathrm{d}^*\omega \rangle \qquad \text{for all } \mu \in \Omega^{p-1}(M),\ \omega \in \Omega^p(M), \tag{2.85}$$

with $\langle\ ,\ \rangle$ the L^2 inner product [$\langle \omega, \nu \rangle = \int_M \omega \wedge \star\nu$, from (2.84)], on a compact M without boundary. *How to compute*:

$$\mathrm{d}^* \;=\; (-1)^{np+n+1} \star \mathrm{d} \star \qquad \text{on } p\text{-forms}. \tag{2.86}$$

Both lines are in Nash's (2.43), stated without proof (the order of the arguments in the inner product does not matter, since it is symmetric). The formula is a two-line consequence of Stokes' theorem, and we derive it because the sign is the kind of thing one forgets. Take $\mu \in \Omega^{p-1}$ and $\omega \in \Omega^p$. Then $\mu \wedge \star\omega$ is an $(n-1)$-form and

$$\begin{aligned} 0 &\overset{(1)}{=} \int_M \mathrm{d}\big(\mu \wedge \star\omega\big) \overset{(2)}{=} \int_M \mathrm{d}\mu \wedge \star\omega + (-1)^{p-1} \int_M \mu \wedge \mathrm{d} \star \omega \\ &\overset{(3)}{=} \langle \mathrm{d}\mu, \omega \rangle + (-1)^{p-1}(-1)^{(p-1)(n-p+1)} \int_M \mu \wedge \star\star\, \mathrm{d} \star \omega \\ &\overset{(4)}{=} \langle \mathrm{d}\mu, \omega \rangle - (-1)^{np+n+1} \big\langle \mu, \star \mathrm{d} \star \omega \big\rangle. \end{aligned} \tag{2.87}$$

(1) Stokes' theorem on a manifold without boundary: $\int_M \mathrm{d}(\cdot) = \int_{\partial M}(\cdot) = 0$.

(2) The Leibniz rule $\mathrm{d}(\mu \wedge \beta) = \mathrm{d}\mu \wedge \beta + (-1)^{\deg \mu} \mu \wedge \mathrm{d}\beta$, with $\deg \mu = p - 1$.

(3) The first term is the inner product by (2.84). In the second, the form $\mathrm{d} \star \omega$ has degree $n - p + 1$; insert $1 = (-1)^{r(n-r)} \star\star$ with $r = n - p + 1$, valid on r-forms.

(4) Collect signs: the exponent is $(p-1) + (p-1)(n-p+1) = (p-1)(n-p+2)$. Modulo 2 this is $(p-1)(n-p) = pn - p^2 - n + p \equiv pn + n$, because $p^2 \equiv p$. So the total sign is $(-1)^{pn+n} = -(-1)^{np+n+1}$. The remaining integral is $\int \mu \wedge \star(\star \mathrm{d} \star \omega) = \langle \mu, \star \mathrm{d} \star \omega \rangle$.

Setting the last line to zero and comparing with (2.85) gives (2.86). Note that d^* lowers the degree by one, and $(\mathrm{d}^*)^2 = \pm \star\, \mathrm{dd} \star = 0$.

Worked example 2.32: d^* is minus the divergence; Δ_p is minus ∇^2

Work on $\mathbb{R}^3$ (or on a flat torus T^3, so that "compact without boundary" holds and all the integrations by parts are honest). Let $A = A_1 \mathrm{d}x + A_2 \mathrm{d}y + A_3 \mathrm{d}z$. With $n = 3$, $p = 1$ the sign in (2.86) is $(-1)^{3+3+1} = -1$, so

$$\begin{aligned} \mathrm{d}^* A &\overset{(1)}{=} - \star\, \mathrm{d}\left(A_1 \mathrm{d}y \wedge \mathrm{d}z + A_2 \mathrm{d}z \wedge \mathrm{d}x + A_3 \mathrm{d}x \wedge \mathrm{d}y\right) \\ &\overset{(2)}{=} - \star \left((\partial_1 A_1 + \partial_2 A_2 + \partial_3 A_3)\, \mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z\right) \overset{(3)}{=} -\nabla \cdot \vec{A}. \end{aligned}$$

(1) (2.86) and the table of Example 2.31 for $\star A$.

(2) $\mathrm{d}(A_1 \mathrm{d}y \wedge \mathrm{d}z) = \partial_1 A_1\, \mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}z$ (the ∂_2, ∂_3 terms die because $\mathrm{d}y \wedge \mathrm{d}y = 0$ etc.), and cyclically.

(3) $\star$ of the volume form is 1.

The minus sign is exactly what integration by parts demands: $\langle \mathrm{d}f, A \rangle = \int \nabla f \cdot \vec{A} = -\int f\, \nabla \cdot \vec{A} = \langle f, \mathrm{d}^* A \rangle$.

Now the Laplacians. On functions, $\Delta_0 f = \mathrm{d}^* \mathrm{d} f$ (there is no dd^* term since $\mathrm{d}^* f = 0$), and

$$\Delta_0 f \;=\; \mathrm{d}^*(\nabla f) \;=\; -\nabla \cdot \nabla f \;=\; -\nabla^2 f,$$

the positive Laplacian (2.7). On 1-forms, the term $\mathrm{dd}^* A = -\nabla(\nabla \cdot \vec{A})$. For the other, $\mathrm{d}A$ is the 2-form $\star(\nabla \times \vec{A})$ (the flux form of the curl), and on 2-forms in $\mathbb{R}^3$ the sign in (2.86) is $(-1)^{6+3+1} = +1$, so $\mathrm{d}^* \mathrm{d} A = \star \mathrm{d} \star \star (\nabla \times \vec{A}) = \star \mathrm{d}(\nabla \times \vec{A}) = \nabla \times (\nabla \times \vec{A})$, read as a 1-form. Hence

$$\Delta_1 \vec{A} \overset{(1)}{=} \nabla \times (\nabla \times \vec{A}) - \nabla(\nabla \cdot \vec{A}) \overset{(2)}{=} -\nabla^2 \vec{A}. \tag{2.88}$$

(1) The two pieces just computed.

(2) The vector identity $\nabla \times (\nabla \times \vec{A}) = \nabla(\nabla \cdot \vec{A}) - \nabla^2 \vec{A}$.

So in flat space the form Laplacian is minus the ordinary Laplacian applied to each component. This is why Nash's notation Δ_p generalises Δ. On a curved manifold the same is true up to terms with fewer derivatives: the **Weitzenböck formula** reads $\Delta_1 = \nabla^*\nabla + \mathrm{Ric}$, where $\nabla^*\nabla$ is $-g^{ij}\nabla_i\nabla_j$ and the Ricci curvature is a zeroth-order term. That lower-order term is invisible to the symbol.

2.3.8 The symbol calculus

We could compute $\sigma_2(\Delta_p)$ by brute force: write out $\mathrm{d}\mathrm{d}^* + \mathrm{d}^*\mathrm{d}$ in coordinates and replace $\partial \to ik$. For $p = 1$ in $\mathbb{R}^3$ we just did, in effect. Nash recommends a cleverer way: two rules that let us compute symbols of composite operators from symbols of their pieces.

Rule 1: the symbol of a composition is the product of the symbols

Key idea: Composition (Nash's (2.45))

Let $Q : \Gamma(U) \to \Gamma(V)$ have order j and $P : \Gamma(V) \to \Gamma(W)$ have order i. Then $P \circ Q : \Gamma(U) \to \Gamma(W)$ has order (at most) $i + j$, and its leading symbol is the product of the leading symbols, composed as linear maps:

$$\sigma^{P\circ Q}_{i+j}(x,k) \;=\; \sigma^P_i(x,k)\,\sigma^Q_j(x,k) \qquad \big(U_x \xrightarrow{\sigma^Q_j} V_x \xrightarrow{\sigma^P_i} W_x\big). \tag{2.89}$$

Nash says "the reader can easily verify" this. Here is the verification, using the oscillatory test (2.47). Fix a phase ψ with $\mathrm{d}\psi(x) = k$ and a vector v. Apply Q first:

$$\begin{aligned} Q\big(e^{i\lambda\psi}v\big) &\overset{(1)}{=} e^{i\lambda\psi}\big(\lambda^j\,\sigma^Q_j(x,\mathrm{d}\psi)\,v + O(\lambda^{j-1})\big) \;=:\; e^{i\lambda\psi}\,w_\lambda(x), \\ P\big(e^{i\lambda\psi}w_\lambda\big) &\overset{(2)}{=} e^{i\lambda\psi}\big(\lambda^i\,\sigma^P_i(x,\mathrm{d}\psi)\,w_\lambda(x) + \lambda^{i-1}(\dots)\big) \\ &\overset{(3)}{=} e^{i\lambda\psi}\big(\lambda^{i+j}\,\sigma^P_i(x,\mathrm{d}\psi)\,\sigma^Q_j(x,\mathrm{d}\psi)\,v + O(\lambda^{i+j-1})\big). \end{aligned}$$

(1) The oscillatory test for Q; the correction terms have fewer powers of λ. Call the bracket $w_\lambda(x)$: it is a smooth vector function, of size λ^j.

(2) The oscillatory test for P, now applied to $e^{i\lambda\psi}$ times a non-constant vector. Every derivative that lands on w_λ instead of on the exponential costs one power of λ, so those terms are in the "$\lambda^{i-1}(\dots)$". The derivatives of w_λ are still of size λ^j (differentiating does not bring down λ because w_λ has no oscillating phase).

(3) Insert w_λ and keep the top power λ^{i+j}; everything else is at most λ^{i+j-1}.

Reading off the coefficient of λ^{i+j} gives (2.89). The mechanism: *to leading order, derivatives only hit the fast phase, and on the fast phase each operator is just multiplication by its symbol.* The order of the product is the order of the maps: σ^P after σ^Q, as matrices. (If $\sigma^P_i\,\sigma^Q_j = 0$, the order of PQ is less than $i + j$; we shall see this happen for $\mathrm{d} \circ \mathrm{d}$.)

Worked example 2.33: The lower-order symbol does *not* multiply

On $\mathbb{R}$ take $P = \partial_x$ and $Q = x\,\partial_x$. Then, by the product rule,

$$PQ\,u \;=\; \partial_x(xu') \;=\; xu'' + u', \qquad \text{so} \qquad PQ \;=\; x\partial_x^2 + \partial_x.$$

The symbols (replace $\partial_x \to ik$) are

$$\sigma^P = ik, \qquad \sigma^Q = ixk, \qquad \sigma^{PQ} \overset{(1)}{=} x(ik)^2 + ik = -xk^2 + ik,$$
$$\sigma^P \sigma^Q \overset{(2)}{=} (ik)(ixk) = -xk^2.$$

(1) The plane-wave test applied to $x\partial_x^2 + \partial_x$.
(2) Multiply the two symbols.

The leading symbols agree ($-xk^2$), as (2.89) promises. The full symbols do not: σ^{PQ} has the extra first-order piece ik, produced by ∂_x differentiating the coefficient x. For the record, the exact rule is $\sigma^{PQ} = \sum_\alpha \frac{1}{\alpha!}(\partial_k^\alpha \sigma^P)(D_x^\alpha \sigma^Q)$. Here the $\alpha = 1$ term is $(\partial_k\, ik)\big(-i\partial_x(ixk)\big) = (i)(k) = ik$, which is exactly the discrepancy. Part 2b develops this rule for pseudo-differential operators.

Dictionary: Commutators and Poisson brackets

If P and Q are scalar operators, PQ and QP have the same leading symbol, so the commutator $[P, Q]$ has order at most $i + j - 1$. Its symbol of that order is

$$\sigma_{i+j-1}\big([P, Q]\big) = -i\,\{\sigma_i^P, \sigma_j^Q\}, \qquad \{f, h\} = \sum_l \Big(\frac{\partial f}{\partial k_l}\frac{\partial h}{\partial x_l} - \frac{\partial f}{\partial x_l}\frac{\partial h}{\partial k_l}\Big), \tag{2.90}$$

the Poisson bracket on phase space T^*M. Check with Example 2.33: $[\partial_x, x\partial_x] = \partial_x$ has symbol ik, and $-i\{ik, ixk\} = -i\,(i)(ik) = ik$. This is Dirac's rule "commutator $\leftrightarrow$ $i\hbar\times$ Poisson bracket" with $\hbar = 1$ and $D = -i\partial$ playing the momentum. The symbol calculus *is* the classical limit of quantum mechanics: the leading symbol is the classical observable, the operator is its quantisation.

Rule 2: the symbol of the adjoint is the adjoint of the symbol

If U and V carry fibre inner products and M carries a volume form, every operator $L : \Gamma(U) \to \Gamma(V)$ has a **formal adjoint** $L^* : \Gamma(V) \to \Gamma(U)$, defined by $\langle Lu, v\rangle = \langle u, L^*v\rangle$ for all sections, where $\langle\cdot,\cdot\rangle$ integrates the fibre inner product over M. Nash's (2.47) states that

$$\sigma_m^{L^*}(x, k) = \big(\sigma_m^L(x, k)\big)^*, \tag{2.91}$$

the adjoint (conjugate transpose, in orthonormal frames) of the matrix $\sigma_m^L(x, k)$.

Why. Compute L^* for a scalar first-order operator $L = a(x)\partial_x$ on $\mathbb{R}$, using the inner product $\langle u, v\rangle = \int u\,\bar v\,\mathrm{d}x$ (linear in the first slot, antilinear in the second). For compactly supported u, v:

$$\langle Lu, v\rangle \overset{(1)}{=} \int a\,u'\,\bar v\,\mathrm{d}x \overset{(2)}{=} -\int u\,(a\bar v)'\,\mathrm{d}x \overset{(3)}{=} \int u\,\overline{\big(-\bar a\,v' - \bar a'\,v\big)}\,\mathrm{d}x,$$

(1) Definition, with $\langle f, h\rangle = \int f\bar h$.
(2) Integrate by parts; the boundary terms vanish.
(3) Take the complex conjugate outside.

So $L^*v = -\bar a\,v' - \bar a'\,v$, and its leading symbol is $-\bar a\,(ik) = \overline{a\,(ik)}$ (as k is real). That is $\overline{\sigma_1^L}$. The mechanism in general: each integration by parts moves one derivative across, costing a minus sign, and $-ik = \overline{ik}$; the coefficients get conjugated (and, for matrices, transposed) because they move to the other side of the inner product; derivatives that land on the coefficients

give lower-order terms. This is the symbol-level version of $(\hat{p})^\dagger = \hat{p}$: the momentum operator $D = -i\partial$ is self-adjoint, and its symbol k is real.

Assembling the Laplacian

Put $P = \mathrm{d}$, $Q = \mathrm{d}^*$ (and the other way round) into (2.89). Both are first order, and Δ_p is the sum of two compositions, so

$$\sigma_2^{\Delta_p}(x,k) \;=\; \sigma_1^{\mathrm{d}}(x,k)\,\sigma_1^{\mathrm{d}^*}(x,k) + \sigma_1^{\mathrm{d}^*}(x,k)\,\sigma_1^{\mathrm{d}}(x,k), \tag{2.92}$$

Nash's (2.46). (The first product is $\Lambda^p \to \Lambda^{p-1} \to \Lambda^p$, the second $\Lambda^p \to \Lambda^{p+1} \to \Lambda^p$.) We need the two first-order symbols.

2.3.9 The symbols of d, d^* and Δ_p

The symbol of d. Because we are working with forms, Nash writes the Fourier variable as a 1-form, $k = k_i\,\mathrm{d}x^i$. Apply d to a plane wave times a form v with constant components:

$$\mathrm{d}\big(e^{ik\cdot x}v\big) \overset{(1)}{=} \mathrm{d}\big(e^{ik\cdot x}\big)\wedge v + e^{ik\cdot x}\,\mathrm{d}v \overset{(2)}{=} \big(ik_i\,e^{ik\cdot x}\,\mathrm{d}x^i\big)\wedge v \;=\; e^{ik\cdot x}\,\big(i\,k\wedge v\big). \tag{2.93}$$

(1) Leibniz rule for d (a function times a form).
(2) $\mathrm{d}v = 0$ since v has constant components; $\mathrm{d}(e^{ik\cdot x}) = \partial_i e^{ik\cdot x}\,\mathrm{d}x^i$.

Hence

$$\sigma_1^{\mathrm{d}}(x,k) : \Lambda^p T_x^*M \longrightarrow \Lambda^{p+1}T_x^*M, \qquad v \longmapsto i\,k\wedge v, \tag{2.94}$$

which is Nash's (2.48) up to the factor i (see the pitfall below). The symbol of d is *wedging with the momentum*. It needs no metric at all, and it is a map between different bundles, $\Lambda^p \to \Lambda^{p+1}$, of generally different ranks $\binom{n}{p}$ and $\binom{n}{p+1}$. So it is a non-square matrix, and d alone is never elliptic (except in trivial cases).

The symbol of d^*. By Rule 2, $\sigma_1^{\mathrm{d}^*}(x,k) = (ik\wedge)^* = -i\,(k\wedge)^*$. The adjoint of "wedge with k" with respect to the pointwise inner product of forms is "contract with k":

$$(k\wedge)^* \;=\; \iota_k, \qquad \sigma_1^{\mathrm{d}^*}(x,k) : \Lambda^p T_x^*M \to \Lambda^{p-1}T_x^*M, \quad v \longmapsto -i\,\iota_k v. \tag{2.95}$$

Here ι_k is the **interior product** (contraction) with the vector $k^\sharp = g^{ij}k_j\,\partial_i$ obtained by raising the index of k with the metric. (Nash writes $v \mapsto k \lrcorner v$.) In an orthonormal coframe it acts on basis forms by

$$\iota_{e_1}\big(e^1\wedge e^J\big) = e^J, \qquad \iota_{e_1}\big(e^J\big) = 0 \quad (1\notin J),$$

and it is an anti-derivation: $\iota_k(\alpha\wedge\beta) = (\iota_k\alpha)\wedge\beta + (-1)^{\deg\alpha}\alpha\wedge\iota_k\beta$. To see that it is the adjoint of $k\wedge$, rotate the coframe so that $k = |k|\,e^1$. Then $k\wedge$ sends $e^J \mapsto |k|\,e^1\wedge e^J$ for $1\notin J$ and kills $e^1\wedge e^J$. Its matrix in the orthonormal basis of forms has entries 0 and $|k|$; the transpose sends $e^1\wedge e^J \mapsto |k|\,e^J$ and kills the e^J with $1\notin J$: that is $|k|\,\iota_{e_1} = \iota_k$. Unlike σ^{d}, this symbol uses the metric (through $k^\sharp$), as d^* itself does.

A quick check against Example 2.32: on 1-forms in $\mathbb{R}^3$, $\mathrm{d}^* = -\nabla\cdot$ has symbol $\vec{A} \mapsto -i\,\vec{k}\cdot\vec{A}$, and indeed $\iota_k(A_j\mathrm{d}x^j) = k_jA_j$.

The key identity. For any covector k and any form v,

$$k \wedge \iota_k v + \iota_k(k \wedge v) \;=\; |k|^2\, v, \qquad |k|^2 = g^{ij}k_i k_j. \tag{2.96}$$

Derivation:

$$\iota_k(k \wedge v) \overset{(1)}{=} (\iota_k k)\, v - k \wedge \iota_k v \overset{(2)}{=} |k|^2\, v - k \wedge \iota_k v.$$

(1) The anti-derivation rule with $\alpha = k$, a 1-form ($\deg \alpha = 1$).
(2) $\iota_k k = k(k^\sharp) = g^{ij}k_i k_j = |k|^2$.

Moving the last term to the left gives (2.96).

The symbol of the Laplacian. Now assemble:

$$\begin{aligned} \sigma_2^{\Delta_p}(x,k)\, v &\overset{(1)}{=} (ik\wedge)(-i\iota_k)v + (-i\iota_k)(ik\wedge)v \\ &\overset{(2)}{=} k \wedge \iota_k v + \iota_k(k \wedge v) \overset{(3)}{=} |k|^2\, v. \end{aligned} \tag{2.97}$$

(1) (2.92) with the symbols (2.94) and (2.95).
(2) $(i)(-i) = 1$ in both terms.
(3) The key identity (2.96).

So

$$\boxed{\sigma_2^{\Delta_p}(x,k) \;=\; |k|^2\, I} \tag{2.98}$$

where I is the identity on the $\binom{n}{p}$-dimensional fibre $\Lambda^p T_x^* M$. This is Nash's (2.49), written there as $k^2 I$. It is invertible for every $k \neq 0$, with inverse $|k|^{-2}\, I$. *The Hodge Laplacian is elliptic*, on every Riemannian manifold, for every p. Nash's remark: although σ^{Δ_p} is a matrix, it is a multiple of the identity, whereas σ^{d} is not even square. The two non-invertible pieces $k \wedge \iota_k$ and $\iota_k k\wedge$ are complementary projections (times $|k|^2$): one keeps the part of v containing k, the other the part not containing it. Adding them gives everything.

Worked example 2.34: All the symbols of the de Rham complex of $\mathbb{R}^3$, as explicit matrices

Take the bases $\{1\}$, $\{\mathrm{d}x^1, \mathrm{d}x^2, \mathrm{d}x^3\}$, $\{\mathrm{d}x^2 \wedge \mathrm{d}x^3, \mathrm{d}x^3 \wedge \mathrm{d}x^1, \mathrm{d}x^1 \wedge \mathrm{d}x^2\}$, $\{\mathrm{d}x^1 \wedge \mathrm{d}x^2 \wedge \mathrm{d}x^3\}$ of $\Lambda^0, \ldots, \Lambda^3$. These are the bases that make $\star$ look like the identity. Then $k\wedge$ becomes:

$$\begin{aligned} &\Lambda^0 \to \Lambda^1: && f \mapsto k\, f = (k_1 f, k_2 f, k_3 f) && \text{matrix } \vec{k}\ (3\times 1), \\ &\Lambda^1 \to \Lambda^2: && k \wedge (A_j \mathrm{d}x^j) \overset{(1)}{=} (k_2 A_3 - k_3 A_2)\, \mathrm{d}x^2 \wedge \mathrm{d}x^3 + \ldots && \text{matrix } [\vec{k}]_\times\ (3\times 3), \\ &\Lambda^2 \to \Lambda^3: && k \wedge \big(B_1 \mathrm{d}x^2 \wedge \mathrm{d}x^3 + \ldots\big) \overset{(2)}{=} (\vec{k}\cdot\vec{B})\, \mathrm{d}x^1 \wedge \mathrm{d}x^2 \wedge \mathrm{d}x^3 && \text{matrix } \vec{k}^T\ (1\times 3). \end{aligned}$$

(1) Expand $k_i \mathrm{d}x^i \wedge A_j \mathrm{d}x^j$ and collect the coefficient of each basis 2-form: the three coefficients are the components of $\vec{k} \times \vec{A}$.
(2) $\mathrm{d}x^1 \wedge (\mathrm{d}x^2 \wedge \mathrm{d}x^3)$ is the volume form, as are its cyclic versions; all other products vanish.

So $\sigma(\mathrm{d})$ on $\Lambda^0, \Lambda^1, \Lambda^2$ is i times $\vec{k}$, $[\vec{k}]_\times$, $\vec{k}^T$: the symbols of *grad, curl, div* from Example 2.26. The de Rham complex of $\mathbb{R}^3$ is the grad–curl–div sequence of vector calculus. The symbols of d^* are the conjugate transposes: $-i\vec{k}^T$, $-i[\vec{k}]_\times^T = i[\vec{k}]_\times$, $-i\vec{k}$.
Now verify (2.98) degree by degree, using $[\vec{k}]_\times^2 = \vec{k}\vec{k}^T - |k|^2\, I_3$ from Example 2.29:

$$\Lambda^0: \quad (-i\vec{k}^T)(i\vec{k}) \;=\; \vec{k}^T\vec{k} \;=\; |k|^2,$$

$$\Lambda^1: \quad (i\vec{k})(-i\vec{k}^T) + (i[\vec{k}]_\times)(i[\vec{k}]_\times) \overset{(1)}{=} \vec{k}\vec{k}^T - \left(\vec{k}\vec{k}^T - |k|^2 I_3\right) \;=\; |k|^2 I_3,$$
$$\Lambda^2: \quad (i[\vec{k}]_\times)(i[\vec{k}]_\times) + (-i\vec{k})(i\vec{k}^T) \overset{(2)}{=} -\left(\vec{k}\vec{k}^T - |k|^2 I_3\right) + \vec{k}\vec{k}^T \;=\; |k|^2 I_3,$$
$$\Lambda^3: \quad (i\vec{k}^T)(-i\vec{k}) \;=\; |k|^2 .$$

(1) On 1-forms the first term is $\sigma(\mathrm{d})\sigma(\mathrm{d}^*)$ through Λ^0 (the $\nabla\nabla\cdot$ term), the second is $\sigma(\mathrm{d}^*)\sigma(\mathrm{d})$ through Λ^2 (the curl curl term); then $i \cdot i = -1$ and BAC–CAB.
(2) On 2-forms, the same two terms in the other order.

On 1-forms this is the symbol of the identity $\nabla \times \nabla \times - \nabla\nabla\cdot = -\nabla^2$ from (2.88): $\vec{k} \times (\vec{k} \times \vec{A})$ removes the transverse part of $\vec{A}$, $\vec{k}(\vec{k} \cdot \vec{A})$ supplies the longitudinal part, and together they give $|k|^2 \vec{A}$.

Watch out: Nash's $\sigma(\mathrm{d}) = k\wedge$ versus our $ik\wedge$

Nash drops the factor i in (2.48), as he did for the Cauchy–Riemann operator in (2.30). With our Fourier convention $\partial_j \to ik_j$ (Section 2.1.2), the symbol of any first-order operator carries an i. Nothing depends on it: invertibility is unchanged by a factor i, and in products like (2.97) the factors combine to $(i)(-i) = 1$, so Nash's (2.49) is correct as it stands. The i does matter for signs of second-order symbols built from a single first-order operator squared, like $\sigma(\not{\partial})^2 = -|k|^2$ in (2.79).

2.3.10 Generalisations: twisting, gauge fields and gauge fixing

Nash mentions that "various generalisations of the Laplacian can be formed by tensoring the bundle $\Lambda^p T^*M$ with another vector bundle E, and by studying the equations of motion for Yang–Mills theories". Let us make both concrete.

Twisting by a bundle with a connection. Let E be a vector bundle with a connection A (a gauge field, anti-Hermitian matrix-valued 1-form, in the conventions of Chapter T3). Forms with values in E are sections of $\Lambda^p T^*M \otimes E$, and the covariant exterior derivative is

$$\mathrm{d}_A \omega \;=\; \mathrm{d}\omega + A \wedge \omega.$$

The extra term $A \wedge \omega$ has no derivatives, so it is zeroth order and invisible to the leading symbol:

$$\sigma_1^{\mathrm{d}_A}(x,k) \;=\; i\,k\wedge \otimes I_E, \qquad \sigma_2^{\Delta_A}(x,k) \;=\; |k|^2\, I, \tag{2.99}$$

where $\Delta_A = \mathrm{d}_A\mathrm{d}_A^* + \mathrm{d}_A^*\mathrm{d}_A$ is the gauge-covariant Laplacian. It is elliptic, whatever the gauge field. (This is the bundle version of the "gauge fields are lower order" remark in Section 2.3.4.)

Worked example 2.35: Maxwell's operator is not elliptic; gauge fixing makes it so

The Euclidean Maxwell action is $\frac{1}{2}\|\mathrm{d}A\|^2 = \frac{1}{2}\langle A, \mathrm{d}^*\mathrm{d}A\rangle$, so the field equation in vacuum is $\mathrm{d}^*\mathrm{d}A = 0$. Is $\mathrm{d}^*\mathrm{d}$ (on 1-forms) elliptic? Its symbol, by Rule 1, is

$$\sigma_2^{\mathrm{d}^*\mathrm{d}}(k)\,A \overset{(1)}{=} (-i\iota_k)(ik\wedge A) \overset{(2)}{=} |k|^2 A - k\,(\iota_k A) \;=\; |k|^2 A - k\,(k\cdot A).$$

(1) The composition rule with the symbols of d and d^*.
(2) The key identity (2.96), with $\iota_k A = k \cdot A$ a function.

As a matrix this is $|k|^2 I - kk^T$, which is $|k|^2$ times the projection onto the plane transverse

to k. It kills the longitudinal polarisation $A \propto k$, so *Maxwell's operator is not elliptic*. The reason is gauge invariance: $A \mapsto A + \mathrm{d}\chi$ does not change $\mathrm{d}A$, and at the symbol level $\mathrm{d}\chi$ has symbol $ik\chi$, exactly the longitudinal direction. Every gauge theory has this problem, and it is why the photon propagator $(|k|^2\,\delta_{\mu\nu} - k_\mu k_\nu)^{-1}$ does not exist until we fix a gauge. Fixing the gauge repairs it. Add the gauge-fixing term $\frac{1}{2}\,\|\mathrm{d}^*A\|^2$ (the Lorenz/Feynman gauge). The operator becomes $\mathrm{d}^*\mathrm{d} + \mathrm{d}\mathrm{d}^* = \Delta_1$, with symbol

$$|k|^2\, I - kk^T + kk^T \;=\; |k|^2\, I,$$

elliptic, and the propagator is the familiar Feynman-gauge $\delta_{\mu\nu}/|k|^2$. More general R_ξ gauges give $|k|^2\, I - (1-\xi^{-1})kk^T$, which is elliptic for any finite $\xi > 0$. For Yang–Mills theory the same happens with d_A in place of d: the linearised field equation $\mathrm{d}_A^*\mathrm{d}_A$ is not elliptic, and adding the background-field gauge-fixing term $\left\|\mathrm{d}_A^*\,\delta A\right\|^2$ produces the elliptic operator Δ_A. This is the step that makes the one-loop determinant $\det \Delta_A$ and the instanton moduli counting of Chapter 8 meaningful.

2.3.11 The de Rham complex and its symbol sequence

We end the section on operators with an idea that Nash postpones to Chapter 4 but that is already visible here, and that explains *why* the symbol of Δ_p came out invertible. The operators d form a chain

$$0 \longrightarrow \Omega^0(M) \xrightarrow{\mathrm{d}} \Omega^1(M) \xrightarrow{\mathrm{d}} \Omega^2(M) \xrightarrow{\mathrm{d}} \cdots \xrightarrow{\mathrm{d}} \Omega^n(M) \longrightarrow 0, \tag{2.100}$$

the **de Rham complex**, with $\mathrm{d}\circ\mathrm{d} = 0$. Take its leading symbols at a point (x,k) with $k \neq 0$. We get a chain of finite-dimensional vector spaces and linear maps,

$$0 \longrightarrow \Lambda^0 T_x^* \xrightarrow{ik\wedge} \Lambda^1 T_x^* \xrightarrow{ik\wedge} \Lambda^2 T_x^* \xrightarrow{ik\wedge} \cdots \xrightarrow{ik\wedge} \Lambda^n T_x^* \longrightarrow 0, \tag{2.101}$$

the **symbol sequence**. It is again a complex, $(ik\wedge)(ik\wedge) = -k\wedge k\wedge = 0$, because $k\wedge k = 0$. (This is also Rule 1 applied to $\mathrm{d}\circ\mathrm{d} = 0$.)

Working definition 2.19 (exact sequence)

A sequence of linear maps $\cdots \to A_{p-1} \xrightarrow{f_{p-1}} A_p \xrightarrow{f_p} A_{p+1} \to \cdots$ with $f_p \circ f_{p-1} = 0$ is **exact at** A_p if $\operatorname{Ker} f_p = \operatorname{Im} f_{p-1}$: everything killed by the next map comes from the previous one. It is *exact* if it is exact at every place. *How to test*: check $\dim \operatorname{Ker} f_p = \dim \operatorname{Im} f_{p-1}$ at each place, i.e. $\dim A_p = \operatorname{rank} f_{p-1} + \operatorname{rank} f_p$ [$\operatorname{rank} f = \dim \operatorname{Im} f$, the number of independent output vectors].

Key idea: The symbol sequence of the de Rham complex is exact

For every $k \neq 0$ the sequence (2.101) is exact: a form v with $k \wedge v = 0$ is of the form $v = k \wedge w$.

Why. The key identity (2.96) does all the work. Suppose $k \wedge v = 0$. Then

$$v \overset{(1)}{=} \frac{1}{|k|^2}\Big(k \wedge \iota_k v + \iota_k(k \wedge v)\Big) \overset{(2)}{=} \frac{1}{|k|^2}\, k \wedge \iota_k v \;=\; k \wedge w, \qquad w = \frac{\iota_k v}{|k|^2}.$$

(1) The key identity, divided by $|k|^2 \neq 0$. This is where $k \neq 0$ is used.

(2) The second term vanishes because $k \wedge v = 0$.

So $\mathrm{Ker}(k\wedge) \subset \mathrm{Im}(k\wedge)$, and $\mathrm{Im} \subset \mathrm{Ker}$ because $k \wedge k = 0$. The operator $h = \iota_k/|k|^2$ is a **contracting homotopy**: $h\,(k\wedge) + (k\wedge)\,h = 1$. At the ends: at Λ^0, $k \wedge f = 0$ forces $f = 0$ (injective), and at Λ^n every n-form is $k \wedge$ (something) (surjective). At $k = 0$, on the other hand, all maps are zero and the sequence is as far from exact as possible.

Worked example 2.36: Exactness for grad–curl–div, by hand

In $\mathbb{R}^3$ the symbol sequence is (dropping the i's, which do not change kernels and images)

$$0 \to \mathbb{R} \xrightarrow{\vec{k}} \mathbb{R}^3 \xrightarrow{\vec{k}\times} \mathbb{R}^3 \xrightarrow{\vec{k}\cdot} \mathbb{R} \to 0.$$

Check each place with $\vec{k} \neq 0$:

- at the first $\mathbb{R}$: $f\vec{k} = 0 \Rightarrow f = 0$. Kernel 0, image of the incoming map 0. Exact.
- at the first $\mathbb{R}^3$: $\vec{k}\times\vec{v} = 0 \iff \vec{v} \parallel \vec{k} \iff \vec{v} = f\vec{k}$. Kernel = image = the line through $\vec{k}$. (Physics: a curl-free plane wave is a gradient.)
- at the second $\mathbb{R}^3$: $\vec{k}\cdot\vec{w} = 0 \iff \vec{w} \perp \vec{k}$, and the image of $\vec{k}\times$ is exactly the plane $\vec{k}^\perp$ (Figure 2.18). Kernel = image. (Physics: a divergence-free plane wave is a curl.)
- at the last $\mathbb{R}$: $\vec{k}\cdot$ is onto, since $\vec{k}\cdot(c\,\vec{k}/|k|^2) = c$. Exact.

Count dimensions: $1 = 0+1, 3 = 1+2, 3 = 2+1, 1 = 1+0$ (rank in plus rank out). Figure 2.21 shows this bookkeeping. A necessary consequence of exactness is that the alternating sum of the dimensions vanishes: $1 - 3 + 3 - 1 = 0$. In general $\sum_p (-1)^p \binom{n}{p} = (1-1)^n = 0$.

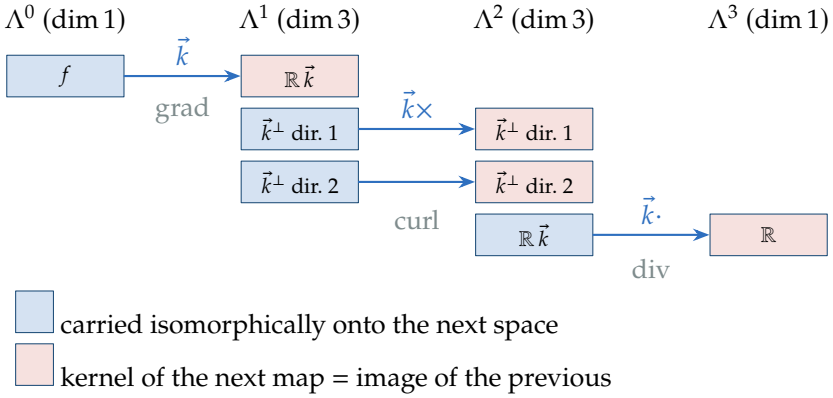


Figure 2.21: Exactness of the symbol sequence of the de Rham complex of $\mathbb{R}^3$, for a fixed $\vec{k} \neq 0$. Each fibre Λ^p splits into a part hit by the incoming map (red) and a part carried isomorphically onto the next space (blue). Nothing is left over anywhere: that is exactness. The dimensions obey $\binom{3}{p}$ = (red) + (blue), which is why $1 - 3 + 3 - 1 = 0$.

Elliptic complexes

The exactness we just found is precisely the reason that Δ_p is elliptic, and it is the pattern for a whole class of objects.

Working definition 2.20 (elliptic complex)

A sequence of differential operators $0 \to \Gamma(E_0) \xrightarrow{D_0} \Gamma(E_1) \xrightarrow{D_1} \cdots \xrightarrow{D_{N-1}} \Gamma(E_N) \to 0$ [$\Gamma(E_p)$ is the space of smooth sections of the bundle E_p, like $\Omega^p(M)$ for d], all of the same order, with $D_{p+1} \circ D_p = 0$, is an **elliptic complex** if its symbol sequence [the leading symbols at one point x, linear maps between the fibres, like $ik\wedge$ for d]

$$0 \to E_{0,x} \xrightarrow{\sigma(D_0)(x,k)} E_{1,x} \to \cdots \to E_{N,x} \to 0$$

is exact for every x and every $k \neq 0$. *How to test*: compute the symbol matrices, and at each place check $\dim E_p = \operatorname{rank}\sigma(D_{p-1}) + \operatorname{rank}\sigma(D_p)$.

A single elliptic operator $L : \Gamma(U) \to \Gamma(V)$ is the special case $N = 1$: $0 \to U_x \xrightarrow{\sigma(L)} V_x \to 0$ is exact exactly when $\sigma(L)$ is injective and surjective, i.e. invertible. So elliptic complexes generalise elliptic operators. Conversely, every elliptic complex produces elliptic operators, its Laplacians.

Key idea: Exact symbol sequence $\Longleftrightarrow$ elliptic Laplacians

For a complex with symbols $s_p = \sigma(D_p)(x,k)$, the symbol of the Laplacian $\Delta_p = D_{p-1}D_{p-1}^* + D_p^*D_p$ is $s_{p-1}s_{p-1}^* + s_p^*s_p$, and this is invertible for all $k \neq 0$ if and only if the symbol sequence is exact at E_p.

Why (the direction we need). Suppose the sequence is exact and $(s_{p-1}s_{p-1}^* + s_p^*s_p)v = 0$. Then

$$0 \overset{(1)}{=} \left(v, (s_{p-1}s_{p-1}^* + s_p^*s_p)v\right) \overset{(2)}{=} \left|s_{p-1}^*v\right|^2 + \left|s_pv\right|^2,$$

(1) Take the inner product of the equation with v.
(2) Move each s across the inner product: $(v, ss^*v) = (s^*v, s^*v)$ and $(v, s^*sv) = (sv, sv)$.

So $s_pv = 0$ and $s_{p-1}^*v = 0$. By exactness, $s_pv = 0$ means $v = s_{p-1}w$ for some w. Then $0 = (w, s_{p-1}^*v) = (s_{p-1}w, v) = |v|^2$, so $v = 0$. The symbol of Δ_p is injective, hence (being a square matrix) invertible. This is the *same* argument as Nash's (2.50) for harmonic forms, which we meet in the next subsection, but performed on the finite-dimensional fibres at one momentum.

Worked example 2.37: Rolling up the complex: $\mathrm{d} + \mathrm{d}^*$ and the Euler characteristic operator

Collect the even and odd forms, $\Omega^{\mathrm{ev}} = \bigoplus_p \Omega^{2p}$ and $\Omega^{\mathrm{odd}} = \bigoplus_p \Omega^{2p+1}$, and consider

$$\mathrm{d} + \mathrm{d}^* : \Omega^{\mathrm{ev}}(M) \longrightarrow \Omega^{\mathrm{odd}}(M).$$

Its symbol is $s = i(k \wedge -\iota_k)$, and, using $k \wedge k\wedge = 0 = \iota_k\iota_k$,

$$s^*s \overset{(1)}{=} (-i)(k \wedge -\iota_k)^*\, i(k \wedge -\iota_k) \overset{(2)}{=} (\iota_k - k\wedge)(k \wedge -\iota_k) \overset{(3)}{=} \iota_k k \wedge + k \wedge \iota_k \overset{(4)}{=} |k|^2.$$

(1) $s^* = \bar{i}\,(k \wedge -\iota_k)^* = -i(k \wedge -\iota_k)^*$.
(2) $(k\wedge)^* = \iota_k$ and $\iota_k^* = k\wedge$; and $(-i)(i) = 1$.
(3) Expand; the terms $\iota_k\iota_k$ and $k \wedge k\wedge$ vanish.
(4) The key identity.

So $\mathrm{d} + \mathrm{d}^*$ is an elliptic operator between the bundles of even and odd forms. (They have equal rank, $\sum_p \binom{n}{2p} = \sum_p \binom{n}{2p+1} = 2^{n-1}$, as they must.) It is a first-order square root of the Laplacian, like the Dirac operator; indeed the div–grad–curl operator of Example 2.29

is exactly $\mathrm{d}+\mathrm{d}^*$ on $\Omega^0\oplus\Omega^2\to\Omega^1\oplus\Omega^3$ in $\mathbb{R}^3$, with 2-forms stored as vectors via $\star$. In Chapter 4 we shall see that the number of its zero modes minus the number of zero modes of its adjoint is the Euler characteristic, $\mathrm{index}(\mathrm{d}+\mathrm{d}^*)=\sum_p(-1)^p b_p=\chi(M)$.

Worked example 2.38: The instanton deformation complex is elliptic

Here is an elliptic complex that appears in Chapter 8. On a 4-manifold, 2-forms split into self-dual and anti-self-dual parts, $\Lambda^2=\Lambda^2_+\oplus\Lambda^2_-$, with $\star=\pm1$ on $\Lambda^2_\pm$ (each of rank 3). Consider

$$0\to\Omega^0\xrightarrow{\mathrm{d}}\Omega^1\xrightarrow{\mathrm{d}^+}\Omega^2_+\to 0,\qquad \mathrm{d}^+\omega=\tfrac12(\mathrm{d}\omega+\star\mathrm{d}\omega).$$

(With a gauge field, d becomes d_A; the first map is an infinitesimal gauge transformation and the second is the linearised self-duality equation $F_A^+=0$.) The symbols are $f\mapsto ikf$ and $a\mapsto i(k\wedge a)^+$. Take $k=e^1$ (rotate and rescale). A basis of Λ^2_+ is $\{e^{12}+e^{34},\ e^{13}+e^{42},\ e^{14}+e^{23}\}$ (with $e^{ij}=e^i\wedge e^j$). Then

$$(e^1\wedge e^2)^+=\tfrac12(e^{12}+e^{34}),\quad (e^1\wedge e^3)^+=\tfrac12(e^{13}+e^{42}),\quad (e^1\wedge e^4)^+=\tfrac12(e^{14}+e^{23}),\quad (e^1\wedge e^1)^+=0.$$

So $a\mapsto(k\wedge a)^+$ maps Λ^1 (rank 4) *onto* Λ^2_+ (rank 3), with kernel spanned by $e^1=k$, which is exactly the image of Λ^0. The first map is injective. Exact everywhere:

$$\text{ranks } 1,\ 4,\ 3:\qquad 1=0+1,\quad 4=1+3,\quad 3=3+0.$$

The index of this complex (computed in Chapter 8) gives the dimension $8k-3$ of the moduli space of SU(2) instantons of charge k on S^4. Its exactness at Λ^1 is, once more, gauge invariance: the only polarisations the linearised self-duality equation cannot see are pure gauge.

Physical picture: Complexes: ellipticity "modulo gauge"

The Maxwell operator $\mathrm{d}^*\mathrm{d}$ failed to be elliptic only because of gauge directions. The complex $\Omega^0\xrightarrow{\mathrm{d}}\Omega^1\xrightarrow{\mathrm{d}}\Omega^2$ says: "the only directions the field strength cannot see are the gauge directions, and nothing else." That is exactly the statement that the symbol sequence is exact at Λ^1. An elliptic complex is the right notion of ellipticity for a gauge theory, and the first map is the ghost of BRST quantisation.

2.3.12 Harmonic forms and Hodge's theorem

We have shown that Δ_p is elliptic. Now we cash in. Nash ends §3 with the result that opened this chapter, (2.1): the zero modes of Δ_p count the p-dimensional holes of M. The question is: *why should solutions of a PDE know about topology?* The answer has two halves. The first is a short computation with inner products (Nash's (2.50)). It shows that a zero mode of Δ_p is a closed form, and so it defines a cohomology class. The second is Hodge's theorem proper: every cohomology class contains exactly one zero mode. Ellipticity is what makes the second half work.

Throughout, M is *compact*, *without boundary*, *oriented* and *Riemannian*. Each of the four words will be used, and we will point out where.

Working definition 2.21 (harmonic form)

A p-form ω on M is **harmonic** if $\Delta_p\omega = 0$ [$\Delta_p = \mathrm{dd}^* + \mathrm{d}^*\mathrm{d}$, the Laplacian on p-forms], i.e. if it lies in $\operatorname{Ker}\Delta_p$. We write

$$\mathcal{H}^p(M) \;:=\; \operatorname{Ker}\Delta_p \;=\; \{\omega \in \Omega^p(M) :\; \Delta_p\omega = 0\}.$$

How to test: check the two first-order equations $\mathrm{d}\omega = 0$ and $\mathrm{d}^*\omega = 0$ [d^* is the adjoint of d, lowering the degree by one] (this is equivalent on a compact manifold, as we show next). Physically: ω is a *static, source-free field configuration*, a zero mode of the Laplacian.

For functions ($p = 0$) this is the harmonic-function notion of Section 2.1.4, with $\Delta_0 = -\nabla^2$ in flat space. For $p = 1$ in $\mathbb{R}^3$, (2.88) says $\Delta_1\vec{A} = -\nabla^2\vec{A}$, so a harmonic 1-form is a vector field whose components are harmonic functions.

Zero modes are closed and co-closed: Nash's (2.50)

Suppose $\Delta_p\omega = 0$. Pair the equation with ω itself:

$$\begin{aligned} 0 &\overset{(1)}{=} \langle\omega, \Delta_p\omega\rangle \overset{(2)}{=} \langle\omega, \mathrm{dd}^*\omega\rangle + \langle\omega, \mathrm{d}^*\mathrm{d}\omega\rangle \\ &\overset{(3)}{=} \langle\mathrm{d}^*\omega, \mathrm{d}^*\omega\rangle + \langle\mathrm{d}\omega, \mathrm{d}\omega\rangle \overset{(4)}{=} \|\mathrm{d}^*\omega\|^2 + \|\mathrm{d}\omega\|^2 . \end{aligned} \tag{2.102}$$

(1) $\Delta_p\omega = 0$, and the inner product with 0 is 0.
(2) The definition (2.81) of Δ_p, and linearity of the inner product.
(3) The defining property (2.85) of d^*, used twice: once to move d off the second slot as d^* onto the first, once to move d^* off as d (the inner product is symmetric). This is an integration by parts; there are no boundary terms because $\partial M = \varnothing$.
(4) The definition $\|\eta\|^2 = \langle\eta, \eta\rangle = \int_M \eta \wedge \star\eta$.

Both terms in the last line are *non-negative*, because the metric is Riemannian (Section 2.3.7: in an orthonormal coframe $\eta \wedge \star\eta = \sum_I \eta_I^2\,\mathrm{vol}$). A sum of two non-negative numbers is zero only if both are zero. So $\mathrm{d}\omega = 0$ and $\mathrm{d}^*\omega = 0$. The converse is immediate: if $\mathrm{d}\omega = \mathrm{d}^*\omega = 0$ then $\Delta_p\omega = \mathrm{d}(\mathrm{d}^*\omega) + \mathrm{d}^*(\mathrm{d}\omega) = 0$. Hence Nash's (2.51):

$$\omega \in \operatorname{Ker}\Delta_p \quad\Longleftrightarrow\quad \mathrm{d}\omega = 0 \;\text{ and }\; \mathrm{d}^*\omega = 0. \tag{2.103}$$

One second-order equation has become two first-order equations. In vector-calculus language in $\mathbb{R}^3$ (Example 2.32): a harmonic 1-form is a field with $\nabla\times\vec{A} = 0$ and $\nabla\cdot\vec{A} = 0$; a field with neither curl nor divergence, i.e. with no sources of either kind.

Physical picture: The energy argument

Line (2.102) is the familiar statement that the energy of a static field is a sum of squares. For $p = 1$ read ω as a vector potential A: $\|\mathrm{d}A\|^2$ is the magnetic energy $\int \left|\vec{B}\right|^2$, and $\|\mathrm{d}^*A\|^2 = \int(\nabla\cdot\vec{A})^2$ is the gauge-fixing term of Example 2.35. The quadratic form $\langle\omega, \Delta_p\omega\rangle$ is *positive semi-definite*, and its null vectors are the configurations with zero energy. Positivity is also why Δ_p has no negative eigenvalues: if $\Delta_p\omega = \lambda\omega$ then $\lambda\,\|\omega\|^2 = \|\mathrm{d}\omega\|^2 + \|\mathrm{d}^*\omega\|^2 \geq 0$.

Watch out: Each hypothesis is needed

- *Compactness / no boundary.* On $\mathbb{R}$ the function $u(x) = x$ has $\Delta_0 u = -u'' = 0$ but $\mathrm{d}u = \mathrm{d}x \neq 0$. The integration by parts in step (3) fails: $\int_{\mathbb{R}} u\,u''$ and $\int_{\mathbb{R}} (u')^2$ are both undefined, and x is not normalisable. On a compact M harmonic functions are constant, as we are about to see.
- *Riemannian signature.* On the Lorentzian torus T^2 with coordinates (t, x), both of period 2π, the analogue of Δ_0 is the wave operator $\Box = \partial_t^2 - \partial_x^2$. The "energy" becomes $\int \big((\partial_t u)^2 - (\partial_x u)^2\big)$, which is not positive. The plane waves $e^{i(mt+nx)}$ have $\Box e^{i(mt+nx)} = (n^2 - m^2)e^{i(mt+nx)}$, so every pair with $|m| = |n|$ is a zero mode: $\operatorname{Ker}\Box$ is *infinite-dimensional* (all $f(t-x) + g(t+x)$ with f, g periodic). There is no hope of counting holes with it. This is the global face of the local fact (Section 2.1) that the wave operator is hyperbolic, not elliptic.

A harmonic form defines a cohomology class: Nash's (2.52)

Recall from Chapter T2 that the p-th de Rham cohomology group is

$$H^p_{\mathrm{dR}}(M) \;=\; \frac{\operatorname{Ker}\big(\mathrm{d} : \Omega^p \to \Omega^{p+1}\big)}{\operatorname{Im}\big(\mathrm{d} : \Omega^{p-1} \to \Omega^p\big)} \;=\; \frac{\text{closed } p\text{-forms}}{\text{exact } p\text{-forms}},$$

whose elements are classes $[\alpha] = \{\alpha + \mathrm{d}\beta\}$ of closed forms, two closed forms being equivalent if they differ by an exact form. Its dimension is the **Betti number** b_p. (Nash writes $H^p(M;\mathbb{R})$, or $H^p_{\text{de Rham}}(M)$, and quotes his (1.43) of Chapter 1 for the definition.) Since a harmonic form is closed by (2.103), it has a class, and we get a linear map

$$\mathcal{H}^p(M) \longrightarrow H^p_{\mathrm{dR}}(M), \qquad \omega \longmapsto [\omega], \tag{2.104}$$

Nash's (2.52). The content of Hodge's theorem is that this map is an isomorphism.

Key idea: Hodge's theorem (Nash p. 41)

On a compact, closed, oriented Riemannian manifold there is *precisely one* harmonic p-form in each de Rham cohomology class. Equivalently, (2.104) is an isomorphism, and

$$\dim \operatorname{Ker} \Delta_p \;=\; \dim H^p_{\mathrm{dR}}(M) \;=\; b_p, \tag{2.105}$$

Nash's (2.53). The left side needs a metric; the right side does not. So the *number* of harmonic forms is independent of the metric, even though the harmonic forms themselves move when the metric changes.

Nash refers to Wells for the proof. We give it in two steps: "at most one" (injectivity) is three lines, and "at least one" (surjectivity) is where ellipticity enters.

At most one harmonic form per class. Suppose ω and ω' are harmonic and $[\omega] = [\omega']$. Then $\eta = \omega - \omega'$ is harmonic (the kernel is a vector space) and exact, $\eta = \mathrm{d}\beta$. Now

$$\|\eta\|^2 \overset{(1)}{=} \langle \eta, \mathrm{d}\beta \rangle \overset{(2)}{=} \langle \mathrm{d}^*\eta, \beta \rangle \overset{(3)}{=} 0.$$

(1) $\eta = \mathrm{d}\beta$ in the second slot.
(2) Move d across as d^*, by (2.85).

(3) $\mathrm{d}^*\eta = 0$, since η is harmonic, by (2.103).

So $\eta = 0$ and $\omega = \omega'$. In words: *a harmonic form is orthogonal to every exact form*, so the only exact harmonic form is zero. The map (2.104) is injective.

At least one: the variational (Dirichlet) picture. Fix a closed form α. Among all forms in its class, $\alpha + \mathrm{d}\beta$, look for the one of *least energy* $E(\beta) = \|\alpha + \mathrm{d}\beta\|^2$. At a minimum, the first variation in every direction $\mathrm{d}\eta$ vanishes:

$$0 \overset{(1)}{=} \frac{\mathrm{d}}{\mathrm{d}\epsilon}\Big|_{\epsilon=0} \|\alpha + \mathrm{d}\beta + \epsilon\,\mathrm{d}\eta\|^2 \overset{(2)}{=} 2\,\langle \alpha + \mathrm{d}\beta, \mathrm{d}\eta\rangle \overset{(3)}{=} 2\,\langle \mathrm{d}^*(\alpha + \mathrm{d}\beta), \eta\rangle \qquad \text{for all } \eta \in \Omega^{p-1}.$$

(1) The condition for β to be a critical point of E.
(2) Expand $\|a + \epsilon b\|^2 = \|a\|^2 + 2\epsilon\langle a, b\rangle + \epsilon^2\,\|b\|^2$ (real forms).
(3) Move d across as d^*.

Since η is arbitrary, the Euler–Lagrange equation is $\mathrm{d}^*(\alpha + \mathrm{d}\beta) = 0$. The minimiser $h = \alpha + \mathrm{d}\beta$ is closed (it is in the class of α) and co-closed, so it is harmonic. This is the physicist's derivation: *the harmonic representative is the minimum-energy field with the prescribed topological fluxes* (the periods $\int_c \alpha$ over closed cycles c, which are the same for every form in the class, by Stokes). A magnetic field on a torus with given fluxes through the two holes settles into the harmonic one.

What the variational argument does *not* show is that a minimiser exists, and that it is smooth. The infimum of E could be approached by a sequence of forms that gets rougher and rougher and converges to nothing smooth. That is exactly the gap that ellipticity closes, by the mechanism we now describe.

Where ellipticity enters: discrete spectrum and the Green's operator

Ellipticity gives three facts about Δ_p on a compact manifold. We have seen their local versions in Sections 2.1.4 and 2.2; the global versions are proved with the Sobolev spaces of part 2b.

1. *Regularity.* If $\Delta_p\omega$ is smooth then ω is smooth (even if ω was only known to be a weak, square-integrable solution). This is Weyl's lemma (Section 2.1.4) for systems. In particular every eigenform is smooth.

2. *Discrete spectrum.* Δ_p has an orthonormal basis of eigenforms ω_j with eigenvalues $0 \le \lambda_0 \le \lambda_1 \le \cdots \to \infty$, each of *finite multiplicity*. The mechanism is the symbol: $\sigma_2 = |k|^2\, I$ grows in *every* direction of momentum space, so there are only finitely many modes below any energy (Weyl's law: roughly $\binom{n}{p}\,\mathrm{vol}(M)\,\lambda^{n/2}/(4\pi)^{n/2}\Gamma(\frac{n}{2}+1)$ of them below λ). Compare the Lorentzian torus of the pitfall above, where the symbol $n^2 - m^2$ vanishes along the light cone and infinitely many modes sit at zero.

3. *A spectral gap.* In particular $\operatorname{Ker}\Delta_p = \mathcal{H}^p$ is finite-dimensional, and the smallest *non-zero* eigenvalue λ_* is strictly positive.

With these facts, let $P_{\mathcal{H}}$ be the orthogonal projection onto $\mathcal{H}^p$ and define the **Green's operator** G by inverting Δ_p on the orthogonal complement of its kernel:

$$G\,\omega_j \;=\; \begin{cases} \lambda_j^{-1}\,\omega_j, & \lambda_j > 0, \\ 0, & \lambda_j = 0. \end{cases} \qquad \text{so} \qquad \Delta_p G \;=\; G\Delta_p \;=\; 1 - P_{\mathcal{H}}. \tag{2.106}$$

The gap makes G bounded ($\|G\| \le 1/\lambda_*$), and regularity makes it map smooth forms to smooth forms. (In physics language: G is the propagator with the zero modes removed, just as one

removes zero modes before inverting the fluctuation operator in an instanton calculation.) Now every form splits in one line. For any $\omega \in \Omega^p(M)$,

$$\omega \overset{(1)}{=} P_{\mathcal{H}}\omega + \Delta_p G\omega \overset{(2)}{=} P_{\mathcal{H}}\omega + \mathrm{d}\big(\mathrm{d}^* G\omega\big) + \mathrm{d}^*\big(\mathrm{d}G\omega\big). \tag{2.107}$$

(1) (2.106): $1 = P_{\mathcal{H}} + \Delta_p G$.
(2) The definition $\Delta_p = \mathrm{d}\mathrm{d}^* + \mathrm{d}^*\mathrm{d}$.

The three pieces are harmonic, exact and *co-exact* (in the image of d^*).

Key idea: The Hodge decomposition

On a compact closed Riemannian manifold, every p-form splits uniquely and orthogonally as

$$\Omega^p(M) \;=\; \mathcal{H}^p(M) \;\oplus\; \mathrm{d}\Omega^{p-1}(M) \;\oplus\; \mathrm{d}^*\Omega^{p+1}(M), \qquad \omega \;=\; h + \mathrm{d}\beta + \mathrm{d}^*\gamma, \tag{2.108}$$

with $h = P_{\mathcal{H}}\omega$, $\beta = \mathrm{d}^* G\omega$, $\gamma = \mathrm{d}G\omega$. The closed forms are exactly $\mathcal{H}^p \oplus \mathrm{d}\Omega^{p-1}$.

Why the pieces are orthogonal. For any $h \in \mathcal{H}^p$, $\beta \in \Omega^{p-1}$, $\gamma \in \Omega^{p+1}$:

$$\langle \mathrm{d}\beta, \mathrm{d}^*\gamma\rangle \overset{(1)}{=} \langle \mathrm{d}\mathrm{d}\beta, \gamma\rangle = 0, \qquad \langle h, \mathrm{d}\beta\rangle \overset{(2)}{=} \langle \mathrm{d}^* h, \beta\rangle = 0, \qquad \langle h, \mathrm{d}^*\gamma\rangle \overset{(3)}{=} \langle \mathrm{d}h, \gamma\rangle = 0.$$

(1) Move d^* across as d; then $\mathrm{d}\circ\mathrm{d} = 0$.
(2) Move d across; $\mathrm{d}^* h = 0$ because h is harmonic.
(3) Move d^* across; $\mathrm{d}h = 0$ because h is harmonic.

Orthogonal subspaces meet only in 0, which gives uniqueness of the splitting.

Why closed forms have no co-exact part. Let α be closed with decomposition $\alpha = h + \mathrm{d}\beta + \mathrm{d}^*\gamma$. Apply d: $0 = \mathrm{d}\alpha = \mathrm{d}\mathrm{d}^*\gamma$ (as $\mathrm{d}h = 0$ and $\mathrm{d}\mathrm{d}\beta = 0$). Then

$$\|\mathrm{d}^*\gamma\|^2 \overset{(1)}{=} \langle \mathrm{d}^*\gamma, \mathrm{d}^*\gamma\rangle \overset{(2)}{=} \langle \gamma, \mathrm{d}\mathrm{d}^*\gamma\rangle \overset{(3)}{=} 0,$$

(1) Definition of the norm.
(2) Move the first d^* across as d.
(3) $\mathrm{d}\mathrm{d}^*\gamma = \mathrm{d}\alpha = 0$.

so $\mathrm{d}^*\gamma = 0$ and $\alpha = h + \mathrm{d}\beta$. That is surjectivity of (2.104): $[\alpha] = [h]$ with h harmonic. Together with injectivity, this proves Hodge's theorem, modulo the three analytic facts supplied by ellipticity.

The picture also explains the minimisation: since $h \perp \mathrm{d}\beta$, $\|h + \mathrm{d}\beta\|^2 = \|h\|^2 + \|\mathrm{d}\beta\|^2 \geq \|h\|^2$, with equality only at $\mathrm{d}\beta = 0$. The harmonic representative is the *foot of the perpendicular* from the origin to the class.

Dictionary: Hodge decomposition in physics

- In $\mathbb{R}^3$ ($p = 1$, suitably decaying fields): the **Helmholtz decomposition** $\vec{A} = \nabla\phi + \nabla\times\vec{C}$ is $\mathrm{d}\beta + \mathrm{d}^*\gamma$ (longitudinal + transverse). On a compact space a third, harmonic, piece appears.
- On a torus T^3 (a box with periodic boundary conditions, as in lattice gauge theory), $\vec{A} = \vec{a} + \nabla\phi + \nabla\times\vec{C}$, where $\vec{a}$ is a *constant* vector: the zero-momentum mode, which is neither a gradient nor a curl of a periodic field.

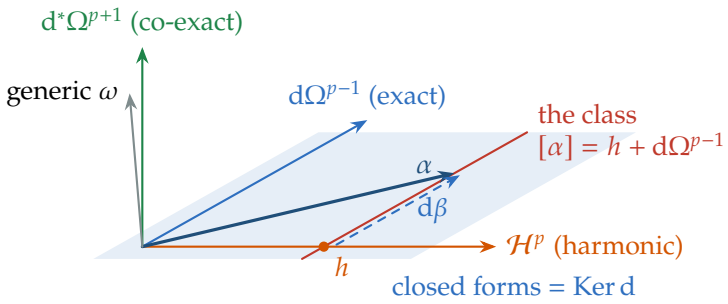


Figure 2.22: The Hodge decomposition (2.108), drawn as three orthogonal axes (each really an infinite-dimensional space, except $\mathcal{H}^p$, which is finite-dimensional). The closed forms fill the shaded plane spanned by the harmonic and exact directions; a generic form (grey) sticks out of it into the co-exact direction. A cohomology class is a line (red) parallel to the exact axis. It meets the harmonic axis in exactly one point h, the harmonic representative, which is also the point of the line closest to the origin: the minimum-energy member of the class.

- In Maxwell theory the three pieces are: the pure-gauge part $\mathrm{d}\beta$, removed by gauge fixing; the transverse photons $\mathrm{d}^*\gamma$; and the harmonic part, the topological "zero modes" (constant potentials whose Wilson loops around the cycles of M are gauge invariant).

2.3.13 A finite-dimensional model: Hodge theory on a triangle

Everything above is linear algebra, except the analysis that guarantees the infinite-dimensional spaces behave. So we can watch Hodge's theorem work *exactly* in a finite-dimensional model, where the analysis is trivial. Replace the circle by a triangle: three vertices v_1, v_2, v_3 and three oriented edges $e_1 = v_1 \to v_2$, $e_2 = v_2 \to v_3$, $e_3 = v_3 \to v_1$ (Figure 2.23).

- A "0-form" is a function on vertices, $f = (f_1, f_2, f_3) \in \mathbb{R}^3$.
- A "1-form" (a **cochain**) assigns a number to each edge, $a = (a_1, a_2, a_3) \in \mathbb{R}^3$: think of a_j as $\int_{e_j} a$.
- The "exterior derivative" is the discrete Stokes theorem: $(\mathrm{d}f)(e) = f(\text{head}) - f(\text{tail})$.

So $\mathrm{d}f = (f_2 - f_1,\ f_3 - f_2,\ f_1 - f_3)$, and with the standard inner products on $\mathbb{R}^3$,

$$\mathrm{d} = \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 1 & 0 & -1 \end{pmatrix}, \qquad \mathrm{d}^* = \mathrm{d}^T = \begin{pmatrix} -1 & 0 & 1 \\ 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix}. \tag{2.109}$$

There are no 2-cells, so d on 1-forms is zero and every 1-cochain is "closed".

Worked example 2.39: The discrete Laplacians and their zero modes

The two Laplacians are $\Delta_0 = \mathrm{d}^*\mathrm{d} = \mathrm{d}^T\mathrm{d}$ (on vertices) and $\Delta_1 = \mathrm{d}\mathrm{d}^* = \mathrm{d}\,\mathrm{d}^T$ (on edges; the other term $\mathrm{d}^*\mathrm{d}$ vanishes since $\mathrm{d} = 0$ on 1-forms). Multiply out:

$$\Delta_0 \overset{(1)}{=} \begin{pmatrix} -1 & 0 & 1 \\ 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix}\begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 1 & 0 & -1 \end{pmatrix} \overset{(2)}{=} \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix},$$

$$\Delta_1 \overset{(3)}{=} \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 1 & 0 & -1 \end{pmatrix} \begin{pmatrix} -1 & 0 & 1 \\ 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix} \overset{(4)}{=} \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}.$$

(1) $\Delta_0 = \mathrm{d}^T \mathrm{d}$ with the matrices (2.109).

(2) Row times column: e.g. the $(1,1)$ entry is $(-1)(-1) + 0 + (1)(1) = 2$ and the $(1,2)$ entry is $(-1)(1) + 0 + (1)(0) = -1$. This is the **graph Laplacian**: degree on the diagonal, -1 for each neighbour. It is the finite-difference version of $-\partial_\theta^2$, $-(f_{j+1} - 2f_j + f_{j-1})$.

(3) $\Delta_1 = \mathrm{d}\,\mathrm{d}^T$.

(4) Row times column again; e.g. the $(1,2)$ entry pairs edge e_1 with edge e_2, which share the vertex v_2 with opposite orientations, giving -1.

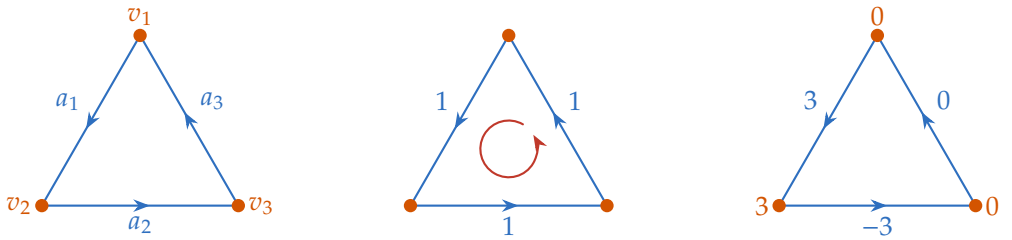
Both matrices are $3I - J$, where J is the all-ones matrix. J has eigenvalues 3 (on $(1,1,1)$) and 0 (twice, on vectors with zero sum), so both Laplacians have spectrum $\{0,3,3\}$. The zero modes:

- $\operatorname{Ker}\Delta_0 = \mathbb{R}(1,1,1)$: the *constant* functions. So $b_0 = 1$: the triangle is connected.
- $\operatorname{Ker}\Delta_1 = \mathbb{R}(1,1,1)$: the cochain assigning 1 to every edge. This is the discrete $\mathrm{d}\theta$, a uniform circulation around the loop. So $b_1 = 1$: the triangle has one hole.

Now check Hodge's theorem directly. The exact cochains are $\operatorname{Im}\mathrm{d} = \{a : a_1 + a_2 + a_3 = 0\}$, a plane (the sum telescopes: $(f_2 - f_1) + (f_3 - f_2) + (f_1 - f_3) = 0$). All cochains are closed, so $H^1 = \mathbb{R}^3/\operatorname{Im}\mathrm{d} \cong \mathbb{R}$, the isomorphism being the total circulation $a \mapsto a_1 + a_2 + a_3$ (the discrete $\oint a$). The harmonic line $\mathbb{R}(1,1,1)$ is *orthogonal* to the exact plane, and meets each class (each plane $a_1 + a_2 + a_3 = c$) in exactly one point, $\frac{c}{3}(1,1,1)$. That is Figure 2.22 with actual numbers. For instance

$$a = (5,-1,2) \;=\; \underbrace{(2,2,2)}_{\text{harmonic}} + \underbrace{(3,-3,0)}_{\text{exact}}, \qquad (3,-3,0) = \mathrm{d}f \;\text{ with }\; f = (0,3,0),$$

and the harmonic part has the same circulation, 6, as a.



cochain $a = (a_1, a_2, a_3)$ on the edges

harmonic: $(1,1,1)$, circulation 3

exact: $\mathrm{d}f$, $f = (0,3,0)$, circulation 0

Figure 2.23: Hodge theory on a triangle, the discrete circle of Example 2.39. Left: the data (numbers on oriented edges). Middle: the unique harmonic cochain up to scale, a uniform circulation around the hole (red arrow); it is the discrete $\mathrm{d}\theta$. Right: an exact cochain, the differences of a vertex function (orange), with zero circulation. Every cochain is uniquely a multiple of the middle one plus one of the right kind.

Try it: A square instead of a triangle

Do the same for a square with four vertices and four edges in a cycle. *Answer*: Δ_0 and Δ_1 are both the 4×4 cycle Laplacian with 2 on the diagonal and -1 for the two neighbours. Its eigenvalues are $2-2\cos(2\pi j/4) = 0,2,4,2$ ($j=0,1,2,3$), with the zero mode $(1,1,1,1)$ in both degrees. Again $b_0 = b_1 = 1$. With N vertices the eigenvalues are $2-2\cos(2\pi j/N) \approx (2\pi j/N)^2$ for small j: after rescaling by $N^2/(2\pi)^2$ they converge to j^2, the spectrum of $-\partial_\theta^2$ on the circle. The discrete model becomes the smooth one.

2.3.14 Examples of Hodge's theorem

Worked example 2.40: The circle: zero modes, spectrum and pairing

Take $M = S^1$ with coordinate $\theta \in [0,2\pi)$ and the round metric $\mathrm{d}\theta^2$. Then $\star 1 = \mathrm{d}\theta$ and $\star \mathrm{d}\theta = 1$. On functions and on 1-forms $f\,\mathrm{d}\theta$,

$$\Delta_0 u \overset{(1)}{=} \mathrm{d}^*\mathrm{d}u \overset{(2)}{=} -u'', \qquad \Delta_1(f\,\mathrm{d}\theta) \overset{(3)}{=} \mathrm{d}\mathrm{d}^*(f\,\mathrm{d}\theta) \overset{(4)}{=} -f''\,\mathrm{d}\theta.$$

(1) On 0-forms $\mathrm{d}^* = 0$, so only $\mathrm{d}^*\mathrm{d}$ survives.
(2) $\mathrm{d}u = u'\mathrm{d}\theta$, and on 1-forms in dimension $n=1$ the sign in (2.86) is $(-1)^{1+1+1} = -1$, so $\mathrm{d}^*(u'\mathrm{d}\theta) = -\star\,\mathrm{d}\star(u'\mathrm{d}\theta) = -\star\,\mathrm{d}u' = -\star(u''\mathrm{d}\theta) = -u''$.
(3) On 1-forms (top degree) $\mathrm{d} = 0$, so only $\mathrm{d}\mathrm{d}^*$ survives.
(4) $\mathrm{d}^*(f\mathrm{d}\theta) = -f'$ by the same computation, and $\mathrm{d}(-f') = -f''\mathrm{d}\theta$.

The eigenforms are Fourier modes: $\Delta_0 e^{in\theta} = n^2 e^{in\theta}$ and $\Delta_1(e^{in\theta}\mathrm{d}\theta) = n^2 e^{in\theta}\mathrm{d}\theta$. The kernels are the $n=0$ modes,

$$\operatorname{Ker}\Delta_0 = \{\text{constants}\}, \qquad \operatorname{Ker}\Delta_1 = \{c\,\mathrm{d}\theta\},$$

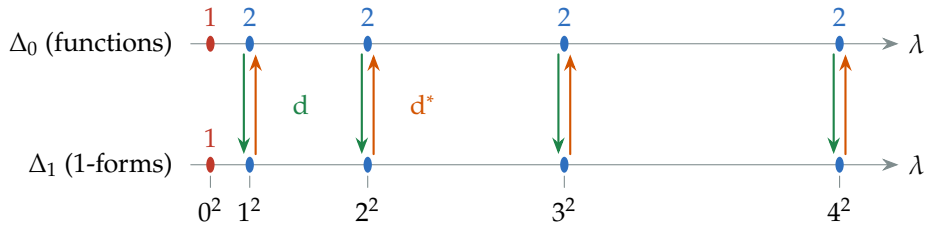
so $b_0 = b_1 = 1$, in agreement with the cohomology of the circle (Chapter T2): $\mathrm{d}\theta$ is closed but not exact, since $\oint \mathrm{d}\theta = 2\pi \neq 0$.
The Hodge decomposition of a 1-form $f(\theta)\,\mathrm{d}\theta$ is its Fourier split into the zero mode and the rest. Let $\bar f = \frac{1}{2\pi}\oint f\,\mathrm{d}\theta$ be the average. Then

$$f\,\mathrm{d}\theta \;=\; \underbrace{\bar f\,\mathrm{d}\theta}_{\text{harmonic}} \;+\; \underbrace{\mathrm{d}g}_{\text{exact}}, \qquad g(\theta) = \int_0^\theta \big(f(s)-\bar f\big)\,\mathrm{d}s,$$

and g is periodic precisely because we subtracted the average. (There is no co-exact part: there are no 2-forms on a circle.) The harmonic part carries all the "flux": $\oint f\mathrm{d}\theta = 2\pi\bar f$.
Finally, notice the *pairing* of the non-zero modes (Figure 2.24). The map d sends the eigenfunction $e^{in\theta}$ of Δ_0 to $in\,e^{in\theta}\mathrm{d}\theta$, an eigenform of Δ_1 with the same eigenvalue n^2; and d^* goes back. For $n\neq 0$ this is a bijection between eigenspaces. Only the zero modes are unpaired: d kills the constants, and d^* kills $\mathrm{d}\theta$. Hence

$$\dim\operatorname{Ker}\Delta_0 - \dim\operatorname{Ker}\Delta_1 \;=\; 1-1 \;=\; 0 \;=\; \chi(S^1),$$

the Euler characteristic. This is the first glimpse of the index theorem of Chapter 4: the non-zero spectrum cancels in pairs, and only topology survives.



non-zero modes paired by d, d*; zero modes (red) unpaired: $\dim\operatorname{Ker}\Delta_0 - \dim\operatorname{Ker}\Delta_1 = \chi(S^1) = 0$

Figure 2.24: The spectra of Δ_0 and Δ_1 on the round circle. Each dot is an eigenvalue n^2; the number above it is its multiplicity ($e^{\pm in\theta}$, or $\cos n\theta$ and $\sin n\theta$). The exterior derivative (green) and its adjoint (orange) map the λ-eigenspace of one Laplacian isomorphically onto that of the other whenever $\lambda \neq 0$. Only the zero modes are left over, and they count the cohomology: $b_0 = b_1 = 1$. Ellipticity is visible as the discreteness of the spectrum and the gap between 0 and the first non-zero eigenvalue.

Worked example 2.41: A lumpy circle: the harmonic form moves, its count does not

Now give S^1 a non-uniform metric $g = h(\theta)^2\,\mathrm{d}\theta^2$ with $h > 0$ smooth and periodic, for example $h = 1 + \frac{1}{2}\cos\theta$. The total length is $\ell = \oint h\,\mathrm{d}\theta$. The oriented orthonormal coframe is $e = h\,\mathrm{d}\theta$ (the arc-length element $\mathrm{d}s$), so

$$\star 1 = e = h\,\mathrm{d}\theta, \qquad \star e = 1 \;\Rightarrow\; \star\mathrm{d}\theta = \frac{1}{h}.$$

Harmonic 1-forms. Every 1-form $\omega = f\,\mathrm{d}\theta$ is closed (top degree). Co-closed means

$$0 \;=\; \mathrm{d}^*\omega \overset{(1)}{=} -\star\,\mathrm{d}\,\star\,(f\,\mathrm{d}\theta) \overset{(2)}{=} -\star\,\mathrm{d}\Big(\frac{f}{h}\Big) \overset{(3)}{=} -\frac{1}{h}\Big(\frac{f}{h}\Big)',$$

(1) (2.86) with $n = p = 1$: sign -1.
(2) $\star(f\,\mathrm{d}\theta) = f\,\star\mathrm{d}\theta = f/h$.
(3) $\mathrm{d}(f/h) = (f/h)'\mathrm{d}\theta$ and $\star\mathrm{d}\theta = 1/h$.

So f/h is constant: the harmonic 1-forms are $\omega = c\,h\,\mathrm{d}\theta = c\,\mathrm{d}s$. The harmonic form *has changed*: it is no longer $\mathrm{d}\theta$, it is the arc-length form, largest where the circle is most stretched. But the space of harmonic 1-forms is still one-dimensional, $b_1 = 1$.
Harmonic functions. $\Delta_0 u = \mathrm{d}^*\mathrm{d}u = -\frac{1}{h}\big(\frac{u'}{h}\big)'$ by the same steps. Its zero modes satisfy $u'/h = a$, i.e. $u(\theta) = a\,s(\theta) + b$ with $s(\theta) = \int_0^\theta h$. Periodicity, $u(2\pi) = u(0)$, forces $a\ell = 0$, so $a = 0$: only constants, $b_0 = 1$. (Here compactness entered: on the line the non-constant solution $a\,s$ would be allowed.)
The harmonic representative of a class. Given any 1-form $f\,\mathrm{d}\theta$ with period $P = \oint f\mathrm{d}\theta$, its harmonic representative is $\frac{P}{\ell}\,h\,\mathrm{d}\theta$: it is the unique harmonic form with the same period. The difference $\big(f - \frac{P}{\ell}h\big)\mathrm{d}\theta$ has zero period, hence is $\mathrm{d}g$ for a periodic g. Figure 2.25 plots this.

Worked example 2.42: The flat torus T^n: harmonic forms are constant forms

Let $T^n = \mathbb{R}^n/(2\pi\mathbb{Z})^n$ with the flat metric. In the flat coordinates the basis forms $\mathrm{d}x^I = \mathrm{d}x^{i_1}\wedge\cdots\wedge\mathrm{d}x^{i_p}$ are parallel, and the computation of Example 2.32 generalises verbatim:

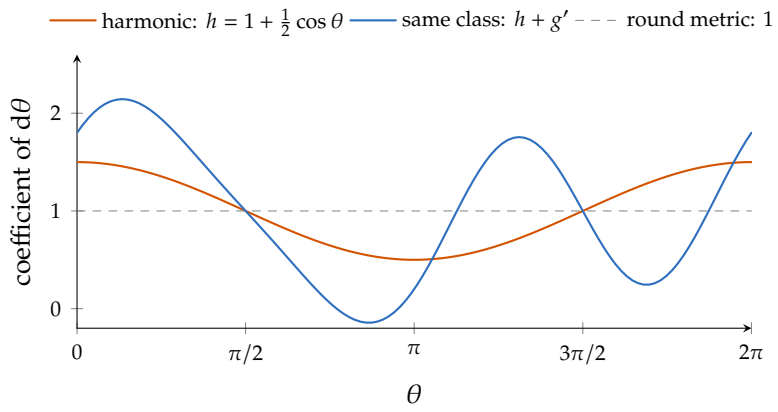


Figure 2.25: The lumpy circle of Example 2.41, with metric $h(\theta)^2 d\theta^2$ and $h = 1 + \frac{1}{2}\cos\theta$. The orange curve is the coefficient of its harmonic 1-form $h\,d\theta$ (the arc-length form); for the round metric it would be the flat dashed line. The blue curve is another closed form in the same class, $h\,d\theta + dg$ with $g = -0.4\cos 2\theta + 0.1\sin 3\theta$: it has the same area under it (the same period 2π), but is not co-closed. Changing the metric moves the harmonic representative, but there is always exactly one per class.

Δ_p acts on each component as $-\nabla^2$,

$$\Delta_p\Big(\sum_I \omega_I\,dx^I\Big) = \sum_I \big(-\nabla^2\omega_I\big)\,dx^I.$$

Expand each component in Fourier modes, $\omega_I = \sum_{m\in\mathbb{Z}^n} c_{I,m} e^{im\cdot x}$. Then

$$\Delta_p\omega = 0 \overset{(1)}{\iff} \sum_{I,m} |m|^2\, c_{I,m}\, e^{im\cdot x} dx^I = 0 \overset{(2)}{\iff} c_{I,m} = 0 \text{ for all } m \neq 0.$$

(1) $-\nabla^2 e^{im\cdot x} = |m|^2 e^{im\cdot x}$: here the symbol $|k|^2$ is literally the eigenvalue, with $k = m$ restricted to the lattice.

(2) The functions $e^{im\cdot x}dx^I$ are linearly independent; $|m|^2 \neq 0$ unless $m = 0$.

So the harmonic p-forms are exactly the forms with *constant* coefficients, and there are $\binom{n}{p}$ independent ones:

$$b_p(T^n) = \binom{n}{p}, \qquad \sum_p b_p\, t^p = (1+t)^n.$$

For T^2: $b = (1, 2, 1)$, with harmonic forms $1;\ dx, dy;\ dx\wedge dy$. The two harmonic 1-forms are dual to the two independent loops. (Compare the discrete model: the constant cochain is the harmonic one.) For a *non-rectangular* flat torus, or a curved metric on T^2, the harmonic forms change, but by (2.105) there are still $1, 2, 1$ of them.

Dictionary: Flat connections and Wilson lines on a torus

A harmonic 1-form on T^2, $A = a\,dx + b\,dy$ with constant a, b, is a U(1) gauge potential with $F = dA = 0$ and $d^*A = 0$: a *flat connection in Lorenz gauge*. It is not pure gauge: its Wilson loops around the two cycles, $W_x = \exp\big(ie\oint_x A\big) = e^{2\pi iea}$ and $W_y = e^{2\pi ieb}$ (physicists' Hermitian convention, charge e), are gauge-invariant and non-trivial. Gauge transformations $e^{ie\chi}$ with $\chi = nx/e$, single-valued as e^{inx}, shift $a \mapsto a + n/e$. So the

space of flat connections modulo gauge transformations is the torus $(a, b) \in (\mathbb{R}/\frac{1}{e}\mathbb{Z})^2$, of dimension $b_1(T^2) = 2$. In general the harmonic 1-forms are the tangent space to the moduli space of flat U(1) connections, which has dimension $b_1(M)$. These are the Aharonov–Bohm phases of the manifold. (Conventions for anti-Hermitian gauge fields are in Chapter T3: $A_{\text{math}} = -ieA_{\text{phys}}$.)

Worked example 2.43: The 2-sphere, and the Dirac monopole as a harmonic form

Take S^2 with the round metric $\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2$ and area form $\mathrm{vol} = \sin\theta\,\mathrm{d}\theta \wedge \mathrm{d}\phi = \star 1$.

- $p = 0$. If $\Delta_0 u = 0$ then $\|\mathrm{d}u\|^2 = 0$ by (2.102), so u is constant. $b_0 = 1$. (In terms of spherical harmonics: $\Delta_0 Y_{lm} = l(l+1)Y_{lm}$, and only $l = 0$ is a zero mode.)
- $p = 2$. By the duality below, $\star$ maps $\mathcal{H}^0$ to $\mathcal{H}^2$, so the harmonic 2-forms are $c\,\mathrm{vol}$. Check directly: $\mathrm{d}(c\,\mathrm{vol}) = 0$ (a 3-form on a surface) and $\mathrm{d}^*(c\,\mathrm{vol}) = \pm\star\,\mathrm{d}\star(c\,\mathrm{vol}) = \pm\star\,\mathrm{d}c = 0$. $b_2 = 1$.
- $p = 1$. A harmonic 1-form ω is closed. On S^2 every closed 1-form is exact (every loop can be shrunk, so the integral of ω from a base point is path-independent; see Chapter T2). So $\omega = \mathrm{d}u$ is harmonic and exact, and the "at most one" argument gives $\omega = 0$. $b_1 = 0$.

So $(b_0, b_1, b_2) = (1, 0, 1)$ and $\chi(S^2) = 1 - 0 + 1 = 2$.
Now the physics. The field of a magnetic monopole of strength g at the origin of $\mathbb{R}^3$ is $\vec{B} = \frac{g}{4\pi}\frac{\hat{r}}{r^2}$; restricted to the unit sphere its flux 2-form is

$$F \;=\; \frac{g}{4\pi}\sin\theta\,\mathrm{d}\theta\wedge\mathrm{d}\phi \;=\; \frac{g}{4\pi}\,\mathrm{vol}, \qquad \int_{S^2} F = g.$$

This is *the* harmonic 2-form with flux g. The source-free Maxwell equations on a closed surface, $\mathrm{d}F = 0$ and $\mathrm{d}^*F = 0$, are exactly harmonicity (2.103). So Hodge's theorem says: *on a compact space, source-free static field strengths are classified by their fluxes, one field for each value of the flux*. For the monopole, Dirac quantisation $eg \in 2\pi\mathbb{Z}$ makes the class $[\frac{e}{2\pi}F]$ integral: it is the first Chern class of the monopole bundle, with harmonic representative $\frac{eg}{8\pi^2}\mathrm{vol}$.

Two consequences: Poincaré duality and the Euler characteristic

The Hodge star preserves harmonicity. Let ω be a harmonic p-form. Using $\mathrm{d}^* = \pm\star\,\mathrm{d}\star$ and $\star\star = \pm 1$ (the signs do not matter for vanishing):

$$\mathrm{d}^*(\star\omega) \overset{(1)}{=} \pm\star\,\mathrm{d}\star\star\omega \overset{(2)}{=} \pm\star\,\mathrm{d}\omega \overset{(3)}{=} 0,$$
$$\mathrm{d}(\star\omega) \overset{(4)}{=} \pm\star\star\mathrm{d}\star\,\omega \overset{(5)}{=} \pm\star\;\mathrm{d}^*\omega \overset{(6)}{=} 0.$$

(1) (2.86) applied to the $(n-p)$-form $\star\omega$.
(2) $\star\star\,\omega = \pm\omega$ by (2.84).
(3) ω is closed.
(4) Insert $1 = \pm\star\star$ in front of the $(n-p+1)$-form $\mathrm{d}\star\omega$.
(5) $\star\mathrm{d}\star\,\omega = \pm\mathrm{d}^*\omega$ by (2.86).
(6) ω is co-closed.

So $\star : \mathcal{H}^p \to \mathcal{H}^{n-p}$, and it is invertible ($\star\star = \pm 1$). With (2.105):

$$b_p \;=\; b_{n-p} \qquad (\textbf{Poincaré duality}). \tag{2.110}$$

A purely topological statement, proved with a metric. The metric is a tool, and the answer does not depend on the tool. We saw it in every example: $(1,1)$ for S^1, $(1,2,1)$ for T^2, $(1,0,1)$ for S^2, and $\binom{n}{p} = \binom{n}{n-p}$ for T^n. It requires M to be *oriented* (so that $\star$ exists globally).

The Euler characteristic. The alternating sum of the Betti numbers is the Euler characteristic, and by (2.105)

$$\chi(M) \;=\; \sum_{p=0}^{n} (-1)^p b_p \;=\; \sum_{p=0}^{n} (-1)^p \dim \operatorname{Ker} \Delta_p \;=\; \dim \operatorname{Ker} \Delta_{\text{ev}} - \dim \operatorname{Ker} \Delta_{\text{odd}}. \tag{2.111}$$

The last expression is the index of the elliptic operator $\mathrm{d} + \mathrm{d}^* : \Omega^{\text{ev}} \to \Omega^{\text{odd}}$ of Example 2.37: its kernel is $\bigoplus_p \mathcal{H}^{2p}$, because $\|(\mathrm{d}+\mathrm{d}^*)\omega\|^2 = \|\mathrm{d}\omega\|^2 + \|\mathrm{d}^*\omega\|^2 + 2\langle \mathrm{d}\omega, \mathrm{d}^*\omega\rangle$ and the cross term is $2\langle \mathrm{dd}\omega, \omega\rangle = 0$, so $(\mathrm{d}+\mathrm{d}^*)\omega = 0$ forces $\mathrm{d}\omega = \mathrm{d}^*\omega = 0$, and the kernel of its adjoint is $\bigoplus_p \mathcal{H}^{2p+1}$. In odd dimensions Poincaré duality pairs b_p with b_{n-p} of opposite parity, so $\chi = 0$; for S^1 we found this by pairing eigenvalues in Figure 2.24.

Generalisations

Nash closes the section by noting that Hodge's theorem "generalises to include cases where the exterior derivative d is replaced by other differential operators". The recipe is the one of Definition 2.20: take an *elliptic complex* $\Gamma(E_0) \xrightarrow{D_0} \Gamma(E_1) \xrightarrow{D_1} \cdots$. Its cohomology $H^p = \operatorname{Ker} D_p / \operatorname{Im} D_{p-1}$ is isomorphic to the kernel of the Laplacian $\Delta_p = D_{p-1}D_{p-1}^* + D_p^* D_p$, by exactly the arguments above:

- the energy identity $\langle \omega, \Delta_p \omega\rangle = \left\| D_{p-1}^* \omega \right\|^2 + \left\| D_p \omega \right\|^2$ gives $\operatorname{Ker} \Delta_p = \operatorname{Ker} D_p \cap \operatorname{Ker} D_{p-1}^*$;
- ellipticity of Δ_p (equivalent to exactness of the symbol sequence, by the key idea of Section 2.3.11) gives a discrete spectrum, a finite-dimensional kernel and a Green's operator, hence a Hodge decomposition.

In particular the cohomology of an elliptic complex on a compact manifold is *finite-dimensional*. Examples we shall meet:

- the **Dolbeault complex** $\bar\partial : \Omega^{p,q} \to \Omega^{p,q+1}$ on a complex manifold, whose symbol is built from the Cauchy–Riemann symbol of Section 2.2.10; its harmonic forms compute the sheaf cohomology of Chapters 3 and 5;
- the **spin complex**, the Dirac operator $\not{D} : \Gamma(S^+) \to \Gamma(S^-)$ of Example 2.30, whose zero modes are the fermion zero modes of anomalies (Chapter 10);
- the **instanton deformation complex** of Example 2.38, whose first cohomology is the tangent space to the instanton moduli space (Chapter 8).

In all of them the index, the alternating sum of the dimensions of the cohomology groups, is topological. That is the subject of Chapter 4.

Reading Nash (§3, pp. 36–41, (2.38)–(2.53))

- pp. 36–37, (2.38)–(2.41) and footnote 4: operators $L : \Gamma(M, U) \to \Gamma(M, V)$ between bundles of ranks q, p; locally a $p \times q$ matrix of operators, with a matrix symbol and leading symbol $\sigma_m(x, k) : U_x \to V_x$ depending on a covector k. Our Section 2.3. Nash's footnote says the local expression depends on the chart; we worked out how it transforms (frame changes act by conjugation, like gauge transformations).
- pp. 37–38: the global picture with $\pi : T^*M \to M$, the pull-back bundles π^*U, π^*V, and ellipticity as "σ_m is an isomorphism $\pi^*U \to \pi^*V$ off the zero section" (which forces $p = q$). We added the examples of curl, div–grad–curl, Cauchy–Riemann as a real system and the Dirac operator, with the winding of the symbol.
- p. 39, (2.42)–(2.44): the Hodge Laplacian $\Delta_p = (\mathrm{d}\mathrm{d}^* + \mathrm{d}^*\mathrm{d})_p$ and the summary (2.43) of inner product, adjoint and Hodge star. Our Section 2.3.7, where we derived the sign of $\mathrm{d}^* = (-1)^{np+n+1} \star \mathrm{d} \star$ from Stokes' theorem. Nash's formula for $\star\omega$ contains a factor $\sqrt{g}/(n-p)!$ and uses the sum-over-all-indices convention on the right; see (2.83) for our reading.
- pp. 39–40, (2.45)–(2.49): composition and adjoint rules for symbols, which Nash calls "easily verifiable"; our Section 2.3.8, with the oscillatory-test proof of (2.45), the integration-by-parts proof of (2.47) and the Poisson-bracket remark. His $\sigma_1(\mathrm{d}) = k\wedge$ lacks our factor i (pitfall in Section 2.3.9); his conclusion $\sigma_2(\Delta_p) = k^2 I$ (2.49) is unaffected. His "elementary calculation" is our identity (2.96).
- p. 40: "generalisations of the Laplacian ... by tensoring with another vector bundle E, and by studying ... Yang–Mills theories"; our Section 2.3.10, with gauge fixing of Maxwell theory. Nash does not discuss elliptic complexes here (they come in his Chapter IV); we previewed them in Section 2.3.11 because they explain why Δ_p is elliptic.
- pp. 40–41, (2.50)–(2.53): harmonic forms, Hodge's theorem, $\dim \operatorname{Ker} \Delta_p = b_p$, metric independence, and the remark that the theorem extends to other operators. Our Section 2.3.12. Nash cites Wells for the proof; we gave the injectivity argument in full, the variational argument for existence, and the Green's-operator proof of the Hodge decomposition, isolating the three analytic inputs that ellipticity provides.

Recap: Ellipticity and vector bundles

- An operator between bundles $U \to V$ is locally a matrix of differential operators. Its leading symbol $\sigma_m(x, k)$ is a matrix $U_x \to V_x$, depending on a covector $k \in T^*_x M$, polynomial of degree m in k. Globally it is a bundle map $\pi^*U \to \pi^*V$ over T^*M.
- *Elliptic*: $\sigma_m(x, k)$ is invertible for every $k \neq 0$. So U and V have equal rank. The curl and the Maxwell operator fail (longitudinal/gauge modes); the div–grad–curl operator, the Dirac operator and Δ_p pass.
- Symbol calculus: $\sigma(PQ) = \sigma(P)\sigma(Q)$ and $\sigma(L^*) = \sigma(L)^*$ at leading order. Then $\sigma(\mathrm{d}) = ik\wedge$, $\sigma(\mathrm{d}^*) = -i\iota_k$, and $\sigma(\Delta_p) = k \wedge \iota_k + \iota_k k\wedge = |k|^2 I$.
- The symbol sequence of the de Rham complex is exact for $k \neq 0$; this is the same fact as ellipticity of all the Δ_p, and makes de Rham an *elliptic complex*. Gauge theories are elliptic "modulo gauge" in exactly this sense.

- Harmonic forms: $\Delta_p\omega = 0 \iff \mathrm{d}\omega = 0 = \mathrm{d}^*\omega$ (energy argument, needs compactness and Riemannian signature). Hodge: each class of $H^p_{\mathrm{dR}}(M)$ contains exactly one harmonic form, the minimum-energy representative; $\dim \operatorname{Ker} \Delta_p = b_p$ for every metric.
- Behind the scenes, ellipticity gives regularity, a discrete spectrum and a spectral gap, hence a Green's operator and the Hodge decomposition $\Omega^p = \mathcal{H}^p \oplus \mathrm{d}\Omega^{p-1} \oplus \mathrm{d}^*\Omega^{p+1}$.
- Examples: S^1 $(1,1)$, lumpy S^1 (harmonic form = arc-length form), T^n $(\binom{n}{p})$, constant forms), S^2 $(1,0,1$; the monopole field is the harmonic 2-form), and the triangle, where it is all 3×3 matrices. Consequences: Poincaré duality $b_p = b_{n-p}$ and $\chi(M) = \operatorname{index}(\mathrm{d}+\mathrm{d}^*)$, the first index theorem.

2.4 Pseudo-differential operators

Here is the question this section answers. Differential operators are the operators of classical physics: ∇^2, $\partial_t - \nabla^2$, $\gamma^\mu\partial_\mu$. But the moment we try to *solve* an equation $Lu = f$, or build a Hamiltonian like $\sqrt{p^2+m^2}$, or regularise a determinant with ζ-functions, we meet operators that are *not* differential operators:

- the Green's operator $(-\nabla^2 + m^2)^{-1}$, which is an integral against the Yukawa potential $e^{-mr}/4\pi r$;
- the relativistic kinetic energy $\sqrt{-\nabla^2+m^2}$;
- the heat operator $e^{t\nabla^2}$, which smears a function out with a Gaussian;
- complex powers $(-\nabla^2+m^2)^{-s}$, whose traces are ζ-functions.

All four have one thing in common. In Fourier space each of them is simply *multiplication by a function of momentum*: $1/(p^2+m^2)$, $\sqrt{p^2+m^2}$, e^{-tp^2}, $(p^2+m^2)^{-s}$. A differential operator is multiplication by a *polynomial* in momentum. A **pseudo-differential operator** is what you get when you allow the polynomial to be replaced by any function that grows or decays like a power of momentum. That single move gives a class of operators that contains every differential operator, is closed under taking inverses (up to harmless errors), square roots and complex powers, and in which ellipticity has exactly the same meaning as before. This is the class in which the index theorem of Chapter 4 is naturally stated and proved.

Before we make the leap, we tie off the last result of the previous section, because it is the reason we care about kernels of elliptic operators at all.

2.4.1 Where we stand: Hodge's theorem and the question of kernels

Recall the setting at the end of the previous section. M is a compact Riemannian manifold without boundary, of dimension n. On p-forms we have the exterior derivative d, its adjoint d^* with respect to the L^2 inner product

$$\langle \omega, \nu\rangle = \int_M \omega \wedge \star\nu,$$

and the Hodge Laplacian

$$\Delta_p = \mathrm{d}\mathrm{d}^* + \mathrm{d}^*\mathrm{d} : \ \Omega^p(M) \to \Omega^p(M). \tag{2.112}$$

(See Chapter T1 for $\star$ and $\mathrm{d}^* = (-1)^{n(p+1)+1}\star\mathrm{d}\star$ in Riemannian signature.) A form with $\Delta_p\omega = 0$ is called **harmonic**.

Harmonic means closed and co-closed. Take the inner product of $\Delta_p\omega$ with ω itself:

$$\begin{aligned}\langle\omega, \Delta_p\omega\rangle &\overset{(1)}{=} \langle\omega, \mathrm{d}\mathrm{d}^*\omega\rangle + \langle\omega, \mathrm{d}^*\mathrm{d}\omega\rangle \\ &\overset{(2)}{=} \langle\mathrm{d}^*\omega, \mathrm{d}^*\omega\rangle + \langle\mathrm{d}\omega, \mathrm{d}\omega\rangle \\ &\overset{(3)}{=} \|\mathrm{d}^*\omega\|^2 + \|\mathrm{d}\omega\|^2 .\end{aligned} \tag{2.113}$$

(1) Definition (2.112) and linearity of the inner product.
(2) Move one d (resp. d^*) across the inner product; this is the defining property $\langle\alpha, \mathrm{d}\beta\rangle = \langle\mathrm{d}^*\alpha, \beta\rangle$ of the adjoint. No boundary terms appear because $\partial M = \varnothing$ (Stokes).
(3) Each term is the squared L^2 norm, which is ≥ 0 because the metric is positive definite.

If $\Delta_p\omega = 0$ the left side vanishes, so a sum of two non-negative numbers is zero and each vanishes:

$$\omega \in \operatorname{Ker}\Delta_p \iff \mathrm{d}\omega = 0 \ \text{ and } \ \mathrm{d}^*\omega = 0. \tag{2.114}$$

(The arrow $\Leftarrow$ is obvious from (2.112).) In particular a harmonic form is closed, so it defines a de Rham class $[\omega] \in H^p_{\mathrm{dR}}(M) = \operatorname{Ker}\mathrm{d}/\operatorname{Im}\mathrm{d}$.

Key idea: Hodge's theorem

On a compact Riemannian manifold without boundary, each de Rham class contains *exactly one* harmonic form. Hence

$$\dim\operatorname{Ker}\Delta_p = \dim H^p_{\mathrm{dR}}(M) = b_p, \tag{2.115}$$

the p-th Betti number. The left side needs a metric; the right side does not. So the *number* of zero modes of the Laplacian is a topological invariant, although the zero modes themselves change when we change the metric.

Why at most one? If ω and ω' are harmonic and in the same class, then $\omega - \omega' = \mathrm{d}\beta$ is exact and harmonic, hence co-closed: $\mathrm{d}^*\mathrm{d}\beta = 0$. Then $\|\mathrm{d}\beta\|^2 = \langle\beta, \mathrm{d}^*\mathrm{d}\beta\rangle = 0$, so $\omega = \omega'$. *Why at least one?* Given a closed α, we want to subtract the exact piece $\mathrm{d}\beta$ that minimises $\|\alpha - \mathrm{d}\beta\|^2$; the minimiser is orthogonal to all exact forms, i.e. co-closed. That the minimum is actually *attained* is an analytic statement: it needs the machinery of this part (Sobolev spaces, compactness, the Fredholm property). This is one reason Nash now turns to analysis.

Worked example 2.44: Harmonic forms on the circle and on the flat torus

Circle. Let $M = S^1$ with coordinate $\theta \in [0, 2\pi)$ and metric $\mathrm{d}\theta^2$. On functions $\Delta_0 f = -f''$, and on 1-forms $\Delta_1(f\,\mathrm{d}\theta) = -f''\,\mathrm{d}\theta$. Expanding $f = \sum_{n\in\mathbb{Z}} c_n e^{in\theta}$ gives $\Delta_0 f = \sum_n n^2 c_n e^{in\theta}$, which vanishes iff only $c_0 \neq 0$. So $\operatorname{Ker}\Delta_0 = \{\text{constants}\}$ and $\operatorname{Ker}\Delta_1 = \{c\,\mathrm{d}\theta\}$: $b_0 = b_1 = 1$.
Torus. Let $M = T^2 = \mathbb{R}^2/2\pi\mathbb{Z}^2$ with the flat metric $\mathrm{d}x^2 + \mathrm{d}y^2$. In flat coordinates $\star$ has constant coefficients, so Δ_p acts on each component of a form as $-(\partial_x^2 + \partial_y^2)$. A general 1-form is $\omega = f\,\mathrm{d}x + g\,\mathrm{d}y$ and

$$\begin{aligned}\Delta_1\omega &\overset{(1)}{=} -(\partial_x^2 + \partial_y^2)f\,\mathrm{d}x - (\partial_x^2 + \partial_y^2)g\,\mathrm{d}y \\ &\overset{(2)}{=} \sum_{k\in\mathbb{Z}^2} |k|^2 \left(f_k\,\mathrm{d}x + g_k\,\mathrm{d}y\right)e^{ik\cdot x},\end{aligned}$$

(1) Flat metric: $\mathrm{d}\mathrm{d}^* + \mathrm{d}^*\mathrm{d} = -\nabla^2$ componentwise (the Weitzenböck curvature term vanishes).
(2) Expand $f = \sum_k f_k e^{ik\cdot x}$, $g = \sum_k g_k e^{ik\cdot x}$ with $k = (k_1, k_2) \in \mathbb{Z}^2$ and use $-\nabla^2 e^{ik\cdot x} = |k|^2 e^{ik\cdot x}$.

The kernel is the $k = 0$ part: $\omega = a\,\mathrm{d}x + b\,\mathrm{d}y$ with constants a, b. Doing the same on 0- and 2-forms,

$$\operatorname{Ker}\Delta_0 = \{c\}, \qquad \operatorname{Ker}\Delta_1 = \{a\,\mathrm{d}x + b\,\mathrm{d}y\}, \qquad \operatorname{Ker}\Delta_2 = \{c\,\mathrm{d}x\wedge\mathrm{d}y\},$$

so $(b_0, b_1, b_2) = (1, 2, 1)$. The two harmonic 1-forms are dual to the two independent loops of Figure 2.26: $\oint_a \mathrm{d}x = 2\pi$, $\oint_a \mathrm{d}y = 0$, and so on. If we change the metric to a sheared lattice, the harmonic forms change (they are the constant forms in the new flat coordinates) but there are still exactly two of them.

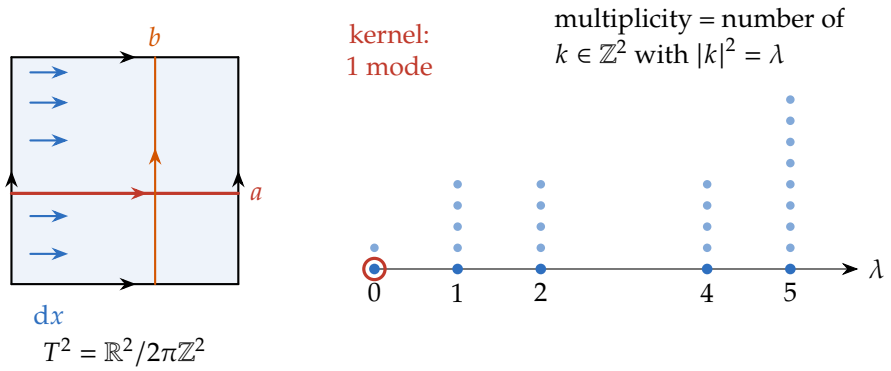


Figure 2.26: Left: the flat torus as a square with opposite sides glued. The constant form $\mathrm{d}x$ (blue arrows) integrates to 2π around the loop a and to 0 around b, so it detects a; $\mathrm{d}y$ detects b. These are the $b_1 = 2$ harmonic 1-forms. Right: the spectrum of Δ_0 on T^2. Eigenvalues are $|k|^2$, $k \in \mathbb{Z}^2$; the dots above each eigenvalue count its multiplicity $(1, 4, 4, 4, 8, \dots)$. The kernel (red circle) is one-dimensional: $b_0 = 1$. The spectrum is discrete and runs off to infinity; both features are consequences of ellipticity that we explain in Section 2.6.

Reading Nash (p. 41, (2.51)–(2.53))

Nash states Hodge's theorem without proof (referring to Wells) and remarks that "the whole question of kernels of elliptic operators is resumed in chapter 4". The finiteness of $\dim\operatorname{Ker}\Delta_p$ is not an accident of the Laplacian: every elliptic operator on a compact manifold has a finite-dimensional kernel. The purpose of §§4–5 is to set up the analysis (pseudo-differential operators, Sobolev spaces) that makes this statement precise. We come back to it in Section 2.6.

2.4.2 Differential operators are Fourier multipliers

We first fix conventions, because factors of 2π and signs in the exponent are the main source of confusion in this subject.

Dictionary: our Fourier conventions (used throughout this part)

On $\mathbb{R}^n$, with $x \cdot \xi = \sum_j x^j \xi_j$:

$$\hat{u}(\xi) = \int_{\mathbb{R}^n} e^{-ix\cdot\xi}\, u(x)\, \mathrm{d}^n x, \qquad u(x) = \int_{\mathbb{R}^n} e^{ix\cdot\xi}\, \hat{u}(\xi)\, \frac{\mathrm{d}^n \xi}{(2\pi)^n},$$
$$D_j = -i\,\frac{\partial}{\partial x^j}, \qquad D^\alpha = D_1^{\alpha_1}\cdots D_n^{\alpha_n},$$
$$\xi^\alpha = \xi_1^{\alpha_1}\cdots\xi_n^{\alpha_n}, \qquad \langle \xi \rangle = (1+|\xi|^2)^{1/2}.$$

Here $\alpha = (\alpha_1, \dots, \alpha_n)$ is a multi-index of non-negative integers with length $|\alpha| = \sum \alpha_j$. Physically, ξ is momentum in units $\hbar = 1$ and $D_j = -i\partial_j$ is the momentum operator $\hat{p}_j$. The **Laplacian** Δ always means the *positive* (geometer's) Laplacian:

$$\Delta = -\sum_j \frac{\partial^2}{\partial (x^j)^2} = -\nabla^2 = \sum_j D_j^2,$$

so that its eigenvalues are ≥ 0. A physicist's $-\nabla^2 + m^2$ is our $\Delta + m^2$.

The whole point of the Fourier transform is that a plane wave is an eigenfunction of every derivative:

$$D_j\, e^{ix\cdot\xi} = -i \cdot i\xi_j\, e^{ix\cdot\xi} = \xi_j\, e^{ix\cdot\xi}, \qquad D^\alpha e^{ix\cdot\xi} = \xi^\alpha e^{ix\cdot\xi}. \tag{2.116}$$

Now let $L = \sum_{|\alpha|\leq m} a_\alpha(x) D^\alpha$ be a differential operator of order m with smooth coefficients. Apply it to u written as a superposition of plane waves:

$$\begin{aligned} Lu(x) &\overset{(1)}{=} \sum_{|\alpha|\leq m} a_\alpha(x)\, D^\alpha \int e^{ix\cdot\xi} \hat{u}(\xi)\, \frac{\mathrm{d}^n\xi}{(2\pi)^n} \\ &\overset{(2)}{=} \sum_{|\alpha|\leq m} a_\alpha(x) \int \xi^\alpha e^{ix\cdot\xi} \hat{u}(\xi)\, \frac{\mathrm{d}^n\xi}{(2\pi)^n} \\ &\overset{(3)}{=} \int e^{ix\cdot\xi}\, \sigma(x,\xi)\, \hat{u}(\xi)\, \frac{\mathrm{d}^n\xi}{(2\pi)^n}, \qquad \sigma(x,\xi) \coloneqq \sum_{|\alpha|\leq m} a_\alpha(x)\xi^\alpha. \end{aligned} \tag{2.117}$$

(1) Fourier inversion for u.
(2) Differentiate under the integral (fine for u in the Schwartz space, defined in Section 2.5.5) and use (2.116).
(3) Pull the finite sum inside the integral.

The polynomial $\sigma(x,\xi)$ is the **symbol** of L and its top-degree part $\sigma_m(x,\xi) = \sum_{|\alpha|=m} a_\alpha(x)\xi^\alpha$ is the **principal (leading) symbol**. Operationally: *replace every D_j by ξ_j, keeping the coefficients to the left.*

Worked example 2.45: Symbols of three familiar operators

(a) The Laplacian with mass. $\Delta + m^2 = \sum_j D_j^2 + m^2$, so $\sigma(\xi) = |\xi|^2 + m^2$ and $\sigma_2(\xi) = |\xi|^2$. This is the Euclidean propagator denominator $p^2 + m^2$.

(b) A variable-coefficient ODE. On $\mathbb{R}$ let $L = a(x)\dfrac{\mathrm{d}^2}{\mathrm{d}x^2} + b(x)\dfrac{\mathrm{d}}{\mathrm{d}x} + c(x)$. Write it in terms of $D = -i\,\mathrm{d}/\mathrm{d}x$:

$$L \overset{(1)}{=} a(x)(iD)^2 + b(x)(iD) + c(x)$$

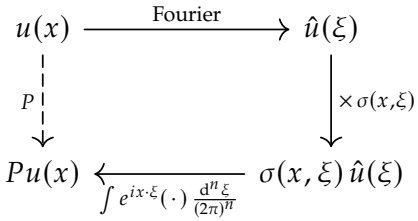


Figure 2.27: The Fourier-multiplier picture of an operator. Go to momentum space, multiply by the symbol, come back. For a differential operator the symbol is a polynomial in ξ; a pseudo-differential operator is anything that fits this diagram with a symbol that behaves like a power of $|\xi|$ at large momentum. When σ depends on x the right-hand arrow is not a "multiplication in momentum space" in the strict sense, but the formula still makes sense: at each x we weight each plane wave by $\sigma(x,\xi)$.

$$\overset{(2)}{=} -a(x)D^2 + i\,b(x)D + c(x),$$

(1) $\mathrm{d}/\mathrm{d}x = iD$ because $D = -i\,\mathrm{d}/\mathrm{d}x$.
(2) $i^2 = -1$.

so $\sigma(x,\xi) = -a(x)\xi^2 + ib(x)\xi + c(x)$ and $\sigma_2(x,\xi) = -a(x)\xi^2$. The operator is elliptic exactly where $a(x) \neq 0$. Bessel's operator $x^2\partial_x^2 + x\partial_x + (x^2 - \lambda^2)$ has $\sigma_2 = -x^2\xi^2$, which dies at $x = 0$: non-elliptic there, as found in the previous section.

(c) A matrix symbol: the 2d Dirac operator. On $\mathbb{R}^2$ take the Pauli matrices $\sigma^1 = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, $\sigma^2 = \left(\begin{smallmatrix} 0 & -i \\ i & 0 \end{smallmatrix}\right)$ as Euclidean gamma matrices and let $\not{D} = \sigma^1 D_1 + \sigma^2 D_2$ act on $u = (u_1, u_2)^T$. Then

$$\sigma_{\not{D}}(\xi) = \sigma^1\xi_1 + \sigma^2\xi_2 = \begin{pmatrix} 0 & \xi_1 - i\xi_2 \\ \xi_1 + i\xi_2 & 0 \end{pmatrix}, \qquad \sigma_{\not{D}}(\xi)^2 = |\xi|^2\,\mathbb{1}.$$

The square identity is the Clifford relation $\sigma^j\sigma^k + \sigma^k\sigma^j = 2\delta^{jk}$ (Chapter T4); it shows $\sigma_{\not{D}}(\xi)$ is invertible for $\xi \neq 0$, with inverse $\sigma_{\not{D}}(\xi)/|\xi|^2$. So $\not{D}$ is elliptic, and $\not{D}^2 = \Delta\,\mathbb{1}$.

Watch out: Nash's sign and 2π conventions

Nash writes $u(x) = \int \hat{u}(p)\,e^{-ipx}\,\mathrm{d}p$ (so his $\hat{u}$ absorbs the $(2\pi)^{-n}$) and $D^\alpha = (-i)^{|\alpha|}\partial^\alpha$. With these two choices $D^\alpha e^{-ipx} = (-p)^\alpha e^{-ipx}$, so his claim that the symbol is $\sum a_\alpha p^\alpha$ is off by a sign for odd $|\alpha|$. Nothing important depends on it: his p is our $-\xi$, and every symbol Nash actually uses in §4 is even in p. We use the conventions in the box above and never rely on Nash's signs. Similarly, overall constants in Nash's kernels (factors of 2π or $\sqrt{2\pi}$) do not affect orders or ellipticity; we point out the discrepancies as we meet them.

2.4.3 The leap: allow any symbol

Formula (2.117) makes sense for *any* reasonable function $\sigma(x,\xi)$, not just polynomials in ξ. That is the definition Nash starts from.

Working definition 2.22 (operator with symbol σ)

Given a function $\sigma(x,\xi)$ on $\mathbb{R}^n \times \mathbb{R}^n$, the operator $P = \mathrm{Op}(\sigma)$ **with symbol** σ acts by

$$Pu(x) = \int_{\mathbb{R}^n} e^{ix\cdot\xi}\,\sigma(x,\xi)\,\hat{u}(\xi)\,\frac{\mathrm{d}^n\xi}{(2\pi)^n}. \tag{2.118}$$

To compute Pu: Fourier transform u, multiply by $\sigma(x,\xi)$, transform back (Figure 2.27). *To find the symbol of a given P:* apply P to a plane wave; by (2.118) with $\hat{u}(\eta) = (2\pi)^n \delta(\eta - \xi)$,

$$\sigma(x,\xi) = e^{-ix\cdot\xi}\, P\big(e^{ix\cdot\xi}\big). \tag{2.119}$$

P is a **pseudo-differential operator** (ΨDO) of order m when σ lies in the symbol class S^m [symbols that grow at most like $|\xi|^m$, each ξ-derivative lowering the growth by one power] defined in Section 2.4.4.

If σ does not depend on x, then (2.118) says $\widehat{Pu}(\xi) = \sigma(\xi)\hat{u}(\xi)$: P is a **Fourier multiplier**, and we write $P = \sigma(D)$. For example $\sigma(\xi) = |\xi|^2 + m^2$ gives $P = \Delta + m^2$, and $\sigma(\xi) = \sqrt{|\xi|^2 + m^2}$ *defines* $\sqrt{\Delta + m^2}$. Functions of the momentum operator are Fourier multipliers, exactly as in quantum mechanics.

Kernels. Every such P is also an integral operator. Substitute $\hat{u}(\xi) = \int e^{-iy\cdot\xi} u(y)\, \mathrm{d}^n y$ into (2.118):

$$\begin{aligned} Pu(x) &\overset{(1)}{=} \int \frac{\mathrm{d}^n\xi}{(2\pi)^n}\, e^{ix\cdot\xi}\sigma(x,\xi) \int \mathrm{d}^n y\, e^{-iy\cdot\xi} u(y) \\ &\overset{(2)}{=} \int \mathrm{d}^n y \left[\int \frac{\mathrm{d}^n\xi}{(2\pi)^n}\, e^{i(x-y)\cdot\xi}\sigma(x,\xi)\right] u(y) \\ &\overset{(3)}{=} \int K(x,y)\, u(y)\, \mathrm{d}^n y, \qquad K(x,y) = \int e^{i(x-y)\cdot\xi}\, \sigma(x,\xi)\, \frac{\mathrm{d}^n\xi}{(2\pi)^n}. \end{aligned} \tag{2.120}$$

(1) Definition of $\hat{u}$.
(2) Exchange the order of integration. For a polynomial or growing σ this exchange is formal, and the ξ-integral is a distribution rather than a function; for decaying σ it is literal.
(3) Name the bracket: the **Schwartz kernel** of P.

Inverting the Fourier transform in $z = x - y$ recovers the symbol from the kernel:

$$\sigma(x,\xi) = \int K(x, x-z)\, e^{-iz\cdot\xi}\, \mathrm{d}^n z. \tag{2.121}$$

For a Fourier multiplier $K(x,y) = k(x-y)$ is a convolution kernel and $\sigma = \hat{k}$. This is Nash's pair (2.57)–(2.58): solving an integral equation $\int K(x,y)u(y)\,\mathrm{d}y = f(x)$ is the same as inverting the pseudo-differential operator with symbol (2.121).

Dictionary: symbol and kernel

Operator side	Kernel side
symbol polynomial in ξ (differential op.)	K = derivatives of $\delta(x-y)$, supported on $x = y$
symbol decays like $\lvert\xi\rvert^m$, $m < 0$	K singular only at $x = y$, like $\lvert x-y\rvert^{-n-m}$
symbol decays faster than any power	K smooth everywhere (smoothing operator)
symbol smooth in ξ	K smooth (even decaying) away from $x = y$

The last line is the heart of *pseudo-locality* (Section 2.4.15): smoothness in momentum is decay in position.

The next three examples are the operators from the opening of the section. Each has an explicit kernel, and each kernel tells a physical story.

Worked example 2.46: The Yukawa potential is the kernel of $(\Delta + m^2)^{-1}$ in $\mathbb{R}^3$

The symbol is $\sigma(\xi) = 1/(|\xi|^2 + m^2)$, a Fourier multiplier. By (2.120) its kernel is $k(x-y)$ with, writing $r = |x|$,

$$k(x) \overset{(1)}{=} \int_0^\infty \int_0^\pi \int_0^{2\pi} \frac{e^{i\rho r\cos\vartheta}}{\rho^2+m^2}\,\frac{\rho^2\sin\vartheta\,\mathrm{d}\varphi\,\mathrm{d}\vartheta\,\mathrm{d}\rho}{(2\pi)^3}$$
$$\overset{(2)}{=} \frac{1}{4\pi^2}\int_0^\infty \frac{\rho^2}{\rho^2+m^2}\left[\frac{e^{i\rho r t}}{i\rho r}\right]_{t=-1}^{t=1}\mathrm{d}\rho = \frac{1}{2\pi^2 r}\int_0^\infty \frac{\rho\sin\rho r}{\rho^2+m^2}\,\mathrm{d}\rho$$
$$\overset{(3)}{=} \frac{1}{2\pi^2 r}\cdot\frac{\pi}{2}e^{-mr} = \frac{e^{-mr}}{4\pi r}.$$

(1) Spherical coordinates in ξ-space with the polar axis along x, so $x\cdot\xi = \rho r\cos\vartheta$ where $\rho = |\xi|$.

(2) Do the φ integral (2π) and substitute $t = \cos\vartheta$, $\mathrm{d}t = -\sin\vartheta\,\mathrm{d}\vartheta$; then $(e^{i\rho r} - e^{-i\rho r})/(i\rho r) = 2\sin(\rho r)/(\rho r)$.

(3) The standard integral $\int_0^\infty \rho\sin(\rho r)/(\rho^2+m^2)\,\mathrm{d}\rho = \frac{\pi}{2}e^{-mr}$ for $r>0$ (close the contour of $\int_{-\infty}^{\infty}\rho e^{i\rho r}/(\rho^2+m^2)$ in the upper half plane; the residue at $\rho = im$ is $e^{-mr}/2$).

So $(\Delta+m^2)^{-1}u(x) = \int \frac{e^{-m|x-y|}}{4\pi|x-y|}u(y)\,\mathrm{d}^3y$: the Yukawa potential of a source u. At $m=0$ it is the Coulomb potential. The kernel is singular only at $x=y$, where it behaves like $|x-y|^{-1} = |x-y|^{-n-m}$ with $n=3$ and order $m=-2$, as the dictionary predicts. And the operator is *not local*: a source confined to a small ball produces a potential everywhere, decaying on the length scale $1/m$ (the Compton wavelength of the exchanged particle).

Worked example 2.47: The one-dimensional resolvent, checked directly

In one dimension the same computation gives

$$(D^2+m^2)^{-1}: \qquad k(x) = \int_{-\infty}^{\infty}\frac{e^{ix\xi}}{\xi^2+m^2}\frac{\mathrm{d}\xi}{2\pi} = \frac{e^{-m|x|}}{2m}.$$

Let us *check* that this kernel inverts $D^2+m^2 = -\mathrm{d}^2/\mathrm{d}x^2 + m^2$, i.e. $(-\partial_x^2+m^2)k = \delta$. Away from 0, $k'' = m^2k$, so the left side vanishes. At 0, k' jumps from $+\frac12$ to $-\frac12$:

$$\int_{-\varepsilon}^{\varepsilon}(-k''+m^2k)\mathrm{d}x \overset{(1)}{=} -\big[k'(\varepsilon)-k'(-\varepsilon)\big] + O(\varepsilon)$$
$$\overset{(2)}{=} -\Big[-\tfrac12 e^{-m\varepsilon} - \tfrac12 e^{-m\varepsilon}\Big] + O(\varepsilon) \xrightarrow{\varepsilon\to0} 1.$$

(1) Fundamental theorem of calculus for k''; k is bounded, so $\int m^2k = O(\varepsilon)$.

(2) $k'(x) = -\frac12\operatorname{sgn}(x)e^{-m|x|}$.

A unit-weight spike at 0 and zero elsewhere: that is $\delta(x)$.

Worked example 2.48: The kernel of the relativistic kinetic energy $\sqrt{D^2+m^2}$

In one dimension, $H = \sqrt{D^2+m^2}$ is the Hamiltonian of a free spinless relativistic particle in the "square-root Klein–Gordon" formulation. Its symbol $\sqrt{\xi^2+m^2}$ is not a polynomial, so H is not a differential operator. What does it do in position space? Start from the inverse,

$\sigma = (\xi^2 + m^2)^{-1/2}$, whose kernel is a classical integral:

$$k_{-1/2}(x) \overset{(1)}{=} \int_{-\infty}^{\infty} \frac{e^{ix\xi}}{\sqrt{\xi^2+m^2}} \frac{\mathrm{d}\xi}{2\pi} = \frac{1}{\pi}\int_0^{\infty} \frac{\cos(x\xi)}{\sqrt{\xi^2+m^2}}\,\mathrm{d}\xi$$
$$\overset{(2)}{=} \frac{1}{\pi} K_0(m\,|x|),$$

(1) The integrand's odd part (sin) integrates to zero.
(2) Substitute $\xi = mt$ and use the integral representation $K_0(z) = \int_0^{\infty} \cos(zt)(1+t^2)^{-1/2}\mathrm{d}t$ of the modified Bessel function K_0.

Now $H = (D^2 + m^2) \circ (D^2 + m^2)^{-1/2}$, because Fourier multipliers compose by multiplying symbols. So away from $x = 0$, writing $z = m\,|x|$,

$$k_{1/2}(x) \overset{(1)}{=} \Big(-\frac{\mathrm{d}^2}{\mathrm{d}x^2} + m^2\Big)\frac{K_0(m\,|x|)}{\pi} = \frac{m^2}{\pi}\big(-K_0''(z) + K_0(z)\big)$$
$$\overset{(2)}{=} \frac{m^2}{\pi}\frac{K_0'(z)}{z} \overset{(3)}{=} -\frac{m}{\pi}\frac{K_1(m\,|x|)}{|x|} \qquad (x \neq 0).$$

(1) Apply $D^2 + m^2$ to the kernel of $(D^2 + m^2)^{-1/2}$; chain rule with $\mathrm{d}z/\mathrm{d}x = \pm m$.
(2) Modified Bessel equation $K_0'' + K_0'/z - K_0 = 0$.
(3) $K_0' = -K_1$.

Two lessons. First, H is *non-local*: $Hu(x)$ depends on u at all points, with a weight $\sim e^{-m|x-y|}$ at distances beyond the Compton wavelength $1/m$ (since $K_1(z) \sim \sqrt{\pi/2z}\,e^{-z}$). This is the classic obstacle to a local one-particle relativistic quantum mechanics. Second, as $m \to 0$, $K_1(z) \approx 1/z$ gives $k_{1/2} \to -1/(\pi x^2)$: the kernel of $|D|$, singular like $|x-y|^{-n-m}$ with $n = 1$, order $m = 1$ (interpreted as a finite-part distribution at $x = y$).

Worked example 2.49: The heat operator: a smoothing operator

For $t > 0$ let $\sigma_t(\xi) = e^{-t|\xi|^2}$, so $P_t = e^{-t\Delta}$, the operator that evolves an initial temperature by the heat equation $\partial_t u = -\Delta u = \nabla^2 u$. Its kernel is a Gaussian:

$$k_t(x) \overset{(1)}{=} \prod_{j=1}^{n} \int_{-\infty}^{\infty} e^{ix_j\xi_j - t\xi_j^2}\frac{\mathrm{d}\xi_j}{2\pi}$$
$$\overset{(2)}{=} \prod_{j=1}^{n} \frac{1}{2\pi}\sqrt{\frac{\pi}{t}}\,e^{-x_j^2/4t} = (4\pi t)^{-n/2} e^{-|x|^2/4t}.$$

(1) The integrand factorises over coordinates.
(2) Complete the square, $-t\xi^2 + ix\xi = -t(\xi - \frac{ix}{2t})^2 - \frac{x^2}{4t}$, and use $\int e^{-t\eta^2}\mathrm{d}\eta = \sqrt{\pi/t}$.

In one dimension with $t = \frac{1}{2}$ this is Nash's example (2.64): kernel $(2\pi)^{-1/2}e^{-(x-y)^2/2}$, symbol $e^{-\xi^2/2}$. The symbol decays faster than any power of ξ, and correspondingly $P_t u$ is C^∞ for any integrable u (differentiate under the integral: every derivative lands on the smooth Gaussian). Even $u = \delta$ becomes the smooth function k_t. As $t \to 0$ the kernel sharpens into δ and $P_t \to \mathrm{id}$ (Figure 2.28).

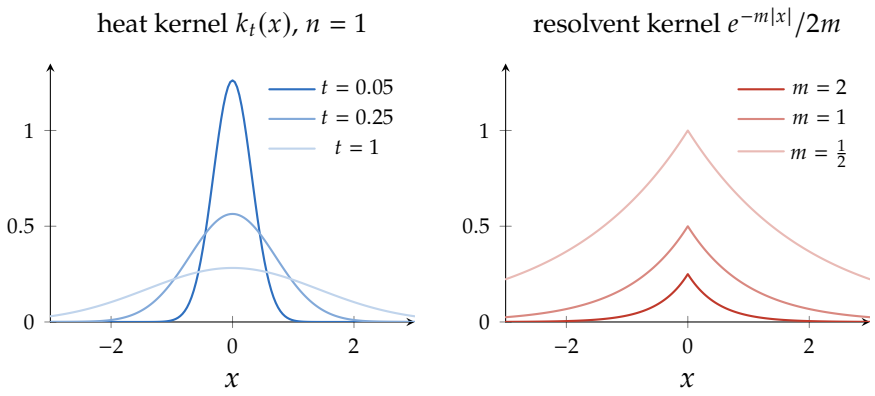


Figure 2.28: Kernels of two pseudo-differential operators on $\mathbb{R}$. Left: the heat kernel is smooth everywhere; its symbol $e^{-t\xi^2}$ decays faster than any power, so $e^{-t\Delta}$ has order $-\infty$ (smoothing). As $t \to 0$ it concentrates into $\delta(x)$. Right: the kernel of $(D^2 + m^2)^{-1}$ (order -2) is smooth except for a kink at $x = 0$ (the jump in k' that produces the δ). Neither kernel is supported at $x = 0$ alone, so neither operator is local; both are smooth off the diagonal, which is pseudo-locality.

Reading Nash (pp. 41, 44, (2.54), (2.57)–(2.58))

Nash's (2.54) is our (2.118) in his conventions (e^{-ipx} and $(2\pi)^{-n}$ absorbed into $\hat{u}$). His (2.58) gives the symbol in terms of the kernel with a prefactor $(2\pi)^{-n}$; in the conventions of his own (2.54) the correct formula has no such prefactor, as in our (2.121). You can see the slip in his Gaussian example: the kernel $e^{-(x-y)^2/2}$ has symbol $\sqrt{2\pi}\, e^{-p^2/2}$, not $e^{-p^2/2}/\sqrt{2\pi}$. The constant is irrelevant for everything he does with it (the order is $-\infty$ either way), but it is worth knowing if you check his formulas.

2.4.4 Symbol classes: what "behaves like a polynomial of degree m" means

We cannot allow *any* function $\sigma(x, \xi)$; the operator would have no useful properties. The trick is to ask what is special about a polynomial of degree m in ξ, and to keep exactly that. Look at $\sigma(\xi) = \xi^3$ in one variable:

$$\sigma = \xi^3, \qquad \partial_\xi \sigma = 3\xi^2, \qquad \partial_\xi^2 \sigma = 6\xi, \qquad \partial_\xi^3 \sigma = 6, \qquad \partial_\xi^4 \sigma = 0.$$

Each ξ-derivative lowers the growth rate at large $|\xi|$ by one power. Derivatives in x act only on the coefficients $a_\alpha(x)$ and do not change the growth rate at all. That is the whole definition.

Working definition 2.23 (symbol class S^m)

Let $m \in \mathbb{R}$ and $\Omega \subset \mathbb{R}^n$ open. A smooth function $\sigma(x, \xi)$ on $\Omega \times \mathbb{R}^n$ belongs to $S^m(\Omega)$ if for every compact [closed and bounded, staying away from the edge of Ω] $K \subset \Omega$ and all multi-indices [α_j counts the derivatives in ξ_j, and β_j those in x_j] α, β there is a constant $C_{\alpha\beta}(K)$ with

$$\left|\partial_x^\beta \partial_\xi^\alpha \sigma(x, \xi)\right| \le C_{\alpha\beta}(K)\, \langle \xi \rangle^{m-|\alpha|} \qquad (x \in K,\ \xi \in \mathbb{R}^n), \tag{2.122}$$

where $\langle \xi \rangle = (1 + |\xi|^2)^{1/2}$ [like $|\xi|$ at large ξ, but never zero]. *Test:* (i) bound $|\sigma|$ by a constant times $\langle \xi \rangle^m$; (ii) check that every ξ-derivative buys one extra power of decay; (iii) check that x-derivatives cost nothing. The number m is an **order** of σ and of $P = \mathrm{Op}(\sigma)$.

Nash writes $(1 + |p|)^{m-|\beta|}$ instead of $\langle \xi \rangle^{m-|\alpha|}$ (and his α, β are swapped relative to ours). The

two weights are interchangeable, because for every ξ

$$\langle\xi\rangle \;\le\; 1+|\xi| \;\le\; \sqrt{2}\,\langle\xi\rangle : \tag{2.123}$$

square both sides; the left inequality is $1+|\xi|^2 \le 1+2|\xi|+|\xi|^2$, and the right one is $1+2|\xi|+|\xi|^2 \le 2+2|\xi|^2$, i.e. $(1-|\xi|)^2 \ge 0$. We prefer $\langle\xi\rangle$ because it is smooth.

Three features of the definition deserve comment.

- *Any real m is allowed:* fractional, negative, and (by intersection) $-\infty$. Operators like $\sqrt{\Delta}$ (order 1) or $\Delta^{-1/3}$ (order $-2/3$) are now on the same footing as ∇^2.
- *The classes are nested:* if $m \le m'$ then $\langle\xi\rangle^{m-|\alpha|} \le \langle\xi\rangle^{m'-|\alpha|}$, so $S^m \subset S^{m'}$. An operator of order m is also of every higher order. We write
$$S^{-\infty} = \bigcap_{m\in\mathbb{R}} S^m, \qquad S^{\infty} = \bigcup_{m\in\mathbb{R}} S^m.$$
- *"The" order is not defined as the smallest m.* One could try $\inf\{m : \sigma \in S^m\}$, but the infimum need not be attained: there are symbols in $S^{m+\varepsilon}$ for every $\varepsilon > 0$ but not in S^m (Example 2.63). So one always says "of order m", meaning "in S^m".

Worked example 2.50: Polynomials, Japanese brackets, and a resolvent

(a) A polynomial of degree m is in S^m. Take $\sigma = a(x)\xi^2$ in one variable, with a smooth. On a compact K, $\left|\partial_x^\beta a\right| \le A_\beta$. Then

$$\left|\partial_x^\beta \sigma\right| = \left|\partial_x^\beta a\right| \xi^2 \overset{(1)}{\le} A_\beta\langle\xi\rangle^2, \qquad \left|\partial_x^\beta \partial_\xi \sigma\right| = 2\left|\partial_x^\beta a\right| |\xi| \overset{(2)}{\le} 2A_\beta\langle\xi\rangle^1,$$
$$\left|\partial_x^\beta \partial_\xi^2 \sigma\right| = 2\left|\partial_x^\beta a\right| \le 2A_\beta\langle\xi\rangle^0, \qquad \partial_\xi^k \sigma = 0 \;\; (k \ge 3).$$

(1) $\xi^2 \le 1+\xi^2 = \langle\xi\rangle^2$.
(2) $|\xi| \le \langle\xi\rangle$.

For $k \ge 3$ the bound $0 \le C\langle\xi\rangle^{2-k}$ is trivially true.
(b) $\langle\xi\rangle^s \in S^s$ for every real s. In one variable,

$$\partial_\xi\langle\xi\rangle^s \overset{(1)}{=} \tfrac{s}{2}(1+\xi^2)^{\frac{s}{2}-1}\cdot 2\xi = s\,\xi\,\langle\xi\rangle^{s-2},$$
$$\partial_\xi^2\langle\xi\rangle^s \overset{(2)}{=} s\langle\xi\rangle^{s-2} + s(s-2)\,\xi^2\langle\xi\rangle^{s-4},$$

(1) Chain rule on $(1+\xi^2)^{s/2}$.
(2) Product rule on $\xi\cdot\langle\xi\rangle^{s-2}$, using the first line with $s \to s-2$.

Using $|\xi| \le \langle\xi\rangle$: $|\partial_\xi\langle\xi\rangle^s| \le |s|\,\langle\xi\rangle^{s-1}$ and $\left|\partial_\xi^2\langle\xi\rangle^s\right| \le (|s|+|s(s-2)|)\langle\xi\rangle^{s-2}$. Inductively, $\partial_\xi^k\langle\xi\rangle^s$ is a finite sum of terms $\xi^j\langle\xi\rangle^{s-k-j}$, each bounded by $\langle\xi\rangle^{s-k}$. So $\langle\xi\rangle^s \in S^s$. The operator $\langle D\rangle^s = (1+\Delta)^{s/2}$, the **Bessel potential** of order s, will define the Sobolev spaces in Section 2.5.
(c) The resolvent symbol is in S^{-2}. $\sigma = (|\xi|^2+m^2)^{-1}$ with $m > 0$. The same computation as (b) works, because $|\xi|^2+m^2$ is comparable to $\langle\xi\rangle^2$: $\min(1,m^2)\langle\xi\rangle^2 \le |\xi|^2+m^2 \le \max(1,m^2)\langle\xi\rangle^2$. For instance

$$\left|\partial_{\xi_j}\sigma\right| = \frac{2\left|\xi_j\right|}{(|\xi|^2+m^2)^2} \le \frac{2\langle\xi\rangle}{\min(1,m^2)^2\langle\xi\rangle^4} = C\langle\xi\rangle^{-3}.$$

Worked example 2.51: Why the derivative condition matters: translation is not a ΨDO

Let $\sigma(\xi) = e^{ia\xi}$ in one variable ($a \neq 0$ real). It is smooth and bounded, $|\sigma| = 1 \leq \langle\xi\rangle^0$, so it passes test (i) for $m = 0$. But $\partial_\xi \sigma = ia\, e^{ia\xi}$ has modulus $|a|$, which is *not* $\leq C\langle\xi\rangle^{-1}$ for large ξ. So $\sigma \notin S^0$. In fact $\sigma \notin S^m$ for any finite m: the k-th derivative has constant modulus $|a|^k$, and $|a|^k \leq C\langle\xi\rangle^{m-k}$ for all large ξ would force $m \geq k$ for every k. What operator is it?

$$\sigma(D)u(x) \overset{(1)}{=} \int e^{ix\xi} e^{ia\xi} \hat{u}(\xi) \frac{\mathrm{d}\xi}{2\pi} \overset{(2)}{=} u(x+a).$$

(1) Definition (2.118).

(2) Fourier inversion at the point $x + a$.

It is the translation $u \mapsto u(\cdot + a)$, i.e. $e^{iaD} = e^{a\,\mathrm{d}/\mathrm{d}x}$, the familiar $e^{ia\hat{p}}$ of quantum mechanics. Translation *moves* singularities (a kink at 0 goes to $-a$), whereas we will see that genuine ΨDOs never create singularities in new places. The condition "each ξ-derivative gains decay" is exactly what forbids this.

Watch out: Symbols that are singular at $\xi = 0$

$|\xi|$, $|\xi|^{2\lambda}$ and $\xi/|\xi|$ are not smooth at $\xi = 0$, so strictly they are not in any S^m. This is Nash's reason for excluding $a = 0$ in his example (2.59). The standard cure: only large $|\xi|$ matters. Pick a smooth cutoff $\chi(\xi)$ with $\chi = 0$ for $|\xi| \leq \frac{1}{2}$ and $\chi = 1$ for $|\xi| \geq 1$, and use $\chi(\xi)\,|\xi|$ instead of $|\xi|$. The difference $(1 - \chi)\,|\xi|$ is bounded with compact support in ξ, which is in $S^{-\infty}$ (all derivatives are bounded and vanish for $|\xi| \geq 1$). Operators with symbols in $S^{-\infty}$ are *smoothing* and are "negligible" in the theory. So we can say: $|D|$ is a ΨDO of order 1 *modulo smoothing operators*.

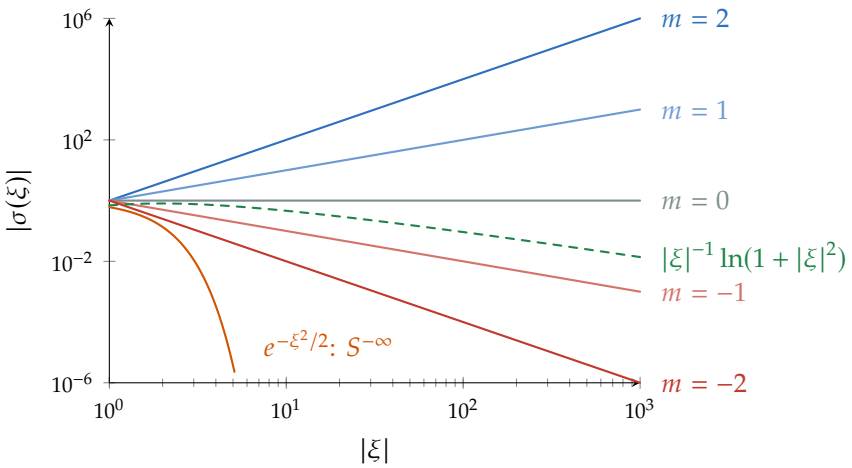


Figure 2.29: Growth of symbols at large momentum on a log–log plot. A symbol of order m is bounded by a line of slope m. The Gaussian (orange) falls below every line: it is in $S^{-\infty}$, and the operator is smoothing. The dashed green symbol $|\xi|^{-1}\ln(1 + |\xi|^2)$ lies above the slope -1 line (the log grows) but below the line of slope $-1 + \varepsilon$ for any $\varepsilon > 0$ eventually: it is in $S^{-1+\varepsilon}$ for all $\varepsilon > 0$ but not in S^{-1}.

A larger family of symbol classes. Nash mentions in a footnote the classes $S^m_{\rho,\delta}$, defined by

$$\left|\partial_x^\beta \partial_\xi^\alpha \sigma\right| \leq C_{\alpha\beta}(K)\langle\xi\rangle^{m-\rho|\alpha|+\delta|\beta|}, \qquad 0 \leq \delta < \rho \leq 1.$$

Here ρ is how much decay each ξ-derivative buys and δ is how much growth each x-derivative costs. Our S^m is $S^m_{1,0}$. For example $\sigma(\xi) = e^{i\langle\xi\rangle^{1/2}}$ has $\partial_\xi\sigma = \frac{i}{2}\xi\langle\xi\rangle^{-3/2}\sigma$, of size $\langle\xi\rangle^{-1/2}$: each derivative gains only half a power, so $\sigma \in S^0_{1/2,0}$ but $\sigma \notin S^0$. Such classes appear for parametrices of hypo-elliptic but non-elliptic operators (for example the heat operator, where time counts as two powers of space). We will not need them.

2.4.5 Pseudo-differential operators on manifolds and between bundles

On a compact manifold we cannot Fourier transform globally, but a ΨDO is defined chart by chart, exactly as a differential operator is.

Working definition 2.24 (ΨDO between vector bundles (Nash's definition))

Let M be a compact n-manifold and $U, V \to M$ vector bundles of ranks s and r. A linear map $P : \Gamma(M,U) \to \Gamma(M,V)$ is a **pseudo-differential operator of order** m if in every chart $\Omega \subset M$ with coordinates x, dual momenta ξ, and local trivialisations [local frames, in which sections become columns of functions] of U and V, P acts on vector-valued functions by (2.118) with an $r \times s$ *matrix* symbol $\sigma(x,\xi)$ each of whose entries lies in $S^m(\Omega)$, up to a smoothing operator. *Operationally:* cover M by charts Ω_i, choose a partition of unity [smooth $\phi_i \geq 0$, each zero outside its chart Ω_i] $\sum_i \phi_i = 1$ and functions ψ_i with $\psi_i = 1$ on $\operatorname{supp}\phi_i$; then $P = \sum_i \phi_i \operatorname{Op}(\sigma_i)\,\psi_i + R$ with R smoothing (a smooth kernel on $M \times M$).

The phrase "up to a smoothing operator" handles the fact that a ΨDO is not local: Pu on a chart Ω depends on u outside Ω too, but only through a smooth kernel (pseudo-locality again).

How the symbol changes under a change of coordinates. The full symbol $\sigma(x,\xi)$ changes in a messy way under changes of coordinates (lower-order terms mix in). The principal symbol does not: it transforms as a function on the cotangent bundle T^*M, with ξ transforming as a covector $\xi_j\,\mathrm{d}x^j$. Here is the simplest check.

Worked example 2.52: The principal symbol lives on T^*M

On an interval let $L = \mathrm{d}^2/\mathrm{d}x^2$, with $\sigma_2 = (i\xi)^2 = -\xi^2$ (from $\mathrm{d}/\mathrm{d}x = iD_x$). Change coordinates $x = g(t)$ with $g' > 0$. Then

$$\frac{\mathrm{d}}{\mathrm{d}x} \overset{(1)}{=} \frac{1}{g'(t)}\frac{\mathrm{d}}{\mathrm{d}t}, \qquad \frac{\mathrm{d}^2}{\mathrm{d}x^2} \overset{(2)}{=} \frac{1}{g'(t)^2}\frac{\mathrm{d}^2}{\mathrm{d}t^2} - \frac{g''(t)}{g'(t)^3}\frac{\mathrm{d}}{\mathrm{d}t},$$

(1) Chain rule, $\mathrm{d}t/\mathrm{d}x = 1/g'$.

(2) Apply the first formula twice; the derivative of $1/g'$ produces the first-order term.

In the t-chart the principal symbol is $-\tau^2/g'(t)^2$, where τ is the momentum dual to t, and the first-order term has changed. Now compare: since $\mathrm{d}x = g'\,\mathrm{d}t$, the covector $\xi\,\mathrm{d}x$ equals $\tau\,\mathrm{d}t$ exactly when $\tau = g'\xi$. Substituting, $-\tau^2/g'^2 = -\xi^2$: the principal symbol is the *same function* on T^*M, written in two coordinate systems. The lower-order part of the symbol is not.

So the principal symbol of $P : \Gamma(U) \to \Gamma(V)$ is a globally defined object: for each point $x \in M$ and each covector $\xi \in T^*_xM$, a linear map $\sigma_m(P)(x,\xi) : U_x \to V_x$. Pulling U and V back along the projection $\pi : T^*M \to M$ (so the fibre of π^*U over (x,ξ) is U_x) we can say it in one breath:

$$\sigma_m(P) :\ \pi^*U \longrightarrow \pi^*V \qquad \text{is a bundle map over } T^*M.$$

2.4.6 The principal symbol of a pseudo-differential operator

For a differential operator the principal symbol is the top-degree part of a polynomial. For a general $\sigma \in S^m$ there is no "top degree", so Nash defines it by zooming out to large momentum.

Working definition 2.25 (m-th order (principal) symbol, Nash (2.56))

If P has symbol $\sigma \in S^m$ [growing at most like $|\xi|^m$ at large momentum], its **m-th order symbol** is

$$\sigma_m(x,\xi) = \lim_{\mu\to+\infty} \frac{\sigma(x,\mu\xi)}{\mu^m}, \qquad \xi \neq 0, \tag{2.124}$$

when the limit exists. *To compute:* expand σ at large $|\xi|$ and keep the leading term. It is **positively homogeneous** of degree m: $\sigma_m(x,\mu\xi) = \mu^m \sigma_m(x,\xi)$ for $\mu > 0$ (replace ξ by $\mu\xi$ in the limit and rescale).

Worked example 2.53: Principal symbols by zooming out

(a) Polynomials. $\sigma = -a(x)\xi^2 + ib(x)\xi + c(x)$ gives $\sigma(x,\mu\xi)/\mu^2 = -a\xi^2 + ib\xi/\mu + c/\mu^2 \to -a(x)\xi^2$: the top-degree part, as before.

(b) The relativistic kinetic energy. For $\sigma = \sqrt{|\xi|^2 + m^2} \in S^1$,

$$\frac{\sigma(\mu\xi)}{\mu} \overset{(1)}{=} \sqrt{|\xi|^2 + m^2/\mu^2} \xrightarrow{\mu\to\infty} |\xi|\,.$$

(1) Pull μ^2 out of the square root ($\mu > 0$).

So $\sigma_1 = |\xi|$: at high momentum every particle is massless. Keep going and we get the full large-momentum (**classical** or **polyhomogeneous**) expansion, a sum of terms of decreasing homogeneity:

$$\sqrt{|\xi|^2 + m^2} \overset{(1)}{=} |\xi| \left(1 + \frac{m^2}{|\xi|^2}\right)^{1/2} \overset{(2)}{=} |\xi| + \frac{m^2}{2\,|\xi|} - \frac{m^4}{8\,|\xi|^3} + \frac{m^6}{16\,|\xi|^5} - \cdots$$

(1) Factor out $|\xi|$ (valid for $\xi \neq 0$).

(2) Binomial series $(1+u)^{1/2} = 1 + \frac{u}{2} - \frac{u^2}{8} + \frac{u^3}{16} - \cdots$ with $u = m^2/|\xi|^2 < 1$ for $|\xi| > m$.

Each term is positively homogeneous: of degree $1, -1, -3, -5, \ldots$ This is the non-relativistic expansion read backwards: $E = |p| + m^2/2\,|p| + \ldots$ is the ultra-relativistic expansion.

(c) A symbol of order 0 that is only positively homogeneous. Take $\sigma(\xi) = \xi/\langle\xi\rangle$ on $\mathbb{R}$. Then $\sigma(\mu\xi) = \mu\xi/\sqrt{1+\mu^2\xi^2} \to \operatorname{sgn}(\xi)$, so $\sigma_0(\xi) = \operatorname{sgn}\xi$. Now $\sigma_0(\mu\xi) = \operatorname{sgn}(\mu)\operatorname{sgn}(\xi)$, which equals $\mu^0\sigma_0(\xi)$ for $\mu > 0$ but not for $\mu < 0$. Similarly $|\xi|$ is positively homogeneous of degree 1 but $|-\xi| \neq -|\xi|$. Symbols of differential operators are polynomials, so they are homogeneous for all real μ; for ΨDOs only $\mu > 0$ is guaranteed. The operator $\operatorname{sgn}(D)$ is i times the Hilbert transform; on the circle it multiplies $e^{in\theta}$ by $\operatorname{sgn} n$.

Every differential operator is a ΨDO. If $\sigma = \sum_{|\alpha|\le m} a_\alpha(x)\xi^\alpha$, then each monomial ξ^α with $|\alpha| \le m$ obeys (2.122) (as in Example 2.50(a)), so $\sigma \in S^m$, and (2.124) returns the top-degree part. So the new definitions extend the old ones without changing them.

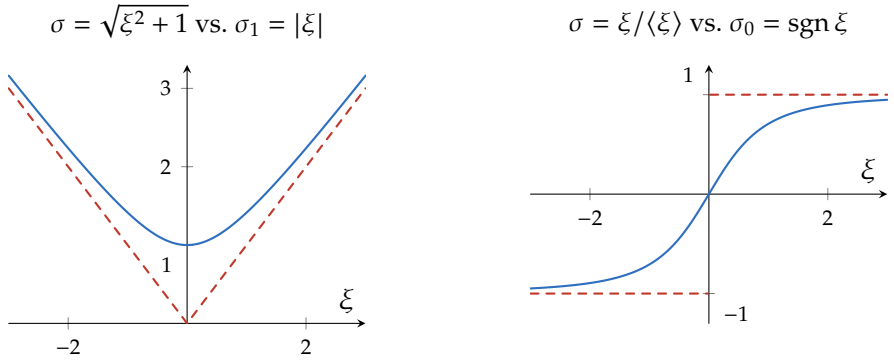


Figure 2.30: The principal symbol is what the full symbol looks like at large momentum. Left: $\sqrt{\xi^2+1}$ (blue) approaches $|\xi|$ (red dashed); the difference $\sim 1/2\,|\xi|$ is of order -1. Right: $\xi/\langle\xi\rangle$ approaches $\operatorname{sgn}\xi$. Both principal symbols have a corner or jump at $\xi = 0$, which is harmless: ellipticity and the whole theory only use $\xi \neq 0$ ("off the zero section"), and only large $|\xi|$ matters.

2.4.7 Ellipticity for pseudo-differential operators

Nothing new is needed: ellipticity is still invertibility of the principal symbol away from $\xi = 0$.

Working definition 2.26 (elliptic ΨDO)

Let U, V have the same rank r. $P : \Gamma(U) \to \Gamma(V)$ with symbol in S^m is **elliptic of order** m if its m-th order symbol [the large-momentum limit $\lim_{\mu\to\infty} \sigma(x,\mu\xi)/\mu^m$] $\sigma_m(x,\xi)$ exists and is an invertible $r \times r$ matrix for every x and every $\xi \neq 0$. Equivalently (Nash's "alternative definition"):

$$\sigma_m(P) : \pi^*U \to \pi^*V \text{ is a bundle isomorphism off the zero section of } T^*M.$$

Test: compute σ_m and its determinant; check $\det \sigma_m(x,\xi) \neq 0$ for $\xi \neq 0$. Quantitatively (by homogeneity and compactness of the unit sphere in ξ), ellipticity means $\left|\sigma_m(x,\xi)^{-1}\right| \leq C\,|\xi|^{-m}$: *the symbol is bounded below by its order.*

Watch out: Elliptic of order m is not elliptic of order $m+1$

An operator of order m is also of order $m+1$, since $S^m \subset S^{m+1}$. But its $(m+1)$-th order symbol is $\lim \sigma(x,\mu\xi)/\mu^{m+1} = \lim\big(\sigma(x,\mu\xi)/\mu^m\big)/\mu = \sigma_m/\mu \to 0$, which is not invertible. So ellipticity pins down the order. For example, Δ is elliptic of order 2 and not of order 3.

Worked example 2.54: Which of our operators are elliptic?

Operator	symbol	σ_m	elliptic?
$\Delta + m^2$	$\lvert\xi\rvert^2 + m^2$	$\lvert\xi\rvert^2$, $m = 2$	yes
$\sqrt{\Delta + m^2}$	$\sqrt{\lvert\xi\rvert^2 + m^2}$	$\lvert\xi\rvert$, $m = 1$	yes
$(\Delta + m^2)^{-1}$	$(\lvert\xi\rvert^2 + m^2)^{-1}$	$\lvert\xi\rvert^{-2}$, $m = -2$	yes
$\operatorname{sgn}(D)$ on $\mathbb{R}$	$\xi/\langle\xi\rangle$	$\operatorname{sgn}\xi$, $m = 0$	yes
$\not{D}$ on $\mathbb{R}^2$	$\sigma^1\xi_1 + \sigma^2\xi_2$	same, $m = 1$	yes ($\det = -\lvert\xi\rvert^2$)
∂_1 on $\mathbb{R}^2$	$i\xi_1$	$i\xi_1$, $m = 1$	no: vanishes on $\xi = (0, \xi_2)$
$e^{-t\Delta}$	$e^{-t\lvert\xi\rvert^2}$	0 for every m	no (order $-\infty$)

Inverses of elliptic operators are elliptic (of the negative order), and their principal symbols are the inverse matrices. Smoothing operators are never elliptic: an elliptic operator must "see" arbitrarily high momenta with full strength.

2.4.8 The symbol calculus: composing, adjoining, commuting

The single most useful fact about ΨDOs is that their symbols compose *almost* like numbers. This is where the theory earns its keep, and it is also where the physicist's intuition about quantisation (x and p almost commute, up to $\hbar$) is made precise.

Constant coefficients: exact multiplication. If $P = p(D)$ and $Q = q(D)$ are Fourier multipliers, then

$$\widehat{PQu}(\xi) \overset{(1)}{=} p(\xi)\,\widehat{Qu}(\xi) \overset{(2)}{=} p(\xi)q(\xi)\,\hat{u}(\xi),$$

(1) P multiplies the Fourier transform of whatever it acts on by $p(\xi)$.
(2) Likewise for Q.

so $\sigma_{PQ} = pq$ exactly, and $PQ = QP$. For instance $\sqrt{\Delta + m^2} \circ \sqrt{\Delta + m^2} = \Delta + m^2$ and $(\Delta + m^2)^z(\Delta + m^2)^w = (\Delta + m^2)^{z+w}$.

Variable coefficients: the composition formula. When coefficients depend on x, derivatives in P also hit the x-dependence of Q's symbol. Let us derive the formula when P is a differential operator, where everything is a finite sum. Write $P = \sum_\alpha a_\alpha(x)D^\alpha$ with symbol $p(x,\xi) = \sum_\alpha a_\alpha(x)\xi^\alpha$, and Q with symbol $q(x,\xi)$. Then

$$\begin{aligned} PQu(x) &\overset{(1)}{=} \sum_\alpha a_\alpha(x) D_x^\alpha \int e^{ix\cdot\xi} q(x,\xi)\hat{u}(\xi)\,\frac{\mathrm{d}^n\xi}{(2\pi)^n} \\ &\overset{(2)}{=} \int e^{ix\cdot\xi} \sum_\alpha a_\alpha(x)(\xi + D_x)^\alpha q(x,\xi)\,\hat{u}(\xi)\,\frac{\mathrm{d}^n\xi}{(2\pi)^n} \\ &\overset{(3)}{=} \int e^{ix\cdot\xi} \sum_\alpha a_\alpha(x) \sum_{\gamma\le\alpha} \binom{\alpha}{\gamma} \xi^{\alpha-\gamma} D_x^\gamma q(x,\xi)\,\hat{u}(\xi)\,\frac{\mathrm{d}^n\xi}{(2\pi)^n} \\ &\overset{(4)}{=} \int e^{ix\cdot\xi} \sum_\gamma \frac{1}{\gamma!}\,\partial_\xi^\gamma p(x,\xi)\,D_x^\gamma q(x,\xi)\,\hat{u}(\xi)\,\frac{\mathrm{d}^n\xi}{(2\pi)^n}. \end{aligned} \tag{2.125}$$

(1) Definition of Q, then apply P.

(2) Leibniz: $D_j(e^{ix\cdot\xi}f) = e^{ix\cdot\xi}(\xi_j + D_j)f$, since $D_j e^{ix\cdot\xi} = \xi_j e^{ix\cdot\xi}$.
(3) Binomial expansion in each variable (ξ_j and D_j commute when acting on functions of x): $(\xi + D)^\alpha = \sum_{\gamma\le\alpha}\binom{\alpha}{\gamma}\xi^{\alpha-\gamma}D^\gamma$, with $\binom{\alpha}{\gamma} = \prod_j \binom{\alpha_j}{\gamma_j}$.
(4) $\frac{1}{\gamma!}\partial_\xi^\gamma \xi^\alpha = \binom{\alpha}{\gamma}\xi^{\alpha-\gamma}$; then sum over α to rebuild $\partial_\xi^\gamma p$. Here $\gamma! = \prod_j \gamma_j!$.

Key idea: Composition of symbols

The symbol of PQ is

$$\sigma_{PQ}(x,\xi) \sim \sum_\gamma \frac{1}{\gamma!}\,\partial_\xi^\gamma \sigma_P(x,\xi)\, D_x^\gamma \sigma_Q(x,\xi), \qquad D_x = -i\partial_x. \tag{2.126}$$

For differential P the sum is finite and exact. For general $P \in S^m$, $Q \in S^{m'}$ it is an asymptotic series: the γ-th term is in $S^{m+m'-|\gamma|}$ (each ∂_ξ lowers the order by one and D_x does not raise it), and truncating after $|\gamma| < N$ leaves an error in $S^{m+m'-N}$. In particular PQ has order $m + m'$ and

$$\sigma_{m+m'}(PQ) = \sigma_m(P)\,\sigma_{m'}(Q). \tag{2.127}$$

Equation (2.127) is the ΨDO version of Nash's (2.45) for differential operators, which he used to compute the symbol of the Hodge Laplacian from those of d and d*.

Worked example 2.55: The canonical commutation relation from the symbol calculus

On $\mathbb{R}$ take $P = D$ (symbol ξ) and $Q = x$ (multiplication, symbol x). By (2.126),

$$\sigma_{DQ} \overset{(1)}{=} \xi\cdot x + \partial_\xi(\xi)\,D_x(x) = x\xi - i, \qquad \sigma_{xD} \overset{(2)}{=} x\cdot\xi = x\xi,$$

(1) Terms $\gamma = 0$ and $\gamma = 1$; $D_x(x) = -i$; higher terms vanish since $\partial_\xi^2\xi = 0$.
(2) Here $P = x$ has no ξ-dependence, so only $\gamma = 0$ survives.

Hence $[D, x]$ has symbol $-i$: $[D, x] = -i$, i.e. $[x, \hat{p}] = i$ with $\hat{p} = D$. Directly: $D(xu) - xDu = -iu$. The product of symbols $x\xi$ is the same for both orders; the commutator lives entirely in the first correction.

Commutators and Poisson brackets. Apply (2.126) to PQ and QP and subtract. The $\gamma = 0$ terms cancel (symbols commute as scalars), and the $|\gamma| = 1$ terms give

$$\begin{aligned}\sigma_{[P,Q]} &\overset{(1)}{=} \sum_j \Big(\partial_{\xi_j} p\, D_{x^j} q - \partial_{\xi_j} q\, D_{x^j} p\Big) + \dots \\ &\overset{(2)}{=} i\sum_j \Big(\partial_{x^j} p\,\partial_{\xi_j} q - \partial_{\xi_j} p\,\partial_{x^j} q\Big) + \cdots = i\,\{p,q\} + \dots\end{aligned} \tag{2.128}$$

(1) First-order terms of (2.126) for PQ minus those for QP (scalar symbols).
(2) $D_x = -i\partial_x$; reorder. The bracket is the classical Poisson bracket $\{p,q\} = \sum_j(\partial_{x^j}p\,\partial_{\xi_j}q - \partial_{\xi_j}p\,\partial_{x^j}q)$, with the physicists' sign ($\{x,\xi\} = 1$).

The dots are of order $m + m' - 2$. So $[P, Q]$ has order $m + m' - 1$, one less than naive counting, and its principal symbol is $i\{\sigma_m(P), \sigma_{m'}(Q)\}$.

Dictionary: symbol calculus ↔ quantum mechanics

ΨDO theory	Quantum mechanics ($\hbar = 1$)
symbol $\sigma(x,\xi)$	classical observable $f(x,p)$ on phase space
$\mathrm{Op}(\sigma)$	its quantisation ("x to the left, p to the right")
T^*M	phase space
$\sigma_{PQ} = \sigma_P\sigma_Q +$ lower order	$\widehat{fg} = \hat f\hat g + O(\hbar)$
$\sigma_{[P,Q]} = i\{\sigma_P,\sigma_Q\} + \dots$	Dirac's rule $[\hat f,\hat g] = i\hbar\widehat{\{f,g\}} + O(\hbar^2)$
"lower order" (one fewer power of ξ)	one more power of $\hbar$ (semiclassical expansion)

Large momentum is the classical limit: at $|\xi| \to \infty$ the quantum corrections (lower-order terms) are negligible relative to the leading term. That is why ellipticity, a statement about the principal symbol, is a "classical" condition with "quantum" consequences.

Worked example 2.56: A commutator with the Laplacian

Let $f(x)$ be a smooth function on $\mathbb{R}^n$, viewed as a multiplication operator of order 0. By (2.128), $[\Delta, f]$ has order $2+0-1=1$ and principal symbol

$$i\{|\xi|^2, f\} \overset{(1)}{=} i\Big(\sum_j \partial_{x^j}|\xi|^2\,\partial_{\xi_j} f - \partial_{\xi_j}|\xi|^2\,\partial_{x^j} f\Big) \overset{(2)}{=} -2i\,\xi\cdot\nabla f.$$

(1) Definition of the bracket.
(2) $|\xi|^2$ does not depend on x, f does not depend on ξ, and $\partial_{\xi_j}|\xi|^2 = 2\xi_j$.

Check directly: $[\Delta,f]u = -\nabla^2(fu) + f\nabla^2 u = -(\nabla^2 f)u - 2\nabla f\cdot\nabla u$. The first-order part is $-2\nabla f\cdot\nabla u = -2i\,\nabla f\cdot Du$, whose symbol is $-2i\,\xi\cdot\nabla f$. It agrees.

Adjoints. With respect to the L^2 inner product $\langle u,v\rangle = \int \bar u\, v$ (or $\int u^\dagger v$ for vector-valued functions), the formal adjoint P^* has symbol

$$\sigma_{P^*}(x,\xi) \sim \sum_\gamma \frac{1}{\gamma!}\partial_\xi^\gamma D_x^\gamma\,\sigma_P(x,\xi)^\dagger, \qquad \text{so} \qquad \sigma_m(P^*) = \sigma_m(P)^\dagger, \tag{2.129}$$

which is Nash's (2.47). The principal symbol of the adjoint is the conjugate-transpose matrix, so *P is elliptic iff P^* is elliptic*. This small fact will give us finite cokernels.

Worked example 2.57: Adjoint of a first-order operator

On $\mathbb{R}$ let $P = a(x)D$ with a complex-valued. For test functions u, v,

$$\langle Pu, v\rangle \overset{(1)}{=} \int \overline{a\,(-i)u'}\,v\,\mathrm{d}x = \int i\,\bar a\,\bar u'\,v\,\mathrm{d}x \overset{(2)}{=} -i\int \bar u\,(\bar a v)'\,\mathrm{d}x \overset{(3)}{=} \langle u, D(\bar a v)\rangle,$$

(1) $D = -i\,\mathrm{d}/\mathrm{d}x$ and complex conjugation.
(2) Integrate by parts (test functions vanish at infinity).
(3) $-i(\bar a v)' = D(\bar a v)$.

So $P^*v = D(\bar a v) = \bar a\,Dv + (D\bar a)v$, with symbol $\bar a\xi + D_x\bar a$. The formula (2.129) gives $\gamma=0$: $\bar a\xi$; $\gamma=1$: $\partial_\xi D_x(\bar a\xi) = D_x\bar a$. It agrees, and the principal symbol $\bar a\xi$ is the complex conjugate of $a\xi$.

2.4.9 Inverting an elliptic operator: the parametrix

Now the payoff. An elliptic operator cannot always be inverted exactly (it may have a kernel). But it can always be inverted *up to a smoothing operator*, and the inverse is again a ΨDO. The construction is a bootstrap that uses nothing but (2.126).

Working definition 2.27 (parametrix)

A **parametrix** for P is an operator Q with

$$PQ = I - R_1, \qquad QP = I - R_2, \qquad R_1, R_2 \text{ smoothing (order } -\infty).$$

Here smoothing means [given by a smooth kernel $R(x,y)$, so Ru is smooth for every input u]. *To build one for an elliptic P of order m:*

(i) First guess: $q_0(x,\xi) = \chi(\xi)\,\sigma_m(P)(x,\xi)^{-1} \in S^{-m}$, with χ a cutoff vanishing near $\xi = 0$ (where σ_m may fail to be invertible or smooth).

(ii) Compute $PQ_0 = I - R$ by (2.126). The leading term is $\sigma_m(P)q_0 = 1$ for large ξ, so R has order -1.

(iii) Correct: $Q = Q_0(I + R + R^2 + R^3 + \cdots)$, summed asymptotically [one operator whose difference from the first N terms has order $-m-N$, for every N] (the k-th term has order $-m-k$). Then $PQ = (I-R)(I+R+R^2+\cdots) = I$ up to order $-\infty$.

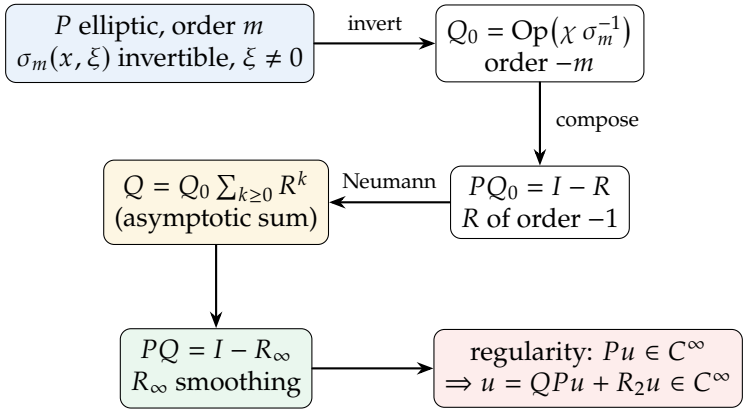


Figure 2.31: The logic of the parametrix construction. Only the principal symbol has to be inverted by hand (a matrix inversion at each (x,ξ)); the symbol calculus does the rest, one order at a time. The end product, invertibility modulo smoothing operators, gives elliptic regularity at once (bottom right) and, on Sobolev spaces, the Fredholm property (Section 2.6).

Worked example 2.58: Parametrix of a variable-coefficient second-order operator

On $\mathbb{R}$ let $P = a(x)D^2 + b(x)D + c(x)$ with $a(x) > 0$ smooth, so $p = a\xi^2 + b\xi + c$ and $\sigma_2(P) = a\xi^2$ is invertible for $\xi \neq 0$. *Step 1.* $q_0 = \chi(\xi)/(a(x)\xi^2)$. For $|\xi| \geq 1$ (where $\chi = 1$):

$$\begin{aligned}\sigma_{PQ_0} &\overset{(1)}{=} p\,q_0 + \partial_\xi p\, D_x q_0 + \tfrac{1}{2}\partial_\xi^2 p\, D_x^2 q_0 \\ &\overset{(2)}{=} \left(1 + \frac{b}{a\xi} + \frac{c}{a\xi^2}\right) + (2a\xi + b)\,\frac{i\,a'}{a^2\xi^2} + a\,D_x^2\!\left(\frac{1}{a}\right)\frac{1}{\xi^2} \\ &\overset{(3)}{=} 1 + \frac{b + 2i\,a'}{a\,\xi} + O(\xi^{-2}).\end{aligned}$$

(1) Composition formula (2.126); it stops at $\gamma = 2$ because p is quadratic in ξ.

(2) $p\,q_0$ multiplied out; $D_x q_0 = -i\,\partial_x(a^{-1})\xi^{-2} = i\,a'a^{-2}\xi^{-2}$; $\frac{1}{2}\partial_\xi^2 p = a$.

(3) Collect the terms of order ξ^{-1}: $b/(a\xi)$ and $2i\,a'/(a\xi)$. Everything else is $O(\xi^{-2})$.

So $R = I - PQ_0$ has order -1 with principal symbol $-(b + 2ia')/(a\xi)$. *Step 2.* Add q_1 of order -3 chosen so that $\sigma_2(P)\,q_1$ cancels it:

$$a\xi^2\, q_1 = -\frac{b + 2i\,a'}{a\,\xi} \quad\Longrightarrow\quad q_1 = -\frac{b + 2i\,a'}{a^2\,\xi^3}.$$

Then $P(Q_0 + Q_1) = I + (\text{order } -2)$, and we could continue. *Sanity check with constant coefficients* ($a = 1, a' = 0$): the exact inverse symbol is $1/(\xi^2 + b\xi + c) = \xi^{-2}(1 + b/\xi + c/\xi^2)^{-1} = \xi^{-2} - b\xi^{-3} + O(\xi^{-4})$, and indeed $q_0 + q_1 = \xi^{-2} - b\xi^{-3}$. The term $2ia'$ is the new, purely "quantum" correction from the x-dependence of a.

Worked example 2.59: On the circle the parametrix is an inverse up to rank one

On S^1 let $P = D^2 = -\mathrm{d}^2/\mathrm{d}\theta^2$, which multiplies $e^{in\theta}$ by n^2. Define Q by $Qe^{in\theta} = n^{-2}e^{in\theta}$ for $n \neq 0$ and $Qe^0 = 0$. Its symbol is n^{-2} for $n \neq 0$: order -2. Let $\Pi_0 u = \frac{1}{2\pi}\int_0^{2\pi} u\,\mathrm{d}\theta$ (projection onto constants). Then for $u = \sum c_n e^{in\theta}$,

$$PQu \overset{(1)}{=} \sum_{n\neq 0} n^2 \cdot n^{-2} c_n e^{in\theta} = u - c_0 \overset{(2)}{=} (I - \Pi_0)u,$$

(1) Apply Q then P mode by mode.

(2) $c_0 = \frac{1}{2\pi}\int u\,\mathrm{d}\theta = \Pi_0 u$.

and likewise $QP = I - \Pi_0$. The error Π_0 has the smooth kernel $K(\theta, \theta') = 1/2\pi$, so it is smoothing; it is also of rank one. It is exactly the projection onto $\operatorname{Ker} P$. This is the whole Fredholm story in miniature: invertible except on a finite-dimensional space.

Key idea: Elliptic ⇔ invertible modulo smoothing; elliptic regularity

A ΨDO is elliptic if and only if it has a parametrix. Consequence: if P is elliptic and $Pu = f$ with f smooth on an open set, then u is smooth there. Reason: $u = QPu + R_2 u = Qf + R_2 u$. Q, being a ΨDO, maps smooth functions to smooth functions (pseudo-locality, next subsection), and R_2 maps *anything* to a smooth function. So elliptic ΨDOs are hypo-elliptic, which generalises the result of the previous section from differential operators to ΨDOs.

2.4.10 Nash's gallery I: solving an integral equation with a Bessel potential

Nash motivates ΨDOs with a practical problem. Suppose we are handed an **integral equation of the first kind**,

$$Pu(x) = \int_{\mathbb{R}^n} K(x, y)\, u(y)\, \mathrm{d}^n y = f(x), \tag{2.130}$$

with the kernel K and the right-hand side f known, and u unknown. By (2.120)–(2.121), P is the operator with symbol $\sigma(x, \xi) = \int K(x, x - z)e^{-iz\cdot\xi}\,\mathrm{d}^n z$. If this symbol lies in some S^m and is elliptic, the machinery of Section 2.4.9 inverts P, at least modulo smoothing operators. When K depends only on $x - y$ and the symbol has no zeros, we can do better: the inverse is *exactly* the Fourier multiplier $1/\sigma(\xi)$, and

$$u(x) = \int e^{ix\cdot\xi}\,\frac{\hat f(\xi)}{\sigma(\xi)}\,\frac{\mathrm{d}^n\xi}{(2\pi)^n}. \tag{2.131}$$

Nash's example is the one-dimensional family of symbols

$$\sigma_\lambda(\xi) = (a^2 + \xi^2)^\lambda, \qquad a > 0, \quad -\tfrac{1}{2} < \lambda < 0. \tag{2.132}$$

By Example 2.50(b),(c) (the same computation with a^2 in place of 1), $\sigma_\lambda \in S^{2\lambda}$. Its principal symbol is $|\xi|^{2\lambda}$, which never vanishes for $\xi \neq 0$: P_λ is elliptic of order $2\lambda \in (-1, 0)$. It is a smoothing-by-a-fraction-of-a-derivative operator, a **Bessel potential**.

Its kernel. We need one classical integral, Basset's representation of the modified Bessel function K_ν: for $x > 0$ and $\nu > -\frac{1}{2}$,

$$\int_0^\infty \frac{\cos(xt)}{(t^2 + a^2)^{\nu+1/2}}\,\mathrm{d}t = \frac{\sqrt{\pi}}{\Gamma(\nu + \frac{1}{2})}\Big(\frac{x}{2a}\Big)^\nu K_\nu(ax). \tag{2.133}$$

(Recall that K_ν is the decaying solution of $z^2w'' + zw' - (z^2 + \nu^2)w = 0$, that $K_{-\nu} = K_\nu$, that $K_\nu(z) \approx \sqrt{\pi/2z}\,e^{-z}$ for large z, and that $K_\nu(z) \approx \frac{1}{2}\Gamma(\nu)(2/z)^\nu$ for small z and $\nu > 0$. The special case $K_{1/2}(z) = \sqrt{\pi/2z}\,e^{-z}$ is elementary.)

Worked example 2.60: The kernel of the Bessel potential $(a^2 + D^2)^\lambda$

By (2.120) the kernel is $k_\lambda(x - y)$ with

$$\begin{aligned} k_\lambda(x) &\overset{(1)}{=} \int_{-\infty}^{\infty} e^{ix\xi}(a^2 + \xi^2)^\lambda \frac{\mathrm{d}\xi}{2\pi} = \frac{1}{\pi}\int_0^\infty \cos(|x|\,\xi)\,(a^2 + \xi^2)^\lambda\,\mathrm{d}\xi \\ &\overset{(2)}{=} \frac{1}{\pi}\cdot\frac{\sqrt{\pi}}{\Gamma(-\lambda)}\Big(\frac{|x|}{2a}\Big)^{-\lambda-\frac{1}{2}} K_{-\lambda-\frac{1}{2}}(a\,|x|) \\ &\overset{(3)}{=} \frac{(2a)^{\lambda+\frac{1}{2}}}{\sqrt{\pi}\,\Gamma(-\lambda)}\,\frac{K_{\lambda+\frac{1}{2}}(a\,|x|)}{|x|^{\lambda+\frac{1}{2}}}. \end{aligned} \tag{2.134}$$

(1) The symbol is even in ξ, so only the cosine part of $e^{ix\xi}$ survives, and the integral over $\mathbb{R}$ is twice that over $(0, \infty)$. The integral converges (conditionally) because $\lambda < 0$.
(2) Basset's formula (2.133) with $\nu + \frac{1}{2} = -\lambda$, i.e. $\nu = -\lambda - \frac{1}{2}$, which is $> -\frac{1}{2}$ exactly because $\lambda < 0$.
(3) Collect the constants and use $K_{-\nu} = K_\nu$.

This is Nash's first formula in (2.60): his prefactor $2^{1+\lambda}a^{\frac{1}{2}+\lambda}/(\sqrt{2\pi}\,\Gamma(-\lambda))$ equals ours, since $2^{1+\lambda}/\sqrt{2\pi} = 2^{\lambda+\frac{1}{2}}/\sqrt{\pi}$.
Check at $\lambda = -1$ (outside Nash's range, but the formula is valid for all $\lambda < 0$). Then $\nu = \frac{1}{2}$ and $K_{1/2}(z) = \sqrt{\pi/2z}\,e^{-z}$, so

$$k_{-1}(x) \overset{(1)}{=} \frac{(2a)^{-1/2}}{\sqrt{\pi}\cdot 1}\,|x|^{1/2}\sqrt{\frac{\pi}{2a\,|x|}}\,e^{-a|x|} \overset{(2)}{=} \frac{e^{-a|x|}}{2a},$$

(1) Put $\lambda = -1$ into (2.134); $\Gamma(1) = 1$.
(2) The powers of $|x|$ cancel and $(2a)^{-1/2}(2a)^{-1/2} = 1/2a$.

which is the resolvent kernel of Example 2.47. Good.
Short-distance behaviour. For $|x| \to 0$, with $\nu = \lambda + \frac{1}{2} \in (0, \frac{1}{2})$,

$$k_\lambda(x) \overset{(1)}{=} \frac{(2a)^\nu}{\sqrt{\pi}\,\Gamma(-\lambda)}\,|x|^{-\nu}\cdot\frac{\Gamma(\nu)}{2}\Big(\frac{2}{a\,|x|}\Big)^\nu\big(1 + o(1)\big) \overset{(2)}{=} \frac{2^{2\lambda}\,\Gamma(\lambda + \frac{1}{2})}{\sqrt{\pi}\,\Gamma(-\lambda)}\,\frac{1 + o(1)}{|x|^{1+2\lambda}}.$$

(1) Small-argument form of K_ν.

(2) $(2a)^{\nu}2^{\nu}a^{-\nu}/2 = 2^{2\nu-1} = 2^{2\lambda}$ and $|x|^{-2\nu} = |x|^{-1-2\lambda}$.

So the kernel is singular only at $x = y$, like $|x-y|^{-n-m}$ with $n = 1$ and order $m = 2\lambda$, exactly as the symbol–kernel dictionary predicts. Because $-1 < -1-2\lambda < 0$ the singularity is integrable, and $P_\lambda u$ is a genuine integral. At large distance k_λ decays like $|x|^{-\lambda-1}\, e^{-a|x|}$: the mass a sets a screening length $1/a$.

The solution. The inverse of P_λ is $P_{-\lambda}$, whose symbol $(a^2+\xi^2)^{-\lambda}$ has positive order $-2\lambda \in (0,1)$. Formally its kernel is (2.134) with $\lambda \to -\lambda$, and that is Nash's second formula in (2.60):

$$u(x) = \frac{(2a)^{\frac{1}{2}-\lambda}}{\sqrt{\pi}\,\Gamma(\lambda)} \int_{\mathbb{R}} \frac{K_{\frac{1}{2}-\lambda}(a\,|x-y|)}{|x-y|^{\frac{1}{2}-\lambda}}\, f(y)\,\mathrm{d}y. \tag{2.135}$$

But look at it closely. $\Gamma(\lambda) < 0$ for $\lambda \in (-\frac{1}{2}, 0)$, and the kernel behaves like $|x-y|^{-1+2\lambda}$ near the diagonal, with $-1+2\lambda < -1$: *not integrable*. This is no accident. An operator of positive order differentiates, and differentiation cannot be written as an honest integral against a locally integrable kernel. The integral in (2.135) has to be read as a distribution (Hadamard's *finite part*: subtract $f(x)$ from $f(y)$ inside the integral, which makes it converge). Nash alludes to this with "we take an appropriate definition of the fractional powers". In momentum space there is no subtlety at all: (2.131) with $1/\sigma_\lambda = (a^2+\xi^2)^{-\lambda}$.

Watch out: Positive order means a non-integrable kernel

Kernels of negative-order operators are locally integrable (in n dimensions, order $m < 0$ gives $|x-y|^{-n-m}$ with $-n-m > -n$). Kernels of operators of order $m \geq 0$ are not functions near the diagonal: they are δ and its derivatives (differential operators) or finite-part distributions like f.p. $|x-y|^{-n-m}$. This is why symbols, not kernels, are the convenient way to handle ΨDOs.

2.4.11 Nash's gallery II: fractional integration and differentiation

Now set $a = 0$ in (2.132). The symbol becomes

$$\sigma_\lambda(\xi) = |\xi|^{2\lambda}, \tag{2.136}$$

and $P_\lambda = |D|^{2\lambda} = \Delta^\lambda$ is a fractional power of the Laplacian. For $\lambda = 1$ it is $\Delta = -\mathrm{d}^2/\mathrm{d}x^2$ itself; for $\lambda = \frac{1}{2}$ it is $|D|$; for $\lambda < 0$ it is **fractional integration** of order -2λ; its inverse $\Delta^{-\lambda}$ is **fractional differentiation** of order -2λ. Strictly $|\xi|^{2\lambda}$ is not in $S^{2\lambda}$ because it is not smooth at $\xi = 0$ (the pitfall in Section 2.4.4); but only large $|\xi|$ matters for the order, and the operator is perfectly well defined.

Worked example 2.61: The Riesz kernel and its constant

For $-\frac{1}{2} < \lambda < 0$,

$$\begin{aligned} k^0_\lambda(x) &\overset{(1)}{=} \frac{1}{\pi}\int_0^\infty \xi^{2\lambda}\cos(|x|\,\xi)\,\mathrm{d}\xi \overset{(2)}{=} \frac{1}{\pi\,|x|^{1+2\lambda}}\int_0^\infty t^{2\lambda}\cos t\,\mathrm{d}t \\ &\overset{(3)}{=} \frac{\Gamma(1+2\lambda)\cos\!\left(\frac{\pi}{2}(1+2\lambda)\right)}{\pi\;|x|^{1+2\lambda}} \overset{(4)}{=} \frac{1}{2\cos(\pi\lambda)\,\Gamma(-2\lambda)}\cdot\frac{1}{|x|^{1+2\lambda}}. \end{aligned} \tag{2.137}$$

(1) As in Example 2.60: even symbol, cosine part only.

(2) Substitute $t = |x|\,\xi$; this is where the homogeneity of the symbol turns into a pure power of $|x|$.
(3) The Euler integral $\int_0^\infty t^{s-1}\cos t\,\mathrm{d}t = \Gamma(s)\cos(\pi s/2)$, valid for $0 < s < 1$; here $s = 1 + 2\lambda \in (0,1)$.
(4) $\cos(\frac{\pi}{2} + \pi\lambda) = -\sin\pi\lambda$, then the reflection formula $\Gamma(1+2\lambda)\Gamma(-2\lambda) = \pi/\sin(-2\pi\lambda)$ and $\sin 2\pi\lambda = 2\sin\pi\lambda\cos\pi\lambda$.

Nash's (2.61) has $\sqrt{2}\cos(\lambda\pi)\Gamma(-2\lambda)$ in the denominator instead of our $2\cos(\lambda\pi)\Gamma(-2\lambda)$: an overall factor of $\sqrt{2}$ from his normalisation of the Fourier transform. The $|x|$-dependence, which is what matters, agrees. It is also the $a \to 0$ limit of the short-distance form of Example 2.60: one can check with the duplication formula that $2^{2\lambda}\Gamma(\lambda+\frac{1}{2})/(\sqrt{\pi}\,\Gamma(-\lambda)) = 1/(2\cos\pi\lambda\,\Gamma(-2\lambda))$. (At $\lambda = -\frac{1}{4}$ both are $1/\sqrt{2\pi} \approx 0.399$.)

So fractional integration of order $-2\lambda =: \alpha \in (0,1)$ is

$$\Delta^{-\alpha/2}u(x) = \frac{1}{2\cos(\pi\alpha/2)\,\Gamma(\alpha)}\int_{\mathbb{R}} \frac{u(y)}{|x-y|^{1-\alpha}}\,\mathrm{d}y,$$

the symmetric (Riesz) cousin of the Riemann–Liouville fractional integral

$$I^{\alpha}u(x) = \frac{1}{\Gamma(\alpha)}\int_{-\infty}^{x}(x-y)^{\alpha-1}u(y)\,\mathrm{d}y.$$

As $\alpha \to 1$ the kernel flattens to a constant and the prefactor blows up: fractional integration of order one is not quite "$\int$" on the whole line, because $1/|\xi|$ is not locally integrable at $\xi = 0$ (an infrared problem, not an ultraviolet one).

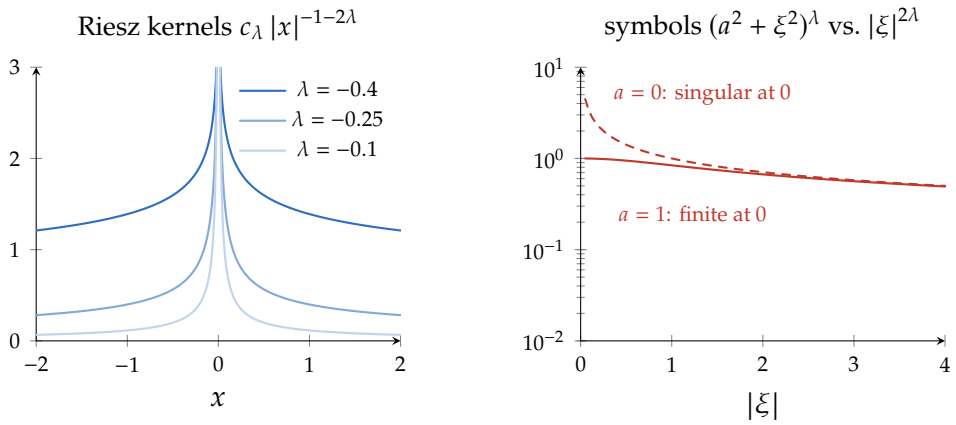


Figure 2.32: Left: kernels of fractional integration Δ^λ on $\mathbb{R}$ for three values of λ. The closer the order 2λ is to 0, the more sharply the kernel concentrates at $x = 0$ (as $\lambda \to 0$ it becomes δ, the kernel of the identity). All three are integrable at 0. Right: the Bessel symbol with $a = 1$ (solid) and the Riesz symbol $a = 0$ (dashed), $\lambda = -\frac{1}{4}$. They agree at large $|\xi|$ (same order, same principal symbol) and differ only near $\xi = 0$, which affects the long-distance tail of the kernel, not its singularity.

The group law $P_\lambda P_\mu = P_{\lambda+\mu}$. Nash notes that "$P$ does possess one of the essential properties of a derivative": the powers compose additively. In momentum space it is one line,

$$\widehat{P_\lambda P_\mu u}(\xi) \overset{(1)}{=} \sigma_\lambda(\xi)\sigma_\mu(\xi)\hat{u}(\xi) \overset{(2)}{=} (a^2+\xi^2)^{\lambda+\mu}\hat{u}(\xi) = \widehat{P_{\lambda+\mu}u}(\xi),$$

(1) Fourier multipliers compose by multiplying symbols (Section 2.4.8).

(2) $x^\lambda x^\mu = x^{\lambda+\mu}$ for $x > 0$.

for any $a \geq 0$. In position space it is a non-trivial statement about convolutions of kernels, and the structure is visible by power counting.

Worked example 2.62: The group law for Riesz kernels, by scaling

With $a = 0$, $P_\lambda P_\mu$ has kernel $k^0_\lambda * k^0_\mu$. Write $\alpha = 1+2\lambda$, $\beta = 1+2\mu$ (both in $(0,1)$) and assume $\alpha+\beta > 1$ so the integral converges at infinity. For $x \neq 0$,

$$\int_{\mathbb{R}} \frac{\mathrm{d}y}{|x-y|^\alpha\,|y|^\beta} \overset{(1)}{=} \int_{\mathbb{R}} \frac{|x|\,\mathrm{d}t}{|x|^\alpha\,|\hat{x}-t|^\alpha\;|x|^\beta\,|t|^\beta} \overset{(2)}{=} \frac{C_{\alpha\beta}}{|x|^{\alpha+\beta-1}}, \qquad C_{\alpha\beta} = \int_{\mathbb{R}} \frac{\mathrm{d}t}{|1-t|^\alpha\,|t|^\beta},$$

(1) Substitute $y = |x|\,t$, and write $x = |x|\,\hat{x}$ with $\hat{x} = \pm 1$.
(2) The remaining integral does not depend on x (by $t \to -t$ it is the same for $\hat{x} = \pm 1$), and converges: near $t = 0$ and $t = 1$ because $\alpha, \beta < 1$, at infinity because $\alpha+\beta > 1$.

Now $\alpha+\beta-1 = 1+2(\lambda+\mu)$: the convolution is proportional to $|x|^{-1-2(\lambda+\mu)}$, the Riesz kernel of $P_{\lambda+\mu}$. Scaling fixes the power; the constant $C_{\alpha\beta}$ (a combination of Beta functions) is guaranteed by the momentum-space argument to come out right. This is the position-space shadow of $|\xi|^{2\lambda}\,|\xi|^{2\mu} = |\xi|^{2(\lambda+\mu)}$.

When the integrals diverge (for instance $\lambda+\mu \geq 0$, where the result has non-negative order), the composition still makes sense as a statement about distributions; Nash refers to Gel'fand–Shilov for this. Our viewpoint: *define* P_λ by its symbol for every $\lambda \in \mathbb{C}$, and the group law holds identically. The kernel is a secondary object.

2.4.12 Nash's gallery III: a logarithm, and an order that is not attained

Worked example 2.63: The symbol $(a^2+\xi^2)^\lambda \ln(a^2+\xi^2)$

Nash's (2.62) is the operator

$$Pu(x) = \int e^{ix\xi}(a^2+\xi^2)^\lambda \ln(a^2+\xi^2)\,\hat{u}(\xi)\,\frac{\mathrm{d}\xi}{2\pi}, \qquad \sigma(\xi) = (a^2+\xi^2)^\lambda \ln(a^2+\xi^2).$$

It is in $S^{2\lambda+\varepsilon}$ for every $\varepsilon > 0$. The logarithm grows more slowly than any power: $\ln(a^2+\xi^2) \leq C_\varepsilon\langle\xi\rangle^\varepsilon$ for all ξ, because $\ln(a^2+\xi^2)/\langle\xi\rangle^\varepsilon \to 0$ as $|\xi| \to \infty$ (l'Hôpital) and it is continuous. So

$$|\sigma(\xi)| \overset{(1)}{\leq} C\langle\xi\rangle^{2\lambda}\cdot C_\varepsilon\langle\xi\rangle^\varepsilon = C'\langle\xi\rangle^{2\lambda+\varepsilon},$$
$$|\partial_\xi\sigma| \overset{(2)}{=} \left|2\lambda\xi(a^2+\xi^2)^{\lambda-1}\ln(a^2+\xi^2) + 2\xi(a^2+\xi^2)^{\lambda-1}\right| \overset{(3)}{\leq} C''\langle\xi\rangle^{2\lambda+\varepsilon-1}.$$

(1) $(a^2+\xi^2)^\lambda$ is comparable to $\langle\xi\rangle^{2\lambda}$ (Example 2.50(c)).
(2) Product rule: the derivative of $\ln(a^2+\xi^2)$ is $2\xi/(a^2+\xi^2)$.
(3) Each term is $|\xi|\,\langle\xi\rangle^{2\lambda-2}$ times a logarithm or a constant; bound the logarithm by $C_\varepsilon\langle\xi\rangle^\varepsilon$.

Higher derivatives work the same way: each ∂_ξ either hits the power (costing one order) or the logarithm (turning it into a rational function of order -1).
It is not in $S^{2\lambda}$. That would need $|\sigma(\xi)| \leq C\langle\xi\rangle^{2\lambda}$, i.e. $\ln(a^2+\xi^2) \leq C'$ for all ξ, which is false. So the set of admissible orders is the open half-line $(2\lambda, \infty)$ and its infimum 2λ is not attained. This is the example promised in Section 2.4.4.

It has no principal symbol in Nash's sense. With $m = 2\lambda$, $\sigma(\mu\xi)/\mu^{2\lambda} \approx |\xi|^{2\lambda}\,(\ln\mu^2 + \ln\xi^2) \to \infty$; with $m = 2\lambda + \varepsilon$ the limit is 0. Neither is an invertible homogeneous function, so P is *not* elliptic of any order, even though its symbol never vanishes for large ξ. Operators like this are not "classical" (their symbols have no expansion in homogeneous terms).

The kernel by differentiating in λ. Nash's trick: $\partial_\lambda (a^2+\xi^2)^\lambda = (a^2+\xi^2)^\lambda \ln(a^2+\xi^2)$. So the kernel G of P is ∂_λ of the Bessel kernel (2.134). Write $z = |x|$, $\nu = \lambda + \frac{1}{2}$, and $k_\lambda(z) = A(\lambda)\,(z/2a)^{-\nu} K_\nu(az)$ with $A = 1/(\sqrt{\pi}\,\Gamma(-\lambda))$. Then

$$G(z) = \partial_\lambda k_\lambda(z) \overset{(1)}{=} A'(\lambda)\Big(\frac{z}{2a}\Big)^{-\nu} K_\nu + A\,\partial_\lambda\Big(\frac{z}{2a}\Big)^{-\nu} K_\nu + A\Big(\frac{z}{2a}\Big)^{-\nu}\partial_\nu K_\nu$$
$$\overset{(2)}{=} \Big[\psi(-\lambda) + \ln(2a) - \ln z\Big] k_\lambda(z) + A\Big(\frac{z}{2a}\Big)^{-\nu}\frac{\partial K_\nu}{\partial \nu}(az), \tag{2.138}$$

(1) Product rule; $\mathrm{d}\nu/\mathrm{d}\lambda = 1$.

(2) $A'(\lambda) = \psi(-\lambda)A(\lambda)$ where $\psi = \Gamma'/\Gamma$ is the digamma function (differentiate $1/\Gamma(-\lambda)$), and $\partial_\lambda (z/2a)^{-\nu} = -\ln(z/2a)\,(z/2a)^{-\nu}$.

Since $\psi(-\lambda) = -\frac{\mathrm{d}}{\mathrm{d}\lambda}\ln\Gamma(-\lambda)$ this is Nash's "formidable looking" (2.63), term by term: the braces $\{\ln(2a) - \frac{\mathrm{d}}{\mathrm{d}\lambda}\ln\Gamma(-\lambda)\}$ and $\{\frac{\mathrm{d}}{\mathrm{d}\lambda}K_{\frac{1}{2}+\lambda} - K_{\frac{1}{2}+\lambda}\ln z\}$. The $-\ln z$ factor is the kernel-side fingerprint of the logarithm in the symbol: it makes the singularity at $z = 0$ slightly worse than $z^{-1-2\lambda}$, but better than $z^{-1-2\lambda-\varepsilon}$ for any $\varepsilon > 0$, in perfect parallel with the order.

2.4.13 Nash's gallery IV: smoothing operators

Nash's next example (2.64) is the Gaussian integral operator

$$Pu(x) = \int_{\mathbb{R}} e^{-(x-y)^2/2}\,u(y)\,\mathrm{d}y.$$

In Example 2.49 we computed that its symbol is (in our conventions) $\sigma(\xi) = \sqrt{2\pi}\,e^{-\xi^2/2}$. Nash's key observation (2.65) is that this symbol is dominated by *every* power of $\langle\xi\rangle$.

Worked example 2.64: The Gaussian symbol is in S^m for every m

We need, for each real m and each $k \ge 0$, a constant with $\left|\partial_\xi^k e^{-\xi^2/2}\right| \le C_{m,k}\langle\xi\rangle^{m-k}$. First $k = 0$, $m = -M < 0$: we want $\sup_\xi e^{-\xi^2/2}(1+\xi^2)^{M/2} < \infty$. Put $s = \xi^2 \ge 0$ and $g(s) = e^{-s/2}(1+s)^{M/2}$. Then

$$g'(s) \overset{(1)}{=} e^{-s/2}(1+s)^{M/2-1}\Big(-\tfrac{1}{2}(1+s) + \tfrac{M}{2}\Big) \overset{(2)}{=} 0 \iff s = M-1,$$

(1) Product rule, then factor out $e^{-s/2}(1+s)^{M/2-1}$.

(2) The bracket vanishes at $1 + s = M$.

so for $M \ge 1$ the maximum is $g(M-1) = e^{-(M-1)/2}M^{M/2}$, a finite number (for $M < 1$ the maximum is $g(0) = 1$). For example with $M = 2$ we get $C = 2e^{-1/2} \approx 1.21$; with $M = 10$, $C = 10^5 e^{-4.5} \approx 1.1 \times 10^3$. The constants grow with M, but each is finite, and that is all the definition asks. For $k \ge 1$: $\partial_\xi^k e^{-\xi^2/2} = (-1)^k \mathrm{He}_k(\xi) e^{-\xi^2/2}$, where He_k is the (probabilists') Hermite polynomial of degree k (for instance $\mathrm{He}_1 = \xi$, $\mathrm{He}_2 = \xi^2 - 1$). A polynomial times a Gaussian is again dominated by every power, by the same argument. Hence $\sigma \in S^m$ for all m: $\sigma \in S^{-\infty}$.

Working definition 2.28 (smoothing operator)

An operator is **smoothing** (of order $-\infty$) if it maps every distribution [a generalised function such as δ or its derivatives], however rough, to a C^∞ function. *Tests:* (a) its symbol lies in $S^{-\infty}$ [it decays faster than every power $\langle\xi\rangle^{-N}$, like the Gaussian]; equivalently (b) its Schwartz kernel [the function K with $Pu(x) = \int K(x,y)\,u(y)\,dy$] $K(x,y)$ is a C^∞ function of both variables. Then $D_x^\beta P u(x) = \int D_x^\beta K(x,y)\,u(y)\,dy$: every derivative lands on the smooth kernel, never on u.

Nash's sufficient condition for the kernel-side test is: $K(x,y)$ smooth in x and compactly supported in y. As he notes, it is not necessary (the Gaussian is not compactly supported), and on a compact manifold every smooth kernel satisfies it automatically. The name is literal: P does to u what infinitely many integrations would do.

Worked example 2.65: Smoothing operators on the circle

On S^1 every operator that is diagonal in Fourier modes, $Pe^{in\theta} = p_n e^{in\theta}$, has kernel

$$K(\theta,\theta') \overset{(1)}{=} \frac{1}{2\pi}\sum_{n\in\mathbb{Z}} p_n\, e^{in(\theta-\theta')},$$

(1) Write $u = \sum_n c_n e^{in\theta}$ with $c_n = \frac{1}{2\pi}\int u(\theta')e^{-in\theta'}d\theta'$, apply P, and exchange sum and integral.

The *symbol* is the sequence p_n. If p_n decays faster than every power of n, the series for K can be differentiated term by term any number of times ($\sum |n|^k\,|p_n| < \infty$ for all k), so K is C^∞: P is smoothing. Three instances:

- $p_n = e^{-tn^2}$: the heat operator $e^{-t\Delta}$ on the circle; K is the periodic Gaussian (a theta function).
- $p_n = 1$ for $n = 0$, else 0: the projection Π_0 onto constants, $K = 1/2\pi$ (Example 2.59). Every finite-rank operator with smooth range and smooth "co-range" is smoothing.
- $p_n = r^{|n|}$, $0 < r < 1$: the Poisson kernel $K = \frac{1}{2\pi}\frac{1-r^2}{1-2r\cos(\theta-\theta')+r^2}$, which maps boundary data on the circle to the value of the harmonic function at radius r.

By contrast $p_n = |n|$ ($|D|$, order 1), $p_n = (1+n^2)^{-1}$ (order -2) and $p_n = \operatorname{sgn} n$ (order 0) decay at most like a power, and their kernels are singular on the diagonal $\theta = \theta'$.

Smoothing operators are the "zero" of the theory: ΨDOs are usually considered modulo them. On a compact manifold, they are also *compact* operators on every Sobolev space (Section 2.5.8), and that is what turns "invertible modulo smoothing" into "Fredholm".

2.4.14 Nash's gallery V: complex powers of an elliptic operator

The last example is the most important for later chapters. Let $a > 0$ and $z \in \mathbb{C}$, and define

$$P = (a^2+\Delta)^z, \qquad \sigma(\xi) = (a^2+|\xi|^2)^z := e^{z\ln(a^2+|\xi|^2)}, \tag{2.139}$$

which is unambiguous because $a^2 + |\xi|^2 > 0$ (we use the real logarithm).

Worked example 2.66: Order and ellipticity of $(a^2+\Delta)^z$

Write $z = u + iv$. The size of the symbol is

$$|\sigma(\xi)| \overset{(1)}{=} \left|e^{(u+iv)\ln(a^2+|\xi|^2)}\right| \overset{(2)}{=} e^{u\ln(a^2+|\xi|^2)} = (a^2+|\xi|^2)^u \overset{(3)}{\le} C\langle\xi\rangle^{2u}.$$

(1) Definition (2.139).
(2) $|e^w| = e^{\operatorname{Re} w}$; the imaginary part $v\ln(a^2+|\xi|^2)$ is a pure phase.
(3) Comparability of $a^2+|\xi|^2$ with $\langle\xi\rangle^2$.

This is Nash's (2.68): $\left|(a^2+p^2)^{iv}\right| = 1$. For the derivatives,

$$\left|\partial_{\xi_j}\sigma\right| \overset{(1)}{=} |z|\;2\left|\xi_j\right|\;(a^2+|\xi|^2)^{u-1} \overset{(2)}{\le} C\,|z|\,\langle\xi\rangle^{2u-1},$$

(1) Chain rule: $\partial_{\xi_j} e^{zL} = z\,(\partial_{\xi_j}L)\,e^{zL}$ with $L = \ln(a^2+|\xi|^2)$ and $\partial_{\xi_j}L = 2\xi_j/(a^2+|\xi|^2)$; take absolute values as in the first chain.
(2) $\left|\xi_j\right| \le \langle\xi\rangle$ and comparability.

and so on for higher derivatives. So $\sigma \in S^{2\operatorname{Re} z}$: *the order is* $2\operatorname{Re} z$; the imaginary part of the exponent does not affect growth. The principal symbol is $|\xi|^{2z}$, homogeneous of the complex degree $2z$, and $\left||\xi|^{2z}\right| = |\xi|^{2u} \neq 0$ for $\xi \neq 0$. So P is elliptic of order $2\operatorname{Re} z$.

The imaginary powers $(a^2+\Delta)^{iv}$ are unitary on L^2 (the symbol has modulus one), of order 0, and elliptic. They are not smoothing and not the identity: they oscillate faster and faster at high momentum, with a phase $v\ln(a^2+|\xi|^2)$. The family $z \mapsto (a^2+\Delta)^z$ is a one-parameter group, $(a^2+\Delta)^z(a^2+\Delta)^w = (a^2+\Delta)^{z+w}$, analytic in z.

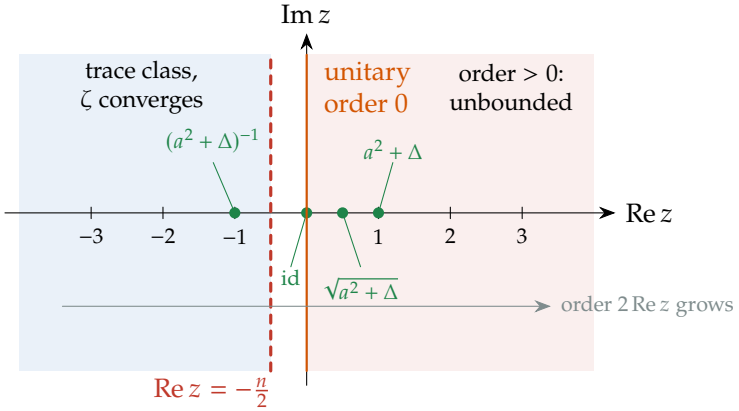


Figure 2.33: The complex powers $(a^2+\Delta)^z$ as a map from the z-plane to elliptic ΨDOs. Only $\operatorname{Re} z$ controls the order (and hence how much the operator smooths or roughens). Integer points are the familiar operators; the imaginary axis (orange) consists of unitary operators of order 0, and to its right (shaded red) the order is positive and the operator is unbounded on L^2. To the left of $\operatorname{Re} z = -n/2$ (shaded blue, drawn for $n = 1$) the operator is trace class on a compact n-manifold, and its trace defines the ζ-function, which is then continued analytically to the whole plane.

Worked example 2.67: Complex powers on the circle, the ζ-function, and a determinant

On S^1 the operator $D_\theta^2 = -\mathrm{d}^2/\mathrm{d}\theta^2$ has eigenvalues n^2 ($n \in \mathbb{Z}$) with eigenfunctions $e^{in\theta}$. Removing the zero mode, its complex power acts by $(D_\theta^2)^{-s}e^{in\theta} = n^{-2s}e^{in\theta}$ ($n \neq 0$). The

trace (sum of the diagonal entries) is

$$\zeta_{D^2}(s) := \operatorname{Tr}'(D_\theta^2)^{-s} \overset{(1)}{=} \sum_{n\neq 0} \frac{1}{|n|^{2s}} \overset{(2)}{=} 2\sum_{n=1}^{\infty} \frac{1}{n^{2s}} = 2\,\zeta_R(2s),$$

(1) The operator is diagonal in the basis $e^{in\theta}$; the prime means the zero mode is omitted.
(2) n and $-n$ contribute equally.

The series converges for $\operatorname{Re} s > \frac{1}{2}$, i.e. when the order $-2\operatorname{Re} s$ is below $-1 = -\dim S^1$: a ΨDO is trace class when its order is less than minus the dimension. Riemann's ζ_R continues analytically to all $s \neq 1$, with $\zeta_R(0) = -\frac{1}{2}$ and $\zeta_R'(0) = -\frac{1}{2}\ln 2\pi$. The **ζ-regularised determinant** is defined by $\ln\det' A = -\zeta_A'(0)$ (for a finite matrix, $-\frac{\mathrm{d}}{\mathrm{d}s}\sum \lambda_j^{-s}\big|_{s=0} = \sum \ln\lambda_j$), so

$$\ln\det' D_\theta^2 \overset{(1)}{=} -\frac{\mathrm{d}}{\mathrm{d}s}\,2\zeta_R(2s)\Big|_{s=0} = -4\zeta_R'(0) \overset{(2)}{=} 2\ln(2\pi), \qquad \det' D_\theta^2 = (2\pi)^2.$$

(1) Chain rule: $\frac{\mathrm{d}}{\mathrm{d}s}\zeta_R(2s) = 2\zeta_R'(2s)$.
(2) $\zeta_R'(0) = -\frac{1}{2}\ln(2\pi)$.

The formally infinite product $\prod_{n\neq 0} n^2$ has been given the finite value $4\pi^2$. This is how one-loop determinants $\det(-\nabla^2 + V)$ in the Euclidean path integral are defined, and it is why Nash cares about complex powers of elliptic operators. Note also $\zeta_{D^2}(0) = 2\zeta_R(0) = -1 = -\dim\operatorname{Ker} D_\theta^2$: a first hint that the value of ζ at 0 knows about zero modes (Chapter 4).

2.4.15 Locality and pseudo-locality

Differential operators are *local*. Pseudo-differential operators, as every example above shows, are not. What survives is a weaker and very useful property: they never create singularities where there were none.

Working definition 2.29 (support and singular support)

The **support** $\operatorname{supp} u$ of a function (or distribution) u is the closure [the set together with its limit points, so a support is always closed] of the set where $u \neq 0$; equivalently the complement of the largest open set on which u vanishes. The **singular support** $\operatorname{sing\,supp} u$ is the complement of the largest open set on which u is C^∞ (Nash's footnote 7). *Test:* $x_0 \notin \operatorname{sing\,supp} u$ iff u agrees with a smooth function on some neighbourhood of x_0.

For example: $u(x) = |x|$ on $\mathbb{R}$ has $\operatorname{supp} u = \mathbb{R}$, $\operatorname{sing\,supp} u = \{0\}$; δ has $\operatorname{supp}\delta = \operatorname{sing\,supp}\delta = \{0\}$; the box function $\mathbf{1}_{[-1,1]}$ has support $[-1,1]$ and singular support $\{-1,1\}$; any smooth u has empty singular support.

Working definition 2.30 (local operator)

P is **local** if $\operatorname{supp} Pu \subset \operatorname{supp} u$ for all u: if u vanishes on an open set, so does Pu. *Test:* feed P a bump function [a smooth function that is nonzero only on a small ball] and see whether the output is supported in the same place.

A differential operator is local: if $u = 0$ on an open set W, all its derivatives vanish on W, so $\sum a_\alpha D^\alpha u = 0$ on W. Nash's footnote 6 points out that the converse holds too, which is **Peetre's theorem**: a linear operator on smooth functions that is local (and continuous) is, locally, a differential operator of finite order. So "local" can be taken as the *definition* of "differential".

A genuinely pseudo-differential operator is never local: in kernel terms, a local operator has $K(x,y)$ supported on the diagonal $x=y$, which forces K to be a finite sum of derivatives of $\delta(x-y)$, and then the symbol is a polynomial.

Working definition 2.31 (pseudo-local operator)

P is **pseudo-local** if

$$\operatorname{sing\,supp} Pu \subset \operatorname{sing\,supp} u \qquad \text{for all } u, \tag{2.140}$$

where sing supp is the singular support [the points near which the function is not smooth]. *Picture:* P may smear u out over all of space, but it cannot create a kink, a jump or a spike where u was smooth.

Key idea: ΨDOs are pseudo-local

Every pseudo-differential operator satisfies (2.140) (Nash's (2.69)). The mechanism: *the kernel of a ΨDO is smooth off the diagonal*, because smoothness of the symbol in ξ becomes decay of the kernel in $x-y$.

Here is the calculation behind the mechanism. For $\sigma \in S^m$ and any multi-index α, with K from (2.120),

$$\begin{aligned}(x-y)^\alpha K(x,y) &\overset{(1)}{=} \int \left(D_\xi^\alpha e^{i(x-y)\cdot\xi}\right)\sigma(x,\xi)\,\frac{\mathrm{d}^n\xi}{(2\pi)^n}\\ &\overset{(2)}{=} \int e^{i(x-y)\cdot\xi}\left((-D_\xi)^\alpha\sigma\right)(x,\xi)\,\frac{\mathrm{d}^n\xi}{(2\pi)^n}.\end{aligned} \tag{2.141}$$

(1) $D_{\xi_j} e^{i(x-y)\cdot\xi} = -i\cdot i(x-y)^j e^{i(x-y)\cdot\xi} = (x-y)^j e^{i(x-y)\cdot\xi}$, applied α_j times in each variable.
(2) Integrate by parts in ξ, $|\alpha|$ times (each moves one D_{ξ_j} onto σ with a minus sign; boundary terms are dropped, rigorously in the sense of distributions).

Now $(-D_\xi)^\alpha\sigma \in S^{m-|\alpha|}$. Once $m-|\alpha|<-n$ the last integrand is bounded by $C\langle\xi\rangle^{m-|\alpha|}$, which is integrable over $\mathbb{R}^n$, so $(x-y)^\alpha K$ is a continuous function. Taking $|\alpha|$ larger still, we may differentiate under the integral sign j times as long as $m-|\alpha|+j<-n$. Hence $(x-y)^\alpha K$ is C^j for $|\alpha|$ large, and away from $x=y$ we can divide by $(x-y)^\alpha$ (choosing α along a coordinate where $x^i\neq y^i$). So K is C^∞ off the diagonal. The more smoothly σ depends on ξ, the faster K decays: that is the content of the last line of the symbol–kernel dictionary.

From smooth kernels to pseudo-locality. Let u be smooth on an open set W and $x_0\in W$. Pick a cutoff ϕ with $\phi=1$ near x_0 and $\operatorname{supp}\phi\subset W$, and split $u=\phi u+(1-\phi)u$.

- ϕu is smooth with compact support, so $\widehat{\phi u}$ decays faster than any power, and $P(\phi u)=\int e^{ix\cdot\xi}\sigma\,\widehat{\phi u}\,\frac{\mathrm{d}^n\xi}{(2\pi)^n}$ can be differentiated under the integral any number of times: it is smooth everywhere.
- $(1-\phi)u$ vanishes near x_0, say for $|y-x_0|<2\delta$. For x with $|x-x_0|<\delta$, $P((1-\phi)u)(x)=\int K(x,y)(1-\phi(y))u(y)\,\mathrm{d}y$ only uses K at $|x-y|>\delta$, where it is smooth: so this piece is smooth near x_0.

So Pu is smooth near x_0: $x_0\notin\operatorname{sing\,supp} Pu$.

Worked example 2.68: $|D|$ applied to a box: not local, but pseudo-local

Let $u=\mathbf{1}_{[-1,1]}$ on $\mathbb{R}$. We use $|D|=H\circ\frac{\mathrm{d}}{\mathrm{d}x}$, where H is the **Hilbert transform** $Hf(x)=$

$\frac{1}{\pi}\,\mathrm{p.v.}\int \frac{f(y)}{x-y}\,\mathrm{d}y$, whose symbol is $-i\,\mathrm{sgn}\,\xi$. (Check: $(-i\,\mathrm{sgn}\,\xi)(i\xi) = |\xi|$.) Then

$$|D|\,u \overset{(1)}{=} H\big(\delta(\cdot+1)-\delta(\cdot-1)\big) \overset{(2)}{=} \frac{1}{\pi}\Big(\frac{1}{x+1}-\frac{1}{x-1}\Big) \overset{(3)}{=} \frac{2}{\pi(1-x^2)}.$$

(1) The box jumps up by 1 at $x=-1$ and down by 1 at $x=+1$, so $u' = \delta(x+1)-\delta(x-1)$.
(2) $H\delta(\cdot - c) = \frac{1}{\pi(x-c)}$, straight from the kernel.
(3) Common denominator: $(x-1)-(x+1) = -2$ over $(x+1)(x-1) = x^2-1$.

(The result is valid for $x \neq \pm 1$.) Read off the two properties: $\mathrm{supp}\,|D|\,u = \mathbb{R}$, much bigger than $\mathrm{supp}\,u = [-1,1]$, so $|D|$ is not local; but $|D|\,u$ is smooth (indeed real-analytic) except at $x=\pm1$, so $\mathrm{sing\,supp}\,|D|\,u = \{-1,1\} = \mathrm{sing\,supp}\,u$. The singularities stayed put. For comparison, $Hu = \frac{1}{\pi}\ln\left|\frac{x+1}{x-1}\right|$: again smooth off $\{\pm1\}$, with a logarithmic singularity (order 0 keeps the jump's strength roughly the same) instead of the pole of the order-1 operator.

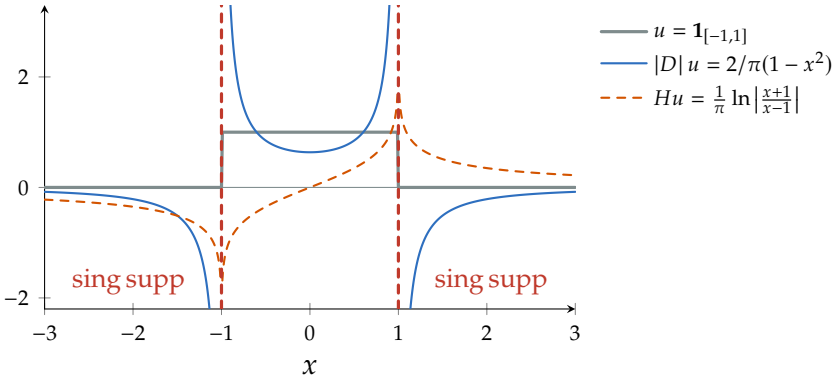


Figure 2.34: Pseudo-locality in action. The box u (grey) is supported in $[-1,1]$ and singular only at ±1 (red dashed lines). After applying the order-1 operator $|D|$ (blue) or the order-0 Hilbert transform H (orange), the result is nonzero on the whole line (the operators are not local), but it is still smooth everywhere except at ±1. Higher order makes the singularity stronger (a pole instead of a logarithm); it never moves it.

When singular supports coincide: hypo-ellipticity again. Pseudo-locality is only an inclusion. It allows P to *remove* singularities. A smoothing operator removes all of them. Nash's (2.70) considers operators that can remove none:

$$\mathrm{sing\,supp}\,Pu = \mathrm{sing\,supp}\,u \qquad \text{for all } u. \tag{2.142}$$

Then "Pu smooth" means $\mathrm{sing\,supp}\,Pu = \varnothing$, hence $\mathrm{sing\,supp}\,u = \varnothing$, hence u smooth: P is hypo-elliptic, in the sense of the previous sections. Conversely, if P is hypo-elliptic on every open set, then u smooth where Pu is smooth gives $\mathrm{sing\,supp}\,u \subset \mathrm{sing\,supp}\,Pu$, and pseudo-locality gives the other inclusion. So (2.142) is a characterisation of hypo-ellipticity (for pseudo-local operators), and elliptic ΨDOs satisfy it by the parametrix argument of Section 2.4.9: $u = Q(Pu) + R_2u$, with Q pseudo-local and R_2 smoothing, so $\mathrm{sing\,supp}\,u \subset \mathrm{sing\,supp}\,Pu$.

Reading Nash (pp. 41–47, (2.54)–(2.70))

Nash's §4 is short and example-driven. The dictionary to these notes: (2.54) is (2.118); (2.55) is (2.122); (2.56) is (2.124); the two definitions of ellipticity are in Section 2.4.7; (2.57)–(2.58) are (2.130) and (2.121); (2.59)–(2.60) are Example 2.60; (2.61) is Example 2.61; (2.62)–(2.63) are Example 2.63 and (2.138); (2.64)–(2.65) are Examples 2.49 and 2.64; (2.66)–

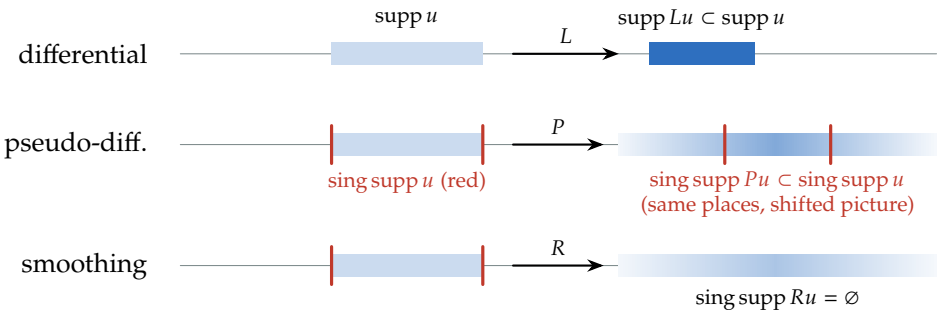


Figure 2.35: Three levels of "respecting" a function. A differential operator keeps supports (local). A ΨDO spreads the support over everything, but keeps singularities where they were (pseudo-local): in the middle row the red marks of the output sit over the same points $x = \pm 1$ as those of the input, drawn to the right only for legibility. A smoothing operator spreads everything and erases all singularities. Elliptic operators are the ones that erase *no* singularity: sing supp $Pu =$ sing supp u.

(2.68) are Example 2.66; (2.69)–(2.70) are (2.140) and (2.142). Nash's constants in (2.58), (2.61) and (2.64) differ from ours by factors of $\sqrt{2\pi}$ or $\sqrt{2}$, coming from his mixed Fourier normalisations; orders, principal symbols and ellipticity are unaffected. What Nash does *not* do, and we added, is the symbol calculus (composition, adjoints, commutators) and the parametrix; these are the tools that make the Fredholm theory of §5 work, and they are what his remark "elliptic operators are invertible modulo smoothing operators" (p. 55) rests on.

Recap: Pseudo-differential operators

- A ΨDO is "Fourier transform, multiply by $\sigma(x,\xi)$, transform back". Differential operators are the case of polynomial symbols.
- $\sigma \in S^m$: each ξ-derivative gains one power of decay; x-derivatives cost nothing. Orders can be any real number, or $-\infty$ (smoothing).
- Principal symbol $\sigma_m = \lim \sigma(x,\mu\xi)/\mu^m$: positively homogeneous, lives on T^*M. Elliptic: $\sigma_m(x,\xi)$ invertible for $\xi \neq 0$.
- Symbols multiply to leading order; commutators give Poisson brackets; elliptic operators have parametrices (inverses modulo smoothing).
- Examples: resolvents and Yukawa kernels, $\sqrt{\Delta + m^2}$, Bessel and Riesz potentials (fractional integration), a logarithmic symbol whose order is not attained, the Gaussian (smoothing), complex powers (order $2\operatorname{Re} z$, the route to ζ-functions and determinants).
- ΨDOs are not local but are pseudo-local: sing supp $Pu \subset$ sing supp u. Elliptic ones are exactly those that also keep every singularity: hypo-ellipticity.

2.5 Pseudo-differential operators and Sobolev spaces

We now have a large class of operators. The next question is: *on which spaces do they act, and how do we measure what they do?* Nash's answer is the Sobolev spaces, and the reasons are worth stating up front.

- A ΨDO of order m "costs m derivatives". We want spaces labelled by a number s that counts derivatives, so that P maps the space with label s to the one with label $s-m$, continuously. C^k spaces almost do this but fail in small, annoying ways; Sobolev spaces do it exactly, for every real s.
- We want Hilbert spaces, so that adjoints, orthogonal complements and spectral theory are available. C^k spaces are not Hilbert spaces; H^s spaces are.
- We want *compactness*: a bounded set of functions with one extra derivative under control should have convergent subsequences. That is what will make elliptic operators Fredholm, with finite-dimensional kernels, and give a finite index.

Physicists meet Sobolev norms constantly without the name. The energy of a static scalar field, $\frac{1}{2}\int(|\nabla\phi|^2 + m^2\phi^2)\,\mathrm{d}^n x$, is (for $m=1$) half the square of the H^1 norm. "Finite-energy configurations" means "$\phi \in H^1$". The δ-function source of a point charge lives in a Sobolev space of negative order. And the paths that dominate a Feynman path integral are continuous but not differentiable: they live in H^s for every $s < \frac{1}{2}$ and no better.

2.5.1 Counting derivatives in L^2 and L^p

The idea is the simplest one possible: a function is "k times differentiable in the L^2 sense" if it and all its derivatives up to order k are square integrable. For $f:\mathbb{R}^n \to \mathbb{C}$ this means (Nash's (2.71))

$$\sum_{|\alpha|\le k} \int_{\mathbb{R}^n} \left|\partial^\alpha f\right|^2 \mathrm{d}^n x < \infty, \tag{2.143}$$

and replacing the square by a p-th power gives the L^p version (2.72).

Working definition 2.32 (Sobolev space L^p_k, integer k; Nash (2.73)–(2.74))

Let $\Omega \subset \mathbb{R}^n$ be open, $k \in \{0,1,2,\dots\}$ and $1 \le p < \infty$. The **Sobolev space** $L^p_k(\Omega)$ consists of the functions f on Ω for which

$$\|f\|_{k,p} = \sum_{|\alpha|\le k} \left\{ \int_\Omega |\partial^\alpha f|^p \; \mathrm{d}^n x \right\}^{1/p} < \infty, \tag{2.144}$$

where the derivatives are taken in the *weak* (distributional) sense [moved onto smooth test functions by integrating by parts, so a kink like $|x|$ still has a first derivative] defined below. For $p=2$ it is a Hilbert space [its norm can be taken to come from an inner product, and every Cauchy sequence converges] and is written $H^k(\Omega) = L^2_k(\Omega)$. *Test:* compute all weak derivatives of order $\le k$ and check that each has finite L^p norm.

Two remarks on the norm. First, Nash adds the L^p norms of the derivatives. For $p=2$ it is more convenient to add their squares, $\|f\|^2_{H^k} = \sum_{|\alpha|\le k}\|\partial^\alpha f\|^2_{L^2}$, because then the norm comes from an inner product. The two choices are **equivalent norms**: for N non-negative numbers a_i, $\sum a_i^2 \le (\sum a_i)^2 \le N\sum a_i^2$ (the right inequality is Cauchy–Schwarz). Equivalent norms define the same space and the same notion of convergence, so we switch freely. Second, for $k=0$ there are no derivatives: $L^p_0 = L^p$ and $H^0 = L^2$.

Weak derivatives. Nash stresses that "the derivatives must be understood in the distributional sense". Here is what that means and why it matters. If f is C^1 and ϕ is a smooth function van-

ishing outside a bounded set (a *test function*), integration by parts gives $\int f'\phi\,\mathrm{d}x = -\int f\phi'\,\mathrm{d}x$, with no boundary terms. The right side makes sense even when f is not differentiable. So:

Working definition 2.33 (weak derivative)

g is the **weak derivative** $\partial^\alpha f$ if

$$\int g\,\phi\,\mathrm{d}^n x = (-1)^{|\alpha|}\int f\,\partial^\alpha\phi\,\mathrm{d}^n x \qquad \text{for every test function } \phi.$$

To compute: differentiate classically on the pieces where f is smooth, then check whether the jumps of f produce extra δ-functions (they do whenever f itself jumps).

Worked example 2.69: Weak derivatives of $|x|$ and of a step

(a) $f(x) = |x|$ on $\mathbb{R}$. For a test function ϕ,

$$\begin{aligned} -\int_{-\infty}^{\infty} |x|\;\phi'(x)\,\mathrm{d}x &\overset{(1)}{=} \int_{-\infty}^{0} x\,\phi'\,\mathrm{d}x - \int_{0}^{\infty} x\,\phi'\,\mathrm{d}x \\ &\overset{(2)}{=} \Big([x\phi]_{-\infty}^{0} - \int_{-\infty}^{0}\phi\,\mathrm{d}x\Big) - \Big([x\phi]_{0}^{\infty} - \int_{0}^{\infty}\phi\,\mathrm{d}x\Big) \\ &\overset{(3)}{=} \int_{-\infty}^{\infty} \operatorname{sgn}(x)\,\phi(x)\,\mathrm{d}x. \end{aligned}$$

(1) $|x| = -x$ for $x < 0$ and x for $x > 0$.
(2) Integrate by parts on each half-line.
(3) The boundary terms vanish: $x\phi = 0$ at $x = 0$, and $\phi = 0$ far away.

So the weak derivative of $|x|$ is $\operatorname{sgn} x$, a bounded function, as the classical derivative away from 0 suggests. *(b)* Now differentiate $\operatorname{sgn} x$:

$$-\int \operatorname{sgn}(x)\,\phi'(x)\,\mathrm{d}x \overset{(1)}{=} \int_{-\infty}^{0}\phi'\,\mathrm{d}x - \int_{0}^{\infty}\phi'\,\mathrm{d}x \overset{(2)}{=} \phi(0) + \phi(0) = 2\phi(0).$$

(1) Split at 0 as before.
(2) Fundamental theorem of calculus, ϕ vanishing at $\pm\infty$.

No function g satisfies $\int g\phi = 2\phi(0)$ for all ϕ; the weak derivative is the distribution 2δ. So, near the origin, $|x|$ has one weak derivative in L^2 but not two: multiplied by a cutoff it is in H^1 and not in H^2.

On manifolds and bundles. Sobolev spaces are defined on a compact manifold M by the partition-of-unity trick Nash mentions: cover M by charts Ω_i, pick $\phi_i \geq 0$ with $\sum\phi_i = 1$ and $\operatorname{supp}\phi_i \subset \Omega_i$, and set $\|f\|_{L^p_k(M)} = \sum_i \|\phi_i f\|_{L^p_k(\Omega_i)}$, computing each term in the coordinates of Ω_i. Different choices give equivalent norms (the chain rule converts derivatives in one chart into combinations of derivatives in another, with smooth, hence bounded, coefficients on the compact supports). For a vector bundle $E \to M$ we do the same with the components of a section in local trivialisations, giving $L^p_k(M,E)$ and $H^k(M,E)$: "vector-valued functions on M". When nothing is lost, we abbreviate to L^p_k and H^k.

2.5.2 Sobolev spaces of any real order via the Fourier transform

Derivatives count as powers of momentum. So instead of adding the norms of derivatives, we can weight the Fourier transform by a power of momentum, and then nothing forces the power to be an integer.

Working definition 2.34 (H^s for real s; Nash (2.75))

For $s \in \mathbb{R}$, $H^s(\mathbb{R}^n)$ is the space of (tempered) distributions [generalised functions, such as δ, that grow at most polynomially, so they have a Fourier transform] f whose Fourier transform is a function with

$$\|f\|_s^2 = \int_{\mathbb{R}^n} \langle\xi\rangle^{2s} \left|\hat f(\xi)\right|^2 \frac{\mathrm{d}^n\xi}{(2\pi)^n} < \infty, \qquad \langle\xi\rangle^{2s} = (1+|\xi|^2)^s. \tag{2.145}$$

Equivalently $f \in H^s$ iff $(1+\Delta)^{s/2} f \in L^2$, and $\|f\|_s = \left\|(1+\Delta)^{s/2} f\right\|_{L^2}$: *apply the Bessel potential of order s (Example 2.50(b)) and ask for square integrability*. For $1 \le p < \infty$, L^p_s is defined by asking that $(1+\Delta)^{s/2} f \in L^p$.

The "equivalently" is one line: $(1+\Delta)^{s/2}$ is the Fourier multiplier $\langle\xi\rangle^s$, and Plancherel's theorem $\int |g|^2 \,\mathrm{d}^n x = \int |\hat g|^2 \frac{\mathrm{d}^n\xi}{(2\pi)^n}$ (in our conventions) turns $\left\|(1+\Delta)^{s/2} f\right\|_{L^2}^2$ into (2.145). At $s = 0$ it is just Plancherel: $H^0 = L^2$.

Agreement with the old definition for integer k. This is Nash's (2.76)–(2.77). Expand the weight with the multinomial theorem:

$$(1+\xi_1^2+\cdots+\xi_n^2)^k = \sum_{|\beta|\le k} \binom{k}{\beta} \xi^{2\beta}, \qquad \binom{k}{\beta} = \frac{k!}{(k-|\beta|)!\,\beta_1!\cdots\beta_n!}, \tag{2.146}$$

a finite sum with positive integer coefficients (Nash's a_α). Then

$$\begin{aligned} \|f\|_k^2 &\overset{(1)}{=} \sum_{|\beta|\le k} \binom{k}{\beta} \int \xi^{2\beta} \left|\hat f(\xi)\right|^2 \frac{\mathrm{d}^n\xi}{(2\pi)^n} \overset{(2)}{=} \sum_{|\beta|\le k} \binom{k}{\beta} \int \left|\widehat{D^\beta f}(\xi)\right|^2 \frac{\mathrm{d}^n\xi}{(2\pi)^n} \\ &\overset{(3)}{=} \sum_{|\beta|\le k} \binom{k}{\beta} \int \left|D^\beta f(x)\right|^2 \mathrm{d}^n x \overset{(4)}{=} \sum_{|\beta|\le k} \binom{k}{\beta} \left\|\partial^\beta f\right\|_{L^2}^2 . \end{aligned} \tag{2.147}$$

(1) Insert (2.146) into (2.145) and integrate term by term.

(2) $\widehat{D^\beta f} = \xi^\beta \hat f$ (derivatives become multiplication), and $\xi^{2\beta}\left|\hat f\right|^2 = \left|\xi^\beta \hat f\right|^2$ since ξ is real.

(3) Plancherel for each $D^\beta f$.

(4) $\left|D^\beta f\right| = \left|(-i)^{|\beta|}\partial^\beta f\right| = \left|\partial^\beta f\right|$.

All coefficients lie between 1 and $k!$, so this norm is equivalent to $\sum \left\|\partial^\beta f\right\|^2$, hence to Nash's (2.144) with $p = 2$. For instance in one dimension

$$\|f\|_1^2 = \|f\|^2 + \|f'\|^2, \qquad \|f\|_2^2 = \|f\|^2 + 2\|f'\|^2 + \|f''\|^2,$$

from $1+\xi^2$ and $(1+\xi^2)^2 = 1 + 2\xi^2 + \xi^4$.

Worked example 2.70: The tent function: both definitions, and its exact smoothness

Let $f(x) = (1-|x|)_+$ (equal to $1-|x|$ on $[-1,1]$, zero outside). Its weak derivative is $f' = +1$ on $(-1,0)$, -1 on $(0,1)$, zero outside (as in Example 2.69, the corners produce no δ). *Position space:*

$$\|f\|^2 = 2\int_0^1 (1-x)^2 dx = \tfrac{2}{3}, \qquad \|f'\|^2 = \int_{-1}^1 1\,dx = 2, \qquad \|f\|_1^2 = \tfrac{8}{3}.$$

Momentum space:

$$\hat f(\xi) \overset{(1)}{=} 2\int_0^1 (1-x)\cos(x\xi)\,dx \overset{(2)}{=} \frac{2(1-\cos\xi)}{\xi^2} = \frac{4\sin^2(\xi/2)}{\xi^2}.$$

(1) f is even, so $e^{-ix\xi}$ may be replaced by $\cos x\xi$ and the integral doubled.

(2) Integrate by parts once: $2\big[(1-x)\frac{\sin x\xi}{\xi}\big]_0^1 + \frac{2}{\xi}\int_0^1 \sin(x\xi)\,dx = \frac{2(1-\cos\xi)}{\xi^2}$.

Then, with the classical integrals $\int_{\mathbb{R}} (\sin x/x)^4 dx = 2\pi/3$ and $\int_{\mathbb{R}} \sin^4 x/x^2\,dx = \pi/2$ (substituting $x = \xi/2$),

$$\int |\hat f|^2 \frac{d\xi}{2\pi} \overset{(1)}{=} \frac{16}{2\pi}\int \frac{\sin^4(\xi/2)}{\xi^4}\,d\xi \overset{(2)}{=} \frac{16}{2\pi}\cdot\frac{2}{16}\cdot\frac{2\pi}{3} = \frac{2}{3},$$
$$\int \xi^2 |\hat f|^2 \frac{d\xi}{2\pi} \overset{(3)}{=} \frac{16}{2\pi}\int \frac{\sin^4(\xi/2)}{\xi^2}\,d\xi \overset{(4)}{=} \frac{16}{2\pi}\cdot\frac{2}{4}\cdot\frac{\pi}{2} = 2.$$

(1) Square $\hat f$.

(2) $\xi = 2x$: $\sin^4(\xi/2)/\xi^4 = \sin^4 x/(16x^4)$ and $d\xi = 2dx$.

(3) Same, with one factor ξ^2 cancelled.

(4) $\xi = 2x$: $\sin^4 x/(4x^2)$, $d\xi = 2dx$.

The two definitions agree: $\frac{2}{3} + 2 = \frac{8}{3}$. But the Fourier side tells us more. Since $|\hat f| \sim 4\sin^2(\xi/2)/\xi^2$ decays like $|\xi|^{-2}$ (on average), $\int \langle\xi\rangle^{2s} |\hat f|^2$ behaves like $\int^\infty \xi^{2s-4} d\xi$, finite iff $2s - 4 < -1$:

$$f \in H^s(\mathbb{R}) \iff s < \tfrac{3}{2}.$$

The tent is in H^1, and even in $H^{1.4}$, but not in $H^{3/2}$. Integer derivatives alone could only have told us "H^1 but not H^2".

Watch out: Normalisation conventions

Nash's (2.75) omits the $(2\pi)^{-n}$ and uses $\langle\xi\rangle^k$ inside a p-th power. Different books use $(1+4\pi^2|\xi|^2)^s$, or $(1+|\xi|)^{2s}$, or $|\xi|^{2s}$ plus the L^2 norm. All of these give the same spaces with equivalent norms (by (2.123) and its relatives); only numerical constants change. The *homogeneous* weight $|\xi|^{2s}$ alone, without the 1, does *not* give an equivalent norm: it ignores low momenta and defines a different space.

2.5.3 Sobolev spaces on the circle, explicitly

On a compact manifold without boundary the cleanest picture is the circle, where the Fourier transform is a Fourier series and every statement becomes a statement about how fast coefficients

decay. Write

$$u(\theta)=\sum_{n\in\mathbb{Z}} c_n\, e^{in\theta}, \qquad c_n=\frac{1}{2\pi}\int_0^{2\pi} u(\theta)e^{-in\theta}\,\mathrm{d}\theta, \qquad \|u\|_{L^2}^2=2\pi\sum_n |c_n|^2,$$

the last being Parseval's identity. The derivative $D_\theta=-i\,\mathrm{d}/\mathrm{d}\theta$ acts by $c_n\mapsto nc_n$, and $\Delta=D_\theta^2$ by $c_n\mapsto n^2c_n$.

Working definition 2.35 ($H^s(S^1)$ and $H^s(T^n)$)

$$H^s(S^1)=\Big\{u=\sum c_n e^{in\theta}:\ \|u\|_s^2=2\pi\sum_{n\in\mathbb{Z}}(1+n^2)^s\,|c_n|^2<\infty\Big\}.$$

On the torus $T^n=\mathbb{R}^n/2\pi\mathbb{Z}^n$ [the cube $[0,2\pi]^n$ with opposite faces identified], replace $n\in\mathbb{Z}$ by $k\in\mathbb{Z}^n$ and $1+n^2$ by $1+|k|^2$. *Test:* find the power law of $|c_n|$ at large $|n|$. If $|c_n|\sim|n|^{-a}$, then $u\in H^s$ iff $2s-2a<-1$, i.e. $s<a-\frac12$.

The test is the p-series: $\sum_{n\ge1} n^{2s}n^{-2a}$ converges iff $2s-2a<-1$.

Worked example 2.71: Triangle, square wave, delta: three rungs of the ladder

Work on $\theta\in(-\pi,\pi]$, extended periodically.
Square wave $q(\theta)=\operatorname{sgn}\theta$. Its coefficients:

$$c_n(q)\overset{(1)}{=}\frac{1}{2\pi}\Big(\int_0^{\pi}e^{-in\theta}\mathrm{d}\theta-\int_{-\pi}^{0}e^{-in\theta}\mathrm{d}\theta\Big)\overset{(2)}{=}\frac{-i}{\pi}\int_0^{\pi}\sin n\theta\,\mathrm{d}\theta\overset{(3)}{=}\frac{-i\,(1-(-1)^n)}{\pi n},$$

(1) Definition, splitting at 0.
(2) Substitute $\theta\to-\theta$ in the second integral; $e^{-in\theta}-e^{in\theta}=-2i\sin n\theta$.
(3) $\int_0^{\pi}\sin n\theta\,\mathrm{d}\theta=(1-\cos n\pi)/n$ and $\cos n\pi=(-1)^n$.

So $c_n=-2i/(\pi n)$ for odd n and 0 for even n: $|c_n|\sim|n|^{-1}$, $a=1$, and $q\in H^s$ iff $s<\frac12$.
Triangle wave $t(\theta)=|\theta|$. Here $c_0=\pi/2$ and for $n\neq0$

$$c_n(t)\overset{(1)}{=}\frac{1}{\pi}\int_0^{\pi}\theta\cos n\theta\,\mathrm{d}\theta\overset{(2)}{=}\frac{(-1)^n-1}{\pi n^2},$$

(1) t is even, so only the cosine part survives.
(2) Integrate by parts: $\big[\theta\frac{\sin n\theta}{n}\big]_0^{\pi}+\big[\frac{\cos n\theta}{n^2}\big]_0^{\pi}$.

so $c_n=-2/(\pi n^2)$ for odd n: $a=2$, and $t\in H^s$ iff $s<\frac32$ (the tent again).
Delta at $\theta=0$: $c_n=\frac{1}{2\pi}$ for every n, so $a=0$ and $\delta\in H^s$ iff $s<-\frac12$.
The ladder. Now differentiate. Term by term, $\frac{\mathrm{d}}{\mathrm{d}\theta}$ multiplies c_n by in:

$$in\,c_n(t)=in\cdot\frac{-2}{\pi n^2}=\frac{-2i}{\pi n}=c_n(q)\quad(n\text{ odd}),$$
$$in\,c_n(q)=in\cdot\frac{-2i}{\pi n}=\frac{2}{\pi}=c_n(2\delta_0-2\delta_\pi)\quad(n\text{ odd}),$$

where δ_π is the delta at $\theta=\pi$, with $c_n(\delta_\pi)=(-1)^n/2\pi$, so $c_n(2\delta_0-2\delta_\pi)=(1-(-1)^n)/\pi$. So $t'=q$ and $q'=2\delta_0-2\delta_\pi$: the square wave jumps up by 2 at 0 and down by 2 at π. Each derivative moves one rung down the Sobolev ladder: thresholds $\frac32\to\frac12\to-\frac12$ (Figure 2.36).

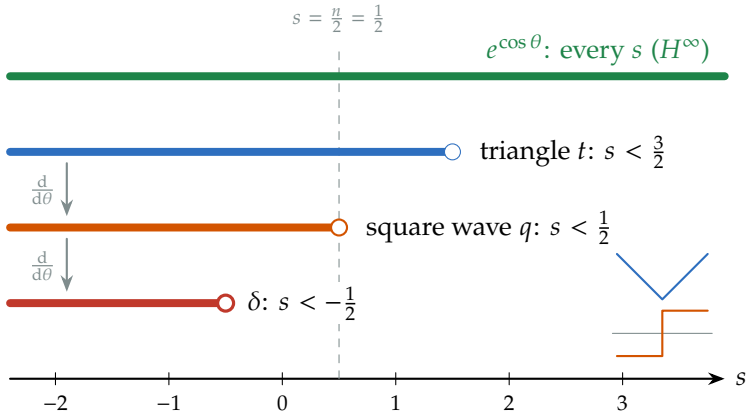


Figure 2.36: The Sobolev ladder on the circle. Each bar shows the values of s for which the function lies in $H^s(S^1)$; the open circle marks the threshold, which is never attained. Differentiating moves a function down one rung (threshold decreases by exactly 1). A jump discontinuity sits just below $s = \frac{1}{2} = \frac{n}{2}$, the dimension-dependent line above which functions are guaranteed to be continuous (Sobolev embedding, Section 2.5.7). Small inset: the triangle (blue) and its derivative, the square wave (orange).

Worked example 2.72: A smooth function is in every H^s

For $u = e^{\cos\theta}$ the coefficients are the modified Bessel numbers $c_n = I_n(1) = \sum_{j\geq 0} \frac{(1/2)^{2j+|n|}}{j!\,(j+|n|)!}$, and $I_n(1) \leq \frac{(1/2)^{|n|}}{|n|!} e^{1/4}$. This decays faster than any power of n, so $\sum (1+n^2)^s\,|c_n|^2 < \infty$ for every s. In general, integrating by parts k times shows $|c_n| \leq \left\|u^{(k)}\right\|_{L^1} / (2\pi\,|n|^k)$ for any C^k function: *smooth functions have rapidly decreasing Fourier coefficients*, and conversely. So $C^\infty(S^1) = \bigcap_s H^s(S^1)$, a first glimpse of $H^\infty = C^\infty$ on compact manifolds.

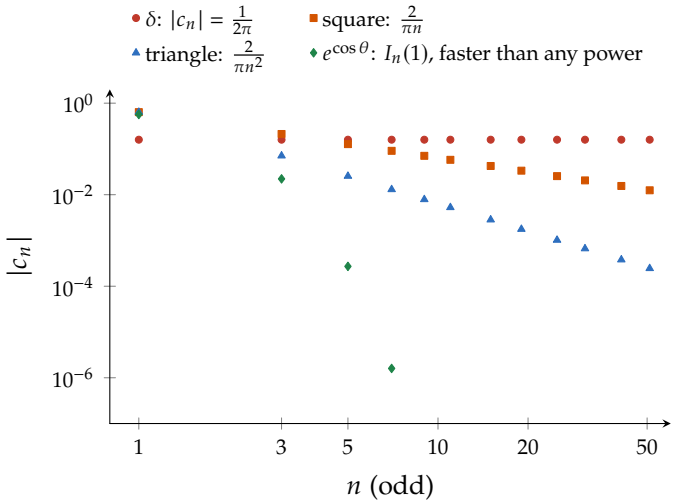


Figure 2.37: Decay of Fourier coefficients on a log–log plot. A power law $|c_n| \sim n^{-a}$ is a straight line of slope $-a$, and the Sobolev threshold is $s = a - \frac{1}{2}$. Slopes $0, -1, -2$ give thresholds $-\frac{1}{2}, \frac{1}{2}, \frac{3}{2}$. The smooth function's coefficients (green) bend down off every straight line: it is in every H^s.

Physical picture: Brownian paths and the path integral

A Brownian path on $[0, 2\pi]$ (with the endpoints pinned) has a random Fourier sine series with independent Gaussian coefficients of size $|c_n| \sim n^{-1}$: this is the free-particle path integral of quantum mechanics in Euclidean time. By the test above, a typical path is in H^s exactly for $s < \frac{1}{2}$: continuous (just), but with no square-integrable derivative. So the kinetic term $\int \dot{x}^2 \, \mathrm{d}t$ of the action is infinite on almost every path that contributes. This is the simplest example of the general fact that the field configurations in a Euclidean path integral are distributions in some H^s with s below the threshold where the action is finite.

2.5.4 Negative orders are dual spaces

What is H^{-1}? The Fourier definition (2.145) works for negative s without change: the weight $\langle \xi \rangle^{2s}$ now *damps* high momenta, so $\hat{f}$ is allowed to grow, and f can be a distribution such as δ. Nash's point is that these spaces are exactly the dual spaces of the positive ones:

$$H^{-s} = (H^s)^*, \tag{2.148}$$

where V^* is the space of continuous linear functionals on V. In words: *an element of H^{-s} is something you can integrate against a function with s derivatives in L^2.*

Take $f \in H^s$ and $g \in H^{-s}$ and pair them by $\langle f, g \rangle = \int f \, \bar{g} \, \mathrm{d}^n x$ (Nash's $g(f)$). Then

$$\begin{aligned} |\langle f, g \rangle| &\overset{(1)}{=} \left| \int \hat{f}(\xi) \overline{\hat{g}(\xi)} \, \frac{\mathrm{d}^n \xi}{(2\pi)^n} \right| \overset{(2)}{=} \left| \int \langle \xi \rangle^s \hat{f}(\xi) \, \langle \xi \rangle^{-s} \overline{\hat{g}(\xi)} \, \frac{\mathrm{d}^n \xi}{(2\pi)^n} \right| \\ &\overset{(3)}{\le} \left(\int \langle \xi \rangle^{2s} \left| \hat{f} \right|^2 \frac{\mathrm{d}^n \xi}{(2\pi)^n} \right)^{1/2} \left(\int \langle \xi \rangle^{-2s} \, |\hat{g}|^2 \, \frac{\mathrm{d}^n \xi}{(2\pi)^n} \right)^{1/2} \overset{(4)}{=} \|f\|_s \, \|g\|_{-s} \, . \end{aligned} \tag{2.149}$$

(1) Parseval (the polarised form of Plancherel).
(2) Multiply and divide by $\langle \xi \rangle^s$: this is Nash's $h^k = (1 + \xi^2)^{k/2}$ trick in (2.78).
(3) Cauchy–Schwarz in $L^2(\frac{\mathrm{d}^n \xi}{(2\pi)^n})$. This is Nash's (2.79).
(4) Definition (2.145) with s and $-s$.

So every $g \in H^{-s}$ defines a bounded functional on H^s, with norm at most $\|g\|_{-s}$. Conversely, every bounded functional ℓ on the Hilbert space H^s is $\ell(f) = (f, h)_s$ for some $h \in H^s$ (Riesz representation), and $(f, h)_s = \int \hat{f} \, \langle \xi \rangle^{2s} \bar{\hat{h}} \, \frac{\mathrm{d}^n \xi}{(2\pi)^n} = \langle f, g \rangle$ with $\hat{g} = \langle \xi \rangle^{2s} \hat{h}$, which has $\|g\|_{-s} = \|h\|_s < \infty$. So the correspondence is onto and (2.148) holds isometrically. (Nash's own argument shows one direction, "$g(f) < \infty$ for all f iff $g \in H^{-k}$"; the Riesz step supplies the other.)

Worked example 2.73: The delta function as an element of H^{-s}

On the circle. Pairing with δ_0 is evaluation, $u \mapsto u(0) = \sum_n c_n$. Is it bounded on $H^s(S^1)$?

$$\begin{aligned} |u(0)| &\overset{(1)}{\le} \sum_n (1 + n^2)^{s/2} \, |c_n| \, (1 + n^2)^{-s/2} \overset{(2)}{\le} \left(\sum_n (1 + n^2)^s \, |c_n|^2 \right)^{1/2} \left(\sum_n (1 + n^2)^{-s} \right)^{1/2} \\ &= \frac{1}{\sqrt{2\pi}} \, \|u\|_s \, \left(\sum_n (1 + n^2)^{-s} \right)^{1/2} . \end{aligned}$$

(1) Triangle inequality, and multiply and divide by $(1 + n^2)^{s/2}$.
(2) Cauchy–Schwarz for sequences.

The last sum is finite iff $2s > 1$. So evaluation at a point is continuous on $H^s(S^1)$ precisely

for $s > \frac{1}{2}$, which matches $\delta \in H^{-s}$ iff $-s < -\frac{1}{2}$ (Example 2.71). *In* $\mathbb{R}^n$. $\hat{\delta} = 1$, so $\|\delta\|_s^2 = \int (1+|\xi|^2)^s \frac{d^n\xi}{(2\pi)^n} = \frac{\mathrm{vol}(S^{n-1})}{(2\pi)^n} \int_0^\infty (1+\rho^2)^s \rho^{n-1} d\rho$, finite iff $2s + n - 1 < -1$:

$$\delta \in H^s(\mathbb{R}^n) \iff s < -\tfrac{n}{2}.$$

A point charge in three dimensions is in $H^{-3/2-\varepsilon}$; its Coulomb potential, $(\Delta)^{-1}$ of it, gains two derivatives and is (locally) in $H^{1/2-\varepsilon}$: indeed $1/r$ is locally square integrable in $\mathbb{R}^3$ but its gradient $\sim 1/r^2$ is not.

The L^p version. For $1 < p < \infty$ Nash quotes (from Treves)

$$(L^p_k)^* = L^{p'}_{-k}, \qquad p' = \frac{p}{p-1}, \qquad \text{i.e. } \frac{1}{p} + \frac{1}{p'} = 1. \tag{2.150}$$

At $k = 0$ this is the familiar statement $(L^p)^* = L^{p'}$, which rests on Hölder's inequality $\left|\int f\bar{g}\right| \le \|f\|_{L^p}\,\|g\|_{L^{p'}}$ (the L^p replacement for Cauchy–Schwarz). For $p = 2$, $p' = 2$ and we are back to (2.148). The conjugate exponent is forced by scaling: the pairing $\int f\bar{g}$ is invariant under $f \to \lambda^{n/p} f(\lambda x)$, $g \to \lambda^{n/p'} g(\lambda x)$ exactly when $1/p + 1/p' = 1$.

2.5.5 Test functions, distributions, and rigged Hilbert space

Since negative Sobolev spaces contain distributions, Nash pauses to recall the standard function and distribution spaces. We give each one with examples, because the whole point is to know what lives where.

Working definition 2.36 (the three test-function spaces; Nash (2.81))

- $\mathcal{D}(\mathbb{R}^n)$: C^∞ functions with compact support [each one is zero outside some bounded set] (**test functions**).
- $\mathcal{S}(\mathbb{R}^n)$: C^∞ functions such that $x^\alpha D^\beta f \to 0$ as $|x| \to \infty$ for all α, β (**Schwartz space**: smooth and rapidly decreasing, with all derivatives).
- $\mathcal{E}(\mathbb{R}^n) = C^\infty(\mathbb{R}^n)$: all smooth functions.

Clearly $\mathcal{D} \subset \mathcal{S} \subset \mathcal{E}$.

Worked example 2.74: Who lives where

A bump function in $\mathcal{D}$. $\psi(x) = e^{-1/(1-x^2)}$ for $|x| < 1$ and 0 otherwise. It is smooth even at $x = \pm 1$: near $x = 1$ it is $e^{-1/t}$ times a smooth function, with $t = 1 - x^2 \to 0^+$, and every derivative of $e^{-1/t}$ is (polynomial in $1/t$) $e^{-1/t} \to 0$ as $t \to 0^+$, because the exponential beats every power. So all derivatives match the zero function from outside. *In $\mathcal{S}$ but not $\mathcal{D}$:* the Gaussian e^{-x^2} (every $x^\alpha D^\beta e^{-x^2}$ is a polynomial times e^{-x^2}), and $1/\cosh x$. *In $\mathcal{E}$ but not $\mathcal{S}$:* $1/(1+x^2)$ (smooth, but $x^2 \cdot \frac{1}{1+x^2} \to 1 \neq 0$), $\sin x$, x^3, e^x.

Distributions are the continuous linear functionals on these spaces:

$$\mathcal{D} \subset \mathcal{S} \subset \mathcal{E} \qquad \Longrightarrow \qquad \mathcal{E}' \subset \mathcal{S}' \subset \mathcal{D}'. \tag{2.151}$$

The inclusions reverse (Nash's (2.82)): the more test functions a functional must accept, the fewer functionals qualify. A functional on $\mathcal{E}$ must cope with e^{x^2}; a functional on $\mathcal{D}$ only with compactly supported functions.

Working definition 2.37 (distribution, operationally)

A **distribution** $T \in \mathcal{D}'$ is a linear rule $\phi \mapsto T(\phi)$ on test functions that is continuous in the sense: for each bounded region K there are C, N with $|T(\phi)| \le C \sum_{|\alpha|\le N} \sup |\partial^\alpha \phi|$ for all ϕ supported in K. *Every locally integrable function f [$\int_K |f| < \infty$ over every bounded region K, e.g. $1/|x|$ in $\mathbb{R}^3$] is one*, via $T_f(\phi) = \int f\phi$. Derivatives are defined by moving them onto ϕ: $(\partial^\alpha T)(\phi) = (-1)^{|\alpha|} T(\partial^\alpha \phi)$, which is the weak derivative again. **Tempered distributions** $\mathcal{S}'$ are those that also accept Schwartz functions; $\mathcal{E}'$ consists of the distributions with compact support [there is a bounded set K with $T(\phi) = 0$ whenever ϕ vanishes near K].

Worked example 2.75: Examples of distributions and where they sit

T	$\mathcal{E}'$ (compact support)	$\mathcal{S}'$ (tempered)	$\mathcal{D}'$
δ, δ', $\mathbf{1}_{[0,1]}$	yes	yes	yes
1, x^3, e^{ipx}, $\lvert x\rvert^{-1/2}$, p.v.$\frac{1}{x}$	no	yes	yes
e^x, e^{x^2}	no	no	yes

Why $e^x \notin \mathcal{S}'$: pair it with the Schwartz function $\phi(x) = e^{-\sqrt{1+x^2}}$ (smooth, and it decays exponentially together with its derivatives). For $x \to +\infty$, $e^x\phi(x) = e^{x-\sqrt{1+x^2}} \to e^0 = 1$, so $\int e^x\phi$ diverges. *Why* p.v.$\frac{1}{x}$ *is a distribution:* $\lim_{\varepsilon\to 0}\int_{|x|>\varepsilon} \frac{\phi(x)}{x}\mathrm{d}x = \int_0^\infty \frac{\phi(x)-\phi(-x)}{x}\mathrm{d}x$, and the integrand is bounded by $2\sup|\phi'|$ near 0. This is the kernel of the Hilbert transform in Example 2.68.

The Fourier transform maps $\mathcal{S}$ onto $\mathcal{S}$ (derivatives and multiplication by x swap roles, and both are harmless there), and hence, by duality, $\mathcal{S}'$ onto $\mathcal{S}'$. That is why tempered distributions are the natural home of Fourier analysis: $\hat{1} = (2\pi)^n\delta$, $\hat{\delta} = 1$, $\widehat{e^{ipx}} = (2\pi)^n\delta(\xi - p)$. It is also why our $H^s(\mathbb{R}^n)$ were defined as subspaces of $\mathcal{S}'$.

The rigged Hilbert space (Nash's (2.83)). $\mathcal{S} \subset L^2$, and L^2 is its own dual (Riesz again: $g \in L^2$ is the functional $f \mapsto \langle g, f\rangle$). Dualising $\mathcal{S} \subset L^2$ gives $(L^2)' \subset \mathcal{S}'$, hence the **Gelfand triple**

$$\mathcal{S} \subset L^2 \subset \mathcal{S}'. \tag{2.152}$$

This is the identification "a function is a distribution" made precise.

Dictionary: Dirac's bras and kets live in a Gelfand triple

Quantum mechanics	Functional analysis
normalisable state $\lvert\psi\rangle$	$\psi \in L^2$
"nice" states (finite $\langle x^a p^b\rangle$ for all a, b)	$\psi \in \mathcal{S}$
momentum eigenstate $\lvert p\rangle$, $\langle x\rvert p\rangle = e^{ipx}$	$e^{ipx} \in \mathcal{S}' \setminus L^2$
position eigenstate $\lvert x_0\rangle$, $\langle x\rvert x_0\rangle = \delta(x - x_0)$	$\delta_{x_0} \in \mathcal{S}' \setminus L^2$
$\langle p\rvert\psi\rangle$ for $\psi \in \mathcal{S}$	a finite number: $\hat{\psi}(p)$
$\langle p\rvert p'\rangle = 2\pi\delta(p - p')$	pairing of two elements of $\mathcal{S}'$: only a distribution

Dirac's formalism works because the generalised eigenvectors sit in $\mathcal{S}'$ and are only ever paired with states in $\mathcal{S}$. In Sobolev terms $\delta \in H^{-1/2-\varepsilon}(\mathbb{R})$: one needs slightly more than half a derivative of smoothness in the state to evaluate it at a point.

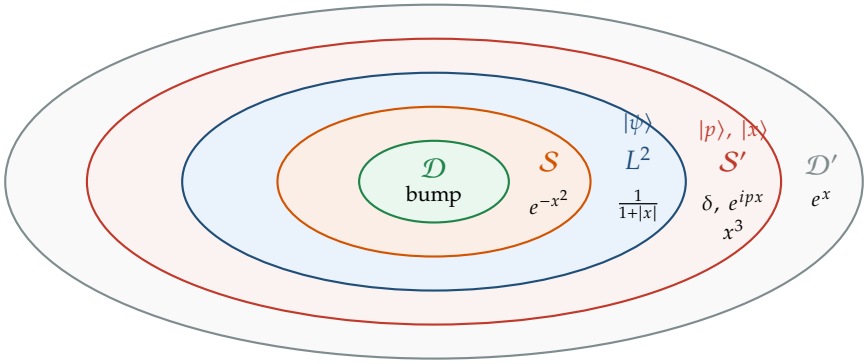


Figure 2.38: The nested spaces of (2.151)–(2.152) on $\mathbb{R}$, with one inhabitant of each shell. Test functions (green) sit inside Schwartz functions, inside L^2 (normalisable states), inside tempered distributions (where plane waves and δ live), inside all distributions (where even e^x lives). The Sobolev spaces H^s interpolate between $\mathcal{S}$ and $\mathcal{S}'$: H^s shrinks towards $\mathcal{S}$-like smoothness as $s \to +\infty$ and grows towards $\mathcal{S}'$ as $s \to -\infty$.

2.5.6 The Sobolev scale

Since $\langle\xi\rangle \geq 1$, a larger exponent gives a larger weight:

$$s > s' \implies \langle\xi\rangle^{2s'} \leq \langle\xi\rangle^{2s} \implies \|f\|_{s'} \leq \|f\|_s \implies H^s \subset H^{s'}.$$

So the spaces *shrink* as s increases (Nash's (2.84)):

$$\cdots \subset H^2 \subset H^1 \subset H^0 = L^2 \subset H^{-1} \subset H^{-2} \subset \cdots, \tag{2.153}$$

with the inclusion maps having norm ≤ 1. The extremes are (Nash's (2.85))

$$H^\infty = \bigcap_{s\in\mathbb{R}} H^s, \qquad H^{-\infty} = \bigcup_{s\in\mathbb{R}} H^s = (H^\infty)^*, \tag{2.154}$$

and the whole hierarchy (Nash's (2.86)) is

$$\mathcal{D} \subset \mathcal{S} \subset H^\infty \subset \cdots \subset H^k \subset \cdots \subset H^0 \subset \cdots \subset H^{-k} \subset \cdots \subset H^{-\infty} \subset \mathcal{S}' \subset \mathcal{D}'. \tag{2.155}$$

Each step of the chain is a real loss or gain: the tent function is in H^1 but not H^2; the square wave in H^0 but not H^1; δ in H^{-1} but not H^0 (in one dimension).

Watch out: $H^\infty(\mathbb{R}^n)$ is not $\mathcal{S}$

$H^\infty(\mathbb{R}^n)$ contains $1/(1+x^2)$: all its derivatives are square integrable, but it does not decay rapidly, so it is not in $\mathcal{S}$. Sobolev norms measure smoothness and L^2-size, not decay. On a *compact* manifold there is no "infinity" to decay at, and the chain simplifies (Nash): drop $\mathcal{S}$ and $\mathcal{S}'$, and

$$C^\infty(M,E) = H^\infty(M,E) \subset \cdots \subset H^k(M,E) \subset \cdots \subset H^{-\infty}(M,E) = \mathcal{D}'(M,E).$$

The first equality is the Sobolev embedding theorem of the next subsection (and Example 2.72 on the circle).

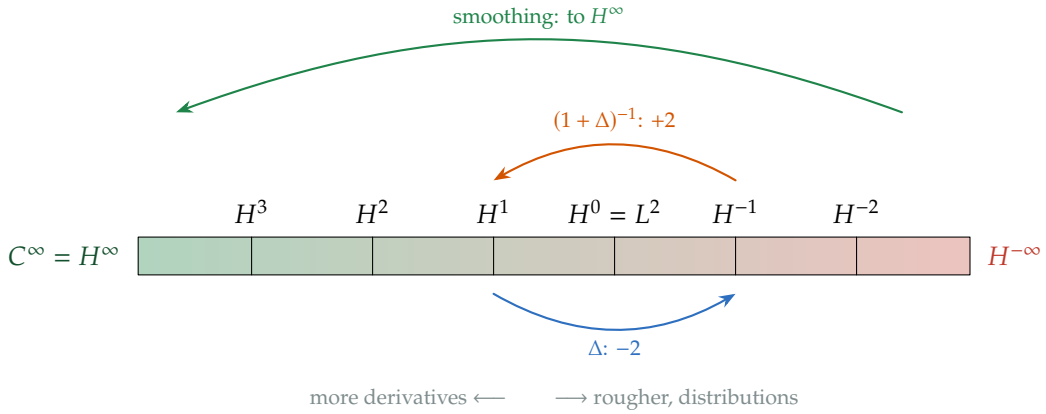


Figure 2.39: The Sobolev scale as a ruler. Smooth functions sit at the left end, distributions at the right. A ΨDO of order m moves everything m steps to the right (for $m > 0$) or $|m|$ steps to the left (for $m < 0$): the Laplacian moves H^1 to H^{-1}, its inverse moves H^{-1} to H^1. A smoothing operator sends everything to the far left end.

2.5.7 How many derivatives buy continuity: the Sobolev embedding theorem

An element of H^k is defined through integrals. Integrals do not see what happens on a set of measure zero, so a priori an element of H^k is not a function with definite values at points; it is a class of functions that agree almost everywhere. The question Nash now answers is: *when is an element of H^k (or L^p_k) an honest continuous function, or an honest C^l function?* Physically: when does "the value of the field at the point x" mean anything?

We have already seen the answer on the circle in one special case. In Example 2.73, evaluation at a point, $u \mapsto u(0)$, was a bounded functional on $H^s(S^1)$ exactly when $s > \frac{1}{2}$. A bounded evaluation functional is what continuity at a point needs. So on the circle *half a derivative in L^2 plus a little bit* is what it takes to be continuous. The general rule replaces $\frac{1}{2}$ by $\frac{n}{p}$.

The rule from dimensional analysis. Before any theorem, here is the physicist's derivation. Suppose an inequality of the form

$$\sup_x |\partial^\alpha f(x)| \le C \, \|f\|_{k,p} \qquad (|\alpha| = l) \tag{2.156}$$

were true for all f, with C independent of f. Test it on a function that is squeezed into a small region: pick any bump f and set $f_\lambda(x) = f(\lambda x)$ with $\lambda \gg 1$ (Figure 2.40). For a multi-index β,

$$\begin{aligned} \left\|\partial^\beta f_\lambda\right\|_{L^p}^p &\overset{(1)}{=} \int \left|\lambda^{|\beta|}(\partial^\beta f)(\lambda x)\right|^p \, \mathrm{d}^n x \overset{(2)}{=} \lambda^{p|\beta|} \int \left|(\partial^\beta f)(y)\right|^p \, \lambda^{-n} \mathrm{d}^n y \\ &\overset{(3)}{=} \lambda^{p|\beta|-n} \left\|\partial^\beta f\right\|_{L^p}^p, \qquad \text{so} \qquad \left\|\partial^\beta f_\lambda\right\|_{L^p} = \lambda^{|\beta|-\frac{n}{p}} \left\|\partial^\beta f\right\|_{L^p}. \end{aligned} \tag{2.157}$$

(1) Chain rule: each derivative of $f(\lambda x)$ brings out a factor λ.
(2) Substitute $y = \lambda x$, $\mathrm{d}^n x = \lambda^{-n} \mathrm{d}^n y$.
(3) Collect powers of λ and take the p-th root.

For large λ the terms with $|\beta| = k$ dominate the norm (2.144), so $\|f_\lambda\|_{k,p} \approx \lambda^{k-n/p} \times$(const). The left side of (2.156) scales as $\sup|\partial^\alpha f_\lambda| = \lambda^l \sup|\partial^\alpha f|$, with no $\lambda^{-n/p}$ because a supremum is not an integral. So (2.156) can only survive $\lambda \to \infty$ if

$$\lambda^l \lesssim \lambda^{k-n/p} \quad \text{for all large } \lambda \qquad \Longleftrightarrow \qquad l \le k - \frac{n}{p}.$$

The number $k - n/p$ is the **Sobolev number** (or "differential dimension") of L^p_k: it counts derivatives, but each L^p integral over n dimensions costs n/p of them. A C^l space has Sobolev number l (the supremum norm is $p = \infty$). The theorem says that the naive bookkeeping is right, provided the inequality is strict.

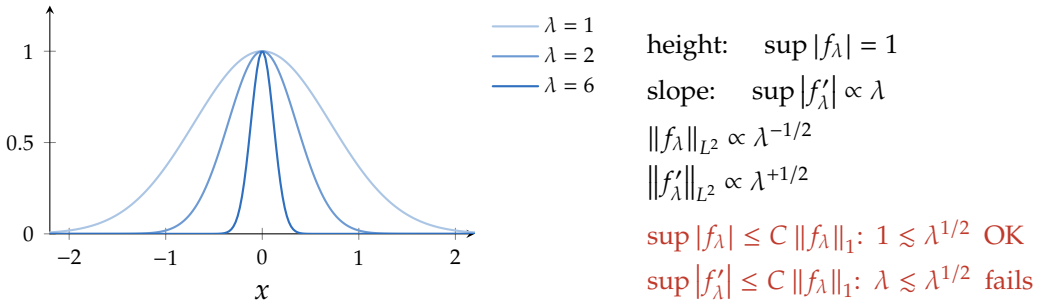


Figure 2.40: The scaling test behind the Sobolev embedding, in $n = 1$, $p = 2$. Squeezing a bump by a factor λ keeps its height, multiplies its slope by λ, and changes its L^2-type norms by $\lambda^{|\beta|-1/2}$. An H^1 bound can control the height (Sobolev number $1 - \frac{1}{2} = \frac{1}{2} > 0$) but not the slope ($\frac{1}{2} < 1$). This is dimensional analysis, with λ playing the role of an inverse length.

Key idea: Nash calls it Sobolev (i), (2.87)

Let $M = \mathbb{R}^n$ or M a compact n-manifold. If

$$k - \frac{n}{p} > l \qquad \text{then} \qquad L^p_k(M) \subset C^l(M), \tag{2.158}$$

where C^l is the space of l times differentiable functions with continuous derivatives. More precisely: every $f \in L^p_k$ agrees almost everywhere with a C^l function, and $\sup_{|\alpha|\le l} \sup_x |\partial^\alpha f| \le C \|f\|_{k,p}$. On a compact M the inclusion map $L^p_k \to C^l$ is moreover a *compact operator* (Section 2.5.8).

Why it is true, for $p = 2$. This is Nash's (2.89)–(2.90), with every step filled in. Let $f \in H^k(\mathbb{R}^n)$ and $|\alpha| = l$. Fourier inversion writes the derivative as

$$\partial^\alpha f(x) = \int e^{ix\cdot\xi} (i\xi)^\alpha \hat f(\xi) \frac{\mathrm{d}^n\xi}{(2\pi)^n}, \tag{2.159}$$

provided the integral converges absolutely. If it does, the right side is automatically a continuous function of x: the integrand is continuous in x and bounded by the fixed integrable function $\left|\xi^\alpha \hat f(\xi)\right|$, so we may pass limits $x \to x_0$ inside the integral (dominated convergence). So everything reduces to one estimate:

$$\begin{aligned}
\int \left|\xi^\alpha \hat f(\xi)\right| \frac{\mathrm{d}^n\xi}{(2\pi)^n} &\overset{(1)}{=} \int \langle\xi\rangle^k \left|\hat f(\xi)\right| \cdot \frac{|\xi^\alpha|}{\langle\xi\rangle^k} \frac{\mathrm{d}^n\xi}{(2\pi)^n} \\
&\overset{(2)}{\le} \left(\int \langle\xi\rangle^{2k} \left|\hat f(\xi)\right|^2 \frac{\mathrm{d}^n\xi}{(2\pi)^n}\right)^{1/2} \left(\int \frac{|\xi^\alpha|^2}{(1+|\xi|^2)^k} \frac{\mathrm{d}^n\xi}{(2\pi)^n}\right)^{1/2} \\
&\overset{(3)}{=} \|f\|_k \left(\int \frac{|\xi^\alpha|^2}{(1+|\xi|^2)^k} \frac{\mathrm{d}^n\xi}{(2\pi)^n}\right)^{1/2} \overset{(4)}{\le} \|f\|_k \left(\int \frac{|\xi|^{2l}}{(1+|\xi|^2)^k} \frac{\mathrm{d}^n\xi}{(2\pi)^n}\right)^{1/2}.
\end{aligned} \tag{2.160}$$

(1) Multiply and divide by $\langle\xi\rangle^k = (1+|\xi|^2)^{k/2}$: the same trick as in the duality argument (2.149).

(2) Cauchy–Schwarz in $L^2(\frac{\mathrm{d}^n\xi}{(2\pi)^n})$.
(3) The first factor is the H^k norm (2.145).
(4) $|\xi^\alpha| = |\xi_1|^{\alpha_1}\cdots|\xi_n|^{\alpha_n} \le |\xi|^{|\alpha|} = |\xi|^l$.

It remains to decide when the last integral is finite. Only large $|\xi|$ can cause trouble (the integrand is bounded near $\xi = 0$). Go to polar coordinates, $\mathrm{d}^n\xi = \rho^{n-1}\mathrm{d}\rho\,\mathrm{d}\Omega_{n-1}$ with $\rho = |\xi|$:

$$\int_{|\xi|>1} \frac{|\xi|^{2l}}{(1+|\xi|^2)^k}\,\mathrm{d}^n\xi \overset{(1)}{=} \mathrm{vol}(S^{n-1}) \int_1^\infty \frac{\rho^{2l+n-1}}{(1+\rho^2)^k}\,\mathrm{d}\rho \overset{(2)}{\le} \mathrm{vol}(S^{n-1}) \int_1^\infty \rho^{2l-2k+n-1}\,\mathrm{d}\rho$$
$$\overset{(3)}{=} \mathrm{vol}(S^{n-1})\,\frac{1}{2k-2l-n} < \infty \qquad \text{iff} \qquad 2l-2k+n<0. \tag{2.161}$$

(1) The integrand depends only on ρ; the angular integral gives the area of the unit sphere.
(2) $(1+\rho^2)^k \ge \rho^{2k}$. For the "only if" direction use $(1+\rho^2)^k \le (2\rho^2)^k$ for $\rho \ge 1$ instead.
(3) $\int_1^\infty \rho^{-1-a}\mathrm{d}\rho = 1/a$ for $a>0$ and $=\infty$ for $a \le 0$; here $a = 2k-2l-n$.

The condition $2l-2k+n<0$ is $k - \frac{n}{2} > l$: exactly (2.158) for $p=2$. This is what Nash calls "power counting". It is the *same* power counting that decides whether a Feynman integral $\int \mathrm{d}^n\xi\,\xi^{2l}/(\xi^2+m^2)^k$ is ultraviolet convergent; here the "loop momentum" is the Fourier variable of f.

For a compact M, cut f into pieces $\phi_i f$ with a partition of unity, each supported in a chart; each piece is a compactly supported function on $\mathbb{R}^n$ to which the estimate applies, and a finite sum of C^l functions is C^l. On the circle or torus we can even skip the charts:

Worked example 2.76: The sharp constant for $H^1(\mathbb{R}) \subset C^0$, and its extremal

Take $n=1$, $k=1$, $l=0$. Then (2.160) with $\alpha=0$ gives

$$\sup_x |f(x)| \overset{(1)}{\le} \int \left|\hat f(\xi)\right| \frac{\mathrm{d}\xi}{2\pi} \overset{(2)}{\le} \|f\|_1 \left(\int_{-\infty}^\infty \frac{1}{1+\xi^2}\frac{\mathrm{d}\xi}{2\pi}\right)^{1/2} \overset{(3)}{=} \frac{1}{\sqrt{2}}\,\|f\|_1 .$$

(1) $\left|\int e^{ix\xi}\hat f\,\mathrm{d}\xi\right| \le \int\left|\hat f\right|\mathrm{d}\xi$.
(2) (2.160).
(3) $\int_{-\infty}^\infty \mathrm{d}\xi/(1+\xi^2) = \pi$, so the bracket is $\frac12$.

Is the constant $1/\sqrt{2}$ best possible? Equality in Cauchy–Schwarz needs $\langle\xi\rangle\hat f \propto \langle\xi\rangle^{-1}$, i.e. $\hat f \propto 1/(1+\xi^2)$. By Example 2.47 (with $m=1$) this is $f(x) = e^{-|x|}$. Check directly:

$$\|f\|_{L^2}^2 = \int e^{-2|x|}\mathrm{d}x = 1, \qquad \|f'\|_{L^2}^2 = \int e^{-2|x|}\mathrm{d}x = 1,$$
$$\|f\|_1 = \sqrt{2}, \qquad \tfrac{1}{\sqrt2}\|f\|_1 = 1 = f(0).$$

Equality holds. So the extremal function of the Sobolev inequality is the Green's function of $1+D^2$: the function with the least "energy" $\|f\|^2+\|f'\|^2$ for a given peak height. Physically, it is the shape of an elastic string on a spring foundation pulled up at one point. The kink at $x=0$ shows the function is continuous but no better: $e^{-|x|} \in H^s$ iff $s < \frac32$, just like the tent.

Worked example 2.77: Continuity on the circle and on the torus

On S^1, for $u = \sum c_n e^{in\theta} \in H^1(S^1)$,

$$\sup_\theta |u(\theta)| \overset{(1)}{\le} \sum_n |c_n| \overset{(2)}{\le} \Big(\sum_n (1+n^2)\,|c_n|^2\Big)^{1/2} \Big(\sum_n \frac{1}{1+n^2}\Big)^{1/2} \overset{(3)}{=} \Big(\frac{\pi\coth\pi}{2\pi}\Big)^{1/2} \|u\|_1 .$$

(1) Triangle inequality on the Fourier series; $\left|e^{in\theta}\right| = 1$.
(2) Cauchy–Schwarz for sequences, as in Example 2.73.
(3) $\sum_{n\in\mathbb{Z}} \frac{1}{1+n^2} = \pi\coth\pi \approx 3.153$ (the Mittag-Leffler expansion of coth), and $\|u\|_1^2 = 2\pi\sum(1+n^2)\,|c_n|^2$.

Since the Fourier series converges absolutely and uniformly, its sum is continuous. On the torus T^n the same argument needs $\sum_{k\in\mathbb{Z}^n}(1+|k|^2)^{-s} < \infty$, which (comparing the lattice sum with the integral (2.161)) holds iff $2s > n$. On T^2, H^1 is *not* enough for continuity: $\sum_{k\in\mathbb{Z}^2}(1+|k|^2)^{-1}$ diverges logarithmically, like $\int \rho\,\mathrm{d}\rho/\rho^2$. The next example shows the divergence is real.

The condition cannot be dropped. Nash says it is "easy enough" to give evidence that $k - n/p > l$ is necessary. Here is the standard evidence. Let χ be a smooth cutoff equal to 1 near 0 and vanishing for $|x| \ge \frac12$, and let $f(x) = \chi(x)\,|x|^\lambda$ on $\mathbb{R}^n$ with λ not an even integer. Near 0, each derivative lowers the power by one: $\left|\partial^\beta f\right| \sim |x|^{\lambda-|\beta|}$. So

$$\int_{|x|<\frac12} \left|\partial^\beta f\right|^p \,\mathrm{d}^n x \overset{(1)}{=} c\int_0^{1/2} r^{p(\lambda-|\beta|)}\, r^{n-1}\mathrm{d}r < \infty$$
$$\iff \quad p(\lambda-|\beta|)+n > 0 \quad \iff \quad \lambda > |\beta| - \frac{n}{p}.$$

(1) Polar coordinates, as in (2.161); convergence at $r = 0$ of $\int r^{a-1}\mathrm{d}r$ needs $a > 0$.

Hence $f \in L^p_k$ iff $\lambda > k - \frac{n}{p}$, while $f \in C^l$ iff $\lambda > l$ (the l-th derivative $\sim |x|^{\lambda-l}$ must stay bounded and continuous at 0). If $k - \frac{n}{p} < l$, pick λ strictly between them: $f \in L^p_k$ but $f \notin C^l$.

Worked example 2.78: The borderline case: an unbounded function of finite energy in two dimensions

When $k - \frac{n}{p} = l$ exactly, the power $|x|^\lambda$ is not subtle enough, but a double logarithm is. Take $n = 2$, $k = 1$, $p = 2$, $l = 0$ (so $k - \frac{n}{p} = 0 = l$), and

$$f(x) = \log\log\frac{1}{|x|} \quad \text{for } 0 < |x| < e^{-1}, \qquad f = 0 \text{ otherwise.}$$

f vanishes on the circle $|x| = e^{-1}$ (since $\log\log e = 0$), so it is continuous there, and it blows up (very slowly) at the origin: f is not bounded. Yet with $r = |x|$:

$$\|\nabla f\|_{L^2}^2 \overset{(1)}{=} \int_0^{1/e} \Big(\frac{1}{r\log(1/r)}\Big)^2 2\pi r\,\mathrm{d}r \overset{(2)}{=} 2\pi\int_1^\infty \frac{\mathrm{d}t}{t^2} \overset{(3)}{=} 2\pi.$$

(1) f is radial with $\partial_r f = \frac{1}{\log(1/r)}\cdot\left(-\frac1r\right)$; area element $2\pi r\,\mathrm{d}r$.
(2) Substitute $t = \log(1/r)$, $\mathrm{d}t = -\mathrm{d}r/r$; $r = 1/e \mapsto t = 1$, $r \to 0 \mapsto t \to \infty$.
(3) Elementary.

And $\int f^2 < \infty$ because $\log\log(1/r)$ grows more slowly than any power. So $f \in H^1(\mathbb{R}^2)$ but $f \notin C^0$. In physics language: in two dimensions a massless scalar with finite Dirichlet energy $\int |\nabla\phi|^2$ need not have a finite value at a point. This is the analytic root of the logarithmic infrared and ultraviolet behaviour of two-dimensional scalar fields (the propagator $-\frac{1}{2\pi}\log|x|$, the absence of spontaneous breaking of continuous symmetries, the need to exponentiate $e^{i\beta\phi}$ with normal ordering).

Watch out: Nash's example for the necessity of $k - n/p > l$

Nash suggests the function $(1 + x_1^2 + \cdots + x_n^2)^{1/2}$ as a test. As printed this cannot be right: that function is C^∞ (so it is in every C^l) and it grows at infinity (so it is in no L^p). The intended example is a function with a power singularity at a point, such as the $\chi(x)\,|x|^\lambda$ above; the missing exponent and cutoff were probably lost in typesetting.

The smoothness deficit. Nash's way of remembering the result (his (2.91)): for $\varepsilon > 0$ and $k \geq n/2$,

$$H^{k+\varepsilon} \subset C^{\,k-n/2}, \qquad \text{and more generally} \qquad L^p_{k+\varepsilon} \subset C^{\,k-n/p}. \tag{2.162}$$

(Check: the Sobolev number of $H^{k+\varepsilon}$ is $k + \varepsilon - \frac{n}{2} > k - \frac{n}{2}$.) Read it as: *an element of H^k is not C^k; it has lost $n/2$ derivatives*, the price of measuring derivatives by integrals instead of suprema. When $k - n/2$ is not an integer, read $C^{k-n/2}$ as $C^{\lfloor k-n/2\rfloor}$ (or as a Hölder space $C^{j,\gamma}$, whose functions have j derivatives that are Hölder continuous of exponent γ, with $j + \gamma = k - n/2$). Examples in one dimension: the tent function and $e^{-|x|}$ are in $H^{3/2-\varepsilon}$, and $\frac{3}{2} - \frac{1}{2} = 1$: they are "almost" C^1, and indeed they are Lipschitz ($C^{0,1}$) but not C^1.

Physical picture: Can a quantum field be evaluated at a point?

A free scalar field of mass m in d Euclidean dimensions is a random distribution ϕ with $\langle|\hat\phi(\xi)|^2\rangle \propto 1/(|\xi|^2 + m^2)$ (per unit volume). Its expected H^s norm density is therefore $\int \langle\xi\rangle^{2s}(|\xi|^2 + m^2)^{-1}\mathrm{d}^d\xi$, finite iff $2s - 2 + d < 0$: typical configurations lie in H^s_{loc} exactly for $s < 1 - \frac{d}{2}$. For $d = 1$ (quantum mechanics, $s < \frac{1}{2}$) the paths are continuous: $H^{1/2-\varepsilon} \subset C^0$ just fails to apply, but Hölder continuity of exponent $< \frac{1}{2}$ does hold, as for Brownian motion. For $d \geq 2$ the threshold $1 - \frac{d}{2} \leq 0$ lies below every Sobolev number that guarantees continuity: the field is a genuine distribution, $\phi(x)$ is meaningless, and only smeared fields $\phi(g) = \int \phi(x)g(x)\,\mathrm{d}^d x$ with test functions g make sense. This is why the Wightman axioms use operator-valued distributions, and why products $\phi(x)^2$ need renormalisation.

Sobolev (ii): embedding one Sobolev space in another. The same scaling bookkeeping predicts when $L^p_k \subset L^q_l$: by (2.157), the norm $\|\partial^\alpha f_\lambda\|_{L^q}$ with $|\alpha| = l$ scales as $\lambda^{l-n/q}$, which must not beat $\lambda^{k-n/p}$.

Key idea: Nash calls it Sobolev (ii), (2.88)

Let $M = \mathbb{R}^n$ or M compact. If

$$k \geq l \qquad \text{and} \qquad k - \frac{n}{p} \;\geq\; l - \frac{n}{q}, \tag{2.163}$$

then $L^p_k \subset L^q_l$, with $\|f\|_{l,q} \leq C\,\|f\|_{k,p}$: *you may trade derivatives for integrability at the exchange rate fixed by the Sobolev number.* On compact M, if both inequalities are strict, the inclusion

is a compact operator (the Rellich–Kondrachov theorem).

Watch out: On $\mathbb{R}^n$ one also needs $q \geq p$

Take $k = l = 0$. Then (2.163) reads $-\frac{n}{p} \geq -\frac{n}{q}$, i.e. $q \leq p$, and asserts $L^p \subset L^q$. On a compact M (finite volume) this is true, by Hölder's inequality: $\int |f|^q \leq (\int |f|^p)^{q/p} \mathrm{vol}(M)^{1-q/p}$. On $\mathbb{R}^n$ it is false: $(1+|x|)^{-n/q}$ is in L^p for $p > q$ but not in L^q. Large distances behave oppositely to small distances. On $\mathbb{R}^n$ the correct statement adds $p \leq q$ (and excludes some endpoint cases such as $q = \infty$ with equality); Nash's (2.88) is stated for both cases at once and silently means the compact case for $q < p$.

Worked example 2.79: Critical exponents and renormalisable interactions

Take $k = 1$, $p = 2$, $l = 0$ on $\mathbb{R}^d$. Then (2.163) gives $H^1 \subset L^q$ for

$$1 - \frac{d}{2} \geq -\frac{d}{q} \qquad \Longleftrightarrow \qquad q \leq q_* = \frac{2d}{d-2} \quad (d \geq 3),$$

(together with $q \geq 2$). This is the **Sobolev inequality** $\|\phi\|_{L^{q_*}} \leq C \|\nabla\phi\|_{L^2}$ (at the endpoint only the gradient term is needed, by scaling). Its physical content: *the kinetic energy* $\int |\nabla\phi|^2$ *controls the interaction* $\int |\phi|^q$ *for all* $q \leq q_*$.

d	$q_* = \frac{2d}{d-2}$	critical interaction	where it appears
3	6	ϕ^6	tricritical point; $H^1(\mathbb{R}^3) \subset L^6$ in stability of matter
4	4	ϕ^4	Higgs self-coupling; scale invariance of 4d instantons (Chapter 8)
6	3	ϕ^3	cubic scalar theory in six dimensions

These are exactly the polynomial interactions that are renormalisable by power counting: with $[\phi] = \frac{d-2}{2}$ in mass units, the coupling of ϕ^q has dimension $d - q\frac{d-2}{2} \geq 0$ iff $q \leq \frac{2d}{d-2}$. Both statements are the same dimensional analysis, (2.157). At the critical exponent the inequality is scale invariant and the embedding is *not* compact (squeezed bumps $\lambda^{(d-2)/2} f(\lambda x)$ keep their norms), which is the analytic face of the "bubbling" of instantons in four dimensions.

2.5.8 Compact operators and Rellich's theorem

Nash's footnote 9 introduces *compact operators* in two lines. They are the single most important idea in the rest of this section, so we give them room. The motivating fact is from linear algebra: in finite dimensions every bounded sequence has a convergent subsequence (Bolzano–Weierstrass). In infinite dimensions this fails. In ℓ^2 (or in $L^2(S^1)$ with the basis $e_n = e^{in\theta}/\sqrt{2\pi}$) the orthonormal vectors $e_1, e_2, e_3, \ldots$ all have norm 1, yet

$$\|e_n - e_m\|^2 = \|e_n\|^2 + \|e_m\|^2 = 2 \qquad (n \neq m),$$

so no subsequence is Cauchy. The unit ball of an infinite-dimensional Hilbert space is not compact. Compact operators are the operators that repair this.

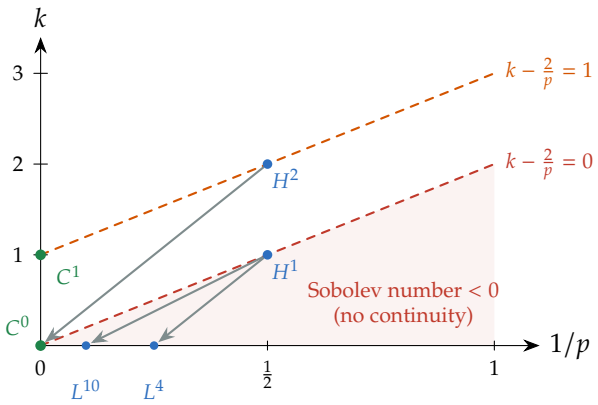


Figure 2.41: The Sobolev diagram in $n = 2$ dimensions. Each space L^p_k is a point $(1/p, k)$; C^l spaces sit on the vertical axis ($p = \infty$). The Sobolev number $k - n/p$ is constant along lines of slope $n = 2$ (dashed). Grey arrows: $H^2 \subset C^0$ (Sobolev number $1 > 0$) and $H^1 \subset L^4, L^{10}$. An embedding $L^p_k \subset L^q_l$ goes *down and to the left* without crossing to a higher Sobolev number; strictly below the dashed line through the start it is compact (on compact M). H^1 lies exactly on the Sobolev-number-0 line, which is why it embeds in every L^q, $q < \infty$, but not in C^0 (Example 2.78).

Working definition 2.38 (compact operator)

A bounded operator [$\|Tu\| \le C \|u\|$ for one constant C and all u] $T : H_1 \to H_2$ between Hilbert spaces is **compact** if it maps bounded sequences to sequences with a convergent subsequence (equivalently, the image of the unit ball has compact closure). *Tests, in order of usefulness:*

(i) T has **finite rank** (finite-dimensional image) $\Rightarrow$ compact.

(ii) T is a norm limit of finite-rank operators, $\|T - T_N\| \to 0$ [$\|T - T_N\|$ is the smallest C with $\|(T - T_N)u\| \le C \|u\|$] $\Leftrightarrow$ compact (for Hilbert spaces).

(iii) T is diagonal, $Te_n = \lambda_n f_n$ in orthonormal bases: compact $\Leftrightarrow \lambda_n \to 0$.

Test (iii) follows from (ii). Let T_N keep only the first N modes, $T_N e_n = \lambda_n f_n$ for $n \le N$ and 0 otherwise. For $u = \sum u_n e_n$,

$$\|(T - T_N)u\|^2 \overset{(1)}{=} \Big\| \sum_{n>N} \lambda_n u_n f_n \Big\|^2 \overset{(2)}{=} \sum_{n>N} |\lambda_n|^2 |u_n|^2 \overset{(3)}{\le} \Big(\sup_{n>N} |\lambda_n|\Big)^2 \sum_{n>N} |u_n|^2 \overset{(4)}{\le} \Big(\sup_{n>N} |\lambda_n|\Big)^2 \|u\|^2 . \tag{2.164}$$

(1) Only the modes beyond N survive in $T - T_N$.

(2) Pythagoras: the f_n are orthonormal.

(3) Bound each $|\lambda_n|$ by the largest one in the tail.

(4) The tail of $\sum |u_n|^2$ is at most the whole sum.

So $\|T - T_N\| \le \sup_{n>N} |\lambda_n|$ (with equality, testing on $u = e_n$), which tends to 0 iff $\lambda_n \to 0$. Conversely if $\lambda_n \not\to 0$, the images of e_n along a subsequence with $|\lambda_n| \ge c > 0$ stay $\ge c\sqrt{2}$ apart, as for the identity.

Physical picture: Compact means "effectively finite-dimensional"

A compact operator squeezes the unit ball into an "ellipsoid" whose semi-axes $|\lambda_n|$ shrink to zero (Figure 2.42). Up to any error ε, only finitely many directions survive. This is Nash's remark that a compact operator "can always be approximated by a sequence of operators with a finite dimensional range". For a physicist: a compact operator is one

that can be truncated to a finite matrix with controllable error, like keeping only the modes below a cutoff. Integral operators with square-integrable kernels ($\iint |K(x,y)|^2 < \infty$, **Hilbert–Schmidt** operators) are the typical examples; every smoothing operator on a compact manifold is one.

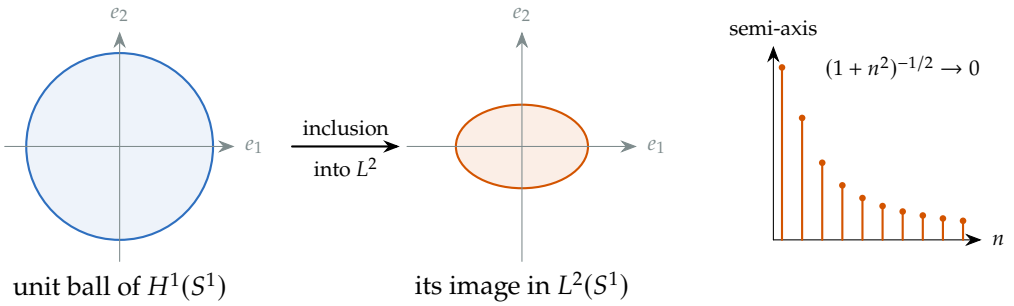


Figure 2.42: Rellich's theorem on the circle, seen as geometry. The inclusion $H^1(S^1) \hookrightarrow L^2(S^1)$ maps the unit ball of H^1 to an ellipsoid in L^2 whose semi-axis along the mode $e^{in\theta}$ is $(1+n^2)^{-1/2}$ (right). The semi-axes shrink to zero, so the image is "thin" in all but finitely many directions, up to any tolerance: the inclusion is compact. Two of the infinitely many directions are drawn.

Key idea: Rellich's theorem

On a compact manifold M and for a vector bundle $E \to M$, the inclusion

$$H^s(M,E) \hookrightarrow H^t(M,E) \qquad (s > t)$$

is a compact operator. In words: *a bounded family of functions with a little extra smoothness has a convergent subsequence in the rougher norm.* Together with Sobolev (i), $H^s \hookrightarrow C^l$ is compact for $s - \frac{n}{2} > l$.

Why, on the circle. Use the orthonormal bases $e_n^{(s)} = e^{in\theta}/\left(\sqrt{2\pi}(1+n^2)^{s/2}\right)$ of $H^s(S^1)$ and $e_n^{(t)}$ of $H^t(S^1)$ (orthonormal by the definition of $\|\cdot\|_s$ in Section 2.5.3). The inclusion ι sends

$$\iota\left(e_n^{(s)}\right) \overset{(1)}{=} \frac{e^{in\theta}}{\sqrt{2\pi}(1+n^2)^{s/2}} \overset{(2)}{=} (1+n^2)^{\frac{t-s}{2}}\, e_n^{(t)},$$

(1) Definition of $e_n^{(s)}$; the inclusion does nothing to the function.

(2) Rewrite in terms of $e_n^{(t)}$.

a diagonal operator with $\lambda_n = (1+n^2)^{(t-s)/2} \to 0$ because $t - s < 0$. By test (iii) it is compact, and by (2.164) the error of truncating at $|n| \le N$ is exactly $(1+(N+1)^2)^{(t-s)/2}$. On the torus T^n the same computation holds with $k \in \mathbb{Z}^n$. On a general compact M, a partition of unity cuts f into finitely many pieces supported in coordinate balls; each ball can be placed inside a large torus, and the torus result applies piece by piece.

Watch out: Rellich fails on $\mathbb{R}^n$ and for $s = t$

Both hypotheses are needed. *Equal orders:* the identity $H^s \to H^s$ is not compact (its λ_n are all 1). *Non-compact space:* on $\mathbb{R}^n$ take one bump f and translate it, $f_j(x) = f(x - 3j\,e_1)$. All f_j have the same H^1 norm, but they have disjoint supports if f is supported in a unit ball, so $\|f_i - f_j\|_{L^2}^2 = 2\,\|f\|_{L^2}^2$ and no subsequence converges. Mass can escape to infinity. This is the mathematical reason why a free particle on $\mathbb{R}^n$ has continuous spectrum while a

particle in a box, or on a compact manifold, has discrete spectrum (next paragraph).

Compact self-adjoint operators have discrete spectra. We need one more standard fact, the spectral theorem for compact self-adjoint operators: *if $T = T^*$ is compact on a Hilbert space, then H has an orthonormal basis of eigenvectors of T; the eigenvalues are real, accumulate only at 0, and every non-zero eigenvalue has finite multiplicity.* It is the infinite-dimensional version of diagonalising a Hermitian matrix; compactness is what guarantees the largest $|\lambda|$ is attained (maximise $|\langle u, Tu\rangle|$ over the unit ball: a maximising sequence has a convergent subsequence), after which we restrict to the orthogonal complement and repeat.

Worked example 2.80: Why the Laplacian on a compact manifold has discrete spectrum

Let M be compact and Δ the (positive) Laplacian on functions. We will see in Section 2.6.5 that $\Delta + 1 : H^2(M) \to L^2(M)$ is invertible (on the circle this is obvious: it multiplies c_n by $1 + n^2 \geq 1$). Consider

$$T = (\Delta + 1)^{-1} : \; L^2 \xrightarrow{\text{bounded}} H^2 \xrightarrow{\text{compact (Rellich)}} L^2.$$

A bounded operator followed by a compact one is compact, and T is self-adjoint and positive. So T has eigenvalues $\mu_1 \geq \mu_2 \geq \cdots > 0$ with $\mu_j \to 0$, each of finite multiplicity, and an orthonormal eigenbasis. Since $\Delta u = \lambda u \iff Tu = (1+\lambda)^{-1}u$, the Laplacian has eigenvalues $\lambda_j = \mu_j^{-1} - 1 \to +\infty$, each of finite multiplicity, with a complete set of eigenfunctions. This is the spectrum of dots on a line in Figure 2.26: discreteness is Rellich, and "running off to infinity" is compactness of the inverse. In quantum mechanics this is the statement that a particle confined to a compact region has quantised energy levels. The rate at which the λ_j grow is **Weyl's law**: on $T^2 = \mathbb{R}^2/2\pi\mathbb{Z}^2$ the number of eigenvalues $|k|^2 \leq \Lambda$ is the number of lattice points in a disc of radius $\sqrt{\Lambda}$, about $\pi\Lambda = \frac{\text{Area}(T^2)}{4\pi}\Lambda$ (the area is $4\pi^2$).

2.5.9 Pseudo-differential operators act on the Sobolev scale

Now we can say precisely what "a ΨDO of order m costs m derivatives" means.

Key idea: ΨDOs are bounded on the Sobolev scale; Nash's (2.92)

If P is a pseudo-differential operator of order m (on $\mathbb{R}^n$ with symbol in S^m satisfying the estimates uniformly in x, or on a compact manifold between bundles U and V), then for every $s \in \mathbb{R}$

$$P : \; H^s \longrightarrow H^{s-m} \quad \text{is bounded:} \qquad \|Pu\|_{s-m} \leq C_s \, \|u\|_s \, . \tag{2.165}$$

Since orders can be fractional or negative (Section 2.4), this is one reason we needed H^s for all real s: $\sqrt{\Delta+1}$ maps H^s to H^{s-1}, $|D|^{1/3}$ maps H^s to $H^{s-1/3}$, and the resolvent $(\Delta+1)^{-1}$ maps H^s *up* to H^{s+2}.

Why, for a Fourier multiplier. If $P = \sigma(D)$ with $|\sigma(\xi)| \leq C\langle\xi\rangle^m$, then

$$\|Pu\|_{s-m}^2 \overset{(1)}{=} \int \langle\xi\rangle^{2(s-m)} \, |\sigma(\xi)\hat{u}(\xi)|^2 \, \frac{\mathrm{d}^n\xi}{(2\pi)^n} \overset{(2)}{\leq} C^2 \int \langle\xi\rangle^{2(s-m)}\langle\xi\rangle^{2m} \, |\hat{u}(\xi)|^2 \, \frac{\mathrm{d}^n\xi}{(2\pi)^n} \overset{(3)}{=} C^2 \, \|u\|_s^2 \, . \tag{2.166}$$

(1) Definition (2.145), with $\widehat{Pu} = \sigma\hat{u}$.

(2) The symbol estimate (2.122) with $\alpha = \beta = 0$.
(3) The powers of $\langle\xi\rangle$ combine to $\langle\xi\rangle^{2s}$.

The best constant is $C = \sup_\xi |\sigma(\xi)|\,\langle\xi\rangle^{-m}$. Only the size of the symbol entered, not its derivatives: those are needed for composition and for pseudo-locality, not for boundedness of multipliers.

Why, with x-dependence. The new ingredient is multiplication by a smooth function $a(x)$, which is an operator of order 0 (its symbol is $a(x)$). The Fourier transform turns a product into a convolution, $\widehat{au}(\xi) = \int \hat a(\xi-\eta)\hat u(\eta)\,\frac{\mathrm{d}^n\eta}{(2\pi)^n}$, and the weights are moved through the convolution by **Peetre's inequality**

$$\langle\xi\rangle^s \le 2^{|s|/2}\,\langle\xi-\eta\rangle^{|s|}\,\langle\eta\rangle^s \qquad (s\in\mathbb{R}), \tag{2.167}$$

whose proof is two lines:

$$\langle\xi\rangle^2 = 1+|(\xi-\eta)+\eta|^2 \overset{(1)}{\le} 1+2\,|\xi-\eta|^2+2\,|\eta|^2 \overset{(2)}{\le} 2\,(1+|\xi-\eta|^2)(1+|\eta|^2) = 2\langle\xi-\eta\rangle^2\langle\eta\rangle^2.$$

(1) $|a+b|^2 \le 2\,|a|^2+2\,|b|^2$ (parallelogram law).
(2) Expand the right side: $2+2\,|\xi-\eta|^2+2\,|\eta|^2+2\,|\xi-\eta|^2\,|\eta|^2$ is larger.

Raising to the power $s/2 \ge 0$ gives (2.167) for $s \ge 0$; for $s<0$ swap the roles of ξ and η (write $\langle\eta\rangle^2 \le 2\langle\eta-\xi\rangle^2\langle\xi\rangle^2$ and raise to the power $|s|/2$). With it,

$$\langle\xi\rangle^s\,\big|\widehat{au}(\xi)\big| \overset{(1)}{\le} \int \langle\xi\rangle^s\,|\hat a(\xi-\eta)|\;|\hat u(\eta)|\,\frac{\mathrm{d}^n\eta}{(2\pi)^n} \overset{(2)}{\le} 2^{|s|/2}\int \langle\xi-\eta\rangle^{|s|}\,|\hat a(\xi-\eta)|\;\langle\eta\rangle^s\,|\hat u(\eta)|\,\frac{\mathrm{d}^n\eta}{(2\pi)^n},$$

(1) Triangle inequality for the convolution integral.
(2) Peetre (2.167).

a convolution of the L^1 function $A(\zeta) = \langle\zeta\rangle^{|s|}\,|\hat a(\zeta)|$ (integrable because $\hat a$ decays rapidly for smooth compactly supported a) with the L^2 function $\langle\eta\rangle^s\,|\hat u(\eta)|$. Young's inequality $\|A*B\|_{L^2} \le \|A\|_{L^1}\,\|B\|_{L^2}$ gives $\|au\|_s \le 2^{|s|/2}\,\|A\|_{L^1}\,\|u\|_s$ (with the $(2\pi)^{-n}$ absorbed in the normalisation of $\|A\|_{L^1}$). A general ΨDO is handled by the same idea applied to the x-Fourier transform of its symbol; the derivative estimates in x provide the decay of that transform.

Worked example 2.81: Operators on the circle and their Sobolev norms

On S^1 an operator diagonal in Fourier modes, $e^{in\theta} \mapsto p_n e^{in\theta}$, satisfies, by the same computation as (2.166),

$$\|P\|_{H^s\to H^{s-m}} = \sup_n\;|p_n|\;(1+n^2)^{-m/2}.$$

- *Hilbert transform* $p_n = -i\,\operatorname{sgn} n$ ($p_0=0$), order 0: norm 1 on every H^s.
- $|D|$, $p_n = |n|$, order 1: norm $\sup_n |n|/\sqrt{1+n^2} = 1$ as a map $H^s \to H^{s-1}$.
- $\sqrt{\Delta+\mu^2}$, $p_n = \sqrt{n^2+\mu^2}$: norm $\max(1,\mu)$ from H^s to H^{s-1}.
- The heat operator $e^{-t\Delta}$, $p_n = e^{-tn^2}$: bounded from H^s to H^{s+N} for every N, with norm $\sup_n e^{-tn^2}(1+n^2)^{N/2}$, which blows up like $(N/2et)^{N/2}$ as $t\to 0$ (maximise over n^2): the smoothing gets weaker as the Gaussian sharpens.

Smoothing operators, again (Nash's (2.93)). A smoothing operator has order $-\infty$, i.e. order $-N$ for every N. By (2.165) it maps $H^s \to H^{s+N}$ for all N, hence

$$P:\; H^s \longrightarrow H^\infty \subset C^\infty, \tag{2.168}$$

the last inclusion being Sobolev (i) applied for every l. This is the Sobolev-space proof of the claim in Section 2.4.13 that a smoothing operator "maps any function onto a smooth function", and even any distribution in $H^{-\infty}$: the heat operator turns δ into a Gaussian.

Unbounded on L^2, bounded on the scale. Nash raises an apparent paradox. The Laplacian has eigenvalues $0, 1, 4, 9, \dots$ on the circle. The norm of an operator $L : H \to H$ is

$$\|L\| = \sup_{v \neq 0} \frac{\|Lv\|}{\|v\|}, \tag{2.169}$$

and testing on eigenvectors gives $\|L\| \geq |\lambda_n|$ for every n. So as a map $L^2 \to L^2$ the Laplacian has infinite norm: it is **unbounded** (and not even defined on all of L^2). How can it then be a "bounded operator between Sobolev spaces"? Because the target is measured with two fewer derivatives.

Worked example 2.82: Nash's example: $P = -\mathrm{d}^2/\mathrm{d}\theta^2$ on the circle has norm 1 from H^k to H^{k-2}

Let $P = D_\theta^2$, so $Pe_n = n^2 e_n$ with $e_n = e^{in\theta}$. *With the Fourier norm* $\|u\|_s^2 = 2\pi \sum (1+n^2)^s |c_n|^2$ and any real k:

$$\|Pu\|_{k-2}^2 \overset{(1)}{=} 2\pi \sum_n (1+n^2)^{k-2} n^4 |c_n|^2 \overset{(2)}{=} 2\pi \sum_n \left(\frac{n^2}{1+n^2}\right)^2 (1+n^2)^k |c_n|^2$$
$$\overset{(3)}{\leq} \sup_n \left(\frac{n^2}{1+n^2}\right)^2 \cdot \|u\|_k^2 = \|u\|_k^2 .$$

(1) Pu has coefficients $n^2 c_n$.
(2) $(1+n^2)^{k-2} n^4 = \left(\frac{n^2}{1+n^2}\right)^2 (1+n^2)^k$.
(3) Bound the factor by its supremum, which is 1 (approached as $|n| \to \infty$, never attained).

And on eigenvectors $\|Pe_n\|_{k-2} / \|e_n\|_k = n^2/(1+n^2) \to 1$. So $\|P\|_{H^k \to H^{k-2}} = 1$ exactly. *With Nash's norm* $\|u\|_{k,2} = \sum_{j=0}^k \left\|D_\theta^j u\right\|$ (integer $k \geq 2$), the bound is even simpler:

$$\|Pu\|_{k-2,2} \overset{(1)}{=} \sum_{j=0}^{k-2} \left\|D_\theta^{j+2} u\right\| \overset{(2)}{=} \sum_{j=2}^{k} \left\|D_\theta^j u\right\| \overset{(3)}{\leq} \sum_{j=0}^{k} \left\|D_\theta^j u\right\| = \|u\|_{k,2} ,$$

(1) Definition of the norm applied to $Pu = D_\theta^2 u$.
(2) Shift the summation index $j \to j-2$.
(3) Add the two missing non-negative terms $j = 0, 1$.

and on e_n, where $\left\|D_\theta^j e_n\right\| = \sqrt{2\pi}\, |n|^j$, the ratio is $\sum_{j=2}^k n^j \,/\, \sum_{j=0}^k n^j \to 1$. Either way the norm is 1: finite, as befits an operator that will turn out to be Fredholm.

Watch out: A slip in Nash's (2.96), and a caution about eigenvectors

(a) In the middle expression of (2.96) Nash re-indexes $\sum_{j=0}^{k-2} \left\|D^j (D^2 e_n)\right\|$ as $\sum_{j=0}^k \left\|D^j e_n\right\|$; the correct range is $j = 2, \dots, k$, as in step (2) above. With his version the ratio would be identically 1; with the correct one it is < 1 for each n and tends to 1. The supremum is 1 either way. (b) "Take the supremum over the eigenvectors" is legitimate here only because *P and the norms* are diagonal in the same orthogonal basis $\{e_n\}$. For a general operator the norm is not the supremum over eigenvectors: the matrix $\left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right)$ has only the eigenvalue 0 but norm 1.

Dictionary: unbounded Hamiltonians and Sobolev spaces

Quantum mechanics	Sobolev language
domain of $H = -\nabla^2 + V$ (smooth bounded V)	H^2: $H : H^2 \to L^2$ bounded
finite-energy states, $\langle \psi, H\psi \rangle < \infty$	form domain H^1: $H : H^1 \to H^{-1}$ bounded
"H is unbounded"	$H : L^2 \to L^2$ has infinite norm (2.169)
resolvent $(H + c)^{-1}$	bounded $L^2 \to H^2$; compact on L^2 if space compact
point spectrum only (box, compact M)	compact resolvent (Rellich)

The elliptic estimate. Boundedness says $\|Pu\|_{s-m} \le C \|u\|_s$: P costs at most m derivatives. Ellipticity gives the reverse: P costs *exactly* m derivatives, up to a harmless lower-order error. Let Q be a parametrix (Section 2.4.9), so $QP = I - R$ with R smoothing. For any u in some H^t, $t < s$:

$$\|u\|_s \overset{(1)}{=} \|QPu + Ru\|_s \overset{(2)}{\le} \|Q(Pu)\|_s + \|Ru\|_s \overset{(3)}{\le} C_1 \|Pu\|_{s-m} + C_2 \|u\|_t . \tag{2.170}$$

(1) $u = QPu + Ru$ from $QP = I - R$.
(2) Triangle inequality.
(3) Q has order $-m$, so it is bounded $H^{s-m} \to H^s$ by (2.165); R is smoothing, so it is bounded $H^t \to H^s$ for any t.

This is the **elliptic estimate** (in the Laplacian case, Gårding's inequality). Two immediate consequences:

- *Elliptic regularity in Sobolev form:* if $u \in H^t$ and $Pu \in H^{s-m}$, then $u \in H^s$. (The right side of (2.170) is finite.) Taking $s \to \infty$: $Pu \in C^\infty \Rightarrow u \in C^\infty$, which is hypo-ellipticity once more.
- *Zero modes are smooth:* if $Pu = 0$ with $u \in H^t$ for some t, then $u \in H^s$ for every s, i.e. $u \in C^\infty$. In particular $\operatorname{Ker} P$ does not depend on which Sobolev space we let P act on.

Worked example 2.83: The elliptic estimate for $D^2 + 1$ on the circle, with constant 1

For $P = D_\theta^2 + 1$ the Fourier norms make (2.170) an identity:

$$\left\|(D_\theta^2 + 1)u\right\|_{s-2}^2 = 2\pi \sum_n (1 + n^2)^{s-2}(1 + n^2)^2 |c_n|^2 = \|u\|_s^2 .$$

P is an isometry $H^s \to H^{s-2}$, invertible, with no error term needed. For $P = D_\theta^2$ (no mass) the estimate needs the error term: the constants are in the kernel and $\left\|D_\theta^2 \cdot 1\right\|_{s-2} = 0$ while $\|1\|_s \neq 0$. Split off the $n = 0$ mode:

$$\|u\|_s^2 \overset{(1)}{=} 2\pi |c_0|^2 + 2\pi \sum_{n \neq 0} (1 + n^2)^s |c_n|^2 \overset{(2)}{\le} 2\pi |c_0|^2 + 4 \cdot 2\pi \sum_{n \neq 0} (1 + n^2)^{s-2} n^4 |c_n|^2$$
$$\overset{(3)}{\le} \|u\|_t^2 + 4 \left\|D_\theta^2 u\right\|_{s-2}^2 .$$

(1) Definition of the norm; the $n = 0$ weight is 1.
(2) For $n \neq 0$, $n^4/(1 + n^2)^2 \ge \frac{1}{4}$ (the worst case is $n = \pm 1$), so $(1 + n^2)^s \le 4(1 + n^2)^{s-2} n^4$.
(3) $2\pi |c_0|^2 \le \|u\|_t^2$ for any t, and the sum is $\left\|D_\theta^2 u\right\|_{s-2}^2$ (the $n = 0$ term vanishes anyway).

So $\|u\|_s \le 2\left\|D_\theta^2 u\right\|_{s-2} + \|u\|_t$: the estimate holds, but a single low mode must be controlled by the weak norm, and that mode is exactly the kernel. This is the Fredholm theory of the next section in miniature.

Reading Nash (pp. 47–54, (2.71)–(2.96))

Nash's §5 up to (2.96) maps onto these notes as follows: (2.71)–(2.74) are Section 2.5.1; (2.75)–(2.77) are (2.145) and (2.147); (2.78)–(2.80) are (2.149) and (2.150); (2.81)–(2.86) are Sections 2.5.5 and 2.5.6; (2.87)–(2.91) are (2.158)–(2.162); (2.92)–(2.93) are (2.165) and (2.168); (2.94)–(2.96) are Example 2.82. Nash's H_k in the proof of Sobolev (i) is his H^k. His remark that $L^p_{k+\varepsilon} \subset C^{k-n/p}$ is (2.162). Rellich's theorem is implicit in his footnote 9 and his claim that the inclusions are compact; we have made it explicit because it is what the Fredholm property rests on.

Recap: Pseudo-differential operators and Sobolev spaces

- H^s measures "s derivatives in L^2" by the Fourier weight: $\|u\|_s^2 = \int \langle\xi\rangle^{2s} |\hat{u}|^2 \frac{d^n\xi}{(2\pi)^n}$; on S^1 it is $2\pi \sum (1+n^2)^s |c_n|^2$, a statement about how fast Fourier coefficients decay. For integer k it agrees with "all derivatives up to order k in L^2".
- Negative orders are duals, $H^{-s} = (H^s)^*$; $\delta \in H^{-s}$ iff $s > \frac{n}{2}$. The chain $\mathcal{S} \subset \cdots \subset H^s \subset \cdots \subset L^2 \subset \cdots \subset H^{-s} \subset \cdots \subset \mathcal{S}'$ is the Sobolev scale; L^2 in the middle is the physicist's Hilbert space of kets.
- Sobolev (i): $H^s \subset C^l$ if $s - \frac{n}{2} > l$ (power counting of $\int \langle\xi\rangle^{2l-2s}$); functions in H^s have a smoothness deficit $n/2$. Sobolev (ii) compares two L^p_k spaces by the Sobolev number $k - n/p$.
- Rellich: on a compact manifold $H^s \hookrightarrow H^t$ $(s > t)$ is compact, because the ratio of weights $(1+n^2)^{(t-s)/2}$ tends to 0. This is why elliptic operators on compact spaces have discrete spectra.
- A ΨDO of order m is bounded $H^s \to H^{s-m}$ for every s; smoothing operators map everything into $H^\infty \subset C^\infty$. The Laplacian is unbounded on L^2 but has norm 1 as a map $H^k \to H^{k-2}$ on the circle.
- Elliptic estimate: $\|u\|_s \le C\,\|Pu\|_{s-m} + C\,\|u\|_t$. Elliptic regularity and smoothness of zero modes follow at once.

2.6 Elliptic operators are Fredholm

Here are the two questions every physicist asks about a linear equation $Pu = f$.

(1) *How many solutions does the homogeneous equation $Pu = 0$ have?* These are the zero modes: the harmonic forms of Hodge theory, the fermion zero modes in an instanton background, the flat directions of a moduli space.

(2) *For which sources f can we solve $Pu = f$ at all?* On a compact space there are often constraints. The simplest is Gauss's law: on a closed universe the total charge must vanish.

In finite dimensions, linear algebra answers both at once: for a matrix, (number of zero modes) − (number of constraints) is fixed by the sizes of the matrix, whatever its entries. Nash's last

page of the chapter says that elliptic operators on compact manifolds behave in the same way: they have finitely many zero modes and finitely many constraints, and the difference, the **index**, is robust. He says it in one paragraph and two equations, (2.96)–(2.97). In this section we unpack that paragraph. Every tool of the previous two sections gets used: the parametrix, the smoothing operators, Rellich's compactness theorem and the boundedness of ΨDOs on the Sobolev scale.

2.6.1 Warm-up: the index of a matrix

Let $A : \mathbb{C}^n \to \mathbb{C}^m$ be an $m \times n$ matrix. Two subspaces measure how far it is from being invertible.

- The **kernel** $\operatorname{Ker} A = \{u : Au = 0\} \subset \mathbb{C}^n$: the solutions of the homogeneous equation.
- The **cokernel** $\operatorname{Coker} A = \mathbb{C}^m / \operatorname{Im} A$: the directions in the target that A cannot reach. Its dimension is the number of independent linear conditions that f must satisfy for $Au = f$ to be solvable.

The **rank–nullity theorem** says $\dim \operatorname{Ker} A + \dim \operatorname{Im} A = n$: the domain splits into the part A kills and a part that A maps isomorphically onto its image. Hence

$$\begin{aligned} \dim \operatorname{Ker} A - \dim \operatorname{Coker} A &\overset{(1)}{=} \dim \operatorname{Ker} A - \big(m - \dim \operatorname{Im} A\big) \\ &\overset{(2)}{=} \big(\dim \operatorname{Ker} A + \dim \operatorname{Im} A\big) - m \\ &\overset{(3)}{=} n - m. \end{aligned} \tag{2.171}$$

(1) The dimension of a quotient is the dimension of the big space minus that of the subspace.
(2) Regroup.
(3) Rank–nullity.

The left side depends on the entries of A; the right side does not. That is the whole phenomenon of this section in its simplest form.

The cokernel is the kernel of the adjoint. With the standard inner products, a vector $f \in \mathbb{C}^m$ is orthogonal to the image of A if and only if $A^\dagger f = 0$:

$$f \perp \operatorname{Im} A \iff \langle f, Au\rangle = 0 \ \forall u \overset{(1)}{\iff} \langle A^\dagger f, u\rangle = 0 \ \forall u \iff A^\dagger f = 0. \tag{2.172}$$

(1) Definition of the adjoint, $\langle f, Au\rangle = \langle A^\dagger f, u\rangle$.

(The last step: take $u = A^\dagger f$.) So $\operatorname{Coker} A \cong (\operatorname{Im} A)^\perp = \operatorname{Ker} A^\dagger$, and

$$\operatorname{index} A = \dim \operatorname{Ker} A - \dim \operatorname{Ker} A^\dagger.$$

This version is the one that generalises: it counts zero modes of A minus zero modes of $A^\dagger$, and $Au = f$ is solvable exactly when f is orthogonal to every zero mode of $A^\dagger$ (the **Fredholm alternative** in finite dimensions).

Worked example 2.84: The kernel jumps, the index does not

Take the 2×3 family

$$A_\varepsilon = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 6+\varepsilon \end{pmatrix} : \mathbb{C}^3 \to \mathbb{C}^2.$$

At $\varepsilon = 0$ the second row is twice the first, so the rank is 1. The kernel is the plane $u_1 + 2u_2 + 3u_3 = 0$, of dimension 2. The image is the line spanned by $(1,2)^T$, so the cokernel has dimension $2-1=1$; concretely $\operatorname{Ker} A_0^\dagger$ is spanned by $(2,-1)^T$, and $A_0 u = f$ is solvable iff $2f_1 - f_2 = 0$. Index: $2-1=1$.
At $\varepsilon \neq 0$ the rows are independent, the rank is 2, the kernel is the line through the vector

$$v = (1,2,3)\times(2,4,6+\varepsilon) \overset{(1)}{=} (2\varepsilon,\ -\varepsilon,\ 0) \ \propto\ (2,-1,0),$$

(1) The cross product of the rows is orthogonal to both rows, hence in the kernel: $\big(2(6+\varepsilon)-3\cdot 4,\ 3\cdot 2 - 1\cdot(6+\varepsilon),\ 1\cdot 4 - 2\cdot 2\big)$.

and the cokernel is 0. Index: $1-0=1$. As $\varepsilon \to 0$ one kernel vector and one cokernel vector appear *together*; their difference never changes, and equals $n-m=3-2$ as (2.171) demands.

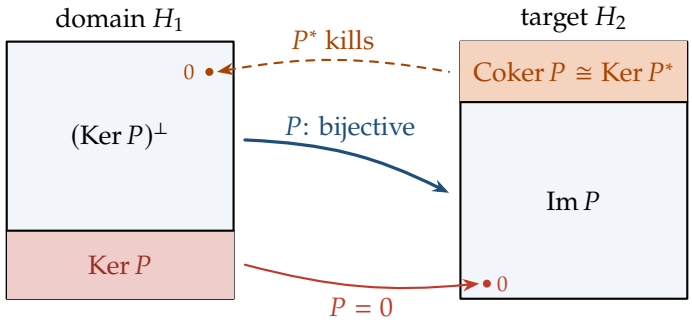


Figure 2.43: What a Fredholm operator looks like. The domain splits into the zero modes Ker P (red, finite-dimensional) and their orthogonal complement, which P maps bijectively onto its image. The image is closed, and its orthogonal complement (orange) is finite-dimensional: it is the cokernel, and it consists of the zero modes of the adjoint. The index is (red dimension) − (orange dimension). For a matrix this is $n-m$; for an operator on infinite-dimensional spaces it is a number that the sizes alone do not determine.

Watch out: A square matrix always has index 0; an operator need not

For $A : \mathbb{C}^n \to \mathbb{C}^n$, (2.171) gives index 0: a square matrix has as many zero modes as constraints. In infinite dimensions "$n-m$" reads $\infty - \infty$, and an operator from a Hilbert space to itself can have any integer index (Section 2.6.3). This is not a pathology; it is exactly why the index of an elliptic operator carries information.

2.6.2 Kernel, cokernel and the adjoint in Hilbert space

Now let $T : H_1 \to H_2$ be a bounded operator between Hilbert spaces (for us: $P : H^s(M,U) \to H^{s-m}(M,V)$). Its kernel Ker T is a closed subspace, because T is continuous. Its **cokernel** is again the quotient Coker $T = H_2/\operatorname{Im} T$. The computation (2.172) goes through word for word and gives

$$(\operatorname{Im} T)^\perp = \operatorname{Ker} T^*, \tag{2.173}$$

where $T^* : H_2 \to H_1$ is the Hilbert-space adjoint. But there is a new subtlety. $(\operatorname{Im} T)^\perp$ only sees the *closure* of the image, because $(\operatorname{Im} T)^\perp = (\overline{\operatorname{Im} T})^\perp$. So

$$\operatorname{Coker} T \cong \operatorname{Ker} T^* \qquad \text{only when } \operatorname{Im} T \text{ is closed.}$$

Worked example 2.85: An operator whose range is not closed

On ℓ^2 (sequences $u = (u_1, u_2, \dots)$ with $\sum |u_n|^2 < \infty$, or equivalently $L^2(S^1)$ through Fourier coefficients) let T be diagonal with $Te_n = \frac{1}{n} e_n$.

- $\operatorname{Ker} T = 0$: if $u_n/n = 0$ for all n then $u = 0$.
- $T = T^*$, so $\operatorname{Ker} T^* = 0$ too, and $(\operatorname{Im} T)^\perp = 0$: the image is dense.
- But T is not onto. For $f = (1, \frac{1}{2}, \frac{1}{3}, \dots) \in \ell^2$ the only candidate solution is $u_n = n f_n = 1$ for all n, which is not in ℓ^2.

In fact $f^{(a)} = (n^{-a})_n$ for $\frac{1}{2} < a \le 1$ gives infinitely many independent elements of ℓ^2 outside the image (no finite combination of them lies in $\operatorname{Im} T$), so $H_2/\operatorname{Im} T$ is infinite-dimensional even though $\operatorname{Ker} T^* = 0$. Such a T is compact (its eigenvalues $1/n \to 0$, test (iii) of Section 2.5.8), and this is typical: *a compact operator on an infinite-dimensional space never has closed range of finite codimension*. Compact operators are the opposite of Fredholm ones.

Working definition 2.39 (Fredholm operator, index)

A bounded operator $T : H_1 \to H_2$ is **Fredholm** if

$$\dim \operatorname{Ker} T < \infty \qquad \text{and} \qquad \dim \operatorname{Coker} T < \infty.$$

Then $\operatorname{Im} T$ is automatically closed, $\operatorname{Coker} T \cong \operatorname{Ker} T^*$ [T^* is the adjoint, $\langle Tu, v\rangle = \langle u, T^*v\rangle$. Each independent v with $T^*v = 0$ is one obstruction to solving $Tu = f$], and the **index** is the integer

$$\operatorname{index} T = \dim \operatorname{Ker} T - \dim \operatorname{Coker} T = \dim \operatorname{Ker} T - \dim \operatorname{Ker} T^*. \tag{2.174}$$

Operational test: count the zero modes [independent nonzero solutions of $Tu = 0$, respectively $T^*v = 0$] of T and of T^*; check that both counts are finite and that T is bounded below on the complement of its kernel, $\|Tu\| \ge c\,\|u\|$ for $u \perp \operatorname{Ker} T$ (this is closed range).

Why the image is automatically closed. Pick a finite-dimensional subspace $C \subset H_2$ complementary to $\operatorname{Im} T$ (possible because the quotient is finite-dimensional). The map $(u, c) \mapsto Tu + c$ from $(\operatorname{Ker} T)^\perp \oplus C$ to H_2 is bounded, injective and onto. By the open mapping theorem (a bounded bijection between Hilbert spaces has a bounded inverse) it is an isomorphism, so it maps the closed subspace $(\operatorname{Ker} T)^\perp \oplus 0$ onto a closed subspace, which is $\operatorname{Im} T$. Nash lists only the two dimension conditions; this remark is why that is enough.

Why "bounded below" is the same as closed range. If $\|Tu\| \ge c\,\|u\|$ on $(\operatorname{Ker} T)^\perp$ and $Tu_j \to f$ with $u_j \perp \operatorname{Ker} T$, then

$$\|u_i - u_j\| \overset{(1)}{\le} \frac{1}{c}\,\|T(u_i - u_j)\| = \frac{1}{c}\,\|Tu_i - Tu_j\| \longrightarrow 0, \tag{2.175}$$

(1) The lower bound, applied to $u_i - u_j \perp \operatorname{Ker} T$.

so (u_j) is Cauchy, converges to some u, and $Tu = f$ by continuity: $f \in \operatorname{Im} T$. The diagonal T of Example 2.85 fails exactly here: $\|Te_n\| = \frac{1}{n}\,\|e_n\|$, so no $c > 0$ works.

2.6.3 A gallery on ℓ^2: shifts and the Hilbert hotel

The simplest operators with non-zero index are shifts. Work in ℓ^2 with orthonormal basis $e_0, e_1, e_2, \dots$ (think of $e_n = e^{in\theta}$, $n \geq 0$, on the circle: this is the **Hardy space** of boundary values of holomorphic functions on the disc). Define the **unilateral shift**

$$Se_n = e_{n+1}, \qquad S\Big(\sum_{n\geq 0} u_n e_n\Big) = \sum_{n\geq 0} u_n e_{n+1}.$$

On the circle, S is multiplication by $e^{i\theta}$. Its adjoint moves the other way:

$$\langle e_m, S^* e_n\rangle \overset{(1)}{=} \langle Se_m, e_n\rangle \overset{(2)}{=} \langle e_{m+1}, e_n\rangle = \delta_{m+1,n} \quad\Longrightarrow\quad S^* e_n = e_{n-1}\ (n \geq 1),\ \ S^* e_0 = 0. \tag{2.176}$$

(1) Definition of the adjoint (with $\langle\cdot,\cdot\rangle$ conjugate-linear in the first slot, and real matrix elements here, so the order is harmless).
(2) Definition of S.

Now count.

- $\operatorname{Ker} S = 0$: S preserves norms, $\|Su\| = \|u\|$.
- $\operatorname{Im} S = \{u : u_0 = 0\} = e_0^{\perp}$, closed, with cokernel $\operatorname{span}(e_0) = \operatorname{Ker} S^*$.
- So $\operatorname{index} S = 0 - 1 = -1$, and $\operatorname{index} S^* = 1 - 0 = +1$.
- S^k shifts by k: $\operatorname{Ker} S^k = 0$, $\operatorname{Coker} S^k = \operatorname{span}(e_0, \dots, e_{k-1})$, $\operatorname{index} S^k = -k$.

Also $S^*S = I$ while $SS^* = I - \Pi_0$ with Π_0 the projection onto e_0: S is invertible *modulo a rank-one operator*, which is the Fredholm property in its purest form.

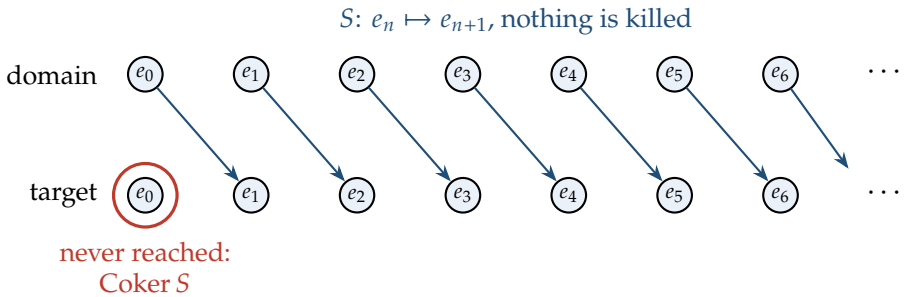


Figure 2.44: The unilateral shift, drawn as a map from the basis of the domain (top) to the basis of the target (bottom). Every basis vector is sent to a distinct basis vector, so the kernel is zero; but e_0 in the target is never hit, so the cokernel is one-dimensional and $\operatorname{index} S = -1$. This is Hilbert's hotel: every guest moves up one room, and room 0 becomes free. A matrix cannot do this, because a finite hotel has a last room.

Worked example 2.86: A table of operators on ℓ^2: Fredholm or not?

operator	dim Ker	dim Coker	index	comment
I	0	0	0	invertible
S	0	1	-1	isometry, not onto
S^*	1	0	$+1$	onto, kills e_0
S^k, $(S^*)^k$	$0, k$	$k, 0$	$\mp k$	every integer occurs
$I - \Pi_0$	1	1	0	projection of finite corank
Π_{even}	∞	∞	–	kills all odd modes: not Fredholm
$e_n \mapsto \frac{1}{n} e_n$	0	∞	–	compact; range dense, not closed
$e_n \mapsto (1 + \frac{1}{n}) e_n$	0	0	0	I plus a compact operator

For the last line: the inverse is $e_n \mapsto \frac{n}{n+1} e_n$, bounded by 1. It is the identity plus the compact diagonal operator $e_n \mapsto \frac{1}{n} e_n$, an example of the general rule "Fredholm plus compact is Fredholm, with the same index" (Section 2.6.6).

Physical picture: The shift is the Dirac sea moving by one level

In second quantisation the one-particle states are labelled by energy levels, and the vacuum fills every negative level. If an external field makes the levels flow so that each one moves up by one step, the new vacuum differs from the old one by one particle: one state has crossed zero. The operator that relabels the levels is a shift, and its index (± 1) is the charge created. This "spectral flow equals index" mechanism is how anomalies appear in Chapter 10; we see a concrete instance on the circle in Figure 2.47.

2.6.4 Fredholm means invertible modulo compact operators

Nash states: "A further well known property of Fredholm operators is that they are invertible modulo a compact operator." This is **Atkinson's theorem**, and it is the bridge between the parametrix of Section 2.4.9 and the Fredholm property.

Key idea: Atkinson's theorem; Nash's (2.97)

A bounded operator $T : H_1 \to H_2$ is Fredholm if and only if there is a bounded $Q : H_2 \to H_1$ (a **parametrix**) and compact operators K_1 on H_2 and K_2 on H_1 with

$$TQ = I_{H_2} - K_1, \qquad QT = I_{H_1} - K_2. \tag{2.177}$$

Moreover Q is then itself Fredholm, with $\operatorname{index} Q = -\operatorname{index} T$.

From Fredholm to a parametrix (with finite-rank errors). Suppose T is Fredholm. By Figure 2.43, T restricts to a bounded bijection $T_0 : (\operatorname{Ker} T)^\perp \to \operatorname{Im} T$ between closed subspaces, whose inverse is bounded (open mapping theorem again). Let Π_{Ker} be the orthogonal projection onto $\operatorname{Ker} T$ and Π_{Coker} the projection onto $(\operatorname{Im} T)^\perp$. Define $Q = T_0^{-1}(I - \Pi_{\text{Coker}})$: project onto the image, then invert. Then for $u \in H_1$,

$$QTu \overset{(1)}{=} T_0^{-1} T u \overset{(2)}{=} T_0^{-1} T_0 (I - \Pi_{\text{Ker}}) u = (I - \Pi_{\text{Ker}}) u, \tag{2.178}$$

(1) Tu already lies in $\operatorname{Im} T$, so the projection $I - \Pi_{\text{Coker}}$ does nothing to it.

(2) Split $u = \Pi_{\mathrm{Ker}} u + (I - \Pi_{\mathrm{Ker}})u$; T kills the first piece and acts as T_0 on the second.

and similarly $TQ = I - \Pi_{\mathrm{Coker}}$. So (2.177) holds with $K_2 = \Pi_{\mathrm{Ker}}$ and $K_1 = \Pi_{\mathrm{Coker}}$, which are *finite rank*, hence compact. We met this for $P = D_\theta^2$ on the circle in Example 2.59, where both errors were the projection Π_0 onto the constants.

From a parametrix to Fredholm. Now suppose (2.177) holds with K_1, K_2 compact. Three things must be checked.

(a) The kernel is finite-dimensional. If $Tu = 0$ then $0 = QTu = u - K_2 u$, so K_2 acts as the identity on $\operatorname{Ker} T$. Let B be the closed unit ball of $\operatorname{Ker} T$. Then

$$B \overset{(1)}{=} K_2(B) \overset{(2)}{\subset} K_2(\text{unit ball of } H_1),$$

(1) $K_2 u = u$ for every $u \in \operatorname{Ker} T$.
(2) B is part of the unit ball of H_1.

and the right side has compact closure because K_2 is compact. So B is a closed subset of a compact set, hence compact. But the unit ball of an infinite-dimensional Hilbert space is never compact: an orthonormal sequence has $\|e_i - e_j\| = \sqrt{2}$ for all $i \neq j$ (Section 2.5.8). Therefore $\dim \operatorname{Ker} T < \infty$.

(b) T is bounded below on $(\operatorname{Ker} T)^\perp$, so its image is closed. Suppose not. Then there are $u_j \perp \operatorname{Ker} T$ with $\|u_j\| = 1$ and $Tu_j \to 0$. By compactness of K_2, a subsequence has $K_2 u_j \to v$. Along it,

$$u_j \overset{(1)}{=} QTu_j + K_2 u_j \longrightarrow Q(0) + v = v,$$

(1) $u = QTu + K_2 u$ is (2.177) rearranged; Q is bounded, so $QTu_j \to 0$.

so $u_j \to v$ in norm. Then $\|v\| = 1$, $v \perp \operatorname{Ker} T$ (a closed condition), and $Tv = \lim Tu_j = 0$, so $v \in \operatorname{Ker} T$. A unit vector cannot be both in $\operatorname{Ker} T$ and orthogonal to it: contradiction. By (2.175) the image is closed.

(c) The cokernel is finite-dimensional. Take adjoints in $TQ = I - K_1$: $Q^*T^* = I - K_1^*$, and K_1^* is compact (adjoints of compact operators are compact). Step (a) applied to T^* with parametrix Q^* gives $\dim \operatorname{Ker} T^* < \infty$. Since the image is closed, $\operatorname{Coker} T \cong \operatorname{Ker} T^*$ by (2.173). Done.

Physical picture: Why "modulo compact" is the right notion of "almost invertible"

A compact operator is, up to any ε, a finite matrix (Section 2.5.8). "Invertible modulo compact" therefore means "invertible except on an effectively finite-dimensional piece". Whatever goes wrong goes wrong in finitely many dimensions, and in finitely many dimensions rank–nullity (2.171) controls the count. The Fredholm index is the infinite-dimensional shadow of $n - m$.

Worked example 2.87: Fredholm's integral equation on the circle

Fredholm introduced these ideas in 1903 for integral equations of the second kind, $u - \lambda K u = f$. On S^1 take the smooth kernel $K(\theta, \theta') = \frac{1}{\pi}\cos(\theta - \theta')$:

$$(Ku)(\theta) = \frac{1}{\pi}\int_0^{2\pi} \cos(\theta - \theta')\, u(\theta')\, \mathrm{d}\theta'.$$

Act on a Fourier mode:

$$Ke^{in\theta} \overset{(1)}{=} \frac{1}{2\pi}\int_0^{2\pi} \left(e^{i(\theta-\theta')} + e^{-i(\theta-\theta')}\right) e^{in\theta'}\, \mathrm{d}\theta' \overset{(2)}{=} \frac{1}{2\pi}\left(e^{i\theta}\cdot 2\pi\,\delta_{n,1} + e^{-i\theta}\cdot 2\pi\,\delta_{n,-1}\right)$$

$$= \begin{cases} e^{in\theta}, & n = \pm 1, \\ 0, & \text{otherwise.} \end{cases}$$

(1) $\cos x = \frac{1}{2}(e^{ix} + e^{-ix})$.

(2) $\int_0^{2\pi} e^{i(n-1)\theta'} \mathrm{d}\theta' = 2\pi\delta_{n,1}$, and similarly for the other term.

So K is the orthogonal projection onto $\operatorname{span}(e^{i\theta}, e^{-i\theta})$: finite rank, smoothing, compact. The operator $I - \lambda K$ multiplies $e^{\pm i\theta}$ by $1 - \lambda$ and every other mode by 1.

- $\lambda \neq 1$: $I - \lambda K$ is invertible, with $(I - \lambda K)^{-1} = I + \frac{\lambda}{1-\lambda} K$ (check: $K^2 = K$). Index 0.
- $\lambda = 1$: $\operatorname{Ker}(I - K) = \operatorname{span}(\cos\theta, \sin\theta)$, and since $I - K$ is self-adjoint, the cokernel is the same space. The equation $u - Ku = f$ is solvable iff $\int f \cos\theta = \int f \sin\theta = 0$, and then the solution is unique up to adding $a\cos\theta + b\sin\theta$. Index $2 - 2 = 0$.

This is the classical **Fredholm alternative**: *either* $u - \lambda K u = f$ has a unique solution for every f, *or* the homogeneous equation has finitely many solutions and f must satisfy the same finite number of orthogonality conditions. In both cases the index of $I - \lambda K$ is 0; we will see in Section 2.6.6 that this is true for every compact K.

Watch out: Nash's (2.97) has its identities swapped

Nash writes, for $P : H^k \to H^{k-m}$ with parametrix $Q : H^{k-m} \to H^k$, "$P \circ Q_1 - K_1 = I_{H^k}$ and $Q \circ P - K_2 = I_{H^{k-m}}$". Follow the arrows: $Q \circ P$ starts and ends in H^k, while $P \circ Q$ starts and ends in H^{k-m}. The correct statement is

$$P \circ Q = I_{H^{k-m}} + K_1, \qquad Q \circ P = I_{H^k} + K_2,$$

with K_1 compact on H^{k-m} and K_2 compact on H^k (the "Q_1" is a misprint for Q). The sign in front of K_i is immaterial, since $-K$ is compact when K is. In (2.177) we write $I - K$ to match Section 2.4.9.

2.6.5 Elliptic operators on compact manifolds are Fredholm

We can now assemble Nash's claim. Throughout, M is a compact manifold without boundary, $U, V \to M$ are vector bundles of the same rank, and P is an elliptic ΨDO of order m from sections of U to sections of V (a differential operator is the special case of polynomial symbol).

Key idea: Elliptic ⇒ Fredholm (Nash, p. 54)

For every real s:

(i) $P : H^s(M, U) \to H^{s-m}(M, V)$ is a Fredholm operator.

(ii) Its kernel consists of smooth sections and does not depend on s.

(iii) Its cokernel is isomorphic to $\operatorname{Ker} P^*$, where P^* is the formal adjoint (again elliptic); these are smooth sections too. Concretely, for $f \in H^{s-m}$:

$$Pu = f \text{ has a solution} \iff \int_M \langle v, f \rangle \,\mathrm{dvol} = 0 \ \text{ for every } v \in \operatorname{Ker} P^*. \tag{2.179}$$

(iv) Hence the index $\operatorname{index} P = \dim \operatorname{Ker} P - \dim \operatorname{Ker} P^*$ is a finite integer, the same for all s.

The mechanism: parametrix plus Rellich. By Section 2.4.9, P has a parametrix Q of order $-m$ with

$$QP = I - R_2, \qquad PQ = I - R_1, \qquad R_1, R_2 \text{ smoothing.}$$

By (2.165), Q is bounded $H^{s-m} \to H^s$. The errors are compact on the relevant Sobolev spaces, because

$$R_2 : H^s \xrightarrow[\text{bounded}]{\text{(2.165), order } -1} H^{s+1} \xrightarrow[\text{compact}]{\text{Rellich}} H^s, \tag{2.180}$$

and a bounded operator followed by a compact one is compact: a bounded sequence u_j in H^s goes to a bounded sequence $R_2 u_j$ in H^{s+1}, which has a subsequence converging in H^s. (We used only that a smoothing operator has order ≤ -1; any negative order would do.) The same holds for R_1 on H^{s-m}. Atkinson's theorem (2.177) now gives (i). This is exactly Nash's sentence "this is nothing other than the Sobolev space version of the property that elliptic operators possess of being invertible modulo a smoothing operator": *smoothing becomes compact* once we measure in Sobolev norms on a compact space.

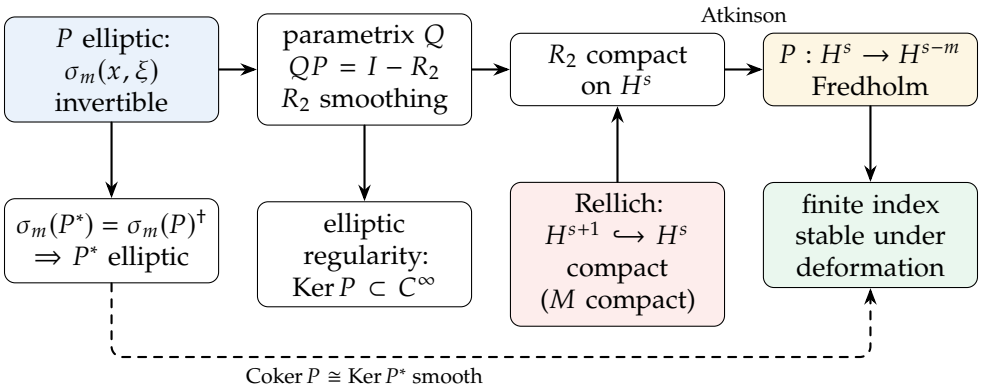


Figure 2.45: The logic of "elliptic implies Fredholm". The top row is the proof: invert the principal symbol, build a parametrix, observe that the smoothing error is compact because of Rellich's theorem (which needs M compact), and apply Atkinson's theorem. The bottom row gives the extra information: zero modes are smooth, the cokernel is the kernel of the (elliptic) adjoint, and the index is a finite integer that survives deformations.

The adjoint is elliptic. Nash's reason: "the leading symbol of P^* is the adjoint of that of P". In formulas, $\sigma_m(P^*)(x,\xi) = \sigma_m(P)(x,\xi)^\dagger$ (the first term of (2.129)). If $\sigma = \sigma_m(P)(x,\xi)$ is an invertible matrix, then

$$\sigma^\dagger (\sigma^{-1})^\dagger \overset{(1)}{=} (\sigma^{-1}\sigma)^\dagger = I^\dagger = I,$$

(1) $(AB)^\dagger = B^\dagger A^\dagger$.

and similarly in the other order, so $\sigma^\dagger$ is invertible with inverse $(\sigma^{-1})^\dagger$. Hence P^* is elliptic of order m, from V to U, and everything we say about P applies to it.

Why the cokernel is $\operatorname{Ker} P^*$, on every Sobolev space. There is a subtlety: the Hilbert-space adjoint of $P : H^s \to H^{s-m}$ is taken with respect to the H^s and H^{s-m} inner products, not the L^2 one, so it is not literally the differential (or pseudo-differential) operator P^*. The clean way round is duality. The L^2 pairing identifies the dual of H^{s-m} with H^{m-s} (Section 2.5.4). A closed subspace is the set of vectors on which every continuous functional vanishing on the subspace

also vanishes (Hahn–Banach, or just orthogonal projection in a Hilbert space). So, for $f \in H^{s-m}$,

$$\begin{aligned} f \in \operatorname{Im} P &\overset{(1)}{=}\Longleftrightarrow \ \langle v, f\rangle_{L^2} = 0 \ \text{ for all } v \in H^{m-s} \text{ with } \langle v, Pu\rangle_{L^2} = 0 \ \forall u \in H^s \\ &\overset{(2)}{=}\Longleftrightarrow \ \langle v, f\rangle_{L^2} = 0 \ \text{ for all } v \in H^{m-s} \text{ with } P^*v = 0 \\ &\overset{(3)}{=}\Longleftrightarrow \ \langle v, f\rangle_{L^2} = 0 \ \text{ for all } v \in \operatorname{Ker} P^* \subset C^\infty. \end{aligned} \tag{2.181}$$

(1) $\operatorname{Im} P$ is closed by (i); describe it by the functionals $v \in H^{m-s} = (H^{s-m})^*$ that annihilate it.
(2) $\langle v, Pu\rangle = \langle P^*v, u\rangle$ for all smooth u is the definition of P^*v as a distribution; vanishing for all u means $P^*v = 0$.
(3) P^* is elliptic, so its distributional zero modes are smooth (the elliptic estimate (2.170)), and $\operatorname{Ker} P^*$ is the same finite-dimensional space of smooth sections whichever Sobolev space we start from.

This is (2.179): the constraints on the source are "orthogonality to the zero modes of the adjoint", measured in the ordinary L^2 inner product. The number of independent constraints is $\dim \operatorname{Ker} P^*$, the same for every s. With (ii), the index is independent of s.

Worked example 2.88: $\Delta + 1$ is invertible on any compact manifold

We promised this in Example 2.80. Let $\Delta = \mathrm{d}^*\mathrm{d}$ be the positive Laplacian on functions on a compact Riemannian M, and $P = \Delta + 1 : H^{s+2} \to H^s$. It is elliptic ($\sigma_2 = |\xi|^2$), so Fredholm. Its kernel is zero: if $Pu = 0$ then u is smooth and

$$0 = \langle u, (\Delta + 1)u\rangle \overset{(1)}{=} \langle u, \mathrm{d}^*\mathrm{d}u\rangle + \langle u, u\rangle \overset{(2)}{=} \|\mathrm{d}u\|^2 + \|u\|^2 ,$$

(1) Definition of Δ on functions (d^* of a function vanishes).
(2) Move d^* across; no boundary terms since $\partial M = \varnothing$.

so $\|u\| = 0$. The operator is formally self-adjoint, $P^* = P$, so $\operatorname{Ker} P^* = 0$ as well, and by (2.179) there are no constraints. A bounded bijection between Hilbert spaces has a bounded inverse, so $(\Delta + 1)^{-1} : H^s \to H^{s+2}$ exists and is bounded for every s. Composed with the Rellich inclusion it is compact on L^2, which is what gave the discrete spectrum in Example 2.80. The same argument works for $\Delta + \mu^2$ with any $\mu \neq 0$: a mass term makes the Laplacian invertible.

Worked example 2.89: Gauss's law on a closed universe

Now drop the mass. For $P = \Delta$ on functions on a connected compact M:

- $\operatorname{Ker} \Delta$: as above $\|\mathrm{d}u\|^2 = 0$, so $\mathrm{d}u = 0$ and u is constant (M connected). One zero mode.
- $\Delta^* = \Delta$, so $\operatorname{Ker} \Delta^* = \{\text{constants}\}$: one constraint.
- By (2.179), $\Delta u = f$ is solvable iff $\int_M f \,\mathrm{dvol} = 0$, and the solution is unique up to an additive constant. Index $1 - 1 = 0$.

The necessity of the constraint is Stokes' theorem: $\int_M \Delta u = \int_M \mathrm{d}^*\mathrm{d}u = -\int_M \mathrm{d}\star\mathrm{d}u \cdot (\pm 1) = 0$, since $\partial M = \varnothing$. The sufficiency is the Fredholm theory.
Physics. Electrostatics with our positive Laplacian reads $\Delta\phi = \rho/\varepsilon_0$. On a three-torus (a periodic box, as in lattice QED or a cosmological model with compact space) this has a solution iff the total charge $\int \rho$ vanishes. You cannot put a single point charge in a closed universe: its field lines have nowhere to go. This is why numerical simulations of charged

systems in periodic boxes add a uniform "neutralising background" $-Q/\mathrm{Vol}$, and why the Green's function on a compact manifold solves $\Delta G = \delta - 1/\mathrm{Vol}$ rather than $\Delta G = \delta$.
On the circle, explicitly. Solve $-u'' = \cos\theta + b$. In modes, $n^2 c_n(u) = c_n(f)$. For $n = \pm 1$: $c_{\pm 1}(u) = \frac{1}{2}$; for $n = 0$: $0 = b$. So a solution exists iff $b = 0$, and then $u = \cos\theta + c$ with c arbitrary.

Watch out: On $\mathbb{R}^n$ elliptic operators need not be Fredholm

Compactness of M entered through Rellich, and it cannot be dropped. Take $P = \mathrm{d}/\mathrm{d}x : H^1(\mathbb{R}) \to L^2(\mathbb{R})$. Its kernel is 0 (constants are not in L^2) and so is the kernel of its adjoint $-\mathrm{d}/\mathrm{d}x$. If P were Fredholm it would be invertible. It is not bounded below: with $u_\lambda(x) = \lambda^{1/2}\phi(\lambda x)$ for a fixed bump ϕ,

$$\|u_\lambda\|_{L^2} = \|\phi\|_{L^2}, \qquad \|u_\lambda'\|_{L^2} = \lambda\,\|\phi'\|_{L^2} \xrightarrow{\lambda\to 0} 0.$$

Wider and wider bumps have smaller and smaller derivatives, so the range is dense but not closed. In physics language: a massless field in infinite volume has modes of arbitrarily low momentum, a continuum reaching down to zero, and no finite count of "zero modes" makes sense. A mass gap ($\Delta + \mu^2$ on $\mathbb{R}^n$ is invertible) or a confining potential (Example 2.96) restores the Fredholm property.

2.6.6 The index and its stability

Working definition 2.40 (analytical index of an elliptic operator)

For an elliptic P [its principal symbol $\sigma_m(P)(x,\xi)$ is an invertible matrix for every $\xi \neq 0$] on a compact manifold,

$$\operatorname{index} P = \dim \operatorname{Ker} P - \dim \operatorname{Ker} P^*,$$

where P^* is the adjoint [$\langle Pu, v\rangle = \langle u, P^*v\rangle$ in the L^2 inner product], computed on smooth sections (by Section 2.6.5 it is the same on every H^s). *To compute it:* solve $Pu = 0$ and $P^*v = 0$, count, subtract. *To avoid computing it:* deform P to something simpler without leaving the elliptic operators; the index cannot change.

The second half of that box is the reason the index is interesting. It rests on the following properties.

Key idea: Stability of the index

(i) *Local constancy.* If T is Fredholm and $\|S - T\|$ is small enough, then S is Fredholm, $\operatorname{index} S = \operatorname{index} T$, and $\dim \operatorname{Ker} S \le \dim \operatorname{Ker} T$.
(ii) *Compact perturbations.* If T is Fredholm and K compact, $\operatorname{index}(T + K) = \operatorname{index} T$. In particular $\operatorname{index}(I - K) = 0$.
(iii) *Elliptic operators.* $\operatorname{index} P$ depends only on the principal symbol $\sigma_m(P)$, and only on its homotopy class through elliptic (invertible off $\xi = 0$) symbols.
(iv) *Products and adjoints.* $\operatorname{index}(T_1T_2) = \operatorname{index} T_1 + \operatorname{index} T_2$ and $\operatorname{index} T^* = -\operatorname{index} T$. A self-adjoint Fredholm operator has index 0.

Why (i): reduction to a finite matrix. Let $N = \operatorname{Ker} T$ and $C = (\operatorname{Im} T)^{\perp} \cong \operatorname{Coker} T$, both finite-dimensional. Split the domain and target:

$$H_1 = N^{\perp} \oplus N, \qquad H_2 = \operatorname{Im} T \oplus C.$$

Any bounded $S : H_1 \to H_2$ becomes a 2×2 block operator matrix, and T itself is block diagonal with an invertible top-left block T_0:

$$S = \begin{pmatrix} A & B \\ \Gamma & E_0 \end{pmatrix}, \qquad T = \begin{pmatrix} T_0 & 0 \\ 0 & 0 \end{pmatrix}, \qquad \begin{array}{ll} A : N^{\perp} \to \operatorname{Im} T, & B : N \to \operatorname{Im} T, \\ \Gamma : N^{\perp} \to C, & E_0 : N \to C. \end{array}$$

If $\|S - T\| < \left\|T_0^{-1}\right\|^{-1}$, then A is invertible, because $A = T_0\big(I + T_0^{-1}(A - T_0)\big)$ and $\left\|T_0^{-1}(A - T_0)\right\| < 1$, so the Neumann series $\sum_k \big(-T_0^{-1}(A - T_0)\big)^k$ inverts the bracket. Now clear the off-diagonal blocks by "row and column operations":

$$\begin{aligned} \begin{pmatrix} I & 0 \\ -\Gamma A^{-1} & I \end{pmatrix} \begin{pmatrix} A & B \\ \Gamma & E_0 \end{pmatrix} \begin{pmatrix} I & -A^{-1}B \\ 0 & I \end{pmatrix} &\overset{(1)}{=} \begin{pmatrix} I & 0 \\ -\Gamma A^{-1} & I \end{pmatrix} \begin{pmatrix} A & 0 \\ \Gamma & E_0 - \Gamma A^{-1} B \end{pmatrix} \\ &\overset{(2)}{=} \begin{pmatrix} A & 0 \\ 0 & E \end{pmatrix}, \qquad E \coloneqq E_0 - \Gamma A^{-1} B : \; N \to C. \end{aligned} \tag{2.182}$$

(1) Right multiplication: the second column becomes $-AA^{-1}B + B = 0$ and $-\Gamma A^{-1}B + E_0$.
(2) Left multiplication: the bottom-left block becomes $-\Gamma A^{-1}A + \Gamma = 0$; the rest is unchanged.

The outer factors are invertible (their inverses just flip the sign of the off-diagonal block), so S has the same kernel and cokernel dimensions as $\operatorname{diag}(A, E)$. Since A is invertible, all of the kernel and cokernel sit in the finite matrix E, of size $\dim C \times \dim N$. Therefore

$$\operatorname{index} S \overset{(1)}{=} \dim \operatorname{Ker} E - \dim \operatorname{Coker} E \overset{(2)}{=} \dim N - \dim C \overset{(3)}{=} \operatorname{index} T, \tag{2.183}$$

(1) $\operatorname{diag}(A, E)$ with A invertible: only E contributes.
(2) Rank–nullity (2.171) for the matrix $E : N \to C$.
(3) $N = \operatorname{Ker} T$, $C \cong \operatorname{Coker} T$.

and $\dim \operatorname{Ker} S = \dim \operatorname{Ker} E \le \dim N$. When we perturb, the infinite-dimensional part stays invertible; all the action happens in a finite matrix, and a finite matrix cannot change its index.

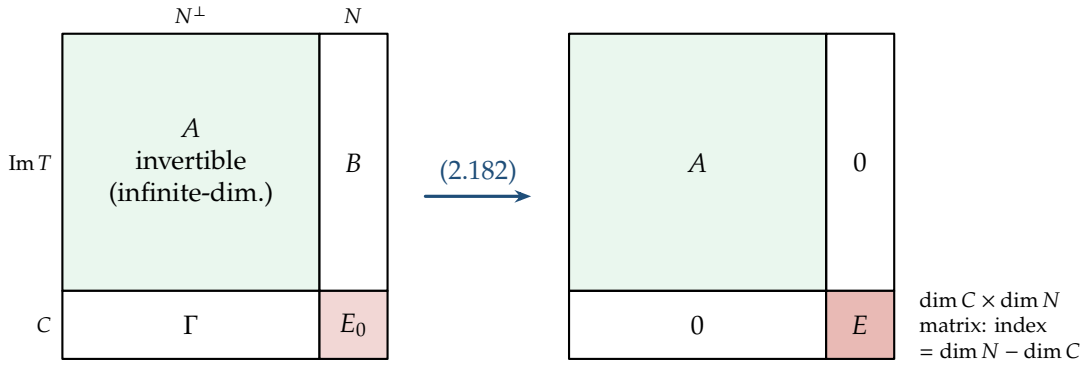


Figure 2.46: Why the index is stable. Split domain and target so that T is invertible on the big block. A small perturbation S keeps the big block A invertible. Row and column operations then push everything that can go wrong into the small red corner E, a finite matrix from the old kernel N to the old cokernel C. The kernel of S may shrink (if E has rank), but the index of E is $\dim N - \dim C$ by rank–nullity, whatever its entries.

Why (ii). For $t \in [0,1]$ consider $T_t = T + tK$. Each T_t is Fredholm, because a parametrix Q of T still works: $QT_t = I - K_2 + tQK$ and QK is compact (bounded after compact), and similarly on the other side; Atkinson applies. The function $t \mapsto \operatorname{index} T_t$ is locally constant by (i), since $\|T_t - T_{t'}\| = |t - t'|\,\|K\|$. A locally constant integer-valued function on the connected interval $[0,1]$ is constant:

$$\operatorname{index}(T+K) = \operatorname{index} T_1 \overset{(1)}{=} \operatorname{index} T_0 = \operatorname{index} T.$$

(1) Local constancy along the connected path $t \mapsto T_t$.

With $T = I$ this proves the claim made after Example 2.87: $I - K$ always has index 0, for every compact K. That is the full *Fredholm alternative*: for $u - Ku = f$, the number of independent solutions of the homogeneous equation equals the number of constraints on f.

Why (iii). If P_0 and P_1 have the same principal symbol, their difference has order $m-1$, so as a map $H^s \to H^{s-m}$ it factors as

$$P_1 - P_0 :\ H^s \xrightarrow{\text{bounded}} H^{s-m+1} \xrightarrow{\text{compact (Rellich)}} H^{s-m},$$

which is compact, and (ii) gives $\operatorname{index} P_1 = \operatorname{index} P_0$. *Lower-order terms never change the index.* If instead the principal symbols are joined by a continuous path σ_t of elliptic symbols, we can choose operators P_t with these symbols depending continuously on t in the norm of bounded maps $H^s \to H^{s-m}$; by (i) the index is constant along the path. So the index is a function of the *homotopy class* of the principal symbol: a topological quantity attached to an analytical object. Classifying these homotopy classes is what K-theory does, and computing the function is the Atiyah–Singer index theorem (Chapters 3 and 4).

Why (iv). For adjoints, $\operatorname{index} T^* = \dim\operatorname{Ker} T^* - \dim\operatorname{Ker} T = -\operatorname{index} T$, and if $T = T^*$ this forces $\operatorname{index} T = 0$. For products, one checks that the sequence

$$0 \to \operatorname{Ker} T_2 \to \operatorname{Ker} T_1T_2 \xrightarrow{T_2} \operatorname{Ker} T_1 \to \operatorname{Coker} T_2 \xrightarrow{T_1} \operatorname{Coker} T_1T_2 \to \operatorname{Coker} T_1 \to 0$$

(inclusion; apply T_2; take the class of v modulo $\operatorname{Im} T_2$; apply T_1; project) is exact, and the alternating sum of dimensions in a finite exact sequence is zero. Rearranged, that is $\operatorname{index} T_1T_2 = \operatorname{index} T_1 + \operatorname{index} T_2$. Check on shifts: $S^k(S^*)^k = I - \Pi_{<k}$ has index $0 = -k + k$.

Watch out: The number of zero modes is not stable; only the difference is

$\dim\operatorname{Ker} P$ by itself can jump under an arbitrarily small deformation (Example 2.84, and Example 2.90 below). What is protected is $\dim\operatorname{Ker} P - \dim\operatorname{Ker} P^*$. Zero modes can only be lifted *in pairs*, one of P together with one of P^*, exactly as the red corner E in Figure 2.46 can only gain rank by using up a direction of N and a direction of C at once. In a theory of chiral fermions, the index of the chiral Dirac operator is therefore a number of zero modes that no small perturbation (masses aside, which couple the two chiralities) can remove.

2.6.7 Worked operators on the circle

On S^1 we write $e_n = e^{in\theta}$, $D = -i\,\mathrm{d}/\mathrm{d}\theta$ (so $De_n = ne_n$) and $|D|\,e_n = |n|\,e_n$, as in Section 2.5.3. All operators below are Fredholm by Section 2.6.5; the point is to see their kernels, cokernels and indices with our own eyes.

Worked example 2.90: The twisted derivative: the kernel jumps, the index stays 0

Let $a \in \mathbb{R}$ and $D_a = D + a = -i\,\mathrm{d}/\mathrm{d}\theta + a : H^1(S^1) \to L^2(S^1)$. Physically this is the momentum of a charged particle on a ring threaded by a magnetic flux (Aharonov–Bohm), with a the flux in units of the flux quantum; mathematically it is the covariant derivative $-i(\mathrm{d}/\mathrm{d}\theta + ia)$ of a flat U(1) connection with holonomy $e^{-2\pi i a}$ around the circle.

- Spectrum: $D_a e_n = (n + a)e_n$. The eigenvalues are $n + a$, $n \in \mathbb{Z}$.
- Kernel: $n + a = 0$ has a solution iff $a \in \mathbb{Z}$. So $\dim \operatorname{Ker} D_a = 1$ for integer a, and 0 otherwise.
- Adjoint: $D_a^* = D_a$ (D is formally self-adjoint and a is real), so $\dim \operatorname{Coker} D_a = \dim \operatorname{Ker} D_a$.
- Index: 0 for every a, as (iv) demands.

The reduction (2.182) made explicit. At $a = 0$, $N = C = \{\text{constants}\}$. For small a, D_a maps the constants to the constants ($D_a 1 = a \cdot 1$) and non-constant modes to non-constant modes, so $B = \Gamma = 0$, $A = D_a$ on the non-constant modes (invertible, eigenvalues $n + a$ with $n \neq 0$), and the red corner is the 1×1 matrix $E = (a)$. Its kernel is everything when $a = 0$ and nothing otherwise, and its index is $1 - 1 = 0$ always.

A non-constant potential is a gauge transformation away. For a smooth real $a(\theta)$ let $\bar{a} = \frac{1}{2\pi}\int_0^{2\pi} a\,\mathrm{d}\theta$ and $\chi(\theta) = \int_0^\theta \big(a(\theta') - \bar{a}\big)\mathrm{d}\theta'$, which is periodic because $\chi(2\pi) = 2\pi\bar{a} - 2\pi\bar{a} = 0$. For any v,

$$\begin{aligned}\big(D + a\big)\big(e^{-i\chi}v\big) &\overset{(1)}{=} -i\big(-i\chi' e^{-i\chi}v + e^{-i\chi}v'\big) + a\,e^{-i\chi}v \\ &\overset{(2)}{=} e^{-i\chi}\big(-iv' + (a - \chi')v\big) \overset{(3)}{=} e^{-i\chi}\,(D + \bar{a})v.\end{aligned}$$

(1) Product rule for $-i\,\mathrm{d}/\mathrm{d}\theta$.
(2) Collect terms: $-i \cdot (-i) = -1$.
(3) $\chi' = a - \bar{a}$.

So $D + a(\theta) = e^{-i\chi}(D + \bar{a})e^{i\chi}$: conjugate to the constant case. The spectrum is $n + \bar{a}$, and a zero mode exists iff the holonomy $e^{-2\pi i\bar{a}}$ is 1. Only the gauge-invariant flux $\bar{a}$ matters, as it must.

Worked example 2.91: Every scalar elliptic differential operator on S^1 has index 0

Let $P = \sum_{j=0}^m a_j(\theta)D^j$ with $a_m(\theta) \neq 0$ everywhere (elliptic, since $\sigma_m = a_m(\theta)\xi^m$). Write

$$P = a_m\,(D^m + i) \;+\; \underbrace{\Big(\sum_{j<m} a_j D^j - i\,a_m\Big)}_{\text{order } \leq m-1}.$$

Then

$$\operatorname{index} P \overset{(1)}{=} \operatorname{index}\big(a_m(D^m + i)\big) \overset{(2)}{=} \operatorname{index}(a_m) + \operatorname{index}(D^m + i) \overset{(3)}{=} 0 + 0.$$

(1) Stability (iii): the underbraced lower-order part is compact $H^s \to H^{s-m}$.
(2) Product rule (iv), with a_m the multiplication operator $H^{s-m} \to H^{s-m}$.

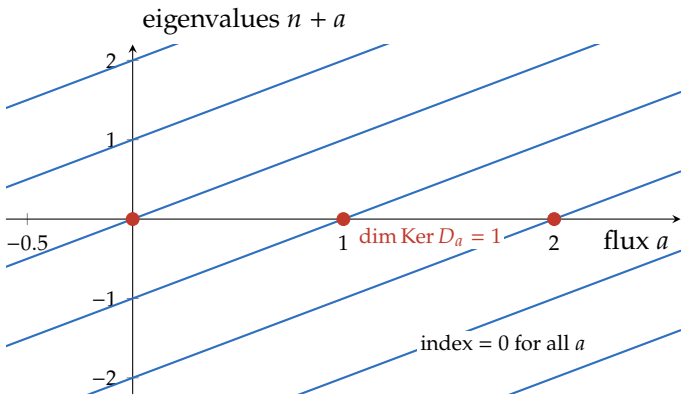


Figure 2.47: Spectral flow of the twisted derivative $D_a = -i\,\mathrm{d}/\mathrm{d}\theta + a$ on the circle. Each Fourier mode $e^{in\theta}$ gives a line $n + a$. As the flux a increases by one unit, every eigenvalue moves up by one: the spectrum returns to itself, but one eigenvalue has crossed zero from below (red dots mark the crossings, where a zero mode exists). The kernel jumps at each integer a; the index, $\dim\mathrm{Ker} - \dim\mathrm{Ker}^*$, is 0 throughout because D_a is self-adjoint. The net number of crossings per unit flux (here +1) is the *spectral flow*; in the second-quantised theory it is the charge pumped out of the Dirac sea (the intuition box after Figure 2.44).

(3) Multiplication by the nowhere-zero smooth function a_m is invertible (its inverse is multiplication by $1/a_m$), so its index is 0; and $D^m + i$ multiplies e_n by $n^m + i \neq 0$, so it is invertible too.

A concrete check. $P = D + e^{i\theta}$. Solve $Pu = 0$: $-iu' + e^{i\theta}u = 0$, i.e. $u' = -ie^{i\theta}u$, whose solution $u = \exp(-e^{i\theta})$ is periodic (note $\frac{\mathrm{d}}{\mathrm{d}\theta}(-e^{i\theta}) = -ie^{i\theta}$). So $\dim\mathrm{Ker}\,P = 1$. The formal adjoint is $P^* = D + e^{-i\theta}$, and $P^*v = 0$ gives $v = \exp(e^{-i\theta})$: $\dim\mathrm{Ker}\,P^* = 1$. Index $1 - 1 = 0$. (Both zero modes exist because $\int_0^{2\pi} e^{\pm i\theta}\mathrm{d}\theta = 0$: the complex "flux" $\bar{a}$ of the previous example is zero.) Nash proves the vanishing of the index of elliptic differential operators on *every* odd-dimensional manifold in Chapter 4, using the antipodal map $\xi \mapsto -\xi$. On the circle we have just seen the elementary version.

Worked example 2.92: A pseudo-differential operator with non-zero index

ΨDOs escape the last example. Let $\Pi_\pm$ be the projections onto the non-negative and the negative modes:

$$\Pi_+ e_n = \begin{cases} e_n, & n \geq 0 \\ 0, & n < 0 \end{cases}, \qquad \Pi_- = I - \Pi_+.$$

Π_+ is a Fourier multiplier whose symbol is 1 for large positive ξ and 0 for large negative ξ (smooth it arbitrarily near $\xi = 0$): a ΨDO of order 0 with principal symbol $\sigma_0(\Pi_+) = 1$ on $\xi > 0$ and 0 on $\xi < 0$. (It is the Szegő projection onto boundary values of holomorphic functions, and $\Pi_+ - \Pi_-$ is i times the Hilbert transform of Example 2.81, up to the $n = 0$ mode.) For $k \in \mathbb{Z}$ define the order-0 operator

$$T_k = e^{ik\theta}\,\Pi_+ + \Pi_-, \qquad \sigma_0(T_k)(\theta, \xi) = \begin{cases} e^{ik\theta}, & \xi > 0, \\ 1, & \xi < 0. \end{cases}$$

The principal symbol never vanishes, so T_k is elliptic. Its action on the basis is

$$T_k e_n = \begin{cases} e_{n+k}, & n \geq 0, \\ e_n, & n < 0. \end{cases}$$

For $k > 0$: distinct basis vectors go to distinct basis vectors, so $\operatorname{Ker} T_k = 0$. The image contains e_m for $m < 0$ and for $m \geq k$, and misses $e_0, \dots, e_{k-1}$. So $\dim \operatorname{Coker} T_k = k$ and $\operatorname{index} T_k = -k$. (On the non-negative modes T_k is the shift S^k of Section 2.6.3.)
For $k = -j < 0$: $T_{-j} e_n = e_{n-j}$ for $n \geq 0$, so the j vectors $e_0, \dots, e_{j-1}$ land on $e_{-j}, \dots, e_{-1}$, which are also the images of themselves under Π_-. Hence the j vectors $e_n - e_{n-j}$ $(0 \leq n < j)$ are killed, every basis vector is hit, and $\operatorname{index} T_{-j} = j - 0 = j$.
In both cases

$$\operatorname{index} T_k = -k = -\big(\text{winding number of } \sigma_0(T_k) \text{ on } \xi > 0\big) + \big(\text{winding number on } \xi < 0\big).$$

This is the **Toeplitz index theorem** in its simplest form, and a first glimpse of the Atiyah–Singer philosophy: the index is the topology (here, a winding number) of the principal symbol. For a *differential* operator the symbols on $\xi > 0$ and $\xi < 0$ are $a_m(\theta)\xi^m$ and $a_m(\theta)(-|\xi|)^m$: the same function up to a constant sign, with equal winding numbers, which cancel. That is the circle version of Nash's antipodal argument.
Nash's example (4.35). In Chapter 4 Nash exhibits the order-1 operator with symbol $e^{-i\theta}(\xi + |\xi|) + (\xi - |\xi|)$. In our notation it is

$$P_N = e^{-i\theta}(D + |D|) + (D - |D|) = T_{-1} \circ 2D,$$

because $D + |D| = 2D\Pi_+$ and $D - |D| = 2D\Pi_-$ (check on e_n: $n + |n|$ is $2n$ for $n \geq 0$ and 0 for $n < 0$). Directly: $P_N e_n = 2n e_{n-1}$ for $n \geq 0$ and $P_N e_n = 2n e_n$ for $n < 0$. So $\operatorname{Ker} P_N = \operatorname{span}(e_0)$, every e_m is hit, and $\operatorname{index} P_N = 1$, in agreement with $\operatorname{index} T_{-1} + \operatorname{index}(2D) = 1 + 0$.

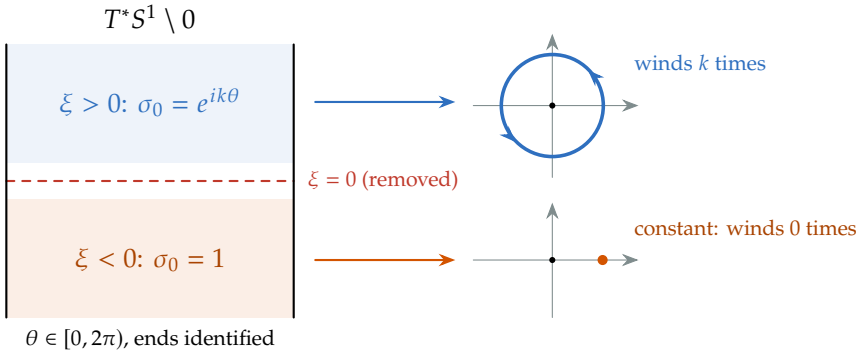


Figure 2.48: The principal symbol of $T_k = e^{ik\theta}\Pi_+ + \Pi_-$ lives on T^*S^1 with the zero section removed, which has two components, $\xi > 0$ and $\xi < 0$. On each component the symbol is a map from a circle to $\mathbb{C} \setminus \{0\}$, with a winding number: k on top, 0 below. The index is $-k$. A differential operator cannot do this: its symbols on the two halves are tied together by $\xi \mapsto -\xi$ and have equal winding numbers, so its index is 0.

2.6.8 Worked operators on the torus

Now $M = T^2 = \mathbb{R}^2/2\pi\mathbb{Z}^2$ with coordinates (x, y), flat metric, and Fourier modes $e_k = e^{i(k_1 x + k_2 y)}$, $k = (k_1, k_2) \in \mathbb{Z}^2$. Write $z = x + iy$ and

$$\partial = \tfrac{1}{2}(\partial_x - i\partial_y), \qquad \bar\partial = \tfrac{1}{2}(\partial_x + i\partial_y), \qquad \bar\partial e_k = \tfrac{i}{2}(k_1 + ik_2)\, e_k, \qquad \partial e_k = \tfrac{i}{2}(k_1 - ik_2)\, e_k.$$

Worked example 2.93: The Cauchy–Riemann operator on the torus

$\bar\partial : H^1(T^2) \to L^2(T^2)$ is elliptic (its symbol $\frac{i}{2}(\xi_1 + i\xi_2)$ vanishes only at $\xi = 0$; see the Cauchy–Riemann operator in the section *Ellipticity and hypo-ellipticity*).

- Kernel: $\frac{i}{2}(k_1 + ik_2) = 0$ iff $k = 0$. So $\operatorname{Ker}\bar\partial = \{\text{constants}\}$: a holomorphic function on a compact Riemann surface is constant (Liouville).
- Adjoint: $\partial_x^* = -\partial_x$ and $(i\partial_y)^* = (-i)(-\partial_y) = i\partial_y$, so $\bar\partial^* = \frac{1}{2}(-\partial_x + i\partial_y) = -\partial$. Its kernel is again the constants.
- Index: $1 - 1 = 0$.

This matches the Riemann–Roch formula index $\bar\partial = 1 - g$ for the trivial bundle on a surface of genus g (here $g = 1$), which we meet in Chapter 5.
Twisting. For a constant $c \in \mathbb{C}$ (a flat connection on the trivial line bundle), $\bar\partial + c$ multiplies e_k by $\frac{i}{2}(k_1 + ik_2) + c$, so it has a zero mode iff $c = -\frac{i}{2}(k_1 + ik_2) = \frac{1}{2}(k_2 - ik_1)$ for some $k \in \mathbb{Z}^2$, i.e. iff c lies on the lattice $\frac{1}{2}(\mathbb{Z} + i\mathbb{Z})$. Off the lattice there are no zero modes of either $\bar\partial + c$ or its adjoint; on it, one of each. The index is 0 throughout, as it must be since c is a compact (order $0 < 1$) perturbation.

Worked example 2.94: The Dirac operator on the torus, in explicit matrices

Use the Pauli matrices $\sigma_1 = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, $\sigma_2 = \left(\begin{smallmatrix} 0 & -i \\ i & 0 \end{smallmatrix}\right)$, which satisfy $\sigma_a\sigma_b + \sigma_b\sigma_a = 2\delta_{ab}$ (the Clifford relations of Chapter T4), and set $\not{D} = -i(\sigma_1\partial_x + \sigma_2\partial_y)$. Multiplying out,

$$\not{D} \overset{(1)}{=} -i\begin{pmatrix} 0 & \partial_x - i\partial_y \\ \partial_x + i\partial_y & 0 \end{pmatrix} \overset{(2)}{=} \begin{pmatrix} 0 & -2i\,\partial \\ -2i\,\bar\partial & 0 \end{pmatrix}.$$

(1) $\sigma_1\partial_x + \sigma_2\partial_y$ has off-diagonal entries $\partial_x - i\partial_y$ (top right) and $\partial_x + i\partial_y$ (bottom left).
(2) The definitions of ∂ and $\bar\partial$.

So $\not{D}$ is off-diagonal in the chiral splitting $\psi = (\psi_+, \psi_-)$, with chiral part $\not{D}_+ = -2i\bar\partial : \psi_+ \mapsto$ lower component, and $\not{D}_- = -2i\partial = (\not{D}_+)^*$. Its square is

$$\not{D}^2 \overset{(1)}{=} -\sum_{a,b} \sigma_a\sigma_b\,\partial_a\partial_b \overset{(2)}{=} -\tfrac{1}{2}\sum_{a,b}(\sigma_a\sigma_b + \sigma_b\sigma_a)\partial_a\partial_b \overset{(3)}{=} -(\partial_x^2 + \partial_y^2)\,I,$$

(1) Expand $\left(-i\sum_a \sigma_a\partial_a\right)^2$; partial derivatives commute with constant matrices.
(2) $\partial_a\partial_b$ is symmetric, so only the symmetric part of $\sigma_a\sigma_b$ survives.
(3) Clifford relation.

so $\not{D}$ is elliptic (a square root of the Laplacian). With periodic boundary conditions for both spinor components (one of the four spin structures on T^2), the zero modes are the constant spinors: $\operatorname{Ker}\not{D}_+ = \mathbb{C}$, $\operatorname{Ker}\not{D}_- = \mathbb{C}$, and index $\not{D}_+ = 1 - 1 = 0$. With an antiperiodic direction the momenta are shifted to $\mathbb{Z} + \frac{1}{2}$, the symbol $\frac{i}{2}(k_1 + ik_2)$ never vanishes on the allowed k, and there are no zero modes at all. The index is 0 either way; the number of zero modes depends on the spin structure. A non-zero index on a surface needs curvature or a background gauge field with non-zero flux: that is the content of the index theorem, and of the anomaly in Chapter 10.

Worked example 2.95: The operator $\mathrm{d}+\mathrm{d}^*$ and the Euler characteristic

On a compact oriented M consider $\mathrm{d}+\mathrm{d}^*$ from even forms to odd forms,

$$\mathrm{d}+\mathrm{d}^*:\ \Omega^{\text{even}}(M)=\bigoplus\nolimits_p \Omega^{2p}\longrightarrow \Omega^{\text{odd}}(M)=\bigoplus\nolimits_p \Omega^{2p+1}.$$

Its square (on all forms) is $\mathrm{d}\mathrm{d}^*+\mathrm{d}^*\mathrm{d}=\Delta$, because $\mathrm{d}^2=0=(\mathrm{d}^*)^2$; so its principal symbol squares to $|\xi|^2\,I$ and it is elliptic.
Its zero modes are the harmonic forms. For a form ω (of mixed degree),

$$\begin{aligned}\|(\mathrm{d}+\mathrm{d}^*)\omega\|^2 &\overset{(1)}{=} \|\mathrm{d}\omega\|^2+\|\mathrm{d}^*\omega\|^2+2\,\mathrm{Re}\langle \mathrm{d}\omega,\mathrm{d}^*\omega\rangle \overset{(2)}{=} \|\mathrm{d}\omega\|^2+\|\mathrm{d}^*\omega\|^2+2\,\mathrm{Re}\langle \mathrm{d}\mathrm{d}\omega,\omega\rangle\\ &\overset{(3)}{=} \|\mathrm{d}\omega\|^2+\|\mathrm{d}^*\omega\|^2,\end{aligned}$$

(1) Expand the square of a sum.
(2) Move d^* back across the inner product as d.
(3) $\mathrm{d}^2=0$.

so $(\mathrm{d}+\mathrm{d}^*)\omega=0$ iff $\mathrm{d}\omega=\mathrm{d}^*\omega=0$ iff ω is harmonic, by (2.114). The adjoint is $\mathrm{d}+\mathrm{d}^*$ from odd to even forms, whose kernel is the odd harmonic forms. By Hodge's theorem (2.115):

$$\operatorname{index}\big(\mathrm{d}+\mathrm{d}^*\big)_{\text{even}\to\text{odd}} \overset{(1)}{=} \sum_p \dim\operatorname{Ker}\Delta_{2p}-\sum_p \dim\operatorname{Ker}\Delta_{2p+1} \overset{(2)}{=} \sum_q (-1)^q\, b_q \overset{(3)}{=} \chi(M). \tag{2.184}$$

(1) Kernel = even harmonic forms; cokernel = kernel of the adjoint = odd harmonic forms.
(2) Hodge: $\dim\operatorname{Ker}\Delta_q=b_q$.
(3) The alternating sum of Betti numbers is the Euler characteristic.

On T^2, explicitly. An even form is $(f,h)\leftrightarrow f+h\,\mathrm{d}x\wedge\mathrm{d}y$ and an odd form is $(a,b)\leftrightarrow a\,\mathrm{d}x+b\,\mathrm{d}y$. Using $\mathrm{d}^*=-\star\,\mathrm{d}\star$ on 2-forms in two dimensions, $\star(\mathrm{d}x\wedge\mathrm{d}y)=1$, $\star\mathrm{d}x=\mathrm{d}y$, $\star\mathrm{d}y=-\mathrm{d}x$:

$$\mathrm{d}^*(h\,\mathrm{d}x\wedge\mathrm{d}y)\overset{(1)}{=}-\star\,\mathrm{d}h\overset{(2)}{=}-\star\,(h_x\,\mathrm{d}x+h_y\,\mathrm{d}y)\overset{(3)}{=}h_y\,\mathrm{d}x-h_x\,\mathrm{d}y,$$

(1) $\star(h\,\mathrm{d}x\wedge\mathrm{d}y)=h$.
(2) Exterior derivative of a function.
(3) $\star\mathrm{d}x=\mathrm{d}y$ and $\star\mathrm{d}y=-\mathrm{d}x$.

so, as a matrix of operators acting on the column $(f,h)^T$,

$$\mathrm{d}+\mathrm{d}^*=\begin{pmatrix}\partial_x & \partial_y\\ \partial_y & -\partial_x\end{pmatrix},\qquad \sigma_1(\mathrm{d}+\mathrm{d}^*)(\xi)=i\begin{pmatrix}\xi_1 & \xi_2\\ \xi_2 & -\xi_1\end{pmatrix},\qquad \det\begin{pmatrix}\xi_1 & \xi_2\\ \xi_2 & -\xi_1\end{pmatrix}=-|\xi|^2.$$

On the mode e_k the operator is i times the same matrix with $\xi=k$, invertible for $k\neq 0$. So the kernel is $k=0$: the constants (f,h), dimension $2=b_0+b_2$. The cokernel is the constant 1-forms $a\,\mathrm{d}x+b\,\mathrm{d}y$, dimension $2=b_1$. Index $0=\chi(T^2)$. On the sphere S^2 the same operator has index $b_0-b_1+b_2=2$; on a surface of genus g, $2-2g$. By Gauss–Bonnet (Chapter T2) this equals $\frac{1}{2\pi}\int_M K\,\mathrm{d}A$: an analytical count of zero modes equals an integral of curvature. That sentence, for every elliptic operator, is the index theorem of Chapter 4.

Try it: Index of a matrix-valued operator on the circle

On S^1 let $P = \left(\begin{smallmatrix} D & 0 \\ 0 & D+\frac{1}{2} \end{smallmatrix}\right)$ acting on pairs of functions. What are $\dim\operatorname{Ker} P$, $\dim\operatorname{Ker} P^*$ and index P? What happens if you add the off-diagonal term $\left(\begin{smallmatrix} 0 & \varepsilon \\ \varepsilon & 0 \end{smallmatrix}\right)$?
Answer. D has the constants as kernel; $D+\frac{1}{2}$ has none (eigenvalues $n+\frac{1}{2}$). P is self-adjoint, so $\dim\operatorname{Ker} P = \dim\operatorname{Ker} P^* = 1$ and index $P = 0$. The perturbation is order 0, so the index stays 0. On the constants the 2×2 block becomes $\left(\begin{smallmatrix} 0 & \varepsilon \\ \varepsilon & \frac{1}{2} \end{smallmatrix}\right)$, with determinant $-\varepsilon^2 \neq 0$ for $\varepsilon \neq 0$, and on the mode e_n it is $\left(\begin{smallmatrix} n & \varepsilon \\ \varepsilon & n+\frac{1}{2} \end{smallmatrix}\right)$, with determinant $n^2 + \frac{n}{2} - \varepsilon^2$, which for small $\varepsilon \neq 0$ vanishes for no integer n. So the zero mode is lifted, together with the zero mode of P^* (it is the same one): a pair, as the pitfall in Section 2.6.6 says.

2.6.9 Counting zero modes with the heat kernel

The index can be computed without solving $Pu = 0$. The idea is that the non-zero modes of P^*P and PP^* come in pairs, and cancel in a suitable difference of traces.

Pairing. Suppose $P^*Pu = \lambda u$ with $\lambda \neq 0$. Then

$$PP^*(Pu) = P(P^*Pu) \overset{(1)}{=} \lambda\, Pu, \qquad \|Pu\|^2 \overset{(2)}{=} \langle u, P^*Pu\rangle = \lambda\,\|u\|^2 \neq 0. \tag{2.185}$$

(1) u is an eigenvector of P^*P.
(2) Definition of the adjoint.

So Pu is a non-zero eigenvector of PP^* with the same eigenvalue, and $P/\sqrt{\lambda}$ maps the λ-eigenspace of P^*P isometrically onto the λ-eigenspace of PP^* (the inverse is $P^*/\sqrt{\lambda}$). At $\lambda = 0$ the pairing fails: $\operatorname{Ker} P^*P = \operatorname{Ker} P$ (since $\langle u, P^*Pu\rangle = \|Pu\|^2$) and $\operatorname{Ker} PP^* = \operatorname{Ker} P^*$, and these have different dimensions in general.

Key idea: McKean–Singer formula

For an elliptic P on a compact manifold and every $t > 0$,

$$\operatorname{index} P = \operatorname{Tr} e^{-tP^*P} - \operatorname{Tr} e^{-tPP^*}. \tag{2.186}$$

Why. P^*P and PP^* are elliptic, self-adjoint and non-negative, so (by the argument of Example 2.80) they have discrete spectra $\lambda_j \to \infty$ with finite multiplicities; e^{-tP^*P} is a smoothing operator, and its trace $\sum_j e^{-t\lambda_j}$ converges (Weyl's law makes λ_j grow like a power of j). Then

$$\begin{aligned}\operatorname{Tr} e^{-tP^*P} - \operatorname{Tr} e^{-tPP^*} &\overset{(1)}{=} \sum_{\lambda} \big(m_{P^*P}(\lambda) - m_{PP^*}(\lambda)\big) e^{-t\lambda} \overset{(2)}{=} m_{P^*P}(0) - m_{PP^*}(0) \\ &\overset{(3)}{=} \dim\operatorname{Ker} P - \dim\operatorname{Ker} P^*,\end{aligned}$$

(1) Sum over distinct eigenvalues with multiplicities $m(\lambda)$.
(2) The pairing (2.185): multiplicities agree for $\lambda \neq 0$.
(3) $\operatorname{Ker} P^*P = \operatorname{Ker} P$ and $\operatorname{Ker} PP^* = \operatorname{Ker} P^*$.

The left side of (2.186) looks t-dependent, but is not. As $t \to \infty$ it reduces to the zero modes (the definition of the index). As $t \to 0$ the heat kernels become local, and their diagonal values have an expansion in the curvature at each point: the index becomes an integral of a local density. That is the heat-kernel proof of the index theorem (Chapter 4), and the Fujikawa computation

of anomalies (Chapter 10). The smoothing property of the heat operator, which in this chapter we met as Nash's Gaussian example, is what makes it work.

Worked example 2.96: Supersymmetric harmonic oscillator: index 1

Take $P = \frac{\mathrm{d}}{\mathrm{d}x} + x$ on $L^2(\mathbb{R})$. (This is on $\mathbb{R}$, but the potential x^2 confines, the resolvent is compact, and all the above goes through.) Its adjoint is $P^* = -\frac{\mathrm{d}}{\mathrm{d}x} + x$, and

$$P^*Pu \overset{(1)}{=} \Big(-\frac{\mathrm{d}}{\mathrm{d}x} + x\Big)(u' + xu) \overset{(2)}{=} -u'' - u - xu' + xu' + x^2u = \Big(-\frac{\mathrm{d}^2}{\mathrm{d}x^2} + x^2 - 1\Big)u,$$
$$PP^*u \overset{(3)}{=} \Big(\frac{\mathrm{d}}{\mathrm{d}x} + x\Big)(-u' + xu) \overset{(4)}{=} -u'' + u + xu' - xu' + x^2u = \Big(-\frac{\mathrm{d}^2}{\mathrm{d}x^2} + x^2 + 1\Big)u.$$

(1) Apply P, then P^*.
(2) Product rule: $(xu)' = u + xu'$.
(3) Apply P^*, then P.
(4) Product rule again, now with the opposite sign.

With $H = -\frac{\mathrm{d}^2}{\mathrm{d}x^2} + x^2$ (eigenvalues $2n+1$, $n \geq 0$), $P^*P = H - 1$ has eigenvalues $0, 2, 4, \ldots$ and $PP^* = H + 1$ has eigenvalues $2, 4, 6, \ldots$. The zero modes: $Pu = 0$ gives $u' = -xu$, $u = e^{-x^2/2} \in L^2$; $P^*v = 0$ gives $v = e^{x^2/2} \notin L^2$. So index $P = 1 - 0 = 1$, and

$$\operatorname{Tr} e^{-tP^*P} - \operatorname{Tr} e^{-tPP^*} \overset{(1)}{=} \sum_{n\geq 0} e^{-2nt} - \sum_{n\geq 0} e^{-2(n+1)t} \overset{(2)}{=} 1,$$

(1) The two spectra.
(2) The second sum is the first sum minus its $n = 0$ term; both are geometric series.

independent of t, as promised. In the language of supersymmetric quantum mechanics (Chapter 7): P^*P and PP^* are the bosonic and fermionic Hamiltonians, P is the supercharge that pairs their excited states, and the index is the **Witten index** $\operatorname{Tr}(-1)^F e^{-tH}$, which counts supersymmetric ground states.

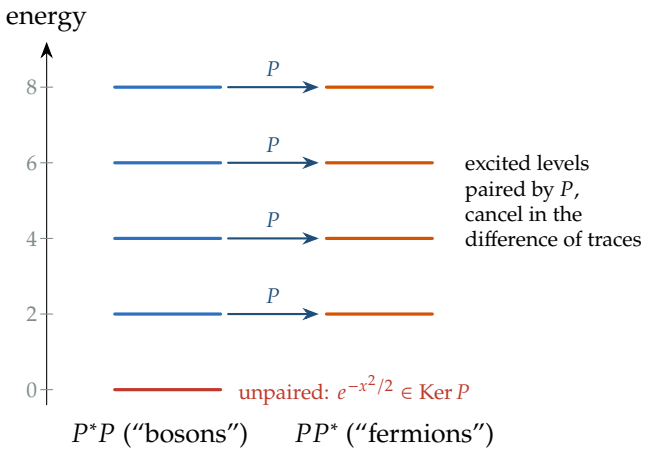


Figure 2.49: The McKean–Singer mechanism for the supersymmetric oscillator $P = \mathrm{d}/\mathrm{d}x + x$. Every excited level of P^*P is carried by P onto a level of PP^* with the same energy, so these cancel in $\operatorname{Tr} e^{-tP^*P} - \operatorname{Tr} e^{-tPP^*}$ for every t. Only the zero mode of P (red) is left unpaired; PP^* has none. The difference of traces is the index, 1.

2.6.10 Looking ahead: the space of Fredholm operators

Let $\mathcal{F}(H)$ be the set of all Fredholm operators on a Hilbert space H. Stability (i) says that the index is constant on each connected component of $\mathcal{F}(H)$; the shifts S^k and $(S^*)^k$ show that every integer occurs. In fact two Fredholm operators can be joined by a path in $\mathcal{F}(H)$ exactly when they have the same index, so

$$\text{index} : \ \pi_0\big(\mathcal{F}(H)\big) \xrightarrow{\cong} \mathbb{Z}.$$

Nash takes this one step further at the end of Chapter 3: for a family of Fredholm operators parametrised by a compact space X, the "index" is no longer an integer but an element of the K-theory group $K(X)$, and $[X, \mathcal{F}] \cong K(X)$. For X a point this is the statement above. Families of Dirac operators parametrised by gauge fields are exactly such families, and their K-theoretic index is how global anomalies are detected (Chapter 10).

Dictionary: Fredholm theory and physics

Mathematics	Physics
$\operatorname{Ker} P$, finite-dimensional	normalisable zero modes (harmonic forms, fermion zero modes)
$\operatorname{Ker} P^*$ and (2.179)	constraints on sources: Gauss's law on a closed space
$\operatorname{index} P = \dim \operatorname{Ker} P - \dim \operatorname{Ker} P^*$	net number of chiral zero modes; Witten index
stability of the index	zero modes are lifted only in pairs; "protected" zero modes
compact perturbation	adding lower-order terms, masses in P^*P, smooth potentials
parametrix	propagator up to smooth (non-singular) corrections
spectral flow of a self-adjoint family	charge pumped from the Dirac sea
McKean–Singer (2.186)	$\operatorname{Tr}(-1)^F e^{-tH}$; Fujikawa's regularised Jacobian
$\pi_0(\mathcal{F}) = \mathbb{Z}$, $[X, \mathcal{F}] = K(X)$	topological classes of families; global anomalies

Reading Nash (pp. 54–55, (2.94)–(2.97))

Nash's final page and a half is dense. Here is the map. "If P is an elliptic mth order differential operator on a compact manifold then P has a finite dimensional kernel": our Section 2.6.5, via the parametrix, Rellich (Section 2.5.8) and part (a) of Atkinson's theorem in Section 2.6.4; it holds for ΨDOs as well. "P elliptic implies P^* elliptic": the adjoint-symbol computation in Section 2.6.5. "Bounded linear operators between Hilbert spaces with this property are called Fredholm operators": our working definition in Section 2.6.2, with the remark that the cokernel is $\operatorname{Ker} P^*$ only once the range is known to be closed. The boundedness discussion (2.94)–(2.96) is Example 2.82. "Fredholm operators ... are invertible modulo a compact operator" is Atkinson's theorem (2.177), and "the Sobolev space version of ... invertible modulo a smoothing operator" is (2.180). The parametrix (2.97) is (2.177), with the identities corrected (pitfall in Section 2.6.4). Nash never writes down the index in this chapter; he postpones it to his Chapter 4 (our Chapter 4) and, for Fredholm families, to the end of his Chapter 3. We have introduced it here, with its stability, because it is the reason for the whole chapter.

Recap: Elliptic operators are Fredholm

- For a matrix, $\dim\mathrm{Ker} - \dim\mathrm{Coker} = n - m$ regardless of the entries. A Fredholm operator is one where both are finite; its index $\dim\mathrm{Ker}\,T - \dim\mathrm{Ker}\,T^*$ is the infinite-dimensional replacement for $n - m$, and can be any integer (shifts).
- Atkinson: Fredholm $\iff$ invertible modulo compact operators. Compact kernels act as the identity only on finite-dimensional spaces; that is where finiteness comes from.
- On a compact manifold, an elliptic ΨDO of order m is Fredholm $H^s \to H^{s-m}$: parametrix errors are smoothing, smoothing is compact by Rellich. Kernel and cokernel $\mathrm{Ker}\,P^*$ are smooth and independent of s; $Pu = f$ is solvable iff $f \perp \mathrm{Ker}\,P^*$ (Gauss's law).
- The index is locally constant, unchanged by compact (in particular lower-order) perturbations, and depends only on the homotopy class of the principal symbol. The proof reduces everything to a finite matrix. Self-adjoint operators have index 0.
- Examples: $D + a$ on S^1 (kernel jumps, index 0; spectral flow), scalar differential operators on S^1 (index 0), Toeplitz operators (index = − winding number), $\bar\partial$ and $\not{D}$ on T^2 (index 0), $\mathrm{d} + \mathrm{d}^*$ (index $\chi(M)$), the SUSY oscillator (index 1).
- McKean–Singer: $\mathrm{index}\,P = \mathrm{Tr}\,e^{-tP^*P} - \mathrm{Tr}\,e^{-tPP^*}$ for all t, because non-zero modes pair up. This is the bridge to the heat-kernel proof of the index theorem and to anomalies.

Chapter summary

This chapter asked one question: *which differential operators behave like the Laplacian?* "Behave like the Laplacian" turned out to mean three things. Smooth sources give smooth solutions (regularity). On a compact space there are only finitely many zero modes, and only finitely many obstructions to solving $Pu = f$ (finiteness). And the difference of those two numbers does not change when we deform the operator (stability). The answer was a single algebraic test on a single matrix of polynomials: the operator is *elliptic* when its principal symbol $\sigma_m(x,\xi)$ is invertible for every $\xi \neq 0$.

The story, in eight sentences:

1. The Fourier transform turns $-i\partial_j$ into multiplication by the momentum ξ_j, so a differential operator becomes a polynomial $\sigma(x,\xi)$, its *symbol*.
2. Its top-degree part $\sigma_m(x,\xi)$ controls the high-momentum (short-distance) behaviour. When it never vanishes for $\xi \neq 0$ (elliptic), the operator can be inverted at high momentum; when it vanishes on a cone (hyperbolic), singularities travel along that cone.
3. Nothing changes for systems of equations, i.e. operators between vector bundles: σ_m becomes a matrix, and ellipticity means that the matrix is invertible.
4. The Hodge Laplacian has $\sigma_2 = |\xi|^2\, I$, so it is elliptic. Hodge's theorem follows: its zero modes on p-forms count the p-dimensional holes, $\dim\mathrm{Ker}\,\Delta_p = b_p$.
5. To invert an elliptic operator we have to allow symbols that are not polynomials, such as $1/|\xi|^2$ and $\sqrt{|\xi|^2 + m^2}$: these give the pseudo-differential operators, with a calculus in which principal symbols multiply.

6. The inverse of the principal symbol quantises to a *parametrix* Q, an inverse up to smoothing errors: $PQ = I - R_1$, $QP = I - R_2$.
7. Sobolev spaces H^s count derivatives in L^2 with the weight $(1 + |\xi|^2)^s$. On them ΨDOs of order m are bounded maps $H^s \to H^{s-m}$, and on a compact manifold smoothing operators are compact (Rellich).
8. Invertible modulo compact operators means Fredholm (Atkinson). So an elliptic operator on a compact manifold has a finite index $\dim \operatorname{Ker} P - \dim \operatorname{Ker} P^*$, which depends only on the homotopy class of its principal symbol. Computing that integer from the symbol and the topology is the index theorem of Chapter 4.

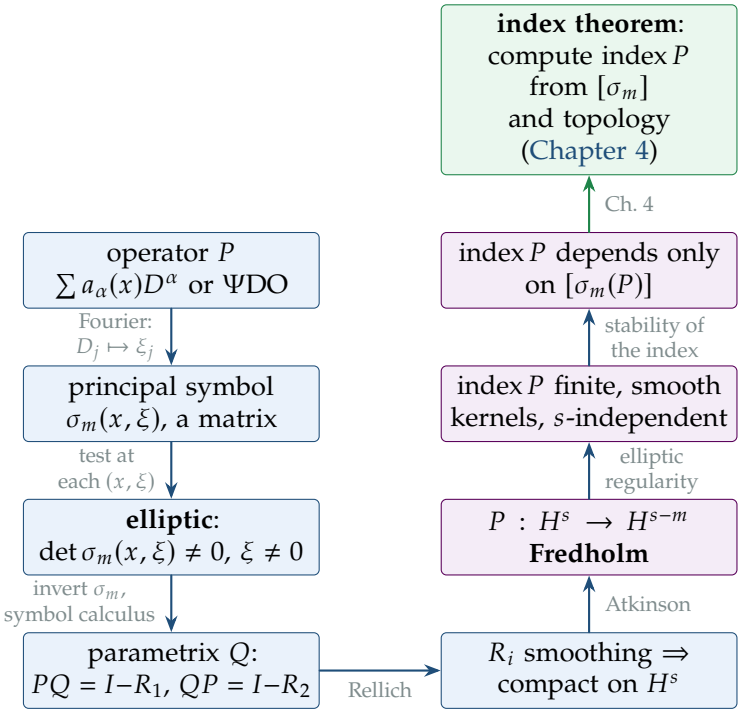


Figure 2.50: The logical spine of the chapter: from a polynomial to an integer. Down the left column we stay in *symbol space*: plane waves, polynomials in ξ, and inverting a matrix at each point of T^*M. At the bottom we cross into *function space*: the smoothing error of the parametrix is compact on Sobolev spaces (Rellich), and Atkinson's theorem turns "invertible modulo compact" into "Fredholm". Up the right column come the consequences. Each arrow is one tool of the chapter, named in grey.

One operator through every lens: the Laplacian on the circle

Almost every idea of the chapter can be seen, with every number explicit, on the simplest elliptic operator of all: $\Delta = -\mathrm{d}^2/\mathrm{d}\theta^2 = D_\theta^2$ on $S^1 = \mathbb{R}/2\pi\mathbb{Z}$, acting on $u = \sum_n c_n e^{in\theta}$ by $c_n \mapsto n^2 c_n$. Table 2.2 collects the facts. Three of the entries are worth deriving once more, because together they show the whole mechanism at work.

The Green's function, by solving an ODE. We want G with $\Delta_\theta G(\theta - \varphi) = \delta(\theta - \varphi) - \frac{1}{2\pi}$. The constant $\frac{1}{2\pi}$ has to be subtracted: by the solvability condition, $\Delta u = f$ needs $\int f = 0$, and

Concept	For $\Delta = D_\theta^2$ on S^1
symbol, principal symbol	$\sigma(\theta,\xi) = \xi^2 = \sigma_2(\theta,\xi)$: no lower-order terms
ellipticity	$\xi^2 \neq 0$ for $\xi \neq 0$; characteristic set empty
harmonic functions, Hodge	$\operatorname{Ker}\Delta = \{\text{constants}\}$, $\dim = 1 = b_0(S^1)$; on 1-forms, $\operatorname{Ker}\Delta_1 = \{c\,\mathrm{d}\theta\}$, $b_1 = 1$
Green's operator, kernel	$G(x) = \frac{\pi}{6} - \frac{x}{2} + \frac{x^2}{4\pi}$ on $[0,2\pi]$, $x = \theta - \varphi$: smooth off the diagonal, a kink on it (pseudo-locality)
parametrix	$Q = \sum_{n\neq 0} n^{-2}\,\Pi_n$; $Q\Delta = \Delta Q = I - \Pi_0$ with $\Pi_0 u = \frac{1}{2\pi}\int u$: smoothing, rank 1
Sobolev norm	$\|u\|_s^2 = 2\pi\sum_n (1+n^2)^s\,\lvert c_n\rvert^2$; $G \in H^s$ iff $s < \frac{3}{2}$
boundedness (Nash (2.95)–(2.96))	$\Delta : H^s \to H^{s-2}$ has norm exactly 1
Sobolev embedding	$H^s(S^1) \subset C^0$ for $s > \frac{1}{2}$; so G is continuous but G' jumps
Rellich	$H^s \hookrightarrow H^t$ $(s > t)$ is diagonal with entries $(1+n^2)^{(t-s)/2} \to 0$: compact
Fredholm, index	$\operatorname{Ker}\Delta = \operatorname{Ker}\Delta^* = \text{constants}$; $\operatorname{index}\Delta = 1 - 1 = 0$
solvability (Gauss's law)	$\Delta u = f$ solvable iff $\int_0^{2\pi} f\,\mathrm{d}\theta = 0$
spectrum, Weyl law	eigenvalues n^2 (multiplicity 2 for $n \neq 0$); $N(\lambda) = \#\{n^2 \leq \lambda\} \approx 2\sqrt{\lambda}$
heat trace	$\operatorname{Tr} e^{-t\Delta} = \sum_n e^{-tn^2} = \sqrt{\pi/t}\,\sum_k e^{-\pi^2k^2/t} \approx \sqrt{\pi/t}$ as $t \to 0$
complex powers, ζ, det	$\zeta_\Delta(s) = 2\zeta_R(2s)$, $\zeta_\Delta(0) = -1$, $\det'\Delta = (2\pi)^2$ (Section 2.4.14)

Table 2.2: The Laplacian on the circle as a one-page model of Chapter 2. Read down the right column: every row is a concept of the chapter reduced to a statement about the numbers n^2.

$\int \delta = 1$. Put $x = \theta - \varphi \in (0, 2\pi)$. Away from $x = 0$ the equation is $-G'' = -\frac{1}{2\pi}$, so

$$G(x) \overset{(1)}{=} \frac{x^2}{4\pi} + bx + c, \qquad G(2\pi) - G(0) \overset{(2)}{=} \pi + 2\pi b = 0 \;\Rightarrow\; b = -\tfrac{1}{2},$$

$$G'(0^+) - G'(2\pi^-) \overset{(3)}{=} b - (1 + b) = -1, \qquad \int_0^{2\pi} G \,\mathrm{d}x \overset{(4)}{=} \frac{2\pi^2}{3} - \pi^2 + 2\pi c = 0 \;\Rightarrow\; c = \frac{\pi}{6}.$$

(1) Integrate $G'' = \frac{1}{2\pi}$ twice; b, c are constants of integration.
(2) G must be continuous around the circle (periodic).
(3) The derivative jumps by -1 across $x = 0$, so $-G''$ contains $+\delta(x)$: this is automatic once b is fixed.
(4) We choose G orthogonal to the zero mode (the constants), which fixes c: $\int_0^{2\pi} \frac{x^2}{4\pi} = \frac{2\pi^2}{3}$ and $\int_0^{2\pi} bx = 2\pi^2 b = -\pi^2$.

So $G(x) = \frac{\pi}{6} - \frac{x}{2} + \frac{x^2}{4\pi}$. Its Fourier coefficients are $\frac{1}{2\pi n^2}$ for $n \neq 0$ and 0 for $n = 0$, the symbol $1/\xi^2$ sampled at the integers. The operator Q with this kernel is an integral operator of order -2. Its kernel is a polynomial away from the diagonal and has a kink on the diagonal, which is pseudo-locality (Section 2.4.15) in its simplest form.

The parametrix identity. For $u = \sum c_n e^{in\theta}$,

$$Q\Delta u \overset{(1)}{=} Q \sum_n n^2 c_n e^{in\theta} \overset{(2)}{=} \sum_{n\neq 0} \frac{n^2}{n^2}\, c_n e^{in\theta} \overset{(3)}{=} u - c_0 \overset{(4)}{=} \Big(I - \Pi_0\Big)u, \qquad \Pi_0 u = \frac{1}{2\pi}\int_0^{2\pi} u \,\mathrm{d}\theta.$$

(1) Δ multiplies the n-th coefficient by n^2.
(2) Q divides the n-th coefficient by n^2 for $n \neq 0$, and kills $n = 0$.
(3) Add and subtract the $n = 0$ term.
(4) c_0 is the mean of u.

The error Π_0 has the smooth kernel $\frac{1}{2\pi}$ and rank one. It is compact, so Δ is Fredholm by Atkinson's theorem (2.177). Its range is the image of $I - \Pi_0$, the functions of mean zero, which is the Gauss-law condition in Table 2.2.

The norm, Nash's (2.95)–(2.96).

$$\|\Delta u\|_{s-2}^2 \overset{(1)}{=} 2\pi \sum_n (1 + n^2)^{s-2}\, n^4\, |c_n|^2 \overset{(2)}{\leq} 2\pi \sum_n (1 + n^2)^s\, |c_n|^2 = \|u\|_s^2 ,$$

(1) Definition of the H^{s-2} norm, applied to the coefficients $n^2 c_n$ of Δu.
(2) $n^4 \leq (1 + n^2)^2$.

The ratio $n^4/(1 + n^2)^2$ tends to 1 as $n \to \infty$, so no constant smaller than 1 works: the norm is exactly 1 (Section 2.5.9). The operator loses exactly two derivatives, no more and no less, at every rung of the Sobolev ladder.

Every concept of the chapter

The tables below list every concept introduced in Chapter 2, one table per section of Nash. The middle column is the *operational* meaning: how to compute it or test for it. The last column says where the concept is used later in the book.

Nash §1, The meaning of ellipticity (Nash pp. 27–32)

Concept	Operational meaning	Used later
Fourier transform	$\partial_j \mapsto i\xi_j$: derivatives become multiplication by momentum; $\hat{u}(\xi) = \int e^{-ix\cdot\xi} u \, \mathrm{d}^n x$	everywhere: symbols, Sobolev norms, heat kernels
symbol	act on a plane wave $e^{ix\cdot\xi}$ and divide by it: $\sigma(x,\xi) = e^{-ix\cdot\xi} L e^{ix\cdot\xi}$	index theorem (Chapter 4)
elliptic (2nd order)	the quadratic form $a^{ij}\xi_i\xi_j$ is definite; its level sets are ellipsoids	Laplacians, $\not{D}^2$, Yang–Mills (Chapter 8)
hyperbolic, parabolic	form of Lorentzian signature / degenerate form plus a first-order term	wave and heat equations; heat kernel (Chapter 4)
elliptic regularity	Lu smooth $\Rightarrow u$ smooth; test: $\sigma_m \neq 0$ off $\xi = 0$	zero modes and harmonic forms are smooth
Green's function	$LG = \delta$; Δ on $\mathbb{R}^3$: $G = 1/(4\pi r)$; the propagator	parametrices; propagators in QFT
characteristic covector	$\sigma_m(x,\xi) = 0$ with $\xi \neq 0$: a direction singularities can travel along	light cones; elliptic = no characteristics
propagation of singularities	kinks move along light rays (Huygens) for hyperbolic L	contrast with pseudo-locality
Wick rotation	$t \mapsto -i\tau$ turns $\partial_t^2 - \nabla^2$ into $-\partial_\tau^2 - \nabla^2 = \Delta$	Euclidean QFT, instantons (Chapter 8), anomalies (Chapter 10)

Nash §2, Ellipticity and hypo-ellipticity (Nash pp. 32–36)

Concept	Operational meaning	Used later
multi-index	$\alpha = (\alpha_1, \dots, \alpha_n)$, $\lvert\alpha\rvert = \sum \alpha_j$, $\xi^\alpha = \prod \xi_j^{\alpha_j}$, $D^\alpha = \prod D_j^{\alpha_j}$	all symbol formulas
leading symbol σ_m	keep the terms of top degree m in ξ; homogeneous: $\sigma_m(x, \lambda\xi) = \lambda^m \sigma_m(x, \xi)$	symbol class in K-theory (Chapter 3, Chapter 4)
ellipticity (general)	$\sigma_m(x, \xi) \neq 0$ for all x and all $\xi \neq 0$	every index computation
σ_m on T^*M	ξ transforms as a covector under changes of coordinates, so σ_m is a function on T^*M	the index as a function of $[\sigma_m]$
hypo-ellipticity	$\operatorname{sing\,supp} Lu = \operatorname{sing\,supp} u$: L creates no smoothness and destroys none	heat-kernel methods
heat operator	$\partial_t - \nabla^2$: hypo-elliptic but not elliptic; the heat kernel is smooth for $t > 0$	McKean–Singer, heat-kernel proof (Chapter 4)
Schrödinger operator	$i\partial_t + \nabla^2$: not hypo-elliptic; singularities persist	real-time vs Euclidean QFT
Cauchy–Riemann $\bar\partial$	$\bar\partial = \frac{1}{2}(\partial_x + i\partial_y)$, $\sigma = \frac{i}{2}(\xi_1 + i\xi_2) \neq 0$: elliptic	Riemann surfaces (Chapter 5), CFT (Chapter 11)
analytic hypo-ellipticity	analytic data give analytic solutions (holomorphic functions)	holomorphic sections
elliptic ODE	order-m ODE with nowhere-zero leading coefficient	supersymmetric QM (Chapter 7)

Nash §3, Ellipticity and vector bundles (Nash pp. 36–41)

Concept	Operational meaning	Used later
vector bundle, section	a field with internal indices, glued by transition matrices $g_{\alpha\beta}(x)$	gauge fields (Chapter T3), every chapter
operator between bundles	a matrix of differential operators in each trivialisation	Dirac, $\mathrm{d}+\mathrm{d}^*$, $\bar\partial$, twisted operators
matrix symbol	replace D_j by ξ_j entrywise; the leading part is an $r\times s$ matrix	index theorem (Chapter 4)
pullback bundle π^*U	over each point (x,ξ) of T^*M, put the fibre U_x	the symbol class in $K(TM)$ (Chapter 3)
elliptic (bundles)	$\sigma_m(x,\xi): U_x \to V_x$ invertible for $\xi\neq 0$; needs $\operatorname{rank} U = \operatorname{rank} V$	every index problem
Hodge star, d^*	$\star$ from the metric; $\mathrm{d}^* = \pm\star\,\mathrm{d}\star$, the L^2 adjoint of d	Maxwell, Yang–Mills, gauge fixing
Hodge Laplacian	$\Delta_p = \mathrm{d}\mathrm{d}^* + \mathrm{d}^*\mathrm{d}$; $\sigma_2(\Delta_p) = \lvert\xi\rvert^2 I$	Hodge theorem; heat kernel
symbols of d, d^*	$\sigma(\mathrm{d}) = i\xi\wedge$, $\sigma(\mathrm{d}^*) = -i\,\iota_\xi$; composition multiplies leading symbols	de Rham, Dolbeault, signature complexes (Chapter 4)
elliptic complex	$\cdots\to\Gamma(E_i)\xrightarrow{P_i}\Gamma(E_{i+1})\to\ldots$, $P_{i+1}P_i=0$, symbol sequence exact off $\xi=0$	Dolbeault (Chapter 3), instanton deformations (Chapter 8)
harmonic form	$\Delta\omega=0 \iff \mathrm{d}\omega=0=\mathrm{d}^*\omega$ (on compact M)	zero modes; Chern–Simons, TQFT (Chapter 12)
Hodge theorem	one harmonic form per de Rham class: $\dim\operatorname{Ker}\Delta_p = b_p$	Euler characteristic and signature as indices
Hodge decomposition	$\Omega^p = \mathcal{H}^p \oplus \operatorname{Im}\mathrm{d}\oplus\operatorname{Im}\mathrm{d}^*$	gauge fixing, BRST
twisting	replace d by d_A with a connection; leading symbol unchanged	Yang–Mills, instanton moduli (Chapter 8)

Nash §4, Pseudo-differential operators (Nash pp. 41–47)

Concept	Operational meaning	Used later
operator with symbol	$Pu(x) = \int e^{ix\cdot\xi}\sigma(x,\xi)\hat{u}(\xi)\,\frac{\mathrm{d}^n\xi}{(2\pi)^n}$ (Section 2.4.3); Nash's (2.54)	inverses, square roots, propagators
Schwartz kernel	$K(x,y) = \int e^{i(x-y)\cdot\xi}\sigma(x,\xi)\frac{\mathrm{d}^n\xi}{(2\pi)^n}$: the inverse Fourier transform in ξ	propagators, heat kernels
symbol class S^m	$\left\lvert\partial_x^\beta\partial_\xi^\alpha\sigma\right\rvert \le C\langle\xi\rangle^{m-\lvert\alpha\rvert}$: each ξ-derivative gains one power of decay	the calculus
ΨDO on a manifold	locally a symbol in S^m in each chart, matrices for bundles (Section 2.4.5)	index theory of ΨDOs
principal symbol	$\sigma_m = \lim_{\mu\to\infty}\mu^{-m}\sigma(x,\mu\xi)$; positively homogeneous	symbol classes in K-theory
elliptic ΨDO	$\sigma_m(x,\xi)$ invertible for $\xi\neq 0$; order m elliptic is not order $m+1$ elliptic	Fredholm theory
composition	$\sigma_{PQ}\sim\sum_\gamma\frac{1}{\gamma!}\partial_\xi^\gamma\sigma_P\,D_x^\gamma\sigma_Q$; principal symbols multiply	parametrices
commutator	$\sigma_{[P,Q]} = i\{\sigma_P,\sigma_Q\}+\ldots$: quantum commutator ↔ Poisson bracket	semiclassics; anomalies
adjoint	$\sigma_m(P^*) = \sigma_m(P)^\dagger$; P elliptic $\iff$ P^* elliptic	finite cokernels
parametrix	Q with symbol $\sigma_m^{-1}+\ldots$; $PQ = I-R_1$, $QP = I-R_2$, R_i smoothing	Fredholm property, regularity
Bessel potential	$(a^2+\Delta)^\lambda$, kernel $\propto r^{-\nu}K_\nu(ar)$, $\nu=\frac{n}{2}+\lambda$; Yukawa e^{-ar}/r for $n=3$, $\lambda=-1$	massive propagators
Riesz potential	Δ^λ, kernel $\propto\lvert x\rvert^{-n-2\lambda}$: fractional derivatives and integrals	fractional powers, conformal dimensions
logarithmic symbol	$(a^2+\lvert\xi\rvert^2)^\lambda\ln(a^2+\lvert\xi\rvert^2)$: in $S^{2\lambda+\varepsilon}$ for all $\varepsilon>0$, not in $S^{2\lambda}$	ζ', determinants
smoothing operator	symbol in $S^{-\infty}$, i.e. smooth kernel; e.g. the Gaussian $e^{-t\Delta}$	heat kernel, McKean–Singer
complex powers	$(a^2+\Delta)^z$ has order $2\operatorname{Re}z$, elliptic; trace gives $\zeta(s)$	det′ (Chapter 9, Chapter 12, Chapter 10)
locality	$\operatorname{supp}Pu\subset\operatorname{supp}u$: true exactly for differential operators (Peetre)	
pseudo-locality	$\operatorname{sing\,supp}Pu\subset\operatorname{sing\,supp}u$: kernel smooth off the diagonal	regularity, heat-kernel localisation

Nash §5, Pseudo-differential operators and Sobolev spaces (Nash pp. 47–55)

Concept	Operational meaning	Used later
weak derivative	$\int(\partial^\alpha u)\phi = (-1)^{\lvert\alpha\rvert}\int u\,\partial^\alpha\phi$ for all test functions ϕ	distributions
L^p_k, H^k	derivatives up to order k in L^p (in L^2 for H^k), Nash's (2.73)–(2.74)	function spaces for gauge fields (Chapter 6, Chapter 8)
H^s, real s	$\lVert u\rVert_s^2 = \int(1+\lvert\xi\rvert^2)^s\,\lvert\hat u\rvert^2\,\frac{\mathrm{d}^n\xi}{(2\pi)^n}$; on S^1: $2\pi\sum(1+n^2)^s\,\lvert c_n\rvert^2$	domain of Fredholm operators
negative orders	$H^{-s} = (H^s)^*$ via the L^2 pairing; $\delta\in H^{-s}(\mathbb{R}^n)$ iff $s>\frac{n}{2}$	sources, distributional fields
$\mathcal{D},\mathcal{S},\mathcal{E}$ and duals	test functions (compact support, rapid decay, all smooth) and the distributions dual to them	Green's functions, QFT fields
Gelfand triple	$\mathcal{S}\subset L^2\subset\mathcal{S}'$: kets, the Hilbert space, bras (plane waves, δ)	Dirac notation made honest
Sobolev scale	$H^\infty\subset\cdots\subset H^1\subset H^0 = L^2\subset H^{-1}\subset\cdots\subset H^{-\infty}$	bookkeeping for orders
Sobolev (i)	$k-\frac{n}{p}>l\Rightarrow L^p_k\subset C^l$; power counting $\int\langle\xi\rangle^{2l-2k}\mathrm{d}^n\xi<\infty$	regularity of zero modes; UV divergences
Sobolev (ii)	$k\ge l$, $k-\frac{n}{p}\ge l-\frac{n}{q}\Rightarrow L^p_k\subset L^q_l$	critical dimensions of Yang–Mills
compact operator	maps bounded sets to precompact sets; a norm limit of finite-rank operators	Fredholm theory
Rellich	$H^s\hookrightarrow H^t$ compact for $s>t$ on compact M; so smoothing operators are compact	discrete spectrum, finite index
ΨDO boundedness	order $m\Rightarrow P: H^s\to H^{s-m}$ bounded for every s, Nash's (2.92)	Fredholm property
elliptic estimate	$\lVert u\rVert_s\le C(\lVert Pu\rVert_{s-m}+\lVert u\rVert_t)$; regularity: $Pu\in H^{s-m}\Rightarrow u\in H^s$	smoothness of kernels
operator norm	$\Delta: H^k\to H^{k-2}$ on S^1 has norm 1 (Nash (2.94)–(2.96))	

Fredholm operators and the index (Nash pp. 54–55, expanded in Section 2.6)

Concept	Operational meaning	Used later
index of a matrix	$m \times n$ matrix: $\dim\mathrm{Ker} - \dim\mathrm{Coker} = n - m$ whatever the entries	the model for everything
cokernel	$H_2/\mathrm{Im}\,T \cong \mathrm{Ker}\,T^*$ when $\mathrm{Im}\,T$ is closed: the obstructions to solving $Tu = f$	solvability conditions
Fredholm operator	bounded, $\dim\mathrm{Ker}$ and $\dim\mathrm{Coker}$ finite (so closed range)	Chapter 4, Chapter 3
index	$\mathrm{index}\,T = \dim\mathrm{Ker}\,T - \dim\mathrm{Ker}\,T^*$	the analytical side of the index theorem
shift operators	$Se_n = e_{n+1}$ on ℓ^2: $\mathrm{index}\,S = -1$, $\mathrm{index}\,S^* = +1$; every integer occurs	Toeplitz operators, spectral flow
Atkinson's theorem	Fredholm $\iff$ invertible modulo compact operators; Nash's (2.97)	elliptic $\Rightarrow$ Fredholm
elliptic $\Rightarrow$ Fredholm	$P : H^s \to H^{s-m}$ Fredholm; kernel and $\mathrm{Ker}\,P^*$ smooth, independent of s	every index
solvability	$Pu = f$ solvable $\iff$ $f \perp \mathrm{Ker}\,P^*$; Gauss's law on a closed space	constraints, Gauss law
stability	index locally constant, unchanged by compact (lower-order) terms, depends only on $[\sigma_m]$	topological invariance of the index
spectral flow	for $D_\theta + a$ on S^1, eigenvalues $n + a$ cross zero as a varies; the index stays 0	anomalies (Chapter 10)
Toeplitz operator	$e^{ik\theta}\Pi_+ + \Pi_-$: index $= -k$, minus the winding number of the symbol	a first index theorem
$\mathrm{d} + \mathrm{d}^*$	even $\to$ odd forms: index $= \sum(-1)^q b_q = \chi(M)$	Gauss–Bonnet as an index theorem
McKean–Singer	$\mathrm{index}\,P = \mathrm{Tr}\,e^{-tP^*P} - \mathrm{Tr}\,e^{-tPP^*}$ for every $t > 0$	heat-kernel proof, Fujikawa
SUSY oscillator	$P = \frac{\mathrm{d}}{\mathrm{d}x} + x$: index 1, the Witten index	Morse theory (Chapter 7)
space of Fredholm ops	$\pi_0(\mathcal{F}) \cong \mathbb{Z}$ via the index; families give $K(X)$	families index, global anomalies

Formulas to keep at your fingertips

F1. *Plane-wave test.* $D^\alpha e^{ix\cdot\xi} = \xi^\alpha e^{ix\cdot\xi}$, so $L = \sum a_\alpha D^\alpha$ has $\sigma_L(x,\xi) = \sum a_\alpha(x)\xi^\alpha$ and $\sigma_m(x,\xi) = \sum_{|\alpha|=m} a_\alpha(x)\xi^\alpha$.

F2. *Ellipticity.* $\det\sigma_m(x,\xi) \neq 0$ for $\xi \neq 0$; equivalently $\left|\sigma_m(x,\xi)^{-1}\right| \leq C\,|\xi|^{-m}$.

F3. *Hodge.* $\sigma_2(\Delta_p)(\xi) = (i\xi\wedge)(-i\iota_\xi) + (-i\iota_\xi)(i\xi\wedge) = |\xi|^2\, I$, and $\dim\mathrm{Ker}\,\Delta_p = b_p(M)$.

F4. *Composition.* $\sigma_{PQ} \sim \sum_\gamma \frac{1}{\gamma!}\,\partial_\xi^\gamma \sigma_P\, D_x^\gamma \sigma_Q$; $\sigma_{m+m'}(PQ) = \sigma_m(P)\sigma_{m'}(Q)$; $\sigma_{[P,Q]} = i\{\sigma_P, \sigma_Q\} + \ldots$; $\sigma_m(P^*) = \sigma_m(P)^\dagger$.

F5. *Parametrix.* $q \sim \sigma_m^{-1} + \ldots$, $PQ = I - R_1$, $QP = I - R_2$, with R_1, R_2 smoothing.

F6. *Sobolev norm.* $\|u\|_s^2 = \int (1 + |\xi|^2)^s\, |\hat{u}(\xi)|^2\, \frac{\mathrm{d}^n\xi}{(2\pi)^n}$; on T^n: $(2\pi)^n \sum_k (1 + |k|^2)^s\, |c_k|^2$.

F7. *Sobolev embedding.* $H^s \subset C^l$ if $s > l + \frac{n}{2}$; the smoothness deficit is $\frac{n}{p}$ ($\frac{n}{2}$ in L^2).

F8. *Mapping property.* Order $m \Rightarrow H^s \to H^{s-m}$ bounded; smoothing $\Rightarrow H^s \to H^\infty = C^\infty$ (compact M).

F9. *Fredholm.* $\operatorname{index} P = \dim \operatorname{Ker} P - \dim \operatorname{Ker} P^*$, finite, s-independent, unchanged by lower-order terms.

F10. *McKean–Singer.* $\operatorname{index} P = \operatorname{Tr} e^{-tP^*P} - \operatorname{Tr} e^{-tPP^*}$ for all $t > 0$.

Where the chapter is used later

The tools of this chapter are used in almost every later chapter. Figure 2.51 shows the main routes.

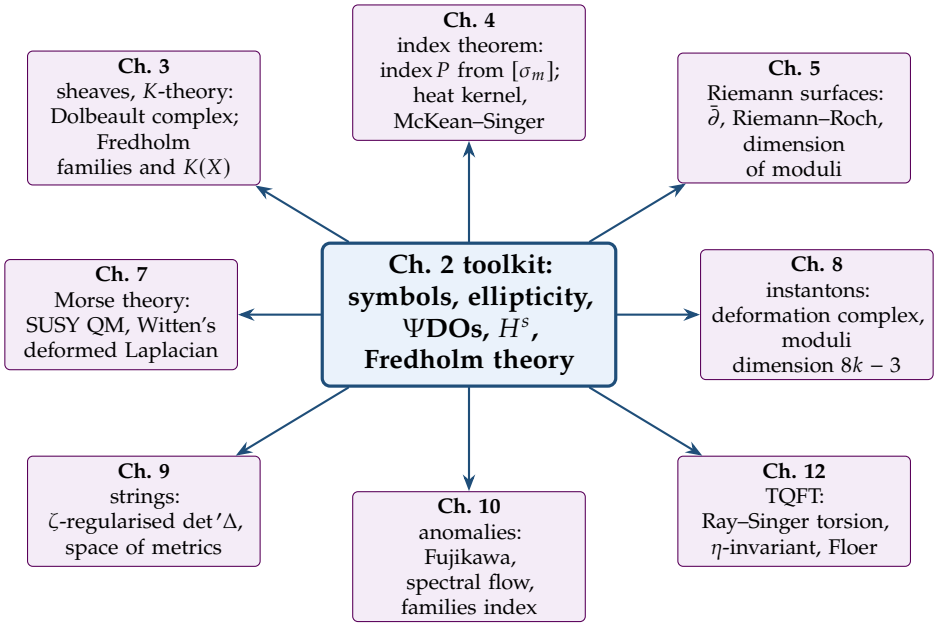


Figure 2.51: Where Chapter 2 is used. Each purple box names a later chapter and the tool it takes from this one. The index theorem (top) uses the whole spine of Figure 2.50. The other chapters mostly use one piece of it: complex powers and ζ-determinants (strings, TQFT), heat kernels and spectral flow (anomalies), the index of an elliptic complex (instantons, Riemann surfaces), or Fredholm families (K-theory).

- **Counting moduli.** The dimension of a moduli space (of instantons in Chapter 8, of complex structures on a Riemann surface in Chapter 5) is the index of the elliptic complex that linearises the defining equations. The index is stable, so the dimension can be computed from topology alone.
- **Counting zero modes.** Fermion zero modes in an instanton background are $\operatorname{Ker} \not{D}_A$; their net chirality is an index (Chapter 4), and it drives the anomalies of Chapter 10.
- **Regularising determinants.** One-loop amplitudes are det′ of elliptic operators, defined through ζ-functions of complex powers (Chapter 9, Chapter 12).
- **Function spaces.** Spaces of connections $\mathcal{A}$ and gauge groups $\mathcal{G}$ are completed in Sobolev norms, so that they become Hilbert manifolds and Lie groups (Chapter 6, Chapter 8).

Watch out: The five confusions of this chapter

1. *Elliptic is about the top order only.* $\Delta + V(x)$ is elliptic for every potential V, and $\Delta - m^2$ is elliptic although it has zero modes for special m. Ellipticity does not mean invertible. It means *invertible modulo smoothing*.
2. *Elliptic is not positive.* $\bar\partial$ and $\not{D}$ are elliptic with complex or indefinite symbols. What matters is that $\det \sigma_m \neq 0$, not that $\sigma_m > 0$.
3. *Signs and 2π's.* Nash's Fourier conventions differ from ours by the sign in the exponent (see the Fourier dictionary in Section 2.4.2). Symbols of odd-order operators change sign accordingly. Ellipticity and the index do not.
4. *Cokernel is not kernel of the adjoint in general.* The identification $\operatorname{Coker} P \cong \operatorname{Ker} P^*$ needs a closed range. For elliptic operators on compact manifolds the range is closed, and the identification holds.
5. *Compactness is essential.* On $\mathbb{R}^n$, Δ is elliptic but not Fredholm on L^2: its spectrum $[0, \infty)$ reaches 0 continuously. Finiteness of the index needs a compact manifold, or a confining potential as in the SUSY oscillator.

Try it: Self-test on the whole chapter

(a) Is $\partial_x^2 - \partial_y^2 + \partial_x$ elliptic? Is $\partial_x^4 + \partial_y^4$?
(b) What is the principal symbol of $\sqrt{\Delta + m^2}$, and its order?
(c) For which s is the square wave in $H^s(S^1)$? Is it continuous?
(d) On T^2, what is index $\bar\partial$? Why does adding a smooth potential $V(z)$ not change it?
(e) Why is $\operatorname{index}(\Delta + 1) = 0$ on a compact M without computing anything?

Answers. (a) The first has $\sigma_2 = -\xi_1^2 + \xi_2^2$, which vanishes on the lines $\xi_1 = \pm\xi_2$, so it is not elliptic (it is hyperbolic). The second has $\sigma_4 = \xi_1^4 + \xi_2^4 > 0$ for $\xi \neq 0$, so it is elliptic. (b) $\sigma_1 = |\xi|$, order 1; it is elliptic, and non-local but pseudo-local. (c) Its coefficients fall like $1/|n|$, so it is in H^s exactly for $s < \frac{1}{2}$. Consistently with Sobolev embedding (which needs $s > \frac{1}{2}$), it is not continuous. (d) Zero: the kernel and cokernel are the constants. V is of order 0, a compact perturbation, so the index is unchanged. (e) $\Delta + 1$ is self-adjoint, so its index is 0. In fact it is invertible, since $\langle u, (\Delta + 1)u\rangle = \|\mathrm{d}u\|^2 + \|u\|^2 > 0$ for $u \neq 0$.

Reading Nash (Chapter II, pp. 27–55: section map)

Nash section	Nash pages, equations	In these notes
§1 The meaning of ellipticity	pp. 27–32, (2.1)–(2.18)	part 2a, first section
§2 Ellipticity and hypo-ellipticity	pp. 32–36, (2.19)–(2.37)	part 2a, second section
§3 Ellipticity and vector bundles	pp. 36–41, (2.38)–(2.53)	part 2a, third section; Section 2.4.1
§4 Pseudo-differential operators	pp. 41–47, (2.54)–(2.70)	Section 2.4
§5 ΨDOs and Sobolev spaces	pp. 47–54, (2.71)–(2.93)	Section 2.5
§5, last page	pp. 54–55, (2.94)–(2.97)	Section 2.6

Nash's chapter ends abruptly, with the parametrix identity (2.97) and the remark that elliptic operators are Fredholm. He takes up the index itself in his Chapter 4 (Chapter 4) and the space of Fredholm operators at the end of his Chapter 3 (Chapter 3). Our Section 2.6 bridges that gap, so that when the index theorem arrives, every word in its statement already has an operational meaning.

Chapter 3

Cohomology of Sheaves and Bundles

Where this chapter is going

Physics is full of objects that are easy to write down *locally* and hard to write down *globally*. A vector potential A for a magnetic monopole exists on the northern hemisphere and on the southern hemisphere, but not on the whole sphere. A wavefunction of a charged particle in the monopole field is an honest function on each hemisphere, but only a "twisted" function on the sphere. A solution of a differential equation exists near every point, but may fail to exist everywhere at once. In each case the local pieces are fine, and the interesting physics lives in the *mismatch* between them.

This chapter builds the machinery that measures such mismatches. It has two halves. The first half (this part, Nash §1–§2) introduces *sheaves*, which are the bookkeeping device for "data defined on open sets that can be restricted and glued", and *sheaf cohomology*, which is the group that counts the obstructions to gluing. The payoff is a pair of theorems that say that two very different-looking computations give the same answer: the de Rham theorem (combinatorial cohomology with real coefficients equals closed forms modulo exact forms) and the Dolbeault theorem (sheaf cohomology of holomorphic forms equals $\bar\partial$-closed forms modulo $\bar\partial$-exact forms). Along the way we will see that complex line bundles, and hence $U(1)$ gauge fields such as the Dirac monopole, are classified by a sheaf cohomology group, and that Dirac's quantisation condition is literally the cocycle condition of that cohomology.

The second half of the chapter (Nash §3–§6, later parts) turns to *vector bundles as algebraic objects*: K-theory, Bott periodicity, characteristic classes (Chern classes, the Chern character, the Todd class, the $\hat A$-genus), and the relation between Fredholm operators and $K(X)$. All of that is the raw material of the Atiyah–Singer index theorem in Chapter 4. Figure 3.1 shows how the pieces feed into each other.

Physical picture: One sentence for the whole chapter

A sheaf remembers what you are allowed to do *locally*; its cohomology measures what goes wrong when you try to do it *globally*. For a physicist: H^0 counts the global solutions, H^1 counts the ways local solutions can be patched together with a twist (gauge transformations on overlaps), and H^2 counts the twists of twists (the quantised magnetic charges).

3.1 Sheaves

Nash opens the chapter with a question that every physicist has met without naming it: *when does something that is true locally become true globally?*

Some properties pass from local to global for free. If every small open set of a connected

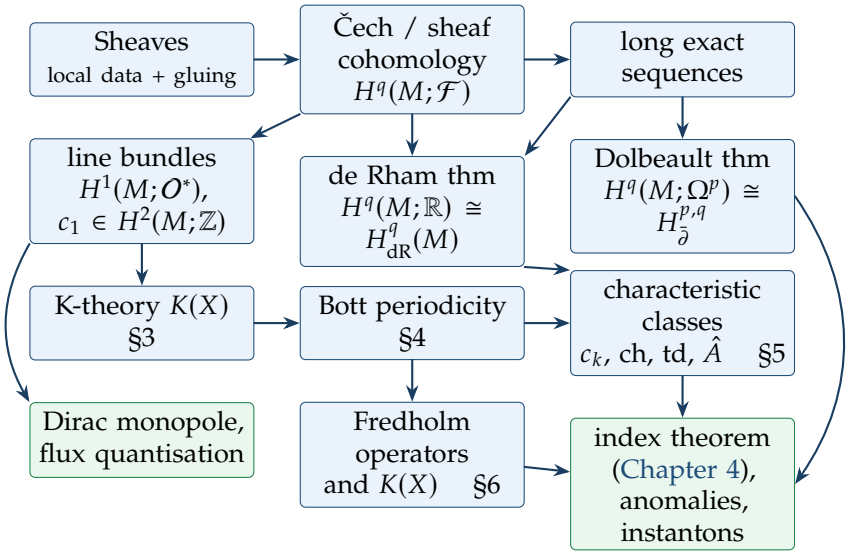


Figure 3.1: The logic of Chapter 3. Blue boxes are mathematics, green boxes are where the physics comes out. This part covers the top two rows (sheaves, their cohomology, the de Rham and Dolbeault theorems, line bundles and c_1). The later parts of the chapter build K-theory and characteristic classes, which feed, together with the Dolbeault complex, into the index theorem.

manifold looks like a piece of $\mathbb{R}^n$, the manifold has dimension n: there is nothing more to check. Other properties do not pass at all. Nash's example is scalar curvature. On a flat torus the scalar curvature is zero everywhere, so it is "locally constant"; but on a surface built from a flat cylinder capped by two round hemispheres, the curvature is locally constant near every point (zero on the cylinder, positive on the caps) and yet not constant globally. "Locally constant" and "constant" differ, and the difference is detected by the global shape of the space.

The physicist's version of the same question is sharper.

- Maxwell: $\mathrm{d}F = 0$ means that near every point there is a potential with $F = \mathrm{d}A$ (Poincaré lemma). Is there a single A on the whole of space? Not if there is a monopole.
- Quantum mechanics: the Schrödinger equation can be solved near every point. Is there a single-valued global wavefunction? Not always: Aharonov–Bohm phases and monopoles force the wavefunction to be a section of a nontrivial line bundle.
- Complex analysis: every nonvanishing holomorphic function has a logarithm near every point. Is there a global logarithm? Not for z on $\mathbb{C} \setminus \{0\}$.

A **sheaf** is the device that packages "the things we are allowed to write down on each open set" together with the two operations that every one of these examples uses: *restricting* data from a big set to a small set, and *gluing* data on small sets into data on a big set. Sheaf *cohomology* (Section 3.3) then measures the failure of gluing. Nash warns the reader not to be put off by the abstract appearance of the definition. We follow his advice and his order: definition, then lots of examples.

3.1.1 Local data on open sets: presheaves

Fix a topological space M; for us it will almost always be a smooth manifold. The basic example to keep in mind is:

$$\mathcal{E}(U) \;=\; \{\text{smooth real functions } f : U \to \mathbb{R}\}, \qquad U \subset M \text{ open}.$$

Three facts about this assignment are so obvious that we normally never say them.

(a) Each $\mathcal{E}(U)$ is an abelian group under pointwise addition (in fact a vector space, even a ring).

(b) If $U \subset V$, a smooth function on the big set V can be *restricted* to the small set U. Restriction respects addition: $(f+g)|_U = f|_U + g|_U$.

(c) Restricting in two steps is the same as restricting in one step, and "restricting" U to itself does nothing.

Anything with these three properties is a presheaf.

Working definition 3.1 (presheaf)

A **presheaf** $\mathcal{F}$ of abelian groups on M assigns

- to every open set $U \subset M$ an abelian group $\mathcal{F}(U)$ [the allowed local objects on U, which can be added, like the smooth functions on U], whose elements are called **sections of $\mathcal{F}$ over U**;
- to every pair of open sets $U \subset V$ a group homomorphism [a map that respects addition, $r(s+t) = r(s) + r(t)$] $r^V_U : \mathcal{F}(V) \to \mathcal{F}(U)$, called **restriction**,

such that

$$r^U_U = \mathrm{id}, \qquad r^W_U = r^V_U \circ r^W_V \quad \text{whenever } U \subset V \subset W. \tag{3.1}$$

We usually write $s|_U$ for $r^V_U(s)$. *Operational test:* decide what the "allowed local objects" on U are, check they can be added (or multiplied, for a multiplicative sheaf), and check that restricting an allowed object gives an allowed object.

Figure 3.2 shows the picture. Restriction goes from big sets to small sets. Nash's gloss is worth repeating: the first rule says restricting from U to U leaves s unchanged; the second says restricting from W to U may be done via an intermediate set V.

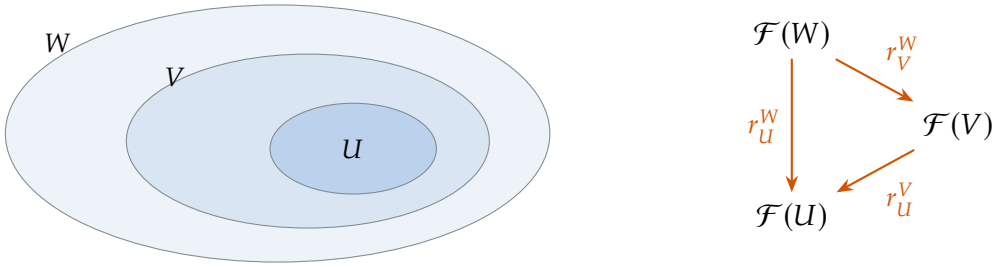


Figure 3.2: Nested open sets $U \subset V \subset W$ (left) and the restriction maps between their groups of sections (right). Restriction always goes from a bigger set to a smaller one, and the triangle commutes: $r^W_U = r^V_U \circ r^W_V$.

Watch out: Arrows go the "wrong" way

The open sets are included $U \hookrightarrow V$, but the sections go $\mathcal{F}(V) \to \mathcal{F}(U)$. Bigger sets carry *fewer* sections (there are more conditions to satisfy), smaller sets carry *more*. Keeping this reversal straight saves a lot of confusion later.

We also set $\mathcal{F}(\varnothing) = 0$ (the trivial group). This is the only sensible choice: there is exactly one function on the empty set.

3.1.2 The two gluing axioms: sheaves

A presheaf only knows how to go from big to small. The interesting direction is from small to big: given data on the pieces of a cover, can we assemble global data? For smooth functions the answer is "yes, if the pieces agree on overlaps, and then uniquely". This is what makes a presheaf a sheaf.

Working definition 3.2 (sheaf)

A presheaf $\mathcal{F}$ is a **sheaf** if for every open set U and every open cover $U = \bigcup_i U_i$ the following hold.

(S1) **Locality** [a section is determined by its pieces]. If $s_1, s_2 \in \mathcal{F}(U)$ and $s_1|_{U_i} = s_2|_{U_i}$ for every i, then $s_1 = s_2$.

(S2) **Gluing** [pieces that agree on overlaps assemble into one section]. If $s_i \in \mathcal{F}(U_i)$ is given for every i and $s_i|_{U_i \cap U_j} = s_j|_{U_i \cap U_j}$ for all i, j, then there is a section $s \in \mathcal{F}(U)$ with $s|_{U_i} = s_i$ for every i.

(By (S1) the glued section in (S2) is unique.) This is Nash's condition (3.1). *Operational test:* the defining condition for the allowed objects must be checkable point by point, or at least in an arbitrarily small neighbourhood of each point. Conditions of that kind are called **local**, and local conditions always give sheaves.

The two axioms pull in opposite directions, as Nash points out. (S1) says a section over U is *determined* by its restrictions: global $\to$ local loses no information. (S2) says compatible local data can be *assembled*: local $\to$ global is possible whenever the pieces are compatible. What sheaf cohomology will measure is the more interesting situation where local pieces exist but are *not* compatible, and we ask whether they can be *made* compatible.

For groups there is a neat way to package (S1) and (S2) at once. Write $U_{ij} = U_i \cap U_j$. Consider

$$0 \longrightarrow \mathcal{F}(U) \xrightarrow{\;r\;} \prod_i \mathcal{F}(U_i) \xrightarrow{\;\delta\;} \prod_{i<j} \mathcal{F}(U_{ij}), \qquad r(s) = (s|_{U_i})_i, \quad \delta(s_i)_{ij} = s_j|_{U_{ij}} - s_i|_{U_{ij}}. \tag{3.2}$$

Then (S1) says r is injective, and (S2) says every element killed by δ is in the image of r. In other words: $\mathcal{F}$ *is a sheaf exactly when* (3.2) *is exact*. We will meet the map δ again: it is the first Čech coboundary.

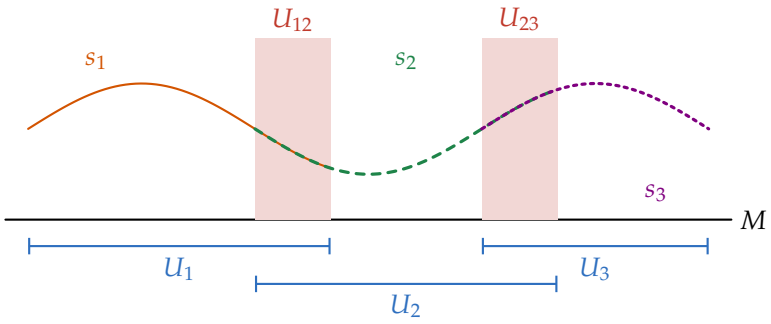


Figure 3.3: The gluing axiom for smooth functions on an interval covered by three open sets. The local sections s_1, s_2, s_3 (solid, dashed, dotted) agree on the overlaps U_{12} and U_{23} (shaded), so they are pieces of one global smooth function, and that function is unique.

Worked example 3.1: Smooth functions form a sheaf

Nash leaves as an exercise the check that $\mathcal{E}(U) = C^\infty(U, \mathbb{R})$ with ordinary restriction is a sheaf. Let us do it.
Presheaf. $C^\infty(U)$ is a group under addition. The restriction of a smooth function to a smaller open set is smooth (smoothness at a point only involves a neighbourhood of the point). The rules (3.1) hold because restriction of functions is literally "forget the values outside U".
Locality (S1). If $s_1|_{U_i} = s_2|_{U_i}$ for every i, then for any $x \in U$ pick i with $x \in U_i$; then $s_1(x) = s_2(x)$. So $s_1 = s_2$ as functions.
Gluing (S2). Given compatible s_i, define $s : U \to \mathbb{R}$ by

$$s(x) \overset{(1)}{=} s_i(x) \quad \text{for any } i \text{ with } x \in U_i.$$

(1) This is well defined: if also $x \in U_j$, then $x \in U_{ij}$ and $s_i(x) = s_j(x)$ by compatibility.

Near each x the function s coincides with the smooth function s_i on the open set U_i, so all derivatives of s exist at x. Hence $s \in C^\infty(U)$ and $s|_{U_i} = s_i$. The key point is that smoothness is a *local* condition: it can be checked in an arbitrarily small neighbourhood of each point.

Exactly the same argument works for many other local conditions, which gives Nash's first list of sheaves.

- C^0: continuous functions.
- $\mathcal{E}$: smooth (C^∞) functions. Nash's notation, after Wells; the letter $\mathcal{E}$ is traditional (it goes back to Schwartz's $\mathcal{E}$ for smooth test functions).
- $\mathcal{A}$: real-analytic functions (those equal to their convergent Taylor series near each point).
- $\mathcal{O}$: holomorphic functions on a complex manifold of complex dimension n (in local coordinates $z^1, \dots, z^n$, functions that are complex-differentiable in each variable, equivalently $\partial f / \partial \bar{z}^i = 0$; see Section 3.7.3).

Each is a sheaf because continuity, smoothness, analyticity and holomorphicity are all checked point by point.

Worked example 3.2: Constant functions: a presheaf that is not a sheaf

Let $M = \mathbb{R}$ and let $\mathcal{P}(U) = \{\text{constant functions } U \to \mathbb{R}\} \cong \mathbb{R}$ for $U \neq \varnothing$, with the obvious restrictions. This is a presheaf. Take $U = (0,1) \cup (2,3)$, covered by $U_1 = (0,1)$ and $U_2 = (2,3)$, and let $s_1 = 0$, $s_2 = 1$. The overlap is empty, so the compatibility condition $s_1|_{U_{12}} = s_2|_{U_{12}}$ is vacuous (both sides live in $\mathcal{P}(\varnothing) = 0$). Gluing would need a *constant* function on U equal to 0 on one interval and 1 on the other. None exists, so (S2) fails.
The repair is to ask for **locally constant** functions: functions that are constant near each point. Then the glued function "0 on $(0,1)$, 1 on $(2,3)$" is allowed, and we get a sheaf. We denote it by $\underline{\mathbb{R}}$, or just $\mathbb{R}$, and call it the **constant sheaf**. Its sections are

$$\underline{\mathbb{R}}(U) \;=\; \mathbb{R}^{\#\{\text{connected components of } U\}}. \tag{3.3}$$

The same construction with values in $\mathbb{Z}$, $\mathbb{C}$ or any group G gives the constant sheaves $\underline{\mathbb{Z}}$, $\underline{\mathbb{C}}$, $\underline{G}$.

Watch out: "The sheaf of constant functions"

Nash (p. 59–60) says that $\mathbb{R}$ can be thought of as "the sheaf of constant functions". He means *locally* constant functions, as in Example 3.2. Being constant is not a local property (it compares values at far-away points); being locally constant is. On a connected M the two agree for global sections, which is why the sloppy name survives.

Worked example 3.3: Bounded holomorphic functions are not a sheaf

This is Nash's non-example (taken from Wells). Let $M = \mathbb{C}$ and

$$\mathcal{F}(U) = \{f : U \to \mathbb{C} \text{ holomorphic and bounded}\}.$$

Sums and restrictions of bounded holomorphic functions are bounded and holomorphic, so this is a presheaf. Now cover $\mathbb{C}$ by the discs $D_R = \{|z| < R\}$, $R = 1, 2, 3, \ldots$, and on each disc take

$$s_R(z) = z \in \mathcal{F}(D_R), \qquad \text{bounded because } |s_R(z)| < R \text{ on } D_R.$$

On any overlap $D_R \cap D_{R'} = D_{\min(R,R')}$ the functions agree. If (S2) held they would glue to a bounded holomorphic s on $\mathbb{C}$ with $s(z) = z$ everywhere. But

$$\sup_{z\in\mathbb{C}} |s(z)| \overset{(1)}{=} \sup_{z\in\mathbb{C}} |z| \overset{(2)}{=} \infty,$$

(1) s restricts to $s_R(z) = z$ on every disc and the discs cover $\mathbb{C}$.
(2) $|z|$ is unbounded on $\mathbb{C}$.

So the gluing axiom fails. Nash's version of the argument uses Liouville's theorem: a bounded holomorphic function on all of $\mathbb{C}$ is constant, so the only global sections are constants, while there are many non-constant local sections that agree on overlaps. *Why it fails:* "bounded" is not a local condition. Near any single point every continuous function is bounded; boundedness is about behaviour far away.

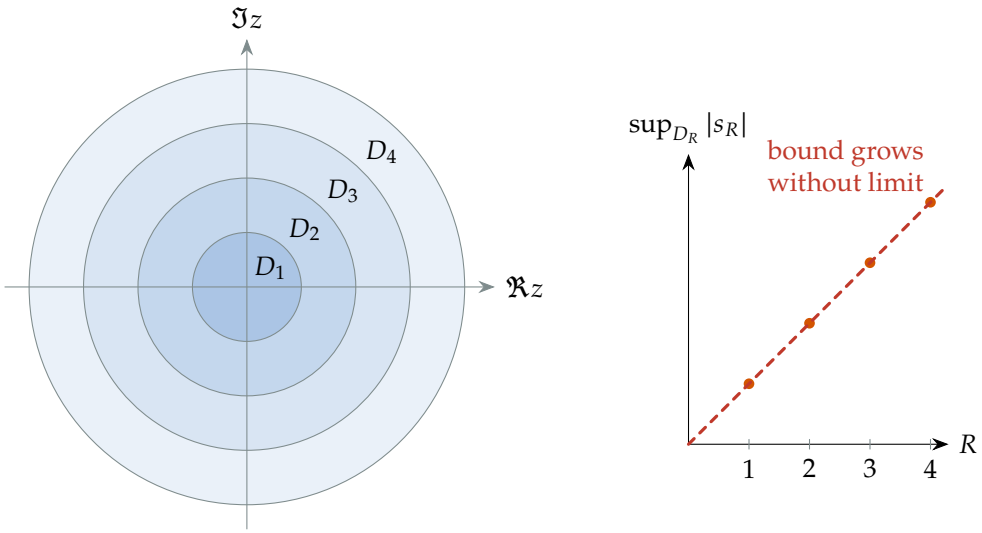


Figure 3.4: Left: the discs D_R exhaust $\mathbb{C}$. On each one, $s_R(z) = z$ is a bounded holomorphic function, and neighbouring s_R agree on overlaps. Right: the bound on D_R is R, which grows without limit, so the glued function is not bounded. "Bounded" is not a local condition, and the gluing axiom fails.

Physical picture: Local conditions are the ones physics writes down

Differential equations, regularity conditions and gauge conditions are checked at points; they give sheaves. Conditions such as "bounded", "normalisable", "finite total energy" or "constant" are global; they do not. This is why the local-field-theory language of a physicist sits so naturally in sheaf language: a local Lagrangian produces equations of motion whose solutions form a sheaf (the sheaf of local solutions), and all the global subtleties are then pushed into cohomology.

3.1.3 Germs and stalks

As U gets bigger, $\mathcal{F}(U)$ gets smaller, because there are more conditions to satisfy. Conversely, as U shrinks down around a point x, $\mathcal{F}(U)$ gets bigger. Nash asks what we get "in the limit" when U shrinks to the point x. The answer is not "the value at x"; it is something richer, called the **germ** at x.

Here is the idea for smooth functions. Two functions f and g defined near x have the *same germ at x* if they agree on *some* neighbourhood of x, however small. So the germ remembers the value $f(x)$, all derivatives of f at x, and more, but it forgets what f does at any fixed distance from x.

Working definition 3.3 (germ and stalk)

Let $x \in M$. Consider pairs (U, s) with U an open neighbourhood of x and $s \in \mathcal{F}(U)$. Declare

$$(U,s) \sim (V,t) \quad \Longleftrightarrow \quad s|_W = t|_W \text{ for some open } W \ni x,\ W \subset U \cap V.$$

The equivalence class of (U, s) is the **germ** s_x of s at x. The set of germs is the **stalk**

$$\mathcal{F}_x \;=\; \varinjlim_{U \ni x} \mathcal{F}(U). \tag{3.4}$$

This is Nash's (3.2); the symbol $\varinjlim$ is the **direct limit** [here it just produces the set of germs defined above, as U shrinks towards x] over neighbourhoods of x ordered by reverse inclusion. *Operational test:* to decide whether two local sections have the same germ at x, ask whether they agree on some small ball around x.

The stalk is a group: add germs by adding representatives on a common neighbourhood, $s_x + t_x = (s|_W + t|_W)_x$. This does not depend on the representatives chosen, because shrinking W further does not change the germ.

Worked example 3.4: Three stalks, three sizes

(a) Locally constant sheaf. For $\underline{\mathbb{R}}$ and any x, a small connected neighbourhood carries exactly one real number's worth of sections, so

$$\underline{\mathbb{R}}_x \cong \mathbb{R}.$$

(b) Holomorphic functions. A holomorphic function near $x = 0 \in \mathbb{C}$ is determined on a small disc by its Taylor series (identity theorem), and any power series with positive radius of convergence defines one. Hence

$$\mathcal{O}_0 \;\cong\; \mathbb{C}\{z\} \;=\; \Big\{ \sum_{n \geq 0} a_n z^n \;:\; \text{radius of convergence} > 0 \Big\}.$$

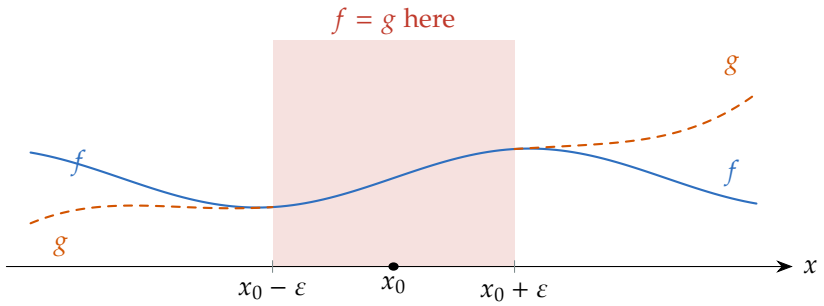


Figure 3.5: Two functions with the same germ at x_0. They agree on the shaded neighbourhood $(x_0 - \varepsilon, x_0 + \varepsilon)$ and differ outside it. The stalk $\mathcal{E}_{x_0}$ cannot tell them apart: a germ remembers everything about f infinitely close to x_0 and nothing at a finite distance.

The germ at a point knows the function on its whole disc of convergence, and (by analytic continuation) essentially on the whole connected domain. *(c) Smooth functions.* The map "germ $\mapsto$ Taylor series" $\mathcal{E}_0 \to \mathbb{R}[[x]]$ is *not* injective. The function

$$\phi(x) = \begin{cases} e^{-1/x^2} & x \neq 0 \\ 0 & x = 0 \end{cases} \qquad \text{has} \qquad \phi^{(k)}(0) \overset{(1)}{=} \lim_{x \to 0} P_k(1/x)\, e^{-1/x^2} \overset{(2)}{=} 0$$

(1) By induction, the k-th derivative away from 0 is a polynomial P_k in $1/x$ times e^{-1/x^2} (differentiate once more and use the product rule).

(2) e^{-1/x^2} goes to zero faster than any power of $1/x$ blows up (substitute $u = 1/x^2$ and compare $u^m e^{-u} \to 0$).

So ϕ has zero Taylor series at 0 but a nonzero germ there (it is positive at every $x \neq 0$, arbitrarily close to 0). Smooth germs carry information invisible to all derivatives.

This contrast (b) versus (c) is the root of one of the most important differences in the chapter: smooth functions are "floppy" and can be cut off, holomorphic functions are "rigid" and cannot. The floppy sheaves will turn out to have no higher cohomology (Section 3.2.5); all the interesting cohomology comes from rigid sheaves such as $\underline{\mathbb{Z}}$, $\underline{\mathbb{R}}$, $\mathcal{O}$ and $\mathcal{O}^*$.

The sheaf as a space of germs

Nash's footnote on p. 58 says that a sheaf is sometimes called a "sheaf of germs". This refers to an equivalent, more geometric picture. Collect all the stalks into one set,

$$\mathrm{Et}(\mathcal{F}) = \bigsqcup_{x \in M} \mathcal{F}_x, \qquad \pi : \mathrm{Et}(\mathcal{F}) \to M, \quad \pi(s_x) = x.$$

This is the **étalé space** ("spread-out space"). Each section $s \in \mathcal{F}(U)$ gives a map $x \mapsto s_x$ from U into $\mathrm{Et}(\mathcal{F})$, which is a section of π in the ordinary sense of bundles: a choice of one point in each "fibre" $\mathcal{F}_x$ (Figure 3.6). One topologises $\mathrm{Et}(\mathcal{F})$ so that these maps are continuous, and then *the continuous sections of π over U are exactly $\mathcal{F}(U)$*. So a sheaf is "a bundle-like object whose sections are the things we care about", except that the fibres $\mathcal{F}_x$ can be huge and the space is usually very non-Hausdorff. For calculations we never need the étalé space, but it explains the name: a sheaf (French *faisceau*, a bundle of stalks of wheat) is literally a bundle of stalks.

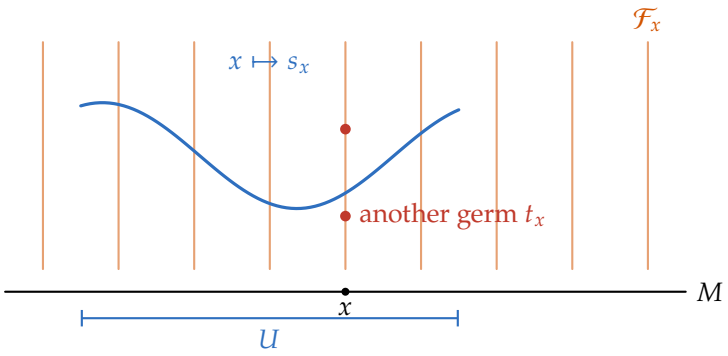


Figure 3.6: The étalé space of a sheaf: above each point x sits the stalk $\mathcal{F}_x$ (orange). A section s over U becomes a continuous choice of one germ in each stalk (blue curve). Two sections that agree near x pass through the same point of $\mathcal{F}_x$; a section with a different germ at x passes through a different point (red).

Reading Nash (p. 57, (3.2))

Nash introduces the stalk with the single line $\mathcal{F}_x = \varinjlim_{x \in U} \mathcal{F}(U)$ and the remark that stalks are used when we discuss sheaf cohomology. We will use them in exactly one place, but it is a crucial one: *a sequence of sheaves is called exact if it is exact on every stalk* (Section 3.2.3). Exactness on stalks is exactness "infinitely close to each point", which is the right notion of "locally exact".

3.1.4 Sheaves of sections of vector bundles

Nash's second rich source of sheaves is vector bundles. Recall (Chapter 1, Chapter T2) that a rank-k vector bundle $E \to M$ is described by an open cover $\{U_i\}$, local trivialisations $E|_{U_i} \cong U_i \times \mathbb{R}^k$ (or $\mathbb{C}^k$), and **transition functions** $g_{ij} : U_{ij} \to \mathrm{GL}(k)$ telling us how to translate between trivialisations. With the convention we will use throughout this part,

$$s_i = g_{ij}\, s_j \quad \text{on } U_{ij}, \tag{3.5}$$

where $s_i : U_i \to \mathbb{R}^k$ is the local component vector of a section s in the i-th trivialisation. Consistency on triple overlaps ($s_i = g_{ij}s_j = g_{ij}g_{jk}s_k$ must equal $g_{ik}s_k$) forces the **cocycle condition**

$$g_{ij}\, g_{jk} = g_{ik} \quad \text{on } U_{ijk} = U_i \cap U_j \cap U_k, \qquad g_{ii} = 1, \quad g_{ji} = g_{ij}^{-1}. \tag{3.6}$$

Remember this equation: it will reappear as "$\delta g = 1$" in Čech cohomology.

Working definition 3.4 (sheaf of sections)

For a vector bundle $E \to M$ let $\mathcal{F}(U) = \Gamma(U, E)$ be the smooth (or continuous, or, if M and E are complex, holomorphic) sections of E over U [maps that pick one vector $s(x)$ in the fibre E_x over each point $x \in U$], with restriction the ordinary restriction of sections. In terms of a trivialising cover, a section over U is a collection of vector-valued functions s_i on $U \cap U_i$ obeying (3.5).

Being a smooth section is a local condition, so this is a sheaf by the same argument as Example 3.1. Nash's point is that almost every sheaf we need is of this kind, for a bundle E "reflecting the geometry of M".

- $E = M \times \mathbb{R}$ (trivial line bundle): sections are functions, and we get back $\mathcal{E}$.

- $E = T^*M$: sections are smooth 1-forms. For $E = \Lambda^p T^*M$ the sections are smooth p-forms; Nash writes $\mathcal{E}^p$ for this sheaf, so $\mathcal{E}^p(U) = \Omega^p(U)$ and $\mathcal{E}^0 = \mathcal{E}$.

- M complex, E the holomorphic cotangent bundle (spanned by $\mathrm{d}z^1, \dots, \mathrm{d}z^n$) and its exterior powers: the sheaf Ω^p of *holomorphic* p-forms, $\omega = \frac{1}{p!}\omega_{i_1 \dots i_p}(z)\,\mathrm{d}z^{i_1} \wedge \cdots \wedge \mathrm{d}z^{i_p}$ with holomorphic coefficients. In particular $\Omega^0 = \mathcal{O}$.

- $\mathcal{O}^*$: nowhere-vanishing holomorphic functions. Sums of nonvanishing functions can vanish ($1 + (-1) = 0$), so $\mathcal{O}^*$ is not a group under addition. It is a group under *multiplication*, and Nash stresses that $\mathcal{O}^*$ is to be viewed as a **multiplicative sheaf**: its "zero" is the constant function 1, and "$f - g$" means f/g. Similarly $\mathcal{E}^*$ denotes nowhere-vanishing smooth complex functions.

Worked example 3.5: The Möbius bundle as a sheaf

Cover S^1 (angle θ) by two arcs $U_1 = \{-\varepsilon < \theta < \pi + \varepsilon\}$ and $U_2 = \{\pi - \varepsilon < \theta < 2\pi + \varepsilon\}$. The overlap has two components, A near $\theta = 0$ and B near $\theta = \pi$. The trivial line bundle has $g_{12} = +1$ on both components; the **Möbius bundle** has

$$g_{12} = +1 \text{ on } B, \qquad g_{12} = -1 \text{ on } A.$$

There are no triple overlaps, so the cocycle condition is empty. A section is a pair of functions s_1 on U_1 and s_2 on U_2 with $s_1 = s_2$ on B and $s_1 = -s_2$ on A. Near any point the sheaf of sections looks exactly like $\mathcal{E}$ (on a small arc the section is just a function). Globally, a continuous section must vanish somewhere:

$$\begin{aligned} s_1(\text{near } 0) &\overset{(1)}{=} -s_2(\text{near } 2\pi) \\ &\overset{(2)}{=} -\big(\text{value obtained by continuing } s \text{ once round}\big), \end{aligned}$$

(1) The transition rule on A, where the angle 0 in U_1 is the angle 2π in U_2.
(2) Going from $\theta = 0$ in U_1 to $\theta = \pi$, across B with no sign change, then on through U_2 up to 2π, we follow the section continuously round the circle.

So a continuous section returns with its sign flipped, and by the intermediate value theorem it has a zero. The sheaves "smooth functions" and "smooth sections of the Möbius bundle" agree on every small open set (all their stalks are the same), but their global sections differ. Telling such sheaves apart is exactly what H^1 will do (Example 3.15).

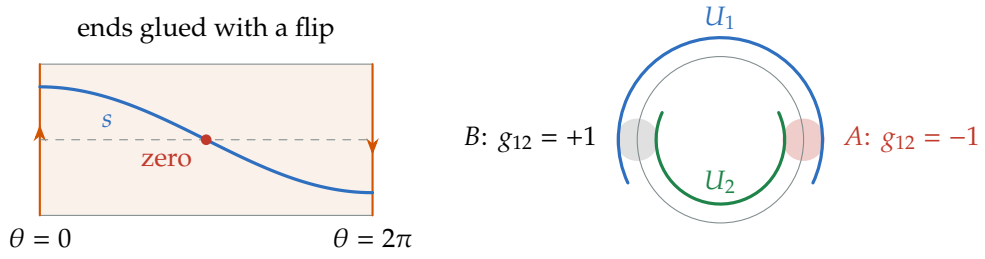


Figure 3.7: Left: the Möbius bundle over S^1, cut open along $\theta = 0$. The two ends of the strip are glued with a flip (the arrows point opposite ways), so a section must return to minus its starting value. The blue section $s(\theta) = 0.7\cos(\theta/2)$ therefore crosses zero (red dot). Right: the same bundle as gluing data. Two arcs overlap in two pieces; on A the transition function is -1, on B it is $+1$.

Worked example 3.6: The line bundles $O(n)$ over $\mathbb{CP}^1$ and their sections

Cover the Riemann sphere $\mathbb{CP}^1 = \mathbb{C} \cup \{\infty\}$ by $U_0 = \mathbb{C}$ with coordinate z and $U_1 = \mathbb{CP}^1 \setminus \{0\}$ with coordinate $w = 1/z$. The overlap is $U_{01} = \mathbb{C}^*$. For each integer n define a holomorphic line bundle $\mathcal{O}(n)$ by the single transition function

$$g_{01}(z) = z^{\,n} \in \mathcal{O}^*(U_{01}), \qquad \text{i.e.} \quad s_0(z) = z^n\, s_1(1/z) \text{ on } \mathbb{C}^*. \tag{3.7}$$

(Again no triple overlaps, so no cocycle condition.) Which global holomorphic sections are there? Write $s_0(z) = \sum_{k\ge 0} a_k z^k$, an entire function. Then

$$\begin{aligned} s_1(w) &\overset{(1)}{=} w^{\,n}\, s_0(1/w) \\ &\overset{(2)}{=} \sum_{k\ge 0} a_k\, w^{\,n-k}, \end{aligned}$$

(1) Invert (3.7): $s_1(1/z) = z^{-n} s_0(z)$, and put $z = 1/w$.
(2) Substitute the Taylor series of s_0.

For s_1 to be holomorphic at $w = 0$ we need $a_k = 0$ whenever $n - k < 0$. So

$$\Gamma(\mathbb{CP}^1, \mathcal{O}(n)) = \{\, a_0 + a_1 z + \cdots + a_n z^n \,\}, \qquad \dim \Gamma(\mathbb{CP}^1, \mathcal{O}(n)) = \begin{cases} n+1, & n \ge 0, \\ 0, & n < 0. \end{cases}$$

All the bundles $\mathcal{O}(n)$ look alike locally (every stalk is $\mathcal{O}_x$), but the dimensions of their global sections differ. For $n = 0$ we recover Liouville's theorem on the Riemann sphere: the only global holomorphic functions are the constants. We will compute the "higher" version of this count, $H^1(\mathbb{CP}^1; \mathcal{O}(n))$, in Example 3.18; the difference $\dim H^0 - \dim H^1 = n + 1$ is the simplest instance of the Riemann–Roch theorem, which Chapter 4 and Chapter 5 develop.

Coefficients in a bundle

Nash next combines the geometric sheaves with an *additional* bundle E. The sheaf $\mathcal{E}^p(E)$ of **p-forms with values in E** has sections

$$\mathcal{E}^p(E)(U) = \Gamma\big(U, \Lambda^p T^*M \otimes E\big). \tag{3.8}$$

This is Nash's (3.3). Concretely: over an open set U on which E is trivial, with a frame $e_1, \dots, e_k$, such a section is

$$\omega = \sum_{a=1}^{k} \omega^a \otimes e_a, \qquad \omega^a \in \Omega^p(U),$$

that is, a column vector of ordinary p-forms. On overlaps the column vectors are related by the transition matrices of E. This is familiar from gauge theory: the curvature of a connection on E is a section of $\mathcal{E}^2(\operatorname{End} E)$, a matrix of 2-forms that transforms by conjugation (Chapter T3). The holomorphic analogue is $\Omega^p(E)$. Using direct sums and tensor products of bundles we manufacture further sheaves such as $\mathcal{E}(E \otimes F)$ and $\mathcal{O}(E \oplus E \oplus \cdots)$.

Dictionary: Sheaves a physicist already uses

Sheaf	Physics meaning
$\mathcal{E}$, $\mathcal{E}^p$	smooth scalar fields, smooth p-form fields
$\underline{\mathbb{R}}$, $\underline{\mathbb{Z}}$	locally constant quantities: coupling constants, quantised charges
$\mathcal{E}^1(\mathfrak{g})$	local gauge potentials A on each patch
$\mathcal{E}^*$, $\mathcal{O}^*$	local $U(1)$ gauge transformations $e^{i\lambda}$ (smooth, or holomorphic)
$\Gamma(\,\cdot\,, L)$ for a line bundle L	wavefunctions of a charged particle in a background field
$\mathcal{O}$, Ω^p	holomorphic fields, e.g. chiral fields on a worldsheet

3.1.5 Why analysts like sheaves

Nash (p. 58) explains why sheaves are valuable in analysis. Two situations make differential forms awkward.

- *Non-compact manifolds.* Global integrals may diverge, and compactness arguments fail; but local statements ("near every point there is a solution") remain true.
- *Singular spaces.* Complex analytic spaces such as the cone $\{z_1^2 + z_2^2 + z_3^2 = 0\} \subset \mathbb{C}^3$ have singular points where there are no good local coordinates, so "differential form" is not even well defined there. But "holomorphic function on an open set" still makes sense (restrict holomorphic functions of z_1, z_2, z_3), so the sheaf $\mathcal{O}$ still exists.

The methods of partial differential equations are essentially local: they produce solutions near points. Sheaves are exactly the tool that takes local solvability and asks what survives globally. Combining the two gives a theory that works on this wider class of spaces. For us the main beneficiaries will be the Dolbeault theorem (Section 3.7) and, in Chapter 5, algebraic geometry, where sheaves are the basic language.

Reading Nash (pp. 56–59, (3.1)–(3.3))

Nash's §1 is a definition, three examples ($\mathcal{E}$, $\mathcal{A}$, $\mathcal{O}$), one non-example (bounded holomorphic functions, via Liouville), and a list of sheaves of sections: $\mathcal{E}^p$, Ω^p, $\Omega^0 = \mathcal{O}$, the multiplicative sheaf $\mathcal{O}^*$, and forms with values in a bundle, $\mathcal{E}^p(E)$ with sections $\Gamma(M, \Lambda^p T^*M \otimes E)$, his (3.3). His condition (ii), equation (3.1), is our pair of axioms (S1) locality and (S2) gluing, written for a cover $U = \bigcup U_i$. His "stalk" (3.2) is our Eq. (3.4); his footnote identifies stalks with germs. Everything in the section is local bookkeeping; the global content arrives in §2.

Recap

- A *presheaf* assigns a group $\mathcal{F}(U)$ to each open set, with restriction maps from big sets to small sets.
- A *sheaf* also satisfies locality (a section is determined by its restrictions) and gluing (compatible local sections glue). Local conditions (continuity, smoothness, holomorphicity, being a section of a bundle, satisfying a differential equation) give sheaves; global conditions (boundedness, constancy) do not.

- Main examples: $\mathcal{C}^0, \mathcal{E}, \mathcal{A}, \mathcal{O}$; $\mathcal{E}^p$, Ω^p; $\mathcal{O}^*$ and $\mathcal{E}^*$ (multiplicative); locally constant $\underline{\mathbb{R}}, \underline{\mathbb{Z}}$; sections of any bundle, e.g. Möbius or $\mathcal{O}(n)$; $\mathcal{E}^p(E)$.
- The *stalk* $\mathcal{F}_x$ is the group of germs at x: sections near x modulo agreement on a smaller neighbourhood.
- Different sheaves can have identical stalks everywhere but different global sections (trivial vs Möbius, $\mathcal{O}$ vs $\mathcal{O}(n)$). Measuring this is the job of cohomology.

3.2 Morphisms and exact sequences of sheaves

Nash's §2 begins by recalling ordinary cohomology $H^i(M; G)$ with coefficients in an abelian group G such as $\mathbb{Z}$, $\mathbb{R}$ or $\mathbb{Q}$ (Chapter 1). Sheaf cohomology replaces the fixed group G by a sheaf such as $\mathcal{E}$ or $\mathcal{O}$. The two theories overlap: the locally constant sheaf $\underline{\mathbb{R}}$ gives back $H^i(M; \mathbb{R})$.

Before defining anything, Nash makes a remark that organises the whole section: *cohomology is a measure of how far some sequence is from being exact*. So we first need sequences of sheaves, and a notion of exactness for them. This section supplies both, and then shows the basic phenomenon: a sequence of sheaves can be exact (locally) while its global sections are not. That failure is what sheaf cohomology measures.

3.2.1 Reminder: exact sequences of groups

A sequence of abelian groups and homomorphisms

$$A \xrightarrow{\alpha} B \xrightarrow{\beta} C$$

is **exact at** B if $\operatorname{Im}\alpha = \operatorname{Ker}\beta$. Read it as two statements: $\beta \circ \alpha = 0$ ("everything coming from A dies in C") and, conversely, everything that dies in C came from A. A **short exact sequence**

$$0 \longrightarrow A \xrightarrow{\alpha} B \xrightarrow{\beta} C \longrightarrow 0 \tag{3.9}$$

is exact at A, B and C. Exactness at A means α is injective (its kernel is the image of 0). Exactness at C means β is surjective (its image is the kernel of $C \to 0$, which is all of C). So A is a subgroup of B and $C \cong B/A$.

Worked example 3.7: Three short exact sequences to keep in mind

(a) $0 \to \mathbb{Z} \xrightarrow{\times 2} \mathbb{Z} \to \mathbb{Z}_2 \to 0$: the even integers inside the integers, with quotient "parity".

(b) $0 \to \mathbb{Z} \xrightarrow{i} \mathbb{R} \xrightarrow{e} U(1) \to 0$ with $e(x) = e^{2\pi i x}$. Check exactness in the middle:

$$e(x) = 1 \overset{(1)}{=} \iff e^{2\pi i x} = 1 \overset{(2)}{=} \iff x \in \mathbb{Z} = \operatorname{Im} i.$$

(1) Definition of e.

(2) $e^{2\pi i x} = \cos 2\pi x + i \sin 2\pi x$ equals 1 exactly when x is an integer.

And e is surjective because every unit complex number is $e^{i\phi}$ for some real ϕ; take $x = \phi/2\pi$. This sequence is the "one-point" version of the exponential sequence below: a phase is a real number modulo integers.

(c) $0 \to \mathbb{R} \to C^\infty(\mathbb{R}) \xrightarrow{\mathrm{d}/\mathrm{d}x} C^\infty(\mathbb{R}) \to 0$: the kernel of differentiation is the constants, and every smooth function is a derivative (integrate it). On the circle, the last map is *not* surjective, which is the first sign of topology; see Example 3.11.

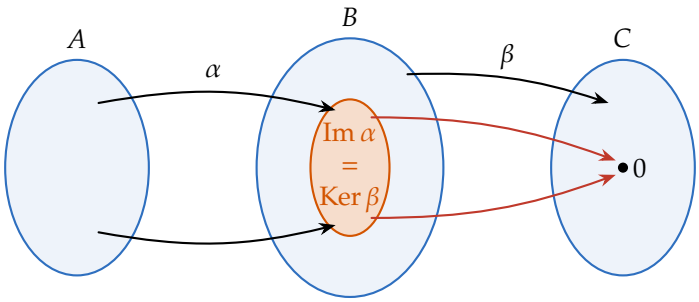


Figure 3.8: Exactness at B: the image of α (orange) is precisely the set of elements that β sends to 0 (red arrows). Elements of B outside the orange region have nonzero image in C.

3.2.2 Sheaf morphisms

A map between sheaves must act on sections over every open set, and it must not care whether we restrict before or after applying it.

Working definition 3.5 (sheaf morphism)

A **sheaf morphism** $m : \mathcal{A} \to \mathcal{B}$ is a collection of group homomorphisms [maps respecting addition, $m_U(s+t) = m_U(s) + m_U(t)$] $m_U : \mathcal{A}(U) \to \mathcal{B}(U)$, one for each open U, that commute with restriction: for $U \subset V$,

$$\begin{array}{ccc} \mathcal{A}(V) & \xrightarrow{m_V} & \mathcal{B}(V) \\ {\scriptstyle r_U^V}\downarrow & & \downarrow{\scriptstyle r_U^V} \\ \mathcal{A}(U) & \xrightarrow[m_U]{} & \mathcal{B}(U) \end{array} \qquad \text{i.e.} \qquad m_V(s)\big|_U = m_U\big(s|_U\big). \tag{3.10}$$

This is Nash's diagram (3.5). *Operational test:* the operation must be *local*: the value of $m(s)$ near a point depends only on s near that point. Pointwise operations and differential operators pass; integral operators (which smear over distances) usually fail.

Worked example 3.8: Morphisms we will use

- The exterior derivative $\mathrm{d} : \mathcal{E}^p \to \mathcal{E}^{p+1}$. Commuting with restriction is the statement that $\mathrm{d}(\omega|_U) = (\mathrm{d}\omega)|_U$: derivatives at a point only see a neighbourhood of the point.
- The inclusion $i : \underline{\mathbb{Z}} \to \mathcal{O}$ (a locally constant integer-valued function is a holomorphic function) and the inclusion $\underline{\mathbb{R}} \to \mathcal{E}$.
- The exponential $\exp : \mathcal{O} \to \mathcal{O}^*$, $f \mapsto e^{2\pi i f}$. It turns addition into multiplication: $e^{2\pi i(f+g)} = e^{2\pi i f} e^{2\pi i g}$, so it is a homomorphism from the additive group $\mathcal{O}(U)$ to the multiplicative group $\mathcal{O}^*(U)$.
- The operator $\bar\partial : \mathcal{E}^{p,q} \to \mathcal{E}^{p,q+1}$ on a complex manifold (Section 3.7).

A *non-example*: $\mathcal{E}(\mathbb{R}) \ni f \mapsto F(x) = \int_0^x f$ is not defined open set by open set in a way that commutes with restriction: on $U = (5, 6)$ the integral from 0 needs values of f outside U.

A morphism m induces maps on stalks, $m_x : \mathcal{A}_x \to \mathcal{B}_x$, $m_x(s_x) = (m_U s)_x$. This is well defined precisely because m commutes with restriction: if s and s' agree near x, so do $m(s)$ and $m(s')$.

3.2.3 Exact sequences of sheaves: exact on stalks

What should it mean for a sequence of sheaves

$$\mathcal{E} \xrightarrow{e} \mathcal{F} \xrightarrow{f} \mathcal{G} \tag{3.11}$$

to be exact? (This is Nash's (3.4)/(3.6); his $\mathcal{E}$ here is a generic sheaf, not the sheaf of smooth functions.) The naive answer, "exact on U for every open U", is too strong: it would already fail for the most basic example (the Poincaré lemma sequence below on the circle). The right answer is "exact near every point", and the precise version uses stalks.

Working definition 3.6 (exact sequence of sheaves)

The sequence (3.11) of sheaf morphisms is **exact** if for every $x \in M$ the induced sequence of stalks [germs at x, i.e. sections remembered only on arbitrarily small neighbourhoods of x]

$$\mathcal{E}_x \xrightarrow{e_x} \mathcal{F}_x \xrightarrow{f_x} \mathcal{G}_x \tag{3.12}$$

is an exact sequence of groups [the image of e_x is exactly the set of germs that f_x sends to 0] (Nash's (3.7)). *Operational test:* (i) $f \circ e = 0$; (ii) if $f(t) = 0$ for a section t near x, then *on some smaller neighbourhood of* x, $t = e(s)$ for some s.

The phrase "on some smaller neighbourhood" is the whole point. It is the sheaf-theoretic version of the physicist's "locally". Surjectivity of a sheaf map $f : \mathcal{F} \to \mathcal{G}$ means: every section of $\mathcal{G}$ can be lifted *near each point*, not necessarily globally.

Worked example 3.9: The exponential sequence

On any complex manifold M (for instance an open set in $\mathbb{C}$), the sequence

$$0 \longrightarrow \underline{\mathbb{Z}} \xrightarrow{i} \mathcal{O} \xrightarrow{\exp} \mathcal{O}^* \longrightarrow 0, \qquad \exp(f) = e^{2\pi i f}, \tag{3.13}$$

is exact. This is Nash's (3.8). Recall that $\mathcal{O}^*$ is multiplicative, so its "zero" is the constant function 1. We check the three exactness conditions at a point x.

Exact at $\underline{\mathbb{Z}}$ *(*i *injective).* A locally constant integer function that is zero as a holomorphic function is zero. Nothing to do.

Exact at $\mathcal{O}$ ($\operatorname{Ker}\exp = \operatorname{Im} i$). Let f be holomorphic near x with $\exp f = 1$. Then on a connected neighbourhood W of x

$$\begin{aligned} e^{2\pi i f(z)} = 1 \ \ \forall z \in W &\overset{(1)}{\Longrightarrow} \ f(z) \in \mathbb{Z} \ \ \forall z \in W \\ &\overset{(2)}{\Longrightarrow} \ f \text{ constant on } W \\ &\overset{(3)}{\Longrightarrow} \ f_x \in \operatorname{Im} i_x. \end{aligned}$$

(1) $e^{2\pi i w} = 1$ exactly when $w \in \mathbb{Z}$ (Example 3.7(b), now for complex w: $\left|e^{2\pi i w}\right| = e^{-2\pi \Im w} = 1$ forces $\Im w = 0$).

(2) A continuous function from a connected set to the discrete set $\mathbb{Z}$ is constant. (Nash's phrase "the exactness follows trivially from $(\exp \circ i)(n) = e^{2\pi i n} = 1$" covers the other inclusion, $\operatorname{Im} i \subset \operatorname{Ker}\exp$.)

(3) A constant integer function is in the image of i.

Exact at $\mathcal{O}^*$ (exp *surjective on stalks).* Let g be holomorphic and nonzero near x. Choose a small disc (or polydisc) W around x on which g is nonzero. On a disc, g'/g is holomorphic

and has a primitive, and

$$f(z) := \frac{1}{2\pi i}\Big(\log g(x) + \int_x^z \frac{g'(w)}{g(w)}\,\mathrm{d}w\Big) \quad \text{satisfies}$$
$$\frac{\mathrm{d}}{\mathrm{d}z}\Big(g\,e^{-2\pi i f}\Big) \overset{(1)}{=} (g' - 2\pi i f' g)e^{-2\pi i f} \overset{(2)}{=} 0,$$

(1) Product and chain rules.
(2) By construction $2\pi i f' = g'/g$.

so $g\,e^{-2\pi i f}$ is constant on W, and it equals 1 at $z = x$ by the choice of the constant $\log g(x)$ (any branch). Hence $g = e^{2\pi i f}$ on W. The integral is path-independent because a disc is simply connected (Cauchy's theorem).

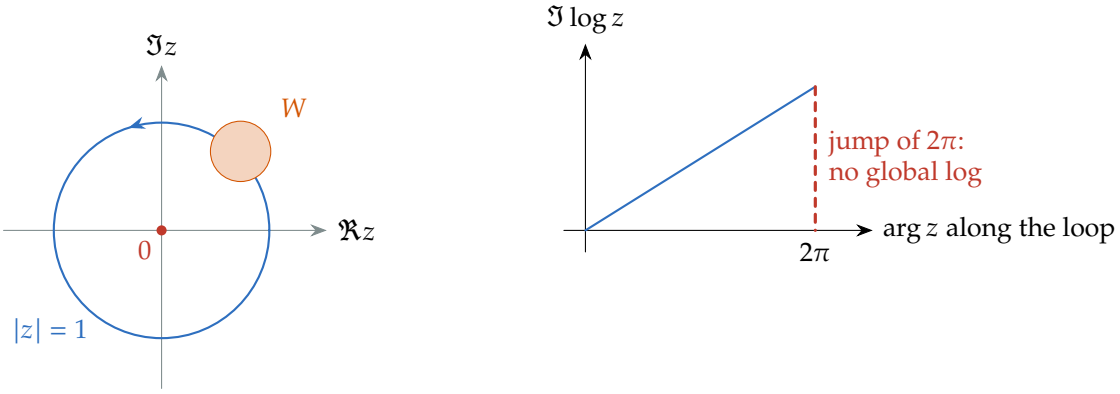


Figure 3.9: The function $g(z) = z$ on $M = \mathbb{C}^*$. Near any point, e.g. on the small disc W (orange), z has a logarithm, so $z = e^{2\pi i f}$ locally and exp is surjective on stalks. Following the logarithm once round the loop $|z| = 1$ (left) adds $2\pi i$ (right), so there is no global f with $z = e^{2\pi i f}$: exp is not surjective on global sections.

Worked example 3.10: Poincaré-lemma sequences

Let $\mathcal{Z}^p \subset \mathcal{E}^p$ be the sheaf of *closed* p-forms ($\mathrm{d}\omega = 0$). Then for $p \geq 0$

$$0 \to \underline{\mathbb{R}} \to \mathcal{E}^0 \xrightarrow{\mathrm{d}} \mathcal{Z}^1 \to 0, \qquad 0 \to \mathcal{Z}^p \hookrightarrow \mathcal{E}^p \xrightarrow{\mathrm{d}} \mathcal{Z}^{p+1} \to 0 \tag{3.14}$$

are exact sequences of sheaves. Exactness at the left and middle is easy: d lands in closed forms because $\mathrm{d}^2 = 0$; the kernel of d on functions is the locally constant functions (so $\mathcal{Z}^0 = \underline{\mathbb{R}}$, as Nash notes); and the kernel of $\mathrm{d} : \mathcal{E}^p \to \mathcal{Z}^{p+1}$ is $\mathcal{Z}^p$ by definition. Surjectivity on stalks is the **Poincaré lemma**: *on a ball, every closed form is exact.*
Here is the explicit primitive for a closed 1-form $\alpha = \alpha_\mu\,\mathrm{d}x^\mu$ on a ball centred at 0:

$$f(x) = \int_0^1 x^\mu\,\alpha_\mu(tx)\,\mathrm{d}t. \tag{3.15}$$

Check:

$$\partial_\nu f(x) \overset{(1)}{=} \int_0^1 \Big(\alpha_\nu(tx) + t\,x^\mu(\partial_\nu\alpha_\mu)(tx)\Big)\mathrm{d}t$$
$$\overset{(2)}{=} \int_0^1 \Big(\alpha_\nu(tx) + t\,x^\mu(\partial_\mu\alpha_\nu)(tx)\Big)\mathrm{d}t$$

$$\overset{(3)}{=} \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t}\Big(t\,\alpha_\nu(tx)\Big)\mathrm{d}t \;=\; \alpha_\nu(x).$$

(1) Differentiate under the integral; ∂_ν hits the explicit x^μ (giving δ^μ_ν) and the argument tx (giving a factor t by the chain rule).
(2) Closedness, $\mathrm{d}\alpha = 0$, means $\partial_\nu\alpha_\mu = \partial_\mu\alpha_\nu$.
(3) $\frac{\mathrm{d}}{\mathrm{d}t}\alpha_\nu(tx) = x^\mu(\partial_\mu\alpha_\nu)(tx)$, so the integrand is a total t-derivative; evaluate between 0 and 1.

The same trick for a divergence-free magnetic field $\vec{B}$ in $\mathbb{R}^3$ gives the familiar formula $\vec{A}(\vec{x}) = \int_0^1 t\,\vec{B}(t\vec{x})\times\vec{x}\,\mathrm{d}t$ with $\nabla\times\vec{A} = \vec{B}$: the Poincaré lemma for 2-forms. The formula needs every segment from 0 to x to stay in the domain (a ball, or any star-shaped set). On a domain with a hole it can fail, and that is where topology enters.

There is also a smooth version of the exponential sequence. Let $\mathcal{E}_{\mathbb{C}}$ be smooth complex functions and $\mathcal{E}^*$ smooth nowhere-zero complex functions. Then

$$0 \longrightarrow \underline{\mathbb{Z}} \longrightarrow \mathcal{E}_{\mathbb{C}} \xrightarrow{\exp} \mathcal{E}^* \longrightarrow 0, \qquad \exp(f) = e^{2\pi i f}, \tag{3.16}$$

is exact on any manifold, by the same argument (a smooth nonvanishing function on a ball has a smooth logarithm, obtained by integrating $\mathrm{d}g/g$ along radial segments as in (3.15)).

3.2.4 Global sections: exact on the left, not on the right

Now take a short exact sequence of sheaves

$$0 \longrightarrow \mathcal{E} \xrightarrow{e} \mathcal{F} \xrightarrow{f} \mathcal{G} \longrightarrow 0 \tag{3.17}$$

(Nash's (3.9)) and look at global sections, i.e. at the groups of sections over all of M:

$$0 \longrightarrow \mathcal{E}(M) \xrightarrow{e_M} \mathcal{F}(M) \xrightarrow{f_M} \mathcal{G}(M) \longrightarrow 0. \tag{3.18}$$

This is Nash's (3.10). He says "it is easy to check" that it can only fail to be exact at the right-hand end. Here is the check.

e_M is injective. If $e_M(s) = 0$, then every germ $e_x(s_x) = 0$, so $s_x = 0$ for every x (stalks are exact at $\mathcal{E}_x$). A section all of whose germs vanish is zero, by the locality axiom (S1).

Exact at $\mathcal{F}(M)$. Clearly $f_M \circ e_M = 0$ (it is zero on every stalk, hence zero by (S1)). Conversely let $t \in \mathcal{F}(M)$ with $f_M(t) = 0$. By stalk exactness, each x has a neighbourhood U_x and a section $s^{(x)} \in \mathcal{E}(U_x)$ with $e(s^{(x)}) = t|_{U_x}$. On an overlap $U_x \cap U_y$,

$$e\big(s^{(x)} - s^{(y)}\big) \overset{(1)}{=} t|_{U_x\cap U_y} - t|_{U_x\cap U_y} = 0 \quad \overset{(2)}{\Longrightarrow} \quad s^{(x)} = s^{(y)} \text{ on } U_x \cap U_y,$$

(1) Both local lifts map to the same global t.
(2) e is injective (on stalks, hence on sections over any open set, by the argument just given).

So the local lifts are compatible and glue, by (S2), to $s \in \mathcal{E}(M)$ with $e_M(s) = t$.

f_M need not be surjective. Every $g \in \mathcal{G}(M)$ can be lifted locally, $g|_{U_i} = f(t_i)$. The local lifts t_i are ambiguous up to adding something from $\mathcal{E}$, and they may not agree on overlaps. If we cannot choose them to agree, there is no global lift. Two examples show that this really happens.

Worked example 3.11: Global surjectivity fails: two classic cases

(a) The exponential sequence on $M = \mathbb{C}^*$. The function $g(z) = z$ is a global section of $\mathcal{O}^*(\mathbb{C}^*)$. Suppose $z = e^{2\pi i f(z)}$ for a holomorphic f on $\mathbb{C}^*$. Differentiating, $1 = 2\pi i f'(z)\, z$, so $f'(z) = 1/(2\pi i z)$. Integrate round the unit circle:

$$0 \overset{(1)}{=} \oint_{|z|=1} f'(z)\,\mathrm{d}z \overset{(2)}{=} \frac{1}{2\pi i}\oint_{|z|=1}\frac{\mathrm{d}z}{z} \overset{(3)}{=} \frac{1}{2\pi i}\int_0^{2\pi}\frac{i e^{i\phi}\,\mathrm{d}\phi}{e^{i\phi}} = 1,$$

(1) The integral of the derivative of a single-valued function round a closed loop is zero (fundamental theorem of calculus).
(2) Substitute $f' = 1/(2\pi i z)$.
(3) Parametrise $z = e^{i\phi}$, $\mathrm{d}z = i e^{i\phi}\mathrm{d}\phi$.

Contradiction: $0 = 1$. So $\exp_M : \mathcal{O}(\mathbb{C}^*) \to \mathcal{O}^*(\mathbb{C}^*)$ is not surjective. The obstruction is the winding number of g round the origin (Figure 3.9).
(b) The Poincaré sequence on $M = S^1$. The 1-form $\mathrm{d}\theta$ is a global section of $\mathcal{Z}^1(S^1)$ (it is well defined although θ is not: $\mathrm{d}\theta = (x\,\mathrm{d}y - y\,\mathrm{d}x)/(x^2+y^2)$). If $\mathrm{d}\theta = \mathrm{d}F$ for a smooth function F on S^1, then $2\pi = \oint \mathrm{d}\theta = \oint \mathrm{d}F = 0$, again a contradiction. So $\mathrm{d} : \mathcal{E}^0(S^1) \to \mathcal{Z}^1(S^1)$ is not surjective, although it is surjective on every stalk (Poincaré lemma on small arcs).

Both obstructions are integers, or real numbers, attached to loops. Sheaf cohomology will package them: in (a) the obstruction lives in $H^1(\mathbb{C}^*;\underline{\mathbb{Z}}) \cong \mathbb{Z}$, in (b) in $H^1(S^1;\underline{\mathbb{R}}) \cong \mathbb{R}$.

Physical picture: Gauge theory reading

In (b), think of $\mathrm{d}\theta$ as a flat $U(1)$ gauge field a on a circle (an Aharonov–Bohm flux line piercing the plane). Locally it is pure gauge, $a = \mathrm{d}\lambda$, because every closed form is locally exact. Globally it is not, and the obstruction $\oint a$ is the Aharonov–Bohm phase. Sheaf cohomology is the systematic way of computing all such "locally pure gauge but globally not" configurations.

3.2.5 Soft sheaves, fine sheaves and partitions of unity

When is (3.18) exact after all? Nash gives a sufficient condition: the left-hand sheaf $\mathcal{E}$ should be able to extend sections from small sets to all of M. Sheaves of smooth objects can do this, because smooth functions can be cut off with bump functions. Holomorphic functions and locally constant functions cannot.

Bump functions

The basic tool is a smooth function that is 1 on one set and 0 outside a slightly larger one. Start from

$$\phi(t) = \begin{cases} e^{-1/t}, & t > 0, \\ 0, & t \le 0, \end{cases} \tag{3.19}$$

which is smooth by the argument of Example 3.4(c). The smooth step

$$h(t) = \frac{\phi(t)}{\phi(t) + \phi(1-t)} \tag{3.20}$$

is 0 for $t \le 0$, 1 for $t \ge 1$, smooth everywhere (the denominator never vanishes, since at least one of $t, 1-t$ is positive). Products such as $h\big(\frac{x-a}{\varepsilon}\big)\,h\big(\frac{b-x}{\varepsilon}\big)$ give bumps equal to 1 on $[a+\varepsilon, b-\varepsilon]$

and 0 outside (a,b). In $\mathbb{R}^n$ the function $h\big(2-|x|\big)$ is 1 on the ball of radius 1 and 0 outside the ball of radius 2.

Working definition 3.7 (partition of unity)

A **partition of unity** subordinate to an open cover $\{U_i\}$ of M is a family of smooth functions $\rho_i : M \to [0,1]$ with

(i) $\operatorname{supp}\rho_i \subset U_i$ [the closure of the set where $\rho_i \neq 0$ lies inside U_i];

(ii) locally finite: each point has a neighbourhood meeting only finitely many $\operatorname{supp}\rho_i$;

(iii) $\sum_i \rho_i = 1$ everywhere.

Construction: pick bumps $\psi_i \geq 0$ supported in U_i whose positivity sets still cover M, and set $\rho_i = \psi_i / \sum_j \psi_j$. Every paracompact [every open cover can be refined to a locally finite one. Every manifold we meet is paracompact] manifold (Section 3.3.5) admits partitions of unity subordinate to any open cover.

Worked example 3.12: A partition of unity on the circle

Use the two-arc cover of Example 3.5: $U_1 = (-\varepsilon, \pi+\varepsilon)$, $U_2 = (\pi-\varepsilon, 2\pi+\varepsilon)$, overlaps A near 0 and B near π. Define

$$\rho_1(\theta) = h\!\left(\frac{\theta+\varepsilon}{2\varepsilon}\right) h\!\left(\frac{\pi+\varepsilon-\theta}{2\varepsilon}\right) \quad \left(-\tfrac{\pi}{2} \leq \theta \leq \tfrac{3\pi}{2}\right), \qquad \rho_2 = 1-\rho_1.$$

Then ρ_1 rises from 0 to 1 across A, equals 1 on the middle of U_1, and falls back to 0 across B; so $\operatorname{supp}\rho_1 \subset U_1$. Its complement ρ_2 is 1 on the middle of U_2 and vanishes on the middle of U_1; so $\operatorname{supp}\rho_2 \subset U_2$ (Figure 3.10). They add to 1 by construction.

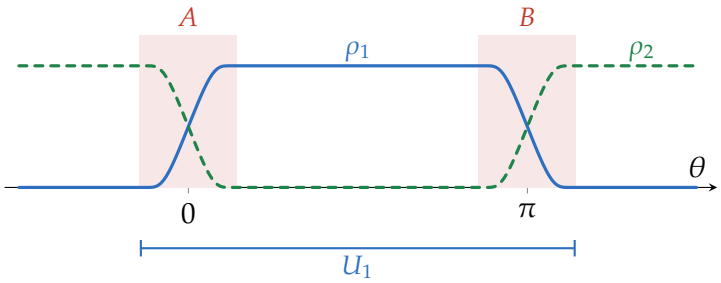


Figure 3.10: A smooth partition of unity on S^1 (the angle runs from $-\pi/2$ to $3\pi/2$, one full turn). ρ_1 (solid) lives inside U_1, $\rho_2 = 1-\rho_1$ (dashed) lives inside U_2, and all the switching happens inside the overlaps A and B (shaded). No holomorphic or locally constant function can do this.

Soft sheaves

Working definition 3.8 (soft sheaf)

A sheaf $\mathcal{F}$ is **soft** if every section defined on (a neighbourhood of) a *closed* set $K \subset M$ agrees on K with some global section [a section over all of M]. Nash's (3.11) expresses this as surjectivity of restriction $r_U^M : \mathcal{F}(M) \to \mathcal{F}(U)$. *Operational test:* can you multiply a local section by a cut-off function and extend by zero?

Watch out: Open versus closed sets in (3.11)

Taken literally for *open* U, Nash's condition (3.11) fails even for smooth functions: $1/x$ on $U = (0, \infty)$ is smooth but is not the restriction of any smooth function on $\mathbb{R}$. The correct notion of softness uses closed sets (or, equivalently for manifolds, "given s on an open U and a closed $K \subset U$, there is a global section equal to s near K"). This is Wells' definition, and it is what the bump-function argument proves: multiply s by a bump ψ equal to 1 near K and supported in U, and extend ψs by zero.

With this definition $\mathcal{E}$ and every $\mathcal{E}^p$ are soft (by the bump-function argument), and so is $\mathcal{E}^p(E)$ for any bundle E. The sheaves $\mathcal{O}$, $\underline{\mathbb{Z}}$, $\underline{\mathbb{R}}$ are not soft: a holomorphic function on a small disc in $\mathbb{CP}^1$ is not the restriction of a global holomorphic function unless it is constant; a locally constant function equal to 0 on one small disc and 1 on another cannot be extended to a connected M.

Fine sheaves

In practice we use a slightly stronger and more checkable property.

Working definition 3.9 (fine sheaf)

A sheaf $\mathcal{F}$ is **fine** if it admits partitions of unity: for every locally finite open cover $\{U_i\}$ there are sheaf morphisms [maps acting on sections over every open set, compatible with restriction] $\eta_i : \mathcal{F} \to \mathcal{F}$ with $\eta_i(s) = 0$ outside a closed subset of U_i and $\sum_i \eta_i = \mathrm{id}$. *Operational test:* if $\mathcal{F}$ is a sheaf of modules over smooth functions (you can multiply its sections by smooth functions), it is fine: take $\eta_i(s) = \rho_i s$ for a partition of unity $\{\rho_i\}$.

Fine (hence soft)	Not fine	Why not
$\mathcal{E}$, $\mathcal{E}_{\mathbb{C}}$, $\mathcal{E}^p$	$\underline{\mathbb{Z}}$, $\underline{\mathbb{R}}$	ρf is not locally constant
$\mathcal{E}^p(E)$, $\mathcal{E}^{p,q}$	$\mathcal{O}$, Ω^p	ρf is not holomorphic
$\Gamma(\cdot, E)$ smooth sections	$\mathcal{O}^*$, $\mathcal{E}^*$	multiplicative; ρg may vanish

Nash states, without proof (he refers to Wells), that fine sheaves are soft and that *soft sheaves have no higher cohomology*: $H^q(M; \mathcal{F}) = 0$ for $q > 0$. The proof for fine sheaves is short and important enough that we give it now in the simplest case, and in general in Section 3.3.7.

Key idea: Fine on the left makes global sections exact

If in $0 \to \mathcal{E} \to \mathcal{F} \to \mathcal{G} \to 0$ the sheaf $\mathcal{E}$ is fine, then $\mathcal{F}(M) \to \mathcal{G}(M)$ is surjective, so (3.18) is exact.

Why. Take $g \in \mathcal{G}(M)$. Stalk surjectivity gives a cover $\{U_i\}$ (which we may take locally finite) and local lifts $t_i \in \mathcal{F}(U_i)$ with $f(t_i) = g|_{U_i}$. On overlaps the lifts differ by something that f kills, hence by the image of a section of $\mathcal{E}$:

$$t_j - t_i = e(c_{ij}), \qquad c_{ij} \in \mathcal{E}(U_{ij}). \tag{3.21}$$

On a triple overlap, $(t_k - t_j) - (t_k - t_i) + (t_j - t_i) = 0$, and e is injective, so

$$c_{jk} - c_{ik} + c_{ij} = 0 \quad \text{on } U_{ijk}. \tag{3.22}$$

We want to correct each t_i by $e(h_i)$ so that the corrected lifts agree. Use the partition of unity $\{\rho_k\}$ and set

$$h_i = \sum_k \rho_k \, c_{ki} \ \in \mathcal{E}(U_i) \tag{3.23}$$

(each term is defined on $U_k \cap U_i$ and vanishes near the edge of U_k, so it extends by zero to all of U_i). Then on U_{ij}

$$h_j - h_i \overset{(1)}{=} \sum_k \rho_k (c_{kj} - c_{ki}) \overset{(2)}{=} \sum_k \rho_k \, c_{ij} \overset{(3)}{=} c_{ij}.$$

(1) Definition (3.23), subtracted.
(2) The cocycle identity (3.22) for the triple (k, i, j) reads $c_{ij} - c_{kj} + c_{ki} = 0$, i.e. $c_{kj} - c_{ki} = c_{ij}$.
(3) $\sum_k \rho_k = 1$.

Now put $t_i' = t_i - e(h_i)$. Then $t_j' - t_i' = e(c_{ij}) - e(h_j - h_i) = 0$, so the t_i' glue to a global $t \in \mathcal{F}(M)$, and $f(t) = g$ because $f \circ e = 0$. The whole argument is: *local lifts disagree by a cocycle; a partition of unity writes the cocycle as a difference; correct the lifts.* The next section turns "cocycle" and "difference" into the definition of Čech cohomology.

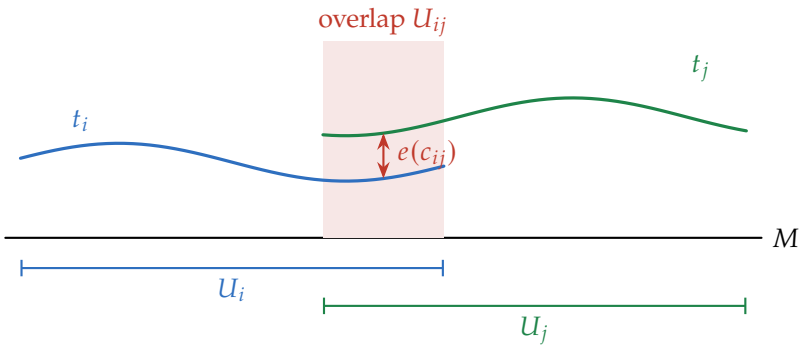


Figure 3.11: Local lifts t_i and t_j of the same global section g disagree on the overlap by $e(c_{ij})$, a section of the kernel sheaf. If the c_{ij} can be written as differences $h_j - h_i$ (which a partition of unity achieves when $\mathcal{E}$ is fine), shifting each lift by $e(h_i)$ removes the disagreement and the lifts glue.

Reading Nash (pp. 59–61, (3.4)–(3.11))

Nash's sequence $\mathcal{E} \xrightarrow{e} \mathcal{F} \xrightarrow{f} \mathcal{G}$ in (3.4) uses $\mathcal{E}$ as a generic sheaf, not the sheaf of smooth functions; beware the clash of notation. His (3.5) is our commuting square (3.10); (3.6)–(3.7) define exactness by stalks; (3.8) is the exponential sequence, with $\exp f = e^{2\pi i f}$ so that the kernel is exactly $\mathbb{Z}$. His claim that (3.10) "can only fail to be exact at its right-hand end" is Section 3.2.4; the extension property (3.11) should be read with closed sets (see the pitfall). He also asserts that $H^q(M; \mathcal{F}) = 0$ for $q > 0$ when $\mathcal{F}$ is soft, that $\mathcal{E}$ and $\mathcal{E}^p$ are soft, and that fine sheaves (those admitting a locally finite partition of unity) are a subclass of soft sheaves. We have explained the mechanism above; Section 3.3.7 gives it in general.

Recap

- A sheaf morphism is a family of homomorphisms commuting with restriction, i.e. a *local* operation (d, $\bar{\partial}$, exp, inclusions).
- A sequence of sheaves is exact if it is exact on every stalk: "locally exact". Key examples: the exponential sequence $0 \to \mathbb{Z} \to \mathcal{O} \to \mathcal{O}^* \to 0$ and the Poincaré-lemma sequences $0 \to \mathcal{Z}^p \to \mathcal{E}^p \to \mathcal{Z}^{p+1} \to 0$.
- On global sections, a short exact sequence of sheaves gives a sequence that is exact except possibly at the right end. Failures: z on $\mathbb{C}^*$ has no global logarithm; $\mathrm{d}\theta$ on S^1 has no global primitive.

- Smooth objects can be cut off with bump functions: the sheaves $\mathcal{E}$, $\mathcal{E}^p$, $\mathcal{E}^p(E)$ are fine (and soft). If the left sheaf is fine, global sections are exact. The mechanism is a partition-of-unity correction of a cocycle.

3.3 Čech cohomology

At the end of the last section we met a very specific piece of bookkeeping. Local lifts t_i of a global section disagreed on overlaps by amounts c_{ij}; these amounts satisfied $c_{jk} - c_{ik} + c_{ij} = 0$ on triple overlaps; and if we could write $c_{ij} = h_j - h_i$, the disagreement could be removed. Exactly the same pattern appears for vector bundles: transition functions g_{ij} obey $g_{ij}g_{jk} = g_{ik}$, and changing trivialisations replaces g_{ij} by $h_i g_{ij} h_j^{-1}$.

So there is a recurring question:

> *Given data c_{ij} on the overlaps of a cover that is consistent on triple overlaps, is it just the "difference" of data h_i on the open sets themselves?*

Čech cohomology is the systematic answer, in every degree at once. It is the construction Nash uses to define sheaf cohomology (pp. 61–62, equations (3.12)–(3.17)). Its great virtue for us is that it is *computable*: for a sensible cover, everything reduces to finite matrices.

3.3.1 Covers, overlaps and the nerve

Fix an open cover $\mathcal{U} = \{U_i\}_{i \in I}$ of M (Nash writes $O = \{O_\alpha\}$). We label the index set by integers and use the shorthand

$$U_{ij} = U_i \cap U_j, \qquad U_{ijk} = U_i \cap U_j \cap U_k, \qquad U_{i_0 i_1 \cdots i_q} = U_{i_0} \cap U_{i_1} \cap \cdots \cap U_{i_q}.$$

A list $(i_0, \dots, i_q)$ with $U_{i_0 \cdots i_q} \neq \varnothing$ is what Nash calls a Δ: "a collection $(O_1, \dots, O_{q+1})$ of $q+1$ of the O_α with non-empty intersection".

A useful picture is the **nerve** of the cover. Draw one vertex for each open set, an edge between i and j whenever $U_{ij} \neq \varnothing$, a filled triangle on i, j, k whenever $U_{ijk} \neq \varnothing$, a solid tetrahedron for each non-empty fourfold overlap, and so on. The result is a combinatorial skeleton of M (Figure 3.12). For a good cover (defined in Section 3.3.5) the nerve has the same shape as M, up to deformation, and Čech cohomology with constant coefficients is just the cohomology of this skeleton.

3.3.2 Cochains and the coboundary

Working definition 3.10 (Čech cochains)

A **Čech q-cochain** of the cover $\mathcal{U}$ with coefficients in the sheaf $\mathcal{F}$ is a choice of one section [one allowed local object, for $\underline{\mathbb{R}}$ one real number per connected component]

$$c_{i_0 i_1 \cdots i_q} \in \mathcal{F}\left(U_{i_0 i_1 \cdots i_q}\right)$$

for every $(q+1)$-fold overlap [the intersection $U_{i_0} \cap \cdots \cap U_{i_q}$ of $q+1$ sets of the cover]. We require the choice to be **alternating**: swapping two indices flips the sign ($c_{ji} = -c_{ij}$,

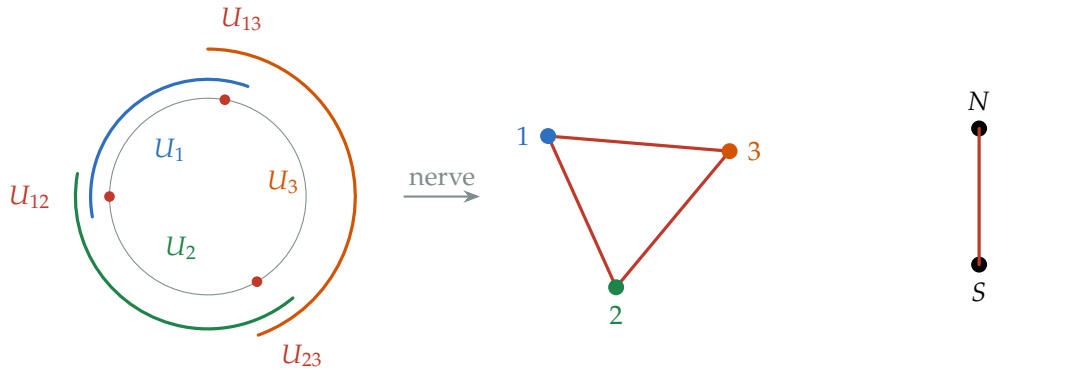


Figure 3.12: The nerve of a cover. Left: three arcs on a circle overlap in pairs (red dots) but never all three at once, so the nerve (middle) is a hollow triangle, which is a circle again. Right: two caps covering S^2 overlap in an annulus; the nerve is a single edge, which has lost the "hole" of the sphere. The first cover is good; the second is not, and it will give the wrong H^2.

$c_{jik} = -c_{ijk}, \dots$), and a cochain with a repeated index is zero. The q-cochains form a group

$$C^q(\mathcal{U};\mathcal{F}) = \prod_{i_0<i_1<\cdots<i_q} \mathcal{F}\left(U_{i_0\cdots i_q}\right). \tag{3.24}$$

This is Nash's (3.12), $C^q(O,\mathcal{F}) = \mathcal{F}(\Delta)$. *Operational recipe:* list the non-empty overlaps with increasing indices; a q-cochain is a column vector with one entry (a section) per $(q+1)$-fold overlap.

In low degrees:

- a 0-cochain is a section f_i on each open set U_i (no compatibility required);
- a 1-cochain is a section c_{ij} on each double overlap U_{ij};
- a 2-cochain is a section n_{ijk} on each triple overlap U_{ijk}.

If an overlap $U_{i_0\cdots i_q}$ has several connected components and $\mathcal{F} = \underline{\mathbb{R}}$, the "section" is one real number per component. This will matter in almost every example.

Working definition 3.11 (Čech coboundary)

The **coboundary** $\delta : C^q(\mathcal{U};\mathcal{F}) \to C^{q+1}(\mathcal{U};\mathcal{F})$ is

$$(\delta c)_{i_0 i_1 \cdots i_{q+1}} = \sum_{k=0}^{q+1} (-1)^k \, c_{i_0\cdots \widehat{i_k} \cdots i_{q+1}}\Big|_{U_{i_0\cdots i_{q+1}}}, \tag{3.25}$$

where the hat means "omit this index". This is Nash's (3.13), with his restriction map $r^{\Delta}_{\Delta_i}$ written as "|" [cutting a section down to the smaller overlap]. *Operational recipe:* to evaluate δc on an overlap of $q+2$ sets, drop each index in turn, restrict the corresponding entry of c to the smaller set, and add with alternating signs.

Written out in the three degrees we will use constantly (Nash calls δ simply d; we reserve d

for the exterior derivative):

$$(\delta f)_{ij} = f_j - f_i, \tag{3.26}$$
$$(\delta c)_{ijk} = c_{jk} - c_{ik} + c_{ij}, \tag{3.27}$$
$$(\delta n)_{ijkl} = n_{jkl} - n_{ikl} + n_{ijl} - n_{ijk}. \tag{3.28}$$

(Each right-hand side is restricted to the overlap named on the left.) These are exactly the combinations we met in (3.21) and (3.22).

Key idea: $\delta^2 = 0$

Applying the coboundary twice gives zero: $\delta \circ \delta = 0$.

Check it in the lowest degree, for a 0-cochain f:

$$\begin{aligned}(\delta\delta f)_{ijk} &\overset{(1)}{=} (\delta f)_{jk} - (\delta f)_{ik} + (\delta f)_{ij} \\ &\overset{(2)}{=} (f_k - f_j) - (f_k - f_i) + (f_j - f_i) \\ &= 0.\end{aligned}$$

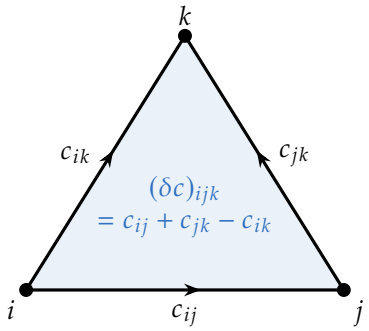
(1) Formula (3.27) applied to the 1-cochain $c = \delta f$.
(2) Formula (3.26) for each of the three entries.

And one degree up, for a 1-cochain c:

$$\begin{aligned}(\delta\delta c)_{ijkl} &\overset{(1)}{=} (\delta c)_{jkl} - (\delta c)_{ikl} + (\delta c)_{ijl} - (\delta c)_{ijk} \\ &\overset{(2)}{=} (c_{kl} - c_{jl} + c_{jk}) - (c_{kl} - c_{il} + c_{ik}) + (c_{jl} - c_{il} + c_{ij}) - (c_{jk} - c_{ik} + c_{ij}) \\ &= 0.\end{aligned}$$

(1) Formula (3.28) applied to $n = \delta c$.
(2) Formula (3.27) for each entry; now every c appears exactly twice with opposite signs.

The general reason is the same bookkeeping: in $\delta\delta c$ each term has two indices removed, say i_a and i_b with $a < b$. It arises once by removing i_b first and then i_a (sign $(-1)^b(-1)^a$), and once by removing i_a first, after which i_b has moved to position $b-1$ (sign $(-1)^a(-1)^{b-1}$). The two cancel. This is the same sign trick that makes $\partial^2 = 0$ for the boundary of a simplex, and δ is its dual; Nash simply says "it is straightforward to verify that $d^2 = 0$".



Walk $i \to j \to k$ and then back from k to i, adding the value on each edge, with a minus sign when an edge is traversed against its orientation. A *cocycle* ($\delta c = 0$) is data whose total round every triangle is zero, like a curl-free field. A *coboundary* ($c = \delta f$) is a difference of vertex values, like a gradient.

Figure 3.13: The coboundary of a 1-cochain, read on the nerve. 1-cochains live on edges, δc lives on triangles and is the signed sum round the triangle. This is a discrete "curl"; δ on 0-cochains is a discrete "gradient", and $\delta^2 = 0$ is "curl grad = 0".

Physical picture: Discrete vector calculus

On the nerve, δ on 0-cochains is a discrete gradient (difference of values at the ends of an edge), on 1-cochains a discrete curl (circulation round a triangle), on 2-cochains a discrete divergence (flux out of a tetrahedron). The Čech complex is lattice vector calculus on the skeleton of the cover, with the values replaced by sections of a sheaf.

3.3.3 The cohomology of a cover

Because $\delta^2 = 0$ we have a **cochain complex** (Nash's (3.14))

$$0 \longrightarrow C^0(\mathcal{U};\mathcal{F}) \xrightarrow{\delta} C^1(\mathcal{U};\mathcal{F}) \xrightarrow{\delta} C^2(\mathcal{U};\mathcal{F}) \xrightarrow{\delta} \cdots \tag{3.29}$$

in which the image of each map lies inside the kernel of the next.

Working definition 3.12 (Čech cohomology of a cover)

A q-cochain c is a **cocycle** if $\delta c = 0$ [for $q = 1$: $c_{jk} - c_{ik} + c_{ij} = 0$ on every triple overlap] and a **coboundary** if $c = \delta b$ for some $(q-1)$-cochain b [for $q = 1$: $c_{ij} = b_j - b_i$ on every double overlap]. Every coboundary is a cocycle. The **Čech cohomology of the cover** is

$$H^q(\mathcal{U};\mathcal{F}) = \frac{\operatorname{Ker}\left(\delta : C^q \to C^{q+1}\right)}{\operatorname{Im}\left(\delta : C^{q-1} \to C^q\right)} = \frac{\{q\text{-cocycles}\}}{\{q\text{-coboundaries}\}}, \tag{3.30}$$

Nash's (3.15). *Operational recipe:* write the matrices of δ, compute kernels and images, take the quotient. For finite covers and $\mathcal{F} = \underline{\mathbb{R}}$ this is linear algebra: $\dim H^q = \dim C^q - \operatorname{rank}\delta_q - \operatorname{rank}\delta_{q-1}$.

Key idea: Nash (3.17): H^0 is global sections

For any sheaf $\mathcal{F}$ and any cover, $H^0(\mathcal{U};\mathcal{F}) = \mathcal{F}(M) = \Gamma(M,\mathcal{F})$.

There are no (-1)-cochains, so H^0 is just the 0-cocycles:

$$\begin{aligned} H^0(\mathcal{U};\mathcal{F}) &\overset{(1)}{=} \{(f_i) \;:\; (\delta f)_{ij} = 0 \;\forall i,j\} \\ &\overset{(2)}{=} \{(f_i) \;:\; f_i|_{U_{ij}} = f_j|_{U_{ij}} \;\forall i,j\} \\ &\overset{(3)}{=} \mathcal{F}(M). \end{aligned}$$

(1) Definition (3.30) with $\operatorname{Im}\delta_{-1} = 0$.
(2) Formula (3.26).
(3) The sheaf axioms: compatible local sections glue to a unique global section (S2, S1), and every global section restricts to a compatible family.

So H^0 is the familiar thing (global solutions, global wavefunctions), and the gluing axiom is literally the statement "0-cocycles are global sections". The higher H^q are new.

Physical picture: What H^1 means

A 1-cocycle is a set of "transition data" c_{ij} on overlaps that is consistent on triple overlaps. It is a coboundary if it can be undone by a change of local data, $c_{ij} = f_j - f_i$. So H^1 counts *twisted gluings modulo re-gaugings*. For the multiplicative sheaf of $U(1)$-valued functions

the c_{ij} are gauge transformations between patches, and H^1 is the set of $U(1)$ bundles (Section 3.5). In the fine-sheaf argument of Section 3.2.5, H^1 of the kernel sheaf is precisely the obstruction to lifting a global section.

Watch out: Multiplicative sheaves

For $\mathcal{O}^*$, $\mathcal{E}^*$ or $U(1)$-valued functions the group operation is multiplication. Then "$-$" becomes "$\div$", and the formulas read $(\delta h)_{ij} = h_j h_i^{-1}$ and $(\delta g)_{ijk} = g_{jk} g_{ik}^{-1} g_{ij}$. The alternating condition becomes $g_{ji} = g_{ij}^{-1}$. A 1-cocycle is therefore $g_{ij} g_{jk} = g_{ik}$: the cocycle condition (3.6) for transition functions of a line bundle.

3.3.4 First computations: the circle

Worked example 3.13: S^1 with three arcs, coefficients $\underline{\mathbb{R}}$

Take the cover of Figure 3.12: three open arcs U_1, U_2, U_3 with connected pairwise overlaps and $U_{123} = \varnothing$. Every overlap is connected, so each section of $\underline{\mathbb{R}}$ is one real number:

$$C^0 = \mathbb{R}^3 \ni (a_1, a_2, a_3), \qquad C^1 = \mathbb{R}^3 \ni (c_{12}, c_{13}, c_{23}), \qquad C^2 = 0.$$

By (3.26), $\delta(a) = (a_2 - a_1,\ a_3 - a_1,\ a_3 - a_2)$, i.e.

$$\delta_0 = \begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & -1 & 1 \end{pmatrix} \quad (\text{rows } 12, 13, 23;\ \text{columns } a_1, a_2, a_3). \tag{3.31}$$

The kernel of δ_0 is the constants (a, a, a), so

$$H^0(\mathcal{U}; \underline{\mathbb{R}}) \overset{(1)}{=} \operatorname{Ker} \delta_0 \cong \mathbb{R},$$
$$H^1(\mathcal{U}; \underline{\mathbb{R}}) \overset{(2)}{=} C^1 / \operatorname{Im} \delta_0 \overset{(3)}{=} \mathbb{R}^3 / \mathbb{R}^2 \cong \mathbb{R},$$

(1) H^0 is the kernel of δ_0: a locally constant function on a connected circle is constant.
(2) Every 1-cochain is a cocycle because $C^2 = 0$.
(3) $\operatorname{rank} \delta_0 = 3 - \dim \operatorname{Ker} \delta_0 = 2$.

An explicit isomorphism $H^1 \to \mathbb{R}$: the image of δ_0 is the plane $c_{12} - c_{13} + c_{23} = 0$ (check: $(a_2 - a_1) - (a_3 - a_1) + (a_3 - a_2) = 0$). So

$$[c] \longmapsto c_{12} + c_{23} - c_{13} \tag{3.32}$$

is well defined on classes and is an isomorphism. It is the signed sum round the triangle $1 \to 2 \to 3 \to 1$: a discrete line integral round the circle. The class of $c = (0, 0, -1)$, for instance, is the generator "one unit of circulation".
With $\underline{\mathbb{Z}}$ coefficients the same matrix gives $H^0 = \mathbb{Z}$ and $H^1 = \mathbb{Z}^3 / \operatorname{Im} \delta_0 \cong \mathbb{Z}$, again via (3.32), which is now onto $\mathbb{Z}$. This matches the topologist's $H^1(S^1; \mathbb{Z}) = \mathbb{Z}$ (Chapter 1).

Worked example 3.14: S^1 with two arcs: disconnected overlaps

Now use the two-arc cover U_1, U_2 of Example 3.5. The overlap $U_{12} = A \sqcup B$ has two

components, so a section of $\underline{\mathbb{R}}$ over U_{12} is a pair of numbers:

$$C^0 = \mathbb{R}^2 \ni (a_1, a_2), \qquad C^1 = \underline{\mathbb{R}}(A \sqcup B) = \mathbb{R}^2 \ni (c_A, c_B), \qquad C^2 = 0.$$

The coboundary restricts $a_2 - a_1$ to each component:

$$\delta(a_1, a_2) \overset{(1)}{=} \big((a_2 - a_1)|_A,\ (a_2 - a_1)|_B\big) = (a_2 - a_1,\ a_2 - a_1), \qquad \delta_0 = \begin{pmatrix} -1 & 1 \\ -1 & 1 \end{pmatrix}.$$

(1) Formula (3.26); a constant restricted to each component is the same constant.

So $\operatorname{Ker}\delta_0 = \{(a, a)\} \cong \mathbb{R}$ and $\operatorname{Im}\delta_0 = \{(c, c)\}$, giving

$$H^0 = \mathbb{R}, \qquad H^1 = \mathbb{R}^2/\{(c, c)\} \cong \mathbb{R}, \quad [c] \mapsto c_A - c_B.$$

The same answer as with three arcs, although the nerve is now a single edge. The information about the hole is hiding in the fact that the edge carries *two* numbers: the overlap is disconnected. A disconnected overlap is harmless here, because each component is still contractible.

Worked example 3.15: The Möbius bundle in Čech language

The Möbius bundle of Example 3.5 has transition function $g_{12} = (-1 \text{ on } A,\ +1 \text{ on } B)$. Two related Čech computations tell it apart from the trivial bundle.

(a) The transition function as a class in $H^1(S^1; \underline{\mathbb{Z}_2})$. Take the multiplicative sheaf $\underline{\mathbb{Z}_2}$ of locally constant functions with values in $\{\pm 1\}$. With the two-arc cover a 1-cochain is a pair (g_A, g_B) of signs, and every cochain is a cocycle ($C^2 = 0$). The coboundaries are

$$(\delta\epsilon)_{12} \overset{(1)}{=} \epsilon_2\epsilon_1^{-1}\big|_{A\sqcup B} = (\epsilon_2\epsilon_1,\ \epsilon_2\epsilon_1),$$

(1) Multiplicative form of (3.26), with constants $\epsilon_1, \epsilon_2 \in \{\pm 1\}$ on the connected arcs; $\epsilon^{-1} = \epsilon$.

So a coboundary has *equal* signs on A and B. The trivial bundle $(+1, +1)$ is a coboundary; the Möbius bundle $(-1, +1)$ is not. Hence $H^1(\mathcal{U}; \underline{\mathbb{Z}_2}) = \{\pm 1\}^2/\{(\epsilon, \epsilon)\} \cong \mathbb{Z}_2$, the class is $g_A g_B$, and the Möbius bundle is the nontrivial element. No change of trivialisation ($s_i \mapsto \epsilon_i s_i$) can untwist it. Real line bundles over M are classified by $H^1(M; \underline{\mathbb{Z}_2})$ in general: this class is the first **Stiefel–Whitney class**.

(b) Cohomology of the twisted constant sheaf. Let $\underline{\mathbb{R}}^{\mathrm{tw}}$ be the sheaf of locally constant sections of the Möbius bundle. Over each arc a section is a constant a_i in the i-th trivialisation. To compare on the overlap we express everything in trivialisation 1, using $s_1 = g_{12}s_2$:

$$(\delta a)|_A \overset{(1)}{=} a_1 - g_{12}|_A\, a_2 = a_1 + a_2, \qquad (\delta a)|_B = a_1 - a_2, \qquad \delta_0 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

(1) The difference of the two local sections on A, written in the first trivialisation. (The overall sign convention for δ does not affect kernels and images.)

$\det\delta_0 = -2 \neq 0$, so δ_0 is invertible: $H^0(S^1; \underline{\mathbb{R}}^{\mathrm{tw}}) = 0$ (no global flat section: it would have to change sign) and $H^1(S^1; \underline{\mathbb{R}}^{\mathrm{tw}}) = 0$. Compare the untwisted answer $H^0 = H^1 = \mathbb{R}$. Two sheaves with isomorphic stalks everywhere, and completely different cohomology.

Physical picture: Aharonov–Bohm in a line

Replace $\{\pm 1\}$ by $U(1)$ in Example 3.15(a): a flat $U(1)$ bundle on a circle is a pair of phases (g_A, g_B) modulo $(e^{i\alpha}, e^{i\alpha})$, i.e. one phase $g_A g_B^{-1}$. This is the Aharonov–Bohm phase $e^{ie\oint A}$ of a charged particle encircling a flux tube. A wavefunction in this background is a section of the flat bundle, and the twisted-coefficient computation (b) says a flat section exists only if the phase is 1.

3.3.5 Refinements, the limit over covers, and good covers

Nash writes $H^q(O;\mathcal{F})$, with the cover in the notation, "to draw attention to the fact that our construction could depend on the covering used". It really can.

Worked example 3.16: S^2 with two caps: the wrong answer

Cover S^2 by a northern cap U_N and a southern cap U_S, each a little bigger than a hemisphere; U_{NS} is an open annulus round the equator (connected). With $\underline{\mathbb{R}}$ coefficients

$$C^0 = \mathbb{R}^2, \qquad C^1 = \mathbb{R}, \qquad C^2 = 0, \qquad \delta(a_N, a_S) = a_S - a_N.$$

The map $\delta_0 = (-1 \;\; 1)$ is onto, so $H^0 = \mathbb{R}$, $H^1 = 0$ and $H^2 = 0$ (because $C^2 = 0$). But the true answer (Chapter 1, Chapter T2) is $H^2(S^2;\mathbb{R}) = \mathbb{R}$: the sphere encloses a volume, and the area form is closed but not exact. The cover is too coarse to see this: a cover with only two sets has no triple overlaps and can never see H^2.

The culprit is that the overlap annulus is not contractible: it has its own H^1, which the cover does not resolve. Nash's remedy is to use finer and finer covers.

Working definition 3.13 (refinement)

A cover $\mathcal{V} = \{V_\beta\}_{\beta\in J}$ is a **refinement** of $\mathcal{U} = \{U_\alpha\}_{\alpha\in I}$ if every V_β lies in some U_α (Nash's footnote 2). Choose such an $\alpha = \tau(\beta)$ for each β [the choice is not unique, but the induced map on cohomology does not depend on it]. Then cochains can be pulled back,

$$(\tau^* c)_{\beta_0\cdots\beta_q} = c_{\tau(\beta_0)\cdots\tau(\beta_q)}\big|_{V_{\beta_0\cdots\beta_q}},$$

which makes sense because $V_{\beta_0\cdots\beta_q} \subset U_{\tau(\beta_0)\cdots\tau(\beta_q)}$. *Operational recipe:* a cochain on the coarse cover gives one on the fine cover by restricting each entry to the smaller overlaps.

The map τ^* commutes with δ (both are built from restrictions), so it maps cocycles to cocycles and coboundaries to coboundaries, and induces $\tau^* : H^q(\mathcal{U};\mathcal{F}) \to H^q(\mathcal{V};\mathcal{F})$. One checks that a different choice of τ changes $\tau^* c$ by a coboundary, so the map on cohomology depends only on the two covers. Nash then defines

$$H^q(M;\mathcal{F}) = \varinjlim_{\mathcal{U}} H^q(\mathcal{U};\mathcal{F}), \tag{3.33}$$

his (3.16). In words: a class in $H^q(M;\mathcal{F})$ is represented by a cocycle on *some* cover, and two such are identified if they become cohomologous on a common refinement. This is the **sheaf cohomology** of M with coefficients in $\mathcal{F}$.

Taking a limit over all covers sounds hopeless for computation. In practice one never does it, because of the following fact.

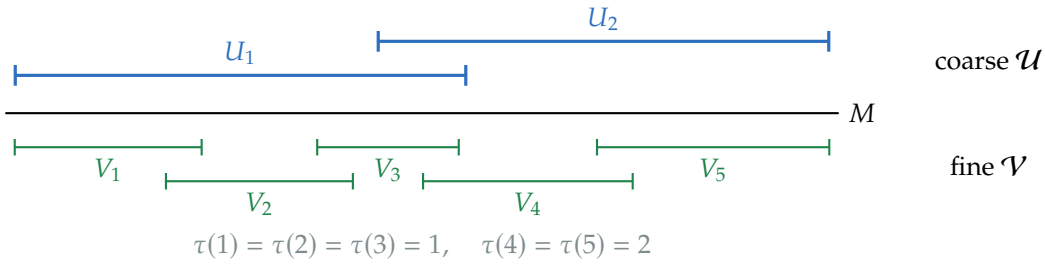


Figure 3.14: A refinement: every green set V_β sits inside some blue set $U_{\tau(\beta)}$. Cochains of the coarse cover restrict to cochains of the fine cover. Sheaf cohomology is what remains when we pass to ever finer covers; for a good cover nothing changes any more.

Key idea: Leray's theorem: good covers compute sheaf cohomology

If the sheaf $\mathcal{F}$ has no higher cohomology on every finite overlap, $H^p(U_{i_0 \cdots i_q}; \mathcal{F}) = 0$ for all $p > 0$, then $H^q(\mathcal{U}; \mathcal{F}) = H^q(M; \mathcal{F})$. Such a cover is called a **Leray cover** for $\mathcal{F}$. For $\underline{\mathbb{R}}$, $\underline{\mathbb{Z}}$ or any locally constant sheaf, a **good cover**, one in which every non-empty finite overlap is contractible (or a disjoint union of contractible pieces), is a Leray cover.

Why it is plausible. Refining a cover changes the cochains only if the finer cover detects something new inside the old overlaps. If every overlap is already "cohomologically trivial" for $\mathcal{F}$, there is nothing new to detect. Every manifold has good covers: take small geodesically convex balls for some Riemannian metric; intersections of convex sets are convex. This is why the three-arc cover of S^1 gave the right answer, and the two-cap cover of S^2 did not (its overlap, an annulus, has $H^1 = \mathbb{R} \neq 0$). For $\mathcal{O}$ on a complex manifold, open sets such as discs, polydiscs, $\mathbb{C}$ and $\mathbb{C}^*$ have no higher cohomology (this follows from the Dolbeault lemma of Section 3.7.3), so covers by such sets are Leray covers for $\mathcal{O}$.

Watch out: The cover must be fine enough for the sheaf and the degree

A cover by m open sets has no q-cochains for $q \geq m$ (with alternating cochains), so it cannot see H^q for $q \geq m$. More generally, a cover whose overlaps have hidden topology gives answers that are simply wrong, as Example 3.16 shows. Always check the Leray condition before trusting a finite computation.

Paracompactness and the other definition

Nash mentions that there is a second, more abstract definition of sheaf cohomology using *resolutions* of the sheaf (Wells). One replaces $\mathcal{F}$ by an exact sequence $0 \to \mathcal{F} \to \mathcal{L}^0 \to \mathcal{L}^1 \to \cdots$ of fine (or soft, or injective) sheaves and takes the cohomology of the complex of global sections $\mathcal{L}^0(M) \to \mathcal{L}^1(M) \to \cdots$. We will see exactly this happen in the de Rham theorem, where the resolution is $0 \to \underline{\mathbb{R}} \to \mathcal{E}^0 \to \mathcal{E}^1 \to \cdots$ (Section 3.6). The two definitions agree when M is **paracompact**: every open cover has a **locally finite** refinement, meaning that each point has a neighbourhood meeting only finitely many sets of the refinement. Every manifold is paracompact (Spivak), and paracompactness is also exactly what guarantees partitions of unity subordinate to any cover. For us: *on manifolds, all definitions of $H^q(M; \mathcal{F})$ agree, and a good cover computes it.*

3.3.6 The sphere, done right

Worked example 3.17: S^2 with a tetrahedral cover

Inflate a tetrahedron into a sphere. Its four faces become four closed spherical triangles; let U_0, U_1, U_2, U_3 be slightly enlarged open versions of them, with U_i the face *opposite* vertex i (Figure 3.15). Then:

- each pair of faces shares an edge, so U_{ij} is a thin strip along that edge: contractible, 6 of them;
- each triple of faces shares a vertex, so U_{ijk} is a small disc round that vertex: contractible, 4 of them;
- no point lies in all four faces, so $U_{0123} = \varnothing$.

This is a good cover, and its nerve is the (hollow) boundary of a tetrahedron, which is a sphere. With $\underline{\mathbb{R}}$ coefficients:

$$C^0 = \mathbb{R}^4, \qquad C^1 = \mathbb{R}^6, \qquad C^2 = \mathbb{R}^4, \qquad C^3 = 0.$$

Order the pairs $01, 02, 03, 12, 13, 23$ and the triples $012, 013, 023, 123$. The formulas (3.26)–(3.27) give

$$\delta_0 = \begin{pmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \\ 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 1 \end{pmatrix}, \qquad \delta_1 = \begin{pmatrix} 1 & -1 & 0 & 1 & 0 & 0 \\ 1 & 0 & -1 & 0 & 1 & 0 \\ 0 & 1 & -1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & -1 & 1 \end{pmatrix}. \tag{3.34}$$

(Row 012 of δ_1 reads $c_{12} - c_{02} + c_{01}$, and so on.)
Check $\delta_1\delta_0 = 0$ on the first row. Row 012 of δ_1 combines rows 01, 02, 12 of δ_0 with signs $+, -, +$:

$$\begin{aligned}(\delta_1\delta_0)_{012,\bullet} &\overset{(1)}{=} (-1,1,0,0) - (-1,0,1,0) + (0,-1,1,0) \\ &= (0,0,0,0).\end{aligned}$$

(1) Row 012 of δ_1 has entries $+1, -1, +1$ in the columns 01, 02, 12.

The other rows work the same way.
Ranks. $\operatorname{Ker}\delta_0$ is the constants, so $\operatorname{rank}\delta_0 = 3$. For δ_1: its rows $r_{012}, r_{013}, r_{023}$ are independent (each is the only one with an entry in column 12, 13, 23 respectively), and

$$r_{012} - r_{013} + r_{023} \overset{(1)}{=} (0,0,0,1,-1,1) = r_{123},$$

(1) Add the rows entry by entry: columns 01, 02, 03 cancel.

so $\operatorname{rank}\delta_1 = 3$. Therefore

$$\begin{aligned}\dim H^0 &= 4 - 3 = 1, \\ \dim H^1 &\overset{(1)}{=} (6 - \operatorname{rank}\delta_1) - \operatorname{rank}\delta_0 = 3 - 3 = 0, \\ \dim H^2 &\overset{(2)}{=} 4 - \operatorname{rank}\delta_1 = 1.\end{aligned}$$

(1) $\dim\operatorname{Ker}\delta_1 - \dim\operatorname{Im}\delta_0$.

(2) Every 2-cochain is a cocycle ($C^3 = 0$).

So $H^0(S^2;\mathbb{R}) = \mathbb{R}$, $H^1(S^2;\mathbb{R}) = 0$, $H^2(S^2;\mathbb{R}) = \mathbb{R}$: the correct answer. The alternating sum $4 - 6 + 4 = 2$ is the Euler characteristic of the sphere, and it equals $1 - 0 + 1$, as it must (the alternating sum of dimensions is the same for a complex and for its cohomology).

An explicit generator. The row relation above says that for every 1-cochain c, $(\delta c)_{123} - (\delta c)_{023} + (\delta c)_{013} - (\delta c)_{012} = 0$. So

$$\ell(n) = n_{123} - n_{023} + n_{013} - n_{012} \tag{3.35}$$

vanishes on coboundaries and defines a map $H^2 \to \mathbb{R}$. It is an isomorphism. It even works over $\mathbb{Z}$: if $\ell(n) = 0$, set $c_{12} = n_{012}$, $c_{13} = n_{013}$, $c_{23} = n_{023}$, all other $c = 0$; then δc agrees with n on 012, 013, 023 by construction, and on 123

$$(\delta c)_{123} \overset{(1)}{=} c_{23} - c_{13} + c_{12} = n_{023} - n_{013} + n_{012} \overset{(2)}{=} n_{123},$$

(1) Formula (3.27).

(2) The assumption $\ell(n) = 0$.

so $n = \delta c$. Hence $H^2(S^2;\mathbb{Z}) \cong \mathbb{Z}$ via ℓ. The integer $\ell(n)$ is the discrete analogue of $\int_{S^2}$; in Section 3.5 it will be the magnetic charge of a monopole.

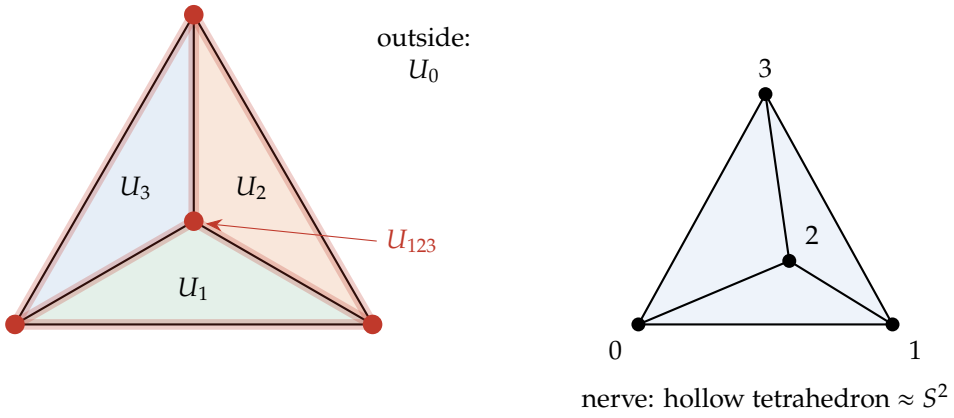


Figure 3.15: Left: the sphere, stereographically flattened so that face U_0 is everything outside the big triangle. Each pair of faces meets along an edge (red strips: the double overlaps U_{ij}) and each triple meets at a vertex (red dots: the triple overlaps U_{ijk}). Right: the nerve, with one vertex per face, is the surface of a tetrahedron. Four vertices, six edges, four triangles: $4 - 6 + 4 = 2$.

3.3.7 Fine sheaves have no higher cohomology

Nash asserts that for a soft sheaf $H^q(M;\mathcal{F}) = 0$ for $q > 0$ and that $\mathcal{E}$, $\mathcal{E}^p$ are soft (indeed fine). For fine sheaves the proof is one formula, the general version of the correction we used in Section 3.2.5.

Let $\mathcal{F}$ be fine, $\mathcal{U}$ a locally finite cover and $\{\rho_k\}$ a partition of unity subordinate to it. Define the **homotopy operator** $K : C^q(\mathcal{U};\mathcal{F}) \to C^{q-1}(\mathcal{U};\mathcal{F})$ by

$$(Kc)_{i_0\cdots i_{q-1}} = \sum_k \rho_k \, c_{k\, i_0\cdots i_{q-1}}. \tag{3.36}$$

Each term $\rho_k c_{k i_0\cdots i_{q-1}}$ is a section on $U_k \cap U_{i_0\cdots i_{q-1}}$ that vanishes near the edge of U_k, so it extends by zero to a section on all of $U_{i_0\cdots i_{q-1}}$; local finiteness makes the sum finite near each point.

Key idea: The homotopy formula

For $q \geq 1$ and every $c \in C^q(\mathcal{U};\mathcal{F})$: $\delta Kc + K\delta c = c$. Hence every cocycle is a coboundary, $c = \delta(Kc)$, and $H^q(\mathcal{U};\mathcal{F}) = 0$ for $q \geq 1$. Taking the limit, $H^q(M;\mathcal{F}) = 0$ for $q \geq 1$.

The check, for $q = 1$ (the general case is identical, with more indices):

$$\begin{aligned}(\delta Kc)_{ij} + (K\delta c)_{ij} &\overset{(1)}{=} \big((Kc)_j - (Kc)_i\big) + \sum_k \rho_k (\delta c)_{kij} \\ &\overset{(2)}{=} \sum_k \rho_k \big(c_{kj} - c_{ki}\big) + \sum_k \rho_k \big(c_{ij} - c_{kj} + c_{ki}\big) \\ &= \sum_k \rho_k\, c_{ij} \overset{(3)}{=} c_{ij}.\end{aligned}$$

(1) Formula (3.26) for δ of the 0-cochain Kc, and definition (3.36) for K of the 2-cochain δc.
(2) Definition of K again, and formula (3.27) for $(\delta c)_{kij}$.
(3) $\sum_k \rho_k = 1$.

For general q the same thing happens: the terms of $(K\delta c)$ in which the index k has been removed give $c_{i_0\cdots i_q} \sum \rho_k = c$, and the remaining terms cancel exactly against δKc.

Physical picture: Why floppy sheaves are cohomologically trivial

The formula says: *a partition of unity lets you spread any mismatch on overlaps back onto the open sets.* This is only possible if you are allowed to multiply sections by bump functions. That is why $\mathcal{E}$, $\mathcal{E}^p$, $\mathcal{E}^{p,q}$ and $\Gamma(\,\cdot\,, E)$ have $H^q = 0$ for $q > 0$, while the rigid sheaves $\underline{\mathbb{Z}}$, $\underline{\mathbb{R}}$, $\mathcal{O}$, $\mathcal{O}^*$ carry all the topology. The whole strategy of the next sections is to connect a rigid sheaf to floppy ones by exact sequences, and let the floppy ones do the work.

3.3.8 Holomorphic example: the Riemann sphere

Our last example uses the rigid sheaf $\mathcal{O}$ and shows H^1 doing genuinely analytic work: it counts the obstructions to solving a Mittag-Leffler type problem.

Worked example 3.18: $H^q(\mathbb{CP}^1;\mathcal{O}(n))$ by Laurent series

Use the cover $U_0 = \mathbb{C} \ni z$, $U_1 = \mathbb{CP}^1 \setminus \{0\} \ni w = 1/z$ of Example 3.6, with $U_{01} = \mathbb{C}^*$. Both $U_i \cong \mathbb{C}$ and $U_{01} = \mathbb{C}^*$ have no higher $\mathcal{O}$-cohomology (we justify this with the $\bar\partial$ equation in Section 3.7.3), so this two-set cover is a Leray cover and computes $H^q(\mathbb{CP}^1;\mathcal{O}(n))$. With two sets, $C^q = 0$ for $q \geq 2$, so $H^q = 0$ for $q \geq 2$.
A 0-cochain is a pair (s_0, s_1) of entire functions of z and of w. A 1-cochain is one holomorphic function $c(z)$ on $\mathbb{C}^*$, i.e. a convergent Laurent series $c(z) = \sum_{k\in\mathbb{Z}} c_k z^k$. The coboundary, written in the trivialisation over U_0 and using $s_0 = z^n s_1$ (3.7), is

$$\begin{aligned}(\delta s)_{01}(z) &\overset{(1)}{=} s_0(z) - z^n\, s_1(1/z) \\ &\overset{(2)}{=} \sum_{k\geq 0} a_k z^k - \sum_{m\geq 0} b_m\, z^{\,n-m},\end{aligned}$$

(1) Compare the two local sections in the same trivialisation (the sign convention is irrelevant for kernels and images).
(2) Taylor series $s_0 = \sum a_k z^k$, $s_1(w) = \sum b_m w^m$, and $w = 1/z$.

So the image of δ consists of all Laurent series whose nonzero coefficients have exponent $k \geq 0$ (from s_0) or $k \leq n$ (from s_1). (Convergent series split this way: the part with $k \geq 0$ converges on $\mathbb{C}$, the part with $k < 0$ converges on $\mathbb{CP}^1 \setminus \{0\}$.)
Case $n \geq -1$. Every integer k satisfies $k \geq 0$ or $k \leq n$, since $n \geq -1$. So every Laurent series is a coboundary and $H^1(\mathbb{CP}^1; \mathcal{O}(n)) = 0$.
Case $n \leq -2$. The exponents $k = n+1, n+2, \dots, -1$ are missed: they are negative, and larger than n. So

$$H^1(\mathbb{CP}^1; \mathcal{O}(n)) \cong \operatorname{span}\{z^{n+1}, \dots, z^{-1}\}, \qquad \dim H^1 = -n-1.$$

Together with $\dim H^0 = n+1$ for $n \geq 0$ and 0 for $n < 0$ (Example 3.6):

n	-4	-3	-2	-1	0	1	2
$\dim H^0(\mathbb{CP}^1; \mathcal{O}(n))$	0	0	0	0	1	2	3
$\dim H^1(\mathbb{CP}^1; \mathcal{O}(n))$	3	2	1	0	0	0	0
$\dim H^0 - \dim H^1$	-3	-2	-1	0	1	2	3

The last row is always $n+1$. This is the Riemann–Roch theorem for the sphere, $\dim H^0 - \dim H^1 = \deg + 1 - g$ with genus $g = 0$ and degree n, and it is the first example of an *index*: the individual dimensions jump around, but their difference is topological. The case $n = -2$ is the bundle of holomorphic 1-forms ($dz = -w^{-2}dw$ transforms with z^{-2} in our convention), and $H^1(\mathbb{CP}^1; \Omega^1) \cong \mathbb{C}$ is spanned by the class of z^{-1}: the residue. In particular (case $n = 0$) $H^1(\mathbb{CP}^1; \mathcal{O}) = 0$.

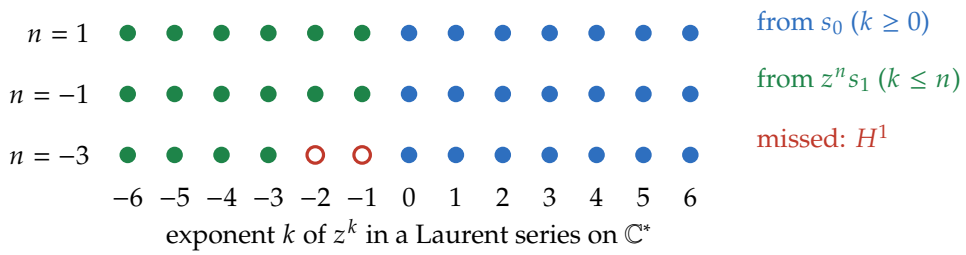


Figure 3.16: Which Laurent monomials on the overlap $\mathbb{C}^*$ are coboundaries for $\mathcal{O}(n)$ on $\mathbb{CP}^1$. Filled dots are hit by entire functions of z (blue) or by $z^n \times$ entire functions of $1/z$ (green). Open red circles are missed, and span $H^1(\mathbb{CP}^1; \mathcal{O}(n))$. For $n \geq -1$ nothing is missed; for $n = -3$ the two monomials z^{-2}, z^{-1} are missed.

Reading Nash (pp. 61–62, (3.12)–(3.17))

Nash's $\Delta = (O_1, \dots, O_{q+1})$ is our overlap $U_{i_0 \cdots i_q}$, and his (3.12) $C^q(O, \mathcal{F}) = \mathcal{F}(\Delta)$ means "one section for each such overlap", our (3.24). In (3.13) the sum runs over the $q+2$ sets of a $(q+1)$-cochain's overlap (his indexing is slightly loose), and $r^{\Delta}_{\Delta_i}$ is the restriction from the overlap with the i-th set omitted; this is our (3.25), and his d is our δ. His (3.14)–(3.15) are (3.29)–(3.30). The limit (3.16) over ever finer covers is (3.33), with refinement as in his footnote 2; we have added Leray's theorem, which is how one actually computes. His (3.17), $H^0 = \Gamma(M, \mathcal{F})$, is our first key idea. His remarks on paracompactness and the alternative definition via resolutions are expanded in Section 3.3.5; the vanishing of higher cohomology of soft and fine sheaves, which he states without proof, is Section 3.3.7.

Recap

- A Čech q-cochain is a section on every $(q+1)$-fold overlap of a cover. The coboundary δ is the alternating sum of restrictions; $\delta^2 = 0$. On the nerve, δ is discrete grad, curl, div.
- $H^q(\mathcal{U};\mathcal{F})$ = cocycles/coboundaries. $H^0 = \Gamma(M,\mathcal{F})$ always. H^1 = twisted gluing data modulo re-gauging.
- Sheaf cohomology $H^q(M;\mathcal{F})$ is the limit over refinements; in practice a Leray (good) cover computes it. Too coarse a cover gives wrong answers (two caps on S^2).
- Worked answers: $H^*(S^1;\mathbb{R}) = (\mathbb{R},\mathbb{R})$ from two or three arcs; Möbius = nontrivial class in $H^1(S^1;\mathbb{Z}_2)$; $H^*(S^2;\mathbb{R}) = (\mathbb{R},0,\mathbb{R})$ from the tetrahedral cover, with $H^2(S^2;\mathbb{Z}) \cong \mathbb{Z}$ via $\ell(n)$; $H^1(\mathbb{CP}^1;\mathcal{O}(n))$ of dimension $\max(0,-n-1)$ by Laurent series.
- Fine sheaves have $H^q = 0$ for $q > 0$, by the homotopy formula $\delta K + K\delta = 1$ built from a partition of unity.

3.4 The long exact sequence

We now have two facts that pull in opposite directions. A short exact sequence of sheaves $0 \to \mathcal{E} \to \mathcal{F} \to \mathcal{G} \to 0$ is exact near every point, but its global sections need not be exact at the right-hand end (Section 3.2.4). And we have a cohomology theory $H^q(M;\,\cdot\,)$ whose degree-zero part is global sections. Nash's (3.18)–(3.19) connects them: the failure of exactness at the right end of the global-section sequence is measured by H^1 of the left-hand sheaf, the failure one step further along by H^2, and so on. The result is the single most useful computational tool in the subject.

Key idea: The long exact cohomology sequence

A short exact sequence of sheaves $0 \to \mathcal{E} \xrightarrow{e} \mathcal{F} \xrightarrow{f} \mathcal{G} \to 0$ (Nash's (3.18)) produces a long exact sequence (Nash's (3.19))

$$\begin{aligned} 0 \to\ & H^0(M;\mathcal{E}) \xrightarrow{e_*} H^0(M;\mathcal{F}) \xrightarrow{f_*} H^0(M;\mathcal{G}) \\ & \xrightarrow{\delta^*} H^1(M;\mathcal{E}) \xrightarrow{e_*} H^1(M;\mathcal{F}) \xrightarrow{f_*} H^1(M;\mathcal{G}) \\ & \xrightarrow{\delta^*} H^2(M;\mathcal{E}) \xrightarrow{e_*} H^2(M;\mathcal{F}) \xrightarrow{f_*} \cdots \end{aligned} \tag{3.37}$$

The maps e_*, f_* apply e, f to cochains. The new ingredient is the **connecting homomorphism** $\delta^* : H^q(M;\mathcal{G}) \to H^{q+1}(M;\mathcal{E})$, which raises the degree by one.

The first row is exactly the global-section sequence (3.18). So the long exact sequence says, in its first two rows:

$$\operatorname{Coker}\big(\mathcal{F}(M) \to \mathcal{G}(M)\big) \ \hookrightarrow\ H^1(M;\mathcal{E}). \tag{3.38}$$

If $H^1(M;\mathcal{E}) = 0$, then $\mathcal{F}(M) \to \mathcal{G}(M)$ is onto. That is the key idea of Section 3.2.5 again: a fine $\mathcal{E}$ has $H^1 = 0$ (Section 3.3.7).

3.4.1 Building the connecting homomorphism

The recipe is a "zig-zag" through the diagram of cochain groups. Fix a cover $\mathcal{U}$. Applying e and f entry by entry gives, in each degree, a sequence of cochain groups, and e, f commute with δ (both are built from restrictions, and sheaf morphisms commute with restriction):

$$\begin{array}{ccccccc} 0 & \longrightarrow & C^{q+1}(\mathcal{U};\mathcal{E}) & \xrightarrow{e} & C^{q+1}(\mathcal{U};\mathcal{F}) & \xrightarrow{f} & C^{q+1}(\mathcal{U};\mathcal{G}) \\ & & \uparrow{\scriptstyle\delta} & & \uparrow{\scriptstyle\delta} & & \uparrow{\scriptstyle\delta} \\ 0 & \longrightarrow & C^{q}(\mathcal{U};\mathcal{E}) & \xrightarrow{e} & C^{q}(\mathcal{U};\mathcal{F}) & \xrightarrow{f} & C^{q}(\mathcal{U};\mathcal{G}) \end{array} \tag{3.39}$$

Working definition 3.14 (connecting homomorphism, by zig-zag)

Start with a cocycle $\gamma \in C^q(\mathcal{U};\mathcal{G})$, $\delta\gamma = 0$.

(1) *Lift* (go left): find $\beta \in C^q(\mathcal{U};\mathcal{F})$ with $f(\beta) = \gamma$. This may require refining the cover [replacing it by smaller open sets, each inside one of the old ones], because f is only surjective *locally*.

(2) *Go up:* form $\delta\beta \in C^{q+1}(\mathcal{U};\mathcal{F})$.

(3) *Observe:* $f(\delta\beta) = \delta f(\beta) = \delta\gamma = 0$. By exactness in the middle [what f sends to 0 is exactly what e produces, and e is injective], $\delta\beta = e(\alpha)$ for a unique $\alpha \in C^{q+1}(\mathcal{U};\mathcal{E})$ (go left again).

(4) *Declare* $\delta^*[\gamma] = [\alpha] \in H^{q+1}(\mathcal{U};\mathcal{E})$ [the class of α, i.e. α up to adding a coboundary. Other choices of β change α only by such a term].

Operational summary: lift, take the Čech coboundary, and read off the answer, which automatically lies in the smaller sheaf.

Three things need checking, and each is one line.

- α *is a cocycle:* $e(\delta\alpha) = \delta e(\alpha) = \delta\delta\beta = 0$, and e is injective, so $\delta\alpha = 0$.
- *Independent of the lift:* another lift is $\beta' = \beta + e(\eta)$ (two lifts differ by something f kills), and then $\alpha' = \alpha + \delta\eta$, which is cohomologous to α.
- *Independent of the representative:* replacing γ by $\gamma + \delta\kappa$ we may lift κ to $\tilde{\kappa}$ and take $\beta + \delta\tilde{\kappa}$ as the new lift; its coboundary is unchanged.

Why the sequence is exact (one spot, carefully)

Let us check exactness at $H^1(M;\mathcal{E})$, i.e. $\operatorname{Im}\delta^* = \operatorname{Ker} e_*$. The other spots are similar diagram chases.

$e_* \circ \delta^* = 0$: by construction $e(\alpha) = \delta\beta$ is a coboundary in $C^1(\mathcal{U};\mathcal{F})$, so $e_*[\alpha] = 0$.

$\operatorname{Ker} e_* \subset \operatorname{Im}\delta^*$: suppose α is a 1-cocycle of $\mathcal{E}$ and $e(\alpha) = \delta\beta$ for some 0-cochain β of $\mathcal{F}$. Put $\gamma = f(\beta)$. Then

$$\delta\gamma \overset{(1)}{=} f(\delta\beta) \overset{(2)}{=} f(e(\alpha)) \overset{(3)}{=} 0,$$

(1) f commutes with δ.

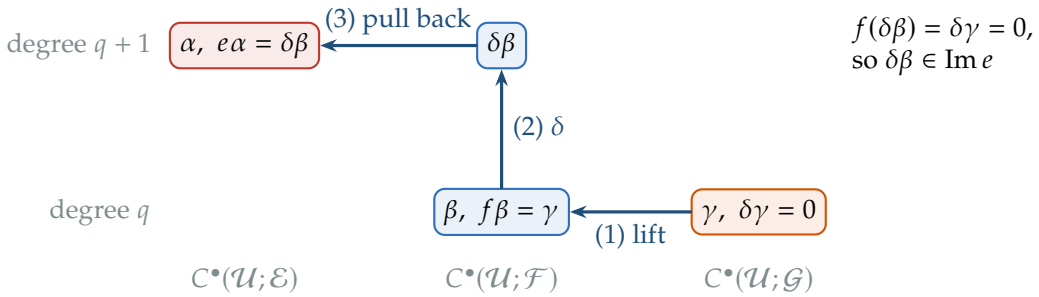


Figure 3.17: The zig-zag that defines δ^*. Start at the right with a cocycle γ of $\mathcal{G}$ (orange), lift it to $\mathcal{F}$, go up one degree with δ, and come back into $\mathcal{E}$ (red). The answer lands one degree higher than where we started. Every connecting homomorphism in this chapter, including the Chern class and the de Rham isomorphism, is this picture.

(2) The assumption $e(\alpha) = \delta\beta$.
(3) $f \circ e = 0$.

so γ is a 0-cocycle, i.e. a global section of $\mathcal{G}$, and running the zig-zag on γ with the lift β returns exactly α. So $[\alpha] = \delta^*[\gamma]$.

In words: *a class in $H^1(M;\mathcal{E})$ dies in $H^1(M;\mathcal{F})$ exactly when it is the obstruction to lifting some global section of $\mathcal{G}$.* This is the precise version of Nash's remark that sheaf cohomology "can measure any obstruction to exactness".

3.4.2 The exponential sequence on the punctured plane

Worked example 3.19: δ^* is the winding number

Take the exponential sequence $0 \to \underline{\mathbb{Z}} \to \mathcal{O} \to \mathcal{O}^* \to 0$ on $M = \mathbb{C}^*$ and the global section $g(z) = z$ of $\mathcal{O}^*$. We showed in Example 3.11 that g has no global logarithm. Now we compute the obstruction $\delta^*[g] \in H^1(\mathbb{C}^*;\underline{\mathbb{Z}})$.
Cover $\mathbb{C}^*$ by two slit planes (Figure 3.18):

$$U_1 = \mathbb{C} \setminus (-\infty, 0], \ \arg z \in (-\pi, \pi); \qquad U_2 = \mathbb{C} \setminus [0, \infty), \ \arg z \in (0, 2\pi).$$

Each is simply connected, and $U_{12} = H_+ \sqcup H_-$ is the union of the open upper and lower half-planes. This is a good cover.
(1) Lift. On each U_i, z has a logarithm. Take $f_i(z) = \frac{1}{2\pi i}(\ln|z| + i \arg_i z)$, with $\arg_i$ the branch in the stated range, so that $e^{2\pi i f_i} = z$ on U_i.
(2) Coboundary. On the two components of the overlap:

$$(\delta f)_{12}\big|_{H_+} \overset{(1)}{=} \frac{1}{2\pi i}\, i\,(\arg_2 z - \arg_1 z)\Big|_{H_+} \overset{(2)}{=} \frac{1}{2\pi}(\phi - \phi) = 0,$$
$$(\delta f)_{12}\big|_{H_-} \overset{(1)}{=} \frac{1}{2\pi}(\arg_2 z - \arg_1 z)\Big|_{H_-} \overset{(3)}{=} \frac{1}{2\pi}((\phi + 2\pi) - \phi) = 1.$$

(1) $(\delta f)_{12} = f_2 - f_1$; the $\ln|z|$ terms cancel.
(2) In H_+ both branches give the same angle $\phi \in (0, \pi)$.
(3) In H_-, $\arg_1 z = \phi \in (-\pi, 0)$ while $\arg_2 z = \phi + 2\pi$.

(3) Read off. The result $(0, 1)$ is integer-valued, as it must be: it lies in $\underline{\mathbb{Z}}$. By Example 3.14 (the same combinatorics: two sets, overlap with two contractible pieces), $H^1(\mathcal{U};\underline{\mathbb{Z}}) = \mathbb{Z}^2/\{(m, m)\} \cong \mathbb{Z}$ via $(c_+, c_-) \mapsto c_- - c_+$. So

$$\delta^*[z] = 1 \in H^1(\mathbb{C}^*;\underline{\mathbb{Z}}) \cong \mathbb{Z}.$$

Repeating with $g(z) = z^m$ multiplies every argument by m and gives $\delta^*[z^m] = m$. For a general nonvanishing g, the jump of $\frac{1}{2\pi}\arg g$ between the two half-planes is the winding number $\frac{1}{2\pi i}\oint g'/g\,\mathrm{d}z$. So the connecting homomorphism is the winding number, and it vanishes exactly when g has a global logarithm.

Push the example one step further along the long exact sequence:

$$\mathcal{O}(\mathbb{C}^*) \xrightarrow{\exp} \mathcal{O}^*(\mathbb{C}^*) \xrightarrow{\delta^*} H^1(\mathbb{C}^*;\underline{\mathbb{Z}}) \longrightarrow H^1(\mathbb{C}^*;\mathcal{O}).$$

It turns out that $H^1(\mathbb{C}^*;\mathcal{O}) = 0$ (the $\bar\partial$ equation can be solved on any open subset of $\mathbb{C}$; see Section 3.7.3). Exactness then says δ^* is onto $\mathbb{Z}$ and that its kernel is $\exp\mathcal{O}(\mathbb{C}^*)$. In plain terms: *every nonvanishing holomorphic function on $\mathbb{C}^*$ is $z^m e^{2\pi i h(z)}$ for a unique integer m (its winding number) and some holomorphic h.* A structure theorem, obtained without touching a single integral.

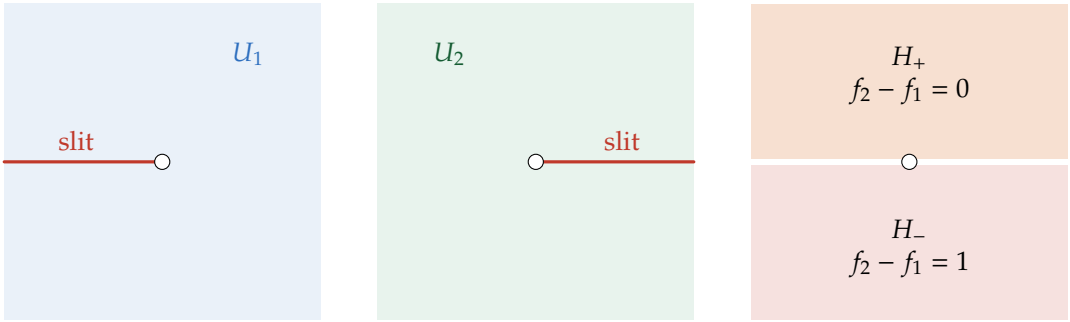


Figure 3.18: Computing $\delta^*[z]$ on $\mathbb{C}^*$. Each slit plane carries a branch of $\frac{1}{2\pi i}\log z$. On the overlap (right) the two branches agree in the upper half-plane and differ by exactly 1 in the lower half-plane. That integer mismatch is the Čech 1-cocycle representing the winding number.

3.5 Line bundles, $H^1(M;\mathcal{O}^*)$ and the first Chern class

This section is not in Nash's §2, but the brief for these notes asks for it, and it is the reason a physicist should care about H^1 and H^2. It also prepares for the characteristic classes of §5 of Nash's chapter. The statement is:

> *Complex line bundles (equivalently $U(1)$ gauge fields' "topological sectors") are classified by H^1 of a multiplicative sheaf; the exponential sequence turns this into an integer class $c_1 \in H^2(M;\mathbb{Z})$; and for the Dirac monopole c_1 is the quantised magnetic charge.*

3.5.1 Line bundles are 1-cocycles

A complex line bundle $L \to M$ is given by a cover $\{U_i\}$ and transition functions $g_{ij} : U_{ij} \to \mathbb{C}^*$ with $s_i = g_{ij}s_j$ (3.5) and $g_{ij}g_{jk} = g_{ik}$ (3.6). In the language of the previous section:

- the collection (g_{ij}) is a 1-cochain with coefficients in the multiplicative sheaf $\mathcal{E}^*$ (smooth L) or $\mathcal{O}^*$ (holomorphic L);
- the cocycle condition $g_{ij}g_{jk}g_{ik}^{-1} = 1$ says exactly that $\delta g = 1$ (the multiplicative form of (3.27)).

Now change the trivialisations: $s'_i = h_i s_i$ with $h_i : U_i \to \mathbb{C}^*$. Then

$$s'_i \overset{(1)}{=} h_i\, g_{ij}\, s_j \overset{(2)}{=} h_i\, g_{ij}\, h_j^{-1}\, s'_j, \qquad \text{so} \qquad g'_{ij} = g_{ij}\, h_i h_j^{-1} = g_{ij}\,(\delta h)_{ij}^{-1},$$

(1) Definition of s'_i and the transition rule $s_i = g_{ij} s_j$.
(2) Invert $s'_j = h_j s_j$.

using the multiplicative coboundary $(\delta h)_{ij} = h_j h_i^{-1}$. So isomorphic bundles have transition cocycles that differ by a coboundary.

Key idea: Line bundles = H^1 of the multiplicative sheaf

Isomorphism classes of smooth complex line bundles on M are in one-to-one correspondence with $H^1(M;\mathcal{E}^*)$; holomorphic line bundles on a complex manifold with $H^1(M;\mathcal{O}^*)$. Tensor product of bundles is multiplication of cocycles, $(L \otimes L')_{ij} = g_{ij} g'_{ij}$, and the dual bundle has cocycle g_{ij}^{-1}. So line bundles form an abelian group, the **Picard group** $\mathrm{Pic}(M) = H^1(M;\mathcal{O}^*)$.

For $U(1)$ gauge theory one may use $U(1)$-valued transition functions instead of $\mathbb{C}^*$-valued ones. Nothing changes: any $g : U \to \mathbb{C}^*$ can be written $g = |g| \cdot (g/|g|)$, and the positive factor $|g| = e^{\ln|g|}$ is the exponential of a smooth real function, which a partition of unity can always absorb into the h_i. So smooth complex line bundles = $U(1)$ bundles, and the transition functions are the gauge transformations between patches.

Worked example 3.20: The bundles $\mathcal{O}(n)$ as classes

On $\mathbb{CP}^1$ with the two-set cover, $\mathcal{O}(n)$ has the single transition function $g_{01} = z^n$ (3.7). Tensor products: $\mathcal{O}(m) \otimes \mathcal{O}(n)$ has $z^m z^n = z^{m+n}$, so $\mathcal{O}(m) \otimes \mathcal{O}(n) \cong \mathcal{O}(m+n)$, and $\mathcal{O}(n)^* \cong \mathcal{O}(-n)$. Are the $\mathcal{O}(n)$ different from each other? A coboundary on this cover is $h_1 h_0^{-1}$ with h_0 nonvanishing entire in z and h_1 nonvanishing entire in $w = 1/z$. A nonvanishing entire function has winding number 0 round any circle (its logarithm is entire), so every coboundary has winding number 0 round $|z| = 1$, while z^n has winding number n. Hence the classes $[z^n]$ are all distinct. We will now see that the winding number is the first Chern class.

3.5.2 The first Chern class from the exponential sequence

Apply the long exact sequence to the *smooth* exponential sequence (3.16), $0 \to \underline{\mathbb{Z}} \to \mathcal{E}_{\mathbb{C}} \to \mathcal{E}^* \to 0$. The middle sheaf is fine, so $H^q(M;\mathcal{E}_{\mathbb{C}}) = 0$ for $q \geq 1$, and the relevant stretch of (3.37) is

$$0 = H^1(M;\mathcal{E}_{\mathbb{C}}) \longrightarrow H^1(M;\mathcal{E}^*) \xrightarrow{\delta^*} H^2(M;\underline{\mathbb{Z}}) \longrightarrow H^2(M;\mathcal{E}_{\mathbb{C}}) = 0. \tag{3.40}$$

So δ^* is an *isomorphism*: smooth complex line bundles are classified by an integer cohomology class, their **first Chern class**.

Working definition 3.15 (Čech formula for c_1)

Given transition functions g_{ij} on a good cover [every set and every nonempty overlap is contractible], choose on each (connected, contractible) overlap a branch [one value of the multivalued log, so λ_{ij} is fixed up to adding an integer] of

$$\lambda_{ij} = \frac{i}{2\pi} \log g_{ij}, \qquad \text{so that} \qquad g_{ij} = e^{-2\pi i \lambda_{ij}}, \tag{3.41}$$

with $\lambda_{ji} = -\lambda_{ij}$. Then

$$n_{ijk} = \lambda_{jk} - \lambda_{ik} + \lambda_{ij} = (\delta\lambda)_{ijk} \tag{3.42}$$

is a locally constant *integer* on each triple overlap, $\delta n = 0$, and $c_1(L) = [n] \in H^2(M;\mathbb{Z})$. *Operational recipe:* take logs of the transition functions, add them round each triangle of the nerve [each triple i, j, k with U_{ijk} nonempty], the integers you get are the Chern cocycle.

Why the n_{ijk} are integers. Exponentiate:

$$e^{-2\pi i\, n_{ijk}} \overset{(1)}{=} e^{-2\pi i\lambda_{jk}}\, e^{+2\pi i\lambda_{ik}}\, e^{-2\pi i\lambda_{ij}} \overset{(2)}{=} g_{jk}\, g_{ik}^{-1}\, g_{ij} \overset{(3)}{=} 1,$$

(1) Definition (3.42); exponentials turn sums into products.
(2) Definition (3.41).
(3) The cocycle condition $g_{ij}g_{jk} = g_{ik}$ (the g's are numbers, so they commute).

A continuous function n_{ijk} with $e^{-2\pi i n_{ijk}} = 1$ takes integer values, and is constant on each connected component. It is a Čech cocycle because it is a coboundary ($\delta n = \delta\delta\lambda = 0$) of real-valued cochains; it is *not* in general the coboundary of an integer-valued cochain, and that is the whole point. This is exactly the zig-zag of Figure 3.17: $\gamma = g$, lift $\beta = \lambda$, $\alpha = n = \delta\lambda$.

Why the class is well defined. Two kinds of choice were made.

- *Branches of the logarithm:* $\lambda_{ij} \mapsto \lambda_{ij} + m_{ij}$ with $m_{ij} \in \mathbb{Z}$ changes n by δm, an integer coboundary.
- *Trivialisations:* $g_{ij} \mapsto g_{ij} h_i h_j^{-1}$ changes λ_{ij} by $\mu_i - \mu_j$ with $\mu_i = \frac{i}{2\pi}\log h_i$ (a logarithm exists on the contractible U_i), i.e. by $-(\delta\mu)_{ij}$, and then n changes by $-\delta\delta\mu = 0$.

Watch out: The sign of c_1 is a convention

Nash's exponential map is $f \mapsto e^{2\pi i f}$, so his connecting homomorphism uses $f_{ij} = \frac{1}{2\pi i}\log g_{ij} = -\lambda_{ij}$ and produces $-[n]$. Different books also differ on whether g_{ij} acts on components ($s_i = g_{ij}s_j$, as here) or on frames, which inverts g. The only convention-independent statement is the normalisation: *c_1 is fixed so that $c_1(\mathcal{O}(1)) = +1$ on $\mathbb{CP}^1$ with its complex orientation, and so that c_1 agrees with the curvature integral $\frac{i}{2\pi}\int F$* (Section 3.6.4). Our $\lambda_{ij} = \frac{i}{2\pi}\log g_{ij}$ is chosen to make both hold, and it parallels the real curvature form $\frac{i}{2\pi}F$ of Section 3.6.4.

The holomorphic version

For holomorphic bundles the middle sheaf $\mathcal{O}$ is *not* fine, and the long exact sequence of $0 \to \mathbb{Z} \to \mathcal{O} \to \mathcal{O}^* \to 0$ reads

$$H^1(M;\mathbb{Z}) \to H^1(M;\mathcal{O}) \to H^1(M;\mathcal{O}^*) \xrightarrow{c_1} H^2(M;\mathbb{Z}) \to H^2(M;\mathcal{O}). \tag{3.43}$$

So a holomorphic line bundle still has an integer Chern class, but now

- bundles with $c_1 = 0$ need not be holomorphically trivial: they form the group obtained from $H^1(M;\mathcal{O})$ by dividing out the image of $H^1(M;\mathbb{Z})$. This is a complex torus when M is compact Kähler, called the **Picard variety** or **Jacobian**;
- not every integer class need come from a holomorphic bundle: the obstruction lies in $H^2(M;\mathcal{O})$.

Worked example 3.21: Holomorphic line bundles on $\mathbb{CP}^1$ and on a torus

$\mathbb{CP}^1$. $H^1(\mathbb{CP}^1;\mathcal{O}) = 0$ (Example 3.18, $n = 0$) and $H^2(\mathbb{CP}^1;\mathcal{O}) = 0$ (a two-set Leray cover has no 2-cochains). By (3.43), $c_1 : \mathrm{Pic}(\mathbb{CP}^1) \to H^2(\mathbb{CP}^1;\mathbb{Z}) = \mathbb{Z}$ is an isomorphism. Every holomorphic line bundle on the Riemann sphere is $\mathcal{O}(n)$ for exactly one n, and $n = c_1$.
Torus $T = \mathbb{C}/(\mathbb{Z} + \tau\mathbb{Z})$. Here $H^1(T;\mathbb{Z}) = \mathbb{Z}^2$ and $H^1(T;\mathcal{O}) \cong \mathbb{C}$ (Example 3.33). The bundles with $c_1 = 0$ form $\mathbb{C}/\mathbb{Z}^2$, a torus again: a flat $U(1)$ bundle on T is specified by the two holonomies ("Wilson lines") $e^{i\theta_1}, e^{i\theta_2}$ round the two cycles, and each such flat bundle carries a unique holomorphic structure. The moduli space of flat line bundles is the dual torus.

3.5.3 The Dirac monopole as a Čech cocycle

Now the physics. A magnetic monopole of strength g_m at the origin of $\mathbb{R}^3$ has field $\vec{B} = g_m\,\hat{r}/r^2$, i.e. the closed 2-form

$$F^{\mathrm{phys}} = g_m \sin\theta\,\mathrm{d}\theta\wedge\mathrm{d}\phi \qquad \text{on } S^2 \text{ (any sphere round the origin)}, \qquad \int_{S^2} F^{\mathrm{phys}} = 4\pi g_m. \tag{3.44}$$

(We write g_m for the magnetic charge, to avoid a clash with the transition functions g_{ij}; units $\hbar = c = 1$.) Because $\int_{S^2} F \neq 0$, Stokes' theorem forbids a global potential: if $F = \mathrm{d}A$ on all of S^2 then $\int_{S^2} F = \int_{\partial S^2} A = 0$. Wu and Yang's resolution is to use two patches.

Worked example 3.22: Two patches, and quantisation as a well-defined transition function

Let $U_N = \{\theta < \frac{\pi}{2} + \varepsilon\}$ and $U_S = \{\theta > \frac{\pi}{2} - \varepsilon\}$, and put

$$A_N^{\mathrm{phys}} = g_m(1-\cos\theta)\,\mathrm{d}\phi, \qquad A_S^{\mathrm{phys}} = -g_m(1+\cos\theta)\,\mathrm{d}\phi. \tag{3.45}$$

Each is smooth on its patch: $\mathrm{d}\phi$ is singular at the poles, but its coefficient vanishes at $\theta = 0$ for A_N and at $\theta = \pi$ for A_S (the would-be Dirac strings are pushed into the other patch). Both give the monopole field:

$$\mathrm{d}A_N^{\mathrm{phys}} \overset{(1)}{=} g_m\,\mathrm{d}(1-\cos\theta)\wedge\mathrm{d}\phi \overset{(2)}{=} g_m\sin\theta\,\mathrm{d}\theta\wedge\mathrm{d}\phi = F^{\mathrm{phys}},$$

(1) $\mathrm{d}(f\,\mathrm{d}\phi) = \mathrm{d}f\wedge\mathrm{d}\phi$ since $\mathrm{d}\mathrm{d}\phi = 0$ away from the poles.
(2) $\mathrm{d}(-\cos\theta) = \sin\theta\,\mathrm{d}\theta$.

and similarly for A_S. On the overlap they differ by a gauge transformation:

$$A_N^{\mathrm{phys}} - A_S^{\mathrm{phys}} \overset{(1)}{=} g_m\big[(1-\cos\theta)+(1+\cos\theta)\big]\mathrm{d}\phi = 2g_m\,\mathrm{d}\phi = \mathrm{d}\Lambda, \qquad \Lambda = 2g_m\phi.$$

(1) Subtract (3.45).

A particle of charge e couples through $D = \mathrm{d} - ieA^{\mathrm{phys}}$, and the wavefunctions in the two gauges are related by $\psi_N = e^{ie\Lambda}\psi_S = e^{2ieg_m\phi}\psi_S$. The transition function of the line bundle in which ψ lives is therefore

$$g_{NS} = e^{2ieg_m\phi} \quad \text{on the equatorial annulus } U_{NS}.$$

This must be a single-valued function on the annulus, i.e. a section of $\mathcal{E}^*(U_{NS})$. Going once round, $\phi \to \phi + 2\pi$:

$$e^{2ieg_m(\phi+2\pi)} = e^{2ieg_m\phi} \iff e^{4\pi ieg_m} = 1 \iff 2eg_m = n \in \mathbb{Z}. \tag{3.46}$$

This is **Dirac's quantisation condition**. The bundle is then the one with $g_{NS} = e^{in\phi}$, which on $\mathbb{CP}^1 = S^2$ is $\mathcal{O}(n)$ restricted to the equator (with $z = \tan(\theta/2)e^{i\phi}$, $U_N = U_0$, and $z^n = e^{in\phi}$ on $|z| = 1$).

The two-patch cover has no triple overlaps, so it cannot display c_1 as a Čech 2-cocycle: in this cover the integer is hidden in the requirement that g_{NS} be single-valued on a non-contractible overlap, just as the two-cap cover of S^2 hid H^2 (Example 3.16). To see the cocycle we refine.

Worked example 3.23: A three-set cover of S^2 and the monopole cocycle

Keep $V_1 = U_N$ and split the southern region into two overlapping halves that share a neighbourhood of the south pole (Figure 3.19):

$$V_2 = U_S \cap \big(\{\phi \in (-\varepsilon, \pi + \varepsilon)\} \cup D\big), \qquad V_3 = U_S \cap \big(\{\phi \in (\pi - \varepsilon, 2\pi + \varepsilon)\} \cup D\big),$$

where D is a small disc round the south pole. The overlaps are:

- V_{12}, V_{13}: the two halves of the equatorial annulus (contractible);
- V_{23}: a band through the south pole along the meridians $\phi = 0$ and $\phi = \pi$, together with D (contractible);
- V_{123}: *two* small squares on the equator, one near $\phi = 0$ and one near $\phi = \pi$.

All pieces are contractible, so this is a Leray cover for $\underline{\mathbb{Z}}$.
Cohomology of S^2 from this cover. $C^0 = \mathbb{Z}^3$, $C^1 = \mathbb{Z}^3$, $C^2 = \underline{\mathbb{Z}}(V_{123}) = \mathbb{Z}^2 \ni (m_0, m_\pi)$ (one integer per square). A 1-cochain (c_{12}, c_{13}, c_{23}) consists of constants on connected sets, so

$$(\delta c)_{123} = c_{23} - c_{13} + c_{12} \text{ takes the same value on both squares:} \qquad \operatorname{Im} \delta_1 = \{(m, m)\}.$$

Hence $H^2 = \mathbb{Z}^2/\{(m,m)\} \cong \mathbb{Z}$ via $\ell(m) = m_0 - m_\pi$. (One also checks $H^0 = \mathbb{Z}$, $H^1 = 0$ as for the tetrahedron.) The choice of sign in ℓ is the one compatible with integration, as the Čech–de Rham check in Example 3.27 confirms.
The monopole cocycle. The transition functions from V_1 (north gauge) to V_2, V_3 (south gauge) are $e^{in\phi}$, and V_2, V_3 use the same gauge:

$$g_{12} = e^{in\phi} \text{ on } V_{12}, \qquad g_{13} = e^{in\phi} \text{ on } V_{13}, \qquad g_{23} = 1.$$

Take logarithms with ϕ in the range natural to each set, $\phi \in (-\varepsilon, \pi + \varepsilon)$ on V_{12} and $\phi \in (\pi - \varepsilon, 2\pi + \varepsilon)$ on V_{13}:

$$\lambda_{12} \overset{(1)}{=} \frac{i}{2\pi}(in\phi) = -\frac{n\phi}{2\pi}, \qquad \lambda_{13} = -\frac{n\phi}{2\pi}, \qquad \lambda_{23} = 0.$$

(1) Definition (3.41); $\log e^{in\phi} = in\phi$ is a smooth branch on each (contractible) half-annulus.

Now evaluate (3.42) on the two squares:

$$n_{123}\big|_{\phi\approx\pi} \overset{(1)}{=} 0 - \left(-\frac{n\pi}{2\pi}\right) + \left(-\frac{n\pi}{2\pi}\right) = 0,$$
$$n_{123}\big|_{\phi\approx 0} \overset{(2)}{=} 0 - \left(-\frac{n\cdot 2\pi}{2\pi}\right) + \left(-\frac{n\cdot 0}{2\pi}\right) = n.$$

(1) Near $\phi = \pi$ both half-annuli use the same value of the angle.
(2) Near $\phi = 0$ the square is at $\phi \approx 0$ as seen from V_{12} but at $\phi \approx 2\pi$ as seen from V_{13}.

So $n = (n_0, n_\pi) = (n, 0)$ and $c_1 = \ell(n) = n$: the Chern class is the quantised magnetic charge $2eg_m$.
Dirac's condition as the cocycle condition. Suppose we did not know (3.46) and wrote $g_{12} = e^{2ieg_m\phi}$ on V_{12}, $g_{13} = e^{2ieg_m\phi}$ on V_{13}, $g_{23} = 1$. On this cover each of these is a perfectly good function, because each overlap is contractible and we never go round a loop. The only condition left is the cocycle condition $g_{12}g_{23} = g_{13}$ on V_{123}. Near $\phi = \pi$ it holds trivially. Near $\phi = 0$:

$$g_{12}\,g_{23} = g_{13} \overset{(1)}{\iff} e^{2ieg_m\phi} = e^{2ieg_m(\phi+2\pi)} \iff e^{4\pi ieg_m} = 1 \iff 2eg_m \in \mathbb{Z}.$$

(1) On the square near $\phi = 0$ the angle is ϕ in V_{12} and $\phi + 2\pi$ in V_{13}.

Dirac quantisation is literally the Čech cocycle condition on the triple overlap, and the resulting integer is the class of the cocycle in $H^2(S^2;\mathbb{Z})$.

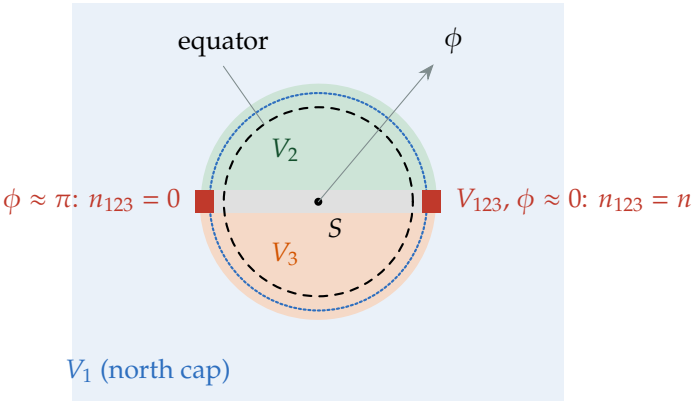


Figure 3.19: The three-set cover of S^2 used for the monopole, drawn by stereographic projection from the north pole (south pole at the centre, equator dashed, north cap V_1 everything outside the blue dotted circle). The southern halves V_2 (green, $0 < \phi < \pi$) and V_3 (orange, $\pi < \phi < 2\pi$) overlap along the grey band through S. The triple overlap is the two red squares on the equator. The angle ϕ is continuous on each half but jumps by 2π between V_{12} and V_{13} at $\phi \approx 0$: that is where the Chern integer n sits.

Dictionary: Monopole physics ↔ Čech data

Physics	Sheaf cohomology
local gauge potential A_i on patch U_i	0-cochain of $\mathcal{E}^1$
gauge transformation $e^{ie\Lambda_{ij}}$ between patches	1-cochain g_{ij} of $\mathcal{E}^*$
consistency on triple overlaps	$\delta g = 1$
global gauge change $\psi_i \mapsto e^{ie\chi_i}\psi_i$	$g \mapsto g \cdot (\delta h)^{-1}$
topological sector (bundle)	class in $H^1(M;\mathcal{E}^*)$
Dirac quantisation $2e g_m = n$	$\delta g = 1$ on a refined cover; $n = [\delta\lambda]$
magnetic charge n	$c_1 \in H^2(S^2;\mathbb{Z}) \cong \mathbb{Z}$
flux $\frac{e}{2\pi}\int F^{\text{phys}} = n$	de Rham image of c_1

Physical picture: Why magnetic charge is an integer

On each patch you may shift the phase of the wavefunction as you like. What you cannot change is the *mismatch round a triangle of patches*: the product of three gauge transformations round a triple overlap must be 1, so the sum of their logarithms is 2π times an integer. Summed over the sphere, those integers give the total flux in units of $2\pi/e$. Quantisation of magnetic charge is the statement that c_1 lives in $H^2(M;\mathbb{Z})$ rather than $H^2(M;\mathbb{R})$.

Reading Nash (pp. 62–63, (3.18)–(3.19))

Nash states the long exact sequence (3.19) for $0 \to \mathcal{E} \to \mathcal{F} \to \mathcal{G} \to 0$ (3.18) without constructing the connecting maps; he calls it "one of the most fundamental calculational tools". Our Section 3.4.1 builds δ^* by the zig-zag and checks exactness at one spot. Nash's

only application in §2 is the de Rham and Dolbeault theorems, which follow. Line bundles, c_1 via the exponential sequence (his (3.8)), and the monopole are our additions; Nash returns to Chern classes in §5 of this chapter (via curvature) and to monopoles and instantons in Chapter 8.

Recap

- A short exact sequence of sheaves gives a long exact sequence in cohomology. The connecting map δ^* is the zig-zag: lift, apply δ, pull back.
- δ^* for the exponential sequence on $\mathbb{C}^*$ is the winding number; with $H^1(\mathbb{C}^*;\mathcal{O}) = 0$ this gives every nonvanishing holomorphic function as $z^m e^{2\pi i h}$.
- Line bundles are 1-cocycles of $\mathcal{E}^*$ or $\mathcal{O}^*$; isomorphism classes form $H^1(M;\mathcal{E}^*)$ or $\mathrm{Pic}(M) = H^1(M;\mathcal{O}^*)$.
- The exponential sequence gives $c_1 : H^1(M;\mathcal{E}^*) \xrightarrow{\cong} H^2(M;\mathbb{Z})$, computed by $n_{ijk} = \delta\lambda$, $\lambda_{ij} = \frac{i}{2\pi}\log g_{ij}$. On $\mathbb{CP}^1$, $\mathrm{Pic} = \mathbb{Z}$ via $c_1(\mathcal{O}(n)) = n$.
- Dirac monopole: two patches, $g_{NS} = e^{2ieg_m\phi}$; on a refined cover Dirac's condition $2eg_m \in \mathbb{Z}$ is the cocycle condition, and $c_1 = 2eg_m$.

3.6 The de Rham theorem

We now have two completely different ways of computing "the cohomology of a circle".

- *Combinatorially.* Cover S^1 by arcs, write one real number on each overlap, and compute the kernel and image of a small integer matrix (Example 3.13, Example 3.14). Answer: $H^1(S^1;\mathbb{R}) \cong \mathbb{R}$.
- *Analytically.* Take closed 1-forms on S^1 modulo exact ones. Every 1-form is closed; $f(\theta)\,d\theta$ is exact precisely when $\oint f\,d\theta = 0$. Answer: $H^1_{\mathrm{dR}}(S^1) \cong \mathbb{R}$, via $\oint$.

The answers agree. Is that a coincidence? The **de Rham theorem** says it is not: for every manifold and every degree,

$$H^q(M;\mathbb{R}) \;\cong\; H^q_{\mathrm{dR}}(M) = \frac{\{\text{closed } q\text{-forms}\}}{\{\text{exact } q\text{-forms}\}}.$$

The left side knows only about open sets and locally constant numbers. The right side knows about derivatives and integrals. Nash proves the theorem in half a page, using the long exact sequence and the fact that the sheaves of forms are soft. In this section we unpack his argument step by step, then make the isomorphism completely explicit with the Čech–de Rham double complex, and finally run it on the monopole, where it turns "field strength" into "potentials on patches" into "gauge transformations on overlaps" into "an integer on triple overlaps".

Physical picture: Why a physicist should care

De Rham cohomology is where *fluxes* live: $\int_\Sigma F$ over closed surfaces, $\oint A$ round closed loops. Čech cohomology is where *patching data* live: gauge transformations between coordinate patches. The de Rham theorem says these are the same information. It is the reason that a flux through a closed surface can be read off from how the gauge potentials

are glued, and conversely.

3.6.1 The de Rham resolution

Recall the sheaves involved (Example 3.10):

- $\mathcal{E}^p$: smooth p-forms. Its sections over U are the p-forms on U. It is fine (multiply by a partition of unity).
- $\mathcal{Z}^p \subset \mathcal{E}^p$: *closed* p-forms, $\mathrm{d}\omega = 0$. It is not fine: $\rho\,\omega$ is usually not closed.
- $\mathcal{Z}^0 = \underline{\mathbb{R}}$: a closed 0-form is a function with $\mathrm{d}f = 0$, i.e. a locally constant function.

The exterior derivative is a sheaf morphism $\mathrm{d} : \mathcal{E}^p \to \mathcal{E}^{p+1}$ (it is local: $\mathrm{d}(\omega|_U) = (\mathrm{d}\omega)|_U$). Stringing these together we get a long sequence of sheaves

$$0 \longrightarrow \underline{\mathbb{R}} \xrightarrow{\iota} \mathcal{E}^0 \xrightarrow{\mathrm{d}} \mathcal{E}^1 \xrightarrow{\mathrm{d}} \mathcal{E}^2 \xrightarrow{\mathrm{d}} \cdots \xrightarrow{\mathrm{d}} \mathcal{E}^n \longrightarrow 0, \tag{3.47}$$

where ι regards a locally constant function as a smooth one.

Working definition 3.16 (fine resolution)

A **resolution** of a sheaf $\mathcal{F}$ is an exact sequence of sheaves [what one map kills, the previous map produces, on some small neighbourhood of each point] $0 \to \mathcal{F} \to \mathcal{L}^0 \to \mathcal{L}^1 \to \cdots$. It is a **fine resolution** if every $\mathcal{L}^p$ is fine [its sections can be cut up by a partition of unity, like smooth forms]. *Operational test for exactness:* on a small ball round any point, every section of $\mathcal{L}^p$ killed by the next map comes from a section of $\mathcal{L}^{p-1}$. For (3.47) that is the Poincaré lemma: *on a ball every closed form is exact.*

So (3.47) is a fine resolution of $\underline{\mathbb{R}}$. The content is entirely the Poincaré lemma, which we checked for 1-forms in (3.15). For higher forms the same radial trick works. On a star-shaped domain centred at 0 define, for a p-form ω,

$$(h\omega)(x) = \int_0^1 t^{\,p-1}\, \iota_x\, \omega(tx)\, \mathrm{d}t, \tag{3.48}$$

where ι_x inserts the radial vector $x^\mu \partial_\mu$ into the first slot. One checks $\mathrm{d}h\omega + h\,\mathrm{d}\omega = \omega$ for $p \geq 1$, so a closed ω is $\mathrm{d}(h\omega)$. Rather than prove the general identity, we check it where every physicist has used it.

Worked example 3.24: The Poincaré lemma for 2-forms: a vector potential for any divergence-free $\vec{B}$

In $\mathbb{R}^3$ a 2-form is a vector field $\vec{B}$ (via $B_x\,\mathrm{d}y \wedge \mathrm{d}z + \cdots$), and it is closed exactly when $\nabla \cdot \vec{B} = 0$. The formula (3.48) with $p = 2$ becomes

$$\vec{A}(\vec{x}) = \int_0^1 t\, \vec{B}(t\vec{x}) \times \vec{x}\, \mathrm{d}t. \tag{3.49}$$

We check $\nabla \times \vec{A} = \vec{B}$ when $\nabla \cdot \vec{B} = 0$. Write $\vec{b} = \vec{B}(t\vec{x})$ for the field evaluated at the scaled point, and use the identity $\nabla \times (\vec{u} \times \vec{v}) = \vec{u}(\nabla \cdot \vec{v}) - \vec{v}(\nabla \cdot \vec{u}) + (\vec{v} \cdot \nabla)\vec{u} - (\vec{u} \cdot \nabla)\vec{v}$ with $\vec{u} = \vec{b}$, $\vec{v} = \vec{x}$:

$$\nabla \times (\vec{b} \times \vec{x}) \overset{(1)}{=} \vec{b}\,(\nabla \cdot \vec{x}) - \vec{x}\,(\nabla \cdot \vec{b}) + (\vec{x} \cdot \nabla)\vec{b} - (\vec{b} \cdot \nabla)\vec{x}$$

$$\overset{(2)}{=} 3\,\vec{b} - \vec{x}\; t\,(\nabla\cdot\vec{B})(t\vec{x}) + t\,\frac{\mathrm{d}}{\mathrm{d}t}\vec{B}(t\vec{x}) - \vec{b}$$
$$\overset{(3)}{=} 2\,\vec{B}(t\vec{x}) + t\,\frac{\mathrm{d}}{\mathrm{d}t}\vec{B}(t\vec{x}).$$

(1) The vector identity for the curl of a cross product.

(2) $\nabla\cdot\vec{x} = 3$; the chain rule gives $\partial_j[B_k(t\vec{x})] = t(\partial_j B_k)(t\vec{x})$, so $\nabla\cdot\vec{b} = t(\nabla\cdot\vec{B})(t\vec{x})$ and $(\vec{x}\cdot\nabla)\vec{b} = t\,x^j(\partial_j\vec{B})(t\vec{x}) = t\frac{\mathrm{d}}{\mathrm{d}t}\vec{B}(t\vec{x})$; finally $(\vec{b}\cdot\nabla)\vec{x} = \vec{b}$.

(3) $\nabla\cdot\vec{B} = 0$ kills the second term; $3\vec{b} - \vec{b} = 2\vec{b}$.

Now multiply by t and integrate:

$$\nabla\times\vec{A} \overset{(1)}{=} \int_0^1 \Big(2t\,\vec{B}(t\vec{x}) + t^2\frac{\mathrm{d}}{\mathrm{d}t}\vec{B}(t\vec{x})\Big)\mathrm{d}t \overset{(2)}{=} \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t}\Big(t^2\,\vec{B}(t\vec{x})\Big)\mathrm{d}t \overset{(3)}{=} \vec{B}(\vec{x}).$$

(1) Differentiate (3.49) under the integral and use the previous chain.

(2) Product rule: $\frac{\mathrm{d}}{\mathrm{d}t}(t^2\vec{B}) = 2t\vec{B} + t^2\frac{\mathrm{d}\vec{B}}{\mathrm{d}t}$.

(3) Fundamental theorem of calculus; the $t = 0$ end vanishes because of the factor t^2.

For a uniform field $\vec{B} = B\hat{z}$ this gives $\vec{A} = \frac{1}{2}B\,\hat{z}\times\vec{x} = \frac{1}{2}B(-y, x, 0)$, the symmetric gauge of Landau-level problems. For the monopole field $\vec{B} = g_m\hat{r}/r^2$ the formula fails: the domain $\mathbb{R}^3\setminus\{0\}$ is not star-shaped, the segment from 0 to $\vec{x}$ passes through the monopole, and, as we know, no global $\vec{A}$ exists. Exactness of (3.47) is a *local* statement, true on balls; what fails globally is the business of cohomology.

Nash does not use the long resolution (3.47) directly. He chops it into short exact pieces, one for each degree:

$$\begin{gathered} 0\to\underline{\mathbb{R}}\to\mathcal{E}^0\xrightarrow{\mathrm{d}}\mathcal{Z}^1\to 0, \qquad 0\to\mathcal{Z}^1\to\mathcal{E}^1\xrightarrow{\mathrm{d}}\mathcal{Z}^2\to 0, \qquad \dots, \\ 0\to\mathcal{Z}^p\to\mathcal{E}^p\xrightarrow{\mathrm{d}}\mathcal{Z}^{p+1}\to 0, \qquad \dots \end{gathered} \tag{3.50}$$

This is his (3.20). Each piece is exact by the Poincaré lemma, and each has a fine sheaf in the middle. Figure 3.20 shows how the pieces fit back together: every $\mathcal{Z}^{p+1}$ is at the same time the image of $\mathcal{E}^p$ and the kernel in $\mathcal{E}^{p+1}$.

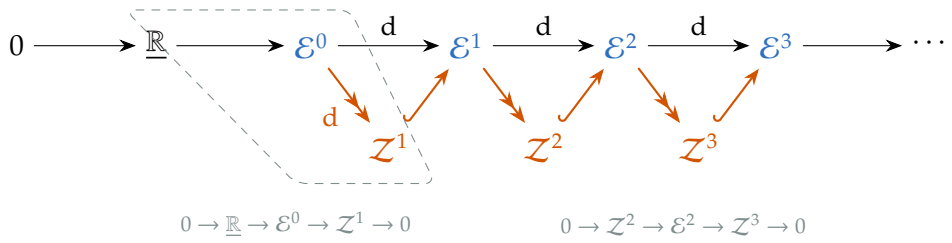


Figure 3.20: Splicing. The long exact de Rham resolution (top row) is built from the short exact sequences (3.50) (the "V" shapes, one of which is outlined): $\mathcal{E}^p$ maps *onto* the closed forms $\mathcal{Z}^{p+1}$ (Poincaré lemma), which then sit *inside* $\mathcal{E}^{p+1}$. Blue sheaves are fine and cohomologically invisible; orange sheaves carry the topology.

3.6.2 Nash's dimension-shifting argument

Feed one short sequence $0 \to \mathcal{Z}^p \to \mathcal{E}^p \to \mathcal{Z}^{p+1} \to 0$ into the long exact sequence (3.37). A typical stretch is

$$\cdots \to H^k(M;\mathcal{E}^p) \to H^k(M;\mathcal{Z}^{p+1}) \xrightarrow{\delta^*} H^{k+1}(M;\mathcal{Z}^p) \to H^{k+1}(M;\mathcal{E}^p) \to \cdots \tag{3.51}$$

Because $\mathcal{E}^p$ is fine, $H^k(M;\mathcal{E}^p) = 0$ for every $k \geq 1$ (Section 3.3.7). Two cases follow.

Case $k \geq 1$. Both outer groups in (3.51) vanish:

$$0 \to H^k(M;\mathcal{Z}^{p+1}) \xrightarrow{\delta^*} H^{k+1}(M;\mathcal{Z}^p) \to 0,$$

and a map sandwiched between zeros in an exact sequence is an isomorphism (injective because the kernel is the image of 0; surjective because the image is the kernel of the map to 0, which is everything). So

$$H^{k+1}(M;\mathcal{Z}^p) \cong H^k(M;\mathcal{Z}^{p+1}) \qquad (k \geq 1). \tag{3.52}$$

Moving one step up in the sheaf lowers the degree by one.

Case $k = 0$. Now the left end is $H^0(M;\mathcal{E}^p) = \mathcal{E}^p(M)$, the global p-forms, which is certainly not zero. The sequence is

$$\mathcal{E}^p(M) \xrightarrow{\mathrm{d}} \mathcal{Z}^{p+1}(M) \xrightarrow{\delta^*} H^1(M;\mathcal{Z}^p) \longrightarrow H^1(M;\mathcal{E}^p) = 0,$$

so δ^* is onto, with kernel the image of d:

$$H^1(M;\mathcal{Z}^p) \cong \frac{\mathcal{Z}^{p+1}(M)}{\mathrm{d}\,\mathcal{E}^p(M)} = \frac{\text{closed } (p+1)\text{-forms on } M}{\text{exact } (p+1)\text{-forms on } M} = H^{p+1}_{\mathrm{dR}}(M). \tag{3.53}$$

This is Nash's (3.22), with $H^0(M;\mathcal{Z}^{p+1}) = \mathcal{Z}^{p+1}(M)$ and $H^0(M;\mathcal{E}^p) = \mathcal{E}^p(M)$ by (3.17).

Now chain the isomorphisms. For $q \geq 1$,

$$\begin{aligned} H^q(M;\underline{\mathbb{R}}) &\overset{(1)}{=} H^q(M;\mathcal{Z}^0) \\ &\overset{(2)}{=} H^{q-1}(M;\mathcal{Z}^1) \overset{(2)}{=} H^{q-2}(M;\mathcal{Z}^2) \overset{(2)}{=} \cdots \overset{(2)}{=} H^1(M;\mathcal{Z}^{q-1}) \\ &\overset{(3)}{=} \frac{\mathcal{Z}^q(M)}{\mathrm{d}\,\mathcal{E}^{q-1}(M)} \overset{(4)}{=} H^q_{\mathrm{dR}}(M). \end{aligned} \tag{3.54}$$

(1) $\underline{\mathbb{R}} = \mathcal{Z}^0$: locally constant functions are the closed 0-forms.
(2) Each step is (3.52) with $(k+1, p) = (q-j, j)$, which needs $k = q-j-1 \geq 1$; we can go on until the degree has dropped to 1. (Here "=" means "is canonically isomorphic to". For $q = 1$ this line is empty.)
(3) The last step (3.53) with $p = q-1$.
(4) Definition of de Rham cohomology.

For $q = 0$ both sides are the locally constant functions: $H^0(M;\underline{\mathbb{R}}) = \underline{\mathbb{R}}(M)$, and $H^0_{\mathrm{dR}}(M) = \{f : \mathrm{d}f = 0\}$.

Key idea: The de Rham theorem (Nash's (3.23))

For every smooth manifold M and every $q \geq 0$,

$$H^q(M;\mathbb{R}) \;\cong\; H^q_{\mathrm{dR}}(M).$$

The mechanism: $\underline{\mathbb{R}}$ has a resolution by fine sheaves (the Poincaré lemma), fine sheaves have

no higher cohomology (partitions of unity), and the long exact sequence shifts degrees until everything is expressed through global sections, which are forms.

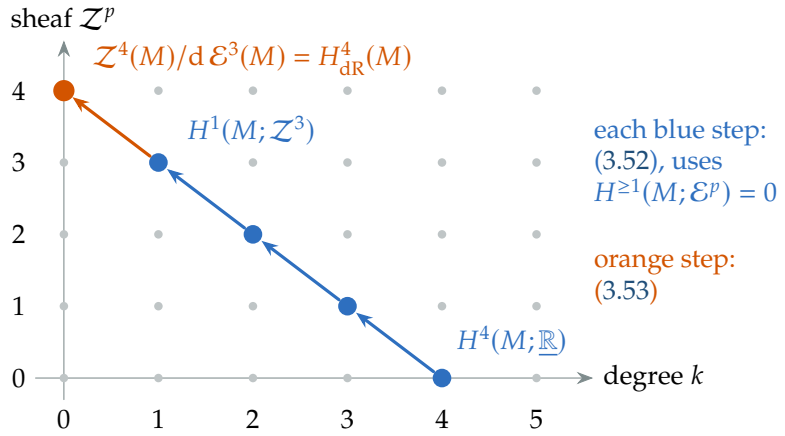


Figure 3.21: Nash's induction (3.54) for $q = 4$. Dots are the groups $H^k(M;\mathcal{Z}^p)$. Each blue arrow is an isomorphism that trades one unit of cohomological degree for one unit of form degree, staying on the antidiagonal $k + p = q$. The orange step, at $k = 1$, lands in global forms, where cohomology becomes "closed modulo exact".

Three remarks make the statement sharper.

- *Vanishing above the dimension.* For $q > n = \dim M$, the chain reaches $H^1(M;\mathcal{Z}^{q-1})$ with $q - 1 \geq n$. But $\mathcal{Z}^n = \mathcal{E}^n$ (every top form is closed), which is fine, and $\mathcal{Z}^p = 0$ for $p > n$. So $H^q(M;\mathbb{R}) = 0$ for $q > n$, as it must be.

- *Integers.* The argument uses real numbers essentially: there is no fine resolution of $\underline{\mathbb{Z}}$ by forms, since $\int$ of a form is a real number. The de Rham theorem computes $H^q(M;\mathbb{Z}) \otimes \mathbb{R}$ and loses torsion (for instance the $\mathbb{Z}_2$ in $H^2(\mathbb{RP}^3;\mathbb{Z})$). The integer-valued Čech cocycles of Section 3.5 are finer information than forms can hold.

- *Other resolutions.* The same argument works for any fine resolution of any sheaf. That is Nash's remark about the "alternative definition" via resolutions: whichever fine resolution you choose, the cohomology of its complex of global sections is $H^q(M;\mathcal{F})$. We use this again, with $\bar{\partial}$ in place of d, in Section 3.7.

Watch out: A garbled line in Nash

Just before (3.22) Nash writes "since $H^0(M;\mathcal{E}^p) \neq 0$, we cannot conclude that $H^1(M;\mathcal{Z}^{q-1}) \neq H^0(M;\mathcal{Z}^p)$". The intended statement is: *we cannot conclude that* $H^1(M;\mathcal{Z}^{q-1}) \cong H^0(M;\mathcal{Z}^q)$, because at $k = 0$ the left-hand term of (3.51) is the nonzero group of global forms. Instead one gets the quotient (3.53). Also, in (3.22) the p should be q: $H^1(M;\mathcal{Z}^{q-1}) \cong H^0(M;\mathcal{Z}^q)/\mathrm{d}H^0(M;\mathcal{E}^{q-1})$.

3.6.3 The isomorphism, made explicit on the circle

Abstract isomorphisms are hard to trust. Every map in (3.54) is a connecting homomorphism, i.e. a zig-zag (Figure 3.17), so we can run it by hand.

Worked example 3.25: From $\mathrm{d}\theta$ to a Čech cocycle, and back

Take $M = S^1$ with the two-arc cover $U_1 = (-\varepsilon, \pi + \varepsilon)$, $U_2 = (\pi - \varepsilon, 2\pi + \varepsilon)$, overlaps A near $\theta = 0$ and B near $\theta = \pi$ (Example 3.14). For $q = 1$, (3.54) is the single connecting map of $0 \to \underline{\mathbb{R}} \to \mathcal{E}^0 \to \mathcal{Z}^1 \to 0$:

$$\delta^* : \; H^1_{\mathrm{dR}}(S^1) = \frac{\mathcal{Z}^1(S^1)}{\mathrm{d}\,\mathcal{E}^0(S^1)} \xrightarrow{\;\cong\;} H^1(S^1; \underline{\mathbb{R}}).$$

Forward: from a form to numbers. Start with $\omega = \mathrm{d}\theta$.

(1) *Lift* to $\mathcal{E}^0$ on each arc: $f_1 = \theta_1$, the angle with values in $(-\varepsilon, \pi + \varepsilon)$, and $f_2 = \theta_2$, with values in $(\pi - \varepsilon, 2\pi + \varepsilon)$. Then $\mathrm{d}f_i = \omega$ on U_i.

(2) *Coboundary:* $(\delta f)_{12} = f_2 - f_1$ on $U_{12} = A \sqcup B$.

(3) *Read off:* the difference is locally constant, as it must be ($\mathrm{d}(f_2 - f_1) = \omega - \omega = 0$):

$$c_A \overset{(1)}{=} (\theta_2 - \theta_1)\big|_A = 2\pi, \qquad c_B \overset{(2)}{=} (\theta_2 - \theta_1)\big|_B = 0.$$

(1) Near $\theta = 0$, a point with $\theta_1 = \theta \in (-\varepsilon, \varepsilon)$ has $\theta_2 = \theta + 2\pi$.
(2) Near $\theta = \pi$ both arcs use the same value of the angle.

Under the isomorphism $[c] \mapsto c_A - c_B$ of Example 3.14, the class of $\mathrm{d}\theta$ goes to $2\pi = \oint \mathrm{d}\theta$. More generally, for any closed $\omega = g(\theta)\mathrm{d}\theta$, the jump $c_A - c_B$ of its local primitives is $\oint \omega$: *the Čech class of a closed* 1*-form is its period.*

Backward: from numbers to a form. Start with a cocycle (c_A, c_B) and the partition of unity $\rho_1, \rho_2 = 1 - \rho_1$ of Example 3.12. The homotopy operator (3.36) gives a 0-cochain of functions with $\delta f = c$:

$$f_1 \overset{(1)}{=} \rho_2\, c_{21} = -\rho_2\, c_{12}, \qquad f_2 \overset{(1)}{=} \rho_1\, c_{12},$$
$$f_2 - f_1 \overset{(2)}{=} (\rho_1 + \rho_2)\, c_{12} \overset{(3)}{=} c_{12} \quad \text{on } U_{12}.$$

(1) $(Kc)_i = \sum_k \rho_k c_{ki}$; the term $k = i$ vanishes because $c_{ii} = 0$; each product is extended by zero outside U_{12}.
(2) Subtract.
(3) $\rho_1 + \rho_2 = 1$.

The form $\omega = \mathrm{d}f_i$ is then globally defined, since $\mathrm{d}f_2 - \mathrm{d}f_1 = \mathrm{d}c_{12} = 0$. Explicitly, on U_2,

$$\omega = \mathrm{d}(\rho_1 c_{12}) = c_A\, \rho_1'(\theta)\, \mathrm{d}\theta\big|_A + c_B\, \rho_1'(\theta)\, \mathrm{d}\theta\big|_B, \tag{3.55}$$

a form supported in the overlaps. Integrating,

$$\oint \omega \overset{(1)}{=} c_A \int_A \rho_1'\, \mathrm{d}\theta + c_B \int_B \rho_1'\, \mathrm{d}\theta \overset{(2)}{=} c_A(1 - 0) + c_B(0 - 1) = c_A - c_B.$$

(1) (3.55); $\omega = 0$ outside $A \cup B$.
(2) ρ_1 rises from 0 to 1 across A and falls from 1 to 0 across B (Figure 3.10).

So the backward map sends (c_A, c_B) to a form with period $c_A - c_B$: it is the inverse of the forward map. For $(c_A, c_B) = (2\pi, 0)$ we get $\omega = 2\pi\rho_1'\, \mathrm{d}\theta$, a bump concentrated near $\theta = 0$ (Figure 3.22). It is *not* equal to $\mathrm{d}\theta$, but it is cohomologous to it: the difference has zero period, hence is d of a periodic function.

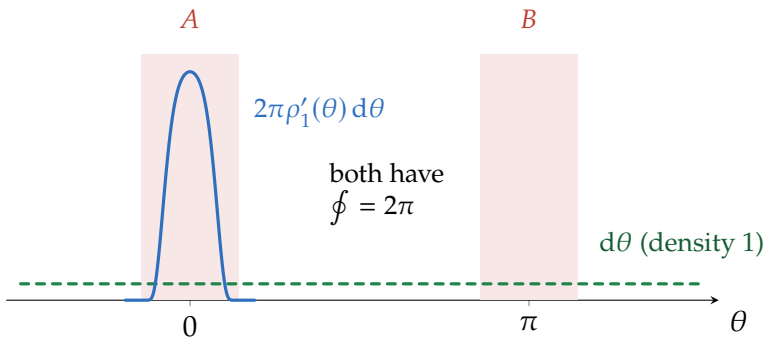


Figure 3.22: Two representatives of the same de Rham class on S^1. The uniform form $d\theta$ (dashed) and the bump form $2\pi\rho_1' d\theta$ (solid) that the partition of unity manufactures from the Čech cocycle $(c_A, c_B) = (2\pi, 0)$. All of the second form's "circulation" is squeezed into the overlap A, where the two local primitives disagree.

Worked example 3.26: The area form of S^2 by dimension shifting

For $q = 2$ on $M = S^2$ the chain (3.54) has two steps:

$$H^2(S^2;\underline{\mathbb{R}}) \xleftarrow{\delta^*} H^1(S^2;\mathcal{Z}^1) \xleftarrow{\delta^*} \frac{\mathcal{Z}^2(S^2)}{d\,\mathcal{E}^1(S^2)}.$$

Let us run the second (rightmost) map on $\omega = \sin\theta\, d\theta \wedge d\phi$, using the two caps U_N, U_S with the equatorial annulus U_{NS}.

Why two caps are enough here. We saw (Example 3.16) that two caps cannot compute $H^2(S^2;\underline{\mathbb{R}})$. But they *can* compute $H^1(S^2;\mathcal{Z}^1)$: by (3.52), $H^k(U;\mathcal{Z}^1) \cong H^{k+1}(U;\underline{\mathbb{R}})$ for $k \geq 1$, which vanishes on a cap and on an annulus (both have no cohomology above degree 1). So the caps form a Leray cover for $\mathcal{Z}^1$. Shifting the sheaf changes which covers are good.

The zig-zag.

(1) *Lift:* on each cap ω is exact (the caps are contractible). The monopole potentials of Example 3.22 with $g_m = 1$ do it: $\alpha_N = (1 - \cos\theta)\,d\phi$, $\alpha_S = -(1 + \cos\theta)\,d\phi$, $d\alpha_N = d\alpha_S = \omega$.

(2) *Coboundary:* $(\delta\alpha)_{NS} = \alpha_S - \alpha_N = -2\,d\phi$ on the annulus.

(3) *Read off:* $-2\,d\phi$ is a closed 1-form on U_{NS}, i.e. a 1-cochain of $\mathcal{Z}^1$, automatically a cocycle (no triple overlaps).

Which classes are there? On this cover $H^1(\mathcal{U};\mathcal{Z}^1) = \mathcal{Z}^1(U_{NS})/\{\beta_S - \beta_N\}$ with β_N, β_S closed on the caps. A closed form on a cap is exact, $\beta = dh$, and every function on the annulus is a difference $h_S - h_N$ (use a partition of unity, as in Example 3.25). So the coboundaries are exactly the exact 1-forms on the annulus, and

$$H^1(S^2;\mathcal{Z}^1) \cong \frac{\mathcal{Z}^1(U_{NS})}{d\,\mathcal{E}^0(U_{NS})} = H^1_{\mathrm{dR}}(\text{annulus}) \cong \mathbb{R}, \qquad [\beta] \mapsto -\oint_{\mathrm{eq}} \beta,$$

with the equator traversed in the direction of increasing ϕ. The class of ω goes to $-\oint(-2\,d\phi) = 4\pi$. Why the minus sign? Stokes on the two hemispheres:

$$\int_{S^2} \omega \overset{(1)}{=} \int_{\theta\leq\pi/2} d\alpha_N + \int_{\theta\geq\pi/2} d\alpha_S \overset{(2)}{=} \oint_{\mathrm{eq}} \alpha_N - \oint_{\mathrm{eq}} \alpha_S \overset{(3)}{=} -\oint_{\mathrm{eq}} (\delta\alpha)_{NS} = 4\pi.$$

(1) Split the sphere along the equator; on each half use the local primitive.

(2) Stokes. With the orientation $d\theta \wedge d\phi$, the boundary of the northern hemisphere is the equator traversed with ϕ increasing, and the boundary of the southern one is the same circle traversed backwards.
(3) $(\delta\alpha)_{NS} = \alpha_S - \alpha_N$; and $-\oint(-2\,d\phi) = 4\pi$.

In physics language: *the flux through the sphere equals the circulation of the difference of the two gauge potentials round the equator.* That is exactly the step "field strength → gauge transformation on the overlap". The first map of the chain, from $H^1(S^2;\mathcal{Z}^1)$ to $H^2(S^2;\underline{\mathbb{R}})$, requires a cover with triple overlaps, and we carry it out in Example 3.27.

3.6.4 The Čech–de Rham double complex

Nash's induction proves that the groups are isomorphic, but it hides the map in a sequence of connecting homomorphisms. There is a way to see all the steps at once, due to Weil and popularised by Bott and Tu: put Čech cochains and differential forms into one grid.

Working definition 3.17 (Čech–de Rham double complex)

Fix a good cover $\mathcal{U}$ of M [every set and every nonempty overlap is contractible]. Let

$$K^{p,q} = C^p(\mathcal{U};\mathcal{E}^q) = \left\{\text{a } q\text{-form } \omega_{i_0\cdots i_p} \text{ on each } (p+1)\text{-fold overlap } U_{i_0\cdots i_p}\right\},$$

with two commuting differentials [maps that each give 0 when applied twice]: the Čech coboundary $\delta : K^{p,q} \to K^{p+1,q}$ (alternating sum of restrictions, applied to forms) and the exterior derivative $d : K^{p,q} \to K^{p,q+1}$ (applied on each overlap separately). They commute, $\delta d = d\delta$, because d commutes with restriction. *Operational picture:* a grid whose horizontal direction is "number of patches" and whose vertical direction is "form degree" (Figure 3.23).

Two facts make the grid useful. Both are things we have already proved.

- *The rows are exact* (except at the left edge). Fix a form degree q. The row

$$0 \to \mathcal{E}^q(M) \xrightarrow{r} C^0(\mathcal{U};\mathcal{E}^q) \xrightarrow{\delta} C^1(\mathcal{U};\mathcal{E}^q) \xrightarrow{\delta} C^2(\mathcal{U};\mathcal{E}^q) \to \cdots \tag{3.56}$$

  (with r restricting a global form to each patch) is exact. Exactness at C^0 is the gluing axiom; exactness further right is the homotopy formula $\delta K + K\delta = 1$ of Section 3.3.7, since $\mathcal{E}^q$ is fine. *In words: any Čech cocycle of forms is a Čech coboundary of forms.*

- *The columns are exact* (except at the bottom). Fix p. The column

$$0 \to C^p(\mathcal{U};\underline{\mathbb{R}}) \xrightarrow{\iota} C^p(\mathcal{U};\mathcal{E}^0) \xrightarrow{d} C^p(\mathcal{U};\mathcal{E}^1) \xrightarrow{d} C^p(\mathcal{U};\mathcal{E}^2) \to \cdots \tag{3.57}$$

  is exact, because every overlap of a good cover is contractible and the Poincaré lemma holds there. *In words: on each overlap, a closed form is exact, and a closed function is a constant.*

So the grid has two "edges": the left edge holds global forms $\mathcal{E}^q(M)$ (the de Rham complex), the bottom edge holds Čech cochains with constant coefficients $C^p(\mathcal{U};\underline{\mathbb{R}})$ (the Čech complex). The zig-zag through the interior connects them.

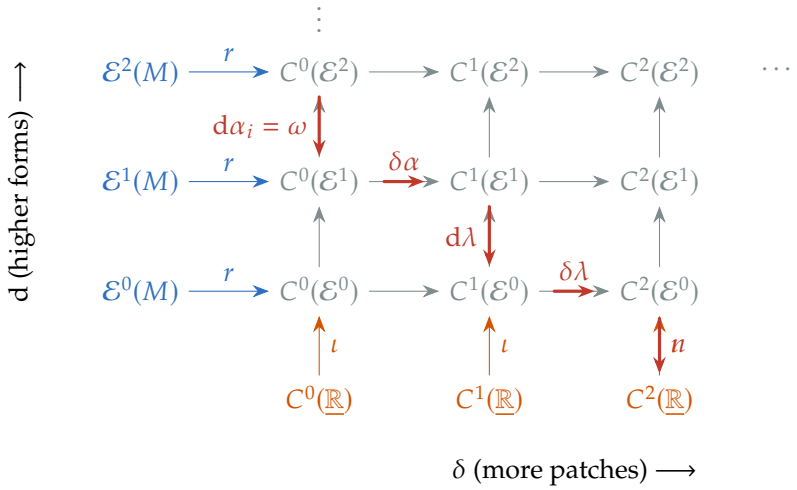


Figure 3.23: The Čech–de Rham double complex $K^{p,q} = C^p(\mathcal{U}; \mathcal{E}^q)$. The left edge (blue) is the de Rham complex of global forms; the bottom edge (orange) is the Čech complex with constant coefficients. The red staircase is the zig-zag (3.58) for a closed 2-form ω: primitives α_i on patches, their mismatch $\delta\alpha = \mathrm{d}\lambda$ on double overlaps, and the mismatch $\delta\lambda = n$ of those, which is a constant on triple overlaps. Every step moves one square down or one square right.

Working definition 3.18 (the zig-zag from a closed form to a Čech cocycle)

Let ω be a closed 2-form on M and $\mathcal{U}$ a good cover.

(1) On each U_i: $\omega = \mathrm{d}\alpha_i$ (Poincaré lemma [on a contractible set every closed form is exact]).

(2) On each U_{ij}: $\mathrm{d}(\alpha_j - \alpha_i) = \omega - \omega = 0$, so $\alpha_j - \alpha_i = \mathrm{d}\lambda_{ij}$ for a function λ_{ij} (Poincaré lemma again), chosen with $\lambda_{ji} = -\lambda_{ij}$.

(3) On each U_{ijk}: $n_{ijk} = (\delta\lambda)_{ijk} = \lambda_{jk} - \lambda_{ik} + \lambda_{ij}$ has $\mathrm{d}n_{ijk} = 0$, so it is a constant.

In one line:

$$\omega = \mathrm{d}\alpha, \qquad \delta\alpha = \mathrm{d}\lambda, \qquad \delta\lambda = n \in C^2(\mathcal{U}; \underline{\mathbb{R}}). \tag{3.58}$$

The class $[n] \in H^2(\mathcal{U}; \underline{\mathbb{R}})$ is the Čech image [the Čech class matched with $[\omega]$ by the de Rham isomorphism] of $[\omega] \in H^2_{\mathrm{dR}}(M)$. For a closed 1-form the recipe stops one step earlier ($\omega = \mathrm{d}f$, $n = \delta f$), which is Example 3.25; for a closed q-form it takes q steps.

Check that step (3) is really a constant:

$$\mathrm{d}n \overset{(1)}{=} \mathrm{d}\,\delta\lambda \overset{(2)}{=} \delta\,\mathrm{d}\lambda \overset{(3)}{=} \delta\delta\alpha \overset{(4)}{=} 0.$$

(1) Definition of n.
(2) d and δ commute.
(3) Step (2): $\mathrm{d}\lambda = \delta\alpha$.
(4) $\delta^2 = 0$ (Section 3.3.2).

And n is a Čech cocycle because $\delta n = \delta\delta\lambda = 0$.

Why the class does not depend on the choices. Each choice is made "up to something one step further along":

- $\alpha_i \mapsto \alpha_i + \mathrm{d}\mu_i$ changes $\delta\alpha$ by $\mathrm{d}\delta\mu$, which we absorb by $\lambda \mapsto \lambda + \delta\mu$; then $n \mapsto n + \delta\delta\mu = n$.

- $\lambda_{ij} \mapsto \lambda_{ij} + m_{ij}$ with constants m_{ij} (the only freedom once $\mathrm{d}\lambda$ is fixed) changes n by δm, a Čech coboundary.
- $\omega \mapsto \omega + \mathrm{d}\beta$ with global β: take $\alpha_i + \beta$; then $\delta\alpha$ is unchanged.

The inverse map, by partitions of unity. Start instead with a Čech 2-cocycle n of constants and a partition of unity $\{\rho_k\}$. Walk the staircase the other way, each time using row exactness in the form "a Čech cocycle of forms is δ of K of itself":

$$\lambda \overset{(1)}{=} Kn, \quad \text{i.e.} \quad \lambda_{ij} = \sum_k \rho_k\, n_{kij}, \qquad \text{so } \delta\lambda = n;$$

$$\alpha \overset{(2)}{=} K(\mathrm{d}\lambda), \quad \text{i.e.} \quad \alpha_i = \sum_k \rho_k\, \mathrm{d}\lambda_{ki} = \sum_{k,l} n_{lki}\, \rho_k\, \mathrm{d}\rho_l, \qquad \text{so } \delta\alpha = \mathrm{d}\lambda;$$

$$\omega \overset{(3)}{=} \mathrm{d}\alpha_i = \sum_{k,l} n_{lki}\, \mathrm{d}\rho_k \wedge \mathrm{d}\rho_l \qquad \text{on } U_i.$$

(1) Homotopy formula: $\delta Kn = n - K\delta n = n$ because $\delta n = 0$.
(2) $\mathrm{d}\lambda$ is a Čech 1-cocycle of 1-forms ($\delta\mathrm{d}\lambda = \mathrm{d}\delta\lambda = \mathrm{d}n = 0$, n constant), so again $\delta K(\mathrm{d}\lambda) = \mathrm{d}\lambda$. Then substitute $\lambda_{ki} = \sum_l \rho_l n_{lki}$ and use $\mathrm{d}n = 0$.
(3) $\delta(\mathrm{d}\alpha) = \mathrm{d}\delta\alpha = \mathrm{d}\mathrm{d}\lambda = 0$, so the $\mathrm{d}\alpha_i$ agree on overlaps and glue to a global closed 2-form.

Running the forward recipe on this ω with these very α and λ returns n, so the two maps are inverse to each other. The formula $\omega = \sum n_{lki}\, \mathrm{d}\rho_k \wedge \mathrm{d}\rho_l$ is worth a look: it is a closed 2-form supported where three partition functions change at once, i.e. near the triple overlaps, with "strength" given by the integers n_{lki}. It is the 2-dimensional version of the bump form of Figure 3.22.

Key idea: De Rham theorem, Čech–de Rham version

On a good cover $\mathcal{U}$, the zig-zag (3.58) gives an isomorphism $H^q_{\mathrm{dR}}(M) \cong H^q(\mathcal{U}; \underline{\mathbb{R}})$, whose inverse is built from a partition of unity. Rows exact = fine sheaves; columns exact = Poincaré lemma. This is the same theorem as Nash's (3.23), with every map written out.

3.6.5 The monopole through the double complex

We now close the loop with Section 3.5. There we defined c_1 of a line bundle by taking logarithms of transition functions. Here we start from the field strength and walk down the staircase. The two must agree, and checking that they do fixes all the signs.

Conventions for $U(1)$ connections

A connection on a line bundle with transition functions g_{ij} (so that local components obey $s_i = g_{ij} s_j$) is a collection of 1-forms A_i, one per patch, with covariant derivative $Ds_i = \mathrm{d}s_i + A_i s_i$. Following Nash and the house convention, A_i is *anti-Hermitian*: for $U(1)$, an imaginary-valued 1-form. The covariant derivative of a section must transform like the section, $Ds_i = g_{ij} Ds_j$. Then

$$\mathrm{d}s_i + A_i s_i \overset{(1)}{=} \mathrm{d}(g_{ij} s_j) + A_i\, g_{ij} s_j \overset{(2)}{=} g_{ij}\big(\mathrm{d}s_j + (g_{ij}^{-1}\mathrm{d}g_{ij} + A_i)\, s_j\big),$$

(1) Substitute $s_i = g_{ij} s_j$.
(2) Leibniz rule $\mathrm{d}(gs) = g\,\mathrm{d}s + \mathrm{d}g\, s$; factor out g_{ij}.

and comparing with $g_{ij}(\mathrm{d}s_j + A_j s_j)$ gives

$$A_j = A_i + g_{ij}^{-1}\mathrm{d}g_{ij} = A_i + \mathrm{d}\log g_{ij} \qquad \text{on } U_{ij}. \tag{3.59}$$

The curvature $F = \mathrm{d}A_i$ is the same on every patch (abelian case: $\mathrm{d}\,\mathrm{d}\log g = 0$), so it is a global closed 2-form. Its real normalisation is the **first Chern form**

$$\omega = \frac{i}{2\pi}F, \qquad a_i = \frac{i}{2\pi}A_i \text{ (real 1-forms)}, \qquad \omega = \mathrm{d}a_i. \tag{3.60}$$

Dictionary: Mathematicians' versus physicists' $U(1)$ potential

For a particle of charge e the physicist writes $D = \mathrm{d} - ieA^{\mathrm{phys}}$ with real A^{phys}, so $A = -ieA^{\mathrm{phys}}$ and $F = -ieF^{\mathrm{phys}}$. Then

$$a_i = \frac{e}{2\pi}A_i^{\mathrm{phys}}, \qquad \omega = \frac{i}{2\pi}F = \frac{e}{2\pi}F^{\mathrm{phys}}.$$

So $\int \omega$ is the magnetic flux measured in units of the flux quantum $2\pi/e$ (with $\hbar = 1$).

Now multiply (3.59) by $\frac{i}{2\pi}$:

$$a_j - a_i \overset{(1)}{=} \frac{i}{2\pi}\,\mathrm{d}\log g_{ij} \overset{(2)}{=} \mathrm{d}\lambda_{ij}, \qquad \lambda_{ij} = \frac{i}{2\pi}\log g_{ij}.$$

(1) (3.59) times $\frac{i}{2\pi}$, with (3.60).
(2) Definition (3.41) of λ_{ij} in Section 3.5.2.

So *the zig-zag* (3.58) *of the Chern form, with the gauge potentials as the first lift, produces exactly the λ_{ij} of Section 3.5.2, and hence exactly the Chern cocycle $n = \delta\lambda$.* This proves the promised normalisation: the image of $c_1(L) \in H^2(M;\mathbb{Z})$ in $H^2(M;\mathbb{R}) \cong H^2_{\mathrm{dR}}(M)$ is the class of $\frac{i}{2\pi}F$. In particular, for any closed surface $\Sigma \subset M$,

$$\int_\Sigma \frac{i}{2\pi}F \in \mathbb{Z}, \tag{3.61}$$

whatever connection we choose. This is *flux quantisation* in its general form.

Worked example 3.27: The monopole's Chern form, walked down the staircase

Take the monopole bundle with $n = 2e g_m$ and the three-set cover V_1, V_2, V_3 of Example 3.23 (Figure 3.19). By the dictionary,

$$\omega \overset{(1)}{=} \frac{e}{2\pi}\,g_m \sin\theta\,\mathrm{d}\theta\wedge\mathrm{d}\phi = \frac{n}{4\pi}\sin\theta\,\mathrm{d}\theta\wedge\mathrm{d}\phi,$$
$$a_N \overset{(2)}{=} \frac{n}{4\pi}(1-\cos\theta)\,\mathrm{d}\phi, \qquad a_S = -\frac{n}{4\pi}(1+\cos\theta)\,\mathrm{d}\phi.$$

(1) $\omega = \frac{e}{2\pi}F^{\mathrm{phys}}$ with (3.44); $e g_m = n/2$.
(2) $a = \frac{e}{2\pi}A^{\mathrm{phys}}$ with (3.45).

Step (1), potentials on patches: $a_1 = a_N$ on V_1 and $a_2 = a_3 = a_S$ on V_2, V_3 (both southern sets use the south gauge).
Step (2), gauge functions on double overlaps.

$$a_2 - a_1 \overset{(1)}{=} -\frac{n}{4\pi}\big[(1+\cos\theta) + (1-\cos\theta)\big]\mathrm{d}\phi = -\frac{n}{2\pi}\,\mathrm{d}\phi \overset{(2)}{=} \mathrm{d}\lambda_{12}, \qquad \lambda_{12} = -\frac{n\phi}{2\pi},$$

(1) Subtract a_N from a_S.
(2) On the half-annulus V_{12}, ϕ is a smooth function with values in $(-\varepsilon, \pi+\varepsilon)$.

and in the same way $\lambda_{13} = -n\phi/2\pi$ on V_{13} with $\phi \in (\pi-\varepsilon, 2\pi+\varepsilon)$, while $a_3 - a_2 = 0$ gives $\lambda_{23} = 0$. These are the logarithms of $g_{12} = g_{13} = e^{in\phi}$, $g_{23} = 1$ found in Example 3.23.

Step (3), integers on the triple overlaps. As computed there, $n_{123} = n$ on the square near $\phi = 0$ and $n_{123} = 0$ on the square near $\phi = \pi$. Write P_0, P_π for points in these two squares.

The check against integration. We now show that for *any* zig-zag data $(\omega, a_i, \lambda_{ij}, n_{123})$ on this cover, $\int_{S^2}\omega = n_{123}(P_0) - n_{123}(P_\pi) = \ell(n)$, which confirms the sign of ℓ chosen in Example 3.23. Cut the sphere into three closed regions (Figure 3.24): D_1 the northern hemisphere, D_2 the southern quarter-sphere with $0 \le \phi \le \pi$, D_3 the one with $\pi \le \phi \le 2\pi$, so $D_i \subset V_i$. Their common edges are the equatorial arcs E_2 ($D_1|D_2$, from P_0 to P_π) and E_3 ($D_1|D_3$, from P_π round to P_0), both traversed with ϕ increasing, and the meridian arcs M_0 and M_π ($D_2|D_3$, at $\phi = 0$ and $\phi = \pi$), both traversed from the equator down to the south pole S. With the orientation $\mathrm{d}\theta \wedge \mathrm{d}\phi$, i.e. counter-clockwise in the (θ, ϕ) rectangle, one reads off

$$\partial D_1 = E_2 + E_3, \qquad \partial D_2 = M_0 - M_\pi - E_2, \qquad \partial D_3 = M_\pi - M_0 - E_3.$$

Then

$$\begin{aligned}
\int_{S^2}\omega &\overset{(1)}{=} \sum_{i=1}^{3}\int_{D_i}\mathrm{d}a_i \overset{(2)}{=} \sum_{i=1}^{3}\oint_{\partial D_i} a_i \\
&\overset{(3)}{=} \int_{E_2}(a_1 - a_2) + \int_{E_3}(a_1 - a_3) + \int_{M_0}(a_2 - a_3) + \int_{M_\pi}(a_3 - a_2) \\
&\overset{(4)}{=} -\int_{E_2}\mathrm{d}\lambda_{12} - \int_{E_3}\mathrm{d}\lambda_{13} - \int_{M_0}\mathrm{d}\lambda_{23} + \int_{M_\pi}\mathrm{d}\lambda_{23} \\
&\overset{(5)}{=} -\big[\lambda_{12}(P_\pi) - \lambda_{12}(P_0)\big] - \big[\lambda_{13}(P_0) - \lambda_{13}(P_\pi)\big] \\
&\qquad - \big[\lambda_{23}(S) - \lambda_{23}(P_0)\big] + \big[\lambda_{23}(S) - \lambda_{23}(P_\pi)\big] \\
&\overset{(6)}{=} \big(\lambda_{23} - \lambda_{13} + \lambda_{12}\big)(P_0) - \big(\lambda_{23} - \lambda_{13} + \lambda_{12}\big)(P_\pi) \\
&\overset{(7)}{=} n_{123}(P_0) - n_{123}(P_\pi).
\end{aligned}$$

(1) Split the integral over the three regions and use $\omega = \mathrm{d}a_i$ on $D_i \subset V_i$.
(2) Stokes' theorem on each region.
(3) Insert the boundaries listed above and collect the terms on each common edge.
(4) $a_j - a_i = \mathrm{d}\lambda_{ij}$, so $a_1 - a_2 = -\mathrm{d}\lambda_{12}$ etc.
(5) Fundamental theorem of calculus along each oriented arc: end minus start. E_3 runs from P_π to P_0 inside V_{13}, where λ_{13} is a single smooth function.
(6) The two $\lambda_{23}(S)$ terms cancel; regroup the rest point by point.
(7) Definition $n = \delta\lambda$.

For the monopole this gives $n - 0 = n$, and directly $\int\omega = \frac{n}{4\pi}\cdot 4\pi = n$. The two computations agree, and the Chern number of the monopole bundle is the flux in units of $2\pi/e$:

$$c_1 = \int_{S^2}\frac{e}{2\pi}F^{\mathrm{phys}} = \frac{e}{2\pi}\cdot 4\pi g_m = 2e g_m = n.$$

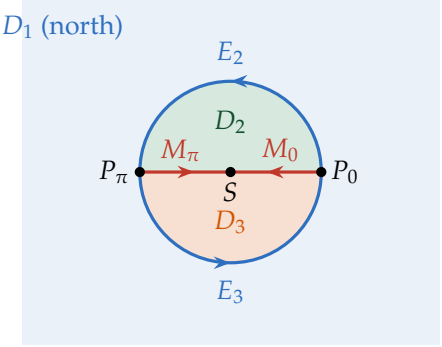


Figure 3.24: The three regions used to integrate the Chern form, drawn as in Figure 3.19 (south pole at the centre, equator the circle). Stokes' theorem turns $\int \omega$ into line integrals of the mismatches $a_j - a_i = \mathrm{d}\lambda_{ij}$ along the edges; the fundamental theorem of calculus then turns those into values of λ at the vertices P_0, P_π, S. The contributions at S cancel, and what survives at P_0 and P_π is exactly the Čech cocycle n_{123}.

Physical picture: Stokes, three times

The integration check is Stokes' theorem applied in every dimension of the staircase: a 2-dimensional integral of $\mathrm{d}a$ becomes 1-dimensional integrals of $a_j - a_i = \mathrm{d}\lambda$, which become 0-dimensional evaluations of $\delta\lambda = n$. The double complex is a bookkeeping device for doing Stokes' theorem "patch by patch". The same computation on any surface with any good cover gives

$$\int_\Sigma \omega = \sum_{\text{vertices } v} \pm\, n_{ijk}(v),$$

a signed count of the integers at the triple points of a triangulation dual to the cover. The integer c_1 is literally a sum of integers.

Reading Nash (pp. 63, (3.20)–(3.23))

Nash's (3.20) is our (3.50): the short exact sequences $0 \to \mathcal{Z}^p \to \mathcal{E}^p \xrightarrow{\mathrm{d}} \mathcal{Z}^{p+1} \to 0$, with $\mathcal{Z}^0 = \mathbb{R}$, each followed by the isomorphism $H^{q-1}(M;\mathcal{Z}^{p+1}) \cong H^q(M;\mathcal{Z}^p)$ it produces (our (3.52)). His (3.21) is the chain (3.54); his (3.22) is (3.53) (read the p there as q, see the pitfall); his (3.23) is the de Rham theorem. He uses exactly two facts: the long exact sequence (3.19) and the softness of $\mathcal{E}^p$. We have added the explicit form of the maps (Example 3.25, Example 3.26), the Čech–de Rham double complex, and the monopole check that ties Section 3.5 to the curvature: $c_1 \mapsto [\frac{i}{2\pi}F]$.

Recap

- $0 \to \underline{\mathbb{R}} \to \mathcal{E}^0 \xrightarrow{\mathrm{d}} \mathcal{E}^1 \xrightarrow{\mathrm{d}} \cdots$ is a fine resolution of $\underline{\mathbb{R}}$. Exactness is the Poincaré lemma (explicit radial homotopy, e.g. $\vec{A} = \int_0^1 t\, \vec{B}(t\vec{x}) \times \vec{x}\, \mathrm{d}t$).
- Chop it into $0 \to \mathcal{Z}^p \to \mathcal{E}^p \to \mathcal{Z}^{p+1} \to 0$; fine middle terms give $H^{k+1}(\mathcal{Z}^p) \cong H^k(\mathcal{Z}^{p+1})$ for $k \geq 1$ and $H^1(\mathcal{Z}^p) \cong H^{p+1}_{\mathrm{dR}}$; chaining gives $H^q(M;\mathbb{R}) \cong H^q_{\mathrm{dR}}(M)$.
- The isomorphism is a zig-zag: $\omega = \mathrm{d}\alpha$, $\delta\alpha = \mathrm{d}\lambda$, $\delta\lambda = n$. On S^1 the Čech class of ω is its period; the inverse uses a partition of unity and produces bump forms.
- For a $U(1)$ bundle, the zig-zag of $\frac{i}{2\pi}F$ through the gauge potentials reproduces $n = \delta\lambda$,

$\lambda = \frac{i}{2\pi}\log g$: the de Rham image of c_1 is $[\frac{i}{2\pi}F]$, and $\int_\Sigma \frac{i}{2\pi}F \in \mathbb{Z}$.

- Monopole: Stokes three times gives $\int \omega = n_{123}(P_0) - n_{123}(P_\pi) = n = 2eg_m$.

3.7 Complex manifolds and the Dolbeault theorem

The de Rham theorem answered a question about the sheaf $\underline{\mathbb{R}}$ of locally constant functions: its Čech cohomology can be computed by solving the differential equation $\mathrm{d}\alpha = \omega$. On a complex manifold there is a second sheaf that we care about just as much, the sheaf $\mathcal{O}$ of holomorphic functions (and its cousins Ω^p, the holomorphic p-forms). We have already met its cohomology several times:

- $H^1(M;\mathcal{O})$ sat in the exponential sequence next to $H^1(M;\mathcal{O}^*)$, the group of holomorphic line bundles (Section 3.5).
- $H^1(\mathbb{CP}^1;\mathcal{O}(n))$ was computed with Laurent series and produced the first index, $\dim H^0 - \dim H^1 = n+1$ (Example 3.18).
- Several times we *assumed* that discs, $\mathbb{C}$ and $\mathbb{C}^*$ have no higher $\mathcal{O}$-cohomology, so that covers by such sets are Leray covers.

The question of this section is: *is there a differential equation whose solvability computes $H^q(M;\mathcal{O})$, the way $\mathrm{d}\alpha = \omega$ computes $H^q(M;\mathbb{R})$?* The answer is yes. The equation is the Cauchy–Riemann equation in its inhomogeneous form,

$$\frac{\partial f}{\partial \bar z} = g,$$

and the resulting statement is the **Dolbeault theorem**. Nash obtains it by rerunning the de Rham argument word for word with $\bar\partial$ in place of d. To make sense of that we first need the objects $\bar\partial$ acts on, the forms of type (p,q).

Physical picture: Why a physicist should care

In two dimensions, written in the complex coordinate $z = x + iy$, the Euclidean Dirac operator splits into ∂_z and $\partial_{\bar z}$ (we show this explicitly below). Zero modes of $\partial_{\bar z}$ are holomorphic functions; the obstruction to solving $\partial_{\bar z} f = g$ is a count of zero modes of the adjoint. So Dolbeault cohomology counts *chiral zero modes*: of fermions on a Riemann surface, of b,c ghosts on a string worldsheet (Chapter 9), of the operators whose index the Riemann–Roch and Atiyah–Singer theorems compute (Chapter 4). It also counts *moduli*: $H^1(M;\mathcal{O})$ is the space of deformations of a holomorphic line bundle, which on a torus are the Wilson lines of a flat $U(1)$ field.

3.7.1 Complex coordinates and Wirtinger derivatives

Start in one complex dimension, on an open set of $\mathbb{C}$ with coordinate $z = x + iy$. Instead of the real basis $\mathrm{d}x, \mathrm{d}y$ of 1-forms we use the complex basis

$$\mathrm{d}z = \mathrm{d}x + i\,\mathrm{d}y, \qquad \mathrm{d}\bar z = \mathrm{d}x - i\,\mathrm{d}y, \qquad \text{so} \qquad \mathrm{d}x = \tfrac{1}{2}(\mathrm{d}z + \mathrm{d}\bar z), \quad \mathrm{d}y = \tfrac{1}{2i}(\mathrm{d}z - \mathrm{d}\bar z). \tag{3.62}$$

This is only a change of basis of the complexified cotangent space $T^*_x \otimes \mathbb{C} \cong \mathbb{C}^2$; nothing has been lost. The dual basis of (complexified) vectors is given by the **Wirtinger derivatives**

$$\frac{\partial}{\partial z} = \frac{1}{2}\Big(\frac{\partial}{\partial x} - i\frac{\partial}{\partial y}\Big), \qquad \frac{\partial}{\partial \bar z} = \frac{1}{2}\Big(\frac{\partial}{\partial x} + i\frac{\partial}{\partial y}\Big). \tag{3.63}$$

They are defined so that $\mathrm{d}z(\partial_z) = 1$, $\mathrm{d}z(\partial_{\bar z}) = 0$, and so on. For example $\mathrm{d}z(\partial_{\bar z}) = \tfrac12(\mathrm{d}x + i\,\mathrm{d}y)(\partial_x + i\partial_y) = \tfrac12(1 + i\cdot i) = 0$. With them the total differential of any smooth complex function $f(x,y)$ reads

$$\begin{aligned} \mathrm{d}f &\overset{(1)}{=} \frac{\partial f}{\partial x}\,\mathrm{d}x + \frac{\partial f}{\partial y}\,\mathrm{d}y \\ &\overset{(2)}{=} \frac{\partial f}{\partial x}\,\tfrac12(\mathrm{d}z + \mathrm{d}\bar z) + \frac{\partial f}{\partial y}\,\tfrac{1}{2i}(\mathrm{d}z - \mathrm{d}\bar z) \\ &\overset{(3)}{=} \tfrac12\Big(\frac{\partial f}{\partial x} - i\frac{\partial f}{\partial y}\Big)\mathrm{d}z + \tfrac12\Big(\frac{\partial f}{\partial x} + i\frac{\partial f}{\partial y}\Big)\mathrm{d}\bar z \\ &\overset{(4)}{=} \frac{\partial f}{\partial z}\,\mathrm{d}z + \frac{\partial f}{\partial \bar z}\,\mathrm{d}\bar z. \end{aligned} \tag{3.64}$$

(1) The usual total differential in real coordinates.
(2) Substitute (3.62) for $\mathrm{d}x, \mathrm{d}y$.
(3) Collect the $\mathrm{d}z$ and $\mathrm{d}\bar z$ terms; $1/(2i) = -i/2$.
(4) The definitions (3.63).

So we may treat z and $\bar z$ "as if they were independent variables". This is a bookkeeping rule, not a claim that they are independent: it works because $\partial_z z = 1$, $\partial_z \bar z = 0$, $\partial_{\bar z} z = 0$, $\partial_{\bar z}\bar z = 1$, which you can check from (3.63).

Working definition 3.19 (holomorphic function)

A smooth function f on an open set of $\mathbb{C}$ is **holomorphic** when

$$\frac{\partial f}{\partial \bar z} = 0,$$

[$\partial_{\bar z} f = \tfrac12(\partial_x f + i\partial_y f)$, the Wirtinger derivative defined just above] i.e. when $\mathrm{d}f$ has no $\mathrm{d}\bar z$ component. *Operational test:* write f as a function of z and $\bar z$ and check that $\bar z$ does not appear. In n complex variables f is holomorphic when $\partial f/\partial \bar z^i = 0$ for $i = 1,\dots,n$.

The condition $\partial_{\bar z} f = 0$ is the Cauchy–Riemann equations in disguise. Write $f = u + iv$ with u, v real:

$$\begin{aligned} 2\,\frac{\partial f}{\partial \bar z} &\overset{(1)}{=} \Big(\frac{\partial}{\partial x} + i\frac{\partial}{\partial y}\Big)(u + iv) \\ &\overset{(2)}{=} \Big(\frac{\partial u}{\partial x} - \frac{\partial v}{\partial y}\Big) + i\Big(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\Big). \end{aligned}$$

(1) Definition (3.63).
(2) Multiply out; $i\cdot i\,\partial_y v = -\partial_y v$.

The real and imaginary parts vanish exactly when $u_x = v_y$ and $u_y = -v_x$, the Cauchy–Riemann equations. This is Nash's remark after (3.25): "the familiar Cauchy–Riemann equations for a holomorphic function f are simply the statement $\bar\partial f = 0$".

Worked example 3.28: Which functions are holomorphic?

Treat $z, \bar z$ as independent and differentiate.

- $f = z^2$: $\partial_{\bar z} f = 0$. Holomorphic. Indeed $u = x^2 - y^2$, $v = 2xy$ satisfy $u_x = 2x = v_y$ and $u_y = -2y = -v_x$.

- $f = |z|^2 = z\bar{z}$: $\partial_{\bar{z}} f = z$, which is zero only at the origin. Not holomorphic.
- $f = \bar{z}$: $\partial_{\bar{z}} f = 1$. Not holomorphic; it is *antiholomorphic* ($\partial_z f = 0$).
- $f = \log|z|^2 = \log z + \log \bar{z}$ on $\mathbb{C}^*$: $\partial_z f = 1/z$, $\partial_{\bar{z}} f = 1/\bar{z}$. Neither.

Also the Laplacian factorises:

$$4\,\frac{\partial}{\partial z}\frac{\partial}{\partial \bar{z}} \overset{(1)}{=} \Big(\frac{\partial}{\partial x} - i\frac{\partial}{\partial y}\Big)\Big(\frac{\partial}{\partial x} + i\frac{\partial}{\partial y}\Big) \overset{(2)}{=} \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} = \nabla^2.$$

(1) Definition (3.63), twice.
(2) The cross terms $i\partial_x\partial_y - i\partial_y\partial_x$ cancel, and $-i \cdot i = 1$.

So $\partial_z\partial_{\bar{z}} \log|z|^2 = \partial_z(1/\bar{z}) = 0$ away from 0: $\log|z|^2$ is harmonic, the 2d Coulomb potential. We shall see in Section 3.7.4 that the charge at the origin is also visible in this language.

3.7.2 Complex manifolds and forms of type (p, q)

A **complex manifold** of complex dimension n is a manifold covered by charts with coordinates $z^1, \dots, z^n$ in open sets of $\mathbb{C}^n$, such that on every overlap the new coordinates z'^j are *holomorphic* functions of the old ones. As a real manifold it has dimension $2n$. Examples: any open subset of $\mathbb{C}^n$; the Riemann sphere $\mathbb{CP}^1$ with charts z and $w = 1/z$ (and w is holomorphic in z on $\mathbb{C}^*$); a torus $\mathbb{C}/\Lambda$ for a lattice Λ, with transition maps $z \mapsto z + \lambda$; $\mathbb{CP}^n$ with its standard charts; every smooth complex algebraic curve.

On such a manifold the complexified 1-forms have the local basis $\mathrm{d}z^1, \dots, \mathrm{d}z^n, \mathrm{d}\bar{z}^1, \dots, \mathrm{d}\bar{z}^n$. Nash's (3.24) is the general monomial built from them:

$$\omega = \frac{1}{p!\,q!}\,\omega_{i_1\cdots i_p \bar{j}_1\cdots\bar{j}_q}\,\mathrm{d}z^{i_1} \wedge \cdots \wedge \mathrm{d}z^{i_p} \wedge \mathrm{d}\bar{z}^{j_1} \wedge \cdots \wedge \mathrm{d}\bar{z}^{j_q}, \tag{3.65}$$

summed over indices, with coefficients antisymmetric in the i's and in the j's separately (the factor $1/(p!\,q!)$ is our usual normalisation of forms; Nash omits it). Such an ω is a $(p+q)$-form; we say it has **type** (p, q): p holomorphic differentials and q antiholomorphic ones.

Working definition 3.20 (type of a form)

To find the type decomposition of a complex k-form [a k-form with complex coefficients], rewrite it in the basis $\mathrm{d}z^i, \mathrm{d}\bar{z}^j$ (3.62) and sort the terms by how many $\mathrm{d}\bar{z}$'s they contain. The piece with q of them is the $(k-q, q)$ component. Formally,

$$\Lambda^k T^*M \otimes \mathbb{C} = \bigoplus_{p+q=k} \Lambda^{p,q} T^*M, \tag{3.66}$$

[$\Lambda^{p,q}T^*M$ is spanned by the monomials with exactly p factors $\mathrm{d}z^i$ and q factors $\mathrm{d}\bar{z}^j$] and $\mathcal{E}^{p,q}$ denotes the sheaf of smooth forms of type (p, q) [its sections over U are the smooth (p, q)-forms on U]. This is Nash's (3.27).

Why is "type" meaningful, i.e. independent of the chart? Because the transition functions are holomorphic. Under $z \mapsto z'(z)$,

$$\mathrm{d}z'^j \overset{(1)}{=} \sum_i \frac{\partial z'^j}{\partial z^i}\,\mathrm{d}z^i + \sum_i \frac{\partial z'^j}{\partial \bar{z}^i}\,\mathrm{d}\bar{z}^i \overset{(2)}{=} \sum_i \frac{\partial z'^j}{\partial z^i}\,\mathrm{d}z^i,$$

(1) The chain rule (3.64), in n variables.
(2) z'^j is a holomorphic function of the z^i.

So dz''s are combinations of dz's only, and (conjugating) $d\bar{z}'$'s of $d\bar{z}$'s only. A (p,q)-form in one chart is a (p,q)-form in every chart. On $\mathbb{CP}^1$ with $w = 1/z$: $dw = -z^{-2}dz$ and $d\bar{w} = -\bar{z}^{-2}d\bar{z}$, and the type is preserved.

Watch out: Type needs a complex structure

On a real $2n$-manifold there is no preferred way to split k-forms into types. The splitting depends on which functions we declare holomorphic. If we used the chart $z' = \bar{z}$ on part of $\mathbb{C}$, then $dz' = d\bar{z}$ and types $(1,0)$ and $(0,1)$ would swap. Complex manifolds forbid this: their transition functions are holomorphic, never antiholomorphic.

Counting. At each point $\Lambda^{p,q}$ has a basis of monomials with $i_1 < \cdots < i_p$ and $j_1 < \cdots < j_q$, so

$$\dim_{\mathbb{C}} \Lambda^{p,q} = \binom{n}{p}\binom{n}{q}, \qquad \sum_{p+q=k} \binom{n}{p}\binom{n}{q} = \binom{2n}{k} = \dim_{\mathbb{C}} \Lambda^k(\mathbb{R}^{2n}) \otimes \mathbb{C},$$

the last identity being Vandermonde's (choose k objects out of n "z"s and n "$\bar{z}$"s). For $n = 2$:

degree k	(p,q) pieces	dimensions	total	basis
0	$(0,0)$	1	1	1
1	$(1,0),(0,1)$	$2+2$	4	$dz^1, dz^2 \mid d\bar{z}^1, d\bar{z}^2$
2	$(2,0),(1,1),(0,2)$	$1+4+1$	6	$dz^1dz^2 \mid dz^i d\bar{z}^j \mid d\bar{z}^1 d\bar{z}^2$
3	$(2,1),(1,2)$	$2+2$	4	$dz^1dz^2d\bar{z}^j \mid dz^i d\bar{z}^1 d\bar{z}^2$
4	$(2,2)$	1	1	$dz^1dz^2d\bar{z}^1d\bar{z}^2$

(wedges suppressed in the last column).

Worked example 3.29: Type decomposition of real forms

(a) The area form on $\mathbb{C}$. Using (3.62),

$$dz \wedge d\bar{z} \overset{(1)}{=} (dx + i\,dy) \wedge (dx - i\,dy) \overset{(2)}{=} -i\,dx \wedge dy + i\,dy \wedge dx \overset{(3)}{=} -2i\,dx \wedge dy.$$

(1) Definition of $dz, d\bar{z}$.
(2) Expand; $dx \wedge dx = dy \wedge dy = 0$.
(3) $dy \wedge dx = -dx \wedge dy$.

So $dx \wedge dy = \frac{i}{2}\,dz \wedge d\bar{z}$: the real area form is pure type $(1,1)$. We use this constantly.
(b) A real 1-form. $\alpha = a\,dx + b\,dy$ with a, b real. Substituting (3.62), $\alpha = \frac{1}{2}(a - ib)\,dz + \frac{1}{2}(a + ib)\,d\bar{z}$. The $(1,0)$ and $(0,1)$ parts are complex conjugates of each other; a real form always has $\overline{\omega^{p,q}} = \omega^{q,p}$. The electromagnetic potential $A = A_x dx + A_y dy$ on a surface splits as $A = A_z dz + A_{\bar{z}} d\bar{z}$ with $A_z = \frac{1}{2}(A_x - iA_y)$ and $A_{\bar{z}} = \overline{A_z}$.
(c) A 2-form on $\mathbb{C}^2$. The symplectic form of phase space, $\omega = dx^1 \wedge dy^1 + dx^2 \wedge dy^2$ (identifying $z^k = x^k + iy^k$ with $q^k + ip^k$), is by (a) $\omega = \frac{i}{2}(dz^1 \wedge d\bar{z}^1 + dz^2 \wedge d\bar{z}^2)$, type $(1,1)$. The form $dx^1 \wedge dx^2$ on the other hand is

$$dx^1 \wedge dx^2 \overset{(1)}{=} \tfrac{1}{4}(dz^1 + d\bar{z}^1) \wedge (dz^2 + d\bar{z}^2)$$

$$\overset{(2)}{=} \underbrace{\tfrac{1}{4}\mathrm{d}z^1 \wedge \mathrm{d}z^2}_{(2,0)} + \underbrace{\tfrac{1}{4}(\mathrm{d}z^1 \wedge \mathrm{d}\bar{z}^2 + \mathrm{d}\bar{z}^1 \wedge \mathrm{d}z^2)}_{(1,1)} + \underbrace{\tfrac{1}{4}\mathrm{d}\bar{z}^1 \wedge \mathrm{d}\bar{z}^2}_{(0,2)},$$

(1) $\mathrm{d}x^k = \frac{1}{2}(\mathrm{d}z^k + \mathrm{d}\bar{z}^k)$ from (3.62).
(2) Expand and sort by the number of $\mathrm{d}\bar{z}$'s.

which has all three types.

3.7.3 The operators ∂ and $\bar{\partial}$

Equation (3.64) split $\mathrm{d}f$ into a $(1,0)$ part and a $(0,1)$ part. Nash's (3.25) gives these parts names:

$$\partial = \sum_{i=1}^{n} \mathrm{d}z^i \wedge \frac{\partial}{\partial z^i}, \qquad \bar{\partial} = \sum_{i=1}^{n} \mathrm{d}\bar{z}^i \wedge \frac{\partial}{\partial \bar{z}^i}, \tag{3.67}$$

acting on a form (3.65) by differentiating its coefficients and wedging the differential on from the left:

$$\bar{\partial}\omega = \frac{1}{p!\,q!} \sum_{k} \frac{\partial \omega_{i_1 \cdots i_p \bar{j}_1 \cdots \bar{j}_q}}{\partial \bar{z}^k}\, \mathrm{d}\bar{z}^k \wedge \mathrm{d}z^{i_1} \wedge \cdots \wedge \mathrm{d}\bar{z}^{j_q},$$

and similarly for ∂. So ∂ raises the type from (p,q) to $(p+1,q)$ and $\bar{\partial}$ raises it to $(p,q+1)$.

Key idea: $\mathrm{d} = \partial + \bar{\partial}$ and its square (Nash's (3.26))

On a complex manifold,

$$\mathrm{d} = \partial + \bar{\partial}, \qquad \partial^2 = 0, \qquad \partial\bar{\partial} + \bar{\partial}\partial = 0, \qquad \bar{\partial}^2 = 0.$$

The first identity is the chain rule; the other three are $\mathrm{d}^2 = 0$ sorted by type.

Why $\mathrm{d} = \partial + \bar{\partial}$. For a (p,q)-form $\omega = f\,\mathrm{d}z^I \wedge \mathrm{d}\bar{z}^J$ (multi-index notation, one monomial suffices by linearity),

$$\begin{aligned} \mathrm{d}\omega &\overset{(1)}{=} \mathrm{d}f \wedge \mathrm{d}z^I \wedge \mathrm{d}\bar{z}^J \\ &\overset{(2)}{=} \Big(\sum_k \frac{\partial f}{\partial z^k}\mathrm{d}z^k + \sum_k \frac{\partial f}{\partial \bar{z}^k}\mathrm{d}\bar{z}^k\Big) \wedge \mathrm{d}z^I \wedge \mathrm{d}\bar{z}^J \\ &\overset{(3)}{=} \partial\omega + \bar{\partial}\omega. \end{aligned}$$

(1) The coordinate differentials are closed: $\mathrm{d}(\mathrm{d}z^i) = \mathrm{d}\mathrm{d}x^i + i\,\mathrm{d}\mathrm{d}y^i = 0$, so d only hits the coefficient.
(2) Equation (3.64) in n variables.
(3) Definition (3.67); the first sum is $(p+1,q)$, the second $(p,q+1)$.

(If the transition functions were not holomorphic, $\mathrm{d}(\mathrm{d}z^i)$ would still vanish, but a (p,q)-form in one chart could acquire other types in the next and $\partial, \bar{\partial}$ would not be chart-independent. Holomorphy is what makes the splitting (3.67) global.)

Why the three identities. Apply d twice to a (p,q)-form:

$$\begin{aligned} 0 &\overset{(1)}{=} \mathrm{d}^2\omega \\ &\overset{(2)}{=} (\partial+\bar\partial)(\partial+\bar\partial)\omega \\ &\overset{(3)}{=} \underbrace{\partial^2\omega}_{(p+2,q)} + \underbrace{(\partial\bar\partial+\bar\partial\partial)\omega}_{(p+1,q+1)} + \underbrace{\bar\partial^2\omega}_{(p,q+2)}. \end{aligned}$$

(1) $\mathrm{d}^2=0$ (Chapter T1).
(2) $\mathrm{d}=\partial+\bar\partial$.
(3) Expand and record the type of each piece.

The decomposition (3.66) is a direct sum, so a form is zero only if each of its type components is zero. All three brackets vanish separately.

Worked example 3.30: Checking $\partial\bar\partial=-\bar\partial\partial$ on a function

Let $f(z,\bar z)$ be a smooth function on $\mathbb{C}$. Then

$$\begin{aligned} \partial\bar\partial f &\overset{(1)}{=} \partial\big(\partial_{\bar z} f\,\mathrm{d}\bar z\big) \overset{(2)}{=} \partial_z\partial_{\bar z} f\;\mathrm{d}z\wedge\mathrm{d}\bar z, \\ \bar\partial\partial f &\overset{(3)}{=} \bar\partial\big(\partial_z f\,\mathrm{d}z\big) \overset{(4)}{=} \partial_{\bar z}\partial_z f\;\mathrm{d}\bar z\wedge\mathrm{d}z \overset{(5)}{=} -\partial_z\partial_{\bar z} f\;\mathrm{d}z\wedge\mathrm{d}\bar z. \end{aligned}$$

(1) $\bar\partial f$ is the $\mathrm{d}\bar z$ part of $\mathrm{d}f$.
(2) Differentiate the coefficient in z and wedge $\mathrm{d}z$ on the left.
(3) ∂f is the $\mathrm{d}z$ part of $\mathrm{d}f$.
(4) Differentiate in $\bar z$ and wedge $\mathrm{d}\bar z$ on the left.
(5) Mixed partials commute; $\mathrm{d}\bar z\wedge\mathrm{d}z=-\mathrm{d}z\wedge\mathrm{d}\bar z$.

The sum is zero, as promised. With Example 3.29(a) and $4\partial_z\partial_{\bar z}=\nabla^2$,

$$2i\,\partial\bar\partial f = 2i\cdot\tfrac14\nabla^2 f\cdot(-2i)\,\mathrm{d}x\wedge\mathrm{d}y = \nabla^2 f\;\mathrm{d}x\wedge\mathrm{d}y.$$

So $2i\,\partial\bar\partial$ is "the Laplacian times the area form". In Kähler geometry and in the theory of instantons this operator appears everywhere; for instance, $\frac{i}{2\pi}\partial\bar\partial\log(1+|z|^2)$ is the Fubini–Study area form of $\mathbb{CP}^1$, which we meet in Example 3.34.

Holomorphic p-forms. A $(p,0)$-form $\omega=\frac{1}{p!}\omega_{i_1\cdots i_p}\mathrm{d}z^{i_1}\wedge\cdots\wedge\mathrm{d}z^{i_p}$ is called **holomorphic** if its coefficients are holomorphic functions. Since $\bar\partial\omega$ collects exactly the $\bar z$-derivatives of the coefficients, this is the same as $\bar\partial\omega=0$:

$$\Omega^p = \mathcal{Z}^{p,0} := \operatorname{Ker}\big(\bar\partial:\mathcal{E}^{p,0}\to\mathcal{E}^{p,1}\big), \qquad \Omega^0=\mathcal{O}. \tag{3.68}$$

For example $f(z)\,\mathrm{d}z$ on $\mathbb{C}$ is holomorphic iff $\bar\partial(f\,\mathrm{d}z)=\partial_{\bar z}f\,\mathrm{d}\bar z\wedge\mathrm{d}z=0$ iff f is holomorphic. Note that a holomorphic 1-form is automatically d-closed in one variable: $\mathrm{d}(f\,\mathrm{d}z)=\partial_{\bar z}f\,\mathrm{d}\bar z\wedge\mathrm{d}z=0$. That is why contour integrals of holomorphic functions depend only on the homology class of the contour: Cauchy's theorem is Stokes' theorem for the closed form $f\,\mathrm{d}z$.

Watch out: Nash's "type $(0,p)$"

On p. 64 Nash writes that holomorphic p-forms are "necessarily of type $(0,p)$". This is a slip: they are of type $(p,0)$ (all $\mathrm{d}z$'s, no $\mathrm{d}\bar z$'s), as his own next sentence $\Omega^p=\mathcal{Z}^{p,0}$ says.

Also, in the first line of his (3.28) the arrow $\Omega^p \to \mathcal{E}^{p,0}$ is the inclusion, not $\bar\partial$; the $\bar\partial$ label belongs only on the second arrow.

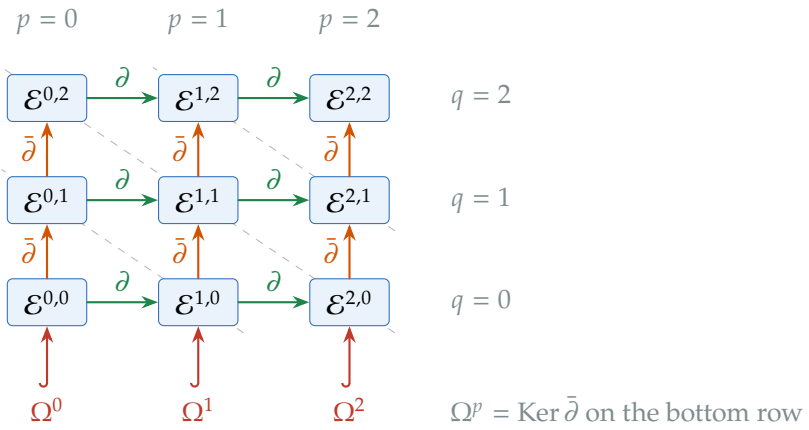


Figure 3.25: The Dolbeault double complex of a complex surface ($n = 2$). The sheaf $\mathcal{E}^{p,q}$ sits at position (p, q); ∂ (green) moves right, $\bar\partial$ (orange) moves up, and $\mathrm{d} = \partial + \bar\partial$ moves from one dashed antidiagonal (fixed total degree $k = p + q$) to the next. Each column, read upwards, is the resolution of the holomorphic p-forms Ω^p (red) that computes $H^q(M; \Omega^p)$.

Dictionary: $\bar\partial$ is half of the 2d Dirac operator

Take the Euclidean gamma matrices $\gamma^1 = \sigma_1$, $\gamma^2 = \sigma_2$ on $\mathbb{R}^2$ (Pauli matrices). Then

$$\not{D} \overset{(1)}{=} \sigma_1\partial_x + \sigma_2\partial_y \overset{(2)}{=} \begin{pmatrix} 0 & \partial_x - i\partial_y \\ \partial_x + i\partial_y & 0 \end{pmatrix} \overset{(3)}{=} \begin{pmatrix} 0 & 2\partial_z \\ 2\partial_{\bar z} & 0 \end{pmatrix}.$$

(1) $\not{D} = \gamma^\mu\partial_\mu$ (Chapter T4).

(2) $\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, $\sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}$.

(3) The Wirtinger derivatives (3.63).

So on a spinor $\psi = (\psi_+, \psi_-)^T$ the Dirac equation $\not{D}\psi = 0$ says $\partial_{\bar z}\psi_+ = 0$ and $\partial_z\psi_- = 0$: the positive-chirality component is holomorphic, the negative one antiholomorphic. "Chiral fermion in 2d" = "$\bar\partial$". Counting chiral zero modes on a compact Riemann surface is exactly counting $\dim \mathrm{Ker}\,\bar\partial - \dim \mathrm{Coker}\,\bar\partial$, which by the Dolbeault theorem is $\dim H^0 - \dim H^1$ of a sheaf: the Riemann–Roch index we met in Example 3.18.

3.7.4 The $\bar\partial$-Poincaré lemma: solving $\partial_{\bar z} f = g$

In the de Rham story, the local input was the Poincaré lemma: on a ball every closed form is exact. Here we need its analogue, the **Dolbeault lemma** (also called the Dolbeault–Grothendieck or $\bar\partial$-Poincaré lemma). Nash uses it silently when he calls the sequences (3.28) exact. We now supply it, starting with the one-variable case, where it amounts to inverting $\partial_{\bar z}$.

Step 1: the Green's function of $\partial_{\bar z}$. We claim

$$\frac{\partial}{\partial \bar z}\,\frac{1}{\pi z} = \delta^2(z), \tag{3.69}$$

where $\delta^2(z) = \delta(x)\delta(y)$ and we integrate against the real area element $\mathrm{d}^2z := \mathrm{d}x\,\mathrm{d}y$. Away from $z = 0$ the left side vanishes, since $1/z$ is holomorphic there. To find the strength of the

singularity, integrate over a disc $D_R = \{|z| < R\}$ using Stokes' theorem in complex form. For any smooth h on a region D,

$$\oint_{\partial D} h\,\mathrm{d}z \overset{(1)}{=} \int_D \mathrm{d}(h\,\mathrm{d}z) \overset{(2)}{=} \int_D \partial_{\bar z} h\;\mathrm{d}\bar z \wedge \mathrm{d}z \overset{(3)}{=} 2i \int_D \partial_{\bar z} h\;\mathrm{d}^2 z. \tag{3.70}$$

(1) Stokes' theorem (Chapter T2).
(2) $\mathrm{d}(h\,\mathrm{d}z) = \mathrm{d}h \wedge \mathrm{d}z$ and only the $\mathrm{d}\bar z$ part of $\mathrm{d}h$ survives the wedge with $\mathrm{d}z$.
(3) $\mathrm{d}\bar z \wedge \mathrm{d}z = +2i\,\mathrm{d}x \wedge \mathrm{d}y$ by Example 3.29(a).

Apply this to $h = 1/z$ on the annulus $\varepsilon < |z| < R$, where h is smooth and $\partial_{\bar z} h = 0$: the outer and inner circle integrals agree, and both equal $\oint \mathrm{d}z/z = 2\pi i$. The "missing" $2\pi i$ must therefore be produced by a point source at the origin. Treating the disc as a whole in the sense of distributions,

$$2i \int_{D_R} \frac{\partial}{\partial \bar z}\frac{1}{z}\,\mathrm{d}^2 z = \oint_{|z|=R} \frac{\mathrm{d}z}{z} = 2\pi i \quad\Longrightarrow\quad \int_{D_R} \frac{\partial}{\partial \bar z}\frac{1}{z}\,\mathrm{d}^2 z = \pi,$$

for every R. A distribution supported at 0 with total integral π is $\pi\delta^2$, which is (3.69). (The same computation with $\partial_z \partial_{\bar z} \log|z|^2 = \partial_{\bar z}(1/z) = \pi\delta^2$ recovers the 2d Coulomb law $\nabla^2 \log|z| = 2\pi\delta^2$, completing Example 3.28.)

Step 2: the Cauchy integral. Given a smooth g with compact support in $\mathbb{C}$, set

$$f(z) := \frac{1}{\pi}\int_{\mathbb{C}} \frac{g(w)}{z-w}\,\mathrm{d}^2 w. \tag{3.71}$$

The integrand has an integrable $1/|z-w|$ singularity, so f is well defined. Then

$$\frac{\partial f}{\partial \bar z}(z) \overset{(1)}{=} \int_{\mathbb{C}} g(w)\,\frac{\partial}{\partial \bar z}\Big(\frac{1}{\pi(z-w)}\Big)\mathrm{d}^2 w \overset{(2)}{=} \int_{\mathbb{C}} g(w)\,\delta^2(z-w)\,\mathrm{d}^2 w \overset{(3)}{=} g(z).$$

(1) Differentiate under the integral sign (in the sense of distributions; the honest version substitutes $u = z - w$ first, differentiates $g(z-u)$, and then undoes the derivative with (3.70) on $\varepsilon < |u| < R$, letting $\varepsilon \to 0$).
(2) The Green's function (3.69), translated to w.
(3) The defining property of δ^2.

So (3.71) solves $\partial_{\bar z} f = g$. Every smooth $(0,1)$-form $g\,\mathrm{d}\bar z$ with compact support is $\bar\partial$-exact on $\mathbb{C}$.

Worked example 3.31: Solving $\partial_{\bar z} f = 1$ on a disc with the Cauchy kernel

Of course $f = \bar z$ solves $\partial_{\bar z} f = 1$. Let us check that the Cauchy integral finds it. Take $g = 1$ on the disc D_R and 0 outside, and evaluate f from (3.71) at a point $|z| < R$. In polar coordinates $w = re^{i\theta}$, first do the angular integral:

$$\int_0^{2\pi} \frac{\mathrm{d}\theta}{z - re^{i\theta}} = \begin{cases} 2\pi/z, & r < |z|\,, \\ 0, & r > |z|\,. \end{cases}$$

For $r < |z|$ expand $\frac{1}{z-w} = \frac{1}{z}\sum_{k\ge 0}(w/z)^k$; only $k = 0$ survives the θ-integral. For $r > |z|$ expand $\frac{1}{z-w} = -\frac{1}{w}\sum_{k\ge 0}(z/w)^k$; every term carries $e^{-i(k+1)\theta}$ and integrates to zero. Then

$$f(z) \overset{(1)}{=} \frac{1}{\pi}\int_0^R r\,\mathrm{d}r \int_0^{2\pi} \frac{\mathrm{d}\theta}{z - re^{i\theta}} \overset{(2)}{=} \frac{1}{\pi}\int_0^{|z|} \frac{2\pi}{z}\,r\,\mathrm{d}r \overset{(3)}{=} \frac{|z|^2}{z} \overset{(4)}{=} \bar z.$$

(1) Definition (3.71) with $\mathrm{d}^2 w = r\,\mathrm{d}r\,\mathrm{d}\theta$.

(2) The angular integral just computed: only the circles inside $|z|$ contribute.
(3) $\int_0^{|z|} r\,\mathrm{d}r = |z|^2/2$.
(4) $|z|^2 = z\bar{z}$.

Inside the disc the Cauchy integral returns exactly $\bar{z}$. (Outside, for $|z| > R$, every circle contributes and $f = R^2/z$, which is holomorphic there, as it must be since $g = 0$ outside the disc.) Electrostatically, this is Gauss's law for a uniformly charged disc, written in complex form: the "field" f grows linearly inside and falls off like $1/z$ outside.

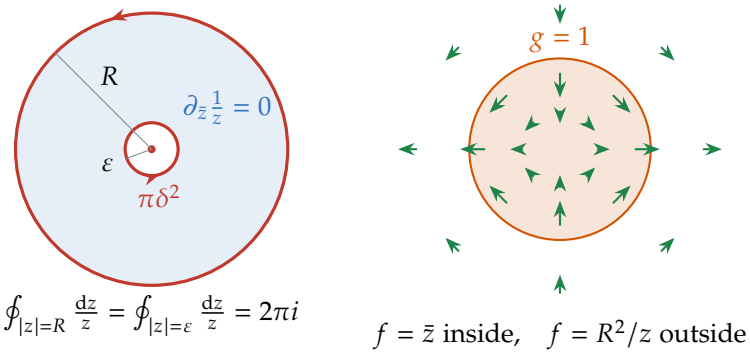


Figure 3.26: Inverting $\partial_{\bar{z}}$. Left: $1/z$ is holomorphic on the shaded annulus, so its contour integrals over the two red circles agree; the whole $2\pi i$ is carried by a point source $\pi\delta^2$ at the origin, which makes $1/(\pi z)$ the Green's function of $\partial_{\bar{z}}$. Right: the Cauchy integral of a uniform source $g = 1$ on a disc (Example 3.31). Green arrows show the complex number $f(z)$ as a vector: $\bar{z}$ inside (growing, reflected in the real axis), and the holomorphic R^2/z outside (decaying).

Step 3: from compact support to any disc. Nash's sheaves are local, so we need to solve $\partial_{\bar{z}} f = g$ near a point for g that need not have compact support. That is easy: to solve on a smaller disc D_r, multiply g by a bump function χ equal to 1 on D_r and supported in D_R, and apply (3.71) to χg. The solution f satisfies $\partial_{\bar{z}} f = \chi g = g$ on D_r. This proves the one-variable Dolbeault lemma: *every point has a neighbourhood on which every* $(0,1)$*-form is* $\bar{\partial}$*-exact.* With a little more work (solve on an exhausting sequence of discs and correct the discrepancies with holomorphic functions, which can be approximated by polynomials by Runge's theorem) one solves $\partial_{\bar{z}} f = g$ on the whole of any open set $U \subset \mathbb{C}$. We use this stronger form below; we do not prove it.

Step 4: several variables. On a polydisc $P = D_1 \times \cdots \times D_n \subset \mathbb{C}^n$ take a $\bar{\partial}$-closed $(0,q)$-form. The standard argument solves the equation one variable at a time, using the Cauchy integral in one variable with the others as parameters. The case $n = 2$, $q = 1$ shows every idea. Let $\omega = g_1\,\mathrm{d}\bar{z}^1 + g_2\,\mathrm{d}\bar{z}^2$ with compact support and $\bar{\partial}\omega = 0$, i.e. (the coefficient of $\mathrm{d}\bar{z}^1 \wedge \mathrm{d}\bar{z}^2$)

$$\frac{\partial g_2}{\partial \bar{z}^1} = \frac{\partial g_1}{\partial \bar{z}^2}. \tag{3.72}$$

Solve the first component with the Cauchy integral in z^1 only: $f(z^1, z^2) = \frac{1}{\pi}\int \frac{g_1(w, z^2)}{z^1 - w}\,\mathrm{d}^2 w$. Then $\partial f/\partial \bar{z}^1 = g_1$ by Step 2, and

$$\frac{\partial f}{\partial \bar{z}^2} \overset{(1)}{=} \frac{1}{\pi}\int \frac{\partial_{\bar{z}^2} g_1(w, z^2)}{z^1 - w}\,\mathrm{d}^2 w \overset{(2)}{=} \frac{1}{\pi}\int \frac{\partial_{\bar{w}} g_2(w, z^2)}{z^1 - w}\,\mathrm{d}^2 w \overset{(3)}{=} g_2(z^1, z^2).$$

(1) The kernel $1/(z^1 - w)$ does not depend on z^2; differentiate the numerator.
(2) The closedness condition (3.72).
(3) The Cauchy–Pompeiu formula: for compactly supported h, $\frac{1}{\pi}\int \frac{\partial_{\bar{w}} h(w)}{z - w}\mathrm{d}^2 w = h(z)$. (Proof: write $h = \frac{1}{\pi}\int \frac{\partial_{\bar{w}} h}{z-w} +$ (holomorphic) by Step 2; the holomorphic remainder is entire and vanishes at infinity, hence is zero by Liouville.)

So $\bar\partial f = g_1 \mathrm{d}\bar z^1 + g_2 \mathrm{d}\bar z^2 = \omega$. For higher q and n one inducts on the number of $\mathrm{d}\bar z^k$'s that appear, and for (p,q)-forms one simply carries the $\mathrm{d}z^{i_1}\wedge\cdots\wedge\mathrm{d}z^{i_p}$ along as spectators ($\bar\partial$ does not touch them).

Key idea: The Dolbeault lemma

On a polydisc (in particular, in a small neighbourhood of any point of a complex manifold), every $\bar\partial$-closed (p,q)-form with $q \geq 1$ is $\bar\partial$-exact. In one variable the inverse of $\partial_{\bar z}$ is convolution with the Green's function $1/(\pi z)$. Consequently, for each p the sequence of sheaves

$$0 \longrightarrow \Omega^p \hookrightarrow \mathcal{E}^{p,0} \xrightarrow{\bar\partial} \mathcal{E}^{p,1} \xrightarrow{\bar\partial} \mathcal{E}^{p,2} \xrightarrow{\bar\partial} \cdots \xrightarrow{\bar\partial} \mathcal{E}^{p,n} \longrightarrow 0 \tag{3.73}$$

is exact: it is a resolution of Ω^p, the **Dolbeault resolution**.

Two structural facts complete the input to Nash's argument.

- *$\mathcal{E}^{p,q}$ is fine.* If $\{\rho_\alpha\}$ is a smooth (real) partition of unity and ω has type (p,q), then $\rho_\alpha\omega$ still has type (p,q): multiplying by a function does not change which differentials appear. So the partition-of-unity argument of Section 3.2.5 applies verbatim, and $H^k(M;\mathcal{E}^{p,q}) = 0$ for $k \geq 1$ (Section 3.3.7).
- *$\mathcal{Z}^{p,q}$ and Ω^p are not fine.* Multiplying a holomorphic function by a bump function destroys holomorphy (by the identity theorem, a holomorphic function vanishing on an open set of a connected domain vanishes everywhere, so holomorphic bump functions do not exist). This rigidity is why $\mathcal{O}$ has interesting cohomology.

3.7.5 Nash's argument: the Dolbeault theorem

We now have everything Nash needs, and his proof is the de Rham proof (Section 3.6.2) with three substitutions:

$$\mathrm{d} \rightsquigarrow \bar\partial, \qquad \mathcal{E}^q \rightsquigarrow \mathcal{E}^{p,q}, \qquad \mathcal{Z}^q \rightsquigarrow \mathcal{Z}^{p,q} := \operatorname{Ker}\bigl(\bar\partial : \mathcal{E}^{p,q} \to \mathcal{E}^{p,q+1}\bigr),$$

with the holomorphic degree p fixed throughout, a passive spectator. Chop the Dolbeault resolution (3.73) into short exact sequences, Nash's (3.28):

$$\begin{gathered} 0 \to \Omega^p \to \mathcal{E}^{p,0} \xrightarrow{\bar\partial} \mathcal{Z}^{p,1} \to 0, \qquad 0 \to \mathcal{Z}^{p,1} \to \mathcal{E}^{p,1} \xrightarrow{\bar\partial} \mathcal{Z}^{p,2} \to 0, \qquad \dots, \\ 0 \to \mathcal{Z}^{p,q} \to \mathcal{E}^{p,q} \xrightarrow{\bar\partial} \mathcal{Z}^{p,q+1} \to 0, \qquad \dots \end{gathered} \tag{3.74}$$

Why is each one exact? On stalks (which is what exactness of sheaves means, Section 3.2.3):

- *at the left*: $\mathcal{Z}^{p,q} \to \mathcal{E}^{p,q}$ is an inclusion, hence injective;
- *in the middle*: the kernel of $\bar\partial$ on $\mathcal{E}^{p,q}$ is $\mathcal{Z}^{p,q}$ by definition;
- *at the right*: $\bar\partial$ maps $\mathcal{E}^{p,q}$ into $\mathcal{Z}^{p,q+1}$ because $\bar\partial^2 = 0$, and *onto* it near every point by the Dolbeault lemma (Section 3.7.4).

Only the last point has content, exactly as in the de Rham case.

Feed one of these into the long exact sequence (3.37):

$$\cdots \to H^k(M;\mathcal{E}^{p,q}) \to H^k(M;\mathcal{Z}^{p,q+1}) \xrightarrow{\delta^*} H^{k+1}(M;\mathcal{Z}^{p,q}) \to H^{k+1}(M;\mathcal{E}^{p,q}) \to \cdots$$

Because $\mathcal{E}^{p,q}$ is fine, the outer terms vanish for $k \geq 1$, and

$$H^{k+1}(M;\mathcal{Z}^{p,q}) \cong H^k(M;\mathcal{Z}^{p,q+1}) \qquad (k \geq 1). \tag{3.75}$$

For $k = 0$ the left-hand term is the space of global (p,q)-forms, and exactness of $\mathcal{E}^{p,q}(M) \xrightarrow{\bar\partial} \mathcal{Z}^{p,q+1}(M) \xrightarrow{\delta^*} H^1(M;\mathcal{Z}^{p,q}) \to 0$ gives

$$H^1(M;\mathcal{Z}^{p,q-1}) \cong \frac{H^0(M;\mathcal{Z}^{p,q})}{\bar\partial\, H^0(M;\mathcal{E}^{p,q-1})} = \frac{\operatorname{Ker}\big(\bar\partial : \mathcal{E}^{p,q}(M) \to \mathcal{E}^{p,q+1}(M)\big)}{\operatorname{Im}\big(\bar\partial : \mathcal{E}^{p,q-1}(M) \to \mathcal{E}^{p,q}(M)\big)}. \tag{3.76}$$

The left equality is Nash's (3.29); the right one, which is (3.30), uses $H^0 = \Gamma$ (3.17) in numerator and denominator.

Working definition 3.21 (Dolbeault cohomology)

The **Dolbeault cohomology** of a complex manifold [charts in $\mathbb{C}^n$ glued by holomorphic coordinate changes] is

$$H^{p,q}_{\bar\partial}(M) := \frac{\{\text{global } (p,q)\text{-forms } \omega \text{ with } \bar\partial\omega = 0\}}{\{\bar\partial\beta : \ \beta \text{ a global } (p,q-1)\text{-form}\}}.$$

[here $\bar\partial\omega$ is the $(p,q+1)$ part of $\mathrm{d}\omega$, obtained by differentiating the coefficients in the $\bar z^j$ only] *To compute it:* find all global solutions of $\bar\partial\omega = 0$ in type (p,q), and decide which of them are $\bar\partial$ of something global. For $q = 0$ there is nothing to divide by: $H^{p,0}_{\bar\partial}(M)$ is the space of global holomorphic p-forms [$(p,0)$-forms whose coefficients are holomorphic functions].

Now chain the isomorphisms, exactly as in (3.54). For $q \geq 1$,

$$\begin{aligned} H^q(M;\Omega^p) &\overset{(1)}{=} H^q(M;\mathcal{Z}^{p,0}) \\ &\overset{(2)}{=} H^{q-1}(M;\mathcal{Z}^{p,1}) \overset{(2)}{=} H^{q-2}(M;\mathcal{Z}^{p,2}) \overset{(2)}{=} \cdots \overset{(2)}{=} H^1(M;\mathcal{Z}^{p,q-1}) \\ &\overset{(3)}{=} \frac{\mathcal{Z}^{p,q}(M)}{\bar\partial\,\mathcal{E}^{p,q-1}(M)} \overset{(4)}{=} H^{p,q}_{\bar\partial}(M). \end{aligned} \tag{3.77}$$

(1) Holomorphic p-forms are the $\bar\partial$-closed $(p,0)$-forms (3.68).
(2) The degree shift (3.75), used $q-1$ times, each time with $k \geq 1$ ("=" means canonically isomorphic).
(3) The last step (3.76).
(4) Definition of Dolbeault cohomology.

For $q = 0$ both sides are the global holomorphic p-forms.

Key idea: The Dolbeault theorem (Nash's (3.31))

For every complex manifold M and all $p, q \geq 0$,

$$H^q(M;\Omega^p) \ \cong \ H^{p,q}_{\bar\partial}(M), \qquad \text{in particular} \qquad H^q(M;\mathcal{O}) \cong H^{0,q}_{\bar\partial}(M).$$

The Čech cohomology of the rigid sheaf of holomorphic forms is computed by solving the inhomogeneous Cauchy–Riemann equation $\bar\partial\beta = \omega$ globally. Mechanism: the same three ingredients as de Rham, namely a local solvability lemma (Dolbeault), fineness of the smooth forms (partitions of unity), and the long exact sequence.

	de Rham	**Dolbeault**
sheaf being resolved	$\underline{\mathbb{R}}$ (locally constant)	Ω^p (holomorphic p-forms)
fine sheaves	$\mathcal{E}^q$ (q-forms)	$\mathcal{E}^{p,q}$ ((p,q)-forms)
differential	d	$\bar\partial$
local lemma	Poincaré: $\mathrm{d}\omega = 0 \Rightarrow \omega = \mathrm{d}\alpha$	Dolbeault: $\bar\partial\omega = 0 \Rightarrow \omega = \bar\partial\beta$
local inverse	radial homotopy (3.48)	Cauchy kernel $1/(\pi z)$ (3.71)
theorem	$H^q(M;\mathbb{R}) \cong H^q_{\mathrm{dR}}(M)$	$H^q(M;\Omega^p) \cong H^{p,q}_{\bar\partial}(M)$
vanishes for	$q > \dim_{\mathbb{R}} M = 2n$	$q > \dim_{\mathbb{C}} M = n$
Nash	(3.20)–(3.23)	(3.28)–(3.31)

Three remarks.

- *Vanishing above the complex dimension.* There are no (p,q)-forms with $q > n$ (a wedge of more than n of the $\mathrm{d}\bar z^j$ is zero). So $H^q(M;\Omega^p) = 0$ for $q > n$, although the real dimension is $2n$. On a Riemann surface ($n = 1$) only H^0 and H^1 of $\mathcal{O}$ can be nonzero, which matches what the two-set cover of $\mathbb{CP}^1$ gave us (Example 3.18).

- *Bundle coefficients.* If $E \to M$ is a holomorphic vector bundle (holomorphic transition functions $g_{\alpha\beta}$), then $\bar\partial$ acts on E-valued forms, the sheaves $\mathcal{E}^{p,q}(E)$ of (3.8). The reason is one line of Leibniz: if $s_\alpha = g_{\alpha\beta}s_\beta$ on an overlap, then

$$\bar\partial s_\alpha \overset{(1)}{=} \bar\partial(g_{\alpha\beta}s_\beta) \overset{(2)}{=} (\bar\partial g_{\alpha\beta})\,s_\beta + g_{\alpha\beta}\,\bar\partial s_\beta \overset{(3)}{=} g_{\alpha\beta}\,\bar\partial s_\beta,$$

  (1) The local components of the section are related by the transition function.
  (2) Leibniz rule for the first-order operator $\bar\partial$.
  (3) $g_{\alpha\beta}$ is holomorphic.

  so the local $\bar\partial s_\alpha$ glue into a global E-valued $(0,1)$-form. (For d this fails: $\mathrm{d}g_{\alpha\beta} \neq 0$, and one needs a connection.) The whole argument goes through and gives $H^q(M;\Omega^p(E)) \cong H^{p,q}_{\bar\partial}(M;E)$. For $E = \mathcal{O}(n)$ on $\mathbb{CP}^1$ this turns the Laurent-series counting of Example 3.18 into a statement about the operator $\bar\partial_{\mathcal{O}(n)}$: its kernel has dimension $\dim H^0$, its cokernel dimension $\dim H^1$, and its index is $n+1$. This is the prototype of the index theorems of Chapter 4.

- *Dolbeault versus de Rham.* On a compact Kähler manifold (for instance any compact Riemann surface, or $\mathbb{CP}^n$) Hodge theory gives $H^k_{\mathrm{dR}}(M)\otimes\mathbb{C} \cong \bigoplus_{p+q=k} H^{p,q}_{\bar\partial}(M)$, so the Betti numbers split as $b_k = \sum_{p+q=k} h^{p,q}$ with $h^{p,q} = \dim H^{p,q}_{\bar\partial}$. We shall check this in the examples below; it is not part of Nash's argument, which works for any complex manifold, compact or not.

3.7.6 Examples

Worked example 3.32: Open subsets of $\mathbb{C}$: no higher $\mathcal{O}$-cohomology

Let $U \subset \mathbb{C}$ be open (a disc, $\mathbb{C}$ itself, $\mathbb{C}^*$, an annulus, ...). Since $n = 1$, the only possibly nonzero higher group is

$$H^1(U;\mathcal{O}) \overset{(1)}{=} H^{0,1}_{\bar\partial}(U) \overset{(2)}{=} \frac{\{g\,\mathrm{d}\bar z :\ g \in C^\infty(U)\}}{\{\partial_{\bar z} f\,\mathrm{d}\bar z :\ f \in C^\infty(U)\}} \overset{(3)}{=} 0.$$

(1) The Dolbeault theorem with $p = 0$, $q = 1$.

(2) In one complex dimension every $(0,1)$-form is $\bar\partial$-closed (its $\bar\partial$ would be a $(0,2)$-form, and there are none), and $\bar\partial f = \partial_{\bar z} f \, d\bar z$.
(3) $\partial_{\bar z} f = g$ is solvable on all of U (Step 3 of Section 3.7.4).

The same argument with $f\,dz$ in place of f gives $H^1(U;\Omega^1) = 0$. This is the fact we borrowed earlier: discs, $\mathbb{C}$ and $\mathbb{C}^*$ are acyclic for $\mathcal{O}$, so the covers of $\mathbb{CP}^1$ and of $\mathbb{C}^*$ used in Example 3.18 and Example 3.19 are Leray covers.
Combine it with the exponential sequence (3.43):

$$0 = H^1(U;\mathcal{O}) \to H^1(U;\mathcal{O}^*) \xrightarrow{c_1} H^2(U;\mathbb{Z}) = 0,$$

where $H^2(U;\mathbb{Z}) = 0$ because an open subset of the plane has no closed surfaces in it to carry flux. Hence $H^1(U;\mathcal{O}^*) = 0$: *every holomorphic line bundle over an open subset of $\mathbb{C}$ is trivial.* In gauge language: on a planar domain every holomorphic transition function g_{01} can be written as h_0/h_1 with h_i holomorphic and nowhere zero on the patches.

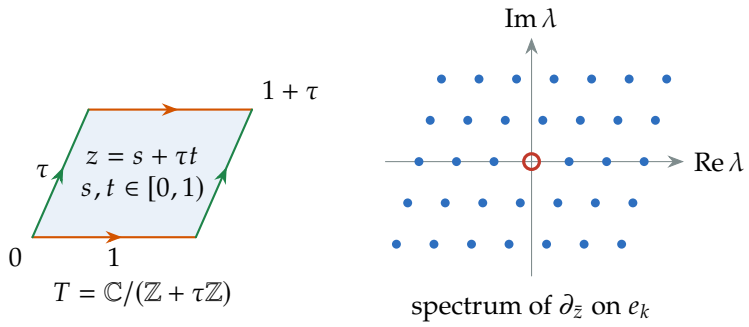


Figure 3.27: Left: a complex torus is the parallelogram with opposite sides identified (orange with orange, green with green); s,t are periodic real coordinates. Right: $\partial_{\bar z}$ is diagonal on the Fourier modes e_k, with eigenvalues $\lambda_k = \pi(k_1\tau - k_2)/\operatorname{Im}\tau$ forming a lattice in $\mathbb{C}$ (drawn for $\tau = 0.3 + 1.1i$). Only $k = 0$ gives zero (red): one zero mode (constants, $H^{0,0}$), and one obstruction to solving $\partial_{\bar z} f = g$ (the average of g, $H^{0,1}$).

Worked example 3.33: The torus: $H^{0,1}_{\bar\partial}(T) \cong \mathbb{C}$ by Fourier analysis

Let $T = \mathbb{C}/\Lambda$ with lattice $\Lambda = \mathbb{Z} + \tau\mathbb{Z}$, $\tau = \tau_1 + i\tau_2$, $\tau_2 > 0$. Every $z \in \mathbb{C}$ is uniquely $z = s + \tau t$ with s,t real, and translating by the lattice shifts s or t by an integer. So a smooth function on T is a smooth function of (s,t) periodic with period 1 in each, and it has a Fourier series

$$f = \sum_{k\in\mathbb{Z}^2} f_k\, e_k, \qquad e_k := e^{2\pi i(k_1 s + k_2 t)},$$

with rapidly decreasing coefficients f_k.
Step 1: express s,t through $z,\bar z$. From $z = s + \tau t$ and $\bar z = s + \bar\tau t$ (s,t are real):

$$t \overset{(1)}{=} \frac{z - \bar z}{\tau - \bar\tau}, \qquad s \overset{(2)}{=} z - \tau t = \frac{\tau\bar z - \bar\tau z}{\tau - \bar\tau},$$

(1) Subtract the two equations.
(2) Substitute back; over the common denominator, $z(\tau - \bar\tau) - \tau(z - \bar z) = \tau\bar z - \bar\tau z$.

hence $\partial_{\bar z} s = \tau/(\tau - \bar\tau)$ and $\partial_{\bar z} t = -1/(\tau - \bar\tau)$, with $\tau - \bar\tau = 2i\tau_2$.
Step 2: $\partial_{\bar z}$ is diagonal on Fourier modes.

$$\partial_{\bar z} e_k \overset{(1)}{=} 2\pi i\,(k_1\partial_{\bar z}s + k_2\partial_{\bar z}t)\, e_k \overset{(2)}{=} 2\pi i\,\frac{k_1\tau - k_2}{2i\tau_2}\, e_k \overset{(3)}{=} \lambda_k\, e_k, \qquad \lambda_k := \frac{\pi\,(k_1\tau - k_2)}{\tau_2}. \tag{3.78}$$

(1) Chain rule on the exponential.
(2) Step 1.
(3) Cancel $2i$ and name the eigenvalue.

When is $\lambda_k = 0$? Its imaginary part is πk_1, so we need $k_1 = 0$, and then $\lambda_k = -\pi k_2/\tau_2$ forces $k_2 = 0$. So $\lambda_k = 0$ only for $k = 0$, and in fact $|\lambda_k| \geq c\ |k|$ for some $c > 0$ (the λ_k form a lattice, Figure 3.27).

Step 3: the cohomology. The form $\mathrm{d}\bar{z}$ is invariant under translations, so it is a global $(0,1)$-form on T, and every $(0,1)$-form is $g\,\mathrm{d}\bar{z}$ with g a smooth function on T. We want to solve $\bar{\partial} f = g\,\mathrm{d}\bar{z}$, i.e. $\partial_{\bar{z}} f = g$. In Fourier components, by (3.78),

$$\lambda_k f_k = g_k \qquad \text{for all } k \in \mathbb{Z}^2.$$

- For $k \neq 0$: $f_k = g_k/\lambda_k$. Since $|\lambda_k| \geq c\,|k|$, these still decrease rapidly, and f is smooth.
- For $k = 0$: $0 \cdot f_0 = g_0$. Solvable only if $g_0 = 0$, and then f_0 is arbitrary.

So

$$H^{0,0}_{\bar{\partial}}(T) = \operatorname{Ker} \partial_{\bar{z}} = \{f :\ f_k = 0\ \forall k \neq 0\} = \mathbb{C} \quad \text{(constants)},$$
$$H^{0,1}_{\bar{\partial}}(T) = \frac{\{g\,\mathrm{d}\bar{z}\}}{\{\partial_{\bar{z}} f\,\mathrm{d}\bar{z}\}} \xrightarrow[\cong]{g\,\mathrm{d}\bar{z} \mapsto g_0} \mathbb{C}, \qquad g_0 = \int_0^1\!\!\int_0^1 g\,\mathrm{d}s\,\mathrm{d}t.$$

The class is represented by $\mathrm{d}\bar{z}$ itself. By the Dolbeault theorem, $H^1(T;\mathcal{O}) \cong \mathbb{C}$, as claimed in Example 3.21. The first line is Liouville's theorem for a compact Riemann surface (holomorphic functions on T are constant), obtained here by Fourier analysis. The same computation with $\mathrm{d}z$ carried along gives $H^{1,0}_{\bar{\partial}}(T) = \mathbb{C}\,\mathrm{d}z$ and $H^{1,1}_{\bar{\partial}}(T) = \mathbb{C}\,[\mathrm{d}z \wedge \mathrm{d}\bar{z}]$. All four Hodge numbers are 1, and $b_1(T) = 2 = h^{1,0} + h^{0,1}$: the two real classes $[\mathrm{d}s]$, $[\mathrm{d}t]$ are recombined into $[\mathrm{d}z]$ and $[\mathrm{d}\bar{z}]$.

Physical picture: $H^{0,1}(T)$ is the space of Wilson lines

Take a flat $U(1)$ connection on T with constant potential (in our anti-Hermitian convention A is i times a real 1-form, $A = -iA_{\text{phys}}$): $A = 2\pi i(\theta_1\,\mathrm{d}s + \theta_2\,\mathrm{d}t)$, whose holonomies round the two cycles are $e^{\pm 2\pi i\theta_1}$ and $e^{\pm 2\pi i\theta_2}$ (the sign depends on the direction of transport). Using Step 1, $\mathrm{d}s = (\tau\mathrm{d}\bar{z} - \bar{\tau}\mathrm{d}z)/(\tau - \bar{\tau})$ and $\mathrm{d}t = (\mathrm{d}z - \mathrm{d}\bar{z})/(\tau - \bar{\tau})$, so its $(0,1)$ part is

$$A^{0,1} = 2\pi i\,\frac{\theta_1\tau - \theta_2}{\tau - \bar{\tau}}\,\mathrm{d}\bar{z} = \lambda_\theta\,\mathrm{d}\bar{z}, \qquad \lambda_\theta = \frac{\pi(\theta_1\tau - \theta_2)}{\tau_2},$$

the *same* formula as (3.78) with θ in place of k. The operator $\bar{\partial}_A = \bar{\partial} + A^{0,1}$ defines a holomorphic structure, and $A^{0,1}$ is a class in $H^{0,1}(T) \cong \mathbb{C}$. A large gauge transformation $g = e_k$ shifts A by $g^{-1}\mathrm{d}g = 2\pi i(k_1\mathrm{d}s + k_2\mathrm{d}t)$, i.e. $\theta \mapsto \theta + k$, and shifts $A^{0,1}$ by $\lambda_k\mathrm{d}\bar{z}$. So the gauge-inequivalent flat connections form $\mathbb{C}/\{\lambda_k\}$, a torus dual to T: the *Jacobian* of T, or in physics language the moduli space of Wilson lines.

Try it: Solve $\partial_{\bar{z}} f = \cos 2\pi s$ on the square torus

Take $\tau = i$. Find a smooth f on T with $\partial_{\bar{z}} f = \cos 2\pi s$. Why is there no solution for $\partial_{\bar{z}} f = 1 + \cos 2\pi s$?

Answer. $\cos 2\pi s = \frac{1}{2}(e_{(1,0)} + e_{(-1,0)})$ and $\lambda_{(\pm 1,0)} = \pm\pi i$, so $f = \frac{1}{2\pi i}(e_{(1,0)} - e_{(-1,0)}) = \frac{\sin 2\pi s}{\pi}$.

Check: $\partial_{\bar z} s = \tau/(\tau-\bar\tau) = \frac12$, so $\partial_{\bar z} f = 2\cos(2\pi s)\cdot\frac12$. The source $1+\cos 2\pi s$ has average $g_0 = 1 \neq 0$, the nonzero class in $H^{0,1}(T)$.

Worked example 3.34: $\mathbb{CP}^1$: the Fubini–Study form and the residue

On the Riemann sphere, with charts z on $U_0 = \mathbb{C}$ and $w = 1/z$ on U_1, we know from Čech cohomology (Example 3.18, cases $n = 0$ and $n = -2$): $H^1(\mathbb{CP}^1; \mathcal{O}) = 0$ and $H^1(\mathbb{CP}^1; \Omega^1) \cong \mathbb{C}$. By the Dolbeault theorem, therefore,

$$H^{0,1}_{\bar\partial}(\mathbb{CP}^1) = 0, \qquad H^{1,1}_{\bar\partial}(\mathbb{CP}^1) \cong \mathbb{C}.$$

The first says: every $(0,1)$-form on the sphere is $\bar\partial$ of a global function. Let us see the second concretely, and watch the Dolbeault class turn into the Čech class.

A nonzero class. The **Fubini–Study form**

$$\omega = \frac{i}{2\pi}\,\frac{\mathrm{d}z\wedge\mathrm{d}\bar z}{(1+|z|^2)^2} = \frac{i}{2\pi}\,\partial\bar\partial\log(1+|z|^2) \tag{3.79}$$

is a $(1,1)$-form on U_0 (the second expression: $\partial_{\bar z}\log(1+z\bar z) = z/(1+z\bar z)$ and ∂_z of that is $1/(1+z\bar z)^2$). It is automatically $\bar\partial$-closed, since there are no $(1,2)$-forms. It extends smoothly over $z=\infty$: with $z = 1/w$, $\mathrm{d}z\wedge\mathrm{d}\bar z = \mathrm{d}w\wedge\mathrm{d}\bar w/|w|^4$ and $(1+|z|^2)^2 = (1+|w|^2)^2/|w|^4$, so ω has the same form in w. Its integral is

$$\int_{\mathbb{CP}^1}\omega \overset{(1)}{=} \frac{i}{2\pi}\int_{\mathbb{C}}\frac{-2i\,\mathrm{d}x\,\mathrm{d}y}{(1+r^2)^2} \overset{(2)}{=} \frac{1}{\pi}\int_0^{2\pi}\!\!\int_0^{\infty}\frac{r\,\mathrm{d}r\,\mathrm{d}\phi}{(1+r^2)^2} \overset{(3)}{=} \frac{1}{\pi}\cdot 2\pi\cdot\frac12 = 1.$$

(1) The point ∞ has measure zero, and $\mathrm{d}z\wedge\mathrm{d}\bar z = -2i\,\mathrm{d}x\wedge\mathrm{d}y$ (Example 3.29(a)).
(2) $\frac{i}{2\pi}(-2i) = \frac{1}{\pi}$; polar coordinates.
(3) $\int_0^\infty \frac{r\,\mathrm{d}r}{(1+r^2)^2} = \left[-\frac{1}{2(1+r^2)}\right]_0^\infty = \frac12$.

If ω were $\bar\partial\beta$ for a global $(1,0)$-form β, then $\omega = \mathrm{d}\beta$ (because $\partial\beta$ would be of type $(2,0)$, which is zero in one dimension), and Stokes would give $\int\omega = 0$. So $[\omega] \neq 0$ in $H^{1,1}_{\bar\partial}(\mathbb{CP}^1)$. Up to normalisation, ω is the round area form $\frac{1}{4\pi}\sin\theta\,\mathrm{d}\theta\wedge\mathrm{d}\phi$ pulled back by stereographic projection, and the curvature form $\frac{i}{2\pi}F$ of the unit monopole.

The zig-zag to Čech. Now run the connecting map of the Dolbeault resolution, as in Section 3.6.4. First solve $\bar\partial\alpha_i = \omega$ on each patch. Try $\alpha_0 = h\,\mathrm{d}z$; since $\bar\partial(h\,\mathrm{d}z) = -\partial_{\bar z}h\,\mathrm{d}z\wedge\mathrm{d}\bar z$, we need $\partial_{\bar z}h = -\frac{i}{2\pi}(1+z\bar z)^{-2}$, and $h = -\frac{i}{2\pi}\,\bar z/(1+z\bar z)$ works (quotient rule: $\partial_{\bar z}\frac{\bar z}{1+z\bar z} = \frac{(1+z\bar z)-\bar z z}{(1+z\bar z)^2}$). By the $z\leftrightarrow w$ symmetry,

$$\alpha_0 = -\frac{i}{2\pi}\,\frac{\bar z\,\mathrm{d}z}{1+|z|^2}\ \text{on } U_0, \qquad \alpha_1 = -\frac{i}{2\pi}\,\frac{\bar w\,\mathrm{d}w}{1+|w|^2}\ \text{on } U_1.$$

On the overlap $\mathbb{C}^*$ the difference is $\bar\partial$-closed (both have $\bar\partial = \omega$), hence a holomorphic 1-form. Compute it in the z chart:

$$\alpha_1 \overset{(1)}{=} -\frac{i}{2\pi}\,\frac{\bar z^{-1}(-z^{-2}\mathrm{d}z)}{1+|z|^{-2}} \overset{(2)}{=} \frac{i}{2\pi}\,\frac{\mathrm{d}z}{z\,(1+|z|^2)},$$
$$\alpha_0 - \alpha_1 \overset{(3)}{=} -\frac{i}{2\pi}\,\frac{\mathrm{d}z}{1+|z|^2}\left(\bar z + \frac1z\right) \overset{(4)}{=} -\frac{i}{2\pi}\,\frac{\mathrm{d}z}{z}.$$

(1) $w = 1/z$, $\bar w = 1/\bar z$, $\mathrm{d}w = -z^{-2}\mathrm{d}z$.

(2) Multiply numerator and denominator by $|z|^2 = z\bar{z}$.
(3) Subtract, taking out the common factor.
(4) $\bar{z} + 1/z = (1 + z\bar{z})/z$.

The Čech 1-cocycle representing $[\omega]$ in $H^1(\mathbb{CP}^1;\Omega^1)$ is $c_{01} = -\frac{i}{2\pi}\frac{dz}{z}$: a multiple of $z^{-1}dz$, the "residue" monomial that Example 3.18 (case $n = -2$) found to span $H^1(\mathbb{CP}^1;\Omega^1)$. And the numbers match. With $D_0 = \{|z| \le 1\}$, $D_1 = \{|z| \ge 1\}$ and $d\alpha_i = \bar{\partial}\alpha_i = \omega$,

$$1 = \int_{\mathbb{CP}^1} \omega \overset{(1)}{=} \int_{D_0} d\alpha_0 + \int_{D_1} d\alpha_1 \overset{(2)}{=} \oint_{|z|=1} \alpha_0 - \oint_{|z|=1} \alpha_1 \overset{(3)}{=} \oint_{|z|=1} c_{01} \overset{(4)}{=} -\frac{i}{2\pi}\cdot 2\pi i = 1.$$

(1) Split the sphere along the unit circle; on each half $\omega = d\alpha_i$.
(2) Stokes; the unit circle is the boundary of D_0 counterclockwise and of D_1 clockwise.
(3) $c_{01} = \alpha_0 - \alpha_1$.
(4) $\oint dz/z = 2\pi i$.

The Dolbeault class is detected by $\int \omega$, the Čech class by the residue of c_{01}, and Stokes' theorem says they are the same number. This is the monopole flux computation of Example 3.27 in holomorphic dress.

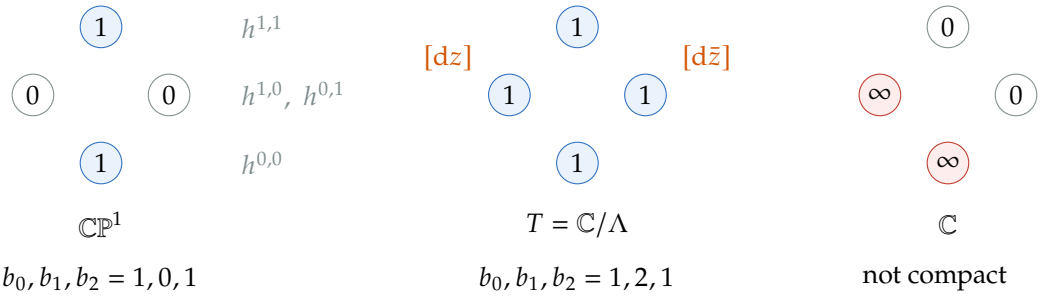


Figure 3.28: Hodge diamonds: the numbers $h^{p,q} = \dim H^{p,q}_{\bar{\partial}}$ arranged with $h^{0,0}$ at the bottom, $h^{1,1}$ at the top, and $h^{1,0}$ (left), $h^{0,1}$ (right) in the middle row. For the compact surfaces each horizontal row adds up to a Betti number: $b_1(T) = h^{1,0} + h^{0,1} = 2$ from $[dz]$ and $[d\bar{z}]$. For the non-compact $\mathbb{C}$ the holomorphic functions and holomorphic 1-forms are infinite-dimensional ($h^{0,0} = h^{1,0} = \infty$, red), while the Cauchy kernel kills $H^{0,1}$ and $H^{1,1}$.

Watch out: Dolbeault cohomology is not topological

Unlike de Rham cohomology, $H^{p,q}_{\bar{\partial}}$ depends on the complex structure, not only on the underlying manifold. On a non-compact manifold it can be infinite-dimensional: on $\mathbb{C}$, $H^{0,0}_{\bar{\partial}}(\mathbb{C})$ is the space of all entire functions. On compact manifolds it is finite-dimensional (a consequence of ellipticity, Chapter 2), and only certain combinations, such as $\sum_q (-1)^q h^{p,q}$, are topological invariants. That alternating sum is exactly the kind of number the index theorem computes.

Reading Nash (pp. 63–65, (3.24)–(3.31))

After the de Rham theorem (3.23) Nash introduces complex forms. His (3.24) is our (3.65) (without the $1/(p!\,q!)$), his (3.25) the operators (3.67), and his (3.26) is our key idea on $d = \partial + \bar{\partial}$; we added the one-line proof "sort $d^2 = 0$ by type". His (3.27) is the type decomposition (3.66); he states that $\mathcal{E}^{p,q}$ is fine, which we justified (real partitions of unity preserve type). His (3.28) is (3.74); its exactness rests on the Dolbeault lemma, which

he does not state and we proved with the Cauchy kernel (Section 3.7.4). His "exactly similar inductive argument" is (3.77), his (3.29)–(3.30) are (3.76), and his theorem (3.31) is the Dolbeault theorem. Two slips (see the pitfall in Section 3.7.3): holomorphic p-forms are of type $(p,0)$, not $(0,p)$; and the first arrow in (3.28) is the inclusion. The examples ($\mathbb{C}$, the torus, $\mathbb{CP}^1$ and the Fubini–Study form), the Dirac operator dictionary and the bundle-valued version are ours. Nash says that "Dolbeault's theorem will be of use later"; it is, in Chapter 4 and Chapter 5, where the index of $\bar\partial$ becomes Riemann–Roch.

Recap

- Complex coordinates: $\mathrm{d}z = \mathrm{d}x + i\mathrm{d}y$, $\partial_{\bar z} = \frac{1}{2}(\partial_x + i\partial_y)$; $\mathrm{d}f = \partial_z f\,\mathrm{d}z + \partial_{\bar z} f\,\mathrm{d}\bar z$; holomorphic $\iff \partial_{\bar z} f = 0 \iff$ Cauchy–Riemann; $4\partial_z\partial_{\bar z} = \nabla^2$.
- On a complex manifold forms split by type, $\Lambda^k \otimes \mathbb{C} = \bigoplus_{p+q=k} \Lambda^{p,q}$, because transition functions are holomorphic; $\mathrm{d} = \partial + \bar\partial$ and $\partial^2 = \bar\partial^2 = \partial\bar\partial + \bar\partial\partial = 0$. Holomorphic p-forms: $\Omega^p = \operatorname{Ker}\bar\partial$ on $(p,0)$-forms.
- Dolbeault lemma: locally, $\bar\partial\omega = 0 \Rightarrow \omega = \bar\partial\beta$. In one variable $\partial_{\bar z}^{-1}$ is convolution with $1/(\pi z)$, because $\partial_{\bar z}(1/\pi z) = \delta^2$ (Stokes).
- Dolbeault theorem: $H^q(M;\Omega^p) \cong H^{p,q}_{\bar\partial}(M)$, by Nash's de Rham argument with $\bar\partial$ for d. Also with coefficients in a holomorphic bundle.
- Examples: open $U \subset \mathbb{C}$ have $H^1(U;\mathcal{O}) = 0$ (so they are Leray and carry only trivial holomorphic line bundles); the torus has $h^{p,q} = 1$ for all $p, q \le 1$, with $H^{0,1}(T) = \mathbb{C}[\mathrm{d}\bar z]$ the space of Wilson lines; $\mathbb{CP}^1$ has $H^{1,1} = \mathbb{C}[\omega_{FS}]$, and the zig-zag turns ω_{FS} into the residue cocycle $-\frac{i}{2\pi}\frac{\mathrm{d}z}{z}$.
- Physics: $\not{D}_{2d} = \begin{pmatrix} 0 & 2\partial_z \\ 2\partial_{\bar z} & 0 \end{pmatrix}$, so Dolbeault cohomology counts chiral zero modes, and $\dim H^0 - \dim H^1$ is an index.

3.8 K-theory: why count bundles, and what we count with

In the previous sections we measured the shape of a manifold M by counting *forms*: closed forms modulo exact ones gave the de Rham groups, and sheaf cohomology generalised that. Now we switch the object being counted. We want to count *vector bundles* over M.

Why would a physicist care? Because a gauge field is a connection on a vector bundle, and a matter field is a section of one. Two field configurations that live on non-isomorphic bundles can never be deformed into each other: they sit in different *topological sectors*. The Dirac monopole of charge 1 and the monopole of charge 2 live on different line bundles over the sphere at infinity. The BPST instanton and the trivial vacuum live on different SU(2) bundles over S^4. So the question "how many different bundles are there over M?" is the question "how many topological sectors does a gauge theory on M have?".

That question is hard in general. K-theory answers an easier, coarser version of it, and the coarse version turns out to be both computable and exactly what the index theorem (Chapter 4) needs. Nash puts it in one line:

Key idea: K-theory in one sentence

K-theory is the linear algebra of *very large matrices*: it keeps the information about a vector bundle that survives when we are allowed to add as many trivial directions as we like. In return for that coarsening, bundles acquire an addition, a subtraction and a multiplication, just like numbers.

Before we can do any of that we need the raw material: vector bundles, written down concretely by transition functions, and the ways of combining them. This section supplies that background, which Nash assumes. Readers who know it well can skim to Section 3.9.

Standing conventions for this part.

- M (or X) is a *compact* manifold, or more generally a compact Hausdorff space. Nash says "Hausdorff" in a footnote; we really need compactness (see Section 3.9).
- Unless stated otherwise, M is *connected*, so every bundle has a single rank.
- $\mathbb{F}$ stands for $\mathbb{R}$, $\mathbb{C}$ or $\mathbb{H}$ (real numbers, complex numbers, quaternions). A bundle with fibre $\mathbb{F}^k$ is a real, complex or quaternionic bundle of **rank** k.
- I^j denotes the **trivial bundle** of rank j: $I^j = M \times \mathbb{R}^j$ (real case) or $I^j = M \times \mathbb{C}^j$ (complex case). We write $I = I^1$ and $I^0 = M \times \{0\}$.
- $\mathrm{Vect}_k(M)$ is the set of isomorphism classes of rank-k bundles over M, and $\mathrm{Vect}(M)$ is the set of isomorphism classes of all bundles. When the field matters we write $\mathrm{Vect}_k(M, \mathbb{F})$.

3.8.1 Vector bundles by transition functions

A vector bundle over M is a family of vector spaces E_x, one for each point $x \in M$, glued together smoothly. Locally it always looks like a product $U \times \mathbb{F}^k$. All the interesting information is in *how the local products are glued*. The gluing data are matrices.

Working definition 3.22 (vector bundle via transition functions)

To specify a rank-k bundle E over M:

1. Cover M by open sets U_a ("patches").
2. Over each patch, use a fibre coordinate $v_a \in \mathbb{F}^k$: a point of E over $x \in U_a$ is a pair (x, v_a).
3. On each overlap $U_a \cap U_b$ give a matrix-valued function $g_{ab} : U_a \cap U_b \to \mathrm{GL}(k, \mathbb{F})$ [invertible $k \times k$ matrices with entries in $\mathbb{F}$], the **transition function**, saying how to convert fibre coordinates: $v_a = g_{ab}(x)\, v_b$.
4. Consistency on triple overlaps (the **cocycle condition**): $g_{ab}\, g_{bc} = g_{ac}$ on $U_a \cap U_b \cap U_c$, and $g_{aa} = 1$.

Two sets of transition functions $\{g_{ab}\}$ and $\{g'_{ab}\}$ describe **isomorphic** bundles, $E \cong E'$, if there are functions $h_a : U_a \to \mathrm{GL}(k, \mathbb{F})$ (a change of fibre coordinates on each patch, a "gauge transformation") with

$$g'_{ab} = h_a\, g_{ab}\, h_b^{-1}.$$

Formally: a vector bundle is a space E with a projection $\pi : E \to M$ whose fibres [the sets $E_x = \pi^{-1}(x)$ sitting over single points x] are vector spaces and which is locally isomorphic to $U \times \mathbb{F}^k$; the g_{ab} are the comparison maps between local isomorphisms.

The cocycle condition simply says that converting from c to b and then from b to a is the same as converting directly from c to a. The isomorphism rule says that we may re-choose fibre coordinates patch by patch: $v'_a = h_a v_a$. Let us check that this really produces the stated rule for g':

$$\begin{aligned} v'_a &\overset{(1)}{=} h_a\, v_a \\ &\overset{(2)}{=} h_a\, g_{ab}\, v_b \\ &\overset{(3)}{=} h_a\, g_{ab}\, h_b^{-1}\, v'_b. \end{aligned}$$

(1) Definition of the new fibre coordinate on U_a.
(2) Old transition rule $v_a = g_{ab} v_b$.
(3) Invert the new coordinate on U_b: $v_b = h_b^{-1} v'_b$.

So the new transition function is $g'_{ab} = h_a g_{ab} h_b^{-1}$, as claimed.

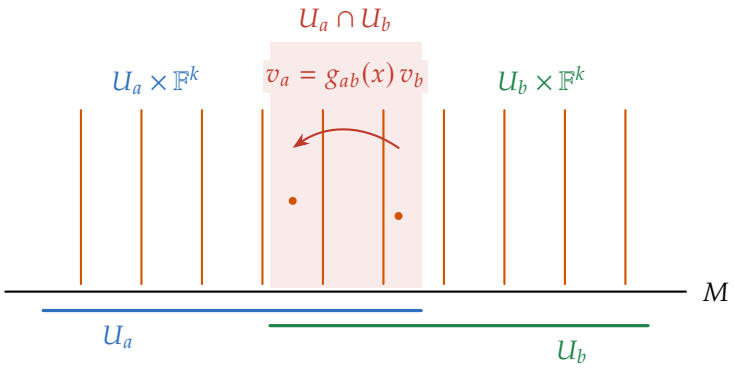


Figure 3.29: A vector bundle drawn as fibres (orange) standing over the base M. Over each patch the bundle is a product. On the overlap (shaded) the same point of the bundle has two fibre coordinates, related by the matrix $g_{ab}(x)$. All of the topology of E lives in these matrices.

Two quick consequences are worth recording.

- **Trivial bundle.** I^k is described by a single patch $U = M$, or by any cover with all $g_{ab} = 1$. A bundle is **trivial** if it is isomorphic to I^k, i.e. if we can find h_a with $h_a g_{ab} h_b^{-1} = 1$, i.e. $g_{ab} = h_a^{-1} h_b$.
- **Sections.** A **section** s of E is a choice of vector $s(x) \in E_x$ for every x, depending smoothly on x. In coordinates it is a collection of functions $s_a : U_a \to \mathbb{F}^k$ with $s_a = g_{ab} s_b$ on overlaps. Physically: a matter field in a background gauge field is a section.

Key idea: A nonvanishing section trivialises a line bundle

If a line bundle L (rank 1) has a section s with $s(x) \neq 0$ for every x, then $L \cong I$.

Why: in each patch $s_a(x)$ is a nonzero number. Use it as the change of fibre coordinate, $h_a = s_a^{-1}$, so that the new fibre coordinate of s is 1 on every patch. Then

$$\begin{aligned} g'_{ab} &\overset{(1)}{=} h_a\, g_{ab}\, h_b^{-1} \\ &\overset{(2)}{=} s_a^{-1}\, g_{ab}\, s_b \\ &\overset{(3)}{=} s_a^{-1}\, s_a = 1. \end{aligned}$$

(1) The isomorphism rule from the working definition.
(2) Our choice $h_a = s_a^{-1}$ (a nonzero number, so invertible).
(3) s is a section, so $g_{ab}s_b = s_a$ (for rank 1 everything commutes).

All transition functions become 1: the bundle is trivial. The same argument with k pointwise linearly independent sections (a **frame**) shows that a rank-k bundle with a global frame is trivial; the h_a is then the inverse of the matrix whose columns are the k sections.

Three bundles we will use over and over

Worked example 3.35: The Möbius bundle over S^1

Cover the circle $S^1 = \{e^{i\theta}\}$ by two arcs U_1 and U_2 which overlap in two small intervals, one near $\theta = 0$ and one near $\theta = \pi$. Define a real line bundle $\mathcal{M}$ by

$$g_{12} = +1 \text{ near } \theta = 0, \qquad g_{12} = -1 \text{ near } \theta = \pi.$$

There are no triple overlaps, so the cocycle condition is automatic. Is $\mathcal{M}$ trivial? We would need nonzero functions h_1, h_2 with $g_{12} = h_1^{-1}h_2$. Follow h_2/h_1 around the circle:

$$\operatorname{sgn}(h_2/h_1)\big|_{\theta\approx 0} \overset{(1)}{=} \operatorname{sgn}(g_{12})\big|_{\theta\approx 0} = +1,$$
$$\operatorname{sgn}(h_2/h_1)\big|_{\theta\approx \pi} \overset{(1)}{=} \operatorname{sgn}(g_{12})\big|_{\theta\approx \pi} = -1.$$

(1) Triviality would require $g_{12} = h_1^{-1}h_2$ on each overlap interval.

But h_1 is a nonzero real function on the connected arc U_1, so it has a fixed sign on all of U_1; the same holds for h_2 on U_2. Then $\operatorname{sgn}(h_2/h_1) = \operatorname{sgn}(h_2)\operatorname{sgn}(h_1)$ is the same number on both overlap intervals. Contradiction. So $\mathcal{M}$ is *not* trivial. The total space is the Möbius band (Figure 3.30). Equivalently: every section of $\mathcal{M}$ must change sign somewhere, so it has a zero.

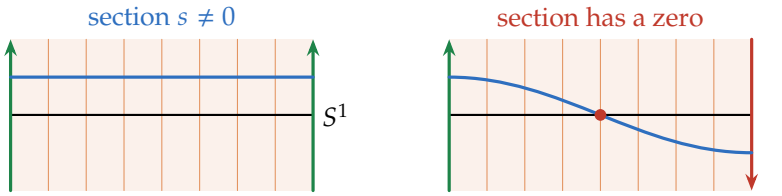


Figure 3.30: Two real line bundles over the circle. On the left every fibre is glued to its neighbour with +1. On the right one of the two gluings is −1 (red arrow reversed), so the fibre comes back upside down after one trip around the base. A section (blue curve crossing every fibre once) that starts above the zero section must end below it, so it vanishes somewhere (red dot).

Worked example 3.36: The tautological (Hopf) line bundle H over $\mathbb{CP}^1 = S^2$

The complex projective line $\mathbb{CP}^1$ is the set of complex lines through the origin in $\mathbb{C}^2$. A line is named by any nonzero vector on it, $[z_0 : z_1]$, with $[z_0 : z_1] = [\lambda z_0 : \lambda z_1]$ for $\lambda \in \mathbb{C}^\times$. Two patches cover it:

$$U_0 = \{z_0 \neq 0\} \text{ with coordinate } z = z_1/z_0, \qquad U_1 = \{z_1 \neq 0\} \text{ with coordinate } w = z_0/z_1.$$

On the overlap, $w = 1/z$. Topologically $\mathbb{CP}^1$ is the Riemann sphere $\mathbb{C} \cup \{\infty\} = S^2$, with z the stereographic coordinate.
The **tautological line bundle** H has as its fibre over the point $[z_0 : z_1]$ *the line itself*, $\{\lambda(z_0, z_1)\} \subset \mathbb{C}^2$. Nash calls it η, "the canonical line bundle over $\mathbb{CP}^1$". We find its transition function by choosing a nonvanishing local section on each patch:

$$s_0(z) = (1, z) \text{ on } U_0, \qquad s_1(w) = (w, 1) \text{ on } U_1.$$

Each is a nonzero vector lying on the line in question. On the overlap,

$$\begin{aligned} s_0 &\overset{(1)}{=} (1, z) \\ &\overset{(2)}{=} z\,(1/z,\, 1) \\ &\overset{(3)}{=} z\, s_1. \end{aligned}$$

(1) Definition of s_0.
(2) Pull out the factor z, which is allowed because $z \neq 0$ on the overlap.
(3) $1/z = w$, so $(1/z, 1) = (w, 1) = s_1$.

A point of H over the overlap can be written $\lambda_0 s_0 = \lambda_1 s_1$; using $s_0 = z s_1$ gives $\lambda_1 = z\lambda_0$. So the fibre coordinates are related by

$$\boxed{\lambda_1 = g_{10}(z)\,\lambda_0, \qquad g_{10}(z) = z} \qquad (z \in U_0 \cap U_1 = \mathbb{C}^\times).$$

On the equator $|z| = 1$, $z = e^{i\phi}$: the transition function winds *once* around U(1) as we go once around the equator. That winding number is the topological charge of H.

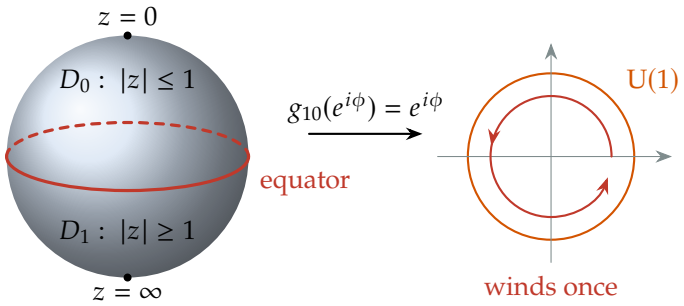


Figure 3.31: The Riemann sphere split into two discs along the equator. Over each disc the Hopf bundle H is trivial. The gluing along the equator is the map $e^{i\phi} \mapsto e^{i\phi}$ into U(1), which winds once. The winding number of the gluing map is the topological invariant of the bundle. Later (Section 3.12) this is called a clutching function.

Worked example 3.37: The tangent bundle of S^2 as a complex line bundle

A tangent vector to the Riemann sphere at a point of U_0 can be written $v = v^z\,\partial_z$, and at a point of U_1 as $v = v^w\,\partial_w$. (Because S^2 is a complex curve, TS^2 is a complex line bundle: a real rank-2 bundle on which rotation by 90° acts as multiplication by i.) The chain rule relates the two:

$$v^z\,\partial_z \overset{(1)}{=} v^z\,\frac{\mathrm{d}w}{\mathrm{d}z}\,\partial_w$$

$$\overset{(2)}{=} v^z \left(-z^{-2}\right) \partial_w ,$$

(1) Chain rule: $\partial_z = (\mathrm{d}w/\mathrm{d}z)\,\partial_w$.
(2) $w = 1/z$, so $\mathrm{d}w/\mathrm{d}z = -1/z^2$.

so $v^w = -z^{-2}\, v^z$ and the transition function is $g_{10}(z) = -z^{-2}$. On the equator it is $-e^{-2i\phi}$, which winds *twice*, in the opposite sense to H. The constant -1 can be removed by a homotopy (rotate it to 1 inside U(1)), so up to isomorphism TS^2 has transition function z^{-2}. Comparing with H (transition z) this says $TS^2 \cong (H^*)^{\otimes 2}$, once we have defined $\otimes$ and duals below. The "2" is the Euler characteristic $\chi(S^2) = 2$.

Watch out: Signs and names for the Hopf bundle

Different books call H the tautological bundle, the Hopf bundle, the canonical bundle, or the monopole bundle, and some use its dual H^* (transition z^{-1}) under the same name. With the complex orientation of S^2, the tautological bundle has first Chern class $c_1(H) = -1$ and H^* (the "hyperplane bundle") has $c_1 = +1$. For everything in this part the distinction is harmless: H and H^* both generate the K-theory of S^2. We keep the name H for the tautological bundle and write Nash's η for the same thing when quoting him.

Physical picture: The Dirac monopole is a bundle

Put a magnetic monopole of charge n at the origin and look at the sphere around it. Use the anti-Hermitian convention of these notes: a U(1) gauge field is $A = -i\,eA_\mu^{\text{phys}}\,\mathrm{d}x^\mu$, and a gauge transformation $g = e^{i\chi}$ acts as $A \mapsto A + g^{-1}\mathrm{d}g$. The potentials

$$A_N = -\tfrac{in}{2}(1-\cos\theta)\,\mathrm{d}\phi, \qquad A_S = +\tfrac{in}{2}(1+\cos\theta)\,\mathrm{d}\phi$$

are regular on the northern and southern hemispheres respectively (the coefficient of $\mathrm{d}\phi$ vanishes at the pole where $\mathrm{d}\phi$ is singular). On the overlap

$$A_S - A_N = \tfrac{in}{2}\big[(1+\cos\theta)+(1-\cos\theta)\big]\mathrm{d}\phi = in\,\mathrm{d}\phi = g^{-1}\mathrm{d}g, \qquad g = e^{in\phi}.$$

So the two patches are glued by the gauge transformation $e^{in\phi}$. It is single-valued on the equator only if $n \in \mathbb{Z}$: that is Dirac's quantisation condition. With the rule $A \mapsto g^{-1}Ag + g^{-1}\mathrm{d}g$, a charged field transforms as $\psi \mapsto g^{-1}\psi$, so $\psi_S = e^{-in\phi}\psi_N$: the fibre-coordinate transition function is $e^{-in\phi}$, which winds n times (clockwise). With these conventions $\frac{i}{2\pi}\int_{S^2} F = n$, and comparing with H (whose fibre coordinates obey $\lambda_1 = e^{+i\phi}\lambda_0$) the charge-$n$ monopole bundle is $(H^*)^{\otimes n}$.

3.8.2 Operations on bundles: sum, product, dual, pullback

Everything one can do to vector spaces one can do to bundles, fibre by fibre. At the level of transition functions each operation is an operation on matrices. Suppose E has rank k and transition functions g_{ab}, and F has rank l and transition functions f_{ab} (take a common cover for both; one can always refine the covers until they agree).

Working definition 3.23 (operations on bundles)

- **Direct (Whitney) sum** $E \oplus F$: fibre $E_x \oplus F_x$, rank $k + l$, transition functions the

block-diagonal matrices

$$(g \oplus f)_{ab} = \begin{pmatrix} g_{ab} & 0 \\ 0 & f_{ab} \end{pmatrix}.$$

- **Tensor product** $E \otimes F$: fibre $E_x \otimes F_x$, rank kl, transition functions the Kronecker products [the $kl \times kl$ matrices sending $v \otimes w$ to $g_{ab}v \otimes f_{ab}w$] $(g \otimes f)_{ab} = g_{ab} \otimes f_{ab}$. For line bundles this is just the product of the two numbers.
- **Dual** E^*: fibre E_x^* (linear functionals), rank k, transition functions $\left(g_{ab}^T\right)^{-1}$. For a U(1) line bundle this is complex conjugation: $e^{i\chi} \mapsto e^{-i\chi}$.
- **Pullback** f^*E along a map $f : N \to M$: the fibre over $y \in N$ is $E_{f(y)}$. Transition functions $g_{ab} \circ f$ on the cover $f^{-1}(U_a)$ [the points of N that f sends into U_a] of N.

Let us check that the block-diagonal and Kronecker matrices really satisfy the cocycle condition, so these are genuine bundles:

$$(g \oplus f)_{ab}(g \oplus f)_{bc} \overset{(1)}{=} \begin{pmatrix} g_{ab}g_{bc} & 0 \\ 0 & f_{ab}f_{bc} \end{pmatrix} \overset{(2)}{=} \begin{pmatrix} g_{ac} & 0 \\ 0 & f_{ac} \end{pmatrix} = (g \oplus f)_{ac},$$

$$(g \otimes f)_{ab}(g \otimes f)_{bc} \overset{(3)}{=} (g_{ab}g_{bc}) \otimes (f_{ab}f_{bc}) \overset{(2)}{=} g_{ac} \otimes f_{ac} = (g \otimes f)_{ac}.$$

(1) Block-diagonal matrices multiply block by block.
(2) The cocycle conditions for E and F separately.
(3) The mixed-product rule for Kronecker products, $(A \otimes B)(C \otimes D) = AC \otimes BD$.

For the dual, $\left((g_{ab}g_{bc})^T\right)^{-1} = (g_{ab}^T)^{-1}(g_{bc}^T)^{-1}$ because transposing reverses the order and inverting reverses it again.

Worked example 3.38: Powers of the Hopf bundle

Using $g_{10} = z$ for H:

$$\begin{aligned} H \otimes H: &\quad g_{10} = z \cdot z = z^2, \\ H^*: &\quad g_{10} = (z)^{-1} = z^{-1}, \\ H \otimes H^*: &\quad g_{10} \overset{(1)}{=} z \cdot z^{-1} = 1 \implies H \otimes H^* \cong I, \\ H \oplus H: &\quad g_{10} = \begin{pmatrix} z & 0 \\ 0 & z \end{pmatrix}. \end{aligned}$$

(1) Tensor product of line bundles multiplies transition functions; the result is identically 1, which is the trivial bundle.

In general $H^{\otimes n}$ has transition function z^n for any $n \in \mathbb{Z}$ (negative powers meaning powers of H^*). So the line bundles over S^2 form a group under $\otimes$, isomorphic to $\mathbb{Z}$ via the winding number; we justify "every line bundle is one of these" in Section 3.12. By Example 3.37, $TS^2 \cong H^{\otimes(-2)}$.

Worked example 3.39: Möbius ⊗ Möbius is trivial

For $\mathcal{M}$ the transition function is ± 1, so for $\mathcal{M} \otimes \mathcal{M}$ it is $(\pm 1)^2 = +1$ on both overlap intervals. Hence $\mathcal{M} \otimes \mathcal{M} \cong I$. Geometrically: flipping twice is not flipping.

The operations obey the rules one expects from arithmetic, because they hold fibre by fibre

for vector spaces:

$$E \oplus F \cong F \oplus E, \qquad (E \oplus F) \oplus G \cong E \oplus (F \oplus G), \qquad E \oplus I^0 \cong E,$$
$$E \otimes F \cong F \otimes E, \qquad E \otimes (F \oplus G) \cong (E \otimes F) \oplus (E \otimes G), \qquad E \otimes I \cong E,$$

and ranks behave like numbers: $\operatorname{rk}(E \oplus F) = \operatorname{rk} E + \operatorname{rk} F$, $\operatorname{rk}(E \otimes F) = \operatorname{rk} E \cdot \operatorname{rk} F$. Pullback respects everything: $f^*(E \oplus F) \cong f^*E \oplus f^*F$ and $f^*(E \otimes F) \cong f^*E \otimes f^*F$, because composing with f commutes with forming block-diagonal or Kronecker matrices.

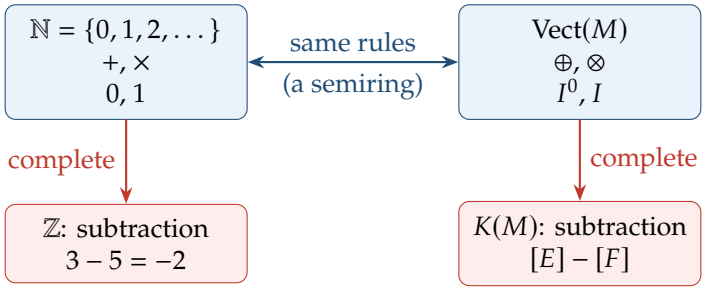


Figure 3.32: The analogy that drives K-theory. Bundles under $\oplus$ and $\otimes$ behave like natural numbers under $+$ and $\times$, with the zero bundle and the trivial line bundle playing 0 and 1. What is missing is subtraction. Completing $\mathbb{N}$ to $\mathbb{Z}$ supplies it for numbers; the same construction applied to bundles gives $K(M)$ (Section 3.10).

Worked example 3.40: Pullback of H along a degree-n map

Let $f_n : S^2 \to S^2$ be $f_n(z) = z^n$ ($n \geq 1$). It maps $U_0 = \{z \neq \infty\}$ into U_0 and U_1 into U_1. The pulled-back transition function is

$$(f_n^* g)_{10}(z) \overset{(1)}{=} g_{10}\big(f_n(z)\big) \overset{(2)}{=} z^n,$$

(1) Definition of pullback: compose the transition function with f.
(2) $g_{10}(u) = u$ for H, evaluated at $u = z^n$.

so $f_n^* H \cong H^{\otimes n}$: pulling back along a map of degree n multiplies the charge by n.

3.8.3 What "generalised cohomology theory" means

Nash calls K-theory "probably the most important example of a *generalised* cohomology theory" and says the adjective refers to the failure of one of the Eilenberg–Steenrod axioms. Here is what that means, at the level we need.

A cohomology theory assigns abelian groups $h^n(X)$ to spaces X, for each integer n, and to each map $f : X \to Y$ a *pullback* $f^* : h^n(Y) \to h^n(X)$ (note the reversed direction: this is the **contravariance** we already know from pulling back forms). Eilenberg and Steenrod listed the properties that the ordinary theories (de Rham, singular, Čech) share:

1. **Homotopy.** If $f \simeq g$ then $f^* = g^*$. (For de Rham: homotopic maps pull back closed forms to forms differing by an exact form.)
2. **Exactness.** A pair $A \subset X$ gives a long exact sequence $\cdots \to h^n(X, A) \to h^n(X) \to h^n(A) \to h^{n+1}(X, A) \to \cdots$.
3. **Excision.** Cutting out a small piece deep inside A does not change $h^n(X, A)$. (This is the formal version of "locality": cohomology can be computed patch by patch, as in Mayer–Vietoris.)

4. **Additivity.** h^n of a disjoint union is the product of the h^n's.

5. **Dimension.** For a single point, $h^0(\mathrm{pt}) = \mathbb{Z}$ (or $\mathbb{R}$) and $h^n(\mathrm{pt}) = 0$ for $n \neq 0$.

The dimension axiom says: "a point has no cohomology except the constants in degree zero". The de Rham theory satisfies it: on a point, the only forms are 0-forms (numbers), so $H^0_{\mathrm{dR}}(\mathrm{pt}) = \mathbb{R}$ and nothing else.

Key idea: Why K-theory is "generalised"

K-theory satisfies axioms 1–4 but *not* the dimension axiom. When we have built the groups K^{-n} in Section 3.13, we will find

$$K^{-n}(\mathrm{pt}) = \begin{cases} \mathbb{Z}, & n \text{ even}, \\ 0, & n \text{ odd}. \end{cases}$$

A point carries K-theory in infinitely many degrees. Everything else about K-theory behaves like cohomology: it pulls back, it is homotopy invariant, it has exact sequences and a product.

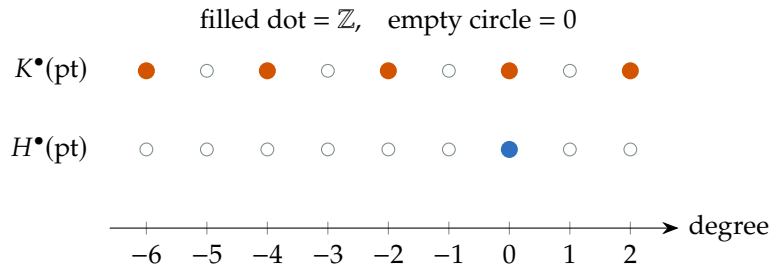


Figure 3.33: The dimension axiom in a picture. Ordinary cohomology of a point is a single $\mathbb{Z}$ in degree 0 (blue). K-theory of a point is a $\mathbb{Z}$ in every even degree (orange). The repetition with period 2 is Bott periodicity (Section 3.13); the extra groups are why K-theory is called a *generalised* cohomology theory.

Dictionary: K-theory words

Mathematics	Physics
vector bundle E over M	arena for a gauge field and its matter fields
transition function g_{ab}	gauge transformation relating two local gauges
isomorphism $E \cong E'$	the same physics in different gauges
isomorphism class	topological sector
trivial bundle I^k	no topological charge; global gauge exists
$E \oplus F$	two decoupled multiplets side by side
$E \otimes F$	a field carrying both charges (e.g. quark colour ⊗ flavour)
element of $K(M)$	a charge that survives adding neutral junk (e.g. D-brane charge)

Reading Nash (p. 65, opening of §3)

Nash introduces K-theory in one paragraph: a "cohomology theory for vector bundles", a "kind of linear algebra for large matrices", the most important generalised cohomology theory, failing the Eilenberg–Steenrod dimension axiom. He then fixes the notation $I^j \cong M \times \mathbb{R}^j$ or $M \times \mathbb{C}^j$ and announces that bundles will be sorted into "two sorts of equivalence classes". The two sorts are *stable equivalence classes* (Section 3.9) and *formal differences* (Section 3.10). Nash takes all the bundle background above for granted.

Recap

- A bundle is encoded by transition functions g_{ab} obeying $g_{ab}g_{bc} = g_{ac}$; isomorphism means $g' = hgh^{-1}$ patchwise.
- Key examples: Möbius ($g = \pm 1$ on S^1), Hopf H ($g = z$ on S^2), TS^2 ($g = -z^{-2} \simeq z^{-2}$), monopole of charge n ($g = e^{in\phi}$).
- $\oplus$ = block-diagonal, $\otimes$ = Kronecker product, dual = inverse transpose, pullback = composition.
- Bundles form a semiring like $\mathbb{N}$; K-theory will add subtraction.
- K-theory is a cohomology theory except that a point has $K^{-n}(\mathrm{pt}) = \mathbb{Z}$ for all even n.

3.9 Stable equivalence: ignoring trivial junk

Here is the problem that K-theory is built to solve. For numbers, $a + c = b + c$ implies $a = b$: we can cancel. For vector bundles we cannot. There are bundles E, F, G with $E \oplus G \cong F \oplus G$ but $E \not\cong F$. So "adding the same thing to both sides" loses information, and bundle arithmetic is not like number arithmetic.

K-theory's response is pragmatic: if adding trivial bundles can make two bundles isomorphic, we agree to *call them equivalent*. This section makes that precise, shows the classic example (the tangent bundle of a sphere), and proves Nash's stable equivalence theorem, which says that cancellation *does* work once we allow ourselves to add trivial bundles.

3.9.1 The definition

Working definition 3.24 (stable equivalence)

Two bundles E and F over M (possibly of *different* ranks) are **stably equivalent**, written $E \overset{s}{\sim} F$, if

$$E \oplus I^j \cong F \oplus I^k \qquad \text{for some } j, k \geq 0. \tag{3.80}$$

[I^j is the trivial bundle $M \times \mathbb{R}^j$ or $M \times \mathbb{C}^j$, and $\cong$ means isomorphic as bundles] *Test:* add trivial lines to both until the ranks agree and some large enough number of extra trivial lines have been added; then check for an isomorphism. A bundle stably equivalent to a trivial bundle is called **stably trivial**.

Note that if E has rank m and F rank n, then (3.80) forces $m + j = n + k$, so $j - k = n - m$ is fixed by the ranks. Only the "overall amount of padding" is free.

Nash says "it is easy to verify" that this is an equivalence relation. Let us verify it.

- *Reflexive:* $E \oplus I^0 \cong E \oplus I^0$ (take $j = k = 0$).
- *Symmetric:* the definition is symmetric in E and F (swap j and k).
- *Transitive:* suppose $E \oplus I^j \cong F \oplus I^k$ and $F \oplus I^l \cong G \oplus I^m$. Then

$$\begin{aligned} E \oplus I^{j+l} &\overset{(1)}{=} (E \oplus I^j) \oplus I^l \\ &\overset{(2)}{\cong} (F \oplus I^k) \oplus I^l \\ &\overset{(3)}{=} (F \oplus I^l) \oplus I^k \\ &\overset{(4)}{\cong} (G \oplus I^m) \oplus I^k = G \oplus I^{m+k}. \end{aligned}$$

  (1) $I^{j+l} = I^j \oplus I^l$ (block-diagonal identity transition functions), and $\oplus$ is associative.
  (2) First hypothesis, then add I^l to both sides (adding the same bundle preserves isomorphism: use the block-diagonal isomorphism $\mathrm{diag}(h, 1)$).
  (3) Commutativity and associativity of $\oplus$.
  (4) Second hypothesis, then add I^k to both sides.

  So $E \overset{s}{\sim} G$.

The equivalence classes are **stable equivalence classes**.

Worked example 3.41: All trivial bundles are one stable class

I^3 and I^7 have different ranks, so they are not isomorphic. But $I^3 \oplus I^4 \cong I^7 \oplus I^0$, so $I^3 \overset{s}{\sim} I^7$. In fact every I^j is stably equivalent to I^0, the zero bundle. Stable equivalence forgets rank completely; rank will be restored separately in Section 3.10.

3.9.2 Cancellation fails: the tangent bundle of a sphere

Nash states

$$E \oplus G \cong F \oplus G \not\Longrightarrow E \cong F \tag{3.81}$$

and gives the standard counterexample. Let $S^n \subset \mathbb{R}^{n+1}$ be the unit sphere. At a point $x \in S^n$ (a unit vector) every vector $u \in \mathbb{R}^{n+1}$ splits uniquely into a part tangent to the sphere and a part along the normal:

$$u = \underbrace{\big(u - (u \cdot x)\, x\big)}_{\in T_x S^n} + \underbrace{(u \cdot x)\, x}_{\in N_x}.$$

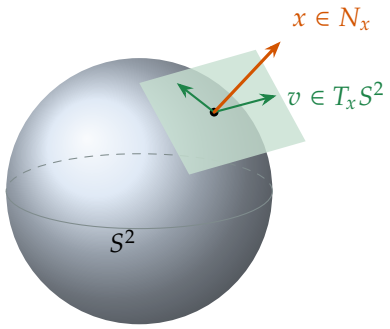
The tangent part is perpendicular to x: $(u - (u \cdot x)x) \cdot x = u \cdot x - (u \cdot x)(x \cdot x) = 0$ because $x \cdot x = 1$. Doing this at every point gives the bundle statement

$$T\mathbb{R}^{n+1}\big|_{S^n} \cong TS^n \oplus N(S^n), \tag{3.82}$$

where $N(S^n)$ is the **normal bundle**, the line bundle whose fibre at x is the line $\mathbb{R}x$.

Now we assemble the counterexample in labelled steps.

$$\begin{aligned} TS^n \oplus I &\overset{(1)}{\cong} TS^n \oplus N(S^n) \\ &\overset{(2)}{\cong} T\mathbb{R}^{n+1}\big|_{S^n} \\ &\overset{(3)}{\cong} I^{n+1} \\ &\overset{(4)}{=} I^n \oplus I. \end{aligned}$$



$T_xS^2 \oplus N_x = \mathbb{R}^3$
$TS^2 \oplus N \cong I^3$
$N \cong I$ (section $x \mapsto x$)
so $TS^2 \oplus I \cong I^2 \oplus I$
but $TS^2 \not\cong I^2$

Figure 3.34: At each point x of the sphere, $\mathbb{R}^3$ splits into the tangent plane (green) and the normal line (orange). The normal line bundle has the nowhere-zero section $x \mapsto x$, so it is trivial. Adding it to TS^2 gives the trivial bundle I^3, yet TS^2 itself is not trivial: we cannot cancel the I.

(1) $N(S^n)$ has the global nowhere-vanishing section $x \mapsto x$ (the outward normal); by the key idea of Section 3.8.1, a line bundle with a nonvanishing section is trivial.
(2) The pointwise splitting (3.82).
(3) $T\mathbb{R}^{n+1} = \mathbb{R}^{n+1} \times \mathbb{R}^{n+1}$ is trivial (use the constant frame $\partial_1, \dots, \partial_{n+1}$), and so is its restriction.
(4) Splitting off one trivial direction.

So $TS^n \oplus I \cong I^n \oplus I$ (Nash's (3.36)). With $E = TS^n$, $F = I^n$, $G = I$ this is $E \oplus G \cong F \oplus G$. If cancellation held, TS^n would be trivial. It is not, for most n:

Key idea: Which tangent bundles of spheres are trivial

TS^n is trivial ("S^n is **parallelisable**") only for $n = 1, 3, 7$ (and trivially $n = 0$). This is a deep theorem of Adams (1960), closely tied to the existence of the real division algebras $\mathbb{C}$, $\mathbb{H}$ and the octonions $\mathbb{O}$.

We will not prove Adams' theorem, but we can do the two things a physicist wants: show the trivial cases explicitly, and prove non-triviality in the most important case $n = 2$.

Worked example 3.42: Explicit frames on S^1 and S^3

S^1: at $x = (x_1, x_2)$ the vector $e(x) = (-x_2, x_1)$ is tangent ($e \cdot x = -x_2x_1 + x_1x_2 = 0$) and never zero. One nonvanishing section trivialises the line bundle TS^1.
S^3: write $x \in S^3 \subset \mathbb{R}^4$ as a unit quaternion $q = a + b\,i + c\,j + d\,k$, with $i^2 = j^2 = k^2 = ijk = -1$. Right-multiply by i, j, k:

$$q\,i \overset{(1)}{=} ai + bi^2 + c\,ji + d\,ki \overset{(2)}{=} -b + ai + dj - ck,$$
$$q\,j \overset{(1)}{=} aj + b\,ij + cj^2 + d\,kj \overset{(2)}{=} -c - di + aj + bk,$$
$$q\,k \overset{(1)}{=} ak + b\,ik + c\,jk + dk^2 \overset{(2)}{=} -d + ci - bj + ak.$$

(1) Distribute the product over the four components of q.
(2) Quaternion multiplication table: $ij = k = -ji$, $jk = i = -kj$, $ki = j = -ik$.

As vectors in $\mathbb{R}^4$ (components along $1, i, j, k$):

$$e_1 = (-b, a, d, -c), \qquad e_2 = (-c, -d, a, b), \qquad e_3 = (-d, c, -b, a).$$

Each is perpendicular to $x = (a, b, c, d)$; for instance $e_1 \cdot x = -ab + ab + dc - cd = 0$. They are mutually perpendicular; for instance $e_1 \cdot e_2 = bc - ad + ad - bc = 0$. Each has length

1. So (e_1, e_2, e_3) is a global orthonormal frame of TS^3, and $TS^3 \cong I^3$. (The same trick with octonions frames S^7.) Physically: $S^3 = \mathrm{SU}(2)$, and e_1, e_2, e_3 are the left-invariant vector fields generated by the three Pauli directions.

Worked example 3.43: The hairy-ball theorem from transition functions

We show TS^2 is not trivial. First we need a fact about complex line bundles over $S^2 = D_0 \cup D_1$ glued along the equator $|z| = 1$ with transition function $g(e^{i\phi}) \in \mathrm{U}(1)$: *the winding number $w(g)$ of g around the equator is an isomorphism invariant.* Suppose $g' = h_1\, g\, h_0^{-1}$ with h_0 defined (and nonzero) on all of D_0 and h_1 on all of D_1. Then

$$\begin{aligned} w(g') &\overset{(1)}{=} w(h_1) + w(g) - w(h_0) \\ &\overset{(2)}{=} 0 + w(g) - 0 = w(g). \end{aligned}$$

(1) The winding number of a product of loops in $\mathbb{C}^\times$ is the sum of their winding numbers (phases add), and an inverse has minus the winding.

(2) h_0 restricted to the equator is the boundary of the map h_0 on the disc D_0. A loop in $\mathbb{C}^\times$ that extends over a disc can be shrunk to a point through the disc, so its winding number is zero. The same holds for h_1 on D_1.

In Example 3.37 we found that TS^2 has transition function $-z^{-2}$, winding number $-2 \neq 0$, while I has transition function 1, winding number 0. So $TS^2 \not\cong I$ as complex line bundles. As a real rank-2 bundle it is also non-trivial: a real trivialisation would give a nowhere-zero tangent vector field s, and then s and is (the 90° rotation of s) would be a complex frame, trivialising it as a complex line bundle too.

Conclusion: *every tangent vector field on S^2 has a zero* (you cannot comb a hairy ball flat). The number 2 we found is the Euler characteristic $\chi(S^2)$, which by Poincaré–Hopf is also the sum of the indices of the zeros of any vector field.

So TS^2 is *stably trivial but not trivial*: the simplest example of information that stable equivalence throws away.

Watch out: Stably trivial is not trivial, and not every bundle is stably trivial

Stable equivalence throws some information away (the winding number of TS^2 disappears), but not all. The Möbius bundle $\mathcal{M}$ and the Hopf bundle H are *not* stably trivial. The tool is the **determinant line bundle** $\det E = \Lambda^{\mathrm{top}} E$, whose transition functions are $\det g_{ab}$. Since

$$\begin{aligned} \det(g_{ab} \oplus 1_j) &\overset{(1)}{=} \det g_{ab} \cdot \det 1_j \\ &= \det g_{ab}, \end{aligned}$$

(1) The determinant of a block-diagonal matrix is the product of the determinants of the blocks.

we have $\det(E \oplus I^j) \cong \det E$: the determinant line bundle is a *stable invariant*. Now $\det \mathcal{M} = \mathcal{M}$ is non-trivial, while $\det I^k = I$ is trivial; so $\mathcal{M} \oplus I^j$ is never trivial. Likewise $\det H = H$ has winding number 1, so H is not stably trivial. (And for TS^2 viewed as a real bundle, det is the orientation bundle, which is trivial because S^2 is orientable: consistent.)

3.9.3 Complements and the stable equivalence theorem

The failure of cancellation is annoying. The saving fact is that every bundle has a "complement" that makes it trivial.

Key idea: Every bundle is a summand of a trivial bundle

If M is compact, then for every bundle G there is a bundle G' and an integer n with

$$G \oplus G' \cong I^n. \tag{3.83}$$

Why this is true. The idea is to embed G inside a big trivial bundle and take the orthogonal complement.

1. Because M is compact, finitely many patches $U_1, \dots, U_N$ trivialise G, with local fibre-coordinate maps $\varphi_a : G|_{U_a} \to U_a \times \mathbb{F}^k$.
2. Take a partition of unity ρ_a subordinate to the cover (smooth functions ≥ 0, ρ_a vanishing outside U_a, $\sum_a \rho_a = 1$).
3. Map a vector $v \in G_x$ to the big vector $\Phi(v) = \big(\rho_1(x)\varphi_1(v), \dots, \rho_N(x)\varphi_N(v)\big) \in \mathbb{F}^{kN}$. This is linear on each fibre, and injective: if $\Phi(v) = 0$, pick a with $\rho_a(x) > 0$; then $\varphi_a(v) = 0$, so $v = 0$.
4. So G sits inside $I^{kN} = M \times \mathbb{F}^{kN}$ as a subbundle. Let G' be its orthogonal complement with respect to the standard inner product on $\mathbb{F}^{kN}$. Pointwise $\mathbb{F}^{kN} = G_x \oplus G'_x$, so $G \oplus G' \cong I^{kN}$.

Compactness was used in step 1: with infinitely many patches, Φ would need infinitely many components. Nash's footnote says "we assume M to be Hausdorff", but compactness (or finite-dimensionality of the base with finitely many patches) is what matters. For a physicist: work on a compact spacetime (a sphere or a torus, after imposing boundary conditions at infinity).

Worked example 3.44: The complement of the Hopf bundle is its dual

The Hopf bundle H is *defined* as a subbundle of $I^2 = \mathbb{CP}^1 \times \mathbb{C}^2$: its fibre over $[z_0 : z_1]$ is the line through (z_0, z_1). Its orthogonal complement $H^\perp$ (using the Hermitian inner product $\langle u, v\rangle = \bar{u}_0 v_0 + \bar{u}_1 v_1$) has fibre spanned by $(-\bar{z}_1, \bar{z}_0)$, because $\langle (z_0, z_1), (-\bar{z}_1, \bar{z}_0)\rangle = -\bar{z}_0\bar{z}_1 + \bar{z}_1\bar{z}_0 = 0$. Choose local sections as for H:

$$t_0 = (-\bar{z},\, 1) \text{ on } U_0, \qquad t_1 = (-1,\, \bar{w}) \text{ on } U_1.$$

(Put $(z_0, z_1) = (1, z)$ and $(w, 1)$ respectively into $(-\bar{z}_1, \bar{z}_0)$.) On the overlap,

$$\begin{aligned} t_0 &\overset{(1)}{=} (-\bar{z},\, 1) \\ &\overset{(2)}{=} \bar{z}\,\big(-1,\, 1/\bar{z}\big) \\ &\overset{(3)}{=} \bar{z}\, t_1, \end{aligned}$$

(1) Definition of t_0.
(2) Pull out the nonzero factor $\bar{z}$.
(3) $1/\bar{z} = \overline{1/z} = \bar{w}$.

So fibre coordinates obey $\lambda_1 = \bar{z}\,\lambda_0$. On the equator $\bar{z} = e^{-i\phi} = z^{-1}$: winding number -1, the same as the dual H^*. Hence

$$H \oplus H^* \cong H \oplus H^\perp \cong I^2.$$

A complement of H is H^*, with $n = 2$. This single identity will give us $[H^*] = 2 - [H]$ in K-theory.

Worked example 3.45: Möbius ⊕ Möbius is trivial

Realise $\mathcal{M}$ inside $S^1 \times \mathbb{R}^2$: over the angle θ take the line spanned by $u(\theta) = (\cos\frac{\theta}{2}, \sin\frac{\theta}{2})$. Going once around, $\theta \to \theta + 2\pi$, the spanning vector goes to $-u$: the same line, but the direction has flipped. That is exactly the Möbius twist. The orthogonal complement is spanned by $u^\perp(\theta) = (-\sin\frac{\theta}{2}, \cos\frac{\theta}{2})$, which also flips sign after one trip, so it is again a Möbius bundle. Hence

$$\mathcal{M} \oplus \mathcal{M} \cong I^2.$$

Check with transition functions: $\mathcal{M} \oplus \mathcal{M}$ is glued by $\mathrm{diag}(1,1)$ near $\theta = 0$ and $\mathrm{diag}(-1,-1)$ near $\theta = \pi$. But $\mathrm{diag}(-1,-1)$ is the rotation by π, which can be continuously rotated back to $\mathrm{diag}(1,1)$ inside $\mathrm{SO}(2)$. The gluing is homotopic to the identity, so the bundle is trivial.

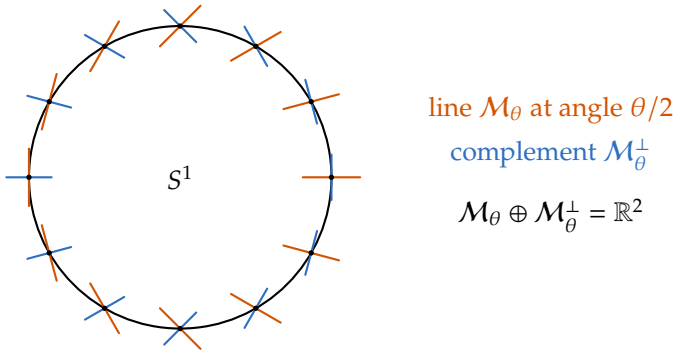


Figure 3.35: The Möbius bundle realised inside the trivial bundle $S^1 \times \mathbb{R}^2$. At angle θ the orange line points at angle $\theta/2$, so after one trip it has turned by only π: the same line with its direction reversed. The blue perpendicular lines form a second Möbius bundle, and together they fill $\mathbb{R}^2$ at every point: $\mathcal{M} \oplus \mathcal{M} \cong I^2$.

Now the theorem.

Key idea: Nash's stable equivalence theorem

If $E \oplus G \cong F \oplus G$, then $E \overset{s}{\sim} F$. Cancellation fails for isomorphism, but it holds for stable equivalence.

The proof is a two-line calculation once we have complements. Choose G' with $G \oplus G' \cong I^n$. Then

$$\begin{aligned} E \oplus I^n &\overset{(1)}{\cong} E \oplus (G \oplus G') \\ &\overset{(2)}{=} (E \oplus G) \oplus G' \\ &\overset{(3)}{\cong} (F \oplus G) \oplus G' \\ &\overset{(2)}{=} F \oplus (G \oplus G') \\ &\overset{(1)}{\cong} F \oplus I^n. \end{aligned}$$

(1) The complement relation (3.83).
(2) Associativity of $\oplus$.

(3) The hypothesis $E \oplus G \cong F \oplus G$, with G' added to both sides.

So $E \oplus I^n \cong F \oplus I^n$, which is stable equivalence with $j = k = n$. The whole trick is: *adding G' converts the arbitrary common summand G into a trivial one*, and trivial summands are exactly what stable equivalence allows us to ignore.

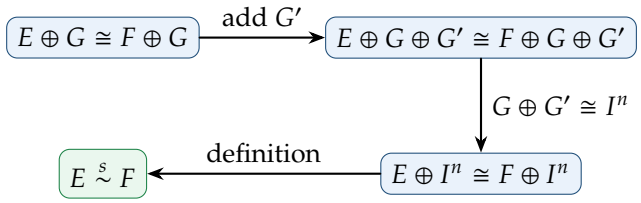


Figure 3.36: The logic of the stable equivalence theorem. The only input beyond bookkeeping is the existence of a complement G', which needs a compact base.

Worked example 3.46: Applying the theorem to TS^2

We found $TS^2 \oplus I \cong I^2 \oplus I$. The theorem (with $G = I$, $G' = I^0$, $n = 1$) gives $TS^2 \overset{s}{\sim} I^2$, which is the definition of stably trivial. Similarly, $H \oplus H^* \cong I^2 = I \oplus I$ does *not* say $H \overset{s}{\sim} I$ (the common summand would have to be the same on both sides, and $H^* \neq I$). Indeed H is not stably trivial.

Try it: Stable classes of line bundles on S^1

Show that every real line bundle on S^1 is isomorphic to I or $\mathcal{M}$, and that these are in different stable classes. *Answer:* glue two arcs by nonzero real numbers on the two overlap intervals; rescaling on one arc lets us make both gluings ± 1, and only the product of the two signs is invariant. So there are two bundles, I (product $+1$) and $\mathcal{M}$ (product -1). They are stably distinct because their determinant bundles, I and $\mathcal{M}$, differ.

Reading Nash (pp. 66–67, (3.32)–(3.39))

Nash defines stable equivalence (3.32)–(3.33), states that cancellation fails (3.34) and gives the example $T\mathbb{R}^{n+1} \cong TS^n \oplus N(S^n)$, (3.35)–(3.36), citing parallelisability only for $n = 1, 3, 7$. His theorem (3.37) is proved in (3.38)–(3.39) using a complement G' with $G \oplus G' \cong I^n$, whose existence he relegates to a footnote ("we assume M to be Hausdorff"). We have supplied: the verification that stable equivalence is an equivalence relation; the construction of G' (which needs *compactness*); explicit frames for S^1, S^3; a self-contained proof that TS^2 is non-trivial; explicit complements for H and $\mathcal{M}$; and the determinant bundle as a stable invariant.

Recap

- $E \overset{s}{\sim} F$ means $E \oplus I^j \cong F \oplus I^k$; ranks may differ.
- Cancellation fails: $TS^2 \oplus I \cong I^3$ but $TS^2 \not\cong I^2$ (winding number 2; hairy-ball theorem).
- Over a compact base every G has a complement: $G \oplus G' \cong I^n$ (embed and take $\perp$). Examples: $H \oplus H^* \cong I^2$, $\mathcal{M} \oplus \mathcal{M} \cong I^2$.
- Stable equivalence theorem: $E \oplus G \cong F \oplus G \Rightarrow E \overset{s}{\sim} F$.
- Stable invariants survive: e.g. $\det E$, so $\mathcal{M}$ and H are not stably trivial.

3.10 The group $K(M)$: subtracting bundles

We can add bundles with $\oplus$, but we cannot subtract them. Given E and F, there is usually no bundle X with $F \oplus X \cong E$ (think of $E = I$ and $F = H$ over S^2: no bundle of rank 0 does it, obviously). In algebraic language, $\mathrm{Vect}(M)$ with $\oplus$ is an abelian **semigroup** (really a monoid: it has the zero bundle I^0 as identity), not a group. And by (3.81) it does not even have cancellation.

This is exactly the situation of the natural numbers $\mathbb{N} = \{0, 1, 2, \ldots\}$: you can add, you cannot subtract, and the fix is to invent negative numbers as *formal differences*. Nash's second kind of equivalence class is precisely this: formal differences of bundles. The resulting group is $K(M)$.

3.10.1 Completing a semigroup: the Grothendieck construction

Let S be any abelian semigroup with operation $+$ and identity 0. We want a group $G(S)$ in which every pair (s, t) stands for "$s - t$".

Working definition 3.25 (Grothendieck group)

Elements of $G(S)$ are pairs $(s, t) \in S \times S$, thought of as $s - t$, modulo the relation

$$(s_1, t_1) \sim (s_2, t_2) \iff s_1 + t_2 + u = s_2 + t_1 + u \quad \text{for some } u \in S. \tag{3.84}$$

[the extra u matters only when S lacks cancellation, as $\mathrm{Vect}(M)$ does] Add componentwise: $(s, t) + (s', t') = (s + s',\ t + t')$. The zero is $(0, 0)$ (equivalently any (s, s)) and the negative of (s, t) is (t, s). *Test for equality:* two formal differences $s_1 - t_1$ and $s_2 - t_2$ are equal when "cross-adding" gives equal answers *after adding some common* u.

Without the u, the relation $s_1 + t_2 = s_2 + t_1$ is what you would write for numbers ($5 - 3 = 7 - 5$ because $5 + 5 = 7 + 3$). The extra u is needed because S may lack cancellation; it is what makes $\sim$ transitive. Let us check transitivity, since it shows exactly where u earns its keep. Suppose $(s_1, t_1) \sim (s_2, t_2)$ via u and $(s_2, t_2) \sim (s_3, t_3)$ via v. Then

$$\begin{aligned}(s_1 + t_3) + (t_2 + u + v + s_2) &\overset{(1)}{=} (s_1 + t_2 + u) + (s_2 + t_3 + v)\\ &\overset{(2)}{=} (s_2 + t_1 + u) + (s_3 + t_2 + v)\\ &\overset{(1)}{=} (s_3 + t_1) + (t_2 + u + v + s_2).\end{aligned}$$

(1) Rearranging a sum in the abelian semigroup S.
(2) The two hypotheses, $s_1 + t_2 + u = s_2 + t_1 + u$ and $s_2 + t_3 + v = s_3 + t_2 + v$.

So $(s_1, t_1) \sim (s_3, t_3)$ via $w = t_2 + u + v + s_2$. We could not have cancelled $t_2 + s_2$ from both sides; we did not need to, because w is allowed.

That (t, s) really is the negative of (s, t):

$$\begin{aligned}(s, t) + (t, s) &\overset{(1)}{=} (s + t,\ t + s)\\ &\overset{(2)}{\sim} (0, 0),\end{aligned}$$

(1) Componentwise addition.
(2) Test (3.84) with $u = 0$: $(s + t) + 0 = 0 + (t + s)$ by commutativity.

so every element has an inverse and $G(S)$ is an abelian group.

Worked example 3.47: From $\mathbb{N}$ to $\mathbb{Z}$

Take $S = \mathbb{N}$. The pairs $(5,3)$, $(7,5)$, $(2,0)$ are all equivalent: $5+5 = 7+3$, $5+0 = 2+3$. Each equivalence class is a diagonal line of lattice points in $\mathbb{N}\times\mathbb{N}$ (Figure 3.37), and the classes are labelled by $s-t \in \mathbb{Z}$. So $G(\mathbb{N}) = \mathbb{Z}$. Here $\mathbb{N}$ has cancellation, so u is never needed.

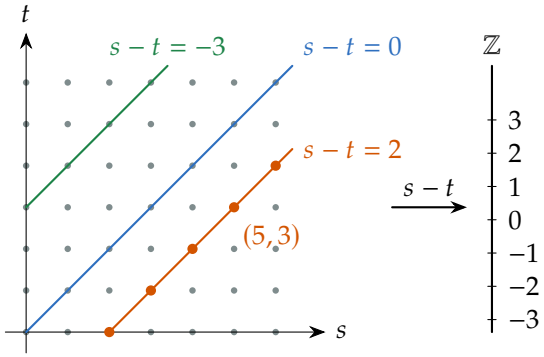


Figure 3.37: The Grothendieck construction for $\mathbb{N}$. A pair (s,t) is a lattice point; pairs on the same diagonal are declared equal. The diagonals are labelled by $s-t$, which runs over all of $\mathbb{Z}$. For bundles the lattice is $\mathrm{Vect}(M)\times\mathrm{Vect}(M)$ and the "diagonals" are the classes $[E]-[F]$.

Worked example 3.48: What the u does when cancellation fails

Take $S = \mathbb{N}\cup\{\infty\}$ with $n+\infty = \infty$. For *any* two pairs choose $u = \infty$: then both sides of (3.84) equal ∞. So every pair is equivalent to $(0,0)$ and $G(S) = 0$. The lesson: in $G(S)$, s and s' become equal as soon as $s+u = s'+u$ for *some* u. The Grothendieck group identifies things that cannot be told apart after adding something. For bundles, this is precisely stable equivalence, as we now see.

3.10.2 Definition of $K(M)$

Apply the construction to $S = \mathrm{Vect}(M)$ with $+ = \oplus$. The relation (3.84) becomes

$$(E,F)\sim(G,H) \iff E\oplus H\oplus J \cong G\oplus F\oplus J \text{ for some bundle } J. \tag{3.85}$$

Working definition 3.26 (the K-group $K(M)$)

$$K(M) = \big(\mathrm{Vect}(M)\times\mathrm{Vect}(M)\big)\,/\sim .$$

[$\mathrm{Vect}(M)$ is the set of isomorphism classes of bundles over M] The class of (E,F) is written $[E]-[F]$, and the class of (E,I^0) is written $[E]$. Elements are called **virtual bundles** [formal differences $[E]-[F]$, in general not the class of an actual bundle]. Addition is $([E]-[F])+([G]-[H]) = [E\oplus G]-[F\oplus H]$. We write n for $[I^n]$. *Test for equality:* $[E]-[F] = [G]-[H]$ iff $E\oplus H\oplus I^N \cong G\oplus F\oplus I^N$ for some N.

The last line of the box (with a *trivial* J) is not the definition but follows from it using complements. If $E\oplus H\oplus J \cong G\oplus F\oplus J$, choose J' with $J\oplus J' \cong I^N$ and add J' to both sides:

$$\begin{aligned} E\oplus H\oplus I^N &\overset{(1)}{\cong} E\oplus H\oplus J\oplus J' \\ &\overset{(2)}{\cong} G\oplus F\oplus J\oplus J' \\ &\overset{(1)}{\cong} G\oplus F\oplus I^N. \end{aligned}$$

(1) Complement (3.83) of J.
(2) Hypothesis, with J' added to both sides.

In particular $[E] = [F]$ in $K(M)$ iff $E \oplus I^N \cong F \oplus I^N$ for some N: stably equivalent *with the same rank*.

Worked example 3.49: The K-group of a point

A vector bundle over a single point is just a vector space, determined up to isomorphism by its dimension. So $\mathrm{Vect}(\mathrm{pt}) \cong \mathbb{N}$ (with $\oplus$ corresponding to $+$), and by Example 3.47

$$K(\mathrm{pt}) = G(\mathbb{N}) = \mathbb{Z}.$$

A virtual vector space $[\mathbb{C}^a] - [\mathbb{C}^b]$ is remembered only through the integer $a - b$: its *super-dimension*, if we think of $\mathbb{C}^a$ as bosonic and $\mathbb{C}^b$ as fermionic states.

3.10.3 Every virtual bundle is "bundle minus trivial"

A formal difference $[E] - [F]$ has a bundle on both sides. Nash shows that we can always make the subtracted bundle trivial. Choose F' with $F \oplus F' \cong I^n$. Then

$$\begin{aligned} [E] - [F] &\overset{(1)}{=} \big([E] + [F']\big) - \big([F] + [F']\big) \\ &\overset{(2)}{=} [E \oplus F'] - [F \oplus F'] \\ &\overset{(3)}{=} [E \oplus F'] - [I^n] \\ &\overset{(4)}{=} [H] - [I^n], \qquad H \coloneqq E \oplus F'. \end{aligned} \tag{3.86}$$

(1) Add and subtract $[F']$ (legitimate in a group).
(2) Addition in $K(M)$ is induced by $\oplus$.
(3) $F \oplus F' \cong I^n$, and isomorphic bundles have the same class.
(4) Name the bundle $E \oplus F'$.

Worked example 3.50: Negatives in $K(S^2)$

From $H \oplus H^* \cong I^2$ (Example 3.44),

$$[H] + [H^*] \overset{(1)}{=} [H \oplus H^*] \overset{(2)}{=} [I^2] = 2,$$

(1) Addition in K is $\oplus$.
(2) The complement relation.

so $[H^*] = 2 - [H]$ and $-[H] = [H^*] - 2$. The "negative of the Hopf bundle" is represented by H^* minus two trivial lines. Similarly,

$$\begin{aligned} 1 - [H] &\overset{(1)}{=} \big(1 + [H^*]\big) - \big([H] + [H^*]\big) \\ &\overset{(2)}{=} [I \oplus H^*] - [I^2], \end{aligned}$$

(1) Add and subtract $[H^*]$.
(2) $1 + [H^*] = [I \oplus H^*]$, and $[H] + [H^*] = 2$ from the chain above.

which is of the form (3.86) with $H \to I \oplus H^*$ and $n = 2$.

3.10.4 Virtual dimension

Working definition 3.27 (virtual dimension)

The **virtual dimension** (or virtual rank) of $[E]-[F]$ is $\operatorname{rk}E-\operatorname{rk}F\in\mathbb{Z}$ [rk is the dimension of the fibres]. It can be positive, negative or zero. By (3.86) [every class can be written $[H]-n$ for one bundle H] it equals $\operatorname{rk}H-n$.

It is worth checking that this is well defined: two representatives of the same class must give the same integer. If $(E,F)\sim(G,H)$ then $E\oplus H\oplus J\cong G\oplus F\oplus J$, and taking ranks,

$$\operatorname{rk}E+\operatorname{rk}H+\operatorname{rk}J\overset{(1)}{=}\operatorname{rk}G+\operatorname{rk}F+\operatorname{rk}J$$
$$\Longrightarrow\ \operatorname{rk}E-\operatorname{rk}F\overset{(2)}{=}\operatorname{rk}G-\operatorname{rk}H.$$

(1) Isomorphic bundles have equal rank, and rank is additive under $\oplus$.
(2) Subtract $\operatorname{rk}J+\operatorname{rk}F+\operatorname{rk}H$ from both sides (these are integers, so cancellation is fine).

Physical picture: Virtual bundles in physics

A virtual bundle $E\ominus F$ is a $\mathbb{Z}_2$-graded object: think of E as bosonic and F as fermionic, with a rule that equal pieces on the two sides cancel. The virtual dimension is then "number of bosons minus number of fermions", like a Witten index. The most important virtual bundle in this book is the *index bundle* of a family of Dirac operators D_y parametrised by $y\in Y$: roughly $[\operatorname{Ker}D_y]-[\operatorname{Coker}D_y]$. The kernel and cokernel can jump in dimension as y varies (zero modes appear in pairs), but the difference is stable, and its virtual dimension is the index (Chapter 4, Chapter 10).

3.10.5 Stable classes are the virtual bundles of dimension zero

Nash now links his two kinds of equivalence classes. Assign to a bundle E_m of rank m the virtual bundle

$$E_m\longmapsto[E_m]-[I^m],$$

which has virtual dimension $m-m=0$. *Claim:* this map gives a one-to-one correspondence between stable equivalence classes of bundles and virtual bundles of dimension zero. There are three things to check.

(a) Same image $\Rightarrow$ stably equivalent. Suppose E_m (rank m) and F_n (rank n) have the same image. Then

$$[E_m]-[I^m]=[F_n]-[I^n]$$
$$\Longrightarrow\ E_m\oplus I^n\oplus K\overset{(1)}{\cong}F_n\oplus I^m\oplus K\quad\text{for some } K$$
$$\Longrightarrow\ E_m\oplus I^n\overset{(2)}{\overset{s}{\sim}}F_n\oplus I^m$$
$$\Longrightarrow\ E_m\overset{(3)}{\overset{s}{\sim}}F_n.$$

(1) The equality test (3.85) with $(E,F)=(E_m,I^m)$, $(G,H)=(F_n,I^n)$.
(2) Nash's stable equivalence theorem, cancelling the common summand K.
(3) $E\overset{s}{\sim}E\oplus I^j$ for any j (definition), and $\overset{s}{\sim}$ is transitive.

(b) Stably equivalent ⇒ same image (so the map is well defined on stable classes). Suppose $E_m \oplus I^j \cong F_n \oplus I^k$; ranks give $m + j = n + k$. Then

$$\begin{aligned}[E_m] - [I^m] &\overset{(1)}{=} [E_m \oplus I^j] - [I^{m+j}] \\ &\overset{(2)}{=} [F_n \oplus I^k] - [I^{n+k}] \\ &\overset{(1)}{=} [F_n] - [I^n].\end{aligned}$$

(1) Add and subtract $[I^j]$ (respectively $[I^k]$).
(2) The hypothesis $E_m \oplus I^j \cong F_n \oplus I^k$, and $m + j = n + k$.

(c) Every virtual bundle of dimension zero is hit. By (3.86) it is $[H] - [I^n]$ with $\operatorname{rk} H = n$, which is the image of H. (Nash leaves this step implicit.)

Key idea: Reduced K-theory

The **reduced K-group** $\widetilde{K}(M)$ is the set of stable equivalence classes of bundles over M. By (a)–(c) it is the same thing as the subgroup of $K(M)$ of virtual bundles of virtual dimension zero:

$$\widetilde{K}(M) = \{\, [E] - [I^{\operatorname{rk} E}] \,\} = \ker\big(\mathrm{Rk} : K(M) \to \mathbb{Z}\big),$$

where $\mathrm{Rk}([E] - [F]) = \operatorname{rk} E - \operatorname{rk} F$. In this form it is automatically a group.

Watch out: The group law on stable classes

Defined as "stable classes", it is not obvious that $\widetilde{K}(M)$ is a group: $\oplus$ adds, but what is the inverse of the stable class of E? It is the stable class of a complement E', since $E \oplus E' \cong I^n \overset{s}{\sim} I^0$. The identification with $\ker \mathrm{Rk}$ makes this automatic.

3.10.6 $K(M) = \widetilde{K}(M) \oplus \mathbb{Z}$: pulling back to a point

Rk is a group homomorphism onto $\mathbb{Z}$ (onto because $\mathrm{Rk}[I^n] = n$), and $\widetilde{K}(M)$ is its kernel. Nash repackages this using pullbacks, which is the form that generalises.

Pullback on K-theory. A map $\alpha : N \to M$ pulls bundles back, $E \mapsto \alpha^* E$ (Section 3.8.2), and respects $\oplus$ and isomorphism. So it induces a group homomorphism

$$\alpha^* : K(M) \to K(N), \qquad \alpha^*\big([E] - [F]\big) = [\alpha^* E] - [\alpha^* F]. \tag{3.87}$$

It is well defined: if $E \oplus H \oplus J \cong G \oplus F \oplus J$ then pulling back gives $\alpha^* E \oplus \alpha^* H \oplus \alpha^* J \cong \alpha^* G \oplus \alpha^* F \oplus \alpha^* J$, which is the same relation on N. It is functorial, $(\alpha \circ \beta)^* = \beta^* \circ \alpha^*$, because composing transition functions with α and then with β is composing with $\alpha \circ \beta$. And it is homotopy invariant: *homotopic maps pull back isomorphic bundles* over a compact base. The reason: along a homotopy α_t the transition functions $g_{ab} \circ \alpha_t$ vary continuously in t, and two sets of transition functions that are close enough are related by h_a's close to 1, hence define isomorphic bundles; by compactness of $[0, 1]$ finitely many small steps connect $t = 0$ to $t = 1$.

The point and its two maps. Let $m_0 \in M$ be a base point. There are two natural maps

$$i : \{m_0\} \hookrightarrow M \ \text{(inclusion)}, \qquad c : M \to \{m_0\} \ \text{(collapse everything to } m_0),$$

with $c \circ i = \mathrm{id}_{\{m_0\}}$. On K-theory:

- $i^* : K(M) \to K(\{m_0\}) = \mathbb{Z}$ restricts a virtual bundle to the single fibre over m_0: $i^*([E]-[F]) = \dim E_{m_0} - \dim F_{m_0}$. For connected M this is Rk. (Nash's footnote 4: on a disconnected M ranks can differ from component to component, and i^* only sees the component of m_0.)
- $c^* : \mathbb{Z} \to K(M)$ pulls back $\mathbb{C}^n$ (or $\mathbb{R}^n$) over the point to the constant family: $c^*(n) = [I^n]$.
- $i^* \circ c^* = (c \circ i)^* = \mathrm{id}$: restricting a trivial bundle of rank n to a point gives $\mathbb{C}^n$.

This is summarised by the **split exact sequence**

$$0 \longrightarrow \widetilde{K}(M) \xrightarrow{I} K(M) \underset{c^*}{\overset{i^*}{\rightleftarrows}} K(\{m_0\}) \longrightarrow 0, \tag{3.88}$$

where I is the inclusion. "Exact" means: at each spot, the image of the incoming arrow equals the kernel of the outgoing one. Here: I is injective, $\operatorname{Im} I = \widetilde{K}(M) = \ker i^*$, and i^* is onto. "Split" means that c^* is a right inverse of i^*.

$$0 \longrightarrow \widetilde{K}(M) \xrightarrow{I} K(M) \underset{c^*}{\overset{i^*}{\rightleftarrows}} K(\mathrm{pt}) = \mathbb{Z} \longrightarrow 0$$

$$[E] - [I^{\mathrm{rk}\,E}] \longmapsto [E] \longmapsto \mathrm{rk}\,E$$

Figure 3.38: The split exact sequence for $K(M)$. Going right, i^* remembers only the rank (the fibre over one point). Going back, c^* turns an integer n into the trivial bundle I^n. What i^* forgets is exactly $\widetilde{K}(M)$, the stable class. The bottom row follows one bundle through the maps.

A split exact sequence of abelian groups always gives a direct sum. Every $x \in K(M)$ decomposes as

$$x \overset{(1)}{=} \big(x - c^*i^*x\big) + c^*i^*x,$$

(1) Add and subtract c^*i^*x.

and the first piece lies in $\ker i^*$ because $i^*(x - c^*i^*x) = i^*x - (i^*c^*)(i^*x) = i^*x - i^*x = 0$. The decomposition is unique: if $y \in \ker i^*$ and $y = c^*n$, then $n = i^*c^*n = i^*y = 0$. Hence

$$K(M) = \ker i^* \oplus c^*K(\{m_0\}) = \widetilde{K}(M) \oplus \mathbb{Z}. \tag{3.89}$$

Concretely, $[E]-[F]$ splits as $\big([E]-[F]-(\mathrm{rk}\,E-\mathrm{rk}\,F)\big)+(\mathrm{rk}\,E-\mathrm{rk}\,F)$: "stable class" plus "virtual rank".

Worked example 3.51: K of the circle, complex case

Every complex bundle over S^1 is trivial. Glue two arcs with transition matrices $g \in \mathrm{GL}(k,\mathbb{C})$ on the two overlap intervals. Rechoosing the fibre coordinate on one arc lets us make one gluing the identity, leaving one matrix g. Since $\mathrm{GL}(k,\mathbb{C})$ is *connected* (any invertible complex matrix can be joined to 1: diagonalise or triangularise it and move each eigenvalue to 1 along a path in $\mathbb{C}^\times$ avoiding 0), g can be deformed to 1, and a homotopy of transition functions gives an isomorphic bundle. So $\mathrm{Vect}_k(S^1,\mathbb{C}) = \{I^k\}$ and

$$KU(S^1) = \mathbb{Z}, \qquad \widetilde{KU}(S^1) = 0.$$

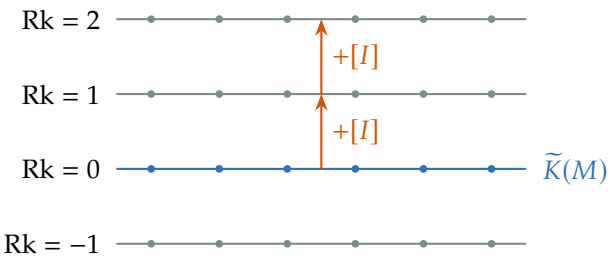


Figure 3.39: A picture of $K(M) = \widetilde{K}(M) \oplus \mathbb{Z}$. Virtual bundles sit in rows according to their virtual rank. Adding a trivial line moves straight up to the next row without changing the stable class. The reduced group $\widetilde{K}(M)$ is the row of rank zero (blue).

Worked example 3.52: K of the circle, real case

For real bundles the gluing matrix lies in $\mathrm{GL}(k,\mathbb{R})$, which has *two* components, $\det > 0$ and $\det < 0$. So there are exactly two bundles of each rank $k \geq 1$: I^k and $\mathcal{M} \oplus I^{k-1}$ (the latter glued by $\mathrm{diag}(-1,1,\dots,1)$). Hence $\widetilde{KO}(S^1)$ has two elements, 0 and $x := [\mathcal{M}] - 1$, and

$$2x \overset{(1)}{=} [\mathcal{M} \oplus \mathcal{M}] - 2 \overset{(2)}{=} [I^2] - 2 = 0,$$

(1) $2x = 2[\mathcal{M}] - 2$ and addition is $\oplus$.
(2) $\mathcal{M} \oplus \mathcal{M} \cong I^2$ (Example 3.45).

So $\widetilde{KO}(S^1) = \mathbb{Z}_2$ and $KO(S^1) = \mathbb{Z} \oplus \mathbb{Z}_2$. The $\mathbb{Z}_2$ is the orientability of the bundle (its determinant line), invisible to complex K-theory.

Worked example 3.53: A disconnected space

Let $M = \{p, q\}$ be two points. A bundle is a pair of vector spaces of dimensions (a, b), so $\mathrm{Vect}(M) = \mathbb{N} \times \mathbb{N}$ and $K(M) = \mathbb{Z} \oplus \mathbb{Z}$. With base point p, $i^*(a,b) = a$, so $\widetilde{K}(M) = \mathbb{Z}$ (generated by the bundle that is 0 over p and $\mathbb{C}$ over q). Here "Rk" is not a single number, which is why Nash assumes M connected.

3.10.7 The ring structure

K-theory has a multiplication, induced by $\otimes$. For formal differences we multiply out as for numbers:

$$\big([E] - [F]\big) \cdot \big([G] - [H]\big) := [E \otimes G \oplus F \otimes H] - [E \otimes H \oplus F \otimes G]. \tag{3.90}$$

This is what $(e - f)(g - h) = (eg + fh) - (eh + fg)$ says. It is well defined because $\otimes$ distributes over $\oplus$ and respects isomorphism (so if we change (E, F) to an equivalent pair, the product changes to an equivalent pair). The unit is $1 = [I]$, since $E \otimes I \cong E$.

Key idea: K-theory is a ring

$K(M)$ is a commutative ring with unit $[I]$, and α^* is a ring homomorphism. Rank is multiplicative, $\mathrm{Rk}(xy) = \mathrm{Rk}(x)\,\mathrm{Rk}(y)$, so the product of anything with an element of rank zero has rank zero: $\widetilde{K}(M)$ is an *ideal*. Having a product is, as Nash says, another sign that we are dealing with a cohomology theory (compare the wedge product on de Rham cohomology).

Worked example 3.54: The ring $KO(S^1)$

With $x = [\mathcal{M}] - 1$:

$$\begin{aligned} x^2 &\overset{(1)}{=} [\mathcal{M}]^2 - 2[\mathcal{M}] + 1 \\ &\overset{(2)}{=} [\mathcal{M} \otimes \mathcal{M}] - 2[\mathcal{M}] + 1 \\ &\overset{(3)}{=} 1 - 2[\mathcal{M}] + 1 \\ &= -2\big([\mathcal{M}] - 1\big) = -2x \overset{(4)}{=} 0. \end{aligned}$$

(1) Expand using the ring axioms.
(2) Multiplication in K is induced by $\otimes$.
(3) $\mathcal{M} \otimes \mathcal{M} \cong I$ (Example 3.39).
(4) $2x = 0$ (Example 3.52).

So $KO(S^1) = \mathbb{Z}[x]/(2x,\ x^2)$: additively $\mathbb{Z} \oplus \mathbb{Z}_2$, and the reduced part squares to zero.

Worked example 3.55: A preview: the ring $K(S^2)$

In Section 3.12 we will show $\widetilde{KU}(S^2) = \mathbb{Z}$, generated by $b = [H] - 1$, and that $H \oplus H \cong H^{\otimes 2} \oplus I$. Granting that,

$$\begin{aligned} b^2 &\overset{(1)}{=} [H \otimes H] - 2[H] + 1 \\ &\overset{(2)}{=} \big(2[H] - 1\big) - 2[H] + 1 = 0. \end{aligned}$$

(1) Expand $([H] - 1)^2$ and use $[H]^2 = [H \otimes H]$.
(2) $[H \otimes H] + [I] = [H \oplus H] = 2[H]$.

So $K(S^2) = \mathbb{Z}[b]/(b^2)$, i.e. $(H - 1)^2 = 0$. This innocent relation is the seed of Bott periodicity.

Reading Nash (pp. 67–70, (3.40)–(3.55))

Nash observes that $\mathrm{Vect}(M)$ is only a semigroup (3.40), defines the Grothendieck group by (3.41), specialises to bundles (3.42) and defines $K(M)$ (3.43). He proves $[E] - [F] = [H] - [I^n]$ (3.44), defines virtual dimension (3.45), and proves that stable classes correspond to elements of virtual dimension zero (3.46)–(3.47). Reduced K-theory is (3.48), Rk is (3.49), and $K(M) = \widetilde{K}(M) \oplus \mathbb{Z}$ is (3.50), proved via pullbacks (3.51)–(3.55). Two corrections: on p. 70 Nash writes "$\ker i^* = \widetilde{K}(\{m_0\})$", which should read $\ker i^* = \widetilde{K}(M)$ (the reduced K-group of a point is zero); and the surjectivity in (c) above is left implicit. The ring structure is mentioned in one sentence; we supplied the formula (3.90) and the examples.

Recap

- Grothendieck group: pairs (s, t) with $(s_1, t_1) \sim (s_2, t_2)$ iff $s_1 + t_2 + u = s_2 + t_1 + u$; the u absorbs the failure of cancellation.
- $K(M)$: formal differences $[E] - [F]$; $[E] = [F]$ iff $E \oplus I^N \cong F \oplus I^N$. Every element is $[H] - [I^n]$.
- Virtual dimension $\operatorname{rk} E - \operatorname{rk} F$; $\widetilde{K}(M)$ = stable classes = virtual dimension zero.

- $K(M) = \widetilde{K}(M) \oplus \mathbb{Z}$ via i^* (restrict to a point) and c^* (trivial bundles).
- $\otimes$ makes $K(M)$ a ring. $KU(S^1) = \mathbb{Z}$; $KO(S^1) = \mathbb{Z}[x]/(2x, x^2)$; $KU(S^2) = \mathbb{Z}[b]/(b^2)$ with $b = [H] - 1$.

3.11 Classifying spaces and the stable range

So far we have described bundles one at a time, by transition functions. There is a completely different way to describe *all* rank-k bundles at once. Build one special space B with one special bundle Q over it, so that every bundle over every M is obtained by pulling Q back along some map $f : M \to B$. Then

$$\text{bundles over } M \quad \longleftrightarrow \quad \text{maps } M \to B \text{ up to homotopy.}$$

A question about bundles becomes a question about maps, which is a question in homotopy theory. For spheres $M = S^n$ it becomes a question about homotopy groups, and those are known. This is how Nash computes $\widetilde{K}(S^n)$, and it is the gateway to Bott periodicity.

We met this idea in Chapter 1 as the classifying space BG of a group G. Here we make it completely concrete: B will be a *Grassmannian*, a space of planes, and Q will be the bundle whose fibre over a plane is that plane.

A physicist has met this construction without the name. Take a Hamiltonian $\hat{H}(x)$ depending on parameters $x \in M$, with a band of k states separated by a gap from the rest. At each x the band spans a k-dimensional subspace of the Hilbert space. The map $x \mapsto$ (that subspace) is a map from M into a Grassmannian, and the bundle of band states, whose Berry connection gives Berry phases, is the pullback of Q. We will see this explicitly for a spin-$\frac{1}{2}$ in a magnetic field.

3.11.1 Reducing the structure group: $\mathrm{GL}(k)$ to $\mathrm{O}(k)$

Nash's first remark is that a real rank-k bundle has structure group $\mathrm{GL}(k, \mathbb{R})$, "reducible to $\mathrm{O}(k)$ on choosing a metric". Here is what that means.

A **fibre metric** on E is an inner product $\langle\,,\rangle_x$ on each fibre E_x, varying smoothly with x. One always exists: on each patch use the standard dot product of the fibre coordinates, and average these with a partition of unity $\{\rho_a\}$ ($\rho_a \geq 0$, supported in U_a, $\sum_a \rho_a = 1$):

$$\langle v, w\rangle_x = \sum_a \rho_a(x)\, v_a^T w_a.$$

A positive combination of inner products is an inner product, so this is positive definite.

Given a metric, apply Gram–Schmidt to any local frame. The result is an **orthonormal frame** on each patch. Use orthonormal fibre coordinates, so that on every patch $\langle v, w\rangle = v_a^T w_a$. Then the transition functions are forced to be orthogonal:

$$\begin{aligned} v_b^T w_b &\overset{(1)}{=} \langle v, w\rangle \\ &\overset{(2)}{=} v_a^T w_a \\ &\overset{(3)}{=} (g_{ab} v_b)^T (g_{ab} w_b) \\ &\overset{(4)}{=} v_b^T \left(g_{ab}^T g_{ab}\right) w_b. \end{aligned}$$

(1) The coordinates on U_b are orthonormal.

(2) So are the coordinates on U_a.
(3) The transition rule $v_a = g_{ab} v_b$.
(4) Transpose of a product reverses the order.

Since this holds for all v_b, w_b, we get $g_{ab}^T g_{ab} = 1$: every g_{ab} lies in $\mathrm{O}(k)$. The same argument with a Hermitian metric gives $g_{ab}^\dagger g_{ab} = 1$ for complex bundles, i.e. structure group $\mathrm{U}(k)$, and with a quaternionic metric gives $\mathrm{Sp}(k)$.

Key idea: Only the compact group matters

Every real, complex or quaternionic bundle can be described by transition functions in $\mathrm{O}(k)$, $\mathrm{U}(k)$ or $\mathrm{Sp}(k)$ respectively. Topologically nothing is lost, because $\mathrm{GL}(k,\mathbb{R})$ deformation-retracts onto $\mathrm{O}(k)$ (and $\mathrm{GL}(k,\mathbb{C})$ onto $\mathrm{U}(k)$).

The retraction is the **polar decomposition**. Every invertible matrix factors uniquely as $g = O\,P$ with O orthogonal and P symmetric positive definite, namely $P = (g^T g)^{1/2}$ and $O = g P^{-1}$. The path

$$g_t = O\,P^{1-t} = O\exp\big((1-t)\log P\big), \qquad t \in [0,1],$$

runs inside $\mathrm{GL}(k,\mathbb{R})$ from $g_0 = g$ to $g_1 = O \in \mathrm{O}(k)$, and fixes the points already in $\mathrm{O}(k)$ (where $P = 1$).

Worked example 3.56: Polar decomposition of a 2×2 matrix

Take $g = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ (a shear). Then

$$g^T g = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}.$$

This has eigenvalues $\lambda_\pm = \frac{3\pm\sqrt{5}}{2}$, both positive, so $P = (g^T g)^{1/2}$ exists. Using $\sqrt{\lambda_\pm} = \frac{\sqrt{5}\pm 1}{2}$ (check: $\big(\frac{\sqrt{5}\pm1}{2}\big)^2 = \frac{6\pm 2\sqrt{5}}{4} = \lambda_\pm$) one finds

$$P = \frac{1}{\sqrt{5}}\begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}, \qquad O = g P^{-1} \overset{(1)}{=} \frac{1}{\sqrt{5}}\begin{pmatrix} 2 & 1 \\ -1 & 2 \end{pmatrix}.$$

(1) $\det P = (6-1)/5 = 1$, so $P^{-1} = \frac{1}{\sqrt{5}}\begin{pmatrix} 3 & -1 \\ -1 & 2 \end{pmatrix}$, and $gP^{-1} = \frac{1}{\sqrt{5}}\begin{pmatrix} 3-1 & -1+2 \\ -1 & 2 \end{pmatrix}$.

Check: $P^2 = \frac{1}{5}\begin{pmatrix} 5 & 5 \\ 5 & 10 \end{pmatrix} = g^T g$, and O is a rotation by $\arctan(-\frac{1}{2})$, since its columns are orthonormal and $\det O = (4+1)/5 = 1$. The shear has been deformed to a rotation without ever becoming singular.

Watch out: The metric is extra structure, but it costs nothing

Different metrics give different orthonormal frames, but the resulting $\mathrm{O}(k)$-bundles are isomorphic: two metrics are joined by the straight line $(1-t)\langle,\rangle_0 + t\langle,\rangle_1$, which is a metric for every t. So "real bundle" and "$\mathrm{O}(k)$-bundle" are interchangeable for counting purposes, and from now on we use whichever is convenient.

3.11.2 Grassmannians, concretely

Working definition 3.28 (real Grassmannian)

$\mathrm{Gr}(k,m,\mathbb{R})$ is the set of k-dimensional linear subspaces ("k-planes through the origin") of $\mathbb{R}^m$. Three equivalent ways to hold a point of it in your hand:

1. a k-plane $P \subset \mathbb{R}^m$;
2. an $m \times k$ matrix A of rank k whose columns span P, modulo $A \sim AG$ for $G \in \mathrm{GL}(k,\mathbb{R})$ [an invertible $k \times k$ real matrix] (change of basis of P);
3. the orthogonal projector Π onto P [the matrix sending each vector of $\mathbb{R}^m$ to its nearest point in P]: an $m \times m$ matrix with $\Pi^T = \Pi$, $\Pi^2 = \Pi$, $\mathrm{tr}\,\Pi = k$.

The third description is the most useful in practice: it has no redundancy, and $\Pi = A(A^TA)^{-1}A^T$ converts from the second.

The simplest case is $k = 1$: a line through the origin in $\mathbb{R}^m$. That is real projective space, $\mathrm{Gr}(1,m,\mathbb{R}) = \mathbb{RP}^{m-1}$. For $m = 2$ it is a circle, as we now check with projectors.

Worked example 3.57: $\mathrm{Gr}(1,2,\mathbb{R}) = \mathbb{RP}^1$ is a circle

The line at angle θ is spanned by $u = (\cos\theta, \sin\theta)^T$. Its projector is

$$\Pi(\theta) \overset{(1)}{=} u\,u^T = \begin{pmatrix} \cos^2\theta & \cos\theta\sin\theta \\ \cos\theta\sin\theta & \sin^2\theta \end{pmatrix} \overset{(2)}{=} \frac{1}{2}\begin{pmatrix} 1+\cos 2\theta & \sin 2\theta \\ \sin 2\theta & 1-\cos 2\theta \end{pmatrix} \overset{(3)}{=} \frac{1}{2}\left(1 + \cos 2\theta\,\sigma_3 + \sin 2\theta\,\sigma_1\right).$$

(1) For a unit vector u, uu^T is the orthogonal projector onto its span ($uu^Tu = u$, and $uu^Tw = 0$ for $w \perp u$).
(2) Double-angle formulas.
(3) In terms of the Pauli matrices $\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, $\sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$.

The projector depends only on 2θ. As θ runs from 0 to π (and the line turns through half a revolution, returning to itself), the point $(\cos 2\theta, \sin 2\theta)$ runs once round a circle. So $\mathbb{RP}^1$ is a circle, parametrised by 2θ.

As a homogeneous space. The orthogonal group $\mathrm{O}(m)$ moves k-planes around: $P \mapsto RP$ for $R \in \mathrm{O}(m)$. It moves any k-plane to any other (rotate an orthonormal basis of one onto an orthonormal basis of the other and extend). The rotations that fix the reference plane $\mathbb{R}^k \subset \mathbb{R}^m$ as a set are the block-diagonal ones, $\begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}$ with $A \in \mathrm{O}(k)$ acting inside the plane and $B \in \mathrm{O}(m-k)$ acting on its orthogonal complement. Hence

$$\mathrm{Gr}(k,m,\mathbb{R}) = \frac{\mathrm{O}(m)}{\mathrm{O}(m-k) \times \mathrm{O}(k)}. \tag{3.91}$$

This is Nash's (3.56). Counting dimensions, with $\dim \mathrm{O}(p) = p(p-1)/2$:

$$\dim \mathrm{Gr}(k,m,\mathbb{R}) \overset{(1)}{=} \tfrac{m(m-1)}{2} - \tfrac{k(k-1)}{2} - \tfrac{(m-k)(m-k-1)}{2}$$

$$\overset{(2)}{=} \tfrac{1}{2}\left[m^2 - m - k^2 + k - (m-k)^2 + (m-k)\right]$$
$$\overset{(3)}{=} \tfrac{1}{2}\left[2mk - 2k^2\right] = k(m-k).$$

(1) Dimension of a quotient group/subgroup is the difference of dimensions.
(2) Expand.
(3) $(m-k)^2 = m^2 - 2mk + k^2$; the linear terms $-m + k + (m-k)$ cancel.

There is a direct way to see $k(m-k)$: a k-plane near $\mathbb{R}^k$ is the graph of a linear map $\mathbb{R}^k \to \mathbb{R}^{m-k}$, i.e. an $(m-k)\times k$ matrix of free parameters. For $k=1$, $m=2$ we get 1, the circle.

Worked example 3.58: $\mathrm{Gr}(2,3,\mathbb{R}) \cong \mathbb{RP}^2$

A plane through the origin of $\mathbb{R}^3$ is determined by its normal line, so $\mathrm{Gr}(2,3,\mathbb{R}) \cong \mathrm{Gr}(1,3,\mathbb{R}) = \mathbb{RP}^2$, and indeed $\dim = 2\cdot 1 = 1\cdot 2$. In projector language, $\Pi_{\text{plane}} = 1 - \Pi_{\text{normal}} = 1 - nn^T$ for a unit normal n; the sign of n drops out, which is why we get $\mathbb{RP}^2 = S^2/\pm$ rather than S^2. In general $\mathrm{Gr}(k,m) \cong \mathrm{Gr}(m-k,m)$ via $P \mapsto P^\perp$, $\Pi \mapsto 1-\Pi$.

The universal bundle. Over $\mathrm{Gr}(k,m,\mathbb{R})$ there is a tautological rank-k bundle, Nash's $Q(k,m,\mathbb{R})$:

$$Q(k,m,\mathbb{R}) = \{\,(P,v) \;:\; P \in \mathrm{Gr}(k,m,\mathbb{R}),\; v \in P\,\}.$$

The fibre over the plane P *is* the plane P. It sits inside the trivial bundle $\mathrm{Gr}\times\mathbb{R}^m$, and its fibre at Π is the image of the projector Π.

Worked example 3.59: The universal line bundle over $\mathbb{RP}^1$ is the Möbius bundle

Over $\mathrm{Gr}(1,2,\mathbb{R})$, the fibre over the line at angle θ is that line, with the local nonvanishing section $s(\theta) = u(\theta) = (\cos\theta, \sin\theta)$. Follow s once around the base, i.e. from θ to $\theta+\pi$ (by Example 3.57, that is one full loop in $\mathbb{RP}^1$):

$$s(\theta+\pi) \overset{(1)}{=} (\cos(\theta+\pi), \sin(\theta+\pi))$$
$$\overset{(2)}{=} -(\cos\theta, \sin\theta) = -s(\theta).$$

(1) Definition of s.
(2) cos and sin change sign under a shift by π.

The frame comes back reversed: the gluing is -1. So $Q(1,2,\mathbb{R})$ is the Möbius bundle $\mathcal{M}$ of Example 3.35. Similarly, over $\mathrm{Gr}(1,2,\mathbb{C}) = \mathbb{CP}^1$ the universal bundle is, by definition, the tautological bundle H of Example 3.36.

3.11.3 Every bundle is a pullback of Q

Here is the construction that makes Q universal. It uses only the complement of Section 3.9.3.

Working definition 3.29 (classifying map)

Let E be a rank-k bundle over M. Choose a complement [a second bundle E' with $E\oplus E'$ trivial, which exists because M is compact], $E \oplus E' \cong I^m = M\times\mathbb{R}^m$. Then each fibre E_x is a k-plane in the *fixed* vector space $\mathbb{R}^m$. Define

$$f : M \to \mathrm{Gr}(k,m,\mathbb{R}), \qquad f(x) = E_x \subset \mathbb{R}^m.$$

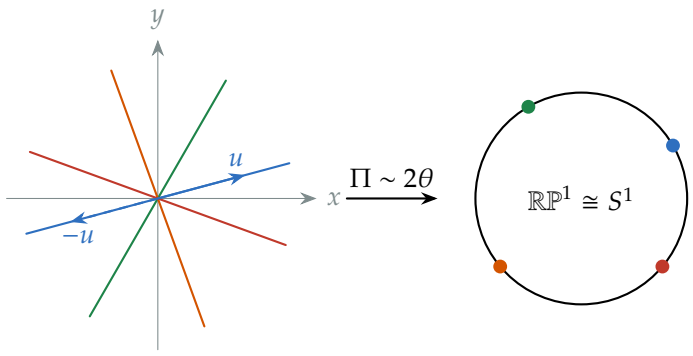


Figure 3.40: The Grassmannian of lines in the plane. Each coloured line through the origin is one point of $\mathrm{Gr}(1,2,\mathbb{R})$. The line at angle θ and the line at $\theta+\pi$ are the same line, so the natural coordinate is 2θ, and the space of lines is a circle (right). The unit vector u spanning a line comes back as $-u$ after one loop: the tautological bundle over this circle is the Möbius bundle.

In projector form, $f(x)=\Pi(x)$ is the orthogonal projector onto E_x. The map f is the **classifying map** of E.

The pullback of Q along f is E again, tautologically:

$$\begin{aligned}(f^*Q)_x &\overset{(1)}{=} Q_{f(x)}\\ &\overset{(2)}{=} \text{the plane } f(x)\\ &\overset{(3)}{=} E_x.\end{aligned}$$

(1) Definition of pullback: the fibre over x is the fibre of Q over $f(x)$.
(2) Definition of Q: its fibre over a plane is that plane.
(3) Definition of f.

Both are subbundles of $M\times\mathbb{R}^m$ with the same fibre at every point, so $f^*Q\cong E$. This is Nash's (3.57).

Worked example 3.60: The Gauss map of the sphere

For $TS^2\subset S^2\times\mathbb{R}^3$ (complement: the normal bundle, $m=3$), the classifying map sends x to its tangent plane:

$$f: S^2\to \mathrm{Gr}(2,3,\mathbb{R}),\qquad f(x)=x^{\perp},\qquad \Pi(x)=1-xx^T.$$

Composing with $\mathrm{Gr}(2,3)\cong\mathbb{RP}^2$ (plane $\mapsto$ normal line), f is the map $S^2\to\mathbb{RP}^2$, $x\mapsto\{\pm x\}$. This is the classical **Gauss map** of a surface, which records the direction of the normal at each point. Its degree (suitably defined) is what the Gauss–Bonnet theorem integrates.

Worked example 3.61: Spin-$\frac{1}{2}$ in a magnetic field: the Hopf bundle as a pullback

Take $\hat H(\vec n)=-\vec n\cdot\vec\sigma$ for a unit vector $\vec n\in S^2$. The ground state is the $+1$ eigenvector of $\vec n\cdot\vec\sigma$, and the projector onto it is

$$\Pi(\vec n)\overset{(1)}{=}\tfrac{1}{2}\big(1+\vec n\cdot\vec\sigma\big)$$

$$\overset{(2)}{=} \frac{1}{2}\begin{pmatrix} 1+n_3 & n_1 - in_2 \\ n_1 + in_2 & 1-n_3 \end{pmatrix}.$$

(1) $(\vec{n}\cdot\vec{\sigma})^2 = |\vec{n}|^2 = 1$, so the eigenvalues are ± 1, and $\frac{1}{2}(1 \pm \vec{n}\cdot\vec{\sigma})$ are the projectors onto the two eigenspaces (they are idempotent: $\frac{1}{4}(1 + 2\vec{n}\cdot\vec{\sigma} + 1) = \frac{1}{2}(1 + \vec{n}\cdot\vec{\sigma})$; their trace is 1).
(2) Write out the Pauli matrices.

So $\vec{n} \mapsto \Pi(\vec{n})$ is a map $S^2 \to \mathrm{Gr}(1,2,\mathbb{C}) = \mathbb{CP}^1$. It is a bijection (the Bloch sphere: every pure state of a qubit is $\Pi(\vec{n})$ for exactly one $\vec{n}$). The ground-state bundle is the pullback of $Q(1,2,\mathbb{C}) = H$ along a bijection, i.e. it is H itself. The Berry phase of the ground state around a loop on S^2 is half the solid angle, and its integrated Berry curvature over S^2 is 2π times the Chern number ∓ 1 (the sign depends on orientation conventions): a monopole of unit charge in parameter space.

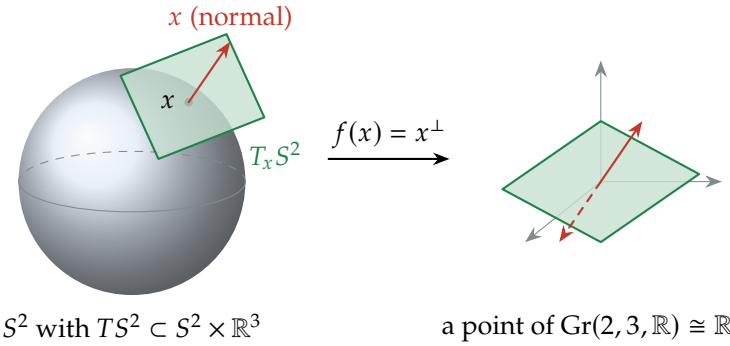


Figure 3.41: The classifying map of the tangent bundle of S^2. At each point x the tangent plane is a 2-plane in $\mathbb{R}^3$; translate it to the origin and you have a point of the Grassmannian of planes in $\mathbb{R}^3$. The plane is determined by its normal line (red), and $\pm x$ give the same line, so the target is $\mathbb{RP}^2$. Pulling back the tautological plane bundle along this map returns TS^2.

How big must m be? A dimension count

Nash states the universal property for $m \geq n + k + 1$, where $n = \dim M$. Where does that number come from? Two facts are needed.

(i) Every rank-k bundle embeds in I^m once $m \geq n + k$. An embedding is the same as m global sections $s_1, \ldots, s_m$ which span E_x at every x (then E' is the kernel of the resulting map, and $E_x = \mathrm{span}\{s_i(x)\}$ sits in $\mathbb{R}^m$ via the dual map). Written in a local frame, the sections form a $k \times m$ matrix $S(x)$, and they fail to span exactly where $S(x)$ has rank $< k$. Count the codimension of the bad matrices:

$$\begin{aligned} \mathrm{codim}\{S : \mathrm{rank}\, S \leq k-1\} &\overset{(1)}{=} km - (k-1)\big(k + m - (k-1)\big) \\ &\overset{(2)}{=} km - (k-1)(m+1) \\ &= m - k + 1. \end{aligned}$$

(1) The $k \times m$ matrices of rank r form a space of dimension $r(k+m-r)$ (choose the r-dimensional row space, $r(m-r)$ parameters, then k rows in it, kr parameters); take $r = k-1$.
(2) Simplify.

A generic n-parameter family of matrices $S(x)$ misses a set of codimension $m - k + 1$ as soon as $m - k + 1 > n$, i.e. $m \geq n + k$. So generic sections span everywhere.

(ii) Any two classifying maps are homotopic once $m \geq n + k + 1$. Two embeddings are joined by a family of embeddings over $M \times [0, 1]$, which has dimension $n + 1$. Repeating the count with $n \to n + 1$ gives the condition $m \geq n + k + 1$. A homotopy between embeddings is a homotopy between the classifying maps.

Conversely, homotopic maps pull back isomorphic bundles (the homotopy invariance of Section 3.10.6). Putting (i), (ii) and this together:

$$\mathrm{Vect}_k(M) \;\cong\; [M, \mathrm{Gr}(k, m, \mathbb{R})] \qquad (m \geq n + k + 1), \tag{3.92}$$

where $[A, B]$ denotes homotopy classes of maps $A \to B$. Only the homotopy class of f matters.

3.11.4 The infinite Grassmannian and $BO(k)$

The bound on m depends on $\dim M$. To get one space that works for all M at once, let $m \to \infty$. The inclusion $\mathbb{R}^m \subset \mathbb{R}^{m+1}$ (as the first m coordinates) makes every k-plane in $\mathbb{R}^m$ a k-plane in $\mathbb{R}^{m+1}$:

$$\mathrm{Gr}(k, m, \mathbb{R}) \subset \mathrm{Gr}(k, m + 1, \mathbb{R}) \subset \mathrm{Gr}(k, m + 2, \mathbb{R}) \subset \cdots \tag{3.93}$$

The union is the **infinite Grassmannian**

$$\mathrm{Gr}(k, \infty, \mathbb{R}) = \bigcup_m \mathrm{Gr}(k, m, \mathbb{R}), \tag{3.94}$$

the space of k-planes in $\mathbb{R}^\infty = \bigcup_m \mathbb{R}^m$ (sequences with finitely many nonzero entries). A map from a compact M lands in some finite $\mathrm{Gr}(k, m, \mathbb{R})$, so (3.92) gives, for every compact M,

$$\mathrm{Vect}_k(M) \cong [M, \mathrm{Gr}(k, \infty, \mathbb{R})] = [M, BO(k)]. \tag{3.95}$$

These are Nash's (3.60)–(3.61). The notation $BO(k)$ is the classifying space of the group $\mathrm{O}(k)$ from Chapter 1.

Physical picture: Why this is the classifying space BG

In Chapter 1, BG was the base of a principal G-bundle $EG \to BG$ with *contractible* total space. Here EG is the space of orthonormal k-frames in $\mathbb{R}^\infty$ (the Stiefel manifold $V(k, \infty)$). $\mathrm{O}(k)$ acts on frames by rotating them within their plane, freely, and the quotient ("forget the frame, keep the plane") is $\mathrm{Gr}(k, \infty)$. The frame space is contractible because there is infinite room: any frame can be slid, one coordinate at a time, into the unused far-away coordinates and then rotated to a standard frame. For $k = 1$ this says: the unit sphere $S^\infty \subset \mathbb{R}^\infty$ is contractible, which is false for every finite S^{m-1}.

Worked example 3.62: Real line bundles and $\mathbb{RP}^\infty$

For $k = 1$: $BO(1) = \mathbb{RP}^\infty$, and real line bundles over M correspond to $[M, \mathbb{RP}^\infty]$. For $M = S^1$, a based loop in $\mathbb{RP}^\infty$ lifts to a path in S^∞ which either closes up or ends at the antipode; these two cases are the two elements of $\pi_1(\mathbb{RP}^\infty) = \mathbb{Z}_2$. So there are exactly two real line bundles over S^1, I and $\mathcal{M}$, matching Example 3.35 and Example 3.59. (In general $[M, \mathbb{RP}^\infty] = H^1(M; \mathbb{Z}_2)$, and the class of a line bundle there is its first Stiefel–Whitney class w_1, the obstruction to orienting it.)

Worked example 3.63: Complex line bundles and $\mathbb{CP}^\infty$: monopoles

For complex line bundles, $BU(1) = \mathrm{Gr}(1,\infty,\mathbb{C}) = \mathbb{CP}^\infty$. Line bundles over S^2 are classified by $[S^2, \mathbb{CP}^\infty] = \pi_2(\mathbb{CP}^\infty)$. The fibration $S^1 \to S^\infty \to \mathbb{CP}^\infty$ with contractible S^∞ gives $\pi_2(\mathbb{CP}^\infty) \cong \pi_1(S^1) = \mathbb{Z}$: one line bundle $H^{\otimes n}$ for each integer n, the monopole charges. (In general $[M, \mathbb{CP}^\infty] = H^2(M;\mathbb{Z})$, and the class is the first Chern class c_1.)

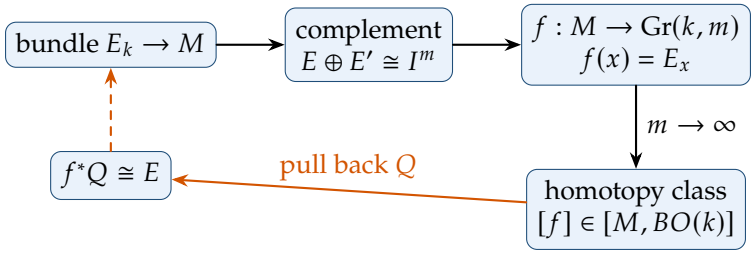


Figure 3.42: The logic of classifying spaces. A bundle has a complement inside a big trivial bundle; that makes each fibre a plane in a fixed $\mathbb{R}^m$, i.e. a point of the Grassmannian. The resulting map, up to homotopy, is a complete invariant: pulling back the tautological bundle Q recovers the bundle up to isomorphism.

3.11.5 The stable range theorem

Now raise the rank. What happens to $\mathrm{Vect}_k(M)$ as k grows? Nash's answer: it stops changing once $k > \dim M$, and it then *is* the reduced K-group.

Key idea: Stable range theorem, real case (Nash, p. 72)

Let $n = \dim M$. If E_k is a real bundle of rank $k > n$, then

$$E_k \cong F_n \oplus I^{k-n}$$

for some bundle F_n of rank n. Moreover, for $k > n$, two rank-k bundles that are stably equivalent are isomorphic. Such E_k are said to be in the **stable range**.

Why: a section of a high-rank bundle need not vanish. Pick a generic section s of E_k. In a local frame s is k real functions of n variables. Its zeros are where all k functions vanish at once:

$$\dim\{x : s(x) = 0\} \overset{(1)}{=} n - k.$$

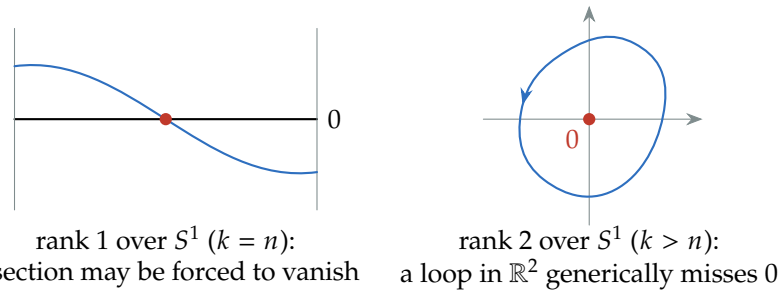
(1) k independent equations in n unknowns cut out a set of dimension $n - k$ (generic transversality); a negative dimension means the set is empty.

So if $k > n$ a generic section never vanishes. A nonvanishing section spans a trivial line subbundle $I \subset E_k$ (Section 3.8.1), and with a fibre metric $E_k = I \oplus s^\perp$. Now $s^\perp$ has rank $k-1$. If still $k - 1 > n$, repeat. We stop at rank n:

$$E_k \cong E_{k-1} \oplus I \cong E_{k-2} \oplus I^2 \cong \cdots \cong F_n \oplus I^{k-n}.$$

Why stable equivalence implies isomorphism for $k > n$. Suppose $E \oplus I \cong F \oplus I$ with $\mathrm{rk}\, E = \mathrm{rk}\, F = k > n$. The bundle $G := E \oplus I$ has two nonvanishing sections: $s_0 =$ the I in $E \oplus I$, and $s_1 =$ the I in $F \oplus I$ transported by the isomorphism, with $E \cong s_0^\perp$ and $F \cong s_1^\perp$. Join them by a family s_t, $t \in [0,1]$, of sections. This is a section of a rank-$(k+1)$ bundle over the

$(n+1)$-dimensional space $M \times [0,1]$; generically it has no zeros if $k+1 > n+1$, i.e. $k > n$. Then $s_t^{\perp}$ is a bundle over $M \times [0,1]$ restricting to E at $t=0$ and F at $t=1$, and homotopy invariance gives $E \cong F$. Induction handles $E \oplus I^j \cong F \oplus I^j$.



rank 1 over S^1 ($k = n$): a section may be forced to vanish

rank 2 over S^1 ($k > n$): a loop in $\mathbb{R}^2$ generically misses 0

Figure 3.43: The dimension count behind the stable range. Left: a section of a real line bundle over a circle is one function of one variable; for the Möbius bundle it must change sign and so it vanishes (red dot). Right: a section of a rank-2 bundle over a circle traces a loop in the fibre $\mathbb{R}^2$; a small push moves it off the origin, so the section is nowhere zero and splits off a trivial line.

The K-theory meaning. Combine the theorem with Section 3.10.5. The map

$$\mathrm{Vect}_k(M) \longrightarrow \widetilde{K}(M), \qquad E_k \longmapsto [E_k] - [I^k],$$

is *onto* for $k \geq n$ (every stable class is $[H] - [I^N]$; if $\mathrm{rk}\, H < k$ add trivial lines, if $\mathrm{rk}\, H > k$ split them off with the theorem) and *one-to-one* for $k > n$ (equal images mean stably equivalent, hence isomorphic). So

$$\widetilde{K}(M) \;\cong\; \mathrm{Vect}_k(M) \;\cong\; [M, \mathrm{Gr}(k,\infty,\mathbb{R})] \qquad (k > \dim M). \tag{3.96}$$

This is Nash's (3.62). Nash sums it up: "there is nothing to be gained by considering bundles of very high rank". In the stable range, bundles and their K-theory classes are the same thing.

Worked example 3.64: Inside and outside the stable range over S^1

Over S^1 ($n=1$) every real bundle of rank $k \geq 2$ is in the stable range. By Example 3.52 there are exactly two of each rank, I^k and $\mathcal{M} \oplus I^{k-1}$, and $\widetilde{KO}(S^1) = \mathbb{Z}_2$ also has two elements. The theorem says $\mathcal{M} \oplus \mathcal{M}$ (rank $2 > 1$) must be one of these two; its determinant is $\det(\mathrm{diag}(-1,-1)) = +1$, so it is I^2, as we found by hand in Example 3.45.

Worked example 3.65: Outside the stable range bundles are finer than K-theory

Take oriented real rank-2 bundles over S^2 ($k = n = 2$, not in the stable range). Such a bundle is the same as a complex line bundle (rotation by 90° is multiplication by i), so by Example 3.63 there is one for each integer: $H^{\otimes d}$, with TS^2 at $d=-2$. There are infinitely many. But we saw in Section 3.9 that $TS^2 \oplus I \cong I^3$, so TS^2 and I^2 become equal in $\widetilde{KO}(S^2)$. Adding one trivial line kills the even charges. We will find in Section 3.12 that $\widetilde{KO}(S^2) = \mathbb{Z}_2$: only d mod 2 survives stabilisation. Outside the stable range, the set of bundles is bigger than the K-group.

Watch out: Stable versus unstable information

K-theory deliberately throws away information that is only visible at low rank, such as the Euler class of TS^2 (which is $\chi(S^2) = 2$ and detects that TS^2 is non-trivial). Invariants that

survive adding I (Chern classes, Pontryagin classes, Stiefel–Whitney classes, the Chern character) are *stable*; the Euler class is not. The index theorem involves only stable invariants, which is one reason K-theory is the right language for it.

Try it: How many trivial lines can a bundle over S^3 shed?

A real rank-5 bundle over S^3: by the theorem it is $F_3 \oplus I^2$. Can one always go further, to $F_2 \oplus I^3$?
Answer: Not by the dimension count, which only guarantees a nonvanishing section when rk > dim. (In fact every bundle over S^3 is trivial, because $\pi_2(\mathrm{O}(k)) = 0$; see Section 3.12. But that is special to S^3.)

Reading Nash (pp. 70–72, (3.56)–(3.62))

Nash reduces $\mathrm{GL}(k,\mathbb{R})$ to $\mathrm{O}(k)$ with a metric, introduces the Grassmannian (3.56) and the universal bundle $Q(k,m,\mathbb{R})$, states the classifying property (3.57) for $m \geq n+k+1$, passes to $\mathrm{Gr}(k,\infty,\mathbb{R}) = BO(k)$ via the inclusions (3.58)–(3.61), and states the stable range theorem, referring to Husemoller for the proof by induction on k. Its K-theory translation is (3.62). He writes that E_k is stably equivalent to "F_n of lower rank n where $n = \dim M + 1$"; the rank of F_n is $n = \dim M$, as in the statement of the theorem. We have added the polar decomposition, three concrete descriptions of the Grassmannian with the dimension count $k(m-k)$, the explicit classifying maps for TS^2 and the Hopf bundle, the dimension counts behind $m \geq n+k+1$ and $k > n$, and the $k = 1$ cases $\mathbb{RP}^\infty$, $\mathbb{CP}^\infty$.

Recap

- A fibre metric reduces $\mathrm{GL}(k,\mathbb{R})$ to $\mathrm{O}(k)$ ($\mathrm{U}(k)$, $\mathrm{Sp}(k)$ for $\mathbb{C}$, $\mathbb{H}$); polar decomposition shows nothing topological is lost.
- $\mathrm{Gr}(k,m,\mathbb{R}) = \mathrm{O}(m)/(\mathrm{O}(m-k)\times\mathrm{O}(k))$, dimension $k(m-k)$; points are projectors $\Pi = \Pi^T = \Pi^2$, $\operatorname{tr}\Pi = k$.
- Tautological bundle Q: fibre over a plane is the plane. Over $\mathbb{RP}^1$ it is Möbius; over $\mathbb{CP}^1$ it is H.
- Classifying map $f(x) = E_x \subset \mathbb{R}^m$; $f^*Q \cong E$; $\mathrm{Vect}_k(M) = [M, BO(k)]$, $BO(k) = \mathrm{Gr}(k,\infty,\mathbb{R})$.
- Stable range $k > \dim M$: $E_k \cong F_n \oplus I^{k-n}$, and $\mathrm{Vect}_k(M) \cong \widetilde{K}(M) = [M, \mathrm{Gr}(k,\infty)]$.

3.12 Bundles over spheres: clutching functions and $\widetilde{K}(S^n)$

We now have a formula, (3.96), that turns K-theory into homotopy theory. Nash applies it to the simplest spaces with interesting topology, the spheres S^n. This section computes $\widetilde{K}(S^n)$ for $n \leq 4$ in the real, complex and quaternionic cases, and shows where the answers come from. Two ideas do all the work:

1. a bundle over a sphere is determined by *one* transition function on the equator, the **clutching function**;
2. the homotopy groups of $\mathrm{O}(k)$, $\mathrm{U}(k)$, $\mathrm{Sp}(k)$ stop depending on k when k is large.

For a physicist this is the familiar statement that the topological sectors of a gauge theory on a sphere are classified by the winding of the gauge transformation that relates the "north" and "south" gauges. The monopole (S^2, winding in U(1)) and the instanton (S^4, winding in SU(2)) are the two famous cases.

3.12.1 $\widetilde{K}(S^n)$ as a homotopy group

Put $M = S^n$ in (3.96). For $k > n$,

$$\widetilde{K}(S^n) = [S^n, \mathrm{Gr}(k, \infty, \mathbb{R})] \cong \pi_n\big(\mathrm{Gr}(k, \infty, \mathbb{R})\big). \tag{3.97}$$

This is Nash's (3.63). Recall that the **homotopy group** $\pi_n(X)$ is the set of maps $S^n \to X$ sending a chosen base point of S^n to a chosen base point x_0 of X, up to homotopies that keep the base point fixed. The group law glues two spheres at their base points.

Watch out: Free versus based homotopy

$[S^n, X]$ (free homotopy classes) and $\pi_n(X)$ (based classes) are not quite the same. A based class can be dragged around a loop in X while the base point travels with it; this is an action of $\pi_1(X)$ on $\pi_n(X)$, and the free classes are the orbits. If X is simply connected (as $BU(k) = \mathrm{Gr}(k, \infty, \mathbb{C})$ is) nothing changes. For $BO(k)$, $\pi_1 = \mathbb{Z}_2$ (orientation), and the identification in (3.97) really counts bundles *with a chosen trivialisation of the fibre over the base point*. Nash calls this "the minor technical point". For the groups we compute below the distinction never changes the answer.

3.12.2 The clutching construction

Split S^n into two closed hemispheres ("caps") D_+ and D_-, each a disc D^n, meeting along the equator S^{n-1}. Thicken them slightly to open sets $U \supset D_+$ and $V \supset D_-$; the overlap $U \cap V$ is a band around the equator, which deformation-retracts onto S^{n-1}.

Step 1: every bundle over a disc is trivial. A disc is contractible: the map $r_t(x) = (1-t)x$ shrinks it to its centre. Pulling a bundle E back along the homotopy r_t from $r_0 = \mathrm{id}$ to $r_1 =$(constant map) gives, by homotopy invariance,

$$E \overset{(1)}{=} r_0^* E \overset{(2)}{\cong} r_1^* E \overset{(3)}{=} D \times E_0 = I^k.$$

(1) r_0 is the identity.
(2) Homotopic maps pull back isomorphic bundles.
(3) The pullback along a constant map with value 0 is the constant family of vector spaces E_0: a product.

Step 2: glue. So $E|_U \cong U \times \mathbb{R}^k$ and $E|_V \cong V \times \mathbb{R}^k$, and E is determined by a single transition function

$$g : U \cap V \to \mathrm{O}(k) \qquad \rightsquigarrow \qquad g : S^{n-1} \to \mathrm{O}(k)$$

(restrict to the equator; the band retracts onto it). This g is the **clutching function**.

Working definition 3.30 (clutching)

Given a map $g : S^{n-1} \to G$, where $G = \mathrm{O}(k)$, $\mathrm{U}(k)$ or $\mathrm{Sp}(k)$ [the matrices preserving lengths in $\mathbb{F}^k$ for $\mathbb{F} = \mathbb{R}, \mathbb{C}, \mathbb{H}$ respectively], the bundle E_g over S^n is obtained by taking $D_+ \times \mathbb{F}^k$ and $D_- \times \mathbb{F}^k$ and identifying $(x, v) \in \partial D_- \times \mathbb{F}^k$ with $(x, g(x)v) \in \partial D_+ \times \mathbb{F}^k$ for each x on the equator. *To test whether* $E_g \cong E_{g'}$: check whether g and g' are homotopic [one can be deformed continuously into the other] as maps $S^{n-1} \to G$ (after normalising both to send a base point to 1).

Why only the homotopy class of g matters, and why constants can be normalised away: a change of trivialisation on the two caps is a pair of maps $h_\pm : D_\pm \to G$, and changes g to $g'(x) = h_+(x)\, g(x)\, h_-(x)^{-1}$. Since each disc is contractible, h_+ restricted to the equator is homotopic to the constant $h_+(0)$ (restrict h_+ to the circles of radius $1 - t$), and likewise for h_-. So

$$g' \;\simeq\; h_+(0)\, g\, h_-(0)^{-1},$$

a constant times g times a constant. If G is connected, the constants can be moved to 1 and drop out. For $G = \mathrm{O}(k)$ we fix them by demanding that g send a base point to 1; this is the "chosen trivialisation at the base point" of the pitfall above. Conversely, a homotopy g_t gives a bundle E_{g_t} over $S^n \times [0,1]$, so $E_{g_0} \cong E_{g_1}$. Hence

$$\mathrm{Vect}_k(S^n) \;\cong\; \pi_{n-1}\big(\mathrm{O}(k)\big) \quad \text{(based, i.e. with a framing at the base point).} \tag{3.98}$$

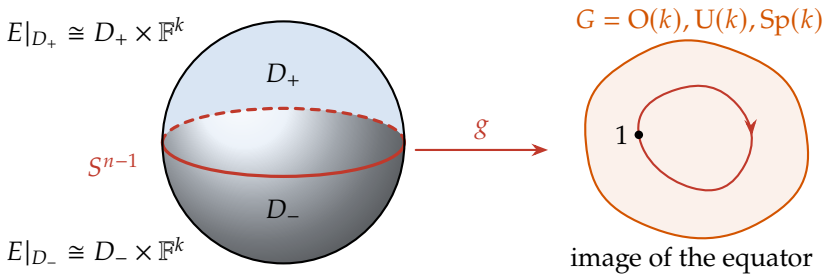


Figure 3.44: The clutching construction. Over each cap the bundle is a product. All the information is in how the two products are glued along the equator S^{n-1}: a map g from the equator into the structure group. Deforming the loop (red) inside G does not change the bundle, so the bundle is labelled by the homotopy class of g, an element of $\pi_{n-1}(G)$.

Worked example 3.66: Clutching over S^1 and S^2

$n = 1$. The "equator" of S^1 is S^0, two points. Normalising $g = 1$ at one point, a real bundle is fixed by g(other point) $\in \mathrm{O}(k)$, up to homotopy, i.e. by the component $\det g = \pm 1$. So $\mathrm{Vect}_k(S^1, \mathbb{R}) \cong \pi_0(\mathrm{O}(k)) = \mathbb{Z}_2$: I^k or $\mathcal{M} \oplus I^{k-1}$, exactly Example 3.52.
$n = 2$, *complex line bundles.* $g : S^1 \to \mathrm{U}(1)$ is a loop $e^{i\phi} \mapsto g(e^{i\phi})$ on the unit circle, classified by its winding number

$$d = \frac{1}{2\pi i} \oint_{S^1} g^{-1} \mathrm{d}g \in \mathbb{Z} = \pi_1(\mathrm{U}(1)).$$

From Section 3.8: H has $g = e^{i\phi}$, $d = 1$; $H^{\otimes d}$ has $g = e^{id\phi}$; TS^2 has $g = -e^{-2i\phi} \simeq e^{-2i\phi}$, $d = -2$. Every line bundle over S^2 is some $H^{\otimes d}$, as promised in Example 3.38.

Let us verify the winding-number integral for $g = e^{id\phi}$, since it is the prototype of every

topological charge in this book:

$$\frac{1}{2\pi i}\oint g^{-1}\mathrm{d}g \overset{(1)}{=} \frac{1}{2\pi i}\int_0^{2\pi} e^{-id\phi}\,(id)\,e^{id\phi}\,\mathrm{d}\phi$$
$$\overset{(2)}{=} \frac{id}{2\pi i}\cdot 2\pi = d.$$

(1) $\mathrm{d}(e^{id\phi}) = id\,e^{id\phi}\mathrm{d}\phi$.
(2) The integrand is the constant id.

The integral is unchanged by any deformation of g (it is an integer depending continuously on g), which is the homotopy invariance in action.

3.12.3 $[S^n, BG] = [S^{n-1}, G]$: loop spaces

We now have two descriptions of $\mathrm{Vect}_k(S^n)$: maps $S^n \to BG$ (classifying maps) and maps $S^{n-1} \to G$ (clutching functions). Nash digresses to explain why they agree, for any group G.

Working definition 3.31 (loop space and suspension)

The **loop space** ΩX of a based space (X, x_0) [a space with one chosen point x_0] is the space of loops $\gamma : [0,1] \to X$ with $\gamma(0) = \gamma(1) = x_0$. The **reduced suspension** SX is $X \times [0,1]$ with the top $X \times \{1\}$, the bottom $X \times \{0\}$ and the segment $\{x_0\} \times [0,1]$ all collapsed to a single point. For spheres, $SS^{n-1} \cong S^n$.

The basic relation (Nash's (1.22) in Chapter 1) is the **adjunction**

$$[SX, Y] \cong [X, \Omega Y]. \tag{3.99}$$

Why: a point of SX is a pair (x,t). A map $F : SX \to Y$ can be read as: for each x, the path $t \mapsto F(x,t)$, which starts and ends at the base point (because the top and bottom are collapsed). That is a map $X \to \Omega Y$, $x \mapsto F(x,\cdot)$, and conversely. With $X = S^{n-1}$ this gives $\pi_n(Y) = [S^n, Y]_0 = [S^{n-1}, \Omega Y]_0 = \pi_{n-1}(\Omega Y)$.

Now the chain Nash writes, for a group G:

$$[S^{n-1}, G] \overset{(1)}{=} [S^n, BG]$$
$$\overset{(2)}{=} [SS^{n-1}, BG]$$
$$\overset{(3)}{=} [S^{n-1}, \Omega BG]. \tag{3.100}$$

(1) Both sides classify G-bundles over S^n: clutching functions on the left, classifying maps on the right.
(2) $S^n \cong SS^{n-1}$.
(3) The adjunction (3.99).

This holds for every n, which suggests (and it is true) that

$$\Omega BG \;\simeq\; G.$$

Nash: "think of the loop space operand Ω as being a kind of homotopic inverse to the classifying space operand B".

Physical picture: A loop in BG is a group element: holonomy

Recall $EG \to BG$ is a principal G-bundle with contractible total space. Start at a point e in the fibre over the base point and lift a loop γ in BG to a path in EG. It ends in the same fibre, at some point $e \cdot g$. The group element g is the *holonomy* of the loop, just as a Wilson loop in gauge theory records the group element picked up by parallel transport around a closed path (Chapter T3). Contractibility of EG makes this a homotopy equivalence $\Omega BG \to G$. For a bundle over $S^n = SS^{n-1}$, each point x of the equator gives a loop in BG (up over D_+ and back down over D_- through x), and its holonomy is the clutching value $g(x)$.

Combining (3.97) and (3.100) with $G = \mathrm{O}(k)$:

$$\boxed{\widetilde{K}(S^n) \cong \pi_{n-1}\big(\mathrm{O}(k)\big), \qquad k > n.} \tag{3.101}$$

This is Nash's (3.64). All the K-theory of spheres is in the homotopy groups of the classical groups.

3.12.4 Stability of the homotopy groups of $\mathrm{O}(k)$

Nash quotes from Husemoller that $\pi_{n-1}(\mathrm{O}(k)) = \pi_{n-1}(\mathrm{O}(k+1))$ for $k > n$. This is the group-theoretic face of the stable range theorem, and it has a short explanation.

The group $\mathrm{O}(k+1)$ acts on the unit sphere $S^k \subset \mathbb{R}^{k+1}$, transitively, and the rotations fixing the north pole e_{k+1} form $\mathrm{O}(k)$. So the map

$$\mathrm{O}(k+1) \longrightarrow S^k, \qquad R \longmapsto R\, e_{k+1}$$

("where does the north pole go?") has fibre $\mathrm{O}(k)$: it is a **fibration** $\mathrm{O}(k) \to \mathrm{O}(k+1) \to S^k$. A fibration $F \to E \to B$ has a long exact sequence of homotopy groups,

$$\cdots \to \pi_{j+1}(B) \to \pi_j(F) \to \pi_j(E) \to \pi_j(B) \to \pi_{j-1}(F) \to \cdots .$$

(A physicist's reading: a sphere in E projects to a sphere in B; if B has no spheres of that dimension to wrap, the sphere can be pushed into a single fibre.) For $B = S^k$, $\pi_j(S^k) = 0$ whenever $j < k$ (a lower- dimensional sphere in S^k can always be moved off a point and then shrunk through the resulting $\mathbb{R}^k$). The segment of the sequence

$$\underbrace{\pi_{j+1}(S^k)}_{0 \text{ if } j+1<k} \to \pi_j(\mathrm{O}(k)) \to \pi_j(\mathrm{O}(k+1)) \to \underbrace{\pi_j(S^k)}_{0 \text{ if } j<k}$$

then says that the middle map is an isomorphism whenever $j \le k-2$. For $j = n-1$ this is $k \ge n+1$, i.e. $k > n$: precisely the stable range. The common value is written $\pi_j(\mathrm{O})$, the homotopy of the **infinite orthogonal group** $\mathrm{O} = \mathrm{O}(\infty) = \bigcup_k \mathrm{O}(k)$.

The same argument with $\mathrm{U}(k) \to \mathrm{U}(k+1) \to S^{2k+1}$ (acting on the unit sphere of $\mathbb{C}^{k+1}$) and $\mathrm{Sp}(k) \to \mathrm{Sp}(k+1) \to S^{4k+3}$ (unit sphere of $\mathbb{H}^{k+1}$) gives

$$\pi_j(\mathrm{U}(k)) \cong \pi_j(\mathrm{U}(k+1)) \ \ (j \le 2k-1), \qquad \pi_j(\mathrm{Sp}(k)) \cong \pi_j(\mathrm{Sp}(k+1)) \ \ (j \le 4k+1). \tag{3.102}$$

The spheres are bigger, so the groups stabilise sooner.

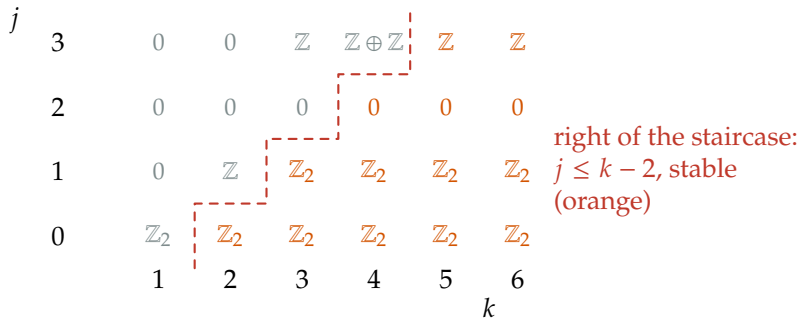


Figure 3.45: The homotopy groups $\pi_j(\mathrm{O}(k))$ for small j and k. Reading along a row, the entries settle down once $k \geq j+2$ (orange, to the right of the red staircase). The unstable entries (grey) are accidents of small groups: $\pi_1(\mathrm{O}(2)) = \mathbb{Z}$ because SO(2) is a circle, and $\pi_3(\mathrm{O}(4)) = \mathbb{Z} \oplus \mathbb{Z}$ because SO(4) is locally $\mathrm{SU}(2) \times \mathrm{SU}(2)$. K-theory sees only the stable values.

3.12.5 The real table: $\widetilde{KO}(S^n)$ for $n \leq 4$

Nash lists the result (his (3.65)):

$$\begin{aligned} & & \widetilde{KO}(S^0) &= \mathbb{Z} \\ \pi_0(\mathrm{O}(k)) = \mathbb{Z}_2,\ k \geq 1 &\Rightarrow & \widetilde{KO}(S^1) &= \mathbb{Z}_2 \\ \pi_1(\mathrm{O}(k)) = \mathbb{Z}_2,\ k \geq 3 &\Rightarrow & \widetilde{KO}(S^2) &= \mathbb{Z}_2 \\ \pi_2(\mathrm{O}(k)) = 0,\ k \geq 4 &\Rightarrow & \widetilde{KO}(S^3) &= 0 \\ \pi_3(\mathrm{O}(k)) = \mathbb{Z},\ k \geq 5 &\Rightarrow & \widetilde{KO}(S^4) &= \mathbb{Z} \end{aligned} \tag{3.103}$$

Each line deserves an explanation.

$n = 0$. S^0 is two points $\{p, q\}$. A bundle is a pair of vector spaces of dimensions (a, b), $K(S^0) = \mathbb{Z} \oplus \mathbb{Z}$, and with base point p the reduced group is $\{(0, b)\} \cong \mathbb{Z}$ (the disconnected example in Section 3.10.6). Here the "rank" is not constant, which is why $\widetilde{K}(S^0)$ is not zero.

$n = 1$: $\pi_0(\mathrm{O}(k)) = \mathbb{Z}_2$. $\mathrm{O}(k)$ has two components, det $= +1$ ($\mathrm{SO}(k)$) and det $= -1$. The generator of $\widetilde{KO}(S^1)$ is $[\mathcal{M}] - 1$.

$n = 2$: $\pi_1(\mathrm{O}(k)) = \pi_1(\mathrm{SO}(k)) = \mathbb{Z}_2$ for $k \geq 3$. This is the famous fact that a 2π rotation is not deformable to the identity but a 4π rotation is: the reason spinors exist (Chapter T4). We show it in Example 3.67 below. The generator of $\widetilde{KO}(S^2)$ is the underlying real rank-2 bundle of H (clutching $e^{i\phi} \in \mathrm{U}(1) = \mathrm{SO}(2) \subset \mathrm{SO}(k)$, a single 2π turn), minus 2.

$n = 3$: $\pi_2(\mathrm{O}(k)) = 0$. In fact $\pi_2(G) = 0$ for every compact Lie group G (a theorem of É. Cartan). So every real bundle over S^3 is trivial (with a framing at the base point).

$n = 4$: $\pi_3(\mathrm{O}(k)) = \mathbb{Z}$ for $k \geq 5$. Start with SO(3). Its double cover is $\mathrm{SU}(2) = S^3$, and a covering map does not change π_j for $j \geq 2$ (a sphere S^j with $j \geq 2$ is simply connected, so it lifts), so $\pi_3(\mathrm{SO}(3)) = \pi_3(S^3) = \mathbb{Z}$, the degree of a map $S^3 \to S^3$. This $\mathbb{Z}$ survives into SO(5) and beyond. The generator of $\widetilde{KO}(S^4)$ can be taken to be the underlying real rank-4 bundle of the quaternionic Hopf bundle (the BPST instanton bundle; see Example 3.70), minus 4.

Worked example 3.67: $\pi_1(\mathrm{SO}(3)) = \mathbb{Z}_2$, explicitly

The 2π loop is not contractible. Let $R(t)$ be rotation about the z-axis by angle $2\pi t$, $t \in [0,1]$, a closed loop at 1 in SO(3). The map $\mathrm{SU}(2) \to \mathrm{SO}(3)$, $U \mapsto R_U$ with $R_U(\vec{x}\cdot\vec{\sigma}) = U(\vec{x}\cdot\vec{\sigma})U^\dagger$, is two-to-one ($\pm U \mapsto$ the same rotation). Lift $R(t)$ to SU(2):

$$U(t) \overset{(1)}{=} \exp\left(-i\pi t\,\sigma_3\right) = \begin{pmatrix} e^{-i\pi t} & 0 \\ 0 & e^{i\pi t} \end{pmatrix}, \qquad U(1) \overset{(2)}{=} -1.$$

(1) $e^{-i\theta\sigma_3/2}$ rotates by angle θ about z; here $\theta = 2\pi t$.
(2) $e^{\mp i\pi} = -1$.

The lift starts at 1 and ends at -1: it is not a closed loop in SU(2). If $R(t)$ could be shrunk to the constant loop, the lift would deform along with it (lifts vary continuously) while keeping its endpoints in the discrete set $\{\pm 1\}$, so the endpoint would stay -1; but the constant loop lifts to the constant path, ending at $+1$. Contradiction.
The 4π loop is contractible. Let $R_{\vec{a}}(\theta)$ denote rotation by θ about the unit axis $\vec{a}$. Consider the family of loops, for $s \in [0,1]$,

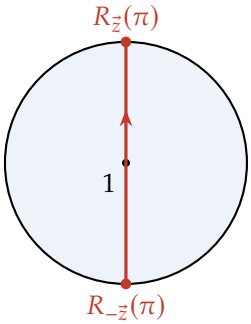
$$\gamma_s: \quad \text{first } R_{\vec{z}}(2\pi t),\ t\in[0,1]; \quad \text{then } R_{\vec{a}(s)}(2\pi t),\ t\in[0,1],$$

where $\vec{a}(s)$ tilts from $\vec{a}(0) = \vec{z}$ to $\vec{a}(1) = -\vec{z}$ through the xz-plane. Each half is a closed loop, so each γ_s is a closed loop. At $s = 0$ it is the 4π rotation about z. At $s = 1$,

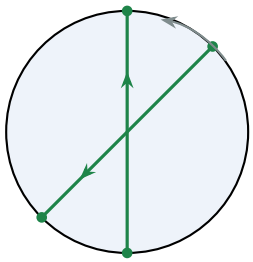
$$\begin{aligned} \gamma_1 &\overset{(1)}{=} R_{\vec{z}}(2\pi t) \text{ followed by } R_{-\vec{z}}(2\pi t) \\ &\overset{(2)}{=} R_{\vec{z}}(2\pi t) \text{ followed by } R_{\vec{z}}(-2\pi t), \end{aligned}$$

(1) Definition of γ_s at $s = 1$.
(2) Rotating by θ about $-\vec{z}$ is rotating by $-\theta$ about $\vec{z}$.

which is a loop followed by its own reverse: it retracts to the constant loop. So $2\cdot[2\pi \text{ loop}] = 0$, and $\pi_1(\mathrm{SO}(3)) = \mathbb{Z}_2$. In SU(2): the lift of the 4π loop ends at $(-1)^2 = +1$.



2π loop: a diameter with antipodal ends identified



4π loop: two diameters; rotate one onto the other

Figure 3.46: SO(3) as a solid ball of radius π: the point at vector $\theta\vec{a}$ is rotation by θ about $\vec{a}$, and antipodal points of the boundary sphere are identified (rotation by π about $\vec{a}$ equals rotation by π about $-\vec{a}$). Left: the 2π loop runs along a diameter and jumps from one end to the other; it cannot be shrunk, since its two ends must stay antipodal. Right: the 4π loop crosses twice; moving the second diameter until the crossing points meet lets the loop unhook and shrink.

Worked example 3.68: Stably, TS^2 is trivial and $H_{\mathbb{R}}$ is not

The clutching function of the real rank-2 bundle underlying $H^{\otimes d}$ is the loop $e^{id\phi} \in \mathrm{SO}(2)$, a rotation by $d\phi$ in a fixed plane. Put it into SO(3) by adding a trivial line: that is d times the 2π loop, so its class in $\pi_1(\mathrm{SO}(3)) = \mathbb{Z}_2$ is d mod 2. Hence

$$[TS^2] - 2 \overset{(1)}{=} (-2 \bmod 2) \cdot x = 0, \qquad [H_{\mathbb{R}}] - 2 \overset{(2)}{=} (1 \bmod 2) \cdot x = x \neq 0,$$

(1) TS^2 has $d = -2$ (Example 3.66); x is the generator of $\widetilde{KO}(S^2) = \mathbb{Z}_2$.
(2) H has $d = 1$.

This recovers $TS^2 \oplus I \cong I^3$ (Section 3.9) from a completely different direction.

3.12.6 Complex and quaternionic bundles

Everything so far works with $\mathbb{R}$ replaced by $\mathbb{C}$ or by the quaternions $\mathbb{H}$. Nash notes that only the Grassmannian and the stable range change.

Quaternions in one paragraph. $\mathbb{H}$ is the real algebra with basis $1, \mathbf{i}, \mathbf{j}, \mathbf{k}$ and $\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = \mathbf{ijk} = -1$ (so $\mathbf{ij} = \mathbf{k} = -\mathbf{ji}$, etc.). It is a Clifford algebra, $\mathbb{H} = \mathrm{Cl}_{0,2}$ (Chapter T4). A faithful representation by 2×2 complex matrices is

$$1 \mapsto \mathbb{1}, \quad \mathbf{i} \mapsto -i\sigma_1, \quad \mathbf{j} \mapsto -i\sigma_2, \quad \mathbf{k} \mapsto -i\sigma_3,$$

so a quaternion $q = a + b\,\mathbf{i} + c\,\mathbf{j} + d\,\mathbf{k}$ becomes $a\mathbb{1} - i(b\sigma_1 + c\sigma_2 + d\sigma_3)$. Check one relation:

$$\mathbf{ij} \overset{(1)}{=} (-i\sigma_1)(-i\sigma_2) = -\sigma_1\sigma_2 \overset{(2)}{=} -i\sigma_3 \overset{(3)}{=} \mathbf{k}.$$

(1) $(-i)^2 = -1$.
(2) $\sigma_1\sigma_2 = i\sigma_3$.
(3) The assignment $\mathbf{k} \mapsto -i\sigma_3$.

The norm is $|q|^2 = a^2 + b^2 + c^2 + d^2 = \det(\text{matrix})$, and the unit quaternions $|q| = 1$ are exactly the matrices in SU(2). Topologically they form the sphere $S^3 \subset \mathbb{R}^4$. The **compact symplectic group** $\mathrm{Sp}(k)$ is the group of $k \times k$ quaternionic matrices A with $A^\dagger A = 1$ (quaternionic conjugate transpose); $\mathrm{Sp}(1) = \mathrm{SU}(2) = S^3$.

The Grassmannians. By the same argument as (3.91),

$$\mathrm{Gr}(k, m, \mathbb{C}) = \frac{\mathrm{U}(m)}{\mathrm{U}(m-k) \times \mathrm{U}(k)}, \qquad \mathrm{Gr}(k, m, \mathbb{H}) = \frac{\mathrm{Sp}(m)}{\mathrm{Sp}(m-k) \times \mathrm{Sp}(k)}, \tag{3.104}$$

Nash's (3.66), of real dimensions $2k(m-k)$ and $4k(m-k)$. The classifying statements become

$$\begin{aligned} \mathrm{Vect}_k(M, \mathbb{C}) &\cong [M, \mathrm{Gr}(k, \infty, \mathbb{C})] = [M, BU(k)], \\ \mathrm{Vect}_k(M, \mathbb{H}) &\cong [M, \mathrm{Gr}(k, \infty, \mathbb{H})] = [M, BSp(k)], \end{aligned} \tag{3.105}$$

Nash's (3.67).

Worked example 3.69: Projective lines are spheres: $\mathbb{CP}^1 = S^2$ and $\mathbb{H}P^1 = S^4$

$\mathrm{Gr}(1,2,\mathbb{C}) = \mathbb{CP}^1$ has real dimension $2\cdot 1\cdot 1 = 2$; it is the Riemann sphere, $\mathbb{C}\cup\{\infty\}$, via $[z_0 : z_1] \mapsto z_1/z_0$. In exactly the same way $\mathrm{Gr}(1,2,\mathbb{H}) = \mathbb{H}P^1$, of real dimension 4, is $\mathbb{H}\cup\{\infty\} = S^4$, via $[q_0 : q_1] \mapsto q_1 q_0^{-1}$ (with quaternionic scalars acting on the right). The tautological quaternionic line bundle over $\mathbb{H}P^1$ has, by the computation of Example 3.36 word for word, clutching function

$$g(q) = q, \qquad q \in S^3 = \{|q| = 1\} = \mathrm{Sp}(1) = \mathrm{SU}(2),$$

the identity map $S^3 \to \mathrm{SU}(2)$, of degree 1. This is a generator of $\pi_3(\mathrm{SU}(2)) = \mathbb{Z}$.

Worked example 3.70: The instanton bundle

On $\mathbb{R}^4$ (compactified to S^4), an SU(2) gauge field of finite action approaches pure gauge at infinity, $A \to g^{-1}\mathrm{d}g$ on the large sphere S^3_∞, with $g : S^3_\infty \to \mathrm{SU}(2)$. Viewed on S^4 this g is the clutching function between the gauge near the origin and the gauge near infinity. The bundle, hence the topological sector, is fixed by $k = \deg g \in \pi_3(\mathrm{SU}(2)) = \mathbb{Z}$, the instanton number $k = \frac{1}{8\pi^2}\int \mathrm{tr}\, F\wedge F$ (up to the orientation sign; see Chapter 8). The BPST instanton lives on the bundle with $g(q) = q$ of Example 3.69. So $\widetilde{KU}(S^4) = \mathbb{Z}$ and $\widetilde{KSp}(S^4) = \mathbb{Z}$ below are the instanton number.

The stable range, complex and quaternionic. The dimension count of Section 3.11.5 changes only in that a section of a complex rank-k bundle is $2k$ real functions, and of a quaternionic one $4k$ real functions. A generic section is nowhere zero if $2k > n$ (respectively $4k > n$). Nash's theorem, with $[m]$ the integer part of m:

Key idea: Stable range, complex and quaternionic cases (Nash, p. 74)

Let $\dim M = n$, $k(c) = [(n+1)/2]$, $k(h) = [(n+1)/4]$. A complex bundle of rank $k > k(c)$ is isomorphic to $F_{k(c)} \oplus I^{k-k(c)}$; a quaternionic bundle of rank $k > k(h)$ is isomorphic to $F_{k(h)} \oplus I^{k-k(h)}$. In particular it is enough that $k > n/2$ (complex) or $k > n/4$ (quaternionic). Then

$$\mathrm{Vect}_k(M,\mathbb{C}) \cong \widetilde{KU}(M) \ \ (k > \tfrac{n}{2}), \qquad \mathrm{Vect}_k(M,\mathbb{H}) \cong \widetilde{KSp}(M) \ \ (k > \tfrac{n}{4}). \tag{3.106}$$

We write $\widetilde{KO}$, $\widetilde{KU}$, $\widetilde{KSp}$ for the real, complex and quaternionic theories. (When no letter is given, "K" usually means KU.) Clutching gives, exactly as in (3.101),

$$\widetilde{KU}(S^n) \cong \pi_{n-1}(\mathrm{U}(k)) \ \ (k > \tfrac{n}{2}), \qquad \widetilde{KSp}(S^n) \cong \pi_{n-1}(\mathrm{Sp}(k)) \ \ (k > \tfrac{n}{4}), \tag{3.107}$$

Nash's (3.69), and the stability ranges (3.102) confirm that these k are large enough.

Watch out: Nash's bound is sufficient, not sharp

The condition $k > k(c)$ is sometimes stronger than needed. For example, on S^2 the complex line bundles ($k = 1$, not $> n/2 = 1$) already map bijectively to $\widetilde{KU}(S^2)$: $H^{\otimes d} \mapsto [H^{\otimes d}] - 1 = d\,([H]-1)$ (see (3.110) below). For isomorphism one needs a nonvanishing section over $M\times[0,1]$ for the next rank, $2(k+1) > n+1$, and for the splitting only $2(k+1) > n$; for $n = 2$, $k = 1$ both hold ($4 > 3$). Nash's rounded bounds are always safe.

The complex and quaternionic tables

Inserting the homotopy groups (Nash's (3.70)):

$\pi_{n-1}(\mathrm{U}(k)),\ \pi_{n-1}(\mathrm{Sp}(k))$	$\widetilde{KU}(S^n)$	$\widetilde{KSp}(S^n)$
$n=0$	$\mathbb{Z}$	$\mathbb{Z}$
$n=1:\ \pi_0(\mathrm{U})=0,\ \pi_0(\mathrm{Sp})=0$	0	0
$n=2:\ \pi_1(\mathrm{U})=\mathbb{Z},\ \pi_1(\mathrm{Sp})=0$	$\mathbb{Z}$	0
$n=3:\ \pi_2(\mathrm{U})=0,\ \pi_2(\mathrm{Sp})=0$	0	0
$n=4:\ \pi_3(\mathrm{U})=\mathbb{Z},\ \pi_3(\mathrm{Sp})=\mathbb{Z}$	$\mathbb{Z}$	$\mathbb{Z}$

(3.108)

Why each entry:

- π_0: $\mathrm{U}(k)$ and $\mathrm{Sp}(k)$ are *connected* (every unitary matrix is e^{iX} with X Hermitian; shrink X to 0). So there is no Möbius-type bundle over S^1 in the complex world (Example 3.51).
- $\pi_1(\mathrm{U}(k))=\mathbb{Z}$: the determinant $\det:\mathrm{U}(k)\to\mathrm{U}(1)$ has simply connected fibre $\mathrm{SU}(k)$, so loops in $\mathrm{U}(k)$ are classified by the winding of their determinant. $\pi_1(\mathrm{Sp}(k))=0$ because $\mathrm{Sp}(1)=S^3$ is simply connected and the group is stable already.
- $\pi_2=0$: Cartan's theorem again.
- $\pi_3=\mathbb{Z}$: for both, generated by $\mathrm{SU}(2)=\mathrm{Sp}(1)=S^3\subset\mathrm{U}(k)$ or $\mathrm{Sp}(k)$, and measured by the instanton number (second Chern class).

Nash points out: *the* $\widetilde{KU}(S^n)$ *have period* 2 *in* n, $\mathbb{Z},0,\mathbb{Z},0,\mathbb{Z},\ldots$, and this persists for all n. That is Bott periodicity, the subject of the next section.

	S^0	S^1	S^2	S^3	S^4	S^5	S^6	S^7	S^8
$\widetilde{KO}$	$\mathbb{Z}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	0	$\mathbb{Z}$	0	0	0	$\mathbb{Z}$
$\widetilde{KU}$	$\mathbb{Z}$	0	$\mathbb{Z}$	0	$\mathbb{Z}$	0	$\mathbb{Z}$	0	$\mathbb{Z}$
$\widetilde{KSp}$	$\mathbb{Z}$	0	0	0	$\mathbb{Z}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	0	$\mathbb{Z}$

Nash (3.65), (3.70) — beyond Nash's table (Bott periodicity)

Figure 3.47: The reduced K-groups of spheres. The columns left of the dashed line are Nash's tables, computed from homotopy groups of $\mathrm{O}(k)$, $\mathrm{U}(k)$ and $\mathrm{Sp}(k)$. To the right, the values that Bott periodicity predicts: $\widetilde{KU}$ repeats with period 2 (blue), $\widetilde{KO}$ with period 8 (orange), and $\widetilde{KSp}$ is $\widetilde{KO}$ shifted by 4 (green).

3.12.7 The ring $K(S^2)$ from clutching: $(H-1)^2=0$

We can now prove the relation $H\oplus H\cong H^{\otimes 2}\oplus I$ promised in Example 3.55, and with it the full ring structure of $K(S^2)$. This is the one computation on which complex Bott periodicity rests.

Rank-k complex bundles over S^2. By (3.98) with $\mathrm{U}(k)$, they correspond to $\pi_1(\mathrm{U}(k))=\mathbb{Z}$, detected by the winding number of $\det g$. So two clutching functions $g,g':S^1\to\mathrm{U}(k)$ give isomorphic bundles if and only if $\det g$ and $\det g'$ have the same winding number. In particular, every rank-k bundle is $H^{\otimes d}\oplus I^{k-1}$ for a unique $d\in\mathbb{Z}$ (clutching $\mathrm{diag}(z^d,1,\ldots,1)$, determinant z^d).

The key isomorphism. Compare

$$H \oplus H:\ g = \begin{pmatrix} z & 0 \\ 0 & z \end{pmatrix}, \qquad H^{\otimes 2} \oplus I:\ g' = \begin{pmatrix} z^2 & 0 \\ 0 & 1 \end{pmatrix} \qquad (z = e^{i\phi}).$$

Both determinants are z^2, so the bundles are isomorphic. It is instructive to see the homotopy explicitly. Let $R(t) = \begin{pmatrix} \cos\frac{\pi t}{2} & -\sin\frac{\pi t}{2} \\ \sin\frac{\pi t}{2} & \cos\frac{\pi t}{2} \end{pmatrix}$ and

$$g_t(z) = \begin{pmatrix} z & 0 \\ 0 & 1 \end{pmatrix} R(t) \begin{pmatrix} z & 0 \\ 0 & 1 \end{pmatrix} R(t)^{-1}, \qquad t \in [0,1].$$

Each $g_t(z)$ is a product of unitary matrices, hence unitary, and depends continuously on (t, z). At the two ends:

$$\begin{aligned} g_0(z) &\overset{(1)}{=} \begin{pmatrix} z & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} z & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} z^2 & 0 \\ 0 & 1 \end{pmatrix} = g'(z), \\ g_1(z) &\overset{(2)}{=} \begin{pmatrix} z & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} z & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \\ &\overset{(3)}{=} \begin{pmatrix} z & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & z \end{pmatrix} = \begin{pmatrix} z & 0 \\ 0 & z \end{pmatrix} = g(z). \end{aligned}$$

(1) $R(0) = 1$.
(2) $R(1)$ is the rotation by $90°$; its inverse is its transpose.
(3) Conjugating a diagonal matrix by the $90°$ rotation swaps its diagonal entries:

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} z & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ z & 0 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & z \end{pmatrix}.$$

So $g \simeq g'$ and

$$H \oplus H \ \cong\ H^{\otimes 2} \oplus I. \tag{3.109}$$

Key idea: The ring $K(S^2)$

In $K(S^2)$ write H for $[H]$. Then (3.109) says $2H = H^2 + 1$, i.e.

$$(H-1)^2 = 0, \qquad K(S^2) = \mathbb{Z}[H]/(H-1)^2, \qquad \widetilde{KU}(S^2) = \mathbb{Z}\cdot b,\ \ b = H-1,\ \ b^2 = 0.$$

Let us derive H^n for every integer n from this. With $H = 1 + b$ and $b^2 = 0$, for $n \geq 0$:

$$\begin{aligned} H^n &\overset{(1)}{=} (1+b)^n \\ &\overset{(2)}{=} 1 + nb + \tbinom{n}{2} b^2 + \cdots \\ &\overset{(3)}{=} 1 + n\,b = 1 + n(H-1). \end{aligned} \tag{3.110}$$

(1) $H = 1 + b$ by definition of b.
(2) Binomial theorem (the ring is commutative).
(3) $b^2 = 0$ kills every term beyond the linear one.

For negative powers, $(1+b)(1-b) = 1 - b^2 = 1$, so $H^{-1} = [H^*] = 1 - b = 2 - H$, agreeing with $[H] + [H^*] = 2$ from Example 3.50. Then $H^{-n} = (1-b)^n = 1 - nb$, so (3.110) holds for all $n \in \mathbb{Z}$. Consequences:

- The line bundle $H^{\otimes d}$ (monopole charge d, up to the sign convention) has K-class $1+d\,b$: *the reduced K-class of a line bundle over S^2 is its monopole charge times the generator*. In particular $\widetilde{KU}(S^2)=\mathbb{Z}$ is detected by c_1.
- $[TS^2]=H^{-2}=1-2b$ in $KU(S^2)$ (complex structure), so TS^2 is *not* stably trivial as a complex bundle, while its underlying real bundle is (Example 3.68).
- Tensor products are computed by multiplying out. For example $[H^{\otimes 2}\oplus I]\cdot[H^*]=(H^2+1)H^{-1}=H+H^{-1}=(1+b)+(1-b)=2$, and indeed $(H^{\otimes 2}\oplus I)\otimes H^*\cong H\oplus H^*\cong I^2$ (the last step by Example 3.44).

Physical picture: Why $b^2=0$ is "obvious" in cohomology

The Chern character (the subject of Nash's next section, §5) sends $H\mapsto e^{c_1(H)}=1+c_1(H)+\frac{1}{2}c_1(H)^2+\cdots$. On S^2 there are no 4-forms, so $c_1^2=0$ and $\mathrm{ch}(H)=1+c_1(H)$, $\mathrm{ch}(b)=c_1(H)$, $\mathrm{ch}(b)^2=0$. The K-theory relation $b^2=0$ is the shadow of "no forms of degree higher than $\dim S^2$". What is remarkable is that it holds in K-theory itself, with integer coefficients, where it is proved by the explicit homotopy above.

Worked example 3.71: A stably trivial but non-trivial complex bundle

Stability can fail below the stable range even for complex bundles. Take S^5 and rank 2 (note $2<5/2$). Rank-2 bundles correspond to $\pi_4(\mathrm{U}(2))$. Topologically $\mathrm{U}(2)\cong\mathrm{U}(1)\times\mathrm{SU}(2)=S^1\times S^3$, so $\pi_4(\mathrm{U}(2))=\pi_4(S^3)=\mathbb{Z}_2$: there is exactly one non-trivial rank-2 complex bundle E over S^5. But $\pi_4(\mathrm{U}(k))$ for $k\geq 3$ is the stable group $\pi_4(\mathrm{U})=\widetilde{KU}(S^5)=0$ (from Bott periodicity: $\widetilde{KU}$ of an odd sphere vanishes). So $E\oplus I\cong I^3$: E is invisible to K-theory.

Reading Nash (pp. 72–75, (3.63)–(3.70))

Nash sets $M=S^n$ in (3.62) to get (3.63), mentions the free/based subtlety, and then describes clutching in one paragraph: cover S^n by two hemispheres, the bundle is trivial on each, and the transition function on the overlap gives an element of $\pi_{n-1}(\mathrm{O}(k))$. He digresses on $[S^n,BG]=[S^{n-1},G]$ and $\Omega BG\simeq G$, arrives at (3.64), quotes the stability of $\pi_{n-1}(\mathrm{O}(k))$ and the table (3.65) from Husemoller, then repeats everything for $\mathbb{C}$ and $\mathbb{H}$: Grassmannians (3.66), classification (3.67), stable range theorem with $k(c)$, $k(h)$, (3.68), (3.69) and the table (3.70), remarking on the period 2 of $\widetilde{KU}(S^n)$. We have supplied: why bundles over discs are trivial, why only the homotopy class of the clutching function matters, the winding integral, the adjunction and the holonomy picture of $\Omega BG\simeq G$, the fibration argument for stability, a line-by-line justification of both tables, the explicit $\pi_1(\mathrm{SO}(3))=\mathbb{Z}_2$ computation, quaternions and $\mathbb{H}P^1=S^4$, the instanton interpretation, and the explicit proof of $(H-1)^2=0$.

Recap

- A bundle over S^n is trivial on each hemisphere and is fixed by the clutching function $g:S^{n-1}\to G$ up to homotopy: $\mathrm{Vect}_k(S^n)=\pi_{n-1}(G)$.
- $[S^n,BG]=[S^{n-1},G]$ because $\Omega BG\simeq G$ (holonomy).
- $\widetilde{K}(S^n)=\pi_{n-1}(\mathrm{O}(k))$, $\pi_{n-1}(\mathrm{U}(k))$, $\pi_{n-1}(\mathrm{Sp}(k))$ in the stable range; the groups stabilise by the fibrations $G(k)\to G(k+1)\to S^\bullet$.
- Tables: $\widetilde{KO}(S^n)=\mathbb{Z},\mathbb{Z}_2,\mathbb{Z}_2,0,\mathbb{Z}$; $\widetilde{KU}(S^n)=\mathbb{Z},0,\mathbb{Z},0,\mathbb{Z}$; $\widetilde{KSp}(S^n)=\mathbb{Z},0,0,0,\mathbb{Z}$ for

$n = 0, \dots, 4$.

- $H \oplus H \cong H^2 \oplus I$ by an explicit homotopy of clutching functions, so $K(S^2) = \mathbb{Z}[H]/(H-1)^2$ and $H^n = 1 + n(H-1)$.

3.13 Bott periodicity

The tables of the previous section showed a pattern: $\widetilde{KU}(S^n) = \mathbb{Z}, 0, \mathbb{Z}, 0, \mathbb{Z}$ for $n = 0, \dots, 4$. Bott's theorem says the pattern never stops. It is one of the deepest facts in topology, and it is the reason K-theory is so computable: once you know K^0 and K^{-1} you know everything.

Nash first states it as a fact about homotopy groups of the unitary group (the original form, due to Bott, 1959), and then reformulates it as a statement internal to K-theory: multiplication by the Hopf class $b = [H] - 1$ is an isomorphism $\widetilde{KU}^{-n}(X) \to \widetilde{KU}^{-n-2}(X)$. We follow the same route, filling in the definitions of suspension, wedge and smash product on the way, and adding a physicist's sketch of *why* it is true. As in Nash, we treat the complex case first and leave the real and quaternionic cases to Section 3.14.

3.13.1 Periodicity of the homotopy of $\mathrm{U}(\infty)$

Embed each unitary group in the next as the upper-left block,

$$\mathrm{U}(k) \hookrightarrow \mathrm{U}(k+1), \qquad A \longmapsto \begin{pmatrix} A & 0 \\ 0 & 1 \end{pmatrix},$$

and let $\mathrm{U} = \mathrm{U}(\infty) = \bigcup_k \mathrm{U}(k)$, the **infinite unitary group**: unitary matrices of any size that differ from the identity in finitely many entries. By the stability (3.102), $\pi_j(\mathrm{U}(k))$ does not change once $k > j/2$, and that common value is $\pi_j(\mathrm{U})$. Letting $k \to \infty$ in (3.107):

$$\widetilde{KU}(S^n) = \pi_{n-1}(\mathrm{U}(\infty)). \tag{3.111}$$

This is Nash's (3.71).

Key idea: Bott periodicity, homotopy form (Nash (3.72))

$$\pi_n(\mathrm{U}(\infty)) = \pi_{n+2}(\mathrm{U}(\infty)) \qquad \text{for all } n \geq 0, \qquad \text{i.e.} \qquad \pi_n(\mathrm{U}) = \begin{cases} 0, & n \text{ even}, \\ \mathbb{Z}, & n \text{ odd}. \end{cases}$$

Combined with (3.111) this gives Nash's (3.73) immediately:

$$\begin{aligned} \widetilde{KU}(S^n) &\overset{(1)}{=} \pi_{n-1}(\mathrm{U}) \\ &\overset{(2)}{=} \pi_{n+1}(\mathrm{U}) \\ &\overset{(1)}{=} \widetilde{KU}(S^{n+2}). \end{aligned} \tag{3.112}$$

(1) Clutching and stability, (3.111) (with n and $n+2$).
(2) Bott periodicity with $n \to n-1$.

The periodicity is a *stable* phenomenon. For small k the groups $\pi_j(\mathrm{U}(k))$ are wild: $\pi_4(\mathrm{U}(2)) = \mathbb{Z}_2$ (Example 3.71), $\pi_5(\mathrm{U}(2)) = \pi_5(S^3) = \mathbb{Z}_2$, $\pi_6(\mathrm{U}(2)) = \pi_6(S^3) = \mathbb{Z}_{12}$, and so on, reflecting the notoriously complicated homotopy groups of spheres. Only after enough stabilisation does the simple pattern emerge (Figure 3.48).

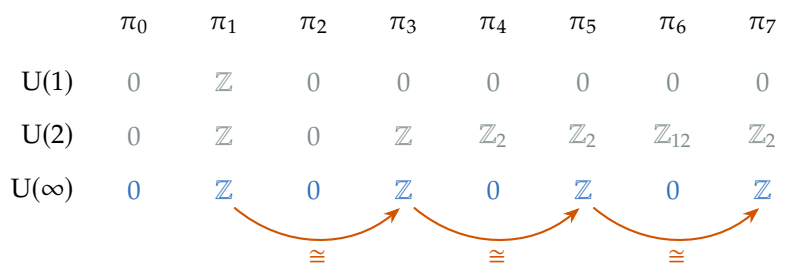


Figure 3.48: Homotopy groups of unitary groups. The small groups U(1) and U(2) (grey) have irregular homotopy, inherited from spheres ($\mathrm{U}(1) = S^1$; $\mathrm{U}(2) \cong S^1 \times S^3$). The infinite unitary group (blue) has the simple pattern $0, \mathbb{Z}, 0, \mathbb{Z}, \ldots$ with period two (orange arrows). This regularity is Bott periodicity.

Worked example 3.72: The odd homotopy of U as winding numbers

The $\mathbb{Z}$'s in odd degree are detected by integrals of the Maurer–Cartan form $g^{-1}\mathrm{d}g$, the same object as a pure-gauge field. For $g : S^1 \to \mathrm{U}(k)$,

$$w_1(g) = \frac{1}{2\pi i} \oint_{S^1} \operatorname{tr}\left(g^{-1}\mathrm{d}g\right) \in \mathbb{Z},$$

which equals the winding of $\det g$ (since $\operatorname{tr}(g^{-1}\mathrm{d}g) = \mathrm{d}\log\det g$). For $g : S^3 \to \mathrm{U}(k)$,

$$w_3(g) = \frac{1}{24\pi^2} \int_{S^3} \operatorname{tr}\left(g^{-1}\mathrm{d}g\right)^3 \in \mathbb{Z}$$

(up to an orientation sign). Let us check the normalisation on the generator $g(q) = q$ of Example 3.69, $S^3 = \mathrm{SU}(2)$. The integrand is left-invariant ($g \mapsto hg$ leaves $g^{-1}\mathrm{d}g$ unchanged), so it is a constant multiple of the volume form, and we evaluate it at $g = 1$. There, with $q = x^0 - i\,x^a\sigma_a$ near $x = (1,0,0,0)$, we have $g^{-1}\mathrm{d}g = -i\sigma_a\,\mathrm{d}x^a$, and

$$\begin{aligned}\operatorname{tr}\left(g^{-1}\mathrm{d}g\right)^3 &\overset{(1)}{=} (-i)^3\,\operatorname{tr}(\sigma_a\sigma_b\sigma_c)\,\mathrm{d}x^a\wedge\mathrm{d}x^b\wedge\mathrm{d}x^c\\ &\overset{(2)}{=} i\cdot 2i\,\varepsilon_{abc}\,\mathrm{d}x^a\wedge\mathrm{d}x^b\wedge\mathrm{d}x^c\\ &\overset{(3)}{=} -2\cdot 3!\,\mathrm{d}x^1\wedge\mathrm{d}x^2\wedge\mathrm{d}x^3 = -12\,\mathrm{vol}.\end{aligned}$$

(1) Multiply out; the $\mathrm{d}x$'s anticommute, the σ's do not.
(2) $(-i)^3 = i$, and only the totally antisymmetric part of $\operatorname{tr}(\sigma_a\sigma_b\sigma_c) = 2i\varepsilon_{abc} +$ (symmetric in two indices) survives against the wedge product.
(3) $\varepsilon_{abc}\,\mathrm{d}x^a\wedge\mathrm{d}x^b\wedge\mathrm{d}x^c = 3!\,\mathrm{d}x^1\wedge\mathrm{d}x^2\wedge\mathrm{d}x^3$, which at $x = (1,0,0,0)$ is the volume form of the unit S^3.

Since $\mathrm{Vol}(S^3) = 2\pi^2$, $w_3 = \frac{-12\cdot 2\pi^2}{24\pi^2} = -1$: a unit winding, with a sign fixed by orientation. This is exactly the instanton number of the BPST instanton (Example 3.70), computed from its behaviour at infinity. The next one, w_5 for $\pi_5(\mathrm{U}) = \mathbb{Z}$, governs the non-abelian anomaly in four dimensions (Chapter 10).

3.13.2 Loop spaces: $\Omega^2\mathrm{U} \simeq \mathrm{U}$

Why should π_n and π_{n+2} agree? Nash recalls, from Chapter 1 and the adjunction (3.99),

$$\pi_n(X) = \pi_{n-1}(\Omega X) \qquad\Longrightarrow\qquad \pi_n(X) = \pi_{n-2}(\Omega^2 X). \tag{3.113}$$

With $X = \mathrm{U}$ this gives $\pi_{n+2}(\mathrm{U}) = \pi_n(\Omega^2\mathrm{U})$. So periodicity would follow if the double loop space of U looked like U itself. That is what Bott proved:

$$\Omega^2\mathrm{U} \;\simeq\; \mathrm{U},$$

more precisely in two steps,

$$\Omega\mathrm{U} \;\simeq\; \mathbb{Z}\times BU, \qquad \Omega(\mathbb{Z}\times BU) = \Omega BU \;\simeq\; \mathrm{U}. \tag{3.114}$$

The second step is $\Omega BG \simeq G$ from Section 3.12.3 (loops based at a point only see one component, so the $\mathbb{Z}$ drops out). The first step is the real content. Here is Bott's idea, which uses Morse theory (Chapter 7) on the loop space.

Physical picture: Bott's proof: minimal geodesics from 1 to -1

Think of $\Omega\mathrm{SU}(2m)$ as the space of paths from 1 to -1 in $\mathrm{SU}(2m)$ (a path space is homotopy equivalent to a loop space). Give each path its energy $\int |\dot\gamma|^2\,\mathrm{d}t$. Critical points are geodesics $\gamma(t) = \exp(i\pi t A)$ with A Hermitian and $\exp(i\pi A) = -1$, so the eigenvalues of A are odd integers, summing to zero (for $\det = 1$). The energy is proportional to $\mathrm{tr}\,A^2$, so the *minimal* geodesics have all eigenvalues ±1, m of each:

$$A = 2\Pi - 1, \qquad \Pi \text{ a projector of rank } m \text{ on } \mathbb{C}^{2m}.$$

So the space of minimal geodesics is exactly the Grassmannian $\mathrm{Gr}(m, 2m, \mathbb{C})$. Morse theory says the loop space is built from the minimum by attaching cells of dimension equal to the Morse indices of the other geodesics, and Bott computed that these indices grow like $2m$. As $m \to \infty$ the non-minimal geodesics are pushed to infinite dimension and become invisible to homotopy groups of any fixed degree:

$$\Omega\mathrm{SU} \;\simeq\; \mathrm{Gr}(\infty, \infty, \mathbb{C}) = BU.$$

For U instead of SU the paths also have a winding number of the determinant, giving the extra $\mathbb{Z}$ in $\Omega\mathrm{U} \simeq \mathbb{Z}\times BU$.

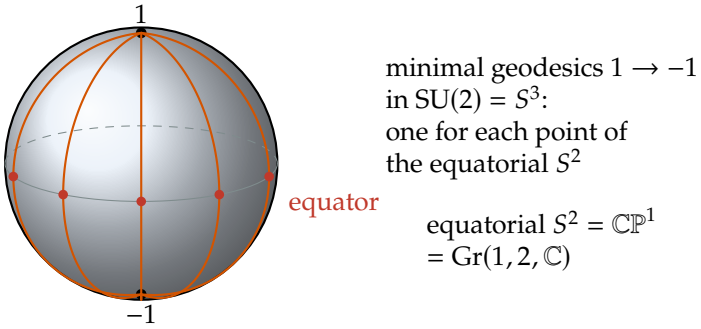


Figure 3.49: A cartoon of Bott's proof in the smallest case, drawn one dimension down. In $\mathrm{SU}(2) = S^3$ the shortest paths from 1 (north pole) to -1 (south pole) are the great semicircles (orange). Each crosses the equatorial 2-sphere at one point (red), so the space of minimal geodesics is $S^2 = \mathbb{CP}^1 = \mathrm{Gr}(1,2,\mathbb{C})$. In $\mathrm{SU}(2m)$ the same count gives $\mathrm{Gr}(m,2m,\mathbb{C})$, and as $m \to \infty$ this approximates the whole loop space: $\Omega\mathrm{SU} \simeq BU$.

Worked example 3.73: Checking $\Omega\mathrm{U} \simeq \mathbb{Z} \times BU$ on low homotopy groups

Take π_j of both sides, using $\pi_j(\Omega\mathrm{U}) = \pi_{j+1}(\mathrm{U})$ and $\pi_j(BU) = \pi_{j-1}(\mathrm{U})$ (the latter from $\Omega BU \simeq \mathrm{U}$):

$$\begin{aligned} \pi_0(\Omega\mathrm{U}) &\overset{(1)}{=} \pi_1(\mathrm{U}) = \mathbb{Z} && \text{vs} \quad \pi_0(\mathbb{Z} \times BU) = \mathbb{Z}, \\ \pi_1(\Omega\mathrm{U}) &\overset{(1)}{=} \pi_2(\mathrm{U}) = 0 && \text{vs} \quad \pi_1(BU) \overset{(2)}{=} \pi_0(\mathrm{U}) = 0, \\ \pi_2(\Omega\mathrm{U}) &\overset{(1)}{=} \pi_3(\mathrm{U}) = \mathbb{Z} && \text{vs} \quad \pi_2(BU) \overset{(2)}{=} \pi_1(\mathrm{U}) = \mathbb{Z}. \end{aligned}$$

(1) (3.113).
(2) $\Omega BU \simeq \mathrm{U}$.

The two sides agree, as they must. Note that $\pi_2(BU) = \mathbb{Z}$ is the monopole charge of Example 3.63 and $\pi_2(\Omega\mathrm{U}) = \pi_3(\mathrm{U}) = \mathbb{Z}$ is the instanton winding: Bott periodicity identifies them.

3.13.3 Suspension and the groups K^{-n}

To state periodicity for any space X, not just spheres, Nash defines the negative-degree K-groups using suspension.

Recall the **reduced suspension** SX (Section 3.12.3): $X \times [0,1]$ with $X \times \{0\}$, $X \times \{1\}$ and $\{x_0\} \times [0,1]$ collapsed to one point. Picture it as a double cone on X, pinched along a line (Figure 3.50). Iterate: $S^n X = S(S^{n-1}X)$. The key example is $SS^m \cong S^{m+1}$: the double cone on a circle is a sphere.

Working definition 3.32 (negative K-groups)

$$\widetilde{KU}^{-n}(X) := \widetilde{KU}(S^n X), \qquad n = 0, 1, 2, \dots \tag{3.115}$$

[$\widetilde{KU}$ is reduced K-theory, the virtual bundles of virtual dimension 0, and $S^n X$ is the n-fold suspension] (Nash's (3.75)). For the unreduced groups use $X_+ = X \sqcup \{\mathrm{pt}\}$ (add a disjoint base point) and set $KU^{-n}(X) = \widetilde{KU}^{-n}(X_+)$; for $n = 0$ this gives back $KU(X)$, since $\widetilde{K}(X_+) = K(X \sqcup \mathrm{pt})/K(\mathrm{pt}) = K(X)$ [$K(X \sqcup \mathrm{pt}) = K(X) \oplus \mathbb{Z}$, the $\mathbb{Z}$ recording the rank on the extra point].

For spheres, repeatedly using $SS^m = S^{m+1}$:

$$\widetilde{KU}^{-n}(S^m) \overset{(1)}{=} \widetilde{KU}(S^n S^m) \overset{(2)}{=} \widetilde{KU}(S^{m+n}), \tag{3.116}$$

Nash's (3.76), where (1) is the definition (3.115) and (2) is $SS^m \cong S^{m+1}$ applied n times.

The K-theory of a point in every degree. Take $X = \mathrm{pt}$, so $X_+ = S^0$ and $S^n S^0 = S^n$:

$$KU^{-n}(\mathrm{pt}) \overset{(1)}{=} \widetilde{KU}(S^n S^0) \overset{(2)}{=} \widetilde{KU}(S^n) \overset{(3)}{=} \begin{cases} \mathbb{Z}, & n \text{ even}, \\ 0, & n \text{ odd}. \end{cases}$$

(1) Definition, with $\mathrm{pt}_+ = S^0$.
(2) $S^n S^0 \cong S^n$.
(3) Bott periodicity and the table (3.108).

This is the failure of the dimension axiom announced in Section 3.8.3 and drawn in Figure 3.33.

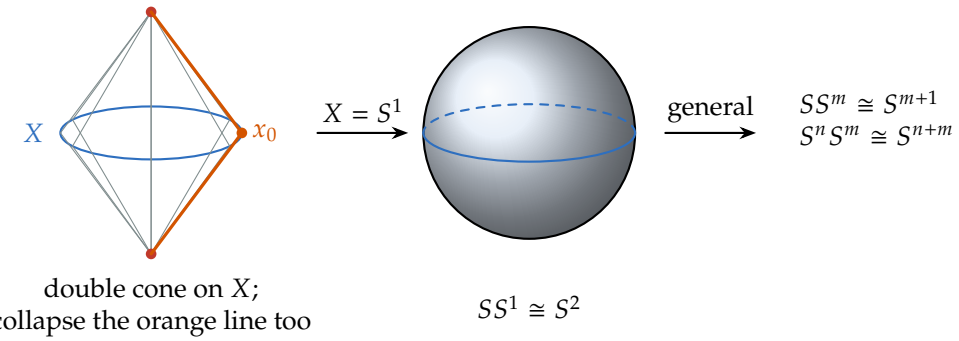


Figure 3.50: The reduced suspension. Stack copies of X over the interval $[0,1]$, shrink the top and bottom copies to points (red), and also collapse the segment over the base point x_0 (orange). For X a circle the result is a sphere: each copy of the circle becomes a line of latitude. Collapsing the orange segment in addition changes nothing up to homotopy for nice spaces, but makes the construction compatible with base points.

K^{-1} is maps into U. For a based space X the classifying space picture gives a neat description:

$$\begin{aligned}\widetilde{KU}^{-1}(X) &\overset{(1)}{=} \widetilde{KU}(SX)\\ &\overset{(2)}{=} [SX, BU]\\ &\overset{(3)}{=} [X, \Omega BU]\\ &\overset{(4)}{=} [X, \mathrm{U}].\end{aligned} \tag{3.117}$$

(1) Definition (3.115).
(2) Stable classes are classified by maps to $BU = \mathrm{Gr}(\infty, \infty, \mathbb{C})$ (the stable range statement (3.106) with $k \to \infty$; SX is connected).
(3) Adjunction (3.99).
(4) $\Omega BU \simeq \mathrm{U}$.

Similarly $KU(X) = [X, \mathbb{Z} \times BU]$, the $\mathbb{Z}$ recording the virtual rank. So the two basic K-groups are homotopy classes of maps into the two spaces $\mathbb{Z} \times BU$ and U, and by (3.114) each is the loop space of the other. An element of $K^{-1}(X)$ is concretely a family of unitary matrices $g(x)$, i.e. a clutching function; for $X = S^1$ it is a loop of unitaries, and $\widetilde{KU}^{-1}(S^1) = \widetilde{KU}(S^2) = \mathbb{Z}$ counts its winding.

Dictionary: K^0 and K^{-1} in physics

$KU^0(X) = [X, \mathbb{Z} \times BU]$	families of subspaces/projectors: filled bands of a gapped Hamiltonian, Chern insulators; D-brane charges in type IIB
$KU^{-1}(X) = [X, \mathrm{U}]$	families of unitaries: gauge transformations, Floquet operators, chiral-symmetric Hamiltonians; D-brane charges in type IIA
Bott map $K^{-n} \to K^{-n-2}$	dimensional shift of topological phases by 2 (class A: Chern number in $d = 2$, second Chern number in $d = 4$)

3.13.4 Products: internal and external

The K-theoretic form of Bott periodicity uses a product. Nash distinguishes two.

Internal product. The ring structure of Section 3.10.7, induced by $\otimes$:

$$\beta : KU(X) \otimes KU(X) \to KU(X), \qquad a \otimes b \mapsto a \bullet b.$$

Nash's footnote 5 explains the notation: one usually writes a product on a set S as a map $S \times S \to S$, but because $KU(X)$ is an abelian group and the product is bilinear, it is a linear map on the tensor product of groups $KU(X) \otimes_{\mathbb{Z}} KU(X)$. (Here $\otimes$ of abelian groups: $\mathbb{Z} \otimes A = A$, $\mathbb{Z}^a \otimes \mathbb{Z}^b = \mathbb{Z}^{ab}$, $\mathbb{Z}_2 \otimes \mathbb{Z} = \mathbb{Z}_2$.)

External product. This involves two spaces. Let $\pi_X : X \times Y \to X$ and $\pi_Y : X \times Y \to Y$ be the projections, which induce pullbacks

$$\pi_X^* : KU(X) \to KU(X \times Y), \qquad \pi_Y^* : KU(Y) \to KU(X \times Y) \tag{3.118}$$

(Nash's (3.78)). Then (Nash's (3.77), (3.79))

$$\alpha : KU(X) \otimes KU(Y) \to KU(X \times Y), \qquad c \otimes d \longmapsto \pi_X^*(c) \bullet \pi_Y^*(d). \tag{3.119}$$

Working definition 3.33 (external tensor product of bundles)

For bundles $E \to X$ and $F \to Y$, the **external tensor product** $E \boxtimes F := \pi_X^* E \otimes \pi_Y^* F$ [$\pi_X^* E$ has fibre E_x at (x, y), whatever y is] is the bundle over $X \times Y$ with fibre $E_x \otimes F_y$ at (x, y). Its transition functions are the Kronecker products $g^E_{ab}(x) \otimes g^F_{cd}(y)$ on the product patches $U_a \times V_c$. On K-classes, $\alpha([E] \otimes [F]) = [E \boxtimes F]$.

Worked example 3.74: $K(S^2 \times S^2)$ and the Bott class

Take $X = Y = S^2$, and let $b_1 = \pi_1^* b$, $b_2 = \pi_2^* b$ be the Hopf class pulled back from each factor. Then $b_1^2 = \pi_1^*(b^2) = 0$, $b_2^2 = 0$, and the periodicity theorem below says α is an isomorphism, so

$$K(S^2 \times S^2) \cong K(S^2) \otimes K(S^2) = \mathbb{Z}[b_1, b_2]/(b_1^2, b_2^2),$$

free abelian of rank 4 with basis $1, b_1, b_2, b_1 b_2$. (Compare $H^\bullet(S^2 \times S^2)$: also rank 4, in degrees $0, 2, 2, 4$.) The line bundle $H^{\otimes p} \boxtimes H^{\otimes q}$ has class

$$\begin{aligned} [H^{\otimes p} \boxtimes H^{\otimes q}] &\overset{(1)}{=} \pi_1^*[H^{\otimes p}] \bullet \pi_2^*[H^{\otimes q}] \\ &\overset{(2)}{=} (1 + p\, b_1)(1 + q\, b_2) \\ &= 1 + p\, b_1 + q\, b_2 + pq\, b_1 b_2. \end{aligned}$$

(1) Definition of $\boxtimes$ and (3.119).
(2) $[H^{\otimes n}] = 1 + nb$, (3.110), pulled back.

The top class $b_1 b_2 = b \boxtimes b$ restricts to zero on $S^2 \vee S^2$ and will turn out to be the generator of $\widetilde{KU}(S^4)$: the **Bott class**. Under the Chern character it maps to $c_1 \boxtimes c_1$, which integrates to ± 1 over $S^2 \times S^2$: it is an "instanton number one" class.

3.13.5 Wedges, smashes, and the reduced external product

Nash states the key fact first: *for $Y = S^2$, α is an isomorphism* (his (3.80)),

$$KU(X) \otimes KU(S^2) \cong KU(X \times S^2). \tag{3.120}$$

To connect this with periodicity of $\widetilde{KU}(S^n)$ he rewrites both sides in reduced K-theory. That needs two constructions.

Working definition 3.34 (wedge and smash)

For based spaces [spaces with one chosen point] (X, x_0) and (Y, y_0):

- the **wedge** $X \vee Y$ is the disjoint union with the two base points identified; inside $X \times Y$ it is $X \times \{y_0\} \cup \{x_0\} \times Y$ (the "axes");
- the **smash** $X \wedge Y = (X \times Y)/(X \vee Y)$ collapses the axes to a single point.

Worked example 3.75: $S^1 \wedge S^1 = S^2$

$S^1 \times S^1$ is the torus, a square with opposite sides identified. $S^1 \vee S^1$ is the figure-eight formed by the two circles $\theta_1 = 0$ and $\theta_2 = 0$, which in the square are the boundary edges. Collapsing the whole boundary of a square to one point gives a sphere (Figure 3.51). In general $S^n \wedge S^m \cong S^{n+m}$: S^n is $\mathbb{R}^n$ plus a point at infinity, $S^n \times S^m$ minus the axes is $\mathbb{R}^n \times \mathbb{R}^m = \mathbb{R}^{n+m}$, and collapsing the axes adds back one point at infinity. Also $SX \cong S^1 \wedge X$, so $S^n X \cong S^n \wedge X$.

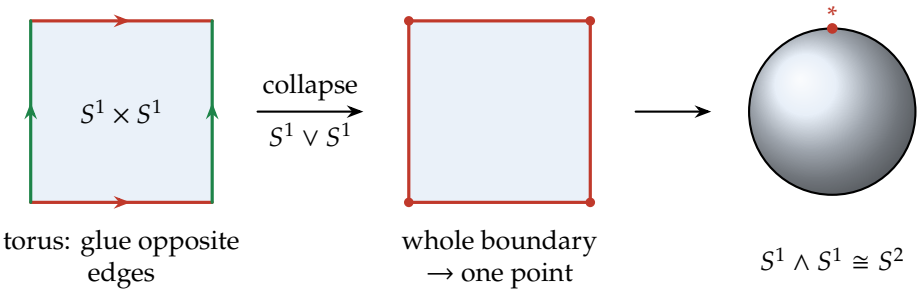


Figure 3.51: The smash product of two circles. The torus is a square with opposite edges glued (red to red, green to green). The two glued edge circles form the wedge $S^1 \vee S^1$. Collapsing them means collapsing the entire boundary of the square to one point, which turns the open square ($\cong \mathbb{R}^2$) into a sphere with the collapsed boundary as its north pole.

Reducing the left side of α. Write $KU(X) = \widetilde{KU}(X) \oplus \mathbb{Z}$ ((3.89), with the $\mathbb{Z}$ coming from the base point x_0) and similarly for Y. Tensor products distribute over direct sums, and $\mathbb{Z} \otimes A = A$:

$$\begin{aligned} KU(X) \otimes KU(Y) &\overset{(1)}{=} \left(\widetilde{KU}(X) \oplus \mathbb{Z}\right) \otimes \left(\widetilde{KU}(Y) \oplus \mathbb{Z}\right) \\ &\overset{(2)}{=} \left(\widetilde{KU}(X) \otimes \widetilde{KU}(Y)\right) \oplus \left(\widetilde{KU}(X) \otimes \mathbb{Z}\right) \oplus \left(\mathbb{Z} \otimes \widetilde{KU}(Y)\right) \oplus (\mathbb{Z} \otimes \mathbb{Z}) \\ &\overset{(3)}{=} \left(\widetilde{KU}(X) \otimes \widetilde{KU}(Y)\right) \oplus \widetilde{KU}(X) \oplus \widetilde{KU}(Y) \oplus \mathbb{Z}. \end{aligned} \tag{3.121}$$

(1) The splitting (3.89) for X and for Y.
(2) $\otimes$ distributes over $\oplus$.
(3) $A \otimes \mathbb{Z} \cong A$.

This is Nash's (3.81).

Reducing the right side. The inclusion and the collapse

$$X \vee Y \xrightarrow{i} X \times Y \longrightarrow \frac{X \times Y}{X \vee Y} = X \wedge Y \tag{3.122}$$

(Nash's (3.82)) induce maps on K-theory in the reverse direction, forming a short exact sequence

$$0 \to \widetilde{KU}(X \wedge Y) \to \widetilde{KU}(X \times Y) \to \widetilde{KU}(X \vee Y) \to 0 \tag{3.123}$$

(Nash's (3.83)). What the three maps mean:

- $\widetilde{KU}(X \times Y) \to \widetilde{KU}(X \vee Y)$ is restriction to the axes. It is onto, and even split: a class on $X \vee Y$ is a pair (class on X, class on Y), and $\pi_X^* a + \pi_Y^* b$ restricts to it. This also shows $\widetilde{KU}(X \vee Y) \cong \widetilde{KU}(X) \oplus \widetilde{KU}(Y)$: a bundle over a wedge is a bundle over each piece, glued along the single fibre at the common point.
- $\widetilde{KU}(X \wedge Y) \to \widetilde{KU}(X \times Y)$ pulls back along the collapse. A bundle pulled back from $X \wedge Y$ is trivial on the axes (they map to a single point).
- Exactness in the middle is the converse: a (stable) bundle on $X \times Y$ that is trivial on $X \vee Y$ can be trivialised there and then *descends* to the quotient $X \wedge Y$, where the collapsed axes carry the chosen trivialisation.

A split short exact sequence gives a direct sum (as in Section 3.10.6):

$$\begin{aligned} \widetilde{KU}(X \times Y) &\overset{(1)}{=} \widetilde{KU}(X \wedge Y) \oplus \widetilde{KU}(X \vee Y) \\ &\overset{(2)}{=} \widetilde{KU}(X \wedge Y) \oplus \widetilde{KU}(X) \oplus \widetilde{KU}(Y), \end{aligned} \tag{3.124}$$

(1) The split exact sequence (3.123).
(2) $\widetilde{KU}(X \vee Y) \cong \widetilde{KU}(X) \oplus \widetilde{KU}(Y)$.

which is Nash's (3.84). Adding the $\mathbb{Z}$ from the base point, $KU(X \times Y) = \widetilde{KU}(X \wedge Y) \oplus \widetilde{KU}(X) \oplus \widetilde{KU}(Y) \oplus \mathbb{Z}$.

The reduced product. Comparing (3.121) and (3.124), α has the form

$$\alpha : \big(\widetilde{KU}(X) \otimes \widetilde{KU}(Y)\big) \oplus R \longrightarrow \widetilde{KU}(X \wedge Y) \oplus R, \qquad R = \widetilde{KU}(X) \oplus \widetilde{KU}(Y) \oplus \mathbb{Z} \tag{3.125}$$

(Nash's (3.85)), and it is the identity on R (e.g. $a \otimes 1 \mapsto \pi_X^* a$). Discarding R leaves the **reduced external product**

$$\tilde{\alpha} : \widetilde{KU}(X) \otimes \widetilde{KU}(Y) \to \widetilde{KU}(X \wedge Y), \qquad a \otimes b \mapsto \pi_X^* a \bullet \pi_Y^* b \tag{3.126}$$

(Nash's (3.86)). Why does $\pi_X^* a \bullet \pi_Y^* b$ come from $X \wedge Y$? Restricted to $X \times \{y_0\}$ it is $a \bullet (\text{rank of } b) = a \cdot 0 = 0$, since b has virtual rank zero; similarly on $\{x_0\} \times Y$. So it vanishes on the axes, and by exactness of (3.123) it comes from $X \wedge Y$.

Worked example 3.76: $\tilde{\alpha}$ for $X = Y = S^2$

$\widetilde{KU}(S^2) \otimes \widetilde{KU}(S^2) = \mathbb{Z}b \otimes \mathbb{Z}b = \mathbb{Z}(b \otimes b)$, and $\tilde{\alpha}(b \otimes b) = b_1 b_2$, the class of Example 3.74. Since $S^2 \wedge S^2 = S^4$, the periodicity theorem says $b_1 b_2$ generates $\widetilde{KU}(S^4) = \mathbb{Z}$. So the reduced

K-theory of S^4 is generated by the "square" of the Hopf class, and by induction

$$\widetilde{KU}(S^{2j}) = \mathbb{Z} \cdot b^{\boxtimes j}, \qquad b^{\boxtimes j} = \underbrace{b \boxtimes \cdots \boxtimes b}_{j} \text{ on } S^2 \wedge \cdots \wedge S^2 = S^{2j}.$$

3.13.6 The Bott map and the periodicity theorem

Now put $Y = S^2$ in $\tilde{\alpha}$ and replace X by $S^n X$. Bott periodicity says the result is an isomorphism:

$$\begin{aligned} \widetilde{KU}(S^n X) \otimes \widetilde{KU}(S^2) &\overset{(1)}{\cong} \widetilde{KU}(S^n X \wedge S^2) \\ &\overset{(2)}{=} \widetilde{KU}(S^{n+2} \wedge X) \\ &\overset{(3)}{=} \widetilde{KU}^{-(n+2)}(X). \end{aligned} \tag{3.127}$$

(1) $\tilde{\alpha}$ with $Y = S^2$ is an isomorphism (the reduced form of (3.120)).
(2) $S^n X \wedge S^2 = S^n \wedge X \wedge S^2 \cong S^n \wedge S^2 \wedge X \cong S^{n+2} \wedge X$ (the smash product is commutative and associative up to homeomorphism, and $S^n \wedge S^2 = S^{n+2}$).
(3) $S^{n+2} \wedge X = S^{n+2}X$ and the definition (3.115).

The left side is $\widetilde{KU}^{-n}(X) \otimes \widetilde{KU}(S^2)$ (Nash's (3.87)).

The generator b. By Section 3.12.7, $\widetilde{KU}(S^2) = \mathbb{Z}$, generated by $b = [\eta] - [I]$ where $\eta = H$ is the canonical line bundle over $\mathbb{CP}^1$. Nash names it after the Hopf fibration: the unit vectors in the fibres of H,

$$\{(z_0, z_1) \in \mathbb{C}^2 : |z_0|^2 + |z_1|^2 = 1\} = S^3 \longrightarrow \mathbb{CP}^1 = S^2, \qquad (z_0, z_1) \mapsto [z_0 : z_1],$$

form the Hopf fibration, with circle fibres (the phases $e^{i\chi}$ of a unit vector on a given line). It is also the bundle of the Dirac monopole (Section 3.8). Since $\widetilde{KU}(S^2)$ is free on one generator, $A \otimes \widetilde{KU}(S^2) \cong A$ via $a \otimes b \mapsto a$, and we may drop the tensor factor.

Key idea: Bott periodicity, K-theory form (Nash (3.88)–(3.89))

The **Bott map**

$$\gamma : \widetilde{KU}^{-n}(X) \longrightarrow \widetilde{KU}^{-(n+2)}(X), \qquad a \longmapsto \tilde{\alpha}(a \otimes b),$$

"multiply externally by the Hopf class $b = [H] - 1$", is an isomorphism:

$$\widetilde{KU}^{-n}(X) \cong \widetilde{KU}^{-(n+2)}(X) \qquad \text{for every } X,\ n \geq 0.$$

Recovering the sphere statement. Take $X = S^0$. Then $S^n S^0 = S^n$ and

$$\widetilde{KU}(S^n) \overset{(1)}{=} \widetilde{KU}^{-n}(S^0) \overset{(2)}{\cong} \widetilde{KU}^{-(n+2)}(S^0) \overset{(1)}{=} \widetilde{KU}(S^{n+2}),$$

(1) Definition (3.115) and $S^n S^0 = S^n$.
(2) The Bott map γ.

which is (3.112), Nash's (3.73). Explicitly, the generator of $\widetilde{KU}(S^{n+2})$ is b times the generator of $\widetilde{KU}(S^n)$.

$$\widetilde{KU}(S^0)=\mathbb{Z} \xrightarrow{\cdot b} \widetilde{KU}(S^2)=\mathbb{Z} \xrightarrow{\cdot b} \widetilde{KU}(S^4)=\mathbb{Z} \xrightarrow{\cdot b} \widetilde{KU}(S^6)=\mathbb{Z} \xrightarrow{\cdot b} \cdots$$

$$1 \longmapsto b \longmapsto b\boxtimes b \longmapsto b^{\boxtimes 3} \longmapsto \cdots$$

$$\widetilde{KU}(S^1)=0 \xrightarrow{\cdot b} \widetilde{KU}(S^3)=0 \xrightarrow{\cdot b} \widetilde{KU}(S^5)=0 \xrightarrow{\cdot b} \widetilde{KU}(S^7)=0 \xrightarrow{\cdot b} \cdots$$

Figure 3.52: The Bott map at work on spheres. External multiplication by the Hopf class $b=[H]-1$ raises the dimension of the sphere by two and is an isomorphism at every step. Starting from $\widetilde{KU}(S^0)=\mathbb{Z}$ (generator 1) and $\widetilde{KU}(S^1)=0$ it produces the whole pattern. The generator of $\widetilde{KU}(S^{2j})$ is the j-fold external power of b.

Consequences.

- Since $K^{-n}=K^{-n-2}$, we can *define* K^n for positive n as well, by periodicity. Complex K-theory then has only two distinct groups, $K^0(X)$ and $K^1(X):=K^{-1}(X)$: it is a $\mathbb{Z}_2$-graded theory, like the even/odd split of fermion number.
- The long exact sequences of the Eilenberg–Steenrod exactness axiom close up into a hexagon (the *six-term exact sequence*) linking K^0 and K^1 of a pair.
- K of any space built from spheres can be computed: e.g. $\widetilde{KU}(S^{2j})=\mathbb{Z}$, $\widetilde{KU}(S^{2j+1})=0$, $K(S^2\times S^2)=\mathbb{Z}^4$, and for the torus T^2, $K^0(T^2)=\mathbb{Z}^2$, $K^1(T^2)=\mathbb{Z}^2$ (build T^2 from the wedge $S^1\vee S^1$ by attaching one 2-cell, the square of Figure 3.51, and use the exact sequences).

Why it is true: clutching functions become polynomials

Here is the idea of the Atiyah–Bott proof of (3.120), at the physicist's level. A bundle on $X\times S^2$ is obtained, as in Section 3.12.2, by gluing two trivial pieces over $X\times D_+$ and $X\times D_-$ with a clutching function

$$g(x,z):\ X\times S^1\to \mathrm{GL}(N,\mathbb{C}),\qquad z=e^{i\phi}.$$

1. *Fourier.* Expand g in Fourier modes in ϕ and truncate: up to homotopy, g is a Laurent polynomial $\sum_{j=-L}^{L} a_j(x)z^j$. Multiplying by z^L (tensoring with a power of H) makes it a polynomial.
2. *Linearise.* A polynomial clutching function of degree L in z can be traded, by enlarging the matrix size (a "companion matrix" trick, as when a high-order ODE is written as a first-order system), for a *linear* one, $a(x)z+c(x)$.
3. *Projectors.* A linear clutching function can be deformed to the form $g(x,z)=z\,\Pi(x)+(1-\Pi(x))$ with $\Pi(x)$ a family of projectors. It multiplies the image of Π by z and leaves the kernel alone, so the bundle is $(\mathrm{Im}\,\Pi)\boxtimes H\ \oplus\ (\mathrm{Ker}\,\Pi)\boxtimes I$.

So every bundle on $X\times S^2$ is, stably, of the form $E\boxtimes H\oplus F\boxtimes I$ with E,F bundles on X, i.e. it lies in the image of α, and a similar argument for homotopies shows there are no extra relations. For $X=\mathrm{pt}$, step 3 says: every bundle on S^2 is $H^{\oplus r}\oplus I^s$ stably, exactly what Section 3.12.7 found.

Worked example 3.77: The linear clutching function on S^2

For $X = \mathrm{pt}$ and $\Pi = \mathrm{diag}(1,0)$,

$$g(z) = z\,\Pi + (1-\Pi) = \begin{pmatrix} z & 0 \\ 0 & 1 \end{pmatrix},$$

which is the clutching function of $H \oplus I$. For the rank-one projector $\Pi(\vec{n}) = \tfrac{1}{2}(1 + \vec{n}\cdot\vec{\sigma})$ of Example 3.61 the same formula gives $g(z) = z\Pi(\vec{n}) + (1-\Pi(\vec{n}))$, which is conjugate to $\mathrm{diag}(z,1)$ for each fixed $\vec{n}$; as $\vec{n}$ varies over a parameter space X we get exactly the external product of the ground-state bundle with H, plus the excited-state bundle times I: $(\mathrm{Im}\,\Pi)\boxtimes H \oplus (\ker\Pi)\boxtimes I$. This is the Bott map made concrete.

Reading Nash (pp. 75–78, (3.71)–(3.89))

Nash's §4 begins with $\widetilde{KU}(S^n) = \pi_{n-1}(\mathrm{U}(\infty))$ (3.71) and Bott's theorem $\pi_n(\mathrm{U}(\infty)) = \pi_{n+2}(\mathrm{U}(\infty))$ (3.72), giving (3.73). He explains it via $\pi_n(X) = \pi_{n-2}(\Omega^2 X)$ (3.74) and $\Omega^2\mathrm{U} \simeq \mathrm{U}$. He defines $\widetilde{KU}^{-n}(X) = \widetilde{KU}(S^n X)$ (3.75)–(3.76), the internal product β (footnote 5) and the external product α (3.77)–(3.79), and states that α is an isomorphism for $Y = S^2$ (3.80), citing Husemoller. He reduces α to $\tilde{\alpha}$ using (3.81)–(3.86), and obtains the Bott map γ (3.88) and periodicity (3.89) by setting $Y = S^2$ and $X \to S^n X$ (3.87). In (3.84) the last term "$K(Y)$" should read $\widetilde{K}(Y)$. We have added the table of unstable groups, the winding-number integrals, the Morse-theory sketch of $\Omega\mathrm{U} \simeq \mathbb{Z}\times BU$, the computation of $K^{-n}(\mathrm{pt})$ and $K^{-1}(X) = [X,\mathrm{U}]$, pictures of suspension and smash product, the example $K(S^2\times S^2)$ with the Bott class, and the Atiyah–Bott idea of polynomial clutching functions.

Recap

- $\widetilde{KU}(S^n) = \pi_{n-1}(\mathrm{U})$ and $\pi_n(\mathrm{U}) = \pi_{n+2}(\mathrm{U})$: $0,\mathbb{Z},0,\mathbb{Z},\ldots$; the $\mathbb{Z}$'s are winding numbers $\propto \int \mathrm{tr}(g^{-1}\mathrm{d}g)^{2j-1}$.
- Mechanism: $\Omega\mathrm{U} \simeq \mathbb{Z}\times BU$ and $\Omega BU \simeq \mathrm{U}$, so $\Omega^2\mathrm{U} \simeq \mathrm{U}$ (Bott: minimal geodesics form Grassmannians).
- $\widetilde{KU}^{-n}(X) = \widetilde{KU}(S^n X)$; $K^{-n}(\mathrm{pt}) = \mathbb{Z},0,\mathbb{Z},0,\ldots$; $K^0 = [X,\mathbb{Z}\times BU]$, $K^{-1} = [X,\mathrm{U}]$.
- External product $\alpha(c\otimes d) = \pi_X^* c \bullet \pi_Y^* d$; $\widetilde{K}(X\times Y) = \widetilde{K}(X\wedge Y)\oplus\widetilde{K}(X)\oplus\widetilde{K}(Y)$; reduced product $\tilde{\alpha}$ lands in $\widetilde{K}(X\wedge Y)$.
- Bott map $\gamma(a) = \tilde{\alpha}(a\otimes b)$, $b = [H]-1$, is an isomorphism $\widetilde{KU}^{-n}(X) \to \widetilde{KU}^{-n-2}(X)$.

3.14 Real and quaternionic periodicity: the period eight

Nash ends his section on Bott periodicity in four lines. He "simply quotes" the real and quaternionic versions and remarks that we can now compute $\widetilde{KO}(S^n)$, $\widetilde{KU}(S^n)$ and $\widetilde{KSp}(S^n)$ for every n. Those four lines hide one of the most striking patterns in mathematics. The real theory has period *eight*, the same eight that controls Clifford algebras, Majorana spinors, and the dimensions where indices are only defined mod 2.

This section unpacks the four lines. We state the theorem and tabulate it. We then see where the eight comes from in three ways: from Bott's chain of loop spaces, from Clifford

modules (the Atiyah–Bott–Shapiro construction, which also *writes down* the generators), and from the comparison between real and complex K-theory. We finish with applications: vector fields on spheres and a dictionary of places where the period eight appears in physics.

Throughout, $\mathrm{Cl}_{0,k}$ is the real Clifford algebra with k generators squaring to -1, as in Chapter T4:

$$e_i e_j + e_j e_i = -2\delta_{ij}, \qquad i, j = 1, \dots, k.$$

3.14.1 Nash's statement and the full tables

Write $\mathrm{O} = \mathrm{O}(\infty)$ and $\mathrm{Sp} = \mathrm{Sp}(\infty)$ for the infinite orthogonal and symplectic groups, defined like $\mathrm{U}(\infty)$ by upper-left block inclusions $\mathrm{O}(k) \hookrightarrow \mathrm{O}(k+1)$, $\mathrm{Sp}(k) \hookrightarrow \mathrm{Sp}(k+1)$. By the stability results of Section 3.12.4, $\pi_j(\mathrm{O}) = \pi_j(\mathrm{O}(k))$ once $k > j+1$, and $\pi_j(\mathrm{Sp}) = \pi_j(\mathrm{Sp}(k))$ once $j \le 4k+1$ ((3.102)).

Key idea: Real and quaternionic Bott periodicity (Nash (3.90))

$$\pi_n(\mathrm{O}) = \pi_{n+8}(\mathrm{O}), \qquad \widetilde{KO}^{-n}(X) \cong \widetilde{KO}^{-n-8}(X),$$
$$\pi_n(\mathrm{Sp}) = \pi_{n+4}(\mathrm{O}), \qquad \widetilde{KSp}^{-n}(X) \cong \widetilde{KO}^{-n-4}(X).$$

Here $\widetilde{KO}^{-n}(X) := \widetilde{KO}(S^n X)$ and $\widetilde{KSp}^{-n}(X) := \widetilde{KSp}(S^n X)$, exactly as in (3.115). By symmetry also $\pi_n(\mathrm{O}) = \pi_{n+4}(\mathrm{Sp})$ and $\widetilde{KO}^{-n}(X) \cong \widetilde{KSp}^{-n-4}(X)$.

So the quaternionic theory is the real theory shifted by half a period. Going round twice, $\mathrm{Sp} \to \mathrm{O} \to \mathrm{Sp}$, is a shift by 8, consistent with the first line.

The homotopy groups. One period of each stable group:

n mod 8	0	1	2	3	4	5	6	7
$\pi_n(\mathrm{O})$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	0	$\mathbb{Z}$	0	0	0	$\mathbb{Z}$
$\pi_n(\mathrm{U})$	0	$\mathbb{Z}$	0	$\mathbb{Z}$	0	$\mathbb{Z}$	0	$\mathbb{Z}$
$\pi_n(\mathrm{Sp})$	0	0	0	$\mathbb{Z}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	0	$\mathbb{Z}$

(3.128)

The U row has period 2 (Section 3.13). The Sp row is the O row read four places further along. Let us check the second line of Nash's (3.90) entry by entry:

$$\pi_0(\mathrm{Sp}) \overset{(1)}{=} 0 = \pi_4(\mathrm{O}), \qquad \pi_1(\mathrm{Sp}) \overset{(1)}{=} 0 = \pi_5(\mathrm{O}),$$
$$\pi_2(\mathrm{Sp}) \overset{(2)}{=} 0 = \pi_6(\mathrm{O}), \qquad \pi_3(\mathrm{Sp}) \overset{(3)}{=} \mathbb{Z} = \pi_7(\mathrm{O}),$$
$$\pi_4(\mathrm{Sp}) \overset{(4)}{=} \mathbb{Z}_2 = \pi_8(\mathrm{O}) \overset{(5)}{=} \pi_0(\mathrm{O}), \qquad \pi_5(\mathrm{Sp}) \overset{(4)}{=} \mathbb{Z}_2 = \pi_9(\mathrm{O}) \overset{(5)}{=} \pi_1(\mathrm{O}).$$

(1) $\mathrm{Sp}(1) = S^3$ is connected and simply connected, and these groups are already stable (table (3.108)).
(2) Cartan: π_2 of any compact Lie group vanishes.
(3) $\pi_3(\mathrm{Sp}(1)) = \pi_3(S^3) = \mathbb{Z}$, the instanton number.
(4) $\pi_4(\mathrm{Sp}) = \pi_4(\mathrm{Sp}(1)) = \pi_4(S^3) = \mathbb{Z}_2$, and similarly $\pi_5(\mathrm{Sp}) = \pi_5(S^3) = \mathbb{Z}_2$; for these j the group $\mathrm{Sp}(1)$ is already in the stable range, since $j \le 4 \cdot 1 + 1$.
(5) The first line of (3.90).

The first four entries are the tables of Section 3.12.5; the $\mathbb{Z}_2$'s in degrees 4, 5 are the classical homotopy groups $\pi_4(S^3) = \pi_5(S^3) = \mathbb{Z}_2$ of the three-sphere, which we quote.

The K-groups of every sphere. Clutching gives $\widetilde{KO}(S^n) = \pi_{n-1}(\mathrm{O})$, $\widetilde{KU}(S^n) = \pi_{n-1}(\mathrm{U})$, $\widetilde{KSp}(S^n) = \pi_{n-1}(\mathrm{Sp})$ for $n \geq 1$ ((3.101), (3.107) with $k \to \infty$). Shifting the index by one:

n mod 8	0	1	2	3	4	5	6	7
$\widetilde{KO}(S^n) = KO^{-n}(\mathrm{pt})$	$\mathbb{Z}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	0	$\mathbb{Z}$	0	0	0
$\widetilde{KU}(S^n) = KU^{-n}(\mathrm{pt})$	$\mathbb{Z}$	0	$\mathbb{Z}$	0	$\mathbb{Z}$	0	$\mathbb{Z}$	0
$\widetilde{KSp}(S^n) = KSp^{-n}(\mathrm{pt})$	$\mathbb{Z}$	0	0	0	$\mathbb{Z}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	0

(3.129)

The identification with the groups of a point is the one of Section 3.13.3: $KO^{-n}(\mathrm{pt}) = \widetilde{KO}(S^n S^0) = \widetilde{KO}(S^n)$. The $n = 0$ column is $\widetilde{K}(S^0) = \mathbb{Z}$ (a bundle over two points modulo its rank at the base point), and is consistent with periodicity: $\widetilde{KO}(S^8) = \pi_7(\mathrm{O}) = \mathbb{Z}$. This is Nash's closing remark: there is now enough information to calculate any of these groups.

Worked example 3.78: Reading off K-groups of large spheres

$$\widetilde{KO}(S^{27}) \overset{(1)}{=} \pi_{26}(\mathrm{O}) \overset{(2)}{=} \pi_2(\mathrm{O}) \overset{(3)}{=} 0,$$
$$\widetilde{KSp}(S^{13}) \overset{(4)}{=} \widetilde{KO}(S^{17}) \overset{(1)}{=} \pi_{16}(\mathrm{O}) \overset{(2)}{=} \pi_0(\mathrm{O}) \overset{(3)}{=} \mathbb{Z}_2,$$
$$\widetilde{KO}(S^{12}) \overset{(1)}{=} \pi_{11}(\mathrm{O}) \overset{(2)}{=} \pi_3(\mathrm{O}) \overset{(3)}{=} \mathbb{Z}, \qquad \widetilde{KU}(S^{11}) \overset{(5)}{=} \pi_{10}(\mathrm{U}) \overset{(6)}{=} \pi_0(\mathrm{U}) = 0.$$

(1) Clutching: $\widetilde{KO}(S^n) = \pi_{n-1}(\mathrm{O})$.
(2) Real Bott periodicity, $26 = 3 \cdot 8 + 2$, $16 = 2 \cdot 8$, $11 = 8 + 3$.
(3) The table (3.128).
(4) The second line of (3.90) with $X = S^0$: $\widetilde{KSp}(S^n) = \widetilde{KO}(S^{n+4})$.
(5) Clutching for U.
(6) Complex Bott periodicity, five times.

As a cross-check of the second line, directly: $\widetilde{KSp}(S^{13}) = \pi_{12}(\mathrm{Sp}) = \pi_4(\mathrm{Sp}) = \mathbb{Z}_2$ from (3.128) (using that the Sp row has period 8 too). Concretely: there is exactly one stably non-trivial quaternionic bundle over S^{13}, and there are infinitely many stably distinct real bundles over S^{12}, labelled by an integer.

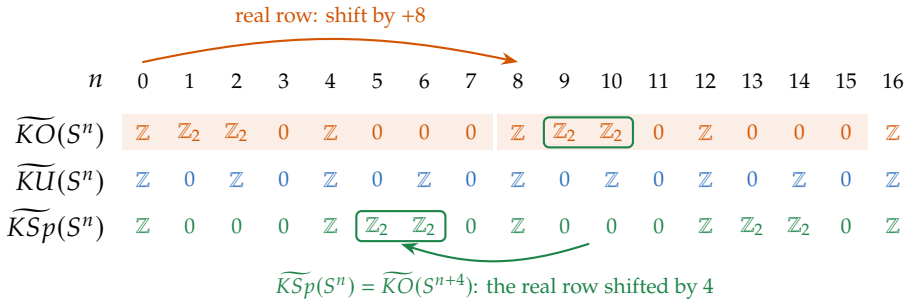


Figure 3.53: The reduced K-groups of spheres up to S^{16}. The real row (orange) repeats every 8 (the shaded blocks are identical). The complex row (blue) repeats every 2. The quaternionic row (green) is the real row shifted by 4: for example the $\mathbb{Z}_2$'s of $\widetilde{KSp}$ at $n = 5, 6$ (green box) are the $\mathbb{Z}_2$'s of $\widetilde{KO}$ at $n = 9, 10$ (green box).

3.14.2 From unstable to stable: π_3 and π_4 of $\mathrm{SO}(k)$

Periodicity is a statement about $k = \infty$. For a physicist it is worth seeing how the small groups settle down. The tool is the list of *exceptional isomorphisms* of spin groups from Chapter T4:

$$\mathrm{Spin}(3) = \mathrm{SU}(2), \quad \mathrm{Spin}(4) = \mathrm{SU}(2) \times \mathrm{SU}(2), \quad \mathrm{Spin}(5) = \mathrm{Sp}(2), \quad \mathrm{Spin}(6) = \mathrm{SU}(4).$$

$\mathrm{Spin}(k) \to \mathrm{SO}(k)$ is a double cover, and a covering map does not change π_j for $j \geq 2$ (a map from a simply connected sphere lifts uniquely once a base point is chosen). So $\pi_j(\mathrm{SO}(k)) = \pi_j(\mathrm{Spin}(k))$ for $j \geq 2$.

Worked example 3.79: $\pi_3(\mathrm{SO}(k))$ and $\pi_4(\mathrm{SO}(k))$ for $k = 3, \ldots, 7$

Using $\pi_3(S^3) = \mathbb{Z}$, $\pi_4(S^3) = \mathbb{Z}_2$, and the stability ranges:

k	$\mathrm{Spin}(k)$	$\pi_3(\mathrm{SO}(k))$	$\pi_4(\mathrm{SO}(k))$
3	$\mathrm{SU}(2) = S^3$	$\mathbb{Z}$	$\mathbb{Z}_2$
4	$S^3 \times S^3$	$\mathbb{Z} \oplus \mathbb{Z}$	$\mathbb{Z}_2 \oplus \mathbb{Z}_2$
5	$\mathrm{Sp}(2)$	$\mathbb{Z}$	$\mathbb{Z}_2$
6	$\mathrm{SU}(4)$	$\mathbb{Z}$	0
7	(stable)	$\mathbb{Z}$	0

The entries for $k = 5$ and $k = 6$ come from chains like

$$\pi_4(\mathrm{SO}(5)) \overset{(1)}{=} \pi_4(\mathrm{Sp}(2)) \overset{(2)}{=} \pi_4(\mathrm{Sp}(1)) \overset{(3)}{=} \pi_4(S^3) = \mathbb{Z}_2,$$
$$\pi_4(\mathrm{SO}(6)) \overset{(1)}{=} \pi_4(\mathrm{SU}(4)) \overset{(4)}{=} \pi_4(\mathrm{U}) \overset{(5)}{=} 0.$$

(1) Double cover and the exceptional isomorphism.
(2) The inclusion $\mathrm{Sp}(1) \to \mathrm{Sp}(2)$ has quotient $\mathrm{Sp}(2)/\mathrm{Sp}(1) = S^7$, which has no homotopy below degree 7, so it is an isomorphism on π_j for $j < 6$.
(3) $\mathrm{Sp}(1) = S^3$.
(4) $\mathrm{SU}(4) \to \mathrm{U}(4)$ does not change π_j for $j \geq 2$ (the quotient is $\mathrm{U}(1)$), and $\mathrm{U}(4)$ is in the stable range for π_4 ((3.102): $j \leq 2k - 1$, here $4 \leq 7$).
(5) Bott: $\pi_{\mathrm{even}}(\mathrm{U}) = 0$.

Reading the columns: π_3 passes through $\mathbb{Z} \oplus \mathbb{Z}$ at $k = 4$ and settles at $\mathbb{Z} = \pi_3(\mathrm{O})$; π_4 settles at $0 = \pi_4(\mathrm{O})$ from $k = 6$ on, exactly when the stability bound $k > j + 1 = 5$ is met.

Physical picture: The $\mathbb{Z} \oplus \mathbb{Z}$ at $k = 4$

$\mathrm{SO}(4)$ bundles over S^4 are the frame bundles of four-dimensional Euclidean gauge theory with gauge group $\mathrm{SO}(4) = (\mathrm{SU}(2)_L \times \mathrm{SU}(2)_R)/\mathbb{Z}_2$. The two integers are the instanton numbers of the two $\mathrm{SU}(2)$ factors (self-dual and anti-self-dual parts). Adding a trivial line, $\mathrm{SO}(4) \subset \mathrm{SO}(5)$, identifies them into a single integer: stably only one combination survives. This is the topological shadow of the fact that $\Lambda^2\mathbb{R}^4$ splits into self-dual and anti-self-dual pieces only in dimension four (Chapter T1).

3.14.3 Where the eight comes from, I: Bott's chain of loop spaces

In the complex case the mechanism was $\Omega\mathrm{U} \simeq \mathbb{Z} \times BU$ and $\Omega(\mathbb{Z} \times BU) \simeq \mathrm{U}$: two steps, period two (Section 3.13.2). In the real case Bott's Morse-theory argument produces eight steps before

returning to the start:

$$\begin{aligned} &\Omega(\mathbb{Z}\times BO)\simeq \mathrm{O}, \quad \Omega\mathrm{O}\simeq \mathrm{O/U}, \quad \Omega(\mathrm{O/U})\simeq \mathrm{U/Sp}, \quad \Omega(\mathrm{U/Sp})\simeq \mathbb{Z}\times BSp,\\ &\Omega(\mathbb{Z}\times BSp)\simeq \mathrm{Sp}, \quad \Omega\mathrm{Sp}\simeq \mathrm{Sp/U}, \quad \Omega(\mathrm{Sp/U})\simeq \mathrm{U/O}, \quad \Omega(\mathrm{U/O})\simeq \mathbb{Z}\times BO. \end{aligned} \tag{3.130}$$

Here O/U means $\mathrm{O}(2m)/\mathrm{U}(m)$ with $m\to\infty$ (orthogonal complex structures on $\mathbb{R}^{2m}$), U/Sp means $\mathrm{U}(2m)/\mathrm{Sp}(m)$ (quaternionic structures on $\mathbb{C}^{2m}$), Sp/U means $\mathrm{Sp}(m)/\mathrm{U}(m)$, and U/O means $\mathrm{U}(m)/\mathrm{O}(m)$ (real forms of $\mathbb{C}^m$, i.e. Lagrangian planes). Composing eight loopings, $\Omega^8\mathrm{O}\simeq\mathrm{O}$, and $\Omega^4\mathrm{O}\simeq\mathbb{Z}\times BSp$, $\Omega^4\mathrm{Sp}\simeq\mathbb{Z}\times BO$, which is (3.90).

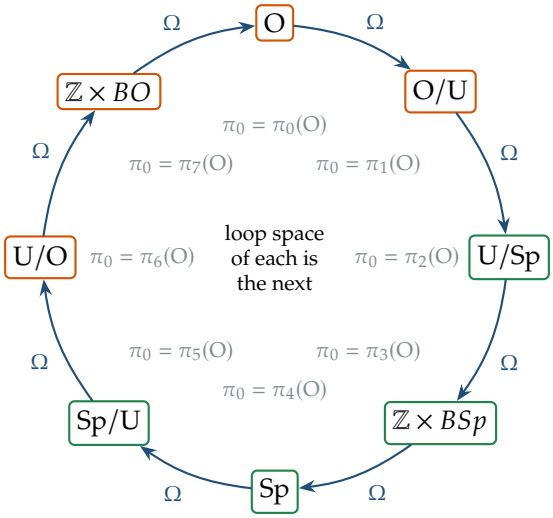


Figure 3.54: Bott's eight-step clock. Following an arrow means "take the loop space"; after eight steps we are back at O, so $\Omega^8\mathrm{O}\simeq\mathrm{O}$. The inner labels say what each space contributes: its set of components π_0 is $\pi_k(\mathrm{O})$, because $\pi_k(\mathrm{O})=\pi_0(\Omega^k\mathrm{O})$. Real-flavoured spaces are orange, quaternionic-flavoured ones green; the clock is half real, half quaternionic, and the quaternionic half is the real half shifted by four steps.

Let us check this chain against the table (3.128), using $\pi_k(\mathrm{O})=\pi_{k-1}(\Omega\mathrm{O})=\cdots=\pi_0(\Omega^k\mathrm{O})$ ((3.113) repeatedly):

$$\begin{aligned} &\pi_0(\mathrm{O})\overset{(1)}{=}\mathbb{Z}_2, && \pi_1(\mathrm{O})\overset{(2)}{=}\pi_0(\mathrm{O/U})\overset{(3)}{=}\mathbb{Z}_2,\\ &\pi_2(\mathrm{O})\overset{(2)}{=}\pi_0(\mathrm{U/Sp})\overset{(4)}{=}0, && \pi_3(\mathrm{O})\overset{(2)}{=}\pi_0(\mathbb{Z}\times BSp)\overset{(5)}{=}\mathbb{Z},\\ &\pi_4(\mathrm{O})\overset{(2)}{=}\pi_0(\mathrm{Sp})\overset{(6)}{=}0, && \pi_5(\mathrm{O})\overset{(2)}{=}\pi_0(\mathrm{Sp/U})\overset{(6)}{=}0,\\ &\pi_6(\mathrm{O})\overset{(2)}{=}\pi_0(\mathrm{U/O})\overset{(6)}{=}0, && \pi_7(\mathrm{O})\overset{(2)}{=}\pi_0(\mathbb{Z}\times BO)\overset{(5)}{=}\mathbb{Z}. \end{aligned}$$

(1) Two components, det $=\pm1$.
(2) $\pi_k(\mathrm{O})=\pi_0(\Omega^k\mathrm{O})$ and the chain (3.130).
(3) A complex structure J on $\mathbb{R}^{2m}$ ($J^2=-1$, J orthogonal) is either orientation-compatible with the standard one or not: $\mathrm{O}(2m)/\mathrm{U}(m)$ has two components. (For $m=1$: $J=\pm\left(\begin{smallmatrix}0&-1\\1&0\end{smallmatrix}\right)$, a rotation by $\pm90°$.)
(4) $\mathrm{U}(2m)$ is connected, so its quotient is.
(5) The components of $\mathbb{Z}\times BG$ are labelled by the virtual rank.
(6) $\mathrm{Sp}(m)$ and $\mathrm{U}(m)$ are connected, so they and their quotients are.

All eight entries agree with the table. The eight spaces are exactly the compact symmetric spaces that appear in Cartan's classification, and they return in physics as the eight real classes of random matrices and of topological insulators (Section 3.14.7).

Physical picture: Complex structures that anticommute: Clifford algebras appear

Milnor's version of Bott's proof describes the k-th space in the chain as the set of orthogonal matrices J on $\mathbb{R}^N$ (N large) with

$$J^2 = -1, \qquad JJ_i = -J_iJ \quad (i = 1, \dots, k-1),$$

for fixed $J_1, \dots, J_{k-1}$ satisfying the same relations. In words: $J_1, \dots, J_{k-1}, J$ make $\mathbb{R}^N$ into a module over $\mathrm{Cl}_{0,k}$. The minimal geodesics from J_k to $-J_k$ in the k-th space are parametrised by the $(k+1)$-st space (compare Figure 3.49). So each loop step adds one anticommuting generator, and *the period eight of Bott is the period eight of Clifford algebras*, $\mathrm{Cl}_{0,k+8} \cong \mathrm{Cl}_{0,k} \otimes \mathbb{R}(16)$ (Chapter T4).

Worked example 3.80: $\pi_1(\mathrm{U}/\mathrm{O}) = \mathbb{Z}$ is the Maslov index

The chain says $\pi_1(\mathrm{U}/\mathrm{O}) = \pi_7(\mathrm{O}) = \mathbb{Z}$. We can see this $\mathbb{Z}$ directly. $\mathrm{U}(m)/\mathrm{O}(m)$ is the space of Lagrangian planes in $\mathbb{C}^m = \mathbb{R}^{2m}$: the image $g\mathbb{R}^m$ of the real subspace under a unitary g, where g and gh ($h \in \mathrm{O}(m)$) give the same plane. The function

$$\mu : \mathrm{U}(m)/\mathrm{O}(m) \to \mathrm{U}(1), \qquad [g] \mapsto \det(g)^2$$

is well defined, because

$$\det(gh)^2 \overset{(1)}{=} \det(g)^2 \det(h)^2 \overset{(2)}{=} \det(g)^2.$$

(1) det is multiplicative.
(2) h is real orthogonal, so $\det h = \pm 1$ and $\det(h)^2 = 1$.

The winding number of $\det^2$ around a loop of Lagrangian planes is the **Maslov index**. For $m = 1$ the Lagrangian "planes" are the lines $e^{i\theta}\mathbb{R} \subset \mathbb{C}$, $\theta \in [0, \pi)$, and $\mu = e^{2i\theta}$ winds once as the line turns through 180° back to itself. Physically: in semiclassical (WKB) quantisation a periodic orbit carries a loop of Lagrangian planes (the tangent planes of the invariant torus), and its Maslov index μ corrects the Bohr–Sommerfeld rule to $\oint p\,\mathrm{d}q = 2\pi\hbar(n + \mu/4)$. For the harmonic oscillator $\mu = 2$ (two turning points), which gives $E = \hbar\omega(n + \frac{1}{2})$: the zero-point energy is a $\pi_7(\mathrm{O})$ in disguise.

3.14.4 Where the eight comes from, II: real versus complex

A second way to see the pattern, and to see that the $\mathbb{Z}$'s in $\widetilde{KO}(S^4)$ and $\widetilde{KO}(S^8)$ are *not* the same kind of $\mathbb{Z}$, is to compare real and complex bundles.

Working definition 3.35 (complexification, realification, conjugation)

- **Complexification** $c : KO(X) \to KU(X)$ [K-groups built from real, respectively complex, bundles], $[E] \mapsto [E \otimes_{\mathbb{R}} \mathbb{C}]$. Transition functions: the same real matrices, now regarded as complex.
- **Realification** $r : KU(X) \to KO(X)$, $[F] \mapsto [F_{\mathbb{R}}]$: forget the complex structure [the rule for multiplying fibre vectors by i]. A complex rank-k bundle becomes a real rank-$2k$ bundle; a transition function $A + iB$ becomes the real $2k \times 2k$ matrix $\left(\begin{smallmatrix} A & -B \\ B & A \end{smallmatrix}\right)$.
- **Conjugation** $\psi : KU(X) \to KU(X)$, $[F] \mapsto [\bar{F}]$: complex conjugate the transition functions. For a line bundle, $\bar{L} \cong L^* = L^{-1}$ (with a Hermitian metric [a complex inner

product on each fibre, which lets one take g unitary], $\bar{g} = g^{-1}$ for $g \in \mathrm{U}(1)$).

They satisfy

$$r \circ c = 2, \qquad c \circ r = 1 + \psi. \tag{3.131}$$

["2" is multiplication by 2: $E \otimes_{\mathbb{R}} \mathbb{C}$, viewed as a real bundle, is $E \oplus E$]

Why (3.131): for a real E, $(E \otimes_{\mathbb{R}} \mathbb{C})_{\mathbb{R}} = E \oplus iE \cong E \oplus E$. For a complex F, $F_{\mathbb{R}} \otimes_{\mathbb{R}} \mathbb{C} \cong F \oplus \bar{F}$: the complex structure J of F extends $\mathbb{C}$-linearly to $F_{\mathbb{R}} \otimes \mathbb{C}$, where $J^2 = -1$ splits it into the $+i$ eigenspace ($\cong F$) and the $-i$ eigenspace ($\cong \bar{F}$). This is the same splitting as $T_{\mathbb{C}}\mathbb{R}^2 = T^{1,0} \oplus T^{0,1}$, spanned by ∂_z and $\partial_{\bar{z}}$.

Worked example 3.81: c kills $\widetilde{KO}(S^2)$

The generator of $\widetilde{KO}(S^2) = \mathbb{Z}_2$ is $r(b) = [H_{\mathbb{R}}] - 2$, the Hopf bundle regarded as a real plane bundle (Section 3.12.5). Its complexification is

$$c\big(r(b)\big) \overset{(1)}{=} b + \psi(b) \overset{(2)}{=} ([H] - 1) + ([H^{-1}] - 1)$$
$$\overset{(3)}{=} (1 + b - 1) + (1 - b - 1) = 0.$$

(1) (3.131).
(2) $\psi(b) = [\bar{H}] - 1 = [H^{-1}] - 1$ for the line bundle H.
(3) $[H^n] = 1 + nb$ for all $n \in \mathbb{Z}$, (3.110).

So $c = 0$ on $\widetilde{KO}(S^2)$. This had to be: c is a group homomorphism $\mathbb{Z}_2 \to \mathbb{Z}$, and the only one is zero. Physically, the monopole of charge +1 and the monopole of charge −1 are the same *real* bundle (complex conjugation is a real isomorphism $H_{\mathbb{R}} \cong \bar{H}_{\mathbb{R}}$), and together they form a trivial complex bundle. Also $r(c(r(b))) = 2r(b) = 0$, consistent with $r(b)$ having order 2.

Worked example 3.82: c doubles on $\widetilde{KO}(S^4)$

Let E be the quaternionic Hopf line bundle over $S^4 = \mathbb{H}P^1$ (Example 3.69), viewed as a complex rank-2 bundle with clutching $g(q) = q \in \mathrm{SU}(2)$. By Example 3.70, $y = [E] - 2$ generates $\widetilde{KU}(S^4) = \mathbb{Z}$, and $\alpha_4 = r(y) = [E_{\mathbb{R}}] - 4$ generates $\widetilde{KO}(S^4) = \mathbb{Z}$. Quaternionic multiplication by $\mathbf{j}$ on the right is a complex-antilinear isomorphism $E \to E$, so $\bar{E} \cong E$ and $\psi(y) = y$. Then

$$c(\alpha_4) \overset{(1)}{=} c\big(r(y)\big) \overset{(2)}{=} y + \psi(y) \overset{(3)}{=} 2y.$$

(1) Definition of α_4.
(2) (3.131).
(3) $\bar{E} \cong E$.

So $c : \widetilde{KO}(S^4) = \mathbb{Z} \to \widetilde{KU}(S^4) = \mathbb{Z}$ is multiplication by 2. In terms of characteristic classes: the real generator has first Pontryagin number ±2, twice the instanton number of the complex generator, because $p_1(E_{\mathbb{R}}) = -c_2(E \oplus \bar{E}) = -2c_2(E)$.

In degree 8, by contrast, $c : \widetilde{KO}(S^8) \to \widetilde{KU}(S^8)$ is an isomorphism: the real generator (built from octonions in Section 3.14.5) complexifies to the generator $b^{\boxtimes 4}$. The full comparison over

one period is:

n	0	1	2	3	4	5	6	7
$\widetilde{KO}(S^n)$	$\mathbb{Z}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	0	$\mathbb{Z}$	0	0	0
c	$\cong$	0	0	0	$\times 2$	0	0	0
$\widetilde{KU}(S^n)$	$\mathbb{Z}$	0	$\mathbb{Z}$	0	$\mathbb{Z}$	0	$\mathbb{Z}$	0

(3.132)

The ring $KO^{-*}(\mathrm{pt})$. Products (external, as in Section 3.13.4) make $\bigoplus_n KO^{-n}(\mathrm{pt})$ a ring. Its generators are

$$\eta \in KO^{-1}\ (\text{Möbius} - 1), \qquad \alpha_4 \in KO^{-4}, \qquad \beta_8 \in KO^{-8}\ (\text{the real Bott class}),$$

with $\eta^2 = r(b)$ generating $KO^{-2} = \mathbb{Z}_2$, and relations

$$2\eta = 0, \qquad \eta^3 = 0, \qquad \eta\,\alpha_4 = 0, \qquad \alpha_4^2 = 4\beta_8.$$

The first three hold because the groups they land in are $\mathbb{Z}_2$ (with $2x = 0$), $KO^{-3} = 0$ and $KO^{-5} = 0$. The last one is a genuine calculation, done by complexifying. Since c is a ring homomorphism and is injective on $KO^{-8} = \mathbb{Z}$,

$$c(\alpha_4^2) \overset{(1)}{=} c(\alpha_4)^2 \overset{(2)}{=} (2b^{\boxtimes 2})^2 = 4\,b^{\boxtimes 4} \overset{(3)}{=} 4\,c(\beta_8) \overset{(4)}{=} c(4\beta_8),$$

(1) c preserves products.
(2) Example 3.82, with the generator of $\widetilde{KU}(S^4)$ written as $b^{\boxtimes 2}$ (up to a sign, which squares away).
(3) c is an isomorphism in degree 8, $c(\beta_8) = b^{\boxtimes 4}$.
(4) c is additive.

and injectivity gives $\alpha_4^2 = 4\beta_8$. Multiplication by β_8 is the real Bott map $\widetilde{KO}^{-n}(X) \to \widetilde{KO}^{-n-8}(X)$ of (3.90), the analogue of γ in (3.127). The quaternionic isomorphism of (3.90) is similarly multiplication by the class of the quaternionic Hopf bundle in $\widetilde{KSp}(S^4)$ (a real bundle tensored with a quaternionic one is quaternionic).

Watch out: Period eight does not mean "the same $\mathbb{Z}$"

The groups $\widetilde{KO}(S^4)$ and $\widetilde{KO}(S^8)$ are both $\mathbb{Z}$, but they sit differently relative to complex K-theory: one complexifies with a factor 2, the other isomorphically. The honest period is 8, not 4: $\widetilde{KO}(S^5) = 0$ but $\widetilde{KO}(S^9) = \mathbb{Z}_2$. Nash's statement $\widetilde{KSp}^{-n} \cong \widetilde{KO}^{-n-4}$ is a relation between two *different* theories, not a period 4 of either.

3.14.5 Where the eight comes from, III: Clifford modules build the generators

The brief of this chapter promised a link with Clifford algebras. Here it is, in the form found by Atiyah, Bott and Shapiro (1964). It does something the homotopy tables cannot: it *writes down* a clutching function for the generator of every $\widetilde{KO}(S^k)$, and it explains the whole table by linear algebra.

Clutching by Clifford multiplication

Recall the clutching construction (Section 3.12.2): a bundle over S^k is two trivial pieces over the hemispheres $D^k_\pm$, glued along the equator S^{k-1} by a map $g : S^{k-1} \to \mathrm{GL}(N)$. Clifford algebras supply such maps for free.

Working definition 3.36 (the ABS clutching function)

Let $M = M^0 \oplus M^1$ be a **graded module** of $\mathrm{Cl}_{0,k}$: real matrices $e_1, \dots, e_k$ on M with $e_i e_j + e_j e_i = -2\delta_{ij}$, each mapping $M^0 \to M^1$ and $M^1 \to M^0$ ("odd"). For a unit vector $x = (x_1, \dots, x_k) \in S^{k-1}$ put

$$g(x) := x\cdot = \sum_{i=1}^{k} x_i e_i\Big|_{M^0} : M^0 \longrightarrow M^1.$$

It is invertible, because

$$(x\cdot)^2 \overset{(1)}{=} \sum_{i,j} x_i x_j\, e_i e_j \overset{(2)}{=} \tfrac{1}{2}\sum_{i,j} x_i x_j\,(e_i e_j + e_j e_i) \overset{(3)}{=} -\sum_i x_i^2 \overset{(4)}{=} -1,$$

(1) Expand; the x_i are numbers.
(2) $x_i x_j$ is symmetric in i, j.
(3) The Clifford relation.
(4) x is a unit vector.

so $g(x)^{-1} = -x\cdot$ (restricted to M^1). Gluing $D^k_+ \times M^0$ to $D^k_- \times M^1$ with g gives a real bundle over S^k; its class minus the trivial bundle is the **ABS class** $[M] \in \widetilde{KO}(S^k)$ [real virtual bundles over S^k of virtual dimension 0].

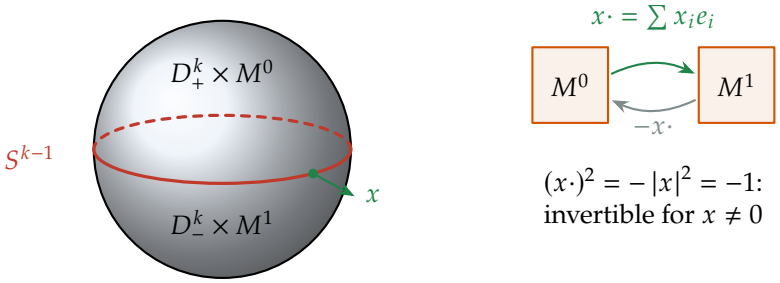


Figure 3.55: The Atiyah–Bott–Shapiro construction. Over the northern hemisphere the fibre is M^0, over the southern it is M^1. At a point x of the equator (red) they are glued by Clifford multiplication by x (green), which is invertible because a unit vector squares to -1. The Clifford algebra hands us a clutching function without any choices.

This is the same construction as the symbol of a Dirac operator: on $\mathbb{R}^k$, $\not{D} = \sum e_i \partial_i$ has symbol $i\sum \xi_i e_i$, invertible for $\xi \neq 0$. The ABS class of the spinor module is the class of the Dirac symbol, which is how the index theorem (Chapter 4) enters K-theory.

Worked example 3.83: The generators in degrees $1, 2, 4, 8$

$k = 1$. $\mathrm{Cl}_{0,1} = \mathbb{C}$ with $e_1 = i$; grade it as $M^0 = \mathbb{R}$, $M^1 = i\mathbb{R}$. The equator of S^1 is $S^0 = \{\pm 1\}$, and $g(\pm 1) = \pm i$. Identifying M^1 with M^0 via e_1, the clutching function is $+1$ at one point and -1 at the other: a flip. This is the Möbius bundle (Example 3.35), the generator η of $\widetilde{KO}(S^1) = \mathbb{Z}_2$.

$k = 2$. $\mathrm{Cl}_{0,2} = \mathbb{H}$ with $e_1 = \mathbf{i}$, $e_2 = \mathbf{j}$ and $J := e_1 e_2 = \mathbf{k}$. Grade it by $M^0 = \mathrm{span}(1, J) \cong \mathbb{C}$ and $M^1 = \mathrm{span}(e_1, e_2) = e_1 M^0$. For $x = (\cos\phi, \sin\phi)$,

$$\cos\phi\, e_1 + \sin\phi\, e_2 \overset{(1)}{=} e_1\big(\cos\phi - \sin\phi\, e_1 e_2\big) \overset{(2)}{=} e_1\, e^{-\phi J}.$$

(1) $e_1(-e_1 e_2) = -e_1^2 e_2 = e_2$.

(2) $J^2 = e_1e_2e_1e_2 = -e_1^2e_2^2 = -1$, so J behaves like i and $\cos\phi - \sin\phi\, J = e^{-\phi J}$.

After identifying $M^1 = e_1M^0$ with $M^0 \cong \mathbb{C}$, the clutching function is multiplication by the phase $e^{-i\phi}$: winding number ±1 (the sign is orientation). This is the Hopf bundle regarded as a real plane bundle, the generator $\eta^2 = r(b)$ of $\widetilde{KO}(S^2)$.

$k = 4$. Let $u_1, \dots, u_4 = 1, \mathbf{i}, \mathbf{j}, \mathbf{k}$ and let $\mathrm{Cl}_{0,4}$ act on $M = \mathbb{H} \oplus \mathbb{H}$ (real dimension 8) by

$$e_\mu = \begin{pmatrix} 0 & -\bar{u}_\mu \\ u_\mu & 0 \end{pmatrix} \quad \text{(left multiplication by quaternions).}$$

The relations hold:

$$e_\mu e_\nu + e_\nu e_\mu \overset{(1)}{=} \begin{pmatrix} -(\bar{u}_\mu u_\nu + \bar{u}_\nu u_\mu) & 0 \\ 0 & -(u_\mu \bar{u}_\nu + u_\nu \bar{u}_\mu) \end{pmatrix} \overset{(2)}{=} -2\,\mathrm{Re}(\bar{u}_\mu u_\nu) \overset{(3)}{=} -2\delta_{\mu\nu}.$$

(1) Block multiplication.

(2) For quaternions, $\bar{p}q + \bar{q}p = 2\,\mathrm{Re}(\bar{p}q) = p\bar{q} + q\bar{p}$ (the real part of $\bar{p}q$ is the Euclidean dot product in $\mathbb{R}^4$).

(3) $1, \mathbf{i}, \mathbf{j}, \mathbf{k}$ are orthonormal.

For $x \in S^3$ the clutching function $M^0 \to M^1$ is left multiplication by the unit quaternion $q = \sum x_\mu u_\mu$: $g(q) = q$, the clutching function of the quaternionic Hopf bundle (Example 3.69). It gives the generator α_4 of $\widetilde{KO}(S^4)$, the BPST instanton bundle. The matrices e_μ are the Euclidean gamma matrices in a chiral basis, $\gamma_\mu = \left(\begin{smallmatrix} 0 & \bar{\sigma}_\mu \\ \sigma_\mu & 0 \end{smallmatrix}\right)$ (Chapter T4), up to factors of i.

$k = 8$. Replace quaternions by octonions $\mathbb{O}$ (real dimension 8), with unit basis $u_1, \dots, u_8$. The same block formula on $M = \mathbb{O} \oplus \mathbb{O}$ gives sixteen-dimensional real gamma matrices for $\mathrm{Cl}_{0,8} = \mathbb{R}(16)$. The octonions are not associative, but they are *alternative*: $\bar{u}(ux) = |u|^2 x$ for all u, x, and polarising this in u gives $\bar{u}_\mu(u_\nu x) + \bar{u}_\nu(u_\mu x) = 2\delta_{\mu\nu}x$, which is all the computation above needs. The clutching function is left octonion multiplication, $S^7 \to \mathrm{GL}(8,\mathbb{R})$, $x \mapsto L_x$, and its class is the real Bott class β_8 generating $\widetilde{KO}(S^8) = \mathbb{Z}$.

The complex story is the same with $\mathrm{Cl}_{0,k} \otimes \mathbb{C}$. For $k = 2$ use $e_1 = i\sigma_1$, $e_2 = i\sigma_2$, graded by the eigenvalues of σ_3. Then

$$x\cdot = i \begin{pmatrix} 0 & x_1 - ix_2 \\ x_1 + ix_2 & 0 \end{pmatrix},$$

and the block $M^0 \to M^1$ is $i(x_1 + ix_2) = ie^{i\phi}$: the Hopf bundle H, generator b of $\widetilde{KU}(S^2)$. For $k = 2m$ the chiral block of $x \cdot \gamma$ on Weyl spinors $S^+ \to S^-$ gives the generator $b^{\boxtimes m}$ of $\widetilde{KU}(S^{2m})$. Bott periodicity is "tensoring with one more pair of gamma matrices".

Which modules give trivial bundles

Not every module gives a new class. Here is the one mechanism that kills classes.

Key idea: Extendable modules give trivial bundles

If the graded $\mathrm{Cl}_{0,k}$-module M extends to a graded $\mathrm{Cl}_{0,k+1}$-module (an extra odd e_{k+1} with $e_{k+1}^2 = -1$ anticommuting with $e_1, \dots, e_k$), then its ABS bundle over S^k is trivial.

The reason is a rotation into the new direction. For $t \in [0, \pi/2]$ put

$$g_t(x) := \cos t\,(x\cdot) + \sin t\, e_{k+1} : \; M^0 \to M^1.$$

Each $g_t(x)$ is Clifford multiplication by the vector $(\cos t\, x, \sin t) \in \mathbb{R}^{k+1}$, which has length one, so by the same computation as before $g_t(x)^2 = -1$ and $g_t(x)$ is invertible. At $t = 0$ it is g; at $t = \pi/2$ it is the constant e_{k+1}, which clutches a trivial bundle. A homotopy of clutching functions does not change the bundle (Section 3.12.2), so the ABS bundle is trivial.

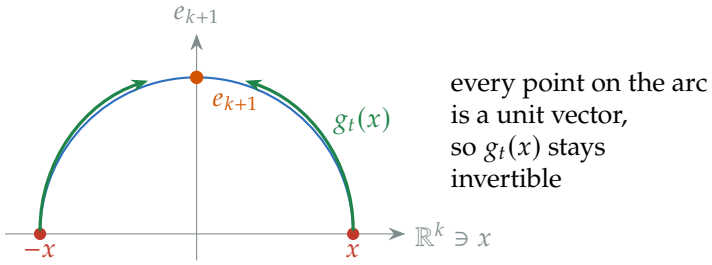


Figure 3.56: Why an extra Clifford generator trivialises the bundle. For each equator point x we rotate the unit vector x towards the new direction e_{k+1} (green). Every point of the quarter circle is a unit vector, so Clifford multiplication by it stays invertible. At the end all equator points have been moved to the same vector e_{k+1}, and the clutching function is constant.

So the ABS classes form a quotient

$$A_k := \frac{\{\text{graded } \mathrm{Cl}_{0,k}\text{-modules}\}}{\{\text{those that extend to } \mathrm{Cl}_{0,k+1}\}} \longrightarrow \widetilde{KO}(S^k),$$

where "modules" means the Grothendieck group of modules under $\oplus$, exactly as in Section 3.10.

Key idea: Atiyah–Bott–Shapiro

The map $A_k \to \widetilde{KO}(S^k)$ is an isomorphism for every k. The complex version, with $\mathrm{Cl}_{0,k} \otimes \mathbb{C}$, gives $A_k^{\mathbb{C}} \cong \widetilde{KU}(S^k)$.

Counting modules: the table from linear algebra

We now compute A_k by hand. Two facts make this pure bookkeeping.

(a) Graded $\mathrm{Cl}_{0,k}$-modules are ungraded $\mathrm{Cl}_{0,k-1}$-modules. Given a graded M, the even elements $f_i := e_i e_k$ $(i < k)$ map M^0 to itself, and

$$f_i^2 \overset{(1)}{=} e_i e_k e_i e_k \overset{(2)}{=} -e_i e_i e_k e_k \overset{(3)}{=} -(-1)(-1) = -1,$$
$$f_i f_j \overset{(2)}{=} e_i e_k e_j e_k = -e_i e_j e_k e_k \overset{(3)}{=} e_i e_j \qquad (i \neq j),$$

(1) Definition of f_i.
(2) Move one e_k past one e_i or e_j $(i, j \neq k)$: one sign.
(3) $e_i^2 = e_k^2 = -1$.

so $f_i f_j = e_i e_j = -e_j e_i = -f_j f_i$ for $i \neq j$. Thus $f_1, \dots, f_{k-1}$ make M^0 a $\mathrm{Cl}_{0,k-1}$-module. Conversely, from a $\mathrm{Cl}_{0,k-1}$-module W one rebuilds $M = W \oplus W$. So graded $\mathrm{Cl}_{0,k}$-modules correspond to ungraded $\mathrm{Cl}_{0,k-1}$-modules, and extending the graded module to $\mathrm{Cl}_{0,k+1}$ corresponds to extending W to $\mathrm{Cl}_{0,k}$. Hence

$$A_k \cong \frac{N_{k-1}}{\text{image of } N_k \text{ under restriction}}, \qquad N_j := \{\text{ungraded } \mathrm{Cl}_{0,j}\text{-modules}\}. \tag{3.133}$$

(b) The modules of $\mathrm{Cl}_{0,j}$ are read off from its matrix type (Chapter T4, table of Clifford algebras). A matrix algebra $\mathbb{F}(m)$ ($\mathbb{F} = \mathbb{R}, \mathbb{C}, \mathbb{H}$) has exactly one irreducible module, $\mathbb{F}^m$; a sum $\mathbb{F}(m) \oplus \mathbb{F}(m)$

has two. So $N_j = \mathbb{Z}$ or $\mathbb{Z}^2$, and we record the real dimension d_j of the irreducible modules:

j	0	1	2	3	4	5	6	7	8
$\mathrm{Cl}_{0,j}$	$\mathbb{R}$	$\mathbb{C}$	$\mathbb{H}$	$\mathbb{H}\oplus\mathbb{H}$	$\mathbb{H}(2)$	$\mathbb{C}(4)$	$\mathbb{R}(8)$	$\mathbb{R}(8)\oplus\mathbb{R}(8)$	$\mathbb{R}(16)$
N_j	$\mathbb{Z}$	$\mathbb{Z}$	$\mathbb{Z}$	$\mathbb{Z}^2$	$\mathbb{Z}$	$\mathbb{Z}$	$\mathbb{Z}$	$\mathbb{Z}^2$	$\mathbb{Z}$
d_j	1	2	4	4	8	8	8	8	16

(3.134)

When there are two irreducible modules ($j = 3, 7$), they are distinguished by the action of the volume element $\omega = e_1 e_2 \cdots e_j$, which is central with $\omega^2 = +1$, acting as $+1$ on one and -1 on the other. (In general $\omega^2 = (-1)^{j(j-1)/2}(-1)^j$ for $\mathrm{Cl}_{0,j}$, which is $+1$ for $j = 3, 7$ and -1 for $j = 1, 5$.)

Now restrict. An irreducible $\mathrm{Cl}_{0,k}$-module of dimension d_k, viewed as a $\mathrm{Cl}_{0,k-1}$-module, decomposes into irreducibles of dimension d_{k-1}. When $N_{k-1} = \mathbb{Z}$ it contains d_k/d_{k-1} copies. When $N_{k-1} = \mathbb{Z}^2$ ($k - 1 = 3, 7$), the extra generator e_k anticommutes with the volume element $\omega = e_1 \cdots e_{k-1}$ (moving e_k past $k-1$ odd generators gives $(-1)^{k-1} = -1$ for $k = 4, 8$), so e_k swaps the $\omega = \pm 1$ eigenspaces: they have equal size, and the restriction contains one of each.

Worked example 3.84: The ABS count for $k = 1, \ldots, 8$

Combining (3.133) and (3.134):

$$A_1 \overset{(1)}{=} \mathbb{Z}/(d_1/d_0)\mathbb{Z} = \mathbb{Z}/2\mathbb{Z} = \mathbb{Z}_2, \qquad A_2 \overset{(1)}{=} \mathbb{Z}/(d_2/d_1)\mathbb{Z} = \mathbb{Z}/2\mathbb{Z} = \mathbb{Z}_2,$$
$$A_3 \overset{(2)}{=} \mathbb{Z}/\langle 1, 1\rangle = 0, \qquad A_4 \overset{(3)}{=} \mathbb{Z}^2/\langle (1,1)\rangle = \mathbb{Z},$$
$$A_5 \overset{(1)}{=} \mathbb{Z}/(d_5/d_4)\mathbb{Z} = \mathbb{Z}/\mathbb{Z} = 0, \qquad A_6 \overset{(1)}{=} \mathbb{Z}/(d_6/d_5)\mathbb{Z} = 0,$$
$$A_7 \overset{(2)}{=} \mathbb{Z}/\langle 1, 1\rangle = 0, \qquad A_8 \overset{(3)}{=} \mathbb{Z}^2/\langle (1,1)\rangle = \mathbb{Z}.$$

(1) $N_{k-1} = N_k = \mathbb{Z}$, and the irreducible $\mathrm{Cl}_{0,k}$-module restricts to d_k/d_{k-1} copies of the irreducible $\mathrm{Cl}_{0,k-1}$-module; for example the $\mathbb{H}$-module $\mathbb{H} = \mathbb{C}^2$ restricts to two copies of $\mathbb{C}$.

(2) $N_{k-1} = \mathbb{Z}$, $N_k = \mathbb{Z}^2$: each of the two irreducible $\mathrm{Cl}_{0,k}$-modules has dimension $d_k = d_{k-1}$ and restricts to the irreducible $\mathrm{Cl}_{0,k-1}$-module; the image is generated by 1 and 1, which is all of $\mathbb{Z}$.

(3) $N_{k-1} = \mathbb{Z}^2$, $N_k = \mathbb{Z}$: the irreducible module restricts to one copy of each $\omega = \pm 1$ module, so the image is the diagonal $\mathbb{Z}(1,1)$, and $\mathbb{Z}^2/\mathbb{Z}(1,1) \cong \mathbb{Z}$ via $(a, b) \mapsto a - b$.

The result $\mathbb{Z}_2, \mathbb{Z}_2, 0, \mathbb{Z}, 0, 0, 0, \mathbb{Z}$ is exactly the $\widetilde{KO}(S^k)$ row of (3.129). The $\mathbb{Z}_2$'s come from the *doubling* of irreducible dimension ($1 \to 2 \to 4$), the $\mathbb{Z}$'s from the *splitting* into two modules by the volume element, and the zeros from everything else. Since $\mathrm{Cl}_{0,k+8} = \mathrm{Cl}_{0,k} \otimes \mathbb{R}(16)$ changes each d_j by the same factor 16 and nothing else, the count repeats with period 8: this is real Bott periodicity, derived from linear algebra.

Try it

Do the complex count. $\mathrm{Cl}_{0,j} \otimes \mathbb{C}$ is $\mathbb{C}(2^{j/2})$ for j even (one module, complex dimension $2^{j/2}$) and $\mathbb{C}(2^{(j-1)/2}) \oplus \mathbb{C}(2^{(j-1)/2})$ for j odd (two modules). Show that $A_k^{\mathbb{C}} = \mathbb{Z}$ for k even and 0 for k odd.

Answer. k even: $N_{k-1}^{\mathbb{C}} = \mathbb{Z}^2$, $N_k^{\mathbb{C}} = \mathbb{Z}$ restricts to $(1,1)$, so $A_k^{\mathbb{C}} = \mathbb{Z}^2/\mathbb{Z}(1,1) = \mathbb{Z}$. k odd: $N_{k-1}^{\mathbb{C}} = \mathbb{Z}$, and each of the two modules of $N_k^{\mathbb{C}}$ restricts to the single irreducible, so the image is all of $\mathbb{Z}$ and $A_k^{\mathbb{C}} = 0$. This is $\widetilde{KU}(S^k) = \mathbb{Z}, 0, \mathbb{Z}, 0, \ldots$: complex Bott periodicity with period 2.

3.14.6 Clifford modules at work: vector fields on spheres

In Example 3.42 we framed S^1 and S^3 by explicit vector fields, and noted that S^n is parallelisable only for $n = 1, 3, 7$. Clifford modules answer the more refined question: *how many* everywhere linearly independent vector fields does S^{n-1} carry?

Working definition 3.37 (vector fields from a Clifford module)

Suppose $\mathbb{R}^n$ is a module over $\mathrm{Cl}_{0,k}$ [there are real $n \times n$ matrices $e_1, \dots, e_k$ with $e_i e_j + e_j e_i = -2\delta_{ij}$], with the e_i orthogonal matrices (always possible: average an inner product over the finite group generated by the e_i). Then for $x \in S^{n-1}$ the k vectors

$$v_1(x) = e_1 x, \ \dots, \ v_k(x) = e_k x$$

are orthonormal and tangent to S^{n-1} at x [perpendicular to x, because an orthogonal e_i with $e_i^2 = -1$ is antisymmetric, so $\langle x, e_i x\rangle = 0$].

Check: e_i orthogonal with $e_i^2 = -1$ means $e_i^T = e_i^{-1} = -e_i$ (antisymmetric). Then

$$\langle e_i x, x\rangle \overset{(1)}{=} \langle x, e_i^T x\rangle \overset{(2)}{=} -\langle x, e_i x\rangle \overset{(3)}{=} -\langle e_i x, x\rangle \quad \Longrightarrow \quad \langle e_i x, x\rangle = 0,$$
$$\langle e_i x, e_j x\rangle \overset{(1)}{=} \langle x, e_i^T e_j x\rangle \overset{(2)}{=} -\langle x, e_i e_j x\rangle \overset{(4)}{=} \tfrac{1}{2}\langle x, -(e_i e_j + e_j e_i)x\rangle \overset{(5)}{=} \delta_{ij}\,|x|^2 .$$

(1) Definition of the transpose.
(2) $e_i^T = -e_i$.
(3) Symmetry of the inner product.
(4) $\langle x, Ax\rangle$ only sees the symmetric part of A.
(5) The Clifford relation.

The first line says $v_i(x) \perp x$ (tangent), the second that the v_i are orthonormal.

Worked example 3.85: S^1, S^3, S^7, S^{15}

$n = 2$: $\mathbb{R}^2 = \mathbb{C}$ is a $\mathrm{Cl}_{0,1}$-module ($e_1 = i$), giving $v_1(x) = ix$: the rotation field $(-x_2, x_1)$ on S^1 of Example 3.42.
$n = 4$: $\mathbb{R}^4 = \mathbb{H}$ is a $\mathrm{Cl}_{0,3}$-module via right multiplication by $\mathbf{i}, \mathbf{j}, \mathbf{k}$ (right multiplications reverse the order of products but preserve anticommutation), giving the three fields $q\mathbf{i}$, $q\mathbf{j}$, $q\mathbf{k}$ of Example 3.42: a full frame of TS^3.
$n = 8$: $\mathbb{R}^8 = \mathbb{O}$ is a $\mathrm{Cl}_{0,7}$-module (an irreducible module of $\mathbb{R}(8) \oplus \mathbb{R}(8)$) via the seven imaginary octonion units: seven fields, a frame of TS^7.
$n = 16$: from (3.134), the largest k with d_k dividing 16 is $k = 8$ ($d_8 = 16$, $d_9 = 32$). So S^{15} carries 8 independent vector fields, not 15: it is not parallelisable.

The general count is the **Hurwitz–Radon number**: write $n = 2^{4a+b} \cdot (\text{odd})$ with $0 \le b \le 3$; then $\rho(n) = 2^b + 8a$, and the construction gives $\rho(n) - 1$ fields on S^{n-1}. For $n = 2, 4, 8, 16, 32$ this is $1, 3, 7, 8, 9$ fields. For odd n it is 0: the hairy ball theorem on even-dimensional spheres (Example 3.43). That no sphere carries *more* fields than this was proved by Adams (1962), using $\widetilde{KO}$ of real projective spaces and operations on K-theory. So the question "which spheres are parallelisable?", raised in Section 3.9, is answered by the same period-eight structure: $\rho(n) = n$ only for $n = 1, 2, 4, 8$.

3.14.7 The period eight in physics

The eight real K-groups of a point, $KO^{-n}(\mathrm{pt}) = \mathbb{Z}, \mathbb{Z}_2, \mathbb{Z}_2, 0, \mathbb{Z}, 0, 0, 0$, keep reappearing. A few places where they do:

Dictionary: $KO^{-n}(\mathrm{pt})$ in physics

Majorana spinors	Real spinors exist when the Clifford algebra is a real matrix algebra, which depends on the dimension and signature mod 8 (Chapter T4). This is the same period eight, seen in the algebra before the topology.
Mod-2 indices	A real skew-adjoint Dirac-type operator has an index defined only mod 2 (the number of zero modes mod 2). It lives in $KO^{-1}(\mathrm{pt}) = \mathbb{Z}_2$ or $KO^{-2}(\mathrm{pt}) = \mathbb{Z}_2$, so it occurs in dimensions $8m+1$ and $8m+2$.
Global anomalies	Witten's SU(2) anomaly uses $\pi_4(\mathrm{SU}(2)) = \pi_4(\mathrm{Sp}(1)) = \mathbb{Z}_2$; stably this is $\widetilde{KSp}(S^5) = \widetilde{KO}(S^9) = \mathbb{Z}_2$, a mod-2 index (Chapter 10).
Topological insulators and superconductors	The "ten-fold way" (Kitaev, 2009): gapped free-fermion phases in d dimensions in real symmetry class s ($s = 0, \dots, 7$ for AI, BDI, D, DIII, AII, CII, C, CI) are classified by $KO^{-(s-d)}(\mathrm{pt})$. For example: Kitaev's Majorana chain (D, $d=1$) is $\mathbb{Z}_2$; the $p+ip$ superconductor (D, $d=2$) is $\mathbb{Z}$; the quantum spin Hall insulator (AII, $d=2$) is $\mathbb{Z}_2$. The two complex classes follow KU, with period 2.
Type I D-branes	Charges of Dp-branes in type I string theory lie in $\widetilde{KO}(S^{9-p})$ (Witten, 1998): $\mathbb{Z}$ for $p = 9, 5, 1$, and $\mathbb{Z}_2$ for $p = 8, 7, 0, -1$, the stable non-BPS branes. There is no D3-brane charge: $\widetilde{KO}(S^6) = 0$.

Worked example 3.86: Checking two entries of the ten-fold way

Class D has $s = 2$. For the Kitaev chain ($d = 1$) and the $p + ip$ superconductor ($d = 2$):

$$KO^{-(2-1)}(\mathrm{pt}) \overset{(1)}{=} \widetilde{KO}(S^1) \overset{(2)}{=} \mathbb{Z}_2, \qquad KO^{-(2-2)}(\mathrm{pt}) \overset{(1)}{=} \widetilde{KO}(S^0) \overset{(2)}{=} \mathbb{Z}.$$

(1) $KO^{-n}(\mathrm{pt}) = \widetilde{KO}(S^n)$.
(2) The table (3.129).

The $\mathbb{Z}_2$ of the Kitaev chain is the presence or absence of unpaired Majorana zero modes at its ends, exactly like the Möbius bundle's "twist or no twist", the generator η of $\widetilde{KO}(S^1)$. The $\mathbb{Z}$ of the $p + ip$ superconductor is a Chern number. For class AII ($s = 4$) in $d = 3$ one gets $KO^{-1} = \mathbb{Z}_2$: the strong invariant of a three-dimensional topological insulator.

Reading Nash (p. 78, (3.90))

Nash says: "We simply quote the corresponding facts for the real and the quaternionic cases. Both theories possess periodicities and these can be traced back to the stable homotopy of the infinite orthogonal and symplectic groups respectively." He gives (3.90): $\pi_n(\mathrm{O}(\infty)) = \pi_{n+8}(\mathrm{O}(\infty))$ with $\widetilde{KO}^{-n}(X) = \widetilde{KO}^{-n-8}(X)$, and $\pi_n(\mathrm{Sp}(\infty)) = \pi_{n+4}(\mathrm{O}(\infty))$ with $\widetilde{KSp}^{-n}(X) = \widetilde{KO}^{-n-4}(X)$, and concludes that there is now enough information to calculate $\widetilde{KO}(S^n)$, $\widetilde{KU}(S^n)$ and $\widetilde{KSp}(S^n)$ for all n. We have added: the full tables (3.128) and (3.129) with checks; the unstable groups of $\mathrm{SO}(k)$ via spin groups; Bott's eight-step chain of loop

spaces and the Maslov index; complexification and realification, including $c = 0$ on $\widetilde{KO}(S^2)$, $c = \times 2$ on $\widetilde{KO}(S^4)$, and the ring relation $\alpha_4^2 = 4\beta_8$; the Atiyah–Bott–Shapiro clutching functions and the Clifford module count that reproduces the real table; vector fields on spheres; and the physics dictionary.

Recap

- Real K-theory has period 8, quaternionic K-theory is real K-theory shifted by 4, complex K-theory has period 2: $\widetilde{KO}(S^n) = \mathbb{Z}, \mathbb{Z}_2, \mathbb{Z}_2, 0, \mathbb{Z}, 0, 0, 0$ (n mod 8), $\widetilde{KSp}(S^n) = \widetilde{KO}(S^{n+4})$.
- Mechanism 1: Bott's chain of eight loop spaces $\mathrm{O} \to \mathrm{O/U} \to \mathrm{U/Sp} \to \mathbb{Z} \times BSp \to \mathrm{Sp} \to \mathrm{Sp/U} \to \mathrm{U/O} \to \mathbb{Z} \times BO \to \mathrm{O}$; each step adds an anticommuting complex structure.
- Mechanism 2: $rc = 2$, $cr = 1 + \psi$. c kills $\widetilde{KO}(S^2)$, doubles on $\widetilde{KO}(S^4)$, is an isomorphism on $\widetilde{KO}(S^8)$; $\alpha_4^2 = 4\beta_8$.
- Mechanism 3 (ABS): Clifford multiplication $x \mapsto x\cdot$ is a clutching function; $\widetilde{KO}(S^k) =$ graded $\mathrm{Cl}_{0,k}$-modules modulo those extending to $\mathrm{Cl}_{0,k+1}$. The generators are the Möbius, Hopf, quaternionic Hopf (instanton) and octonionic bundles in degrees $1, 2, 4, 8$.
- Clifford modules also give $\rho(n) - 1$ vector fields on S^{n-1}; the period eight appears in physics as Majorana conditions, mod-2 indices, the ten-fold way, and type I D-brane charges.

3.15 Characteristic classes: what they measure and where they live

The last two sections built a machine, K-theory, that sorts vector bundles into classes and adds and multiplies those classes. Bott periodicity then told us what the K-groups of spheres are. But a physicist holding a particular bundle in their hands (the bundle of a monopole, of an instanton, the tangent bundle of a curved space) still has a practical problem: *how do we tell whether this bundle is twisted, and by how much?*

We already know the kind of answer we want. For the Dirac monopole the answer is a single integer, the magnetic charge n, which we get by integrating the field strength over a sphere that surrounds the monopole. For an SU(2) instanton on $\mathbb{R}^4$ it is again an integer, the instanton number $k = \frac{1}{8\pi^2}\int \operatorname{tr} F \wedge F$. In both cases we integrate a polynomial in the curvature over a closed manifold and get an integer that does not change when we deform the field.

This section and the next five explain the general machine behind such integers. The objects it produces are called **characteristic classes**. They are cohomology classes on the base space, built from the bundle, with three properties:

- isomorphic bundles give the same classes;
- they are compatible with pulling bundles back along maps;
- they vanish (apart from a trivial constant) on a trivial bundle.

So a nonzero characteristic class *proves* that a bundle is twisted, exactly as a nonzero magnetic flux through a sphere proves that no global vector potential exists.

There are two complementary ways in. Nash starts with the *topological* definition, pulling classes back from a classifying space, and only at the end of the section (his p. 87) turns to the *curvature* formulas. We follow the same order for the definition, then move the curvature formulas forward to Section 3.16, because curvature is where a physicist actually computes.

Reading Nash (p. 78, (3.87)–(3.90), the end of §4)

Nash's page 78 opens with the last step of Bott periodicity, which the previous part treated: $\widetilde{KU}^{-n}(X) \cong \widetilde{KU}^{-(n+2)}(X)$, proved by multiplying by the generator $b = [\eta] - [I]$ of $\widetilde{KU}(S^2) \cong \mathbb{Z}$. Here η is the Hopf line bundle over $\mathbb{CP}^1 = S^2$, which is also the bundle of the Dirac monopole, and I is the trivial line bundle. Two facts from there matter to us now. First, K-theory of spheres is made of integers. Second, those integers were found by homotopy theory, not by a formula. Characteristic classes are the formula: in Section 3.18 we will see that the Chern character turns b into the integer ± 1 by integrating a curvature over S^2.

3.15.1 The universal bundle and the classifying map

Here is the organising idea. There is one "mother" bundle of rank k from which every other rank-k complex bundle is obtained by pulling back. Any invariant we define on the mother bundle is then automatically defined on every bundle.

The mother bundle. Take a very large complex vector space $\mathbb{C}^N$. The **Grassmannian** $\mathrm{Gr}(k, N, \mathbb{C})$ is the set of all k-dimensional complex subspaces ("k-planes") of $\mathbb{C}^N$. It is a compact manifold. For $k = 1$ it is the space of lines through the origin, $\mathrm{Gr}(1, N, \mathbb{C}) = \mathbb{CP}^{N-1}$. Over the Grassmannian there is a bundle with a completely tautological description:

$$Q(k, N, \mathbb{C}) = \big\{(P, v) \;:\; P \in \mathrm{Gr}(k, N, \mathbb{C}),\; v \in P\big\}.$$

The fibre over the point P is the plane P itself. Letting $N \to \infty$ gives the **classifying space** $BU(k) = \mathrm{Gr}(k, \infty, \mathbb{C}) = \bigcup_N \mathrm{Gr}(k, N, \mathbb{C})$ and the **universal bundle** $E_{U(k)} = Q(k, \infty, \mathbb{C})$ over it. For $k = 1$ we get $BU(1) = \mathbb{CP}^\infty$ with its tautological line bundle.

The classifying map. Now let $E \to X$ be any rank-k complex bundle over a compact space X. It is always possible to put E inside a trivial bundle $X \times \mathbb{C}^N$ for some large N (choose N global sections that span every fibre, or choose a complement E' with $E \oplus E'$ trivial, as in the stable-equivalence argument of the previous part). Then each fibre E_x *is* a k-plane in $\mathbb{C}^N$, and we can define

$$f : X \longrightarrow \mathrm{Gr}(k, N, \mathbb{C}) \subset BU(k), \qquad f(x) = E_x. \tag{3.135}$$

By construction the fibre of the universal bundle over $f(x)$ is E_x again. In other words

$$E \;\cong\; f^* E_{U(k)}. \tag{3.136}$$

Physical picture: The Gauss map

The real version of this construction is an old friend. A surface $\Sigma \subset \mathbb{R}^3$ has at each point a tangent plane $T_x\Sigma$, which is a 2-plane in $\mathbb{R}^3$. The map $x \mapsto T_x\Sigma$ into the real Grassmannian $\mathrm{Gr}(2, 3, \mathbb{R})$ of planes in $\mathbb{R}^3$ is (essentially) Gauss's map $x \mapsto$ unit normal. The tangent bundle $T\Sigma$ is the pullback of the tautological plane bundle along this map. Gauss discovered that the total curvature $\int_\Sigma K \, \mathrm{d}A$ equals the signed area that the unit normal sweeps out on the unit sphere. Dividing by 2π gives the Euler characteristic of Σ. That is our first

characteristic class (the Euler class of Section 3.20), computed through the classifying map.

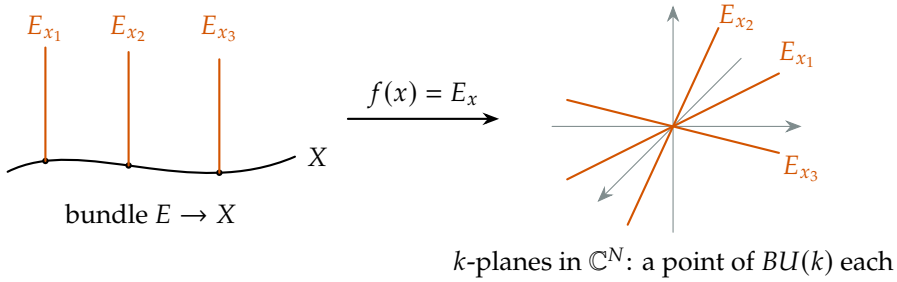


Figure 3.57: The classifying map. Once E sits inside a trivial bundle $X \times \mathbb{C}^N$, every fibre E_x is a k-plane in $\mathbb{C}^N$, that is, a point of the Grassmannian. The map f sends x to that point. The universal bundle has the plane itself as its fibre, so pulling it back along f gives E back.

Two facts make this useful. We state them and explain why they are plausible.

Key idea: Classification of bundles

1. Every rank-k complex bundle over X is $f^*E_{U(k)}$ for some map $f : X \to BU(k)$.
2. $f^*E_{U(k)} \cong g^*E_{U(k)}$ if and only if f and g are homotopic.

So isomorphism classes of rank-k bundles are the same thing as homotopy classes of maps: $\mathrm{Vect}_k(X, \mathbb{C}) \cong [X, BU(k)]$.

Part 1 is the construction above. For part 2, a homotopy f_t is a family of classifying maps, that is, a bundle over $X \times [0,1]$; bundles over a cylinder are "constant in the t direction" (one can parallel transport along t with any connection), so the two ends are isomorphic. The infinite size of $BU(k)$ is what makes the choices of N and of embedding irrelevant: any two embeddings can be rotated into each other inside a big enough $\mathbb{C}^N$.

$$\begin{array}{ccc} E = f^*E_{U(k)} & \longrightarrow & E_{U(k)} \\ \downarrow & & \downarrow \\ X & \xrightarrow{\ f\ } & BU(k) \end{array} \qquad \begin{array}{ccc} H^{2i}(X;\mathbb{Z}) & \xleftarrow{\ f^*\ } & H^{2i}(BU(k);\mathbb{Z}) \\ \Psi & & \Psi \\ c_i(E) = f^*c_i & \longleftarrow\!\shortmid & c_i \end{array}$$

Figure 3.58: Left: every bundle is a pullback of the universal bundle. Right: cohomology goes the other way. The universal Chern classes c_i living on $BU(k)$ are pulled back by f^* to classes on X. Homotopic maps induce the same f^*, so the result depends only on the isomorphism class of E.

3.15.2 Chern classes defined by pulling back

The cohomology of the universal space is known exactly. We state the result and then make it believable for $k = 1$.

Key idea: Cohomology of $BU(k)$

$H^*(BU(k);\mathbb{Z}) = \mathbb{Z}[c_1, \dots, c_k]$, a polynomial ring on k generators with $c_i \in H^{2i}(BU(k);\mathbb{Z})$. There is nothing in odd degrees, and no relations between the generators.

For $k = 1$, $BU(1) = \mathbb{CP}^\infty$. Build $\mathbb{CP}^n$ by induction: the points $[z_0 : \cdots : z_n]$ with $z_n \neq 0$ can be written $[z_0/z_n : \cdots : 1]$, a copy of $\mathbb{C}^n = \mathbb{R}^{2n}$, an open $2n$-dimensional cell; the points with $z_n = 0$ form $\mathbb{CP}^{n-1}$. So

$$\mathbb{CP}^n = \mathrm{pt} \cup \mathbb{R}^2 \cup \mathbb{R}^4 \cup \cdots \cup \mathbb{R}^{2n},$$

one cell in each even dimension. With no odd cells, there is nothing for a coboundary map to do, and $H^{2j}(\mathbb{CP}^n; \mathbb{Z}) = \mathbb{Z}$ for $0 \le j \le n$. The generator of H^2 is called x (it will be $-c_1$ of the tautological bundle), and Poincaré duality on $\mathbb{CP}^n$ shows that x^j generates H^{2j}. Letting $n \to \infty$ gives $H^*(\mathbb{CP}^\infty; \mathbb{Z}) = \mathbb{Z}[x]$.

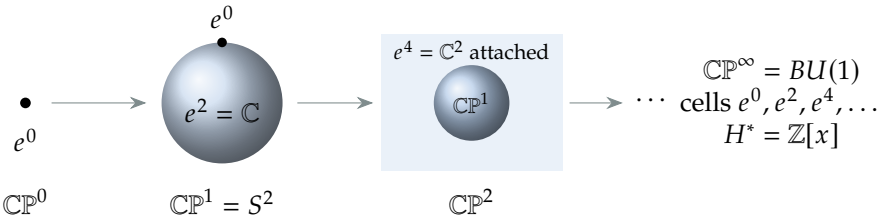


Figure 3.59: The cell structure of complex projective space: one cell in each even dimension and none in odd dimensions. This is why $BU(1) = \mathbb{CP}^\infty$ has cohomology $\mathbb{Z}$ in every even degree, generated by powers of a single degree-2 class. That class, pulled back, is the first Chern class.

Working definition 3.38 (Chern classes, topological version)

Let $E \to X$ be a complex vector bundle of rank k and $f : X \to BU(k)$ its classifying map [the map with $E \cong f^* E_{U(k)}$, unique up to homotopy]. The ***i*-th Chern class** of E is

$$c_i(E) = f^*\big(c_i(E_{U(k)})\big) \in H^{2i}(X; \mathbb{Z}), \qquad i = 1, \ldots, k. \tag{3.137}$$

[$c_i(E_{U(k)})$ is the degree-$2i$ generator of $H^*(BU(k); \mathbb{Z}) = \mathbb{Z}[c_1, \ldots, c_k]$] We set $c_0(E) = 1$ and $c_i(E) = 0$ for $i > k$. *Operationally:* you never compute f. You use the properties below (naturality, the Whitney formula, normalisation), or the curvature formula of Section 3.16, which produces a closed differential form representing $c_i(E)$ in de Rham cohomology.

Three properties follow almost immediately, and together they pin the classes down completely.

1. **Naturality.** If $g : Y \to X$ is a map, then $c_i(g^* E) = g^* c_i(E)$. Reason: the classifying map of $g^* E$ is $f \circ g$, and $(f \circ g)^* = g^* f^*$.

2. **Whitney sum formula.** $c(E \oplus F) = c(E)\, c(F)$ for the total class defined next. (We will see it with curvature in Section 3.16: the curvature of a direct sum is block diagonal, and a determinant of a block-diagonal matrix factorises.)

3. **Normalisation.** For the tautological line bundle γ over $\mathbb{CP}^1$, $\int_{\mathbb{CP}^1} c_1(\gamma) = -1$, where $\mathbb{CP}^1$ carries its complex orientation. (Some books normalise the sign the other way; ours agrees with the curvature formula $c_1 = \frac{i}{2\pi} \operatorname{tr} F$ and with Milnor and Stasheff. We check the sign by a calculation in Example 3.92.)

Working definition 3.39 (Total Chern class)

The **total Chern class** of a rank-k bundle is the inhomogeneous [a sum of pieces of different degrees $0, 2, \ldots, 2k$] cohomology class

$$c(E) = 1 + c_1(E) + c_2(E) + \cdots + c_k(E) \ \in \ H^{\text{even}}(X; \mathbb{Z}). \tag{3.138}$$

It is a bookkeeping device: the Whitney formula $c(E \oplus F) = c(E)c(F)$ unpacks, degree by degree, into

$$c_n(E \oplus F) = \sum_{i+j=n} c_i(E)\, c_j(F).$$

Worked example 3.87: The trivial bundle has $c = 1$

Let $E = X \times \mathbb{C}^k = I^k$. Its classifying map can be chosen constant: every fibre is the same plane $\mathbb{C}^k \subset \mathbb{C}^N$. A constant map factors through a point, $X \to \{\mathrm{pt}\} \to BU(k)$. Then for $i \geq 1$

$$\begin{aligned} c_i(I^k) &\overset{(1)}{=} f^* c_i \\ &\overset{(2)}{=} (\mathrm{const})^* \left(c_i\big|_{\mathrm{pt}}\right) \\ &\overset{(3)}{=} 0. \end{aligned}$$

(1) Definition (3.137).
(2) The constant map factors through a point, and pullback is functorial.
(3) $H^{2i}(\mathrm{pt};\mathbb{Z}) = 0$ for $i \geq 1$: a point has no cohomology above degree zero.

So $c(I^k) = 1$. Combined with Whitney: $c(E \oplus I^j) = c(E) \cdot 1 = c(E)$. *Chern classes cannot see trivial summands*, which is exactly the behaviour of K-theory's stable equivalence. Stably equivalent bundles have the same Chern classes.

3.15.3 Line bundles over the sphere: c_1 is the clutching degree

Let us see what the definition says in the first nontrivial case, complex line bundles over S^2. The previous part classified bundles over spheres by **clutching**: cover S^2 by a northern disc U_N and a southern disc U_S, trivialise the bundle over each (discs are contractible, so every bundle is trivial over them), and record how the two trivialisations are glued on the overlap, a band around the equator. For a line bundle, the gluing is a function

$$g : \text{ equator} \cong S^1 \longrightarrow U(1), \qquad \phi \mapsto g(\phi),$$

and only its winding number $n \in \pi_1(U(1)) = \mathbb{Z}$ matters. Write L_n for the bundle glued with a map of winding $-n$, for instance $g(\phi) = e^{-in\phi}$. (The sign is chosen so that L_n will have $c_1 = n$ below; with this choice L_{-1} is the tautological bundle, which we will check.)

Worked example 3.88: Chern classes of $L_m \oplus L_n$ over S^2

Let $u \in H^2(S^2;\mathbb{Z})$ be the generator with $\int_{S^2} u = 1$. Anticipating the calculation of Example 3.91, $c(L_n) = 1 + nu$. Then

$$\begin{aligned} c(L_m \oplus L_n) &\overset{(1)}{=} (1 + mu)(1 + nu) \\ &= 1 + (m+n)u + mn\, u^2 \\ &\overset{(2)}{=} 1 + (m+n)u. \end{aligned}$$

(1) The Whitney sum formula.
(2) $u^2 \in H^4(S^2;\mathbb{Z}) = 0$: a 2-dimensional space has no cohomology in degree 4.

So $c_1(L_m \oplus L_n) = (m+n)u$ and $c_2 = 0$. This matches the clutching picture: a rank-2 bundle over S^2 is glued by a loop in $U(2)$, and $\pi_1(U(2)) = \mathbb{Z}$ is detected by the determinant,

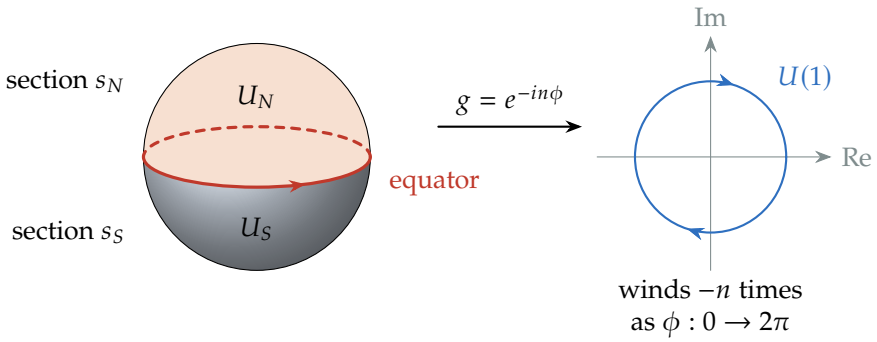


Figure 3.60: Clutching a line bundle over S^2. The bundle is trivial over each hemisphere. All the twisting is recorded by the gluing function on the equator, a loop in $U(1)$. Its winding number is the only invariant, and it is the first Chern class (up to the sign convention stated in the text).

$\det\big(\mathrm{diag}(e^{-im\phi}, e^{-in\phi})\big) = e^{-i(m+n)\phi}$, of winding $-(m+n)$. In particular $L_m \oplus L_n \cong L_{m+n} \oplus I$, the same total winding. On S^2 the first Chern class is a *complete* invariant of complex bundles, rank by rank.

Worked example 3.89: c_2 lives on four-dimensional spaces

Why do we need higher Chern classes at all? Because $c_i \in H^{2i}(X)$, and on a space of real dimension d only c_i with $2i \le d$ can be nonzero. On S^2 only c_1 survives (Example 3.88). On S^4, $H^2(S^4) = 0$, so $c_1 = 0$ for every bundle, and the only possible invariant is $c_2 \in H^4(S^4;\mathbb{Z}) = \mathbb{Z}$. Clutching agrees: bundles over S^4 are glued along the equatorial S^3 by maps $S^3 \to \mathrm{SU}(2)$, classified by $\pi_3(\mathrm{SU}(2)) = \mathbb{Z}$. We will show in Example 3.93 that c_2 equals this winding number. It is the instanton number.

Dictionary: Characteristic classes and charges

Mathematics	Physics
c_1 of a $U(1)$ bundle over S^2	magnetic charge of a Dirac monopole
c_2 of an SU(2) bundle over S^4	instanton number (topological charge)
clutching function on the equator	large gauge transformation between patches
classifying map $X \to BU(k)$	"which vacuum/gauge sector we are in"
$c(E \oplus I) = c(E)$	adding a free, untwisted field changes no charge

Watch out: Characteristic classes do not classify

Equal Chern classes do not imply isomorphic bundles. Two examples. (i) Over S^5, rank-2 complex bundles are clutched along the equatorial S^4 by maps $S^4 \to U(2)$, and $\pi_4(U(2)) = \pi_4(S^3) = \mathbb{Z}_2$, so there is a nontrivial rank-2 bundle over S^5 (the previous part met it as a stably trivial but nontrivial bundle). Yet all its Chern classes vanish, simply because $H^2(S^5) = H^4(S^5) = 0$. (ii) The real tangent bundle TS^2 is not trivial (hairy-ball theorem), but all its Pontryagin classes vanish (Example 3.114); only the Euler class of Section 3.20 sees it. The classes are *obstructions*: nonzero means twisted; zero means "no information from this test". For *line* bundles c_1 is always complete (line bundles over X correspond one-to-one to $H^2(X;\mathbb{Z})$, because $\mathbb{CP}^\infty$ has only one nonzero homotopy group, $\pi_2 = \mathbb{Z}$), and over S^2 it is complete in every rank. Beyond that, completeness is luck of low dimension.

Watch out: Integer classes versus differential forms

The topological definition gives classes in $H^{2i}(X;\mathbb{Z})$. Differential forms only see $H^{2i}(X;\mathbb{R})$, which forgets *torsion*: elements t with $mt = 0$ for some integer m. On all the manifolds we compute with (spheres, tori, $\mathbb{CP}^n$, K3) the integer cohomology has no torsion, and nothing is lost. When Nash writes "modulo elements of order 2" (Section 3.19.6), he is warning about precisely this.

Reading Nash (pp. 78–79, (3.91)–(3.92))

Nash states the definition in one display (3.91): for a rank-k bundle E with classifying map $f : X \to BU(k)$ and $E = f^*E_{U(k)}$, $H^*(BU(k);\mathbb{Z})$ has even generators $c_i(E_{U(k)}) \in H^{2i}(BU(k);\mathbb{Z})$, and $c_i(E) = f^*(c_i(E_{U(k)})) \in H^{2i}(X;\mathbb{Z})$. Then (3.92) introduces the total class $c(E) = 1 + c_1(E) + \cdots + c_k(E)$ and states $c(E \oplus F) = c(E)c(F)$. His $\mathrm{Vect}(X,\mathbb{C})$ is the set of complex vector bundles over X. The classifying spaces themselves were introduced in his Chapter 1, §5 and in (3.66)–(3.67) as infinite Grassmannians $BU(k) = \mathrm{Gr}(k,\infty,\mathbb{C})$.

Recap

- Every rank-k bundle is the pullback $f^*E_{U(k)}$ of the universal bundle over $BU(k) = \mathrm{Gr}(k,\infty,\mathbb{C})$; homotopic f give isomorphic bundles.
- $H^*(BU(k);\mathbb{Z}) = \mathbb{Z}[c_1,\dots,c_k]$; pulling back defines $c_i(E) \in H^{2i}(X;\mathbb{Z})$.
- Properties: naturality, Whitney $c(E \oplus F) = c(E)c(F)$, normalisation $\int_{\mathbb{CP}^1} c_1(\gamma) = -1$, and $c(\text{trivial}) = 1$.
- Over S^2: c_1 = winding of the clutching function; over S^4: c_2 = winding of a map $S^3 \to \mathrm{SU}(2)$.
- Characteristic classes detect twisting; they do not always classify.

3.16 Chern–Weil theory: characteristic classes from curvature

The definition of the last section is clean but useless for computation: nobody writes down classifying maps. Nash closes his section (p. 87) by recalling that characteristic classes "are often calculated using curvature methods", and in this book those are the methods we will actually use, for anomalies, instantons and index densities. So we develop them now, in full.

The puzzle to keep in mind is this. Curvature is *local*: it is a field strength, it depends on the connection, and it changes from point to point. A characteristic class is *global and rigid*: it is an integer that does not change under deformations. How can the one produce the other? The answer has two halves.

1. Suitable polynomials in the curvature are *closed* forms, so they define de Rham cohomology classes.

2. Changing the connection changes those forms by an *exact* form, so the cohomology class, and hence every integral over a closed cycle, does not change.

The exact form in half 2 has a name that every physicist knows: for $\operatorname{tr} F^2$ it is the *Chern–Simons form*.

3.16.1 Reminder: connections and curvature on a vector bundle

A full treatment is in Chapter T2 and Chapter T3; here is what we need, with the conventions fixed.

Let $E \to X$ be a complex vector bundle of rank k. Over an open set U choose a **local frame**: k sections $s = (s_1, \dots, s_k)$ that form a basis of each fibre. A **connection** ∇ differentiates sections. On the frame it is recorded by a $k \times k$ matrix $A = (A_{ij})$ of 1-forms:

$$\nabla s_j = \sum_i s_i \, A_{ij}, \qquad \text{or in matrix shorthand} \quad \nabla s = s\, A. \tag{3.139}$$

A general section $\psi = s\, u$ (with u a column of functions) then has $\nabla\psi = s\,(\mathrm{d}u + Au)$. This is the covariant derivative $D = \mathrm{d} + A$ of gauge theory. The **curvature** is the matrix of 2-forms

$$F = \mathrm{d}A + A \wedge A. \tag{3.140}$$

Conventions. If E carries a Hermitian metric and we use orthonormal frames and a metric-compatible connection, then A and F take values in $\mathfrak{u}(k)$, the *anti-Hermitian* matrices. This is Nash's convention and ours. Physicists usually write $D_\mu = \partial_\mu - i g\, A^a_\mu T^a$ with Hermitian generators T^a.

Dictionary: Anti-Hermitian (maths) versus Hermitian (physics) gauge fields

Maths / Nash	Physics
$A = A_\mu \,\mathrm{d}x^\mu$, $A_\mu^\dagger = -A_\mu$	$A_\mu = -ig\, A^a_\mu T^a$, T^a Hermitian, $\operatorname{tr} T^a T^b = \tfrac{1}{2}\delta^{ab}$
$F = \mathrm{d}A + A \wedge A$	$F = -ig\, F^a T^a$, $F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu + g f^{abc} A^b_\mu A^c_\nu$
$\operatorname{tr} F \wedge F$	$-\frac{g^2}{2} F^a \wedge F^a = -\frac{g^2}{4} F^a_{\mu\nu} \tilde{F}^{a\,\mu\nu}\, \mathrm{d}^4x$
iF Hermitian	field strength as an observable
$U(1)$: $A = -ieA^{\mathrm{phys}}$	$F = -ie\, F^{\mathrm{phys}}$, $F^{\mathrm{phys}} = \mathrm{d}A^{\mathrm{phys}}$

Here $\tilde{F}^{a\,\mu\nu} = \frac{1}{2}\varepsilon^{\mu\nu\rho\sigma} F^a_{\rho\sigma}$ and we used $F^a \wedge F^a = \frac{1}{2} F^a_{\mu\nu} \tilde{F}^{a\,\mu\nu} \mathrm{d}^4x$.

Changing the frame (a gauge transformation). On an overlap of two patches, or simply by choice, we may replace the frame s by $s' = s\, g$ with $g : U \to \mathrm{GL}(k, \mathbb{C})$ (or $U(k)$ for orthonormal frames). Then

$$\begin{aligned} \nabla s' &\overset{(1)}{=} \nabla(s)\, g + s\, \mathrm{d}g \\ &\overset{(2)}{=} s\, A\, g + s\, \mathrm{d}g \\ &\overset{(3)}{=} s' \left(g^{-1} A g + g^{-1} \mathrm{d}g \right), \end{aligned} \tag{3.141}$$

(1) The Leibniz rule for ∇; the entries of g are functions.
(2) The definition (3.139) of A.
(3) Substitute $s = s' g^{-1}$.

so $A' = g^{-1}Ag + g^{-1}\mathrm{d}g$, the familiar gauge transformation. The curvature transforms *homogeneously*:

$$\begin{aligned} F' &\overset{(1)}{=} \mathrm{d}\big(g^{-1}Ag + g^{-1}\mathrm{d}g\big) + \big(g^{-1}Ag + g^{-1}\mathrm{d}g\big) \wedge \big(g^{-1}Ag + g^{-1}\mathrm{d}g\big) \\ &\overset{(2)}{=} \big(-g^{-1}\mathrm{d}g\, g^{-1}Ag + g^{-1}\mathrm{d}A\, g - g^{-1}A\,\mathrm{d}g - g^{-1}\mathrm{d}g\, g^{-1}\mathrm{d}g\big) \\ &\qquad + \big(g^{-1}A^2 g + g^{-1}A\,\mathrm{d}g + g^{-1}\mathrm{d}g\, g^{-1}Ag + g^{-1}\mathrm{d}g\, g^{-1}\mathrm{d}g\big) \\ &\overset{(3)}{=} g^{-1}(\mathrm{d}A + A \wedge A)\, g = g^{-1}Fg. \end{aligned} \tag{3.142}$$

(1) Definition (3.140) applied to A'.
(2) $\mathrm{d}(g^{-1}) = -g^{-1}\mathrm{d}g\, g^{-1}$ (differentiate $g^{-1}g = 1$); moving d past the 1-form A costs a sign; expand the product.
(3) The three pairs of terms containing $\mathrm{d}g$ cancel in pairs.

This one line is the reason Chern–Weil theory works: F changes by *conjugation*, and there are functions of matrices that do not care about conjugation.

The Bianchi identity. Differentiating (3.140),

$$\begin{aligned} \mathrm{d}F &\overset{(1)}{=} \mathrm{d}A \wedge A - A \wedge \mathrm{d}A \\ &\overset{(2)}{=} (F - A^2) \wedge A - A \wedge (F - A^2) \\ &= F \wedge A - A \wedge F, \end{aligned} \tag{3.143}$$

(1) $\mathrm{d}^2 A = 0$, and $\mathrm{d}(A \wedge A) = \mathrm{d}A \wedge A - A \wedge \mathrm{d}A$ because A is a 1-form.
(2) Substitute $\mathrm{d}A = F - A^2$; the A^3 terms cancel.

In the language of covariant derivatives, $DF := \mathrm{d}F + A \wedge F - F \wedge A = 0$.

Graded cyclicity of the trace. We will move matrix-valued forms around inside traces all the time. If α is a matrix of p-forms and β a matrix of q-forms, then

$$\operatorname{tr}(\alpha \wedge \beta) = (-1)^{pq} \operatorname{tr}(\beta \wedge \alpha). \tag{3.144}$$

The ordinary cyclicity of the trace moves the matrices; moving a p-form past a q-form costs $(-1)^{pq}$. In particular 2-forms such as F can be moved freely, and $\operatorname{tr}(A \wedge A) = -\operatorname{tr}(A \wedge A) = 0$ for a 1-form A. A consequence we use repeatedly: for any matrix-valued p-form ω that transforms by conjugation, with $D\omega = \mathrm{d}\omega + A \wedge \omega - (-1)^p \omega \wedge A$,

$$\mathrm{d} \operatorname{tr} \omega = \operatorname{tr} D\omega, \tag{3.145}$$

because $\operatorname{tr}(A \wedge \omega) = (-1)^p \operatorname{tr}(\omega \wedge A)$ makes the extra terms cancel.

3.16.2 Invariant polynomials

Working definition 3.40 (Invariant polynomial)

A polynomial function P of the entries of a $k \times k$ matrix X is **invariant** if $P(g^{-1}Xg) = P(X)$ for every invertible g. *Test:* it is enough to check $P(XY) = P(YX)$ for all X, Y. The basic examples are the traces of powers, $\operatorname{tr} X^m$, and the coefficients $\sigma_j(X)$ [$\sigma_1 = \operatorname{tr} X$ and

$\sigma_k = \det X$] of the characteristic polynomial,

$$\det(t\,\mathbb{1} + X) = \sum_{j=0}^{k} t^{\,k-j}\,\sigma_j(X). \tag{3.146}$$

Every invariant polynomial is a polynomial in the σ_j (equivalently, in the $\operatorname{tr} X^m$, $m \le k$).

Why the σ_j are enough: an invariant polynomial is unchanged by diagonalising X (when X is diagonalisable), so it is a function of the eigenvalues $\lambda_1, \dots, \lambda_k$. It is also unchanged by permuting the eigenvalues (a permutation matrix is a conjugation). A symmetric polynomial in the λ's is a polynomial in the elementary symmetric ones, and $\sigma_j(X) = \sigma_j(\lambda_1, \dots, \lambda_k)$. Diagonalisable matrices are dense, so the conclusion holds for all X. We return to this argument in Section 3.17; it is exactly Nash's justification of the splitting principle on p. 88.

Worked example 3.90: The 2×2 case by hand

Let $X = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. Then

$$\begin{aligned} \det(t\mathbb{1} + X) &= (t + a)(t + d) - bc \\ &= t^2 + (a + d)\,t + (ad - bc), \end{aligned}$$

so $\sigma_1 = \operatorname{tr} X$ and $\sigma_2 = \det X$. Also

$$\begin{aligned} \tfrac{1}{2}\big[(\operatorname{tr} X)^2 - \operatorname{tr} X^2\big] &\overset{(1)}{=} \tfrac{1}{2}\big[(a + d)^2 - (a^2 + 2bc + d^2)\big] \\ &= ad - bc = \sigma_2. \end{aligned}$$

(1) X^2 has diagonal entries $a^2 + bc$ and $bc + d^2$.

This is the first Newton identity, $\sigma_2 = \frac{1}{2}(\sigma_1^2 - \operatorname{tr} X^2)$, which we will meet again in Section 3.17.3.

Plugging in curvature. Now put $X = \frac{i}{2\pi}F$. The entries of F are 2-forms, and 2-forms commute with each other under $\wedge$. So all the algebra of ordinary polynomials (in particular det, and its expansion) works verbatim with the entries of F, with products read as wedge products. Because $F' = g^{-1}Fg$, an invariant polynomial gives the *same* form in every frame:

$$P(F') = P(g^{-1}Fg) = P(F).$$

So $P(F)$ is a globally defined differential form on X, even though F itself is only defined patch by patch.

3.16.3 Half 1: invariant polynomials of F are closed

It suffices to treat $P = \operatorname{tr} F^m$, since the others are polynomials in these, and a product of closed forms is closed.

$$\begin{aligned} \mathrm{d}\operatorname{tr} F^m &\overset{(1)}{=} \sum_{r=0}^{m-1} \operatorname{tr}\big(F^r \wedge \mathrm{d}F \wedge F^{m-1-r}\big) \\ &\overset{(2)}{=} m\ \operatorname{tr}\big(F^{m-1} \wedge \mathrm{d}F\big) \\ &\overset{(3)}{=} m\ \operatorname{tr}\big(F^{m-1} \wedge (F \wedge A - A \wedge F)\big) \end{aligned}$$

$$
\begin{aligned}
&= m\ \mathrm{tr}\big(F^m \wedge A\big) - m\ \mathrm{tr}\big(F^{m-1} \wedge A \wedge F\big) \\
&\overset{(4)}{=} 0.
\end{aligned}
\tag{3.147}
$$

(1) Leibniz rule; each F is a 2-form, so no signs appear.
(2) Graded cyclicity (3.144): the 2-forms F^r can be cycled to the back without sign.
(3) The Bianchi identity (3.143).
(4) Cycle the final 2-form F in the second term to the front (no sign), giving $\mathrm{tr}(F^m \wedge A)$, which cancels the first term.

So $\mathrm{tr}\, F^m$ is closed, and it defines a class in $H^{2m}_{\mathrm{dR}}(X)$.

3.16.4 Half 2: the class does not depend on the connection

Take two connections ∇_0, ∇_1 on the same bundle, with matrices A_0, A_1 in a frame. Their difference $\eta = A_1 - A_0$ transforms homogeneously, $\eta' = g^{-1}\eta g$ (the inhomogeneous $g^{-1}\mathrm{d}g$ cancels in the difference), so η is a globally defined $\mathrm{End}(E)$-valued 1-form. Connections form an *affine space*: the straight line

$$
A_t = A_0 + t\,\eta, \qquad 0 \le t \le 1,
\tag{3.148}
$$

consists of connections, with curvatures $F_t = \mathrm{d}A_t + A_t^2$. (Physicists know this as the statement that the space $\mathcal{A}$ of gauge fields is contractible.) Differentiating in t,

$$
\frac{\mathrm{d}F_t}{\mathrm{d}t} = \mathrm{d}\eta + A_t \wedge \eta + \eta \wedge A_t = D_t\eta,
\tag{3.149}
$$

the covariant derivative of η with respect to A_t. Now watch $\mathrm{tr}\, F_t^m$ change:

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\,\mathrm{tr}\, F_t^m &\overset{(1)}{=} m\ \mathrm{tr}\Big(F_t^{m-1} \wedge \frac{\mathrm{d}F_t}{\mathrm{d}t}\Big) \\
&\overset{(2)}{=} m\ \mathrm{tr}\big(F_t^{m-1} \wedge D_t\eta\big) \\
&\overset{(3)}{=} m\ \mathrm{tr}\big(D_t(F_t^{m-1} \wedge \eta)\big) \\
&\overset{(4)}{=} \mathrm{d}\Big[m\ \mathrm{tr}\big(F_t^{m-1} \wedge \eta\big)\Big].
\end{aligned}
\tag{3.150}
$$

(1) Leibniz in t and cyclicity, exactly as in step (1)–(2) of (3.147).
(2) Equation (3.149).
(3) Leibniz for D_t: $D_t(F_t^{m-1} \wedge \eta) = D_t(F_t^{m-1}) \wedge \eta + F_t^{m-1} \wedge D_t\eta$, and $D_t F_t = 0$ by Bianchi, so the first term vanishes.
(4) Identity (3.145): the trace of a covariant derivative is an ordinary derivative of the trace.

Integrating from $t = 0$ to $t = 1$ gives the **transgression formula**

$$
\boxed{\mathrm{tr}\, F_1^m - \mathrm{tr}\, F_0^m = \mathrm{d}\, Q_m(A_1, A_0), \qquad Q_m(A_1, A_0) = m \int_0^1 \mathrm{tr}\big(F_t^{m-1} \wedge (A_1 - A_0)\big)\, \mathrm{d}t.}
\tag{3.151}
$$

Q_m is a globally defined $(2m-1)$-form (it is built from η and F_t, both of which transform by conjugation). So $\mathrm{tr}\, F_1^m$ and $\mathrm{tr}\, F_0^m$ differ by an exact form, and they define the same de Rham class.

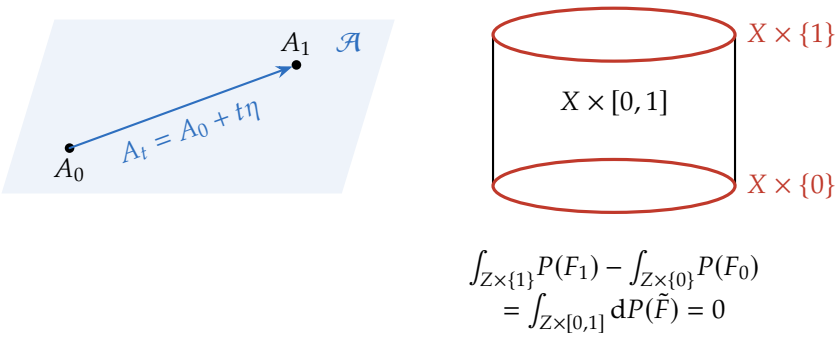


Figure 3.61: Two pictures of the same argument. Left: connections form an affine space, so any two can be joined by a straight line A_t, along which the Chern–Weil form changes by an exact form, $\mathrm{d}Q$. Right: equivalently, put the connection A_t on the bundle pulled back to the cylinder $X \times [0,1]$. For any closed cycle Z in X, Stokes' theorem on $Z \times [0,1]$ (whose boundary is the two ends) and closedness of $P(\tilde{F})$ show that the two ends give the same integral.

Key idea: The Chern–Weil theorem

For an invariant polynomial P and any connection on E:

1. $P(F)$ is a closed, globally defined form;
2. its de Rham class $[P(F)]$ is independent of the connection;
3. for $P(X) = \det\!\left(\mathbb{1} + \frac{i}{2\pi}X\right)$ the class is the image in real cohomology of the topological total Chern class $c(E)$ of Section 3.15. In particular all its periods (integrals over closed cycles) are integers.

Items 1 and 2 we have proved. Item 3 is plausible from two facts. The Chern–Weil classes are natural under pullback (the pullback connection has the pulled-back curvature), and they satisfy the Whitney formula (below). So they obey the same axioms as the topological classes, and axioms plus the normalisation on $\mathbb{CP}^1$ (checked in Example 3.92) determine characteristic classes uniquely, via the splitting principle of Section 3.17. The integrality we will see directly in every example: transgression pushes the integral onto an equator, where it becomes a winding number.

3.16.5 The Chern forms

Working definition 3.41 (Chern forms and Chern classes from curvature)

For a connection [a matrix A of 1-forms on each patch, the gauge potential] with curvature F [the matrix of 2-forms $F = \mathrm{d}A + A \wedge A$] on a rank-k complex bundle, define the **Chern forms** by expanding a determinant:

$$c(F) = \det\!\left(\mathbb{1} + \frac{i}{2\pi}F\right) = 1 + c_1(F) + c_2(F) + \cdots + c_k(F), \tag{3.152}$$

with $c_j(F)$ a closed $2j$-form. Their classes are the Chern classes, $c_j(E) = [c_j(F)] \in H^{2j}_{\mathrm{dR}}(X)$ [closed $2j$-forms modulo exact ones]. *Operationally:* pick any connection (the simplest you can find), compute F, expand the determinant, integrate.

Nash writes the same thing with a bookkeeping variable t (his (3.125)–(3.126)):

$$P(t) = \det\Big(t\,\mathbb{1} + \frac{iF}{2\pi}\Big) = \sum_{j=0}^{k} t^{\,j}\, c_{k-j}(F), \tag{3.153}$$

which is (3.146) with $X = iF/2\pi$; setting $t = 1$ recovers (3.152).

Let us extract the first two explicitly. Write $X = \frac{i}{2\pi}F$ and expand $\det(\mathbb{1} + X)$ to second order:

$$\begin{aligned} \det(\mathbb{1} + X) &\overset{(1)}{=} \exp \operatorname{tr} \log(\mathbb{1} + X) \\ &\overset{(2)}{=} \exp\big(\operatorname{tr} X - \tfrac{1}{2}\operatorname{tr} X^2 + \dots\big) \\ &\overset{(3)}{=} 1 + \operatorname{tr} X + \tfrac{1}{2}\big[(\operatorname{tr} X)^2 - \operatorname{tr} X^2\big] + \dots \end{aligned} \tag{3.154}$$

(1) $\det e^{Y} = e^{\operatorname{tr} Y}$ (true for diagonal Y, hence for all Y by conjugation-invariance and density); take $Y = \log(\mathbb{1} + X)$.
(2) Power series $\log(1+X) = X - \frac{1}{2}X^2 + \dots$; the entries of X are commuting 2-forms, so formal manipulations are legitimate and the series terminate.
(3) $e^{u} = 1 + u + \frac{1}{2}u^2 + \dots$ with $u = \operatorname{tr} X - \frac{1}{2}\operatorname{tr} X^2$, keeping terms of degree ≤ 4.

Hence

$$c_1(F) = \frac{i}{2\pi}\operatorname{tr} F, \tag{3.155}$$

$$c_2(F) = \frac{1}{2}\Big(\frac{i}{2\pi}\Big)^2\big[(\operatorname{tr} F)^2 - \operatorname{tr} F^2\big] = \frac{1}{8\pi^2}\big[\operatorname{tr}(F\wedge F) - \operatorname{tr} F \wedge \operatorname{tr} F\big]. \tag{3.156}$$

For an $\mathrm{SU}(k)$ bundle, $\operatorname{tr} F = 0$ ($\mathfrak{su}(k)$ is traceless), so $c_1(F) = 0$ and

$$c_2(F) = \frac{1}{8\pi^2}\operatorname{tr}(F\wedge F) \qquad (\mathrm{SU}(k)\text{ bundles}). \tag{3.157}$$

This is the instanton-number density.

Reality. For a unitary connection $F^\dagger = -F$, so iF is a Hermitian matrix of 2-forms and $\overline{\det(\mathbb{1} + \frac{i}{2\pi}F)} = \det(\mathbb{1} + \frac{i}{2\pi}F)^T = \det(\mathbb{1} + \frac{i}{2\pi}F)$. The Chern forms are *real* forms, as they must be to represent classes that are ultimately integers.

Whitney formula from curvature. For $E\oplus E'$ with the direct sum connection, $F_{E\oplus E'} = \begin{pmatrix} F_E & 0 \\ 0 & F_{E'} \end{pmatrix}$ and

$$\begin{aligned} c(F_{E\oplus E'}) &\overset{(1)}{=} \det\begin{pmatrix} \mathbb{1} + \frac{i}{2\pi}F_E & 0 \\ 0 & \mathbb{1} + \frac{i}{2\pi}F_{E'} \end{pmatrix} \\ &\overset{(2)}{=} \det\Big(\mathbb{1} + \frac{i}{2\pi}F_E\Big)\det\Big(\mathbb{1} + \frac{i}{2\pi}F_{E'}\Big) = c(F_E)\,c(F_{E'}). \end{aligned}$$

(1) The direct sum connection does not mix the two summands.
(2) The determinant of a block-diagonal matrix is the product of the blocks' determinants (the entries commute, so ordinary algebra applies).

The Chern–Simons form. Take $m = 2$ in the transgression formula, with $A_0 = 0$ (the trivial flat connection in a local trivialisation, so the formula holds on that patch) and $A_1 = A$. Then $\eta = A$ and $F_t = t\,\mathrm{d}A + t^2A^2$:

$$\begin{aligned} Q_2(A,0) &\overset{(1)}{=} 2\int_0^1 \operatorname{tr}\big((t\,\mathrm{d}A + t^2A^2)\wedge A\big)\,\mathrm{d}t \\ &\overset{(2)}{=} 2\Big(\tfrac{1}{2}\operatorname{tr}(\mathrm{d}A\wedge A) + \tfrac{1}{3}\operatorname{tr}(A^3)\Big) \\ &\overset{(3)}{=} \operatorname{tr}\Big(A\wedge \mathrm{d}A + \tfrac{2}{3}A\wedge A\wedge A\Big) \;:=\; \mathrm{CS}(A). \end{aligned} \tag{3.158}$$

(1) (3.151) with $m = 2$.
(2) $\int_0^1 t\,\mathrm{d}t = \frac{1}{2}$ and $\int_0^1 t^2\mathrm{d}t = \frac{1}{3}$.
(3) Graded cyclicity: $\operatorname{tr}(\mathrm{d}A\wedge A) = (-1)^{2\cdot 1}\operatorname{tr}(A\wedge \mathrm{d}A)$.

It is reassuring to check $\mathrm{d}\,\mathrm{CS} = \operatorname{tr}F^2$ directly:

$$\begin{aligned} \mathrm{d}\,\mathrm{CS}(A) &\overset{(1)}{=} \operatorname{tr}(\mathrm{d}A\wedge \mathrm{d}A) + \tfrac{2}{3}\cdot 3\,\operatorname{tr}(\mathrm{d}A\wedge A^2) \\ &\overset{(2)}{=} \operatorname{tr}\big(\mathrm{d}A\wedge \mathrm{d}A + \mathrm{d}A\wedge A^2 + A^2\wedge \mathrm{d}A + A^4\big) \\ &= \operatorname{tr}\big((\mathrm{d}A + A^2)\wedge(\mathrm{d}A + A^2)\big) = \operatorname{tr}F^2. \end{aligned} \tag{3.159}$$

(1) $\mathrm{d}\operatorname{tr}(A\,\mathrm{d}A) = \operatorname{tr}(\mathrm{d}A\,\mathrm{d}A)$; and $\mathrm{d}\operatorname{tr}A^3 = \operatorname{tr}(\mathrm{d}A\,A^2) - \operatorname{tr}(A\,\mathrm{d}A\,A) + \operatorname{tr}(A^2\mathrm{d}A) = 3\operatorname{tr}(\mathrm{d}A\,A^2)$, each term cycled by (3.144).
(2) Split $2\operatorname{tr}(\mathrm{d}A\,A^2) = \operatorname{tr}(\mathrm{d}A\,A^2) + \operatorname{tr}(A^2\,\mathrm{d}A)$, and add $\operatorname{tr}A^4 = 0$ (graded cyclicity gives $\operatorname{tr}(A\cdot A^3) = -\operatorname{tr}(A^3\cdot A)$).

Two limits are worth remembering. For $U(1)$ (or the trace part), $\operatorname{tr}F = \mathrm{d}\operatorname{tr}A$: the transgression form of c_1 is just the gauge field. For a *pure gauge* $A = g^{-1}\mathrm{d}g$ we have $F = 0$, so $\mathrm{d}A = -A^2$, and

$$\mathrm{CS}(g^{-1}\mathrm{d}g) = \operatorname{tr}\big(-A^3 + \tfrac{2}{3}A^3\big) = -\tfrac{1}{3}\operatorname{tr}\big(g^{-1}\mathrm{d}g\big)^3. \tag{3.160}$$

3.16.6 Worked examples

Worked example 3.91: The monopole bundles: $c_1 = n$

Cover S^2 by $U_N = \{\theta < \pi\}$ and $U_S = \{\theta > 0\}$. On each patch take a unitary frame and the $U(1)$ gauge fields

$$A_N = -\frac{in}{2}(1-\cos\theta)\,\mathrm{d}\phi, \qquad A_S = +\frac{in}{2}(1+\cos\theta)\,\mathrm{d}\phi.$$

Each is smooth on its own patch: $(1-\cos\theta)\,\mathrm{d}\phi$ vanishes like $\theta^2\mathrm{d}\phi$ at the north pole, where $\mathrm{d}\phi$ is singular, and similarly for A_S at the south pole. On the overlap

$$A_N - A_S = -in\,\mathrm{d}\phi = g^{-1}\mathrm{d}g \qquad \text{with} \qquad g = e^{-in\phi},$$

so by (3.141) the frames are related by $s_N = s_S\,e^{-in\phi}$: this is the bundle L_n of Section 3.15.3. The curvature is the same on both patches, $F = \mathrm{d}A_N = \mathrm{d}A_S = -\frac{in}{2}\sin\theta\,\mathrm{d}\theta\wedge\mathrm{d}\phi$. Then

$$\begin{aligned} \int_{S^2} c_1(F) &\overset{(1)}{=} \frac{i}{2\pi}\int_{S^2} F \\ &\overset{(2)}{=} \frac{i}{2\pi}\cdot\Big(-\frac{in}{2}\Big)\int_0^{2\pi}\!\!\int_0^{\pi} \sin\theta\,\mathrm{d}\theta\,\mathrm{d}\phi \end{aligned}$$

$$\overset{(3)}{=} \frac{n}{4\pi}\cdot 4\pi = n.$$

(1) (3.155) for a line bundle (tr of a 1×1 matrix).
(2) Substitute F; (θ,ϕ) cover the sphere once with the outward orientation $\mathrm{d}\theta\wedge\mathrm{d}\phi$.
(3) $i\cdot(-i)=1$ and the area of the unit sphere is 4π.

Now the same answer by transgression, which shows *why* it is an integer. Split the sphere along the equator C and use $F=\mathrm{d}A_N$ on the north, $F=\mathrm{d}A_S$ on the south:

$$\begin{aligned}\frac{i}{2\pi}\int_{S^2}F &\overset{(1)}{=} \frac{i}{2\pi}\Big(\int_{\text{north}}\mathrm{d}A_N+\int_{\text{south}}\mathrm{d}A_S\Big)\\ &\overset{(2)}{=} \frac{i}{2\pi}\Big(\oint_C A_N-\oint_C A_S\Big)\\ &\overset{(3)}{=} \frac{i}{2\pi}\oint_C g^{-1}\mathrm{d}g \overset{(4)}{=} \frac{i}{2\pi}(-in)(2\pi)=n.\end{aligned}$$

(1) On each hemisphere F is exact, with its own potential.
(2) Stokes. C with increasing ϕ is the boundary of the north hemisphere and, with the opposite orientation, of the south one.
(3) $A_N-A_S=g^{-1}\mathrm{d}g$ on the overlap.
(4) $g^{-1}\mathrm{d}g=-in\,\mathrm{d}\phi$ and $\oint\mathrm{d}\phi=2\pi$.

The magnetic charge is the winding number of the gauge transformation between patches: $\frac{i}{2\pi}\oint g^{-1}\mathrm{d}g=-\frac{1}{2\pi i}\oint\frac{\mathrm{d}g}{g}=-\mathrm{wind}(g)$. It is an integer because g is single valued. In physical units ($A=-ieA^{\mathrm{phys}}$) this is Dirac's quantisation: the flux $\int F^{\mathrm{phys}}=2\pi n/e$.

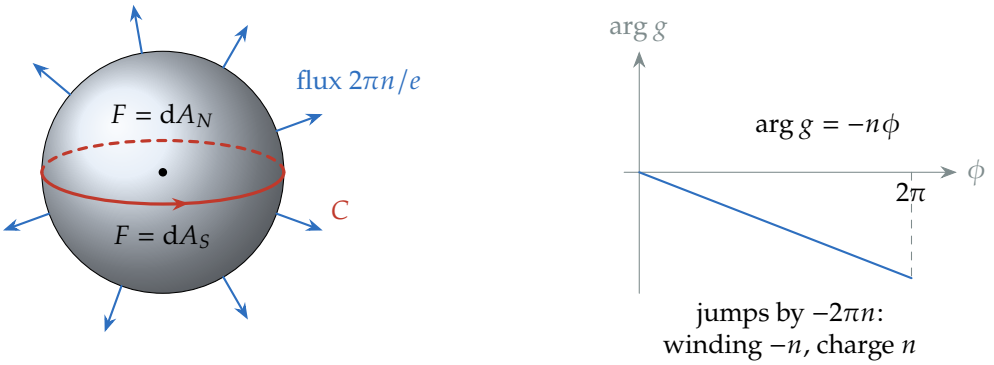


Figure 3.62: The monopole as a bundle. Left: the field strength is exact on each hemisphere separately, with different potentials; all the flux is "paid for" on the equator C. Right: the gauge transformation $g=e^{-in\phi}$ relating the two potentials on C. Its phase must return to itself modulo 2π, so the winding, and hence the flux, is quantised.

Worked example 3.92: The tautological (Hopf) line bundle over $\mathbb{CP}^1$ has $c_1=-1$

The tautological bundle $\gamma\to\mathbb{CP}^1$ has as fibre over the line $\ell=[z_0:z_1]$ the line $\ell\subset\mathbb{C}^2$ itself. On the patch $z_0\neq 0$ use the affine coordinate $z=z_1/z_0$, and the nowhere-zero section

$$s(z)=(1,z)\in\ell.$$

It inherits the length $h(z)=\|s\|^2=1+|z|^2$ from $\mathbb{C}^2$. The natural connection is orthogonal projection of the ordinary derivative back onto the line: $\nabla s=\frac{\langle s,\mathrm{d}s\rangle}{\langle s,s\rangle}s$, with $\langle u,v\rangle=$

$\bar{u}_0 v_0 + \bar{u}_1 v_1$. Since $\mathrm{d}s = (0, \mathrm{d}z)$,

$$A = \frac{\bar{z}\,\mathrm{d}z}{1+|z|^2} \qquad \text{(in the frame } s\text{)}.$$

This frame is not unitary. Normalise: $\hat{s} = s/\sqrt{h}$. By (3.141) with $g = h^{-1/2}$,

$$\begin{aligned}\hat{A} &\overset{(1)}{=} A + h^{1/2}\,\mathrm{d}\big(h^{-1/2}\big) = A - \tfrac{1}{2}\mathrm{d}\log h \\ &\overset{(2)}{=} \frac{\bar{z}\,\mathrm{d}z}{1+|z|^2} - \frac{\bar{z}\,\mathrm{d}z + z\,\mathrm{d}\bar{z}}{2(1+|z|^2)} = \frac{\bar{z}\,\mathrm{d}z - z\,\mathrm{d}\bar{z}}{2(1+|z|^2)} \\ &\overset{(3)}{=} \frac{i\,r^2\,\mathrm{d}\phi}{1+r^2}.\end{aligned}$$

(1) Abelian gauge transformation; $h^{1/2}\mathrm{d}h^{-1/2} = -\frac{1}{2}\mathrm{d}\log h$.
(2) $\mathrm{d}h = \bar{z}\,\mathrm{d}z + z\,\mathrm{d}\bar{z}$.
(3) Polar coordinates $z = re^{i\phi}$: $\bar{z}\,\mathrm{d}z - z\,\mathrm{d}\bar{z} = 2ir^2\mathrm{d}\phi$.

$\hat{A}$ is imaginary, i.e. anti-Hermitian, as it should be for a unitary frame. Its curvature is $F = \mathrm{d}\hat{A} = \frac{2ir}{(1+r^2)^2}\,\mathrm{d}r \wedge \mathrm{d}\phi$, and

$$\begin{aligned}\int_{\mathbb{CP}^1} c_1(F) &\overset{(1)}{=} \frac{i}{2\pi}\int_0^{2\pi}\int_0^{\infty} \frac{2ir}{(1+r^2)^2}\,\mathrm{d}r\,\mathrm{d}\phi \\ &\overset{(2)}{=} -\frac{1}{2\pi}\cdot 2\pi\cdot\left[-\frac{1}{1+r^2}\right]_0^{\infty} \\ &= -1.\end{aligned}$$

(1) The patch $z_0 \neq 0$ is all of $\mathbb{CP}^1$ except one point, which has zero area; $\mathrm{d}r \wedge \mathrm{d}\phi$ is the complex orientation ($\mathrm{d}x \wedge \mathrm{d}y = r\,\mathrm{d}r \wedge \mathrm{d}\phi$).
(2) $i \cdot 2i = -2$; $\int_0^{\infty} \frac{2r\,\mathrm{d}r}{(1+r^2)^2} = 1$.

In spherical coordinates, via stereographic projection $r = \tan(\theta/2)$, $\frac{r^2}{1+r^2} = \sin^2\frac{\theta}{2} = \frac{1}{2}(1-\cos\theta)$, so $\hat{A} = \frac{i}{2}(1-\cos\theta)\,\mathrm{d}\phi$. This is A_N of Example 3.91 with $n = -1$. *The tautological bundle is the monopole bundle L_{-1}*, confirming the normalisation we announced in Section 3.15.2. Its dual, the hyperplane bundle $\gamma^* = L_1$, has $c_1 = +1$ (Section 3.19.1).

Physical picture: Where the Hopf fibration comes in

The unit vectors in the fibres of γ form the unit sphere $S^3 = \{|z_0|^2 + |z_1|^2 = 1\} \subset \mathbb{C}^2$, and forgetting the phase gives the Hopf map $S^3 \to \mathbb{CP}^1 = S^2$. The connection of Example 3.92 comes from the single 1-form $a = \bar{z}_0\mathrm{d}z_0 + \bar{z}_1\mathrm{d}z_1$ on S^3. (It is imaginary on S^3 because $\mathrm{d}(|z_0|^2 + |z_1|^2) = a + \bar{a} = 0$ there.) Pulling a back by the local section $z \mapsto (1, z)/\sqrt{1+|z|^2}$ gives exactly $\hat{A}$. So the monopole of charge -1 *is* the Hopf fibration with its round connection: the fibre circles are the Dirac-string-free description of the monopole's $U(1)$ phase.

Worked example 3.93: An SU(2) bundle over S^4: c_2 is the winding number

Identify $S^4 = \mathbb{R}^4 \cup \{\infty\}$. A connection on an SU(2) bundle over S^4 can be described in a single gauge on $\mathbb{R}^4$, provided that as $|x| \to \infty$ it approaches a pure gauge, $A \to g^{-1}\mathrm{d}g$, for a map $g : S^3_\infty \to \mathrm{SU}(2)$. (That map is the clutching function to the patch around ∞.) The

standard choice, used by Belavin, Polyakov, Schwartz and Tyupkin, is

$$g(x) = \frac{x_4 \mathbb{1} + i\, x_k \sigma_k}{|x|}, \qquad A = f(r)\, g^{-1} \mathrm{d} g, \qquad f(r) = \frac{r^2}{r^2 + \lambda^2},$$

with σ_k the Pauli matrices, $r = |x|$ and $\lambda > 0$ the instanton size. The matrix g is in SU(2): its determinant is $x_4^2 + x_1^2 + x_2^2 + x_3^2 = r^2$ divided by r^2, and it is unitary because it is a unit quaternion. We orient $\mathbb{R}^4$ by $\mathrm{d}x_1 \wedge \mathrm{d}x_2 \wedge \mathrm{d}x_3 \wedge \mathrm{d}x_4$ and the sphere S^3_R of radius R as the boundary of the ball B_R (outward normal first).

Step 1: reduce to the sphere at infinity. By (3.157), the transgression (3.159) and Stokes,

$$\int_{B_R} c_2(F) \overset{(1)}{=} \frac{1}{8\pi^2} \int_{B_R} \operatorname{tr} F^2 \overset{(2)}{=} \frac{1}{8\pi^2} \int_{S^3_R} \mathrm{CS}(A).$$

(1) SU(2) bundle, so $c_1(F) = 0$ and (3.157) applies.

(2) A is a single smooth 1-form on all of $\mathbb{R}^4$ (at $r \to 0$, $f \sim r^2/\lambda^2$ cancels the $1/r$ of $g^{-1}\mathrm{d}g$), so $\operatorname{tr} F^2 = \mathrm{d}\,\mathrm{CS}(A)$ globally on B_R.

Step 2: the Chern–Simons form of $A = f\theta$, where $\theta = g^{-1}\mathrm{d}g$ satisfies $\mathrm{d}\theta = -\theta^2$ and f depends only on r, so on S^3_R it is a constant:

$$\begin{aligned} \mathrm{CS}(f\theta)\big|_{S^3_R} &\overset{(1)}{=} \operatorname{tr}\!\left(f\theta \wedge f\,\mathrm{d}\theta + \tfrac{2}{3} f^3 \theta^3\right) \\ &\overset{(2)}{=} \left(-f^2 + \tfrac{2}{3} f^3\right) \operatorname{tr} \theta^3. \end{aligned}$$

(1) Definition (3.158); $\mathrm{d}f = 0$ along the sphere.

(2) $\mathrm{d}\theta = -\theta^2$ (differentiate $g^{-1}g = \mathbb{1}$).

Step 3: compute $\operatorname{tr}\theta^3$. At the "north pole" $p = (0,0,0,R)$ we have $g(p) = \mathbb{1}$ and, along the sphere, $\mathrm{d}x_4 = 0$ and $\mathrm{d}r = 0$ to first order, so $\theta = \frac{i}{R}\sigma_k \mathrm{d}x_k$. Then

$$\begin{aligned} \operatorname{tr}\theta^3\big|_p &\overset{(1)}{=} \frac{i^3}{R^3} \operatorname{tr}(\sigma_a\sigma_b\sigma_c)\, \mathrm{d}x_a \wedge \mathrm{d}x_b \wedge \mathrm{d}x_c \\ &\overset{(2)}{=} \frac{-i}{R^3}\, \varepsilon_{abc} \operatorname{tr}(\sigma_a\sigma_b\sigma_c)\, \mathrm{d}x_1 \wedge \mathrm{d}x_2 \wedge \mathrm{d}x_3 \\ &\overset{(3)}{=} \frac{-i}{R^3}\,(12i)\, \mathrm{d}x_1 \wedge \mathrm{d}x_2 \wedge \mathrm{d}x_3 = \frac{12}{R^3}\, \mathrm{d}x_1 \wedge \mathrm{d}x_2 \wedge \mathrm{d}x_3. \end{aligned}$$

(1) Substitute θ and pull the constants out.

(2) $\mathrm{d}x_a \wedge \mathrm{d}x_b \wedge \mathrm{d}x_c = \varepsilon_{abc}\, \mathrm{d}x_1 \wedge \mathrm{d}x_2 \wedge \mathrm{d}x_3$.

(3) $\operatorname{tr}(\sigma_a\sigma_b\sigma_c) = 2i\varepsilon_{abc}$ and $\varepsilon_{abc}\varepsilon_{abc} = 6$.

Now two observations finish the job. (i) θ is unchanged under $g \mapsto hg$ for a constant $h \in \mathrm{SU}(2)$, and these left multiplications move p to any point of S^3_R while preserving the round volume form; so $\operatorname{tr}\theta^3$ is a *constant* multiple of the volume form of S^3_R. (ii) At p the outward normal is e_4, and (e_4, e_1, e_2, e_3) is an odd permutation of (e_1, e_2, e_3, e_4); so the boundary orientation at p is $-\mathrm{d}x_1 \wedge \mathrm{d}x_2 \wedge \mathrm{d}x_3$, and $\mathrm{d}x_1 \wedge \mathrm{d}x_2 \wedge \mathrm{d}x_3 = -R^3\, \mathrm{vol}_{S^3}$ in terms of the unit-sphere volume element. Therefore

$$\int_{S^3_R} \operatorname{tr}\theta^3 \overset{(1)}{=} -12\,\mathrm{Vol}(S^3) \overset{(2)}{=} -12 \cdot 2\pi^2 = -24\pi^2.$$

(1) Observations (i) and (ii); the factors of R cancel.

(2) The unit 3-sphere has volume $2\pi^2$.

Step 4: assemble. Combining steps 1–3,

$$\int_{B_R} c_2(F) = \frac{1}{8\pi^2}\left(-f^2 + \tfrac{2}{3}f^3\right)(-24\pi^2) = 3f(R)^2 - 2f(R)^3,$$

and as $R \to \infty$, $f \to 1$:

$$\boxed{\int_{S^4} c_2 = \frac{1}{8\pi^2}\int_{\mathbb{R}^4} \operatorname{tr} F \wedge F = 1.}$$

The profile f only affects *where* the charge sits (Figure 3.63); the total is fixed by g alone, as Chern–Weil theory promises. With $A \to g^{-1}\mathrm{d}g$ at infinity for a general g, the same steps give

$$\int_{S^4} c_2 = -\frac{1}{24\pi^2}\int_{S^3} \operatorname{tr}\left(g^{-1}\mathrm{d}g\right)^3 := w(g) \in \mathbb{Z},$$

the **winding number** (degree) of $g : S^3 \to \mathrm{SU}(2) \cong S^3$. For our g (the identity map of the unit quaternions) $w = 1$; for g^m one finds $w = m$.

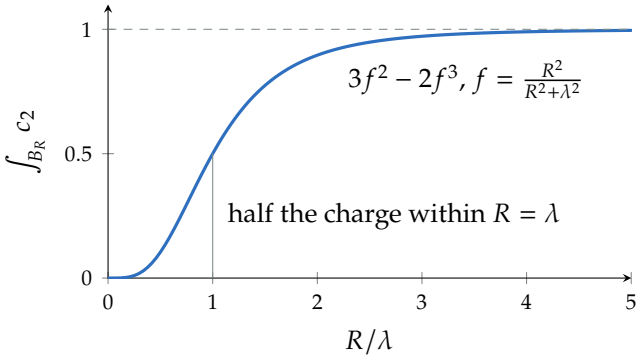


Figure 3.63: How much instanton number lies inside the ball of radius R for the BPST connection. It rises from 0 to exactly 1; the size λ sets the scale (at $R = \lambda$, $f = \frac{1}{2}$ and the enclosed charge is $\frac{1}{2}$). Changing f changes the curve but never its limit: the total is the winding number of g on the sphere at infinity.

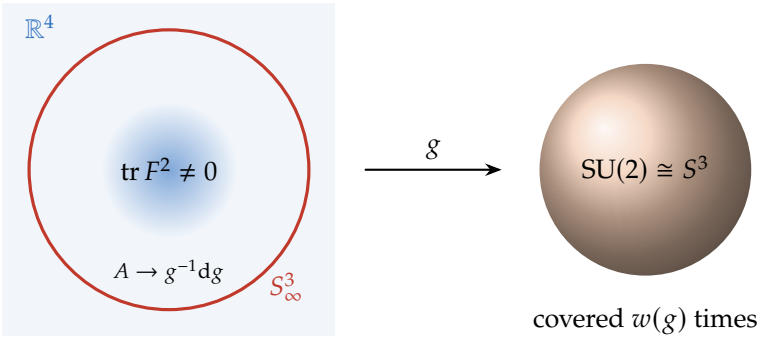


Figure 3.64: The instanton number as a winding number. The curvature is concentrated in a region of size λ; far away the field is pure gauge, $A \approx g^{-1}\mathrm{d}g$, and $\operatorname{tr} F^2 = \mathrm{d}\,\mathrm{CS}$ turns the 4-volume integral into a surface integral over the sphere at infinity. There $\mathrm{CS} = -\frac{1}{3}\operatorname{tr}(g^{-1}\mathrm{d}g)^3$ counts how many times g wraps S^3_∞ around the group manifold $\mathrm{SU}(2) \cong S^3$.

Watch out: Signs of c_2 and of "the" instanton number

With our conventions (anti-Hermitian F, $c_2 = \frac{1}{8\pi^2}\operatorname{tr} F\wedge F$, standard orientation of $\mathbb{R}^4$), write $F = F^a T^a$ with $T^a = -\frac{i}{2}\sigma_a$, so $\operatorname{tr} F\wedge F = -\frac{1}{2}F^a\wedge F^a$. A self-dual field ($\star F = F$) has $F^a\wedge F^a = |F^a|^2\,\mathrm{vol} \geq 0$, hence $c_2 \leq 0$; the BPST field above, with $c_2 = +1$, is therefore *anti*-self-dual in this orientation. This is the convention of Donaldson theory (anti-self-dual connections on bundles with $c_2 = k > 0$). Physics texts usually define the topological charge $Q = \frac{g^2}{32\pi^2}\int F^a_{\mu\nu}\tilde F^{a\,\mu\nu}\mathrm{d}^4x$; by the dictionary above, $Q = -c_2[S^4]$. Nothing physical depends on the choice, but always check which one a formula uses.

Worked example 3.94: The Chern–Simons form under a gauge transformation

$\mathrm{CS}(A)$ is *not* gauge invariant, and how it fails is physics. Let $A' = g^{-1}(A+\rho)g$ with $\rho = \mathrm{d}g\,g^{-1}$ (this is the same as $g^{-1}Ag + g^{-1}\mathrm{d}g$). Write $\mathrm{CS}(A) = \operatorname{tr}(A\wedge F - \frac{1}{3}A^3)$, which equals (3.158) because $F = \mathrm{d}A + A^2$. Then

$$\begin{aligned}
\mathrm{CS}(A') &\overset{(1)}{=} \operatorname{tr}\big((A+\rho)F\big) - \tfrac{1}{3}\operatorname{tr}(A+\rho)^3\\
&\overset{(2)}{=} \mathrm{CS}(A) + \operatorname{tr}(\rho F) - \operatorname{tr}(A^2\rho) - \operatorname{tr}(A\rho^2) - \tfrac{1}{3}\operatorname{tr}\rho^3\\
&\overset{(3)}{=} \mathrm{CS}(A) + \operatorname{tr}(\rho\,\mathrm{d}A) - \operatorname{tr}(A\,\mathrm{d}\rho) - \tfrac{1}{3}\operatorname{tr}\rho^3\\
&\overset{(4)}{=} \mathrm{CS}(A) - \mathrm{d}\operatorname{tr}(\rho\wedge A) - \tfrac{1}{3}\operatorname{tr}\big(g^{-1}\mathrm{d}g\big)^3.
\end{aligned}$$

(1) $F' = g^{-1}Fg$, $A'^3 = g^{-1}(A+\rho)^3 g$, and the trace kills the conjugation.
(2) Expand the cube; by graded cyclicity the three orderings of A, A, ρ have equal traces, as do those of A, ρ, ρ.
(3) $\operatorname{tr}(\rho A^2)$ cancels $\operatorname{tr}(A^2\rho)$; and $\mathrm{d}\rho = \rho^2$ (from $\mathrm{d}(g^{-1}) = -g^{-1}\mathrm{d}g\,g^{-1}$), so $\operatorname{tr}(A\rho^2) = \operatorname{tr}(A\,\mathrm{d}\rho)$.
(4) $\mathrm{d}\operatorname{tr}(\rho A) = \operatorname{tr}(\mathrm{d}\rho\,A) - \operatorname{tr}(\rho\,\mathrm{d}A)$ and $\operatorname{tr}(\mathrm{d}\rho\,A) = \operatorname{tr}(A\,\mathrm{d}\rho)$; also $\operatorname{tr}\rho^3 = \operatorname{tr}(g\theta g^{-1})^3 = \operatorname{tr}\theta^3$.

On a closed 3-manifold M the exact term integrates to zero, and

$$\int_M \mathrm{CS}(A') - \int_M \mathrm{CS}(A) = -\frac{1}{3}\int_M \operatorname{tr}(g^{-1}\mathrm{d}g)^3 = 8\pi^2\, w(g).$$

So the Chern–Simons action $S = \frac{\kappa}{8\pi^2}\cdot 2\pi\int_M \mathrm{CS}(A) = \frac{\kappa}{4\pi}\int_M \mathrm{CS}(A)$ shifts by $2\pi\kappa\, w(g)$, and e^{iS} is gauge invariant only if the level κ is an integer. This is the level quantisation of Chern–Simons theory, which we meet again in Chapter 12; the "secondary" nature of CS is the subject of Nash's Chapter 8, Chapter 8.

Try it: Abelian Chern–Simons and linking

For $U(1)$, $\mathrm{CS}(A) = A\wedge\mathrm{d}A$. Show that under $A\to A+\mathrm{d}\chi$ it changes by the exact form $\mathrm{d}(\chi\,\mathrm{d}A)$, so on a closed 3-manifold it is gauge invariant for single-valued χ.
Answer: $(A+\mathrm{d}\chi)\wedge\mathrm{d}A = A\wedge\mathrm{d}A + \mathrm{d}\chi\wedge\mathrm{d}A = A\wedge\mathrm{d}A + \mathrm{d}(\chi\,\mathrm{d}A)$.

Reading Nash (pp. 87–88, (3.125)–(3.127))

Nash introduces curvature methods only at the end of his §5: F is a curvature 2-form of a rank-k bundle with group $U(k)$, $O(k)$ or $\mathrm{SO}(k)$, and $P(t) = \det(t + iF/2\pi)$ (3.125) is a polynomial in t whose coefficients are closed forms, representing the characteristic

classes; for complex bundles $\det(t + iF/2\pi) = \sum_j t^j c_{k-j}(F)$ (3.126). He does not prove closedness or independence of the connection, and refers to Nash and Sen. We have supplied both (Sections 3.16.3 and 3.16.4), with the explicit Chern–Simons form. His (3.127), $\mathrm{ch}(F) = \operatorname{tr}\exp(iF/2\pi)$, is in Section 3.18.

Recap

- Under a gauge transformation $F \mapsto g^{-1}Fg$, so conjugation-invariant polynomials of F are global forms.
- Bianchi $\Rightarrow \mathrm{d}\operatorname{tr} F^m = 0$. Transgression $\Rightarrow \operatorname{tr} F_1^m - \operatorname{tr} F_0^m = \mathrm{d}Q_m$: the class does not depend on the connection.
- $c(F) = \det(1 + \frac{i}{2\pi}F)$; $c_1 = \frac{i}{2\pi}\operatorname{tr} F$; $c_2 = \frac{1}{8\pi^2}(\operatorname{tr} F^2 - \operatorname{tr} F \wedge \operatorname{tr} F)$.
- $\operatorname{tr} F^2 = \mathrm{d}\,\mathrm{CS}$, $\mathrm{CS} = \operatorname{tr}(A\mathrm{d}A + \frac{2}{3}A^3)$; for pure gauge $\mathrm{CS} = -\frac{1}{3}\operatorname{tr}(g^{-1}\mathrm{d}g)^3$.
- Monopole: c_1 = minus the winding of the patching function; tautological bundle has $c_1 = -1$. SU(2) on S^4: c_2 = winding of $g : S^3 \to \mathrm{SU}(2)$; BPST has $c_2 = 1$.

3.17 The splitting principle: pretend every bundle is a sum of lines

The Whitney formula tells us the Chern classes of a *direct sum*. But physics hands us other combinations all the time. A quark field in the fundamental representation coupled to a $U(1)$ charge lives in $E \otimes L$. Gluons live in the adjoint bundle $\mathrm{End}_0 E$. Antiparticles live in the conjugate bundle. Fermion bilinears live in $\Lambda^2 E$. We want the Chern classes of all of these in terms of those of E and L. Is there a single trick that handles every such question?

There is, and it is the linear algebra trick every physicist uses without thinking: *diagonalise*. If E were a sum of line bundles, $E = L_1 \oplus \cdots \oplus L_k$, then everything would be easy, because line bundles are described by one number each (their c_1), tensor products of lines add those numbers, and Whitney handles the sum. The splitting principle says we may always compute *as if* E split this way.

3.17.1 Chern roots

Suppose first that E really is a sum of line bundles. Write $x_i = c_1(L_i) \in H^2(X;\mathbb{Z})$. For a line bundle $c(L_i) = 1 + x_i$, since a rank-1 bundle has only c_0 and c_1. Then

$$\begin{aligned} c(E) &\overset{(1)}{=} c(L_1)\,c(L_2)\cdots c(L_k) \\ &\overset{(2)}{=} \prod_{i=1}^{k}(1 + x_i) \\ &\overset{(3)}{=} 1 + \sum_i x_i + \sum_{i<j} x_i x_j + \cdots + x_1 x_2 \cdots x_k. \end{aligned} \tag{3.161}$$

(1) The Whitney sum formula, applied $k - 1$ times.
(2) $c(L_i) = 1 + c_1(L_i)$ for a line bundle.
(3) Multiply out. Cohomology classes of even degree commute, so the product behaves like an ordinary polynomial.

Comparing degrees with $c(E) = 1 + c_1(E) + \cdots + c_k(E)$ gives

$$\begin{aligned} c_1(E) &= \sum_i x_i, \qquad c_2(E) = \sum_{i<j} x_i x_j, \qquad \dots \\ c_j(E) &= \sum_{i_1<\cdots<i_j} x_{i_1}\cdots x_{i_j}, \qquad \dots, \qquad c_k(E) = x_1 \cdots x_k. \end{aligned} \tag{3.162}$$

These are the **elementary symmetric functions** $\sigma_j(x_1, \dots, x_k)$ of the x_i. Nash's (3.93) and (3.94) are exactly (3.161) and (3.162).

Working definition 3.42 (Chern roots)

For any complex bundle E of rank k, the **Chern roots** $x_1, \dots, x_k$ are formal degree-2 classes [symbols that need not be classes on X itself. Only their symmetric polynomials are] defined by the single rule

$$c(E) = \prod_{i=1}^{k} (1 + x_i),$$

that is, $c_j(E) = \sigma_j(x_1, \dots, x_k)$. *Operationally:* to compute any characteristic class of anything built from E, (1) pretend $E = \bigoplus L_i$ with $c_1(L_i) = x_i$; (2) work out the answer as a polynomial in the x_i; (3) check it is symmetric in the x_i [unchanged by every permutation of $x_1, \dots, x_k$]; (4) rewrite it in terms of the $\sigma_j = c_j(E)$. The result is correct for every E, split or not.

This is exactly like the characteristic polynomial of a matrix. The coefficients of $\det(t + X)$ are honest functions of the matrix entries; the eigenvalues are convenient fictions (they may not even be real or distinct), but any *symmetric* function of the eigenvalues is a polynomial in the coefficients and therefore honest too. The Chern classes are the coefficients; the Chern roots are the eigenvalues. In Section 3.17.5 this analogy becomes literal: the Chern roots are the eigenvalues of $\frac{i}{2\pi}F$.

Worked example 3.95: Rank 2 and rank 3 written out

For rank 2,

$$(1 + x_1)(1 + x_2) = 1 + (x_1 + x_2) + x_1 x_2,$$

so $c_1 = x_1 + x_2$ and $c_2 = x_1 x_2$. For rank 3,

$$\begin{aligned} (1 + x_1)(1 + x_2)(1 + x_3) &\overset{(1)}{=} \big(1 + (x_1 + x_2) + x_1 x_2\big)(1 + x_3) \\ &= 1 + (x_1 + x_2 + x_3) + (x_1 x_2 + x_1 x_3 + x_2 x_3) + x_1 x_2 x_3, \end{aligned}$$

(1) The rank-2 result, then multiply by $1 + x_3$ and collect by degree.

so $c_1 = \sigma_1$, $c_2 = \sigma_2$, $c_3 = \sigma_3$, each a sum over subsets of $\{1, 2, 3\}$ of the corresponding size.

3.17.2 Why the principle is true: the flag bundle

Of course most bundles do *not* split. The tangent bundle of S^2 is not a sum of two real line bundles, and the instanton bundle over S^4 is not a sum of two complex line bundles (if it were, $c_2 = x_1 x_2$ with $x_i \in H^2(S^4) = 0$ would vanish). The trick is to *move to a bigger base space* on which the pulled-back bundle does split, and to lose no information in the process.

Step 1: split off one line. Given $E \to X$ of rank k, form the **projectivisation** $\mathbb{P}(E)$: a point of $\mathbb{P}(E)$ is a pair (x, ℓ) with $x \in X$ and ℓ a complex line in the fibre E_x. The map $\pi : \mathbb{P}(E) \to X$, $(x, \ell) \mapsto x$, has fibre $\mathbb{CP}^{k-1}$ (the lines in $E_x \cong \mathbb{C}^k$). Pull E back to $\mathbb{P}(E)$: the fibre of π^*E over (x, ℓ) is E_x, and it contains a *preferred* line, namely ℓ itself. These preferred lines form a line bundle $\gamma_E \subset \pi^*E$, the tautological bundle again. With a Hermitian metric,

$$\pi^*E = \gamma_E \oplus \gamma_E^{\perp},$$

where $\gamma_E^{\perp}$ has rank $k-1$.

Step 2: repeat. Apply step 1 to $\gamma_E^{\perp}$ over $\mathbb{P}(E)$, and so on. After $k-1$ steps we reach a space $\mathrm{Fl}(E)$, the **flag bundle**, whose points are $x \in X$ together with an ordered splitting of E_x into k orthogonal lines. Over it the pullback splits completely:

$$\pi^*E = L_1 \oplus L_2 \oplus \cdots \oplus L_k. \tag{3.163}$$

Step 3: nothing is lost. The key topological fact is that $\pi^* : H^*(X) \to H^*(\mathrm{Fl}(E))$ is *injective*. For one step, the reason is the **Leray–Hirsch theorem**: the class $y = c_1(\gamma_E^*)$ restricts on each fibre $\mathbb{CP}^{k-1}$ to the generator of $H^2(\mathbb{CP}^{k-1})$, whose powers $1, y, \dots, y^{k-1}$ give all the cohomology of the fibre (Figure 3.59). So $H^*(\mathbb{P}(E))$ is a free module over $H^*(X)$ with basis $1, y, \dots, y^{k-1}$, exactly as the cohomology of a product $X \times \mathbb{CP}^{k-1}$ would be. In particular a class $a \in H^*(X)$ with $\pi^*a = 0$ must vanish, because $\pi^*a = a \cdot 1$ is a coefficient in that basis.

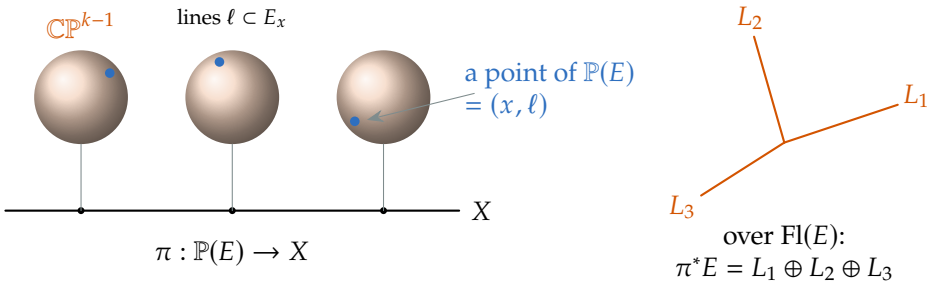


Figure 3.65: The splitting construction. Left: over each point x of the base, the projectivisation $\mathbb{P}(E)$ has a whole $\mathbb{CP}^{k-1}$ of points, one for each line ℓ in the fibre E_x. Pulled back to $\mathbb{P}(E)$, the bundle contains the chosen line as a sub-line-bundle. Right: repeating the construction (the flag bundle) chooses a complete orthogonal frame of lines, so the pulled-back bundle splits. The cohomology of X injects into that of the bigger space, so an identity checked upstairs holds downstairs.

Key idea: The splitting principle

For every complex vector bundle $E \to X$ there is a space Y and a map $\pi : Y \to X$ such that

1. π^*E is a direct sum of line bundles, and
2. $\pi^* : H^*(X; \mathbb{Z}) \to H^*(Y; \mathbb{Z})$ is injective.

Consequently, a polynomial identity among characteristic classes which is natural under pullback and holds for all sums of line bundles holds for all bundles.

The last sentence is a two-line argument. Let $P(E)$ and $Q(E)$ be two natural expressions in

characteristic classes that agree for split bundles. Then

$$\pi^*\big(P(E)-Q(E)\big) \overset{(1)}{=} P(\pi^*E)-Q(\pi^*E) \overset{(2)}{=} 0,$$

(1) Naturality: characteristic classes commute with pullback.
(2) π^*E splits, and $P = Q$ on split bundles.

and injectivity of π^* gives $P(E) = Q(E)$.

Watch out: The splitting is a device, not a fact about E

Upstairs on $\mathrm{Fl}(E)$ the classes $x_i = c_1(L_i)$ really exist. Downstairs on X they do *not*: only their symmetric combinations descend. So never write "x_1" as a class on X unless the bundle genuinely splits. An expression like $x_1 - x_2$ is meaningless downstairs; $(x_1-x_2)^2 = c_1^2 - 4c_2$ is fine.

3.17.3 Power sums and the Newton identities

The elementary symmetric functions are one natural basis of the symmetric polynomials. Another is the **power sums**

$$s_m(x_1,\dots,x_k) = x_1^m + x_2^m + \cdots + x_k^m.$$

They matter because the Chern character (Section 3.18) is built from them, and because in curvature language they are the simplest invariants, $s_m \leftrightarrow \mathrm{tr}\big(\frac{i}{2\pi}F\big)^m$. The dictionary between the two bases is the set of Newton identities. Here is a clean way to derive them all at once. Take the generating function $\prod_i(1+x_it)$ and its logarithm:

$$\log\sum_{j\ge 0}\sigma_j\,t^j \overset{(1)}{=} \log\prod_{i=1}^{k}(1+x_it) \overset{(2)}{=} \sum_{i=1}^{k}\log(1+x_it)$$
$$\overset{(3)}{=} \sum_{i=1}^{k}\sum_{m\ge 1}\frac{(-1)^{m-1}}{m}x_i^m t^m \overset{(4)}{=} \sum_{m\ge 1}\frac{(-1)^{m-1}}{m}s_m\,t^m. \tag{3.164}$$

(1) The product expands to $\sum_j \sigma_j t^j$, as in (3.161).
(2) The log of a product is the sum of logs.
(3) Taylor series $\log(1+u) = u - \frac{u^2}{2} + \frac{u^3}{3} - \dots$.
(4) Exchange the sums and collect $\sum_i x_i^m = s_m$.

Now expand the left side in powers of t, using $\log(1+u) = u - \frac{1}{2}u^2 + \frac{1}{3}u^3 - \dots$ with $u = \sigma_1 t + \sigma_2 t^2 + \sigma_3 t^3 + \dots$, and compare coefficients:

$$\begin{aligned} t^1:&\quad s_1 = \sigma_1,\\ t^2:&\quad -\tfrac{1}{2}s_2 = \sigma_2 - \tfrac{1}{2}\sigma_1^2 \;\Rightarrow\; s_2 = \sigma_1^2 - 2\sigma_2,\\ t^3:&\quad \tfrac{1}{3}s_3 = \sigma_3 - \sigma_1\sigma_2 + \tfrac{1}{3}\sigma_1^3 \;\Rightarrow\; s_3 = \sigma_1^3 - 3\sigma_1\sigma_2 + 3\sigma_3. \end{aligned}$$

The t^3 coefficient of $u - \frac{1}{2}u^2 + \frac{1}{3}u^3$ collects σ_3 from u, $-\frac{1}{2}\cdot 2\sigma_1\sigma_2$ from u^2 and $\frac{1}{3}\sigma_1^3$ from u^3. Going one order further (or differentiating (3.164) in t to get the recursion $s_m = \sigma_1 s_{m-1} - \sigma_2 s_{m-2} + \cdots + (-1)^{m-1} m\,\sigma_m$),

$$s_4 = \sigma_1^4 - 4\sigma_1^2\sigma_2 + 2\sigma_2^2 + 4\sigma_1\sigma_3 - 4\sigma_4. \tag{3.165}$$

Worked example 3.96: Checking s_2 and s_3 on numbers

Take $k = 3$ and $(x_1, x_2, x_3) = (1, 2, 3)$. Then $\sigma_1 = 6$, $\sigma_2 = 2 + 3 + 6 = 11$, $\sigma_3 = 6$. Directly $s_2 = 1 + 4 + 9 = 14$ and $s_3 = 1 + 8 + 27 = 36$. From the identities,

$$\sigma_1^2 - 2\sigma_2 = 36 - 22 = 14,$$
$$\sigma_1^3 - 3\sigma_1\sigma_2 + 3\sigma_3 = 216 - 198 + 18 = 36.$$

Both agree. (For a matrix with eigenvalues $1, 2, 3$, these say $\operatorname{tr} X^2 = (\operatorname{tr} X)^2 - 2\sigma_2(X)$ and similarly for $\operatorname{tr} X^3$.)

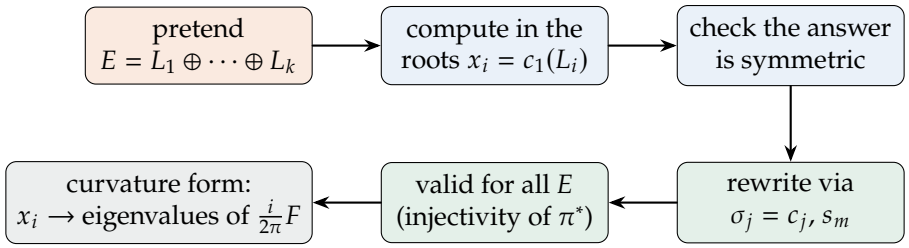


Figure 3.66: The splitting-principle recipe as a flowchart. Every computation in the rest of this chapter follows these arrows: work with the fictitious roots, confirm that only symmetric combinations appear, translate back to Chern classes, and (if a differential form is wanted) replace the roots by the eigenvalues of the curvature.

3.17.4 Recipes: duals, tensor products, exterior powers, $\mathbb{CP}^n$

To use Chern roots for tensor constructions we need one fact about line bundles: c_1 *turns tensor products into sums*, $c_1(L \otimes M) = c_1(L) + c_1(M)$. With connections this is one line. Give L and M connections with curvatures F_L, F_M (imaginary 2-forms). The tensor product connection $\nabla(s \otimes t) = \nabla s \otimes t + s \otimes \nabla t$ has $A_{L\otimes M} = A_L + A_M$, hence $F_{L\otimes M} = F_L + F_M$ (abelian, so no $A \wedge A$ terms), and c_1 is linear in F. In the clutching picture, transition functions multiply, $g_{L\otimes M} = g_L g_M$, and winding numbers add. In particular $L \otimes L^*$ is trivial, so $c_1(L^*) = -c_1(L)$.

Now the recipes. In each, E has roots x_i; if $E = \bigoplus L_i$, then the construction is again a sum of lines, whose first Chern classes we read off.

Worked example 3.97: The dual bundle

If $E = \bigoplus L_i$ then $E^* = \bigoplus L_i^*$, with roots $-x_i$. So

$$c(E^*) \overset{(1)}{=} \prod_i (1 - x_i)$$
$$\overset{(2)}{=} 1 - \sigma_1 + \sigma_2 - \cdots + (-1)^k \sigma_k,$$

(1) $c_1(L_i^*) = -c_1(L_i)$ and Whitney.

(2) Replacing $x_i \to -x_i$ multiplies a degree-j symmetric function by $(-1)^j$.

that is, $c_j(E^*) = (-1)^j c_j(E)$. Complex conjugation gives the same answer, since $\bar{E} \cong E^*$ for a unitary bundle (Section 3.19.1). *Physics:* antiparticles carry the opposite charges.

Worked example 3.98: Twisting by a line bundle

Let L have $c_1(L) = y$. If $E = \bigoplus L_i$ then $E \otimes L = \bigoplus (L_i \otimes L)$, with roots $x_i + y$. For rank 2,

$$\begin{aligned} c(E \otimes L) &\overset{(1)}{=} (1 + x_1 + y)(1 + x_2 + y) \\ &= 1 + (x_1 + x_2 + 2y) + \big(x_1 x_2 + (x_1 + x_2)y + y^2\big) \\ &\overset{(2)}{=} 1 + \big(c_1 + 2y\big) + \big(c_2 + c_1 y + y^2\big). \end{aligned}$$

(1) Roots of $E \otimes L$ are $x_i + y$.
(2) $x_1 + x_2 = c_1(E)$ and $x_1 x_2 = c_2(E)$.

In general $c_1(E \otimes L) = c_1(E) + k\,y$. A charged multiplet of k fields each carrying charge y has total "charge" ky added to its own.

Worked example 3.99: The adjoint bundle $\mathrm{End}\,E$ and the $\mathrm{SO}(3)$ instanton number

$\mathrm{End}\,E = E \otimes E^*$ has roots $x_i - x_j$ for all pairs (i, j), including the k zero roots $i = j$. For rank 2 the roots are $0, 0, x_1 - x_2, x_2 - x_1$, so

$$\begin{aligned} c(\mathrm{End}\,E) &\overset{(1)}{=} (1 + 0)^2\,(1 + x_1 - x_2)(1 - x_1 + x_2) \\ &= 1 - (x_1 - x_2)^2 \\ &\overset{(2)}{=} 1 - \big((x_1 + x_2)^2 - 4x_1 x_2\big) = 1 + \big(4c_2 - c_1^2\big). \end{aligned}$$

(1) Whitney over the four roots.
(2) $(x_1 - x_2)^2 = (x_1 + x_2)^2 - 4x_1 x_2$, which is symmetric, as it must be.

So $c_1(\mathrm{End}\,E) = 0$ and $c_2(\mathrm{End}\,E) = 4c_2(E) - c_1(E)^2$. For an SU(2) bundle, $c_1 = 0$ and $c_2(\mathrm{End}\,E) = 4c_2(E)$. The gluons of an SU(2) gauge theory live in the traceless part $\mathrm{End}_0\,E = \mathfrak{su}(2) \otimes \mathbb{C}$, and the trivial trace part does not change c. So in the adjoint representation the instanton number looks four times bigger: the familiar factor relating $\mathrm{tr}_{\mathrm{adj}}$ and $\mathrm{tr}_{\mathrm{fund}}$, $\mathrm{tr}_{\mathrm{adj}}\,F^2 = 4\,\mathrm{tr}_{\mathrm{fund}}\,F^2$ for SU(2) (that is, $2N$ with $N = 2$).

Worked example 3.100: Exterior powers and the determinant line

$\Lambda^p E$ has roots $x_{i_1} + \cdots + x_{i_p}$, one for each subset $i_1 < \cdots < i_p$. Summing all roots, each x_i appears in $\binom{k-1}{p-1}$ subsets, so

$$c_1(\Lambda^p E) = \binom{k-1}{p-1} c_1(E).$$

Two special cases: the **determinant line bundle** $\det E = \Lambda^k E$ has the single root $x_1 + \cdots + x_k$, hence $c_1(\det E) = c_1(E)$; and for rank 3, $\Lambda^2 E$ has roots $x_1 + x_2$, $x_1 + x_3$, $x_2 + x_3$, with $c_1(\Lambda^2 E) = 2c_1(E)$ and

$$\begin{aligned} c_2(\Lambda^2 E) &\overset{(1)}{=} (x_1 + x_2)(x_1 + x_3) + (x_1 + x_2)(x_2 + x_3) + (x_1 + x_3)(x_2 + x_3) \\ &\overset{(2)}{=} \sigma_1^2 + \sigma_2 = c_1^2 + c_2. \end{aligned}$$

(1) σ_2 of the three roots.
(2) Expanding gives $x_1^2 + x_2^2 + x_3^2 + 3(x_1 x_2 + x_1 x_3 + x_2 x_3) = s_2 + 3\sigma_2 = \sigma_1^2 - 2\sigma_2 + 3\sigma_2$.

Why $c_1(\det E) = c_1(E)$ matters: the determinant line of a family of Dirac operators is where the global anomaly lives (Chapter 10), and its first Chern class is computed from the index bundle by exactly this rule.

Worked example 3.101: The tangent bundle of $\mathbb{CP}^n$

Let $\gamma \to \mathbb{CP}^n$ be the tautological line bundle and $h = c_1(\gamma^*)$, the generator of $H^2(\mathbb{CP}^n;\mathbb{Z})$ with $\int_{\mathbb{CP}^1} h = 1$ on a line $\mathbb{CP}^1 \subset \mathbb{CP}^n$ (Example 3.92). A tangent vector at the point $\ell \in \mathbb{CP}^n$ is an infinitesimal tilt of the line ℓ inside $\mathbb{C}^{n+1}$, that is, a linear map $\ell \to \ell^\perp$:

$$T_\ell \mathbb{CP}^n \cong \mathrm{Hom}(\ell, \ell^\perp).$$

Add the trivial line bundle $\mathrm{Hom}(\gamma,\gamma) = \gamma^* \otimes \gamma \cong I$ (it has the nowhere-zero section id):

$$\begin{aligned} T\mathbb{CP}^n \oplus I &\overset{(1)}{=} \mathrm{Hom}(\gamma, \gamma^\perp) \oplus \mathrm{Hom}(\gamma,\gamma) \\ &\overset{(2)}{=} \mathrm{Hom}(\gamma, \mathbb{C}^{n+1}) \\ &\overset{(3)}{=} \gamma^* \oplus \cdots \oplus \gamma^* \qquad (n+1 \text{ copies}). \end{aligned}$$

(1) The identification above, plus the trivial line.
(2) $\gamma^\perp \oplus \gamma = \mathbb{C}^{n+1}$ fibrewise.
(3) A linear map from ℓ to $\mathbb{C}^{n+1}$ is $n+1$ linear functionals on ℓ, one per coordinate.

This is the **Euler sequence**. By Whitney and $c(I) = 1$,

$$c(T\mathbb{CP}^n) = (1+h)^{n+1} = \sum_{j=0}^{n} \binom{n+1}{j} h^j. \tag{3.166}$$

Checks: for $n = 1$, $c_1(T\mathbb{CP}^1) = 2h$, and $\int_{\mathbb{CP}^1} c_1 = 2 = \chi(S^2)$ (so TS^2, viewed as a complex line bundle, is the monopole bundle L_2). For $n = 2$, $c_1 = 3h$, $c_2 = 3h^2$ and $\int_{\mathbb{CP}^2} c_2 = 3 = \chi(\mathbb{CP}^2)$, the number of cells in Figure 3.59. That the top Chern class of the tangent bundle integrates to the Euler characteristic is the Gauss–Bonnet–Chern theorem of Section 3.20.

Try it: Symmetric square

Show that for rank 2, $\mathrm{Sym}^2 E$ (roots $2x_1$, x_1+x_2, $2x_2$) has $c_1 = 3c_1(E)$ and $c_2 = 2c_1^2 + 4c_2$.
Answer: $c_1 = 3(x_1+x_2)$. $c_2 = 2x_1(x_1+x_2) + 4x_1x_2 + 2x_2(x_1+x_2) = 2(x_1+x_2)^2 + 4x_1x_2$.

3.17.5 The curvature version: roots are eigenvalues of $\frac{i}{2\pi}F$

Nash ends his §5 with the observation that every splitting-principle formula has a curvature counterpart, obtained by a rule of substitution. Here is why.

For a unitary connection, $\frac{i}{2\pi}F$ is at each point a Hermitian matrix whose entries are commuting 2-forms. Pretend for a moment that it is diagonal,

$$\frac{i}{2\pi}F = \begin{pmatrix} \frac{i}{2\pi}F_1 & & \\ & \ddots & \\ & & \frac{i}{2\pi}F_k \end{pmatrix}. \tag{3.167}$$

This is exactly the curvature of a direct sum of line bundles with curvatures F_i, each contributing the Chern form $\frac{i}{2\pi}F_i$. So in this case

$$\det\Big(\mathbb{1} + \frac{i}{2\pi}F\Big) \overset{(1)}{=} \prod_{i=1}^{k}\Big(1 + \frac{i}{2\pi}F_i\Big), \qquad \mathrm{tr}\Big(\frac{i}{2\pi}F\Big)^m \overset{(2)}{=} \sum_{i=1}^{k}\Big(\frac{i}{2\pi}F_i\Big)^m,$$

(1) The determinant of a diagonal matrix is the product of its entries.
(2) Powers and traces of a diagonal matrix act entry by entry.

which are (3.161) and the power sums with $x_i \to \frac{i}{2\pi}F_i$. The general rule is:

Key idea: From roots to curvature

Let $f(x_1, \dots, x_k)$ be a symmetric (formal power series) expression in the Chern roots. Its curvature representative is obtained by replacing the x_i by the eigenvalues of $\frac{i}{2\pi}F$; equivalently, write f as a function of the matrix X that is conjugation invariant, and set $X = \frac{i}{2\pi}F$. Examples: $\prod(1 + x_i) \to \det(\mathbb{1} + \frac{i}{2\pi}F)$, $\sum x_i^m \to \operatorname{tr}(\frac{i}{2\pi}F)^m$, $\sum e^{x_i} \to \operatorname{tr}\exp(\frac{i}{2\pi}F)$.

Why may we pretend F is diagonal? Nash's argument (p. 88, following Eguchi, Gilkey and Hanson) is purely algebraic. An invariant polynomial is a *continuous* function on the space of matrices. The diagonalisable matrices are *dense*: a matrix with distinct eigenvalues is diagonalisable, and any matrix can be perturbed slightly to make its eigenvalues distinct. Two continuous functions that agree on a dense set agree everywhere. So an identity between invariant polynomials that holds for diagonal matrices holds for all matrices. This is "a kind of linear algebraic analogue of the validity of the splitting principle".

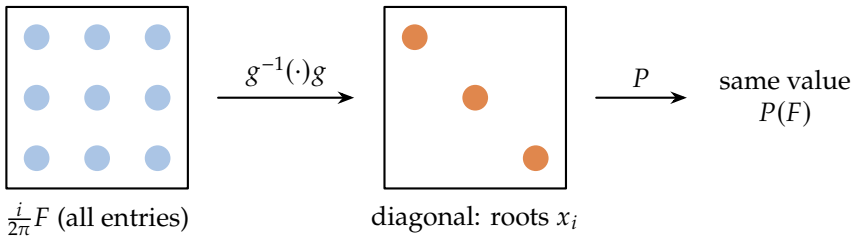


Figure 3.67: Why roots are eigenvalues. An invariant polynomial P does not change under conjugation, so (where possible) we may diagonalise $\frac{i}{2\pi}F$ first; the diagonal entries play the role of the Chern roots. Where the matrix cannot be diagonalised, continuity and density of the diagonalisable matrices finish the job. No global splitting of the bundle is ever needed.

Worked example 3.102: The curvature form of c_2 from the roots

From $c_2 = \sigma_2 = \frac{1}{2}(\sigma_1^2 - s_2)$ (Newton) and the substitution rule,

$$c_2(F) \overset{(1)}{=} \tfrac{1}{2}\left[\left(\operatorname{tr}\tfrac{i}{2\pi}F\right)^2 - \operatorname{tr}\left(\tfrac{i}{2\pi}F\right)^2\right]$$
$$\overset{(2)}{=} \frac{1}{8\pi^2}\left[\operatorname{tr}F\wedge F - \operatorname{tr}F\wedge\operatorname{tr}F\right],$$

(1) $\sigma_1 \to \operatorname{tr}\frac{i}{2\pi}F$ and $s_2 \to \operatorname{tr}(\frac{i}{2\pi}F)^2$.
(2) $(i/2\pi)^2 = -1/4\pi^2$, and the overall $\frac{1}{2}$.

which reproduces (3.156) of the previous section without expanding any determinant.

Reading Nash (pp. 79–80, (3.93)–(3.94), and p. 88)

Nash introduces the splitting principle in words (p. 79): to prove an identity it suffices to prove it for $E = L_1 \oplus \cdots \oplus L_k$, because any bundle F over X "can always be regarded as the pullback of another rank k bundle E over Y which is a sum of line bundles" (Bott and Tu). He does not mention the injectivity of $H^*(X) \to H^*(Y)$, which is the essential point; we have supplied it with the flag bundle. His (3.93) is $c(E) = \prod(1 + x_i)$ with $x_i = c_1(L_i)$ and

(3.94) lists the c_j as elementary symmetric functions. On p. 88 he gives the curvature rule $x_i \to iF_i/2\pi$ and justifies it by density of diagonalisable matrices.

Recap

- Chern roots: $c(E) = \prod(1 + x_i)$, $c_j = \sigma_j(x)$; they are the "eigenvalues" of the bundle.
- Splitting principle: pull back to the flag bundle, where E splits and $H^*(X)$ injects; identities true for sums of lines are true.
- Newton: $s_1 = c_1$, $s_2 = c_1^2 - 2c_2$, $s_3 = c_1^3 - 3c_1c_2 + 3c_3$.
- Recipes: E^* has roots $-x_i$; $E \otimes L$ has $x_i + y$; $\operatorname{End} E$ has $x_i - x_j$; $\Lambda^p E$ has sums of p distinct roots; $c(T\mathbb{CP}^n) = (1 + h)^{n+1}$.
- Curvature: replace the roots by the eigenvalues of $\frac{i}{2\pi}F$.

3.18 The Chern character and the Todd class

K-theory has two operations, $\oplus$ and $\otimes$. Cohomology also has two, $+$ and the cup (wedge) product. The previous two parts built K-theory; this chapter is about linking it to cohomology, where we can integrate and get numbers. So the question is: *is there a characteristic class that turns $\oplus$ into $+$ and $\otimes$ into the product?*

The total Chern class does not do it: it turns $\oplus$ into a *product*, $c(E \oplus F) = c(E)c(F)$, and it behaves messily under $\otimes$ (Example 3.98). We need a function that converts sums of roots into products and is additive over the roots. With Chern roots the answer is forced on us. The roots of $E \oplus F$ are the x_i together with the y_j; the roots of $E \otimes F$ are all the sums $x_i + y_j$. A function $\phi(E) = \sum_i f(x_i)$ is automatically additive under $\oplus$. It is multiplicative under $\otimes$ exactly when $\sum_{i,j} f(x_i + y_j) = \sum_i f(x_i) \sum_j f(y_j)$, that is, when $f(x + y) = f(x)f(y)$. The solution is the exponential. This is the Chern character.

The Todd class comes along for a different reason, which Nash flags in advance: together with the Chern character it makes up the index formula of Chapter 4.

3.18.1 Definition and first terms

Working definition 3.43 (Chern character)

For a complex bundle E of rank k with Chern roots $x_1, \dots, x_k$ [formal classes with $c(E) = \prod_i(1 + x_i)$, as if E were a sum of line bundles with $c_1 = x_i$],

$$\operatorname{ch}(E) = \sum_{i=1}^{k} e^{x_i} = \sum_{m \geq 0} \frac{1}{m!}\, s_m(x), \tag{3.168}$$

where $s_m = \sum_i x_i^m$ are the power sums. The degree-$2m$ part is $\operatorname{ch}_m(E) = s_m/m!$. The series stops at $m = \lfloor \dim X/2 \rfloor$ because higher cohomology vanishes. *Operationally:* use the Newton identities [$s_1 = c_1$, $s_2 = c_1^2 - 2c_2$, and so on] to write s_m in Chern classes; or, with a connection, use the curvature formula $\operatorname{ch}(F) = \operatorname{tr} \exp\left(\frac{i}{2\pi}F\right)$ of Section 3.18.5.

Nash calls (3.168) "a kind of generating function": the m-th power sum is generated with

weight $1/m!$. Using the Newton identities of Section 3.17.3,

$$\begin{aligned}\operatorname{ch}(E) &\overset{(1)}{=} k + s_1 + \tfrac{1}{2}s_2 + \tfrac{1}{6}s_3 + \dots \\ &\overset{(2)}{=} k + c_1 + \tfrac{1}{2}\big(c_1^2 - 2c_2\big) + \tfrac{1}{6}\big(c_1^3 - 3c_1c_2 + 3c_3\big) + \dots\end{aligned} \tag{3.169}$$

(1) $e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \dots$ for each root; summing over k roots gives k from the constant terms and $s_m/m!$ from the rest.

(2) $s_1 = \sigma_1$, $s_2 = \sigma_1^2 - 2\sigma_2$, $s_3 = \sigma_1^3 - 3\sigma_1\sigma_2 + 3\sigma_3$, and $\sigma_j = c_j$.

The first three terms are Nash's (3.97). Note what the degree-0 part is: the *rank*. The Chern character remembers the dimension of the fibre, which the total Chern class forgets.

Worked example 3.103: Line bundles and monopoles

For a line bundle L with $c_1(L) = x$, $\operatorname{ch}(L) = e^x = 1 + x + \frac{x^2}{2} + \dots$. Over S^2, $x^2 = 0$, so for the monopole bundles

$$\operatorname{ch}(L_n) = 1 + n\,u, \qquad \int_{S^2} \operatorname{ch}(L_n) = n.$$

Nash's Bott generator $b = [\eta] - [I] \in \widetilde{K}(S^2)$ then has $\operatorname{ch}(b) = \operatorname{ch}(\eta) - 1 = c_1(\eta)$, and $\int_{S^2} \operatorname{ch}(b) = \pm 1$: $+1$ if η is the hyperplane bundle L_1, -1 if η is the tautological bundle L_{-1} (the two conventions for "the Hopf bundle" both occur in the literature). Either way the Chern character identifies $\widetilde{K}(S^2) \cong \mathbb{Z}$ with $H^2(S^2;\mathbb{Z}) \cong \mathbb{Z}$: the K-theory class is literally the monopole charge.

Worked example 3.104: The instanton bundle over S^4

Let $E \to S^4$ be an SU(2) bundle with $\int c_2(E) = k$. Then $c_1 = 0$ and there is no degree-6 or higher cohomology, so

$$\begin{aligned}\operatorname{ch}(E) &\overset{(1)}{=} 2 + c_1 + \tfrac{1}{2}\big(c_1^2 - 2c_2\big) \\ &\overset{(2)}{=} 2 - c_2, \qquad \int_{S^4} \operatorname{ch}(E) = -k.\end{aligned}$$

(1) (3.169) with rank 2, truncated at degree 4.

(2) $c_1 = 0$ for SU(2).

The class $[E] - [I^2] \in \widetilde{K}(S^4)$ has $\operatorname{ch} = -c_2$. For the BPST bundle of Example 3.93, $\int \operatorname{ch}_2 = -1$: the generator of $\widetilde{K}(S^4) \cong \mathbb{Z}$ (which 3b built as the external product $b \boxtimes b$ on $S^2 \times S^2$ collapsed to S^4) is, up to sign, the instanton.

Worked example 3.105: The tangent bundle of $\mathbb{CP}^2$, two ways

From the Euler sequence $T\mathbb{CP}^n \oplus I = (n+1)\gamma^*$ (Example 3.101) and additivity (proved below),

$$\operatorname{ch}(T\mathbb{CP}^n) + 1 \overset{(1)}{=} (n+1)\,e^h \quad \Rightarrow \quad \operatorname{ch}(T\mathbb{CP}^2) = 3e^h - 1 = 2 + 3h + \tfrac{3}{2}h^2.$$

(1) ch of I is 1 and $\operatorname{ch}(\gamma^*) = e^{c_1(\gamma^*)} = e^h$.

Check against (3.169) with $k = 2$, $c_1 = 3h$, $c_2 = 3h^2$:

$$2 + 3h + \tfrac{1}{2}(9h^2 - 6h^2) = 2 + 3h + \tfrac{3}{2}h^2. \quad \checkmark$$

The coefficient $\frac{3}{2}$ is not an integer. Integrated over $\mathbb{CP}^2$ (where $\int h^2 = 1$), $\int_{\mathbb{CP}^2} \mathrm{ch}_2(T\mathbb{CP}^2) = \frac{3}{2}$.

3.18.2 Rational, not integral

The factorials make $\mathrm{ch}(E)$ a class with *rational* coefficients, Nash's (3.96):

$$\mathrm{ch}(E) \in H^*(X;\mathbb{Q}).$$

Two consequences follow, and Nash mentions both.

The values on cycles need not be integers. Example 3.105 already showed this. An even simpler case: for the hyperplane bundle $\gamma^* = L_1$ over $\mathbb{CP}^2$, $\mathrm{ch}(\gamma^*) = 1 + h + \frac{1}{2}h^2$ and $\int_{\mathbb{CP}^2} \mathrm{ch}_2(\gamma^*) = \frac{1}{2}$. On spheres, by contrast, the top component ch_n on S^{2n} *is* integral (a theorem of Bott: ch maps $\widetilde{K}(S^{2n})$ onto $H^{2n}(S^{2n};\mathbb{Z})$), which is what Examples 3.103 and 3.104 found.

Torsion is invisible. A class $t \in H^*(X;\mathbb{Z})$ with $mt = 0$ for some integer m becomes 0 in rational cohomology, since $t = \frac{1}{m}(mt) = 0$ once we may divide. So the Chern character cannot distinguish bundles that differ only in torsion.

Worked example 3.106: A bundle the Chern character cannot see

On the real projective plane $\mathbb{RP}^2$ (a disc with antipodal boundary points identified) take the real Möbius line bundle μ and complexify it, $L = \mu \otimes_{\mathbb{R}} \mathbb{C}$. Its transition functions are the constants ± 1, so L carries a *flat* connection (constant transition functions are compatible with $A = 0$ in every patch). Then

$$\mathrm{ch}(L) \overset{(1)}{=} \mathrm{tr}\exp\Big(\frac{i}{2\pi}F\Big) \overset{(2)}{=} \mathrm{tr}\exp(0) = 1 = \mathrm{ch}(I).$$

(1) The curvature formula (3.173) below, with any connection (the class does not depend on it).
(2) The flat connection has $F = 0$.

Yet L is not trivial: $L \otimes L$ has transition functions $(\pm 1)^2 = 1$, so $2c_1(L) = c_1(L \otimes L) = 0$, and in fact $c_1(L)$ is the nonzero element of $H^2(\mathbb{RP}^2;\mathbb{Z}) = \mathbb{Z}_2$. The Chern class sees L; the Chern character, like any curvature integral, cannot. Physically: a flat gauge field with holonomy -1 around a non-contractible loop has no field strength anywhere, so no local density can detect it, but it is a different sector (an Aharonov–Bohm phase).

Watch out: Rational does not mean "approximate"

$\mathrm{ch}(E)$ is an exact cohomology class; only its coefficients live in $\mathbb{Q}$. Integrality comes back in index theory: $\int_X \mathrm{ch}(E)\ \mathrm{td}(X)$ and $\int_X \mathrm{ch}(E)\hat{A}(X)$ are always integers (they are indices), even though neither factor alone need be.

3.18.3 Sums and products

Nash says it is "straightforward to check" (3.98). Here is the check, by the splitting principle. Let E and F have roots x_i ($i \le k$) and y_j ($j \le l$). Then

$$\operatorname{ch}(E \oplus F) \overset{(1)}{=} \sum_{i=1}^{k} e^{x_i} + \sum_{j=1}^{l} e^{y_j} = \operatorname{ch} E + \operatorname{ch} F, \tag{3.170}$$

$$\operatorname{ch}(E \otimes F) \overset{(2)}{=} \sum_{i,j} e^{x_i + y_j} \overset{(3)}{=} \sum_{i,j} e^{x_i} e^{y_j} = \Big(\sum_i e^{x_i}\Big)\Big(\sum_j e^{y_j}\Big) = \operatorname{ch} E \, \operatorname{ch} F. \tag{3.171}$$

(1) The roots of $E \oplus F$ are the x_i together with the y_j (Whitney).
(2) If $E = \bigoplus L_i$ and $F = \bigoplus M_j$ then $E \otimes F = \bigoplus_{i,j} L_i \otimes M_j$, and $c_1(L_i \otimes M_j) = x_i + y_j$.
(3) The exponential of a sum of commuting classes is the product of exponentials (the classes have even degree, so they commute).

Worked example 3.107: Checking multiplicativity in low degree

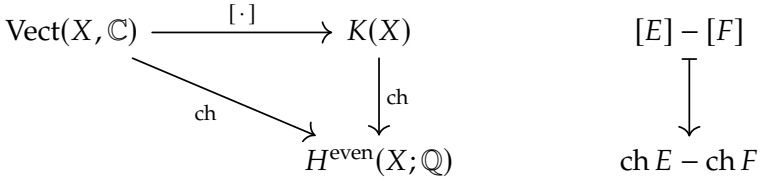
Take E of rank 2 and L a line bundle with $c_1(L) = y$. By (3.171),

$$\operatorname{ch}(E \otimes L) = \big(2 + c_1 + \tfrac{1}{2}(c_1^2 - 2c_2)\big)\big(1 + y + \tfrac{1}{2}y^2\big) = 2 + (c_1 + 2y) + \big(\tfrac{1}{2}c_1^2 - c_2 + c_1 y + y^2\big) + \dots$$

From Example 3.98 instead, $c_1(E \otimes L) = c_1 + 2y$ and $c_2(E \otimes L) = c_2 + c_1 y + y^2$, so

$$\begin{aligned} \tfrac{1}{2}\big(c_1(E \otimes L)^2 - 2c_2(E \otimes L)\big) &= \tfrac{1}{2}\big(c_1^2 + 4c_1 y + 4y^2\big) - c_2 - c_1 y - y^2 \\ &= \tfrac{1}{2}c_1^2 - c_2 + c_1 y + y^2. \end{aligned}$$

The two computations agree.



in $K(X)$	$\xrightarrow{\text{ch}}$	in $H^{\text{even}}(X;\mathbb{Q})$
$\oplus$ (sum)		$+$
$\otimes$ (product)		$\wedge$ (cup product)
rank k		degree-0 part k
$[L_n] - [I]$ on S^2		$n\,u$
torsion elements		0

Figure 3.68: The Chern character as a ring homomorphism from K-theory to even rational cohomology. Top: it is defined on actual bundles and extended to formal differences. Bottom: what it does to each piece of structure. The last row is the price of rational coefficients.

3.18.4 The Chern character on $K(X)$

Since ch is additive, it extends from bundles to the K-group. Recall from the previous part that every element of $K(X)$ is a formal difference $[E] - [F]$. Define (Nash's (3.99))

$$\operatorname{ch} : K(X) \longrightarrow H^*(X;\mathbb{Q}), \qquad [E] - [F] \longmapsto \operatorname{ch} E - \operatorname{ch} F. \tag{3.172}$$

Well defined? The same element has many representatives, and we must check that they all give the same answer, Nash's (3.100). Recall the rule (3.42) of the previous part: $[E]-[F]=[G]-[H]$ exactly when $E\oplus H\oplus J\cong G\oplus F\oplus J$ for some bundle J. Then

$$\begin{aligned}[E]-[F]=[G]-[H] &\overset{(1)}{=} E\oplus H\oplus J\cong G\oplus F\oplus J\\ &\overset{(2)}{=} \operatorname{ch}(E\oplus H\oplus J)=\operatorname{ch}(G\oplus F\oplus J)\\ &\overset{(3)}{=} \operatorname{ch}E+\operatorname{ch}H+\operatorname{ch}J=\operatorname{ch}G+\operatorname{ch}F+\operatorname{ch}J\\ &\overset{(4)}{=} \operatorname{ch}E-\operatorname{ch}F=\operatorname{ch}G-\operatorname{ch}H.\end{aligned}$$

(1) The definition of equality in the Grothendieck group (Nash's (3.42)). Here "=" in the chain means "implies".
(2) Isomorphic bundles have the same characteristic classes.
(3) Additivity (3.170).
(4) Cancel $\operatorname{ch}J$ (cohomology is a group, so cancellation is allowed there, unlike for bundles).

This is Nash's (3.101). The same additivity and multiplicativity give

$$\begin{aligned}\operatorname{ch}\big(([E]-[F])([G]-[H])\big) &\overset{(1)}{=} \operatorname{ch}\big([E\otimes G]-[E\otimes H]-[F\otimes G]+[F\otimes H]\big)\\ &\overset{(2)}{=} \operatorname{ch}E\operatorname{ch}G-\operatorname{ch}E\operatorname{ch}H-\operatorname{ch}F\operatorname{ch}G+\operatorname{ch}F\operatorname{ch}H\\ &=(\operatorname{ch}E-\operatorname{ch}F)(\operatorname{ch}G-\operatorname{ch}H),\end{aligned}$$

(1) The product in $K(X)$ is induced by $\otimes$, expanded bilinearly.
(2) Additivity on each term, then multiplicativity.

so ch is a **ring homomorphism** (Nash's (3.102)). Since all Chern roots have degree 2, ch lands in even degrees (Nash's (3.103)):

$$\operatorname{ch}: K(X)\longrightarrow \bigoplus_{i\geq 0} H^{2i}(X;\mathbb{Q}).$$

A theorem beyond Nash's text makes the link precise: for a finite cell complex, ch becomes an *isomorphism* after tensoring with $\mathbb{Q}$, $K(X)\otimes\mathbb{Q}\cong H^{\text{even}}(X;\mathbb{Q})$. K-theory and even cohomology carry the same rational information; K-theory additionally knows about torsion and integrality.

3.18.5 The curvature formula

By the substitution rule of Section 3.17.5, $\sum_i e^{x_i}$ becomes the trace of the exponential of the matrix $\frac{i}{2\pi}F$, Nash's (3.127):

$$\operatorname{ch}(F)=\operatorname{tr}\exp\Big(\frac{i}{2\pi}F\Big)=\sum_{m\geq 0}\frac{1}{m!}\Big(\frac{i}{2\pi}\Big)^m\operatorname{tr}F^m. \tag{3.173}$$

The exponential series terminates because F^m is a $2m$-form. The first terms are

$$\operatorname{ch}(F)=k+\frac{i}{2\pi}\operatorname{tr}F-\frac{1}{8\pi^2}\operatorname{tr}F\wedge F-\frac{i}{48\pi^3}\operatorname{tr}F\wedge F\wedge F+\dots \tag{3.174}$$

Each $\operatorname{tr}F^m$ is closed and its class is independent of the connection (Sections 3.16.3 and 3.16.4), so no further proof is needed that (3.173) represents a characteristic class.

Physical picture: Why physicists meet $\operatorname{tr} e^{iF/2\pi}$

The Chern character is the trace of the exponential of the field strength: the same structure as a partition function $\operatorname{tr} e^{-\beta H}$. This is not a coincidence. In the heat-kernel proof of the index theorem (Chapter 4) the operator $e^{-t\not{D}^2}$ contains, through the Lichnerowicz formula, the factor $e^{-t\,\sigma^{\mu\nu}F_{\mu\nu}/2}$, and its supertrace as $t \to 0$ produces exactly $\operatorname{ch}(F)$. In the anomaly computation (Chapter 10) the four-dimensional chiral anomaly density is $\operatorname{ch}_2(F) \propto \operatorname{tr} F \wedge F$.

Worked example 3.108: The instanton again, from curvature

For an SU(2) connection, $\operatorname{tr} F = 0$, so by (3.174) and (3.157)

$$\operatorname{ch}_2(F) \overset{(1)}{=} -\frac{1}{8\pi^2}\operatorname{tr} F \wedge F \overset{(2)}{=} -c_2(F),$$

(1) The $m = 2$ term: $\frac{1}{2}(\frac{i}{2\pi})^2 = -\frac{1}{8\pi^2}$.

(2) $c_2(F) = \frac{1}{8\pi^2}\operatorname{tr} F \wedge F$ for traceless F.

in agreement with Example 3.104. For the BPST connection, $\int_{\mathbb{R}^4} \operatorname{ch}_2(F) = -1$.

3.18.6 The Todd class

The second class of the section is also defined by a generating function in the roots. Its motivation will only be fully visible in Chapter 4: the Riemann–Roch–Hirzebruch theorem says that the index of the Dolbeault operator on a complex manifold X, twisted by a bundle E, is $\int_X \operatorname{ch}(E)\,\operatorname{td}(TX)$. We preview this below on $\mathbb{CP}^1$.

Working definition 3.44 (Todd class)

For a complex bundle E with Chern roots x_i, the **total Todd class** is (Nash's (3.104))

$$\operatorname{td}(E) = \prod_{i=1}^{k} \frac{x_i}{1 - e^{-x_i}}. \tag{3.175}$$

Each factor is a power series starting with 1 $[\,1 + \frac{x_i}{2} + \frac{x_i^2}{12} + \cdots\,]$, so $\operatorname{td}(E) = 1 + (\text{higher degree})$, and it is invertible [it has an inverse power series, because its degree-0 part is 1]. *Operationally:* expand each factor to the degree you need, multiply, and rewrite in Chern classes; or replace x by the matrix $\frac{i}{2\pi}F$ and take a determinant, $\operatorname{td}(F) = \det\!\left(\frac{iF/2\pi}{1-e^{-iF/2\pi}}\right)$.

The one-variable series. First expand $x/(1-e^{-x})$. Write $\frac{1-e^{-x}}{x} = 1 - \frac{x}{2} + \frac{x^2}{6} - \frac{x^3}{24} + \frac{x^4}{120} - \dots$ and look for the inverse series $1 + a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4 + \dots$. Their product must be 1, so each coefficient of x^n, $n \geq 1$, must vanish:

$$\begin{aligned}
x^1: &\quad a_1 - \tfrac{1}{2} = 0 && \Rightarrow\ a_1 = \tfrac{1}{2},\\
x^2: &\quad a_2 - \tfrac{1}{2}a_1 + \tfrac{1}{6} = 0 && \Rightarrow\ a_2 = \tfrac{1}{4} - \tfrac{1}{6} = \tfrac{1}{12},\\
x^3: &\quad a_3 - \tfrac{1}{2}a_2 + \tfrac{1}{6}a_1 - \tfrac{1}{24} = 0 && \Rightarrow\ a_3 = \tfrac{1}{24} - \tfrac{1}{12} + \tfrac{1}{24} = 0,\\
x^4: &\quad a_4 - \tfrac{1}{2}a_3 + \tfrac{1}{6}a_2 - \tfrac{1}{24}a_1 + \tfrac{1}{120} = 0 && \Rightarrow\ a_4 = -\tfrac{1}{72} + \tfrac{1}{48} - \tfrac{1}{120} = -\tfrac{1}{720}.
\end{aligned}$$

So

$$\frac{x}{1-e^{-x}} = 1 + \frac{x}{2} + \frac{x^2}{12} - \frac{x^4}{720} + \dots \tag{3.176}$$

The vanishing of all odd terms beyond the first has a neat reason:

$$\frac{x}{1-e^{-x}} - \frac{x}{2} = \frac{x}{2}\cdot\frac{1+e^{-x}}{1-e^{-x}} = \frac{x}{2}\coth\frac{x}{2},$$

an even function. Equivalently $\frac{x}{1-e^{-x}} = e^{x/2}\,\frac{x/2}{\sinh(x/2)}$, a factorisation that becomes the relation $\mathrm{td} = e^{c_1/2}\hat{A}$ in Section 3.20.8.

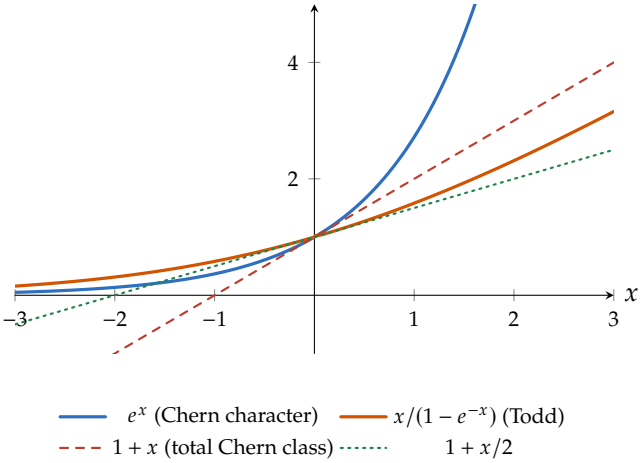


Figure 3.69: The one-variable generating functions of this section. A multiplicative class is fixed by a series $Q(x)$ through $\prod_i Q(x_i)$: $1+x$ gives the total Chern class, $x/(1-e^{-x})$ the Todd class. The Chern character is additive instead, $\sum_i e^{x_i}$. Near $x=0$ the Todd series hugs its tangent line $1+x/2$ (its degree-1 coefficient is $\frac{1}{2}$) and bends up with curvature $\frac{1}{12}$; it is smooth at $x=0$ despite the $0/0$ in its formula.

In Chern classes. Multiply the series for all roots and collect by degree. The degree-1 part is $\frac{1}{2}\sum x_i = \frac{1}{2}c_1$. The degree-2 part has $x_i^2/12$ from one factor and $\frac{x_i}{2}\frac{x_j}{2}$ from two:

$$\mathrm{td}_2 \overset{(1)}{=} \frac{1}{12}\sum_i x_i^2 + \frac{1}{4}\sum_{i<j} x_i x_j$$
$$\overset{(2)}{=} \frac{1}{12}\left(c_1^2 - 2c_2\right) + \frac{1}{4}c_2 = \frac{1}{12}\left(c_1^2 + c_2\right).$$

(1) Pick the x^2 term from one factor, or the x term from two different factors.
(2) $s_2 = c_1^2 - 2c_2$ and $\sigma_2 = c_2$.

The degree-3 part has no x_i^3 terms ($a_3 = 0$), terms $x_i^2 x_j$ ($i \neq j$) with coefficient $\frac{1}{12}\cdot\frac{1}{2}$, and terms $x_i x_j x_l$ ($i<j<l$) with coefficient $\frac{1}{8}$:

$$\mathrm{td}_3 \overset{(1)}{=} \frac{1}{24}\sum_{i\neq j} x_i^2 x_j + \frac{1}{8}\sum_{i<j<l} x_i x_j x_l$$
$$\overset{(2)}{=} \frac{1}{24}\left(c_1c_2 - 3c_3\right) + \frac{1}{8}c_3 = \frac{1}{24}c_1c_2.$$

(1) Choose which factors contribute which powers.
(2) $\sigma_1\sigma_2 = \sum_{i\neq j} x_i^2 x_j + 3\sigma_3$, since each triple product $x_i x_j x_l$ arises three times in $\sigma_1\sigma_2$.

Altogether, Nash's (3.106):

$$\mathrm{td}(E) = 1 + \frac{1}{2}c_1 + \frac{1}{12}\left(c_1^2 + c_2\right) + \frac{1}{24}c_1c_2 + \dots \tag{3.177}$$

Being a product over roots, td is multiplicative, $\mathrm{td}(E \oplus F) = \mathrm{td}(E)\,\mathrm{td}(F)$ (Nash's (3.105)), and like ch it is a rational class.

Worked example 3.109: The Todd genus of $\mathbb{CP}^n$ is 1

By the Euler sequence and multiplicativity ($\mathrm{td}(I) = 1$), $\mathrm{td}(T\mathbb{CP}^n) = \left(\frac{h}{1-e^{-h}}\right)^{n+1}$. The **Todd genus** is $\int_{\mathbb{CP}^n} \mathrm{td}(T\mathbb{CP}^n)$, the coefficient of h^n. For $n = 1$: $\mathrm{td} = 1 + \frac{1}{2}c_1 = 1 + h$, and $\int = 1$. For $n = 2$: $c_1 = 3h$, $c_2 = 3h^2$, and $\mathrm{td}_2 = \frac{1}{12}(9+3)h^2 = h^2$, so again $\int = 1$. For general n, a residue does it:

$$\int_{\mathbb{CP}^n} \mathrm{td} \overset{(1)}{=} \frac{1}{2\pi i}\oint \left(\frac{h}{1-e^{-h}}\right)^{n+1}\frac{\mathrm{d}h}{h^{n+1}} = \frac{1}{2\pi i}\oint \frac{\mathrm{d}h}{(1-e^{-h})^{n+1}}$$
$$\overset{(2)}{=} \frac{1}{2\pi i}\oint \frac{\mathrm{d}w}{w^{n+1}(1-w)} \overset{(3)}{=} 1.$$

(1) The coefficient of h^n in a power series $f(h)$ is $\frac{1}{2\pi i}\oint f(h)\,h^{-n-1}\mathrm{d}h$ around a small circle.
(2) Substitute $w = 1 - e^{-h}$, so $\mathrm{d}w = e^{-h}\mathrm{d}h = (1-w)\mathrm{d}h$; a small circle around $h = 0$ becomes a small circle around $w = 0$.
(3) The coefficient of w^n in $\frac{1}{1-w} = \sum w^j$ is 1.

In Chapter 4 this "1" will be the statement that the only holomorphic functions on $\mathbb{CP}^n$ are constants: the Dolbeault index is $\dim H^{0,0} = 1$.

Worked example 3.110: Riemann–Roch on $\mathbb{CP}^1$: a preview

Take $X = \mathbb{CP}^1$ and the line bundle $L_n = \mathcal{O}(n)$ (the monopole bundle, $c_1 = nh$). Then

$$\int_{\mathbb{CP}^1} \mathrm{ch}(L_n)\,\mathrm{td}(T\mathbb{CP}^1) \overset{(1)}{=} \int_{\mathbb{CP}^1} (1+nh)(1+h)$$
$$\overset{(2)}{=} \int_{\mathbb{CP}^1} \big(1 + (n+1)h\big) = n+1.$$

(1) $\mathrm{ch}(L_n) = e^{nh} = 1 + nh$ and $\mathrm{td}(T\mathbb{CP}^1) = 1 + \frac{1}{2}c_1(T\mathbb{CP}^1) = 1 + h$, since $h^2 = 0$ on $\mathbb{CP}^1$.
(2) Only the degree-2 part integrates.

For $n \geq 0$ this is exactly the number of holomorphic sections of $\mathcal{O}(n)$: the homogeneous polynomials of degree n in (z_0, z_1), spanned by $z_0^n, z_0^{n-1}z_1, \dots, z_1^n$, which is $n+1$ of them (computed with Čech cohomology in part 3a). For $n = -2$ the formula gives -1: no sections, and one-dimensional H^1. The Chern character counts the twisting, the Todd class corrects for the curvature of the base, and the product integrates to an integer that counts zero modes. This is the index theorem in miniature.

Reading Nash (pp. 78–82, (3.95)–(3.106))

Nash introduces the Chern character as "an extremely valuable tool in K-theory" linking $K(X)$ and $H^*(X)$, defines it by (3.95), $\mathrm{ch}(E) = \sum e^{x_i}$, stresses rational coefficients and blindness to torsion (3.96), gives the first terms (3.97), states additivity and multiplicativity (3.98) as "straightforward", and defines ch on $K(X)$ (3.99), checking well-definedness (3.100)–(3.101) with the Grothendieck-group relation (3.42) and concluding it is a ring

homomorphism into even rational cohomology (3.102)–(3.103). His closing remark, that $\mathrm{ch}(E)$ evaluated on a cycle "need only be rational rather than integral", is Example 3.105. The Todd class (3.104) comes with multiplicativity (3.105) and the expansion (3.106), which we derived term by term. His curvature formula (3.127), $\mathrm{ch}(F) = \operatorname{tr}\exp(iF/2\pi)$, is (3.173).

Recap

- $\mathrm{ch}(E) = \sum e^{x_i} = k + c_1 + \frac{1}{2}(c_1^2 - 2c_2) + \frac{1}{6}(c_1^3 - 3c_1c_2 + 3c_3) + \dots$; with curvature $\mathrm{ch}(F) = \operatorname{tr} e^{iF/2\pi}$.
- $\mathrm{ch}(E \oplus F) = \mathrm{ch}\,E + \mathrm{ch}\,F$, $\mathrm{ch}(E \otimes F) = \mathrm{ch}\,E\ \mathrm{ch}\,F$; so $\mathrm{ch} : K(X) \to H^{\mathrm{even}}(X;\mathbb{Q})$ is a well-defined ring homomorphism.
- Rational coefficients: values on cycles can be fractions ($\frac{1}{2}$ on $\mathbb{CP}^2$); torsion (the flat bundle on $\mathbb{RP}^2$) is invisible.
- Monopole: $\mathrm{ch}(L_n) = 1 + nu$. Instanton: $\mathrm{ch} = 2 - c_2$, $\int \mathrm{ch}_2 = -k$.
- $\mathrm{td}(E) = \prod \frac{x_i}{1-e^{-x_i}} = 1 + \frac{1}{2}c_1 + \frac{1}{12}(c_1^2 + c_2) + \frac{1}{24}c_1c_2 + \dots$; Todd genus of $\mathbb{CP}^n$ is 1; $\int_{\mathbb{CP}^1} \mathrm{ch}(\mathcal{O}(n))\,\mathrm{td} = n + 1$.

3.19 Real bundles, complexification and Pontryagin classes

The tangent bundle of spacetime, or of any real manifold, is a *real* vector bundle. Chern classes are defined only for complex bundles. So how do we measure the twisting of a real bundle, the kind of twisting that gravitational anomalies and the $\hat{A}$-genus of the Dirac index see? Nash's answer is the obvious one: turn the real bundle into a complex one by allowing complex coefficients in each fibre, then take Chern classes. The result is the Pontryagin classes. Before we get there we need a small toolkit of operations on bundles, each of which is one line in terms of transition functions: dual, conjugate, complexification and realification.

Throughout, a bundle is described by transition functions $g_{\alpha\beta}$ on overlaps, relating the local frames: $s_\beta = s_\alpha\, g_{\alpha\beta}$ on $U_\alpha \cap U_\beta$. A section with components u_β in the frame s_β has components $u_\alpha = g_{\alpha\beta}u_\beta$ in the frame s_α.

3.19.1 The dual and the conjugate bundle

Dual. The fibre of E^* at x is the space of linear functionals on E_x. If $s = (s_1, \dots, s_k)$ is a frame of E, the **dual frame** $s^* = (s^1, \dots, s^k)$ is defined by $s^i(s_j) = \delta^i_j$. How does it change when $s' = s\,g$? Writing the dual frame as a column, we need $s'^* = h\,s^*$ for some matrix h with

$$\delta^i_j = s'^{\,i}(s'_j) \overset{(1)}{=} \sum_{l,m} h_{il}\, s^l\big(s_m\, g_{mj}\big) \overset{(2)}{=} \sum_l h_{il}\, g_{lj} = (hg)_{ij},$$

(1) Substitute $s'^{\,i} = \sum_l h_{il}s^l$ and $s'_j = \sum_m s_m g_{mj}$.

(2) Linearity and $s^l(s_m) = \delta^l_m$.

so $h = g^{-1}$. The *components* of a covector then transform by $(g^{-1})^T$, and the transition functions of E^* are

$$g^{E^*}_{\alpha\beta} = \left(g^{-1}_{\alpha\beta}\right)^T.$$

This is the "inverse transpose" that every physicist knows from upper and lower indices: $v^\mu \mapsto \Lambda^\mu{}_\nu v^\nu$ but $w_\mu \mapsto w_\nu (\Lambda^{-1})^\nu{}_\mu$.

Conjugate. The **conjugate bundle** $\bar{E}$ has the same underlying real bundle as E but with scalar multiplication by λ redefined as multiplication by $\bar{\lambda}$. In coordinates this just means complex-conjugating everything:

$$g^{\bar{E}}_{\alpha\beta} = \bar{g}_{\alpha\beta}.$$

Unitary bundles: $E^* \cong \bar{E}$. Choose a Hermitian inner product on the fibres (always possible, by a partition of unity) and use only orthonormal frames. Then the transition functions are unitary, $g^{-1} = g^\dagger = \bar{g}^T$, and therefore

$$\left(g^{-1}\right)^T = \bar{g} \qquad \Longrightarrow \qquad E^* \cong \bar{E}.$$

Concretely, the isomorphism is $v \mapsto \langle v, \cdot \rangle$: the inner product turns a vector into a covector, antilinearly in v, which is exactly a complex-*linear* map from $\bar{E}$ to E^*. (This is the bra–ket correspondence $|v\rangle \mapsto \langle v|$.)

Nash's footnote 6 is about the step "use only orthonormal frames". Given any frame, Gram–Schmidt makes it orthonormal by a smooth upper-triangular change of frame, so every bundle has a cover with unitary transition functions: the **structure group** reduces from $\mathrm{GL}(k,\mathbb{C})$ to $U(k)$. Nothing topological is lost, because $\mathrm{GL}(k,\mathbb{C})$ deformation retracts onto $U(k)$ (polar decomposition $g = u\,p$ with p positive, and p^t, $0 \le t \le 1$, shrinks p to $\mathbb{1}$).

Worked example 3.111: Duals and conjugates of monopole bundles

The monopole bundle L_n is glued by $g = e^{-in\phi}$ on the equator. Its dual has transition function $(g^{-1})^T = e^{+in\phi}$, and so does its conjugate, $\bar{g} = e^{+in\phi}$. Both are L_{-n}:

$$L_n^* \cong \bar{L}_n \cong L_{-n}, \qquad c_1(L_n^*) = -c_1(L_n).$$

In particular the dual of the tautological bundle $\gamma = L_{-1}$ is the **hyperplane bundle** $\gamma^* = L_1$. Its fibre over the line ℓ is ℓ^*, so each linear function on $\mathbb{C}^2$ restricts to a section: the coordinate z_0 gives a section of γ^* that vanishes only on the line $\{z_0 = 0\}$, a single point of $\mathbb{CP}^1$, with multiplicity one. One zero, and $c_1(\gamma^*) = +1$: for holomorphic line bundles c_1 counts the zeros of a section, a first glimpse of the Euler class of Section 3.20.

3.19.2 Line bundles form a group

For line bundles the transition functions are 1×1, so they commute and multiply: $L \otimes M$ has transition functions $g^L g^M$. The trivial bundle (transition functions 1) is a unit, and L^* (transition functions $1/g^L$) is an inverse, since $L \otimes L^*$ has transition functions 1. So isomorphism classes of line bundles form an **abelian group** under $\otimes$. This justifies Nash's notation L^{-1} for L^* and L^n for $\otimes^n L$.

Key idea: c_1 is a group isomorphism for line bundles

$c_1(L \otimes M) = c_1(L) + c_1(M)$ and $c_1(L^{-1}) = -c_1(L)$. For smooth complex line bundles, $L \mapsto c_1(L)$ is an isomorphism from the group of line bundles to $H^2(X;\mathbb{Z})$.

The homomorphism property was shown in Section 3.17.4. That it is an isomorphism is the statement $[X, \mathbb{CP}^\infty] = H^2(X;\mathbb{Z})$; part 3a obtained it from the exponential sequence of sheaves, $0 \to \mathbb{Z} \to \mathcal{E} \to \mathcal{E}^* \to 0$. On S^2 the group is $\mathbb{Z}$, generated by L_1, and $L_1^n = L_n$: this is the statement that a monopole of charge n is the n-th tensor power of the unit monopole. (In K-theory, by contrast, $[L_1]^n = [L_n]$ but also $[L_n] = n[L_1] - (n-1)$, because $([L_1]-1)^2 = 0$ in $K(S^2)$: the ring structure of K and the group structure of line bundles are different things.)

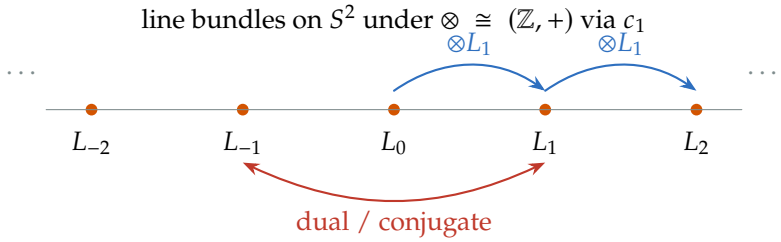


Figure 3.70: Line bundles over S^2 form the group $\mathbb{Z}$. Tensoring with the unit monopole L_1 steps one place to the right; the dual (equivalently, for unitary bundles, the conjugate) reflects $L_n \mapsto L_{-n}$. The trivial bundle L_0 is the identity element.

3.19.3 Real bundles

Now let V be a *real* vector bundle of rank k. A Riemannian (fibre) metric reduces the structure group from $\mathrm{GL}(k,\mathbb{R})$ to $\mathrm{O}(k)$, by the same Gram–Schmidt argument. For orthogonal g, $(g^{-1})^T = g$, so

$$V^* \cong V,$$

and conjugation does nothing on real fibres, $\bar{V} = V$. So for real bundles the dual and conjugate bring nothing new *topologically*. Nash adds a sensible caution: for $V = TX$ we still keep TX and T^*X apart in analysis. The isomorphism $TX \cong T^*X$ depends on the choice of metric (it is the index-lowering map $v^\mu \mapsto g_{\mu\nu}v^\nu$), while the distinction between a velocity and a gradient does not.

3.19.4 Complexification

Working definition 3.45 (Complexification)

The **complexification** $V \otimes_{\mathbb{R}} \mathbb{C}$ (also written $V_{\mathbb{C}}$) of a real bundle V of rank k is the complex bundle of rank k whose fibre is $V_x \otimes_{\mathbb{R}} \mathbb{C} = V_x \oplus iV_x$, the vectors $v_1 + iv_2$ with $v_1, v_2 \in V_x$. Its transition functions [the matrices g_{ab} converting fibre coordinates between patches] are the *same real matrices* g as for V, now acting on $\mathbb{C}^k$ (Nash writes $g \otimes 1$). *Operationally:* keep the transition functions, allow complex coefficients.

Nash's (3.107) writes this as $V \otimes_{\mathbb{R}} \mathbb{C} = V \otimes (\mathbb{R} \oplus i\mathbb{R})$, and the conjugate bundle as $V \otimes (\mathbb{R} \oplus -i\mathbb{R})$. The key point is that *nothing is new*:

Key idea: Nash's (3.108)

For a real bundle V,

$$V \otimes_{\mathbb{R}} \mathbb{C} \;\cong\; \overline{V \otimes_{\mathbb{R}} \mathbb{C}} \;\cong\; \left(V \otimes_{\mathbb{R}} \mathbb{C}\right)^*.$$

Why: the transition functions g of $V_{\mathbb{C}}$ are real, so $\bar{g} = g$, which gives the first isomorphism; and with an orthogonal choice $(g^{-1})^T = g$, which gives the second. Concretely, the fibrewise

map $v_1 + iv_2 \mapsto v_1 - iv_2$ is a complex-linear isomorphism $\overline{V_{\mathbb{C}}} \to V_{\mathbb{C}}$ (antilinear on $V_{\mathbb{C}}$, hence linear on the conjugate).

Consequence: odd Chern classes of $V_{\mathbb{C}}$ are 2-torsion. Combine (3.108) with the dual rule $c_j(E^*) = (-1)^j c_j(E)$ of Example 3.97:

$$c_j(V_{\mathbb{C}}) \overset{(1)}{=} c_j\big((V_{\mathbb{C}})^*\big) \overset{(2)}{=} (-1)^j\, c_j(V_{\mathbb{C}}),$$

(1) Isomorphic bundles have equal Chern classes, and $V_{\mathbb{C}} \cong (V_{\mathbb{C}})^*$.
(2) The roots of a dual bundle are minus the roots.

so for odd j, $2c_j(V_{\mathbb{C}}) = 0$. In real (de Rham) cohomology the odd classes simply vanish. Only the even ones, $c_2, c_4, \ldots$, carry information. That is why the Pontryagin classes below are built from c_{2i} alone.

Worked example 3.112: The complexified tangent bundle of S^2

TS^2 is a real rank-2 bundle. The complex structure of $S^2 = \mathbb{CP}^1$ makes it a complex line bundle, $T\mathbb{CP}^1 = L_2$ (Example 3.101). By the result of the next subsection, $TS^2 \otimes_{\mathbb{R}} \mathbb{C} \cong L_2 \oplus \bar{L}_2 = L_2 \oplus L_{-2}$, so

$$c(TS^2 \otimes \mathbb{C}) \overset{(1)}{=} (1 + 2u)(1 - 2u) \overset{(2)}{=} 1 - 4u^2 \overset{(3)}{=} 1.$$

(1) Whitney, with $c_1(L_{\pm 2}) = \pm 2u$.
(2) Multiply out.
(3) $u^2 \in H^4(S^2) = 0$.

$c_1(TS^2 \otimes \mathbb{C}) = 0$, as the torsion argument demands ($H^2(S^2;\mathbb{Z}) = \mathbb{Z}$ has no torsion, so the 2-torsion class c_1 must be zero). In the local complex coordinate z, the splitting is the familiar one: $T_{\mathbb{C}}S^2 = \operatorname{span}\{\partial_z\} \oplus \operatorname{span}\{\partial_{\bar{z}}\}$.

3.19.5 Realify, then complexify: $E_r \otimes \mathbb{C} = E \oplus \bar{E}$

Going the other way, a complex bundle E of rank k can be viewed as a real bundle E_r of rank $2k$ by forgetting that we can multiply by i. What is left of i is a real bundle map $J : E_r \to E_r$, $J(v) = iv$, with $J^2 = -1$. Complexifying E_r gives a complex bundle of rank $2k$, and Nash's (3.109) identifies it:

$$E_r \otimes_{\mathbb{R}} \mathbb{C} \;\cong\; E \oplus \bar{E}. \tag{3.178}$$

Nash derives this by writing $\mathbb{C} = \mathbb{R} \oplus i\mathbb{R}$. The cleanest way to see it is through the eigenvalues of J. Extend J complex-linearly to $E_r \otimes \mathbb{C}$. Since $J^2 = -1$, its eigenvalues are $\pm i$, and the fibre splits into the two eigenspaces

$$E^{1,0} = \{w : Jw = +iw\}, \qquad E^{0,1} = \{w : Jw = -iw\}.$$

Each vector $v \in E_r$ produces one vector in each:

$$J\big(v - iJv\big) \overset{(1)}{=} Jv - iJ^2v = Jv + iv \overset{(2)}{=} i\big(v - iJv\big),$$
$$J\big(v + iJv\big) \overset{(3)}{=} Jv + iJ^2v = Jv - iv = -i\big(v + iJv\big).$$

(1) J is complex-linear on $E_r \otimes \mathbb{C}$, so it passes the scalar i.
(2) $J^2 = -1$; factor out i.

(3) Same computation with the opposite sign.

The map $\phi_+(v) = \frac{1}{2}(v - iJv)$ from E to $E^{1,0}$ is complex-linear: multiplying v by i in E means applying J, and $\phi_+(Jv) = \frac{1}{2}(Jv + iv) = i\phi_+(v)$. The map $\phi_-(v) = \frac{1}{2}(v + iJv)$ has $\phi_-(Jv) = \frac{1}{2}(Jv - iv) = -i\phi_-(v)$: it is antilinear, that is, a linear isomorphism from $\bar{E}$ onto $E^{0,1}$. Hence $E_r \otimes \mathbb{C} = E^{1,0} \oplus E^{0,1} \cong E \oplus \bar{E}$.

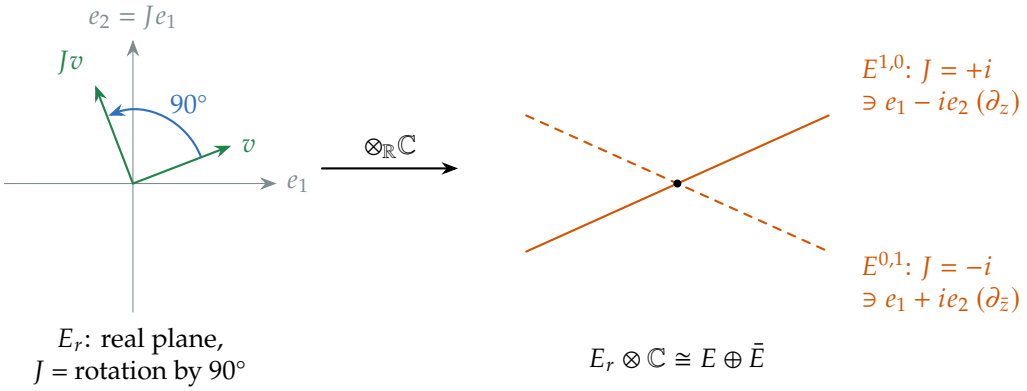


Figure 3.71: Realify, then complexify. Left: a complex line, seen as a real plane, remembers i only as the rotation J. Right: after allowing complex coefficients, J can be diagonalised; its $+i$ and $-i$ eigenlines are copies of E and of $\bar{E}$. For the tangent plane of a Riemann surface these are spanned by $\partial_z = \frac{1}{2}(\partial_x - i\partial_y)$ and $\partial_{\bar{z}}$, the splitting behind the (p,q)-forms of part 3a.

Worked example 3.113: J and its eigenvectors for $\mathbb{C}$ itself

Take $E = \mathbb{C}$, a trivial line. As a real plane with basis $e_1 = 1$, $e_2 = i$, multiplication by i is $J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$. Then

$$J\begin{pmatrix} 1 \\ -i \end{pmatrix} \overset{(1)}{=} \begin{pmatrix} i \\ 1 \end{pmatrix} = i\begin{pmatrix} 1 \\ -i \end{pmatrix}, \qquad J\begin{pmatrix} 1 \\ i \end{pmatrix} \overset{(2)}{=} \begin{pmatrix} -i \\ 1 \end{pmatrix} = -i\begin{pmatrix} 1 \\ i \end{pmatrix}.$$

(1) First row: $0 \cdot 1 + (-1)(-i) = i$; second row: $1 \cdot 1 + 0 = 1$.
(2) First row: $-i$; second row: 1.

So $e_1 - ie_2$ spans $E^{1,0}$ and $e_1 + ie_2$ spans $E^{0,1}$. With $e_1 = \partial_x$, $e_2 = \partial_y$ these are $2\partial_z$ and $2\partial_{\bar{z}}$.

3.19.6 Pontryagin classes

Working definition 3.46 (Pontryagin classes)

For a real vector bundle V of rank k, the ***i*-th Pontryagin class** is (Nash's (3.111))

$$p_i(V) = (-1)^i\, c_{2i}(V \otimes_{\mathbb{R}} \mathbb{C}) \;\in\; H^{4i}(X;\mathbb{Z}), \tag{3.179}$$

[only even Chern classes are used, because the odd ones of $V \otimes_{\mathbb{R}} \mathbb{C}$ satisfy $2c_{2i+1} = 0$ and vanish in real cohomology] and the total Pontryagin class is (Nash's (3.112))

$$p(V) = 1 + p_1(V) + \cdots + p_{[k/2]}(V). \tag{3.180}$$

Operationally: complexify, compute the even Chern classes, insert the signs. With an $\mathrm{O}(k)$ connection [its matrix A of 1-forms is real antisymmetric, so parallel transport preserves

the metric] whose curvature is the real antisymmetric matrix of 2-forms Ω,

$$p(\Omega) = \det\Big(\mathbb{1} + \frac{\Omega}{2\pi}\Big), \qquad p_1(\Omega) = -\frac{1}{8\pi^2}\operatorname{tr}(\Omega \wedge \Omega). \tag{3.181}$$

The top class is $p_{[k/2]}$, where $[k/2]$ is the integer part of $k/2$: the complexification has rank k, so $c_{2i}(V_{\mathbb{C}}) = 0$ once $2i > k$. As always, some classes can vanish just because X is too small: $p_i \in H^{4i}(X)$ is zero whenever $4i > \dim X$. On a 4-manifold only p_1 can be nonzero.

The curvature formula. An $\mathrm{O}(k)$ connection on V is also a connection on $V_{\mathbb{C}}$, with the same real antisymmetric curvature matrix Ω (anti-Hermitian, since real and antisymmetric). So $c(V_{\mathbb{C}}) = \det(\mathbb{1} + \frac{i}{2\pi}\Omega)$. The degree-$2j$ part carries $i^{2j} = (-1)^j$ times the corresponding coefficient of $\det(\mathbb{1} + \frac{\Omega}{2\pi})$, which cancels the sign in (3.179); this gives the first formula in (3.181). The odd terms vanish identically, as forms:

$$\det\Big(\mathbb{1} + \frac{\Omega}{2\pi}\Big) \overset{(1)}{=} \det\Big(\mathbb{1} + \frac{\Omega}{2\pi}\Big)^{T} \overset{(2)}{=} \det\Big(\mathbb{1} - \frac{\Omega}{2\pi}\Big),$$

(1) A determinant equals that of the transpose (the entries commute).
(2) $\Omega^T = -\Omega$.

and $\Omega \to -\Omega$ flips the sign of every odd-degree coefficient. For p_1, by (3.156) with $\operatorname{tr}\Omega = 0$,

$$p_1(\Omega) \overset{(1)}{=} -c_2(\Omega) \overset{(2)}{=} -\frac{1}{8\pi^2}\operatorname{tr}(\Omega \wedge \Omega) \overset{(3)}{=} \frac{1}{8\pi^2}\sum_{a,b} \Omega_{ab} \wedge \Omega_{ab},$$

(1) Definition (3.179) with $i = 1$.
(2) (3.156), and $\operatorname{tr}\Omega = 0$ for an antisymmetric matrix.
(3) $\operatorname{tr}(\Omega \wedge \Omega) = \sum_{a,b} \Omega_{ab} \wedge \Omega_{ba} = -\sum_{a,b} \Omega_{ab} \wedge \Omega_{ab}$.

For the Levi-Civita connection this is the gravitational Pontryagin density $\frac{1}{8\pi^2} R_{ab} \wedge R_{ab}$, which appears in the gravitational anomaly and in the signature theorem.

Whitney sums, modulo 2-torsion. Pontryagin classes are "not quite as well behaved" as Chern classes, Nash warns (3.113). Let us see exactly what goes wrong at the first step. Since $(V \oplus W)_{\mathbb{C}} = V_{\mathbb{C}} \oplus W_{\mathbb{C}}$,

$$\begin{aligned} p_1(V \oplus W) &\overset{(1)}{=} -c_2(V_{\mathbb{C}} \oplus W_{\mathbb{C}}) \\ &\overset{(2)}{=} -\big(c_2(V_{\mathbb{C}}) + c_1(V_{\mathbb{C}})c_1(W_{\mathbb{C}}) + c_2(W_{\mathbb{C}})\big) \\ &\overset{(3)}{=} p_1(V) + p_1(W) - c_1(V_{\mathbb{C}})\, c_1(W_{\mathbb{C}}). \end{aligned}$$

(1) Definition.
(2) Whitney for Chern classes, degree 4.
(3) Definition again for V and W.

The leftover $r = -c_1(V_{\mathbb{C}})c_1(W_{\mathbb{C}})$ satisfies $2r = 0$, because $2c_1(V_{\mathbb{C}}) = 0$. So

$$p(V \oplus W) = p(V)\, p(W) \quad \text{modulo elements of order 2 in } H^*(X;\mathbb{Z}), \tag{3.182}$$

which is Nash's (3.114). If $H^*(X;\mathbb{Z})$ has no torsion, or if we only use de Rham cohomology, the correction is invisible and p is honestly multiplicative. With connections it is automatic: the direct-sum curvature is block diagonal and $\det(\mathbb{1} + \Omega/2\pi)$ factorises.

Worked example 3.114: Pontryagin classes of spheres

Nash's example (3.115). The sphere $S^n \subset \mathbb{R}^{n+1}$ has at each point a tangent space and a normal line, together filling $\mathbb{R}^{n+1}$:

$$TS^n \oplus N(S^n) \cong I^{n+1}.$$

The normal bundle $N(S^n)$ is trivial: the outward unit normal $x \mapsto x$ is a nowhere-zero section, and a line bundle with a nowhere-zero section is trivial. The cohomology of S^n has no torsion, so Whitney holds exactly:

$$p(TS^n) \overset{(1)}{=} p(TS^n)\,p(I) \overset{(2)}{=} p\big(TS^n \oplus N(S^n)\big) \overset{(3)}{=} p(I^{n+1}) \overset{(4)}{=} 1.$$

(1) $p(I) = 1$: a trivial bundle has trivial complexification, whose Chern classes vanish (Example 3.87).
(2) Whitney (3.182), exact here, and $N(S^n) \cong I$.
(3) The isomorphism above.
(4) Trivial bundle again.

So $p_i(S^n) := p_i(TS^n) = 0$ for all $i > 0$. The cause is stable triviality: TS^n becomes trivial after adding one trivial line, and Pontryagin classes, like all classes built from Chern classes, cannot see trivial summands. Yet TS^2 is not trivial. The Pontryagin classes are blind to it; we need a class that is *not* stable, the Euler class.

A curvature check on the round S^4: in an orthonormal frame the Levi-Civita curvature of the unit sphere is $\Omega_{ab} = e^a \wedge e^b$ (constant curvature 1). Then

$$\sum_{a,b} \Omega_{ab} \wedge \Omega_{ab} = \sum_{a,b} e^a \wedge e^b \wedge e^a \wedge e^b = 0,$$

since each term repeats a 1-form. The Pontryagin *form* of the round metric vanishes pointwise, not just in cohomology.

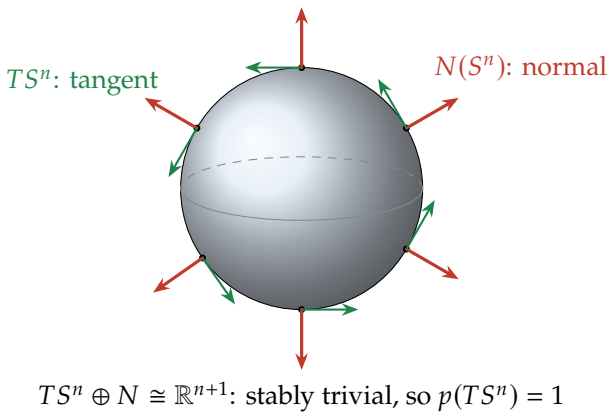


Figure 3.72: Why the Pontryagin classes of a sphere vanish. At each point the tangent space (green) and the normal line (red) together span $\mathbb{R}^{n+1}$, and the normal line is trivialised by the outward normal. Adding a trivial summand to TS^n makes it trivial, and Pontryagin classes cannot see trivial summands. (They are therefore also blind to the hairy-ball twisting of TS^2.)

Pontryagin classes of complex bundles. Nash extends the definition to a complex bundle E of rank k by applying it to the underlying real bundle: $p(E) \coloneqq p(E_r)$. Then

$$p_i(E) \overset{(1)}{=} (-1)^i\, c_{2i}(E_r \otimes \mathbb{C}) \overset{(2)}{=} (-1)^i\, c_{2i}(E \oplus \bar{E}), \tag{3.183}$$

(1) Definition (3.179) applied to E_r.
(2) (3.178).

which is Nash's (3.116). Now expand with Whitney and $c_j(\bar{E}) = (-1)^j c_j(E)$, Nash's (3.117):

$$c(E \oplus \bar{E}) = \big(1 + c_1 + c_2 + \cdots + c_k\big)\big(1 - c_1 + c_2 - \cdots + (-1)^k c_k\big). \tag{3.184}$$

The degree-$2i$ part is

$$c_{2i}(E \oplus \bar{E}) \overset{(1)}{=} \sum_{j=0}^{2i} c_j\,(-1)^{2i-j} c_{2i-j} \overset{(2)}{=} \sum_{j=0}^{2i} (-1)^j c_j\, c_{2i-j},$$

(1) Pick c_j from the first factor and $(-1)^{2i-j}c_{2i-j}$ from the second.
(2) $(-1)^{2i-j} = (-1)^j$.

and multiplying by $(-1)^i$ gives Nash's (3.118):

$$p_i(E) = \sum_{j=0}^{2i} (-1)^{i+j} c_j(E)\, c_{2i-j}(E) = c_i^2 - 2c_{i-1}c_{i+1} + 2c_{i-2}c_{i+2} - \cdots + (-1)^i\, 2c_{2i}. \tag{3.185}$$

(The terms j and $2i-j$ are equal, which produces the factors of 2; the middle term $j = i$ is c_i^2 with sign $(-1)^{2i} = +1$.) The first two cases are

$$p_1(E) = c_1^2 - 2c_2, \qquad p_2(E) = c_2^2 - 2c_1c_3 + 2c_4.$$

The splitting version. With Chern roots x_i of E, $\bar{E}$ has roots $-x_i$, so (Nash's (3.119))

$$\begin{aligned} c(E_r \otimes \mathbb{C}) &\overset{(1)}{=} \prod_{i=1}^{k} (1 + x_i)(1 - x_i) = \prod_{i=1}^{k} \big(1 - x_i^2\big), \\ p_j(E) &\overset{(2)}{=} \sigma_j\big(x_1^2, \ldots, x_k^2\big) : \qquad p_1 = \sum_i x_i^2, \quad p_2 = \sum_{i<j} x_i^2 x_j^2, \quad \ldots, \quad p_k = x_1^2 \cdots x_k^2. \end{aligned} \tag{3.186}$$

(1) (3.178) and the dual/conjugate rule for roots.
(2) The degree-$4j$ part of $\prod(1 - x_i^2)$ is $(-1)^j \sigma_j(x^2)$, and the sign cancels against $(-1)^j$ in the definition.

So the Pontryagin roots are the *squares* of the Chern roots. For a real bundle V of rank $2m$ or $2m+1$ the same holds with $\pm x_1, \ldots, \pm x_m$ (and a zero root if the rank is odd) as the roots of $V_{\mathbb{C}}$: an antisymmetric real matrix has eigenvalues in pairs $\pm i\lambda$.

Watch out: A typo in Nash's (3.119)

Nash writes $p_2(E) = \sum_{i,j} x_i^2 x_j^2$. The sum must run over $i < j$ (each unordered pair once, no $i = j$ terms), as the elementary symmetric function $\sigma_2(x^2)$ requires. With $\sum_{i,j}$ unrestricted one would get p_1^2 instead.

Worked example 3.115: Pontryagin classes of $\mathbb{CP}^n$ and of $\mathbb{CP}^2$

By the Euler sequence, $T\mathbb{CP}^n \oplus I = (n+1)\gamma^*$, so $T\mathbb{CP}^n$ has, "up to a zero root", the $n+1$ roots all equal to h. The Pontryagin roots are their squares, and

$$p(T\mathbb{CP}^n) = (1+h^2)^{n+1}. \tag{3.187}$$

For $\mathbb{CP}^2$: $p_1 = 3h^2$ and $\int_{\mathbb{CP}^2} p_1 = 3$. Check with (3.185): $p_1 = c_1^2 - 2c_2 = 9h^2 - 6h^2 = 3h^2$. ✓ This 3 will become the signature $\sigma(\mathbb{CP}^2) = 1$ after dividing by 3 (Section 3.20.9). For $\mathbb{CP}^4$: $p_1 = 5h^2$, $p_2 = 10h^4$.

Worked example 3.116: Instanton bundles as real bundles; the $\mathrm{SO}(3)$ bundle

Let $E \to S^4$ be an SU(2) bundle with $\int c_2 = k$. As a real rank-4 bundle,

$$p_1(E_r) \overset{(1)}{=} c_1^2 - 2c_2 \overset{(2)}{=} -2c_2, \qquad \int_{S^4} p_1(E_r) = -2k.$$

(1) (3.185) with $i = 1$.
(2) $c_1 = 0$ for SU(2).

The adjoint bundle $\mathrm{ad}E$ is the real rank-3 bundle of $\mathfrak{su}(2)$ values, whose complexification is $\mathrm{End}_0 E$. From Example 3.99, $c_2(\mathrm{End}_0 E) = 4c_2(E)$, so

$$p_1(\mathrm{ad}E) = -c_2(\mathrm{ad}E \otimes \mathbb{C}) = -4c_2(E), \qquad \int_{S^4} p_1(\mathrm{ad}E) = -4k.$$

This is how instanton numbers are phrased for SO(3) gauge theory, where p_1 of the adjoint bundle replaces c_2 (and can be any integer $\equiv w_2^2$ mod 4 on a general 4-manifold).

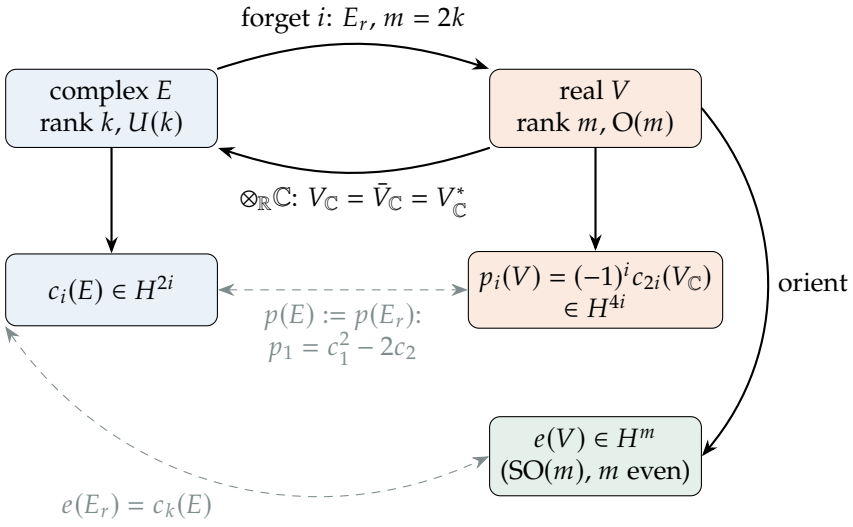


Figure 3.73: How the classes are related. A complex bundle has Chern classes; forgetting i makes it real, and a real bundle has Pontryagin classes, obtained by complexifying and taking the even Chern classes. Orientation (structure group $\mathrm{SO}(m)$, m even) adds the Euler class, the subject of the next section. The dashed arrows are the conversion formulae.

Reading Nash (pp. 82–86, (3.107)–(3.119))

Nash's "brief summary of a few useful properties" (p. 82) covers: the dual E^* with transition functions $(g^{-1})^T$ and the conjugate $\bar{E}$ with $\bar{g}$, equal for unitary g (footnote 6 on reduction to $U(k)$); line bundles as an abelian group with $L^{-1} = L^*$; real bundles with $V \cong V^*$, $\bar{V} = V$; complexification $V \otimes_{\mathbb{R}} \mathbb{C}$ with (3.107)–(3.108); realification and $E_r \otimes \mathbb{C} = E \oplus \bar{E}$ (3.109)–(3.110); Pontryagin classes (3.111)–(3.112); the Whitney formula up to 2-torsion (3.113)–(3.114); the sphere example (3.115), which rests on his (3.36), $TS^n \oplus N(S^n) \cong I^{n+1}$; Pontryagin classes of complex bundles (3.116)–(3.118); and the splitting form (3.119). We added the J-eigenspace derivation, the explanation of why odd $c_j(V_{\mathbb{C}})$ are 2-torsion, the curvature formula, and the examples $\mathbb{CP}^n$, ad$E$ and round S^4.

Recap

- E^*: transition $(g^{-1})^T$; $\bar{E}$: $\bar{g}$; for unitary g they agree. $c_j(E^*) = (-1)^j c_j(E)$.
- Line bundles form a group under $\otimes$; c_1 identifies it with $H^2(X;\mathbb{Z})$.
- Real V: $V \cong V^*$; $V_{\mathbb{C}} \cong \bar{V}_{\mathbb{C}}$, so odd $c_j(V_{\mathbb{C}})$ are 2-torsion.
- $E_r \otimes \mathbb{C} \cong E \oplus \bar{E}$: the $\pm i$ eigenspaces of J.
- $p_i(V) = (-1)^i c_{2i}(V_{\mathbb{C}}) \in H^{4i}$; $p(\Omega) = \det(1 + \Omega/2\pi)$; Whitney up to 2-torsion; $p(TS^n) = 1$; $p(T\mathbb{CP}^n) = (1 + h^2)^{n+1}$.
- For complex E: $p_1 = c_1^2 - 2c_2$, $p_2 = c_2^2 - 2c_1c_3 + 2c_4$; roots x_i^2.

3.20 The Euler class, Gauss–Bonnet–Chern, and the genera $\hat{A}$ and L

The last section ended with a puzzle. The tangent bundle of S^2 is twisted: the hairy-ball theorem says every vector field on S^2 has a zero, so TS^2 has no nowhere-zero section and cannot be trivial. Yet all its Pontryagin classes vanish (Example 3.114). The Pontryagin classes are *stable*: they cannot tell TS^2 from $TS^2 \oplus I$, which is trivial. We want a class that sees the hairy-ball twisting.

The missing ingredient is *orientation*. Nash puts it in one sentence (p. 86): on an orientable space, an oriented real bundle of even rank has one extra characteristic class, the Euler class. This section builds it from a single piece of linear algebra, the Pfaffian, computes it by curvature, and integrates it to get the Euler characteristic: the Gauss–Bonnet–Chern theorem. We then complete the toolbox of the index theorem with the two remaining generating-function classes, the $\hat{A}$-genus (Dirac operator) and the L-genus (signature).

3.20.1 Why orientation gives a new invariant polynomial

Chern–Weil theory (Section 3.16) turns every *invariant polynomial* of the structure group into a characteristic class. For $\mathrm{O}(k)$ the invariant polynomials of an antisymmetric matrix Ω are generated by the coefficients of $\det(\mathbb{1} + \Omega/2\pi)$: these are the Pontryagin forms. Why not the other coefficients, the odd ones? They vanish identically for antisymmetric matrices (Section 3.19.6).

Now shrink the group from $\mathrm{O}(k)$ to $\mathrm{SO}(k)$. That is what an **orientation** does: it lets us use only frames that are positively oriented, so the transition functions have $\det g = +1$. With fewer allowed transformations there can be more invariants. Is there a polynomial P with $P(g^{-1}\Omega g) = P(\Omega)$ for every $g \in \mathrm{SO}(k)$, but not for reflections? For even $k = 2m$ there is, and it

is essentially unique: the *square root of the determinant*.

Physical picture: The simplest case, $k = 2$

An antisymmetric 2×2 matrix is $\Omega = \begin{pmatrix} 0 & \omega \\ -\omega & 0 \end{pmatrix}$, an infinitesimal rotation of the plane with angular speed ω. Its determinant is ω^2. A rotation of the plane does not change the angular speed of another rotation, since SO(2) is abelian: ω itself is invariant. A reflection reverses the sense of rotation, $\omega \to -\omega$; only ω^2 survives. So SO(2) has the linear invariant ω, O(2) only the quadratic one ω^2. The Euler class of a rank-2 bundle is built from ω.

3.20.2 The Pfaffian

Working definition 3.47 (Pfaffian)

For an antisymmetric $2m \times 2m$ matrix A with commuting entries [numbers, or even-degree forms such as curvature 2-forms], the **Pfaffian** is

$$\begin{aligned} \mathrm{Pf}(A) &= \frac{1}{2^m\, m!} \sum_{\sigma \in S_{2m}} \operatorname{sgn}(\sigma)\, A_{\sigma(1)\sigma(2)} A_{\sigma(3)\sigma(4)} \cdots A_{\sigma(2m-1)\sigma(2m)} \\ &= \frac{1}{2^m\, m!}\, \varepsilon_{a_1 a_2 \ldots a_{2m}} A_{a_1 a_2} \cdots A_{a_{2m-1} a_{2m}}. \end{aligned} \tag{3.188}$$

[$\operatorname{sgn}(\sigma) = \pm 1$ for even or odd σ. For $m = 1$ this gives $\mathrm{Pf}(A) = A_{12}$, and always $\mathrm{Pf}(A)^2 = \det A$] *Operationally:* sum over all ways of splitting $\{1, \ldots, 2m\}$ into unordered pairs, each pairing contributing the product of the corresponding entries A_{ij} $(i < j)$ times the sign of the permutation that lists the pairs in order. There are $(2m-1)!! = 1 \cdot 3 \cdot 5 \cdots (2m-1)$ pairings.

The factor $2^m m!$ removes overcounting: swapping the two members of a pair (2 ways per pair, 2^m in all) and reordering the pairs ($m!$ ways) give the same term, because A is antisymmetric and the entries commute.

The first two cases, written out.

$$m = 1: \quad \mathrm{Pf}\begin{pmatrix} 0 & a \\ -a & 0 \end{pmatrix} = a, \tag{3.189}$$

$$m = 2: \quad \mathrm{Pf}(A) = A_{12}A_{34} - A_{13}A_{24} + A_{14}A_{23}. \tag{3.190}$$

For $m = 2$ the three pairings of $\{1,2,3,4\}$ are $\{12,34\}$, $\{13,24\}$ and $\{14,23\}$, and the permutations 1234, 1324, 1423 have signs $+, -, +$.

Worked example 3.117: $\mathrm{Pf}^2 = \det$ for a 4×4 matrix

Take the general antisymmetric matrix with entries $a = A_{12}$, $b = A_{13}$, $c = A_{14}$, $d = A_{23}$, $e = A_{24}$, $f = A_{34}$:

$$A = \begin{pmatrix} 0 & a & b & c \\ -a & 0 & d & e \\ -b & -d & 0 & f \\ -c & -e & -f & 0 \end{pmatrix}, \qquad \mathrm{Pf}(A) = af - be + cd.$$

Expand the determinant along the first row. Each 3×3 minor is quickly computed, and

one finds

$$\begin{aligned}\det A &\overset{(1)}{=} a^2f^2 + b^2e^2 + c^2d^2 - 2abef + 2acdf - 2bcde \\ &\overset{(2)}{=} (af - be + cd)^2 = \mathrm{Pf}(A)^2.\end{aligned}$$

(1) Cofactor expansion along the first row and collection of terms (tedious but mechanical; a computer algebra system confirms it).
(2) Multiply out $(af - be + cd)^2 = a^2f^2 + b^2e^2 + c^2d^2 - 2abef + 2acdf - 2bcde$.

A quick numerical check: $a = f = 1$, all others 0 is two independent rotations in the (12) and (34) planes; $\det A = 1$, $\mathrm{Pf}(A) = 1$.

A physicist's representation: Grassmann integrals. The cleanest way to prove everything about Pfaffians is the fermionic Gaussian integral. Let $\theta_1, \dots, \theta_{2m}$ be anticommuting (Grassmann) variables, with the Berezin rule $\int \mathrm{d}\theta_a\, \theta_a = 1$, $\int \mathrm{d}\theta_a\, 1 = 0$. Then

$$\int \mathrm{d}\theta_{2m} \cdots \mathrm{d}\theta_1 \; \exp\!\Big(\tfrac{1}{2}\,\theta^T A\,\theta\Big) = \mathrm{Pf}(A), \tag{3.191}$$

because only the term with every θ_a exactly once survives, and that term is the m-th term of the exponential series. Check for $m = 1$:

$$\begin{aligned}\tfrac{1}{2}\theta^T A\theta &\overset{(1)}{=} \tfrac{1}{2}\big(a\,\theta_1\theta_2 - a\,\theta_2\theta_1\big) \overset{(2)}{=} a\,\theta_1\theta_2, \\ \int \mathrm{d}\theta_2\mathrm{d}\theta_1\, e^{a\theta_1\theta_2} &\overset{(3)}{=} \int \mathrm{d}\theta_2\mathrm{d}\theta_1\,(1 + a\,\theta_1\theta_2) \overset{(4)}{=} a.\end{aligned}$$

(1) $A_{12} = a$, $A_{21} = -a$.
(2) $\theta_2\theta_1 = -\theta_1\theta_2$.
(3) $(\theta_1\theta_2)^2 = 0$, so the series stops.
(4) Do the inner integral first: $\int \mathrm{d}\theta_1\, \theta_1\theta_2 = \theta_2$, then $\int \mathrm{d}\theta_2\, \theta_2 = 1$.

The Dirac-fermion version, $\int \mathrm{d}\bar\psi\, \mathrm{d}\psi\, e^{\bar\psi M\psi} = \det M$, is familiar from the QFT path integral; (3.191) is its *Majorana* version. A Majorana (real) fermion gives Pf, a Dirac (complex) fermion, which is two Majoranas, gives det = Pf^2.

How the Pfaffian transforms. Change variables $\theta = B\eta$ with B an invertible real matrix. The Berezin measure transforms with the *inverse* Jacobian, $\mathrm{d}\theta_{2m} \cdots \mathrm{d}\theta_1 = (\det B)^{-1}\mathrm{d}\eta_{2m} \cdots \mathrm{d}\eta_1$. So

$$\begin{aligned}\mathrm{Pf}(A) &\overset{(1)}{=} \int \mathrm{d}^{2m}\theta\, e^{\frac{1}{2}\theta^T A\theta} \overset{(2)}{=} (\det B)^{-1} \int \mathrm{d}^{2m}\eta\, e^{\frac{1}{2}\eta^T (B^T AB)\eta} \overset{(3)}{=} (\det B)^{-1}\,\mathrm{Pf}(B^T AB), \\ &\Longrightarrow \quad \mathrm{Pf}(B^T AB) = \det(B)\,\mathrm{Pf}(A).\end{aligned} \tag{3.192}$$

(1) (3.191).
(2) Substitute $\theta = B\eta$: the exponent becomes $\frac{1}{2}\eta^T B^T AB\eta$, and the measure picks up $(\det B)^{-1}$ (Berezin integration is "differentiation", so Jacobians appear inverted).
(3) (3.191) again, for the matrix $B^T AB$.

This one formula contains the whole story:

- For $g \in \mathrm{SO}(2m)$, $g^{-1} = g^T$ and $\det g = 1$, so $\mathrm{Pf}(g^{-1}Ag) = \mathrm{Pf}(A)$: the Pfaffian is an $\mathrm{SO}(2m)$-invariant polynomial.

- For a reflection, $\det g = -1$ and Pf changes sign: it is *not* $\mathrm{O}(2m)$-invariant. Orientation is essential.
- Scaling: $\mathrm{Pf}(tA) = t^m \mathrm{Pf}(A)$, a homogeneous polynomial of degree m, half the degree of $\det A$.

Normal form. Every real antisymmetric matrix can be brought by some $g \in \mathrm{SO}(2m)$ to block-diagonal form,

$$g^T A g = \begin{pmatrix} 0 & \lambda_1 & & & \\ -\lambda_1 & 0 & & & \\ & & \ddots & & \\ & & & 0 & \lambda_m \\ & & & -\lambda_m & 0 \end{pmatrix}, \qquad \mathrm{Pf}(A) = \lambda_1 \lambda_2 \cdots \lambda_m, \qquad \det A = \lambda_1^2 \cdots \lambda_m^2. \tag{3.193}$$

Geometrically: an infinitesimal rotation of $\mathbb{R}^{2m}$ is a set of independent rotations in m orthogonal planes, with angular speeds λ_j. The determinant forgets the senses of rotation; the Pfaffian remembers their product (one sign can be flipped by flipping the orientation of one plane, which is why the overall orientation matters).

$\mathrm{Pf}^2 = \det$ in general. Bring A to normal form with $g \in \mathrm{SO}(2m)$. Then

$$\mathrm{Pf}(A)^2 \overset{(1)}{=} \mathrm{Pf}(g^T A g)^2 \overset{(2)}{=} (\lambda_1 \cdots \lambda_m)^2 \overset{(3)}{=} \det(g^T A g) \overset{(4)}{=} \det A,$$

(1) (3.192) with $\det g = 1$.
(2) (3.193): the Pfaffian of a block-diagonal matrix is the product of the 2×2 Pfaffians (3.189), since a pairing that joins two different blocks picks up a zero entry.
(3) The determinant of the block-diagonal matrix is $\prod_j \lambda_j^2$.
(4) $\det g^T = \det g = 1$.

Example 3.117 is the brute-force check of this for $m = 2$. The identity holds as a polynomial identity in the entries, so it also holds when the entries are commuting 2-forms, which is how we will use it.

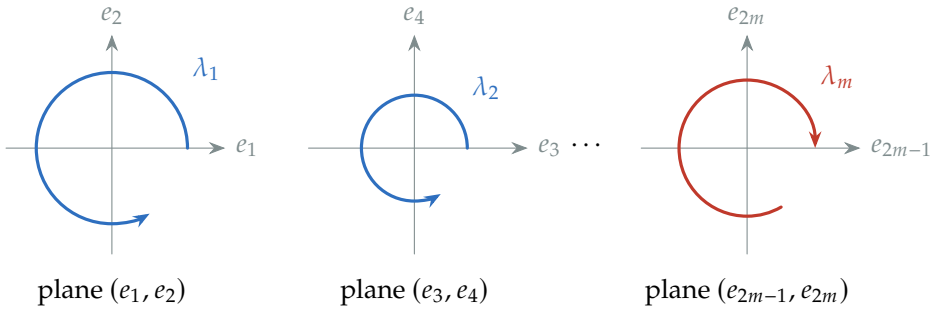


Figure 3.74: What the Pfaffian measures. An antisymmetric matrix is an infinitesimal rotation, and after an $\mathrm{SO}(2m)$ change of basis it rotates m orthogonal planes independently. The determinant is the product of the squared angular speeds. The Pfaffian is the signed product; reversing one plane (red, drawn clockwise) flips its sign, which is why it is only defined once an orientation is fixed.

3.20.3 The Euler class: definition and properties

Working definition 3.48 (Euler class)

Let $V \to X$ be an oriented real bundle [frames chosen so that every transition function has $\det g = +1$] of even rank $k = 2m$, with a metric and an $\mathrm{SO}(k)$ connection [a real antisymmetric matrix of 1-forms in each oriented orthonormal frame] whose curvature, in an oriented orthonormal frame, is the antisymmetric matrix of 2-forms $\Omega = (\Omega_{ab})$. The **Euler class** $e(V) \in H^k(X)$ is represented by

$$e(\Omega) = \mathrm{Pf}\Big(\frac{\Omega}{2\pi}\Big) = \frac{1}{(4\pi)^m\, m!}\, \varepsilon_{a_1 \dots a_{2m}}\, \Omega_{a_1 a_2} \wedge \cdots \wedge \Omega_{a_{2m-1} a_{2m}}. \tag{3.194}$$

Operationally: write the curvature in an oriented orthonormal frame and take its Pfaffian, treating the 2-forms as commuting numbers. For odd rank there is no Pfaffian and we set $e(V) = 0$ in real cohomology.

The second form of (3.194) comes from (3.188): $\frac{1}{2^m m!} \cdot \frac{1}{(2\pi)^m} = \frac{1}{(4\pi)^m m!}$. The two most used cases are

$$k = 2: \quad e = \frac{1}{2\pi}\, \Omega_{12}, \tag{3.195}$$

$$k = 4: \quad e = \frac{1}{32\pi^2}\, \varepsilon_{abcd}\, \Omega_{ab} \wedge \Omega_{cd} = \frac{1}{4\pi^2}\big(\Omega_{12} \wedge \Omega_{34} - \Omega_{13} \wedge \Omega_{24} + \Omega_{14} \wedge \Omega_{23}\big). \tag{3.196}$$

Why it is a characteristic class. Everything from Section 3.16 applies, because the argument there used only invariance under conjugation. Under a change of oriented orthonormal frame, $\Omega \mapsto g^{-1}\Omega g$ with $g \in \mathrm{SO}(k)$, and Pf does not change by (3.192); so $e(\Omega)$ is a globally defined k-form. The Bianchi identity makes it closed, and the transgression formula makes its class independent of the connection. Nash does not use curvature here; he defines $e(V)$ "up to a sign" by its square. The Pfaffian fixes the sign: it is the orientation.

The properties Nash lists. Each is now a one-line consequence.

1. *Square root of the top Pontryagin class* (Nash's (3.120)):

$$e(V)^2 \overset{(1)}{=} \mathrm{Pf}\Big(\frac{\Omega}{2\pi}\Big)^2 \overset{(2)}{=} \det\Big(\frac{\Omega}{2\pi}\Big) \overset{(3)}{=} p_m(V).$$

   (1) Definition (3.194); forms of even degree commute.
   (2) $\mathrm{Pf}^2 = \det$ for antisymmetric matrices.
   (3) The top-degree part of $p(\Omega) = \det(\mathbb{1} + \Omega/2\pi)$ (3.181) is $\det(\Omega/2\pi)$, of degree $2k = 4m$.

2. *Degree* (Nash's (3.121)): $e(V) \in H^k(X; \mathbb{Z})$ for rank k. With integer coefficients this needs the topological definition (the Thom class, Chapter 4); with forms we see $H^k(X; \mathbb{R})$.

3. *Whitney* (Nash's (3.122)): $e(V \oplus W) = e(V)\, e(W)$. With the direct sum connection the curvature is block diagonal, and the Pfaffian of a block-diagonal matrix is the product of the Pfaffians (the pairings never mix the blocks, since the off-block entries are zero).

4. *Complex bundles* (Nash's (3.123)): $e(E) := e(E_r)$, with E_r oriented by the complex structure, frames $(s_1, Js_1, \dots, s_k, Js_k)$.

5. *Top Chern class* (Nash's (3.124)): $e(E_r) = c_k(E)$. This is computed in Section 3.20.4 below.

6. *Orientation reversal* flips the sign: $e(\bar{V}) = -e(V)$, where $\bar{V}$ is V with the opposite orientation. This is (3.192) with a reflection.

7. *A nowhere-zero section kills it.* If V has a nowhere-zero section s, then $V = V' \oplus I$ with I spanned by $s/|s|$. Choose a connection that preserves this splitting (the direct sum of a connection on V' and the trivial one on I). In a frame whose last vector is $s/|s|$, the last row and column of Ω vanish; every pairing in the Pfaffian uses the last index, so $e(\Omega) = 0$ identically. In particular $e(W \oplus I) = 0$ for every W. This is the property the Pontryagin classes lacked: *e is not stable*.

Watch out: Odd rank

For odd k the reflection $v \mapsto -v$ has $\det = -1$ and reverses the orientation of V, yet it is an automorphism of V. So $e(V) = -e(V)$ and $2e(V) = 0$: the integral Euler class of an odd-rank bundle is at most 2-torsion, and it is zero in de Rham cohomology. Similarly $\chi(M) = 0$ for every closed odd-dimensional manifold.

3.20.4 The Euler class of a complex bundle is its top Chern class

Nash's argument (3.124) is: $e^2 = p_k$ and, for a complex bundle of rank k, $p_k(E) = x_1^2 \cdots x_k^2 = (x_1 \cdots x_k)^2 = c_k(E)^2$; so $e = \pm c_k$, and the sign "can be justified". With curvature we can see the sign directly.

Rank one. Let L be a Hermitian line bundle with unitary frame s and connection $\nabla s = s\,A$, $A = i\alpha$ with α a real 1-form (anti-Hermitian 1×1 matrix). The curvature is $F = i\beta$ with $\beta = \mathrm{d}\alpha$. As a real bundle L_r has the oriented orthonormal frame $(e_1, e_2) = (s, is)$. Then

$$\nabla e_1 \overset{(1)}{=} s\, i\alpha = e_2\, \alpha,$$
$$\nabla e_2 \overset{(2)}{=} i\, s\, i\alpha = -e_1\, \alpha,$$

(1) $\nabla s = s\, i\alpha$, and $is = e_2$.
(2) ∇ is complex-linear, so $\nabla(is) = i\nabla s$, and $i \cdot i = -1$.

In the convention $\nabla e_b = \sum_a e_a\, A^{\mathbb{R}}_{ab}$ this says $A^{\mathbb{R}}_{21} = \alpha$, $A^{\mathbb{R}}_{12} = -\alpha$:

$$A^{\mathbb{R}} = \begin{pmatrix} 0 & -\alpha \\ \alpha & 0 \end{pmatrix} = \alpha\, J, \qquad \Omega = \mathrm{d}A^{\mathbb{R}} = \begin{pmatrix} 0 & -\beta \\ \beta & 0 \end{pmatrix},$$

with J the rotation by $+90°$ of Example 3.113. So

$$e(L_r) \overset{(1)}{=} \frac{1}{2\pi}\,\Omega_{12} \overset{(2)}{=} -\frac{\beta}{2\pi} \overset{(3)}{=} \frac{i}{2\pi}\,(i\beta) \overset{(4)}{=} c_1(L).$$

(1) (3.195).
(2) Read off $\Omega_{12} = -\beta$.
(3) $i \cdot i = -1$.
(4) $c_1 = \frac{i}{2\pi} F$ with $F = i\beta$ (3.155).

The signs agree: with the complex orientation, $e = c_1$ exactly, as forms.

Any rank. By the splitting principle (or by diagonalising the curvature, Section 3.17.5) we may assume $F = \operatorname{diag}(i\beta_1, \dots, i\beta_k)$, a sum of line bundles with roots $x_j = \frac{i}{2\pi}(i\beta_j) = -\beta_j/2\pi$. The realified curvature is block diagonal with blocks $\begin{pmatrix} 0 & -\beta_j \\ \beta_j & 0 \end{pmatrix}$, so by the normal form (3.193),

$$e(E_r) \overset{(1)}{=} \prod_{j=1}^{k} \frac{-\beta_j}{2\pi} \overset{(2)}{=} x_1 x_2 \cdots x_k \overset{(3)}{=} c_k(E), \tag{3.197}$$

(1) The Pfaffian of a block-diagonal matrix is the product of the block Pfaffians, each being its $(1,2)$ entry $-\beta_j/2\pi$.
(2) $x_j = -\beta_j/2\pi$.
(3) $c_k = \sigma_k(x)$, the product of all roots (3.162).

and $e(E_r)^2 = \prod x_j^2 = p_k(E)$ reproduces Nash's (3.124). The same block computation also recovers the Pontryagin roots: each block contributes $\det \begin{pmatrix} 1 & -\beta_j/2\pi \\ \beta_j/2\pi & 1 \end{pmatrix} = 1 + x_j^2$ to $p = \det(\mathbb{1} + \Omega/2\pi)$.

Key idea: Three faces of one class

For a complex bundle E of rank k with Chern roots x_j:

$$e(E_r) = c_k(E) = x_1 \cdots x_k, \qquad p_k(E_r) = e(E_r)^2 = x_1^2 \cdots x_k^2.$$

For the tangent bundle of a complex n-manifold, $e(TX) = c_n(TX)$.

Worked example 3.118: Monopole bundles as real plane bundles

The monopole bundle $L_n \to S^2$ viewed as an oriented real plane bundle has $\int_{S^2} e(L_n) = \int c_1(L_n) = n$. What does n count? A *section*. For $n \geq 0$ the homogeneous polynomial z_1^n (a holomorphic section of $L_n = \mathcal{O}(n)$, Example 3.110) vanishes only at $[1:0]$, where in the local coordinate $w = z_1/z_0$ it looks like w^n: a zero of order n. A generic smooth perturbation, such as $\prod_{j=1}^{n}(z_1 - a_j z_0)$, splits it into n simple zeros, each where the section winds once around the origin of the fibre. The Euler number counts zeros of a section, with signs. For $n < 0$ the zeros must wind the other way, and there are $|n|$ of them.

3.20.5 Zeros of sections: the geometric meaning

Example 3.118 is the general rule. Let s be a section of an oriented rank-k bundle V over a closed oriented k-manifold M, with isolated zeros. Near a zero p, in a local trivialisation, s is a map $\mathbb{R}^k \to \mathbb{R}^k$ with $s(p) = 0$; restricted to a small sphere around p it gives a map $s/|s| : S^{k-1} \to S^{k-1}$, whose degree is the **index** of the zero, $\operatorname{index}_p(s)$. (For $k = 2$: the winding number of the arrow s as we walk once around p.) Then

$$\int_M e(V) = \sum_{s(p)=0} \operatorname{index}_p(s). \tag{3.198}$$

For $V = TM$ this is the **Poincaré–Hopf theorem**: the sum of the indices of the zeros of any vector field is a topological invariant, and it equals the Euler characteristic $\chi(M) = \sum_i (-1)^i b_i(M)$.

Why (3.198) holds (Chern's argument, in words). Away from the zeros, s picks out a unit vector $\hat{s} = s/|s|$, i.e. a nowhere-zero section, and a nowhere-zero section makes e *exact* there:

on the complement of the zeros one can write $e(\Omega) = d\Pi$ for a $(k-1)$-form Π built from $\hat{s}$ (for $k = 2$, Π is the connection 1-form in a frame whose first vector is $\hat{s}$, divided by 2π). Cut out small balls B_p around the zeros and use Stokes:

$$\int_M e \overset{(1)}{=} \lim_{\varepsilon\to 0} \int_{M\setminus \cup_p B_p} d\Pi \overset{(2)}{=} -\lim_{\varepsilon\to 0} \sum_p \int_{\partial B_p} \Pi \overset{(3)}{=} \sum_p \operatorname{index}_p(s).$$

(1) The balls have vanishing volume as $\varepsilon \to 0$, and $e = d\Pi$ outside them.
(2) Stokes; the boundary of $M \setminus \cup B_p$ is $\cup \partial B_p$ with the inward orientation, hence the minus sign.
(3) On a tiny sphere around p, $-\Pi$ integrates to the degree of $\hat{s}$ measured against a frame that extends across p. For $k = 2$: if ψ is the angle of $\hat{s}$ in such a frame, then in the frame $(\hat{s}, J\hat{s})$ one has $\Pi = -d\psi/2\pi$ up to terms that vanish as $\varepsilon \to 0$, and $\oint d\psi = 2\pi \operatorname{index}_p(s)$.

This is the Stokes-theorem mechanism of the whole subject: a closed form that is exact away from a few points integrates to a sum of local contributions at those points.

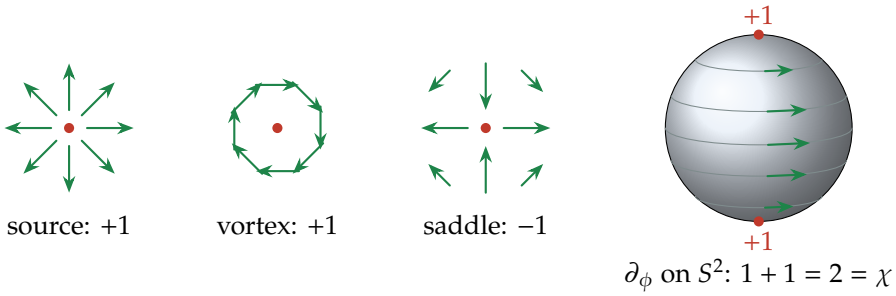


Figure 3.75: Indices of zeros of a vector field in the plane: a source and a vortex both wind once (+1), a saddle winds backwards (−1). Right: the rotation field ∂_ϕ on the sphere vanishes only at the two poles, each a vortex of index +1; the total 2 is $\chi(S^2) = \int_{S^2} e(TS^2)$. No vector field on S^2 avoids zeros, because the total must be 2, not 0.

3.20.6 The Gauss–Bonnet–Chern theorem

Put (3.194) and (3.198) together for $V = TM$, with the Levi-Civita connection:

Key idea: Gauss–Bonnet–Chern

For a closed oriented Riemannian manifold M of even dimension $2m$, with Levi-Civita curvature Ω_{ab} in an oriented orthonormal frame,

$$\chi(M) = \int_M e(TM) = \int_M \operatorname{Pf}\Big(\frac{\Omega}{2\pi}\Big) = \frac{1}{(4\pi)^m m!} \int_M \varepsilon_{a_1\dots a_{2m}} \Omega_{a_1a_2} \wedge \cdots \wedge \Omega_{a_{2m-1}a_{2m}}. \quad (3.199)$$

The left side is a count $(\sum(-1)^i b_i$, or vertices minus edges plus faces); the right side is an integral of curvature. Neither side depends on the metric.

Reminder: Cartan's structure equations. (See Chapter T2 and [Dray Ch. 18].) Choose an orthonormal coframe e^a, so the metric is $\sum_a (e^a)^2$. The Levi-Civita connection 1-forms $\omega^a{}_b = -\omega^b{}_a$ are fixed by the torsion-free condition, and the curvature 2-forms follow:

$$de^a = -\omega^a{}_b \wedge e^b, \qquad \Omega^a{}_b = d\omega^a{}_b + \omega^a{}_c \wedge \omega^c{}_b. \quad (3.200)$$

Indices are raised and lowered with δ_{ab}, so $\Omega_{ab} = \Omega^a{}_b$. On a surface, $\Omega_{12} = K\, e^1 \wedge e^2$ with K the Gaussian curvature.

Worked example 3.119: Gauss–Bonnet for the round S^2

On the sphere of radius R take $e^1 = R\,\mathrm{d}\theta$, $e^2 = R\sin\theta\,\mathrm{d}\phi$ (orientation $\mathrm{d}\theta \wedge \mathrm{d}\phi$, outward normal). Then

$$\mathrm{d}e^2 \overset{(1)}{=} R\cos\theta\,\mathrm{d}\theta \wedge \mathrm{d}\phi \overset{(2)}{=} -\big(\cos\theta\,\mathrm{d}\phi\big) \wedge \big(R\,\mathrm{d}\theta\big), \qquad \text{so} \quad \omega^2{}_1 = \cos\theta\,\mathrm{d}\phi,$$

$$\Omega^1{}_2 \overset{(3)}{=} \mathrm{d}\big(-\cos\theta\,\mathrm{d}\phi\big) \overset{(4)}{=} \sin\theta\,\mathrm{d}\theta \wedge \mathrm{d}\phi \overset{(5)}{=} \frac{1}{R^2}\, e^1 \wedge e^2.$$

(1) Differentiate $R\sin\theta\,\mathrm{d}\phi$.
(2) Reorder the wedge; compare with $\mathrm{d}e^2 = -\omega^2{}_1 \wedge e^1$. The other equation, $\mathrm{d}e^1 = 0 = -\omega^1{}_2 \wedge e^2$, holds because $\omega^1{}_2 \propto \mathrm{d}\phi$ and $e^2 \propto \mathrm{d}\phi$.
(3) $\omega^1{}_2 = -\omega^2{}_1$; the quadratic term $\omega \wedge \omega$ vanishes for a 2×2 antisymmetric matrix of 1-forms.
(4) $\mathrm{d}(-\cos\theta) = \sin\theta\,\mathrm{d}\theta$.
(5) $e^1 \wedge e^2 = R^2 \sin\theta\,\mathrm{d}\theta \wedge \mathrm{d}\phi$.

So $K = 1/R^2$, and

$$\int_{S^2} e(TS^2) \overset{(1)}{=} \frac{1}{2\pi}\int_{S^2} \Omega_{12} \overset{(2)}{=} \frac{1}{2\pi}\cdot\frac{1}{R^2}\cdot 4\pi R^2 \overset{(3)}{=} 2 = \chi(S^2).$$

(1) (3.195).
(2) $\Omega_{12} = K\,\mathrm{d}A$ and the area is $4\pi R^2$.
(3) $\chi(S^2) = V - E + F = 2$ (e.g. the tetrahedron: $4 - 6 + 4$).

The radius cancels, as it must: squash the sphere into an egg and the curvature moves around, but its total stays 4π. And since $TS^2 = T\mathbb{CP}^1 = L_2$ as a complex line bundle, $e = c_1(L_2)$ and $\int = 2$ again.

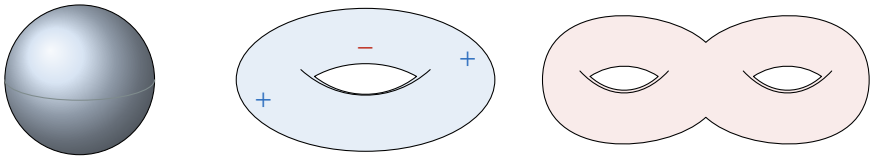
Worked example 3.120: Surfaces of every genus

Torus. The flat torus $\mathbb{R}^2/\mathbb{Z}^2$ has $e^1 = \mathrm{d}x$, $e^2 = \mathrm{d}y$, $\omega = 0, \Omega = 0$; so $\chi(T^2) = 0$. The embedded doughnut has $K > 0$ on the outside and $K < 0$ on the inside, and the two cancel exactly.
Genus $g \geq 2$. Such a surface carries a hyperbolic metric, $K = -1$. Then

$$2 - 2g \overset{(1)}{=} \chi(\Sigma_g) \overset{(2)}{=} \frac{1}{2\pi}\int_{\Sigma_g} K\,\mathrm{d}A \overset{(3)}{=} -\frac{\mathrm{Area}(\Sigma_g)}{2\pi} \quad\Longrightarrow\quad \mathrm{Area}(\Sigma_g) = 4\pi(g-1).$$

(1) Each handle lowers χ by 2 ([Dray §18.10]).
(2) Gauss–Bonnet, (3.199) with $m = 1$.
(3) $K = -1$.

Topology fixes the area of every hyperbolic surface. This is the fact behind the dimension counts of moduli spaces of Riemann surfaces in string theory (Chapter 9).



S^2: $K > 0$ $\chi = 2$ | T^2: + and − cancel $\chi = 0$ | Σ_2: mostly $K < 0$ $\chi = -2$

Figure 3.76: Gauss–Bonnet: $\frac{1}{2\pi}\int K\,\mathrm{d}A = \chi = 2 - 2g$. The sphere is positively curved everywhere. On the torus the outer (positive) and inner (negative) curvature cancel exactly. A surface with two handles must be negatively curved on balance, with total curvature -4π. However the surface is deformed, the totals do not change.

Worked example 3.121: Gauss–Bonnet–Chern on the round S^4

On the unit sphere S^n (constant curvature 1) the Levi-Civita curvature in any orthonormal frame is $\Omega_{ab} = e^a \wedge e^b$; on S^2 this is Example 3.119 with $R = 1$. For S^4:

$$\varepsilon_{abcd}\,\Omega_{ab}\wedge\Omega_{cd} \overset{(1)}{=} \varepsilon_{abcd}\,e^a\wedge e^b\wedge e^c\wedge e^d \overset{(2)}{=} 4!\;e^1\wedge e^2\wedge e^3\wedge e^4 = 24\,\mathrm{vol},$$

$$\int_{S^4} e(TS^4) \overset{(3)}{=} \frac{24}{32\pi^2}\,\mathrm{Vol}(S^4) \overset{(4)}{=} \frac{3}{4\pi^2}\cdot\frac{8\pi^2}{3} = 2.$$

(1) Constant curvature 1.
(2) Each of the 4! orderings (a,b,c,d) of $(1,2,3,4)$ gives $\varepsilon_{abcd}\,e^a\wedge e^b\wedge e^c\wedge e^d = \varepsilon^2_{abcd}\,e^1\wedge e^2\wedge e^3\wedge e^4$.
(3) (3.196).
(4) $\mathrm{Vol}(S^4) = \frac{8\pi^2}{3}$.

And indeed $\chi(S^4) = 2$ ($b_0 = b_4 = 1$, all others zero). Compare Example 3.114: on the same sphere the Pontryagin form vanishes identically, while the Euler form is a positive multiple of the volume form. The two classes see completely different things.

Try it: All even spheres

Show that $\int_{S^{2m}} e = 2$ for the unit sphere, using $\mathrm{Vol}(S^{2m}) = \frac{2^{2m+1}\pi^m\,m!}{(2m)!}$.
Answer: as in Example 3.121, $e = \frac{(2m)!}{(4\pi)^m m!}\mathrm{vol}$, so $\int e = \frac{(2m)!}{4^m\pi^m m!}\cdot\frac{2^{2m+1}\pi^m m!}{(2m)!} = \frac{2^{2m+1}}{2^{2m}} = 2$.

Worked example 3.122: Complex manifolds: $\chi = \int c_n$

For a complex n-manifold, $e(TX) = c_n(TX)$. For $\mathbb{CP}^n$, by (3.166),

$$\chi(\mathbb{CP}^n) = \int_{\mathbb{CP}^n} c_n(T\mathbb{CP}^n) = \binom{n+1}{n} = n+1,$$

which matches the cell decomposition $\mathbb{CP}^n = \mathbb{C}^0\cup\mathbb{C}^1\cup\cdots\cup\mathbb{C}^n$ (Figure 3.59): one even cell in each dimension $0,2,\ldots,2n$, so $b_0 = b_2 = \cdots = b_{2n} = 1$. For $S^2\times S^2$, Whitney for e gives $e(T(S^2\times S^2)) = e(TS^2)\times e(TS^2)$, so $\chi = 2\cdot 2 = 4$, again right ($b_0 = b_4 = 1$, $b_2 = 2$). For the K3 surface, $c_2 = 24$ and $\chi = 24$.

Physical picture: Why physicists meet the Euler class

(i) In two-dimensional gravity and string theory, the Einstein–Hilbert action $\frac{1}{4\pi}\int \sqrt{g}\,R$ is *topological*: since $R = 2K$ it equals $\chi(\Sigma)$, and $e^{-\lambda\chi}$ weights the genus expansion by g_s^{2g-2}. (ii) In four dimensions the Gauss–Bonnet–Chern density $\varepsilon\, R \wedge R$ is a total derivative locally, which is why the Gauss–Bonnet term does not change the classical equations of motion. (iii) The index of $\mathrm{d}+\mathrm{d}^*$ from even to odd forms is $\chi(M)$; in supersymmetric quantum mechanics it is the Witten index $\mathrm{Tr}(-1)^F e^{-\beta H}$, and the Grassmann-integral formula (3.191) is exactly how the Pfaffian of the curvature emerges from the fermion zero modes (Chapter 4).

Reading Nash (pp. 86–87, (3.120)–(3.124))

Nash introduces the Euler class for real bundles V of *even* rank over an *orientable* space: orientation reduces $\mathrm{O}(k)$ to $\mathrm{SO}(k)$, and for even k the determinant of an antisymmetric matrix has an $\mathrm{SO}(k)$-invariant square root, the Pfaffian (he cites Nash and Sen). He then defines $e(V)$ only up to sign by $e^2(V) = p_{k/2}(V)$ (3.120), so $e(V) \in H^k(X;\mathbb{Z})$ (3.121); states the Whitney rule $e(V\oplus W) = e(V)e(W)$ (3.122); defines $e(E) = e(E_r)$ for complex E (3.123); and shows $e(E) = c_k(E)$ from $p_k = x_1^2\cdots x_k^2$ (3.124), leaving the sign to Bott–Tu and Milnor–Stasheff. We have supplied the Pfaffian explicitly, fixed the sign with the curvature formula $e = \mathrm{Pf}(\Omega/2\pi)$, and added Poincaré–Hopf and Gauss–Bonnet–Chern with examples.

3.20.7 Genera from generating functions

We have now met several classes defined by a one-variable function of the roots. It is worth seeing them as one construction, because the index theorem uses two more of them, $\hat{A}$ and L.

Working definition 3.49 (Multiplicative class from a power series)

Let $Q(z) = 1 + a_1 z + a_2 z^2 + \dots$ be a formal power series [convergence is never needed, since forms above the dimension of X vanish].

- For a complex bundle E with Chern roots [formal classes with $c(E) = \prod_i (1 + x_i)$] x_i, the **multiplicative class** of Q is $Q(E) = \prod_i Q(x_i)$. It is symmetric in the roots, hence a polynomial in the c_j in each degree.
- If Q is *even*, $Q(z) = \tilde{Q}(z^2)$, then $\prod_i \tilde{Q}(x_i^2)$ is symmetric in the squares x_i^2, hence a polynomial in the Pontryagin classes. It therefore makes sense for any *real* bundle V, using the Pontryagin roots x_j^2 of (3.186).
- The **genus** of a closed manifold [compact, without boundary] M is $\int_M Q(TM)$.

Operationally: expand Q to the degree needed, multiply over roots, convert with the symmetric-function rules of Section 3.17; or with curvature, $Q(E) = \det Q\big(\frac{i}{2\pi}F\big)$.

The multiplicative property is automatic: $Q(E\oplus F) = Q(E)Q(F)$, because the roots of a sum are the union of the roots. Here is the catalogue.

class	$Q(z)$	first terms	index theorem it appears in
c	$1+z$	$1+c_1+c_2+\dots$	(top term) Euler characteristic
td	$\dfrac{z}{1-e^{-z}}$	$1+\frac{1}{2}c_1+\frac{1}{12}(c_1^2+c_2)$	Dolbeault (Riemann–Roch)
$\hat{A}$	$\dfrac{z/2}{\sinh(z/2)}$	$1-\frac{1}{24}p_1+\frac{1}{5760}(7p_1^2-4p_2)$	Dirac
L	$\dfrac{z}{\tanh z}$	$1+\frac{1}{3}p_1+\frac{1}{45}(7p_2-p_1^2)$	signature
ch (additive)	$\sum e^z$	$k+c_1+\frac{1}{2}(c_1^2-2c_2)$	twisting by a bundle

3.20.8 The $\hat{A}$ class

Working definition 3.50 ($\hat{A}$-class)

For a real bundle V with Pontryagin roots x_j^2 [formal classes with $p(V)=\prod_j(1+x_j^2)$] (for TM of a $2m$-manifold, $j=1,\dots,m$),

$$\hat{A}(V)=\prod_j \frac{x_j/2}{\sinh(x_j/2)}. \tag{3.201}$$

With curvature, $\hat{A}(\Omega)=\det{}^{1/2}\!\left(\dfrac{i\Omega/4\pi}{\sinh(i\Omega/4\pi)}\right)$. The $\hat{A}$-**genus** of a closed $4k$-manifold is $\hat{A}(M)=\int_M \hat{A}(TM)$.

Where does the curvature form come from? In normal form the curvature is made of blocks $\begin{pmatrix}0 & -\beta\\ \beta & 0\end{pmatrix}$, and the Hermitian matrix $\frac{i}{2\pi}\Omega$ has, on each block, the eigenvalues $\pm x$ with $x=-\beta/2\pi$ (the same x as in (3.197)). The function $\frac{z/2}{\sinh(z/2)}$ is even, so the pair $\pm x$ contributes the same factor twice to the determinant; the square root keeps one copy. The same reasoning gives $L(\Omega)=\det{}^{1/2}\!\left(\frac{i\Omega/2\pi}{\tanh(i\Omega/2\pi)}\right)$ for the L-class below.

The series. Put $y=x/2$. First, $\frac{\sinh y}{y}=1+\frac{y^2}{6}+\frac{y^4}{120}+\dots$. Its inverse $1+b_1y^2+b_2y^4$ satisfies

$$\begin{aligned} y^2:&\quad b_1+\tfrac{1}{6}=0 && \Rightarrow\ b_1=-\tfrac{1}{6},\\ y^4:&\quad b_2+\tfrac{1}{6}b_1+\tfrac{1}{120}=0 && \Rightarrow\ b_2=\tfrac{1}{36}-\tfrac{1}{120}=\tfrac{7}{360}. \end{aligned}$$

So $\frac{y}{\sinh y}=1-\frac{y^2}{6}+\frac{7y^4}{360}-\dots$, and with $y=x/2$,

$$\frac{x/2}{\sinh(x/2)}=1-\frac{x^2}{24}+\frac{7x^4}{5760}-\dots \tag{3.202}$$

(using $\frac{1}{6\cdot 4}=\frac{1}{24}$ and $\frac{7}{360\cdot 16}=\frac{7}{5760}$).

In Pontryagin classes. Multiply over the roots and collect by degree. With $p_1=\sum_j x_j^2$, $p_2=\sum_{i<j}x_i^2x_j^2$:

$$\hat{A}_1 \overset{(1)}{=} -\frac{1}{24}\sum_j x_j^2=-\frac{p_1}{24},$$

$$\hat{A}_2 \overset{(2)}{=} \frac{7}{5760}\sum_j x_j^4 + \frac{1}{576}\sum_{i<j} x_i^2 x_j^2$$
$$\overset{(3)}{=} \frac{7}{5760}(p_1^2 - 2p_2) + \frac{10}{5760}p_2 = \frac{7p_1^2 - 4p_2}{5760}.$$

(1) Degree 4: the x^2 term from one factor.
(2) Degree 8: the x^4 term from one factor, or the x^2 term from two different factors, with coefficient $(\frac{1}{24})^2 = \frac{1}{576}$.
(3) $\sum x_j^4 = p_1^2 - 2p_2$ (Newton, for the variables x_j^2), and $\frac{1}{576} = \frac{10}{5760}$.

This agrees with [Bär Ex. 4.3.6]. So on a 4-manifold, $\hat{A}(M) = -\frac{1}{24}\int_M p_1$, and on an 8-manifold $\hat{A}(M) = \frac{1}{5760}\int_M (7p_1^2 - 4p_2)$.

Relation to Todd. Nash's Todd series factorises (Section 3.18.6):

$$\frac{x}{1-e^{-x}} = e^{x/2}\cdot\frac{x/2}{\sinh(x/2)}.$$

For a complex manifold X, the Pontryagin roots of TX are the squares of the Chern roots, so taking the product over the Chern roots of TX,

$$\mathrm{td}(TX) = e^{c_1(TX)/2}\,\hat{A}(TX). \tag{3.203}$$

When c_1 is even, $e^{c_1/2}$ is the Chern character of a square root $K^{-1/2}$ of the anticanonical bundle. Physically: on a Kähler manifold, the Dirac operator is the Dolbeault operator twisted by $K^{1/2}$, and (3.203) is the statement that the two index formulae agree (Chapter 4).

Worked example 3.123: $\hat{A}$ of K3 and of $\mathbb{CP}^2$: integrality as a test

K3. A K3 surface is a compact complex surface with $c_1 = 0$ and $c_2 = 24$. Then

$$\hat{A}(\mathrm{K3}) \overset{(1)}{=} -\frac{1}{24}\int p_1 \overset{(2)}{=} -\frac{1}{24}\int (c_1^2 - 2c_2) \overset{(3)}{=} -\frac{1}{24}(0 - 48) = 2.$$

(1) $\hat{A}_1 = -p_1/24$.
(2) $p_1 = c_1^2 - 2c_2$ for a complex bundle (3.185).
(3) $c_1 = 0$, $\int c_2 = 24$.

An integer, as it must be: K3 is a spin manifold and $\hat{A}$ is the index of its Dirac operator. (By (3.203) it is also the Todd genus, since $c_1 = 0$: $\frac{1}{12}\int(c_1^2 + c_2) = 2$.)
$\mathbb{CP}^2$. Here $p_1 = 3h^2$ (Example 3.115), so $\hat{A}(\mathbb{CP}^2) = -\frac{3}{24} = -\frac{1}{8}$. Not an integer! So $\mathbb{CP}^2$ cannot carry a spin structure: if it did, $-\frac{1}{8}$ would be the index of a Dirac operator. This is the index theorem used as a no-go theorem.

Worked example 3.124: Zero modes in an instanton background: a preview

The twisted Dirac index is $\int_M \hat{A}(TM)\,\mathrm{ch}(E)$ (Chapter 4). On S^4 with the round metric, $p_1(TS^4) = 0$, so $\hat{A} = 1$. For the SU(2) bundle E with $\int c_2 = k$,

$$\int_{S^4} \hat{A}(TS^4)\,\mathrm{ch}(E) \overset{(1)}{=} \int_{S^4} 1\cdot(2 - c_2(E)) \overset{(2)}{=} -k.$$

(1) $\hat{A} = 1$, and Example 3.104.

(2) Only the 4-form part integrates.

So a Dirac fermion in the fundamental representation, in a charge-k instanton background, has $|k|$ more zero modes of one chirality than of the other (the chirality is fixed by orientation conventions, which Chapter 4 makes precise). These are the 't Hooft zero modes responsible for instanton-induced fermion interactions and, through the anomaly, for baryon-number violation.

Physical picture: Where sinh comes from: Landau levels

Put a unit-charge particle of unit mass in the plane, in a uniform magnetic field B, with $H = \frac{1}{2}(p - A)^2$. Its levels are the Landau levels $E_n = B(n + \frac{1}{2})$, each with $B/2\pi$ states per unit area. The heat trace per unit area is therefore

$$\frac{B}{2\pi}\sum_{n\geq 0} e^{-tB(n+\frac{1}{2})} = \frac{B}{2\pi}\cdot\frac{1}{2\sinh(tB/2)},$$

while for $B = 0$ it is $\frac{1}{2\pi t}$. The ratio is $\frac{tB/2}{\sinh(tB/2)}$: exactly the $\hat{A}$ factor, with tB in place of x. In the heat-kernel proof of the index theorem (Chapter 4) the curvature of each tangent 2-plane acts on spinors like a magnetic field, and this Landau-level (Mehler) factor is where $\hat{A}$ comes from. In anomaly language, the term $\hat{A}_1 = -p_1/24 = \frac{1}{192\pi^2}\operatorname{tr}(\Omega\wedge\Omega)$ is the gravitational contribution to the index density, and hence to the chiral anomaly (Chapter 10).

3.20.9 The L class and the signature

The signature. On a closed oriented $4k$-manifold M the middle cohomology carries the **intersection form**

$$Q(\alpha,\beta) = \int_M \alpha\wedge\beta, \qquad \alpha,\beta\in H^{2k}(M;\mathbb{R}).$$

It is symmetric (2-forms of even degree $2k$ commute) and non-degenerate (Poincaré duality). Its **signature** $\sigma(M) = b^+ - b^-$ is the number of positive minus negative eigenvalues. Examples, with matrices in a basis of H^2:

M	H^2 basis	matrix of Q	σ
$\mathbb{CP}^2$	h	(1)	1
$\overline{\mathbb{CP}^2}$ (reversed orientation)	h	(-1)	-1
$S^2\times S^2$	u_1, u_2	$\begin{pmatrix}0&1\\1&0\end{pmatrix}$	0
S^4	none	–	0
K3	22 classes	$3\begin{pmatrix}0&1\\1&0\end{pmatrix}\oplus 2(-E_8)$	-16

For $S^2\times S^2$: $u_1^2 = u_2^2 = 0$ (each is pulled back from a 2-sphere) and $\int u_1u_2 = 1$; the eigenvalues of the matrix are ± 1.

Working definition 3.51 (L-class)

The **Hirzebruch L-class** of a real bundle with Pontryagin roots x_j^2 is

$$L(V) = \prod_j \frac{x_j}{\tanh x_j}, \tag{3.204}$$

[each factor is even in x_j, so $L(V)$ is a polynomial in the Pontryagin classes] and the **L-genus** of a closed $4k$-manifold is $\int_M L(TM)$ [Hirzebruch's theorem says this number equals the signature $\sigma(M)$].

The series. $\frac{\tanh y}{y} = 1 - \frac{y^2}{3} + \frac{2y^4}{15} - \dots$, and its inverse $1 + b_1 y^2 + b_2 y^4$ has

$$\begin{aligned} y^2: &\quad b_1 - \tfrac{1}{3} = 0 &&\Rightarrow\ b_1 = \tfrac{1}{3},\\ y^4: &\quad b_2 - \tfrac{1}{3}b_1 + \tfrac{2}{15} = 0 &&\Rightarrow\ b_2 = \tfrac{1}{9} - \tfrac{2}{15} = -\tfrac{1}{45}. \end{aligned}$$

So

$$\frac{x}{\tanh x} = 1 + \frac{x^2}{3} - \frac{x^4}{45} + \dots \tag{3.205}$$

Multiplying over roots, exactly as for $\hat{A}$:

$$L_1 = \frac{1}{3}\sum_j x_j^2 = \frac{p_1}{3},$$

$$L_2 \overset{(1)}{=} -\frac{1}{45}\sum_j x_j^4 + \frac{1}{9}\sum_{i<j} x_i^2 x_j^2 \overset{(2)}{=} -\frac{1}{45}(p_1^2 - 2p_2) + \frac{5}{45}p_2 = \frac{7p_2 - p_1^2}{45}.$$

(1) The x^4 term from one factor, or x^2 from two factors, $(\frac{1}{3})^2 = \frac{1}{9}$.
(2) $\sum x_j^4 = p_1^2 - 2p_2$ and $\frac{1}{9} = \frac{5}{45}$.

Key idea: Hirzebruch signature theorem

For a closed oriented $4k$-manifold, $\sigma(M) = \int_M L(TM)$. In four dimensions, $\sigma(M) = \frac{1}{3}\int_M p_1(TM)$.

Nash does not state this here; it is the index theorem for the signature operator $d + d^*$ acting from self-dual to anti-self-dual forms ([Bär Thm. 4.3.9], and Chapter 4). Let us test it.

Worked example 3.125: Checking the signature theorem

$\mathbb{CP}^2$. $p_1 = 3h^2$, so $\frac{1}{3}\int p_1 = 1 = \sigma(\mathbb{CP}^2)$. ✓ (This is the "3 becomes 1" promised in Example 3.115.)
$S^2\times S^2$ *and* S^4. Both have $p_1 = 0$: $p(S^4) = 1$ by Example 3.114, and $p(S^2\times S^2) = p(S^2)\,p(S^2) = 1$ by Whitney (the tangent bundle of a product is the sum of the pulled-back tangent bundles). Both have $\sigma = 0$. ✓
$\mathbb{CP}^4$. $p(T\mathbb{CP}^4) = (1+h^2)^5$, so $p_1 = 5h^2$, $p_2 = 10h^4$, and

$$\int_{\mathbb{CP}^4} L_2 \overset{(1)}{=} \frac{7\cdot 10 - 5^2}{45} \overset{(2)}{=} \frac{45}{45} = 1.$$

(1) $L_2 = \frac{7p_2 - p_1^2}{45}$ and $\int h^4 = 1$.

(2) Arithmetic: $70 - 25 = 45$.

And $\sigma(\mathbb{CP}^4) = 1$: H^4 is spanned by h^2, with $\int h^2 \cdot h^2 = 1$. ✓

K3. $p_1 = -48$ (Example 3.123), so $\frac{1}{3}\int p_1 = -16 = \sigma(\mathrm{K3})$. ✓

Worked example 3.126: Why the L series is forced: all $\mathbb{CP}^{2k}$ at once

Hirzebruch found L by demanding $\int_{\mathbb{CP}^{2k}} L = \sigma(\mathbb{CP}^{2k}) = 1$ for every k. Let us verify it for all k with a residue, exactly as for the Todd genus in Example 3.109. By the Euler sequence, $L(T\mathbb{CP}^{2k}) = \left(\frac{h}{\tanh h}\right)^{2k+1}$, and we need its h^{2k} coefficient:

$$\int_{\mathbb{CP}^{2k}} L \overset{(1)}{=} \frac{1}{2\pi i}\oint \left(\frac{h}{\tanh h}\right)^{2k+1}\frac{\mathrm{d}h}{h^{2k+1}} = \frac{1}{2\pi i}\oint \frac{\mathrm{d}h}{(\tanh h)^{2k+1}}$$
$$\overset{(2)}{=} \frac{1}{2\pi i}\oint \frac{\mathrm{d}w}{w^{2k+1}(1-w^2)} \overset{(3)}{=} 1.$$

(1) Coefficient extraction by a contour integral around $h = 0$.

(2) Substitute $w = \tanh h$, $\mathrm{d}w = (1 - \tanh^2 h)\,\mathrm{d}h = (1 - w^2)\,\mathrm{d}h$.

(3) The coefficient of w^{2k} in $\frac{1}{1-w^2} = \sum_j w^{2j}$ is 1.

The Pontryagin classes of $\mathbb{CP}^{2k}$ determine the coefficients of Q one at a time, so this computation, together with multiplicativity of σ under products and its invariance under cobordism, pins down $x/\tanh x$ uniquely.

Worked example 3.127: Rokhlin-type constraint from $\hat{A}$ and L

On a 4-manifold, $\hat{A} = -\frac{1}{24}\int p_1$ and $\sigma = \frac{1}{3}\int p_1$, so

$$\hat{A}(M) = -\frac{\sigma(M)}{8}.$$

If M is spin, $\hat{A}(M)$ is an index, hence an integer, so $\sigma(M)$ is a multiple of 8. (A finer argument, using the quaternionic structure of spinors in four dimensions, gives Rokhlin's theorem: $\sigma \equiv 0 \bmod 16$.) K3: $\sigma = -16$, $\hat{A} = 2$. $\mathbb{CP}^2$: $\sigma = 1$, not divisible by 8, so not spin, as found in Example 3.123.

Recap

- Orientation reduces $\mathrm{O}(2m)$ to $\mathrm{SO}(2m)$ and allows a new invariant, the Pfaffian: $\mathrm{Pf}(B^T AB) = \det B\,\mathrm{Pf}(A)$, $\mathrm{Pf}^2 = \det$; $\mathrm{Pf}(A) = \int \mathrm{d}\theta\, e^{\theta^T A\theta/2}$.
- $e(V) = \mathrm{Pf}(\Omega/2\pi) \in H^{2m}$; $e^2 = p_m$; $e(V \oplus W) = e(V)e(W)$; $e(E_r) = c_k(E)$; e vanishes if there is a nowhere-zero section.
- $\int_M e(V)$ counts zeros of a section with indices; for TM: Poincaré–Hopf and Gauss–Bonnet–Chern, $\chi(M) = \int \mathrm{Pf}(\Omega/2\pi)$. Checks: S^2, T^2, Σ_g, S^4, $\mathbb{CP}^n$.
- Genera: $\mathrm{td} = \prod \frac{x}{1-e^{-x}}$, $\hat{A} = \prod \frac{x/2}{\sinh(x/2)} = 1 - \frac{p_1}{24} + \frac{7p_1^2 - 4p_2}{5760}$, $L = \prod \frac{x}{\tanh x} = 1 + \frac{p_1}{3} + \frac{7p_2 - p_1^2}{45}$; $\mathrm{td} = e^{c_1/2}\hat{A}$.
- $\sigma = \int L$; $\hat{A}(\mathrm{K3}) = 2$, $\hat{A}(\mathbb{CP}^2) = -\frac{1}{8}$ (not spin); on 4-manifolds $\hat{A} = -\sigma/8$.

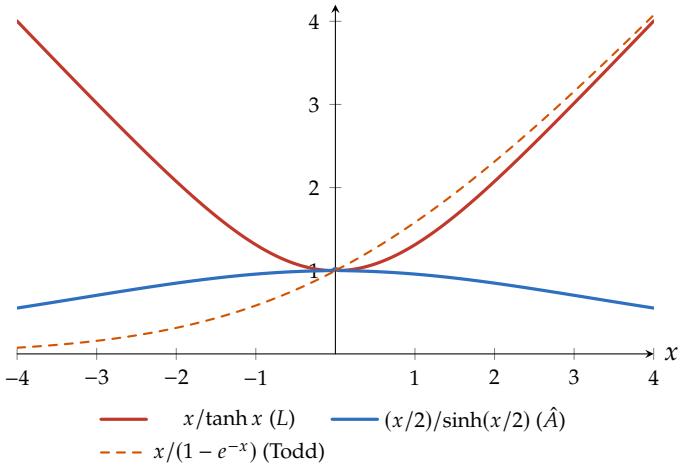


Figure 3.77: The generating functions of the three genera of the index theorem. The L and $\hat{A}$ series are even, so they are functions of x^2 and make sense for real bundles (through Pontryagin classes). The Todd series is not even: it is $e^{x/2}$ times the $\hat{A}$ series, which tilts the blue curve into the orange one; this is td $= e^{c_1/2}\hat{A}$. Near $x = 0$: $L \approx 1 + x^2/3$ bends up, $\hat{A} \approx 1 - x^2/24$ bends down gently.

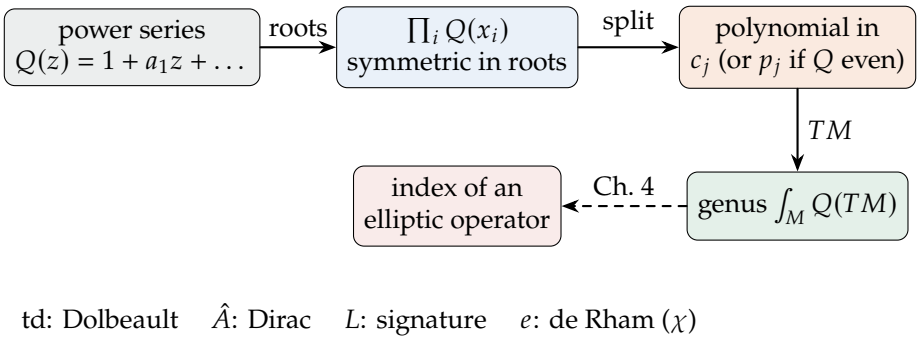


Figure 3.78: The genus machine. A power series fixes a multiplicative class through the roots; the splitting principle rewrites it in Chern or Pontryagin classes; integrating over M gives a number which, for the right series, is the index of a natural elliptic operator.

3.21 Fredholm operators and $K(X)$

Here is a question every field theorist meets sooner or later. A Dirac operator in an instanton background has some zero modes. How many? If we deform the gauge field a little, the zero modes may move away from zero, or new ones may appear. So the *number* of zero modes is not a robust quantity. But something is robust: in an instanton background of charge k the number of left-handed zero modes *minus* the number of right-handed zero modes is fixed by k, whatever we do. That difference controls the 't Hooft vertex, the axial anomaly and the fermion-number violation in the electroweak theory.

The mathematical name for "zero modes of D minus zero modes of $D^\dagger$" is the **index** of D, and the operators for which it makes sense are the **Fredholm operators**. Nash closes his chapter with a one-paragraph section, §6, that says something remarkable about them: *the space of all Fredholm operators is itself a machine for producing K-theory.* A family of Fredholm operators parametrised by a space X has an index that is not a number but an element of $K(X)$, and every element of $K(X)$ arises this way. This is the bridge from the K-theory of the previous sections to the index theory of Chapter 4.

Nash's paragraph uses words (Hilbert space, Fredholm, cokernel, homotopy classes of maps into an operator space) that he does not define. We supply them all. The plan is:

(a) the index of a matrix (finite dimensions: the index exists but is boring);

(b) Hilbert spaces, bounded and compact operators (the background);

(c) Fredholm operators and a gallery of examples: shifts, the harmonic oscillator, Toeplitz operators;

(d) why the index is stable (the heart of the matter);

(e) families of Fredholm operators and the index bundle;

(f) the Atiyah–Jänich theorem $[X, \mathcal{F}] \cong K(X)$ and Kuiper's theorem;

(g) elliptic operators as the source of Fredholm families, and a look ahead to the index theorem.

3.21.1 Warm-up: the index of a matrix

Take an $m \times n$ complex matrix A, viewed as a linear map $A : \mathbb{C}^n \to \mathbb{C}^m$. Two subspaces measure how far A is from being invertible:

- the **kernel** $\operatorname{Ker} A = \{v \in \mathbb{C}^n : Av = 0\}$, the solutions of the homogeneous equation (the "zero modes");
- the **cokernel** $\operatorname{Coker} A = \mathbb{C}^m / \operatorname{Im} A$, which measures how many right-hand sides b the equation $Av = b$ *cannot* reach.

The **index** is the difference of their dimensions. By rank–nullity,

$$\begin{aligned} \operatorname{index} A &\overset{(1)}{=} \dim \operatorname{Ker} A - \dim \operatorname{Coker} A \\ &\overset{(2)}{=} (n - \operatorname{rank} A) - (m - \operatorname{rank} A) \\ &= n - m. \end{aligned} \tag{3.206}$$

(1) The definition of the index.

(2) Rank–nullity: $\dim \operatorname{Ker} A = n - \dim \operatorname{Im} A$; and $\dim \operatorname{Coker} A = m - \dim \operatorname{Im} A$, since a quotient has dimension "big space minus subspace". Both use $\dim \operatorname{Im} A = \operatorname{rank} A$.

The rank cancels. The index does not depend on A at all, only on the dimensions of the two spaces. It is the difference "number of unknowns minus number of equations".

Worked example 3.128: A matrix whose kernel jumps, but whose index does not

Consider the family of 2×3 matrices

$$A_\varepsilon = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 6+\varepsilon \end{pmatrix} : \mathbb{C}^3 \to \mathbb{C}^2.$$

At $\varepsilon = 0$. The second row is twice the first, so the rank is 1. The kernel is the plane $v_1 + 2v_2 + 3v_3 = 0$, of dimension 2. The image is the line spanned by $(1,2)^T$, so the cokernel $\mathbb{C}^2/\operatorname{Im} A_0$ has dimension 1. Index $2 - 1 = 1$.
At $\varepsilon \neq 0$. The columns $(1,2)^T$ and $(3, 6+\varepsilon)^T$ are independent, so the rank is 2, the image is all of $\mathbb{C}^2$, the cokernel is 0, and the kernel is a line (spanned by $(-2,1,0)^T$). Index $1 - 0 = 1$.
As $\varepsilon \to 0$ one extra zero mode appears, and at the same moment one extra cokernel direction appears. They come and go *in pairs*, so the index $3 - 2 = 1$ never moves. See Figure 3.79.

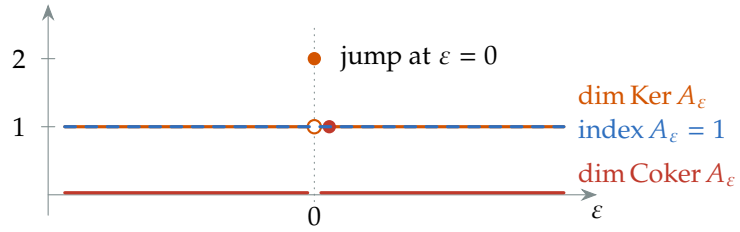


Figure 3.79: The family of Example 3.128. The dimension of the kernel (orange) jumps from 1 to 2 at $\varepsilon = 0$; at the same point the cokernel (red) jumps from 0 to 1. Their difference (blue, dashed) is constant. This pairing of jumps is the whole mechanism behind the stability of the index.

There is a second way to count the cokernel that generalises better. Use the standard inner product $\langle u, v \rangle = \sum_i \bar{u}_i v_i$ (antilinear in the first slot, the physicists' convention, used throughout this section). The cokernel is represented by the vectors orthogonal to the image, and those are the zero modes of the adjoint:

$$\begin{aligned} w \perp \operatorname{Im} A &\overset{(1)}{=} \langle w, Av \rangle = 0 \text{ for all } v \\ &\overset{(2)}{=} \langle A^\dagger w, v \rangle = 0 \text{ for all } v \\ &\overset{(3)}{=} A^\dagger w = 0. \end{aligned}$$

(1) Definition of "orthogonal to every vector of the image". (Each "=" in this chain means "is equivalent to".)
(2) Definition of the adjoint $A^\dagger = \bar{A}^T$.
(3) A vector orthogonal to every v is zero (take $v = A^\dagger w$).

So $\operatorname{Coker} A \cong \operatorname{Ker} A^\dagger$ and

$$\operatorname{index} A = \dim \operatorname{Ker} A - \dim \operatorname{Ker} A^\dagger. \tag{3.207}$$

This is the physicist's form: *zero modes of A minus zero modes of $A^\dagger$*.

Key idea: What finite dimensions teach us

Even for matrices, dim Ker and dim Coker jump under perturbation, but only together, so the index is constant. In finite dimensions the constant is the uninteresting number $n - m$. In infinite dimensions "$n - m$" is "$\infty - \infty$", and the index becomes a genuine invariant of the operator.

3.21.2 Background: Hilbert spaces and bounded operators

Nash takes for granted the functional analysis of operators on Hilbert space. Here is the minimum we need, stated operationally.

Working definition 3.52 (Hilbert space)

A **Hilbert space** H is a complex vector space with an inner product $\langle \cdot, \cdot \rangle$ in which every Cauchy sequence [a sequence whose terms get arbitrarily close to each other] converges (it is *complete*). We only need *separable* ones, which have a countable orthonormal basis $e_0, e_1, e_2, \dots$ [every vector is a convergent infinite sum of the e_n, not only a finite combination] Then every vector is a square-summable sequence of components:

$$v = \sum_{n \geq 0} v_n e_n, \qquad \|v\|^2 = \sum_{n \geq 0} |v_n|^2 < \infty.$$

So every separable infinite-dimensional Hilbert space "is" the sequence space $\ell^2 = \ell^2(\mathbb{N})$ in a chosen basis. Three incarnations will appear repeatedly:

- ℓ^2 itself, with basis vectors $e_n = (0, \dots, 0, 1, 0, \dots)$;
- $L^2(S^1)$, with the Fourier basis $e^{in\theta}/\sqrt{2\pi}$, $n \in \mathbb{Z}$ (a basis indexed by $\mathbb{Z}$ rather than $\mathbb{N}$; relabelling $0, 1, -1, 2, -2, \dots$ turns it into ℓ^2);
- the Fock space of a harmonic oscillator, with basis $|n\rangle$, $n = 0, 1, 2, \dots$

Working definition 3.53 (bounded operator)

A linear map $T : H \to H'$ is **bounded** if $\|T\| = \sup_{v \neq 0} \|Tv\| / \|v\| < \infty$. Bounded is the same as continuous [$v_t \to v$ implies $Tv_t \to Tv$]. The bounded operators form a vector space $B(H, H')$ with the **norm topology**: $T_t \to T$ means $\|T_t - T\| \to 0$. Every bounded T has an adjoint $T^* : H' \to H$ (written $T^\dagger$ by physicists), defined by $\langle T^* w, v \rangle = \langle w, Tv \rangle$.

Watch out: Differential operators are not bounded on L^2, but that is fine

$-i\,\mathrm{d}/\mathrm{d}\theta$ sends $e^{in\theta}$ to $n\,e^{in\theta}$, and n is unbounded, so it is not a bounded operator $L^2 \to L^2$. The cure is to change the domain: as a map from the Sobolev space $H^1(S^1)$ (norm $\|u\|_1^2 = \sum (1 + n^2)\,|u_n|^2$) to $L^2(S^1)$ it is bounded, with norm ≤ 1. Nash's phrase "elliptic operators are Fredholm operators when viewed as operators in Hilbert space" always means "between suitable Sobolev spaces"; see Chapter 2. With this understanding everything below applies to differential operators.

Two things go wrong in infinite dimensions that cannot go wrong for matrices.

- *Injective does not imply surjective.* The shift below is injective and misses a direction.

- *The image need not be closed.* Then $H'/\operatorname{Im} T$ is a wild, infinite-dimensional space with no good topology. Fredholm operators are exactly the ones for which neither pathology is serious.

Working definition 3.54 (compact operator)

An operator K is **compact** if it is a norm limit [$\|K - K_N\| \to 0$ for a sequence K_N] of *finite-rank* operators (operators whose image is finite-dimensional). *Operational test:* in a suitable basis, K acts nontrivially only on a finite number of modes, up to an error that can be made as small as we like. The model is a diagonal operator $Ke_n = \lambda_n e_n$ with $\lambda_n \to 0$.

Why is $Ke_n = \lambda_n e_n$ with $\lambda_n \to 0$ compact? Truncate it: $K_N e_n = \lambda_n e_n$ for $n < N$ and $K_N e_n = 0$ for $n \geq N$. Then K_N has rank $\leq N$ and

$$\|K - K_N\| \overset{(1)}{=} \sup_{v \neq 0} \frac{\left(\sum_{n \geq N} |\lambda_n|^2 \, |v_n|^2\right)^{1/2}}{\left(\sum_n |v_n|^2\right)^{1/2}} \overset{(2)}{=} \sup_{n \geq N} |\lambda_n| \xrightarrow{N \to \infty} 0.$$

(1) $K - K_N$ keeps only the modes $n \geq N$, each multiplied by λ_n.
(2) The ratio is largest when v is concentrated on the mode with the largest $|\lambda_n|$, $n \geq N$; that supremum tends to 0 because $\lambda_n \to 0$.

The identity on an infinite-dimensional space is *not* compact: every truncation misses a mode of eigenvalue 1, so $\|1 - K_N\| \geq 1$.

Physical picture: Compact means "negligible at high mode number"

Think of H as the space of modes of a field. A compact operator affects only finitely many low modes appreciably; at high mode number it is as small as we like. Adding a mass term to a Dirac operator, or any lower-order term to an elliptic operator, is (after composing with a parametrix) a compact perturbation. Things that depend only on the "ultraviolet" behaviour of an operator are therefore blind to compact perturbations. The index will be one of them.

The one fact about compact operators we shall use is the **Fredholm alternative** (Riesz): *if K is compact, then $1 + K$ has a finite-dimensional kernel, a closed image and a cokernel of the same dimension as its kernel.* For diagonal K this is plain to see: $1 + K$ multiplies e_n by $1 + \lambda_n$, which vanishes for at most finitely many n (since $\lambda_n \to 0$), and those same finitely many e_n span both the kernel and the cokernel.

3.21.3 Fredholm operators and the index

Working definition 3.55 (Fredholm operator and its index)

A bounded operator $T : H \to H'$ is **Fredholm** if

$$\dim \operatorname{Ker} T < \infty \qquad \text{and} \qquad \dim \operatorname{Coker} T = \dim\left(H'/\operatorname{Im} T\right) < \infty.$$

[finitely many independent zero modes $Tv = 0$, and only finitely many independent directions of H' missed by T] (The image is then automatically closed.) Its **index** is Nash's (3.128):

$$\operatorname{index} T = \dim \operatorname{Ker} T - \dim \operatorname{Coker} T = \dim \operatorname{Ker} T - \dim \operatorname{Ker} T^*. \tag{3.208}$$

Operational test: solve $Tv = 0$ and $T^*w = 0$. If both solution spaces are finite-dimensional (and the image is closed, which is automatic once the cokernel is finite-dimensional), T is Fredholm, and the index is the difference of the two dimensions. We write $\mathcal{F} = \mathcal{F}(H, H')$ for the set of Fredholm operators, with the norm topology.

The second form of (3.208) holds for the same reason as in finite dimensions: the orthogonal complement of the (closed) image is $\operatorname{Ker} T^*$, by the chain before (3.207).

Worked example 3.129: Easy cases

- Every invertible bounded operator is Fredholm with index 0: kernel and cokernel are both zero. In particular the identity.
- Every linear map between finite-dimensional spaces is Fredholm, with index $n - m$ by (3.206).
- $1 + K$ with K compact is Fredholm of index 0, by the Fredholm alternative.
- The zero operator on an infinite-dimensional H is *not* Fredholm: its kernel is all of H.
- The orthogonal projection onto the even modes, $Pe_{2n} = e_{2n}$, $Pe_{2n+1} = 0$, is not Fredholm: its kernel (all odd modes) is infinite-dimensional.

Worked example 3.130: A compact operator is never Fredholm: the image fails to close

Let $Ke_n = \frac{1}{n+1} e_n$ on ℓ^2. The kernel is 0. The image contains every finite combination of basis vectors, so it is dense. Is it everything? Try to solve $Kv = w$ with $w_n = \frac{1}{n+1}$, which is in ℓ^2 because $\sum \frac{1}{(n+1)^2} = \frac{\pi^2}{6}$:

$$\begin{aligned} Kv = w &\overset{(1)}{=} \frac{v_n}{n+1} = \frac{1}{n+1} \text{ for all } n \\ &\overset{(2)}{=} v_n = 1 \text{ for all } n, \end{aligned}$$

(1) Compare components; K is diagonal.
(2) Multiply by $n + 1$.

but $\sum_n |1|^2 = \infty$, so $v \notin \ell^2$. Thus $w \notin \operatorname{Im} K$ although it is a limit of vectors in the image: the image is not closed, and $\ell^2/\operatorname{Im} K$ is infinite-dimensional (the same argument works for $w_n = (n+1)^{-s}$ for any $\frac{1}{2} < s \leq \frac{3}{2}$). So K is not Fredholm. Every compact operator on an infinite-dimensional space fails in this way: it squashes the high modes, so it cannot be "almost invertible".

The shift: the Hilbert hotel has index -1

The fundamental example is the **unilateral shift** on ℓ^2,

$$Se_n = e_{n+1}, \qquad S(v_0, v_1, v_2, \dots) = (0, v_0, v_1, v_2, \dots).$$

It is the Hilbert hotel: every guest moves one room up, and room 0 is left empty.

Worked example 3.131: Index of the shift and its adjoint

Kernel of S. $Sv = (0, v_0, v_1, \dots) = 0$ forces every $v_n = 0$. So $\operatorname{Ker} S = 0$.
Adjoint. Compute matrix elements:

$$\langle S^* e_m, e_n \rangle \overset{(1)}{=} \langle e_m, S e_n \rangle \overset{(2)}{=} \langle e_m, e_{n+1} \rangle = \delta_{m,n+1}.$$

(1) Definition of the adjoint.
(2) Definition of S.

So $S^* e_m = e_{m-1}$ for $m \geq 1$ and $S^* e_0 = 0$: the adjoint shifts down and kills the bottom vector. As a matrix,

$$S = \begin{pmatrix} 0 & 0 & 0 & \cdots \\ 1 & 0 & 0 & \cdots \\ 0 & 1 & 0 & \cdots \\ 0 & 0 & 1 & \cdots \\ \vdots & & & \ddots \end{pmatrix}, \qquad S^* = \begin{pmatrix} 0 & 1 & 0 & 0 & \cdots \\ 0 & 0 & 1 & 0 & \cdots \\ 0 & 0 & 0 & 1 & \cdots \\ \vdots & & & & \ddots \end{pmatrix}.$$

Kernel of S^*. $S^* v = (v_1, v_2, \dots) = 0$ forces $v_n = 0$ for $n \geq 1$, so $\operatorname{Ker} S^* = \mathbb{C} e_0$.
Index.

$$\operatorname{index} S \overset{(1)}{=} \dim \operatorname{Ker} S - \dim \operatorname{Ker} S^* = 0 - 1 = -1,$$

$$\operatorname{index} S^* \overset{(2)}{=} \dim \operatorname{Ker} S^* - \dim \operatorname{Ker} S^{**} = 1 - 0 = +1.$$

(1) (3.208) with the two kernels just found.
(2) The same, using $S^{**} = S$.

Products. $S^* S = 1$ (up then down: nothing lost), while $S S^* = 1 - |e_0\rangle\langle e_0|$ (down then up: the bottom component is lost). And S^k has kernel 0 and cokernel spanned by $e_0, \dots, e_{k-1}$, so $\operatorname{index} S^k = -k$, $\operatorname{index} (S^*)^k = +k$. *Every integer is the index of some Fredholm operator.*

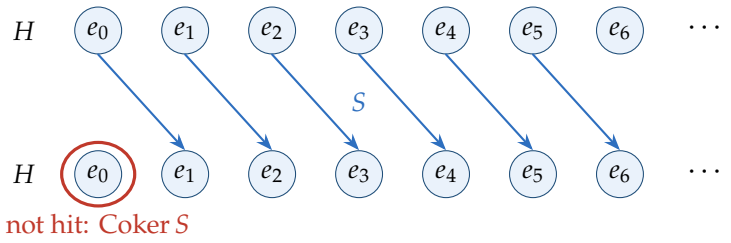


Figure 3.80: The shift S as the Hilbert hotel. Every basis vector moves one step up; nothing is sent to zero, so $\operatorname{Ker} S = 0$; the bottom vector e_0 is never reached, so $\operatorname{Coker} S = \mathbb{C} e_0$. Index $0 - 1 = -1$. For a matrix this is impossible (an injective square matrix is surjective); in infinite dimensions there is always "room at the top".

The harmonic oscillator, and why there is no phase operator

Worked example 3.132: The annihilation operator has index $+1$

On Fock space, $a|n\rangle = \sqrt{n}\,|n-1\rangle$ and $a^\dagger|n\rangle = \sqrt{n+1}\,|n+1\rangle$. These are unbounded (the factors $\sqrt{n}$ grow), so, as for differential operators, we regard a as a bounded map from the domain $D = \{\psi : \sum n\,|\psi_n|^2 < \infty\}$ (with norm $\|\psi\|_D^2 = \sum (1+n)\,|\psi_n|^2$) to H.

- $\operatorname{Ker} a = \mathbb{C}|0\rangle$: the vacuum is the unique state killed by a.
- a is onto: given $\phi = \sum c_n|n\rangle \in H$, the state $\psi = \sum_n \frac{c_n}{\sqrt{n+1}}|n+1\rangle$ satisfies $a\psi = \phi$ and lies in D, since $\sum_n (n+2)\frac{|c_n|^2}{n+1} \le 2\sum_n |c_n|^2 < \infty$.

So $\operatorname{index} a = 1 - 0 = 1$, and likewise $\operatorname{index} a^\dagger = -1$. In fact a is the shift S^* dressed with the factors $\sqrt{n}$. Write $N = a^\dagger a$ for the number operator, $N|n\rangle = n|n\rangle$. Then

$$a|n\rangle \overset{(1)}{=} \sqrt{n}\,|n-1\rangle \overset{(2)}{=} S^*\big(\sqrt{n}|n\rangle\big) \overset{(3)}{=} S^*\sqrt{N}|n\rangle,$$

(1) Definition of a.
(2) S^* lowers $|n\rangle$ to $|n-1\rangle$ (and kills $|0\rangle$, consistent with $\sqrt{0} = 0$), with $e_n = |n\rangle$.
(3) $\sqrt{N}|n\rangle = \sqrt{n}|n\rangle$.

so $a = S^*\sqrt{N}$. This is the *polar decomposition* of a, and S^* is the Susskind–Glogower "exponential of the phase" $\widehat{e^{i\varphi}}$.

Worked example 3.133: No unitary phase operator: an index argument

Dirac hoped to write $a = U\sqrt{N}$ with $U = e^{i\hat{\varphi}}$ *unitary*, which would make $\hat{\varphi}$ a Hermitian phase operator conjugate to N. Suppose it were possible. On every $|n\rangle$ with $n \ge 1$ we must have $U|n\rangle = \frac{1}{\sqrt{n}}a|n\rangle = |n-1\rangle = S^*|n\rangle$; so U and S^* differ only on the single vector $|0\rangle$, i.e. $U - S^*$ has rank at most 1. Then

$$0 \overset{(1)}{=} \operatorname{index} U \overset{(2)}{=} \operatorname{index}\big(S^* + (U - S^*)\big) \overset{(3)}{=} \operatorname{index} S^* \overset{(4)}{=} 1,$$

(1) A unitary operator is invertible.
(2) Trivial rewriting.
(3) The index does not change under a finite-rank (compact) perturbation; we prove this in Section 3.21.4.
(4) Example 3.131.

a contradiction. *The nonexistence of a Hermitian phase operator for the oscillator is the statement* $\operatorname{index} a = 1 \neq 0$. Physically: the spectrum of N is bounded below, so the "lowering" operator must annihilate something, and no unitary can do that.

Toeplitz operators: the first index theorem

Now a family of examples where the index is computed by a *topological* number, exactly as in the Atiyah–Singer theorem, but where everything can be checked by hand.

Inside $L^2(S^1)$ take the **Hardy space** H^2: the closed span of the nonnegative Fourier modes $z^n = e^{in\theta}$, $n \ge 0$. These are the boundary values of functions holomorphic inside the unit disc, $h(z) = \sum_{n\ge 0} h_n z^n$. With $e_n = z^n$, H^2 is just ℓ^2. Let P be the orthogonal projection of $L^2(S^1)$ onto H^2 (throw away the negative modes).

Working definition 3.56 (Toeplitz operator)

For a continuous function $f : S^1 \to \mathbb{C}$ (the **symbol**), the **Toeplitz operator** is

$$T_f : H^2 \to H^2, \qquad T_f h = P(fh) :$$

[H^2 is spanned by the modes z^n with $n \ge 0$, and P discards the negative modes] multiply

by f, then project back to the nonnegative modes. Its matrix in the basis z^n is $(T_f)_{mn} = \hat{f}_{m-n}$, where $\hat{f}_k$ are the Fourier coefficients of f [$f = \sum_{k\in\mathbb{Z}} \hat{f}_k z^k$]: constant along diagonals (a "Toeplitz matrix").

Worked example 3.134: Symbols z^k give the shifts

For $f = z$: $T_z z^n = P(z^{n+1}) = z^{n+1}$, so $T_z = S$. For $f = z^{-1} = \bar{z}$: $T_{\bar{z}} z^n = P(z^{n-1})$, which is z^{n-1} for $n \geq 1$ and 0 for $n = 0$ (since z^{-1} is a negative mode). So $T_{\bar{z}} = S^*$. In general $T_{z^k} = S^k$ and $T_{z^{-k}} = (S^*)^k$ for $k \geq 0$, hence

$$\operatorname{index} T_{z^k} = -k.$$

The function z^k winds k times around the origin. The index is minus the winding number of the symbol.

Worked example 3.135: The symbol $z - a$: a zero inside the disc makes a cokernel

Take $f(z) = z - a$ with $a \in \mathbb{C}$, $|a| \neq 1$, so that f never vanishes on the circle. Then $T_f = S - a$.
Case $|a| > 1$. The symbol winds 0 times around the origin (the zero $z = a$ is outside the circle). Since $\|S\| = 1 < |a|$,

$$S - a = -a\big(1 - a^{-1}S\big), \qquad \big(1 - a^{-1}S\big)^{-1} = \sum_{j\geq 0} a^{-j} S^j,$$

the geometric (Neumann) series converging in norm. So $S - a$ is invertible: index 0.
Case $|a| < 1$. The symbol winds once. Kernel: $(S - a)h = 0$ reads, in components, $-a h_0 = 0$ and $h_{n-1} - a h_n = 0$ for $n \geq 1$. If $a \neq 0$ the first gives $h_0 = 0$ and then the second gives $h_n = h_{n-1}/a = 0$ for all n; if $a = 0$, S is injective anyway. So $\operatorname{Ker}(S - a) = 0$. Cokernel: look for zero modes of the adjoint $S^* - \bar{a}$. Try the vector

$$k_a = \sum_{n\geq 0} \bar{a}^{\,n} e_n, \qquad \|k_a\|^2 = \sum_n |a|^{2n} = \frac{1}{1 - |a|^2} < \infty.$$

Then

$$S^* k_a \overset{(1)}{=} \sum_{n\geq 1} \bar{a}^{\,n} e_{n-1} \overset{(2)}{=} \bar{a} \sum_{m\geq 0} \bar{a}^{\,m} e_m = \bar{a}\, k_a,$$

(1) S^* lowers each e_n and kills e_0.
(2) Put $m = n - 1$ and pull out one factor $\bar{a}$.

so $(S - a)^* k_a = (S^* - \bar{a})k_a = 0$. Conversely any solution of $S^* v = \bar{a} v$ has $v_{n+1} = \bar{a} v_n$, so $v = v_0 k_a$: $\operatorname{Ker}(S - a)^* = \mathbb{C} k_a$. Hence $\operatorname{index}(S - a) = 0 - 1 = -1$.
What the cokernel is. Since $\langle k_a, h\rangle = \sum_n a^n h_n = h(a)$, the vector k_a is "evaluation at a". The image of $S - a$ consists of the functions $(z - a)h(z)$, which all vanish at $z = a$: the cokernel is the single condition $h(a) = 0$ that the image obeys. It exists exactly when a is inside the disc, where holomorphic functions can be evaluated.
Check by stability. Write $P_0 = |e_0\rangle\langle e_0|$. For $|a| < 1$,

$$S(1 - aS^*) \overset{(1)}{=} S - a\, SS^* \overset{(2)}{=} S - a(1 - P_0) = (S - a) + aP_0,$$

(1) Distribute.
(2) $SS^* = 1 - P_0$, Example 3.131.

so $S - a = S(1 - aS^*) - aP_0$: the index -1 operator S, times an invertible operator ($\|aS^*\| < 1$), plus a rank-one operator. Index $-1 + 0 = -1$, agreeing with the direct computation.

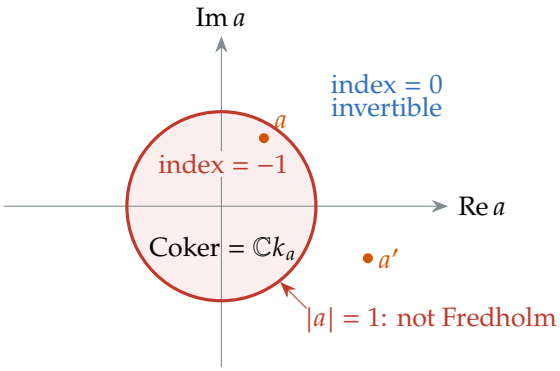


Figure 3.81: The index of the Toeplitz operator $T_{z-a} = S - a$ as a function of a. Inside the unit circle the symbol $z - a$ winds once around 0 and the index is -1; outside it winds 0 times and the operator is invertible. The index can only change by crossing the red wall $|a| = 1$, where the symbol vanishes somewhere on the circle and $S - a$ stops being Fredholm (its image is no longer closed). The index is locally constant on the space of Fredholm operators, and the walls between its values are made of non-Fredholm operators.

Key idea: The Toeplitz index theorem (Noether, Gohberg–Krein)

If $f : S^1 \to \mathbb{C}$ is continuous and never zero, T_f is Fredholm and

$$\operatorname{index} T_f = -\operatorname{wind}(f) = -\frac{1}{2\pi i}\oint_{S^1}\frac{\mathrm{d}f}{f}. \tag{3.209}$$

Left side: an analytic count of solutions. Right side: a topological invariant of the symbol, the class of f in $\pi_1(\mathbb{C}^*) = \mathbb{Z}$.

Why it is true. Any nowhere-zero f can be deformed through nowhere-zero functions into z^k with $k = \operatorname{wind}(f)$ (this is what it means for the winding number to classify loops in $\mathbb{C}^*$). Along the deformation T_f moves continuously in norm (since $\|T_f - T_g\| \le \sup|f - g|$) and stays Fredholm, so its index cannot change (Section 3.21.4), and at the end it is index $S^k = -k$. The step we skip is that T_f is Fredholm when f has no zeros: the parametrix is $T_{1/f}$, because $T_f T_{1/f} - 1$ is compact (it is a product of two "Hankel" operators, which are compact for continuous symbols).

This is the prototype of every index theorem in this book: *an operator whose "symbol" is invertible has an index, and the index is a topological invariant of the symbol.* For elliptic operators the symbol is the principal symbol $\sigma(x, p)$, invertible for $p \neq 0$, and its topology lives in K-theory.

3.21.4 Why the index is stable

We have used several times that the index does not change under small or compact perturbations. This is *the* property that makes the index useful, so we now explain it twice: once with a physicist's spectral argument, once with a two-line block-matrix computation that is actually a proof.

The spectral argument: nonzero modes come in pairs

Two facts about T^*T and TT^*. First, they have the same zero modes as T and T^*:

$$T^*Tv = 0 \overset{(1)}{=} \langle v, T^*Tv\rangle = 0 \overset{(2)}{=} \|Tv\|^2 = 0 \overset{(3)}{=} Tv = 0.$$

(1) Take the inner product with v. (Conversely $Tv = 0$ clearly gives $T^*Tv = 0$, so each "=" here is an equivalence.)
(2) $\langle v, T^*Tv\rangle = \langle Tv, Tv\rangle$ by the definition of the adjoint.
(3) Only the zero vector has norm zero.

So $\operatorname{Ker} T = \operatorname{Ker} T^*T$, and likewise $\operatorname{Ker} T^* = \operatorname{Ker} TT^*$. Second, their *nonzero* eigenvalues coincide, with multiplicity. If $T^*Tv = \lambda v$ with $\lambda \neq 0$, then

$$TT^*(Tv) \overset{(1)}{=} T(T^*Tv) \overset{(2)}{=} \lambda\, Tv, \qquad \|Tv\|^2 \overset{(3)}{=} \lambda\, \|v\|^2 \neq 0.$$

(1) Associativity of composition.
(2) v is a λ-eigenvector of T^*T.
(3) $\|Tv\|^2 = \langle v, T^*Tv\rangle = \lambda\langle v, v\rangle$.

So T maps the λ-eigenspace of T^*T injectively into the λ-eigenspace of TT^*, and T^* maps back the same way; the two eigenspaces have equal dimension. Therefore

$$\operatorname{index} T = \dim \operatorname{Ker} T^*T - \dim \operatorname{Ker} TT^*, \tag{3.210}$$

and the right-hand side counts only the *unpaired* levels. When we deform T, the eigenvalues of T^*T and TT^* move, but the nonzero ones move in locked pairs. A zero mode of T^*T can rise off zero only if a partner rises off zero with it in TT^*, and vice versa. The difference of the numbers of zero modes cannot change.

Worked example 3.136: Supersymmetric quantum mechanics and the Witten index

Package T into a "supercharge" on $H \oplus H'$ (bosons $\oplus$ fermions):

$$Q = \begin{pmatrix} 0 & T^* \\ T & 0 \end{pmatrix}, \qquad \mathcal{H} = Q^2 = \begin{pmatrix} T^*T & 0 \\ 0 & TT^* \end{pmatrix}, \qquad (-1)^F = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

The pairing above is the statement that Q maps bosonic states of energy $\lambda \neq 0$ to fermionic states of the same energy. The **Witten index** is

$$\begin{aligned} \operatorname{Tr}(-1)^F e^{-\beta\mathcal{H}} &\overset{(1)}{=} \operatorname{Tr}_H e^{-\beta T^*T} - \operatorname{Tr}_{H'} e^{-\beta TT^*} \\ &\overset{(2)}{=} \dim \operatorname{Ker} T^*T - \dim \operatorname{Ker} TT^* + \sum_{\lambda\neq 0} (m_\lambda - m_\lambda) e^{-\beta\lambda} \\ &\overset{(3)}{=} \operatorname{index} T, \end{aligned}$$

(1) The supertrace of a block-diagonal operator.
(2) Split each trace into zero modes (contributing $e^0 = 1$ each) and nonzero levels, where m_λ is the common multiplicity of λ in T^*T and in TT^*.
(3) The nonzero levels cancel pairwise, and (3.210).

independent of β (whenever the traces converge). For the oscillator, $T = a$ gives $T^*T = a^\dagger a = N$ with spectrum $\{0, 1, 2, \dots\}$ and $TT^* = aa^\dagger = N + 1$ with spectrum $\{1, 2, 3, \dots\}$:

$$\operatorname{Tr}(-1)^F e^{-\beta\mathcal{H}} \overset{(1)}{=} \sum_{n\geq 0} e^{-\beta n} - \sum_{n\geq 0} e^{-\beta(n+1)} \overset{(2)}{=} \frac{1 - e^{-\beta}}{1 - e^{-\beta}} = 1 = \operatorname{index} a.$$

(1) The two spectra.
(2) Sum both geometric series: $\frac{1}{1-e^{-\beta}} - \frac{e^{-\beta}}{1-e^{-\beta}}$.

This β-independence is the germ of the heat-kernel proof of the index theorem (the McKean–Singer formula, Chapter 4): at large β only zero modes survive, at small β the traces are computed by local curvature integrals, and the answer is the same.

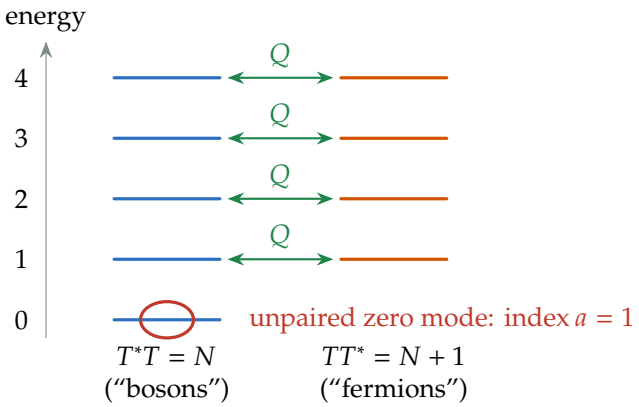


Figure 3.82: The spectra of $a^\dagger a$ and $aa^\dagger$ for the harmonic oscillator. Every nonzero level on the left has a partner on the right (green arrows: the supercharge Q maps one to the other). Only the vacuum on the left is unpaired. Deforming a can move the levels, but only in pairs, so the one unpaired level, the index, survives any deformation.

The block-matrix argument: reduction to a finite matrix

Now the real mechanism, which works for any Fredholm $T : H \to H'$. Split the two spaces into "zero-mode" and "non-zero-mode" parts:

$$H = \underbrace{\operatorname{Ker} T}_{H_0} \oplus \underbrace{(\operatorname{Ker} T)^\perp}_{H_1}, \qquad H' = \underbrace{(\operatorname{Im} T)^\perp}_{H_0' \cong \operatorname{Coker} T} \oplus \underbrace{\operatorname{Im} T}_{H_1'}.$$

H_0 and H_0' are finite-dimensional, of dimensions $\dim \operatorname{Ker} T$ and $\dim \operatorname{Coker} T$. In these blocks

$$T = \begin{pmatrix} 0 & 0 \\ 0 & T_{11} \end{pmatrix},$$

where $T_{11} : H_1 \to H_1'$ is a bijection, hence has a bounded inverse (by the open mapping theorem of functional analysis). Now perturb, $T \to T + R$, and write the perturbed operator in the same blocks:

$$T + R = \begin{pmatrix} A & B \\ C & D \end{pmatrix}, \qquad A = R_{00} : H_0 \to H_0', \quad D = T_{11} + R_{11} : H_1 \to H_1'.$$

If $\|R\| < 1/\left\|T_{11}^{-1}\right\|$, then $D = T_{11}(1 + T_{11}^{-1}R_{11})$ is invertible (Neumann series again). Now do Gaussian elimination with operator entries:

$$\begin{pmatrix} 1 & -BD^{-1} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -D^{-1}C & 1 \end{pmatrix} \overset{(1)}{=} \begin{pmatrix} 1 & -BD^{-1} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} A - BD^{-1}C & B \\ 0 & D \end{pmatrix}$$
$$\overset{(2)}{=} \begin{pmatrix} A - BD^{-1}C & 0 \\ 0 & D \end{pmatrix}. \tag{3.211}$$

(1) Multiply the right two factors: the first column becomes $(A - BD^{-1}C,\ C - DD^{-1}C) = (A - BD^{-1}C,\ 0)$, the second is unchanged.

(2) Multiply by the left factor: the top-right entry becomes $B - BD^{-1}D = 0$; the top-left gets $-BD^{-1} \cdot 0 = 0$ added.

The two outer factors are invertible (triangular with 1 on the diagonal), so they change neither the dimension of the kernel nor that of the cokernel. The block D is invertible. So all the kernel and cokernel of $T + R$ sit in the *finite matrix*

$$M = A - BD^{-1}C : H_0 \longrightarrow H_0',$$

the **Schur complement**, and

$$\operatorname{index}(T+R) \overset{(1)}{=} \operatorname{index} M \overset{(2)}{=} \dim H_0 - \dim H_0' \overset{(3)}{=} \dim \operatorname{Ker} T - \dim \operatorname{Coker} T = \operatorname{index} T. \quad (3.212)$$

(1) (3.211): $\operatorname{Ker}(T+R) \cong \operatorname{Ker} M$ and $\operatorname{Coker}(T+R) \cong \operatorname{Coker} M$, because the invertible block D contributes nothing.
(2) M is a matrix from a space of dimension $\dim H_0$ to one of dimension $\dim H_0'$; its index is the difference, by (3.206).
(3) The definitions of H_0 and H_0'.

Key idea: Finite-dimensional reduction

Near any Fredholm operator T, every nearby operator is, up to invertible changes of basis, "an invertible operator on the non-zero modes" $\oplus$ "a finite matrix M from $\operatorname{Ker} T$ to $\operatorname{Coker} T$". All jumping of kernels happens inside M, where rank–nullity forces the kernel and cokernel to jump together. So:

- the Fredholm operators form an *open* set in $B(H, H')$;
- $\dim \operatorname{Ker}$ can only drop nearby ($\dim \operatorname{Ker} M \leq \dim H_0$): it is *upper semicontinuous*;
- the index is *locally constant*, hence constant along any continuous path of Fredholm operators: it is a *homotopy invariant*.

This is exactly the picture of Figure 3.79, now in infinite dimensions, and it is how one computes an index in practice: identify a finite-dimensional "low-mode" problem and discard the rest.

Composition, adjoints and parametrices

Adjoints. Since $\operatorname{Ker} T^* = (\operatorname{Im} T)^\perp$ and $\operatorname{Coker} T^* \cong \operatorname{Ker} T$, the roles swap:

$$\operatorname{index} T^* = -\operatorname{index} T.$$

Composition. For Fredholm $T : H_1 \to H_2$ and $S : H_2 \to H_3$, the kernels and cokernels fit into a sequence of finite-dimensional spaces

$$0 \to \operatorname{Ker} T \xrightarrow{\iota} \operatorname{Ker} ST \xrightarrow{T} \operatorname{Ker} S \xrightarrow{\pi} \operatorname{Coker} T \xrightarrow{S} \operatorname{Coker} ST \xrightarrow{\pi'} \operatorname{Coker} S \to 0,$$

where ι is inclusion ($Tv = 0 \Rightarrow STv = 0$), the second map applies T (if $STv = 0$ then $Tv \in \operatorname{Ker} S$), π sends $w \in \operatorname{Ker} S \subset H_2$ to its class in $H_2/\operatorname{Im} T$, the next map is induced by S, and π' is the projection $H_3/\operatorname{Im} ST \to H_3/\operatorname{Im} S$ (valid since $\operatorname{Im} ST \subset \operatorname{Im} S$). One checks directly that this sequence is *exact*: the image of each map is the kernel of the next. For instance, at $\operatorname{Ker} S$: an element $w \in \operatorname{Ker} S$ maps to 0 in $\operatorname{Coker} T$ iff $w = Tv$ for some v, and then $STv = Sw = 0$, so $v \in \operatorname{Ker} ST$ and w is in the image of the previous map. The alternating sum of dimensions in a finite exact sequence vanishes (apply rank–nullity at each spot), so

$$\begin{aligned} 0 &\overset{(1)}{=} \dim \operatorname{Ker} T - \dim \operatorname{Ker} ST + \dim \operatorname{Ker} S \\ &\qquad - \dim \operatorname{Coker} T + \dim \operatorname{Coker} ST - \dim \operatorname{Coker} S \\ &\overset{(2)}{=} \operatorname{index} T - \operatorname{index} ST + \operatorname{index} S, \\ \text{i.e.} \quad \operatorname{index}(ST) &= \operatorname{index} S + \operatorname{index} T. \end{aligned} \quad (3.213)$$

(1) Alternating sum over the six-term exact sequence.
(2) Regroup into the three indices.

Checks: $S^*S = 1$ gives $\operatorname{index} S^* + \operatorname{index} S = 1 - 1 = 0 = \operatorname{index} 1$. $SS^* = 1 - P_0$ has kernel and cokernel both $\mathbb{C}e_0$, index $0 = -1 + 1$. And $\operatorname{index}\big(S^j(S^*)^k\big) = k - j$.

Working definition 3.57 (parametrix; Atkinson's theorem)

A **parametrix** for $T : H \to H'$ is a bounded $Q : H' \to H$ such that $QT - 1$ and $TQ - 1$ are compact [norm limits of finite-rank operators, negligible at high mode number]. *Atkinson's theorem:* T is Fredholm [finite-dimensional kernel and cokernel] if and only if it has a parametrix. In words, *Fredholm = invertible modulo compact operators*.

Both directions are short.

- *Fredholm $\Rightarrow$ parametrix.* In the blocks above, let Q be T_{11}^{-1} on $H_1' = \operatorname{Im} T$ and 0 on H_0'. Then $QT = 1 - P_{\operatorname{Ker} T}$ and $TQ = 1 - P_{(\operatorname{Im} T)^\perp}$, where P denotes orthogonal projection. These differ from 1 by finite-rank projections. Physically, Q is the propagator with the zero modes removed.
- *Parametrix $\Rightarrow$ Fredholm.* If $QT = 1 + K$ with K compact, then $\operatorname{Ker} T \subset \operatorname{Ker}(1 + K)$, which is finite-dimensional by the Fredholm alternative. If $TQ = 1 + K'$, then $\operatorname{Im} T \supset \operatorname{Im}(1 + K')$, which has finite codimension. So kernel and cokernel are finite.

Two consequences.

- *Compact perturbations do not change the index.* A parametrix Q of T is also one of $T + tK$ for every t (the extra terms tQK, tKQ are compact), so $T + tK$ is Fredholm for all $t \in [0, 1]$, and by homotopy invariance $\operatorname{index}(T + K) = \operatorname{index} T$. This justifies step (3) of Example 3.133 and the stability check in Example 3.135.
- *Elliptic operators are Fredholm.* In Chapter 2 a parametrix for an elliptic operator is built by inverting its principal symbol $\sigma(x, p)$ for large $|p|$; the error is a smoothing operator, which is compact on a compact manifold. Atkinson's theorem then gives Nash's remark on p. 88.

Key idea: The components of $\mathcal{F}$ are labelled by the index

The index is a continuous map $\operatorname{index} : \mathcal{F} \to \mathbb{Z}$, and it is a bijection on path components:

$$\pi_0(\mathcal{F}) \xrightarrow{\cong} \mathbb{Z}.$$

It is onto because $\operatorname{index} S^k = -k$ and $\operatorname{index}(S^*)^k = k$. It is one-to-one because two Fredholm operators of the same index can be joined by a path in $\mathcal{F}$ (reduce each to "invertible $\oplus$ finite matrix" by (3.211), join the finite matrices through matrices of the same size, and join the invertible parts using the connectedness of $\mathrm{GL}(H)$, discussed in Section 3.21.6). This is Nash's (3.128) read as $[\text{point}, \mathcal{F}] \cong \mathbb{Z} = K(\text{point})$.

3.21.5 Families of Fredholm operators and the index bundle

Now let the operator depend on parameters. Let X be a compact space (think of a sphere of gauge fields, a torus of Wilson lines, a moduli space of instantons) and H a fixed separable infinite-dimensional Hilbert space.

Working definition 3.58 (families and $[X,\mathcal{F}]$)

A **family of Fredholm operators** over X is a continuous map $T : X \to \mathcal{F}$, $x \mapsto T_x$, continuous in the norm topology [$\|T_{x'} - T_x\| \to 0$ as $x' \to x$]. Two families are **homotopic** if they are joined by a family over $X \times [0,1]$. The set of homotopy classes is Nash's $[X,\mathcal{F}]$. Pointwise composition, $(ST)_x = S_x T_x$, is a product on it: the product of Fredholm operators is Fredholm (by (3.213)), and it respects homotopy. So $[X,\mathcal{F}]$ is a semigroup [it has an associative product, but inverses need not exist], with the constant family 1 as unit.

For each x we get an integer index T_x, but on a connected X it is constant, so it carries little information. The natural guess is to use the kernels themselves: the vector spaces $\operatorname{Ker} T_x$, $x \in X$, look like the fibres of a vector bundle, and so do the cokernels. The guess fails, because kernels jump.

Worked example 3.137: A family whose kernels do not form a bundle

On $H = \ell^2 \oplus \mathbb{C}$ take $X = [-1,1]$ and

$$T_t = \begin{pmatrix} S^* & 0 \\ 0 & t \end{pmatrix} : \ (h,z) \mapsto (S^*h,\ tz).$$

For $t \neq 0$: $\operatorname{Ker} T_t = \mathbb{C}(e_0, 0)$ and T_t is onto, so $\dim \operatorname{Ker} = 1$, $\dim \operatorname{Coker} = 0$. At $t = 0$: $\operatorname{Ker} T_0 = \mathbb{C}(e_0,0) \oplus \mathbb{C}(0,1)$ and $\operatorname{Coker} T_0 = \mathbb{C}(0,1)$, so $\dim \operatorname{Ker} = 2$, $\dim \operatorname{Coker} = 1$. The index is 1 throughout, but the kernels have rank 1 except at one point, so $\{\operatorname{Ker} T_t\}$ is not a vector bundle.

The cure is to add enough extra directions to the target that the cokernel disappears everywhere. Then the kernel has constant dimension, by the index.

Working definition 3.59 (the index bundle)

Let $T : X \to \mathcal{F}(H, H')$ with X compact.

1. Choose finitely many vectors $w_1, \dots, w_N \in H'$ such that $\operatorname{Im} T_x + \operatorname{span}(w_1, \dots, w_N) = H'$ for every $x \in X$.
2. Define the *stabilised* family
$$\hat{T}_x : H \oplus \mathbb{C}^N \to H', \qquad \hat{T}_x(h,c) = T_x h + \textstyle\sum_j c_j w_j.$$
It is surjective for every x, so its kernels $K_x = \operatorname{Ker} \hat{T}_x$ all have the same dimension and form a vector bundle $K \to X$.
3. The **index bundle** of the family is the K-theory class
$$\operatorname{index} T = [K] - [\mathbb{C}^N] \ \in K(X). \tag{3.214}$$
[$[\mathbb{C}^N]$ is the trivial bundle $X \times \mathbb{C}^N$, and $K(X)$ holds formal differences of bundles over X]

Special case: if $\dim \operatorname{Ker} T_x$ is constant, then $\operatorname{Ker} T$ and $\operatorname{Coker} T$ are honest bundles and $\operatorname{index} T = [\operatorname{Ker} T] - [\operatorname{Coker} T]$.

Let us check each step.

Step 1 is possible. At a single point x_0, take the w_j to be a basis of $(\operatorname{Im} T_{x_0})^\perp$; then $\hat{T}_{x_0}$ is onto. A

surjective Fredholm operator stays surjective under small perturbations: in (3.211) its cokernel block H_0' is zero, so the Schur complement M maps into the zero space and has no cokernel. So $\hat{T}_x$ is onto for x in a neighbourhood U_{x_0}. Finitely many such neighbourhoods cover the compact space X; take all their w's together.

The rank of K. For each x,

$$\dim K_x \overset{(1)}{=} \operatorname{index} \hat{T}_x \overset{(2)}{=} \operatorname{index}\big(T_x \circ \mathrm{pr}_H\big) \overset{(3)}{=} \operatorname{index} T_x + \operatorname{index} \mathrm{pr}_H \overset{(4)}{=} \operatorname{index} T_x + N.$$

(1) $\hat{T}_x$ is onto, so its cokernel is 0 and its index is the dimension of its kernel.
(2) $\hat{T}_x = T_x \circ \mathrm{pr}_H +$ (the finite-rank map $(h, c) \mapsto \sum c_j w_j$), where $\mathrm{pr}_H : H \oplus \mathbb{C}^N \to H$ is the projection; a finite-rank perturbation does not change the index.
(3) Additivity (3.213).
(4) pr_H is onto with kernel $\mathbb{C}^N$.

Constant, as promised. (That the K_x fit together into a locally trivial bundle follows from the same block picture: near x_0, orthogonal projection onto K_{x_0} restricts to an isomorphism $K_x \to K_{x_0}$.)

The answer does not depend on the choice of w's. Adding M more vectors $w_1', \dots, w_M'$ gives $\hat{T}'$ with kernel K', and there is a short exact sequence of bundles

$$0 \to K \to K' \to \mathbb{C}^M \to 0, \qquad (h, c) \mapsto (h, c, 0), \quad (h, c, c') \mapsto c',$$

exact on the right because for every c' the vector $-\sum c_j' w_j'$ lies in $H' = \operatorname{Im} \hat{T}_x$. Short exact sequences of bundles split (choose a metric), so $K' \cong K \oplus \mathbb{C}^M$ and

$$[K'] - [\mathbb{C}^{N+M}] = [K] + [\mathbb{C}^M] - [\mathbb{C}^N] - [\mathbb{C}^M] = [K] - [\mathbb{C}^N].$$

Any two choices are dominated by their union, so all choices agree.

Homotopy invariance. A homotopy is a family over $X \times [0, 1]$, whose index bundle restricts to the index bundles at the two ends; bundles over $X \times [0, 1]$ have isomorphic restrictions to $X \times \{0\}$ and $X \times \{1\}$ (homotopic maps pull back isomorphic bundles, as in the discussion of classifying maps in the previous part). So index is a well-defined map

$$\operatorname{index} : [X, \mathcal{F}] \longrightarrow K(X).$$

At a point (Nash's (3.128)). If X is a point, $K(\mathrm{pt}) \cong \mathbb{Z}$ by the dimension, and

$$\operatorname{index} T = [K] - [\mathbb{C}^N] \;\longmapsto\; \dim K - N \overset{(1)}{=} \operatorname{index} T + N - N = \dim \operatorname{Ker} T - \dim \operatorname{Coker} T.$$

(1) The rank computation above.

So the index bundle is the promised generalisation of the integer index.

The special case. If $\dim \operatorname{Ker} T_x$ is constant, there is an exact sequence of bundles

$$0 \to \operatorname{Ker} T \to K \to \mathbb{C}^N \to \operatorname{Coker} T \to 0,$$

with maps $h \mapsto (h, 0)$, $(h, c) \mapsto c$, and $c \mapsto [\sum c_j w_j]$. (Exactness at $\mathbb{C}^N$: c comes from K iff $\sum c_j w_j = -T_x h$ for some h, iff its class in the cokernel is zero.) The alternating sum of the classes of an exact sequence of bundles is zero in $K(X)$, so

$$\operatorname{index} T \overset{(1)}{=} [K] - [\mathbb{C}^N] \overset{(2)}{=} [\operatorname{Ker} T] - [\operatorname{Coker} T].$$

(1) Definition (3.214).
(2) $[\operatorname{Ker} T] - [K] + [\mathbb{C}^N] - [\operatorname{Coker} T] = 0$ from the exact sequence.

Worked example 3.138: Stabilising the jumping family

In Section 3.21.5 the only cokernel direction ever needed is $w = (0, 1)$. With $N = 1$,

$$\hat{T}_t(h, z, c) = \big(S^*h,\ tz + c\big), \qquad K_t = \operatorname{Ker} \hat{T}_t = \mathbb{C}(e_0, 0, 0) \oplus \mathbb{C}(0, 1, -t).$$

Now $\dim K_t = 2$ for every t, including $t = 0$: the jump has been absorbed by the extra coordinate c. K is a trivial rank-2 bundle over the interval, and index $T = [\mathbb{C}^2] - [\mathbb{C}^1]$, the class of virtual dimension 1. Nothing twisted here (an interval is contractible); the point is only that the construction removes the jumps.

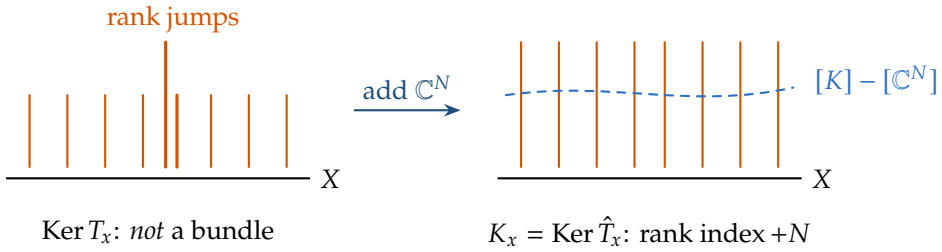


Figure 3.83: Left: the kernels of a Fredholm family have jumping dimension, so they do not form a vector bundle. Right: after adding N auxiliary directions that kill every cokernel, the kernels have constant rank index $T_x + N$ and form a bundle K. The index bundle is the virtual bundle $[K] - [\mathbb{C}^N]$; the N auxiliary directions are subtracted off again.

Watch out: Finite-dimensional families only give trivial classes

If every T_x acts between fixed spaces $\mathbb{C}^n \to \mathbb{C}^m$ (with constant rank, say), the exact sequence $0 \to \operatorname{Ker} T \to \mathbb{C}^n \to \mathbb{C}^m \to \operatorname{Coker} T \to 0$ gives index $T = [\operatorname{Ker} T] - [\operatorname{Coker} T] = [\mathbb{C}^n] - [\mathbb{C}^m]$, which is trivial. The twisting of an index bundle can only come from infinite dimensions: from the "room at the top" of the Hilbert hotel. This is why K-theory needs Hilbert space, and why the index of a matrix was boring.

Every K-class is an index bundle

Here is a family that produces any given bundle $E \to X$. By the stable equivalence theorem of the previous part, E sits inside a trivial bundle $X \times \mathbb{C}^N$ (it has a complement E' with $E \oplus E' \cong \mathbb{C}^N$). Let P_x be the orthogonal projection of $\mathbb{C}^N$ onto the fibre E_x; it depends continuously on x. On $H = \mathbb{C}^N \otimes \ell^2$ define

$$T^E_x = P_x \otimes S^* + (1 - P_x) \otimes 1. \tag{3.215}$$

In words: on $E_x \otimes \ell^2$ it acts as the down-shift, on $E_x^\perp \otimes \ell^2$ as the identity. Its kernel is $E_x \otimes \operatorname{Ker} S^* = E_x \otimes e_0$ and it is onto. The kernels have constant dimension $\operatorname{rank} E$, so

$$\operatorname{index} T^E = [\operatorname{Ker} T^E] - [\operatorname{Coker} T^E] = [E \otimes e_0] - 0 = [E].$$

The adjoint family $(T^E)^* = P_x \otimes S + (1 - P_x) \otimes 1$ has no kernel and cokernel $E \otimes e_0$, so $\operatorname{index}(T^E)^* = -[E]$. Finally, for two bundles E, F,

$$\operatorname{index}\big(T^E \oplus (T^F)^*\big) = [E] - [F].$$

Every element of $K(X)$ has the form $[E] - [F]$, so *every element of $K(X)$ is the index bundle of some Fredholm family*: the map index : $[X, \mathcal{F}] \to K(X)$ is onto.

Worked example 3.139: The Bott class as an operator family on $\mathbb{CP}^1$

Let $L \to \mathbb{CP}^1$ be the tautological line bundle, whose fibre over $[1:z]$ is the line through $u = (1,z)^T$. The projection onto it is

$$P_z = \frac{uu^\dagger}{u^\dagger u} = \frac{1}{1+|z|^2}\begin{pmatrix} 1 & \bar z \\ z & |z|^2 \end{pmatrix},$$

which tends to $\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$ as $z \to \infty$, so it is continuous on the whole sphere. On $H = \ell^2 \oplus \ell^2$ the family (3.215) is the 2×2 matrix of operators

$$\begin{aligned} T_z &\overset{(1)}{=} P_z \otimes S^* + (1-P_z)\otimes 1 \\ &\overset{(2)}{=} \frac{1}{1+|z|^2}\begin{pmatrix} S^* + |z|^2 & \bar z\,(S^*-1) \\ z\,(S^*-1) & |z|^2\,S^* + 1 \end{pmatrix}. \end{aligned}$$

(1) Definition (3.215) with $N = 2$.

(2) $1 - P_z = \frac{1}{1+|z|^2}\begin{pmatrix} |z|^2 & -\bar z \\ -z & 1 \end{pmatrix}$; add the two matrices entry by entry.

Check that (e_0, ze_0) is in the kernel, using $S^* e_0 = 0$:

$$(1+|z|^2)\,T_z\begin{pmatrix} e_0 \\ ze_0 \end{pmatrix} \overset{(1)}{=} \begin{pmatrix} |z|^2 e_0 + \bar z z(-e_0) \\ z(-e_0) + z\,e_0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

(1) Apply each entry: $(S^* + |z|^2)e_0 = |z|^2\,e_0$, $(S^*-1)ze_0 = -ze_0$, and $(|z|^2\,S^* + 1)ze_0 = ze_0$. So $\operatorname{Ker} T_z = \mathbb{C}\,u\otimes e_0 \cong L_z$, each T_z has index 1, and the index bundle is $[L]$. By Example 3.92, $c_1(L) = -1$, so

$$\operatorname{ch}(\operatorname{index} T) = 1 - h, \qquad \int_{\mathbb{CP}^1} \operatorname{ch}(\operatorname{index} T) = -1.$$

The reduced class $\operatorname{index} T - 1 = [L] - [\mathbb{C}]$ is the Bott generator of $\widetilde K(S^2) \cong \mathbb{Z}$ (in the tautological convention of Example 3.103). We have written the generator of Bott periodicity as an explicit family of operators on Hilbert space. Its topology is invisible in any single T_z (each is just "a shift in some direction"), and visible only in how the zero mode $u \otimes e_0$ rotates as z goes around the sphere: it winds like the monopole wavefunction.

3.21.6 The Atiyah–Jänich theorem and Kuiper's theorem

Key idea: Atiyah–Jänich (Nash: "it can be shown that $[X,\mathcal{F}]$ is isomorphic to $K(X)$")

For compact X, the index bundle is an isomorphism

$$\operatorname{index} : [X,\mathcal{F}] \xrightarrow{\cong} K(X), \qquad \operatorname{index}(ST) = \operatorname{index} S + \operatorname{index} T.$$

Composition of families corresponds to addition in K. In particular $[X,\mathcal{F}]$ is not just a semigroup but a group, with the inverse of $[T]$ given by $[T^*]$.

We have shown that index is well defined and onto. Additivity comes from the families version of the six-term sequence (3.213). Here is why it is one-to-one, as a narrative.

Suppose $\operatorname{index} T = 0$ in $K(X)$, i.e. $[K] = [\mathbb{C}^N]$. Adding more trivial directions to the

stabilisation, we may assume $K \cong \mathbb{C}^N$ is actually trivial. Pick a bundle isomorphism $\phi_x : K_x \to \mathbb{C}^N$ and change the family on the finite-rank piece so that the new operator sends K_x isomorphically onto the auxiliary directions. The result is a family of *invertible* operators, homotopic to the original one through Fredholm families. So the question becomes: *is every family of invertible operators homotopic to the constant family* 1*?* In finite dimensions the answer would be no ($\mathrm{GL}(1,\mathbb{C}) = \mathbb{C}^*$ has a loop $e^{i\theta}$ that does not contract, detected by the winding of the determinant). In infinite dimensions the answer is yes:

Key idea: Kuiper's theorem

The group $\mathrm{GL}(H)$ of invertible bounded operators on an infinite-dimensional separable Hilbert space, and its unitary subgroup $\mathrm{U}(H)$, are *contractible* in the norm topology. So every map $X \to \mathrm{GL}(H)$ is homotopic to the constant map 1, and $[X, \mathrm{GL}(H)]$ has just one element.

The core of the proof is the **Eilenberg swindle**, and it is worth seeing because it is the Hilbert hotel once more.

Step 1: $u \oplus u^{-1}$ is connected to 1. For any unitary u on a space V, work on $V \oplus V$ with the rotation $R_t = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix}$ (entries are multiples of 1_V). Define the path $h_t = \begin{pmatrix} u & 0 \\ 0 & 1 \end{pmatrix} R_t \begin{pmatrix} u^{-1} & 0 \\ 0 & 1 \end{pmatrix} R_t^{-1}$, $0 \le t \le \frac{\pi}{2}$. Then $h_0 = 1$ and

$$h_{\pi/2} \overset{(1)}{=} \begin{pmatrix} u & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & u^{-1} \end{pmatrix} = \begin{pmatrix} u & 0 \\ 0 & u^{-1} \end{pmatrix}.$$

(1) Conjugating a diagonal matrix by the quarter turn $R_{\pi/2} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ swaps its entries: $R_{\pi/2}\,\mathrm{diag}(a,b)R_{\pi/2}^{-1} = \mathrm{diag}(b,a)$.

Every h_t is unitary, and h_t depends continuously on t (and on x, if $u = u_x$ is a family).

Step 2: the swindle. Identify H with an infinite direct sum $V \oplus V \oplus V \oplus \cdots$ (possible since $\ell^2 \otimes V \cong H$ for $V = H$). Consider $U = \mathrm{diag}(u, 1, 1, 1, \dots)$. Apply Step 1 simultaneously to the slots $(2,3), (4,5), \dots$, running $1 \oplus 1 \to u^{-1} \oplus u$ in each pair (Step 1 with u^{-1} in place of u):

$$\mathrm{diag}(u, \underbrace{1,1}, \underbrace{1,1}, \dots) \simeq \mathrm{diag}(u, \underbrace{u^{-1}, u}, \underbrace{u^{-1}, u}, \dots) = \mathrm{diag}(\underbrace{u, u^{-1}}, \underbrace{u, u^{-1}}, \dots) \simeq 1.$$

In the middle we only re-bracketed; in the last step we apply Step 1 to the pairs $(1,2), (3,4), \dots$ The same path is used in every block, so the homotopy is uniformly continuous in norm. The hard part of Kuiper's proof, which we skip, is to show that every u can first be deformed into the form $\mathrm{diag}(u', 1, 1, \dots)$.

Watch out: $\mathrm{U}(H)$ is contractible, but $\mathrm{U}(\infty)$ is not

In the previous part, $\mathrm{U}(\infty) = \bigcup_n \mathrm{U}(n)$ had $\pi_1 = \mathbb{Z}$ and the whole of Bott periodicity. There is no contradiction with Kuiper: $\mathrm{U}(\infty)$ consists of unitaries that differ from 1 in only finitely many entries, and the swindle needs infinitely many nontrivial blocks. The larger group $\mathrm{U}(H)$ (all unitaries, norm topology) contains the swindle and is contractible.

Physical picture: $\mathcal{F}$ is a classifying space for K-theory

Combine the pieces. $\mathrm{GL}(H)$ acts on $\mathcal{F}$ by composition and is contractible, and what survives in $\mathcal{F}$ is exactly the finite-dimensional kernel–cokernel data. The Atiyah–Jänich

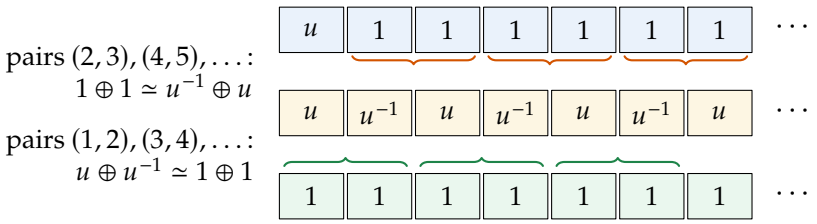


Figure 3.84: The Eilenberg swindle. Top row: u in the first slot, 1 everywhere else. Rotating each orange pair turns the row into the alternating pattern in the middle. Re-bracketing into the green pairs and rotating again gives 1 everywhere. The winding carried by u has been pushed off to infinity, one room at a time. With finitely many slots the last u would be left over, which is why $\mathrm{U}(n)$ is not contractible.

theorem says that $\mathcal{F}$ is a classifying space for K:

$$K(X) = [X, \mathcal{F}], \qquad \mathcal{F} \;\simeq\; \mathbb{Z} \times BU,$$

where the $\mathbb{Z}$ labels the components by the index (the virtual dimension) and each component looks like the infinite Grassmannian BU that classified bundles in the previous part ($\widetilde{K}(X) = [X, BU]$ for connected X). The Grassmannian is the "geometric" model of K-theory; $\mathcal{F}$ is the "analytic" model. The index bundle translates between them.

3.21.7 Where Fredholm families come from: elliptic operators

Nash ends with: "Since elliptic operators are Fredholm operators when viewed as operators in Hilbert space, one might expect $K(X)$ to turn up in a discussion of the index of elliptic operators." Let us see this concretely.

Worked example 3.140: The Dirac operator on the circle is Fredholm

Let $D = -i\,\mathrm{d}/\mathrm{d}\theta : H^1(S^1) \to L^2(S^1)$. In the Fourier basis $De^{in\theta} = n\,e^{in\theta}$, a diagonal operator with entries $n \in \mathbb{Z}$.

- $\operatorname{Ker} D$: the mode $n = 0$, the constants. Dimension 1.
- $\operatorname{Im} D$: every mode except $n = 0$, so $\operatorname{Coker} D$ is again the constants (D is self-adjoint, so $\operatorname{Ker} D^* = \operatorname{Ker} D$). Dimension 1.
- Index $1 - 1 = 0$.

The parametrix is "divide by the symbol": $Qe^{in\theta} = \frac{1}{n}e^{in\theta}$ for $n \neq 0$, $Q1 = 0$. Then

$$QDe^{in\theta} \overset{(1)}{=} Q\,n\,e^{in\theta} \overset{(2)}{=} \begin{cases} e^{in\theta}, & n \neq 0, \\ 0, & n = 0, \end{cases}$$

(1) The action of D on a Fourier mode.
(2) For $n \neq 0$ the factors n and $\frac{1}{n}$ cancel; for $n = 0$ the mode was already killed by D.

so $QD = 1 - P_0$ and similarly $DQ = 1 - P_0$, with P_0 the rank-one projection onto the constants. Atkinson's theorem: D is Fredholm. The mechanism is general: the symbol n (in general $\sigma(x, p)$) is invertible for large $|n|$, and only finitely many low modes can fail.

Worked example 3.141: A non-elliptic operator is not Fredholm

On the torus T^2 let $D = \partial/\partial\theta_1$. On the modes $e^{i(n_1\theta_1+n_2\theta_2)}$ it multiplies by in_1. Every mode with $n_1 = 0$ is in the kernel: the functions of θ_2 alone. The kernel is infinite-dimensional, so D is not Fredholm. The symbol ip_1 vanishes on the nonzero covectors $p = (0, p_2)$: the operator is not elliptic, and nothing controls the modes in that direction.

Worked example 3.142: Operators with nonzero index: $\bar\partial$ on $\mathbb{CP}^1$ and on T^2

Take the Cauchy–Riemann operator $\bar\partial$ acting on sections of the monopole bundle $L_n = \mathcal{O}(n)$ over $\mathbb{CP}^1$ (between the appropriate Sobolev spaces). Its kernel is the space of holomorphic sections, the polynomials of degree $\le n$ in the affine coordinate, of dimension $n+1$ for $n \ge 0$; its cokernel is the Dolbeault group $H^{0,1}(\mathbb{CP}^1; \mathcal{O}(n))$, which vanishes for $n \ge -1$ (both computed with Čech cohomology in part 3a). So

$$\operatorname{index} \bar\partial_{\mathcal{O}(n)} = n + 1 = \int_{\mathbb{CP}^1} \operatorname{ch}(L_n)\, \operatorname{td}(T\mathbb{CP}^1),$$

the last equality being Example 3.110. On the torus $T^2 = \mathbb{C}/(\mathbb{Z} + i\mathbb{Z})$, $\bar\partial$ on functions has kernel the constants and cokernel spanned by $d\bar z$ (the Dolbeault group $H^{0,1}(T^2)$ is one-dimensional), so its index is $0 = \int_{T^2} \operatorname{td}(TT^2)$, because the tangent bundle of the torus is trivial and $c_1 = 0$. In each case the left side is Fredholm theory and the right side is the characteristic-class calculus of this chapter.

Key idea: The index theorem, as a preview of Chapter 4

For an elliptic operator D on a compact manifold M, the *analytical index* $\dim\operatorname{Ker} D - \dim\operatorname{Coker} D$ (this section) equals a *topological index* built from characteristic classes (Sections 3.18 and 3.20). The two most important cases:

$$\operatorname{index} \bar\partial_E = \int_M \operatorname{ch}(E)\, \operatorname{td}(TM), \qquad \operatorname{index} \not{D}_E = \int_M \operatorname{ch}(E)\, \hat A(TM).$$

The Toeplitz theorem (3.209) is the one-dimensional ancestor. For a *family* of elliptic operators D_x, $x \in X$ (a Dirac operator coupled to a family of gauge fields, say), the index is an element of $K(X)$ by this section, and the families index theorem computes its Chern character. Its degree-2 part is the first Chern class of the determinant line bundle, which is the gauge anomaly (Chapter 10).

Aside: self-adjoint families and spectral flow

A self-adjoint Fredholm operator has index 0 ($\operatorname{Ker} D^* = \operatorname{Ker} D$), so a single one carries no integer. But a *loop* of them does. Take the Dirac operator on a circle coupled to a constant U(1) gauge field A (a Wilson line),

$$D_A = -i\frac{\mathrm{d}}{\mathrm{d}\theta} + A, \qquad D_A e^{in\theta} = (n + A)\, e^{in\theta}, \qquad 0 \le A \le 1.$$

The endpoints are gauge equivalent. With $g = e^{i\theta}$,

$$g^{-1} D_0 (gu) \overset{(1)}{=} e^{-i\theta}\left(-i\frac{\mathrm{d}}{\mathrm{d}\theta}\right)\left(e^{i\theta}u\right) \overset{(2)}{=} e^{-i\theta}\left(e^{i\theta}u - ie^{i\theta}u'\right) = D_1 u,$$

(1) Conjugate D_0 by multiplication by g.

(2) Product rule: $-i(e^{i\theta}u)' = e^{i\theta}u - ie^{i\theta}u'$.

so $D_1 = g^{-1}D_0 g$ has the same spectrum as D_0, and $A \in [0,1]$ is really a loop. Yet along the loop the eigenvalue $-1 + A$ rises from -1 to 0 and the eigenvalue A rises from 0 to 1: the spectrum as a whole returns to itself, but it has shifted up by one unit. Exactly one eigenvalue crosses zero from below. This net number of crossings is the **spectral flow**, here +1. It is a homotopy invariant of loops of self-adjoint Fredholm operators, just as the index is for single operators, and such loops (more generally, maps $X \to \mathcal{F}^{\mathrm{sa}}$) classify the K-group $K^{-1}(X)$ in the same way that $\mathcal{F}$ classifies $K(X)$ (Atiyah–Singer).

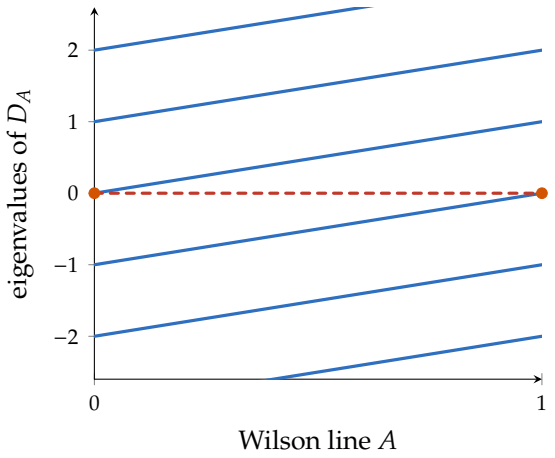


Figure 3.85: Spectral flow. The eigenvalues $n + A$ of the circle Dirac operator D_A as the Wilson line A goes from 0 to 1. The spectrum at $A = 1$ is the same set as at $A = 0$ (the two operators are gauge equivalent), but every level has moved up by one place: one eigenvalue has crossed the red line 0 from below. In a second-quantised theory this is one unit of fermion number created by an adiabatic change of flux: the mechanism behind level-crossing pictures of the chiral anomaly (Chapter 10).

Dictionary: Fredholm theory

Mathematics	Physics
$\operatorname{Ker} D$, $\operatorname{Ker} D^*$	zero modes of D and of $D^\dagger$ (e.g. left- and right-handed fermion zero modes)
$\operatorname{index} D$	net number of zero modes; Witten index $\operatorname{Tr}(-1)^F e^{-\beta\mathcal{H}}$
Fredholm	well-posed up to finitely many zero modes and constraints
compact perturbation	lower-order terms, masses, smooth potentials: UV-irrelevant
parametrix	propagator with zero modes removed
family over X; index bundle in $K(X)$	zero modes varying over a parameter space (moduli, Wilson lines, background fields)
spectral flow	level crossing; fermion number created by an adiabatic process
Kuiper's theorem	in infinite dimensions no winding survives in the invertible operators; only zero modes carry topology

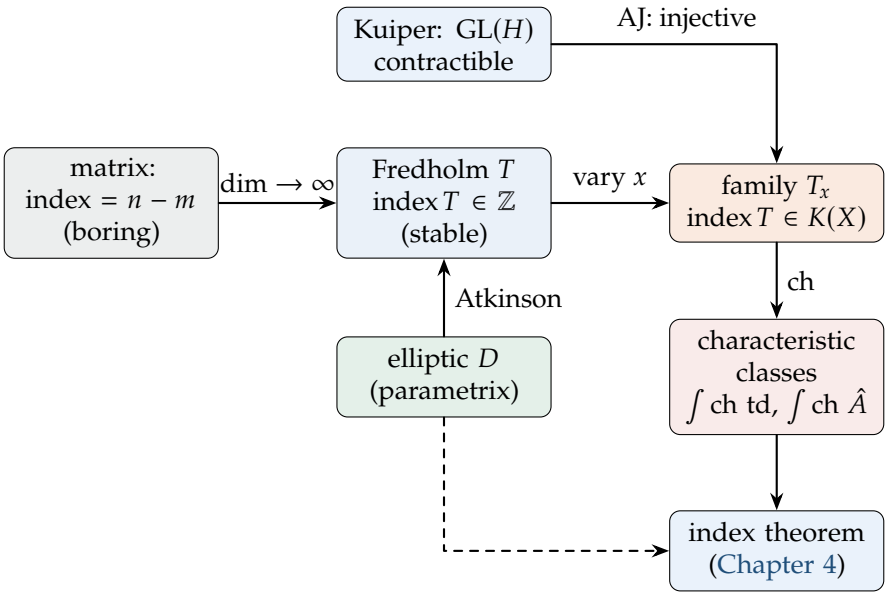


Figure 3.86: The logic of this section. Infinite dimensions make the index a real invariant; parameters promote it to a K-theory class (Atiyah–Jänich, with Kuiper's theorem ensuring nothing else survives); elliptic operators supply the Fredholm operators; the Chern character turns the K-class into cohomology, where the characteristic classes of §5 compute it. The index theorem is the dashed arrow.

Reading Nash (p. 88, §6, (3.128))

Nash's whole section is one paragraph. He introduces $\mathcal{F}$, "the space of all Fredholm operators", without defining Fredholm; we did so in Section 3.21.3 (finite-dimensional kernel and cokernel), with the norm topology understood. His "product of two Fredholm operators is also Fredholm" is our (3.213), which also shows that the index turns products into sums. He calls $[X, \mathcal{F}]$ a semigroup; by the Atiyah–Jänich theorem it is in fact a group, isomorphic to $K(X)$. "Actually it can be shown" hides two ingredients: the index bundle construction (Section 3.21.5) and Kuiper's theorem (Section 3.21.6). His (3.128), $\dim\ker F - \dim\operatorname{coker} F$, is the case $X =$ point. His closing sentence, that elliptic operators are Fredholm "when viewed as operators in Hilbert space", means between Sobolev spaces (see the pitfall in Section 3.21.2), and it is Atkinson's theorem applied to the parametrix of Chapter 2. The standard references for this material are Atiyah's *K-theory* (appendix) and Lawson–Michelsohn, *Spin Geometry*, §III.7.

Recap

- $\operatorname{index} T = \dim\operatorname{Ker} T - \dim\operatorname{Coker} T = \dim\operatorname{Ker} T - \dim\operatorname{Ker} T^*$. For matrices it is $n - m$; in infinite dimensions it is a real invariant.
- Fredholm: finite kernel and cokernel $\Leftrightarrow$ invertible modulo compacts (Atkinson). Examples: shift S (index -1), a (index $+1$, hence no unitary phase operator), Toeplitz T_f (index $-$ wind f).
- Stability: nonzero modes of T^*T and TT^* pair up; near T every operator is "invertible $\oplus$ finite matrix $\operatorname{Ker} T \to \operatorname{Coker} T$" (Schur complement). The index is locally constant, homotopy invariant, blind to compact perturbations, and additive: $\operatorname{index} ST = \operatorname{index} S + \operatorname{index} T$.

- $\pi_0(\mathcal{F}) \cong \mathbb{Z}$ (Nash (3.128)). A family $T : X \to \mathcal{F}$ has an index bundle $[K]-[\mathbb{C}^N] \in K(X)$ (stabilise to kill cokernels).
- Atiyah–Jänich: $[X, \mathcal{F}] \cong K(X)$. Onto via $T^E = P \otimes S^* + (1 - P) \otimes 1$; one-to-one via Kuiper ($\mathrm{GL}(H)$ contractible, Eilenberg swindle). The Bott class is the family of Example 3.139.
- Elliptic operators are Fredholm (parametrix from the symbol); their indices are computed by characteristic classes: the index theorem. Self-adjoint loops carry spectral flow.

Chapter summary

This chapter had one theme, played three times. Physics hands us objects that are easy to write down on small patches and hard to write down globally: a vector potential, a wavefunction in a monopole field, a solution of a differential equation, a family of zero modes. The chapter built three machines that measure the global twisting of such objects.

1. *Sheaf cohomology* (Nash §1–§2) measures the failure of local solutions to glue. Its outputs are groups such as $H^1(M;\mathcal{O}^*)$ (line bundles) and $H^2(M;\mathbb{Z})$ (their magnetic charges), and two theorems, de Rham and Dolbeault, that let us compute them with differential forms.

2. *K-theory* (Nash §3–§4) counts vector bundles up to adding trivial ones, and turns the counting into a ring. Over spheres it is a homotopy group of a Lie group (clutching), and Bott periodicity says these groups repeat with period 2 (complex) or 8 (real).

3. *Characteristic classes* (Nash §5) turn a bundle into cohomology classes that can be computed from any gauge field by integrating invariant polynomials of the curvature. *Fredholm operators* (Nash §6) turn families of operators into K-theory classes. Together they set up the index theorem: an analytic count of zero modes on one side, an integral of characteristic classes on the other.

Figure 3.87 draws the logic, and the tables below list every concept of the chapter with its operational meaning and where it is used again.

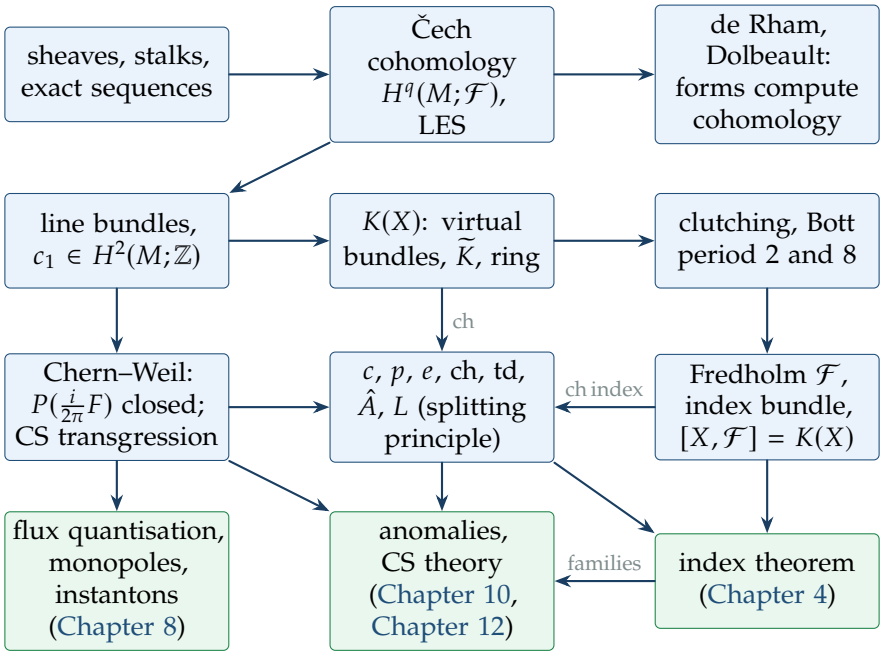


Figure 3.87: Chapter 3 in one picture. Top row: local data and its cohomology (Nash §1–§2). Middle row: bundles and K-theory (§3–§4). Third row: characteristic classes and Fredholm operators (§5–§6). The green boxes are where the later chapters use the machinery. (The de Rham theorem is what lets the curvature forms of the third row represent the integral classes of the second.) Every arrow into the index theorem passes through the third row: the Chern character takes the K-theory class of an operator family into cohomology, where the classes td and $\hat{A}$ evaluate it.

Concept table I: sheaves and their cohomology (Nash §1–§2)

Concept	Operational meaning: how to compute or test it	Used later
presheaf, sheaf $\mathcal{F}$	A group $\mathcal{F}(U)$ for each open U with restrictions. Sheaf: sections are determined by, and glue from, compatible local pieces. Test: is the defining condition *local*? (smooth, holomorphic, solves a PDE: yes; bounded, constant: no).	sheaf cohomology of $\mathcal{O}(D)$ in Chapter 5; strings Chapter 9
stalk, germ $\mathcal{F}_x$	Sections near x, two identified if they agree on a smaller neighbourhood. For $\mathcal{O}$: convergent power series at x.	exactness tests, Chapter 5
sheaf morphism; exact sequence	A local operation commuting with restriction (d, $\bar\partial$, $\exp 2\pi i$). Exact means exact on every stalk, i.e. locally solvable.	resolutions in Chapter 4, Chapter 5
exponential sequence	$0 \to \mathbb{Z} \to \mathcal{O} \xrightarrow{e^{2\pi i\,\cdot}} \mathcal{O}^* \to 0$; it is exact because logarithms exist locally but not globally.	c_1 of line bundles, divisors (Chapter 5)
fine / soft sheaf	Sections can be cut off with bump functions ($\mathcal{E}^p$, sections of a smooth bundle). Then $H^q = 0$ for $q > 0$.	every de Rham-type argument
Čech cochains, δ, $H^q(\mathcal{U};\mathcal{F})$	One section per $(q+1)$-fold overlap; δ = alternating restriction; cocycles modulo coboundaries. H^0 = global sections; H^1 = twisted gluing data modulo re-gauging. Use a good (Leray) cover.	line bundles, instanton patching (Chapter 8)
long exact sequence, δ^*	Zig-zag: lift locally, apply δ, pull back. For the exponential sequence on $\mathbb{C}^*$, δ^* is the winding number.	c_1; Riemann–Roch arguments in Chapter 5
de Rham theorem	$H^q(M;\mathbb{R}) \cong H^q_{\mathrm{dR}}(M)$ via the fine resolution $\underline{\mathbb{R}} \to \mathcal{E}^0 \to \mathcal{E}^1 \to \cdots$. The Čech class of a closed form is read off by the zig-zag $\omega = \mathrm{d}\alpha$, $\delta\alpha = \mathrm{d}\lambda$, $\delta\lambda = n$.	Betti numbers in Morse theory (Chapter 7); index of $\mathrm{d}+\mathrm{d}^*$ (Chapter 4)
complex manifold, (p,q)-forms, $\bar\partial$	Holomorphic transition functions split forms by type; $\partial_{\bar z} = \frac{1}{2}(\partial_x + i\partial_y)$; holomorphic $\iff \bar\partial f = 0$.	Dolbeault complex (Chapter 4), Riemann surfaces (Chapter 5, Chapter 9)
Dolbeault theorem	$H^q(M;\Omega^p) \cong H^{p,q}_{\bar\partial}(M)$, also with coefficients in a holomorphic bundle. Example: $H^{0,1}(T^2) = \mathbb{C}[\mathrm{d}\bar z]$, $\dim H^1(\mathbb{CP}^1;\mathcal{O}(n)) = \max(0,-n-1)$.	Riemann–Roch as an index (Chapter 4, Chapter 5); moduli of Riemann surfaces
line bundles $= H^1(M;\mathcal{E}^*)$; c_1	Transition functions g_{ij} form a 1-cocycle; $c_1 = [\delta\lambda]$, $\lambda_{ij} = \frac{i}{2\pi}\log g_{ij}$; de Rham image $[\frac{i}{2\pi}F]$.	monopoles (Chapter 8), determinant line bundle (Chapter 10)
Dirac quantisation	$2eg_m \in \mathbb{Z}$ is the cocycle condition on triple overlaps; $c_1 = 2eg_m$.	monopoles (Chapter 8)

Concept table II: K-theory and Bott periodicity (Nash §3–§4)

Concept	Operational meaning: how to compute or test it	Used later
vector bundle by transition functions	Matrices g_{ab} on overlaps with $g_{ab}g_{bc} = g_{ac}$; isomorphic if $g' = h_a g h_b^{-1}$. Examples: Möbius (± 1), Hopf (z), TS^2 (z^{-2}), monopole ($e^{in\phi}$).	everywhere: gauge fields are connections on bundles
$\oplus$, $\otimes$, dual, pullback	Block-diagonal, Kronecker product, inverse transpose, composition of the transition functions.	twisted operators (Chapter 4)
stable equivalence	$E \oplus I^j \cong F \oplus I^k$. Cancellation can fail: $TS^2 \oplus I \cong I^3$ but $TS^2 \not\cong I^2$. Every bundle over a compact base has a complement.	$\widetilde{K}$, stable classes
$K(X)$, $\widetilde{K}(X)$	Formal differences $[E] - [F]$ (Grothendieck group); every class is $[E] - [I^n]$; virtual dimension; $K = \widetilde{K} \oplus \mathbb{Z}$; ring under $\otimes$. $K(S^2) = \mathbb{Z}[H]/(H-1)^2$.	symbol class of an elliptic operator (Chapter 4); families (Chapter 10)
reduction of structure group	A fibre metric reduces $\mathrm{GL}(k)$ to $\mathrm{O}(k)$, $\mathrm{U}(k)$ or $\mathrm{Sp}(k)$ (polar decomposition), losing no topology.	unitary gauge groups; orthonormal frames
Grassmannian, classifying map	$x \mapsto E_x \subset \mathbb{C}^m$ (a projector Π_x); $E = f^*Q$; $\mathrm{Vect}_k(X) = [X, BU(k)]$, $BU(k) = \mathrm{Gr}(k, \infty)$.	definition of characteristic classes (Section 3.15)
stable range	If $k > \dim X$ (real), a generic section never vanishes, so $E_k \cong F \oplus I^{k-n}$ and $\mathrm{Vect}_k = \widetilde{K}$.	dimension counts in Chapter 8
clutching	A bundle over S^n = a map $g : S^{n-1} \to G$ up to homotopy: $\mathrm{Vect}_k(S^n) = \pi_{n-1}(G)$. c_1 = winding of $g : S^1 \to \mathrm{U}(1)$; c_2 = degree of $g : S^3 \to \mathrm{SU}(2)$.	instanton number (Chapter 8); global anomalies (Chapter 10)
tables $\widetilde{K}(S^n)$	$\widetilde{KU}$: $\mathbb{Z}, 0, \mathbb{Z}, 0, \ldots$ $\widetilde{KO}$: $\mathbb{Z}, \mathbb{Z}_2, \mathbb{Z}_2, 0, \mathbb{Z}, 0, 0, 0$ (n mod 8) $\widetilde{KSp}(S^n) = \widetilde{KO}(S^{n+4})$.	mod-2 indices, Witten's SU(2) anomaly ($\pi_4(\mathrm{SU}(2)) = \mathbb{Z}_2$)
Bott periodicity	$\Omega^2 \mathrm{U} \simeq \mathrm{U}$, so $\widetilde{KU}^{-n}(X) \cong \widetilde{KU}^{-n-2}(X)$, implemented by multiplication with $b = [H] - 1$. Real: eight loop spaces, or Clifford modules.	Thom isomorphism and proof of the index theorem (Chapter 4); loop groups (Chapter 6)
K^{-n}, external product	$\widetilde{K}^{-n}(X) = \widetilde{K}(S^n X)$; $K^{-1}(X) = [X, \mathrm{U}]$; $\pi_X^* a \cdot \pi_Y^* b$ builds classes on products.	families index, spectral flow (K^{-1})

Concept table III: characteristic classes (Nash §5)

Concept	Operational meaning: how to compute or test it	Used later
Chern classes $c_i(E) \in H^{2i}(X;\mathbb{Z})$	Pull back the generators of $H^*(BU(k)) = \mathbb{Z}[c_1,\dots,c_k]$ by the classifying map (Section 3.15). Naturality, Whitney $c(E\oplus F) = c(E)c(F)$, $\int_{\mathbb{CP}^1} c_1(\text{taut.}) = -1$.	every index formula (Chapter 4)
Chern–Weil forms	$c(F) = \det\big(1 + \frac{i}{2\pi}F\big)$; $c_1 = \frac{i}{2\pi}\operatorname{tr} F$, $c_2 = \frac{1}{8\pi^2}(\operatorname{tr}F^2 - \operatorname{tr}F\wedge\operatorname{tr}F)$. Closed by Bianchi; class independent of A by transgression (Section 3.16).	instanton number $k = \int c_2$ (Chapter 8)
Chern–Simons form	$\operatorname{tr}F^2 = \mathrm{d}\,\operatorname{tr}(A\,\mathrm{d}A + \frac{2}{3}A^3)$. Under a large gauge transformation it shifts by $-\frac{1}{3}\operatorname{tr}(g^{-1}\mathrm{d}g)^3$, a winding number.	secondary classes (Chapter 8); descent and anomalies (Chapter 10); CS theory (Chapter 12)
splitting principle, Chern roots	Pretend $E = L_1\oplus\cdots\oplus L_k$: $c(E) = \prod(1+x_i)$, $c_j = \sigma_j(x)$. With curvature, the x_i are the eigenvalues of $\frac{i}{2\pi}F$ (Section 3.17). Newton: $s_2 = c_1^2 - 2c_2$.	every multiplicative class below
Chern character ch	$\operatorname{ch}E = \sum e^{x_i} = \operatorname{tr}e^{iF/2\pi} = k + c_1 + \frac{1}{2}(c_1^2 - 2c_2) + \dots$; additive and multiplicative, so a ring map $K(X)\to H^{\text{even}}(X;\mathbb{Q})$; blind to torsion (Section 3.18).	the "bundle" factor in every index formula; anomaly polynomial
Todd class td	$\prod \frac{x_i}{1-e^{-x_i}} = 1 + \frac{c_1}{2} + \frac{c_1^2+c_2}{12} + \frac{c_1c_2}{24} + \dots$; Todd genus of $\mathbb{CP}^n$ is 1.	Riemann–Roch–Hirzebruch (Chapter 4, Chapter 5)
dual, conjugate, line bundle group	E^*: $(g^{-1})^T$; $\bar{E}$: $\bar{g}$; equal for unitary g; $c_j(E^*) = (-1)^j c_j(E)$. Line bundles under $\otimes$: c_1 identifies them with $H^2(X;\mathbb{Z})$ (Section 3.19).	antiparticles, charge conjugation; Chapter 5
complexification, realification	$V_{\mathbb{C}} = V\otimes_{\mathbb{R}}\mathbb{C} \cong \bar{V}_{\mathbb{C}}$; $E_r\otimes\mathbb{C}\cong E\oplus\bar{E}$ (the $\pm i$ eigenspaces of J).	real families, mod-2 indices
Pontryagin classes $p_i \in H^{4i}$	$p_i(V) = (-1)^i c_{2i}(V_{\mathbb{C}})$; $p(\Omega) = \det(1+\Omega/2\pi)$; roots $\pm x_i$, $p_j = \sigma_j(x^2)$; $p(TS^n) = 1$, $p(T\mathbb{CP}^n) = (1+h^2)^{n+1}$; for complex E: $p_1 = c_1^2 - 2c_2$.	gravitational anomalies (Chapter 10); framing in Chapter 12
Euler class e	$e(V) = \operatorname{Pf}(\Omega/2\pi)$ for oriented even rank; $e^2 = p_{top}$; $e(E_r) = c_{top}(E)$; $\int e$ counts zeros of a section with signs (Section 3.20).	Poincaré–Hopf in Morse theory (Chapter 7); Gauss–Bonnet as an index (Chapter 4)
Gauss–Bonnet–Chern	$\chi(M) = \int_M \operatorname{Pf}(\Omega/2\pi)$; $\chi(S^2) = 2$, $\chi(\Sigma_g) = 2-2g$, $\chi(S^4) = 2$, $\chi(\mathbb{CP}^n) = n+1$.	string genus expansion (Chapter 9); Chapter 4
$\hat{A}$ and L	$\hat{A} = \prod \frac{x_i/2}{\sinh(x_i/2)} = 1 - \frac{p_1}{24} + \frac{7p_1^2-4p_2}{5760}$; $L = \prod \frac{x_i}{\tanh x_i} = 1 + \frac{p_1}{3} + \frac{7p_2-p_1^2}{45}$; $\operatorname{td} = e^{c_1/2}\hat{A}$; $\sigma(M) = \int L$.	Dirac index, Hirzebruch signature (Chapter 4); Donaldson (Chapter 12)

Concept table IV: Fredholm operators and $K(X)$ (Nash §6)

Concept	Operational meaning: how to compute or test it	Used later
Fredholm operator, index	Bounded T with finite-dimensional kernel and cokernel; index $T = \dim \operatorname{Ker} T - \dim \operatorname{Ker} T^*$. Count zero modes of T and of $T^\dagger$ (Section 3.21).	Chapter 4 throughout
shift, ladder, Toeplitz	index $S = -1$; index $a = +1$ (so no unitary phase operator); index $T_f = -\operatorname{wind}(f)$: the first index theorem.	loop groups and Fock space (Chapter 6)
stability	Nonzero modes of T^*T and TT^* pair up; near T everything reduces to a finite matrix $\operatorname{Ker} T \to \operatorname{Coker} T$ (Schur complement). Index is locally constant, homotopy invariant, blind to compact perturbations; index ST = index S + index T.	heat-kernel proof, Witten index (Chapter 4, Chapter 7)
Atkinson, parametrix	Fredholm $\iff$ invertible modulo compacts. Elliptic operators have a parametrix built from the inverse symbol, so they are Fredholm.	Chapter 2 → Chapter 4
index bundle	Add $\mathbb{C}^N$ to kill all cokernels; kernels then form a bundle K; index $T = [K] - [\mathbb{C}^N] \in K(X)$.	families index theorem; determinant line (Chapter 10)
Atiyah–Jänich, Kuiper	$[X, \mathcal{F}] \cong K(X)$; $\mathrm{GL}(H)$ is contractible (Eilenberg swindle), so only zero-mode data carries topology; $\mathcal{F} \simeq \mathbb{Z} \times BU$.	anomalies from a Hamiltonian view (Chapter 10)
spectral flow	Net number of eigenvalues crossing 0 along a loop of self-adjoint Fredholm operators; e.g. $D_A = -i\partial_\theta + A$, $A : 0 \to 1$, flow +1.	level crossing and anomalies (Chapter 10); Floer theory (Chapter 12)

The formula sheet: every class from one generating function

Most of §5 is one idea: choose a power series $Q(x)$, and define a class by $\prod_i Q(x_i)$ over the Chern roots (or over the roots $\pm x_i$ of a real bundle). The following table collects them, together with the first terms we derived and the curvature formula that computes them from a gauge field.

Class	Roots	First terms	From curvature
c	$\prod(1 + x_i)$	$1 + c_1 + c_2 + \dots$	$\det\big(1 + \frac{i}{2\pi}F\big)$
ch	$\sum e^{x_i}$	$k + c_1 + \frac{1}{2}(c_1^2 - 2c_2) + \frac{1}{6}(c_1^3 - 3c_1c_2 + 3c_3)$	$\operatorname{tr} \exp \frac{iF}{2\pi}$
td	$\prod \frac{x_i}{1-e^{-x_i}}$	$1 + \frac{1}{2}c_1 + \frac{1}{12}(c_1^2 + c_2) + \frac{1}{24}c_1c_2$	$\det \frac{iF/2\pi}{1-e^{-iF/2\pi}}$
p	$\prod(1 + x_i^2)$	$1 + p_1 + p_2 + \dots$	$\det\big(1 + \frac{\Omega}{2\pi}\big)$
e	$\prod x_i$ (half the roots)	top degree only	$\operatorname{Pf}\big(\frac{\Omega}{2\pi}\big)$
$\hat{A}$	$\prod \frac{x_i/2}{\sinh(x_i/2)}$	$1 - \frac{p_1}{24} + \frac{7p_1^2 - 4p_2}{5760}$	$\det{}^{1/2} \frac{i\Omega/4\pi}{\sinh(i\Omega/4\pi)}$
L	$\prod \frac{x_i}{\tanh x_i}$	$1 + \frac{p_1}{3} + \frac{7p_2 - p_1^2}{45}$	$\det{}^{1/2} \frac{i\Omega/2\pi}{\tanh(i\Omega/2\pi)}$

In the last three rows the matrix functions are defined by their power series, and the square root of a determinant of an even power series in Ω is unambiguous term by term. For a real bundle, Ω is the $\mathfrak{so}(k)$-valued curvature in an orthonormal frame. Viewed as a complex matrix it is anti-Hermitian, so $\frac{i}{2\pi}\Omega$ has real eigenvalues $\pm x_j$: these are the roots, exactly as $\frac{i}{2\pi}F$ supplies the Chern roots of a complex bundle. (This is why an i appears inside $\hat{A}$ and L; without it

sinh would turn into sin and the sign of p_1 would flip.) The sign conventions are fixed in Sections 3.19 and 3.20.

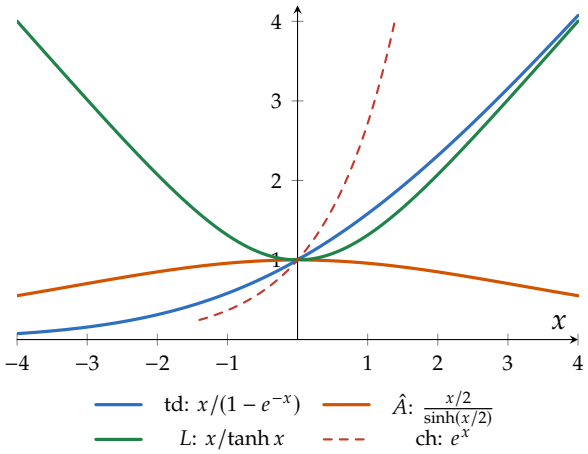


Figure 3.88: The four generating functions of §5, drawn as functions of one Chern root. All equal 1 at $x = 0$, which is why each class starts with 1 (or with the rank, for ch). $\hat{A}$ and L are *even* (orange and green are symmetric), so they involve only x^2: they are Pontryagin classes of a real bundle. The Todd function is not even; its odd part $x/2$ is exactly the factor $e^{c_1/2}$ in $\mathrm{td} = e^{c_1/2}\hat{A}$, and indeed $\frac{x}{1-e^{-x}} = e^{x/2}\frac{x/2}{\sinh(x/2)}$.

We verify the identity quoted in the caption, since it is the bridge between the two index formulas of Chapter 4:

$$e^{x/2}\,\frac{x/2}{\sinh(x/2)} \overset{(1)}{=} e^{x/2}\,\frac{x/2}{\frac{1}{2}\left(e^{x/2}-e^{-x/2}\right)} \overset{(2)}{=} \frac{x\,e^{x/2}}{e^{x/2}\left(1-e^{-x}\right)} = \frac{x}{1-e^{-x}}.$$

(1) The definition $\sinh u = \frac{1}{2}(e^u - e^{-u})$ with $u = x/2$.

(2) Cancel the factors $\frac{1}{2}$, and factor $e^{x/2}$ out of the bracket: $e^{x/2} - e^{-x/2} = e^{x/2}(1 - e^{-x})$.

Taking the product over the roots gives $\mathrm{td}(E) = e^{c_1(E)/2}\,\hat{A}(E)$, because $\prod_i e^{x_i/2} = e^{\sum_i x_i/2} = e^{c_1/2}$. (Here $\hat{A}(E)$ means the same product over the complex roots x_i, which for $E = TM$ of a complex manifold is the $\hat{A}$ of the underlying real bundle.)

The test cases: numbers to check any formula against

The same handful of spaces appeared in every section. Here are their invariants, all computed in this chapter; h is the hyperplane class, with $\int_{\mathbb{CP}^n} h^n = 1$. Any new formula should be checked against this table before it is trusted.

Space	**Classes of** TM	χ	σ	$\int \mathrm{td}$	$\int \hat{A}$
$S^2 = \mathbb{CP}^1$	$c = 1 + 2h, \quad e = c_1$	2	–	1	0
T^2	$c = 1$ (trivial tangent bundle)	0	–	0	0
Σ_g	$\int c_1 = 2 - 2g$	$2-2g$	–	$1-g$	0
S^4	$p = 1$ (stably trivial), $\int e = 2$	2	0	–	0
$\mathbb{CP}^2$	$c = (1+h)^3$, $p = (1+h^2)^3$, $p_1 = 3h^2$	3	1	1	$-\frac{1}{8}$
K3	$c_1 = 0$, $\int c_2 = 24$, $\int p_1 = -48$	24	−16	2	2

Three checks the table passes. (i) On a complex surface $\chi = \int c_2$ (the Euler class is the top Chern class): 3 for $\mathbb{CP}^2$, 24 for K3. (ii) $\sigma = \int L_1 = \frac{1}{3}\int p_1$: $\frac{3}{3} = 1$ and $-\frac{48}{3} = -16$. (iii) On a

4-manifold $\int \hat{A} = -\frac{1}{24}\int p_1 = -\frac{\sigma}{8}$: $-\frac{1}{8}$ for $\mathbb{CP}^2$ (not an integer, because $\mathbb{CP}^2$ is not spin, so this number is not the index of any Dirac operator) and 2 for K3. The Todd genus of a Riemann surface, $\int \frac{c_1}{2} = 1 - g$, is Riemann–Roch for the trivial bundle: $\dim H^0(\mathcal{O}) - \dim H^1(\mathcal{O}) = 1 - g$.

And two bundles that are not tangent bundles:

- the *monopole* bundle L_n over S^2: $c_1(L_n) = nu$ with $\int_{S^2} u = 1$, $\mathrm{ch}(L_n) = 1 + nu$, and index $\bar{\partial}_{L_n} = \int \mathrm{ch}(L_n)\,\mathrm{td}(TS^2) = n + 1$; the tautological (Hopf) bundle has $\int c_1 = -1$;
- the *instanton* bundle over S^4 clutched by $g(x) = (x_4 + i\vec{x}\cdot\vec{\sigma})/|x|$: $c_1 = 0$, $\int c_2 = 1$ in the orientation of Section 3.16, $\mathrm{ch} = 2 - c_2$, and the Dirac index formula $\int \mathrm{ch}\,\hat{A}(TS^4) = -\int c_2$ (as $\hat{A}(TS^4) = 1$), whose absolute value $|k|$ counts the chiral zero modes of the fundamental fermion in the instanton background (Chapter 8).

What the later chapters take from this one

- **Chapter 4 (index theory)** uses almost everything: the symbol of an elliptic operator defines a class in K-theory (§3), Bott periodicity gives the Thom isomorphism that turns it into a number (§4), the Chern character and the Todd and $\hat{A}$ classes write the answer as an integral (§5), and Fredholm theory (§6) makes the analytical index well defined and homotopy invariant. The de Rham, Dolbeault, signature and spin complexes give Gauss–Bonnet–Chern, Riemann–Roch–Hirzebruch, the signature theorem and the $\hat{A}$ theorem, each already met here as a characteristic-class identity.
- **Chapter 5 (algebraic geometry)** uses sheaves of holomorphic sections, $H^1(M;\mathcal{O}^*)$ as the Picard group, divisors via the exponential sequence, and Dolbeault cohomology for Serre duality and Riemann–Roch on Riemann surfaces.
- **Chapter 6 (infinite-dimensional groups)** uses Bott's $\Omega\mathrm{U} \simeq \mathbb{Z}\times BU$, Fredholm and Toeplitz operators on the Hardy space, and c_1 of line bundles for central extensions.
- **Chapter 7 (Morse theory)** uses de Rham cohomology (Betti numbers), Poincaré–Hopf (zeros of sections count χ), and the pairing of nonzero modes behind the Witten index.
- **Chapter 8 (instantons and monopoles)** uses clutching ($\pi_3(\mathrm{SU}(2)) = \mathbb{Z}$), c_2 as the instanton number, c_1 as the monopole charge, and the Chern–Simons form as a secondary characteristic class.
- **Chapter 9 (the elliptic geometry of strings)** uses Riemann surfaces, holomorphic line bundles, Gauss–Bonnet and the Riemann–Roch count of moduli.
- **Chapter 10 (anomalies)** uses the Chern character and $\hat{A}$ in the anomaly polynomial, the Chern–Simons transgression for the descent equations, the index bundle and determinant line of a family of Dirac operators, $\pi_4(\mathrm{SU}(2)) = \mathbb{Z}_2$ for the global anomaly, and spectral flow for the Hamiltonian picture.
- **Chapter 12 (topological field theories)** uses the Chern–Simons functional and its behaviour under large gauge transformations, p_1 and framings, and spectral flow in Floer theory.

Recap: The chapter in five sentences

- Local data glue into global objects with a twist; sheaf cohomology counts the twists, and de Rham and Dolbeault compute it with forms.

- Bundles modulo trivial ones form the ring $K(X)$; over spheres it is a homotopy group of a Lie group, and it repeats with period 2 (or 8).
- Invariant polynomials of the curvature give closed forms whose classes do not depend on the gauge field: c, ch, td, p, e, $\hat{A}$, L, all from generating functions of the Chern roots.
- A family of Fredholm operators has an index in $K(X)$, and every K-class arises this way.
- The index theorem (next chapter) says that the number of zero modes is an integral of characteristic classes.

Chapter 4

Index Theory for Elliptic Operators

Where this chapter is going

A physicist meets the index long before meeting the word. Count the zero modes of a Dirac operator in an instanton background: so many of left-handed chirality, so many of right-handed. Each number separately depends on the details of the gauge field and jumps when we deform it. The *difference* does not. It is an integer that stays put under every smooth deformation, so it must be computable from topology alone. That integer is the **index**, and this chapter is about the theorem of Atiyah and Singer which says exactly how to compute it.

The chapter has a simple logical spine. First we define the index analytically: solve the equation $D\psi = 0$, count solutions, subtract the count for the adjoint equation. Then we notice that this integer depends only on the *leading symbol* of D, a matrix-valued function on phase space T^*X, and only on it up to deformation. Deformation classes of such symbols form a K-theory group $K(T^*X)$. Next we build a second, purely topological integer out of the same symbol class, using embeddings and the Thom isomorphism. The Atiyah–Singer theorem says the two integers agree. The rest of the chapter (§2 onward in Nash) turns the topological side into a formula with characteristic classes, works out the classical examples (Gauss–Bonnet, signature, Riemann–Roch, the Dirac operator), and extends the theorem to families of operators, group actions and fixed points, and manifolds with boundary. Those extensions are exactly what the later physics chapters use: anomalies, the dimension of instanton moduli spaces, moduli of Riemann surfaces and supersymmetric quantum mechanics.

This first part covers Nash's §1. Because Nash is extremely terse here (nine pages which quote rather than explain), we also build the ground floor: finite-dimensional linear algebra, Fredholm operators, explicit operators on the circle, and the heat-kernel formula which is how most physicists actually meet the index theorem.

4.1 Counting solutions: the index of a linear map

Here is the question that starts the whole subject. We have a linear equation $Au = f$. How many solutions does it have, and for which f can it be solved at all? The first question is answered by the kernel of A, the second by its cokernel. Both answers depend delicately on A. We will find a combination of the two that does not, and see that it survives the passage to infinite dimensions, where the differential operators live.

4.1.1 Kernel, image, cokernel

Let $A\colon V \to W$ be a linear map between vector spaces.

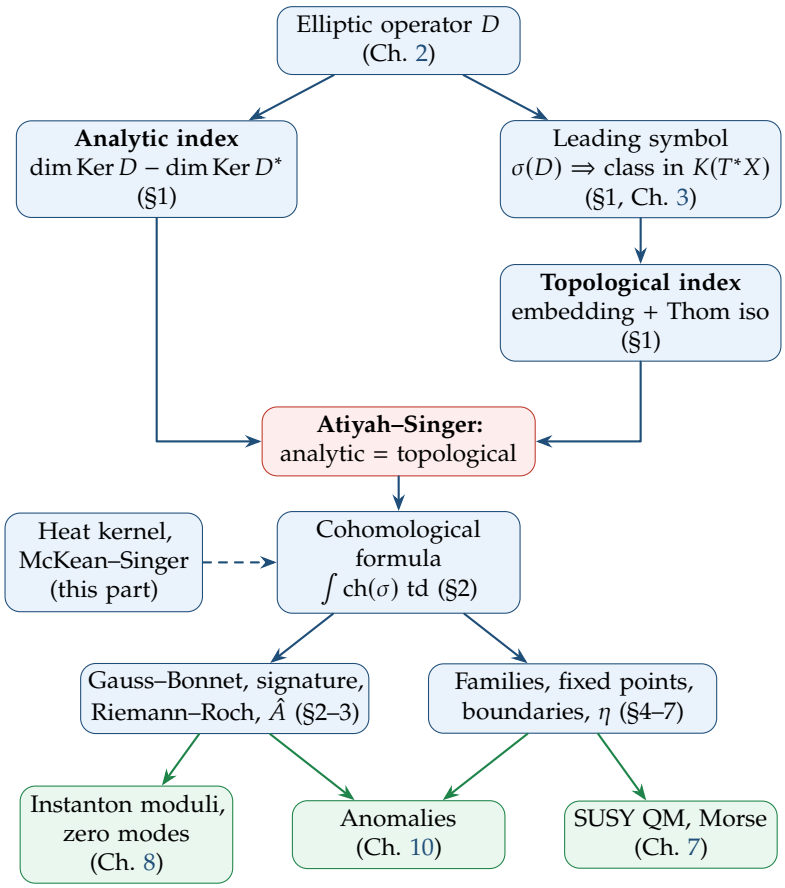


Figure 4.1: The logic of Chapter 4. Two roads leave the elliptic operator D: counting solutions (left) and looking only at the symbol (right). The Atiyah–Singer theorem is the statement that they arrive at the same integer. Everything below the red box is about computing that integer and using it in physics (green).

Working definition 4.1 (kernel, image, cokernel)

- The **kernel** $\operatorname{Ker} A = \{v \in V : Av = 0\}$ is the space of solutions of the homogeneous equation [$Au = f$ with $f = 0$]. Its dimension counts independent solutions ("zero modes").
- The **image** $\operatorname{Im} A = \{Av : v \in V\} \subset W$ is the set of right-hand sides f for which $Au = f$ can be solved.
- The **cokernel** $\operatorname{Coker} A = W/\operatorname{Im} A$ [two vectors of W count as the same if they differ by some Av] measures what is missing from the image. Its dimension counts independent *obstructions* to solving $Au = f$.

Operational test. Write A as a matrix. Then $\dim \operatorname{Ker} A$ is the number of columns minus the rank [the number of linearly independent columns], and $\dim \operatorname{Coker} A$ is the number of rows minus the rank.

The cokernel is a quotient, which is awkward to picture. When W has an inner product there is a concrete stand-in. A vector w is orthogonal to the whole image exactly when

$$0 = \langle w, Av\rangle \ \ \forall v \overset{(1)}{=} \langle A^* w, v\rangle \ \ \forall v \\ \iff A^* w = 0.$$

(1) Definition of the adjoint A^* (the conjugate transpose, for matrices).

So $(\operatorname{Im} A)^\perp = \operatorname{Ker} A^*$, and since $W = \operatorname{Im} A \oplus (\operatorname{Im} A)^\perp$,

$$\operatorname{Coker} A \ \cong\ (\operatorname{Im} A)^\perp \ =\ \operatorname{Ker} A^*. \tag{4.1}$$

The obstructions to solving $Au = f$ are the zero modes of the adjoint equation. This is the "Fredholm alternative" familiar from Green's function problems: $Au = f$ is solvable if and only if f is orthogonal to every solution of $A^* w = 0$.

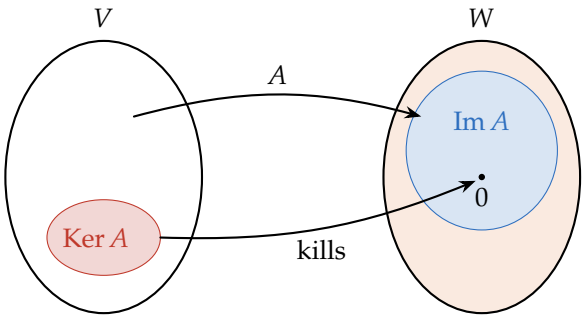


Figure 4.2: A linear map $A\colon V \to W$. The red blob $\operatorname{Ker} A$ is crushed to 0. The image (blue) is where A lands; the part of W outside it (orange) is the cokernel, the directions in which $Au = f$ has no solution.

4.1.2 The index in finite dimensions

Suppose V and W are finite-dimensional. The rank–nullity theorem says $\dim V = \dim \operatorname{Ker} A + \dim \operatorname{Im} A$ (every direction in V is either killed or survives into the image). Then

$$\dim \operatorname{Ker} A - \dim \operatorname{Coker} A \overset{(1)}{=} \dim \operatorname{Ker} A - \big(\dim W - \dim \operatorname{Im} A\big)$$

$$\overset{(2)}{=} \big(\dim \operatorname{Ker} A + \dim \operatorname{Im} A\big) - \dim W$$
$$\overset{(3)}{=} \dim V - \dim W. \tag{4.2}$$

(1) $\operatorname{Coker} A = W/\operatorname{Im} A$, and the dimension of a quotient is the difference of dimensions.
(2) Rearrange.
(3) Rank–nullity.

Working definition 4.2 (index of a linear map)

The **index** of $A\colon V \to W$ is

$$\operatorname{index} A \coloneqq \dim \operatorname{Ker} A - \dim \operatorname{Coker} A \;=\; \dim \operatorname{Ker} A - \dim \operatorname{Ker} A^{*}.$$

[A^{*} is the conjugate-transpose matrix. $A^{*}w = 0$ exactly when w is orthogonal to every Av].
In finite dimensions it equals $\dim V - \dim W$, *whatever* A *is*.

This is the whole philosophy of the chapter in miniature. The two numbers $\dim \operatorname{Ker} A$ and $\dim \operatorname{Coker} A$ depend on the fine details of A. Their difference depends only on the "shape" of the problem: the number of unknowns minus the number of equations. A physicist would call it a counting of degrees of freedom: unknowns minus constraints.

Worked example 4.1: A family of 2×3 matrices

Let $A_\varepsilon\colon \mathbb{C}^3 \to \mathbb{C}^2$ be

$$A_\varepsilon = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \varepsilon & 0 \end{pmatrix}.$$

For $\varepsilon \neq 0$ the two rows are independent, so the rank is 2. For $\varepsilon = 0$ the second row vanishes and the rank is 1. Hence

$$\varepsilon \neq 0: \quad \operatorname{index} A_\varepsilon \overset{(1)}{=} (3-2) - (2-2) = 1 - 0 = 1,$$
$$\varepsilon = 0: \quad \operatorname{index} A_0 \overset{(1)}{=} (3-1) - (2-1) = 2 - 1 = 1.$$

(1) $\dim \operatorname{Ker} = \#\text{columns} - \operatorname{rank}$ and $\dim \operatorname{Coker} = \#\text{rows} - \operatorname{rank}$.

Explicitly, for $\varepsilon \neq 0$ the kernel is spanned by $(0, 0, 1)$ and the map is onto. At $\varepsilon = 0$ the kernel gains $(0, 1, 0)$ and the image loses the direction $(0, 1)$. A zero mode and an obstruction appear *together*, so the index stays $1 = 3 - 2$. See Figure 4.3.

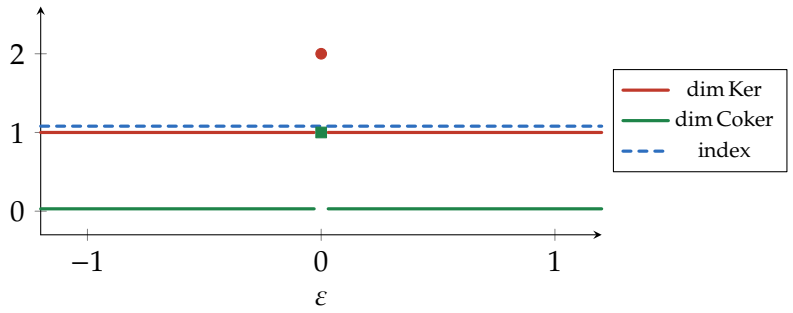


Figure 4.3: The 2×3 family A_ε of the worked example above: as ε passes through 0, $\dim \operatorname{Ker} A_\varepsilon$ (red) jumps from 1 to 2 and $\dim \operatorname{Coker} A_\varepsilon$ (green) from 0 to 1 at the same moment. The index (blue, dashed, drawn slightly offset for visibility) never moves.

Try it: Square matrices

Show that every square matrix has index 0, and find a 2×2 family B_ε whose kernel is 0 for $\varepsilon \neq 0$ and one-dimensional at $\varepsilon = 0$.
Answer. $\dim V = \dim W$, so the index is 0 by (4.2). Take $B_\varepsilon = \operatorname{diag}(\varepsilon, 1)$: at $\varepsilon = 0$ the kernel is spanned by $(1, 0)$ and the cokernel by $(1, 0)$.

4.1.3 Why the index cannot jump: reduction to a finite matrix

In finite dimensions the stability of the index is trivial, because it equals $\dim V - \dim W$. In infinite dimensions that formula is $\infty - \infty$ and means nothing. We need an argument for stability that does not use it. The argument below is worth understanding, because it is the real reason the index of a differential operator is stable. It shows that under a small perturbation, everything interesting happens inside a *finite* matrix.

Let $T\colon H_1 \to H_2$ be a linear map with finite-dimensional kernel $K_1 = \operatorname{Ker} T$ and finite-dimensional cokernel, and suppose the image $R = \operatorname{Im} T$ is closed. Let $C = R^\perp \cong \operatorname{Coker} T$. Split both spaces:

$$H_1 = K_1^\perp \oplus K_1, \qquad H_2 = R \oplus C.$$

In these splittings T is block diagonal,

$$T = \begin{pmatrix} T_0 & 0 \\ 0 & 0 \end{pmatrix}, \qquad T_0\colon K_1^\perp \to R \text{ invertible},$$

because T kills K_1, lands in R, and is one-to-one and onto from the complement of its kernel to its image. Now perturb: $P = T + \varepsilon B$ with B bounded. In blocks,

$$P = \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} T_0 + \varepsilon B_{11} & \varepsilon B_{12} \\ \varepsilon B_{21} & \varepsilon B_{22} \end{pmatrix}.$$

For small ε the corner $a = T_0 + \varepsilon B_{11}$ is still invertible (a small perturbation of an invertible operator is invertible: the geometric series $(T_0 + \varepsilon B_{11})^{-1} = T_0^{-1} \sum_n (-\varepsilon B_{11} T_0^{-1})^n$ converges). Then Gaussian elimination in blocks gives

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \overset{(1)}{=} \begin{pmatrix} 1 & 0 \\ c\,a^{-1} & 1 \end{pmatrix} \begin{pmatrix} a & 0 \\ 0 & d - c\,a^{-1} b \end{pmatrix} \begin{pmatrix} 1 & a^{-1} b \\ 0 & 1 \end{pmatrix}. \tag{4.3}$$

(1) Multiply out the right-hand side: the middle and right factors give $\begin{pmatrix} a & b \\ 0 & d - c a^{-1} b \end{pmatrix}$, and the left factor adds $c a^{-1}$ times the first row to the second, producing $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$.

The outer factors are invertible (triangular with 1's on the diagonal), and invertible factors do not change the dimension of kernel or cokernel. The middle factor is block diagonal with an invertible block a and the **Schur complement**

$$S := d - c\,a^{-1} b\colon K_1 \to C,$$

a map between *finite-dimensional* spaces. Therefore

$$\begin{aligned} \operatorname{index}(T + \varepsilon B) &\overset{(1)}{=} \operatorname{index} a + \operatorname{index} S \\ &\overset{(2)}{=} 0 + \big(\dim K_1 - \dim C\big) \\ &\overset{(3)}{=} \dim \operatorname{Ker} T - \dim \operatorname{Coker} T = \operatorname{index} T. \end{aligned} \tag{4.4}$$

(1) Use (4.3); kernels and cokernels of a block diagonal map are direct sums of those of the blocks.
(2) a is invertible, so its index is 0; S is a map between finite-dimensional spaces, so (4.2) applies.
(3) Definitions $K_1 = \operatorname{Ker} T$ and $C \cong \operatorname{Coker} T$.

Key idea: Stability of the index

Under a small perturbation, the zero modes of T and of T^* can only interact with each other through the finite matrix $S\colon \operatorname{Ker} T \to \operatorname{Coker} T$. The kernel of S may be anything from 0 to $\dim \operatorname{Ker} T$, so $\dim \operatorname{Ker}$ can drop. But whenever a zero mode is lost, an obstruction is lost with it: the index of S is fixed at $\dim \operatorname{Ker} T - \dim \operatorname{Coker} T$.

Worked example 4.2: Watching a pair of zero modes lift

Take $H_1 = H_2 = \mathbb{C}^3$ and $T = \operatorname{diag}(1, 0, 0)$, so $\operatorname{Ker} T$ is spanned by e_2, e_3 and $\operatorname{Coker} T$ by e_2, e_3 as well. Perturb with

$$P_\varepsilon = \begin{pmatrix} 1 & \varepsilon & 0 \\ \varepsilon & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Here $a = 1$, $b = (\varepsilon, 0)$, $c = (\varepsilon, 0)^T$, $d = 0$, so

$$S = d - c\,a^{-1}b \overset{(1)}{=} -\begin{pmatrix} \varepsilon \\ 0 \end{pmatrix}\begin{pmatrix} \varepsilon & 0 \end{pmatrix} = \begin{pmatrix} -\varepsilon^2 & 0 \\ 0 & 0 \end{pmatrix}.$$

(1) $a^{-1} = 1$ and $d = 0$; the product of a column and a row is a 2×2 matrix.

For $\varepsilon \neq 0$, S has a one-dimensional kernel and a one-dimensional cokernel: one zero mode of T and one obstruction of T have "paired up" and left zero together. The index is 0 before and after. Directly: $\det P_\varepsilon$ restricted to the first two coordinates is $-\varepsilon^2 \neq 0$, so $\operatorname{Ker} P_\varepsilon$ is spanned by e_3 alone, and so is its cokernel.

4.1.4 Infinite dimensions: Fredholm operators and the shift

In finite dimensions every index is $\dim V - \dim W$, which is boring. Infinite dimensions are where the index becomes interesting, because now a map from a space to *itself* can have nonzero index.

Working definition 4.3 (Fredholm operator)

A bounded operator [$\|Tv\| \le C\|v\|$ for one constant C] $T\colon H_1 \to H_2$ between Hilbert spaces is **Fredholm** if $\operatorname{Ker} T$ and $\operatorname{Coker} T$ are finite-dimensional (finite dimensionality of the cokernel forces $\operatorname{Im} T$ to be closed [a limit of vectors Tv_n is again of the form Tv]). Its index is $\operatorname{index} T = \dim \operatorname{Ker} T - \dim \operatorname{Coker} T = \dim \operatorname{Ker} T - \dim \operatorname{Ker} T^*$. *Operational test:* T is Fredholm if it is invertible "up to finite-rank errors": there is Q with $QT - 1$ and $TQ - 1$ of finite rank [with finite-dimensional image] (or, more generally, compact [norm limits of finite-rank operators, see below]).

Worked example 4.3: The shift operator: Hilbert's hotel

Let $H = \ell^2$, the square-summable sequences $(x_0, x_1, x_2, \dots)$, with orthonormal basis $e_0, e_1, e_2, \dots$. The **shift** is

$$Se_n = e_{n+1}, \qquad S(x_0, x_1, x_2, \dots) = (0, x_0, x_1, \dots).$$

Its adjoint shifts back: $\langle e_m, Se_n \rangle = \delta_{m,n+1}$, so $S^* e_n = e_{n-1}$ for $n \geq 1$ and $S^* e_0 = 0$. Now

$$\operatorname{index} S \overset{(1)}{=} \dim \operatorname{Ker} S - \dim \operatorname{Ker} S^* \overset{(2)}{=} 0 - 1 = -1.$$

(1) Definition, with $\operatorname{Coker} S \cong \operatorname{Ker} S^*$ by (4.1).

(2) $Sx = 0$ forces every $x_n = 0$; $S^* x = 0$ forces $x_1 = x_2 = \cdots = 0$ and leaves x_0 free, so $\operatorname{Ker} S^* = \mathbb{C} e_0$.

In the same way $\operatorname{index} S^* = +1$ and $\operatorname{index} S^k = -k$. The operator S is a map from ℓ^2 to itself, yet its index is not zero: "unknowns minus equations" is $\infty - \infty$, and the shift makes that difference -1.

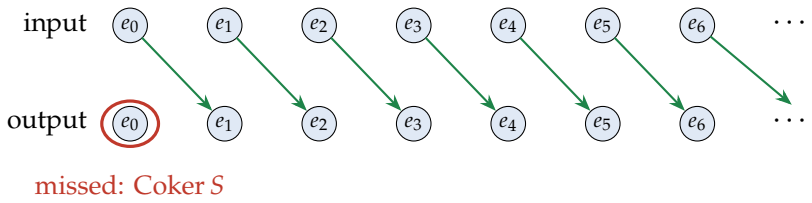


Figure 4.4: The shift S moves every basis vector one step to the right, like guests in Hilbert's hotel each moving to the next room. Nothing is killed ($\operatorname{Ker} S = 0$), but room e_0 is left empty ($\operatorname{Coker} S = \mathbb{C} e_0$), so $\operatorname{index} S = -1$.

4.1.5 Compact perturbations do not change the index

Nash says (p. 89) that the index is "stable under perturbations of the operator by a compact operator". Let us say what that means and why it is true.

Working definition 4.4 (compact operator)

An operator K is **compact** if it is a norm limit [$\|K - K_n\| \to 0$, i.e. the largest stretching $\max_{\|v\|=1} \|(K - K_n)v\|$ goes to 0] of finite-rank operators. *Operational test:* in a suitable orthonormal basis, K is (close to) a matrix whose entries die off, such as $\operatorname{diag}(1, \frac{1}{2}, \frac{1}{3}, \dots)$. Compact operators squash all but finitely many directions to almost nothing.

Two facts, both plausible from the finite-rank case, do the job:

1. If T is Fredholm and K is compact, $T + K$ is Fredholm. (A compact perturbation spoils invertibility-up-to-compacts by only a compact amount.)
2. The index is locally constant on Fredholm operators. This is exactly (4.4).

Now join T to $T + K$ by the straight path $T + sK$, $s \in [0, 1]$. Every operator on the path is Fredholm by fact 1. The index is an integer that is locally constant along the path by fact 2, so it is constant:

$$\operatorname{index}(T + K) = \operatorname{index} T \qquad (K \text{ compact}). \tag{4.5}$$

The same argument shows that the index is constant along *any* continuous path of Fredholm operators. That is the precise sense in which the index is a topological invariant: it labels the connected components of the space of Fredholm operators.

Worked example 4.4: A rank-one perturbation of the shift

Let K be the rank-one operator $Ke_0 = -e_1$, $Ke_n = 0$ for $n \geq 1$. Then $(S+K)e_0 = e_1 - e_1 = 0$ and $(S+K)e_n = e_{n+1}$ for $n \geq 1$. So

$$\begin{aligned} \operatorname{index}(S+K) &\overset{(1)}{=} \dim \operatorname{span}\{e_0\} - \dim \operatorname{span}\{e_0, e_1\} \\ &= 1 - 2 = -1 = \operatorname{index} S. \end{aligned}$$

(1) e_0 is now killed, and the image is spanned by $e_2, e_3, \ldots$, which misses both e_0 and e_1.

Both the kernel and the cokernel grew by one; the index did not move.

Physical picture: Zero modes come and go in pairs

Think of $\operatorname{Ker} T$ as the zero modes of a "left-handed" sector and $\operatorname{Ker} T^*$ as those of a "right-handed" sector. A perturbation can give a mass to a left-handed zero mode only by pairing it with a right-handed one, just as a Dirac mass term $m\bar{\psi}_L\psi_R$ needs both chiralities. The unpaired zero modes are protected. The index counts them. We will see in Section 4.8 that this is literally the supersymmetric pairing of nonzero modes.

Reading Nash (p. 89, (4.1))

Nash opens the chapter with $\operatorname{index} O = \dim \operatorname{Ker} O - \dim \operatorname{Coker} O$ and the remark that the index "is also stable under perturbations of the operator O by a compact operator". The Fredholm-operator picture sits at the end of his Chapter 3 (§6, p. 88), where he notes that homotopy classes of maps into the space of Fredholm operators give $K(X)$, and that for X a point this is just the index $\dim \operatorname{Ker} F - \dim \operatorname{Coker} F$. Everything in this section is the background those two remarks assume.

Recap

- $\operatorname{index} A = \dim \operatorname{Ker} A - \dim \operatorname{Coker} A$, and $\operatorname{Coker} A \cong \operatorname{Ker} A^*$.
- Finite dimensions: $\operatorname{index} A = \dim V - \dim W$ for every A.
- Under perturbation, zero modes and obstructions interact only through a finite Schur-complement matrix, so they appear and disappear in pairs and the index is locally constant.
- Infinite dimensions allow index $\neq 0$ for maps of a space to itself (shift: -1). The index is unchanged by compact perturbations and along continuous paths of Fredholm operators.

4.2 Elliptic operators have an index

A differential operator acts on functions, and function spaces are infinite-dimensional. So why should $D\psi = 0$ have only finitely many independent solutions? For most operators it does not: the wave equation on a torus has infinitely many. The operators with finite-dimensional kernel

and cokernel are the *elliptic* ones, and for them Nash's definition (4.1),

$$\operatorname{index} D = \dim \operatorname{Ker} D - \dim \operatorname{Coker} D,$$

makes sense. In this section we recall what ellipticity is, see with Fourier modes why it forces finiteness, and compute the index of the first honest examples: $\mathrm{d}/\mathrm{d}x + a$ on a circle, $\bar\partial$ on a torus, and the Toeplitz operator, which is our first operator with nonzero index.

4.2.1 Reminder: the leading symbol

Let E and F be vector bundles of ranks q and p over a manifold X of dimension n, and let $\Gamma(X,E)$ denote the smooth sections of E ("vector-valued fields"). A linear differential operator of order m, $D\colon \Gamma(X,E) \to \Gamma(X,F)$, looks in local coordinates x^μ and local trivialisations like

$$(Du)(x) = \sum_{|\alpha| \le m} a_\alpha(x)\, \partial^\alpha u(x), \qquad \partial^\alpha = \partial_1^{\alpha_1} \cdots \partial_n^{\alpha_n}, \tag{4.6}$$

where each coefficient $a_\alpha(x)$ is a $p \times q$ matrix (Nash (2.39)).

Working definition 4.5 (leading symbol)

To get the **leading symbol** of D, keep only the top-order terms $|\alpha| = m$ and replace each derivative ∂_μ by $\mathrm{i}k_\mu$:

$$\sigma_m(D)(x,k) = \sum_{|\alpha|=m} a_\alpha(x)\, (\mathrm{i}k)^\alpha.$$

For each point x and covector $k = k_\mu \mathrm{d}x^\mu \in T^*_x X$ [a momentum at x, with components k_μ] this is a $p \times q$ matrix, i.e. a linear map $E_x \to F_x$ [from the field components at x to the output components at x]. So $\sigma_m(D)$ is a matrix-valued function on phase space T^*X [the pairs (x,k) of a point and a momentum at that point], homogeneous of degree m [$\sigma_m(D)(x,\lambda k) = \lambda^m \sigma_m(D)(x,k)$] in the momentum k.

The rule $\partial_\mu \to \mathrm{i}k_\mu$ is just what a plane wave does: $\partial_\mu e^{\mathrm{i}k\cdot x} = \mathrm{i}k_\mu e^{\mathrm{i}k\cdot x}$. The leading symbol records how D acts on very rapidly oscillating waves, where the highest derivatives dominate. It is a matrix, not a number, because D mixes the components of the field.

Working definition 4.6 (elliptic operator)

D is **elliptic** if $\sigma_m(D)(x,k)\colon E_x \to F_x$ is an invertible matrix for every x and every $k \neq 0$ [every nonzero momentum. So at leading order D sends no rapidly oscillating wave $e^{\mathrm{i}k\cdot x}v$, $v \neq 0$, to zero]. (So in particular $p = q$.) *Operational test:* compute $\det \sigma_m(D)(x,k)$ and check that it vanishes only at $k = 0$.

Globally, Nash (p. 38) packages this as follows. Let $\pi\colon T^*X \to X$ be the projection. Pull E and F back to bundles π^*E, π^*F over T^*X (same fibres, now attached to each point (x,k)). Then $\sigma_m(D)$ is a bundle map $\pi^*E \to \pi^*F$, and ellipticity says it is an isomorphism *off the zero section* $\{k = 0\}$. This geometric phrasing is what makes K-theory enter later.

Worked example 4.5: A table of symbols

operator	on	$\sigma_m(x,k)$	elliptic?
$\mathrm{d}/\mathrm{d}x$	S^1	$\mathrm{i}k$	yes
$-\partial_x^2-\partial_y^2$	$\mathbb{R}^2$	$k_x^2+k_y^2$	yes
$\bar\partial=\frac12(\partial_x+\mathrm{i}\partial_y)$	$\mathbb{C}$	$\frac{\mathrm{i}}{2}(k_x+\mathrm{i}k_y)$	yes
$\sigma_1\partial_x+\sigma_2\partial_y$	$\mathbb{R}^2$, 2-spinors	$\mathrm{i}(\sigma_1k_x+\sigma_2k_y)$	yes
$\partial_t^2-\partial_x^2$	$\mathbb{R}^2$	$-k_t^2+k_x^2$	no (light cone)
$\partial_t-\partial_x^2$	$\mathbb{R}^2$	k_x^2 (order 2)	no ($k_x=0$)

For the two-dimensional Dirac-type operator we use the Pauli matrices $\sigma_1=\left(\begin{smallmatrix}0&1\\1&0\end{smallmatrix}\right)$, $\sigma_2=\left(\begin{smallmatrix}0&-\mathrm{i}\\\mathrm{i}&0\end{smallmatrix}\right)$. Its symbol is invertible for $k\neq0$ because

$$\begin{aligned}\big(\mathrm{i}(\sigma_1k_x+\sigma_2k_y)\big)^2 &\overset{(1)}{=} -\big(\sigma_1^2k_x^2+\sigma_2^2k_y^2+(\sigma_1\sigma_2+\sigma_2\sigma_1)k_xk_y\big)\\ &\overset{(2)}{=} -(k_x^2+k_y^2)\,\mathbb{1}.\end{aligned}$$

(1) Expand the square, keeping the order of the matrices.
(2) Pauli algebra: $\sigma_1^2=\sigma_2^2=\mathbb{1}$ and $\sigma_1\sigma_2+\sigma_2\sigma_1=0$.

A matrix whose square is $-|k|^2\mathbb{1}$ has inverse $-\sigma/|k|^2$. This "square root of the Laplacian" property is exactly Dirac's trick, and it is why Dirac operators are elliptic. The wave operator fails on the light cone $k_t=\pm k_x$, the heat operator fails on the line $k_x=0$ (its leading, second-order part does not see k_t).

Watch out: Nash drops the i

Nash writes the symbol as $\sum a_\alpha(x)k^\alpha$, without the factors of i (his (2.40)–(2.41)). For deciding ellipticity this makes no difference: it multiplies σ_m by the nonzero number i^m. We keep the i because it makes symbols of adjoints and products come out with the right signs, e.g. $\sigma(\mathrm{d}/\mathrm{d}x)=\mathrm{i}k$ and $\sigma(-\mathrm{d}^2/\mathrm{d}x^2)=k^2>0$.

Physical picture: Ellipticity is a dispersion relation with no massless directions

A physicist reads $\sigma_m(D)(x,k)$ as a dispersion relation: it is the matrix that a plane wave of momentum k sees. Ellipticity says this matrix is invertible for every direction of $k\neq0$, i.e. there is no direction in momentum space where the leading term vanishes and a mode can propagate for free at arbitrarily high momentum. That is exactly why the wave operator ($\sigma=k_t^2-k_x^2$, vanishing on the light cone) and the heat operator ($\sigma=k_x^2$, blind to k_t) are *not* elliptic: each has a whole ray of high-momentum directions with no restoring term, which is why they support propagating waves or an unconstrained time direction. An elliptic operator, by contrast, behaves like a field with a mass gap in every direction of T^*X: only finitely many low-momentum modes can be soft, which is the Fourier-mode argument of the next subsection.

4.2.2 Why ellipticity makes the kernel finite

The cleanest way to see finiteness is on a torus with constant coefficients, where Fourier modes diagonalise everything. Let $X=T^n=\mathbb{R}^n/(2\pi\mathbb{Z})^n$ and let $D=\sum_{|\alpha|\le m}a_\alpha\partial^\alpha$ have constant matrix

coefficients. On a plane wave with integer wave-vector $N \in \mathbb{Z}^n$,

$$D\big(v\, e^{\mathrm{i}N\cdot x}\big) \overset{(1)}{=} \sum_{|\alpha|\le m} a_\alpha (\mathrm{i}N)^\alpha v\, e^{\mathrm{i}N\cdot x} \overset{(2)}{=} \sigma_{\mathrm{full}}(N)\, v\, e^{\mathrm{i}N\cdot x}, \tag{4.7}$$

(1) Each ∂_μ brings down $\mathrm{i}N_\mu$; v is a constant vector in the fibre.
(2) Define the **full symbol** $\sigma_{\mathrm{full}}(N) = \sum_{|\alpha|\le m} a_\alpha (\mathrm{i}N)^\alpha$, whose top-degree part is $\sigma_m(N)$.

So D acts on the N-th Fourier mode as the finite matrix $\sigma_{\mathrm{full}}(N)$. Hence

$$\operatorname{Ker} D = \bigoplus_N \operatorname{Ker} \sigma_{\mathrm{full}}(N)\, e^{\mathrm{i}N\cdot x}, \qquad \operatorname{Coker} D \cong \bigoplus_N \operatorname{Coker} \sigma_{\mathrm{full}}(N)\, e^{\mathrm{i}N\cdot x}.$$

Now use ellipticity. By homogeneity and invertibility on the unit sphere, $\|\sigma_m(N)v\| \ge c\,|N|^m\,\|v\|$ for some $c > 0$. The lower-order terms grow at most like $|N|^{m-1}$. So for $|N|$ large enough, the top term wins and $\sigma_{\mathrm{full}}(N)$ is invertible. Only the finitely many lattice points in some ball $|N| \le R$ can contribute to kernel or cokernel. Each contributes a finite-dimensional space. Therefore $\operatorname{Ker} D$ and $\operatorname{Coker} D$ are finite-dimensional.

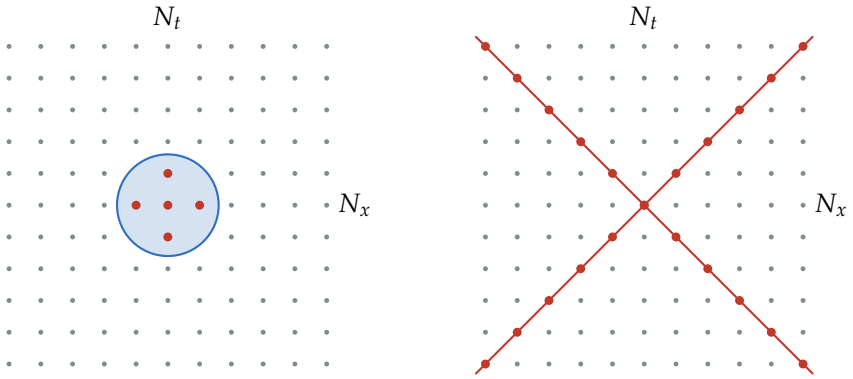
Worked example 4.6: The wave operator has an infinite kernel

On T^2 with coordinates (t, x), take $\Box = \partial_t^2 - \partial_x^2$. By (4.7), on $e^{\mathrm{i}(N_t t + N_x x)}$,

$$\Box\, e^{\mathrm{i}(N_t t+N_x x)} \overset{(1)}{=} \big((\mathrm{i}N_t)^2 - (\mathrm{i}N_x)^2\big) e^{\mathrm{i}(N_t t+N_x x)} = (N_x^2 - N_t^2)\, e^{\mathrm{i}(N_t t+N_x x)}.$$

(1) Each derivative brings down i times the wave number.

Every mode with $N_t = \pm N_x$ is killed: $\operatorname{Ker}\Box$ contains $e^{\mathrm{i}N(t\pm x)}$ for every $N \in \mathbb{Z}$, an infinite-dimensional space (left- and right-moving waves). The symbol's zeros on the light cone are exactly the zero modes that escape to infinity. See Figure 4.5.



elliptic: zero modes only near $N = 0$ wave operator: whole light cone

Figure 4.5: Fourier lattice $\mathbb{Z}^2$ of a two-torus. Red dots are modes that may lie in the kernel. Left: for an elliptic operator the symbol grows in every direction, so only finitely many modes inside a disc (blue) can be zero modes. Right: the wave operator's symbol vanishes along the light cone $N_t = \pm N_x$, and infinitely many modes are killed.

For variable coefficients on a general compact manifold, the Fourier argument is replaced by a **parametrix**: an approximate inverse Q built by inverting the symbol, with $QD = 1 + K_1$,

$DQ = 1 + K_2$ and K_1, K_2 compact (they are "smoothing", of order -1 or lower). By the working definition in Section 4.1.4, D is then Fredholm between the appropriate Sobolev spaces (Nash Chapter 2, §5). The moral is the same as on the torus: the symbol is invertible at high momentum, so all the trouble lives in a finite-dimensional low-momentum corner.

4.2.3 The index, and why the adjoint appears

Working definition 4.7 (index of an elliptic operator)

For an elliptic $D\colon \Gamma(X,E) \to \Gamma(X,F)$ on a compact manifold X without boundary,

$$\operatorname{index} D = \dim \operatorname{Ker} D - \dim \operatorname{Coker} D = \dim \operatorname{Ker} D - \dim \operatorname{Ker} D^*.$$

[D^* is the formal adjoint, $\int \langle D^*\chi, \psi\rangle = \int \langle \chi, D\psi\rangle$. Each independent solution of $D^*\chi = 0$ is one independent obstruction to solving $D\psi = f$] *Operational recipe:* solve $D\psi = 0$ and count independent smooth solutions; solve $D^*\chi = 0$ and count; subtract.

Here D^* is the **formal adjoint**, defined by integration by parts. Choose inner products on the fibres of E, F and a volume form on X, so that $\langle u, v\rangle = \int_X (u, v)\,\mathrm{vol}$. Then D^* is the operator with $\langle \chi, D\psi\rangle = \langle D^*\chi, \psi\rangle$ for all smooth ψ, χ. For example,

$$\begin{aligned} \left\langle \chi, \frac{\mathrm{d}\psi}{\mathrm{d}x} \right\rangle &\overset{(1)}{=} \int_0^{2\pi} \bar{\chi}\,\psi'\,\mathrm{d}x \\ &\overset{(2)}{=} \big[\bar{\chi}\psi\big]_0^{2\pi} - \int_0^{2\pi} \bar{\chi}'\,\psi\,\mathrm{d}x \\ &\overset{(3)}{=} \left\langle -\frac{\mathrm{d}\chi}{\mathrm{d}x}, \psi \right\rangle. \end{aligned} \tag{4.8}$$

(1) The L^2 inner product on the circle, antilinear in the first slot.
(2) Integration by parts.
(3) The circle has no boundary: χ and ψ are periodic, so the boundary term cancels.

So $(\mathrm{d}/\mathrm{d}x)^* = -\mathrm{d}/\mathrm{d}x$. Step (3) is where "$X$ has no boundary" (Nash p. 89) is used. On an interval, the boundary term survives, and one must impose boundary conditions; that is the business of Nash §7 (the Atiyah–Patodi–Singer theorem).

Why is $\operatorname{Coker} D \cong \operatorname{Ker} D^*$? In finite dimensions this was (4.1). Here the same computation shows that the L^2-orthogonal complement of $\operatorname{Im} D$ consists of (weak) solutions of $D^*\chi = 0$. D^* is elliptic too (its symbol is the adjoint matrix $\sigma_m(D)^\dagger$, Nash (2.47)), and solutions of elliptic equations are smooth (Nash Chapter 2, elliptic regularity). So those weak solutions are honest smooth zero modes of D^*.

4.2.4 First example: $\mathrm{d}/\mathrm{d}x + a$ on the circle

Worked example 4.7: $D_a = \mathrm{d}/\mathrm{d}x + a$ with constant $a \in \mathbb{C}$

Take $X = S^1 = \mathbb{R}/2\pi\mathbb{Z}$ and $D_a = \mathrm{d}/\mathrm{d}x + a\colon C^\infty(S^1) \to C^\infty(S^1)$. Its leading symbol is $\mathrm{i}k$, invertible for $k \neq 0$: elliptic. On Fourier modes $e_N = e^{\mathrm{i}Nx}$,

$$D_a e_N \overset{(1)}{=} (\mathrm{i}N + a)\,e_N, \qquad D_a^* e_N \overset{(2)}{=} (-\mathrm{i}N + \bar{a})\,e_N.$$

(1) $\frac{\mathrm{d}}{\mathrm{d}x} e^{\mathrm{i}Nx} = \mathrm{i}N e^{\mathrm{i}Nx}$.
(2) $D_a^* = -\mathrm{d}/\mathrm{d}x + \bar{a}$ by (4.8), since multiplication by a has adjoint multiplication by $\bar{a}$.

So D_a is the infinite diagonal matrix $\operatorname{diag}(\dots, a-2\mathrm{i}, a-\mathrm{i}, a, a+\mathrm{i}, a+2\mathrm{i}, \dots)$. A mode e_N is in the kernel iff $a = -\mathrm{i}N$, and in the kernel of D_a^* iff $\bar{a} = \mathrm{i}N$, which is the same condition. Hence

$$\dim\operatorname{Ker} D_a = \dim\operatorname{Ker} D_a^* = \begin{cases} 1 & a \in \mathrm{i}\mathbb{Z}, \\ 0 & \text{otherwise}, \end{cases} \qquad \text{and} \qquad \operatorname{index} D_a = 0 \text{ for every } a.$$

In position space: $u' + au = 0$ gives $u = Ce^{-ax}$, which is 2π-periodic iff $e^{-2\pi a} = 1$, i.e. $a \in \mathrm{i}\mathbb{Z}$. (Physically: a charged particle on a ring; $a = \mathrm{i}\theta$ is a constant gauge field, and a zero mode exists exactly when the flux θ is an integer.)

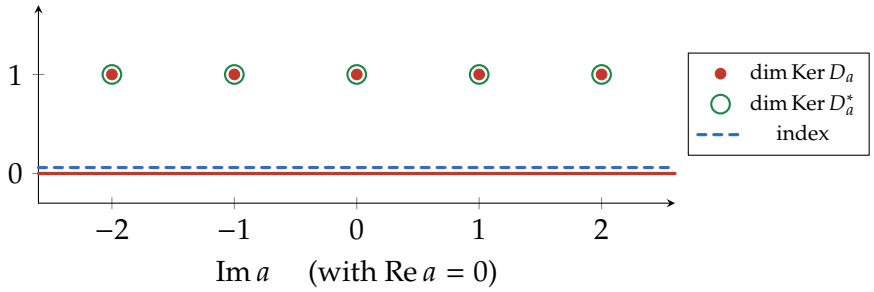


Figure 4.6: $D_a = \mathrm{d}/\mathrm{d}x + a$ on the circle. The kernel (red dots) and the cokernel (green rings) are both 0 except at the integer points $a \in \mathrm{i}\mathbb{Z}$, where both jump to 1 at the same time. The index (blue dashed line) is identically 0.

Worked example 4.8: $D = \mathrm{d}/\mathrm{d}x + a(x)$ with a variable coefficient

Now let $a(x)$ be any smooth periodic complex function, and let $\bar{a}_0 = \frac{1}{2\pi}\int_0^{2\pi} a\,\mathrm{d}x$ be its average. Solve $u' + au = 0$:

$$u(x) \overset{(1)}{=} C\exp\Big(-\int_0^x a(s)\,\mathrm{d}s\Big),$$
$$u(2\pi) \overset{(2)}{=} u(0)\,e^{-2\pi\bar{a}_0}.$$

(1) Separation of variables: $u'/u = -a$.
(2) Put $x = 2\pi$ and use the definition of the average.

So a periodic solution exists iff $e^{-2\pi\bar{a}_0} = 1$, i.e. $\bar{a}_0 \in \mathrm{i}\mathbb{Z}$. The adjoint $D^* = -\mathrm{d}/\mathrm{d}x + \bar{a}(x)$ has solutions $\chi = C\exp(\int_0^x \bar{a})$, periodic iff $e^{2\pi\overline{\bar{a}_0}} = 1$: the same condition. Again $\dim\operatorname{Ker} D = \dim\operatorname{Ker} D^*$ and $\operatorname{index} D = 0$. Notice that only the *average* of a (the holonomy of the gauge field around the ring) decides whether zero modes exist, but no choice of a changes the index.

4.2.5 Lower-order terms never change the index

The last two examples illustrate a general principle that Nash states on p. 93: "two elliptic D with the same leading symbol have the same index". Here is why. If D_0 and D_1 are elliptic of order m with the same leading symbol, then $D_1 - D_0$ has order at most $m-1$. As a map from the Sobolev space H^m (fields with m square-integrable derivatives) to L^2, an operator of order $m-1$ is compact: it only uses $m-1$ derivatives, and the "spare" derivative makes high-momentum modes negligible (this is Rellich's lemma; on the torus it is visible in the

Fourier picture, where the ratio $|N|^{m-1}/|N|^m \to 0$ plays the role of the decaying diagonal entries $1, \frac{1}{2}, \frac{1}{3}, \dots$). Therefore

$$\begin{aligned} \operatorname{index} D_1 &\overset{(1)}{=} \operatorname{index}\big(D_0 + (D_1 - D_0)\big) \\ &\overset{(2)}{=} \operatorname{index} D_0. \end{aligned} \tag{4.9}$$

(1) Trivially rewrite D_1.
(2) $D_1 - D_0 \colon H^m \to L^2$ is compact, and compact perturbations do not change the index, (4.5).

More generally, if the leading symbol is deformed continuously *through elliptic symbols*, the operators form a continuous path of Fredholm operators and the index is constant. So:

Key idea: The index lives in the symbol

The index of an elliptic operator depends only on the leading symbol $\sigma_m(D)$, and only on its deformation class among symbols that stay invertible off the zero section. This is the fact that lets topology in: deformation classes of invertible matrix-valued functions are exactly what K-theory counts.

In Example 4.8 this is visible directly: every $\mathrm{d}/\mathrm{d}x + a(x)$ has symbol $\mathrm{i}k$, the same as $\mathrm{d}/\mathrm{d}x$, whose kernel and cokernel are the constants. So every such operator has index $1 - 1 = 0$.

Worked example 4.9: $\bar\partial$ on a torus

Let $X = T^2 = \mathbb{R}^2/(2\pi\mathbb{Z})^2$ with complex coordinate $z = x + \mathrm{i}y$, and $\bar\partial = \frac{1}{2}(\partial_x + \mathrm{i}\partial_y)$ acting on functions. On $e_{MN} = e^{\mathrm{i}(Mx+Ny)}$,

$$\begin{aligned} \bar\partial e_{MN} &\overset{(1)}{=} \tfrac{1}{2}\big(\mathrm{i}M + \mathrm{i}\cdot\mathrm{i}N\big)e_{MN} = \tfrac{\mathrm{i}}{2}(M + \mathrm{i}N)\, e_{MN}, \\ \bar\partial^* e_{MN} &\overset{(2)}{=} -\tfrac{1}{2}(\partial_x - \mathrm{i}\partial_y)e_{MN} = -\tfrac{\mathrm{i}}{2}(M - \mathrm{i}N)\, e_{MN}. \end{aligned}$$

(1) Differentiate the plane wave.
(2) Integrate by parts as in (4.8): each ∂_μ changes sign and i is conjugated.

$M + \mathrm{i}N = 0$ only for $M = N = 0$. So $\operatorname{Ker}\bar\partial$ = constants (the holomorphic functions on a compact surface are constant) and $\operatorname{Ker}\bar\partial^*$ = constants. Index $= 1 - 1 = 0$. We will see in the next part that the index of $\bar\partial$ on a surface of genus g is $1 - g$, and the torus has $g = 1$.

4.2.6 An operator with nonzero index: the Toeplitz operator

Every example so far has index 0. That is not an accident: we will see that on odd-dimensional manifolds, and for self-adjoint operators, the index of a differential operator always vanishes (for $D = D^*$ it is obvious: $\dim\operatorname{Ker} D - \dim\operatorname{Ker} D^* = 0$). To see a nonzero index on the circle we must leave differential operators and allow **pseudo-differential** ones, whose symbols need not be polynomials in k. This is why Nash insists throughout on "differential or pseudo-differential" operators.

Split $L^2(S^1)$ into positive and negative frequencies. Let P be the projection onto the **Hardy space** $H^2 = \overline{\operatorname{span}}\{e_N : N \geq 0\}$:

$$P\Big(\sum_N c_N e_N\Big) = \sum_{N\geq 0} c_N e_N.$$

P acts on $e^{\mathrm{i}Nx}$ as the step function $\theta(N)$ (with $\theta(0) = 1$). It is a pseudo-differential operator of order 0 with symbol $\theta(k)$: not a polynomial, so not a differential operator.

Working definition 4.8 (Toeplitz operator)

For a continuous function $f\colon S^1 \to \mathbb{C}\setminus\{0\}$ [f never vanishes, so it has a winding number around 0], the **Toeplitz operator** is $T_f = PM_fP\colon H^2 \to H^2$, where M_f is multiplication by f. In the basis $e_0, e_1, \dots$ its matrix is $(T_f)_{MN} = \hat{f}_{M-N}$, constant along diagonals, where $\hat{f}_j$ are the Fourier coefficients of f [$f(x) = \sum_j \hat{f}_j e^{ijx}$].

Worked example 4.10: T_{z^k} is the k-fold shift

Let $z = e^{ix}$ and $f = z^k$ with $k \geq 0$. Then $M_f e_N = e_{N+k}$, which stays in H^2 when $N \geq 0$, so $T_{z^k} e_N = e_{N+k}$: this is S^k, the shift of Example 4.3 applied k times. Hence

$$\operatorname{index} T_{z^k} = -k.$$

For $k < 0$, $T_{z^k} e_N = e_{N+k}$ if $N + k \geq 0$ and 0 otherwise: this is $(S^*)^{|k|}$, with index $+|k| = -k$. In all cases, index $T_{z^k} = -k = -$(winding number of z^k around 0).

Now take any nonvanishing f with winding number w around 0. The loop $f(S^1)$ can be deformed to z^w through nonvanishing loops (that is what equal winding number means). Along the deformation f_s, T_{f_s} is Fredholm, because

$$T_f T_{1/f} - 1 \;=\; PM_f PM_{1/f}P - P \;=\; -PM_f(1-P)M_{1/f}P \tag{4.10}$$

is compact (the commutator of P with multiplication by a smooth function only mixes finitely many modes near $N = 0$, up to rapidly decaying tails), so $T_{1/f}$ is an inverse up to compacts. By the stability of the index along Fredholm paths,

$$\begin{aligned} \operatorname{index} T_f &\overset{(1)}{=} \operatorname{index} T_{z^w} \\ &\overset{(2)}{=} -w = -\frac{1}{2\pi i}\oint_{S^1} \frac{df}{f}. \end{aligned} \tag{4.11}$$

(1) f is homotopic to z^w through nonvanishing functions, and the index is constant along the corresponding path of Fredholm operators.

(2) Example 4.10's computation, and the standard argument-principle formula for the winding number.

Key idea: The first index theorem

The **Noether–Gohberg–Krein** index theorem, $\operatorname{index} T_f = -\operatorname{wind}(f)$, is the Atiyah–Singer theorem on the circle. The left side is analytic (count solutions). The right side is topological (a winding number, an integral of a closed form). Every feature of the general theorem is already here: the index depends only on the symbol f, only through its homotopy class, and the answer is an integral of a characteristic class ($\frac{1}{2\pi i} d\log f$ is the simplest Chern character).

To regard T_f as an operator on all of $L^2(S^1)$ (so that its symbol lives on T^*S^1), extend it by the identity on negative frequencies: $A_f = PM_fP + (1-P)$. It has the same kernel and cokernel as T_f. Its principal symbol is

$$\sigma(A_f)(x,k) = \begin{cases} f(x) & k > 0, \\ 1 & k < 0. \end{cases} \tag{4.12}$$

T^*S^1 minus its zero section is two half-cylinders, $k > 0$ and $k < 0$. The symbol is a nonvanishing function on each. Its winding number on the $k > 0$ half is w, on the $k < 0$ half it is 0. Compare

a differential operator $\mathrm{d}/\mathrm{d}x + a$: its symbol ik winds 0 times on both halves. The index in both cases is minus the difference (winding on $k > 0$) − (winding on $k < 0$): $-w$ for Toeplitz, 0 for $\mathrm{d}/\mathrm{d}x + a$. This is the circle version of the topological index we will construct in Section 4.6.

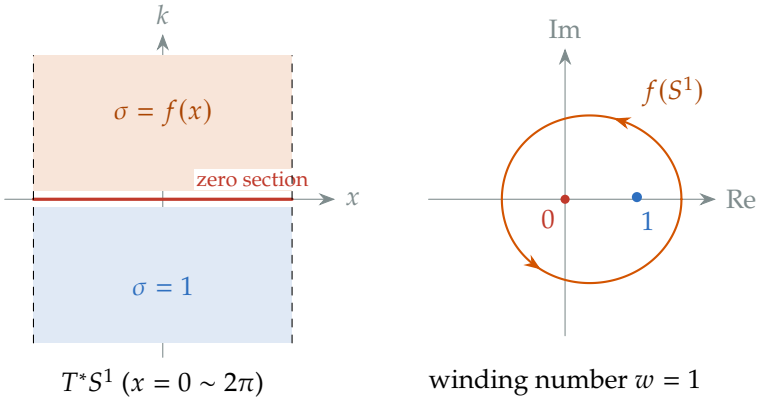


Figure 4.7: Left: the symbol of the Toeplitz-type operator A_f on the two halves of T^*S^1 minus the zero section. Right: the loop traced by f on the upper half winds once around 0, while the constant 1 on the lower half does not wind. The mismatch is what gives index $A_f = -1$.

Reading Nash (p. 89, (4.1))

Nash: "We have already discussed elliptic operators in chapter 2 where we saw that such operators have finite dimensional kernels and cokernels. This being so one can define the index ... index O = dim Ker O − dim Coker O ... Let this operator O act on a topological space X ... For the moment we require X to have no boundary." He also notes that the index theorem applies to "differential or pseudo-differential" operators. The Toeplitz example shows why the pseudo-differential ones are needed: on the circle they are the only source of nonzero index. And his remark on p. 93 that operators with the same leading symbol have the same index is (4.9).

Recap

- Leading symbol: keep top derivatives, $\partial_\mu \to ik_\mu$; a matrix-valued function on T^*X. Elliptic = invertible for $k \neq 0$.
- Ellipticity ⇒ only low-momentum modes can be zero modes ⇒ finite kernel and cokernel; the wave operator fails on its light cone.
- index D = dim Ker D − dim Ker D^* (X closed, so no boundary terms in the adjoint).
- $\mathrm{d}/\mathrm{d}x + a(x)$ on S^1, $\bar\partial$ on T^2: kernels jump, index 0. Lower-order terms never change the index.
- Toeplitz: index $T_f = -\mathrm{wind}(f)$, the first index theorem.

4.3 K-theory from complexes of bundles

We have just seen that the index of an elliptic operator depends only on its leading symbol $\sigma_m(D)$, and only on the deformation class of that symbol. The symbol is a bundle map $\pi^*E \to \pi^*F$ over phase space T^*X which is invertible everywhere except on the zero section. So the question is: *what group classifies such things?* Ordinary K-theory, as we met it in Chapter 3, classifies bundles,

not maps between bundles. Nash (p. 90) therefore switches to a second description of K-theory, due to Segal, in which the basic objects are *complexes* of bundles. This section explains that description from scratch, because it is the language of everything that follows.

4.3.1 Reminder: $K(X)$ as formal differences of bundles

Recall the construction of Chapter 3. Isomorphism classes of complex vector bundles over X form a set $\mathrm{Vect}(X)$, and direct sum makes it an abelian **semigroup**: we can add, but not subtract. We complete it to a group exactly as we build the integers from the natural numbers.

Working definition 4.9 (the group $K(X)$)

An element of $K(X)$ is a formal difference $[E]-[F]$ [a pair of bundles (E,F), read as "E minus F"] of two bundles, with the rule

$$[E]-[F]=[E']-[F'] \iff E\oplus F'\oplus G \cong E'\oplus F\oplus G \text{ for some bundle } G.$$

[G is needed because $E\oplus G\cong E'\oplus G$ need not imply $E\cong E'$] *Operational meaning:* think of $[E]-[F]$ as a "virtual bundle" of rank $\operatorname{rank}E-\operatorname{rank}F$, which may be negative.

Worked example 4.11: From $\mathbb{N}$ to $\mathbb{Z}$, and $K(\text{point})=\mathbb{Z}$

A vector bundle over a single point is just a vector space, and a complex vector space is determined up to isomorphism by its dimension. So $\mathrm{Vect}(\mathrm{pt})=\{0,1,2,\dots\}=\mathbb{N}$ with addition. Pairs (a,b) stand for $a-b$, and

$$(a,b)\sim(a',b') \overset{(1)}{\Leftrightarrow} a+b'+c=a'+b+c \text{ for some } c$$
$$\overset{(2)}{\Leftrightarrow} a-b=a'-b'.$$

(1) The equivalence relation of the working definition, with dimensions in place of bundles.

(2) Cancel c; this is allowed in $\mathbb{N}$.

So $K(\mathrm{pt})\cong\mathbb{Z}$, the isomorphism being "virtual dimension": $[\mathbb{C}^a]-[\mathbb{C}^b]\mapsto a-b$. Keep this in mind: it is the first hint that K-theory and the index are made for each other, since the index of a linear map $\mathbb{C}^a\to\mathbb{C}^b$ is also $a-b$.

Nash notes (p. 89–90) that there are real and quaternionic versions (KO, KSp) and that "$K(X)$" in this chapter always means the complex one, $KU(X)$.

4.3.2 Complexes of bundles

Working definition 4.10 (complex of bundles, Nash (4.2))

A **complex** E over X is a finite sequence of vector bundles and bundle maps ("homomorphisms", i.e. fibrewise linear maps depending continuously on the base point)

$$0\longrightarrow E^0\xrightarrow{\alpha_0}E^1\xrightarrow{\alpha_1}\cdots\xrightarrow{\alpha_n}E^{n+1}\longrightarrow 0, \qquad \alpha_i\circ\alpha_{i-1}=0. \tag{4.13}$$

Concretely: over each point x we have a string of matrices $\alpha_i(x)$, and the product of any two consecutive ones is the zero matrix: "the image of one map lies inside the kernel of the next".

The condition $\alpha_i\alpha_{i-1} = 0$ is the same as $\mathrm{d}\circ\mathrm{d} = 0$ for the exterior derivative. At each point x it lets us form the *cohomology* of the fibre sequence, $\operatorname{Ker}\alpha_i(x)/\operatorname{Im}\alpha_{i-1}(x)$. When all of these vanish the sequence is called **exact** at x.

Working definition 4.11 (support, Nash (4.3))

The **support** of a complex E is the set of points $x \in X$ at which the sequence of vector spaces

$$0 \to E^0_x \xrightarrow{\alpha_0(x)} E^1_x \xrightarrow{\alpha_1(x)} \cdots \xrightarrow{\alpha_n(x)} E^{n+1}_x \to 0$$

fails to be exact. *Operational test* for a two-term complex $0 \to E^0 \xrightarrow{\alpha} E^1 \to 0$: exactness means $\alpha(x)$ is an isomorphism, so the support is where $\det\alpha(x) = 0$ (or where the ranks differ [$\operatorname{rank}E^0 \neq \operatorname{rank}E^1$, so no $\alpha(x)$ can be invertible], in which case the support is everything).

Exactness is an open condition (an invertible matrix stays invertible under small changes), so the support is always a closed set.

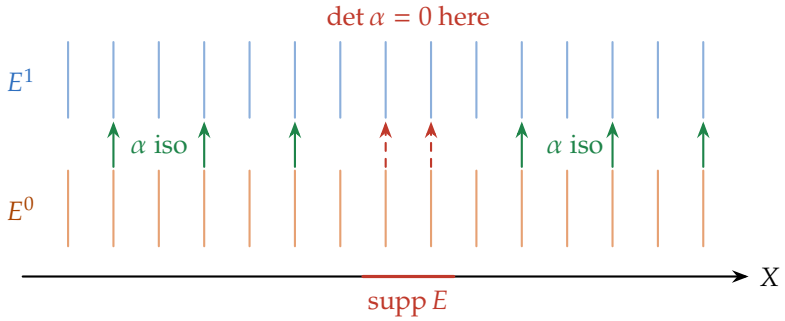


Figure 4.8: A two-term complex $0 \to E^0 \xrightarrow{\alpha} E^1 \to 0$ drawn as two families of fibres over X with a map between them. Where α is an isomorphism (green) the complex is exact and "carries no information"; the red segment where α fails to be invertible is the support. A complex is a virtual bundle $E^0 - E^1$ that has been told it may only live on its support.

Working definition 4.12 (homotopy of complexes)

Two complexes E, F over X are **homotopic** if there is a complex G over $X \times [0,1]$ with $G|_{X\times\{0\}} \cong E$ [G restricted to the bottom slice $X \times \{0\}$ is isomorphic to E, i.e. equal after a pointwise invertible change of basis] and $G|_{X\times\{1\}} \cong F$. *Concretely:* we may continuously deform the maps $\alpha_i(x)$ (and the bundles), keeping $\alpha_i\alpha_{i-1} = 0$ all the way.

Let $C(X)$ denote the set of homotopy classes of complexes over X. Direct sum of complexes (take $E^i \oplus F^i$ and $\alpha_i \oplus \beta_i$, block-diagonal matrices) makes $C(X)$ an abelian semigroup. Inside it sits the sub-semigroup $C_\phi(X)$ of complexes with *empty* support: those that are exact at every point.

Key idea: Segal's description of K-theory, Nash (4.4)

$$K(X) \;\cong\; \frac{C(X)}{C_\phi(X)}. \tag{4.14}$$

In words: a K-theory class is a complex of bundles, up to homotopy, where everywhere-exact complexes count as zero. The class of E is written $[E]$.

Why is this the same group as before? The dictionary is:

- A single bundle V is the complex $0 \to V \to 0$ with V in degree 0.
- The "trivially exact" complex $0 \to V \xrightarrow{\mathrm{id}} V \to 0$ has empty support, so it is zero in the quotient. That is the same as saying that V placed in degree 1 is the negative of V placed in degree 0. So degree 1 plays the role of the minus sign.

The next subsection makes this precise and derives Nash's formula (4.5).

Physical picture: Complexes as brane–antibrane pairs

There is a vivid physical picture for "bundle in even degree minus bundle in odd degree, up to exact sequences". In string theory a stack of D-branes wrapping E^0 and a stack of anti-D-branes wrapping E^1 carry a **tachyon field** $\alpha_0 \colon E^0 \to E^1$ (a bifundamental scalar, since the branes have opposite charge). Where α_0 is invertible the branes and antibranes can pair up and annihilate completely, leaving nothing: that is exactly an everywhere-exact two-term complex, and it is zero in $C_\phi(X)$. Where α_0 degenerates (its support), some brane charge survives. Sen's conjecture (brane-antibrane annihilation via tachyon condensation) is precisely the statement that the surviving, stable D-brane charge is classified by K-theory built from such complexes, which is why physicists call this "K-theory charges of D-branes". The Bott class of Example 4.13 below is, in this language, one unit of brane charge localised where the tachyon vanishes.

4.3.3 The alternating-sum formula (compact X)

Suppose X is compact. Given a complex E with maps α_i, consider the family of complexes E_s with the same bundles and maps $s\,\alpha_i$, $s \in [0,1]$. Since $(s\alpha_i)(s\alpha_{i-1}) = s^2\alpha_i\alpha_{i-1} = 0$, each E_s is a complex, so $E = E_1$ is homotopic to E_0, the complex with all maps zero. A complex with zero maps is a direct sum of one-bundle complexes $V[i]$, meaning "V sitting alone in degree i". Finally,

$$\begin{aligned} [V[i]] + [V[i+1]] &\overset{(1)}{=} \big[\, 0 \to V \xrightarrow{0} V \to 0 \,\big] \\ &\overset{(2)}{=} \big[\, 0 \to V \xrightarrow{\mathrm{id}} V \to 0 \,\big] \\ &\overset{(3)}{=} 0. \end{aligned} \tag{4.15}$$

(1) Direct sum of the two one-bundle complexes, with V in degrees i and $i+1$ and the zero map between them.
(2) The maps $s \cdot \mathrm{id}$, $s \in [0,1]$, give a homotopy (any linear map between consecutive terms of a two-term complex is allowed, since there is no condition $\alpha_i\alpha_{i-1} = 0$ to check).
(3) The identity is invertible at every point, so this complex has empty support and lies in $C_\phi(X)$.

So $[V[i+1]] = -[V[i]]$, and by induction $[V[i]] = (-1)^i[V]$. Putting the pieces together gives Nash's (4.5):

$$[E] \overset{(1)}{=} [E_0] \overset{(2)}{=} \Big[\bigoplus_i E^i[i] \Big] \overset{(3)}{=} \sum_i \big[E^i[i]\big] \overset{(4)}{=} \sum_i (-1)^i [E^i]. \tag{4.16}$$

(1) The homotopy $s\alpha_i$ from E to the zero-map complex E_0.
(2) A complex with zero maps is the direct sum of its terms, each in its own degree.
(3) Addition in $C(X)$ is direct sum.
(4) $[V[i]] = (-1)^i[V]$ from (4.15).

Watch out: Where compactness was used

On a non-compact space the homotopy $s\alpha_i$ is not allowed. As we will see next, for non-compact spaces we only consider complexes whose support is compact, and homotopies whose support is compact. At $s = 0$ the support of E_0 is all of X (zero maps are never exact unless the bundles vanish), which is not compact. So on a non-compact space *the maps carry the information*, and (4.16) would destroy it. This is Nash's footnote on p. 90, and it is the whole point of the complex description: on phase space T^*X, the interesting information is in the symbol map, not in the bundles.

Worked example 4.12: Complexes over a point: the Euler characteristic

Over $X = \mathrm{pt}$ a complex is a string of matrices. Take

$$0 \to \mathbb{C} \xrightarrow{\alpha_0} \mathbb{C}^2 \xrightarrow{\alpha_1} \mathbb{C} \to 0, \qquad \alpha_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \alpha_1 = \begin{pmatrix} 1 & -1 \end{pmatrix}.$$

It is a complex, since $\alpha_1\alpha_0 = 1 - 1 = 0$. It is exact: α_0 is injective, $\operatorname{Ker}\alpha_1 = \operatorname{span}\{(1,1)\} = \operatorname{Im}\alpha_0$, and α_1 is onto. So its support is empty, $[E] = 0$, and indeed

$$\sum_i (-1)^i \dim E^i = 1 - 2 + 1 = 0.$$

Now take the two-term complex $0 \to \mathbb{C}^2 \xrightarrow{A} \mathbb{C}^3 \to 0$ for *any* 3×2 matrix A. By (4.16) and Example 4.11,

$$[E] \overset{(1)}{=} [\mathbb{C}^2] - [\mathbb{C}^3] \overset{(2)}{=} 2 - 3 = -1 \overset{(3)}{=} \operatorname{index} A.$$

(1) Alternating sum formula (a point is compact).
(2) $K(\mathrm{pt}) \cong \mathbb{Z}$ by dimension.
(3) Finite-dimensional index (4.2).

Over a point, *the K-class of a two-term complex is the index of the map*. The whole analytic index of Section 4.5 is a vast generalisation of this remark.

4.3.4 Non-compact spaces and compact supports

The space we care about, T^*X, is never compact: its fibres are vector spaces. Nash's footnotes on pp. 90–91 prescribe what to do.

Working definition 4.13 (K-theory with compact supports)

For non-compact Y, $K(Y)$ (more precisely $K_c(Y)$) is defined by (4.14) [complexes of bundles up to homotopy, with everywhere-exact complexes counted as zero] using only complexes *whose support is compact* [the complex fails to be exact only inside some bounded region], and homotopies over $Y \times [0,1]$ whose support is compact. Equivalently,

$$K_c(Y) \;\cong\; \widetilde{K}(Y^+),$$

the reduced K-theory of the one-point compactification $Y^+ = Y \cup \{\infty\}$ [one extra point ∞ that every sequence running off to infinity approaches, e.g. $(\mathbb{R}^n)^+ = S^n$]. ("Reduced" means: virtual bundles of virtual dimension 0, i.e. those that are trivial near the added point.)

Why the two descriptions agree: a complex with compact support is exact outside some large compact set, so near ∞ it is "nothing", and it defines a virtual bundle on Y^+ which is trivial at ∞. We will see exactly how in the next example. For compact X nothing changes: $X^+ = X \sqcup \{\infty\}$ and $\widetilde{K}(X \sqcup \mathrm{pt}) = K(X)$.

Worked example 4.13: The Bott class: $K_c(\mathbb{C}) = \widetilde{K}(S^2)$

Let $Y = \mathbb{C}$ with coordinate z, and consider the two-term complex of trivial line bundles

$$\beta: \qquad 0 \to \mathbb{C}\times\mathbb{C} \xrightarrow{\cdot z} \mathbb{C}\times\mathbb{C} \to 0, \qquad (z,v) \mapsto (z, zv). \tag{4.17}$$

The map at the point z is the 1×1 matrix (z). It is invertible unless $z = 0$, so $\operatorname{supp}\beta = \{0\}$, which is compact. Hence β defines a class in $K_c(\mathbb{C})$, the **Bott class**.

Can we deform it away? Try the homotopy $z \mapsto z - c(s)$. It moves the zero to $z = c(s)$ but never removes it. More convincingly, on a large circle $|z| = R$ the map is the loop $z/|z|$ in $\mathrm{U}(1)$, and its winding number 1 cannot change under any compactly supported homotopy (which leaves the map invertible on a large enough circle for all s). A complex with empty support would give winding number 0 there (an invertible function on the whole disc $|z| \le R$ winds 0 times). So $[\beta] \neq 0$. More generally, the map z^n gives a class $[\beta_n]$ whose large-circle winding number is n.

The bundle on the sphere. Compactify $\mathbb{C}^+ = S^2$. Cover it by the disc $D_0 = \{|z| \le 1\}$ and the cap $D_\infty = \{|z| \ge 1\} \cup \{\infty\}$. On D_∞ the map z is invertible, so we may use it to identify the fibre of E^1 with the fibre of E^0. Doing so, E^0 is the trivial bundle on all of S^2, while E^1 becomes the line bundle L obtained by gluing the trivial bundle on D_0 to the trivial bundle on D_∞ with transition function $g(z) = z$ on the equator $|z| = 1$. So

$$[\beta] \;=\; [\mathbf{1}] - [L] \;\in\; \widetilde{K}(S^2), \tag{4.18}$$

where $\mathbf{1}$ is the trivial line bundle. A transition function $z = e^{i\phi}$ with winding number one produces a line bundle of degree ± 1: the monopole bundle of charge one (the tautological bundle over $\mathbb{CP}^1$ or its dual, depending on orientation conventions). Its first Chern class integrates to ± 1. It is a basic fact (Bott periodicity in its simplest form) that $\widetilde{K}(S^2) \cong \mathbb{Z}$, generated by exactly this class.

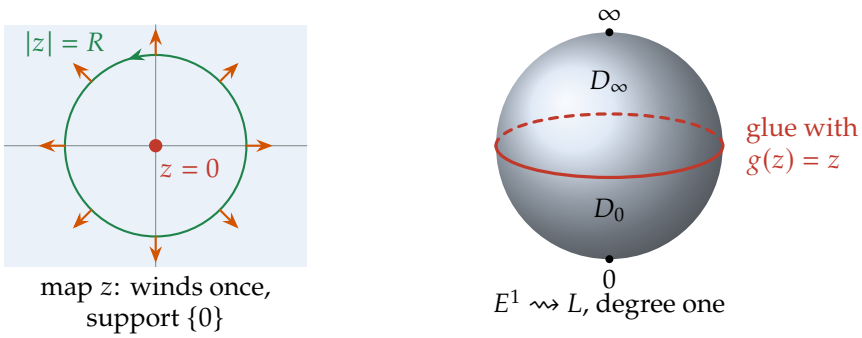


Figure 4.9: Left: the Bott complex on $\mathbb{C}$. Its map z vanishes only at the origin (red), and on a large circle it is the phase $z/|z|$ (orange arrows), which winds once. Right: after one-point compactification to S^2, the invertibility of z near ∞ is used to glue the two halves of E^1 with transition function z along the equator, producing the charge-one monopole line bundle L.

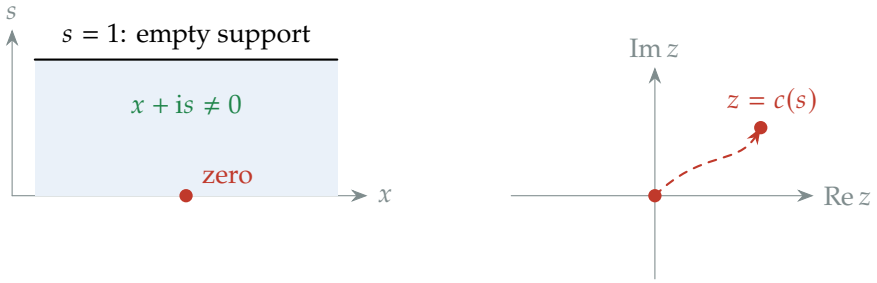
Worked example 4.14: One real dimension is not enough: $K_c(\mathbb{R}) = 0$

On $Y = \mathbb{R}$ take $0 \to \mathbb{C} \xrightarrow{\cdot x} \mathbb{C} \to 0$, multiplication by the real coordinate x. Its support is $\{0\}$. Now use the homotopy

$$\alpha_s(x) = x + \mathrm{i}s, \qquad s \in [0,1].$$

This vanishes only at $(x,s) = (0,0)$, so the support of the homotopy in $\mathbb{R} \times [0,1]$ is a single point: compact, as required. At $s = 1$ the map $x + \mathrm{i}$ never vanishes, so the complex has empty support and is zero. Therefore $[\cdot x] = 0$. The zero was able to escape into the imaginary direction. On $\mathbb{C}$ there is no extra direction to escape into: the zero of z is trapped by its winding number. In the language of one-point compactifications, $\mathbb{R}^+ = S^1$ and $\widetilde{K}(S^1) = 0$, whereas $\mathbb{C}^+ = S^2$ and $\widetilde{K}(S^2) = \mathbb{Z}$. More generally Bott periodicity says

$$K_c(\mathbb{R}^{2n}) = \widetilde{K}(S^{2n}) \cong \mathbb{Z}, \qquad K_c(\mathbb{R}^{2n+1}) = \widetilde{K}(S^{2n+1}) = 0.$$



on $\mathbb{R}$: the zero disappears on $\mathbb{C}$: the zero only moves

Figure 4.10: Left: on the real line the zero of x can be removed by the compactly supported homotopy $x + \mathrm{i}s$, so the class is 0. Right: on the complex plane any homotopy of z merely drags the zero around; it can never be removed, because the winding number on a large circle is fixed.

4.3.5 Products of complexes

$K(X)$ is a ring: $[V] \cdot [W] = [V \otimes W]$. On complexes, the product is the tensor product of complexes, which we need for the Thom homomorphism in Section 4.6. For two-term complexes $A = (A^0 \xrightarrow{a} A^1)$ and $B = (B^0 \xrightarrow{b} B^1)$,

$$A \otimes B: \quad 0 \to A^0 \otimes B^0 \xrightarrow{\begin{pmatrix} a\otimes 1 \\ 1\otimes b \end{pmatrix}} (A^1 \otimes B^0) \oplus (A^0 \otimes B^1) \xrightarrow{(-1\otimes b \;\; a\otimes 1)} A^1 \otimes B^1 \to 0. \tag{4.19}$$

The composite of the two maps is $-(a \otimes b) + (a \otimes b) = 0$, so this is a complex (the minus sign is the same sign as in the Leibniz rule $\mathrm{d}(\alpha \wedge \beta) = \mathrm{d}\alpha \wedge \beta + (-1)^p \alpha \wedge \mathrm{d}\beta$). Two facts make it useful:

1. Its class is the product of classes: by (4.16), $[A \otimes B] = A^0B^0 - (A^1B^0 + A^0B^1) + A^1B^1 = (A^0 - A^1)(B^0 - B^1)$ (writing V for $[V]$).
2. $\operatorname{supp}(A \otimes B) = \operatorname{supp} A \cap \operatorname{supp} B$. At a point where, say, a is invertible, one checks directly that the sequence is exact (the cohomology of a tensor product of complexes of vector spaces is the tensor product of their cohomologies, and one factor has none).

The second fact explains why products of compactly supported classes are compactly supported, and gives a quick proof of a relation we will use.

Worked example 4.15: $(1-L)^2=0$ in $K(S^2)$

The class $\mathbf{1}-L$ of (4.18) is represented by the complex $0\to\mathbf{1}\xrightarrow{s}L\to 0$, where s is a section of $\mathrm{Hom}(\mathbf{1},L)=L$. A degree-one line bundle has a section with a single simple zero, and we may put that zero anywhere, say at p or at $q\neq p$. Choose one representative with support $\{p\}$ and one with support $\{q\}$. Then

$$(\mathbf{1}-L)^2 \overset{(1)}{=} \left[(\mathbf{1}\xrightarrow{s_p}L)\otimes(\mathbf{1}\xrightarrow{s_q}L)\right] \overset{(2)}{=} [\text{complex with support } \{p\}\cap\{q\}=\varnothing] \overset{(3)}{=} 0.$$

(1) Product of classes is the class of the tensor product.
(2) Support of a tensor product is the intersection of the supports.
(3) Empty support means zero in $C(X)/C_\phi(X)$.

Consequently $L^n=\big(\mathbf{1}-(\mathbf{1}-L)\big)^n=\mathbf{1}-n(\mathbf{1}-L)$ by the binomial theorem, i.e. $[\beta_n]=\mathbf{1}-L^n=n(\mathbf{1}-L)=n[\beta]$: the class of the map z^n is n times the Bott class. The K-class sees exactly the winding number.

4.3.6 Why this is the right language for symbols

Now return to an elliptic operator $D\colon\Gamma(X,E)\to\Gamma(X,F)$ on a compact manifold. Its leading symbol gives a two-term complex over T^*X,

$$0\to\pi^*E\xrightarrow{\sigma_m(D)}\pi^*F\to 0. \tag{4.20}$$

Ellipticity says exactly that the support is contained in the zero section $\{k=0\}\cong X$, which is compact. So (4.20) defines a class in $K_c(T^*X)$, which Nash simply writes $K(T^*X)$ (footnote 2, p. 91). Homotopies of elliptic symbols are compactly supported homotopies of complexes. So:

$$\textit{deformation classes of elliptic symbols}\longrightarrow K(T^*X).$$

This is the map through which the index will factor.

Worked example 4.16: Two symbols as complexes

(a) For $D=\mathrm{d}/\mathrm{d}x$ on S^1, $T^*S^1=S^1\times\mathbb{R}$ with coordinates (x,k), and the symbol complex is $0\to\mathbb{C}\xrightarrow{\mathrm{i}k}\mathbb{C}\to 0$, with support $S^1\times\{0\}$: a circle, compact.
(b) For $\bar\partial=\frac12(\partial_x+\mathrm{i}\partial_y)$ on $\mathbb{R}^2$, the symbol at a fixed base point is

$$\sigma(\bar\partial)(k)=\tfrac{\mathrm{i}}{2}(k_x+\mathrm{i}k_y)=\tfrac{\mathrm{i}}{2}\,\kappa,\qquad \kappa:=k_x+\mathrm{i}k_y.$$

As a function of the momentum plane this is (up to the harmless nonzero factor $\mathrm{i}/2$) exactly the Bott complex (4.17) in the variable κ. The Cauchy–Riemann operator *is* the Bott class, one fibre of $T^*\mathbb{R}^2$ at a time. This is why $\bar\partial$ and its relatives (Dolbeault, Dirac) generate everything in index theory.

Reading Nash (pp. 89–91, (4.2)–(4.5) and footnotes 1–2)

Nash recalls $K(X)$ as the group completion of $\mathrm{Vect}(X)$, then introduces Segal's complexes (4.2), their homotopies, the semigroup $C(X)$, the support (4.3), the sub-semigroup $C_\phi(X)$ of complexes with empty support, and the isomorphism $K(X)=C(X)/C_\phi(X)$ (4.4). For

compact X he quotes $[E]=\sum_i(-1)^i[E^i]$ (4.5), which we derived in (4.16) from the homotopy $s\alpha_i$. His footnotes say that for non-compact X one uses complexes with compact support, and that $K(T^*X)$ means compactly supported K-theory, equivalently the reduced K-theory of the one-point compactification. Nash writes ϕ for the empty set in C_ϕ; we have kept his notation.

Recap

- A complex of bundles is a string of bundle maps with $\alpha_i\alpha_{i-1}=0$; its support is where it fails to be exact.
- $K(X)=C(X)/C_\phi(X)$: complexes up to homotopy, exact ones counting as zero. Degree i carries the sign $(-1)^i$.
- Compact X: $[E]=\sum(-1)^i[E^i]$. Non-compact Y: use compact supports; $K_c(Y)=\widetilde{K}(Y^+)$, and the maps carry the information.
- Bott class: $0\to\mathbb{C}\xrightarrow{z}\mathbb{C}\to 0$ on $\mathbb{C}$, the generator of $K_c(\mathbb{C})=\widetilde{K}(S^2)=\mathbb{Z}$; the map z^n gives n times it. $K_c(\mathbb{R})=0$.
- Tensor products of complexes multiply classes and intersect supports.
- An elliptic symbol is a two-term complex over T^*X supported on the zero section, hence a class in $K(T^*X)$.

4.4 Elliptic complexes and the rolled-up operator

The most important operators in geometry do not come alone. The exterior derivative d maps 0-forms to 1-forms, 1-forms to 2-forms, and so on, with $\mathrm{d}\circ\mathrm{d}=0$. In three dimensions this is the familiar chain

$$\text{functions}\xrightarrow{\nabla}\text{vector fields}\xrightarrow{\nabla\times}\text{vector fields}\xrightarrow{\nabla\cdot}\text{functions},\qquad \nabla\times\nabla=0,\quad \nabla\cdot\nabla\times=0.$$

None of these operators is elliptic on its own: the gradient maps a 1-component field to a 3-component field, so its symbol cannot be an invertible square matrix. Yet the chain as a whole is "elliptic" in a sense we now make precise, and it has an index, the Euler characteristic. Nash's (4.6)–(4.16) set this up. The key trick is to roll the whole chain up into a *single* elliptic operator D, so that everything we know about indices of single operators applies.

4.4.1 Complexes of operators and their symbol complexes

Working definition 4.14 (complex of operators, Nash (4.6))

Let $E^0,\dots,E^{n+1}$ be vector bundles over a compact manifold X. A **complex of operators** is a sequence of differential (or pseudo-differential [acting on waves through a symbol that need not be a polynomial in k]) operators, all of the same order m [at most m derivatives, or a symbol growing like $|k|^m$],

$$0\to\Gamma(X,E^0)\xrightarrow{d_0}\Gamma(X,E^1)\xrightarrow{d_1}\cdots\xrightarrow{d_n}\Gamma(X,E^{n+1})\to 0,\qquad d_i\circ d_{i-1}=0. \tag{4.21}$$

[$\Gamma(X,E^i)$ is the space of smooth sections, the fields $x\mapsto e(x)\in E^i_x$]

(Atiyah and Bott allow operators of different orders; Nash, and we, do not need that generality.) Each d_i has a leading symbol $\sigma_m(d_i)(x,k)\colon E^i_x \to E^{i+1}_x$. The first thing to check is that the symbols form a complex too. The cleanest way is to feed the operators a rapidly oscillating wave. For a real function φ with $\mathrm{d}\varphi = k$ at x and a large parameter λ, an operator P of order m satisfies

$$P\big(e^{\mathrm{i}\lambda\varphi}u\big) = \lambda^m\, e^{\mathrm{i}\lambda\varphi}\big(\sigma_m(P)(x,\mathrm{d}\varphi)u + O(\lambda^{-1})\big), \tag{4.22}$$

because every derivative that hits the exponential brings down $\mathrm{i}\lambda\,\partial_\mu\varphi$, and the terms with the most such factors are exactly the top-order terms with $\partial_\mu \to \mathrm{i}k_\mu$. Applying this twice,

$$\begin{aligned} 0 = d_i d_{i-1}\big(e^{\mathrm{i}\lambda\varphi}u\big) &\overset{(1)}{=} d_i\Big(\lambda^m e^{\mathrm{i}\lambda\varphi}\big(\sigma_m(d_{i-1})u + O(\lambda^{-1})\big)\Big) \\ &\overset{(2)}{=} \lambda^{2m} e^{\mathrm{i}\lambda\varphi}\big(\sigma_m(d_i)\sigma_m(d_{i-1})u + O(\lambda^{-1})\big). \end{aligned} \tag{4.23}$$

(1) (4.22) for $P = d_{i-1}$.
(2) (4.22) for $P = d_i$, applied to the leading term.

Dividing by λ^{2m} and letting $\lambda \to \infty$ gives $\sigma_m(d_i)\sigma_m(d_{i-1}) = 0$. So we get Nash's **symbol complex** (4.7), a complex of bundles over T^*X in the sense of Section 4.3:

$$\sigma(E)\colon \quad 0 \to \pi^*E^0 \xrightarrow{\sigma_m(d_0)} \pi^*E^1 \xrightarrow{\sigma_m(d_1)} \cdots \xrightarrow{\sigma_m(d_n)} \pi^*E^{n+1} \to 0. \tag{4.24}$$

Working definition 4.15 (elliptic complex)

The complex (4.21) is **elliptic** if its symbol complex (4.24) is exact [the kernel of each map equals the image of the previous one] off the zero section [the points $(x, k = 0)$] of T^*X: for every x and every $k \neq 0$, the sequence of matrices $\sigma_m(d_i)(x,k)$ is exact. Equivalently, the support of $\sigma(E)$ [the points (x,k) where the symbol sequence fails to be exact] lies in the zero section.
Operational test: at a fixed $k \neq 0$ compute the ranks r_i of the symbol matrices; exactness means $r_i + r_{i-1} = \operatorname{rank} E^i$ for every i (kernel of the i-th map = image of the $(i-1)$-th, given that the composite is zero).

Note that the complex of operators itself has no reason to be exact (its cohomology will turn out to be the interesting part). Only the symbol complex must be exact. If there are only two nonzero terms, $0 \to \Gamma(E^0) \xrightarrow{d_0} \Gamma(E^1) \to 0$, exactness of the symbol complex says that $\sigma_m(d_0)$ is an isomorphism for $k \neq 0$: that is ordinary ellipticity of d_0.

Worked example 4.17: The de Rham complex is elliptic

Take $E^p = \Lambda^p T^*X \otimes \mathbb{C}$ (complex-valued p-forms) and $d_p = \mathrm{d}$. Using (4.22) with $P = \mathrm{d}$,

$$\begin{aligned} \mathrm{d}\big(e^{\mathrm{i}\lambda\varphi}\omega\big) &\overset{(1)}{=} \mathrm{i}\lambda\,\mathrm{d}\varphi \wedge e^{\mathrm{i}\lambda\varphi}\omega + e^{\mathrm{i}\lambda\varphi}\mathrm{d}\omega \\ &\overset{(2)}{=} \lambda\, e^{\mathrm{i}\lambda\varphi}\big(\mathrm{i}k \wedge \omega + O(\lambda^{-1})\big), \end{aligned}$$

(1) Leibniz rule for d acting on a function times a form.
(2) $\mathrm{d}\varphi = k$ at the point; the second term has no factor of λ.

so $\sigma_1(\mathrm{d})(x,k)\omega = \mathrm{i}\,k \wedge \omega$. Exactness for $k \neq 0$: choose an orthonormal basis $e^1, \dots, e^n$ of T^*_xX with $e^1 = k/|k|$. Every form splits uniquely as $\omega = e^1 \wedge \alpha + \beta$ with α, β built from $e^2, \dots, e^n$ only. Then

$$k \wedge \omega \overset{(1)}{=} |k|\, e^1 \wedge e^1 \wedge \alpha + |k|\, e^1 \wedge \beta \overset{(2)}{=} |k|\, e^1 \wedge \beta,$$

(1) Substitute $k = |k|e^1$ and the splitting of ω.
(2) $e^1 \wedge e^1 = 0$.

which vanishes iff $\beta = 0$, i.e. iff $\omega = e^1 \wedge \alpha = k \wedge (\alpha/|k|)$. So $\mathrm{Ker}(k\wedge) = \mathrm{Im}(k\wedge)$: exact. This is Nash's "$k \wedge v = 0 \Rightarrow v = k \wedge w$" (p. 94).

Worked example 4.18: grad, curl, div in momentum space

On $\mathbb{R}^3$ the de Rham complex, with 1- and 2-forms identified with vector fields, is ∇, $\nabla\times$, $\nabla\cdot$. Its symbol complex at momentum $k = (k_1, k_2, k_3)$ is

$$0 \to \mathbb{C} \xrightarrow{\mathrm{i}k} \mathbb{C}^3 \xrightarrow{\mathrm{i}k\times} \mathbb{C}^3 \xrightarrow{\mathrm{i}k\cdot} \mathbb{C} \to 0,$$

with matrices

$$\sigma_0 = \mathrm{i}\begin{pmatrix} k_1 \\ k_2 \\ k_3 \end{pmatrix}, \quad \sigma_1 = \mathrm{i}\begin{pmatrix} 0 & -k_3 & k_2 \\ k_3 & 0 & -k_1 \\ -k_2 & k_1 & 0 \end{pmatrix}, \quad \sigma_2 = \mathrm{i}\begin{pmatrix} k_1 & k_2 & k_3 \end{pmatrix}.$$

The composites vanish: $\sigma_1\sigma_0 = -k \times k = 0$ and $\sigma_2\sigma_1 v = -k \cdot (k \times v) = 0$. For $k \neq 0$ the ranks are $r_0 = 1$ (a nonzero column), $r_1 = 2$ (the kernel of $v \mapsto k \times v$ is the line through k), $r_2 = 1$ (a nonzero row). The exactness test:

$$r_0 = 1 = \mathrm{rank}\, E^0, \qquad r_0 + r_1 = 3 = \mathrm{rank}\, E^1,$$
$$r_1 + r_2 = 3 = \mathrm{rank}\, E^2, \qquad r_2 = 1 = \mathrm{rank}\, E^3.$$

All satisfied, so the complex is elliptic. At $k = 0$ all maps vanish and exactness fails: the support is the zero section. Physically: a plane wave field is longitudinal ($\propto k$) exactly when it is curl-free, and transverse ($\perp k$) exactly when it is divergence-free and then it is a curl.

The Dolbeault complex $\bar\partial\colon \Omega^{0,q} \to \Omega^{0,q+1}$ on a complex manifold is elliptic by the same argument, with $k\wedge$ replaced by $k^{0,1}\wedge$; the next part treats it in detail.

4.4.2 Rolling the complex up into one operator

Choose Hermitian metrics on the bundles E^i and a volume form on X, so that every d_i has a formal adjoint $d_i^*\colon \Gamma(E^{i+1}) \to \Gamma(E^i)$ (Section 4.2.3). Taking adjoints of $d_i d_{i-1} = 0$ gives

$$d_{i-1}^* d_i^* = (d_i d_{i-1})^* = 0. \tag{4.25}$$

Group the bundles into even and odd (Nash (4.8)):

$$E^+ = E^0 \oplus E^2 \oplus E^4 \oplus \cdots, \qquad E^- = E^1 \oplus E^3 \oplus E^5 \oplus \cdots. \tag{4.26}$$

Working definition 4.16 (rolled-up operator, Nash (4.9))

$$D = \sum_i \left(d_{2i} + d_{2i+1}^*\right)\colon \Gamma(X, E^+) \to \Gamma(X, E^-),$$
$$D(e_0, e_2, e_4, \dots) = (d_0 e_0 + d_1^* e_2,\ d_2 e_2 + d_3^* e_4,\ \dots). \tag{4.27}$$

As a matrix of operators, with rows labelled by $E^1, E^3, \dots$ and columns by $E^0, E^2, \dots$ [each d_i runs from E^i up to E^{i+1}, and its formal adjoint d_i^* runs back down from E^{i+1} to E^i],

$$D = \begin{pmatrix} d_0 & d_1^* & 0 & \cdots \\ 0 & d_2 & d_3^* & \cdots \\ 0 & 0 & d_4 & \cdots \\ \vdots & & & \ddots \end{pmatrix}, \qquad D^* = \begin{pmatrix} d_0^* & 0 & 0 & \cdots \\ d_1 & d_2^* & 0 & \cdots \\ 0 & d_3 & d_4^* & \cdots \\ \vdots & & & \ddots \end{pmatrix}.$$

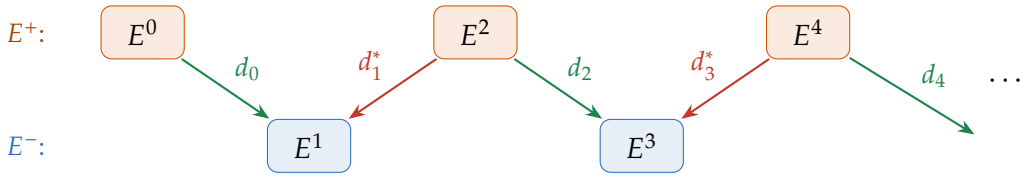


Figure 4.11: The rolled-up operator. The even bundles (top row, orange) make E^+, the odd ones (bottom row, blue) make E^-. D is the sum of all downward arrows: each d_{2i} steps forward (green) and each adjoint d_{2i+1}^* steps back (red). A single elliptic operator $E^+ \to E^-$ remembers the whole complex.

Worked example 4.19: The de Rham complex of a surface

For a surface, $E^+ = \Lambda^0 \oplus \Lambda^2$ and $E^- = \Lambda^1$, and $D(f, \beta) = \mathrm{d}f + \mathrm{d}^*\beta$. On the flat torus $T^2 = \mathbb{R}^2/(2\pi\mathbb{Z})^2$ write $\beta = g\,\mathrm{d}x \wedge \mathrm{d}y$. With the Hodge star $\star \mathrm{d}x = \mathrm{d}y$, $\star \mathrm{d}y = -\mathrm{d}x$, $\star(\mathrm{d}x \wedge \mathrm{d}y) = 1$ and $\mathrm{d}^* = -\star\,\mathrm{d}\star$ on 2-forms in two dimensions,

$$\mathrm{d}^*(g\,\mathrm{d}x \wedge \mathrm{d}y) \overset{(1)}{=} -\star\,\mathrm{d}g \overset{(2)}{=} -\star\,(g_x \mathrm{d}x + g_y \mathrm{d}y) \overset{(3)}{=} g_y\,\mathrm{d}x - g_x\,\mathrm{d}y.$$

(1) $\star(g\,\mathrm{d}x \wedge \mathrm{d}y) = g$.
(2) Exterior derivative of a function.
(3) $\star \mathrm{d}x = \mathrm{d}y$ and $\star \mathrm{d}y = -\mathrm{d}x$.

So, writing a 1-form as the column of its $\mathrm{d}x$ and $\mathrm{d}y$ components,

$$D\begin{pmatrix} f \\ g \end{pmatrix} = \begin{pmatrix} \partial_x & \partial_y \\ \partial_y & -\partial_x \end{pmatrix}\begin{pmatrix} f \\ g \end{pmatrix}, \qquad \sigma_1(D)(k) = \mathrm{i}\begin{pmatrix} k_x & k_y \\ k_y & -k_x \end{pmatrix}.$$

The symbol squares to $-(k_x^2 + k_y^2)\mathbb{1}$ (multiply it out), so it is invertible for $k \neq 0$: D is elliptic, as promised, although the individual d's are not square. The equations $D(f, g) = 0$ read $f_x + g_y = 0$, $f_y - g_x = 0$: these are the Cauchy–Riemann equations for $f - \mathrm{i}g$. On a Fourier mode $e^{\mathrm{i}(Mx+Ny)}$ the matrix of D is $\sigma_1(D)(M, N)$, invertible unless $M = N = 0$. Hence $\operatorname{Ker} D$ = constant pairs (f, g), dimension 2; similarly $\operatorname{Ker} D^*$ = constant 1-forms $a\,\mathrm{d}x + b\,\mathrm{d}y$, dimension 2; and index $D = 2 - 2 = 0$. We will see in a moment that this is $\chi(T^2) = 1 - 2 + 1$.

4.4.3 Why the rolled-up operator is elliptic

Nash (p. 92) says that ellipticity of D "follows" from exactness of $\sigma(E)$. Here is the linear algebra behind it. Fix (x, k) with $k \neq 0$ and write σ_i for the matrix $\sigma_m(d_i)(x, k)$. With our convention (keeping the i's) the symbol of an adjoint is the adjoint matrix, $\sigma_m(d_i^*) = \sigma_i^\dagger$. So the symbol of D is $\sigma(D) = \sum_i (\sigma_{2i} + \sigma_{2i+1}^\dagger)$, and the same block computation we are about to do for D^*D in (4.30)

gives

$$\sigma(D)^\dagger\sigma(D) = \bigoplus_{i \text{ even}} \left(\sigma_i^\dagger\sigma_i + \sigma_{i-1}\sigma_{i-1}^\dagger\right).$$

Now suppose $v \in E_x^i$ is killed by the i-th block. Then

$$0 = \left\langle v, (\sigma_i^\dagger\sigma_i + \sigma_{i-1}\sigma_{i-1}^\dagger)v\right\rangle \overset{(1)}{=} \|\sigma_i v\|^2 + \left\|\sigma_{i-1}^\dagger v\right\|^2, \tag{4.28}$$

(1) Move one factor of each product to the other side of the inner product.

so $\sigma_i v = 0$ and $\sigma_{i-1}^\dagger v = 0$. The first says $v \in \operatorname{Ker}\sigma_i = \operatorname{Im}\sigma_{i-1}$ (exactness). The second says $v \perp \operatorname{Im}\sigma_{i-1}$ (by the same argument that gave (4.1)). A vector in a subspace and orthogonal to it is zero. So $\sigma(D)^\dagger\sigma(D)$ is injective, hence invertible, hence $\sigma(D)$ is injective. The same argument for $\sigma(D)\sigma(D)^\dagger$ shows $\sigma(D)$ is onto. So $\sigma(D)$ is an isomorphism $E_x^+ \to E_x^-$ for $k \neq 0$, i.e. D is elliptic. In particular $\operatorname{rank} E^+ = \operatorname{rank} E^-$, i.e. $\sum_i (-1)^i \operatorname{rank} E^i = 0$, which is the familiar fact that the alternating sum of dimensions of an exact sequence vanishes.

In the language of Section 4.3, the two-term complex $0 \to \pi^*E^+ \xrightarrow{\sigma(D)} \pi^*E^- \to 0$ (Nash (4.10)) is exact off the zero section. It is a "rolled-up" version of $\sigma(E)$, and one can show it represents the same class in $K(T^*X)$.

4.4.4 Laplacians and harmonic sections

Working definition 4.17 (Laplacians of a complex)

The **Laplacians** of the complex are

$$\Delta_i = d_i^*d_i + d_{i-1}d_{i-1}^* \colon \Gamma(X, E^i) \to \Gamma(X, E^i).$$

For the de Rham complex these are the Hodge Laplacians $\Delta = \mathrm{d}^*\mathrm{d} + \mathrm{d}\mathrm{d}^*$ [d^* is the adjoint of d, lowering the form degree by one] on p-forms; on functions on $\mathbb{R}^n$, $\Delta_0 = \mathrm{d}^*\mathrm{d} = -\sum_\mu \partial_\mu^2$.

A section e with $\Delta_i e = 0$ is called **harmonic**. On a compact manifold harmonic is the same as "closed and co-closed":

$$\begin{aligned}\Delta_i e = 0 \implies 0 = \langle e, \Delta_i e\rangle &\overset{(1)}{=} \langle e, d_i^*d_i e\rangle + \langle e, d_{i-1}d_{i-1}^* e\rangle \\ &\overset{(2)}{=} \|d_i e\|^2 + \left\|d_{i-1}^* e\right\|^2\end{aligned} \tag{4.29}$$

(1) Definition of Δ_i.
(2) Definition of the adjoint, $\langle e, d^*de\rangle = \langle de, de\rangle$; no boundary terms because X is closed.

so $d_i e = 0$ and $d_{i-1}^* e = 0$; the converse is obvious. Nash writes this as $\Delta_i e_i = 0 \iff d_i e_i = d_{i-1}^* e_i = 0$.

Now compute D^*D on E^+. Its E^0 component and its E^2 component are

$$\begin{aligned}\left(D^*D(e_0, e_2, \dots)\right)_0 &\overset{(1)}{=} d_0^*\left(d_0e_0 + d_1^*e_2\right) \\ &\overset{(2)}{=} d_0^*d_0e_0 + (d_1d_0)^*e_2 \overset{(3)}{=} \Delta_0 e_0, \\ \left(D^*D(e_0, e_2, \dots)\right)_2 &\overset{(4)}{=} d_1\left(d_0e_0 + d_1^*e_2\right) + d_2^*\left(d_2e_2 + d_3^*e_4\right) \\ &\overset{(5)}{=} d_1d_0e_0 + \left(d_1d_1^* + d_2^*d_2\right)e_2 + (d_3d_2)^*e_4 \overset{(6)}{=} \Delta_2 e_2.\end{aligned} \tag{4.30}$$

(1) Row 0 of D^* is $(d_0^*, 0, 0, \dots)$, applied to the first component of De.

(2) Expand; $d_0^* d_1^* = (d_1 d_0)^*$.
(3) $d_1 d_0 = 0$; and $\Delta_0 = d_0^* d_0$ since there is no d_{-1}.
(4) Row 2 of D^* is $(d_1, d_2^*, 0, \dots)$.
(5) Expand.
(6) $d_1 d_0 = 0$ and $d_3 d_2 = 0$ kill the cross terms; what is left is the definition of Δ_2.

The same pattern continues. So all the cross terms cancel because $d \circ d = 0$, and (Nash (4.11))

$$\Delta^+ := D^* D = \Delta_0 \oplus \Delta_2 \oplus \Delta_4 \oplus \cdots, \qquad \Delta^- := D D^* = \Delta_1 \oplus \Delta_3 \oplus \cdots. \tag{4.31}$$

Since these are block diagonal, their kernels are direct sums (Nash (4.12)): $\operatorname{Ker} \Delta^+ = \bigoplus_i \operatorname{Ker} \Delta_{2i}$ and $\operatorname{Ker} \Delta^- = \bigoplus_i \operatorname{Ker} \Delta_{2i+1}$. And $\operatorname{Ker} D = \operatorname{Ker} \Delta^+$ by the same positivity trick (Nash (4.13)):

$$D^* D e = 0 \implies 0 = \langle e, D^* D e \rangle \overset{(1)}{=} \langle De, De \rangle \implies De = 0, \tag{4.32}$$

(1) Definition of the adjoint.

and conversely $De = 0 \Rightarrow D^* De = 0$. Likewise $\operatorname{Ker} D^* = \operatorname{Ker} \Delta^-$. Therefore (Nash (4.14))

$$\begin{aligned} \operatorname{index} D &\overset{(1)}{=} \dim \operatorname{Ker} D - \dim \operatorname{Ker} D^* \\ &\overset{(2)}{=} \dim \operatorname{Ker} \Delta^+ - \dim \operatorname{Ker} \Delta^- \\ &\overset{(3)}{=} \sum_i \dim \operatorname{Ker} \Delta_{2i} - \sum_i \dim \operatorname{Ker} \Delta_{2i+1} \\ &= \sum_i (-1)^i \dim \operatorname{Ker} \Delta_i. \end{aligned} \tag{4.33}$$

(1) Definition of the index, with $\operatorname{Coker} D \cong \operatorname{Ker} D^*$.
(2) $\operatorname{Ker} D = \operatorname{Ker} \Delta^+$ and $\operatorname{Ker} D^* = \operatorname{Ker} \Delta^-$.
(3) Block-diagonal form (4.31).

The index of D counts harmonic sections, with a sign for parity.

4.4.5 The Hodge theorem: harmonic sections are cohomology

The complex (4.21) has cohomology

$$H^i(E) := \frac{\operatorname{Ker} d_i}{\operatorname{Im} d_{i-1}}.$$

For the de Rham complex this is de Rham cohomology $H^i_{\mathrm{dR}}(X; \mathbb{C})$, whose dimensions are the Betti numbers b_i. Nash quotes (from Wells):

Key idea: Hodge theorem for elliptic complexes, Nash (4.15)

For an elliptic complex over a compact closed manifold,

$$H^i(E) \;\cong\; \operatorname{Ker} \Delta_i :$$

every cohomology class has exactly one harmonic representative.

The mechanism is pure linear algebra, and it is worth seeing in finite dimensions first. Let $0 \to V^0 \xrightarrow{d_0} V^1 \to \cdots$ be a complex of finite-dimensional inner-product spaces. Then each V^i splits orthogonally into three pieces,

$$V^i = \operatorname{Im} d_{i-1} \;\oplus\; \operatorname{Ker} \Delta_i \;\oplus\; \operatorname{Im} d_i^*. \tag{4.34}$$

Why: $\operatorname{Im} d_{i-1} \perp \operatorname{Im} d_i^*$ because $\langle d_{i-1}a, d_i^* b\rangle = \langle d_i d_{i-1} a, b\rangle = 0$. A vector orthogonal to both satisfies $d_{i-1}^* v = 0$ (orthogonal to $\operatorname{Im} d_{i-1}$) and $d_i v = 0$ (orthogonal to $\operatorname{Im} d_i^*$), i.e. it is harmonic by (4.29). Moreover $\operatorname{Ker} d_i = (\operatorname{Im} d_i^*)^\perp = \operatorname{Im} d_{i-1} \oplus \operatorname{Ker} \Delta_i$. Hence

$$H^i \overset{(1)}{=} \frac{\operatorname{Ker} d_i}{\operatorname{Im} d_{i-1}} \overset{(2)}{=} \frac{\operatorname{Im} d_{i-1} \oplus \operatorname{Ker} \Delta_i}{\operatorname{Im} d_{i-1}} \overset{(3)}{=} \operatorname{Ker} \Delta_i. \tag{4.35}$$

(1) Definition of cohomology.
(2) The splitting of $\operatorname{Ker} d_i$ just derived.
(3) Quotienting a direct sum by one summand leaves the other.

In infinite dimensions the same argument works once one knows that the images are closed and the pieces fit together, and *that* is where ellipticity (via the Fredholm property of D and elliptic regularity) is needed.

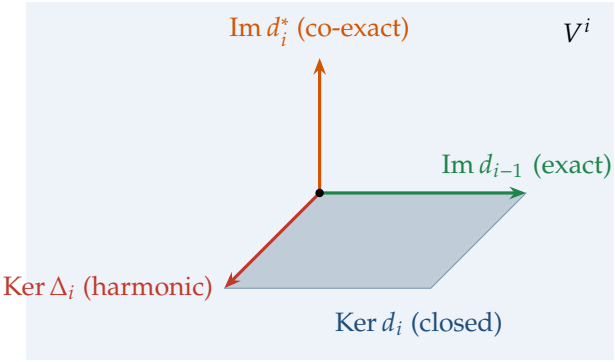


Figure 4.12: The Hodge decomposition (4.34), drawn like three perpendicular coordinate axes. Closed vectors fill the shaded plane spanned by the exact and harmonic axes; modding out by exact ones leaves exactly the harmonic direction. So each cohomology class has one harmonic representative, the one of smallest norm in its class.

Combining (4.33) with the Hodge theorem gives Nash's (4.16):

$$\operatorname{index} D = \sum_i (-1)^i \dim H^i(E) =: \chi(E), \tag{4.36}$$

the **Euler characteristic** of the elliptic complex. For the de Rham complex it is the ordinary Euler characteristic $\chi(X) = \sum_i (-1)^i b_i$.

Worked example 4.20: A finite complex: the triangulated circle

Triangulate the circle as a triangle with vertices v_1, v_2, v_3 and oriented edges $e_1 = [v_1 v_2]$, $e_2 = [v_2 v_3]$, $e_3 = [v_3 v_1]$. A 0-cochain assigns a number to each vertex, a 1-cochain a number to each edge, and the coboundary is "value at the head minus value at the tail":

$$0 \to \mathbb{C}^3 \xrightarrow{\ \delta\ } \mathbb{C}^3 \to 0, \qquad \delta = \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 1 & 0 & -1 \end{pmatrix}.$$

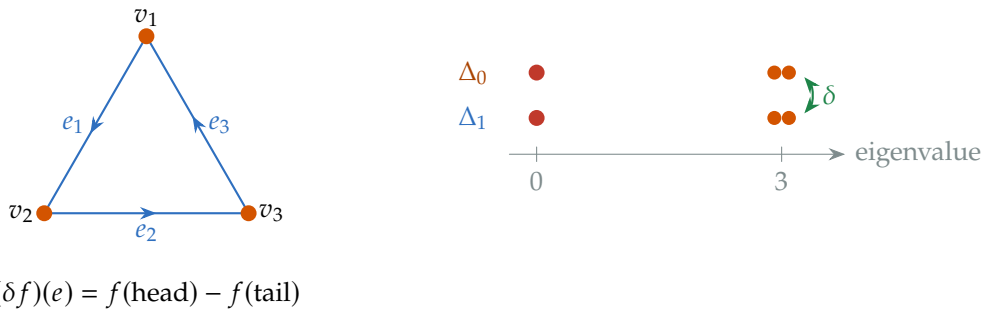
The rows sum to zero and any two are independent, so $\operatorname{rank} \delta = 2$. Then

$$\dim H^0 = \dim \operatorname{Ker} \delta = 3 - 2 = 1, \qquad \dim H^1 = \dim \operatorname{Coker} \delta = 3 - 2 = 1,$$

and $\chi = 1 - 1 = 0 = 3 - 3$. The Laplacians are

$$\Delta_0 = \delta^T\delta = \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}, \qquad \Delta_1 = \delta\delta^T = \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}.$$

Δ_0 is the graph Laplacian; its kernel is the constant cochain $(1,1,1)$, the harmonic representative of H^0. $\operatorname{Ker}\Delta_1$ is spanned by $(1,1,1)$ too: the "uniform circulation" 1-cochain, the harmonic representative of the generator of H^1 (the discrete version of $\mathrm{d}\theta$). The other eigenvalues of both Laplacians are $3, 3$; they occur in pairs, one for Δ_0 and one for Δ_1, matched by δ. Keep this pairing in mind: it is the heart of Section 4.8.



$(\delta f)(e) = f(\text{head}) - f(\text{tail})$

Figure 4.13: Left: the triangulated circle and its coboundary. Right: the spectra of Δ_0 and Δ_1. The nonzero eigenvalues (orange, 3 twice each) are matched by δ; only the zero modes (red), one on each side, are unmatched. Their difference $1 - 1 = 0$ is the Euler characteristic of the circle.

For finite complexes the Euler characteristic can be computed without finding the cohomology at all. Write $r_i = \operatorname{rank} d_i$. Then

$$\begin{aligned} \sum_i (-1)^i \dim H^i &\overset{(1)}{=} \sum_i (-1)^i \big(\dim \operatorname{Ker} d_i - r_{i-1}\big) \\ &\overset{(2)}{=} \sum_i (-1)^i \big(\dim V^i - r_i - r_{i-1}\big) \\ &\overset{(3)}{=} \sum_i (-1)^i \dim V^i. \end{aligned} \tag{4.37}$$

(1) $H^i = \operatorname{Ker} d_i / \operatorname{Im} d_{i-1}$ and $\dim \operatorname{Im} d_{i-1} = r_{i-1}$.
(2) Rank–nullity for d_i.
(3) The sum $\sum_i (-1)^i (r_i + r_{i-1})$ telescopes to zero (each r_j appears once with sign $(-1)^j$ and once with $(-1)^{j+1}$).

This is the finite-dimensional statement "the index depends only on the shape" again, (4.2), now for complexes. For the de Rham complex of a manifold each $V^i = \Omega^i$ is infinite-dimensional, the right-hand side is meaningless, and computing the left-hand side topologically is exactly what the index theorem achieves.

Worked example 4.21: $D = \mathrm{d} + \mathrm{d}^*$ on the circle and on the two-sphere

Circle. $E^+ = \Lambda^0$, $E^- = \Lambda^1$, $D = \mathrm{d}$. $\operatorname{Ker}\mathrm{d}$ = constant functions; $\operatorname{Ker}\mathrm{d}^*$ on 1-forms $h\,\mathrm{d}x$ is $\{h' = 0\}$, the constant multiples of $\mathrm{d}x$. So $\operatorname{index} D = 1 - 1 = 0 = \chi(S^1)$, in agreement with the triangulated circle.
Two-sphere. $E^+ = \Lambda^0 \oplus \Lambda^2$, $D(f,\beta) = \mathrm{d}f + \mathrm{d}^*\beta$. Suppose $D(f,\beta) = 0$. The two terms are

orthogonal:

$$\langle \mathrm{d}f, \mathrm{d}^*\beta\rangle \overset{(1)}{=} \langle \mathrm{d}\mathrm{d}f, \beta\rangle \overset{(2)}{=} 0,$$

(1) Move d^* across as d.
(2) $\mathrm{d}\circ\mathrm{d} = 0$.

so $0 = \|\mathrm{d}f + \mathrm{d}^*\beta\|^2 = \|\mathrm{d}f\|^2 + \|\mathrm{d}^*\beta\|^2$ forces $\mathrm{d}f = 0$ and $\mathrm{d}^*\beta = 0$: f is constant and β is a constant multiple of the area form. $\dim\operatorname{Ker} D = 2$. $\operatorname{Ker} D^* =$ harmonic 1-forms $\cong H^1_{\mathrm{dR}}(S^2) = 0$ (every closed loop on the sphere can be shrunk, so every closed 1-form is exact). So $\operatorname{index} D = 2 - 0 = 2 = \chi(S^2) = 1 - 0 + 1$. Compare the torus of Example 4.19: $\operatorname{index} D = 0 = 1 - 2 + 1$.

Physical picture: Why physicists like the rolled-up operator

Put $Q = \left(\begin{smallmatrix} 0 & D^* \\ D & 0 \end{smallmatrix}\right)$ acting on $E^+ \oplus E^-$. Then Q is self-adjoint and $Q^2 = \left(\begin{smallmatrix} D^*D & 0 \\ 0 & DD^* \end{smallmatrix}\right) = \Delta^+ \oplus \Delta^-$. This is a supercharge whose square is a Hamiltonian, with E^+ "bosonic" and E^- "fermionic". For the de Rham complex, $Q = \mathrm{d} + \mathrm{d}^*$ and the grading is the parity of the form degree: this is Witten's supersymmetric quantum mechanics on X, and (4.36) is the statement that its Witten index is $\chi(X)$. We return to this in Section 4.8 and in Chapter 7.

Reading Nash (pp. 91–93, (4.6)–(4.16))

Nash defines complexes of operators d_i of the same order m (4.6), their symbol complex $\sigma(E)$ over T^*X built from the leading symbols of p. 38 (4.7), and calls the complex elliptic when $\sigma(E)$ is exact off the zero section. He forms $E^{\pm}$ (4.8), the operator $D = \sum(d_{2i} + d^*_{2i+1})$ (4.9), notes the two-term complexes (4.10), and asserts that exactness of $\sigma(E)$ gives ellipticity of D; we supplied the positivity argument (4.28). The Laplacians, the decomposition $D^*D = \Delta_0 \oplus \Delta_2 \oplus \cdots$ (4.11)–(4.12), the kernel argument (4.13), the index as an alternating count of harmonic sections (4.14), the Hodge theorem (4.15) and $\operatorname{index} D = \chi(E)$ (4.16) are all there; we filled in the cross-term cancellation (4.30) and the linear algebra of the Hodge decomposition.

Recap

- A complex of operators d_i with $d_i d_{i-1} = 0$ is elliptic if its symbol complex is exact for $k \neq 0$ (de Rham: $\mathrm{i}k\wedge$; in $\mathbb{R}^3$: $\mathrm{i}k$, $\mathrm{i}k\times$, $\mathrm{i}k\cdot$).
- Rolled-up operator $D = \sum(d_{2i} + d^*_{2i+1})$: $E^+ \to E^-$ is elliptic by a positivity argument.
- $D^*D = \bigoplus \Delta_{2i}$, $DD^* = \bigoplus \Delta_{2i+1}$ because $d \circ d = 0$ kills cross terms.
- $\operatorname{index} D = \sum(-1)^i \dim\operatorname{Ker}\Delta_i = \sum(-1)^i \dim H^i(E) = \chi(E)$ by the Hodge theorem.
- Examples: $\chi(S^1) = 0$, $\chi(S^2) = 2$, $\chi(T^2) = 0$ from explicit kernels; nonzero eigenvalues of the Laplacians pair up.

4.5 The analytical index

We now have all the pieces for the first of Nash's two definitions. The question of this section is: *in what sense is the index a function on K-theory?* The answer is a map

$$\mathrm{ind}_a \colon K(T^*X) \longrightarrow \mathbb{Z},$$

the **analytical index**. Its input is a K-theory class of symbols; its output is the number of solutions of $D\psi = 0$ minus the number for $D^*\chi = 0$, for any operator D with that symbol class.

4.5.1 From an operator to a K-class

Start with an elliptic complex E, or just a single elliptic operator. Roll it up to $D \colon \Gamma(X, E^+) \to \Gamma(X, E^-)$ as in Section 4.4.2. The leading symbol of D is a two-term complex of bundles over T^*X (Nash (4.17)),

$$0 \longrightarrow \pi^*E^+ \xrightarrow{\sigma_m(D)} \pi^*E^- \longrightarrow 0, \tag{4.38}$$

which is exact off the zero section. By Section 4.3.6 it has compact support and defines a class

$$[\sigma_m(D)] \in K(T^*X) \qquad \text{(compactly supported K-theory)}.$$

For a single elliptic operator $D \colon \Gamma(E) \to \Gamma(F)$ it is just $0 \to \pi^*E \xrightarrow{\sigma_m(D)} \pi^*F \to 0$.

4.5.2 Why the index only depends on the class

For the rule $[\sigma_m(D)] \mapsto \operatorname{index} D$ to be a function on $K(T^*X)$, three things must hold. Each is a consequence of something we already know.

1. *Homotopic symbols give equal indices.* A compactly supported homotopy of two-term symbol complexes is a continuous family of elliptic symbols. Choosing operators with these symbols gives a continuous path of Fredholm operators, along which the index is constant (Sections 4.1.5 and 4.2.5). Operators with literally the same leading symbol differ by lower order terms and have the same index by (4.9).

2. *Direct sums add.* The symbol $\sigma_1 \oplus \sigma_2$ is the symbol of $D_1 \oplus D_2$, which acts block-diagonally; its kernel and cokernel are direct sums, so

$$\begin{aligned} \operatorname{index}(D_1 \oplus D_2) &\overset{(1)}{=} \dim(\operatorname{Ker} D_1 \oplus \operatorname{Ker} D_2) - \dim(\operatorname{Ker} D_1^* \oplus \operatorname{Ker} D_2^*) \\ &\overset{(2)}{=} \operatorname{index} D_1 + \operatorname{index} D_2. \end{aligned} \tag{4.39}$$

   (1) The adjoint of a block-diagonal operator is block-diagonal.
   (2) Dimensions add under direct sum; regroup.

3. *Symbols with empty support give index zero.* If the symbol is invertible even on the zero section, it can be deformed (by $\sigma(x,k) \to \sigma(x,sk)$, $s \to 0$, which does not create zeros) to the k-independent symbol $\sigma(x,0)$, which is the symbol of the order-zero operator "multiply by the invertible bundle map $\sigma(x,0)$". An invertible operator has no kernel and no cokernel, so its index is 0. (Such symbols only exist for order-zero, i.e. pseudo-differential, operators: a homogeneous symbol of degree $m > 0$ vanishes at $k = 0$.)

Items 1 and 2 make "index" a homomorphism on the semigroup $C(T^*X)$ of homotopy classes of symbol complexes; item 3 says it vanishes on $C_\phi(T^*X)$. By Segal's description (4.14), it therefore descends to the quotient $K(T^*X)$.

Working definition 4.18 (analytical index, Nash (4.18))

$$\operatorname{ind}_a : K(T^*X) \to \mathbb{Z}, \qquad [\sigma_m(D)] \longmapsto \operatorname{index} D = \dim \operatorname{Ker} D - \dim \operatorname{Ker} D^*. \tag{4.40}$$

Operational recipe: given a class, find any elliptic (pseudo-)differential operator whose symbol represents it [the class of its leading symbol $\sigma_m(D)$ in $K(T^*X)$ is the given class], and count its zero modes [independent solutions of $D\psi = 0$] minus the zero modes of its adjoint.

Watch out: Two hidden facts

The definition quietly assumes two things. First, that *every* element of $K(T^*X)$ is the class of some elliptic symbol; otherwise ind_a would only be defined on part of the group. This is true, but only if we allow pseudo-differential operators (Atiyah–Singer [2]); differential operators alone do not reach every class (on the circle they only reach 0, as we are about to see). Second, that the index depends only on the class, which is items 1–3 above; item 1 is where the Fredholm theory of Section 4.1 is really used.

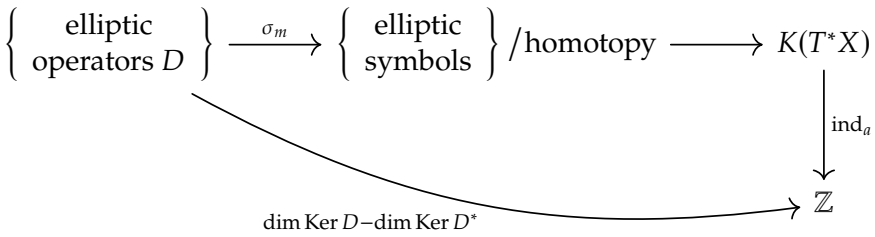


Figure 4.14: The analytical index as a factorisation. Counting zero modes (the curved arrow) is defined on operators, but it factors through the symbol, its homotopy class and finally its K-theory class. The vertical arrow is ind_a. The Atiyah–Singer theorem will give a second, topological description of the same vertical arrow.

Physical picture: The analytical index is renormalisation-group-blind

Physicists compute effective actions and anomalies by integrating out high-momentum modes; the answer can depend on how that is done (the choice of regulator). The factorisation in Figure 4.14 says the *index* cannot: two operators with homotopic leading symbols – however different their low-momentum, "infrared" details – have exactly the same index. Only the ultraviolet data, the leading symbol at large momentum, matters, and even that only through its topological class. This is the same phenomenon as the RG-invariance of an 't Hooft anomaly coefficient: it is fixed by the short-distance theory and cannot change under any smooth deformation of the low-energy physics.

4.5.3 Examples

Worked example 4.22: X a point: ind_a is the identity of $\mathbb{Z}$

If $X = \mathrm{pt}$ then $T^*X = \mathrm{pt}$, bundles are vector spaces, and an "elliptic operator" is simply a linear map $A : \mathbb{C}^a \to \mathbb{C}^b$ (every map is elliptic, since there are no $k \neq 0$ to check). Its symbol class is the complex $0 \to \mathbb{C}^a \xrightarrow{A} \mathbb{C}^b \to 0$, whose class in $K(\mathrm{pt}) = \mathbb{Z}$ is $a - b$ by Example 4.12.

Then

$$\operatorname{ind}_a(a-b) \overset{(1)}{=} \operatorname{index} A \overset{(2)}{=} a-b.$$

(1) Definition (4.40).
(2) Finite-dimensional index (4.2).

So $\operatorname{ind}_a : K(\mathrm{pt}) \to \mathbb{Z}$ is the identity. This "normalisation" is one of the axioms that pins down the index uniquely (Section 4.7).

Worked example 4.23: The circle: $K(T^*S^1) \cong \mathbb{Z}$ and winding numbers

Let $X = S^1$, so $T^*S^1 = S^1 \times \mathbb{R}$ is an infinite cylinder with coordinates (x, k). A symbol of order 0 on the trivial line bundle is a function $\sigma(x, k)$, homogeneous of degree 0 in k, hence determined by its values on the two circles $k = +1$ and $k = -1$. Ellipticity says both loops

$$x \mapsto \sigma(x, +1), \qquad x \mapsto \sigma(x, -1)$$

avoid 0, so each has a winding number, w_+ and w_- (for matrix symbols, the winding numbers of $\det \sigma$). Two observations:

- Multiplying σ by a k-independent invertible function $g(x)$ (the symbol of the invertible operator "multiply by g", which has index 0) shifts w_+ and w_- by the same amount. So only the difference $w_+ - w_-$ can matter.
- The Toeplitz-type operator A_f of (4.12) has $w_+ = \operatorname{wind}(f)$, $w_- = 0$ and index $-\operatorname{wind}(f)$.

Hence, for every elliptic symbol on the circle,

$$\operatorname{ind}_a[\sigma] = -(w_+ - w_-). \tag{4.41}$$

Topologically: the one-point compactification of the cylinder $S^1 \times \mathbb{R}$ is a sphere with its north and south poles glued together, which has the same K-theory as S^2 with a circle attached, so $K(T^*S^1) = \widetilde{K}(S^2) \oplus \widetilde{K}(S^1) = \mathbb{Z} \oplus 0 = \mathbb{Z}$, and $w_+ - w_-$ is the isomorphism. Now look at a differential operator of order m such as $\mathrm{d}/\mathrm{d}x + a$. Its leading symbol is a polynomial homogeneous of degree m, so $\sigma(x, -k) = (-1)^m \sigma(x, k)$: the loop at $k = -1$ is the loop at $k = +1$ times the constant $(-1)^m$, and $w_- = w_+$. So *every differential operator on the circle has index zero*, and only pseudo-differential operators reach the nonzero classes. This is the simplest case of the vanishing theorem in odd dimensions proved in the next part.

Worked example 4.24: Additivity, and the class of the adjoint

On the circle take $D_1 = A_{z^2}$ (index -2) and $D_2 = A_{z^{-3}}$ (index 3). Then $D_1 \oplus D_2$ acts on $\mathbb{C}^2$-valued functions with diagonal symbol $\operatorname{diag}(\sigma_1, \sigma_2)$; the winding number of its determinant on $k = +1$ is $2 + (-3) = -1$, on $k = -1$ it is 0. So by (4.41) its index is $+1$, and by (4.39) it is $-2 + 3 = +1$. Consistent.
The adjoint D^* has symbol $\sigma^\dagger$, whose determinant is the complex conjugate of $\det \sigma$ and so winds the opposite way: $w_\pm \to -w_\pm$ and ind_a changes sign. That matches $\operatorname{index} D^* = \dim \operatorname{Ker} D^* - \dim \operatorname{Ker} D = -\operatorname{index} D$.

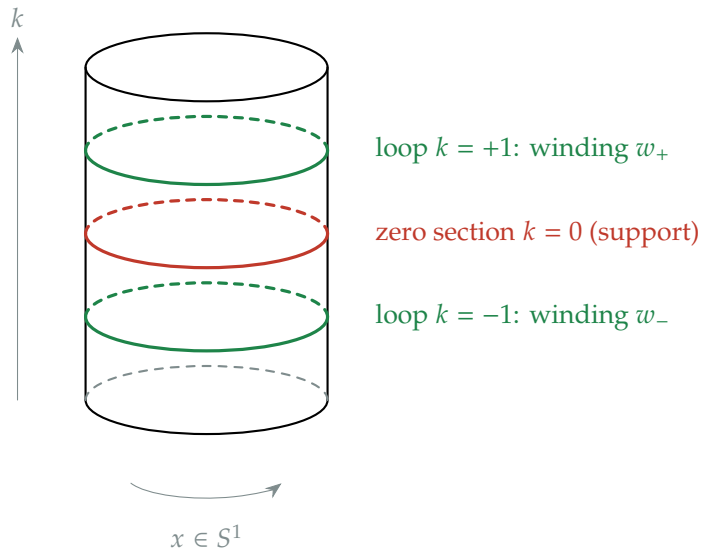


Figure 4.15: The phase space T^*S^1, drawn as a cylinder: the circle direction x runs around the cylinder and k runs up its axis. An elliptic symbol may vanish only on the red zero section. Its class is recorded by the winding numbers of the two loops $k = \pm 1$ (green); the index is $-(w_+ - w_-)$. For a differential operator the two loops differ by the sign $(-1)^m$, so $w_+ = w_-$ and the index vanishes.

Worked example 4.25: The de Rham class

For the de Rham complex on a closed X, the rolled-up symbol $\sigma(\mathrm{d} + \mathrm{d}^*)(x, k) = \mathrm{i}(k \wedge - \iota_k)$ (wedge with k plus its adjoint, contraction with the vector dual to k) gives a class $[\sigma_{\mathrm{dR}}] \in K(T^*X)$, and (4.36) says $\mathrm{ind}_a[\sigma_{\mathrm{dR}}] = \chi(X)$. For $X = S^2$ this is 2, for T^2 it is 0, for a surface of genus g it is $2 - 2g$. On the circle, $\sigma(\mathrm{d}) = \mathrm{i}k$ winds zero times on both $k = \pm 1$ (the loops are the constants $\pm\mathrm{i}$), so $w_+ = w_- = 0$ and the index is $0 = \chi(S^1)$, in agreement with (4.41).

Reading Nash (pp. 93–94, (4.17)–(4.18))

Nash: the symbol complex $0 \to \pi^*E^+ \to \pi^*E^- \to 0$ (4.17) gives an element $[\sigma_m(D)] \in K(T^*X)$; "two elliptic D with the same leading symbol have the same index"; so computing the index is a map from $K(T^*X)$ to $\mathbb{Z}$, the analytical index (4.18). He remarks that it is "tacit" that every element of $K(T^*X)$ arises as $[\sigma_m(D)]$, which Atiyah–Singer [2] guarantee once pseudo-differential operators are allowed. We added the three-step argument for well-definedness and the complete answer on a point and on a circle.

Recap

- Elliptic $D \Rightarrow$ symbol complex $0 \to \pi^*E^+ \to \pi^*E^- \to 0$ with compact support $\Rightarrow$ $[\sigma_m(D)] \in K(T^*X)$.
- Index is homotopy invariant, additive, and zero on everywhere-exact symbols, so it descends to $\mathrm{ind}_a : K(T^*X) \to \mathbb{Z}$.
- Point: $\mathrm{ind}_a = \mathrm{id}_{\mathbb{Z}}$. Circle: $K(T^*S^1) = \mathbb{Z}$, $\mathrm{ind}_a = -(w_+ - w_-)$; differential operators have $w_+ = w_-$ and index 0.
- Every class is realised by a pseudo-differential operator.

4.6 The topological index

The analytical index needs analysis: we must solve a PDE and count solutions. The question of this section is whether the same integer can be extracted from the class $[\sigma] \in K(T^*X)$ by *topology alone*, with no operator in sight. The idea is simple to state:

> Embed X in a big Euclidean space $\mathbb{R}^n$. Push the class $[\sigma]$ from T^*X into $T^*\mathbb{R}^n = \mathbb{R}^{2n}$. There the K-group is just $\mathbb{Z}$, so the pushed class *is* an integer. That integer is the topological index.

To push classes forward we need two tools, both of which Nash introduces in a few lines on pp. 94–95: the *Thom homomorphism*, which moves a class from a space into the total space of a vector bundle over it, and *extension by zero*, which moves a compactly supported class from an open subset into the whole space. And to connect the two we need *tubular neighbourhoods*. We explain all three, then assemble them.

4.6.1 Tubular neighbourhoods

Nash's picture: a curved piece of string S in $\mathbb{R}^3$. Thicken it into a tube U. The tube is a union of small discs, one through each point of the string and perpendicular to it. Each disc is a copy of a neighbourhood of 0 in the plane perpendicular to the string, i.e. in the **normal space** $N_s = (T_sS)^\perp$. So the tube looks like the **normal bundle** $N(S)$, the vector bundle over S whose fibre at s is N_s.

Working definition 4.19 (normal bundle and tubular neighbourhood)

For a submanifold $S \subset M$ of a Riemannian manifold, the **normal bundle** $N(S)$ has fibre $N_s = (T_sS)^\perp \subset T_sM$. A **tubular neighbourhood** is an open set $U \supset S$ with a diffeomorphism [a smooth bijection with smooth inverse] $N(S) \cong U$ (or from a disc bundle [the normal vectors $v \in N_s$ with $|v| < \epsilon$] in $N(S)$) sending the zero section [the vectors $v = 0$, one at each $s \in S$] to S. *Operational recipe* in $\mathbb{R}^n$: send (s, v), with $v \in N_s$ small, to the point $s + v$. For v small enough this is one-to-one (the normal discs do not collide).

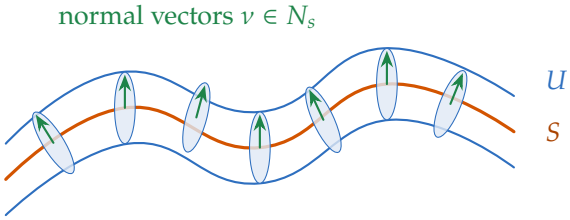


Figure 4.16: A string S (orange) in $\mathbb{R}^3$ and its tubular neighbourhood U (blue). Every point of the tube lies on exactly one small normal disc, centred at a point s of the string, so the tube is a copy of (a disc bundle in) the normal bundle $N(S)$. The green arrows are normal vectors.

Worked example 4.26: Three tubular neighbourhoods

(a) *The unit circle in* $\mathbb{R}^2$. At $s = (\cos\theta, \sin\theta)$ the normal line is spanned by $\hat{n} = (\cos\theta, \sin\theta)$ itself. The map $(\theta, r) \mapsto (1 + r)\hat{n}$, $|r| < 1$, identifies the trivial line bundle $S^1 \times (-1, 1)$ with the open annulus $0 < |x| < 2$. The normal bundle is trivial.

(b) *A point in $\mathbb{R}^n$.* $T_p\{p\} = 0$, so $N_p = \mathbb{R}^n$ and the tubular neighbourhood of a point is all of $\mathbb{R}^n$ (or a ball around it).
(c) *The central circle of a Möbius band.* Going once round the circle, the normal direction comes back reversed. The tubular neighbourhood is the Möbius band itself, and the normal bundle is the non-trivial real line bundle over S^1. So normal bundles carry topology, and that is why the construction below needs the full Thom homomorphism rather than a simple product.

4.6.2 The exterior complex and the Thom class

Let $F \to X$ be a complex vector bundle of rank r, with projection $p\colon F \to X$. Points of the total space are pairs (x, v) with $v \in F_x$. Over the total space F we have the pulled-back bundles $p^*\Lambda^i F$, whose fibre at (x, v) is the exterior power $\Lambda^i F_x$. At (x, v), wedge with v itself:

$$\Lambda(F)\colon \quad 0 \to \Lambda^0 F \xrightarrow{v\wedge} \Lambda^1 F \xrightarrow{v\wedge} \cdots \xrightarrow{v\wedge} \Lambda^r F \to 0, \qquad \omega \mapsto v \wedge \omega. \tag{4.42}$$

This is Nash's **exterior complex** (4.20). It is a complex because $v \wedge v \wedge \omega = 0$. It is exact whenever $v \neq 0$, by exactly the argument of Example 4.17 (choose a basis of F_x with first vector v). At $v = 0$ every map is zero and exactness fails. So

$$\operatorname{supp} \Lambda(F) = \text{zero section} \cong X,$$

which is compact when X is. Its class $\lambda_F := [\Lambda(F)] \in K(F)$ is the **Thom class** of F.

Worked example 4.27: The exterior complex in ranks one and two

Rank one, $F = X \times \mathbb{C}$. $\Lambda^0 = \mathbb{C}$, $\Lambda^1 = \mathbb{C}$, and $v\wedge$ is multiplication by the number v:

$$0 \to \mathbb{C} \xrightarrow{v} \mathbb{C} \to 0.$$

Over a point this is precisely the Bott complex (4.17). So *the Thom class of $\mathbb{C}$ over a point is the Bott class.*
Rank two, $F = X \times \mathbb{C}^2$. Use the basis 1; e_1, e_2; $e_1 \wedge e_2$, and write $v = v_1 e_1 + v_2 e_2$. Then $v \wedge 1 = v_1 e_1 + v_2 e_2$ and

$$v \wedge (a e_1 + b e_2) \overset{(1)}{=} v_1 b\, e_1 \wedge e_2 + v_2 a\, e_2 \wedge e_1 \overset{(2)}{=} (-v_2 a + v_1 b)\, e_1 \wedge e_2,$$

(1) $e_1 \wedge e_1 = e_2 \wedge e_2 = 0$.
(2) $e_2 \wedge e_1 = -e_1 \wedge e_2$.
so the complex is

$$0 \to \mathbb{C} \xrightarrow{\binom{v_1}{v_2}} \mathbb{C}^2 \xrightarrow{(-v_2\ v_1)} \mathbb{C} \to 0.$$

The composite is $-v_2 v_1 + v_1 v_2 = 0$. For $v \neq 0$ both maps have rank 1, and $1+1 = 2 = \dim \mathbb{C}^2$, so it is exact. Now compare with the tensor product formula (4.19) for two Bott complexes, with $a = v_1$ and $b = v_2$: the maps are $\binom{a}{b}$ and $(-b\ a)$. Identical. So

$$\lambda_{\mathbb{C}^2} = \beta \boxtimes \beta, \qquad \text{and in general} \qquad \lambda_{\mathbb{C}^r} = \beta \boxtimes \cdots \boxtimes \beta \ (r \text{ factors}),$$

where $\boxtimes$ is the product of classes pulled back from the separate factors. The Thom class of a trivial bundle is a product of Bott classes, one for each complex direction.

Watch out: Nash's $k = k_i\,\mathrm{d}k^i$

On p. 94 Nash writes the exactness argument as "if v is an i-form and $k = k_i\,\mathrm{d}k^i \neq 0$, then $k \wedge v = 0 \Rightarrow v = k \wedge w$". He is recycling the de Rham notation. In the exterior complex the thing we wedge with is the point $v \in F_x$ of the fibre, not a covector on X, and the forms are elements of $\Lambda^\bullet F_x$, not differential forms. The linear algebra is the same.

4.6.3 The Thom homomorphism

Working definition 4.20 (Thom homomorphism, Nash (4.19), (4.21))

For a complex vector bundle $p\colon F \to X$,

$$\phi\colon K(X) \to K(F), \qquad \phi(a) = p^*a \cdot \lambda_F. \tag{4.43}$$

In words: pull the class a back to the total space [F itself, the set of all vectors in all fibres F_x] (constant along the fibres), then multiply by the Thom class [the class of the exterior complex, exact wherever the fibre vector $v \neq 0$], which cuts it down to a neighbourhood of the zero section.

The product has compact support even though p^*a does not, because the support of a product of complexes is the intersection of the supports (Section 4.3.5), and λ_F is supported on the compact zero section. For non-compact X (such as X replaced by T^*X, which is how we will use it) the same formula works with compactly supported classes a (Atiyah–Singer [2]).

Key idea: Thom isomorphism and Bott periodicity

$\phi\colon K(X) \to K(F)$ is an *isomorphism*. For X a point and $F = \mathbb{C}^r$ this says

$$\mathbb{Z} \overset{(1)}{=} K(\mathrm{pt}) \xrightarrow{\phi} K(\mathbb{C}^r) \overset{(2)}{=} K(\mathbb{R}^{2r}) \overset{(3)}{=} \widetilde{K}(S^{2r}), \tag{4.44}$$

(1) Example 4.11.
(2) $\mathbb{C}^r = \mathbb{R}^{2r}$ as a space.
(3) Compact supports = reduced K-theory of the one-point compactification, and $(\mathbb{R}^{2r})^+ = S^{2r}$ (stereographic projection).

i.e. $\widetilde{K}(S^{2r}) \cong \mathbb{Z}$, generated by $\phi(1) = \lambda_{\mathbb{C}^r} = \beta^{\boxtimes r}$. This is Bott periodicity.

Physical picture: The Thom class is a K-theoretic delta function

In de Rham cohomology the analogue of λ_F is the **Thom form**: a closed form on F, concentrated near the zero section (a Gaussian bump in the fibre directions, times the fibre volume form), whose integral over each fibre is 1. Multiplying by it is "multiplying by $\delta(v)$ in the normal directions". The Thom homomorphism takes a field on X and turns it into a field on the bigger space F localised on X, like a field living on a brane. The K-theory Thom class does the same job, with the Bott complex playing the role of the normalised bump: in each complex fibre direction it contributes one unit of winding.

4.6.4 Extension by zero

Let $i\colon S \hookrightarrow M$ be the inclusion of an *open* subset. Nash (p. 95) first explains the idea with differential forms: a p-form ω on S "can be extended to the whole of M by letting it be zero

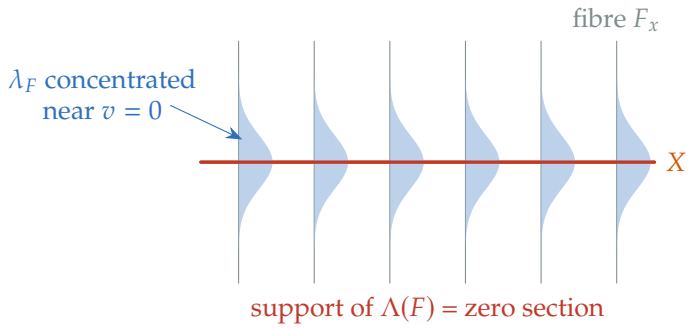


Figure 4.17: The Thom class of a vector bundle $F \to X$, drawn as its de Rham cousin: a bump (blue) in each fibre, centred on the zero section. The exterior complex $\Lambda(F)$ is exact wherever $v \neq 0$, so as a K-class it lives only on the zero section (red). Multiplying by it transplants a class on X into a compactly supported class on F.

outside S".

Worked example 4.28: A bump form on an arc of the circle

Let $S = (a, b) \subset S^1 = M$ be an open arc, and let $\omega = h(x)\,\mathrm{d}x$ with $h \geq 0$ a smooth bump supported inside $[a + \varepsilon, b - \varepsilon]$ and $\int_a^b h\,\mathrm{d}x = 1$. Extended by zero, ω is a smooth 1-form on the whole circle with

$$\int_{S^1} i_*\omega \overset{(1)}{=} \int_a^b h(x)\,\mathrm{d}x \overset{(2)}{=} 1,$$

(1) $i_*\omega$ vanishes outside (a, b).
(2) Normalisation of the bump.

So $i_*\omega$ represents the generator of $H^1_{\mathrm{dR}}(S^1)$, although on the arc itself $\omega = \mathrm{d}\big(\int_a^x h\big)$ is exact. The extension is smooth only because ω vanishes near the ends of the arc.

Watch out: Extension by zero needs compact supports

Nash writes the induced map as $H^p_{\text{de Rham}}(S) \to H^p_{\text{de Rham}}(M)$. As the example shows, it is only defined for forms that vanish near the edge of S, i.e. for compactly supported cohomology: $i_*\colon H^p_c(S) \to H^p(M)$ (or $\to H^p_c(M)$). A general form on S (say $\mathrm{d}x/(x - a)$) does not extend. The same caveat applies in K-theory, and it is harmless there because all our classes have compact support.

Working definition 4.21 (extension by zero in K-theory, Nash (4.22))

For an open inclusion $i\colon S \hookrightarrow M$ [S is an open subset of M and i is the identity on it],

$$i_*\colon K(S) \to K(M)$$

(compactly supported K-theory [classes of complexes that are exact outside some compact set] on both sides) sends the class of a complex E on S with compact support $C \subset S$ to the class of a complex on M that agrees with E near C and is exact everywhere else. *How:* outside C the complex is exact, so its bundles can be identified with each other through the maps (as in the clutching of Example 4.13); this identification lets us continue the complex across the edge of S as a trivially exact complex.

Worked example 4.29: The Bott class on the plane, the sphere and the torus

Identify $\mathbb{C}$ with S^2 minus the north pole by stereographic projection. The Bott complex on $\mathbb{C}$ extends by zero to the class $\mathbf{1} - [L] \in \widetilde{K}(S^2)$ found in (4.18): this is i_* for $\mathbb{C} \hookrightarrow S^2$. Equally, a small open disc in the torus T^2 is a copy of $\mathbb{C}$, and extending the Bott class from it gives $\mathbf{1} - [L_{T^2}]$, where L_{T^2} is a line bundle of degree one on the torus. In both cases the result is "one unit of winding, localised at a point". The first Chern number of the resulting bundle counts the units: this is the K-theory version of the bump form above.

4.6.5 Assembling the topological index

Now we put the pieces together. Let X be compact of dimension d, and choose an embedding $X \subset \mathbb{R}^n$ (Whitney's theorem guarantees one for n large enough). Let $N = N(X)$ be its normal bundle, of real rank $n - d$.

Step 1: the phase space of X sits inside the phase space of $\mathbb{R}^n$. Using the Euclidean metric to identify covectors with vectors, $T^*X \subset T^*\mathbb{R}^n = \mathbb{R}^n \times \mathbb{R}^n$: a point (x, k) of T^*X is a position on X and a momentum tangent to X.

Step 2: its tubular neighbourhood. At (x, k) we may move off T^*X in two independent kinds of normal direction: shift the position along N_x, or add to the momentum a vector in N_x. So the normal bundle of T^*X in $T^*\mathbb{R}^n$ is $\pi^*(N \oplus N)$, the pull-back to T^*X of two copies of N. Nash's trick is to regard the pair $(a, b) \in N_x \oplus N_x$ as the single complex vector $a + \mathrm{i}b$:

$$N(T^*X) \;\cong\; \pi^*(N \oplus N) \;\cong\; \pi^*(N \oplus \mathrm{i}N) \;=\; \pi^*(N \otimes_{\mathbb{R}} \mathbb{C}) \;=:\; F. \tag{4.45}$$

So the tubular neighbourhood of T^*X is a *complex* vector bundle F of complex rank $n - d$. That is what makes a K-theory Thom homomorphism available. (This is the physicist's familiar $z = q + \mathrm{i}p$: a normal position and its conjugate normal momentum combine into one complex coordinate.) Dimension check:

$$\dim_{\mathbb{R}} F \overset{(1)}{=} \dim_{\mathbb{R}} T^*X + 2\,\mathrm{rank}_{\mathbb{C}}\, F \overset{(2)}{=} 2d + 2(n-d) = 2n = \dim_{\mathbb{R}} T^*\mathbb{R}^n. \tag{4.46}$$

(1) Base plus fibre; a complex fibre of rank r has real dimension $2r$.
(2) $\dim T^*X = 2d$, and $\mathrm{rank}_{\mathbb{C}}\, F = \mathrm{rank}_{\mathbb{R}}\, N = n - d$.

As it should be: F is (identified with) an open subset of $T^*\mathbb{R}^n$.

Step 3: the maps. Thom homomorphism for $F \to T^*X$ (Nash (4.23)), then extension by zero for the open inclusion $F \subset T^*\mathbb{R}^n$ (Nash (4.24)), composed (Nash (4.25)):

$$i_!\colon K(T^*X) \xrightarrow{\phi} K(F) \xrightarrow{i_*} K(T^*\mathbb{R}^n), \qquad i_! = i_* \circ \phi. \tag{4.47}$$

Step 4: compare with a point. Let $P = \{x_0\}$ be the origin of $\mathbb{R}^n$ and $j\colon P \to \mathbb{R}^n$ its inclusion (Nash (4.26)). Running the same construction for P: $T^*P = P$, the normal bundle of P is $\mathbb{R}^n$, and $N \otimes \mathbb{C} = \mathbb{C}^n$, whose total space is all of $T^*\mathbb{R}^n = \mathbb{R}^{2n}$. So $j_! = \phi_{\mathbb{C}^n}\colon K(\mathrm{pt}) \to K(\mathbb{C}^n)$ is the Thom isomorphism of a point, i.e. Bott periodicity (4.44).

Step 5: the K-groups at the ends. (Nash (4.29).)

$$K(T^*P) \overset{(1)}{=} K(P) \overset{(2)}{=} \mathbb{Z}, \qquad K(T^*\mathbb{R}^n) \overset{(3)}{=} K(\mathbb{R}^n \times \mathbb{R}^n) = K(\mathbb{R}^{2n}) \overset{(4)}{=} \widetilde{K}(S^{2n}) \overset{(5)}{=} \mathbb{Z}. \tag{4.48}$$

(1) The cotangent space of a point is the point.
(2) Example 4.11.
(3) $T^*\mathbb{R}^n$ is a bundle over a contractible base, hence trivial.

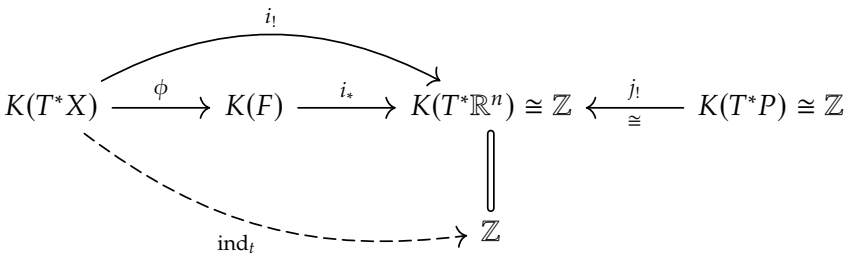


Figure 4.18: Nash's diagram (4.27), with the Thom homomorphism and the extension by zero written out. A symbol class on T^*X is transplanted into a tubular neighbourhood F of T^*X in $T^*\mathbb{R}^n$ by ϕ, then extended by zero to all of $T^*\mathbb{R}^n$. There it is compared, via the Bott isomorphism $j_!$, with the K-theory of a point. The result is the integer $\mathrm{ind}_t[\sigma]$.

(4) Compact supports and one-point compactification.
(5) Bott periodicity.

Working definition 4.22 (topological index, Nash (4.28), (4.30))

$$\mathrm{ind}_t = (j_!)^{-1} \circ i_! \colon K(T^*X) \to K(T^*P) = \mathbb{Z}, \qquad [\sigma_m(D)] \longmapsto (j_!)^{-1}\big(i_!([\sigma_m(D)])\big). \tag{4.49}$$

[$i_!$ is the Thom homomorphism followed by extension by zero, from T^*X into $T^*\mathbb{R}^n$. $j_!$ does the same for a point $P \in \mathbb{R}^n$, and is an isomorphism] *In words:* transplant the symbol class into Euclidean phase space, where every class is an integer multiple of the Bott class [$j_!(1)$, the Thom class of $T^*\mathbb{R}^n = \mathbb{C}^n$ over the point P], and read off the multiple.

Two remarks from Nash (p. 97). First, the construction used an embedding and a tubular neighbourhood, so one must check that the answer does not depend on these choices (it does not; Atiyah–Singer [2]). Any two embeddings become isotopic after enlarging n, and isotopic choices give homotopic complexes. Second, $j_!$ is an isomorphism $\mathbb{Z} \to \mathbb{Z}$ sending generator to generator, so it is "the identity", and we could omit it. It is kept because in the equivariant theory of §6, where a group G acts and $K(\mathrm{pt})$ is replaced by the representation ring $R(G)$, it is no longer trivial.

Worked example 4.30: The topological index of a point

Take $X = P$ itself. Then $i = j$, so $i_! = j_!$ and

$$\mathrm{ind}_t \overset{(1)}{=} (j_!)^{-1} \circ j_! = \mathrm{id}_{\mathbb{Z}} \overset{(2)}{=} \mathrm{ind}_a.$$

(1) Definition (4.49) with $i = j$.
(2) Example 4.22.

The two indices agree on a point. This is the "normalisation axiom" of Section 4.7.

Worked example 4.31: The circle in the plane: the construction in coordinates

Let $X = S^1 \subset \mathbb{R}^2$ ($d = 1$, $n = 2$), with unit tangent $\hat{t} = (-\sin\theta, \cos\theta)$ and unit normal $\hat{n} = (\cos\theta, \sin\theta)$. Then:

- $T^*S^1 \subset T^*\mathbb{R}^2 = \mathbb{R}^4$ consists of the points $(\hat{n},\ k\,\hat{t})$: position on the unit circle, momentum tangent to it. Coordinates (θ, k).

- N is the trivial real line bundle spanned by $\hat{n}$, so $F = \pi^*(N \otimes \mathbb{C})$ is the trivial complex line bundle over the cylinder, with fibre coordinate $w = r + ip$.
- The tubular neighbourhood map is

$$(\theta, k, w) \longmapsto \Big(\underbrace{(1+r)\,\hat{n}}_{\text{position}},\ \underbrace{k\,\hat{t} + p\,\hat{n}}_{\text{momentum}} \Big) \in \mathbb{R}^4, \qquad |r| < 1.$$

  It is one-to-one: from the position we read off θ and r, and from the momentum, decomposed along $\hat{t}$ and $\hat{n}$, we read off k and p. Its image is (annulus)$\times\mathbb{R}^2$, open in $\mathbb{R}^4$.
- The Thom class of F is the rank-one exterior complex $0 \to \mathbb{C} \xrightarrow{w} \mathbb{C} \to 0$. For a symbol class $[\sigma]$, $\phi[\sigma] = [\sigma] \cdot [w]$, the tensor product of the symbol complex with the complex "multiply by w". Its support is $\{k = 0\} \cap \{w = 0\}$, the circle $\{(\hat{n}, 0)\}$ inside $\mathbb{R}^4$: compact, so i_* can extend it to $K(\mathbb{R}^4) = \widetilde{K}(S^4) = \mathbb{Z}$.

The resulting integer is $\mathrm{ind}_t[\sigma]$. The Atiyah–Singer theorem (next section) says it equals $-(w_+ - w_-)$ from (4.41). Doing that last identification by hand is possible but not illuminating; its content is that the class in $\widetilde{K}(S^4)$ measures how many times the combined map (σ, w) "wraps" around its zero set, and that wrapping is the relative winding of the symbol on the two halves of T^*S^1.

Reading Nash (pp. 94–97, (4.19)–(4.30))

Nash introduces the tubular neighbourhood with the string-in-$\mathbb{R}^3$ picture, defines the Thom homomorphism $\phi\colon K(X) \to K(F)$ (4.19) using the exterior complex (4.20) and its class λ_F, as multiplication by λ_F (4.21), and notes that it is an isomorphism which for a point is Bott periodicity $\widetilde{K}(S^{2n}) = \mathbb{Z}$. Then comes extension by zero $i_*\colon K(S) \to K(M)$ for open $S \subset M$ (4.22), motivated by differential forms. The embedding $X \subset \mathbb{R}^n$ gives the complex bundle $F = \pi^*(N(X) \otimes_{\mathbb{R}} \mathbb{C})$ (4.23), the maps i_* (4.24), $i_! = i_* \circ \phi$ (4.25), $j\colon P \to \mathbb{R}^n$ and $j_!$ (4.26), the diagram (4.27), and $\mathrm{ind}_t = (j_!)^{-1} \circ i_!$ (4.28), (4.30), with $K(T^*P) = K(T^*\mathbb{R}^n) = \mathbb{Z}$ (4.29). We added the explicit exterior complexes, their identification with products of Bott classes, the bump-form example, the dimension check (4.46) and the circle in coordinates.

Recap

- Tubular neighbourhood of $S \subset M \cong$ normal bundle $N(S)$.
- Exterior complex $v\wedge\colon \Lambda^i F \to \Lambda^{i+1} F$, exact for $v \neq 0$; its class λ_F is the Thom class; for $F = \mathbb{C}^r$ it is the product of r Bott classes.
- Thom homomorphism $\phi(a) = p^*a \cdot \lambda_F$ is an isomorphism; for a point it is Bott periodicity.
- Extension by zero i_* for open inclusions (compact supports).
- $T^*X \subset T^*\mathbb{R}^n$ has tubular neighbourhood $F = \pi^*(N \otimes \mathbb{C})$ ($q + ip$ in the normal directions); $\mathrm{ind}_t = (j_!)^{-1} \circ i_* \circ \phi\colon K(T^*X) \to \mathbb{Z}$.

4.7 The Atiyah–Singer index theorem

We now have two maps from the same group to the integers:

$$\mathrm{ind}_a,\ \mathrm{ind}_t \colon K(T^*X) \longrightarrow \mathbb{Z}.$$

The first counts zero modes. To evaluate it we must solve a partial differential equation on the whole of X, which is a global and usually hopeless problem. The second pushes the symbol class into Euclidean phase space and reads off a multiple of the Bott class. To evaluate it we need only the symbol, and only up to deformation. The question of this section is the obvious one: *are they the same map?* The answer, yes, is the Atiyah–Singer index theorem. We state it, test it on every example where both sides can be computed by hand, and then explain the two strategies by which it is proved.

4.7.1 The statement

Key idea: The Atiyah–Singer index theorem (Nash p. 97)

For every compact manifold X without boundary,

$$\mathrm{ind}_t = \mathrm{ind}_a \colon K(T^*X) \to \mathbb{Z}. \tag{4.50}$$

Concretely: for every elliptic (pseudo-)differential operator, or elliptic complex, D on X,

$$\underbrace{\dim \operatorname{Ker} D - \dim \operatorname{Ker} D^*}_{\text{analysis: count solutions}} = \underbrace{(j_!)^{-1}\, i_! \,[\sigma_m(D)]}_{\text{topology: symbol class only}} .$$

It is worth saying slowly why this is surprising. The left side depends, a priori, on everything: the coefficients of D at every point, the metric used to define D^*, the global shape of X. The right side depends only on

- the leading symbol, i.e. how D acts on very short wavelengths,
- and only on its homotopy class through elliptic symbols,
- together with the topology of X (through the embedding and the normal bundle).

We already knew from Sections 4.2 and 4.5 that the index is *insensitive* to lower-order terms and to deformations. The theorem goes much further: it says the index can be *computed* from the symbol alone. Short-wavelength data determines a count of zero modes, which are the longest-wavelength states there are.

Physical picture: A conserved charge computed from its flux

Think of Gauss's law. The total charge inside a closed surface is a global quantity, yet it can be read off from the flux of $\vec{E}$ through any surface enclosing it, and it does not change as we deform the surface. The index behaves the same way. The symbol class lives near "infinity in momentum space" (on the sphere bundle $|k| = 1$, where it is invertible), and it measures how many zero modes are trapped inside, just as the flux at infinity measures the charge. In the Toeplitz example this is literal: the index is a winding number $\frac{1}{2\pi \mathrm{i}} \oint \mathrm{d}f/f$, a flux integral of the symbol around the circle at large $|k|$.

In §2 of Nash (the next part of these notes) the right-hand side is turned, via the Chern character, into an integral of characteristic classes:

$$\operatorname{index} D = (-1)^n \big\{ \operatorname{ch}[\sigma_m(D)] \operatorname{td}(TX \otimes \mathbb{C}) \big\}[TX].$$

That is the form in which physicists use the theorem. The K-theory version (4.50) is the sharper one: it is an equality of integers with no denominators, and it survives in situations (families, real operators, group actions) where the cohomological formula loses information.

4.7.2 Checking the theorem where both sides are computable

Before looking at proofs, let us convince ourselves the theorem is true. We run through examples where the analytic side has been computed in earlier sections and the topological side can be found independently. For the topological side we will sometimes use the cohomological form of the theorem (proved in the next part) as a stand-in for $(j_!)^{-1} i_!$; this is honest, since the cohomological form is derived from ind_t alone, without any analysis.

Worked example 4.32: The point, once more

For $X = \mathrm{pt}$ we found $\operatorname{ind}_a = \mathrm{id}_{\mathbb{Z}}$ (Example 4.22) and $\operatorname{ind}_t = (j_!)^{-1} j_! = \mathrm{id}_{\mathbb{Z}}$ (Example 4.30). The theorem on a point is rank–nullity, (4.2): $\dim \operatorname{Ker} A - \dim \operatorname{Coker} A = \dim V - \dim W$. The left side needs the matrix; the right side needs only the "symbol class" $[V]-[W] \in K(\mathrm{pt})$, i.e. the sizes of the two spaces. This is the whole theorem in miniature.

Worked example 4.33: The de Rham operator on a closed surface: Gauss–Bonnet

Let $X = \Sigma_g$ be a closed orientable surface of genus g and $D = \mathrm{d} + \mathrm{d}^* \colon \Omega^0 \oplus \Omega^2 \to \Omega^1$.
Analytic side. By (4.36) and the Hodge theorem, $\operatorname{index} D = b_0 - b_1 + b_2 = 1 - 2g + 1 = 2 - 2g$. (For S^2 and T^2 we found the kernels by hand in Examples 4.19 and 4.21.)
Topological side. For the de Rham symbol class the cohomological index formula (next part, Nash (4.53)–(4.54)) reduces to the integral of the Euler class, which in two dimensions is the Gaussian curvature K times the area form divided by 2π:

$$\operatorname{ind}_t[\sigma_{\mathrm{dR}}] \overset{(1)}{=} \int_{\Sigma_g} e(TX) \overset{(2)}{=} \frac{1}{2\pi} \int_{\Sigma_g} K \, \mathrm{d}A.$$

(1) Cohomological form of ind_t for the de Rham complex.
(2) The Euler form of a surface; see Chapter T2 for Gauss–Bonnet.

For the round sphere of radius R, $K = 1/R^2$ and the area is $4\pi R^2$, so

$$\frac{1}{2\pi} \int_{S^2} K \, \mathrm{d}A \overset{(1)}{=} \frac{1}{2\pi} \cdot \frac{1}{R^2} \cdot 4\pi R^2 = 2 = \chi(S^2).$$

(1) K is constant, so the integral is K times the area.

For the flat torus $K = 0$ and the integral is $0 = \chi(T^2)$. For genus $g \geq 2$ with a hyperbolic metric ($K = -1$) the area must be $4\pi(g-1)$ for the two sides to agree, and it is (a hyperbolic surface of genus g is glued from a $4g$-gon with interior angles summing to 2π, whose area by the hyperbolic Gauss–Bonnet formula is $(4g-2)\pi - 2\pi$). Gauss–Bonnet *is* the Atiyah–Singer theorem for $\mathrm{d} + \mathrm{d}^*$ in two dimensions.

Worked example 4.34: Riemann–Roch on the Riemann sphere

Let $X = \mathbb{CP}^1 = \mathbb{C} \cup \{\infty\}$ and let L_n be the line bundle $\mathcal{O}(n)$: a section is a function $f(z)$ on the z-chart and a function $\tilde f(w)$ on the $w = 1/z$ chart, related on the overlap by $\tilde f(w) = w^n f(1/w)$. Let $D = \bar\partial_{L_n}$, the Cauchy–Riemann operator on sections of L_n, whose kernel is the holomorphic sections.

Analytic side, $n \geq 0$. A holomorphic section is an entire function $f(z)$ such that $w^n f(1/w)$ is holomorphic at $w = 0$, i.e. $f(z) = O(|z|^n)$ as $z \to \infty$. By Liouville's theorem f is a polynomial of degree at most n, so

$$\operatorname{Ker} \bar\partial_{L_n} = \operatorname{span}\{1, z, z^2, \dots, z^n\}, \qquad \dim \operatorname{Ker} \bar\partial_{L_n} = n + 1.$$

The cokernel is $H^{0,1}(L_n)$, which vanishes for $n \geq -1$ (by Serre duality it is dual to the holomorphic sections of L_{-n-2}, and a negative-degree bundle has none). So index $\bar\partial_{L_n} = n+1$.

Analytic side, $n \leq -2$. Now f must be entire and decay like $|z|^n$, so $f = 0$: no kernel. The cokernel is dual to the sections of L_{-n-2}, of dimension $-n-1$. So index $\bar\partial_{L_n} = 0 - (-n-1) = n+1$ again.

Topological side. The cohomological formula for twisted Dolbeault operators is the Riemann–Roch theorem (next part, and Chapter 5):

$$\begin{aligned}\operatorname{ind}_t[\sigma(\bar\partial_{L_n})] &\overset{(1)}{=} \int_{\mathbb{CP}^1} \operatorname{ch}(L_n)\,\operatorname{td}(\mathbb{CP}^1) \overset{(2)}{=} \int_{\mathbb{CP}^1} \big(1 + c_1(L_n)\big)\big(1 + \tfrac12 c_1(T\mathbb{CP}^1)\big) \\ &\overset{(3)}{=} \int_{\mathbb{CP}^1} c_1(L_n) + \tfrac12 \int_{\mathbb{CP}^1} c_1(T\mathbb{CP}^1) \overset{(4)}{=} n + \tfrac12 \cdot 2 = n + 1.\end{aligned}$$

(1) Riemann–Roch form of the index formula.
(2) Expand ch $= 1 + c_1 + \dots$ and td $= 1 + \frac12 c_1 + \dots$; on a surface only two-forms survive integration.
(3) Keep the degree-two part.
(4) $\int c_1(L_n) = n$ (degree of $\mathcal{O}(n)$; its transition function $w^n = z^{-n}$ winds n times), and $\int c_1(T\mathbb{CP}^1) = \chi(\mathbb{CP}^1) = 2$.

Analytic and topological sides agree for every n. Note how the analytic side changed character at $n = -1$ (kernel for $n \geq 0$, cokernel for $n \leq -2$), while the topological side is one linear formula: this is the index doing its job of ignoring which of the two spaces the zero modes live in.

Worked example 4.35: The Toeplitz operator on the circle

Here both sides were found in the previous sections: the analytic index is index $A_f = -\operatorname{wind}(f)$ ((4.11)), and the K-group $K(T^*S^1) \cong \mathbb{Z}$ is detected by $-(w_+ - w_-)$ ((4.41)). The Atiyah–Singer theorem on S^1 is the statement that the topological construction of Example 4.31 produces this same integer. Historically this case came first: it is the Noether–Gohberg–Krein theorem (1950s), and Gel'fand's question of 1960, "can the index of every elliptic operator be expressed through its symbol by a topological formula?", is exactly the question the Atiyah–Singer theorem answered.

4.7.3 The axiomatic proof: why the topological index is forced

Nash describes the proof in one paragraph (p. 97): the topological index is characterised as the *unique* map satisfying certain axioms; the analytical index is then shown, "after considerable amount of work", to satisfy the same axioms. We now make this concrete. The uniqueness half

Operator on X	analytic side	topological side	value
linear map $\mathbb{C}^a \to \mathbb{C}^b$ on a point	$\dim\ker - \dim\operatorname{coker}$	$[\mathbb{C}^a] - [\mathbb{C}^b]$	$a - b$
Toeplitz A_f on S^1	kernel of T_f in H^2	$-(w_+ - w_-)$	$-\operatorname{wind} f$
$\mathrm{d} + \mathrm{d}^*$ on Σ_g	$b_0 - b_1 + b_2$	$\frac{1}{2\pi}\int K\,\mathrm{d}A$	$2 - 2g$
$\bar\partial_{\mathcal{O}(n)}$ on $\mathbb{CP}^1$	polynomials of degree $\le n$	$\int \operatorname{ch}\operatorname{td}$	$n + 1$
$\mathrm{d} + \mathrm{d}^*$ on X^{2m}	$\sum(-1)^i b_i$	$\int e(TX)$	$\chi(X)$
$\bar\partial$ on complex X	$\sum(-1)^q h^{0,q}$	$\int \operatorname{td}(X)$	arithmetic genus

Table 4.1: The Atiyah–Singer theorem in the classical cases. Each row equates a count of solutions (second column) with a symbol-and-topology computation (third column). The last three rows use the cohomological form of the topological index derived in §2 of Nash; the classical theorems they reproduce are Gauss–Bonnet, Riemann–Roch, Chern–Gauss–Bonnet and Hirzebruch–Riemann–Roch.

is short and beautiful, and we give it in full. The existence half is where the analysis lives, and we explain what has to be checked and why each check is plausible.

Index functions and the axioms

Suppose that for every compact manifold X we are given a homomorphism

$$\operatorname{ind}^X \colon K(T^*X) \to \mathbb{Z}.$$

Call this family an **index function** if it satisfies:

(A0) *Naturality.* A diffeomorphism $f \colon X \to X'$ induces $f^* \colon K(T^*X') \to K(T^*X)$, and $\operatorname{ind}^X \circ f^* = \operatorname{ind}^{X'}$. (Relabelling the points of X changes nothing.)

(A1) *Normalisation.* For $X = \mathrm{pt}$, $\operatorname{ind}^{\mathrm{pt}} \colon K(\mathrm{pt}) = \mathbb{Z} \to \mathbb{Z}$ is the identity.

(A2) *Embedding axiom.* For every embedding $i \colon X \hookrightarrow Y$ of compact manifolds, with $i_! = i_* \circ \phi \colon K(T^*X) \to K(T^*Y)$ built from a tubular neighbourhood as in Section 4.6.5,

$$\operatorname{ind}^Y \circ i_! = \operatorname{ind}^X. \tag{4.51}$$

Axiom (A2) says that transplanting a symbol class from X to a neighbourhood of X inside a bigger manifold, using the Thom class in the normal directions, does not change its index. We already met the simplest instance: for $X = \mathrm{pt}$ inside $Y = \mathbb{R}^n$ (or its compactification), (A2) says the Bott class of $\mathbb{R}^{2n} = T^*\mathbb{R}^n$ has index one.

Uniqueness

Key idea: Uniqueness of the index function

Any index function satisfying (A1) and (A2) equals ind_t.

Here is the argument. Embed X in $\mathbb{R}^n$ and compactify $\mathbb{R}^n$ to the sphere $S^n = \mathbb{R}^n \cup \{\infty\}$, with open inclusion $k \colon \mathbb{R}^n \hookrightarrow S^n$. Then $k \circ i \colon X \hookrightarrow S^n$ is an embedding of compact manifolds, and its shriek map is $(k \circ i)_! = k_* \circ i_!$ (extend the class from $T^*\mathbb{R}^n$ to T^*S^n by zero). Similarly, for the origin P, $(k \circ j)_! = k_* \circ j_!$. Let $u \in K(T^*X)$ and let $m = \operatorname{ind}_t(u) \in \mathbb{Z}$. By definition (4.49) of m,

$$i_!(u) = m \cdot j_!(1) \qquad \text{in } K(T^*\mathbb{R}^n) \cong \mathbb{Z}, \tag{4.52}$$

since $j_!(1)$ (the Bott class) generates $K(T^*\mathbb{R}^n)$. Now

$$\begin{aligned}\operatorname{ind}^X(u) &\overset{(1)}{=} \operatorname{ind}^{S^n}\big(k_*\, i_!(u)\big)\\ &\overset{(2)}{=} \operatorname{ind}^{S^n}\big(k_*\big(m\, j_!(1)\big)\big)\\ &\overset{(3)}{=} m\cdot \operatorname{ind}^{S^n}\big(k_*\, j_!(1)\big)\\ &\overset{(4)}{=} m\cdot \operatorname{ind}^{\mathrm{pt}}(1)\\ &\overset{(5)}{=} m \overset{(6)}{=} \operatorname{ind}_t(u).\end{aligned} \tag{4.53}$$

(1) Axiom (A2) for the embedding $k\circ i\colon X\hookrightarrow S^n$.
(2) (4.52).
(3) k_* and ind^{S^n} are homomorphisms of groups.
(4) Axiom (A2) for the embedding $k\circ j\colon \mathrm{pt}\hookrightarrow S^n$.
(5) Axiom (A1).
(6) Definition of m.

That is the whole proof of uniqueness. Look at what it uses: we route the class u through the sphere, where it becomes m copies of the class coming from a point, and the point is where the normalisation fixes everything. The embedding axiom is a statement that the index can be "transported" from X to the sphere and from the sphere to a point.

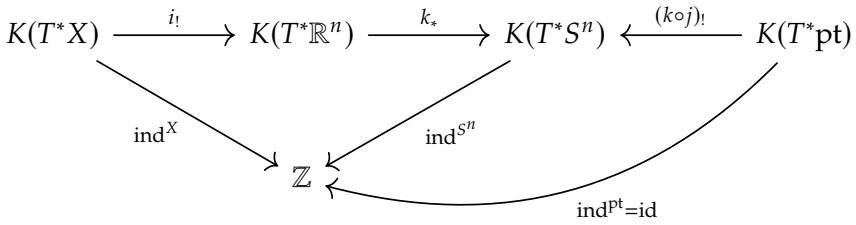


Figure 4.19: The uniqueness argument (4.53). By the embedding axiom both triangles commute. A class u on T^*X is carried to the sphere, where it equals m times the image of the generator from the point; the normalisation on the point then forces $\operatorname{ind}^X(u) = m = \operatorname{ind}_t(u)$.

The topological index satisfies the axioms

(A0) and (A1) are immediate from the definition (for a point, see Example 4.30). For (A2) one needs that shriek maps compose, $(i'\circ i)_! = i'_!\circ i_!$: embedding X in Y and then Y in $\mathbb{R}^n$ is the same as embedding X in $\mathbb{R}^n$, and the Thom classes multiply (the normal bundle of X in $\mathbb{R}^n$ is the normal bundle of X in Y plus the normal bundle of Y, restricted to X; the Thom class of a direct sum is the product of Thom classes, as we saw for $\lambda_{\mathbb{C}^2} = \beta\boxtimes\beta$ in Example 4.27). Then $\operatorname{ind}_t^Y(i_!u) = (j_!)^{-1}i'_!i_!u = (j_!)^{-1}(i'\circ i)_!u = \operatorname{ind}_t^X(u)$ (with $i'\colon Y\hookrightarrow\mathbb{R}^n$). The only real work is the independence of the choices (embedding, tubular neighbourhood), mentioned at the end of Section 4.6.5.

The analytical index satisfies the axioms

This is the "considerable amount of work". (A0) is clear (a diffeomorphism carries operators to operators and zero modes to zero modes) and (A1) is Example 4.22. For (A2), Atiyah and Singer split the embedding axiom into two simpler properties of the analytical index, each of which has a transparent meaning.

Working definition 4.23 (Excision)

Let $U \subset X$ be open, and let $u \in K(T^*U)$ have compact support [u is represented by a symbol complex that is exact outside a compact set of T^*U]. Then $\mathrm{ind}_a^X(i_* u)$, where i_* is extension by zero, depends only on u, not on the compact manifold X into which U is placed. *Analytic meaning:* an elliptic operator which, outside a compact set $C \subset U$, is simply an isomorphism ("multiplication by an invertible bundle map") has its index determined by what happens near C.

The physical content of excision: zero modes and obstructions can only live where the symbol fails to be "trivially invertible". Away from C the operator is an isomorphism and cannot support any unpaired modes, so we may cut X away from U and glue in any other manifold without changing the index. It is the same logic as the statement that the charge enclosed by a surface does not depend on what lies outside it.

Working definition 4.24 (Multiplicativity)

For compact X, Y and classes $u \in K(T^*X)$, $v \in K(T^*Y)$,

$$\mathrm{ind}_a^{X\times Y}(u \boxtimes v) = \mathrm{ind}_a^X(u) \cdot \mathrm{ind}_a^Y(v), \tag{4.54}$$

where $u \boxtimes v$ is the product of the pulled-back classes on $T^*(X \times Y) = T^*X \times T^*Y$ (the tensor product of symbol complexes (4.19)). *Analytic meaning:* there is an explicit "product operator" [built from $D_X \otimes 1$ and $1 \otimes D_Y$ together with their adjoints] on $X \times Y$, and its zero modes are products of zero modes [$\psi_X \otimes \psi_Y$, each factor a zero mode of D_X or D_X^*, and of D_Y or D_Y^*].

Both properties can be checked by hand in the key cases, and we do so now. Once they are known, (A2) follows: by excision the index of $i_! u$ only depends on a tubular neighbourhood of X in Y, which looks locally like $X \times \mathbb{R}^k$, and by multiplicativity the index of $u \boxtimes$ (Bott class on $\mathbb{R}^{2k}$) is $\mathrm{ind}(u) \cdot \mathrm{ind}(\mathrm{Bott}) = \mathrm{ind}(u) \cdot 1$. (When the normal bundle is not trivial, one needs a "twisted" version of multiplicativity, in which the second factor varies over X; this is the step in which Atiyah and Singer genuinely need their pseudo-differential machinery.)

Worked example 4.36: The product operator and multiplicativity

Let $D_1 \colon \Gamma(E_1) \to \Gamma(F_1)$ on X and $D_2 \colon \Gamma(E_2) \to \Gamma(F_2)$ on Y be elliptic, both of order one. On $X \times Y$ let D act from $(E_1 \boxtimes E_2) \oplus (F_1 \boxtimes F_2)$ to $(F_1 \boxtimes E_2) \oplus (E_1 \boxtimes F_2)$ by

$$D = \begin{pmatrix} D_1 \otimes 1 & -1 \otimes D_2^* \\ 1 \otimes D_2 & D_1^* \otimes 1 \end{pmatrix}, \qquad D^* = \begin{pmatrix} D_1^* \otimes 1 & 1 \otimes D_2^* \\ -1 \otimes D_2 & D_1 \otimes 1 \end{pmatrix}.$$

(Here $D_1 \otimes 1$ differentiates in the X variables only, and $1 \otimes D_2$ in the Y variables only; the two commute.) Then

$$(D^*D)_{11} \overset{(1)}{=} (D_1^* \otimes 1)(D_1 \otimes 1) + (1 \otimes D_2^*)(1 \otimes D_2) = D_1^*D_1 \otimes 1 + 1 \otimes D_2^*D_2,$$
$$(D^*D)_{12} \overset{(2)}{=} -(D_1^* \otimes 1)(1 \otimes D_2^*) + (1 \otimes D_2^*)(D_1^* \otimes 1) \overset{(3)}{=} -D_1^* \otimes D_2^* + D_1^* \otimes D_2^* = 0,$$
$$(D^*D)_{22} \overset{(4)}{=} (1 \otimes D_2)(1 \otimes D_2^*) + (D_1 \otimes 1)(D_1^* \otimes 1) = D_1D_1^* \otimes 1 + 1 \otimes D_2D_2^*.$$

(1) Row one of D^* times column one of D.
(2) Row one of D^* times column two of D.
(3) Operators acting on different variables commute: $(A\otimes 1)(1\otimes B) = (1\otimes B)(A\otimes 1) = A\otimes B$.

(4) Row two of D^* times column two of D; the two minus signs cancel.

(The minus sign in D is there exactly to kill the cross term, the same sign as in the tensor product of complexes (4.19).) So D^*D is block diagonal, and each block is a sum of two non-negative operators. A sum $A\otimes 1+1\otimes B$ with $A, B \geq 0$ has kernel $\operatorname{Ker} A\otimes\operatorname{Ker} B$ (expand in eigenvectors: the eigenvalue $\alpha+\beta$ vanishes only if $\alpha=\beta=0$). With $k_i = \dim\operatorname{Ker} D_i$ and $c_i = \dim\operatorname{Ker} D_i^*$,

$$\dim\operatorname{Ker} D \overset{(1)}{=} \dim\operatorname{Ker} D^*D \overset{(2)}{=} k_1k_2 + c_1c_2,$$

(1) $\operatorname{Ker} D = \operatorname{Ker} D^*D$ by the positivity argument (4.32).

(2) First block: $\operatorname{Ker} D_1^*D_1 \otimes \operatorname{Ker} D_2^*D_2 = \operatorname{Ker} D_1 \otimes \operatorname{Ker} D_2$; second block: $\operatorname{Ker} D_1^* \otimes \operatorname{Ker} D_2^*$.

and the same computation for DD^* (whose blocks are $D_1D_1^*\otimes 1 + 1\otimes D_2^*D_2$ and $D_1^*D_1\otimes 1 + 1\otimes D_2D_2^*$) gives $\dim\operatorname{Ker} D^* = c_1k_2 + k_1c_2$. Therefore

$$\operatorname{index} D = k_1k_2 + c_1c_2 - c_1k_2 - k_1c_2 \overset{(1)}{=} (k_1-c_1)(k_2-c_2) = \operatorname{index} D_1 \cdot \operatorname{index} D_2.$$

(1) Factorise.

Its leading symbol is the same block matrix with $D_i \to \sigma(D_i)$; since $\sigma^*\sigma = \sigma_1^*\sigma_1\otimes 1+1\otimes\sigma_2^*\sigma_2$ is positive unless both momenta vanish, D is elliptic, and its symbol class is $[\sigma_1]\boxtimes[\sigma_2]$. This is multiplicativity (4.54), checked with explicit zero modes.

Instance. The kernel computation above used nothing about the orders of the operators, so it applies verbatim to order-zero operators (making the product symbol elliptic then needs a small modification, which Atiyah–Singer supply). Take $D_1 = D_2 = A_{\bar z}$, the Toeplitz-type operator of the conjugate loop $\bar z = e^{-\mathrm{i}x}$, index +1, on each factor of the torus $T^2 = S^1\times S^1$. The product has index $1\cdot 1 = 1$; its single zero mode is the product of the two single zero modes. With $D_1 = A_z$ (index −1) and $D_2 = A_{\bar z}$ we get index −1. And Example 4.19 is also an instance: the de Rham operator of T^2 is the product of two copies of $\mathrm{d}/\mathrm{d}x$ on S^1, and its index is $0\cdot 0 = 0 = \chi(T^2)$.

The analytic Bott class: the harmonic oscillator

The embedding axiom for a point in $\mathbb{R}^n$ asks that the Bott class of $T^*\mathbb{R}^n = \mathbb{R}^{2n}$ have analytical index 1. What operator has the Bott class as its symbol? The Bott complex on $\mathbb{C}$ is "multiply by z" ((4.17)). On $T^*\mathbb{R} = \mathbb{R}^2$ with coordinates (x,ξ) we set $z = x + \mathrm{i}\xi$. We need an operator on functions of x whose symbol is $x+\mathrm{i}\xi$. The rule $\partial_x \to \mathrm{i}\xi$ from Section 4.2.1 gives it immediately:

$$a \coloneqq x + \frac{\mathrm{d}}{\mathrm{d}x}, \qquad \sigma(a)(x,\xi) = x + \mathrm{i}\xi. \tag{4.55}$$

This is the **annihilation operator** of the harmonic oscillator (in units $\hbar = m = \omega = 1$, up to the conventional factor $1/\sqrt{2}$). Its full symbol $x+\mathrm{i}\xi$ vanishes only at the origin of phase space, so it is invertible outside a compact set, which is exactly what a compactly supported class on $T^*\mathbb{R}$ requires. (Strictly, a is not elliptic in the sense of Section 4.2.1, because its symbol is not homogeneous; it is elliptic in the sense appropriate to $\mathbb{R}^n$, where position and momentum are treated alike. That is enough for its index to be well defined and deformation invariant.)

Worked example 4.37: The index of the annihilation operator is one

Kernel of a. Solve $au = 0$:

$$u' + xu = 0 \overset{(1)}{\Longrightarrow} \frac{u'}{u} = -x \overset{(2)}{\Longrightarrow} u(x) = C\,e^{-x^2/2}.$$

(1) Divide by u (separation of variables).
(2) Integrate: $\ln u = -x^2/2 + \text{const}$.

This is square integrable: it is the oscillator ground state. So $\dim \operatorname{Ker} a = 1$.
Kernel of a^*. The adjoint on $L^2(\mathbb{R})$ is $a^* = x - \mathrm{d}/\mathrm{d}x$ (integrate by parts; x is real). Solving $a^*v = 0$ gives $v' = xv$, i.e. $v = Ce^{+x^2/2}$, which is not square integrable unless $C = 0$. So $\dim \operatorname{Ker} a^* = 0$.
Hence

$$\operatorname{index} a = 1 - 0 = 1 = \operatorname{ind}_a(\text{Bott class of } T^*\mathbb{R}),$$

as the embedding axiom requires. In n dimensions take the product operator of Example 4.36 built from $a_1 = x_1 + \partial_1, \dots, a_n = x_n + \partial_n$; its index is $1^n = 1$ and its single zero mode is the Gaussian $e^{-|x|^2/2}$. Its symbol class is $\beta \boxtimes \cdots \boxtimes \beta = \lambda_{\mathbb{C}^n} = j_!(1)$, by Example 4.27. So $\operatorname{ind}_a(j_!1) = 1$: the analytic Bott periodicity.

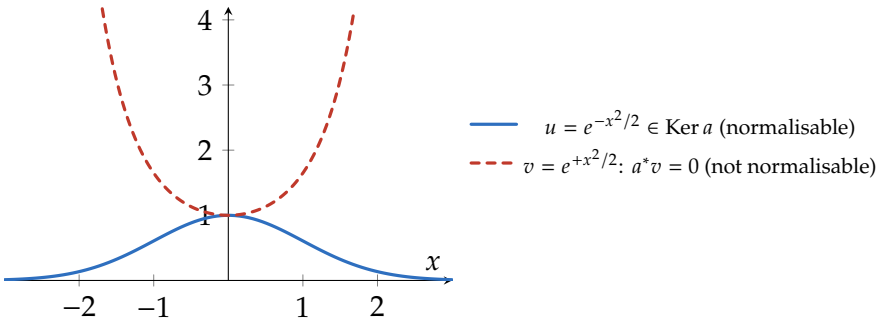


Figure 4.20: Zero modes of the annihilation operator $a = x + \mathrm{d}/\mathrm{d}x$ and its adjoint $a^* = x - \mathrm{d}/\mathrm{d}x$. The equation $au = 0$ has the Gaussian ground state (blue); $a^*v = 0$ has only a growing exponential (red), which is thrown out. One zero mode, no obstruction: $\operatorname{index} a = 1$, the analytic incarnation of the Bott class $z = x + \mathrm{i}\xi$ on phase space.

Watch out: Normalisability is a boundary condition at infinity

On a compact manifold every smooth solution of $Du = 0$ counts. On $\mathbb{R}$, the growing exponential $e^{x^2/2}$ is a perfectly smooth solution of $a^*v = 0$, and we discard it because it is not in L^2. The index of an operator on a non-compact space depends on such conditions at infinity. That is why Atiyah–Singer work with compact manifolds and why the Bott class is placed on the sphere S^n in the uniqueness argument. (The same issue reappears for manifolds with boundary in the last part of this chapter, where the boundary condition changes the index.)

4.7.4 The older proof: cobordism and spheres

Nash also sketches the first proof (Atiyah–Singer [1], 1963, published in detail by Palais' seminar). It follows the model of Hirzebruch's proof of his signature theorem. The problem looks formidable because "both the operator and the manifold are unrestricted". The proof reduces both.

1. *Reduce the manifold.* Embed $X \hookrightarrow \mathbb{R}^n$ and compactify to S^n, exactly as in the construction of ind_t. By excision, the index of D can be computed from an operator on the sphere.

2. *Reduce the operator.* Show that every symbol class is, at least rationally, a combination of symbol classes of the **twisted signature operators** $\mathrm{d} + \mathrm{d}^*$ acting on forms with values in

an auxiliary bundle F (Nash §3). These are "the generalised signature operators".

3. *Cobordism invariance.* Both indices are invariant under cobordism: if $X = \partial W$ is the boundary of a compact manifold W (and the operator extends over W suitably), then its index is 0. So each side is a homomorphism out of a cobordism group, which is known (Thom) and has explicit generators.

4. *Compute on generators.* Check the theorem for the twisted signature operators on spheres and on the complex projective spaces by explicit calculation.

Working definition 4.25 (cobordism)

Two closed n-manifolds [compact and without boundary] X_0, X_1 are **cobordant** if there is a compact $(n+1)$-manifold W with $\partial W = X_0 \sqcup \overline{X_1}$ [the boundary of W is the two manifolds side by side, not touching] (the bar reverses orientation). *Operational test for the simplest case:* X is cobordant to nothing ("null-cobordant") if it bounds [$X = \partial W$ for some compact W].

Worked example 4.38: Why the signature of a boundary is zero

Let $X^{4k} = \partial W$. The signature of X is the signature of the intersection form $(\alpha, \beta) \mapsto \int_X \alpha \wedge \beta$ on $H^{2k}(X)$. Classes that extend over W (restrictions $\alpha = \iota^* A$ of closed forms A on W) are isotropic:

$$\int_X \iota^* A \wedge \iota^* B \overset{(1)}{=} \int_{\partial W} A \wedge B \overset{(2)}{=} \int_W \mathrm{d}(A \wedge B) \overset{(3)}{=} 0.$$

(1) Pullback commutes with wedge, and $\iota : X \to W$ is the inclusion of the boundary.
(2) Stokes' theorem (Chapter T2).
(3) A and B are closed, so $\mathrm{d}(A \wedge B) = \mathrm{d}A \wedge B \pm A \wedge \mathrm{d}B = 0$.

A standard duality argument shows these classes fill exactly half of $H^{2k}(X)$, a Lagrangian subspace, and a quadratic form with a half-dimensional isotropic subspace has signature 0. Since the signature is the index of the signature operator (Nash §2), this is cobordism invariance of the index in its oldest form: *the index of the signature operator of a boundary vanishes.* Example: $S^4 = \partial B^5$ has $H^2 = 0$ and signature 0; $\mathbb{CP}^2$ has signature 1, so it bounds no compact oriented 5-manifold.

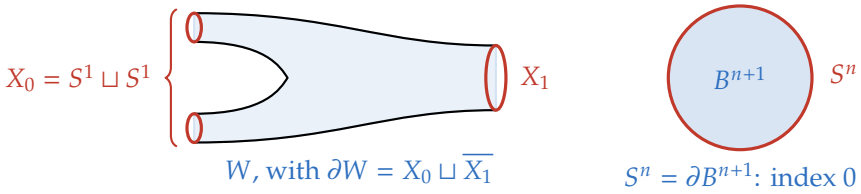


Figure 4.21: Cobordism. Left: a pair of pants W is a cobordism from X_0, a disjoint union of two circles, to one circle X_1. Right: a sphere (red circle) is the boundary of a ball (blue disc), so it is null-cobordant. The index of the (twisted) signature operator vanishes on boundaries, so it only depends on the cobordism class of (X, F), and the old proof reduces the theorem to a check on a few generators.

Physical picture: Why the index should vanish on a boundary

Suppose $X = \partial W$. A zero mode on X ought to be the boundary value of a solution on W (think of harmonic functions on a ball, determined by their boundary values). A careful

analysis shows that the boundary values of solutions of the extended operator on W form a Lagrangian subspace of the zero-mode space of X, which pairs positive-chirality modes with negative ones. Paired modes cancel in the index. The signature example above is the cohomological shadow of this.

4.7.5 A third proof: the heat equation

There is a third route, not described by Nash on these pages but used throughout physics, which proves the cohomological form of the theorem directly: the **heat equation method** of Atiyah–Bott–Patodi, McKean–Singer and Gilkey, streamlined by Getzler and by the supersymmetric path integral of Alvarez-Gaumé and Friedan–Windey. Its starting point is the formula

$$\operatorname{index} D = \operatorname{tr} e^{-tD^*D} - \operatorname{tr} e^{-tDD^*} \qquad \text{for every } t > 0,$$

whose right side, as $t \to 0$, becomes the integral over X of a *local* expression in the curvature. It explains, better than either of the K-theory proofs, why the index is the integral of a characteristic form. It is the subject of the next section.

Reading Nash (p. 97)

Nash: ind_t is independent of the embedding and tubular neighbourhood (Atiyah–Singer [2]); $j_!$ is a ring isomorphism $\mathbb{Z} \to \mathbb{Z}$, hence the identity, but is kept in the definition for the equivariant theory of §6. *Theorem (Atiyah–Singer): the topological and the analytical index are equal.* The proof is axiomatic: the topological index is the *unique* map satisfying certain index axioms, and the analytical index is shown "after considerable amount of work" to satisfy them. The earlier proof (Atiyah–Singer [1]) reduces X to a sphere via $i\colon X \to \mathbb{R}^n$ and D to a generalised signature operator, and checks spheres by calculation. We have written out the axioms (A0)–(A2), the uniqueness argument (4.53), excision and multiplicativity with the explicit product operator, the harmonic oscillator as the analytic Bott class, the cobordism step, and the checks of Table 4.1.

Recap

- Atiyah–Singer: $\operatorname{ind}_a = \operatorname{ind}_t$ on $K(T^*X)$; the number of zero modes minus obstructions is computed by the symbol class and the topology of X.
- Checks: point (rank–nullity), S^1 (Toeplitz, winding number), surfaces (Gauss–Bonnet), $\mathbb{CP}^1$ (Riemann–Roch, $n+1$).
- Uniqueness: normalisation on a point plus the embedding axiom $\operatorname{ind}^Y \circ i_! = \operatorname{ind}^X$ force the index to be $(j_!)^{-1} i_!$; the proof routes a class through the sphere.
- The analytical index satisfies the axioms by excision (zero modes live where the symbol is non-trivial) and multiplicativity (the product operator has index index $D_1 \cdot$ index D_2); the oscillator $a = x + \mathrm{d}/\mathrm{d}x$ has index 1 and realises the Bott class.
- Older proof: reduce to spheres and twisted signature operators; the index vanishes on boundaries (cobordism invariance).

4.8 The heat-kernel view: McKean–Singer

The index is a difference of two dimensions, $\dim\operatorname{Ker} D - \dim\operatorname{Ker} D^*$. Each dimension separately is hard to compute: we must find all solutions of a PDE, and the answer jumps under deformation. The question of this section is: *can we write the index as a sum over all modes of the operator, in such a way that the sum is easy to evaluate?* The answer is yes. Weight every eigenmode of D^*D and of DD^* by a Boltzmann factor $e^{-t\lambda}$ and subtract. The nonzero modes cancel in pairs, for every t, leaving the index. Then let $t \to 0$: the heat operator only feels the geometry within a distance $\sqrt{t}$ of each point, so the index becomes the integral of a *local* density built from the curvature. This is the heat-equation proof of the index theorem, and it is how the theorem enters physics, through the Witten index and Fujikawa's computation of anomalies.

Nash does not develop this on his pp. 89–97; we follow Bär, Chapter 3 [Bär §3.1–3.4], and use the rolled-up operators of Section 4.4 as examples.

4.8.1 Nonzero modes come in pairs

Let $D\colon \Gamma(E^+) \to \Gamma(E^-)$ be elliptic on a compact manifold X (for instance a rolled-up complex), and put

$$\Delta^+ = D^*D \ \text{ on } E^+, \qquad \Delta^- = DD^* \ \text{ on } E^-.$$

Both are self-adjoint, non-negative, elliptic operators, so each has a discrete spectrum of eigenvalues $0 \le \lambda_1 \le \lambda_2 \le \cdots \to \infty$ with finite-dimensional eigenspaces $E_\lambda(\Delta^\pm)$ (Nash Chapter 2). We know from (4.32) that $\operatorname{Ker}\Delta^+ = \operatorname{Ker} D$ and $\operatorname{Ker}\Delta^- = \operatorname{Ker} D^*$, so

$$\operatorname{index} D = \dim E_0(\Delta^+) - \dim E_0(\Delta^-). \tag{4.56}$$

Now let $\lambda > 0$ and $\Delta^+\varphi = \lambda\varphi$. Then

$$\Delta^-(D\varphi) \overset{(1)}{=} DD^*D\varphi \overset{(2)}{=} D(\Delta^+\varphi) \overset{(3)}{=} \lambda\, D\varphi, \tag{4.57}$$

(1) Definition $\Delta^- = DD^*$.
(2) Regroup: $D(D^*D) = (DD^*)D$.
(3) φ is an eigenvector of Δ^+ with eigenvalue λ.

so D carries $E_\lambda(\Delta^+)$ into $E_\lambda(\Delta^-)$. It does not kill anything there, since

$$\|D\varphi\|^2 \overset{(1)}{=} \langle \varphi, D^*D\varphi\rangle \overset{(2)}{=} \lambda\,\|\varphi\|^2 > 0.$$

(1) Definition of the adjoint.
(2) Eigenvalue equation, and $\lambda > 0$.

In the same way D^* carries $E_\lambda(\Delta^-)$ into $E_\lambda(\Delta^+)$, and on $E_\lambda(\Delta^+)$

$$\tfrac{1}{\lambda}D^*(D\varphi) \overset{(1)}{=} \tfrac{1}{\lambda}\Delta^+\varphi \overset{(2)}{=} \varphi,$$

(1) $D^*D = \Delta^+$.
(2) Eigenvalue equation.

so $\frac{1}{\lambda}D^*$ is the inverse of D between the two eigenspaces [Bär p. 149].

Key idea: Supersymmetric pairing

For every $\lambda > 0$, $D\colon E_\lambda(\Delta^+) \to E_\lambda(\Delta^-)$ is an isomorphism. Hence

$$\dim E_\lambda(\Delta^+) = \dim E_\lambda(\Delta^-) \qquad (\lambda > 0)\,: \tag{4.58}$$

Δ^+ and Δ^- have the same nonzero spectrum, with the same multiplicities. Only the zero modes can be unpaired, and their mismatch is the index.

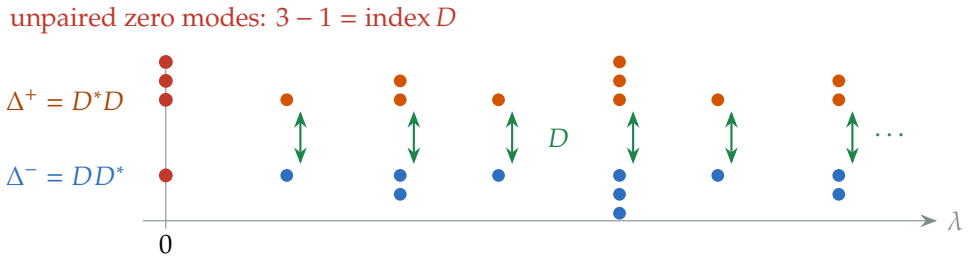


Figure 4.22: The spectra of Δ^+ (top) and Δ^- (bottom). At every nonzero eigenvalue the two eigenspaces have the same dimension (the stacks of dots), because D maps one isomorphically onto the other (green). Only the zero modes (red) can fail to match; here there are three on top and one below, so index $D = 2$.

Dictionary: Index theory ↔ supersymmetric quantum mechanics

Mathematics	Physics
$\mathcal{H} = \Gamma(E^+) \oplus \Gamma(E^-)$	Hilbert space: bosonic ⊕ fermionic states
grading $\Gamma = \mathrm{diag}(1,-1)$	$(-1)^F$, fermion parity
$Q = \left(\begin{smallmatrix} 0 & D^* \\ D & 0 \end{smallmatrix}\right)$	supercharge, $Q = Q^\dagger$, $Q\Gamma = -\Gamma Q$
$Q^2 = \Delta^+ \oplus \Delta^-$	Hamiltonian $H = Q^2 \geq 0$
pairing (4.58)	boson–fermion degeneracy of excited states
$\operatorname{Ker} Q = \operatorname{Ker} D \oplus \operatorname{Ker} D^*$	supersymmetric ground states ($E = 0$)
index D	Witten index $\mathrm{Tr}(-1)^F$

Worked example 4.39: The harmonic oscillator, mode by mode

Take $D = a = x + \mathrm{d}/\mathrm{d}x$ on $L^2(\mathbb{R})$, index 1 (Example 4.37). First compute the two Laplacians:

$$a^*a\,u \overset{(1)}{=} \Big(x - \frac{\mathrm{d}}{\mathrm{d}x}\Big)(xu + u') \overset{(2)}{=} x^2u + xu' - u - xu' - u'' = \big(-\tfrac{\mathrm{d}^2}{\mathrm{d}x^2} + x^2 - 1\big)u,$$

(1) $a^* = x - \mathrm{d}/\mathrm{d}x$ applied to $au = xu + u'$.
(2) Product rule: $\frac{\mathrm{d}}{\mathrm{d}x}(xu) = u + xu'$.

and in the same way $aa^* = -\frac{\mathrm{d}^2}{\mathrm{d}x^2} + x^2 + 1$. The oscillator Hamiltonian $-\frac{\mathrm{d}^2}{\mathrm{d}x^2} + x^2$ has eigenvalues $2n + 1$ with Hermite functions ψ_n, $n = 0, 1, 2, \ldots$. Hence

$$\operatorname{spec}\Delta^+ = \{0, 2, 4, 6, \ldots\}, \qquad \operatorname{spec}\Delta^- = \{2, 4, 6, \ldots\},$$

each eigenvalue with multiplicity one. The pairing is the ladder: $a\psi_n = \sqrt{2n}\,\psi_{n-1}$ takes the level-$2n$ state of Δ^+ (which is ψ_n) to the level-$2n$ state of Δ^- (which is ψ_{n-1}, since $aa^*\psi_{n-1} = (2(n-1) + 1 + 1)\psi_{n-1} = 2n\,\psi_{n-1}$). The ground state ψ_0 at $\lambda = 0$ has no partner. One unpaired mode: index 1.

4.8.2 The heat operator and the heat kernel

To turn (4.58) into a formula we need a way of summing over all eigenvalues that converges. The Boltzmann weight $e^{-t\lambda}$ is the natural choice.

Working definition 4.26 (heat operator, heat kernel, heat trace)

Let Δ be a non-negative [$\langle u, \Delta u\rangle \geq 0$, so every $\lambda_j \geq 0$] self-adjoint elliptic operator of order two (a Laplacian) on sections of $E \to X$, with orthonormal eigensections $\Delta\varphi_j = \lambda_j\varphi_j$.

- The **heat operator** $e^{-t\Delta}$, $t > 0$, multiplies each eigensection by $e^{-t\lambda_j}$. Applied to initial data u_0 it solves the heat equation $(\partial_t + \Delta)u = 0$, $u|_{t=0} = u_0$.
- It is an integral operator, $(e^{-t\Delta}u)(x) = \int_X K_t(x,y)\,u(y)\,\mathrm{dvol}(y)$, with the **heat kernel**

$$K_t(x,y) = \sum_j e^{-t\lambda_j}\,\varphi_j(x)\otimes\varphi_j(y)^\dagger, \tag{4.59}$$

  a matrix $E_y \to E_x$ [$\varphi_j(y)^\dagger$ is the conjugate-transpose row vector, so each term is a column times a row] for each pair of points [Bär Def. 3.1.3].
- The **heat trace** is $\operatorname{Tr} e^{-t\Delta} = \sum_j e^{-t\lambda_j} = \int_X \operatorname{tr} K_t(x,x)\,\mathrm{dvol}(x)$ [tr is the matrix trace over the fibre E_x, Tr the trace over all sections].

Physically: $K_t(x,y)$ is the temperature at x at time t when a unit of heat is released at y; equivalently, the Euclidean propagator $\langle x|e^{-tH}|y\rangle$ of a particle with Hamiltonian $H = \Delta$.

Why the trace formula holds:

$$\int_X \operatorname{tr} K_t(x,x)\,\mathrm{dvol} \overset{(1)}{=} \sum_j e^{-t\lambda_j}\int_X \varphi_j(x)^\dagger\varphi_j(x)\,\mathrm{dvol} \overset{(2)}{=} \sum_j e^{-t\lambda_j}.$$

(1) Insert (4.59) at $y = x$; the trace of the matrix $v\otimes v^\dagger$ is $v^\dagger v = |v|^2$.
(2) The φ_j are normalised.

The sum converges for every $t > 0$ because the eigenvalues grow: by Weyl's law, on an n-manifold λ_j grows like $j^{2/n}$, and $\sum_j e^{-tj^{2/n}} < \infty$. It diverges as $t \to 0$, where it counts every mode; how it diverges is the subject of Section 4.8.4.

Worked example 4.40: The heat kernel of the line

On $X = \mathbb{R}$ with $\Delta = -\mathrm{d}^2/\mathrm{d}x^2$, the heat kernel is the Gaussian

$$K_t(x,y) = (4\pi t)^{-1/2}\,e^{-(x-y)^2/4t}.$$

Check the heat equation. Write $K = K_t(x,y)$ and $s = x - y$:

$$\partial_t K \overset{(1)}{=} \Big(-\frac{1}{2t} + \frac{s^2}{4t^2}\Big)K,$$
$$\partial_x^2 K \overset{(2)}{=} \partial_x\Big(-\frac{s}{2t}K\Big) \overset{(3)}{=} \Big(-\frac{1}{2t} + \frac{s^2}{4t^2}\Big)K,$$

(1) Differentiate $t^{-1/2}$ (giving $-\frac{1}{2t}$) and the exponent $-s^2/4t$ (giving $+s^2/4t^2$).
(2) $\partial_x e^{-s^2/4t} = -\frac{s}{2t}e^{-s^2/4t}$.
(3) Product rule: $\partial_x(-\frac{s}{2t}) = -\frac{1}{2t}$, plus $-\frac{s}{2t}\cdot(-\frac{s}{2t})K$.

so $\partial_t K = \partial_x^2 K$, i.e. $(\partial_t + \Delta)K = 0$. Also $\int K\,\mathrm{d}x = 1$ (a normalised Gaussian of variance $2t$), and as $t \to 0$ it tends to $\delta(x-y)$. In n dimensions the kernel is the product $(4\pi t)^{-n/2} e^{-|x-y|^2/4t}$. Its value on the diagonal, $(4\pi t)^{-n/2}$, is the leading behaviour of every heat kernel on every n-manifold [Bär Def. 3.2.3].

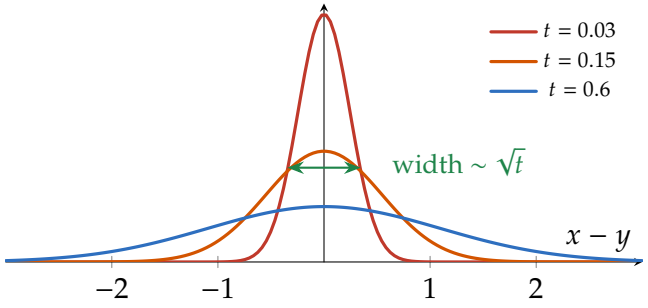


Figure 4.23: The heat kernel $K_t(x,y) = (4\pi t)^{-1/2} e^{-(x-y)^2/4t}$ at three times (the red curve is cut off at the top of the plot). Heat released at y spreads a distance of order $\sqrt{t}$. For small t the kernel is sharply peaked, so $K_t(x,x)$ can only depend on the geometry within a distance $\sim \sqrt{t}$ of x: this is why the small-t behaviour of the heat trace is local.

Worked example 4.41: The circle: local and global pieces of the heat trace

On $S^1 = \mathbb{R}/2\pi\mathbb{Z}$ with $\Delta = -\mathrm{d}^2/\mathrm{d}x^2$ the eigenfunctions are $e^{\mathrm{i}Nx}/\sqrt{2\pi}$, $N \in \mathbb{Z}$, with eigenvalues N^2. So

$$K_t(x,y) = \frac{1}{2\pi} \sum_{N\in\mathbb{Z}} e^{-tN^2} e^{\mathrm{i}N(x-y)}, \qquad \operatorname{Tr} e^{-t\Delta} = \sum_{N\in\mathbb{Z}} e^{-tN^2}.$$

This sum is a theta function; it is easy for large t (only $N = 0, \pm 1$ matter) but not for small t. Use Poisson summation, $\sum_N f(N) = \sum_m \hat{f}(2\pi m)$ with $\hat{f}(\omega) = \int f(s) e^{-\mathrm{i}\omega s}\mathrm{d}s$:

$$\begin{aligned} \sum_{N\in\mathbb{Z}} e^{-tN^2} &\overset{(1)}{=} \sum_{m\in\mathbb{Z}} \int_{-\infty}^{\infty} e^{-ts^2} e^{-2\pi \mathrm{i} m s}\,\mathrm{d}s \\ &\overset{(2)}{=} \sqrt{\frac{\pi}{t}} \sum_{m\in\mathbb{Z}} e^{-\pi^2 m^2/t} \\ &\overset{(3)}{=} (4\pi t)^{-1/2} \cdot 2\pi \cdot \left(1 + 2e^{-\pi^2/t} + 2e^{-4\pi^2/t} + \cdots\right). \end{aligned} \tag{4.60}$$

(1) Poisson summation with $f(s) = e^{-ts^2}$.

(2) Gaussian integral: $\int e^{-ts^2 - \mathrm{i}\omega s}\mathrm{d}s = \sqrt{\pi/t}\, e^{-\omega^2/4t}$, at $\omega = 2\pi m$.

(3) $\sqrt{\pi/t} = 2\pi/\sqrt{4\pi t}$; group $\pm m$.

Read the answer. The leading factor $(4\pi t)^{-1/2} \cdot 2\pi$ is "flat-space diagonal value times length of the circle": local. The corrections $e^{-\pi^2 m^2/t} = e^{-(2\pi m)^2/4t}$ are the flat heat kernel evaluated at distance $2\pi m$, i.e. heat that has gone m times round the circle and come back (the method of images). They are global, and they are exponentially small as $t \to 0$. So the power-series part of the small-t expansion is purely local, with no powers of t beyond the leading one here because the circle is flat.

4.8.3 The McKean–Singer formula

Now combine the heat trace with the pairing.

Key idea: McKean–Singer formula

For every $t > 0$,

$$\operatorname{index} D = \operatorname{Tr} e^{-tD^*D} - \operatorname{Tr} e^{-tDD^*} = \operatorname{Str} e^{-tQ^2}, \tag{4.61}$$

where $\operatorname{Str} A = \operatorname{Tr}(\Gamma A)$ is the **supertrace** (trace over E^+ minus trace over E^-).

Derivation by pairing [Bär (3.17)]:

$$\begin{aligned} \operatorname{Tr} e^{-t\Delta^+} - \operatorname{Tr} e^{-t\Delta^-} &\overset{(1)}{=} \sum_{\lambda \geq 0} e^{-t\lambda} \big(\dim E_\lambda(\Delta^+) - \dim E_\lambda(\Delta^-)\big) \\ &\overset{(2)}{=} e^{-t\cdot 0} \big(\dim E_0(\Delta^+) - \dim E_0(\Delta^-)\big) \\ &\overset{(3)}{=} \operatorname{index} D. \end{aligned} \tag{4.62}$$

(1) Group each trace by distinct eigenvalues, each counted with its multiplicity.
(2) Pairing (4.58): every term with $\lambda > 0$ vanishes.
(3) (4.56).

Derivation by supersymmetry. It is instructive to see the t-independence directly, without diagonalising anything, because this is the argument that generalises to path integrals. Two facts about the supertrace: Γ anticommutes with Q (Q is "odd": it maps E^+ to E^- and back), and the trace is cyclic, $\operatorname{Tr}(AB) = \operatorname{Tr}(BA)$, for the smoothing operators we meet here. For any odd operator B (also anticommuting with Γ),

$$\operatorname{Str}(QB) \overset{(1)}{=} \operatorname{Tr}(\Gamma QB) \overset{(2)}{=} -\operatorname{Tr}(Q\Gamma B) \overset{(3)}{=} -\operatorname{Tr}(\Gamma BQ) \overset{(4)}{=} -\operatorname{Str}(BQ), \tag{4.63}$$

(1) Definition of the supertrace.
(2) $\Gamma Q = -Q\Gamma$.
(3) Cyclicity: move Q from the front to the back.
(4) Definition of the supertrace again.

so $\operatorname{Str}(QB+BQ) = 0$: *the supertrace of an anticommutator of odd operators vanishes*. Now differentiate:

$$\frac{\mathrm{d}}{\mathrm{d}t} \operatorname{Str} e^{-tQ^2} \overset{(1)}{=} -\operatorname{Str}\big(Q^2 e^{-tQ^2}\big) \overset{(2)}{=} -\tfrac{1}{2} \operatorname{Str}\big(Q\,(Qe^{-tQ^2}) + (Qe^{-tQ^2})\,Q\big) \overset{(3)}{=} 0. \tag{4.64}$$

(1) Differentiate the exponential.
(2) Q commutes with e^{-tQ^2}, so $Q^2 e^{-tQ^2} = Q(Qe^{-tQ^2}) = (Qe^{-tQ^2})Q$; write it as half the sum.
(3) (4.63) with the odd operator $B = Qe^{-tQ^2}$ (e^{-tQ^2} is even, commuting with Γ, so B is odd).

So $\operatorname{Str} e^{-tQ^2}$ is independent of t. As $t \to \infty$ every mode with $\lambda > 0$ is exponentially suppressed and the heat operator becomes the projection onto $\operatorname{Ker} Q^2 = \operatorname{Ker} D \oplus \operatorname{Ker} D^*$; the supertrace of that projection is $\dim \operatorname{Ker} D - \dim \operatorname{Ker} D^*$. Since the supertrace does not depend on t, it equals $\operatorname{index} D$ for every t.

Watch out: The ordinary trace is not constant

$\operatorname{Tr} e^{-t\Delta^+}$ and $\operatorname{Tr} e^{-t\Delta^-}$ each depend strongly on t and blow up as $t \to 0$. Only their difference is constant. Likewise, the cancellation in (4.63) needs the operators to be trace

class: Str(Q) or Str(1) make no sense on an infinite-dimensional space (the formal "index $=\dim E^+ - \dim E^-$" is $\infty - \infty$). The factor e^{-tQ^2} is exactly what makes everything trace class; it is a regulator that the answer does not depend on.

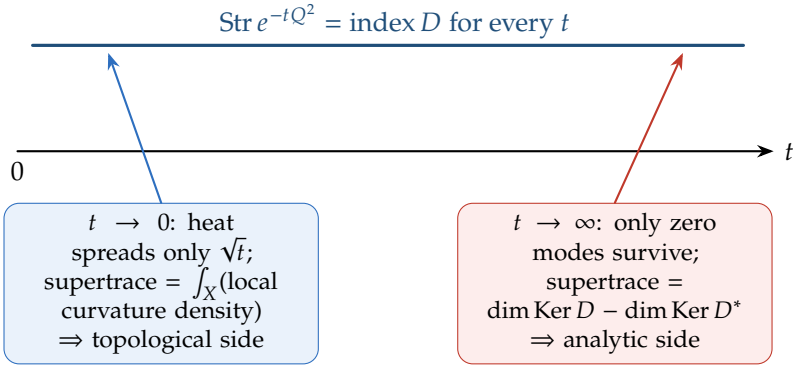


Figure 4.24: The logic of the heat-equation proof. The supertrace of the heat operator is constant in t. At large t it visibly counts zero modes (the analytical index). At small t it becomes an integral of a local expression in the curvature, which turns out to be the characteristic-class integrand of the topological index.

Worked example 4.42: The oscillator: McKean–Singer summed exactly

With the spectra of Example 4.39,

$$\operatorname{Tr} e^{-ta^*a} - \operatorname{Tr} e^{-taa^*} \overset{(1)}{=} \sum_{n=0}^{\infty} e^{-2nt} - \sum_{n=1}^{\infty} e^{-2nt} \overset{(2)}{=} \frac{1}{1-e^{-2t}} - \frac{e^{-2t}}{1-e^{-2t}} \overset{(3)}{=} 1.$$

(1) $\operatorname{spec} a^*a = \{0, 2, 4, \dots\}$, $\operatorname{spec} aa^* = \{2, 4, \dots\}$.
(2) Geometric series.
(3) $1 - e^{-2t}$ over itself.

Each trace separately is $\approx \frac{1}{2t}$ for small t (the classical phase-space volume $\int \frac{dx\,d\xi}{2\pi} e^{-t(x^2+\xi^2)} = \frac{1}{2t}$); the difference is exactly 1 for every t.

Worked example 4.43: The de Rham operator on the circle

On S^1, $E^+ = \Lambda^0$, $E^- = \Lambda^1$, $D = d$. The Laplacian on functions has eigenvalues N^2 ($N \in \mathbb{Z}$); on 1-forms $h\,dx$ it acts on h in the same way, with the same eigenvalues. So $\operatorname{Tr} e^{-t\Delta_0} = \operatorname{Tr} e^{-t\Delta_1} = \sum_N e^{-tN^2}$ and $\operatorname{Str} e^{-tQ^2} = 0 = \chi(S^1)$, mode by mode, for every t, including the zero modes 1 and dx, which pair with each other.

Worked example 4.44: The de Rham operator on the two-sphere, all t

Take the unit sphere, $D = d + d^*$ from $\Omega^0 \oplus \Omega^2$ to Ω^1, so $\Delta^+ = \Delta_0 \oplus \Delta_2$ and $\Delta^- = \Delta_1$.

- Δ_0: spherical harmonics Y_{lm}, eigenvalue $l(l+1)$, multiplicity $2l+1$, $l \geq 0$.
- Δ_2: the Hodge star $\star\colon \Omega^0 \to \Omega^2$ commutes with the Laplacian, so Δ_2 has the same spectrum as Δ_0 (eigenforms $Y_{lm}\,dA$).
- Δ_1: there are no harmonic 1-forms ($b_1 = 0$), and by the Hodge decomposition every 1-form is $df + \star dg$. The forms dY_{lm} and $\star dY_{lm}$, $l \geq 1$, are eigenforms with eigenvalue

$l(l+1)$ (because Δ commutes with d and $\star$), giving multiplicity $2(2l+1)$.

Write $\vartheta(t) = \sum_{l\geq 0}(2l+1)e^{-tl(l+1)}$. Then

$$\begin{aligned}\operatorname{Str} e^{-tQ^2} &\overset{(1)}{=} \operatorname{Tr} e^{-t\Delta_0} + \operatorname{Tr} e^{-t\Delta_2} - \operatorname{Tr} e^{-t\Delta_1} \\ &\overset{(2)}{=} 2\vartheta(t) - 2\big(\vartheta(t) - 1\big) \\ &= 2 = \chi(S^2).\end{aligned}$$

(1) $\Delta^+ = \Delta_0 \oplus \Delta_2$, $\Delta^- = \Delta_1$.
(2) The spectra above; the Δ_1 sum is 2ϑ minus its $l = 0$ term, which is $2 \cdot 1$.

The pairing is visible: d maps Y_{lm} to $\mathrm{d}Y_{lm}$, and d^* maps $Y_{lm}\mathrm{d}A$ to $\pm \star\, \mathrm{d}Y_{lm}$. The two unpaired zero modes are the constant function and the area form. See Figure 4.25.

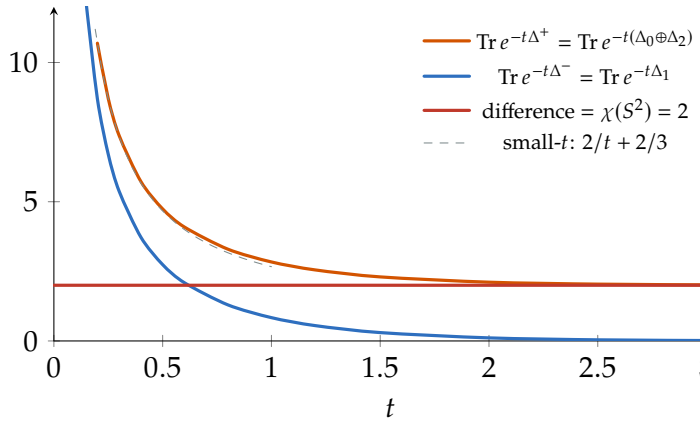


Figure 4.25: Heat traces on the unit two-sphere, computed from the exact spectra. Each trace blows up like $2/t$ as $t \to 0$ (Weyl's law: it counts all modes) and decays to its number of zero modes as $t \to \infty$ (2 for Δ^+, 0 for Δ^-). Their difference is exactly 2 at every t. The dashed curve is the small-t expansion of the upper trace derived in Example 4.45.

4.8.4 Small t: the index is a local integral

We now use the freedom to take t small. For small t, heat has spread only a distance $\sim \sqrt{t}$ (Figure 4.23), so the heat kernel on the diagonal can only know about the geometry near x. Making this precise gives the central analytic fact of the method.

Key idea: Short-time expansion of the heat kernel [Bär Cor. 3.2.10–3.2.11]

Let $\Delta = \nabla^*\nabla + \mathcal{K}$ be a Laplace-type operator on a bundle E of rank r over a compact Riemannian n-manifold ($\nabla^*\nabla$ is the connection Laplacian and $\mathcal{K}$ a symmetric endomorphism field). Then as $t \to 0$,

$$K_t(x,x) \sim (4\pi t)^{-n/2} \sum_{j=0}^{\infty} t^j\, \Phi_j(x), \qquad \Phi_0 = \mathrm{id}_E, \quad \Phi_1 = \tfrac{1}{6}\operatorname{scal}\mathrm{id}_E - \mathcal{K}, \tag{4.65}$$

where scal is the scalar curvature and each $\Phi_j(x)$ is a universal polynomial in the curvature of X, the curvature of ∇, $\mathcal{K}$, and their covariant derivatives at x. Integrating,

$$\operatorname{Tr} e^{-t\Delta} \sim (4\pi t)^{-n/2}\Big(r\operatorname{vol}(X) + t\int_X \big(\tfrac{r}{6}\operatorname{scal} - \operatorname{tr}\mathcal{K}\big)\mathrm{dvol} + O(t^2)\Big). \tag{4.66}$$

A dimensional check: t has dimensions of length^2 (because $\partial_t \sim \partial_x^2$), so Φ_j has dimensions of length^{-2j}, i.e. it contains $2j$ derivatives of the metric. Φ_1 is linear in curvature (two derivatives), Φ_2 quadratic in curvature, and so on. The leading term of (4.66) is Weyl's law: the number of eigenvalues below Λ grows like $r\,\mathrm{vol}(X)\Lambda^{n/2}$ times a universal constant. ("One can hear the volume of a drum.")

Now feed (4.65) into McKean–Singer. Write $a_j(x) = \operatorname{tr}\Phi_j^{\Delta^+}(x) - \operatorname{tr}\Phi_j^{\Delta^-}(x)$. Then [Bär (3.18)–(3.19)]

$$\begin{aligned}\operatorname{index} D &\overset{(1)}{=} \int_X \big(\operatorname{tr} K_t^+(x,x) - \operatorname{tr} K_t^-(x,x)\big)\,\mathrm{dvol}\\ &\overset{(2)}{=} (4\pi t)^{-n/2}\sum_{j\ge 0} t^j \int_X a_j\,\mathrm{dvol} + (\text{terms vanishing faster than any power of } t)\\ &\overset{(3)}{=} (4\pi)^{-n/2}\int_X a_{n/2}\,\mathrm{dvol}\quad (n \text{ even}),\qquad 0\quad (n \text{ odd}).\end{aligned}\tag{4.67}$$

(1) McKean–Singer (4.61) written with heat kernels.
(2) Short-time expansion (4.65) for $\Delta^\pm$.
(3) The left side does not depend on t. So on the right the coefficient of every negative power $t^{j-n/2}$, $j < n/2$, must vanish, and the t^0 coefficient is the index. For n even that is $j = n/2$. For n odd there is no t^0 term at all ($j - n/2$ is a half-integer), and letting $t \to 0$ gives 0.

Key idea: What the heat equation proves

1. $\operatorname{index} D = (4\pi)^{-n/2}\int_X a_{n/2}(x)\,\mathrm{dvol}$: the index is the integral of a *local* density built from the curvature and the coefficients of D [Bär Thm. 3.4.4].
2. On odd-dimensional manifolds the index of every such (Dirac-type) operator vanishes. (Compare the circle: differential operators on S^1 have index 0, Example 4.23.)
3. The index is constant under smooth deformations of the metric and of D, because it is an integer that depends smoothly on the parameters [Bär Cor. 3.4.7].

Identifying $a_{n/2}$ with the characteristic-class integrand of the topological index (ch · td, or the Euler form, or $\hat{A}$) is the remaining, and hardest, step. For $n = 2$ we can do it by hand.

Worked example 4.45: The sphere: the small-t expansion from the exact spectrum

Let us check (4.66) on the unit S^2 for Δ_0, where $r = 1$, $\mathcal{K} = 0$, vol $= 4\pi$, and scal $= 2$ (twice the Gaussian curvature $K = 1$). The prediction is

$$\vartheta(t) \overset{(1)}{=} (4\pi t)^{-1}\Big(4\pi + t\cdot\tfrac{1}{6}\cdot 2\cdot 4\pi + O(t^2)\Big) = \frac{1}{t} + \frac{1}{3} + O(t).$$

(1) (4.66) with $n = 2$.

Now compute from the spectrum. Put $f(l) = (2l+1)e^{-tl(l+1)}$, so $\vartheta = \sum_{l\ge 0} f(l)$. The Euler–Maclaurin formula $\sum_{l=0}^{\infty} f(l) = \int_0^\infty f\,\mathrm{d}l + \frac{1}{2}f(0) - \frac{1}{12}f'(0) + \frac{1}{720}f'''(0) - \cdots$ gives

$$\begin{aligned}\vartheta(t) &\overset{(1)}{=} \Big[-\tfrac{1}{t}e^{-tl(l+1)}\Big]_0^\infty + \tfrac{1}{2} - \tfrac{1}{12}(2-t) + O(t)\\ &\overset{(2)}{=} \frac{1}{t} + \frac{1}{2} - \frac{1}{6} + O(t) = \frac{1}{t} + \frac{1}{3} + O(t).\end{aligned}$$

(1) $f = -\frac{1}{t}\frac{\mathrm{d}}{\mathrm{d}l}e^{-tl(l+1)}$ is an exact derivative; $f(0) = 1$; $f'(l) = \big(2 - t(2l+1)^2\big)e^{-tl(l+1)}$, so $f'(0) = 2 - t$; the t^0 part of f is the linear polynomial $2l + 1$, so $f'''(0) = O(t)$.

(2) Evaluate the integral: $0 - (-\frac{1}{t}) = \frac{1}{t}$.

The two agree. (Numerically, $\vartheta(0.1) = 10.3401$ against $\frac{1}{0.1} + \frac{1}{3} + \frac{0.1}{15} = 10.3400$; the next coefficient is $\frac{1}{15}$.)

For Δ_1 we need the Weitzenböck formula: on 1-forms, $\Delta_1 = \nabla^*\nabla + \mathrm{Ric}$, so $r = 2$ and $\operatorname{tr}\mathcal{K} = \mathrm{scal} = 2$. Then

$$\operatorname{Tr} e^{-t\Delta_1} \overset{(1)}{=} (4\pi t)^{-1}\Big(2\cdot 4\pi + t\big(\tfrac{2}{6}\cdot 2 - 2\big)4\pi + O(t^2)\Big) = \frac{2}{t} - \frac{4}{3} + O(t),$$

(1) (4.66) with $r = 2$, $\mathcal{K} = \mathrm{Ric}$; all curvatures are constant on the round sphere.

which agrees with the exact $2(\vartheta - 1) = \frac{2}{t} + \frac{2}{3} - 2$. The supertrace at small t:

$$\Big(\frac{1}{t} + \frac{1}{3}\Big) + \Big(\frac{1}{t} + \frac{1}{3}\Big) - \Big(\frac{2}{t} - \frac{4}{3}\Big) + O(t) = 0\cdot\frac{1}{t} + 2 + O(t).$$

The $1/t$ terms cancel (ranks $1 + 1 - 2 = 0$), exactly as the argument of (4.67) demands, and the constant is $\chi(S^2) = 2$.

Worked example 4.46: Gauss–Bonnet for every surface, from the heat kernel

Now let X be any closed oriented surface with any metric. For the de Rham operator [Bär Ex. 3.4.6]:

$$\begin{aligned} a_1 &\overset{(1)}{=} \operatorname{tr}\Phi_1^{\Delta_0} + \operatorname{tr}\Phi_1^{\Delta_2} - \operatorname{tr}\Phi_1^{\Delta_1} \\ &\overset{(2)}{=} \tfrac{1}{6}\mathrm{scal} + \tfrac{1}{6}\mathrm{scal} - \big(\tfrac{2}{6}\mathrm{scal} - \mathrm{scal}\big) \\ &= \mathrm{scal}. \end{aligned} \tag{4.68}$$

(1) $\Delta^+ = \Delta_0 \oplus \Delta_2$, $\Delta^- = \Delta_1$.

(2) $\Phi_1 = \frac{1}{6}\mathrm{scal}\,\mathrm{id} - \mathcal{K}$: on functions and on 2-forms (isomorphic to functions via $\star$) $\mathcal{K} = 0$, $r = 1$; on 1-forms $r = 2$ and $\operatorname{tr}\mathcal{K} = \operatorname{tr}\mathrm{Ric} = \mathrm{scal}$.

With $n = 2$, (4.67) and $\mathrm{scal} = 2K$ give

$$\chi(X) \overset{(1)}{=} \operatorname{index}(\mathrm{d} + \mathrm{d}^*) \overset{(2)}{=} \frac{1}{4\pi}\int_X \mathrm{scal}\,\mathrm{d}A \overset{(3)}{=} \frac{1}{2\pi}\int_X K\,\mathrm{d}A.$$

(1) (4.36).

(2) (4.67) with $n/2 = 1$ and $a_1 = \mathrm{scal}$.

(3) For a surface the scalar curvature is twice the Gaussian curvature.

This is the Gauss–Bonnet theorem, derived with no triangulation and no Euler class: only the pairing of eigenmodes and the first heat coefficient. Notice what happened: the *pointwise* density $a_1(x) = \mathrm{scal}(x)$ depends on the metric, but its integral cannot, because it equals an integer that does not move when the metric is deformed.

Worked example 4.47: Coverings multiply the index

If $\tilde{X} \to X$ is a k-sheeted covering and $\tilde{D}$ is the pull-back of D, then the local density $a_{n/2}$ of $\tilde{D}$ is the pull-back of that of D, and $\tilde{X}$ has k times the volume [Bär Cor. 3.4.8]:

$$\operatorname{index}\tilde{D} \overset{(1)}{=} (4\pi)^{-n/2}\int_{\tilde{X}} \tilde{a}_{n/2}\,\mathrm{dvol} \overset{(2)}{=} k\,(4\pi)^{-n/2}\int_X a_{n/2}\,\mathrm{dvol} = k\operatorname{index} D.$$

(1) (4.67) on $\tilde{X}$.
(2) Locally $\tilde{X}$ looks exactly like X, and each small patch of X is covered by k identical patches.

Example: $S^2 \to \mathbb{RP}^2$ is a double cover, so $2 = \chi(S^2) = 2\chi(\mathbb{RP}^2)$ and $\chi(\mathbb{RP}^2) = 1$ ($b_0 = 1$, $b_1 = b_2 = 0$ for the non-orientable $\mathbb{RP}^2$ with real coefficients, as it must be). Local formulas make such statements immediate; counting zero modes directly would not.

Physical picture: The miraculous cancellation

In (4.67) the coefficients $\int a_j$, $j < n/2$, vanish because the left side does not depend on t. For Dirac operators much more is true: the densities $a_j(x)$ themselves vanish *pointwise* for $j < n/2$, and $a_{n/2}(x)$ is exactly the degree-n part of the characteristic form $\hat{A}(X)\,\mathrm{ch}(F)$. This is the **local index theorem** (Patodi, Gilkey, Getzler). Getzler's proof rescales $t \to \varepsilon^2 t$ and Clifford generators $\gamma^\mu \to \varepsilon^{-1}\gamma^\mu$; in the limit $\varepsilon \to 0$ the Dirac Laplacian becomes a harmonic oscillator on the Clifford algebra, whose heat kernel (Mehler's formula) produces the $\hat{A}$ genus $\prod \frac{x_i/2}{\sinh(x_i/2)}$. Chapter T4 has the Clifford algebra background.

4.8.5 Physics: the Witten index, path integrals and anomalies

Physical picture: The Witten index as a path integral

In supersymmetric quantum mechanics with supercharge Q and Hamiltonian $H = Q^2$, the **Witten index** is $\mathrm{Tr}(-1)^F e^{-\beta H}$. By McKean–Singer it is independent of β and equals the index of D. Written as a Euclidean path integral, the trace becomes an integral over loops of period β, and the insertion of $(-1)^F$ makes the fermions *periodic* rather than antiperiodic around the loop:

$$\mathrm{Tr}(-1)^F e^{-\beta H} = \int_{\text{periodic}} [\mathrm{d}x][\mathrm{d}\psi]\, e^{-S_E[x,\psi]}.$$

As $\beta \to 0$ the loops shrink (their size is $\sim \sqrt{\beta}$, the same $\sqrt{t}$ as in Figure 4.23), and the path integral localises onto constant loops. What remains is an ordinary integral over X of the one-loop determinant around each constant loop: a characteristic form. For the de Rham complex (fermions $\psi^\mu \leftrightarrow \mathrm{d}x^\mu$, $Q = \mathrm{d} + \mathrm{d}^*$) this gives $\chi(X)$ as the integral of the Euler form; for the $N = 1$ theory whose supercharge is the Dirac operator it gives $\int \hat{A}$ (Alvarez-Gaumé, Friedan–Windey). This is the physicist's proof of the index theorem, and Example 4.46 is its rigorous two-dimensional shadow. Chapter 7 develops Witten's version for Morse theory.

Physical picture: Fujikawa and anomalies

In a Euclidean gauge theory with massless fermions, a chiral rotation $\psi \to e^{\mathrm{i}\alpha\gamma_5}\psi$ changes the fermion measure by a Jacobian $\exp\bigl(-2\mathrm{i}\alpha \sum_n \varphi_n^\dagger \gamma_5 \varphi_n\bigr)$, summed over eigenmodes of $\not{D}$. The sum is ill-defined and Fujikawa regularises it with exactly our factor: $\sum_n \varphi_n^\dagger \gamma_5 e^{-t\not{D}^2}\varphi_n = \mathrm{Str}\, e^{-t\not{D}^2}$, with grading γ_5. By McKean–Singer this is $n_+ - n_- = \mathrm{index}\,\not{D}$, independent of the regulator, and its small-t limit is the local density $\frac{1}{8\pi^2}\,\mathrm{tr}\, F \wedge F$ in four dimensions (up to the sign fixed by orientation). So the chiral anomaly *is* the index density. Chapter 10 and [QFT notes Ch. 22] have the details.

Try it: McKean–Singer for a finite complex

For the triangulated circle of Example 4.20, compute $\operatorname{Tr} e^{-t\Delta_0} - \operatorname{Tr} e^{-t\Delta_1}$ from the spectra and show it is 0 for all t.
Answer. Both Laplacians have spectrum $\{0, 3, 3\}$, so both traces equal $1 + 2e^{-3t}$ and the difference is $0 = \chi(S^1)$. In finite dimensions McKean–Singer at $t = 0$ is just (4.37): $\operatorname{Str}(1) = \dim V^0 - \dim V^1$.

Reading Nash (context)

Nash's §1 (pp. 89–97) states the theorem and the two definitions of the index but not the heat-equation method; he mentions the heat-equation proof of Atiyah–Bott–Patodi only in passing later in the chapter, in connection with zeta functions and the η-invariant (§7). We included it because it is the version physicists use, because it makes the "pairing" mechanism behind stability explicit, and because it is needed for the fixed-point formulas (heat-kernel localisation) and for anomalies in later parts. Sources: Bär §3.1–3.4, in particular (3.17)–(3.19), Theorem 3.4.4, Example 3.4.6 and Corollaries 3.4.7–3.4.8.

Recap

- D maps the λ-eigenspace of D^*D isomorphically onto that of DD^* for $\lambda > 0$: nonzero modes pair up (supersymmetry).
- Heat kernel $K_t(x, y) = \sum e^{-t\lambda_j} \varphi_j(x) \varphi_j(y)^\dagger$; heat trace $= \int \operatorname{tr} K_t(x, x)$; on $\mathbb{R}^n$ it is $(4\pi t)^{-n/2} e^{-|x-y|^2/4t}$.
- McKean–Singer: $\operatorname{index} D = \operatorname{Tr} e^{-tD^*D} - \operatorname{Tr} e^{-tDD^*} = \operatorname{Str} e^{-tQ^2}$ for all t; t-independence because the supertrace of an anticommutator of odd operators is zero.
- Small t: $K_t(x, x) \sim (4\pi t)^{-n/2} \sum t^j \Phi_j$ with local Φ_j; so $\operatorname{index} D = (4\pi)^{-n/2} \int a_{n/2}$, and 0 in odd dimensions.
- Checks: oscillator (1 for all t), S^1 (0), S^2 (2, including the $1/t + 1/3$ expansion); $a_1 = \operatorname{scal}$ gives Gauss–Bonnet; coverings multiply the index.
- Physics: Witten index $\operatorname{Tr}(-1)^F e^{-\beta H}$ and Fujikawa's regulated Jacobian are this supertrace.

4.9 From K-theory to numbers: the index formula in characteristic classes

The Atiyah–Singer theorem, as stated at the end of the previous section, says that two integers attached to an elliptic operator D agree: the *analytical index* $\dim \operatorname{Ker} D - \dim \operatorname{Coker} D$, which counts solutions, and the *topological index*, which is built out of the K-theory class $[\sigma_m(D)] \in K(T^*X)$ of the principal symbol by a chain of Thom isomorphisms and push-forwards. That is a beautiful statement, but as it stands it is not a formula anyone can evaluate. Nobody computes Thom homomorphisms in K-theory on a Tuesday afternoon.

What a physicist wants is something of the form

$$\operatorname{index} D \;=\; \int_X \big(\text{some polynomial in curvatures}\big),$$

because integrals of curvature polynomials are things we know how to do: we have done them

for the Gauss–Bonnet theorem, for the monopole charge and for the instanton number. This section explains how to get there. The tool is the **Chern character**, which turns K-theory classes into de Rham cohomology classes, i.e. into closed differential forms. Before we can state the formula we have to set up the characteristic classes it uses. Nash assumes all of this from his Chapter 3 (Chapter 3); we rebuild what we need, concretely, so that every later example in this part can be computed by hand.

4.9.1 Characteristic classes as curvature polynomials

Let $E \to X$ be a complex vector bundle of rank k with a connection. In a local trivialisation the connection is a matrix of 1-forms A and the curvature is the matrix of 2-forms

$$F = \mathrm{d}A + A \wedge A. \tag{4.69}$$

We use the mathematicians' convention, which is also Nash's: for a unitary connection A and F are *anti-Hermitian* matrices. The dictionary to the physicists' Hermitian gauge field is $A_{\text{math}} = -i\,g\,A^a_{\text{phys}}T^a$ with Hermitian generators T^a (see Chapter T3). Under a change of trivialisation $F \mapsto g^{-1}Fg$, so any *conjugation-invariant* polynomial in F is a globally defined form. The two basic ones are the determinant and the trace of the exponential.

Working definition 4.27 (Chern classes and Chern character)

Given the curvature F of a connection on E, form the matrix of 2-forms $\frac{i}{2\pi}F$ (it is Hermitian, so its "eigenvalues" are real). Then

$$c(E) = \det\!\Big(1 + \tfrac{i}{2\pi}F\Big) = 1 + c_1(E) + c_2(E) + \cdots, \qquad c_j(E) \in \Omega^{2j}(X),$$
$$\mathrm{ch}(E) = \operatorname{tr}\exp\!\Big(\tfrac{i}{2\pi}F\Big) = \mathrm{ch}_0(E) + \mathrm{ch}_1(E) + \mathrm{ch}_2(E) + \cdots, \qquad \mathrm{ch}_j(E) \in \Omega^{2j}(X).$$

Expand, keep the part of each degree, and you have explicit closed forms. Their de Rham classes [two closed forms count as the same if they differ by d of a form] do not depend on the connection chosen. *Formally:* c_j and ch_j are the Chern–Weil images of the invariant polynomials "j-th elementary symmetric function" [the sum of all products of j distinct eigenvalues] and "$\frac{1}{j!}\operatorname{tr}(\cdot)^j$".

Why are these closed, and why is their class independent of A? The mechanism is the Bianchi identity $\mathrm{d}F = F\wedge A - A\wedge F$. For example

$$\begin{aligned}\mathrm{d}\,\operatorname{tr}(F\wedge F) &\overset{(1)}{=} \operatorname{tr}(\mathrm{d}F\wedge F) + \operatorname{tr}(F\wedge \mathrm{d}F)\\ &\overset{(2)}{=} \operatorname{tr}\big((FA-AF)F\big) + \operatorname{tr}\big(F(FA-AF)\big)\\ &\overset{(3)}{=} \operatorname{tr}(FAF) - \operatorname{tr}(AFF) + \operatorname{tr}(FFA) - \operatorname{tr}(FAF) = 0.\end{aligned}$$

(1) The Leibniz rule; F is a 2-form, so no sign appears when d passes it.
(2) The Bianchi identity, obtained by applying d to (4.69).
(3) Expand; then use cyclicity of the trace, which holds for matrices of forms up to the sign $(-1)^{\deg\cdot\deg}$; moving a 1-form past a 4-form costs no sign, so $\operatorname{tr}(FFA) = \operatorname{tr}(AFF)$ and everything cancels.

The independence of the connection is the same computation done for a one-parameter family A_t: the t-derivative of $\operatorname{tr}F_t^j$ turns out to be d of something (the Chern–Simons form), so the

class does not move. We will use this repeatedly: *we may compute characteristic numbers with any connection we like*, usually the most symmetric one.

Expanding the definitions to low order gives the formulas we will use again and again. With $\Phi := \frac{i}{2\pi}F$,

$$\begin{aligned}\det(1+\Phi) &\overset{(1)}{=} \exp \operatorname{tr}\log(1+\Phi)\\ &\overset{(2)}{=} \exp\Big(\operatorname{tr}\Phi - \tfrac{1}{2}\operatorname{tr}\Phi^2 + \cdots\Big)\\ &\overset{(3)}{=} 1 + \operatorname{tr}\Phi + \tfrac{1}{2}\big[(\operatorname{tr}\Phi)^2 - \operatorname{tr}\Phi^2\big] + \cdots .\end{aligned} \tag{4.70}$$

(1) The identity $\det M = \exp\operatorname{tr}\log M$, valid for matrices near the identity; 2-forms commute, so the matrix entries behave like ordinary commuting numbers here.
(2) The Taylor series $\log(1+\Phi) = \Phi - \Phi^2/2 + \cdots$.
(3) Expand the exponential and collect terms of form degree 2 and 4.

Reading off,

$$c_1 = \frac{i}{2\pi}\operatorname{tr}F, \qquad c_2 = \frac{1}{2}\Big(\frac{i}{2\pi}\Big)^2\big[\operatorname{tr}F\wedge\operatorname{tr}F - \operatorname{tr}(F\wedge F)\big] \;=\; \frac{1}{8\pi^2}\big[\operatorname{tr}(F\wedge F) - \operatorname{tr}F\wedge\operatorname{tr}F\big], \tag{4.71}$$

and, from $\operatorname{tr}e^{\Phi} = k + \operatorname{tr}\Phi + \frac{1}{2}\operatorname{tr}\Phi^2 + \cdots$,

$$\mathrm{ch}_0 = k, \qquad \mathrm{ch}_1 = c_1, \qquad \mathrm{ch}_2 = \tfrac{1}{2}\operatorname{tr}\Phi^2 = -\frac{1}{8\pi^2}\operatorname{tr}(F\wedge F) = \tfrac{1}{2}c_1^2 - c_2. \tag{4.72}$$

For an SU(2) gauge field $\operatorname{tr}F = 0$, so $c_2 = \frac{1}{8\pi^2}\int \operatorname{tr}F\wedge F$ is the instanton number k of Chapter 8 (with our orientation), and $\int \mathrm{ch}_2 = -k$.

4.9.2 Chern roots and the splitting principle

Every formula above is a symmetric function of the eigenvalues of the Hermitian matrix $\Phi = \frac{i}{2\pi}F$. At each point we can (formally) diagonalise it:

$$\Phi \sim \begin{pmatrix} x_1 & & \\ & \ddots & \\ & & x_k \end{pmatrix},$$

where the x_j are commuting 2-forms, the **Chern roots** of E. They need not exist as honest global forms, but every *symmetric* expression in them does, because a symmetric polynomial in eigenvalues is a polynomial in the traces $\operatorname{tr}\Phi^j$.

Working definition 4.28 (Chern roots; splitting principle)

To compute any characteristic class of E, pretend that $E = L_1 \oplus \cdots \oplus L_k$ is a sum of line bundles [rank-one bundles, one complex dimension in each fibre] with $c_1(L_j) = x_j$. Then

$$c(E) = \prod_{j=1}^{k}(1 + x_j), \qquad \mathrm{ch}(E) = \sum_{j=1}^{k} e^{x_j},$$

compute whatever you need as a symmetric function [unchanged by any permutation of the x_j] of the x_j, and at the end rewrite it in terms of the elementary symmetric functions $c_1 = \sum x_j$, $c_2 = \sum_{i<j} x_i x_j$, ... Any identity proved this way holds for all bundles (this is

the **splitting principle**).

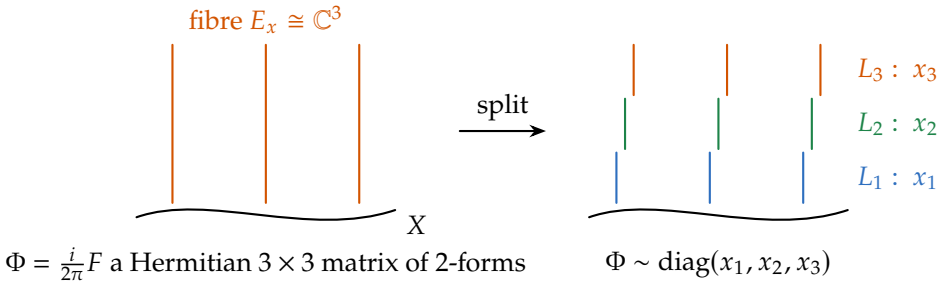


Figure 4.26: The splitting principle. Left: a rank-3 bundle, whose curvature is a Hermitian matrix of 2-forms. Right: for the purpose of computing characteristic classes we may pretend the bundle is a sum of three line bundles whose first Chern classes x_1, x_2, x_3 are the eigenvalues of Φ. Every symmetric function of the x_j is a genuine global form.

Worked example 4.48: The splitting principle reproduces (4.72)

For a rank-2 bundle with Chern roots x_1, x_2,

$$\begin{aligned} \mathrm{ch}_2(E) &\overset{(1)}{=} \tfrac{1}{2}\left(x_1^2 + x_2^2\right) \\ &\overset{(2)}{=} \tfrac{1}{2}\left[(x_1+x_2)^2 - 2x_1x_2\right] \\ &\overset{(3)}{=} \tfrac{1}{2}c_1^2 - c_2. \end{aligned}$$

(1) $\mathrm{ch}(E) = e^{x_1} + e^{x_2}$; the degree-4 part of e^x is $x^2/2$.
(2) The algebraic identity $a^2 + b^2 = (a+b)^2 - 2ab$.
(3) $c_1 = x_1 + x_2$ and $c_2 = x_1x_2$ are the elementary symmetric functions.

The answer agrees with the curvature computation (4.72), as it must: both are $\frac{1}{2}\operatorname{tr}\Phi^2$.

Two properties of ch make it the right bridge from K-theory, and the splitting principle makes them obvious:

$$\begin{aligned} \mathrm{ch}(E \oplus E') &\overset{(1)}{=} \sum_j e^{x_j} + \sum_l e^{x'_l} = \mathrm{ch}\,E + \mathrm{ch}\,E', \\ \mathrm{ch}(E \otimes E') &\overset{(2)}{=} \sum_{j,l} e^{x_j + x'_l} = \Big(\sum_j e^{x_j}\Big)\Big(\sum_l e^{x'_l}\Big) = \mathrm{ch}\,E\ \mathrm{ch}\,E'. \end{aligned} \tag{4.73}$$

(1) The Chern roots of a direct sum are the union of the two lists (the curvature is block diagonal).
(2) The Chern roots of a tensor product are all sums $x_j + x'_l$, because $L \otimes L'$ has $c_1(L\otimes L') = c_1(L) + c_1(L')$ (the connections add: $A_{L\otimes L'} = A_L + A_{L'}$).

So ch turns $\oplus$ into $+$ and $\otimes$ into $\times$. A K-theory class is a formal difference $[E]-[E']$ of bundles, so $\mathrm{ch}([E]-[E']) = \mathrm{ch}\,E - \mathrm{ch}\,E'$ is well defined, and ch is a ring homomorphism $K(X) \to H^{\mathrm{even}}(X;\mathbb{Q})$. This is precisely the mechanism Nash asks for on [Nash p. 97–98]: a way of passing from K-theory to cohomology.

Two more rules we will need. For the dual bundle E^* the connection is $-A^T$, so its Chern roots are $-x_j$:

$$c_j(E^*) = (-1)^j c_j(E), \qquad \mathrm{ch}(E^*) = \sum_j e^{-x_j}. \tag{4.74}$$

The complex conjugate bundle $\bar{E}$ has connection $\bar{A} = -A^T$ (for a unitary connection), so $\bar{E} \cong E^*$ as far as characteristic classes are concerned. And for exterior powers,

$$E = L_1 \oplus \cdots \oplus L_k \implies \Lambda^p E = \bigoplus_{i_1 < \cdots < i_p} L_{i_1} \otimes \cdots \otimes L_{i_p}, \tag{4.75}$$

which is just the statement that $e_{i_1} \wedge \cdots \wedge e_{i_p}$, $i_1 < \cdots < i_p$, is a basis of $\Lambda^p \mathbb{C}^k$ when $e_1, \dots, e_k$ is a basis of $\mathbb{C}^k$. We use (4.75) heavily in Section 4.12.

4.9.3 Multiplicative classes: Todd, $\hat{A}$ and L

The Chern character is *additive*: a sum over Chern roots. The other classes in the index theorem are *multiplicative*: a product over Chern roots of one fixed power series.

Working definition 4.29 (multiplicative class of a power series)

Pick a power series $Q(x) = 1 + q_1 x + q_2 x^2 + \cdots$. For a bundle with Chern roots $x_1, \dots, x_k$ define

$$Q(E) := \prod_{j=1}^{k} Q(x_j),$$

expand, and rewrite each degree in terms of $c_1, c_2, \dots$. Then $Q(E \oplus E') = Q(E)\,Q(E')$ [the Chern roots of $E \oplus E'$ are those of E together with those of E'].

The **Todd class** is the multiplicative class of

$$\frac{x}{1 - e^{-x}} = 1 + \frac{x}{2} + \frac{x^2}{12} - \frac{x^4}{720} + \cdots. \tag{4.76}$$

(The x^3 coefficient vanishes because $\frac{x}{1-e^{-x}} - \frac{x}{2} = \frac{x}{2} \coth \frac{x}{2}$ is even.) Multiplying out and collecting,

$$\begin{aligned} \operatorname{td}(E) &\overset{(1)}{=} \prod_j \left(1 + \tfrac{1}{2}x_j + \tfrac{1}{12}x_j^2 + 0 \cdot x_j^3 + \cdots\right) \\ &\overset{(2)}{=} 1 + \tfrac{1}{2}\sum_j x_j + \left[\tfrac{1}{12}\sum_j x_j^2 + \tfrac{1}{4}\sum_{i<j} x_i x_j\right] + \left[\tfrac{1}{24}\sum_{i \neq j} x_i^2 x_j + \tfrac{1}{8}\sum_{i<j<l} x_i x_j x_l\right] + \cdots \\ &\overset{(3)}{=} 1 + \tfrac{1}{2}c_1 + \tfrac{1}{12}\left(c_1^2 + c_2\right) + \tfrac{1}{24}c_1 c_2 + \cdots. \end{aligned} \tag{4.77}$$

(1) Definition of the multiplicative class with the series (4.76).

(2) Collect terms of polynomial degree 1, 2, 3 in the x_j (form degrees 2, 4, 6): either one factor contributes its x^2 or x^3 term, or several factors contribute their linear term.

(3) Use $\sum x_j^2 = c_1^2 - 2c_2$, so $\frac{1}{12}(c_1^2 - 2c_2) + \frac{1}{4}c_2 = \frac{1}{12}(c_1^2 + c_2)$; and $\sum_{i \neq j} x_i^2 x_j = c_1 c_2 - 3c_3$, so $\frac{1}{24}(c_1 c_2 - 3c_3) + \frac{1}{8}c_3 = \frac{1}{24}c_1 c_2$.

For the dual bundle, (4.74) gives

$$\operatorname{td}(E^*) = \prod_{j=1}^{k} \frac{-x_j}{1 - e^{x_j}}. \tag{4.78}$$

Real bundles, Pontryagin classes and the Euler class

The index formula involves the *complexified* tangent bundle $TX_{\mathbb{C}} = TX \otimes_{\mathbb{R}} \mathbb{C}$ of a real manifold. A complexified real bundle is isomorphic to its own conjugate (complex conjugation of the $\mathbb{C}$

factor), hence to its own dual. By (4.74) this forces $c_j(TX_{\mathbb{C}}) = (-1)^j c_j(TX_{\mathbb{C}})$, so all odd Chern classes vanish (as de Rham classes). In terms of Chern roots: they come in pairs $\pm x_j$.

We can see this directly from curvature. Take a Riemannian metric on X (dimension $n = 2m$) and its Levi-Civita connection in an orthonormal frame. The curvature $R = (R^a{}_b)$ is a *real antisymmetric* $n \times n$ matrix of 2-forms. A real antisymmetric matrix can be brought to block-diagonal form by a rotation:

$$\frac{R}{2\pi} \sim \begin{pmatrix} 0 & x_1 & & & & \\ -x_1 & 0 & & & & \\ & & \ddots & & & \\ & & & 0 & x_m \\ & & & -x_m & 0 \end{pmatrix}. \tag{4.79}$$

Each block $\left(\begin{smallmatrix} 0 & x \\ -x & 0 \end{smallmatrix}\right)$ has eigenvalues $\pm ix$, so $\Phi = \frac{i}{2\pi}R$ has eigenvalues $\mp x$: the Chern roots of $TX_{\mathbb{C}}$ are $\pm x_1, \dots, \pm x_m$. Physically, each x_j is the "rotation angle" (really a 2-form) of the curvature in the j-th 2-plane.

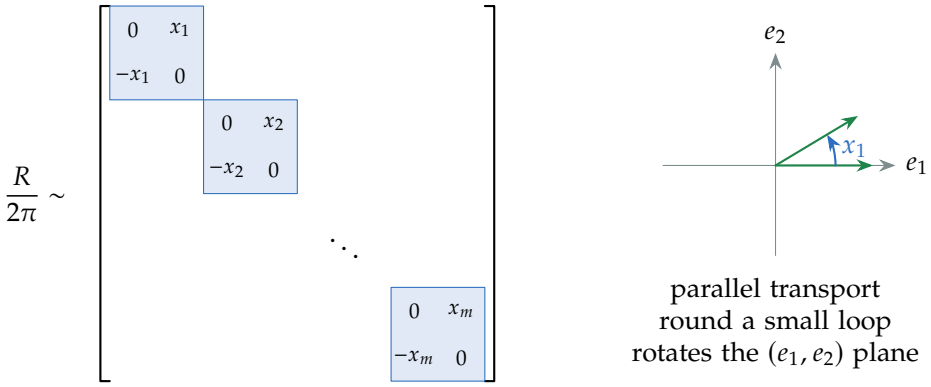


Figure 4.27: Left: the curvature of a Riemannian manifold, divided by 2π, is a real antisymmetric matrix of 2-forms; it can be brought to 2×2 blocks. Right: the j-th block describes an infinitesimal rotation of the j-th 2-plane, by an "angle" x_j that is a 2-form. The x_j play the role of Chern roots for real bundles.

Working definition 4.30 (Pontryagin and Euler classes)

For a real oriented bundle V [with a chosen orientation of every fibre, which fixes the sign of $e(V)$] of rank $2m$ with curvature in the block form (4.79) [$R/2\pi$ rotated into 2×2 blocks with entries $\pm x_j$]:

$$p(V) = \prod_{j=1}^{m}(1 + x_j^2), \qquad p_i(V) = (-1)^i c_{2i}(V_{\mathbb{C}}), \quad \text{e.g. } p_1(V) = \sum_j x_j^2 = -\frac{1}{8\pi^2}\operatorname{tr}(R \wedge R),$$

$$e(V) = x_1 x_2 \cdots x_m = \operatorname{Pf}\left(\frac{R}{2\pi}\right), \qquad \text{so } e(V)^2 = p_m(V).$$

Here $V_{\mathbb{C}} = V \otimes \mathbb{C}$ [the same fibres and transition matrices, now acting on complex vectors]. The Pfaffian of an antisymmetric $2m \times 2m$ matrix is $\operatorname{Pf}(M) = \frac{1}{2^m m!}\epsilon_{a_1 \cdots a_{2m}} M^{a_1 a_2} \cdots M^{a_{2m-1} a_{2m}}$; on the block form it is the product of the upper-right entries.

Let us check the formula for p_1 in terms of R, because we will use it for the signature and

the $\hat{A}$-genus:

$$\begin{aligned} p_1(V) &\overset{(1)}{=} -c_2(V_{\mathbb{C}}) \\ &\overset{(2)}{=} -\frac{1}{8\pi^2}\left[\operatorname{tr}(R\wedge R) - \operatorname{tr} R \wedge \operatorname{tr} R\right] \\ &\overset{(3)}{=} -\frac{1}{8\pi^2}\operatorname{tr}(R\wedge R). \end{aligned}$$

(1) Definition of p_1 with $i = 1$.
(2) Formula (4.71) with F replaced by the real curvature R (which is also anti-Hermitian, being real antisymmetric).
(3) R is antisymmetric, so $\operatorname{tr} R = 0$.

And in terms of the roots: $\operatorname{tr}(R\wedge R)/(2\pi)^2$ on the block form is $\sum_j 2\cdot x_j\cdot(-x_j) = -2\sum_j x_j^2$, so $p_1 = -\frac{1}{8\pi^2}\cdot(-8\pi^2)\sum_j x_j^2 = \sum_j x_j^2$. The two descriptions agree.

Finally, the relation between the Euler class and the top Chern class of the complexification (Nash's (3.124)), which the de Rham calculation will need:

$$\begin{aligned} c_{2m}(V_{\mathbb{C}}) &\overset{(1)}{=} \prod_{j=1}^{m} x_j(-x_j) \\ &\overset{(2)}{=} (-1)^m\Big(\prod_j x_j\Big)^2 = (-1)^m e(V)^2. \end{aligned} \tag{4.80}$$

(1) The top Chern class is the product of all Chern roots, and the roots of $V_{\mathbb{C}}$ are $\pm x_j$.
(2) Pull out one sign per pair; the remaining product is $e(V)$ squared.

The genera that appear in the index theorem

Here is the key simplification that Nash states as (4.32). The Todd class of a complexified tangent bundle is a perfect square:

$$\begin{aligned} \operatorname{td}(TX_{\mathbb{C}}) &\overset{(1)}{=} \prod_{j=1}^{m} \frac{x_j}{1-e^{-x_j}}\cdot\frac{-x_j}{1-e^{x_j}} \\ &\overset{(2)}{=} \prod_{j=1}^{m} \frac{x_j^2}{(e^{x_j/2}-e^{-x_j/2})^2} \\ &\overset{(3)}{=} \prod_{j=1}^{m}\Big(\frac{x_j/2}{\sinh(x_j/2)}\Big)^2 \;=\; \hat{A}(X)^2. \end{aligned} \tag{4.81}$$

(1) The Chern roots of $TX_{\mathbb{C}}$ are $\pm x_j$; insert them into the Todd series (4.76).
(2) $(1-e^{-x})(1-e^{x}) = 2 - e^{x} - e^{-x} = -(e^{x/2}-e^{-x/2})^2$, and the numerator is $-x^2$, so the two minus signs cancel.
(3) $e^{x/2}-e^{-x/2} = 2\sinh(x/2)$; the product of $(x/2)/\sinh(x/2)$ is by definition the $\hat{A}$-class, which we meet properly in Section 4.15.

Now expand. With $t = x/2$, $t/\sinh t = 1 - \frac{t^2}{6} + \frac{7t^4}{360} - \cdots$, so

$$\frac{x/2}{\sinh(x/2)} = 1 - \frac{x^2}{24} + \frac{7x^4}{5760} - \cdots, \tag{4.82}$$

and therefore

$$\hat{A}(X) \overset{(1)}{=} 1 - \frac{1}{24}\sum_j x_j^2 + \Big[\frac{7}{5760}\sum_j x_j^4 + \frac{1}{576}\sum_{i<j} x_i^2 x_j^2\Big] + \cdots$$
$$\overset{(2)}{=} 1 - \frac{p_1}{24} + \frac{7(p_1^2 - 2p_2) + 10p_2}{5760} + \cdots = 1 - \frac{p_1}{24} + \frac{7p_1^2 - 4p_2}{5760} + \cdots . \tag{4.83}$$

(1) Multiply the series (4.82) over j and collect degrees.
(2) In the variables x_j^2, $p_1 = \sum x_j^2$ and $p_2 = \sum_{i<j} x_i^2 x_j^2$, so $\sum x_j^4 = p_1^2 - 2p_2$; and $\frac{1}{576} = \frac{10}{5760}$.

Squaring,

$$\mathrm{td}(TX_{\mathbb{C}}) \overset{(1)}{=} 1 - 2\cdot\frac{p_1}{24} + \Big[2\cdot\frac{7p_1^2 - 4p_2}{5760} + \frac{p_1^2}{576}\Big] + \cdots$$
$$\overset{(2)}{=} 1 - \frac{p_1}{12} + \frac{24p_1^2 - 8p_2}{5760} + \cdots = 1 - \frac{1}{12}p_1 + \frac{1}{720}(3p_1^2 - p_2) + \cdots , \tag{4.84}$$

(1) $(1 + a + b)^2 = 1 + 2a + (2b + a^2) + \cdots$ with $a = -p_1/24$ and b the degree-8 term.
(2) $14 + 10 = 24$ and $24/5760 = 3/720$, $8/5760 = 1/720$.

which is exactly Nash's (4.32). Only Pontryagin classes appear, as they must for a real bundle.

class	series (per root)	low-order expansion	where it appears
$\mathrm{ch}(E)$ (additive)	$\sum e^{x_j}$	$k + c_1 + \frac{1}{2}(c_1^2 - 2c_2)$	every twisted index
$\mathrm{td}(E)$	$\dfrac{x}{1 - e^{-x}}$	$1 + \frac{1}{2}c_1 + \frac{1}{12}(c_1^2 + c_2) + \frac{1}{24}c_1c_2$	Dolbeault, (4.31)
$\hat{A}(V)$	$\dfrac{x/2}{\sinh(x/2)}$	$1 - \frac{1}{24}p_1 + \frac{1}{5760}(7p_1^2 - 4p_2)$	Dirac
$\mathcal{L}(V)$ (Nash)	$\dfrac{x/2}{\tanh(x/2)}$	$1 + \frac{1}{12}p_1 + \frac{1}{720}(7p_2 - p_1^2)$	signature
$e(V)$	x (product of m roots)	$x_1 \cdots x_m = \mathrm{Pf}(R/2\pi)$	Gauss–Bonnet

Table 4.2: The characteristic classes of the four classical index theorems, each given by one power series in the Chern roots. For real bundles, the roots are the block "angles" x_j of (4.79). The $\mathcal{L}$ row uses Nash's normalisation; we derive it in Section 4.14.

4.9.4 The index formula

We can now read Nash's cohomological formula. Recall from the previous section that an elliptic operator D of order m on an n-manifold X determines a class $[\sigma_m(D)] \in K(T^*X)$ (K-theory with compact support on the cotangent bundle). Apply the Chern character to get a compactly supported cohomology class $\mathrm{ch}[\sigma_m(D)]$ on T^*X. Pull back the Todd class of $TX_{\mathbb{C}}$ to T^*X (or TX; for our purposes the metric identifies them). Multiply, and integrate over the $2n$-dimensional total space.

Key idea: The cohomological index formula, Nash (4.31)

For an elliptic operator D of order m on a compact n-manifold X,

$$\mathrm{index}\, D = (-1)^n \big\{\mathrm{ch}\big([\sigma_m(D)]\big)\cdot \mathrm{td}(TX_{\mathbb{C}})\big\}[TX], \tag{4.85}$$

where $[TX]$ is the fundamental class of the total space of the tangent bundle and $\{\cdots\}[TX]$ means "evaluate on it", i.e. integrate the top-degree ($2n$-form) part of the integrand over TX. The integrand has compact support in the fibre directions because the symbol complex is exact away from the zero section.

It is worth pausing on what this formula claims. The right-hand side is built from ch and td, whose power series have denominators $2, 12, 24, 720, \ldots$. It is a number in rational cohomology. Nothing in its definition makes it an integer. Yet it equals $\dim \operatorname{Ker} D - \dim \operatorname{Coker} D$, which is an integer. Nash calls this "far from obvious". Indeed, whole theorems will fall out of this integrality alone: Rokhlin's theorem in Section 4.15.9 is one. Historically, the same combination $\mathrm{ch} \cdot \mathrm{td}$ had already appeared in Hirzebruch's Riemann–Roch theorem for complex manifolds (1954), which is the special case we derive in Section 4.13.

Physical picture: Why characteristic classes at all

The index is stable: it does not change when D is deformed continuously through elliptic operators. So it can only depend on topological data, and the only topological data around are the bundles E^p the operator acts on and the tangent bundle of X. Characteristic classes are exactly the topological invariants of bundles that can be written as local integrals. The index theorem says: the global count of zero modes is a local integral. This is the same philosophy as Gauss–Bonnet, and, in physics, as the anomaly: a global quantity (a count of states, a phase) fixed by a local density.

In practice we will never integrate over TX directly. The next two sections show how to push the computation down to X itself, where it becomes $\int_X$(curvature polynomial). Before that, three computations that we will reuse.

Worked example 4.49: Characteristic classes of $S^2 = \mathbb{CP}^1$

Think of S^2 as the Riemann sphere. Its holomorphic tangent bundle $T = T^{1,0}S^2$ is a complex line bundle, with one Chern root $y = c_1(T)$. For the round metric of radius 1, the Levi-Civita connection on T has curvature $F = -i\,K\,\mathrm{d}A$ with Gaussian curvature $K = 1$ (this is just the rotation of the tangent plane under parallel transport), so

$$\int_{S^2} c_1(T) \overset{(1)}{=} \int_{S^2} \frac{i}{2\pi}\left(-iK\,\mathrm{d}A\right) \overset{(2)}{=} \frac{1}{2\pi}\int_{S^2} \mathrm{d}A = \frac{4\pi}{2\pi} = 2.$$

(1) Definition of c_1 for a line bundle (tr of a 1×1 matrix).
(2) $K = 1$ for the unit sphere; the area is 4π.

So $T \cong \mathcal{O}(2)$, the line bundle of degree 2. The underlying real bundle TS^2 has one block angle $x_1 = y$, so $e(TS^2) = y$ with $\int e = 2$, and $p_1(TS^2) = y^2 = 0$ since there are no 4-forms on a surface. The Todd class $\mathrm{td}(T) = 1 + \frac{1}{2}y$ integrates to 1.

Worked example 4.50: Pontryagin class of $\mathbb{CP}^2$

Let h be the positive generator of $H^2(\mathbb{CP}^2;\mathbb{Z})$, the Poincaré dual of a line $\mathbb{CP}^1 \subset \mathbb{CP}^2$, so $\int_{\mathbb{CP}^2} h^2 = 1$ (two lines meet in one point). A standard fact (the Euler sequence; Bär's Example

4.3.4) is that the holomorphic tangent bundle T has total Chern class $c(T) = (1+h)^3$. Then

$$\begin{aligned} p_1(\mathbb{CP}^2) &\overset{(1)}{=} \sum_j y_j^2 \\ &\overset{(2)}{=} c_1(T)^2 - 2c_2(T) \\ &\overset{(3)}{=} (3h)^2 - 2\cdot 3h^2 = 3h^2, \qquad \text{so} \quad \int_{\mathbb{CP}^2} p_1 = 3. \end{aligned}$$

(1) For a complex manifold, $TX_{\mathbb{C}} = T \oplus \bar{T}$ has Chern roots y_j (of T) and $-y_j$ (of $\bar{T} \cong T^*$), so the "block angles" are the y_j and $p_1 = \sum y_j^2$.

(2) $\sum y_j^2 = (\sum y_j)^2 - 2\sum_{i<j} y_i y_j$.

(3) Expand $(1+h)^3 = 1 + 3h + 3h^2$: $c_1 = 3h$, $c_2 = 3h^2$.

Also $\int e(\mathbb{CP}^2) = \int c_2(T) = 3$, which is $\chi(\mathbb{CP}^2) = 1+1+1$ (cells in dimensions 0, 2, 4). Both numbers will be used below.

Worked example 4.51: The round S^4: $p_1 = 0$, $\int e = 2$

For the unit sphere S^n the curvature 2-forms in an orthonormal coframe e^a are $R^{ab} = e^a \wedge e^b$ (constant sectional curvature 1). Then

$$\operatorname{tr}(R \wedge R) = \sum_{a,b} R^{ab} \wedge R^{ba} = \sum_{a,b} e^a \wedge e^b \wedge e^b \wedge e^a = 0,$$

because $e^b \wedge e^b = 0$. So $p_1(S^4) = 0$. (More generally $TS^n \oplus \mathbb{R} = \mathbb{R}^{n+1}$ is trivial, so $p(TS^n) = 1$.) The Euler form is not zero:

$$\begin{aligned} \int_{S^4} e &\overset{(1)}{=} \int_{S^4} \frac{1}{2^2\, 2!\, (2\pi)^2} \epsilon_{abcd} R^{ab} \wedge R^{cd} \\ &\overset{(2)}{=} \frac{1}{32\pi^2} \int_{S^4} \epsilon_{abcd}\, e^a \wedge e^b \wedge e^c \wedge e^d \\ &\overset{(3)}{=} \frac{24}{32\pi^2} \operatorname{vol}(S^4) = \frac{24}{32\pi^2} \cdot \frac{8\pi^2}{3} = 2. \end{aligned}$$

(1) The Pfaffian formula with $m = 2$ applied to $R/2\pi$.

(2) $R^{ab} = e^a \wedge e^b$ for the unit sphere.

(3) Each of the $4! = 24$ orderings of $(1,2,3,4)$ contributes $+e^1 \wedge e^2 \wedge e^3 \wedge e^4$, the volume form; the volume of the unit 4-sphere is $8\pi^2/3$.

And $\chi(S^4) = 2$. We have just checked the four-dimensional Gauss–Bonnet theorem on one example; Section 4.12 derives it from the index theorem.

Watch out: Nash's $[TX]$ and $[X]$

Nash writes $\{\cdots\}[TX]$ in (4.31) and later $\{\cdots\}[X]$. Both mean "take the top-degree component and integrate". The difference is which space you integrate over: the $2n$-dimensional total space TX (whose orientation is canonical; this is why no orientation of X is needed in (4.31)), or X itself (which requires X oriented). The sign factors $(-1)^n$, $(-1)^{n(n+1)/2}$, $(-1)^{n/2}$ that Nash carries around are bookkeeping of these orientation conventions. We will always fix them at the end by checking a known case.

Reading Nash (pp. 97–98, (4.31)–(4.32))

Nash's §2 opens by saying that the K-theoretic index can be translated into cohomology with the Chern character, and quotes the result (4.31) from Atiyah–Singer III, calling the translation "slightly lengthy but routine". He then notes that $TX_{\mathbb{C}}$ is self-conjugate, so odd Chern classes vanish, and quotes the Todd expansion (4.32) in Pontryagin classes, referring to his (3.106) and (3.108). We have derived (4.32) from $\mathrm{td}(TX_{\mathbb{C}}) = \hat{A}(X)^2$, eq. (4.84).

Recap

- Characteristic classes are closed curvature polynomials: $c(E) = \det(1 + \frac{i}{2\pi}F)$, $\mathrm{ch}(E) = \operatorname{tr} e^{iF/2\pi}$.
- Splitting principle: compute with Chern roots x_j as if E were a sum of line bundles; ch is additive and multiplicative, so it maps K-theory to cohomology.
- Real bundles: roots $\pm x_j$ (block angles of $R/2\pi$); $p = \prod(1 + x_j^2)$, $e = \prod x_j = \mathrm{Pf}(R/2\pi)$, $c_{2m}(V_{\mathbb{C}}) = (-1)^m e^2$.
- $\mathrm{td}(TX_{\mathbb{C}}) = \hat{A}(X)^2 = 1 - \frac{1}{12}p_1 + \frac{1}{720}(3p_1^2 - p_2) + \cdots$.
- Index formula: $\operatorname{index} D = (-1)^n\{\mathrm{ch}[\sigma(D)]\ \mathrm{td}(TX_{\mathbb{C}})\}[TX]$, an integer although it is built from rational classes.

4.10 Two vanishing results: odd dimensions and operators on functions

Before computing indices we should know when not to bother. Nash's first two examples are *vanishing theorems*: situations where the index formula forces $\operatorname{index} D = 0$ without any calculation, purely by a symmetry or a dimension count. They are also good practice in reading the formula (4.85): what depends on what, and in which degree.

The two statements are:

1. On an *odd*-dimensional manifold every elliptic *differential* operator has index zero.

2. An elliptic differential operator acting on (scalar) *functions*, i.e. between trivial line bundles, always has index zero; and even a *pseudo-differential* one has index zero once $\dim X > 1$.

Both have an exception for pseudo-differential operators in dimension one, and we shall compute that exception explicitly.

4.10.1 Symbols are homogeneous polynomials in the momentum

Recall (Chapter 2) how the principal symbol is read off. Write the operator in local coordinates, $D = \sum_{|\alpha|\le m} a_\alpha(x)\,\partial^\alpha$ with multi-index notation $\partial^\alpha = \partial_1^{\alpha_1}\cdots\partial_n^{\alpha_n}$, replace each ∂_μ by ip_μ (the Fourier rule: $\partial_\mu e^{ip\cdot x} = ip_\mu e^{ip\cdot x}$), and keep only the top-order terms:

$$\sigma_m^D(x,p) = \sum_{|\alpha|=m} a_\alpha(x)\,(ip)^\alpha. \tag{4.86}$$

(Nash absorbs the powers of i into a_α and writes $\sum_{|\alpha|=m} a_\alpha(x)p^\alpha$.) Each term is a monomial of degree exactly m in p. Hence the **homogeneity** property

$$\sigma_m^D(x,\lambda p) = \lambda^m\,\sigma_m^D(x,p) \qquad \text{for all real } \lambda. \tag{4.87}$$

The important case for us is $\lambda = -1$:

$$\sigma_m^D(x,-p) = (-1)^m\,\sigma_m^D(x,p). \tag{Nash 4.33}$$

For example, the Laplacian $-\sum_\mu \partial_\mu^2$ has symbol $|p|^2$ (even, $m=2$); the operator $\mathrm{d}/\mathrm{d}\theta$ on the circle has symbol ip (odd, $m=1$); the Dirac operator has symbol $i\gamma^\mu p_\mu$ (odd, $m=1$). All obey (Nash 4.33). What fails for pseudo-differential operators is precisely this: their symbols are only required to be homogeneous for $\lambda > 0$, so $|p|$ is allowed, and $|-p| = |p|$ is not $(-1)^1\,|p|$.

4.10.2 Why the index vanishes in odd dimensions

The idea is to apply the **antipodal map**

$$R : T^*X \to T^*X, \qquad (x,p) \mapsto (x,-p)$$

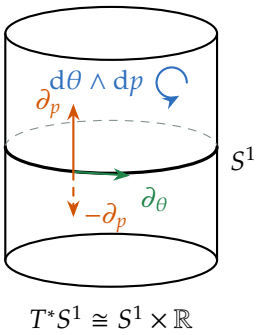
to the index formula. It changes the integrand in a controlled way (by (Nash 4.33)), and it may reverse the orientation of the space we integrate over. Let us see when.

In local coordinates $(x^1,\dots,x^n,p_1,\dots,p_n)$ on T^*X, the canonical orientation is given by the volume form $\mathrm{d}x^1\wedge\cdots\wedge\mathrm{d}x^n\wedge\mathrm{d}p_1\wedge\cdots\wedge\mathrm{d}p_n$ (any other ordering convention changes it by a fixed sign, which does not matter here). Pull it back by R:

$$\begin{aligned} R^*\big(\mathrm{d}x^1\wedge\cdots\wedge\mathrm{d}x^n\wedge\mathrm{d}p_1\wedge\cdots\wedge\mathrm{d}p_n\big) &\overset{(1)}{=} \mathrm{d}x^1\wedge\cdots\wedge\mathrm{d}x^n\wedge\mathrm{d}(-p_1)\wedge\cdots\wedge\mathrm{d}(-p_n) \\ &\overset{(2)}{=} (-1)^n\,\mathrm{d}x^1\wedge\cdots\wedge\mathrm{d}x^n\wedge\mathrm{d}p_1\wedge\cdots\wedge\mathrm{d}p_n. \end{aligned}$$

(1) R leaves the x^μ alone and sends $p_\mu \mapsto -p_\mu$; pull-back commutes with d.
(2) Pull out one minus sign from each of the n factors $\mathrm{d}p_\mu$.

So R preserves orientation when n is even and *reverses* it when n is odd. On fundamental classes: $R_*[T^*X] = (-1)^n[T^*X]$.



$R : (x,p) \mapsto (x,-p)$
$n=1$: $\mathrm{d}\theta\wedge\mathrm{d}p \mapsto -\mathrm{d}\theta\wedge\mathrm{d}p$
(flips the cylinder's orientation)
$n=2$: $\mathrm{d}^2x\,\mathrm{d}^2p \mapsto (-1)^2\,\mathrm{d}^2x\,\mathrm{d}^2p$
general: factor $(-1)^n$

Figure 4.28: The antipodal map on the cotangent bundle of the circle. The cotangent bundle is a cylinder with coordinates (θ,p); reversing p reverses the orientation $\mathrm{d}\theta\wedge\mathrm{d}p$. In n dimensions the sign is $(-1)^n$, so the antipodal map is orientation-reversing exactly when $\dim X$ is odd.

Next, what does R do to the K-theory class of the symbol? By (Nash 4.33), $R^*\sigma_m(D) = (-1)^m\sigma_m(D)$. If m is even there is nothing to say. If m is odd, the symbol has been multiplied by the constant -1. But the constant matrix -1 can be joined to $+1$ through invertible matrices, for example by the path $t \mapsto e^{i\pi t}\cdot 1$, $t\in[0,1]$. Multiplying the symbol by $e^{i\pi t}$ keeps it invertible

off the zero section at every t, so it is a homotopy of elliptic symbols, and homotopic symbols define the same K-class:

$$\big[(-1)^m \sigma_m(D)\big] = \big[\sigma_m(D)\big] \in K(T^*X). \tag{4.88}$$

Now we can run the argument. Suppose n is odd and D is a differential operator. Integrals are invariant under diffeomorphisms up to the orientation sign, $\int_M R^*\omega = \pm \int_M \omega$ with + for orientation-preserving R. So

$$\begin{aligned}
\operatorname{index} D &\overset{(1)}{=} (-1)^n \big\{\operatorname{ch}[\sigma_m(D)] \cdot \operatorname{td}(TX_{\mathbb{C}})\big\}[TX] \\
&\overset{(2)}{=} (-1)^n \big\{\operatorname{ch}[R^*\sigma_m(D)] \cdot \operatorname{td}(TX_{\mathbb{C}})\big\}\big(-[TX]\big) \\
&\overset{(3)}{=} (-1)^n \big\{\operatorname{ch}[(-1)^m\sigma_m(D)] \cdot \operatorname{td}(TX_{\mathbb{C}})\big\}\big(-[TX]\big) \\
&\overset{(4)}{=} -(-1)^n \big\{\operatorname{ch}[\sigma_m(D)] \cdot \operatorname{td}(TX_{\mathbb{C}})\big\}[TX] \\
&\overset{(5)}{=} -\operatorname{index} D.
\end{aligned} \tag{4.89}$$

(1) The cohomological index formula (4.85).
(2) Change variables by the diffeomorphism R. The integrand is pulled back (the Todd class is pulled back from X, and R covers the identity on X, so only the symbol part changes); because n is odd, R reverses orientation, which is the sign on $[TX]$.
(3) Homogeneity of the symbol of a differential operator, (Nash 4.33).
(4) Equation (4.88): the sign $(-1)^m$ does not change the K-class, hence not its Chern character; pull out the orientation sign.
(5) The index formula again.

An integer equal to its own negative is zero.

Key idea: Odd dimensions (Nash (4.34))

If $\dim X$ is odd and D is an elliptic *differential* operator, then $\operatorname{index} D = 0$. The mechanism: the antipodal map $p \mapsto -p$ reverses the orientation of T^*X but leaves the K-class of a polynomial symbol unchanged.

Worked example 4.52: Every differential operator on the circle has index zero

Take $X = S^1$ with coordinate θ and the first-order operator $D_c = \dfrac{\mathrm{d}}{\mathrm{d}\theta} - ic$ on complex functions, with $c \in \mathbb{R}$ a constant. Its symbol ip is invertible for $p \neq 0$, so it is elliptic. Expand $f = \sum_{k\in\mathbb{Z}} a_k e^{ik\theta}$. Then

$$D_c f = \sum_k i(k-c)\, a_k e^{ik\theta}.$$

Kernel. $D_c f = 0$ iff $(k-c)a_k = 0$ for all k. If $c \notin \mathbb{Z}$ this forces every $a_k = 0$; if $c = k_0 \in \mathbb{Z}$ the mode $e^{ik_0\theta}$ survives. So $\dim \operatorname{Ker} D_c = 1$ if $c \in \mathbb{Z}$ and 0 otherwise.
Cokernel. The equation $D_c f = g$ with $g = \sum b_k e^{ik\theta}$ is solved by $a_k = b_k/(i(k-c))$, which works for every k unless $k = c$. So $\dim \operatorname{Coker} D_c = 1$ if $c \in \mathbb{Z}$ and 0 otherwise.
Hence

$$\begin{aligned}
\operatorname{index} D_c &\overset{(1)}{=} \dim \operatorname{Ker} D_c - \dim \operatorname{Coker} D_c \\
&\overset{(2)}{=} \begin{cases} 1-1 & c \in \mathbb{Z} \\ 0-0 & c \notin \mathbb{Z} \end{cases} = 0.
\end{aligned}$$

(1) Definition of the analytical index.

(2) The two mode counts above.

This is the stability of the index in action. As c moves, the kernel jumps from 0 to 1 at integer c, but the cokernel jumps with it: zero modes can only appear in pairs that the index does not see. The odd-dimensional vanishing theorem says this is not a special feature of D_c: *any* elliptic differential operator on S^1, of any order and acting on any bundle, has kernel and cokernel of equal dimension.

4.10.3 The pseudo-differential exception on the circle

A **pseudo-differential operator** on S^1 is, for our purposes, an operator that acts diagonally (or nearly so) on Fourier modes with a multiplier that grows like a power of $|k|$, not necessarily polynomially. The basic example is

$$|D_\theta| := \sqrt{-\partial_\theta^2}: \qquad |D_\theta|\, e^{ik\theta} = |k|\, e^{ik\theta},$$

the "absolute value of the momentum". Its symbol is $|p|$, homogeneous of degree 1 for $\lambda > 0$ but even under $p \to -p$. We also write $D_\theta := -i\partial_\theta$, so that $D_\theta e^{ik\theta} = k e^{ik\theta}$ and its symbol is p.

Nash's example (4.35) is built from the two combinations $D_\theta \pm |D_\theta|$, which are (twice) the projections onto positive and negative frequencies:

$$(D_\theta + |D_\theta|)e^{ik\theta} = \begin{cases} 2k\, e^{ik\theta} & k > 0 \\ 0 & k \le 0 \end{cases}, \qquad (D_\theta - |D_\theta|)e^{ik\theta} = \begin{cases} 0 & k \ge 0 \\ 2k\, e^{ik\theta} & k < 0 \end{cases}.$$

Define

$$D := e^{-i\theta}\big(D_\theta + |D_\theta|\big) + \big(D_\theta - |D_\theta|\big), \qquad \sigma_1^D(\theta, p) = e^{-i\theta}(p + |p|) + (p - |p|). \tag{4.90}$$

The symbol equals $2p\, e^{-i\theta}$ for $p > 0$ and $2p$ for $p < 0$: never zero for $p \neq 0$, so D is elliptic. And $\sigma(\theta, -p)$ bears no simple relation to $\sigma(\theta, p)$, so the argument (4.89) does not apply.

Watch out: A sign typo in Nash (4.35)

Nash prints the operator as $e^{-i\theta}(-i\partial_\theta + \sqrt{-\partial_\theta^2}) - (\partial_\theta + \sqrt{-\partial_\theta^2})$ but its symbol as $e^{-i\theta}(p + |p|) + (p - |p|)$. These do not match (the second bracket would have symbol $-(ip + |p|)$). The symbol is the intended object; the operator (4.90) is the one with that symbol, and it is the one we analyse.

Worked example 4.53: Nash's pseudo-differential operator has index one

Apply D to a single mode, using the two projection formulas:

$$D\, e^{ik\theta} = \begin{cases} 2k\, e^{i(k-1)\theta} & k > 0, \\ 0 & k = 0, \\ 2k\, e^{ik\theta} & k < 0. \end{cases}$$

So positive modes are *shifted down by one* and rescaled, negative modes are rescaled in place, and the constant mode is killed.

Kernel. For $f = \sum a_k e^{ik\theta}$, $Df = \sum_{k>0} 2k a_k e^{i(k-1)\theta} + \sum_{k<0} 2k a_k e^{ik\theta}$. The first sum only contains modes $k - 1 \ge 0$, the second only modes $k < 0$, so the two cannot cancel each other. Thus $Df = 0$ iff $a_k = 0$ for all $k \neq 0$: $\operatorname{Ker} D = \operatorname{span}\{1\}$.

Cokernel. Every target mode is hit: a mode $e^{ij\theta}$ with $j \geq 0$ is D of $e^{i(j+1)\theta}/(2(j+1))$, and a mode with $j < 0$ is D of $e^{ij\theta}/(2j)$. So D is onto and $\operatorname{Coker} D = 0$.
Hence

$$\operatorname{index} D \overset{(1)}{=} \dim \operatorname{Ker} D - \dim \operatorname{Coker} D \overset{(2)}{=} 1 - 0 = 1.$$

(1) Definition of the analytical index.
(2) The kernel is the constants; the operator is surjective.

A nonzero index on an odd-dimensional manifold: pseudo-differential operators really do escape the theorem.

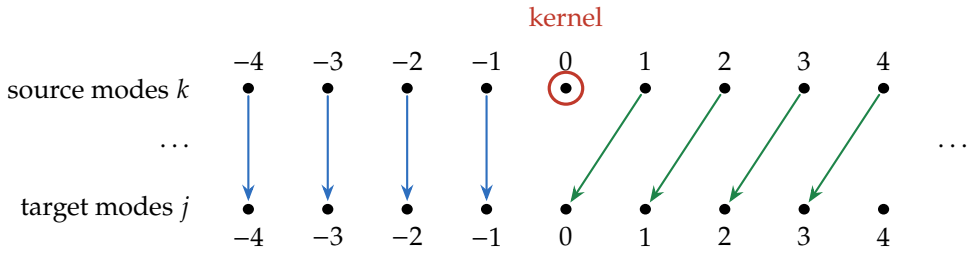


Figure 4.29: How Nash's pseudo-differential operator acts on Fourier modes. Negative modes go straight down (blue); positive modes are shifted one step to the left (green). Every target mode is hit exactly once, so the cokernel is zero, while the source mode $k = 0$ has nowhere to go: it is the kernel. Index $= 1 - 0 = 1$. A differential operator could never do this: it cannot treat $k > 0$ and $k < 0$ differently.

Physical picture: A winding number in disguise

Restrict the symbol to the unit cosphere bundle $|p| = 1$, which over S^1 is two circles, $p = +1$ and $p = -1$. On them $\sigma = 2e^{-i\theta}$ (winding number -1 around the origin of $\mathbb{C}$) and $\sigma = -2$ (winding number 0). The index is the difference $0 - (-1) = 1$. This is the classical Noether–Toeplitz index formula: for operators of this "positive frequencies times a, negative frequencies times b" type, the index is $w(b) - w(a)$, the difference of the winding numbers of the two symbol branches. A differential operator has $\sigma(\theta, -1) = \pm\sigma(\theta, +1)$, so the two windings agree and the index vanishes, which is (4.89) again, seen from the circle.

Try it: More winding

Replace $e^{-i\theta}$ by $e^{-iN\theta}$, $N \geq 1$, in (4.90). Show that the index is N.
Answer. Now $e^{ik\theta} \mapsto 2k\, e^{i(k-N)\theta}$ for $k > 0$ and $e^{ik\theta} \mapsto 2ke^{ik\theta}$ for $k < 0$. Every target mode is hit (modes $j \geq 1 - N$ from $k = j + N > 0$, and all $j < 0$), so the cokernel is zero. The target modes $j = 1 - N, \ldots, -1$ are hit twice, once from $k = j + N$ and once from $k = j$; the combination $a_{j+N}(j + N) + a_j j = 0$ gives one kernel direction for each such j, i.e. $N - 1$ directions, plus the constant. So index $= N - 0 = N = w(-2) - w(2e^{-iN\theta}) = 0 - (-N)$.

4.10.4 Operators acting on functions

Now suppose $D : \Gamma(X, E) \to \Gamma(X, F)$ with E and F both *trivial* of rank k (the ranks agree because D is elliptic, so its symbol is an isomorphism $E_x \to F_x$). Then D acts on k-component functions, and its leading symbol at (x, p) is literally a $k \times k$ matrix $\sigma_m(D)(x, p)$, invertible for $p \neq 0$.

By homogeneity, the symbol is determined by its values on unit covectors $|p| = 1$. The

set of unit covectors is the **unit sphere bundle** (cosphere bundle) $S(X) \subset T^*X$, a bundle of $(n-1)$-spheres over X, of total dimension $2n-1$. So the symbol is a map

$$\sigma_m(D) : S(X) \longrightarrow \mathrm{GL}(k, \mathbb{C}). \tag{Nash 4.36}$$

All the topology of the symbol is in how this map wraps $S(X)$ around $\mathrm{GL}(k,\mathbb{C})$, which is recorded by the pull-back on cohomology,

$$\sigma_m(D)^* : H^*(\mathrm{GL}(k,\mathbb{C});\mathbb{Z}) \longrightarrow H^*(S(X);\mathbb{Z}). \tag{Nash 4.37}$$

Two simplifications. First, $\mathrm{GL}(k,\mathbb{C})$ deformation-retracts onto $\mathrm{U}(k)$ (polar decomposition $M = U\,(M^\dagger M)^{1/2}$, and the positive factor can be shrunk to 1 through positive matrices), so we may replace $\mathrm{GL}(k,\mathbb{C})$ by $\mathrm{U}(k)$ in (Nash 4.37). Second, specialise to a single function, $k = 1$. Then $\mathrm{U}(1) = S^1$, whose cohomology is $\mathbb{Z}$ in degrees 0 and 1, with degree-1 generator h: the "angle form" $\frac{1}{2\pi}\mathrm{d}\arg z$. The pull-back $\sigma^*h = \frac{1}{2\pi}\mathrm{d}\arg\sigma$ is a closed 1-form on $S(X)$ whose integral over any loop is the winding number of the symbol around that loop.

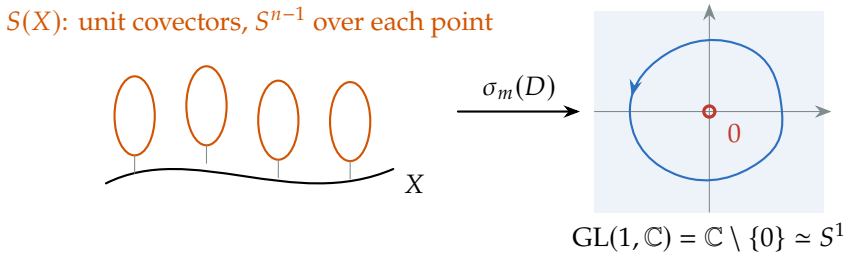


Figure 4.30: The symbol of an elliptic operator on functions ($k = 1$) is a map from the unit cosphere bundle $S(X)$ to the punctured plane. Its only topological content is how many times loops in $S(X)$ wind around the origin: a class in $H^1(S(X);\mathbb{Z})$. That single degree-1 class is what the dimension count in the text works with.

Now Nash argues that the index formula can be evaluated on $S(X)$ rather than on TX (the compactly supported cohomology of T^*X is reached from $S(X)$ by a suspension, the "clutching" construction of Chapter 4's first part):

$$\mathrm{index}\, D = (-1)^n \left\{ \mathrm{ch}[\sigma_m(D)] \cdot \mathrm{td}(TX_{\mathbb{C}}) \right\} [S(X)]. \tag{Nash 4.38}$$

Now count degrees.

- The symbol part: the only cohomology σ can carry is the pulled back 1-dimensional class, so $\mathrm{ch}[\sigma_m(D)]$ is proportional to $\sigma^*h \in H^1(S(X);\mathbb{Z})$: it has degree 1.
- The Todd part: by (4.84), $\mathrm{td}(TX_{\mathbb{C}})$ is a polynomial in Pontryagin classes, which live in degrees $0, 4, 8, \ldots$: pick its degree-$4i$ piece.
- The evaluation on $[S(X)]$ only sees forms of degree $\dim S(X) = 2n-1$.

So a nonzero contribution requires

$$4i + 1 = 2n - 1 \quad \Longrightarrow \quad n = 2i + 1. \tag{Nash 4.39}$$

In words: the index can be nonzero only if n is odd. But in odd dimensions a differential operator has index zero by (4.89). Both alternatives give zero.

For a pseudo-differential operator we lose the odd-dimensional theorem, but we gain a second constraint. The Todd class is pulled back from X, and a form on X of degree bigger than n vanishes. So $4i \le n$ as well:

$$0 \le 4i \le n, \qquad n = 2i + 1 \quad \Longrightarrow \quad 4i \le 2i + 1 \Longrightarrow\ i = 0,\ n = 1. \tag{Nash 4.40}$$

The only escape is $\dim X = 1$, and the circle example (4.90) shows the escape is real.

Key idea: Operators on functions

The index of an elliptic *differential* operator acting on a single function (trivial line bundles $E = F = X \times \mathbb{C}$) is always zero. For a *pseudo-differential* operator on functions the index is zero whenever $\dim X > 1$; in dimension 1 it is a winding number and can be anything.

Worked example 4.54: Why the hypothesis "functions" matters: $\bar\partial$ on a Riemann surface

The Cauchy–Riemann operator $\bar\partial : f \mapsto \partial_{\bar z} f \, d\bar z$ on a compact Riemann surface Σ_g of genus g takes functions to $(0,1)$-forms. Its kernel is the holomorphic functions, which on a compact surface are constants (maximum principle), so $\dim \operatorname{Ker} \bar\partial = 1$; its cokernel has dimension g (the number of independent holomorphic 1-forms, as we shall see in Section 4.13). So

$$\operatorname{index} \bar\partial = 1 - g.$$

This is nonzero for $g \neq 1$, but there is no contradiction: the target bundle $\Lambda^{0,1}$ is a nontrivial line bundle of degree $2g-2$ (it is $\bar K$), except on the torus $g = 1$, where $\Lambda^{0,1}$ *is* trivial (spanned by $d\bar z$) and indeed $\operatorname{index} \bar\partial = 1 - 1 = 0$, as the theorem demands.

Watch out: Systems of functions

For $k > 1$ the theorem changes. The symbol is now a map $S(X) \to \mathrm{U}(k)$, and $\mathrm{U}(k)$ has cohomology in degrees $1, 3, 5, \dots, 2k-1$, so the degree count (Nash 4.39) admits more solutions. Atiyah and Singer (paper III) show that vanishing statements survive for a certain range of k and n, but not in general. The Dirac operator on $\mathbb{R}^4$ is a 2×2 system and can have nonzero index once it is coupled to a nontrivial gauge field, which makes the bundle nontrivial (Section 4.16).

Reading Nash (pp. 98–101, (4.33)–(4.40))

Nash's "index when dim X is odd" is our Section 4.10.2; his (4.33) is (Nash 4.33), (4.34) is (4.89). The counterexample (4.35) is from Eguchi–Gilkey–Hanson; we have computed its index (one) mode by mode and corrected a sign in the printed operator. His "index when D acts on functions" is Section 4.10.4, with (4.36)–(4.40) reproduced as displayed. He says "we can now argue informally" before (4.38): the evaluation on $S(X)$ is indeed a sketch, and so is ours. From p. 101 on, Nash restricts to orientable X and differential operators; we follow him in the next section.

Recap

- Symbols of differential operators are homogeneous polynomials in p: $\sigma(x,-p) = (-1)^m \sigma(x,p)$, and the sign does not change the K-class.
- The antipodal map on T^*X has orientation sign $(-1)^n$; in odd dimensions it forces $\operatorname{index} D = -\operatorname{index} D = 0$.
- Pseudo-differential operators, like $e^{-i\theta}(D_\theta + |D_\theta|) + (D_\theta - |D_\theta|)$ on S^1, can have nonzero index (here 1) because $|p|$ is even.
- On functions ($k = 1$) the symbol only carries a degree-1 winding class; degree counting then kills the index except for pseudo-differential operators in dimension 1.

4.11 Orientable manifolds: the index formula on X itself

The formula (4.85) integrates over the $2n$-dimensional total space of the tangent bundle. That is awkward: we know how to integrate curvature forms over X, not over TX. In this section we push the integral down to X. The price is two mild restrictions, which Nash imposes from here on:

- X is *orientable*, so that it has a fundamental class $[X]$ and "$\int_X$" makes sense;
- D is a *differential* operator, and in fact part of an *elliptic complex* $0 \to \Gamma(E^0) \to \Gamma(E^1) \to \cdots$. (By Section 4.10 the interesting case is then n even.)

The tool for passing between TX and X is the **Thom isomorphism**. It is nothing more exotic than *integration over the fibres*, and the object that makes it work, the **Thom class**, is a closed form concentrated near the zero section whose integral over each fibre is 1: a normalised "bump" or "Gaussian" in the fibre directions. Its value on the zero section turns out to be the Euler class, and this is why $e(X)$ appears in Nash's second index formula (4.48).

4.11.1 Orientation and the fundamental class

An **orientation** of an n-manifold is a consistent choice of "positive" ordering of bases of every tangent space, equivalently a nowhere-vanishing n-form up to positive multiples, equivalently an atlas whose transition functions have positive Jacobian determinant. The circle, the spheres, the tori and every complex manifold are orientable; the Möbius strip, $\mathbb{RP}^2$ and the Klein bottle are not.

Working definition 4.31 (fundamental class; evaluation)

On a compact oriented n-manifold X, the **fundamental class** $[X] \in H_n(X;\mathbb{Z})$ [integer combinations of closed n-dimensional pieces of X, up to boundaries] is "X itself, counted once with its orientation". For a cohomology class a represented by a closed form α [$\mathrm{d}\alpha = 0$, a sum of pieces of several degrees] we write

$$a[X] \;:=\; \int_X \alpha^{(n)},$$

where $\alpha^{(n)}$ is the degree-n component. Components of other degrees are simply ignored. *Operationally*: "evaluate on $[X]$" means "integrate the top-degree part".

Stokes' theorem guarantees this is well defined on classes: if $\alpha^{(n)}$ changes by $\mathrm{d}\beta$, the integral changes by $\int_X \mathrm{d}\beta = \int_{\partial X} \beta = 0$. The same notation $b[TX]$ was used in (4.85) for a compactly supported class b on the non-compact manifold TX; compact support is exactly what makes Stokes' theorem (and so well-definedness) work there.

4.11.2 The Thom class: a unit bump in the fibres

Start with a single fibre $\mathbb{R}^n$, coordinates $v^1, \dots, v^n$. Among all n-forms on $\mathbb{R}^n$ with compact support (or rapid decay) consider

$$u_{\mathbb{R}^n} \;=\; \pi^{-n/2}\, e^{-|v|^2}\, \mathrm{d}v^1 \wedge \cdots \wedge \mathrm{d}v^n. \tag{4.91}$$

It is closed (it is a top form) and its integral is

$$\int_{\mathbb{R}^n} u_{\mathbb{R}^n} \overset{(1)}{=} \pi^{-n/2} \prod_{j=1}^{n} \int_{-\infty}^{\infty} e^{-(v^j)^2}\, \mathrm{d}v^j$$
$$\overset{(2)}{=} \pi^{-n/2} \big(\sqrt{\pi}\big)^n = 1.$$

(1) The Gaussian factorises, $e^{-|v|^2} = \prod_j e^{-(v^j)^2}$, and the orientation $\mathrm{d}v^1 \wedge \cdots \wedge \mathrm{d}v^n$ is the standard one, so the form integral is an ordinary iterated integral.

(2) The one-dimensional Gaussian integral $\int e^{-t^2}\mathrm{d}t = \sqrt{\pi}$.

Any other rapidly decaying n-form with total integral 1 differs from $u_{\mathbb{R}^n}$ by d of a rapidly decaying $(n-1)$-form (this is the statement that the compactly supported cohomology $H^n_c(\mathbb{R}^n)$ is one-dimensional, detected by the total integral). So the *class* of $u_{\mathbb{R}^n}$ is canonical: the generator of $H^n_c(\mathbb{R}^n) \cong \mathbb{R}$.

Now let $V \to X$ be an oriented real vector bundle of rank n (for us, $V = TX$). We want a closed n-form u on the total space of V which, on every fibre, looks like (4.91).

Working definition 4.32 (Thom class)

A **Thom form** of an oriented rank-n bundle $\pi : V \to X$ is a closed n-form u on the total space [V itself, the set of all vectors in all fibres], with compact support (or Gaussian decay) in the fibre directions, whose integral over every fibre V_x is 1:

$$\mathrm{d}u = 0, \qquad \int_{V_x} u = 1 \quad \text{for all } x \in X.$$

Its class $u \in H^n_c(V)$ is the **Thom class**. *Formally:* u is the unique class in compactly supported cohomology [closed forms vanishing outside a compact set, up to d of such forms] restricting to the orientation generator [the class whose integral over V_x, with its orientation, is +1] of each $H^n_c(V_x)$.

On a product $X \times \mathbb{R}^n$ we can take $u = \pi_2^* u_{\mathbb{R}^n}$. On a twisted bundle the naive Gaussian is not closed, because the fibre coordinates rotate from patch to patch. One must add correction terms involving the connection; the result (the Mathai–Quillen form) is closed, and its correction terms carry the curvature. We work this out completely in rank 2 below; it is exactly these correction terms that produce the Euler class.

4.11.3 The Thom isomorphism is integration over the fibre

Nash's map (4.41) sends a class a on X to a compactly supported class on TX:

$$\psi : H^*(X) \xrightarrow{\pi^*} H^*(TX) \xrightarrow{\cdot u} H^{*+n}_c(TX), \qquad \psi(a) = \pi^*(a) \cdot u. \tag{Nash 4.41}$$

Pull a back to the total space (a form constant along each fibre), then multiply by the unit bump to make it compactly supported. The inverse goes the other way, by integrating over the fibres. The **fibre integral** π_* of a form on TX is the form on X obtained by integrating out the fibre coordinates. For a product form,

$$\pi_*\big(\psi(a)\big) \overset{(1)}{=} \pi_*\big(\pi^* a \wedge u\big)$$
$$\overset{(2)}{=} a \cdot \int_{V_x} u$$
$$\overset{(3)}{=} a. \tag{4.92}$$

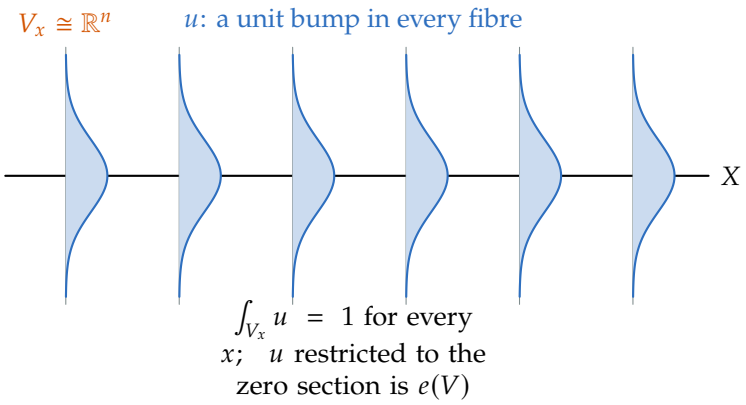


Figure 4.31: The Thom form of a vector bundle $V \to X$, drawn as a bump function of the fibre coordinate over each point of the base (the fibres are the vertical lines, the zero section is the horizontal line X). The area under each bump is 1. Integrating a form of the type $\pi^* a \wedge u$ over the total space therefore gives $\int_X a$: the Thom class "converts" integrals over V into integrals over X.

(1) Definition (Nash 4.41).
(2) $\pi^* a$ does not depend on the fibre coordinates, so it comes out of the fibre integral (the **projection formula**); only the degree-n fibre part of u survives the integration.
(3) Normalisation of the Thom form.

So $\psi^{-1} = \pi_*$ on the image of ψ. The content of the *Thom isomorphism theorem* is that ψ is onto: *every* compactly supported class on TX is of the form $\pi^* a \cdot u$. In words, the only topology a compactly supported class on a vector bundle can have is "topology of the base, times a bump".

And evaluation on fundamental classes is compatible:

$$\begin{aligned} \psi(a)[TX] &\overset{(1)}{=} \int_{TX} \pi^* a \wedge u \\ &\overset{(2)}{=} \int_X a \cdot \left(\int_{T_x X} u \right) = a[X]. \end{aligned} \tag{4.93}$$

(1) Definition of evaluation for a compactly supported class.
(2) Fubini: integrate first over the fibre, then over the base. This holds with the "base first, then fibre" orientation of TX; with another orientation convention a sign appears (see below).

4.11.4 The Thom class restricted to the zero section is the Euler class

Let $s_0 : X \to TX$ be the zero section, $x \mapsto (x, 0)$. Restricting u to it gives an n-form on X. For a trivial bundle this is zero (the Gaussian form $u_{\mathbb{R}^n}$ contains $\mathrm{d}v^1 \wedge \cdots \wedge \mathrm{d}v^n$, which vanishes when pulled back to $v = 0$). For a twisted bundle, the connection terms survive:

Key idea: Thom class on the zero section

For an oriented rank-n bundle V,

$$s_0^* u = e(V), \tag{4.94}$$

the Euler class. In particular the Euler class of an odd-rank bundle vanishes: the fibrewise map $v \mapsto -v$ reverses the orientation of odd-dimensional fibres, so it sends $u \mapsto -u$, while it fixes the zero section; hence $e(V) = -e(V) = 0$.

Let us see (4.94) with our own eyes in rank 2, where everything can be written down.

Worked example 4.55: Thom form of an oriented plane bundle

Let V have rank 2 with a metric and a compatible (SO(2)) connection. Over a patch, choose an orthonormal frame and polar fibre coordinates (r, φ). The connection is a single real 1-form ω (the SO(2) matrix is $\left(\begin{smallmatrix} 0 & -\omega \\ \omega & 0 \end{smallmatrix}\right)$ up to sign conventions), and under a change of frame by a rotation through angle $\lambda(x)$,

$$\varphi \mapsto \varphi - \lambda, \qquad \omega \mapsto \omega - \mathrm{d}\lambda.$$

Hence the combination $\Psi := \mathrm{d}\varphi - \omega$ is a globally defined 1-form on V minus its zero section (the **global angular form**), and

$$\mathrm{d}\Psi = -\mathrm{d}\omega = -\pi^*\big(2\pi\, e\big), \qquad e := \frac{\mathrm{d}\omega}{2\pi}. \tag{4.95}$$

Here $\mathrm{d}\omega$ is the curvature 2-form, and $e = \mathrm{d}\omega/2\pi$ is the Pfaffian of $R/2\pi$ for rank 2: the Euler form. Pick a function $\rho(r)$ with $\rho(0) = -1$ and $\rho(r) = 0$ for large r (for example $\rho = -e^{-r^2}$), and set

$$u := \mathrm{d}\Big(\rho(r)\,\frac{\Psi}{2\pi}\Big) = \frac{\rho'(r)}{2\pi}\,\mathrm{d}r \wedge \Psi + \frac{\rho(r)}{2\pi}\,\mathrm{d}\Psi. \tag{4.96}$$

It is exact on V minus the zero section, but it is *closed* on all of V (it is smooth across $r = 0$ because ρ is constant to first order there, and $\mathrm{d} \circ \mathrm{d} = 0$). Its support is near the zero section. Now check the two properties.

Fibre integral. On a fibre ω and $\mathrm{d}\omega$ restrict to zero (they are pulled back from the base), so

$$\begin{aligned} \int_{V_x} u &\overset{(1)}{=} \int_0^\infty \int_0^{2\pi} \frac{\rho'(r)}{2\pi}\,\mathrm{d}r\,\mathrm{d}\varphi \\ &\overset{(2)}{=} \rho(\infty) - \rho(0) = 0 - (-1) = 1. \end{aligned}$$

(1) On a single fibre $\Psi = \mathrm{d}\varphi$, and $\mathrm{d}\Psi = 0$; the orientation $\mathrm{d}r \wedge \mathrm{d}\varphi$ is the standard one of the plane.

(2) The φ integral gives 2π; the r integral is a total derivative.

Restriction to the zero section. At $r = 0$ we have $\mathrm{d}r = 0$ along the zero section and $\rho(0) = -1$, so

$$\begin{aligned} s_0^* u &\overset{(1)}{=} \frac{\rho(0)}{2\pi}\, s_0^* \mathrm{d}\Psi \\ &\overset{(2)}{=} \frac{-1}{2\pi}\big(-2\pi e\big) = e. \end{aligned}$$

(1) The first term of (4.96) contains $\mathrm{d}r$, which restricts to zero on $r = 0$.

(2) Equation (4.95).

For $V = TS^2$ with the round metric: in the orthonormal coframe $e^1 = \mathrm{d}\theta$, $e^2 = \sin\theta\,\mathrm{d}\varphi$ the Levi-Civita connection is $\omega = -\cos\theta\,\mathrm{d}\varphi$ (up to orientation of the frame rotation), so $\mathrm{d}\omega = \sin\theta\,\mathrm{d}\theta \wedge \mathrm{d}\varphi$ and $\int_{S^2} e = 4\pi/2\pi = 2$. The Thom form of TS^2 carries the Euler number 2 on its zero section.

Worked example 4.56: Euler number from zeros of a vector field

The Euler class has a second, very physical, meaning. A vector field s is a section of TX. Pull the Thom form back by s instead of the zero section. Since s is homotopic to s_0 (through ts, $t \in [0,1]$) and u is closed, $\int_X s^* u = \int_X s_0^* u = \int_X e$. But if we scale $s \to \lambda s$ with

$\lambda \to \infty$, the bump u becomes a delta function, and s^*u localises at the *zeros* of s:

$$\int_X e \overset{(1)}{=} \int_X (\lambda s)^* u \overset{(2)}{=} \sum_{s(x_k)=0} \int_{B_k} (\lambda s)^* u \overset{(3)}{=} \sum_{s(x_k)=0} \operatorname{ind}_{x_k}(s).$$

(1) Homotopy invariance: λs is homotopic to the zero section, and integrals of pull-backs of a closed form are homotopy invariant.

(2) For large λ the integrand vanishes except in small balls B_k around the zeros, because u decays away from the zero section.

(3) Near an isolated zero, λs maps a small ball onto a large region of the fibre; the integral of the unit bump over it counts how many times (with sign) s covers the fibre: the local **index** (degree) of the zero, ± 1 for a non-degenerate zero.

On S^2 take the "rotation" field $s = \partial_\varphi$, which vanishes at the two poles. Near each pole it looks like $(x, y) \mapsto (-y, x)$, a rotation, of index +1. So $\int e = 1 + 1 = 2$. This is the **Poincaré–Hopf theorem**, $\sum \text{ind} = \chi(X)$ once we know $\int e = \chi$ (which the de Rham example below proves from the index theorem). On a torus the constant field ∂_x has no zeros: $\int_{T^2} e = 0 = \chi(T^2)$.

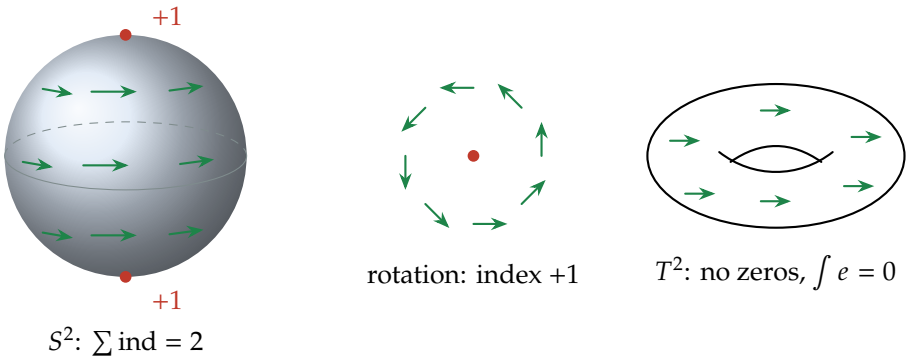


Figure 4.32: The Euler class counts zeros of vector fields. Left: the rotation field on S^2 vanishes only at the poles (red dots), each a zero of index +1 (middle: the local picture). Right: the torus carries a nowhere-vanishing field. Pulling the Thom form back by a large multiple of the vector field concentrates it at the zeros, which is why $\int_X e$ equals the sum of the indices.

4.11.5 The K-class of an elliptic complex

Now the operator side. An **elliptic complex** is a sequence

$$0 \longrightarrow \Gamma(X, E^0) \xrightarrow{d_0} \Gamma(X, E^1) \xrightarrow{d_1} \cdots \xrightarrow{d_{N-1}} \Gamma(X, E^N) \longrightarrow 0, \qquad d_{i+1} \circ d_i = 0, \tag{4.97}$$

of differential operators whose *symbol sequence*

$$0 \longrightarrow \pi^* E^0 \xrightarrow{\sigma(d_0)} \pi^* E^1 \xrightarrow{\sigma(d_1)} \cdots \longrightarrow \pi^* E^N \longrightarrow 0 \tag{Nash 4.44}$$

is *exact* at every (x,p) with $p \neq 0$. (Here π^*E^i is the bundle over T^*X whose fibre at (x,p) is E^i_x.) A single elliptic operator is the case $N=1$: exactness of $0 \to E^0_x \to E^1_x \to 0$ just says the symbol is an isomorphism. The index of the complex is the alternating sum of the dimensions of its cohomology groups, index $= \sum_i (-1)^i \dim H^i$; rolling up the complex into the single operator $d + d^* : \Gamma(\oplus E^{\text{even}}) \to \Gamma(\oplus E^{\text{odd}})$ gives an elliptic operator with the same index (Chapter 2).

For a complex of *vector bundles* $0 \to E^0 \to E^1 \to \cdots$ over a space, the K-theory class is the alternating sum

$$[E] = \sum_i (-1)^i [E^i]. \tag{Nash 4.43}$$

The logic: if the complex is exact, it splits, $E^i \cong B^i \oplus B^{i+1}$ with B^i the image of the previous map, and the alternating sum telescopes to zero. So (Nash 4.43) measures the failure of exactness, and a complex which is exact outside a compact set defines a compactly supported class. The symbol complex (Nash 4.44) is exact off the zero section, so it defines

$$[\sigma_m(D)] = \sum_p (-1)^p [\pi^* E^p] \in K(T^*X),$$

$$\operatorname{ch}[\sigma_m(D)] \overset{(1)}{=} \operatorname{ch}\Big(\sum_p (-1)^p [\pi^* E^p]\Big) \overset{(2)}{=} \pi^* \operatorname{ch}\Big(\sum_p (-1)^p [E^p]\Big). \tag{Nash 4.45}$$

(1) ch is additive, so it can be applied term by term.

(2) ch commutes with pull-back (the curvature of π^*E is the pull-back of the curvature of E).

We must be careful about what (Nash 4.45) means. The right-hand side, π^* of a class on X, is not compactly supported. The left-hand side is: it is the compactly supported class of the whole complex including its maps $\sigma(d_i)$, which is what cuts it off at large $|p|$. The honest statement is that the two agree *on the zero section*:

$$s_0^* \operatorname{ch}[\sigma_m(D)] = \operatorname{ch}\Big(\sum_p (-1)^p [E^p]\Big) \in H^{\text{even}}(X). \tag{4.98}$$

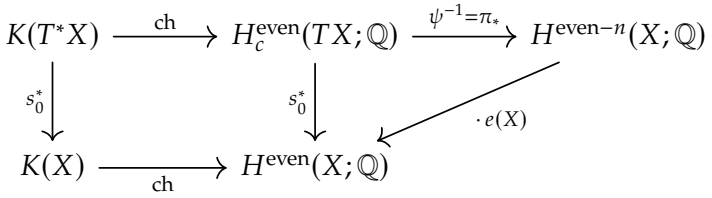


Figure 4.33: The bookkeeping behind Nash's (4.47). A compactly supported class on TX (top middle) can be pushed down in two ways: integrate over the fibres (ψ^{-1}), or restrict to the zero section (s_0^*). Because $\psi(b) = \pi^*b \cdot u$ and u restricts to $e(X)$, the two routes differ by multiplication by the Euler class: $s_0^*\psi(b) = b \cdot e(X)$.

4.11.6 Nash's second index formula

We now have all the pieces. Write the compactly supported class $\operatorname{ch}[\sigma_m(D)]$ as $\psi(b) = \pi^*b \cdot u$ for a class b on X (possible by the Thom isomorphism). Restrict to the zero section:

$$\operatorname{ch}\Big(\sum_p (-1)^p [E^p]\Big) \overset{(1)}{=} s_0^* \operatorname{ch}[\sigma_m(D)]$$

$$\overset{(2)}{=} s_0^*\big(\pi^*b \cdot u\big)$$

$$\overset{(3)}{=} b \cdot s_0^* u$$
$$\overset{(4)}{=} b \cdot e(X). \tag{4.99}$$

(1) Equation (4.98).
(2) The Thom isomorphism: $\mathrm{ch}[\sigma_m(D)] = \psi(b)$.
(3) $\pi \circ s_0 = \mathrm{id}_X$, so $s_0^* \pi^* b = b$; pull-back is multiplicative.
(4) The Thom class restricts to the Euler class, (4.94).

Dividing, $b = \psi^{-1}\,\mathrm{ch}[\sigma_m(D)] = \mathrm{ch}(\sum(-1)^p[E^p])/e(X)$, which is Nash's (4.47).

Watch out: Dividing by the Euler class

Cohomology classes cannot in general be divided: $e(X)$ lives in top degree n, so $b \cdot e(X)$ only remembers the degree-0 part of b! The quotient in (4.47)–(4.48) must be read as a *formal identity in the Chern roots*. In every example the numerator $\mathrm{ch}(\sum_p (-1)^p E^p)$ turns out to be a polynomial in the roots that is divisible by $x_1 \cdots x_m$ (e.g. it is $\prod_j (1 - e^{x_j})(1 - e^{-x_j})$ for the de Rham complex); one cancels the factor $x_1 \cdots x_m$ *before* evaluating. The result is the correct formula for b because the splitting-principle computation is really done on the universal bundle, where there are no zero divisors.

Now the index. Start from (4.85). Two orientation conventions for TX are in play: the canonical one used by Atiyah and Singer in (4.31), and the "base, then fibre" one used in (4.93). They differ by a sign we can compute. In local coordinates, the canonical (symplectic) orientation of T^*X is $\frac{1}{n!}\big(\sum_\mu \mathrm{d}x^\mu \wedge \mathrm{d}p_\mu\big)^n = \mathrm{d}x^1 \wedge \mathrm{d}p_1 \wedge \mathrm{d}x^2 \wedge \mathrm{d}p_2 \wedge \cdots$, and

$$\mathrm{d}x^1 \wedge \mathrm{d}p_1 \wedge \cdots \wedge \mathrm{d}x^n \wedge \mathrm{d}p_n \overset{(1)}{=} (-1)^{0+1+\cdots+(n-1)}\,\mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^n \wedge \mathrm{d}p_1 \wedge \cdots \wedge \mathrm{d}p_n$$
$$\overset{(2)}{=} (-1)^{n(n-1)/2}\,\mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^n \wedge \mathrm{d}p_1 \wedge \cdots \wedge \mathrm{d}p_n.$$

(1) Move $\mathrm{d}x^2$ to the left past one $\mathrm{d}p$, then $\mathrm{d}x^3$ past two, and so on; each transposition of 1-forms costs a sign.
(2) $0 + 1 + \cdots + (n-1) = n(n-1)/2$.

So $[TX] = (-1)^{n(n-1)/2}[TX]_{\text{base-fibre}}$. Together with the explicit $(-1)^n$ in (4.31) the total sign is $(-1)^n(-1)^{n(n-1)/2} = (-1)^{n(n+1)/2}$, which is the factor in Nash's (4.42). (Different books orient TX differently; what matters is that the final formula is right, and we will confirm the sign against the de Rham complex, whose index we know independently.) Hence

$$\begin{aligned}
\mathrm{index}\, D &\overset{(1)}{=} (-1)^n \big\{\mathrm{ch}[\sigma_m(D)] \cdot \mathrm{td}(TX_{\mathbb{C}})\big\}[TX] \\
&\overset{(2)}{=} (-1)^{n(n+1)/2} \big\{\psi(b) \cdot \pi^*\,\mathrm{td}(TX_{\mathbb{C}})\big\}[TX]_{\text{base-fibre}} \\
&\overset{(3)}{=} (-1)^{n(n+1)/2} \big\{b \cdot \mathrm{td}(TX_{\mathbb{C}})\big\}[X] \\
&\overset{(4)}{=} (-1)^{n/2} \left\{\frac{\mathrm{ch}\big(\sum_p (-1)^p [E^p]\big)}{e(X)} \cdot \mathrm{td}(TX_{\mathbb{C}})\right\}[X].
\end{aligned} \tag{4.100}$$

(1) The cohomological index formula (4.85).
(2) Change the orientation convention (sign computed above); the Todd class on TX is pulled back from X.
(3) $\psi(b) \cdot \pi^*\,\mathrm{td} = \pi^*(b \cdot \mathrm{td}) \cdot u = \psi(b \cdot \mathrm{td})$, and (4.93).
(4) Insert b from (4.99). Now n is even (in odd dimensions the index of a differential operator vanishes, and $e(X) = 0$ anyway), $n = 2l$, and $n(n+1)/2 = l(2l+1) \equiv l \pmod 2$, so the sign is $(-1)^{n/2}$.

Key idea: Index formula for elliptic complexes, Nash (4.48)

For an elliptic complex $(E^\bullet, d)$ of differential operators on a compact oriented manifold X of even dimension n,

$$\operatorname{index} D = (-1)^{n/2} \int_X \frac{\operatorname{ch}\big(\sum_p (-1)^p E^p\big)}{e(X)} \operatorname{td}(TX_{\mathbb{C}}). \tag{4.101}$$

Recipe: (i) write $\operatorname{ch}(\sum_p (-1)^p E^p)$ in Chern roots; (ii) cancel the product $x_1 \cdots x_m$ against $e(X)$; (iii) multiply by $\operatorname{td}(TX_{\mathbb{C}}) = \prod_j \frac{x_j}{1-e^{-x_j}} \frac{-x_j}{1-e^{x_j}}$; (iv) pick out the degree-n part and integrate.

Physical picture: What the division by $e(X)$ is doing

The symbol class is concentrated near the zero section of TX, like the Thom form. It is "some cohomology on X times a unit bump". The formula asks for the "some cohomology" part. Restricting to the zero section gives it multiplied by the value of the bump at its centre, $e(X)$; the division removes that factor. In the four examples that follow, the numerator will always factorise as $e(X)$ times something nice, which is the index density.

Worked example 4.57: A first run of the recipe: $\bar\partial$ on a Riemann surface

Let $X = \Sigma_g$ be a compact Riemann surface ($n = 2$) and take the two-term complex $\bar\partial : \Gamma(E^0) \to \Gamma(E^1)$ with $E^0 = X \times \mathbb{C}$ (functions) and $E^1 = \Lambda^{0,1}$ (forms $f\, d\bar z$). The real tangent bundle has one block angle x, which is also the first Chern class of the holomorphic tangent line bundle $T^{1,0}$, and $e(X) = x$. The bundle $\Lambda^{0,1}$ is the complex conjugate of $\Lambda^{1,0} = (T^{1,0})^*$, so it has Chern root $-(-x) = x$. Therefore $\operatorname{ch}(E^0 - E^1) = 1 - e^x$. In dimension 2, $\operatorname{td}(TX_{\mathbb{C}}) = 1 - \frac{p_1}{12} + \cdots = 1$ (there are no 4-forms). Then

$$\begin{aligned}
\operatorname{index} \bar\partial &\overset{(1)}{=} (-1)^1 \int_X \frac{1 - e^x}{x} \cdot 1 \\
&\overset{(2)}{=} -\int_X \frac{-x - \frac{1}{2}x^2 - \cdots}{x} \\
&\overset{(3)}{=} \int_X \Big(1 + \tfrac{1}{2}x + \cdots\Big) \\
&\overset{(4)}{=} \tfrac{1}{2} \int_X e(X) = \tfrac{1}{2}(2 - 2g) = 1 - g.
\end{aligned}$$

(1) The recipe (4.101) with $n/2 = 1$.
(2) Expand $e^x = 1 + x + \frac{1}{2}x^2 + \cdots$. The degree-0 part of the numerator cancels, which is the ellipticity of the complex showing up.
(3) Cancel the common factor x as a formal polynomial division.
(4) Only the degree-2 piece $\frac{1}{2}x$ is integrated; $x = e(X)$ and $\int_{\Sigma_g} e = 2 - 2g$ (Gauss–Bonnet, which the de Rham example below derives from the same formula).

This agrees with the direct count $\operatorname{index} \bar\partial = \dim\{\text{holomorphic functions}\} - \dim\{\text{holomorphic 1-forms}\} = 1 - g$ of Section 4.10. The sign $(-1)^{n/2}$ was essential: without it we would have obtained $g - 1$.

Watch out: The formula needs to know how E^p is built from TX

If E^0 and E^1 were arbitrary line bundles with Chern classes a and b, the recipe would ask for "$(a-b)/x$", a degree-0 quantity not determined by cohomology. The index depends on the *symbol*, not only on the bundles. In each classical complex, E^p is a natural bundle (exterior powers, spinors) whose Chern roots are functions of the x_j, and the division by $x_1 \cdots x_m$ is an honest polynomial division. This is why the rest of the section treats the four classical complexes one at a time.

Reading Nash (pp. 101–103, (4.41)–(4.48))

Nash now assumes X orientable and D differential, introduces the Thom isomorphism $\psi(a) = \pi^*(a)\cdot u$ (4.41) ("for homology", with the symbol ϕ reserved for its K-theory cousin), applies ψ^{-1} to the index formula (4.42), recalls the K-class of a complex (4.43), writes the symbol complex (4.44) and its Chern character (4.45)–(4.46), and quotes $\psi^{-1}\{\pi^*\,\mathrm{ch}(\cdots)\} = \mathrm{ch}(\cdots)/e(X)$ from Atiyah–Singer (4.47), arriving at (4.48). We have supplied: an explicit Thom form, the proof that it restricts to the Euler class (in rank 2), the Poincaré–Hopf interpretation, the derivation of (4.47) via the zero section, and the origin of the sign $(-1)^{n(n+1)/2} \to (-1)^{n/2}$. Nash's "$\psi^{-1}\{\pi^*u\} = u/e(X)$" uses u for a generic class on X, not the Thom class.

Recap

- The Thom class u of an oriented bundle is a closed unit bump in the fibres; $\psi(a) = \pi^*a\cdot u$ and ψ^{-1} is integration over the fibre.
- u restricted to the zero section is the Euler class e; pulled back by a vector field it counts zeros (Poincaré–Hopf).
- An elliptic complex has symbol class $\sum(-1)^p[\pi^*E^p]$; on the zero section its Chern character is $\mathrm{ch}(\sum(-1)^pE^p) = b\cdot e(X)$.
- Nash (4.48): $\operatorname{index} D = (-1)^{n/2}\int_X \mathrm{ch}(\sum(-1)^pE^p)\ \mathrm{td}(TX_{\mathbb{C}})/e(X)$, the division done formally in Chern roots.

4.12 The de Rham complex and the Gauss–Bonnet theorem

The remaining examples are the four classical elliptic complexes, where geometry and topology meet. The simplest is the de Rham complex. Its index is a number every physicist has met: the Euler characteristic $\chi(X) = V - E + F$ of a triangulated surface, or more generally the alternating sum of Betti numbers. The index theorem will express it as the integral of a curvature density, the Euler form. In two dimensions this is the Gauss–Bonnet theorem, $\int_X K\,\mathrm{d}A = 2\pi\chi(X)$. In four dimensions it is the Chern–Gauss–Bonnet theorem, whose integrand is the "Gauss–Bonnet term" $R_{abcd}R^{abcd} - 4R_{ab}R^{ab} + R^2$ familiar from higher-curvature gravity.

4.12.1 The complex and its symbol

We use *complex-valued* forms, because the index theorem is phrased with the Chern character, which needs complex vector bundles. Nothing changes except that coefficients are allowed to

be complex:

$$\cdots \xrightarrow{\mathrm{d}_{p-1}} \Omega^p(X) \xrightarrow{\mathrm{d}_p} \Omega^{p+1}(X) \xrightarrow{\mathrm{d}_{p+1}} \cdots, \qquad \Omega^p(X) = \Gamma\big(X, \Lambda^p T^*X_{\mathbb{C}}\big), \tag{Nash 4.49}$$

where $T^*X_{\mathbb{C}} = T^*X \otimes_{\mathbb{R}} \mathbb{C}$ is the complexified cotangent bundle. A p-form is

$$\omega = \tfrac{1}{p!}\, \omega_{\mu_1 \cdots \mu_p}(x)\, \mathrm{d}x^{\mu_1} \wedge \cdots \wedge \mathrm{d}x^{\mu_p}$$

with complex component functions, and $\mathrm{d} \circ \mathrm{d} = 0$ makes this a complex.

The symbol of d**.** To read off a symbol, feed the operator a rapidly oscillating form $e^{it\phi(x)}\omega$ and keep the leading power of t:

$$\begin{aligned} \mathrm{d}\big(e^{it\phi}\omega\big) &\overset{(1)}{=} it\, e^{it\phi}\, \mathrm{d}\phi \wedge \omega + e^{it\phi}\, \mathrm{d}\omega \\ &\overset{(2)}{=} e^{it\phi}\big(it\, \mathrm{d}\phi \wedge \omega + O(t^0)\big). \end{aligned}$$

(1) The Leibniz rule for d acting on a function times a form.
(2) Collect powers of t.

With $p = t\, \mathrm{d}\phi$ the covector, the principal symbol is

$$\sigma_1(\mathrm{d})(x, p)\, \omega = i\, p \wedge \omega. \tag{4.102}$$

So the symbol sequence is "wedge with ip":

$$\cdots \xrightarrow{ip\wedge} \pi^*\big(\Lambda^p T^*X_{\mathbb{C}}\big) \xrightarrow{ip\wedge} \pi^*\big(\Lambda^{p+1} T^*X_{\mathbb{C}}\big) \xrightarrow{ip\wedge} \cdots. \tag{Nash 4.50}$$

Exactness of the symbol sequence. Let ι_p denote the interior product with the vector $p^\sharp$ dual to p (using a metric). It is an antiderivation, $\iota_p(\alpha \wedge \beta) = (\iota_p\alpha) \wedge \beta + (-1)^{\deg \alpha} \alpha \wedge \iota_p\beta$, and $\iota_p p = |p|^2$. Then for any form ω,

$$\begin{aligned} \big(p \wedge \iota_p + \iota_p\, p\wedge\big)\omega &\overset{(1)}{=} p \wedge \iota_p\omega + (\iota_p p)\, \omega - p \wedge \iota_p\omega \\ &\overset{(2)}{=} |p|^2\, \omega. \end{aligned} \tag{4.103}$$

(1) The antiderivation rule applied to $\iota_p(p \wedge \omega)$, with $\deg p = 1$ giving the minus sign.
(2) The two $p \wedge \iota_p\omega$ terms cancel; $\iota_p p = |p|^2$.

Now suppose $p \neq 0$ and $p \wedge \omega = 0$. Then (4.103) gives $\omega = p \wedge \big(\iota_p\omega/|p|^2\big)$: ω is in the image of $p\wedge$. So kernel = image at every step: the symbol sequence is exact away from $p = 0$, and the de Rham complex is elliptic. (This is the "same argument as for (4.20)" that Nash refers to.)

Worked example 4.58: The symbol complex of d in two dimensions, as matrices

On a surface use the bases $\{1\}$, $\{\mathrm{d}x^1, \mathrm{d}x^2\}$, $\{\mathrm{d}x^1 \wedge \mathrm{d}x^2\}$ for $\Lambda^0, \Lambda^1, \Lambda^2$. Then $ip \wedge 1 = i(p_1\mathrm{d}x^1 + p_2\mathrm{d}x^2)$ and $ip \wedge (a\, \mathrm{d}x^1 + b\, \mathrm{d}x^2) = i(p_1 b - p_2 a)\, \mathrm{d}x^1 \wedge \mathrm{d}x^2$. So the symbol complex is

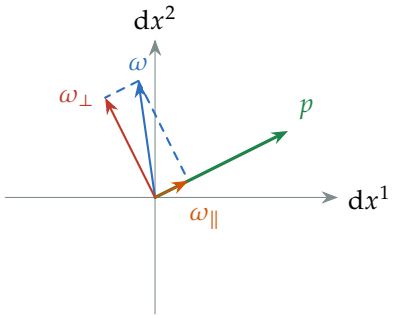
$$0 \longrightarrow \mathbb{C} \xrightarrow{i\,(p_1,\ p_2)^T} \mathbb{C}^2 \xrightarrow{i\,(-p_2\ \ p_1)} \mathbb{C} \longrightarrow 0.$$

Check that it is a complex and that it is exact:

$$\sigma(\mathrm{d}_1)\,\sigma(\mathrm{d}_0) \overset{(1)}{=} i^2 \begin{pmatrix} -p_2 & p_1 \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \end{pmatrix} \overset{(2)}{=} -(-p_2 p_1 + p_1 p_2) = 0.$$

(1) Compose the two matrices.
(2) Multiply out.

For $p \neq 0$: the first map is injective (a nonzero column), the last is surjective (a nonzero row), and the kernel of the row $(-p_2, p_1)$ is the line spanned by $(p_1, p_2)^T$, which is the image of the first map. The alternating sum of the ranks is $1 - 2 + 1 = 0$, as it must be for an exact sequence.



$p \wedge \omega = p \wedge \omega_\perp$: only the part of ω *not* along p survives.

$p \wedge \omega = 0 \iff \omega = \omega_\parallel = p \wedge (\iota_p \omega / |p|^2)$.

Kernel of $p\wedge$ = image of $p\wedge$: the symbol complex is exact for every $p \neq 0$, and fails to be exact only at $p = 0$.

Figure 4.34: Why the symbol complex of d is exact. Wedging with a nonzero covector p kills exactly the components along p (orange) and is injective on the rest (red). So anything killed by $p\wedge$ is itself a multiple of p. The identity $p \wedge \iota_p + \iota_p\, p\wedge = |p|^2$ is the algebraic form of this picture.

4.12.2 The index of the de Rham complex is the Euler characteristic

The underlying elliptic operator is obtained by rolling up the complex:

$$\mathrm{d} + \mathrm{d}^* : \left(\Omega^0 \oplus \Omega^2 \oplus \cdots\right) \longrightarrow \left(\Omega^1 \oplus \Omega^3 \oplus \cdots\right). \tag{Nash 4.51}$$

Here d^* is the adjoint of d with respect to the L^2 inner product $\langle \omega, \nu \rangle = \int_X \omega \wedge \star\bar{\nu}$; it lowers degree by one. Its symbol is the adjoint of $ip\wedge$, namely $-i\,\iota_p$. So

$$\begin{aligned} \sigma(\mathrm{d} + \mathrm{d}^*)^2 &\overset{(1)}{=} \left(i\, p \wedge -i\, \iota_p\right)^2 \\ &\overset{(2)}{=} -\left(p \wedge p \wedge - p \wedge \iota_p - \iota_p\, p \wedge + \iota_p \iota_p\right) \\ &\overset{(3)}{=} p \wedge \iota_p + \iota_p\, p\wedge = |p|^2 . \end{aligned}$$

(1) The symbols of d and d^*.
(2) Expand the square, with $i^2 = -1$.
(3) $p \wedge p = 0$ and $\iota_p \iota_p = 0$ (both are antisymmetric in a repeated slot); then (4.103).

The square of the symbol is $|p|^2$ times the identity: $\mathrm{d} + \mathrm{d}^*$ is a "square root of the Laplacian" $\Delta = (\mathrm{d} + \mathrm{d}^*)^2 = \mathrm{d}\mathrm{d}^* + \mathrm{d}^*\mathrm{d}$, which is the statement that $p \wedge -\iota_p$ generates a Clifford algebra (Chapter T4). In particular $\sigma(\mathrm{d} + \mathrm{d}^*)$ is invertible for $p \neq 0$: the operator is elliptic.

Hodge theory. On a compact manifold, each de Rham class has a unique **harmonic** representative, $\Delta\omega = 0$, equivalently $\mathrm{d}\omega = 0$ and $\mathrm{d}^*\omega = 0$. (Reason: $\langle \Delta\omega, \omega\rangle = \|\mathrm{d}\omega\|^2 + \|\mathrm{d}^*\omega\|^2$, so harmonic means closed and co-closed; and the harmonic representative minimises $\|\omega\|^2$ within its class.) Hence

$$\operatorname{Ker}(\mathrm{d}+\mathrm{d}^*)\big|_{\Omega^p} \;=\; \mathcal{H}^p \;\cong\; H^p_{\mathrm{dR}}(X;\mathbb{C}),$$

the cohomology $H^p(E)$ of the complex is de Rham cohomology, and its dimension is the p-th **Betti number** b_p. (With complex coefficients, $b_p = \dim_{\mathbb{C}} H^p(X;\mathbb{C})$, but this equals the usual $\dim_{\mathbb{R}} H^p(X;\mathbb{R})$: complexifying a real vector space does not change its dimension. This is Nash's footnote 3.) The index of the complex is therefore

$$\begin{aligned}
\operatorname{index}\mathrm{d} &\overset{(1)}{=} \dim\operatorname{Ker}(\mathrm{d}+\mathrm{d}^*)_{\text{even}} - \dim\operatorname{Ker}(\mathrm{d}+\mathrm{d}^*)_{\text{odd}} \\
&\overset{(2)}{=} \sum_i (-1)^i \dim_{\mathbb{C}} H^i_{\mathrm{dR}}(X;\mathbb{C}) \\
&\overset{(3)}{=} \sum_i (-1)^i b_i = \chi(X). \qquad \text{(Nash 4.52)}
\end{aligned}$$

(1) The index of the rolled-up operator (Nash 4.51): its cokernel on odd forms is the kernel of its adjoint, which is $\mathrm{d}+\mathrm{d}^*$ on odd forms.
(2) Hodge theory: harmonic forms of degree i are the de Rham classes.
(3) Definition of Betti numbers and of the Euler–Poincaré characteristic.

Worked example 4.59: Harmonic forms on S^2 and T^2

Torus $T^2 = \mathbb{R}^2/\mathbb{Z}^2$ with the flat metric. The harmonic forms are those with constant coefficients: $1;\ \mathrm{d}x, \mathrm{d}y;\ \mathrm{d}x\wedge\mathrm{d}y$. So $b_0 = 1$, $b_1 = 2$, $b_2 = 1$ and

$$\operatorname{index}\mathrm{d} = \underbrace{(1+1)}_{\text{even: } 1,\ \mathrm{d}x\wedge\mathrm{d}y} - \underbrace{2}_{\text{odd: } \mathrm{d}x,\mathrm{d}y} = 0 = \chi(T^2).$$

Sphere S^2 with the round metric. Harmonic 0-forms are constants. Harmonic 2-forms are constant multiples of the area form $\sin\theta\,\mathrm{d}\theta\wedge\mathrm{d}\varphi$. There are no harmonic 1-forms: a closed 1-form on S^2 is exact, $\omega = \mathrm{d}f$ (every loop bounds), and then $\mathrm{d}^*\mathrm{d}f = 0$ forces f constant, so $\omega = 0$. Hence

$$\operatorname{index}\mathrm{d} = (1+1) - 0 = 2 = \chi(S^2).$$

In both cases the count agrees with $V - E + F$ for a triangulation: a tetrahedron has $4 - 6 + 4 = 2$; the standard square torus with one vertex, two edges and one face has $1 - 2 + 1 = 0$.

Physical picture: Supersymmetric quantum mechanics and the Witten index

Think of differential forms on X as the states of a quantum particle moving on X with fermionic partners: a p-form has fermion number p (one fermion for each $\mathrm{d}x^\mu$). The supercharge is $Q = \mathrm{d}+\mathrm{d}^*$, the Hamiltonian $H = Q^2 = \Delta$, and $(-1)^F$ is $(-1)^p$. States with $H > 0$ come in boson–fermion pairs (Q maps one to the other, preserving the energy), so they cancel in the **Witten index**

$$\operatorname{Tr}(-1)^F e^{-\beta H} = \sum_p (-1)^p \dim\mathcal{H}^p = \chi(X),$$

for every β. Taking $\beta \to 0$ (high temperature, short paths) turns the trace into a local integral over X; its density is the Euler form. This is the path-integral proof of Gauss–Bonnet due to

Alvarez-Gaumé and Witten, and the heat-kernel proof of the index theorem is its rigorous version.

4.12.3 The characteristic-class side

Now evaluate the right-hand side of (4.101). The bundles are $E^p = \Lambda^p T^*X_{\mathbb{C}}$, so we need

$$\mathrm{ch}\big([\sigma_1(\mathrm{d})]\big) \;\to\; \mathrm{ch}\Big(\sum_p (-1)^p [E^p]\Big) = \sum_p (-1)^p \,\mathrm{ch}\big(\Lambda^p T^*X_{\mathbb{C}}\big). \tag{Nash 4.55}$$

Exterior powers by the splitting principle. Let F be any complex bundle of rank N with Chern roots $y_1, \dots, y_N$, i.e. formally $F = L_1 \oplus \cdots \oplus L_N$, $c_1(L_i) = y_i$. By (4.75),

$$\Lambda^p F = \bigoplus_{1 \le i_1 < \cdots < i_p \le N} L_{i_1} \otimes \cdots \otimes L_{i_p}. \tag{Nash 4.56}$$

Apply ch, using additivity and multiplicativity (4.73):

$$\begin{aligned} \mathrm{ch}(\Lambda^p F) &\overset{(1)}{=} \sum_{i_1<\cdots<i_p} \mathrm{ch}(L_{i_1}) \cdots \mathrm{ch}(L_{i_p}) \\ &\overset{(2)}{=} \sum_{i_1<\cdots<i_p} e^{y_{i_1}} \cdots e^{y_{i_p}} \;=\; \sum_{i_1<\cdots<i_p} e^{y_{i_1}+\cdots+y_{i_p}}. \end{aligned} \tag{Nash 4.57–4.58}$$

(1) ch of a direct sum is the sum; ch of a tensor product is the product.
(2) $\mathrm{ch}(L) = e^{c_1(L)}$ for a line bundle.

Now the alternating sum. The key observation is that $\sum_p (-1)^p \sum_{i_1<\cdots<i_p} t_{i_1} \cdots t_{i_p} = \prod_i (1 - t_i)$: expanding the product, each subset $\{i_1, \dots, i_p\}$ of factors from which we take $-t_i$ contributes $(-1)^p t_{i_1} \cdots t_{i_p}$. With $t_i = e^{y_i}$:

$$\sum_{p=0}^{N} (-1)^p \,\mathrm{ch}(\Lambda^p F) = \prod_{i=1}^{N} \big(1 - e^{y_i}\big). \tag{Nash 4.59}$$

Worked example 4.60: Checking (Nash 4.59) in rank 1 and rank 2

Rank 1: $\Lambda^0 F = \mathbb{C}$, $\Lambda^1 F = F$, so the left side is $1 - e^{y}$. ✓
Rank 2: $\Lambda^0 F = \mathbb{C}$, $\Lambda^1 F = L_1 \oplus L_2$, $\Lambda^2 F = L_1 \otimes L_2$ (the determinant line bundle), so

$$\begin{aligned} \sum_p (-1)^p \,\mathrm{ch}(\Lambda^p F) &\overset{(1)}{=} 1 - \big(e^{y_1} + e^{y_2}\big) + e^{y_1+y_2} \\ &\overset{(2)}{=} \big(1 - e^{y_1}\big)\big(1 - e^{y_2}\big). \end{aligned}$$

(1) (Nash 4.57–4.58) for $p = 0, 1, 2$.
(2) Factorise: $(1-a)(1-b) = 1 - a - b + ab$.

The lowest-order term is $(-y_1)(-y_2) = y_1 y_2 = c_2(F)$: the alternating sum starts in degree $2N$, the top Chern class. That is the "divisibility by the Euler class" promised in Section 4.11.

Recognising the Todd class. Pull out the lowest-order factor of each term:

$$\begin{aligned}\prod_{i=1}^{N}\bigl(1-e^{y_i}\bigr) &\overset{(1)}{=} \prod_{i=1}^{N}(-y_i)\cdot\prod_{i=1}^{N}\frac{1-e^{y_i}}{-y_i}\\ &\overset{(2)}{=} (-1)^N y_1\cdots y_N\cdot\Bigl(\prod_{i=1}^{N}\frac{-y_i}{1-e^{y_i}}\Bigr)^{-1}\\ &\overset{(3)}{=} \frac{c_N(F^*)}{\mathrm{td}(F^*)}.\end{aligned}\tag{Nash 4.60}$$

(1) Multiply and divide each factor by $-y_i$.
(2) Rearrange the signs and invert the second product.
(3) The dual bundle has roots $-y_i$: $c_N(F^*)=\prod(-y_i)=(-1)^N y_1\cdots y_N$, and $\mathrm{td}(F^*)=\prod\frac{-y_i}{1-e^{y_i}}$ by (4.78).

Putting it together. Take $F=T^*X_{\mathbb{C}}$, which has complex rank $N=n$. It is the complexification of a real bundle, so $F\cong F^*$ (Section 4.9), and $T^*X_{\mathbb{C}}\cong TX_{\mathbb{C}}$ via the metric. Its Chern roots are $\pm x_1,\dots,\pm x_m$, $n=2m$. Then

$$\begin{aligned}\operatorname{index}\mathrm{d} &\overset{(1)}{=} (-1)^{n/2}\int_X \frac{c_n(TX_{\mathbb{C}})}{e(X)\ \mathrm{td}(TX_{\mathbb{C}})}\ \mathrm{td}(TX_{\mathbb{C}})\\ &\overset{(2)}{=} (-1)^{n/2}\int_X \frac{c_n(TX_{\mathbb{C}})}{e(X)}\\ &\overset{(3)}{=} (-1)^{n/2}\int_X \frac{(-1)^{n/2}\,e(X)^2}{e(X)}\\ &\overset{(4)}{=} \int_X e(X).\end{aligned}\tag{Nash 4.61–4.62}$$

(1) Nash's formula (4.101) with the numerator (Nash 4.60), $F^*=F=TX_{\mathbb{C}}$.
(2) The Todd classes cancel.
(3) The top Chern class of a complexified real bundle, (4.80): $c_{2m}(V_{\mathbb{C}})=(-1)^m e(V)^2$.
(4) $(-1)^{n/2}(-1)^{n/2}=1$; cancel one factor $e(X)$ (formally, in the roots).

It is instructive to redo this directly in the roots, without the intermediate c_N/td step, to see the division by $e(X)$ happen honestly:

$$\begin{aligned}\prod_{j=1}^{m}\bigl(1-e^{x_j}\bigr)\bigl(1-e^{-x_j}\bigr) &\overset{(1)}{=} \prod_{j=1}^{m}\Bigl(-\bigl(e^{x_j/2}-e^{-x_j/2}\bigr)^2\Bigr)\\ &\overset{(2)}{=} (-1)^m\prod_j x_j^2\cdot\prod_j\Bigl(\frac{\sinh(x_j/2)}{x_j/2}\Bigr)^2 \;=\; (-1)^m\,\frac{e(X)^2}{\mathrm{td}(TX_{\mathbb{C}})}.\end{aligned}$$

(1) $(1-e^{x})(1-e^{-x})=2-e^{x}-e^{-x}=-(e^{x/2}-e^{-x/2})^2$.
(2) $e^{x/2}-e^{-x/2}=2\sinh(x/2)=x\cdot\frac{\sinh(x/2)}{x/2}$; then $\prod x_j=e(X)$ and (4.81).

Dividing by $e(X)$ leaves $(-1)^m e(X)/\mathrm{td}(TX_{\mathbb{C}})$, a genuine polynomial in the roots; multiplying by $(-1)^m\,\mathrm{td}(TX_{\mathbb{C}})$ leaves exactly $e(X)$.

Key idea: The Gauss–Bonnet theorem from the index theorem (Nash (4.53)–(4.54))

For a compact oriented even-dimensional manifold,

$$\chi(X) = \sum_i (-1)^i b_i = \int_X e(X), \qquad e(X) = \mathrm{Pf}\Big(\frac{R}{2\pi}\Big). \tag{4.104}$$

A topological count (the alternating number of harmonic forms) equals the integral of a local curvature density.

4.12.4 Explicit Euler densities

Dimension 2. The curvature of an oriented surface in an orthonormal coframe (e^1, e^2) is the single 2-form $R^1{}_2 = K\, e^1 \wedge e^2 = K\, \mathrm{d}A$, with K the Gaussian curvature. The Pfaffian of a 2×2 antisymmetric matrix is its upper-right entry, so

$$e = \frac{1}{2\pi} K\, \mathrm{d}A, \qquad \int_X K\, \mathrm{d}A = 2\pi\chi(X) = 2\pi(2-2g). \tag{4.105}$$

Dimension 4. With $R^{ab} = \tfrac{1}{2} R^{ab}{}_{cd}\, e^c \wedge e^d$ and the Pfaffian formula of Section 4.9,

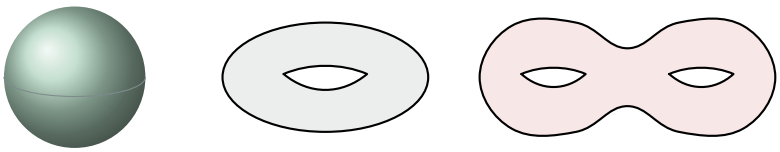
$$\begin{aligned} e &\overset{(1)}{=} \frac{1}{2^2\, 2!}\, \epsilon_{abcd}\, \frac{R^{ab}}{2\pi} \wedge \frac{R^{cd}}{2\pi} = \frac{1}{32\pi^2}\, \epsilon_{abcd}\, R^{ab} \wedge R^{cd} \\ &\overset{(2)}{=} \frac{1}{128\pi^2}\, \epsilon_{abcd}\, \epsilon^{efgh}\, R^{ab}{}_{ef}\, R^{cd}{}_{gh}\ \mathrm{vol} \\ &\overset{(3)}{=} \frac{1}{32\pi^2}\Big(R_{abcd}R^{abcd} - 4R_{ab}R^{ab} + R^2\Big)\mathrm{vol}. \end{aligned} \tag{4.106}$$

(1) The Pfaffian with $m = 2$: $\mathrm{Pf}(M) = \frac{1}{2^m m!}\epsilon_{a_1\cdots a_4} M^{a_1a_2} M^{a_3a_4}$.

(2) Insert $R^{ab} = \frac{1}{2} R^{ab}{}_{ef} e^e \wedge e^f$ (two factors of $\frac{1}{2}$) and $e^e \wedge e^f \wedge e^g \wedge e^h = \epsilon^{efgh}\mathrm{vol}$.

(3) The contraction identity $\epsilon_{abcd}\epsilon^{efgh} R^{ab}{}_{ef} R^{cd}{}_{gh} = 4\big(R_{abcd}R^{abcd} - 4R_{ab}R^{ab} + R^2\big)$, obtained by writing $\epsilon\epsilon$ as a determinant of Kronecker deltas and using $R^a{}_{bad} = R_{bd}$, $R^a{}_a = R$.

Check on the unit S^4: $R_{abcd} = \delta_{ac}\delta_{bd} - \delta_{ad}\delta_{bc}$ gives $R_{abcd}R^{abcd} = 2\cdot 4\cdot 3 = 24$, $R_{ab} = 3\delta_{ab}$ so $R_{ab}R^{ab} = 36$, and $R = 12$. The bracket is $24 - 144 + 144 = 24$, and $\int e = \frac{24}{32\pi^2}\cdot\frac{8\pi^2}{3} = 2 = \chi(S^4)$, agreeing with Example 4.51.



$g = 0$, $\chi = 2$
$K > 0$: $\int K\mathrm{d}A = 4\pi$

$g = 1$, $\chi = 0$
flat: $K = 0$

$g = 2$, $\chi = -2$
$K = -1$: area $4\pi(g-1)$

Figure 4.35: Gauss–Bonnet for surfaces: $\int_X K\, \mathrm{d}A = 2\pi\chi = 2\pi(2-2g)$. The sphere is positively curved (green), the torus can be made flat, and a surface of genus $g \geq 2$ can be given constant curvature -1 (red), whose area must then be $4\pi(g-1)$. However we deform the metric, the total curvature cannot change.

Worked example 4.61: Gauss–Bonnet on the three model geometries

Round sphere of radius a. $K = 1/a^2$ and area $4\pi a^2$: $\int K\,\mathrm{d}A = 4\pi = 2\pi\cdot 2$. The radius drops out, as it must for a topological invariant.
Flat torus. $K = 0$ everywhere: $\int K\,\mathrm{d}A = 0 = 2\pi\cdot 0$.
Genus g hyperbolic surface. Uniformisation provides a metric with $K = -1$. Then

$$\mathrm{Area}(\Sigma_g) \overset{(1)}{=} -\int_{\Sigma_g} K\,\mathrm{d}A \overset{(2)}{=} -2\pi\chi(\Sigma_g) = -2\pi(2-2g) = 4\pi(g-1).$$

(1) $K = -1$, so $\mathrm{d}A = -K\,\mathrm{d}A$.
(2) Gauss–Bonnet (4.105).

Geometry (the area) is fixed by topology (the genus).

Worked example 4.62: Four-dimensional Euler numbers

$S^2\times S^2$. The curvature is block diagonal (one block per factor), so the Pfaffian factorises: $e(S^2\times S^2) = e(S^2)\wedge e(S^2)$, and $\int e = 2\cdot 2 = 4$. Directly: $b_0 = 1, b_2 = 2, b_4 = 1$, $\chi = 4$.
$\mathbb{CP}^2$. By Example 4.50, $e(\mathbb{CP}^2) = c_2(T) = 3h^2$, so $\int e = 3 = 1+1+1$.
T^4 *and* $S^1\times S^3$. Both admit nowhere-vanishing vector fields (the circle direction), and both have $\chi = 0$; T^4 is flat, so the Euler density vanishes pointwise.
K3 surface. Its Betti numbers are $1, 0, 22, 0, 1$, so $\chi = 24$, and this is the integral of (4.106) for its Ricci-flat (Calabi–Yau) metric, where only the $R_{abcd}R^{abcd}$ term survives: $\frac{1}{32\pi^2}\int_{K3} R_{abcd}R^{abcd}\,\mathrm{vol} = 24$.

Physical picture: Why string perturbation theory is a genus expansion

On a two-dimensional worldsheet Σ, the Einstein–Hilbert action $\frac{1}{4\pi}\int_\Sigma \sqrt{g}\,R\,\mathrm{d}^2\sigma$ is (with $R = 2K$) exactly $\chi(\Sigma)$ by (4.105): it has no dynamics, only topology. Adding $\lambda\chi(\Sigma)$ to the string action weights a worldsheet of genus g by $e^{-\lambda(2-2g)} = g_s^{2g-2}$ with $g_s = e^{\lambda}$. This is why the string coupling counts handles (Chapter 9). In four dimensions the Gauss–Bonnet density (4.106) plays the same role: adding it to the gravitational action changes nothing in the equations of motion, only the weight $e^{-\lambda\chi}$ of each topology in the path integral.

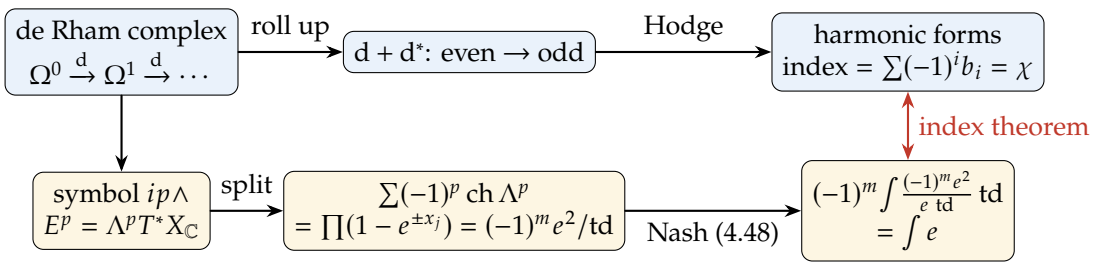


Figure 4.36: The logic of the de Rham example. Top row: the analytic side (count harmonic forms). Bottom row: the topological side (characteristic classes of the symbol). The index theorem equates the right-hand ends: this is the Gauss–Bonnet theorem $\chi(X) = \int_X e(X)$.

Watch out: Odd dimensions

In odd dimensions both sides vanish separately: $e(X) = 0$ for an odd-rank bundle (Section 4.11), and $\chi(X) = 0$ by Poincaré duality ($b_i = b_{n-i}$ and $n - i$ has the opposite parity from i, so the terms cancel in pairs). The index theorem is consistent with the odd-dimensional vanishing of Section 4.10, but tells us nothing new there.

Reading Nash (pp. 103–106, (4.49)–(4.62))

Nash takes complex-valued forms (4.49), states that the symbol complex (4.50) is exact "by the same argument as for (4.20)" (we gave it: $p \wedge \iota_p + \iota_p p \wedge = |p|^2$), rolls up to $\mathrm{d} + \mathrm{d}^*$ (4.51), and gets $\operatorname{index} \mathrm{d} = \chi(X)$ (4.52) from Hodge theory. On the topological side he computes (4.55)–(4.60) by the splitting principle exactly as we did, then uses $F = F^*$ and his relation (3.124) $c_n(TX_{\mathbb{C}}) = (-1)^{n/2} e^2(X)$ (our (4.80)) to obtain $\int_X e(X) = \chi(X)$ (4.53)–(4.54), (4.61)–(4.62). In (4.61) his c_n is the top Chern class of the rank-n complex bundle $TX_{\mathbb{C}}$; he writes e^2 for $e \wedge e$.

Recap

- de Rham complex: symbol $ip\wedge$, exact for $p \neq 0$ because $p \wedge \iota_p + \iota_p p \wedge = |p|^2$; rolled-up operator $\mathrm{d} + \mathrm{d}^*$.
- Analytic index $= \sum (-1)^i b_i = \chi(X)$ by Hodge theory (the Witten index of SUSY quantum mechanics on X).
- Topological side: $\sum (-1)^p \operatorname{ch} \Lambda^p F = \prod (1 - e^{y_i}) = c_N(F^*)/\operatorname{td}(F^*)$, and with $c_n(TX_{\mathbb{C}}) = (-1)^{n/2} e^2$ the formula gives $\int_X e$.
- Gauss–Bonnet: $\int K \mathrm{d}A = 2\pi\chi$ in 2D; $\chi = \frac{1}{32\pi^2} \int (R^2_{abcd} - 4R^2_{ab} + R^2)$ in 4D.

4.13 The Dolbeault complex and the Riemann–Roch theorem

The de Rham complex counts *closed forms modulo exact ones*. On a complex manifold there is a finer question: how many *holomorphic* objects are there? How many holomorphic functions, holomorphic 1-forms, holomorphic sections of a line bundle? On a compact complex manifold these spaces are finite dimensional, and the numbers are among the most important invariants in algebraic geometry. They are also physics numbers: the number of holomorphic sections of a line bundle over a surface is the degeneracy of the lowest Landau level of a charged particle on that surface, and zero modes of two-dimensional chiral fermions are holomorphic sections.

Counting holomorphic sections one by one is hard. The individual numbers can jump when the complex structure is deformed. The index theorem computes an alternating sum of them, which cannot jump, as the integral of a curvature polynomial: the Todd class. This is the Riemann–Roch–Hirzebruch theorem. Historically it came first (Hirzebruch, 1954), and it was the model that led Atiyah and Singer to their general theorem.

4.13.1 Complex coordinates and (p, q)-forms

We first set up what Nash takes for granted from his Chapter 3 (Chapter 3).

A **complex manifold** of complex dimension m is a manifold of real dimension $n = 2m$ covered by charts with complex coordinates $z^j = x^j + iy^j$, $j = 1, \ldots, m$, whose transition

functions are holomorphic. (Nash writes $n/2$ for the complex dimension; we write $m = n/2$.) In each chart we have the complex 1-forms and the dual vector fields

$$\mathrm{d}z^j = \mathrm{d}x^j + i\,\mathrm{d}y^j, \quad \mathrm{d}\bar z^j = \mathrm{d}x^j - i\,\mathrm{d}y^j, \qquad \frac{\partial}{\partial z^j} = \tfrac12\Big(\frac{\partial}{\partial x^j} - i\frac{\partial}{\partial y^j}\Big), \quad \frac{\partial}{\partial \bar z^j} = \tfrac12\Big(\frac{\partial}{\partial x^j} + i\frac{\partial}{\partial y^j}\Big). \tag{4.107}$$

One checks $\mathrm{d}z^j(\partial/\partial z^k) = \delta^j_k$, $\mathrm{d}z^j(\partial/\partial \bar z^k) = 0$, and so on. A function is holomorphic exactly when $\partial f/\partial \bar z^j = 0$ for all j: this is the Cauchy–Riemann system, written compactly.

Working definition 4.33 ((p,q)-forms and the operators ∂, $\bar\partial$)

A (p,q)**-form** is a complex form that in every holomorphic chart [complex coordinates z^j whose changes $z \mapsto z'(z)$ are holomorphic] contains exactly p factors $\mathrm{d}z$ and q factors $\mathrm{d}\bar z$:

$$\omega = \frac{1}{p!\,q!}\,\omega_{j_1\cdots j_p \bar k_1\cdots \bar k_q}\,\mathrm{d}z^{j_1}\wedge\cdots\wedge\mathrm{d}z^{j_p}\wedge\mathrm{d}\bar z^{k_1}\wedge\cdots\wedge\mathrm{d}\bar z^{k_q}.$$

Holomorphic coordinate changes preserve this type, so (p,q) is a global notion. The bundle of (p,q)-forms is $\Lambda^{p,q}T^*X$, of complex rank $\binom{m}{p}\binom{m}{q}$ [choose p of the m differentials $\mathrm{d}z^j$ and q of the m differentials $\mathrm{d}\bar z^k$], and $\Lambda^k T^*X_{\mathbb C} = \bigoplus_{p+q=k}\Lambda^{p,q}$. The exterior derivative splits as $\mathrm{d} = \partial + \bar\partial$ with

$$\partial\omega = \frac{\partial \omega_{\dots}}{\partial z^j}\,\mathrm{d}z^j\wedge(\cdots), \qquad \bar\partial\omega = \frac{\partial\omega_{\dots}}{\partial\bar z^j}\,\mathrm{d}\bar z^j\wedge(\cdots),$$

so ∂ raises p by one and $\bar\partial$ raises q by one.

Why do ∂ and $\bar\partial$ separately square to zero? Expand $0 = \mathrm{d}^2 = (\partial+\bar\partial)^2$ and sort by type:

$$0 = \mathrm{d}^2\omega \overset{(1)}{=} \partial^2\omega + (\partial\bar\partial + \bar\partial\partial)\omega + \bar\partial^2\omega \overset{(2)}{=} \underbrace{\partial^2\omega}_{(p+2,q)} + \underbrace{(\partial\bar\partial+\bar\partial\partial)\omega}_{(p+1,q+1)} + \underbrace{\bar\partial^2\omega}_{(p,q+2)}.$$

(1) Expand the square of $\mathrm{d} = \partial + \bar\partial$.
(2) Each term has a definite type, and forms of different types are linearly independent.

So each piece vanishes separately: $\partial^2 = 0$, $\bar\partial^2 = 0$ and $\partial\bar\partial = -\bar\partial\partial$. (The splitting $\mathrm{d} = \partial + \bar\partial$ with no extra pieces is where the holomorphy of the transition functions, i.e. the integrability of the complex structure, is used.)

Worked example 4.63: One complex dimension: $\bar\partial$ is the Cauchy–Riemann operator

On $\mathbb C$ with $z = x + iy$, a function $f = u + iv$ has

$$\bar\partial f \overset{(1)}{=} \frac{\partial f}{\partial \bar z}\,\mathrm{d}\bar z \overset{(2)}{=} \tfrac12\big(u_x + iv_x + iu_y - v_y\big)\mathrm{d}\bar z \overset{(3)}{=} \tfrac12\big[(u_x - v_y) + i(v_x + u_y)\big]\mathrm{d}\bar z.$$

(1) Definition of $\bar\partial$ on a $(0,0)$-form.
(2) Insert $\partial/\partial\bar z = \frac12(\partial_x + i\partial_y)$ from (4.107) and $f = u + iv$.
(3) Collect real and imaginary parts.

So $\bar\partial f = 0$ is exactly $u_x = v_y$, $u_y = -v_x$: the Cauchy–Riemann equations. The complex $0 \to \Omega^{0,0} \xrightarrow{\bar\partial} \Omega^{0,1} \to 0$ has kernel "holomorphic functions" and cokernel "$(0,1)$-forms that are not $\partial_{\bar z}$ of anything".

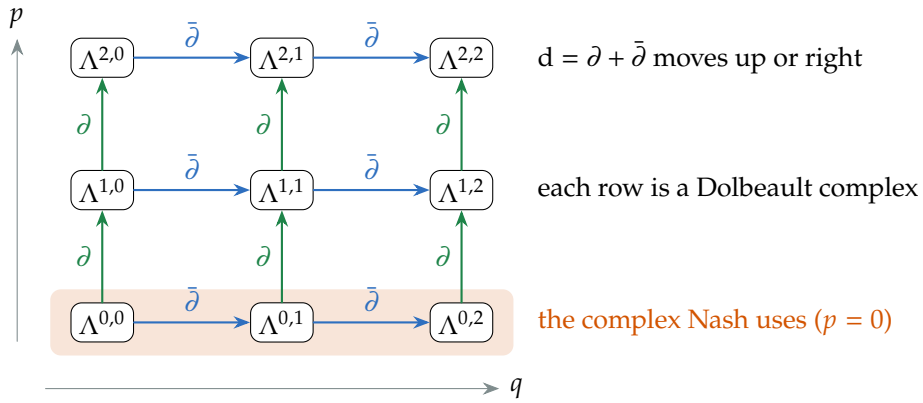


Figure 4.37: Forms on a complex surface ($m = 2$) sorted by type. The operator $\bar\partial$ moves one step right, ∂ one step up, and d is their sum, which changes the total degree $p + q$ by one. Each row $\Lambda^{p,0} \to \Lambda^{p,1} \to \Lambda^{p,2}$ is an elliptic complex; its cohomology groups are the $H^{p,q}_{\bar\partial}$. Nash sets $p = 0$ (orange).

4.13.2 The Dolbeault complex, its cohomology and its symbol

For each fixed p we get a complex, and its cohomology

$$\cdots \xrightarrow{\bar\partial_{p,q-1}} \Omega^{p,q}(X) \xrightarrow{\bar\partial_{p,q}} \Omega^{p,q+1}(X) \longrightarrow \cdots, \qquad H^{p,q}_{\bar\partial}(X) = \frac{\operatorname{Ker}\left(\bar\partial : \Omega^{p,q} \to \Omega^{p,q+1}\right)}{\operatorname{Im}\left(\bar\partial : \Omega^{p,q-1} \to \Omega^{p,q}\right)} \tag{Nash 4.63}$$

is **Dolbeault cohomology**. Its dimensions $h^{p,q} = \dim_{\mathbb{C}} H^{p,q}_{\bar\partial}(X)$ are the **Hodge numbers**. (Nash writes $h_{p,q}$.) Two of them are easy to interpret:

- $H^{0,0}_{\bar\partial} = \{f : \bar\partial f = 0\}$ is the space of global holomorphic functions. On a compact connected X these are constant (the maximum principle), so $h^{0,0} = 1$.
- $H^{p,0}_{\bar\partial} = \{\omega \in \Omega^{p,0} : \bar\partial\omega = 0\}$ (there is nothing to divide by) is the space of holomorphic p-forms.

The symbol. Feed $\bar\partial$ a rapidly oscillating form, exactly as for d in Section 4.12:

$$\begin{aligned}\bar\partial\left(e^{it\phi}\omega\right) &\overset{(1)}{=} it\, e^{it\phi}\,\bar\partial\phi \wedge \omega + e^{it\phi}\,\bar\partial\omega \\ &\overset{(2)}{=} e^{it\phi}\left(i\,p^{0,1} \wedge \omega + O(t^0)\right), \qquad p^{0,1} := t\,\bar\partial\phi.\end{aligned}$$

(1) Leibniz rule; $\bar\partial$ of a function ϕ is the $(0,1)$ part of $\mathrm{d}\phi$.
(2) Set the covector $p = t\,\mathrm{d}\phi$; then $t\,\bar\partial\phi$ is its $(0,1)$-part $p^{0,1}$.

So $\sigma_1(\bar\partial)(x,p) = i\,p^{0,1}\wedge$. The crucial point: for a *real* nonzero covector p, the $(0,1)$-part is never zero. Indeed $p = p^{1,0} + p^{0,1}$ with $p^{1,0} = \overline{p^{0,1}}$ (because p is real), so $p^{0,1} = 0$ would force $p = 0$. In

one complex dimension, $p = p_x\mathrm{d}x + p_y\mathrm{d}y$ gives

$$\begin{aligned} p &\overset{(1)}{=} p_x\,\tfrac{1}{2}(\mathrm{d}z + \mathrm{d}\bar z) + p_y\,\tfrac{1}{2i}(\mathrm{d}z - \mathrm{d}\bar z) \\ &\overset{(2)}{=} \tfrac{1}{2}(p_x - ip_y)\,\mathrm{d}z + \tfrac{1}{2}(p_x + ip_y)\,\mathrm{d}\bar z, \end{aligned}$$

(1) Invert (4.107): $\mathrm{d}x = \frac{1}{2}(\mathrm{d}z + \mathrm{d}\bar z)$, $\mathrm{d}y = \frac{1}{2i}(\mathrm{d}z - \mathrm{d}\bar z)$.
(2) Collect the coefficients of $\mathrm{d}z$ and $\mathrm{d}\bar z$.

and $p^{0,1} = \frac{1}{2}(p_x + ip_y)\mathrm{d}\bar z$ vanishes only at $p = 0$. The symbol of $\partial_{\bar z}$ is the complex number $\frac{i}{2}(p_x + ip_y)$, whose modulus is $\frac{1}{2}\,|p|$: the Cauchy–Riemann operator is a "square root" of the Laplacian, $4\partial_z\partial_{\bar z} = \partial_x^2 + \partial_y^2$. With $p^{0,1} \neq 0$ the Koszul argument (4.103) (with ι_p replaced by contraction with the vector dual to $p^{0,1}$) shows that the symbol sequence

$$\cdots \xrightarrow{\sigma_1(\bar\partial_{0,q-1})} \pi^*\big(\Lambda^{0,q}T^*X\big) \xrightarrow{\sigma_1(\bar\partial_{0,q})} \pi^*\big(\Lambda^{0,q+1}T^*X\big) \xrightarrow{\sigma_1(\bar\partial_{0,q+1})} \cdots \tag{Nash 4.66}$$

is exact for $p \neq 0$. The Dolbeault complex is elliptic.

The index. Following Nash we set $p = 0$ and call the resulting complex $\Omega^{0,0} \to \Omega^{0,1} \to \cdots \to \Omega^{0,m}$, and its index, simply index $\bar\partial$. Rolling it up as for de Rham gives the elliptic operator $\bar\partial + \bar\partial^* : \Omega^{0,\mathrm{even}} \to \Omega^{0,\mathrm{odd}}$, and the Hodge theorem for the Laplacian $\Box = \bar\partial\bar\partial^* + \bar\partial^*\bar\partial$ identifies $H^{0,q}_{\bar\partial}$ with harmonic $(0,q)$-forms. Therefore

$$\operatorname{index}\bar\partial = \sum_{q=0}^{m}(-1)^q \dim H^{0,q}_{\bar\partial}(X) = \sum_{q=0}^{m}(-1)^q h^{0,q} \;=:\; \chi(\mathcal{O}_X). \tag{Nash 4.64–4.65}$$

Nash writes $\chi(\mathcal{E})$, with $\mathcal{E}$ standing for the Dolbeault complex; in algebraic geometry it is written $\chi(X, \mathcal{O}_X)$, the Euler characteristic of the sheaf of holomorphic functions (Chapter 3), and it is called the **arithmetic genus** of X.

Worked example 4.64: The arithmetic genus of a torus and of a Riemann surface

Complex torus $X = \mathbb{C}/(\mathbb{Z} + \tau\mathbb{Z})$, $\operatorname{Im}\tau > 0$. The flat metric makes the harmonic forms those with constant coefficients: 1 in degree $(0,0)$ and $\mathrm{d}\bar z$ in degree $(0,1)$. So $h^{0,0} = h^{0,1} = 1$ and $\chi(\mathcal{O}) = 1 - 1 = 0$. In complex dimension m, $X = \mathbb{C}^m/\Lambda$, the harmonic $(0,q)$-forms are spanned by the $\binom{m}{q}$ constant forms $\mathrm{d}\bar z^{k_1} \wedge \cdots \wedge \mathrm{d}\bar z^{k_q}$, so

$$\chi(\mathcal{O}_{T^{2m}}) = \sum_q (-1)^q \binom{m}{q} = (1-1)^m = 0.$$

Riemann surface of genus g. Here $h^{0,0} = 1$ and $h^{0,1} = g$. The second number is the classical statement that a genus-g surface carries g independent holomorphic 1-forms: $h^{1,0} = g$, and complex conjugation maps harmonic $(1,0)$-forms to harmonic $(0,1)$-forms, so $h^{0,1} = h^{1,0} = g$. (For $g = 1$: $\mathrm{d}z$ and $\mathrm{d}\bar z$.) So $\chi(\mathcal{O}_{\Sigma_g}) = 1 - g$, which is what we found in Example 4.57.

4.13.3 The characteristic-class side: the Todd class appears

Now we evaluate Nash's formula (4.101) with $E^q = \Lambda^{0,q}T^*X$. Two small facts about complex bundles are needed.

Fact 1: $\Lambda^{0,1} \cong T^{1,0}$. Let us write T for the *holomorphic tangent bundle* $T^{1,0}X$, spanned by the $\partial/\partial z^j$. (Nash calls it TX in this example; see the pitfall below.) Its dual $T^* = \Lambda^{1,0}$ is spanned by the dz^j, and its conjugate $\Lambda^{0,1} = \overline{T^*}$ by the $d\bar z^j$. Given a Hermitian metric, the conjugate of a bundle is isomorphic to its dual ((4.74) and the remark after it), so $\Lambda^{0,1} = \overline{T^*} \cong (T^*)^* = T$. Therefore

$$\Lambda^{0,q}T^*X = \Lambda^q\big(\overline{T^*}\big) \cong \Lambda^q T.$$

Concretely, the metric turns $d\bar z^j$ into $g^{\bar j k}\,\partial/\partial z^k$.

Fact 2: $TX_{\mathbb C} = T \oplus \bar T$. A complexified real tangent vector $v = v^j\partial_{x^j} + w^j\partial_{y^j}$ with complex v^j, w^j can be rewritten in the basis $\partial/\partial z^j, \partial/\partial \bar z^j$: the complexified tangent bundle splits into the ∂_z-part T and the $\partial_{\bar z}$-part $\bar T \cong T^*$. If $x_1, \dots, x_m$ are the Chern roots of T, then $TX_{\mathbb C}$ has roots $x_1, \dots, x_m, -x_1, \dots, -x_m$. Comparing with the real picture of Section 4.9, the block angles of the underlying real bundle are exactly the x_j, and so

$$e(X) = x_1 \cdots x_m = c_m(T), \qquad \operatorname{td}(TX_{\mathbb C}) = \operatorname{td}(T)\,\operatorname{td}(\bar T) = \operatorname{td}(T)\,\operatorname{td}(T^*). \tag{Nash 4.69}$$

(On $\mathbb{CP}^2$ this is Example 4.50: $\int e = \int c_2(T) = 3$.)

The computation. Put $F = T$. By Fact 1 and the de Rham computation (Nash 4.59)–(Nash 4.60),

$$\operatorname{ch}\big(\sigma_1(\bar\partial)\big) \to \sum_{q=0}^{m}(-1)^q \operatorname{ch}\big(\Lambda^{0,q}\big) = \sum_{q=0}^{m}(-1)^q \operatorname{ch}\big(\Lambda^q F\big) = \prod_{j=1}^{m}\big(1 - e^{x_j}\big) = \frac{c_m(F^*)}{\operatorname{td}(F^*)}. \tag{Nash 4.67}$$

Now insert into the index formula:

$$\begin{aligned}
\operatorname{index}\bar\partial &\overset{(1)}{=} (-1)^m \int_X \frac{c_m(F^*)}{e(X)\,\operatorname{td}(F^*)}\,\operatorname{td}(TX_{\mathbb C}) \\
&\overset{(2)}{=} (-1)^m \int_X \frac{(-1)^m e(X)}{e(X)\,\operatorname{td}(F^*)}\,\operatorname{td}(F)\,\operatorname{td}(F^*) \\
&\overset{(3)}{=} \int_X \operatorname{td}(F) \;=\; \int_X \operatorname{td}(X).
\end{aligned} \tag{Nash 4.68, 4.70–4.71}$$

(1) Nash's formula (4.101) with $n/2 = m$ and the numerator (Nash 4.67).
(2) $c_m(F^*) = (-1)^m c_m(F)$ by (4.74), and $c_m(F) = e(X)$ by (Nash 4.69); also $\operatorname{td}(TX_{\mathbb C}) = \operatorname{td}(F)\operatorname{td}(F^*)$.
(3) $(-1)^m(-1)^m = 1$; the Euler classes cancel, and so does $\operatorname{td}(F^*)$. We abbreviate $\operatorname{td}(X) := \operatorname{td}(T^{1,0}X)$, as Nash does.

Let us also do it directly in the roots, to see the cancellation happen:

$$\begin{aligned}
(-1)^m \prod_{j=1}^{m} \frac{1 - e^{x_j}}{x_j} \cdot \frac{x_j}{1 - e^{-x_j}} \cdot \frac{-x_j}{1 - e^{x_j}} &\overset{(1)}{=} (-1)^m \prod_{j=1}^{m} \frac{-x_j}{1 - e^{-x_j}} \\
&\overset{(2)}{=} \prod_{j=1}^{m} \frac{x_j}{1 - e^{-x_j}} = \operatorname{td}(T).
\end{aligned}$$

(1) Cancel $1 - e^{x_j}$ and one x_j in each factor.
(2) Pull out one -1 per factor; $(-1)^m(-1)^m = 1$.

Key idea: The Riemann–Roch–Hirzebruch theorem (Nash (4.72))

For a compact complex manifold X of complex dimension m,

$$\chi(\mathcal{O}_X) = \sum_{q=0}^{m} (-1)^q h^{0,q} = \int_X \mathrm{td}(X), \qquad \mathrm{td}(X) = \prod_{j=1}^{m} \frac{x_j}{1 - e^{-x_j}}. \tag{4.108}$$

More generally, for a holomorphic vector bundle $V \to X$ (twist every $\Lambda^{0,q}$ by V; the Chern character just multiplies, as the next section explains in general):

$$\chi(X, V) = \sum_{q} (-1)^q \dim H^{0,q}_{\bar\partial}(X; V) = \int_X \mathrm{ch}(V)\ \mathrm{td}(X). \tag{4.109}$$

A count of holomorphic objects equals the integral of a curvature polynomial.

4.13.4 The Todd integrand in low dimensions

By (4.77), the top-degree part of $\mathrm{td}(X)$ in complex dimension $m = 1, 2, 3$ is

$$\chi(\mathcal{O}_X) = \begin{cases} \dfrac{1}{2}\displaystyle\int_X c_1, & m = 1, \\ \dfrac{1}{12}\displaystyle\int_X (c_1^2 + c_2), & m = 2 \quad \text{(Noether's formula)}, \\ \dfrac{1}{24}\displaystyle\int_X c_1 c_2, & m = 3, \end{cases} \tag{4.110}$$

where $c_j = c_j(T^{1,0}X)$. In curvature language: for a Kähler manifold the curvature of T is the Hermitian matrix of $(1,1)$-forms $\Phi = \frac{i}{2\pi} F$, and $c_1 = \mathrm{tr}\,\Phi$ is $\frac{1}{2\pi}$ times the Ricci form. For $m = 1$, $c_1(T) = \frac{1}{2\pi} K\,\mathrm{d}A$ (Example 4.49), so

$$\begin{aligned} \chi(\mathcal{O}_{\Sigma_g}) &\overset{(1)}{=} \frac{1}{2}\int_{\Sigma_g} c_1(T) \\ &\overset{(2)}{=} \frac{1}{4\pi}\int_{\Sigma_g} K\,\mathrm{d}A \\ &\overset{(3)}{=} \frac{1}{4\pi}\cdot 2\pi(2 - 2g) = 1 - g. \end{aligned}$$

(1) (4.110) with $m = 1$.
(2) The curvature of the tangent line bundle is the Gaussian curvature.
(3) Gauss–Bonnet (4.105).

This reproduces $h^{0,0} - h^{0,1} = 1 - g$ of Example 4.64. On a curve, Riemann–Roch for $\mathcal{O}$ *is* half of Gauss–Bonnet.

Worked example 4.65: Noether's formula on three complex surfaces

$\mathbb{CP}^2$. From Example 4.50, $c(T) = (1 + h)^3$, so $c_1 = 3h$, $c_2 = 3h^2$, $\int h^2 = 1$:

$$\chi(\mathcal{O}_{\mathbb{CP}^2}) = \tfrac{1}{12}(9 + 3) = 1.$$

Directly: $h^{0,0} = 1$, and $h^{0,1} = h^{0,2} = 0$ ($\mathbb{CP}^2$ has no holomorphic 1- or 2-forms; its only cohomology is in the classes $1, h, h^2$, all of type (p,p)).

$\mathbb{CP}^1 \times \mathbb{CP}^1$. Let a, b be the area classes of the two factors, $\int a = \int b = 1$ on the respective factor, so $a^2 = b^2 = 0$ and $\int ab = 1$. The tangent bundle is $T = \mathcal{O}(2) \boxplus \mathcal{O}(2)$, one line for each factor, with Chern roots $2a$ and $2b$. Then

$$\begin{aligned}\chi(\mathcal{O}) &\overset{(1)}{=} \tfrac{1}{12}\int \big(c_1^2 + c_2\big)\\ &\overset{(2)}{=} \tfrac{1}{12}\int \big((2a+2b)^2 + (2a)(2b)\big)\\ &\overset{(3)}{=} \tfrac{1}{12}\int \big(8ab + 4ab\big) = 1.\end{aligned}$$

(1) Noether's formula (4.110).
(2) $c_1 = x_1 + x_2$ and $c_2 = x_1 x_2$ with roots $x_1 = 2a$, $x_2 = 2b$.
(3) $a^2 = b^2 = 0$; $\int ab = 1$.

Also $\int c_2 = 4 = \chi(S^2 \times S^2)$, as it must be.
The K3 surface. This is a compact complex surface with $c_1 = 0$ (a nowhere-vanishing holomorphic 2-form $\Omega = \mathrm{d}z^1 \wedge \mathrm{d}z^2$ trivialises $\Lambda^{2,0}$) and $\chi = \int c_2 = 24$. So

$$\chi(\mathcal{O}_{K3}) = \tfrac{1}{12}(0 + 24) = 2,$$

and directly: $h^{0,0} = 1$, $h^{0,1} = 0$ (K3 is simply connected), $h^{0,2} = h^{2,0} = 1$ (spanned by $\bar{\Omega}$): $1 - 0 + 1 = 2$. ✓

Worked example 4.66: Integrality forbids a complex structure on S^4

Noether's formula says that $c_1^2 + c_2$ is divisible by 12 on every compact complex surface. Suppose S^4 were a complex surface. Then $c_2 = e$ integrates to $\chi(S^4) = 2$, and $c_1 \in H^2(S^4) = 0$, so $c_1^2 = 0$. The formula would give

$$\chi(\mathcal{O}_{S^4}) = \tfrac{1}{12}(0 + 2) = \tfrac{1}{6} \notin \mathbb{Z},$$

a contradiction. So S^4 carries no complex structure. (The same argument, run through the spin^c Dirac operator, shows that S^4 does not even carry an *almost* complex structure.) This is the "far from obvious" integrality of the index formula doing real work.

Worked example 4.67: Complex dimension 3: $\mathbb{CP}^3$ and Calabi–Yau threefolds

$\mathbb{CP}^3$: $c(T) = (1 + h)^4$, so $c_1 = 4h$, $c_2 = 6h^2$, $\int h^3 = 1$, and $\chi(\mathcal{O}) = \frac{1}{24} \cdot 4 \cdot 6 = 1$, matching $h^{0,0} = 1$, $h^{0,1} = h^{0,2} = h^{0,3} = 0$.
A Calabi–Yau threefold (the internal space of many string compactifications) has $c_1 = 0$, so $\chi(\mathcal{O}) = \frac{1}{24}\int c_1 c_2 = 0$. Directly: $h^{0,0} = 1$, $h^{0,1} = h^{0,2} = 0$, $h^{0,3} = 1$ (the holomorphic 3-form), and $1 - 0 + 0 - 1 = 0$. ✓

4.13.5 Riemann–Roch for curves, checked by hand

The original Riemann–Roch theorem (Riemann 1857, Roch 1865) is about line bundles on a Riemann surface. It is the case $m = 1$ of (4.109), and it is worth checking completely.

Let $L \to \Sigma_g$ be a holomorphic line bundle, with $\deg L := \int_{\Sigma_g} c_1(L)$ (an integer, the "number of zeros minus poles" of any meromorphic section). Its Chern character is $1 + c_1(L)$ and

$\mathrm{td}(\Sigma_g) = 1 + \frac{1}{2}c_1(T)$. Then

$$\begin{aligned} h^0(L) - h^1(L) &\overset{(1)}{=} \int_{\Sigma_g} \big(1 + c_1(L)\big)\big(1 + \tfrac{1}{2}c_1(T)\big) \\ &\overset{(2)}{=} \int_{\Sigma_g} \Big(c_1(L) + \tfrac{1}{2}c_1(T)\Big) \\ &\overset{(3)}{=} \deg L + 1 - g. \end{aligned} \tag{4.111}$$

(1) Riemann–Roch (4.109) for $V = L$; here $h^q(L) = \dim H^{0,q}(\Sigma_g; L)$.
(2) Keep only the 2-form part: the product $c_1(L)c_1(T)$ is a 4-form and vanishes on a surface.
(3) $\int c_1(T) = \chi(\Sigma_g) = 2 - 2g$.

The second number has a classical meaning: by **Serre duality**, $h^1(L) = h^0(K \otimes L^{-1})$, where $K = \Lambda^{1,0} = T^*$ is the canonical bundle (holomorphic 1-forms), of degree $2g - 2$. (The pairing $H^{0,1}(L) \times H^0(K \otimes L^{-1}) \to \mathbb{C}$, $(\alpha, \beta) \mapsto \int \alpha \wedge \beta$, is perfect.)

Worked example 4.68: Line bundles on $\mathbb{CP}^1$: homogeneous polynomials

On $\mathbb{CP}^1$ ($g = 0$) the line bundle $\mathcal{O}(d)$ has holomorphic sections the homogeneous polynomials of degree d in (z_0, z_1). In the chart $z = z_1/z_0$ a section is a polynomial $s(z) = a_0 + a_1 z + \cdots + a_d z^d$; in the other chart $w = 1/z$ it is $w^d s(1/w) = a_0 w^d + \cdots + a_d$, which is holomorphic at $w = 0$ precisely because $\deg s \le d$. So

$$h^0(\mathcal{O}(d)) = d + 1 \quad (d \ge 0), \qquad h^0(\mathcal{O}(d)) = 0 \quad (d < 0).$$

For the h^1: $K = \mathcal{O}(-2)$ on $\mathbb{CP}^1$ (Example 4.49: $T = \mathcal{O}(2)$), so $h^1(\mathcal{O}(d)) = h^0(\mathcal{O}(-2-d))$, which is $-d - 1$ for $d \le -2$ and 0 otherwise. Now the check:

d	-4	-3	-2	-1	0	1	2
h^0	0	0	0	0	1	2	3
h^1	3	2	1	0	0	0	0
$h^0 - h^1$	-3	-2	-1	0	1	2	3
$d + 1 - g$	-3	-2	-1	0	1	2	3

The individual numbers have a kink at $d = -1$; the difference is the straight line $d + 1$ (Figure 4.38).

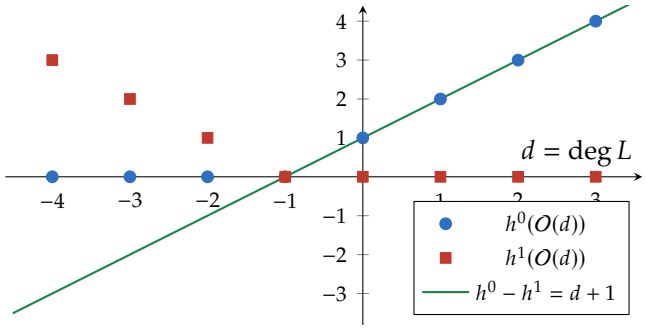


Figure 4.38: Riemann–Roch on $\mathbb{CP}^1$. The number of holomorphic sections h^0 (blue) and the obstruction count h^1 (red) each have a corner, but their difference is exactly the linear function $d + 1 - g$ (green) predicted by the index theorem. Only the difference is a topological invariant.

Worked example 4.69: The canonical bundle and the elliptic curve

$L = K$ *on* Σ_g. $h^0(K) = g$ (holomorphic 1-forms) and, by Serre duality, $h^1(K) = h^0(\mathcal{O}) = 1$. Riemann–Roch predicts $h^0 - h^1 = \deg K + 1 - g = (2g-2) + 1 - g = g - 1$, and indeed $g - 1 = g - 1$. Turned around, this *proves* $\deg K = 2g - 2$, i.e. Gauss–Bonnet again.
Elliptic curve ($g = 1$, K trivial). For $\deg L = d > 0$ we have $h^1(L) = h^0(L^{-1}) = 0$ (a bundle of negative degree has no sections: a section would have $d < 0$ zeros). So $h^0(L) = d$. For $d = 1$ the unique section (up to scale) is a Jacobi theta function; the d sections for $\deg L = d$ are the theta functions of level d.

Physical picture: Landau levels on a sphere and on a torus

A charged particle on a surface Σ in a uniform magnetic field with N_ϕ flux quanta, $\frac{1}{2\pi}\int B = N_\phi$, is a section of a line bundle L of degree N_ϕ. The lowest-Landau-level wavefunctions are the solutions of $\bar\partial_A \psi = 0$, i.e. holomorphic sections of L (up to a spin/curvature shift, which on the sphere replaces L by $L \otimes K^{1/2}$ or not, depending on the Hamiltonian). For strong enough field $h^1 = 0$ and Riemann–Roch gives the degeneracy

$$\dim(\mathrm{LLL}) = N_\phi + 1 - g.$$

On the torus this is N_ϕ, the familiar "one state per flux quantum". On Haldane's sphere, with our $\mathcal{O}(N_\phi)$, it is $N_\phi + 1$; the shift by $1 - g$ is a topological property of the surface, which is the origin of the "shift" in quantum Hall physics.

4.13.6 Hodge numbers refine Betti numbers

Forgetting the complex structure, X has a de Rham complex and an Euler characteristic. Since $\mathrm{d} = \partial + \bar\partial$, one expects the Betti numbers to decompose into Hodge numbers. Nash states (4.73)

$$\chi(X) = \sum_k (-1)^k b_k = \sum_{p,q} (-1)^{p+q} h^{p,q}. \tag{Nash 4.73}$$

Why is this true? Both sides are indices. The left side is the index of the de Rham complex. The right side is $\sum_p (-1)^p$ times the index of the Dolbeault complex in row p of Figure 4.37, i.e. the index of $\bar\partial + \bar\partial^*$ on all forms, graded by total degree. The symbols of $\mathrm{d} + \mathrm{d}^*$ and of $\sqrt{2}(\bar\partial + \bar\partial^*)$ can be deformed into each other through elliptic symbols (on a Kähler manifold the operators themselves agree: $\Delta_d = 2\Box_{\bar\partial}$), and the index cannot change under such a deformation. On a Kähler manifold we have more: each harmonic k-form splits into harmonic (p,q)-pieces, so

$$b_k = \sum_{p+q=k} h^{p,q} \qquad \text{(Kähler)},$$

and (Nash 4.73) follows by summing. The Hodge numbers also obey $h^{p,q} = h^{q,p}$ (conjugation) and $h^{p,q} = h^{m-p,m-q}$ (Serre duality), which gives the **Hodge diamond** its symmetry.

Worked example 4.70: Checking (Nash 4.73) on K3

From the diamond in Figure 4.39:

$$\sum_{p,q}(-1)^{p+q}h^{p,q} \overset{(1)}{=} h^{0,0} - (h^{1,0}+h^{0,1}) + (h^{2,0}+h^{1,1}+h^{0,2}) - (h^{2,1}+h^{1,2}) + h^{2,2}$$

$$\overset{(2)}{=} 1 - 0 + (1 + 20 + 1) - 0 + 1 = 24,$$

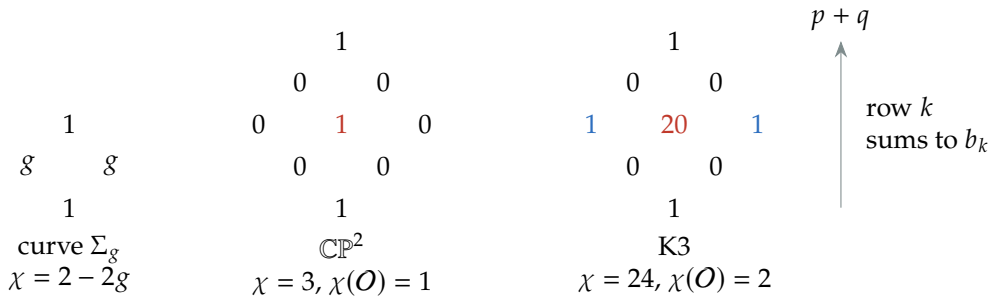


Figure 4.39: Hodge diamonds: $h^{p,q}$ is placed at height $p+q$, with $h^{m,0}$ at the far left and $h^{0,m}$ at the far right of the middle row. Summing row k gives the Betti number b_k (e.g. for K3, $b_2 = 1+20+1 = 22$); the alternating sum over all entries is $\chi(X)$; the alternating sum along the right-hand edge $(h^{0,0}, h^{0,1}, \dots, h^{0,m})$ is the arithmetic genus $\chi(\mathcal{O})$ computed by the Todd class.

(1) Group the Hodge numbers by total degree $p+q$.
(2) Read off the diamond.

which is $\chi(K3) = \int c_2 = 24$, the value we used in Example 4.62.

Watch out: Two meanings of TX

In this example Nash writes TX for the *holomorphic* tangent bundle $T^{1,0}X$, a complex bundle of rank $m = n/2$, and $\overline{TX}$ for its conjugate. Elsewhere TX is the real tangent bundle of rank n. The two are related by $TX_{\mathbb{C}} = T^{1,0}X \oplus \overline{T^{1,0}X}$, which is exactly Nash's (4.69). Also, Nash writes $\mathrm{td}(X)$ for $\mathrm{td}(T^{1,0}X)$; it is *not* $\mathrm{td}(TX_{\mathbb{C}})$, which is $\hat{A}(X)^2$ and involves only Pontryagin classes.

Watch out: The sign in Nash (4.67)

The Chern roots of $\Lambda^{0,1}$ are $+x_j$ (not $-x_j$): $\mathrm{d}\bar{z}$ transforms like $\partial/\partial z$, not like $\mathrm{d}z$. With the wrong sign one obtains $\int \mathrm{td}(T^*) = (-1)^m \int \mathrm{td}(T)$, i.e. $g-1$ instead of $1-g$ on a curve. The check $\chi(\mathcal{O}_{\mathbb{CP}^m}) = 1$ is a quick way to catch this.

Reading Nash (pp. 106–108, (4.63)–(4.73))

Nash introduces the Dolbeault complex and its cohomology (4.63), notes that there is one complex for each p and sets $p = 0$, and identifies the index with the arithmetic genus $\chi(\mathcal{E}) = \sum (-1)^q h_{0,q}$ (4.64)–(4.65). The symbol complex is (4.66). Using $\overline{T^*X} \cong TX$ (holomorphic tangent bundle) he reuses the de Rham manipulation to get (4.67)–(4.68), then (4.69) $c_{n/2}(F^*) = (-1)^{n/2} e(X)$ and $\mathrm{td}(TX_{\mathbb{C}}) = \mathrm{td}(F)\,\mathrm{td}(F^*)$, and arrives at index $\bar{\partial} = \mathrm{td}(X)[X]$ (4.70)–(4.71), the Riemann–Roch theorem (4.72). Finally (4.73) relates the Euler characteristic to the Hodge numbers. We added: explicit (p,q)-calculus, the symbol and its ellipticity, low dimensional Todd integrands, and checks on curves, $\mathbb{CP}^1$ line bundles, $\mathbb{CP}^2$, $\mathbb{CP}^1 \times \mathbb{CP}^1$, K3, $\mathbb{CP}^3$ and Calabi–Yau threefolds.

Recap

- On a complex manifold $\mathrm{d} = \partial + \bar{\partial}$; each row of the (p,q)-grid is an elliptic complex with symbol $ip^{0,1}\wedge$.
- index $\bar{\partial} = \sum_q (-1)^q h^{0,q} = \chi(\mathcal{O}_X)$, the arithmetic genus.

- $\Lambda^{0,q} \cong \Lambda^q T^{1,0}$ gives ch $= \prod(1 - e^{x_j}) = c_m(T^*)/\mathrm{td}(T^*)$, and the index formula collapses to $\int_X \mathrm{td}(X)$: Riemann–Roch–Hirzebruch.
- Low dimensions: $\frac{1}{2}c_1$; $\frac{1}{12}(c_1^2 + c_2)$; $\frac{1}{24}c_1c_2$. On curves with a line bundle: $h^0 - h^1 = \deg L + 1 - g$.
- Integrality has teeth: S^4 is not a complex surface.
- $\chi(X) = \sum(-1)^{p+q} h^{p,q}$: Hodge numbers refine Betti numbers.

4.14 The signature complex and the Hirzebruch signature theorem

The Euler characteristic uses all the Betti numbers but forgets how the cohomology classes multiply. The *signature* is an invariant that uses the multiplication. On a $4k$-manifold two middle-dimensional classes can be multiplied to a top-degree class and integrated, giving a number: geometrically, the number of points (with signs) in which two $2k$-dimensional submanifolds intersect. This **intersection form** is a symmetric matrix of integers, and its signature (number of positive minus number of negative eigenvalues) is a topological invariant.

Physicists meet it in several places: the classification of four-manifolds (Donaldson and Seiberg–Witten theory, Chapter 12), the gravitational contribution to the axial anomaly, and chiral 2-form fields, whose zero modes are the self-dual harmonic forms. The index theorem will express the signature as the integral of a curvature polynomial, the L-polynomial. In four dimensions,

$$\mathrm{Sign}(X) = -\frac{1}{24\pi^2}\int_X \mathrm{tr}\big(R \wedge R\big).$$

An integer (a count of eigenvalue signs) equals an integral of curvature.

4.14.1 The intersection form

Let X be compact and oriented, $\dim X = n = 2l$. Cohomology classes of degree l can be multiplied (cup product; for de Rham classes, the wedge product of representatives) to degree $2l = n$ and then integrated:

$$q(u, v) = (u \cup v)[X] = \int_X u \wedge v, \qquad u, v \in H^l(X). \tag{Nash 4.74}$$

With integer coefficients, q is an integer-valued bilinear form on the lattice $H^l(X;\mathbb{Z})$ (modulo torsion). By Poincaré duality it is *unimodular*: in an integral basis its matrix has determinant ± 1, in particular it is non-degenerate.

Working definition 4.34 (Signature)

Choose a basis of $H^l(X;\mathbb{R})$ and write q as a real matrix Q. Diagonalise it. Let b^+ and b^- be the numbers of positive and negative eigenvalues ($b^+ + b^- = b_l$ [the Betti number $\dim H^l(X;\mathbb{R})$] since q is non-degenerate). The **signature** is

$$\mathrm{Sign}(X) = b^+ - b^-. \tag{Nash 4.75}$$

By Sylvester's law of inertia [under a change of basis $Q \mapsto P^T Q P$ the numbers of positive and of negative eigenvalues stay the same], $b^\pm$ do not depend on the basis.

Why only $n = 4k$. The symmetry of q is that of the wedge product:

$$q(u,v) \overset{(1)}{=} \int_X u \wedge v \overset{(2)}{=} (-1)^{l\cdot l} \int_X v \wedge u \overset{(3)}{=} (-1)^l\, q(v,u). \tag{Nash 4.76}$$

(1) Definition (Nash 4.74).
(2) Commuting an l-form past an l-form costs $(-1)^{l\cdot l}$.
(3) $l^2 \equiv l \pmod 2$.

If l is odd, q is antisymmetric. A real antisymmetric matrix has purely imaginary eigenvalues $\pm i\lambda$ (it is i times a Hermitian matrix), so there are no positive or negative ones, and the signature is zero by fiat. The interesting case is l even, i.e. $n = 4k$. Reversing the orientation of X flips the sign of $\int_X$, hence $q \mapsto -q$ and $\mathrm{Sign}(\bar{X}) = -\mathrm{Sign}(X)$: the signature is an invariant of *oriented* manifolds.

Physical picture: Counting intersection points

If u and v are Poincaré dual to closed oriented submanifolds A and B of dimension l (i.e. $\int_A u' = \int_X u' \wedge u$ for all closed u'), then $q(u,v)$ is the number of intersection points of A and B, each counted $+1$ or -1 according to whether the orientations of A and B at the point add up to that of X. The self-intersection $q(u,u)$ is computed by pushing A off itself to a nearby copy A' and counting $A \cap A'$.

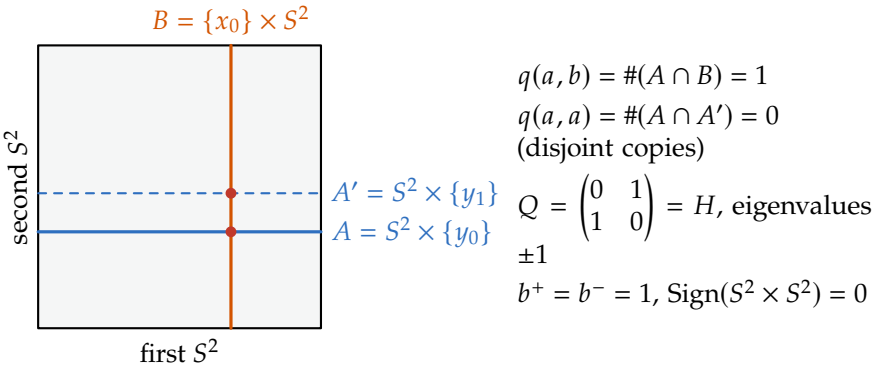


Figure 4.40: The intersection form of $S^2 \times S^2$, drawn as a square (each side stands for one sphere factor). A horizontal sphere A and a vertical sphere B meet in one point (red), so $q(a,b) = 1$; A can be pushed off itself to a disjoint parallel copy A', so $q(a,a) = 0$. The matrix is the hyperbolic form H, with one positive and one negative eigenvalue.

Worked example 4.71: Intersection forms of the basic four-manifolds

$\mathbb{CP}^2$. $H^2 = \mathbb{Z}h$ with $\int h^2 = 1$ (two lines meet in a point). $Q = (1)$, Sign $= 1$. For $\overline{\mathbb{CP}^2}$ (reversed orientation) $Q = (-1)$, Sign $= -1$.
$S^2 \times S^2$. $Q = H$ as in Figure 4.40, Sign $= 0$.
$T^4 = \mathbb{R}^4/\mathbb{Z}^4$. H^2 is spanned by the six forms $\mathrm{d}x^{ij} := \mathrm{d}x^i \wedge \mathrm{d}x^j$, $i < j$. The only nonzero pairings are between complementary pairs:

$$\int \mathrm{d}x^{12} \wedge \mathrm{d}x^{34} = +1,$$
$$\int \mathrm{d}x^{13} \wedge \mathrm{d}x^{24} \overset{(1)}{=} -\int \mathrm{d}x^1 \wedge \mathrm{d}x^2 \wedge \mathrm{d}x^3 \wedge \mathrm{d}x^4 = -1,$$
$$\int \mathrm{d}x^{14} \wedge \mathrm{d}x^{23} \overset{(2)}{=} +\int \mathrm{d}x^1 \wedge \mathrm{d}x^2 \wedge \mathrm{d}x^3 \wedge \mathrm{d}x^4 = +1.$$

(1) Swap dx^3 and dx^2: one transposition.
(2) Move dx^4 to the end, past dx^2 and dx^3: two transpositions.

So $Q = H \oplus (-H) \oplus H$. Since $-H$ becomes H after changing the sign of one basis vector, $Q \cong 3H$; either way each block has eigenvalues ± 1, $b^+ = b^- = 3$ and $\mathrm{Sign}(T^4) = 0$.
K3. $b_2 = 22$ and (by a theorem of Freedman type, or by explicit construction as a Kummer surface) $Q = 3H \oplus 2(-E_8)$, where E_8 is the Cartan matrix of the Lie algebra E_8:

$$E_8 = \begin{pmatrix} 2 & -1 & & & & & & \\ -1 & 2 & -1 & & & & & \\ & -1 & 2 & -1 & & & & \\ & & -1 & 2 & -1 & & & \\ & & & -1 & 2 & -1 & & -1 \\ & & & & -1 & 2 & -1 & \\ & & & & & -1 & 2 & \\ & & & & -1 & & & 2 \end{pmatrix}.$$

It is positive definite, has $\det E_8 = 1$ (unimodular) and even diagonal. So $b^+ = 3$, $b^- = 3 + 16 = 19$, and $\mathrm{Sign}(K3) = 3 - 19 = -16$.

4.14.2 The involution τ

To realise $\mathrm{Sign}(X)$ as an index we need a way to tell "positive" from "negative" middle-dimensional forms using an operator. In four dimensions the answer is familiar from Yang–Mills theory: for a real 2-form ω,

$$q(\omega,\omega) = \int_X \omega \wedge \omega = \begin{cases} +\int_X \omega \wedge \star\omega = +\|\omega\|^2 & \text{if } \star\,\omega = \omega \text{ (self-dual)}, \\ -\int_X \omega \wedge \star\omega = -\|\omega\|^2 & \text{if } \star\,\omega = -\omega \text{ (anti-self-dual)}. \end{cases}$$

So the Hodge star on 2-forms sorts classes into positive and negative ones. Nash extends this to an involution on *all* forms.

The Hodge star on complex forms. Complex forms carry the Hermitian inner product

$$\langle \omega, \nu \rangle = \int_X \omega \wedge \star\bar{\nu}, \qquad \omega, \nu \in \Omega^p(X),$$

with $\star$ extended complex-linearly. On an n-manifold with Riemannian metric $\star\star = (-1)^{p(n-p)}$ on p-forms. For n even, $p(n-p) \equiv p^2 \equiv p \pmod 2$, so

$$\star^2 = (-1)^p \qquad \text{on } \Omega^p,\ n \text{ even}.$$

Working definition 4.35 (The involution τ (Nash (4.77)))

On an even-dimensional oriented Riemannian manifold define

$$\tau : \Omega^p(X) \to \Omega^{n-p}(X), \qquad \tau\omega = i^{\lambda(p)} \star \omega, \qquad \lambda(p) = p(p-1) + \tfrac{n}{2}. \qquad \text{(Nash 4.77)}$$

[the forms here are complex-valued, since τ contains powers of i] It is the Hodge star dressed with the phase needed to make it square to one.

Check that $\tau^2 = 1$ on $\omega \in \Omega^p$:

$$\tau^2\omega \overset{(1)}{=} i^{\lambda(p)}\,\tau(\star\omega) \overset{(2)}{=} i^{\lambda(p)+\lambda(n-p)}\,\star^2\,\omega$$
$$\overset{(3)}{=} i^{2p}\ \star^2\,\omega \overset{(4)}{=} (-1)^p(-1)^p\omega = \omega. \tag{Nash 4.78}$$

(1) Definition of τ on Ω^p; τ is complex linear.
(2) $\star\omega$ has degree $n-p$, so the second τ brings $i^{\lambda(n-p)}$.
(3) $\lambda(p)+\lambda(n-p) = p^2 - p + (n-p)^2 - (n-p) + n = 2p^2 - 2np + n^2$. With n even, $n^2 \equiv 0$ and $2np \equiv 0 \pmod 4$, and $2p^2 \equiv 2p \pmod 4$ (since $2p(p-1)$ is divisible by 4). So $i^{\lambda(p)+\lambda(n-p)} = i^{2p}$.
(4) $i^{2p} = (-1)^p$ and $\star^2 = (-1)^p$.

Since $\tau^2 = 1$, its eigenvalues are ± 1, and every form splits uniquely as $\omega = \frac{1}{2}(\omega+\tau\omega)+\frac{1}{2}(\omega-\tau\omega)$:

$$\Omega^*(X) = \bigoplus_p \Omega^p(X) = \Omega^+ \oplus \Omega^-, \qquad \Omega^\pm = \{\omega : \tau\omega = \pm\omega\}. \tag{Nash 4.79}$$

The corresponding bundles are $\Lambda^\pm T^*X_{\mathbb{C}}$, each of rank $\frac{1}{2}\cdot 2^n = 2^{n-1}$ (the pairing $\omega \mapsto \omega \pm \tau\omega$ is a bijection between Λ^p and each $\Lambda^\pm \cap (\Lambda^p \oplus \Lambda^{n-p})$ for $p < n/2$, and the middle degree splits in half).

Worked example 4.72: τ in two and four dimensions, explicitly

$n = 2$ (so $n/2 = 1$), orthonormal coframe e^1, e^2, $\mathrm{vol} = e^1 \wedge e^2$, $\star 1 = \mathrm{vol}$, $\star e^1 = e^2$, $\star e^2 = -e^1$, $\star\mathrm{vol} = 1$. The phases are $i^{\lambda(0)} = i$, $i^{\lambda(1)} = i$, $i^{\lambda(2)} = i^3 = -i$. So

$$\tau 1 = i\,\mathrm{vol}, \quad \tau\,\mathrm{vol} = -i, \quad \tau e^1 = ie^2, \quad \tau e^2 = -ie^1.$$

The eigenvectors: $\tau(1 + i\,\mathrm{vol}) = i\,\mathrm{vol} + 1$ and $\tau(e^1 + ie^2) = ie^2 + e^1$ have eigenvalue $+1$; $1 - i\,\mathrm{vol}$ and $e^1 - ie^2$ have eigenvalue -1. With $z = x^1 + ix^2$ along the plane,

$$\Lambda^+ = \mathrm{span}\{1 + i\,\mathrm{vol},\ \mathrm{d}z\}, \qquad \Lambda^- = \mathrm{span}\{1 - i\,\mathrm{vol},\ \mathrm{d}\bar{z}\}.$$

$n = 4$ ($n/2 = 2$). The phases are $i^{\lambda(p)} = i^{p(p-1)+2}$: -1 for $p = 0, 1$; $+1$ for $p = 2, 3$; -1 for $p = 4$. So $\tau = -\star$ on 0-, 1- and 4-forms and $\tau = +\star$ on 2- and 3-forms. In particular

$$\Lambda^+ = \underbrace{\mathrm{span}\{1 - \mathrm{vol}\}}_{1} \oplus \underbrace{\{\alpha - \star\alpha : \alpha \in \Lambda^1\}}_{4} \oplus \underbrace{\Lambda^2_+}_{3},$$

where Λ^2_+ are the self-dual 2-forms, e.g. $e^{12} + e^{34}$, $e^{13} - e^{24}$, $e^{14} + e^{23}$. (Check one piece: $\tau(\alpha - \star\alpha) = -\star\alpha + \star\star\alpha = -\star\alpha + \alpha$, using $\star^2 = -1$ on 1-forms.) The rank is $1+4+3 = 8 = 2^{4-1}$.

4.14.3 The signature operator

The key property is that τ *anticommutes* with $\mathrm{d} + \mathrm{d}^*$. First recall that on an even-dimensional manifold $\mathrm{d}^* = -\star\,\mathrm{d}\star$. Then for $\omega \in \Omega^p$:

$$\tau\,\mathrm{d}\,\tau\,\omega \overset{(1)}{=} i^{\lambda(p)}\,\tau\big(\mathrm{d}\star\omega\big)$$
$$\overset{(2)}{=} i^{\lambda(p)+\lambda(n-p+1)}\ \star\,\mathrm{d}\star\,\omega$$
$$\overset{(3)}{=} -\,i^{\lambda(p)+\lambda(n-p+1)}\,\mathrm{d}^*\omega \overset{(4)}{=} -\mathrm{d}^*\omega. \tag{4.112}$$

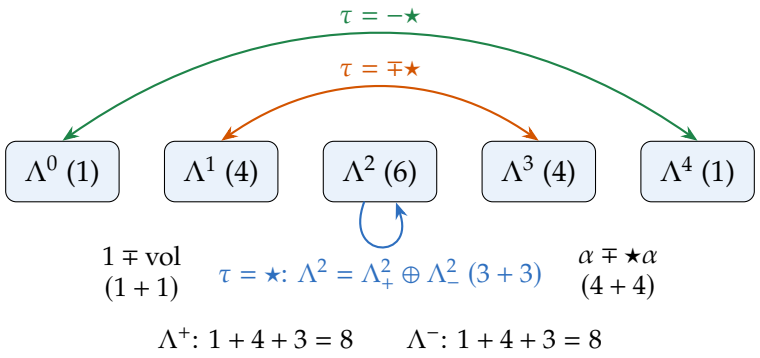


Figure 4.41: How τ acts on forms in four dimensions (ranks in brackets). It exchanges Λ^p with Λ^{4-p}, so each pair $\Lambda^p \oplus \Lambda^{4-p}$ splits evenly into ± 1 eigenspaces; on the middle degree it is the Hodge star itself, and its eigenspaces are the self-dual and anti-self-dual 2-forms. Only the middle degree can contribute an imbalance between + and −.

(1) $\tau\omega = i^{\lambda(p)} \star \omega$ and d is linear.
(2) $\mathrm{d} \star \omega$ has degree $n - p + 1$.
(3) $\mathrm{d}^* = - \star \mathrm{d} \star$.
(4) $\lambda(p) + \lambda(n-p+1) = p^2 - p + (n-p+1)(n-p) + n = n^2 - 2np + 2p^2 + 2n - 2p$. For n even, n^2, $2np$ and $2n$ are divisible by 4, and so is $2p^2 - 2p = 2p(p-1)$. So the phase is $i^0 = 1$.

Multiplying (4.112) on the right by τ (using $\tau^2 = 1$) gives $\tau\mathrm{d} = -\mathrm{d}^*\tau$, and taking adjoints ($\tau$ is unitary and squares to one, hence self-adjoint) gives $\tau\mathrm{d}^* = -\mathrm{d}\tau$. Adding:

$$\tau(\mathrm{d} + \mathrm{d}^*) = -(\mathrm{d} + \mathrm{d}^*)\tau. \tag{4.113}$$

So if $\tau\omega = \omega$ then $\tau(\mathrm{d} + \mathrm{d}^*)\omega = -(\mathrm{d} + \mathrm{d}^*)\omega$: $\mathrm{d} + \mathrm{d}^*$ maps Ω^+ into Ω^- and vice versa.

Working definition 4.36 (Signature operator and signature complex)

The **signature operator** is the restriction

$$\mathrm{d}_+ = (\mathrm{d} + \mathrm{d}^*)\big|_{\Omega^+} : \Omega^+(X) \longrightarrow \Omega^-(X), \tag{Nash 4.80}$$

[d^* is the adjoint of d, lowering the form degree by one] and the **signature complex** is the two-term complex

$$0 \longrightarrow \Gamma\big(X, \Lambda^+ T^*X \otimes_{\mathbb{R}} \mathbb{C}\big) \xrightarrow{\mathrm{d}_+} \Gamma\big(X, \Lambda^- T^*X \otimes_{\mathbb{R}} \mathbb{C}\big) \longrightarrow 0, \tag{Nash 4.81}$$

with symbol complex [the leading symbol, here $\sigma_1(\mathrm{d} + \mathrm{d}^*)(x,k) = i(k \wedge - \iota_k)$, acting on the same fibre at every $(x,k) \in T^*X$]

$$0 \longrightarrow \pi^*\Lambda^+ T^*X \otimes_{\mathbb{R}} \mathbb{C} \xrightarrow{\sigma_1(\mathrm{d}_+)} \pi^*\Lambda^- T^*X \otimes_{\mathbb{R}} \mathbb{C} \longrightarrow 0. \tag{Nash 4.82}$$

Its adjoint is $\mathrm{d}_- = (\mathrm{d} + \mathrm{d}^*)|_{\Omega^-}$.

Ellipticity is inherited from $\mathrm{d} + \mathrm{d}^*$: its symbol $\sigma_1 = i(p \wedge - \iota_p)$ squares to $|p|^2$ (Section 4.12), so its restriction $\Lambda^+ \to \Lambda^-$ is invertible for $p \neq 0$. The only difference from the de Rham operator is *how the forms are split* into two halves: by τ instead of by the parity of the degree.

4.14.4 The index is the signature

Since d_+ is elliptic, Hodge theory identifies kernel and cokernel with harmonic forms. The Laplacian $\Delta = (\mathrm{d}+\mathrm{d}^*)^2$ commutes with τ (two anticommutations), so τ preserves the harmonic forms $\mathcal{H}^* \cong H^*(X;\mathbb{C})$, and

$$H^{\pm}(X;\mathbb{C}) = \{\omega \in \mathcal{H}^* : \tau\omega = \pm\omega\}, \qquad h_\pm = \dim H^\pm, \qquad H^*(X;\mathbb{C}) = H^+ \oplus H^-. \quad \text{(Nash 4.83)}$$

The kernel of d_+ is H^+ and the kernel of its adjoint d_- is H^-, so

$$\operatorname{index} \mathrm{d}_+ = h_+ - h_-. \quad \text{(Nash 4.84)}$$

Now we sort the harmonic forms by degree, as in Figure 4.41. For $0 \le p < n/2$ put

$$I^p = H^p(X;\mathbb{C}) \oplus H^{n-p}(X;\mathbb{C}). \quad \text{(Nash 4.85)}$$

The involution τ maps I^p to itself, exchanging the two summands (it is an isomorphism $H^p \to H^{n-p}$: this is Poincaré duality realised by the Hodge star). For any $\alpha \in H^p$, the forms $\alpha \pm \tau\alpha$ are ±1 eigenvectors, and every eigenvector in I^p has this form. Hence

$$I^p = I^p_+ \oplus I^p_-, \qquad \dim I^p_+ = \dim I^p_- = b_p. \quad \text{(Nash 4.86)}$$

What is left is the middle degree $p = n/2 = l$, where τ maps H^l to itself:

$$H^l(X;\mathbb{C}) = H^l_+(X;\mathbb{C}) \oplus H^l_-(X;\mathbb{C}), \qquad h^l_\pm = \dim H^l_\pm. \quad \text{(Nash 4.87)}$$

On middle forms the phase is $i^{\lambda(l)} = i^{l(l-1)+l} = i^{l^2} = 1$ because l is even. So $\tau = \star$ on H^l, a *real* operator, and its eigenspaces are spanned by real forms. For a real harmonic ω with $\star\omega = \pm\omega$,

$$q(\omega,\omega) \overset{(1)}{=} \int_X \omega \wedge \omega \overset{(2)}{=} \pm \int_X \omega \wedge \star\omega \overset{(3)}{=} \pm \|\omega\|^2,$$

(1) The intersection form on de Rham representatives.
(2) $\omega = \pm \star \omega$.
(3) For real forms, $\int \omega \wedge \star\omega$ is the L^2 norm squared.

and for $\omega_+ \in H^l_+$, $\omega_- \in H^l_-$, $q(\omega_+,\omega_-) = -\int \omega_+ \wedge \star\omega_- = -\langle \omega_+, \omega_-\rangle = 0$ (eigenvectors of the self-adjoint $\star$ with different eigenvalues are orthogonal). So q is positive definite on H^l_+, negative definite on H^l_-, and these are q-orthogonal: by Sylvester, $h^l_\pm = b^\pm$.

Now collect everything:

$$\begin{aligned} H^*(X;\mathbb{C}) &\overset{(1)}{=} I^0 \oplus \cdots \oplus I^{l-1} \oplus H^l(X;\mathbb{C}) \\ &\overset{(2)}{=} \left(I^0_+ \oplus \cdots \oplus I^{l-1}_+ \oplus H^l_+\right) \oplus \left(I^0_- \oplus \cdots \oplus I^{l-1}_- \oplus H^l_-\right) = H^+ \oplus H^-, \\ \implies h_\pm &\overset{(3)}{=} h^l_\pm + \sum_{p=0}^{l-1} \dim I^p_\pm, \end{aligned} \quad \text{(Nash 4.88)}$$

(1) Group the degrees in pairs $(p, n-p)$ plus the middle degree.
(2) Split each piece into τ-eigenspaces, by (Nash 4.86) and (Nash 4.87).
(3) Take dimensions.

and therefore

$$\operatorname{index} \mathrm{d}_+ = h_+ - h_- \overset{(1)}{=} h^l_+ - h^l_- \overset{(2)}{=} b^+ - b^- = \operatorname{Sign}(X). \quad \text{(Nash 4.89)}$$

(1) (Nash 4.88): the $I^p_\pm$ have equal dimensions and cancel.
(2) The identification $h^l_\pm = b^\pm$ just made.

Worked example 4.73: Two harmonic checks

Surfaces ($n = 2$, $l = 1$ odd). By Example 4.72, $\Lambda^+ = \{1 + i\,\mathrm{vol}\} \oplus \Lambda^{1,0}$. Harmonic forms in Ω^+: the constant $1 + i$ vol and the g holomorphic 1-forms; in Ω^-: $1 - i$ vol and the g antiholomorphic 1-forms. So $\operatorname{index} \mathrm{d}_+ = (1 + g) - (1 + g) = 0$, as it must be for l odd.
$S^2 \times S^2$ with the product of unit round metrics. The harmonic 2-forms are spanned by the two area forms ω_1, ω_2. The Hodge star of the product metric swaps them: $\star\omega_1 = \omega_2$ (the orthogonal complement of the first factor's tangent plane is the second's). So $\omega_1 + \omega_2$ is self-dual and $\omega_1 - \omega_2$ anti-self-dual: $b^+ = b^- = 1$, Sign $= 0$, matching $Q = H$. Note $q(\omega_1 + \omega_2, \omega_1 + \omega_2) = 2\int \omega_1 \wedge \omega_2 > 0$. ✓

4.14.5 The characteristic-class side: the L-polynomial

We now need

$$\mathrm{ch}\big(\sigma_1(\mathrm{d}_+)\big) \longrightarrow \mathrm{ch}\big(\Lambda^+ T^*X \otimes \mathbb{C}\big) - \mathrm{ch}\big(\Lambda^- T^*X \otimes \mathbb{C}\big). \tag{Nash 4.90}$$

Nash quotes the answer from Atiyah–Singer ("the splitting principle and a little group theory"). Here is the argument.

Step 1: τ is a product over 2-planes. Use the orthonormal frame in which the curvature has the block form (4.79), so that $\mathbb{R}^n = P_1 \oplus \cdots \oplus P_m$ is a sum of oriented 2-planes, $n = 2m$. Let $c(v) = v \wedge - \,\iota_v$ act on forms; it satisfies $c(v)^2 = -|v|^2$, the Clifford relation (Chapter T4). One checks (as we do in a moment in two dimensions, and as in Lawson–Michelsohn or Bär) that

$$\tau = i^m\, c(e_1)c(e_2)\cdots c(e_n) = \prod_{j=1}^{m} \underbrace{\big(i\, c(e_{2j-1})c(e_{2j})\big)}_{\tau_j}. \tag{4.114}$$

In two dimensions: $i\,c(e_1)c(e_2)1 = i\,c(e_1)e^2 = i\,e^1 \wedge e^2 = \tau 1$, and $i\,c(e_1)c(e_2)e^1 = i\,c(e_1)(e^2 \wedge e^1) = -i\,c(e_1)\mathrm{vol} = -i(-e^2) = ie^2 = \tau e^1$. Each τ_j involves only the plane P_j, the τ_j commute (they are products of an even number of c's from disjoint planes), and each squares to one. Under $\Lambda^*\mathbb{R}^n = \Lambda^* P_1 \otimes \cdots \otimes \Lambda^* P_m$, τ_j acts on the j-th factor only. The $+1$ eigenspace of the product consists of tensor products with an even number of -1 factors. So, as virtual bundles,

$$\Lambda^+ - \Lambda^- = \bigotimes_{j=1}^{m} \big(\Lambda^+_{P_j} - \Lambda^-_{P_j}\big). \tag{4.115}$$

Step 2: one plane. By Example 4.72, $\Lambda^\pm_{P_j} = \mathbb{C}\cdot(1 \pm i\,\mathrm{vol}_j) \oplus \mathbb{C}\cdot(e^{2j-1} \pm ie^{2j})$. The first summand is rotation invariant: a trivial line. The second is a line on which a rotation of P_j by angle θ acts by $e^{\mp i\theta}$. In the curvature language of Section 4.9, $\mathrm{d}z_j = e^{2j-1} + ie^{2j}$ spans a line with Chern root $-x_j$ (it is dual to $\partial/\partial z_j$, which has root $+x_j$), and $\mathrm{d}\bar z_j$ a line with root $+x_j$. So

$$\mathrm{ch}\,\Lambda^+_{P_j} - \mathrm{ch}\,\Lambda^-_{P_j} = (1 + e^{-x_j}) - (1 + e^{x_j}) = e^{-x_j} - e^{x_j}.$$

Step 3: multiply. Since ch is multiplicative,

$$\mathrm{ch}\big(\Lambda^+ T^*X \otimes \mathbb{C}\big) - \mathrm{ch}\big(\Lambda^- T^*X \otimes \mathbb{C}\big) = \prod_{j=1}^{m} \big(e^{-x_j} - e^{x_j}\big). \tag{Nash 4.91}$$

Its lowest-order term is $\prod(-2x_j) = (-2)^m e(X)$: again divisible by the Euler class, as the recipe requires.

Step 4: the index formula. Insert (Nash 4.91) into (4.101), with $e = \prod x_j$ and $\mathrm{td}(TX_{\mathbb{C}}) = \prod_j \frac{x_j}{1-e^{-x_j}} \frac{-x_j}{1-e^{x_j}}$:

$$\begin{aligned}
\operatorname{index} \mathrm{d}_+ &\overset{(1)}{=} (-1)^m \int_X \prod_{j=1}^m \frac{e^{-x_j} - e^{x_j}}{x_j} \cdot \frac{x_j}{1-e^{-x_j}} \cdot \frac{-x_j}{1-e^{x_j}} \\
&\overset{(2)}{=} (-1)^m \int_X \prod_{j=1}^m \left(-x_j \frac{e^{x_j/2} + e^{-x_j/2}}{e^{x_j/2} - e^{-x_j/2}}\right) \\
&\overset{(3)}{=} \int_X \prod_{j=1}^m \frac{x_j\left(e^{x_j}+1\right)}{e^{x_j}-1} \;=\; \int_X \prod_{j=1}^m \frac{x_j \cosh(x_j/2)}{\sinh(x_j/2)} \\
&\overset{(4)}{=} 2^m \int_X \prod_{j=1}^m \frac{x_j/2}{\tanh(x_j/2)}.
\end{aligned} \tag{Nash 4.92}$$

(1) The index formula for complexes, $n/2 = m$.
(2) Write $a = e^{x/2} - e^{-x/2}$, $b = e^{x/2} + e^{-x/2}$. Then $e^{-x} - e^{x} = -ab$ and $(1-e^{-x})(1-e^{x}) = -a^2$, so each factor is $\frac{-ab}{x} \cdot \frac{-x^2}{-a^2} = -x\, b/a$.
(3) Pull out $(-1)^m$; multiply numerator and denominator by $e^{x/2}$; $b/a = \coth(x/2)$.
(4) $x \coth(x/2) = 2 \cdot \frac{x/2}{\tanh(x/2)}$, one factor 2 per root.

Key idea: The Hirzebruch signature theorem (Nash (4.93)–(4.95))

For a compact oriented manifold of dimension $n = 4k$,

$$\operatorname{Sign}(X) = \operatorname{index} \mathrm{d}_+ = 2^{n/2} \int_X \mathcal{L}(X) = \int_X L(X), \tag{4.116}$$

where

$$\mathcal{L} = \prod_j \frac{x_j/2}{\tanh(x_j/2)}, \qquad L = \prod_j \frac{x_j}{\tanh x_j}.$$

The number $L(X) := 2^{n/2}\mathcal{L}(X)[X]$ is the **L-genus**.

Two normalisations. Nash's $\mathcal{L}$ uses $x/2$; Hirzebruch's L uses x. They agree on the top-degree term after the factor $2^{n/2}$:

$$\begin{aligned}
2^m \Big[\prod_j \frac{x_j/2}{\tanh(x_j/2)}\Big]_{\deg 2m} &\overset{(1)}{=} 2^m \Big[\prod_j \frac{y_j}{\tanh y_j}\Big]_{\deg 2m}\Big|_{y=x/2} \\
&\overset{(2)}{=} 2^m\, 2^{-m} \Big[\prod_j \frac{x_j}{\tanh x_j}\Big]_{\deg 2m} = \big[L(X)\big]_{\deg 2m}.
\end{aligned}$$

(1) Substitute $y_j = x_j/2$.
(2) The top-degree part is a homogeneous polynomial of degree m in the roots (form degree $2m = n$), so rescaling every root by $\frac{1}{2}$ multiplies it by 2^{-m}.

The lower-degree terms differ. They do not matter here, but they do when $\mathcal{L}$ is multiplied by another class, as in the twisted signature operator of the next section; then one must use Nash's $2^{n/2}\mathcal{L}$.

4.14.6 The L-polynomials explicitly

Expand $x/\tanh x = 1 + \frac{x^2}{3} - \frac{x^4}{45} + \frac{2x^6}{945} - \cdots$ (this follows from $\coth x = \frac{1}{x} + \frac{x}{3} - \frac{x^3}{45} + \cdots$). Then, with $p_1 = \sum x_j^2$ and $p_2 = \sum_{i<j} x_i^2 x_j^2$,

$$\begin{aligned} L(X) &\overset{(1)}{=} \prod_j \left(1 + \tfrac{1}{3}x_j^2 - \tfrac{1}{45}x_j^4 + \cdots\right) \\ &\overset{(2)}{=} 1 + \tfrac{1}{3}\sum_j x_j^2 + \left[-\tfrac{1}{45}\sum_j x_j^4 + \tfrac{1}{9}\sum_{i<j} x_i^2 x_j^2\right] + \cdots \\ &\overset{(3)}{=} 1 + \tfrac{1}{3}p_1 + \tfrac{1}{45}(7p_2 - p_1^2) + \cdots . \end{aligned} \tag{4.117}$$

(1) Definition of L with the series of $x/\tanh x$.
(2) Collect form degrees 4 and 8.
(3) $\sum x_j^4 = p_1^2 - 2p_2$, so $-\frac{1}{45}(p_1^2 - 2p_2) + \frac{5}{45}p_2 = \frac{1}{45}(7p_2 - p_1^2)$.

The next term is $L_3 = \frac{1}{945}(62p_3 - 13p_2p_1 + 2p_1^3)$. In dimension 4 the signature theorem reads, using $p_1 = -\frac{1}{8\pi^2}\operatorname{tr}(R \wedge R)$ from Section 4.9,

$$\begin{aligned} \operatorname{Sign}(X) &\overset{(1)}{=} \frac{1}{3}\int_X p_1 \overset{(2)}{=} -\frac{1}{24\pi^2}\int_X \operatorname{tr}\big(R \wedge R\big) \\ &\overset{(3)}{=} \frac{1}{96\pi^2}\int_X \epsilon^{cdef}\, R_{abcd}\, R^{ab}{}_{ef}\ \mathrm{vol}. \end{aligned} \tag{4.118}$$

(1) The degree-4 part of L is $\frac{1}{3}p_1$.
(2) $p_1 = -\frac{1}{8\pi^2}\operatorname{tr}(R \wedge R)$.
(3) Insert $R^a{}_b = \frac{1}{2}R^a{}_{bcd}\, e^c \wedge e^d$, so $\operatorname{tr}(R \wedge R) = R^a{}_b \wedge R^b{}_a = \frac{1}{4}R^a{}_{bcd}R^b{}_{aef}\,\epsilon^{cdef}\mathrm{vol}$; in an orthonormal frame $R^b{}_{aef} = -R_{abef}$, which turns $-\frac{1}{24\pi^2}\cdot\frac{1}{4}$ into $+\frac{1}{96\pi^2}$.

The last form is the "Pontryagin density" $R\tilde{R}$ of gravity, the gravitational analogue of $F\tilde{F}$ in the axial anomaly.

Worked example 4.74: Checking $\operatorname{Sign} = p_1/3$ in four dimensions

$\mathbb{CP}^2$: $\int p_1 = 3$ (Example 4.50), so Sign $= 1$, matching $Q = (1)$.
S^4, $S^2 \times S^2$, T^4: $p_1 = 0$ (for S^4 see Example 4.51; $S^2 \times S^2$ has $p = p(S^2)p(S^2) = 1$; T^4 is flat). All have Sign $= 0$, matching $Q = 0, H, 3H$.
K3: $c_1 = 0$ and $c_2 = 24$, so $p_1 = c_1^2 - 2c_2 = -48$ (the formula of Example 4.50) and Sign $= -48/3 = -16$, matching $Q = 3H \oplus 2(-E_8)$.
A consequence: for every closed oriented 4-manifold, $\int p_1$ is divisible by 3. Nothing in the definition $p_1 = -\frac{1}{8\pi^2}\operatorname{tr} R \wedge R$ suggests this.

Worked example 4.75: Eight dimensions: $\mathbb{CP}^4$ and $\mathbb{H}P^2$

$\mathbb{CP}^4$. With h the hyperplane class, $\int h^4 = 1$, the total Pontryagin class is $p = (1 + h^2)^5$ (Bär, Example 4.3.4): $p_1 = 5h^2$, $p_2 = 10h^4$. Then

$$\operatorname{Sign}(\mathbb{CP}^4) = \tfrac{1}{45}(7 \cdot 10 - 25) = \tfrac{45}{45} = 1,$$

matching $H^4(\mathbb{CP}^4) = \mathbb{Z}h^2$ with $\int h^2 \cdot h^2 = 1$.
Quaternionic projective plane $\mathbb{H}P^2$. Its cohomology is $\mathbb{Z}[u]/(u^3)$ with $\deg u = 4$, $\int u^2 = 1$,

and $p(\mathbb{H}P^2) = 1 + 2u + 7u^2$. So

$$\mathrm{Sign}(\mathbb{H}P^2) = \tfrac{1}{45}(7 \cdot 7 - 2^2) = \tfrac{45}{45} = 1.$$

Two quite different manifolds, and in both cases the 45 in the denominator is exactly cancelled. Hirzebruch's formula is how Milnor (1956) proved that certain 7-dimensional manifolds homeomorphic to S^7 are not diffeomorphic to it: a hypothetical smooth 8-manifold bounding them would have to have a non-integral $\int p_1^2$.

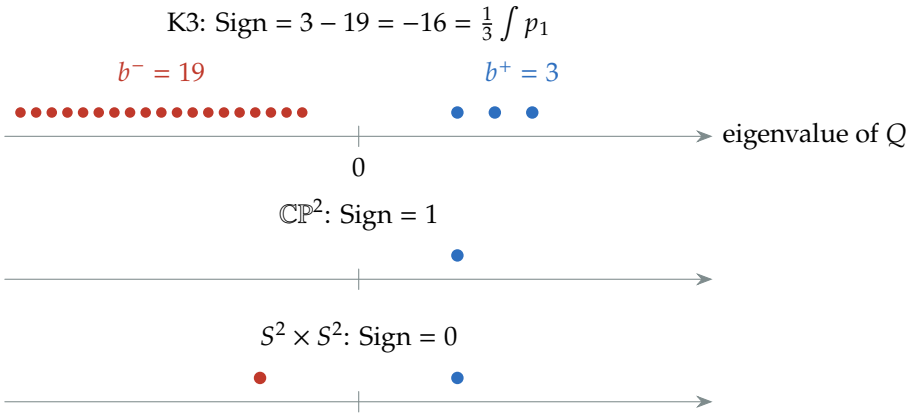


Figure 4.42: The signature as an eigenvalue count. Each dot is an eigenvalue of the intersection form on $H^2(X;\mathbb{R})$ (positions schematic; only the sign is invariant). The index theorem turns "blue dots minus red dots" into the curvature integral $\frac{1}{3}\int p_1 = -\frac{1}{24\pi^2}\int \mathrm{tr}\, R \wedge R$.

Physical picture: Self-dual forms and chiral fields

The self-dual harmonic 2-forms on a four-manifold are the zero modes of a chiral (self-dual) 2-form field; in six dimensions the analogous self-dual 3-forms appear in the M5-brane and in type IIB supergravity's 5-form in ten dimensions. The signature operator is the "Dirac operator" of such fields, and the L-polynomial is its anomaly polynomial. We will see in Section 4.16 that d_+ is literally the Dirac operator twisted by the spinor bundle, which is why $\mathcal{L}$ and $\hat{A}$ look so alike.

Watch out: Manifolds with boundary

The formula $\mathrm{Sign} = \int L$ is for *closed* manifolds. For a manifold with boundary the curvature integral is not an integer in general, and Atiyah–Patodi–Singer showed that the correction is a spectral invariant of the boundary, the η-invariant: $\mathrm{Sign}(X) = \int_X L - \eta(\partial X)$. This is the mathematics behind parity anomalies and the boundary states of topological insulators (Chapter 10).

Reading Nash (pp. 108–112, (4.74)–(4.95))

Nash defines the intersection form q (4.74), its signature (4.75), shows q is antisymmetric for l odd (4.76), and introduces $\tau = i^{p(p-1)+n/2}\star$ (4.77) with $\tau^2 = 1$ (4.78). (In (4.78) he writes the key step as $i^{\lambda(p)+\lambda(n-p)} = i^{2p}$; this is true mod 4, as we showed.) He defines $\Omega^{\pm}$ (4.79), notes $\tau \mathrm{d}_+ = -\mathrm{d}_+ \tau$ (we proved (4.113)), and writes the signature complex and symbol (4.80)–(4.82). The index is $h_+ - h_-$ (4.83)–(4.84); the bookkeeping (4.85)–(4.88) with I^p gives $\mathrm{index}\, \mathrm{d}_+ = \mathrm{Sign}(X)$ (4.89). Then (4.90)–(4.92) evaluate the index formula, (4.93) defines

the L-genus, and he remarks on the two normalisations. The theorem is (4.94)–(4.95). Note that in (4.79) Nash says "eigenvalues ∓ 1" while labelling the spaces $\Omega^{\pm}$; the labelling $\Omega^{\pm} = \{\tau = \pm 1\}$ is the one used throughout.

Recap

- Intersection form $q(u,v) = \int u \wedge v$ on $H^{n/2}$; symmetric for $n = 4k$; Sign $= b^+ - b^-$. Examples: $\mathbb{CP}^2$: (1); $S^2 \times S^2$: H; T^4: $3H$; K3: $3H \oplus 2(-E_8)$.
- $\tau = i^{p(p-1)+n/2}\star$ squares to 1 and anticommutes with $\mathrm{d} + \mathrm{d}^*$; $\mathrm{d}_+ : \Omega^+ \to \Omega^-$ is the signature operator.
- Paired degrees I^p cancel; the middle degree gives $\operatorname{index} \mathrm{d}_+ = b^+ - b^-$, with $\tau = \star$ there (self-dual vs anti-self-dual).
- $\operatorname{ch}\Lambda^+ - \operatorname{ch}\Lambda^- = \prod(e^{-x_j} - e^{x_j})$ (one 2-plane at a time), leading to Sign $= \int \prod \frac{x_j}{\tanh x_j}$.
- $L_1 = \frac{1}{3}p_1$, $L_2 = \frac{1}{45}(7p_2 - p_1^2)$; in 4D Sign $= -\frac{1}{24\pi^2}\int \operatorname{tr} R \wedge R$.

4.15 The spin complex, the Dirac operator and the $\hat{A}$-genus

The three complexes we have met so far act on differential forms. Forms are built out of the tangent bundle, so they exist on every manifold, and the operators d, d^*, $\bar{\partial}$ come for free. Nash's last example is different, and for a physicist it is the most important one. It is the operator that governs the electron: the **Dirac operator**. The question we want to answer is

How many zero modes does a massless fermion have on a compact curved space, and how many of each chirality?

The index theorem answers with a curvature integral, the $\hat{A}$-genus. On the way we have to face a problem that forms never posed: spinors are not built from the tangent bundle. A spinor changes sign under a rotation by 2π, and a manifold must carry extra structure, a *spin structure*, before spinor fields make sense on it. So the section has three parts: what a spin structure is and when it exists; the Dirac operator and its chiral halves; and the characteristic-class computation, which ends in some of the most striking consequences of the whole theory, such as the fact that $\mathbb{CP}^2$ admits no spinors and Rokhlin's theorem that the signature of a smooth spin 4-manifold is divisible by 16.

Throughout this section X is compact, oriented, Riemannian, of dimension n, and from Section 4.15.3 on $n = 2m$ is even.

4.15.1 Why spinors need more than a metric

A vector at a point transforms under a rotation $R \in \mathrm{SO}(n)$. A spinor transforms under an element g of the group $\mathrm{Spin}(n)$, which maps onto $\mathrm{SO}(n)$ two-to-one: g and $-g$ give the same rotation. The simplest case shows everything.

The two-dimensional case, explicitly. Take two anticommuting matrices with Nash's normalisation $\gamma_a^2 = -1$ (we write them down in Section 4.15.3; all we need now is $\gamma_1\gamma_2 = -\gamma_2\gamma_1$

and $\gamma_1^2 = \gamma_2^2 = -1$). Then $(\gamma_1\gamma_2)^2 = -\gamma_1^2\gamma_2^2 = -1$, so $\gamma_1\gamma_2$ behaves like the imaginary unit and

$$g_\theta := \exp\Big(\tfrac{\theta}{2}\gamma_1\gamma_2\Big) = \cos\tfrac{\theta}{2} + \sin\tfrac{\theta}{2}\,\gamma_1\gamma_2.$$

Let us see which rotation g_θ produces when it acts on "vectors" $v = v^1\gamma_1 + v^2\gamma_2$ by conjugation, $v \mapsto g_\theta v g_\theta^{-1}$:

$$\begin{aligned} g_\theta\,\gamma_1\,g_\theta^{-1} &\overset{(1)}{=} \gamma_1\,g_\theta^{-1}g_\theta^{-1} = \gamma_1\,g_{-2\theta} \\ &\overset{(2)}{=} \gamma_1\big(\cos\theta - \sin\theta\,\gamma_1\gamma_2\big) \\ &\overset{(3)}{=} \cos\theta\,\gamma_1 + \sin\theta\,\gamma_2. \end{aligned} \tag{4.119}$$

(1) $\gamma_1\gamma_2$ anticommutes with γ_1 ($\gamma_1\gamma_1\gamma_2 = -\gamma_1\gamma_2\gamma_1$), so moving γ_1 to the left through g_θ flips the sign of the sin term: $g_\theta\gamma_1 = \gamma_1 g_{-\theta} = \gamma_1 g_\theta^{-1}$.
(2) $g_{-\theta}g_{-\theta} = g_{-2\theta}$, and the half-angle formula with $\theta/2 \to -\theta$.
(3) $\gamma_1\gamma_1\gamma_2 = -\gamma_2$, so $-\sin\theta\,\gamma_1\gamma_1\gamma_2 = +\sin\theta\,\gamma_2$.

The same computation gives $g_\theta\gamma_2 g_\theta^{-1} = \cos\theta\,\gamma_2 - \sin\theta\,\gamma_1$. So g_θ rotates the plane by the angle θ, while it itself depends on $\theta/2$. Going once round, $\theta = 2\pi$, the rotation is the identity but

$$g_{2\pi} = \cos\pi + \sin\pi\,\gamma_1\gamma_2 = -1.$$

Only $\theta = 4\pi$ brings g back to $+1$. The group $\mathrm{Spin}(2) = \{g_\theta\}$ is a circle that wraps twice around the circle $\mathrm{SO}(2) = \{R_\theta\}$ (Figure 4.43).

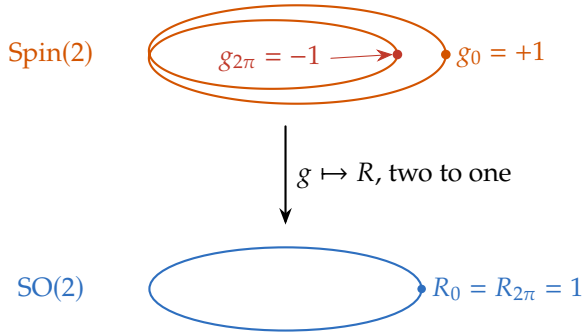


Figure 4.43: $\mathrm{Spin}(2) \to \mathrm{SO}(2)$ is the connected double cover of a circle. Follow the orange curve once round ($\theta : 0 \to 2\pi$): the rotation below comes back to the identity, but the spin element arrives at -1, on the other sheet. A second turn is needed to return to $+1$.

In three dimensions the same thing is the familiar statement $\mathrm{Spin}(3) = \mathrm{SU}(2)$: a 2×2 unitary matrix U rotates the vector $\vec{v}$ by $U(\vec{v}\cdot\vec{\sigma})U^\dagger = (R\vec{v})\cdot\vec{\sigma}$, and $\pm U$ give the same R. In four Euclidean dimensions $\mathrm{Spin}(4) = \mathrm{SU}(2)\times\mathrm{SU}(2)$, which we will see explicitly below. All of this is developed in Chapter T4; here we only need the two-to-one map and the sign $g_{2\pi} = -1$.

Physical picture: What the sign means physically

A spinor field is a quantity whose sign is flipped by a 2π rotation of the frame. On a manifold we describe fields in local orthonormal frames, and neighbouring frames are related by rotations. To define a spinor field globally we must lift *every* such frame rotation from $\mathrm{SO}(n)$ to $\mathrm{Spin}(n)$, consistently. Each lift is ambiguous by a sign, and the question is whether the signs can be chosen consistently round every loop and over every triple overlap. Sometimes they cannot.

4.15.2 Spin structures and the Stiefel–Whitney classes

Cover X by charts U_i, each with an orthonormal frame of TX. On an overlap $U_i \cap U_j$ the two frames differ by a rotation-or-reflection $g_{ij}(x) \in \mathrm{O}(n)$; these are the **transition functions** of the tangent bundle, and they satisfy $g_{ij}g_{jk}g_{ki} = 1$ on triple overlaps.

Working definition 4.37 (Orientation, spin structure, w_1, w_2)

Orientation. The signs $\det g_{ij} = \pm 1$ form a $\mathbb{Z}_2$-valued Čech 1-cocycle [a sign on each overlap, with product $+1$ around every triple overlap]. Its class [two such sign assignments count as the same if they differ by flipping the frames on some charts] is the **first Stiefel–Whitney class** $w_1(X) \in H^1(X;\mathbb{Z}_2)$. X is orientable iff we can flip frames so that all $\det g_{ij} = +1$, iff $w_1(X) = 0$. Then $g_{ij} \in \mathrm{SO}(n)$.
Spin structure. Choose, for every overlap, one of the two lifts [$\mathrm{Spin}(n) \to \mathrm{SO}(n)$ is two-to-one, so each g_{ij} has exactly two preimages $\pm\tilde g_{ij}$] $\tilde g_{ij} \in \mathrm{Spin}(n)$ of $g_{ij} \in \mathrm{SO}(n)$, with $\tilde g_{ji} = \tilde g_{ij}^{-1}$. On a triple overlap $\tilde g_{ij}\tilde g_{jk}\tilde g_{ki}$ lies over $g_{ij}g_{jk}g_{ki} = 1$, so it is a sign:

$$\tilde g_{ij}\tilde g_{jk}\tilde g_{ki} = \varepsilon_{ijk} \in \{\pm 1\}.$$

The signs ε_{ijk} form a Čech 2-cocycle whose class is the **second Stiefel–Whitney class** $w_2(X) \in H^2(X;\mathbb{Z}_2)$. If $w_2(X) = 0$ we can re-choose the lifts so that all $\varepsilon_{ijk} = +1$. Such a consistent choice of lifts (up to the obvious equivalence [changing $\tilde g_{ij}$ to $\eta_i\tilde g_{ij}\eta_j$, with one sign $\eta_i = \pm 1$ per chart]) is a **spin structure**, and X with a spin structure is a **spin manifold**.
Test in practice: for a complex manifold, $w_2 = c_1 \bmod 2$; for a simply connected 4-manifold, $w_2 = 0$ iff the intersection form is even [$q(a,a)$ is even for every integral class a] (see Section 4.15.9).

This is Nash's (4.96):

$$w_1(X) = 0 \iff X \text{ orientable}, \qquad w_2(X) = 0 \iff \text{spinors are globally defined on } X. \tag{Nash 4.96}$$

How many spin structures? Suppose one spin structure $\{\tilde g_{ij}\}$ has been found. Any other one is $\{\eta_{ij}\tilde g_{ij}\}$ with signs $\eta_{ij} = \pm 1$, and the cocycle condition for the new lifts says $\eta_{ij}\eta_{jk}\eta_{ki} = 1$: η is a $\mathbb{Z}_2$ 1-cocycle. Changing the lift in one chart by a sign (which is a mere relabelling) changes η by a coboundary. So the spin structures are in one-to-one correspondence with $H^1(X;\mathbb{Z}_2)$, once a reference one is chosen. A simply connected manifold (such as S^n or $\mathbb{CP}^n$) has $H^1(X;\mathbb{Z}_2) = 0$, hence at most one spin structure.

Watch out: A torsor, not a group

Nash says that each spin structure "determines an element of $H^1(X;\mathbb{Z}_2)$". More precisely the *difference* of two spin structures is an element of $H^1(X;\mathbb{Z}_2)$. There is no preferred "zero" spin structure in general, so the set of spin structures is a copy of $H^1(X;\mathbb{Z}_2)$ with the origin forgotten (a torsor). On the circle, for example, the two spin structures are genuinely different (one bounds a disc, one does not; see below), and neither is "the trivial one" in any invariant sense.

Worked example 4.76: Line bundles on S^2, and why TS^2 is spin

Cover S^2 by a northern and a southern disc overlapping in a band around the equator, with angular coordinate ϕ. An oriented real 2-plane bundle (equivalently a complex line

bundle L) is specified by one transition function on the band, a rotation

$$g_{NS}(\phi) = R_{d\phi}, \qquad d \in \mathbb{Z},$$

and the integer d is the degree, $d = \int c_1(L)$. For the tangent bundle $d = 2$ (Example 4.49): the frame "∂_x, ∂_y" of the northern chart turns twice relative to that of the southern chart as we go once round the equator.
A lift to Spin(2) must be a continuous function $\tilde{g}_{NS}(\phi)$ on the band with $\tilde{g}_{NS} \mapsto R_{d\phi}$. By (4.119), the lift of R_α is $\pm g_\alpha = \pm\exp(\frac{\alpha}{2}\gamma_1\gamma_2)$, so the only candidates are $\pm g_{d\phi}$. Following ϕ once round the band:

$$\begin{aligned} g_{d(\phi+2\pi)} &\overset{(1)}{=} \exp\left(\tfrac{d\phi}{2}\gamma_1\gamma_2\right)\exp\left(d\pi\gamma_1\gamma_2\right) \\ &\overset{(2)}{=} (-1)^d\, g_{d\phi}. \end{aligned}$$

(1) $\gamma_1\gamma_2$ commutes with itself, so the exponential of a sum is the product of exponentials.
(2) $\exp(d\pi\gamma_1\gamma_2) = \cos(d\pi) + \sin(d\pi)\gamma_1\gamma_2 = (-1)^d$.

So the lift is single-valued on the band iff d is even. With only two charts there are no triple overlaps, and this single-valuedness is the whole obstruction: $w_2(L) = d \bmod 2 = c_1(L) \bmod 2$. In particular TS^2 ($d = 2$) is spin, with the lift g_ϕ that turns *once* while the frames turn twice (Figure 4.44).

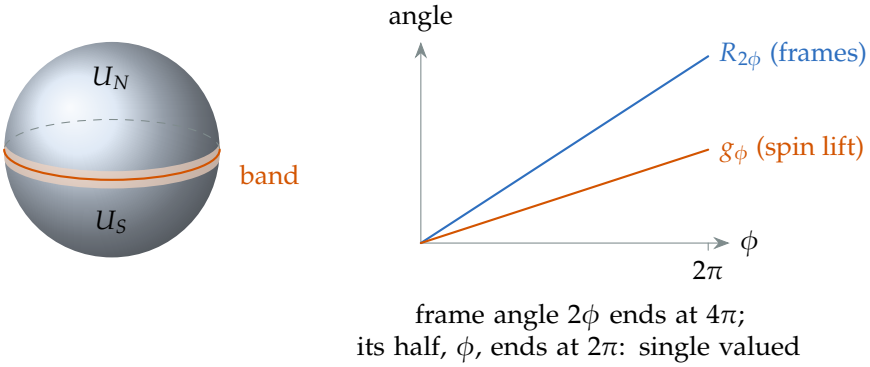


Figure 4.44: Lifting the transition function of TS^2. Across the equatorial band the northern frame is rotated by 2ϕ relative to the southern one (degree $d = 2$). The spin lift needs half that angle, ϕ, and returns to itself after $\phi \to \phi + 2\pi$. For a bundle of odd degree the half angle would end at an odd multiple of π, i.e. at -1: no spin lift.

Complex projective spaces. The computation above is the prototype of a general rule: for a complex manifold, $w_2(TX)$ is the reduction mod 2 of $c_1(TX)$. (Split TX into line bundles; each contributes its degree mod 2.) For $\mathbb{CP}^n$ the tangent bundle has $c(T\mathbb{CP}^n) = (1+h)^{n+1}$, with h the hyperplane class (the Euler sequence; the case $n = 2$ was used in Example 4.50), so

$$\begin{aligned} w_2(\mathbb{CP}^n) &\overset{(1)}{=} c_1(T\mathbb{CP}^n) \bmod 2 \\ &\overset{(2)}{=} (n+1)\,h \bmod 2, \end{aligned}$$

(1) The rule $w_2 = c_1 \bmod 2$ for complex manifolds.
(2) The degree-2 part of $(1+h)^{n+1}$ is $(n+1)h$.

and h mod 2 generates $H^2(\mathbb{CP}^n;\mathbb{Z}_2)$. So $\mathbb{CP}^n$ is spin iff n is odd, as Nash says: $\mathbb{CP}^1 = S^2$ and $\mathbb{CP}^3$ are spin, $\mathbb{CP}^2$ and $\mathbb{CP}^4$ are not. Being simply connected, the spin ones have exactly one spin structure. We shall see in Example 4.85 that the index theorem itself "knows" that $\mathbb{CP}^2$ is not spin.

Worked example 4.77: The two spin structures on a circle

For $X = S^1$ the frame is just the unit tangent vector, and there are no rotations to lift: $\mathrm{SO}(1) = \{1\}$ and $\mathrm{Spin}(1) = \{\pm 1\}$. A spin structure is then a double cover of the circle, and there are $\left|H^1(S^1;\mathbb{Z}_2)\right| = 2$ of them (Figure 4.45):

- the trivial cover (two disjoint circles): spinor fields are **periodic**, $\psi(\theta + 2\pi) = \psi(\theta)$. String theorists call this the **Ramond** sector;
- the connected cover (one circle of length 4π): spinor fields are **antiperiodic**, $\psi(\theta + 2\pi) = -\psi(\theta)$, the **Neveu–Schwarz** sector. Fermions at finite temperature live here: the Euclidean time circle carries antiperiodic fermions.

Which one extends over the disc D^2 whose boundary is the circle? Take a constant spinor ψ_0 on the flat disc, in the constant frame (∂_x, ∂_y). On the boundary circle the natural frame is the tangent frame $(\partial_r, \partial_\theta/r)$, rotated by the angle θ relative to (∂_x, ∂_y). Re-expressed in the rotating frame the same spinor has components $g_{-\theta}\psi_0$, and $g_{-(\theta+2\pi)} = -g_{-\theta}$. So the spin structure that bounds the disc is the *antiperiodic* one; the periodic one does not bound. On a torus $T^2 = S^1 \times S^1$ there are $2 \times 2 = 4$ spin structures, labelled by periodic (P) or antiperiodic (A) in each direction; on a surface of genus g there are 2^{2g}.

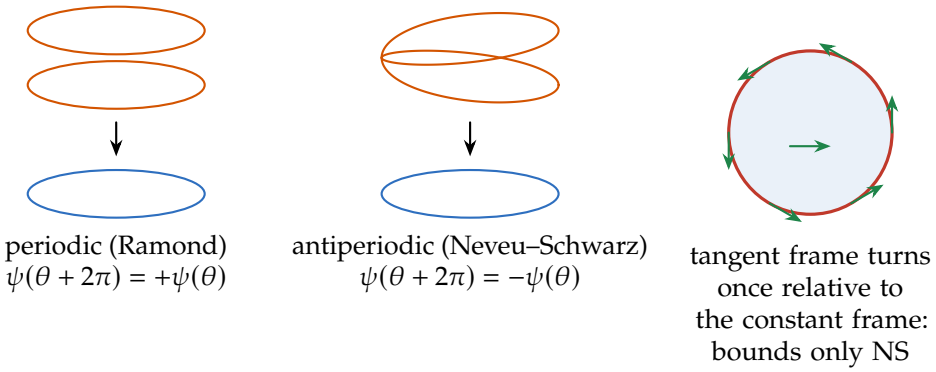


Figure 4.45: The two spin structures on S^1 are the two double covers of the circle (blue). Spinor fields are functions on the cover that change sign under the deck transformation. Right: along the boundary of a disc the tangent frame (green) makes one full turn relative to a constant frame, so the spin structure inherited from the disc is the antiperiodic one.

4.15.3 Spinors: the representation and the gamma matrices

From now on $n = 2m$ is even. Nash (4.97)–(4.99) collects the facts about the spin representation. We state them and then exhibit them.

Working definition 4.38 (Gamma matrices, chirality, Weyl spinors)

The **gamma matrices** are $N \times N$ complex matrices $\gamma_1, \dots, \gamma_n$, $N = 2^{n/2}$, with (Nash's convention, Euclidean)

$$\gamma_a\gamma_b + \gamma_b\gamma_a = \{\gamma_a, \gamma_b\} = -2\delta_{ab}\, I, \qquad \gamma_5 := i^{n(n+1)/2}\gamma_1\gamma_2\cdots\gamma_n. \tag{Nash 4.99}$$

They can be chosen anti-Hermitian, $\gamma_a^\dagger = -\gamma_a$. The **chirality operator** γ_5 satisfies $\gamma_5^2 = 1$ and $\{\gamma_5, \gamma_a\} = 0$. The space of spinors $S = \mathbb{C}^N$ splits into the ± 1 eigenspaces of γ_5, the **Weyl** or **chiral spinors**:

$$S = S^+ \oplus S^-, \qquad \dim S = 2^{n/2}, \qquad \dim S^\pm = 2^{n/2-1}. \tag{Nash 4.97}$$

For odd n one has $N = 2^{(n-1)/2}$ and no chirality splitting. *Formally:* S is the unique irreducible module [no proper nonzero subspace is mapped into itself by all γ_a] of the complex Clifford algebra [polynomials in the γ_a, with only the anticommutation rule imposed] $\mathrm{Cl}_n \otimes \mathbb{C}$; restricted to $\mathrm{Spin}(n) \subset \mathrm{Cl}_n$ it splits into the two inequivalent half-spin representations $S^\pm$ [$\mathrm{Spin}(n)$ is generated by the products $\gamma_a\gamma_b$, which commute with γ_5, so it maps S^+ and S^- to themselves].

Physicists usually take $\{\gamma_a, \gamma_b\} = +2\delta_{ab}$. The two conventions differ by $\gamma_a^{\text{phys}} = i\gamma_a^{\text{Nash}}$; nothing below depends on which is used, except that signs in intermediate formulas change.

Let us verify $\gamma_5^2 = 1$, because the odd-looking phase is chosen exactly for it:

$$\begin{aligned}
\gamma_5^2 &\overset{(1)}{=} i^{n(n+1)}\,(\gamma_1\cdots\gamma_n)(\gamma_1\cdots\gamma_n) \\
&\overset{(2)}{=} i^{n(n+1)}\,(-1)^{n(n-1)/2}\,\gamma_1^2\gamma_2^2\cdots\gamma_n^2 \\
&\overset{(3)}{=} (-1)^{n(n+1)/2}(-1)^{n(n-1)/2}(-1)^n \\
&\overset{(4)}{=} (-1)^{n^2+n} = 1.
\end{aligned}$$

(1) Definition of γ_5.
(2) Reverse the order of the second product to bring equal matrices next to each other. Reversing n anticommuting objects takes $n(n-1)/2$ transpositions, each costing a sign.
(3) $i^{n(n+1)} = (i^2)^{n(n+1)/2}$, and each $\gamma_a^2 = -1$.
(4) Add the exponents: $\frac{n(n+1)+n(n-1)}{2} + n = n^2 + n$, which is even.

And $\gamma_a\gamma_5 = -\gamma_5\gamma_a$ for n even: moving γ_a through the product $\gamma_1\cdots\gamma_n$ passes $n-1$ matrices that anticommute with it and one (itself) that commutes, a total sign $(-1)^{n-1} = -1$.

The chirality operator is a product over 2-planes. Group the product in pairs:

$$\begin{aligned}
\gamma_5 &\overset{(1)}{=} i^{m(2m+1)}\,(\gamma_1\gamma_2)(\gamma_3\gamma_4)\cdots(\gamma_{n-1}\gamma_n) \\
&\overset{(2)}{=} \prod_{j=1}^{m}\big(-i\,\gamma_{2j-1}\gamma_{2j}\big).
\end{aligned} \tag{4.120}$$

(1) $n(n+1)/2 = m(2m+1)$ for $n = 2m$.
(2) $i^{m(2m+1)} = (i^2)^{m^2} i^m = (-1)^m i^m = (-i)^m$, using $m^2 \equiv m \pmod 2$; distribute one factor $-i$ to each pair.

Each factor $-i\gamma_{2j-1}\gamma_{2j}$ squares to 1 (since $(\gamma_{2j-1}\gamma_{2j})^2 = -1$), and factors for different planes commute. So chirality is the product of m independent two-dimensional chiralities. This observation is all we will need for the characteristic classes.

Worked example 4.78: Gamma matrices in two dimensions

Take

$$\gamma_1 = i\sigma_2 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \qquad \gamma_2 = i\sigma_1 = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}.$$

Then $\gamma_1^2 = -\sigma_2^2 = -1$, $\gamma_2^2 = -\sigma_1^2 = -1$, and $\{\gamma_1, \gamma_2\} = -\{\sigma_2, \sigma_1\} = 0$. For the chirality:

$$\begin{aligned} \gamma_5 &\overset{(1)}{=} i^3 \gamma_1\gamma_2 = -i\,(i\sigma_2)(i\sigma_1) \\ &\overset{(2)}{=} i\,\sigma_2\sigma_1 = i(-i\sigma_3) = \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \end{aligned}$$

(1) Nash's definition with $n = 2$: $i^{n(n+1)/2} = i^3 = -i$.
(2) $\sigma_2\sigma_1 = -i\sigma_3$.

So S^+ is the upper component and S^- the lower. The rotation generator is $\gamma_1\gamma_2 = i\gamma_5 = i\sigma_3$, hence

$$g_\theta = \exp\!\left(\tfrac{\theta}{2}\gamma_1\gamma_2\right) = \begin{pmatrix} e^{i\theta/2} & 0 \\ 0 & e^{-i\theta/2} \end{pmatrix}.$$

A rotation of the plane by θ multiplies a positive-chirality spinor by $e^{i\theta/2}$ and a negative-chirality one by $e^{-i\theta/2}$. The tangent plane viewed as a complex line ($e_1 \leftrightarrow 1$, $e_2 \leftrightarrow i$) is multiplied by $e^{i\theta}$. So S^+ is a "square root" of the tangent line, $S^+ \otimes S^+ \cong T$, and S^- is a square root of its dual. In the Chern-root language of Section 4.9: if x is the root of T, then $S^\pm$ have roots $\pm x/2$.

Worked example 4.79: Gamma matrices in four dimensions and $\mathrm{Spin}(4) = \mathrm{SU}(2) \times \mathrm{SU}(2)$

Let s_a, $a = 1, \dots, 4$, be the 2×2 matrices

$$s_1 = i\sigma_1, \quad s_2 = i\sigma_2, \quad s_3 = i\sigma_3, \quad s_4 = I,$$

which represent the quaternions $\mathbf{i}, \mathbf{j}, \mathbf{k}, 1$. Their Hermitian conjugates are $\bar{s}_a := s_a^\dagger$ (so $\bar{s}_k = -s_k$ for $k \le 3$ and $\bar{s}_4 = I$). The key identity is

$$s_a\bar{s}_b + s_b\bar{s}_a = 2\delta_{ab} I = \bar{s}_a s_b + \bar{s}_b s_a, \tag{4.121}$$

which is the quaternion statement $p\bar{q} + q\bar{p} = 2\langle p, q\rangle$; for example $s_1\bar{s}_2 + s_2\bar{s}_1 = -(i\sigma_1)(i\sigma_2) - (i\sigma_2)(i\sigma_1) = \{\sigma_1, \sigma_2\} = 0$ and $s_1\bar{s}_1 = \sigma_1^2 = I$. Define the 4×4 matrices

$$\gamma_a = \begin{pmatrix} 0 & s_a \\ -\bar{s}_a & 0 \end{pmatrix}.$$

Then

$$\begin{aligned} \{\gamma_a, \gamma_b\} &\overset{(1)}{=} \begin{pmatrix} -s_a\bar{s}_b - s_b\bar{s}_a & 0 \\ 0 & -\bar{s}_a s_b - \bar{s}_b s_a \end{pmatrix} \\ &\overset{(2)}{=} -2\delta_{ab}\, I_4. \end{aligned}$$

(1) Multiply the block matrices: the product of two off-diagonal blocks is block diagonal.

(2) The identity (4.121).

For the chirality ($n = 4$, $i^{10} = -1$):

$$\gamma_5 \overset{(1)}{=} -\gamma_1\gamma_2\gamma_3\gamma_4 = -\begin{pmatrix} s_1\bar{s}_2s_3\bar{s}_4 & 0 \\ 0 & \bar{s}_1s_2\bar{s}_3s_4 \end{pmatrix}$$
$$\overset{(2)}{=} -\begin{pmatrix} (i\sigma_1)(-i\sigma_2)(i\sigma_3) & 0 \\ 0 & (-i\sigma_1)(i\sigma_2)(-i\sigma_3) \end{pmatrix}$$
$$\overset{(3)}{=} -\begin{pmatrix} i\cdot i & 0 \\ 0 & -i\cdot i \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix}.$$

(1) $\gamma_1\gamma_2 = \mathrm{diag}(-s_1\bar{s}_2, -\bar{s}_1s_2)$ and similarly for $\gamma_3\gamma_4$; the two minus signs cancel.
(2) Insert the matrices; $s_4 = \bar{s}_4 = I$.
(3) $\sigma_1\sigma_2\sigma_3 = iI$; the prefactors are $i(-i)i = i$ and $(-i)i(-i) = -i$.

This is Nash's block form (4.101). The generators of rotations, $\gamma_a\gamma_b$ with $a \neq b$, are block diagonal: $\gamma_a\gamma_b = \mathrm{diag}(-s_a\bar{s}_b, -\bar{s}_as_b)$, and each block is a traceless anti-Hermitian 2×2 matrix, i.e. an element of $\mathfrak{su}(2)$. So Spin(4) acts on $S^+ = \mathbb{C}^2$ and $S^- = \mathbb{C}^2$ by two separate SU(2) matrices: $\mathrm{Spin}(4) = \mathrm{SU}(2)_+ \times \mathrm{SU}(2)_-$.

Which rotations act on which half? Take the self-dual combination $\gamma_1\gamma_2 + \gamma_3\gamma_4$ (it corresponds to the 2-form $e^1\wedge e^2 + e^3\wedge e^4 = \star(e^1\wedge e^2 + e^3\wedge e^4)$). On S^- it is $-\bar{s}_1s_2 - \bar{s}_3s_4 = -\sigma_1\sigma_2 + i\sigma_3 = -i\sigma_3 + i\sigma_3 = 0$. On S^+ it is $-s_1\bar{s}_2 - s_3\bar{s}_4 = -\sigma_1\sigma_2 - i\sigma_3 = -2i\sigma_3 \neq 0$. *Self-dual 2-forms act only on S^+, anti-self-dual ones only on S^-* (check $\gamma_1\gamma_2 - \gamma_3\gamma_4$ yourself). This is the spinor form of the splitting $\mathfrak{so}(4) = \mathfrak{su}(2)_+ \oplus \mathfrak{su}(2)_- \cong \Lambda^+ \oplus \Lambda^-$, and it will decide the chirality of instanton zero modes in Section 4.16.

Weights in general. Put the rotation by angles θ_j in the planes (e_{2j-1}, e_{2j}) together. Its spin lift is $\prod_j \exp(\frac{\theta_j}{2}\gamma_{2j-1}\gamma_{2j})$. By (4.120) we may diagonalise all the commuting operators $-i\gamma_{2j-1}\gamma_{2j}$ at once; their eigenvalues are signs $\varepsilon_j = \pm 1$, and on the joint eigenvector the lift acts by

$$\prod_j \exp\Big(\tfrac{\theta_j}{2}\, i\varepsilon_j\Big) = \exp\Big(\tfrac{i}{2}\sum_j \varepsilon_j\theta_j\Big), \qquad \gamma_5 = \prod_j \varepsilon_j.$$

There are $2^m = \dim S$ choices of signs, one per basis spinor, so

$$S \text{ has weights } \tfrac{1}{2}(\pm\theta_1 \pm \cdots \pm \theta_m); \qquad S^+ : \text{ an even number of } - \text{ signs}, \quad S^- : \text{ odd}. \tag{4.122}$$

(In two dimensions: S^+ has weight $+\theta/2$, S^- has $-\theta/2$, as in Example 4.78.)

The spinor bundles. On a spin manifold the frame bundle P of TX (a principal $\mathrm{SO}(n)$ bundle) has a double cover $\tilde{P}$ with group $\mathrm{Spin}(n)$; concretely $\tilde{P}$ is glued from the lifts $\tilde{g}_{ij}$. Nash defines

$$E^\pm = \tilde{P} \times_{\mathrm{Spin}(n)} S^\pm, \qquad E = E^+ \oplus E^-. \tag{Nash 4.98}$$

In words: a spinor field is, in each chart U_i, a column vector $\psi_i(x) \in \mathbb{C}^{2^m}$, and on overlaps $\psi_i = \tilde{g}_{ij}\psi_j$. The cocycle condition $\tilde{g}_{ij}\tilde{g}_{jk}\tilde{g}_{ki} = 1$ is exactly what makes this consistent. Since γ_5 commutes with every $\tilde{g}_{ij}$ (spin elements are even products of gammas), the splitting into $E^\pm$ is global.

4.15.4 The Dirac operator and the spin complex

To differentiate a spinor field we need a connection on E, and the metric provides one. Let $e_a = e_a^\mu\partial_\mu$ be a local orthonormal frame (the **vielbein** or **tetrad**), with dual coframe $e^a = e^a_\mu \mathrm{d}x^\mu$,

so that

$$e^a_\mu e^b_\nu \delta_{ab} = g_{\mu\nu}.$$

The Levi-Civita connection in this frame is a matrix of 1-forms $\omega^a{}_b = -\omega^b{}_a$, defined by $\nabla e_b = e_a\,\omega^a{}_b$, and determined by Cartan's first structure equation $de^a = -\omega^a{}_b \wedge e^b$ (Chapter T2). At each point $\omega_\mu = (\omega_\mu{}^a{}_b)$ is an infinitesimal rotation, an element of $\mathfrak{so}(n)$. To make it act on spinors we lift it to the Lie algebra of $\mathrm{Spin}(n)$.

Lifting an infinitesimal rotation. Let $A = (A^a{}_b)$ be an antisymmetric matrix, $Ae_b = e_a A^a{}_b$. We claim its lift is

$$\rho(A) = -\tfrac{1}{4} A^a{}_b\, \gamma_a \gamma_b \qquad \text{(sum over } a, b\text{)}. \tag{4.123}$$

Test it on the rotation of the (e_1, e_2) plane by the angle θ: $A^2{}_1 = \theta$, $A^1{}_2 = -\theta$. Then

$$\begin{aligned}\rho(A) &\overset{(1)}{=} -\tfrac{1}{4}\big(A^1{}_2\gamma_1\gamma_2 + A^2{}_1\gamma_2\gamma_1\big)\\ &\overset{(2)}{=} -\tfrac{1}{4}\big(-\theta\gamma_1\gamma_2 - \theta\gamma_1\gamma_2\big) = \tfrac{\theta}{2}\gamma_1\gamma_2,\end{aligned}$$

(1) Only the entries $A^1{}_2$ and $A^2{}_1$ are nonzero.
(2) $\gamma_2\gamma_1 = -\gamma_1\gamma_2$.

which is the generator of g_θ from Section 4.15.1, as required. (Equivalently: $[\rho(A), \gamma_c] = \gamma_a A^a{}_c$, the infinitesimal version of (4.119).)

Working definition 4.39 (Dirac operator)

The **spin connection** is $\nabla_\mu \psi = (\partial_\mu + \omega_\mu)\psi$ with $\omega_\mu = -\tfrac{1}{4}\omega_\mu{}^a{}_b \gamma_a \gamma_b$ [$\omega_\mu{}^a{}_b$ is the Levi-Civita connection written in the orthonormal frame e_a]. The **Dirac operator** is Clifford multiplication [contracting the free frame index with γ^a] composed with it:

$$\not{D} : \Gamma(X, E) \to \Gamma(X, E), \qquad \not{D} = \gamma^a e^\mu_a(x)\big(\partial_\mu + \omega_\mu\big), \tag{Nash 4.100}$$

where Latin indices are frame indices (raised and lowered with δ_{ab}, so $\gamma^a = \gamma_a$) and Greek indices are coordinate indices [E is the spinor bundle, with fibre $S = \mathbb{C}^N$, and $e_a = e^\mu_a \partial_\mu$ are the orthonormal frame vectors]. *To compute:* find e^a [the dual coframe, $e^a(e_b) = \delta^a_b$], solve $de^a = -\omega^a{}_b \wedge e^b$ for ω, insert.

In flat space with the constant frame, $\omega = 0$ and $\not{D} = \gamma^a \partial_a$. Squaring it recovers the Laplacian, which is the reason Dirac introduced it:

$$\begin{aligned}\not{D}^2 &\overset{(1)}{=} \gamma^a\gamma^b\partial_a\partial_b = \tfrac{1}{2}\{\gamma^a, \gamma^b\}\partial_a\partial_b\\ &\overset{(2)}{=} -\delta^{ab}\partial_a\partial_b = -\sum_a \partial_a^2.\end{aligned} \tag{4.124}$$

(1) $\partial_a\partial_b$ is symmetric in a, b, so only the symmetric part of $\gamma^a\gamma^b$ survives.
(2) Nash's Clifford relation (Nash 4.99).

With Nash's sign convention $\not{D}^2$ is the *positive* Laplacian, and $\not{D}$ is formally self-adjoint: γ^a and ∂_a are both anti-Hermitian, so their product is Hermitian.

The chiral Dirac operator. Since γ_5 anticommutes with every γ^a and commutes with ω_μ (an even product of gammas),

$$\not{D}\gamma_5 = -\gamma_5 \not{D}.$$

In the block form $\gamma_5 = \mathrm{diag}(I, -I)$ of Nash's (4.101) this means $\not{D}$ is block off-diagonal:

$$\not{D} = \begin{pmatrix} 0 & \not{D}^- \\ \not{D}^+ & 0 \end{pmatrix}, \qquad \not{D}^+ = \not{D}\,\tfrac{1+\gamma_5}{2} : \Gamma(E^+) \to \Gamma(E^-), \tag{4.125}$$

and $\not{D}^- : \Gamma(E^-) \to \Gamma(E^+)$. Nash writes $\not{\partial}$ for the chiral operator $\not{D}^+$ and calls the two-term complex

$$0 \longrightarrow \Gamma(X, E^+) \xrightarrow{\not{D}^+} \Gamma(X, E^-) \longrightarrow 0, \qquad 0 \longrightarrow \pi^* E^+ \xrightarrow{\sigma_1(\not{D}^+)} \pi^* E^- \longrightarrow 0 \tag{Nash 4.102}$$

the **spin complex** and its symbol complex.

Ellipticity, and why it needs Euclidean signature. The symbol is obtained by $\partial_\mu \to i p_\mu$ in the highest-order part (the connection term ω_μ is lower order and drops out):

$$\sigma_1(\not{D})(x, p) = i\gamma^a e_a^\mu p_\mu =: i\,\not{p}.$$

Its square is

$$\begin{aligned} (i\not{p})^2 &\overset{(1)}{=} -\gamma^a\gamma^b\, p_a p_b \\ &\overset{(2)}{=} \delta^{ab} p_a p_b = |p|^2 , \end{aligned}$$

(1) $i^2 = -1$; $p_a = e_a^\mu p_\mu$ are the frame components.
(2) Symmetrise and use (Nash 4.99), as in (4.124).

so $\sigma_1(\not{D})(x,p)^{-1} = i\not{p}/|p|^2$ exists for every $p \neq 0$: $\not{D}$, and hence each block $\not{D}^\pm$, is elliptic. In Lorentzian signature the same computation gives $p_0^2 - |\vec{p}|^2$, which vanishes on the light cone; there the symbol is not invertible and the operator is hyperbolic, not elliptic (Figure 4.46). The Dirac index is a Euclidean notion, which is why physicists meet it after a Wick rotation, in instanton and anomaly calculations.

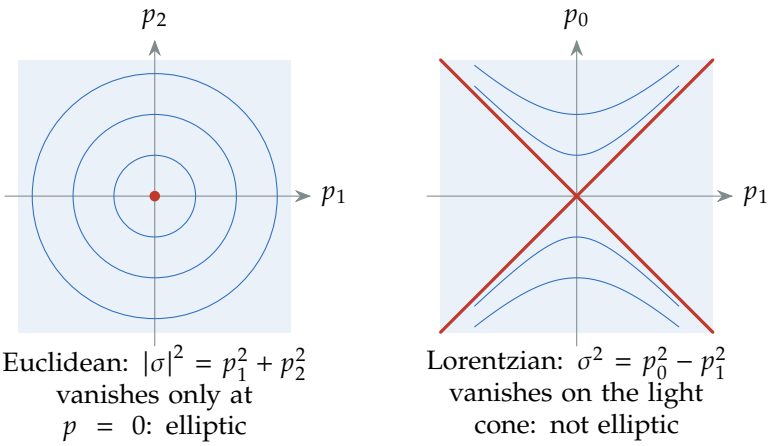


Figure 4.46: Where the Dirac symbol fails to be invertible (red). With Euclidean signature the level sets of σ^2 are circles and only the origin is bad, which is what ellipticity requires. With Lorentzian signature the whole light cone is bad; the Dirac equation then propagates waves along it instead of having a finite-dimensional space of zero modes.

The adjoint. Give E the Hermitian inner product $\langle \psi, \chi \rangle = \int_X \psi^\dagger \chi \,\mathrm{vol}$. The Dirac operator is formally self-adjoint (the spin connection is metric, and the flat-space argument above goes through with ∇ in place of ∂; see Bär §2.5). Taking the adjoint of the block matrix (4.125) then gives $(\not{D}^+)^* = \not{D}^-$. Nash writes $\not{\partial}^*$ for $\not{D}^-$ and notes that it is also the operator one gets by interchanging the roles of E^+ and E^-.

Worked example 4.80: The flat Dirac operator in two dimensions

With the matrices of Example 4.78 and $z = x^1 + ix^2$, $\partial_z = \tfrac{1}{2}(\partial_1 - i\partial_2)$, $\partial_{\bar z} = \tfrac{1}{2}(\partial_1 + i\partial_2)$:

$$\not{D} \overset{(1)}{=} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \partial_1 + \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix} \partial_2 = \begin{pmatrix} 0 & \partial_1 + i\partial_2 \\ -\partial_1 + i\partial_2 & 0 \end{pmatrix} \overset{(2)}{=} \begin{pmatrix} 0 & 2\partial_{\bar z} \\ -2\partial_z & 0 \end{pmatrix}.$$

(1) Insert γ_1, γ_2; $\omega = 0$ in the constant frame.
(2) $\partial_1 + i\partial_2 = 2\partial_{\bar z}$ and $-\partial_1 + i\partial_2 = -2\partial_z$.

So $\not{D}^+ = -2\partial_z$ and $\not{D}^- = 2\partial_{\bar z}$. The Dirac equation in two dimensions is the Cauchy–Riemann equation: negative-chirality zero modes are holomorphic functions, positive-chirality ones antiholomorphic. On the plane there are infinitely many of each (e.g. z^k); on a compact surface only finitely many survive, and counting them is the index problem. This is why the two-dimensional Dirac index and the Riemann–Roch theorem of Section 4.13 are the same subject.

Worked example 4.81: The Dirac operator on the round two-sphere

Use the metric $\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2$ and the coframe $e^1 = \mathrm{d}\theta$, $e^2 = \sin\theta\,\mathrm{d}\phi$. *Spin connection:*

$$\mathrm{d}e^2 \overset{(1)}{=} \cos\theta\,\mathrm{d}\theta \wedge \mathrm{d}\phi \overset{(2)}{=} -\omega^2{}_1 \wedge e^1 \quad \text{with} \quad \omega^2{}_1 = \cos\theta\,\mathrm{d}\phi,$$

(1) Differentiate $e^2 = \sin\theta\,\mathrm{d}\phi$.
(2) Cartan's equation $\mathrm{d}e^2 = -\omega^2{}_1 \wedge e^1$: $-\cos\theta\,\mathrm{d}\phi \wedge \mathrm{d}\theta = \cos\theta\,\mathrm{d}\theta \wedge \mathrm{d}\phi$. The other equation, $\mathrm{d}e^1 = 0 = -\omega^1{}_2 \wedge e^2$, holds because $\omega^1{}_2 = -\cos\theta\,\mathrm{d}\phi$ is proportional to e^2.

Then by (4.123), only ω_ϕ is nonzero:

$$\omega_\phi \overset{(1)}{=} -\tfrac{1}{4}\big(\omega_\phi{}^1{}_2 \gamma_1\gamma_2 + \omega_\phi{}^2{}_1 \gamma_2\gamma_1\big) \overset{(2)}{=} -\tfrac{1}{4}\big(-\cos\theta\,\gamma_1\gamma_2 - \cos\theta\,\gamma_1\gamma_2\big) = \tfrac{1}{2}\cos\theta\,\gamma_1\gamma_2.$$

(1) The lift formula; only the $(1,2)$ and $(2,1)$ entries are nonzero.
(2) $\omega^1{}_2 = -\cos\theta\,\mathrm{d}\phi$, $\omega^2{}_1 = +\cos\theta\,\mathrm{d}\phi$, and $\gamma_2\gamma_1 = -\gamma_1\gamma_2$.

Dirac operator: with $e_1 = \partial_\theta$, $e_2 = \frac{1}{\sin\theta}\partial_\phi$,

$$\not{D} \overset{(1)}{=} \gamma_1\partial_\theta + \gamma_2 \frac{1}{\sin\theta}\Big(\partial_\phi + \tfrac{1}{2}\cos\theta\,\gamma_1\gamma_2\Big) \overset{(2)}{=} \gamma_1\Big(\partial_\theta + \tfrac{1}{2}\cot\theta\Big) + \gamma_2 \frac{1}{\sin\theta}\partial_\phi. \tag{4.126}$$

(1) Definition (Nash 4.100) with the frame above.
(2) $\gamma_2\gamma_1\gamma_2 = -\gamma_1\gamma_2\gamma_2 = \gamma_1$.

The spin connection has turned into the term $\frac{1}{2}\cot\theta$: the "spinor version" of the $\frac{1}{r}$ in the flat divergence $\partial_r + \frac{1}{r}$. In components (with Example 4.78)

$$\not{D}^+ = -\left(\partial_\theta + \tfrac{1}{2}\cot\theta\right) + \frac{i}{\sin\theta}\partial_\phi, \qquad \not{D}^- = \left(\partial_\theta + \tfrac{1}{2}\cot\theta\right) + \frac{i}{\sin\theta}\partial_\phi.$$

One subtlety: the frame (e_1, e_2) is singular at the poles, where it rotates once as we circle the pole. By the argument of Example 4.77, the spinor components in this frame are then *antiperiodic* in ϕ: $\psi \propto e^{ik\phi}$ with $k \in \mathbb{Z} + \frac{1}{2}$.

4.15.5 Harmonic spinors and the index

Working definition 4.40 (Harmonic spinors and the Dirac index)

A spinor with $\not{D}\psi = 0$ is a **harmonic spinor**, i.e. a zero mode of the massless Dirac operator. Write Ψ for the space of harmonic spinors. Since $\not{D}$ anticommutes with γ_5, γ_5 maps Ψ to itself, and Ψ splits by chirality:

$$\Psi = \Psi^+ \oplus \Psi^-, \qquad n^\pm := \dim \Psi^\pm. \tag{Nash 4.104}$$

Then

$$\operatorname{index} \not{D}^+ = \dim \operatorname{Ker} \not{D}^+ - \dim \operatorname{Ker}(\not{D}^+)^* = \dim \Psi^+ - \dim \Psi^- = n^+ - n^-.$$

(Nash 4.103–4.106)

[$\not{D}^+$ is $\not{D}$ acting on positive-chirality spinors. Its adjoint is $\not{D}^-$, since $\not{D}$ is self-adjoint, so $\operatorname{Ker}(\not{D}^+)^* = \Psi^-$] *Operationally:* count the zero modes of each chirality and subtract.

The middle step is Nash's (4.105): $\operatorname{Ker}\not{D}^+ = \Psi^+$ is clear, and $\operatorname{Ker}(\not{D}^+)^* = \operatorname{Ker}\not{D}^- = \Psi^-$ because $\not{D}\psi = 0$ for $\psi = \psi^+ + \psi^-$ means $\not{D}^+\psi^+ = 0$ and $\not{D}^-\psi^- = 0$ separately (they land in different chiralities). "Hodge theory" here is the observation that $\operatorname{Ker}\not{D} = \operatorname{Ker}\not{D}^2$ for the self-adjoint operator $\not{D}$: if $\not{D}^2\psi = 0$ then $\|\not{D}\psi\|^2 = \langle\psi, \not{D}^2\psi\rangle = 0$.

Physical picture: Why only the difference is topological

Nonzero modes of $\not{D}$ come in pairs: if $\not{D}\psi = \lambda\psi$ then $\not{D}(\gamma_5\psi) = -\lambda\gamma_5\psi$. Under a deformation of the metric, a pair $\pm\lambda$ can descend to zero and become one zero mode of each chirality, or a zero-mode pair can leave. Either way n^+ and n^- change by the same amount. So n^+ and n^- separately depend on the metric, while $n^+ - n^-$ does not. This is the McKean–Singer mechanism of the previous section, specialised to the Dirac operator.

The Lichnerowicz formula. There is one more piece of general information about harmonic spinors, and it is extremely useful. In curved space the square of the Dirac operator is not quite the Laplacian: a curvature term appears. The result (Lichnerowicz, 1963; Bär Theorem 2.5.11) is

$$\not{D}^2 = \nabla^*\nabla + \frac{R}{4}, \tag{4.127}$$

where R is the scalar curvature (normalised so that the unit S^n has $R = n(n-1)$) and $\nabla^*\nabla = -\sum_a \nabla_a\nabla_a$ + (connection terms) is the non-negative "rough Laplacian". Here is why. Write $\not{D}^2 = \gamma^a\gamma^b\nabla_a\nabla_b$ at a point where the frame is parallel, and split $\gamma^a\gamma^b$ into its symmetric

and antisymmetric parts:

$$\begin{aligned} \not{D}^2 &\overset{(1)}{=} \tfrac{1}{2}\{\gamma^a,\gamma^b\}\nabla_a\nabla_b + \tfrac{1}{2}\gamma^a\gamma^b[\nabla_a,\nabla_b] \\ &\overset{(2)}{=} -\nabla_a\nabla_a + \tfrac{1}{2}\gamma^a\gamma^b\,\rho\big(R(e_a,e_b)\big) \\ &\overset{(3)}{=} \nabla^*\nabla - \tfrac{1}{8}R_{ab}{}^{c}{}_{d}\,\gamma^a\gamma^b\gamma_c\gamma^d \overset{(4)}{=} \nabla^*\nabla + \tfrac{1}{4}R. \end{aligned}$$

(1) $\gamma^a\gamma^b = \frac{1}{2}\{\gamma^a,\gamma^b\} + \frac{1}{2}[\gamma^a,\gamma^b]$, and the antisymmetric part only sees the antisymmetric part $\frac{1}{2}[\nabla_a,\nabla_b]$ of $\nabla_a\nabla_b$.

(2) The Clifford relation for the first term. The commutator of covariant derivatives is the curvature, acting on spinors through the lift ρ of (4.123).

(3) At the chosen point $-\nabla_a\nabla_a = \nabla^*\nabla$; and $\rho(R(e_a,e_b)) = -\frac{1}{4}R_{ab}{}^{c}{}_{d}\gamma_c\gamma^d$.

(4) The product of four gammas contracted with the Riemann tensor collapses, using the first Bianchi identity $R_{a[bcd]} = 0$ to kill the totally antisymmetric part and the Clifford relation to contract the rest, to $-2R$. (This index gymnastics is done in full in Bär §2.5; we check the overall sign on S^2 just below.)

Key idea: Positive scalar curvature kills harmonic spinors

If $R > 0$ everywhere then, for any spinor ψ,

$$\big\|\not{D}\psi\big\|^2 = \|\nabla\psi\|^2 + \int_X \frac{R}{4}\,|\psi|^2\ \mathrm{vol} > 0 \quad (\psi \neq 0),$$

so $\operatorname{Ker}\not{D} = 0$: $n^+ = n^- = 0$ and $\operatorname{index}\not{D}^+ = 0$. Combined with the index theorem below: *a compact spin manifold with $\hat{A}(X) \neq 0$ admits no metric of positive scalar curvature.*

Worked example 4.82: No zero modes on S^2, by hand

Look for $\psi^+ = e^{ik\phi}f(\theta)$, $k \in \mathbb{Z} + \frac{1}{2}$, with $\not{D}^+\psi^+ = 0$. From Example 4.81:

$$\begin{aligned} 0 &\overset{(1)}{=} -\Big(f' + \tfrac{1}{2}\cot\theta\, f\Big) + \frac{i\cdot ik}{\sin\theta}f \\ &\overset{(2)}{=} -\Big(f' + \tfrac{1}{2}\cot\theta\, f + \frac{k}{\sin\theta}f\Big), \end{aligned}$$

(1) Insert ψ^+ into $\not{D}^+$; $\partial_\phi \to ik$.

(2) $i\cdot i = -1$; collect.

so $(\ln f)' = -\frac{1}{2}\cot\theta - \frac{k}{\sin\theta}$. Since $\int \cot\theta\,\mathrm{d}\theta = \ln\sin\theta$ and $\int \frac{\mathrm{d}\theta}{\sin\theta} = \ln\tan\frac{\theta}{2}$,

$$f(\theta) = (\sin\theta)^{-1/2}\big(\tan\tfrac{\theta}{2}\big)^{-k}.$$

Near the north pole $f \sim \theta^{-1/2-k}$, which is bounded only if $k \le -\frac{1}{2}$. Near the south pole, with $u = \pi - \theta$, $\sin\theta \approx u$ and $\tan\frac{\theta}{2} \approx \frac{2}{u}$, so $f \sim u^{-1/2+k}$, bounded only if $k \ge \frac{1}{2}$. No k does both: $\Psi^+ = 0$. The same computation for $\not{D}^-$ gives $f = (\sin\theta)^{-1/2}(\tan\frac{\theta}{2})^k$ and again no solution. So $n^+ = n^- = 0$.

This agrees with Lichnerowicz ($R = 2 > 0$ on the unit sphere). In fact the whole spectrum is known: the eigenvalues of $\not{D}$ on the unit S^2 are $\pm(\ell+1)$, $\ell = 0,1,2,\dots$, each with multiplicity $2(\ell+1)$ (Bär §2.5). The smallest eigenvalue of $\not{D}^2$ is 1, and the bound from (4.127) is $\not{D}^2 \ge R/4 = \frac{1}{2}$: consistent, and a check that the sign of the curvature term is right.

Worked example 4.83: The four spin structures on a flat torus

Let $T^2 = \mathbb{R}^2/(2\pi\mathbb{Z})^2$ with the flat metric and constant frame, so $\not{D}$ is the flat operator of Example 4.80 and $\not{D}^2 = -(\partial_1^2 + \partial_2^2)$. A spin structure is a choice of periodic or antiperiodic boundary condition in each direction. Plane waves $\psi = u\, e^{i(k_1 x^1 + k_2 x^2)}$ ($u \in \mathbb{C}^2$ constant) satisfy these conditions when

$$k_j \in \mathbb{Z}\ (\mathrm{P}) \qquad \text{or} \qquad k_j \in \mathbb{Z} + \tfrac{1}{2}\ (\mathrm{A}),$$

and they have $\not{D}^2\psi = (k_1^2 + k_2^2)\psi$. A zero mode needs $k = 0$, which is allowed only for the (P,P) structure. There the zero modes are the constant spinors: $u = (1,0)^T$ of positive chirality and $u = (0,1)^T$ of negative chirality. Hence

$$(\mathrm{P},\mathrm{P}):\ n^+ = n^- = 1; \qquad (\mathrm{P},\mathrm{A}),\ (\mathrm{A},\mathrm{P}),\ (\mathrm{A},\mathrm{A}):\ n^+ = n^- = 0.$$

In all four cases $\operatorname{index} \not{D}^+ = 0$, as it must be ($T^2$ is flat, and we will see that the two-dimensional index always vanishes). But the numbers $n^\pm$ themselves depend on the spin structure: they are not topological for the index theorem, only their difference is. (Their parity *is* a deformation invariant, the mod 2 index of the next part of this chapter: (P,P) is the "odd" spin structure.)

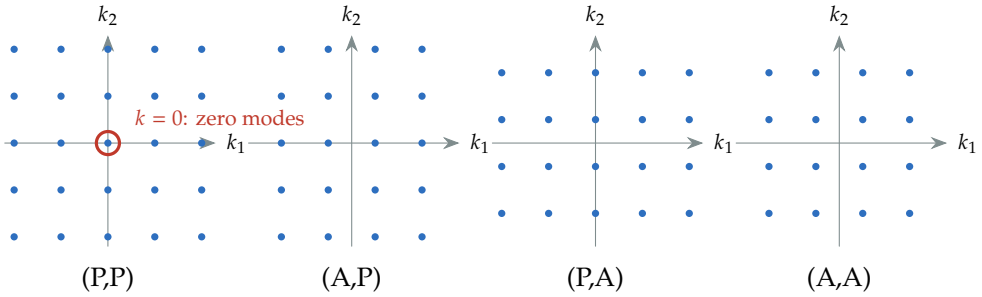


Figure 4.47: Allowed momenta for spinors on the square torus, for the four spin structures. $\not{D}^2$ has eigenvalue $|k|^2$, so zero modes sit at $k = 0$. Only the periodic–periodic structure contains the origin (red circle), and there the two constant spinors (one of each chirality) are the zero modes. The index $n^+ - n^-$ is zero in every case.

4.15.6 The Chern character of the spinor bundles

We now run the recipe (4.101) on the spin complex. The complex has two terms, $E^0 = E^+$ and $E^1 = E^-$, so the numerator we need is

$$\operatorname{ch}\big(\textstyle\sum_p (-1)^p E^p\big) = \operatorname{ch}(E^+) - \operatorname{ch}(E^-).$$

Nash quotes the answer (4.107) from Atiyah–Singer, saying it needs "some properties of the characters of the maximal torus of Spin(n)". Those properties are exactly the weights (4.122) that we have already found by hand. What remains is to turn weights into Chern roots.

Step 1: the curvature of the spinor bundle. The spinor bundle carries the connection $\nabla = \mathrm{d} + \omega$ with $\omega = \rho(\omega^a{}_b)$, the lift (4.123) of the Levi-Civita connection. Its curvature is the lift of the Riemann curvature, $\rho(R) = -\frac{1}{4} R^a{}_b \gamma_a \gamma_b$, because ρ is a Lie algebra homomorphism (it maps commutators of rotation generators to commutators of the $\gamma_a\gamma_b$). Now put R into the block

form (4.79): in the j-th plane $R^{2j-1}{}_{2j} = 2\pi x_j$ and $R^{2j}{}_{2j-1} = -2\pi x_j$, all other entries zero. Then

$$\begin{aligned}\rho(R) &\overset{(1)}{=} -\tfrac{1}{4}\sum_{j=1}^{m}\Big(R^{2j-1}{}_{2j}\,\gamma_{2j-1}\gamma_{2j} + R^{2j}{}_{2j-1}\,\gamma_{2j}\gamma_{2j-1}\Big)\\ &\overset{(2)}{=} -\tfrac{1}{4}\sum_{j=1}^{m}\big(2\pi x_j + 2\pi x_j\big)\gamma_{2j-1}\gamma_{2j} = -\pi\sum_{j=1}^{m} x_j\,\gamma_{2j-1}\gamma_{2j}\\ &\overset{(3)}{=} -i\pi\sum_{j=1}^{m} x_j\,\big(-i\gamma_{2j-1}\gamma_{2j}\big) = -i\pi\sum_{j=1}^{m}\varepsilon_j x_j \quad \text{on the joint eigenvector.}\end{aligned} \tag{4.128}$$

(1) The lift formula, keeping only the nonzero entries of the block matrix.
(2) $\gamma_{2j}\gamma_{2j-1} = -\gamma_{2j-1}\gamma_{2j}$ and $R^{2j}{}_{2j-1} = -2\pi x_j$, so both terms contribute $2\pi x_j\,\gamma_{2j-1}\gamma_{2j}$.
(3) Write $\gamma_{2j-1}\gamma_{2j} = i\,(-i\gamma_{2j-1}\gamma_{2j})$ and replace the commuting operators $-i\gamma_{2j-1}\gamma_{2j}$ by their eigenvalues $\varepsilon_j = \pm 1$, as in (4.122).

The Chern roots of a bundle are the eigenvalues of $\frac{i}{2\pi}\times$(curvature) (Section 4.9). So the Chern root belonging to the sign choice $\varepsilon = (\varepsilon_1, \dots, \varepsilon_m)$ is

$$\frac{i}{2\pi}\big(-i\pi\textstyle\sum_j \varepsilon_j x_j\big) = \tfrac{1}{2}\big(\varepsilon_1 x_1 + \varepsilon_2 x_2 + \cdots + \varepsilon_m x_m\big). \tag{4.129}$$

The same normalisation applied to $TX_{\mathbb{C}}$ gives roots $\pm x_j$ (Section 4.9). So: *the spinor bundle has the "half" roots* $\frac{1}{2}(\pm x_1 \pm \cdots \pm x_m)$, the positive-chirality part E^+ taking the choices with $\prod_j \varepsilon_j = +1$.

Step 2: sum over the roots. Since ch of a bundle is the sum of e^{root} over its roots, and $\gamma_5 = \prod_j \varepsilon_j$ is $+1$ on E^+ and -1 on E^-,

$$\begin{aligned}\operatorname{ch}(E^+) - \operatorname{ch}(E^-) &\overset{(1)}{=} \sum_{\varepsilon\in\{\pm1\}^m}\Big(\prod_{j=1}^{m}\varepsilon_j\Big)\exp\Big(\tfrac{1}{2}\sum_{j=1}^{m}\varepsilon_j x_j\Big)\\ &\overset{(2)}{=} \sum_{\varepsilon_1=\pm1}\cdots\sum_{\varepsilon_m=\pm1}\prod_{j=1}^{m}\big(\varepsilon_j\, e^{\varepsilon_j x_j/2}\big)\\ &\overset{(3)}{=} \prod_{j=1}^{m}\Big(\sum_{\varepsilon_j=\pm1}\varepsilon_j\, e^{\varepsilon_j x_j/2}\Big) \;=\; \prod_{j=1}^{m}\big(e^{x_j/2} - e^{-x_j/2}\big).\end{aligned} \tag{Nash 4.107}$$

(1) Each root (4.129) contributes e^{root} to $\operatorname{ch}(E^+)$ if $\prod \varepsilon_j = +1$ and to $\operatorname{ch}(E^-)$ if $\prod \varepsilon_j = -1$; the factor $\prod \varepsilon_j$ supplies the relative minus sign.
(2) The exponential of a sum is a product of exponentials, and the sign $\prod \varepsilon_j$ is a product too, so the summand factorises over j.
(3) The signs ε_j are summed independently, so the sum of products is the product of sums (like $(a+b)(c+d) = ac + ad + bc + bd$ read backwards).

The lowest-order term of each factor is x_j, so the lowest-order term of the product is $x_1 \cdots x_m = e(X)$. The numerator is divisible by the Euler class, as the recipe needs (Section 4.11). This divisibility is ellipticity again: on the zero section of TX the symbol complex is not exact, and the non-exactness is measured by exactly one power of $e(X)$.

Worked example 4.84: Checking (4.107) in two and four dimensions

Two dimensions ($m = 1$). From Example 4.78, E^+ is a line bundle with root $+x/2$ and E^- has root $-x/2$. Directly, $\operatorname{ch} E^+ - \operatorname{ch} E^- = e^{x/2} - e^{-x/2} = x + \frac{x^3}{24} + \cdots$, and on a surface only the x survives: $\int_X (\operatorname{ch} E^+ - \operatorname{ch} E^-) = \int_X e(X) = \chi(X)$. For the round S^2, $E^\pm$ are line bundles of degree ± 1 ("square roots" of $T = \mathcal{O}(2)$ and of its dual), i.e. $E^+ \cong \mathcal{O}(1)$, $E^- \cong \mathcal{O}(-1)$, and $1 - (-1) = 2 = \chi(S^2)$.

Four dimensions ($m = 2$). The four sign choices give

$$E^+ : \ \pm\tfrac{1}{2}(x_1 + x_2), \qquad E^- : \ \pm\tfrac{1}{2}(x_1 - x_2).$$

This matches Example 4.79: E^+ is the spin-$\frac{1}{2}$ representation of $\mathrm{SU}(2)_+$, whose Cartan generator is the self-dual rotation $\gamma_1\gamma_2 + \gamma_3\gamma_4$ (angles $\theta_1 = \theta_2$), and E^- that of $\mathrm{SU}(2)_-$ ($\theta_1 = -\theta_2$). Then

$$\begin{aligned} \operatorname{ch} E^+ - \operatorname{ch} E^- &\overset{(1)}{=} 2\cosh\tfrac{x_1+x_2}{2} - 2\cosh\tfrac{x_1-x_2}{2} \\ &\overset{(2)}{=} 4\sinh\tfrac{x_1}{2}\sinh\tfrac{x_2}{2} = \left(e^{x_1/2} - e^{-x_1/2}\right)\left(e^{x_2/2} - e^{-x_2/2}\right). \end{aligned}$$

(1) $e^a + e^{-a} = 2\cosh a$ for each pair of opposite roots.

(2) The identity $\cosh(u+v) - \cosh(u-v) = 2\sinh u \sinh v$ with $u = x_1/2$, $v = x_2/2$; and $2\sinh u = e^u - e^{-u}$.

Expanding, $4\sinh\frac{x_1}{2}\sinh\frac{x_2}{2} = x_1x_2 + \frac{1}{24}x_1x_2(x_1^2 + x_2^2) + \cdots = e(X)\left(1 + \frac{p_1}{24} + \cdots\right)$. The bracket is $1/\hat{A}$ to this order; we will see in a moment that this is no accident.

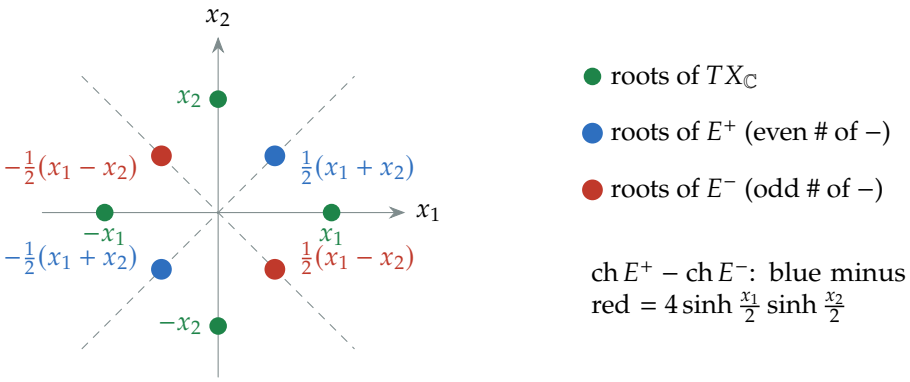


Figure 4.48: Chern roots in four dimensions, drawn as points in the (x_1, x_2) plane. The tangent bundle has the roots $\pm x_1, \pm x_2$ (green). The spinor roots sit at the half-way points $\frac{1}{2}(\pm x_1 \pm x_2)$: the positive-chirality ones on the diagonal (the self-dual direction), the negative ones on the anti-diagonal. Every spinor root is "half" a tangent root: this is the square root that forced us to introduce spin structures.

4.15.7 The index of the Dirac operator

Insert (Nash 4.107) into the index formula for complexes (4.101), with $e(X) = \prod_j x_j$ and $\operatorname{td}(TX_{\mathbb{C}}) = \prod_j \frac{x_j}{1-e^{-x_j}}\frac{-x_j}{1-e^{x_j}}$:

$$\operatorname{index} \not{D}^+ \overset{(1)}{=} (-1)^m \int_X \prod_{j=1}^{m} \frac{e^{x_j/2} - e^{-x_j/2}}{x_j} \cdot \frac{x_j}{1-e^{-x_j}} \cdot \frac{-x_j}{1-e^{x_j}}$$

$$\overset{(2)}{=} (-1)^m \int_X \prod_{j=1}^{m} \frac{e^{x_j/2} - e^{-x_j/2}}{x_j} \cdot \frac{x_j^2}{\left(e^{x_j/2} - e^{-x_j/2}\right)^2}$$

$$\overset{(3)}{=} (-1)^m \int_X \prod_{j=1}^{m} \frac{x_j}{e^{x_j/2} - e^{-x_j/2}}$$

$$\overset{(4)}{=} (-1)^m \int_X \prod_{j=1}^{m} \frac{x_j/2}{\sinh(x_j/2)}. \qquad \text{(Nash 4.108)}$$

(1) The recipe (4.101) with $n/2 = m$: numerator (Nash 4.107), one factor x_j of $e(X)$ per plane, and the Todd class of $TX_{\mathbb{C}}$ with roots $\pm x_j$.

(2) $(1 - e^{-x})(1 - e^{x}) = 2 - e^{x} - e^{-x} = -(e^{x/2} - e^{-x/2})^2$; with the numerator $-x^2$ the two minus signs cancel. (This is the step of (4.81).)

(3) Cancel one factor $e^{x_j/2} - e^{-x_j/2}$ and one factor x_j in each plane.

(4) $e^{x/2} - e^{-x/2} = 2\sinh(x/2)$; move the 2 under x_j.

Nash writes the two Todd factors in the other order; the product is the same.

There is a slicker way to see the same thing. Compare (Nash 4.107) with the definition of $\hat{A}$: per plane, $\frac{e^{x/2}-e^{-x/2}}{x} = \frac{\sinh(x/2)}{x/2}$, so

$$\operatorname{index} \not{D}^+ \overset{(1)}{=} (-1)^m \int_X \frac{\operatorname{ch} E^+ - \operatorname{ch} E^-}{e(X)} \operatorname{td}(TX_{\mathbb{C}})$$

$$\overset{(2)}{=} (-1)^m \int_X \frac{1}{\hat{A}(X)} \cdot \hat{A}(X)^2 = (-1)^m \int_X \hat{A}(X).$$

(1) The recipe.

(2) $\frac{\operatorname{ch} E^+ - \operatorname{ch} E^-}{e(X)} = \prod_j \frac{\sinh(x_j/2)}{x_j/2} = \hat{A}(X)^{-1}$, and $\operatorname{td}(TX_{\mathbb{C}}) = \hat{A}(X)^2$ by (4.81).

Physical picture: The Dirac operator is the square root of everything

The spinor roots are halves of the tangent roots, and correspondingly the spin complex is a "square root" of the de Rham complex: we will see in Section 4.16 that $\Lambda^* \otimes \mathbb{C} \cong E \otimes E$. At the level of characteristic classes, $\hat{A} = \sqrt{\operatorname{td}(TX_{\mathbb{C}})}$. The Dirac index density is the most basic of all the index densities; the others are obtained from it by twisting.

4.15.8 The $\hat{A}$-genus

Working definition 4.41 ($\hat{A}$-class and $\hat{A}$-genus (Nash (4.109)–(4.111)))

The **$\hat{A}$-class** ("A-roof" or "A-hat") of a Riemannian manifold is the multiplicative class [the product of $Q(x_j)$ over the x_j of the block form of $R/2\pi$] of the even power series [only even powers of x occur, so the result depends only on the x_j^2, i.e. on the p_i] $\frac{x/2}{\sinh(x/2)}$:

$$\hat{A}(X) = \prod_{j=1}^{n/2} \frac{x_j/2}{\sinh(x_j/2)} = \det{}^{1/2}\left(\frac{R/4\pi}{\sin(R/4\pi)}\right). \qquad \text{(Nash 4.109)}$$

The **$\hat{A}$-genus** is the number $\hat{A}(X) := \int_X \hat{A}(X)$ (Nash writes $\hat{A}[X]$). *To compute:* expand

$$\hat{A}(X) = 1 - \frac{1}{24}p_1 + \frac{1}{5760}(7p_1^2 - 4p_2) + \cdots \qquad \text{(Nash 4.110, corrected)}$$

and integrate the degree-n term.

Let us justify the curvature form of (Nash 4.109), which is what one actually integrates. The real antisymmetric matrix $R/4\pi$ has, in the block form, eigenvalues $\pm\frac{i}{2}x_j$ (each block $\left(\begin{smallmatrix}0 & x/2\\ -x/2 & 0\end{smallmatrix}\right)$ has eigenvalues $\pm ix/2$). Hence

$$\det{}^{1/2}\!\left(\frac{R/4\pi}{\sin(R/4\pi)}\right) \overset{(1)}{=} \Big[\prod_{j=1}^{m}\frac{ix_j/2}{\sin(ix_j/2)}\cdot\frac{-ix_j/2}{\sin(-ix_j/2)}\Big]^{1/2}$$
$$\overset{(2)}{=} \Big[\prod_{j=1}^{m}\Big(\frac{ix_j/2}{i\sinh(x_j/2)}\Big)^2\Big]^{1/2} = \prod_{j=1}^{m}\frac{x_j/2}{\sinh(x_j/2)}.$$

(1) The determinant of a function of a diagonalisable matrix is the product of the function over the eigenvalues; there are $2m$ eigenvalues $\pm ix_j/2$.
(2) $y/\sin y$ is even, so both members of a pair give the same factor; and $\sin(iu) = i\sinh u$.

The square root is needed because each plane contributes two eigenvalues but only one factor. Since $y/\sin y = 1 + \frac{y^2}{6} + \cdots$ is a power series in y^2, the determinant is a polynomial in $\operatorname{tr}(R^2), \operatorname{tr}(R^4), \ldots$, which is the Chern–Weil representative. Only even powers of R appear.

The expansion, by logarithms. The quickest route to (Nash 4.110, corrected) (the one Bär uses, [Bär §4.3]) goes through the logarithm, because the logarithm of a product over roots is a sum over roots, i.e. a symmetric power sum. First, with $t = x/2$,

$$\ln\frac{t}{\sinh t} \overset{(1)}{=} -\ln\Big(1+\frac{t^2}{6}+\frac{t^4}{120}+\cdots\Big)$$
$$\overset{(2)}{=} -\frac{t^2}{6}-\frac{t^4}{120}+\frac{1}{2}\Big(\frac{t^2}{6}\Big)^2+\cdots = -\frac{t^2}{6}+\frac{t^4}{180}-\cdots$$
$$\overset{(3)}{=} -\frac{x^2}{24}+\frac{x^4}{2880}-\cdots.$$

(1) $\sinh t/t = 1 + t^2/3! + t^4/5! + \cdots$.
(2) $\ln(1+u) = u - \frac{u^2}{2} + \cdots$ with $u = \frac{t^2}{6} + \frac{t^4}{120}$; and $-\frac{1}{120} + \frac{1}{72} = \frac{-3+5}{360} = \frac{1}{180}$.
(3) $t^2 = x^2/4$ and $t^4 = x^4/16$; $6\cdot 4 = 24$, $180\cdot 16 = 2880$.

Now sum over roots and exponentiate, using $\sum_j x_j^2 = p_1$ and $\sum_j x_j^4 = p_1^2 - 2p_2$:

$$\hat{A}(X) \overset{(1)}{=} \exp\Big(-\frac{1}{24}\sum_j x_j^2 + \frac{1}{2880}\sum_j x_j^4 - \cdots\Big)$$
$$\overset{(2)}{=} \exp\Big(-\frac{p_1}{24}+\frac{p_1^2-2p_2}{2880}-\cdots\Big)$$
$$\overset{(3)}{=} 1-\frac{p_1}{24}+\frac{p_1^2-2p_2}{2880}+\frac{1}{2}\Big(\frac{p_1}{24}\Big)^2+\cdots$$
$$\overset{(4)}{=} 1-\frac{p_1}{24}+\frac{2p_1^2-4p_2+5p_1^2}{5760}+\cdots = 1-\frac{p_1}{24}+\frac{7p_1^2-4p_2}{5760}+\cdots. \tag{4.130}$$

(1) $\hat{A} = \prod_j \frac{x_j/2}{\sinh(x_j/2)} = \exp\sum_j \ln\frac{x_j/2}{\sinh(x_j/2)}$.
(2) Newton's identity for the power sums in the variables x_j^2, whose elementary symmetric functions are $p_1, p_2, \ldots$.

(3) $e^u = 1 + u + \frac{1}{2}u^2 + \cdots$, keeping terms up to degree 8.
(4) Common denominator $5760 = 2 \cdot 2880 = 10 \cdot 576$: $\frac{1}{2880} = \frac{2}{5760}$ and $\frac{1}{2 \cdot 576} = \frac{5}{5760}$.

This agrees with the direct multiplication (4.83) of Section 4.9. The next term, for completeness, is $\hat{A}_3 = \frac{1}{967680}\left(-31p_1^3 + 44p_1p_2 - 16p_3\right)$.

Watch out: A misprint in Nash's (4.110)

Nash prints the degree-8 term as $\frac{1}{5760}(7p_1^2 - p_2)$. The correct coefficient of p_2 is -4, as the two independent derivations above show. The misprint matters: the quaternionic projective plane $\mathbb{H}P^2$ (a spin manifold, see Example 4.87) has $\int p_1^2 = 4$ and $\int p_2 = 7$, which gives $\frac{28-28}{5760} = 0$ with the correct formula, but the non-integer $\frac{28-7}{5760}$ with the misprinted one.

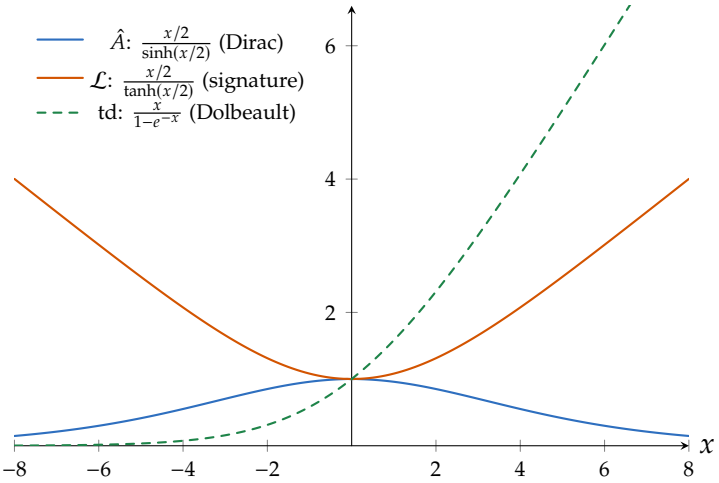


Figure 4.49: The generating functions of three genera, as ordinary functions of one variable. $\hat{A}$ and $\mathcal{L}$ are even (mirror-symmetric), so their classes involve only x_j^2, i.e. Pontryagin classes, and live in degrees $0, 4, 8, \ldots$: this is why the Dirac and signature indices can be nonzero only when $4 \mid n$. The Todd function is not even; its odd part $x/2$ is what lets c_1 enter the Riemann–Roch formula. All three equal 1 at $x = 0$, so each class starts with 1.

Why the index vanishes unless $4 \mid n$. The series $\frac{x/2}{\sinh(x/2)}$ is even, so $\hat{A}(X)$ is a polynomial in the x_j^2, i.e. in the Pontryagin classes, whose form degrees are multiples of 4 (Figure 4.49). If $n \equiv 2 \pmod 4$ the integrand has no component of degree n and the integral is zero. If $n \equiv 0 \pmod 4$ then $m = n/2$ is even and the sign $(-1)^m$ in (Nash 4.108) is $+1$. This gives Nash's final formula.

Key idea: The index theorem for the Dirac operator (Nash (4.111))

On a compact spin manifold X of even dimension n,

$$n^+ - n^- = \operatorname{index} \not{D}^+ = \begin{cases} \hat{A}(X) = \displaystyle\int_X \hat{A}(X), & n \equiv 0 \pmod 4, \\ 0, & n \equiv 2 \pmod 4. \end{cases} \qquad \text{(Nash 4.111)}$$

In particular the $\hat{A}$-genus of a spin manifold is an *integer*, although its definition has denominators $24, 5760, 967680, \ldots$

Dictionary: Low-dimensional forms of the Dirac index

n	index $\not{D}^+$	in terms of curvature / other invariants
2	0	(no 4-forms; $n^+ = n^-$ always)
4	$-\frac{1}{24}\int_X p_1$	$= \frac{1}{192\pi^2}\int_X \operatorname{tr}(R\wedge R) = -\frac{\operatorname{Sign}(X)}{8}$
6	0	
8	$\frac{1}{5760}\int_X (7p_1^2 - 4p_2)$	$= \frac{1}{896}\Big(\int_X p_1^2 - 4\operatorname{Sign}(X)\Big)$

The four-dimensional entry follows from $p_1 = -\frac{1}{8\pi^2}\operatorname{tr}(R\wedge R)$ (Section 4.9) and $\operatorname{Sign} = \frac{1}{3}\int p_1$ (Section 4.14):

$$\operatorname{index}\not{D}^+ \overset{(1)}{=} -\frac{1}{24}\int_X p_1 = -\frac{1}{24}\cdot\Big(-\frac{1}{8\pi^2}\Big)\int_X \operatorname{tr}(R\wedge R) = \frac{1}{192\pi^2}\int_X \operatorname{tr}(R\wedge R)$$
$$\overset{(2)}{=} -\frac{1}{24}\cdot 3\operatorname{Sign}(X) = -\frac{\operatorname{Sign}(X)}{8}.$$

(1) The degree-4 term of $\hat{A}$, and the Chern–Weil form of p_1.
(2) The signature theorem in dimension 4, $\int p_1 = 3\operatorname{Sign}(X)$.

For the eight-dimensional entry, eliminate p_2 with the signature theorem $\operatorname{Sign} = \frac{1}{45}\int(7p_2 - p_1^2)$, i.e. $7\int p_2 = 45\operatorname{Sign} + \int p_1^2$:

$$5760\,\hat{A}(X) \overset{(1)}{=} 7\int p_1^2 - \frac{4}{7}\Big(45\operatorname{Sign} + \int p_1^2\Big)$$
$$\overset{(2)}{=} \frac{45}{7}\int p_1^2 - \frac{180}{7}\operatorname{Sign} = \frac{45}{7}\Big(\int p_1^2 - 4\operatorname{Sign}\Big),$$

(1) Substitute $\int p_2 = \frac{1}{7}(45\operatorname{Sign} + \int p_1^2)$.
(2) $7 - \frac{4}{7} = \frac{45}{7}$.

and $5760\cdot\frac{7}{45} = 896$. So on an 8-dimensional spin manifold $\int p_1^2 \equiv 4\operatorname{Sign} \pmod{896}$: again a strong integrality constraint invisible from the definitions.

Worked example 4.85: $\mathbb{CP}^2$ is not spin, and the index theorem knows it

By Example 4.50, $\int_{\mathbb{CP}^2} p_1 = 3$. So the would-be index is

$$\int_{\mathbb{CP}^2}\hat{A} = -\frac{3}{24} = -\frac{1}{8} \notin \mathbb{Z}.$$

If $\mathbb{CP}^2$ had a spin structure, this number would be $n^+ - n^-$, an integer. So $\mathbb{CP}^2$ has no spin structure, in agreement with the direct computation $w_2(\mathbb{CP}^2) = 3h \bmod 2 \neq 0$ of Section 4.15.2. The $\hat{A}$-integrand is still a perfectly good closed 4-form on $\mathbb{CP}^2$; it is only its interpretation as an index that fails, because there is no Dirac operator to take the index of. (For the physicist: $\mathbb{CP}^2$ does admit a Dirac operator *coupled to a* U(1) *gauge field of half-integral flux*, a so-called spin^c structure. The half-integral flux cancels the half-integral gravitational contribution and the index becomes an integer; this is the Dolbeault operator in disguise, see Section 4.16.)

Worked example 4.86: Four-dimensional spin manifolds: S^4, T^4, $S^2 \times S^2$, K3

S^4. $p_1 = 0$ (Example 4.51), so index $\not{D}^+ = 0$. Stronger: the round metric has $R = 12 > 0$, so by Lichnerowicz there are no harmonic spinors at all, $n^+ = n^- = 0$.
T^4. Flat, so the integrand vanishes identically and the index is 0. With the flat metric and the periodic spin structure in all four directions the harmonic spinors are the constant ones, exactly as in Example 4.83: $n^+ = n^- = 2$ (a basis of $\mathbb{C}^2 \oplus \mathbb{C}^2$).
$S^2 \times S^2$. Spin (each factor is), $p_1 = 0$, index 0; with the product of round metrics $R = 2+2 > 0$, so again no harmonic spinors.
K3. This is the interesting one. It is spin: it is a complex surface with $c_1 = 0$, so $w_2 = c_1 \bmod 2 = 0$. From Example 4.74, $\int p_1 = -48$ and Sign $= -16$, hence

$$\begin{aligned} \operatorname{index} \not{D}^+ &\overset{(1)}{=} -\frac{1}{24}\int_{\mathrm{K3}} p_1 = -\frac{-48}{24} = 2 \\ &\overset{(2)}{=} -\frac{\operatorname{Sign}(\mathrm{K3})}{8} = -\frac{-16}{8} = 2. \end{aligned}$$

(1) The four-dimensional $\hat{A}$-genus.
(2) Cross-check through the signature.

So every metric on K3 has at least two positive-chirality harmonic spinors. For the Ricci-flat (Calabi–Yau) metric we can see them: the holonomy of this metric is SU(2), sitting inside $\mathrm{Spin}(4) = \mathrm{SU}(2)_+ \times \mathrm{SU}(2)_-$ as one of the two factors, and it acts trivially on the other half-spin bundle. That half is then flat, and a basis of it at one point extends to two covariantly constant spinors: these are the two zero modes. (Lichnerowicz with $R = 0$ says every harmonic spinor is parallel, so there are no others.) This is why K3 compactifications preserve half of the supersymmetry: covariantly constant spinors are unbroken supersymmetry parameters.
A consequence: since $\hat{A}(\mathrm{K3}) \neq 0$, K3 admits *no* metric of positive scalar curvature (Section 4.15.5). The index theorem has turned a topological number into a statement about all possible geometries.

Worked example 4.87: Eight dimensions: $\mathbb{H}P^2$ and $\mathrm{K3} \times \mathrm{K3}$

$\mathbb{H}P^2$. It is simply connected with $H^2 = 0$, so $w_2 = 0$ and it is spin. With the data of Example 4.75 ($p_1 = 2u$, $p_2 = 7u^2$, $\int u^2 = 1$):

$$\hat{A}(\mathbb{H}P^2) = \frac{7 \cdot 4 - 4 \cdot 7}{5760} = 0.$$

Consistent with Lichnerowicz: the standard metric of $\mathbb{H}P^2$ has positive scalar curvature. The alternative formula gives $\frac{1}{896}(4 - 4 \cdot 1) = 0$ too.
$\mathrm{K3} \times \mathrm{K3}$. Let a and b be the pull-backs of $p_1(\mathrm{K3})$ from the two factors, $\int_{\mathrm{K3}} p_1 = -48$. The Pontryagin class is multiplicative (up to 2-torsion, invisible here): $p = (1 + a)(1 + b)$, so $p_1 = a + b$ and $p_2 = ab$, with $a^2 = b^2 = 0$ (8-forms on a 4-manifold vanish) and $\int ab = (-48)^2 = 2304$. Then

$$\begin{aligned} \hat{A}(\mathrm{K3} \times \mathrm{K3}) &\overset{(1)}{=} \frac{1}{5760}\int \big(7(a+b)^2 - 4ab\big) = \frac{1}{5760}\int \big(14ab - 4ab\big) \\ &\overset{(2)}{=} \frac{10 \cdot 2304}{5760} = 4 = \hat{A}(\mathrm{K3})^2. \end{aligned}$$

(1) The degree-8 term of $\hat{A}$; $(a+b)^2 = 2ab$.

(2) $\int ab = 2304$, and $23040/5760 = 4$.

The answer $2 \cdot 2$ is what multiplicativity of $\hat{A}$ predicts ($\hat{A}(X \times Y) = \hat{A}(X)\hat{A}(Y)$, since the Chern roots of $T(X \times Y)$ are the union of those of X and of Y). Both checks fail with Nash's misprinted coefficient.

Try it: $\hat{A}$ of products and a divisibility

(a) Show that $\hat{A}(X \times Y) = \hat{A}(X)\hat{A}(Y)$, and deduce that $\hat{A}(\mathrm{K3} \times T^4) = 0$. (b) Using Rokhlin's theorem (next subsection), show that $\int_X p_1$ is divisible by 48 on every closed smooth spin 4-manifold.

Answer. (a) Roots of a product are the union of the roots, and $\hat{A}$ is multiplicative; $\hat{A}(T^4) = 0$. (b) $\int p_1 = 3\,\mathrm{Sign}$ and $16 \mid \mathrm{Sign}$.

4.15.9 Rokhlin's theorem

Here is a striking consequence of the integrality in (Nash 4.111). The question is:

Which integers occur as the signature of a closed smooth spin 4-manifold?

The answer, due to Rokhlin (1952), long before the index theorem, is: only multiples of 16. We give the index-theory proof, which has three steps. The first is algebra, the second is the index theorem, the third is a symmetry of four-dimensional spinors.

Step 0: spin means even intersection form. Let X be a closed simply connected 4-manifold with intersection form q on $H^2(X;\mathbb{Z})$ (Section 4.14). The **Wu formula** says that for every $x \in H^2(X;\mathbb{Z})$

$$q(x,x) \equiv \int_X w_2(X) \cup x \pmod 2. \tag{4.131}$$

So $w_2 = 0$ exactly when $q(x,x)$ is even for every x, i.e. when the intersection form is **even**. We can check this on the examples of Section 4.14: $S^2 \times S^2$ has $q = H = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, even, and is spin; $\mathbb{CP}^2$ has $q = (1)$, odd, and is not spin; K3 has $q = 3H \oplus 2(-E_8)$, even, and is spin. (Why the Wu formula holds, in one sentence for complex surfaces: $w_2 = c_1 \bmod 2$, and the adjunction formula for a curve C representing x says $x \cdot x - c_1 \cdot x = 2g(C) - 2$, which is even, so $x \cdot x \equiv c_1 \cdot x \pmod 2$.)

Step 1: the index theorem gives $8 \mid \mathrm{Sign}$. For a spin 4-manifold index $\not{D}^+ = -\mathrm{Sign}(X)/8$ is an integer, so $\mathrm{Sign}(X) \equiv 0 \pmod 8$. Honesty requires a remark: this step alone is not news. A purely algebraic theorem (Serre) says that every even unimodular symmetric form has signature divisible by 8; the form E_8 has signature exactly 8. The index theorem re-proves this for forms that come from spin manifolds. The new information is in the next step.

Step 2: four-dimensional spinors are quaternionic. We show that $n^+ = \dim \mathrm{Ker}\, \not{D}^+$ is *even*. The reason is an antilinear symmetry. Take the gamma matrices of Example 4.79, built from the real quaternion matrices s_a, $\bar{s}_a$. Let K denote complex conjugation of components and $\varepsilon = i\sigma_2 = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$.

Working definition 4.42 (Quaternionic structure on four-dimensional spinors)

Define the antilinear map [$J(c\psi) = \bar{c}\, J\psi$ for every complex number c]

$$J = \begin{pmatrix} \varepsilon K & 0 \\ 0 & \varepsilon K \end{pmatrix} : \mathbb{C}^2 \oplus \mathbb{C}^2 \to \mathbb{C}^2 \oplus \mathbb{C}^2, \qquad J \begin{pmatrix} z_1 \\ z_2 \\ w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} \bar{z}_2 \\ -\bar{z}_1 \\ \bar{w}_2 \\ -\bar{w}_1 \end{pmatrix}.$$

It preserves chirality, satisfies $J^2 = -1$ [so $J\psi$ is never a multiple of ψ when $\psi \neq 0$, and zero modes come in independent pairs $\psi, J\psi$], and commutes with every γ_a, with every $\gamma_a\gamma_b$, and hence with $\not{D}$.

Let us verify these claims. First, a real quaternion matrix $q = q^0 I + q^1(i\sigma_1) + q^2(i\sigma_2) + q^3(i\sigma_3)$ ($q^\alpha \in \mathbb{R}$) commutes with εK. On a vector v:

$$\varepsilon K(qv) \overset{(1)}{=} \varepsilon\, \bar{q}\, \bar{v} = \left(\varepsilon \bar{q} \varepsilon^{-1}\right) \varepsilon \bar{v} \overset{(2)}{=} q\, \varepsilon \bar{v} = q\,(\varepsilon K v).$$

(1) Complex conjugation of a product is the product of the conjugates ($\bar{q}$ is the entrywise conjugate); insert $\varepsilon^{-1}\varepsilon = 1$.

(2) $\varepsilon \bar{q} \varepsilon^{-1} = q$. Check on each basis element: $\bar{I} = I$; $\overline{i\sigma_1} = -i\sigma_1$ and $\sigma_2\sigma_1\sigma_2 = -\sigma_1$, so $\varepsilon(-i\sigma_1)\varepsilon^{-1} = -i\sigma_2\sigma_1\sigma_2 = i\sigma_1$ (using $\varepsilon = i\sigma_2$, $\varepsilon^{-1} = -i\sigma_2$); likewise for $i\sigma_3$; and $\overline{i\sigma_2} = i\sigma_2$ commutes with ε.

The blocks of every γ_a are $\pm s_a$, $\pm \bar{s}_a$, real quaternion matrices, and J acts by the same εK on both chiral halves, so $J\gamma_a = \gamma_a J$; products of gammas then commute with J as well. Second,

$$J^2 v \overset{(1)}{=} \varepsilon\, \overline{\varepsilon \bar{v}} = \varepsilon\, \bar{\varepsilon}\, v \overset{(2)}{=} \varepsilon^2 v = -v.$$

(1) Apply εK twice; $\bar{\bar{v}} = v$.

(2) ε is real, and $\varepsilon^2 = (i\sigma_2)^2 = -I$.

Finally $\not{D} = \gamma^a e_a^\mu(\partial_\mu + \omega_\mu)$: the vielbein e_a^μ is real, ∂_μ commutes with complex conjugation, and ω_μ is a real combination of the $\gamma_a\gamma_b$. So $J\not{D} = \not{D}J$ and J maps $\operatorname{Ker} \not{D}^+$ into itself.

Why J forces even dimension. Let $V = \operatorname{Ker} \not{D}^+$ (finite dimensional). If $0 \neq \psi \in V$, then $J\psi \in V$ and $J\psi$ is not a complex multiple of ψ: if $J\psi = \lambda\psi$ then

$$-\psi \overset{(1)}{=} J^2\psi = J(\lambda\psi) \overset{(2)}{=} \bar{\lambda} J\psi = \bar{\lambda}\lambda\psi = |\lambda|^2\, \psi,$$

(1) $J^2 = -1$, and the assumption.

(2) J is antilinear, so it conjugates scalars.

which is impossible since $|\lambda|^2 \geq 0$. So $\{\psi, J\psi\}$ spans a 2-dimensional subspace, invariant under J (as $J(J\psi) = -\psi$). J also preserves the L^2 inner product up to conjugation, so the orthogonal complement of this subspace in V is again J-invariant, and we repeat. The dimension of V is $2 + 2 + \cdots$: even. In the language of algebra, V is a vector space over the quaternions $\mathbb{H}$, with J

as right multiplication by the quaternion $\mathbf{j}$. The same holds for $\operatorname{Ker} \not{D}^-$. Hence n^+ and n^- are both even, and

$$\hat{A}(X) = n^+ - n^- \in 2\mathbb{Z} \qquad (\dim X = 4). \tag{4.132}$$

Key idea: Rokhlin's theorem

If X is a closed smooth spin 4-manifold then

$$\operatorname{Sign}(X) = -8\,\hat{A}(X) \equiv 0 \pmod{16}. \tag{4.133}$$

Mechanism: the index theorem ($\operatorname{Sign} = -8\hat{A}$) plus the quaternionic structure of four-dimensional spinors ($\hat{A}$ even).

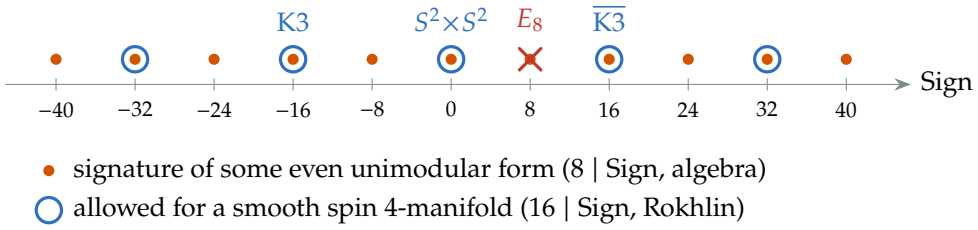


Figure 4.50: Signatures of spin 4-manifolds. Algebra alone allows every multiple of 8 (orange dots) for an even form. Smoothness and the index theorem cut this down to multiples of 16 (blue circles). The form E_8, of signature 8, is even but is not the intersection form of any smooth spin 4-manifold (red cross); K3 and K3 with reversed orientation realise ∓ 16.

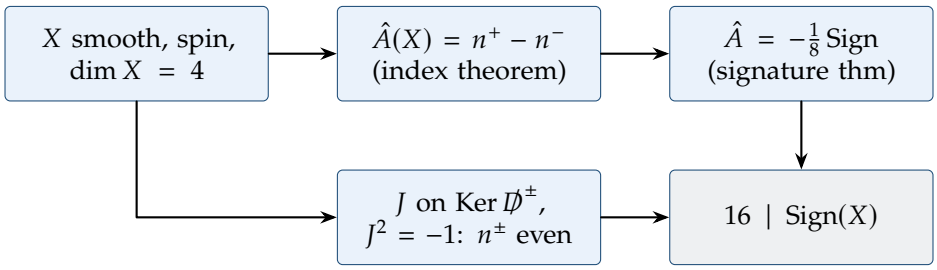


Figure 4.51: The logic of Rokhlin's theorem. Two independent inputs meet: the index theorem expresses the signature through the number of spinor zero modes, and the quaternionic structure of four-dimensional spinors says that zero modes come in pairs.

Worked example 4.88: What Rokhlin rules out

The E_8 form. E_8 is the 8×8 Cartan matrix of the Lie algebra $\mathfrak{e}_8$; it is even, unimodular and positive definite, so its signature is 8. A simply connected smooth 4-manifold with intersection form E_8 would be spin (even form, Step 0) with signature 8, contradicting (4.133). Such a manifold does not exist. (Freedman showed that a *topological* 4-manifold with form E_8 does exist; so it admits no smooth structure at all. This was an early sign of how strange smooth 4-manifold topology is, later made systematic by Donaldson's instanton methods of Chapter 8.)

K3 is optimal. $\operatorname{Sign}(\text{K3}) = -16$ and $\hat{A}(\text{K3}) = 2$: the smallest nonzero values allowed. Its form $3H \oplus 2(-E_8)$ contains two copies of $-E_8$; neither can be "split off" into a smooth spin manifold of its own.

Non-spin examples are unconstrained. $\mathbb{CP}^2$ has signature 1. Rokhlin does not apply: the form (1) is odd.

Watch out: Where the factor 2 comes from, and in which dimensions

The quaternionic structure exists because the spin representations in dimension 4 are pseudo-real: SU(2) matrices commute with εK. In dimension $8k+4$ the half-spin representations are again quaternionic and $\hat{A}$ is even; in dimension $8k$ they are real and no such doubling occurs (the Bott manifold in dimension 8 has $\hat{A}=1$). The mod-2 phenomena in dimensions $8k+1$, $8k+2$ are the mod-2 index of the next part of this chapter. Physicists meet the same pseudo-reality in Witten's SU(2) anomaly: a single Weyl doublet of SU(2) has its zero modes in pairs, and an odd number of doublets makes the theory inconsistent.

Physical picture: Rokhlin as a statement about fermions

Think of X as a Euclidean spacetime, a "gravitational instanton". The index n^+-n^- is the net number of fermion zero modes in this background, and it controls the violation of fermion number (chiral charge) by $\Delta Q_5 = 2\,\mathrm{index}\,\not{D}^+$ for one Dirac fermion (Chapter 10). Rokhlin's theorem says: in a smooth spin spacetime the gravitational chiral anomaly always comes in units of 2 zero modes per Weyl fermion, and the signature of spacetime is -8 times it.

Reading Nash (pp. 113–116, (4.96)–(4.111))

Nash introduces Spin(n), spin structures and w_1, w_2 (4.96); states that spin structures are parametrised by $H^1(X;\mathbb{Z}_2)$ (more precisely, as we noted, a torsor over it) and that $\mathbb{CP}^n$ is spin iff n is odd; describes the spin representation and its chiral halves (4.97) and the spinor bundles (4.98); gives the Clifford relations and γ_5 (4.99) (with the mathematicians' sign $\{\gamma_a,\gamma_b\}=-2\delta_{ab}$), the Dirac operator (4.100), the block form of γ_5 (4.101) and the spin complex (4.102). He then notes that the symbol $i\gamma^a e_a^\mu p_\mu$ is invertible only in Euclidean signature, defines harmonic spinors and $n^\pm$ (4.103)–(4.106), quotes the Chern character (4.107) and derives (4.108), defines $\hat{A}$ (4.109)–(4.110) and states the theorem (4.111). We have supplied: explicit Spin(2) and gamma matrices, the construction of spin structures from transition functions, the lift of the connection and explicit Dirac operators on $\mathbb{R}^2$, S^2 and T^2, Lichnerowicz, the derivation of (4.107) from the weights, two derivations of (4.110) (which has a misprint: read $-4p_2$ for $-p_2$), and the applications to $\mathbb{CP}^2$, K3, $\mathbb{H}P^2$ and Rokhlin's theorem, which Nash does not discuss in this section.

Recap

- Spinors need a spin structure: consistent lifts of the frame transition functions to Spin(n); obstruction $w_2(X)$; choices form a torsor over $H^1(X;\mathbb{Z}_2)$. $\mathbb{CP}^n$ is spin iff n is odd.
- Nash's gammas: $\{\gamma_a,\gamma_b\}=-2\delta_{ab}$, $\gamma_5 = i^{n(n+1)/2}\gamma_1\cdots\gamma_n = \prod_j(-i\gamma_{2j-1}\gamma_{2j})$; $S = S^+\oplus S^-$ of dimensions $2^{n/2-1}$ each.
- $\not{D} = \gamma^a e_a^\mu(\partial_\mu+\omega_\mu)$, with $\omega = -\frac{1}{4}\omega^a{}_b\gamma_a\gamma_b$; $\not{D}^2 = \nabla^*\nabla + R/4$; elliptic only in Euclidean signature.
- $\mathrm{index}\,\not{D}^+ = n^+-n^-$ (harmonic spinors of each chirality).
- Spinor Chern roots $\frac{1}{2}(\pm x_1\pm\cdots\pm x_m)$ give $\mathrm{ch}\,E^+-\mathrm{ch}\,E^- = \prod(e^{x_j/2}-e^{-x_j/2})$, and $\mathrm{index}\,\not{D}^+ = \int\hat{A}$ for $4\mid n$, zero otherwise.
- $\hat{A} = 1-\frac{p_1}{24}+\frac{7p_1^2-4p_2}{5760}$; in 4D $\mathrm{index}\,\not{D}^+ = \frac{1}{192\pi^2}\int \mathrm{tr}\,R\wedge R = -\,\mathrm{Sign}/8$.

- Examples: $\mathbb{CP}^2$: $-\frac{1}{8}$ (not spin); K3: 2 (no psc metric); $\mathbb{H}P^2$: 0. Rokhlin: smooth spin 4-manifolds have 16 | Sign.

4.16 Coupling to gauge fields: counting fermion zero modes

So far the Dirac operator has felt only gravity, through the spin connection. In physics a fermion also carries charge: it is a section of $E \otimes F$, where F is a vector bundle carrying a gauge field. An electron in a magnetic field, a quark in an instanton background. The question of this section is the physicist's version of the index problem:

Given a gauge field configuration, how many zero modes does a massless charged fermion have, and of which chirality?

The answer controls some of the most important non-perturbative effects in field theory: the degeneracy of the lowest Landau level, the 't Hooft vertex and the breaking of chiral symmetry by instantons, and the dimension of the instanton moduli space. Nash treats the general theory of "twisted" complexes in his next section (§3, which is the subject of the next part of this chapter). Here we derive the one formula we need, count zero modes by hand in two and four dimensions, and then use the same formula to show that the four classical complexes of this section are all Dirac operators in disguise.

4.16.1 The Dirac operator with a gauge field

Conventions. A gauge field is a connection on a Hermitian vector bundle $F \to X$ of rank r, with structure group $\mathrm{U}(r)$ or a subgroup such as $\mathrm{SU}(2)$. In a local trivialisation it is a matrix of 1-forms $A = A_\mu \mathrm{d}x^\mu$ with *anti-Hermitian* values, and its curvature is $F_A = \mathrm{d}A + A \wedge A$ (we write F_A for the curvature to keep F for the bundle). Under a gauge transformation $A \mapsto g^{-1}Ag + g^{-1}\mathrm{d}g$. The Chern character is $\mathrm{ch}(F) = \operatorname{tr} \exp(\frac{i}{2\pi}F_A)$, so that

$$\mathrm{ch}(F) = r + c_1(F) + \tfrac{1}{2}\big(c_1^2 - 2c_2\big)(F) + \cdots, \qquad c_1 = \frac{i}{2\pi}\operatorname{tr} F_A, \quad \mathrm{ch}_2(F) = -\frac{1}{8\pi^2}\operatorname{tr}(F_A \wedge F_A). \tag{4.134}$$

The last expression is $\frac{1}{2}\big(\frac{i}{2\pi}\big)^2 \operatorname{tr}(F_A \wedge F_A)$. For an $\mathrm{SU}(r)$ bundle $c_1 = 0$ and $\mathrm{ch}_2 = -c_2$; we define the **instanton number** as $k := c_2(F)[X] = \frac{1}{8\pi^2}\int_X \operatorname{tr}(F_A \wedge F_A)$.

Dictionary: Gauge fields, mathematicians vs physicists

here (Nash)	physics textbooks
A_μ anti-Hermitian	$A_\mu = -ig\, A_\mu^a T^a$, T^a Hermitian, $\operatorname{tr} T^a T^b = \frac{1}{2}\delta^{ab}$
$\nabla_\mu = \partial_\mu + A_\mu$	$D_\mu = \partial_\mu - igA_\mu^a T^a$
$U(1)$: $A = -i\,a$, $F_A = -i\,\mathrm{d}a$	a_μ real (charge absorbed), $\mathrm{d}a = B\,\mathrm{d}x \wedge \mathrm{d}y$ in 2D
$k = c_2 = \frac{1}{8\pi^2}\int \operatorname{tr} F_A \wedge F_A$	$k_{\text{phys}} = \frac{g^2}{32\pi^2}\int F^a_{\mu\nu}\tilde{F}^{a\mu\nu}\,\mathrm{d}^4x = -k$
anti-self-dual ($\star F_A = -F_A$): $k > 0$	"anti-instanton" in many physics texts

(The last two rows use $\operatorname{tr}(F_A \wedge F_A) = -\frac{g^2}{2}F^a \wedge F^a$ and $F^a \wedge F^a = \frac{1}{2}F^a_{\mu\nu}\tilde{F}^{a\mu\nu}\mathrm{d}^4x$. Signs of k are orientation conventions; the invariant statements below are phrased so that they do not depend on them.)

Working definition 4.43 (Twisted Dirac operator)

On a spin manifold with a gauge field A on F [a connection on the colour bundle F, a matrix of 1-forms A_μ acting on the colour index], the **twisted** (or **coupled**) Dirac operator acts on sections of $E \otimes F$ (spinor index $\otimes$ colour index):

$$\not{D}_A = \gamma^a e_a^\mu \big(\partial_\mu + \omega_\mu \otimes 1 + 1 \otimes A_\mu\big), \qquad \not{D}_A^+ : \Gamma(E^+ \otimes F) \to \Gamma(E^- \otimes F). \tag{4.135}$$

Everything said about $\not{D}$ goes through: γ_5 still anticommutes, the symbol is still $i\not{p} \otimes 1$ [$\partial_\mu \to i p_\mu$ gives $i\gamma^a p_a$] (the gauge field is lower order [A_μ comes with no derivative, so it drops out of the leading symbol]), and index $\not{D}_A^+ = n^+ - n^-$ counts charged zero modes by chirality.

The index density. The symbol is unchanged, but the bundles are: E^p is replaced by $E^p \otimes F$. The curvature of $E^\pm \otimes F$ is $\rho(R) \otimes 1 + 1 \otimes F_A$; the two terms commute, so its Chern roots are the sums (spinor root) + (root y_k of F). Therefore

$$\begin{aligned} \operatorname{ch}(E^+ \otimes F) - \operatorname{ch}(E^- \otimes F) &\overset{(1)}{=} \sum_{\varepsilon} \Big(\prod_j \varepsilon_j\Big) \sum_{k=1}^{r} \exp\Big(\tfrac{1}{2}\sum_j \varepsilon_j x_j + y_k\Big) \\ &\overset{(2)}{=} \Big[\sum_{\varepsilon} \Big(\prod_j \varepsilon_j\Big) e^{\frac{1}{2}\sum_j \varepsilon_j x_j}\Big]\Big[\sum_k e^{y_k}\Big] = \big(\operatorname{ch} E^+ - \operatorname{ch} E^-\big)\operatorname{ch}(F), \end{aligned} \tag{4.136}$$

(1) Chern roots of a tensor product are the pairwise sums of roots.
(2) $e^{a+b} = e^a e^b$, and the double sum factorises.

and running the recipe exactly as in (Nash 4.108), the extra factor $\operatorname{ch}(F)$ rides along:

Key idea: Index of the twisted Dirac operator (Nash (4.122), next part)

$$\operatorname{index} \not{D}_A^+ = (-1)^{n/2} \int_X \hat{A}(X)\, \operatorname{ch}(F). \tag{4.137}$$

Unlike the untwisted case, this is generally nonzero in every even dimension, because $\operatorname{ch}(F)$ has components in all even degrees.

Explicitly, in the two cases that matter most:

$$n = 2: \qquad \operatorname{index} \not{D}_A^+ = -\int_X c_1(F) = -\frac{i}{2\pi}\int_X \operatorname{tr} F_A, \tag{4.138}$$

$$n = 4: \qquad \operatorname{index} \not{D}_A^+ = -\frac{r}{24}\int_X p_1 + \int_X \operatorname{ch}_2(F) = \frac{r}{192\pi^2}\int_X \operatorname{tr}(R \wedge R) - \frac{1}{8\pi^2}\int_X \operatorname{tr}(F_A \wedge F_A). \tag{4.139}$$

For (4.138): $\hat{A} = 1$ on a surface and the 2-form part of $\operatorname{ch} F$ is c_1; the sign is $(-1)^1$. For (4.139): the 4-form part of $\big(1 - \frac{p_1}{24}\big)\big(r + c_1 + \operatorname{ch}_2\big)$ is $-\frac{r}{24}p_1 + \operatorname{ch}_2$, and the sign is $(-1)^2 = +1$. The gravitational term is r times the untwisted index; the gauge term is the "instanton contribution". The pure gauge term $-\frac{1}{8\pi^2}\int \operatorname{tr} F_A \wedge F_A$ is, up to normalisation, the integrated chiral anomaly $\int \partial_\mu J_5^\mu$ (Chapter 10).

Watch out: Index zero untwisted, nonzero twisted

On S^2 the untwisted index vanishes (as it does on every surface), but with a magnetic monopole it is $-d$ for monopole charge d (below). Nash stresses this in §3: a manifold on

which index $\not{D} = 0$ can still have index $\not{D}_A \neq 0$. The index of the twisted operator measures the topology of the gauge field, not only of the manifold.

4.16.2 Two dimensions: monopoles and Landau levels

Take a U(1) gauge field, $A = -i\,a$ with a a real 1-form. Then $F_A = -i\,\mathrm{d}a$, $c_1 = \frac{1}{2\pi}\mathrm{d}a$, and on a surface $\int c_1 = \Phi/2\pi$ with Φ the total magnetic flux. The index theorem (4.138) says $n^+ - n^- = -\Phi/2\pi$ on a compact surface, where $\Phi/2\pi = d \in \mathbb{Z}$ is the monopole charge (the degree of the line bundle). We now find the zero modes explicitly, first on the plane, then on the sphere.

The Dirac operator with a magnetic field. With the gamma matrices of Example 4.78 and $\partial_\mu \to \partial_\mu - i a_\mu$, the blocks of Example 4.80 become

$$\not{D}_A^+ = -2\big(\partial_z - i a_z\big), \qquad \not{D}_A^- = 2\big(\partial_{\bar z} - i a_{\bar z}\big), \qquad a_z = \tfrac{1}{2}(a_1 - i a_2), \quad a_{\bar z} = \tfrac{1}{2}(a_1 + i a_2).$$

Choose the Coulomb gauge $a_1 = -\partial_2\phi$, $a_2 = \partial_1\phi$, so that $B = \partial_1 a_2 - \partial_2 a_1 = \nabla^2\phi$. Then

$$\begin{aligned} a_{\bar z} &\overset{(1)}{=} \tfrac{1}{2}\big(-\partial_2\phi + i\partial_1\phi\big) = \tfrac{i}{2}\big(\partial_1\phi + i\partial_2\phi\big) \\ &\overset{(2)}{=} i\,\partial_{\bar z}\phi, \qquad \text{and similarly} \qquad a_z = -i\,\partial_z\phi. \end{aligned}$$

(1) Insert the gauge choice; factor out i (note $i\cdot i\partial_2\phi = -\partial_2\phi$).
(2) $\partial_{\bar z} = \tfrac{1}{2}(\partial_1 + i\partial_2)$.

Hence $\partial_{\bar z} - i a_{\bar z} = \partial_{\bar z} + \partial_{\bar z}\phi = e^{-\phi}\partial_{\bar z} e^{\phi}$, and $\partial_z - i a_z = \partial_z - \partial_z\phi = e^{\phi}\partial_z e^{-\phi}$. The zero-mode equations are solved by "gauging away" the field:

$$\not{D}_A^- \psi^- = 0 \iff \psi^- = e^{-\phi} f(z), \qquad \not{D}_A^+ \psi^+ = 0 \iff \psi^+ = e^{\phi} g(\bar z), \tag{4.140}$$

with f holomorphic and g antiholomorphic. Which of these are acceptable depends on the behaviour of ϕ at infinity.

Worked example 4.89: Aharonov–Casher: zero modes on the plane

Let B be localised with total flux $\Phi > 0$. The solution of $\nabla^2\phi = B$ is $\phi(x) = \frac{1}{2\pi}\int \ln|x - x'|\, B(x')\,\mathrm{d}^2x'$, so $\phi \approx \frac{\Phi}{2\pi}\ln r$ for large r. Then $\psi^+ = e^{\phi} g$ grows like $r^{\Phi/2\pi}\,|g|$ and is never normalisable: $n^+ = 0$. For $\psi^- = e^{-\phi} z^k$ (f must be a polynomial, or the growth of an entire function beats any power):

$$\begin{aligned} \int |\psi^-|^2\,\mathrm{d}^2x &\overset{(1)}{=} \int^{\infty} r^{2k}\, r^{-\Phi/\pi}\, r\,\mathrm{d}r \cdot (\text{finite}) \\ &\overset{(2)}{=} \text{finite} \iff 2k - \frac{\Phi}{\pi} + 1 < -1 \iff k < \frac{\Phi}{2\pi} - 1. \end{aligned}$$

(1) $\big|z^k\big|^2 = r^{2k}$, $e^{-2\phi} \approx r^{-\Phi/\pi}$, and $\mathrm{d}^2x = r\,\mathrm{d}r\,\mathrm{d}\theta$; the region of finite r gives a finite contribution.
(2) Convergence of $\int^{\infty} r^s \mathrm{d}r$ requires $s < -1$.

So the zero modes are $e^{-\phi} z^k$ with $0 \le k < \frac{\Phi}{2\pi} - 1$:

$$n^- = \left\lceil \frac{\Phi}{2\pi} \right\rceil - 1, \qquad n^+ = 0 \qquad (\Phi > 0),$$

all of one chirality (for $\Phi < 0$ the roles of $\psi^{\pm}$ swap). This is the Aharonov–Casher theorem (1979). Notice: *not* $\Phi/2\pi$. For $\Phi = 2\pi d$ exactly, the would-be d-th mode $z^{d-1}e^{-\phi} \sim r^{-1}$ is only logarithmically divergent. The plane is not compact, so the index theorem does not apply as it stands; the deficit is a boundary effect at infinity, of the Atiyah–Patodi–Singer type treated in the next part.

The sphere, by a conformal map. The round sphere is conformally flat: in the stereographic coordinate z the metric is $g = \rho^2 |\mathrm{d}z|^2$ with $\rho = \frac{2}{1+|z|^2}$. The Dirac operator transforms simply under $g = \rho^2\delta$. Take the coframe $e^a = \rho\,\mathrm{d}x^a$ and write $\sigma = \ln\rho$. Cartan's equations give $\omega^1{}_2 = \partial_2\sigma\,\mathrm{d}x - \partial_1\sigma\,\mathrm{d}y$, hence $\omega_\mu = -\frac{1}{2}\omega_\mu{}^1{}_2\,\gamma_1\gamma_2$, and

$$\begin{aligned} \not{D}_g &\overset{(1)}{=} \rho^{-1}\gamma^a\big(\partial_a + \omega_a + A_a\big) \\ &\overset{(2)}{=} \rho^{-1}\Big(\gamma^a(\partial_a + A_a) - \tfrac{1}{2}\partial_2\sigma\,\gamma_1\gamma_1\gamma_2 + \tfrac{1}{2}\partial_1\sigma\,\gamma_2\gamma_1\gamma_2\Big) \\ &\overset{(3)}{=} \rho^{-1}\Big(\gamma^a(\partial_a + A_a) + \tfrac{1}{2}\gamma^a\partial_a\sigma\Big) \\ &\overset{(4)}{=} \rho^{-3/2}\,\not{D}_{\text{flat}}\,\rho^{1/2}. \end{aligned} \tag{4.141}$$

(1) The definition (4.135) with $e_a = \rho^{-1}\partial_a$.
(2) $\omega_1 = -\frac{1}{2}\partial_2\sigma\,\gamma_1\gamma_2$ and $\omega_2 = +\frac{1}{2}\partial_1\sigma\,\gamma_1\gamma_2$. (Cartan: $\mathrm{d}e^1 = \partial_2\rho\,\mathrm{d}y\wedge\mathrm{d}x$ must equal $-\omega^1{}_2\wedge\rho\,\mathrm{d}y$, and similarly for e^2.)
(3) $\gamma_1\gamma_1\gamma_2 = -\gamma_2$ and $\gamma_2\gamma_1\gamma_2 = \gamma_1$.
(4) $\gamma^a\partial_a(\rho^{1/2}\psi) = \rho^{1/2}\big(\gamma^a\partial_a\psi + \frac{1}{2}\gamma^a\partial_a\sigma\,\psi\big)$, so the bracket is $\rho^{-1/2}\not{D}_{\text{flat}}\rho^{1/2}$.

So $\not{D}_g\psi = 0$ iff $\not{D}_{\text{flat}}(\rho^{1/2}\psi) = 0$: *zero modes on the sphere are $\rho^{-1/2}$ times zero modes on the plane*, with the same gauge field. (For $A = 0$ this reproduces Example 4.82: $\rho^{-1/2}f(z)$ grows like r^{1+k}, while the area element is $\rho^2\mathrm{d}^2x \sim r^{-4}\mathrm{d}^2x$, so $\int |\psi|^2\,\rho^2\mathrm{d}^2x$ converges only for $k < 0$: no zero modes.)

Worked example 4.90: Zero modes of a charge-d monopole on S^2

The rotationally symmetric monopole of charge $d > 0$ has $\phi = \frac{d}{2}\ln(1+|z|^2)$. Check the flux:

$$\begin{aligned} \int_{\mathbb{R}^2} \nabla^2\phi\,\mathrm{d}^2x &\overset{(1)}{=} \int_{\mathbb{R}^2} \frac{2d}{(1+r^2)^2}\,\mathrm{d}^2x \\ &\overset{(2)}{=} 2d\cdot 2\pi\int_0^\infty \frac{r\,\mathrm{d}r}{(1+r^2)^2} = 2d\cdot 2\pi\cdot\tfrac{1}{2} = 2\pi d. \end{aligned}$$

(1) $\nabla^2\ln(1+r^2) = \frac{1}{r}\big(r\cdot\frac{2r}{1+r^2}\big)' = \frac{4}{(1+r^2)^2}$, times $\frac{d}{2}$.
(2) Polar coordinates; $\int_0^\infty \frac{r\,\mathrm{d}r}{(1+r^2)^2} = \frac{1}{2}$.

So $\int c_1 = d$. The field is uniform on the sphere: $B\,\mathrm{d}^2x = \frac{d}{2}\rho^2\mathrm{d}^2x$ is $\frac{d}{2}$ times the round area form. By (4.140) and (4.141), the negative-chirality candidates are

$$\psi_k^- = \rho^{-1/2}e^{-\phi}z^k \propto (1+|z|^2)^{\frac{1-d}{2}}\,z^k, \qquad k = 0,1,2,\dots$$

Normalisability on the sphere:

$$\begin{aligned} \int_{S^2} \big|\psi_k^-\big|^2\,\mathrm{d}A &\overset{(1)}{=} \int_{\mathbb{R}^2} (1+r^2)^{1-d}\,r^{2k}\,\frac{4\,\mathrm{d}^2x}{(1+r^2)^2} \\ &\overset{(2)}{=} \text{finite} \iff 2k + 2(1-d) - 4 + 1 < -1 \iff k < d. \end{aligned}$$

(1) The area element of the unit sphere is $\rho^2 \mathrm{d}^2x$.
(2) Large-r power counting, as in Example 4.89.

(Equivalently, in the southern chart $w = 1/z$ the mode is smooth at $w = 0$ exactly when $k \le d-1$.) The positive-chirality candidates $\rho^{-1/2} e^{\phi} g(\bar{z})$ grow even faster and none survive. So

$$n^- = d, \qquad n^+ = 0, \qquad \operatorname{index} \not{D}_A^+ = -d = -\int_{S^2} c_1,$$

exactly as (4.138) predicts. On the plane the same flux gave $d-1$ modes; compactness supplies the missing one.

Physical picture: Zero modes are the lowest Landau level

Squaring the Dirac operator gives the Pauli Hamiltonian. On flat $\mathbb{R}^2$ the twisted Lichnerowicz formula (4.143) below reads $\not{D}_A^2 = -\nabla_A^2 + \gamma_5 B$: the curvature term is $\gamma_1\gamma_2 F_{A,12} = (i\gamma_5)(-iB)$. For $B > 0$ the operator $-\nabla_A^2$ has lowest eigenvalue B (the zero-point energy of the cyclotron motion), and on negative chirality the Zeeman term $-B$ cancels it exactly. So the Dirac zero modes are the lowest Landau level of a spin-$\frac{1}{2}$ particle with gyromagnetic ratio $g = 2$. On a sphere pierced by d flux quanta (Haldane's sphere of the quantum Hall effect) this level has degeneracy d, one state per flux quantum. (A *spinless* particle on the same sphere has $d+1$ states in its lowest level, as the Riemann–Roch count of Section 4.13 with $\mathcal{O}(d)$ showed; the difference is the $K^{1/2}$ carried by the spinor.) The index theorem is the statement that this degeneracy is topological and survives any deformation of B and of the metric. In graphene, whose electrons obey a two-dimensional Dirac equation, this is the famous zero-energy Landau level.

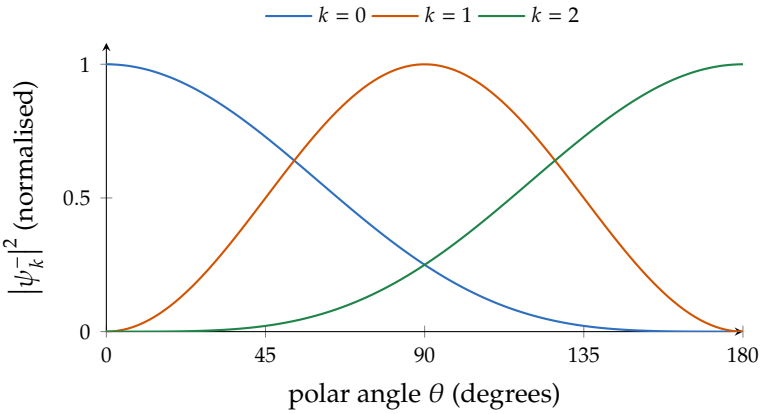


Figure 4.52: The $d = 3$ zero modes of a monopole on the unit sphere, as functions of the polar angle ($|z| = \tan\frac{\theta}{2}$, so $|\psi_k^-|^2 \propto \sin^{2k}\frac{\theta}{2}\cos^{2(d-1-k)}\frac{\theta}{2}$; each curve scaled to maximum 1). They are rings of charge centred at latitudes that move from the north pole ($k = 0$) to the south pole ($k = d-1$): the lowest Landau level on the sphere, one state per flux quantum.

Riemann–Roch cross-check. By Example 4.80 negative-chirality zero modes solve a $\bar{\partial}$-equation, and with the field they are holomorphic sections of $E^- \otimes L$, where $E^- \cong K^{1/2}$ (a square root of the canonical bundle, degree $g-1$) and L is the monopole line bundle, of degree d. On $\mathbb{CP}^1$, $\deg(K^{1/2} \otimes L) = d-1$, and the Riemann–Roch count of Section 4.13.5 gives $h^0 = (d-1)+1-0 = d$ holomorphic sections when $d \ge 1$ (they are the polynomials of degree $\le d-1$, i.e. our z^k). The Dirac index and Riemann–Roch are the same computation.

4.16.3 Four dimensions: instanton zero modes

Now let X be S^4 (or $\mathbb{R}^4$ with fields decaying at infinity, which by conformal invariance is the same problem) and F an SU(2) bundle with instanton number k. Then $p_1(S^4) = 0$ and $c_1(F) = 0$, and (4.139) becomes

$$\operatorname{index} \not{D}_A^+ = \int_{S^4} \operatorname{ch}_2(F) = -c_2(F)[S^4] = -k \qquad \text{(fundamental representation)}. \tag{4.142}$$

Which chirality? A vanishing theorem. The index gives $n^+ - n^-$. To get $n^\pm$ separately we need the twisted Lichnerowicz formula ([Bär Thm. 5.1.1])

$$\not{D}_A^2 = \nabla^*\nabla + \frac{R}{4} + \frac{1}{2}\gamma^a\gamma^b \otimes F_{A,ab}. \tag{4.143}$$

The derivation is the one of (4.127): the commutator $[\nabla_a, \nabla_b]$ now also contains the gauge curvature $F_{A,ab}$. The new term involves $\gamma^a\gamma^b F_{A,ab}$, i.e. the 2-form F_A acting by Clifford multiplication, and by Example 4.79 self-dual 2-forms act only on E^+ and anti-self-dual ones only on E^-. So for an *anti-self-dual* field ($\star F_A = -F_A$) the curvature term vanishes on $E^+ \otimes F$, and there

$$\not{D}_A^- \not{D}_A^+ = \nabla^*\nabla + \frac{R}{4}.$$

On S^4 ($R = 12$) this is strictly positive, so $n^+ = 0$; on $\mathbb{R}^4$ ($R = 0$) a normalisable ψ^+ with $\not{D}_A^+\psi^+ = 0$ would have $\nabla\psi^+ = 0$ and hence constant norm, so again $n^+ = 0$. Combined with the index:

Key idea: Counting instanton zero modes

In an anti-self-dual SU(2) instanton of charge $k > 0$, a massless Weyl fermion in the fundamental representation has exactly $n^- = k$ zero modes, all of negative chirality, and none of positive chirality. For a self-dual instanton ($k < 0$) the roles are exchanged: $|k|$ zero modes of positive chirality.

For $k = 1$ and the BPST instanton of size λ centred at the origin, the zero mode was found by 't Hooft (1976). In the regular gauge it has the form

$$\psi^-_{\alpha i}(x) \propto \frac{\lambda}{(|x|^2 + \lambda^2)^{3/2}}\, \varepsilon_{\alpha i},$$

where α is the spinor index of $E^- = \mathbb{C}^2$ and i the colour index of $F = \mathbb{C}^2$: the zero mode *locks* spin to colour, being the singlet $\varepsilon_{\alpha i}$ of the diagonal SU(2). We do not verify this here (it is a pleasant calculation with the BPST connection of Chapter 8), but the power law is easy to understand: far from the instanton the field is pure gauge, so the zero mode must look like a flat-space solution, and $\not{D}(\not{x}/|x|^4) = 0$ away from the origin in four dimensions. Its decay $|x|^{-3}$ is exactly that of $\lambda(|x|^2 + \lambda^2)^{-3/2}$, and $\int |x|^{-6}\, \mathrm{d}^4x$ converges at infinity: the mode is normalisable.

Worked example 4.91: Other representations: the Dynkin index

If the fermion transforms in a representation $\mathbf{R}$ of SU(2), the bundle is $F_{\mathbf{R}}$ and $\operatorname{ch}_2(F_{\mathbf{R}}) = -\frac{1}{8\pi^2}\operatorname{tr}_{\mathbf{R}}(F_A \wedge F_A)$. Write $F_A = F^a T_a$ with anti-Hermitian generators; then $\operatorname{tr}_{\mathbf{R}}(T_a T_b) = -T(\mathbf{R})\,\delta_{ab}$ with $T(\mathbf{R})$ the **Dynkin index**. For isospin j, $T(j) = \frac{1}{3}j(j+1)(2j+1)$ (so $T(\frac{1}{2}) = \frac{1}{2}$,

$T(1) = 2$). Hence

$$\operatorname{index} \not{D}^+_{A,\mathbf{R}} \overset{(1)}{=} -\frac{1}{8\pi^2}\int \operatorname{tr}_{\mathbf{R}}(F_A \wedge F_A) = \frac{T(\mathbf{R})}{8\pi^2}\int F^a \wedge F^a$$
$$\overset{(2)}{=} \frac{T(\mathbf{R})}{T(\frac12)}\cdot\Big(-\frac{1}{8\pi^2}\int \operatorname{tr}_{\frac12}(F_A\wedge F_A)\Big) = -\frac{T(\mathbf{R})}{T(\frac12)}\,k = -\tfrac{2}{3}j(j+1)(2j+1)\,k.$$

(1) Definition of ch_2 and of the Dynkin index.
(2) Compare with the fundamental, for which the bracket is $-k$ by (4.142).

Check: $j=\frac12$ gives $\frac23\cdot\frac12\cdot\frac32\cdot 2 = 1$. For the adjoint ($j=1$): $\frac23\cdot1\cdot2\cdot3 = 4$, so a gluino has $4k$ zero modes in a charge-k instanton. For $k=1$ in $\mathcal{N}=1$ super Yang–Mills, two of the four are generated by the broken supersymmetries and two by the broken superconformal symmetries: the index knows about symmetries the operator never mentions. For $\mathrm{SU}(N)$ one finds $1\cdot k$ modes in the fundamental and $2N\cdot k$ in the adjoint.

Worked example 4.92: The dimension of the instanton moduli space

The infinitesimal deformations $a\in\Omega^1(X;\operatorname{ad}F)$ of an anti-self-dual connection satisfy the linearised equations $\mathrm{d}^+_A a = 0$ (stay anti-self-dual) and $\mathrm{d}^*_A a = 0$ (not a gauge transformation). So the moduli space has dimension $\operatorname{index}\big(\mathrm{d}^*_A+\mathrm{d}^+_A : \Omega^1\to\Omega^0\oplus\Omega^+\big)$ when the cokernel vanishes (it does on S^4, by a vanishing theorem like the one above). This operator is a twisted Dirac operator. In four dimensions the representation theory of $\mathrm{Spin}(4)=\mathrm{SU}(2)_+\times\mathrm{SU}(2)_-$ gives

$$\Lambda^1\otimes\mathbb{C} = E^+\otimes E^-, \qquad (\Lambda^0\oplus\Lambda^+)\otimes\mathbb{C} = E^+\otimes E^+$$

(a vector is $(\frac12,\frac12)$; $(\frac12,0)\otimes(\frac12,0)=(0,0)\oplus(1,0)$, the scalar and the self-dual 2-forms). So $(\Lambda^0\oplus\Lambda^+)\to\Lambda^1$ is $\not{D}^+$ twisted by $W = E^+\otimes\operatorname{ad}F$, and the operator we want is its adjoint. On S^4 with $\operatorname{ad}F$ of rank 3:

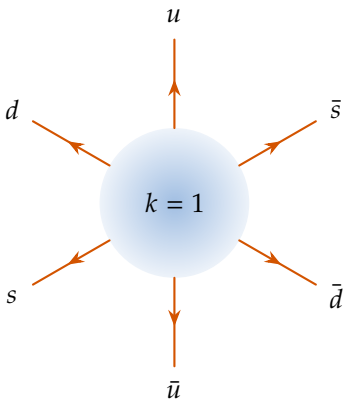
$$\operatorname{index}\big(\mathrm{d}^*_A+\mathrm{d}^+_A\big) \overset{(1)}{=} -\operatorname{index}\not{D}^+_W = -\int_{S^4}\hat{A}\,\mathrm{ch}(E^+)\,\mathrm{ch}(\operatorname{ad}F)$$
$$\overset{(2)}{=} -\int_{S^4}\Big(2\,\mathrm{ch}_2(\operatorname{ad}F) + 3\,\mathrm{ch}_2(E^+)\Big)$$
$$\overset{(3)}{=} -\big(2\cdot(-4k)+3\cdot1\big) = 8k-3.$$

(1) The index of an adjoint is minus the index; (4.137) with $n/2=2$ and $\mathrm{ch}(W) = \mathrm{ch}(E^+)\,\mathrm{ch}(\operatorname{ad}F)$.
(2) $\hat{A}=1$ on S^4 ($p_1=0$); E^+ has rank 2, $\operatorname{ad}F$ rank 3, and both have $c_1=0$, so the 4-form part of the product is $2\,\mathrm{ch}_2(\operatorname{ad}F)+3\,\mathrm{ch}_2(E^+)$.
(3) $\int\mathrm{ch}_2(\operatorname{ad}F) = -4k$ by Example 4.91. For E^+, whose roots are $\pm\frac12(x_1+x_2)$ (Example 4.84), $\mathrm{ch}_2(E^+) = \frac14(x_1+x_2)^2 = \frac14 p_1+\frac12 e$, which integrates on S^4 to $0+\frac12\cdot2=1$.

This is the celebrated $8k-3$ of Atiyah–Hitchin–Singer: for $k=1$, five parameters, the centre (4) and the size (1) of the BPST instanton. (On $\mathbb{R}^4$ one adds the 3 global gauge orientations and gets $8k$.) The moduli spaces are the subject of Chapter 8.

Physical picture: Why physicists care: anomalies from zero modes

In the Euclidean path integral, a fermion zero mode is a Grassmann variable that does not appear in the action, and $\int \mathrm{d}\theta\,1 = 0$. So in an instanton background the partition



each of the N_f Dirac flavours contributes one zero mode of each Weyl component, ψ_L and $\bar{\psi}_R$

Grassmann integration over the $2N_f k$ zero modes gives an effective $2N_f$-fermion vertex: $\Delta Q_5 = 2N_f k$

Figure 4.53: The 't Hooft vertex. In a one-instanton background the fermion path integral vanishes unless every zero mode is soaked up by an external fermion line. With three light flavours (QCD) the instanton induces a six-quark interaction that violates the axial U(1) charge by $2N_f = 6$. The number of legs is fixed by the index theorem.

function vanishes, and only correlation functions with enough fermion insertions to absorb every zero mode survive. Each flavour gives $n^- - n^+ = k$ zero modes of one chirality, so the surviving correlators carry net chiral charge $2N_f k$ (a Dirac fermion has two Weyl components). This is the integrated form of the Adler–Bell–Jackiw anomaly: the index theorem is the statement that $\int \partial_\mu J_5^\mu = 2N_f \ \mathrm{index}\, \not{D}_A^+$ up to sign, and the anomaly density is the index density. It solves the $\mathrm{U}(1)_A$ problem of QCD and is the subject of Chapter 10.

4.16.4 The four classical complexes as twisted Dirac operators

We end this section by closing a circle. The de Rham, signature and Dolbeault complexes were introduced independently, and their index densities (the Euler class, the L-class, the Todd class) were computed separately. But all three are Dirac operators twisted by a suitable bundle, and their densities are $\hat{A} \cdot \mathrm{ch}(\mathrm{twist})$.

Forms are spinors times spinors. On a spin manifold of dimension $n = 2m$ the bundle of complex forms is

$$\Lambda^* T^* X \otimes \mathbb{C} \ \cong \ E \otimes E^*, \tag{4.144}$$

with ranks $2^n = 2^m \cdot 2^m$. This is the linear-algebra fact that $\mathrm{End}(S)$, the $2^m \times 2^m$ matrices, has the products $\gamma_{a_1} \cdots \gamma_{a_p}$ $(a_1 < \cdots < a_p)$ as a basis, one for each basic p-form $e^{a_1} \wedge \cdots \wedge e^{a_p}$. For $n = 2$: $1, \gamma_1, \gamma_2, \gamma_1\gamma_2$ span all 2×2 matrices, matching $1, e^1, e^2, e^1 \wedge e^2$. Under this identification, Clifford multiplication by e^a from the left is $e^a \wedge - \iota_{e_a}$ on forms (up to a sign convention), and so $\mathrm{d} + \mathrm{d}^*$ is the Dirac operator on E twisted by E^*. The different complexes differ only in how the 2^n-dimensional space is split into "even" and "odd":

- grading by $\gamma_5 \otimes 1$ (chirality of the left factor): this is τ, and gives the *signature* complex;
- grading by $\gamma_5 \otimes \gamma_5$ (both factors): this is the parity of the form degree, and gives the *de Rham* complex.

Signature. With the twist $E^* \cong E$ (for even n the spin representation is self-dual up to isomorphism, so $\mathrm{ch}\, E^* = \mathrm{ch}\, E$), whose Chern character is $\mathrm{ch}\, E = \prod_j (e^{x_j/2} + e^{-x_j/2}) = \prod_j 2\cosh \frac{x_j}{2}$

(the sum over all sign choices, now without the factor $\prod \varepsilon_j$):

$$(-1)^m \int_X \hat{A}\,\mathrm{ch}(E) \overset{(1)}{=} (-1)^m \int_X \prod_{j=1}^{m} \frac{x_j/2}{\sinh(x_j/2)} \cdot 2\cosh\frac{x_j}{2}$$
$$\overset{(2)}{=} (-1)^m\, 2^m \int_X \prod_{j=1}^{m} \frac{x_j/2}{\tanh(x_j/2)} = (-1)^m\, 2^m \int_X \mathcal{L}(X). \tag{4.145}$$

(1) The twisted index formula (4.137) with $F = E$.
(2) $\cosh/\sinh = 1/\tanh$; one factor 2 per plane.

For $n = 4k$ the sign is $+1$ and this is exactly Nash's $\mathrm{Sign}(X) = 2^{n/2}\mathcal{L}(X)[X]$ of Section 4.14, now with the normalisation $\mathcal{L} = \prod \frac{x/2}{\tanh(x/2)}$ explained: the $\frac{x}{2}$ comes from the spinor roots. (This is Bär's proof of the signature theorem, [Bär §5.2].)

de Rham. Now the twist is graded as well, and the numerator is $(\mathrm{ch}\,E^+ - \mathrm{ch}\,E^-)(\mathrm{ch}\,E^{+*} - \mathrm{ch}\,E^{-*})$. The dual bundles have the negated roots, so

$$(\mathrm{ch}\,E^+ - \mathrm{ch}\,E^-)\big(\mathrm{ch}\,E^{+*} - \mathrm{ch}\,E^{-*}\big) \overset{(1)}{=} \prod_j \big(e^{x_j/2} - e^{-x_j/2}\big)\big(e^{-x_j/2} - e^{x_j/2}\big)$$
$$\overset{(2)}{=} \prod_j \big(1 - e^{x_j}\big)\big(1 - e^{-x_j}\big), \tag{4.146}$$

(1) (Nash 4.107) for $E^{\pm}$, and the same with $x_j \to -x_j$ for the duals.
(2) Multiply out: $(a - a^{-1})(a^{-1} - a) = 2 - a^2 - a^{-2} = (1 - a^2)(1 - a^{-2})$ with $a = e^{x/2}$.

which is Nash's (4.59) for $\sum_p (-1)^p\,\mathrm{ch}\,\Lambda^p$, the alternating sum that led to Gauss–Bonnet in Section 4.12. The corresponding index density is

$$(-1)^m\,\hat{A} \cdot \big(\mathrm{ch}\,E^{+*} - \mathrm{ch}\,E^{-*}\big) \overset{(1)}{=} (-1)^m \prod_j \frac{x_j/2}{\sinh(x_j/2)} \cdot \big(-2\sinh\tfrac{x_j}{2}\big)$$
$$\overset{(2)}{=} (-1)^m (-1)^m \prod_j x_j = e(X),$$

(1) The graded twist has $\mathrm{ch} = \prod_j (e^{-x_j/2} - e^{x_j/2}) = \prod_j (-2\sinh\frac{x_j}{2})$.
(2) The sinh cancels, leaving $-x_j$ per plane.

the Euler class, and $\mathrm{index}(\mathrm{d} + \mathrm{d}^*)_{\mathrm{even}\to\mathrm{odd}} = \chi(X)$.

Dolbeault. On a complex manifold with holomorphic Chern roots $y_1, \dots, y_m$ (so that the real block angles are $x_j = y_j$ and $c_1 = \sum y_j$), compare the Todd and $\hat{A}$ series root by root:

$$\mathrm{td}(X) \overset{(1)}{=} \prod_{j=1}^{m} \frac{y_j}{1 - e^{-y_j}} = \prod_{j=1}^{m} \frac{y_j\, e^{y_j/2}}{e^{y_j/2} - e^{-y_j/2}}$$
$$\overset{(2)}{=} \prod_{j=1}^{m} e^{y_j/2}\,\frac{y_j/2}{\sinh(y_j/2)} = e^{c_1(X)/2}\,\hat{A}(X). \tag{4.147}$$

(1) Definition of the Todd class; multiply numerator and denominator by $e^{y/2}$.
(2) $e^{y/2} - e^{-y/2} = 2\sinh(y/2)$; the product of $e^{y_j/2}$ is $e^{\sum y_j/2} = e^{c_1/2}$.

And $e^{c_1/2} = \mathrm{ch}(K^{-1/2})$, the Chern character of a line bundle with first Chern class $\frac{1}{2}c_1(X)$: a square root of the anticanonical bundle $K^{-1} = \Lambda^m T^{1,0}X$. So *the Dolbeault operator is the Dirac operator twisted by* $K^{-1/2}$, and Riemann–Roch–Hirzebruch $\chi(X,\mathcal{O}) = \int \mathrm{td}(X)$ is the Dirac index theorem $\int \hat{A}\ \mathrm{ch}(K^{-1/2})$ (up to the chirality convention; in complex dimension 1 we checked all signs in the Riemann–Roch cross-check above).

Worked example 4.93: Two wrongs make a right: $\mathbb{CP}^2$

On $\mathbb{CP}^2$ neither E nor $K^{-1/2}$ exists ($c_1 = 3h$ is odd, so no square root). But their product $E \otimes K^{-1/2} \cong \Lambda^{0,*}$ does exist: the transition-function signs that obstruct each factor cancel in the product. Such a structure is a **spin^c structure**. The index density still makes sense, and

$$\int_{\mathbb{CP}^2} \hat{A}\, e^{c_1/2} \overset{(1)}{=} \int_{\mathbb{CP}^2} \Big(1 - \frac{p_1}{24}\Big)\Big(1 + \frac{c_1}{2} + \frac{c_1^2}{8}\Big) = \int_{\mathbb{CP}^2} \Big(-\frac{p_1}{24} + \frac{c_1^2}{8}\Big)$$
$$\overset{(2)}{=} -\frac{3}{24} + \frac{9}{8} = -\frac{1}{8} + \frac{9}{8} = 1.$$

(1) Expand both factors and keep the 4-form part.
(2) $\int p_1 = 3$ and $\int c_1^2 = \int 9h^2 = 9$ (Example 4.50).

The non-integral gravitational part $-\frac{1}{8}$ (Example 4.85) is compensated by the half-integral "flux" of $K^{-1/2}$, and the total is the integer $\chi(\mathbb{CP}^2,\mathcal{O}) = h^{0,0} - h^{0,1} + h^{0,2} = 1 - 0 + 0 = 1$. Physically: on a non-spin spacetime a fermion can still be defined if it carries half-integral charge under a U(1) gauge field. This is the mechanism behind spin^c structures in Seiberg–Witten theory.

complex	as twisted Dirac	index density	index	§
de Rham	$E \otimes E^*$, graded by $\gamma_5 \otimes \gamma_5$	$e(X)$	$\chi(X) = \sum(-1)^p b_p$	4.12
signature	$E \otimes E^*$, graded by $\gamma_5 \otimes 1$	$2^{n/2}\mathcal{L}(X)$	$\mathrm{Sign}(X) = b^+ - b^-$	4.14
Dolbeault	$E \otimes K^{-1/2}$	$\mathrm{td}(X) = e^{c_1/2}\hat{A}$	$\chi(X,\mathcal{O}) = \sum(-1)^q h^{0,q}$	4.13
spin	E	$\hat{A}(X)$	$n^+ - n^-$	4.15
gauge-coupled spin	$E \otimes F$	$\hat{A}(X)\,\mathrm{ch}(F)$	charged zero modes	4.16

Table 4.3: The classical complexes of Nash's §2, seen as one operator. Each is the Dirac operator on spinors tensored with a "coefficient" bundle, and each index density is $\hat{A}$ times the Chern character of the coefficients (with the appropriate grading). The de Rham and signature complexes exist even on non-spin manifolds, because $E \otimes E^*$ always exists although E may not; the local computation is the same.

Watch out: Non-spin manifolds

The identification (4.144) needs spinors, which may not exist. The de Rham and signature operators do exist on every oriented manifold, and their index theorems hold there too. The reason the spin computation still gives the right answer is that index densities are *local*: every point has a neighbourhood that is spin, and on it the local integrand is $\hat{A}\,\mathrm{ch}(E)$, a polynomial in the curvature. The global identity then follows for all manifolds ([Bär §5.2] uses exactly this argument).

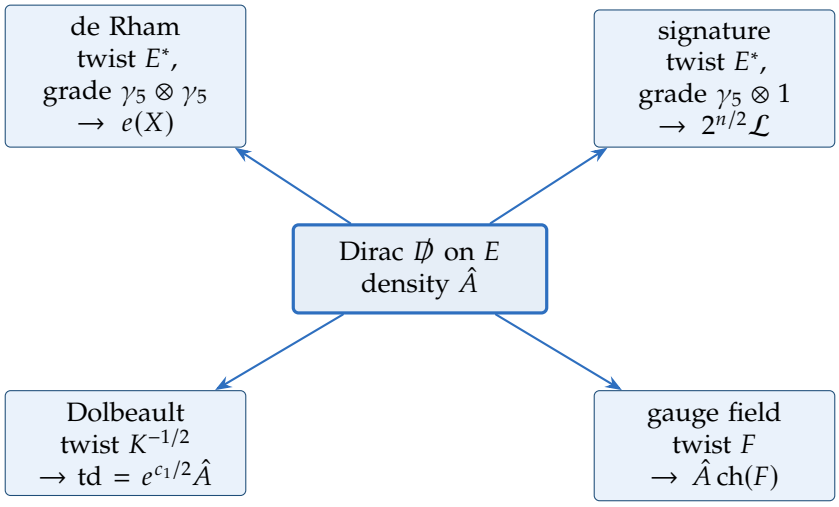


Figure 4.54: All roads lead from the Dirac operator. Twisting the spinor bundle by a coefficient bundle multiplies the index density $\hat{A}$ by the Chern character of the coefficients; choosing how to grade the result selects the complex. The Euler class, L-class and Todd class are all of the form $\hat{A} \cdot \mathrm{ch}(\cdots)$.

Reading Nash (beyond §2; compare Nash pp. 116–119, (4.112)–(4.122))

Nash's §2 ends with (4.111) and does not couple the Dirac operator to gauge fields. His §3 ("Twisted complexes", treated in full in the next part of these notes) replaces each E^p by $F \otimes E^p$ and finds, among others, the twisted Dirac index $(-1)^{n/2}\int \mathrm{ch}(F)\hat{A}(X)$, his (4.122), and the remark that this is nonzero also in dimensions not divisible by 4. This section has used that formula to count zero modes explicitly (monopoles on $\mathbb{R}^2$ and S^2, instantons on S^4), and has shown how the de Rham, signature and Dolbeault complexes of §2 are themselves twisted Dirac operators, which explains why all the index densities are built from the same few power series.

Recap

- Charged fermions: $\not{D}_A = \gamma^a e_a^\mu(\partial_\mu + \omega_\mu + A_\mu)$ on $E \otimes F$; index $\not{D}_A^+ = (-1)^{n/2}\int \hat{A}\ \mathrm{ch}(F)$.
- 2D: index $\not{D}_A^+ = -\int c_1(F)$. On S^2 with d flux quanta there are d zero modes, $(1+|z|^2)^{(1-d)/2}z^k$, $0 \le k < d$: the lowest Landau level. On $\mathbb{R}^2$ only $\lceil \Phi/2\pi \rceil - 1$ (Aharonov–Casher).
- 4D: index $\not{D}_A^+ = \frac{r}{192\pi^2}\int \mathrm{tr}\, R \wedge R - \frac{1}{8\pi^2}\int \mathrm{tr}\, F_A \wedge F_A$. ASD SU(2) instanton of charge k: k zero modes (fundamental), $4k$ (adjoint), all of one chirality by the twisted Lichnerowicz formula.
- The moduli space of SU(2) instantons on S^4 has dimension $8k-3$, a twisted Dirac index.
- $\Lambda^* \otimes \mathbb{C} = E \otimes E^*$: de Rham and signature are Dirac twisted by spinors, Dolbeault is Dirac twisted by $K^{-1/2}$; e, $2^{n/2}\mathcal{L}$, td are all $\hat{A} \cdot \mathrm{ch}(\cdots)$.

4.17 Twisted complexes

In physics, almost nothing is "bare". An electron field is not just a spinor: it is a spinor that also carries an electric charge, so it lives in a spinor bundle *tensored* with a U(1) line bundle. A quark is a spinor tensored with a colour triplet. A gluino is a spinor tensored with the adjoint of SU(3). The operator that acts on such a field is not the bare Dirac operator $\not{D}$ but the gauge-covariant one, $\gamma^\mu(\partial_\mu + \omega_\mu + A_\mu)$.

So the question of this section is:

> *If we couple a classical elliptic operator to a gauge field, how does its index change?*

The answer turns out to be pleasingly simple. The index formula of the untwisted operator gets multiplied, under the integral sign, by the Chern character $\mathrm{ch}(F)$ of the gauge bundle F. Whether this changes anything depends on the operator. For the de Rham complex it changes almost nothing (only the rank of F survives). For the Dolbeault, signature and Dirac operators the gauge field changes the index completely. This is the mathematical origin of the fact that fermion zero modes in an instanton or monopole background are counted by the topological charge.

Nash calls the construction **twisting**, and the new complexes **twisted complexes**.

4.17.1 The tools: index of a complex and the Chern character

We first collect, in one place, the two tools the whole section relies on. Both were introduced earlier in the chapter; we restate them so that this part can be read on its own.

Elliptic complexes and the index formula

An **elliptic complex** is a sequence of vector bundles $E^0, E^1, \dots, E^N$ over a closed n-manifold X and differential operators between their sections,

$$0 \longrightarrow \Gamma(X, E^0) \xrightarrow{d_0} \Gamma(X, E^1) \xrightarrow{d_1} \cdots \xrightarrow{d_{N-1}} \Gamma(X, E^N) \longrightarrow 0, \qquad d_p \circ d_{p-1} = 0, \tag{4.148}$$

such that the corresponding sequence of *symbols* (replace each derivative ∂_μ by ip_μ and keep only the highest-order terms) is exact whenever $p \neq 0$. The cohomology $H^p(E) = \operatorname{Ker} d_p / \operatorname{Im} d_{p-1}$ is finite dimensional, and the **index of the complex** is

$$\operatorname{index} E \;=\; \sum_p (-1)^p \dim H^p(E). \tag{4.149}$$

One can always "roll up" a complex into a single operator: choose inner products, and let

$$D = d + d^* : \bigoplus_{p \text{ even}} \Gamma(E^p) \longrightarrow \bigoplus_{p \text{ odd}} \Gamma(E^p). \tag{4.150}$$

By Hodge theory $\operatorname{Ker} D$ is the sum of the even cohomologies and $\operatorname{Ker} D^*$ the sum of the odd ones, so $\operatorname{index} D = \operatorname{index} E$.

The index formula for complexes that we will use repeatedly (Nash's (4.48), derived in the earlier part of this chapter) reads, for X even dimensional and oriented:

$$\operatorname{index} E \;=\; (-1)^{n/2} \left\{ \frac{\mathrm{ch}\big(\sum_p (-1)^p E^p\big)}{e(X)} \; \mathrm{td}(TX_{\mathbb{C}}) \right\} [X]. \tag{4.151}$$

Here $TX_{\mathbb{C}} = TX \otimes_{\mathbb{R}} \mathbb{C}$, $e(X)$ is the Euler class of TX, td is the Todd class, and "$\{\cdots\}[X]$" means: take the part of the inhomogeneous differential form of degree exactly n and integrate it over X. The ratio makes sense because the numerator always turns out to be divisible by $e(X)$ in the examples.

Key idea: How to read the index formula

The integrand is a polynomial in curvatures. We expand it, keep only the piece of total form degree $n = \dim X$, and integrate. Everything of lower or higher degree is thrown away. Most of this section consists of doing this bookkeeping carefully.

The Chern character

Let $F \to X$ be a complex vector bundle of rank m with a connection A (an anti-Hermitian $m \times m$ matrix of 1-forms in a local trivialisation) and curvature $F_A = \mathrm{d}A + A \wedge A$.

Working definition 4.44 (Chern character)

The **Chern character** of F is the closed, inhomogeneous [a sum of pieces of degrees $0, 2, 4, \ldots$] differential form

$$\mathrm{ch}(F) = \operatorname{tr} \exp\!\left(\frac{iF_A}{2\pi}\right) = m + \frac{i}{2\pi}\operatorname{tr} F_A + \frac{1}{2}\left(\frac{i}{2\pi}\right)^2 \operatorname{tr}(F_A \wedge F_A) + \cdots .$$

To compute it: write down the curvature matrix, take traces of its wedge powers, and sort by degree: $\mathrm{ch}_0 = m$ (the rank), $\mathrm{ch}_1 = c_1$, $\mathrm{ch}_2 = \frac{1}{2}(c_1^2 - 2c_2)$, and so on. *Splitting-principle form:* if we pretend $F = L_1 \oplus \cdots \oplus L_m$ is a sum of line bundles [rank-one bundles] with first Chern classes $x_1, \ldots, x_m$ (the **Chern roots**), then $\mathrm{ch}(F) = \sum_i e^{x_i}$.

Dictionary: anti-Hermitian versus Hermitian gauge fields

We use the mathematicians' convention (as Nash does): A is anti-Hermitian, $F_A = \mathrm{d}A + A \wedge A$, and gauge transformations act as $A \mapsto g^{-1}Ag + g^{-1}\mathrm{d}g$. The physicists' Hermitian field $A^a_\mu T^a$ with coupling g is related by $A_{\text{math}} = -i\, g\, A^a_{\text{phys}} T^a$. For U(1) with a real potential $a = a_\mu \mathrm{d}x^\mu$ (charge absorbed), $A = -ia$, $F_A = -i\,\mathrm{d}a$ and $c_1 = \frac{i}{2\pi}F_A = \frac{1}{2\pi}\mathrm{d}a$: the first Chern class is the magnetic field divided by 2π.

The two properties that make ch useful are that it turns $\oplus$ into $+$ and $\otimes$ into $\times$. The second one is the only thing Nash uses in [Nash p. 117, (4.115)], so let us see why it holds. Take a connection A on F and B on G. The natural connection on $F \otimes G$ is $A \otimes 1 + 1 \otimes B$, and its curvature is $F_{A\otimes G} = F_A \otimes 1 + 1 \otimes F_B$. The two terms commute (they act on different tensor factors, and they are even forms). Hence

$$\begin{aligned}
\mathrm{ch}(F \otimes G) &\overset{(1)}{=} \operatorname{tr}_{F\otimes G} \exp\!\left(\tfrac{i}{2\pi}(F_A \otimes 1 + 1 \otimes F_B)\right) \\
&\overset{(2)}{=} \operatorname{tr}_{F\otimes G}\left[\exp\!\left(\tfrac{i}{2\pi}F_A\right) \otimes \exp\!\left(\tfrac{i}{2\pi}F_B\right)\right] \\
&\overset{(3)}{=} \operatorname{tr}_F \exp\!\left(\tfrac{i}{2\pi}F_A\right)\ \operatorname{tr}_G \exp\!\left(\tfrac{i}{2\pi}F_B\right) \\
&= \mathrm{ch}(F)\ \mathrm{ch}(G).
\end{aligned} \tag{4.152}$$

(1) Definition of ch applied to the tensor product connection.
(2) $e^{X+Y} = e^X e^Y$ when X and Y commute; $X \otimes 1$ and $1 \otimes Y$ always commute, and $e^{X\otimes 1} = e^X \otimes 1$.

(3) The trace of a tensor product of matrices is the product of the traces: $\operatorname{tr}(M \otimes N) = \sum_{i,a} M_{ii} N_{aa} = (\operatorname{tr} M)(\operatorname{tr} N)$.

In terms of Chern roots this is the statement that the roots of $F \otimes G$ are all the sums $x_i + y_a$, and $\sum_{i,a} e^{x_i + y_a} = \left(\sum_i e^{x_i}\right)\left(\sum_a e^{y_a}\right)$. Similarly $\operatorname{ch}(F \oplus G) = \operatorname{ch} F + \operatorname{ch} G$ because the curvature of a direct sum is block diagonal and the trace of a block-diagonal matrix is the sum of the traces.

Worked example 4.94: The Chern character of the monopole bundle

On S^2 with polar angles (θ, ϕ), the charge-k monopole bundle L_k has the two local connection forms

$$A_N = -\tfrac{ik}{2}(1 - \cos\theta)\,\mathrm{d}\phi \quad (\theta < \pi), \qquad A_S = +\tfrac{ik}{2}(1 + \cos\theta)\,\mathrm{d}\phi \quad (\theta > 0).$$

Each is non-singular in its patch ($1 - \cos\theta$ vanishes at the north pole, where $\mathrm{d}\phi$ is ill-defined, and similarly in the south). On the overlap $A_N - A_S = -ik\,\mathrm{d}\phi = g^{-1}\mathrm{d}g$ with $g = e^{-ik\phi}$, which is single-valued precisely because $k \in \mathbb{Z}$. The curvature is the same in both patches:

$$\begin{aligned} F_A &\overset{(1)}{=} \mathrm{d}A_N = -\tfrac{ik}{2}\sin\theta\,\mathrm{d}\theta \wedge \mathrm{d}\phi, \\ \int_{S^2} \operatorname{ch}_1(L_k) &\overset{(2)}{=} \frac{i}{2\pi}\int_{S^2}\left(-\frac{ik}{2}\right)\sin\theta\,\mathrm{d}\theta \wedge \mathrm{d}\phi \\ &\overset{(3)}{=} \frac{k}{4\pi}\cdot 4\pi = k. \end{aligned}$$

(1) For a U(1) bundle $A \wedge A = 0$, so $F_A = \mathrm{d}A$; and $\mathrm{d}(1 - \cos\theta) = \sin\theta\,\mathrm{d}\theta$.
(2) $\operatorname{ch}_1 = c_1 = \frac{i}{2\pi}F_A$ for a line bundle.
(3) $i \cdot (-i) = 1$ and the area of the unit sphere is 4π.

So $\operatorname{ch}(L_k) = 1 + c_1(L_k)$ with $\int_{S^2} c_1(L_k) = k$. There is no degree-4 term because S^2 has no 4-forms.

For reference, here are the characteristic classes that appear in the four classical complexes, written in terms of the Chern roots x_i of TX (for a real 2ℓ-manifold, $TX_{\mathbb{C}}$ has roots $\pm x_1, \dots, \pm x_\ell$).

class	as a product over roots	first terms
Euler $e(X)$	$\prod_{i=1}^{\ell} x_i$	top degree only
Todd $\operatorname{td}(X)$ (complex X)	$\prod_i \dfrac{x_i}{1 - e^{-x_i}}$	$1 + \frac{1}{2}c_1 + \frac{1}{12}(c_1^2 + c_2) + \cdots$
$\hat{A}(X)$	$\prod_i \dfrac{x_i/2}{\sinh(x_i/2)}$	$1 - \frac{1}{24}p_1 + \frac{1}{5760}(7p_1^2 - 4p_2) + \cdots$
Hirzebruch $L(X)$	$\prod_i \dfrac{x_i}{\tanh x_i}$	$1 + \frac{1}{3}p_1 + \frac{1}{45}(7p_2 - p_1^2) + \cdots$

Here $p_1 = \sum_i x_i^2$ and $p_2 = \sum_{i<j} x_i^2 x_j^2$ are the Pontrjagin classes.

4.17.2 Matrix-valued forms and the twisted de Rham complex

What a matrix-valued form is

The ordinary de Rham complex [Nash p. 116, (4.112)] is

$$\cdots \xrightarrow{d_{p-1}} \Omega^p(X) \xrightarrow{d_p} \Omega^{p+1}(X) \xrightarrow{d_{p+1}} \cdots, \qquad \Omega^p(X) = \Gamma(X, \Lambda^p T^* X_{\mathbb{C}}).$$

In gauge theory we constantly meet forms that carry an extra index: the gauge potential $A^a{}_b$, the field strength $F^a{}_b$, a charged scalar ϕ^a, a covariant derivative $D\phi^a$. These are **vector-valued** (or matrix-valued) forms.

Working definition 4.45 (forms with values in a bundle)

Let $F \to X$ be a complex vector bundle of rank m. An **F-valued p-form** is, in each local trivialisation U_α of F, a column of m ordinary p-forms $\psi_\alpha = (\psi_\alpha^1, \dots, \psi_\alpha^m)^T$, such that on overlaps $\psi_\alpha = g_{\alpha\beta}\,\psi_\beta$, where $g_{\alpha\beta}$ are the transition functions [the $m \times m$ matrices $g_{\alpha\beta}(x)$ relating the two trivialisations on $U_\alpha \cap U_\beta$] of F. The space of them is

$$\Omega^p(X;F) = \Gamma(X, F \otimes \Lambda^p T^*X_{\mathbb{C}}).$$

[smooth sections, with one F-index and p antisymmetric form indices at each point] *Test:* if you can write it as "(an index in the gauge representation) × (a differential form)", and it transforms under gauge transformations only through the gauge index, it is an F-valued form.

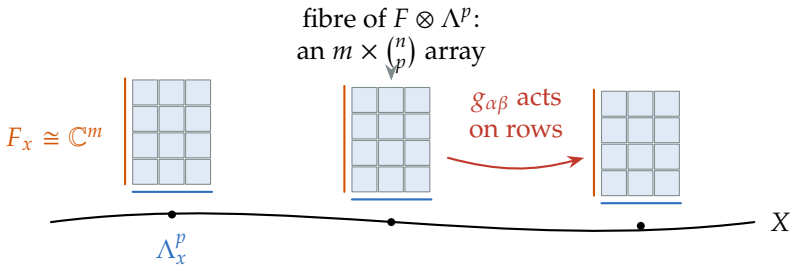


Figure 4.55: An F-valued p-form assigns to each point an array: one column index for the p-form components ($\binom{n}{p}$ of them) and one row index for the gauge components (m of them). Gauge transformations $g_{\alpha\beta}$ mix only the rows. The ordinary de Rham complex is the case of a single row.

The twisted exterior derivative

To differentiate an F-valued form in a gauge-covariant way we need a connection A on F. The **covariant exterior derivative** is

$$d_A\psi = \mathrm{d}\psi + A \wedge \psi, \tag{4.153}$$

where $A \wedge \psi$ means matrix multiplication in the gauge index and wedge product in the form index. Nash writes d_F for it. Let us check it is covariant. Under a gauge transformation $\psi' = g^{-1}\psi$ and $A' = g^{-1}Ag + g^{-1}\mathrm{d}g$:

$$\begin{aligned} d_{A'}\psi' &\overset{(1)}{=} \mathrm{d}(g^{-1}\psi) + (g^{-1}Ag + g^{-1}\mathrm{d}g)\,g^{-1}\psi \\ &\overset{(2)}{=} -g^{-1}(\mathrm{d}g)\,g^{-1}\psi + g^{-1}\mathrm{d}\psi + g^{-1}A\psi + g^{-1}(\mathrm{d}g)\,g^{-1}\psi \\ &\overset{(3)}{=} g^{-1}(\mathrm{d}\psi + A\psi) = g^{-1}\,d_A\psi. \end{aligned}$$

(1) Definition (4.153) in the new gauge.
(2) Leibniz rule and $\mathrm{d}(g^{-1}) = -g^{-1}(\mathrm{d}g)\,g^{-1}$ (differentiate $g^{-1}g = 1$).
(3) The two $\mathrm{d}g$ terms cancel.

So $d_A\psi$ transforms exactly like ψ: it is again an F-valued form. Replacing d by d_A in the de Rham complex gives Nash's **twisted de Rham complex** [Nash p. 116, (4.113)]:

$$\cdots \xrightarrow{d_A} \Omega^{p-1}(X;F) \xrightarrow{d_A} \Omega^p(X;F) \xrightarrow{d_A} \Omega^{p+1}(X;F) \xrightarrow{d_A} \cdots . \tag{4.154}$$

Watch out: The twisted de Rham sequence is a complex only for flat F

Unlike $\mathrm{d}^2 = 0$, the square of d_A is not zero:

$$\begin{aligned} d_A(d_A\psi) &\overset{(1)}{=} \mathrm{d}(\mathrm{d}\psi + A\wedge\psi) + A\wedge(\mathrm{d}\psi + A\wedge\psi) \\ &\overset{(2)}{=} (\mathrm{d}A)\wedge\psi - A\wedge\mathrm{d}\psi + A\wedge\mathrm{d}\psi + A\wedge A\wedge\psi \\ &\overset{(3)}{=} (\mathrm{d}A + A\wedge A)\wedge\psi = F_A\wedge\psi. \end{aligned}$$

(1) Apply (4.153) twice.
(2) $\mathrm{d}^2 = 0$, and the graded Leibniz rule $\mathrm{d}(A\wedge\psi) = \mathrm{d}A\wedge\psi - A\wedge\mathrm{d}\psi$ (the minus sign because A is a 1-form).
(3) The $A\wedge\mathrm{d}\psi$ terms cancel; what remains is the curvature.

So (4.154) is literally a complex only when $F_A = 0$, i.e. for a flat bundle. Nash passes over this. It does not matter for the *index*: the operator whose index we compute is the rolled-up $D_A = d_A + d_A^*$ from even to odd F-valued forms, and this is elliptic for *every* connection. Its leading symbol is $\sigma(D_A)(x,p) = i(p\wedge - \iota_p)\otimes 1_F$, independent of A: the connection only enters at order zero. Since the index depends only on the symbol class, the formula below holds for any F and any A.

The index of the twisted de Rham operator

Recall how the untwisted index came out [Nash p. 117, (4.114)]: with $E^p = \Lambda^p T^*X_{\mathbb{C}}$,

$$\operatorname{index} d = (-1)^{n/2}\,\frac{\operatorname{ch}\big(\sum_p (-1)^p[E^p]\big)}{e(X)}\,\operatorname{td}(TX_{\mathbb{C}})[X] = \int_X e(X) = \chi(X). \tag{4.155}$$

Passing to d_A changes only one thing: E^p becomes $F\otimes E^p$. Then

$$\begin{aligned} \operatorname{index} d_A &\overset{(1)}{=} (-1)^{n/2}\,\frac{\operatorname{ch}\big(\sum_p (-1)^p[F\otimes E^p]\big)}{e(X)}\,\operatorname{td}(TX_{\mathbb{C}})[X] \\ &\overset{(2)}{=} (-1)^{n/2}\operatorname{ch}(F)\,\frac{\operatorname{ch}\big(\sum_p (-1)^p[E^p]\big)}{e(X)}\,\operatorname{td}(TX_{\mathbb{C}})[X] \\ &\overset{(3)}{=} \operatorname{ch}(F)\,e(X)\,[X] = \int_X \operatorname{ch}(F)\wedge e(X) \qquad (4.156) \\ &\overset{(4)}{=} \int_X \big(m + c_1(F) + \operatorname{ch}_2(F) + \cdots\big)\wedge e(X) \\ &\overset{(5)}{=} m\int_X e(X) = m\,\chi(X). \qquad (4.157) \end{aligned}$$

(1) The general formula (4.151) for the complex with bundles $F\otimes E^p$.
(2) ch is additive, so $\operatorname{ch}(\sum(-1)^p F\otimes E^p) = \sum(-1)^p \operatorname{ch}(F)\operatorname{ch}(E^p) = \operatorname{ch}(F)\operatorname{ch}(\sum(-1)^p E^p)$ by (4.152).
(3) The remaining factor is exactly the untwisted integrand, which (4.155) simplifies to $e(X)$. This is Nash's (4.116).
(4) Expand $\operatorname{ch}(F)$ by degree; $m = \operatorname{rk} F$.
(5) $e(X)$ is already a form of top degree n; wedging it with anything of positive degree gives a form of degree $> n$, which vanishes on an n-manifold. Only the degree-0 part m of $\operatorname{ch}(F)$ survives. This is Nash's (4.117).

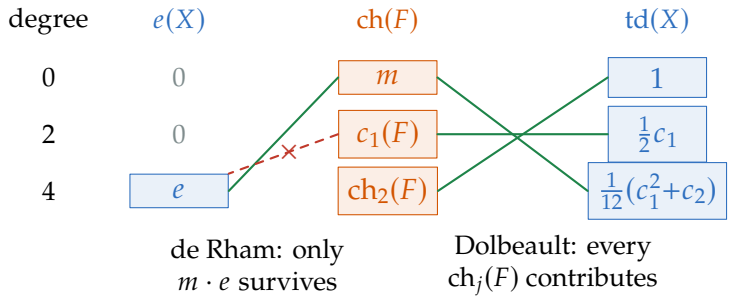


Figure 4.56: Degree bookkeeping on a real 4-manifold. We need products of total degree 4. On the left, the Euler class sits entirely in degree 4, so it can only be paired with the degree-0 part m of $\mathrm{ch}(F)$ (green line; the dashed red pairing would have degree 6 and vanishes). On the right, the Todd class has pieces in every degree, so every component of $\mathrm{ch}(F)$ finds a partner. This is why twisting changes nothing essential for de Rham but a lot for Dolbeault.

Physical picture: Why a gauge field cannot change the Euler characteristic

The de Rham index counts (with signs) the harmonic forms. Twisting by a rank-m bundle is, as far as topology is concerned, like taking m copies of the forms. The individual cohomology groups of the twisted complex can be wildly different from m copies of the untwisted ones (see the examples), but the alternating sum is rigid.

Worked example 4.95: A flat line bundle on the circle

Let $X = S^1$ with coordinate $\theta \in [0, 1)$ and let F be the trivial line bundle with the flat connection $A = 2\pi i a \, \mathrm{d}\theta$, $a \in \mathbb{R}$. Its holonomy around the circle is $\exp(-\oint A) = e^{-2\pi i a}$. The twisted complex is $\Omega^0 \xrightarrow{d_A} \Omega^1$ with $d_A f = (f' + 2\pi i a f)\,\mathrm{d}\theta$. Expand $f = \sum_n f_n e^{2\pi i n\theta}$ and $g\,\mathrm{d}\theta = \sum_n g_n e^{2\pi i n\theta}\mathrm{d}\theta$. Then

$$d_A f \overset{(1)}{=} \sum_n 2\pi i\,(n + a)\, f_n\, e^{2\pi i n\theta}\,\mathrm{d}\theta.$$

(1) $\frac{\mathrm{d}}{\mathrm{d}\theta} e^{2\pi i n\theta} = 2\pi i n\, e^{2\pi i n\theta}$, plus the term $2\pi i a f$.

So d_A is diagonal in the Fourier basis with eigenvalue $2\pi i(n + a)$.

- If $a \notin \mathbb{Z}$, every eigenvalue is non-zero: $\ker d_A = 0$, and every g is hit ($f_n = g_n/2\pi i(n + a)$), so $H^1 = 0$. Both cohomologies vanish.
- If $a \in \mathbb{Z}$, the mode $n = -a$ is killed: $H^0 = \mathbb{C}$ and the corresponding g_{-a} cannot be reached, so $H^1 = \mathbb{C}$.

In both cases $\dim H^0 - \dim H^1 = 0 = 1 \cdot \chi(S^1)$. The groups jump as the holonomy passes through 1; their alternating sum does not.

Worked example 4.96: Flat line bundles on the torus

On $T^2 = \mathbb{R}^2/\mathbb{Z}^2$ with coordinates (x, y) take $A = 2\pi i(a\,\mathrm{d}x + b\,\mathrm{d}y)$. Exactly as on the circle, d_A acts on the Fourier mode $e^{2\pi i(mx+ny)}$ by multiplying the $\mathrm{d}x$ and $\mathrm{d}y$ components by $2\pi i(m + a)$ and $2\pi i(n + b)$. The twisted Laplacian $\Delta_A = d_A d_A^* + d_A^* d_A$ therefore acts on every mode, in every form degree, by the number

$$4\pi^2\big[(m + a)^2 + (n + b)^2\big].$$

For $(a,b) \notin \mathbb{Z}^2$ this never vanishes, so there are no harmonic forms at all: $h^0 = h^1 = h^2 = 0$. For $(a,b) \in \mathbb{Z}^2$ one mode is harmonic in each form component: $h^0 = 1$, $h^1 = 2$ (the $\mathrm{d}x$ and $\mathrm{d}y$ components), $h^2 = 1$. In both cases

$$h^0 - h^1 + h^2 = 0 = 1 \cdot \chi(T^2),$$

as (4.157) predicts. This example is worth remembering: the parameter (a,b) lives on a torus of its own, and the kernel jumps at one point of it. We will meet this "parameter torus" again when we discuss families in Section 4.18.

Worked example 4.97: The monopole bundle on S^2: kernel grows, index does not

Now take $X = S^2$ and $F = L_k$, the charge-k monopole bundle of Example 4.94. It is not flat, so we work with the operator $D_A = d_A + d_A^* : \Omega^0(L_k) \oplus \Omega^2(L_k) \to \Omega^1(L_k)$. The prediction (4.157) is $\operatorname{index} D_A = 1 \cdot \chi(S^2) = 2$ for every k.

On a surface there is a neat way to see this. Use $\star$ to write a 2-form as $g\,\mathrm{vol}$, with g a function (section of L_k). Use a local complex coordinate $z = x + iy$ and split $d_A = \partial_A + \bar\partial_A$ into its $\mathrm{d}z$ and $\mathrm{d}\bar z$ parts. With orientation $\mathrm{d}x \wedge \mathrm{d}y$ we have $\star \mathrm{d}z = -i\,\mathrm{d}z$ and $\star \mathrm{d}\bar z = i\,\mathrm{d}\bar z$, and on 2-forms in two dimensions $d_A^* = - \star d_A \star$. Then

$$\begin{aligned} D_A(f, g\,\mathrm{vol}) &\overset{(1)}{=} d_A f - \star d_A g \\ &\overset{(2)}{=} \partial_A f + \bar\partial_A f - \left(-i\,\partial_A g + i\,\bar\partial_A g\right) \\ &\overset{(3)}{=} \partial_A (f + ig) + \bar\partial_A (f - ig). \end{aligned}$$

(1) $d_A^*(g\,\mathrm{vol}) = - \star d_A \star (g\,\mathrm{vol}) = - \star d_A g$.
(2) Split $d_A g = \partial_A g + \bar\partial_A g$ and apply the Hodge star to each piece.
(3) Collect the $\mathrm{d}z$ and $\mathrm{d}\bar z$ components.

So D_A is two decoupled operators: $\bar\partial_A$ acting on $u = f - ig$ and ∂_A acting on $v = f + ig$. The first is the Cauchy–Riemann operator of the holomorphic line bundle $\mathcal{O}(k)$; the second, after complex conjugation, is the Cauchy–Riemann operator of $\overline{L_k} = L_{-k}$, i.e. of $\mathcal{O}(-k)$. Anticipating the count of holomorphic sections in Example 4.98 below (index $\bar\partial_{\mathcal{O}(j)} = j + 1$):

$$\operatorname{index} D_A = \underbrace{(k+1)}_{\mathcal{O}(k)} + \underbrace{(-k+1)}_{\mathcal{O}(-k)} = 2.$$

For $k > 0$ the kernel of D_A has dimension $k + 1$ and the cokernel $k - 1$: both grow with the magnetic charge, but their difference is always 2.

4.17.3 The twisted Dolbeault complex and Riemann–Roch

Now let X be a complex manifold of complex dimension ℓ ($n = 2\ell$ real dimensions) with local holomorphic coordinates $z^1, \dots, z^\ell$. The Dolbeault complex uses forms of type $(0,p)$, i.e. built from $\mathrm{d}\bar z^{j_1} \wedge \cdots \wedge \mathrm{d}\bar z^{j_p}$, with the operator $\bar\partial = \mathrm{d}\bar z^j\, \partial/\partial \bar z^j$. Its cohomology $H^{0,p}(X)$ has dimensions $h^{0,p}$ and its index is the arithmetic genus $\sum_p (-1)^p h^{0,p} = \int_X \mathrm{td}(X)$ (Nash's (4.72)).

Twisting by a *holomorphic* vector bundle F (one whose transition functions are holomorphic) gives $(0,p)$-forms with values in F, which Nash writes $\Omega^p(F)$ (Chapter 3 called them the sheaf of p-forms with coefficients in F). The operator $\bar\partial_F$ acts componentwise in any holomorphic trivialisation; this is consistent because holomorphic transition functions are annihilated by

$\bar\partial$. Now $\bar\partial_F^2 = 0$ honestly, so this is a genuine complex, with cohomology groups $H^{0,p}(X;F) = H^p(X;\mathcal{O}(F))$, the sheaf cohomology of the sheaf of holomorphic sections of F.

The calculation is identical in shape to (4.156):

$$\begin{aligned}\operatorname{index}\bar\partial_F &\overset{(1)}{=} (-1)^\ell \frac{\operatorname{ch}\big(\sum_p (-1)^p [F\otimes\Lambda^{0,p}]\big)}{e(X)}\ \operatorname{td}(TX_{\mathbb{C}})[X]\\ &\overset{(2)}{=} \operatorname{ch}(F)\,(-1)^\ell \frac{\operatorname{ch}\big(\sum_p (-1)^p [\Lambda^{0,p}]\big)}{e(X)}\ \operatorname{td}(TX_{\mathbb{C}})[X]\\ &\overset{(3)}{=} \int_X \operatorname{ch}(F)\wedge\operatorname{td}(X) = \chi(X,F).\end{aligned} \tag{4.158}$$

(1) General formula (4.151) for the bundles $E^p = F\otimes\Lambda^{0,p}$.
(2) Additivity and multiplicativity of ch.
(3) The untwisted combination reduces to $\operatorname{td}(X)$, the Todd class of the holomorphic tangent bundle (Nash's (4.70)–(4.71)); and the alternating sum of dimensions of sheaf cohomology is by definition the **holomorphic Euler characteristic** $\chi(X,F) = \sum_p (-1)^p \dim H^p(X;\mathcal{O}(F))$.

Equation (4.158) is Nash's (4.118): the **Hirzebruch–Riemann–Roch theorem**. Unlike the de Rham case, $\operatorname{td}(X)$ has components in every even degree (Figure 4.56), so every Chern class of F contributes.

Complex dimension one: the classical Riemann–Roch theorem

Let X be a compact Riemann surface of genus g and F a holomorphic bundle of rank r and degree $\deg F = \int_X c_1(F)$. Then

$$\begin{aligned}\operatorname{index}\bar\partial_F &\overset{(1)}{=} \int_X \big(r + c_1(F)\big)\wedge\big(1+\tfrac12 c_1(X)\big)\\ &\overset{(2)}{=} \int_X c_1(F) + \frac{r}{2}\int_X c_1(X)\\ &\overset{(3)}{=} \deg F + \frac{r}{2}(2-2g) = \deg F + r(1-g).\end{aligned} \tag{4.159}$$

(1) $\operatorname{ch}(F) = r + c_1(F)$ and $\operatorname{td}(X) = 1 + \frac12 c_1(X)$, truncated at real degree 2 because X has no forms of higher degree.
(2) Keep only the 2-form part: $r\cdot\frac12 c_1(X) + c_1(F)\cdot 1$.
(3) For a curve $c_1(X) = c_1(TX) = e(X)$, and Gauss–Bonnet gives $\int_X e(X) = \chi(X) = 2-2g$.

In words: $\dim H^0(X;F) - \dim H^1(X;F) = \deg F + r(1-g)$. This is the classical Riemann–Roch theorem, now seen as an index theorem.

Worked example 4.98: Holomorphic sections of $\mathcal{O}(k)$ on $\mathbb{CP}^1$

On $\mathbb{CP}^1 = S^2$ use the two coordinates z (near 0) and $w = 1/z$ (near ∞). The line bundle $\mathcal{O}(k)$ is defined by the transition function z^k: a section is a pair of holomorphic functions $s_0(z)$ and $s_\infty(w)$ with $s_\infty(w) = w^k s_0(1/w)$. It is the holomorphic version of the monopole bundle L_k, with $\deg\mathcal{O}(k) = k$.

Direct count. Let $s_0(z) = \sum_{j\ge 0} a_j z^j$ be entire. Then $s_\infty(w) = \sum_j a_j w^{k-j}$ must be holomorphic at $w=0$, which forces $a_j = 0$ for $j > k$. Hence the global holomorphic sections are exactly the polynomials of degree $\le k$:

$$H^0(\mathbb{CP}^1;\mathcal{O}(k)) = \operatorname{span}\{1, z, z^2, \dots, z^k\}, \qquad h^0 = k+1 \quad (k\ge 0).$$

One can show $h^1 = 0$ for $k \geq -1$ (and by Serre duality $h^1(\mathcal{O}(k)) = h^0(\mathcal{O}(-2-k))$ in general). *Index theorem.* With $g = 0$, $r = 1$, (4.159) gives index $\bar\partial_{\mathcal{O}(k)} = k + 1$, matching the count. For $k \leq -2$ there are no sections, and the index $k + 1 < 0$ is carried entirely by $h^1 = -k - 1$.

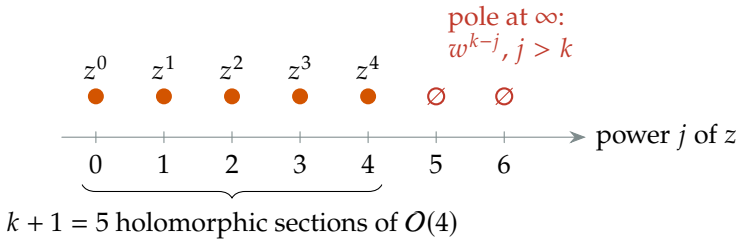


Figure 4.57: Global holomorphic sections of $\mathcal{O}(k)$ on $\mathbb{CP}^1$, here $k = 4$. In the z-chart a section is a power series $\sum a_j z^j$; in the $w = 1/z$ chart it becomes $\sum a_j w^{k-j}$, which is regular at $w = 0$ only if $j \leq k$ (orange dots). The index $k + 1$ is literally the number of allowed monomials.

Physical picture: Lowest Landau level = index

A charged particle on a torus T^2 in a uniform magnetic field with N flux quanta has a lowest Landau level of degeneracy exactly N. The lowest-Landau-level wave functions are the zero modes of $\bar\partial_A$, i.e. holomorphic sections of a degree-N line bundle (they are theta functions). Riemann–Roch with $g = 1$ gives index $= N + 1 \cdot (1 - 1) = N$, and one checks $h^1 = 0$ for $N > 0$. The celebrated "degeneracy = number of flux quanta" is a twisted index theorem.

Complex dimension two

Nash works out the next case [Nash p. 118, (4.119)–(4.120)]. On a complex surface ($\ell = 2$, real dimension 4) we need everything up to real degree 4:

$$\operatorname{ch}(F) = \operatorname{rk} F + c_1(F) + \tfrac{1}{2}\big(c_1^2(F) - 2c_2(F)\big) + \cdots,$$
$$\operatorname{td}(X) = 1 + \tfrac{1}{2}c_1(X) + \tfrac{1}{12}\big(c_1^2(X) + c_2(X)\big) + \cdots.$$

Multiplying and keeping the degree-4 part:

$$\begin{aligned}\operatorname{index}\bar\partial_F &\overset{(1)}{=} \int_X \Big[\operatorname{rk} F \cdot \tfrac{1}{12}(c_1^2 + c_2)(X) + c_1(F)\cdot\tfrac{1}{2}c_1(X) + \tfrac{1}{2}\big(c_1^2(F) - 2c_2(F)\big)\cdot 1\Big] \\ &\overset{(2)}{=} \frac{\operatorname{rk} F}{12}\int_X \big(c_1^2(X) + c_2(X)\big) + \frac{1}{2}\int_X \big(c_1^2(F) - 2c_2(F) + c_1(F)c_1(X)\big).\end{aligned} \tag{4.160}$$

(1) The three ways to make degree 4: (degree 0 of ch)×(degree 4 of td), (degree 2)×(degree 2), (degree 4)×(degree 0).
(2) Regroup; this is Nash's (4.120).

For F trivial of rank 1 this gives Noether's formula $\chi(X, \mathcal{O}) = \frac{1}{12}\int (c_1^2 + c_2)$.

Worked example 4.99: $\mathbb{CP}^2$ and homogeneous polynomials

Let h be the hyperplane class of $\mathbb{CP}^2$, normalised by $\int_{\mathbb{CP}^2} h^2 = 1$. The total Chern class of $T\mathbb{CP}^2$ is $(1 + h)^3$, so $c_1(X) = 3h$ and $c_2(X) = 3h^2$. The bundle $\mathcal{O}(k)$ has $c_1 = kh$, $c_2 = 0$, rank

1. Then

$$\operatorname{index}\bar\partial_{\mathcal{O}(k)} \overset{(1)}{=} \frac{1}{12}(9+3)+\frac{1}{2}\big(k^2+3k\big) \overset{(2)}{=} 1+\frac{k^2+3k}{2}=\frac{(k+1)(k+2)}{2}.$$

(1) Insert into (4.160): $c_1^2(X)+c_2(X)=9h^2+3h^2$, and $c_1^2(F)+c_1(F)c_1(X)=k^2h^2+3kh^2$; integrate using $\int h^2=1$.

(2) Arithmetic: $2+k^2+3k=(k+1)(k+2)$.

The holomorphic sections of $\mathcal{O}(k)$ on $\mathbb{CP}^2$ are the homogeneous polynomials of degree k in (z_0,z_1,z_2), and there are $\binom{k+2}{2}=\frac{(k+1)(k+2)}{2}$ monomials $z_0^a z_1^b z_2^c$ with $a+b+c=k$. For $k\geq 0$ the higher cohomology vanishes, and the index theorem reproduces the count exactly. For $k=0$ we get 1: the only holomorphic functions are constants.

Try it: A K3 surface

A K3 surface has $c_1=0$ and $\int c_2=\chi=24$. Compute $\chi(X,\mathcal{O})=h^{0,0}-h^{0,1}+h^{0,2}$.
Answer: Noether gives $\frac{1}{12}(0+24)=2$; indeed $h^{0,0}=1$, $h^{0,1}=0$, $h^{0,2}=1$ (the holomorphic 2-form).

4.17.4 The twisted signature complex

Recall the signature operator. On an oriented Riemannian n-manifold (n even) there is an involution τ on complex forms, $\tau=i^{p(p-1)+n/2}\star$ on p-forms, with $\tau^2=1$. Its ±1 eigenspaces $\Omega^{\pm}$ are exchanged by $d+d^*$, and the **signature operator** is $d_+=d+d^*:\Omega^+\to\Omega^-$. When $n\equiv 0 \bmod 4$, $\operatorname{index} d_+=\operatorname{Sign}(X)$, the signature of the intersection form on $H^{n/2}$. Twisting by F with a connection gives $d_+^F=d_A+d_A^*:\Omega^+(F)\to\Omega^-(F)$.

The index calculation needs $\operatorname{ch}\big(\Lambda^+-\Lambda^-\big)$. Nash showed [Nash p. 106, (4.91)] that $\operatorname{ch}(\Lambda^+-\Lambda^-)=\prod_i(e^{-x_i}-e^{x_i})$ up to sign, and inserting it into (4.151) with the extra factor $\operatorname{ch}(F)$ one finds

$$\operatorname{index} d_+^F=\int_X \operatorname{ch}(F)\wedge\prod_{i=1}^{n/2}\frac{x_i}{\tanh(x_i/2)}. \tag{4.161}$$

Nash writes this as $2^{n/2}\int_X \operatorname{ch}(F)\wedge\mathcal{L}(X)$ [Nash p. 118, (4.121)].

Watch out: Which "L" is Nash's $\mathcal{L}$?

Pull a factor of 2 out of each root: $\frac{x}{\tanh(x/2)}=2\cdot\frac{x/2}{\tanh(x/2)}$. So (4.161) equals $2^{n/2}\int\operatorname{ch}(F)\,\mathcal{L}$ exactly when $\mathcal{L}=\prod_i\frac{x_i/2}{\tanh(x_i/2)}$. This is the normalisation Nash uses (it is what makes his (4.93)–(4.95) consistent). The Hirzebruch polynomial $L=\prod_i\frac{x_i}{\tanh x_i}$ is different. The two agree in top degree after the factor $2^{n/2}$, because the degree-$4j$ part of $\mathcal{L}$ is $2^{-2j}L_j$ and for $4j=n$ this is $2^{-n/2}L_{n/4}$. But the *lower*-degree parts, which pair with $\operatorname{ch}(F)$, differ by powers of 2. The safe form is (4.161), or equivalently

$$\operatorname{index} d_+^F=\sum_{j\geq 0}2^{\,n/2-2j}\int_X \operatorname{ch}_{n/2-2j}(F)\wedge L_j(X).$$

Let us expand (4.161) in low dimensions. We need the power series

$$\frac{x}{\tanh(x/2)} \overset{(1)}{=} 2 \cdot \frac{y}{\tanh y}\Big|_{y=x/2} \overset{(2)}{=} 2\Big(1 + \frac{y^2}{3} - \frac{y^4}{45} + \cdots\Big)_{y=x/2}$$
$$= 2 + \frac{x^2}{6} - \frac{x^4}{360} + \cdots . \tag{4.162}$$

(1) Write $x = 2y$.

(2) $\tanh y = y - \frac{y^3}{3} + \frac{2y^5}{15} - \cdots$, so $\frac{y}{\tanh y} = \big(1 - \frac{y^2}{3} + \frac{2y^4}{15}\big)^{-1} = 1 + \frac{y^2}{3} - \frac{y^4}{45} + \cdots$.

Dimension 2 ($n = 2$, one root x). The degree-2 part of $\mathrm{ch}(F)\,(2 + \frac{x^2}{6} + \cdots)$ is $2c_1(F)$ (note x^2 has degree 4, too high). So

$$\operatorname{index} d_+^F = 2\int_X c_1(F) \qquad (n = 2). \tag{4.163}$$

Dimension 4 ($n = 4$, roots x_1, x_2). Here $\prod_{i=1}^2 (2 + \frac{x_i^2}{6}) = 4 + \frac{1}{3}(x_1^2 + x_2^2) + \cdots = 4 + \frac{1}{3}p_1 + \cdots$, so

$$\operatorname{index} d_+^F = 4\int_X \mathrm{ch}_2(F) + \operatorname{rk} F \cdot \frac{1}{3}\int_X p_1 = 4\int_X \mathrm{ch}_2(F) + \operatorname{rk} F \cdot \operatorname{Sign}(X). \tag{4.164}$$

For trivial F of rank 1 we recover Hirzebruch's $\operatorname{Sign}(X) = \frac{1}{3}\int p_1$.

Worked example 4.100: Twisted signature operator on S^2 and S^4

S^2 with the monopole bundle L_k. The untwisted signature operator in dimension 2 has index 0 (there is no signature in dimension 2). By (4.163) and Example 4.94:

$$\operatorname{index} d_+^{L_k} = 2\int_{S^2} c_1(L_k) = 2k \neq 0.$$

S^4 with an SU(2) *instanton.* Let F be the rank-2 bundle of a charge-k instanton; $c_1(F) = 0$ and $\int c_2(F) = k$ (this fixes our orientation convention). Then $\mathrm{ch}_2(F) = \frac{1}{2}(c_1^2 - 2c_2) = -c_2$ and $\operatorname{Sign}(S^4) = 0$, so (4.164) gives

$$\operatorname{index} d_+^F = 4 \cdot (-k) + 2 \cdot 0 = -4k.$$

Nash's remark [Nash p. 118] that $\operatorname{index} d_+^F$ can be non-zero in any even dimension, unlike $\operatorname{index} d_+$, is illustrated by the first case.

4.17.5 The twisted spin complex: Dirac operators in gauge fields

This is the case physicists care about most. Let X be a spin manifold of even dimension n with spinor bundle $E = E^+ \oplus E^-$ (split by the chirality operator γ_5). A Dirac fermion coupled to a Yang–Mills field in the representation carried by F is a section of $E \otimes F$, and the covariant Dirac operator is, in local coordinates [Nash p. 118],

$$\not{D}_F = \gamma^a e_a{}^\mu(x)\big(\partial_\mu + \omega_\mu + A_\mu\big), \qquad \not{D}_F^+ = \not{D}_F\,\tfrac{1}{2}(1 + \gamma_5) : \Gamma(E^+ \otimes F) \to \Gamma(E^- \otimes F). \tag{4.165}$$

Here $e_a{}^\mu$ is the vielbein, ω_μ the spin connection and A_μ the Yang–Mills connection (an anti-Hermitian matrix in the representation of F).

An explicit representation in two dimensions

Let us be concrete. In two Euclidean dimensions take

$$\gamma^1 = \sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \gamma^2 = \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \qquad \gamma_5 = -i\gamma^1\gamma^2 = \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \tag{4.166}$$

One checks $(\gamma^1)^2 = (\gamma^2)^2 = 1$, $\gamma^1\gamma^2 = i\sigma_3 = -\gamma^2\gamma^1$, and γ_5 anticommutes with both. (Nash normalises $\{\gamma_a, \gamma_b\} = -2\delta_{ab}$, which amounts to multiplying our γ's by i; this and the choice of orientation are the source of his sign $(-1)^{n/2}$ below.) On flat $\mathbb{R}^2$ with complex coordinate $z = x + iy$, $\partial_z = \tfrac{1}{2}(\partial_x - i\partial_y)$, $\partial_{\bar z} = \tfrac{1}{2}(\partial_x + i\partial_y)$:

$$\not{D} \overset{(1)}{=} \sigma_1\partial_x + \sigma_2\partial_y \overset{(2)}{=} \begin{pmatrix} 0 & \partial_x - i\partial_y \\ \partial_x + i\partial_y & 0 \end{pmatrix} \overset{(3)}{=} \begin{pmatrix} 0 & 2\partial_z \\ 2\partial_{\bar z} & 0 \end{pmatrix}. \tag{4.167}$$

(1) Flat space: ${e_a}^\mu = \delta_a^\mu$, $\omega = 0$.
(2) Insert the Pauli matrices (4.166).
(3) Definitions of ∂_z and $\partial_{\bar z}$.

So on $\psi = (\psi_+, \psi_-)^T$, $\not{D}^+\psi_+ = 2\partial_{\bar z}\psi_+$: a positive-chirality zero mode is a holomorphic function, and $\not{D}^-\psi_- = 2\partial_z\psi_-$: a negative-chirality zero mode is antiholomorphic. With a gauge field, $\partial_{\bar z} \to \partial_{\bar z} + A_{\bar z}$.

The index formula

For the twisted spin complex the same manipulation as before gives [Nash p. 118, (4.122)]

$$\operatorname{index} \not{D}_F^+ = \pm \int_X \operatorname{ch}(F) \wedge \hat{A}(X), \tag{4.168}$$

where Nash's sign is $(-1)^{n/2}$. With the representation (4.166), the orientation $\mathrm{d}x \wedge \mathrm{d}y$ and the gauge convention of the Dictionary above, the sign is $+$, as the explicit computation below confirms; we use $+$ from now on.

Key idea: What twisting does to the Dirac index

$\hat{A}(X) = 1 - \frac{1}{24}p_1 + \cdots$ has a non-zero degree-0 term. So every component $\operatorname{ch}_j(F)$ of the gauge bundle contributes. In particular in dimension n the pure gauge term is $\int_X \operatorname{ch}_{n/2}(F)$: *in flat space, the index of the Dirac operator in a gauge field is the $(n/2)$-th Chern character number of the gauge field.* In $n = 2$ this is the magnetic flux over 2π; in $n = 4$ it is minus the instanton number.

Worked example 4.101: Aharonov–Casher zero modes in the plane

Let a U(1) magnetic field $B(x, y) \geq 0$ with total flux $\Phi = \int B\,\mathrm{d}x\,\mathrm{d}y$ be localised near the origin of $\mathbb{R}^2$. Choose a potential ϕ with $\nabla^2\phi = B$, and the gauge $a = (-\partial_y\phi, \partial_x\phi)$, so $\partial_x a_y - \partial_y a_x = \nabla^2\phi = B$. In our convention $A = -ia$. Then

$$A_{\bar z} \overset{(1)}{=} \tfrac{1}{2}(A_x + iA_y) \overset{(2)}{=} -\tfrac{i}{2}\big(-\partial_y\phi + i\partial_x\phi\big) \overset{(3)}{=} \tfrac{1}{2}(\partial_x + i\partial_y)\phi = \partial_{\bar z}\phi,$$

$$\not{D}_A^+\psi_+ \overset{(4)}{=} 2(\partial_{\bar z} + \partial_{\bar z}\phi)\psi_+ \overset{(5)}{=} 2e^{-\phi}\,\partial_{\bar z}\big(e^{\phi}\psi_+\big).$$

(1) $A_{\bar z}\,\mathrm{d}\bar z + A_z\,\mathrm{d}z = A_x\mathrm{d}x + A_y\mathrm{d}y$ with $\mathrm{d}\bar z = \mathrm{d}x - i\mathrm{d}y$ gives $A_{\bar z} = \tfrac{1}{2}(A_x + iA_y)$.
(2) $A = -ia$ with $a_x = -\partial_y\phi$, $a_y = \partial_x\phi$.

(3) Multiply out: $-\frac{i}{2}\cdot i\,\partial_x\phi = \frac{1}{2}\partial_x\phi$ and $-\frac{i}{2}(-\partial_y\phi) = \frac{i}{2}\partial_y\phi$.
(4) The positive-chirality block of (4.167) with $\partial_{\bar z} \to \partial_{\bar z} + A_{\bar z}$.
(5) Product rule: $\partial_{\bar z}(e^{\phi}\psi) = e^{\phi}(\partial_{\bar z}\psi + \psi\,\partial_{\bar z}\phi)$.

So the zero modes are $\psi_+ = e^{-\phi}f(z)$ with f entire. Far away, $\phi \approx \frac{\Phi}{2\pi}\ln r$ (the 2D Green's function: $\nabla^2 \ln r = 2\pi\delta^2$), so $e^{-\phi} \approx r^{-\Phi/2\pi}$. Taking $f = z^j$, the mode is square-integrable iff $\int^{\infty} r^{2j-\Phi/\pi}\, r\,\mathrm{d}r < \infty$, i.e. $j < \frac{\Phi}{2\pi} - 1$. The analogous computation for $\psi_- = e^{\phi}\overline{g(z)}$ gives nothing normalisable. Therefore

$$n_+ = \#\Big\{j \in \mathbb{Z}_{\geq 0} :\ j < \tfrac{\Phi}{2\pi} - 1\Big\} \approx \frac{\Phi}{2\pi}, \qquad n_- = 0.$$

All zero modes have the same chirality, and their number is the flux in units of 2π, i.e. $\int c_1$, up to a boundary effect of at most one that comes from the plane not being compact.

Worked example 4.102: Dirac operator on S^2 in a monopole field

On the closed sphere the boundary ambiguity disappears. With $\hat{A}(S^2) = 1$ (there are no degree-4 forms), (4.168) and Example 4.94 give

$$\operatorname{index} \not{D}^{+}_{L_k} \overset{(1)}{=} \int_{S^2} \big(1 + c_1(L_k)\big) \wedge 1 \overset{(2)}{=} \int_{S^2} c_1(L_k) = k.$$

(1) $\operatorname{ch}(L_k) = 1 + c_1(L_k)$, $\hat{A}(S^2) = 1$.
(2) Only the 2-form part integrates; $\int c_1 = k$.

We can confirm this by counting. On a Riemann surface the positive spinors are $E^+ = K^{1/2}$, a square root of the canonical bundle ($K = \mathcal{O}(-2)$ on $\mathbb{CP}^1$, so $K^{1/2} = \mathcal{O}(-1)$), and $\not{D}^{+}_{L_k} = \sqrt{2}\,\bar\partial$ acting on sections of $K^{1/2} \otimes L_k = \mathcal{O}(k-1)$. By Example 4.98 for $k \geq 1$ there are exactly k holomorphic sections (polynomials of degree $\leq k-1$) and no cokernel. So there are k zero modes, all of positive chirality. Note the pleasant cross-check between two index theorems: Dirac on L_k equals Dolbeault on $\mathcal{O}(k-1)$, and $k = (k-1) + 1$.

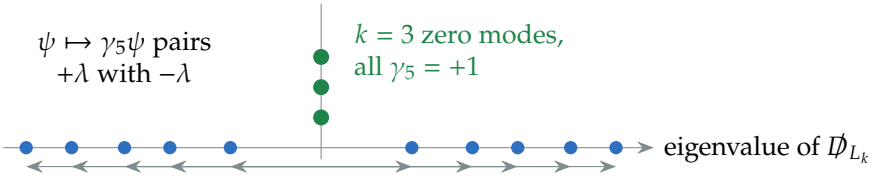


Figure 4.58: Spectrum of the Dirac operator on S^2 coupled to a charge-3 monopole. Because γ_5 anticommutes with $\not{D}$, non-zero eigenvalues come in pairs $\pm\lambda$ (blue), each pair mixing both chiralities. The zero modes (green) are unpaired, and they all have positive chirality: $\operatorname{index} \not{D}^{+} = n_+ - n_- = 3$.

Worked example 4.103: Instantons and fermion zero modes in four dimensions

On S^4 (or $\mathbb{R}^4$ with finite action) $p_1 = 0$, so $\hat{A}(S^4) = 1$ and the untwisted Dirac operator has index 0. Couple it to a charge-k SU(2) instanton in the fundamental representation ($\int c_2 = k$, $c_1 = 0$):

$$\operatorname{index} \not{D}^{+}_{F} \overset{(1)}{=} \int_{S^4} \operatorname{ch}_2(F) \overset{(2)}{=} \int_{S^4} \tfrac{1}{2}\big(c_1^2 - 2c_2\big)(F) \overset{(3)}{=} -k.$$

(1) $\hat{A} = 1$ on S^4; the degree-4 part of $\operatorname{ch}(F)\cdot 1$.

(2) $\mathrm{ch}_2 = \frac{1}{2}(c_1^2 - 2c_2)$.
(3) $c_1 = 0$ and $\int c_2 = k$.

One can show the kernel of $\not{D}_F^+$ vanishes for a self-dual connection, so there are exactly k zero modes, all of negative chirality. These are 't Hooft's zero modes, responsible for the instanton-induced violation of axial charge. For the adjoint representation of SU(2) one finds $\int \mathrm{ch}_2(\mathrm{ad}) = -4k$: $4k$ gluino zero modes.

Worked example 4.104: When twisting creates an index where there was none

Nash notes [Nash pp. 118–119] two ways in which $\not{D}_F$ differs from $\not{D}$:

- In dimensions $n \not\equiv 0 \bmod 4$, $\int \hat{A} = 0$ automatically (it is a polynomial in the p_i, of degree divisible by 4), so index $\not{D} = 0$. But index $\not{D}_F \neq 0$ is possible: S^2 with a monopole, index k.
- Even when $4 \mid n$, one may have index $\not{D} = 0$ but index $\not{D}_F \neq 0$: S^4 with an instanton, index $-k$.

For contrast, an untwisted index that is non-zero: for a K3 surface $p_1 = -48$, so index $\not{D} = \int \hat{A} = -\frac{1}{24}(-48) = 2$.

complex	untwisted index	twisted index	does F matter?
de Rham	$\int e(X) = \chi(X)$	$\int \mathrm{ch}(F)\, e(X) = m\,\chi(X)$	only its rank
Dolbeault	$\int \mathrm{td}(X)$	$\int \mathrm{ch}(F)\ \mathrm{td}(X) = \chi(X, F)$	yes, all $c_j(F)$
signature	$\int L(X) = \mathrm{Sign}\, X$	$\int \mathrm{ch}(F) \prod \frac{x_i}{\tanh(x_i/2)}$	yes, all $c_j(F)$
spin (Dirac)	$\int \hat{A}(X)$	$\int \mathrm{ch}(F)\, \hat{A}(X)$	yes, all $c_j(F)$

Figure 4.59: The four classical complexes and their twisted versions. In every case twisting multiplies the integrand by $\mathrm{ch}(F)$. Only for de Rham is the untwisted integrand concentrated in top degree, which is why only there the twist is invisible (apart from the factor $m = \mathrm{rk}\, F$).

Reading Nash (pp. 116–119, (4.112)–(4.122))

Nash introduces matrix-valued forms $\Omega^p(X; F) = \Gamma(X, F \otimes \Lambda^p T^* X_{\mathbb{C}})$ and the twisted de Rham complex (4.113). He then observes that in the index formula (4.114) the only change is $E^p \to F \otimes E^p$ and uses $\mathrm{ch}(A \otimes B) = \mathrm{ch}\, A\ \mathrm{ch}\, B$ to get (4.115)–(4.117): index $d_F = m\,\chi(X)$. He repeats the argument for Dolbeault (4.118), expands in complex dimension 2 (4.119)–(4.120), states the twisted signature (4.121) and Dirac (4.122) formulas, and remarks that these twisted indices can be non-zero in dimensions where the untwisted ones vanish. What we added: the covariance of d_A, the fact that $d_A^2 = F_A$ (so (4.113) is a complex only for flat F), the explicit Fourier-mode examples, the exact normalisation of $\mathcal{L}$ in (4.121), and the explicit zero-mode counts.

Recap

- Twisting = tensoring every bundle in a complex with a gauge bundle F and replacing d by the covariant d_A.

- Because ch is multiplicative, the index integrand is multiplied by $\mathrm{ch}(F)$.
- de Rham: $\operatorname{index} d_A = \operatorname{rk} F \cdot \chi(X)$; the gauge field is invisible since $e(X)$ is top degree.
- Dolbeault: Hirzebruch–Riemann–Roch $\chi(X,F) = \int \mathrm{ch}(F)\,\mathrm{td}(X)$; on curves $\deg F + r(1-g)$; on $\mathbb{CP}^1$, $k+1$ sections of $\mathcal{O}(k)$.
- Dirac: $\operatorname{index} \not{D}_F^+ = \int \mathrm{ch}(F)\hat{A}(X)$; the number of fermion zero modes is the magnetic flux (2D) or instanton number (4D).

4.18 The index theorem for families of operators

So far we have always had *one* elliptic operator and computed *one* integer. Physics rarely hands us a single operator. The Dirac operator of a fermion depends on the background gauge field A; as A varies over the space $\mathcal{A}$ of all gauge fields we get a whole *family* $\{\not{D}_A\}_{A\in\mathcal{A}}$. The Laplacian on a torus depends on the shape of the torus; as the shape varies over the moduli space we again get a family. The Hamiltonian of an electron in a crystal depends on the Bloch momentum k in the Brillouin zone. In each case there is a **parameter space** Y, and for each $y \in Y$ an operator D_y.

The question of this section is:

> *For a family of elliptic operators, is there a topological invariant that is richer than the (constant) integer* $\operatorname{index} D_y$*? And can we compute it?*

The answer is yes. The kernels and cokernels of the D_y fit together into (virtual) vector bundles over Y, and their difference is an element $\operatorname{index} D$ of the K-theory $K(Y)$. The integer index is only the rank of this class. The rest of it, for example its Chern classes, carries information that no single operator can see. The **families index theorem** of Atiyah and Singer computes it by a formula that is the single-operator formula with the final integral over X replaced by an integration along the fibres. In physics this is the machinery behind the geometry of the fermion determinant, the Berry phase of the lowest Landau level, and the topological analysis of gauge anomalies in Chapter 10.

4.18.1 A family of operators and what varies with it

Working definition 4.46 (family of elliptic operators)

A **family of elliptic operators** parametrised by a (compact) space Y is a map $y \mapsto D_y$ that assigns to each point $y \in Y$ an elliptic operator [leading symbol invertible at every nonzero momentum] $D_y : \Gamma(X,E) \to \Gamma(X,F)$ on a fixed closed manifold [compact and without boundary] X, such that the coefficients of D_y (in local coordinates) depend continuously on y. *In practice:* write down the operator with some parameters in it; the set of allowed parameter values is Y.

Because the index is a deformation invariant, $\operatorname{index} D_y$ is the same integer for all y in a connected Y. But the kernel and cokernel themselves vary. Their *dimensions* may jump, and even when the dimensions stay fixed, the *subspaces* $\ker D_y$ rotate inside the space of sections as y moves. It is this rotation that the families index captures.

Worked example 4.105: The simplest jumping kernel

Take X to be a point, so that sections are just vectors, and let $Y = \mathbb{C}$ with the family of 1×1 matrices

$$D_y : \mathbb{C} \to \mathbb{C}, \qquad D_y(v) = y\, v.$$

For $y \neq 0$ this is invertible: $\operatorname{Ker} D_y = 0$, $\operatorname{Coker} D_y = 0$. At $y = 0$ it is the zero map: $\operatorname{Ker} D_0 = \mathbb{C}$, $\operatorname{Coker} D_0 = \mathbb{C}$. So

$$\dim \operatorname{Ker} D_y = \begin{cases} 1, & y = 0, \\ 0, & y \neq 0, \end{cases} \qquad \operatorname{index} D_y = \dim \operatorname{Ker} D_y - \dim \operatorname{Coker} D_y = 0 \text{ for all } y.$$

The kernel jumps, but it jumps *together* with the cokernel, so the index does not move. This is the general mechanism: a zero mode can only appear or disappear in company with a zero mode of the adjoint.

Worked example 4.106: A family on the circle: the Aharonov–Bohm phase

Take $X = S^1$ ($\theta \in [0, 1)$) and the family of Example 4.95,

$$D_a = \frac{\mathrm{d}}{\mathrm{d}\theta} + 2\pi i a : \ \Omega^0(S^1) \to \Omega^0(S^1), \qquad a \in \mathbb{R}.$$

Physically, a is the magnetic flux (in units of the flux quantum) threading a ring on which a charged particle moves. On the Fourier mode $e^{2\pi i n\theta}$ the operator acts by multiplication by $\lambda_n(a) = 2\pi i(n + a)$. The parameter is really periodic: the gauge transformation $g = e^{-2\pi i\theta}$ maps D_a to

$$g^{-1} D_a\, g \overset{(1)}{=} e^{2\pi i\theta}\Big(\frac{\mathrm{d}}{\mathrm{d}\theta} + 2\pi i a\Big)e^{-2\pi i\theta} \overset{(2)}{=} \frac{\mathrm{d}}{\mathrm{d}\theta} - 2\pi i + 2\pi i a = D_{a-1}.$$

(1) Conjugate by the gauge transformation (a unitary operator on $L^2(S^1)$, so kernels and cokernels are carried into each other).
(2) Product rule: $\frac{\mathrm{d}}{\mathrm{d}\theta}(e^{-2\pi i\theta} f) = e^{-2\pi i\theta}(f' - 2\pi i f)$.

So D_a and D_{a-1} are unitarily equivalent, and the genuine parameter space is the circle $Y = \mathbb{R}/\mathbb{Z}$ of Aharonov–Bohm phases $e^{2\pi i a}$. On this circle the kernel is zero except at the single point $a = 0$, where it is spanned by the constants. Again $\operatorname{index} D_a = 0$ for all a.

4.18.2 Reminder: the K-theory of a parameter space

Earlier in the chapter we met K-theory as the natural home of symbols. Here we need it for the parameter space Y, so we recall the essentials.

Working definition 4.47 (virtual bundles and $K(Y)$)

An element of $K(Y)$ is a formal difference $[E] - [F]$ of complex vector bundles over Y (a **virtual bundle**). Two differences are equal, $[E]-[F] = [E']-[F']$, when $E \oplus F' \oplus G \cong E' \oplus F \oplus G$ for some bundle G [G is needed because $E \oplus G \cong E' \oplus G$ need not imply $E \cong E'$]. Addition is $\oplus$, multiplication is $\otimes$. *To detect an element:* compute its rank $\operatorname{rk} E - \operatorname{rk} F$ and its Chern character $\operatorname{ch} E - \operatorname{ch} F \in H^{\text{even}}(Y; \mathbb{Q})$ [cohomology classes of degrees $0, 2, 4, \ldots$, with rational

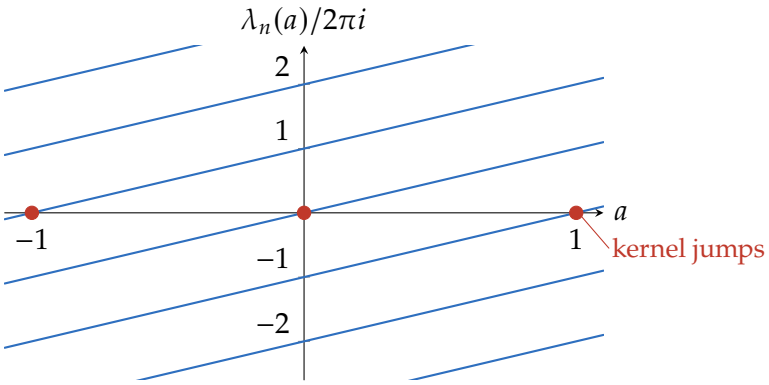


Figure 4.60: The eigenvalues $\lambda_n(a)/2\pi i = n + a$ of the family $D_a = \mathrm{d}/\mathrm{d}\theta + 2\pi i a$ on the circle, one straight line for each Fourier mode n. A zero mode exists exactly when a line crosses the axis (red dots, at $a \in \mathbb{Z}$). The picture is periodic under $a \to a + 1$ (the lines are relabelled $n \to n - 1$), which is why the parameter space is a circle. Although no line stays at zero, one line passes through zero per period: this "one eigenvalue crosses per loop" will reappear as spectral flow in Section 4.21.10.

coefficients].

Two examples fix ideas.

- $K(\mathrm{pt}) = \mathbb{Z}$: a vector bundle over a point is a vector space, classified by its dimension, so $[V] - [W] \mapsto \dim V - \dim W$. *The ordinary index lives in* $K(\mathrm{pt})$.
- $K(S^2) = \mathbb{Z} \oplus \mathbb{Z}$, generated by the trivial line 1 and $[H] - 1$, where H is the monopole (Hopf) line bundle with $\int c_1(H) = \pm 1$. The rank detects the first $\mathbb{Z}$, and $\int_{S^2} c_1$ detects the second. Since $c_1(H)^2 = 0$ on S^2, $([H] - 1)^2 = 0$ in $K(S^2)$.

The Chern character is additive and multiplicative (Eq. (4.152)), so it extends to a ring homomorphism

$$\mathrm{ch} : K(Y) \longrightarrow H^{\mathrm{even}}(Y; \mathbb{Q}), \qquad \mathrm{ch}\big([E] - [F]\big) = \mathrm{ch}\, E - \mathrm{ch}\, F.$$

It loses exactly the torsion part of $K(Y)$ (see Section 4.18.9 below); after tensoring with $\mathbb{Q}$ it is an isomorphism.

4.18.3 The index bundle when the kernel does not jump

Suppose first that $\dim \operatorname{Ker} D_y$ and $\dim \operatorname{Coker} D_y$ are constant in y. Then the kernels form a vector bundle over Y,

$$\operatorname{Ker} D = \bigcup_{y \in Y} \operatorname{Ker} D_y \longrightarrow Y,$$

whose fibre over y is the finite-dimensional vector space $\operatorname{Ker} D_y$. (Constancy of the dimension is what makes the kernels vary smoothly: the orthogonal projector onto $\operatorname{Ker} D_y$ can be written as a contour integral of the resolvent of $D_y^* D_y$ around 0, and this is continuous in y as long as no other eigenvalue enters the contour.) The same holds for $\operatorname{Coker} D = \bigcup_y \operatorname{Coker} D_y$. Nash then defines [Nash p. 119, (4.123)]

$$\operatorname{index} D = [\operatorname{Ker} D] - [\operatorname{Coker} D] \ \in K(Y). \tag{4.169}$$

Key idea: The index of a family is a virtual bundle

For one operator, the index is a number: $\dim\operatorname{Ker} - \dim\operatorname{Coker}$. For a family, it is a K-theory class: [bundle of kernels] − [bundle of cokernels]. Its rank is the old integer index. Its Chern classes are new invariants, which measure how the zero modes twist as we go around loops and surfaces in the parameter space.

Physical picture: Berry phases

A physicist already knows one instance. If $H(y)$ is a family of Hamiltonians with an isolated, non-degenerate ground state $|0(y)\rangle$, the ground states form a line bundle over Y. Its Berry connection is $\mathcal{A} = \langle 0(y)|\mathrm{d}\,0(y)\rangle$ and its Berry curvature integrated over a closed surface in Y, divided by $2\pi i$, is an integer: the first Chern number of the ground-state bundle. The families index is the same idea applied to the bundle of zero modes of an elliptic operator. When the cokernel is zero, $\operatorname{index} D$ *is* the bundle of zero modes and its c_1 is literally a Berry-curvature integral.

Watch out: Finite-dimensional families between fixed spaces are boring

Let $D_y : \mathbb{C}^N \to \mathbb{C}^M$ be a family of matrices with constant rank. The exact sequence $0 \to \operatorname{Ker} D_y \to \mathbb{C}^N \to \mathbb{C}^M \to \operatorname{Coker} D_y \to 0$ holds fibrewise, and for vector bundles an exact sequence splits, so $\operatorname{Ker} D \oplus \underline{\mathbb{C}}^M \cong \underline{\mathbb{C}}^N \oplus \operatorname{Coker} D$. Hence

$$\operatorname{index} D = [\operatorname{Ker} D] - [\operatorname{Coker} D] = [\underline{\mathbb{C}}^N] - [\underline{\mathbb{C}}^M] = N - M,$$

a trivial class: no Chern classes survive. Kernels of matrices can twist, but the cokernels twist in exactly the compensating way. A non-trivial index bundle needs *infinite*-dimensional spaces of sections. This is one reason the families index is a genuinely analytic object.

4.18.4 Jumping kernels and the stabilisation trick

Nash points out [Nash p. 119] that in practice $\dim\operatorname{Ker} D_y$ does jump (as in Examples 4.105 and 4.106), and then $\bigcup_y \operatorname{Ker} D_y$ is not a vector bundle. He refers to Atiyah and Singer for the fix. The fix is a piece of linear algebra worth seeing, because it is also how one proves that $\operatorname{index} D$ is well defined.

The idea: add a finite number of extra directions

Fix $y_0 \in Y$. Since $\operatorname{Coker} D_{y_0}$ is finite dimensional, we can choose a finite-dimensional subspace $W \subset \Gamma(X, F)$ that, together with the image, spans everything:

$$\operatorname{Im} D_{y_0} + W = \Gamma(X, F).$$

(Take for W the cokernel itself, realised as $\operatorname{Ker} D_{y_0}^*$.) Being surjective is an *open* condition, so for y in a neighbourhood U of y_0 we still have $\operatorname{Im} D_y + W = \Gamma(X, F)$. Now enlarge the operator:

$$\widetilde{D}_y : \Gamma(X, E) \oplus W \longrightarrow \Gamma(X, F), \qquad \widetilde{D}_y(u, w) = D_y u + w. \tag{4.170}$$

This is surjective for $y \in U$. Its kernel is

$$\operatorname{Ker} \widetilde{D}_y = \{(u, -D_y u) : D_y u \in W\} \cong D_y^{-1}(W),$$

the set of sections that D_y maps into the finite-dimensional space W. Its dimension is

$$\dim\operatorname{Ker} \widetilde{D}_y \overset{(1)}{=} \operatorname{index} \widetilde{D}_y \overset{(2)}{=} \operatorname{index} D_y + \dim W. \tag{4.171}$$

(1) $\widetilde{D}_y$ is surjective, so its cokernel is zero and its index equals the dimension of its kernel.
(2) Adding the finite-dimensional space W to the domain (and mapping it in by the inclusion) raises the index by $\dim W$: the extra piece is a finite-rank perturbation of $D_y \oplus 0_W$, and $\operatorname{index}(D_y \oplus 0_W) = \operatorname{index} D_y + \dim W - 0$.

The right side of (4.171) does not depend on y. So the spaces $D_y^{-1}(W)$ have *constant* dimension and form an honest vector bundle over U, even though $\operatorname{Ker} D_y$ jumps. We then define

$$\operatorname{index} D\big|_U = [D^{-1}(W)] - [\underline{W}] \ \in K(U), \tag{4.172}$$

where $\underline{W} = U \times W$ is a trivial bundle. By compactness of Y one can use a single W for all of Y (take the span of finitely many local choices).

Working definition 4.48 (index bundle of a family)

To compute $\operatorname{index} D \in K(Y)$: find a finite-dimensional space W of sections of F such that $\operatorname{Im} D_y + W$ is everything [every f can be written as $D_y u + w$ with $w \in W$] for all y. The index is [$\{u : D_y u \in W\}$] $-$ [W] [the first term is the vector bundle over Y whose fibre at y is this finite-dimensional space of u]. If the kernel dimension is constant, one may take $W = \operatorname{Coker} D_y$ [realised as the zero modes of D_y^*, a finite-dimensional space of sections of F] (varying with y) and recover $[\operatorname{Ker} D] - [\operatorname{Coker} D]$.

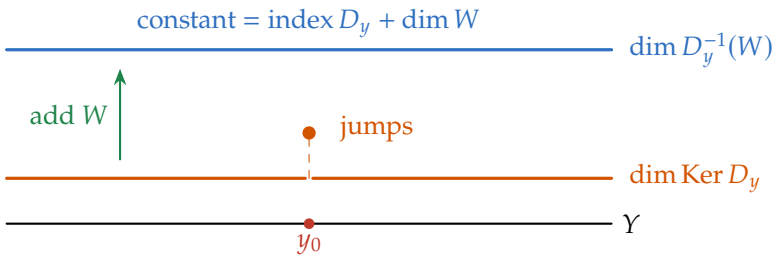


Figure 4.61: The stabilisation trick. The kernel dimension (orange) jumps at y_0, so the kernels do not form a vector bundle. After enlarging the domain by a fixed finite-dimensional space W that fills in the cokernel, the kernels $D_y^{-1}(W)$ of the enlarged operator have constant dimension (blue) and do form a bundle. Subtracting the trivial bundle W gives the index class.

Worked example 4.107: Stabilising the two jumping families

The 1×1 family $D_y = y$. Take $W = \mathbb{C}$, the whole target. Then $D_y^{-1}(W) = \mathbb{C}$ for every y, and $\operatorname{index} D = [\underline{\mathbb{C}}] - [\underline{\mathbb{C}}] = 0$. The jump at $y = 0$ has been absorbed.
The circle family D_a near $a = 0$. Take $W = \operatorname{span}\{1\}$, the constant functions in the target. For a near 0 the modes $n \neq 0$ are mapped isomorphically onto themselves (their eigenvalues $2\pi i(n + a)$ stay away from zero), and the mode $n = 0$ is mapped into W. Thus

$$D_a^{-1}(W) \overset{(1)}{=} \left\{ f = \textstyle\sum_n f_n e^{2\pi i n\theta} : \ 2\pi i(n+a) f_n = 0 \text{ for } n \neq 0 \right\}$$
$$\overset{(2)}{=} \operatorname{span}\{1\},$$

(1) $D_a f$ lies in W iff all its non-constant Fourier coefficients vanish.
(2) For $n \neq 0$ and $|a| < 1$, $n + a \neq 0$, so $f_n = 0$; f_0 is free.

a trivial line. Hence $\operatorname{index} D = [\underline{\mathbb{C}}] - [\underline{\mathbb{C}}] = 0$ near $a = 0$, consistent with $\operatorname{index} D_a = 0$ elsewhere.

4.18.5 Operators on a varying manifold: fibre bundles of manifolds

Nash's second generalisation [Nash pp. 119–120, (4.124)] is that the manifold itself may vary with y. The Laplacian on a torus depends on the torus's shape; in string theory one integrates over all shapes of a Riemann surface, and the operators on each surface form a family whose manifold changes from point to point.

Working definition 4.49 (family on a fibration)

Let $\pi : Z \to Y$ be a fibre bundle whose fibres $X_y = \pi^{-1}(y)$ are closed n-manifolds, all diffeomorphic [identified by a smooth bijection with smooth inverse] to a model X. Over Z let there be vector bundles E, F, with restrictions E_y, F_y to X_y. A family is a collection of elliptic operators

$$D_y : \Gamma(X_y, E_y) \longrightarrow \Gamma(X_y, F_y), \qquad y \in Y, \tag{4.173}$$

varying continuously with y. *Locally* (over a small $U \subset Y$) $Z \cong U \times X$ and this is just a family on the fixed manifold X; globally, going around a loop in Y may bring X back to itself by a non-trivial diffeomorphism [e.g. on a torus, $(\theta_1, \theta_2) \mapsto (\theta_1 + \theta_2, \theta_2)$, which cannot be deformed to the identity].

The structure group of $Z \to Y$ is $\mathrm{Diff}(X)$, the diffeomorphism group of the fibre. Nash's point is that this is exactly what is required for the local analysis to be meaningful: whatever we compute on X_y in local coordinates must be invariant under diffeomorphisms of X, so that it does not matter which identification $X_y \cong X$ we used. Kernels and cokernels are invariantly defined, so (4.169) and the stabilisation trick carry over unchanged.

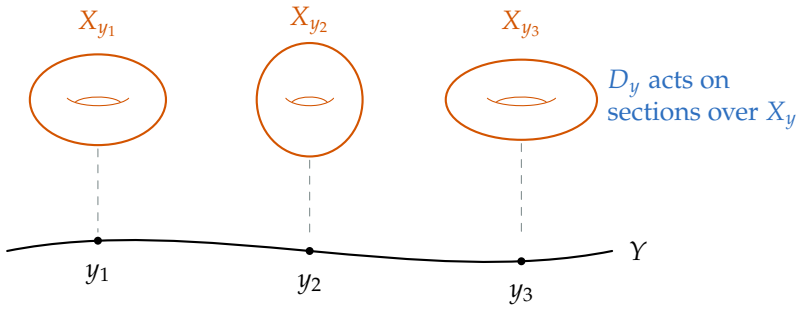


Figure 4.62: A family on a fibration. Over each point y of the parameter space sits a manifold X_y (here a torus whose shape changes with y), and an operator D_y acting on fields on X_y. Locally in Y this is the same as a family on a fixed torus; globally, transporting a torus around a loop in Y may return it to itself by a diffeomorphism (for example a Dehn twist).

Worked example 4.108: The family of flat tori

For τ in the upper half plane let $X_\tau = \mathbb{C}/(\mathbb{Z} + \tau\mathbb{Z})$, with the flat metric $|\mathrm{d}z|^2$. Every flat torus (up to scale) is of this form, and $X_\tau \cong X_{\tau'}$ isometrically iff $\tau' = \frac{a\tau+b}{c\tau+d}$ with $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}(2,\mathbb{Z})$. Identify each X_τ with the fixed square torus $X = \mathbb{R}^2/\mathbb{Z}^2$ through $z = x + \tau y$. Then $\mathrm{d}z = \mathrm{d}x + \tau\,\mathrm{d}y$ and

$$\frac{\partial}{\partial \bar z} \overset{(1)}{=} \frac{1}{\tau - \bar\tau}\left(\tau\frac{\partial}{\partial x} - \frac{\partial}{\partial y}\right) \overset{(2)}{=} \frac{1}{2i\,\mathrm{Im}\,\tau}\left(\tau\,\partial_x - \partial_y\right).$$

(1) Invert $x = \frac{\tau\bar z - \bar\tau z}{\tau - \bar\tau}$, $y = \frac{z - \bar z}{\tau - \bar\tau}$ and use the chain rule: $\partial x/\partial\bar z = \tau/(\tau - \bar\tau)$, $\partial y/\partial\bar z = -1/(\tau - \bar\tau)$.
(2) $\tau - \bar\tau = 2i\,\mathrm{Im}\,\tau$.

So on the fixed torus X we get the family of operators $D_\tau = (\tau\partial_x - \partial_y)/2i\,\mathrm{Im}\,\tau$, parametrised by τ. The honest parameter space is the moduli space $Y = \{\mathrm{Im}\,\tau > 0\}/\mathrm{SL}(2,\mathbb{Z})$, and going around a loop in Y (for example $\tau \to \tau + 1$) brings X back to itself by the diffeomorphism $(x, y) \mapsto (x + y, y)$, a Dehn twist. This is exactly the situation of (4.173) with structure group $\mathrm{Diff}(X)$. On each fibre $\mathrm{Ker}\, D_\tau = \mathbb{C}$ (the constants), and the families index is the statement that these kernels, and the corresponding cokernels (spanned by the class of $\mathrm{d}\bar z = \mathrm{d}x + \bar\tau\,\mathrm{d}y$, which does change with τ), form line bundles over moduli space. Their determinant is the object behind the "measure on moduli space" in string theory (Chapter 9).

4.18.6 Analytic and topological index for families

The proof of the families index theorem follows the same pattern as for one operator: we define an analytic index and a topological index and show they agree. What changes is where the symbol lives.

The symbol class

Each D_y has a leading symbol $\sigma_m(D_y)(x, p)$, defined for $x \in X_y$ and $p \in T^*_x X_y$. For a single operator the symbol defines a class in $K(T^*X)$ (compactly supported K-theory, because the symbol is an isomorphism off the zero section). For a family we collect all the cotangent spaces *along the fibres*:

$$T_F Z = \bigcup_{y \in Y} T^* X_y,$$

the **cotangent bundle along the fibres** of Z (Nash's $T_F Z$). A point of $T_F Z$ is a triple (y, x, p) with $x \in X_y$ and p a covector tangent to the fibre. Its dimension is $\dim Y + 2n$. The symbols of all the D_y together give an element

$$[\sigma_m(D)] \in K(T_F Z).$$

The analytic index is then the map [Nash p. 120, (4.125)]

$$\mathrm{index}_a : K(T_F Z) \longrightarrow K(Y), \qquad [\sigma_m(D)] \longmapsto \mathrm{index}\, D. \tag{4.174}$$

It is well defined for the same reason as before: two families with the same symbol class can be deformed into each other, and the index bundle is invariant under deformations (a homotopy of families over Y is a family over $Y \times [0,1]$, and vector bundles over $Y \times [0,1]$ are pulled back from Y).

The topological index

For one operator the topological index was built by embedding X in some $\mathbb{R}^N$, pushing the symbol class forward by the Thom isomorphism $K(T^*X) \to K(T^*\mathbb{R}^N)$, and then using Bott periodicity $K(T^*\mathbb{R}^N) = K(\mathbb{R}^{2N}) \cong K(\mathrm{pt}) = \mathbb{Z}$. For a family, we embed the whole fibration fibrewise,

$$Z \hookrightarrow Y \times \mathbb{R}^N, \qquad X_y \hookrightarrow \{y\} \times \mathbb{R}^N,$$

push $[\sigma_m(D)]$ forward to $K(Y \times \mathbb{R}^{2N})$ and use Bott periodicity *with parameters*: $K(Y \times \mathbb{R}^{2N}) \cong K(Y)$. The result is a homomorphism $\mathrm{index}_t : K(T_F Z) \to K(Y)$.

Key idea: The families index theorem (Atiyah–Singer IV)

For every elliptic family,

$$\operatorname{index}_a = \operatorname{index}_t : \ K(T_F Z) \longrightarrow K(Y).$$

The bundle of zero modes, modulo the bundle of zero modes of the adjoint, is computed by a purely topological construction from the symbols. *Why it is true:* the proof for a single operator used only properties (functoriality under embeddings, excision, normalisation on a point) that hold "with parameters"; everything goes through fibre by fibre, and the constructions are continuous in y. Setting $Y = \mathrm{pt}$ recovers the ordinary index theorem.

4.18.7 The cohomological formula and integration along the fibre

Applying the Chern character to the K-theory statement gives a formula in cohomology [Nash p. 120, (4.126)]:

$$\operatorname{ch}(\operatorname{index} D) = (-1)^n \, \pi_* \Big\{ \operatorname{ch}\big([\sigma_m(D)]\big) \ \operatorname{td}(T_F Z_{\mathbb{C}}) \Big\} \ \in H^{\mathrm{even}}(Y; \mathbb{Q}), \tag{4.175}$$

where $T_F Z_{\mathbb{C}}$ is the complexified tangent bundle along the fibres and π_* is **integration along the fibre**. This is the single-operator formula with one change: the final "evaluate on $[X]$" becomes "integrate over the fibres, leaving a form on Y". The output is not a number but a cohomology class on Y: its degree-0 part is the integer index, its degree-2 part is $c_1(\operatorname{index} D)$, and so on.

What integration along the fibre is

Working definition 4.50 (integration along the fibre)

Let $\pi : Z \to Y$ have oriented [with a consistent handedness of fibre coordinates, which fixes the sign of $\int_X$] k-dimensional fibres. Locally $Z = U \times X$ with fibre coordinates $x^1, \dots, x^k$ and base coordinates y^α. Write a form on Z as a sum of terms

$$\omega = f(x,y)\, \mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^k \wedge \mathrm{d}y^{\alpha_1} \wedge \cdots \wedge \mathrm{d}y^{\alpha_q} + (\text{terms with fewer than } k \text{ fibre differentials}).$$

Then

$$\pi_* \omega = \Big(\int_X f(x,y)\, \mathrm{d}^k x \Big) \mathrm{d}y^{\alpha_1} \wedge \cdots \wedge \mathrm{d}y^{\alpha_q},$$

and the terms with fewer than k fibre differentials are thrown away. *Rule of thumb:* "integrate out the fibre directions, keep the base directions". The degree drops by k: $\pi_* : \Omega^p(Z) \to \Omega^{p-k}(Y)$.

Nash writes this as [Nash p. 121, (4.127)]: $\pi_*[\omega] = [\nu]$ with $\nu = \int_X \omega \in \Omega^{p-n}(Y)$. Two facts make it work in cohomology.

- *It commutes with* d: $\mathrm{d}(\pi_* \omega) = \pi_*(\mathrm{d}\omega)$ when the fibres are closed. (The x-derivative part of $\mathrm{d}\omega$ integrates to zero over the closed fibre by Stokes; the y-derivative part can be pulled out of the integral.) So closed forms go to closed forms and exact to exact.
- *Projection formula*: $\pi_*(\pi^* \beta \wedge \omega) = \beta \wedge \pi_* \omega$ for any form β on Y, up to the sign fixed by where we put the fibre differentials. (Base forms are constant along the fibre, so they come out of the integral.)

For $Y = \mathrm{pt}$ it is ordinary integration over X.

Worked example 4.109: Integration along the fibre in coordinates

Let $Z = S^1_\theta \times \mathbb{R}^2_{(u,v)} \to Y = \mathbb{R}^2_{(u,v)}$, and

$$\omega = \cos^2\theta\,\mathrm{d}\theta \wedge \mathrm{d}u + u\,\mathrm{d}\theta \wedge \mathrm{d}u \wedge \mathrm{d}v + \sin\theta\,\mathrm{d}u \wedge \mathrm{d}v.$$

Then (with $\theta \in [0, 2\pi)$)

$$\pi_*\omega \overset{(1)}{=} \Big(\int_0^{2\pi} \cos^2\theta\,\mathrm{d}\theta\Big)\mathrm{d}u + \Big(\int_0^{2\pi} u\,\mathrm{d}\theta\Big)\mathrm{d}u \wedge \mathrm{d}v + 0$$
$$\overset{(2)}{=} \pi\,\mathrm{d}u + 2\pi u\,\mathrm{d}u \wedge \mathrm{d}v.$$

(1) The first two terms contain the single fibre differential $\mathrm{d}\theta$ (in front); the last term has no $\mathrm{d}\theta$ and is discarded.
(2) $\int_0^{2\pi} \cos^2\theta\,\mathrm{d}\theta = \pi$; u is constant on the fibre.

The degrees drop by one: the 2-form part became a 1-form and the 3-form part a 2-form. Notice the discarded term: a form that has no leg along the fibre carries no information about the integral over the fibre.

Watch out: Which fibre is integrated in (4.126)?

The class $\mathrm{ch}([\sigma_m(D)])\ \mathrm{td}(T_F Z_{\mathbb{C}})$ lives on $T_F Z$, whose fibre over $y \in Y$ is T^*X_y, of dimension $2n$. So π_* in (4.175) really integrates over $2n$ dimensions, and lowers degree by $2n$. Nash's (4.127) describes integration over the n-dimensional X. The two are reconciled by doing the integral in two stages: first integrate over each cotangent fibre $T^*_x X_y \cong \mathbb{R}^n$ (this is the Thom isomorphism, which is how the single-operator formula turned into $\{\mathrm{ch}(\cdot)\ \mathrm{td}/e\}[X]$), then over X_y. In the examples below we always use the second, "integrate over X" form.

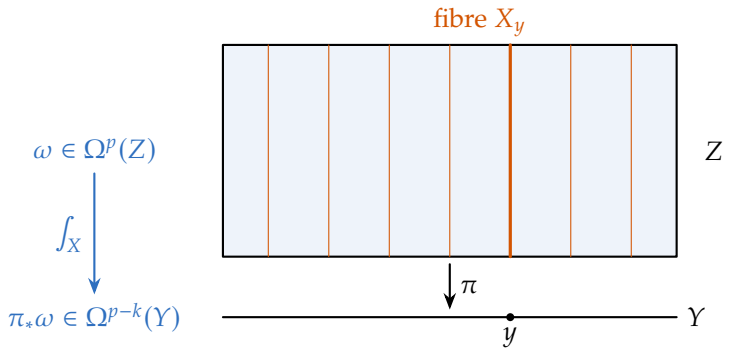


Figure 4.63: Integration along the fibre. A form on the total space Z is integrated over each fibre X_y (orange), with the base point y held fixed. What is left is a form on Y whose degree is lower by the fibre dimension k. For Y a point this is ordinary integration over X.

The families Riemann–Roch formula

For the twisted operators of Section 4.17 the formula simplifies exactly as it did for one operator. Suppose $Z = X \times Y$ and the family is $\bar\partial$ on a complex curve or surface X, twisted by a family of holomorphic bundles, i.e. by one bundle $P \to X \times Y$ whose restriction to $X \times \{y\}$ is the twisting

bundle P_y. Then the chain that led to (4.158) goes through verbatim, with $[X]$ replaced by π_*:

$$\operatorname{ch}(\operatorname{index}\bar\partial_P) = \pi_*\big(\operatorname{ch}(P)\,\operatorname{td}(X)\big) = \int_X \operatorname{ch}(P)\wedge\operatorname{td}(X). \tag{4.176}$$

The same holds for twisted Dirac families, $\operatorname{ch}(\operatorname{index}\not{D}_P) = \int_X \operatorname{ch}(P)\wedge\hat A(X)$. This is the **families index theorem** in the form physicists use. The only new feature is that $\operatorname{ch}(P)$ now has mixed components (some legs along X, some along Y), and the Y-legs survive the integral.

4.18.8 Worked example: flat line bundles on a torus

We now compute a non-trivial index bundle from beginning to end. It continues Example 4.96.

The family

Let $X = \mathbb{R}^2/\mathbb{Z}^2$ with coordinates (x,y) and $z = x+iy$ (the square torus, flat metric). For $(a,b)\in\mathbb{R}^2$ consider the flat connection $A_{a,b} = 2\pi i(a\,\mathrm{d}x + b\,\mathrm{d}y)$ on the trivial line bundle and the twisted Cauchy–Riemann operator

$$\bar\partial_{a,b} f = \big(\partial_{\bar z} + (A_{a,b})_{\bar z}\big) f\,\mathrm{d}\bar z.$$

Exactly as for the circle (Example 4.106), shifting a or b by an integer is a gauge transformation ($g = e^{-2\pi i x}$ or $e^{-2\pi i y}$), so the parameter space is a second torus, the **dual torus** $\hat X = \{(a,b)\}/\mathbb{Z}^2$.

The spectrum and the jumping kernel

We have $(A_{a,b})_{\bar z} = \tfrac12(A_x + iA_y) = \pi i(a+ib)$ and $\partial_{\bar z} = \tfrac12(\partial_x + i\partial_y)$. On a Fourier mode $e_{m,n} = e^{2\pi i(mx+ny)}$:

$$\begin{aligned}\bar\partial_{a,b}\, e_{m,n} &\overset{(1)}{=} \Big[\tfrac12\big(2\pi i m + i\cdot 2\pi i n\big) + \pi i(a+ib)\Big] e_{m,n}\,\mathrm{d}\bar z \\ &\overset{(2)}{=} \pi i\big[(m+a) + i(n+b)\big] e_{m,n}\,\mathrm{d}\bar z.\end{aligned} \tag{4.177}$$

(1) $\partial_x e_{m,n} = 2\pi i m\, e_{m,n}$, $\partial_y e_{m,n} = 2\pi i n\, e_{m,n}$.
(2) Collect: $\tfrac12(2\pi i m - 2\pi n) = \pi i(m+in)$.

So the operator is diagonal, with eigenvalue $\pi i\,(w+m+in)$ where $w = a+ib$. The kernel is zero unless $w\in\mathbb{Z}+i\mathbb{Z}$, i.e. unless (a,b) is the origin of $\hat X$; there, one mode is annihilated and one mode of the target is missed. So $\operatorname{index}\bar\partial_{a,b} = 0$ for all (a,b), and the kernel jumps at a single point of $\hat X$. Riemann–Roch agrees: degree 0, genus 1, index $= 0 + 1\cdot(1-1) = 0$.

The index bundle from the families formula

To use (4.176) we need the line bundle $P\to X\times\hat X$ (the **Poincaré bundle**) whose restriction to $X\times\{(a,b)\}$ is the flat bundle with connection $A_{a,b}$. The natural connection on P is the same expression, now read as a 1-form on the 4-manifold $X\times\hat X$ with a,b as coordinates:

$$\mathbb{A} = 2\pi i(a\,\mathrm{d}x + b\,\mathrm{d}y).$$

Its curvature has mixed components only:

$$c_1(P) \overset{(1)}{=} \frac{i}{2\pi}\,\mathrm{d}\mathbb{A} \overset{(2)}{=} \frac{i}{2\pi}\cdot 2\pi i\,(\mathrm{d}a\wedge\mathrm{d}x + \mathrm{d}b\wedge\mathrm{d}y) = -(\mathrm{d}a\wedge\mathrm{d}x + \mathrm{d}b\wedge\mathrm{d}y). \tag{4.178}$$

(1) $c_1 = \frac{i}{2\pi} F$ and $F = \mathrm{d}\mathbb{A}$ for a line bundle.
(2) $\mathrm{d}(a\,\mathrm{d}x) = \mathrm{d}a \wedge \mathrm{d}x$; now a varies, which is the whole point.

(On $X \times \{(a,b)\}$ with a, b fixed the curvature vanishes: each P_y is flat, as it should be.) The torus is flat, so $\mathrm{td}(X) = 1$. Now

$$\begin{aligned}\mathrm{ch}(\operatorname{index}\bar\partial) &\overset{(1)}{=} \int_X \Big(1 + c_1(P) + \tfrac{1}{2}c_1(P)^2\Big) \\ &\overset{(2)}{=} \int_X \tfrac{1}{2}\big(\mathrm{d}a \wedge \mathrm{d}x + \mathrm{d}b \wedge \mathrm{d}y\big)^2 \\ &\overset{(3)}{=} \int_X \mathrm{d}a \wedge \mathrm{d}x \wedge \mathrm{d}b \wedge \mathrm{d}y \\ &\overset{(4)}{=} \int_X (-\mathrm{d}x \wedge \mathrm{d}y \wedge \mathrm{d}a \wedge \mathrm{d}b) \overset{(5)}{=} -\mathrm{d}a \wedge \mathrm{d}b. \end{aligned} \tag{4.179}$$

(1) Families Riemann–Roch (4.176) with $\mathrm{td}(X) = 1$; higher powers of $c_1(P)$ are 6-forms and vanish on a 4-manifold.
(2) The 1 has no X-legs and $c_1(P)$ has only one X-leg in each term; integrating over the 2-dimensional X needs two, so both drop out.
(3) $(\alpha+\beta)^2 = \alpha^2 + 2\alpha\wedge\beta + \beta^2$ for 2-forms, and $\alpha^2 = \beta^2 = 0$ here because each contains a repeated differential.
(4) Reorder: moving $\mathrm{d}x$ past $\mathrm{d}a$ costs a sign; moving $\mathrm{d}y$ past $\mathrm{d}b$ and $\mathrm{d}a$ costs two signs.
(5) Integration along the fibre with $\int_X \mathrm{d}x \wedge \mathrm{d}y = 1$.

Reading off by degree: the rank of $\operatorname{index}\bar\partial$ is 0 (as it must be), and

$$c_1(\operatorname{index}\bar\partial) = -\mathrm{d}a \wedge \mathrm{d}b, \qquad \int_{\hat X} c_1(\operatorname{index}\bar\partial) = -1.$$

So *the index bundle of this family is a non-trivial element of* $K(\hat X)$, even though every single operator has index zero. (The sign depends on the orientation of $\hat X$; the fact that the Chern number is ± 1, not 0, does not.)

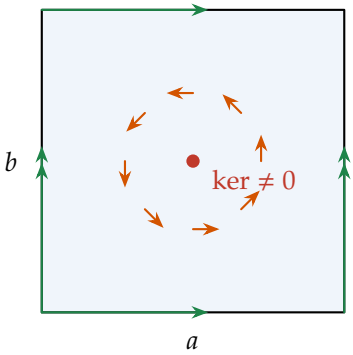
A check via a twisted family with no jumping

Now twist in addition by a fixed holomorphic line bundle L of degree 1 on X, with $c_1(L) = h$, $\int_X h = 1$. The family is $\bar\partial_{L\otimes P_{a,b}}$, of index $1 + 1\cdot(1-1) = 1$ at every point. For a degree-1 bundle on a torus $h^1 = 0$, so the cokernel vanishes and the kernel is a genuine line bundle over $\hat X$: at each (a,b) it is spanned by one theta function, the unique lowest-Landau-level state of a particle on the torus with one flux quantum and Aharonov–Bohm fluxes a, b through the two cycles. The families formula gives

$$\mathrm{ch}(\mathrm{index}) \overset{(1)}{=} \int_X e^{h}\, e^{c_1(P)} \overset{(2)}{=} \int_X \Big(h + h c_1(P) + \tfrac{1}{2}c_1(P)^2\Big) \overset{(3)}{=} 1 - \mathrm{d}a \wedge \mathrm{d}b.$$

(1) $\mathrm{ch}(L\otimes P) = \mathrm{ch}\,L\ \mathrm{ch}\,P = e^h e^{c_1(P)}$ and $\mathrm{td}(X) = 1$.
(2) Keep terms with exactly two X-legs. Terms with h^2 vanish (h is already a top form on X).
(3) $\int_X h = 1$; $h \wedge c_1(P) = 0$ because every term of $c_1(P)$ contains $\mathrm{d}x$ or $\mathrm{d}y$ and $h \propto \mathrm{d}x\wedge\mathrm{d}y$; the last term is (4.179).

So the bundle of lowest-Landau-level states over the flux torus is a line bundle of Chern number -1. This is precisely the Niu–Thouless–Wu computation of the quantised Hall conductance: threading one flux quantum through each cycle of the torus sweeps the ground state around a Berry-phase monopole of unit charge, and $\sigma_{xy} = \frac{e^2}{h}\,|c_1| = \frac{e^2}{h}$.



dual torus $\hat{X}$ of flat connections
(origin moved to the centre)

index $\bar{\partial}_{a,b} = 0$ at every point,
but $\int_{\hat{X}} c_1(\text{index}\,\bar{\partial}) = -1$:
the phase of the small eigenvalue
$\pi i(a + ib)$ winds once (orange).

Figure 4.64: The Poincaré family on the dual torus. Only at one point (red) does $\bar{\partial}_{a,b}$ have a kernel. Near it, the only small eigenvalue is $\pi i(a + ib)$, whose phase winds once around the red point (orange arrows). This winding cannot be undone globally on $\hat{X}$, and it is what the families index theorem detects as a unit first Chern class.

4.18.9 What the Chern character cannot see: torsion

Nash remarks [Nash p. 121] that (4.175) computes $\text{ch}(\text{index}\,D) \in H^*(Y;\mathbb{Q})$, not $\text{index}\,D \in K(Y)$ itself, "thus we lose any torsion that may be present". What does that mean concretely?

A **torsion** element $\xi \in K(Y)$ is one with $N\xi = 0$ for some integer $N \geq 1$. Since ch is additive, $N\,\text{ch}(\xi) = \text{ch}(N\xi) = 0$, and in rational cohomology this forces $\text{ch}(\xi) = 0$. So the cohomological formula is blind to torsion.

Worked example 4.110: A torsion class on $\mathbb{RP}^2$

On $Y = \mathbb{RP}^2$ (the sphere with antipodal points identified) take the real line bundle ℓ whose sections are odd functions on S^2, and its complexification $\ell_{\mathbb{C}}$. Then $\ell_{\mathbb{C}} \otimes \ell_{\mathbb{C}}$ is trivial (the product of two odd functions is even), and one shows that $\xi = [\ell_{\mathbb{C}}] - 1$ satisfies $\xi^2 = -2\xi$ and $2\xi = 0$ in $K(\mathbb{RP}^2)$: $K(\mathbb{RP}^2) = \mathbb{Z} \oplus \mathbb{Z}_2$. The Chern class $c_1(\ell_{\mathbb{C}})$ is a torsion class in $H^2(\mathbb{RP}^2;\mathbb{Z}) = \mathbb{Z}_2$, so in real (de Rham) cohomology $c_1(\ell_{\mathbb{C}}) = 0$ and $\text{ch}(\xi) = 0$. Concretely: $\ell_{\mathbb{C}}$ admits a flat connection (holonomy -1 around the non-contractible loop), so its curvature form is identically zero. Any Chern–Weil formula gives zero for it, yet it is not trivial. A family whose index bundle were ξ would have $\text{ch}(\text{index}\,D) = 0$ with $\text{index}\,D \neq 0$.

This is not an academic worry. Torsion index classes are precisely what the next section (real families) is about, and in physics they correspond to *global* anomalies, such as Witten's SU(2) anomaly, which no Feynman diagram (no curvature integral) can detect.

4.18.10 The determinant line bundle and anomalies

Nash defers examples of families to Chapter 10. We give here the one construction that makes the link to physics, because it is the main reason a physicist should care about families.

The fermion determinant is not a function

In a Euclidean path integral, integrating out a fermion in a background gauge field gives

$$Z[A] = \int \mathcal{D}\bar{\psi}\,\mathcal{D}\psi\; e^{-\int \bar{\psi} D_A \psi} = \det D_A .$$

For a chiral (Weyl) fermion, $D_A = \not{D}_A^+ : \Gamma(E^+ \otimes F) \to \Gamma(E^- \otimes F)$ maps one space to a *different* space. The determinant of a map between different vector spaces is not a number: it depends on choices of bases in both. What it naturally is, is an element of a line.

Working definition 4.51 (determinant line)

For a linear map $T : V \to W$ between finite-dimensional spaces of equal dimension d, the determinant is an element

$$\det T \in \Lambda^d V^* \otimes \Lambda^d W,$$

[a one-dimensional space, since $\Lambda^d V$ is spanned by $v_1 \wedge \cdots \wedge v_d$ for any basis] namely the map $v_1 \wedge \cdots \wedge v_d \mapsto Tv_1 \wedge \cdots \wedge Tv_d$. Choosing bases identifies this line with $\mathbb{C}$ and $\det T$ with the usual number. For a Fredholm operator [finite-dimensional kernel and cokernel] D one sets

$$\mathrm{DET}(D) = \left(\Lambda^{\max} \operatorname{Ker} D\right)^* \otimes \Lambda^{\max} \operatorname{Coker} D.$$

[$\Lambda^{\max}$ is the top exterior power, Λ^k with k the dimension, again a line] *Recipe:* if D is invertible, $\mathrm{DET}(D)$ is canonically $\mathbb{C}$ and $\det D$ is (a regularised version of) the product of eigenvalues; if not, $\det D = 0$ and the line is "spanned by the zero modes".

For a family $\{D_y\}$ these lines fit together into a line bundle $\mathrm{DET} \to Y$, the **determinant line bundle** (Quillen), even where the kernels jump. The reason is the finite-dimensional identity: for $T : V \to W$ there is an exact sequence $0 \to \operatorname{Ker} T \to V \xrightarrow{T} W \to \operatorname{Coker} T \to 0$, and exactness gives a canonical isomorphism

$$(\Lambda^{\max} \operatorname{Ker} T)^* \otimes \Lambda^{\max} \operatorname{Coker} T \cong \Lambda^{\max} V^* \otimes \Lambda^{\max} W.$$

Applied to the stabilised operators of Section 4.18.4 (with $V = D_y^{-1}(W)$, which varies nicely), the right side varies smoothly with y, so the left side does too. Comparing with $\operatorname{index} D = [\mathrm{Ker}] - [\mathrm{Coker}]$,

$$c_1(\mathrm{DET}) \overset{(1)}{=} -c_1(\operatorname{Ker} D) + c_1(\operatorname{Coker} D) \overset{(2)}{=} -c_1(\operatorname{index} D). \tag{4.180}$$

(1) $c_1(\Lambda^{\max} E) = c_1(E)$ (the top exterior power has Chern root $\sum x_i$), $c_1(L^*) = -c_1(L)$, and c_1 adds under $\otimes$ of line bundles.

(2) c_1 is additive on K-theory: $c_1([E] - [F]) = c_1(E) - c_1(F)$.

So *the families index theorem computes the topology of the determinant line bundle*: $c_1(\mathrm{DET}) = -[\mathrm{ch}(\operatorname{index} D)]_{(2)}$, the degree-2 part of the families formula.

Worked example 4.111: The determinant on the dual torus

For the Poincaré family, (4.179) and (4.180) give $\int_{\hat{X}} c_1(\mathrm{DET}) = +1$. Directly: $\det \bar{\partial}_{a,b}$ is the (zeta-regularised) product of the eigenvalues (4.177),

$$\det \bar{\partial}_{a,b} \text{ “} = \text{”} \prod_{m,n \in \mathbb{Z}} \pi i \,(w + m + in), \qquad w = a + ib.$$

As a function of w this is (up to a non-vanishing factor) a theta-function-like object: it vanishes exactly when w is a lattice point, with a simple zero, and it is holomorphic in w, so near each zero it behaves like w itself and its phase winds by $+1$. It is not periodic under $w \to w + 1$, $w \to w + i$ (the regularisation spoils periodicity by a non-vanishing factor), which is exactly what one expects of a section of a non-trivial line bundle. A section of a line bundle over a closed surface with a single simple positively oriented zero has Chern number $+1$. The analytic count and the families formula agree.

Anomalies as the topology of the determinant bundle

Now take $Y = \mathcal{A}/\mathcal{G}$, the space of gauge fields modulo gauge transformations, and the family of chiral Dirac operators $\not{D}_A^+$ in n dimensions. The partition function $\det \not{D}_A^+$ is a section of $\mathrm{DET} \to \mathcal{A}/\mathcal{G}$.

- If DET is a *trivial* line bundle (with its natural connection flat), we can pick a nowhere-vanishing reference section and turn $\det \not{D}_A^+$ into a genuine gauge-invariant function of A. The theory is consistent.
- If DET is *non-trivial*, no such choice exists: the effective action cannot be made gauge invariant. This is a **gauge anomaly**.

The families index theorem computes $c_1(\mathrm{DET})$. To evaluate it on a 2-sphere $S^2 \subset \mathcal{A}/\mathcal{G}$, form the $(n+2)$-manifold $X \times S^2$ with the gauge bundle P whose restriction to $X \times \{s\}$ carries the gauge field A_s. Then

$$\int_{S^2} c_1(\mathrm{DET}) \overset{(1)}{=} -\int_{S^2}\int_X \mathrm{ch}(P)\,\hat{A}(X) \overset{(2)}{=} -\int_{X\times S^2} \mathrm{ch}(P)\,\hat{A}(X\times S^2). \tag{4.181}$$

(1) (4.180) and the families formula for twisted Dirac operators.
(2) Integration along the fibre followed by integration over the base is integration over the total space; $\hat{A}(S^2) = 1$.

This is the celebrated statement that *the gauge anomaly in n dimensions is governed by the index density in $n+2$ dimensions*: the anomaly of a 4D chiral fermion is controlled by $\int \mathrm{ch}_3(F) \propto \int \mathrm{tr}\, F^3$ in six dimensions, which is where the famous $d^{abc} = \mathrm{tr}\, T^a\{T^b, T^c\}$ anomaly coefficient comes from. The perturbative (triangle-diagram) form of this, and the torsion (global) anomalies invisible to ch, are the subject of Chapter 10.

Dictionary: families index ↔ physics

parameter space Y	space of background fields, moduli, Bloch momenta
bundle of kernels $\mathrm{Ker}\, D$	bundle of zero modes / ground states
$c_1(\mathrm{Ker}\, D)$ when $\mathrm{Coker}\, D = 0$	integrated Berry curvature
$\mathrm{DET} \to Y$	where the fermion determinant $Z[A]$ lives
$c_1(\mathrm{DET}) \neq 0$	perturbative gauge anomaly
torsion part of index D	global anomaly (invisible to diagrams)
integration along the fibre	descent from $n+2$ to n dimensions

Reading Nash (pp. 119–121, (4.123)–(4.127))

Nash defines the index of a family as $[\mathrm{Ker}\, D] - [\mathrm{Coker}\, D] \in K(Y)$ (4.123), notes that this needs modification when $\dim \mathrm{Ker}\, D_y$ jumps (we supplied the stabilisation trick of Atiyah–Singer), generalises to operators on the fibres X_y of a bundle $Z \to Y$ with structure group $\mathrm{Diff}(X)$ (4.124), introduces the symbol class in $K(T_F Z)$ and the analytic index (4.125), sketches the topological index via an embedding $Z \hookrightarrow Y \times \mathbb{R}^n$, and states the cohomological formula (4.126) with integration along the fibre (4.127). He closes by noting that the formula only computes $\mathrm{ch}(\mathrm{index}\, D) \in H^*(Y;\mathbb{Q})$, losing torsion. He defers examples to Chapter 10 (Chapter 10). We added: the explicit jumping examples, the Poincaré family on the torus with its Chern number, the Hall-conductance interpretation, the determinant line bundle and its link to anomalies. Beware the dimension count in (4.127) (see the pitfall on which

fibre is integrated).

Recap

- A family $\{D_y\}_{y\in Y}$ has index $\operatorname{index} D = [\operatorname{Ker} D] - [\operatorname{Coker} D] \in K(Y)$; its rank is the ordinary index, its Chern classes are new.
- When the kernel jumps, stabilise: $\operatorname{index} D = [D^{-1}(W)] - [W]$ for a finite-dimensional W filling the cokernels.
- The manifold may vary: families live on fibrations $Z \to Y$ with structure group $\mathrm{Diff}(X)$; symbols give a class in $K(T_F Z)$.
- Families index theorem: $\operatorname{index}_a = \operatorname{index}_t : K(T_F Z) \to K(Y)$; cohomologically $\operatorname{ch}(\operatorname{index} D) = \int_{\text{fibre}}(\cdots)$, e.g. $\int_X \operatorname{ch}(P)\operatorname{td}(X)$.
- Example: flat bundles on T^2 give an index bundle over the dual torus with $c_1 = \mp 1$ (Hall conductance).
- $c_1(\mathrm{DET}) = -c_1(\operatorname{index} D)$: the families index controls whether the fermion determinant is a function (no anomaly) or only a section.
- ch loses torsion.

4.19 The index for real families

Many operators in physics are *real*. The Laplacian has real coefficients. The Dirac operator of a Majorana fermion, written in a Majorana basis of gamma matrices, is a real matrix of real derivatives. The Bogoliubov–de Gennes Hamiltonian of a superconductor, in a Majorana basis, is i times a real antisymmetric matrix. For such operators there is a natural question that the complex theory cannot even ask:

> *A real skew-adjoint operator always has index zero. Is there nevertheless a robust integer, or at least a robust parity, attached to it?*

The answer is the **mod 2 index**: the number of zero modes, counted modulo 2, is a deformation invariant. Physically it counts unpaired Majorana zero modes; its parity cannot change under any continuous deformation because zero modes of a real skew operator can only be created or destroyed in pairs. Nash explains [Nash pp. 121–123] that this invariant is an index of a *real family*, living in a real K-theory group that is pure torsion, and so is invisible to any curvature formula.

4.19.1 Real elliptic operators

The reality condition on the symbol

Recall that the symbol of a differential operator is obtained by replacing ∂_μ with ip_μ. If the operator has real coefficients, complex conjugation of the symbol flips the sign of every i, which is the same as flipping $p \to -p$. Nash takes this as the definition [Nash p. 121].

Working definition 4.52 (real elliptic operator)

An elliptic operator D is **real** if its symbol satisfies

$$\overline{\sigma^D(x,p)} = \sigma^D(x,-p) \tag{4.182}$$

[the bar conjugates every matrix entry] in every local coordinate system. For a differential operator this just means *it has real coefficients*: it maps real-valued fields to real-valued fields. A family $\{D_y\}$ is **real** if every D_y is real; otherwise it is called complex.

The condition (4.182) is the symbol-level shadow of a familiar fact from Fourier analysis: a function $f(x)$ is real if and only if its Fourier transform satisfies $\overline{\hat{f}(p)} = \hat{f}(-p)$.

Worked example 4.112: Which operators are real?

- $D = \mathrm{d}/\mathrm{d}x$ on functions: $\sigma(p) = ip$, so $\overline{\sigma(p)} = -ip = \sigma(-p)$. Real.
- The Laplacian $\Delta = -\sum_\mu \partial_\mu^2$: $\sigma(p) = |p|^2$, real and even. Real.
- $D = \mathrm{d}/\mathrm{d}x + ia$ with $a \in \mathbb{R}$, $a \neq 0$: the leading symbol ip satisfies (4.182), but the operator does not map real functions to real functions. The symbol condition only sees the leading part; to have a real *operator* we need all coefficients real.
- The 2D Dirac operator with the gamma matrices $\gamma^1 = \sigma_1$, $\gamma^2 = \sigma_3$ (both real):

$$\not{D} = \sigma_1\partial_x + \sigma_3\partial_y = \begin{pmatrix} \partial_y & \partial_x \\ \partial_x & -\partial_y \end{pmatrix},$$

  a matrix of real operators. This is the *Majorana representation*: it acts on real two-component spinors. In the representation $\gamma^2 = \sigma_2$ of (4.166), by contrast, $\not{D}$ contains i and is not real.

A single real operator: nothing new

For a single real operator D, complexify: let it act on complex fields. Then $\mathrm{Ker}_{\mathbb{C}} D = \mathrm{Ker}_{\mathbb{R}} D \otimes \mathbb{C}$ (take real and imaginary parts of a complex zero mode separately; both are real zero modes since D is real), so $\dim_{\mathbb{C}} \mathrm{Ker}_{\mathbb{C}} D = \dim_{\mathbb{R}} \mathrm{Ker}_{\mathbb{R}} D$, and similarly for the cokernel. The index is unchanged, and the index theorem for the complexified operator computes it. Nash's first point [Nash p. 121] is that this is no longer true for families.

4.19.2 Real families and real K-theory

Complexification loses information for families

For a real family, the kernels are *real* vector spaces, so the index bundle is a class of *real* vector bundles over Y: an element of real K-theory $KO(Y)$. Complexifying gives a map

$$c : KO(Y) \longrightarrow K(Y), \qquad [E] \longmapsto [E \otimes_{\mathbb{R}} \mathbb{C}].$$

Nash's observation is that this map is *not injective*. The simplest example shows what gets lost.

Worked example 4.113: The Möbius bundle dies under complexification

Let $Y = S^1$ with angle $t \in [0, 2\pi]$. The **Möbius bundle** μ is the real line bundle obtained from $[0, 2\pi] \times \mathbb{R}$ by gluing $(2\pi, v) \sim (0, -v)$. A section is a function $s(t)$ with $s(2\pi) = -s(0)$; it must vanish somewhere (intermediate value theorem), so μ has no nowhere-zero section and is not trivial.

Its complexification $\mu_{\mathbb{C}}$ is glued by the same rule with $v \in \mathbb{C}$. Now $s(t) = e^{it/2}$ satisfies $s(2\pi) = e^{i\pi} = -1 = -s(0)$ and never vanishes. So $\mu_{\mathbb{C}}$ has a nowhere-zero section: $\mu_{\mathbb{C}}$ is trivial. The real class $[\mu] - 1 \neq 0$ in $KO(S^1)$ maps to $[\mu_{\mathbb{C}}] - 1 = 0$ in $K(S^1)$.

Moreover $\mu \oplus \mu$ *is* trivial as a real bundle: it is glued by $-1_2 = \left(\begin{smallmatrix} \cos\pi & -\sin\pi \\ \sin\pi & \cos\pi \end{smallmatrix}\right)$, and the rotation matrices $R(t/2)$, $t \in [0, 2\pi]$, give two sections

$$e_1(t) = \left(\cos\tfrac{t}{2}, \sin\tfrac{t}{2}\right), \qquad e_2(t) = \left(-\sin\tfrac{t}{2}, \cos\tfrac{t}{2}\right),$$

which satisfy $e_j(2\pi) = -e_j(0)$ and are everywhere linearly independent. Hence $2([\mu]-1) = 0$ in $KO(S^1)$. With some more work one shows

$$\widetilde{KO}(S^1) = \mathbb{Z}_2, \text{ generated by } \eta := [\mu] - 1, \qquad \widetilde{K}(S^1) = 0.$$

(Here $\widetilde{K}$ denotes the **reduced** group: classes of virtual rank zero.) A real family over S^1 can therefore carry a $\mathbb{Z}_2$ index that its complexification cannot see.

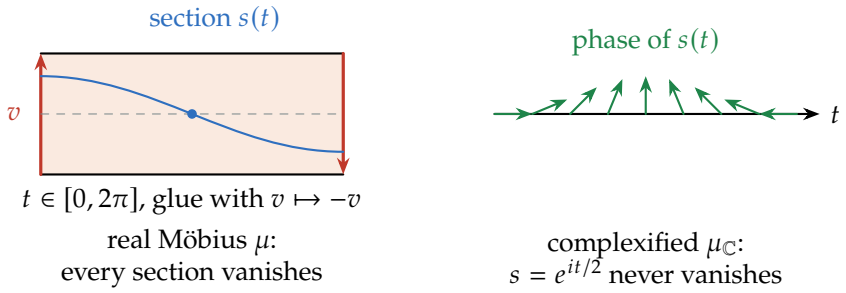


Figure 4.65: Left: the real Möbius bundle; a section starting at $+v$ must end at $-v$, so it crosses zero (blue dot). Right: after complexification the fibre is $\mathbb{C}$, and the section can rotate its phase from 1 to -1 (green arrows) without ever vanishing, so $\mu_{\mathbb{C}}$ is trivial. This is the simplest instance of information lost by complexifying a real family.

Why KR and not KO

One might expect the symbol class of a real family to live in real K-theory $KO(T_F Z)$. Nash notes [Nash pp. 121–122] that the correct group is $KR(T_F Z)$, Atiyah's K-theory of spaces with involution. The reason is visible in (4.182): the symbol is not a real matrix; complex conjugation relates the symbol at p to the symbol at $-p$.

Working definition 4.53 (Real spaces and KR)

A **Real space** is a space T with an involution $\tau : T \to T$, $\tau^2 = \mathrm{id}$. A **Real vector bundle** over it is a complex vector bundle $E \to T$ together with an *antilinear* [$\kappa(cv) = \bar{c}\,\kappa v$ for every complex number c] map $\kappa : E_t \to E_{\tau(t)}$ with $\kappa^2 = 1$ (a "complex conjugation that moves the base point"). $KR(T)$ is the K-theory built from Real vector bundles.

- If $\tau = \mathrm{id}$, then κ is an honest complex conjugation on each fibre, its fixed vectors form a real bundle $E_{\mathbb{R}}$ with $E = E_{\mathbb{R}} \otimes \mathbb{C}$, and $KR(T) = KO(T)$ [the K-theory built from real

vector bundles].

- For $T = T_F Z$ [the momenta along the fibres, points (y, x, p) with $x \in X_y$] with $\tau(y, x, p) = (y, x, -p)$, a real symbol is exactly a Real bundle map: (4.182) says that $\kappa\, \sigma(x, p) = \sigma(x, -p)\, \kappa$ with κ fibrewise complex conjugation.

So a real family has a symbol class $[\sigma(D)] \in KR(T_F Z)$, and the topological index is a map

$$\text{index}_t : KR(T_F Z) \longrightarrow KR(Y) = KO(Y),$$

where Y carries the trivial involution (each D_y is itself real). The real families index theorem says again $\text{index}_a = \text{index}_t$.

4.19.3 Skew-adjoint real operators and the mod 2 index

Their ordinary index is zero

Nash now focuses on real operators of *odd* order, such as first-order Dirac-type operators, which we take to be **skew-adjoint**: $D^* = -D$. For them the ordinary index is trivially zero:

$$\text{index}\, D \overset{(1)}{=} \dim \text{Ker}\, D - \dim \text{Ker}\, D^* \overset{(2)}{=} \dim \text{Ker}\, D - \dim \text{Ker}(-D) \overset{(3)}{=} 0. \tag{4.183}$$

(1) The cokernel of D is identified with the kernel of its adjoint.
(2) Skew-adjointness.
(3) $-D$ and D have the same kernel.

Watch out: "Automatically skew-adjoint"

Nash says that real operators of odd order are "automatically skew-adjoint". The accurate statement is about the leading symbol. For a real operator of odd order m, (4.182) and homogeneity give $\overline{\sigma(x,p)} = \sigma(x,-p) = (-1)^m \sigma(x,p) = -\sigma(x,p)$, so the symbol is purely imaginary. If in addition it is a symmetric matrix (as for scalar operators, or for Dirac operators with real symmetric gamma matrices), then $\sigma^\dagger = \bar{\sigma}^T = -\sigma$: the symbol is skew-Hermitian. The operator itself then differs from its skew-adjoint part $\frac{1}{2}(D - D^*)$ only by lower-order terms, which can be deformed away. For example $\frac{\mathrm{d}}{\mathrm{d}\theta} + V(\theta)$ with real V is real and first order but not skew-adjoint; its skew-adjoint part is $\frac{\mathrm{d}}{\mathrm{d}\theta}$.

Eigenvalues come in pairs

The key linear algebra: a real antisymmetric matrix K ($K^T = -K$) has purely imaginary eigenvalues, and they come in conjugate pairs $\pm i\lambda$. Indeed, if $Kv = i\lambda v$ with $\lambda \in \mathbb{R}$, then taking complex conjugates, $K\bar{v} = -i\lambda \bar{v}$ because K is real. By a real orthogonal change of basis, every real antisymmetric matrix takes the block form

$$K \sim \begin{pmatrix} 0 & \lambda_1 \\ -\lambda_1 & 0 \end{pmatrix} \oplus \begin{pmatrix} 0 & \lambda_2 \\ -\lambda_2 & 0 \end{pmatrix} \oplus \cdots \oplus \underbrace{0 \oplus \cdots \oplus 0}_{\dim \text{Ker}\, K}, \qquad \lambda_j > 0. \tag{4.184}$$

The non-zero eigenvalues occupy an even number of dimensions. So for an $N \times N$ real antisymmetric matrix,

$$\dim \text{Ker}\, K \equiv N \pmod 2.$$

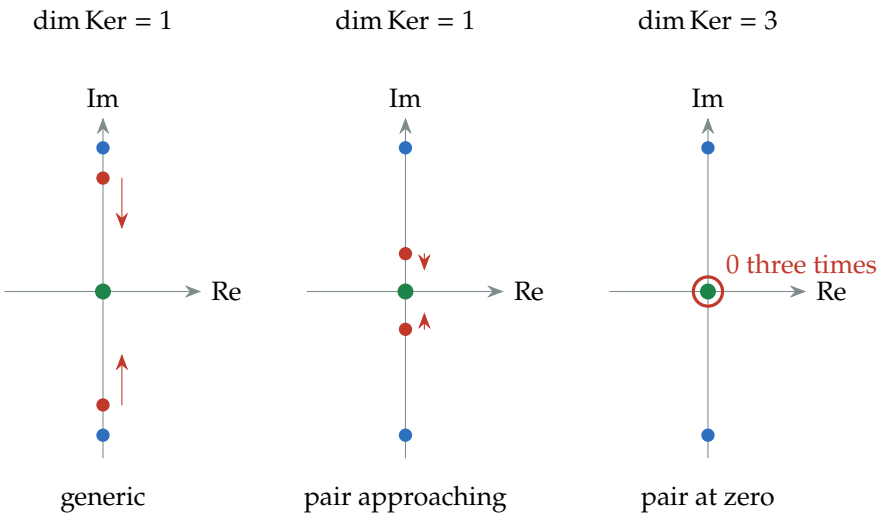


Figure 4.66: Spectrum of a real skew-adjoint operator: eigenvalues lie on the imaginary axis in pairs $\pm i\lambda$ (blue, red). The kernel (green) here is one-dimensional. When a pair (red) is deformed down to zero, the kernel jumps by two, from 1 to 3. Its parity, the mod 2 index, never changes.

As K is deformed, a pair $\pm i\lambda$ can shrink to zero, adding *two* dimensions to the kernel, or a two-dimensional piece of the kernel can split off into a pair. The kernel can never change by one.

Working definition 4.54 (mod 2 index)

For a real skew-adjoint [$D^* = -D$, so its eigenvalues are purely imaginary] elliptic operator D,

$$\operatorname{index}_1 D := \dim \operatorname{Ker} D \pmod 2 \in \mathbb{Z}_2.$$

To compute it: count the real zero modes [independent real solutions of $D\psi = 0$. Since D is real, their number equals $\dim_{\mathbb{C}} \operatorname{Ker} D$] and take the parity. *Why it is invariant:* non-zero eigenvalues of D come in pairs $\pm i\lambda$ (the finite-dimensional argument above applies to the eigenspaces), so under a continuous deformation through real skew-adjoint operators, zero modes appear and disappear two at a time.

Worked example 4.114: A family of skew matrices

Consider the 3×3 family

$$K(s) = \begin{pmatrix} 0 & s & 0 \\ -s & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad s \in \mathbb{R}.$$

Its characteristic polynomial is

$$\det(K(s) - \mu) \overset{(1)}{=} -\mu \cdot \det \begin{pmatrix} -\mu & s \\ -s & -\mu \end{pmatrix} \overset{(2)}{=} -\mu(\mu^2 + s^2),$$

(1) Expand along the third row, which has only the entry $-\mu$.
(2) $(-\mu)(-\mu) - s(-s) = \mu^2 + s^2$.

so the eigenvalues are $0, \pm is$. For $s \neq 0$, $\dim \operatorname{Ker} K(s) = 1$; at $s = 0$, $\dim \operatorname{Ker} K(0) = 3$. Both are odd, as they must be for $N = 3$.

Examples on the circle

Worked example 4.115: Majorana fermions on a circle: two spin structures

Let $D = \frac{\mathrm{d}}{\mathrm{d}\theta}$ act on real functions on the circle $\theta \in [0,1)$. It is real, first order and skew-adjoint:

$$\langle f, Dg\rangle \overset{(1)}{=} \int_0^1 f\, g'\,\mathrm{d}\theta \overset{(2)}{=} \big[f g\big]_0^1 - \int_0^1 f'\, g\,\mathrm{d}\theta \overset{(3)}{=} -\langle Df, g\rangle.$$

(1) The real L^2 inner product.
(2) Integration by parts.
(3) The boundary term vanishes for periodic functions (and also for antiperiodic ones, since then $f(1)g(1) = (-f(0))(-g(0))$).

Its kernel depends on the boundary condition, which in 1D is the choice of *spin structure*:

- *Periodic* (Ramond): $f(\theta+1) = f(\theta)$. The kernel is the constants: $\operatorname{index}_1 D = 1$.
- *Antiperiodic* (Neveu–Schwarz): $f(\theta+1) = -f(\theta)$, i.e. f is a section of the Möbius bundle. A constant antiperiodic function is zero, so $\operatorname{index}_1 D = 0$.

The mod 2 index distinguishes the two spin structures. Physically: a single real (Majorana) fermion on a periodic circle has one zero mode ψ_0 with $\psi_0^2 = \frac{1}{2}$, which cannot be paired into a complex fermion; the ground state then carries an ambiguous fermion parity. This is the one-dimensional ancestor of Majorana zero modes at the ends of a Kitaev chain.

Worked example 4.116: Deforming without changing the parity

(a) Changing the metric. If the circle has metric $a(\theta)^{-2}\mathrm{d}\theta^2$ with $a > 0$ periodic, the skew-adjoint operator is $D_a = a\frac{\mathrm{d}}{\mathrm{d}\theta} + \frac{1}{2}a'$ (the $\frac{1}{2}a'$ restores skew-adjointness with respect to $\int f g\,\mathrm{d}\theta$, since $(a\,\mathrm{d}/\mathrm{d}\theta)^* = -a\,\mathrm{d}/\mathrm{d}\theta - a'$). Its kernel:

$$a u' + \tfrac{1}{2}a' u = 0 \overset{(1)}{=} \frac{u'}{u} = -\frac{a'}{2a} \overset{(2)}{=} u = C\, a^{-1/2},$$

(1) Divide by au (where $u \neq 0$).
(2) Integrate: $\ln u = -\frac{1}{2}\ln a + \text{const}$.

which is periodic. So $\dim\operatorname{Ker} D_a = 1$ for every metric: $\operatorname{index}_1 = 1$.

(b) A rotating frame. Let $D_t = \frac{\mathrm{d}}{\mathrm{d}\theta} + tL$ act on $\mathbb{R}^3$-valued periodic functions, where $L = \left(\begin{smallmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{smallmatrix}\right)$ generates rotations about the third axis. D_t is real and skew-adjoint ($L^T = -L$). The third component obeys $u_3' = 0$: always one constant zero mode. For the first two, set $w = u_1 + iu_2$; since $L(u_1, u_2) = (-u_2, u_1)$ corresponds to $w \mapsto iw$,

$$D_t w = w' + itw = 0 \quad\Longrightarrow\quad w = Ce^{-it\theta},$$

which is periodic iff $t \in 2\pi\mathbb{Z}$, giving two more real zero modes. Hence

$$\dim\operatorname{Ker} D_t = \begin{cases} 3, & t \in 2\pi\mathbb{Z}, \\ 1, & \text{otherwise}, \end{cases} \qquad \operatorname{index}_1 D_t = 1 \text{ for all } t.$$

The kernel jumps by two, exactly as in Figure 4.66.

4.19.4 The mod 2 index as the index of a real family over the circle

Nash now explains [Nash p. 122, (4.128)–(4.129)] how the mod 2 index, which is an invariant of a *single* operator, fits into the families framework. Atiyah and Singer attach to D a real family $\widetilde{D}$ parametrised by the circle S^1, so that

$$\operatorname{index} \widetilde{D} \in KR(S^1) = KO(S^1) = \mathbb{Z} \oplus \widetilde{KO}(S^1) = \mathbb{Z} \oplus \mathbb{Z}_2.$$

The associated family

A natural candidate is the path of real operators

$$\widetilde{D}_t = \cos t \cdot 1 + \sin t \cdot D, \qquad t \in [0, \pi], \tag{4.185}$$

which starts at the identity and ends at minus the identity. It is Fredholm of index zero throughout. To see where it fails to be invertible, use the normal form (4.184): on a 2-plane where $D = \lambda \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$,

$$\det\big(\widetilde{D}_t\big|_{\text{plane}}\big) \overset{(1)}{=} \det \begin{pmatrix} \cos t & \lambda \sin t \\ -\lambda \sin t & \cos t \end{pmatrix} \overset{(2)}{=} \cos^2 t + \lambda^2 \sin^2 t > 0,$$

(1) Insert the 2×2 block of D.
(2) $\cos t$ and $\sin t$ never vanish together, and $\lambda > 0$.

while on $\operatorname{Ker} D$ the operator is $\cos t \cdot 1_m$, which vanishes at $t = \pi/2$. So $\widetilde{D}_t$ is invertible except at $t = \pi/2$, where its kernel is exactly $\operatorname{Ker} D$. In finite dimensions (D an $N \times N$ matrix, $m = \dim \operatorname{Ker} D$) the determinant along the path is

$$\det \widetilde{D}_t = (\cos t)^m \prod_j \big(\cos^2 t + \lambda_j^2 \sin^2 t\big),$$

which *changes sign* at $t = \pi/2$ if and only if m is odd.

To close the path into a loop we must return from -1 to $+1$ through invertible real operators. In infinite dimensions this is always possible (the group of invertible operators on a real Hilbert space is connected, by Kuiper's theorem). The result is a loop of real Fredholm operators: the family $\widetilde{D}$ over S^1. Its index bundle is computed by stabilisation (Section 4.18.4) near $t = \pi/2$: the kernel $\operatorname{Ker} D$ appears and disappears, and whether the real line $\Lambda^{\max}$ of the index bundle comes back to itself with or without a sign flip is decided by the sign change of the determinant: once per kernel dimension. Each zero mode of D contributes one copy of the Möbius twist.

Key idea: The mod 2 index as a family index (Nash's (4.128)–(4.129))

The index of the associated real family is

$$\operatorname{index} \widetilde{D} = m\,[\eta] \in \widetilde{KO}(S^1) = \mathbb{Z}_2, \qquad m = \dim \operatorname{Ker} D,$$

where $[\eta] = [\mu] - 1$ is the Möbius generator. Since $2[\eta] = 0$, only the parity of m survives:

$$\operatorname{index} \widetilde{D} = \dim \operatorname{Ker} D \bmod 2 = \operatorname{index}_1 D.$$

The mod 2 index of a single skew-adjoint operator is the index of a real family over a circle.

Physical picture: Counting sign flips

Think of the determinant of $\widetilde{D}_t$ as an orientation of the space of fields. Going once around the loop, each zero mode of D flips the orientation once (the factor $\cos t$ passes through zero). An even number of flips is no flip. The $\mathbb{Z}_2$ index says whether, after transporting around the circle, the "orientation of the fermion measure" comes back reversed. This is the mechanism of global anomalies: Witten's SU(2) anomaly is precisely a mod 2 index of a five-dimensional real operator, detecting that the fermion determinant changes sign along a loop of gauge transformations (Chapter 10).

4.19.5 No cohomological formula

The ordinary and families index theorems both have cohomological versions obtained by applying ch. Nash's final remark [Nash p. 123] is that this fails here. The index of a real family lies in $KR(Y)$, not in $H^*(Y)$, and in the present case it is a *torsion* element ($2[\eta] = 0$). As we saw in Section 4.18.9, the Chern character kills torsion:

$$\mathrm{ch}\big(c(\eta)\big) \overset{(1)}{=} \mathrm{ch}(\mu_{\mathbb{C}}) - \mathrm{ch}(1) \overset{(2)}{=} 1 - 1 = 0.$$

(1) Complexify and use additivity.

(2) $\mu_{\mathbb{C}}$ is trivial (Example 4.113); even without that, it has a flat connection, so its Chern form vanishes.

So no formula of the form "integral of a curvature polynomial" can compute $\mathrm{index}_1 D$. One might hope for a formula in mod 2 cohomology (Stiefel–Whitney classes: $w_1(\mu) \neq 0$ does detect the Möbius bundle), and in special cases such formulas exist, but there is no general local-integrand formula. This is a real limitation: *mod 2 indices are global, not local, invariants*.

$$\begin{array}{ccccc} KR(T_F Z) & \xrightarrow{\mathrm{index}_t = \mathrm{index}_a} & KO(Y) & \xrightarrow{\text{rank, } w_1, \ldots} & \mathbb{Z},\ \mathbb{Z}_2, \ldots \\ \downarrow{\scriptstyle c} & & \downarrow{\scriptstyle c} & & \\ K(T_F Z) & \xrightarrow{\mathrm{index}_t = \mathrm{index}_a} & K(Y) & \xrightarrow{\mathrm{ch}} & H^{\mathrm{even}}(Y;\mathbb{Q}) \end{array}$$

Figure 4.67: Where information is lost. A real family has an index in $KO(Y)$ (top row). Complexifying (c, down arrows) forgets torsion such as the Möbius class $\eta \in \widetilde{KO}(S^1)$, and the Chern character forgets any remaining torsion. The mod 2 index lives only in the top row, which is why it has no curvature formula.

Reading Nash (pp. 121–123, (4.128)–(4.129))

Nash defines real operators by $\overline{\sigma(x,p)} = \sigma(x,-p)$, notes that complexifying a single real operator loses nothing but that for families the map from real to complex index is not injective (our Example 4.113), and that the index of a real family lives in $KR(Y)$ (Atiyah's K-theory with involution, reducing to $KO(Y)$ for trivial involutions), with the involution $p \mapsto -p$ on $T_F Z$. He then introduces the mod 2 index $\mathrm{index}_1 D = \dim \mathrm{Ker}\, D \bmod 2$ of a real skew-adjoint operator, states that it equals the index of an associated real family $\widetilde{D}$ over S^1, (4.128)–(4.129), with $\widetilde{KR}(S^1) = \mathbb{Z}_2$ generated by $[\eta]$, and observes that torsion kills any cohomological formula. We supplied the eigenvalue-pairing argument, explicit examples on the circle (including the spin-structure dependence), the Möbius computations, and a concrete version of the family $\widetilde{D}$.

Recap

- Real operator: $\overline{\sigma(x,p)} = \sigma(x,-p)$, i.e. real coefficients.
- For families, complexification loses information: the Möbius class in $\widetilde{KO}(S^1) = \mathbb{Z}_2$ becomes trivial in $\widetilde{K}(S^1) = 0$.
- Symbols of real families live in $KR(T_F Z)$ (involution $p \to -p$); indices in $KR(Y) = KO(Y)$.
- Real skew-adjoint operators have index 0 but a mod 2 index $\dim\operatorname{Ker} D$ mod 2, invariant because eigenvalues pair as $\pm i\lambda$.
- $d/d\theta$ on the circle: $\operatorname{index}_1 = 1$ (periodic), 0 (antiperiodic).
- $\operatorname{index}_1 D = \operatorname{index}\widetilde{D} = m[\eta]$ for a family over S^1; torsion, hence no curvature formula.

4.20 Index theory and fixed points

Symmetry is the physicist's favourite tool. If a symmetry g acts on a system, we do not just count states; we count them *weighted by the eigenvalue of* g. The partition function becomes a character $\operatorname{Tr}(g\, e^{-\beta H})$, the Witten index becomes the twisted index $\operatorname{Tr}\big((-1)^F g\, e^{-\beta H}\big)$, and a degeneracy becomes a representation. The question of this section is:

> *If a symmetry g acts on a manifold and on an elliptic complex, can we compute the "g-weighted index" $\sum_p (-1)^p \operatorname{tr}(g|_{H^p})$ topologically? And what does the answer look like?*

The answer, due to Atiyah, Bott and Segal, is one of the most beautiful results in the subject. The weighted index is computed entirely by the *fixed points* of g: everything away from the fixed points cancels. For the de Rham complex this is the classical Lefschetz fixed point theorem; for the Dolbeault complex on $\mathbb{CP}^1$ it is the Weyl character formula for SU(2). In physics this is the prototype of **localisation**: the Duistermaat–Heckman formula, equivariant localisation of supersymmetric path integrals, and the evaluation of twisted indices by fixed points all follow the same pattern.

4.20.1 Maps act on cohomology: the Lefschetz number

Pullback on cohomology

A smooth map $g : X \to X$ pulls forms back: $(g^*\omega)(x) = \omega(g(x))$ with all differentials transformed by the chain rule. Pullback commutes with d (Chapter T1), so it maps closed forms to closed forms and exact forms to exact forms, and therefore acts on cohomology:

$$g_p^* : H^p(X) \longrightarrow H^p(X).$$

Choosing a basis of H^p, g_p^* is a matrix. For integer cohomology $H^p(X;\mathbb{Z})$ its entries are integers; in de Rham language, the basis is dual to a set of p-cycles and the entries are integrals of pulled-back forms over them.

Working definition 4.55 (Lefschetz number)

The **Lefschetz number** of $g : X \to X$ is [Nash p. 123, (4.130)]

$$L(g) = \sum_{p=0}^{n} (-1)^p \operatorname{tr}\big(g_p^*\big|_{H^p(X)}\big). \tag{4.186}$$

To compute it: find a basis of each cohomology group (for example a set of closed forms, one per independent cycle [a closed p-dimensional submanifold, or combination of them, with no boundary]), pull each back by g, express the result in the basis, and take traces.

Three properties follow at once.

- *Identity.* If $g = \mathrm{id}$, each g_p^* is the identity matrix, of trace $b_p = \dim H^p$, so [Nash p. 123, (4.131)]
$$L(\mathrm{id}) = \sum_p (-1)^p b_p = \chi(X). \tag{4.187}$$
- *Homotopy invariance.* Homotopic maps induce the same map on cohomology, so $L(g)$ depends only on the homotopy class of g. In particular $L(g) = \chi(X)$ for every g homotopic to the identity.
- *The fixed point theorem.* If $L(g) \neq 0$ then g has at least one fixed point (a point with $g(x) = x$). We explain why in Section 4.20.2.

Combining the last two: if $\chi(X) \neq 0$, every map homotopic to the identity has a fixed point. A rotation of S^2 ($\chi = 2$) must fix something; a rotation of the torus ($\chi = 0$) need not, and indeed a translation of T^2 fixes nothing.

Worked example 4.117: Maps of the circle

On S^1 with angle θ, $H^0 = \mathbb{R}$ (constants) and $H^1 = \mathbb{R}$, spanned by the class of $\mathrm{d}\theta$ (with $\oint \mathrm{d}\theta = 2\pi \neq 0$). Constants pull back to constants, so $\operatorname{tr} g_0^* = 1$ for every g. Hence $L(g) = 1 - \operatorname{tr} g_1^*$, and we need only the action on $\mathrm{d}\theta$.

- *Rotation* $g(\theta) = \theta + \alpha$: $g^*\mathrm{d}\theta = \mathrm{d}(\theta + \alpha) = \mathrm{d}\theta$, so $L = 1 - 1 = 0$. For $\alpha \notin 2\pi\mathbb{Z}$ there are indeed no fixed points.
- *Reflection* $g(\theta) = -\theta$: $g^*\mathrm{d}\theta = -\mathrm{d}\theta$, so $L = 1 - (-1) = 2$. Fixed points: $\theta = 0$ and $\theta = \pi$. Two of them.
- *Degree-k map* $g(e^{i\theta}) = e^{ik\theta}$: $g^*\mathrm{d}\theta = k\,\mathrm{d}\theta$, so $L = 1 - k$. Fixed points solve $e^{ik\theta} = e^{i\theta}$, i.e. $e^{i(k-1)\theta} = 1$: there are $|k - 1|$ of them. For $k = -1$ this is the reflection again.

In the last case $|L|$ equals the number of fixed points, but the sign of L is negative for $k \geq 2$. Fixed points come with signs; we now find out which.

Worked example 4.118: Maps of the sphere

On S^2, $H^0 = \mathbb{R}$ and $H^2 = \mathbb{R}$ spanned by the area form ω (with $\int \omega = 4\pi$), and $H^1 = 0$. For any g, $\int_{S^2} g^*\omega = \deg(g) \int \omega$, so $g_2^* = \deg g$ and

$$L(g) = 1 + \deg g.$$

- *Rotation* by angle α about the z-axis: deg = 1, so $L = 2$. Fixed points: the two poles.

- *Antipodal map* $x \mapsto -x$: it reverses orientation on S^2 ($\deg = -1$), so $L = 0$. Indeed it has no fixed points.
- *Any map homotopic to the identity* has $L = 2 = \chi(S^2)$ and so a fixed point. This is the "hairy ball theorem": flowing a little way along a tangent vector field gives a map homotopic to the identity, and its fixed points are the zeros of the vector field.

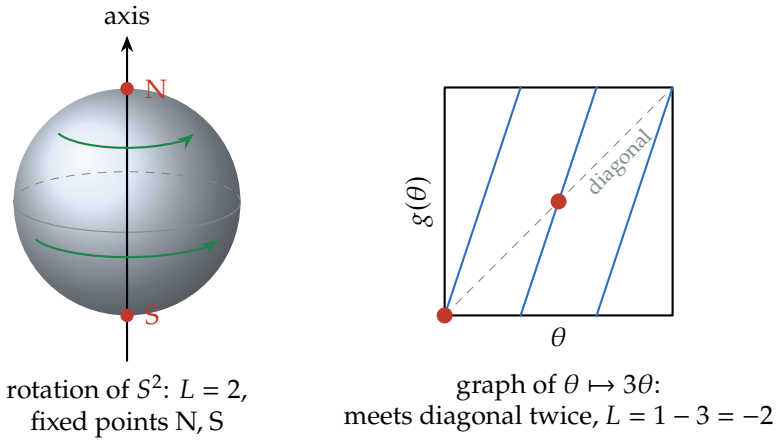


Figure 4.68: Left: a rotation of the sphere fixes exactly the two poles, and $L(g) = 2 = \chi(S^2)$. Right: the graph of the degree-3 map of the circle (blue) crosses the diagonal (fixed points, red) at $|k-1| = 2$ points, counting $\theta = 0 \equiv 2\pi$ once. Each crossing is from below to above, steeper than the diagonal, and contributes -1, giving $L = -2$.

4.20.2 Why fixed points: the Lefschetz–Hopf formula

Nash states that $L(g) \neq 0$ implies a fixed point and, for isolated fixed points, that $\chi(X)$ equals the number of fixed points. The precise statement assigns a sign to each fixed point.

Key idea: Lefschetz–Hopf fixed point formula

If $g : X \to X$ has only isolated fixed points P, each *non-degenerate* ($\det(1 - \mathrm{d}g_P) \neq 0$, i.e. 1 is not an eigenvalue of the Jacobian at P), then

$$L(g) = \sum_{P:\, g(P)=P} \operatorname{sgn} \det\bigl(1 - \mathrm{d}g_P\bigr). \tag{4.188}$$

In particular if there are no fixed points, $L(g) = 0$; contrapositively, $L(g) \neq 0$ forces a fixed point.

Why it is true (the physicist's picture). Consider the graph $\Gamma_g = \{(x, g(x))\}$ and the diagonal $\Delta = \{(x, x)\}$ in $X \times X$. Fixed points are the intersection points $\Gamma_g \cap \Delta$. The Lefschetz number is the *signed intersection number* $\Gamma_g \cdot \Delta$: it can be computed in cohomology (giving (4.186): the Poincaré dual of Δ is $\sum_i e_i \otimes e^i$ in terms of dual bases, and pairing it with the graph produces the traces), or geometrically by counting intersections with orientation signs (giving (4.188)). The sign at P compares the orientation of $T\Gamma_g \oplus T\Delta$ with that of $T(X \times X)$; in coordinates $T\Gamma_g$ is spanned by $(v, \mathrm{d}g\, v)$ and $T\Delta$ by (w, w), and the determinant of the matrix $\left(\begin{smallmatrix} 1 & 1 \\ \mathrm{d}g & 1 \end{smallmatrix}\right)$ is $\det(1 - \mathrm{d}g)$.

Worked example 4.119: Checking the signs

Degree-k map of the circle. At each fixed point $\mathrm{d}g = k$, so each contributes $\operatorname{sgn}(1-k)$. With $|k-1|$ fixed points,

$$\sum_P \operatorname{sgn}\det(1-\mathrm{d}g_P) \overset{(1)}{=} |k-1|\,\operatorname{sgn}(1-k) \overset{(2)}{=} 1-k,$$

(1) All fixed points have the same Jacobian k.
(2) $|x|\operatorname{sgn}(x) = x$ with $x = 1-k$.

matching $L = 1-k$ from Example 4.117.
Rotation of S^2 by $\alpha \notin 2\pi\mathbb{Z}$. At the north pole, in a tangent plane with the outward orientation, $\mathrm{d}g$ is the rotation matrix $R(\alpha)$. At the south pole it is $R(-\alpha)$ in the outward orientation. Either way

$$\det\bigl(1-R(\pm\alpha)\bigr) \overset{(1)}{=} \det\begin{pmatrix} 1-\cos\alpha & \pm\sin\alpha \\ \mp\sin\alpha & 1-\cos\alpha \end{pmatrix} \overset{(2)}{=} (1-\cos\alpha)^2+\sin^2\alpha \overset{(3)}{=} 2-2\cos\alpha > 0.$$

(1) $R(\alpha) = \left(\begin{smallmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{smallmatrix}\right)$.
(2) 2×2 determinant.
(3) $\cos^2+\sin^2 = 1$.

Each pole contributes $+1$: total $2 = L(g)$. *Identity perturbed by a vector field.* If g is the time-ε flow of a vector field V with non-degenerate zeros, then $\mathrm{d}g_P \approx 1+\varepsilon\,\partial V(P)$ and $\operatorname{sgn}\det(1-\mathrm{d}g_P) = \operatorname{sgn}\det(-\partial V(P)) = (-1)^n \operatorname{sgn}\det\partial V(P)$. Up to the overall sign $(-1)^n$ (which is $+1$ in even dimensions) this is the index of the zero of V, and (4.188) with (4.187) becomes the Poincaré–Hopf theorem: the sum of the indices of the zeros of a vector field is $\chi(X)$.

Watch out: "The number of fixed points equals χ"

Nash says that for isolated fixed points of a map homotopic to the identity, $\chi(X)$ is the number of fixed points. This is true when every fixed point has sign $+1$ (for instance, rotations of S^2, or generic isometries). In general each fixed point counts with the sign (4.188): the flow of $V = x\partial_x - y\partial_y$ near a saddle contributes -1.

4.20.3 Lefschetz numbers for elliptic complexes

The Euler characteristic is the index of the de Rham complex, and $L(\mathrm{id}) = \chi$. This suggests generalising from de Rham to any elliptic complex [Nash p. 124, (4.132)]

$$0 \to \Gamma(X,E^0) \xrightarrow{d_0} \Gamma(X,E^1) \xrightarrow{d_1} \cdots \xrightarrow{d_{N-1}} \Gamma(X,E^N) \to 0, \qquad d_i \circ d_{i-1} = 0.$$

Working definition 4.56 (invariant complex and generalised Lefschetz number)

Let $g : X \to X$ be a map and suppose we are given linear maps $T_i : \Gamma(X,E^i) \to \Gamma(X,E^i)$ that commute with the operators [Nash p. 124, (4.133)]:

$$d_i T_i = T_{i+1} d_i.$$

Then the complex is **invariant** under g. Because T commutes with d, it maps kernels to kernels and images to images, and so acts on each cohomology group $H^i(E)$ [$\operatorname{Ker} d_i/\operatorname{Im} d_{i-1}$, solutions of $d_i s = 0$ modulo those of the form $d_{i-1}t$]; call this action $g_i^*(E)$. The **generalised**

Lefschetz number is [Nash p. 124, (4.134)]

$$L(g,E)=\sum_i (-1)^i \operatorname{tr}\big(g_i^*(E)\big). \tag{4.189}$$

Typical source of the T_i: g lifts to the bundles ($\phi_i : E^i_{g(x)} \to E^i_x$, linear), and $T_i s(x) = \phi_i\, s(g(x))$. For the de Rham complex, $\phi_i = \Lambda^i(\mathrm{d}g)^T$ [the transposed Jacobian of g acting on each covector factor of an i-form] and $T_i = g^*$, giving back (4.186).

For $g = \mathrm{id}$ and $T_i = \mathrm{id}$, $L(\mathrm{id},E) = \operatorname{index} E$: *the generalised Lefschetz number is a g-weighted index.*

Group actions and characters

To have many g's at once, let a compact Lie group G act on X, i.e. a continuous map $G\times X\to X$, $(g,x)\mapsto g\cdot x$, with $(gh)\cdot x = g\cdot(h\cdot x)$. If the complex is invariant under every $g\in G$, each $H^i(E)$ becomes a finite-dimensional *representation* of G, and

$$g \longmapsto L(g,E)=\sum_i (-1)^i \chi_{H^i(E)}(g)$$

is an alternating sum of *characters*. Nash's remark that $L(g,E)$ "is a little reminiscent of the character of a group representation" [Nash p. 124] is exactly right: it *is* a (virtual) character. Nash's footnote observes that the maps are then all invertible (g^{-1} is in G too), which is more special than general fixed point theory but is precisely the situation in which index theory is powerful.

Working definition 4.57 (representation ring)

The **representation ring** $R(G)$ consists of formal differences $[V]-[W]$ of finite-dimensional complex representations, with $\oplus$ and $\otimes$. An element is determined by its *character* $g\mapsto \operatorname{tr}_V(g)-\operatorname{tr}_W(g)$, a class function [unchanged under $g\mapsto hgh^{-1}$, since $\operatorname{tr}(hgh^{-1})=\operatorname{tr} g$] on G.

Worked example 4.120: Representation rings

- $G=\mathrm{U}(1)$, $g=\lambda=e^{i\alpha}$. The irreducible representations are one-dimensional, $\lambda\mapsto\lambda^j$, $j\in\mathbb{Z}$. So $R(\mathrm{U}(1))=\mathbb{Z}[\lambda,\lambda^{-1}]$: characters are Laurent polynomials. The virtual representation $2-\lambda^3$ is "two copies of the trivial representation minus one of charge 3".
- $G=\mathbb{Z}_2=\{1,g\}$: two irreducibles (trivial, sign), characters $(1,1)$ and $(1,-1)$. $R(\mathbb{Z}_2)=\mathbb{Z}[\varepsilon]/(\varepsilon^2-1)$.
- $G=\mathrm{SU}(2)$: irreducibles V_s of spin $s=0,\tfrac12,1,\dots$, with character on $\operatorname{diag}(e^{i\alpha/2},e^{-i\alpha/2})$

$$\chi_s(\alpha)=\sum_{m=-s}^{s} e^{im\alpha}=\frac{\sin\big((s+\tfrac12)\alpha\big)}{\sin(\alpha/2)}.$$

Equivariant K-theory

To formulate an index theorem one needs K-theory that remembers the group action [Nash p. 125].

Working definition 4.58 (G-vector bundle and $K_G(X)$)

Let X be a G-space [each $g \in G$ moves points by a map $x \mapsto g \cdot x$, compatibly with the group law]. A **G-vector bundle** is a vector bundle $\pi : F \to X$ on which G also acts, such that

- the action covers the action on X: $\pi(g \cdot f) = g \cdot \pi(f)$;
- g maps fibres *linearly*: $F_x \to F_{g\cdot x}$ is linear [Nash (4.135)].

$K_G(X)$ is the K-theory of G-vector bundles (formal differences, as for $K(X)$).

Two special cases identify it with familiar objects.

- $G = \{e\}$ trivial: every bundle is a G-bundle, so $K_{\{e\}}(X) = K(X)$.
- $X = \{x_0\}$ a point: a G-bundle over a point is a vector space with a linear G-action, i.e. a representation. So $K_G(\mathrm{pt}) = R(G)$.

Worked example 4.121: The monopole bundles as $\mathrm{U}(1)$-bundles

Let U(1) act on $\mathbb{CP}^1$ by $\lambda \cdot z = \lambda z$ (rotation about the z-axis). The bundle $\mathcal{O}(k)$ (sections: pairs $s_0(z)$, $s_\infty(w) = w^k s_0(1/w)$) becomes a U(1)-bundle by

$$(\lambda \cdot s)_0(z) = s_0(\lambda z).$$

In the other chart this reads, for $s_0 = z^j$, $(\lambda \cdot s)_\infty(w) = \lambda^j w^{k-j} = \lambda^k\, s_\infty(\lambda^{-1} w)$: the action pulls back along $w \mapsto \lambda^{-1} w$ and multiplies the fibre by λ^k. At the two fixed points the group acts on the fibre by

$$\text{at } z = 0 :\ 1, \qquad \text{at } z = \infty :\ \lambda^k.$$

These two numbers are all the fixed point formula below will need. (The lift is not unique: multiplying by an overall character λ^c gives another lift. Choosing $c = -k/2$, when allowed, makes the picture symmetric under $z \to 1/z$ and turns the action into an SU(2) one.)

4.20.4 The G-index theorem

Analytic index

Let D be a G-invariant elliptic operator (it commutes with the G action on sections). Its symbol, being built from G-invariant data, defines a class $[\sigma_m(D)] \in K_G(T^*X)$. The kernel and cokernel are finite-dimensional representations, and Nash defines [Nash p. 126, (4.136)–(4.137)]

$$\mathrm{index}_G D = [\mathrm{Ker}\, D] - [\mathrm{Coker}\, D] \in R(G), \qquad \mathrm{index}_a : K_G(T^*X) \to R(G),\ [\sigma_m(D)] \mapsto \mathrm{index}_G D. \tag{4.190}$$

Evaluating the character at $g = e$ gives back the ordinary index, $\dim \mathrm{Ker}\, D - \dim \mathrm{Coker}\, D$.

Topological index

The topological index is built as before, with one change: X must be embedded G-equivariantly, and $\mathbb{R}^N$ is replaced by a real *representation space* V of G (by a theorem of Mostow and Palais this is always possible). With $i : X \hookrightarrow V$ and $j : P \hookrightarrow V$ the inclusion of the origin (a point,

necessarily fixed by G), the diagram is [Nash p. 126, (4.138)]

$$K_G(T^*X) \xrightarrow{i_!} K_G(T^*V) \xleftarrow[\cong]{j_!} K_G(T^*P)\ , \qquad \operatorname{index}_t = (j_!)^{-1} \circ i_!. \tag{4.191}$$

Since P is a point, $T^*P = P$ and $K_G(P) = R(G)$ rather than $\mathbb{Z}$ [Nash (4.139)]. So the topological index is automatically character-valued: $\operatorname{index}_t : K_G(T^*X) \to R(G)$. The map $j_!$ is the equivariant Bott (Thom) isomorphism, which is where the representation theory enters.

Worked example 4.122: An equivariant embedding

$S^2 \subset \mathbb{R}^3$ as the unit sphere is SO(3)-equivariant: SO(3) acts on $V = \mathbb{R}^3$ by its defining representation, and rotations of the sphere are restrictions of this linear action. The origin $P = 0$ is fixed. For the circle action of Example 4.121, $V = \mathbb{R}^3 = \mathbb{C} \oplus \mathbb{R}$ with λ acting by multiplication on $\mathbb{C}$ and trivially on $\mathbb{R}$.

Key idea: The G-index theorem and the Atiyah–Segal formula

For a G-invariant elliptic operator (or complex),

$$\operatorname{index}_a = \operatorname{index}_t : K_G(T^*X) \longrightarrow R(G).$$

Taking the character at $g \in G$, the generalised Lefschetz number is [Nash p. 126, (4.140)]

$$L(g, E) = \chi(g), \qquad \chi = \operatorname{index}_t(u), \quad u = [\sigma(E)] \in K_G(T^*X). \tag{4.192}$$

The right side can be evaluated by a cohomological formula which involves only the *fixed point set* $X^g = \{x : g \cdot x = x\}$: it is an integral over X^g of characteristic classes of X^g, of its normal bundle, and of the restrictions of the E^i, with g acting on the normal bundle and on the fibres.

For $g = e$, $X^e = X$ and the formula is the ordinary cohomological index formula. For g with isolated fixed points, X^g is a finite set, the "integral over X^g" is a finite sum, and the formula becomes remarkably simple. We now derive it.

4.20.5 Why only fixed points contribute: heat-kernel localisation

Recall the heat-kernel form of the index (McKean–Singer): for an elliptic complex with Laplacians $\Delta_i = d_i^* d_i + d_{i-1} d_{i-1}^*$, $\operatorname{index} E = \sum_i (-1)^i \operatorname{Tr} e^{-t\Delta_i}$ for every $t > 0$. The non-zero eigenvalues cancel in the alternating sum, because d maps a Δ_i-eigenform to a Δ_{i+1}-eigenform with the same eigenvalue. If T commutes with d (and with d^*, e.g. if g is an isometry), the same cancellation holds with T inserted:

$$L(g, E) = \sum_i (-1)^i \operatorname{Tr}\big(T_i\, e^{-t\Delta_i}\big) \qquad \text{for every } t > 0. \tag{4.193}$$

Now write the trace with the heat kernel $K_t^i(x, y)$: $\operatorname{Tr}(T_i e^{-t\Delta_i}) = \int_X \operatorname{tr}\big(\phi_i\, K_t^i(g(x), x)\big)\, \mathrm{d}x$. For small t the heat kernel is sharply peaked on the diagonal:

$$K_t(y, x) \approx \frac{1}{(4\pi t)^{n/2}}\, e^{-|y-x|^2/4t} \cdot \big(1 + O(t)\big).$$

The integrand involves $K_t(g(x), x)$, which is exponentially small unless $g(x)$ is close to x. As $t \to 0$ *only neighbourhoods of the fixed points contribute*. Near a non-degenerate fixed point P, in

coordinates centred at P, $g(x) \approx Ax$ with $A = \mathrm{d}g_P$. Then

$$\begin{aligned}\int_{\mathbb{R}^n} \frac{e^{-|Ax-x|^2/4t}}{(4\pi t)^{n/2}}\,\mathrm{d}^n x &\overset{(1)}{=} \int_{\mathbb{R}^n} \frac{e^{-|y|^2/4t}}{(4\pi t)^{n/2}}\,\frac{\mathrm{d}^n y}{|\det(1-A)|} \\ &\overset{(2)}{=} \frac{1}{|\det(1-A)|}\prod_{k=1}^{n}\int_{\mathbb{R}} \frac{e^{-y_k^2/4t}}{(4\pi t)^{1/2}}\,\mathrm{d}y_k \\ &\overset{(3)}{=} \frac{1}{|\det(1-A)|}.\end{aligned} \tag{4.194}$$

(1) Change variables $y = (A-1)x$; non-degeneracy means $1-A$ is invertible, and $\mathrm{d}^n y = |\det(1-A)|\,\mathrm{d}^n x$.

(2) The Gaussian factorises over the components of y.

(3) $\int_{\mathbb{R}} e^{-y^2/4t}\,\mathrm{d}y = \sqrt{4\pi t}$.

The answer does not depend on t, as it must not. Including the fibre maps $\phi_{i,P}$ at the fixed point, we arrive at:

Key idea: Atiyah–Bott fixed point formula

If g has isolated non-degenerate fixed points, then

$$L(g,E) = \sum_{P:\,g(P)=P} \frac{\sum_i (-1)^i \operatorname{tr}(\phi_{i,P})}{\left|\det(1-\mathrm{d}g_P)\right|}, \tag{4.195}$$

where $\phi_{i,P} : E^i_P \to E^i_P$ is the lifted action on the fibre over the fixed point. Everything away from the fixed points cancels; each fixed point contributes a rational function of the eigenvalues of g there.

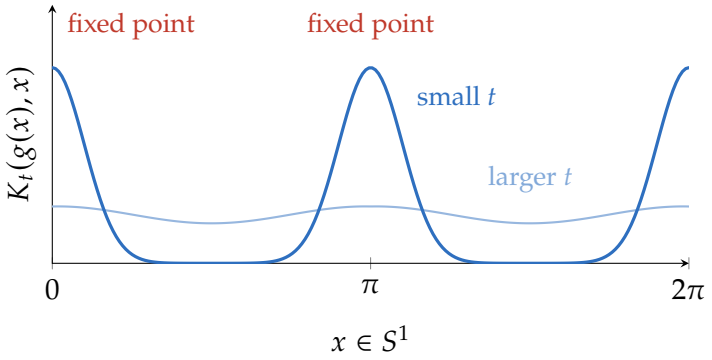


Figure 4.69: Localisation. For the reflection $g(x) = -x$ of the circle, the integrand $K_t(g(x),x)$ of the g-twisted heat trace is peaked where $g(x) = x$, i.e. at the fixed points 0 and π. As $t \to 0$ the peaks narrow (dark curve), but their areas, $1/|\det(1-\mathrm{d}g)| = \frac{1}{2}$ each, stay fixed. The whole Lefschetz number is carried by the fixed points.

Recovering the classical formula

For the de Rham complex, $E^i = \Lambda^i T^*X$ and $\phi_{i,P} = \Lambda^i A^T$, with $A = \mathrm{d}g_P$. The numerator is the alternating sum of traces of the exterior powers of A^T, which is a characteristic polynomial:

$$\sum_i (-1)^i \operatorname{tr}\Lambda^i A^T \overset{(1)}{=} \prod_{k=1}^{n}(1-a_k) \overset{(2)}{=} \det(1-A). \tag{4.196}$$

(1) If A has eigenvalues $a_1, \dots, a_n$, then $\Lambda^i A^T$ has eigenvalues $a_{k_1} \cdots a_{k_i}$ ($k_1 < \cdots < k_i$), so $\operatorname{tr} \Lambda^i A^T = e_i(a)$, the i-th elementary symmetric polynomial, and $\sum_i (-1)^i e_i(a) = \prod_k (1 - a_k)$.
(2) The determinant is the product of the eigenvalues.

So the Atiyah–Bott contribution is $\det(1-A)/|\det(1-A)| = \operatorname{sgn}\det(1-A)$, which is precisely the Lefschetz–Hopf formula (4.188).

Worked example 4.123: The reflection of the circle, by heat kernels

Let $g(\theta) = -\theta$ on S^1 and check (4.193) directly. On functions use the basis $1, \cos n\theta, \sin n\theta$ ($n \geq 1$), eigenfunctions of $\Delta = -\mathrm{d}^2/\mathrm{d}\theta^2$ with eigenvalue n^2. Then $g^* \cos n\theta = \cos n\theta$ and $g^* \sin n\theta = -\sin n\theta$, so

$$\operatorname{Tr}\big(g^* e^{-t\Delta_0}\big) \overset{(1)}{=} 1 + \sum_{n \geq 1} \big(e^{-tn^2} - e^{-tn^2}\big) = 1.$$

(1) Diagonal entries: $+1$ on $\cos n\theta$, -1 on $\sin n\theta$, each weighted by e^{-tn^2}; the constant contributes 1.

On 1-forms, $g^*(f(\theta)\mathrm{d}\theta) = -f(-\theta)\,\mathrm{d}\theta$, so $\mathrm{d}\theta \mapsto -\mathrm{d}\theta$, $\cos n\theta\,\mathrm{d}\theta \mapsto -\cos n\theta\,\mathrm{d}\theta$, $\sin n\theta\,\mathrm{d}\theta \mapsto +\sin n\theta\,\mathrm{d}\theta$, and $\operatorname{Tr}(g^* e^{-t\Delta_1}) = -1$. Hence

$$L(g) = 1 - (-1) = 2 \quad \text{for every } t.$$

By Atiyah–Bott: at each fixed point $A = -1$, $|\det(1-A)| = 2$, and the numerator is $\operatorname{tr}\phi_0 - \operatorname{tr}\phi_1 = 1 - (-1) = 2$. Each fixed point contributes $2/2 = 1$; two fixed points give 2. Figure 4.69 shows how the heat-kernel integrand localises.

The holomorphic Lefschetz formula

For the Dolbeault complex of a complex manifold of complex dimension ℓ, with a holomorphic g, $E^i = \Lambda^{0,i}$ and the numerator involves the *antiholomorphic* part $\bar{A}$ of $\mathrm{d}g_P$, while the real Jacobian has $|\det_{\mathbb{R}}(1 - \mathrm{d}g_P)| = |\det_{\mathbb{C}}(1-A)|^2$. Therefore, for the bundle F twisted in, with fibre action ϕ_P on F_P:

$$\nu(P) \overset{(1)}{=} \frac{\operatorname{tr}\phi_P \ \sum_i (-1)^i \operatorname{tr}\Lambda^i \bar{A}}{|\det_{\mathbb{C}}(1-A)|^2} \overset{(2)}{=} \frac{\operatorname{tr}\phi_P \ \det(1-\bar{A})}{\det(1-A)\,\det(1-\bar{A})} \overset{(3)}{=} \frac{\operatorname{tr}\phi_P}{\det_{\mathbb{C}}(1-A)}. \tag{4.197}$$

(1) Atiyah–Bott (4.195) for $E^i = F \otimes \Lambda^{0,i}$ (traces multiply on tensor products).
(2) (4.196) applied to $\bar{A}$; and $|z|^2 = z\bar{z}$ with $z = \det(1-A)$.
(3) Cancel $\det(1-\bar{A})$ (non-zero by non-degeneracy).

This is the **holomorphic Lefschetz formula**. (Here A is the holomorphic derivative of g at P, i.e. its action on $T^{1,0}_P$.)

Worked example 4.124: Rotating $\mathbb{CP}^1$: the Weyl character formula

Take $g(z) = \lambda z$, $\lambda = e^{i\alpha} \neq 1$, on $\mathbb{CP}^1$ acting on $\mathcal{O}(k)$ as in Example 4.121. *Direct computation.* For $k \geq 0$, $H^0(\mathbb{CP}^1; \mathcal{O}(k)) = \operatorname{span}\{1, z, \dots, z^k\}$ and $H^1 = 0$. The action $(\lambda \cdot s)(z) = s(\lambda z)$ sends $z^j \mapsto \lambda^j z^j$. So

$$L(g, \bar{\partial}_{\mathcal{O}(k)}) = \sum_{j=0}^{k} \lambda^j.$$

Fixed points. g fixes $z = 0$ and $z = \infty$.

- At $z = 0$: $A = \mathrm{d}g = \lambda$, and the fibre action is 1.
- At $z = \infty$, in the coordinate $w = 1/z$: $g(w) = \lambda^{-1}w$, so $A = \lambda^{-1}$, and the fibre action is λ^{k}.

The holomorphic Lefschetz formula (4.197) gives

$$L(g, \bar\partial_{O(k)}) \overset{(1)}{=} \frac{1}{1-\lambda} + \frac{\lambda^{k}}{1-\lambda^{-1}} \overset{(2)}{=} \frac{1}{1-\lambda} - \frac{\lambda^{k+1}}{1-\lambda} \overset{(3)}{=} \frac{1-\lambda^{k+1}}{1-\lambda} \overset{(4)}{=} \sum_{j=0}^{k} \lambda^{j}.$$

(1) One term per fixed point: numerator the fibre action, denominator $1 - A$.
(2) Multiply numerator and denominator of the second term by $-\lambda$: $\frac{\lambda^{k}}{1-\lambda^{-1}} = \frac{-\lambda^{k+1}}{1-\lambda}$.
(3) Common denominator.
(4) Geometric series.

The two fixed points reproduce the full character. Expanding the first term as $1+\lambda+\lambda^{2}+\cdots$ and the second as $-\lambda^{k+1} - \lambda^{k+2} - \cdots$, the "tail" of the north pole's geometric series is cancelled by the south pole's, leaving exactly the $k+1$ weights (Figure 4.70).
With the symmetric lift (multiply by $\lambda^{-k/2}$, $\lambda = e^{i\alpha}$, $s = k/2$) this reads

$$\frac{\lambda^{-s} - \lambda^{s+1}}{1-\lambda} = \frac{\sin\!\big((s+\tfrac{1}{2})\alpha\big)}{\sin(\alpha/2)} = \chi_{s}(\alpha),$$

the Weyl character formula for spin s. Setting $\lambda \to 1$ gives $k+1$, the Riemann–Roch index of Example 4.98.

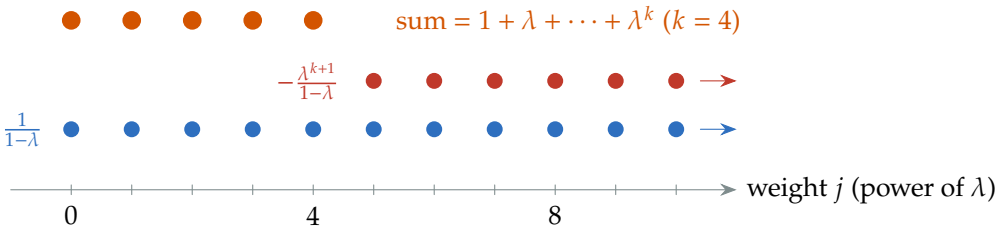


Figure 4.70: The holomorphic Lefschetz formula on $\mathbb{CP}^{1}$ for $\mathcal{O}(4)$. The north pole contributes the infinite geometric series of all weights $j \geq 0$ (blue); the south pole contributes minus all weights $j \geq k+1$ (red). The infinite tails cancel and exactly the $k+1 = 5$ weights of H^{0} survive (orange). This is the Weyl character formula as a fixed point formula.

Worked example 4.125: Spinors on a rotating sphere

The (untwisted) Dirac operator on the round S^{2} has no zero modes (Lichnerowicz: $\not{D}^{2} = \nabla^{*}\nabla + \frac{R}{4}$ with $R > 0$), so its Lefschetz number vanishes for every rotation. Check with Atiyah–Bott. The spin complex has $E^{0} = S^{+}$, $E^{1} = S^{-}$. A rotation by α of a tangent plane lifts to spinors as $e^{\pm i\alpha/2}$ on $S^{\pm}$ (the lift of $R(\alpha)$ to Spin(2), chosen continuously from the identity). At a fixed point where $\mathrm{d}g = R(\alpha)$ in the oriented tangent plane:

$$\nu \overset{(1)}{=} \frac{e^{i\alpha/2} - e^{-i\alpha/2}}{2 - 2\cos\alpha} \overset{(2)}{=} \frac{2i\sin(\alpha/2)}{4\sin^{2}(\alpha/2)} \overset{(3)}{=} \frac{i}{2\sin(\alpha/2)}.$$

(1) Atiyah–Bott with numerator $\operatorname{tr}\phi_{0} - \operatorname{tr}\phi_{1}$ and denominator from Example 4.119.
(2) $e^{ix} - e^{-ix} = 2i\sin x$ and $1 - \cos\alpha = 2\sin^{2}(\alpha/2)$.

(3) Cancel.

At the north pole the rotation is $R(\alpha)$ in the outward orientation; at the south pole it is $R(-\alpha)$. So the total is $\frac{i}{2\sin(\alpha/2)} + \frac{i}{2\sin(-\alpha/2)} = 0$, as it must be. Coupled to the monopole bundle L_k, the two contributions acquire different fibre factors and no longer cancel; the result is the character of the k zero modes of Example 4.102, which form the spin-$\frac{k-1}{2}$ representation of SU(2).

Physical picture: Twisted indices and localisation in physics

In a supersymmetric quantum mechanics with a symmetry g commuting with the supercharge, the twisted Witten index $\mathrm{Tr}\big((-1)^F g\, e^{-\beta H}\big)$ is independent of β (non-zero-energy states cancel in boson–fermion pairs, exactly as in (4.193)). Computed as a path integral over loops $x(\tau)$ with twisted boundary condition $x(\beta) = g\,x(0)$, the $\beta \to 0$ limit forces $x(0)$ to be a fixed point of g, and the Gaussian fluctuation integral around it is (4.194). This is the Atiyah–Bott formula as a path integral, and it is the simplest example of equivariant localisation, the technique behind Duistermaat–Heckman, Nekrasov's instanton partition functions and supersymmetric localisation.

Reading Nash (pp. 123–127, (4.130)–(4.140))

Nash recalls the classical Lefschetz number (4.130) and $L(\mathrm{id}) = \chi(X)$ (4.131); generalises to invariant elliptic complexes (4.132)–(4.134); introduces compact group actions, G-vector bundles (4.135), equivariant K-theory with $K_{\{e\}}(X) = K(X)$ and $K_G(\mathrm{pt}) = R(G)$; the character-valued analytic index (4.136)–(4.137); the topological index via an equivariant embedding into a representation space, diagram (4.138), (4.139); and the Atiyah–Segal theorem $L(g, E) = \chi(g)$ (4.140). He does not write the fixed point formula, saying only that it is an evaluation on X^g that simplifies when X^g is finite. We supplied: the explicit circle and sphere computations, the Lefschetz–Hopf signs, the heat-kernel derivation of the Atiyah–Bott formula, its de Rham, holomorphic and spin versions, and the Weyl character formula on $\mathbb{CP}^1$.

Recap

- $L(g) = \sum (-1)^p \,\mathrm{tr}\, g^*|_{H^p}$; $L(\mathrm{id}) = \chi(X)$; $L(g) \neq 0 \Rightarrow$ fixed point; $L = \sum_P \mathrm{sgn}\det(1 - \mathrm{d}g_P)$.
- Circle: rotation $L = 0$, reflection $L = 2$, degree k: $L = 1 - k$. Sphere: $L = 1 + \deg g$.
- For G-invariant complexes, $L(g, E)$ is a virtual character; $\mathrm{index}_G D \in R(G) = K_G(\mathrm{pt})$; G-index theorem $\mathrm{index}_a = \mathrm{index}_t$.
- Atiyah–Bott: $L(g, E) = \sum_P \frac{\sum (-1)^i \,\mathrm{tr}\, \phi_{i,P}}{|\det(1 - \mathrm{d}g_P)|}$, derived by heat-kernel localisation.
- Holomorphic version on $\mathbb{CP}^1$ gives the Weyl character formula.

4.21 Index theory for manifolds with boundary

Everything so far has lived on a *closed* manifold: compact, with no edge. Physics is full of edges. A bag model confines quarks to a ball. A Hall bar has an edge where chiral modes run. A lattice simulation lives on a box; a domain wall is an edge between two phases; a spacetime region between two time slices has the initial and final slices as its boundary. So the question of

this section is:

> *What becomes of the index theorem when X has a boundary ∂X? What boundary conditions make the index finite, and what extra term does the boundary contribute to the index formula?*

The answer, due to Atiyah, Patodi and Singer (APS), is one of the most beautiful results in the subject. It has three ingredients.

1. For the operators physicists care about (Dirac, signature, $\bar\partial$) there is *no* local boundary condition that works. We must use a *global* one: a spectral projection onto half of the spectrum of an operator L that lives on the boundary.
2. With this condition the index is the usual bulk integral *plus a correction* that depends only on the spectrum of L.
3. The correction is $-\frac{1}{2}(h + \eta(0))$, where $h = \dim\operatorname{Ker} L$ and $\eta(0)$ measures the *asymmetry* between the positive and negative eigenvalues of L: the **eta invariant**.

The eta invariant then turns out to be a physical quantity in its own right: it is the fractional fermion number of a Dirac sea, the phase of a three-dimensional fermion determinant (the parity anomaly), and, through its jumps, it counts the *spectral flow* that underlies anomalous charge non-conservation. We build up to all of this slowly.

4.21.1 Why a boundary forces us to choose boundary conditions

On a closed manifold an elliptic operator has a finite-dimensional kernel. With a boundary, this fails badly.

Worked example 4.126: $\bar\partial$ on a disc has an infinite kernel

Let $X = \{|z| \le 1\}$ and $D = \bar\partial = \frac{1}{2}(\partial_x + i\partial_y)$ acting on functions. Its kernel is the space of holomorphic functions on the disc. This contains $1, z, z^2, z^3, \dots$, which are linearly independent. So $\dim\operatorname{Ker}\bar\partial = \infty$. The same operator on the closed sphere $\mathbb{CP}^1$ has kernel the constants only (Liouville: a holomorphic function on a compact Riemann surface is constant).

What went wrong? On a closed manifold, a zero mode must be globally consistent; this is a strong constraint. On a disc, we can start with *any* boundary data compatible with the equation and extend it inwards. The kernel is as big as a space of functions on the boundary. To get back a finite index, we must *throw away* a suitable half of this boundary data: that is what a boundary condition does.

Local boundary conditions: Dirichlet and Neumann

The classical elliptic boundary problems are *local*: the condition at a boundary point x involves only the values (and finitely many derivatives) of u at x. Nash's two examples [Nash p. 127, (4.141)–(4.142)], for the Laplacian Δ on a region $\Omega \subset \mathbb{R}^n$ with inward or outward normal n:

$$\Delta u = 0, \qquad u = g \text{ on } \partial\Omega \qquad \text{(Dirichlet)}, \tag{4.198}$$

$$\Delta u = 0, \qquad \frac{\partial u}{\partial n} = g \text{ on } \partial\Omega \qquad \text{(Neumann)}. \tag{4.199}$$

In electrostatics: Dirichlet prescribes the potential on a conductor; Neumann prescribes the normal electric field (the surface charge).

An elliptic boundary problem is a *pair* of operators, one in the interior and one on the boundary. For the index we regard it as a single map

$$\mathcal{P} = (\Delta, B):\ C^\infty(\Omega) \longrightarrow C^\infty(\Omega) \oplus C^\infty(\partial\Omega), \qquad u \longmapsto \big(\Delta u,\ Bu|_{\partial\Omega}\big), \tag{4.200}$$

with $B = 1$ (Dirichlet) or $B = \partial/\partial n$ (Neumann). Its kernel is the set of solutions of the *homogeneous* problem; its cokernel is the set of constraints the data (f, g) must satisfy for a solution to exist.

Worked example 4.127: Dirichlet and Neumann on the unit disc, by Fourier series

Take Ω the unit disc, polar coordinates (r, θ), and expand the boundary data $g(\theta) = \sum_{m\in\mathbb{Z}} g_m e^{im\theta}$. The Laplacian in polar coordinates is $\Delta = \partial_r^2 + \frac{1}{r}\partial_r + \frac{1}{r^2}\partial_\theta^2$. Try $u = r^\alpha e^{im\theta}$:

$$\Delta\big(r^\alpha e^{im\theta}\big) \overset{(1)}{=} \big(\alpha(\alpha-1) + \alpha - m^2\big) r^{\alpha-2} e^{im\theta} \overset{(2)}{=} (\alpha^2 - m^2)\, r^{\alpha-2} e^{im\theta}.$$

(1) Each ∂_r lowers the power of r by one; ∂_θ^2 brings down $-m^2$.
(2) Simplify.

So $\alpha = \pm|m|$. Only $\alpha = +|m|$ is regular at the centre (and for $m = 0$ the second solution is $\log r$, also singular). Hence every harmonic function on the disc is $u = \sum_m c_m r^{|m|} e^{im\theta}$.
Dirichlet. $u|_{r=1} = g$ forces $c_m = g_m$ for all m: a unique solution for every g. Kernel 0, cokernel 0, index 0.
Neumann (outward normal ∂_r). $\partial_r u|_{r=1} = \sum_m |m| c_m e^{im\theta} = g$. For $m \neq 0$, $c_m = g_m/|m|$. For $m = 0$ the left side has no term, so we need $g_0 = 0$, i.e. $\int_0^{2\pi} g\, \mathrm{d}\theta = 0$; and c_0 is free. Kernel: the constants (dimension 1). Cokernel: the single constraint $\int g = 0$ (dimension 1). Index $1 - 1 = 0$.
The Neumann constraint is Gauss's law: $\oint \partial_n u\, \mathrm{d}s = \int_\Omega \Delta u = 0$, so with no charge inside the total flux through the boundary must vanish.

Both problems have finite kernel and cokernel, and both are *local*. Nash's point is that this is special. For the operators of index theory it is impossible.

A test for good boundary conditions: count the decaying modes

How do we know whether a boundary condition gives a finite index? There is a very physical test. Zoom in on a boundary point until the boundary looks flat and the coefficients look constant. Take the half-space $y > 0$ with boundary coordinate x, and Fourier transform along the boundary, $u = e^{i\xi x} v(y)$. The equation $Du = 0$ becomes an ODE in the normal variable y. Some of its solutions decay into the interior ($y \to +\infty$); some grow. The decaying ones are the modes that are "localised at the boundary", and those are what the boundary condition must control.

Working definition 4.59 (good (elliptic) boundary condition — decaying-mode test)

For each boundary momentum $\xi \neq 0$ let $M^+(\xi)$ be the space of solutions $v(y)$ of the frozen, flattened equation [coefficients fixed at one boundary point, boundary taken flat, $y \geq 0$ the normal coordinate] that *decay* as $y \to +\infty$. A boundary condition B imposing k conditions is good if, for every $\xi \neq 0$, B maps $M^+(\xi)$ *isomorphically* onto $\mathbb{C}^k$: it kills no decaying mode (uniqueness) and misses no data (existence). *Formally*: this is the Lopatinskii–Shapiro condition.

Worked example 4.128: The test on the Laplacian and on $\bar\partial$

Laplacian. $\Delta = \partial_x^2 + \partial_y^2$ and $u = e^{i\xi x}v(y)$ give $v'' = \xi^2 v$, so $v = e^{-|\xi|y}$ (decays) or $e^{+|\xi|y}$ (grows). Thus $\dim M^+(\xi) = 1$ for every $\xi \neq 0$, and we need exactly $k = 1$ condition. Dirichlet: $B = 1$ sends $e^{-|\xi|y} \mapsto v(0) = 1 \neq 0$. Good. Neumann: $B = \partial_y$ sends it to $-|\xi| \neq 0$. Good.

Cauchy–Riemann. $\bar\partial = \frac{1}{2}(\partial_x + i\partial_y)$ and $u = e^{i\xi x}v(y)$ give

$$0 = \bar\partial u \overset{(1)}{=} \tfrac{1}{2}\big(i\xi v + iv'\big)e^{i\xi x} \quad\Longrightarrow\quad v' \overset{(2)}{=} -\xi v \quad\Longrightarrow\quad v \overset{(3)}{=} e^{-\xi y}.$$

(1) ∂_x acts on $e^{i\xi x}$ as $i\xi$; ∂_y acts on v.
(2) Divide by $\frac{i}{2}e^{i\xi x}$.
(3) Solve the first-order ODE.

So $e^{-\xi y}$ decays for $\xi > 0$ and grows for $\xi < 0$:

$$\dim M^+(\xi) = \begin{cases} 1, & \xi > 0, \\ 0, & \xi < 0. \end{cases}$$

A local condition imposes the same number k of conditions at every ξ (it is a fixed matrix, or a polynomial in ξ, acting on boundary values). We would need $k = 1$ for $\xi > 0$ and $k = 0$ for $\xi < 0$. *No local boundary condition exists for $\bar\partial$.*

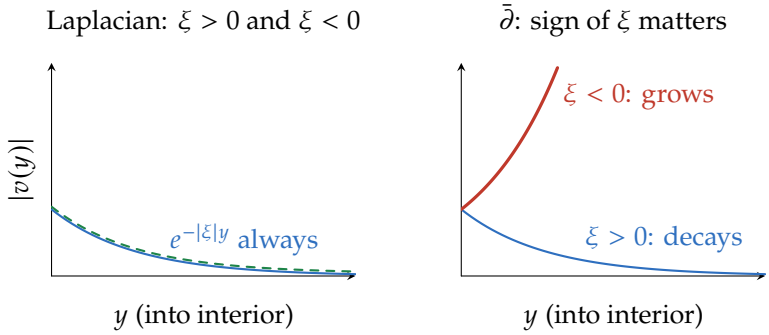


Figure 4.71: The decaying-mode test. Left: for the Laplacian, every boundary frequency ξ has exactly one mode decaying into the interior, so one local condition (Dirichlet or Neumann) is right for all ξ. Right: for $\bar\partial$ only the positive frequencies have a decaying mode. The number of conditions needed jumps from 1 to 0 as ξ changes sign, and no local condition can do that. A condition that depends on the *sign of the boundary frequency* can: this is the APS projection.

This is the simplest instance of what Nash calls the *topological obstruction* to local boundary conditions [Nash p. 127], found by Atiyah and Bott. In general the decaying modes $M^+(x,\xi)$ form a vector bundle over the set of unit boundary covectors (x,ξ) (the cosphere bundle of ∂X), and a local boundary condition exists only if this bundle is, stably, pulled back from ∂X itself, i.e. does not twist as ξ turns around. For $\bar\partial$ on a surface the cosphere bundle of a boundary circle is two circles, $\xi = +1$ and $\xi = -1$, and the bundle has rank 1 on one and rank 0 on the other: the crudest possible failure. For the signature operator d_+ and the chiral Dirac operator in any even dimension the failure is subtler (the ranks agree, but the bundle twists), and the conclusion is the same.

Watch out: The full Dirac operator does have local conditions

Bag-model physicists impose the local MIT condition $i\gamma^n\psi = \psi$ on the full Dirac operator $\not{D}$ on $S^+ \oplus S^-$. That is fine: the full $\not{D}$ is self-adjoint and has index zero. The obstruction concerns the *chiral* half $\not{D}^+ : S^+ \to S^-$, the operator whose index counts chiral zero modes. The MIT condition mixes S^+ with S^- and so does not define a boundary problem for $\not{D}^+$ alone.

Where the symbol lives: relative K-theory

On a closed manifold the symbol of an elliptic operator defines a class in $K(T^*X)$ (Chapter 4, first part). Nash notes [Nash p. 127] that with a boundary the natural home is a **relative K-group** $K(X, \partial X)$. Operationally: an element of $K(X, \partial X)$ is a pair of bundles on X *together with an isomorphism between them over* ∂X. Collapsing ∂X to a point, such a pair becomes a pair of bundles on the quotient $X/\partial X$, so $K(X, \partial X) \cong \tilde{K}(X/\partial X)$. For example $D^2/S^1 = S^2$, so $K(D^2, S^1) \cong \tilde{K}(S^2) \cong \mathbb{Z}$, generated by the monopole bundle. In the boundary problem the interior symbol gives the pair of bundles, and the boundary condition supplies the isomorphism over the boundary. For $\bar{\partial}$ there is no *local* way to supply it (the rank of M^+ jumps), which is the K-theoretic form of Example 4.128.

4.21.2 The signature operator near the boundary

Nash now takes the signature operator $d_+ : \Gamma(\Lambda^+ T^*X \otimes \mathbb{C}) \to \Gamma(\Lambda^- T^*X \otimes \mathbb{C})$ on a manifold with boundary and asks for boundary conditions under which its index is the signature [Nash p. 128]. We follow him, but first set up the geometry that every such problem shares.

A collar and the normal form $D = A(\partial_n + L)$

Near the boundary choose coordinates (n, y) with y coordinates on ∂X and $n \geq 0$ the distance from the boundary measured *inwards*. The region $0 \leq n < \varepsilon$ is a **collar** $\partial X \times [0, \varepsilon)$. Assume the metric is a **product** there:

$$\mathrm{d}s^2 = \mathrm{d}n^2 + h_{ij}(y)\,\mathrm{d}y^i \mathrm{d}y^j, \tag{4.201}$$

with h independent of n. (We relax this at the end, in Section 4.21.11.) Any first-order operator splits into a normal part and a tangential part, Nash's (4.143):

$$D = A\frac{\partial}{\partial n} + B \cdot D^T, \tag{4.202}$$

where A, B are matrices and D^T differentiates only along ∂X. Ellipticity says the symbol $\sigma(n, p) = i(Ap_n + \dots)$ is invertible for $p = (p_n, 0) \neq 0$, so A is invertible. Factor it out [Nash (4.144)]:

$$D \overset{(1)}{=} A\Big(\frac{\partial}{\partial n} + A^{-1}B \cdot D^T\Big) \overset{(2)}{=} A\Big(\frac{\partial}{\partial n} + L\Big), \qquad L := A^{-1}B \cdot D^T. \tag{4.203}$$

(1) A is invertible, so $A\,A^{-1} = 1$ can be inserted.
(2) Definition of the **boundary operator** L.

For Dirac-type operators (Dirac, signature, $\bar{\partial}$) with a product metric, A is Clifford multiplication by the unit normal, L is independent of n, and L is a *self-adjoint elliptic operator on the closed manifold* ∂X. It is essentially "the same kind of operator, one dimension down".

Worked example 4.129: The boundary operator of the 2D chiral Dirac operator

On the flat half-cylinder $\{(n,\theta) : n \geq 0,\ \theta \in S^1\}$, with complex coordinate $w = n + i\theta$, the Cauchy–Riemann operator is

$$\bar\partial_w = \tfrac{1}{2}(\partial_n + i\partial_\theta) = \tfrac{1}{2}\big(\partial_n + L\big), \qquad L = i\partial_\theta.$$

So $A = \frac{1}{2}$ and $L = i\,\mathrm{d}/\mathrm{d}\theta$. Check self-adjointness on $L^2(S^1)$: integrating by parts, $\int \bar f\,(ig')\,\mathrm{d}\theta = -i\int \bar f' g\,\mathrm{d}\theta = \int \overline{(if')}\,g\,\mathrm{d}\theta$, using that there is no boundary term on the closed circle. Its eigenfunctions are $e^{-ik\theta}$ with eigenvalues $k \in \mathbb{Z}$: $i\partial_\theta e^{-ik\theta} = k\,e^{-ik\theta}$. Up to the relabelling $k \to -k$ this is the Dirac operator $-i\,\mathrm{d}/\mathrm{d}\theta$ of the circle. The chiral Dirac operator in the Weyl basis (4.167) of the first section of this part is a multiple of $\bar\partial$ on one chirality, so it has the same boundary operator.

For the signature operator, Nash's d_+^T is the tangential part of $d + d^*$. On a $4k$-manifold, restricting a form in Ω^+ to the collar and keeping its components without $\mathrm{d}n$ identifies $\Omega^+|_{\text{collar}}$ with all forms on ∂X, and L becomes the operator $\phi \mapsto \pm(\star\mathrm{d} - \mathrm{d}\star)\phi$ on $\Omega^*(\partial X)$ (with degree-dependent signs; Atiyah–Patodi–Singer, Part I, §4). For a 3-manifold boundary its most important piece is the *curl operator* $\star\mathrm{d}$ on 1-forms: in $\mathbb{R}^3$, $\star\mathrm{d}A = \nabla \times \vec{A}$.

Separation of variables in the collar

The normal form makes the local analysis explicit. Let $L\phi_\lambda = \lambda\phi_\lambda$ be the eigen-decomposition of L on ∂X (discrete real spectrum, because L is self-adjoint elliptic on a closed manifold). Expand a solution of $Du = 0$ in the collar:

$$u = \sum_\lambda f_\lambda(n)\,\phi_\lambda(y), \qquad 0 = Du \overset{(1)}{=} A\sum_\lambda \big(f_\lambda' + \lambda f_\lambda\big)\phi_\lambda \quad\Longrightarrow\quad f_\lambda(n) \overset{(2)}{=} f_\lambda(0)\,e^{-\lambda n}.$$

(1) ∂_n acts on the coefficients; L acts on the eigenfunctions by λ.
(2) A is invertible and the ϕ_λ are independent, so each coefficient solves $f' = -\lambda f$.

A mode with $\lambda > 0$ decays into the interior; one with $\lambda < 0$ grows into the interior. This is the decaying-mode test again, now globally on ∂X: the boundary frequencies of Example 4.128 have become the eigenvalues of L.

The APS boundary condition

Split the boundary data by the sign of λ. Nash [Nash p. 128] writes

$$H_+ = \overline{\operatorname{span}}\{\phi_\lambda : \lambda \geq 0\}, \qquad H_- = \overline{\operatorname{span}}\{\phi_\lambda : \lambda < 0\}, \qquad L^2(\partial X) = H_+ \oplus H_-.$$

Working definition 4.60 (APS (spectral) boundary condition)

Let Π be the orthogonal projection [keeps the H_+ components and sets the H_- components to zero] of $L^2(\partial X)$ onto H_+, the span of the eigenvectors of L [the boundary operator in $D = A(\partial_n + L)$ near ∂X] with $\lambda \geq 0$. The **Atiyah–Patodi–Singer boundary condition** on u is

$$\Pi\big(u|_{\partial X}\big) = 0, \qquad \text{equivalently} \qquad u|_{\partial X} \in H_-, \tag{4.204}$$

Nash's (4.145)–(4.146). *To impose it*: expand the boundary value of u in eigenfunctions of L and demand that every coefficient with $\lambda \geq 0$ vanish. The condition on the adjoint problem [the equation $D^*v = 0$] is the complementary one, $(1 - \Pi)(v|_{\partial X}) = 0$.

Physical picture: Gluing on an infinite cylinder

Here is why this is the natural condition. Glue a half-infinite cylinder $(-\infty, 0] \times \partial X$ onto X along the boundary, and extend the product metric and the operator $A(\partial_n + L)$ to it (Figure 4.72). A solution in the collar continues onto the cylinder by the same formula $f_\lambda(0)e^{-\lambda n}$, now with $n \to -\infty$. It is square-integrable there if and only if every λ that appears is *negative*. So

u satisfies the APS condition $\iff$ *u extends to an L^2 solution on the elongated manifold $\hat{X} = X \cup$ cylinder,*

except for $\lambda = 0$ modes, which extend as bounded but non-L^2 constants along the cylinder ("extended L^2 solutions"; they are the origin of the h in the index formula). The APS index is thus the honest L^2 index of an operator on a complete manifold with a cylindrical end: a very physical set-up (an infinitely long tube, with no reflection at a wall).

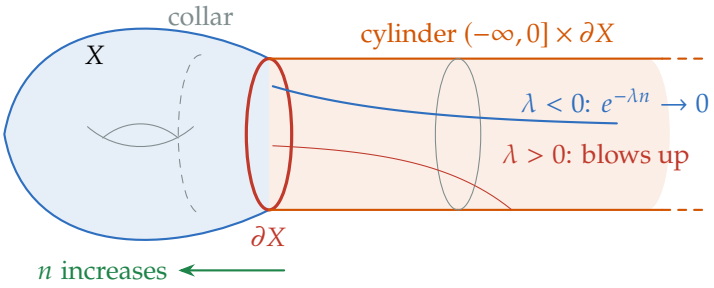


Figure 4.72: The elongated manifold $\hat{X}$. The collar of X (blue) is continued by a half-infinite cylinder (orange) on which the operator is $A(\partial_n + L)$ with $n \to -\infty$ along the tube. Boundary data in an eigenmode of L with $\lambda < 0$ continue as decaying tails; those with $\lambda > 0$ blow up. The APS condition $\Pi(u|_{\partial X}) = 0$ keeps exactly the modes that live happily on the infinite tube.

Worked example 4.130: APS on the half-cylinder: kernel and cokernel

Take the half-infinite cylinder $X = [0, \infty) \times S^1$ itself (now n runs *into* X as $n \to +\infty$) with $D = \partial_n + L$, $L = -i\partial_\theta$, eigenvalues $k \in \mathbb{Z}$ on $e^{ik\theta}$. Solutions of $Du = 0$ are $\sum_k c_k e^{-kn} e^{ik\theta}$. Square-integrability as $n \to +\infty$ needs $c_k = 0$ for $k < 0$ (these blow up) and for $k = 0$ (constant, not L^2 on an infinite tube). The APS condition independently removes $k \geq 0$. So $\operatorname{Ker} D = 0$. The adjoint $D^* = -\partial_n + L$ has solutions $\sum_k d_k e^{kn} e^{ik\theta}$; L^2 needs $k < 0$, and the complementary condition $(1 - \Pi)v| = 0$ removes $k < 0$. So $\operatorname{Coker} D = 0$ and the index is 0. The bulk contributes nothing here (flat metric, trivial bundle), so all the interest of the index formula must come from the spectrum of L, and here $\{k\}$ is symmetric about 0. We will see in Section 4.21.6 that a flat twist makes it asymmetric.

The projection Π is a pseudo-differential operator

Nash shows that Π is not exotic: it is a *zeroth-order pseudo-differential operator* [Nash p. 128–129, (4.147)]. The recipe:

1. Make L invertible: $M := L + N$ where N is the orthogonal projection onto $\operatorname{Ker} L$. On ϕ_λ with $\lambda \neq 0$, M acts as λ; on $\operatorname{Ker} L$ it acts as 1.
2. Take the positive square root $|M| = \sqrt{M^2}$ (Seeley's complex powers of an elliptic operator, Chapter 2).

3. Set
$$\Pi = \tfrac{1}{2}M^{-1}\big(M + |M|\big). \tag{4.205}$$

Check it on the eigenvectors:
$$\Pi\phi_\lambda \overset{(1)}{=} \frac{\lambda + |\lambda|}{2\lambda}\phi_\lambda \overset{(2)}{=} \begin{cases}\phi_\lambda, & \lambda > 0,\\ 0, & \lambda < 0,\end{cases} \qquad \Pi\phi_0 \overset{(3)}{=} \frac{1+1}{2}\phi_0 = \phi_0.$$

(1) For $\lambda \neq 0$, M, $|M|$ and M^{-1} act on ϕ_λ as λ, $|\lambda|$ and $1/\lambda$.
(2) $\lambda + |\lambda|$ is 2λ or 0.
(3) On $\operatorname{Ker} L$, $M = |M| = 1$.

So Π is exactly the projection onto H_+ (including $\operatorname{Ker} L$). The operators M and $|M|$ are first order, M^{-1} is of order -1, so Π has order 0. Its leading symbol is
$$\sigma_0(\Pi)(y,\xi) = \tfrac{1}{2}\Big(1 + \frac{\sigma(L)(y,\xi)}{|\sigma(L)(y,\xi)|}\Big),$$
the projection onto the positive eigenspace of the *symbol* of L at (y,ξ). This is homogeneous of degree zero in ξ but not polynomial: it is not the symbol of any differential operator.

Worked example 4.131: Π on the circle is the Hardy projection, and is non-local

Take $L = -i\,\mathrm{d}/\mathrm{d}\theta$ on S^1, $Le^{ik\theta} = ke^{ik\theta}$. Then Π keeps $k \geq 0$: $\Pi f = \sum_{k\geq 0}\hat f_k e^{ik\theta}$ with $\hat f_k = \frac{1}{2\pi}\int f(\phi)e^{-ik\phi}\mathrm{d}\phi$. Write it as an integral kernel:
$$\begin{aligned}(\Pi f)(\theta) &\overset{(1)}{=} \frac{1}{2\pi}\int_0^{2\pi}\Big(\sum_{k\geq 0} e^{ik(\theta-\phi)}\Big)f(\phi)\,\mathrm{d}\phi\\ &\overset{(2)}{=} \frac{1}{2\pi}\int_0^{2\pi}\Big(\pi\delta(\theta-\phi) + \tfrac{1}{2} + \tfrac{i}{2}\cot\tfrac{\theta-\phi}{2}\Big)f(\phi)\,\mathrm{d}\phi\\ &\overset{(3)}{=} \tfrac{1}{2}f(\theta) + \tfrac{1}{2}\hat f_0 + \frac{i}{4\pi}\,\mathrm{p.v.}\int_0^{2\pi}\cot\tfrac{\theta-\phi}{2}\,f(\phi)\,\mathrm{d}\phi.\end{aligned}$$

(1) Insert the Fourier coefficients and exchange sum and integral.
(2) Split $\sum_{k\geq 0} = \frac{1}{2}\big(\sum_{k\in\mathbb{Z}}+1\big) + \frac{1}{2}\sum_{k\neq 0}\operatorname{sgn}(k)e^{ikx}$, and use $\sum_{k\in\mathbb{Z}} e^{ikx} = 2\pi\delta(x)$ and $\sum_{k\neq 0}\operatorname{sgn}(k)e^{ikx} = 2i\sum_{k>0}\sin kx = i\cot\frac{x}{2}$ (Abel summation: $\operatorname{Im}\frac{e^{ix}}{1-e^{ix}} = \frac{1}{2}\cot\frac{x}{2}$).
(3) Do the δ and constant integrals; the cot is integrated as a principal value.

The last term is the **Hilbert transform** (the conjugate-function operator of Fourier analysis, the Kramers–Kronig transform of optics). Its kernel $\cot\frac{\theta-\phi}{2} \approx \frac{2}{\theta-\phi}$ has a long tail: $(\Pi f)(\theta)$ depends on f at *every* point of the circle. The APS condition $\Pi(u|_{S^1}) = 0$ is therefore genuinely non-local; it says "the boundary value of u has only negative frequencies", a statement about the whole function, not about its values point by point. Its symbol is $\sigma_0(\Pi)(\theta,\xi) = \frac{1}{2}(1+\operatorname{sgn}\xi)$: 1 for $\xi > 0$, 0 for $\xi < 0$. Compare Figure 4.71: that is exactly the rank-jumping behaviour a local condition could not provide.

Watch out: Why a non-local condition does not spoil the index

One might worry that a non-local condition breaks the analysis. It does not: Π is a pseudo-differential operator of order zero, and the Fredholm theory of elliptic boundary problems works just as well with pseudo-differential boundary conditions. What matters

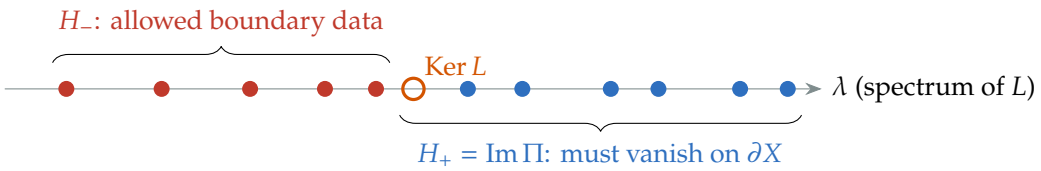


Figure 4.73: The APS condition splits the spectrum of the boundary operator L at zero. The zero modes are put with H_+ (Nash's convention and APS's). Boundary data may contain only the red (negative) modes. The index correction will measure how *unevenly* the dots are spread on the two sides: the eta invariant.

for the index is only the *homotopy class* of the symbol data. If we could deform $\sigma_0(\Pi)(y,\xi)$ through projections into one independent of ξ (the symbol of a multiplication operator, which is local), we would get an equivalent local problem. For d_+ no such homotopy exists, precisely because of the Atiyah–Bott obstruction [Nash p. 129].

4.21.3 A warm-up where the boundary term is local: Gauss–Bonnet

Before computing any index with a boundary, Nash looks at a case where the answer is already known [Nash p. 129, (4.148)]: the Euler characteristic of a surface with boundary. For a compact oriented surface X with Gaussian curvature K (Nash writes R) and boundary geodesic curvature k_g (Nash's R_g),

$$\frac{1}{2\pi}\int_X K\,\mathrm{d}A + \frac{1}{2\pi}\oint_{\partial X} k_g\,\mathrm{d}s = \chi(X). \tag{4.206}$$

The geodesic curvature k_g of a boundary curve is how much it bends *within* the surface (its acceleration projected onto the surface, positive when it bends towards the interior). A geodesic has $k_g = 0$.

Worked example 4.132: Checking (4.206) on discs and caps

Flat disc of radius a. $K = 0$; the boundary circle bends towards the centre with curvature $1/a$, and has length $2\pi a$:

$$0 + \frac{1}{2\pi}\cdot\frac{1}{a}\cdot 2\pi a = 1 = \chi(D^2).$$

Spherical cap. On a sphere of radius a take the cap $0 \le \vartheta \le \vartheta_0$ (polar angle). Then $K = 1/a^2$ and $\mathrm{d}A = a^2\sin\vartheta\,\mathrm{d}\vartheta\,\mathrm{d}\varphi$. The boundary is the latitude circle, of radius $a\sin\vartheta_0$ in space; its curvature vector $1/(a\sin\vartheta_0)$ points to the axis, and its component tangent to the sphere is $\cos\vartheta_0/(a\sin\vartheta_0)$. So:

$$\frac{1}{2\pi}\int_X K\,\mathrm{d}A + \frac{1}{2\pi}\oint k_g\,\mathrm{d}s \overset{(1)}{=} \frac{1}{2\pi}\int_0^{2\pi}\int_0^{\vartheta_0}\sin\vartheta\,\mathrm{d}\vartheta\,\mathrm{d}\varphi + \frac{1}{2\pi}\cdot\frac{\cot\vartheta_0}{a}\cdot 2\pi a\sin\vartheta_0$$
$$\overset{(2)}{=} (1-\cos\vartheta_0) + \cos\vartheta_0 \overset{(3)}{=} 1.$$

(1) $K\,\mathrm{d}A = \sin\vartheta\,\mathrm{d}\vartheta\,\mathrm{d}\varphi$; $k_g = \cot\vartheta_0/a$; the latitude has length $2\pi a\sin\vartheta_0$.
(2) Do the integrals.
(3) The two pieces trade off as ϑ_0 varies; only the sum is topological.

As $\vartheta_0 \to 0$ we recover the flat disc (everything in the boundary term); at $\vartheta_0 = \pi/2$ (hemisphere) the boundary is a geodesic and everything is in the bulk.

Flat annulus $r_1 \le r \le r_2$. $K = 0$. The outer circle bends towards the interior ($k_g = +1/r_2$); the inner circle bends away from it ($k_g = -1/r_1$). Total $\frac{1}{2\pi}(2\pi r_2/r_2 - 2\pi r_1/r_1) = 0 = \chi(\text{annulus})$.

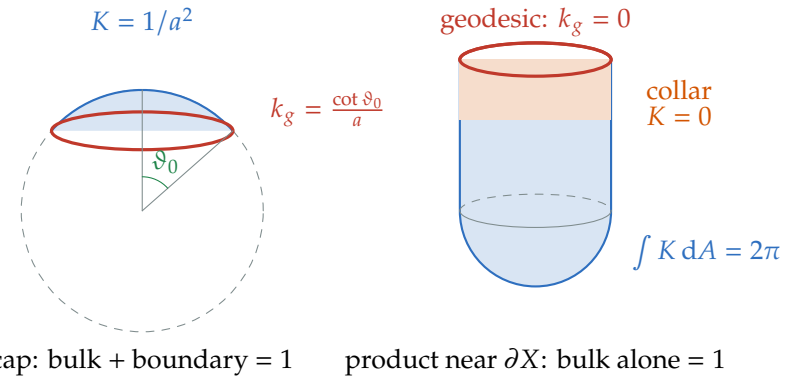


Figure 4.74: Gauss–Bonnet with boundary. Left: on a spherical cap the Euler number 1 is shared between the bulk curvature and the bending of the rim. Right: a "test tube" has a cylindrical collar, i.e. a product metric near the boundary; the rim is a geodesic, and the bulk integral alone gives $\chi = 1$. This is Nash's (4.149): for Euler characteristic, a product metric makes the boundary term disappear.

Why is the boundary term there, and why does it vanish for a product? In two dimensions $K\,\mathrm{d}A = \mathrm{d}\omega$ locally, where ω is the Levi-Civita connection 1-form of an orthonormal frame (Chapter T2). Stokes turns the bulk integral into $\oint \omega$ on the boundary, and $\oint \omega$ measures how the frame rotates relative to parallel transport along the rim, i.e. $2\pi\chi$ minus the turning of the rim's own tangent, which is $\oint k_g\,\mathrm{d}s$. For a product collar $\mathrm{d}n^2 + h$ the rim $n = 0$ is a geodesic ($k_g = 0$), so [Nash (4.149)]

$$\frac{1}{2\pi}\int_X K\,\mathrm{d}A = \chi(X) \qquad \text{if } X \text{ is a product near } \partial X. \tag{4.207}$$

In index language: the Euler characteristic is the index of the de Rham complex; with a product metric the bulk formula holds unchanged, with *no* boundary correction. We will see why in Section 4.21.8.

4.21.4 The signature with boundary: where the eta invariant comes from

The signature of a manifold with boundary

For a closed oriented 4-manifold the signature is that of the intersection form $Q(a,b) = \int_X a \wedge b$ on $H^2(X)$: diagonalise it and count (#positive) − (#negative) (Chapter 4, the signature theorem). With a boundary, use closed 2-forms that vanish near ∂X (compactly supported, relative classes $H^2(X, \partial X)$). The form Q may now be degenerate; $\mathrm{Sign}(X)$ is the signature of its non-degenerate part.

Worked example 4.133: Signatures of some manifolds with boundary

- The ball B^4: $H^2(B^4, S^3) = 0$, so Sign = 0.
- $\mathbb{CP}^2$ minus a small ball: take a projective line $\mathbb{CP}^1 \subset \mathbb{CP}^2$ missing the ball. Its Poincaré dual is a closed 2-form supported in a tube around the line, so it vanishes near the boundary. Two lines meet in one point, positively, so $Q = (+1)$ and Sign = 1, as for $\mathbb{CP}^2$ itself.
- The disc bundle of the line bundle of degree e over S^2: the zero section has self-intersection e, so $Q = (e)$ and $\mathrm{Sign} = \mathrm{sgn}\, e$. Its boundary is S^3 for $e = \pm 1$ (the Hopf fibration), $S^2 \times S^1$ for $e = 0$, and the lens space $L(|e|, 1)$ otherwise.

The bulk formula fails, by a boundary number

For a closed 4-manifold $\tilde{X}$, Hirzebruch's theorem is [Nash p. 129, (4.150)]

$$\operatorname{Sign}(\tilde{X}) = \frac{1}{3}\int_{\tilde{X}} p_1(\tilde{X}), \qquad p_1 = -\frac{1}{8\pi^2}\operatorname{tr}(R\wedge R). \tag{4.208}$$

Nash states that if X has a boundary, even with a product metric near it, (4.208) fails, and the failure defines a real number attached to the boundary [Nash p. 130, (4.151)]:

$$\operatorname{Sign}(X) - \frac{1}{3}\int_X p_1(X) = -\frac{1}{2}\,\eta(\partial X). \tag{4.209}$$

The factor $-\frac{1}{2}$ is a normalisation chosen for later convenience.

The defect depends only on the boundary

Nash's key observation [Nash (4.152)] is that the left side of (4.209) depends only on ∂X (with its metric), not on how X fills it in. The argument is a gluing, and it is worth doing carefully.

Let X and Y be two oriented 4-manifolds with the same boundary $\partial X = \partial Y = W$, with the same metric on W and product metrics on both collars. Reverse the orientation of Y (write $-Y$) and glue it to X along W to form a closed manifold $Z = X \cup_W (-Y)$ (Figure 4.75). The product collars fit together smoothly into a collar $W \times (-\varepsilon, \varepsilon)$, so Z has a smooth metric that agrees with the old ones on each piece. Then

$$\begin{aligned} 0 &\overset{(1)}{=} \operatorname{Sign}(Z) - \frac{1}{3}\int_Z p_1 \\ &\overset{(2)}{=} \Big(\operatorname{Sign}(X) + \operatorname{Sign}(-Y)\Big) - \frac{1}{3}\Big(\int_X p_1 + \int_{-Y} p_1\Big) \\ &\overset{(3)}{=} \Big(\operatorname{Sign}(X) - \frac{1}{3}\int_X p_1\Big) - \Big(\operatorname{Sign}(Y) - \frac{1}{3}\int_Y p_1\Big). \end{aligned}$$

(1) Hirzebruch's theorem (4.208) on the closed manifold Z.
(2) The signature is additive under gluing along a common boundary (Novikov additivity), and an integral over Z is the sum of the integrals over the pieces.
(3) Reversing orientation flips the sign of the intersection form, so $\operatorname{Sign}(-Y) = -\operatorname{Sign}(Y)$, and flips the sign of every integral of a top form.

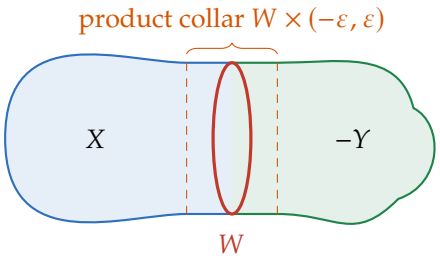
Hence

$$\operatorname{Sign}(X) - \frac{1}{3}\int_X p_1 = \operatorname{Sign}(Y) - \frac{1}{3}\int_Y p_1 \qquad \text{whenever } \partial X = \partial Y, \tag{4.210}$$

which is Nash's (4.152). The defect is a function of W alone: $\eta(W)$ is well defined.

Two consequences follow at once.

- *Orientation reversal.* Replacing X by $-X$ flips both Sign and $\int p_1$, so $\eta(-W) = -\eta(W)$.
- *Symmetric boundaries.* If W has an orientation-reversing isometry (for example the round S^3, via a reflection), then $(W, h) \cong (-W, h)$ and $\eta(W) = -\eta(W) = 0$.



$Z = X \cup_W (-Y)$ closed: $\operatorname{Sign} Z = \frac{1}{3}\int_Z p_1$

Figure 4.75: Why the signature defect depends only on the boundary. Glue X to the orientation reversal of any other filling Y of the same W. Product collars make the glued metric smooth. On the closed Z the defect is zero, and the defects of X and Y enter with opposite signs, so they are equal. The defect is therefore a property of (W, h): the eta invariant.

Worked example 4.134: A consistency check with $\mathbb{CP}^2$

Write $\mathbb{CP}^2 = X \cup_{S^3} B$ with $X = \mathbb{CP}^2 \setminus B$ and B a small ball. Choose a metric that is a product near a round S^3 (deform the Fubini–Study metric in a thin shell). The round S^3 has $\eta = 0$ (it has an orientation-reversing isometry). Then (4.209) applied to each piece, with $\operatorname{Sign}(X) = 1$ and $\operatorname{Sign}(B) = 0$ from Example 4.133, gives $\frac{1}{3}\int_X p_1 = 1$ and $\frac{1}{3}\int_B p_1 = 0$. They add to $\frac{1}{3}\int_{\mathbb{CP}^2} p_1 = 1 = \operatorname{Sign}\mathbb{CP}^2$, as they must. All of the Pontryagin number sits in the piece that carries the topology.

What the number is: the spectral asymmetry of L

Nash stresses [Nash p. 130] that $\eta(\partial X)$ is *not* a local geometric quantity: no integral of curvature over ∂X produces it. Instead it is built out of the spectrum of the boundary operator L of (4.203). Let $\{\lambda_n\}$ be the eigenvalues of L and define, for $\operatorname{Re} s$ large, [Nash (4.153)]

$$\eta_L(s) = \sum_{\lambda_n \neq 0} \frac{\operatorname{sgn}(\lambda_n)}{|\lambda_n|^s}. \tag{4.211}$$

Working definition 4.61 (eta invariant)

For a self-adjoint elliptic operator L on a closed manifold [compact and without boundary, here ∂X], the **eta function** $\eta_L(s)$ is (4.211). It converges for $\operatorname{Re} s > \dim(\partial X)$, continues to a meromorphic function [analytic except at isolated poles] of s, and is finite at $s = 0$. The **eta invariant** is $\eta_L(0)$. *To compute it*: list the eigenvalues, sum their signs with a regulator $|\lambda|^{-s}$, continue to $s = 0$. Informally,

$$\eta_L(0) = \text{“}\#\{\lambda > 0\} - \#\{\lambda < 0\}\text{”},$$

the regularised excess of positive over negative eigenvalues: the **spectral asymmetry**.

If the spectrum is symmetric (λ and $-\lambda$ with equal multiplicities), every term cancels and $\eta(s) \equiv 0$. So η is a direct measure of the failure of $\lambda \to -\lambda$ symmetry.

The connection with (4.209) is [Nash (4.154)–(4.155)]

$$\eta(\partial X) = \eta_L(0), \qquad \operatorname{Sign}(X) = \frac{1}{3}\int_X p_1(X) - \frac{1}{2}\,\eta_L(0). \tag{4.212}$$

Why the spectrum of L should appear. Near the boundary, everything about d_+ is encoded in L (Section 4.21.2). Under $\lambda \to -\lambda$ the roles of H_+ and H_- in the boundary condition swap, which is what reversing the orientation does. A boundary quantity that is odd under orientation reversal, additive over disjoint boundaries, and built from the spectrum of L has to be an odd function of the spectrum; η is the simplest. The actual proof (heat kernels on the elongated manifold) we sketch in Section 4.21.9.

Watch out: Normalisations of η, and the harmonic forms h

Nash's L is the boundary operator of d_+ acting on *all* forms of ∂X. APS work instead with its restriction B^{ev} to even forms; the spectrum of L is two copies of that of B^{ev}, so $\eta_L = 2\eta_{B^{\mathrm{ev}}}$, and (4.212) reads, in APS's form, $\mathrm{Sign}(X) = \int_X L(p) - \eta_{B^{\mathrm{ev}}}(0)$. Also, Nash's remark [Nash p. 131] that "Sign(X) is the index of d_+ subject to the global boundary conditions" needs a correction: L always has a kernel (the harmonic forms on ∂X, including the constants), and APS find $\mathrm{index}_{\mathrm{APS}}\, d_+ = \mathrm{Sign}(X) - h$, with h the dimension of the space of harmonic forms of even degree on ∂X. The index formula of Section 4.21.7 accounts for h explicitly.

Reading Nash (pp. 127–131, (4.141)–(4.155))

Nash introduces boundary problems through the local Dirichlet and Neumann problems (4.141)–(4.142), mentions the Atiyah–Bott obstruction and $K(X, \partial X)$, writes the signature operator near the boundary as $A(\partial/\partial n + L)$ (4.143)–(4.144), imposes $u|_{\partial X} \in H_-$ (4.145)–(4.146), and exhibits $\Pi = \frac{1}{2}M^{-1}(M + |M|)$ (4.147). He then recalls Gauss–Bonnet with boundary (4.148)–(4.149), and uses the failure of Hirzebruch's formula (4.150) on a manifold with boundary to define $\eta(\partial X)$ (4.151), shows it depends only on ∂X (4.152), and identifies it with $\eta_L(0)$ (4.153)–(4.155). We added the Fourier solution of the Dirichlet and Neumann problems, the decaying-mode test explaining *why* $\bar\partial$ has no local condition, the elongated manifold picture, the Hardy-projection form of Π on the circle, the cap and annulus checks of Gauss–Bonnet, the gluing computation behind (4.152), and the normalisation caveats.

Recap

- With a boundary, an elliptic operator has an infinite kernel until we impose boundary conditions. Dirichlet and Neumann are local and fine for Δ.
- Decaying-mode test: for $\bar\partial$ (and chiral Dirac, and d_+) the number of boundary-localised modes depends on the sign of the boundary frequency, so no local condition works.
- Near a product boundary $D = A(\partial_n + L)$ with L self-adjoint elliptic on ∂X. APS condition: boundary data in H_- (negative spectrum of L); equivalently, L^2 extension to an attached infinite cylinder.
- $\Pi = \frac{1}{2}M^{-1}(M+|M|)$ is a zeroth-order pseudo-differential, non-local projection (Hardy projection on S^1).
- Gauss–Bonnet with boundary needs $\oint k_g$; for a product collar it vanishes. The signature needs a genuinely spectral correction: $\mathrm{Sign}\, X = \frac{1}{3}\int p_1 - \frac{1}{2}\eta_L(0)$, and the defect depends only on ∂X (gluing).

4.21.5 Giving a divergent sum a value: zeta functions

The eta function $\eta_L(s) = \sum \operatorname{sgn}(\lambda_n)|\lambda_n|^{-s}$ of (4.211) is a convergent series only when $\operatorname{Re} s$ is large. The number we want is its value at $s = 0$, where the series is the meaningless "$\sum \pm 1$". So we must say precisely how a function defined by a series far to the right in the complex s-plane is continued to $s = 0$. Nash does this by reducing η to zeta functions [Nash p. 131, (4.156)–(4.158)]. The zeta function trick is also the standard way physicists regularise determinants and Casimir energies, so it is worth learning properly.

The zeta function of a positive operator

Working definition 4.62 (zeta function of an operator, Nash (4.156))

Let P be a positive [all eigenvalues ≥ 0] self-adjoint elliptic operator of order $m > 0$ [m derivatives, so that μ_k grows like $k^{m/n}$] on a closed manifold of dimension n, with eigenvalues $0 < \mu_1 \leq \mu_2 \leq \cdots$ (zero eigenvalues, if any, are left out). Its **zeta function** is

$$\zeta_P(s) = \sum_k \frac{1}{\mu_k^s} = \operatorname{Tr} P^{-s}, \tag{4.213}$$

which converges for $\operatorname{Re} s > n/m$. *To compute it*: list the eigenvalues, raise each to the power $-s$, sum where the sum converges, and then continue analytically. Seeley's theorem: ζ_P continues to a meromorphic function [analytic except at isolated poles] on all of $\mathbb{C}$ with only simple poles [near each pole s_0 it behaves like $c/(s - s_0)$], and it is *regular at* $s = 0$.

The convergence bound comes from Weyl's law: the number of eigenvalues below Λ grows like $\Lambda^{n/m}$, so $\mu_k \sim k^{m/n}$ and $\sum \mu_k^{-s} \sim \sum k^{-sm/n}$ converges exactly when $sm/n > 1$.

Worked example 4.135: The Laplacian on the circle gives the Riemann zeta function

Take $P = -\mathrm{d}^2/\mathrm{d}\theta^2$ on S^1. Its eigenfunctions are $e^{ik\theta}$, $k \in \mathbb{Z}$, with eigenvalue k^2. Leaving out $k = 0$,

$$\zeta_P(s) \overset{(1)}{=} \sum_{k \neq 0} \frac{1}{(k^2)^s} \overset{(2)}{=} 2 \sum_{k=1}^{\infty} \frac{1}{k^{2s}} \overset{(3)}{=} 2\,\zeta_R(2s).$$

(1) Definition (4.213) with $\mu = k^2$.
(2) k and $-k$ give the same eigenvalue.
(3) $\zeta_R(s) = \sum_{k \geq 1} k^{-s}$ is the Riemann zeta function.

Here $n = 1$, $m = 2$, so convergence needs $\operatorname{Re} s > \frac{1}{2}$, as it should for $\sum k^{-2s}$. The continued value is $\zeta_P(0) = 2\zeta_R(0) = 2 \cdot (-\frac{1}{2}) = -1$. We will derive $\zeta_R(0) = -\frac{1}{2}$ below from the heat kernel, without quoting it.

How the continuation works: the Mellin transform and the heat trace

The engine behind every zeta-function continuation is one line of calculus, the **Mellin transform** of an exponential. For $\mu > 0$ and $\operatorname{Re} s > 0$,

$$\int_0^\infty t^{s-1} e^{-t\mu}\,\mathrm{d}t \overset{(1)}{=} \int_0^\infty \left(\frac{x}{\mu}\right)^{s-1} e^{-x}\,\frac{\mathrm{d}x}{\mu} \overset{(2)}{=} \mu^{-s}\,\Gamma(s). \tag{4.214}$$

(1) Substitute $x = t\mu$, so $t = x/\mu$ and $\mathrm{d}t = \mathrm{d}x/\mu$.

(2) Pull out μ^{-s}; what is left is Euler's integral $\Gamma(s)=\int_0^\infty x^{s-1}e^{-x}\mathrm{d}x$.

Sum (4.214) over the eigenvalues of P:

$$\zeta_P(s)=\frac{1}{\Gamma(s)}\int_0^\infty t^{s-1}\,\operatorname{Tr}'\!\left(e^{-tP}\right)\mathrm{d}t, \qquad \operatorname{Tr}' e^{-tP}:=\sum_{\mu_k>0} e^{-t\mu_k}=\operatorname{Tr}e^{-tP}-\dim\operatorname{Ker}P. \tag{4.215}$$

The heat trace $\operatorname{Tr}e^{-tP}$ is the object we met in the heat-kernel proof of the index theorem (Chapter 4, McKean–Singer). Its two limits are what we need:

- *Large t*: $\operatorname{Tr}' e^{-tP}\le Ce^{-t\mu_1}$ decays exponentially, so $\int_1^\infty$ is an entire function of s.
- *Small t*: there is an asymptotic expansion in powers of t whose coefficients are integrals of local curvature invariants,

$$\operatorname{Tr}e^{-tP}\sim\sum_{j\ge 0}a_j\,t^{(j-n)/m}\qquad(t\to 0). \tag{4.216}$$

Now split $\int_0^\infty=\int_0^1+\int_1^\infty$ and put the expansion into the first piece:

$$\begin{aligned}\Gamma(s)\zeta_P(s)&\overset{(1)}{=}\int_0^1 t^{s-1}\Big(\sum_{j\le J}a_j t^{(j-n)/m}-\dim\operatorname{Ker}P\Big)\mathrm{d}t+\big(\text{holomorphic for }\operatorname{Re}s>\tfrac{n-J-1}{m}\big)\\&\overset{(2)}{=}\sum_{j\le J}\frac{a_j}{s+(j-n)/m}-\frac{\dim\operatorname{Ker}P}{s}+\big(\text{holomorphic}\big).\end{aligned}$$

(1) Insert (4.216) truncated at order J; the remainder is $O(t^{(J+1-n)/m})$ and its integral converges further to the left, as does $\int_1^\infty$.

(2) $\int_0^1 t^{s-1+c}\mathrm{d}t=1/(s+c)$.

So $\Gamma(s)\zeta_P(s)$ is meromorphic with simple poles at $s=(n-j)/m$. Near $s=0$ we have $\Gamma(s)=\frac{1}{s}-\gamma+O(s)$, so $1/\Gamma(s)=s+O(s^2)$ has a *zero* at $s=0$ that cancels the pole there. Therefore

$$\zeta_P(0)=a_n-\dim\operatorname{Ker}P, \tag{4.217}$$

where a_n is the coefficient of t^0 in the heat expansion. This is the reason $\zeta_P(0)$ is always finite, as Nash asserts: the would-be pole is killed by $1/\Gamma$. Informally, $\zeta_P(0)$ is "the regularised number of nonzero eigenvalues".

Worked example 4.136: $\zeta_R(0)=-\frac{1}{2}$ from the heat kernel on the circle

For $P=-\mathrm{d}^2/\mathrm{d}\theta^2$ on S^1 the heat trace is the Jacobi theta function, and Poisson summation (Section 4.21.12 does this in detail) gives

$$\operatorname{Tr}e^{-tP}=\sum_{k\in\mathbb{Z}}e^{-tk^2}=\sqrt{\frac{\pi}{t}}\sum_{m\in\mathbb{Z}}e^{-\pi^2m^2/t}=\sqrt{\frac{\pi}{t}}+O\big(t^{-1/2}e^{-\pi^2/t}\big).$$

So in (4.216) (with $n=1$, $m=2$) only $a_0=\sqrt{\pi}$ is nonzero; in particular the t^0 coefficient is $a_1=0$. The kernel is the constants, $\dim\operatorname{Ker}P=1$. Hence (4.217) gives $\zeta_P(0)=0-1=-1$, and comparing with Example 4.135, $2\zeta_R(0)=-1$, i.e. $\zeta_R(0)=-\frac{1}{2}$. The "regularised number of all eigenvalues" is $a_1=0$, and removing the one zero mode leaves -1.

Dictionary: zeta regularisation in physics

$\zeta_P(0)$	regularised number of modes (a heat coefficient)
$-\zeta_P'(0) = \ln\det P$	one-loop effective action of a Gaussian field, $\int \mathcal{D}\phi\, e^{-\frac{1}{2}\langle\phi, P\phi\rangle} \propto (\det P)^{-1/2}$
$\frac{1}{2}\zeta_H(-1)$ for H with spectrum ω_k	Casimir (zero-point) energy $\frac{1}{2}\sum\omega_k$; e.g. $\sum_{k\geq 1} k = \zeta_R(-1) = -\frac{1}{12}$ gives the $-\frac{1}{24}$ of string theory (Chapter 9)

Eta from two zeta functions: Nash's trick

The zeta function needs a *positive* operator; L has eigenvalues of both signs. Nash (following APS) builds two positive operators out of L [Nash (4.157)]. Assume for now that L is invertible, and let $|L| = \sqrt{L^2}$:

$$L_+ = \tfrac{3}{2}|L| + \tfrac{1}{2}L, \qquad L_- = \tfrac{3}{2}|L| - \tfrac{1}{2}L. \tag{4.218}$$

Write the positive eigenvalues of L as $\lambda_k^+ > 0$ and the negative ones as $\lambda_k^- < 0$. On an eigenvector with eigenvalue λ, $L_\pm$ acts by $\frac{3}{2}|\lambda| \pm \frac{1}{2}\lambda$:

	$\lambda = \lambda^+ > 0$	$\lambda = \lambda^- < 0$
L_+	$\frac{3}{2}\lambda^+ + \frac{1}{2}\lambda^+ = 2\lambda^+$	$-\frac{3}{2}\lambda^- + \frac{1}{2}\lambda^- = -\lambda^-$
L_-	$\frac{3}{2}\lambda^+ - \frac{1}{2}\lambda^+ = \lambda^+$	$-\frac{3}{2}\lambda^- - \frac{1}{2}\lambda^- = -2\lambda^-$

All four entries are positive, so $\zeta_{L_\pm}$ exist. Now subtract [Nash (4.158)]:

$$\begin{aligned}
\zeta_{L_+}(s) - \zeta_{L_-}(s) &\overset{(1)}{=} \Big\{\sum_{\lambda^+}\frac{1}{(2\lambda^+)^s} + \sum_{\lambda^-}\frac{1}{(-\lambda^-)^s}\Big\} - \Big\{\sum_{\lambda^+}\frac{1}{(\lambda^+)^s} + \sum_{\lambda^-}\frac{1}{(-2\lambda^-)^s}\Big\} \\
&\overset{(2)}{=} (2^{-s}-1)\sum_{\lambda^+}\frac{1}{(\lambda^+)^s} - (2^{-s}-1)\sum_{\lambda^-}\frac{1}{|\lambda^-|^s} \\
&\overset{(3)}{=} (2^{-s}-1)\,\eta_L(s), \\
\Longrightarrow \quad \eta_L(s) &= \frac{\zeta_{L_+}(s) - \zeta_{L_-}(s)}{2^{-s}-1}.
\end{aligned} \tag{4.219}$$

(1) Read the eigenvalues of $L_\pm$ off the table.
(2) Collect the λ^+ sums, $(2\lambda^+)^{-s} - (\lambda^+)^{-s} = (2^{-s}-1)(\lambda^+)^{-s}$, and the λ^- sums, $|\lambda^-|^{-s} - (2|\lambda^-|)^{-s} = -(2^{-s}-1)|\lambda^-|^{-s}$.
(3) The bracket is $\sum \operatorname{sgn}(\lambda)|\lambda|^{-s}$, the definition (4.211).

The right side of (4.219) is a ratio of meromorphic functions, so η_L is meromorphic on $\mathbb{C}$. That is all Nash uses the formula for.

Watch out: The formula does not by itself show that $\eta(0)$ is finite

The denominator $2^{-s} - 1 = -s\ln 2 + O(s^2)$ *vanishes* at $s = 0$ (and at $s = 2\pi i k/\ln 2$). So (4.219) is finite at $s = 0$ only if $\zeta_{L_+}(0) = \zeta_{L_-}(0)$, and then $\eta_L(0) = -\frac{1}{\ln 2}\frac{\mathrm{d}}{\mathrm{d}s}\big(\zeta_{L_+} - \zeta_{L_-}\big)\big|_{s=0}$. The regularity of η at 0 is a separate, deep fact (Section 4.21.12). The zeros of $2^{-s} - 1$ on the imaginary axis away from 0 produce no poles: the heat-kernel form (4.221) below shows that all poles of η lie on the real axis, so the numerator must vanish there too.

Eta as a trace, and its heat-kernel form

Nash remarks [Nash p. 132] that $\eta(s)$ is also a trace. On an eigenvector, $\lambda\,|\lambda|^{-s-1} = \operatorname{sgn}(\lambda)|\lambda| \cdot |\lambda|^{-s-1} = \operatorname{sgn}(\lambda)|\lambda|^{-s}$, so

$$\eta_L(s) = \operatorname{Tr}\big(L\,|L|^{-s-1}\big). \tag{4.220}$$

Applying the Mellin transform (4.214) with $\mu = \lambda^2$ and exponent $\frac{s+1}{2}$ turns this into a heat-kernel integral:

$$\begin{aligned}\eta_L(s) &\overset{(1)}{=} \sum_\lambda \lambda\,(\lambda^2)^{-\frac{s+1}{2}} \overset{(2)}{=} \sum_\lambda \frac{\lambda}{\Gamma(\frac{s+1}{2})}\int_0^\infty t^{\frac{s+1}{2}-1} e^{-t\lambda^2}\,\mathrm{d}t \\ &\overset{(3)}{=} \frac{1}{\Gamma(\frac{s+1}{2})}\int_0^\infty t^{\frac{s-1}{2}}\,\operatorname{Tr}\big(L\,e^{-tL^2}\big)\,\mathrm{d}t.\end{aligned} \tag{4.221}$$

(1) (4.220) with $|\lambda|^{-s-1} = (\lambda^2)^{-(s+1)/2}$.

(2) (4.214) with $\mu = \lambda^2$ and $s \to \frac{s+1}{2}$.

(3) Exchange sum and integral; $\sum \lambda e^{-t\lambda^2} = \operatorname{Tr}(Le^{-tL^2})$. (Zero modes contribute $0 \cdot e^0 = 0$, so they drop out automatically.)

At $s = 0$ the prefactor is $1/\Gamma(\frac{1}{2}) = 1/\sqrt{\pi}$:

$$\eta_L(0) = \frac{1}{\sqrt{\pi}}\int_0^\infty t^{-1/2}\,\operatorname{Tr}\big(L\,e^{-tL^2}\big)\,\mathrm{d}t, \tag{4.222}$$

provided the integral converges at $t \to 0$. Unlike $\operatorname{Tr} e^{-tL^2}$, the "odd" heat trace $\operatorname{Tr}(Le^{-tL^2})$ weighs each eigenvalue with its sign, so the big positive and negative contributions cancel. Whether they cancel well enough as $t \to 0$ is exactly the finiteness question.

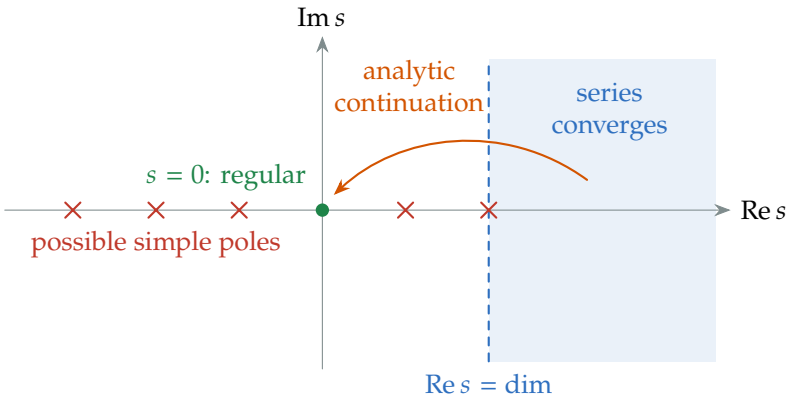


Figure 4.76: The life of a zeta or eta function in the complex s-plane. The defining series converges only in the shaded half-plane. The heat-kernel formulas (4.215) and (4.221) continue it to the left; small-t powers of the heat trace produce simple poles on the real axis (crosses). At $s = 0$ the function is regular: for ζ because $1/\Gamma(s)$ vanishes there, for η because of a deeper cancellation (Section 4.21.12).

Key idea: Zeta functions compute indices: Nash's footnote 5

Let D be elliptic on a closed manifold. The nonzero eigenvalues of D^*D and DD^* coincide (if $D^*D\phi = \mu\phi$ with $\mu \neq 0$, then $DD^*(D\phi) = \mu D\phi$ and $D\phi \neq 0$). So their zeta functions are *identical*, and (4.217) at $s = 0$ gives

$$0 \overset{(1)}{=} \zeta_{D^*D}(0) - \zeta_{DD^*}(0) \overset{(2)}{=} \big(a_n(D^*D) - \dim\operatorname{Ker} D^*D\big) - \big(a_n(DD^*) - \dim\operatorname{Ker} DD^*\big),$$

$$\operatorname{index} D \overset{(3)}{=} \dim\operatorname{Ker} D - \dim\operatorname{Ker} D^* \overset{(4)}{=} a_n(D^*D) - a_n(DD^*).$$

(1) Equal nonzero spectra give equal zeta functions.
(2) (4.217) for each operator.
(3) Definition of the index; $\operatorname{Ker} D^*D = \operatorname{Ker} D$ and $\operatorname{Ker} DD^* = \operatorname{Ker} D^*$ (if $D^*D\phi = 0$ then $\|D\phi\|^2 = \langle\phi, D^*D\phi\rangle = 0$).
(4) Rearrange the first line.

The index is a difference of *local* heat coefficients. This is the heat-equation derivation of the index theorem (Atiyah–Bott–Patodi) that Nash's footnote mentions; it is the same statement as McKean–Singer in Chapter 4. With a boundary the same idea works, and the boundary produces η: see Section 4.21.9.

4.21.6 Explicit computation: η for the Dirac operator on the circle

Nash computes one eta invariant explicitly [Nash p. 132, (4.159)–(4.161)], and it is the one every physicist should know: a Dirac fermion on a circle threaded by a magnetic flux. We do the computation three ways.

The operator and its spectrum

Nash's operator is

$$L_t = i\frac{\mathrm{d}}{\mathrm{d}\theta} + t \qquad \text{on functions on } S^1, \quad 0 < t < 1. \tag{4.223}$$

It is the Dirac operator of the circle in the presence of a constant gauge potential. With the U(1) connection $A = -it\,\mathrm{d}\theta$ (anti-Hermitian, our convention), the covariant derivative is $\nabla_\theta = \partial_\theta - it$, and $i\nabla_\theta = i\partial_\theta + t = L_t$. Physically, t is the magnetic flux through the ring in units of the flux quantum 2π (an Aharonov–Bohm flux), since the holonomy is $\exp(-\oint A) = e^{2\pi i t}$. Equivalently, by the gauge transformation $\psi \to e^{-it\theta}\psi$ one can remove t from the operator and put it into a twisted boundary condition $\psi(\theta + 2\pi) = e^{2\pi i t}\psi(\theta)$.

The eigenfunctions are Fourier modes:

$$L_t\, e^{-in\theta} \overset{(1)}{=} \big(i\cdot(-in) + t\big)e^{-in\theta} \overset{(2)}{=} (n+t)\, e^{-in\theta}, \qquad n \in \mathbb{Z}.$$

(1) $\frac{\mathrm{d}}{\mathrm{d}\theta}e^{-in\theta} = -in\, e^{-in\theta}$.
(2) $i\cdot(-i) = 1$.

So the spectrum is $\lambda_n = n + t$, $n \in \mathbb{Z}$, Nash's (4.159). It is the integer lattice shifted by t (Figure 4.77). For $t = 0$ it is symmetric about 0 and contains 0; for $t = \frac{1}{2}$ it is $\{\pm\frac{1}{2}, \pm\frac{3}{2}, \dots\}$, again symmetric; for any other t the dots sit unevenly around 0.

First way: Hurwitz zeta functions (Nash's route)

Split the spectrum by sign. With $0 < t < 1$, $n + t > 0$ exactly when $n \geq 0$, and $n + t < 0$ exactly when $n \leq -1$. Then

$$\begin{aligned}\eta_{L_t}(s) &\overset{(1)}{=} \sum_{n\geq 0}\frac{1}{(n+t)^s} - \sum_{n\leq -1}\frac{1}{|n+t|^s} \overset{(2)}{=} \sum_{n\geq 0}\frac{1}{(n+t)^s} - \sum_{m\geq 1}\frac{1}{(m-t)^s}\\ &\overset{(3)}{=} \sum_{n\geq 0}\frac{1}{(n+t)^s} - \sum_{m'\geq 0}\frac{1}{(m'+1-t)^s} \overset{(4)}{=} \zeta(s,t) - \zeta(s,1-t).\end{aligned} \tag{4.224}$$

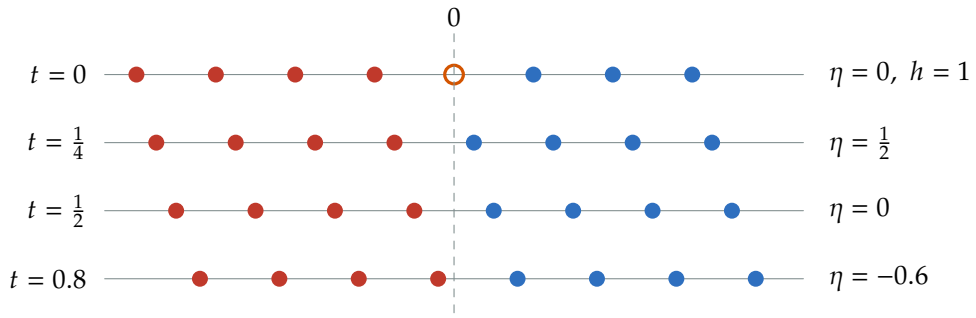


Figure 4.77: The spectrum $\{n+t\}$ of $L_t = i\,\mathrm{d}/\mathrm{d}\theta + t$ for four values of the flux t: negative eigenvalues red, positive blue, a zero mode as an orange ring. Shifting the lattice by t moves every dot the same distance to the right. Near 0 the nearest positive eigenvalue is t and the nearest negative one is $t-1$; that imbalance is what $\eta(0) = 1 - 2t$ measures. The far-away dots are balanced and cancel in the regularised count.

(1) Definition (4.211): positive eigenvalues with $+$, negative with $-$.
(2) Put $m = -n \geq 1$; then $|n+t| = m - t$.
(3) Put $m' = m - 1 \geq 0$, so $m - t = m' + (1-t)$.
(4) The **Hurwitz zeta function** $\zeta(s,t) = \sum_{n\geq 0}(n+t)^{-s}$, Nash's (4.160); it is Riemann's ζ_R when $t = 1$.

Nash quotes (from Whittaker and Watson) the value $\zeta(0,t) = \frac{1}{2} - t$. Let us derive it, since it is the heart of the matter. Apply the Mellin transform (4.214) to each term and sum the geometric series:

$$\Gamma(s)\,\zeta(s,t) \overset{(1)}{=} \int_0^\infty \varepsilon^{s-1} \sum_{n\geq 0} e^{-\varepsilon(n+t)}\,\mathrm{d}\varepsilon \overset{(2)}{=} \int_0^\infty \varepsilon^{s-1}\,\frac{e^{-\varepsilon t}}{1-e^{-\varepsilon}}\,\mathrm{d}\varepsilon.$$

(1) (4.214) with $\mu = n + t$, summed over n.
(2) $\sum_{n\geq 0} e^{-\varepsilon n} = 1/(1-e^{-\varepsilon})$.

The small-ε behaviour of the integrand controls everything. Expand:

$$\frac{e^{-\varepsilon t}}{1-e^{-\varepsilon}} \overset{(1)}{=} \Big(1 - \varepsilon t + \tfrac{1}{2}\varepsilon^2 t^2 - \cdots\Big)\Big(\frac{1}{\varepsilon} + \frac{1}{2} + \frac{\varepsilon}{12} + \cdots\Big) \overset{(2)}{=} \frac{1}{\varepsilon} + \Big(\frac{1}{2} - t\Big) + O(\varepsilon).$$

(1) Taylor series of $e^{-\varepsilon t}$; and $\frac{1}{1-e^{-\varepsilon}} = \frac{1}{\varepsilon - \varepsilon^2/2 + \varepsilon^3/6 - \cdots} = \frac{1}{\varepsilon}\big(1 + \frac{\varepsilon}{2} + \frac{\varepsilon^2}{12} + \cdots\big)$.
(2) Multiply out and collect the ε^{-1} and ε^0 terms.

Now split the integral at $\varepsilon = 1$ as in Section 4.21.5. The piece $\int_1^\infty$ is entire. In $\int_0^1$ the first two terms give $\int_0^1 \varepsilon^{s-2}\mathrm{d}\varepsilon = \frac{1}{s-1}$ and $(\frac{1}{2} - t)\int_0^1 \varepsilon^{s-1}\mathrm{d}\varepsilon = \frac{1/2 - t}{s}$, and the $O(\varepsilon)$ remainder is holomorphic for $\operatorname{Re} s > -1$. So near $s = 0$,

$$\zeta(s,t) \overset{(1)}{=} \frac{1}{\Gamma(s)}\Big(\frac{\frac{1}{2}-t}{s} + \text{regular}\Big) \overset{(2)}{=} \big(s + O(s^2)\big)\Big(\frac{\frac{1}{2}-t}{s} + \text{regular}\Big) \xrightarrow{s\to 0} \frac{1}{2} - t. \tag{4.225}$$

(1) Divide by $\Gamma(s)$; the $\frac{1}{s-1}$ term is regular at $s = 0$.
(2) $1/\Gamma(s) = s + O(s^2)$.

(The same bookkeeping at higher order gives the classical $\zeta(-k,t) = -B_{k+1}(t)/(k+1)$ with Bernoulli polynomials; here $B_1(t) = t - \frac{1}{2}$.) Putting (4.225) into (4.224):

$$\eta_{L_t}(0) \overset{(1)}{=} \Big(\frac{1}{2} - t\Big) - \Big(\frac{1}{2} - (1-t)\Big) \overset{(2)}{=} 1 - 2t, \qquad 0 < t < 1. \tag{4.226}$$

(1) $\zeta(0,t) = \frac{1}{2} - t$ and $\zeta(0,1-t) = \frac{1}{2} - (1-t)$.
(2) Arithmetic.

This is Nash's (4.161). At $t = \frac{1}{2}$ the answer is 0, as the symmetric spectrum demands. As $t \to 0^+$ the answer tends to 1: the eigenvalue t is about to become a zero mode, and it is the one unpaired positive eigenvalue.

Second way: a smooth cutoff

The ε-integrand above suggests a quicker route. Regularise the divergent "$\sum \operatorname{sgn}\lambda$" with a cutoff $e^{-\varepsilon|\lambda|}$ and let $\varepsilon \to 0$:

$$\sum_n \operatorname{sgn}(\lambda_n) e^{-\varepsilon|\lambda_n|} \overset{(1)}{=} \sum_{n\geq 0} e^{-\varepsilon(n+t)} - \sum_{m\geq 1} e^{-\varepsilon(m-t)} \overset{(2)}{=} \frac{e^{-\varepsilon t} - e^{-\varepsilon(1-t)}}{1-e^{-\varepsilon}}$$
$$\overset{(3)}{=} \frac{\varepsilon(1-2t) + O(\varepsilon^2)}{\varepsilon + O(\varepsilon^2)} \xrightarrow{\varepsilon\to 0} 1 - 2t.$$

(1) Split by sign as in (4.224).
(2) Two geometric series with ratio $e^{-\varepsilon}$; the second starts at $e^{-\varepsilon(1-t)}$.
(3) Expand numerator and denominator to first order in ε: $e^{-\varepsilon t} - e^{-\varepsilon(1-t)} = \varepsilon(1-t) - \varepsilon t + O(\varepsilon^2)$.

The two infinite sums each diverge like $1/\varepsilon$; the divergences cancel exactly, and the finite remainder is the eta invariant. This is the sense in which $\eta(0)$ is "$\#(\lambda > 0) - \#(\lambda < 0)$". (For this operator any reasonable cutoff gives the same answer; in general the zeta/heat version is the one guaranteed to be well defined.)

Third way: the heat kernel and Poisson summation

This is the method that generalises, and it shows directly why nothing diverges. We use (4.222), so we need the odd heat trace $\operatorname{Tr}(L_t e^{-\tau L_t^2}) = \sum_n (n+t) e^{-\tau(n+t)^2}$ (we write τ for the heat time to keep t for the flux). The **Poisson summation formula** says that for a rapidly decreasing f,

$$\sum_{n\in\mathbb{Z}} f(n+t) = \sum_{m\in\mathbb{Z}} \hat{f}(2\pi m)\, e^{2\pi i m t}, \qquad \hat{f}(k) = \int_{-\infty}^{\infty} f(x) e^{-ikx}\, \mathrm{d}x. \tag{4.227}$$

(The left side is a periodic function of t with period 1; the right side is its Fourier series.)

Worked example 4.137: $\eta(0) = 1 - 2t$ from the heat kernel

Take $f(x) = x\, e^{-\tau x^2}$. Its Fourier transform is

$$\hat{f}(k) \overset{(1)}{=} i\frac{\mathrm{d}}{\mathrm{d}k} \int e^{-\tau x^2} e^{-ikx}\, \mathrm{d}x \overset{(2)}{=} i\frac{\mathrm{d}}{\mathrm{d}k}\left(\sqrt{\frac{\pi}{\tau}} e^{-k^2/4\tau}\right) \overset{(3)}{=} -\frac{ik}{2\tau}\sqrt{\frac{\pi}{\tau}}\, e^{-k^2/4\tau}.$$

(1) $i\frac{\mathrm{d}}{\mathrm{d}k} e^{-ikx} = x\, e^{-ikx}$.
(2) The Gaussian integral.
(3) Differentiate.

Note $\hat{f}(0) = 0$ because f is odd. Hence by (4.227)

$$\operatorname{Tr}\big(L_t e^{-\tau L_t^2}\big) = -i\pi^{3/2}\tau^{-3/2} \sum_{m\neq 0} m\, e^{-\pi^2 m^2/\tau} e^{2\pi i m t}. \tag{4.228}$$

Every term is *exponentially small* as $\tau \to 0$: the odd heat trace of the circle has no small-τ expansion at all. So the integral (4.222) converges at $\tau = 0$ and

$$\eta_{L_t}(0) \overset{(1)}{=} \frac{1}{\sqrt{\pi}}\left(-i\pi^{3/2}\right) \sum_{m\neq 0} m\, e^{2\pi i m t} \int_0^\infty \tau^{-2} e^{-\pi^2 m^2/\tau}\, \mathrm{d}\tau \overset{(2)}{=} -i\pi \sum_{m\neq 0} \frac{m\, e^{2\pi i m t}}{\pi^2 m^2}$$
$$\overset{(3)}{=} -\frac{i}{\pi} \sum_{m\geq 1} \frac{e^{2\pi i m t} - e^{-2\pi i m t}}{m} \overset{(4)}{=} \frac{2}{\pi} \sum_{m\geq 1} \frac{\sin 2\pi m t}{m} \overset{(5)}{=} \frac{2}{\pi} \cdot \pi\left(\frac{1}{2} - t\right) = 1 - 2t.$$

(1) Insert (4.228) into (4.222); $\tau^{-1/2}\tau^{-3/2} = \tau^{-2}$.
(2) Substitute $u = 1/\tau$: $\int_0^\infty \tau^{-2} e^{-c/\tau} \mathrm{d}\tau = \int_0^\infty e^{-cu}\mathrm{d}u = 1/c$ with $c = \pi^2 m^2$.
(3) Pair m with $-m$.
(4) $e^{ix} - e^{-ix} = 2i \sin x$ and $-\frac{i}{\pi} \cdot 2i = \frac{2}{\pi}$.
(5) The Fourier series of the sawtooth: $\sum_{m\geq 1} \frac{\sin mx}{m} = \frac{\pi - x}{2}$ for $0 < x < 2\pi$, with $x = 2\pi t$.

The third method also tells us what happens at $t = 0$: the Fourier series converges to 0, the average of the two one-sided limits ± 1.

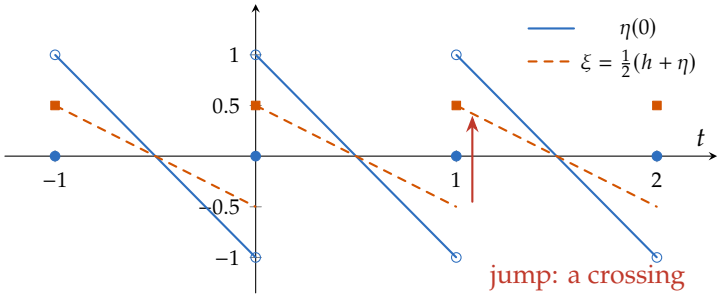


Figure 4.78: The eta invariant of $L_t = i\,\mathrm{d}/\mathrm{d}\theta + t$ as a function of the flux t (solid): the sawtooth $1 - 2t$ on $(0, 1)$, repeated with period 1. Between integers it changes smoothly with slope -2; at each integer it jumps by 2, because one eigenvalue $n + t$ passes through zero; at the integers themselves (filled dots) the spectrum is symmetric and $\eta = 0$. The dashed line is $\xi = \frac{1}{2}(h + \eta)$, with $h = \dim \mathrm{Ker}\, L_t$; it jumps by exactly 1 and takes the right-hand value at the integers (squares). The jumps are the spectral flow of Section 4.21.10.

What the answer teaches

- *η is a measure of asymmetry.* It vanishes exactly at $t = \frac{1}{2}$ (symmetric spectrum) and at $t = 0$ (symmetric, with a zero mode), just as Nash says.

- *η is not local.* A local invariant of the circle with a flat connection could only be an integral of a local density, and a flat connection has no local field strength at all. Yet η depends on t: it sees the holonomy $e^{2\pi i t}$, a global Aharonov–Bohm quantity.

- *η is gauge invariant but discontinuous.* Under a large gauge transformation $t \to t + 1$ the spectrum is relabelled and η is unchanged: η is a function of the holonomy. But it jumps by 2 whenever an eigenvalue crosses zero.

- *$e^{i\pi\eta}$ is smooth.* Because the jumps are by 2, $e^{i\pi\eta(0)} = e^{i\pi(1-2t)} = -e^{-2\pi i t}$ is a smooth function of t, and it is exactly (minus) the inverse holonomy. Phases of this type are the phases of chiral fermion determinants, which is why η enters the theory of anomalies (Chapter 10).

- *η does not care about the size of the circle.* On a circle of length ℓ the operator is $\frac{2\pi}{\ell}L_t$, and $\eta_{cL}(s) = c^{-s}\eta_L(s)$, so $\eta(0)$ is unchanged. In general $\eta(0)$ is invariant under rescaling of the metric.

Worked example 4.138: Spin structures, and a matrix of fluxes

The plain Dirac operator $-i\,\mathrm{d}/\mathrm{d}\theta$ on the circle depends on a choice of **spin structure** (Chapter T4): spinors may be periodic (the "Ramond" or non-bounding structure) or antiperiodic (the "Neveu–Schwarz" structure, the one that extends over a disc). Antiperiodic spinors are $e^{i(n+\frac12)\theta}$: this is $t = \frac12$, so $\eta = 0$, $h = 0$. Periodic spinors are $t = 0$: $\eta = 0$, $h = 1$. So $\xi = \frac12(h+\eta)$ is 0 or $\frac12$; that $\frac12$ is the zero-mode contribution that distinguishes the two spin structures.

Now take an N-component fermion with a constant diagonal potential, $L = i\,\mathrm{d}/\mathrm{d}\theta \otimes \mathbf{1}_N + \mathrm{diag}(t_1,\dots,t_N)$, i.e. a flat $\mathrm{U}(N)$ connection with holonomy $\mathrm{diag}(e^{2\pi i t_j})$. The spectrum is the union of the N shifted lattices, and η is additive over direct sums, so

$$\eta_L(0) = \sum_{j=1}^{N} \bigl(1 - 2\{t_j\}\bigr) \quad (t_j \notin \mathbb{Z}), \qquad \{t\} = t - \lfloor t \rfloor .$$

For example the $\mathrm{SU}(2)$ holonomy $\mathrm{diag}(e^{2\pi i t}, e^{-2\pi i t})$ with $0 < t < 1$ gives $(1-2t) + (1-2(1-t)) = 0$: the $\mathrm{SU}(2)$ case is symmetric, because $\mathrm{diag}(t,-t)$ has a spectrum symmetric under $\lambda \to -\lambda$ combined with $n \to -n$.

Try it: Scaling and orientation

(a) Show that $\eta_{cL}(0) = \eta_L(0)$ for $c > 0$ and $\eta_{-L}(0) = -\eta_L(0)$. (b) Reversing the orientation of the circle replaces θ by $-\theta$. What does it do to L_t and to η? *Answer:* (a) $\eta_{cL}(s) = \sum \mathrm{sgn}(c\lambda)|c\lambda|^{-s} = c^{-s}\eta_L(s)$, and $c^0 = 1$; flipping every sign flips the sum. (b) $i\,\mathrm{d}/\mathrm{d}\theta \to -i\,\mathrm{d}/\mathrm{d}\theta$, so $L_t \to -L_{-t}$, with spectrum $\{-(n-t)\} = \{n+t\}$ again after relabelling. So this particular η is unchanged; the orientation-odd invariant is η of the operator that also keeps track of the chirality, as in Section 4.21.4.

4.21.7 The Atiyah–Patodi–Singer index theorem

Nash now leaves the signature operator and states the general theorem [Nash pp. 132–133, (4.162)–(4.164)]. Everything we have built so far fits into it.

The set-up

Let X be a compact manifold with boundary, $E, F \to X$ vector bundles, and

$$D : \Gamma(X,E) \to \Gamma(X,F) \tag{4.229}$$

a first-order elliptic operator. Assume that near the boundary X is isometric to a product $\partial X \times [0,\varepsilon)$ and D has the form

$$D = A\Bigl(\frac{\partial}{\partial n} + L\Bigr) \quad \text{near } \partial X, \tag{4.230}$$

where n is the inward normal coordinate, A is an invertible bundle map (independent of n), and L is a self-adjoint elliptic operator on ∂X, independent of n. Impose the global boundary condition

$$\Pi\bigl(u|_{\partial X}\bigr) = 0, \qquad u \in \Gamma(X,E), \tag{4.231}$$

where Π is the spectral projection of L onto its eigenvalues $\lambda \geq 0$. This is Nash's (4.162). The zero eigenvalue now needs care. Nash defines [Nash (4.163)]

$$h = \dim \operatorname{Ker} L, \qquad \eta(s) = \sum_{\lambda_n \neq 0} \frac{\operatorname{sgn} \lambda_n}{|\lambda_n|^s}. \tag{4.232}$$

Key idea: The APS index theorem (Nash's "index theory for boundary value problems", (4.164))

The boundary problem (4.230)–(4.231) is Fredholm, and its index is

$$\operatorname{index} D = \int_X P \; - \; \frac{h + \eta(0)}{2}. \tag{4.233}$$

Here $P = P(R, F)$ is the *same* local index density (a combination of characteristic forms built from the curvature) that gives index D on a closed manifold: $\operatorname{ch}(F)\,\operatorname{td}(X)$ for a twisted $\bar{\partial}$, $\operatorname{ch}(F)\hat{A}(X)$ for a twisted Dirac operator, $2^{n/2}L(X)$ for the signature operator (in the normalisations of Chapter 4). The combination $\xi = \frac{1}{2}(h + \eta(0))$ is often called the **xi invariant** of L.

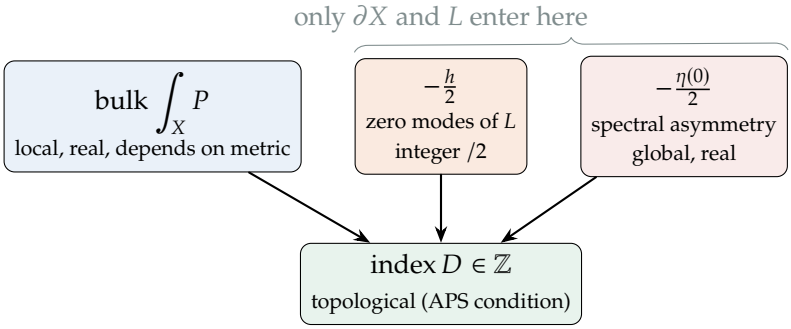


Figure 4.79: The three pieces of the APS formula. The bulk integral is a local density and changes continuously with the metric and gauge field. The integer index D does not. The boundary pieces absorb the difference: η is non-local, and h records zero modes of L. Only their sum with the bulk integral is an integer.

How to read the formula

- *On a closed manifold* there is no boundary, $h = \eta = 0$, and (4.233) is the ordinary index theorem $\operatorname{index} D = \int_X P$.
- *The bulk term is not an integer.* On a manifold with boundary $\int_X P$ is just a real number, which varies continuously with the geometry. The boundary term $-\frac{1}{2}(h + \eta)$ is what makes the total an integer. In particular (4.233) determines $\eta(0)$ modulo 2 from the bulk: $\eta(0) \equiv 2\int_X P - h \pmod 2$.
- *The signature operator.* With $D = d_+$, $\int_X P = \frac{1}{3}\int_X p_1$ in four dimensions, and (4.233) is (4.212) corrected by h (Section 4.21.4).
- *Why the zero modes are counted with* $-\frac{1}{2}$. The APS condition puts $\operatorname{Ker} L$ into H_+, i.e. it forces the zero-mode components of $u|_{\partial X}$ to vanish. That is "half" of the freedom a zero mode has (on the elongated manifold it is a bounded but not square-integrable solution, sitting on the fence between allowed and forbidden). The $-\frac{h}{2}$ is the price.

Worked example 4.139: Consistency check: the adjoint problem

Take $A = 1$ for simplicity, so $D = \partial_n + L$ near the boundary. Its formal adjoint is $D^* = -\partial_n + L = -(\partial_n - L)$, whose boundary operator is $-L$. The *true* adjoint of the APS problem for D carries the complementary condition $(1 - \Pi)v|_{\partial X} = 0$ (boundary data of v in the span of $\lambda \geq 0$), and its index is $-\operatorname{index} D$. The APS problem for D^* in its own right instead demands boundary data in $H_-(-L)$, the span of $\lambda > 0$: a stricter condition, differing by the h-dimensional $\operatorname{Ker} L$. The theorem is consistent with this:

$$\operatorname{index}_{\mathrm{APS}}(D^*) \overset{(1)}{=} \int_X P^* - \frac{h + \eta_{-L}(0)}{2} \overset{(2)}{=} -\int_X P - \frac{h - \eta_L(0)}{2}$$
$$\overset{(3)}{=} -\Big(\int_X P - \frac{h + \eta_L(0)}{2}\Big) - h \overset{(4)}{=} -\operatorname{index}_{\mathrm{APS}}(D) - h.$$

(1) (4.233) for D^*, whose boundary operator is $-L$ with the same kernel.
(2) The local index density of the adjoint is $P^* = -P$ (the roles of D^*D and DD^* swap), and $\eta_{-L} = -\eta_L$.
(3) Rearrange: $-\frac{h-\eta}{2} = \frac{h+\eta}{2} - h$.
(4) (4.233) for D.

So imposing the APS condition on D^* loses exactly h from the index relative to the true adjoint, which is what the h extra constraints should do. We check it by hand in the next example.

The simplest example: a domain wall on an interval

The smallest possible case of the theorem is one-dimensional: $X = [0, 1]$, whose boundary is two points. It is worth doing completely, because it is the Jackiw–Rebbi zero mode of a fermion on a domain wall, and because it shows the theorem working with nothing hidden.

Worked example 4.140: The APS theorem on an interval: Jackiw–Rebbi

Let $D = \dfrac{\mathrm{d}}{\mathrm{d}x} + a(x)$ acting on complex functions on $[0, 1]$, with a a smooth real function (a position-dependent mass). *The boundary operators.* At $x = 0$ the inward normal coordinate is $n = x$, so $D = \partial_n + a(0)$: $A = 1$ and $L_0 = a(0)$, a 1×1 matrix (an operator on the 0-dimensional boundary component). At $x = 1$ the inward coordinate is $n = 1 - x$, so $\partial_x = -\partial_n$ and $D = -(\partial_n - a(1))$: $A = -1$ and $L_1 = -a(1)$. The eta invariant of a 1×1 matrix (λ) is simply $\operatorname{sgn}\lambda$, and $h = 1$ if $\lambda = 0$.
Direct count. Solutions of $Du = 0$ are $u(x) = u(0)\exp\big(-\int_0^x a\big)$, never zero unless identically zero. The APS condition allows $u(0) \neq 0$ only if $L_0 = a(0) < 0$ and $u(1) \neq 0$ only if $L_1 = -a(1) < 0$. So

$$\dim \operatorname{Ker} D = \begin{cases} 1 & a(0) < 0 < a(1), \\ 0 & \text{otherwise.} \end{cases}$$

The adjoint $D^* = -\frac{\mathrm{d}}{\mathrm{d}x} + a$ has solutions $v = v(0)\exp\big(+\int_0^x a\big)$, with the complementary condition: $v(0) \neq 0$ is allowed iff $a(0) \geq 0$ and $v(1) \neq 0$ iff $-a(1) \geq 0$. So $\dim \operatorname{Coker} D = 1$ iff $a(1) \leq 0 \leq a(0)$.

The theorem. In one dimension there is no local index density ($P = 0$: the heat-kernel coefficients of $D^*D = -\partial^2 + a^2 - a'$ and $DD^* = -\partial^2 + a^2 + a'$ agree at order t^0). So

$$\operatorname{index} D \overset{(1)}{=} 0 - \frac{h_0 + \eta(L_0)}{2} - \frac{h_1 + \eta(L_1)}{2} \overset{(2)}{=} \frac{\operatorname{sgn} a(1) - \operatorname{sgn} a(0)}{2} - \frac{h_0 + h_1}{2}.$$

(1) (4.233), one boundary term per boundary point.
(2) $\eta(L_0) = \operatorname{sgn} a(0)$, $\eta(L_1) = \operatorname{sgn}(-a(1)) = -\operatorname{sgn} a(1)$.

Check all cases: $a(0) < 0 < a(1)$ gives $\frac{1+1}{2} = 1$ (one zero mode, no cokernel) ✓; $a(0) > 0 > a(1)$ gives -1 (cokernel only) ✓; equal signs give 0 ✓; and $a \equiv 0$ gives $0 - \frac{1+1}{2} = -1$, matching the direct count (no kernel; the constant v is in the cokernel) ✓. For $a \equiv 0$ one also sees Example 4.139: the APS problem for $D^* = -\mathrm{d}/\mathrm{d}x$ has index $-1 = -(-1) - 2$.

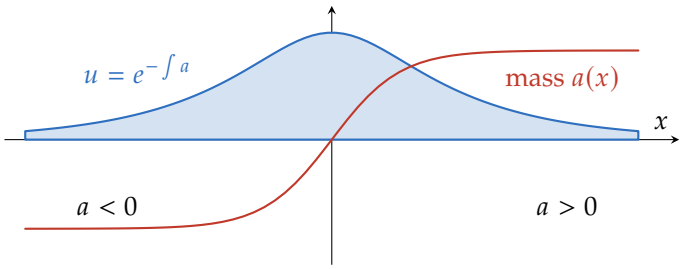


Figure 4.80: The Jackiw–Rebbi zero mode. A mass profile $a(x)$ that changes sign from negative to positive traps exactly one normalisable solution of $(\mathrm{d}/\mathrm{d}x + a)u = 0$, $u = \exp(-\int a)$, localised on the wall. On a finite interval the APS condition reproduces this count through the boundary eta invariants $\operatorname{sgn} a(0)$ and $-\operatorname{sgn} a(1)$; on the whole line (the elongated manifold) it is literally the L^2 kernel. The index is the number of eigenvalues of the "boundary family" $a(x)$ that cross from negative to positive: a first example of spectral flow.

Physical picture: Domain walls and fermion zero modes

In the Jackiw–Rebbi model a Dirac fermion in $1+1$ dimensions couples to a scalar kink $\phi(x)$ through a mass term $g\phi\bar\psi\psi$. With a suitable choice of gamma matrices, the zero-energy Dirac equation for one spinor component reduces to $(\mathrm{d}/\mathrm{d}x + g\phi)u = 0$. A kink, with $\phi(\pm\infty) = \pm v$, therefore carries one fermion zero mode, and the soliton's fermion number is $\pm\frac{1}{2}$ depending on whether the mode is filled. The number of zero modes depends only on the signs of the mass at the two ends, which is exactly the data $\eta(L_0), \eta(L_1)$. The same mechanism puts chiral fermions on domain walls in lattice QCD (domain-wall fermions) and gives the edge modes of topological insulators.

A disc threaded by magnetic flux

Our next example has all three terms of the APS formula nonzero, and it uses the circle computation of Section 4.21.6 as its boundary term.

Worked example 4.141: The $\bar\partial$ operator on a disc with a vortex

Let X be the unit disc $|z| \le 1$, with a metric that equals $|\mathrm{d}z/z|^2 = \mathrm{d}r^2/r^2 + \mathrm{d}\theta^2$ near $|z| = 1$ (so the collar is a flat cylinder) and is smooth inside. Couple $\bar\partial$ on functions to a U(1) gauge field as in Section 4.17: $\bar\partial_A = \partial_{\bar z} + A_{\bar z}$ with $A_{\bar z} = \partial_{\bar z}\phi$ for a real radial function $\phi(r)$, smooth at the origin, and equal to $\alpha \ln r$ near the boundary. The field strength is $F = -i\nabla^2\phi\,\mathrm{d}x \wedge \mathrm{d}y$,

so the flux is confined to the inside and the connection is flat on the collar. The Chern number of the flux is

$$\int_X c_1(F) \overset{(1)}{=} \frac{1}{2\pi}\int_X \nabla^2\phi \,\mathrm{d}x\,\mathrm{d}y \overset{(2)}{=} \frac{1}{2\pi}\oint_{r=1} \partial_r\phi \,\mathrm{d}s \overset{(3)}{=} \frac{1}{2\pi}\cdot\alpha\cdot 2\pi = \alpha.$$

(1) $c_1 = \frac{i}{2\pi}F = \frac{1}{2\pi}\nabla^2\phi\,\mathrm{d}x\wedge\mathrm{d}y$.
(2) The divergence theorem in the plane.
(3) $\partial_r(\alpha\ln r) = \alpha$ at $r=1$, and the circle has length 2π.

So α is the magnetic flux in units of 2π, and it need not be an integer.

The boundary operator. In polar coordinates $\partial_{\bar z} = \frac{e^{i\theta}}{2}\big(\partial_r + \frac{i}{r}\partial_\theta\big)$ (check: it kills $z = re^{i\theta}$ and sends $\bar z$ to 1). On the collar $A_{\bar z} = \frac{e^{i\theta}}{2}\phi'(r) = \frac{e^{i\theta}}{2}\frac{\alpha}{r}$. The inward normal coordinate for the metric $\mathrm{d}r^2/r^2 + \mathrm{d}\theta^2$ is $n = -\ln r$, so $\partial_n = -r\partial_r$. Therefore

$$\bar\partial_A \overset{(1)}{=} \frac{e^{i\theta}}{2r}\big(r\partial_r + i\partial_\theta + \alpha\big) \overset{(2)}{=} -\frac{e^{i\theta}}{2r}\Big(\partial_n - \big(i\partial_\theta + \alpha\big)\Big) \overset{(3)}{=} -\frac{e^{i\theta}}{2r}\big(\partial_n + L\big), \quad L = -i\partial_\theta - \alpha.$$

(1) Add the two pieces and factor out $\frac{e^{i\theta}}{2r}$.
(2) $r\partial_r = -\partial_n$.
(3) Definition of L; the factor $A = -\frac{e^{i\theta}}{2r}$ is invertible.

L acts on $e^{ik\theta}$ by $k-\alpha$, $k\in\mathbb{Z}$.

Direct count of the kernel. Solutions of $\bar\partial_A\psi = 0$ are $\psi = e^{-\phi}f(z)$ with f holomorphic on the disc (Section 4.17), so $\psi = e^{-\phi}\sum_{k\ge0}a_k z^k$. On the boundary ($\phi(1)=0$) this is $\sum_{k\ge0}a_k e^{ik\theta}$. The APS condition kills every coefficient with $k-\alpha\ge 0$. So the kernel is spanned by $e^{-\phi}z^k$ with $0\le k<\alpha$:

$$\dim\operatorname{Ker}\bar\partial_A = \#\{k\in\mathbb{Z}:\ 0\le k<\alpha\} = \lfloor\alpha\rfloor + 1 \qquad (\alpha>0,\ \alpha\notin\mathbb{Z}).$$

The analogous computation for the adjoint (antiholomorphic modes $e^{\phi}\bar z^k$, with the complementary condition) finds nothing when $\alpha>0$, so $\operatorname{index}\bar\partial_A = \lfloor\alpha\rfloor+1$.

The theorem. The local density for $\bar\partial$ twisted by a line bundle is the degree-2 part of $\operatorname{ch}(F)\operatorname{td}(X)$, namely $c_1(F)+\frac12 c_1(X)$, and $c_1(X)$ is the Euler form. The spectrum $\{k-\alpha\}$ is Nash's $\{n+t\}$ with $t = 1-\{\alpha\}\in(0,1)$ (write $-\alpha = -(\lfloor\alpha\rfloor+1)+(1-\{\alpha\})$ and relabel $n = k-\lfloor\alpha\rfloor-1$). Hence $h=0$ and $\eta(0) = 1-2t = 2\{\alpha\}-1$, by (4.226). Then

$$\operatorname{index}\bar\partial_A \overset{(1)}{=} \int_X c_1(F) + \frac12\int_X e(X) - \frac{h+\eta(0)}{2} \overset{(2)}{=} \alpha + \frac12 - \frac{0+2\{\alpha\}-1}{2}$$
$$\overset{(3)}{=} \alpha - \{\alpha\} + 1 \overset{(4)}{=} \lfloor\alpha\rfloor + 1.$$

(1) (4.233) with $P = c_1(F) + \frac12 e(X)$.
(2) The flux integral above; $\int_X e = \chi(\text{disc}) = 1$ by Gauss–Bonnet for a product collar (4.207); the circle eta invariant.
(3) Arithmetic.
(4) $\alpha = \lfloor\alpha\rfloor + \{\alpha\}$.

This agrees with the direct count. If α is an integer, L has the zero mode $e^{i\alpha\theta}$ ($h=1$), the spectrum is symmetric ($\eta=0$), and the formula gives $\alpha+\frac12-\frac12=\alpha$; the direct count ($0\le k<\alpha$) gives α too. With no flux ($\alpha=0$) the index is 0: the constant function is killed by the boundary condition.

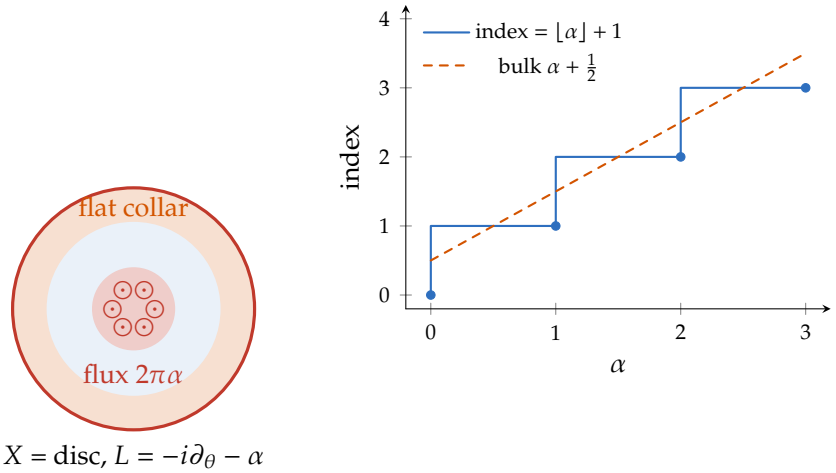


Figure 4.81: Left: a disc threaded by a vortex of flux $2\pi\alpha$, with a flat collar near the rim. Right: the index of $\bar\partial_A$ with APS boundary conditions (solid staircase) and the bulk integral $\alpha + \frac{1}{2}$ (dashed). The vertical gap between them is $\frac{1}{2}(h + \eta(0)) = \{\alpha\} - \frac{1}{2}$ for non-integer α and $\frac{1}{2}$ at the integers: the boundary eta invariant turns a continuous bulk number into an integer.

Nash's example: $\bar\partial$ on a Riemann surface with holes

Now the example Nash works out [Nash pp. 133–135, (4.165)–(4.174)]. Take a closed Riemann surface Σ_g of genus g and remove p disjoint discs $D_1, \dots, D_p$ [Nash (4.165)]:

$$X = \Sigma_g - \bigcup_{i=1}^{p} D_i. \tag{4.234}$$

Then X is a complex manifold whose boundary is p disjoint circles.

Making the metric a product near the boundary. Choose a local complex coordinate z_i on a neighbourhood of each D_i with $z_i = 0$ at its centre, and let $D_i = \{|z_i| < \rho\}$. Near ∂D_i choose the metric [Nash (4.166)]

$$\mathrm{d}s^2 = \left|\frac{\mathrm{d}z_i}{z_i}\right|^2 = \frac{\mathrm{d}r_i^2}{r_i^2} + \mathrm{d}\theta_i^2 = (\mathrm{d}\ln r_i)^2 + \mathrm{d}\theta_i^2, \tag{4.235}$$

using $z_i = r_i e^{i\theta_i}$ and $\mathrm{d}z/z = \mathrm{d}r/r + i\,\mathrm{d}\theta$. In the coordinate $n = \ln(r_i/\rho)$ this is $\mathrm{d}n^2 + \mathrm{d}\theta^2$: a flat cylinder of circumference 2π, i.e. a product. The metric is singular at $z_i = 0$ (the cylinder is infinitely long there), but $z_i = 0$ is not in X, so on X it is perfectly regular. Glue it to any metric on the rest of X (Figure 4.82).

The operator near each boundary circle. As in the flux disc (now with $\alpha = 0$ and X on the *outside* of D_i, so the inward normal is $\partial_n = r\partial_r$):

$$\bar\partial \overset{(1)}{=} \frac{e^{i\theta}}{2}\left(\frac{\partial}{\partial r} + \frac{i}{r}\frac{\partial}{\partial\theta}\right) \overset{(2)}{=} \frac{e^{i\theta}}{2r}\left(\frac{\partial}{\partial n} + i\frac{\partial}{\partial\theta}\right), \tag{4.236}$$

(1) Polar form of $\partial_{\bar z}$, Nash's (4.167).
(2) $r\partial_r = \partial_n$.

So $A = \frac{e^{i\theta}}{2r}$ and $L = i\,\partial_\theta$ on each circle.

Watch out: Nash's $-\frac{i}{r}\partial_\theta$

Nash writes the tangential operator as $(-i/r)\partial/\partial\theta$. The factor $1/r$ disappears once one uses the arc-length normal coordinate $n = \ln r$ of the product metric (4.235), as the normal form (4.230) requires; the overall sign depends on which way the normal points and on how $\Omega^{0,1}$ is identified with functions. None of this matters here: the spectrum is $\mathbb{Z}$ either way.

The boundary invariants. $L = i\partial_\theta$ has eigenfunctions $e^{-ik\theta}$ with eigenvalues $k \in \mathbb{Z}$ (the $t = 0$ case of (4.223)). Its only zero mode is the constant, so $h = 1$ for each circle. The nonzero eigenvalues $\pm1, \pm2, \dots$ pair off, so $\eta(s) \equiv 0$ and $\eta(0) = 0$. Each of the p circles contributes the same $\frac{1}{2}(h + \eta(0)) = \frac{1}{2}$.

The bulk term. For $\bar\partial$ on functions the local index density is the Todd form, $P = \mathrm{td}(X)$ (Chapter 4), whose top component in complex dimension one is [Nash (4.169)]

$$\mathrm{td}(X) = 1 + \frac{c_1(X)}{2}, \qquad c_1(X) = e(X)$$

(the top Chern class of a complex manifold is its Euler class). So [Nash (4.168), (4.170)]

$$\operatorname{index}\bar\partial = \int_X \mathrm{td}(X) - p\,\frac{h + \eta(0)}{2} = \frac{1}{2}\int_X e(X) - \frac{p}{2}. \tag{4.237}$$

The Euler integral by Gauss–Bonnet. Nash computes $\int_X e$ by putting the discs back [Nash (4.171)–(4.173)]. Extend the metric of X over each D_i, so that $\Sigma_g = X \cup \bigcup D_i$ has a smooth metric (cap each cylindrical collar with a hemisphere, as in the test tube of Figure 4.74). Then

$$\begin{aligned}\int_X e(X) &\overset{(1)}{=} \int_{\Sigma_g} e(\Sigma_g) - \sum_{i=1}^{p}\int_{D_i} e(D_i) \overset{(2)}{=} (2-2g) - \sum_{i=1}^{p}\Big(\chi(D_i) - \oint_{\partial D_i}\frac{k_g}{2\pi}\,\mathrm{d}s\Big)\\ &\overset{(3)}{=} (2-2g) - \sum_{i=1}^{p}(1-0) \overset{(4)}{=} 2 - 2g - p.\end{aligned}$$

(1) The integral over Σ_g is the sum of the integrals over the pieces, [Nash (4.171)].
(2) Gauss–Bonnet on the closed Σ_g [Nash (4.172)] and with boundary on each D_i, (4.206) = [Nash (4.173)].
(3) $\chi(D_i) = 1$, and ∂D_i is a geodesic of the product collar, so $k_g = 0$.
(4) Sum.

(This is also $\chi(X)$, directly by (4.207): removing a disc lowers the Euler characteristic by one.) Putting it into (4.237) [Nash (4.174)]:

$$\operatorname{index}\bar\partial \overset{(1)}{=} \frac{1}{2}\big(\chi(\Sigma_g) - p\,\chi(D_i)\big) - \frac{p}{2}\,h \overset{(2)}{=} \frac{1}{2}(2-2g-p) - \frac{p}{2} \overset{(3)}{=} 1 - g - p. \tag{4.238}$$

(1) The Euler integral just computed, and $\eta(0) = 0$.
(2) $\chi(\Sigma_g) = 2 - 2g$, $\chi(D_i) = 1$, $h = 1$.
(3) Arithmetic.

Worked example 4.142: Checking $1 - g - p$ by hand on a cylinder

Take $g = 0$, $p = 2$: the sphere minus two discs is an annulus, and with the metric (4.235) it is exactly the flat cylinder $[0, \ell] \times S^1$ with coordinate $w = u + i\theta$, where $u = \ln r$. The formula predicts $\operatorname{index}\bar\partial = 1 - 0 - 2 = -1$. Directly: on the cylinder $\bar\partial = \partial_{\bar w} = \frac{1}{2}(\partial_u + i\partial_\theta)$. *Kernel*:

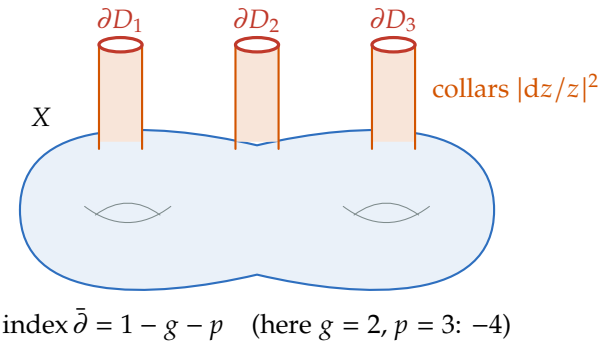


Figure 4.82: Nash's example: a genus-g surface with p discs removed. The metric $|dz_i/z_i|^2$ turns a neighbourhood of each hole into a flat cylinder (orange), so the product-collar hypothesis of the APS theorem holds. Each boundary circle has $L = i\partial_\theta$ with $h = 1$, $\eta = 0$ and contributes $-\frac{1}{2}$; the bulk Todd integral contributes $\frac{1}{2}\chi(X) = \frac{1}{2}(2 - 2g - p)$.

holomorphic functions $\sum_k c_k e^{kw} = \sum_k c_k e^{ku} e^{ik\theta}$. At the end $u = 0$ the inward normal is $+\partial_u$ and $L = i\partial_\theta$, with eigenvalue $-k$ on $e^{ik\theta}$; APS needs $-k < 0$, i.e. $k > 0$. At $u = \ell$ the inward normal is $-\partial_u$, $L = -i\partial_\theta$, eigenvalue k; APS needs $k < 0$. No k satisfies both, so $\operatorname{Ker}\bar\partial = 0$. *Cokernel*: identifying $(0,1)$-forms with functions, the adjoint is $-\partial_w$, with antiholomorphic solutions $\sum_k d_k e^{k\bar w}$. The complementary conditions now keep boundary data with eigenvalue ≥ 0 at each end: $k \leq 0$ at one end and $k \geq 0$ at the other. Only $k = 0$, the constant, survives. So $\dim\operatorname{Coker} = 1$ and the index is -1 ✓. The single cokernel element is the constant: it is "half-killed" at each end by the zero mode of L, which is the $-\frac{h}{2} - \frac{h}{2} = -1$ of the formula (the bulk is flat, $\int \mathrm{td} = 0$).
For $g = 0$, $p = 1$ the formula gives 0, which is the $\alpha = 0$ case of Example 4.141.

Physical picture: What each hole does to Riemann–Roch

On the closed surface Σ_g, Riemann–Roch gives $\operatorname{index}\bar\partial = \dim H^0 - \dim H^1 = 1 - g$ (Chapter 4). Each hole lowers the index by exactly one, and the formula splits that unit into two equal halves: $\frac{1}{2}$ from the bulk (removing a disc lowers χ by one, and the Todd density is $\frac{1}{2}e$) and $\frac{1}{2}$ from the zero mode of the new boundary circle. No eta invariant appears, because the round circle with no flux has a symmetric spectrum. Adding a flux through a hole (Example 4.141) is what switches η on.

4.21.8 Which indices need a boundary correction?

Nash makes a short but important remark [Nash p. 133]. For the signature operator the boundary corrections are genuinely there: the eta invariant of the boundary is in general nonzero. For the Euler characteristic, on the other hand, we saw in (4.207) that a product metric makes the bulk formula exact, with no correction at all. Nash adds that this persists in higher dimensions, and draws the conclusion that "the index of the de Rham complex does not depend on non-local boundary conditions". Why should the two complexes behave so differently? There are two separate reasons, and both are worth seeing concretely.

- *The de Rham complex admits local boundary conditions.* The topological obstruction of Section 4.21.1 vanishes for it, and there are two natural local elliptic conditions, called **absolute** and **relative**. Their indices are the Euler characteristics $\chi(X)$ and $\chi(X, \partial X)$.
- *Even with the APS condition, no eta invariant appears*: the boundary operator of the de Rham

complex has a spectrum that is symmetric under $\lambda \to -\lambda$, so $\eta \equiv 0$ and only the integer h survives.

The signature operator has neither property. We now check both statements by explicit computation.

Absolute and relative conditions on an interval

Working definition 4.63 (absolute and relative boundary conditions)

Near ∂X write a form as $\omega = \omega_{\tan} + \mathrm{d}n \wedge \omega_{\mathrm{nor}}$, where neither piece contains $\mathrm{d}n$ [n is the normal distance from the boundary].

- **Relative** (Dirichlet-type) condition: the tangential part vanishes on the boundary, $\omega_{\tan}|_{\partial X} = 0$ (and the same for $\mathrm{d}^*\omega$). For functions it is $f|_{\partial X} = 0$.
- **Absolute** (Neumann-type) condition: the normal part vanishes, $\omega_{\mathrm{nor}}|_{\partial X} = 0$ (and the same for $\mathrm{d}\omega$). For functions it is $\partial_n f|_{\partial X} = 0$.

Both are *local*: they are pointwise linear conditions on the boundary values. With them, the harmonic forms compute $H^p(X, \partial X)$ [closed forms whose tangential part vanishes on ∂X, modulo d of such forms] (relative) and $H^p(X)$ (absolute), so $\mathrm{index}_{\mathrm{rel}}(\mathrm{d}+\mathrm{d}^*) = \chi(X, \partial X)$ [$\sum_p (-1)^p \dim H^p(X, \partial X)$] and $\mathrm{index}_{\mathrm{abs}}(\mathrm{d}+\mathrm{d}^*) = \chi(X)$.

The relative Euler characteristic is computed from the exact sequence of the pair, which gives $\chi(X, \partial X) = \chi(X) - \chi(\partial X)$. In even dimensions ∂X is a closed odd-dimensional manifold, so $\chi(\partial X) = 0$ and the two answers agree. In odd dimensions they differ.

Worked example 4.143: The Euler operator on an interval, three ways

Let $X = [0, 1]$. Even forms are functions f and odd forms are $g\,\mathrm{d}x$, so the Euler operator is simply

$$D = \mathrm{d} + \mathrm{d}^* : \ \Omega^0 \to \Omega^1, \qquad Df = f'\,\mathrm{d}x, \qquad D^*(g\,\mathrm{d}x) = -g'.$$

Near $x = 0$ the inward normal is $n = x$, so $D = \partial_n + 0$: $A = 1$ and the boundary operator is $L_0 = 0$ on the one-dimensional space of values at the point. Near $x = 1$, $n = 1 - x$ and $D = -(\partial_n - 0)$: again $L_1 = 0$. So each end has $h = 1$ and $\eta = 0$.
Absolute condition ($f'(0) = f'(1) = 0$ for f; the normal part $g\,\mathrm{d}x$ must vanish, $g(0) = g(1) = 0$). Kernel: $f' = 0$, so f is constant; this satisfies $f' = 0$ at the ends, $\dim \mathrm{Ker} = 1$. Cokernel: $g' = 0$ with $g(0) = 0$ forces $g = 0$. Index $= 1 - 0 = 1 = \chi([0, 1])$.
Relative condition ($f(0) = f(1) = 0$; no condition on g). Kernel: constant f with $f(0) = 0$ is zero. Cokernel: $g' = 0$, g constant, allowed. Index $= 0 - 1 = -1 = \chi([0, 1]) - \chi(\{0, 1\}) = 1 - 2$.
APS condition. Π projects onto $\lambda \geq 0$, which includes the zero mode, so $\Pi(f|_{\partial X}) = 0$ means $f(0) = f(1) = 0$: here APS coincides with the relative condition, and the index is -1. The APS theorem (4.233) agrees:

$$\mathrm{index}_{\mathrm{APS}}\, D \overset{(1)}{=} \int_{[0,1]} P - \frac{h_0 + \eta(L_0)}{2} - \frac{h_1 + \eta(L_1)}{2} \overset{(2)}{=} 0 - \frac{1+0}{2} - \frac{1+0}{2} = -1.$$

(1) The APS formula with one boundary term for each end point.
(2) In one dimension there is no local index density ($e(X) = 0$ in odd dimensions); each end has $L = 0$, so $h = 1$ and $\eta = 0$.

All three indices are integers built from h alone. No eta invariant enters, because a 1×1 zero matrix has a symmetric spectrum.

The boundary operator of the Euler operator on a surface

In two dimensions we can write the boundary operator out completely. Take a product collar with coordinates (n, θ), metric $\mathrm{d}n^2 + \mathrm{d}\theta^2$, orientation $\mathrm{d}n \wedge \mathrm{d}\theta$. The Euler operator maps even forms to odd forms:

$$D = \mathrm{d} + \mathrm{d}^* : \Omega^0 \oplus \Omega^2 \to \Omega^1, \qquad (f,\ g\,\mathrm{d}n \wedge \mathrm{d}\theta) \mapsto \mathrm{d}f + \mathrm{d}^*(g\,\mathrm{d}n \wedge \mathrm{d}\theta).$$

In two dimensions $\mathrm{d}^* = - \star \mathrm{d} \star$, and $\star \mathrm{d}n = \mathrm{d}\theta$, $\star \mathrm{d}\theta = -\mathrm{d}n$, $\star(\mathrm{d}n \wedge \mathrm{d}\theta) = 1$. So

$$\mathrm{d}^*(g\,\mathrm{d}n \wedge \mathrm{d}\theta) \overset{(1)}{=} - \star\, \mathrm{d}g \overset{(2)}{=} - \star \left(g_n\,\mathrm{d}n + g_\theta\,\mathrm{d}\theta\right) \overset{(3)}{=} -g_n\,\mathrm{d}\theta + g_\theta\,\mathrm{d}n,$$

(1) $\mathrm{d}^* = - \star \mathrm{d} \star$ on a surface, and $\star(g\,\mathrm{d}n \wedge \mathrm{d}\theta) = g$.
(2) Expand $\mathrm{d}g$; subscripts are partial derivatives.
(3) $\star \mathrm{d}n = \mathrm{d}\theta$ and $\star \mathrm{d}\theta = -\mathrm{d}n$.

and $D(f, g) = (f_n + g_\theta)\,\mathrm{d}n + (f_\theta - g_n)\,\mathrm{d}\theta$. Record a 1-form by its components (a, b) along $(\mathrm{d}n, \mathrm{d}\theta)$. Then

$$D \begin{pmatrix} f \\ g \end{pmatrix} \overset{(1)}{=} \begin{pmatrix} \partial_n & \partial_\theta \\ \partial_\theta & -\partial_n \end{pmatrix} \begin{pmatrix} f \\ g \end{pmatrix} \overset{(2)}{=} \underbrace{\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}}_{A} \left(\partial_n + \underbrace{\begin{pmatrix} 0 & \partial_\theta \\ -\partial_\theta & 0 \end{pmatrix}}_{L} \right) \begin{pmatrix} f \\ g \end{pmatrix}. \tag{4.239}$$

(1) Read off the components of $D(f, g)$.
(2) Factor out $A = \mathrm{diag}(1, -1)$, using $A^{-1} = A$: $A \begin{pmatrix} 0 & \partial_\theta \\ -\partial_\theta & 0 \end{pmatrix} = \begin{pmatrix} 0 & \partial_\theta \\ \partial_\theta & 0 \end{pmatrix}$.

L is self-adjoint on the circle: ∂_θ is anti-self-adjoint, and transposing the matrix while taking adjoints of the entries returns L. On a Fourier mode $e^{ik\theta}$,

$$L = \begin{pmatrix} 0 & ik \\ -ik & 0 \end{pmatrix} = -k\,\sigma_y, \qquad \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix},$$

with eigenvalues $\pm k$. So the spectrum of L is $\{\pm k :\ k \in \mathbb{Z}\}$, perfectly symmetric: $\eta_L(s) = 0$ for all s. The zero modes are the two constant vectors at $k = 0$, so $h = 2$.

Key idea: Why the de Rham boundary operator is symmetric

In (4.239) the matrix $\sigma_z = \mathrm{diag}(1, -1)$ anticommutes with L (it anticommutes with σ_y). So if $L\phi = \lambda\phi$ then $L(\sigma_z\phi) = -\lambda\,\sigma_z\phi$: every eigenvalue comes with its negative, and $\eta_L \equiv 0$. Geometrically σ_z is the *parity of the form degree on the boundary*: f restricts to a 0-form on ∂X and g corresponds to the 1-form $g\,\mathrm{d}\theta$. The same is true in every dimension: the boundary operator of $\mathrm{d} + \mathrm{d}^*$ is built from $\mathrm{d} + \mathrm{d}^*$ of ∂X, which exchanges even and odd forms, so $(-1)^{\text{degree}}$ anticommutes with it. The signature operator is different, because its bundles $\Omega^\pm$ are defined with the Hodge star, and the boundary operator has no such symmetry.

With this, the APS theorem for the Euler operator on a surface with p boundary circles and a product collar reads

$$\mathrm{index}_{\mathrm{APS}}(\mathrm{d} + \mathrm{d}^*) \overset{(1)}{=} \int_X e(X) - p\,\frac{h + \eta(0)}{2} \overset{(2)}{=} \chi(X) - p.$$

(1) (4.233) with the Euler density $P = e(X)$, one boundary term per circle.
(2) Gauss–Bonnet for a product collar (4.207); $h = 2$ and $\eta = 0$ for each circle.

The APS answer differs from the local answers $\chi(X) = \chi(X, \partial X)$ only by the integer p, which records the zero modes. There is no spectral asymmetry to correct for, and nothing depends on the geometry of the boundary. That is the content of Nash's remark.

Worked example 4.144: The signature operator on a 3-manifold boundary is asymmetric

For a 4-manifold with boundary Y, the boundary operator of d_+ is, on the even forms of Y and up to a sign convention,

$$B^{\mathrm{ev}}\phi = (-1)^p\big(\star\mathrm{d} - \mathrm{d}\star\big)\phi, \qquad \phi \in \Omega^{2p}(Y),$$

which is often called the **odd signature operator**. Its interesting part is $\star\mathrm{d}$ acting on co-exact 1-forms of Y (the other pieces pair off symmetrically). The operator $\star\mathrm{d}$ on a 3-manifold is the *curl*: for $Y = \mathbb{R}^3$ and a 1-form $\alpha = \vec{a}\cdot\mathrm{d}\vec{x}$, $\star\mathrm{d}\alpha = (\nabla\times\vec{a})\cdot\mathrm{d}\vec{x}$. Its eigenfields are Beltrami fields, $\nabla\times\vec{a} = \lambda\vec{a}$, and there is no reason for right-handed and left-handed eigenfields to come in equal numbers. For the round S^3 the spectrum *is* symmetric (an orientation-reversing isometry exists), so $\eta = 0$, consistent with $\mathrm{Sign}(B^4) = 0$. For lens spaces $L(p,q) = S^3/\mathbb{Z}_p$ APS computed (in their second paper) a finite trigonometric sum: in one common convention

$$\eta\big(L(p,q)\big) = -\frac{1}{p}\sum_{k=1}^{p-1}\cot\frac{\pi k}{p}\,\cot\frac{\pi k q}{p}.$$

For $p = 2$ ($\mathbb{RP}^3$) every term has $\cot\frac{\pi}{2} = 0$, so $\eta = 0$. For $L(3,1)$, $\cot^2\frac{\pi}{3} = \cot^2\frac{2\pi}{3} = \frac{1}{3}$ gives $\eta = -\frac{1}{3}\cdot\frac{2}{3} = -\frac{2}{9}$: a genuinely nonzero, and non-integer, spectral asymmetry. So for the signature operator the correction $-\frac{1}{2}\eta$ really is there.

Watch out: Two different reasons for "no correction"

Do not confuse the two mechanisms. The Gauss–Bonnet formula (4.207) with no boundary term uses *local* boundary conditions and a product metric: the local boundary contribution (geodesic curvature) vanishes. The APS index of $\mathrm{d}+\mathrm{d}^*$ uses the *global* condition and has the integer correction $-\frac{h}{2}$ per boundary component, but no η. In both cases the answer is independent of the metric on the boundary, which is the point.

4.21.9 Where eta comes from: the heat kernel on the collar

We have stated the APS theorem and checked it on examples. We now show where the $-\frac{1}{2}(h+\eta)$ actually comes from. The argument is the heat-kernel proof of Chapter 4 (McKean–Singer), with one new ingredient: an exactly solvable model for the collar. It is the proof APS gave, stripped to its essentials.

The strategy

Let $D = \partial_u + L$ on the collar, where $u \geq 0$ is the inward normal coordinate (we take $A = 1$; a unitary A changes nothing below). With the APS condition, the nonzero eigenvalues of D^*D and DD^* still pair up (D maps an eigenvector of D^*D in the domain to an eigenvector of DD^* in the adjoint domain). So McKean–Singer still holds:

$$\mathrm{index}\,D = \mathrm{Tr}\,e^{-tD^*D} - \mathrm{Tr}\,e^{-tDD^*} \qquad \text{for every } t > 0. \tag{4.240}$$

Now let $t \to 0$. The heat kernel only sees a neighbourhood of size $\sqrt{t}$ of each point, so the trace splits into two kinds of contribution:

- *interior points*, where the kernel is the same as on a closed manifold, and the $t \to 0$ limit of the local supertrace is the index density P (Chapter 4);

- *points within $\sqrt{t}$ of the boundary*, where the kernel feels the boundary condition. Here the collar is exactly the product $[0,\varepsilon)\times\partial X$ with $D=\partial_u+L$, so the kernel is (up to exponentially small errors) that of the half-infinite cylinder $[0,\infty)\times\partial X$ with the APS condition at $u=0$.

The first gives $\int_X P$. On the cylinder the local density vanishes, because $D^*D=-\partial_u^2+L^2=DD^*$ as differential operators there (L does not depend on u), so everything new comes from the boundary condition. We compute that exactly.

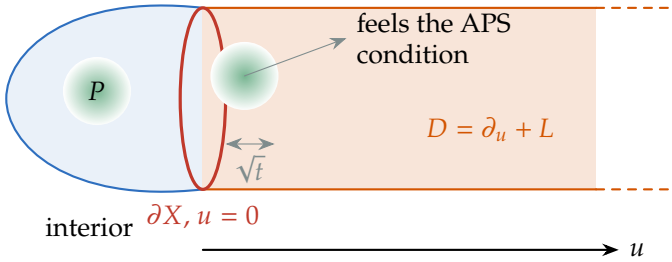


Figure 4.83: The heat-kernel proof of the APS theorem. For small t the heat kernel spreads over a distance $\sqrt{t}$. Deep inside X it cannot tell that there is a boundary, and the local supertrace tends to the index density P. Near ∂X it feels the boundary condition; there the collar is a product, and the problem reduces to one-dimensional problems $\partial_u+\lambda$ on a half-line, one for each eigenvalue λ of L.

One mode on a half-line

Expand in eigenvectors of L: $\phi(u,y)=\sum_\lambda f_\lambda(u)\psi_\lambda(y)$ with $L\psi_\lambda=\lambda\psi_\lambda$. Then D acts on each coefficient as the ordinary differential operator

$$D_\lambda=\frac{\mathrm{d}}{\mathrm{d}u}+\lambda \quad \text{on } [0,\infty), \qquad D_\lambda^*=-\frac{\mathrm{d}}{\mathrm{d}u}+\lambda.$$

The adjoint comes from an integration by parts, which also tells us what boundary condition the adjoint must carry:

$$\langle D_\lambda f,g\rangle \overset{(1)}{=} \int_0^\infty (f'+\lambda f)\bar g\,\mathrm{d}u \overset{(2)}{=} \big[f\bar g\big]_0^\infty+\int_0^\infty f\,\overline{(-g'+\lambda g)}\,\mathrm{d}u \overset{(3)}{=} -f(0)\bar g(0)+\langle f,D_\lambda^* g\rangle.$$

(1) Definition of the L^2 inner product, λ real.
(2) Integrate $f'\bar g$ by parts.
(3) Square-integrable functions vanish at $u=\infty$.

So the boundary term $f(0)\bar g(0)$ must vanish: if the condition on f fixes $f(0)=0$, then $g(0)$ is free, and if $f(0)$ is free, then $g(0)=0$.

The APS condition kills the components with $\lambda\ge 0$. Combining, the domains of $D_\lambda^*D_\lambda$ and $D_\lambda D_\lambda^*$ (both equal to $-\mathrm{d}^2/\mathrm{d}u^2+\lambda^2$ in the interior) are:

	$D_\lambda^*D_\lambda$ (acts on f)	$D_\lambda D_\lambda^*$ (acts on g)
$\lambda>0$	$f(0)=0$ (Dirichlet)	$(-g'+\lambda g)(0)=0$ (Robin, $g'=\lambda g$)
$\lambda=0$	$f(0)=0$ (Dirichlet)	$g'(0)=0$ (Neumann)
$\lambda<0$	$(f'+\lambda f)(0)=0$ (Robin, $f'=\lvert\lambda\rvert f$)	$g(0)=0$ (Dirichlet)

(The second condition in each entry is "$D_\lambda f$ lies in the domain of D_λ^*", respectively "$D_\lambda^* g$ lies in the domain of D_λ".) Define the *boundary defect* of one mode,

$$F_\lambda(t) = \operatorname{Tr} e^{-tD_\lambda^* D_\lambda} - \operatorname{Tr} e^{-tD_\lambda D_\lambda^*},$$

the difference of the two heat traces on the half-line. Neither trace is finite (the half-line is infinite), but their difference is, because far from $u = 0$ both kernels are the same free kernel.

Worked example 4.145: The defect of a zero mode is exactly $-\frac{1}{2}$

For $\lambda = 0$ we need Dirichlet minus Neumann on the half-line. The method of images gives both kernels from the free kernel $G_t(x) = (4\pi t)^{-1/2} e^{-x^2/4t}$: $K_t^{D/N}(u,u') = G_t(u-u') \mp G_t(u+u')$ (an image of opposite or equal sign at $-u'$). So

$$F_0(t) \overset{(1)}{=} \int_0^\infty \big(K_t^D(u,u) - K_t^N(u,u)\big)\,\mathrm{d}u \overset{(2)}{=} -2\int_0^\infty G_t(2u)\,\mathrm{d}u$$
$$\overset{(3)}{=} -\frac{2}{\sqrt{4\pi t}}\int_0^\infty e^{-u^2/t}\,\mathrm{d}u \overset{(4)}{=} -\frac{2}{\sqrt{4\pi t}}\cdot\frac{\sqrt{\pi t}}{2} = -\frac{1}{2}.$$

(1) A trace is the integral of the kernel on the diagonal.
(2) The direct terms $G_t(0)$ cancel; the image terms give $-G_t(2u) - G_t(2u)$.
(3) $G_t(2u) = (4\pi t)^{-1/2} e^{-4u^2/4t}$.
(4) The half Gaussian integral $\int_0^\infty e^{-u^2/t}\mathrm{d}u = \frac{1}{2}\sqrt{\pi t}$.

The answer does not depend on t. It is "half a state": the constant function is a bounded (not normalisable) solution of $D_0 g = 0$ sitting exactly on the fence, and the Neumann problem has half a state more than the Dirichlet problem.

For $\lambda \neq 0$ the Robin condition makes the image method awkward, and a physicist's tool is quicker: count states in a large box and use **phase shifts**. Take $\lambda > 0$. Scattering states of $-\partial_u^2 + \lambda^2$ with energy $k^2 + \lambda^2$ are

$$\text{Dirichlet: } \sin ku = \cos\big(ku - \tfrac{\pi}{2}\big), \qquad \text{Robin: } \cos ku + \frac{\lambda}{k}\sin ku \;\propto\; \cos\big(ku - \delta(k)\big), \quad \tan\delta(k) = \frac{\lambda}{k}.$$

(Check the Robin state: at $u = 0$ its value is 1 and its slope is λ, so $g' = \lambda g$.) Put both problems in a large box $[0, R]$ and impose, for both, that the wave function vanishes at the far end $u = R$. The Dirichlet states need $\sin kR = 0$, i.e. $kR = \pi m$; the Robin states need $\cos(kR - \delta) = 0$, i.e. $kR - \delta(k) = \pi(m - \frac{1}{2})$, with $m = 1, 2, \ldots$ in both cases. So the numbers of states with momentum below k are

$$N_D(k) \approx \frac{kR}{\pi}, \qquad N_R(k) \approx \frac{kR - \delta(k)}{\pi} + \frac{1}{2}, \qquad N_D(k) - N_R(k) \approx \frac{\delta(k)}{\pi} - \frac{1}{2}.$$

The R-dependence cancels in the difference, and differentiating in k gives the difference of the densities of states:

$$\rho_D(k) - \rho_R(k) = \frac{1}{\pi}\frac{\mathrm{d}\delta}{\mathrm{d}k} = \frac{1}{\pi}\frac{\mathrm{d}}{\mathrm{d}k}\arctan\frac{\lambda}{k} = -\frac{1}{\pi}\frac{\lambda}{k^2 + \lambda^2}. \tag{4.241}$$

This is the one-dimensional **Friedel sum rule** (or Levinson's theorem): a boundary condition acts like a scatterer, and it changes the density of states by $\frac{1}{\pi}$ times the derivative of the phase shift. Now

$$F_\lambda(t) \overset{(1)}{=} \int_0^\infty \big(\rho_D(k) - \rho_R(k)\big)e^{-t(k^2+\lambda^2)}\,\mathrm{d}k \overset{(2)}{=} -\frac{\lambda}{\pi}\int_0^\infty \frac{e^{-t(k^2+\lambda^2)}}{k^2 + \lambda^2}\,\mathrm{d}k. \tag{4.242}$$

(1) A heat trace is $\sum e^{-tE}$ over states, i.e. $\int \rho(k) e^{-tE(k)} \mathrm{d}k$. (There are no bound states: $e^{\lambda u}$ is not normalisable.)
(2) (4.241).

Two values are easy. At $t = 0$, $F_\lambda(0) = -\frac{\lambda}{\pi} \cdot \frac{\pi}{2\lambda} = -\frac{1}{2}$, using $\int_0^\infty \frac{\mathrm{d}k}{k^2+\lambda^2} = \frac{\pi}{2\lambda}$. As $t \to \infty$, $F_\lambda \to 0$. In between, differentiate:

$$\frac{\mathrm{d}F_\lambda}{\mathrm{d}t} \overset{(1)}{=} \frac{\lambda}{\pi} \int_0^\infty e^{-t(k^2+\lambda^2)}\, \mathrm{d}k \overset{(2)}{=} \frac{\lambda}{\pi} e^{-t\lambda^2} \frac{1}{2}\sqrt{\frac{\pi}{t}} = \frac{\lambda\, e^{-t\lambda^2}}{2\sqrt{\pi t}}.$$

(1) Differentiating e^{-tE} brings down $-E = -(k^2 + \lambda^2)$, which cancels the denominator.
(2) The half Gaussian integral.

The **complementary error function** $\operatorname{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-y^2} \mathrm{d}y$ has $\frac{\mathrm{d}}{\mathrm{d}t} \operatorname{erfc}(\lambda\sqrt{t}) = -\frac{\lambda e^{-\lambda^2 t}}{\sqrt{\pi t}}$ (chain rule, $\lambda > 0$). Comparing derivatives and the value at $t = \infty$ gives $F_\lambda(t) = -\frac{1}{2}\operatorname{erfc}(\lambda\sqrt{t})$ for $\lambda > 0$.

For $\lambda < 0$ the table shows that the roles are swapped: now $D_\lambda^* D_\lambda$ is the Robin problem with parameter $|\lambda|$ and $D_\lambda D_\lambda^*$ is Dirichlet. So F_λ is minus the previous answer with $\lambda \to |\lambda|$, i.e. $+\frac{1}{2}\operatorname{erfc}(|\lambda|\sqrt{t})$. All three cases are one formula:

$$F_\lambda(t) = \begin{cases} -\frac{1}{2}\operatorname{sgn}(\lambda)\, \operatorname{erfc}\big(|\lambda|\sqrt{t}\big), & \lambda \neq 0, \\ -\frac{1}{2}, & \lambda = 0. \end{cases} \tag{4.243}$$

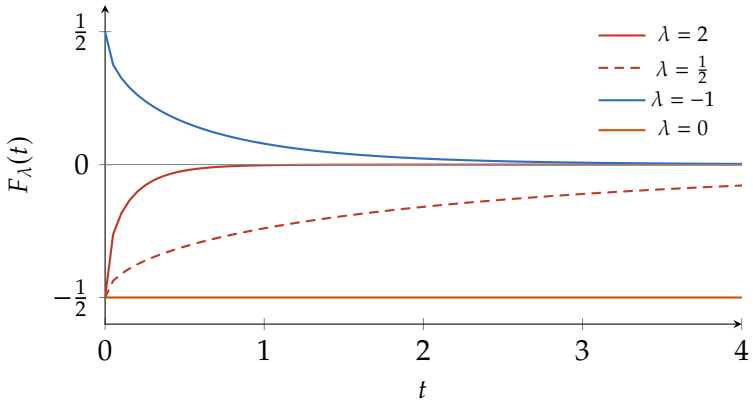


Figure 4.84: The boundary defect $F_\lambda(t)$ of a single mode of L, (4.243). At $t \to 0$ every nonzero mode contributes $-\frac{1}{2}\operatorname{sgn}\lambda$; as t grows the contribution dies off on the time scale $1/\lambda^2$, so large $|\lambda|$ (short wavelength) modes die first. A zero mode (orange) contributes $-\frac{1}{2}$ for all t. The sum over all modes at $t \to 0$ is the regularised $-\frac{1}{2}\sum \operatorname{sgn}\lambda$, i.e. $-\frac{1}{2}\eta(0)$.

Summing the modes: the eta invariant appears

Summing (4.243) over the spectrum of L gives the total boundary contribution to (4.240). For fixed $t > 0$ the sum converges, because $\operatorname{erfc}(|\lambda|\sqrt{t})$ decays like a Gaussian in λ. To take $t \to 0$, rewrite each term as an integral over the heat time:

$$\sum_{\lambda\neq 0} \operatorname{sgn}(\lambda) \operatorname{erfc}\big(|\lambda|\sqrt{t}\big) \overset{(1)}{=} \sum_{\lambda\neq 0} \operatorname{sgn}(\lambda) \int_t^\infty \frac{|\lambda|\, e^{-\lambda^2\tau}}{\sqrt{\pi\tau}}\, \mathrm{d}\tau \overset{(2)}{=} \frac{1}{\sqrt{\pi}} \int_t^\infty \tau^{-1/2} \operatorname{Tr}\big(L\, e^{-\tau L^2}\big)\, \mathrm{d}\tau$$
$$\xrightarrow{t\to 0} \frac{1}{\sqrt{\pi}} \int_0^\infty \tau^{-1/2} \operatorname{Tr}\big(L\, e^{-\tau L^2}\big)\, \mathrm{d}\tau \overset{(3)}{=} \eta_L(0). \tag{4.244}$$

(1) $\operatorname{erfc}(|\lambda|\sqrt{t}) = -\int_t^\infty \frac{\mathrm{d}}{\mathrm{d}\tau}\operatorname{erfc}(|\lambda|\sqrt{\tau})\,\mathrm{d}\tau$ and the derivative computed above; $\operatorname{erfc}(\infty) = 0$.
(2) $\operatorname{sgn}(\lambda)|\lambda| = \lambda$, and $\sum_\lambda \lambda e^{-\lambda^2\tau} = \operatorname{Tr}(Le^{-\tau L^2})$ (zero modes contribute nothing).
(3) The heat-kernel formula (4.222) for $\eta(0)$, valid because the integral converges at $\tau = 0$ (Section 4.21.12).

Putting the pieces into (4.240):

$$\begin{aligned}\operatorname{index} D &\overset{(1)}{=} \lim_{t\to 0}\left(\operatorname{Tr} e^{-tD^*D} - \operatorname{Tr} e^{-tDD^*}\right) \overset{(2)}{=} \int_X P + \lim_{t\to 0}\sum_\lambda F_\lambda(t) \\ &\overset{(3)}{=} \int_X P - \frac{h}{2} - \frac{1}{2}\lim_{t\to 0}\sum_{\lambda\neq 0}\operatorname{sgn}(\lambda)\operatorname{erfc}\left(|\lambda|\sqrt{t}\right) \overset{(4)}{=} \int_X P - \frac{h+\eta_L(0)}{2}.\end{aligned} \tag{4.245}$$

(1) McKean–Singer (4.240): the right side does not depend on t, so we may take any limit.
(2) Interior points give the local index density; the collar gives the half-line defects, one for each eigenvalue of L (errors are $O(e^{-c/t})$).
(3) (4.243): each of the h zero modes gives $-\frac{1}{2}$.
(4) (4.244).

This is the APS theorem (4.233).

Physical picture: What the proof says in words

Each eigenmode of the boundary operator is a little one-dimensional scattering problem. Positive and negative modes are scattered by the APS condition in opposite ways and change the "number of states" of D^*D relative to DD^* by $\mp\frac{1}{2}$. If the spectrum were symmetric, these half-states would cancel. The eta invariant is precisely the net, regularised count of the uncancelled half-states, and the zero modes each leave one half-state behind. The index, being an honest integer count, is the bulk density plus this boundary bookkeeping.

Watch out: What is and is not proved here

The sketch hides two technical points. First, one must show that the APS problem is Fredholm and that (4.240) holds; this is where the spectral projection is essential (a local condition would not make the half-line problems decouple mode by mode). Second, the claim that the interior and collar contributions simply add, with exponentially small error, uses a patching argument (Duhamel's principle). Both are standard and are carried out in APS I; nothing in them changes the answer.

Try it: Why $\mp\frac{1}{2}$ at short times

Explain without any integrals why $F_\lambda(t) \to -\frac{1}{2}\operatorname{sgn}\lambda$ as $t \to 0$. *Answer:* at short times the heat kernel explores only a distance $\sqrt{t}$ from the wall, over which the Robin condition $g' = \lambda g$ is indistinguishable from $g' = 0$ (the term λg is smaller than $g' \sim g/\sqrt{t}$ by a factor $|\lambda|\sqrt{t}$). So for $\lambda > 0$ the defect is Dirichlet minus Neumann, which is $-\frac{1}{2}$ by Example 4.145, and for $\lambda < 0$ it is Neumann minus Dirichlet, $+\frac{1}{2}$. Only the *sign* of λ decides which side of the pair carries the Dirichlet condition.

4.21.10 Spectral flow

The sawtooth of Figure 4.78 showed that η has two kinds of dependence on a parameter: a smooth drift, and jumps by 2 when an eigenvalue crosses zero. The jumps are integers and are

therefore topological. Counting them is so useful that it has its own name. This is not in Nash's Chapter IV, but it is the form in which the APS theorem is most often used in physics (and it is behind the families picture of Section 4.18 for self-adjoint operators).

Working definition 4.64 (spectral flow)

Let L_u, $0 \le u \le 1$, be a continuous family of self-adjoint elliptic operators. Their eigenvalues are continuous functions of u (draw them as curves). The **spectral flow** $\mathrm{SF}(L_u)$ is the net number of eigenvalues that cross zero as u goes from 0 to 1:

$$\mathrm{SF} = \#\{\text{crossings from } \lambda < 0 \text{ to } \lambda \ge 0\} - \#\{\text{crossings from } \lambda \ge 0 \text{ to } \lambda < 0\}.$$

Operationally: draw the spectral curves, and count signed intersections with the horizontal axis. We count an eigenvalue that reaches 0 from below as having crossed (the convention matching ξ [$\xi(L) = \frac{1}{2}\big(h(L) + \eta_L(0)\big)$, with $h(L) = \dim \ker L$] below).

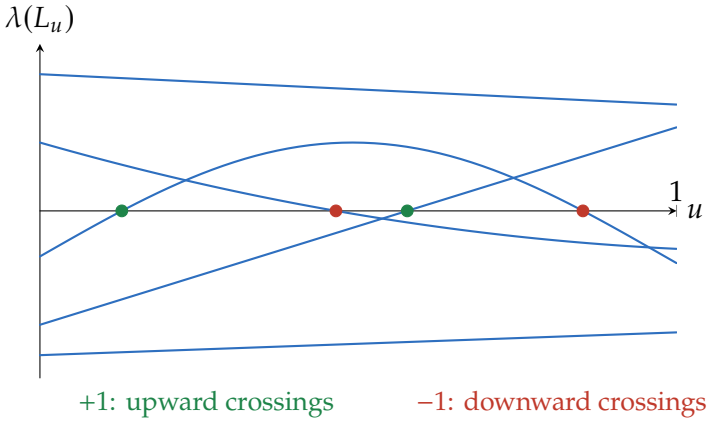


Figure 4.85: Spectral flow is a signed count of zero crossings. Each curve is an eigenvalue of L_u. Green dots: an eigenvalue crosses from negative to positive (+1); red dots: from positive to negative (−1). Here the net flow is $2 - 2 = 0$. Eigenvalues that stay on one side contribute nothing, however much they move; that motion is what the smooth part of η records.

How η changes along a family

Write $\xi(L) = \frac{1}{2}\big(h(L) + \eta_L(0)\big)$ as before. When an eigenvalue moves from negative to 0, η loses a -1 and h gains 1, so ξ jumps by $+\frac{1}{2} + \frac{1}{2} = 1$; as it continues to positive values, h loses 1 and η gains +1, and ξ is unchanged. So ξ jumps by exactly +1 at an upward crossing and -1 at a downward one, and otherwise changes smoothly:

$$\xi(L_1) - \xi(L_0) = \mathrm{SF}(L_u) + \int_0^1 \frac{\mathrm{d}\xi}{\mathrm{d}u}\Big|_{\text{smooth}} \mathrm{d}u. \tag{4.246}$$

The smooth part is *local*. To see this, differentiate the eta function with respect to u at a value of u with no crossing. For an eigenvalue $\lambda(u) \neq 0$, first-order perturbation theory gives $\dot\lambda = \langle \psi_\lambda, \dot L \psi_\lambda \rangle$ (dots are u-derivatives), and

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}u}\eta_{L_u}(s) &\overset{(1)}{=} \sum_\lambda \operatorname{sgn}(\lambda)\,\frac{\mathrm{d}}{\mathrm{d}u}|\lambda|^{-s} \overset{(2)}{=} \sum_\lambda \operatorname{sgn}(\lambda)\,(-s)|\lambda|^{-s-1}\operatorname{sgn}(\lambda)\dot\lambda \\ &\overset{(3)}{=} -s\sum_\lambda \langle \psi_\lambda, \dot L \psi_\lambda\rangle |\lambda|^{-s-1} \overset{(4)}{=} -s\ \operatorname{Tr}\big(\dot L\,|L|^{-s-1}\big). \end{aligned} \tag{4.247}$$

(1) Definition of $\eta(s)$; $\operatorname{sgn}\lambda$ is constant away from crossings.
(2) $\frac{\mathrm{d}}{\mathrm{d}u}|\lambda| = \operatorname{sgn}(\lambda)\dot\lambda$.
(3) $\operatorname{sgn}^2 = 1$, and the Hellmann–Feynman formula for $\dot\lambda$.
(4) The sum of diagonal matrix elements in an eigenbasis is the trace.

At $s = 0$ the explicit factor s kills everything except a *pole* of $\operatorname{Tr}(\dot L|L|^{-s-1})$ at $s = 0$. Poles of such traces come from the small-time expansion of heat traces, which is local. Concretely, $\operatorname{Tr}(\dot L|L|^{-s-1}) = \frac{1}{\Gamma(\frac{s+1}{2})}\int_0^\infty \tau^{\frac{s-1}{2}}\operatorname{Tr}(\dot L e^{-\tau L^2})\mathrm{d}\tau$ (the Mellin transform again), and if $\operatorname{Tr}(\dot L e^{-\tau L^2}) \sim \cdots + b\,\tau^{-1/2} + \dots$ as $\tau \to 0$, the b term gives $\frac{b}{\sqrt\pi}\int_0^1 \tau^{\frac{s}{2}-1}\mathrm{d}\tau = \frac{2b}{\sqrt\pi\, s}$. So

$$\left.\frac{\mathrm{d}\eta_{L_u}(0)}{\mathrm{d}u}\right|_{\text{smooth}} = -\frac{2}{\sqrt\pi}\,b(u), \qquad b(u) = \text{coefficient of } \tau^{-1/2} \text{ in } \operatorname{Tr}\big(\dot L_u e^{-\tau L_u^2}\big), \tag{4.248}$$

and $b(u)$ is the integral over the boundary of a local density built from $\dot L$ and the symbol of L.

Worked example 4.146: Checking (4.248) on the circle

For $L_u = i\,\mathrm{d}/\mathrm{d}\theta + t(u)$ we have $\dot L = \dot t$ (a constant), and the heat trace was computed by Poisson summation in Example 4.136: $\operatorname{Tr} e^{-\tau L^2} = \sum_n e^{-\tau(n+t)^2} = \sqrt{\pi/\tau} + (\text{exponentially small})$. So $\operatorname{Tr}(\dot L e^{-\tau L^2}) = \dot t\sqrt\pi\,\tau^{-1/2} + \dots$, i.e. $b = \dot t\sqrt\pi$, and

$$\left.\frac{\mathrm{d}\eta}{\mathrm{d}u}\right|_{\text{smooth}} = -\frac{2}{\sqrt\pi}\,\dot t\sqrt\pi = -2\dot t.$$

This is the slope -2 of the sawtooth $1-2t$, now obtained with no knowledge of the Hurwitz zeta function. The same slope follows from (4.247) directly: $\operatorname{Tr}|L|^{-s-1} = \zeta(s+1,t) + \zeta(s+1,1-t)$ has a simple pole of residue 2 at $s=0$ (each Hurwitz zeta function has residue 1 at $s+1=1$), so $-s\cdot\dot t\cdot\frac{2}{s} = -2\dot t$.

The index on a cylinder is the spectral flow

Now apply the APS theorem to the cylinder $X = [0,1]\times Y$ with

$$D = \partial_u + L_u,$$

where L_u is constant near $u=0$ and near $u=1$ (so that both ends are product collars). At $u=0$ the inward normal is ∂_u and the boundary operator is L_0. At $u=1$ the inward normal is $-\partial_u$, $D = -(\partial_n - L_1)$, and the boundary operator is $-L_1$. So

$$\begin{aligned}\operatorname{index} D &\overset{(1)}{=} \int_X P - \xi(L_0) - \xi(-L_1) \overset{(2)}{=} \int_X P - \xi(L_0) - \big(h_1 - \xi(L_1)\big)\\ &\overset{(3)}{=} \int_X P + \big(\xi(L_1) - \xi(L_0)\big) - h_1 \overset{(4)}{=} \int_X P + \mathrm{SF} + \int_0^1 \dot\xi_{\text{smooth}}\,\mathrm{d}u - h_1.\end{aligned} \tag{4.249}$$

(1) (4.233) with two boundary components, written with $\xi = \frac12(h+\eta)$.
(2) $\xi(-L) = \frac12(h-\eta_L) = h - \xi(L)$, since $-L$ has the same kernel and the opposite asymmetry.
(3) Rearrange.
(4) (4.246).

Both $\int_X P$ and $\int \dot\xi_{\text{smooth}}$ are integrals of local densities along the family. To compare them, take a family with *no* crossings and L_0, L_1 invertible. Then $\operatorname{index} D = 0$: on each eigenmode the

solution of $Du = 0$ is $f(u) = f(0)\exp(-\int_0^u \lambda)$, and the APS conditions demand $\lambda(0) < 0$ at one end and $-\lambda(1) < 0$ at the other, which is a crossing; similarly for the cokernel. So (4.249) gives $\int_X P = -\int_0^1 \dot{\xi}_{\text{smooth}}\,\mathrm{d}u$ for such families, and since both sides are local this identity holds for all families. Hence

Key idea: Index = spectral flow (APS III)

For $D = \partial_u + L_u$ on $[0,1]\times Y$ with APS conditions,

$$\operatorname{index} D = \operatorname{SF}(L_u) - h(L_1), \tag{4.250}$$

and in particular $\operatorname{index} D = \operatorname{SF}(L_u)$ if L_1 is invertible. The bulk integral and the smooth drift of η cancel identically; what remains is purely the integer count of zero crossings.

Worked example 4.147: A cylinder with a growing magnetic flux

Let $Y = S^1$ and $L_u = i\,\mathrm{d}/\mathrm{d}\theta + Nu$ with N a positive integer (a flux that grows by N quanta along the cylinder). On the Fourier mode with eigenvalue $\lambda_n(u) = n + Nu$ the operator is $\partial_u + \lambda_n(u)$. *Direct count.* The kernel has one solution $f_n = \exp\bigl(-\int_0^u \lambda_n\bigr)$ per n with $\lambda_n(0) < 0 < \lambda_n(1)$, i.e. $n < 0 < n + N$: $n = -N+1, \dots, -1$, giving $N-1$ solutions. The cokernel needs $\lambda_n(1) \le 0 \le \lambda_n(0)$, i.e. $n + N \le 0 \le n$, impossible. So $\operatorname{index} D = N - 1$. *Spectral flow.* The lines $n + Nu$ that go from < 0 to ≥ 0 are $n = -N, \dots, -1$: $\operatorname{SF} = N$. The end $L_1 = i\,\mathrm{d}/\mathrm{d}\theta + N$ has the zero mode $n = -N$, so $h_1 = 1$ and (4.250) gives $N - 1$ ✓. *The full APS formula.* The bulk density is the Chern form of the flux: $\int_X P = \int_0^1 \dot{t}\,\mathrm{d}u = N$, which is minus the smooth drift $\int \dot{\xi}_{\text{smooth}} = \frac{1}{2}\int(-2\dot{t}) = -N$ ✓; both ends have integer flux, so $\eta = 0$ and $h = 1$, and $\operatorname{index} D = N - \frac{1}{2} - \frac{1}{2} = N - 1$ ✓.

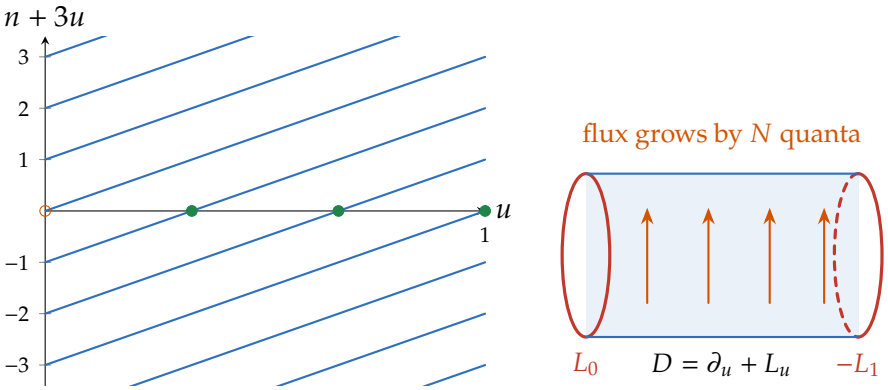


Figure 4.86: Left: the spectrum of $L_u = i\,\mathrm{d}/\mathrm{d}\theta + 3u$ on the circle. Each Fourier mode is a line of slope 3; three lines cross zero upwards (green), so SF = 3 (the one ending exactly on the axis at $u = 1$ counts, and is also the zero mode $h_1 = 1$ of L_1; the orange circle marks the zero mode of L_0, which does not cross). Right: the corresponding cylinder, on which $\operatorname{index}(\partial_u + L_u) = \operatorname{SF} - h_1 = 2$ by Example 4.147.

Physical picture: Closing the cylinder into a torus

If L_1 is gauge equivalent to L_0 (here: integer N, since $t \to t+1$ is a large gauge transformation), the ends of the cylinder can be glued with the gauge transformation, and D becomes $\bar{\partial}$ (up to normalisation) coupled to a line bundle of degree N on the torus T^2. There is no boundary, and the ordinary index theorem (Riemann–Roch with $g = 1$) gives

index $= N + 1 - 1 = N =$ SF around the loop. For a loop of self-adjoint operators the spectral flow is the "index" of the family, and it plays for self-adjoint families the role that the index bundle plays in Section 4.18. It is exactly the "one eigenvalue crosses per period" of Figure 4.60.

Spectral flow in physics: the chiral anomaly by level crossing

Spectral flow is the cleanest way to see an anomaly. The QFT notes do the following argument for two-dimensional QED [QFT notes §22.2.4, pp. 1374–1376]; we rephrase it in our language.

Put a massless Dirac fermion on a spatial circle of length 2π, with a constant, slowly varying gauge potential $a(\tau)$ along the circle (physicists' Hermitian convention; the anti-Hermitian $A = -ia\,\mathrm{d}\theta$). The one-particle Hamiltonians of the two chiralities are

$$H_\pm = \mp i\,\partial_\theta \mp a(\tau) \quad \text{with eigenvalues} \quad E_{\pm,n} = \pm(n - a),\ n \in \mathbb{Z},$$

on the modes $e^{in\theta}$. As a increases by one unit (one flux quantum, $\oint a\,\mathrm{d}\theta$ changes by 2π), every level of H_+ moves down by one and every level of H_- moves up by one: $\mathrm{SF}(H_+) = -1$ and $\mathrm{SF}(H_-) = +1$. The spectrum returns to itself, but one filled level of H_- has risen out of the Dirac sea, and one empty level of H_+ has sunk into it. The charges therefore change by

$$\Delta N_+ = \mathrm{SF}(H_+) = -1, \qquad \Delta N_- = \mathrm{SF}(H_-) = +1, \qquad \Delta(N_+ - N_-) = -2,$$

while the total charge $N_+ + N_-$ is conserved. This is the integrated chiral anomaly, the QFT notes' (22.2.48). The anomaly is literally the spectral flow of the Dirac Hamiltonian, and by (4.250) it is also the index of the Euclidean Dirac operator $\partial_\tau + H$ on the space-time cylinder, which is the Chern number of the gauge field on the time-slab. In the path-integral language of Chapter 10 this index is the number of zero modes that make the fermion measure non-invariant (Fujikawa).

Dictionary: spectral flow

$\mathrm{SF}(L_u)$	net number of levels crossing zero energy
$\mathrm{index}(\partial_u + L_u) = \mathrm{SF}$	number of fermions created from the Dirac sea during an adiabatic process (Euclidean time u)
smooth drift of η	polarisation of the Dirac sea: charge that moves continuously, not in units
closed loop of L_u	periodic adiabatic process; SF is a topological invariant of the loop (a "Thouless pump")

4.21.11 Relaxing the product metric: the second fundamental form

Everything so far assumed that near ∂X the manifold looks exactly like a cylinder $\partial X \times [0, \varepsilon)$ (4.201). That is convenient but unnatural: the flat unit disc, the round hemisphere and a ball in $\mathbb{R}^4$ are not products near their boundaries. Nash ends the section [Nash pp. 135–136, (4.175)–(4.177)] by explaining what changes when we drop the assumption. The answer is simple to state:

> *the index formula acquires one more term, an integral over ∂X of a local expression in the second fundamental form of the boundary.*

We have already met the prototype. In Gauss–Bonnet (4.206) the extra term is $\frac{1}{2\pi}\oint k_g \, \mathrm{d}s$, and the geodesic curvature k_g is the second fundamental form of a curve. Our job is to see where such a term comes from in general, and to compute it in examples.

Coordinates adapted to the boundary

Nash writes the metric near the boundary as [Nash (4.175)]

$$\mathrm{d}s^2 = g_{11}\,\mathrm{d}n^2 + \sum_{i,j\geq 2} h_{ij}\,\mathrm{d}x^i\mathrm{d}x^j, \tag{4.251}$$

where n is a normal coordinate and x^i, $i = 2, \dots, \dim X$, are coordinates on ∂X. We can always make $g_{11} = 1$ by choosing n to be the geodesic distance from the boundary (walk inwards along the geodesic that leaves ∂X at right angles, and let n be the arc length). Then the metric is

$$\mathrm{d}s^2 = \mathrm{d}n^2 + h_{ij}(n, x)\,\mathrm{d}x^i\mathrm{d}x^j,$$

and the only difference from a product is that the metric h_{ij} of the slices $n = \text{const}$ *depends on* n. How fast it changes as we move off the boundary is measured by one tensor.

Working definition 4.65 (second fundamental form)

In the coordinates above, with n the inward normal distance, the **second fundamental form** of ∂X is the symmetric tensor on ∂X

$$K_{ij} = -\frac{1}{2}\frac{\partial h_{ij}}{\partial n}\Big|_{n=0}.$$

Operationally: slide the boundary inwards by a distance δn and see how lengths on it change; $K_{ij}\,\delta n$ is (minus) the change of the metric. The sign is chosen so that $K > 0$ when the boundary is convex, i.e. when it shrinks as we move inwards, as for the rim of a disc. Its trace $H = h^{ij}K_{ij}$ [h^{ij} is the inverse matrix of h_{ij}] is the **mean curvature**. For a curve on a surface, K is a 1×1 matrix, and $h^{\theta\theta}K_{\theta\theta}$ is the geodesic curvature k_g [how fast the curve turns within the surface, e.g. $1/r$ for a circle of radius r in the plane].

Nash describes the same object differently: as the *normal component* of the Levi-Civita connection on the boundary. The two descriptions agree. Use the standard formula $\Gamma^a{}_{bc} = \frac{1}{2}g^{ad}(\partial_b g_{dc} + \partial_c g_{db} - \partial_d g_{bc})$ with $g_{nn} = 1$, $g_{ni} = 0$, $g_{ij} = h_{ij}$:

$$\Gamma^n{}_{ij} \overset{(1)}{=} \frac{1}{2}g^{nn}\big(\partial_i g_{nj} + \partial_j g_{ni} - \partial_n g_{ij}\big) \overset{(2)}{=} -\frac{1}{2}\,\partial_n h_{ij} \overset{(3)}{=} K_{ij} \quad \text{at } n = 0,$$

$$\Gamma^i{}_{nj} \overset{(4)}{=} \frac{1}{2}h^{ik}\big(\partial_n g_{kj} + \partial_j g_{kn} - \partial_k g_{nj}\big) \overset{(5)}{=} \frac{1}{2}h^{ik}\partial_n h_{kj} \overset{(6)}{=} -K^i{}_j \quad \text{at } n = 0.$$

(1) The Christoffel formula; g^{nd} is nonzero only for $d = n$.
(2) $g_{nj} = 0$ identically, so its derivatives vanish.
(3) Definition of K_{ij}.
(4) The Christoffel formula; g^{id} is nonzero only for $d = k$ tangential.
(5) Again $g_{kn} = g_{nj} = 0$.
(6) Raise an index with h^{ik}.

All the other mixed components ($\Gamma^n{}_{nn}$, $\Gamma^n{}_{ni}$, $\Gamma^i{}_{nn}$) vanish because $g_{nn} = 1$ is constant. So the parts of the connection that mix the normal direction with the tangential ones are *exactly* $\pm K$. Nash's statement "the metric is a product near the boundary, so Γ has no normal component" is the statement $K = 0$ (for all n in the collar).

Worked example 4.148: Surfaces of revolution: the geodesic curvature from K_{ij}

Take a disc with the rotationally symmetric metric

$$\mathrm{d}s^2 = \mathrm{d}r^2 + f(r)^2\,\mathrm{d}\theta^2, \qquad 0 \le r \le R,$$

with $f(r) \approx r$ near $r = 0$ (smooth at the centre). The inward normal distance from the rim is $n = R - r$, so $h_{\theta\theta} = f(R-n)^2$ and

$$K_{\theta\theta} \overset{(1)}{=} -\frac{1}{2}\frac{\partial}{\partial n} f(R-n)^2\Big|_{n=0} \overset{(2)}{=} f(R)f'(R), \qquad k_g \overset{(3)}{=} h^{\theta\theta}K_{\theta\theta} \overset{(4)}{=} \frac{f'(R)}{f(R)}.$$

(1) Definition of K.
(2) Chain rule: $\partial_n f(R-n)^2 = -2ff'$.
(3) The geodesic curvature is the trace of K on a curve.
(4) $h^{\theta\theta} = 1/f(R)^2$.

Three cases, all familiar from Example 4.132:

- *Flat disc*, $f = r$: $k_g = 1/R$, the curvature of a circle of radius R.
- *Spherical cap* on the unit sphere, $f = \sin r$, $R = \vartheta_0$: $k_g = \cot\vartheta_0$. It vanishes at the equator $\vartheta_0 = \pi/2$, which is a great circle (a geodesic).
- *Test tube*: f constant for r near R (a hemisphere glued to a straight cylinder). Then $f'(R) = 0$, so $K = 0$: this is a product collar.

Why the extra term is a boundary integral: transgression

Here is the mechanism in one sentence: *characteristic forms of two different connections differ by an exact form*. Nash quotes it as [Nash (4.176)]: if Γ and Γ' are two connections with curvatures R and R', and P is an invariant polynomial, then

$$P(R) - P(R') = \mathrm{d}Q. \tag{4.252}$$

The form Q is called the **transgression form** (or secondary characteristic form). Integrating over X and using Stokes,

$$\int_X P(R) = \int_X P(R') + \int_{\partial X} Q. \tag{4.253}$$

On a closed manifold the last term is absent, which is the familiar statement that characteristic numbers do not depend on the connection (Chapter 4). With a boundary it is the only thing that changes.

Key idea: The transgression formula, derived

Join the two connections by the straight line $A_t = A' + t\,a$, $a = A - A'$, $0 \le t \le 1$, and write $F_t = \mathrm{d}A_t + A_t \wedge A_t$. Take $P(F) = \operatorname{tr} F^k$ (every invariant polynomial is built from these). Then

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\operatorname{tr}F_t^k &\overset{(1)}{=} k\,\operatorname{tr}\big(\dot F_t \wedge F_t^{k-1}\big) \overset{(2)}{=} k\,\operatorname{tr}\big((\mathrm{d}a + a\wedge A_t + A_t\wedge a)\wedge F_t^{k-1}\big)\\ &\overset{(3)}{=} k\,\operatorname{tr}\big(D_t a \wedge F_t^{k-1}\big) \overset{(4)}{=} k\,\mathrm{d}\,\operatorname{tr}\big(a\wedge F_t^{k-1}\big).\end{aligned} \tag{4.254}$$

(1) Product rule; the cyclicity of the trace (with the signs for even forms, which are all +) lets us collect the k terms.

(2) $\dot{A}_t = a$, so $\dot{F}_t = \mathrm{d}a + a \wedge A_t + A_t \wedge a$.
(3) This combination is the covariant derivative $D_t a$ of the adjoint-valued 1-form a (Chapter T3).
(4) For any adjoint-valued form ω, $\mathrm{d}\,\mathrm{tr}\,\omega = \mathrm{tr}\, D_t\omega$; and the Bianchi identity $D_t F_t = 0$ lets D_t act on a alone.

Integrating from $t = 0$ to $t = 1$ gives (4.252) with the explicit

$$Q = k \int_0^1 \mathrm{tr}\big(a \wedge F_t^{k-1}\big)\,\mathrm{d}t, \qquad a = A - A'. \tag{4.255}$$

Q is built out of the difference a of the connections and the curvatures; it is local and gauge covariant.

Worked example 4.149: The Chern–Simons form

Take $k = 2$, $A' = 0$ (on a region where the bundle is trivialised), so $a = A$, $A_t = tA$ and $F_t = t\,\mathrm{d}A + t^2 A \wedge A$. Then

$$Q \overset{(1)}{=} 2\int_0^1 \mathrm{tr}\big(A \wedge (t\,\mathrm{d}A + t^2 A \wedge A)\big)\mathrm{d}t \overset{(2)}{=} 2\Big(\frac{1}{2}\,\mathrm{tr}(A \wedge \mathrm{d}A) + \frac{1}{3}\,\mathrm{tr}(A \wedge A \wedge A)\Big)$$
$$\overset{(3)}{=} \mathrm{tr}\Big(A \wedge \mathrm{d}A + \frac{2}{3} A \wedge A \wedge A\Big).$$

(1) (4.255) with $k = 2$.
(2) $\int_0^1 t\,\mathrm{d}t = \frac{1}{2}$ and $\int_0^1 t^2 \mathrm{d}t = \frac{1}{3}$.
(3) Multiply out.

This is the **Chern–Simons form**, and we can check $\mathrm{d}Q = \mathrm{tr}\, F \wedge F$ directly: $\mathrm{d}\,\mathrm{tr}(A\,\mathrm{d}A) = \mathrm{tr}(\mathrm{d}A\,\mathrm{d}A)$, and $\mathrm{d}\,\mathrm{tr}(A^3) = 3\,\mathrm{tr}(\mathrm{d}A\,A^2)$, while $\mathrm{tr}\,F^2 = \mathrm{tr}(\mathrm{d}A\,\mathrm{d}A) + 2\,\mathrm{tr}(\mathrm{d}A\,A^2) + \mathrm{tr}(A^4)$ with $\mathrm{tr}\,A^4 = 0$ (a cyclic permutation of four 1-forms costs a sign). So $\mathrm{d}Q = \mathrm{tr}(\mathrm{d}A)^2 + \frac{2}{3}\cdot 3\,\mathrm{tr}(\mathrm{d}A\,A^2) = \mathrm{tr}\,F^2$ ✓. For an abelian field $A = -ia$ it is $Q = -a \wedge \mathrm{d}a$. The same form will be the action of Chern–Simons theory in Chapter 12, and the anomaly polynomial's descendant in Chapter 10.

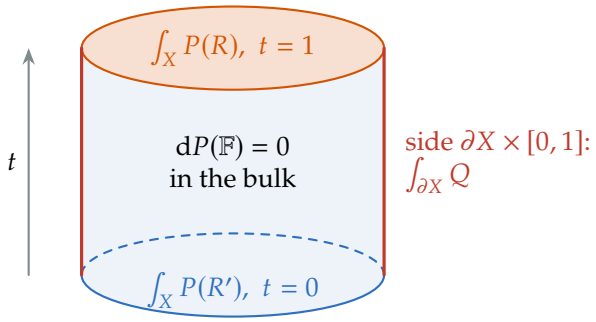


Figure 4.87: Transgression as Stokes' theorem. Put the interpolating connection A_t on the cylinder $X \times [0,1]$; its characteristic form $P(\mathbb{F})$ is closed. Stokes says the integrals over the whole boundary of the cylinder add to zero: top minus bottom equals the side. On a closed X there is no side, and $\int_X P$ is independent of the connection. With $\partial X \neq \varnothing$ the side contributes $\int_{\partial X} Q$, where Q is $P(\mathbb{F})$ integrated along t (integration along the fibre, as in Section 4.18).

The general index formula

Now apply (4.253) to the Levi-Civita connections. Let g be any metric on X, and let g' be a metric that agrees with g away from a thin collar and is a product near ∂X (deform $h_{ij}(n)$ to $h_{ij}(0)$ in the collar). For g' the APS theorem holds as stated. The boundary operator L only sees the metric *on* ∂X, which is the same for both. Hence

$$\operatorname{index} D \overset{(1)}{=} \int_X P(R') - \frac{h+\eta(0)}{2} \overset{(2)}{=} \int_X P(R) - \int_{\partial X} Q - \frac{h+\eta(0)}{2}. \tag{4.256}$$

(1) The APS theorem (4.233) for the product metric g'. The index itself does not change under the deformation: with the APS condition fixed, a continuous deformation of the operator keeps it Fredholm and cannot change the integer.

(2) (4.253) with R the curvature of g and R' that of g'.

In the collar the two connections differ by $a = \Gamma - \Gamma'$, whose only nonzero components at the boundary are the normal–tangential ones we computed, $\pm K_{ij}$ (for g' they are zero). So Q restricted to ∂X is a polynomial in the second fundamental form and the curvature of ∂X: this is Nash's "appropriate Q expressed in terms of the second fundamental form". Writing the boundary term with the opposite sign gives Nash's [Nash (4.177)],

$$\operatorname{index} D = \int_X P(R) + \int_{\partial X} \widetilde{Q} - \frac{h+\eta(0)}{2}, \qquad \widetilde{Q} = -Q\big|_{\partial X}. \tag{4.257}$$

Worked example 4.150: Gauss–Bonnet from the transgression

We re-derive Gauss–Bonnet with boundary from (4.253), for the metric $\mathrm{d}r^2 + f^2\mathrm{d}\theta^2$ of Example 4.148. Use the orthonormal coframe $e^1 = \mathrm{d}r$, $e^2 = f\,\mathrm{d}\theta$ (Chapter T2). The first structure equation $\mathrm{d}e^a = -\omega^a{}_b \wedge e^b$ gives the connection 1-form:

$$\mathrm{d}e^2 \overset{(1)}{=} f'\,\mathrm{d}r \wedge \mathrm{d}\theta \overset{(2)}{=} -\omega^2{}_1 \wedge \mathrm{d}r \quad\Longrightarrow\quad \omega^2{}_1 \overset{(3)}{=} f'\,\mathrm{d}\theta.$$

(1) Differentiate $e^2 = f(r)\,\mathrm{d}\theta$.

(2) The structure equation for $a = 2$ ($\omega^2{}_2 = 0$).

(3) $-f'\mathrm{d}\theta \wedge \mathrm{d}r = f'\mathrm{d}r \wedge \mathrm{d}\theta$ ✓; also $\mathrm{d}e^1 = 0 = -\omega^1{}_2 \wedge e^2$ holds because $\mathrm{d}\theta \wedge \mathrm{d}\theta = 0$.

The curvature is $\Omega^2{}_1 = \mathrm{d}\omega^2{}_1 = f''\,\mathrm{d}r \wedge \mathrm{d}\theta$, and comparing with $\Omega^2{}_1 = -K_{\mathrm{G}}\, e^1 \wedge e^2$ gives the Gaussian curvature $K_{\mathrm{G}} = -f''/f$ (= 1 for $f = \sin r$). The Euler form is

$$e(g) = \frac{1}{2\pi} K_{\mathrm{G}}\,\mathrm{d}A = -\frac{1}{2\pi} f''\,\mathrm{d}r \wedge \mathrm{d}\theta = \mathrm{d}\Big(-\frac{f'}{2\pi}\mathrm{d}\theta\Big).$$

Let g' have the profile $\tilde{f}$, equal to f near $r = 0$ and constant near $r = R$. Then

$$e(g) - e(g') \overset{(1)}{=} \mathrm{d}Q, \qquad Q = -\frac{f' - \tilde{f}'}{2\pi}\,\mathrm{d}\theta,$$
$$\int_X e(g) \overset{(2)}{=} \int_X e(g') + \oint_{r=R} Q \overset{(3)}{=} \chi(X) - f'(R) \overset{(4)}{=} \chi(X) - \frac{1}{2\pi}\oint k_g\,\mathrm{d}s.$$

(1) Subtract the two Euler forms. Q is a smooth 1-form on the whole disc, because $f' - \tilde{f}'$ vanishes near the centre, where $\mathrm{d}\theta$ is singular.

(2) Stokes, (4.253).

(3) Gauss–Bonnet for the product metric (4.207); $\tilde{f}'(R) = 0$ and $\oint \mathrm{d}\theta = 2\pi$.

(4) $k_g = f'(R)/f(R)$ from Example 4.148, and the rim has length $2\pi f(R)$.

This is Gauss–Bonnet with boundary (4.206). The boundary term $\widetilde{Q} = -Q = \frac{1}{2\pi}k_g\,\mathrm{d}s$ is linear in the second fundamental form, as promised.

Worked example 4.151: $\bar\partial$ on a flat disc versus a test tube

Consider $\bar\partial$ on functions on the unit disc with APS conditions. We computed in Example 4.141 (with $\alpha = 0$) that the index is 0 for a metric with a product collar. Now take the *flat* metric $\mathrm{d}x^2 + \mathrm{d}y^2$, which is not a product.

The boundary operator. With $n = 1 - r$, so $\partial_r = -\partial_n$,

$$\partial_{\bar z} = \frac{e^{i\theta}}{2}\Big(\partial_r + \frac{i}{r}\partial_\theta\Big) = -\frac{e^{i\theta}}{2}\Big(\partial_n - \frac{i}{1-n}\partial_\theta\Big).$$

The tangential operator now depends on n. At $n = 0$ it is $L = -i\partial_\theta$, with eigenvalue k on $e^{ik\theta}$: spectrum $\mathbb{Z}$, $h = 1$, $\eta = 0$, as before.

Direct count. A holomorphic $f = \sum_{k\ge 0} a_k z^k$ has boundary values $\sum_{k\ge 0} a_k e^{ik\theta}$. The APS condition removes all modes with $k \ge 0$, so Ker = 0. The index is conformally invariant data here (the kernel does not depend on the conformal factor, and neither does the boundary condition), so it is still 0.

The formula. The index density is the degree-2 part of $\mathrm{td}\, X$, $\frac12 c_1(TX) = \frac12 e$, so Q is half the Gauss–Bonnet transgression:

$$\operatorname{index}\bar\partial \overset{(1)}{=} \frac12\int_X e(g_{\mathrm{flat}}) + \frac12\cdot\frac{1}{2\pi}\oint k_g\,\mathrm{d}s - \frac{h+\eta(0)}{2} \overset{(2)}{=} 0 + \frac12\cdot 1 - \frac{1+0}{2} = 0\ \checkmark$$

(1) (4.257) with $P = \frac12 e$ and $\widetilde Q = \frac12\cdot\frac{1}{2\pi}k_g\,\mathrm{d}s$ (Example 4.150).
(2) The flat disc has $K_{\mathrm G} = 0$; its rim has $k_g = 1$ and length 2π.

For the test tube all of the $\frac12$ sits in the bulk (the curvature of the hemispherical cap) and $\widetilde Q = 0$. For a spherical cap of angle ϑ_0 it is split as $\frac12(1-\cos\vartheta_0) + \frac12\cos\vartheta_0$. In each case the total of the *local* terms is $\frac12$, and the boundary spectral term $-\frac12$ is unchanged.

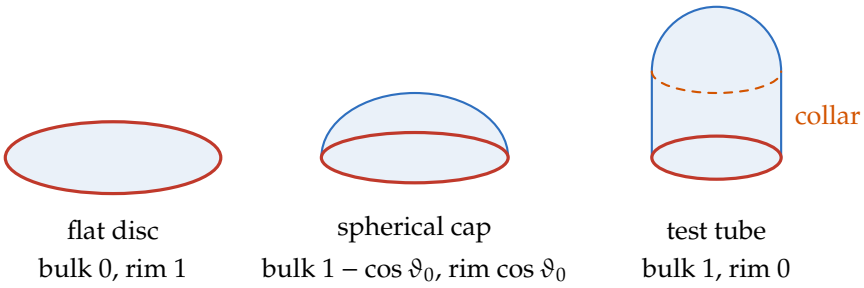


Figure 4.88: Three metrics on the disc, and where the Euler characteristic $\chi = 1$ is stored. For the flat disc all of it is in the geodesic curvature of the rim (a boundary transgression term); for the test tube, whose collar is a product, all of it is in the bulk curvature of the cap. The index of $\bar\partial$ with APS conditions is 0 in every case: $\frac12\chi$ from the local terms, $-\frac12$ from the zero mode of the boundary circle.

Watch out: The sign of Q in Nash's (4.177), and what L means

Whether Q enters with a plus or a minus depends only on which metric is called R and which R' in (4.176); Nash avoids the issue by saying "an appropriate Q". In our conventions (4.256), the correction is $-\int_{\partial X} Q$ with Q the transgression from the product metric to the actual one. A second subtlety: when the metric is not a product, $D = A_n(\partial_n + L_n)$ with A_n and L_n both depending on n, and there is freedom to move zeroth-order terms between them. The APS condition uses the boundary operator $L = L_0$ defined intrinsically by the

geometry of ∂X (for Dirac operators, the Dirac operator of ∂X). Any other choice shifts η by a local amount, which is compensated by a change in $\widetilde{Q}$.

Physical picture: Why a product collar is the natural normalisation

A product collar is precisely the situation in which the bulk index density "does not know" where the boundary is: the geometry near ∂X is translation invariant in n, so nothing local can accumulate there. Every local boundary effect then lives in the transgression term, and every non-local one in η. APS chose this normalisation because it separates the two kinds of correction cleanly; a physicist would call $\widetilde{Q}$ a *surface counterterm*.

4.21.12 Is $\eta(0)$ always finite?

Our treatment of η has quietly assumed that $\eta_L(s)$, continued from the half-plane where its series converges, is finite at $s = 0$. For the zeta function this was automatic: the factor $1/\Gamma(s)$ in (4.215) vanishes at $s = 0$ and kills the only possible pole. For eta the corresponding prefactor is $1/\Gamma(\frac{s+1}{2})$, which is $1/\sqrt{\pi}$ at $s = 0$, finite and nonzero. Nothing kills a pole. Nash raises the question at the very end of the chapter [Nash p. 136]:

> *If we take any self-adjoint elliptic operator B on a closed manifold Y and form $\eta_B(s)$, is $\eta_B(0)$ finite?*

His answer has three parts. For the boundary operators L of the APS theorem, yes: the index theorem itself proves it, because every other term in (4.233) is finite. For general B it is also yes, if $\dim Y$ is odd, or if $\dim Y$ is even and B is a differential operator of odd order; but the proof is delicate. And the delicacy is visible: the *operator* $B|B|^{-s-1}$ can have a pole at $s = 0$ with a nonzero matrix-valued residue whose *trace* is zero. We now make each of these statements concrete.

Where a pole at $s = 0$ would come from

Let L be a first-order self-adjoint elliptic operator on a closed manifold Y of dimension m. As $t \to 0$, the odd heat trace has an asymptotic expansion in powers of $\sqrt{t}$, like the ordinary heat trace (4.216):

$$\operatorname{Tr}\big(L\, e^{-tL^2}\big) \sim \sum_{k\geq 0} c_k\, t^{\frac{k-m-1}{2}}, \qquad c_k = \int_Y \gamma_k(y)\, \mathrm{d}^m y, \tag{4.258}$$

where each $\gamma_k(y)$ is a *local* density: a polynomial in the coefficients of L and their derivatives at y (with k derivatives in total, counting a coefficient of order zero as one derivative). The power is fixed by dimensional analysis: L has dimension $1/\text{length}$, t has dimension length^2, the diagonal of a kernel on Y has dimension length^{-m}, so the leading term is $\text{length}^{-m-1} \sim t^{-\frac{m+1}{2}}$, and each extra derivative contributes another $t^{1/2}$.

Put (4.258) into the heat-kernel formula (4.221). Split the t integral at $t = 1$: the piece $\int_1^\infty$ is an entire function of s (the integrand decays exponentially if we remove the zero modes, which contribute nothing to $\operatorname{Tr}(Le^{-tL^2})$ anyway). The piece $\int_0^1$ gives, term by term,

$$\frac{1}{\Gamma(\frac{s+1}{2})}\int_0^1 t^{\frac{s-1}{2}}\, c_k\, t^{\frac{k-m-1}{2}}\, \mathrm{d}t \overset{(1)}{=} \frac{c_k}{\Gamma(\frac{s+1}{2})}\int_0^1 t^{\frac{s+k-m}{2}-1}\, \mathrm{d}t \overset{(2)}{=} \frac{c_k}{\Gamma(\frac{s+1}{2})}\cdot\frac{2}{s+k-m}. \tag{4.259}$$

(1) Add the exponents: $\frac{s-1}{2} + \frac{k-m-1}{2} = \frac{s+k-m}{2} - 1$.

(2) $\int_0^1 t^{a-1}\mathrm{d}t = 1/a$, continued to all $a \neq 0$.

So $\eta_L(s)$ has (at most) simple poles at $s = m-k$, $k = 0,1,2,\dots$, with residue $2c_k/\Gamma(\frac{m-k+1}{2})$. (When $\frac{m-k+1}{2}$ is a non-positive integer, $1/\Gamma$ vanishes and that pole is absent.) The point $s = 0$ corresponds to $k = m$:

Key idea: The finiteness criterion

$$\eta_L(s) \text{ is regular at } s = 0 \iff c_m = 0, \qquad \operatorname*{Res}_{s=0} \eta_L = \frac{2c_m}{\sqrt{\pi}},$$

where c_m is the coefficient of $t^{-1/2}$ in $\mathrm{Tr}(Le^{-tL^2})$. The residue is the integral over Y of a local density $\gamma_m(y)$. So the question "is $\eta(0)$ finite?" is a question about one local heat coefficient.

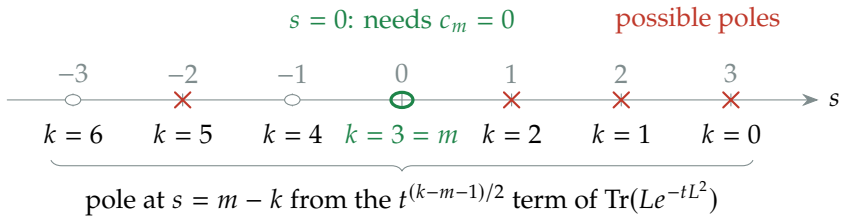


Figure 4.89: The ladder of possible poles of $\eta_L(s)$ for $\dim Y = m$ (drawn for $m = 3$). Each term $c_k t^{(k-m-1)/2}$ in the small-t expansion of the odd heat trace can produce a simple pole at $s = m-k$. The point $s = 0$ is hit by exactly one term, $k = m$, the coefficient of $t^{-1/2}$. (Grey circles: at $s = -1, -3, \dots$ the prefactor $1/\Gamma(\frac{s+1}{2})$ vanishes and cancels the would-be pole.)

Even dimensions and differential operators: a parity argument

The first mechanism that makes c_m vanish is a symmetry of momentum space. Start with the simplest case: constant coefficients on a flat torus T^m, $L = \gamma^j p_j$ with $p_j = -i\partial_j$ and Hermitian matrices γ^j satisfying $\gamma^j\gamma^k + \gamma^k\gamma^j = 2\delta^{jk}$ (Chapter T4). Then $L^2 = p^2$ and, up to exponentially small corrections (Poisson summation, as in Example 4.136), the trace is a momentum integral:

$$\mathrm{Tr}\big(L\,e^{-tL^2}\big) \overset{(1)}{=} \mathrm{Vol}(T^m) \int \frac{\mathrm{d}^m p}{(2\pi)^m}\, \mathrm{tr}\big(\gamma^j p_j\big) e^{-tp^2} \overset{(2)}{=} 0.$$

(1) The trace over plane waves $e^{ip\cdot y}$ becomes a momentum integral; tr is the matrix trace.
(2) The integrand is odd under $p \to -p$ (and also $\mathrm{tr}\,\gamma^j = 0$).

The general statement is a refinement of step (2). For a *differential* operator the full symbol is a polynomial in p, and its homogeneous pieces have definite parity: the order-1 piece is odd in p, the order-0 piece even. When we build the local heat coefficient $\gamma_k(y)$ from the symbol, every term is a momentum integral whose integrand has parity $(-1)^{k+1}$ under $p \to -p$. (Each derivative of a coefficient trades one power of p for one power of ∂_y, flipping the parity; the leading term, $k = 0$, is odd as we just saw.) So

$$\gamma_k(y) = 0 \quad \text{pointwise, for every even } k.$$

If m is even, c_m has $k = m$ even and vanishes identically: $\eta_L(0)$ is finite. The same argument works for any differential operator of odd order. It fails for pseudo-differential operators, whose symbols need not be polynomials, and it fails for m odd, where $k = m$ is odd and the integrand is even.

Worked example 4.152: The first local coefficient, and a cancellation in one dimension

Add a constant Hermitian matrix potential: $L = \gamma^j p_j + V$ on T^m. We compute the term linear in V of the local odd heat trace (this is the $k = 1$ coefficient, since V counts as one derivative). Since p^2 is a scalar matrix it commutes with everything, and $L^2 = p^2 + \{\gamma \cdot p, V\} + V^2$, so to first order $e^{-tL^2} \approx e^{-tp^2}\big(1 - t\{\gamma \cdot p, V\}\big)$. Then, per unit volume,

$$\gamma_1 \overset{(1)}{=} \int \frac{\mathrm{d}^m p}{(2\pi)^m}\, e^{-tp^2}\ \mathrm{tr}\Big[(\gamma \cdot p + V)\big(1 - t\{\gamma \cdot p, V\}\big)\Big]_{\text{linear in } V}$$
$$\overset{(2)}{=} \int \frac{\mathrm{d}^m p}{(2\pi)^m}\, e^{-tp^2}\big(\mathrm{tr}\, V - 2t\, p^2\ \mathrm{tr}\, V\big) \overset{(3)}{=} \mathrm{tr}\, V\,(4\pi t)^{-m/2}\Big(1 - 2t \cdot \frac{m}{2t}\Big) \overset{(4)}{=} (1-m)\,\frac{\mathrm{tr}\, V}{(4\pi t)^{m/2}}.$$

(1) The trace over plane waves, keeping first order in V.
(2) $\mathrm{tr}(\gamma \cdot p\{\gamma \cdot p, V\}) = 2\,\mathrm{tr}\big((\gamma \cdot p)^2 V\big) = 2p^2\,\mathrm{tr}\, V$ by cyclicity and $(\gamma \cdot p)^2 = p^2$.
(3) Gaussian integrals: $\int \frac{\mathrm{d}^m p}{(2\pi)^m} e^{-tp^2} = (4\pi t)^{-m/2}$ and $\int \frac{\mathrm{d}^m p}{(2\pi)^m} p^2 e^{-tp^2} = \frac{m}{2t}(4\pi t)^{-m/2}$.
(4) Arithmetic.

This is $c_1\, t^{(1-m-1)/2} = c_1 t^{-m/2}$, as (4.258) predicts. It produces a pole of η at $s = m - 1$ with residue proportional to $(1 - m)\,\mathrm{tr}\, V$. For $m = 3$ that is a genuine pole at $s = 2$, harmless for $\eta(0)$. For $m = 1$ the pole would sit exactly at $s = 0$; and there the factor $1 - m$ *vanishes*. This is our circle operator $i\,\mathrm{d}/\mathrm{d}\theta + t$ in disguise, whose odd heat trace we found to be exponentially small (Example 4.137). The cancellation is between the explicit V and the V hidden in e^{-tL^2}: a small instance of the kind of cancellation that makes $\eta(0)$ finite in general.

Odd dimensions: the residue is a matrix, its trace is zero

In odd dimensions the parity argument gives nothing, and the local density $\gamma_m(y)$ need not vanish. Nash's last remark describes what actually happens. Think of $\eta_L(s) = \mathrm{Tr}\big(L|L|^{-s-1}\big)$ as the trace of an operator-valued function of s. Its Schwartz kernel on the diagonal, $K_s(y, y)$, is a *matrix* (an endomorphism of the fibre at y), and it continues meromorphically in s. APS found examples where $K_s(y, y)$ has a pole at $s = 0$ with a nonzero matrix-valued residue $\mathcal{R}(y)$. But the residue of η is $\int_Y \mathrm{tr}\,\mathcal{R}(y)\,\mathrm{d}^m y$, and this turned out to be zero:

$$\mathcal{R}(y) \neq 0, \qquad \text{but} \qquad \int_Y \mathrm{tr}\,\mathcal{R}(y)\,\mathrm{d}^m y = 0.$$

The argument for the vanishing is not a local cancellation. It goes as follows:

1. The residue $\mathrm{Res}_{s=0}\,\eta_L$ is the integral of a local density, so it varies smoothly (in fact, polynomially in the derivatives) under deformations of L.
2. Its variation is computed by (4.247): the derivative of $\eta_L(s)$ is $-s\,\mathrm{Tr}(\dot{L}|L|^{-s-1})$, which has at most a simple pole times s at $s = 0$, i.e. it is *finite*. So $\frac{\mathrm{d}}{\mathrm{d}u}\,\mathrm{Res}_{s=0}\,\eta_{L_u} = 0$: the residue is a *homotopy invariant* of L.
3. A homotopy invariant that is additive and local must be a topological invariant of the symbol, in K-theory. One then checks enough examples (for which $\eta(0)$ is known to be finite, e.g. boundary operators, by the APS theorem) to see that it vanishes on a set of generators.

This is the proof, due to APS for odd dimensions and completed by Gilkey in general. It uses the index theorem, which is why Nash says that the finiteness of $\eta(0)$ is implied by its validity.

Try it: Pole or no pole?

For which of these is $\eta(0)$ finite by the parity argument alone? (a) The Dirac operator on T^2. (b) The odd signature operator on a 3-manifold. (c) A first-order differential operator on S^1. (d) The pseudo-differential operator $|D|^{1/2}\operatorname{sgn} D$ on T^2. *Answer:* (a) yes ($m = 2$ even, order 1). (b) no, $m = 3$ is odd; it is finite by the APS theorem, since it is a boundary operator. (c) no, $m = 1$ is odd; but Example 4.152 and the reparametrisation $\mathrm{d}\phi = \mathrm{d}\theta/a(\theta)$, which turns any such operator into $i\,\mathrm{d}/\mathrm{d}\phi + \text{const}$ up to a unitary transformation and a rescaling, show it is finite. (d) no, it is not a differential operator.

4.21.13 The eta invariant in physics

We promised at the start of this section that the eta invariant is not only a correction term in an index formula. Here are the three places a physicist meets it. In each case L is a Dirac-type operator (a one-particle Hamiltonian, or a Euclidean Dirac operator in odd dimension), and η measures the asymmetry between its positive and negative spectrum.

The fermion number of the Dirac sea

Second-quantise a Dirac field whose one-particle Hamiltonian H has discrete spectrum $\{\lambda\}$ with no zero modes. The vacuum fills every negative level. With the charge-conjugation symmetric definition of the charge operator, $Q = \frac{1}{2}\sum_\lambda [\psi_\lambda^\dagger, \psi_\lambda]$, the vacuum value is

$$\langle Q\rangle_{\mathrm{vac}} \overset{(1)}{=} \frac{1}{2}\Big(\sum_{\lambda<0} 1 - \sum_{\lambda>0} 1\Big) \overset{(2)}{=} -\frac{1}{2}\sum_{\lambda} \operatorname{sgn}\lambda \overset{(3)}{=} -\frac{1}{2}\,\eta_H(0). \tag{4.260}$$

(1) $\frac{1}{2}[\psi^\dagger,\psi] = \psi^\dagger\psi - \frac{1}{2}$ has expectation $+\frac{1}{2}$ on a filled level and $-\frac{1}{2}$ on an empty one.
(2) Collect the terms: a filled (negative) level has $\operatorname{sgn}\lambda = -1$.
(3) The divergent sum is regularised by the eta function, the only regularisation that respects the symmetry $\lambda \to -\lambda$ and locality.

(If H has h zero modes, each can be filled or empty, and the charge is $-\frac{1}{2}\eta_H(0) \pm \frac{1}{2}$ per zero mode.) So the vacuum charge is minus half the spectral asymmetry. It is *fractional* whenever η is.

Worked example 4.153: Fractional charge on a ring with flux

A single chiral fermion on a ring threaded by flux t (in units of the flux quantum) has $H = i\,\mathrm{d}/\mathrm{d}\theta + t$, eigenvalues $n + t$. By the circle computation (4.226), $\eta_H(0) = 1 - 2t$ for $0 < t < 1$, so

$$\langle Q\rangle_{\mathrm{vac}} = -\frac{1}{2}(1-2t) = t - \frac{1}{2}.$$

As the flux increases from 0 to 1, the vacuum charge increases continuously from $-\frac{1}{2}$ to $+\frac{1}{2}$. At $t = 1$ a level crosses zero, the old vacuum becomes an excited state with charge $+\frac{1}{2}$, and the new vacuum has charge $-\frac{1}{2}$ again. The charge transported by one flux quantum is the continuous drift $\int_0^1 \frac{\mathrm{d}\langle Q\rangle}{\mathrm{d}t}\mathrm{d}t = 1$, which equals the spectral flow: this is the chiral anomaly of Section 4.21.10 seen from the vacuum's point of view. For the Jackiw–Rebbi soliton of Example 4.140, charge-conjugation symmetry makes the continuum part of η vanish and the single zero mode gives charge $\pm\frac{1}{2}$: the famous half-integer fermion number of a soliton (Jackiw–Rebbi; Goldstone–Wilczek for the general formula $\langle Q\rangle = -\frac{1}{2}\eta$).

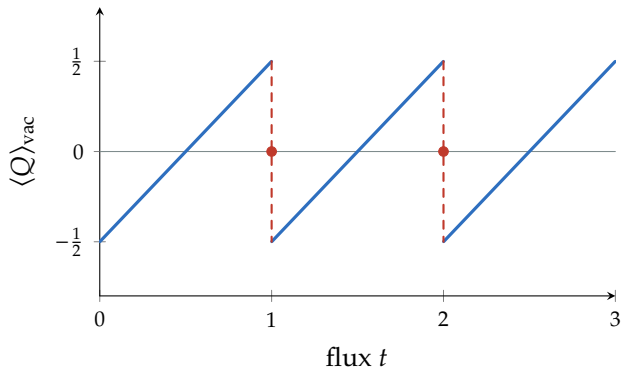


Figure 4.90: The vacuum charge $-\frac{1}{2}\eta$ of a chiral fermion on a ring, as a function of the magnetic flux. It drifts up continuously by one unit per flux quantum, the charge pumped by the anomaly, and drops by one each time a level crosses zero (red dots, where $h = 1$ and the charge is 0 on the symmetric average). The sawtooth is $-\frac{1}{2}$ times the eta sawtooth of Figure 4.78.

The phase of a fermion determinant: the parity anomaly

In odd space-time dimension n, the Euclidean Dirac operator $i\not{D}_A$ is self-adjoint and there is no chirality to split it. The fermion path integral gives $\det(i\not{D}_A)$, a product of real eigenvalues of both signs, and it must be regularised. A gauge-invariant regularisation (for instance Pauli–Villars with a heavy fermion of mass $M \to \infty$) gives

$$\det{}_{\text{reg}}(i\not{D}_A) = \left|\det(i\not{D}_A)\right| \exp\!\left(\mp \frac{i\pi}{2}\, \eta_{i\not{D}_A}(0)\right). \tag{4.261}$$

The reason is simple. The modulus of the determinant is the (regularised) product of the $|\lambda|$. Its sign is $(-1)^{N_-}$, where N_- is the (infinite) number of negative eigenvalues. Write $(-1)^{N_-} = e^{\pm i\pi N_-}$ and split $N_- = \frac{1}{2}(N_+ + N_-) - \frac{1}{2}(N_+ - N_-)$. The first half is a constant that a local counterterm can absorb; the second is $-\frac{1}{2}\sum_\lambda \operatorname{sgn}\lambda$, regularised to $-\frac{1}{2}\eta(0)$. So the phase is $e^{\mp \frac{i\pi}{2}\eta(0)}$ up to a constant, and which sign occurs is fixed by the sign of the regulator mass.

Now compare two gauge fields in $n = 3$. The variation of η is local (Section 4.21.10), and the APS theorem on a 4-manifold X with $\partial X = Y$ identifies it: for a U(1) field with the Hermitian potential a and field strength $f = \mathrm{d}a$ (so $F = -if$), and ignoring the gravitational part,

$$\begin{aligned} \frac{\eta(0)}{2} &\overset{(1)}{=} \int_X \mathrm{ch}_2(F) - \operatorname{index} D_X - \frac{h}{2} \overset{(2)}{=} \frac{1}{8\pi^2}\int_X f \wedge f \quad (\mathrm{mod}\ \mathbb{Z}\ \text{and}\ h) \\ &\overset{(3)}{=} \frac{1}{8\pi^2}\int_Y a \wedge \mathrm{d}a \quad (\mathrm{mod}\ \mathbb{Z}\ \text{and}\ h). \end{aligned}$$

(1) The APS theorem (4.233) for the Dirac operator on X, solved for η (the $\hat{A}$ part omitted).
(2) $\mathrm{ch}_2 = \frac{1}{2}\left(\frac{iF}{2\pi}\right)^2$ with $F = -if$, $f = \mathrm{d}a$; the index is an integer.
(3) $f \wedge f = \mathrm{d}(a \wedge \mathrm{d}a)$ and Stokes; this is Example 4.149 for U(1).

So the phase in (4.261) is $\exp\!\left(\mp\frac{i}{2}\cdot\frac{1}{4\pi}\int_Y a\wedge \mathrm{d}a\right)$ up to integer-level pieces: a single Dirac fermion in three dimensions induces a **Chern–Simons term at level $\frac{1}{2}$**. A level-$\frac{1}{2}$ Chern–Simons term is not gauge invariant under large gauge transformations (only integer levels are), and this clash between gauge invariance and parity is the **parity anomaly** (Redlich; Niemi–Semenoff; Alvarez-Gaumé, Della Pietra and Moore). It is the reason the surface of a topological insulator carries a half-quantised Hall conductance $\sigma_{xy} = \frac{1}{2}\frac{e^2}{h}$.

Global anomalies: eta of a mapping torus

The determinant line bundle of Section 4.18 carries a natural connection (Quillen, Bismut–Freed). Its curvature is the degree-2 part of the families index formula, the local anomaly. Its *holonomy* around a loop in the space of gauge fields is a global quantity, and Witten and Bismut–Freed showed that it is an eta invariant: if the loop of gauge fields A_u on a $2k$-manifold X is glued into a gauge field on the $(2k+1)$-dimensional mapping torus $X \times [0,1]/\sim$, then

$$\text{holonomy of DET} = \lim_{\text{adiabatic}} \exp\bigl(-i\pi\,\eta(0)\bigr),$$

with η that of the odd-dimensional Dirac operator on the mapping torus. A nontrivial holonomy is a **global anomaly**; Witten's SU(2) anomaly (an odd number of Weyl doublets is inconsistent) is the case where the relevant η is 1 modulo 2, and it is equivalent to the mod 2 index of Section 4.19. These are developed in Chapter 10.

Dictionary: the eta invariant

$\eta_H(0)$	$-2\times$ vacuum charge of the Dirac sea
jump of η by 2	a level crossing zero; one unit of spectral flow
smooth drift of η	charge pumped continuously by the anomaly
$\exp(-\frac{i\pi}{2}\eta)$	phase of the odd-dimensional fermion determinant (parity anomaly, Chern–Simons at level $\frac{1}{2}$)
$\exp(-i\pi\eta)$ on a mapping torus	holonomy of the determinant line bundle (global anomaly)
$-\frac{1}{2}(h+\eta)$ in APS	the boundary's share of the index; zero modes count as half-states

Reading Nash (pp. 131–136, (4.156)–(4.177) and footnote 5)

After identifying the signature defect with $\eta(0)$, Nash defines the zeta function of a positive operator (4.156), notes that $\zeta_P(0)$ is finite, and derives η from two zeta functions of $L_\pm = \frac{3}{2}|L| \pm \frac{1}{2}L$ (4.157)–(4.158). He computes $\eta(0) = 1 - 2t$ for $L = i\,\mathrm{d}/\mathrm{d}\theta + t$ on the circle via the Hurwitz zeta function (4.159)–(4.161). He then states the APS theorem for a general first-order $D = A(\partial_n + L)$ with global boundary conditions (4.162)–(4.164), remarks that the Euler characteristic needs no correction, and works out $\bar\partial$ on Σ_g minus p discs, finding $1 - g - p$ (4.165)–(4.174). The last page relaxes the product metric (4.175)–(4.177), with a transgression term Q built from the second fundamental form, and discusses the finiteness of $\eta_B(0)$ for general self-adjoint elliptic B. Footnote 5 mentions the heat-equation proof. We added the Mellin/heat-kernel continuation of ζ and η with explicit pole structure; two further computations of the circle η (smooth cutoff and Poisson summation); the domain wall (Jackiw–Rebbi) and flux-disc examples; the explicit cylinder check of Nash's surface example; the absolute/relative/APS comparison for the Euler operator and the symmetry that forces $\eta = 0$ there; lens-space values of η; the half-line heat-kernel derivation of $-\frac{1}{2}(h+\eta)$; spectral flow and index = spectral flow; the explicit Christoffel symbols, transgression and Chern–Simons forms, and the flat-disc check of (4.177); the pole ladder and parity argument for finiteness; and the physics of η (fractional charge, parity anomaly, global anomalies).

Recap: Section 7

- Zeta functions $\zeta_P(s) = \sum \mu^{-s} = \operatorname{Tr} P^{-s}$ are defined by analytic continuation through the Mellin transform of the heat trace; $\zeta_P(0)$ is always finite. Eta is

$\eta_L(s) = \sum \operatorname{sgn}\lambda |\lambda|^{-s} = \operatorname{Tr}(L|L|^{-s-1})$.

- On the circle, $L = i\,\mathrm{d}/\mathrm{d}\theta + t$ has $\eta(0) = 1 - 2t$ for $0 < t < 1$: a sawtooth with slope -2 (local) and jumps of 2 (level crossings).
- APS: $\operatorname{index} D = \int_X P - \frac{1}{2}(h + \eta(0))$ for $D = A(\partial_n + L)$ with the spectral boundary condition. Examples: domain wall (Jackiw–Rebbi), flux disc ($\lfloor \alpha \rfloor + 1$), $\bar\partial$ on Σ_g minus p discs $(1 - g - p)$.
- Mechanism: each mode of L is a half-line scattering problem contributing $-\frac{1}{2}\operatorname{sgn}\lambda \operatorname{erfc}(|\lambda|\sqrt{t})$ to the supertrace; the zero modes give $-\frac{1}{2}$ each.
- The de Rham complex needs no η (its boundary spectrum is symmetric); the signature operator does.
- Spectral flow: $\operatorname{index}(\partial_u + L_u) = \mathrm{SF}(L_u) - h(L_1)$ on a cylinder; it is the level-crossing picture of the chiral anomaly.
- Non-product metrics add $\int_{\partial X} \widetilde{Q}$, a transgression (Chern–Simons-type) form built from the second fundamental form.
- $\eta(0)$ is finite: by parity in even dimensions, by a global (index-theoretic) argument in odd dimensions, where the local residue is a nonzero matrix of zero trace.
- Physics: vacuum charge $-\frac{1}{2}\eta$, fermion determinant phase $e^{\mp i\pi\eta/2}$ (parity anomaly), global anomalies as $e^{-i\pi\eta}$.

Chapter summary

This chapter answered one question and then asked it again in six new settings. The question: *how many zero modes does an elliptic operator have, net of the zero modes of its adjoint?* The answer (Atiyah–Singer) is that this integer, the index, depends only on the leading symbol up to deformation, and it can be computed by integrating a polynomial in the curvatures over the manifold. The rest of the chapter kept the question and changed the setting.

1. *One operator on a closed manifold* (Nash §1). The index is defined analytically ($\dim\operatorname{Ker} D - \dim\operatorname{Ker} D^*$), repackaged as a map $\operatorname{index}_a : K(T^*X) \to \mathbb{Z}$ on symbol classes, and compared with a topological map index_t built from an embedding and the Thom isomorphism. The theorem is $\operatorname{index}_a = \operatorname{index}_t$. The heat kernel (McKean–Singer) gives a second road to the same integer, and shows why it is a *local* integral.

2. *The classical complexes* (Nash §2). Applying the Chern character turns index_t into $\int_X(\text{characteristic classes})$. For the de Rham, Dolbeault, signature and spin complexes this gives Gauss–Bonnet ($\chi = \int e$), Riemann–Roch ($\chi(X, \mathcal{O}) = \int \mathrm{td}$), Hirzebruch ($\mathrm{Sign} = \int L$) and the $\hat{A}$-genus ($\operatorname{index} \not{D} = \int \hat{A}$).

3. *Coupling to a gauge field* (Nash §3, Section 4.17). Tensor every bundle with a gauge bundle F. The integrand is multiplied by $\mathrm{ch}(F)$. This is the formula that counts fermion zero modes in monopole and instanton backgrounds.

4. *Families* (Nash §4, Section 4.18). Let the operator depend on parameters $y \in Y$. The index becomes a virtual bundle over Y, and its Chern classes (computed by integrating along the fibre) measure how the zero modes twist over Y. The determinant line bundle, where the fermion determinant lives, is its top exterior power; its curvature is the perturbative anomaly.

5. *Real families* (Nash §5, Section 4.19). For real operators the index lives in $KR(Y)$. Torsion appears, and with it invariants such as $\dim\operatorname{Ker} D \bmod 2$ that no curvature integral can see. These are behind global anomalies and Majorana zero modes.

6. *Symmetry* (Nash §6, Section 4.20). If a group G acts, the kernel and cokernel are representations of G and the index is a character. Evaluated at g it is a Lefschetz number, and it is a sum of contributions from the fixed points of g alone (Atiyah–Bott). The Weyl character formula is an example.

7. *Boundaries* (Nash §7, Section 4.21). The boundary condition that keeps the index topological is non-local: keep only the negative spectrum of the boundary operator L. The price is a correction $-\frac{1}{2}(h + \eta)$, where η measures the asymmetry of the spectrum of L. Its jumps are spectral flow; its phase is the parity anomaly.

Figure 4.91 draws the logic. After it come a formula sheet, one example that runs through several of the theorems at once, a picture of the four spectral mechanisms of the chapter, the concept tables, and a guide to which tool answers which physics question.

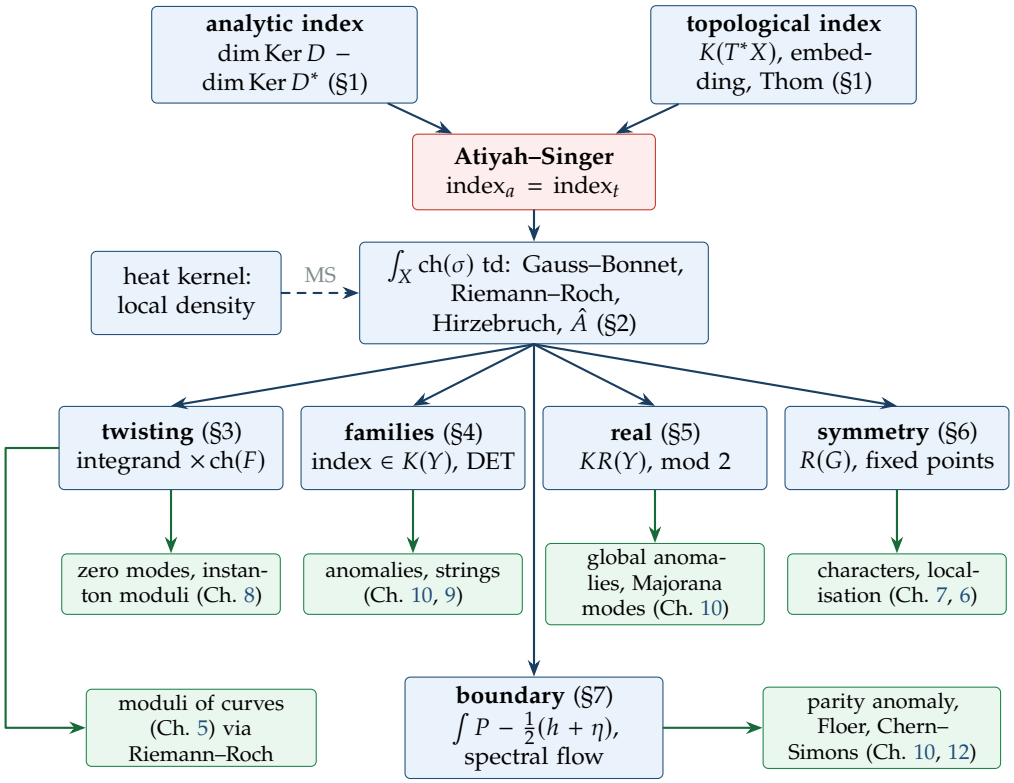


Figure 4.91: Chapter 4 in one picture. The red box is the theorem of §1, reached from the analytic side and the topological side. The cohomological formula (§2) turns it into integrals of curvature; the heat kernel (McKean–Singer, "MS") explains why the integrand is local. The four blue boxes in the middle row are the generalisations of §3–§6; each changes *where the index lives* ($\mathbb{Z}$, $K(Y)$, $KR(Y)$, $R(G)$). The boundary theorem (§7) uses the same local density plus a spectral correction; its proof is the heat kernel on a collar, and its jumps are the spectral flow of a family. Green boxes: where the later chapters use each piece.

The formula sheet

Every index formula of the chapter is the same recipe. *Write the integrand in Chern roots, keep the part of form degree $n = \dim X$, and integrate.* The master formulas of §1 and §2 are

$$\operatorname{index} D = (-1)^n \big\{ \operatorname{ch}[\sigma(D)]\ \operatorname{td}(TX_{\mathbb{C}}) \big\}[TX], \qquad \operatorname{index} D = (-1)^{n/2} \int_X \frac{\operatorname{ch}\big(\sum_p (-1)^p E^p\big)}{e(X)}\ \operatorname{td}(TX_{\mathbb{C}}), \tag{4.262}$$

the first for any elliptic operator (Nash (4.31)), the second for an elliptic complex $E^\bullet$ on an oriented even-dimensional X (Nash (4.48)). The quotient by $e(X)$ is a formal cancellation in the Chern roots. Worked out for the classical complexes, and then twisted by a gauge bundle F of rank r, they give the table below. Here x_i are the Chern roots of TX (roots $\pm x_i$ for the complexification of a real 2ℓ-manifold), $c_1 = c_1(TX)$ for a complex manifold, and $p_1 = \sum x_i^2$.

Complex	Index	Integrand (twisted by F)	Low dimensions
de Rham d_F	$r\,\chi(X)$	$\operatorname{ch}(F)\,e(X) \to r\,e(X)$ (only the rank survives)	$n=2$: $\frac{r}{2\pi}\int K\,\mathrm{d}A$
Dolbeault $\bar\partial_F$	$\chi(X,F) = \sum(-1)^q h^{0,q}(F)$	$\operatorname{ch}(F)\ \operatorname{td}(X)$, $\operatorname{td} = \prod \frac{x_i}{1-e^{-x_i}}$	curve: $\deg F + r(1-g)$; surface: $\int \operatorname{ch}_2(F) + \frac{1}{2}c_1\,c_1(F) + \frac{r}{12}(c_1^2 + c_2)$
signature d_+^F	$\operatorname{Sign}(X)$ if F trivial	$\operatorname{ch}(F)\prod \frac{x_i}{\tanh(x_i/2)}$ (Nash: $2^{n/2}\operatorname{ch}(F)\mathcal{L}$)	$n=2$: $2\int c_1(F)$; $n=4$: $4\int \operatorname{ch}_2(F) + r\operatorname{Sign} X$
spin $\not{D}_F^+$	$n_+ - n_-$ (chiral zero modes)	$\operatorname{ch}(F)\,\hat A(X)$, $\hat A = \prod \frac{x_i/2}{\sinh(x_i/2)}$	$n=2$: $\int c_1(F)$; $n=4$: $\int \operatorname{ch}_2(F) - \frac{r}{24}\int p_1$
family $\{D_y\}$	$[\mathrm{Ker}] - [\mathrm{Coker}] \in K(Y)$	$\operatorname{ch}(\operatorname{index} D) = \pi_*\big(\operatorname{ch}(P)\ \operatorname{td}\big)$ or $\pi_*(\operatorname{ch}(P)\hat A)$	$c_1(\mathrm{DET}) = -[\operatorname{ch}\operatorname{index} D]_{(2)}$
real skew D	$\dim\operatorname{Ker} D \bmod 2$	none (torsion)	Majorana on S^1: R 1, NS 0
G-invariant	$L(g,E) = \sum(-1)^i \operatorname{tr}(g\vert H^i)$	fixed points: $\sum_P \frac{\sum(-1)^i \operatorname{tr}\phi_{i,P}}{\vert\det(1-\mathrm{d}g_P)\vert}$	$\mathbb{CP}^1$, $O(k)$: $\sum_{j=0}^k \lambda^j$
boundary (APS)	index D with $\Pi_{\geq 0}(u\vert_{\partial X}) = 0$	$\int_X P - \frac{1}{2}\big(h + \eta_L(0)\big)$ $(+\int_{\partial X} Q)$	$L = i\frac{\mathrm{d}}{\mathrm{d}\theta} + t$: $\eta(0) = 1 - 2t$
cylinder	$\mathrm{SF}(L_u) - h(L_1)$	bulk and smooth η-drift cancel	flux N pumped: $N - 1$

Two computations show the recipe at work; both are the kind of thing one does on the back of an envelope once the table is memorised.

Worked example 4.154: Chiral zero modes in four dimensions, in one line

Let X be a compact spin 4-manifold and F a gauge bundle of rank r. Then

$$\operatorname{index} \not{D}_F^+ \overset{(1)}{=} \int_X \Big[\operatorname{ch}(F)\,\hat A(X)\Big]_4 \overset{(2)}{=} \int_X \Big[\big(r + c_1(F) + \operatorname{ch}_2(F)\big)\big(1 - \tfrac{1}{24}p_1\big)\Big]_4$$
$$\overset{(3)}{=} \int_X \operatorname{ch}_2(F) - \frac{r}{24}\int_X p_1 \overset{(4)}{=} \int_X \operatorname{ch}_2(F) - \frac{r}{8}\operatorname{Sign}(X).$$

(1) The twisted spin row of the table, degree-4 part.
(2) $\operatorname{ch}(F) = r + c_1(F) + \operatorname{ch}_2(F) + \cdots$ and $\hat A = 1 - \frac{p_1}{24} + \cdots$; higher terms have degree > 4.
(3) Multiply out and keep degree 4: $r\cdot(-\frac{p_1}{24})$ and $\operatorname{ch}_2(F)\cdot 1$; $c_1(F)$ times a degree-4 class is degree 6.
(4) Hirzebruch in four dimensions: $\operatorname{Sign}(X) = \frac{1}{3}\int p_1$.

Two checks. For F trivial ($r = 1$, $\mathrm{ch}_2 = 0$) on K3, with Sign $= -16$, the index is $+2$: the two covariantly constant spinors. For $X = S^4$ (Sign $= 0$) and F the fundamental SU(2) bundle of an instanton, $\mathrm{ch}_2(F) = -c_2(F)$ and the index is $-c_2(F)$: one zero mode per unit of instanton number (the sign is fixed by the orientation convention for k). Rokhlin's theorem hides in the same line: on a spin 4-manifold with F trivial the index is $-\frac{1}{8}$ Sign, so Sign is divisible by 8 (and in fact by 16, because 4D spinors are quaternionic).

Worked example 4.155: Riemann–Roch for a line bundle on a curve

Let $X = \Sigma_g$ be a compact Riemann surface and F a holomorphic line bundle of degree $d = \int c_1(F)$. Then

$$h^0(F) - h^1(F) \overset{(1)}{=} \int_{\Sigma_g} \Big[\mathrm{ch}(F)\,\mathrm{td}(\Sigma_g)\Big]_2 \overset{(2)}{=} \int_{\Sigma_g} \Big[\big(1 + c_1(F)\big)\big(1 + \tfrac{1}{2}c_1(T\Sigma_g)\big)\Big]_2$$
$$\overset{(3)}{=} \int_{\Sigma_g} c_1(F) + \frac{1}{2}\int_{\Sigma_g} c_1(T\Sigma_g) \overset{(4)}{=} d + \frac{1}{2}(2 - 2g) = d + 1 - g.$$

(1) The twisted Dolbeault row: index $= \chi(\Sigma_g, F)$.
(2) $\mathrm{ch}(F) = e^{c_1(F)}$ and $\mathrm{td} = 1 + \frac{1}{2}c_1 + \cdots$ in complex dimension 1.
(3) Keep the 2-form part.
(4) $\int c_1(F) = d$; $c_1(T\Sigma_g)$ is the Euler class, and Gauss–Bonnet gives $\chi(\Sigma_g) = 2 - 2g$.

For $F = \mathcal{O}(k)$ on $\mathbb{CP}^1$ ($g = 0$, $d = k$) this is $k + 1$, the number of polynomials of degree $\le k$. For $F = K^2$, the quadratic differentials ($d = 4g - 4$, $h^0(K^{-1}) = 0$ for $g \ge 2$), it gives $h^0(K^2) = 3g - 3$, the complex dimension of the moduli space of Riemann surfaces used in Chapter 5 and Chapter 9.

One sphere, several theorems

The best way to see that the generalisations of §3–§7 are one theory is to run them on the same object. Take $X = \mathbb{CP}^1 = S^2$ and the operator $\bar\partial$ twisted by $\mathcal{O}(k)$, $k \ge 0$. We already know three things about it.

- *Closed index* (§2–§3): index $= k + 1$, the polynomials $1, z, \ldots, z^k$ (Example 4.155).
- *Character* (§6): under $z \mapsto \lambda z$ the kernel carries the character $\sum_{j=0}^{k} \lambda^j$, and the Atiyah–Bott formula recovers it from the two poles (Example 4.124).
- *Relation to Dirac* (§3): $\not{D}^+$ in a monopole of charge k is $\bar\partial$ on $\mathcal{O}(k-1)$, so it has k zero modes (Example 4.102), forming a spin-$\frac{k-1}{2}$ multiplet by the previous item.

Now we cut the sphere along a circle into two discs and use the boundary theorem (§7) on each half. Gluing the halves back must reproduce $k + 1$. This tests the whole boundary machinery, including the eta invariant and the half-integer bookkeeping of h.

Worked example 4.156: Gluing two APS indices back into Riemann–Roch

Fix $0 < \alpha < k$. Put a metric on $\mathbb{CP}^1$ which is a flat cylinder $\mathrm{d}n^2 + \mathrm{d}\theta^2$ near the equator, and a U(1) connection on $\mathcal{O}(k)$ which is flat there, with flux α through the northern disc X_1 and $k - \alpha$ through the southern disc X_2 (Figure 4.92). Each disc, in its own complex coordinate (z or $w = 1/z$), is exactly the flux disc of Example 4.141, whose boundary

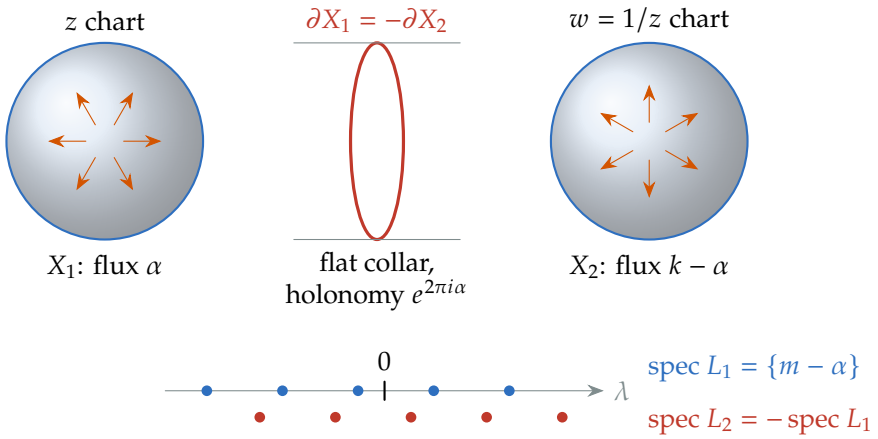


Figure 4.92: Cutting $\mathbb{CP}^1$ into two discs. The flux k of $\mathcal{O}(k)$ is split as α in the northern disc and $k-\alpha$ in the southern one, with a flat collar in between. Seen from the two sides, the collar has boundary operators L_1 and L_2 with *opposite* spectra (blue and red dots, drawn here for $\{\alpha\} \approx 0.35$), because the two discs induce opposite orientations on the circle. So $\eta_{L_2} = -\eta_{L_1}$: the spectral asymmetries cancel on gluing.

operator $L = -i\partial_\theta - \alpha$ has spectrum $\{m - \alpha\}_{m\in\mathbb{Z}}$ and

$$\eta_L(0) = 2\{\alpha\} - 1 \ (\alpha \notin \mathbb{Z}), \qquad \eta_L(0) = 0,\ h = 1 \ (\alpha \in \mathbb{Z}).$$

For X_2 replace α by $\beta = k - \alpha$. Since k is an integer, $\{\beta\} = 1 - \{\alpha\}$ when $\alpha \notin \mathbb{Z}$, so $\eta_{L_2}(0) = 2(1 - \{\alpha\}) - 1 = -\eta_{L_1}(0)$, as the figure says.

Non-integer α ($h = 0$). The APS theorem on each disc gives

$$\begin{aligned}\operatorname{index}_{X_1} + \operatorname{index}_{X_2} &\overset{(1)}{=} \left[\alpha + \tfrac{1}{2} - \tfrac{1}{2}\eta_{L_1}(0)\right] + \left[(k-\alpha) + \tfrac{1}{2} - \tfrac{1}{2}\eta_{L_2}(0)\right] \\ &\overset{(2)}{=} k + 1 - \tfrac{1}{2}\big(\eta_{L_1}(0) + \eta_{L_2}(0)\big) \overset{(3)}{=} k + 1.\end{aligned}$$

(1) (4.233) on each disc, with $\int_{X_i} P = \int c_1(F) + \frac{1}{2}\int e = (\text{flux}) + \frac{1}{2}$ and $h = 0$.

(2) Collect terms; the two half Euler characteristics add up to $\frac{1}{2}\chi(S^2) = 1$.

(3) $\eta_{L_2} = -\eta_{L_1}$.

In terms of the direct counts: $(\lfloor\alpha\rfloor + 1) + (\lfloor k - \alpha\rfloor + 1) = \lfloor\alpha\rfloor + 1 + (k - \lfloor\alpha\rfloor - 1) + 1 = k + 1$. The polynomials z^j with $j < \alpha$ are counted on the north side; those with $j > \alpha$, rewritten as w^{k-j}, on the south side.

Integer α ($h = 1$). Now L_1 has the zero mode $e^{i\alpha\theta}$, the spectra are symmetric, and each disc gives just its flux: $\operatorname{index}_{X_1} = \alpha$, $\operatorname{index}_{X_2} = k - \alpha$. The sum is k, one short. The missing section is z^α itself. On the circle it is the zero mode of L, and both APS conditions project it out (each keeps only strictly negative modes of its own L, and $\lambda = 0$ is not negative for either side). So the gluing law reads

$$\operatorname{index}_X = \operatorname{index}_{X_1} + \operatorname{index}_{X_2} + h, \tag{4.263}$$

and gives $\alpha + (k - \alpha) + 1 = k + 1$. In the non-integer case $h = 0$ and (4.263) is the chain above.

What this shows. The APS pieces are not individually topological: each depends on where we cut (α), through the flux integral and through η. The dependence cancels between the two halves, the local part because the integrals add, and the spectral part because the two sides see L and $-L$. This is the gluing argument of Section 4.21.4 (the defect depends only on the boundary) seen from the index side.

Four spectral pictures

Behind every theorem of the chapter there is a simple statement about how eigenvalues move. Figure 4.93 collects the four mechanisms.

1. *Pairing* (§1). D^*D and DD^* have the same nonzero eigenvalues, because D maps one eigenspace isomorphically onto the other. Only zero modes are unpaired, and their net number cannot change when the nonzero levels move: that is why the index is invariant, and why $\operatorname{Tr} e^{-tD^*D} - \operatorname{Tr} e^{-tDD^*}$ does not depend on t (McKean–Singer).

2. *Jumping* (§4). In a family, a nonzero pair can come down to zero together: dim Ker and dim Coker jump by the same amount and the index does not. The kernels do not form a bundle, but the virtual difference does (stabilisation).

3. *Pairs on the imaginary axis* (§5). A real skew-adjoint operator has eigenvalues $\pm i\lambda$ in pairs. Zero modes can only enter or leave two at a time, so dim Ker mod2 is invariant, even though the complex index is 0.

4. *Asymmetry and flow* (§7). A self-adjoint L (the boundary operator) need not have a symmetric spectrum. The regularised excess of positive over negative eigenvalues is $\eta_L(0)$. When an eigenvalue crosses zero, η jumps by 2; the net number of crossings is the spectral flow, and it equals an index on the cylinder.

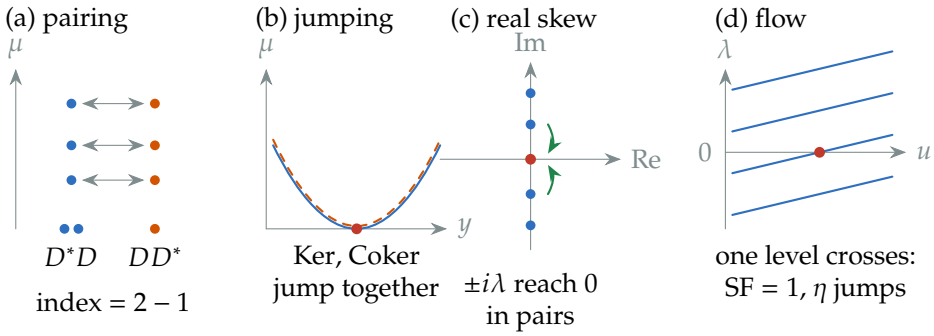


Figure 4.93: The four spectral mechanisms of the chapter. (a) Nonzero levels of D^*D and DD^* are paired; only zero modes (at $\mu = 0$) count, here $2 - 1 = 1$. (b) In a family a paired level can touch zero (red dot): kernel and cokernel jump together, the index does not. (c) For a real skew-adjoint operator eigenvalues come in pairs $\pm i\lambda$ and reach zero two at a time, so dim Ker mod2 is invariant. (d) Eigenvalues of a self-adjoint family L_u drift upward; the level that crosses $\lambda = 0$ is one unit of spectral flow, and it makes $\eta_{L_u}(0)$ jump by 2.

Concept table I: the index and the Atiyah–Singer theorem (Nash §1)

Concept	Operational meaning: how to compute or test it	Used later
kernel, cokernel, index of a linear map	$\operatorname{Coker} T = W/\operatorname{Im} T \cong \operatorname{Ker} T^*$. In finite dimensions $\operatorname{index} T = \dim V - \dim W$ whatever T is (rank–nullity), so the index ignores the details of T.	everywhere
Fredholm operator	Finite-dimensional kernel and cokernel, closed range. Index is locally constant, unchanged by compact perturbations, additive under composition. Test case: the shift on ℓ^2, index -1.	spectral flow (Chapter 10); Fock spaces (Chapter 6)
elliptic operator is Fredholm	Invertible leading symbol $\Rightarrow$ parametrix (inverse up to smoothing, hence compact, errors) $\Rightarrow$ Fredholm. Lower-order terms never change the index ($\mathrm{d}/\mathrm{d}x + a$ on S^1 has index 0 for every a).	all index computations
Toeplitz operator	T_f on the Hardy space of S^1: $\operatorname{index} T_f = -\operatorname{wind}(f)$. The simplest index theorem: analysis (zero modes) equals topology (winding).	loop groups (Chapter 6)
K-theory from complexes	A complex of bundles exact outside a compact set defines a class in K_c; on compact X it is $\sum(-1)^i[E^i]$. Bott class on $\mathbb{C}$: $0 \to \underline{\mathbb{C}} \xrightarrow{z} \underline{\mathbb{C}} \to 0$, i.e. the monopole bundle on S^2.	symbol classes; Thom classes
elliptic complex, rolled-up operator	$\cdots \to \Gamma(E^i) \xrightarrow{d} \Gamma(E^{i+1}) \to \cdots$ with exact symbol sequence. $D = d + d^*$ from even to odd; Laplacians $\Delta_i = d^*d + dd^*$; Hodge: $H^i \cong \operatorname{Ker} \Delta_i$; $\operatorname{index} D = \sum(-1)^i \dim H^i$.	deformation complexes of instantons (Chapter 8), moduli (Chapter 5)
analytic index index_a	$\operatorname{index}_a : K(T^*X) \to \mathbb{Z}$, $[\sigma(D)] \mapsto \operatorname{index} D$: well defined because deformations of the symbol do not change the index.	families, G-index, all of §3–§7
Thom isomorphism, topological index index_t	Embed $X \subset \mathbb{R}^m$; push the symbol class to $T^*\mathbb{R}^m$ with the Thom class of the normal bundle (a Gaussian bump / product of Bott classes); Bott periodicity identifies $K_c(\mathbb{R}^{2m}) = \mathbb{Z}$.	the Thom form in the cohomological formula
Atiyah–Singer theorem	$\operatorname{index}_a = \operatorname{index}_t$. Forced by axioms (excision, multiplicativity, normalisation on a point); checked on Gauss–Bonnet and Riemann–Roch for $\mathbb{CP}^1$.	the whole book
heat kernel, McKean–Singer	$\operatorname{index} D = \operatorname{Tr} e^{-tD^*D} - \operatorname{Tr} e^{-tDD^*}$ for every $t > 0$; as $t \to 0$ this is $\int_X$(local density). Witten index $\operatorname{Tr}(-1)^F e^{-\beta H}$ of SUSY QM.	Morse theory (Chapter 7); Fujikawa anomaly (Chapter 10)

Concept table II: the cohomological formula and the classical complexes (Nash §2)

Concept	Operational meaning: how to compute or test it	Used later
Chern character, Chern roots	$\mathrm{ch}(E) = \mathrm{tr}\, e^{iF/2\pi} = \sum_i e^{x_i}$; additive and multiplicative, so a ring map $K(X) \to H^{\mathrm{even}}(X;\mathbb{Q})$. Splitting principle: compute as if E were a sum of line bundles.	every index formula
td, $\hat{A}$, L	Multiplicative classes from the power series $\frac{x}{1-e^{-x}}$, $\frac{x/2}{\sinh(x/2)}$, $\frac{x}{\tanh x}$. First terms: $1 + \frac{c_1}{2} + \frac{c_1^2+c_2}{12}$; $1 - \frac{p_1}{24}$; $1 + \frac{p_1}{3}$.	Riemann–Roch (Chapter 5); gravitational anomalies (Chapter 10)
cohomological index formula	(4.262). Integrality of $\int \hat{A}$, $\int \mathrm{td}$ is a nontrivial consequence (e.g. Sign $\equiv 0$ mod 16 for spin 4-manifolds).	as above
odd dimensions	Differential operators on odd-dimensional closed manifolds have index 0 (the antipodal map $p \to -p$ reverses the orientation of T^*X). Pseudo-differential operators can evade this (Nash's example on S^1 has index 1).	why η, not the index, is the odd-dimensional invariant (§7)
Thom class, Euler class	The Thom form is a unit-mass Gaussian in the fibres; restricted to the zero section it is $e(X) = \mathrm{Pf}(\Omega/2\pi)$. Poincaré–Hopf: $\int e = \sum$ indices of zeros of a vector field.	Morse theory (Chapter 7)
de Rham complex	$\mathrm{index}(\mathrm{d} + \mathrm{d}^*) = \chi(X) = \int_X e(X)$ (Chern–Gauss–Bonnet); in 2D $\frac{1}{2\pi}\int K\, \mathrm{d}A$.	Morse inequalities (Chapter 7)
Dolbeault complex	$\mathrm{index}(\bar{\partial} + \bar{\partial}^*) = \sum (-1)^q h^{0,q} = \int \mathrm{td}(X)$; Hodge numbers $h^{p,q}$; Noether's formula $\chi(\mathcal{O}) = \frac{1}{12}\int (c_1^2 + c_2)$ on surfaces.	Riemann surfaces, Teichmüller space (Chapter 5, Chapter 9)
signature complex	$\tau = i^{p(p-1)+n/2}\star$ splits forms; $\mathrm{index}\, \mathrm{d}_+ = \mathrm{Sign}(X) = b^+ - b^-$ of the intersection form $= \int L$ (Hirzebruch; $\frac{1}{3}\int p_1$ in 4D).	intersection forms, Donaldson theory (Chapter 8, Chapter 12)
spin structure, Dirac operator	Spin structures exist iff $w_1 = w_2 = 0$; $\not{D} = \gamma^a e_a{}^\mu(\partial_\mu + \omega_\mu)$; $\mathrm{index}\, \not{D}^+ = n_+ - n_- = \int \hat{A}$. Lichnerowicz: positive scalar curvature $\Rightarrow$ no harmonic spinors $\Rightarrow \int \hat{A} = 0$.	gravitational anomalies (Chapter 10), strings (Chapter 9)
twisted Dirac, zero modes	$\mathrm{index}\, \not{D}_F^+ = \int \mathrm{ch}(F)\hat{A}$: $\int c_1(F)$ zero modes in a 2D flux (Aharonov–Casher); $\lvert k\rvert$ fundamental and $4\lvert k\rvert$ adjoint zero modes for an SU(2) instanton; moduli dimension $8k - 3$.	instantons, 't Hooft vertex (Chapter 8, Chapter 10)

Concept table III: twisting, families and real families (Nash §3–§5)

Concept	Operational meaning: how to compute or test it	Used later
matrix-valued forms, d_A	$\Omega^p(X;F)=\Gamma(F\otimes\Lambda^pT^*X_{\mathbb{C}})$; $d_A\omega=\mathrm{d}\omega+A\wedge\omega$ is gauge covariant and $d_A^2=F_A$, so the twisted de Rham sequence is a complex only for flat F (Section 4.17.2).	gauge fields throughout (Chapter 8)
twisting rule	Replace E^p by $F\otimes E^p$; since $\mathrm{ch}(A\otimes B)=\mathrm{ch}\,A\ \mathrm{ch}\,B$ the integrand is multiplied by $\mathrm{ch}(F)$ (Section 4.17.1).	all of the above
twisted de Rham	$\operatorname{index}\mathrm{d}_F=\int\mathrm{ch}(F)e(X)=r\chi(X)$: only the rank survives, because $e(X)$ is already top degree.	flat bundles, local systems
twisted Dolbeault	$\chi(X,F)=\int\mathrm{ch}(F)\,\mathrm{td}(X)$ (Hirzebruch–Riemann–Roch). On $\mathbb{CP}^1$: $\dim H^0(\mathcal{O}(k))=k+1$; on $\mathbb{CP}^2$: $\binom{k+2}{2}$ (Section 4.17.3).	Chapter 5, Chapter 9
twisted signature, twisted Dirac	$\int\mathrm{ch}(F)\prod\frac{x_i}{\tanh(x_i/2)}$ and $\int\mathrm{ch}(F)\hat{A}$ (Sections 4.17.4 and 4.17.5); nonzero even in dimensions where the untwisted index vanishes (e.g. $\not{D}$ on S^2 in a monopole: k).	zero modes (Chapter 8)
family, index bundle	$\{D_y\}_{y\in Y}$: $\operatorname{index}D=[\ker D]-[\operatorname{Coker}D]\in K(Y)$; if the kernel jumps, stabilise by adding $\mathbb{C}^N$ to kill the cokernel (Sections 4.18.3 and 4.18.4).	anomalies (Chapter 10)
fibre bundle of manifolds	$X\to Z\to Y$ with structure group $\mathrm{Diff}(X)$; symbols give a class in $K(T_FZ)$ (tangent along fibres); $\operatorname{index}_a=\operatorname{index}_t:K(T_FZ)\to K(Y)$ (Section 4.18.5).	gravitational anomalies, moduli of metrics (Chapter 10, Chapter 9)
families formula, π_*	$\mathrm{ch}(\operatorname{index}D)=(-1)^n\pi_*\{\mathrm{ch}[\sigma]\,\mathrm{td}(T_FZ_{\mathbb{C}})\}$; π_* integrates the components with $\dim X$ legs along the fibre and lowers degree by $\dim X$ (Section 4.18.7). Flat bundles on T^2: $c_1=\mp1$ over the dual torus (Section 4.18.8).	descent equations (Chapter 10)
torsion	ch is rational and kills torsion in $K(Y)$; a family can be nontrivial with $\mathrm{ch}(\operatorname{index}D)=\text{rank}$ (Section 4.18.9).	global anomalies (Chapter 10)
determinant line bundle	$\mathrm{DET}=(\Lambda^{\max}\ker)^*\otimes\Lambda^{\max}\operatorname{Coker}$, a line bundle even when kernels jump; $c_1(\mathrm{DET})=-c_1(\operatorname{index}D)$. The fermion determinant is a section of it; $c_1\neq0$ is a gauge anomaly (Section 4.18.10).	Quillen metric, anomalies, string measure (Chapter 10, Chapter 9)
real operator, KR	Real symbol: $\overline{\sigma(x,p)}=\sigma(x,-p)$ (real coefficients). Real families have index in $KR(Y)$, K-theory with the involution $p\to-p$; for trivial involution $KR=KO$ (Sections 4.19.1 and 4.19.2).	global anomalies (Chapter 10)
mod 2 index	Real skew-adjoint D: $\operatorname{index}_1D=\dim\ker D \bmod 2$, invariant because eigenvalues $\pm i\lambda$ pair up. It is the index of a real family over S^1, $m[\eta]$ with $[\eta]$ the Möbius generator of $\widetilde{KR}(S^1)=\mathbb{Z}_2$; no cohomological formula (Sections 4.19.3 to 4.19.5).	Witten's SU(2) anomaly (Chapter 10); $(-1)^F$ in SUSY QM (Chapter 7)

Concept table IV: symmetries and fixed points (Nash §6)

Concept	Operational meaning: how to compute or test it	Used later
Lefschetz number $L(g)$	$\sum_p (-1)^p \operatorname{tr}(g^* \mid H^p)$; homotopy invariant, $L(\mathrm{id}) = \chi(X)$; $L(g) \neq 0 \Rightarrow$ fixed point. Circle maps: $L = 1 - \deg g$; sphere maps: $L = 1 + \deg g$ (Section 4.20.1).	Morse theory (Chapter 7)
Lefschetz–Hopf	$L(g) = \sum_{g(P)=P} \operatorname{sgn} \det(1 - \mathrm{d}g_P)$ for non-degenerate fixed points; rotation of S^2: two poles, each +1 (Section 4.20.2).	localisation (Chapter 7)
$L(g, E)$ for an elliptic complex	$\sum_i (-1)^i \operatorname{tr}(g \mid H^i(E))$ for g commuting with the d_i; with G compact it is a character, the *character-valued index* (Section 4.20.3).	symmetric instantons and monopoles (Chapter 8)
equivariant K-theory, $R(G)$	$K_G(X)$ from G-bundles ($F_x \to F_{gx}$); $K_G(\mathrm{pt}) = R(G)$, the representation ring (formal differences of representations, added by $\oplus$, multiplied by $\otimes$); $K_{\{e\}}(X) = K(X)$.	loop group representations (Chapter 6)
G-index theorem, Atiyah–Segal	$\mathrm{index}_a = \mathrm{index}_t : K_G(T^*X) \to R(G)$; $L(g, E) = \chi(g)$ is an integral over the fixed-point set X^g (Section 4.20.4).	equivariant Morse theory (Chapter 7)
Atiyah–Bott formula	Isolated fixed points: $L(g, E) = \sum_P \frac{\sum_i (-1)^i \operatorname{tr} \phi_{i,P}}{\lvert \det(1 - \mathrm{d}g_P) \rvert}$. Holomorphic version on $\mathbb{CP}^1$ gives the Weyl character $\frac{\sin(s+\frac{1}{2})\alpha}{\sin \frac{\alpha}{2}}$ (Example 4.124).	characters, fusion (Chapter 6, Chapter 11)
heat-kernel localisation	$\operatorname{Tr}(g\, e^{-t\Delta})$ has kernel $K_t(gx, x)$, which is exponentially small unless $gx \approx x$; so as $t \to 0$ only fixed points contribute (Section 4.20.5).	SUSY localisation (Chapter 7)

Concept table V: manifolds with boundary (Nash §7)

Concept	Operational meaning: how to compute or test it	Used later
boundary conditions, local vs global	Dirichlet/Neumann are local and fine for Δ; for $\bar\partial$, chiral Dirac and d_+ the number of decaying boundary modes depends on the sign of the boundary frequency, so no local condition gives a Fredholm problem (Section 4.21.1).	domain walls, edge states
boundary operator, APS condition	Near a product boundary $D = A(\partial_n + L)$, L self-adjoint elliptic on ∂X; impose $\Pi_{\geq 0}(u|_{\partial X}) = 0$, i.e. L^2 on the elongated manifold; $\Pi = \frac{1}{2}M^{-1}(M + |M|)$ is non-local (Hardy projection on S^1) (Section 4.21.2).	Chapter 10, Chapter 12
Gauss–Bonnet with boundary	$\chi(X) = \frac{1}{2\pi}(\int K\,\mathrm{d}A + \oint k_g\,\mathrm{d}s)$; the boundary term is local and vanishes for a product collar (Section 4.21.3).	string world-sheets with boundary (Chapter 9)
signature defect	$\mathrm{Sign}(X) - \frac{1}{3}\int p_1 = -\frac{1}{2}\eta_L(0)$ depends only on ∂X (gluing argument) and is not a local integral (Section 4.21.4).	3-manifold invariants (Chapter 12)
zeta function	$\zeta_P(s) = \sum \mu^{-s} = \mathrm{Tr}\,P^{-s} = \frac{1}{\Gamma(s)}\int_0^\infty t^{s-1}\,\mathrm{Tr}'\,e^{-tP}\mathrm{d}t$; finite at $s = 0$; $\zeta_P(0) = a_n - \dim\mathrm{Ker}\,P$, a_n the t^0 heat coefficient (Section 4.21.5).	determinants, conformal anomaly (Chapter 9, Chapter 11)
eta invariant	$\eta_L(s) = \sum \mathrm{sgn}\,\lambda|\lambda|^{-s}$; $\eta_L(0)$ = regularised $\#(\lambda > 0) - \#(\lambda < 0)$. Circle: $L = i\frac{\mathrm{d}}{\mathrm{d}\theta} + t$ gives $\eta(0) = 1 - 2t$ on $0 < t < 1$ (Hurwitz zeta, heat kernel, Poisson) (Section 4.21.6).	parity anomaly, CS framing (Chapter 10, Chapter 12)
APS index theorem	$\mathrm{index}\,D = \int_X P - \frac{1}{2}(h + \eta_L(0))$ with $h = \dim\mathrm{Ker}\,L$. Examples: Jackiw–Rebbi domain wall, flux disc $\lfloor\alpha\rfloor + 1$, $\bar\partial$ on Σ_g minus p discs: $1 - g - p$ (Section 4.21.7).	anomalies (Chapter 10)
which indices need η	The de Rham (Euler) complex needs no correction: its boundary spectrum is symmetric. The signature and Dirac operators do (Section 4.21.8).	
spectral flow	$\mathrm{SF}(L_u)$ = net number of eigenvalues crossing 0 upward; $\mathrm{index}(\partial_u + L_u) = \mathrm{SF} - h(L_1)$ on a cylinder (Section 4.21.10).	Hamiltonian anomalies (Chapter 10); Floer homology (Chapter 12)
non-product metrics	Add $\int_{\partial X} Q$, a transgression (Chern–Simons type) form built from the second fundamental form: $P(R) - P(R') = \mathrm{d}Q$ (Section 4.21.11).	Chern–Simons forms (Chapter 8, Chapter 12)
finiteness of $\eta(0)$	Residue at $s = 0$ is a local integral that vanishes: by parity in even dimensions, by a global argument in odd ones (Section 4.21.12).	
η in physics	Vacuum charge $-\frac{1}{2}\eta_H(0)$; odd-dimensional fermion determinant phase $e^{\mp i\pi\eta/2}$ (parity anomaly); $e^{-i\pi\eta}$ on a mapping torus is a global anomaly (Section 4.21.13).	Chapter 10, Chapter 12

Which tool answers which physics question

The concept tables are organised by mathematics. A physicist usually starts from a question. The table below runs the other way: from the question to the invariant that answers it, the formula that computes it, and the chapter where the answer is used.

Question	Invariant	Formula	Where
How many chiral fermion zero modes in this gauge background?	index $\not{D}_F^+$	$\int \mathrm{ch}(F)\hat{A}$; 2D: flux; 4D: $\int \mathrm{ch}_2(F)$	Chapter 8
How many parameters does the space of instantons / complex structures have?	index of a deformation complex	$8k-3$ for SU(2) on S^4; $3g-3$ for curves	Chapter 8, Chapter 5
Is the fermion determinant a gauge-invariant function?	$c_1(\mathrm{DET})$ over $\mathcal{A}/\mathcal{G}$	families formula in $n+2$ dimensions	Chapter 10
Is there an anomaly no Feynman diagram sees?	torsion of index; mod 2 index; $e^{-i\pi\eta}$	dim Ker mod2 of a 5D operator (Witten SU(2))	Chapter 10
How much charge does a soliton or a flux carry?	$\eta_H(0)$	$\langle Q\rangle = -\frac{1}{2}\eta_H(0)$	Chapter 10
How much charge is pumped when a parameter winds?	spectral flow	$\mathrm{SF} = \mathrm{index}(\partial_u + L_u)$	Chapter 10, Chapter 12
Phase of the fermion path integral in odd dimensions?	$\eta(0)$ of $i\not{D}_A$	$\lvert\det\rvert\, e^{\mp i\pi\eta/2}$	Chapter 10, Chapter 12
How do the zero modes transform under a symmetry?	character-valued index	Atiyah–Bott sum over fixed points	Chapter 7, Chapter 6
Does SUSY break? How many ground states?	Witten index	$\mathrm{Tr}(-1)^F e^{-\beta H} = \mathrm{index}$	Chapter 7
Why is $\sum \lambda^{-s}$ or $\det\Delta$ finite?	ζ function	Mellin transform of the heat trace	Chapter 9, Chapter 11

Reading Nash (Chapter IV at a glance, pp. 89–136 (PDF 99–146))

Nash's chapter, section by section, with where each part of these notes treats it.

- §1 *The index of an elliptic operator*, pp. 89–97, (4.1)–(4.30): Fredholm property, symbol classes, index_a and index_t, the Atiyah–Singer theorem. Nash quotes rather than explains; the first part of this chapter builds it from linear algebra, and adds Toeplitz operators and the heat kernel.
- §2 *Some examples*, pp. 97–116, (4.31)–(4.111): the cohomological formula, odd dimensions, the Thom and Euler classes, and the de Rham, Dolbeault, signature and spin complexes.
- §3 *Twisted complexes*, pp. 116–119, (4.112)–(4.122): Section 4.17.
- §4 *The index theorem for families of operators*, pp. 119–121, (4.123)–(4.127): Section 4.18; the determinant line bundle is our addition, and Nash takes it up in his anomalies chapter.
- §5 *The index for real families*, pp. 121–123, (4.128)–(4.129): Section 4.19.
- §6 *Index theory and fixed points*, pp. 123–127, (4.130)–(4.140): Section 4.20; the Atiyah–Bott formula and its examples are our addition to Nash's statement of the Atiyah–Segal

theorem.

- §7 *Index theory for manifolds with boundary*, pp. 127–136, (4.141)–(4.177): Section 4.21.

Nash ends the chapter without a summary. The formula sheet, the gluing example of Example 4.156 and the tables above are ours.

Recap: What to carry forward from Chapter

- The index is the only part of the zero-mode count that is stable. It depends on the leading symbol up to deformation, so it is topological, and it is a local integral of curvature.
- The integrand is always a product: ch(coefficients) times a genus of X (e, td, L, $\hat{A}$). Keep the top-degree part and integrate.
- With parameters the index becomes a bundle, its top exterior power is the determinant line, and its Chern classes are anomalies. Torsion (mod 2 indices) carries global anomalies, invisible to curvature.
- With a symmetry the index becomes a character, and it localises on fixed points.
- With a boundary the index needs a non-local boundary condition and picks up $-\frac{1}{2}(h+\eta)$; η is the spectral asymmetry, its jumps are spectral flow, and its phase is the parity anomaly.
- The heat kernel is the common engine: it proves McKean–Singer, localisation on fixed points, and the η term of APS.

Chapter 5

Some Algebraic Geometry

Where this chapter is going

Physics keeps producing complex manifolds. The worldsheet of a closed string in Euclidean signature is a two-dimensional surface with a conformal structure, and a surface with a conformal structure is the same thing as a *Riemann surface*. The extra dimensions of superstring theory are Calabi–Yau manifolds, which are Kähler. The self-dual instantons of Chapter 8 are classified by holomorphic data on twistor space. Supersymmetric sigma models with extended supersymmetry need a Kähler target. In all of these places the most efficient way to describe the manifold is not by charts and transition functions but by *polynomial equations*: the torus that a one-loop string amplitude integrates over is the solution set of $y^2 = x(x-1)(x-\lambda)$, and the number λ is the modulus of the torus.

Algebraic geometry is the study of such solution sets. Nash's Chapter V is a compact tour of the pieces of it that a physicist needs later in the book. We go in the following order. First, what an algebraic variety is, and why projective space is the natural place to put it (§5.1). We learn to see projective space in the real projective plane $\mathbb{RP}^2$, where parallel lines meet on the horizon and a parabola closes up into a circle, before passing to $\mathbb{CP}^n$ and the Hopf fibration, and we sort out which pictures of these spaces can be trusted. Then we specialise to complex dimension one, the Riemann surfaces Σ_p of genus p, and learn to describe line bundles on them by recording where their sections vanish or blow up: this bookkeeping device is a *divisor* (§5.2). The integrals of holomorphic 1-forms around the $2p$ independent loops of Σ_p form the *period matrix*, which packages the complex structure into a $p \times p$ symmetric matrix and produces a torus, the *Jacobian*. Next come the global tools (§5.3): Serre duality, the Riemann–Roch theorem that counts holomorphic sections, the classification of line bundles by the exponential sheaf sequence, and finally Kähler manifolds and the Hodge decomposition of their cohomology. The second part of the chapter (Teichmüller space and the moduli space $\mathcal{M}_p$) uses all of this to parametrise the possible complex structures on Σ_p, which is exactly the space a string path integral integrates over in Chapter 9.

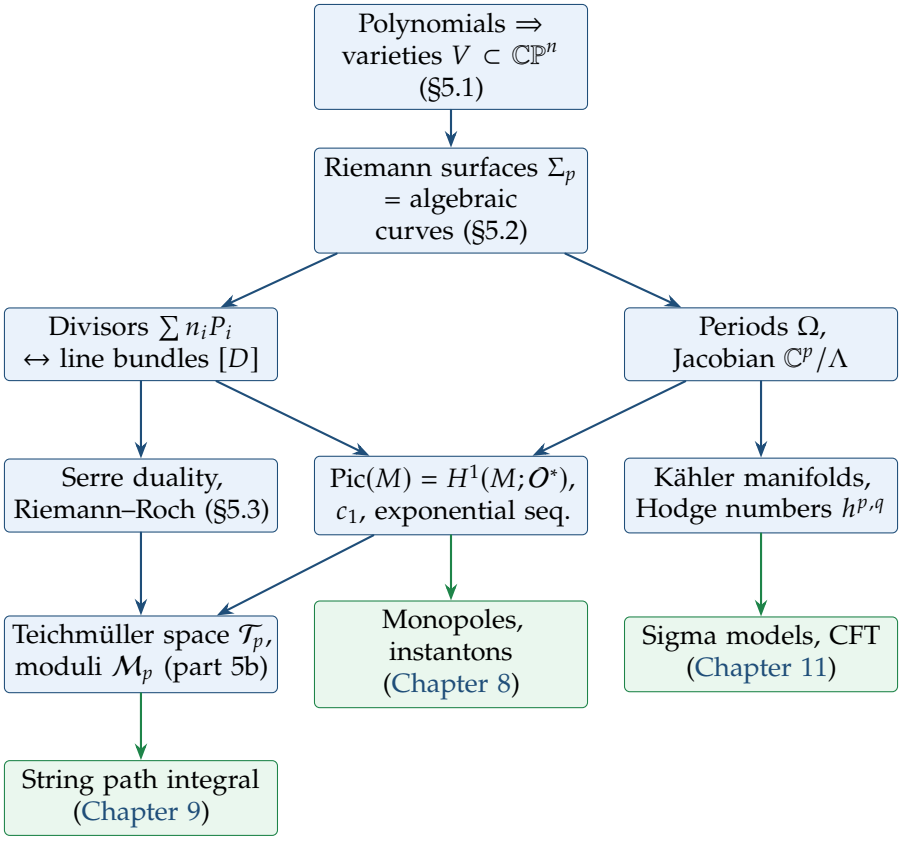


Figure 5.1: The logic of Chapter V. Blue boxes are the mathematics of this part and the next; green boxes are where the physics uses it. The central object is the Riemann surface. Divisors, periods and Kähler geometry are three complementary ways of getting a handle on its complex structure, and Teichmüller theory assembles them into a space of complex structures.

Dictionary: algebraic geometry for physicists

Mathematics	Physics
compact Riemann surface Σ_p	Euclidean closed-string worldsheet with p loops
complex structure modulus (λ, τ, Ω)	shape parameters integrated over in string amplitudes
holomorphic line bundle L	bundle whose sections are charged wavefunctions
degree $\deg L = c_1(L)$	magnetic flux through the surface (Dirac quantisation)
holomorphic section of L	lowest-Landau-level wavefunction
divisor of a section	positions of the zeros (vortices) of that wavefunction
Kähler form ω	symplectic form; phase space of a system; target metric of an $N = 2$ sigma model

5.1 Algebraic varieties

What kind of space do you get if you solve polynomial equations? In real variables the answer can be almost anything: $x^2 + y^2 = 1$ is a circle, $x^2 + y^2 = -1$ is empty, and $y^2 = x^3$ has a sharp corner. Over the complex numbers the answers become much more uniform and much more rigid. A single polynomial equation in two complex variables always has solutions, the solution set is (apart from finitely many bad points) a smooth surface, and once we add the "points at infinity" the surface becomes compact and its topology is fixed by the degree of the polynomial. This section sets up the language: affine varieties; complex functions pictured as maps from a plane to a plane; projective space, first real (where we can draw it: lines through the origin, perspective, conics closing up at infinity) and then complex; projective varieties; singular points; and two theorems (Chow and Kodaira) that say when a complex manifold can be described by polynomials at all.

Throughout, as in Nash, the field of numbers is $\mathbb{C}$.

5.1.1 Zero sets of polynomials

Working definition 5.1 (affine variety)

Pick polynomials $P_1, \dots, P_k$ in m complex variables $x = (x_1, \dots, x_m)$. Their common zero set

$$V(P_1, \dots, P_k) \;=\; \{\, x \in \mathbb{C}^m \;:\; P_1(x) = \cdots = P_k(x) = 0 \,\}$$

is an **affine algebraic variety**. ["affine" means it sits in ordinary $\mathbb{C}^m$, with no points at infinity added] *Operationally*: a point belongs to V if you can substitute it into every P_i and get zero.

A count of equations gives a first guess at the size of V. Each equation cuts one complex dimension, so we expect $\dim_{\mathbb{C}} V = m - k$. When that guess is correct and the equations are "independent" at every point, V is a complex manifold of dimension $m - k$. When it fails, V has *singular points*. We make this precise in Section 5.1.9.

Watch out: complex dimension versus real dimension

A complex curve (complex dimension one) is a real surface (real dimension two). Algebraic geometers say "curve" and draw surfaces. When we draw a picture of a complex curve in $\mathbb{C}^2$ we usually draw only its *real points*, the solutions with $x, y \in \mathbb{R}$. This is a one-dimensional slice of a two-dimensional object. It is useful for spotting singularities, but it can badly misrepresent topology.

Worked example 5.1: The complex circle is a punctured plane

Take $V = \{x^2 + y^2 = 1\} \subset \mathbb{C}^2$. Its real points form the unit circle. Its complex points are a very different space. Introduce $u = x + \mathrm{i}y$ and $v = x - \mathrm{i}y$ (with x, y complex, so u and v are independent complex numbers, not conjugates). Then

$$x^2 + y^2 \overset{(1)}{=} (x + \mathrm{i}y)(x - \mathrm{i}y) \overset{(2)}{=} uv,$$

(1) Expand: $(x + \mathrm{i}y)(x - \mathrm{i}y) = x^2 - \mathrm{i}^2 y^2 = x^2 + y^2$, valid for any complex x, y.
(2) Definition of u and v.

So $V = \{uv = 1\}$. For every $u \neq 0$ there is exactly one $v = 1/u$, so $u \mapsto (u, 1/u)$ identifies V with $\mathbb{C}^* = \mathbb{C} \setminus \{0\}$, a cylinder. The real circle is the loop $|u| = 1$ going once round that cylinder. The complex circle is not compact: it has two ends, $u \to 0$ and $u \to \infty$. Adding one point at each end would close it up into a sphere. That is exactly what projective space will do for us.

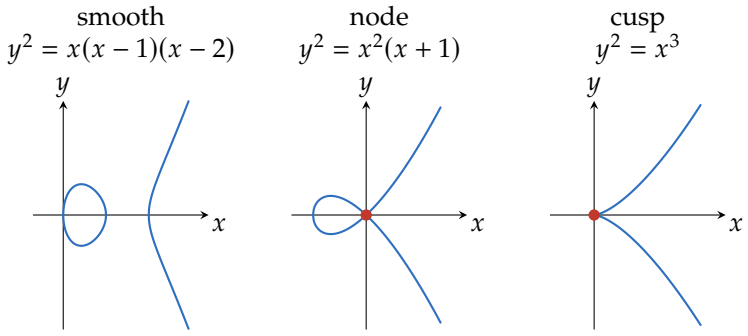


Figure 5.2: Real points of three plane cubic curves. The smooth cubic (left) has two real pieces, an oval and an unbounded branch; its complex points form a torus with one point missing (the point at infinity). The node (middle) has two branches crossing at the red point; the cusp (right) has a single branch with a sharp point. At the red points the Jacobian of the defining polynomial vanishes: those are the singular points.

5.1.2 A complex function is a map from a plane to a plane

From here on the variables are complex, and complex functions appear everywhere: the map $u \mapsto 1/u$ just above, the square root $y = \pm\sqrt{q(x)}$ that solves a cubic, the changes of chart on a Riemann surface. How do we draw a function $y = f(x)$ when x and y are complex?

A complex number $x = a + ib$ is a point (a, b) of a plane, the **complex x-plane**. One complex dimension is two real dimensions. A function $f : \mathbb{C} \to \mathbb{C}$ takes each point of the x-plane to a point of a second plane, the complex y-plane. In real terms it is a pair of real functions of two

real variables,

$$\operatorname{Re} y = U(a,b), \qquad \operatorname{Im} y = W(a,b),$$

that is, a map $\mathbb{R}^2 \to \mathbb{R}^2$, like a change of coordinates or the velocity field of a flow in a plane. Its graph $\{(x, f(x))\}$ lives in $\mathbb{C}^2 = \mathbb{R}^4$, and we cannot draw it. So we draw something else.

Working definition 5.2 (picturing a complex function)

Draw two planes side by side: the x-plane on the left (the inputs) and the y-plane on the right (the outputs). Pick a simple curve in the x-plane, such as a circle around the origin, a ray out of the origin or a straight grid line. Send every point of the curve through f. The image points form a curve in the y-plane; draw it in the same colour. A few such pairs of curves show how f stretches, turns and folds the plane, just as a few grid lines show what a change of coordinates does. *Operationally*: write the curve as $x(t)$, compute $y(t) = f(x(t))$, split it into real and imaginary parts, and eliminate t.

The tool that makes circles and rays easy is the polar form. Write $x = r\,e^{\mathrm{i}\theta}$, where $r = |x|$ is the distance from the origin and $\theta = \arg x$ is the angle from the positive real axis, and $e^{\mathrm{i}\theta} = \cos\theta + \mathrm{i}\sin\theta$ (Euler's formula). Since $e^{\mathrm{i}\alpha}e^{\mathrm{i}\beta} = e^{\mathrm{i}(\alpha+\beta)}$, multiplying two complex numbers multiplies their lengths and adds their angles.

Worked example 5.2: $y = x^2$ doubles angles and squares lengths

Circles and rays. Put $x = r\,e^{\mathrm{i}\theta}$. Then

$$\begin{aligned} y &\overset{(1)}{=} \left(r\,e^{\mathrm{i}\theta}\right)^2 \\ &\overset{(2)}{=} r^2\,e^{2\mathrm{i}\theta}. \end{aligned}$$

(1) Definition of the map.
(2) $(r\,e^{\mathrm{i}\theta})^2 = r^2\,e^{\mathrm{i}\theta}e^{\mathrm{i}\theta}$, and the angles add: $e^{\mathrm{i}\theta}e^{\mathrm{i}\theta} = e^{2\mathrm{i}\theta}$.

So $|y| = r^2$ and $\arg y = 2\theta$. The circle $|x| = r$ goes to the circle $|y| = r^2$. As x runs once around its circle, θ runs from 0 to 2π and 2θ runs from 0 to 4π: the image circle is traced *twice*. The ray $\arg x = \theta$ goes to the ray $\arg y = 2\theta$. The opposite ray, $\arg x = \theta + \pi$, goes to $\arg y = 2\theta + 2\pi$, which is the same ray. In general x and $-x$ have the same image, so every $y \neq 0$ has exactly two preimages: the map is two-to-one.
A grid line. Take the vertical line $\operatorname{Re} x = c$ with $c > 0$, that is, $x = c + \mathrm{i}t$ with $t \in \mathbb{R}$. Then

$$\begin{aligned} y &\overset{(1)}{=} c^2 + 2\mathrm{i}ct + (\mathrm{i}t)^2 \\ &\overset{(2)}{=} c^2 - t^2 + 2\mathrm{i}ct, \end{aligned}$$

(1) Expand $(c + \mathrm{i}t)^2$.
(2) $\mathrm{i}^2 = -1$.

so $\operatorname{Re} y = c^2 - t^2$ and $\operatorname{Im} y = 2ct$. Eliminate t:

$$\begin{aligned} \operatorname{Re} y &\overset{(1)}{=} c^2 - t^2 \\ &\overset{(2)}{=} c^2 - \left(\frac{\operatorname{Im} y}{2c}\right)^2 \\ &= c^2 - \frac{(\operatorname{Im} y)^2}{4c^2}. \end{aligned}$$

(1) The real part found above.

(2) Solve $\operatorname{Im} y = 2ct$ for t, using $c \neq 0$.

This is a parabola that opens to the left. Its tip is at $y = c^2$ (from $t = 0$), and it crosses the imaginary axis at $\operatorname{Im} y = \pm 2c^2$ (from $t = \pm c$). The line $\operatorname{Re} x = -c$ goes to the same parabola, again because $(-x)^2 = x^2$.

Angles. In the x-plane the circles and the rays cross at right angles, and so do their images. This is a general fact: a complex-differentiable map preserves angles wherever $f'(x) \neq 0$. Physicists use it in two-dimensional electrostatics and fluid flow, where such a **conformal map** carries one field pattern to another. Here $f'(x) = 2x$ vanishes at $x = 0$, and exactly there angles are doubled rather than preserved: two rays at angle θ to each other go to rays at angle 2θ.

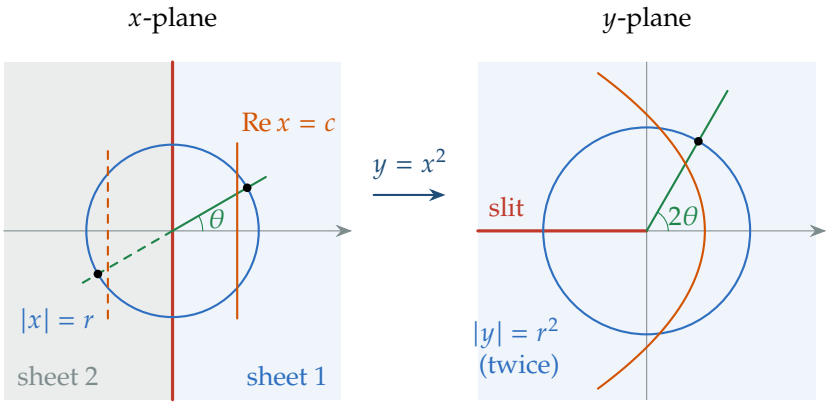


Figure 5.3: The map $y = x^2$, drawn as two planes; curves of the same colour correspond. The circle $|x| = r$ (blue) goes to the circle $|y| = r^2$, traced twice. The ray $\arg x = \theta$ (green, solid) and the opposite ray (green, dashed) both go to the ray $\arg y = 2\theta$. The vertical line $\operatorname{Re} x = c$ (orange, solid) and the line $\operatorname{Re} x = -c$ (orange, dashed) both go to the parabola $\operatorname{Re} y = c^2 - (\operatorname{Im} y)^2/(4c^2)$. The two black dots $\pm x_0$ go to the one black dot x_0^2. Each half of the x-plane, the right half (sheet 1) and the left half (sheet 2), covers the whole y-plane once, except for the negative real axis (red), which is the image of the imaginary axis where the two halves meet. So the square root $x = \sqrt{y}$ has two values, one on each sheet, and crossing the red slit takes it from one sheet to the other.

The square root needs two sheets. Now read the picture from right to left, as the inverse function $x = \sqrt{y}$. Every $y \neq 0$ has two square roots, $\pm x$, one in each half of the x-plane. The right half-plane, $-\frac{\pi}{2} < \theta < \frac{\pi}{2}$, is sent to $-\pi < 2\theta < \pi$: the whole y-plane except the negative real axis. The left half-plane is sent onto the same slit plane. So the two values of $\sqrt{y}$ are organised as two copies of the y-plane, called **sheets**, each cut along the negative real axis (the **slit**). The imaginary axis of the x-plane, where the two halves meet, is sent to the slit: $x = \pm it$ gives $y = -t^2$. Now walk once around the circle $|y| = \rho$, as $y = \rho\, e^{i\varphi}$ with φ running from 0 to 2π, and follow the square root that starts at $\sqrt{\rho}$ on sheet 1:

$$x(\varphi) = \sqrt{\rho}\, e^{i\varphi/2}, \qquad x(2\pi) \overset{(1)}{=} \sqrt{\rho}\, e^{i\pi} \overset{(2)}{=} -\sqrt{\rho} = -x(0).$$

(1) Put $\varphi = 2\pi$; half the angle of y is π.

(2) Euler's formula: $e^{i\pi} = \cos\pi + i\sin\pi = -1$.

One trip around $y = 0$ turns x only halfway around. We come back to the same y but to the other square root: crossing the slit moved us from sheet 1 to sheet 2. A second trip brings us back to sheet 1. So $\sqrt{y}$ is an honest, single-valued, continuous function once we glue the two

sheets *crosswise* along the slit: the upper edge of the slit on each sheet is joined to the lower edge on the other. The point $y = 0$, where the two values coincide and around which they swap, is a **branch point**. So is $y = \infty$: a huge circle also turns x only halfway around. Here the glued sheets are just a copy of the x-plane again.

The same construction, with the letters swapped, is the picture of the cubic curve $y^2 = x(x-1)(x-\lambda)$ in Figure 5.16. There $y = \pm\sqrt{q(x)}$, each of $x = 0, 1, \lambda, \infty$ is a branch point like $y = 0$ here (Example 5.19), and two sheets of the x-sphere are glued crosswise along two slits. With two slits instead of one, the glued sheets form a torus instead of a sphere.

Worked example 5.3: $y = 1/x$ turns the plane inside out

Put $x = r\,e^{i\theta}$ with $r > 0$. Then

$$y \overset{(1)}{=} \frac{1}{r\,e^{i\theta}} \overset{(2)}{=} \frac{1}{r}\,e^{-i\theta}.$$

(1) Definition of the map.
(2) $e^{i\theta}e^{-i\theta} = e^0 = 1$, so $1/e^{i\theta} = e^{-i\theta}$.

So $|y| = 1/r$ and $\arg y = -\theta$. The circle $|x| = r$ goes to the circle $|y| = 1/r$, traced in the opposite direction. The unit circle goes to itself, flipped: $e^{i\theta} \mapsto e^{-i\theta}$ is the reflection in the real axis. The inside of the unit disc goes to the outside, and the outside to the inside. Small circles around $x = 0$ go to huge circles: as $x \to 0$, y runs off to infinity. Unlike x^2, this map is one-to-one on $\mathbb{C} \setminus \{0\}$, and it is its own inverse. It only lacks somewhere to send $x = 0$, and somewhere to receive $x = \infty$. Add one point ∞ to each plane, with $1/0 = \infty$ and $1/\infty = 0$. Then $y = 1/x$ becomes a one-to-one map of the enlarged plane onto itself, which swaps 0 and ∞. The enlarged plane is the Riemann sphere $\mathbb{CP}^1$ of Example 5.9, and $1/x$ is exactly its change of chart near ∞. The complex circle $\{uv = 1\}$ of the example above is the graph of $v = 1/u$.

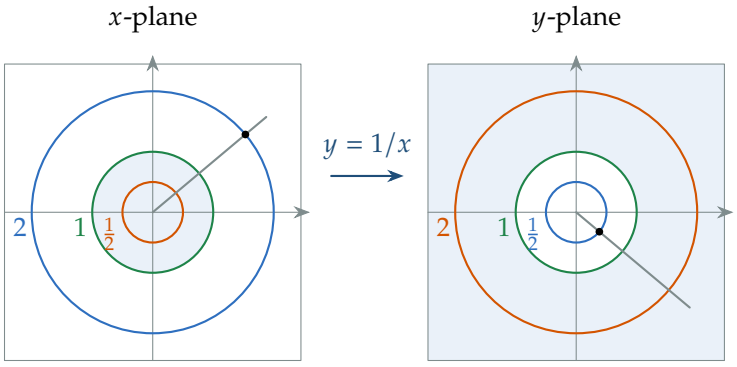


Figure 5.4: The map $y = 1/x$. The circle of radius r goes to the circle of radius $1/r$ (same colour; the radii are written on the negative real axis): the small orange circle goes to the large one, the large blue circle to the small one, and the green unit circle to itself. The shaded unit disc goes to the shaded outside. The gray ray at angle θ goes to the ray at angle $-\theta$, and the black dot $2e^{i\theta}$ goes to $\frac{1}{2}e^{-i\theta}$. The centre $x = 0$ has no image in the plane: it goes to the extra point ∞ of the Riemann sphere.

Try it

Picture $y = x^3$. What happens to circles and rays, and how many sheets does the cube root need?
Answer: $x = r\,e^{i\theta}$ gives $y = r^3 e^{3i\theta}$. Circles $|x| = r$ go to circles $|y| = r^3$, traced three times, and rays $\arg x = \theta$ go to rays $\arg y = 3\theta$. Each of the three sectors $-\frac{\pi}{3} < \theta - \frac{2\pi k}{3} < \frac{\pi}{3}$ ($k = 0, 1, 2$) covers the slit y-plane once, so $\sqrt[3]{y}$ needs three sheets, and one trip around $y = 0$ moves it on to the next sheet.

5.1.3 Real projective space: lines through the origin

Affine varieties are almost never compact: $\mathbb{C}^m$ itself runs off to infinity. Compact spaces are what we want (closed worldsheets, finite-dimensional spaces of sections, integrals that converge), so we enlarge $\mathbb{C}^m$ by adding points at infinity. The systematic way to do this is *projective space*.

Nash works with complex projective space $\mathbb{CP}^n$ from the start. It has real dimension $2n$, so already $\mathbb{CP}^2$ is four-dimensional and cannot be drawn. Its real cousin $\mathbb{RP}^n$ has real dimension n, and $\mathbb{RP}^1$ and $\mathbb{RP}^2$ can be drawn. Every idea we need (homogeneous coordinates, charts, points at infinity, curves that close up) already appears in $\mathbb{RP}^2$. So we build the pictures over $\mathbb{R}$ first, and then replace $\mathbb{R}$ by $\mathbb{C}$ in Section 5.1.6.

The starting question is an old one. A painter stands at a window and paints what she sees onto the glass. Railway tracks never meet, yet on her canvas they meet at a point of the horizon. Straight lines stay straight. A circular pond becomes an ellipse. Where, mathematically, is the point at which the rails meet? Projective geometry grew out of exactly this question (Desargues, in the seventeenth century; cf. the Poppe notes on projective space). The answer is that the rails meet at a point that is not in the plane, but in a slightly larger space. To build that space we change what we mean by "point".

Working definition 5.3 (real projective space)

A point of **real projective space** $\mathbb{RP}^n$ is a line through the origin of $\mathbb{R}^{n+1}$. We name it by any nonzero vector $x = (x_0, \dots, x_n)$ on the line, and remember that proportional vectors name the same point:

$$[x_0 : \cdots : x_n] = [t x_0 : \cdots : t x_n], \qquad t \in \mathbb{R}^* = \mathbb{R} \setminus \{0\}.$$

The x_i are **homogeneous coordinates**. *Operationally*: $[x] = [y]$ exactly when x and y are proportional, that is, when every 2×2 minor $x_i y_j - x_j y_i$ [the determinant of the 2×2 matrix built from entries i and j of x and of y] vanishes.

For $n = 1$ a point is a line through the origin of the plane; for $n = 2$ it is a line through the origin of space. The origin itself is not a line, so $[0 : \cdots : 0]$ is not a point.

Homogeneous coordinates record proportions. The colons are there to remind us that $[x : y : z]$ is a *ratio*, like "flour to sugar to butter = 2 : 1 : 1" in a recipe. Only the proportions $x : y : z$ carry information; the overall size does not. So

$$[x : y : z] = [\lambda x : \lambda y : \lambda z] \qquad \text{for every } \lambda \neq 0,$$

for example $[1 : 2 : 3] = [2 : 4 : 6] = [-\frac{1}{3} : -\frac{2}{3} : -1]$. Geometrically, the vectors $\lambda(x, y, z)$ with $\lambda \neq 0$ fill out the line through the origin in direction (x, y, z) (with the origin removed). So "a proportion class $[x : y : z]$" and "a line through the origin of $\mathbb{R}^3$" are the same thing, said in two ways.

Each line meets the sphere twice. Instead of the whole line we can record where it crosses the unit sphere $S^n = \{|x| = 1\} \subset \mathbb{R}^{n+1}$. The line through $v \neq 0$ is $\{tv : t \in \mathbb{R}\}$, and

$$|tv| = 1 \overset{(1)}{\iff} |t|\,|v| = 1 \overset{(2)}{\iff} t = \pm\frac{1}{|v|}.$$

(1) Euclidean length scales: $|tv| = |t|\,|v|$.
(2) The only real numbers of absolute value $1/|v|$ are $+1/|v|$ and $-1/|v|$.

So every line meets S^n in exactly two points, $\hat{v} = v/|v|$ and $-\hat{v}$. They are **antipodal**: diametrically opposite. The map

$$\pi : S^n \to \mathbb{RP}^n, \qquad x \mapsto [x],$$

is onto, and $\pi(x) = \pi(y)$ exactly when $y = \pm x$. Hence

$$\mathbb{RP}^n = S^n/(x \sim -x). \tag{5.1}$$

Two more facts follow at once. First, π is *locally* one-to-one: a small cap around x (angular radius less than $\pi/2$) never contains a pair of antipodal points, so π maps it bijectively onto its image. A map that is locally a homeomorphism and has exactly k preimages over every point is a **k-fold covering**. So π is 2-to-1, a **double cover**. (Covering spaces are treated in Chapter 1.) Second, $\mathbb{RP}^n$ is compact, because it is a continuous image of the compact sphere.

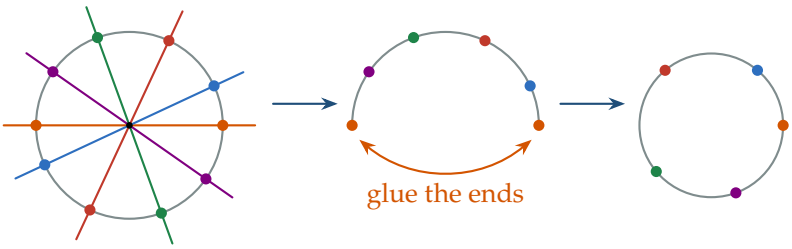


Figure 5.5: $\mathbb{RP}^1$ is a circle. Left: every line through the origin of $\mathbb{R}^2$ meets the unit circle in two antipodal points, drawn in the line's colour. Middle: keep one point per line by keeping the upper half-circle. Only the horizontal (orange) line still has both of its points, the two ends, and they must be glued. Right: an arc with its ends glued is a circle. The map that does this doubles angles (Example 5.4): the line at angle θ lands at 2θ.

Worked example 5.4: $\mathbb{RP}^1$ is a circle: angle doubling

A line through the origin of $\mathbb{R}^2$ has a direction $(\cos\theta, \sin\theta)$. The angles θ and $\theta + \pi$ give the same line, and each line has exactly one angle $\theta \in [0, \pi)$. Identify $\mathbb{R}^2$ with $\mathbb{C}$ and define

$$f : \mathbb{RP}^1 \to S^1 \subset \mathbb{C}, \qquad f([x : y]) = \frac{(x + \mathrm{i}y)^2}{x^2 + y^2}.$$

Well defined. Replacing (x, y) by (tx, ty) multiplies the numerator and the denominator by t^2, so f depends only on the line. *What it does.* Write $x = r\cos\theta$, $y = r\sin\theta$ with $r > 0$. Then

$$f([x : y]) \overset{(1)}{=} \frac{(re^{\mathrm{i}\theta})^2}{r^2} \overset{(2)}{=} e^{2\mathrm{i}\theta}$$

$$\overset{(3)}{=} \cos 2\theta + \mathrm{i}\sin 2\theta.$$

(1) $x + \mathrm{i}y = re^{\mathrm{i}\theta}$ and $x^2 + y^2 = r^2$.
(2) Cancel r^2.
(3) Euler's formula.

So f doubles the angle. As θ runs once over $[0, \pi)$, 2θ runs once over $[0, 2\pi)$. Hence f is a bijection, continuous in both directions, and $\mathbb{RP}^1 \cong S^1$. In real components, $f = \big((x^2 - y^2)/(x^2 + y^2),\ 2xy/(x^2 + y^2)\big)$.
The same map, read in a chart, shows the point at infinity. For lines with $x \neq 0$ use the slope $m = y/x$: it is the height at which the line crosses the vertical line $x = 1$. Then

$$f([1 : m]) \overset{(1)}{=} \frac{(1 + \mathrm{i}m)^2}{1 + m^2} \overset{(2)}{=} \frac{1 - m^2}{1 + m^2} + \mathrm{i}\,\frac{2m}{1 + m^2}.$$

(1) $[x : y] = [1 : m]$; substitute $x = 1$, $y = m$.
(2) Expand $(1 + \mathrm{i}m)^2 = 1 - m^2 + 2\mathrm{i}m$.

As $m \to +\infty$ *and* as $m \to -\infty$ this tends to $-1 = f([0 : 1])$, the image of the vertical line. The chart of slopes is a copy of $\mathbb{R}$ that misses one point, and both ends of $\mathbb{R}$ run into that point: $\mathbb{RP}^1 = \mathbb{R} \cup \{\infty\}$. (With $m = \tan\theta$ these are the half-angle formulas $\cos 2\theta = (1 - \tan^2\theta)/(1 + \tan^2\theta)$ and $\sin 2\theta = 2\tan\theta/(1 + \tan^2\theta)$.) Finally, composing f with π shows that the double cover $S^1 \to \mathbb{RP}^1 \cong S^1$ is $e^{\mathrm{i}\theta} \mapsto e^{2\mathrm{i}\theta}$, that is, $z \mapsto z^2$ on the unit circle.

Watch out: $\mathbb{RP}^1$ is a circle, but $\mathbb{RP}^2$ is not a sphere

The pattern breaks in dimension two. $S^2 \to \mathbb{RP}^2$ is a double cover of a genuinely new surface, one with no inside and no outside. We prove this with an Euler characteristic count in Example 5.5, and we see what it means for drawing in Section 5.1.7.

The disc model of $\mathbb{RP}^2$. Keep only the upper hemisphere $H = \{x \in S^2 : x_3 \geq 0\}$. A line through the origin of $\mathbb{R}^3$ that is not horizontal has exactly one of its two points $\pm\hat{v}$ strictly above the equator, so it meets H once, in the interior. A horizontal line meets H in both of its points $\pm\hat{v}$, and both lie on the equator. Now flatten H onto the unit disc by forgetting x_3.

Working definition 5.4 (disc model of $\mathbb{RP}^2$)

$\mathbb{RP}^2$ is the closed unit disc $D = \{u \in \mathbb{R}^2 : |u| \leq 1\}$ with each boundary point u glued to its antipode $-u$. [the diametrically opposite point of the boundary circle] Interior points of D are the non-horizontal lines; each pair $\pm u$ on the boundary is one horizontal line. *Operationally*: the point $[x_1 : x_2 : x_3]$ with $x_3 \neq 0$ sits at $u = \operatorname{sgn}(x_3)\,(x_1, x_2)/|x|$, [$\operatorname{sgn}(x_3) = \pm 1$ flips the representative so that $x_3 > 0$, the upper hemisphere] and a curve that leaves the disc at u re-enters at $-u$.

This is the construction of Figure 5.5 one dimension up. There the upper half-circle, flattened, is the interval $[-1, 1]$ with its two ends glued, which is a circle.

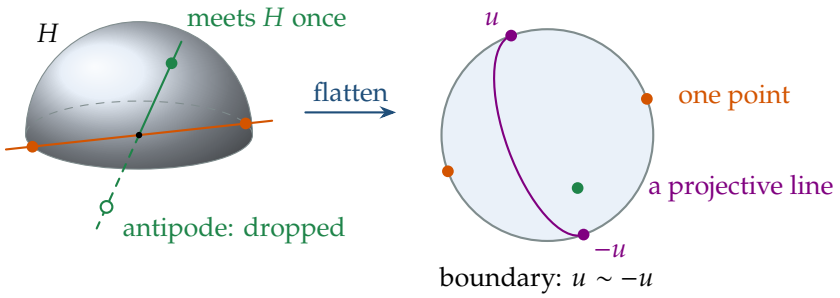


Figure 5.6: The disc model of $\mathbb{RP}^2$. Left: a non-horizontal line through the origin (green) meets the upper hemisphere H once; its other point lies below and is dropped. A horizontal line (orange) meets H twice, at antipodal points of the equator. Right: flattened, H becomes a disc whose boundary points are glued in antipodal pairs, so the two orange dots are a single point of $\mathbb{RP}^2$. A projective line (violet, the image of a great circle) runs from u to $-u$; since those are the same point, it is a closed loop.

Worked example 5.5: The Euler characteristic of $\mathbb{RP}^2$ is 1

From the double cover. The octahedron with vertices $\pm e_1, \pm e_2, \pm e_3$ is a triangulation of S^2 that the antipodal map carries to itself, and no vertex, edge or face is carried to itself. So its 6 vertices, 12 edges and 8 faces pair off into 3 vertices, 6 edges and 4 faces of $\mathbb{RP}^2$:

$$\begin{aligned}\chi(\mathbb{RP}^2) &\overset{(1)}{=} \tfrac{1}{2}V - \tfrac{1}{2}E + \tfrac{1}{2}F \\ &\overset{(2)}{=} \tfrac{1}{2}(6 - 12 + 8) \\ &\overset{(3)}{=} \tfrac{1}{2}\chi(S^2) = 1.\end{aligned}$$

(1) Each cell of $\mathbb{RP}^2$ has exactly two preimages, which $x \mapsto -x$ swaps.
(2) Count the vertices, edges and faces of the octahedron.
(3) $6 - 12 + 8 = 2 = \chi(S^2)$.

From the disc model. The disc model has one 2-cell (the open disc), one 1-cell (the open upper half of the boundary circle, since the lower half is glued onto it) and one 0-cell (the two points $(\pm 1, 0)$, glued into one). So $V - E + F = 1 - 1 + 1 = 1$ again.
We will see in (5.7) that every closed *orientable* surface has $\chi = 2 - 2p$, an even number. Since $\chi(\mathbb{RP}^2) = 1$ is odd, $\mathbb{RP}^2$ is a closed surface that is not orientable. (A double cover without branch points halves χ. With branch points it does not: that is the count of Example 5.19 for the cubic.)

5.1.4 From the plane to $\mathbb{RP}^2$: perspective and the line at infinity

Return to the painter. Put her eye at the origin of $\mathbb{R}^3$, with coordinates (X, Y, Z), and put the canvas in the plane $\{Z = 1\}$. The canvas is a copy of $\mathbb{R}^2$: the point (x, y) of the plane is the point $(x, y, 1)$ of space. She sees it along the ray of light from her eye through $(x, y, 1)$. That ray lies on exactly one line through the origin, so we get a map

$$j : \mathbb{R}^2 \to \mathbb{RP}^2, \qquad (x, y) \mapsto [x : y : 1].$$

Different points of the canvas lie on different lines, so j is one-to-one. A line through the origin crosses the canvas exactly when it is not horizontal, that is, when its Z-component is nonzero, and then it crosses at $[X : Y : Z] = [X/Z : Y/Z : 1]$.

The painter's picture is exactly the construction of $\mathbb{RP}^2$: the eye is the origin, the canvas is the plane $Z = 1$, and the rays of light are the lines through the origin. Here is the construction step by step.

Key idea: building $\mathbb{RP}^2$ from the plane in four steps

1. *Embed.* Put $\mathbb{R}^2$ in $\mathbb{R}^3$ as the plane $Z = 1$: $(x, y) \mapsto (x, y, 1)$.
2. *Replace points by lines.* Identify each point $(x, y, 1)$ with the line through the origin of $\mathbb{R}^3$ that passes through it.
3. *Name the lines by proportions.* Lines through the origin are exactly the proportion classes $[X : Y : Z]$. The point (x, y) becomes $[x : y : 1]$, and any $[X : Y : Z]$ with $Z \neq 0$ comes from exactly one point, $(X/Z, Y/Z)$.
4. *Add what is missing.* The lines through the origin lying in the plane $Z = 0$ never meet the plane $Z = 1$, because every point on them has $Z = 0 \neq 1$. These lines, $[X : Y : 0]$, are the new **points at infinity**. They are the lines through the origin of a plane, so they form a whole $\mathbb{RP}^1$ of directions: the **line at infinity**.

Step 4 is where a point that runs off to infinity ends up. Let the point (t, mt) run off along the line $y = mx$:

$$j(t, mt) \overset{(1)}{=} [t : mt : 1] \overset{(2)}{=} \left[1 : m : \tfrac{1}{t}\right] \overset{(3)}{\longrightarrow} [1 : m : 0] \qquad (t \to \infty).$$

(1) Step 3: the point (t, mt) is the line through $(t, mt, 1)$.
(2) Divide all three entries by $t \neq 0$; the proportions are unchanged.
(3) $1/t \to 0$. The limit $[1 : m : 0]$ is the line in the plane $Z = 0$ with direction $(1, m)$: the point at infinity of every line of slope m.

Working definition 5.5 (affine chart and line at infinity of $\mathbb{RP}^2$)

In homogeneous coordinates $[X : Y : Z]$, [a nonzero triple, with proportional triples naming the same point] the **affine chart** $U_Z = \{Z \neq 0\}$ is identified with $\mathbb{R}^2$ by

$$[X : Y : Z] \mapsto (x, y) = \left(\frac{X}{Z}, \frac{Y}{Z}\right), \qquad \text{with inverse} \qquad (x, y) \mapsto [x : y : 1].$$

The rest of $\mathbb{RP}^2$ is the **line at infinity** $L_\infty = \{Z = 0\} = \{[X : Y : 0]\}$. *Operationally*: to find where a point of $\mathbb{RP}^2$ sits in the plane, divide by Z. If $Z = 0$, the point is not a place in the plane but a direction (X, Y) in it, and (X, Y) and $(-X, -Y)$ count as the same direction.

The chart is well defined: rescaling (X, Y, Z) by t changes neither X/Z nor Y/Z. The line at infinity is itself a projective line. Its points $[X : Y : 0]$ are the lines through the origin of the horizontal plane $\{Z = 0\}$, so $L_\infty \cong \mathbb{RP}^1 \cong S^1$. Altogether

$$\mathbb{RP}^2 = \mathbb{R}^2 \sqcup L_\infty = \mathbb{R}^2 \sqcup \mathbb{RP}^1,$$

just as $\mathbb{RP}^1 = \mathbb{R} \sqcup \{\text{point}\}$. In general $\mathbb{RP}^n = \mathbb{R}^n \sqcup \mathbb{RP}^{n-1}$. The other two charts, $U_X = \{X \neq 0\}$ and $U_Y = \{Y \neq 0\}$, are also copies of $\mathbb{R}^2$, and the three charts together cover $\mathbb{RP}^2$.

Running off to infinity. Walk away from the origin of the canvas in a fixed direction $(a, b) \neq (0, 0)$, that is, along $(x, y) = (ta, tb)$ with $t \to \infty$. Then

$$j(ta, tb) \overset{(1)}{=} [ta : tb : 1]$$

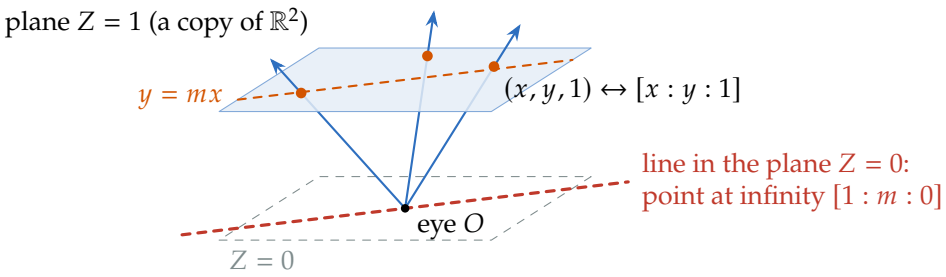


Figure 5.7: Perspective. The eye sits at the origin and the plane $Z = 1$ is a canvas. Each point (x, y) of the canvas (orange) is seen along one ray (blue), so it names one line through the origin, $[x : y : 1]$. As the canvas point runs off to infinity, its ray tilts towards the horizontal. A horizontal ray (red, dashed), lying in the plane $Z = 0$, is still a line through the origin, so it is still a point of $\mathbb{RP}^2$, but it meets no point of the canvas: it is a point at infinity. The red line has the direction $(1, m)$ of the canvas line $y = mx$ (orange, dashed), so it is the point $[1 : m : 0]$ where that line ends up. All the lines through the origin in the plane $Z = 0$ together form the line at infinity L_∞.

$$\overset{(2)}{=} \left[a : b : \tfrac{1}{t}\right]$$
$$\overset{(3)}{\longrightarrow} [a : b : 0] \qquad (t \to \pm\infty).$$

(1) Definition of j.
(2) Divide all three coordinates by $t \neq 0$; this names the same point.
(3) $1/t \to 0$ whether $t \to +\infty$ or $t \to -\infty$, and $[a : b : 0]$ is a genuine point because $(a, b) \neq 0$.

In the picture: as the canvas point runs away, its ray tilts down towards the horizontal, and in the limit it becomes the horizontal line in direction (a, b). Step (3) contains the key point. Walking off in direction (a, b) and walking off in the opposite direction $(-a, -b)$ bring us to the *same* point at infinity, because $[a : b : 0] = [-a : -b : 0]$. Each direction, without its sign, is one point of L_∞.

Lines. Take the line $ax + by + c = 0$ in the plane, with $(a, b) \neq 0$. Substitute $x = X/Z$, $y = Y/Z$ and multiply by Z:

$$aX + bY + cZ = 0.$$

This is a plane through the origin of $\mathbb{R}^3$ with normal vector $n = (a, b, c)$. So a **projective line** is a plane through the origin. As a set of lines through the origin it is an $\mathbb{RP}^1$, a circle. It consists of the old affine line plus one point at infinity: setting $Z = 0$ forces $aX + bY = 0$, so the extra point is $[-b : a : 0]$, the direction of the line. Two distinct planes through the origin always meet in exactly one line through the origin. Therefore:

Key idea: two lines always meet

Any two distinct lines in $\mathbb{RP}^2$ meet in exactly one point. Parallel lines are no exception: they meet on the line at infinity, at the point named by their common direction. *Operationally*: the lines with normals n_1 and n_2 meet at $[n_1 \times n_2]$, because the cross product is orthogonal to both normals and so lies in both planes.

Worked example 5.6: Parallel lines meet at $[1 : m : 0]$

Take $y = mx + c_1$ and $y = mx + c_2$ with $c_1 \neq c_2$. Homogenised, they are

$$Y = mX + c_1 Z, \qquad Y = mX + c_2 Z.$$

A common point $[X : Y : Z]$ satisfies both equations, so

$$\begin{aligned} 0 &\overset{(1)}{=} (mX + c_1 Z) - (mX + c_2 Z) \\ &\overset{(2)}{=} (c_1 - c_2)\, Z, \end{aligned}$$

(1) Subtract the two equations; both left-hand sides equal Y.
(2) Collect terms; the mX cancels.

and $c_1 \neq c_2$ forces $Z = 0$. Then either equation gives $Y = mX$, and

$$\begin{aligned} [X : Y : Z] &\overset{(1)}{=} [X : mX : 0] \\ &\overset{(2)}{=} [1 : m : 0]. \end{aligned}$$

(1) $Y = mX$ and $Z = 0$.
(2) $X \neq 0$, since otherwise all three coordinates would vanish; divide by X.

So the two lines meet in exactly one point, $[1 : m : 0]$. It does not depend on c_1 or c_2: every line of slope m passes through it. Vertical lines $x = c$ homogenise to $X = cZ$, and any two of them meet where $Z = 0$ and $X = 0$, that is, at $[0 : 1 : 0]$, the point of "slope ∞". So the points of L_∞ correspond one-to-one to the slopes $m \in \mathbb{R} \cup \{\infty\}$, which is another way to see that $L_\infty \cong \mathbb{RP}^1$.

As a check, use the cross-product rule. Written as $mx - y + c_i = 0$, the lines have normals $n_i = (m, -1, c_i)$, and

$$\begin{aligned} n_1 \times n_2 &\overset{(1)}{=} \big((-1)c_2 - c_1(-1),\ c_1 m - m c_2,\ m(-1) - (-1)m\big) \\ &\overset{(2)}{=} (c_1 - c_2)\,(1, m, 0). \end{aligned}$$

(1) For $n_i = (a_i, b_i, e_i)$ the cross product is $(b_1 e_2 - e_1 b_2,\ e_1 a_2 - a_1 e_2,\ a_1 b_2 - b_1 a_2)$.
(2) Factor out $c_1 - c_2$.

A tiled floor in perspective. A painter sees the same fact on every tiled floor. Stand on a floor with the eye at height h, looking straight ahead along the y-axis, with x the sideways coordinate. Lay square tiles turned by 45°, so that their edges are two families of parallel lines,

$$y - x = l \quad (\text{slope } 1), \qquad x + y = k \quad (\text{slope } -1), \qquad k, l \in \mathbb{Z}.$$

A tile corner is the crossing point of the edge $x + y = k$ with the edge $y - x = l$. Adding and subtracting the two equations puts it at $x = \frac{k-l}{2}$, $y = \frac{k+l}{2}$; we call it the corner (k, l). The tile with lowest corner (k, l) has the four corners (k, l), $(k+1, l)$, $(k+1, l+1)$, $(k, l+1)$. As in Example 5.11 (with the picture enlarged by a factor S), the ground point (x, y) with $y > 0$ is painted at

$$(u, v) = \Big(S\,\frac{x}{y},\ -\frac{Sh}{y}\Big).$$

A point of the edge $y - x = l$ has the form $(x, y) = (y - l,\ y)$. It is painted at

$$\begin{aligned} (u, v) &\overset{(1)}{=} \Big(S\,\frac{y - l}{y},\ -\frac{Sh}{y}\Big) \\ &\overset{(2)}{=} \Big(S - \frac{l}{h}\,\frac{Sh}{y},\ -\frac{Sh}{y}\Big) \\ &\overset{(3)}{=} \Big(S + \frac{l}{h}\,v,\ v\Big). \end{aligned}$$

(1) The painting map, with $x = y - l$.
(2) Split the fraction, $\frac{y-l}{y} = 1 - \frac{l}{y}$, and write $\frac{l}{y} = \frac{l}{h} \cdot \frac{h}{y}$.
(3) Write $v = -Sh/y$.

So the painted edge is the straight line $u = S + (l/h)\,v$. At eye level, $v = 0$, it passes through $u = S$, whatever l is: every edge of this family runs to the single point $V = (S, 0)$ of the horizon. It gets there only in the limit $y \to \infty$, where $v = -Sh/y \to 0$. The same computation with $x = k - y$ gives $u = -S - (k/h)\,v$, so every edge $x + y = k$ runs to the point $V' = (-S, 0)$.

Now the diagonals. The diagonal of a tile from (k, l) to $(k + 1, l + 1)$ keeps $x = \frac{k-l}{2}$ fixed, so it lies on the ground line $x = \frac{j}{2}$ with $j = k - l$. These lines point straight ahead, and they form a third family of parallels. With $x = \frac{j}{2}$ fixed, the painted point is $(u, v) = \left(\frac{Sj}{2y}, -\frac{Sh}{y}\right) = \left(-\frac{j}{2h}\,v,\ v\right)$, a straight line through $D = (0, 0)$. So the diagonals run to a third vanishing point D, straight ahead. The other diagonal of each tile, from $(k + 1, l)$ to $(k, l + 1)$, keeps y fixed. It is painted as a horizontal line $v = -Sh/y$, parallel to the horizon, and its vanishing point lies at infinity along the horizon. In the language of Example 5.6: each direction on the ground has its own point at infinity, $[1 : m : 0]$ for the lines of slope m and $[0 : 1 : 0]$ for the lines straight ahead, and all of these points lie on one line, the horizon, which is the line at infinity L_∞ of the ground.

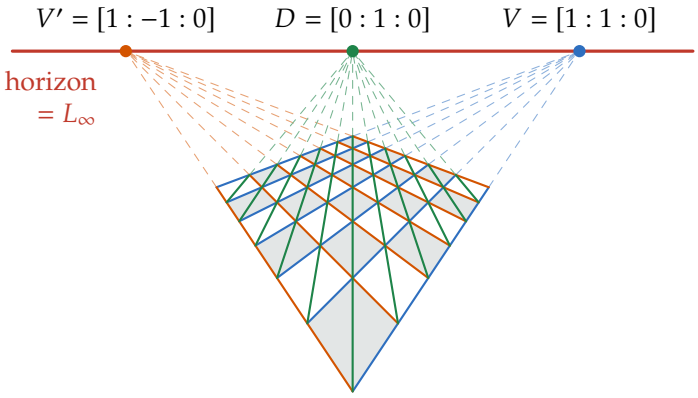


Figure 5.8: Parallel lines meet on the horizon: a tiled floor in two-point perspective. The ground point at distance y ahead and x to the side is painted at $(Sx/y,\ -Sh/y)$, with eye height $h = 3$. The tile edges form two families of parallel lines, $y - x = l$ (blue) and $x + y = k$ (orange). Each family runs, solid on the floor and dashed beyond it, to its own vanishing point: $V = [1 : 1 : 0]$ for the blue edges (slope 1) and $V' = [1 : -1 : 0]$ for the orange ones (slope -1). The green lines are the diagonals of the tiles: each passes through opposite corners of every tile it crosses. They are parallel on the ground (straight ahead, $x = \text{const}$), so they meet at a third vanishing point $D = [0 : 1 : 0]$. The other diagonals, $y = \text{const}$, would be drawn parallel to the horizon; their vanishing point $[1 : 0 : 0]$ is at infinity along the horizon. Each direction on the ground is one point of the horizon, and all of them lie on one line: the horizon is the line at infinity of the ground, made visible (Example 5.11).

Physical picture: the horizon is a genuine line

A painter's horizon is the line at infinity of the ground plane, drawn at a finite place. Every family of parallel lines on the ground runs to its own vanishing point, and all the vanishing points lie on one line. In the projective plane the vanishing points are points like any others. Painters build a whole tiled floor from this one rule. Draw the two families of tile edges to V and V', then draw one diagonal to D. The diagonal crosses the edges exactly at tile corners, which fixes where each new row of tiles begins, and so every tile of the floor follows. We make "drawn at a finite place" precise in Example 5.11.

5.1.5 Conics in the projective plane: closing curves at infinity

Now put a curve in the plane and ask what happens to it at infinity. The recipe is the one we used for lines: substitute $x = X/Z$, $y = Y/Z$, clear the denominators, and then look at $Z = 0$.

Worked example 5.7: The parabola closes up into a circle

Homogenise. On the chart $Z \neq 0$,

$$y = x^2 \overset{(1)}{\iff} \frac{Y}{Z} = \frac{X^2}{Z^2} \overset{(2)}{\iff} YZ = X^2.$$

(1) Substitute $x = X/Z$ and $y = Y/Z$.
(2) Multiply by $Z^2 \neq 0$.

Both sides of $YZ = X^2$ are homogeneous of degree 2 (rescaling by t multiplies each by t^2), so $C = \{YZ = X^2\}$ is a well-defined curve in $\mathbb{RP}^2$.
Points at infinity. Put $Z = 0$. Then $X^2 = 0$, so $X = 0$ and $Y \neq 0$. There is exactly one point, $[0 : 1 : 0]$, the common point of all vertical lines. The equation $X^2 = 0$ has a *double* root there, which means that C touches L_∞ rather than crossing it.
Both arms go there. Along the curve $(x, y) = (t, t^2)$, and

$$[t : t^2 : 1] \overset{(1)}{=} \left[\tfrac{1}{t} : 1 : \tfrac{1}{t^2}\right] \overset{(2)}{\longrightarrow} [0 : 1 : 0] \qquad (t \to \pm\infty).$$

(1) Divide all coordinates by t^2.
(2) $1/t \to 0$ and $1/t^2 \to 0$ for either sign of t.

The right arm ($t \to +\infty$) and the left arm ($t \to -\infty$) both run off in the vertical direction, and "vertically up" and "vertically down" are the same point at infinity. So the two arms are glued at $[0 : 1 : 0]$, and C is the line of parameters t plus one point: a closed loop.
A chart in which the loop is visible. Change coordinates by

$$Y = U + V, \qquad Z = U - V \qquad \left(\text{so } U = \tfrac{Y+Z}{2},\ V = \tfrac{Y-Z}{2}\right).$$

This is an invertible linear map of $\mathbb{R}^3$, so it only relabels the points of $\mathbb{RP}^2$. Then

$$YZ - X^2 \overset{(1)}{=} (U + V)(U - V) - X^2 \overset{(2)}{=} U^2 - V^2 - X^2,$$

(1) Substitute $Y = U + V$ and $Z = U - V$.
(2) Difference of two squares.

so the curve is $U^2 = X^2 + V^2$. In the chart $U \neq 0$, with coordinates $\xi = X/U$ and $\eta = V/U$, divide by U^2:

$$\xi^2 + \eta^2 = 1.$$

The parabola *is* a circle, seen in another chart. To see which point goes where, take the affine point $[x : x^2 : 1]$. Then $U = (x^2 + 1)/2$ and $V = (x^2 - 1)/2$, so

$$\xi = \frac{2x}{x^2 + 1}, \qquad \eta = \frac{x^2 - 1}{x^2 + 1},$$

and indeed

$$\xi^2+\eta^2 \overset{(1)}{=} \frac{4x^2+(x^2-1)^2}{(x^2+1)^2} \overset{(2)}{=} \frac{x^4+2x^2+1}{(x^2+1)^2} \overset{(3)}{=} 1.$$

(1) Put both terms over the common denominator.
(2) Expand $(x^2-1)^2 = x^4-2x^2+1$ and add $4x^2$.
(3) $x^4+2x^2+1 = (x^2+1)^2$.

The vertex $x=0$ goes to $(0,-1)$ and the points $x=\pm1$ go to $(\pm1,0)$. As $x\to\pm\infty$ both arms approach $(0,1)$, which is the point at infinity $[X:Y:Z]=[0:1:0]$: it has $X=0$ and $Z=0$, so $V=U$. The old line at infinity $Z=0$ is $U=V$, the horizontal line $\eta=1$, which touches the circle at its top. The new chart misses only the line $U=0$, that is $Y=-Z$, which in the old chart is the line $y=-1$. The parabola never reaches $y=-1$, and that is why the whole closed curve fits into the new chart. (The map $x\mapsto(\xi,\eta)$ is the inverse of stereographic projection from the top point of the circle. It is the same map that identifies $\mathbb{CP}^1$ with S^2 in Example 5.9, one dimension down.)

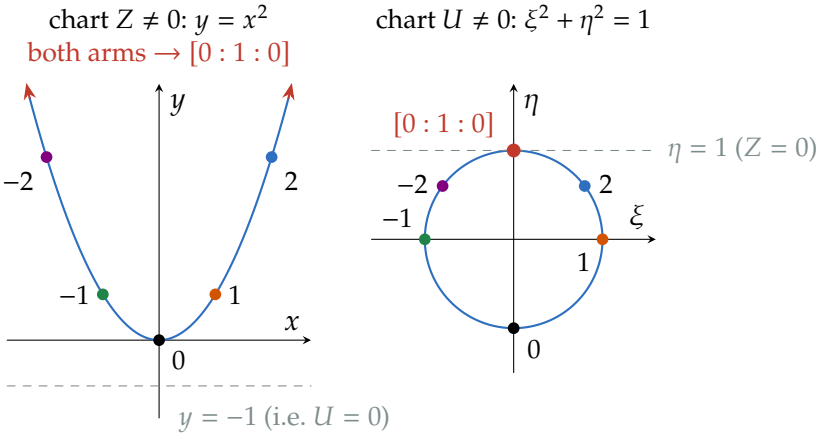


Figure 5.9: One curve, two charts. Left: the parabola $y=x^2$ in the chart $Z\neq0$, with the points $x=-2,\dots,2$ marked. Its point at infinity is not in this chart; both arms run towards it (red arrows). Right: the same curve in the chart $U\neq0$ of Example 5.7, where it is the unit circle. The marked points reappear in the same colours, and the point at infinity $[0:1:0]$ is the red dot at the top. The dashed line on the right is the line at infinity of the left chart, which touches the circle there. The dashed line on the left, $y=-1$, is the line at infinity of the right chart. The parabola never meets it, so the right chart shows the whole curve.

Worked example 5.8: The hyperbola: two branches, one loop

Homogenise $xy=1$:

$$xy=1 \overset{(1)}{\iff} \frac{X}{Z}\cdot\frac{Y}{Z}=1 \overset{(2)}{\iff} XY=Z^2.$$

(1) Substitute $x=X/Z$ and $y=Y/Z$.

(2) Multiply by $Z^2 \neq 0$.

At $Z = 0$ we need $XY = 0$. This gives *two* points at infinity, $[1:0:0]$ and $[0:1:0]$, the directions of the two asymptotes (the x- and y-axes). Parametrise the curve by $(t, 1/t)$ with $t \neq 0$:

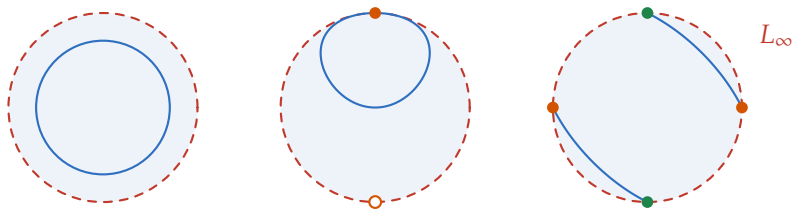
$$\left[t : \tfrac{1}{t} : 1\right] \overset{(1)}{=} \left[1 : \tfrac{1}{t^2} : \tfrac{1}{t}\right] \longrightarrow [1:0:0] \qquad (t \to \pm\infty),$$
$$\left[t : \tfrac{1}{t} : 1\right] \overset{(2)}{=} \left[t^2 : 1 : t\right] \longrightarrow [0:1:0] \qquad (t \to 0^{\pm}).$$

(1) Divide all coordinates by t.

(2) Multiply all coordinates by t.

Now follow the curve. Start on the first-quadrant branch high up near the y-axis ($t \to 0^+$, close to $[0:1:0]$) and move to the right. As $t \to +\infty$ we reach $[1:0:0]$. We pass through that point and come back from the far left on the third-quadrant branch, with t increasing from $-\infty$. As $t \to 0^-$ we run down along the negative y-axis into $[0:1:0]$, where we started. The two branches and the two points at infinity form *one* closed loop.

Two changes of chart make this visible. In the chart $Y \neq 0$, with coordinates $s = X/Y$ and $r = Z/Y$, the equation $XY = Z^2$ becomes $s = r^2$: a parabola, whose only point at infinity is $[1:0:0]$ and whose vertex is $[0:1:0]$. With $X = U + V$ and $Y = U - V$ it becomes $U^2 = V^2 + Z^2$, a circle in the chart $U \neq 0$, exactly as for the parabola.



ellipse $x^2 + y^2 = 1$: misses L_∞ — parabola $y = x^2$: touches L_∞ once — hyperbola $xy = 1$: meets L_∞ twice

Figure 5.10: Three conics in the disc model of $\mathbb{RP}^2$. The chart $Z \neq 0$ is drawn as the open disc, by sending (x, y) to $(x, y)/\sqrt{x^2 + y^2 + 1}$ (the point where the ray through $(x, y, 1)$ meets the upper hemisphere, flattened). The boundary circle (red, dashed) is L_∞, with antipodal points glued. The ellipse stays inside. Both arms of the parabola reach the boundary at the top, and the orange dot at the bottom is the same point. Each branch of the hyperbola runs from an orange boundary point to a green one; since opposite boundary points are glued, the two branches join into one closed loop. In $\mathbb{RP}^2$ all three are the same curve.

Key idea: all non-degenerate conics are the same in $\mathbb{RP}^2$

A conic is the zero set of a quadratic form $Q(X, Y, Z) = w^{\mathsf{T}} M w$, with $w = (X, Y, Z)^{\mathsf{T}}$ and M a real symmetric 3×3 matrix. It is **non-degenerate** if $\det M \neq 0$. By Sylvester's law of inertia, an invertible linear change of coordinates brings M to $\mathrm{diag}(\pm 1, \pm 1, \pm 1)$, and $Q = 0$ and $-Q = 0$ are the same curve. So there are only two cases: $X^2 + Y^2 + Z^2 = 0$, which has no real points, and

$$X^2 + Y^2 = Z^2,$$

the circle. *Every non-degenerate real conic with a real point is the circle, seen in some chart.* Ellipse, parabola and hyperbola differ only in how the circle sits relative to the one line we chose to call L_∞: it misses that line, touches it, or cuts it twice.

For example, $YZ - X^2$ has $M = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 0 & 1/2 \\ 0 & 1/2 & 0 \end{pmatrix}$, with $\det M = \frac{1}{4} \neq 0$ and eigenvalues $-1, \frac{1}{2}, -\frac{1}{2}$: one sign differs from the other two, as for $X^2 + Y^2 - Z^2$. The three affine types compare as follows (cf. the table in the illustrated perspective guide):

affine conic	homogenised	meets L_∞ ($Z = 0$) where	real points at ∞
ellipse $x^2 + y^2 = 1$	$X^2 + Y^2 = Z^2$	$X^2 + Y^2 = 0$: $[1 : \pm i : 0]$	none
parabola $y = x^2$	$YZ = X^2$	$X^2 = 0$: $[0 : 1 : 0]$, double	one (tangent)
hyperbola $xy = 1$	$XY = Z^2$	$XY = 0$: $[1 : 0 : 0]$, $[0 : 1 : 0]$	two

The ellipse has no real points at infinity, but over $\mathbb{C}$ it has two, $[1 : \pm i : 0]$. These are the two points we add to the complex circle in Example 5.13, where they are written $[0 : 1 : \pm i]$ because there the coordinate z_0 (our Z) is listed first. Over $\mathbb{C}$ the distinction disappears completely: every non-degenerate conic in $\mathbb{CP}^2$ is carried to $z_0 z_2 = z_1^2$ by a linear change of coordinates, and so is a copy of $\mathbb{CP}^1$. The degenerate conics, such as $xy = 0$ (two lines) or $x^2 = 0$ (a doubled line), have $\det M = 0$ and really are different.

Physical picture: adding points at infinity closes curves up

Over $\mathbb{R}$: a line gains one point at infinity and becomes a circle. The parabola gains one point and the hyperbola two, and each becomes a single closed loop. Over $\mathbb{C}$ the same thing happens one dimension up, and this is the mechanism behind Nash's torus. The affine cubic $y^2 = x(x-1)(x-\lambda)$ in $\mathbb{C}^2$ is a torus with one point missing. Homogenised, it is $Y^2Z = X(X - Z)(X - \lambda Z)$, and at $Z = 0$ this forces $X^3 = 0$: there is a single point at infinity, $[0 : 1 : 0]$, the vertical direction, just as for the parabola. Every far-away point of the curve runs into it: dividing by y, $[x : y : 1] = [x/y : 1 : 1/y]$, and since $|y| \sim |x|^{3/2}$ both x/y and $1/y$ tend to 0. Adding this one point closes the surface up into a torus (Example 5.19 and Figure 5.16). The real picture shows it too. The unbounded real branch of the smooth cubic in Figure 5.2 is closed up through $[0 : 1 : 0]$, so in $\mathbb{RP}^2$ the real points of that cubic form two loops, the oval and the completed branch.

5.1.6 Complex projective space

Everything in the last three subsections works with $\mathbb{C}$ in place of $\mathbb{R}$. The algebra is unchanged. What changes is the pictures, because each complex dimension counts as two real ones.

Working definition 5.6 (complex projective space)

A point of $\mathbb{CP}^n$ is a complex line through the origin of $\mathbb{C}^{n+1}$. [all complex multiples λv of one nonzero vector v. As a real set it is a plane] We name it by any nonzero vector on that line, $(z_1, \dots, z_{n+1}) \neq 0$, and remember that proportional vectors name the same point:

$$[z_1 : \cdots : z_{n+1}] = [\lambda z_1 : \cdots : \lambda z_{n+1}], \qquad \lambda \in \mathbb{C}^*. \, [\, anynonzerocomplexnumber \,]$$

The z_i are **homogeneous coordinates**. Only their ratios mean anything. *Operationally*: to decide whether two tuples are the same point, check whether they are proportional.

Nash writes the homogeneous coordinates as $(z_1, \dots, z_{n+1})$. Later, for the Fubini–Study metric, he switches to $(z_0, z_1, \dots, z_n)$; we will do the same when we get there.

Charts. Ratios give local coordinates. Let $U_j = \{z_j \neq 0\} \subset \mathbb{CP}^n$. On U_j we can divide by z_j, and the n ratios z_i/z_j ($i \neq j$) determine the point uniquely, because the representative with jth

entry equal to 1 is unique. So

$$\phi_j : U_j \to \mathbb{C}^n, \qquad [z_1 : \cdots : z_{n+1}] \mapsto \left(\tfrac{z_1}{z_j}, \ldots, \widehat{\tfrac{z_j}{z_j}}, \ldots, \tfrac{z_{n+1}}{z_j} \right)$$

(the hat means "omit") is a bijection. Every point has some nonzero coordinate, so the $n+1$ sets U_j cover $\mathbb{CP}^n$. Nash's example is U_{n+1}: if $z_{n+1} \neq 0$ the point is $[z_1/z_{n+1} : \cdots : z_n/z_{n+1} : 1]$.

Transition functions are holomorphic. On the overlap $U_j \cap U_k$ the two sets of coordinates are related by multiplying by z_j/z_k, which is itself one of the k-coordinates. Written out,

$$\begin{aligned} \frac{z_i}{z_k} &\overset{(1)}{=} \frac{z_i}{z_j} \cdot \frac{z_j}{z_k} \\ &\overset{(2)}{=} \frac{z_i/z_j}{z_k/z_j}. \end{aligned}$$

(1) Multiply and divide by z_j, allowed on U_j.
(2) Rewrite $z_j/z_k = 1/(z_k/z_j)$, allowed on U_k.

So each k-coordinate is a ratio of j-coordinates with a denominator z_k/z_j that does not vanish on the overlap: a holomorphic function. This makes $\mathbb{CP}^n$ a **complex manifold** of complex dimension n.

Compactness, and circles instead of pairs. Every complex line through the origin meets the unit sphere $S^{2n+1} = \{\sum |z_i|^2 = 1\} \subset \mathbb{C}^{n+1} = \mathbb{R}^{2n+2}$. In the real case the line met its sphere in a pair of antipodal points. In the complex case it meets it in a whole circle. The complex line through $v \neq 0$ is $\{\lambda v : \lambda \in \mathbb{C}\}$, and

$$\begin{aligned} |\lambda v| = 1 &\overset{(1)}{\iff} |\lambda| = \frac{1}{|v|} \\ &\overset{(2)}{\iff} \lambda = \frac{e^{i\alpha}}{|v|} \text{ for some } \alpha \in [0, 2\pi). \end{aligned}$$

(1) $|\lambda v| = |\lambda|\,|v|$ for the Hermitian length $|v|^2 = \sum_i |v_i|^2$.
(2) A complex number of modulus $1/|v|$ is $1/|v|$ times a phase $e^{i\alpha}$. Over $\mathbb{R}$ the only "phases" are ± 1, which gave the two points of (5.1).

So the line meets the sphere in the circle $\{e^{i\alpha}\hat{v}\}$, and

$$\mathbb{CP}^n = S^{2n+1}/\mathrm{U}(1), \qquad z \sim e^{i\alpha} z. \tag{5.2}$$

In particular $\mathbb{CP}^n$ is a continuous image of a compact sphere, hence compact. The map $S^{2n+1} \to \mathbb{CP}^n$ has circles as its fibres, and counting dimensions,

$$\begin{aligned} \dim_{\mathbb{R}} \mathbb{CP}^n &\overset{(1)}{=} \dim_{\mathbb{R}} S^{2n+1} - \dim_{\mathbb{R}} S^1 \\ &\overset{(2)}{=} (2n+1) - 1 = 2n. \end{aligned}$$

(1) Collapsing each circle fibre to a point removes one real dimension.
(2) S^{2n+1} is the unit sphere in $\mathbb{R}^{2n+2}$.

This agrees with the charts: complex dimension n is real dimension $2n$. Side by side, the two constructions read:

	real	complex
a point is	a line $\mathbb{R}v \subset \mathbb{R}^{n+1}$	a complex line $\mathbb{C}v \subset \mathbb{C}^{n+1}$
unit scalars	$\{\pm 1\} = S^0$	$\{e^{i\alpha}\} = S^1 = \mathrm{U}(1)$
line $\cap$ unit sphere	two antipodal points	a circle
sphere model	$\mathbb{RP}^n = S^n/\{\pm 1\}$	$\mathbb{CP}^n = S^{2n+1}/\mathrm{U}(1)$
real dimension	n	$2n$
$n = 1$	$\mathbb{RP}^1 \cong S^1$, double cover $z \mapsto z^2$	$\mathbb{CP}^1 \cong S^2$, Hopf map $S^3 \to S^2$

Worked example 5.9: $\mathbb{CP}^1$ is the Riemann sphere

Take $n = 1$ and homogeneous coordinates $[z_0 : z_1]$. The two charts are

$$U_0 = \{z_0 \neq 0\},\ z = z_1/z_0; \qquad U_1 = \{z_1 \neq 0\},\ w = z_0/z_1.$$

On the overlap $U_0 \cap U_1$ both are defined and

$$w \overset{(1)}{=} \frac{z_0}{z_1} \overset{(2)}{=} \frac{1}{z_1/z_0} \overset{(3)}{=} \frac{1}{z}.$$

(1) Definition of the chart on U_1.
(2) Divide numerator and denominator by $z_0 \neq 0$.
(3) Definition of the chart on U_0.

The overlap is $\{z \neq 0\} = \mathbb{C}^*$. The only point of $\mathbb{CP}^1$ missing from U_0 is $[0 : 1]$, which has $w = 0$, that is "$z = \infty$". So $\mathbb{CP}^1 = \mathbb{C} \cup \{\infty\}$ with the rule "near ∞ use $w = 1/z$". This is the **Riemann sphere**. Stereographic projection from the north pole of the unit sphere in $\mathbb{R}^3$,

$$(X_1, X_2, X_3) \mapsto z = \frac{X_1 + iX_2}{1 - X_3},$$

identifies it with S^2. The south pole goes to $z = 0$, the north pole to $z = \infty$, and the equator to $|z| = 1$.

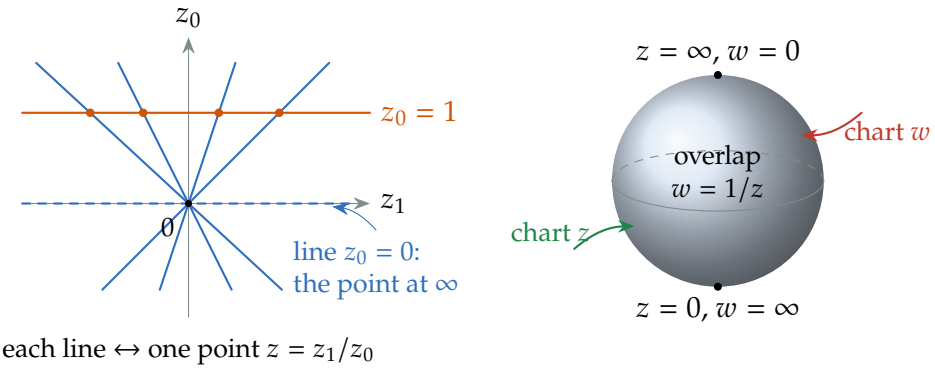
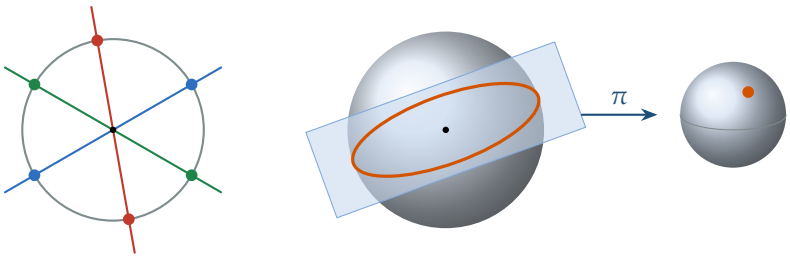


Figure 5.11: Left: a real cartoon of $\mathbb{CP}^1$. Each line through the origin of $\mathbb{C}^2$ (blue) meets the horizontal line $z_0 = 1$ in one point (orange), which we label by $z = z_1/z_0$. The one line that never meets it, $z_0 = 0$ (dashed), is the extra "point at infinity". Right: the result is a sphere covered by two charts, z (everything except the north pole) and w (everything except the south pole), glued by $w = 1/z$ on the overlap.

Try it: the charts of $\mathbb{CP}^2$

Write the three charts of $\mathbb{CP}^2$ with homogeneous coordinates $[z_0 : z_1 : z_2]$ and find the transition function from U_0 to U_1.
Answer. On U_0: $(x, y) = (z_1/z_0, z_2/z_0)$. On U_1: $(u, v) = (z_0/z_1, z_2/z_1)$. On the overlap $x \neq 0$, and $(u, v) = (1/x,\ y/x)$, which is holomorphic there.

The Hopf fibration. For $n = 1$, (5.2) is a map $S^3 \to \mathbb{CP}^1 = S^2$ whose fibres are circles. It is the complex counterpart of the angle-doubling double cover $S^1 \to \mathbb{RP}^1$ of Example 5.4. Two points over each point become a circle over each point. This map is the **Hopf fibration**, which Chapter 1 meets as the first map between spheres of different dimensions that cannot be deformed to a constant.



$\mathbb{R}$: line $\cap$ sphere
$= \{\pm\hat{v}\}$, two points

$\mathbb{C}$: line $\cap$ sphere
$= \{e^{i\alpha}\hat{v}\}$, a circle

one point
$[v] \in \mathbb{CP}^1 = S^2$

Figure 5.12: From two points to a circle. Left: the real cartoon. Each line through the origin meets the sphere in two antipodal points, so $S^n \to \mathbb{RP}^n$ is 2-to-1. Middle: over $\mathbb{C}$ the line $\mathbb{C}v$ is a real *plane* (drawn here one dimension down, as a plane cutting a ball), and it meets the unit sphere S^{2n+1} in a great circle, the phases $e^{i\alpha}\hat{v}$. So in the complex case "two antipodal points" becomes "a circle's worth of points". Right: the whole circle is one point of $\mathbb{CP}^n$. For $n = 1$ this is the Hopf fibration $S^3 \to S^2$ with circle fibres.

Worked example 5.10: The Hopf map, explicitly

Let $S^3 = \{(z_0, z_1) \in \mathbb{C}^2 : |z_0|^2 + |z_1|^2 = 1\}$. Send (z_0, z_1) to $[z_0 : z_1] \in \mathbb{CP}^1$, and then to $S^2 \subset \mathbb{R}^3$ by inverting the stereographic projection of Example 5.9. The result is

$$h(z_0, z_1) = (X_1, X_2, X_3), \qquad X_1 + iX_2 = 2\,\bar{z}_0 z_1, \qquad X_3 = |z_1|^2 - |z_0|^2.$$

It lands on the unit sphere:

$$\begin{aligned} X_1^2 + X_2^2 + X_3^2 &\overset{(1)}{=} 4|z_0|^2|z_1|^2 + \left(|z_1|^2 - |z_0|^2\right)^2 \\ &\overset{(2)}{=} \left(|z_0|^2 + |z_1|^2\right)^2 \\ &\overset{(3)}{=} 1. \end{aligned}$$

(1) $X_1^2 + X_2^2 = |X_1 + iX_2|^2 = 4|\bar{z}_0 z_1|^2$.
(2) $4ab + (b - a)^2 = (a + b)^2$ with $a = |z_0|^2$, $b = |z_1|^2$.
(3) (z_0, z_1) lies on S^3.

It really is the point $[z_0 : z_1]$. For $z_0 \neq 0$, apply the stereographic projection $z = (X_1 + iX_2)/(1 - X_3)$:

$$\frac{X_1 + iX_2}{1 - X_3} \overset{(1)}{=} \frac{2\bar{z}_0 z_1}{1 - |z_1|^2 + |z_0|^2}$$

$$\overset{(2)}{=} \frac{2\bar{z}_0 z_1}{2|z_0|^2}$$
$$\overset{(3)}{=} \frac{z_1}{z_0} = z.$$

(1) Substitute the formulas for $X_1 + \mathrm{i}X_2$ and X_3.
(2) $1 = |z_0|^2 + |z_1|^2$, so $1 - |z_1|^2 + |z_0|^2 = 2|z_0|^2$.
(3) $\bar{z}_0/|z_0|^2 = 1/z_0$.

For $z_0 = 0$ we get $X = (0, 0, 1)$, the north pole, which is $z = \infty$, as it should be.
The fibres are the phase circles. Replace (z_0, z_1) by $(e^{\mathrm{i}\alpha} z_0, e^{\mathrm{i}\alpha} z_1)$. The moduli $|z_i|$ do not change, and $\bar{z}_0 z_1 \mapsto e^{-\mathrm{i}\alpha} e^{\mathrm{i}\alpha} \bar{z}_0 z_1 = \bar{z}_0 z_1$. So h is constant on each circle $\{e^{\mathrm{i}\alpha}(z_0, z_1)\}$. By the compactness computation above, these circles are exactly the fibres.
Chapter 1 writes the same map as the Bloch vector $n_a = \psi^\dagger \sigma_a \psi$ of a normalised spinor ψ. With the components of ψ called z_0, z_1, its third component is $|z_0|^2 - |z_1|^2 = -X_3$. The sign only records which pole we call $z = 0$ (here the south pole, as in Example 5.9). There the map is used to show $\pi_3(S^2) = \mathbb{Z}$: any two fibres are linked circles in S^3.

Physical picture: phases, rays and the monopole

In quantum mechanics S^3 is the set of normalised spin-$\frac{1}{2}$ states, $\mathbb{CP}^1$ is the set of physical states (rays in Hilbert space, the Bloch sphere), and the circle fibre is the unobservable global phase. The fibre over $[z]$ is the unit circle inside the line $[z]$ itself. So S^3 is the unit-circle bundle of the tautological line bundle $\mathcal{O}(-1)$ of Example 5.24, and as a U(1) bundle over S^2 it is the Dirac monopole of unit charge. "Projectivising" is exactly the physicist's move of discarding the overall phase and normalisation of a state.

5.1.7 What we can draw, and what we cannot

We plot curves in $\mathbb{R}^2$ all the time, and we trust the plots. A plot of $y = x^2$ is faithful: every point of the curve has its own place on the page, nothing is glued and nothing is doubled. The scale may stretch distances, but no topology is lost. We have now drawn several pictures of projective space, and none of them is a plot of that kind. This subsection explains why no faithful picture exists, what we draw instead, and which conclusions drawn from such pictures can be trusted.

$\mathbb{RP}^2$ does not fit in $\mathbb{R}^3$. A faithful picture of a closed surface would be an embedding in $\mathbb{R}^3$: a smooth one-to-one map with injective derivative. For $\mathbb{RP}^2$ there is none. Two facts combine.

- A closed surface embedded in $\mathbb{R}^3$ separates space into an inside and an outside (the Jordan–Brouwer separation theorem). So it has an outward unit normal n at every point. A continuous nowhere-zero normal defines an orientation: call a tangent basis (e_1, e_2) positive when (e_1, e_2, n) is right-handed. Hence *every closed surface embedded in $\mathbb{R}^3$ is orientable*.
- $\mathbb{RP}^2$ is not orientable. Example 5.5 showed this by parity ($\chi = 1$ is odd). More visibly, $\mathbb{RP}^2$ contains a Möbius band (Figure 5.13).

The best one can do in $\mathbb{R}^3$ is an **immersion**: a map that is locally an embedding but passes through itself. **Boy's surface** is such an immersion of $\mathbb{RP}^2$. It crosses itself along a closed curve that has one triple point. That curve belongs to the picture, not to $\mathbb{RP}^2$: the two sheets that cross are far apart in $\mathbb{RP}^2$. The other textbook model, the cross-cap, is worse still, since it also has two

pinch points where it is not even an immersion. $\mathbb{RP}^2$ does embed in $\mathbb{R}^4$, for instance through the map $(x, y, z) \mapsto (x^2 - y^2,\ xy,\ yz,\ zx)$ on the unit sphere, which takes the same value at x and $-x$ and so descends to $\mathbb{RP}^2$. But $\mathbb{R}^4$ is no easier to draw.

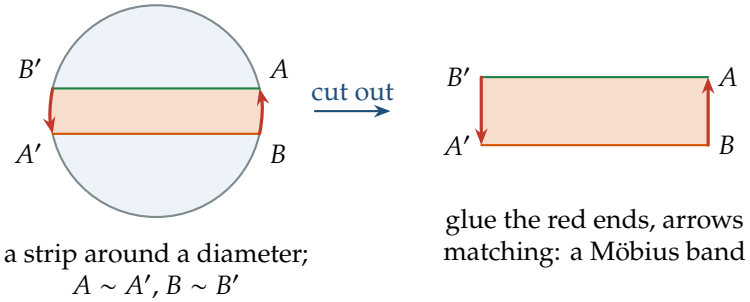


Figure 5.13: $\mathbb{RP}^2$ contains a Möbius band. In the disc model take a strip around a diameter. Its two ends lie on the boundary, and the boundary is glued antipodally: A (top right) is glued to A' (bottom left) and B to B'. So the top (green) edge at one end continues as the bottom (orange) edge at the other end. Cut out and straightened, the strip is a rectangle whose ends are glued with a flip, as the opposite red arrows show. A surface that contains a Möbius band is not orientable, and so it cannot sit in $\mathbb{R}^3$ without crossing itself.

What we draw instead. Three kinds of picture are honest, provided we remember what each one leaves out.

- *One affine chart at a time.* Each of the charts U_X, U_Y, U_Z is an honest copy of $\mathbb{R}^2$, and each shows all of $\mathbb{RP}^2$ except one line. Two charts together show everything. Figure 5.9 shows one curve in two charts.

- *The disc model* (Figures 5.6 and 5.10) is faithful in the interior. On the boundary we must remember the gluing $u \sim -u$.

- *The sphere with antipodal points identified*, (5.1), is faithful up to a factor two. Every figure appears twice, symmetrically under $x \mapsto -x$.

Why chart-by-chart pictures are reliable. Everything we say about curves in projective space is a statement that does not depend on the chart. Such statements can be checked in whichever chart is convenient, and the answer is the same in all of them.

Key idea: intrinsic statements can be checked in any chart

The changes of chart of $\mathbb{RP}^2$ (such as $(x, y) \mapsto (1/x,\ y/x)$ from U_Z to U_X) are smooth with smooth inverses, and so are the projective transformations $[w] \mapsto [Aw]$ with $A \in \mathrm{GL}(3, \mathbb{R})$. A property that is unchanged by all of these is a property of the curve in $\mathbb{RP}^2$, not of the picture. Examples are smoothness, the number of intersection points (counted with multiplicity), tangency, the number of connected components, compactness, and the genus or Euler characteristic.

Here is the check for smoothness. Suppose the curve is $F = 0$ near a point in one chart and $G = 0$ in another, with change of chart ϕ. For a projective curve the two equations are related by $G = h \cdot (F \circ \phi^{-1})$, where h is a nowhere-zero factor (a power of the coordinate we divide by).

At a point p of the curve,

$$\begin{aligned} \mathrm{d}G|_p &\overset{(1)}{=} h(p)\,\mathrm{d}(F\circ\phi^{-1})|_p + (F\circ\phi^{-1})(p)\,\mathrm{d}h|_p \\ &\overset{(2)}{=} h(p)\,\mathrm{d}F|_{\phi^{-1}(p)}\,D\phi^{-1}|_p. \end{aligned}$$

(1) Product rule.
(2) F vanishes at points of the curve, and the chain rule for the first term.

Since $h(p)\neq 0$ and $D\phi^{-1}$ is invertible, $\mathrm{d}G|_p\neq 0$ exactly when $\mathrm{d}F\neq 0$ at the corresponding point. For the parabola, $F=y-x^2$ has $\mathrm{d}F=(-2x,1)\neq 0$ everywhere in the chart $Z\neq 0$, and in the chart $U\neq 0$ the curve is $G=\xi^2+\eta^2-1$ with $\mathrm{d}G=(2\xi,2\eta)\neq 0$ on the circle. So the parabola is smooth everywhere, including at its point at infinity, which only the second chart can see.

Watch out: chart artefacts

Some features of a picture belong to the chart, not to the curve, and must not be taken literally. "The parabola is open, the ellipse is closed" is a statement about the chart $Z\neq 0$: in $\mathbb{RP}^2$ both are the same closed loop (Example 5.7). The same goes for being bounded or unbounded, for the number of branches (the hyperbola has two branches in U_Z and one in U_Y, Example 5.8), for asymptotes, for which points are "at infinity", and for lengths, angles and areas (a circle can be drawn as an ellipse). Even parallelism is chart-dependent: two lines are parallel exactly when they meet on the line we chose to call L_∞.

Intrinsic (trust it in any chart)	**Chart artefact (do not)**
smooth or singular	open or closed; bounded or unbounded
number of intersection points	number of branches; asymptotes
tangency, multiplicity	which points are at infinity
connectedness, compactness	lengths, angles, areas
genus, Euler characteristic	whether two lines are parallel
lines are sent to lines	conic type (ellipse, parabola, hyperbola)

Worked example 5.11: The canvas is another chart

The tiled-floor picture (Figure 5.8) makes the line at infinity of the ground visible as the horizon. Here is why. Put the eye at height h above the ground, looking horizontally along the y-axis, with a vertical canvas at distance 1 in front of the eye. A ground point (x,y) with $y>0$ is seen along the ray from the eye to that point. After moving forward a distance 1 out of y, the ray has covered the fraction $1/y$ of the sideways displacement x and of the drop h. So it crosses the canvas at sideways position u and height v, measured from eye level, given by

$$\begin{aligned} [u:v:1] &\overset{(1)}{=} \left[\frac{x}{y}:-\frac{h}{y}:1\right] \\ &\overset{(2)}{=} [x:-h:y]. \end{aligned}$$

(1) Similar triangles: $u=x/y$ and $v=-h/y$.
(2) Multiply all coordinates by y.

In homogeneous coordinates $[X:Y:Z]$ of the ground (with $[x:y:1]$ for the point (x,y)),

the painting map is therefore

$$[X : Y : Z] \mapsto [X : -hZ : Y], \qquad A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -h \\ 0 & 1 & 0 \end{pmatrix}, \qquad \det A = h \neq 0,$$

a projective transformation of $\mathbb{RP}^2$. The line at infinity of the ground, $Z = 0$, goes to the points $[X : 0 : Y]$, which have $v = 0$: the horizon at eye level. Conversely, the ground line $y = 0$ (the points directly to the left and right of the observer's feet, neither ahead nor behind) goes to $[X : -hZ : 0]$, the line at infinity of the canvas. Rays to those points run parallel to the canvas and are never painted. So the line at infinity is not a special line of $\mathbb{RP}^2$. It is simply the one line that a given chart leaves out, and a change of chart can bring it into view.

The complex case. $\mathbb{CP}^n$ has real dimension $2n$. $\mathbb{CP}^1 = S^2$ can be drawn faithfully, and we do draw it (Figure 5.11). $\mathbb{CP}^2$ is four-dimensional. There is no picture of it on paper, not even of one chart, which is $\mathbb{C}^2 = \mathbb{R}^4$. We picture it in pieces:

- through charts, and real slices of curves in them (with the warning of the pitfall in Section 5.1.1);
- through maps and fibrations: the circle fibration $S^5 \to \mathbb{CP}^2$ of (5.2), or a curve in $\mathbb{CP}^2$ drawn as a branched cover of $\mathbb{CP}^1$, as in Figure 5.16;
- through the moment map below, which draws $\mathbb{CP}^2$ as a triangle with a torus over each interior point.

Worked example 5.12: $\mathbb{CP}^1$ as an interval and $\mathbb{CP}^2$ as a triangle

Define

$$\mu([z_0 : z_1 : z_2]) = \frac{(|z_1|^2,\ |z_2|^2)}{|z_0|^2 + |z_1|^2 + |z_2|^2}.$$

It is well defined, because $z \mapsto \lambda z$ multiplies the numerator and the denominator by $|\lambda|^2$. Its image is the triangle $\Delta = \{(a, b) : a, b \geq 0,\ a + b \leq 1\}$. Take (a, b) in the interior of Δ and choose the representative with $\sum |z_i|^2 = 1$ and z_0 real and positive (this uses up the phase freedom). Then

$$\mu^{-1}(a, b) \overset{(1)}{=} \left\{\left[\sqrt{1 - a - b} : \sqrt{a}\, e^{i\alpha_1} : \sqrt{b}\, e^{i\alpha_2}\right]\right\} \overset{(2)}{=} T^2.$$

(1) The moduli are fixed by $|z_1|^2 = a$, $|z_2|^2 = b$, $|z_0|^2 = 1 - a - b$, with $z_0 > 0$; only the phases of z_1 and z_2 remain free.
(2) Two independent angles α_1, α_2 make a torus.

Over an interior point of an edge, one of the z_i is zero and only one phase remains: the fibre is a circle. Over a vertex only one z_i is nonzero, and the fibre is a single point. Each edge of Δ is one of the lines $z_i = 0$, a copy of $\mathbb{CP}^1$: an interval with a circle over each interior point, where the circles shrink to points at the two ends. That is a sphere. For $\mathbb{CP}^1$ itself the same map is $\mu([z_0 : z_1]) = |z_1|^2/(|z_0|^2 + |z_1|^2)$. On S^3 this is $|z_1|^2$, and in the coordinates of Example 5.10 it equals $(1 + X_3)/2$, the height on the sphere. Its fibres are the circles of latitude.

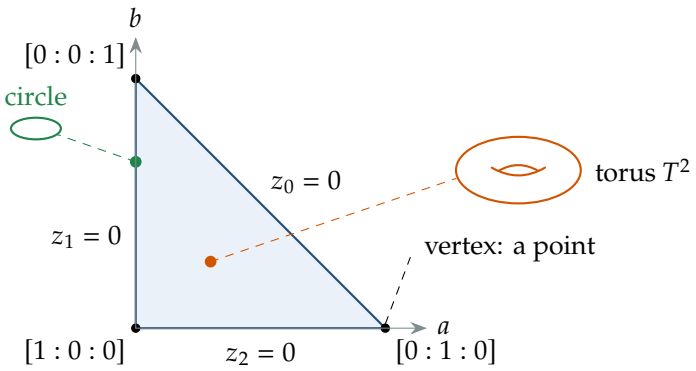


Figure 5.14: A faithful picture of $\mathbb{CP}^2$ by fibres. The moment map $\mu = (|z_1|^2, |z_2|^2)/|z|^2$ sends $\mathbb{CP}^2$ onto a triangle. Over each interior point (orange) sits a torus of phases; over each point of an edge (green) sits a circle; over each vertex sits a single point. Each edge is one of the lines $z_i = 0$, a $\mathbb{CP}^1$ drawn as an interval of circles that shrink to points at the ends. Four real dimensions become a two-dimensional base with the rest of the information stored in the fibres, just as a sphere can be recorded by its height together with a circle of latitude at each height.

Physical picture: no global picture, and nothing lost

We never see all of $\mathbb{RP}^2$ or $\mathbb{CP}^2$ at once, and we do not need to. Every statement we make about a variety in projective space (it is smooth, it meets a line in d points, it has genus p) is intrinsic in the sense of the key idea above. So it can be checked in one chart, or in a fibration picture, and the answer does not depend on the choice. Faithful global pictures are a luxury of low dimension. Charts are the everyday tool, exactly as for any manifold.

Recap: projective space by pictures

- $\mathbb{RP}^n$ is the set of lines through the origin of $\mathbb{R}^{n+1}$. Each line meets S^n in two antipodal points, so $\mathbb{RP}^n = S^n/(x \sim -x)$ and $S^n \to \mathbb{RP}^n$ is a double cover. $\mathbb{RP}^1 \cong S^1$ by angle doubling. $\mathbb{RP}^2$ is a disc with antipodal boundary points glued; $\chi(\mathbb{RP}^2) = 1$, and it is not orientable.
- Perspective puts $\mathbb{R}^2$ inside $\mathbb{RP}^2$ as the chart $Z \neq 0$, $(x, y) \mapsto [x : y : 1]$. What is left over is the line at infinity $Z = 0$, an $\mathbb{RP}^1$ whose points are directions without sign. Any two lines meet in exactly one point; parallels of slope m meet at $[1 : m : 0]$. On a painted floor each direction has its own vanishing point, and all of them lie on the horizon, the line at infinity.
- Homogenising a curve and setting $Z = 0$ finds its points at infinity. The parabola $YZ = X^2$ gains $[0 : 1 : 0]$, and the hyperbola $XY = Z^2$ gains $[1 : 0 : 0]$ and $[0 : 1 : 0]$; each becomes a single loop. Every non-degenerate real conic with real points is the circle $X^2 + Y^2 = Z^2$.
- $\mathbb{CP}^n = S^{2n+1}/\mathrm{U}(1)$: a complex line meets the sphere in a circle. For $n = 1$ this is the Hopf fibration $S^3 \to S^2$.
- No faithful global picture exists for $\mathbb{RP}^2$ (it does not embed in $\mathbb{R}^3$) or for $\mathbb{CP}^2$ (real dimension four). We draw charts, the disc model, the sphere model or fibrations, and we trust only chart-independent statements.

5.1.8 Homogeneous polynomials and projective varieties

A polynomial $P(z_1, \dots, z_{n+1})$ is not a function on $\mathbb{CP}^n$: its value depends on which representative we pick. But its *zero set* can still make sense. This is the key observation of Nash's first page.

Working definition 5.7 (homogeneous polynomial)

P is **homogeneous of degree** d if every monomial in it has total degree d. [the sum of the exponents, e.g. $z_1^2 z_2$ has total degree 3] Equivalently, and this is the operational test,

$$P(\lambda z_1, \dots, \lambda z_{n+1}) \;=\; \lambda^d \, P(z_1, \dots, z_{n+1}) \qquad \text{for all } \lambda \in \mathbb{C}^*. \tag{5.3}$$

This is Nash's (5.1).

Because $\lambda^d \neq 0$, (5.3) says that $P(z) = 0$ if and only if $P(\lambda z) = 0$. So "$P = 0$" is a condition on the line, not on the representative, and

$$V(P) = \{[z] \in \mathbb{CP}^n : P(z) = 0\}$$

is a well-defined subset of $\mathbb{CP}^n$. In the chart U_{n+1} we may choose the representative with $z_{n+1} = 1$, and then

$$\begin{aligned} P(z_1, \dots, z_n, z_{n+1}) = 0 &\overset{(1)}{\iff} z_{n+1}^{-d} \, P(z_1, \dots, z_{n+1}) = 0 \\ &\overset{(2)}{\iff} P\Big(\frac{z_1}{z_{n+1}}, \dots, \frac{z_n}{z_{n+1}}, 1\Big) = 0. \end{aligned} \tag{5.4}$$

(1) Multiply by the nonzero number z_{n+1}^{-d}.
(2) Homogeneity (5.3) with $\lambda = 1/z_{n+1}$.

This is Nash's (5.2) (Nash writes it as a one-way implication; it is in fact an equivalence). In words: *inside a chart, a projective variety is the affine variety of the dehomogenised polynomial*. Conversely, any polynomial $p(x_1, \dots, x_n)$ of degree d can be **homogenised** by setting $x_i = z_i / z_{n+1}$ and multiplying through by z_{n+1}^d:

$$P(z_1, \dots, z_{n+1}) = z_{n+1}^d \, p\big(z_1/z_{n+1}, \dots, z_n/z_{n+1}\big).$$

The zeros of P with $z_{n+1} \neq 0$ are the old affine zeros; the zeros with $z_{n+1} = 0$ are the new **points at infinity**.

Working definition 5.8 (projective variety)

A **projective algebraic variety** is the common zero set in $\mathbb{CP}^n$ of finitely many homogeneous polynomials $P_1, \dots, P_k$ [so whether $P_i = 0$ at a point does not depend on which representative we pick] (their degrees may differ). We write $V \subset \mathbb{CP}^n$. *Operationally*: in each chart U_j [the points with $z_j \neq 0$, a copy of $\mathbb{C}^n$ once we scale to $z_j = 1$] set $z_j = 1$ and solve the resulting ordinary polynomial equations.

Nash phrases this as "V can be thought of as being given by an embedding into $\mathbb{CP}^n$". The point is that we regard V as a space in its own right, and the polynomials describe how it sits inside $\mathbb{CP}^n$.

Worked example 5.13: Completing the complex circle

Homogenise $x^2 + y^2 = 1$ with $x = z_1/z_0$, $y = z_2/z_0$:

$$C = \{\, z_1^2 + z_2^2 - z_0^2 = 0 \,\} \subset \mathbb{CP}^2.$$

The affine part ($z_0 \neq 0$) is the complex circle of the previous example, which we found to be $\mathbb{C}^*$. The points at infinity have $z_0 = 0$, so $z_1^2 + z_2^2 = 0$, that is $z_2 = \pm i z_1$. With $z_1 \neq 0$ (otherwise all coordinates vanish) we get exactly two points,

$$[0 : 1 : \mathrm{i}] \quad \text{and} \quad [0 : 1 : -\mathrm{i}].$$

These are the two missing ends of the cylinder $\mathbb{C}^*$. Adding them gives a sphere. In fact every smooth conic is a copy of $\mathbb{CP}^1$. For the conic $z_0 z_2 = z_1^2$ this is completely explicit:

$$\nu : \mathbb{CP}^1 \to \mathbb{CP}^2, \qquad [s : t] \mapsto [s^2 : st : t^2], \qquad z_0 z_2 \overset{(1)}{=} s^2 t^2 \overset{(2)}{=} (st)^2 = z_1^2.$$

(1) Substitute the components of ν.
(2) Rearrange.

The map ν is well defined, because rescaling $(s, t) \to (\lambda s, \lambda t)$ rescales every component by λ^2. It is injective, because s/t can be read off as z_0/z_1 or z_1/z_2. It is onto, because any point of the conic with $z_0 \neq 0$ is $[1 : a : a^2] = \nu([1 : a])$ and the only point with $z_0 = 0$ is $[0 : 0 : 1] = \nu([0 : 1])$. (Our circle conic is carried to this one by a linear change of coordinates, $z_0 \pm z_1 \to$ new coordinates.)

How many points on a line? A line in $\mathbb{CP}^2$ is the zero set of a linear form, and it is itself a copy of $\mathbb{CP}^1$ (parametrise it by two of its points). Restrict a homogeneous P of degree d to the line: we get a homogeneous polynomial of degree d in the two parameters $[s : t]$, which factors into d linear factors over $\mathbb{C}$. So a degree-d curve meets a line in exactly d points counted with multiplicity (unless it contains the line). This is the simplest case of **Bézout's theorem**: curves of degrees d_1 and d_2 in $\mathbb{CP}^2$ with no common component meet in $d_1 d_2$ points, counted with multiplicity. Over $\mathbb{R}$ none of this is true (a line can miss a circle); over $\mathbb{C}$ and with points at infinity included, counting becomes exact. This is the first sign of the rigidity that makes algebraic geometry so computable.

5.1.9 Singular points and the Jacobian criterion

Nash warns that "an algebraic variety $V \subset \mathbb{CP}^n$ will not in general be a sub-manifold of $\mathbb{CP}^n$": it can have singular points, like the node and the cusp in Figure 5.2. How do we detect them? By the implicit function theorem, which works in the holomorphic setting exactly as in the real one.

Key idea: holomorphic implicit function theorem

Let $P_1, \dots, P_k$ be holomorphic functions of $x \in \mathbb{C}^m$ that vanish at x_0. If the $k \times m$ matrix of first derivatives

$$J(x_0) = \left[\frac{\partial P_i}{\partial x_j}(x_0) \right]_{k \times m}$$

has its maximal rank k, then near x_0 the common zero set is a complex submanifold of dimension $m - k$: we can solve for k of the coordinates as holomorphic functions of the other $m - k$.

Working definition 5.9 (singular point)

Let V be cut out by $P_1, \dots, P_k$ and have the expected dimension $m - k$. [one complex dimension lost per independent equation] A point $x_0 \in V$ is **nonsingular** (or smooth) if the Jacobian matrix $J(x_0) = [\partial P_i / \partial x_j]$ has rank k, [its k rows are linearly independent] and **singular** if the rank drops below k. *Operationally*: compute all $k \times k$ minors of J and solve "all minors vanish" together with $P_1 = \cdots = P_k = 0$. For a single equation ($k = 1$) this means: solve $P = 0$ and $\partial P / \partial x_j = 0$ for all j.

Nash phrases the same thing using the Jacobian of the embedding $e : V \to \mathbb{CP}^n$. Recall that for a smooth map $f : M \to N$ between manifolds, with local coordinates $(x_1, \dots, x_m)$ on M and $(y_1, \dots, y_n)$ on N, the Jacobian $J(f)$ is the $n \times m$ matrix $[\partial y_i / \partial x_j]$. The two versions agree: near a smooth point, the implicit function theorem gives local coordinates on V in which the embedding has a Jacobian of full rank. If the rank drops, no such coordinates exist. When V has no singular points it is a complex submanifold of $\mathbb{CP}^n$, and it is called an **algebraic submanifold** of $\mathbb{CP}^n$.

For a projective hypersurface there is a useful shortcut. If P is homogeneous of degree d, differentiate (5.3) with respect to λ and set $\lambda = 1$:

$$\sum_{i=1}^{n+1} z_i \frac{\partial P}{\partial z_i}(z) \overset{(1)}{=} \frac{\mathrm{d}}{\mathrm{d}\lambda} P(\lambda z)\Big|_{\lambda=1} \overset{(2)}{=} \frac{\mathrm{d}}{\mathrm{d}\lambda} \lambda^d P(z)\Big|_{\lambda=1} = d\, P(z). \tag{5.5}$$

(1) Chain rule: $\frac{\mathrm{d}}{\mathrm{d}\lambda} P(\lambda z) = \sum_i z_i \, (\partial_i P)(\lambda z)$.
(2) Homogeneity (5.3).

This is **Euler's relation**. It implies that if all partial derivatives $\partial P / \partial z_i$ vanish at $z \neq 0$ then $P(z) = 0$ automatically. So: *the singular points of $V(P) \subset \mathbb{CP}^n$ are the points where all $n + 1$ homogeneous partial derivatives vanish.* No chart is needed.

Worked example 5.14: Fermat curves are smooth

Let $F = z_0^d + z_1^d + z_2^d$. Then $\partial F / \partial z_i = d\, z_i^{d-1}$. All three vanish only if $z_0 = z_1 = z_2 = 0$, which is not a point of $\mathbb{CP}^2$. So the Fermat curve of every degree d is a smooth complex curve in $\mathbb{CP}^2$. For $d = 2$ it is a conic (a sphere), for $d = 3$ a torus, and in general its genus is given by the **degree–genus formula**

$$p = \frac{(d-1)(d-2)}{2} \tag{5.6}$$

for any smooth plane curve of degree d. (Plausibility: degenerate the curve into d lines in general position; each line is a sphere, any two meet in one point, so there are $\binom{d}{2}$ junctions; $d - 1$ of them are needed to connect the d spheres into a tree, and each remaining junction adds a handle: $\binom{d}{2} - (d-1) = \frac{(d-1)(d-2)}{2}$. This is the argument in Gathmann's notes, §0.1.)

Worked example 5.15: When is Nash's cubic smooth?

Nash's equation (5.3) is the cubic $F(x, y) = x(x-1)(x-\lambda) - y^2 = 0$. Write $q(x) = x(x-1)(x-\lambda)$. In the affine chart,

$$\frac{\partial F}{\partial y} = -2y, \qquad \frac{\partial F}{\partial x} = q'(x).$$

A singular affine point needs $y = 0$, then $F = 0$ forces $q(x) = 0$, so $x \in \{0, 1, \lambda\}$, and finally $q'(x) = 0$. A root of q is also a root of q' exactly when it is a repeated root, which happens

exactly when $\lambda = 0$ or $\lambda = 1$. Now the point at infinity. Homogenise with $x = X/Z$, $y = Y/Z$:

$$\tilde{F}(X, Y, Z) = X(X - Z)(X - \lambda Z) - Y^2 Z.$$

At $Z = 0$ we need $X^3 = 0$, so the only point at infinity is $[X : Y : Z] = [0 : 1 : 0]$. Use the chart $Y = 1$ with coordinates (X, Z):

$$\frac{\partial \tilde{F}}{\partial Z}(0,1,0) \overset{(1)}{=} \left[\frac{\partial}{\partial Z}\big(X(X - Z)(X - \lambda Z)\big) - 2YZ \cdot 0 - Y^2\right]_{(0,1,0)}$$
$$\overset{(2)}{=} 0 - 1 = -1 \neq 0.$$

(1) Differentiate $\tilde{F}$ term by term; the product rule on $Y^2 Z$ gives Y^2 (with $Y = 1$ in this chart, Y is not a variable).

(2) Every term of the cubic part is at least quadratic in (X, Z), so its derivative vanishes at $X = Z = 0$.

So the point at infinity is always smooth, and *Nash's cubic (5.3) is smooth precisely when* $\lambda \neq 0, 1$. At $\lambda = 0$ the curve becomes $y^2 = x^2(x - 1)$, a node at the origin; at $\lambda = 1$ it becomes $y^2 = x(x - 1)^2$, a node at $(1, 0)$. The parameter space of smooth tori in this family is $\mathbb{C} \setminus \{0, 1\}$. Keep this in mind: it is our first glimpse of a moduli space.

Watch out: a singular cone can have a smooth projectivisation

The affine cone $\{z_1^2 + z_2^2 = z_0^2\} \subset \mathbb{C}^3$ is singular at the origin (all partials vanish there). But the origin is not a point of $\mathbb{CP}^2$, and the projective conic with the same equation is smooth. Singularity is always tested at actual points of the variety.

5.1.10 When is a complex manifold algebraic? Chow and Kodaira

A complex submanifold of $\mathbb{CP}^n$ could, a priori, be cut out by *holomorphic* functions that are not polynomials, such as exp or sin. The first surprise is that compactness rules this out.

Key idea: Chow's theorem

Every closed complex submanifold of $\mathbb{CP}^n$ (more generally, every closed analytic subvariety) is algebraic: it is the zero set of finitely many homogeneous polynomials. Proved in Griffiths–Harris.

Nash puts it as: take a general analytic embedding $e : M \to \mathbb{CP}^n$, not insisting that $e(M)$ be given by polynomials; then $e(M)$ is in fact given by polynomials, and M is a nonsingular algebraic variety. So on compact spaces, "holomorphic" and "algebraic" coincide. The mechanism is visible in the simplest case.

Worked example 5.16: Meromorphic functions on $\mathbb{CP}^1$ are rational

Let f be a meromorphic function on $\mathbb{CP}^1$, that is, holomorphic except for poles, including possibly at $z = \infty$ (tested in the chart $w = 1/z$). Poles are isolated, and $\mathbb{CP}^1$ is compact, so there are finitely many, say at $a_1, \dots, a_r \in \mathbb{C}$, plus perhaps ∞. Near a_j the Laurent series has a finite principal part $\mathrm{pp}_j(z) = \sum_{k=1}^{m_j} c_{jk}(z - a_j)^{-k}$, and near ∞ the principal part is a

polynomial $\mathrm{pp}_\infty(z) = \sum_{k=1}^{m_\infty} c_k z^k$. Consider

$$g(z) = f(z) - \sum_{j=1}^{r} \mathrm{pp}_j(z) - \mathrm{pp}_\infty(z).$$

Then g has no poles in $\mathbb{C}$ and is bounded near ∞. A holomorphic function on all of $\mathbb{C}$ that is bounded is constant (**Liouville's theorem**: the Cauchy estimate $|g'(z_0)| \le \sup|g|/R$ on a circle of radius R gives $g' = 0$ as $R \to \infty$). So

$$f(z) \overset{(1)}{=} \text{const} + \sum_{j=1}^{r}\sum_{k=1}^{m_j} \frac{c_{jk}}{(z-a_j)^k} + \sum_{k=1}^{m_\infty} c_k z^k \overset{(2)}{=} \frac{\text{polynomial}}{\prod_j (z-a_j)^{m_j}}.$$

(1) g is constant by Liouville.
(2) Put everything over a common denominator.

Every meromorphic function on the Riemann sphere is a rational function. The two ingredients are the ones behind Chow's theorem in general: compactness (finitely many poles, controlled growth) and Liouville-type rigidity of holomorphic functions.

Chow's theorem is about subsets of a given $\mathbb{CP}^n$. The second, deeper question is intrinsic: given an abstract compact complex manifold, can it be embedded in some $\mathbb{CP}^n$ at all? To state the answer we need forms of type $(1,1)$, which we treat properly in Section 5.3.1. For now: in local holomorphic coordinates $z^1, \dots, z^n$, a real 2-form of **type** $(1,1)$ is

$$\omega = \frac{\mathrm{i}}{2} \sum_{i,j} h_{i\bar{j}}(z, \bar{z})\, \mathrm{d}z^i \wedge \mathrm{d}\bar{z}^j, \qquad h_{i\bar{j}} = \overline{h_{j\bar{i}}} \text{ (Hermitian)},$$

and it is **positive** if the Hermitian matrix $h_{i\bar{j}}$ is positive definite at every point. (A positive $(1,1)$-form is exactly the Kähler form of a Hermitian metric; see Section 5.3.6.)

Key idea: Kodaira embedding theorem

A compact complex manifold M is an algebraic variety (can be holomorphically embedded in some $\mathbb{CP}^n$) if and only if it carries a closed, positive 2-form ω of type $(1,1)$ whose cohomology class $[\omega]$ lies in $H^2(M;\mathbb{Q})$.

Two remarks from Nash, expanded. First, the class of a closed real 2-form lies automatically in $H^2(M;\mathbb{R})$; the condition is that it is rational. Concretely, $[\omega] \in H^2(M;\mathbb{Q})$ means that $\int_S \omega \in \mathbb{Q}$ for every closed surface S in M. Second, if $[\omega]$ is rational, some integer multiple is integral: $H_2(M;\mathbb{Z})$ is finitely generated, so there are finitely many basis surfaces $S_1, \dots, S_b$; if $\int_{S_a} \omega = r_a/s_a$ then $N = \prod_a s_a$ gives $\int_{S_a} N\omega \in \mathbb{Z}$ for all a. An integral class is exactly the first Chern class of a line bundle (Section 5.3.4), and the embedding in the theorem is built from the holomorphic sections of a high power of that line bundle.

Worked example 5.17: $\mathbb{CP}^1$ and the torus satisfy Kodaira

The sphere. On $\mathbb{CP}^1$ take $\omega = \frac{\mathrm{i}}{2}\frac{\mathrm{d}z\wedge\mathrm{d}\bar{z}}{(1+|z|^2)^2}$ (the Fubini–Study form of Section 5.3.7). With $z = x + \mathrm{i}y$ we have $\mathrm{d}z \wedge \mathrm{d}\bar{z} = -2\mathrm{i}\,\mathrm{d}x \wedge \mathrm{d}y$, so

$$\int_{\mathbb{CP}^1} \omega \overset{(1)}{=} \int_{\mathbb{R}^2} \frac{\mathrm{d}x\,\mathrm{d}y}{(1+x^2+y^2)^2} \overset{(2)}{=} 2\pi \int_0^\infty \frac{r\,\mathrm{d}r}{(1+r^2)^2} \overset{(3)}{=} 2\pi\left[-\frac{1}{2(1+r^2)}\right]_0^\infty = \pi.$$

(1) $\frac{\mathrm{i}}{2}\cdot(-2\mathrm{i})=1$; the chart z misses only one point, which has zero area.
(2) Polar coordinates, $\mathrm{d}x\,\mathrm{d}y = r\,\mathrm{d}r\,\mathrm{d}\theta$.
(3) Antiderivative of $r/(1+r^2)^2$.

Since $H_2(\mathbb{CP}^1;\mathbb{Z})=\mathbb{Z}$ is generated by $\mathbb{CP}^1$ itself, $[\omega/\pi]$ is an integral class. Positivity: $h_{1\bar1}=(1+|z|^2)^{-2}>0$. Closedness: automatic on a surface (see below). So $\mathbb{CP}^1$ is algebraic, as it should be.

The torus. Let $T=\mathbb{C}/\Lambda$ with $\Lambda=\mathbb{Z}+\tau\mathbb{Z}$, $\operatorname{Im}\tau>0$. The flat form $\omega=\frac{\mathrm{i}}{2}\mathrm{d}z\wedge\mathrm{d}\bar z=\mathrm{d}x\wedge\mathrm{d}y$ is positive and closed, and $\int_T\omega$ is the area of the parallelogram spanned by 1 and τ, namely $\operatorname{Im}\tau$. Then $\omega/\operatorname{Im}\tau$ integrates to 1. So *every* complex torus of dimension one is algebraic. We will see it explicitly as the cubic (5.3) in Section 5.2.3. (In complex dimension two and higher this fails: a generic complex torus $\mathbb{C}^2/\Lambda$ has no rational positive $(1,1)$ class, and is not algebraic. The conditions on Λ for it to be algebraic are the Riemann conditions, and they are exactly the symmetry and positivity of the period matrix we meet in Section 5.2.17.)

Physical picture: Kodaira and Dirac quantisation

Think of ω as a magnetic field on M. "$[\omega]$ integral" is Dirac's quantisation condition: the flux through every closed surface is an integer multiple of the flux quantum. It is precisely the condition for a charged particle's wavefunction to be a section of a line bundle L with curvature ω. The holomorphic sections of L^k are the lowest-Landau-level states in the field $k\omega$, and Kodaira's theorem says that for k large there are enough of them to use as homogeneous coordinates: send a point x to $[s_0(x):\cdots:s_N(x)]$. Strong magnetic field embeds the manifold in projective space.

5.1.11 Riemann surfaces are algebraic curves

The most important class of algebraic varieties for us is in complex dimension one. A compact Riemann surface is a compact complex manifold of complex dimension one (we define it carefully in Section 5.2.1). On a real two-dimensional manifold every 2-form is closed, since $\mathrm{d}\omega$ would be a 3-form and there are none. Any Riemann surface carries a Hermitian metric (patch local ones together with a partition of unity), whose $(1,1)$-form is then positive and closed; rescaling it gives total area 1, a rational class. By Kodaira, *every compact Riemann surface is an algebraic curve*. Nash writes Σ_p for the compact Riemann surface of genus p, and so shall we.

Watch out: Nash writes p for the genus

Most books write g for the genus. Nash uses p because he reserves g for a metric on the surface (footnote on p. 138). We follow Nash. Do not confuse this p with the p of a p-form; context always decides, and where both occur we will say so.

Reading Nash (pp. 137–139, (5.1)–(5.2))

Nash defines algebraic varieties as zero loci of homogeneous polynomials in $\mathbb{CP}^n$, explains why the zero locus is well defined (5.1)–(5.2), and detects singular points by the rank of a Jacobian. He then quotes two theorems without proof: Chow's (analytic submanifolds of $\mathbb{CP}^n$ are algebraic) and Kodaira's embedding theorem (algebraic $\iff$ there is a closed positive $(1,1)$-form with rational class). Nash introduces $\mathbb{CP}^n$ directly as the space of lines through the origin of $\mathbb{C}^{n+1}$ and does not discuss $\mathbb{RP}^n$. The real warm-up here (the double cover $S^n\to\mathbb{RP}^n$, the disc model, perspective and the line at infinity, conics in $\mathbb{RP}^2$, which pictures are legitimate) and the Hopf fibration $S^3\to\mathbb{CP}^1$ are background for his

one sentence. Also supplied here: the charts and transition functions of $\mathbb{CP}^n$, the Riemann sphere, the homogenisation dictionary, Euler's relation and the projective Jacobian test, Bézout and the degree–genus formula, the smoothness analysis of Nash's cubic (5.3), the Liouville argument behind Chow's theorem, and the explicit check of Kodaira's hypothesis on $\mathbb{CP}^1$ and on tori.

Recap

- An affine variety is a common zero set in $\mathbb{C}^m$; a projective variety is a common zero set of homogeneous polynomials in $\mathbb{CP}^n$, well defined because $P(\lambda z) = \lambda^d P(z)$.
- A complex function is a map from the x-plane to the y-plane; we draw it by matching curves in the two planes. $y = x^2$ doubles angles and covers each circle twice, so $\sqrt{y}$ lives on two sheets glued crosswise along a slit. $y = 1/x$ swaps inside and outside of the unit circle and wants the extra point ∞ of the Riemann sphere.
- $\mathbb{RP}^n = S^n/(x \sim -x)$; $\mathbb{RP}^2 = \mathbb{R}^2 \sqcup L_\infty$, and adding the line at infinity closes curves up (the parabola and hyperbola become circles). Only chart-independent features of pictures are meaningful.
- $\mathbb{CP}^n = S^{2n+1}/\mathrm{U}(1)$ is covered by $n+1$ charts $z_j \neq 0$ with holomorphic transition functions; it is compact. $\mathbb{CP}^1$ is the Riemann sphere, and $S^3 \to \mathbb{CP}^1$ is the Hopf fibration.
- Singular points are where the Jacobian of the defining polynomials drops rank. For a hypersurface in $\mathbb{CP}^n$: all partials vanish.
- Chow: compact analytic $\Rightarrow$ algebraic. Kodaira: algebraic $\iff$ a closed positive $(1,1)$-form with rational (hence, after scaling, integral) class exists.
- Every compact Riemann surface Σ_p is algebraic. The cubic $y^2 = x(x-1)(x-\lambda)$ is smooth exactly for $\lambda \neq 0, 1$.

5.2 Riemann surfaces and divisors

For the rest of the chapter the main character is a compact Riemann surface Σ_p. Three questions drive this section. What *is* a Riemann surface, concretely, and how do the polynomial descriptions of genus 0, 1 and higher look? How many different complex structures does a surface of genus p admit? And how do we describe holomorphic line bundles on it, which is where all the interesting global information lives? The answer to the last question is the notion of a *divisor*: a finite list of points with integer multiplicities that records where a section vanishes or blows up.

5.2.1 Riemann surfaces: charts and genus

Working definition 5.10 (Riemann surface)

A **Riemann surface** is a connected real 2-manifold [a surface: near each point it looks like a piece of the plane] covered by charts $z_\alpha : U_\alpha \to \mathbb{C}$ [a complex coordinate z_α on each patch U_α] such that on every overlap the change of coordinates $z_\beta = h_{\beta\alpha}(z_\alpha)$ is a *holomorphic* function with [complex differentiable, so near each point a power series in

z_α alone, with no $\bar{z}_\alpha$] nonzero derivative. *Operationally*: to check that some construction makes sense on a Riemann surface, check it in one chart and verify that it is unchanged (or changes in the stated way) under a holomorphic change of variable $z \mapsto w(z)$.

Every holomorphic map with $w'(z) \neq 0$ is conformal: it preserves angles, because locally it is multiplication by the complex number $w'(z)$, a rotation times a dilation. Conversely, on any oriented surface with a metric one can find local **isothermal coordinates** in which $ds^2 = e^{\phi(x,y)}(dx^2 + dy^2) = e^\phi |dz|^2$, and two such coordinate systems differ by a holomorphic map. So

Key idea: complex structure = conformal structure

On an oriented surface, a complex structure is the same thing as a metric up to local rescaling $g \to e^\sigma g$. This is why string theory, whose Polyakov action is invariant under Weyl rescalings of the worldsheet metric, lives on Riemann surfaces.

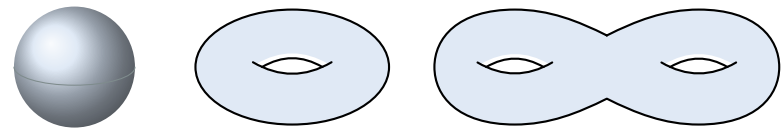
A compact Riemann surface is in particular a compact, orientable, closed surface (orientable because holomorphic changes of coordinates have positive Jacobian determinant $|w'(z)|^2$). Such surfaces are classified topologically by a single integer.

Working definition 5.11 (genus)

The **genus** p of a compact orientable surface is its number of handles: it is homeomorphic to [its points can be matched one-to-one with those of, continuously in both directions,] a sphere with p handles attached. *Operationally*: triangulate it (or divide it into polygons) with V vertices, E edges and F faces; then the **Euler characteristic**

$$\chi(\Sigma_p) = V - E + F = 2 - 2p. \tag{5.7}$$

For example, the surface of a tetrahedron is a sphere with $V - E + F = 4 - 6 + 4 = 2$, so $p = 0$. A square with opposite sides glued is a torus: after gluing there is one vertex, two edges and one face, so $\chi = 1 - 2 + 1 = 0$ and $p = 1$. The $4p$-gon with sides glued in the pattern $a_1 b_1 a_1^{-1} b_1^{-1} \cdots$, which Nash uses in Section 5.2.8, has (after gluing) one vertex, $2p$ edges and one face: $\chi = 1 - 2p + 1 = 2 - 2p$.



$p = 0, \chi = 2$ $p = 1, \chi = 0$ $p = 2, \chi = -2$

Figure 5.15: Compact orientable surfaces are classified by their genus p, the number of handles (holes). The Euler characteristic drops by 2 with every handle. A compact Riemann surface is one of these, with extra structure: a consistent notion of holomorphic function.

5.2.2 Holomorphic and meromorphic functions

A function f on a Riemann surface is **holomorphic** if $f \circ z_\alpha^{-1}$ is holomorphic in every chart. Since changes of chart are holomorphic, it is enough to check one chart around each point.

Working definition 5.12 (meromorphic function; order)

f is **meromorphic** if, near every point P and in a local coordinate z with $z(P) = 0$, it has a Laurent expansion [a power series in z that may also contain negative powers $z^{-1}, z^{-2}, \dots$] with finitely many negative powers:

$$f(z) = \sum_{k \geq n} c_k z^k, \qquad c_n \neq 0.$$

The integer $n = \operatorname{ord}_P f$ is the **order** of f at P. If $n > 0$, f has a zero of order n; if $n < 0$, a pole of order $-n$. Nash's convention is exactly this: *the order of a pole is a negative integer*. *Operationally*: write $f = z^n\, u(z)$ with u holomorphic and $u(0) \neq 0$; then n is the order.

The order does not depend on the coordinate. If $w = w(z)$ is another coordinate centred at P, then $w = z\, v(z)$ with $v(0) = w'(0) \neq 0$, so $z^n = w^n\, v^{-n}$ and $f = w^n \cdot (\text{nonvanishing})$.

Worked example 5.18: Orders of z on the Riemann sphere

Take $f(z) = z$ on $\mathbb{CP}^1$. At $z = 0$ it has a simple zero: $\operatorname{ord}_0 f = 1$. At every other finite point it is holomorphic and nonzero: order 0. At ∞ use the chart $w = 1/z$:

$$f = z = \frac{1}{w} = w^{-1} \cdot 1,$$

so $\operatorname{ord}_\infty f = -1$, a simple pole. The total count is $1 + (-1) = 0$. More generally a rational function $f = a(z)/b(z)$ with $\deg a = m$, $\deg b = k$ has m zeros and k poles in $\mathbb{C}$, and at ∞ it behaves like $z^{m-k} = w^{k-m}$, order $k - m$. Total: $m - k + (k - m) = 0$. We will see in Section 5.2.8 that this is not an accident.

Key idea: holomorphic functions on a compact surface are constant

If Σ is compact and f is holomorphic on all of Σ, then $|f|$ attains a maximum somewhere. A nonconstant holomorphic function is an open map (it maps small discs onto neighbourhoods), so it cannot have a local maximum of $|f|$. Hence f is constant.

So on a compact Riemann surface the interesting functions are *meromorphic*, and the interesting question is: *which pole patterns are allowed?* That is exactly the question divisors and the Riemann–Roch theorem answer.

5.2.3 The three regimes: genus 0, 1 and higher

Nash states that every Σ_p can be embedded in some $\mathbb{CP}^n$ and gives the minimal n in each regime. Let us see each one concretely.

Genus 0. Σ_0 is the Riemann sphere $\mathbb{CP}^1$ of Example 5.9, so "$n = 1$ trivially". (It is a theorem, the uniformisation theorem, that there is only *one* complex structure on the sphere.)

Genus 1. Nash's example (5.3) is the plane cubic

$$x(x-1)(x-\lambda) - y^2 = 0, \qquad \lambda \in \mathbb{C} \setminus \{0, 1\}, \tag{5.8}$$

written in the affine chart of $\mathbb{CP}^2$ and completed by the single point at infinity $[0 : 1 : 0]$ (Example 5.15). Why is this a torus? Two independent arguments follow; the second one is fully explicit.

Worked example 5.19: The cubic as a double cover: counting branch points

Solve (5.8) for y: $y = \pm\sqrt{q(x)}$ with $q(x) = x(x-1)(x-\lambda)$. Over each $x \in \mathbb{CP}^1$ there are two points of the curve, except where the two values of y coincide. That happens at the three roots $x = 0, 1, \lambda$ and also at $x = \infty$: going once around a huge circle $x = Re^{i\theta}$ changes $\sqrt{q} \sim x^{3/2}$ by $e^{3\pi i} = -1$, so the two values are swapped there too. So the curve is a two-sheeted cover of the sphere branched over $b = 4$ points. Now count Euler characteristics with a triangulation of $\mathbb{CP}^1$ that has the four branch points among its vertices. Every face and edge has two preimages; every vertex has two preimages except the branch points, which have one:

$$\begin{aligned}\chi(\text{curve}) &\overset{(1)}{=} 2V - b - 2E + 2F \\ &\overset{(2)}{=} 2\chi(\mathbb{CP}^1) - b \\ &\overset{(3)}{=} 2 \cdot 2 - 4 = 0.\end{aligned}$$

(1) Count preimages of vertices, edges and faces.
(2) $V - E + F = \chi(\mathbb{CP}^1)$.
(3) $\chi(\mathbb{CP}^1) = 2$ and $b = 4$.

By (5.7), $2 - 2p = 0$, so $p = 1$: a torus. The same count for $y^2 = f(x)$ with f of degree $2p+1$ or $2p+2$ and distinct roots gives $b = 2p + 2$ branch points and $\chi = 4 - (2p+2) = 2 - 2p$: genus p. These are the **hyperelliptic curves**.

Figure 5.16 shows how the two sheets fit together. Cut each sheet along the red slits, one from 0 to 1 and one from λ to ∞. Crossing a slit takes you from one value of $\sqrt{q}$ to the other, that is, to the other sheet, so the slits are glued crosswise. Each slit opens into a circle, and two spheres joined along two circles form a torus (bottom of the figure). The red circles on the torus are the former slits. The point $x = \infty$ is one of the four branch points; on the curve in $\mathbb{CP}^2$ it is the single point at infinity $[0:1:0]$, and adding it is what closes the surface up (compare the parabola in Example 5.7).

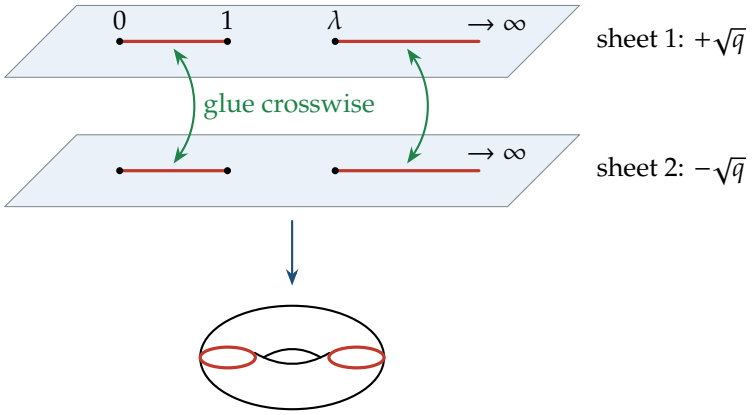
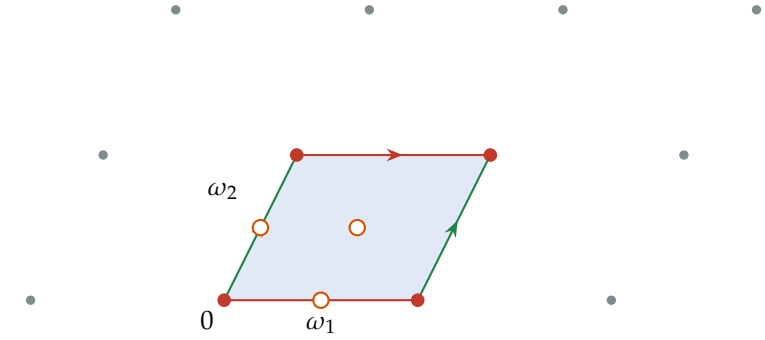


Figure 5.16: The curve $y^2 = x(x-1)(x-\lambda)$ mapped over two copies of the complex x-plane (Riemann spheres), where each sheet represents one of the two possible solutions for y (the two signs of $\sqrt{q}$).

The fully explicit way to see the torus is to start from the torus and write down the cubic. This is the classical theory of the Weierstrass $\wp$-function (Gathmann §6.5).

Working definition 5.13 (complex torus)

Let $\omega_1, \omega_2 \in \mathbb{C}$ be linearly independent over $\mathbb{R}$, [neither is a real multiple of the other, so together they span the plane] and $\Lambda = \mathbb{Z}\omega_1 + \mathbb{Z}\omega_2$ the **lattice** they generate. The quotient $T = \mathbb{C}/\Lambda$, where $z \sim z + m\omega_1 + n\omega_2$, is a **complex torus**. The coordinate z on $\mathbb{C}$ serves as a local coordinate on T, and the changes of chart are translations $z \mapsto z + \omega$, which are holomorphic. Rescaling and rotating ($z \to z/\omega_1$) we may always take $\Lambda = \mathbb{Z} + \tau\mathbb{Z}$ with $\operatorname{Im}\tau > 0$ [here $\tau = \omega_2/\omega_1$, with the two generators swapped if needed so that τ lies in the upper half plane].



• poles of $\wp$ (double) and $\wp'$ (triple): the lattice points
◦ simple zeros of $\wp'$: the half-periods $\frac{\omega_1}{2}$, $\frac{\omega_2}{2}$, $\frac{\omega_1+\omega_2}{2}$

Figure 5.17: The lattice $\Lambda = \mathbb{Z}\omega_1 + \mathbb{Z}\omega_2$ and a fundamental parallelogram. Gluing opposite edges (same colour, same arrow) produces the torus $\mathbb{C}/\Lambda$. The Weierstrass function $\wp$ has a double pole at each lattice point; its derivative $\wp'$ has a triple pole there and three simple zeros at the half-periods (orange circles).

A meromorphic function on T is the same thing as a meromorphic function on $\mathbb{C}$ that is doubly periodic: $f(z+\omega) = f(z)$ for $\omega \in \Lambda$ (an **elliptic function**). There are no nonconstant holomorphic ones (compactness), so we need poles. The simplest choice is a double pole at each lattice point, made convergent by subtracting constants:

$$\wp(z) = \frac{1}{z^2} + \sum_{\omega \in \Lambda \setminus \{0\}} \left(\frac{1}{(z-\omega)^2} - \frac{1}{\omega^2} \right). \tag{5.9}$$

The subtracted $1/\omega^2$ makes each term fall off like $|\omega|^{-3}$, and $\sum |\omega|^{-3}$ converges (there are about $2\pi N$ lattice points at distance $\approx N$, giving $\sum N \cdot N^{-3}$). Its derivative is $\wp'(z) = -2\sum_{\omega\in\Lambda}(z-\omega)^{-3}$, manifestly periodic; $\wp$ itself is even and periodic (Gathmann, Remark 6.5.4).

Worked example 5.20: The torus satisfies a cubic equation

Step 1: Laurent series. Put $G_k = \sum_{\omega \neq 0} \omega^{-k}$ (zero for odd k by $\omega \to -\omega$). Expanding each term of (5.9) in powers of z/ω,

$$\begin{aligned} \wp(z) - \frac{1}{z^2} &\overset{(1)}{=} \sum_{\omega \neq 0} \frac{1}{\omega^2} \left(\sum_{k \geq 0} (k+1) \left(\frac{z}{\omega} \right)^k - 1 \right) \\ &\overset{(2)}{=} \sum_{k \geq 1} (k+1) G_{k+2}\, z^k \\ &\overset{(3)}{=} 3G_4 z^2 + 5G_6 z^4 + O(z^6). \end{aligned}$$

(1) $(z-\omega)^{-2} = \omega^{-2}(1 - z/\omega)^{-2}$ and $(1-u)^{-2} = \sum_{k\geq 0}(k+1)u^k$.

(2) The $k = 0$ term cancels the -1; exchange the sums (absolute convergence).
(3) Odd G's vanish.

Define $g_2 = 60G_4$ and $g_3 = 140G_6$. Then $\wp = z^{-2} + \frac{g_2}{20}z^2 + \frac{g_3}{28}z^4 + \dots$ and $\wp' = -2z^{-3} + \frac{g_2}{10}z + \frac{g_3}{7}z^3 + \dots$.

Step 2: compare poles. Square and cube:

$$\wp'^2 \overset{(1)}{=} 4z^{-6} - \tfrac{2g_2}{5}z^{-2} - \tfrac{4g_3}{7} + O(z^2),$$
$$4\wp^3 \overset{(2)}{=} 4z^{-6} + \tfrac{3g_2}{5}z^{-2} + \tfrac{3g_3}{7} + O(z^2),$$
$$\wp'^2 - 4\wp^3 + g_2\wp + g_3 \overset{(3)}{=} \left(-\tfrac{2}{5} - \tfrac{3}{5} + 1\right)g_2 z^{-2} + \left(-\tfrac{4}{7} - \tfrac{3}{7} + 1\right)g_3 + O(z^2) = O(z^2).$$

(1) $(-2z^{-3})^2 = 4z^{-6}$; cross terms $2(-2z^{-3})(\frac{g_2}{10}z) = -\frac{2g_2}{5}z^{-2}$ and $2(-2z^{-3})(\frac{g_3}{7}z^3) = -\frac{4g_3}{7}$.
(2) $\wp^3 = z^{-6} + 3z^{-4}\left(\frac{g_2}{20}z^2 + \frac{g_3}{28}z^4\right) + O(z^2)$, times 4.
(3) Add, using $g_2\wp = g_2 z^{-2} + O(z^2)$.

Step 3: Liouville. The function $\wp'^2 - 4\wp^3 + g_2\wp + g_3$ is elliptic, has no pole at 0 (hence none anywhere, by periodicity), so it is a holomorphic function on the compact torus, hence constant; and it vanishes at $z = 0$. So

$$\wp'(z)^2 = 4\wp(z)^3 - g_2\wp(z) - g_3. \tag{5.10}$$

The map $z \mapsto [1 : \wp(z) : \wp'(z)]$ (with $0 \mapsto [0 : 0 : 1]$) sends the torus into the cubic $Y^2Z = 4X^3 - g_2XZ^2 - g_3Z^3$ in $\mathbb{CP}^2$, and one checks it is a bijection (Gathmann, Prop. 6.5.7): each value of $\wp$ is taken at two points $\pm z$, and $\wp'$, being odd, separates them.

Worked example 5.21: From Weierstrass to Nash's Legendre form

The half-periods are zeros of $\wp'$: since $\wp'$ is odd and periodic,

$$\wp'\left(\tfrac{\omega_1}{2}\right) \overset{(1)}{=} \wp'\left(\tfrac{\omega_1}{2} - \omega_1\right) = \wp'\left(-\tfrac{\omega_1}{2}\right) \overset{(2)}{=} -\wp'\left(\tfrac{\omega_1}{2}\right) \implies \wp'\left(\tfrac{\omega_1}{2}\right) = 0,$$

(1) Periodicity.
(2) Oddness.

and likewise at $\omega_2/2$ and $\omega_3/2 = (\omega_1 + \omega_2)/2$. So the cubic $4x^3 - g_2x - g_3$ has the three roots $e_i = \wp(\omega_i/2)$, and they are distinct (if $e_1 = e_2$, the function $\wp - e_1$, which has only one double pole per parallelogram, would have two double zeros; but we show in Section 5.2.8 that zeros and poles balance). Substitute $x = e_1 + (e_2 - e_1)X$ and $y = 2(e_2 - e_1)^{3/2}Y$:

$$y^2 \overset{(1)}{=} 4(x - e_1)(x - e_2)(x - e_3)$$
$$\overset{(2)}{=} 4(e_2 - e_1)^3\, X\,(X - 1)\,(X - \lambda), \qquad \lambda = \frac{e_3 - e_1}{e_2 - e_1}.$$

(1) (5.10) with $x = \wp$, $y = \wp'$, and the factorisation by roots.
(2) $x - e_1 = (e_2 - e_1)X$, $x - e_2 = (e_2 - e_1)(X - 1)$, $x - e_3 = (e_2 - e_1)(X - \lambda)$.

Dividing by $4(e_2 - e_1)^3$ gives $Y^2 = X(X - 1)(X - \lambda)$, which is Nash's (5.3). Distinct e_i means $\lambda \neq 0, 1$, matching the smoothness test of Example 5.15. So *the constant λ in (5.3) is a function of the lattice*, that is, of τ.

Watch out: λ is not quite a coordinate on the moduli space

Relabelling the half-periods permutes e_1, e_2, e_3 and replaces λ by one of λ, $1-\lambda$, $1/\lambda$, $1/(1-\lambda)$, $\lambda/(\lambda-1)$, $(\lambda-1)/\lambda$. These six values describe the same torus. The combination that is truly invariant is the j-invariant $j = 256\,(\lambda^2-\lambda+1)^3/\big(\lambda^2(\lambda-1)^2\big)$. Nash calls λ a "local coordinate" on the family, which is the right word: locally, away from the special values, λ determines the torus. The global structure of the family is the subject of part 5b.

Genus 2 and higher. Nash says every Σ_p with $p \geq 2$ can be embedded in $\mathbb{CP}^3$. The reason is general: any smooth curve in $\mathbb{CP}^n$ with $n \geq 4$ can be projected from a point not on it (and not on any of its secant or tangent lines) to $\mathbb{CP}^{n-1}$ injectively; the union of all secant and tangent lines has complex dimension at most 3 (two points on the curve plus one point on the line joining them), so in $\mathbb{CP}^n$ with $n \geq 4$ such a point exists. One cannot in general go down to $\mathbb{CP}^2$: by the degree–genus formula (5.6), only the genera $0, 1, 3, 6, 10, \ldots$ occur for smooth plane curves. A genus-2 surface, for example, is best written as the hyperelliptic curve $y^2 = f(x)$ with f of degree 5 or 6, which is smooth in a weighted projective space rather than $\mathbb{CP}^2$.

5.2.4 How many complex structures? The number $3p-3$

A topological surface of genus p can carry many inequivalent complex structures: we just saw a one-parameter family of tori, labelled by λ (or τ). Nash quotes the answer: for $p = 1$ the family is one dimensional, and for $p > 1$ it has complex dimension $3p-3$. The space of isomorphism classes is the **moduli space** $\mathcal{M}_p$; its covering space is the **Teichmüller space** $\mathcal{T}_p$, studied in part 5b. Riemann's original heuristic for the number $3p-3$ is a nice parameter count.

Worked example 5.22: Riemann's count of moduli

Present Σ_p as a d-sheeted branched cover $f : \Sigma_p \to \mathbb{CP}^1$ (a meromorphic function of degree d), with only simple branch points. The Euler-characteristic count of Example 5.19, done for d sheets and b simple branch points, gives

$$2-2p \overset{(1)}{=} d\,\chi(\mathbb{CP}^1) - b = 2d - b \quad\Longrightarrow\quad b \overset{(2)}{=} 2p-2+2d.$$

(1) Each simple branch point loses one preimage out of d.
(2) Solve for b.

Conversely, a list of b branch points in $\mathbb{CP}^1$ (plus finite combinatorial data about which sheets are joined) determines the pair (Σ_p, f). So pairs (surface, function) form a family of dimension b. How many functions of degree d does a *fixed* surface have? We will derive in Section 5.3.3 that for large d the answer is $2d+1-p$ (choose a degree-d line bundle, p parameters, then two independent sections of it, $2(d+1-p)$ parameters, up to a common scale, -1). So

$$\begin{aligned}\dim \mathcal{M}_p &\overset{(1)}{=} b - (2d+1-p)\\ &\overset{(2)}{=} (2p-2+2d) - (2d+1-p) = 3p-3.\end{aligned}$$

(1) Dimension of the family of pairs minus the dimension of the fibre (functions on a fixed surface).
(2) Substitute b.

The degree d drops out, as it must. For $p = 1$ the formula gives 0, but the torus has a one-dimensional group of automorphisms (translations), each of which can be composed

with f; the correct statement is $\dim \mathcal{M}_p - \dim \mathrm{Aut}(\Sigma_p) = 3p-3$, which gives $1-1=0$ for $p=1$ and $0-3=-3$ for $p=0$ (the sphere has the three-dimensional automorphism group $\mathrm{PSL}(2,\mathbb{C})$). This is exactly the pattern of the string path integral: $3p-3$ complex moduli to integrate over, or conformal Killing vectors to divide by.

5.2.5 Holomorphic line bundles

Nash now says that complex manifolds are "often studied by constructing holomorphic bundles over" them. Let us recall what that means, concretely.

Working definition 5.14 (holomorphic line bundle)

Cover M by open sets U_α. A **holomorphic line bundle** L is specified by **transition functions** $g_{\alpha\beta} : U_\alpha \cap U_\beta \to \mathbb{C}^*$ [$\mathbb{C}^*$ is the nonzero complex numbers. $g_{\alpha\beta}$ converts the patch-β value of a section into the patch-α value], holomorphic and nowhere zero, satisfying the **cocycle condition**

$$g_{\alpha\beta}\, g_{\beta\gamma} = g_{\alpha\gamma} \text{ on } U_\alpha \cap U_\beta \cap U_\gamma, \qquad g_{\alpha\alpha} = 1. \tag{5.11}$$

A **section** s is a collection of local functions $s_\alpha : U_\alpha \to \mathbb{C}$ with $s_\alpha = g_{\alpha\beta}\, s_\beta$ on overlaps. It is holomorphic (meromorphic) if every s_α is. *Operationally*: a line bundle is "a function that picks up a factor $g_{\alpha\beta}$ when you change patch", like a wavefunction that picks up a gauge phase.

Two sets of transition functions describe the same (isomorphic) bundle if $g'_{\alpha\beta} = h_\alpha g_{\alpha\beta} h_\beta^{-1}$ for nowhere-zero holomorphic h_α on U_α: this is a change of local frame, the analogue of a gauge transformation. Line bundles can be multiplied ($L \otimes L'$ has transition functions $g_{\alpha\beta} g'_{\alpha\beta}$) and inverted ($L^{-1} = L^*$ has $g_{\alpha\beta}^{-1}$), so isomorphism classes form an abelian group. The toolkit chapters describe bundles and connections in general (Chapter T2, Chapter T3); sheaves of sections are in Chapter 3.

Worked example 5.23: The line bundles $\mathcal{O}(n)$ on $\mathbb{CP}^1$

Use the two charts of Example 5.9: U_0 with coordinate z, U_1 with $w = 1/z$. For each integer n define $\mathcal{O}(n)$ by the single transition function

$$g_{01}(z) = z^n \qquad \text{on } U_0 \cap U_1 = \mathbb{C}^*.$$

(With only two sets the cocycle condition is empty.) A holomorphic section is a pair $s_0(z)$ holomorphic on $\mathbb{C}$, $s_1(w)$ holomorphic on $\mathbb{C}$, with $s_0 = z^n s_1$. Write $s_0(z) = \sum_{k\ge 0} a_k z^k$. Then

$$s_1(w) \overset{(1)}{=} z^{-n} s_0(z) \overset{(2)}{=} \sum_{k\ge 0} a_k z^{k-n} \overset{(3)}{=} \sum_{k\ge 0} a_k\, w^{n-k}.$$

(1) The transition rule solved for s_1.
(2) Substitute the Taylor series of s_0.
(3) $z = 1/w$.

This is holomorphic at $w=0$ if and only if $a_k = 0$ for $k > n$. So *the holomorphic sections of $\mathcal{O}(n)$ are the polynomials of degree $\le n$ in z*: a vector space of dimension $n+1$ for $n \ge 0$, and $\{0\}$ for $n < 0$. In homogeneous coordinates, $s_0(z) = \sum a_k z^k$ is the dehomogenisation of the degree-n form $\sum a_k z_0^{n-k} z_1^k$ (recall $z = z_1/z_0$): sections of $\mathcal{O}(n)$ are homogeneous polynomials of degree n, which are exactly the objects whose zero sets define varieties.

A section with $s_0 = z^k$ has a zero of order k at $z = 0$ and, since $s_1 = w^{n-k}$, a zero of order $n - k$ at ∞: always n zeros in total.

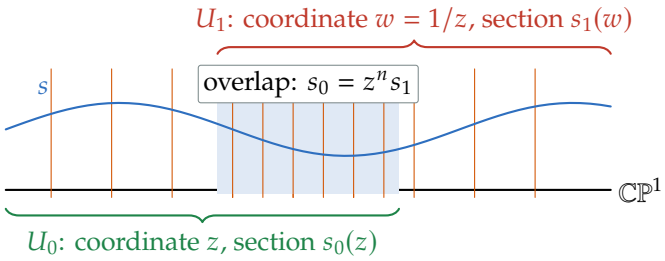


Figure 5.18: A line bundle over $\mathbb{CP}^1$ is two trivial pieces, one over each chart (orange fibres), glued over the overlap by multiplying the fibre coordinate by z^n. A section (blue) is a pair of local functions related by that factor. The integer n is all the information in the gluing: it is the winding number of $g_{01} = z^n$ around the equator $|z| = 1$, and it will turn out to be the degree, or first Chern class, of the bundle.

Worked example 5.24: The tautological bundle is $\mathcal{O}(-1)$

Over each point $[z_0 : z_1]$ of $\mathbb{CP}^1$ put the line it names, $\{\lambda(z_0, z_1)\} \subset \mathbb{C}^2$. This is the **tautological line bundle**. Over U_0 a nonvanishing section is $e_0 = (1, z)$; over U_1 it is $e_1 = (w, 1)$. On the overlap, $e_0 = (1, z) = z\,(1/z, 1) = z\,e_1$. A general point of the fibre is $\xi = s_0 e_0 = s_1 e_1$, so $s_1 e_1 = s_0\, z\, e_1$, that is $s_0 = z^{-1} s_1$. Comparing with Example 5.23, the transition function is z^{-1}: the tautological bundle is $\mathcal{O}(-1)$, and it has no holomorphic sections. Its dual is $\mathcal{O}(1)$, whose sections are linear forms $az_0 + bz_1$: a linear form eats a vector on the line and returns a number.

5.2.6 The canonical bundle

One line bundle is attached to every complex manifold for free. On a complex manifold of dimension n with local coordinates $z^1, \dots, z^n$, the **holomorphic cotangent bundle** T^*M has local frame $\mathrm{d}z^1, \dots, \mathrm{d}z^n$. Under a holomorphic change of coordinates $z \mapsto w(z)$, $\mathrm{d}w^i = \sum_j (\partial w^i / \partial z^j) \mathrm{d}z^j$, so its transition functions are the holomorphic Jacobian matrices. The top exterior power has the single frame $\mathrm{d}z^1 \wedge \cdots \wedge \mathrm{d}z^n$, which transforms by the determinant:

$$\mathrm{d}w^1 \wedge \cdots \wedge \mathrm{d}w^n = \det\left(\frac{\partial w^i}{\partial z^j}\right) \mathrm{d}z^1 \wedge \cdots \wedge \mathrm{d}z^n .$$

Working definition 5.15 (canonical bundle)

The **canonical bundle** is $K_M = \Lambda^n T^*M$, the top exterior power of the holomorphic cotangent bundle. It has rank $\binom{n}{n} = 1$: it is always a line bundle. Its transition functions are the Jacobian determinants $\det(\partial z_\beta / \partial z_\alpha)$. A holomorphic section of K_M is a holomorphic n-form $f(z)\,\mathrm{d}z^1 \wedge \cdots \wedge \mathrm{d}z^n$, that is a form of type $(n, 0)$ [n factors of $\mathrm{d}z$ and none of $\mathrm{d}\bar{z}$]. In sheaf language (Chapter 3), the sheaf of holomorphic sections [the rule that assigns to each open set U the holomorphic sections over U] is $\mathcal{O}(K_M) = \Omega^n$, the sheaf of holomorphic n-forms.

For a Riemann surface $n = 1$, so $T^*\Sigma_p$ is already a line bundle and $K_{\Sigma_p} = T^*\Sigma_p$. Its sections are the expressions $f(z)\,\mathrm{d}z$, and under a change of coordinates $f(z)\,\mathrm{d}z = f(z(w))\,z'(w)\,\mathrm{d}w$.

Worked example 5.25: $K_{\mathbb{CP}^1} = O(-2)$

On the overlap of the two charts of $\mathbb{CP}^1$, $z = 1/w$, so

$$\mathrm{d}z \overset{(1)}{=} -\frac{\mathrm{d}w}{w^2} \overset{(2)}{=} -z^2\,\mathrm{d}w.$$

(1) Differentiate $z = w^{-1}$.
(2) $w^{-2} = z^2$.

A 1-form written $s_0\,\mathrm{d}z$ on U_0 and $s_1\,\mathrm{d}w$ on U_1 satisfies $s_1\,\mathrm{d}w = s_0\,\mathrm{d}z = -z^2 s_0\,\mathrm{d}w$, so $s_0 = -z^{-2}s_1$. The transition function is $-z^{-2}$. The sign is a change of frame on U_1 ($\mathrm{d}w \to -\mathrm{d}w$), so $K_{\mathbb{CP}^1} \cong O(-2)$. By Example 5.23 it has no holomorphic sections: *there is no nonzero holomorphic 1-form on the sphere*. Directly: $\mathrm{d}z$ itself becomes $-\mathrm{d}w/w^2$, a double pole at ∞. The meromorphic form $\mathrm{d}z$ has divisor $-2\cdot\infty$, of degree $-2 = 2\cdot 0 - 2$.

Worked example 5.26: The torus has trivial canonical bundle

On $T = \mathbb{C}/\Lambda$ the changes of chart are translations $z \mapsto z + \omega$, whose derivative is 1. So the transition functions of K_T are all 1, and $\mathrm{d}z$ is a global holomorphic 1-form that never vanishes: K_T is trivial, and its divisor has degree $0 = 2\cdot 1 - 2$. A torus is the one-dimensional example of a Calabi–Yau manifold (trivial canonical bundle, Kähler). For the cubic model, $\mathrm{d}z = \mathrm{d}\wp/\wp' = \mathrm{d}x/y$: the holomorphic 1-form on $y^2 = q(x)$ is $\mathrm{d}x/y$. We will meet its generalisations to higher genus in Section 5.2.16.

The pattern $\deg K = -2, 0, \ldots$ for $p = 0, 1$ continues as $\deg K_{\Sigma_p} = 2p - 2$; we derive this in Section 5.3.3. Nash ends this part of his discussion by noting that one must consider many holomorphic line bundles on Σ_p besides K, and that on algebraic varieties they "can be studied by the closely related notion of divisors". That is next.

5.2.7 Divisors: bookkeeping for zeros and poles

Here is the problem divisors solve. A line bundle L on Σ_p is a collection of transition functions $g_{\alpha\beta}$, and many different collections describe the same bundle. That is a clumsy way to hold on to it. A section is more tangible: in each patch it is a function, and the one thing about a function that does not depend on the patch or the frame is *where it vanishes and where it blows up, and to what order*. That list of points and integers is the divisor. We will see that it remembers the bundle completely.

Take a holomorphic line bundle $L \to \Sigma_p$ and a meromorphic section s (in every patch s_α is meromorphic, and $s_\alpha = g_{\alpha\beta}s_\beta$). Since $g_{\alpha\beta}$ is holomorphic and nowhere zero, s_α and s_β have the same order at every point of the overlap. So the order $\operatorname{ord}_P s := \operatorname{ord}_P s_\alpha$ (computed in any patch containing P) is well defined. On a compact surface a meromorphic section that is not identically zero has only finitely many zeros and poles (they are isolated, and a compact set cannot contain infinitely many isolated points).

Working definition 5.16 (divisor, degree)

A **divisor** on Σ_p is a finite formal sum of points with [a finite list of points, each carrying an integer. The points themselves are never added] integer coefficients,

$$D = \sum_i n_i P_i, \qquad n_i \in \mathbb{Z}.$$

Divisors are added pointwise: $(2P - Q) + (Q + R) = 2P + R$. So they form an abelian group $\mathrm{Div}(\Sigma_p)$, the free abelian group on the points of Σ_p [each element is, in exactly one way, such a finite integer combination of points]. The **degree** is the sum of the coefficients, $\deg D = \sum_i n_i$. A divisor is **effective**, written $D \geq 0$, if all $n_i \geq 0$.
The **divisor of a meromorphic section** $s \not\equiv 0$ is

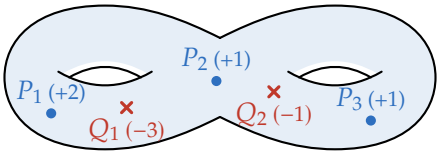
$$\mathrm{div}(s) = \sum_P \mathrm{ord}_P(s)\, P, \tag{5.12}$$

the sum running over the finitely many zeros and poles. Its degree is

$$\deg\big(\mathrm{div}(s)\big) = \sum_P \mathrm{ord}_P(s) = (\text{number of zeros}) - (\text{number of poles}), \tag{5.13}$$

counted with multiplicity. These are Nash's (5.4) and (5.5). *Operationally*: to find $\mathrm{div}(s)$, list the zeros and poles of s in local coordinates and record each with its order (positive for zeros, negative for poles).

A meromorphic function is a meromorphic section of the trivial bundle, so it has a divisor too. A meromorphic 1-form $f(z)\,dz$ is a meromorphic section of K_{Σ_p}, and its divisor records the zeros and poles of the coefficient f (in every chart).



$$D = 2P_1 + P_2 + P_3 - 3Q_1 - Q_2, \qquad \deg D = 4 - 4 = 0$$

Figure 5.19: A divisor is a finite list of points with integer weights: blue dots are zeros (positive weight), red crosses are poles (negative weight). Nothing else about the section is recorded. The divisor shown has degree 0, as the divisor of a meromorphic *function* always must (Section 5.2.8).

Worked example 5.27: Divisors on the Riemann sphere

A rational function. Let $f(z) = \dfrac{z^2(z-1)}{(z+1)^4}$ on $\mathbb{CP}^1$. In $\mathbb{C}$ it has a double zero at 0, a simple zero at 1 and a pole of order 4 at -1. At infinity, in the coordinate $w = 1/z$,

$$f \overset{(1)}{=} \frac{w^{-2}(w^{-1}-1)}{(w^{-1}+1)^4} \overset{(2)}{=} \frac{w^{-3}(1-w)}{w^{-4}(1+w)^4} \overset{(3)}{=} w \cdot \frac{1-w}{(1+w)^4},$$

(1) Substitute $z = 1/w$.
(2) Pull out the powers of w^{-1}: $w^{-1} - 1 = w^{-1}(1-w)$ and $w^{-1} + 1 = w^{-1}(1+w)$.
(3) Collect: $w^{-3}/w^{-4} = w$; the remaining factor is holomorphic and equals 1 at $w = 0$.

so f has a simple zero at ∞. Therefore

$$\mathrm{div}(f) = 2\cdot[0] + [1] - 4\cdot[-1] + [\infty], \qquad \deg = 2 + 1 - 4 + 1 = 0.$$

(We write $[a]$ for the point $z = a$ when a bare number would be ambiguous.)
A section of $\mathcal{O}(n)$. By Example 5.23 a holomorphic section of $\mathcal{O}(n)$ is a polynomial $s_0(z) = c\prod_{k=1}^{m}(z - a_k)$ of degree $m \leq n$ in the chart U_0, and $s_1(w) = w^n s_0(1/w)$ in U_1. Near

$w = 0$, $s_1 = w^n \cdot c\, w^{-m} \prod(1 - a_k w) = c\, w^{n-m} \prod_k (1 - a_k w)$, a zero of order $n - m$. So

$$\operatorname{div}(s) = \sum_{k=1}^{m} [a_k] + (n - m)[\infty], \qquad \deg = n.$$

Every nonzero holomorphic section of $\mathcal{O}(n)$ has exactly n zeros. The degree of the divisor does not depend on the section; it is a property of the bundle.
The 1-form $\mathrm{d}z$. In Example 5.25 we found $\mathrm{d}z = -w^{-2}\mathrm{d}w$: no zeros or poles in $\mathbb{C}$, a double pole at ∞. So $\operatorname{div}(\mathrm{d}z) = -2[\infty]$ and $\deg K_{\mathbb{CP}^1} = -2$.

Worked example 5.28: Divisors on the cubic torus

Let E be Nash's cubic $y^2 = q(x) = x(x-1)(x-\lambda)$ with its point at infinity $O = [0:1:0]$. Write $A_0 = (0,0)$, $A_1 = (1,0)$, $A_\lambda = (\lambda, 0)$.
Local coordinates. At an affine point with $y \neq 0$, $\partial F/\partial y = -2y \neq 0$, so x is a local coordinate (implicit function theorem). At A_0, A_1, A_λ we have $y = 0$ but $\partial F/\partial x = q'(x) \neq 0$, so y is a local coordinate. At O use the chart $Y = 1$ with $X = x/y$, $Z = 1/y$: the curve reads $Z = X(X - Z)(X - \lambda Z)$, so $Z = X^3 + O(X^4)$ and $t = X = x/y$ is a local coordinate at O, with

$$y \overset{(1)}{=} \frac{1}{Z} \overset{(2)}{=} t^{-3}\big(1 + O(t)\big), \qquad x \overset{(3)}{=} \frac{X}{Z} \overset{(4)}{=} t^{-2}\big(1 + O(t)\big).$$

(1) Definition of Z.
(2) $Z = t^3(1 + O(t))$.
(3) $X/Z = (x/y) \cdot y = x$.
(4) $X = t$ and $Z = t^3(1 + O(t))$.

The function y. It vanishes where $q(x) = 0$, i.e. at the three points A_0, A_1, A_λ, and there it is the local coordinate, so each zero is simple. Its only pole is at O, of order 3:

$$\operatorname{div}(y) = A_0 + A_1 + A_\lambda - 3\,O.$$

The function x. It vanishes only at A_0. There the local coordinate is y, and $x = y^2/\big((x-1)(x-\lambda)\big) = y^2/\lambda + O(y^4)$: a double zero. Its only pole is at O, of order 2:

$$\operatorname{div}(x) = 2A_0 - 2\,O.$$

Similarly $\operatorname{div}(x - 1) = 2A_1 - 2O$ and $\operatorname{div}(x - \lambda) = 2A_\lambda - 2O$. All four divisors have degree 0.
The 1-form $\mathrm{d}x/y$. At a point with $y \neq 0$, x is a coordinate and $\mathrm{d}x/y$ has coefficient $1/y$, finite and nonzero. At A_0 (and likewise A_1, A_λ), differentiate $y^2 = q(x)$ to get $2y\,\mathrm{d}y = q'(x)\,\mathrm{d}x$, so $\mathrm{d}x/y = 2\,\mathrm{d}y/q'(x)$, finite and nonzero because $q'(0) = \lambda \neq 0$. At O, $\mathrm{d}x = (-2t^{-3} + \dots)\mathrm{d}t$, so $\mathrm{d}x/y = (-2 + O(t))\mathrm{d}t$, again finite and nonzero. So

$$\operatorname{div}(\mathrm{d}x/y) = 0.$$

This is the nowhere-vanishing holomorphic 1-form of Example 5.26, now seen on the cubic: K_E is trivial, $\deg K_E = 0$.

5.2.8 The divisor of a function has degree zero

In every example so far, the divisor of a meromorphic *function* had degree zero. Nash states this as a general fact and gives a contour-integral argument. We present it in full; it is the model for many arguments on Riemann surfaces (the "residue theorem on a compact surface").

Key idea: zeros and poles balance

If $f \not\equiv 0$ is a meromorphic function on a compact Riemann surface, then

$$\deg \operatorname{div}(f) = 0 :$$

f has as many zeros as poles, counted with multiplicity.

Step 1: the logarithmic derivative counts zeros and poles. Near a point P with $\operatorname{ord}_P f = n$, write $f = z^n u(z)$ with u holomorphic and $u(0) \neq 0$. Then

$$\frac{\mathrm{d}f}{f} \overset{(1)}{=} \frac{nz^{n-1}u\,\mathrm{d}z + z^n u'\,\mathrm{d}z}{z^n u} \overset{(2)}{=} n\,\frac{\mathrm{d}z}{z} + \frac{u'}{u}\,\mathrm{d}z, \tag{5.14}$$

(1) Product rule for $\mathrm{d}(z^n u)$.
(2) Cancel $z^n u$; u'/u is holomorphic near 0 because $u(0) \neq 0$.

So the 1-form $\mathrm{d}f/f$ has a simple pole at every zero and pole of f, with residue equal to the order. Integrating over a small circle $|z| = \varepsilon$ taken anticlockwise,

$$\frac{1}{2\pi\mathrm{i}} \oint_{|z|=\varepsilon} \frac{\mathrm{d}f}{f} \overset{(1)}{=} \frac{n}{2\pi\mathrm{i}} \oint_{|z|=\varepsilon} \frac{\mathrm{d}z}{z} + \frac{1}{2\pi\mathrm{i}} \oint_{|z|=\varepsilon} \frac{u'}{u}\mathrm{d}z \overset{(2)}{=} \frac{n}{2\pi\mathrm{i}} \int_0^{2\pi} \mathrm{i}\,\mathrm{d}\theta + 0 = n.$$

(1) Substitute (5.14).
(2) On the circle $z = \varepsilon e^{\mathrm{i}\theta}$, $\mathrm{d}z/z = \mathrm{i}\,\mathrm{d}\theta$; the second integrand is holomorphic inside the disc, so its integral vanishes (Cauchy, or Stokes: $\oint_{\partial B} \alpha = \int_B \mathrm{d}\alpha$ and $\mathrm{d}(g\,\mathrm{d}z) = \bar\partial g \wedge \mathrm{d}z = 0$ for holomorphic g).

Two further facts make $\mathrm{d}f/f$ useful globally. It is a genuine 1-form on Σ_p minus the zeros and poles, independent of the chart (it is $\mathrm{d}\log f$, and f is a function). And it is *closed* there, because it is locally exact: $\mathrm{d}f/f = \mathrm{d}(\log f)$ for any local branch of the logarithm. So, by Stokes' theorem, its integral over a closed curve depends only on the homology class of the curve in the punctured surface. This is the statement "$\frac{1}{2\pi\mathrm{i}}\int_C \mathrm{d}f/f$ measures $n_0 - n_p$" in Nash: for a curve C bounding a region R that contains some of the zeros and poles, Stokes on R with small discs removed gives

$$\frac{1}{2\pi\mathrm{i}} \oint_C \frac{\mathrm{d}f}{f} = n_0(R) - n_p(R), \tag{5.15}$$

the number of zeros minus the number of poles inside R (the **argument principle**).

Step 2: genus 0. On the sphere Σ_0 let F be the finite set of zeros and poles. Choose a closed curve C that encloses all of F on one side, say in a region R. The complementary region $R' = \Sigma_0 \setminus R$ is a disc containing no point of F (Figure 5.20). Then

$$n_0 - n_p \overset{(1)}{=} \frac{1}{2\pi\mathrm{i}} \oint_C \frac{\mathrm{d}f}{f} \overset{(2)}{=} -\frac{1}{2\pi\mathrm{i}} \oint_{\partial R'} \frac{\mathrm{d}f}{f} \overset{(3)}{=} -\frac{1}{2\pi\mathrm{i}} \int_{R'} \mathrm{d}\Big(\frac{\mathrm{d}f}{f}\Big) \overset{(4)}{=} 0.$$

(1) (5.15) applied to R, which contains all zeros and poles.

(2) C is the boundary of R; as the boundary of R' it carries the opposite orientation.
(3) Stokes' theorem on R', where $\mathrm{d}f/f$ is smooth.
(4) $\mathrm{d}f/f$ is closed.

In Nash's words, C is homologous to a point in $\Sigma_0 - F$: we can slide it over the back of the sphere and shrink it to nothing without crossing F.

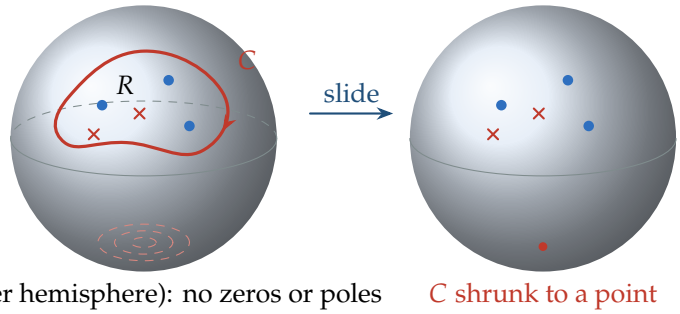


Figure 5.20: The degree-zero argument on the sphere. The red curve C encloses all zeros (blue) and poles (red crosses) in the region R, so $\frac{1}{2\pi\mathrm{i}}\oint_C \mathrm{d}f/f = n_0 - n_p$. But the other side R' is a disc with no zeros or poles, and C can be slid over it and shrunk to a point. The integral of the closed form $\mathrm{d}f/f$ does not change, so it is zero.

Step 3: genus $p > 0$. Now a closed curve need not bound a disc, so the sphere argument fails. Instead Nash uses the standard model of Σ_p as a polygon with $4p$ sides, identified in pairs.

Working definition 5.17 (the $4p$-gon model of Σ_p)

Take a polygon with $4p$ edges, labelled in order around the boundary

$$a_1\, b_1\, a_1^{-1}\, b_1^{-1}\, a_2\, b_2\, a_2^{-1}\, b_2^{-1} \cdots a_p\, b_p\, a_p^{-1}\, b_p^{-1}.$$

Glue the edge a_i to the edge a_i^{-1} and b_i to b_i^{-1}, where the exponent -1 means "the same edge, traversed in the opposite direction" (Figure 5.21). The result is Σ_p. For $p = 1$ this is the square $a\,b\,a^{-1}\,b^{-1}$ with opposite sides identified: the torus. After gluing, all $4p$ vertices become one point, and the $2p$ edges become the $2p$ loops $a_1, b_1, \ldots, a_p, b_p$ that generate the homology [every closed loop on Σ_p counts, up to deformation and cancellation of opposite pieces, as an integer combination of these $2p$ loops] (Section 5.2.15).

Cutting Σ_p along the loops a_i, b_i produces the polygon Π, which is a disc. A meromorphic function f on Σ_p becomes a meromorphic function on Π which takes *equal values at identified boundary points*. Move the loops slightly, if necessary, so that no zero or pole lies on the boundary of Π. Then all zeros and poles lie inside the disc Π, and (5.15) applies to $C = \partial\Pi$:

$$n_0 - n_p \overset{(1)}{=} \frac{1}{2\pi\mathrm{i}} \oint_{\partial\Pi} \frac{\mathrm{d}f}{f} \overset{(2)}{=} \frac{1}{2\pi\mathrm{i}} \sum_{i=1}^{p} \Big(\int_{a_i} + \int_{b_i} + \int_{a_i^{-1}} + \int_{b_i^{-1}} \Big) \frac{\mathrm{d}f}{f} \overset{(3)}{=} 0.$$

(1) (5.15): Π contains all zeros and poles.
(2) The boundary of Π is the union of its $4p$ edges.
(3) On a_i and a_i^{-1} the function f takes the same values at corresponding points, so $\mathrm{d}f/f$ is the same 1-form there; but the two edges are traversed in opposite directions, so $\int_{a_i^{-1}} = -\int_{a_i}$. Likewise for b_i.

Nash phrases this as pushing the contour C out until it coincides with the sides of the polygon, and observing that "the contribution from each side is cancelled by another of opposite orientation". This completes the proof that $\deg \operatorname{div}(f) = 0$ for every genus.

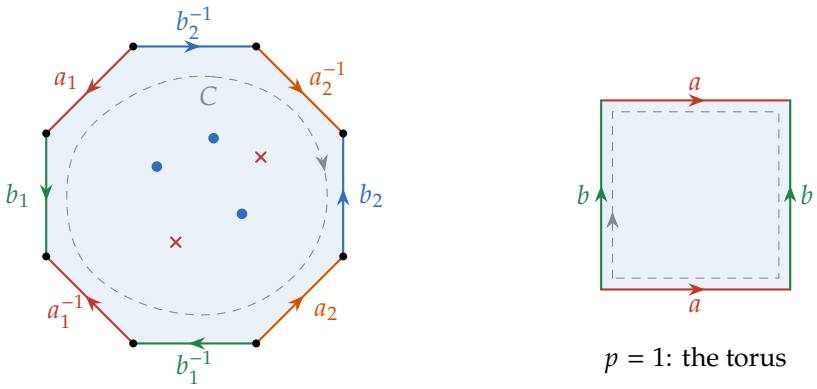


Figure 5.21: Left: the genus-2 surface cut open into an octagon with edges $a_1 b_1 a_1^{-1} b_1^{-1} a_2 b_2 a_2^{-1} b_2^{-1}$; edges of the same colour are glued so that their arrows match. The dashed contour C encloses all zeros and poles; pushing it out to the boundary, the two copies of each edge are traversed in opposite directions and cancel. Right: for $p = 1$ the polygon is the square with opposite sides identified, the torus.

Worked example 5.29: The cancellation on the torus, written out

On $T = \mathbb{C}/\Lambda$, $\Lambda = \mathbb{Z}\omega_1 + \mathbb{Z}\omega_2$, take the fundamental parallelogram Π with corners z_0, $z_0 + \omega_1$, $z_0 + \omega_1 + \omega_2$, $z_0 + \omega_2$, where the base point z_0 is chosen so that no zero or pole of the elliptic function f lies on $\partial\Pi$. Let $g = f'/f$, which is elliptic because f is. The boundary integral splits into two pairs of opposite edges:

$$\begin{aligned}\oint_{\partial\Pi} g(z)\,\mathrm{d}z &\overset{(1)}{=} \int_{z_0}^{z_0+\omega_1} g(z)\,\mathrm{d}z + \int_{z_0+\omega_1}^{z_0+\omega_1+\omega_2} g(z)\,\mathrm{d}z - \int_{z_0+\omega_1+\omega_2}^{z_0+\omega_2} g(z)\,\mathrm{d}z - \int_{z_0}^{z_0+\omega_2} g(z)\,\mathrm{d}z \\ &\overset{(2)}{=} \int_{z_0}^{z_0+\omega_1} \big[g(z) - g(z+\omega_2)\big]\mathrm{d}z + \int_{z_0}^{z_0+\omega_2} \big[g(z+\omega_1) - g(z)\big]\mathrm{d}z \\ &\overset{(3)}{=} 0.\end{aligned}$$

(1) Go round $\partial\Pi$ anticlockwise; the top and left edges are traversed backwards, hence the minus signs.
(2) Substitute $z \to z + \omega_2$ in the third integral and $z \to z + \omega_1$ in the second.
(3) g is periodic, so each bracket vanishes identically.

So every elliptic function has equally many zeros and poles in a period parallelogram. For example $\wp$ has a double pole at 0 and hence exactly two zeros; $\wp - e_1$ has the double pole and a double zero at $\omega_1/2$ (a zero of $\wp - e_1$ that is also a zero of $\wp'$), and no other zeros. This is the fact we used in Example 5.21 to show e_1, e_2, e_3 are distinct.

Try it: residues of a meromorphic 1-form

Repeat the polygon argument for an arbitrary meromorphic 1-form η on Σ_p (not necessarily of the form $\mathrm{d}f/f$). What does it prove?
Answer. η is the same form on identified edges, so $\oint_{\partial\Pi} \eta = 0$; by the residue theorem inside the disc Π this says $\sum_P \operatorname{Res}_P \eta = 0$. *The residues of a meromorphic 1-form on a compact Riemann surface sum to zero.* The degree-zero theorem is the special case $\eta = \mathrm{d}f/f$, whose residues

are the orders.

5.2.9 Changing the section: linear equivalence

The degree-zero theorem has an immediate consequence for line bundles. Let s and s' be two meromorphic sections of the same bundle L, with $s \not\equiv 0$. In each patch the ratio s'_α / s_α is meromorphic, and on overlaps

$$\frac{s'_\alpha}{s_\alpha} \overset{(1)}{=} \frac{g_{\alpha\beta} s'_\beta}{g_{\alpha\beta} s_\beta} = \frac{s'_\beta}{s_\beta},$$

(1) Both sections obey the transition rule with the same $g_{\alpha\beta}$.

so the local ratios patch together into a global meromorphic *function* $f = s'/s$. This is Nash's (5.6):

$$s' = f\, s, \qquad f \text{ a meromorphic function.} \tag{5.16}$$

Orders add under multiplication ($z^m u \cdot z^n v = z^{m+n} uv$), so

$$\operatorname{div}(s') \overset{(1)}{=} \operatorname{div}(fs) \overset{(2)}{=} \operatorname{div}(f) + \operatorname{div}(s), \tag{5.17}$$

$$\deg \operatorname{div}(s') \overset{(3)}{=} \deg \operatorname{div}(f) + \deg \operatorname{div}(s) \overset{(4)}{=} \deg \operatorname{div}(s). \tag{5.18}$$

(1) (5.16).
(2) $\operatorname{ord}_P(fs) = \operatorname{ord}_P f + \operatorname{ord}_P s$ at every point.
(3) deg is additive (it is a sum of coefficients).
(4) The degree-zero theorem of Section 5.2.8.

These are Nash's (5.7) and (5.8). Two conclusions:

- The integer $\deg L := \deg \operatorname{div}(s)$ does not depend on which meromorphic section we use. It is an invariant of the bundle, the **degree** of L.
- The divisor itself does depend on s, but only up to adding the divisor of a function.

Working definition 5.18 (linear equivalence; principal divisor)

A divisor of the form $\operatorname{div}(f)$ for a meromorphic function f is **principal**. [it is the zeros-minus-poles record of one single global function] Two divisors are **linearly equivalent**, $D_1 \sim D_2$, if $D_1 - D_2 = \operatorname{div}(f)$ for some meromorphic f. *Operationally*: to show $D_1 \sim D_2$, exhibit a function with zeros D_1 and poles D_2 (after cancelling common points). Equivalent divisors have equal degree. By (5.17), a line bundle determines a well-defined **divisor class** [the set of all divisors linearly equivalent to one given divisor] $\{\operatorname{div}(s)\}$ modulo $\sim$.

Worked example 5.30: On $\mathbb{CP}^1$ the degree is everything

Let $D = \sum_i n_i [a_i] + n_\infty [\infty]$ be any divisor of degree 0 on $\mathbb{CP}^1$, with $a_i \in \mathbb{C}$. Put $f(z) = \prod_i (z - a_i)^{n_i}$. In $\mathbb{C}$ its divisor is $\sum n_i [a_i]$. At ∞ it behaves like $z^{\sum n_i} = w^{-\sum n_i}$, so

$$\operatorname{div}(f) \overset{(1)}{=} \sum_i n_i [a_i] - \Big(\sum_i n_i\Big)[\infty] \overset{(2)}{=} \sum_i n_i [a_i] + n_\infty [\infty] = D.$$

(1) Orders of f at the finite points and at ∞.
(2) $\deg D = 0$ means $n_\infty = -\sum_i n_i$.

So every degree-zero divisor on the sphere is principal, and two divisors on $\mathbb{CP}^1$ are equivalent if and only if they have the same degree. Combined with the next subsection, this says that *every holomorphic line bundle on $\mathbb{CP}^1$ is $\mathcal{O}(n)$ for exactly one n*, namely its degree.

Worked example 5.31: On the torus the degree is not everything

On $T = \mathbb{C}/\Lambda$ let $P \neq Q$ be two points. Is $P - Q$ principal? It would be the divisor of an elliptic function f with a single simple zero at P and a single simple pole at Q. Consider the meromorphic 1-form $f\,dz$ (dz is a global nowhere-zero holomorphic 1-form on T). Its only pole is at Q, simple, with residue $\operatorname{Res}_Q f \neq 0$. But residues of a meromorphic 1-form sum to zero (the *Try it* above). Contradiction. So

$$P - Q \not\sim 0 \qquad \text{for } P \neq Q,$$

although $\deg(P - Q) = 0$. There are therefore many inequivalent divisor classes of degree 0 on a torus, one for each point P if we fix $Q = O$: we will see in Section 5.2.19 that $P - O \sim P' - O$ only if $P = P'$, and that the degree-zero classes form a copy of the torus itself.

5.2.10 From a divisor to a line bundle

We now go the other way. Given a divisor $D = \sum_i n_i P_i$ (Nash's (5.9)), we build a line bundle $[D]$ and a meromorphic section of it whose divisor is exactly D.

Working definition 5.19 (the line bundle $[D]$)

Cover Σ_p by coordinate discs U_α. [patches shaped like discs, each carrying one complex coordinate z] In each U_α choose a **local defining function**: a meromorphic function f_α on U_α whose divisor is D restricted to U_α. Concretely, with z the coordinate on U_α,

$$f_\alpha(z) = \prod_{P_i \in U_\alpha} \big(z - z(P_i)\big)^{n_i} \qquad (f_\alpha = 1 \text{ if } U_\alpha \text{ contains no } P_i).$$

On an overlap, f_α and f_β have the same divisor, so their ratio has no zeros and no poles. Define

$$g_{\alpha\beta} = \frac{f_\alpha}{f_\beta} \qquad \text{on } U_\alpha \cap U_\beta. \tag{5.19}$$

This is Nash's (5.10). The line bundle with these transition functions [each $g_{\alpha\beta}$ is holomorphic and nowhere zero, and $g_{\alpha\beta}g_{\beta\gamma} = g_{\alpha\gamma}$ holds automatically] is $[D]$, and the collection $s_D = \{f_\alpha\}$ is a meromorphic section of it with $\operatorname{div}(s_D) = D$, the **canonical section** of $[D]$.

Let us check each claim.

- *$g_{\alpha\beta}$ is holomorphic and nowhere zero.* Near any point of $U_\alpha \cap U_\beta$ write $f_\alpha = z^n u$, $f_\beta = z^n v$ with the same n (same divisor) and u, v holomorphic and nonzero; then $g_{\alpha\beta} = u/v$.

- *Cocycle condition.* On triple overlaps $g_{\alpha\beta}g_{\beta\gamma} = \frac{f_\alpha}{f_\beta}\frac{f_\beta}{f_\gamma} = \frac{f_\alpha}{f_\gamma} = g_{\alpha\gamma}$, automatically. This is (5.11).

- *s_D is a section.* We need $f_\alpha = g_{\alpha\beta} f_\beta$, which is (5.19) rearranged.

- *Independence of choices.* Another choice of defining functions is $f'_\alpha = h_\alpha f_\alpha$ with h_α holomorphic and nowhere zero on U_α. Then $g'_{\alpha\beta} = h_\alpha g_{\alpha\beta} h_\beta^{-1}$, which is the same bundle in a different frame (Section 5.2.5).

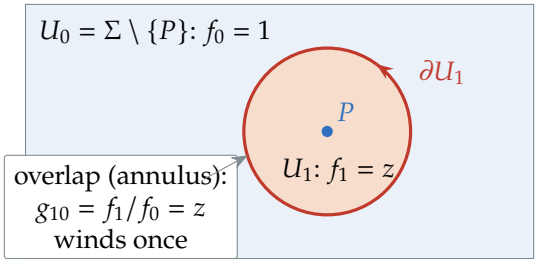


Figure 5.22: The line bundle $[P]$ of a single point. Over the disc U_1 around P the defining function is the coordinate z; everywhere else it is 1. The transition function on the annulus where the two sets overlap is $g_{10} = z$, which winds once around P. That winding number is the degree of $[P]$, and the canonical section (equal to z near P, to 1 elsewhere) has a single simple zero at P.

The dictionary. The following rules follow directly from the construction.

- $[D_1 + D_2] = [D_1] \otimes [D_2]$ and $[-D] = [D]^{-1}$: defining functions multiply, so transition functions multiply.
- $[\operatorname{div} f]$ is trivial: take $f_\alpha = f$ in every patch; then every $g_{\alpha\beta} = 1$.
- Hence $D_1 \sim D_2 \Rightarrow [D_1] \cong [D_2]$.
- Conversely, if L has a meromorphic section $s \not\equiv 0$, then $L \cong [\operatorname{div} s]$: the local functions s_α are themselves local defining functions for $\operatorname{div}(s)$, and $s_\alpha / s_\beta = g_{\alpha\beta}$ are L's own transition functions.

On a compact Riemann surface every holomorphic line bundle has a nonzero meromorphic section (a theorem we will not prove; for $\mathcal{O}(n)$ and K we have seen them). So divisors modulo linear equivalence and holomorphic line bundles up to isomorphism are the same thing:

$$\frac{\operatorname{Div}(\Sigma_p)}{\{\text{principal divisors}\}} \;\cong\; \{\text{holomorphic line bundles on } \Sigma_p\}/\!\cong, \qquad D \mapsto [D]. \tag{5.20}$$

This is what Nash means by saying that "one may work either with (equivalence classes of) divisors or with holomorphic line bundles depending on which is more natural for the problem at hand".

Worked example 5.32: $[n \cdot \infty] = \mathcal{O}(n)$ and $[\operatorname{div} dz] = K$

On $\mathbb{CP}^1$ take $D = n[\infty]$ and the two charts U_0 (z) and U_1 ($w = 1/z$). Local defining functions are $f_0 = 1$ (no point of D in U_0) and $f_1 = w^n$. By (5.19),

$$g_{01} \overset{(1)}{=} \frac{f_0}{f_1} \overset{(2)}{=} w^{-n} \overset{(3)}{=} z^n.$$

(1) Definition (5.19).
(2) $f_0 = 1$, $f_1 = w^n$.
(3) $w = 1/z$.

This is exactly the transition function of $\mathcal{O}(n)$ in Example 5.23. So $[n \cdot \infty] = \mathcal{O}(n)$, and

the canonical section is $s_0 = 1$, $s_1 = w^n$: the constant polynomial 1, vanishing to order n at ∞. Choosing instead $D = n[0]$ gives $f_0 = z^n$, $f_1 = 1$, $g_{01} = z^n$ again, as it must since $n[0] \sim n[\infty]$ (the function z^n has divisor $n[0] - n[\infty]$).
Similarly, since $\operatorname{div}(\mathrm{d}z) = -2[\infty]$, $K_{\mathbb{CP}^1} = [-2 \cdot \infty] = \mathcal{O}(-2)$, confirming Example 5.25. On the cubic torus $\operatorname{div}(\mathrm{d}x/y) = 0$ gives $K_E = [0]$, the trivial bundle. In general $K_{\Sigma_p} = [\operatorname{div}\eta]$ for any meromorphic 1-form η; the class of such divisors is the **canonical divisor class**, usually written K as well.

5.2.11 The degree is the first Chern class

A line bundle has two integer invariants we now know how to compute: the degree of the divisor of any meromorphic section, and the first Chern class $c_1(L) = \frac{\mathrm{i}}{2\pi}F$ integrated over the surface (Chapter T2, Chapter T3). Nash's (5.11) says they agree:

$$c_1([D]) = \deg D, \tag{5.21}$$

where on the left we mean $\int_{\Sigma_p} c_1$. Nash says this "can be proved by starting with D and constructing its transition functions according to (5.10); then one applies the usual curvature formula". Let us do exactly that.

The curvature of a holomorphic line bundle. Give L a Hermitian metric: in each patch a positive function h_α such that the length of a section is $\|s\|^2 = h_\alpha |s_\alpha|^2$. Independence of the patch requires

$$h_\alpha |s_\alpha|^2 = h_\beta |s_\beta|^2 \quad \Longleftrightarrow \quad h_\beta = |g_{\alpha\beta}|^2 h_\alpha. \tag{5.22}$$

There is a unique connection compatible with both the metric and the holomorphic structure (the **Chern connection**); in the holomorphic frame of U_α its connection 1-form is $A_\alpha = \partial \log h_\alpha$, and its curvature is

$$F = \mathrm{d}A_\alpha = \bar\partial\partial \log h_\alpha = -\partial\bar\partial \log h_\alpha, \qquad c_1(L) = \frac{\mathrm{i}}{2\pi}F = -\frac{\mathrm{i}}{2\pi}\,\partial\bar\partial \log h_\alpha. \tag{5.23}$$

Here $\partial = \mathrm{d}z\,\partial_z$ and $\bar\partial = \mathrm{d}\bar z\,\partial_{\bar z}$ are the holomorphic and antiholomorphic parts of $\mathrm{d} = \partial + \bar\partial$ (more on these in Section 5.3.1). Two checks: F is the same in every patch, because by (5.22) $\log h_\beta - \log h_\alpha = \log g_{\alpha\beta} + \log \bar g_{\alpha\beta}$, and $\partial\bar\partial$ kills both a holomorphic and an antiholomorphic function. And F is imaginary (the U(1) curvature is anti-Hermitian, as in our gauge convention), because $\overline{\partial\bar\partial\phi} = \bar\partial\partial\phi = -\partial\bar\partial\phi$ for real ϕ. So c_1 is a real 2-form.

Worked example 5.33: The curvature of $\mathcal{O}(n)$ integrates to n

Give $\mathcal{O}(n)$ the metric $\|s\|^2 = |s_0|^2/(1+|z|^2)^n$, i.e. $h_0 = (1+|z|^2)^{-n}$. (It is consistent: with $s_1 = w^n s_0$ and $|z|^2 = 1/|w|^2$ one finds $h_1 = (1+|w|^2)^{-n}$, the same formula in the other chart.) First the basic derivative:

$$\partial\bar\partial \log(1+z\bar z) \overset{(1)}{=} \partial\Big(\frac{z\,\mathrm{d}\bar z}{1+z\bar z}\Big) \overset{(2)}{=} \frac{(1+z\bar z)-z\bar z}{(1+z\bar z)^2}\,\mathrm{d}z\wedge\mathrm{d}\bar z = \frac{\mathrm{d}z\wedge\mathrm{d}\bar z}{(1+|z|^2)^2}. \tag{5.24}$$

(1) $\partial_{\bar z}\log(1+z\bar z) = z/(1+z\bar z)$.
(2) $\partial_z\big(z/(1+z\bar z)\big) = \big((1+z\bar z)-z\bar z\big)/(1+z\bar z)^2$ by the quotient rule, and the form becomes $\mathrm{d}z\wedge\mathrm{d}\bar z$.

Then

$$\int_{\mathbb{CP}^1} c_1(\mathcal{O}(n)) \overset{(1)}{=} -\frac{\mathrm{i}}{2\pi}\int_{\mathbb{C}} \partial\bar\partial \log(1+|z|^2)^{-n} \overset{(2)}{=} \frac{n\,\mathrm{i}}{2\pi}\int_{\mathbb{C}} \frac{\mathrm{d}z\wedge\mathrm{d}\bar z}{(1+|z|^2)^2}$$
$$\overset{(3)}{=} \frac{n}{\pi}\int_{\mathbb{R}^2}\frac{\mathrm{d}x\,\mathrm{d}y}{(1+x^2+y^2)^2} \overset{(4)}{=} \frac{n}{\pi}\cdot\pi = n.$$

(1) (5.23) in the chart U_0, which misses one point.
(2) $\log h_0 = -n\log(1+|z|^2)$ and (5.24).
(3) $\mathrm{d}z\wedge\mathrm{d}\bar z = -2\mathrm{i}\,\mathrm{d}x\wedge\mathrm{d}y$ and $\mathrm{i}\cdot(-2\mathrm{i}) = 2$.
(4) The integral computed in Example 5.17.

And the divisor of any holomorphic section of $\mathcal{O}(n)$ has degree n (Example 5.27). So $c_1 = \deg$ for all the $\mathcal{O}(n)$. In particular the tautological bundle $\mathcal{O}(-1)$ has $c_1 = -1$: it is the *Dirac monopole* bundle of charge -1 (or $+1$, depending on the sign convention for the charge), and $\frac{\mathrm{i}}{2\pi}F$ is its uniform magnetic field on the sphere.

The general argument. Let s be a meromorphic section of L with zeros and poles at P_i of orders n_i. Away from the P_i, $\|s\|^2$ is a smooth positive function and

$$\log h_\alpha \overset{(1)}{=} \log\|s\|^2 - \log s_\alpha - \log\bar s_\alpha \quad\Longrightarrow\quad \partial\bar\partial\log h_\alpha \overset{(2)}{=} \partial\bar\partial\log\|s\|^2.$$

(1) Take the logarithm of $\|s\|^2 = h_\alpha s_\alpha \bar s_\alpha$ (locally).
(2) $\bar\partial\log s_\alpha = 0$ and $\partial\log\bar s_\alpha = 0$, since s_α is holomorphic and nonzero away from the P_i.

So on $\Sigma^\varepsilon = \Sigma_p \setminus \bigcup_i B_\varepsilon(P_i)$ (small discs removed) the Chern form is the *global* expression $c_1 = -\frac{\mathrm{i}}{2\pi}\partial\bar\partial\log\|s\|^2 = -\frac{\mathrm{i}}{2\pi}\mathrm{d}(\bar\partial\log\|s\|^2)$ (because $\mathrm{d}\bar\partial\phi = \partial\bar\partial\phi + \bar\partial\bar\partial\phi$ and $\bar\partial^2 = 0$). Near P_i, in a coordinate z centred there, $s_\alpha = z^{n_i}u$ and $\log\|s\|^2 = n_i\log|z|^2 + \phi$ with ϕ smooth. Now use Stokes:

$$\int_{\Sigma_p} c_1(L) \overset{(1)}{=} \lim_{\varepsilon\to0}\int_{\Sigma^\varepsilon} -\frac{\mathrm{i}}{2\pi}\,\mathrm{d}\big(\bar\partial\log\|s\|^2\big) \overset{(2)}{=} \lim_{\varepsilon\to0}\frac{\mathrm{i}}{2\pi}\sum_i\oint_{|z|=\varepsilon}\bar\partial\log\|s\|^2$$
$$\overset{(3)}{=} \frac{\mathrm{i}}{2\pi}\sum_i n_i\oint_{|z|=\varepsilon}\frac{\mathrm{d}\bar z}{\bar z} \overset{(4)}{=} \frac{\mathrm{i}}{2\pi}\sum_i n_i\,(-2\pi\mathrm{i}) = \sum_i n_i = \deg\operatorname{div}(s).$$

(1) c_1 is smooth, so removing shrinking discs changes nothing in the limit.
(2) Stokes' theorem; $\partial\Sigma^\varepsilon$ is the union of the small circles traversed *clockwise*, which we rewrite as anticlockwise circles with a minus sign.
(3) $\bar\partial\log|z|^2 = \bar\partial(\log z + \log\bar z) = \mathrm{d}\bar z/\bar z$; the smooth part $\bar\partial\phi$ integrates to $O(\varepsilon)\to0$.
(4) $\oint\mathrm{d}\bar z/\bar z = \overline{\oint\mathrm{d}z/z} = \overline{2\pi\mathrm{i}} = -2\pi\mathrm{i}$.

Applying this to $L = [D]$ and its canonical section s_D gives $c_1([D]) = \deg D$, Nash's (5.11). Applying it to the trivial bundle with $h = 1$ ($F = 0$) and a meromorphic function gives a second proof that $\deg\operatorname{div}(f) = 0$.

Key idea: degree = Chern number = winding number

For a line bundle on a surface,

$$\deg L = \int_\Sigma c_1(L) = \frac{1}{2\pi\mathrm{i}}\oint_{\partial U}\mathrm{d}\log g_{10},$$

where in the last expression Σ is covered by two open sets, a disc $U = U_1$ and the rest U_0,

and g_{10} is the transition function on the annulus between them. For $[P]$ with U_1 a disc around P, $g_{10} = f_1/f_0 = z$ winds once (Figure 5.22). For $\mathcal{O}(n)$ with U_1 the disc around ∞, $g_{10} = z^{-n} = w^n$ winds n times in the coordinate w.

The last equality is the same Stokes argument with the two patches instead of small discs: $\int_{U_1} c_1 + \int_{U_0} c_1 = \frac{i}{2\pi}\oint_{\partial U_1} \bar\partial \log(h_0/h_1)$, and $\log(h_0/h_1) = \log|g_{10}|^2$ by (5.22), whose $\bar\partial$ is $\overline{\mathrm{d}\log g_{10}}$. The physicist will recognise this as the Wu–Yang description of the Dirac monopole: two gauge patches on the sphere, and the monopole charge is the winding number of the U(1) gauge transformation relating them on the equator.

5.2.12 Holomorphic sections and positive degree

Nash next remarks that if $\deg D > 0$ then $[D]$ has a global holomorphic section "and vice versa", and that such a section must vanish somewhere, "otherwise the bundle $[D]$ would be trivial". The second half is simple and important; the first half needs to be stated more carefully. The tool is the following reformulation of sections.

Working definition 5.20 (the space $\mathcal{L}(D)$)

For a divisor D let

$$\mathcal{L}(D) = \{\, f \text{ meromorphic on } \Sigma_p : \ \operatorname{div}(f) + D \geq 0 \,\} \cup \{0\}.$$

If $D = \sum n_i P_i$, this means: f may have a pole of order at most n_i at each P_i with $n_i > 0$, must vanish to order at least $|n_i|$ at each P_i with $n_i < 0$, and is holomorphic everywhere else. It is a complex vector space. [sums and constant multiples of allowed functions are again allowed. That is why 0 is added by hand] *Operationally*: list the allowed poles and the required zeros, and find all functions obeying them.

Key idea: holomorphic sections of $[D]$ = the space $\mathcal{L}(D)$

The map $f \mapsto f\,s_D$ is an isomorphism $\mathcal{L}(D) \cong H^0(\Sigma_p, \mathcal{O}([D]))$, the space of global holomorphic sections of $[D]$. We write $h^0(D) = \dim \mathcal{L}(D)$.

Why: every meromorphic section of $[D]$ is $f\,s_D$ for a meromorphic function f (by (5.16)), and $\operatorname{div}(f\,s_D) = \operatorname{div} f + D$, which is ≥ 0 exactly when $f\,s_D$ is holomorphic. Now the facts:

1. *Negative degree: no sections.* A nonzero holomorphic section has an effective divisor, of degree ≥ 0. So if $\deg L < 0$ then $H^0(L) = 0$. Example: $\mathcal{O}(n)$, $n < 0$; the canonical bundle of the sphere.

2. *Degree zero: sections only if trivial.* If $\deg L = 0$ and $s \neq 0$ is holomorphic, $\operatorname{div}(s)$ is effective of degree 0, hence 0: s vanishes nowhere, and s itself is a nowhere-zero frame, so L is trivial. This is Nash's remark: *a holomorphic section of a nontrivial bundle must vanish somewhere*.

3. *Positive degree.* If D is effective (all $n_i \geq 0$) then the canonical section s_D is holomorphic, so $h^0(D) \geq 1$. More generally $[D]$ has a nonzero holomorphic section if and only if D is linearly equivalent to an effective divisor. Degree ≥ 0 is necessary but, for $p \geq 2$, not sufficient.

Watch out: positive degree does not guarantee a section

Nash's "and vice versa" should be read as: if $[D]$ has a nonzero holomorphic section then $\deg D \geq 0$, with equality only for the trivial bundle; and if $\deg D$ is large enough then there are sections. The precise threshold comes from Riemann–Roch (Section 5.3.3): $\deg D \geq p$ always suffices. Below that it can fail. On a surface of genus 2, the degree-1 bundles form a 2-complex-dimensional family (a copy of the Jacobian, Section 5.3.5), while those with a section are the $[P]$, a 1-dimensional family parametrised by the point P. So a generic degree-1 bundle on Σ_2 has no holomorphic section at all.

Worked example 5.34: $\mathcal{L}(D)$ on the sphere and on the torus

Sphere. $\mathcal{L}(n[\infty])$ consists of the functions holomorphic on $\mathbb{C}$ with at most an nth order pole at ∞. By Example 5.16 these are the polynomials of degree $\leq n$: $h^0 = n + 1$, matching $H^0(\mathcal{O}(n))$ from Example 5.23.
Torus. Let O be the origin of $T = \mathbb{C}/\Lambda$. The space $\mathcal{L}(nO)$ consists of elliptic functions whose only pole is at O, of order at most n. Build it up one n at a time.

- $n = 0$: holomorphic functions on a compact surface are constant: $\mathcal{L}(0) = \mathbb{C} \cdot 1$.
- $n = 1$: a function with a single simple pole would contradict the residue theorem (Example 5.31): $\mathcal{L}(O) = \mathbb{C} \cdot 1$.
- $n = 2, 3, 4, 5, 6$: new functions with poles of exactly that order are $\wp$, $\wp'$, $\wp^2$, $\wp\wp'$, $\wp^3$.

Two functions in $\mathcal{L}(nO)$ with the same leading pole coefficient differ by an element of $\mathcal{L}((n-1)O)$, so each step adds at most one dimension. Hence

$$h^0(nO) = n \qquad (n \geq 1),$$

with basis $1, \wp, \wp', \wp^2, \wp\wp', \wp^3, \ldots$. Now look at $n = 6$. The seven functions $1, \wp, \wp', \wp^2, \wp\wp', \wp^3, \wp'^2$ all lie in the 6-dimensional space $\mathcal{L}(6O)$, so they must satisfy a linear relation. That relation is the Weierstrass equation (5.10), $\wp'^2 = 4\wp^3 - g_2\wp - g_3$. The cubic equation of the torus is *forced* by counting sections. And the three sections $1, \wp, \wp'$ of $[3O]$ give the map $z \mapsto [1 : \wp : \wp']$ into $\mathbb{CP}^2$ of Example 5.20: this is Kodaira's embedding (Section 5.1.10) in its simplest instance.

5.2.13 The sheaf-theoretic viewpoint

Nash rephrases divisors in the language of sheaves (Chapter 3). The point of doing so is that the passage $D \mapsto [D]$ then becomes a standard piece of machinery, the connecting map of a long exact sequence, which generalises to any complex manifold.

Recall two multiplicative sheaves on a complex manifold M:

- $\mathcal{O}^*$: over an open set U, the nowhere-vanishing holomorphic functions on U, with multiplication as the group operation;
- $\mathcal{M}^*$: over U, the meromorphic functions on U that are not identically zero on any component (Nash: "non-trivial"), again under multiplication.

$\mathcal{O}^*$ is a subsheaf of $\mathcal{M}^*$, and we can form the quotient sheaf (Nash's (5.12))

$$\mathcal{D} = \frac{\mathcal{M}^*}{\mathcal{O}^*}. \tag{5.25}$$

Working definition 5.21 (a section of $\mathcal{M}^*/\mathcal{O}^*$)

A global section of $\mathcal{D}$ [a section over all of Σ_p of the quotient sheaf just defined, in which a nonzero meromorphic function counts only up to a holomorphic nowhere-zero factor] is given by an open cover $\{U_\alpha\}$ and nonzero meromorphic functions f_α on U_α such that on each overlap $f_\alpha/f_\beta \in \mathcal{O}^*(U_\alpha \cap U_\beta)$, i.e. is holomorphic and nowhere zero. Two such data define the same section if they differ by multiplying each f_α by an element of $\mathcal{O}^*(U_\alpha)$. *Operationally*: it is exactly a system of local defining functions, and it determines the divisor $\sum_P \operatorname{ord}_P(f_\alpha)\, P$ (which is independent of α since the ratios have order 0).

So a section of $\mathcal{D}$ *is* a divisor (on a Riemann surface, and more generally on a complex manifold: see the next subsection), and Nash's (5.13) is the statement

$$\operatorname{Div}(M) = H^0(M; \mathcal{M}^*/\mathcal{O}^*). \tag{5.26}$$

The exact sequence. The inclusion and the quotient give a short exact sequence of sheaves, and hence (Chapter 3) a long exact sequence in cohomology:

$$0 \to \mathcal{O}^* \to \mathcal{M}^* \to \mathcal{M}^*/\mathcal{O}^* \to 0$$

$$H^0(\mathcal{O}^*) \to H^0(\mathcal{M}^*) \xrightarrow{\text{div}} H^0(\mathcal{M}^*/\mathcal{O}^*) = \operatorname{Div}(M) \xrightarrow{\delta} H^1(\mathcal{O}^*) \to H^1(\mathcal{M}^*)$$

Every map here is something we have already met.

- $H^0(\mathcal{M}^*) \to \operatorname{Div}$ sends a global meromorphic function to its divisor. Its kernel is $H^0(\mathcal{O}^*)$: functions with no zeros or poles, which on a compact M are the nonzero constants.
- The connecting map δ is computed by the standard recipe: lift the section $\{f_\alpha\}$ of the quotient to local sections f_α of $\mathcal{M}^*$, and take the "Čech coboundary" on overlaps, which in multiplicative notation is $f_\alpha f_\beta^{-1}$. The result lies in $\mathcal{O}^*$ and is a cocycle:

  $$\delta\{f_\alpha\} = \{g_{\alpha\beta}\} = \left\{\frac{f_\alpha}{f_\beta}\right\} \in H^1(M; \mathcal{O}^*).$$

  This is precisely (5.19): $\delta(D) = [D]$. A class in $H^1(M; \mathcal{O}^*)$ is a holomorphic line bundle, given by its transition functions modulo changes of frame (Section 5.3.5).
- Exactness at $\operatorname{Div}(M)$ says $\ker\delta = \operatorname{im}(\text{div})$: $[D]$ is trivial if and only if D is principal. This is the injectivity half of (5.20). The surjectivity half is the statement that $H^1(\mathcal{O}^*) \to H^1(\mathcal{M}^*)$ is zero, i.e. every line bundle has a meromorphic section.

Watch out: sections of a quotient sheaf are not quotients of sections

One might guess that $H^0(\mathcal{M}^*/\mathcal{O}^*)$ equals $H^0(\mathcal{M}^*)/H^0(\mathcal{O}^*)$, i.e. global meromorphic functions modulo constants. That is false: the latter gives only the *principal* divisors. A general divisor, like a single point P on a torus, is defined by local functions $(z - z(P))$ near P, 1 elsewhere) that do not come from any global function. The failure of "quotient of sections = sections of quotient" is measured by H^1, and here it is exactly the group of line bundles.

5.2.14 Divisors in higher dimensions

Replace Σ_p by a compact complex manifold M of complex dimension $n > 1$. The zero set of a single holomorphic function is no longer a set of points: one equation cuts one complex dimension, so it is a subvariety of dimension $n - 1$.

Working definition 5.22 (codimension; divisor on M)

The **codimension** of a subvariety $V \subset M$ is $\dim M - \dim V$ (complex dimensions). A **divisor** on M is a finite formal sum

$$D = \sum_i n_i V_i, \qquad n_i \in \mathbb{Z}, \tag{5.27}$$

of irreducible analytic subvarieties V_i [each locally cut out by holomorphic equations, and not the union of two smaller such sets] of codimension one (Nash's (5.14)). The divisor of a meromorphic function f is $\sum_V \operatorname{ord}_V(f)\, V$, where $\operatorname{ord}_V(f)$ is the order of vanishing of f along V (measured in a coordinate transverse to V at a generic smooth point of V) [near such a point $V = \{w = 0\}$ for a coordinate w, and $\operatorname{ord}_V(f)$ is the power of w in f]. For $n = 1$ the codimension-one subvarieties are points, and we recover (5.12).

Locally D is again described by a meromorphic function: a meromorphic function of several complex variables is, near each point, a ratio g/h of holomorphic functions, and its zeros ($g = 0$) and poles ($h = 0$) are each of codimension one. The construction $D \mapsto [D]$ with $g_{\alpha\beta} = f_\alpha / f_\beta$, and the sheaf description $\operatorname{Div}(M) = H^0(M; \mathcal{M}^*/\mathcal{O}^*)$, go through word for word.

Worked example 5.35: Hyperplanes in $\mathbb{CP}^n$ give $\mathcal{O}(1)$

In $\mathbb{CP}^n$ with homogeneous coordinates $[z_0 : \cdots : z_n]$, let $H_j = \{z_j = 0\}$, a copy of $\mathbb{CP}^{n-1}$ of codimension one. The ratio $f = z_1/z_0$ is a meromorphic function on $\mathbb{CP}^n$ (a ratio of two homogeneous polynomials of the same degree is well defined). It vanishes to first order along H_1 and has a first-order pole along H_0:

$$\operatorname{div}(z_1/z_0) = H_1 - H_0,$$

so all hyperplanes are linearly equivalent. Their common line bundle is $\mathcal{O}(1) = [H_0]$: on $U_j = \{z_j \neq 0\}$ a local defining function for H_0 is $f_j = z_0/z_j$, and $g_{jk} = f_j/f_k = z_k/z_j$. A homogeneous polynomial P of degree d defines the hypersurface $V(P)$, and P/z_0^d is a meromorphic function with divisor $V(P) - dH_0$; so $[V(P)] = \mathcal{O}(d) = \mathcal{O}(1)^{\otimes d}$. For $n = 1$ this reduces to Example 5.32.

Worked example 5.36: Zeros and poles can meet

On $\mathbb{C}^2$ take $f = x/y$. Its zero set is the line $\{x = 0\}$ and its pole set the line $\{y = 0\}$, both of codimension one, so $\operatorname{div}(f) = \{x = 0\} - \{y = 0\}$. The two lines meet at the origin, and there f is not even defined as a map to $\mathbb{CP}^1$: along the line $y = kx$ it takes the constant value $1/k$, so every value occurs in every neighbourhood of 0. The origin is a point of **indeterminacy**, of codimension two. In dimension one this never happens (zeros and poles are isolated points, and a point cannot be both).

Nash ends the discussion of divisors with a remark: in high dimensions, working with divisors means dealing with singular sets of codimension one, which are large (for $\dim M = 4$ they are three-dimensional), whereas one might have hoped that the technical difficulties would come from small sets of high codimension. That is the price of the flexibility: codimension one is exactly where the zeros and poles of a single function live.

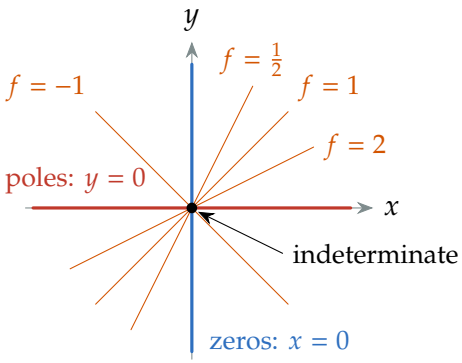


Figure 5.23: The (real picture of the) divisor of $f = x/y$ on $\mathbb{C}^2$: a zero line (blue) and a pole line (red), both of codimension one. The level sets $f = c$ (orange) are the lines $x = cy$, all passing through the origin, where f is indeterminate. Nash's remark that divisors force us to consider "singularities which are not just points" is visible here: the singular set of f is the whole red line.

Reading Nash (pp. 140–142, (5.4)–(5.14))

Nash defines the divisor of a meromorphic section (5.4) and its degree (5.5), and proves that a meromorphic function has a degree-zero divisor with the contour argument: $\frac{1}{2\pi i}\int_C \mathrm{d}f/f = n_0 - n_p$, the contour is contractible in $\Sigma_0 - F$ for genus zero, and for genus $p > 0$ one pushes it to the boundary of the $4p$-gon where opposite sides cancel. Then: $s' = fs$ (5.6), $\mathrm{div}(s') = \mathrm{div}\, f + \mathrm{div}\, s$ (5.7), equal degrees (5.8), linear equivalence, the bundle $[D]$ from $g_{\alpha\beta} = f_\alpha/f_\beta$ (5.9)–(5.10), $c_1([D]) = \deg D$ (5.11), sections of positive-degree bundles, the sheaf $\mathcal{D} = \mathcal{M}^*/\mathcal{O}^*$ (5.12) with $\mathrm{Div}(M) = H^0(M; \mathcal{M}^*/\mathcal{O}^*)$ (5.13), and codimension-one divisors (5.14) in higher dimension. We have added: explicit divisors on the sphere and on the cubic (with local coordinates at infinity), the explicit torus cancellation, the residue theorem for 1-forms, the classification of line bundles on $\mathbb{CP}^1$ by degree, the non-principal divisor $P - Q$ on the torus, a complete curvature computation of $c_1 = \deg$ (including the winding-number form), the spaces $\mathcal{L}(D)$ with the derivation of the Weierstrass cubic by counting sections, the connecting homomorphism, and $\mathcal{O}(1) = [H]$ on $\mathbb{CP}^n$. We have also corrected Nash's "and vice versa" (positive degree does not by itself guarantee a holomorphic section when $p \geq 2$).

5.2.15 Homology of Σ_p and intersection numbers

Divisors and line bundles describe the complex structure of Σ_p from the point of view of functions and sections. Nash now turns to a second, complementary description: integrate the holomorphic 1-forms around the independent loops of the surface. The resulting numbers, the periods, encode the complex structure in a finite matrix. First we need the loops.

Working definition 5.23 (first homology)

$H_1(\Sigma_p; \mathbb{Z})$ is the group of closed oriented curves on Σ_p (and integer combinations of them), where two combinations are regarded as equal if their difference is the boundary of a piece of surface. *Operationally*: a loop is zero in homology if it cuts off a region of the surface; two loops are equal if together (one reversed) they bound a region. The integral of a *closed* 1-form [θ with $\mathrm{d}\theta = 0$] over a loop depends only on its homology class, by Stokes' theorem: if $c_1 - c_2 = \partial S$ then $\int_{c_1}\theta - \int_{c_2}\theta = \int_S \mathrm{d}\theta = 0$.

For the sphere with p handles,

$$H_1(\Sigma_p;\mathbb{Z}) = \underbrace{\mathbb{Z}\oplus\cdots\oplus\mathbb{Z}}_{2p} = \mathbb{Z}^{2p},$$

with a standard basis of $2p$ loops $a_1, b_1, \ldots, a_p, b_p$: a_i goes around the ith handle (around the "tube"), and b_i goes through the ith hole (along the tube and back through the body), as in Figure 5.24. In the $4p$-gon model of Figure 5.21 they are the images of the edges. The count $2p$ also follows from the Euler characteristic: with one vertex, $2p$ edges and one face, the edges are all cycles and the face's boundary $a_1 b_1 a_1^{-1} b_1^{-1}\cdots$ is zero in homology (each edge appears twice with opposite signs), so no relation is imposed and $H_1 = \mathbb{Z}^{2p}$.

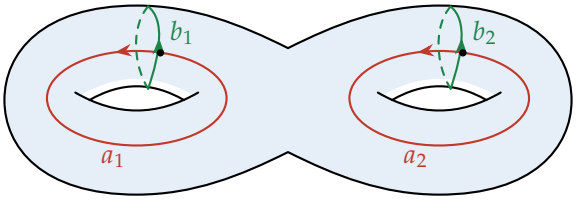


Figure 5.24: The standard homology basis of Σ_2. The red loop a_i goes around the ith hole; the green loop b_i goes over the ith handle, through the hole (dashed part on the back). Each a_i crosses the corresponding b_i exactly once (black dot) and misses all the other loops: $(a_i, b_j) = \delta_{ij}$, $(a_i, a_j) = (b_i, b_j) = 0$.

Working definition 5.24 (intersection number)

For two oriented closed curves a, b meeting transversally, [crossing at a nonzero angle, never touching tangentially] the **intersection number** (a, b) is the number of crossing points, each counted $+1$ if the pair (tangent of a, tangent of b) is a positively oriented basis of the tangent plane there and -1 otherwise. It depends only on the homology classes, and it is bilinear and antisymmetric [$(a + a', b) = (a, b) + (a', b)$, and likewise in b]: $(b, a) = -(a, b)$. *Operationally*: draw both curves, and at each crossing ask whether turning a's direction anticlockwise by less than π brings it to b's direction.

For the standard basis the only crossings are between a_i and b_i, once, which gives Nash's (5.15):

$$(a_i, a_j) = (b_i, b_j) = 0, \qquad (a_i, b_j) = -(b_j, a_i) = \delta_{ij}. \tag{5.28}$$

In matrix form, ordering the basis as $(a_1, \ldots, a_p, b_1, \ldots, b_p)$, the intersection form is

$$J = \begin{pmatrix} 0 & I_p \\ -I_p & 0 \end{pmatrix},$$

the standard symplectic matrix. Changes of homology basis that preserve (5.28) are integer matrices M with $M^T J M = J$: the group $\mathrm{Sp}(2p, \mathbb{Z})$. For $p = 1$ this is $\mathrm{SL}(2, \mathbb{Z})$.

Worked example 5.37: Intersection numbers on the torus

Take $T = \mathbb{R}^2/\mathbb{Z}^2$, a the horizontal loop $t \mapsto (t, 0)$, b the vertical loop $t \mapsto (0, t)$. They cross once at the origin with tangents $(1, 0)$ and $(0, 1)$, a positive basis, so $(a, b) = 1$. A straight loop of slope n/m (winding m times horizontally and n times vertically) is the class $ma + nb$. For two such classes,

$$(ma + nb,\ m'a + n'b) \overset{(1)}{=} mm'(a,a) + mn'(a,b) + nm'(b,a) + nn'(b,b)$$

$$\overset{(2)}{=} mn' - nm' = \det \begin{pmatrix} m & m' \\ n & n' \end{pmatrix}.$$

(1) Bilinearity.
(2) (5.28): $(a,a) = (b,b) = 0$, $(a,b) = 1$, $(b,a) = -1$.

For example the diagonal loop $a + b$ meets a with $(a, a+b) = 1$ and b with $(a+b, b) = 1$: draw it on the square and see one crossing each. The determinant formula also shows that $(ma + nb, m'a + n'b) = \pm 1$ exactly when the two classes form a basis of $\mathbb{Z}^2$.

5.2.16 Holomorphic differentials

Next Nash passes "to the dual": 1-forms, which are integrated over loops. By de Rham's theorem (Chapter T2), $H^1(\Sigma_p; \mathbb{C})$, the closed complex 1-forms modulo exact ones, has complex dimension $2p$, and integration over the $2p$ basis loops identifies it with $\mathbb{C}^{2p}$. On a Riemann surface there is a special class of closed 1-forms.

Working definition 5.25 (holomorphic differential)

A **holomorphic 1-form** (holomorphic differential, **Abelian differential of the first kind**, in Nash's footnote) is a global holomorphic section of K_{Σ_p}: [the canonical bundle, whose local frame is $\mathrm{d}z$] in each chart $\omega = f(z)\,\mathrm{d}z$ with f holomorphic, and $f(z)\,\mathrm{d}z = \tilde{f}(w)\,\mathrm{d}w$ on overlaps. *Operationally*: write it in each chart and check there are no poles, including at the points at infinity [the points added when the affine equation of the curve is completed in projective space].

Every holomorphic 1-form is closed:

$$\mathrm{d}\big(f\,\mathrm{d}z\big) \overset{(1)}{=} \Big(\partial_z f\,\mathrm{d}z + \partial_{\bar{z}} f\,\mathrm{d}\bar{z}\Big) \wedge \mathrm{d}z \overset{(2)}{=} \partial_{\bar{z}} f\,\mathrm{d}\bar{z} \wedge \mathrm{d}z \overset{(3)}{=} 0.$$

(1) $\mathrm{d}f = \partial_z f\,\mathrm{d}z + \partial_{\bar{z}} f\,\mathrm{d}\bar{z}$.
(2) $\mathrm{d}z \wedge \mathrm{d}z = 0$.
(3) f is holomorphic: $\partial_{\bar{z}} f = 0$ (Cauchy–Riemann).

So it defines a class in $H^1(\Sigma_p; \mathbb{C})$ and can be integrated over homology classes. Its complex conjugate $\bar{\omega} = \bar{f}\,\mathrm{d}\bar{z}$ is closed too.

Key idea: there are exactly p holomorphic differentials

The holomorphic 1-forms on Σ_p form a complex vector space of dimension p (the genus). If $\omega_1, \dots, \omega_p$ is a basis, then the $2p$ classes $\omega_1, \dots, \omega_p, \bar{\omega}_1, \dots, \bar{\omega}_p$ form a basis of $H^1(\Sigma_p; \mathbb{C})$.

The dimension count $h^0(K) = p$ is a consequence of Riemann–Roch (Section 5.3.3), or of the Hodge decomposition $H^1 = H^{1,0} \oplus H^{0,1}$ (Section 5.3.9). Here is the part we can see directly: the forms ω and $\bar{\omega}$ can never be cohomologous to zero, and cannot cancel each other. The key is a positivity computation. For $\omega = f\,\mathrm{d}z$, with $z = x + \mathrm{i}y$,

$$\mathrm{i}\,\omega \wedge \bar{\omega} \overset{(1)}{=} \mathrm{i}\,|f|^2\,\mathrm{d}z \wedge \mathrm{d}\bar{z} \overset{(2)}{=} \mathrm{i}\,|f|^2\,(-2\mathrm{i})\,\mathrm{d}x \wedge \mathrm{d}y = 2|f|^2\,\mathrm{d}x \wedge \mathrm{d}y. \tag{5.29}$$

(1) $\omega \wedge \bar{\omega} = f\bar{f}\,\mathrm{d}z \wedge \mathrm{d}\bar{z}$.
(2) $\mathrm{d}z \wedge \mathrm{d}\bar{z} = (\mathrm{d}x + \mathrm{i}\,\mathrm{d}y) \wedge (\mathrm{d}x - \mathrm{i}\,\mathrm{d}y) = -\mathrm{i}\,\mathrm{d}x \wedge \mathrm{d}y + \mathrm{i}\,\mathrm{d}y \wedge \mathrm{d}x = -2\mathrm{i}\,\mathrm{d}x \wedge \mathrm{d}y$.

So $\mathrm{i}\int_\Sigma \omega\wedge\bar\omega>0$ unless $\omega=0$. If ω were exact, $\omega=\mathrm{d}F$, then $\int\omega\wedge\bar\omega=\int\mathrm{d}(F\bar\omega)=0$ by Stokes (and $\mathrm{d}\bar\omega=0$), forcing $\omega=0$. A similar argument (using that $\omega_i\wedge\omega_j=0$, both being of type $(1,0)$) shows that no nonzero combination $\sum c_i\omega_i+\sum d_i\bar\omega_i$ is exact.

Worked example 5.38: Holomorphic differentials on hyperelliptic curves

Let Σ be the curve $y^2=f(x)$ with f a polynomial of degree $2p+1$ with distinct roots; by Example 5.19 it has genus p and one point at infinity, ∞, which is a branch point. Consider

$$\omega_k=\frac{x^k\,\mathrm{d}x}{y},\qquad k=0,1,2,\dots$$

Finite points with $y\neq0$: x is a local coordinate and x^k/y is finite. *Branch points $y=0$:* differentiating $y^2=f(x)$ gives $2y\,\mathrm{d}y=f'(x)\,\mathrm{d}x$, so $\mathrm{d}x/y=2\,\mathrm{d}y/f'(x)$, finite since $f'\neq0$ at a simple root, and y is the local coordinate. *At infinity*: since ∞ is a branch point, a local coordinate is t with $x=t^{-2}$, and then $y^2=f(x)\sim x^{2p+1}=t^{-4p-2}$ gives $y\sim t^{-(2p+1)}$ (up to a nonzero constant and higher-order corrections). So

$$\omega_k\overset{(1)}{=}\frac{t^{-2k}\cdot(-2t^{-3}\,\mathrm{d}t)}{t^{-(2p+1)}\,(1+O(t))}\overset{(2)}{=}-2\,t^{\,2p-2-2k}\,\big(1+O(t)\big)\,\mathrm{d}t.$$

(1) $x^k=t^{-2k}$, $\mathrm{d}x=-2t^{-3}\mathrm{d}t$, and the behaviour of y.
(2) Collect the powers: $-2k-3+2p+1=2p-2-2k$.

This is holomorphic at $t=0$ if and only if $2p-2-2k\geq0$, i.e. $k\leq p-1$. So

$$\frac{\mathrm{d}x}{y},\ \frac{x\,\mathrm{d}x}{y},\ \dots,\ \frac{x^{p-1}\,\mathrm{d}x}{y}$$

are p linearly independent holomorphic differentials, in agreement with the key idea. For $p=1$ this is $\mathrm{d}x/y$ on the cubic (Example 5.28). For genus 2, with $y^2=x^5+\dots$, the basis is $\mathrm{d}x/y$ and $x\,\mathrm{d}x/y$. Bonus: the divisor of $\omega_0=\mathrm{d}x/y$ is $(2p-2)\cdot\infty$ (it has no zeros at finite points and a zero of order $2p-2$ at ∞), so $\deg K_\Sigma=2p-2$, the value Riemann–Roch will predict.

5.2.17 The period matrix

Working definition 5.26 (period matrix)

Choose a basis $\omega_1,\dots,\omega_p$ of holomorphic differentials [there are exactly p independent ones on Σ_p] and the standard homology basis $a_1,\dots,a_p,b_1,\dots,b_p$. [the loops of Figure 5.24, a_i around the ith hole and b_i over the ith handle] The **period matrix** is the $p\times2p$ matrix of all their integrals (Nash's (5.16)):

$$\Omega=\begin{pmatrix}\int_{a_1}\omega_1&\cdots&\int_{a_p}\omega_1&\int_{b_1}\omega_1&\cdots&\int_{b_p}\omega_1\\ \vdots&&\vdots&\vdots&&\vdots\\ \int_{a_1}\omega_p&\cdots&\int_{a_p}\omega_p&\int_{b_1}\omega_p&\cdots&\int_{b_p}\omega_p\end{pmatrix}_{p\times2p}.\tag{5.30}$$

Row j lists the periods of the form ω_j; column k lists the integrals of all the forms over one loop. *Operationally*: parametrise each loop and integrate each form along it.

The $2p$ columns $\Pi_1,\dots,\Pi_{2p}\in\mathbb{C}^p$ are the **periods**. Nash says "it can be verified that the Π_i

are linearly independent over $\mathbb{R}$". Here is the verification. Regard $\mathbb{C}^p$ as $\mathbb{R}^{2p}$; the $2p$ real vectors are independent if and only if the $2p \times 2p$ matrix obtained by stacking Ω on top of its complex conjugate is invertible:

$$\sum_k r_k \Pi_k = 0 \text{ for real } r_k \overset{(1)}{\iff} \sum_k r_k \Pi_k = 0 \text{ and } \sum_k r_k \bar{\Pi}_k = 0 \overset{(2)}{\iff} \begin{pmatrix} \Omega \\ \bar{\Omega} \end{pmatrix} r = 0.$$

(1) For real r_k the second equation is the complex conjugate of the first.
(2) Write the two vector equations as one matrix equation; conversely, any solution $r \in \mathbb{C}^{2p}$ of the stacked system has $\bar{r}$ as a solution too, so the real and imaginary parts of r are real solutions.

The stacked matrix is the matrix of integrals of the $2p$ forms $\omega_1, \dots, \omega_p, \bar{\omega}_1, \dots, \bar{\omega}_p$ over the $2p$ loops. These forms are a basis of $H^1(\Sigma_p; \mathbb{C})$ and the loops a basis of H_1; by de Rham's theorem the pairing between them is nondegenerate, so the matrix is invertible. Hence the periods are $\mathbb{R}$-independent.

Working definition 5.27 (Jacobian variety)

The integer combinations of the periods [$\Pi_k \in \mathbb{C}^p$ is column k of Ω, the integrals of all p forms over one loop] form a lattice of rank $2p$ in $\mathbb{C}^p$, [$2p$ vectors independent over $\mathbb{R}$, together with all their integer combinations]

$$\Lambda = \left\{ \sum_{k=1}^{2p} m_k \Pi_k : m_k \in \mathbb{Z} \right\}, \qquad J(\Sigma_p) = \frac{\mathbb{C}^p}{\Lambda}, \tag{5.31}$$

and the quotient is a compact complex torus of complex dimension p (real dimension $2p$), the **Jacobian variety** of Σ_p. This is Nash's (5.17). $\mathbb{R}$-independence of the $2p$ periods is exactly what makes the quotient compact: the periods span $\mathbb{C}^p = \mathbb{R}^{2p}$ over $\mathbb{R}$, so a fundamental domain [a region that contains one representative of each point of the quotient] is a bounded parallelepiped.

Worked example 5.39: The torus is its own Jacobian

On $T = \mathbb{C}/\Lambda_\tau$, $\Lambda_\tau = \mathbb{Z} + \tau\mathbb{Z}$, the space of holomorphic differentials is spanned by $\omega = \mathrm{d}z$ ($p = 1$). Take a to be the image of the segment from 0 to 1 and b the image of the segment from 0 to τ (each closes up in T). The period matrix is 1×2:

$$\Omega = \left(\int_a \mathrm{d}z \quad \int_b \mathrm{d}z \right) = (1 \;\; \tau).$$

The two periods 1 and τ are $\mathbb{R}$-independent because $\operatorname{Im}\tau \neq 0$. The lattice is $\Lambda = \mathbb{Z} + \tau\mathbb{Z}$ again, so $J(T) = \mathbb{C}/(\mathbb{Z} + \tau\mathbb{Z}) = T$. For the cubic model $y^2 = x(x-1)(x-\lambda)$, the periods of $\mathrm{d}x/y$ are twice the integrals between branch points (the loop a encircles the slit from 0 to 1, and on the two sheets y has opposite signs):

$$\int_a \frac{\mathrm{d}x}{y} = 2\int_0^1 \frac{\mathrm{d}x}{\sqrt{x(x-1)(x-\lambda)}}, \qquad \int_b \frac{\mathrm{d}x}{y} = 2\int_1^\lambda \frac{\mathrm{d}x}{\sqrt{x(x-1)(x-\lambda)}},$$

elliptic integrals of the first kind. Their ratio $\tau(\lambda)$ is the modulus of the torus: this is how the parameter λ of (5.3) determines the lattice.

Normalising the basis. We may change the basis of holomorphic differentials by any invertible $p \times p$ matrix: $\omega' = C\omega$ multiplies Ω on the left by C. Write $\Omega = (A \;\; B)$ with $A_{jk} = \int_{a_k} \omega_j$ the a-periods and $B_{jk} = \int_{b_k} \omega_j$ the b-periods. We show below that A is invertible. Then $C = A^{-1}$ gives a new basis with

$$\int_{a_k} \omega_j = \delta_{jk}, \tag{5.32}$$

Nash's (5.18), and the period matrix takes the form (5.19)

$$\Omega = \begin{pmatrix} I_{p\times p} & \tilde{\Omega}_{p\times p} \end{pmatrix}, \qquad \tilde{\Omega}_{jk} = \int_{b_k} \omega_j. \tag{5.33}$$

The $p \times p$ matrix $\tilde{\Omega}$ of b-periods of the normalised basis is also called the period matrix (Riemann matrix). For the torus, $\tilde{\Omega} = \tau$.

5.2.18 Riemann's bilinear relations

Nash states that $\tilde{\Omega}$ "is symmetric and possesses a positive imaginary part". These are **Riemann's bilinear relations**, and they come from one identity, which is Stokes' theorem on the cut-open polygon.

Key idea: the reciprocity identity

For any two closed 1-forms θ, η on Σ_p,

$$\int_{\Sigma_p} \theta \wedge \eta = \sum_{i=1}^{p} \Big(\int_{a_i} \theta \int_{b_i} \eta - \int_{b_i} \theta \int_{a_i} \eta \Big). \tag{5.34}$$

Why, for the torus. Take $T = \mathbb{C}/(\mathbb{Z} + \tau\mathbb{Z})$ and the fundamental parallelogram Π with corners $0, 1, 1+\tau, \tau$. Pull θ back to $\mathbb{C}$. There it is exact (on a simply connected domain closed forms are exact), $\theta = \mathrm{d}F$, and F is *not* periodic: going along a or b it increases by the period,

$$F(z+1) = F(z) + A(\theta), \qquad F(z+\tau) = F(z) + B(\theta),$$

where $A(\theta) = \int_a \theta$, $B(\theta) = \int_b \theta$. Then

$$\begin{aligned} \int_T \theta \wedge \eta &\overset{(1)}{=} \int_\Pi \mathrm{d}(F\eta) \overset{(2)}{=} \oint_{\partial\Pi} F\eta \\ &\overset{(3)}{=} \int_0^1 \big[F(z) - F(z+\tau)\big]\eta + \int_0^\tau \big[F(z+1) - F(z)\big]\eta \\ &\overset{(4)}{=} -B(\theta)\int_0^1 \eta + A(\theta)\int_0^\tau \eta \overset{(5)}{=} A(\theta)B(\eta) - B(\theta)A(\eta). \end{aligned}$$

(1) $\mathrm{d}(F\eta) = \mathrm{d}F \wedge \eta + F\,\mathrm{d}\eta = \theta \wedge \eta$ since η is closed.
(2) Stokes on the parallelogram.
(3) Pair the bottom edge with the top edge (traversed backwards, shifted by τ) and the right edge with the left edge (backwards, shifted by 1), exactly as in Example 5.29; η is periodic so it takes the same values on paired edges.
(4) The jumps of F are the constant periods, which come out of the integrals.
(5) $\int_0^1 \eta = A(\eta)$ and $\int_0^\tau \eta = B(\eta)$.

Check with $\theta = \mathrm{d}x$, $\eta = \mathrm{d}y$: the left side is the area $\operatorname{Im}\tau$; on the right, $A(\mathrm{d}x) = 1$, $B(\mathrm{d}x) = \operatorname{Re}\tau$, $A(\mathrm{d}y) = 0$, $B(\mathrm{d}y) = \operatorname{Im}\tau$, giving $1 \cdot \operatorname{Im}\tau - \operatorname{Re}\tau \cdot 0 = \operatorname{Im}\tau$. For genus p the same computation is done on the $4p$-gon: each group of four edges $a_i b_i a_i^{-1} b_i^{-1}$ contributes one term of the sum, because across the pair a_i, a_i^{-1} the function F jumps by $\pm\int_{b_i}\theta$ and across b_i, b_i^{-1} by $\pm\int_{a_i}\theta$.

Symmetry. Apply (5.34) to two normalised holomorphic differentials $\theta = \omega_j$, $\eta = \omega_k$:

$$0 \overset{(1)}{=} \int_{\Sigma_p} \omega_j \wedge \omega_k \overset{(2)}{=} \sum_i \bigl(\delta_{ij}\tilde\Omega_{ki} - \tilde\Omega_{ji}\delta_{ik}\bigr) = \tilde\Omega_{kj} - \tilde\Omega_{jk}.$$

(1) Both forms are multiples of $\mathrm{d}z$ locally, and $\mathrm{d}z \wedge \mathrm{d}z = 0$.
(2) (5.34) with the normalisation $\int_{a_i}\omega_j = \delta_{ij}$ and $\int_{b_i}\omega_j = \tilde\Omega_{ji}$.

So $\tilde\Omega^T = \tilde\Omega$.

Positivity. Take $\omega = \sum_j c_j\omega_j$ with $c \in \mathbb{C}^p$ nonzero, and apply (5.34) with $\theta = \omega$, $\eta = \bar\omega$. The periods are $\int_{a_i}\omega = c_i$ and $\int_{b_i}\omega = \sum_j c_j\tilde\Omega_{ji} = (\tilde\Omega c)_i$ (using symmetry), and those of $\bar\omega$ are their conjugates. So

$$0 \overset{(1)}{<} \mathrm{i}\int_{\Sigma_p} \omega\wedge\bar\omega \overset{(2)}{=} \mathrm{i}\sum_i \Bigl(c_i\,\overline{(\tilde\Omega c)_i} - (\tilde\Omega c)_i\,\bar c_i\Bigr) \overset{(3)}{=} \mathrm{i}\bigl(\overline{c^\dagger\tilde\Omega c} - c^\dagger\tilde\Omega c\bigr)$$

$$\overset{(4)}{=} \mathrm{i}\bigl(-2\mathrm{i}\operatorname{Im}(c^\dagger\tilde\Omega c)\bigr) = 2\operatorname{Im}(c^\dagger\tilde\Omega c) \overset{(5)}{=} 2\,c^\dagger(\operatorname{Im}\tilde\Omega)\,c.$$

(1) The positivity (5.29), since $\omega \neq 0$.
(2) (5.34).
(3) $\sum_i \bar c_i(\tilde\Omega c)_i = c^\dagger\tilde\Omega c$, and the other sum is its complex conjugate.
(4) $\bar z - z = -2\mathrm{i}\operatorname{Im} z$.
(5) Write $\tilde\Omega = X + \mathrm{i}Y$ with X, Y real symmetric; then $c^\dagger X c$ and $c^\dagger Y c$ are both real, so $\operatorname{Im}(c^\dagger\tilde\Omega c) = c^\dagger Y c$.

So $\operatorname{Im}\tilde\Omega$ is a positive-definite matrix. For $p = 1$ this is just $\operatorname{Im}\tau > 0$.

The a-periods are invertible. We owe this. Before normalising, suppose $\omega = \sum c_j\omega_j$ has all a-periods zero. Then every term on the right of (5.34) with $\theta = \omega$, $\eta = \bar\omega$ contains a factor $\int_{a_i}\omega$ or its conjugate, so $\mathrm{i}\int\omega\wedge\bar\omega = 0$, and (5.29) forces $\omega = 0$. So $Ac = 0 \Rightarrow c = 0$: A is invertible.

Key idea: Torelli's theorem

Two compact Riemann surfaces whose period matrices $\tilde\Omega$ are equal (for suitable choices of symplectic homology bases) are isomorphic. The complex structure is completely encoded in the $p \times p$ symmetric matrix $\tilde\Omega$ with positive imaginary part.

Watch out: the period matrix depends on the homology basis

Nash's phrase "two Riemann surfaces with the same homology basis are isomorphic if their period matrices are equal" hides a subtlety. A different symplectic basis, related by $\begin{pmatrix} D & C \\ B & A \end{pmatrix} \in \mathrm{Sp}(2p,\mathbb{Z})$ (acting on (b,a)), changes the normalised period matrix by $\tilde\Omega \mapsto (A\tilde\Omega + B)(C\tilde\Omega + D)^{-1}$, a generalised Möbius transformation. For $p = 1$ this is $\tau \mapsto (a\tau+b)/(c\tau+d)$ with $ad - bc = 1$: the action of the modular group that part 5b studies

in detail. So the period matrix determines the surface, but the surface determines the period matrix only up to $\mathrm{Sp}(2p,\mathbb{Z})$. Also, not every symmetric $\tilde{\Omega}$ with positive imaginary part comes from a Riemann surface: such matrices form a space of dimension $p(p+1)/2$, which exceeds $3p-3$ for $p \geq 4$ (e.g. $10 > 9$ at $p = 4$). Characterising the ones that do is the *Schottky problem*.

5.2.19 The Abel–Jacobi map and Abel's theorem

Why build the torus $J(\Sigma_p)$? Because it is where degree-zero divisors live. Fix a base point $P_0 \in \Sigma_p$ and a normalised basis ω_j.

Working definition 5.28 (Abel–Jacobi map)

$$u : \Sigma_p \to J(\Sigma_p), \qquad u(P) = \left(\int_{P_0}^{P} \omega_1, \ \dots, \ \int_{P_0}^{P} \omega_p \right) \bmod \Lambda.$$

[mod Λ: two vectors of $\mathbb{C}^p$ are identified when they differ by a lattice vector] The integral depends on the path from P_0 to P, but two paths differ by a closed loop, which is an integer combination of the a_i, b_i, so the vector of integrals changes by an element of Λ: modulo Λ it is well defined. Extend to divisors by linearity: $u\left(\sum n_i P_i\right) = \sum n_i u(P_i)$. On divisors of degree zero the base point drops out. [changing P_0 shifts every $u(P_i)$ by the same vector, and these shifts cancel when $\sum n_i = 0$]

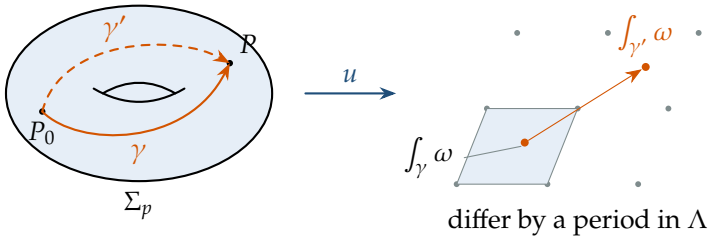


Figure 5.25: The Abel–Jacobi map. Integrating the holomorphic differentials from P_0 to P along two different paths γ, γ' gives vectors in $\mathbb{C}^p$ that differ by the integral over the closed loop $\gamma'\gamma^{-1}$, which is a lattice vector. In the torus $\mathbb{C}^p/\Lambda$ the two answers coincide.

Key idea: Abel's theorem

A divisor D of degree zero is principal, $D = \mathrm{div}(f)$, if and only if $u(D) = 0$ in $J(\Sigma_p)$. Combined with the Jacobi inversion theorem (u is onto on effective divisors of degree p), this identifies the degree-zero divisor classes with the Jacobian:

$$\frac{\{D : \deg D = 0\}}{\{\text{principal}\}} \ \cong \ J(\Sigma_p) = \mathbb{C}^p/\Lambda.$$

We prove the "only if" direction for the torus, where it is a contour integral just like the degree-zero theorem, one power of z higher.

Worked example 5.40: Abel's theorem on the torus

Let f be elliptic on $\mathbb{C}/\Lambda$, $\Lambda = \mathbb{Z}\omega_1 + \mathbb{Z}\omega_2$, with zeros $p_1, \dots, p_n$ and poles $q_1, \dots, q_n$ in a period parallelogram Π (listed with multiplicity; there are equally many by Section 5.2.8).

By the residue theorem applied to $z\,f'(z)/f(z)$, whose residue at a zero or pole c of order m is $m\,c$ (from (5.14)),

$$\frac{1}{2\pi\mathrm{i}}\oint_{\partial\Pi} z\,\frac{f'(z)}{f(z)}\,\mathrm{d}z = \sum_i p_i - \sum_i q_i.$$

Now pair the edges as in Example 5.29. With $g = f'/f$ periodic, the bottom and top edges give

$$\int_{z_0}^{z_0+\omega_1} \big[z\,g(z) - (z+\omega_2)\,g(z+\omega_2)\big]\,\mathrm{d}z \overset{(1)}{=} -\omega_2 \int_{z_0}^{z_0+\omega_1} \frac{f'}{f}\,\mathrm{d}z \overset{(2)}{=} -\omega_2 \cdot 2\pi\mathrm{i}\,k,$$

(1) $g(z+\omega_2) = g(z)$, so the $z\,g$ terms cancel.

(2) $\int f'/f\,\mathrm{d}z$ is the change of $\log f$ along the edge; since $f(z_0+\omega_1) = f(z_0)$ the logarithm returns to its starting value up to $2\pi i k$ for an integer k (the winding number of f along the edge).

and likewise the left and right edges give $+\omega_1 \cdot 2\pi\mathrm{i}\,l$ for an integer l. Dividing by $2\pi\mathrm{i}$:

$$\sum_i p_i - \sum_i q_i = l\,\omega_1 - k\,\omega_2 \in \Lambda.$$

Since $u(P) = z(P) \bmod \Lambda$ on the torus (with $P_0 = 0$ and $\omega = \mathrm{d}z$), this is $u(\operatorname{div} f) = 0$. Consequences:

- $P - O \sim P' - O$ means $P - P'$ is principal, which forces $z(P) - z(P') \in \Lambda$, i.e. $P = P'$. The degree-zero classes on T are in one-to-one correspondence with the points of T, as promised in Example 5.31: $\mathrm{Pic}_0(T) \cong T = J(T)$.
- Check on the cubic: $\operatorname{div}(y) = A_0 + A_1 + A_\lambda - 3O$ with the three branch points at the half-periods $\frac{\omega_1}{2}$, $\frac{\omega_2}{2}$, $\frac{\omega_1+\omega_2}{2}$. Their sum is $\omega_1 + \omega_2 \in \Lambda$. ✓
- A line $\alpha x + \beta y + \gamma = 0$ meets the cubic in three points P, Q, R (Bézout), and the function $\alpha x + \beta y + \gamma$ has divisor $P + Q + R - 3O$. So $z(P) + z(Q) + z(R) \in \Lambda$: *three collinear points on a cubic add up to zero in the group law of the torus*. This is the classical chord construction of addition on an elliptic curve.

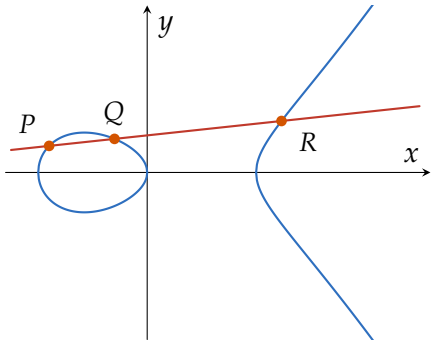


Figure 5.26: Real points of the cubic $y^2 = x(x-1)(x+1)$ (Nash's (5.3) with $\lambda = -1$) and a line meeting it in three points P, Q, R. The line's equation, viewed as a function on the curve, has divisor $P + Q + R - 3O$ (O the point at infinity), so by Abel's theorem $u(P) + u(Q) + u(R) = 0$ in the Jacobian: collinear points sum to zero in the group law of the torus.

Physical picture: periods and Jacobians in physics

Three places where these objects appear later. (1) The Jacobian is the space of *flat* U(1) *connections* on Σ_p (Wilson lines around the $2p$ loops, each a phase): $\mathrm{Hom}(\mathbb{Z}^{2p}, \mathrm{U}(1)) = (S^1)^{2p}$, a real torus of dimension $2p$ which the complex structure turns into $\mathbb{C}^p/\Lambda$ (Section 5.3.5). (2) The one-loop string amplitude integrates over the modulus $\tau = \tilde{\Omega}$ of the worldsheet torus; at p loops the period matrix enters the partition function of a chiral boson through the Riemann theta function $\sum_{n\in\mathbb{Z}^p} \exp(\pi i\, n^T \tilde{\Omega} n + \dots)$, which converges precisely because $\mathrm{Im}\,\tilde{\Omega} > 0$. (3) In Seiberg–Witten theory the matrix of low-energy gauge couplings $\tau_{ij} = \theta_{ij}/2\pi + 4\pi i/g_{ij}^2$ is the period matrix of an auxiliary curve, and positivity of $\mathrm{Im}\,\tilde{\Omega}$ is positivity of the kinetic energy.

Reading Nash (pp. 143–144, (5.15)–(5.19))

Nash recalls $H_1(\Sigma_p;\mathbb{Z}) = \mathbb{Z}^{2p}$ with the basis a_i, b_i and the intersection numbers (5.15); introduces p holomorphic 1-forms (footnote: Abelian differentials, or differentials of the first kind) and their conjugates as a basis of H^1; defines the $p \times 2p$ period matrix (5.16), whose columns are $\mathbb{R}$-independent, the lattice Λ and the Jacobian $J(\Sigma_p) = \mathbb{C}^p/\Lambda$ (5.17); normalises the a-periods (5.18) to get $\Omega = (I, \tilde{\Omega})$ (5.19); and states without proof that $\tilde{\Omega}$ is symmetric with positive imaginary part, and that equal period matrices mean isomorphic surfaces. We have added: the intersection form as a symplectic matrix, the torus computation, the closedness and positivity of holomorphic forms, the explicit basis $x^k \mathrm{d}x/y$ on hyperelliptic curves, the proof of $\mathbb{R}$-independence, the elliptic-integral periods of the cubic, the reciprocity identity with full derivation of symmetry, positivity and invertibility of the a-periods, the $\mathrm{Sp}(2p,\mathbb{Z})$ ambiguity and the Schottky dimension count, and the Abel–Jacobi map with Abel's theorem proved on the torus.

Recap: Section 5.2

- A compact Riemann surface Σ_p is a closed orientable surface of genus p with holomorphic transition functions; complex structure = conformal structure. Genus 0: $\mathbb{CP}^1$. Genus 1: the cubic $y^2 = x(x-1)(x-\lambda) = \mathbb{C}/\Lambda$ (via $\wp$). Genus $p \geq 2$: embeds in $\mathbb{CP}^3$; $3p-3$ complex moduli.
- Line bundles are given by transition functions; $\mathcal{O}(n)$ on $\mathbb{CP}^1$ has sections = polynomials of degree $\leq n$; $K = T^*\Sigma_p$ is the canonical bundle, $K_{\mathbb{CP}^1} = \mathcal{O}(-2)$, K_T trivial.
- The divisor of a meromorphic section lists its zeros and poles with orders. A function's divisor has degree 0 (contour argument). Divisors modulo principal ones = line bundles, via $g_{\alpha\beta} = f_\alpha/f_\beta$, and $c_1([D]) = \deg D$.
- Holomorphic sections of $[D] = \mathcal{L}(D)$; none if $\deg < 0$; for degree 0 only if trivial. $\mathrm{Div} = H^0(\mathcal{M}^*/\mathcal{O}^*)$ and $D \mapsto [D]$ is a connecting homomorphism.
- $H_1 = \mathbb{Z}^{2p}$ with $(a_i, b_j) = \delta_{ij}$; p holomorphic differentials; period matrix $(I, \tilde{\Omega})$ with $\tilde{\Omega}$ symmetric, $\mathrm{Im}\,\tilde{\Omega} > 0$ (bilinear relations); Jacobian $\mathbb{C}^p/\Lambda$ = degree-zero divisor classes (Abel).

5.3 Serre duality, line bundles and Kähler manifolds

In Section 5.2 we collected a list of questions that we kept postponing. How many holomorphic sections does a line bundle of degree d have? Why is $\deg K_{\Sigma_p} = 2p-2$? Why does a surface of genus p have exactly p independent holomorphic 1-forms? Which holomorphic line bundles are trivial, and how big is the set of line bundles with a given Chern class? And, for a general complex manifold: which metrics are compatible with the complex structure, and what does such a metric tell us about the topology?

This section answers all of these with three tools.

1. *Serre duality*, a version of Poincaré duality for holomorphic objects. It pairs holomorphic data of type "(p,q)" with data of type "$(n-p,n-q)$".
2. *The exponential sheaf sequence*, which compares the classification of line bundles up to smooth isomorphism (by an integer, the Chern class) with the classification up to holomorphic isomorphism (by an integer *and* a point of a complex torus).
3. *Kähler metrics*, Hermitian metrics whose 2-form is closed. On a Kähler manifold, harmonic forms split cleanly by type, which forces strong restrictions on Betti numbers.

Along the way we derive the Riemann–Roch theorem for curves, which is the formula that answers the counting questions. Nash does not state it in this section, but it is the engine behind the numbers p, $2p-2$ and $3p-3$ that appear throughout the chapter, and it is the simplest instance of the index theorem of Chapter 4.

From now on, as in Nash, M denotes a closed (compact, no boundary) complex manifold of complex dimension n; for $n=1$ it is a Riemann surface Σ_p.

5.3.1 Forms of type (p,q) and the operators ∂, $\bar\partial$

Serre duality is a statement about the cohomology of holomorphic objects, and the most concrete way to compute that cohomology is with differential forms, exactly as de Rham cohomology computes ordinary cohomology (Chapter T2, Chapter 3). The only new ingredient is that on a complex manifold forms split into "holomorphic" and "antiholomorphic" parts.

One complex variable first. On $\mathbb{C}$ with $z = x + \mathrm{i}y$, the complex 1-forms

$$\mathrm{d}z = \mathrm{d}x + \mathrm{i}\,\mathrm{d}y, \qquad \mathrm{d}\bar z = \mathrm{d}x - \mathrm{i}\,\mathrm{d}y$$

are a basis of the complexified cotangent space, just as $\mathrm{d}x, \mathrm{d}y$ are. The dual basis of vectors is

$$\partial_z = \tfrac12(\partial_x - \mathrm{i}\partial_y), \qquad \partial_{\bar z} = \tfrac12(\partial_x + \mathrm{i}\partial_y),$$

chosen so that $\mathrm{d}z(\partial_z) = 1$, $\mathrm{d}z(\partial_{\bar z}) = 0$ and so on. Every complex 1-form is $a\,\mathrm{d}z + b\,\mathrm{d}\bar z$, and for a function

$$\begin{aligned} \mathrm{d}f &\overset{(1)}{=} f_x\,\mathrm{d}x + f_y\,\mathrm{d}y \overset{(2)}{=} f_x\,\frac{\mathrm{d}z+\mathrm{d}\bar z}{2} + f_y\,\frac{\mathrm{d}z-\mathrm{d}\bar z}{2\mathrm{i}} \\ &\overset{(3)}{=} \tfrac12(f_x - \mathrm{i}f_y)\,\mathrm{d}z + \tfrac12(f_x+\mathrm{i}f_y)\,\mathrm{d}\bar z = \underbrace{\partial_z f\,\mathrm{d}z}_{\partial f} + \underbrace{\partial_{\bar z} f\,\mathrm{d}\bar z}_{\bar\partial f}. \end{aligned} \tag{5.35}$$

(1) The ordinary differential.

(2) Invert: $\mathrm{d}x = \frac{1}{2}(\mathrm{d}z + \mathrm{d}\bar{z})$, $\mathrm{d}y = \frac{1}{2\mathrm{i}}(\mathrm{d}z - \mathrm{d}\bar{z})$.
(3) Collect terms; $1/\mathrm{i} = -\mathrm{i}$.

The condition $\bar{\partial} f = 0$ reads $f_x + \mathrm{i} f_y = 0$. With $f = u + \mathrm{i}v$ its real and imaginary parts are $u_x = v_y$, $u_y = -v_x$: the *Cauchy–Riemann equations*. So

$$\boxed{f \text{ is holomorphic} \iff \bar{\partial} f = 0.}$$

The operator $\bar{\partial}$ is the "obstruction to being holomorphic", and the holomorphic functions are its kernel. This is the sentence to keep in mind for the rest of the section: *holomorphic cohomology is the cohomology of* $\bar{\partial}$.

Working definition 5.29 (forms of type (p,q))

On a complex manifold with local holomorphic coordinates $z^1, \dots, z^n$, [coordinates whose changes of chart are holomorphic functions] a **(p,q)-form** is a form that in every chart can be written

$$\alpha = \frac{1}{p!\,q!}\, \alpha_{i_1 \dots i_p \bar{j}_1 \dots \bar{j}_q}\, \mathrm{d}z^{i_1} \wedge \cdots \wedge \mathrm{d}z^{i_p} \wedge \mathrm{d}\bar{z}^{j_1} \wedge \cdots \wedge \mathrm{d}\bar{z}^{j_q},$$

with p factors $\mathrm{d}z$ and q factors $\mathrm{d}\bar{z}$. The type does not depend on the chart, because a holomorphic change of coordinates sends $\mathrm{d}z$'s to combinations of $\mathrm{d}z$'s ($\mathrm{d}w^i = \frac{\partial w^i}{\partial z^j}\mathrm{d}z^j$, with no $\mathrm{d}\bar{z}$ term). Every complex k-form splits uniquely [sort its terms by how many $\mathrm{d}\bar{z}$ factors each one contains] as a sum of pieces of types (p,q) with $p+q=k$. We write $A^{p,q}(M)$ for the smooth (p,q)-forms. *Operationally*: count the $\mathrm{d}z$'s and the $\mathrm{d}\bar{z}$'s.

The exterior derivative of a (p,q)-form has two pieces. Differentiating a coefficient gives $\sum_k \partial_{z^k}(\cdot)\,\mathrm{d}z^k + \sum_k \partial_{\bar{z}^k}(\cdot)\,\mathrm{d}\bar{z}^k$, so

$$\mathrm{d} = \partial + \bar{\partial}, \qquad \partial : A^{p,q} \to A^{p+1,q}, \qquad \bar{\partial} : A^{p,q} \to A^{p,q+1}.$$

Expanding $0 = \mathrm{d}^2 = \partial^2 + (\partial\bar{\partial} + \bar{\partial}\partial) + \bar{\partial}^2$ and sorting by type (the three pieces have types $(p+2,q)$, $(p+1,q+1)$, $(p,q+2)$, so each must vanish separately) gives

$$\partial^2 = 0, \qquad \bar{\partial}^2 = 0, \qquad \partial\bar{\partial} = -\bar{\partial}\partial. \tag{5.36}$$

Working definition 5.30 (Dolbeault cohomology)

Since $\bar{\partial}^2 = 0$, for each fixed p we have a complex [a chain of maps in which any two applied in a row give zero] $A^{p,0} \xrightarrow{\bar{\partial}} A^{p,1} \xrightarrow{\bar{\partial}} A^{p,2} \to \cdots$, and its cohomology

$$H^{p,q}_{\bar{\partial}}(M) = \frac{\{\alpha \in A^{p,q} : \bar{\partial}\alpha = 0\}}{\{\bar{\partial}\beta : \beta \in A^{p,q-1}\}}$$

is the **Dolbeault cohomology**. *Operationally*: to show a class is zero, solve the equation $\bar{\partial}\beta = \alpha$; the cohomology measures when this $\bar{\partial}$-equation has no global solution. [no β defined on all of M. Locally a solution always exists]

The lowest piece is the one we already know: $H^{p,0}_{\bar{\partial}}(M)$ consists of $(p,0)$-forms α with $\bar{\partial}\alpha = 0$ (there is nothing to divide by), that is, the *holomorphic p-forms*. For $p=0$ these are the holomorphic functions, and $H^{0,0}_{\bar{\partial}} = \mathbb{C}$ on a connected compact M. For $p = n$ they are the holomorphic sections of K_M.

Bundle-valued forms. Let $E \to M$ be a holomorphic vector bundle with transition functions $g_{\alpha\beta}$ (holomorphic matrices). An E-valued (p,q)-form is, in each patch, a column vector α_α of (p,q)-forms, with $\alpha_\alpha = g_{\alpha\beta}\alpha_\beta$. Define $\bar\partial_E \alpha$ patchwise as $\bar\partial \alpha_\alpha$. This is consistent, because

$$\bar\partial \alpha_\alpha \overset{(1)}{=} \bar\partial(g_{\alpha\beta}\alpha_\beta) \overset{(2)}{=} (\bar\partial g_{\alpha\beta}) \wedge \alpha_\beta + g_{\alpha\beta}\,\bar\partial\alpha_\beta \overset{(3)}{=} g_{\alpha\beta}\,\bar\partial\alpha_\beta.$$

(1) The gluing rule.
(2) Leibniz rule for $\bar\partial$ (the matrix g is a 0-form).
(3) $g_{\alpha\beta}$ is holomorphic, so $\bar\partial g_{\alpha\beta} = 0$.

(The same calculation for ∂ or d fails, since $\partial g_{\alpha\beta} \neq 0$: to differentiate sections in the holomorphic directions one needs a connection. The operator $\bar\partial_E$ comes free with the holomorphic structure.) So we get the complex $(A^{p,\bullet}(E), \bar\partial_E)$ and its cohomology $H^{p,q}_{\bar\partial_E}(M)$.

Key idea: the Dolbeault isomorphism

For every holomorphic vector bundle E,

$$H^{p,q}_{\bar\partial_E}(M) \;\cong\; H^q\big(M;\Omega^p(E)\big), \tag{5.37}$$

where $\Omega^p(E)$ is the sheaf of holomorphic E-valued p-forms (sections of $\Lambda^p T^*M \otimes E$ over each open set) and the right side is sheaf cohomology (Chapter 3). This is the form in which Nash quotes it on p. 145. In particular $H^q(M;\mathcal{O}(E)) = H^{0,q}_{\bar\partial_E}(M)$.

Why it is true. It is the same mechanism as the de Rham theorem. There are two ingredients.

- *A local $\bar\partial$-Poincaré lemma* (the Dolbeault–Grothendieck lemma): on a small disc, every $\bar\partial$-closed (p,q)-form with $q \geq 1$ is $\bar\partial$ of something. We verify it below in one variable, where it is the statement that the Green's function of $\partial_{\bar z}$ exists.
- *The sheaves $\mathcal{A}^{p,q}$ of smooth forms are fine*: they admit partitions of unity, so their higher cohomology vanishes (Chapter 3, and Section 5.3.4 below for an explicit computation).

Together they say that $0 \to \Omega^p(E) \to \mathcal{A}^{p,0}(E) \xrightarrow{\bar\partial} \mathcal{A}^{p,1}(E) \xrightarrow{\bar\partial} \cdots$ is a fine resolution of $\Omega^p(E)$, and the cohomology of a sheaf can be computed as the cohomology of global sections of any fine resolution. That is the whole proof, at the level of Chapter 3.

Worked example 5.41: The $\bar\partial$-Poincaré lemma in one variable: the Green's function of $\partial_{\bar z}$

We want to solve $\partial_{\bar z} u = f$ on a disc, for a given smooth f (this is $\bar\partial u = f\,\mathrm{d}\bar z$). The answer is

$$u(z) = \frac{1}{\pi}\int \frac{f(w)}{z-w}\,\mathrm{d}^2 w, \qquad \mathrm{d}^2 w = \mathrm{d}(\mathrm{Re}\,w)\,\mathrm{d}(\mathrm{Im}\,w), \tag{5.38}$$

and the key identity is $\partial_{\bar z}(1/z) = \pi\,\delta^2(z)$. To see it, integrate $\partial_{\bar z}(1/z)$ against the area element over a disc D_ε of radius ε (where $1/z$ is singular, the integral is defined as a limit, i.e. distributionally):

$$\int_{D_\varepsilon} \partial_{\bar z} g\;\mathrm{d}^2 z \overset{(1)}{=} \frac{\mathrm{i}}{2}\int_{D_\varepsilon} \partial_{\bar z} g\;\mathrm{d}z\wedge\mathrm{d}\bar z \overset{(2)}{=} -\frac{\mathrm{i}}{2}\int_{D_\varepsilon} \mathrm{d}(g\,\mathrm{d}z) \overset{(3)}{=} -\frac{\mathrm{i}}{2}\oint_{\partial D_\varepsilon} g\,\mathrm{d}z \overset{(4)}{=} -\frac{\mathrm{i}}{2}\cdot 2\pi\mathrm{i} = \pi.$$

(1) $\mathrm{d}z\wedge\mathrm{d}\bar z = (\mathrm{d}x + \mathrm{i}\,\mathrm{d}y)\wedge(\mathrm{d}x - \mathrm{i}\,\mathrm{d}y) = -2\mathrm{i}\,\mathrm{d}x\wedge\mathrm{d}y$, so $\mathrm{d}^2 z = \frac{\mathrm{i}}{2}\mathrm{d}z\wedge\mathrm{d}\bar z$.
(2) $\mathrm{d}(g\,\mathrm{d}z) = \partial_{\bar z} g\,\mathrm{d}\bar z\wedge\mathrm{d}z = -\partial_{\bar z} g\,\mathrm{d}z\wedge\mathrm{d}\bar z$.

(3) Stokes' theorem.
(4) $g = 1/z$ and $\oint \mathrm{d}z/z = 2\pi\mathrm{i}$.

Away from 0, $\partial_{\bar z}(1/z) = 0$ because $1/z$ is holomorphic there. So $\partial_{\bar z}(1/z)$ is concentrated at the origin with total weight π: $\partial_{\bar z}(1/z) = \pi\delta^2(z)$. Then

$$\partial_{\bar z} u(z) \overset{(1)}{=} \frac{1}{\pi}\int f(w)\,\partial_{\bar z}\frac{1}{z-w}\,\mathrm{d}^2 w \overset{(2)}{=} \frac{1}{\pi}\int f(w)\,\pi\,\delta^2(z-w)\,\mathrm{d}^2 w = f(z).$$

(1) Differentiate under the integral.
(2) The identity just proved, shifted to $z - w$.

A physicist knows this already: since $\partial_z\partial_{\bar z} = \frac{1}{4}\nabla^2$ and $1/z = \partial_z \log|z|^2$, the identity is $\nabla^2 \log|z| = 2\pi\delta^2$, the statement that $\frac{1}{2\pi}\log|z|$ is the Coulomb potential of a point charge in two dimensions. So *locally* the $\bar\partial$-equation can always be solved, and all the obstructions measured by $H^{p,q}_{\bar\partial}$ are global.

Worked example 5.42: Dolbeault cohomology of the torus by Fourier series

Take the square torus $T = \mathbb{C}/(\mathbb{Z} + \mathrm{i}\mathbb{Z})$ with $z = x + \mathrm{i}y$ and $x, y \in \mathbb{R}/\mathbb{Z}$. Which $(0,1)$-forms $f\,\mathrm{d}\bar z$ are $\bar\partial$-exact? (Every $(0,1)$-form is $\bar\partial$-closed, since there are no $(0,2)$-forms in one complex dimension.) Expand in Fourier modes, $f = \sum_{m,n\in\mathbb{Z}} f_{mn}\,e_{mn}$ with $e_{mn} = e^{2\pi\mathrm{i}(mx+ny)}$. The operator $\partial_{\bar z}$ is diagonal on these modes:

$$\partial_{\bar z} e_{mn} \overset{(1)}{=} \tfrac{1}{2}(\partial_x + \mathrm{i}\partial_y)\,e_{mn} \overset{(2)}{=} \tfrac{1}{2}(2\pi\mathrm{i}m + \mathrm{i}\cdot 2\pi\mathrm{i}n)\,e_{mn} = \pi\mathrm{i}\,(m + \mathrm{i}n)\,e_{mn}.$$

(1) Definition of $\partial_{\bar z}$.
(2) $\partial_x e_{mn} = 2\pi\mathrm{i}m\,e_{mn}$, $\partial_y e_{mn} = 2\pi\mathrm{i}n\,e_{mn}$.

The eigenvalue $\pi\mathrm{i}(m + \mathrm{i}n)$ vanishes only for $(m,n) = (0,0)$. So $\partial_{\bar z} u = f$ is solved by $u_{mn} = f_{mn}/\big(\pi\mathrm{i}(m + \mathrm{i}n)\big)$ for every mode except the constant one, and the constant mode can never be hit ($\partial_{\bar z}$ of anything has zero average, by Stokes). Therefore

$$H^{0,1}_{\bar\partial}(T) = \mathbb{C}\cdot[\mathrm{d}\bar z], \qquad H^{0,0}_{\bar\partial}(T) = \ker\partial_{\bar z} = \mathbb{C}\cdot 1.$$

Both are one-dimensional: $h^{0,0} = h^{0,1} = 1$. The same Fourier argument applied to $(1,0)$- and $(1,1)$-forms (multiply everything by $\mathrm{d}z$) gives $h^{1,0} = h^{1,1} = 1$, with representatives $\mathrm{d}z$ and $\mathrm{d}z\wedge\mathrm{d}\bar z$. Notice that $h^{1,0} + h^{0,1} = 2 = b_1(T)$ and $h^{0,0} = h^{1,1} = 1 = b_0 = b_2$: a first glimpse of the Hodge decomposition of Section 5.3.9.

Physical picture: zero modes

Dolbeault cohomology is the language of *zero modes*. The operator $\bar\partial_E$ acting on sections of E is (up to identifying $(0,1)$-forms with spinors of one chirality) the chiral Dirac operator on a surface, coupled to the gauge field of E. Its kernel $H^{0,0}$ is the space of normalisable zero modes of one chirality; its cokernel $H^{0,1}$ is the space of zero modes of the adjoint, of the other chirality. The index $h^0 - h^1$ is what the index theorem computes (Chapter 4); on a surface, that computation is the Riemann–Roch theorem of Section 5.3.3.

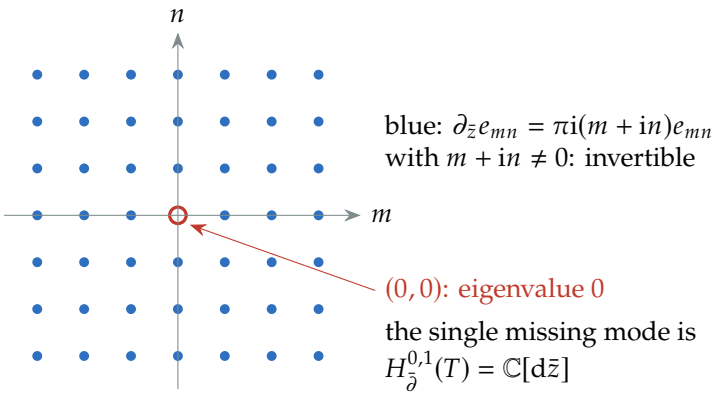


Figure 5.27: The spectrum of $\partial_{\bar z}$ on the square torus. Each Fourier mode e_{mn} is an eigenfunction with eigenvalue $\pi i(m + in)$, the lattice point $m + in$ itself up to a factor. Only the constant mode (open red circle) has eigenvalue zero: it is both the kernel (holomorphic functions, $H^{0,0}$) and the cokernel (the unreachable part of $f\,d\bar z$, $H^{0,1}$). The cohomology of $\bar\partial$ is the zero-mode content of an elliptic operator.

5.3.2 Serre duality

Reminder: Poincaré duality. On a closed oriented real manifold of dimension m, the pairing

$$H^k_{\mathrm{dR}}(M) \times H^{m-k}_{\mathrm{dR}}(M) \to \mathbb{R}, \qquad ([\alpha],[\beta]) \mapsto \int_M \alpha \wedge \beta$$

is well defined (Stokes: $\int d\gamma \wedge \beta = \int d(\gamma \wedge \beta) = 0$ when $d\beta = 0$) and nondegenerate, so $H^k_{\mathrm{dR}} \cong (H^{m-k}_{\mathrm{dR}})^*$ and $b_k = b_{m-k}$. Nondegeneracy has a one-line reason once we have a metric: the harmonic representative α of a nonzero class pairs with $\star\alpha$ (also harmonic) to give $\int \alpha \wedge \star\alpha = \|\alpha\|^2 > 0$.

Serre duality (Nash calls it Kodaira–Serre duality) is the same statement with d replaced by $\bar\partial$. The top-degree forms on a complex n-fold are the (n,n)-forms, so a (p,q)-form must be paired with an $(n-p, n-q)$-form, and a form with values in E must be paired with one with values in the dual bundle E^* so that the values can be contracted to a number.

Key idea: Serre duality

For a holomorphic vector bundle E on a closed complex manifold M of dimension n, the pairing

$$H^{p,q}_{\bar\partial_E}(M) \times H^{n-p,n-q}_{\bar\partial_{E^*}}(M) \to \mathbb{C}, \qquad ([\theta],[\nu]) \mapsto \int_M \langle \theta \wedge \nu \rangle \tag{5.39}$$

is well defined and nondegenerate. Here $\langle \theta \wedge \nu \rangle$ means: wedge the form parts and contract the E values of θ with the E^* values of ν, producing a scalar (n,n)-form. Hence, in sheaf language,

$$H^q\big(M; \Omega^p(E)\big) \;\cong\; \big\{H^{n-q}\big(M; \Omega^{n-p}(E^*)\big)\big\}^*. \tag{5.40}$$

Reading Nash (p. 144, (5.20))

Nash writes (5.20) as $H^p(M, \Omega^q(E)) \simeq \{H^{n-p}(M, \Omega^{n-q}(E^*))\}^*$. His p is the cohomology degree and his q the form degree, the other way round from the Dolbeault labelling $H^{p,q} = H^q(\Omega^p)$ that he uses on the next page and that we use in (5.40). The content is identical: both indices are reflected, (form degree, cohomology degree) $\mapsto (n-\cdot, n-\cdot)$, and E is replaced by E^*. Nash describes the pairing as $\int_M \langle \theta, \nu \rangle$ with "the Hermitian inner

product used to define the ordinary dual $E^{*\prime\prime}$: a Hermitian metric identifies E^* with the complex conjugate of E, which is how one writes $\langle \theta \wedge \nu \rangle$ as an inner product.

The special cases $p = 0$ and $p = n$. A holomorphic E-valued 0-form is just a holomorphic section of E, and a holomorphic E-valued n-form is a holomorphic section of $\Lambda^n T^*M \otimes E = K_M \otimes E$. So (Nash's (5.21))

$$\Omega^0(E) = \mathcal{O}(E), \qquad \Omega^n(E) = \mathcal{O}(K_M \otimes E). \tag{5.41}$$

Putting $p = 0$ in (5.40) gives the most used form of Serre duality, Nash's (5.22):

$$\boxed{H^q\big(M; \mathcal{O}(E)\big) \cong \big\{H^{n-q}\big(M; \mathcal{O}(K_M \otimes E^*)\big)\big\}^*.} \tag{5.42}$$

The canonical bundle appears because it is what turns $(0,q)$-forms into (n,q)-forms. On a Riemann surface ($n = 1$) there are only two groups, and (5.42) says

$$h^1(L) = h^0(K \otimes L^{-1}), \qquad h^0(L) = h^1(K \otimes L^{-1}), \tag{5.43}$$

where $h^q(L) = \dim H^q(\Sigma; \mathcal{O}(L))$ and $L^{-1} = L^*$ for a line bundle. In divisor notation, $h^1(D) = h^0(K - D)$.

Why the pairing is well defined. Take $\theta \in A^{p,q}(E)$ and $\nu \in A^{n-p,n-q}(E^*)$, both $\bar\partial$-closed, and change θ by an exact piece $\bar\partial\alpha$, $\alpha \in A^{p,q-1}(E)$. Then

$$\int_M \langle \bar\partial\alpha \wedge \nu \rangle \overset{(1)}{=} \int_M \bar\partial\langle \alpha \wedge \nu \rangle - (-1)^{p+q-1} \int_M \langle \alpha \wedge \bar\partial\nu \rangle \overset{(2)}{=} \int_M \bar\partial\langle \alpha \wedge \nu \rangle$$
$$\overset{(3)}{=} \int_M \mathrm{d}\langle \alpha \wedge \nu \rangle \overset{(4)}{=} 0.$$

(1) Leibniz rule. It is legitimate to apply $\bar\partial$ to the contracted scalar form: in a local holomorphic frame of E and the dual frame of E^* the contraction is a sum of products of components, and the frames change by holomorphic matrices that $\bar\partial$ does not see.
(2) $\bar\partial_{E^*}\nu = 0$.
(3) $\langle \alpha \wedge \nu \rangle$ has type $(n, n-1)$. Its ∂ would have type $(n+1, n-1)$, which does not exist, so $\mathrm{d} = \partial + \bar\partial$ equals $\bar\partial$ on it.
(4) Stokes' theorem on the closed manifold M.

The same argument works for changing ν. This is word for word the Poincaré duality argument, with the observation in step (3) doing the work of converting $\bar\partial$ into d.

Why it is nondegenerate. Give M a Hermitian metric and E a Hermitian fibre metric. Then there is a conjugate-linear Hodge star $\bar\star_E : A^{p,q}(E) \to A^{n-p,n-q}(E^*)$ with $\langle \theta \wedge \bar\star_E \theta \rangle = |\theta|^2 \,\mathrm{vol}$ (take the Hodge star, complex-conjugate the form part, and use the metric to turn E-values into E^*-values). Hodge theory for the elliptic operator $\Delta_{\bar\partial} = \bar\partial\bar\partial^* + \bar\partial^*\bar\partial$ (the complex version of Chapter 2) says that every Dolbeault class has a unique harmonic representative, and $\bar\star_E$ maps harmonic forms to harmonic forms. So for a nonzero class with harmonic representative θ, $\int \langle \theta \wedge \bar\star_E \theta \rangle = \|\theta\|^2 > 0$: nothing nonzero pairs to zero with everything. In particular $\dim H^{p,q}_{\bar\partial_E} = \dim H^{n-p,n-q}_{\bar\partial_{E^*}}$.

Worked example 5.43: Serre duality on the torus

Take E trivial on $T = \mathbb{C}/\Lambda$, $n = 1$, $p = 0$, $q = 1$. Serre duality pairs $H^{0,1}_{\bar\partial}(T) = H^1(T;\mathcal{O})$ with $H^{1,0}_{\bar\partial}(T) = H^0(T;K)$. From Example 5.42 the first is spanned by $[\mathrm{d}\bar z]$ and from Example 5.26 the second by $\mathrm{d}z$. The pairing is

$$\int_T \mathrm{d}\bar z \wedge \mathrm{d}z \overset{(1)}{=} -\int_T \mathrm{d}z \wedge \mathrm{d}\bar z \overset{(2)}{=} 2\mathrm{i}\int_T \mathrm{d}x \wedge \mathrm{d}y \overset{(3)}{=} 2\mathrm{i}\,\mathrm{Area}(T) \neq 0.$$

(1) Antisymmetry of the wedge product of 1-forms.
(2) $\mathrm{d}z \wedge \mathrm{d}\bar z = -2\mathrm{i}\,\mathrm{d}x \wedge \mathrm{d}y$.
(3) The integral of the area form.

A 1×1 matrix with nonzero entry: the pairing is perfect, and $h^1(\mathcal{O}) = h^0(K) = 1$.

Worked example 5.44: $H^1(\mathbb{CP}^1;\mathcal{O}(n))$ by Laurent series, and Serre duality checked

Here is a computation of H^1 that uses no forms at all, only the Čech description of Chapter 3, and it shows beautifully what H^1 "is". Cover $\mathbb{CP}^1$ by $U_0 = \mathbb{C}$ (coordinate z) and $U_1 = \mathbb{CP}^1 \setminus \{0\}$ (coordinate $w = 1/z$). For a cover by two sets, a Čech 1-cochain of $\mathcal{O}(n)$ is one holomorphic section c_{01} over the overlap $U_0 \cap U_1 = \mathbb{C}^*$ (written in the U_0 frame), there is no cocycle condition, and a coboundary is a difference $c_{01} = s_0 - g_{01}s_1$ of sections over U_0 and U_1. So

$$H^1(\mathbb{CP}^1;\mathcal{O}(n)) = \frac{\mathcal{O}(\mathbb{C}^*)}{\mathcal{O}(U_0) + z^n\,\mathcal{O}(U_1)}.$$

Now write everything as Laurent series in z.

- $c_{01} \in \mathcal{O}(\mathbb{C}^*)$: any convergent Laurent series $\sum_{k\in\mathbb{Z}} c_k z^k$.
- $\mathcal{O}(U_0)$: power series in z, spanned by z^k with $k \geq 0$.
- $z^n\mathcal{O}(U_1)$: z^n times power series in $w = z^{-1}$, spanned by z^{n-j} with $j \geq 0$, i.e. by z^k with $k \leq n$.

The quotient is spanned by the monomials z^k that are in neither list: $k < 0$ and $k > n$. That set is empty if $n \geq -1$, and for $n \leq -2$ it is $k = n+1, \dots, -1$, which has $-n-1$ elements (Figure 5.28). Hence

$$h^1(\mathcal{O}(n)) = \begin{cases} 0, & n \geq -1, \\ -n-1, & n \leq -2, \end{cases} \qquad h^0(\mathcal{O}(n)) = \begin{cases} n+1, & n \geq 0, \\ 0, & n \leq -1. \end{cases} \tag{5.44}$$

(The second formula is Example 5.23: the monomials in *both* lists, $0 \leq k \leq n$, are the global sections.) Now check Serre duality (5.43) with $K = \mathcal{O}(-2)$ (Example 5.25):

$$h^0\big(K \otimes \mathcal{O}(n)^{-1}\big) \overset{(1)}{=} h^0\big(\mathcal{O}(-2-n)\big) \overset{(2)}{=} \begin{cases} -n-1, & -2-n \geq 0, \\ 0, & \text{otherwise}, \end{cases}$$

(1) $\mathcal{O}(a) \otimes \mathcal{O}(b) = \mathcal{O}(a+b)$ (transition functions multiply) and $\mathcal{O}(n)^{-1} = \mathcal{O}(-n)$.
(2) (5.44) with n replaced by $-2-n$: $(-2-n)+1 = -n-1$.

which is exactly $h^1(\mathcal{O}(n))$. One can even see the pairing: the class of z^k ($n < k < 0$) in $H^1(\mathcal{O}(n))$ pairs with the section $z^{-k-1}\,\mathrm{d}z$ of $K \otimes \mathcal{O}(-n)$ by taking the residue of the product at $z = 0$, $\mathrm{Res}_0(z^k \cdot z^{-k-1}\mathrm{d}z) = 1$. (The residue is the Čech form of $\int_M$.) Finally notice

$$h^0(\mathcal{O}(n)) - h^1(\mathcal{O}(n)) = n+1 \qquad \text{for every } n \in \mathbb{Z},$$

a single linear formula valid on both sides of the jump. This is the Riemann–Roch theorem for the sphere.

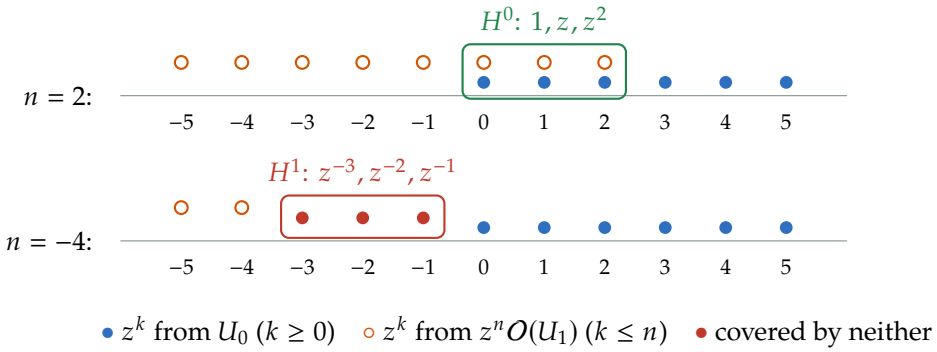


Figure 5.28: Čech cohomology of $\mathcal{O}(n)$ on $\mathbb{CP}^1$ as bookkeeping of Laurent monomials z^k on the overlap $\mathbb{C}^*$. Top ($n = 2$): the two lists overlap in $0 \leq k \leq 2$, giving $h^0 = 3$ global sections, and together they cover everything, so $h^1 = 0$. Bottom ($n = -4$): the lists no longer meet ($h^0 = 0$) and they leave a gap of $-n - 1 = 3$ monomials, which is H^1. Moving n down by one removes one orange dot from the right, and so changes $h^0 - h^1$ by exactly one.

Watch out: Serre duality needs compactness and involves K

On a non-compact manifold (say $\mathbb{C}$) there is no integral over M to pair with, and the statement fails: $H^0(\mathbb{C}; \mathcal{O})$ is infinite-dimensional while $H^1(\mathbb{C}; \mathcal{O}) = 0$. And the dual of $H^q(\mathcal{O}(E))$ is *not* $H^{n-q}(\mathcal{O}(E^*))$: the canonical bundle must be inserted. On $\mathbb{CP}^1$, $h^1(\mathcal{O}(-2)) = 1$ but $h^0(\mathcal{O}(2)) = 3$; the correct partner is $h^0(K \otimes \mathcal{O}(2)) = h^0(\mathcal{O}(0)) = 1$.

5.3.3 The Riemann–Roch theorem for curves

Here is the question that Section 5.2 kept raising. Fix a divisor D on Σ_p: a list of points where we allow poles (positive coefficients) and points where we demand zeros (negative coefficients). How many independent meromorphic functions obey those rules? That number is $h^0(D) = \dim \mathcal{L}(D) = \dim H^0(\Sigma_p; \mathcal{O}([D]))$ (Section 5.2.12). Computing it exactly is hard. Computing it *up to a correction that Serre duality identifies* is easy, and that is the Riemann–Roch theorem.

We write $h^q(D) = \dim H^q(\Sigma_p; \mathcal{O}([D]))$ for $q = 0, 1$ and call

$$\chi(D) = h^0(D) - h^1(D)$$

the **holomorphic Euler characteristic** of $[D]$. On a curve there is no H^2 (by Dolbeault, $H^2(\mathcal{O}(L)) = H^{0,2}_{\bar\partial} = 0$ since there are no $(0,2)$-forms), so χ is the full alternating sum.

Key idea: Riemann–Roch theorem for a compact Riemann surface

For every divisor D on a compact Riemann surface of genus p,

$$h^0(D) - h^1(D) = \deg D + 1 - p. \tag{5.45}$$

By Serre duality (5.43), $h^1(D) = h^0(K - D)$, so equivalently

$$\boxed{h^0(D) - h^0(K - D) = \deg D + 1 - p.} \tag{5.46}$$

The same holds for any holomorphic line bundle L, with $\deg D$ replaced by $\deg L = c_1(L)$.

The left side of (5.46) involves the complex structure (which functions exist); the right side involves only two integers, the degree and the genus. That is the pattern of every index theorem: an analytic count equals a topological number.

Why it is true: three steps. *Step 1: the base case $D = 0$.* Here $h^0(0) = 1$, the constants (Section 5.2.2), and $h^1(0) = h^0(K)$ by Serre duality, the number of independent holomorphic 1-forms. That number is p. We showed in Section 5.2.16 that there are at least "enough" of them to make $\omega, \bar{\omega}$ independent; the exact count comes from the Hodge decomposition $H^1(\Sigma_p;\mathbb{C}) = H^{1,0} \oplus H^{0,1}$ with $H^{0,1} = \overline{H^{1,0}}$ (Section 5.3.9; a Riemann surface is Kähler), which gives $2p = b_1 = 2h^{1,0}$. So

$$\chi(0) = 1 - p. \tag{5.47}$$

Genus p *means*, analytically, $h^0(K) = p$; Riemann–Roch will show it is also the topological genus appearing in $\deg K = 2p - 2$.

Step 2: adding a point raises χ by exactly one. Let P be a point and z a local coordinate centred at P, and suppose D has coefficient m at P. A section of $[D + P]$ is (Section 5.2.12) a meromorphic function f with $\operatorname{div} f + D + P \geq 0$, so near P

$$f = c\, z^{-m-1} + (\text{higher powers of } z).$$

The map "read off the coefficient c" sends $\mathcal{O}([D + P])$ to the **skyscraper sheaf** $\mathbb{C}_P$ (the sheaf whose sections over U are $\mathbb{C}$ if $P \in U$ and 0 otherwise), and its kernel is exactly $\mathcal{O}([D])$ (those f whose pole at P is one order milder). So there is a short exact sequence of sheaves

$$0 \to \mathcal{O}([D]) \to \mathcal{O}([D + P]) \xrightarrow{c} \mathbb{C}_P \to 0,$$

and hence (Chapter 3) a long exact sequence

$$0 \to H^0(D) \to H^0(D + P) \xrightarrow{c} \mathbb{C} \xrightarrow{\delta} H^1(D) \to H^1(D + P) \to H^1(\mathbb{C}_P) = 0. \tag{5.48}$$

($H^1(\mathbb{C}_P) = 0$ because a sheaf concentrated at one point is flabby; concretely, a Čech cocycle of $\mathbb{C}_P$ on a cover where only one set contains P is identically zero.) Now use the elementary fact that in an exact sequence of finite-dimensional vector spaces $0 \to V_1 \to V_2 \to \cdots \to V_k \to 0$ the alternating sum of dimensions is zero. For a short exact sequence $0 \to A \xrightarrow{i} B \xrightarrow{\pi} C \to 0$ this is

$$\dim B \overset{(1)}{=} \dim \ker \pi + \dim \operatorname{im} \pi \overset{(2)}{=} \dim \operatorname{im} i + \dim C \overset{(3)}{=} \dim A + \dim C,$$

(1) Rank–nullity for π.
(2) Exactness: $\ker \pi = \operatorname{im} i$, and π is onto.
(3) i is injective.

and a long sequence is a chain of short ones (each V_j is squeezed between the image coming in and the image going out). Applied to (5.48):

$$\begin{aligned} 0 &\overset{(1)}{=} h^0(D) - h^0(D + P) + 1 - h^1(D) + h^1(D + P) \\ &\overset{(2)}{=} \chi(D) - \chi(D + P) + 1 \quad \Longrightarrow \quad \chi(D + P) = \chi(D) + 1. \end{aligned}$$

(1) Alternating sum of dimensions of the six-term exact sequence; $\dim \mathbb{C} = 1$.
(2) Regroup using $\chi = h^0 - h^1$.

The mechanism is visible: allowing one more pole order at P either produces one new function (the map c is onto) or, if no function with that pole exists, the leftover coefficient is a nonzero obstruction class $\delta(1) \in H^1(D)$ that dies in $H^1(D+P)$. Either way the index goes up by one. We saw exactly this in Figure 5.28, where lowering n by one removed one monomial.

Step 3: induction. Every divisor is reached from 0 by adding and removing points one at a time, and each addition raises both χ and deg by one (each removal lowers both). So $\chi(D) - \deg D$ is the same for all D, and by Step 1 it equals $1 - p$. That is (5.45).

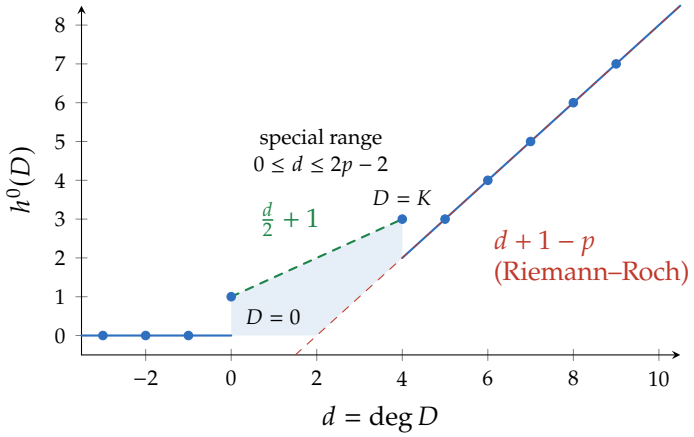


Figure 5.29: What Riemann–Roch says about $h^0(D)$ on a surface of genus $p = 3$. For $d < 0$ there are no sections. For $d > 2p - 2 = 4$ the correction $h^0(K - D)$ vanishes and $h^0 = d + 1 - p$ exactly (solid line). In between (shaded) h^0 depends on the complex structure and on D: it lies between the Riemann–Roch bound $d + 1 - p$ (red dashed) and Clifford's bound $\frac{d}{2} + 1$ (green dashed). The two ends of the special range are pinned: $h^0(0) = 1$ and $h^0(K) = p$.

Consequences. Riemann–Roch is a machine for producing numbers. Here are the ones promised earlier in the chapter.

Worked example 5.45: The degree of the canonical bundle is $2p - 2$

Take $D = K$ in (5.46):

$$h^0(K) - h^0(0) \overset{(1)}{=} \deg K + 1 - p \quad \Longrightarrow \quad p - 1 \overset{(2)}{=} \deg K + 1 - p \quad \Longrightarrow \quad \deg K = 2p - 2.$$

(1) Riemann–Roch with $D = K$, so $K - D = 0$.
(2) $h^0(K) = p$ and $h^0(0) = 1$ (Step 1).

So a holomorphic 1-form on Σ_p has exactly $2p - 2$ zeros, and a meromorphic one has (zeros) − (poles) $= 2p - 2$. Checks: $\mathbb{CP}^1$, dz has a double pole at ∞, degree -2 (Example 5.25); the torus, dz has neither, degree 0 (Example 5.26); genus 2, $y^2 = q(x)$ with $\deg q = 6$, the form dx/y has simple zeros at the two points above $x = \infty$ (Example 5.38), degree 2. Topologically, $K = T^*\Sigma$, so $\deg K = -c_1(T\Sigma) = -\chi(\Sigma) = 2p - 2$: Riemann–Roch has rederived the Gauss–Bonnet theorem (Chapter T2) from function theory.

Worked example 5.46: Large degree: the answer is exact

If $\deg D > 2p - 2$ then $\deg(K - D) < 0$, so $h^0(K - D) = 0$ and

$$h^0(D) = \deg D + 1 - p \qquad (\deg D \ge 2p - 1).$$

In general $h^0(K-D) \geq 0$ gives the **Riemann inequality** $h^0(D) \geq \deg D + 1 - p$, so *every divisor of degree* $\geq p$ *is equivalent to an effective one*: this is the threshold quoted in the pitfall of Section 5.2.12. Two applications.

- *The count* $2d+1-p$ *of Example 5.22.* A meromorphic function of degree d on a fixed Σ_p is a ratio s_0/s_1 of two sections of a line bundle L of degree d (its divisor of poles). The bundles of degree d form a p-dimensional family (Pic_d, a translate of the Jacobian, Section 5.3.5); for $d \geq 2p-1$ each has $h^0 = d+1-p$ sections; the pair (s_0, s_1) matters only up to a common scale. Total: $p + 2(d+1-p) - 1 = 2d+1-p$. ✓
- *Embeddings.* For $\deg D \geq 2p+1$ one checks with (5.46) that for any two points P, Q, $h^0(D-P-Q) = h^0(D)-2$; so sections of $[D]$ separate points (and tangents), and a basis $s_0, \ldots, s_N$ gives an embedding $P \mapsto [s_0(P) : \cdots : s_N(P)]$ into $\mathbb{CP}^N$ with $N = \deg D - p$. This is the curve case of the Kodaira embedding theorem of Section 5.1.10. For the torus with $D = 3O$: $N = 3-1 = 2$, the Weierstrass embedding by $[1 : \wp : \wp']$.

Worked example 5.47: Genus 0 and genus 1 revisited

Genus 0. For $\deg D = 0$ Riemann's inequality gives $h^0(D) \geq 1$, so D is equivalent to an effective divisor of degree 0, i.e. to 0: *every degree-zero divisor on a genus-0 surface is principal*, and holomorphic line bundles are classified by their degree. For $D = P$ a single point, $h^0(P) = 1+1-0 = 2$: there is a nonconstant function f with a single simple pole. It takes every value exactly once (degree 1), so $f : \Sigma_0 \to \mathbb{CP}^1$ is a biholomorphism. *Any Riemann surface of genus* 0 *is* $\mathbb{CP}^1$; there are no moduli, consistent with $3p-3 < 0$.
Genus 1. $h^0(nO) = n+1-1 = n$ for $n \geq 1$ (since $\deg(K-nO) = -n < 0$). This matches the explicit basis $1, \wp, \wp', \wp^2, \ldots$ found by hand in Example 5.34, and explains the gap at $n=1$: $h^0(O) = 1$, only the constants, so there is no elliptic function with a single simple pole.

Worked example 5.48: Quadratic differentials and the number $3p-3$

A **holomorphic quadratic differential** is a holomorphic section of $K^2 = K \otimes K$, locally $\phi(z)\,dz^2$. They are the infinitesimal deformations of the complex structure (their dual space, via Serre duality, is $H^1(T\Sigma)$, the space of first-order changes of the transition functions; this is taken up in part 5b). For $p \geq 2$, $\deg(K-K^2) = \deg K^{-1} = 2-2p < 0$, so

$$h^0(K^2) \overset{(1)}{=} \deg K^2 + 1 - p \overset{(2)}{=} 2(2p-2) + 1 - p = 3p-3.$$

(1) Riemann–Roch with $h^0(K-K^2) = 0$.
(2) $\deg(K \otimes K) = 2\deg K$ (degrees add under $\otimes$).

This is Nash's dimension $3p-3$ of the space of complex structures, now as a theorem. More generally, for $j \geq 2$, $h^0(K^j) = (2j-1)(p-1)$.
Physics: ghost zero modes. In the string path integral the reparametrisation ghosts b and c are sections of K^2 and K^{-1}. The number of b zero modes minus the number of c zero modes is

$$h^0(K^2) - h^0(K^{-1}) \overset{(1)}{=} \deg K^2 + 1 - p \overset{(2)}{=} 3p-3,$$

(1) Riemann–Roch with $D = K^2$, $K-D = K^{-1}$; valid for every p.
(2) As above.

for *every* genus. For $p=0$: $h^0(K^2) = h^0(\mathcal{O}(-4)) = 0$ and $h^0(K^{-1}) = h^0(\mathcal{O}(2)) = 3$, the three conformal Killing vectors $1, z, z^2$ (times ∂_z) generating $\mathrm{PSL}(2,\mathbb{C})$: $0-3 = -3$. For $p=1$:

one modulus and one translation, $1-1=0$. This is the ghost-number anomaly of string theory, and it is Riemann–Roch (Chapter 9).

Worked example 5.49: Theta functions: explicit sections on the torus

We now exhibit the sections that Riemann–Roch counts, for a line bundle of degree N on $T=\mathbb{C}/(\mathbb{Z}+\tau\mathbb{Z})$, $\operatorname{Im}\tau>0$. Describe a line bundle on T by **factors of automorphy**: a section is a holomorphic function ψ on $\mathbb{C}$ with

$$\psi(z+1)=\psi(z), \qquad \psi(z+\tau)=e_\tau(z)\,\psi(z),$$

for a fixed nowhere-zero holomorphic e_τ. (This is the same as transition functions: ψ on $\mathbb{C}$ is the pull-back of the section, and e_τ glues the two sides of the parallelogram.) Take $e_\tau(z)=e^{-\pi\mathrm{i}N\tau-2\pi\mathrm{i}Nz}$. The number of zeros of any section ψ in a parallelogram Π is

$$\begin{aligned}\frac{1}{2\pi\mathrm{i}}\oint_{\partial\Pi}\mathrm{d}\log\psi &\overset{(1)}{=}\frac{1}{2\pi\mathrm{i}}\int_0^1\big[\mathrm{d}\log\psi(x)-\mathrm{d}\log\psi(x+\tau)\big]\\ &\quad+\frac{1}{2\pi\mathrm{i}}\int_0^\tau\big[\mathrm{d}\log\psi(1+t)-\mathrm{d}\log\psi(t)\big]\\ &\overset{(2)}{=}-\frac{1}{2\pi\mathrm{i}}\int_0^1\mathrm{d}\log e_\tau(x)\overset{(3)}{=}-\frac{1}{2\pi\mathrm{i}}\int_0^1(-2\pi\mathrm{i}N)\,\mathrm{d}x=N.\end{aligned}$$

(1) The four edges of Π, anticlockwise: bottom, right, top (backwards), left (backwards), paired as in Example 5.29.
(2) $\psi(1+t)=\psi(t)$ kills the second bracket; $\log\psi(x+\tau)=\log\psi(x)+\log e_\tau(x)$ gives the first.
(3) $\mathrm{d}\log e_\tau=-2\pi\mathrm{i}N\,\mathrm{d}z$.

So this bundle has degree N, a number fixed by the factor of automorphy alone. For $N\geq 1$ Riemann–Roch predicts $h^0=N+1-1=N$ sections. Here they are:

$$\theta_a(z)=\sum_{n\in\mathbb{Z}}\exp\Big(\pi\mathrm{i}N\big(n+\tfrac{a}{N}\big)^2\tau+2\pi\mathrm{i}N\big(n+\tfrac{a}{N}\big)z\Big), \qquad a=0,1,\dots,N-1.$$

The series converges because $|\exp(\pi\mathrm{i}Nk^2\tau)|=e^{-\pi Nk^2\operatorname{Im}\tau}$ is Gaussian in k (positivity of $\operatorname{Im}\tau$ again). Check the factors: $\theta_a(z+1)=\theta_a(z)$ since $e^{2\pi\mathrm{i}(Nn+a)}=1$, and with $k=n+a/N$,

$$\begin{aligned}&\pi\mathrm{i}Nk^2\tau+2\pi\mathrm{i}Nk(z+\tau)\overset{(1)}{=}\pi\mathrm{i}N(k+1)^2\tau-\pi\mathrm{i}N\tau+2\pi\mathrm{i}N(k+1)z-2\pi\mathrm{i}Nz\\ &\quad\Longrightarrow\quad \theta_a(z+\tau)\overset{(2)}{=}e^{-\pi\mathrm{i}N\tau-2\pi\mathrm{i}Nz}\,\theta_a(z).\end{aligned}$$

(1) Complete the square: $k^2+2k=(k+1)^2-1$, and $2kz=2(k+1)z-2z$.
(2) Sum over n; the shift $k\to k+1$ is the relabelling $n\to n+1$ of the summation index, which does not change the sum.

The θ_a are independent (their Fourier modes $e^{2\pi\mathrm{i}(Nn+a)z}$ are disjoint), so they span the N-dimensional space. For $N=1$, θ_0 is the Jacobi theta function $\vartheta(z|\tau)$, whose single zero sits at $z=\frac{1+\tau}{2}$.

Physical picture: Riemann–Roch is the lowest Landau level count

A charged particle on a closed surface Σ_p with N units of magnetic flux lives in a line bundle L with $c_1(L) = N$. With a constant magnetic field, the lowest Landau level wavefunctions are exactly the holomorphic sections of L (the condition $\bar\partial_A \psi = 0$ is the LLL condition $a\psi = 0$ for the lowering operator). So the degeneracy of the lowest Landau level is $h^0(L) = N + 1 - p$ for $N > 2p - 2$. On the sphere (Haldane's geometry) it is $N + 1$; on the torus it is N, and the theta functions of Example 5.49, multiplied by the Gaussian $e^{-\pi N(\operatorname{Im} z)^2/\operatorname{Im}\tau}$ (the Hermitian metric h of Section 5.2.11), are the LLL wavefunctions. The "$1 - p$" is a topological correction to the naive count "one state per flux quantum", the shift that appears in quantum Hall physics.

Riemann–Roch as an index theorem. The number $\chi(L) = h^0(L) - h^1(L)$ is the index of the elliptic operator $\bar\partial_L : A^{0,0}(L) \to A^{0,1}(L)$ (kernel = H^0, cokernel = H^1 by the Dolbeault isomorphism). The Atiyah–Singer index theorem of Chapter 4, in its holomorphic (Hirzebruch–Riemann–Roch) form, says $\chi(L) = \int_M \operatorname{ch}(L)\,\operatorname{td}(TM)$. On a curve only the 2-form parts survive:

$$\chi(L) \overset{(1)}{=} \int_\Sigma \big(1 + c_1(L)\big)\big(1 + \tfrac{1}{2}c_1(T\Sigma)\big) \overset{(2)}{=} \int_\Sigma c_1(L) + \tfrac{1}{2}\int_\Sigma c_1(T\Sigma) \overset{(3)}{=} \deg L + \tfrac{1}{2}(2 - 2p).$$

(1) $\operatorname{ch}(L) = e^{c_1(L)} = 1 + c_1(L) + \dots$ and $\operatorname{td}(T\Sigma) = 1 + \frac{1}{2}c_1(T\Sigma) + \dots$, truncated at 2-forms.
(2) Keep the terms of degree 2.
(3) $c_1(L) = \deg L$ (Section 5.2.11) and $\int c_1(T\Sigma) = \chi(\Sigma) = 2 - 2p$ (Gauss–Bonnet).

That is (5.45). In higher dimensions the same formula, with more terms of ch and td, is the Hirzebruch–Riemann–Roch theorem; for a rank-r bundle E on a curve it reads $h^0(E) - h^1(E) = \deg E + r(1 - p)$.

Try it: Riemann–Roch on $\mathbb{CP}^1$

Check (5.46) on $\mathbb{CP}^1$ for $D = n[\infty]$ using (5.44), for $n = 3$ and for $n = -3$.
Answer. $n = 3$: $h^0 = 4$, $h^0(K - D) = h^0(\mathcal{O}(-5)) = 0$, and $\deg D + 1 - p = 4$. $n = -3$: $h^0 = 0$, $h^0(K - D) = h^0(\mathcal{O}(1)) = 2$, so the left side is $-2 = -3 + 1$. ✓

5.3.4 Line bundles as cohomology classes: the exponential sequence

Nash now turns from counting sections to classifying the bundles themselves. The question is the one a physicist asks about gauge fields on a compact space: *how many inequivalent* U(1) *bundles are there, and what labels them?* For a sphere the answer is Dirac's: an integer, the monopole charge. We now see that this is a general fact about *smooth* line bundles on any manifold, and that it fails, interestingly, for *holomorphic* ones.

Step 0: a line bundle is a Čech 1-cocycle. Recall (Section 5.2.5) that a complex line bundle L on M is given by an open cover $\{U_\alpha\}$ and transition functions $g_{\alpha\beta} : U_\alpha \cap U_\beta \to \mathbb{C}^*$ obeying the cocycle condition $g_{\alpha\beta}g_{\beta\gamma} = g_{\alpha\gamma}$, and that changing the local frames by nowhere-zero functions h_α changes $g_{\alpha\beta}$ to $h_\alpha g_{\alpha\beta} h_\beta^{-1}$. For a merely *smooth* line bundle the $g_{\alpha\beta}$ and h_α are smooth; for a holomorphic one they are holomorphic. Nash writes the cocycle condition in the equivalent form $g_{\alpha\beta}g_{\beta\gamma} = g_{\alpha\gamma} = g_{\gamma\alpha}^{-1}$ (using $g_{\gamma\alpha}g_{\alpha\gamma} = g_{\gamma\gamma} = 1$).

Now read these two sentences in the language of Chapter 3. Let $\mathcal{E}^*$ be the sheaf of smooth nowhere-vanishing complex functions, a sheaf of abelian groups under *multiplication*. Then

- a family $\{g_{\alpha\beta}\}$ with $g_{\alpha\beta} \in \mathcal{E}^*(U_\alpha \cap U_\beta)$ is a Čech 1-cochain, and the cocycle condition says its Čech coboundary $g_{\beta\gamma}\, g_{\alpha\gamma}^{-1}\, g_{\alpha\beta}$ is 1: it is a *1-cocycle* (in multiplicative notation, where "+" is "×" and "0" is "1");
- changing frames multiplies $g_{\alpha\beta}$ by $h_\alpha h_\beta^{-1}$, the Čech coboundary of the 0-cochain $\{h_\alpha\}$.

So isomorphism classes of smooth line bundles are exactly the elements of $H^1(M;\mathcal{E}^*)$, and holomorphic line bundles are the elements of $H^1(M;\mathcal{O}^*)$. The group operation of H^1, multiplying cocycles, is the tensor product of line bundles.

Working definition 5.31 (the class of a line bundle)

To find the element $[g_{\alpha\beta}] \in H^1(M;\mathcal{E}^*)$ [$\mathcal{E}^*$ is the smooth nowhere-zero complex functions] of a line bundle: write down its transition functions on a cover. Two bundles are the same element if and only if you can find frame changes h_α [a nowhere-zero function h_α on each U_α that rescales the local description of sections, $s_\alpha \mapsto h_\alpha s_\alpha$] with $g'_{\alpha\beta} = h_\alpha g_{\alpha\beta} h_\beta^{-1}$. In particular L is *trivial* (the element 1) exactly when it has a nowhere-zero global section: then $h_\alpha = s_\alpha^{-1}$ turns every $g_{\alpha\beta}$ into 1.

This description is concrete but not yet useful: $H^1(M;\mathcal{E}^*)$ is defined through arbitrary covers and arbitrary smooth functions. The trick, which is Nash's (5.23), is to compare the *multiplicative* sheaf $\mathcal{E}^*$ with the *additive* sheaf $\mathcal{E}$ of smooth complex functions by taking logarithms.

Step 1: the exponential sequence. Define

$$0 \longrightarrow \mathbb{Z} \xrightarrow{i} \mathcal{E} \xrightarrow{\exp} \mathcal{E}^* \longrightarrow 0, \qquad i(n) = n, \qquad \exp(f) = e^{2\pi i f}. \tag{5.49}$$

Here $\mathbb{Z}$ is the sheaf of locally constant integer-valued functions. This is a short exact sequence of sheaves. Check each place.

- *Injective at $\mathbb{Z}$:* a constant integer, regarded as a function, is zero only if it is zero.
- *Exact at $\mathcal{E}$:* $e^{2\pi i f} = 1$ on a connected open set means f takes values in $\mathbb{Z}$, and being continuous it is a constant integer. So $\ker \exp = \mathbb{Z}$.
- *Surjective at $\mathcal{E}^*$, as sheaves:* a nowhere-zero function g on a *small* (say contractible) open set has a smooth logarithm $f = \frac{1}{2\pi i} \log g$, so $g = \exp f$ locally. Globally it may not: $g = z$ on the annulus $\{1/2 < |z| < 2\}$ has no single-valued logarithm. This "locally but not globally" is exactly what the long exact sequence measures.

(We follow Nash in putting $2\pi i$ into the exponential; it makes the kernel $\mathbb{Z}$ rather than $2\pi i\mathbb{Z}$.)

Step 2: the long exact sequence. A short exact sequence of sheaves gives a long exact sequence in cohomology (Chapter 3). Here it is Nash's (5.24):

$$\cdots \to H^1(M;\mathcal{E}) \to H^1(M;\mathcal{E}^*) \xrightarrow{\delta} H^2(M;\mathbb{Z}) \to H^2(M;\mathcal{E}) \to \cdots \tag{5.50}$$

The connecting map δ is computed by the standard recipe, and it is worth doing by hand once, because *its output is the first Chern class*. Start from a cocycle $\{g_{\alpha\beta}\}$, on a cover fine enough that all overlaps are contractible.

1. *Lift*: choose logarithms $\lambda_{\alpha\beta} = \frac{1}{2\pi i} \log g_{\alpha\beta} \in \mathcal{E}(U_\alpha \cap U_\beta)$, with $\lambda_{\beta\alpha} = -\lambda_{\alpha\beta}$.

2. *Take the coboundary*: on triple overlaps set $n_{\alpha\beta\gamma} = \lambda_{\beta\gamma} - \lambda_{\alpha\gamma} + \lambda_{\alpha\beta}$.

3. *It lands in $\mathbb{Z}$*:

$$e^{2\pi \mathrm{i}\, n_{\alpha\beta\gamma}} \overset{(1)}{=} e^{2\pi\mathrm{i}\lambda_{\beta\gamma}}\, e^{-2\pi\mathrm{i}\lambda_{\alpha\gamma}}\, e^{2\pi\mathrm{i}\lambda_{\alpha\beta}} \overset{(2)}{=} g_{\beta\gamma}\, g_{\alpha\gamma}^{-1}\, g_{\alpha\beta} \overset{(3)}{=} 1.$$

   (1) The exponential turns sums into products.
   (2) $e^{2\pi\mathrm{i}\lambda} = g$ by the choice of logarithms.
   (3) The cocycle condition $g_{\alpha\beta}g_{\beta\gamma} = g_{\alpha\gamma}$.

   So $n_{\alpha\beta\gamma}$ is a continuous function with values in $\mathbb{Z}$ on the connected set $U_\alpha \cap U_\beta \cap U_\gamma$, hence a constant integer.

4. $\{n_{\alpha\beta\gamma}\}$ is a Čech 2-cocycle with integer coefficients (it is a coboundary of real functions, so its own coboundary vanishes), and changing the choice of logarithms by integers changes it by an integer coboundary. Its class is

$$\delta[g_{\alpha\beta}] = [n_{\alpha\beta\gamma}] = c_1(L) \in H^2(M;\mathbb{Z}).$$

In words: *the first Chern class is the integer ambiguity in the logarithms of the transition functions*. It is zero precisely when the logarithms can be chosen consistently.

Step 3: the de Rham version. To compare with the curvature formula $c_1 = \frac{\mathrm{i}}{2\pi}F$ we translate $[n_{\alpha\beta\gamma}]$ into a 2-form. Take a partition of unity $\{\rho_\gamma\}$ subordinate to the cover ($\sum_\gamma \rho_\gamma = 1$, $\operatorname{supp}\rho_\gamma \subset U_\gamma$) and the closed 1-forms

$$\theta_{\alpha\beta} = \mathrm{d}\lambda_{\alpha\beta} = \frac{1}{2\pi\mathrm{i}}\,\frac{\mathrm{d}g_{\alpha\beta}}{g_{\alpha\beta}} \qquad \text{on } U_\alpha \cap U_\beta.$$

They satisfy $\theta_{\alpha\beta} + \theta_{\beta\gamma} = \theta_{\alpha\gamma}$ (take d log of the cocycle condition) and do not care which branch of the logarithm was chosen. Define on each U_α the 1-form $a_\alpha = \sum_\gamma \rho_\gamma\, \theta_{\alpha\gamma}$. Then on overlaps

$$a_\beta - a_\alpha \overset{(1)}{=} \sum_\gamma \rho_\gamma\big(\theta_{\beta\gamma} - \theta_{\alpha\gamma}\big) \overset{(2)}{=} \sum_\gamma \rho_\gamma\big(-\theta_{\alpha\beta}\big) \overset{(3)}{=} -\theta_{\alpha\beta}. \tag{5.51}$$

(1) Definition of a_α and a_β.
(2) $\theta_{\alpha\beta} + \theta_{\beta\gamma} = \theta_{\alpha\gamma}$ rearranged.
(3) $\sum_\gamma \rho_\gamma = 1$.

Since $\theta_{\alpha\beta}$ is closed, $\mathrm{d}a_\alpha = \mathrm{d}a_\beta$ on overlaps: the local 2-forms $\mathrm{d}a_\alpha$ patch to a global closed 2-form c on M, and its de Rham class is the image of $c_1(L)$ under $H^2(M;\mathbb{Z}) \to H^2(M;\mathbb{R})$. This is exactly the structure of a U(1) gauge field: (5.51) is a gauge transformation between patches and $c = \mathrm{d}a_\alpha$ is the field strength. Indeed, if L is holomorphic with Hermitian metric h_α and Chern connection $A_\alpha = \partial \log h_\alpha$ (Section 5.2.11), then $h_\beta = |g_{\alpha\beta}|^2 h_\alpha$ gives $A_\beta - A_\alpha = \partial \log |g_{\alpha\beta}|^2 = \mathrm{d}\log g_{\alpha\beta} = 2\pi\mathrm{i}\,\theta_{\alpha\beta}$, so $a'_\alpha = \frac{\mathrm{i}}{2\pi}A_\alpha$ obeys the same rule (5.51), and $\mathrm{d}a'_\alpha = \frac{\mathrm{i}}{2\pi}F$. Both recipes give the same class; the curvature formula is just a particular choice of a_α.

Worked example 5.50: The monopole charge as a winding number

Cover $S^2 = \mathbb{CP}^1$ by $U_0 = \{|z| < 2\}$ and $U_1 = \{|z| > \frac{1}{2}\} \cup \{\infty\}$, with overlap the annulus $\{\frac{1}{2} < |z| < 2\}$, and take any smooth line bundle with transition function g_{01} on the annulus. (With two sets there are no triple overlaps, so the Čech recipe needs a finer cover; the

de Rham recipe works directly.) Let ρ_0, ρ_1 be a partition of unity depending only on $r = |z|$, with $\rho_0 = 1$ for $r \le \frac{1}{2}$ and $\rho_0 = 0$ for $r \ge 2$ (Figure 5.30). Then $a_0 = \rho_1\theta_{01}$ and $a_1 = \rho_0\theta_{10} = -\rho_0\theta_{01}$. Let C be the circle $|z| = 1$, anticlockwise:

$$\int_{S^2} c \overset{(1)}{=} \int_{|z|\le 1} \mathrm{d}a_0 + \int_{|z|\ge 1} \mathrm{d}a_1 \overset{(2)}{=} \oint_C a_0 - \oint_C a_1 \overset{(3)}{=} \oint_C (\rho_1 + \rho_0)\,\theta_{01}$$
$$\overset{(4)}{=} \frac{1}{2\pi\mathrm{i}} \oint_C \frac{\mathrm{d}g_{01}}{g_{01}} \overset{(5)}{=} \operatorname{wind}\big(g_{01}|_C\big).$$

(1) Split the sphere along C and use the local expression for c in each half.
(2) Stokes: the disc has boundary C; the outer cap (oriented as part of S^2) has boundary $-C$.
(3) $a_0 - a_1 = \rho_1\theta_{01} + \rho_0\theta_{01}$.
(4) $\rho_0 + \rho_1 = 1$ and the definition of θ_{01}.
(5) The argument principle: $\frac{1}{2\pi\mathrm{i}} \oint \mathrm{d}g/g$ counts how many times $g(C)$ winds around 0.

For $\mathcal{O}(n)$, $g_{01} = z^n$ winds n times, so $\int c = n = \deg \mathcal{O}(n)$, matching Example 5.33. Note that the answer depends only on the homotopy class of the map $g_{01}|_C : S^1 \to \mathbb{C}^*$, that is, on $\pi_1(\mathbb{C}^*) = \mathbb{Z}$. This is the **clutching** description: bundles on S^2 are classified by how the two hemispheres are glued along the equator, and the gluing maps are classified by an integer. For the Dirac monopole, $g_{01} = e^{in\phi}$ is the gauge transformation relating the potentials on the two hemispheres, and single-valuedness forces $n \in \mathbb{Z}$.

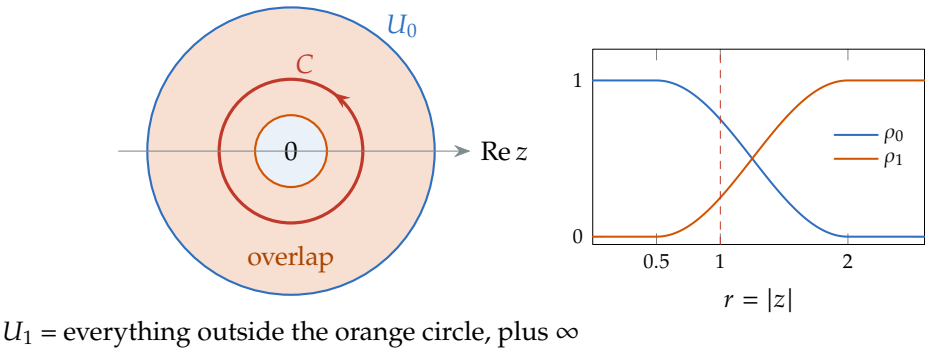


Figure 5.30: Left: the two-set cover of the Riemann sphere used in Example 5.50. U_0 is the blue disc $|z| < 2$, U_1 is the outside of the orange circle $|z| = \frac{1}{2}$ together with ∞, and the overlap is the annulus. All the information in a line bundle is the transition function g_{01} on the annulus; its winding number around the red circle C is the Chern class. Right: a partition of unity depending only on r. The 1-forms $a_0 = \rho_1\theta_{01}$ and $a_1 = -\rho_0\theta_{01}$ are the gauge potentials on the two patches; their curvature lives where ρ changes, in the annulus.

Step 4: $\mathcal{E}$ is fine, so smooth line bundles are classified by c_1. Nash now uses that $\mathcal{E}$ is a *fine* sheaf, so $H^i(M;\mathcal{E}) = 0$ for $i > 0$ (Chapter 3). Let us see why for $i = 1$, since it is a two-line computation and it explains everything that follows. Let $\{f_{\alpha\beta}\}$ be an additive Čech 1-cocycle of smooth functions: $f_{\alpha\beta} + f_{\beta\gamma} = f_{\alpha\gamma}$. Define $k_\alpha = \sum_\gamma \rho_\gamma f_{\alpha\gamma}$ on U_α (each term is extended by zero outside $\operatorname{supp}\rho_\gamma$, which is smooth). Then

$$k_\alpha - k_\beta \overset{(1)}{=} \sum_\gamma \rho_\gamma \big(f_{\alpha\gamma} - f_{\beta\gamma}\big) \overset{(2)}{=} \sum_\gamma \rho_\gamma\, f_{\alpha\beta} \overset{(3)}{=} f_{\alpha\beta}. \tag{5.52}$$

(1) Definition of k_α, k_β.

(2) The cocycle condition: $f_{\alpha\gamma} - f_{\beta\gamma} = f_{\alpha\gamma} + f_{\gamma\beta} = f_{\alpha\beta}$.
(3) $\sum_\gamma \rho_\gamma = 1$.

So every 1-cocycle is a coboundary: $H^1(M;\mathcal{E}) = 0$. (The same trick one degree up gives $H^2(M;\mathcal{E}) = 0$.) The key step is *multiplying by* ρ_γ. A partition of unity is smooth but never holomorphic (a holomorphic function with compact support in a connected open set is zero), so this argument is unavailable for $\mathcal{O}$. Remember that; it is the whole difference between the smooth and the holomorphic stories.

With $H^1(M;\mathcal{E}) = H^2(M;\mathcal{E}) = 0$ the long exact sequence (5.50) becomes

$$0 \longrightarrow H^1(M;\mathcal{E}^*) \xrightarrow{\delta} H^2(M;\mathbb{Z}) \longrightarrow 0.$$

Key idea: smooth line bundles are classified by c_1 (Nash (5.25))

$$H^1(M;\mathcal{E}^*) \;\cong\; H^2(M;\mathbb{Z}), \qquad [L] \longmapsto c_1(L). \tag{5.53}$$

Every class in $H^2(M;\mathbb{Z})$ is the Chern class of some smooth line bundle, and two line bundles are smoothly isomorphic if and only if they have the same c_1. In particular *a line bundle with $c_1(L) = 0$ is (smoothly, and topologically) trivial*.

Why $c_1 = 0$ forces triviality, concretely. If $[n_{\alpha\beta\gamma}] = 0$, adjust the logarithms by integers so that $n_{\alpha\beta\gamma} = 0$ exactly. Then $\{\lambda_{\alpha\beta}\}$ is an additive 1-cocycle of smooth functions, so by (5.52) $\lambda_{\alpha\beta} = k_\alpha - k_\beta$, i.e. $g_{\alpha\beta} = e^{2\pi i k_\alpha}\left(e^{2\pi i k_\beta}\right)^{-1}$. The frame change $h_\alpha = e^{-2\pi i k_\alpha}$ turns all transition functions into 1. On S^2 with the cover of Figure 5.30: if g_{01} has winding number 0 it has a single-valued logarithm λ on the annulus, and $k_0 = \rho_1\lambda$, $k_1 = -\rho_0\lambda$ split it, since $k_0 - k_1 = (\rho_0 + \rho_1)\lambda = \lambda$.

Worked example 5.51: Which smooth line bundles exist?

- $M = S^2$: $H^2 = \mathbb{Z}$. One bundle for each monopole charge n; they are the $\mathcal{O}(n)$ with the holomorphic structure forgotten.
- $M = T^2$ or any closed orientable surface Σ_p: $H^2 = \mathbb{Z}$ again; the bundle is determined by $\int c_1$, the total magnetic flux.
- $M = S^3$, or any M with $H^2(M;\mathbb{Z}) = 0$ (e.g. $\mathbb{R}^n$, S^n for $n \neq 2$): every complex line bundle is trivial, so a U(1) gauge field on S^3 is always a global 1-form. There is no room for a monopole on S^3.
- $M = \mathbb{CP}^n$: $H^2 = \mathbb{Z}$, generated by c_1 of the hyperplane bundle $\mathcal{O}(1)$ (Example 5.35).
- $M = S^2 \times S^2$: $H^2 = \mathbb{Z}^2$; a line bundle is labelled by its two fluxes (n_1, n_2) through the two spheres.
- $M = \mathbb{RP}^3$: $H^2(\mathbb{RP}^3;\mathbb{Z}) = \mathbb{Z}_2$. There is a nontrivial complex line bundle whose c_1 has order two. Its real curvature form is exact, $[c] = 0$ in $H^2(M;\mathbb{R})$, but the bundle is not trivial: $H^2(M;\mathbb{Z}) \to H^2(M;\mathbb{R})$ forgets torsion. Such bundles carry flat connections with nontrivial holonomy.

Watch out: higher rank is different

Nothing like (5.53) holds for vector bundles of rank $k > 1$: it does not follow that a bundle whose Chern classes all vanish is a product. The reason is that the argument

used the *abelian* logarithm; for $\mathrm{GL}(k,\mathbb{C})$-valued transition functions there is no exponential sequence. A concrete example: complex rank-2 bundles on S^5 are classified by clutching maps $S^4 \to \mathrm{U}(2)$ along the equator, that is by $\pi_4(\mathrm{U}(2)) = \pi_4(S^3) = \mathbb{Z}_2$. The nontrivial one has all Chern classes zero, since $H^2(S^5) = H^4(S^5) = 0$. (The same $\pi_4(\mathrm{SU}(2)) = \mathbb{Z}_2$ is behind Witten's global SU(2) anomaly, Chapter 10.) By contrast, rank-2 bundles on S^4 are classified by $\pi_3(\mathrm{SU}(2)) = \mathbb{Z}$, detected by c_2: the instanton number of Chapter 8.

5.3.5 Holomorphic line bundles and the Picard group

Now repeat everything with holomorphic functions. Replace $\mathcal{E}$ by the sheaf $\mathcal{O}$ of holomorphic functions and $\mathcal{E}^*$ by $\mathcal{O}^*$ of nowhere-zero holomorphic functions. The exponential is still locally surjective (a nowhere-zero holomorphic function on a disc has a holomorphic logarithm), so we get the **holomorphic exponential sequence** and its long exact sequence, Nash's (5.26):

$$0 \longrightarrow \mathbb{Z} \xrightarrow{i} \mathcal{O} \xrightarrow{\exp} \mathcal{O}^* \longrightarrow 0,$$
$$\cdots \to H^1(M;\mathbb{Z}) \to H^1(M;\mathcal{O}) \to H^1(M;\mathcal{O}^*) \xrightarrow{c_1} H^2(M;\mathbb{Z}) \to H^2(M;\mathcal{O}) \to \cdots \tag{5.54}$$

The connecting map is computed by exactly the same recipe as before, so it is again the first Chern class. What has changed is that $\mathcal{O}$ is *not* fine: the partition-of-unity trick (5.52) is gone, and indeed $H^1(M;\mathcal{O})$ can be nonzero. We already know an example: on a torus $H^1(T;\mathcal{O}) = H^{0,1}_{\bar\partial}(T) = \mathbb{C}$ (Example 5.42), and on Σ_p it is $\mathbb{C}^p$ (Serre duality: $h^1(\mathcal{O}) = h^0(K) = p$).

Working definition 5.32 (Picard group)

The **Picard group** of a complex manifold is the group of holomorphic line bundles up to holomorphic isomorphism, [two bundles count as the same when holomorphic frame changes turn one set of transition functions into the other] with tensor product as the operation (Nash's (5.27)):

$$\mathrm{Pic}(M) = H^1(M;\mathcal{O}^*). \tag{5.55}$$

Its subgroup $\mathrm{Pic}_0(M)$ consists of the holomorphic line bundles with $c_1 = 0$ [the first Chern class. On a curve it is the degree of the bundle], i.e. those that are topologically trivial [trivial as smooth bundles, though not necessarily as holomorphic ones]. *Operationally*: an element of Pic is a set of holomorphic transition functions modulo holomorphic frame changes; on a curve it is a divisor modulo principal divisors (Section 5.2.10).

Now read off the long exact sequence (5.54) piece by piece, for a compact connected M.

The start. $H^0(M;\mathcal{O}) = \mathbb{C}$ (constants) maps onto $H^0(M;\mathcal{O}^*) = \mathbb{C}^*$ by $c \mapsto e^{2\pi i c}$. Onto means the next map $H^1(M;\mathbb{Z}) \to H^1(M;\mathcal{O})$ is injective. So $H^1(M;\mathbb{Z})$ sits inside $H^1(M;\mathcal{O})$.

The middle. Exactness at $H^1(M;\mathcal{O}^*)$ says that the bundles with $c_1 = 0$ are exactly the image of $H^1(M;\mathcal{O})$. Exactness at $H^1(M;\mathcal{O})$ says the kernel of that map is the image of $H^1(M;\mathbb{Z})$. Hence (Nash's (5.28))

$$\mathrm{Pic}_0(M) = \ker c_1 = \frac{H^1(M;\mathcal{O})}{H^1(M;\mathbb{Z})}. \tag{5.56}$$

In words: a topologically trivial holomorphic line bundle is described by transition functions $g_{\alpha\beta} = e^{2\pi i f_{\alpha\beta}}$ with $\{f_{\alpha\beta}\}$ a holomorphic additive cocycle, and two such describe the same bundle if the f's differ by a holomorphic coboundary plus an integer cocycle.

The end. The image of c_1 is the kernel of $H^2(M;\mathbb{Z}) \to H^2(M;\mathcal{O})$, which (via Dolbeault) sends an integral class to the $(0,2)$ part of its de Rham representative. So the Chern classes of

holomorphic line bundles are the integral classes with no $(0,2)$ part; for real classes this means type $(1,1)$ (the *Lefschetz* $(1,1)$ *theorem*). On a curve $H^2(\Sigma;\mathcal{O}) = H^{0,2} = 0$, so every integer is the degree of some holomorphic line bundle, e.g. $[dP]$.

Key idea: smooth versus holomorphic line bundles

	smooth ($\mathcal{E}^*$)	holomorphic ($\mathcal{O}^*$)
classified by	$c_1 \in H^2(M;\mathbb{Z})$	c_1 *and* a point of Pic_0
$c_1 = 0$ means	trivial	a point of the torus $H^1(\mathcal{O})/H^1(\mathbb{Z})$
reason	$\mathcal{E}$ fine, $H^1(\mathcal{E}) = 0$	$H^1(\mathcal{O})$ may be $\neq 0$
on Σ_p	$\mathbb{Z}$	$0 \to J(\Sigma_p) \to \mathrm{Pic}(\Sigma_p) \to \mathbb{Z} \to 0$

A topologically trivial line bundle can carry a whole continuous family of inequivalent holomorphic structures. On a Riemann surface that family is the Jacobian.

$$\begin{array}{ccccccccc}
0 & \longrightarrow & \mathbb{Z} & \xrightarrow{i} & \mathcal{O} & \xrightarrow{\exp} & \mathcal{O}^* & \longrightarrow & 0 \\
 & & \| & & \downarrow & & \downarrow & & \\
0 & \longrightarrow & \mathbb{Z} & \xrightarrow{i} & \mathcal{E} & \xrightarrow{\exp} & \mathcal{E}^* & \longrightarrow & 0
\end{array}$$

$$\begin{array}{ccccccccc}
H^1(M;\mathbb{Z}) & \hookrightarrow & H^1(M;\mathcal{O}) & \longrightarrow & \mathrm{Pic}(M) & \xrightarrow{c_1} & H^2(M;\mathbb{Z}) & \longrightarrow & H^2(M;\mathcal{O}) \\
 & & \downarrow & & \downarrow{\scriptstyle \text{forget}} & & \| & & \\
0 = H^1(M;\mathcal{E}) & \longrightarrow & H^1(M;\mathcal{E}^*) & \xrightarrow[c_1]{\cong} & H^2(M;\mathbb{Z}) & \longrightarrow & 0 = H^2(M;\mathcal{E}) & &
\end{array}$$

Figure 5.31: The holomorphic exponential sequence (top row) maps into the smooth one (bottom row) by regarding a holomorphic function as a smooth one. In cohomology (lower diagram) the smooth row collapses, because $\mathcal{E}$ is fine: smooth line bundles are just their Chern classes. The holomorphic row does not collapse. The "forget" map from $\mathrm{Pic}(M)$ to smooth bundles is *not* injective; its kernel is the image of $H^1(M;\mathcal{O})$, the group $\mathrm{Pic}_0(M)$ of holomorphic structures on the trivial bundle.

Watch out: $\mathrm{Pic}(M)$ is not a subgroup of the smooth line bundles

Nash writes that the holomorphic line bundles "form a subgroup of the group of all smooth line bundles". Read literally this is not right: there is a homomorphism $\mathrm{Pic}(M) \to H^1(M;\mathcal{E}^*)$ (forget the holomorphic structure), but it is not injective, since many holomorphically different bundles are smoothly the same. What is true is that its *image* is a subgroup of $H^1(M;\mathcal{E}^*) = H^2(M;\mathbb{Z})$ (the integral classes of type $(1,1)$), and that its kernel is $\mathrm{Pic}_0(M)$. The very next sentences of Nash, and (5.28), use it in this corrected sense.

Worked example 5.52: $\mathrm{Pic}(\mathbb{CP}^1) = \mathbb{Z}$: every line bundle is some $\mathcal{O}(n)$

On $\mathbb{CP}^1$ we computed $H^1(\mathbb{CP}^1;\mathcal{O}) = 0$ in Example 5.44 (the case $n = 0$: no Laurent monomials fit between 0 and n), and $H^2(\mathbb{CP}^1;\mathcal{O}) = 0$ because there are no $(0,2)$-forms. The long exact sequence becomes

$$0 = H^1(\mathbb{CP}^1;\mathcal{O}) \longrightarrow \mathrm{Pic}(\mathbb{CP}^1) \xrightarrow{c_1} H^2(\mathbb{CP}^1;\mathbb{Z}) = \mathbb{Z} \longrightarrow H^2(\mathbb{CP}^1;\mathcal{O}) = 0,$$

so $c_1 : \mathrm{Pic}(\mathbb{CP}^1) \to \mathbb{Z}$ is an isomorphism. Every holomorphic line bundle on the sphere is

determined by its degree, and since $\mathcal{O}(n)$ has degree n (Example 5.33), every holomorphic line bundle is isomorphic to exactly one $\mathcal{O}(n)$. This is the sheaf-theoretic version of Example 5.30, where we saw that on $\mathbb{CP}^1$ "the degree is everything". The same argument on $\mathbb{CP}^n$ (where $h^{0,1} = h^{0,2} = 0$, Section 5.3.9) gives $\mathrm{Pic}(\mathbb{CP}^n) = \mathbb{Z}$, generated by $\mathcal{O}(1)$.

Worked example 5.53: Holomorphic structures on the trivial bundle over a torus

This is the example Nash has in mind when he says that line bundles with vanishing Chern class "can have rich holomorphic properties". Let $T = \mathbb{C}/(\mathbb{Z} + \tau\mathbb{Z})$. Describe line bundles by factors of automorphy as in Example 5.49: for a complex number a, a section of L_a is a function ψ on $\mathbb{C}$ with

$$\psi(z+1) = \psi(z), \qquad \psi(z+\tau) = e^{2\pi i a}\,\psi(z).$$

Degree zero. By the computation of Example 5.49, the number of zeros of a section in a fundamental parallelogram is $-\frac{1}{2\pi i}\int_0^1 \mathrm{d}\log e_\tau$; here $e_\tau = e^{2\pi i a}$ is constant, so $\mathrm{d}\log e_\tau = 0$ and $\deg L_a = 0$.

Smoothly trivial (take a real first). The smooth function $\psi_0(z) = e^{2\pi i a\, y/\mathrm{Im}\,\tau}$, with $y = \mathrm{Im}\, z$, is nowhere zero, and

$$\psi_0(z+1) \overset{(1)}{=} \psi_0(z), \qquad \psi_0(z+\tau) \overset{(2)}{=} e^{2\pi i a\,(y+\mathrm{Im}\,\tau)/\mathrm{Im}\,\tau} = e^{2\pi i a}\,\psi_0(z).$$

(1) Shifting z by 1 does not change y.
(2) Shifting z by τ raises y by $\mathrm{Im}\,\tau$.

So ψ_0 is a nowhere-vanishing *smooth* section, and L_a is smoothly trivial, as (5.53) demands for $c_1 = 0$. (For complex a use $\psi_0 = e^{2\pi i a\, y/\mathrm{Im}\,\tau}$ all the same; it is still nowhere zero.)

Holomorphically nontrivial unless $a \in \mathbb{Z} + \tau\mathbb{Z}$. A holomorphic section ψ is an entire function. First let a be real. Then $|\psi(z+1)| = |\psi(z+\tau)| = |\psi(z)|$, so $|\psi|$ is doubly periodic, hence bounded by its maximum on the compact parallelogram, so ψ is constant by Liouville's theorem, and a nonzero constant satisfies $\psi(z+\tau) = e^{2\pi i a}\psi$ only if $e^{2\pi i a} = 1$. So for $a \notin \mathbb{Z}$ the bundle L_a has *no* holomorphic sections at all, while the trivial bundle has the section 1: L_a is not holomorphically trivial.

For complex a, change the description: replace ψ by $\psi' = e^{2\pi i c z}\psi$ for a constant c. This is a global, holomorphic, nowhere-zero frame change, so it does not change the bundle, and it multiplies the two factors of automorphy by $e^{2\pi i c}$ and $e^{2\pi i c\tau}$. Two consequences.

- With $c = 1$ the factors become $(1,\ e^{2\pi i(a+\tau)})$, so $L_a \cong L_{a+\tau}$. Obviously also $L_a = L_{a+1}$. So L_a depends only on a modulo $\mathbb{Z} + \tau\mathbb{Z}$.
- Write $a = s + t\tau$ with s, t real and take $c = -t$: the factors become $(e^{-2\pi i t},\ e^{2\pi i s})$, both of modulus one. The Liouville argument now applies verbatim: a holomorphic section has doubly periodic modulus, so it is constant, and a nonzero constant needs $e^{-2\pi i t} = e^{2\pi i s} = 1$, i.e. $s, t \in \mathbb{Z}$.

So L_a is holomorphically trivial exactly when $a \in \mathbb{Z} + \tau\mathbb{Z}$, and $L_a \cong L_{a'}$ exactly when $a - a' \in \mathbb{Z} + \tau\mathbb{Z}$ (apply the same argument to $L_a \otimes L_{a'}^{-1} = L_{a-a'}$). Therefore

$$\mathrm{Pic}_0(T) \;\cong\; \mathbb{C}/(\mathbb{Z} + \tau\mathbb{Z}) \;=\; T.$$

The torus parametrises its own degree-zero line bundles: this is Abel's theorem (Example 5.40) seen from the other side, with $L_a \cong [P_a - O]$ for the point P_a corresponding to a (up to a sign convention).

Worked example 5.54: (5.56) computed on the torus: the lattice appears

Let us see the quotient $H^1(\mathcal{O})/H^1(\mathbb{Z})$ directly. By Dolbeault, $H^1(T;\mathcal{O}) = H^{0,1}_{\bar\partial}(T) = \mathbb{C}\,[\mathrm{d}\bar z]$; call the coordinate on it c (the class $c\,[\mathrm{d}\bar z]$). The map $H^1(T;\mathbb{Z}) \to H^1(T;\mathcal{O})$ takes a closed real 1-form with integer periods to the $(0,1)$-part of its class. With the cycles a (from 0 to 1) and b (from 0 to τ), the integral basis of $H^1(T;\mathbb{Z})$ is given by the constant forms α_a, α_b with $\int_a \alpha_a = \int_b \alpha_b = 1$ and $\int_b \alpha_a = \int_a \alpha_b = 0$. For a constant form $u\,\mathrm{d}x + v\,\mathrm{d}y$ we have $\int_a = u$ and $\int_b = u\,\mathrm{Re}\,\tau + v\,\mathrm{Im}\,\tau$, so

$$\alpha_a = \mathrm{d}x - \frac{\mathrm{Re}\,\tau}{\mathrm{Im}\,\tau}\,\mathrm{d}y, \qquad \alpha_b = \frac{\mathrm{d}y}{\mathrm{Im}\,\tau}.$$

Their $(0,1)$ parts, using $\mathrm{d}x = \frac12(\mathrm{d}z + \mathrm{d}\bar z)$ and $\mathrm{d}y = \frac{1}{2\mathrm{i}}(\mathrm{d}z - \mathrm{d}\bar z)$, are

$$\alpha_b^{0,1} \overset{(1)}{=} \frac{1}{\mathrm{Im}\,\tau}\cdot\Big(-\frac{1}{2\mathrm{i}}\Big)\mathrm{d}\bar z = \frac{\mathrm{i}}{2\,\mathrm{Im}\,\tau}\,\mathrm{d}\bar z,$$
$$\alpha_a^{0,1} \overset{(2)}{=} \Big(\frac12 - \frac{\mathrm{Re}\,\tau}{\mathrm{Im}\,\tau}\cdot\frac{\mathrm{i}}{2}\Big)\mathrm{d}\bar z \overset{(3)}{=} \frac{\mathrm{Im}\,\tau - \mathrm{i}\,\mathrm{Re}\,\tau}{2\,\mathrm{Im}\,\tau}\,\mathrm{d}\bar z \overset{(4)}{=} \frac{\mathrm{i}}{2\,\mathrm{Im}\,\tau}\,(-\tau)\,\mathrm{d}\bar z.$$

(1) The $\mathrm{d}\bar z$ coefficient of $\mathrm{d}y$ is $-\frac{1}{2\mathrm{i}} = \frac{\mathrm{i}}{2}$.
(2) The $\mathrm{d}\bar z$ coefficient of $\mathrm{d}x$ is $\frac12$, plus the $\mathrm{d}y$ term.
(3) Common denominator.
(4) $\mathrm{i}\cdot(-\tau) = -\mathrm{i}\,\mathrm{Re}\,\tau + \mathrm{Im}\,\tau$.

So the image of $H^1(T;\mathbb{Z})$ is the lattice $\frac{\mathrm{i}}{2\,\mathrm{Im}\,\tau}(\mathbb{Z} + \tau\mathbb{Z})$ in the c-line, and

$$\mathrm{Pic}_0(T) = \frac{H^1(T;\mathcal{O})}{H^1(T;\mathbb{Z})} = \frac{\mathbb{C}}{\frac{\mathrm{i}}{2\,\mathrm{Im}\,\tau}(\mathbb{Z}+\tau\mathbb{Z})} \cong \frac{\mathbb{C}}{\mathbb{Z}+\tau\mathbb{Z}},$$

the last step being multiplication by the nonzero complex number $2\,\mathrm{Im}\,\tau/\mathrm{i}$. The answer agrees with Example 5.53.

General curve. The same computation on Σ_p gives the Jacobian of Section 5.2.17. Identify $H^1(\Sigma_p;\mathcal{O})$ with the dual of $H^0(K)$ by Serre duality, $[\beta] \mapsto \big(\omega \mapsto \int_\Sigma \omega\wedge\beta\big)$. If α is an integral class, Poincaré dual to a cycle γ, then $\int_\Sigma \omega\wedge\alpha^{0,1} = \int_\Sigma \omega\wedge\alpha = \pm\int_\gamma \omega$ (the $(1,0)$ part of α drops out because $\omega\wedge\alpha^{1,0}$ is a $(2,0)$-form, which is zero on a curve). So the image of $H^1(\Sigma_p;\mathbb{Z})$ in $H^0(K)^* \cong \mathbb{C}^p$ is the set of period vectors $\big(\int_\gamma \omega_1, \dots, \int_\gamma \omega_p\big)$, which is the lattice Λ of (5.17). Hence

$$\mathrm{Pic}_0(\Sigma_p) = \mathbb{C}^p/\Lambda = J(\Sigma_p), \qquad \mathrm{Pic}(\Sigma_p) = \bigsqcup_{d\in\mathbb{Z}} \mathrm{Pic}_d(\Sigma_p), \tag{5.57}$$

where Pic_d, the bundles of degree d, is a copy of $J(\Sigma_p)$ (tensor with any fixed bundle of degree d, such as $[dP]$).

Physical picture: flat connections, Wilson lines, and holomorphic structures

Why should a physicist expect Pic_0 to be a torus of real dimension b_1? A degree-zero line bundle carries a *flat* U(1) connection (curvature zero), which is determined by its holonomies (Wilson loops) around the b_1 independent non-contractible loops: b_1 phases, a point of $\mathrm{U}(1)^{b_1} = (\mathbb{R}/\mathbb{Z})^{b_1}$. In Example 5.53 with a real, L_a is exactly the flat bundle with holonomy 1 around a and $e^{2\pi\mathrm{i}a}$ around b: an Aharonov–Bohm flux threading the hole

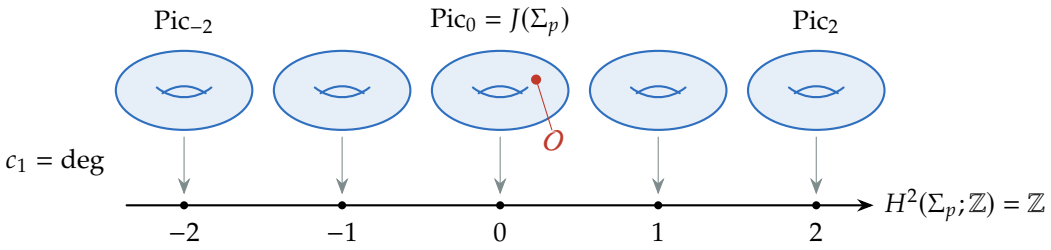


Figure 5.32: The Picard group of a compact Riemann surface. Over each integer d (the Chern class, or degree) sits a whole complex torus Pic_d of holomorphic line bundles, each a translate of the Jacobian $J(\Sigma_p) = \mathbb{C}^p/\Lambda$. Smoothly, everything in one torus is the same bundle; holomorphically, the points are all different. The trivial bundle $\mathcal{O}$ is one point (red) of Pic_0. For $\mathbb{CP}^1$ ($p = 0$) the tori shrink to points and $\mathrm{Pic} = \mathbb{Z}$.

of the torus. Its $\bar\partial$-operator $\bar\partial + A^{0,1}$ is the holomorphic structure. On a compact Kähler manifold every holomorphic line bundle of degree zero arises from exactly one flat unitary connection, so

$$\mathrm{Pic}_0(M) \;=\; \{\text{flat U(1) connections}\}/\text{gauge} \;=\; H^1(M;\mathbb{R})/H^1(M;\mathbb{Z}),$$

a real torus of dimension b_1 whose complex structure comes from $H^1(M;\mathbb{R}) \cong H^{0,1}$. The twisted fermion boundary conditions (spin structures and more general flat twists) in string theory on a torus are points of this torus.

Try it: the Picard group of an abelian surface

Let $M = \mathbb{C}^2/\Lambda$ be a complex 2-torus. Using $h^{0,1} = 2$ and $b_1 = 4$ (constant forms, as in Example 5.42), what is $\mathrm{Pic}_0(M)$ as a real manifold?
Answer. $H^1(M;\mathcal{O}) = \mathbb{C}^2$ modulo the image of $H^1(M;\mathbb{Z}) = \mathbb{Z}^4$, a lattice of rank 4 in $\mathbb{C}^2 = \mathbb{R}^4$: a real 4-torus T^4, carrying a complex structure (the *dual torus*). The discrete part $\mathrm{Pic}/\mathrm{Pic}_0$ is the group of integral $(1,1)$ classes; for generic Λ it is zero, which is one way to see that a generic 2-torus has no positive line bundle and is not algebraic (Example 5.17).

Reading Nash (pp. 144–146, (5.23)–(5.28))

Nash explains that a line bundle is determined by transition functions satisfying the cocycle condition, giving a class in $H^1(M;\mathcal{E}^*)$; writes the exponential sequence (5.23) and its long exact sequence (5.24); uses that $\mathcal{E}$ is fine to get $H^1(M;\mathcal{E}^*) \cong H^2(M;\mathbb{Z})$ (5.25), realised by c_1, so that smooth line bundles are classified by c_1 and $c_1 = 0$ means trivial; warns that this fails in higher rank; repeats the story with $\mathcal{O}$, $\mathcal{O}^*$ (5.26), observing that $\mathcal{O}$ is not fine; defines $\mathrm{Pic}(M) = H^1(M;\mathcal{O}^*)$ (5.27) and $\mathrm{Pic}_0(M) = H^1(M;\mathcal{O})/H^1(M;\mathbb{Z})$ (5.28). Nash's remark that "a global holomorphic section of L is a meromorphic function, and this is not holomorphically trivial" refers to the divisor picture of Section 5.2.12: sections of $[D]$ are meromorphic functions with poles bounded by D, and a bundle of degree 0 other than $\mathcal{O}$ has none. We added: the explicit connecting map (integer ambiguity of logarithms), its de Rham form as a gauge potential, the winding-number computation on S^2, the partition-of-unity proof that $\mathcal{E}$ is fine, a table of $H^2(M;\mathbb{Z})$ with torsion, a rank-2 counterexample on S^5, $\mathrm{Pic}(\mathbb{CP}^n) = \mathbb{Z}$, the explicit family L_a on the torus, the lattice computation, and the identification $\mathrm{Pic}_0(\Sigma_p) = J(\Sigma_p)$.

5.3.6 Hermitian metrics and the Kähler condition

So far we have used almost no metric: divisors, line bundles, Serre duality and the Picard group are all defined by the complex structure alone. But to do analysis (to find harmonic representatives, to define adjoints and Laplacians, to write a sigma-model action) we need a metric, and on a complex manifold we want one that *respects* the complex structure. The question of this subsection is: *which Riemannian metrics are compatible with the complex structure, and what extra condition makes them really useful?* The answer to the second part is Nash's (5.29): the associated 2-form should be closed.

The complex structure as a matrix. A complex manifold has, at every point, an endomorphism J of the real tangent space with $J^2 = -1$: "multiplication by i". In a holomorphic coordinate $z = x + \mathrm{i}y$ on $\mathbb{C}$,

$$J\partial_x = \partial_y, \qquad J\partial_y = -\partial_x, \qquad J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \text{ in the basis } (\partial_x, \partial_y),$$

a rotation by 90°. On $\mathbb{C}^n$ it is this block in each pair (x^k, y^k). Holomorphic changes of coordinates preserve J (their Jacobian is complex linear, which is exactly the statement that it commutes with J: the Cauchy–Riemann equations again). The complex vectors ∂_{z^k} are the $+\mathrm{i}$ eigenvectors of J and $\partial_{\bar z^k}$ the $-\mathrm{i}$ ones, which is why forms split into types.

Working definition 5.33 (Hermitian metric)

A Riemannian metric g on a complex manifold is **Hermitian** if J is an isometry: [J rotates each tangent vector by 90°, $J\partial_x = \partial_y$, and it must not change lengths or angles] $g(JX, JY) = g(X, Y)$ for all tangent vectors. In local holomorphic coordinates it is written (Nash's notation)

$$\mathrm{d}s^2 = g_{i\bar\jmath}\, \mathrm{d}z^i \otimes \mathrm{d}\bar z^j, \qquad g_{i\bar\jmath} = \overline{g_{j\bar\imath}}, \qquad (g_{i\bar\jmath}) \text{ positive definite},$$

[positive definite: $g_{i\bar\jmath} v^i \bar v^j > 0$ for every nonzero complex vector v] and its **associated 2-form** (Kähler form) is

$$\omega = \frac{\mathrm{i}}{2}\, g_{i\bar\jmath}\, \mathrm{d}z^i \wedge \mathrm{d}\bar z^j, \qquad \omega(X, Y) = g(JX, Y). \tag{5.58}$$

Operationally: a Hermitian metric is a positive Hermitian matrix of functions $g_{i\bar\jmath}(z, \bar z)$ in each chart, transforming like $\mathrm{d}z^i \otimes \mathrm{d}\bar z^j$ under holomorphic changes of coordinates. It has only n^2 real components instead of the $n(2n+1)$ of a general metric in real dimension $2n$.

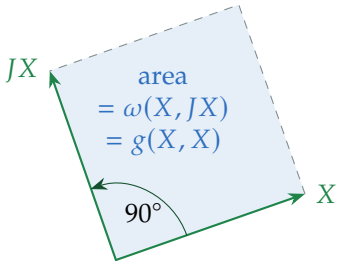
Let us see, in one complex dimension, how the three objects g, J, ω fit together. Take $\mathrm{d}s^2 = \lambda\, \mathrm{d}z \otimes \mathrm{d}\bar z$ with $\lambda > 0$ (a single positive function; every Hermitian metric on a Riemann surface is of this form in a local coordinate). Then

$$\begin{aligned} \lambda\, \mathrm{d}z \otimes \mathrm{d}\bar z &\overset{(1)}{=} \lambda\, (\mathrm{d}x + \mathrm{i}\mathrm{d}y) \otimes (\mathrm{d}x - \mathrm{i}\mathrm{d}y) \\ &\overset{(2)}{=} \lambda\, (\mathrm{d}x \otimes \mathrm{d}x + \mathrm{d}y \otimes \mathrm{d}y) - \mathrm{i}\,\lambda\, (\mathrm{d}x \otimes \mathrm{d}y - \mathrm{d}y \otimes \mathrm{d}x) \\ &\overset{(3)}{=} g - \mathrm{i}\,\omega, \qquad g = \lambda(\mathrm{d}x^2 + \mathrm{d}y^2), \quad \omega = \lambda\, \mathrm{d}x \wedge \mathrm{d}y. \end{aligned} \tag{5.59}$$

(1) $\mathrm{d}z = \mathrm{d}x + \mathrm{i}\mathrm{d}y$, $\mathrm{d}\bar z = \mathrm{d}x - \mathrm{i}\mathrm{d}y$.
(2) Multiply out; the cross terms are $-\mathrm{i}\,\mathrm{d}x \otimes \mathrm{d}y + \mathrm{i}\,\mathrm{d}y \otimes \mathrm{d}x$.

(3) The symmetric part is the Riemannian metric; the antisymmetric part is $dx \wedge dy = dx \otimes dy - dy \otimes dx$.

So *the real part of the Hermitian metric is the Riemannian metric, and minus the imaginary part is the 2-form* ω. Nash's formula for ω agrees: $\frac{i}{2}\lambda\, dz \wedge d\bar{z} = \frac{i}{2}\lambda(-2i)\, dx \wedge dy = \lambda\, dx \wedge dy$. And the formula $\omega(X,Y) = g(JX,Y)$ checks on the basis: $g(J\partial_x, \partial_y) = g(\partial_y, \partial_y) = \lambda = \omega(\partial_x, \partial_y)$. The metric g is conformal to the flat one: a Hermitian metric on a Riemann surface is the same thing as a conformal structure plus a choice of scale (Section 5.2.1).



J: rotate by 90° ("multiply by i")

g: lengths and angles, with $g(JX, JY) = g(X, Y)$

$\omega(X,Y) = g(JX,Y)$: oriented areas

any two of (g, J, ω) determine the third

Figure 5.33: The compatible triple on a single tangent plane. J rotates a vector by a right angle without changing its length. The 2-form ω measures the area of the square spanned by X and JX, which is $|X|^2$; so ω is never degenerate. A Hermitian metric is the Riemannian metric g seen through J; the Kähler condition is the extra requirement that the area form ω, assembled over the whole manifold, is closed.

Working definition 5.34 (Kähler manifold (Nash's (5.29)))

A complex manifold with a Hermitian metric is **Kähler** if the associated 2-form [$\omega(X,Y) = g(JX,Y)$, the area form built from g and J] is closed:

$$ds^2 = g_{i\bar{j}}\, dz^i \otimes d\bar{z}^j, \qquad \omega = \frac{i}{2} g_{i\bar{j}}\, dz^i \wedge d\bar{z}^j, \qquad d\omega = 0. \tag{5.60}$$

The metric is then a **Kähler metric**. *Operational test*: check $\partial_k g_{i\bar{j}} = \partial_i g_{k\bar{j}}$ for all i, j, k (derived next).

The Kähler condition in components. Since ω has type $(1,1)$, $d\omega = \partial\omega + \bar{\partial}\omega$ splits into a $(2,1)$ and a $(1,2)$ part, and each must vanish. Moreover ω is real ($\bar{\omega} = \omega$, because $g_{i\bar{j}}$ is Hermitian), so $\bar{\partial}\omega = \overline{\partial\omega}$ and it is enough to look at $\partial\omega$:

$$\partial\omega \overset{(1)}{=} \frac{i}{2}\, \partial_k g_{i\bar{j}}\, dz^k \wedge dz^i \wedge d\bar{z}^j \overset{(2)}{=} \frac{i}{4}\left(\partial_k g_{i\bar{j}} - \partial_i g_{k\bar{j}}\right) dz^k \wedge dz^i \wedge d\bar{z}^j.$$

(1) ∂ differentiates the coefficient in the holomorphic directions.

(2) $dz^k \wedge dz^i$ is antisymmetric in k, i, so only the antisymmetric part of the coefficient survives; write it out.

The forms $dz^k \wedge dz^i \wedge d\bar{z}^j$ with $k < i$ are independent, so

$$d\omega = 0 \quad \Longleftrightarrow \quad \partial_k g_{i\bar{j}} = \partial_i g_{k\bar{j}} \quad \text{for all } i, j, k. \tag{5.61}$$

This is the integrability condition for $g_{i\bar{j}}$ to be a mixed second derivative:

Key idea: Kähler potential

Locally, a metric is Kähler if and only if there is a real function K, the **Kähler potential**, with

$$g_{i\bar{j}} = \partial_i \partial_{\bar{j}} K, \qquad \omega = \frac{\mathrm{i}}{2}\,\partial\bar{\partial} K.$$

K is determined up to $K \mapsto K + f + \bar{f}$ with f holomorphic (these are killed by $\partial\bar{\partial}$). A Kähler metric in complex dimension n is thus described by *one* real function, instead of n^2.

Why. If $g_{i\bar{j}} = \partial_i \partial_{\bar{j}} K$ then $\partial_k g_{i\bar{j}} = \partial_k \partial_i \partial_{\bar{j}} K$ is symmetric in k, i, so (5.61) holds. Conversely, suppose $\mathrm{d}\omega = 0$ on a ball. By the ordinary Poincaré lemma $\omega = \mathrm{d}\alpha$ with α a real 1-form; split $\alpha = \alpha^{1,0} + \alpha^{0,1}$ with $\alpha^{1,0} = \overline{\alpha^{0,1}}$. Comparing types in $\omega = \partial\alpha^{1,0} + (\bar{\partial}\alpha^{1,0} + \partial\alpha^{0,1}) + \bar{\partial}\alpha^{0,1}$, the $(0,2)$ part gives $\bar{\partial}\alpha^{0,1} = 0$, so by the $\bar{\partial}$-Poincaré lemma (Example 5.41) $\alpha^{0,1} = \bar{\partial} f$ for some complex function f, and then $\alpha^{1,0} = \partial\bar{f}$. The $(1,1)$ part is

$$\omega \overset{(1)}{=} \bar{\partial}\partial\bar{f} + \partial\bar{\partial} f \overset{(2)}{=} \partial\bar{\partial}\big(f - \bar{f}\big) \overset{(3)}{=} \frac{\mathrm{i}}{2}\,\partial\bar{\partial}\big(4\,\mathrm{Im}\, f\big),$$

(1) Substitute $\alpha^{1,0} = \partial\bar{f}$ and $\alpha^{0,1} = \bar{\partial} f$.
(2) $\bar{\partial}\partial = -\partial\bar{\partial}$ (5.36).
(3) $f - \bar{f} = 2\mathrm{i}\,\mathrm{Im}\, f$ and $2\mathrm{i} = \frac{\mathrm{i}}{2}\cdot 4$.

so $K = 4\,\mathrm{Im}\, f$ works. (The $(2,0)$ part $\partial\alpha^{1,0} = \partial\partial\bar{f} = 0$ is automatic.)

Kähler normal coordinates. Nash states that a Hermitian metric is Kähler if and only if around every point there are holomorphic coordinates in which $g_{i\bar{j}} = \delta_{ij} + h_{ij}$ with h_{ij} of order 2: the metric agrees with the flat one *to first order*. In Riemannian geometry this is always possible with *real* coordinates (Riemann normal coordinates); the point is that for a Kähler metric it can be done with *holomorphic* ones. Let us check it. Put the point at $z = 0$ and make a linear change so that $g_{i\bar{j}}(0) = \delta_{ij}$. By Hermiticity the first-order Taylor expansion has the form

$$g_{i\bar{j}}(z) = \delta_{ij} + a_{ijk}\, z^k + \overline{a_{jik}}\,\bar{z}^k + O(|z|^2),$$

and the Kähler condition (5.61) at 0 says $a_{ijk} = a_{kji}$. Try new holomorphic coordinates w with $z^i = w^i + \tfrac{1}{2} c^i_{kl} w^k w^l$, $c^i_{kl} = c^i_{lk}$, so that $\mathrm{d}z^i = \mathrm{d}w^i + c^i_{kl} w^l\,\mathrm{d}w^k$. To first order in w,

$$\begin{aligned} g_{i\bar{j}}\,\mathrm{d}z^i \otimes \mathrm{d}\bar{z}^j &\overset{(1)}{=} \big(\delta_{ij} + a_{ijm} w^m + \overline{a_{jim}}\,\bar{w}^m\big)\big(\mathrm{d}w^i + c^i_{kl} w^l \mathrm{d}w^k\big) \otimes \big(\mathrm{d}\bar{w}^j + \overline{c^j_{kl}}\,\bar{w}^l \mathrm{d}\bar{w}^k\big) \\ &\overset{(2)}{=} \Big[\delta_{ij} + \big(a_{ijm} + c^j_{im}\big) w^m + \big(\overline{a_{jim}} + \overline{c^i_{jm}}\big)\bar{w}^m\Big]\mathrm{d}w^i \otimes \mathrm{d}\bar{w}^j + O(|w|^2). \end{aligned}$$

(1) Substitute the expansion of g and of $\mathrm{d}z^i$; to this order $z^m = w^m$ in the coefficient.
(2) Keep terms of order ≤ 1; in the cross term $\delta_{ij} c^i_{kl} w^l\,\mathrm{d}w^k \otimes \mathrm{d}\bar{w}^j$ rename $k \to i$, $i \to j$, $l \to m$, and similarly for its conjugate.

Choose $c^j_{im} = -a_{ijm}$. This is allowed only if the right side is symmetric in i, m, i.e. $a_{ijm} = a_{mji}$: *exactly the Kähler condition*. With this choice both first-order brackets vanish (the second is the conjugate of the first with $i \leftrightarrow j$), and $g'_{i\bar{j}} = \delta_{ij} + O(|w|^2)$. Conversely, $\mathrm{d}\omega$ involves only first derivatives of g, so if they vanish at a point in some holomorphic coordinates, $\mathrm{d}\omega = 0$ there. So:

Key idea: Kähler $\iff$ flat to first order in holomorphic coordinates

A Hermitian metric is Kähler if and only if at each point there are holomorphic coordinates with $g_{i\bar{j}} = \delta_{ij} + O(|z|^2)$. Consequently, any identity that involves the metric and at most its first derivatives, and that holds on flat $\mathbb{C}^n$, holds on every Kähler manifold.

The second sentence is the working principle behind all of Kähler geometry, and we shall use it for the Laplacians below. Geometrically it says that the Levi-Civita connection preserves J ($\nabla J = 0$), so parallel transport commutes with "multiplication by i" and the holonomy lies in $\mathrm{U}(n) \subset \mathrm{SO}(2n)$.

Kähler manifolds are symplectic. ω is closed, and it is nondegenerate: in coordinates with $g_{i\bar{j}}(x) = \delta_{ij}$ at a point, $\omega = \sum_k \mathrm{d}x^k \wedge \mathrm{d}y^k$ there, and

$$\omega^n \overset{(1)}{=} \Big(\sum_{k=1}^{n} \mathrm{d}x^k \wedge \mathrm{d}y^k\Big)^n \overset{(2)}{=} n!\, \mathrm{d}x^1 \wedge \mathrm{d}y^1 \wedge \cdots \wedge \mathrm{d}x^n \wedge \mathrm{d}y^n \overset{(3)}{=} n!\, \mathrm{vol}_g. \tag{5.62}$$

(1) ω at the point, from (5.59) with $\lambda = 1$ in each complex direction.
(2) The 2-forms $\mathrm{d}x^k \wedge \mathrm{d}y^k$ commute and square to zero, so only products of n distinct ones survive, and there are $n!$ orderings of them.
(3) At that point g is the Euclidean metric in the coordinates (x^k, y^k), whose volume form is the product of all the differentials.

A closed nondegenerate 2-form is a **symplectic form**, so Kähler manifolds are symplectic manifolds, as Nash notes: phase spaces, in physics language. Here is the first topological consequence, straight from Stokes.

Worked example 5.55: The powers of ω are never exact on a compact Kähler manifold

Let M be compact Kähler of complex dimension n, and $1 \le k \le n$. Suppose $\omega^k = \mathrm{d}\beta$. Then

$$n!\,\mathrm{Vol}(M) \overset{(1)}{=} \int_M \omega^n \overset{(2)}{=} \int_M \mathrm{d}\beta \wedge \omega^{n-k} \overset{(3)}{=} \int_M \mathrm{d}\big(\beta \wedge \omega^{n-k}\big) \overset{(4)}{=} \int_{\partial M} \beta \wedge \omega^{n-k} \overset{(5)}{=} 0,$$

(1) Integrate (5.62).
(2) $\omega^n = \omega^k \wedge \omega^{n-k}$ and the assumption.
(3) Leibniz rule: $\mathrm{d}(\beta \wedge \omega^{n-k}) = \mathrm{d}\beta \wedge \omega^{n-k} \pm \beta \wedge \mathrm{d}(\omega^{n-k})$, and $\mathrm{d}\omega = 0$.
(4) Stokes.
(5) M is closed.

a contradiction, since the volume is positive. So $[\omega^k] \neq 0$ in $H^{2k}(M;\mathbb{R})$ and $b_{2k}(M) \ge 1$ for $0 \le k \le n$. In particular $b_2 \ge 1$: *a compact Kähler manifold always has a nonzero second Betti number*. This is item (ii) of Nash's list on p. 148, and it already rules out many complex manifolds (see the Hopf surface in Example 5.64).

Worked example 5.56: Nash's first two examples: flat tori and Riemann surfaces

Complex tori. On $\mathbb{C}^n$ the metric $g_{i\bar{j}} = \delta_{ij}$ has $\omega = \frac{\mathrm{i}}{2}\sum_k \mathrm{d}z^k \wedge \mathrm{d}\bar{z}^k = \sum_k \mathrm{d}x^k \wedge \mathrm{d}y^k$, with constant coefficients, so $\mathrm{d}\omega = 0$. Its Kähler potential is $K = \sum_k |z^k|^2$ (check: $\partial_i \partial_{\bar{j}} \sum_k z^k \bar{z}^k = \delta_{ij}$). The metric is invariant under translations, so it descends to every quotient $T = \mathbb{C}^n/\Lambda$ by a lattice $\Lambda \cong \mathbb{Z}^{2n}$ of translations: every complex torus is Kähler with a flat metric.
Riemann surfaces. On a complex curve ω is a 2-form on a real 2-manifold, so $\mathrm{d}\omega$ is a 3-form on a 2-manifold: zero. *Every* Hermitian metric $\lambda|\mathrm{d}z|^2$ on a Riemann surface is Kähler.

For instance the hyperbolic metric on the upper half plane $\mathbb{H} = \{y > 0\}$ has potential $K = -4\log y$:

$$\partial_z\partial_{\bar z}(-4\log y) \overset{(1)}{=} -4\,\partial_z\partial_{\bar z}\log(z-\bar z) \overset{(2)}{=} -4\,\partial_z\Big(\frac{-1}{z-\bar z}\Big) \overset{(3)}{=} -4\,\frac{1}{(z-\bar z)^2} \overset{(4)}{=} \frac{1}{y^2},$$

(1) $y = (z-\bar z)/2\mathrm{i}$, and the constant $\log(2\mathrm{i})$ is killed by the derivatives.
(2) $\partial_{\bar z}(z-\bar z) = -1$.
(3) $\partial_z(z-\bar z)^{-1} = -(z-\bar z)^{-2}$, and the two minus signs cancel.
(4) $(z-\bar z)^2 = (2\mathrm{i}y)^2 = -4y^2$.

so $\mathrm{d}s^2 = (\mathrm{d}x^2+\mathrm{d}y^2)/y^2$, the Poincaré metric of constant curvature -1. Every surface of genus $p \geq 2$ is a quotient of $\mathbb{H}$ by a group of isometries (uniformisation, part 5b), so it inherits this Kähler metric.

Watch out: Nash's $\mathbb{C}^n/\mathbb{Z}^n$

Nash writes the flat torus as "$\mathbb{C}^n/\mathbb{Z}^n$". For a compact quotient one needs a lattice of *real* rank $2n$, i.e. $\mathbb{C}^n/\Lambda$ with $\Lambda \cong \mathbb{Z}^{2n}$ spanning $\mathbb{C}^n$ over $\mathbb{R}$ (for $n=1$, $\Lambda = \mathbb{Z}+\tau\mathbb{Z}$). The quotient $\mathbb{C}^n/\mathbb{Z}^n$ by the integer points of $\mathbb{R}^n \subset \mathbb{C}^n$ is $(\mathbb{C}^*)^n$, which is not compact. Presumably "$\mathbb{Z}^n$" is shorthand for a lattice.

5.3.7 The Fubini–Study metric on $\mathbb{CP}^n$

The most important example, and the one that makes every projective variety Kähler, is $\mathbb{CP}^n$ with the Fubini–Study metric. Use homogeneous coordinates $[Z_0 : \cdots : Z_n]$ and, on the chart $U_0 = \{Z_0 \neq 0\}$, the affine coordinates $z^i = Z_i/Z_0$, $i = 1,\dots,n$ (Section 5.1.6). Write $|z|^2 = \sum_{k=1}^n z^k\bar z^k$.

Working definition 5.35 (Fubini–Study metric (Nash's (5.30)))

On U_0 the **Fubini–Study metric** is the Kähler metric with potential $K = \log(1+|z|^2)$ [one real function K with $g_{i\bar\jmath} = \partial_i\partial_{\bar\jmath}K$]:

$$g_{i\bar\jmath} = \partial_i\partial_{\bar\jmath}\log\big(1+|z|^2\big) = \frac{\delta_{ij}\big(1+|z|^2\big) - \bar z^i z^j}{\big(1+|z|^2\big)^2}, \qquad \omega = \frac{\mathrm{i}}{2}\,\partial\bar\partial\log\big(1+|z|^2\big). \tag{5.63}$$

In homogeneous coordinates, $\omega = \frac{\mathrm{i}}{2}\partial\bar\partial\log|Z|^2$, with $|Z|^2 = \sum_{a=0}^n |Z_a|^2$ [this is unchanged by $Z \mapsto \lambda Z$, because $\partial\bar\partial$ kills $\log|\lambda|^2$. So ω is defined on all of $\mathbb{CP}^n$].

Computing $g_{i\bar\jmath}$. Two derivatives of the potential:

$$\partial_i\partial_{\bar\jmath}\log\big(1+|z|^2\big) \overset{(1)}{=} \partial_i\Big(\frac{z^j}{1+|z|^2}\Big) \overset{(2)}{=} \frac{\delta_{ij}}{1+|z|^2} - \frac{z^j\,\bar z^i}{\big(1+|z|^2\big)^2} \overset{(3)}{=} \frac{\delta_{ij}\big(1+|z|^2\big) - \bar z^i z^j}{\big(1+|z|^2\big)^2}.$$

(1) $\partial_{\bar\jmath}|z|^2 = \partial_{\bar\jmath}\sum_k z^k\bar z^k = z^j$.
(2) Quotient rule, with $\partial_i z^j = \delta_{ij}$ and $\partial_i|z|^2 = \bar z^i$.
(3) Common denominator.

It is Hermitian: $\overline{g_{j\bar\imath}} = \big(\delta_{ji}(1+|z|^2) - z^j\bar z^i\big)/(1+|z|^2)^2 = g_{i\bar\jmath}$. It is closed, being $\frac{\mathrm{i}}{2}\partial\bar\partial$ of a function ($\mathrm{d}\partial\bar\partial = (\partial+\bar\partial)\partial\bar\partial = 0$ by (5.36)); this is Nash's "it is immediate that the form ω is closed".

It is positive definite. Regard g as the matrix $G = \big[(1+|z|^2)\mathbb{1} - \bar z\, z^T\big]/(1+|z|^2)^2$, where $\bar z\, z^T$ has entries $\bar z^i z^j$. It acts on a vector $v \in \mathbb{C}^n$ by $\bar z\, z^T v = \bar z\,(z^T v)$. Two kinds of eigenvectors:

$$v = \bar z: \quad G\bar z \overset{(1)}{=} \frac{(1+|z|^2)\bar z - \bar z\,|z|^2}{(1+|z|^2)^2} = \frac{1}{(1+|z|^2)^2}\,\bar z,$$
$$z^T v = 0: \quad Gv \overset{(2)}{=} \frac{1}{1+|z|^2}\,v.$$

(1) $z^T\bar z = \sum_k z^k\bar z^k = |z|^2$.
(2) The rank-one term annihilates v.

The eigenvalues $1/(1+|z|^2)^2$ (once, the "radial" direction) and $1/(1+|z|^2)$ ($n-1$ times, the "angular" directions) are positive. They decrease as $|z|\to\infty$, which is how a metric on the noncompact chart $\mathbb{C}^n$ manages to have finite total volume.

It is well defined on all of $\mathbb{CP}^n$. On $U_1 = \{Z_1 \neq 0\}$ use the same formula with Z_1 in place of Z_0. On the overlap the two potentials differ by

$$K_0 - K_1 \overset{(1)}{=} \log\frac{|Z|^2}{|Z_0|^2} - \log\frac{|Z|^2}{|Z_1|^2} \overset{(2)}{=} \log\Big|\frac{Z_1}{Z_0}\Big|^2 \overset{(3)}{=} \log w + \log\bar w, \qquad w = \frac{Z_1}{Z_0},$$

(1) $1+|z|^2 = \sum_a |Z_a|^2/|Z_0|^2$ on U_0, and similarly on U_1.
(2) Subtract the logarithms; $|Z|^2$ cancels.
(3) w is holomorphic and nowhere zero on $U_0\cap U_1$, so it has local logarithms.

which is (locally) a holomorphic plus an antiholomorphic function, killed by $\partial\bar\partial$. So the local forms ω agree on overlaps: ω is a global Kähler form. The potential itself is not global (no nonconstant real function on a compact complex manifold can be a Kähler potential: at its maximum $\partial_i\partial_{\bar j}K$ would be negative semidefinite).

It is homogeneous. The unitary group $\mathrm{U}(n+1)$ acts on $\mathbb{CP}^n$ by $[Z]\mapsto[UZ]$ and preserves $|Z|^2$, hence ω. It acts transitively (any unit vector can be rotated to any other). So the Fubini–Study metric looks the same at every point, and it is enough to understand it at the origin $z=0$ of U_0. There,

$$g_{i\bar j} = \delta_{ij}\big(1-|z|^2\big) - \bar z^i z^j + O(|z|^4), \tag{5.64}$$

from $(1+|z|^2)^{-1} = 1-|z|^2+\dots$ and $(1+|z|^2)^{-2} = 1-2|z|^2+\dots$. There is no first-order term: the affine coordinates are Kähler normal coordinates at the origin, exactly as Nash remarks ("h_{ij} is of order 2 in the coordinates z_i").

Watch out: typos in Nash's (5.30)

Nash's display (5.30) has three slips. (1) The potential is $\log(1+|z_1|^2+\cdots+|z_n|^2)$, not $\log(1+z_1^2+\cdots+z_n^2)$; the latter is not real and not even defined everywhere. (2) The factor $\frac{\mathrm{i}}{2}$ belongs in ω only; in $\mathrm{d}s^2 = g_{i\bar j}\mathrm{d}z^i\otimes\mathrm{d}\bar z^j$ and in $g_{i\bar j}$ itself there is no $\frac{\mathrm{i}}{2}$ (a metric has to be real and positive on the diagonal). (3) With $\mathrm{d}s^2 = g_{i\bar j}\mathrm{d}z^i\otimes\mathrm{d}\bar z^j$, the numerator is $\delta_{ij}(1+|z|^2) - \bar z_i z_j$; Nash's $z_i\bar z_j$ is its transpose (the complex conjugate matrix), which amounts to swapping the roles of $\mathrm{d}z$ and $\mathrm{d}\bar z$. The last line, $\omega = \frac{\mathrm{i}}{2}\partial\bar\partial\log(1+|z|^2)$, is correct as printed, and it is the formula to remember.

Worked example 5.57: The Fubini–Study metric on $\mathbb{CP}^1$ is the round sphere of radius $\frac{1}{2}$

For $n = 1$, (5.63) reads

$$\mathrm{d}s^2 = \frac{\mathrm{d}z\,\mathrm{d}\bar{z}}{(1+|z|^2)^2} = \frac{\mathrm{d}x^2+\mathrm{d}y^2}{(1+x^2+y^2)^2}, \qquad \omega = \frac{\mathrm{i}}{2}\frac{\mathrm{d}z\wedge\mathrm{d}\bar{z}}{(1+|z|^2)^2} = \frac{\mathrm{d}x\wedge\mathrm{d}y}{(1+x^2+y^2)^2},$$

(the 1×1 matrix has $g_{1\bar{1}} = ((1+|z|^2)-|z|^2)/(1+|z|^2)^2$). Which surface is this? Substitute the stereographic coordinate $z = \tan(\theta/2)\,e^{\mathrm{i}\phi}$, with $\theta \in [0,\pi]$ the polar angle from the point $z = 0$:

$$\frac{|\mathrm{d}z|^2}{(1+|z|^2)^2} \overset{(1)}{=} \frac{\frac{1}{4}\sec^4(\theta/2)\,\mathrm{d}\theta^2 + \tan^2(\theta/2)\,\mathrm{d}\phi^2}{\sec^4(\theta/2)}$$
$$\overset{(2)}{=} \tfrac{1}{4}\,\mathrm{d}\theta^2 + \sin^2(\theta/2)\cos^2(\theta/2)\,\mathrm{d}\phi^2 \overset{(3)}{=} \tfrac{1}{4}\big(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2\big).$$

(1) $\mathrm{d}z = e^{\mathrm{i}\phi}\big(\frac{1}{2}\sec^2(\theta/2)\,\mathrm{d}\theta + \mathrm{i}\tan(\theta/2)\,\mathrm{d}\phi\big)$, whose two terms are orthogonal (one real, one imaginary, times the same phase); and $1+\tan^2 = \sec^2$.
(2) Divide; $\tan^2\cos^4 = \sin^2\cos^2$.
(3) $2\sin(\theta/2)\cos(\theta/2) = \sin\theta$.

This is the round metric of a sphere of radius $R = \frac{1}{2}$ ($R^2(\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2)$). Three checks.

- *Area.* $4\pi R^2 = \pi$, matching $\int_{\mathbb{CP}^1}\omega = \pi$ computed in Example 5.17.
- *Curvature.* For a conformal metric $\lambda(\mathrm{d}x^2+\mathrm{d}y^2)$ the Gaussian curvature is $K_G = -\frac{1}{2\lambda}\Delta\log\lambda$ with $\Delta = \partial_x^2+\partial_y^2 = 4\partial_z\partial_{\bar{z}}$. Here

$$K_G \overset{(1)}{=} -\frac{(1+|z|^2)^2}{2}\cdot 4\,\partial_z\partial_{\bar{z}}\big(-2\log(1+|z|^2)\big) \overset{(2)}{=} -\frac{(1+|z|^2)^2}{2}\cdot 4\cdot\frac{-2}{(1+|z|^2)^2} = 4 = \frac{1}{R^2}.$$

  (1) $\lambda = (1+|z|^2)^{-2}$, so $1/\lambda = (1+|z|^2)^2$ and $\log\lambda = -2\log(1+|z|^2)$.
  (2) $\partial_z\partial_{\bar{z}}\log(1+|z|^2) = (1+|z|^2)^{-2}$, from (5.24).

- *Gauss–Bonnet.* $\int K_G\,\mathrm{d}A = 4\cdot\pi = 4\pi = 2\pi\chi(S^2)$. ✓

Finally compare with the curvature of $\mathcal{O}(1)$ with its standard metric (Example 5.33): $c_1(\mathcal{O}(1)) = -\frac{\mathrm{i}}{2\pi}\partial\bar{\partial}\log(1+|z|^2)^{-1} = \frac{1}{\pi}\,\omega$. So

$$\frac{\omega_{FS}}{\pi} = c_1\big(\mathcal{O}(1)\big)\,:$$

the Fubini–Study form is (up to π) the field strength of the unit monopole, and its class is integral. That is why $\mathbb{CP}^n$ passes Kodaira's test (Section 5.1.10) with $\omega = \omega_{FS}/\pi$.

Physical picture: Fubini–Study in quantum mechanics

$\mathbb{CP}^n$ is the space of *pure states* of an $(n+1)$-level quantum system: a state is a ray $[\psi]$ in $\mathbb{C}^{n+1}$. The Fubini–Study line element is the natural distance between rays:

$$\mathrm{d}s^2 = \frac{\langle\mathrm{d}\psi|\mathrm{d}\psi\rangle}{\langle\psi|\psi\rangle} - \frac{|\langle\psi|\mathrm{d}\psi\rangle|^2}{\langle\psi|\psi\rangle^2},$$

the part of $\mathrm{d}\psi$ orthogonal to ψ (the component along ψ only changes the phase or normalisation, which is not physical). For $\psi = (1,z)$ this gives $\big(|\mathrm{d}z|^2(1+|z|^2) - |\bar{z}\,\mathrm{d}z|^2\big)/(1+$

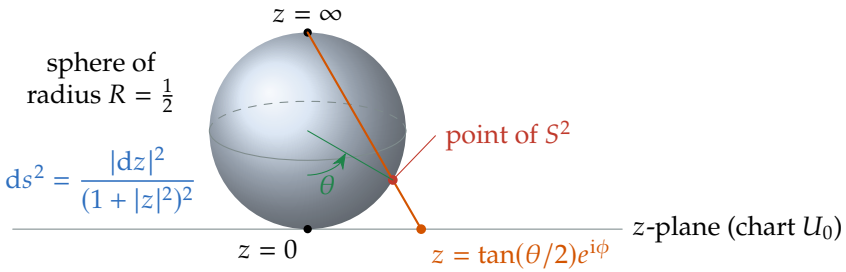


Figure 5.34: The Fubini–Study metric on $\mathbb{CP}^1$ is the round metric on a sphere of radius $\frac{1}{2}$, written in the stereographic coordinate z. Rays from the point $z = \infty$ (top) identify the sphere minus that point with the z-plane; a point at polar angle θ from the bottom lands at $|z| = \tan(\theta/2)$. Far out in the plane the conformal factor $(1 + |z|^2)^{-2}$ is tiny, so the whole infinite plane has finite area π. (The picture is drawn with the sphere enlarged; only the angles matter for the projection.)

$|z|^2)^2 = |dz|^2/(1 + |z|^2)^2$, which is (5.63). For a spin-$\frac{1}{2}$ the ray is a point of the Bloch sphere, and ω is (half) the *Berry curvature*: the Berry phase around a loop on the Bloch sphere is minus half the enclosed solid angle, and 2ω integrates to 2π, the Chern number 1 of the monopole bundle $\mathcal{O}(1)$ (or $\mathcal{O}(-1)$, depending on the convention for the phase). The $\mathbb{CP}^n$ sigma models of Chapter 8 have this metric on their target space.

5.3.8 Complex submanifolds: projective varieties are Kähler

Nash's next observation is short and very powerful. Let $\iota : M \hookrightarrow N$ be a complex submanifold of a Kähler manifold (N, g_N, ω_N): a submanifold whose tangent spaces are J-invariant (locally the zero set of holomorphic functions with independent differentials, Section 5.1.9). Give M the induced metric $g_M = \iota^* g_N$. Then:

- g_M is Hermitian, because J preserves TM and is an isometry of g_N.
- Its 2-form is the restriction of ω_N: for X, Y tangent to M,

$$\omega_M(X,Y) \overset{(1)}{=} g_M(JX,Y) \overset{(2)}{=} g_N(JX,Y) \overset{(3)}{=} \omega_N(X,Y), \qquad \text{so } \omega_M = \iota^*\omega_N.$$

  (1) Definition (5.58) on M.
  (2) JX is again tangent to M, and g_M is g_N on such vectors.
  (3) Definition (5.58) on N.

- It is closed: $d\omega_M = d\iota^*\omega_N = \iota^* d\omega_N = 0$, since pull-back commutes with d (Chapter T1).

(Nash's text has "N is also Kähler" where he means M.) Combining with the Fubini–Study metric:

Key idea: projective ⇒ Kähler

Every complex submanifold of $\mathbb{CP}^n$, in particular every smooth projective variety, is a Kähler manifold, with the restriction of ω_{FS} as Kähler form. Its class $[\omega_{FS}/\pi]$ is integral, being c_1 of the restriction of $\mathcal{O}(1)$.

This is the converse direction of Kodaira's theorem (Section 5.1.10): an algebraic manifold carries a closed positive $(1,1)$-form with rational (indeed integral) class. Kodaira's theorem says that this property also characterises algebraic manifolds.

Worked example 5.58: The area of a complex curve is π times its degree

Take the conic $C = \{Z_0 Z_2 = Z_1^2\} \subset \mathbb{CP}^2$, parametrised by $t \mapsto [1 : t : t^2]$ (together with $[0 : 0 : 1]$ at $t = \infty$). Pull back the Fubini–Study potential: on U_0, $K = \log(1+|Z_1/Z_0|^2+|Z_2/Z_0|^2)$ becomes $K(t) = \log(1 + |t|^2 + |t|^4)$, a function of $r = |t|$ only. For any radial potential,

$$\int_{\mathbb{C}} \frac{\mathrm{i}}{2}\partial\bar\partial K \overset{(1)}{=} \int_{\mathbb{R}^2} \frac{\mathrm{i}}{2}\cdot\frac{\Delta K}{4}\cdot(-2\mathrm{i})\,\mathrm{d}x\,\mathrm{d}y \overset{(2)}{=} \frac{1}{4}\int_0^{2\pi}\int_0^{\infty}\frac{1}{r}\big(rK'\big)'\,r\,\mathrm{d}r\,\mathrm{d}\phi \overset{(3)}{=} \frac{\pi}{2}\Big[rK'(r)\Big]_0^{\infty}.$$

(1) $\partial\bar\partial K = \partial_t\partial_{\bar t}K\,\mathrm{d}t\wedge\mathrm{d}\bar t$ with $\partial_t\partial_{\bar t} = \frac{1}{4}\Delta$ and $\mathrm{d}t\wedge\mathrm{d}\bar t = -2\mathrm{i}\,\mathrm{d}x\wedge\mathrm{d}y$.

(2) The Laplacian of a radial function is $\frac{1}{r}(rK')'$; polar coordinates.

(3) Fundamental theorem of calculus in r, and $\int \mathrm{d}\phi = 2\pi$.

For the conic, $rK' = \dfrac{2r^2+4r^4}{1+r^2+r^4}$, which is 0 at $r = 0$ and tends to 4. So

$$\mathrm{Area}(C) = \frac{\pi}{2}\cdot 4 = 2\pi.$$

For $\mathbb{CP}^1$ itself ($K = \log(1 + r^2)$, $rK' \to 2$) the same formula gives π. In general, for the degree-d curve $t \mapsto [1 : t : \cdots : t^d]$, $rK' \to 2d$ and the area is πd. This is no accident: ω_{FS}/π is $c_1(\mathcal{O}(1))$, and $\int_C c_1(\mathcal{O}(1)|_C)$ is the degree of the divisor cut on C by a hyperplane, i.e. the number of intersection points of C with a line (Section 5.2.11, Section 5.1.8): 2 for a conic.

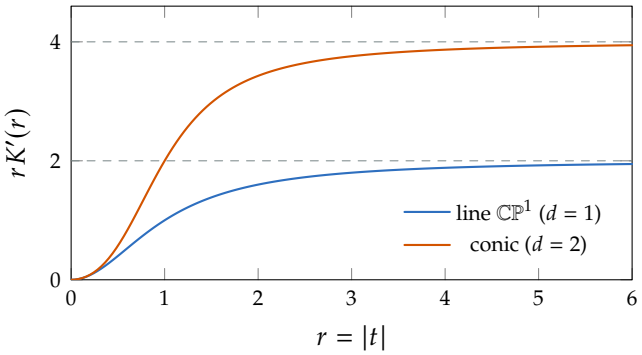


Figure 5.35: The area of a holomorphic curve in $\mathbb{CP}^n$ read off from its Kähler potential. For a radial potential the area is $\frac{\pi}{2}\lim_{r\to\infty} rK'(r)$. For a curve of degree d the potential grows like $2d\log r$, so $rK' \to 2d$ (dashed lines) and the area is πd. In sigma-model language: the action of a holomorphic map (an instanton) is its area, fixed by its topological degree.

Physical picture: holomorphic curves minimise area: BPS bounds

The formula "area = $\pi\times$degree" is a special case of a general fact (Wirtinger's inequality): on a Kähler manifold, $\mathrm{Area}(S) \geq \int_S \omega$ for every surface S, with equality exactly for complex curves. Since $\int_S \omega$ depends only on the homology class of S (ω is closed, Stokes), complex curves are area minimising in their class. In the $\mathbb{CP}^n$ sigma model the action of a map $\Sigma \to \mathbb{CP}^n$ is bounded below by $\int \omega$, a topological charge, and the bound is saturated by holomorphic maps: these are the sigma-model instantons, the two-dimensional analogue of self-dual Yang–Mills fields (Chapter 8).

5.3.9 Hodge theory on Kähler manifolds

We now reach the reason Kähler manifolds are so important: their cohomology splits by type, and the splitting forces strong restrictions on their topology. Recall the question we postponed in Section 5.2.16 and Section 5.3.3: why does a surface of genus p have exactly p independent holomorphic 1-forms? We answer it here.

Three Laplacians. On a compact Hermitian manifold we have an L^2 inner product on forms, $\langle \alpha, \beta \rangle = \int_M \alpha \wedge \star\bar{\beta}$ (the pointwise inner product integrated against the volume form), and hence formal adjoints d^*, ∂^*, $\bar{\partial}^*$, defined by $\langle \mathrm{d}\alpha, \beta \rangle = \langle \alpha, \mathrm{d}^*\beta \rangle$ and so on. They lower the degree: $\bar{\partial}^* : A^{p,q} \to A^{p,q-1}$. From each operator we build a Laplacian, Nash's (5.31):

$$\Delta_d = \mathrm{d}\mathrm{d}^* + \mathrm{d}^*\mathrm{d}, \qquad \Delta_\partial = \partial\partial^* + \partial^*\partial, \qquad \Delta_{\bar{\partial}} = \bar{\partial}\bar{\partial}^* + \bar{\partial}^*\bar{\partial}. \tag{5.65}$$

$\Delta_{\bar{\partial}}$ preserves the type (p, q), while Δ_d in general mixes types. The Hodge theorem (Chapter 2, Chapter 4) says that each Laplacian is elliptic and each cohomology class has a unique harmonic representative:

$$H^k_{\mathrm{dR}}(M; \mathbb{C}) \cong \{\alpha \in A^k : \Delta_d \alpha = 0\}, \qquad H^{p,q}_{\bar{\partial}}(M) \cong \{\alpha \in A^{p,q} : \Delta_{\bar{\partial}} \alpha = 0\}.$$

Operationally, $\Delta\alpha = 0$ is equivalent to the pair of first-order equations $\mathrm{d}\alpha = 0 = \mathrm{d}^*\alpha$ (resp. $\bar{\partial}\alpha = 0 = \bar{\partial}^*\alpha$), because $\langle \Delta\alpha, \alpha \rangle = \|\mathrm{d}\alpha\|^2 + \|\mathrm{d}^*\alpha\|^2$.

In words: a harmonic form is the "most relaxed" representative of its class. Among all closed forms $h + \mathrm{d}\alpha$ in a class, the harmonic one h has the smallest L^2 norm, because $\|h + \mathrm{d}\alpha\|^2 = \|h\|^2 + \|\mathrm{d}\alpha\|^2$ (the cross term is $\langle h, \mathrm{d}\alpha \rangle = \langle \mathrm{d}^* h, \alpha \rangle = 0$). If the form is a magnetic field, the norm is its energy, and the harmonic field is the least-energy field with the given fluxes.

Hodge decomposition at work: Maxwell and Proca

Before refining the Hodge theorem with the complex structure, we ask what the plain (de Rham) version *does* for a physicist. The test case is the simplest gauge theory: Maxwell's equations in the language of forms (Chapter T1), and their massive cousin, the Proca equation. The splitting of forms into three orthogonal pieces will sort the field into gauge, physical and topological parts. It will also show a striking difference: a massless photon can carry topological flux, and a massive one cannot.

The three operators. We use the conventions of Chapter T1. On an oriented n-manifold whose metric has s minus signs, the exterior derivative d raises the degree by one, and the **codifferential** (the adjoint d^* of the previous paragraph) lowers it:

$$\delta = \mathrm{d}^* = (-1)^{n(k+1)+1+s} \star \mathrm{d} \star \quad \text{on } k\text{-forms}, \qquad \Delta = \mathrm{d}\delta + \delta\mathrm{d}. \tag{5.66}$$

The sign is fixed so that $\langle \mathrm{d}\alpha, \beta \rangle = \langle \alpha, \delta\beta \rangle$. Two special cases are worth remembering. On 1-forms in flat Euclidean space δ is minus the divergence, $\delta(A_j \mathrm{d}x^j) = -\partial_j A_j$, so $\Delta = -\nabla^2$ is a *positive* operator. In Minkowski space with $\eta = \mathrm{diag}(-1, +1, +1, +1)$, $\delta = +\star\mathrm{d}\star$ in every degree, and Δ acts on Cartesian components as $\partial_t^2 - \nabla^2$, the wave operator. We write A^k for the smooth k-forms, as above. (The letter Ω^p is taken in this chapter by the sheaf of holomorphic p-forms.)

Three algebraic facts do all the work below: $\mathrm{d}^2 = 0$, $\delta^2 = 0$, and Δ commutes with d. The second holds because $\delta^2 = \pm \star \mathrm{d} \star \star \mathrm{d} \star = \pm \star \mathrm{d}\mathrm{d} \star = 0$ (the middle $\star\star$ is a sign). The third is a two-line check:

$$\mathrm{d}\Delta \overset{(1)}{=} \mathrm{d}\mathrm{d}\delta + \mathrm{d}\delta\mathrm{d} \overset{(2)}{=} \mathrm{d}\delta\mathrm{d} \overset{(3)}{=} \mathrm{d}\delta\mathrm{d} + \delta\mathrm{d}\mathrm{d} \overset{(4)}{=} \Delta\mathrm{d}.$$

(1) Definition (5.66) of Δ.
(2) $\mathrm{dd} = 0$.
(3) Add $\delta\mathrm{dd} = 0$.
(4) Definition of Δ again, now with d on the right.

In the same way $\delta\Delta = \delta\mathrm{d}\delta = \Delta\delta$. So Δ maps exact forms to exact forms ($\Delta\mathrm{d}\alpha = \mathrm{d}(\Delta\alpha)$) and co-exact forms to co-exact forms.

Working definition 5.36 (the Hodge decomposition, and how to compute it)

Let M be compact, oriented and Riemannian, without boundary. Every k-form splits uniquely as

$$\phi = \underbrace{\mathrm{d}\alpha}_{\text{exact}} + \underbrace{\delta\beta}_{\text{co-exact}} + \underbrace{\omega}_{\text{harmonic}}, \qquad A^k = \mathrm{d}A^{k-1} \oplus \delta A^{k+1} \oplus \mathcal{H}^k, \qquad \mathcal{H}^k = \operatorname{Ker}\Delta \cong H^k_{\mathrm{dR}}(M), \tag{5.67}$$

and the three pieces are mutually orthogonal [in the inner product $\int_M \phi \wedge \star\psi$. Co-exact means $\delta\beta$, with δ the adjoint of d. Harmonic means $\Delta\omega = 0$]. *Recipe:* expand ϕ in eigenforms of Δ. The zero modes make up ω. [the eigenforms with $\lambda = 0$, i.e. nonzero solutions of $\Delta\phi = 0$] An eigenform with $\Delta\phi_\lambda = \lambda\phi_\lambda$, $\lambda \neq 0$, splits as $\phi_\lambda = \lambda^{-1}\mathrm{d}(\delta\phi_\lambda) + \lambda^{-1}\delta(\mathrm{d}\phi_\lambda)$, which is just $\Delta = \mathrm{d}\delta + \delta\mathrm{d}$ divided by λ. *Tests:* ϕ is closed exactly when its co-exact part vanishes, and co-closed exactly when its exact part vanishes.

The orthogonality is one line each: $\langle \mathrm{d}\alpha, \delta\beta\rangle = \langle \mathrm{dd}\alpha, \beta\rangle = 0$, $\langle \mathrm{d}\alpha, \omega\rangle = \langle \alpha, \delta\omega\rangle = 0$ and $\langle \delta\beta, \omega\rangle = \langle \beta, \mathrm{d}\omega\rangle = 0$. The isomorphism with cohomology follows from the tests: closed forms are $\mathrm{d}A^{k-1} \oplus \mathcal{H}^k$, so "closed modulo exact" is $\mathcal{H}^k$. The existence of the splitting is the analytic content of the Hodge theorem (Chapter 2). Figure 5.36 shows the picture we will use.

Where this sits in the section. The three Laplacians of (5.65) each have their own version of (5.67). Below, the Kähler identity $\Delta_d = 2\Delta_{\bar\partial}$ says the harmonic sector $\mathcal{H}^k$ is the same for d and for $\bar\partial$, so it splits further by type into pieces of dimension $h_{p,q}$. And counting harmonic forms is counting zero modes, which is what Riemann–Roch (Section 5.3.3) and the index theorem (Chapter 4) do. Maxwell and Proca show what the three sectors mean physically.

Maxwell: one equation is free, the other is dynamics. Recall from Chapter T1 that the field strength is a 2-form F, the potential a 1-form A with $F = \mathrm{d}A$, and that Maxwell's equations are $\mathrm{d}F = 0$ and $\delta F = J$. The first is not an equation of motion at all:

$$\mathrm{d}F \overset{(1)}{=} \mathrm{dd}A \overset{(2)}{=} 0.$$

(1) $F = \mathrm{d}A$ on each patch where the potential is defined.
(2) $\mathrm{d}^2 = 0$. No field equation was used: this holds for every A.

We call this **kinematic**: it is built into the description of the field. The second equation, in vacuum $\delta F = 0$, is **dynamical**: it comes from varying the action. Put together,

$$\Delta F \overset{(1)}{=} \mathrm{d}(\delta F) + \delta(\mathrm{d}F) \overset{(2)}{=} \mathrm{d}(0) + \delta(0) \overset{(3)}{=} 0.$$

(1) Definition (5.66) of Δ.
(2) The vacuum equation $\delta F = 0$ and the Bianchi identity $\mathrm{d}F = 0$.
(3) d and δ are linear.

A source-free Maxwell field is harmonic. On a compact Riemannian (Euclidean) space the converse also holds, by the positivity argument of the previous paragraph. So the space of source-free

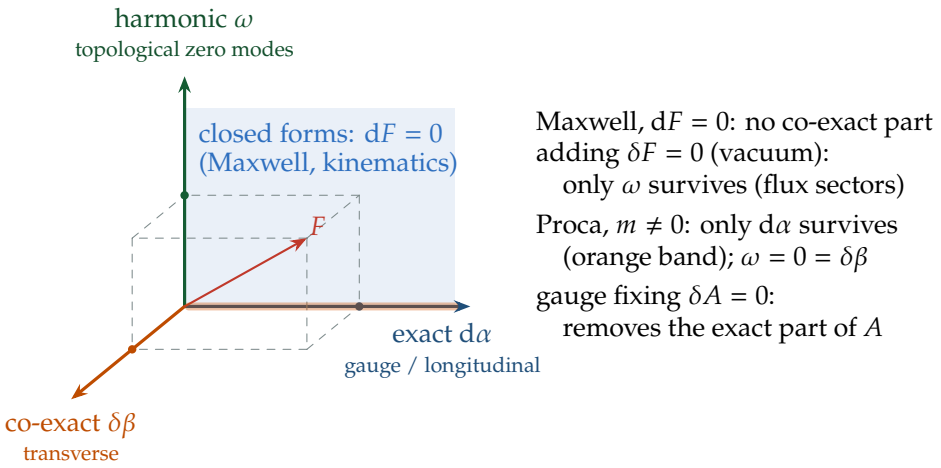


Figure 5.36: The three orthogonal sectors of $A^k(M)$, with the mathematician's names (exact, co-exact, harmonic) and the physicist's (gauge or longitudinal, transverse, topological). A form F (red) is the sum of its three projections (dots). The shaded plane is the space of closed forms: the Bianchi identity $dF = 0$ confines a Maxwell field to it, and the vacuum equation $\delta F = 0$ then leaves only the harmonic axis. The Proca equation instead confines F to the exact axis, so a massive field has no harmonic (topological) part.

fields is exactly $\mathcal{H}^2 \cong H^2_{\mathrm{dR}}(M)$: it is finite-dimensional, and its dimension b_2 is a topological invariant.

There is one subtlety, and it is the whole point. If A were a single global 1-form, $F = dA$ would be exact, its harmonic part would vanish, and the vacuum field would be zero. Harmonic fields exist because $F = dA$ holds only patch by patch: A is a connection on a line bundle that may be nontrivial (Section 5.3.4). Globally, only $dF = 0$ survives.

Flux is topological. For a closed surface Σ (no boundary) and a closed F, write $F = d\alpha + \omega$ (no co-exact part, by the test above). Then

$$\int_\Sigma F \overset{(1)}{=} \int_\Sigma d\alpha + \int_\Sigma \omega \overset{(2)}{=} \int_{\partial\Sigma} \alpha + \int_\Sigma \omega \overset{(3)}{=} \int_\Sigma \omega .$$

(1) The decomposition (5.67), with $\delta\beta = 0$ because F is closed.
(2) Stokes' theorem (Chapter T2); α is a global form, since the exact part is globally exact.
(3) $\partial\Sigma = \varnothing$.

So the flux through any closed surface sees only the harmonic part. It does not change when Σ is deformed (Stokes again, since $d\omega = 0$). With charged matter present it is also quantised, as the next example shows.

Worked example 5.59: Uniform flux on a torus: harmonic, but not exact

Let $T^2 = \mathbb{R}^2/(L_1\mathbb{Z} \times L_2\mathbb{Z})$ with the flat metric $dx^2 + dy^2$ and orientation $dx \wedge dy$. Take a constant magnetic field,

$$F = B\, dx \wedge dy, \qquad B \in \mathbb{R} \text{ constant}.$$

It is harmonic. Closedness is automatic (a 2-form in two dimensions). For co-closedness, (5.66) with $n = 2$, $k = 2$, $s = 0$ gives the sign $(-1)^{2\cdot 3+1} = -1$, and

$$\delta F \overset{(1)}{=} - \star\, d \star (B\, dx \wedge dy) \overset{(2)}{=} - \star\, dB \overset{(3)}{=} 0.$$

(1) The codifferential (5.66) on 2-forms in two Euclidean dimensions.

(2) $\star(\mathrm{d}x \wedge \mathrm{d}y) = 1$, since $\mathrm{d}x \wedge \mathrm{d}y$ is the volume form.
(3) B is constant.

It is not exact. If $F = \mathrm{d}A$ for a global 1-form A on T^2, then

$$BL_1L_2 \overset{(1)}{=} \int_{T^2} F \overset{(2)}{=} \int_{T^2} \mathrm{d}A \overset{(3)}{=} \int_{\partial T^2} A \overset{(4)}{=} 0,$$

(1) $\int_{T^2} \mathrm{d}x \wedge \mathrm{d}y$ is the area L_1L_2.
(2) The assumption $F = \mathrm{d}A$.
(3) Stokes' theorem.
(4) The torus has no boundary.

a contradiction when $B \neq 0$. So there is no global potential. There is a local one, $A = Bx\,\mathrm{d}y$ (indeed $\mathrm{d}A = B\,\mathrm{d}x \wedge \mathrm{d}y$), on the strip $0 \le x < L_1$. It fails to be periodic in x, but only by a gauge transformation:

$$A(x + L_1, y) - A(x, y) \overset{(1)}{=} B(x + L_1)\,\mathrm{d}y - Bx\,\mathrm{d}y \overset{(2)}{=} \mathrm{d}\lambda, \qquad \lambda(y) = BL_1\,y.$$

(1) Substitute into $A = Bx\,\mathrm{d}y$.
(2) $BL_1\,\mathrm{d}y = \mathrm{d}(BL_1y)$.

Flux quantisation. A particle of charge e has wavefunction ψ with $D\psi = (\mathrm{d} - \mathrm{i}eA)\psi$ (units $\hbar = 1$), and a gauge transformation $A \to A + \mathrm{d}\lambda$ acts by $\psi \to e^{\mathrm{i}e\lambda}\psi$. So the gluing across $x \sim x + L_1$ is the transition function $g(y) = e^{\mathrm{i}eBL_1y}$. It must be a well-defined function on the circle $y \sim y + L_2$:

$$1 \overset{(1)}{=} \frac{g(y + L_2)}{g(y)} \overset{(2)}{=} e^{\mathrm{i}eBL_1L_2} \quad \Longrightarrow \quad e\,\Phi = e\,BL_1L_2 \in 2\pi\mathbb{Z}.$$

(1) Single-valuedness of g on the y-circle.
(2) Substitute $g(y) = e^{\mathrm{i}eBL_1y}$.

So $\Phi = 2\pi N/e$: the flux through the torus is an integer number N of flux quanta, and N is the degree (first Chern class) of the line bundle, $N = \frac{e}{2\pi}\int_{T^2} F$ (Section 5.2.11). This is the setting of the lowest Landau level on the torus (Section 5.3.3): the N theta functions of Example 5.49 are the N ground states. Note also that $F = B\,\mathrm{d}x \wedge \mathrm{d}y = \frac{\mathrm{i}}{2}B\,\mathrm{d}z \wedge \mathrm{d}\bar{z}$ is B times the Kähler form of the flat torus, a harmonic form of type $(1,1)$: it is the representative of $h_{1,1} = 1$ in Example 5.62.

The monopole is the same phenomenon. Around a magnetic monopole of charge g, the spatial slice $\mathbb{R}^3 \setminus \{0\}$ has $H^2 = \mathbb{R}$, generated by the sphere. The field $F = \frac{g}{4\pi}\sin\theta\,\mathrm{d}\theta \wedge \mathrm{d}\phi$ (the flux form of $\vec{B} = g\hat{r}/4\pi r^2$) is closed, co-closed ($\star F = \frac{g}{4\pi r^2}\mathrm{d}r$ is closed), and not exact, since its flux through every sphere is $g \neq 0$. On each sphere it is a constant multiple of the area form, the harmonic generator of $H^2_{\mathrm{dR}}(S^2)$. Dirac's condition $eg \in 2\pi\mathbb{Z}$ is the same gluing argument with two hemispheres instead of a strip (Example 5.50, Chapter T2).

Proca: a massive photon. Now give the photon a mass m. The Proca action adds a mass term to the Maxwell action of Chapter T1,

$$S[A] = \int \left(-\tfrac{1}{2}\,F \wedge \star F - \tfrac{1}{2}\,m^2 A \wedge \star A\right), \qquad F = \mathrm{d}A.$$

Varying $A \to A + a$, the mass term contributes $-m^2\, a \wedge \star A$. This is exactly the form of the coupling $+a \wedge \star J$ that produced $\delta F = J$, with J replaced by $-m^2 A$. So the field equation is

$$\delta F + m^2 A = 0. \tag{5.68}$$

(In components, with $\delta F = J \iff \partial_\mu F^{\nu\mu} = J^\nu$: $\partial_\mu F^{\mu\nu} = m^2 A^\nu$, the standard Proca equation in signature $(-,+,+,+)$.) Texts in which Δ has the opposite sign write this as $\delta F - m^2 A = 0$; the sector analysis below is the same in either convention.

A two-form equation, with no potential assumed. Apply d to (5.68) and use $\mathrm{d}A = F$:

$$\mathrm{d}\delta F + m^2 F = 0. \tag{5.69}$$

Now forget A. Treat (5.69) as an equation for an unknown 2-form F on a compact Riemannian M, and do not assume F has a potential, or even that it is closed. Decompose it by (5.67):

$$F = \mathrm{d}\alpha + \delta\beta + \omega, \qquad \mathrm{d}\omega = 0 = \delta\omega.$$

First we compute the left-hand side sector by sector. The codifferential only sees the exact part:

$$\delta F \overset{(1)}{=} \delta\mathrm{d}\alpha + \delta\delta\beta + \delta\omega \overset{(2)}{=} \delta\mathrm{d}\alpha.$$

(1) Linearity of δ.
(2) $\delta^2 = 0$, and $\delta\omega = 0$ because ω is harmonic.

Applying d gives the Laplacian of the exact part:

$$\mathrm{d}\delta F \overset{(1)}{=} \mathrm{d}\delta(\mathrm{d}\alpha) \overset{(2)}{=} \mathrm{d}\delta(\mathrm{d}\alpha) + \delta\mathrm{d}(\mathrm{d}\alpha) \overset{(3)}{=} \Delta(\mathrm{d}\alpha) \overset{(4)}{=} \mathrm{d}(\Delta\alpha).$$

(1) The previous chain.
(2) Add $\delta\mathrm{d}\mathrm{d}\alpha = 0$, which vanishes by $\mathrm{d}^2 = 0$.
(3) Definition of Δ.
(4) Δ commutes with d. So $\mathrm{d}\delta F$ is exact.

Substituting into (5.69) and grouping by sector:

$$0 \overset{(1)}{=} \Delta(\mathrm{d}\alpha) + m^2\big(\mathrm{d}\alpha + \delta\beta + \omega\big) \overset{(2)}{=} \underbrace{(\Delta + m^2)\mathrm{d}\alpha}_{\text{exact}} + \underbrace{m^2\,\delta\beta}_{\text{co-exact}} + \underbrace{m^2\,\omega}_{\text{harmonic}}.$$

(1) The previous chain for $\mathrm{d}\delta F$, and the decomposition of F.
(2) Regroup. The first term is $\mathrm{d}\big((\Delta + m^2)\alpha\big)$, so it is exact.

Three mutually orthogonal vectors add to zero only if each is zero. To see it explicitly, take the inner product of the equation with $\delta\beta$:

$$0 \overset{(1)}{=} \big\langle \delta\beta, (\Delta + m^2)\mathrm{d}\alpha \big\rangle + m^2\langle \delta\beta, \delta\beta\rangle + m^2\langle \delta\beta, \omega\rangle \overset{(2)}{=} m^2\,\|\delta\beta\|^2.$$

(1) Linearity of the inner product.
(2) Co-exact is orthogonal to exact and to harmonic (the one-line checks after (5.67)).

The same with ω and with $\mathrm{d}\alpha$ gives the other two. So the one equation (5.69) is three:

$$(\Delta + m^2)\,\mathrm{d}\alpha = 0, \qquad m^2\,\delta\beta = 0, \qquad m^2\,\omega = 0. \tag{5.70}$$

For $m \neq 0$, the co-exact and harmonic parts vanish, and $F = \mathrm{d}\alpha$ is exact. A potential exists, and it was not put in by hand: the equation of motion forced it. In fact it can be written down directly, in any signature and without Hodge theory:

$$F \overset{(1)}{=} -\frac{1}{m^2}\,\mathrm{d}\delta F \overset{(2)}{=} \mathrm{d}\Big(-\frac{1}{m^2}\,\delta F\Big).$$

(1) Solve (5.69) for F; this needs $m \neq 0$.
(2) m is a constant, so it passes through d.

The potential $A = -m^{-2}\delta F$ is exactly what (5.68) says it should be.

The Lorenz condition comes for free. Return to (5.68) and apply δ:

$$0 \overset{(1)}{=} \delta\big(\delta F + m^2 A\big) \overset{(2)}{=} \delta\delta F + m^2\,\delta A \overset{(3)}{=} m^2\,\delta A.$$

(1) Apply δ to the field equation (5.68).
(2) Linearity.
(3) $\delta^2 = 0$.

So $\delta A = 0$ for $m \neq 0$. In Maxwell theory this is a gauge *choice* (the Lorenz gauge); here it is a *consequence* of the field equation, and there is no gauge freedom left to choose. Then

$$\Delta A \overset{(1)}{=} \mathrm{d}\delta A + \delta \mathrm{d} A \overset{(2)}{=} \delta \mathrm{d} A \overset{(3)}{=} \delta F \overset{(4)}{=} -m^2 A, \qquad \text{so} \qquad (\Delta + m^2)A = 0, \quad \delta A = 0.$$

(1) Definition of Δ.
(2) $\delta A = 0$, just shown.
(3) $F = \mathrm{d}A$.
(4) The field equation (5.68).

In Minkowski space $\Delta + m^2 = \partial_t^2 - \nabla^2 + m^2$ on each Cartesian component: this is the Klein–Gordon equation, and $\delta A = -\partial_\mu A^\mu = 0$ removes one of the four components, leaving the three polarisations of a massive spin-1 particle.

Key idea: kinematic versus dynamical exactness

In Maxwell theory, F is closed *kinematically*: $\mathrm{d}F = 0$ is the identity $\mathrm{d}^2 = 0$ applied to local potentials, true for every configuration. Closed is weaker than exact, and the gap between them, the harmonic sector $\mathcal{H}^2 \cong H^2_{\mathrm{dR}}(M)$, is where topological flux lives.
In Proca theory, F is exact *dynamically*: the field equation projects onto the three sectors and gives $m^2\omega = 0$ and $m^2\delta\beta = 0$. The mass term kills the harmonic sector. *A massive photon cannot carry topological flux.*

Physical picture: harmonic modes are zero modes, and a mass gaps them

On the harmonic sector every derivative term vanishes: $\mathrm{d}\omega = 0$ and $\delta\omega = 0$. So there the Maxwell operator is 0, and harmonic fields cost no gradient energy. They are *zero modes*. The Proca operator $\mathrm{d}\delta + m^2$ acts on the same sector as multiplication by m^2, which is never zero: the mass lifts every zero mode by the same amount, so the field equation forces every harmonic component to vanish.
A superconductor shows this in the laboratory. Inside it the photon is effectively massive (the Higgs mechanism, with $m = 1/\lambda_L$ for the London penetration depth λ_L), and a static field obeys the London equation $(\Delta + m^2)\vec{B} = 0$: it decays like e^{-x/λ_L} from the surface. This is the Meissner effect. Flux can enter a type-II superconductor only through vortices, where the order parameter vanishes and the mass switches off locally. Each vortex carries one flux quantum $2\pi/e_*$ (with $e_* = 2e$ the Cooper-pair charge), the quantised flux of Example 5.59, squeezed into a tube of radius about λ_L around the core.

Worked example 5.60: Proca mode by mode on the torus T^3

Let $M = T^3 = (\mathbb{R}/2\pi\mathbb{Z})^3$ with the flat metric. A form is a sum of Fourier modes $e^{\mathrm{i}k\cdot x} f$, with $k \in \mathbb{Z}^3$ and f a constant-coefficient form. Write k also for the 1-form $k_j \mathrm{d}x^j$, and ι_k for contraction with the vector $k_j \partial_j$.

Step 1: the operators on a mode. Since $\mathrm{d}(e^{\mathrm{i}k\cdot x}) = \mathrm{i}e^{\mathrm{i}k\cdot x}k$ and, in flat space, $\delta = -\sum_j \iota_{\partial_j}\partial_j$ (the general form of "δ is minus the divergence"),

$$\Delta\big(e^{\mathrm{i}k\cdot x} f\big) \overset{(1)}{=} e^{\mathrm{i}k\cdot x}\big[(\mathrm{i}k\wedge)(-\mathrm{i}\iota_k) + (-\mathrm{i}\iota_k)(\mathrm{i}k\wedge)\big] f \overset{(2)}{=} e^{\mathrm{i}k\cdot x}\big[k\wedge\iota_k f + \iota_k(k\wedge f)\big] \overset{(3)}{=} |k|^2\, e^{\mathrm{i}k\cdot x} f.$$

(1) On a mode, d acts as $\mathrm{i}k\wedge$ and δ as $-\mathrm{i}\iota_k$.
(2) $(\mathrm{i})(-\mathrm{i}) = 1$.
(3) $\iota_k(k\wedge f) = (\iota_k k)f - k\wedge\iota_k f$ and $\iota_k k = |k|^2$.

So the harmonic modes are the $k = 0$ modes, as in Example 5.63. For $k \neq 0$ the recipe of the working definition splits f into $f_\parallel = |k|^{-2} k\wedge\iota_k f$ (exact: it contains the factor k) and $f_\perp = |k|^{-2}\iota_k(k\wedge f)$ (co-exact: $\iota_k f_\perp = 0$).

Step 2: the sectors of a 2-form. Take $k = (1,0,0)$, so $k = \mathrm{d}x$. Then

$$e^{\mathrm{i}x}\,\mathrm{d}x\wedge\mathrm{d}y \overset{(1)}{=} \mathrm{d}\big(-\mathrm{i}e^{\mathrm{i}x}\mathrm{d}y\big), \qquad e^{\mathrm{i}x}\,\mathrm{d}y\wedge\mathrm{d}z \overset{(2)}{=} \delta\big(\mathrm{i}e^{\mathrm{i}x}\,\mathrm{d}x\wedge\mathrm{d}y\wedge\mathrm{d}z\big).$$

(1) $\mathrm{d}(e^{\mathrm{i}x}) = \mathrm{i}e^{\mathrm{i}x}\mathrm{d}x$ and $(-\mathrm{i})(\mathrm{i}) = 1$.
(2) $\delta = -\mathrm{i}\iota_k$ on the mode, $\iota_{\partial_x}(\mathrm{d}x\wedge\mathrm{d}y\wedge\mathrm{d}z) = \mathrm{d}y\wedge\mathrm{d}z$, and $(-\mathrm{i})(\mathrm{i}) = 1$. (Directly: $\delta = -\star\mathrm{d}\star$ on 3-forms in three Euclidean dimensions gives the same.)

So for this k the exact 2-forms are spanned by $\mathrm{d}x\wedge\mathrm{d}y$ and $\mathrm{d}x\wedge\mathrm{d}z$ (two of them) and the co-exact ones by $\mathrm{d}y\wedge\mathrm{d}z$ (one). At $k = 0$ there are three harmonic 2-forms, $\mathrm{d}y\wedge\mathrm{d}z$, $\mathrm{d}z\wedge\mathrm{d}x$, $\mathrm{d}x\wedge\mathrm{d}y$: $b_2(T^3) = 3$, three independent fluxes.

Step 3: what survives. On a mode, $\mathrm{d}\delta$ acts as $k\wedge\iota_k$, which is $|k|^2$ on $f_\parallel$ and 0 on $f_\perp$ and on $k = 0$. The two equations become

mode	Maxwell in vacuum: $\mathrm{d}F = 0,\ \delta F = 0$	Proca: $\mathrm{d}\delta F + m^2 F = 0$
$k = 0$, harmonic	$0 = 0$: all 3 survive	$m^2 f = 0$: killed
$k \neq 0$, co-exact $f_\perp$	$k\wedge f_\perp = 0$: killed	$m^2 f_\perp = 0$: killed
$k \neq 0$, exact $f_\parallel$	$\iota_k f_\parallel = 0$: killed	$(\lvert k\rvert^2 + m^2) f_\parallel = 0$: gapped

The last entry is (5.70) mode by mode. In Euclidean signature $|k|^2 + m^2 \geq m^2 > 0$, so this sector is killed too: a static massive field has no source-free configurations on a compact space, and a source produces a field falling off like e^{-mr} (Yukawa). In real time, on $\mathbb{R}_t \times T^3$ with $A = a\, e^{\mathrm{i}(k\cdot x - \omega t)}$, the same factor becomes $-\omega^2 + |k|^2 + m^2$, because $\Delta = \partial_t^2 - \nabla^2$. The exact sector is where the propagating massive photon lives, with $\omega^2 = |k|^2 + m^2$. Its lowest frequency is m, even at $k = 0$: the Maxwell zero modes (constant fluxes, and the constant Wilson lines counted by $b_1(T^3) = 3$) have been lifted by the gap. Figure 5.37 plots both spectra.

Watch out: where the argument is rigorous, and the signs

- *Compactness.* The clean splitting (5.67) holds on a compact Riemannian manifold without boundary. On $\mathbb{R}^n$ one needs fall-off conditions (for example square-integrable forms), and on a manifold with boundary one needs boundary conditions. With such conditions, harmonic forms can disappear: $\mathbb{R}^3$ has no square-integrable harmonic 2-forms, which is why flux sectors need topology.

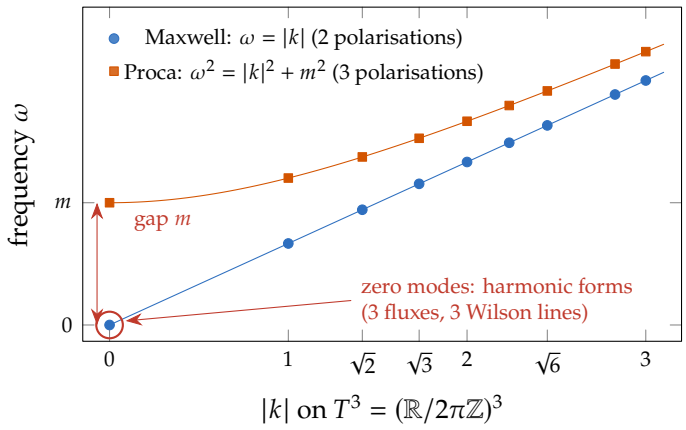


Figure 5.37: The spectrum of photon modes on a spatial torus T^3, one dot per allowed value of $|k|$ (only the values $|k|^2 = 0, 1, 2, 3, 4, 5, 6, 8, 9$ occur). A massless photon has frequency $\omega = |k|$, so the $k = 0$ sector (circled) consists of zero modes: the harmonic forms, which carry the topological fluxes and Wilson lines. A massive photon has $\omega = \sqrt{|k|^2 + m^2}$: every mode, including $k = 0$, is lifted by at least m, and no zero modes remain. This is the spectral meaning of $m^2\omega = 0$ in (5.70).

- *Lorentzian signature.* In Minkowski space the "inner product" $\int F \wedge \star F$ is indefinite (it is $\int (B^2 - E^2)$), so orthogonal pieces need not vanish separately, and $\Delta = \partial_t^2 - \nabla^2$ is not elliptic: its kernel contains every free wave, not a finite-dimensional space. There the sector argument is formal. The conclusions $F = \mathrm{d}(-m^{-2}\delta F)$, $\delta A = 0$ and $(\Delta + m^2)A = 0$ were derived directly, with only $\mathrm{d}^2 = \delta^2 = 0$, so they hold in every signature.

- *Signs.* With δ as in (5.66) and signature $(-,+,+,+)$, $\Delta = -\nabla^2 \geq 0$ in Euclidean signature and $\Delta = \partial_t^2 - \nabla^2$ in Minkowski signature. In both, the Klein–Gordon operator is $\Delta + m^2$, and the Proca equations are $\delta F + m^2 A = 0$ and $\mathrm{d}\delta F + m^2 F = 0$. Books whose Δ is the d'Alembertian $\Box = -\partial_t^2 + \nabla^2$ write $\Delta - m^2$ and $\mathrm{d}\delta F - m^2 F = 0$. The sector equations change only by the sign of m^2, and the conclusions (exactness, $\delta A = 0$, no harmonic part) are the same.

Dictionary: one splitting, two uses

sector	mathematician	physicist
exact $\mathrm{d}\alpha$	zero in cohomology	pure gauge, $A \to A + \mathrm{d}\lambda$; longitudinal modes
co-exact $\delta\beta$	not closed (unless zero); the orthogonal complement of the closed forms	transverse, propagating photons
harmonic ω	$H^k_{\mathrm{dR}}(M)$; Betti numbers b_k; Hodge numbers $h_{p,q}$	zero modes: flux sectors, Wilson lines, moduli, supersymmetric ground states

The mathematician uses (5.67) to *compute* cohomology: solve $\Delta\omega = 0$ and count. The physicist uses the same splitting to *separate* the degrees of freedom. Gauge fixing is a projection. Write $A = \mathrm{d}\lambda + \delta\beta + h$; then $\delta A = \delta\mathrm{d}\lambda = \Delta\lambda$, and on a compact connected Euclidean space $\Delta\lambda = 0$ forces λ to be constant. So the Lorenz condition $\delta A = 0$ is exactly "the exact part of A is zero": it throws away the pure-gauge sector and keeps the transverse and topological ones.

This is also how Nash uses harmonic forms. For him they are the kernels of elliptic operators (Chapter 2), and an index is a difference of numbers of zero modes. The operator $\mathrm{d}+\delta$ from even to odd forms has index $\sum_k (-1)^k b_k = \chi(M)$, because its zero modes are the harmonic forms (Chapter 4). Riemann–Roch counts the zero modes of $\bar\partial_L$ in the same way (Section 5.3.3), and the lowest Landau level of Example 5.59 is such a count. The harmonic sector that a mass removes is precisely the sector an index theorem counts.

Recap: Hodge decomposition in Maxwell and Proca theory

- $A^k = \mathrm{d}A^{k-1} \oplus \delta A^{k+1} \oplus \mathcal{H}^k$ orthogonally, with $\mathcal{H}^k = \operatorname{Ker}\Delta \cong H^k_{\mathrm{dR}}$; closed means no co-exact part, co-closed means no exact part.
- Maxwell: $\mathrm{d}F = 0$ is kinematic ($\mathrm{d}^2 = 0$), $\delta F = 0$ is dynamical, and together $\Delta F = 0$. Source-free fields on a compact space are the harmonic 2-forms: b_2 flux sectors. Flux through a closed surface sees only the harmonic part, and charged matter quantises it. Example: $B\,\mathrm{d}x \wedge \mathrm{d}y$ on T^2, with $eBL_1L_2 \in 2\pi\mathbb{Z}$.
- Proca: $\mathrm{d}\delta F + m^2 F = 0$ splits into $(\Delta + m^2)\mathrm{d}\alpha = 0$, $m^2\delta\beta = 0$, $m^2\omega = 0$. For $m \neq 0$, $F = \mathrm{d}(-m^{-2}\delta F)$ is exact, $\delta A = 0$ is automatic, and $(\Delta + m^2)A = 0$ is Klein–Gordon.
- A mass gaps out the harmonic zero modes: a massive photon carries no topological flux (the Meissner effect).

The Kähler refinement

On a general Hermitian manifold there is no particular relation between these Laplacians. On a Kähler manifold there is (Nash's (5.32)):

Key idea: Kähler identity for Laplacians

$$\Delta_d = 2\Delta_\partial = 2\Delta_{\bar\partial} \qquad \text{on a Kähler manifold.} \tag{5.71}$$

Physically: the zero modes of the full Laplacian Δ_d (the topological sector of Figure 5.36) and the zero modes of $\Delta_{\bar\partial}$ (the holomorphic data, such as lowest-Landau-level states and holomorphic differentials) are the *same* forms. On a Kähler manifold, topology and holomorphy count the same things.

Worked example 5.61: $\Delta_d = 2\Delta_{\bar\partial}$ on functions on $\mathbb{C}$

Take flat $\mathbb{C}$ with $\mathrm{d}s^2 = |\mathrm{d}z|^2$ and compactly supported functions. We need $\bar\partial^*$ on $(0,1)$-forms. The pointwise norm of $\mathrm{d}\bar z = \mathrm{d}x - i\mathrm{d}y$ is $|\mathrm{d}x|^2 + |\mathrm{d}y|^2 = 2$. For a function f and a $(0,1)$-form $\beta = b\,\mathrm{d}\bar z$,

$$\begin{aligned}\langle \bar\partial f, \beta\rangle &\overset{(1)}{=} \int \partial_{\bar z} f\ \bar b\ |\mathrm{d}\bar z|^2\,\mathrm{d}A \overset{(2)}{=} 2\int \partial_{\bar z} f\ \bar b\,\mathrm{d}A \\ &\overset{(3)}{=} -2\int f\ \partial_{\bar z}\bar b\,\mathrm{d}A \overset{(4)}{=} \int f\ \overline{(-2\partial_z b)}\,\mathrm{d}A = \langle f, -2\partial_z b\rangle,\end{aligned}$$

(1) $\bar\partial f = \partial_{\bar z} f\,\mathrm{d}\bar z$, and the inner product of two multiples of $\mathrm{d}\bar z$ is the product of coefficients (one conjugated) times $|\mathrm{d}\bar z|^2$.
(2) $|\mathrm{d}\bar z|^2 = 2$.
(3) Integrate by parts; no boundary term for compact support.

(4) $\partial_{\bar z}\bar b = \overline{\partial_z b}$.

so $\bar\partial^*(b\,\mathrm{d}\bar z) = -2\partial_z b$. Therefore

$$\Delta_{\bar\partial} f \overset{(1)}{=} \bar\partial^*\bar\partial f \overset{(2)}{=} -2\,\partial_z\partial_{\bar z} f \overset{(3)}{=} -\tfrac{1}{2}\big(\partial_x^2+\partial_y^2\big) f \overset{(4)}{=} \tfrac{1}{2}\,\Delta_d f\,,$$

(1) On functions $\bar\partial^* f = 0$ (there are no $(0,-1)$-forms).
(2) The adjoint just computed, with $b = \partial_{\bar z} f$.
(3) $\partial_z\partial_{\bar z} = \frac{1}{4}(\partial_x^2+\partial_y^2)$.
(4) $\Delta_d = \mathrm{d}^*\mathrm{d} = -(\partial_x^2+\partial_y^2)$ on functions (the positive Laplacian).

The same computation with ∂ gives $\Delta_\partial f = -2\partial_{\bar z}\partial_z f$, the same thing. So $\Delta_d = 2\Delta_\partial = 2\Delta_{\bar\partial}$ on flat $\mathbb{C}$. On flat $\mathbb{C}^n$ and on forms of any type the calculation is the same in each coordinate direction.

Why it holds on every Kähler manifold. The standard proof goes through the *Kähler identities*, such as $[\Lambda,\bar\partial] = -\mathrm{i}\,\partial^*$, where Λ is the adjoint of wedging with ω. These are first-order identities: they involve the metric and only its first derivatives. They hold on flat $\mathbb{C}^n$ by a computation like the one above, hence on every Kähler manifold by the principle in the key idea on Kähler normal coordinates. The relation (5.71) then follows from them by algebra (one also uses $\partial\bar\partial^* + \bar\partial^*\partial = 0$, another consequence). Without the Kähler condition the first derivatives of g spoil the identities, and Δ_d, Δ_∂, $\Delta_{\bar\partial}$ differ by first-order terms.

The Hodge decomposition. Now combine. Let α be a Δ_d-harmonic k-form and split it by type, $\alpha = \sum_{p+q=k}\alpha^{p,q}$. Then

$$0 = \Delta_d\alpha = 2\Delta_{\bar\partial}\alpha = 2\sum_{p+q=k}\Delta_{\bar\partial}\alpha^{p,q},$$

and since $\Delta_{\bar\partial}$ preserves type, the terms on the right have different types and must vanish separately. So every component $\alpha^{p,q}$ of a harmonic form is itself harmonic (for all three Laplacians). With the **Hodge numbers** (Nash's (5.33))

$$h_{p,q} = \dim H^{p,q}_{\bar\partial}(M) \tag{5.72}$$

we get the **Hodge decomposition** (Nash's (5.34))

$$H^k(M;\mathbb{C}) = \bigoplus_{p+q=k} H^{p,q}(M;\mathbb{C}), \tag{5.73}$$

where $H^{p,q}(M;\mathbb{C})$ is the space of classes represented by a closed form of type (p,q), and $H^{p,q}(M;\mathbb{C}) \cong H^{p,q}_{\bar\partial}(M)$ because both are the space of harmonic (p,q)-forms. In the language of Section 5.3.9: each topological zero mode carries a definite type (p,q). For example, on the torus the flux sector $B\,\mathrm{d}x\wedge\mathrm{d}y = \frac{\mathrm{i}}{2}B\,\mathrm{d}z\wedge\mathrm{d}\bar z$ of Example 5.59 is pure type $(1,1)$, while the two Wilson-line zero modes $\mathrm{d}x, \mathrm{d}y$ recombine into $\mathrm{d}z$ (type $(1,0)$) and $\mathrm{d}\bar z$ (type $(0,1)$).

Consequences. Nash lists five. Each is now a one-line argument.

(i) $b_k = \sum_{p+q=k} h_{p,q}$: take dimensions in (5.73).

(ii) $b_2 > 0$ (for M compact Kähler): Example 5.55, or (v) with $p = 1$.

(iii) b_{2i+1} is even: by (i) and (iv),

$$b_{2i+1} = \sum_{p+q=2i+1} h_{p,q} = \sum_{p<q} \big(h_{p,q} + h_{q,p}\big) = 2 \sum_{p<q,\ p+q=2i+1} h_{p,q},$$

because $p+q$ odd means $p \neq q$, so the terms pair off.

(iv) $h_{p,q} = h_{q,p}$: Δ_d is a real operator, so the complex conjugate of a harmonic form is harmonic, and conjugation turns dz's into $d\bar{z}$'s, i.e. type (p,q) into (q,p). It is a real-linear bijection between the two spaces of harmonic forms.

(v) $h_{p,p} > 0$ for $0 \le p \le n$: the form ω^p has type (p,p); it is closed; and it is co-closed, because $\star\,\omega^p = \frac{p!}{(n-p)!}\,\omega^{n-p}$ (a pointwise identity, checked in a unitary frame as in (5.62)) and so $d^*\omega^p = \pm \star d \star \omega^p \propto \star d\omega^{n-p} = 0$. So ω^p is a nonzero harmonic (p,p)-form, and $h_{p,p} \ge 1$. For $p = 1$ this says that a compact Kähler manifold always admits a nonzero source-free magnetic field with topological flux: the Kähler form itself, the "uniform" field of the Landau problem.

In addition, Serre duality (5.40) gives $h_{p,q} = h_{n-p,n-q}$. So the Hodge numbers, arranged in a diamond, have a left–right symmetry (conjugation) and an up–down symmetry (Serre duality), and the Betti numbers are the sums along the horizontal rows.

Worked example 5.62: Hodge numbers of a compact Riemann surface, and the genus

For Σ_p ($n = 1$): $h_{0,0} = 1$ (constants), $h_{1,1} = h_{0,0} = 1$ (Serre duality; the class of ω), and $h_{1,0} = \dim H^0(K)$, the number of holomorphic 1-forms. By (iv), $h_{0,1} = h_{1,0}$, and by (i),

$$2p \overset{(1)}{=} b_1(\Sigma_p) \overset{(2)}{=} h_{1,0} + h_{0,1} \overset{(3)}{=} 2h_{1,0} \quad \Longrightarrow \quad h^0(K) = h_{1,0} = p.$$

(1) A closed orientable surface of genus p has $H_1 = \mathbb{Z}^{2p}$ (Section 5.2.15).
(2) Hodge decomposition (5.73) with $k = 1$.
(3) $h_{0,1} = h_{1,0}$ by complex conjugation.

This is the fact used in Step 1 of the proof of Riemann–Roch (Section 5.3.3): *a Riemann surface of topological genus p has exactly p independent holomorphic differentials*, and $h^1(\mathcal{O}) = h_{0,1} = p$. The Hodge diamond is

$$\begin{array}{ccc} & 1 & \\ p & & p \\ & 1 & \end{array} \qquad \big(\text{rows: } h_{1,1};\ h_{1,0},\ h_{0,1};\ h_{0,0}\big).$$

Worked example 5.63: Complex tori and projective spaces

Tori. On a flat torus $T = \mathbb{C}^n/\Lambda$, harmonic forms are exactly the forms with constant coefficients: the Laplacian is $-\sum(\partial^2_{x^k} + \partial^2_{y^k})$ on each coefficient, and on a Fourier mode $e^{2\pi i\langle \kappa, x\rangle}$ it is multiplication by $4\pi^2|\kappa|^2$, which vanishes only for $\kappa = 0$ (as in Example 5.42). A constant (p,q)-form is a combination of $dz^{i_1} \wedge \cdots \wedge dz^{i_p} \wedge d\bar{z}^{j_1} \wedge \cdots \wedge d\bar{z}^{j_q}$ with $i_1 < \cdots < i_p$, $j_1 < \cdots < j_q$, so

$$h_{p,q}(T) = \binom{n}{p}\binom{n}{q}, \qquad \sum_{p+q=k} \binom{n}{p}\binom{n}{q} = \binom{2n}{k} = b_k(T^{2n}),$$

the second being Vandermonde's identity: consistent with (i), since the real torus T^{2n} has $b_k = \binom{2n}{k}$.

Projective space. $\mathbb{CP}^n$ is a union of cells $\mathbb{C}^n \cup \mathbb{C}^{n-1} \cup \cdots \cup \mathbb{C}^0$ (the points with $Z_0 \neq 0$, then those with $Z_0 = 0 \neq Z_1$, and so on), one cell in each even real dimension, so $b_{2k} = 1$ for $0 \le k \le n$ and all odd Betti numbers vanish. Since $h_{k,k} \ge 1$ by (v) and $b_{2k} = \sum_{p+q=2k} h_{p,q} = 1$, we get $h_{k,k} = 1$ and every other Hodge number is zero. In particular $h_{0,1} = h_{0,2} = 0$, which is what Example 5.52 needed to conclude $\mathrm{Pic}(\mathbb{CP}^n) = \mathbb{Z}$.

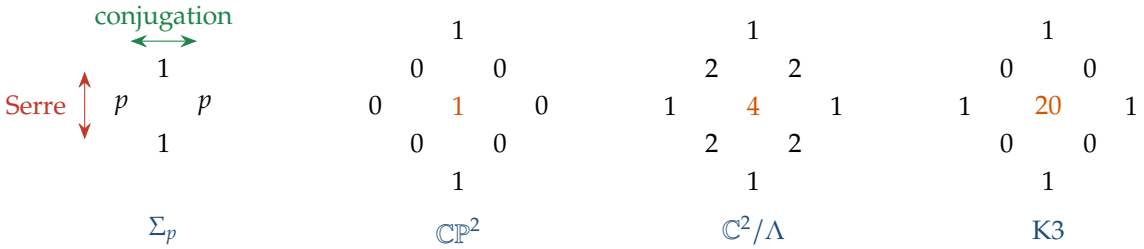


Figure 5.38: Hodge diamonds $h_{p,q}$, with $h_{0,0}$ at the bottom, $h_{n,n}$ at the top, $h_{p,0}$ on the lower-left edge and $h_{0,q}$ on the lower-right edge (for Σ_p the middle row is $h_{1,0}, h_{0,1}$). Row sums are Betti numbers: Σ_p has $b = (1, 2p, 1)$; $\mathbb{CP}^2$ has $(1, 0, 1, 0, 1)$; the complex 2-torus $(1, 4, 6, 4, 1)$; the K3 surface $(1, 0, 22, 0, 1)$. Every diamond is symmetric under left–right reflection (complex conjugation, $h_{p,q} = h_{q,p}$) and under up–down reflection (Serre duality, $h_{p,q} = h_{n-p,n-q}$), and the central column (orange) is never zero (the powers of ω).

Worked example 5.64: A complex manifold that is not Kähler: the Hopf surface

Let $M = (\mathbb{C}^2 \setminus \{0\})/\langle Z \mapsto 2Z \rangle$: identify each nonzero vector Z with $2Z$. The map $Z \mapsto 2Z$ is holomorphic, and the group it generates acts freely and properly, so M is a compact complex surface (complex dimension 2). Topologically, write $Z = e^s u$ with $s \in \mathbb{R}$ and $u \in S^3$; the identification is $s \sim s + \log 2$, so

$$M \cong S^1 \times S^3, \qquad b_0 = b_1 = b_3 = b_4 = 1, \quad b_2 = 0.$$

Both $b_1 = 1$ (odd) and $b_2 = 0$ violate Nash's consequences (iii) and (ii). So *the Hopf surface admits no Kähler metric at all*, and hence (by Section 5.3.8) it cannot be embedded in any $\mathbb{CP}^N$: it is a complex manifold that is not algebraic. For a Hermitian metric on it, such as $\mathrm{d}s^2 = |\mathrm{d}Z|^2/|Z|^2$ (invariant under $Z \mapsto 2Z$), one can check directly that $\mathrm{d}\omega \neq 0$.

Physical picture: Hodge numbers as counts of massless fields

In supersymmetric quantum mechanics on M (Chapter 7), the supercharge is $Q = \mathrm{d} + \mathrm{d}^*$ and the ground states are the harmonic forms, so b_k counts ground states of fermion number k. On a Kähler manifold there are *two* supercharges, $\bar\partial + \bar\partial^*$ and $\partial + \partial^*$, and (5.71) says they generate the same Hamiltonian: the extra supersymmetry refines fermion number into the pair (p, q), and the ground states are counted by $h_{p,q}$. In string compactification on a Calabi–Yau threefold, $h_{1,1}$ and $h_{2,1}$ count massless moduli fields, and the number of quark–lepton generations is $|h_{1,1} - h_{2,1}| = |\chi|/2$. Supersymmetric sigma models with $N = 2$ supersymmetry require a Kähler target, for the same algebraic reason (Zumino's theorem).

Reading Nash (pp. 146–148, (5.29)–(5.34))

Nash defines a Kähler manifold by a Hermitian metric $g_{ij}\mathrm{d}z^i \otimes \mathrm{d}\bar{z}^j$ whose 2-form $\omega = \frac{\mathrm{i}}{2} g_{ij}\mathrm{d}z^i \wedge \mathrm{d}\bar{z}^j$ is closed (5.29); notes $\omega(X,Y) = g(JX,Y)$ and that Kähler manifolds are symplectic; gives the examples of flat tori, Riemann surfaces, and $\mathbb{CP}^n$ with the Fubini–Study metric (5.30) (whose typos we corrected); remarks that Kähler metrics are characterised by $g_{ij} = \delta_{ij} + h_{ij}$ with h of second order; observes that complex submanifolds of Kähler manifolds are Kähler, so projective manifolds are Kähler; introduces the three Laplacians (5.31), states $\Delta_d = 2\Delta_\partial = 2\Delta_{\bar\partial}$ (5.32), the Hodge numbers (5.33) and decomposition (5.34), and lists five consequences (i)–(v) without proof. We added: the matrix J and the compatible triple, the computation $h = g - \mathrm{i}\omega$, the component form of $\mathrm{d}\omega = 0$, the Kähler potential with its derivation, the full construction of Kähler normal coordinates, $\omega^n = n!\,\mathrm{vol}$ and non-exactness of ω^k, the hyperbolic metric, the explicit Fubini–Study matrix with its eigenvalues, overlap consistency and $\mathrm{U}(n+1)$ invariance, the identification of $\mathbb{CP}^1$ with a sphere of radius $\frac{1}{2}$ with curvature and Gauss–Bonnet checks, $\omega_{FS} = \pi c_1(\mathcal{O}(1))$, the area $= \pi\times$degree formula, the Laplacian identity on $\mathbb{C}$ computed, the Hodge decomposition applied to Maxwell and Proca theory (flux sectors, dynamical exactness, the mass gap), proofs of (i)–(v), Hodge diamonds, and the Hopf surface.

Recap: Section 5.3

- Forms split by type, $\mathrm{d} = \partial + \bar\partial$, and holomorphic cohomology is the cohomology of $\bar\partial$: $H^q(M;\Omega^p(E)) \cong H^{p,q}_{\bar\partial_E}(M)$ (Dolbeault).
- Serre duality: $H^q(M;\Omega^p(E)) \cong H^{n-q}(M;\Omega^{n-p}(E^*))^*$, via $\int_M \theta\wedge\nu$; on a curve $h^1(L) = h^0(K\otimes L^{-1})$.
- Riemann–Roch: $h^0(D) - h^0(K-D) = \deg D + 1 - p$. It gives $h^0(K) = p$, $\deg K = 2p-2$, $h^0(K^2) = 3p-3$, and $h^0(D) = \deg D + 1 - p$ for $\deg D > 2p-2$; it is the index theorem for $\bar\partial_L$ and the lowest-Landau-level count.
- Line bundles are classes in $H^1(M;\mathcal{E}^*)$ (smooth) or $H^1(M;\mathcal{O}^*)$ (holomorphic). The exponential sequence gives c_1 as a connecting map; smooth line bundles are classified by $c_1 \in H^2(M;\mathbb{Z})$; holomorphic ones by c_1 and a point of $\mathrm{Pic}_0 = H^1(\mathcal{O})/H^1(\mathbb{Z})$, which for a curve is the Jacobian.
- Kähler: Hermitian metric with $\mathrm{d}\omega = 0$, equivalently $g_{i\bar\jmath} = \partial_i\partial_{\bar\jmath}K$ locally, equivalently flat to first order in holomorphic coordinates. Examples: tori, all Riemann surfaces, $\mathbb{CP}^n$ with Fubini–Study ($\omega = \frac{\mathrm{i}}{2}\partial\bar\partial\log(1+|z|^2)$; on $\mathbb{CP}^1$ the sphere of radius $\frac{1}{2}$), and all projective varieties.
- Hodge decomposition $A^k = \mathrm{d}A^{k-1} \oplus \delta A^{k+1} \oplus \mathcal{H}^k$ in physics: source-free Maxwell fields are harmonic and carry topological flux (quantised, as for $B\,\mathrm{d}x\wedge\mathrm{d}y$ on T^2); the Proca equation forces F to be exact and kills the harmonic sector, so a massive photon has no flux sectors.
- On compact Kähler manifolds $\Delta_d = 2\Delta_{\bar\partial}$, so $H^k = \bigoplus_{p+q=k} H^{p,q}$, $h_{p,q} = h_{q,p} = h_{n-p,n-q}$, b_{odd} even, $h_{p,p} \geq 1$. The Hopf surface $S^1\times S^3$ is complex but not Kähler.

5.4 The Teichmüller space $\mathcal{T}_p$

5.4.1 Why a physicist needs the space of shapes

Here is the question that drives the rest of this chapter. A closed string moving through space-time sweeps out a closed surface Σ, its **worldsheet**. In Polyakov's path integral the worldsheet carries its own metric g_{ab}, and the partition function is, schematically,

$$Z \;=\; \sum_{p=0}^{\infty} \int \mathcal{D}g\, \mathcal{D}\phi \;\exp\!\Big(-\tfrac{1}{2}\int_{\Sigma} \mathrm{d}^2x\, \sqrt{g}\, g^{ab}\, \partial_a \phi^{\mu}\, \partial_b \phi_{\mu}\Big). \tag{5.74}$$

The sum is over the genus p of the surface (the number of handles, i.e. the number of string loops). For each genus we integrate over all metrics g on Σ and over all embeddings $\phi : \Sigma \to \mathbb{R}^d$. Chapter 9 (Chapter 9) does this properly. Here we only ask the preliminary question:

How many genuinely different worldsheets of genus p are there?

The action does not see every detail of g. It is unchanged if we rescale g by a positive function (Weyl invariance, checked in Section 5.4.3 below). It is also unchanged if we move the points of Σ around by a diffeomorphism, because that is just a relabelling of the worldsheet. What survives once we throw away these two redundancies is a *finite-dimensional* space of "shapes". For the torus it is a single complex number τ. For genus $p \geq 2$ it has $3p-3$ complex dimensions. So the enormous integral $\int \mathcal{D}g$ collapses to an ordinary integral over a finite-dimensional space. That space is the **moduli space** $\mathcal{M}_p$, and its simply connected cover is the **Teichmüller space** $\mathcal{T}_p$. This section builds $\mathcal{T}_p$. Section 5.5 builds $\mathcal{M}_p$, Section 5.6 counts its dimension with the index theorem, and Section 5.7 looks at a finer invariant of each individual surface, its Weierstrass points.

Watch out: Notation for the genus

Nash writes the genus as p and a genus-p surface as Σ_p (or just Σ). Most modern books write g for the genus, which clashes with the metric g. We follow Nash: p **is always the genus** in this part, and g is always a metric.

Figure 5.39 is the map of this section and the next. Each arrow is a quotient by a group of redundancies.

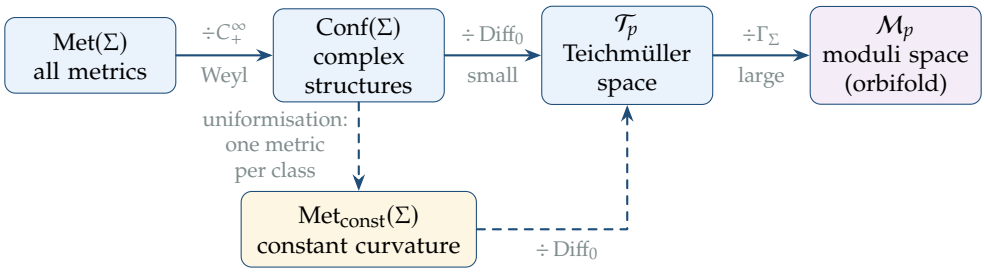


Figure 5.39: The chain of quotients. Reading left to right we throw away Weyl rescalings, then diffeomorphisms that can be deformed to the identity, then the remaining "large" diffeomorphisms. The dashed route is Nash's second definition (5.43): first use uniformisation to pick a canonical metric in each conformal class, then divide by diffeomorphisms.

5.4.2 From a metric to a complex structure

We start with a surface Σ and a Riemannian metric on it. In local coordinates (x, y) the most general metric is

$$\mathrm{d}s^2 \;=\; A\,\mathrm{d}x^2 + 2B\,\mathrm{d}x\,\mathrm{d}y + C\,\mathrm{d}y^2, \qquad A > 0, \quad AC - B^2 > 0. \tag{5.75}$$

The two inequalities say that the matrix $\left(\begin{smallmatrix} A & B \\ B & C \end{smallmatrix}\right)$ is positive definite, so lengths are positive. Nash's first claim [Nash p. 149, (5.35)] is that one can always rewrite this in terms of $z = x + iy$ and $\bar z = x - iy$ as

$$\mathrm{d}s^2 \;=\; \lambda\,(\mathrm{d}z + \mu\,\mathrm{d}\bar z)(\mathrm{d}\bar z + \bar\mu\,\mathrm{d}z) \;=\; \lambda\,|\mathrm{d}z + \mu\,\mathrm{d}\bar z|^2\,, \qquad \lambda > 0,\ |\mu| < 1, \tag{5.76}$$

with λ a positive real function and μ a complex function. Nash says "it is always possible". Let us actually do it, because the explicit formulas are useful.

Expanding the right-hand side

We need three products of differentials. With $\mathrm{d}z = \mathrm{d}x + i\,\mathrm{d}y$:

$$\begin{aligned} \mathrm{d}z\,\mathrm{d}\bar z &= (\mathrm{d}x + i\mathrm{d}y)(\mathrm{d}x - i\mathrm{d}y) = \mathrm{d}x^2 + \mathrm{d}y^2, \\ \mathrm{d}z^2 &= \mathrm{d}x^2 - \mathrm{d}y^2 + 2i\,\mathrm{d}x\,\mathrm{d}y, \\ \mathrm{d}\bar z^2 &= \mathrm{d}x^2 - \mathrm{d}y^2 - 2i\,\mathrm{d}x\,\mathrm{d}y. \end{aligned}$$

(These are *symmetric* products of 1-forms, the kind that make up a metric, not wedge products.) Write $\mu = a + ib$ with a, b real. Then

$$\begin{aligned} |\mathrm{d}z + \mu\,\mathrm{d}\bar z|^2 &\overset{(1)}{=} (1 + |\mu|^2)\,\mathrm{d}z\,\mathrm{d}\bar z + \bar\mu\,\mathrm{d}z^2 + \mu\,\mathrm{d}\bar z^2 \\ &\overset{(2)}{=} (1 + |\mu|^2)(\mathrm{d}x^2 + \mathrm{d}y^2) + 2\,\mathrm{Re}\big[(a - ib)(\mathrm{d}x^2 - \mathrm{d}y^2 + 2i\,\mathrm{d}x\,\mathrm{d}y)\big] \\ &\overset{(3)}{=} (1 + |\mu|^2 + 2a)\,\mathrm{d}x^2 + 4b\,\mathrm{d}x\,\mathrm{d}y + (1 + |\mu|^2 - 2a)\,\mathrm{d}y^2. \end{aligned} \tag{5.77}$$

(1) Multiply out $(\mathrm{d}z + \mu\mathrm{d}\bar z)(\mathrm{d}\bar z + \bar\mu\mathrm{d}z)$; the cross terms are $\bar\mu\,\mathrm{d}z^2$ and $\mu\,\mathrm{d}\bar z^2$, and the diagonal terms give $(1 + \mu\bar\mu)\mathrm{d}z\mathrm{d}\bar z$.

(2) Insert the three products above; the two cross terms are complex conjugates of each other, so together they are twice the real part of one of them.

(3) $\mathrm{Re}[(a - ib)(X + 2iY)] = aX + 2bY$ with $X = \mathrm{d}x^2 - \mathrm{d}y^2$, $Y = \mathrm{d}x\,\mathrm{d}y$; collect the coefficients of $\mathrm{d}x^2$, $\mathrm{d}x\,\mathrm{d}y$ and $\mathrm{d}y^2$.

Solving for λ and μ

Comparing (5.77) (times λ) with (5.75) gives three real equations:

$$A = \lambda(1 + |\mu|^2 + 2a), \qquad B = 2\lambda b, \qquad C = \lambda(1 + |\mu|^2 - 2a).$$

Three real unknowns (λ, a, b), three equations. Form the combinations

$$\begin{aligned} A - C + 2iB &\overset{(1)}{=} 4\lambda a + 4i\lambda b = 4\lambda\mu, \\ A + C &\overset{(2)}{=} 2\lambda(1 + |\mu|^2), \\ AC - B^2 &\overset{(3)}{=} \lambda^2\big[(1 + |\mu|^2)^2 - 4a^2 - 4b^2\big] \overset{(4)}{=} \lambda^2\,(1 - |\mu|^2)^2. \end{aligned} \tag{5.78}$$

(1) Subtract the first and third equations, and add $2i$ times the second.
(2) Add the first and third equations.
(3) $(P+Q)(P-Q) = P^2 - Q^2$ with $P = \lambda(1+|\mu|^2)$ and $Q = 2\lambda a$, then subtract $B^2 = 4\lambda^2 b^2$.
(4) $4a^2 + 4b^2 = 4|\mu|^2$ and $(1+|\mu|^2)^2 - 4|\mu|^2 = (1-|\mu|^2)^2$.

Write $\Delta := AC - B^2 > 0$. Taking the positive square root (we want $|\mu| < 1$) gives $\lambda(1-|\mu|^2) = \sqrt{\Delta}$. Adding this to half of the second line gives $2\lambda = \frac{1}{2}(A+C) + \sqrt{\Delta}$. So

$$\boxed{\lambda = \frac{A + C + 2\sqrt{AC - B^2}}{4}, \qquad \mu = \frac{A - C + 2iB}{A + C + 2\sqrt{AC - B^2}}.} \tag{5.79}$$

Is $|\mu| < 1$ automatic? Yes:

$$\begin{aligned} |A - C + 2iB|^2 &\overset{(1)}{=} (A+C)^2 - 4(AC - B^2) \\ &\overset{(2)}{=} \big(A + C - 2\sqrt{\Delta}\big)\big(A + C + 2\sqrt{\Delta}\big) \overset{(3)}{<} \big(A + C + 2\sqrt{\Delta}\big)^2. \end{aligned}$$

(1) $(A-C)^2 + 4B^2 = (A+C)^2 - 4AC + 4B^2$.
(2) Difference of two squares.
(3) $\sqrt{\Delta} > 0$, so the first factor is smaller than the second.

Dividing by $(A + C + 2\sqrt{\Delta})^2$ gives $|\mu|^2 < 1$. The condition $|\mu| < 1$ is exactly the positivity of the metric.

Working definition 5.37 (Beltrami coefficient of a metric)

Given a metric $A\,\mathrm{d}x^2 + 2B\,\mathrm{d}x\mathrm{d}y + C\,\mathrm{d}y^2$ in local coordinates, its **Beltrami coefficient** is the complex function μ of (5.79) [the metric is $\lambda\,|\mathrm{d}z + \mu\,\mathrm{d}\bar{z}|^2$ with $z = x + \mathrm{i}y$, so μ measures how far it is from round], and its **conformal factor** is λ. *Test:* $\mu \equiv 0$ exactly when $A = C$ and $B = 0$, that is, when the metric is already a multiple of $\mathrm{d}x^2 + \mathrm{d}y^2$. Always $|\mu| < 1$.

Worked example 5.65: A skew constant metric

Take $\mathrm{d}s^2 = \mathrm{d}x^2 + \mathrm{d}x\,\mathrm{d}y + \mathrm{d}y^2$, so $A = C = 1$, $B = \frac{1}{2}$. Then $\Delta = 1 - \frac{1}{4} = \frac{3}{4}$ and $\sqrt{\Delta} = \frac{\sqrt{3}}{2}$, and

$$\lambda \overset{(1)}{=} \frac{2+\sqrt{3}}{4}, \qquad \mu \overset{(2)}{=} \frac{0+i}{2+\sqrt{3}} \overset{(3)}{=} i\,(2-\sqrt{3}) \approx 0.268\,i.$$

(1) Formula (5.79) with $A + C = 2$.
(2) Same formula with $A - C = 0$, $2B = 1$.
(3) Rationalise: $(2+\sqrt{3})(2-\sqrt{3}) = 1$.

Check by expanding back: $1+|\mu|^2 = 1+(2-\sqrt{3})^2 = 8-4\sqrt{3}$ and $\lambda(8-4\sqrt{3}) = (2+\sqrt{3})(2-\sqrt{3}) = 1 = A$, with $a = 0$; and $2\lambda b = \frac{2+\sqrt{3}}{2}(2-\sqrt{3}) = \frac{1}{2} = B$. Good.

What μ looks like: circles become ellipses

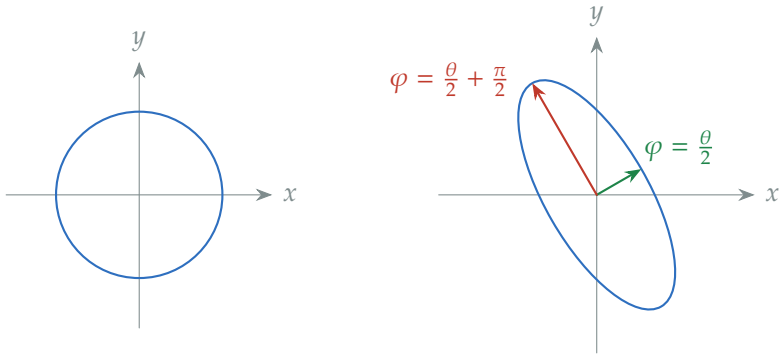
Fix a point and look at the set of tangent vectors of unit length in the metric. If $\mu = 0$, it is a round circle in the (x, y) plane. If $\mu \neq 0$, the condition $\lambda|\mathrm{d}z + \mu\mathrm{d}\bar{z}|^2 = 1$ is an ellipse. Write $\mu = |\mu|e^{i\theta}$ and a tangent direction as $\mathrm{d}z = re^{i\varphi}$. Then

$$|\mathrm{d}z + \mu\,\mathrm{d}\bar{z}| = r\left|e^{i\varphi} + |\mu|e^{i(\theta-\varphi)}\right| = r\left|1 + |\mu|e^{i(\theta-2\varphi)}\right|.$$

This is largest, $r(1+|\mu|)$, when $\varphi = \theta/2$, and smallest, $r(1-|\mu|)$, when $\varphi = \theta/2 + \pi/2$. So a vector of unit length is *short* in the $\theta/2$ direction and *long* in the perpendicular direction. The ellipse has

$$\frac{\text{major axis}}{\text{minor axis}} = \frac{1+|\mu|}{1-|\mu|}, \tag{5.80}$$

a number called the **dilatation**. Figure 5.40 shows this. The coefficient μ is a field of infinitesimal ellipses on Σ: at each point it tells us how the metric's circles are squashed and in which direction. The overall size of the ellipse is λ, and it does not matter for angles.



$\mu = 0$: unit circle is round

$\mu = |\mu|e^{i\theta}$: unit "circle" is an ellipse, axis ratio $(1+|\mu|)/(1-|\mu|)$

Figure 5.40: The Beltrami coefficient as a field of ellipses. Left: a conformally flat metric; the vectors of unit length form a round circle. Right: with $\mu \neq 0$ the unit vectors form an ellipse, short along the direction $\theta/2$ (green) and long perpendicular to it (red). Rescaling the metric by λ changes the size of the ellipse but not its shape; only μ matters for angles.

Why a metric gives a complex structure

A **complex structure** on Σ is an atlas of complex coordinates w_α whose changes of coordinates $w_\beta(w_\alpha)$ are holomorphic. Here is how a metric produces one. Suppose we can find, near each point, a function $w(z,\bar z)$ with

$$\frac{\partial w}{\partial \bar z} = \mu \frac{\partial w}{\partial z}, \qquad \frac{\partial w}{\partial z} \neq 0. \tag{5.81}$$

This is the **Beltrami equation**. Then

$$dw = w_z\,dz + w_{\bar z}\,d\bar z \overset{(1)}{=} w_z\,(dz + \mu\,d\bar z), \qquad \text{so} \qquad ds^2 \overset{(2)}{=} \frac{\lambda}{|w_z|^2}\,|dw|^2.$$

(1) Use (5.81) to replace $w_{\bar z}$ by μw_z.
(2) Take $|\cdot|^2$ and compare with (5.76).

In the new coordinate the metric is a positive function times the flat metric $|dw|^2$. Coordinates with this property are called **isothermal** (or conformal) coordinates. That solutions of (5.81) exist locally, for any smooth μ with $|\mu| < 1$, is a classical theorem of Gauss (for real-analytic metrics), Korn and Lichtenstein, and in full generality Ahlfors and Bers. We take it as given.

Now take two overlapping isothermal coordinates w_1 and w_2. On the overlap $ds^2 = \rho_1|dw_1|^2 = \rho_2|dw_2|^2$ with $\rho_1, \rho_2 > 0$. Then

$$\begin{aligned}\rho_2\,|dw_2|^2 &\overset{(1)}{=} \rho_2\,\left|\partial w_2\,dw_1 + \bar\partial w_2\,d\bar w_1\right|^2 \\ &\overset{(2)}{=} \rho_2\Big[(|\partial w_2|^2 + |\bar\partial w_2|^2)\,|dw_1|^2 + 2\,\mathrm{Re}\big(\overline{\partial w_2}\,\bar\partial w_2\,d\bar w_1^{\,2}\big)\Big].\end{aligned}$$

(1) Chain rule, with $\partial = \partial/\partial w_1$ and $\bar\partial = \partial/\partial \bar w_1$.
(2) Multiply out as in (5.77).

For this to equal $\rho_1 |\mathrm{d}w_1|^2$, the $\mathrm{d}\bar w_1^2$ term must vanish, so $\partial w_2 \cdot \bar\partial w_2 = 0$ everywhere. Hence (on a connected overlap) either $\bar\partial w_2 = 0$, so w_2 is a holomorphic function of w_1, or $\partial w_2 = 0$, so it is antiholomorphic. The antiholomorphic case reverses orientation. If we insist that all coordinates are compatible with a fixed orientation of Σ, the coordinate changes are holomorphic. So:

Key idea: A metric plus an orientation is a complex structure

Isothermal coordinates of a metric, all positively oriented, form a holomorphic atlas. In other words, a metric on an oriented surface determines a complex structure, and so turns Σ into a Riemann surface. The complex structure is encoded locally by the Beltrami coefficient μ.

Worked example 5.66: Flat tori and the Cayley transform

Fix a complex number $\tau = \tau_1 + i\tau_2$ with $\tau_2 > 0$. On the square torus $\mathbb{R}^2/\mathbb{Z}^2$, with coordinates (x, y) taken mod 1, put the metric

$$\mathrm{d}s^2 = |\mathrm{d}x + \tau\,\mathrm{d}y|^2 = \mathrm{d}x^2 + 2\tau_1\,\mathrm{d}x\,\mathrm{d}y + |\tau|^2\,\mathrm{d}y^2. \tag{5.82}$$

Geometrically this is the flat torus $\mathbb{C}/(\mathbb{Z} + \tau\mathbb{Z})$ pulled back along the linear map $(x, y) \mapsto w = x + \tau y$, which sends the unit square to the parallelogram with sides 1 and τ. Read off $A = 1$, $B = \tau_1$, $C = |\tau|^2$, so $\Delta = |\tau|^2 - \tau_1^2 = \tau_2^2$. Then

$$\lambda \overset{(1)}{=} \frac{1 + |\tau|^2 + 2\tau_2}{4} \overset{(2)}{=} \frac{|\tau + i|^2}{4},$$
$$\mu \overset{(3)}{=} \frac{1 - |\tau|^2 + 2i\tau_1}{|\tau + i|^2} \overset{(4)}{=} \frac{(i - \tau)(\overline{i + \tau})}{|i + \tau|^2} = \frac{i - \tau}{i + \tau}.$$

(1) (5.79) with $\sqrt{\Delta} = \tau_2$.
(2) $|\tau + i|^2 = \tau_1^2 + (\tau_2 + 1)^2 = |\tau|^2 + 2\tau_2 + 1$.
(3) (5.79) again, with $A - C + 2iB = 1 - |\tau|^2 + 2i\tau_1$.
(4) $(i - \tau)(-i + \bar\tau) = 1 + i\bar\tau + i\tau - |\tau|^2 = 1 - |\tau|^2 + 2i\tau_1$.

The map $\tau \mapsto \mu = (i - \tau)/(i + \tau)$ is (minus) the **Cayley transform**. It sends the upper half-plane $\tau_2 > 0$ one-to-one onto the unit disc $|\mu| < 1$: the point τ is closer to i than to $-i$ exactly when $\tau_2 > 0$. The square torus $\tau = i$ has $\mu = 0$. So the constant Beltrami coefficient of a flat torus *is* the modular parameter τ, in disguise. We will meet τ again in Section 5.5.

5.4.3 Conformal classes: Weyl rescalings do not change the shape

Look at (5.79) again. If we multiply the metric by a positive function f, then A, B, C all get multiplied by f. The formula for μ is homogeneous of degree zero in (A, B, C), so μ does not change. Only λ changes, to $f\lambda$. The ellipses of Figure 5.40 change size but not shape. Therefore:

Key idea: Complex structures are conformal classes

Two metrics $g_1 = \lambda_1 |\mathrm{d}z + \mu_1 \mathrm{d}\bar z|^2$ and $g_2 = \lambda_2 |\mathrm{d}z + \mu_2 \mathrm{d}\bar z|^2$ define the same complex structure exactly when $\mu_1 = \mu_2$, that is, exactly when $g_1 = f g_2$ for a positive function $f = \lambda_1/\lambda_2$. Complex structures on an oriented surface are in one-to-one correspondence

with **conformal classes** of metrics [Nash p. 149].

The converse direction (every complex structure comes from some metric) is easy: in each holomorphic chart w_α take the flat metric $|\mathrm{d}w_\alpha|^2$, and glue these together with a partition of unity $\sum_\alpha \chi_\alpha = 1$ into $g = \sum_\alpha \chi_\alpha |\mathrm{d}w_\alpha|^2$. On an overlap $|\mathrm{d}w_\beta|^2 = |w'_\beta(w_\alpha)|^2 |\mathrm{d}w_\alpha|^2$, so g is a positive function times $|\mathrm{d}w_\alpha|^2$ in every chart. It is conformal to the flat metric of every chart, which is what we wanted.

Nash packages this as a group action. Let $\mathrm{Met}(\Sigma)$ be the space of all smooth metrics on Σ, and $C^\infty_+(\Sigma)$ the group (under pointwise multiplication) of smooth positive functions. The group acts by multiplication, and the quotient is the space of conformal classes [Nash p. 149, (5.36)]:

$$A : C^\infty_+(\Sigma) \times \mathrm{Met}(\Sigma) \to \mathrm{Met}(\Sigma), \quad (f, g) \mapsto fg; \qquad \mathrm{Conf}(\Sigma) = \frac{\mathrm{Met}(\Sigma)}{C^\infty_+(\Sigma)}. \tag{5.83}$$

Working definition 5.38 (conformal class, $\mathrm{Conf}(\Sigma)$)

Two metrics are **conformally equivalent** if their ratio is a positive function. *Operational test:* compute μ for each with (5.79) in the same coordinates; they are conformally equivalent if and only if the two μ's agree at every point. $\mathrm{Conf}(\Sigma)$ is the set of equivalence classes; [each class is all the metrics conformally equivalent to one given metric] each point of it is a complex structure on Σ [a choice of which local coordinates z count as holomorphic].

Worked example 5.67: Weyl invariance of the Polyakov action

Why does string theory care about conformal classes? Take the integrand of the worldsheet action (5.74) and replace g by fg. In two dimensions,

$$\sqrt{\det(fg)}\,(fg)^{ab} \overset{(1)}{=} \sqrt{f^2 \det g}\; f^{-1} g^{ab} \overset{(2)}{=} f\sqrt{\det g}\; f^{-1} g^{ab} = \sqrt{\det g}\; g^{ab}.$$

(1) The determinant of a 2×2 matrix scales as f^2; the inverse matrix scales as f^{-1}.

(2) $f > 0$, so $\sqrt{f^2} = f$; the two factors cancel.

So the classical action depends only on the conformal class of g. The cancellation is special to dimension two: in dimension n one would get $f^{n/2} f^{-1}$. Quantum mechanically the Weyl invariance survives only in the critical dimension $d = 26$ (the conformal anomaly, Chapter 10). Granting that, the integral over $\mathrm{Met}(\Sigma)$ reduces to an integral over $\mathrm{Conf}(\Sigma)$.

Watch out: $\mathrm{Conf}(\Sigma)$ is still infinite dimensional

Dividing by $C^\infty_+(\Sigma)$ removes one function's worth of freedom out of the three functions A, B, C. What is left is one complex function μ, still an infinite-dimensional space. The diffeomorphisms remove another complex function's worth (two real vector field components), and that is why the final answer is finite dimensional.

5.4.4 Dividing out by diffeomorphisms

If $\alpha : \Sigma \to \Sigma$ is a diffeomorphism and g is a metric, then the pulled-back metric $\alpha^* g$ makes $(\Sigma, \alpha^* g)$ isometric to (Σ, g): the map α is the isometry. Physically, α is only a relabelling of the points of the worldsheet, and the Polyakov action is invariant under it (the fields ϕ are

pulled back too). So we should not count g and $\alpha^* g$ as different. We need to divide $\mathrm{Conf}(\Sigma)$ by diffeomorphisms as well. Three groups appear.

Working definition 5.39 ($\mathrm{Diff}(\Sigma)$, $\mathrm{Diff}^+(\Sigma)$, $\mathrm{Diff}_0(\Sigma)$)

- $\mathrm{Diff}(\Sigma)$: all diffeomorphisms of Σ. [smooth invertible maps $\Sigma \to \Sigma$ with smooth inverse]
- $\mathrm{Diff}^+(\Sigma)$: the orientation-preserving ones (Jacobian determinant > 0 in oriented charts).
- $\mathrm{Diff}_0(\Sigma)$: those that can be joined to the identity by a continuous path α_t of diffeomorphisms, $\alpha_0 = \mathrm{id}$, $\alpha_1 = \alpha$. It is the connected component of the identity.

Test for Diff_0*:* if α acts non-trivially on the first homology (it moves some closed curve to a curve that is not homologous to it) [the two curves together, one reversed, do not bound a region], then $\alpha \notin \mathrm{Diff}_0$, because a continuous deformation cannot change a homology class.

Why only orientation-preserving maps? Consider the reflection $c(x, y) = (x, -y)$, i.e. $c(z) = \bar{z}$. It pulls back $\mathrm{d}z$ to $\mathrm{d}\bar{z}$. So

$$c^*\left(\lambda\,|\mathrm{d}z + \mu\,\mathrm{d}\bar{z}|^2\right) \overset{(1)}{=} (\lambda \circ c)\,|\mathrm{d}\bar{z} + (\mu \circ c)\,\mathrm{d}z|^2 \overset{(2)}{=} (\lambda \circ c)\left|\mathrm{d}z + \overline{(\mu \circ c)}\,\mathrm{d}\bar{z}\right|^2.$$

(1) Pull back: $c^*\mathrm{d}z = \mathrm{d}\bar{z}$ and $c^*\mathrm{d}\bar{z} = \mathrm{d}z$, and functions get composed with c.
(2) A complex number and its conjugate have the same modulus, so $|\mu\,\mathrm{d}z + \mathrm{d}\bar{z}| = |\bar{\mu}\,\mathrm{d}\bar{z} + \mathrm{d}z|$.

The new Beltrami coefficient is $\bar{\mu}$: the *complex conjugate structure*. For the flat torus of Example 5.66, conjugating $\mu = (i-\tau)/(i+\tau)$ gives $\bar{\mu} = (i+\bar{\tau})/(i-\bar{\tau})$, which is the same formula with τ replaced by $-\bar{\tau}$: the mirror-image parallelogram. Orientation-reversing maps identify a Riemann surface with its mirror image, which is in general a different Riemann surface. We keep the two apart by using only $\mathrm{Diff}^+(\Sigma)$ [Nash p. 150].

$\mathrm{Diff}_0(\Sigma)$ is a **normal subgroup** of $\mathrm{Diff}^+(\Sigma)$: if α_t joins id to α, then $h\alpha_t h^{-1}$ joins id to $h\alpha h^{-1}$ for any h. Normality means the quotient $\mathrm{Diff}^+/\mathrm{Diff}_0$ is again a group, the mapping class group of Section 5.5. In this section we use only Diff_0.

Worked example 5.68: Small and large diffeomorphisms of the torus

On $T^2 = \mathbb{R}^2/\mathbb{Z}^2$:

- The translation $\alpha(x, y) = (x + a,\ y + b)$ lies in Diff_0: the path $\alpha_t(x, y) = (x + ta,\ y + tb)$ joins it to the identity.
- The shear $\beta(x, y) = (x + y,\ y)$ is a well-defined diffeomorphism (it maps $\mathbb{Z}^2$ to itself). It sends the closed curve $b : t \mapsto (0, t)$ to $t \mapsto (t, t)$, which winds once around each direction. In homology $[b] \mapsto [a] + [b]$. So $\beta \notin \mathrm{Diff}_0$. It is orientation preserving (Jacobian determinant 1), so $\beta \in \mathrm{Diff}^+$.
- The rotation $\rho(x, y) = (-y,\ x)$ maps $[a] \mapsto [b]$ and $[b] \mapsto -[a]$. Again $\rho \in \mathrm{Diff}^+$ but $\rho \notin \mathrm{Diff}_0$.

The large diffeomorphisms β and ρ are exactly the ones that will generate the modular group in Section 5.5.

Nash's first definition of Teichmüller space is then [Nash p. 150, (5.37)]

$$\mathcal{T}_p \;=\; \frac{\mathrm{Conf}(\Sigma)}{\mathrm{Diff}_0(\Sigma)}, \qquad p \geq 2. \tag{5.84}$$

Why the restriction $p \geq 2$? To see what goes wrong for the sphere and the torus we need a small detour on group actions.

5.4.5 Free actions, fixed points and conformal Killing vectors

Working definition 5.40 (free action, orbit, stabiliser)

A group G **acts** on a space X if each $h \in G$ gives a map $x \mapsto h \cdot x$ with $e \cdot x = x$ and $(h_1 h_2) \cdot x = h_1 \cdot (h_2 \cdot x)$. The **orbit** of x is $G \cdot x = \{h \cdot x : h \in G\}$; the quotient X/G is the set of orbits. The **stabiliser** of x is $G_x = \{h : h \cdot x = x\}$. The action is **free** if every stabiliser is trivial: $h \cdot x = x$ implies $h = e$. *Dimension count:* if G and X are manifolds, the orbit of x has dimension $\dim G - \dim G_x$. When all stabilisers have the same dimension (for instance, all trivial), the quotient is a manifold of dimension $\dim X - \dim G$. Where a stabiliser jumps, [some point x is fixed by more of G than its neighbours are] the quotient is singular there [Nash p. 150, footnote 4].

Worked example 5.69: Three quotients of the plane

1. $\mathbb{R}$ acting on $\mathbb{R}^2$ by horizontal translation $t \cdot (x, y) = (x + t, y)$. Free. Every orbit is a horizontal line, and $\mathbb{R}^2/\mathbb{R} \cong \mathbb{R}$ (label each line by y). A manifold of dimension $2 - 1 = 1$.

2. SO(2) acting by rotations. The origin has stabiliser all of SO(2), of dimension 1; every other point has trivial stabiliser. The orbits are circles of radius $r > 0$ plus the single point $\{0\}$. The quotient is the half-line $[0, \infty)$, which has an edge at $r = 0$. The generic orbits have dimension 1 and the special one has dimension $0 = 1 - 1$.

3. $\mathbb{Z}_2$ acting by rotation through π, $(x, y) \mapsto (-x, -y)$. Only the origin is fixed. The quotient is a *cone* with cone angle π (Figure 5.41): a smooth surface everywhere except at the tip. This is the prototype of an *orbifold point*, and it is exactly what happens in moduli space.

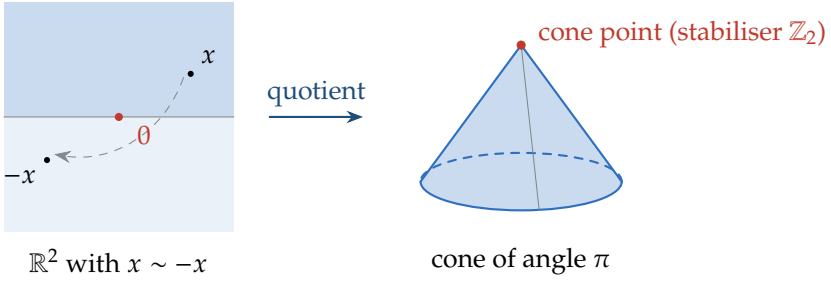


Figure 5.41: Dividing the plane by the rotation through π. The upper half plane (darker) is a fundamental region; the two halves of its edge are glued by $x \sim -x$, rolling it into a cone. Everywhere except the tip the quotient is a smooth surface. At the tip the stabiliser is $\mathbb{Z}_2$ and the space is singular. Moduli spaces of Riemann surfaces have exactly such points.

Now apply this to $\mathrm{Diff}_0(\Sigma)$ acting on $\mathrm{Conf}(\Sigma)$. A point $[g]$ of $\mathrm{Conf}(\Sigma)$ has a nontrivial continuous stabiliser if there is a one-parameter family of diffeomorphisms α_t with $\alpha_t^* g$

conformal to g for all t. Differentiating at $t = 0$, the generating vector field V satisfies

$$\mathcal{L}_V\, g = C\, g \qquad \text{for some function } C. \tag{5.85}$$

Such a V is a **conformal Killing vector** [Nash p. 160, footnote 5]. (If $C = 0$ it is an ordinary Killing vector, an infinitesimal isometry.) So: *the action of* Diff_0 *on* Conf *fails to be free exactly when there are conformal Killing vectors.* Let us find them.

The conformal Killing equation in complex coordinates

We need the Lie derivative of a metric (Chapter T1). In coordinates x^a,

$$(\mathcal{L}_V g)_{ab} = V^c\,\partial_c g_{ab} + g_{cb}\,\partial_a V^c + g_{ac}\,\partial_b V^c. \tag{5.86}$$

The first term is the change of g along the flow, and the other two account for the stretching of the coordinate basis by the flow. We use the complex coordinates $z, \bar z$ as the x^a. A metric in isothermal form $\mathrm{d}s^2 = \lambda\,\mathrm{d}z\,\mathrm{d}\bar z$ has components

$$g_{z\bar z} = g_{\bar z z} = \tfrac{1}{2}\lambda, \qquad g_{zz} = g_{\bar z\bar z} = 0, \tag{5.87}$$

because $\mathrm{d}s^2 = g_{z\bar z}\mathrm{d}z\mathrm{d}\bar z + g_{\bar z z}\mathrm{d}\bar z\mathrm{d}z$. A real vector field is $V = V^z\partial_z + V^{\bar z}\partial_{\bar z}$ with $V^{\bar z} = \overline{V^z}$. Then

$$\begin{aligned}(\mathcal{L}_V g)_{\bar z\bar z} &\overset{(1)}{=} V^c\,\partial_c g_{\bar z\bar z} + g_{c\bar z}\,\partial_{\bar z}V^c + g_{\bar z c}\,\partial_{\bar z}V^c \\ &\overset{(2)}{=} 0 + g_{z\bar z}\,\partial_{\bar z}V^z + g_{\bar z z}\,\partial_{\bar z}V^z \\ &\overset{(3)}{=} \lambda\,\partial_{\bar z}V^z.\end{aligned} \tag{5.88}$$

(1) (5.86) with $a = b = \bar z$.
(2) $g_{\bar z\bar z} \equiv 0$, and the sums over c only keep $c = z$ because $g_{\bar z\bar z} = 0$.
(3) $g_{z\bar z} = g_{\bar z z} = \lambda/2$.

The right side of (5.85), Cg, has zero $\bar z\bar z$ component. The zz component of (5.85) is the complex conjugate of the $\bar z\bar z$ one, and the $z\bar z$ component just determines C. So the conformal Killing equation is simply

$$\boxed{\partial_{\bar z}V^z = 0} \tag{5.89}$$

Conformal Killing vectors are holomorphic vector fields $V = V^z(z)\,\partial_z + \text{c.c.}$

Worked example 5.70: Conformal Killing vectors on the sphere

On $S^2 = \mathbb{C}\cup\{\infty\}$ use z near 0 and $w = 1/z$ near ∞. Since $z = 1/w$, the chain rule gives $\partial_w = (\mathrm{d}z/\mathrm{d}w)\,\partial_z = -w^{-2}\partial_z$, so $\partial_z = -w^2\partial_w$. Try $V = (a + bz + cz^2)\partial_z$:

$$\begin{aligned}V &\overset{(1)}{=} \big(a + bw^{-1} + cw^{-2}\big)(-w^2)\,\partial_w \\ &= -\big(aw^2 + bw + c\big)\,\partial_w.\end{aligned}$$

(1) Substitute $z = 1/w$ and $\partial_z = -w^2\partial_w$.

This is holomorphic at $w = 0$, so V is a global holomorphic vector field. A term $z^3\partial_z = -w^{-1}\partial_w$ would blow up at ∞. In general $V^z(z)$ must be entire, and regularity of $-w^2V^z(1/w)$ at $w = 0$ says $|V^z(z)| \le \text{const}\cdot|z|^2$ for large $|z|$; by the extended Liouville theorem (an entire function bounded by a polynomial of degree k is a polynomial of degree $\le k$), V^z is a quadratic polynomial. So the conformal Killing vectors form a space of complex dimension 3 (real dimension 6). They generate the Möbius maps $z \mapsto (az + b)/(cz + d)$:

∂_z translates, $z\partial_z$ dilates and rotates, $z^2\partial_z$ is a special conformal transformation. In string language they are the $\mathrm{SL}(2,\mathbb{C})$ symmetry that lets one fix three vertex operators on the sphere.

Worked example 5.71: Conformal Killing vectors on the torus

On $T^2 = \mathbb{C}/(\mathbb{Z} + \tau\mathbb{Z})$ a holomorphic vector field is $V^z(z)\partial_z$ with V^z holomorphic on $\mathbb{C}$ and doubly periodic. A continuous doubly periodic function is bounded (it is determined by its values on the compact parallelogram), so by Liouville V^z is constant. The conformal Killing vectors are the translations $V = c\,\partial_z + \bar{c}\,\partial_{\bar{z}}$: complex dimension 1 (real dimension 2). They are actually Killing vectors of the flat metric. In a one-loop string amplitude they let one fix the position of one vertex operator.

For genus $p \geq 2$ there are no holomorphic vector fields at all. We prove this in Section 5.6 (it follows from a degree count: a holomorphic vector field is a section of a line bundle of degree $2 - 2p < 0$). Table 5.1 summarises.

Table 5.1: Conformal Killing vectors (holomorphic vector fields) by genus. A non-zero count means Diff_0 does not act freely on $\mathrm{Conf}(\Sigma)$.

genus p	complex dim	real dim	what they are
0 (sphere)	3	6	∂_z, $z\partial_z$, $z^2\partial_z$: Möbius group $\mathrm{PSL}(2,\mathbb{C})$
1 (torus)	1	2	translations ∂_z
≥ 2	0	0	none

Physical picture: Fixed points are symmetric worldsheets

A conformal Killing vector is a continuous symmetry of the worldsheet's shape. A round sphere can be rotated and still look the same; a flat torus can be slid along itself. The orbit of such a symmetric shape under Diff_0 is smaller than a generic orbit, so the naive quotient $\mathrm{Conf}/\mathrm{Diff}_0$ is singular there. For $p \geq 2$ the only symmetries are discrete (Section 5.5), and Diff_0 acts freely.

5.4.6 Uniformisation: one best metric in each conformal class

Nash's fix for $p < 2$ is to choose a canonical representative in each conformal class: a metric of *constant* Gaussian curvature. That such a metric exists is the content of the uniformisation theorem. We need two pieces of background first.

Background: universal covers

A **covering map** $\pi : \tilde{X} \to X$ is a surjective map that looks locally like a stack of copies of X: every point of X has a neighbourhood U with $\pi^{-1}(U)$ a disjoint union of copies of U. The **universal cover** $\tilde{X}$ is the covering space that is simply connected. Standard examples:

- $\mathbb{R} \to S^1$, $t \mapsto e^{2\pi i t}$. The points t and $t + n$ map to the same point. The group $\mathbb{Z}$ acts on $\mathbb{R}$ by $t \mapsto t + n$, and $S^1 = \mathbb{R}/\mathbb{Z}$.
- $\mathbb{C} \to T^2 = \mathbb{C}/\Lambda$ for a lattice $\Lambda = \mathbb{Z}\omega_1 + \mathbb{Z}\omega_2$. Here $\mathbb{Z} \oplus \mathbb{Z}$ acts by translations $z \mapsto z + m\omega_1 + n\omega_2$.

In general the group of **deck transformations** (maps $\tilde X \to \tilde X$ commuting with π) of the universal cover is the fundamental group $\pi_1(X)$, it acts freely, and [Nash p. 151, (5.38)]

$$\Sigma \;=\; \tilde\Sigma/\pi_1(\Sigma). \tag{5.90}$$

If Σ is a Riemann surface, $\tilde\Sigma$ inherits a complex structure (pull back the charts along π), and the deck transformations are holomorphic.

Background: curvature of a conformally flat metric

For $ds^2 = \lambda|dz|^2 = \lambda(dx^2+dy^2)$ the Gaussian curvature is

$$K \;=\; -\frac{1}{2\lambda}\left(\partial_x^2+\partial_y^2\right)\log\lambda \;=\; -\frac{2}{\lambda}\,\partial_z\partial_{\bar z}\log\lambda, \tag{5.91}$$

using $\partial_x^2+\partial_y^2 = 4\partial_z\partial_{\bar z}$. (This is the standard formula for a conformally flat metric; it follows from the curvature 2-form of the orthonormal coframe $\sqrt{\lambda}\,dx$, $\sqrt{\lambda}\,dy$, see Chapter T2.) The area element is $dA = \lambda\,dx\,dy$. We will also need the Gauss–Bonnet theorem

$$\int_\Sigma K\,dA \;=\; 2\pi\chi(\Sigma) \;=\; 2\pi(2-2p). \tag{5.92}$$

Worked example 5.72: The three model geometries

Round sphere. Stereographic projection gives $\lambda = 4/(1+|z|^2)^2$. Then

$$\partial_z\partial_{\bar z}\log\lambda \overset{(1)}{=} -2\,\partial_z\partial_{\bar z}\log(1+z\bar z) \overset{(2)}{=} -2\,\partial_z\left(\frac{z}{1+z\bar z}\right) \overset{(3)}{=} -\frac{2}{(1+z\bar z)^2},$$

$$K \overset{(4)}{=} -\frac{2}{\lambda}\cdot\left(-\frac{2}{(1+|z|^2)^2}\right) = \frac{(1+|z|^2)^2}{2}\cdot\frac{2}{(1+|z|^2)^2} = 1.$$

(1) $\log\lambda = \log 4 - 2\log(1+z\bar z)$; the constant is killed by derivatives.
(2) $\partial_{\bar z}\log(1+z\bar z) = z/(1+z\bar z)$.
(3) Quotient rule: $\partial_z[z(1+z\bar z)^{-1}] = (1+z\bar z)^{-1} - z\bar z(1+z\bar z)^{-2} = (1+z\bar z)^{-2}$.
(4) (5.91) with $2/\lambda = (1+|z|^2)^2/2$.

Flat plane. $\lambda = 1$, $\log\lambda = 0$, $K = 0$.

Upper half-plane $H = \{\operatorname{Im} z > 0\}$ with $\lambda = 1/y^2$, i.e. $ds^2 = (dx^2+dy^2)/y^2$:

$$K \overset{(1)}{=} -\frac{y^2}{2}\,\partial_y^2(-2\log y) \overset{(2)}{=} -\frac{y^2}{2}\cdot\frac{2}{y^2} = -1.$$

(1) (5.91); $\log\lambda = -2\log y$ does not depend on x.
(2) $\partial_y^2(-2\log y) = 2/y^2$.

Watch out: Nash's formula (5.40) for the hyperbolic metric

Nash writes the curvature -1 metric as $ds^2 = dz\,d\bar z/(z-\bar z)^2$. Since $z-\bar z = 2iy$, this is $dz d\bar z/(-4y^2) = -\tfrac14|dz|^2/y^2$: negative, and a quarter of the correct size. The metric that is positive and has $K=-1$ is

$$ds^2 = \frac{|dz|^2}{y^2} = -\frac{4\,dz\,d\bar z}{(z-\bar z)^2}.$$

Treat (5.40) as "proportional to" and use this corrected version.

Key idea: Riemann uniformisation (Nash (5.39))

Up to biholomorphism there are exactly three simply connected Riemann surfaces: the sphere $S^2 = \mathbb{CP}^1$, the plane $\mathbb{C}$, and the upper half-plane H (equivalently the unit disc). They carry the complete metrics of constant curvature $+1$, 0 and -1 of Example 5.72. The universal cover of a compact Riemann surface of genus p is

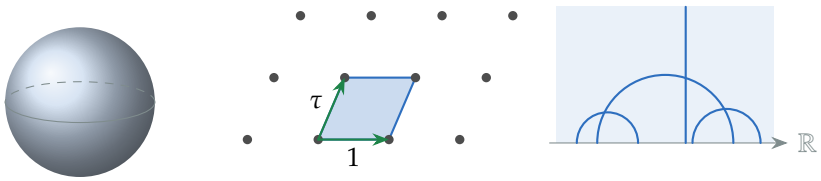
$$\tilde{\Sigma} = S^2 \ (p = 0), \qquad \tilde{\Sigma} = \mathbb{C} \ (p = 1), \qquad \tilde{\Sigma} = H \ (p \geq 2),$$

and the deck group $\pi_1(\Sigma)$ acts by isometries of that metric, so the metric descends to a constant-curvature metric on Σ in the given conformal class.

Which of the three occurs is forced by Gauss–Bonnet. If K is a constant, then (5.92) reads

$$K \cdot \mathrm{Area}(\Sigma) = 2\pi(2 - 2p). \tag{5.93}$$

Area is positive, so $K > 0$ for $p = 0$, $K = 0$ for $p = 1$ and $K < 0$ for $p \geq 2$. With $K = +1$ the sphere has area 4π (the round unit sphere). With $K = -1$ a genus-p surface has area $4\pi(p - 1)$: for example 4π for genus 2. With $K = 0$ the equation says $0 = 0$ and puts no constraint on the area. Figure 5.42 shows the three pictures.



$p = 0$: $\tilde{\Sigma} = \Sigma = S^2$
$K = +1$, $\pi_1 = 1$

$p = 1$: $\tilde{\Sigma} = \mathbb{C}$, $K = 0$
$\pi_1 = \mathbb{Z} \oplus \mathbb{Z}$ (lattice)

$p \geq 2$: $\tilde{\Sigma} = H$, $K = -1$
geodesics: semicircles

Figure 5.42: The three model geometries of uniformisation. Every compact Riemann surface is one of these divided by a group of isometries acting freely: nothing for the sphere, a lattice of translations for the torus (the shaded parallelogram with sides 1 and τ is a copy of the torus), and a discrete group of Möbius maps (a Fuchsian group) for $p \geq 2$. Geodesics of the hyperbolic metric are vertical lines and semicircles centred on the real axis.

5.4.7 The space of constant curvature metrics and Nash's second definition

Uniformisation gives, for each conformal class $c \in \mathrm{Conf}(\Sigma)$, a metric g_c in that class with constant curvature. Nash thinks of $\mathrm{Met}(\Sigma)$ as a bundle over $\mathrm{Conf}(\Sigma)$ (the fibre over c is the set of metrics in the class c, a copy of C^∞_+), and of $c \mapsto g_c$ as a section [Nash p. 151, (5.41)]:

$$s : \mathrm{Conf}(\Sigma) \to \mathrm{Met}(\Sigma), \qquad s(c) = g_c. \tag{5.94}$$

How unique is g_c? Suppose g and $e^{2\sigma} g$ are both flat on a torus. With $g = \lambda |\mathrm{d}z|^2$ and (5.91), flatness of both says $\partial\bar{\partial} \log \lambda = 0$ and $\partial\bar{\partial}(\log \lambda + 2\sigma) = 0$, so $\partial_z \partial_{\bar{z}} \sigma = 0$. Thus σ is a harmonic function on a compact surface. It attains its maximum, and by the maximum principle it is constant. So the flat metric in a conformal class on the torus is unique *up to a constant scale*, and we fix the scale by demanding area 1. For the hyperbolic case the curvature -1 condition fixes the scale as well, and the constant-curvature metric in each conformal class is unique. For the sphere every metric is conformally equivalent to the round one, which is unique up to the conformal automorphisms (the Möbius maps); since we divide by diffeomorphisms anyway this ambiguity is harmless.

So Nash defines [Nash p. 152, (5.42)]

$$\mathrm{Met}_{\mathrm{const}}(\Sigma) = \begin{cases} \{g \in \mathrm{Met}(\Sigma) : R(g) = 1\}, & p = 0, \\ \{g \in \mathrm{Met}(\Sigma) : R(g) = 0,\ \mathrm{vol}(\Sigma) = 1\}, & p = 1, \\ \{g \in \mathrm{Met}(\Sigma) : R(g) = -1\}, & p \geq 2, \end{cases} \tag{5.95}$$

where $R(g)$ is the Gaussian curvature. The second definition of Teichmüller space, which works in every genus, is [Nash p. 152, (5.43)]

$$\boxed{\mathcal{T}_p = \frac{\mathrm{Met}_{\mathrm{const}}(\Sigma)}{\mathrm{Diff}_0(\Sigma)}, \qquad p = 0, 1, 2, \ldots} \tag{5.96}$$

Watch out: A slip in Nash's explanation of the volume condition

Nash says the condition $\mathrm{vol}(\Sigma) = 1$ is imposed "for the $p = 0$ case because in that case Σ is the torus". He means $p = 1$. The logic is (5.93): for $K = 0$ Gauss–Bonnet does not fix the area, so we fix it by hand. For $K = \pm 1$ Gauss–Bonnet fixes it automatically.

Working definition 5.41 (Teichmüller space)

A point of $\mathcal{T}_p$ is a constant-curvature metric [curvature $K = +1, 0, -1$ for $p = 0, 1, \geq 2$, with area fixed to 1 when $K = 0$] on the genus-p surface Σ, where two such metrics count as the same if one is the pull-back of the other by a diffeomorphism isotopic to the identity. [joined to id by a continuous path of diffeomorphisms, i.e. an element of Diff_0] Equivalently it is a complex structure on Σ up to Diff_0. *Operational version:* a complex structure together with a choice of "which curve is which" (a **marking**: a basis a_i, b_i of H_1 up to deformation), since Diff_0 acts trivially on curves up to isotopy, while the larger group Diff^+ can reshuffle them.

Worked example 5.73: Genus zero: $\mathcal{T}_0$ is a point

By uniformisation every metric on S^2 is conformal to one with $K = 1$, and any two metrics with $K = 1$ on S^2 are isometric (both are the round unit sphere). The isometry can be chosen orientation preserving, and $\mathrm{Diff}^+(S^2)$ is connected (every orientation-preserving diffeomorphism of the sphere can be deformed to the identity), so the isometry lies in Diff_0. Hence $\mathrm{Met}_{\mathrm{const}}(S^2)/\mathrm{Diff}_0$ is a single point. The sphere has no shape moduli.

Worked example 5.74: Genus one: $\mathcal{T}_1$ is the upper half-plane

Here is a direct computation of $\mathcal{T}_1$ from (5.96); Section 5.5 redoes it Nash's way with lattices. Work on $T^2 = \mathbb{R}^2/\mathbb{Z}^2$. Take a flat metric of area one. Its universal cover is the flat plane, so in suitable coordinates the metric has *constant* coefficients; after pulling back by an element of Diff_0 we may assume this happens in our coordinates (x, y). (Sketch: a flat torus is $\mathbb{R}^2/\Lambda$ with the Euclidean metric; the linear map sending $\mathbb{Z}^2$ to Λ with the correct marking pulls it back to a constant-coefficient metric on $\mathbb{R}^2/\mathbb{Z}^2$, and the difference between two such identifications is homotopic to the identity.) So we must classify

$$G = \begin{pmatrix} A & B \\ B & C \end{pmatrix}, \qquad A > 0, \quad \det G = AC - B^2 = 1,$$

the determinant being the area of the unit square. Parametrise by $\tau = \tau_1 + i\tau_2$ with $\tau_2 > 0$:

$$G(\tau) = \frac{1}{\tau_2} \begin{pmatrix} 1 & \tau_1 \\ \tau_1 & |\tau|^2 \end{pmatrix}, \qquad \text{i.e.} \qquad \mathrm{d}s^2 = \frac{1}{\tau_2} |\mathrm{d}x + \tau\, \mathrm{d}y|^2. \tag{5.97}$$

Check: $\det G = (|\tau|^2 - \tau_1^2)/\tau_2^2 = \tau_2^2/\tau_2^2 = 1$. The map is invertible: given G with $\det G = 1$, set $\tau_2 = 1/A$ and $\tau_1 = B/A$. Then

$$\frac{|\tau|^2}{\tau_2} \overset{(1)}{=} \left(\frac{B^2}{A^2} + \frac{1}{A^2}\right)A \overset{(2)}{=} \frac{B^2+1}{A} \overset{(3)}{=} \frac{AC}{A} = C,$$

(1) $|\tau|^2 = \tau_1^2 + \tau_2^2$ with the values just chosen, and $1/\tau_2 = A$.
(2) Algebra.
(3) $\det G = 1$ means $B^2 + 1 = AC$.

so $G = G(\tau)$ exactly. Finally, two different τ's are not related by Diff_0: an isometry between two flat tori lifts to an affine map of $\mathbb{R}^2$ preserving $\mathbb{Z}^2$, whose linear part is an integer matrix; being in Diff_0 it acts trivially on H_1, so the linear part is the identity and the map is a translation, which does not change a constant G. Hence

$$\mathcal{T}_1 \;\cong\; \{\tau : \tau_2 > 0\} = H, \tag{5.98}$$

one complex dimension. The space of unit-determinant positive 2×2 matrices is also $\mathrm{SL}(2,\mathbb{R})/\mathrm{SO}(2)$; its natural geometry is the hyperbolic plane, a first hint of the Weil–Petersson metric of chapter 9. Comparing with Example 5.66, the Beltrami coefficient of $G(\tau)$ is $\mu = (i - \tau)/(i + \tau)$ (the overall factor $1/\tau_2$ changes only λ).

Physical picture: What $\mathcal{T}_p$ remembers

Teichmüller space remembers the shape *and* a labelling of the curves on the surface. Two tori with τ and $\tau + 1$ have the same shape (the parallelogram with sides $1, \tau + 1$ tiles the same lattice as the one with sides $1, \tau$), but they are different points of $\mathcal{T}_1$, because the curve we called b has been sent to $a + b$. Forgetting the labelling is the passage to moduli space.

Reading Nash (pp. 148–152, (5.35)–(5.43))

Nash writes the general metric as $\lambda|\mathrm{d}z + \mu\mathrm{d}\bar{z}|^2$ (5.35), observes that μ labels complex structures, and forms $\mathrm{Conf}(\Sigma) = \mathrm{Met}/C^\infty_+$ (5.36). His first definition $\mathcal{T}_p = \mathrm{Conf}/\mathrm{Diff}_0$ (5.37) is valid for $p \geq 2$; footnote 4 explains free actions and orbit dimensions, and the failure for $p < 2$ is blamed on conformal Killing vectors. Uniformisation (5.38)–(5.40) supplies the constant curvature section s (5.41), leading to $\mathrm{Met}_{\mathrm{const}}$ (5.42) and the all-genus definition (5.43). We added: the explicit solution (5.79) for λ, μ; the ellipse picture; the proof that isothermal charts are holomorphic; the conformal Killing equation (5.89) and its solutions on S^2 and T^2; the curvature computations; the corrected hyperbolic metric; and the direct identification $\mathcal{T}_1 = H$.

Recap

- Any metric on a surface is $\lambda|\mathrm{d}z + \mu\mathrm{d}\bar{z}|^2$ with $|\mu| < 1$; μ is a field of infinitesimal ellipses and is unchanged by Weyl rescaling.
- Metric + orientation $\Rightarrow$ complex structure (isothermal charts), and complex structures = conformal classes.
- Teichmüller space $\mathcal{T}_p = \mathrm{Conf}/\mathrm{Diff}_0$ for $p \geq 2$; conformal Killing vectors (holomorphic vector fields, $\partial_{\bar{z}} V^z = 0$) spoil freeness for $p = 0, 1$.
- Uniformisation: each class has a constant-curvature metric ($K = +1, 0, -1$ for $p =$

$0, 1, \geq 2$, forced by Gauss–Bonnet), giving $\mathcal{T}_p = \mathrm{Met}_{\mathrm{const}}/\mathrm{Diff}_0$ in all genera.

- $\mathcal{T}_0$ is a point; $\mathcal{T}_1 = H$, the upper half-plane of τ.

5.5 The moduli space $\mathcal{M}_p$

Teichmüller space still overcounts. In Example 5.74 the tori with parameters τ and $\tau + 1$ were different points of $\mathcal{T}_1$. Yet they are the same flat torus: the lattice $\mathbb{Z} + \tau\mathbb{Z}$ equals the lattice $\mathbb{Z} + (\tau + 1)\mathbb{Z}$, since each basis is an integer combination of the other. They differ only in which closed curve we decided to call "b". A string path integral must count each physically distinct worldsheet exactly once, so it must forget this labelling. The space of Riemann surfaces with the labels forgotten is the moduli space.

5.5.1 Large diffeomorphisms and the mapping class group

The fix is to divide by all orientation-preserving diffeomorphisms, not only those in the identity component [Nash p. 152, (5.44)]:

$$\mathcal{T}_p = \frac{\mathrm{Met}_{\mathrm{const}}(\Sigma)}{\mathrm{Diff}_0(\Sigma)}, \qquad \mathcal{M}_p = \frac{\mathrm{Met}_{\mathrm{const}}(\Sigma)}{\mathrm{Diff}^+(\Sigma)}. \tag{5.99}$$

Because Diff_0 is a normal subgroup of Diff^+, we can take the quotient in two steps. Define the **mapping class group**

$$\Gamma_\Sigma = \frac{\mathrm{Diff}^+(\Sigma)}{\mathrm{Diff}_0(\Sigma)}. \tag{5.100}$$

Then [Nash p. 152, (5.45)]

$$\mathcal{M}_p \overset{(1)}{=} \frac{\mathrm{Met}_{\mathrm{const}}}{\mathrm{Diff}^+} \overset{(2)}{=} \frac{\left(\mathrm{Met}_{\mathrm{const}}/\mathrm{Diff}_0\right)}{\left(\mathrm{Diff}^+/\mathrm{Diff}_0\right)} \overset{(3)}{=} \frac{\mathcal{T}_p}{\Gamma_\Sigma}. \tag{5.101}$$

(1) Definition (5.99).
(2) Two metrics lie in the same Diff^+-orbit iff their Diff_0-orbits are related by some $h \in \mathrm{Diff}^+$; and h and $h\alpha$ with $\alpha \in \mathrm{Diff}_0$ move Diff_0-orbits in the same way (normality: $h\alpha = (h\alpha h^{-1})h$ with $h\alpha h^{-1} \in \mathrm{Diff}_0$), so only the class of h in $\mathrm{Diff}^+/\mathrm{Diff}_0$ matters.
(3) Definitions (5.99) and (5.100).

Working definition 5.42 (mapping class group)

Γ_Σ is the group of orientation-preserving diffeomorphisms of Σ modulo those that can be continuously deformed to the identity. Its elements are the connected components of $\mathrm{Diff}^+(\Sigma)$, so it is a *discrete* group. [no two different elements can be joined by a continuous path] *Operational handle:* let a diffeomorphism act on the first homology $H_1(\Sigma;\mathbb{Z}) \cong \mathbb{Z}^{2p}$. This gives an integer $2p \times 2p$ matrix preserving the intersection form of Nash's (5.15), [(a, b), the signed count of crossings of two loops] i.e. an element of $\mathrm{Sp}(2p, \mathbb{Z})$. [integer matrices M with $M^T J M = J$, where J is the matrix of intersection numbers of the basis loops] For the torus this matrix determines the mapping class completely, and $\Gamma_{T^2} \cong \mathrm{SL}(2, \mathbb{Z})$. For $p \geq 2$ the map $\Gamma_\Sigma \to \mathrm{Sp}(2p, \mathbb{Z})$ is onto, but it has a large kernel (the Torelli group) [mapping classes that send every loop to a homologous loop], so the homology action does not see everything.

The mapping class group acts on $\mathcal{T}_p$ by changing the marking. It does *not* act freely: a Riemann surface with a discrete symmetry is a fixed point. So $\mathcal{M}_p$ is not a manifold. It is an orbifold.

Working definition 5.43 (orbifold)

An **orbifold** is a space that looks locally like $\mathbb{R}^n/G_x$ [points of $\mathbb{R}^n$ related by an element of G_x count as one point] for some finite group G_x acting linearly on $\mathbb{R}^n$. At points with G_x trivial it is an ordinary manifold. At points with $G_x \neq 1$ it has a singularity of the type seen in the cone of Figure 5.41. *Test:* if a space is X/Γ with X a manifold and Γ a discrete group acting with finite stabilisers [each point is fixed by only finitely many elements of Γ], it is an orbifold, and its singular points are the images of points with nontrivial stabiliser [Nash p. 152].

5.5.2 Surfaces as quotients of their universal cover

Nash gets a concrete picture of $\mathcal{M}_p$ from uniformisation (5.90): every Σ is $\tilde{\Sigma}/\pi_1(\Sigma)$ [Nash p. 153, (5.46)]. To build a Riemann surface this way we need an action of $\pi_1(\Sigma)$ on $\tilde{\Sigma}$ that is

- *free*, so that the quotient is a smooth surface (no cone points);
- *holomorphic*, so that the complex structure of $\tilde{\Sigma}$ descends to the quotient.

Classifying genus-p Riemann surfaces then means classifying such group actions up to holomorphic change of coordinates on $\tilde{\Sigma}$.

Genus zero

Here $\tilde{\Sigma} = S^2$ and $\pi_1(S^2) = 1$, so there is nothing to divide by. It is consistent that S^2 has no nontrivial free holomorphic actions at all. The holomorphic automorphisms of $\mathbb{CP}^1$ are the Möbius maps $z \mapsto (az+b)/(cz+d)$ with $ad - bc \neq 0$, and every one has a fixed point:

$$z = \frac{az+b}{cz+d} \iff z(cz+d) = az+b \iff cz^2 + (d-a)z - b = 0.$$

If $c \neq 0$ this quadratic has a root in $\mathbb{C}$ (fundamental theorem of algebra). If $c = 0$ the map is $z \mapsto (az+b)/d$, which fixes $z = \infty$. So every automorphism of the sphere has a fixed point, no nontrivial group acts freely, and *the only Riemann surface of genus zero is* $\mathbb{CP}^1$:

$$\mathcal{M}_0 = \{\text{point}\}.$$

Genus one: $\mathcal{T}_1 = \mathbb{C}^+$

Now $\tilde{\Sigma} = \mathbb{C}$ and $\pi_1(T^2) = \mathbb{Z} \oplus \mathbb{Z}$. The fixed-point-free holomorphic automorphisms of $\mathbb{C}$ are the translations $z \mapsto z + \omega$ (an automorphism $z \mapsto az + b$ with $a \neq 1$ fixes $b/(1-a)$). So $\mathbb{Z} \oplus \mathbb{Z}$ must act by two independent translations,

$$z \mapsto z + m\,\omega_1 + n\,\omega_2, \qquad m, n \in \mathbb{Z},$$

with ω_1, ω_2 linearly independent over $\mathbb{R}$ (otherwise the orbits are not discrete and the quotient is not a surface). Rescaling the coordinate, $z \mapsto z/\omega_1$, is holomorphic and turns the translations into $z \mapsto z + m + n\lambda$ with $\lambda = \omega_2/\omega_1$. If $\operatorname{Im}\lambda < 0$ we replace ω_2 by $-\omega_2$ (this is just choosing the

other generator), so we may take $\lambda \in \mathbb{C}^+ = \{\operatorname{Im}\lambda > 0\}$. Every such λ gives a torus $\mathbb{C}/(\mathbb{Z} + \lambda\mathbb{Z})$, and Nash concludes [Nash p. 153, (5.47)]

$$\mathcal{T}_1 = \mathbb{C}^+, \tag{5.102}$$

in agreement with Example 5.74. Nash calls λ the **modular parameter**; physicists write τ. We will use both.

5.5.3 When are two tori the same? The modular group

Let $T_1^2 = \mathbb{C}/(\mathbb{Z} + \lambda_1\mathbb{Z})$ and $T_2^2 = \mathbb{C}/(\mathbb{Z} + \lambda_2\mathbb{Z})$, and suppose $f : T_1^2 \to T_2^2$ is a biholomorphic (conformal, invertible) map [Nash p. 153, (5.48)]. We want the condition on λ_1, λ_2.

Step 1: the lift is affine

Because $\mathbb{C}$ is simply connected, $f \circ p_1 : \mathbb{C} \to T_2^2$ lifts through the covering $p_2 : \mathbb{C} \to T_2^2$ to a holomorphic map $\tilde f : \mathbb{C} \to \mathbb{C}$ with $f \circ p_1 = p_2 \circ \tilde f$ [Nash p. 154, (5.49)]:

$$\begin{array}{ccc} \mathbb{C} & \xrightarrow{\ \tilde f\ } & \mathbb{C} \\ {\scriptstyle p_1}\downarrow & & \downarrow{\scriptstyle p_2} \\ T_1^2 & \xrightarrow[\ f\ ]{} & T_2^2 \end{array}$$

Nash says "it is immediate that $\tilde f(z) = pz + q$". Here is why. Write $\Lambda_i = \mathbb{Z} + \lambda_i\mathbb{Z}$. For each lattice vector $\omega \in \Lambda_1$, the points z and $z + \omega$ map to the same point of T_1^2, so $\tilde f(z + \omega) - \tilde f(z)$ lies in Λ_2. It depends continuously on z and takes values in a discrete set, so it is a constant $c(\omega)$. Differentiating,

$$\tilde f'(z + \omega) = \tilde f'(z) \qquad \text{for all } \omega \in \Lambda_1.$$

So $\tilde f'$ is an entire function that is doubly periodic, hence bounded, hence constant by Liouville's theorem. Therefore $\tilde f(z) = pz + q$ with $p \neq 0$ (since f is invertible).

Step 2: lattices must correspond

Nash phrases the rest with group actions [Nash p. 154, (5.50)]. Let G_i be the translation group of Λ_i. The lift intertwines them: for each $g_2 \in G_2$ there is $g_1 \in G_1$ with $g_1 \cdot z = \tilde f^{-1}(g_2 \cdot \tilde f(z))$. Take g_2 to be the two generators $w \mapsto w + \lambda_2$ and $w \mapsto w + 1$. With $\tilde f^{-1}(w) = (w - q)/p$:

$$\begin{aligned} \tilde f^{-1}\big(\tilde f(z) + \lambda_2\big) &\overset{(1)}{=} \frac{(pz + q) + \lambda_2 - q}{p} = z + \frac{\lambda_2}{p} \overset{(2)}{=} z + b + a\lambda_1, \\ \tilde f^{-1}\big(\tilde f(z) + 1\big) &\overset{(1)}{=} \frac{(pz + q) + 1 - q}{p} = z + \frac{1}{p} \overset{(2)}{=} z + d + c\lambda_1, \end{aligned} \tag{5.103}$$

with $a, b, c, d \in \mathbb{Z}$ [Nash p. 154, (5.51)].

(1) Insert $\tilde f(z) = pz + q$ and $\tilde f^{-1}(w) = (w - q)/p$.
(2) The result must be a translation by an element of Λ_1, i.e. by $m + n\lambda_1$ for some integers; we name them (b, a) and (d, c).

So $\lambda_2 = p(a\lambda_1 + b)$ and $1 = p(c\lambda_1 + d)$. Divide the first by the second to eliminate p [Nash p. 154, (5.52)]:

$$\lambda_2 = \frac{a\lambda_1 + b}{c\lambda_1 + d}, \qquad a, b, c, d \in \mathbb{Z}. \tag{5.104}$$

Step 3: the determinant is $+1$

The same argument applied to f^{-1} gives integers expressing $(\lambda_1, 1)$ in terms of $(\lambda_2, 1)$. So the integer matrix $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ has an inverse that is also an integer matrix. Its determinant is a nonzero integer (5.53) whose inverse is also an integer, so $ad - bc = \pm 1$. To fix the sign, take imaginary parts in (5.104):

$$\operatorname{Im}\frac{a\lambda + b}{c\lambda + d} \overset{(1)}{=} \operatorname{Im}\frac{(a\lambda + b)(c\bar{\lambda} + d)}{|c\lambda + d|^2} \overset{(2)}{=} \frac{\operatorname{Im}\big(ac|\lambda|^2 + ad\lambda + bc\bar{\lambda} + bd\big)}{|c\lambda + d|^2} \overset{(3)}{=} \frac{(ad - bc)\operatorname{Im}\lambda}{|c\lambda + d|^2}. \tag{5.105}$$

(1) Multiply numerator and denominator by $\overline{c\lambda + d}$; the integers are real.
(2) Expand.
(3) $ac|\lambda|^2$ and bd are real; $\operatorname{Im}(\bar{\lambda}) = -\operatorname{Im}\lambda$.

Both λ_1 and λ_2 are in $\mathbb{C}^+$, so $ad - bc > 0$, hence $ad - bc = +1$. Conversely, if λ_2 is related to λ_1 by such a matrix, then $\tilde{f}(z) = z/(c\lambda_1 + d)$ maps Λ_1 onto Λ_2 and descends to a biholomorphism. We have Nash's criterion [Nash p. 155, (5.54)]:

Key idea: When two tori are the same

$\mathbb{C}/(\mathbb{Z} + \lambda_1\mathbb{Z})$ and $\mathbb{C}/(\mathbb{Z} + \lambda_2\mathbb{Z})$ are biholomorphic if and only if

$$\lambda_2 = \frac{a\lambda_1 + b}{c\lambda_1 + d} \quad \text{with} \quad \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}(2, \mathbb{Z}),$$

i.e. $a, b, c, d \in \mathbb{Z}$ and $ad - bc = 1$.

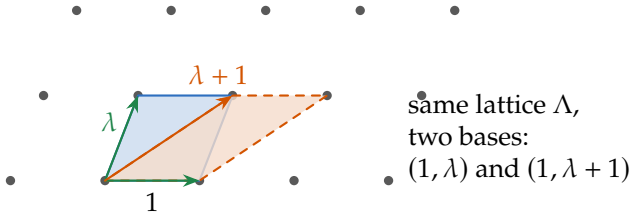


Figure 5.43: One lattice, two bases. The blue parallelogram (sides $1, \lambda$) and the dashed orange one (sides $1, \lambda + 1$) have the same area and tile the plane with the same set of corner points. Both describe the same torus $\mathbb{C}/\Lambda$. Changing basis by an $\mathrm{SL}(2, \mathbb{Z})$ matrix changes λ by the corresponding Möbius map; this is the redundancy that the modular group removes.

The modular group inside the Möbius group

The **Möbius group** is the group of maps [Nash p. 155, (5.55)]

$$z \mapsto \frac{az + b}{cz + d}, \qquad a, b, c, d \in \mathbb{C}, \quad ad - bc = 1.$$

Composition of Möbius maps is matrix multiplication. Indeed, if $M_1 = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ and $M_2 = \left(\begin{smallmatrix} a' & b' \\ c' & d' \end{smallmatrix}\right)$, then

$$M_1 \cdot (M_2 \cdot z) \overset{(1)}{=} \frac{a\,\frac{a'z + b'}{c'z + d'} + b}{c\,\frac{a'z + b'}{c'z + d'} + d} \overset{(2)}{=} \frac{(aa' + bc')z + (ab' + bd')}{(ca' + dc')z + (cb' + dd')} \overset{(3)}{=} (M_1M_2) \cdot z.$$

(1) Apply M_2 first, then M_1.
(2) Multiply numerator and denominator by $c'z + d'$.
(3) The coefficients are exactly the entries of the product matrix $M_1 M_2$.

The matrices M and $-M$ give the same map (numerator and denominator both change sign). So the Möbius group is $\mathrm{SL}(2,\mathbb{C})/\{\pm I\}$. Restricting to integer entries gives the **modular group**

$$\mathrm{PSL}(2,\mathbb{Z}) = \mathrm{SL}(2,\mathbb{Z})/\{\pm I\}, \tag{5.106}$$

which Nash writes $\mathrm{SL}(2,\mathbb{Z})/\{\mp I\}$. It is a discrete subgroup of the Möbius group. Our criterion says precisely that complex structures on the torus are points of $\mathbb{C}^+$ modulo the modular group [Nash p. 155, (5.56)]:

$$\boxed{\mathcal{M}_1 = \frac{\mathcal{T}_1}{\Gamma_{T^2}} = \frac{\mathbb{C}^+}{\mathrm{SL}(2,\mathbb{Z})/\{\pm I\}}.} \tag{5.107}$$

In particular the mapping class group of the torus is $\Gamma_{T^2} \cong \mathrm{SL}(2,\mathbb{Z})/\{\pm I\}$ as far as its action on $\mathcal{T}_1$ goes. (The element $-I$ is the map $z \mapsto -z$, which is a nontrivial mapping class but acts trivially on $\mathbb{C}^+$: every torus has this symmetry.)

Worked example 5.75: The torus automorphisms, explicitly

Write a point of the torus as $x + \lambda y$ with $(x,y) \in \mathbb{R}^2/\mathbb{Z}^2$, so the torus is the unit square with opposite sides glued. Nash's discrete automorphisms [Nash p. 155, (5.57)] are

$$(x,y) \mapsto (ax + by,\ cx + dy), \qquad \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}(2,\mathbb{Z}). \tag{5.108}$$

Integer entries make the map well defined mod $\mathbb{Z}^2$; determinant 1 makes it invertible with integer inverse and orientation preserving. It sends the curve a (the x-axis, $t \mapsto (t,0)$) to $t \mapsto (at, ct)$, which winds a times around the x-direction and c times around the y-direction: $[a] \mapsto a[a] + c[b]$. Likewise $[b] \mapsto b[a] + d[b]$. The action on $H_1 = \mathbb{Z}^2$ is the matrix itself, so for $M \neq I$ the map is not in Diff_0. These maps belong to disconnected components of $\mathrm{Diff}^+(T^2)$, and they are the whole mapping class group.

5.5.4 The generators S and T and the fundamental domain

Two elements do most of the work [Nash p. 156]:

$$S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} :\ \lambda \mapsto -\frac{1}{\lambda}, \qquad T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} :\ \lambda \mapsto \lambda + 1. \tag{5.109}$$

Their relations are short matrix multiplications:

$$S^2 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} = -I,$$

$$ST = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & 1 \end{pmatrix}, \qquad (ST)^2 = \begin{pmatrix} -1 & -1 \\ 1 & 0 \end{pmatrix},$$

$$(ST)^3 = \begin{pmatrix} -1 & -1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 0 & -1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} = -I.$$

So in $\mathrm{PSL}(2,\mathbb{Z})$, $S^2 = 1$ and $(ST)^3 = 1$. In fact these are the only relations, and S, T generate: $\mathrm{PSL}(2,\mathbb{Z}) \cong \mathbb{Z}_2 * \mathbb{Z}_3$. The generation is Euclid's algorithm in disguise: multiplying $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ on the left by T^n replaces a by $a + nc$, and by S swaps the rows up to sign, so one can reduce $|c|$ step by step until $c = 0$, and then the matrix is $\pm T^m$.

Physical picture: S swaps the two radii

Take a rectangular torus with side lengths R_1 (along x) and R_2 (along y), i.e. the lattice $R_1\mathbb{Z} + iR_2\mathbb{Z}$, so $\lambda = iR_2/R_1$. Then $S\lambda = -1/\lambda = iR_1/R_2$: the same torus rotated by 90° and rescaled, i.e. with the roles of the two circles exchanged [Nash p. 156]. In string theory this is the root of T-duality-type statements on a torus target; on the worldsheet it says the two cycles of a torus are on an equal footing. T changes the basis $(1, \lambda)$ to $(1, \lambda + 1)$ (Figure 5.43): a "twist" of the torus by one full turn before regluing.

The fundamental domain

A **fundamental domain** for $\mathrm{PSL}(2,\mathbb{Z})$ acting on $\mathbb{C}^+$ is a region F containing exactly one representative of almost every orbit. The standard choice is

$$F = \left\{\tau \in \mathbb{C}^+ : |\mathrm{Re}\,\tau| \le \tfrac{1}{2},\ |\tau| \ge 1\right\}, \tag{5.110}$$

drawn in Figure 5.44. Its two vertical edges are glued by T, and the two halves of its bottom arc are glued by S (which maps the unit circle to itself, $e^{i\theta} \mapsto -e^{-i\theta}$, reflecting it about i).

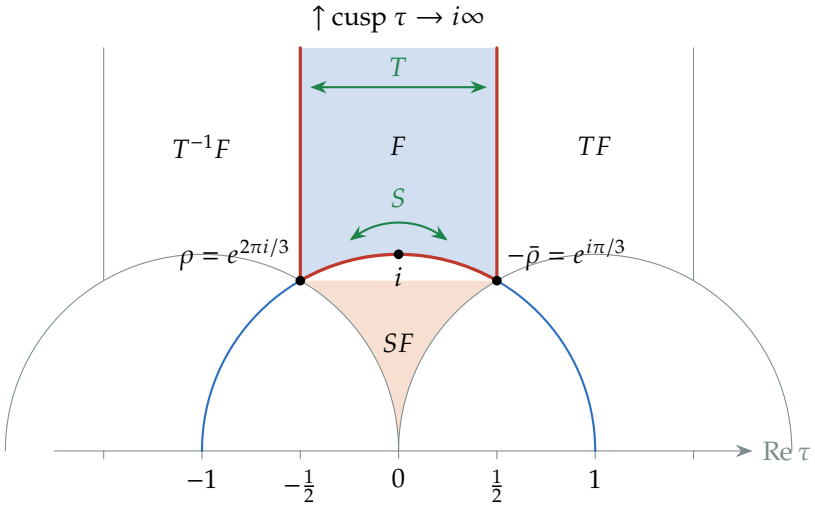


Figure 5.44: The fundamental domain F (blue, red boundary) of the modular group in the upper half-plane. Every torus has a modular parameter in F. The vertical sides are identified by $T : \tau \mapsto \tau + 1$ and the two halves of the bottom arc by $S : \tau \mapsto -1/\tau$. Images SF (orange, below the arc), TF and $T^{-1}F$ tile the half-plane. The special points i and ρ (equivalently $e^{i\pi/3} = T\rho$) are the orbifold points; the cusp at $i\infty$ is the place where the torus degenerates.

Key idea: Every torus has a representative in F

Every $\tau \in \mathbb{C}^+$ can be moved into F by an element of the modular group; two interior points of F are never equivalent; and boundary points are equivalent only via the gluings in Figure 5.44. So, as a set, $\mathcal{M}_1 = F/\!\sim$.

Why can we always get into F? Fix τ and look at the imaginary parts along its orbit. By (5.105), for $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}(2,\mathbb{Z})$,

$$\mathrm{Im}(\gamma\tau) = \frac{\mathrm{Im}\,\tau}{|c\tau + d|^2}. \tag{5.111}$$

The numbers $c\tau + d$ with c, d integers form a lattice, so only finitely many of them have $|c\tau + d| < 1$. Hence $\mathrm{Im}(\gamma\tau)$ takes only finitely many values larger than $\mathrm{Im}\,\tau$, and there is a γ

making $\operatorname{Im}(\gamma\tau)$ as large as possible. Call the result τ'. Apply a power T^n to get $|\operatorname{Re}\tau'| \le \frac{1}{2}$; this does not change the imaginary part. Now if $|\tau'| < 1$, then

$$\operatorname{Im}(S\tau') \overset{(1)}{=} \frac{\operatorname{Im}\tau'}{|\tau'|^2} \overset{(2)}{>} \operatorname{Im}\tau',$$

(1) (5.111) with $c = 1$, $d = 0$.
(2) $|\tau'| < 1$.

contradicting maximality. So $|\tau'| \ge 1$ and $\tau' \in F$. The proof is also an algorithm: *translate into the strip, invert if inside the unit circle, repeat.*

Worked example 5.76: Reducing a modular parameter

(a) $\tau = (2+i)/5 = 0.4 + 0.2i$. Then $|\tau|^2 = 0.2 < 1$, so apply S:

$$S\tau \overset{(1)}{=} -\frac{5}{2+i} \overset{(2)}{=} -\frac{5(2-i)}{5} = -2 + i, \qquad T^2 S\tau = i.$$

(1) $S\tau = -1/\tau$.
(2) Multiply top and bottom by $2 - i$; $|2+i|^2 = 5$.

So the skewed torus with $\tau = (2+i)/5$ is the *square* torus. The group element is $T^2S = \left(\begin{smallmatrix} 2 & -1 \\ 1 & 0 \end{smallmatrix}\right)$; check: $(2\tau - 1)/\tau = 2 - 1/\tau = 2 - (2 - i) = i$.

(b) $\tau = (2+3i)/10$. Now $|\tau|^2 = 13/100 < 1$, so

$$S\tau \overset{(1)}{=} -\frac{10}{2+3i} = -\frac{10(2-3i)}{13} = \frac{-20+30i}{13} \approx -1.54 + 2.31i,$$
$$T^2 S\tau \overset{(2)}{=} \frac{6+30i}{13} \approx 0.46 + 2.31i.$$

(1) $|2+3i|^2 = 13$.
(2) Add $2 = 26/13$; now $|\operatorname{Re}| \le \frac{1}{2}$ and $|\tau| > 1$, so we are in F.

A small $\operatorname{Im}\tau$ (a torus that looks very thin in one direction) becomes a large $\operatorname{Im}\tau$ (a long thin torus in the other direction): they are the same torus.

Fixed points: the orbifold points of $\mathcal{M}_1$

Nash notes [Nash p. 156] that $\lambda \mapsto -1/\lambda$ and $\lambda \mapsto 1 - 1/\lambda$ have fixed points. Let us find them.

- S: $\tau = -1/\tau \iff \tau^2 = -1 \iff \tau = i$ (in $\mathbb{C}^+$). The stabiliser of i in $\mathrm{PSL}(2,\mathbb{Z})$ is $\{1, S\} \cong \mathbb{Z}_2$.

- $TS : \tau \mapsto T(-1/\tau) = 1 - 1/\tau$, with matrix $TS = \left(\begin{smallmatrix} 1 & -1 \\ 1 & 0 \end{smallmatrix}\right)$. Fixed points: $\tau^2 - \tau + 1 = 0$, so $\tau = \frac{1+i\sqrt{3}}{2} = e^{i\pi/3}$. Here $(TS)^3 = -I$ (the same computation as for $(ST)^3$), so the stabiliser is $\mathbb{Z}_3$. The point $e^{i\pi/3}$ is $T\rho$ with $\rho = e^{2\pi i/3}$: the two bottom corners of F are the same point of $\mathcal{M}_1$.

These fixed points are the cone points of $\mathcal{M}_1$ (cone angles π and $2\pi/3$). Together with the cusp at $i\infty$ they are the "singularities" of $\mathcal{M}_1$ that Nash refers to.

Worked example 5.77: The square torus is a fixed point in $\mathrm{Met}_{\mathrm{const}}$

Nash also sees the orbifold point upstairs, in the space of metrics [Nash p. 156]. Take the square torus, the unit square with the metric δ_{ij}, so $\lambda = i$. The automorphism corresponding to S has $a = d = 0$, $b = -c = 1$ in (5.108), i.e. $(x, y) \mapsto (y, -x)$: a rotation by $-90°$. It pulls back $\mathrm{d}x^2 + \mathrm{d}y^2$ to $\mathrm{d}y^2 + \mathrm{d}x^2$, the same metric. So this element of $\mathrm{Diff}^+(T^2)$ fixes the point $\delta_{ij} \in \mathrm{Met}_{\mathrm{const}}(T^2)$, the action of Diff^+ is not free there, and $\mathrm{Met}_{\mathrm{const}}(T^2)/\mathrm{Diff}^+(T^2) = \mathcal{M}_1$ is an orbifold. Similarly the hexagonal torus ($\lambda = e^{i\pi/3}$) is invariant under rotation by $60°$.

5.5.5 Integrating over $\mathcal{M}_1$: the one-loop string amplitude

For the torus, the Polyakov integral (5.74) reduces (chapter 9) to an integral of the form

$$Z_1 \;\propto\; \int_F \frac{\mathrm{d}^2\tau}{\tau_2^2}\, \mathcal{Z}(\tau, \bar\tau), \qquad \mathrm{d}^2\tau = \mathrm{d}\tau_1\, \mathrm{d}\tau_2, \tag{5.112}$$

where $\mathcal{Z}$ is a modular-invariant function built from the determinants of chapter 4. Integrating over F rather than over all of $\mathbb{C}^+$ is exactly "count each torus once". For this to make sense the measure must itself be modular invariant.

Worked example 5.78: The invariant measure on the upper half-plane

Let $\tau' = (a\tau + b)/(c\tau + d)$ with $ad - bc = 1$. Then

$$\frac{\mathrm{d}\tau'}{\mathrm{d}\tau} \overset{(1)}{=} \frac{a(c\tau + d) - c(a\tau + b)}{(c\tau + d)^2} \overset{(2)}{=} \frac{1}{(c\tau + d)^2},$$
$$\mathrm{d}^2\tau' \overset{(3)}{=} \left|\frac{\mathrm{d}\tau'}{\mathrm{d}\tau}\right|^2 \mathrm{d}^2\tau = \frac{\mathrm{d}^2\tau}{|c\tau + d|^4},$$
$$\frac{\mathrm{d}^2\tau'}{(\tau_2')^2} \overset{(4)}{=} \frac{\mathrm{d}^2\tau}{|c\tau + d|^4} \cdot \frac{|c\tau + d|^4}{\tau_2^2} = \frac{\mathrm{d}^2\tau}{\tau_2^2}.$$

(1) Quotient rule.
(2) $ad - bc = 1$.
(3) For a holomorphic map the real Jacobian is $|f'|^2$ (Cauchy–Riemann).
(4) (5.105) with $ad - bc = 1$: $\tau_2' = \tau_2/|c\tau + d|^2$.

The same computation shows that the hyperbolic metric $|\mathrm{d}\tau|^2/\tau_2^2$ is invariant under the whole of $\mathrm{SL}(2, \mathbb{R})$: the Möbius maps with real coefficients are isometries of H, and $\mathrm{d}^2\tau/\tau_2^2$ is its area form.

Worked example 5.79: The volume of $\mathcal{M}_1$

$$\int_F \frac{\mathrm{d}\tau_1\, \mathrm{d}\tau_2}{\tau_2^2} \overset{(1)}{=} \int_{-1/2}^{1/2} \mathrm{d}\tau_1 \int_{\sqrt{1-\tau_1^2}}^{\infty} \frac{\mathrm{d}\tau_2}{\tau_2^2} \overset{(2)}{=} \int_{-1/2}^{1/2} \frac{\mathrm{d}\tau_1}{\sqrt{1 - \tau_1^2}}$$
$$\overset{(3)}{=} \arcsin\tfrac{1}{2} - \arcsin\left(-\tfrac{1}{2}\right) = \frac{\pi}{6} + \frac{\pi}{6} = \frac{\pi}{3}.$$

(1) For fixed $\tau_1 \in [-\frac{1}{2}, \frac{1}{2}]$, F runs from the unit circle $\tau_2 = \sqrt{1 - \tau_1^2}$ up to ∞.
(2) $\int_u^\infty \tau_2^{-2} \mathrm{d}\tau_2 = 1/u$.
(3) $\int \mathrm{d}t/\sqrt{1 - t^2} = \arcsin t$.

Although F reaches up to the cusp, its hyperbolic area is finite: $\pi/3$. (Gauss–Bonnet for the hyperbolic triangle with angles $\pi/3, \pi/3, 0$ gives the same: $\pi - \frac{\pi}{3} - \frac{\pi}{3} - 0 = \frac{\pi}{3}$.)

Worked example 5.80: A function on moduli space

A function on $\mathcal{M}_1$ is a function on $\mathbb{C}^+$ invariant under $\mathrm{SL}(2,\mathbb{Z})$. A natural example is the lattice sum (non-holomorphic Eisenstein series), for $\operatorname{Re} s > 1$,

$$E(\tau, s) = \sum_{(m,n)\neq(0,0)} \frac{\tau_2^s}{|m + n\tau|^{2s}}.$$

It is a sum over the nonzero vectors of the lattice $\mathbb{Z} + \tau\mathbb{Z}$ normalised to unit area. Check invariance under $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$:

$$|m + n\gamma\tau|^2 \overset{(1)}{=} \frac{|m(c\tau + d) + n(a\tau + b)|^2}{|c\tau + d|^2} \overset{(2)}{=} \frac{|m' + n'\tau|^2}{|c\tau + d|^2}, \quad (m', n') = (md + nb,\ mc + na),$$

$$E(\gamma\tau, s) \overset{(3)}{=} \sum_{(m,n)\neq 0} \frac{\tau_2^s\,|c\tau + d|^{-2s}}{|m' + n'\tau|^{2s}\,|c\tau + d|^{-2s}} \overset{(4)}{=} E(\tau, s).$$

(1) Put everything over the common denominator $c\tau + d$.
(2) Collect the coefficients of 1 and τ.
(3) (5.111) for $\operatorname{Im}(\gamma\tau)^s$, and step 2.
(4) The factors cancel, and $(m, n) \mapsto (m', n')$ is a bijection of $\mathbb{Z}^2 \setminus \{0\}$ (its matrix $\left(\begin{smallmatrix} d & b \\ c & a \end{smallmatrix}\right)$ has determinant 1, so has an integer inverse).

Such lattice sums appear in one-loop amplitudes of strings on tori and in the Casimir energy of a field on a torus.

Physical picture: Modular invariance cuts out the ultraviolet

In field theory a one-loop vacuum amplitude is an integral over the Schwinger proper time t of a particle going round a loop, from $t = 0$ to ∞, and the region $t \to 0$ produces the ultraviolet divergence. For the string, τ_2 plays the role of t, but the integral runs only over F, where $\tau_2 \geq \sqrt{3}/2$. The dangerous region $\tau_2 \to 0$ is not integrated over at all: it is a copy of a region near the cusp, already counted. This is the simplest instance of the statement that string loop amplitudes are ultraviolet finite, and it rests entirely on dividing $\mathcal{T}_1$ by the mapping class group.

5.5.6 Genus $p \geq 2$: Fuchsian groups and orbifold points

For $p \geq 2$ the universal cover is H. Its holomorphic automorphisms are the real Möbius maps [Nash p. 156, (5.58)]

$$z \mapsto \frac{az + b}{cz + d}, \qquad a, b, c, d \in \mathbb{R}, \quad ad - bc = 1,$$

forming $\mathrm{PSL}(2,\mathbb{R}) = \mathrm{SL}(2,\mathbb{R})/\{\pm I\}$. By (5.105) they preserve H, and by Example 5.78 they are isometries of the hyperbolic metric. A genus-p surface is H/G with $G \cong \pi_1(\Sigma)$ a discrete subgroup of $\mathrm{PSL}(2,\mathbb{R})$ acting freely; such discrete groups are called **Fuchsian groups**. The group $\pi_1(\Sigma)$ has $2p$ generators $a_1, b_1, \ldots, a_p, b_p$ and one relation,

$$a_1 b_1 a_1^{-1} b_1^{-1} \cdots a_p b_p a_p^{-1} b_p^{-1} = 1, \tag{5.113}$$

which is read off from the $4p$-gon of Nash's p. 140 (Figure 5.45): walking around its boundary traverses $a_1 b_1 a_1^{-1} b_1^{-1} \cdots$, and the boundary is contractible in the polygon.

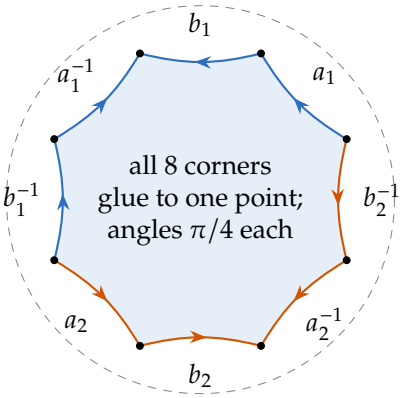


Figure 5.45: A genus-2 surface as a hyperbolic octagon (drawn schematically in the Poincaré disc, whose boundary is the dashed circle) with sides glued in pairs as labelled; the arrows show the gluing directions. Reading the boundary gives the relation $a_1 b_1 a_1^{-1} b_1^{-1} a_2 b_2 a_2^{-1} b_2^{-1} = 1$. In the hyperbolic metric the sides are geodesic arcs and each interior angle is $\pi/4$, so the eight angles add up to 2π around the single glued vertex.

Worked example 5.81: Area of the $4p$-gon and a parameter count

Area. For a hyperbolic polygon with n geodesic sides and interior angles θ_i, Gauss–Bonnet gives area $(n-2)\pi - \sum \theta_i$. For the $4p$-gon all corners glue to one point, so the angles add to 2π:

$$\text{Area} \overset{(1)}{=} (4p-2)\pi - 2\pi \overset{(2)}{=} 4\pi(p-1),$$

(1) $n = 4p$ and $\sum \theta_i = 2\pi$.
(2) Algebra.

which agrees with (5.93) for $K = -1$.
Parameter count. A Fuchsian group for genus p is given by $2p$ matrices in $\mathrm{PSL}(2,\mathbb{R})$, a 3-dimensional group: $6p$ real parameters. The relation (5.113) is an equation in $\mathrm{PSL}(2,\mathbb{R})$: 3 real conditions. Conjugating all generators by the same element of $\mathrm{PSL}(2,\mathbb{R})$ is a change of coordinate on H and gives the same surface: 3 more parameters to remove. So

$$\dim_{\mathbb{R}} \mathcal{T}_p = 6p - 3 - 3 = 6p - 6, \qquad \dim_{\mathbb{C}} \mathcal{T}_p = 3p - 3.$$

This is a heuristic count (one must check that the relation's 3 conditions are independent and that the conjugation acts with finite stabilisers, which uses $p \geq 2$), but it anticipates the index theorem result of Section 5.6.

Symmetric surfaces and the singularities of $\mathcal{M}_p$

A fixed point of Γ_Σ on $\mathcal{T}_p$ is a Riemann surface with a nontrivial holomorphic automorphism that is not isotopic to the identity (at genus $p \geq 2$ none are, apart from the identity). So the orbifold points of $\mathcal{M}_p$ are exactly the surfaces with symmetries, just as for the torus. Nash quotes three facts [Nash p. 156]:

- For $p \geq 2$ the automorphism group of any Σ is finite, of order at most $84(p-1)$ (the **Hurwitz bound**). We will see in Section 5.7 why it is finite: automorphisms must permute the finitely many Weierstrass points.

- For $p \geq 3$ the singular set of $\mathcal{M}_p$ has (complex) codimension at least two (Griffiths and Harris).
- $\mathcal{T}_p$ is contractible and homeomorphic to $\mathbb{C}^{3p-3}$, and it is the universal covering space of $\mathcal{M}_p$ (in the orbifold sense). For $p = 1$ this says $\mathcal{T}_1 = \mathbb{C}^+$ is contractible and $\pi_1(\mathcal{M}_1) = \mathrm{SL}(2,\mathbb{Z})/\{\pm I\}$.

Worked example 5.82: Surfaces with many symmetries

Every genus-2 surface is symmetric. Every genus-2 Riemann surface can be written as $y^2 = (x - e_1)\cdots(x - e_6)$, and $(x, y) \mapsto (x, -y)$ is an automorphism (the hyperelliptic involution) with six fixed points (Section 5.7). So every point of $\mathcal{M}_2$ has at least a $\mathbb{Z}_2$ stabiliser. This $\mathbb{Z}_2$ acts trivially on all of $\mathcal{T}_2$, which is why it does not produce a singularity by itself; the singularities of $\mathcal{M}_2$ come from surfaces with extra symmetry. This is one reason Nash states the codimension result only for $p \geq 3$.
The Hurwitz bound is sharp. The Klein quartic $x^3y + y^3z + z^3x = 0$ in $\mathbb{CP}^2$ is a smooth plane curve of degree 4, so of genus $\frac{(4-1)(4-2)}{2} = 3$, and it has $168 = 84 \cdot (3-1)$ automorphisms, forming the simple group $\mathrm{PSL}(2, \mathbb{F}_7)$.

5.5.7 Dehn twists

How does one get a handle on Γ_Σ for general genus? Nash describes generators [Nash pp. 156–157].

Working definition 5.44 (Dehn twist)

Take a simple closed curve C on Σ [a loop that does not cross itself] and thicken it to an annulus $A \cong S^1 \times [0,1]$. Cut Σ along one boundary circle of A, twist that end through a full 2π rotation, and glue back. In coordinates (θ, t) on A the map is

$$t_C : (\theta, t) \mapsto (\theta + 2\pi\,\chi(t),\ t),$$

where χ increases smoothly from 0 at $t = 0$ to 1 at $t = 1$; outside A it is the identity. The result is a diffeomorphism of Σ, the **Dehn twist** along C. *Test for nontriviality:* a curve that crosses C once is dragged once around C (Figure 5.46); if C is not contractible [it cannot be shrunk to a point on Σ] this changes the homology class of the crossing curve, so $t_C \notin \mathrm{Diff}_0$.

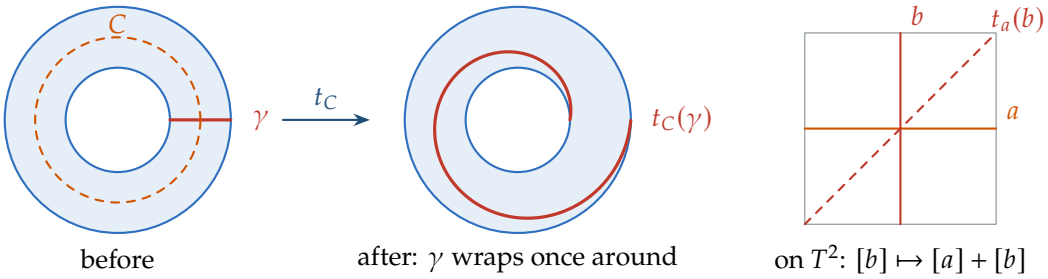


Figure 5.46: A Dehn twist. Left: an annulus around the curve C (dashed orange), crossed by a radial arc γ (red). Middle: after the twist the arc winds once around the annulus; outside the annulus nothing moved. Right: on the torus drawn as a square, the twist along a takes the vertical curve b to the diagonal curve, which is the shear $(x, y) \mapsto (x + y, y)$, i.e. the modular transformation T.

Worked example 5.83: Dehn twists generate the modular group

On $T^2 = \mathbb{R}^2/\mathbb{Z}^2$ let a be the horizontal curve and b the vertical one. The twist along a is (isotopic to) the shear $(x, y) \mapsto (x + y, y)$ and the twist along b is $(x, y) \mapsto (x, y - x)$. By (5.108) their matrices are

$$t_a = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = T, \qquad t_b = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}.$$

Then

$$t_a\, t_b\, t_a \overset{(1)}{=} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$
$$\overset{(2)}{=} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = -S.$$

(1) Multiply the first two matrices.
(2) Multiply by the third; compare with (5.109).

Since $-S$ and S are the same element of $\mathrm{PSL}(2,\mathbb{Z})$, and S and T generate, the two Dehn twists t_a, t_b generate the whole mapping class group of the torus. (The relation $t_a t_b t_a = t_b t_a t_b$, the braid relation, is also easy to verify by multiplying matrices.)

For genus p, the Dehn twists along the curves that go around a handle, through a handle, and between adjacent handles generate Γ_Σ (Dehn, Lickorish: $3p - 1$ twists suffice; Humphries: $2p + 1$ is the minimum). Nash notes that the set of all Dehn twists is far from a minimal generating set, and that the important curves are the non-contractible ones: those around a handle and, for surfaces with punctures, those around a puncture. A twist along a contractible curve is isotopic to the identity.

Physical picture: Dehn twists in string theory

Cut a worldsheet along a closed curve, rotate one side by 2π and reglue: nothing local has changed, so the string amplitude cannot notice. This is the statement that the amplitude must be invariant under Γ_Σ. In the operator language on the torus, the twist T is implemented by $e^{2\pi i(L_0 - \bar{L}_0)}$, and invariance under it is the level-matching condition $L_0 = \bar{L}_0$ of closed strings (Chapter 11).

Nash closes the section by pointing ahead: in chapter 9 Teichmüller space receives a natural Riemannian metric, the **Weil–Petersson metric**, induced from the metric on $\mathrm{Met}(\Sigma)$, and with respect to it $\mathcal{T}_p$ is a Kähler manifold. For $p = 1$ it is the hyperbolic metric $|d\tau|^2/\tau_2^2$ of Example 5.78, whose area form we just integrated over F.

Reading Nash (pp. 152–157, (5.44)–(5.58))

Nash defines $\mathcal{M}_p$ (5.44) and the mapping class group (5.45), notes that $\mathcal{M}_p$ is an orbifold, and uses uniformisation (5.46) to compute $\mathcal{M}_0$ and $\mathcal{M}_1$. For the torus he lifts a conformal map (5.48)–(5.49), derives the $\mathrm{SL}(2,\mathbb{Z})$ criterion (5.50)–(5.54), identifies the modular group inside the Möbius group (5.55), obtains $\mathcal{M}_1 = \mathbb{C}^+/(\mathrm{SL}(2,\mathbb{Z})/\{\mp I\})$ (5.56), writes the torus automorphisms (5.57), and mentions S, T, their fixed points and the fixed point δ_{ij} in $\mathrm{Met}_{\mathrm{const}}$. He then sketches genus ≥ 2 via $\mathrm{SL}(2,\mathbb{R})$ (5.58) and Fuchsian groups, states the orbifold, Hurwitz and contractibility facts, and defines Dehn twists. We filled in: why the lift is affine (Liouville), the imaginary part formula, the matrix relations of S, T, the

fundamental domain and reduction algorithm, the invariant measure and $\mathrm{vol}(F) = \pi/3$, a modular function, the $4p$-gon area and parameter count, explicit symmetric surfaces, and the torus Dehn twists as matrices.

Recap

- $\mathcal{M}_p = \mathrm{Met}_{\mathrm{const}}/\mathrm{Diff}^+ = \mathcal{T}_p/\Gamma_\Sigma$, with $\Gamma_\Sigma = \mathrm{Diff}^+/\mathrm{Diff}_0$ the (discrete) mapping class group. It forgets the labelling of curves.
- $\mathcal{M}_0$ is a point: every Möbius map has a fixed point.
- Tori λ_1, λ_2 are equivalent iff $\lambda_2 = (a\lambda_1 + b)/(c\lambda_1 + d)$ with $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}(2,\mathbb{Z})$, so $\mathcal{M}_1 = \mathbb{C}^+/\mathrm{PSL}(2,\mathbb{Z})$, with fundamental domain F.
- $S : \tau \mapsto -1/\tau$ and $T : \tau \mapsto \tau + 1$ generate; $S^2 = (ST)^3 = -I$. Orbifold points at $\tau = i$ ($\mathbb{Z}_2$) and $\tau = e^{i\pi/3}$ ($\mathbb{Z}_3$); a cusp at $i\infty$.
- $\mathrm{d}^2\tau/\tau_2^2$ is the invariant measure; $\mathrm{vol}(F) = \pi/3$; integrating over F removes the UV region of one-loop amplitudes.
- $p \geq 2$: $\Sigma = H/G$, G Fuchsian; $\dim_{\mathbb{C}} = 3p - 3$ by a parameter count; finitely many automorphisms ($\leq 84(p-1)$); Dehn twists generate Γ_Σ.

5.6 The dimension of the moduli space

In Section 5.5 we found $\mathcal{M}_0$ (a point) and $\mathcal{M}_1$ (the upper half-plane modulo $\mathrm{SL}(2,\mathbb{Z})$, one complex dimension) by hand, and for $p \geq 2$ we only *counted parameters* of Fuchsian groups (Example 5.81) to guess $3p - 3$. This section derives

$$\dim_{\mathbb{C}} \mathcal{M}_p \;=\; 3p - 3 \qquad (p \geq 2)$$

properly. The method is the one we will use again and again in this book (for instantons in Chapter 8, for the string measure in Chapter 9):

(i) write down the *tangent space* to the moduli space as "all infinitesimal deformations modulo infinitesimal gauge transformations";

(ii) recognise that quotient as the cohomology of an elliptic operator;

(iii) compute its dimension with an index theorem plus a vanishing theorem.

Physical picture: Why a string theorist counts moduli

After Weyl and diffeomorphism invariance have been used to fix the worldsheet metric, the Polyakov integral $\int \mathcal{D}g$ of (5.74) becomes an ordinary integral over $\mathcal{M}_p$. The number of integrals left over is $\dim \mathcal{M}_p$. In the Faddeev–Popov language of Chapter 9 the gauge fixing introduces ghosts b and c. The zero modes of c are conformal Killing vectors (leftover gauge symmetry), and the zero modes of b are the *quadratic differentials* we are about to meet (leftover moduli). The difference

$$\#(b \text{ zero modes}) - \#(c \text{ zero modes}) \;=\; 3p - 3$$

is exactly the index we are going to compute. For the torus it is $1 - 1 = 0$; for the sphere $0 - 3 = -3$.

5.6.1 The strategy: count deformations of the metric

Nash starts from the form (5.76) of an arbitrary metric [Nash p. 157, (5.59)],

$$ds^2 = \lambda\, |dz + \mu\, d\bar{z}|^2 ,$$

in which the positive function λ is the Weyl factor and the complex function μ labels the complex structure (Section 5.4.2). A point of $\mathcal{M}_p$ is a metric up to Weyl rescalings and diffeomorphisms. A tangent vector to $\mathcal{M}_p$ at the point $[g]$ is therefore an infinitesimal change δg modulo the changes produced by infinitesimal Weyl rescalings and infinitesimal diffeomorphisms. Since Weyl rescalings only move λ, the tangent vector is carried by $\delta\mu$:

$$T_{[g]}\mathcal{M}_p = \{\textit{perturbations } \delta\mu\} / \{\textit{those produced by diffeomorphisms}\}.$$

If we can compute the dimension of this quotient we have $\dim T_{[g]}\mathcal{M}_p = \dim \mathcal{M}_p$.

Key idea: A heuristic count that already smells of an index

Count real functions. A metric on a surface has 3 components (g_{xx}, g_{xy}, g_{yy}). A Weyl rescaling removes 1. A diffeomorphism is generated by a vector field with 2 real components. So

$$3 - 1 - 2 = 0 \quad \text{functions' worth of moduli.}$$

"Zero functions" does not mean "nothing". It means that what survives is a *finite* number of parameters, and a finite number that arises as "functions minus functions" is precisely what an index measures: the difference between the kernel and the cokernel of an operator that maps 2 real functions (a vector field) to 2 real functions (a $\delta\mu$).

To make this precise we need to know how λ and μ behave when we change coordinates, so that we know what kind of global object $\delta\mu$ is.

5.6.2 How λ and μ transform: Beltrami differentials

Suppose we change to a new holomorphic coordinate $w = f(z)$ (so $f_{\bar{z}} = 0$ and $f_z \neq 0$). The line element ds^2 is a geometric object, so it must look the same in both coordinates:

$$\lambda(z)\, |dz + \mu(z)\, d\bar{z}|^2 \;=\; \tilde{\lambda}(w)\, |dw + \tilde{\mu}(w)\, d\bar{w}|^2 .$$

Here $\tilde{\lambda}$ and $\tilde{\mu}$ are the functions that describe the metric in the w chart. Nash writes them as $\lambda(f(z))$ and $\mu(f(z))$, which is a common abuse of notation. Let us find the relation.

$$\begin{aligned} \tilde{\lambda}\, |dw + \tilde{\mu}\, d\bar{w}|^2 &\overset{(1)}{=} \tilde{\lambda}\, \left| f_z\, dz + \tilde{\mu}\, \overline{f_z}\, d\bar{z} \right|^2 \\ &\overset{(2)}{=} \tilde{\lambda}\, |f_z|^2 \left| dz + \tilde{\mu}\, \frac{\overline{f_z}}{f_z}\, d\bar{z} \right|^2 . \end{aligned} \tag{5.114}$$

(1) Chain rule: $dw = f_z\, dz$ because f is holomorphic, and $d\bar{w} = \overline{f_z}\, d\bar{z}$ by complex conjugation.
(2) Pull the factor f_z out of the absolute value.

Comparing with the left side (the decomposition (5.76) is unique, since (5.79) determines λ and μ from the metric) gives Nash's transformation laws [Nash p. 157, (5.60)]:

$$\tilde{\lambda}(f(z))\, f_z\, \overline{f_z} = \lambda(z), \qquad \tilde{\mu}(f(z))\, \frac{\overline{f_z}}{f_z} = \mu(z). \tag{5.115}$$

The first says that $\lambda\, dz\, d\bar{z}$ is invariant: λ is the component of a $(1,1)$ tensor, a density. The second is more interesting.

Working definition 5.45 (Beltrami differential)

A **Beltrami differential** on a Riemann surface is an object that in each holomorphic chart is written

$$\mu(z,\bar z)\,\frac{\partial}{\partial z}\otimes \mathrm{d}\bar z, \tag{5.116}$$

with the components in two charts related by $\tilde\mu(w)\,\overline{f_z}/f_z = \mu(z)$ [Nash p. 157, (5.61)]. *Operationally:* it is a $(0,1)$-form ($\propto \mathrm{d}\bar z$) whose coefficient is a holomorphic tangent vector ($\propto \partial_z$). To test whether a locally given function μ defines one, check the transformation law on every overlap.

Let us verify that the tensor (5.116) is chart independent exactly when (5.115) holds.

$$\tilde\mu(w)\,\frac{\partial}{\partial w}\otimes \mathrm{d}\bar w \overset{(1)}{=} \tilde\mu(w)\,\frac{1}{f_z}\frac{\partial}{\partial z}\otimes \overline{f_z}\,\mathrm{d}\bar z$$
$$\overset{(2)}{=} \left(\tilde\mu(w)\frac{\overline{f_z}}{f_z}\right)\frac{\partial}{\partial z}\otimes \mathrm{d}\bar z \overset{(3)}{=} \mu(z)\,\frac{\partial}{\partial z}\otimes \mathrm{d}\bar z.$$

(1) $\partial_z = f_z\,\partial_w$ (chain rule, f holomorphic), so $\partial_w = f_z^{-1}\partial_z$; and $\mathrm{d}\bar w = \overline{f_z}\,\mathrm{d}\bar z$.
(2) Collect the scalar factors.
(3) The transformation law (5.115).

Two remarks about (5.115).

- $|\overline{f_z}/f_z| = 1$, so $|\mu|$ is a genuine function on Σ. That fits Figure 5.40: $|\mu|$ fixes the eccentricity of the ellipse field, which cannot depend on the chart. Only the *direction* of the ellipse, encoded in $\arg\mu$, rotates with the chart.
- $\mu = 0$ is preserved by holomorphic coordinate changes. A metric that is conformally flat in one holomorphic chart is conformally flat in all of them.

The Beltrami equation for maps

Nash adds a parenthetical remark [Nash p. 158]: for a general metric with parameters λ,μ, a diffeomorphism $f(z,\bar z)$ is conformal (onto a surface with the standard flat structure $|\mathrm{d}w|^2$) exactly when

$$\frac{f_{\bar z}}{f_z} = \mu. \tag{5.117}$$

This is the Beltrami equation (5.81) of Section 5.4.2, now read as an equation for a map instead of for a coordinate: $\mathrm{d}f = f_z(\mathrm{d}z + \mu\,\mathrm{d}\bar z)$, so $|\mathrm{d}f|^2 = |f_z|^2|\mathrm{d}z+\mu\,\mathrm{d}\bar z|^2$ is a multiple of $\mathrm{d}s^2$.

Worked example 5.84: Solving the Beltrami equation for a constant μ

Take the square torus $\mathbb{C}/(\mathbb{Z}+i\mathbb{Z})$ with coordinate z and a *constant* Beltrami coefficient μ, $|\mu|<1$. Try the affine map $f(z) = z + \mu\bar z$:

$$f_z = 1, \qquad f_{\bar z} = \mu, \qquad \frac{f_{\bar z}}{f_z} = \mu.\checkmark$$

So f is conformal from $(\Sigma, \lambda|\mathrm{d}z+\mu\mathrm{d}\bar z|^2)$ to the flat plane. It sends the lattice generators to

$$f(1) = 1+\mu, \qquad f(i) = i - i\mu = i(1-\mu),$$

so the surface is conformally the flat torus with modular parameter

$$\tau \overset{(1)}{=} \frac{f(i)}{f(1)} = \frac{i(1-\mu)}{1+\mu}.$$

(1) The torus $\mathbb{C}/(\mathbb{Z}\omega_1 + \mathbb{Z}\omega_2)$ is conformal to $\mathbb{C}/(\mathbb{Z} + \mathbb{Z}\,\omega_2/\omega_1)$ by the rescaling $z \mapsto z/\omega_1$ (Section 5.5).

Solving $\tau(1+\mu) = i(1-\mu)$ for μ gives $\mu = (i-\tau)/(i+\tau)$: exactly the Cayley transform found in Example 5.66 from the opposite direction. Two independent calculations agree. Note $|\mu| < 1$ guarantees $\operatorname{Im}\tau > 0$, as it must.

5.6.3 Perturbing the metric

Now we perturb. Choose isothermal coordinates for the unperturbed metric g (Section 5.4.2), so that locally

$$\mathrm{d}s^2 = \lambda\ |\mathrm{d}z|^2 = \lambda\,\mathrm{d}z\,\mathrm{d}\bar{z}.$$

A nearby metric has a slightly different Weyl factor and a small Beltrami coefficient:

$$\mathrm{d}\tilde{s}^2 = (\lambda + \delta\lambda)\ |\mathrm{d}z + \delta\mu\,\mathrm{d}\bar{z}|^2\,.$$

Expand to first order in small quantities:

$$\begin{aligned}\delta(\mathrm{d}s^2) := \mathrm{d}\tilde{s}^2 - \mathrm{d}s^2 &\overset{(1)}{=} (\lambda+\delta\lambda)\big(\mathrm{d}z\,\mathrm{d}\bar{z} + \delta\mu\,\mathrm{d}\bar{z}^2 + \overline{\delta\mu}\,\mathrm{d}z^2 + |\delta\mu|^2\,\mathrm{d}z\,\mathrm{d}\bar{z}\big) - \lambda\,\mathrm{d}z\,\mathrm{d}\bar{z} \\ &\overset{(2)}{=} \delta\lambda\,\mathrm{d}z\,\mathrm{d}\bar{z} + \lambda\,\delta\mu\,\mathrm{d}\bar{z}\,\mathrm{d}\bar{z} + \lambda\,\overline{\delta\mu}\,\mathrm{d}z\,\mathrm{d}z + O(\delta^2). \end{aligned} \tag{5.118}$$

(1) Multiply out $|\mathrm{d}z + \delta\mu\,\mathrm{d}\bar{z}|^2$ as in (5.77).
(2) Drop products of two small quantities ($\delta\lambda\,\delta\mu$, $|\delta\mu|^2$, …); the $\lambda\,\mathrm{d}z\,\mathrm{d}\bar{z}$ terms cancel.

This is Nash's formula [Nash p. 158] for $\delta(\mathrm{d}s^2)$. Now read off the components. In complex coordinates a symmetric tensor is written $\delta(\mathrm{d}s^2) = \delta g_{ab}\,\mathrm{d}x^a\mathrm{d}x^b$ with $a, b \in \{z, \bar{z}\}$:

$$\delta(\mathrm{d}s^2) = \delta g_{zz}\,\mathrm{d}z^2 + 2\,\delta g_{z\bar{z}}\,\mathrm{d}z\,\mathrm{d}\bar{z} + \delta g_{\bar{z}\bar{z}}\,\mathrm{d}\bar{z}^2.$$

Comparing with (5.118):

$$\delta g_{z\bar{z}} = \tfrac{1}{2}\,\delta\lambda, \qquad \delta g_{\bar{z}\bar{z}} = \lambda\,\delta\mu, \qquad \delta g_{zz} = \lambda\,\overline{\delta\mu}. \tag{5.119}$$

Since $g_{z\bar{z}} = \lambda/2$ ((5.87)),

$$\boxed{\delta\mu = \frac{\delta g_{\bar{z}\bar{z}}}{\lambda} = \frac{\delta g_{\bar{z}\bar{z}}}{2\,g_{z\bar{z}}}} \tag{5.120}$$

This is Nash's (5.62) [Nash p. 158]. The Weyl factor sits entirely in $\delta g_{z\bar{z}}$, the trace part, and the complex structure entirely in $\delta g_{\bar{z}\bar{z}}$.

Watch out: A factor of 2 in (5.62) and (5.72)

Nash writes $\delta\mu = \delta g_{\bar{z}\bar{z}}/g_{z\bar{z}}$. This is right if "$g_{z\bar{z}}$" means the coefficient λ of $\mathrm{d}z\,\mathrm{d}\bar{z}$ in $\mathrm{d}s^2$. With the standard tensor convention $\mathrm{d}s^2 = g_{ab}\mathrm{d}x^a\mathrm{d}x^b$ one has $g_{z\bar{z}} = g_{\bar{z}z} = \lambda/2$, and then $\delta\mu = \delta g_{\bar{z}\bar{z}}/(2g_{z\bar{z}})$ as in (5.120). The same factor reappears in his (5.72), $\delta\mu = 2\,\partial_{\bar{z}}V^z$, where we will get $\delta\mu = \partial_{\bar{z}}V^z$. Nash himself notes that "the factor 2 in 5.72 is not of course significant": it only rescales V and cannot change any dimension count.

What $\delta\mu$ means in Cartesian language

It helps to see $\delta g_{\bar z\bar z}$ in terms of the familiar components $\delta g_{xx}, \delta g_{xy}, \delta g_{yy}$. Since $x = (z+\bar z)/2$ and $y = (z-\bar z)/(2i)$, we have $\partial x/\partial \bar z = \frac12$ and $\partial y/\partial \bar z = \frac{i}{2}$. Then

$$\begin{aligned}\delta g_{\bar z\bar z} &\overset{(1)}{=} \frac{\partial x^a}{\partial \bar z}\frac{\partial x^b}{\partial \bar z}\,\delta g_{ab}\\ &\overset{(2)}{=} \tfrac14\,\delta g_{xx} + 2\cdot\tfrac{i}{4}\,\delta g_{xy} + \left(\tfrac{i}{2}\right)^2\delta g_{yy}\\ &= \tfrac14\big(\delta g_{xx} - \delta g_{yy} + 2i\,\delta g_{xy}\big).\end{aligned} \tag{5.121}$$

(1) Tensor transformation law for the components of a symmetric 2-tensor from (x, y) to $(z, \bar z)$.
(2) Sum over $a, b \in \{x, y\}$; the mixed terms xy and yx are equal.

So $\delta\mu$ is (up to $1/4\lambda$) the complex combination of the two *traceless* components of δg: the "stretch in x minus stretch in y" and the "shear". The trace $\delta g_{xx} + \delta g_{yy}$ does not appear. This is the precise meaning of "Weyl rescalings do not change the complex structure" at the infinitesimal level.

Worked example 5.85: Stretching the plane

Take the flat metric $\mathrm{d}x^2 + \mathrm{d}y^2$ ($\lambda = 1$) and stretch it in the x direction: $\mathrm{d}\tilde s^2 = (1+\epsilon)\mathrm{d}x^2 + \mathrm{d}y^2$. Then $\delta g_{xx} = \epsilon$ and the other components vanish, so by (5.121) and (5.120)

$$\delta\mu = \frac{\delta g_{\bar z\bar z}}{\lambda} = \frac{\epsilon}{4}.$$

Cross-check with the exact formula (5.79) of Section 5.4.2, with $A = 1+\epsilon$, $B = 0$, $C = 1$:

$$\mu \overset{(1)}{=} \frac{A - C + 2iB}{A + C + 2\sqrt{AC - B^2}} = \frac{\epsilon}{2 + \epsilon + 2\sqrt{1+\epsilon}} \overset{(2)}{=} \frac{\epsilon}{4} + O(\epsilon^2).$$

(1) (5.79), written as $\mu = (A - C + 2iB)/(4\lambda)$.
(2) $2\sqrt{1+\epsilon} = 2 + \epsilon + O(\epsilon^2)$, so the denominator is $4 + O(\epsilon)$.

Real positive μ squashes the unit circle along the x axis ($\varphi = \theta/2 = 0$ in Figure 5.40): with the metric $(1+\epsilon)\mathrm{d}x^2 + \mathrm{d}y^2$ a unit vector along x is indeed shorter in coordinate terms. A pure shear $\delta g_{xy} = \epsilon/2$ gives instead $\delta\mu = i\epsilon/4$, an ellipse tilted at 45°.

$\delta\mu$ as a global object

Nash then writes $\delta g_{ab} = \epsilon\, h_{ab}$ with ϵ small, so [Nash p. 158, (5.63)]

$$\delta\mu = \epsilon\,\frac{h_{\bar z\bar z}}{g_{z\bar z}}$$

(in his normalisation). The ratio $h_{\bar z\bar z}/g_{z\bar z}$ has the same tensor character as μ itself: $h_{\bar z\bar z}$ carries two lower $\bar z$ indices, like $\mathrm{d}\bar z \otimes \mathrm{d}\bar z$, and dividing by $g_{z\bar z}$ (like $\mathrm{d}z \otimes \mathrm{d}\bar z$) turns one $\mathrm{d}\bar z$ into a $(\mathrm{d}z)^{-1} = \partial_z$. So the perturbation is a Beltrami differential [Nash p. 158, (5.64)]:

$$\nu(z,\bar z)\,\frac{\partial}{\partial z} \otimes \mathrm{d}\bar z. \tag{5.122}$$

To say which space these objects live in we need two line bundles.

Working definition 5.46 (The canonical bundle K_Σ and its dual K_Σ^*)

On a Riemann surface, the **canonical bundle** $K_\Sigma = T^*\Sigma$ (holomorphic cotangent bundle) is the complex line bundle whose local sections are $\phi(z)\,dz$. On an overlap with $w = f(z)$ the local frame changes by $dw = f_z\,dz$, so the transition function is the holomorphic function f_z. Its dual $K_\Sigma^* = T\Sigma$ (holomorphic tangent bundle) has local sections $V^z(z,\bar z)\,\partial_z$ and transition function $1/f_z$. More generally K_Σ^q has sections $\phi(z)\,dz^{\otimes q}$ and transition functions $(f_z)^q$, for any $q \in \mathbb{Z}$ ($K^{-1} = K^*$). *Operationally:* to decide which K^q a local expression belongs to, count powers of dz (with ∂_z counting as -1).

A Beltrami differential is a $(0,1)$-form with values in K_Σ^*. We write the space of smooth ones as $\Omega^{0,1}(K_\Sigma^*)$. Nash's (5.65) [Nash p. 159] is the same statement in local form:

$$\nu\,d\bar z \;\equiv\; \nu(z,\bar z)\,\frac{\partial}{\partial z}\otimes d\bar z. \tag{5.123}$$

Key idea: Every Beltrami differential is $\bar\partial$-closed

On a complex curve there are no $(0,2)$-forms: $d\bar z \wedge d\bar z = 0$. So $\bar\partial$ of any $(0,1)$-form, with or without values in a bundle, is zero. Hence every perturbation $\delta\mu$ lies in the kernel of

$$\bar\partial^{(0,1)}_{K_\Sigma^*} :\ \Omega^{0,1}(K_\Sigma^*) \to \Omega^{0,2}(K_\Sigma^*) = 0.$$

In Nash's words: any element of $\operatorname{Ker}\bar\partial^{(0,1)}_{K_\Sigma^*}$ gives rise to a perturbation in μ, and conversely.

Dictionary: Several names, one object

$\mu\,\partial_z\otimes d\bar z$	Beltrami differential (complex analysis)
$\Omega^{0,1}(K_\Sigma^*)$ or $\Omega^{0,1}(T\Sigma)$	$(0,1)$-forms with values in the tangent bundle
$\delta g_{\bar z\bar z}$, δg_{zz}	traceless part of δg_{ab} (relativity)
$\delta\mu$	infinitesimal change of complex structure
$\int_\Sigma \delta g_{ab}\,T^{ab}$, $\mu\,T_{zz}$	how $\delta\mu$ couples to the stress tensor (CFT)

5.6.4 Removing the diffeomorphisms

Not every $\delta\mu$ changes the point in $\mathcal{M}_p$ (or $\mathcal{T}_p$). If the new metric is just the old one moved around by a diffeomorphism close to the identity, we have gone nowhere. We must find these "fake" perturbations and quotient them out.

The Lie algebra of $\mathrm{Diff}_0(\Sigma)$

An element of $\mathrm{Diff}_0(\Sigma)$ near the identity is generated by a vector field. Nash writes $\mathrm{vect}(\Sigma)$ for the space of smooth vector fields on Σ; it is the Lie algebra of $\mathrm{Diff}_0(\Sigma)$ (and has nothing to do with $\mathrm{Vect}(\Sigma)$, the set of vector bundles over Σ used in K-theory). Given $V \in \mathrm{vect}(\Sigma)$ we "exponentiate" it to a one-parameter family of diffeomorphisms $\alpha_t = \exp(tV)$ [Nash p. 159, (5.66)]:

$$\alpha_t : \Sigma \to \Sigma, \qquad x \mapsto x(t), \tag{5.124}$$

where $x(t)$ is the point reached at time t by moving along the integral curve of V that starts at x. So α_t is the flow of V, the same construction as the time evolution generated by a Hamiltonian vector field in classical mechanics.

The flow acts on metrics [Nash p. 159, (5.67)]:

$$A_t : \mathrm{Diff}_0(\Sigma) \times \mathrm{Met}(\Sigma) \to \mathrm{Met}(\Sigma), \qquad (\alpha_t, g) \mapsto \alpha_t \cdot g, \tag{5.125}$$

with Nash's definition $(\alpha_t \cdot g)(x) = g(\alpha_t(x)) = g(x(t))$: we compare lengths using the metric at the moved point, carried back to x (this is the pull-back $\alpha_t^* g$; Nash calls it a push-forward, and the difference is only $t \to -t$, which does not matter here). As t varies, $g(t) := \alpha_t \cdot g$ is a curve in $\mathrm{Met}(\Sigma)$ through $g(0) = g$ (Figure 5.47). Taylor expanding [Nash p. 159, (5.68)–(5.69)],

$$g(t) = g(0) + t\,\mathcal{L}_V g(0) + \cdots, \qquad g'(0) = \mathcal{L}_V g, \tag{5.126}$$

where $\mathcal{L}_V$ is the Lie derivative (Chapter T1): the rate of change of a tensor as it is dragged along the flow. So $\mathcal{L}_V g$ is the velocity of the curve $g(t)$ at g, a tangent vector to the space of metrics [Nash p. 160, (5.70)]:

$$\mathcal{L}_V\, g \in T_g \mathrm{Met}(\Sigma). \tag{5.127}$$

Nash's finite-dimensional comparison: for a curve $x(t)$ in a manifold generated by a tangent vector $Y = Y^i \partial_i$, the coordinate functions change at the rate $\mathcal{L}_Y x^j = Y(x^j) = Y^i\, \partial x^j/\partial x^i = Y^j$. The Lie derivative of the coordinate functions *is* the velocity. Here the "point" is a whole metric, and its "velocity" is the tensor $\mathcal{L}_V g$.

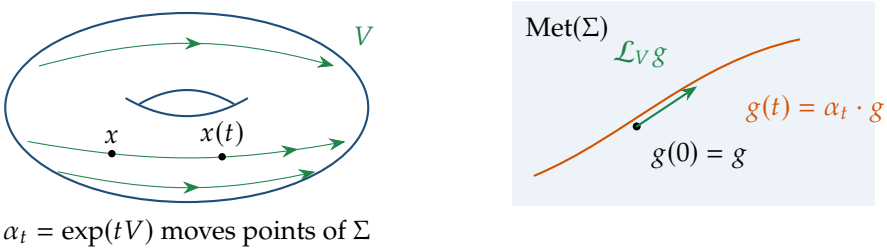


Figure 5.47: Left: a vector field V on Σ generates a flow α_t that slides points along its integral curves. Right: acting on a metric, the flow traces a curve $g(t)$ in the infinite-dimensional space $\mathrm{Met}(\Sigma)$. Its velocity at $t = 0$ is $\mathcal{L}_V g$ (5.68)–(5.70). The curve lies in a single Diff_0-orbit, so every point on it describes the same geometry.

The perturbations produced by a vector field

So we must quotient out every $\delta\mu$ that comes from a $\delta g_{ab} = \mathcal{L}_V g_{ab}$. By (5.120) these are

$$\delta\mu = \frac{(\mathcal{L}_V g)_{z\bar z}}{2 g_{z\bar z}}.$$

The Lie derivative of a symmetric 2-tensor is [Nash p. 160, (5.71)] (this is (5.86))

$$g'_{ab}(0) = \mathcal{L}_V g_{ab} = \frac{\partial g_{ab}}{\partial x^i} V^i + g_{ib} \frac{\partial V^i}{\partial x^a} + g_{ai} \frac{\partial V^i}{\partial x^b}. \tag{5.128}$$

Take $a = b = \bar z$ and sum over $i \in \{z, \bar z\}$, writing $V^z_{,\bar z} = \partial_{\bar z} V^z$ and so on:

$$\begin{aligned}(\mathcal{L}_V g)_{\bar z\bar z} &\overset{(1)}{=} g_{\bar z\bar z, z} V^z + g_{\bar z\bar z, \bar z} V^{\bar z} + g_{z\bar z} V^z_{,\bar z} + g_{\bar z\bar z} V^{\bar z}_{,\bar z} + g_{\bar z z} V^z_{,\bar z} + g_{\bar z\bar z} V^{\bar z}_{,\bar z} \\ &\overset{(2)}{=} 2\, g_{z\bar z}\, \partial_{\bar z} V^z.\end{aligned} \tag{5.129}$$

(1) (5.128): two terms from $V^i \partial_i g_{\bar z\bar z}$, two from $g_{i\bar z}\partial_{\bar z} V^i$, two from $g_{\bar z i}\partial_{\bar z} V^i$. This is the six-term numerator displayed by Nash on p. 160.

(2) In isothermal coordinates the only non-zero components are $g_{z\bar{z}} = g_{\bar{z}z}$, so four of the six terms vanish (including the derivatives of $g_{\bar{z}\bar{z}} \equiv 0$).

Therefore

$$\boxed{\delta\mu = \partial_{\bar{z}} V^z} \qquad \big(\text{Nash's (5.72): } \delta\mu = 2V^z_{,\bar{z}} = 2\partial_{\bar{z}} V^z\big). \tag{5.130}$$

For completeness, the trace part of the same Lie derivative is

$$\begin{aligned}(\mathcal{L}_V g)_{z\bar{z}} &\overset{(1)}{=} V^z \partial_z g_{z\bar{z}} + V^{\bar{z}} \partial_{\bar{z}} g_{z\bar{z}} + g_{z\bar{z}}\, \partial_z V^z + g_{z\bar{z}}\, \partial_{\bar{z}} V^{\bar{z}} \\ &\overset{(2)}{=} \tfrac{1}{2}\big(V^z \partial_z \lambda + V^{\bar{z}} \partial_{\bar{z}} \lambda + \lambda\, \partial_z V^z + \lambda\, \partial_{\bar{z}} V^{\bar{z}}\big),\end{aligned}$$

(1) (5.128) with $a = z$, $b = \bar{z}$, dropping terms with $g_{zz} = g_{\bar{z}\bar{z}} = 0$.
(2) $g_{z\bar{z}} = \lambda/2$.

which is a change of λ only. It is absorbed by a Weyl rescaling, so it does not constrain anything. The whole content of "quotient by diffeomorphisms" is therefore:

Two Beltrami differentials $\delta\mu$ and $\delta\mu'$ describe the same tangent vector to $\mathcal{T}_p$ if $\delta\mu' - \delta\mu = \partial_{\bar{z}} V^z$ for some smooth vector field V.

Worked example 5.86: A diffeomorphism that fakes a change of shape

On the flat plane ($\lambda = 1$) take the real vector field $V = x\,\partial_x - y\,\partial_y$, which generates $\alpha_t(x, y) = (e^t x, e^{-t} y)$: stretch along x, squeeze along y. In complex notation $V^z = V^x + iV^y = x - iy = \bar{z}$. Then (5.130) gives

$$\delta\mu = \partial_{\bar{z}} \bar{z} = 1 \quad \text{(per unit } t\text{)}.$$

Directly: $\alpha_t^*(\mathrm{d}x^2 + \mathrm{d}y^2) = e^{2t}\mathrm{d}x^2 + e^{-2t}\mathrm{d}y^2 \approx (1 + 2t)\mathrm{d}x^2 + (1 - 2t)\mathrm{d}y^2$. So $\delta g_{xx} = 2t$, $\delta g_{yy} = -2t$, $\delta g_{xy} = 0$, and (5.121) gives

$$\delta\mu \overset{(1)}{=} \tfrac{1}{4}\big(2t - (-2t)\big) = t. \checkmark$$

(1) (5.120) with $\lambda = 1$ and (5.121).

So on the plane a constant $\delta\mu$ is *fake*: it is undone by a linear stretch. On a torus the same $V = \bar{z}\,\partial_z$ is *not* allowed, because $\bar{z}$ is not periodic. That single fact is why the torus has a modulus.

Locally, every $\delta\mu$ is fake

On a disc D, every smooth $\delta\mu$ can be written as $\partial_{\bar{z}} V^z$. The reason is the Green's function of $\partial_{\bar{z}}$, familiar from 2d electrostatics:

$$\partial_{\bar{z}} \frac{1}{\pi z} = \delta^2(z) \qquad \text{so} \qquad V^z(z) = \frac{1}{\pi} \int_D \frac{\delta\mu(w)}{z - w}\, \mathrm{d}^2 w \ \text{ solves } \ \partial_{\bar{z}} V^z = \delta\mu. \tag{5.131}$$

(To check the first formula, integrate $\partial_{\bar{z}}(1/\pi z)$ against a test function and use Stokes: $\oint \mathrm{d}z/z = 2\pi i$ around the origin.) So there are no *local* moduli at all: every deformation of complex structure on a disc is a coordinate change. This matches Section 5.4.2, where we used the Beltrami equation to find isothermal coordinates for any metric. Moduli exist because on a *compact* surface the solution V^z has to be a global section of K_Σ^*, and the integral formula does not produce one. The obstruction is global, and global obstructions to solving $\bar{\partial} V = \mu$ are measured by cohomology.

5.6.5 The tangent space as a Dolbeault cohomology group

Background: the operator $\bar\partial_E$

Let E be a holomorphic line bundle on Σ, with local frames related by holomorphic transition functions $g_{\alpha\beta}$. A smooth section is a collection of local functions s_α with $s_\alpha = g_{\alpha\beta}s_\beta$. Because $\partial_{\bar z} g_{\alpha\beta} = 0$,

$$\partial_{\bar z} s_\alpha = g_{\alpha\beta}\,\partial_{\bar z} s_\beta,$$

so the local $(0,1)$-forms $\partial_{\bar z} s_\alpha\,\mathrm{d}\bar z$ glue to a global E-valued $(0,1)$-form. This defines the operator

$$\bar\partial_E = \bar\partial_E^{(0,0)} :\ \Omega^{0,0}(E) \to \Omega^{0,1}(E), \qquad s \mapsto \partial_{\bar z} s\ \mathrm{d}\bar z.$$

It is a first-order elliptic operator (its symbol is multiplication by $\xi_{\bar z}$, which is invertible for $\xi \neq 0$); chapter 4 (Chapter 4) computed its index.

Working definition 5.47 (Dolbeault cohomology of a line bundle on a curve)

For a holomorphic line bundle E on a Riemann surface,

$$H^{0,0}_{\bar\partial_E}(\Sigma) = \operatorname{Ker}\bar\partial_E^{(0,0)} \ \text{(the holomorphic sections)}, \qquad H^{0,1}_{\bar\partial_E}(\Sigma) = \frac{\operatorname{Ker}\bar\partial_E^{(0,1)}}{\operatorname{Im}\bar\partial_E^{(0,0)}} = \frac{\Omega^{0,1}(E)}{\bar\partial_E\,\Omega^{0,0}(E)}.$$

Operationally: $H^{0,0}$ counts solutions of $\bar\partial s = 0$; $H^{0,1}$ counts the right-hand sides β for which $\bar\partial s = \beta$ has *no* solution, modulo those that do. Their dimensions are finite (ellipticity), and $\dim H^{0,0} - \dim H^{0,1} = \operatorname{index}\bar\partial_E$. By Dolbeault's theorem (Chapter 3), $H^{0,q}_{\bar\partial_E}(\Sigma) \cong H^q(\Sigma, \mathcal{O}(E))$, the sheaf cohomology of the sheaf of holomorphic sections of E.

Nash's identification

Take $E = K_\Sigma^*$. A vector field $V = V^z\partial_z + V^{\bar z}\partial_{\bar z}$ has as its $(1,0)$ part $V^z\partial_z$, a smooth section of K_Σ^*. The fake perturbations $\partial_{\bar z}V^z\,\partial_z \otimes \mathrm{d}\bar z$ are exactly the image of $\bar\partial$ acting on sections of K_Σ^* [Nash p. 160, (5.73)]:

$$\{\text{fake } \delta\mu\} = \operatorname{Im}\bar\partial^{(0,0)}_{K_\Sigma^*}, \qquad \text{locally } V^z_{,\bar z}\,\frac{\partial}{\partial z}\otimes\mathrm{d}\bar z. \tag{5.132}$$

The genuine infinitesimal deformations are the quotient [Nash p. 161, (5.74)–(5.76)]:

$$\boxed{T_{[g]}\mathcal{M}_p \;=\; \frac{\operatorname{Ker}\bar\partial^{(0,1)}_{K_\Sigma^*}}{\operatorname{Im}\bar\partial^{(0,0)}_{K_\Sigma^*}} \;=\; H^{0,1}_{\bar\partial_{K_\Sigma^*}}(\Sigma)} \tag{5.133}$$

(Strictly this is the tangent space to $\mathcal{T}_p$, which equals that of $\mathcal{M}_p$ at points where Γ_Σ acts freely; the dimension is the same everywhere.)

Worked example 5.87: The torus in Fourier modes: a complete calculation

Let us compute $H^{0,1}_{\bar\partial_{K^*}}$ for the square torus $\Sigma = \mathbb{C}/(\mathbb{Z}+i\mathbb{Z})$ with $z = x+iy$, x, y mod 1, by brute force. The transition functions of the torus are translations $z \mapsto z+1$, $z \mapsto z+i$, with derivative 1, so ∂_z is a global frame of K^* and a section is just a doubly periodic function $V^z(x,y)$. Likewise $\partial_z \otimes \mathrm{d}\bar z$ is a global frame for Beltrami differentials. Expand in Fourier

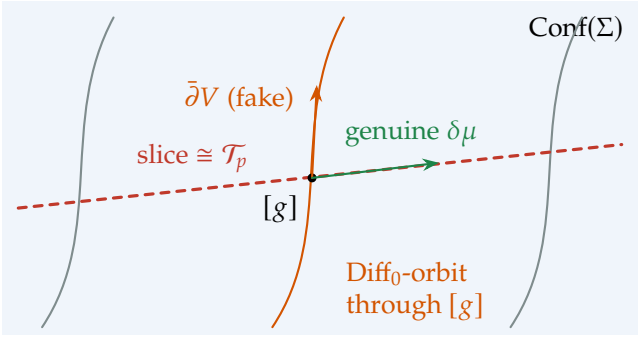


Figure 5.48: The quotient (5.74) as a picture. Each orange curve is one Diff_0-orbit in the space of complex structures: all its points are the same surface. Perturbations tangent to the orbit, $\delta\mu = \bar\partial V$, are fake. The genuine directions are the tangent space modulo the orbit directions, $H^{0,1}(K^*)$; a transverse slice (red dashed) through the orbits is a local model of Teichmüller space.

modes $e_{mn}(x,y) = e^{2\pi i(mx+ny)}$, $(m,n) \in \mathbb{Z}^2$:

$$V^z = \sum_{m,n} v_{mn}\, e_{mn}, \qquad \delta\mu = \sum_{m,n} \mu_{mn}\, e_{mn}.$$

The operator $\bar\partial$ is diagonal in this basis:

$$\partial_{\bar z} e_{mn} \overset{(1)}{=} \tfrac{1}{2}(\partial_x + i\partial_y)\, e^{2\pi i(mx+ny)} \overset{(2)}{=} \tfrac{1}{2}(2\pi i m + i\cdot 2\pi i n)\, e_{mn} \overset{(3)}{=} \pi i\,(m+in)\, e_{mn}.$$

(1) $\partial_{\bar z} = \tfrac{1}{2}(\partial_x + i\partial_y)$, from $x = (z+\bar z)/2$, $y = (z-\bar z)/2i$.
(2) Differentiate the exponential.
(3) $\tfrac{1}{2}(2\pi i m - 2\pi n) = \pi i m - \pi n = \pi i(m+in)$.

So $\bar\partial$ multiplies the (m,n) mode by $\pi i(m+in)$, which is non-zero unless $m = n = 0$ (Figure 5.49). Therefore:

- **Kernel.** $\partial_{\bar z} V^z = 0$ forces $v_{mn} = 0$ for $(m,n) \neq (0,0)$: V^z is constant. $\dim_{\mathbb{C}} \operatorname{Ker} \bar\partial_{K^*} = 1$: the translations, i.e. the conformal Killing vectors of Example 5.71.
- **Image.** Given $\delta\mu$, solve mode by mode: $v_{mn} = \mu_{mn}/(\pi i(m+in))$ for $(m,n) \neq (0,0)$. This is always possible, and the resulting series converges to a smooth V^z because μ_{mn} decays faster than any power. The only obstruction is the zero mode: μ_{00} can never be produced.
- **Cokernel.** $H^{0,1} = \Omega^{0,1}(K^*)/\operatorname{Im}\bar\partial \cong \mathbb{C}$, and the class of $\delta\mu$ is its average, $[\delta\mu] = \mu_{00} = \int_{T^2} \delta\mu\, \mathrm{d}x\, \mathrm{d}y$.

So $\dim_{\mathbb{C}} T\mathcal{T}_1 = 1$, and $\operatorname{index}\bar\partial_{K^*} = 1 - 1 = 0$. Every Beltrami differential on the torus is equivalent to a *constant* one, and the constant is the modulus.

Worked example 5.88: The constant $\delta\mu$ is a change of τ

Take the flat torus $\mathbb{C}/(\mathbb{Z}+\tau\mathbb{Z})$, write points as $z = x + \tau y$ with x, y mod 1, and change $\tau \to \tau + \delta\tau$ while keeping x, y as coordinates. The metric changes from $|\mathrm{d}x + \tau\,\mathrm{d}y|^2 = |\mathrm{d}z|^2$

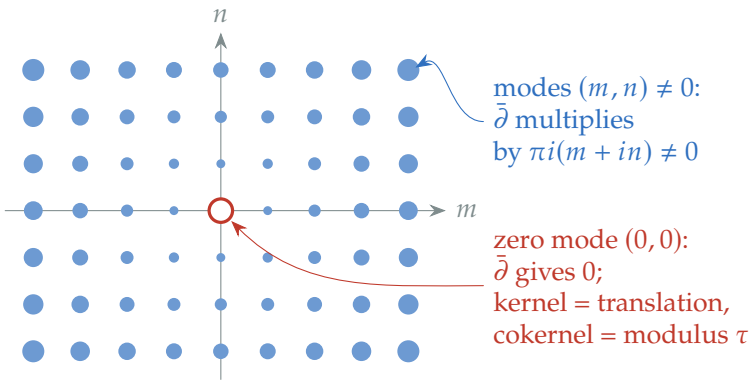


Figure 5.49: The operator $\bar\partial_{K^*}$ on the square torus in the Fourier basis. Each blue dot is a mode $e^{2\pi i(mx+ny)}$; its size shows $|\pi i(m+in)|$, the factor by which $\bar\partial$ rescales it, which never vanishes. Only the zero mode (red circle) is special: a constant V^z is annihilated (a conformal Killing vector) and a constant $\delta\mu$ cannot be reached (a genuine modulus). Kernel and cokernel are both one-dimensional, so the index is 0.

to $|\mathrm{d}x + (\tau+\delta\tau)\mathrm{d}y|^2 = |\mathrm{d}z + \delta\tau\,\mathrm{d}y|^2$. Express $\mathrm{d}y$ in terms of $\mathrm{d}z, \mathrm{d}\bar z$:

$$\mathrm{d}y \overset{(1)}{=} \frac{\mathrm{d}z - \mathrm{d}\bar z}{\tau - \bar\tau} = \frac{\mathrm{d}z - \mathrm{d}\bar z}{2i\tau_2},$$
$$\mathrm{d}z + \delta\tau\,\mathrm{d}y \overset{(2)}{=} \Big(1 + \frac{\delta\tau}{2i\tau_2}\Big)\mathrm{d}z - \frac{\delta\tau}{2i\tau_2}\,\mathrm{d}\bar z,$$
$$\delta\mu \overset{(3)}{=} -\frac{\delta\tau}{2i\tau_2} + O(\delta\tau^2) = \frac{i\,\delta\tau}{2\tau_2}.$$

(1) From $z = x + \tau y$ and $\bar z = x + \bar\tau y$: $z - \bar z = (\tau - \bar\tau)y$.
(2) Substitute.
(3) Divide by the coefficient of $\mathrm{d}z$ to put the metric in the form $\lambda|\mathrm{d}z + \mu\,\mathrm{d}\bar z|^2$, and keep first order.

So $\delta\tau \mapsto [\delta\mu] = i\,\delta\tau/(2\tau_2)$ is a linear isomorphism $T_\tau\mathcal{T}_1 = \mathbb{C} \to H^{0,1}_{\bar\partial_{K^*}}$: Nash's (5.133) for $p = 1$, completely explicitly. (At $\tau = i$ this is $\delta\mu = \frac{i}{2}\delta\tau$, which is also the derivative of the Cayley transform $\mu = (i-\tau)/(i+\tau)$ of Example 5.66 at $\tau = i$: $\mathrm{d}\mu/\mathrm{d}\tau = -2i/(\tau+i)^2 = i/2$.)

5.6.6 Serre duality and quadratic differentials

We now know that $T_{[g]}\mathcal{M}_p = H^{0,1}(K^*)$. This is a quotient space, awkward to compute with. Serre duality replaces it by the dual of a space of honest holomorphic objects, which is much easier to count.

Background: Serre duality

For a holomorphic vector bundle E over a compact complex manifold M of complex dimension n, Serre duality (Chapter 3, and Nash's (5.20), (5.22)) states

$$H^q\big(M, \Omega^p(E)\big) \;\cong\; H^{n-q}\big(M, \Omega^{n-p}(E^*)\big)^*, \tag{5.134}$$

where $\Omega^p(E)$ is the sheaf of holomorphic E-valued p-forms. The pairing that realises it takes a Dolbeault representative θ of type (p,q) and ν of type $(n-p, n-q)$ and integrates their product,

$\int_M \langle \theta, v \rangle$, contracting E with E^*. For a curve ($n = 1$), with $\Omega^0(E) = \mathcal{O}(E)$ and $\Omega^1(E) = \mathcal{O}(K \otimes E)$:

$$H^1\big(\Sigma, \mathcal{O}(E)\big) \;\cong\; H^0\big(\Sigma, \mathcal{O}(K_\Sigma \otimes E^*)\big)^*. \tag{5.135}$$

Applying it to $E = K_\Sigma^*$

Nash's chain [Nash p. 161, (5.77)]:

$$\begin{aligned} H^{0,1}_{\bar\partial_{K_\Sigma^*}}(\Sigma) &\overset{(1)}{=} H^1\big(\Sigma, \mathcal{O}(K_\Sigma^*)\big) \equiv H^1\big(\Sigma, \Omega^0(K_\Sigma^*)\big) \\ &\overset{(2)}{\cong} H^0\big(\Sigma, \Omega^1(K_\Sigma)\big)^* \overset{(3)}{=} H^0\big(\Sigma, \mathcal{O}(K_\Sigma^2)\big)^* \overset{(4)}{=} \big(H^{0,0}_{\bar\partial_{K_\Sigma^2}}(\Sigma)\big)^*. \end{aligned} \tag{5.136}$$

(1) Dolbeault's theorem: $\bar\partial$-cohomology equals sheaf cohomology of holomorphic sections.
(2) Serre duality (5.135) with $E = K^*$, so $E^* = K$.
(3) A holomorphic 1-form with values in K is a holomorphic section of $K \otimes K = K^2$.
(4) H^0 of the sheaf of holomorphic sections is the kernel of $\bar\partial$ on smooth sections.

So [Nash p. 161, (5.78)]:

$$\boxed{T_{[g]}\mathcal{M}_p = H^{0,1}_{\bar\partial_{K_\Sigma^*}}(\Sigma), \qquad T^*_{[g]}\mathcal{M}_p = H^{0,0}_{\bar\partial_{K_\Sigma^2}}(\Sigma)} \tag{5.137}$$

The cotangent space to moduli space is the space of holomorphic sections of K^2.

Working definition 5.48 (Holomorphic quadratic differential)

A **holomorphic quadratic differential** is a holomorphic section of K_Σ^2: in each chart

$$\phi = \phi(z)\,\mathrm{d}z \otimes \mathrm{d}z, \qquad \bar\partial\phi = 0,$$

with $\phi(z) = \tilde\phi(h(z))\,h'(z)^2$ under a holomorphic change of coordinate $w = h(z)$ [Nash p. 162, footnote 6]. *Operationally:* write down a candidate $\phi(z)$ in one chart, transform it to every other chart with the factor $h'(z)^2$, and check it is holomorphic everywhere, including at points added at infinity.

More generally a **holomorphic q-differential** is $\phi = \phi(z)\,\mathrm{d}z^{\otimes q}$, a holomorphic section of K^q, with $\phi(z) = \tilde\phi(h(z))\,h'(z)^q$. Holomorphic 1-differentials are the Abelian differentials of Chapter 5 (there are p of them). The transformation law is the chain rule: $\tilde\phi(w)\,\mathrm{d}w^q = \tilde\phi(h(z))\,(h'(z)\mathrm{d}z)^q$.

Worked example 5.89: Quadratic differentials on the sphere and the torus

Sphere. On $\mathbb{CP}^1 = \mathbb{C} \cup \{\infty\}$ a candidate is $\phi(z)\,\mathrm{d}z^2$ with ϕ entire. Near ∞ use $w = 1/z$, so $z = 1/w$ and $\mathrm{d}z = -w^{-2}\mathrm{d}w$:

$$\phi(z)\,\mathrm{d}z^2 \overset{(1)}{=} \phi(1/w)\,w^{-4}\,\mathrm{d}w^2.$$

(1) $\mathrm{d}z^2 = (-w^{-2}\mathrm{d}w)^2 = w^{-4}\mathrm{d}w^2$.

For this to be holomorphic at $w = 0$ we need $\phi(1/w) = O(w^4)$, i.e. $\phi(z) \to 0$ like z^{-4} as $z \to \infty$. An entire function that tends to 0 at infinity is bounded, hence constant (Liouville), hence zero. So

$$H^0(\mathbb{CP}^1, \mathcal{O}(K^2)) = 0: \quad \text{the sphere has no moduli.}$$

More generally, a holomorphic q-differential with $q \geq 1$ on the sphere vanishes identically,

while for $q \le 0$ ($K^q = (K^*)^{|q|}$) the polynomials of degree $\le 2|q|$ survive: $2|q|+1$ of them. For $q=-1$ these are the three conformal Killing vectors of Example 5.70.
Torus. On $\mathbb{C}/(\mathbb{Z}+\tau\mathbb{Z})$ the translations have $h'=1$, so a quadratic differential is $\phi(z)\,\mathrm{d}z^2$ with ϕ holomorphic and doubly periodic, hence bounded, hence constant:

$$H^0(T^2, \mathcal{O}(K^2)) = \mathbb{C}\cdot \mathrm{d}z^2,$$

one-dimensional, dual to the one-dimensional $H^{0,1}(K^*)$ of Example 5.87. In fact every K^q on the torus is trivial with frame $\mathrm{d}z^q$, and has exactly one holomorphic section.

Nash's example: the duality pairing, explicitly

Nash pauses [Nash p. 162, Example] to show the duality at work. Take a Beltrami differential and a quadratic differential,

$$\mu = \mu(z,\bar z)\,\frac{\partial}{\partial z}\otimes \mathrm{d}\bar z, \qquad \phi = \phi(z)\,\mathrm{d}z\otimes \mathrm{d}z.$$

Contract the ∂_z of μ with one $\mathrm{d}z$ of ϕ ($\langle \partial_z, \mathrm{d}z\rangle = 1$). What is left is $\mu\phi\,\mathrm{d}z\otimes\mathrm{d}\bar z$, a $(1,1)$ object that we integrate as the 2-form $\mu\phi\,\mathrm{d}z\wedge\mathrm{d}\bar z$. We normalise with $\mathrm{d}^2z := \frac{i}{2}\mathrm{d}z\wedge\mathrm{d}\bar z = \mathrm{d}x\wedge\mathrm{d}y$:

$$\langle \mu, \phi\rangle \;:=\; \int_\Sigma \mu(z,\bar z)\,\phi(z)\;\mathrm{d}^2z. \tag{5.138}$$

(Nash writes $\int_\Sigma \mu\phi\,\mathrm{d}z\,\mathrm{d}\bar z$; the constant $i/2$ is immaterial.) First, the integrand is chart independent. Under $w = h(z)$,

$$\begin{aligned}\tilde\mu\,\tilde\phi\;\mathrm{d}w\wedge\mathrm{d}\bar w &\overset{(1)}{=} \Big(\mu\,\frac{h'}{\overline{h'}}\Big)\Big(\phi\,\frac{1}{(h')^2}\Big)\;|h'|^2\;\mathrm{d}z\wedge\mathrm{d}\bar z\\ &\overset{(2)}{=} \mu\,\phi\,\frac{h'\,\overline{h'}}{\overline{h'}\,h'}\;\mathrm{d}z\wedge\mathrm{d}\bar z = \mu\,\phi\;\mathrm{d}z\wedge\mathrm{d}\bar z.\end{aligned}$$

(1) Transformation laws: $\tilde\mu = \mu\,h'/\overline{h'}$ from (5.115); $\tilde\phi = \phi/(h')^2$ from the definition of a quadratic differential; $\mathrm{d}w\wedge\mathrm{d}\bar w = |h'|^2\mathrm{d}z\wedge\mathrm{d}\bar z$.
(2) Collect factors; they cancel exactly.

Second, the pairing vanishes on the fake perturbations, so it is well defined on the quotient $H^{0,1}(K^*)$. If $\mu = \partial_{\bar z}V^z$ then

$$\langle \partial_{\bar z}V^z, \phi\rangle \overset{(1)}{=} \int_\Sigma \partial_{\bar z}\big(V^z\phi\big)\,\mathrm{d}^2z \overset{(2)}{=} -\int_\Sigma V^z\,\partial_{\bar z}\phi\;\mathrm{d}^2z \;+\; 0 \overset{(3)}{=} 0.$$

(1) ϕ is holomorphic, so $\partial_{\bar z}(V^z\phi) = (\partial_{\bar z}V^z)\phi$.
(2) Nash's form of the argument: integrate by parts. The boundary term is zero because Σ is closed. Precisely, $V^z\phi\,\mathrm{d}z$ is a global $(1,0)$-form (a section of $K^*\otimes K^2 = K$), $\mathrm{d}(V^z\phi\,\mathrm{d}z) = \partial_{\bar z}(V^z\phi)\,\mathrm{d}\bar z\wedge\mathrm{d}z$, and Stokes on a surface with $\partial\Sigma=\varnothing$ gives zero.
(3) $\partial_{\bar z}\phi = 0$.

So μ and ϕ "know" about each other in exactly the way Serre duality requires: the pairing ignores the part of μ that is pure gauge.

Why the pairing is perfect: harmonic Beltrami differentials

Nash does not show that the pairing identifies $H^{0,1}(K^*)$ with the *full* dual of $H^0(K^2)$. Here is a concrete way to see it, which also gives the best representative of each class. Choose the metric $\lambda|\mathrm{d}z|^2$ (for instance the constant-curvature one). Given a holomorphic quadratic differential ϕ, set

$$\mu_\phi \;:=\; \frac{\overline{\phi(z)}}{\lambda}\,\frac{\partial}{\partial z}\otimes \mathrm{d}\bar z. \tag{5.139}$$

This is a legitimate Beltrami differential: $\bar\phi\,\mathrm{d}\bar z^2$ divided by $\lambda\,\mathrm{d}z\,\mathrm{d}\bar z$ has exactly the index structure $\mathrm{d}\bar z\otimes(\mathrm{d}z)^{-1}$. Now

$$\langle \mu_\phi,\phi\rangle \overset{(1)}{=} \int_\Sigma \frac{|\phi|^2}{\lambda}\,\mathrm{d}^2z \;>\; 0 \qquad \text{for } \phi\neq 0.$$

(1) (5.138) with (5.139); the integrand is positive wherever $\phi\neq 0$, and a non-zero holomorphic ϕ vanishes only at isolated points.

So no non-zero ϕ pairs to zero with everything: the map $H^{0,1}(K^*)\to H^0(K^2)^*$ is onto. Conversely, the μ_ϕ are exactly the Beltrami differentials orthogonal to all fake ones in the natural inner product $\int \mu_1\bar\mu_2\,\lambda\,\mathrm{d}^2z$, so each class contains exactly one of them (Hodge theory for $\bar\partial_{K^*}$), and the map is also one-to-one. These μ_ϕ are called **harmonic Beltrami differentials**. On the flat torus, $\phi=\mathrm{d}z^2$ gives the constant $\mu=1/\lambda$, in agreement with Example 5.87.

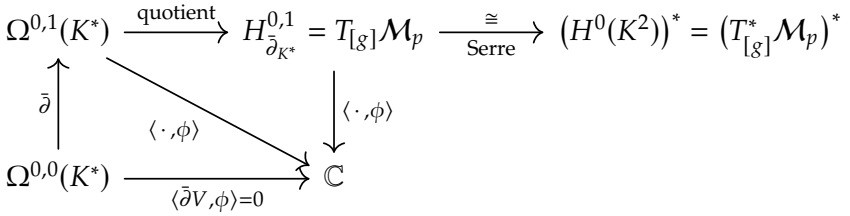


Figure 5.50: The logic of (5.73)–(5.78) as a diagram. Beltrami differentials $\Omega^{0,1}(K^*)$ contain the fake ones, $\bar\partial\Omega^{0,0}(K^*)$. Pairing with a holomorphic quadratic differential ϕ kills the fake ones (bottom row), so it descends to the quotient $H^{0,1}$, and Serre duality says that the resulting map to $H^0(K^2)^*$ is an isomorphism.

Automorphic forms and modular forms

Nash's footnote 6 [Nash p. 162] connects quadratic differentials to a classical subject. Suppose $\Sigma=\tilde\Sigma/\Gamma$ with Γ a discrete group of holomorphic automorphisms of $\tilde\Sigma$ (a Fuchsian group acting on H, as in Section 5.5.6). A holomorphic q-differential on Σ is the same thing as a holomorphic q-differential $\phi(z)\,\mathrm{d}z^q$ on $\tilde\Sigma$ that is invariant under Γ:

$$\phi(z)=\phi(\gamma(z))\,\gamma'(z)^q \qquad \text{for all } \gamma\in\Gamma.$$

Such a ϕ is called an **automorphic form of weight q** for Γ. When Γ is the modular group $\mathrm{SL}(2,\mathbb{Z})/\{\mp I\}$ it is called a **modular form** of weight q.

Worked example 5.90: A modular form as a quadratic differential on H

For $\gamma(\tau)=(a\tau+b)/(c\tau+d)$ with $ad-bc=1$ we found in Example 5.78 that $\gamma'(\tau)=(c\tau+d)^{-2}$. So the invariance condition for $q=2$ reads

$$\phi(\gamma\tau)=(c\tau+d)^4\,\phi(\tau).$$

The holomorphic Eisenstein series $G_4(\tau) = \sum_{(m,n)\neq(0,0)}(m+n\tau)^{-4}$ (the holomorphic cousin of Example 5.80; the sum converges absolutely) satisfies it:

$$G_4(\gamma\tau) \overset{(1)}{=} \sum_{(m,n)\neq 0} \frac{(c\tau+d)^4}{\big(m(c\tau+d)+n(a\tau+b)\big)^4}$$
$$\overset{(2)}{=} (c\tau+d)^4 \sum_{(m',n')\neq 0} \frac{1}{(m'+n'\tau)^4} = (c\tau+d)^4\, G_4(\tau).$$

(1) Put $m + n\gamma\tau$ over the common denominator $c\tau + d$.
(2) Relabel $(m', n') = (md + nb,\ mc + na)$, a bijection of $\mathbb{Z}^2 \setminus 0$ as in Example 5.80.

So $G_4(\tau)\,\mathrm{d}\tau^2$ is invariant under the modular group: it is a quadratic differential on the upper half-plane that descends to the (orbifold) quotient. It is a "modular form of weight 2" in Nash's convention.

Watch out: Two conventions for the weight

Nash's weight q counts powers of $\mathrm{d}z$. Number theorists count powers of $(c\tau + d)$, which is $2q$. So Nash's weight-2 form $G_4\,\mathrm{d}\tau^2$ is the number theorists' "weight 4 modular form G_4". Check which convention a source uses before comparing formulas.

At this point the tangent space problem has become a counting problem:

$$\dim_{\mathbb{C}} \mathcal{M}_p = \textit{number of linearly independent holomorphic quadratic differentials on } \Sigma.$$

The sphere has none and the torus has one (Example 5.89). For higher genus we need a machine, and the machine is the index theorem.

5.6.7 The cotangent space as the kernel of an elliptic operator

We have reduced the problem to counting holomorphic quadratic differentials. Nash now observes [Nash p. 163, (5.79)] that the space of holomorphic sections of K_Σ^2 is, by its definition, the kernel of an operator:

$$H^{0,0}_{\bar\partial_{K_\Sigma^2}}(\Sigma) = \operatorname{Ker}\bar\partial_{K_\Sigma^2}, \qquad \bar\partial_{K_\Sigma^2}:\ \Omega^{0,0}(K_\Sigma^2) \to \Omega^{0,1}(K_\Sigma^2), \quad \phi(z,\bar z)\,\mathrm{d}z^2 \ \mapsto\ \partial_{\bar z}\phi\ \mathrm{d}z^2\otimes\mathrm{d}\bar z. \tag{5.140}$$

Combined with (5.137) this gives Nash's (5.80):

$$\dim \mathcal{M}_p \ =\ \dim T^*_{[g]}\mathcal{M}_p \ =\ \dim \operatorname{Ker}\bar\partial_{K_\Sigma^2}. \tag{5.141}$$

This is the key observation. A question about the geometry of an infinite-dimensional quotient has become a question about the zero modes of a first-order elliptic operator on a compact surface. Zero modes of elliptic operators are what the index theorem of Chapter 4 is built to count.

Physical picture: What we are going to count, in string language

In the Faddeev–Popov treatment of the Polyakov integral (Chapter 9) the ghost c is a vector field, a section of K^*, and the antighost b is a quadratic differential, a section of K^2. Their action is $\int b_{zz}\,\partial_{\bar z}c^z\,\mathrm{d}^2z$, so the equations of motion are $\partial_{\bar z}c^z = 0$ and (after integrating by parts) $\partial_{\bar z}b_{zz} = 0$. Zero modes of c are conformal Killing vectors, zero modes of b are holomorphic quadratic differentials. The path integral over the ghosts vanishes unless we

insert one b for every b zero mode and one c for every c zero mode. Counting them is exactly the calculation below.

The adjoint operator, written out

The index of $\bar\partial_{K^2}$ is $\dim\operatorname{Ker}\bar\partial_{K^2} - \dim\operatorname{Ker}\bar\partial^*_{K^2}$. To see what the second kernel is, we need the adjoint $\bar\partial^*_{K^2}$. Nash never writes it down. Let us do so, because it makes the later Serre-duality step (5.87) completely concrete.

Choose a metric $ds^2 = \lambda\,dz\,d\bar z$. It induces a pointwise length on every tensor. Since $|dx|^2 = |dy|^2 = 1/\lambda$, we have $|dz|^2 = |d\bar z|^2 = 2/\lambda$, and the area element is $\lambda\,d^2z$. Dropping harmless constant factors, the L^2 inner products are

$$\begin{aligned}\langle\phi_1,\phi_2\rangle &= \int_\Sigma \phi_1\bar\phi_2\,\lambda^{-q}\,\lambda\,d^2z && \text{for sections } \phi\,dz^q \text{ of } K^q,\\ \langle\psi_1,\psi_2\rangle &= \int_\Sigma \psi_1\bar\psi_2\,\lambda^{-q-1}\,\lambda\,d^2z && \text{for } \psi\,dz^q\otimes d\bar z\in\Omega^{0,1}(K^q).\end{aligned}\tag{5.142}$$

The adjoint is defined by $\langle\bar\partial\phi,\psi\rangle = \langle\phi,\bar\partial^*\psi\rangle$. Compute the left side:

$$\begin{aligned}\langle\bar\partial\phi,\psi\rangle &\overset{(1)}{=} \int_\Sigma (\partial_{\bar z}\phi)\,\bar\psi\,\lambda^{-q}\,d^2z\\ &\overset{(2)}{=} -\int_\Sigma \phi\,\partial_{\bar z}\big(\lambda^{-q}\bar\psi\big)\,d^2z\\ &\overset{(3)}{=} \int_\Sigma \phi\,\overline{\Big(-\lambda^{q-1}\partial_z\big(\lambda^{-q}\psi\big)\Big)}\,\lambda^{1-q}\,d^2z.\end{aligned}$$

(1) Second line of (5.142); the powers of λ combine to λ^{-q}.

(2) Integrate by parts on the closed surface; the boundary term vanishes by Stokes, exactly as in Nash's Example on p. 162.

(3) Multiply and divide by λ^{1-q} and use $\overline{\partial_z f} = \partial_{\bar z}\bar f$ and λ real.

Comparing with the first line of (5.142):

$$\bar\partial^*_{K^q}\big(\psi\,dz^q\otimes d\bar z\big) = -\lambda^{q-1}\,\partial_z\big(\lambda^{-q}\psi\big)\,dz^q.\tag{5.143}$$

This is the "covariant ∂_z" of the metric. So $\bar\partial^*_{K^q}\psi = 0$ exactly when $\lambda^{-q}\psi$ is *anti*-holomorphic. Now look at the index structure of $\lambda^{-q}\psi$. The coefficient ψ belongs to $dz^q\otimes d\bar z$ and λ belongs to $dz\otimes d\bar z$, so $\lambda^{-q}\psi$ belongs to $d\bar z^{\,1-q}$. Its complex conjugate $v := \overline{\lambda^{-q}\psi}$ belongs to dz^{1-q} and is holomorphic. Hence

$$\operatorname{Ker}\bar\partial^*_{K^q} \xrightarrow{\ \cong\ } \overline{H^0\big(\Sigma,\mathcal{O}(K^{1-q})\big)},\qquad \psi\ \mapsto\ v = \overline{\lambda^{-q}\psi}.\tag{5.144}$$

For $q=2$: *the cokernel of $\bar\partial_{K^2}$ is (the complex conjugate of) the space of holomorphic vector fields*, the conformal Killing vectors of Section 5.4.5. This is Nash's (5.86)–(5.87), which he reaches through Hodge theory and Serre duality; here we reached it by one integration by parts.

Watch out: Holomorphic, anti-holomorphic, dual

The map (5.144) is *conjugate*-linear, because it involves complex conjugation. Serre duality gives a complex-linear isomorphism with the *dual* space instead. For counting dimensions the difference is irrelevant: V, $\bar V$ and V^* all have the same dimension.

5.6.8 The index: Riemann–Roch for quadratic differentials

Background: Chern classes and degrees on a surface

For a complex line bundle E over Σ the **first Chern class** $c_1(E)$ is represented by $\frac{i}{2\pi}F$, where F is the curvature 2-form of any connection on E (Chapter T2). Its integral is an integer, the **degree**:

$$\deg E \;:=\; \int_\Sigma c_1(E) \;=\; \#\{\text{zeros}\} - \#\{\text{poles}\} \text{ of any meromorphic section of } E, \tag{5.145}$$

counted with multiplicity (Chapter 5, $c_1([D]) = \deg D$). Three rules make degrees easy.

- $c_1(E \otimes F) = c_1(E) + c_1(F)$, so $\deg K^q = q \deg K$. (Transition functions multiply, their logarithms add.)
- $c_1(E^*) = -c_1(E)$, so $\deg K^* = -\deg K$.
- $K^* = T\Sigma$, and $c_1(T\Sigma)$ is the Euler class. Gauss–Bonnet (5.92) says $\int_\Sigma c_1(T\Sigma) = \frac{1}{2\pi}\int_\Sigma K\,\mathrm{d}A = 2 - 2p$.

The last rule is Nash's (5.83) [Nash p. 163]:

$$\int_\Sigma c_1(K_\Sigma^*) \;=\; 2-2p \;=\; -\int_\Sigma c_1(K_\Sigma), \qquad \text{i.e.} \quad \deg K_\Sigma = 2p-2. \tag{5.146}$$

Only the 2-form part of any characteristic class can be integrated over a surface, and products of two 2-forms vanish. So on Σ

$$\mathrm{ch}(E) = e^{c_1(E)} = 1 + c_1(E), \qquad \mathrm{td}(E) = \frac{c_1(E)}{1 - e^{-c_1(E)}} = 1 + \tfrac{1}{2}c_1(E).$$

Worked example 5.91: Three ways to see $\deg K = 2p-2$

Sphere, by a meromorphic section. The 1-form $\mathrm{d}z$ on $\mathbb{C}\cup\{\infty\}$ has no zeros or poles on $\mathbb{C}$. Near ∞, with $w = 1/z$,

$$\mathrm{d}z \overset{(1)}{=} -\frac{\mathrm{d}w}{w^2},$$

(1) $z = w^{-1}$, so $\mathrm{d}z = -w^{-2}\mathrm{d}w$.

a double pole. So $\deg K_{S^2} = 0 - 2 = -2 = 2\cdot 0 - 2$. ✓

Sphere, by Poincaré–Hopf. The vector field $\partial_z = -w^2\partial_w$ (Example 5.70) has a double zero at ∞ and no other zeros, so $\deg K^* = 2 = \chi(S^2)$. The degree of the tangent bundle is the sum of the indices of a vector field: that is the Poincaré–Hopf theorem.

Genus p, hyperelliptic. Take $y^2 = f(x)$ with f a polynomial of degree $2p+2$ with distinct roots $e_1, \dots, e_{2p+2}$ (Chapter 5). Consider $\omega = \mathrm{d}x/y$.

- At an ordinary point ($y \neq 0$, x finite), x is a good coordinate and ω is finite and non-zero.
- At a branch point $x = e_i$ the good coordinate is s with $x - e_i = s^2$, so $y = s\sqrt{f'(e_i) + O(s^2)}$ and

$$\omega = \frac{2s\,\mathrm{d}s}{s\sqrt{f'(e_i)}\,(1+O(s^2))} = \frac{2\,\mathrm{d}s}{\sqrt{f'(e_i)}}\,(1+O(s^2)),$$

finite and non-zero.

- Over $x = \infty$ there are two points ($y \approx \pm x^{p+1}$). With $t = 1/x$, $\mathrm{d}x = -t^{-2}\mathrm{d}t$ and $y \approx \pm t^{-p-1}$, so $\omega \approx \mp t^{p+1}t^{-2}\mathrm{d}t = \mp t^{p-1}\mathrm{d}t$: a zero of order $p-1$ at each point.

Total: $\deg K = 2(p-1) = 2p-2$. ✓ (For $p = 1$ this is the nowhere-vanishing $\mathrm{d}z$ of the torus, since $\mathrm{d}x/y = \mathrm{d}\wp/\wp' = \mathrm{d}z$.)

Nash's calculation

The index theorem for $\bar\partial_E$ on a curve is the Riemann–Roch theorem (Chapter 4, Nash's (4.118)). For $E = K_\Sigma^2$ it reads [Nash p. 163, (5.81)]

$$\operatorname{index}\bar\partial_{K_\Sigma^2} = \dim\operatorname{Ker}\bar\partial_{K_\Sigma^2} - \dim\operatorname{Ker}\bar\partial^*_{K_\Sigma^2} = \int_\Sigma \operatorname{ch}(K_\Sigma^2)\wedge\operatorname{td}(T\Sigma). \tag{5.147}$$

Evaluate the integrand [Nash p. 163, (5.82)]:

$$\begin{aligned}\operatorname{ch}(K_\Sigma^2)\wedge\operatorname{td}(T\Sigma) &\overset{(1)}{=} \big(1 + c_1(K_\Sigma^2)\big)\wedge\big(1 + \tfrac12 c_1(K_\Sigma^*)\big)\\ &\overset{(2)}{=} 1 + c_1(K_\Sigma^2) + \tfrac12 c_1(K_\Sigma^*) + \tfrac12 c_1(K_\Sigma^2)\wedge c_1(K_\Sigma^*)\\ &\overset{(3)}{=} 1 + 2c_1(K_\Sigma) - \tfrac12 c_1(K_\Sigma)\\ &= 1 + \tfrac32\, c_1(K_\Sigma).\end{aligned} \tag{5.148}$$

(1) $\operatorname{ch} = 1 + c_1$ and $\operatorname{td} = 1 + \frac12 c_1$ for line bundles on a surface; $T\Sigma = K_\Sigma^*$.
(2) Multiply out.
(3) The last term is a 4-form on a 2-manifold, so it vanishes; $c_1(K^2) = 2c_1(K)$ and $c_1(K^*) = -c_1(K)$.

Only the 2-form part contributes to the integral:

$$\operatorname{index}\bar\partial_{K_\Sigma^2} \overset{(1)}{=} \tfrac32\int_\Sigma c_1(K_\Sigma) \overset{(2)}{=} -\tfrac32\,(2-2p) = 3p-3. \tag{5.149}$$

(1) Integrate (5.148); the constant 1 is a 0-form and contributes nothing.
(2) Gauss–Bonnet in the form (5.146).

This is Nash's (5.84). In the language of Section 5.6.7,

$$\dim H^0(\Sigma, \mathcal{O}(K^2)) - \dim H^0(\Sigma, \mathcal{O}(K^*)) = 3p-3, \tag{5.150}$$

"b zero modes minus c zero modes". This holds for *every* genus. What we still need is the size of one of the two kernels.

Key idea: Riemann–Roch for every K^q

The same algebra with $E = K^q$ gives

$$\operatorname{index}\bar\partial_{K^q} \overset{(1)}{=} \int_\Sigma \big(1 + q\,c_1(K)\big)\big(1 - \tfrac12 c_1(K)\big) \overset{(2)}{=} \big(q - \tfrac12\big)(2p-2) = (2q-1)(p-1).$$

(1) $\operatorname{ch}(K^q) = 1 + qc_1(K)$, $\operatorname{td}(T\Sigma) = 1 - \frac12 c_1(K)$.
(2) Keep the 2-form part and use $\deg K = 2p-2$.

Equivalently $\operatorname{index}\bar\partial_E = \deg E + 1 - p$ for any line bundle E: this is the classical Riemann–Roch theorem of Chapter 4.

q	object	$\deg K^q$	index $\bar\partial_{K^q}$	meaning
-1	vector fields $V^z\partial_z$	$2-2p$	$3-3p$	CKVs minus moduli
0	functions	0	$1-p$	constants (1) minus $h^1(\mathcal{O})=p$
1	Abelian differentials $\omega(z)\mathrm{d}z$	$2p-2$	$p-1$	p differentials minus 1
2	quadratic differentials	$4p-4$	$3p-3$	moduli minus CKVs
q	q-differentials	$2q(p-1)$	$(2q-1)(p-1)$	

Note the symmetry index $\bar\partial_{K^q} = -\,\mathrm{index}\,\bar\partial_{K^{1-q}}$. It is Serre duality (5.135) again: H^0 and H^1 of K^q are H^1 and H^0 of K^{1-q} (up to duals), so the alternating sums change sign. The rows $q=-1$ and $q=2$ are the same statement, and so are $q=0$ and $q=1$.

Worked example 5.92: Checking Riemann–Roch on the sphere

On $\mathbb{CP}^1$ the bundle K^q is $\mathcal{O}(-2q)$, because $\deg K=-2$. A holomorphic section of $K^{-m}=(K^*)^m$ $(m\ge 0)$ is $V(z)\,\partial_z^{\otimes m}$ with V entire. Near ∞, $\partial_z=-w^2\partial_w$, so

$$V(z)\,\partial_z^{\otimes m} \overset{(1)}{=} (-1)^m\,w^{2m}\,V(1/w)\,\partial_w^{\otimes m}.$$

(1) Substitute $\partial_z=-w^2\partial_w$ in each of the m factors.

This is holomorphic at $w=0$ iff V is a polynomial of degree $\le 2m$. So $h^0(K^{-m})=2m+1$. On the other hand $h^1(K^{-m})=h^0(K^{1+m})=0$ by Example 5.89. Riemann–Roch predicts $(2q-1)(p-1)=(-2m-1)(-1)=2m+1$. ✓ For $m=1$ we recover the three conformal Killing vectors, and index $\bar\partial_{K^2}=0-3=-3$.

5.6.9 Vanishing theorems

Riemann–Roch gives a difference of two dimensions. To get $\dim\mathrm{Ker}\,\bar\partial_{K^2}$ by itself we must show that the other kernel vanishes. The tool is a vanishing theorem.

Working definition 5.49 (Positive and negative line bundles)

A holomorphic line bundle E on a compact complex manifold is **positive** if it carries a Hermitian metric whose curvature form $\frac{i}{2\pi}F$ is a positive $(1,1)$-form, and **negative** if E^* is positive. *Operationally, on a Riemann surface:* E is positive iff $\deg E>0$ and negative iff $\deg E<0$. (Given the degree, one can always choose the metric so that the curvature is a constant multiple of the area form, of the same sign as the degree.) Nash writes this as $c_1(E)>0$.

For the bundles we care about:

$$\deg K^2 = 4p-4, \qquad \deg K^* = 2-2p.$$

So for $p\ge 2$, K^2 is positive and K^* is negative; for $p=0$ it is the other way round; and for $p=1$ both are trivial, neither positive nor negative.

The Kodaira vanishing theorem

The result (Nash calls it the Kodaira vanishing theorem) is [Nash p. 164, (5.85)]:

Key idea: Kodaira vanishing

If E is a positive line bundle on a compact Kähler manifold M of complex dimension n, then

$$H^p\big(M, \Omega^q(E)\big) = 0 \qquad \text{for } p+q > n. \tag{5.151}$$

Here p is the cohomological degree and q the form degree ($\Omega^q(E)$ is the sheaf of holomorphic E-valued q-forms). By Serre duality this is equivalent to Nash's second form [Nash p. 165, (5.89)]: if E is *negative*,

$$H^i\big(M, \Omega^j(E)\big) = 0 \qquad \text{for } i+j < n. \tag{5.152}$$

Nash's (5.88) is the intermediate step. Apply Serre duality (5.134) to (5.85):

$$H^{n-p}\big(M, \Omega^{n-q}(E^*)\big) = 0 \qquad \text{for } p+q > n,$$

then rename $i = n-p$, $j = n-q$ (so $i+j < n$) and note that E^* is negative exactly when E is positive.

On a curve ($n = 1$) the two forms say:

- E positive $\Rightarrow H^1(\Sigma, \Omega^1(E)) = H^1(\Sigma, \mathcal{O}(K \otimes E)) = 0$ $\quad (p = q = 1)$;
- E negative $\Rightarrow H^0(\Sigma, \mathcal{O}(E)) = 0$ $\quad (i = j = 0)$.

Why the negative version is obvious on a curve

The second statement has a one-line proof on a Riemann surface, and it is the one we actually need. Suppose s is a non-zero holomorphic section of E. It has no poles, and its zeros are isolated, each with a positive multiplicity. Then by (5.145)

$$\deg E = \#\{\text{zeros of } s\} \;\geq 0.$$

So if $\deg E < 0$ there is no such s: $H^0(\Sigma, \mathcal{O}(E)) = 0$. Moreover, if $\deg E = 0$ then a non-zero section has no zeros at all, so it trivialises E, and $h^0(E) \leq 1$ with equality iff E is holomorphically trivial.

Why it is true in general: the Weitzenböck mechanism

Nash summarises the general proof in words [Nash p. 164]: write the cohomology as the kernel of a Laplacian; write the Laplacian as a positive "kinetic" term plus a curvature term; positivity of $c_1(E)$ makes the curvature term positive; so the Laplacian is strictly positive and has no kernel. The cleanest place to see every step is a physics problem we already know: a charged particle in a magnetic field.

Worked example 5.93: Landau levels are Kodaira vanishing

Put a uniform magnetic field B on the flat plane (or on a flat torus of area $\mathcal{A}$ with $B\mathcal{A} = 2\pi N$). Use the physicists' covariant derivative $D_a = \partial_a - iA_a$ with real A (in the mathematicians' convention $A_{\text{math}} = -iA$) and symmetric gauge $A = \frac{B}{2}(-y\,\mathrm{d}x + x\,\mathrm{d}y)$, so $F_{xy} = \partial_x A_y - \partial_y A_x = B$. The line bundle has $c_1 = \frac{1}{2\pi} B\,\mathrm{d}x \wedge \mathrm{d}y$ and degree $N = B\mathcal{A}/2\pi$.
Step 1: the operators. Set $D_z = \frac{1}{2}(D_x - iD_y)$ and $D_{\bar z} = \frac{1}{2}(D_x + iD_y)$. Since $A_x + iA_y = \frac{B}{2}(-y + ix) = \frac{iB}{2} z$,

$$D_{\bar z} \overset{(1)}{=} \partial_{\bar z} - \tfrac{i}{2}(A_x + iA_y) \overset{(2)}{=} \partial_{\bar z} + \tfrac{B}{4} z, \qquad D_z = \partial_z - \tfrac{B}{4}\bar z.$$

(1) Definition of $D_{\bar z}$ with $\partial_{\bar z} = \frac{1}{2}(\partial_x + i\partial_y)$.
(2) $-\frac{i}{2}\cdot\frac{iB}{2}z = \frac{B}{4}z$. The formula for D_z is the same computation with $A_x - iA_y = -\frac{iB}{2}\bar z$.

$D_{\bar z}$ is the operator $\bar\partial_E$ of this line bundle.
Step 2: the commutator is the curvature.

$$[D_z, D_{\bar z}] \overset{(1)}{=} [\partial_z, \tfrac{B}{4}z] + [-\tfrac{B}{4}\bar z, \partial_{\bar z}] \overset{(2)}{=} \tfrac{B}{4} + \tfrac{B}{4} = \tfrac{B}{2}.$$

(1) $[\partial_z, \partial_{\bar z}] = 0$ and $[z, \bar z] = 0$.
(2) $[\partial_z, z] = 1$ and $[\bar z, \partial_{\bar z}] = -1$ as operators.

Step 3: adjoints. On $L^2(\mathrm{d}^2 z)$, integration by parts gives $\partial_z^\dagger = -\partial_{\bar z}$, and multiplication by $\bar z$ has adjoint multiplication by z. Hence $D_{\bar z}^\dagger = -D_z$.
Step 4: the Weitzenböck identity. The two Laplacians $\Delta_0 = \bar\partial_E^\dagger\bar\partial_E$ (on sections) and $\Delta_1 = \bar\partial_E\bar\partial_E^\dagger$ (on $(0,1)$-forms, which on the flat plane we identify with sections using the flat frame $\mathrm{d}\bar z$) satisfy

$$\Delta_1 - \Delta_0 \overset{(1)}{=} -D_{\bar z}D_z + D_zD_{\bar z} \overset{(2)}{=} [D_z, D_{\bar z}] \overset{(3)}{=} \tfrac{B}{2}. \tag{5.153}$$

(1) Step 3: $\Delta_0 = -D_zD_{\bar z}$ and $\Delta_1 = -D_{\bar z}D_z$.
(2) Definition of the commutator.
(3) Step 2.

Step 5: vanishing. Both Δ_0 and Δ_1 are non-negative (each is $T^\dagger T$ or $TT^\dagger$). If $B > 0$ then $\Delta_1 = \Delta_0 + \frac{B}{2} \geq \frac{B}{2} > 0$: Δ_1 has no kernel, so $\operatorname{Ker}\bar\partial_E^\dagger = 0$, i.e. $H^1(\mathcal{O}(E)) = 0$. If $B < 0$ then $\Delta_0 = \Delta_1 + \frac{|B|}{2} > 0$: no holomorphic sections, $H^0(\mathcal{O}(E)) = 0$. The sign of the magnetic field decides which kernel is killed.
Step 6: the spectrum. For $B > 0$ put $a = \sqrt{2/B}\,D_{\bar z}$, so $a^\dagger = -\sqrt{2/B}\,D_z$ and $[a, a^\dagger] = -\frac{2}{B}[D_{\bar z}, D_z] = 1$. Then $\Delta_0 = \frac{B}{2}a^\dagger a$ and $\Delta_1 = \frac{B}{2}(a^\dagger a + 1)$. This is a harmonic oscillator: the spectrum of Δ_0 is $\frac{B}{2}\{0, 1, 2, \dots\}$ and that of Δ_1 is $\frac{B}{2}\{1, 2, 3, \dots\}$ (Figure 5.51). The zero modes of $D_{\bar z}$ are $f(z)\,e^{-B|z|^2/4}$ with f holomorphic, the lowest Landau level. On the torus there are exactly N of them (theta functions), and Riemann–Roch with $h^1 = 0$ gives index $= \deg E + 1 - p = N$. ✓

On a curved surface the same computation produces an extra term: the $(0,1)$-forms themselves are sections of $E \otimes \bar K$, and the frame $\mathrm{d}\bar z$ is no longer flat. The curvature that appears in $\Delta_1 - \Delta_0$ is then that of $E \otimes K^{-1}$, not of E. This is why the clean statement is about $\Omega^1(E) = \mathcal{O}(K \otimes E)$, as in (5.151): for $K \otimes E$ the offending K^{-1} cancels.

Watch out: Positive does not mean $H^1(\mathcal{O}(E)) = 0$

It is tempting to read Kodaira as "$H^1(\Sigma, \mathcal{O}(E)) = 0$ for every positive E". That is false on a curve of genus $p \geq 2$. Take $E = [P]$, the degree-one bundle of a point. Serre duality gives $h^1(\mathcal{O}([P])) = h^0(\mathcal{O}(K - P))$, the number of holomorphic differentials vanishing at P, which is $p - 1 \neq 0$ (at most one linear condition on p differentials, and it is a genuine condition because not all differentials vanish at P; see Section 5.7). What is true is $H^1(\mathcal{O}(E)) = 0$ whenever $\deg E > 2p - 2$, because then $K \otimes E^*$ is negative.

5.6.10 Putting it together: $\dim_{\mathbb{C}} \mathcal{M}_p = 3p - 3$ for $p \geq 2$

Now we follow Nash's chain [Nash pp. 164–165, (5.86)–(5.92)].

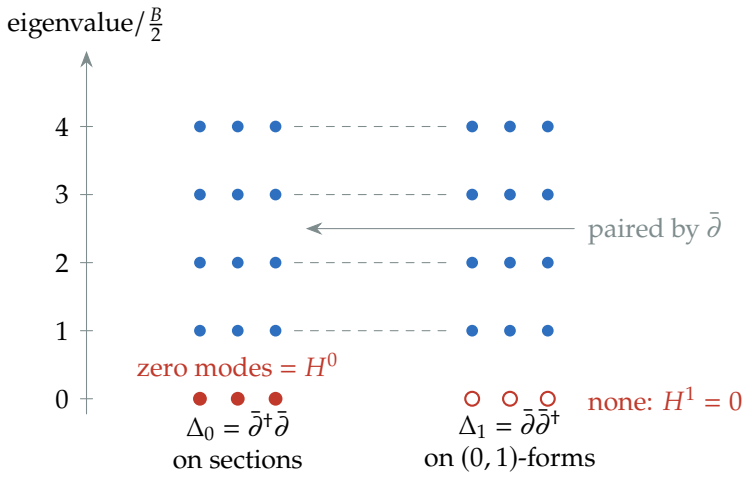


Figure 5.51: The Weitzenböck mechanism for a positive line bundle (Example 5.93, $B > 0$). Every non-zero level of Δ_0 is matched by a level of Δ_1 with the same eigenvalue ($\bar\partial$ maps one eigenspace isomorphically onto the other). The identity $\Delta_1 = \Delta_0 + \frac{B}{2}$ pushes the whole spectrum of Δ_1 up by one unit, so only Δ_0 has zero modes: holomorphic sections exist and H^1 vanishes. For $B < 0$ the picture is reflected.

Step 1: identify the cokernel as a cohomology group. By Hodge theory (Chapter 4, Chapter 3) each Dolbeault class has exactly one representative annihilated by $\bar\partial^*$, so

$$\operatorname{Ker}\bar\partial^*_{K_\Sigma^2} \;\cong\; H^{0,1}_{\bar\partial_{K_\Sigma^2}}(\Sigma) \;\cong\; H^1\big(\Sigma, \mathcal{O}(K_\Sigma^2)\big). \tag{5.154}$$

Step 2: Serre duality.

$$\operatorname{Ker}\bar\partial^*_{K_\Sigma^2} \overset{(1)}{=} H^1\big(\Sigma,\mathcal{O}(K_\Sigma^2)\big) \overset{(2)}{\cong} \big\{H^0\big(\Sigma,\mathcal{O}(K_\Sigma^{-2}\otimes K_\Sigma)\big)\big\}^* \overset{(3)}{=} \big\{H^0\big(\Sigma,\mathcal{O}(K_\Sigma^*)\big)\big\}^* \overset{(4)}{=} \big\{\operatorname{Ker}\bar\partial_{K_\Sigma^*}\big\}^*. \tag{5.155}$$

(1) (5.154).
(2) Serre duality (5.135) with $E = K^2$, $E^* = K^{-2}$.
(3) $K^{-2}\otimes K = K^{-1} = K^*$.
(4) Holomorphic sections are the kernel of $\bar\partial$.

This is the same conclusion as our explicit (5.144). Nash's comment: Serre duality has replaced the *positive* bundle K^2 (for which Kodaira's theorem, in the form Nash uses, gives nothing about $H^1(\mathcal{O}(K^2))$) by the *negative* bundle K^*.

Step 3: vanishing. Apply (5.152) with $E = K^*$, $i = j = 0$, $n = 1$ [Nash p. 165, (5.90)–(5.92)]:

$$H^0\big(\Sigma,\mathcal{O}(K_\Sigma^*)\big) = 0 \quad \text{provided } c_1(K_\Sigma^*) < 0, \qquad c_1(K_\Sigma^*) = 2-2p, \tag{5.156}$$

and $2-2p < 0$ exactly when $p \geq 2$. In words: *a surface of genus $p \geq 2$ has no non-zero holomorphic vector fields*, because such a field would have $2-2p < 0$ zeros. So $\operatorname{Ker}\bar\partial^*_{K^2} = 0$ for $p \geq 2$.

Step 4: conclude.

$$\dim_{\mathbb{C}}\mathcal{M}_p \overset{(1)}{=} \dim\operatorname{Ker}\bar\partial_{K_\Sigma^2} \overset{(2)}{=} \operatorname{index}\bar\partial_{K_\Sigma^2} + \dim\operatorname{Ker}\bar\partial^*_{K_\Sigma^2} \overset{(3)}{=} (3p-3) + 0. \tag{5.157}$$

(1) (5.141).
(2) Definition of the index.
(3) (5.149) and Step 3.

Key idea: The dimension of moduli space (Nash's (5.93))

$$\dim_{\mathbb{C}} \mathcal{M}_p = 3p - 3 \qquad (p \geq 2).$$

Mechanism: the tangent space is $H^{0,1}(K^*)$; its dual is the space of holomorphic quadratic differentials; Riemann–Roch says $h^0(K^2) - h^0(K^*) = 3p - 3$; and $h^0(K^*) = 0$ because a vector field on a surface of negative Euler characteristic must have a negative number of zeros.

Watch out: Nash's "$= \emptyset$"

Nash writes $\operatorname{Ker} \bar{\partial}^*_{K^2_\Sigma} = \emptyset$ and $\operatorname{Ker} \bar{\partial}_{K^2_\Sigma} = \emptyset$. A kernel is a vector space and always contains 0; he means the zero space $\{0\}$, of dimension 0.

A second route, directly through Kodaira

There is a shortcut Nash does not mention. Write $\mathcal{O}(K^2) = \mathcal{O}(K \otimes K) = \Omega^1(K)$. Then

$$H^1\big(\Sigma, \mathcal{O}(K^2)\big) = H^1\big(\Sigma, \Omega^1(K)\big),$$

which has $p + q = 1 + 1 = 2 > 1 = n$. For $p \geq 2$ the bundle K has degree $2p - 2 > 0$, so it is positive, and (5.151) kills this group directly. The two routes agree, as they must: they are Serre duals of each other.

Worked example 5.94: Genus two, by hand

Let us exhibit the $3 \cdot 2 - 3 = 3$ quadratic differentials on a genus-2 surface $y^2 = f(x)$, $\deg f = 6$. The holomorphic differentials are $\omega_0 = \mathrm{d}x/y$ and $\omega_1 = x\,\mathrm{d}x/y$ (Chapter 5; by Example 5.91 ω_0 has simple zeros at the two points over $x = \infty$, and x has simple poles there, so ω_1 is holomorphic too). Products of two holomorphic differentials are holomorphic quadratic differentials:

$$\omega_0^2 = \frac{\mathrm{d}x^2}{y^2}, \qquad \omega_0\omega_1 = \frac{x\,\mathrm{d}x^2}{y^2}, \qquad \omega_1^2 = \frac{x^2\,\mathrm{d}x^2}{y^2}.$$

They are linearly independent (as $1, x, x^2$ are), so $h^0(K^2) \geq 3$; with Riemann–Roch and $h^0(K^*) = 0$, exactly 3. Is there any other candidate, like $x^3\mathrm{d}x^2/y^2$? At a point over ∞, ω_0^2 has a double zero and x^3 a triple pole, so $x^3\omega_0^2$ has a simple pole: not holomorphic. ✓ The count also matches the hyperelliptic parameter count: 6 branch points on $\mathbb{CP}^1$ minus 3 Möbius parameters leaves 3.

5.6.11 Genus zero and genus one

For $p < 2$ the Kodaira argument for K^* fails, and Nash treats the two cases separately [Nash pp. 165–166, (5.94)–(5.99)].

Genus zero

The index is [Nash p. 165, (5.94)]

$$\operatorname{index} \bar{\partial}_{K^2_\Sigma} \overset{(1)}{=} 3 \cdot 0 - 3 = -3, \tag{5.158}$$

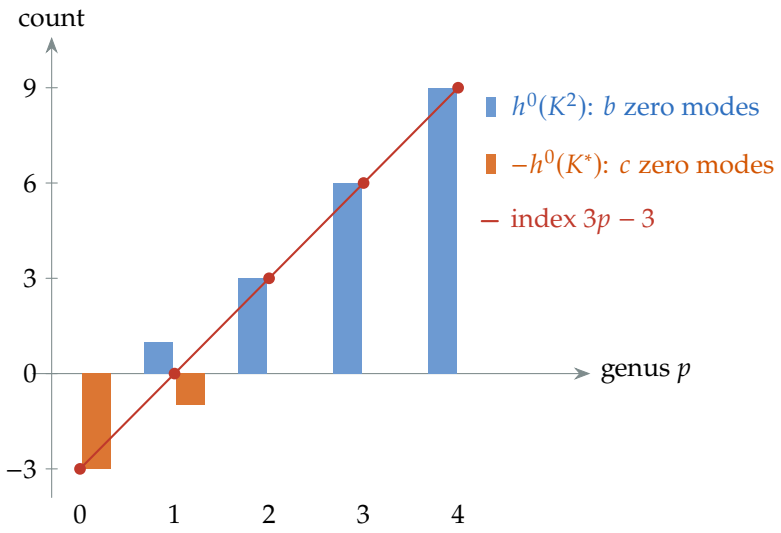


Figure 5.52: Moduli and conformal Killing vectors by genus. Blue bars (up): holomorphic quadratic differentials, $\dim \mathcal{M}_p$. Orange bars (down): holomorphic vector fields, the leftover symmetries. The red line is their difference, the index $3p-3$, which Riemann–Roch computes for every genus. Only for $p \geq 2$ is the orange bar absent, so that the index alone gives the dimension; for $p = 0$ and $p = 1$ we need extra information.

(1) (5.149) with $p = 0$.

so $\dim \operatorname{Ker} \bar\partial^*_{K^2} \geq 3 > 0$. This time we apply the vanishing theorem to the other kernel [Nash (5.95)–(5.96)]:

$$\operatorname{Ker} \bar\partial_{K^2_\Sigma} = H^0\big(\Sigma, \mathcal{O}(K^2_\Sigma)\big) = 0,$$

because $\deg K^2 = 4p - 4 = -4 < 0$: for genus zero, K^2 is a *negative* line bundle. (Example 5.89 proved the same thing with Liouville's theorem.) Hence [Nash (5.97)]

$$\dim_{\mathbb{C}} \mathcal{M}_0 = 0, \tag{5.159}$$

in agreement with Section 5.5: there is only one Riemann sphere. The index is carried entirely by the other kernel: $\dim \operatorname{Ker} \bar\partial^*_{K^2} = 3 = h^0(K^*)$. These are the three conformal Killing vectors (Nash's footnote 7).

Worked example 5.95: The conformal Killing vectors of the sphere form $\mathfrak{sl}(2,\mathbb{C})$

Nash's footnote 7 says the three complex parameters are given by the action of the Möbius group $\mathrm{SL}(2,\mathbb{C})/\{\mp I\}$. Let us exhibit the Lie algebra. Take the basis

$$\ell_{-1} = -\partial_z, \qquad \ell_0 = -z\,\partial_z, \qquad \ell_1 = -z^2\,\partial_z,$$

that is, $\ell_n = -z^{n+1}\partial_z$. Commute two of them, acting on a test function f:

$$\begin{aligned}
[\ell_m, \ell_n] f &\overset{(1)}{=} z^{m+1}\partial_z\big(z^{n+1}f'\big) - z^{n+1}\partial_z\big(z^{m+1}f'\big) \\
&\overset{(2)}{=} (n+1)z^{m+n+1}f' - (m+1)z^{m+n+1}f' \\
&\overset{(3)}{=} (m-n)\,\ell_{m+n} f.
\end{aligned}$$

(1) Definition, with the two minus signs cancelling in each product.
(2) Product rule; the terms with f'' cancel.
(3) $(n-m)z^{m+n+1}f' = (m-n)\cdot(-z^{m+n+1}\partial_z f)$.

So $[\ell_0, \ell_{\pm1}] = \mp\ell_{\pm1}$ and $[\ell_1, \ell_{-1}] = 2\ell_0$: the algebra $\mathfrak{sl}(2,\mathbb{C})$. Compare with the 2×2 matrices

$$\ell_{-1} \leftrightarrow \begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix}, \quad \ell_0 \leftrightarrow \begin{pmatrix} -\frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}, \quad \ell_1 \leftrightarrow \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix},$$

which generate the Möbius maps $z \mapsto (az+b)/(cz+d)$: for instance $\exp\left(t\left(\begin{smallmatrix} 0 & -1 \\ 0 & 0 \end{smallmatrix}\right)\right) = \left(\begin{smallmatrix} 1 & -t \\ 0 & 1 \end{smallmatrix}\right)$ is $z \mapsto z - t$, the flow of $-\partial_z$. In general the matrix $\left(\begin{smallmatrix} \alpha & \beta \\ \gamma & -\alpha \end{smallmatrix}\right)$ generates the vector field $(\beta + 2\alpha z - \gamma z^2)\partial_z$ (differentiate $((1+t\alpha)z + t\beta)/(t\gamma z + 1 - t\alpha)$ at $t = 0$). The matrix commutators reproduce the relations up to an overall sign, e.g. $[M_1, M_{-1}] = -2M_0$: the map from matrices to vector fields of a *left* action is an anti-homomorphism, a standard sign that one can remove by using $-M^T$ instead. These ℓ_n, $n \in \{-1, 0, 1\}$, are the globally defined members of the Witt algebra $\ell_n = -z^{n+1}\partial_z$, $n \in \mathbb{Z}$, whose central extension is the Virasoro algebra of Chapter 11.

Genus one

For the torus [Nash p. 166, (5.98)]

$$\operatorname{index} \bar\partial_{K_\Sigma^2} = 0 \quad \Longrightarrow \quad \dim \operatorname{Ker} \bar\partial^*_{K_\Sigma^2} = \dim \operatorname{Ker} \bar\partial_{K_\Sigma^2}. \tag{5.160}$$

Both K^2 and K^* are trivial ($\mathrm{d}z$ is a global nowhere-zero section of K), so neither vanishing theorem applies. But we already know the answer from Section 5.5: $\mathcal{T}_1 = \mathbb{C}^+$ is one-dimensional, so $\dim \operatorname{Ker} \bar\partial_{K^2} = 1$ (the quadratic differential $\mathrm{d}z^2$).

Nash then turns the logic around. The index theorem now tells us something new: $\dim \operatorname{Ker} \bar\partial^*_{K^2} = 1$, and that kernel is dual to the space of conformal Killing vectors. His argument, spelled out:

(a) A conformal Killing vector V is one whose flow changes the metric only by a Weyl factor, so it produces $\delta\mu = 0$.

(b) By (5.130), $\delta\mu = \partial_{\bar z} V^z$ locally. So $\partial_{\bar z} V^z = 0$: $V^z\partial_z$ is a holomorphic section of $T\Sigma = K^*$.

(c) These form $H^0(\Sigma, \mathcal{O}(K^*))$, which is dual to $\operatorname{Ker} \bar\partial^*_{K^2}$ by (5.155).

(d) Hence $\dim\{\text{conformal Killing vectors}\} = 1$ for genus one.

This agrees with Example 5.71: the only conformal Killing vectors of a flat torus are the translations $c\,\partial_z$.

Physical picture: What the torus numbers mean for a one-loop amplitude

The one modulus is τ, integrated over the fundamental domain F (Section 5.5.5). The one conformal Killing vector is the translation of the torus. In a one-loop amplitude with n vertex operators it is used to fix the position of one of them; if there are no vertex operators (the vacuum amplitude) nothing can be fixed, and we must divide instead by the volume of the translation group, which is the area of the torus, $\int \mathrm{d}^2 z = \tau_2$. That factor $1/\tau_2$, combined with powers of τ_2 from the determinants, assembles the modular-invariant measure $\mathrm{d}^2\tau/\tau_2^2$ of (5.112).

Summary

Nash's final table [Nash p. 166, (5.99)], with the kernels made explicit:

$$\dim_{\mathbb{C}} \mathcal{M}_p = \begin{cases} 0 & p = 0 \ \ (\mathcal{M}_0 \text{ is a point}), \\ 1 & p = 1, \\ 3p - 3 & p \geq 2. \end{cases} \tag{5.161}$$

p	$h^0(K^2)$ (moduli)	$h^0(K^*)$ (CKVs)	index	which vanishing is used
0	0	3	-3	K^2 negative $\Rightarrow h^0(K^2) = 0$
1	1	1	0	none; use $\mathcal{T}_1 = \mathbb{C}^+$
≥ 2	$3p - 3$	0	$3p - 3$	K^* negative $\Rightarrow h^0(K^*) = 0$

A compact way to remember all three rows: $\dim_{\mathbb{C}} \mathcal{M}_p - \dim_{\mathbb{C}}\{\text{CKVs}\} = 3p - 3$, and at most one of the two terms is non-zero except at $p = 1$.

5.6.12 Beyond Nash: three more ways to see $3p - 3$

The number $3p - 3$ is so important that it is worth seeing it appear from completely different directions. Each of the following is a standard physicist's count.

Punctures: $\dim \mathcal{M}_{p,n} = 3p - 3 + n$

String amplitudes with n external states are computed on surfaces with n marked points (the positions of the vertex operators). The relevant space is $\mathcal{M}_{p,n}$, genus-p Riemann surfaces with n distinct labelled points, up to biholomorphisms carrying marked points to marked points. Each marked point can move over the surface, adding one complex parameter, but for $p = 0$ and $p = 1$ the conformal Killing vectors can be used to fix some of them. The index theorem handles all cases at once. Let $D = P_1 + \cdots + P_n$. The deformations are now counted by quadratic differentials allowed to have simple poles at the P_i, i.e. holomorphic sections of $K^2 \otimes [D]$; the infinitesimal symmetries are vector fields vanishing at the P_i, sections of $K^* \otimes [-D]$. Riemann–Roch:

$$\begin{aligned} h^0(K^2 \otimes [D]) - h^0(K^* \otimes [-D]) &\overset{(1)}{=} \deg(K^2 \otimes [D]) + 1 - p \\ &\overset{(2)}{=} (4p - 4 + n) + 1 - p = 3p - 3 + n. \end{aligned}$$

(1) Riemann–Roch, together with Serre duality $h^1(K^2 \otimes [D]) = h^0(K^{-1} \otimes [-D])$.
(2) $\deg K^2 = 4p - 4$, $\deg[D] = n$.

The vector fields term is zero as soon as $\deg(K^* \otimes [-D]) = 2 - 2p - n < 0$, i.e. $2p - 2 + n > 0$, the condition for a **stable** marked surface. Then $\dim_{\mathbb{C}} \mathcal{M}_{p,n} = 3p - 3 + n$.

Worked example 5.96: Four points on the sphere and the Virasoro–Shapiro amplitude

$\mathcal{M}_{0,4}$ has dimension $0 - 3 + 4 = 1$. Explicitly: a Möbius map can send any three distinct points to $0, 1, \infty$, and it is then unique, so a configuration (z_1, z_2, z_3, z_4) is determined by the image of z_4, the **cross-ratio**

$$\lambda = \frac{(z_4 - z_1)(z_2 - z_3)}{(z_4 - z_3)(z_2 - z_1)},$$

which is invariant under all Möbius maps and sends $(z_1, z_2, z_3) \mapsto (0, 1, \infty)$ when evaluated

at z_4. (Check: $z_4 = z_1$ gives 0, $z_4 = z_2$ gives 1, $z_4 = z_3$ gives ∞.) This is why the tree-level closed-string four-point amplitude (Virasoro–Shapiro) is a single integral over the complex plane,

$$A_4 \propto \int_{\mathbb{C}} \mathrm{d}^2\lambda \; |\lambda|^{\alpha' k_1\cdot k_4}\, |1-\lambda|^{\alpha' k_2\cdot k_4}$$

(for tachyons of momenta k_i, in Polchinski's conventions), with the other three vertex operators fixed at $z_1 = 0$, $z_2 = 1$, $z_3 = \infty$ by the three conformal Killing vectors. Moduli (1) plus conformal Killing vectors (3) equals the number of vertex positions (4).

Pants decompositions

Cut a surface of genus $p \geq 2$ along disjoint simple closed curves until every piece is a **pair of pants** (a sphere with three holes). The Euler characteristic of a pair of pants is -1, and cutting along circles does not change χ (a circle has $\chi = 0$), so

$$\#\{\text{pants}\} = -\chi(\Sigma) = 2p - 2.$$

Each pair of pants has 3 boundary circles and each cut circle is shared by two pants, so

$$\#\{\text{cut curves}\} = \tfrac{3}{2}(2p-2) = 3p - 3.$$

A hyperbolic pair of pants with geodesic boundary is determined up to isometry by its three boundary lengths. When we glue, each cut curve carries two real parameters: its length $\ell_i > 0$ and a **twist** angle θ_i saying how far one side is rotated before gluing. These **Fenchel–Nielsen coordinates** give

$$\dim_{\mathbb{R}} \mathcal{T}_p = 2(3p-3) = 6p - 6,$$

and in fact $\mathcal{T}_p \cong (\mathbb{R}_{>0} \times \mathbb{R})^{3p-3}$, which makes the contractibility quoted in Section 5.5 visible. Twisting by a full 2π gives the same surface with the curves relabelled: that is precisely a Dehn twist (Section 5.5.7), an element of the mapping class group.

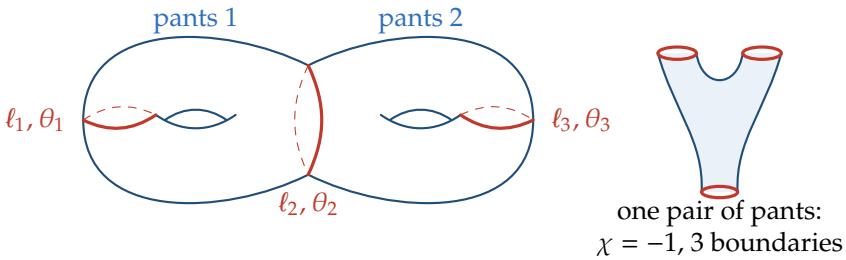


Figure 5.53: A pants decomposition of a genus-2 surface. Cutting along the $3 = 3p - 3$ red curves leaves $2 = 2p - 2$ pairs of pants (right). Each cut curve carries a length ℓ_i and a twist θ_i: $6 = 6p - 6$ real parameters, the Fenchel–Nielsen coordinates on $\mathcal{T}_2$. Twisting θ_i by 2π is a Dehn twist.

Riemann's own count, by branched coverings

Riemann obtained $3p - 3$ in 1857 by counting meromorphic functions. A non-constant meromorphic function f of degree n (it takes each value n times) presents Σ as an n-sheeted cover of $\mathbb{CP}^1$, branched at the points where two sheets meet. For generic f each branch point is simple.

Step 1: the number of branch points (Riemann–Hurwitz). Triangulate $\mathbb{CP}^1$ with the b branch values among the vertices, and lift. Every cell has n preimages, except that each simple branch value has only $n-1$ preimages. So

$$\chi(\Sigma) \overset{(1)}{=} n\,\chi(\mathbb{CP}^1) - b \quad\Longrightarrow\quad 2-2p \overset{(2)}{=} 2n-b \quad\Longrightarrow\quad b = 2n+2p-2. \tag{5.162}$$

(1) Count vertices, edges and faces upstairs: n times the count downstairs, minus one vertex for each simple branch value.
(2) $\chi(\mathbb{CP}^1) = 2$ and $\chi(\Sigma) = 2-2p$.

Step 2: the pairs (Σ, f). By Riemann's existence theorem a branched cover is determined, up to finitely many choices of how the sheets are connected, by its branch values. So the pairs (Σ, f) form a space of complex dimension $b = 2n+2p-2$.

Step 3: functions on a fixed surface. On a fixed Σ, a degree-n function is fixed by its polar divisor D (n points: n parameters) and then by an element of $\mathcal{L}(D)$, the space of functions with poles bounded by D. For $n > 2p-2$ Riemann–Roch gives $\dim \mathcal{L}(D) = n+1-p$ (this is exactly the calculation of Section 5.7 below). So the functions of degree n on a fixed surface form a family of dimension $n + (n+1-p) = 2n+1-p$.

Step 4: subtract.

$$\dim_{\mathbb{C}} \mathcal{M}_p \overset{(1)}{=} b - (2n+1-p) \overset{(2)}{=} (2n+2p-2) - (2n+1-p) = 3p-3.$$

(1) Pairs (Σ, f) fibre over $\mathcal{M}_p$ with fibres the functions on a fixed Σ.
(2) (5.162) and Step 3.

The answer does not depend on the auxiliary degree n, which is a strong consistency check. For the hyperelliptic surfaces of Example 5.94 ($n=2$) the count gives $b-3 = 2p+2-3 = 2p-1$, the dimension of the hyperelliptic locus; for $p=2$ this equals $3p-3=3$, so every genus-2 surface is hyperelliptic, while for $p \geq 3$ ($2p-1 < 3p-3$) most are not. (For $n=2$ the dimension of the functions is not $2n+1-p$, because $n \leq 2p-2$: this is one of the special cases Riemann–Roch warns about, and the reason we subtracted only the 3 Möbius parameters.)

Reading Nash (pp. 157–166, (5.59)–(5.99))

Nash computes $\dim \mathcal{M}_p$ in five moves: (i) write the metric as $\lambda|\mathrm{d}z + \mu\,\mathrm{d}\bar{z}|^2$ (5.59), derive the transformation laws (5.60) and identify $\delta\mu$ as a Beltrami differential (5.61)–(5.65); (ii) remove diffeomorphisms using the Lie derivative (5.66)–(5.73), getting $T\mathcal{M}_p = H^{0,1}_{\bar\partial_{K^*}}$ (5.74)–(5.76); (iii) Serre duality turns this into holomorphic quadratic differentials (5.77)–(5.78), with a worked pairing example on p. 162; (iv) the index of $\bar\partial_{K^2}$ is $3p-3$ by Riemann–Roch and Gauss–Bonnet (5.79)–(5.84); (v) Kodaira vanishing, after Serre duality, kills the cokernel for $p \geq 2$ (5.85)–(5.93), and $p = 0, 1$ are handled separately (5.94)–(5.99). We added: the explicit (λ, μ) decomposition checks, the Cartesian meaning of $\delta\mu$, the torus in Fourier modes, the explicit adjoint (5.143) (which exhibits (5.87) without Serre duality), the Landau-level proof of the vanishing mechanism, the correction that Kodaira on curves concerns $K \otimes E$, the $\mathfrak{sl}(2,\mathbb{C})$ of footnote 7, the genus-2 quadratic differentials, and three independent counts of $3p-3$. Note Nash's factor 2 in (5.72) and his "$= \emptyset$" for the zero space.

Recap

- A tangent vector to $\mathcal{M}_p$ is a Beltrami differential $\delta\mu\,\partial_z \otimes \mathrm{d}\bar{z}$ modulo $\partial_{\bar z} V^z$: $T\mathcal{M}_p = H^{0,1}(K^*)$. On the torus it is the average of $\delta\mu$, i.e. $\delta\tau$.

- Serre duality: $T^*\mathcal{M}_p = H^0(K^2)$, holomorphic quadratic differentials $\phi(z)\mathrm{d}z^2$, paired with $\delta\mu$ by $\int \mu\phi\, \mathrm{d}^2z$.
- Riemann–Roch: $h^0(K^2) - h^0(K^*) = \int \mathrm{ch}(K^2)\,\mathrm{td}(T\Sigma) = 3p-3$ for every p ("b zero modes minus c zero modes").
- Vanishing: a negative line bundle has no holomorphic sections (degree count, or the Weitzenböck/Landau-level argument). For $p \geq 2$, K^* is negative, so $\dim_{\mathbb{C}} \mathcal{M}_p = 3p-3$.
- $p = 0$: 0 moduli, 3 CKVs ($\mathfrak{sl}(2,\mathbb{C})$). $p = 1$: 1 modulus, 1 CKV.
- The same $3p-3$ (or $3p-3+n$ with punctures) comes from pants decompositions and from Riemann's branched-cover count.

5.7 Weierstrass gaps and Weierstrass points

Up to now this part has asked a question about the *collection* of all Riemann surfaces of genus p: how many shapes are there? The answer was a space $\mathcal{M}_p$ of complex dimension $3p-3$. Nash's last section asks a question about *one* surface at a time:

> *Which meromorphic functions can live on a given compact Riemann surface, and in particular, which kinds of pole can a function have if it is allowed only one pole?*

On the complex plane the question is empty: $1/z^n$ has a single pole of order n, for every n. On a compact surface of genus $p \geq 1$ the global topology forbids some of these. A function on a torus, for example, cannot have just one simple pole. The forbidden orders are called **Weierstrass gaps**. There are always exactly p of them, and at most points of the surface they are the first p integers $1, 2, \ldots, p$. The finitely many special points where the pattern is different are the **Weierstrass points**.

Why should a physicist care?

- *Ward identities are pole-counting problems.* In a conformal field theory on a worldsheet Σ, a correlation function $\langle J(z)\,\mathcal{O}_1(z_1)\cdots\mathcal{O}_n(z_n)\rangle$ is, as a function of z, a meromorphic object on Σ with poles only at the insertion points. Whether it is determined by its singular parts, and which singular parts are allowed, is exactly the question of this section. We will see that for a current (a 1-form) the only restriction is the residue theorem, which is charge conservation, and that for the stress tensor (a quadratic differential) the ambiguity is the $3p-3$ quadratic differentials of Section 5.6.
- *Weierstrass points are landmarks.* They are defined by the complex structure alone, so any symmetry of the surface must permute them. Since there are finitely many, a surface of genus $p \geq 2$ has only finitely many symmetries. This is the fact behind the mild (orbifold) singularities of $\mathcal{M}_p$ found in Section 5.5.
- *They reappear in string and conformal field theory* (Chapter 11). Nash notes that the gap theorem constrains correlation functions of conformal field theories on Σ. The determinants of zero-mode wavefunctions of b–c ghost systems, which appear in bosonisation formulas for loop amplitudes, are built from the same Wronskian determinants whose zeros define the Weierstrass points.

The whole section is a single application of the Riemann–Roch theorem, which is the index theorem for $\bar\partial$ that we used in Section 5.6. This time we apply it not to K^2 but to the line bundle of a divisor nz_0.

5.7.1 Poles on the Riemann sphere: no restrictions

Start where everything is allowed. On the sphere $\mathbb{CP}^1 = \mathbb{C} \cup \{\infty\}$ a meromorphic function with a single pole at $z_0 \in \mathbb{C}$ may behave near z_0 like [Nash p. 166, (5.100)]

$$f(z) \;\to\; \frac{c}{(z - z_0)^n} + \cdots, \qquad c \neq 0, \tag{5.163}$$

for *any* $n \geq 1$. The dots are less singular terms.

Watch out: Two sign conventions for the order of a pole

In this section (Nash's footnote 8) a pole of order n has $n > 0$: $1/(z - z_0)^3$ has "a pole of order 3". In the divisor bookkeeping of Chapter 5 the *order* of f at z_0 was a signed integer, negative at poles, so the same function has $\operatorname{ord}_{z_0} f = -3$ and contributes $-3z_0$ to $\operatorname{div} f$. The two are related by

$$f \text{ has a pole of order at most } n \text{ at } z_0 \quad \Longleftrightarrow \quad \operatorname{ord}_{z_0} f \;\geq\; -n.$$

We use the positive convention for pole orders and the signed one inside divisors.

Worked example 5.97: Every pole order occurs on the sphere

Take $f_n(z) = (z - z_0)^{-n}$ with $z_0 \in \mathbb{C}$. On $\mathbb{C}$ it is holomorphic except at z_0, where it has a pole of order n. Is it harmless at ∞? Use the coordinate $w = 1/z$ there:

$$f_n \overset{(1)}{=} \left(\frac{1}{w} - z_0\right)^{-n} \overset{(2)}{=} \frac{w^n}{(1 - z_0 w)^n}.$$

(1) Substitute $z = 1/w$.
(2) Multiply numerator and denominator by w^n.

At $w = 0$ the denominator is 1, so f_n is holomorphic at ∞, with a zero of order n there. So f_n has exactly one pole, of order n, and every $n \geq 1$ is allowed: *the sphere has no gaps*.
What is the full space of functions with a pole of order at most n at z_0 and no other poles? Let f be such a function, and let $P(z) = \sum_{k=1}^{n} a_k (z - z_0)^{-k}$ be its principal part (the singular terms of its Laurent series at z_0). Then:

$$\begin{aligned} f - P \text{ is bounded on } \mathbb{CP}^1 &\overset{(1)}{\Longrightarrow} f - P = a_0 \text{ (a constant)} \\ &\overset{(2)}{\Longrightarrow} f = a_0 + \frac{a_1}{z - z_0} + \cdots + \frac{a_n}{(z - z_0)^n}. \end{aligned}$$

(1) $f - P$ is holomorphic on $\mathbb{C}$ (the singular terms have been removed at z_0) and has a finite limit at ∞ (both f and P do). A bounded entire function is constant: Liouville's theorem.
(2) Rearrange.

So the space is spanned by $1, (z - z_0)^{-1}, \ldots, (z - z_0)^{-n}$, which are linearly independent: its dimension is $n + 1$, and the $n + 1$ Laurent coefficients $a_0, \ldots, a_n$ can be chosen freely. For $z_0 = \infty$ the same space is the space of polynomials of degree at most n.

This "free choice of every coefficient" is exactly what fails at higher genus. It is worth fixing the notation for the space we just computed.

Working definition 5.50 (The space $\mathcal{L}(nz_0)$ and its dimension $\ell(n)$)

For a point z_0 of a compact Riemann surface Σ and an integer $n \geq 0$,

$$\mathcal{L}(nz_0) = \{f \text{ meromorphic on } \Sigma : f \text{ is holomorphic on } \Sigma \setminus \{z_0\} \text{ and has a pole of order} \leq n \text{ at } z_0\}$$

[meromorphic means holomorphic except at isolated poles. A pole of order $\leq n$ means $z^n f$ stays bounded near z_0]. In divisor language $\mathcal{L}(nz_0) = \{f : \operatorname{div} f + nz_0 \geq 0\} \cup \{0\}$ [$\operatorname{div} f$ lists the zeros of f with + their order and the poles with − their order. ≥ 0 says no point is left with a negative count], as in Chapter 5. We write $\ell(n) := \dim_{\mathbb{C}} \mathcal{L}(nz_0)$. *Operationally:* write down the most general Laurent tail $a_0 + a_1 z^{-1} + \cdots + a_n z^{-n}$ in a local coordinate z centred at z_0 ($z(z_0) = 0$) and ask how many of these tails extend to a function on all of Σ with no other poles. On $\mathbb{CP}^1$, $\ell(n) = n + 1$.

The spaces are nested, $\mathcal{L}(0) \subset \mathcal{L}(z_0) \subset \mathcal{L}(2z_0) \subset \cdots$, because a pole of order at most $n - 1$ is in particular a pole of order at most n.

5.7.2 The first restriction: no function has a single simple pole

Nash begins with the simplest gap [Nash p. 167]: on a surface of genus $p \geq 1$ the order $n = 1$ is never allowed. Here are three ways to see it, from the most general to the most explicit.

The degree argument

Key idea: A single simple pole forces genus zero

If a meromorphic function f on a compact Riemann surface Σ has exactly one pole, and that pole is simple, then $f : \Sigma \to \mathbb{CP}^1$ is a biholomorphism, so Σ is a sphere. Hence on a surface of genus $p \geq 1$, the order 1 is a gap at every point.

The mechanism is that "number of zeros equals number of poles" holds not only for f but for $f - c$, for every constant c. Fix $c \in \mathbb{C}$:

$$\#\{\text{zeros of } f - c\} \overset{(1)}{=} \#\{\text{poles of } f - c\} \overset{(2)}{=} \#\{\text{poles of } f\} \overset{(3)}{=} 1.$$

(1) The divisor of a meromorphic function has degree zero (Chapter 5). Recall how: integrate $\mathrm{d}g/g$ with $g = f - c$ around the boundary of the $4p$-gon obtained by cutting Σ open; the sides are glued in pairs with opposite orientation and cancel, so $\frac{1}{2\pi i} \oint \mathrm{d}g/g = \#\text{zeros} - \#\text{poles} = 0$. Both counts include multiplicity.

(2) Subtracting a constant does not move or change the poles.

(3) Assumption: one pole, of order one.

So f takes every finite value c exactly once, counted with multiplicity, and takes the value ∞ exactly once (at z_0). "With multiplicity one" means that $f - c$ has a simple zero, so $f' \neq 0$ there. So f is a bijection $\Sigma \to \mathbb{CP}^1$ with non-vanishing derivative in every local coordinate; its inverse is holomorphic by the inverse function theorem. Therefore $\Sigma \cong \mathbb{CP}^1$ and $p = 0$.

Physical picture: A function is a branched covering

The same count shows more: a function whose poles add up to n (with multiplicity) takes *every* value n times. It presents Σ as an n-sheeted covering of the sphere, branched where

sheets meet. We used this in Riemann's count of $3p-3$ (Section 5.6.12). A 1-sheeted covering is no covering at all: it is the sphere itself. A 2-sheeted covering is a hyperelliptic surface, and we will meet it again as the surfaces whose Weierstrass points are the branch points.

The residue argument on the torus

Nash's argument is via residues. It is cleanest on the torus, where there is a global coordinate z and a global 1-form $\mathrm{d}z$. A meromorphic function on $\Sigma = \mathbb{C}/(\mathbb{Z}\omega_1 \oplus \mathbb{Z}\omega_2)$ is an **elliptic function**: a meromorphic function on $\mathbb{C}$ with $f(z+\omega_1) = f(z+\omega_2) = f(z)$. Choose the orientation so that $\mathrm{Im}(\omega_2/\omega_1) > 0$ (in the notation of Section 5.5, $\omega_1 = 1$ and $\omega_2 = \tau$). Let Π be a fundamental parallelogram with corners a, $a+\omega_1$, $a+\omega_1+\omega_2$, $a+\omega_2$, shifted by a generic a so that no pole lies on its boundary (Figure 5.54).

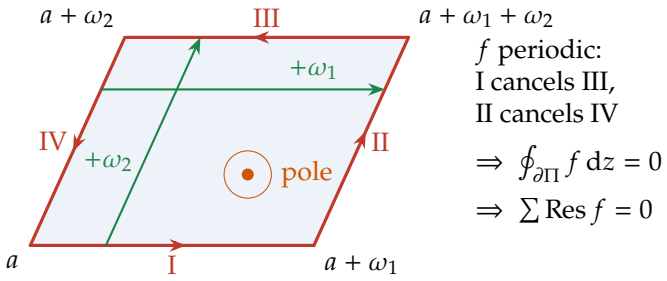


Figure 5.54: The residue theorem on a torus. The boundary of a fundamental parallelogram consists of two pairs of opposite sides. Translation by ω_2 carries side I onto side III, and by ω_1 carries IV onto II. A periodic function takes the same values on both sides of each pair, but the sides are traversed in opposite directions, so their contributions cancel. Hence the residues of an elliptic function add up to zero, and a single simple pole (residue $c \neq 0$) is impossible.

Write the boundary integral side by side:

$$
\begin{aligned}
2\pi i \sum_{\text{poles in } \Pi} \mathrm{Res}\, f &\overset{(1)}{=} \oint_{\partial\Pi} f(z)\,\mathrm{d}z \\
&\overset{(2)}{=} \int_a^{a+\omega_1} f\,\mathrm{d}z + \int_{a+\omega_1}^{a+\omega_1+\omega_2} f\,\mathrm{d}z - \int_{a+\omega_2}^{a+\omega_1+\omega_2} f\,\mathrm{d}z - \int_a^{a+\omega_2} f\,\mathrm{d}z \\
&\overset{(3)}{=} \int_a^{a+\omega_1} \big[f(z) - f(z+\omega_2)\big]\mathrm{d}z + \int_a^{a+\omega_2} \big[f(z+\omega_1) - f(z)\big]\mathrm{d}z \\
&\overset{(4)}{=} 0.
\end{aligned}
\tag{5.164}
$$

(1) The residue theorem of complex analysis for the meromorphic 1-form $f\,\mathrm{d}z$ inside Π.
(2) The boundary is traversed counter-clockwise: I, II, then III and IV backwards.
(3) Shift the variable: on III put $z \to z+\omega_2$, on II put $z \to z+\omega_1$; then pair I with III and II with IV.
(4) Periodicity makes each bracket vanish identically.

If f had only one pole in Π, simple with principal part $c/(z-z_0)$, the left side would be $2\pi i c \neq 0$. Contradiction.

Worked example 5.98: The Weierstrass ζ-function: as close as one can get

Is there anything on $\mathbb{C}$ with a simple pole at each lattice point and no other poles, that is "almost" periodic? Yes. With $\Lambda = \mathbb{Z}\omega_1 \oplus \mathbb{Z}\omega_2$ and $\sum'$ the sum over $\omega \in \Lambda \setminus \{0\}$, let

$$\wp(z) = \frac{1}{z^2} + \sum_{\omega}{}' \Big[\frac{1}{(z-\omega)^2} - \frac{1}{\omega^2} \Big], \qquad \zeta(z) = \frac{1}{z} + \sum_{\omega}{}' \Big[\frac{1}{z-\omega} + \frac{1}{\omega} + \frac{z}{\omega^2} \Big].$$

Both series converge (the terms are $O(|\omega|^{-3})$), and differentiating term by term gives $\zeta' = -\wp$. Since $\wp$ is elliptic, $\frac{\mathrm{d}}{\mathrm{d}z}\big[\zeta(z+\omega_i) - \zeta(z)\big] = -\wp(z+\omega_i) + \wp(z) = 0$, so

$$\zeta(z+\omega_i) = \zeta(z) + \eta_i, \qquad i = 1, 2,$$

with constants η_i. The function ζ has exactly one simple pole per parallelogram, with residue 1, and it is *not* periodic. Run (5.164) for ζ instead of f:

$$\begin{aligned} 2\pi i \cdot 1 &\overset{(1)}{=} \int_a^{a+\omega_1} \big[\zeta(z) - \zeta(z+\omega_2)\big]\mathrm{d}z + \int_a^{a+\omega_2} \big[\zeta(z+\omega_1) - \zeta(z)\big]\mathrm{d}z \\ &\overset{(2)}{=} \int_a^{a+\omega_1} (-\eta_2)\,\mathrm{d}z + \int_a^{a+\omega_2} \eta_1\,\mathrm{d}z \\ &\overset{(3)}{=} \eta_1\omega_2 - \eta_2\omega_1. \end{aligned}$$

(1) Steps (1)–(3) of (5.164), which did not use periodicity; the residue of ζ at its one pole in Π is 1.
(2) Quasi-periodicity.
(3) The integrands are constants.

This is **Legendre's relation** $\eta_1\omega_2 - \eta_2\omega_1 = 2\pi i$. It says precisely that ζ *cannot* be periodic: if $\eta_1 = \eta_2 = 0$ the left side would vanish. The would-be function with one simple pole exists only as a multivalued object, and its failure to be single valued is forced to be exactly the amount the residue demands.
On the other hand the *difference*

$$g(z) = \zeta(z - b_1) - \zeta(z - b_2), \qquad b_1 \neq b_2,$$

is elliptic, because the η_i cancel. It has two simple poles with residues $+1$ and -1, adding to zero as (5.164) requires. So on the torus the pole order 1 is a gap, but the pole divisor $b_1 + b_2$ is allowed; and at a single point, order 2 is allowed: $\wp(z - z_0)$.

Physical picture: Gauss's law on a compact surface

Think of $f\,\mathrm{d}z$ on the torus as the complex electric field of point charges in two dimensions: a simple pole $c/(z - z_0)$ is a point charge (together with a point vortex) of strength $\propto c$. The residue theorem (5.164) says that the total charge on a compact surface is zero, because the flux has nowhere to go. That is the same reason a periodic box in lattice electrostatics needs a neutralising background: a single charge on a torus is inconsistent. A dipole $(\zeta(z - b_1) - \zeta(z - b_2))$ is fine. The quasi-periodic ζ is the field of a single charge *plus* a uniform background, which is not periodic.

What Nash's residue argument means in higher genus

Nash says [Nash p. 167] that the sum of the residues of a meromorphic function on Σ_p is zero, by running the $4p$-gon contour argument of Chapter 5 on f instead of on $\mathrm{d}f/f$. This needs one repair.

Watch out: A residue belongs to a 1-form, not to a function

The residue of f at z_0 is the coefficient of $(z-z_0)^{-1}$ in a local coordinate z, and it *changes* if we change coordinate. Example: $f = 1/z^2$ has residue 0 in the coordinate z; in the coordinate $w = z+z^2$ (so $z = w-w^2+O(w^3)$) it is $f = w^{-2}(1-w+\dots)^{-2} = w^{-2}+2w^{-1}+\dots$, with residue 2. What *is* coordinate independent is the residue of a meromorphic 1-form $\alpha = a(z)\,\mathrm{d}z$, because $\frac{1}{2\pi i}\oint \alpha$ around a small loop does not care about coordinates. The residue theorem on a compact surface is a statement about 1-forms:

$$\sum_P \operatorname{Res}_P \alpha = 0 \qquad \text{for every meromorphic 1-form } \alpha \text{ on } \Sigma.$$

(Proof: α is closed away from its poles; Stokes' theorem on Σ minus small discs around the poles, $\int_\Sigma \mathrm{d}\alpha = 0 = \sum_P \oint_{\partial D_P} \alpha$.) On the torus there is a preferred 1-form $\mathrm{d}z$, and "the residue of f" means the residue of $f\,\mathrm{d}z$. On a surface of higher genus we must pick a holomorphic 1-form ω and apply the theorem to $f\omega$.

With that repair the argument reads as follows. Suppose f has a single simple pole at z_0, $f = c/z + O(1)$ with $c \neq 0$. Take any holomorphic differential $\omega = h(z)\,\mathrm{d}z$. Then $f\omega$ is a meromorphic 1-form with at most a simple pole at z_0 and residue $c\,h(0)$. The residue theorem gives $c\,h(0) = 0$, i.e. *every* holomorphic differential vanishes at z_0. This is a contradiction as soon as some holomorphic differential is non-zero at z_0. On a torus $\omega = \mathrm{d}z$ never vanishes. In genus $p \geq 2$ it is true that the p holomorphic differentials never all vanish at the same point, but that fact is itself equivalent to 1 being a gap, as we will see in Section 5.7.6. The degree argument above avoids this circle.

5.7.3 Gaps, non-gaps and the gap theorem

Nash now states the general result. We first fix the vocabulary.

Working definition 5.51 (Gap, non-gap, gap sequence, weight)

Fix a point z_0 on a compact Riemann surface Σ of genus p [p is the number of handles]. An integer $n \geq 1$ is a **non-gap** at z_0 if there is a meromorphic function on Σ that is holomorphic on $\Sigma \setminus \{z_0\}$ and has a pole of order *exactly* n at z_0. Otherwise n is a **(Weierstrass) gap** at z_0. The list of gaps $n_1 < n_2 < \cdots$ is the **gap sequence** of z_0, and

$$w(z_0) := \sum_{i=1}^{p} (n_i - i)$$

[there are exactly p gaps, by the theorem below. w measures how far they sit above $1, 2, \dots, p$] is its **weight**. *Operational test:* n is a non-gap iff $\ell(n) = \ell(n-1) + 1$ [$\ell(n)$ is the number of independent functions with no poles except one of order $\leq n$ at z_0], i.e. allowing one more order of pole produces a new function (proved in Section 5.7.5).

Key idea: The Weierstrass gap theorem (Nash, p. 167)

Let Σ be a compact Riemann surface of genus $p \geq 1$ and $z_0 \in \Sigma$.

(i) There are exactly p gaps at z_0, and they satisfy

$$1 = n_1 < n_2 < \cdots < n_p \leq 2p - 1.$$

Every $n \geq 2p$ is a non-gap.

(ii) For all but finitely many z_0 the gaps are $1, 2, \ldots, p$ (weight 0). Such z_0 are **normal points**; at them no function with only one pole has a pole of order $\leq p$.

(iii) The finitely many exceptional points, where the gap sequence is anything else (weight > 0), are the **Weierstrass points**. At a Weierstrass point some $n \leq p$ is a non-gap.

For genus 0 there are no gaps; for genus 1 the single gap is $n_1 = 1$ at every point, so every point is normal and there are no Weierstrass points.

The bound $n_p \leq 2p - 1$ is Nash's "$n_p < 2p$". Two quick consequences follow from the multiplicative structure of functions.

Key idea: The non-gaps form a semigroup

If n and m are non-gaps at z_0, so is $n + m$.

Take f and g with single poles of exact orders n and m at z_0. In a local coordinate centred at z_0,

$$f\,g \overset{(1)}{=} \left(az^{-n} + \ldots\right)\left(bz^{-m} + \ldots\right) \overset{(2)}{=} ab\,z^{-(n+m)} + \ldots,$$

(1) Leading terms, with $a \neq 0$, $b \neq 0$; the dots are less singular.
(2) Multiply; $ab \neq 0$.

and fg is holomorphic wherever both f and g are, i.e. away from z_0. So fg has a single pole of order exactly $n + m$. The semigroup property is already a strong constraint.

Worked example 5.99: If 2 is a non-gap, the gaps are the odd numbers

Suppose some function has a single double pole at z_0 (on a surface of genus $p \geq 1$). Then $2, 4, 6, \ldots$ are all non-gaps by the semigroup property. Now count using part (i) of the gap theorem:

$$\#\{\text{gaps}\} \overset{(1)}{=} p,$$

$$\#\{\text{gaps}\} \overset{(2)}{\leq} \#\{\text{odd numbers in } 1, \ldots, 2p-1\} \overset{(3)}{=} p.$$

(1) The gap theorem.
(2) All gaps lie in $\{1, \ldots, 2p-1\}$ and none is even.
(3) The odd numbers there are $1, 3, \ldots, 2p - 1$.

Equality forces the gap sequence to be exactly $1, 3, 5, \ldots, 2p - 1$, with weight

$$w(z_0) = \sum_{i=1}^{p} \left((2i - 1) - i\right) = \sum_{i=1}^{p} (i - 1) = \frac{p(p-1)}{2}.$$

This is the largest weight a point can have (a standard fact we do not prove). For $p = 1$ the weight is 0: on the torus $\wp(z - z_0)$ has a double pole at every point, and the gap sequence

{1} is also the normal one. For $p \geq 2$, a point with non-gap 2 is a Weierstrass point. A function with a single double pole is a degree-2 map to $\mathbb{CP}^1$, so the surface is **hyperelliptic**; we will identify these points as the branch points in Section 5.7.8.

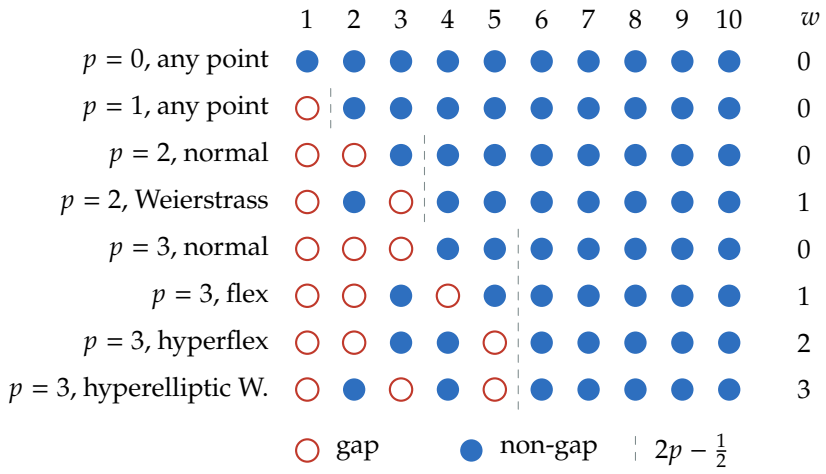


Figure 5.55: Gap sequences. Each row lists the pole orders $n = 1, \ldots, 10$ at one point: open red circles are gaps (no function with only that pole exists), filled blue circles are non-gaps. Every row has exactly p gaps, all to the left of the dashed line at $2p - \frac{1}{2}$, and the gap 1 is always present for $p \geq 1$. Normal points have the gaps packed at the start; at Weierstrass points a gap has moved to the right, and the weight w in the last column measures how far. Every row respects the semigroup property: sums of non-gaps are non-gaps. The genus-2 and genus-3 examples are worked out in Section 5.7.8.

Try it: Which gap sequences are possible in genus 2?

Using only the gap theorem and the semigroup property, list all possible gap sequences in genus 2.
Answer. We need $1 = n_1 < n_2 \leq 3$, so $\{1, 2\}$ or $\{1, 3\}$. Both satisfy the semigroup property: for $\{1, 2\}$ the non-gaps $3, 4, 5, \ldots$ are closed under addition; for $\{1, 3\}$ the non-gaps $2, 4, 5, 6, \ldots$ are too. Both occur (Section 5.7.8).

5.7.4 Counting functions with Riemann–Roch

Nash's outline of the proof [Nash pp. 167–169, (5.101)–(5.110)] turns the question "which f exist?" into "how many holomorphic sections does a certain line bundle have?", and then answers it with the index theorem.

The divisor nz_0 and its line bundle

Let [Nash p. 167, (5.101)]

$$D = nz_0, \qquad n \geq 1, \qquad \deg D = n > 0. \tag{5.165}$$

Recall from Chapter 5 how a divisor defines a line bundle $[D]$. We make it completely explicit for this D.

- *Cover.* Let U_0 be a coordinate disc around z_0, with coordinate z, $z(z_0) = 0$. Cover the rest of Σ by patches U_α ($\alpha \neq 0$) that do not contain z_0.

- *Local equations of D.* On U_0 the divisor is "a zero of order n at $z = 0$", with local equation $f_0 = z^n$. On U_α the divisor is empty, with local equation $f_\alpha = 1$.
- *Transition functions.* $g_{0\alpha} = f_0/f_\alpha = z^n$ on $U_0 \cap U_\alpha$. This is holomorphic and non-zero there, because $z_0 \notin U_\alpha$. The cocycle condition holds automatically, since the g's are ratios.
- *The canonical section.* The local functions $s_0 = z^n$, $s_\alpha = 1$ satisfy $s_0 = g_{0\alpha} s_\alpha$, so they define a global holomorphic section s of $[D]$. It vanishes exactly at z_0, to order n. This is Nash's "s is really the localisation of the divisor D".

In particular $[D]$ has a non-zero holomorphic section; this is Nash's remark that "because the degree of D is positive, $[D]$ has a global holomorphic section". (For a general line bundle of positive degree that is not automatic; here we constructed the section by hand.) Its degree is

$$c_1([nz_0]) = \deg(nz_0) = n,$$

the number of zeros of s (Chapter 5, Nash's (5.11)).

Sections of $[D]$ are functions with a restricted pole

Now let f be meromorphic with its only singularity at z_0, a pole of order at most n. Consider the product fs, with local components

$$(fs)_0 = f \cdot z^n \ \text{ on } U_0, \qquad (fs)_\alpha = f \ \text{ on } U_\alpha.$$

On U_α this is holomorphic because f is. On U_0, $f = O(z^{-n})$, so $z^n f$ is holomorphic too. So fs is a holomorphic section of $[D]$. Conversely, any holomorphic section σ of $[D]$ gives the meromorphic function $f = \sigma/s$ (the ratio of two sections of the same bundle is a function), holomorphic away from z_0 and with a pole of order at most n at z_0 (dividing by z^n). So [Nash p. 168, (5.102)]

$$\mathcal{L}(nz_0) \xrightarrow{\ f \mapsto fs\ } H^0\big(\Sigma_p, \mathcal{O}([D])\big) = H^0\big(\Sigma_p, \mathcal{O}([nz_0])\big) \tag{5.166}$$

is an isomorphism. Two remarks from Nash:

- A function in $\mathcal{L}(nz_0)$ need not be singular. If it is holomorphic at z_0 too, it is holomorphic on a compact surface, so it is constant (maximum modulus principle). Hence the constants are always in $\mathcal{L}(nz_0)$ and $\ell(n) \geq 1$.
- If *all* of $\mathcal{L}(nz_0)$ were holomorphic, $\ell(n)$ would take its minimum value 1.

The naive count and why it is wrong

A function $f \in \mathcal{L}(nz_0)$ has a Laurent expansion at z_0 [Nash (5.103)]

$$f(z) \ \to\ a_0 + \frac{a_1}{z - z_0} + \cdots + \frac{a_n}{(z - z_0)^n} + O(z - z_0). \tag{5.167}$$

It depends on the $n + 1$ parameters $a_0, \ldots, a_n$, and on the sphere the function is determined by them and they are free (Example 5.97). So one might conjecture [Nash (5.104)]

$$\dim \mathcal{L}(nz_0) = \dim H^0\big(\Sigma_p, \mathcal{O}([nz_0])\big) \stackrel{?}{=} n + 1. \tag{5.168}$$

The first half is right: f is determined by its principal part up to a constant, because the difference of two functions in $\mathcal{L}(nz_0)$ with the same principal part is holomorphic, hence

constant. So $\ell(n) \le n+1$. What fails is the *freedom* of the coefficients. The gap theorem contradicts (5.168) in two ways [Nash (5.105)–(5.106)]:

$$\dim \mathcal{L}(nz_0) = 1 \qquad \text{if } n \le p \text{ and } z_0 \text{ is a normal point,} \tag{5.169}$$

$$\dim \mathcal{L}(nz_0) = n+1-p \qquad \text{if } n \ge 2p-1. \tag{5.170}$$

In (5.169) only the constant survives: none of $a_1, \dots, a_n$ can be non-zero. In (5.170) "there are p gaps in the parameters a_i": the $n+1$ coefficients are subject to p independent linear conditions. We will see exactly what these conditions are in Section 5.7.6. First, Nash's derivation of (5.170) from the index theorem.

The index theorem for $\bar\partial_{[nz_0]}$

The space (5.166) is the kernel of the Dolbeault operator $\bar\partial_{[nz_0]}$ of the line bundle, the same kind of operator as $\bar\partial_{K^2}$ in Section 5.6. The Riemann–Roch (index) theorem [Nash p. 169, (5.107)] gives

$$\operatorname{index} \bar\partial_{[nz_0]} = \dim \operatorname{Ker} \bar\partial_{[nz_0]} - \dim \operatorname{Ker} \bar\partial^*_{[nz_0]} = \int_{\Sigma_p} \operatorname{ch}([nz_0]) \wedge \operatorname{td}(T\Sigma_p). \tag{5.171}$$

The integrand is evaluated exactly as in (5.148) [Nash (5.108)]:

$$\begin{aligned} \operatorname{index} \bar\partial_{[nz_0]} &\overset{(1)}{=} \int_{\Sigma_p} \big(1 + c_1([nz_0])\big) \wedge \big(1 + \tfrac{1}{2}c_1(T\Sigma_p)\big) \\ &\overset{(2)}{=} \int_{\Sigma_p} c_1([nz_0]) + \frac{1}{2}\int_{\Sigma_p} c_1(T\Sigma_p) \\ &\overset{(3)}{=} n + \tfrac{1}{2}(2-2p) \\ &= n+1-p. \end{aligned} \tag{5.172}$$

(1) For a line bundle on a surface $\operatorname{ch} = 1 + c_1$ and $\operatorname{td} = 1 + \frac{1}{2}c_1$ (Section 5.6.8).
(2) Only the 2-form part integrates; the product of the two c_1's is a 4-form, zero on a surface.
(3) $\int c_1([nz_0]) = \deg(nz_0) = n$, and Gauss–Bonnet $\int c_1(T\Sigma_p) = \chi(\Sigma_p) = 2-2p$ (5.146).

This is the familiar $\deg D + 1 - p$ of Riemann–Roch.

Killing the cokernel: Serre duality and vanishing

The index is the dimension we want *minus* $\dim \operatorname{Ker} \bar\partial^*_{[nz_0]}$. By Hodge theory that kernel is $H^1(\Sigma_p, \mathcal{O}([nz_0]))$, and then [Nash (5.109)]

$$\begin{aligned} H^1\big(\Sigma_p, \mathcal{O}([nz_0])\big) &\overset{(1)}{\cong} \big\{H^0\big(\Sigma_p, \mathcal{O}(K_{\Sigma_p} \otimes [nz_0]^*)\big)\big\}^*, \\ H^0\big(\Sigma_p, \mathcal{O}(K_{\Sigma_p} \otimes [nz_0]^*)\big) &\overset{(2)}{=} 0 \qquad \text{if } n \ge 2p-1. \end{aligned} \tag{5.173}$$

(1) Serre duality (5.135) with $E = [nz_0]$.
(2) The bundle $K \otimes [nz_0]^*$ has degree $c_1(K) + c_1([nz_0]^*) = (2p-2) - n$, which is negative exactly when $n > 2p-2$, i.e. $n \ge 2p-1$. A line bundle of negative degree has no non-zero holomorphic sections (Section 5.6.9: a section would have a negative number of zeros). Nash quotes this as the Kodaira vanishing theorem for negative line bundles.

What are the sections of $K \otimes [nz_0]^*$ concretely? The bundle $[nz_0]^* = [-nz_0]$ has transition functions z^{-n}, so a section of $K \otimes [-nz_0]$ is a collection of local 1-forms ω_α with $\omega_0 = z^{-n}\omega_\alpha$ on

overlaps; equivalently a global holomorphic 1-form ω (= ω_α away from z_0) such that $z^{-n}\omega$ is still holomorphic at z_0. That is:

$$H^0\big(\Sigma_p, \mathcal{O}(K \otimes [-nz_0])\big) \;\cong\; \big\{\omega \in H^0(\Sigma_p, \mathcal{O}(K)) : \ \omega \text{ vanishes to order} \geq n \text{ at } z_0\big\}, \tag{5.174}$$

whose dimension we call $\kappa(n)$. A holomorphic differential has $\deg K = 2p-2$ zeros, so it cannot vanish to order $n \geq 2p-1$ unless it is zero: this is the vanishing in (5.173) again, seen directly.

Putting (5.172) and (5.173) together gives Nash's (5.110): for $n \geq 2p-1$,

$$\dim H^0\big(\Sigma_p, \mathcal{O}([nz_0])\big) = \dim \mathcal{L}(nz_0) = n+1-p, \tag{5.175}$$

in agreement with (5.170). And for every $n \geq 0$ we have the exact statement

$$\boxed{\ell(n) = n+1-p+\kappa(n)} \tag{5.176}$$

where

$$\kappa(n) = \dim\big\{\omega \text{ holomorphic differential} : \ \omega \text{ vanishes to order} \geq n \text{ at } z_0\big\}.$$

At $n=0$: $\ell(0)=1$ and $\kappa(0)=p$ (all differentials), and indeed $1 = 0+1-p+p$. ✓

Watch out: What the vanishing theorem does and does not give

For $n \leq 2p-2$ the bundle $K \otimes [-nz_0]$ has degree ≥ 0, no vanishing theorem applies, and $\kappa(n)$ genuinely depends on the point z_0. That dependence is the whole content of the distinction between normal points and Weierstrass points. The index theorem alone knows only the topological number $n+1-p$; the fine structure lives in the cokernel.

5.7.5 The staircase: proof of the gap theorem

Nash says "we can outline a proof" and stops at (5.175). The rest of the proof is short, and it is worth writing out because it shows how much information is in the single equation (5.175).

Step 1: the staircase starts at 1. $\ell(0)=1$: a function with no poles on a compact surface is constant.

Step 2: each step has height 0 *or* 1. For $n \geq 1$ consider the linear map "read off the coefficient of z^{-n}":

$$c_n : \ \mathcal{L}(nz_0) \to \mathbb{C}, \qquad f = a_n z^{-n} + \ldots \ \mapsto \ a_n.$$

Its kernel is the set of f whose pole has order at most $n-1$, i.e. $\mathcal{L}((n-1)z_0)$. So

$$\ell(n) \overset{(1)}{=} \dim \operatorname{Ker} c_n + \dim \operatorname{Im} c_n \overset{(2)}{=} \ell(n-1) + \dim \operatorname{Im} c_n \overset{(3)}{\leq} \ell(n-1)+1.$$

(1) Rank–nullity for the linear map c_n.
(2) $\operatorname{Ker} c_n = \mathcal{L}((n-1)z_0)$.
(3) The image is a subspace of $\mathbb{C}$.

Moreover $\dim \operatorname{Im} c_n = 1$ exactly when some $f \in \mathcal{L}(nz_0)$ has $a_n \neq 0$, i.e. has a pole of order exactly n. So

$$n \text{ is a non-gap} \iff \ell(n) = \ell(n-1)+1, \qquad n \text{ is a gap} \iff \ell(n) = \ell(n-1). \tag{5.177}$$

This is the operational test promised in the working definition.

Step 3: count the steps. By (5.175) at $n = 2p-1$, $\ell(2p-1) = (2p-1)+1-p = p$. Between $n = 0$ and $n = 2p-1$ the staircase climbs from 1 to p in $2p-1$ steps of height 0 or 1:

$$\#\{\text{non-gaps in } 1,\dots,2p-1\} \overset{(1)}{=} \sum_{n=1}^{2p-1} \big(\ell(n)-\ell(n-1)\big) \overset{(2)}{=} \ell(2p-1)-\ell(0) \overset{(3)}{=} p-1,$$

$$\#\{\text{gaps in } 1,\dots,2p-1\} \overset{(4)}{=} (2p-1)-(p-1) = p.$$

(1) Each non-gap contributes a step of height 1, each gap a step of height 0, by (5.177).
(2) The sum telescopes.
(3) $\ell(2p-1) = p$ and $\ell(0) = 1$.
(4) Every integer in $1,\dots,2p-1$ is either a gap or a non-gap.

Step 4: no gaps beyond $2p-1$. For $n \geq 2p$ both n and $n-1$ satisfy $\geq 2p-1$, so by (5.175) $\ell(n)-\ell(n-1) = (n+1-p)-(n-p) = 1$: every $n \geq 2p$ is a non-gap.

Step 5: the first gap. 1 is a gap for $p \geq 1$ by Section 5.7.2. (Alternatively: if $\ell(1) = 2$, some function of degree 1 exists, which makes Σ a sphere.)

That proves part (i) of the gap theorem: exactly p gaps, starting at 1, all at most $2p-1$. For $p = 0$ the same argument, with $\ell(n) = n+1$ for all n from (5.176) (there are no differentials, so $\kappa = 0$), shows every step has height one: no gaps.

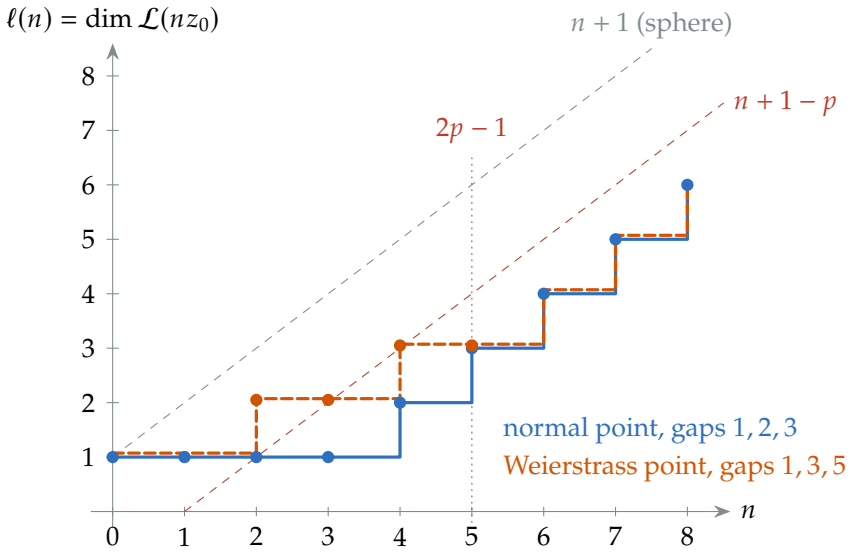


Figure 5.56: The staircase $\ell(n)$ for genus $p = 3$. At each n the staircase either stays flat (a gap) or rises by one (a non-gap). The grey line $n+1$ is the naive count, correct only on the sphere. The red line $n+1-p$ is index $\bar{\partial}_{[nz_0]}$; the staircase lies on or above it, the excess being $\kappa(n)$, and it joins the line for good at $n = 2p-1 = 5$ at the latest, where the cokernel is killed by the vanishing theorem (at a normal point it joins already at $n = p+1$). A normal point (blue) stays flat for the first p steps; a Weierstrass point (orange, here a branch point of a hyperelliptic surface) rises earlier. Both must reach height $p = 3$ at $n = 5$, so both have exactly 3 flat steps.

5.7.6 Where the gaps come from: residues of $f\omega$

We now know how many gaps there are. Which ones they are is decided by the holomorphic differentials, through the residue theorem. This makes Nash's phrase "p gaps in the parameters a_i" precise, and it is the most concrete way to see Riemann–Roch at work.

p conservation laws for the principal part

Let $f \in \mathcal{L}(nz_0)$ with principal part $\sum_{k=1}^{n} a_k z^{-k}$ in a local coordinate z centred at z_0. Let $\omega = h(z)\,dz$ be any holomorphic differential, with Taylor series $h(z) = \sum_{m\geq 0} h_m z^m$ near z_0. Then $f\omega$ is a meromorphic 1-form whose only pole is at z_0, and

$$\begin{aligned} 0 &\overset{(1)}{=} \operatorname{Res}_{z_0}(f\omega) \\ &\overset{(2)}{=} \operatorname{Res}_{z=0}\Big[\Big(\sum_{k=1}^{n} a_k z^{-k} + (\text{holomorphic})\Big)\Big(\sum_{m\geq 0} h_m z^m\Big)dz\Big] \\ &\overset{(3)}{=} \sum_{k=1}^{n} a_k\, h_{k-1}. \end{aligned} \tag{5.178}$$

(1) The residue theorem for the meromorphic 1-form $f\omega$ on the compact surface: the sum of its residues is zero, and it has only one pole.
(2) Insert the Laurent series of f and the Taylor series of h.
(3) The coefficient of z^{-1} comes from $z^{-k}\cdot z^m$ with $m = k-1$; the holomorphic part of f times h has no z^{-1} term.

Choose a basis $\omega_1, \dots, \omega_p$ of holomorphic differentials (Chapter 5: there are exactly p), $\omega_j = h^{(j)}(z)\,dz$. Each gives one linear condition on the principal part. Collect them in the $p \times n$ **residue matrix**

$$R_n = \begin{pmatrix} h_0^{(1)} & h_1^{(1)} & \cdots & h_{n-1}^{(1)} \\ \vdots & & & \vdots \\ h_0^{(p)} & h_1^{(p)} & \cdots & h_{n-1}^{(p)} \end{pmatrix}, \qquad R_n \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} = 0.$$

The (j,k) entry is the $(k-1)$-th Taylor coefficient of the j-th differential at z_0.

Key idea: Riemann–Roch as p conservation laws

A principal part $(a_1, \dots, a_n)$ at z_0 comes from a meromorphic function with no other poles *if and only if* $R_n a = 0$. Hence

$$\ell(n) = 1 + \big(n - \operatorname{rank} R_n\big),$$

the 1 being the constant a_0.

"Only if" is (5.178). "If" is the statement that the residue pairings are the *only* obstruction to solving the Mittag-Leffler problem "find f with a prescribed principal part". On $\mathbb{C}$ there is no obstruction (take f = the principal part). On a compact surface the obstruction lives in $H^1(\Sigma, \mathcal{O})$, and Serre duality (5.135) identifies $H^1(\Sigma, \mathcal{O})$ with the dual of $H^0(\Sigma, \mathcal{O}(K))$, the pairing being exactly $a \mapsto \sum_k a_k h_{k-1}$. So the conditions are necessary by the residue theorem and sufficient by Serre duality. The dimension count matches (5.176): the kernel of R_n^T acting on combinations of the ω_j consists of the differentials with $h_0 = \cdots = h_{n-1} = 0$, i.e. those vanishing to order $\geq n$, so $p - \operatorname{rank} R_n = \kappa(n)$ and $\ell(n) = 1 + n - p + \kappa(n)$. ✓

Physical picture: Each holomorphic differential is a conserved charge

In a conformal field theory with a holomorphic current $j(z)$ (a 1-form), the product $f\omega$ plays the role of j: the residue theorem $\sum \text{Res} = 0$ is the Ward identity for the corresponding charge, which forces the charges of the insertions to add up to zero. A function on Σ multiplied by each of the p holomorphic differentials gives p such conservation laws. On the sphere there are no holomorphic differentials and so no constraints on f; on a torus

there is one ($\omega = \mathrm{d}z$, condition $a_1 = 0$: "no net charge"); in genus p there are p. The number of constraints that are *independent*, rank R_n, depends on where the pole sits, and that is what distinguishes the points of Σ.

Gaps are vanishing orders plus one

Now choose the basis cleverly. Order the differentials by how fast they vanish at z_0. By Gaussian elimination we may assume that their orders of vanishing are all different: if two had the same order m, a suitable combination would vanish to order $> m$. So we get a basis with

$$\omega_j = \big(c_j z^{m_j} + O(z^{m_j+1})\big)\mathrm{d}z, \quad c_j \neq 0, \qquad 0 \le m_1 < m_2 < \cdots < m_p \le 2p-2.$$

The upper bound holds because a non-zero holomorphic differential has exactly $\deg K = 2p-2$ zeros. The integers m_j are the **vanishing orders** of the holomorphic differentials at z_0; they depend only on z_0, not on the basis.

In this basis row j of R_n is $(0,\dots,0,c_j,*,*,\dots)$, with the first non-zero entry in column $k = m_j + 1$. The matrix is in echelon form, and its rank is the number of rows whose pivot fits into the first n columns:

$$\operatorname{rank} R_n = \#\{j : \ m_j + 1 \le n\}.$$

Therefore

$$\ell(n) - \ell(n-1) \overset{(1)}{=} 1 - \big(\operatorname{rank} R_n - \operatorname{rank} R_{n-1}\big) \overset{(2)}{=} 1 - \#\{j : \ m_j + 1 = n\}.$$

(1) The keyidea formula at n and at $n-1$.
(2) The only rows that gain a pivot when the n-th column is added are those with $m_j + 1 = n$; since the m_j are distinct there is at most one.

By (5.177) this is zero exactly when $n = m_j + 1$ for some j.

Key idea: Noether's characterisation of the gaps

n is a gap at z_0 if and only if there is a holomorphic differential vanishing to order *exactly* $n-1$ at z_0. Thus the gap sequence is

$$\{n_1,\dots,n_p\} = \{m_1+1,\ \dots,\ m_p+1\},$$

the vanishing orders of the holomorphic differentials, shifted by one.

This reproves the whole gap theorem in one stroke. There are p gaps because there are p differentials; $n_p \le 2p-1$ because $m_p \le 2p-2$; and $n_1 = 1$ exactly when $m_1 = 0$, i.e. when some differential does not vanish at z_0. (Combined with Section 5.7.2, this shows that for $p \ge 1$ the p differentials never vanish simultaneously: the canonical bundle has no "base points". This is the fact that Nash's residue argument needed.)

It also makes explicit what Nash's "a_i" are doing. Suppose $n = n_j$ is a gap and f had a pole of order exactly n. The condition from ω_j reads

$$0 = \sum_k a_k\, h^{(j)}_{k-1} = c_j\, a_n + \big(\text{terms with } a_k,\ k > n\big) = c_j\, a_n,$$

because $h^{(j)}_{k-1} = 0$ for $k-1 < m_j = n-1$ and $a_k = 0$ for $k > n$. So $a_n = 0$: the pole cannot have order n. In general the j-th condition expresses the coefficient a_{n_j} in terms of the higher coefficients: the p coefficients $a_{n_1},\dots,a_{n_p}$ are not free parameters.

Normal points and Weierstrass points in terms of differentials

A point is normal iff its gaps are $1, \dots, p$, iff the vanishing orders are $0, 1, \dots, p-1$. In other words: *z_0 is a Weierstrass point iff some non-zero holomorphic differential vanishes to order at least p at z_0.* This is a condition we can test with a determinant. With any basis $\omega_j = h_j(z)\,\mathrm{d}z$ (no adaptation needed), a combination $\sum_j \lambda_j \omega_j$ vanishes to order $\geq p$ iff

$$\sum_{j=1}^{p} \lambda_j \, h_j^{(k)}(z_0) = 0 \quad \text{for } k = 0, 1, \dots, p-1, \tag{5.179}$$

where $h_j^{(k)}$ is the k-th derivative. This is p equations in p unknowns; a non-zero solution exists iff the determinant vanishes. So

$$z_0 \text{ is a Weierstrass point} \iff W(z_0) \coloneqq \det \begin{pmatrix} h_1(z_0) & h_2(z_0) & \cdots & h_p(z_0) \\ h_1'(z_0) & h_2'(z_0) & \cdots & h_p'(z_0) \\ \vdots & & & \vdots \\ h_1^{(p-1)}(z_0) & h_2^{(p-1)}(z_0) & \cdots & h_p^{(p-1)}(z_0) \end{pmatrix} = 0. \tag{5.180}$$

W is the **Wronskian** of the p functions h_j. It is a holomorphic function of z_0 in each chart, and it is not identically zero (the Wronskian of linearly independent analytic functions never is). Its zeros are isolated, and on a compact surface there are finitely many of them. That is the finiteness of the set of Weierstrass points, which Nash quotes from Griffiths and Harris. In the next part we turn the Wronskian into a section of a line bundle and count its zeros.

Worked example 5.100: The residue matrix on the torus and in genus two

Torus. $p = 1$, $\omega_1 = \mathrm{d}z$, so $h^{(1)} = 1$: $h_0 = 1$, $h_m = 0$ for $m \geq 1$. At every point

$$R_n = (1, 0, 0, \dots, 0), \qquad R_n a = 0 \iff a_1 = 0.$$

So $\ell(n) = 1 + n - 1 = n$ for $n \geq 1$: the only condition is "zero residue", the only gap is 1, and every point is normal. Explicitly $\mathcal{L}(n z_0)$ is spanned by

$$1,\ \wp(z - z_0),\ \wp'(z - z_0),\ \wp(z - z_0)^2,\ \wp(z - z_0)\wp'(z - z_0),\ \dots$$

with pole orders $0, 2, 3, 4, 5, \dots$: products $\wp^a \wp'^b$ with $b \in \{0, 1\}$ have pole order $2a + 3b$, which takes every value ≥ 2 exactly once. ($\wp'^2$ is not needed: it is a cubic polynomial in $\wp$.)
Genus two at a branch point. Take $y^2 = f(x)$ with $f(x) = \prod_{i=1}^{6}(x - e_i)$, and z_0 the point $(e_1, 0)$. The local coordinate is s with $x - e_1 = s^2$, and $y = s\,u(s)$ with $u(0)^2 = f'(e_1) \neq 0$ (Example 5.91). A basis of holomorphic differentials is $\omega_1 = \mathrm{d}x/y$ and $\omega_2 = (x - e_1)\,\mathrm{d}x/y$. In the coordinate s

$$\omega_1 \overset{(1)}{=} \frac{2s\,\mathrm{d}s}{s\,u(s)} = \frac{2}{u(s)}\,\mathrm{d}s, \qquad \omega_2 \overset{(2)}{=} s^2 \cdot \frac{2}{u(s)}\,\mathrm{d}s.$$

(1) $\mathrm{d}x = 2s\,\mathrm{d}s$ and $y = s\,u(s)$.
(2) $x - e_1 = s^2$.

Now u is even in s (because $y^2/s^2 = f(e_1 + s^2)/s^2$ is a function of s^2), so $2/u(s) = \gamma_0 + \gamma_2 s^2 + O(s^4)$ with $\gamma_0 \neq 0$. The vanishing orders are $m_1 = 0$, $m_2 = 2$, and

$$R_4 = \begin{pmatrix} \gamma_0 & 0 & \gamma_2 & 0 \\ 0 & 0 & \gamma_0 & 0 \end{pmatrix}.$$

The conditions are $\gamma_0 a_1 + \gamma_2 a_3 = 0$ and $\gamma_0 a_3 = 0$, so $a_1 = a_3 = 0$ while a_2, a_4 are free: *the gaps are* 1 *and* 3. Indeed $1/(x - e_1) = s^{-2}$ has a single double pole at z_0 (it vanishes at the two points over $x = \infty$). So branch points are Weierstrass points in genus two, with weight $(1 - 1) + (3 - 2) = 1$.

5.7.7 How many Weierstrass points? The Wronskian as a section

Nash closes his outline with a statement he does not prove [Nash p. 169, (5.111)]: the Weierstrass points are finite in number, which he phrases as

$$\text{if } n \geq p, \qquad \dim \mathcal{L}(n z_0) > 1 \quad \text{for only finitely many } z_0. \tag{5.181}$$

First let us see why this *is* the finiteness of the Weierstrass points. At a normal point the gaps are $1, \dots, p$, so the staircase of Figure 5.56 is flat up to $n = p$ and $\ell(p) = 1$. At a Weierstrass point some $n \leq p$ is a non-gap, so the staircase has risen at least once by $n = p$ and $\ell(p) \geq 2$. Hence

$$\ell(p) = \dim \mathcal{L}(p z_0) > 1 \iff z_0 \text{ is a Weierstrass point},$$

and (5.181) with $n = p$ says there are finitely many of them. (For $n > p$ the statement as Nash prints it is not what he means: once $n \geq p + 1$ we have $\ell(n) \geq n + 1 - p \geq 2$ at *every* point, by the index bound $\ell(n) \geq \operatorname{index} \bar{\partial}_{[n z_0]}$. The meaningful statement is the one at $n = p$.)

Nash refers the proof to Griffiths and Harris, who count singular points of certain curves attached to Σ. We give the argument in the physicist's form: the Weierstrass points are the zeros of one holomorphic object, the Wronskian (5.180), and we can count the zeros of a holomorphic section by computing the degree of its bundle. This is the same move as "$\deg K = 2p - 2$" in Example 5.91: the number of zeros of a section is a Chern number.

Step 1: the order of vanishing of W is the weight

Fix z_0 and a coordinate z centred at z_0. Use the adapted basis of Section 5.7.6,

$$\omega_j = h_j(z)\, \mathrm{d}z, \qquad h_j(z) = c_j z^{m_j}\big(1 + O(z)\big), \qquad c_j \neq 0, \qquad m_1 < \cdots < m_p.$$

First the Wronskian of pure monomials.

Key idea: Wronskian of monomials

For distinct integers $a_1, \dots, a_p \geq 0$,

$$W\big[z^{a_1}, \dots, z^{a_p}\big] = \prod_{i<j} (a_j - a_i)\; z^{\,a_1 + \cdots + a_p - \frac{p(p-1)}{2}}. \tag{5.182}$$

The (k, i) entry of the Wronskian matrix (row $k = 0, \dots, p-1$ for the k-th derivative, column i for the function) is $\frac{\mathrm{d}^k}{\mathrm{d}z^k} z^{a_i} = (a_i)_k\, z^{a_i - k}$, where $(a)_k = a(a-1)\cdots(a-k+1)$ is the falling factorial. Then

$$W \overset{(1)}{=} \det\big[(a_i)_k\, z^{a_i - k}\big] \overset{(2)}{=} z^{\sum_i a_i}\, z^{-\sum_k k}\, \det\big[(a_i)_k\big] \overset{(3)}{=} z^{\sum_i a_i - \frac{p(p-1)}{2}}\, \det\big[a_i^{\,k}\big] \overset{(4)}{=} \prod_{i<j} (a_j - a_i)\; z^{\sum_i a_i - \frac{p(p-1)}{2}}.$$

(1) The entries, differentiated.

(2) Pull the factor z^{a_i} out of column i and z^{-k} out of row k; a determinant is linear in each row and each column.
(3) $\sum_{k=0}^{p-1} k = p(p-1)/2$. The falling factorial $(a)_k$ is a polynomial in a of degree k with leading coefficient 1: $(a)_k = a^k + (\text{lower powers})$. Subtracting suitable multiples of earlier rows from row k removes the lower powers without changing the determinant.
(4) The Vandermonde determinant.

Two checks. For $p = 2$, $a = (0, 2)$: $\det\left(\begin{smallmatrix} 1 & z^2 \\ 0 & 2z \end{smallmatrix}\right) = 2z$, and the formula gives $(2-0)z^{2-1} = 2z$. ✓ For $a = (0, 1, \dots, p-1)$ the exponent is 0: the Wronskian of $1, z, \dots, z^{p-1}$ is the non-zero constant $\prod_{k<p} k!$.

Now the general case. The Wronskian is linear in each function. Expand every h_j in its Taylor series, $h_j = \sum_{b \geq m_j} h_{j,b} z^b$. Then $W[h_1, \dots, h_p]$ is a sum of terms $\left(\prod_j h_{j,b_j}\right) W[z^{b_1}, \dots, z^{b_p}]$ with $b_j \geq m_j$. By (5.182) each such term is zero or has order $\sum_j b_j - \frac{p(p-1)}{2} \geq \sum_j m_j - \frac{p(p-1)}{2}$, with equality only for $b_j = m_j$ for all j. That lowest term is not zero, because the m_j are distinct. So

$$\operatorname{ord}_{z_0} W \overset{(1)}{=} \sum_{j=1}^{p} m_j - \frac{p(p-1)}{2} \overset{(2)}{=} \sum_{j=1}^{p} (n_j - 1) - \sum_{j=1}^{p} (j-1) \overset{(3)}{=} \sum_{j=1}^{p} (n_j - j) = w(z_0). \tag{5.183}$$

(1) The lowest term of the expansion just described.
(2) Noether's characterisation (Section 5.7.6): the gaps are $n_j = m_j + 1$; and $\frac{p(p-1)}{2} = \sum_j (j-1)$.
(3) Combine the sums.

The Wronskian vanishes at z_0 to order exactly the weight of z_0. It is non-zero at normal points (weight 0) and zero at Weierstrass points, as (5.180) already said, but now we know the multiplicity too. Changing the basis multiplies W by the (constant, non-zero) determinant of the change of basis, so the order does not depend on the basis.

Step 2: W transforms like a $\frac{p(p+1)}{2}$-differential

So far W was computed in one chart. How does it change if we use another coordinate w, with $z = z(w)$? In the new chart the same differentials read $\omega_j = h_j(z(w))\, z'(w)\, \mathrm{d}w \coloneqq \tilde{h}_j(w)\, \mathrm{d}w$, and the new Wronskian is built from the $\tilde{h}_j$ and w-derivatives. We need two facts about Wronskians.

(a) *Multiplying all functions by φ:* $W[\varphi f_1, \dots, \varphi f_p] = \varphi^p\, W[f_1, \dots, f_p]$. By the Leibniz rule $(\varphi f)^{(k)} = \varphi f^{(k)} + \sum_{l<k} \binom{k}{l} \varphi^{(k-l)} f^{(l)}$, so the new matrix is (lower-triangular matrix with φ on the diagonal) $\times$ (old matrix). Its determinant is φ^p times the old one.

(b) *Changing the variable:* $W_w[f_1 \circ z, \dots, f_p \circ z] = (z')^{p(p-1)/2}\, W_z[f_1, \dots, f_p] \circ z$. By the chain rule $\frac{\mathrm{d}^k}{\mathrm{d}w^k} f(z(w)) = (z')^k f^{(k)}(z(w)) + \sum_{l<k} (\cdots) f^{(l)}(z(w))$, so again the matrix changes by a lower-triangular factor, now with diagonal $1, z', (z')^2, \dots, (z')^{p-1}$, whose determinant is $(z')^{0+1+\cdots+(p-1)}$.

Therefore

$$\begin{aligned} \tilde{W}(w) &\overset{(1)}{=} W_w\left[z' \cdot (h_1 \circ z), \dots, z' \cdot (h_p \circ z)\right] \\ &\overset{(2)}{=} (z')^p\, W_w\left[h_1 \circ z, \dots, h_p \circ z\right] \\ &\overset{(3)}{=} (z')^p\, (z')^{\frac{p(p-1)}{2}}\, W(z(w)) = \left(\frac{\mathrm{d}z}{\mathrm{d}w}\right)^{\frac{p(p+1)}{2}} W(z(w)). \end{aligned} \tag{5.184}$$

(1) Definition of the new Wronskian, with $\tilde{h}_j = z' \cdot (h_j \circ z)$.

(2) Fact (a) with $\varphi = z'$.
(3) Fact (b).

Compare with the transformation of a q-differential $\phi(z)\,dz^q$ (Nash's footnote 6, Section 5.6.6): $\tilde{\phi}(w) = \phi(z(w))(z')^q$. So (5.184) says exactly that

$$\mathbb{W} := W(z)\,(dz)^{\frac{p(p+1)}{2}} \quad \text{is a global holomorphic section of } K_\Sigma^{\frac{p(p+1)}{2}}. \tag{5.185}$$

It is not identically zero (it is non-zero at any normal point, and normal points exist because the zeros of a non-zero holomorphic function in a chart are isolated).

Step 3: count the zeros

A non-zero holomorphic section of a line bundle L has exactly $\deg L$ zeros, counted with multiplicity (Chapter 5). Then

$$\sum_{z_0\in\Sigma} w(z_0) \overset{(1)}{=} \sum_{z_0\in\Sigma} \operatorname{ord}_{z_0}\mathbb{W} \overset{(2)}{=} \deg K^{\frac{p(p+1)}{2}} \overset{(3)}{=} \frac{p(p+1)}{2}(2p-2) = (p-1)\,p\,(p+1) = p^3 - p. \tag{5.186}$$

(1) (5.183) at every point.
(2) Number of zeros of a holomorphic section = degree.
(3) $\deg K^N = N \deg K$ (transition functions are N-th powers, so c_1 multiplies by N), and $\deg K = 2p-2$.

Key idea: The Weierstrass points are the zeros of a section of $K^{p(p+1)/2}$

On a compact Riemann surface of genus p the Wronskian of a basis of holomorphic differentials is a holomorphic section of $K^{p(p+1)/2}$. It vanishes exactly at the Weierstrass points, to order equal to their weights. Hence the Weierstrass points are finite in number and

$$\sum_{z_0} w(z_0) = p^3 - p.$$

Since $1 \le w(z_0) \le \frac{p(p-1)}{2}$ at a Weierstrass point (Example 5.99), for $p \ge 2$

$$2p+2 = \frac{p^3-p}{p(p-1)/2} \;\le\; \#\{\text{Weierstrass points}\} \;\le\; p^3 - p.$$

For $p = 1$ the count gives 0: no Weierstrass points on a torus, as the gap theorem said. For $p = 0$ it gives 0 again (there are no differentials and nothing to count). For $p = 2$: between 6 and 6 points, so *every* genus-2 surface has exactly six Weierstrass points, each of weight 1. For $p = 3$: between 8 and 24.

Physical picture: Counting zeros by topology, once more

The logic of (5.186) is the logic of this whole chapter: to count the special points of a geometric structure, package the condition as the vanishing of a holomorphic section of a line bundle, and read the count off a Chern number. Poincaré–Hopf counts zeros of a vector field by χ; Gauss–Bonnet counts curvature by χ; here the Weierstrass condition is a determinant, the determinant is a section of a power of K, and the count is $\frac{p(p+1)}{2}\deg K$. Physicists meet the same move when counting zero modes of a Dirac operator by an index, or instantons by $\frac{1}{8\pi^2}\int \operatorname{tr} F \wedge F$.

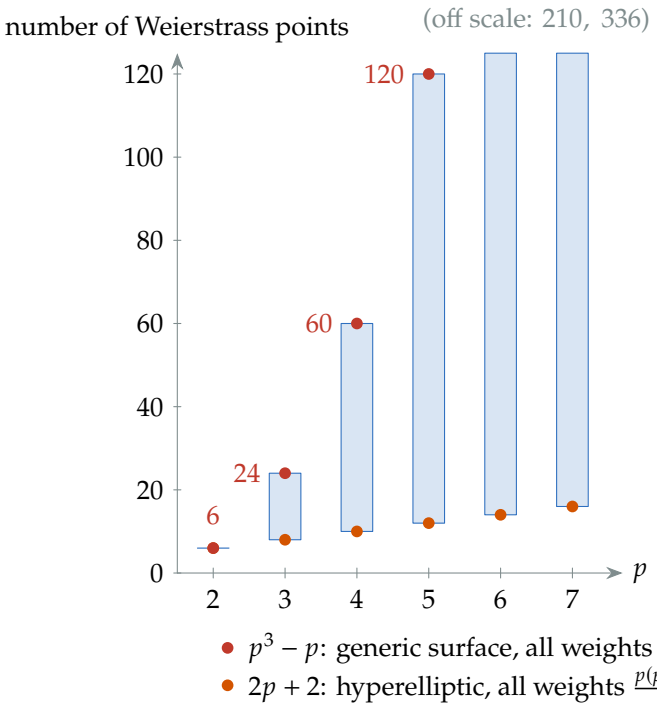


Figure 5.57: The number of Weierstrass points in genus p lies in the shaded range between $2p + 2$ and $p^3 - p$. The upper end is reached when every Weierstrass point has weight 1, which is the case for a generic surface; the lower end exactly for hyperelliptic surfaces, where every Weierstrass point has the maximal weight $p(p-1)/2$. In genus 2 the two ends meet: every genus-2 surface has exactly six Weierstrass points.

Watch out: Two standard facts we quote

Two statements in this subsection are true but not proved here: that $w(z_0) \le p(p-1)/2$ with equality only at the Weierstrass points of a hyperelliptic surface, and that a generic surface has all weights equal to 1. Both are in Griffiths and Harris (Ch. 2, §4). Everything else, in particular the finiteness and the total $p^3 - p$, follows from (5.183)–(5.186), which we derived.

5.7.8 Weierstrass points in examples

The count $p^3 - p$ is a strong check on any explicit calculation. We now find the Weierstrass points of the three classes of surface we know explicitly.

Worked example 5.101: The torus has no Weierstrass points

$p = 1$, one differential $\omega = \mathrm{d}z$, so $h = 1$ and the Wronskian is the 1×1 determinant $W = 1$. It never vanishes, so there are no Weierstrass points, and the count is $\deg K^1 = 0 = 1^3 - 1$. ✓ In terms of functions: at every point z_0 the gap sequence is $\{1\}$, the only obstruction is the single residue condition $a_1 = 0$ of Example 5.100, and $\wp(z - z_0)$ realises the non-gap 2. The torus is homogeneous (translations act transitively), so all its points must look alike: there is no room for special points.

Worked example 5.102: Hyperelliptic surfaces: the branch points

Let Σ be $y^2 = f(x)$ with f a polynomial of degree $2p+2$ with distinct roots $e_1, \dots, e_{2p+2}$ ($p \geq 2$). By Example 5.91, $\mathrm{d}x/y$ has no zeros in the finite part and a zero of order $p-1$ at each of the two points $\infty_\pm$ over $x = \infty$, where x has a simple pole. So

$$\omega_k = x^k \frac{\mathrm{d}x}{y}, \qquad k = 0, 1, \dots, p-1,$$

is holomorphic (order $p-1-k \geq 0$ at $\infty_\pm$), and these are p independent differentials: a basis.

At a branch point $P_i = (e_i, 0)$ use the basis $\omega'_k = (x-e_i)^k\,\mathrm{d}x/y$, $k = 0, \dots, p-1$. It spans the same space, since $(x-e_i)^k$ is a polynomial of degree k in x with leading coefficient 1 (a triangular change of basis). With the local coordinate s, $x - e_i = s^2$, $y = s\,u(s)$, $u(0) \neq 0$ (Example 5.100):

$$\omega'_k \overset{(1)}{=} s^{2k} \cdot \frac{2s\,\mathrm{d}s}{s\,u(s)} \overset{(2)}{=} \frac{2}{u(s)}\, s^{2k}\,\mathrm{d}s.$$

(1) $(x-e_i)^k = s^{2k}$, $\mathrm{d}x = 2s\,\mathrm{d}s$, $y = s\,u(s)$.
(2) Cancel s.

Since $2/u(s)$ is non-zero at $s = 0$, the vanishing orders are $m_k = 2k$: $0, 2, 4, \dots, 2p-2$, all distinct, so this is already the adapted basis. By Noether's characterisation the gaps are

$$1, 3, 5, \dots, 2p-1, \qquad w(P_i) = \sum_{k=0}^{p-1} \big((2k+1) - (k+1)\big) = \sum_{k=0}^{p-1} k = \frac{p(p-1)}{2}.$$

At any other point the pattern is normal. At a finite point $x = a$ with $y \neq 0$ the coordinate is $t = x - a$, $1/y$ is holomorphic and non-zero, and $(x-a)^k \mathrm{d}x/y$ vanishes to order k: orders $0, \dots, p-1$. At $\infty_\pm$ the orders of ω_k are $p-1-k$, again $0, \dots, p-1$. So no other point is a Weierstrass point.

The count.

$$\sum_{z_0} w(z_0) \overset{(1)}{=} (2p+2) \cdot \frac{p(p-1)}{2} \overset{(2)}{=} (p+1)\,p\,(p-1) = p^3 - p. \checkmark$$

(1) $2p+2$ branch points, each of weight $p(p-1)/2$; all other points have weight 0.
(2) Arithmetic.

The functions. The non-gap 2 is realised by $1/(x-e_i) = s^{-2}$: its only pole is the double pole at P_i (at $\infty_\pm$ it vanishes). The odd non-gap $2p+1$ is realised by $y/(x-e_i)^{p+1}$, which behaves like $s\,u/s^{2p+2}$ at P_i and like x^{p+1}/x^{p+1} (finite) at $\infty_\pm$. Products of these give every non-gap, as the semigroup property predicts.

Genus three: plane quartics and their flexes

Most genus-3 surfaces are not hyperelliptic (Section 5.6.12: the hyperelliptic locus has dimension $2p-1 = 5 < 6 = 3p-3$). The generic genus-3 surface is a smooth **plane quartic**, a curve $C = \{F(X,Y,Z) = 0\} \subset \mathbb{CP}^2$ with F homogeneous of degree 4 (Chapter 5: the degree–genus formula gives $\frac{(4-1)(4-2)}{2} = 3$). Its holomorphic differentials have a beautiful description.

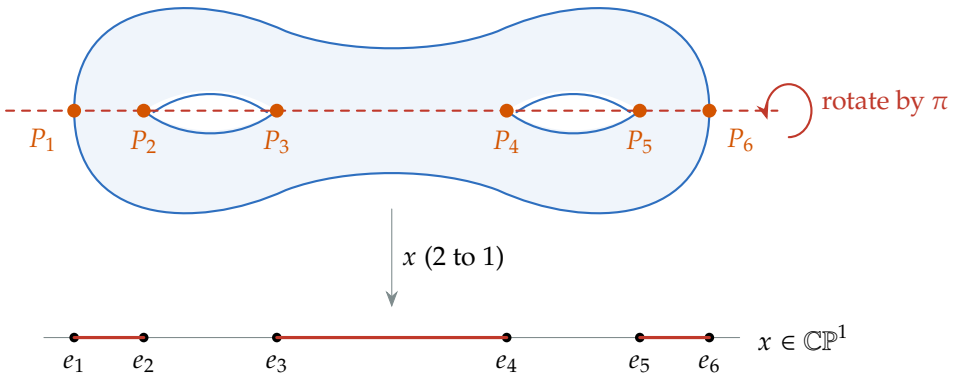


Figure 5.58: A genus-2 surface $y^2 = \prod_{i=1}^{6}(x - e_i)$ drawn so that the hyperelliptic involution $(x, y) \mapsto (x, -y)$ is the rotation by π about the dashed axis. The axis pierces the surface in exactly six points $P_1, \dots, P_6$: the fixed points of the involution, the branch points of the 2-to-1 map x to the sphere (red cuts join them in pairs), and the six Weierstrass points. At each of them $1/(x - e_i)$ is a function with a single double pole, so the gap sequence is $\{1, 3\}$ instead of the normal $\{1, 2\}$. Their weights add up to $6 = 2^3 - 2$.

Key idea: Differentials on a plane quartic are lines

In the affine chart $Z = 1$, with $F(x, y) = F(x, y, 1)$, the holomorphic differentials of a smooth quartic are

$$\omega_P = P(x, y)\,\frac{dx}{\partial F/\partial y}, \qquad P \in \operatorname{span}\{1, x, y\},$$

and the zeros of ω_P are the four intersection points of C with the line $\{P = 0\}$, with multiplicity. So *the order of vanishing of ω_P at z_0 is the order of contact of the line $P = 0$ with C at z_0.*

We take this as known (it is the adjunction formula $K_C = \mathcal{O}(d-3)|_C$ for a plane curve of degree d; see Griffiths and Harris, Ch. 2, §1). Two consistency checks: the space of linear P is $3 = p$ dimensional, and a line meets a quartic in $4 = 2p - 2 = \deg K$ points, by Bézout.

Now read off the vanishing orders at a point $z_0 \in C$. In the 3-dimensional space of lines P:

- a line not through z_0 gives order 0;
- a line through z_0 but not tangent gives order 1;
- the tangent line T_{z_0} gives order 2, or more.

So the vanishing orders are $0, 1, m_3$ with $m_3 =$ the contact order of the tangent line. Three cases:

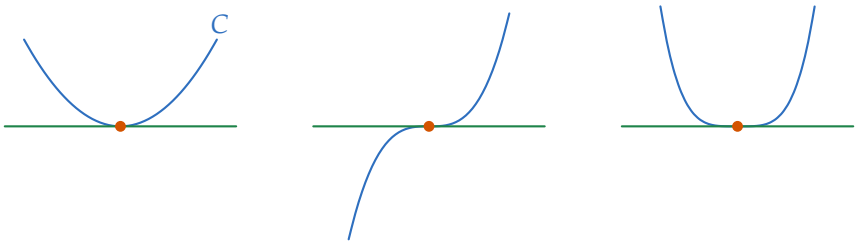
point z_0	tangent line meets C	orders m_j	gaps n_j	weight
ordinary	$2z_0 + Q + R$	0, 1, 2	1, 2, 3	0
flex (inflection point)	$3z_0 + Q$	0, 1, 3	1, 2, 4	1
hyperflex	$4z_0$	0, 1, 4	1, 2, 5	2

The Weierstrass points of a plane quartic are its inflection points. Note that the second vanishing order is always 1, so 2 and 3 are always gaps: no function on a plane quartic has a single pole of order 2, so plane quartics are never hyperelliptic. The classical count of flexes of a plane

curve of degree d is $3d(d-2)$ (intersect C with its Hessian curve, of degree $3(d-2)$, by Bézout), counting a hyperflex twice. For $d = 4$:

$$\#\text{flexes} + 2\,\#\text{hyperflexes} = 3 \cdot 4 \cdot 2 = 24 = 3^3 - 3. \checkmark$$

The Hessian count and the Wronskian count agree, as they must.



ordinary point
contact 2: orders $0, 1, 2$

flex
contact 3: orders $0, 1, 3$

hyperflex
contact 4: orders $0, 1, 4$

Figure 5.59: Real pictures of the three kinds of point on a plane quartic. The green line is the tangent line at the orange point; the differential ω_P attached to it vanishes there to the order of contact. At an ordinary point the curve stays on one side of its tangent (contact 2). At a flex the curve crosses its tangent (contact 3), and at a hyperflex it is flattened to fourth order (contact 4). Flexes and hyperflexes are the Weierstrass points, of weights 1 and 2.

Worked example 5.103: The Fermat quartic: twelve hyperflexes

Take $C: X^4 + Y^4 + Z^4 = 0$. It is smooth: the gradient $(4X^3, 4Y^3, 4Z^3)$ vanishes only at $X = Y = Z = 0$, which is not a point of $\mathbb{CP}^2$. Let ζ be one of the four roots of $\zeta^4 = -1$ and consider the line $X = \zeta Y$. On it,

$$X^4 + Y^4 + Z^4 \overset{(1)}{=} \zeta^4 Y^4 + Y^4 + Z^4 \overset{(2)}{=} Z^4.$$

(1) Substitute $X = \zeta Y$.
(2) $\zeta^4 = -1$.

So the line meets C only where $Z = 0$, i.e. at the single point $[\zeta : 1 : 0]$, and with multiplicity 4 (the restriction of F to the line is Z^4, a fourth-order zero). This point is a hyperflex, and the line is its tangent. The same works with the roles of the coordinates permuted: lines $X = \zeta Z$ and $Y = \zeta Z$. That gives $3 \times 4 = 12$ hyperflexes, each of weight 2:

$$12 \cdot 2 = 24 = 3^3 - 3. \checkmark$$

So the count is exhausted: the Fermat quartic has exactly twelve Weierstrass points, all hyperflexes, and no ordinary flexes. Its automorphisms (permutations of X, Y, Z and multiplication of coordinates by fourth roots of unity, a group of order $6 \cdot 16 = 96$) permute these twelve points, as Section 5.7.11 says they must.

Worked example 5.104: The Klein quartic: twenty-four flexes

The Klein quartic $X^3Y + Y^3Z + Z^3X = 0$ (Example 5.82) has 24 ordinary flexes and no hyperflexes, so it has the generic number $24 = 3^3 - 3$ of Weierstrass points, each of weight 1. Its 168 automorphisms act on these 24 points transitively. The stabiliser of a point then has order $168/24 = 7$. Explicitly, the order-7 symmetry $(X, Y, Z) \mapsto (\epsilon X, \epsilon^4 Y, \epsilon^2 Z)$ with $\epsilon = e^{2\pi i/7}$ maps the curve to itself (each monomial picks up $\epsilon^{3+4} = \epsilon^{12+2} = \epsilon^{6+1} = 1$) and fixes the three points $[1 : 0 : 0], [0 : 1 : 0], [0 : 0 : 1]$. At $[0 : 0 : 1]$, in the chart $Z = 1$, the

curve is $x^3y + y^3 + x = 0$, i.e. $x = -y^3 - x^3y = -y^3 + O(y^{10})$: the tangent line $x = 0$ has contact 3. So this fixed point is a flex. ✓

surface	genus	Weierstrass points	weights
sphere	0	none	–
torus	1	none	–
any genus-2 surface (y^2 = sextic)	2	6 branch points	1 each
hyperelliptic, genus p	p	$2p+2$ branch points	$p(p-1)/2$ each
Klein quartic	3	24 flexes	1 each
Fermat quartic	3	12 hyperflexes	2 each
generic surface of genus p	p	$p^3 - p$	1 each

In every row the weights add up to $p^3 - p$.

Try it: Weierstrass points of the genus-3 hyperelliptic surface

For $y^2 = x^8 - 1$ find the Weierstrass points, their gap sequences, and check the count. *Answer.* $p = 3$. The branch points are the eight points $(\zeta, 0)$, $\zeta^8 = 1$. At each, the vanishing orders of $\mathrm{d}x/y$, $(x-\zeta)\mathrm{d}x/y$, $(x-\zeta)^2\mathrm{d}x/y$ are $0, 2, 4$, so the gaps are $1, 3, 5$ and the weight is $0 + 1 + 2 = 3$. Total $8 \cdot 3 = 24 = 3^3 - 3$. Every other point is normal. So this surface has 8 Weierstrass points, the minimum $2p + 2$ allowed, compared with 24 for the Klein quartic.

5.7.9 q-differentials: the same theorem for every K^q

Nash ends the section with a remark [Nash p. 169]: instead of meromorphic functions one may consider meromorphic q-differentials, and "a very similar proof" gives a gap theorem for them, with the sheaf $\mathcal{O}([nz_0])$ replaced by $\mathcal{O}(K^q \otimes [nz_0])$. Let us carry this out, because one case of it, $q = -1$, is the $3p - 3$ of Section 5.6 seen from a completely different angle.

The count

Recall (Nash's footnote 6, Section 5.6.6) that a **q-differential** is a section of K^q, locally $\phi(z)\,\mathrm{d}z^q$, with $\tilde{\phi}(w) = \phi(z(w))\,(z')^q$. For $q = 0, 1, 2$ these are functions, 1-forms and quadratic differentials; for $q = -1$ they are vector fields $v(z)\,\partial_z$, since $K^{-1} = K^* = T\Sigma$. Let

$$\mathcal{L}_q(nz_0) := \Big\{\text{meromorphic } q\text{-differentials, holomorphic on } \Sigma \setminus \{z_0\},\ \text{with a pole of order} \le n \text{ at } z_0\Big\} \;\cong\; H^0\big(\Sigma, \mathcal{O}(K^q \otimes [nz_0])\big),$$

by the same argument as (5.166): multiply by the canonical section s of $[nz_0]$. Write $\ell_q(n)$ for its dimension. Then

$$\begin{aligned}\ell_q(n) &\overset{(1)}{=} \operatorname{index} \bar{\partial}_{K^q \otimes [nz_0]} + \dim H^1\big(\Sigma, \mathcal{O}(K^q \otimes [nz_0])\big)\\ &\overset{(2)}{=} \deg\big(K^q \otimes [nz_0]\big) + 1 - p + \dim H^0\big(\Sigma, \mathcal{O}(K^{1-q} \otimes [-nz_0])\big)\\ &\overset{(3)}{=} (2q-1)(p-1) + n + \kappa_{1-q}(n). \end{aligned} \tag{5.187}$$

(1) The index is $\dim\ker - \dim\operatorname{coker}$, and the cokernel of $\bar\partial_E$ is $H^1(\Sigma, \mathcal{O}(E))$ (Hodge theory, as in Section 5.6.7).
(2) Riemann–Roch (5.172) for a line bundle of any degree; Serre duality (5.135) with $K \otimes (K^q \otimes [nz_0])^* = K^{1-q} \otimes [-nz_0]$.
(3) $\deg = q(2p-2)+n$, and $q(2p-2)+1-p = (2q-1)(p-1)$. We write $\kappa_{1-q}(n)$ for the dimension of the space of holomorphic $(1-q)$-differentials vanishing to order $\geq n$ at z_0, by the same reasoning as (5.174).

For $q = 0$ this is (5.176), with $\kappa_1 = \kappa$. The staircase argument of Section 5.7.5 goes through word for word (read off the leading coefficient a_n of the pole in a local frame $\mathrm{d}z^q$): each step of ℓ_q has height 0 or 1, and n is a **q-gap** iff $\ell_q(n) = \ell_q(n-1)$. The residue argument of Section 5.7.6 also goes through, because a q-differential times a $(1-q)$-differential is a 1-form, $\phi(z)\mathrm{d}z^q \cdot \psi(z)\mathrm{d}z^{1-q} = \phi\psi\,\mathrm{d}z$, to which the residue theorem applies. The result:

Key idea: Gap theorem for q-differentials

A pole order $n \geq 1$ at z_0 is a gap for meromorphic q-differentials iff some holomorphic $(1-q)$-differential vanishes to order exactly $n-1$ at z_0. So the number of q-gaps is $\dim H^0(\Sigma, \mathcal{O}(K^{1-q}))$, and the gaps are the vanishing orders of the holomorphic $(1-q)$-differentials, plus one.

For $p \geq 2$ this gives the following table (for $p \geq 2$, $\dim H^0(K^r) = (2r-1)(p-1)$ when $r \geq 2$, p when $r = 1$, 1 when $r = 0$, and 0 when $r < 0$, all from (5.187) and vanishing):

q	meromorphic objects with one pole	dual holomorphic objects	number of gaps
$q \geq 2$	q-differentials	$(1-q)$-differentials: none	0
1	1-forms (currents)	functions: constants	1 (namely $n = 1$)
0	functions (Nash's case)	1-forms	p
-1	vector fields	quadratic differentials	$3p-3$
$q \leq -1$	q-differentials	$(1-q)$-differentials	$(1-2q)(p-1)$

Watch out: Where the content of Nash's remark lies

For $q \geq 2$ the "gap theorem" is empty: $K^{1-q} \otimes [-nz_0]$ has negative degree, so $\kappa_{1-q} = 0$ and every pole order occurs. The interesting q-differential gap problems are those with $q \leq 0$, whose gaps are controlled by the holomorphic $(1-q)$-differentials with $1-q \geq 1$. Equivalently, one speaks of the **q-Weierstrass points** of the space $H^0(K^q)$, $q \geq 1$: the points where some holomorphic q-differential vanishes to order at least $\dim H^0(K^q)$. For $q = 1$ these are the ordinary Weierstrass points. Nash's sentence covers both readings.

The case $q = 1$ in the table is the physicist's charge conservation: a meromorphic 1-form (a current $j(z)\mathrm{d}z$) with a single pole must have a pole of order at least 2, because a simple pole carries a residue (a charge) and the total charge on a closed surface vanishes. This is the intuition box of Section 5.7.6, now as a theorem with exactly one gap, dual to the one constant function.

How many q-Weierstrass points?

The Wronskian count generalises directly. Let $d = \dim H^0(K^q)$ and let $\phi_j = h_j(z)\,\mathrm{d}z^q$, $j = 1, \dots, d$, be a basis. The Wronskian $W[h_1, \dots, h_d]$ vanishes at z_0 to order $\sum_j \big(m_j - (j-1)\big)$, the

q-weight of z_0 (the proof of (5.183) did not use $q = 1$). Its transformation law changes in one place only: fact (a) is now used with $\varphi = (z')^q$, giving $(z')^{qd}$. So W is a section of K^N with

$$N = qd + \frac{d(d-1)}{2} = \frac{d\,(2q+d-1)}{2}.$$

For $q \geq 2$ and $p \geq 2$, with $d = (2q-1)(p-1)$:

$$\sum_{z_0} w_q(z_0) \overset{(1)}{=} N\,(2p-2) \overset{(2)}{=} d\,(2q+d-1)\,(p-1)$$

$$\overset{(3)}{=} (2q-1)(p-1)\cdot(2q-1)p\cdot(p-1) = p\,(p-1)^2\,(2q-1)^2. \tag{5.188}$$

(1) Number of zeros of the section $W\,\mathrm{d}z^N$ of K^N.
(2) Insert $N = d(2q+d-1)/2$.
(3) $d = (2q-1)(p-1)$, so $2q+d-1 = (2q-1)+(2q-1)(p-1) = (2q-1)p$.

For $q = 1$ the same line with $d = p$ gives $p(p+1)(p-1)$, which is (5.186). For quadratic differentials ($q = 2$) the total is $9p(p-1)^2$; in genus 2 it is 18.

Worked example 5.105: Quadratic Weierstrass points in genus two

On $y^2 = \prod_{i=1}^{6}(x-e_i)$ the three quadratic differentials are $\mathrm{d}x^2/y^2$, $x\,\mathrm{d}x^2/y^2$, $x^2\,\mathrm{d}x^2/y^2$ (Example 5.94). At the branch point P_1 use instead the basis $(x-e_1)^k\,\mathrm{d}x^2/y^2$, $k = 0,1,2$ (a triangular change). With $x - e_1 = s^2$, $y = s\,u(s)$:

$$(x-e_1)^k\frac{\mathrm{d}x^2}{y^2} \overset{(1)}{=} s^{2k}\,\frac{(2s\,\mathrm{d}s)^2}{s^2u(s)^2} \overset{(2)}{=} \frac{4}{u(s)^2}\,s^{2k}\,\mathrm{d}s^2.$$

(1) Substitute; $\mathrm{d}x^2$ means $\mathrm{d}x\otimes\mathrm{d}x = (2s\,\mathrm{d}s)^2$.
(2) Cancel s^2.

The vanishing orders are $0,2,4$ instead of the normal $0,1,2$, so the $q = 2$ weight is $(0-0)+(2-1)+(4-2) = 3$. The six branch points contribute $6\cdot 3 = 18 = 9\cdot 2\cdot 1^2$: the whole of (5.188). So in genus two the quadratic Weierstrass points are again exactly the six branch points. ✓

5.7.10 Vector fields with one pole: the $3p-3$ moduli as gaps

Take $q = -1$. A meromorphic vector field on Σ with a single pole of order at most n at z_0 lies in $\mathcal{V}(n) := \mathcal{L}_{-1}(nz_0)$, and (5.187) reads

$$\dim\mathcal{V}(n) = n+3-3p+\kappa_2(n), \qquad \kappa_2(n) = \dim\{\phi\in H^0(K^2) :\ \phi \text{ vanishes to order} \geq n \text{ at } z_0\}. \tag{5.189}$$

For $p \geq 2$:

- at $n = 0$: $\dim\mathcal{V}(0) = 3-3p+(3p-3) = 0$. No holomorphic vector fields (Nash's (5.90)). ✓
- for $n \geq 4p-3$: a non-zero quadratic differential has $\deg K^2 = 4p-4$ zeros, so none vanishes to order $\geq 4p-3$, $\kappa_2(n) = 0$, and $\dim\mathcal{V}(n) = n+3-3p$.

The staircase climbs from 0 at $n = 0$ to $(4p-3)+3-3p = p$ at $n = 4p-3$ in $4p-3$ steps of height 0 or 1. So it has p rising steps and

$$(4p-3)-p = 3p-3 \text{ flat steps.}$$

There are exactly $3p-3$ pole orders that no global vector field with a single pole at z_0 can have. By the keyidea above they are the vanishing orders of the $3p-3$ quadratic differentials plus one; at a point where these orders are $0, 1, \dots, 3p-4$ (a "normal point for K^2") the gaps are $1, 2, \dots, 3p-3$.

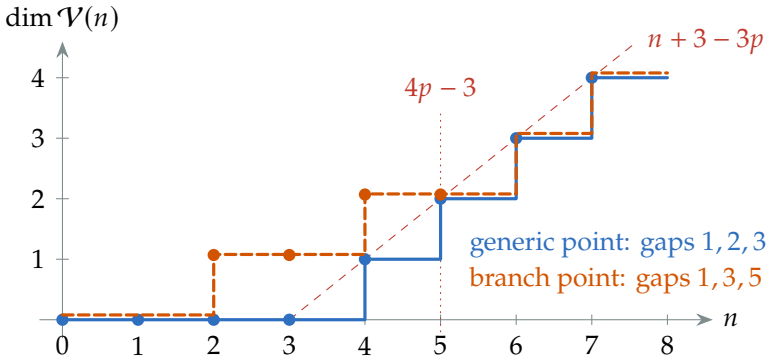


Figure 5.60: Meromorphic vector fields with a single pole on a genus-2 surface. $\dim \mathcal{V}(n)$ starts at 0 (no holomorphic vector fields), and must reach $p=2$ at $n=4p-3=5$, so it has exactly $3=3p-3$ flat steps: three forbidden pole orders, one for each complex modulus. At a generic point the forbidden orders are $1,2,3$; at a branch point of $y^2=$ sextic they are $1,3,5$, and the field $y\,\partial_x/(x-e_1)$ realises the order 2.

Worked example 5.106: Vector fields with one pole in genus two

On $y^2 = f(x) = \prod_{i=1}^{6}(x-e_i)$ consider

$$v_k = \frac{y}{(x-e_1)^k}\,\partial_x, \qquad k \geq 1.$$

Where is it singular? We need ∂_x in good local coordinates.

- *At an ordinary finite point* x is a coordinate, y and $(x-e_1)^{-k}$ are holomorphic (if $x \neq e_1$): no pole.
- *At a branch point* P_i, $x = e_i + s^2$, so $\partial_s = 2s\,\partial_x$ and

$$v_k \overset{(1)}{=} \frac{s\,u(s)}{(e_i-e_1+s^2)^k}\cdot\frac{1}{2s}\,\partial_s \overset{(2)}{=} \frac{u(s)}{2\,(e_i-e_1+s^2)^k}\,\partial_s.$$

  (1) $y = s\,u(s)$ and $\partial_x = (2s)^{-1}\partial_s$.
  (2) Cancel s.

  For $i \neq 1$ this is holomorphic. For $i=1$ it is $\frac{u(s)}{2}s^{-2k}\partial_s$: a pole of order exactly $2k$.
- *At* $\infty_\pm$, $t = 1/x$, $\partial_x = -t^2\partial_t$ and $y \approx \pm x^3 = \pm t^{-3}$, so $v_k \approx \mp t^{-3}\,t^k\,t^2\,\partial_t = \mp t^{k-1}\partial_t$: holomorphic for $k \geq 1$.

So $v_k \in \mathcal{V}(2k)$ has a single pole, of order $2k$, at the branch point P_1. The non-gaps $2,4,6$ are realised, and v_1, v_2, v_3 are independent, in agreement with $\dim \mathcal{V}(6) = 6+3-6+0 = 3$. The odd orders $1,3,5$ are the three gaps, one for each of the three moduli of a genus-2 surface.

The gap modes are the tangent vectors to $\mathcal{M}_p$

Why "one gap for each modulus"? Because the gap modes give an explicit basis of $T_{[\Sigma]}\mathcal{M}_p = H^{0,1}_{\bar\partial_{K^*}}$. The construction is called a **Schiffer variation**. Let D be a coordinate disc around z_0, and χ a smooth cut-off function equal to 1 on $|z| < r_1$ and to 0 on $|z| > r_2$, with $r_1 < r_2$ inside D. For $n \geq 1$ put

$$\mu_n := \bar\partial\big(\chi\, z^{-n}\partial_z\big) = \big(\partial_{\bar z}\chi\big)\, z^{-n}\ \partial_z \otimes \mathrm{d}\bar z. \tag{5.190}$$

This is a smooth Beltrami differential, supported in the annulus $A = \{r_1 < |z| < r_2\}$ (the singular vector field $z^{-n}\partial_z$ is only used where χ has already killed the pole). Its pairing (5.138) with a quadratic differential $\phi = \sum_{m\geq 0}\phi_m z^m\,\mathrm{d}z^2$ is a residue:

$$\begin{aligned}\langle \mu_n, \phi\rangle &\overset{(1)}{=} \int_A \partial_{\bar z}\big(\chi\, z^{-n}\phi(z)\big)\,\mathrm{d}^2z \\ &\overset{(2)}{=} -\frac{i}{2}\oint_{\partial A}\chi\, z^{-n}\phi(z)\,\mathrm{d}z \\ &\overset{(3)}{=} \frac{i}{2}\oint_{|z|=r_1} z^{-n}\phi(z)\,\mathrm{d}z \overset{(4)}{=} \frac{i}{2}\cdot 2\pi i\,\phi_{n-1} = -\pi\,\phi_{n-1}.\end{aligned} \tag{5.191}$$

(1) $z^{-n}\phi$ is holomorphic on A, so $\partial_{\bar z}(\chi z^{-n}\phi) = (\partial_{\bar z}\chi)z^{-n}\phi$; outside A the integrand is zero.
(2) With $\mathrm{d}^2z = \frac{i}{2}\mathrm{d}z\wedge\mathrm{d}\bar z$ and $\mathrm{d}(F\,\mathrm{d}z) = \partial_{\bar z}F\,\mathrm{d}\bar z\wedge\mathrm{d}z$ we have $\partial_{\bar z}F\,\mathrm{d}^2z = -\frac{i}{2}\mathrm{d}(F\,\mathrm{d}z)$; then Stokes.
(3) ∂A is the outer circle (counter-clockwise), where $\chi = 0$, minus the inner circle, where $\chi = 1$.
(4) Residue of $z^{-n}\sum\phi_m z^m$ at 0 is ϕ_{n-1}.

Now take the $3p-3$ gap orders $n_i = m_i + 1$ and the adapted basis $\phi_j = c_j z^{m_j}(1 + O(z))\mathrm{d}z^2$ with $m_1 < m_2 < \cdots$. The matrix of pairings $\langle \mu_{n_i}, \phi_j\rangle = -\pi\,(\phi_j)_{m_i}$ vanishes for $i < j$ (since ϕ_j starts at z^{m_j} and $m_i < m_j$) and equals $-\pi c_j \neq 0$ on the diagonal. A triangular matrix with non-zero diagonal is invertible. Since the pairing is perfect (Section 5.6.6),

Key idea: Moduli from gaps

The $3p-3$ Schiffer variations $\mu_{n_1},\dots,\mu_{n_{3p-3}}$ at the vector-field gaps n_i of any point z_0 form a basis of $T_{[\Sigma]}\mathcal{M}_p$. At a generic point they are the modes $z^{-1}\partial_z,\dots,z^{-(3p-3)}\partial_z$. The non-gap modes give nothing new: if n is a non-gap, a global field $v \in \mathcal{V}(n)$ with leading term $z^{-n}\partial_z$ exists, and μ_n differs from $\bar\partial(\chi(z^{-n}\partial_z - v)) + \bar\partial v$ only by lower modes and the fake $\bar\partial v$... more precisely $\bar\partial v = 0$ away from z_0, so μ_n is cohomologous to a combination of the μ_k with $k < n$.

Physical picture: Cutting, twisting and regluing a disc

A Schiffer variation is a concrete way to change the shape of a surface: cut out a small disc around z_0 and glue it back after moving the rim by the flow of the vector field $\varepsilon z^{-n}\partial_z$, defined on the annulus. The change is trivial if the displacement can be absorbed by moving the disc (a field holomorphic inside) or by moving the rest of the surface (a field holomorphic outside). What cannot be absorbed is exactly one mode per gap. In conformal field theory the vector field $-z^{m+1}\partial_z$ is represented by the Virasoro generator L_m, so the modes $z^{-n}\partial_z$, $n \geq 1$, are the L_{-n-1} ($L_{-2}, L_{-3},\dots$), and a change of moduli is implemented by inserting the corresponding contour integral of the stress tensor $T(z)$ around z_0. This is the germ of the "operator formalism" of string perturbation theory, in which the $3p-3$ moduli of a genus-p worldsheet are generated by $3p-3$ Virasoro modes at a single point.

Worked example 5.107: The torus and the sphere

Torus. K is trivial, so a vector field is $f(z)\partial_z$ with f elliptic, and $\mathcal{V}(n) \cong \mathcal{L}(nz_0)$. Its dimensions are $1, 1, 2, 3, \dots$ for $n = 0, 1, 2, 3, \dots$ (the constant field ∂_z at $n = 0$, the conformal Killing vector of Example 5.71). The only flat step is at $n = 1$: one gap, one modulus. Formula (5.189) agrees: $\kappa_2(0) = 1$ ($\mathrm{d}z^2$) and $\kappa_2(n \geq 1) = 0$, so $\dim \mathcal{V}(n) = n + \kappa_2(n)$. The Schiffer variation μ_1 pairs with $\mathrm{d}z^2$ to $-\pi \neq 0$ by (5.191): it moves τ.
Sphere. There are no quadratic differentials, $\kappa_2 = 0$, and $\dim \mathcal{V}(n) = n + 3$: every step rises, there are no gaps and no moduli. At $n = 0$ we recover the three conformal Killing vectors ∂_z, $z\partial_z$, $z^2\partial_z$ of Example 5.70 (Nash's footnote 7), the Lie algebra of $\mathrm{SL}(2, \mathbb{C})$.

5.7.11 Automorphisms: finiteness and the Hurwitz bound

Nash's final paragraph [Nash p. 169] uses the Weierstrass points to show that a surface of genus $p \geq 2$ has only finitely many automorphisms, at most $84(p-1)$ of them, the bound quoted in Section 5.5. We fill in the argument.

Working definition 5.52 (Automorphism of a Riemann surface)

An **automorphism** of Σ is a biholomorphic map $\varphi : \Sigma \to \Sigma$: in local coordinates $w = \varphi(z)$ is holomorphic with $\varphi' \neq 0$. They form a group $\mathrm{Aut}(\Sigma)$. *Test:* write the surface as an equation (a lattice, a curve in $\mathbb{CP}^2$, $y^2 = f(x)$) and look for coordinate changes that preserve it [e.g. on the torus $\mathbb{C}/\Lambda$, $z \mapsto -z$ sends the lattice Λ to itself, so it is an automorphism].

Automorphisms permute the Weierstrass points

Let $\varphi \in \mathrm{Aut}(\Sigma)$ and ω a holomorphic differential. The pull-back $\varphi^*\omega$ is again a holomorphic differential (in coordinates $\varphi^*(h(w)\mathrm{d}w) = h(\varphi(z))\varphi'(z)\,\mathrm{d}z$), and φ^* is an invertible linear map of $H^0(K)$ to itself. Since φ is a local biholomorphism, orders of vanishing are preserved:

$$\mathrm{ord}_{z_0}(\varphi^*\omega) \overset{(1)}{=} \mathrm{ord}_{z_0}\big(h(\varphi(z))\,\varphi'(z)\big) \overset{(2)}{=} \mathrm{ord}_{z_0}\big(h \circ \varphi\big) \overset{(3)}{=} \mathrm{ord}_{\varphi(z_0)}\, h.$$

(1) The pull-back in coordinates.
(2) $\varphi'(z_0) \neq 0$.
(3) φ is a local coordinate change at z_0, and the order of a zero does not depend on the coordinate.

So the set of vanishing orders $\{m_j\}$ of all holomorphic differentials at $\varphi(z_0)$ equals that at z_0. By Noether's characterisation the gap sequences agree too:

$$\text{gaps at } \varphi(z_0) = \text{gaps at } z_0, \qquad w(\varphi(z_0)) = w(z_0). \tag{5.192}$$

In particular φ maps the finite set $\mathcal{W}$ of Weierstrass points to itself: it permutes them, preserving weights. This is what Nash means by "Weierstrass points are landmarks".

A non-trivial automorphism has few fixed points

Key idea: Fixed-point bound

If $p \geq 1$... more precisely for any genus p, an automorphism $\varphi \neq \mathrm{id}$ of Σ has at most $2p+2$ fixed points.

Pick z_0 with $\varphi(z_0) \neq z_0$ (fixed points of $\varphi \neq \mathrm{id}$ are isolated, so such z_0 exist), and choose a non-constant $f \in \mathcal{L}((p+1)z_0)$. It exists because $\ell(p+1) \geq (p+1)+1-p = 2$, the index lower bound. Put $g = f - f \circ \varphi$.

- $f \circ \varphi$ has its only pole at $\varphi^{-1}(z_0) \neq z_0$, while f has its only pole at z_0. So $g \not\equiv 0$, and g has poles at the two points $z_0, \varphi^{-1}(z_0)$, each of order at most $p+1$: at most $2p+2$ poles in total.

- At a fixed point a of φ (necessarily $a \neq z_0, \varphi^{-1}(z_0)$), $g(a) = f(a) - f(\varphi(a)) = 0$.

Then

$$\#\{\text{fixed points}\} \;\leq\; \#\{\text{zeros of } g\} \overset{(1)}{=} \#\{\text{poles of } g\} \;\leq\; 2p+2,$$

with (1) the degree-zero property of $\operatorname{div} g$. The bound is sharp: the hyperelliptic involution $(x, y) \mapsto (x, -y)$ fixes exactly the $2p+2$ branch points (Figure 5.58).

Finiteness

Argument 1 (Nash's). By (5.192) we have a homomorphism $\mathrm{Aut}(\Sigma) \to \mathrm{Perm}(\mathcal{W})$ into the finite group of permutations of the Weierstrass points. If Σ is not hyperelliptic, then $\#\mathcal{W} > 2p+2$ (the lower bound $2p+2$ is attained only in the hyperelliptic case), and an element of the kernel fixes more than $2p+2$ points, so it is the identity: $\mathrm{Aut}(\Sigma)$ embeds in a finite group. (In the hyperelliptic case the kernel turns out to be $\{\mathrm{id}, J\}$, J the hyperelliptic involution; we do not prove this.)

Argument 2 (with the tools of Sections 5.4 and 5.6). An automorphism maps the unique constant-curvature metric g_c of area $4\pi(p-1)$ in the conformal class to a constant-curvature metric of the same area in the same class, hence to g_c itself: $\mathrm{Aut}(\Sigma)$ is a group of isometries of g_c. The isometry group of a compact Riemannian manifold is a compact Lie group. Its Lie algebra consists of Killing vector fields; those that preserve the complex structure are holomorphic vector fields; and there are none for $p \geq 2$ (Nash's (5.90)). A compact Lie group with zero Lie algebra is discrete and compact, hence finite.

The Hurwitz bound $84(p-1)$

Let $G = \mathrm{Aut}(\Sigma)$, of order N. The quotient $\Sigma' = \Sigma/G$ is a compact Riemann surface of some genus h (near a point with stabiliser of order m the map to the quotient looks like $z \mapsto z^m$, and z^m is a good coordinate downstairs). The projection $\pi : \Sigma \to \Sigma'$ is an N-sheeted branched covering. Let $b_1, \ldots, b_r$ be the branch values, and $m_i \geq 2$ the order of the stabiliser of any point over b_i; over b_i there are N/m_i points instead of N.

Triangulate Σ' with the b_i among the vertices, and lift the triangulation. Faces and edges

have N preimages each; an ordinary vertex has N preimages; b_i has N/m_i. Then

$$\begin{aligned} 2-2p &\overset{(1)}{=} N\,(V-E+F) - \sum_{i=1}^{r}\Big(N - \frac{N}{m_i}\Big) \\ &\overset{(2)}{=} N\,(2-2h) - N\sum_{i=1}^{r}\Big(1-\frac{1}{m_i}\Big), \\ \Longrightarrow \quad 2p-2 &\overset{(3)}{=} N\,R, \qquad R \coloneqq 2h-2+\sum_{i=1}^{r}\Big(1-\frac{1}{m_i}\Big). \end{aligned} \tag{5.193}$$

(1) Euler characteristic upstairs, counted from the lifted triangulation: N copies of everything, minus the missing preimages over each branch value.
(2) $V-E+F=\chi(\Sigma')=2-2h$.
(3) Rearrange, with a sign change.

(This is the Riemann–Hurwitz formula (5.162) for a Galois covering.) For $p \geq 2$ the left side is positive, so $R>0$ and $N=(2p-2)/R$. To bound N we find the *smallest positive* value R can take. Each term $1-\frac{1}{m_i}$ is at least $\frac{1}{2}$.

- $h \geq 2$: $R \geq 2$.
- $h=1$: $R=\sum(1-\frac{1}{m_i})$, positive only if $r \geq 1$, then $R \geq \frac{1}{2}$.
- $h=0$, $r \geq 5$: $R \geq -2+\frac{5}{2}=\frac{1}{2}$.
- $h=0$, $r=4$: $R>0$ excludes $(2,2,2,2)$; the smallest is $(2,2,2,3)$, $R=-2+\frac{3}{2}+\frac{2}{3}=\frac{1}{6}$.
- $h=0$, $r=3$: $R=1-\big(\frac{1}{m_1}+\frac{1}{m_2}+\frac{1}{m_3}\big)$. We want $\frac{1}{m_1}+\frac{1}{m_2}+\frac{1}{m_3}$ as close to 1 as possible but below it. With $m_1 \leq m_2 \leq m_3$: $m_1=2$ (else the sum is at most 1 only with $R \geq 1-\frac{3}{4}\cdot\ldots$, far away); $m_2=3$ (the choice $m_2=4$ gives at best $(2,4,5)$, $R=\frac{1}{20}$); and then the smallest m_3 with $\frac{1}{2}+\frac{1}{3}+\frac{1}{m_3}<1$ is $m_3=7$:

$$R_{\min}=1-\frac{1}{2}-\frac{1}{3}-\frac{1}{7}=\frac{42-21-14-6}{42}=\frac{1}{42}.$$

  For comparison $(2,3,8)$ gives $\frac{1}{24}$, $(3,3,4)$ gives $\frac{1}{12}$.

So $R \geq \frac{1}{42}$ and

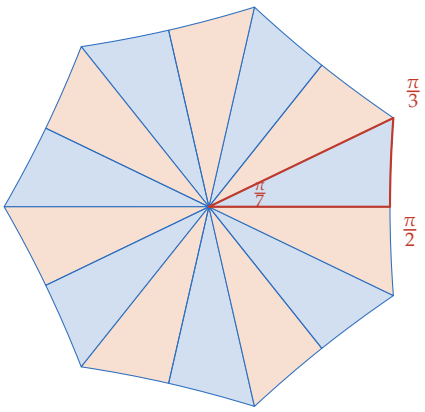
$$\#\operatorname{Aut}(\Sigma)=N=\frac{2p-2}{R} \;\leq\; 42\,(2p-2)=84\,(p-1), \tag{5.194}$$

with equality exactly when $\Sigma \to \Sigma/G$ is a covering of the sphere branched over three points with orders $(2,3,7)$.

The area version in the caption is the same calculation in the language of uniformisation (Section 5.4.6): the Gauss–Bonnet theorem for the hyperbolic metric turns (5.193) into

$$\operatorname{Area}(\Sigma)=|G|\cdot\operatorname{Area}(\Sigma/G), \qquad 4\pi(p-1)=|G|\cdot 2\pi R,$$

and the area of a hyperbolic triangle with angles α,β,γ is $\pi-\alpha-\beta-\gamma$ (Gauss–Bonnet for a geodesic triangle with $K=-1$).



centre of the Poincaré disc $\mathbb{H}$ (magnified)

hyperbolic triangle,
angles $\frac{\pi}{2}, \frac{\pi}{3}, \frac{\pi}{7}$:
area $\pi - \frac{\pi}{2} - \frac{\pi}{3} - \frac{\pi}{7} = \frac{\pi}{42}$

Σ/G has area $\geq 2 \cdot \frac{\pi}{42}$,
Σ has area $4\pi(p-1)$
$\Rightarrow\ |G| \leq 84(p-1)$

Figure 5.61: The Hurwitz bound as an area count. With the constant-curvature metric of curvature -1, a genus-p surface has area $4\pi(p-1)$ (Gauss–Bonnet). Its automorphisms are isometries, and the quotient Σ/G is a hyperbolic orbifold of area $4\pi(p-1)/|G|$, which is $2\pi R$ in the notation of (5.193). The smallest possible orbifold is made of two copies of the $(2,3,7)$ triangle drawn in red (one blue, one orange copy in each kite), of total area $\pi/21$. Fourteen such triangles meet at the vertex with angle $\pi/7$. So $|G| \leq 4\pi(p-1)/(\pi/21) = 84(p-1)$.

Worked example 5.108: The bound in genus 3, and in genus 2

Genus 3. The bound is $84 \cdot 2 = 168$, attained by the Klein quartic: $168/2\ldots$ indeed $NR = 2p-2$ gives $R = 4/168 = 1/42$, so the quotient is the $(2,3,7)$ orbifold. The three branch values have $168/2 = 84$, $168/3 = 56$ and $168/7 = 24$ preimages. The last 24 are the points with stabiliser of order 7: the 24 flexes, i.e. the Weierstrass points of Example 5.104. The Fermat quartic has 96 automorphisms: $R = 4/96 = 1/24$, the $(2,3,8)$ orbifold, and the $96/8 = 12$ points with stabiliser of order 8 are its twelve hyperflexes.
Genus 2. The bound 84 is not attained. The largest group is of order 48, for $y^2 = x(x^4-1)$: $R = 2/48 = 1/24$, again $(2,3,8)$. In every genus-2 case the hyperelliptic involution is in the group and fixes the six Weierstrass points.
Genus 1 and 0: no bound. For $p = 1$, $2p-2 = 0$ and (5.193) gives no constraint: the translations form a continuous group T^2 of automorphisms. For $p = 0$, $\mathrm{Aut}(\mathbb{CP}^1) = \mathrm{SL}(2,\mathbb{C})/\{\pm I\}$. Both reflect the conformal Killing vectors of Section 5.4.5: automorphisms are finite in number exactly when there are no holomorphic vector fields.

5.7.12 Why physicists meet the Wronskian: zero-mode determinants

Nash mentions [Nash p. 318] that the gap theorem constrains correlation functions of conformal field theories on Σ. Here is the simplest instance. Consider a chiral system whose zero modes are the p holomorphic differentials: for instance a b–c system with b of weight 1 and c of weight 0 (a "$\lambda = 1$" ghost system), or a chiral boson current. The path integral over the zero modes of b vanishes unless p fields $b(z_1), \ldots, b(z_p)$ are inserted, and the result is proportional to

$$\big\langle b(z_1)\cdots b(z_p)\big\rangle \ \propto\ \det\big[\, h_i(z_j)\,\big]_{i,j=1}^{p}, \tag{5.195}$$

the determinant of the zero-mode wavefunctions at the insertion points. (This is the Grassmann integral $\int \prod \mathrm{d}\beta_i \ \prod_j (\sum_i \beta_i h_i(z_j))$ over the zero-mode coefficients β_i in $b = \sum_i \beta_i \omega_i + \ldots$.) What happens when all insertions come together at one point z_0? Put $z_j = z_0 + \varepsilon t_j$ with distinct t_j

and expand:

$$\begin{aligned}\det\big[h_i(z_0+\varepsilon t_j)\big] &\overset{(1)}{=} \det\Big[\sum_{k\ge 0}\frac{h_i^{(k)}(z_0)}{k!}\,\varepsilon^k t_j^{\,k}\Big]\\ &\overset{(2)}{=} \det\Big[\frac{h_i^{(k)}(z_0)}{k!}\Big]_{i,\,k<p}\ \det\big[\varepsilon^k t_j^{\,k}\big]_{k<p,\,j}+O\big(\varepsilon^{\frac{p(p-1)}{2}+1}\big)\\ &\overset{(3)}{=} \frac{W(z_0)}{\prod_{k<p}k!}\prod_{i<j}(t_j-t_i)\;\varepsilon^{\frac{p(p-1)}{2}}+O\big(\varepsilon^{\frac{p(p-1)}{2}+1}\big).\end{aligned}\tag{5.196}$$

(1) Taylor expansion of each wavefunction about z_0.

(2) The matrix is a product $B\,C$ with $B_{ik}=h_i^{(k)}/k!$ and $C_{kj}=\varepsilon^k t_j^{\,k}$, $k=0,1,2,\dots$. By the Cauchy–Binet formula its determinant is a sum over p-element sets of rows k of products of minors; the set $\{0,\dots,p-1\}$ has the lowest power of ε, namely $\varepsilon^{0+1+\cdots+(p-1)}$.

(3) The first determinant is the Wronskian with row k divided by $k!$ (and transposed); the second is $\varepsilon^{p(p-1)/2}$ times a Vandermonde determinant.

For $p=2$ this is simply $h_1(z)h_2(z+\varepsilon)-h_2(z)h_1(z+\varepsilon)=\varepsilon\,(h_1h_2'-h_2h_1')+O(\varepsilon^2)=\varepsilon\,W(z)+O(\varepsilon^2)$.

Physical picture: Weierstrass points as places where zero modes pile up

(5.196) says: the leading short-distance behaviour of the zero-mode correlator, as all insertions collide, is the Wronskian. At a normal point it is the generic $\varepsilon^{p(p-1)/2}$ (the Pauli-like repulsion of p fermionic insertions); at a Weierstrass point of weight w it is suppressed by an extra ε^{w}, because a combination of zero modes vanishes to unusually high order there. For the string, the relevant ghost b has weight 2 and its zero modes are the $3p-3$ quadratic differentials; the measure on $\mathcal{M}_p$ contains $\big|\det\phi_i(z_j)\big|^2$, and its coincidence limit is the $q=2$ Wronskian, which vanishes at the quadratic Weierstrass points of (5.188). Bosonisation formulas for loop amplitudes, which trade such determinants for theta functions, keep track of these zeros explicitly (Alvarez-Gaumé, Gomez and Vafa; Verlinde and Verlinde).

Reading Nash (pp. 166–169, (5.100)–(5.111))

Nash's §7 is three pages. (i) On the sphere a pole of any order is allowed (5.100); footnote 8 switches to positive pole orders. (ii) He states the Weierstrass gap theorem: p gaps $1=n_1<\cdots<n_p<2p$, normal points with gaps $1,\dots,p$, finitely many Weierstrass points, none for $p=0,1$. (iii) He explains $n_1=1$ by the residue theorem on the $4p$-gon; we noted that the residue belongs to a 1-form and repaired the argument, and gave the degree argument, which needs no repair. (iv) He outlines the proof by Riemann–Roch: $D=nz_0$ (5.101), $\mathcal{L}(nz_0)=H^0(\mathcal{O}([nz_0]))$ (5.102), the naive count $n+1$ (5.103)–(5.104) contradicted by (5.105)–(5.106), the index $n+1-p$ (5.107)–(5.108), and Serre duality plus vanishing for $n\ge 2p-1$ (5.109)–(5.110). We completed the proof (staircase), identified the p conditions on the Laurent coefficients as residue conditions (Noether), and computed examples. (v) Finiteness (5.111) is quoted from Griffiths and Harris; we proved it with the Wronskian, a section of $K^{p(p+1)/2}$, which also gives the count p^3-p. (vi) The q-differential version with $\mathcal{O}(K^q\otimes[nz_0])$: we worked it out, and for $q=-1$ it recovers the $3p-3$ moduli as gaps of vector fields. (vii) Automorphisms permute the Weierstrass points, so they are finite in number, at most $84(p-1)$: we proved the fixed-point bound, finiteness, and the Hurwitz bound. Note that Nash's (5.111) should be read at $n=p$; for $n>p$, $\dim\mathcal{L}(nz_0)>1$ holds at every point.

Recap: Weierstrass gaps and points

- $\ell(n) = \dim \mathcal{L}(nz_0)$ counts functions with a single pole of order $\leq n$. Riemann–Roch: $\ell(n) = n + 1 - p + \kappa(n)$, where $\kappa(n)$ counts holomorphic differentials vanishing to order $\geq n$ at z_0; $\kappa(n) = 0$ for $n \geq 2p - 1$.
- The staircase $\ell(n)$ rises by 0 or 1; flat steps are the gaps. Exactly p gaps, $1 = n_1 < \cdots < n_p \leq 2p - 1$.
- Gaps = vanishing orders of holomorphic differentials +1 (residue conditions $\sum_k a_k h_{k-1} = 0$). Normal points: gaps $1, \ldots, p$. Weierstrass points: some differential vanishes to order $\geq p$; the Wronskian vanishes.
- The Wronskian is a section of $K^{p(p+1)/2}$; its zeros, weighted, number $p^3 - p$. Between $2p+2$ (hyperelliptic: the branch points) and $p^3 - p$ (generic) Weierstrass points. Plane quartics: the flexes.
- q-differentials: gaps of q-differentials are the vanishing orders of holomorphic $(1-q)$-differentials +1. For vector fields ($q = -1$) there are $3p - 3$ gaps, and the corresponding Schiffer variations form a basis of $T\mathcal{M}_p$.
- Automorphisms permute Weierstrass points; a non-trivial one fixes at most $2p + 2$ points; $|\operatorname{Aut}\Sigma| \leq 84(p-1)$ for $p \geq 2$, from Riemann–Hurwitz and the $(2, 3, 7)$ triangle.
- Physically: residue conditions are Ward identities; the Wronskian is the coincidence limit of zero-mode determinants.

Chapter summary

The chapter asked one question: *what are the possible shapes of a closed string worldsheet, and how many numbers does it take to specify one?* A "shape" here is a complex structure on a closed orientable surface of genus p, because the Polyakov action only sees the metric up to Weyl rescaling and diffeomorphisms. The answer came in two halves.

The first half (§5.1–§5.3) built the tools. A compact Riemann surface is an algebraic curve, so it can be written down as the zero set of polynomials in $\mathbb{CP}^n$. Its holomorphic line bundles are recorded by divisors, bookkeeping for zeros and poles, and their topology is the single integer $\deg D = c_1([D])$. Serre duality and the Riemann–Roch theorem count holomorphic sections. The exponential sequence and the Picard group say which line bundles exist, and Kähler geometry and the Hodge decomposition tie the complex structure to the topology.

The second half (§5.4–§5.7) used the tools. The space of shapes is the Teichmüller space $\mathcal{T}_p$, and dividing by the large diffeomorphisms gives the moduli space $\mathcal{M}_p$. Its dimension is computed by one application of Riemann–Roch, to the quadratic differentials. The Weierstrass points are the places where Riemann–Roch "jumps", and they make the automorphism group finite.

The whole second half fits into one labelled chain, which is worth memorising because every step is one idea of the chapter. For $p \geq 2$:

$$\begin{aligned}\dim_{\mathbb{C}} \mathcal{M}_p &\overset{(1)}{=} \dim_{\mathbb{C}} \frac{\{\text{Beltrami differentials } \mu\, \partial_z \otimes \mathrm{d}\bar{z}\}}{\{\bar{\partial} V : V \text{ a vector field}\}} \overset{(2)}{=} \dim_{\mathbb{C}} H^{0,1}_{\bar{\partial}}(\Sigma, K^*) \\ &\overset{(3)}{=} \dim_{\mathbb{C}} H^0\big(\Sigma, \mathcal{O}(K^2)\big) \overset{(4)}{=} \operatorname{index} \bar{\partial}_{K^2} \\ &\overset{(5)}{=} \deg K^2 + 1 - p \overset{(6)}{=} 2(2p-2) + 1 - p = 3p - 3. \end{aligned} \tag{5.197}$$

(1) A tangent vector to the space of shapes is a change $\delta\mu$ of the Beltrami coefficient of $ds^2 = \lambda\,|dz + \mu\,d\bar{z}|^2$ (Weyl rescalings change only λ), modulo the changes $\delta\mu = 2\,\partial_{\bar{z}} V^z$ produced by infinitesimal diffeomorphisms (Sections 5.6.3 and 5.6.4).
(2) That quotient is by definition the Dolbeault cohomology of $\bar{\partial}$ acting on sections of $K^* = T\Sigma$; every $(0,1)$-form on a curve is $\bar{\partial}$-closed (Section 5.6.5).
(3) Serre duality pairs $H^{0,1}(K^*)$ with the holomorphic sections of $K \otimes (K^*)^* = K^2$: the cotangent space of $\mathcal{M}_p$ is the space of holomorphic quadratic differentials (Section 5.6.6).
(4) The cokernel of $\bar{\partial}_{K^2}$ is dual to the holomorphic vector fields, and there are none for $p \geq 2$ (vanishing theorem, Section 5.6.9), so the kernel is the whole index.
(5) Riemann–Roch, i.e. the index theorem for $\bar{\partial}$ on a line bundle over a curve (Section 5.6.8).
(6) $\deg K = 2p - 2$ by Gauss–Bonnet (Example 5.91).

For $p = 0$ and $p = 1$ step (4) fails: there are 3 and 1 holomorphic vector fields (conformal Killing vectors), the index is -3 and 0, and the true dimensions are 0 and 1 (Section 5.6.11).

Figure 5.62 draws the same argument as a tower of spaces. Each level is obtained from the one above by dividing out a symmetry, and each quotient is "measured" by a piece of the machinery of the chapter.

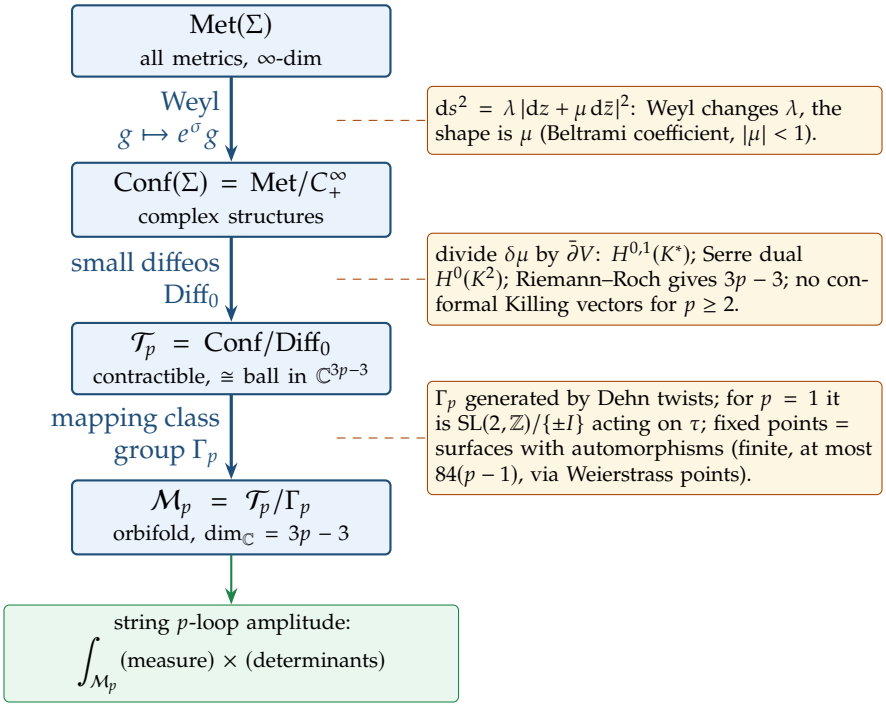


Figure 5.62: The chapter's second half as a tower of quotients. Read down the left: at each arrow a symmetry of the Polyakov path integral is divided out. Read the orange boxes: the tool that measures each quotient. What is left at the bottom is a finite-dimensional orbifold, and the string path integral at p loops becomes an ordinary integral over it (Chapter 9).

The rest of this summary has five parts: a table of every concept of the chapter with its operational meaning and where it returns; the three genus regimes side by side; the important numbers of a genus-p surface; a dictionary to physics with a map of where the chapter goes next; and a self-test.

Every concept, how to use it, and where it returns

The tables follow Nash's sections. The middle column is the *operational* meaning: what we actually do to compute or test the thing. The right column says where in the book the idea comes back. Sections §5.1–§5.3 are the first part of this chapter; §5.4–§5.7 are Sections 5.4 to 5.7.

Nash §1: algebraic varieties

Concept	Operational meaning	Returns in
homogeneous polynomial	every monomial has the same degree d, so $P(\lambda z) = \lambda^d P(z)$ and "$P = 0$" is a well-defined condition on lines through the origin	§5.2, Example 5.103
$\mathbb{CP}^n$	lines in $\mathbb{C}^{n+1}$; charts $U_i = \{z_i \neq 0\}$ with coordinates z_j/z_i, transition functions ratios of coordinates; a compact complex n-manifold	Chapter 8 (twistors, $\mathbb{CP}^1$ models), Chapter 6 (flag manifolds)
projective variety	common zero set of homogeneous $P_1, \dots, P_k$ in $\mathbb{CP}^n$; draw its real points in an affine chart	§5.2 (curves), Section 5.7.8
singular point	a point where the Jacobian $\left[\partial P_i/\partial z_j\right]$ has rank $< k$; if there are none the variety is a complex submanifold of dimension $n - k$	Example 5.104, Chapter 11 (nodal surfaces)
Chow's theorem	a closed complex submanifold of $\mathbb{CP}^n$ is automatically cut out by polynomials: "analytic = algebraic"	§5.2
Kodaira embedding	a compact complex manifold embeds in some $\mathbb{CP}^N$ iff it carries a closed positive $(1,1)$-form with rational (hence, after scaling, integral) class	§5.3 (Kähler), Chapter 10
degree–genus formula	a smooth plane curve of degree d has genus $\frac{1}{2}(d-1)(d-2)$: cubics are tori, quartics have genus 3	Section 5.7.8

Nash §2: Riemann surfaces and divisors

Concept	Operational meaning	Returns in
compact Riemann surface Σ_p	a surface with holomorphic transition functions; topologically a sphere with p handles, $\chi = 2 - 2p$; always an algebraic curve	everywhere below, Chapter 9, Chapter 11
the three regimes	$p = 0$: $\mathbb{CP}^1$; $p = 1$: the cubic $y^2 = x(x-1)(x-\lambda)$, i.e. $\mathbb{C}/(\mathbb{Z} + \tau\mathbb{Z})$; $p \geq 2$: $3p - 3$ complex moduli	Sections 5.4 and 5.5
holomorphic line bundle	cover by charts, give nowhere-zero holomorphic transition functions $g_{\alpha\beta}$; sections are local holomorphic s_α with $s_\alpha = g_{\alpha\beta} s_\beta$	§5.3, Section 5.6.7
$O(n)$ on $\mathbb{CP}^1$	transition function z^n; holomorphic sections are polynomials of degree $\leq n$, so $n+1$ of them for $n \geq 0$	Example 5.92, Chapter 6 (Borel–Weil)
canonical bundle K	$\Lambda^n T^*M$; on a curve $K = T^*\Sigma$, sections are holomorphic 1-forms $h(z)\,\mathrm{d}z$; $\deg K = 2p - 2$	Section 5.6, Chapter 9, Chapter 10
q-differential	a section of K^q, locally $\phi(z)\,\mathrm{d}z^q$ with $\tilde{\phi} = \phi\,(z')^q$; $q = -1, 0, 1, 2$: vector fields, functions, 1-forms, quadratic differentials	Sections 5.6.6 and 5.7.9, Chapter 11 (primary fields)
divisor $D = \sum n_i P_i$	list the zeros ($n_i > 0$) and poles ($n_i < 0$) of a meromorphic section with their orders; $\deg D = \sum n_i$	Section 5.7.4, Chapter 11
$\deg(\mathrm{div}\, f) = 0$	$\frac{1}{2\pi i}\oint \mathrm{d}f/f$ around the boundary of the cut-open $4p$-gon vanishes because the sides cancel in pairs	Section 5.7.2

Nash §2 (continued): divisors, homology and periods

Concept	Operational meaning	Returns in
linear equivalence	$D_1 \sim D_2$ iff $D_1 - D_2 = \operatorname{div} f$; divisors of different meromorphic sections of one bundle are equivalent	Section 5.7.4
$[D]$ and $c_1([D]) = \deg D$	build transition functions f_α / f_β from local equations of D; its first Chern number is the degree (flux = number of zeros)	Eq. (5.166), Chapter 8 (monopole)
sections and degree	a non-zero holomorphic section forces $\deg \geq 0$ (and a trivial bundle if $\deg = 0$); the converse needs Riemann–Roch ($\deg \geq p$ suffices)	Section 5.6.9
$\mathrm{Div} = H^0(\mathcal{M}^*/\mathcal{O}^*)$	a divisor is a local meromorphic function up to local holomorphic units: the sheaf viewpoint	§5.3 (Picard)
higher-dimensional divisors	formal sums of codimension-one subvarieties; where meromorphic functions have zeros and poles	Chapter 11 (compactification divisor)
a_i, b_i and intersection numbers	$H_1(\Sigma_p;\mathbb{Z}) = \mathbb{Z}^{2p}$ with $a_i \cdot b_j = \delta_{ij}$, $a \cdot a = b \cdot b = 0$; the $4p$-gon $a_1 b_1 a_1^{-1} b_1^{-1} \cdots$	Example 5.81, Section 5.5.7
holomorphic differentials	p independent holomorphic 1-forms ω_i (first kind), normalised by $\oint_{a_i} \omega_j = \delta_{ij}$	Sections 5.7.6 and 5.7.7
period matrix Ω	$\Omega_{ij} = \oint_{b_i} \omega_j$; symmetric with $\operatorname{Im}\Omega > 0$ (Riemann bilinear relations); for $p = 1$ it is τ	Section 5.5.3, Chapter 9, Chapter 10
Jacobian $J(\Sigma) = \mathbb{C}^p/\Lambda$	the lattice of periods; the Abel–Jacobi map $P \mapsto \int^P \omega$; Abel: D is a divisor of a function iff $\deg D = 0$ and its image is 0; Torelli: Ω determines Σ	§5.3 (Pic_0), Chapter 10

Nash §3: Serre duality, line bundles and Kähler manifolds

Concept	Operational meaning	Returns in
(p,q)-forms, ∂, $\bar\partial$	split $\mathrm{d} = \partial + \bar\partial$ by counting $\mathrm{d}z$'s and $\mathrm{d}\bar z$'s; $\bar\partial f = 0$ is the Cauchy–Riemann equation	Section 5.6.5, Chapter 4
Dolbeault cohomology	$H^{p,q}_{\bar\partial}(E) = \operatorname{Ker}\bar\partial/\operatorname{Im}\bar\partial$; isomorphic to sheaf cohomology $H^q(M,\Omega^p(E))$	Eq. (5.133), Chapter 3
Serre duality	$H^q(\Omega^p(E)) \cong H^{n-q}(\Omega^{n-p}(E^*))^*$ by integrating the wedge product; on a curve $H^1(\mathcal{O}(L)) \cong H^0(\mathcal{O}(K\otimes L^*))^*$	Section 5.6.6, Section 5.7.4, Chapter 11
Riemann–Roch	$h^0(L) - h^0(K\otimes L^*) = \deg L + 1 - p$; the index theorem for $\bar\partial_L$	Sections 5.6.8 and 5.7.4, Chapter 9
line bundle as a cocycle	transition functions with $g_{\alpha\beta}g_{\beta\gamma} = g_{\alpha\gamma}$ define a class in $H^1(M;\mathcal{E}^*)$	Chapter 10 (determinant bundles)
exponential sequence	$0 \to \mathbb{Z} \to \mathcal{E} \to \mathcal{E}^* \to 0$ with $f \mapsto e^{2\pi i f}$; the connecting map is c_1; smooth line bundles $\leftrightarrow H^2(M;\mathbb{Z})$	Chapter 8, Chapter 10
Picard group	$\mathrm{Pic}(M) = H^1(M;\mathcal{O}^*)$: holomorphic line bundles; $\mathrm{Pic}_0 = H^1(\mathcal{O})/H^1(\mathbb{Z})$, the Jacobian for a curve; flat U(1) bundles, Wilson lines	Chapter 6 (Borel–Weil), Chapter 10
Kähler manifold	Hermitian $g_{i\bar\jmath}$ with $\omega = \frac{i}{2}g_{i\bar\jmath}\mathrm{d}z^i\wedge\mathrm{d}\bar z^j$ closed; locally $g_{i\bar\jmath} = \partial_i\partial_{\bar\jmath}K$	Section 5.5.1 (Weil–Petersson), Chapter 9, Chapter 10
Fubini–Study metric	$\omega = \frac{i}{2}\partial\bar\partial\ln(1+\lvert z\rvert^2)$ on $\mathbb{CP}^n$; on $\mathbb{CP}^1$ the round sphere	Chapter 8, Chapter 6
submanifolds are Kähler	restrict ω; closedness is inherited, so every projective variety is Kähler	Section 5.7.8
Kähler Hodge theory	$\Delta_d = 2\Delta_{\bar\partial}$, so harmonic forms split by type: $b_k = \sum_{p+q=k} h^{p,q}$, $h^{p,q} = h^{q,p}$, b_{odd} even, $b_2 > 0$	Chapter 7 (SUSY QM), Chapter 11

Nash §4: the Teichmüller space $\mathcal{T}_p$ (Section 5.4)

Concept	Operational meaning	Returns in
metric → complex structure	write $\mathrm{d}s^2 = A\mathrm{d}x^2 + 2B\mathrm{d}x\mathrm{d}y + C\mathrm{d}y^2$ as $\lambda\,\lvert\mathrm{d}z + \mu\mathrm{d}\bar{z}\rvert^2$; solve for $\lambda > 0$ and $\lvert\mu\rvert < 1$ (Eq. (5.79)); μ is the shape	Section 5.6, Chapter 9
Beltrami equation	$f_{\bar{z}} = \mu f_z$: a solution f is a holomorphic (isothermal) coordinate for the metric; infinitesimal circles are ellipses of axis ratio $\frac{1+\lvert\mu\rvert}{1-\lvert\mu\rvert}$	Section 5.6.2
conformal class, Conf	Met/C^∞_+: metrics up to Weyl rescaling $g \mapsto e^\sigma g$; one-to-one with complex structures (given an orientation)	Chapter 9, Chapter 10 (Weyl anomaly)
Diff^+, Diff_0	orientation-preserving diffeomorphisms; those isotopic to the identity ("small"); reversing orientation conjugates the complex structure	Section 5.5
conformal Killing vector	$\mathcal{L}_V g = Cg$, i.e. $\partial_{\bar{z}} V^z = 0$: holomorphic vector fields; 3, 1, 0 of them for $p = 0, 1, \geq 2$ (Table 5.1)	Section 5.6.11, Chapter 9 (ghost zero modes)
uniformisation	$\Sigma = \tilde{\Sigma}/\pi_1$ with $\tilde{\Sigma} = S^2$, $\mathbb{C}$ or $\mathbb{H}$; each conformal class has a constant-curvature metric of sign $\operatorname{sgn}\chi$ (Gauss–Bonnet)	Section 5.5.6, Section 5.7.11
$\mathrm{Met}_{\mathrm{const}}$ and $\mathcal{T}_p$	pick the constant-curvature representative (area 1 for the torus); $\mathcal{T}_p = \mathrm{Conf}/\mathrm{Diff}_0 = \mathrm{Met}_{\mathrm{const}}/\mathrm{Diff}_0$	Chapter 9
$\mathcal{T}_1 = \mathbb{C}^+$	flat unit-area metrics on T^2 modulo small diffeomorphisms $\leftrightarrow \tau$ in the upper half plane $\leftrightarrow \mathrm{SL}(2,\mathbb{R})/\mathrm{SO}(2)$ (Eq. (5.98))	Section 5.5, Chapter 11

Nash §5: the moduli space $\mathcal{M}_p$ (Section 5.5)

Concept	Operational meaning	Returns in
mapping class group Γ_p	$\mathrm{Diff}^+/\mathrm{Diff}_0$: diffeomorphisms up to isotopy; for the torus the integer matrices $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ of determinant 1 acting on H_1	Chapter 9, Chapter 12
$\mathcal{M}_p = \mathcal{T}_p/\Gamma_p$	shapes counted once; $\mathrm{Met}_{\mathrm{const}}/\mathrm{Diff}^+$	Chapter 9, Chapter 11
orbifold	quotient by a group acting with finite stabilisers; locally $\mathbb{C}^n/G$; a cone point of angle $2\pi/\lvert G\rvert$ (Figure 5.41)	Section 5.7.11, Chapter 11
modular group	$\mathrm{SL}(2,\mathbb{Z})/\{\pm I\}$ acting by $\tau \mapsto \frac{a\tau+b}{c\tau+d}$; generated by $S : \tau \mapsto -1/\tau$, $T : \tau \mapsto \tau + 1$	Chapter 11 (modular invariance), Chapter 12
fundamental domain F	$\lvert\Re\tau\rvert \leq \frac{1}{2}$, $\lvert\tau\rvert \geq 1$; reduce any τ into F by the S/T algorithm (Example 5.76); orbifold points i, $e^{i\pi/3}$ and a cusp at $i\infty$	Chapter 9, Chapter 11
invariant measure	$\mathrm{d}^2\tau/\tau_2^2$ (hyperbolic area); $\mathrm{vol}(F) = \pi/3$ (Example 5.79); the one-loop amplitude $\int_F \frac{\mathrm{d}^2\tau}{\tau_2^2}\mathcal{Z}$	Chapter 9, Chapter 10
modular forms	$f(\gamma\tau) = (c\tau + d)^q f(\tau)$: q-differentials on $\mathcal{M}_1$; Eisenstein series G_4	Example 5.90, Chapter 11
Fuchsian group	a discrete subgroup $G \subset \mathrm{SL}(2,\mathbb{R})/\{\pm I\}$ acting freely on $\mathbb{H}$; $\Sigma_p = \mathbb{H}/G$ for $p \geq 2$ with a $4p$-gon fundamental domain	Section 5.7.11
topology of $\mathcal{M}_p$	$\mathcal{T}_p$ is contractible and is the orbifold universal cover; $\pi_1^{\mathrm{orb}}(\mathcal{M}_p) = \Gamma_p$	Chapter 10 (global anomalies)
Dehn twist	cut along a simple closed curve, twist by 2π, reglue; on T^2 the twists about a and b are T and a conjugate of T; they generate Γ_p	Chapter 12 (surgery, framings)
Weil–Petersson metric	$\langle\mu,\nu\rangle = \int \mu\bar{\nu}\, g_{z\bar{z}}\, \mathrm{d}^2 z$ (up to normalisation) on Beltrami differentials; Kähler on $\mathcal{T}_p$	Chapter 9

Nash §6: the dimension of the moduli space (Section 5.6)

Concept	Operational meaning	Returns in
transformation laws of λ, μ	under $z \mapsto w$ holomorphic, $\mu \mapsto \mu\, w'/\overline{w'}$ and $\lambda \mapsto \lambda\, \lvert w' \rvert^{-2}$; so μ is a section of $K^{-1} \otimes \bar{K}$ (Eq. (5.114))	Chapter 9
Beltrami differential	$\mu\, \partial_z \otimes \mathrm{d}\bar{z}$; its variation $\delta\mu = \delta g_{\bar{z}\bar{z}}/g_{z\bar{z}}$ is a tangent vector to the space of shapes (Eq. (5.120))	Chapter 9
diffeomorphism directions	$\delta g = \mathcal{L}_V g$ gives $\delta\mu = 2\partial_{\bar{z}} V^z$ (Eq. (5.130)): the "pure gauge" deformations	Chapter 9 (Faddeev–Popov)
$T_{[g]}\mathcal{M}_p = H^{0,1}(K^*)$	Beltrami differentials modulo $\bar{\partial}$(vector fields); represented by harmonic Beltrami differentials $\bar{\phi}/g_{z\bar{z}}$	Chapter 9
quadratic differentials	holomorphic sections $\phi(z)\mathrm{d}z^2$ of K^2; $T^*\mathcal{M}_p = H^0(K^2)$ via the pairing $\int \mu\phi\, \mathrm{d}^2z$ (Eq. (5.138)); explicit for $p = 0, 1, 2$	Chapter 9 (b-ghost zero modes)
index of $\bar{\partial}_{K^2}$	Riemann–Roch: $\deg K^2 + 1 - p = 3p - 3$; equivalently index $= \int \mathrm{ch}(K^2)\, \mathrm{td}(T\Sigma)$ (Eq. (5.149))	Chapter 9, Chapter 10 (ghost number anomaly)
vanishing theorems	a bundle of negative degree has no holomorphic sections; hence no holomorphic vector fields for $p \geq 2$ and $H^1(K^2) = 0$; Weitzenböck/Landau-level picture	Chapter 6 (Borel–Weil), Chapter 9
$\dim_{\mathbb{C}} \mathcal{M}_p$	$0, 1, 3p - 3$ for $p = 0, 1, \geq 2$; in general $\dim \mathcal{M}_p - \dim(\mathrm{CKV}) = 3p - 3$ (Eq. (5.150))	Chapter 9, Chapter 10 (critical dimension)
other counts of $3p - 3$	branched covers of $\mathbb{CP}^1$ (Riemann), pants decompositions ($3p - 3$ lengths and twists), punctures ($3p - 3 + n$), vector-field gaps	Sections 5.6.12 and 5.7.10, Chapter 11

Nash §7: Weierstrass gaps and points (Section 5.7)

Concept	Operational meaning	Returns in
$\mathcal{L}(nz_0)$, $\ell(n)$	meromorphic functions with a single pole, of order $\le n$, at z_0; $\cong H^0(\mathcal{O}([nz_0]))$ (Eq. (5.166))	Chapter 11
sphere: no restriction	on $\mathbb{CP}^1$ the function $(z-z_0)^{-n}$ exists for every n; $\ell(n) = n+1$ (Eq. (5.163))	Section 5.7.1
no single simple pole	for $p \ge 1$ a function with one simple pole would be a degree-one map to $\mathbb{CP}^1$, an isomorphism; residue theorem for $f\omega$	Section 5.7.2
gap, gap theorem	n is a gap if $\ell(n) = \ell(n-1)$; there are exactly p gaps, $1 = n_1 < \cdots < n_p \le 2p-1$ (the staircase, Figure 5.56)	Chapter 11 (correlator constraints)
Riemann–Roch for $\ell(n)$	$\ell(n) = n+1-p+\kappa(n)$ with $\kappa(n)$ the number of holomorphic differentials vanishing to order $\ge n$ at z_0; $\kappa = 0$ for $n \ge 2p-1$ (Eq. (5.176))	Section 5.7.10
residue conditions	$f = \sum_k a_k (z-z_0)^{-k} + \dots$ exists iff $\sum_k a_k h_{k-1} = 0$ for each differential $h\,\mathrm{d}z$: gaps = vanishing orders +1	Chapter 11 (Ward identities)
normal and Weierstrass points	normal: gaps $1, \dots, p$; Weierstrass: a differential vanishes to order $\ge p$; weight $w = \sum_i (n_i - i)$	Section 5.7.11
Wronskian	$W = \det[h_i^{(k)}]$, a section of $K^{p(p+1)/2}$ vanishing to order w; so $\sum w = p^3 - p$ and there are between $2p+2$ and p^3-p Weierstrass points	Section 5.7.12
examples	torus: none; hyperelliptic: the $2p+2$ branch points, weight $\frac{p(p-1)}{2}$; plane quartics: flexes (24 ordinary or 12 hyperflexes)	Section 5.7.8
q-differential gaps	$\ell_q(n) = (2q-1)(p-1) + n + \kappa_{1-q}(n)$; number of q-Weierstrass points $p(p-1)^2(2q-1)^2$ (Eq. (5.188))	Chapter 11
vector-field gaps	for $q=-1$ there are $3p-3$ forbidden pole orders; their Schiffer variations span $T\mathcal{M}_p$	Chapter 9
automorphisms	permute Weierstrass points, fix $\le 2p+2$ points if non-trivial; hence finite, and $\lvert\mathrm{Aut}\,\Sigma\rvert \le 84(p-1)$ (Hurwitz; equality for the Klein quartic, 168)	Section 5.5.1 (orbifold points), Chapter 9
zero-mode determinants	$\langle b(z_1)\cdots b(z_p)\rangle \propto \det h_i(z_j)$; its coincidence limit is the Wronskian (Eq. (5.196))	Chapter 9, Chapter 11

The three genus regimes side by side

Almost every statement of the chapter splits into the cases $p=0$, $p=1$ and $p \ge 2$. The split is always for the same reason: the sign of $\chi = 2-2p = \deg K^*$, which decides the curvature of the uniformising metric and whether holomorphic vector fields exist.

	$p = 0$	$p = 1$	$p \geq 2$
model	$\mathbb{CP}^1$, $y = \text{const}$	$y^2 = x(x-1)(x-\lambda)$, $\mathbb{C}/(\mathbb{Z}+\tau\mathbb{Z})$	$y^2 =$ degree $2p+2$ (hyperelliptic), plane quartic ($p=3$), $\mathbb{H}/G$
universal cover, curvature	S^2, $+1$	$\mathbb{C}$, 0	$\mathbb{H}$, -1
$\deg K = 2p-2$	-2	0	> 0
holomorphic differentials $h^0(K)$	0	1 (dz)	p
conformal Killing vectors $h^0(K^*)$	3 ($\mathfrak{sl}(2,\mathbb{C})$)	1 (∂_z)	0
quadratic differentials $h^0(K^2)$	0	1 (dz^2)	$3p-3$
index $\bar\partial_{K^2} = 3p-3$	-3	0	$3p-3$
$\mathcal{T}_p$	point	$\mathbb{C}^+$	ball in $\mathbb{C}^{3p-3}$
Γ_p	trivial	$\mathrm{SL}(2,\mathbb{Z})/\{\pm I\}$	generated by Dehn twists
$\mathcal{M}_p$	point	F (Figure 5.44)	orbifold of $\dim_{\mathbb{C}} 3p-3$
$\mathrm{Aut}\,\Sigma$	$\mathrm{SL}(2,\mathbb{C})/\{\pm I\}$	$T^2\rtimes$ (finite)	finite, $\leq 84(p-1)$
Weierstrass points	none	none	$2p+2$ to p^3-p
string amplitude	tree level: fix 3 points	one loop: $\int_F d^2\tau/\tau_2^2$	p loops: $\int_{\mathcal{M}_p}$

Key idea: One sign decides everything

The sign of $\deg T\Sigma = 2-2p$ controls the whole table. If it is positive, $T\Sigma$ has holomorphic sections (a group of automorphisms that is too big and fixes 3 points), and there are no deformations. If it is zero, there is one of each. If it is negative, there are no holomorphic vector fields, automorphisms are finite, and Riemann–Roch counts the deformations exactly: $3p-3$.

The numbers of a genus-p surface

It helps to see the chapter's integers together. Each is a degree or a dimension, and each follows from Riemann–Roch plus Gauss–Bonnet. Figure 5.63 plots them for small p.

quantity	value	where
Euler characteristic χ	$2-2p$	§5.2
$\deg K$, zeros of a holomorphic differential	$2p-2$	Example 5.91
holomorphic differentials $h^0(K)$	p	§5.3
real rank of $H_1(\Sigma;\mathbb{Z})$	$2p$	§5.2
$\dim_{\mathbb{C}} \mathcal{M}_p = h^0(K^2)$	$3p-3$	Eq. (5.157)
$h^0(K^q)$, $q \geq 2$	$(2q-1)(p-1)$	Eq. (5.187)
Weierstrass gaps at any point	p, all $\leq 2p-1$	Section 5.7.5
total Weierstrass weight	p^3-p	Eq. (5.186)
Weierstrass points on a hyperelliptic surface	$2p+2$	Example 5.102
$\lvert\mathrm{Aut}\,\Sigma\rvert$	$\leq 84(p-1)$	Eq. (5.194)
area of the constant-curvature ($K=-1$) metric	$4\pi(p-1)$	Eq. (5.92)

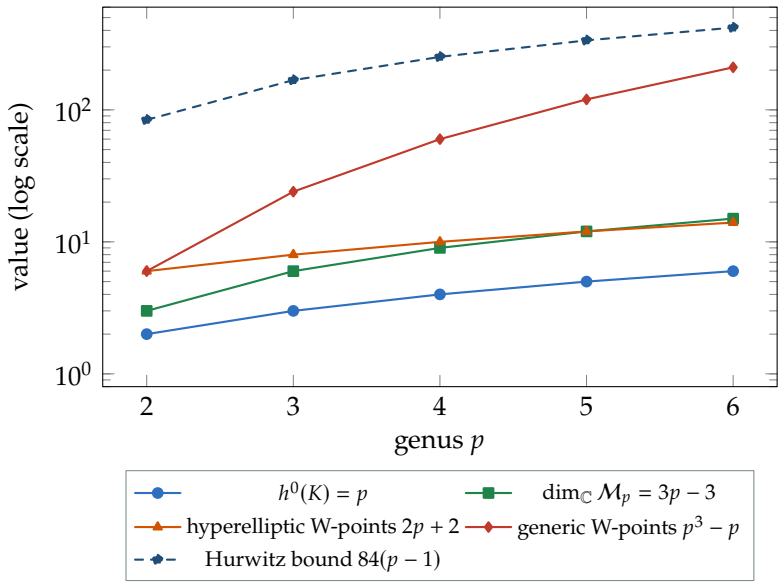


Figure 5.63: The integers of a genus-p surface on a log scale. The linear quantities (p, $3p - 3$, $2p + 2$) are dimensions and degrees. The number of Weierstrass points grows like p^3 because the Wronskian is a section of the high power $K^{p(p+1)/2}$. The Hurwitz bound stays above the generic Weierstrass count only for small p; for $p \geq 9$ the cubic overtakes the linear bound, which is consistent because a generic surface has no automorphisms at all.

Worked example 5.109: Checking the table at genus 2

Take $y^2 = \prod_{i=1}^{6}(x - e_i)$. The holomorphic differentials are $\mathrm{d}x/y$ and $x\,\mathrm{d}x/y$ (Example 5.102), so $h^0(K) = 2 = p$. The quadratic differentials are products of these, $\{\mathrm{d}x^2/y^2,\ x\,\mathrm{d}x^2/y^2,\ x^2\,\mathrm{d}x^2/y^2\}$, and the count agrees with the table:

$$h^0(K^2) \overset{(1)}{=} \deg K^2 + 1 - p \overset{(2)}{=} 4 + 1 - 2 = 3 \overset{(3)}{=} 6 - 3.$$

(1) Riemann–Roch plus vanishing of $H^1(K^2)$ (Section 5.6.9).
(2) $\deg K = 2p - 2 = 2$.
(3) Three complex moduli: the six branch points e_i modulo the three complex parameters of $\mathrm{SL}(2, \mathbb{C})$ acting on x.

The Weierstrass points are the six branch points, each of weight $\frac{p(p-1)}{2} = 1$, total $6 = 2^3 - 2$, so the generic count $p^3 - p$ and the hyperelliptic count $2p + 2$ coincide. The hyperelliptic involution $(x, y) \mapsto (x, -y)$ is an automorphism fixing exactly these six points, the maximum $2p + 2$ allowed, and $|\operatorname{Aut}\Sigma| \leq 84$ is respected (the actual maximum at genus 2 is 48).

Dictionary to physics

The left column is the mathematics of the chapter; the right column is the name the same object carries in string theory and conformal field theory, where Chapters 9 and 11 will use it.

Dictionary: Chapter V for string theorists

Mathematics	Physics
Σ_p, genus p	closed-string worldsheet at p loops
conformal class, Beltrami coefficient μ	the metric modulo Weyl: what the classical Polyakov action sees
Diff_0, $\mathcal{L}_V g$	gauge (reparametrisation) symmetry and its Faddeev–Popov directions
Γ_p, $\mathrm{SL}(2,\mathbb{Z})$	large diffeomorphisms; modular invariance of the torus partition function
$\mathcal{T}_p$, $\mathcal{M}_p$	the integration region of the p-loop amplitude: $\mathcal{T}_p$ overcounts, $\mathcal{M}_p$ counts once
Beltrami differentials $H^{0,1}(K^*)$	moduli insertions $\int \mu\, T$ / $\int \mu\, b$ in the measure
quadratic differentials $H^0(K^2)$	zero modes of the b ghost
conformal Killing vectors $H^0(K^*)$	zero modes of the c ghost; "fix three vertex operators on the sphere"
index $\bar\partial_{K^2} = 3p - 3$	ghost number anomaly: $\#b$ zero modes $-\#c$ zero modes
q-differential	a primary field of conformal weight $(q, 0)$
modular form of weight q	a torus quantity transforming like $(c\tau + d)^q$
$\mathrm{d}^2\tau/\tau_2^2$ on F	one-loop measure; the cusp $\tau_2 \to \infty$ is the infrared (long, thin torus)
Dehn twist	large diffeomorphism; in Chapter 12 a change of framing
residue conditions, gap theorem	Ward identities: which pole structures a correlator may have
Wronskian, Weierstrass points	coincidence limit of zero-mode determinants
Kähler, Weil–Petersson	$\mathcal{M}_p$ carries a natural Kähler metric; holomorphic factorisation of Z_p

Where Chapter V goes next

Figure 5.64 shows where the ideas of this chapter are used later. The main customer is Chapter 9, where the Polyakov integral over metrics is reduced, using exactly the tower of Figure 5.62, to a finite-dimensional integral over $\mathcal{M}_p$. The zero modes of $\bar\partial_{K^2}$ and $\bar\partial_{K^*}$ counted here become the Faddeev–Popov ghost zero modes, and the Weil–Petersson metric on $\mathcal{T}_p$ is built from the Beltrami and quadratic differentials of Section 5.6. Chapter 10 uses the determinant line bundle over $\mathcal{M}_p$ and Riemann–Roch (the Mumford isomorphism) to find the critical dimension 26, and the Kähler structure of $\mathcal{T}_p$ to explain why the partition function is the modulus squared of a holomorphic function (Belavin–Knizhnik). Chapter 11 uses modular invariance of the torus partition function, the Weierstrass gap theorem as a constraint on correlation functions, Serre duality to turn the stress tensor into a connection on a line bundle over $\mathcal{M}_p$, and the compactified moduli space $\overline{\mathcal{M}}_p$ whose boundary, the compactification divisor, consists of surfaces with nodes. Chapter 6 uses the Kodaira vanishing theorem and the exponential sequence on the Kähler manifold G/T to prove the Borel–Weil theorem. Chapter 8 calls the analogy between instanton moduli spaces and Teichmüller theory explicit: in both cases the moduli are counted by the index of a linearised operator. Chapter 12 uses Dehn twists and the mapping class group in surgery and in the modular functor.

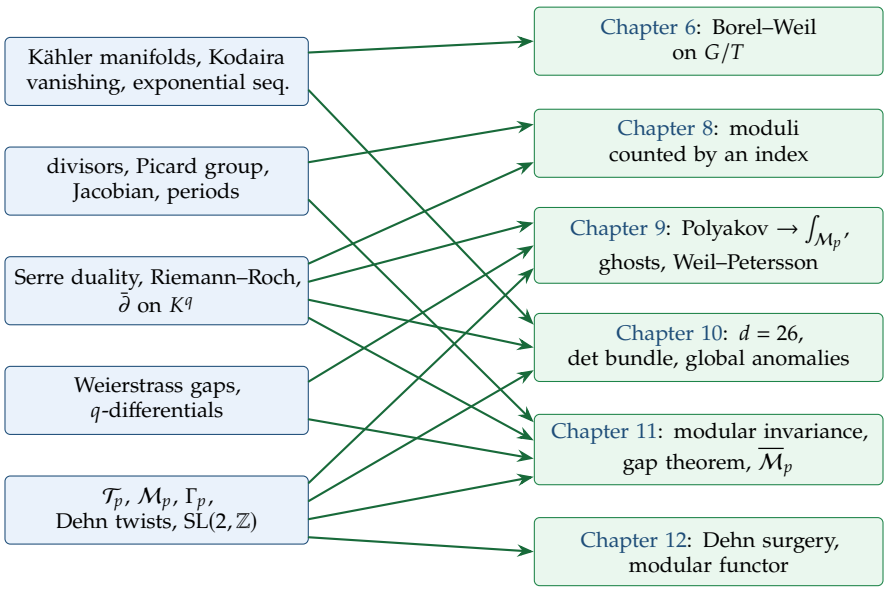


Figure 5.64: Where the chapter goes. Blue: the mathematics of Chapter V. Green: the later chapters that use it. Every route to physics passes through counting holomorphic sections (Riemann–Roch) or through dividing by diffeomorphisms (the tower of Figure 5.62).

Self-test

Try it: Ten questions on Chapter V

1. Why is $x^3 + y^2z - z^3 = 0$ a well-defined curve in $\mathbb{CP}^2$ but $x^3 + y^2 - z = 0$ not?
   Answer: the first is homogeneous of degree 3, so its zero set is a union of lines through 0; the second changes under $(x, y, z) \mapsto \lambda(x, y, z)$ by different powers of λ.

2. A meromorphic function on a torus has poles of order 2 at 0 and nowhere else. How many zeros does it have?
   Answer: two, counted with multiplicity, since $\deg \operatorname{div} f = 0$. ($\wp(z) - c$ is such a function.)

3. What is $\dim H^0(\mathbb{CP}^1, \mathcal{O}(n))$ for $n \geq -1$, and how does Riemann–Roch confirm it?
   Answer: $n + 1$; $\deg +1 - p = n + 1$, and $H^1 = 0$ because its dual $H^0(\mathcal{O}(-2 - n))$ vanishes for $n \geq -1$.

4. Write the metric $\mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}x\,\mathrm{d}y$ as $\lambda\,|\mathrm{d}z + \mu\,\mathrm{d}\bar z|^2$.
   Answer: $A = C = 1$, $B = \frac{1}{2}$, so $2\sqrt{AC - B^2} = \sqrt{3}$, and Eq. (5.79) gives

$$\mu = \frac{A - C + 2iB}{A + C + 2\sqrt{AC - B^2}} = \frac{i}{2 + \sqrt{3}}, \qquad \lambda = \frac{A + C + 2\sqrt{AC - B^2}}{4} = \frac{2 + \sqrt{3}}{4}.$$

   Check: $\lambda(1 + |\mu|^2) = \frac{2+\sqrt{3}}{4} \cdot 4(2 - \sqrt{3}) = 1 = \frac{A+C}{2}$.

5. Reduce $\tau = \frac{3}{10} + \frac{i}{10}$ into the fundamental domain.
   Answer: S: $-1/\tau = -3 + i$; T^3: i. So $\tau \sim i$, the square torus, an orbifold point of order 2.

6. Why does the torus have a 1-dimensional moduli space although the index of $\bar\partial_{K^2}$ is 0?
   Answer: index $= h^0(K^2) - h^0(K^*) = 1 - 1$; the one holomorphic vector field ∂_z cancels the one quadratic differential $\mathrm{d}z^2$ in the index, but not in the dimension.

7. How many real parameters does a pants decomposition of a genus-4 surface give?
Answer: $2p-2=6$ pairs of pants, $3p-3=9$ cuffs, a length and a twist each: $18=6p-6$ real, 9 complex.

8. List the gap sequences at a normal point and at a branch point of a genus-3 hyperelliptic surface.
Answer: $\{1,2,3\}$ and $\{1,3,5\}$; the weight of the latter is $0+1+2=3=\frac{p(p-1)}{2}$.

9. Verify that 8 hyperelliptic Weierstrass points of weight 3 give the total p^3-p at genus 3.
Answer: $8\times 3=24=27-3$.

10. Why must an automorphism group of a genus-$p\geq 2$ surface be finite, in one sentence?
Answer: it permutes the finitely many Weierstrass points, and an automorphism fixing more than $2p+2$ points is the identity, so the group maps to a finite permutation group with a kernel of order at most 2 (the hyperelliptic involution).

Recap: Chapter V in five sentences

- A compact Riemann surface is an algebraic curve; its line bundles are divisors up to equivalence, and their topology is $\deg = c_1$.
- Riemann–Roch $h^0(L)-h^0(K\otimes L^*)=\deg L+1-p$, with Serre duality and vanishing for negative degree, counts every space of holomorphic objects we met.
- The space of worldsheet shapes is $\mathcal{M}_p=\mathrm{Met}/(\mathrm{Weyl}\ltimes\mathrm{Diff}^+)=\mathcal{T}_p/\Gamma_p$; for the torus it is $\mathbb{C}^+/\mathrm{SL}(2,\mathbb{Z})$, the fundamental domain F.
- Its tangent space is $H^{0,1}(K^*)$, its cotangent space the quadratic differentials $H^0(K^2)$, and its dimension $3p-3$ for $p\geq 2$ (with 0 and 1 for $p=0,1$).
- Weierstrass points, where the gap sequence is not $1,\ldots,p$, number between $2p+2$ and p^3-p and make automorphism groups finite ($\leq 84(p-1)$), which is why $\mathcal{M}_p$ is only an orbifold.

Chapter 6

Infinite Dimensional Groups

Where this chapter is going

Every symmetry group we met in a first course on quantum mechanics has a finite number of parameters. A rotation needs three angles. A Lorentz transformation needs six numbers. Quantum field theory, though, keeps producing symmetries with a parameter *at every point of space*. A gauge transformation $g(x)$ is a whole function. So is a change of coordinates $x \mapsto x'(x)$. So is a conformal transformation of a two-dimensional worldsheet $z \mapsto w(z)$. The groups formed by such transformations have infinitely many parameters: they are *infinite-dimensional groups*.

This chapter is a first tour of these groups and of their quantum mechanics. We proceed in three steps, following Nash's three sections.

(1) *The examples* (§6.1). Groups of maps $\mathrm{Map}(X, G)$, with the loop group $LG = \mathrm{Map}(S^1, G)$ as the star; the diffeomorphism group $\mathrm{Diff}(X)$, with $\mathrm{Diff}(S^1)$ and its Lie algebra of vector fields on a circle as the star; and the gauge group $\mathcal{G}$ of a principal bundle. We compute their Lie algebras, count their connected components, and write down explicit bases.

(2) *Extensions* (§6.2). In quantum mechanics a symmetry acts on rays, not on vectors, so the operators multiply only up to phases. For finite-dimensional semisimple groups the phases can always be removed. For loop groups and for $\mathrm{Diff}(S^1)$ they cannot: the commutators acquire a *central term*, a c-number proportional to the identity. This is where the *affine Kac–Moody algebra*, the *Virasoro algebra* with its $\frac{c}{12}(m^3 - m)$, and the *Schwinger term* of current algebra come from. We derive the central terms three ways: from the consistency (cocycle) condition, from normal ordering in free field theories, and from the Sugawara construction.

(3) *Representations* (§6.3). Requiring the energy L_0 to be bounded below forces the central term to be nonzero and leads to highest-weight representations. Unitarity then quantises the central charges: the discrete series $c = 1 - 6/((m+1)(m+2))$ of the Virasoro algebra and the integer level of a Kac–Moody algebra. Finally the Borel–Weil construction realises representations as spaces of holomorphic sections of line bundles, exactly as wavefunctions of the lowest Landau level are holomorphic sections.

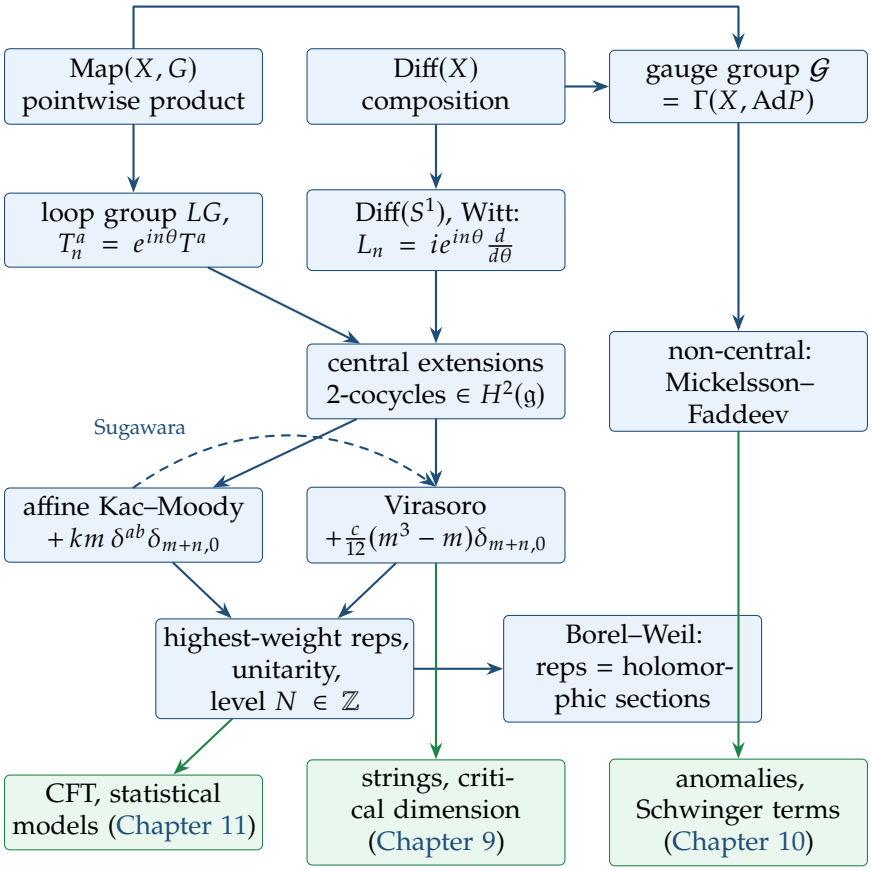


Figure 6.1: The logic of Chapter 6. Top row: three families of infinite-dimensional groups. Middle: in quantum theory their Lie algebras acquire central (or, for Map(X, G) with dim $X > 1$, non-central) terms. Bottom: the representation theory of the centrally extended algebras, and the physics chapters that use it (green).

Dictionary: infinite-dimensional groups for physicists

Mathematics	Physics
Map(X, G), gauge group $\mathcal{G}$	local gauge transformations $g(x)$
Diff(X)	general coordinate transformations
loop algebra Map$(S^1, \mathfrak{g})$	Fourier modes of a current $j^a(\theta)$ on a spatial circle
affine Kac–Moody algebra	equal-time current algebra with Schwinger term
Witt algebra vect(S^1)	classical conformal generators of a closed string
Virasoro algebra, central charge c	quantum conformal generators; c counts degrees of freedom
2-cocycle	the c-number (Schwinger) term in a commutator
central extension	symmetry realised projectively on the Hilbert space
highest-weight representation	spectrum with a lowest-energy state $\lvert h\rangle$
level k	the coefficient of the anomaly/Schwinger term

Watch out: conventions for generators in this chapter

In the toolkit chapters (Chapter T3) and in the gauge theory chapters we use the mathematicians' convention: Lie algebra elements are *anti-Hermitian*, and a gauge field is $A = A_\mu \, \mathrm{d}x^\mu$ with $A_\mu \in \mathfrak{g}$ anti-Hermitian. In this chapter Nash switches to the physicists' convention for generators, and so do we. The basis elements T^a of $\mathfrak{g}$ are *Hermitian*, a group element near the identity is $g = \exp(i\varepsilon_a T^a)$, and

$$[T^a, T^b] = i f^{abc}\, T^c, \qquad f^{abc} \text{ real and totally antisymmetric.}$$

The dictionary is $X_{\text{math}} = iX_{\text{phys}}$: the anti-Hermitian element $i\varepsilon_a T^a$ is what the mathematician calls a Lie algebra element. We always normalise $\operatorname{tr}(T^a T^b) = \frac{1}{2}\delta^{ab}$ in the defining representation of SU(N); for SU(2), $T^a = \sigma^a/2$ with the Pauli matrices and $f^{abc} = \varepsilon^{abc}$. We set $\hbar = 1$ except in Section 6.2.7, where the powers of $\hbar$ are the point.

6.1 Some infinite dimensional groups

What does it mean for a group to be infinite-dimensional, and where do such groups show up in physics? The short answer is: whenever a symmetry transformation is allowed to depend on position. A global phase rotation $\psi \mapsto e^{i\lambda}\psi$ has one parameter λ. A local one, $\psi(x) \mapsto e^{i\lambda(x)}\psi(x)$, has one parameter *per point*, that is, a whole function $\lambda(x)$. The group of all such local phase rotations is the group of maps from space into U(1). Almost every example in this chapter is, at least locally, a group of maps Map(X, Y) from a finite-dimensional manifold X into a finite-dimensional manifold Y. That observation is our guide.

6.1.1 The groups we already know

Before going infinite, it helps to lay out the finite-dimensional groups that physicists use, since everything below is built out of them.

Type	Examples	Where physics uses them
finite (dimension 0)	point groups, S_n, $\mathbb{Z}_n$	crystallography, identical particles
infinite discrete (dim. 0)	$\mathbb{Z}$, $\mathrm{PSL}(2,\mathbb{Z}) = \mathrm{SL}(2,\mathbb{Z})/\{\pm I\}$	lattice translations; modular invariance of the torus in string theory
compact Lie	$\mathrm{U}(1)$, $\mathrm{U}(n)$, $\mathrm{O}(n)$, $\mathrm{Sp}(n)$, E_6, E_7, E_8, F_4, G_2	internal (gauge, flavour) symmetries, GUTs, heterotic string
non-compact Lie	$\mathrm{SL}(2,\mathbb{C})$, Poincaré, $\mathrm{SL}(2,\mathbb{R})$	Lorentz invariance; conformal symmetry

Worked example 6.1: the modular group, and why $\pm I$ is divided out

A torus is $\mathbb{C}$ modulo the lattice generated by 1 and a complex number τ with $\operatorname{Im}\tau > 0$. A change of lattice basis is a matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}(2,\mathbb{Z})$, and it changes τ by

$$\tau \;\mapsto\; \tau' \overset{(1)}{=} \frac{a\tau + b}{c\tau + d} \overset{(2)}{=} \frac{(-a)\tau + (-b)}{(-c)\tau + (-d)}.$$

(1) The new basis vectors are $a\tau + b$ and $c\tau + d$; rescaling the lattice so the second becomes 1 gives the new modulus.

(2) Multiplying numerator and denominator by -1 changes nothing.

So M and $-M$ act identically, and the group that acts effectively is $\mathrm{SL}(2,\mathbb{Z})/\{\pm I\}$. It is generated by $T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ ($\tau \mapsto \tau + 1$) and $S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ ($\tau \mapsto -1/\tau$). It is infinite but discrete: zero-dimensional. A one-loop string amplitude must be invariant under it (Chapter 9).

These groups have a complete representation theory, especially the compact ones: every unitary irreducible representation of a compact group is finite dimensional and is labelled by a highest weight. For infinite-dimensional groups the theory is much younger. Nash's aim, and ours, is to meet the infinite-dimensional groups that the rest of the book needs, to see their structure concretely, and to learn the basic facts about their representations.

6.1.2 Infinite-dimensional manifolds

A finite-dimensional Lie group is a manifold, so we can talk about curves in it, tangent vectors and a Lie algebra. To do the same for $\mathrm{Map}(X, G)$ we need a notion of an infinite-dimensional manifold. The idea is the obvious one.

Working definition 6.1 (infinite-dimensional manifold)

An **infinite-dimensional manifold M modelled on** a topological vector space E [a vector space, here usually of functions, in which it makes sense to say that two vectors are close]

is a space covered by open sets, each of which is identified (by a chart) with an open set of E, such that the changes of chart are smooth maps between open subsets of E [where two charts overlap, converting one chart's coordinates into the other's is a smooth map]. *Operationally:* near every point you can describe M by "coordinates" that are elements of E, and small moves in M are small vectors in E.

In finite dimensions $E = \mathbb{R}^n$ or $\mathbb{C}^n$. In infinite dimensions there is a choice to make, and the name of the manifold records it:

- a **Hilbert manifold** is modelled on a Hilbert space, for example $L^2(S^1)$, the square-integrable functions on a circle, with $\|f\|^2 = \int_0^{2\pi} |f|^2 \, \mathrm{d}\theta$;
- a **Sobolev manifold** is modelled on a Sobolev space H^s, functions whose first s derivatives are square integrable, $\|f\|_s^2 = \sum_n (1+n^2)^s \, |f_n|^2$ in terms of Fourier coefficients;
- a **Banach manifold** is modelled on a complete normed space, for example $C^k(S^1)$ with $\|f\| = \sum_{j \le k} \sup \left|f^{(j)}\right|$;
- a **Fréchet manifold** is modelled on a complete space whose topology is defined by countably many seminorms. The main example is $C^\infty(S^1)$ with the seminorms $\sup \left|f^{(j)}\right|$, $j = 0, 1, 2, \ldots$ (Nash says "locally convex space").

The smooth maps we care about live naturally in the Fréchet world. The Hilbert and Sobolev versions are technically easier (the inverse function theorem holds), and physicists often work in them without saying so.

Key idea: the Lie algebra is the model space

For a finite-dimensional Lie group, the Lie algebra (the tangent space at the identity) is a copy of the model space $\mathbb{R}^n$. The same holds in infinite dimensions: if M is a Lie group modelled on E, its Lie algebra is E. For groups of maps $\mathrm{Map}(X, G)$ the model space is $\mathrm{Map}(X, \mathfrak{g})$.

6.1.3 The group $\mathrm{Map}(X, G)$

Let X be a compact manifold (space, or spacetime in Euclidean signature) and G a finite-dimensional Lie group. The set $\mathrm{Map}(X, G)$ of smooth maps $\alpha : X \to G$ is a group in the most obvious way: multiply the values.

Working definition 6.2 (pointwise product)

For $\alpha, \beta \in \mathrm{Map}(X, G)$ define

$$(\alpha \bullet \beta)(x) := \alpha(x)\,\beta(x), \qquad \alpha^{-1}(x) := \alpha(x)^{-1}, \qquad e(x) := 1 \text{ for all } x. \tag{6.1}$$

Operationally: to multiply two maps, go to each point x and multiply the two matrices sitting there. This is Nash's (6.1).

The group axioms hold because they hold at each point separately. The product and the inverse are smooth, so $\mathrm{Map}(X, G)$ is an (infinite-dimensional) Lie group.

Worked example 6.2: local phase rotations: $\mathrm{Map}(X, \mathrm{U}(1))$

Take $G = \mathrm{U}(1)$. An element is a function $\alpha(x) = e^{i\lambda(x)}$ with λ real (defined modulo 2π). Then

$$(\alpha \bullet \beta)(x) \overset{(1)}{=} e^{i\lambda(x)} e^{i\mu(x)} \overset{(2)}{=} e^{i(\lambda(x)+\mu(x))}.$$

(1) Definition (6.1) with $\beta = e^{i\mu}$.
(2) Phases commute and add.

The group is abelian and "is" the space of functions $\lambda(x)$, at least near the identity. These are the gauge transformations of electromagnetism. One parameter per point: infinitely many parameters.

Worked example 6.3: a non-abelian group of maps: $\mathrm{Map}(S^2, \mathrm{SU}(2))$

Let $\hat{n}(x)$ be a unit vector field on S^2 and $\phi(x)$ a function. Then $\alpha(x) = \exp\big(i\phi(x)\,\hat{n}(x)\cdot\vec{\sigma}/2\big) = \cos\frac{\phi}{2} + i\sin\frac{\phi}{2}\,\hat{n}\cdot\vec{\sigma}$ is an element of $\mathrm{Map}(S^2, \mathrm{SU}(2))$: at each point a rotation of spinors by angle $\phi(x)$ about the axis $\hat{n}(x)$. Two such maps multiply by multiplying 2×2 matrices at each point, and in general $\alpha \bullet \beta \neq \beta \bullet \alpha$ because the axes differ from point to point.

The Lie algebra of $\mathrm{Map}(X, G)$

The Lie algebra is the tangent space at the identity map e. A curve through e is a family α_t with $\alpha_0 = e$, and its velocity is the function $x \mapsto \frac{d}{dt}\alpha_t(x)\big|_{t=0}$, which at each point is a tangent vector to G at 1, that is an element of $\mathfrak{g}$. So

$$\mathrm{Lie}\ \mathrm{Map}(X, G) = \mathrm{Map}(X, \mathfrak{g}), \tag{6.2}$$

Nash's (6.2). The bracket is also pointwise. We check this by the usual trick: the commutator of two group elements close to the identity. Write $\alpha_s(x) = \exp(s\,\xi(x))$ and $\beta_t(x) = \exp(t\,\eta(x))$ with $\xi, \eta \in \mathrm{Map}(X, \mathfrak{g})$ (anti-Hermitian values, the mathematician's $\mathfrak{g}$). At each point x,

$$\begin{aligned}\big(\alpha_s \bullet \beta_t \bullet \alpha_s^{-1} \bullet \beta_t^{-1}\big)(x) &\overset{(1)}{=} e^{s\xi(x)} e^{t\eta(x)} e^{-s\xi(x)} e^{-t\eta(x)} \\ &\overset{(2)}{=} 1 + st\,\big(\xi(x)\eta(x) - \eta(x)\xi(x)\big) + O(s^2 t, st^2) \\ &\overset{(3)}{=} \exp\big(st\,[\xi(x), \eta(x)]\big) + O(s^2 t, st^2).\end{aligned}$$

(1) The group operations of $\mathrm{Map}(X, G)$ are pointwise, so we just evaluate at x.
(2) Expand each exponential to second order; the terms of order s^2 and t^2 alone cancel between $e^{s\xi}$ and $e^{-s\xi}$ (and similarly for η), leaving the mixed term.
(3) Resum to first order in st.

Hence the bracket on $\mathrm{Map}(X, \mathfrak{g})$ is

$$[\xi, \eta](x) = [\xi(x), \eta(x)]. \tag{6.3}$$

There are no derivatives of ξ or η anywhere: the Lie algebra of a group of maps is a *current algebra* without Schwinger terms, at least classically. The Schwinger terms of §6.2 are a quantum effect.

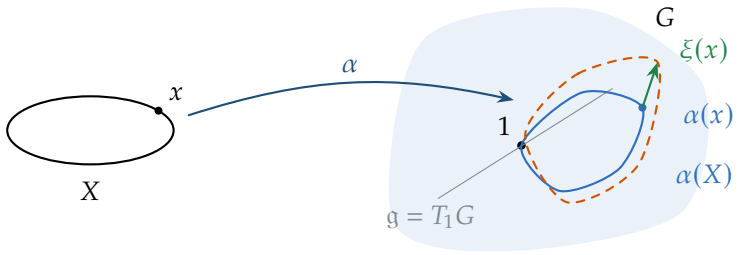


Figure 6.2: A point of $\mathrm{Map}(X, G)$ is a whole map α, drawn as its image (blue) in G. A nearby point (dashed orange) is obtained by moving each $\alpha(x)$ a little, in the direction of a tangent vector $\xi(x)$ (green). A tangent vector to $\mathrm{Map}(X, G)$ is therefore a $\mathfrak{g}$-valued function on X: the model space is $\mathrm{Map}(X, \mathfrak{g})$.

6.1.4 How many pieces does $\mathrm{Map}(X, G)$ have?

A finite-dimensional group like SU(2) is connected: any element can be joined to the identity by a path. A group of maps usually is not. The map $e^{i\theta} : S^1 \to \mathrm{U}(1)$ winds once around the circle, and no continuous deformation can unwind it. Physically, the connected components of the gauge group are the "large gauge transformations", and they are responsible for the θ-vacua of Yang–Mills theory (Chapter 8). Let us count them.

Based maps

Fix a base point $x_0 \in X$. A **based map** is one with $\alpha(x_0) = 1$. The based maps form a subgroup $\mathrm{Map}_0(X, G)$ (Nash introduced it in his Chapter 1). Two based maps are **based-homotopic** if one can be deformed into the other through based maps; the set of classes is written $[X, G]_0$.

The full group splits, as a space, into constants times based maps. For $g \in G$ let α_g denote the constant map $\alpha_g(x) = g$. Given any α, let $g = \alpha(x_0)$. Then

$$\alpha \overset{(1)}{=} \alpha_g \bullet (\alpha_g^{-1} \bullet \alpha),$$
$$(\alpha_g^{-1} \bullet \alpha)(x_0) \overset{(2)}{=} g^{-1}\alpha(x_0) \overset{(3)}{=} g^{-1}g = 1.$$

(1) Insert $\alpha_g \bullet \alpha_g^{-1} = e$ on the left.
(2) Pointwise product evaluated at the base point.
(3) g was defined as $\alpha(x_0)$.

So every α is uniquely "a constant g" times "a based map", and the assignment $\alpha \mapsto (g, \alpha_g^{-1} \bullet \alpha)$ is continuous with continuous inverse. As topological spaces (not as groups),

$$\mathrm{Map}(X, G) \;\cong\; G \times \mathrm{Map}_0(X, G), \tag{6.4}$$

Nash's (6.3). Connected components of a product are pairs of components, so

$$\begin{aligned}\pi_0\big(\mathrm{Map}(X, G)\big) &\overset{(1)}{=} \pi_0(G) \times \pi_0\big(\mathrm{Map}_0(X, G)\big),\\ \pi_0\big(\mathrm{Map}_0(X, G)\big) &\overset{(2)}{=} [S^0, \mathrm{Map}_0(X, G)]_0\\ &\overset{(3)}{=} [S^0 \wedge X, G]_0\\ &\overset{(4)}{=} [X, G]_0.\end{aligned} \tag{6.5}$$

(1) The splitting (6.4). (Nash writes $\oplus$; for the abelian groups that occur below the two notations agree.)

(2) S^0 is two points; a based map from S^0 picks out one point of the target, and based-homotopy classes of such picks are exactly path components.
(3) The adjunction $[A, \mathrm{Map}_0(X, G)]_0 = [A \wedge X, G]_0$: a based map from A into based maps is the same as a based map of two variables, which factors through the smash product $A \wedge X = A \times X/(A \vee X)$ (Nash's (1.18)).
(4) $S^0 \wedge X$ is X itself (smashing with two points, one of which is the base point, changes nothing).

Maps from spheres

For $X = S^m$, based-homotopy classes of maps $S^m \to G$ are by definition the homotopy group $\pi_m(G)$. For $G = \mathrm{U}(n)$, which is connected ($\pi_0(\mathrm{U}(n)) = 0$), we get Nash's (6.5):

$$\pi_0\big(\mathrm{Map}(S^m, \mathrm{U}(n))\big) = \pi_m(\mathrm{U}(n)). \tag{6.6}$$

The standard values (for n large enough, the "stable range" $n > m/2$) are $\pi_m(\mathrm{U}(n)) = \mathbb{Z}$ for m odd and 0 for m even (Bott periodicity). In particular $\pi_3(\mathrm{U}(n)) = \mathbb{Z}$ for $n \geq 2$, so $\mathrm{Map}(S^3, \mathrm{U}(n))$ has infinitely many components, one for each integer. Two explicit examples make the integer visible.

Worked example 6.4: winding number of a loop in $\mathrm{U}(1)$

Take $X = S^1$, $G = \mathrm{U}(1)$ and $\alpha_w(\theta) = e^{iw\theta}$ with $w \in \mathbb{Z}$ (an integer is needed for α_w to be single-valued). Define $W[\alpha] = \frac{1}{2\pi i} \oint \alpha^{-1} \mathrm{d}\alpha$. Then

$$W[\alpha_w] \overset{(1)}{=} \frac{1}{2\pi i} \int_0^{2\pi} e^{-iw\theta}\,(iw)\,e^{iw\theta}\,\mathrm{d}\theta$$
$$\overset{(2)}{=} \frac{1}{2\pi i}\,(iw)(2\pi) = w.$$

(1) $\alpha^{-1}\mathrm{d}\alpha = e^{-iw\theta}\,\mathrm{d}(e^{iw\theta})$ and $\mathrm{d}(e^{iw\theta}) = iw\,e^{iw\theta}\mathrm{d}\theta$.
(2) The integrand is the constant iw.

For a general $\alpha = e^{i\lambda(\theta)}$, $W = \frac{1}{2\pi}\big(\lambda(2\pi) - \lambda(0)\big)$, the number of times the phase goes round. Under a small deformation $\lambda \to \lambda + \delta\lambda$ with $\delta\lambda$ periodic, W changes by $\frac{1}{2\pi}(\delta\lambda(2\pi) - \delta\lambda(0)) = 0$. An integer that cannot change continuously labels the components: $\pi_0(\mathrm{Map}(S^1, \mathrm{U}(1))) = \pi_1(\mathrm{U}(1)) = \mathbb{Z}$.

Worked example 6.5: the degree of a map $S^3 \to \mathrm{SU}(2)$

SU(2) is itself a three-sphere: every element can be written $g = x_4 \mathbb{1} + i\,\vec{x}\cdot\vec{\sigma}$ with $x_4^2 + \big|\vec{x}\big|^2 = 1$. So the identity map $g(x) = x_4 + i\vec{x}\cdot\vec{\sigma}$ is an element of $\mathrm{Map}(S^3, \mathrm{SU}(2))$. Its integer invariant is the degree

$$w[g] = \frac{1}{24\pi^2} \int_{S^3} \mathrm{tr}\big(g^{-1}\mathrm{d}g\big)^3.$$

We evaluate the integrand at the identity element $x = (0, 0, 0, 1)$, where $\vec{x}$ are good local coordinates:

$$g^{-1}\mathrm{d}g\big|_{g=1} \overset{(1)}{=} i\sigma_j\,\mathrm{d}x^j,$$
$$\mathrm{tr}\big(g^{-1}\mathrm{d}g\big)^3\big|_{g=1} \overset{(2)}{=} i^3\,\mathrm{tr}(\sigma_j\sigma_k\sigma_l)\,\mathrm{d}x^j \wedge \mathrm{d}x^k \wedge \mathrm{d}x^l$$

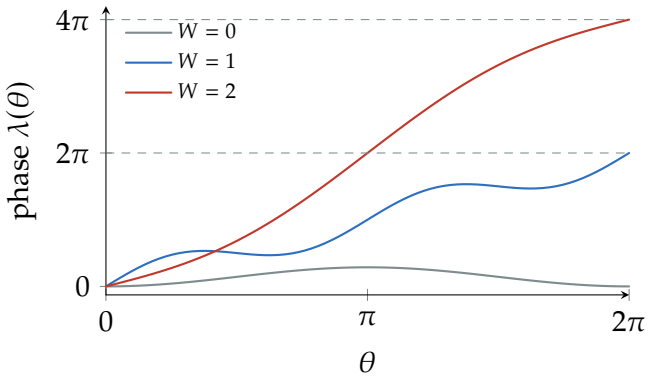


Figure 6.3: Three loops in U(1), drawn as the lifted phase $\lambda(\theta)$ of $\alpha = e^{i\lambda}$. The endpoint $\lambda(2\pi)$ must equal $\lambda(0)$ plus a multiple $2\pi W$. Wiggling a curve (a small deformation) cannot move its endpoint from one dashed line to another, so the winding number W labels the connected components of $LU(1)$.

$$\overset{(3)}{=} (-i)(2i\,\varepsilon_{jkl})\,\mathrm{d}x^j \wedge \mathrm{d}x^k \wedge \mathrm{d}x^l$$
$$\overset{(4)}{=} 12\,\mathrm{d}x^1 \wedge \mathrm{d}x^2 \wedge \mathrm{d}x^3.$$

(1) Near $g = 1$, $g \approx 1 + i\vec{x}\cdot\vec{\sigma}$, so $\mathrm{d}g = i\sigma_j \mathrm{d}x^j$ and $g^{-1} \approx 1$.
(2) Cube the matrix-valued 1-form; the wedge makes the coefficient antisymmetric in j, k, l.
(3) $\mathrm{tr}(\sigma_j\sigma_k\sigma_l) = 2i\varepsilon_{jkl}$.
(4) $\varepsilon_{jkl}\mathrm{d}x^j \wedge \mathrm{d}x^k \wedge \mathrm{d}x^l = 6\,\mathrm{d}x^1 \wedge \mathrm{d}x^2 \wedge \mathrm{d}x^3$ and $(-i)(2i) = 2$.

The form $\mathrm{tr}(g^{-1}\mathrm{d}g)^3$ is invariant under left multiplication $g \mapsto hg$ (since $(hg)^{-1}\mathrm{d}(hg) = g^{-1}\mathrm{d}g$), so it equals 12×(the volume form of the unit S^3) everywhere. The unit three-sphere has volume $2\pi^2$, hence $w[g] = \frac{1}{24\pi^2}\cdot 12 \cdot 2\pi^2 = 1$. The identity map has degree one, and g^w (pointwise power) has degree w. This is the integer that labels the components of $\mathrm{Map}(S^3, \mathrm{SU}(2))$, and, in Chapter 8, the instanton number.

General X and K-theory

For $G = \mathrm{U}(n)$ and any compact X, Nash rewrites $[X, \mathrm{U}(n)]_0$ in the language of vector bundles. The chain is his (6.6):

$$\begin{aligned}\pi_0\big(\mathrm{Map}(X, \mathrm{U}(n))\big) &\overset{(1)}{=} [S^0 \wedge X, \mathrm{U}(n)]_0 \\ &\overset{(2)}{=} [S^0 \wedge X, \Omega B\mathrm{U}(n)]_0 \\ &\overset{(3)}{=} [SX, B\mathrm{U}(n)]_0 \\ &\overset{(4)}{=} \tilde{K}^{-1}(X) \qquad (n > \tfrac{1}{2}\dim X).\end{aligned} \tag{6.7}$$

(1) Equation (6.5), together with $\pi_0(\mathrm{U}(n)) = 0$.
(2) The classifying space $B\mathrm{U}(n)$ has the property that its based loop space $\Omega B\mathrm{U}(n)$ is homotopy equivalent to $\mathrm{U}(n)$. (A loop in $B\mathrm{U}(n)$ is a way of going round the base of the universal bundle, and its holonomy is an element of $\mathrm{U}(n)$.)
(3) Loop–suspension adjunction: a map from Y into loops of Z is a map from the suspension SY into Z. With $Y = X$ this is $[SX, B\mathrm{U}(n)]_0$.

(4) Maps $SX \to BU(n)$ classify rank-n vector bundles on the suspension. Such a bundle is built by gluing trivial bundles on the two cones of SX along X with a clutching function $X \to U(n)$. In the stable range these bundles, up to isomorphism and adding trivial bundles, form $\tilde{K}(SX)$, which is by definition $\tilde{K}^{-1}(X)$ (Chapter 1).

So the components of the gauge group of a trivial $U(n)$ bundle are counted by odd K-theory. Nash's footnote adds that $\tilde{K}^{-1}(X)$ is a sum of copies of $\mathbb{Z}$ and a finite (torsion) group T, and that the number of copies of $\mathbb{Z}$ is $b_1 + b_3 + b_5 + \cdots$, the sum of the odd Betti numbers of X. (This is the statement that the Chern character identifies $K^{-1}(X) \otimes \mathbb{Q}$ with the odd rational cohomology.) Checks:

X	odd Betti numbers	$\pi_0 \operatorname{Map}(X, U(n))$	the integers are
S^1	$b_1 = 1$	$\mathbb{Z}$	winding of $\det \alpha$ (Example 6.4)
S^2	none	0	$\operatorname{Map}(S^2, U(n))$ connected
S^3	$b_3 = 1$	$\mathbb{Z}$	degree (Example 6.5)
T^2	$b_1 = 2$	$\mathbb{Z}^2$	windings of $\det \alpha$ round the two cycles
T^3	$b_1 = 3,\ b_3 = 1$	$\mathbb{Z}^4$	three windings and one degree

6.1.5 The loop group LG

The simplest nontrivial choice of X is a circle. This case is special enough to have its own name and its own beautiful theory.

Working definition 6.3 (loop group)

The **loop group** of G is

$$LG := \operatorname{Map}(S^1, G), \tag{6.8}$$

smooth loops $\gamma : S^1 \to G$, $\theta \mapsto \gamma(\theta)$, with the pointwise product [$(\gamma_1\gamma_2)(\theta) = \gamma_1(\theta)\gamma_2(\theta)$, multiply the values at each θ] (Nash's (6.7)). *Physically:* a gauge transformation [a choice of group element $\gamma(\theta) \in G$ at each point θ of space] at fixed time in a $(1+1)$-dimensional theory whose space is a circle of circumference 2π.

By (6.5), $\pi_0(LG) = \pi_0(G) \times \pi_1(G)$. For a connected G, the components of LG are the elements of $\pi_1(G)$:

- $G = SU(n), Sp(n), E_8$ (simply connected): LG is connected, as Nash notes;
- $G = U(1)$: $\pi_1 = \mathbb{Z}$, components labelled by the winding number;
- $G = SO(3)$: $\pi_1 = \mathbb{Z}_2$. The loop "rotate by 2π about the z-axis as θ goes round" is not contractible, but going round twice (rotation by 4π) is. This is the same $\mathbb{Z}_2$ that makes spinors change sign under a 2π rotation (Chapter T4).

The loop algebra and its Fourier basis

By (6.2) the Lie algebra of LG is

$$\operatorname{Lie} LG = \operatorname{Map}(S^1, \mathfrak{g}), \tag{6.9}$$

Nash's (6.8), the **loop algebra**. Its elements are $\mathfrak{g}$-valued periodic functions $\xi(\theta)$, so we can Fourier expand them. In the physicists' convention (Hermitian T^a) write

$$T^a_n := e^{in\theta}\, T^a, \qquad n \in \mathbb{Z}, \quad a = 1, \ldots, \dim G, \tag{6.10}$$

which is Nash's (6.10). A general element of the (complexified) loop algebra is $\xi(\theta) = \sum_{n,a} \xi_{a,n} T^a_n$, and it is Hermitian at every θ precisely when $\xi_{a,-n} = \overline{\xi_{a,n}}$. The generators T^a_0 are the *constant* loops; they span a copy of $\mathfrak{g}$. The loops with $n \neq 0$ are the "oscillating" directions, the based part of the splitting (6.4).

The bracket of basis elements follows from (6.3):

$$\begin{aligned}[T^a_m, T^b_n](\theta) &\overset{(1)}{=} [e^{im\theta}T^a,\ e^{in\theta}T^b] \\ &\overset{(2)}{=} e^{i(m+n)\theta}\,[T^a, T^b] \\ &\overset{(3)}{=} i f^{abc}\, e^{i(m+n)\theta} T^c \overset{(4)}{=} i f^{abc}\, T^c_{m+n}(\theta). \end{aligned} \tag{6.11}$$

(1) The bracket is pointwise, so we compute at each θ.
(2) The functions $e^{im\theta}$ are numbers at each θ and pull out of the commutator.
(3) Commutation relations of $\mathfrak{g}$.
(4) Definition (6.10) with index $m + n$.

So the mode numbers simply add. This is the **loop algebra** in its mode form. It is "$\mathfrak{g}$ tensored with Laurent polynomials in $z = e^{i\theta}$": $\mathfrak{g} \otimes \mathbb{C}[z, z^{-1}]$.

Worked example 6.6: the SU(2) loop algebra with Pauli matrices

Let $T^a = \sigma^a/2$, so $f^{abc} = \varepsilon^{abc}$. Then $T^1_m = \frac{1}{2}e^{im\theta}\begin{pmatrix}0&1\\1&0\end{pmatrix}$ and $T^2_n = \frac{1}{2}e^{in\theta}\begin{pmatrix}0&-i\\i&0\end{pmatrix}$, and

$$\begin{aligned}[T^1_m, T^2_n] &\overset{(1)}{=} \tfrac{1}{4}e^{i(m+n)\theta}\left[\begin{pmatrix}0&1\\1&0\end{pmatrix}\begin{pmatrix}0&-i\\i&0\end{pmatrix} - \begin{pmatrix}0&-i\\i&0\end{pmatrix}\begin{pmatrix}0&1\\1&0\end{pmatrix}\right] \\ &\overset{(2)}{=} \tfrac{1}{4}e^{i(m+n)\theta}\left[\begin{pmatrix}i&0\\0&-i\end{pmatrix} - \begin{pmatrix}-i&0\\0&i\end{pmatrix}\right] \\ &\overset{(3)}{=} i\,\tfrac{1}{2}e^{i(m+n)\theta}\begin{pmatrix}1&0\\0&-1\end{pmatrix} = i\,T^3_{m+n}. \end{aligned}$$

(1) Pointwise bracket; the phases multiply.
(2) Matrix multiplication.
(3) The difference is $2i\sigma^3$, and $\frac{1}{4}\cdot 2i = \frac{i}{2}$.

This is (6.11) with $f^{123} = 1$. Note what the modes look like as matrices: T^1_1 is a 2×2 matrix whose entries rotate once round the circle as θ goes from 0 to 2π.

The affine Kac–Moody algebra (first look)

Nash now makes a statement that the rest of the chapter explains: *all the interesting representations of* $\mathrm{Map}(S^1, \mathfrak{g})$ *are really representations of a central extension* $\widetilde{L\mathfrak{g}}$. In the extended algebra the modes obey

$$\boxed{[T^a_m, T^b_n] = i f^{abc}\, T^c_{m+n} + k\, m\, \delta^{ab}\, \delta_{m+n,0}, \qquad m, n \in \mathbb{Z}} \tag{6.12}$$

Nash's (6.9). Here k is a **central element**: it commutes with every T^a_n, and in an irreducible representation it is a number, the **level** (up to normalisation, see Section 6.3.9). This is the **untwisted affine Kac–Moody algebra** $\hat{\mathfrak{g}}$. Three things to notice at once:

- the extra term is present only when $m + n = 0$, and it is proportional to m. So it vanishes when $m = n = 0$: the constant loops T^a_0 still close into an ordinary copy of $\mathfrak{g}$;

- it is antisymmetric, as a commutator must be: swapping (a, m) with (b, n) and using $\delta_{m+n,0}$ turns $km\delta^{ab}$ into $kn\delta^{ab} = -km\delta^{ab}$;
- it is a c-number. In a quantum field theory it is the famous *Schwinger term* (Section 6.1.11).

In §6.2 we show that (6.12) is consistent (it satisfies the Jacobi identity), that the extra term cannot be removed by redefining the generators, and that it appears automatically when the algebra is realised on a Fock space with a Dirac sea.

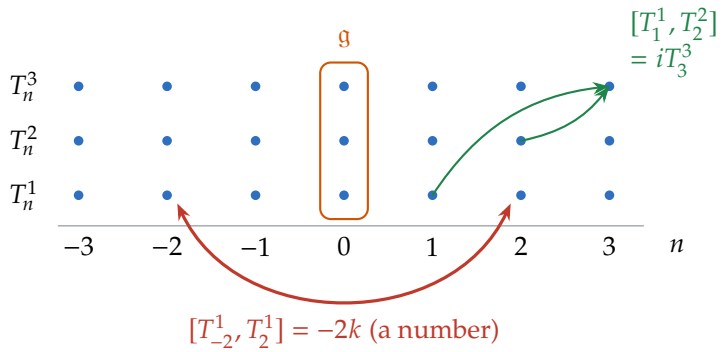


Figure 6.4: The generators of the loop algebra of SU(2) as a lattice: one column per Fourier mode n, one row per direction a in $\mathfrak{g}$. Brackets add mode numbers (green). The column $n = 0$ (orange box) is a copy of $\mathfrak{g}$. In the affine Kac–Moody algebra, a bracket that returns to total mode number zero from a pair $\pm n$ in the same row also produces a c-number (red): the central term $km\,\delta^{ab}\delta_{m+n,0}$.

Reading Nash (pp. 170–174, (6.1)–(6.10))

Nash's treatment of $\mathrm{Map}(X, G)$ and LG is two pages. The formulae (6.1), (6.2), (6.7), (6.8) and (6.10) are the definitions above. His (6.3)–(6.6) are our (6.4)–(6.7); the "using 1.18" refers to the smash-product adjunction of his Chapter 1, and "cf. p. 73 and pp. 74, 76" refers to $\Omega B\mathrm{U}(n) \simeq \mathrm{U}(n)$ and to the definition of $\tilde{K}^{-1}$ by suspension. Nash writes $\pi_3(\mathrm{U}(n)) = \mathbb{Z}$, $n > 1$. The Kac–Moody relations (6.9) are quoted here and justified only in §2; we have followed him in using the same symbol k for the central element and its value. The references he gives for the mathematics and the physics (Pressley and Segal; Kac; Goddard and Olive) are still the standard ones.

Where loop groups turn up

Nash lists the places where loop groups and their Kac–Moody algebras have proved their worth. It is worth pausing on the list, because it explains why a physicist should care about such an abstract-looking object.

- *Two-dimensional statistical mechanics and conformal field theory* (Chapter 11). At a critical point a two-dimensional system with an internal symmetry G (for instance a chain of quantum spins with SU(2) symmetry) has conserved currents $j^a(\theta)$ on a spatial circle. Their Fourier modes obey (6.12). The level k and the central charge of §6.2 are then measurable numbers: they fix the finite-size corrections to the ground-state energy and the critical exponents.
- *Current algebra* (Section 6.1.11). The equal-time commutators of currents, first studied for the strong interactions, are loop-algebra relations with a Schwinger term.
- *String theory* (Chapter 9). A closed string is a loop; a string moving on a group manifold G (the Wess–Zumino–Witten model) has the loop group LG as its symmetry group.

- *Knot invariants* (Chapter 12). The Jones polynomial comes out of the representation theory of the affine SU(2) algebra at integer level, through Chern–Simons theory.

- *Pure mathematics*. Kac–Moody algebras are tied to completely integrable systems (the KdV equation has the Virasoro algebra as its Poisson bracket algebra), to modular forms (characters of Kac–Moody representations are modular forms, which is why string partition functions are modular invariant), and to the Monster, the largest sporadic finite simple group, whose representations are organised by a $c = 24$ conformal field theory.

6.1.6 The diffeomorphism group $\mathrm{Diff}(X)$

Our second family of infinite-dimensional groups has a different product. In $\mathrm{Map}(X, G)$ we multiplied the *values* of two maps. Now the elements are maps of a space to itself, and we multiply them by doing one after the other.

Working definition 6.4 (diffeomorphism group)

A **diffeomorphism** of a manifold X is a smooth map $\alpha : X \to X$ with a smooth inverse. The diffeomorphisms of X form a group $\mathrm{Diff}(X)$ with product *composition*:

$$\alpha \bullet \beta := \alpha \circ \beta, \qquad (\alpha \bullet \beta)(x) = \alpha\big(\beta(x)\big), \tag{6.13}$$

Nash's (6.11). The identity element is the identity map and the inverse is the inverse map. *Physically:* $\mathrm{Diff}(X)$ is the group of general coordinate transformations, the gauge group of general relativity [moving all fields along by the same diffeomorphism does not change the physics]. An active diffeomorphism moves every point of space to a new place.

Note the difference with $\mathrm{Map}(X, G)$. There, the product $\alpha \bullet \beta$ at x only needs the values at x. Here it needs the value of α at the *different* point $\beta(x)$. This nonlocality is what makes $\mathrm{Diff}(X)$ much harder than $\mathrm{Map}(X, G)$, as we will see when we discuss analyticity in Section 6.1.12.

Worked example 6.7: diffeomorphisms of the line and the circle

On $X = \mathbb{R}$, the maps $\alpha_a(x) = x + a$ (translations) and $\delta_\lambda(x) = e^\lambda x$ (dilatations) are diffeomorphisms. Compose them:

$$(\alpha_a \bullet \delta_\lambda)(x) \overset{(1)}{=} \alpha_a(e^\lambda x) = e^\lambda x + a,$$
$$(\delta_\lambda \bullet \alpha_a)(x) \overset{(2)}{=} \delta_\lambda(x + a) = e^\lambda x + e^\lambda a.$$

(1) Definition (6.13): first dilate, then translate.
(2) First translate, then dilate.

The two differ unless $a = 0$ or $\lambda = 0$: the group is non-abelian already for these two one-parameter families. The map $x \mapsto x^3$ is a smooth bijection of $\mathbb{R}$ but not a diffeomorphism, because its inverse $x \mapsto x^{1/3}$ is not differentiable at 0.
On $X = S^1$ with angle θ, a diffeomorphism is a map $\theta \mapsto \alpha(\theta)$ with $\alpha(\theta + 2\pi) = \alpha(\theta) \pm 2\pi$ and $\alpha'(\theta)$ never zero. The rigid rotations $\theta \mapsto \theta + a$ are the simplest. A less rigid example is $\alpha_\varepsilon(\theta) = \theta + \varepsilon \sin\theta$, which is a diffeomorphism for $|\varepsilon| < 1$ because then $\alpha'_\varepsilon = 1 + \varepsilon\cos\theta > 0$. It squeezes the points near $\theta = \pi$ together and spreads out those near $\theta = 0$.

The Lie algebra: vector fields and their flows

What is a diffeomorphism "close to the identity"? It moves every point a little. Record the little displacement at each point: that is a vector at each point, a *vector field*. Conversely, a vector field generates diffeomorphisms by flowing along it.

Working definition 6.5 (flow of a vector field)

Let $V = \sum_i v^i(x)\,\frac{\partial}{\partial x^i}$ be a smooth vector field on X. Its **integral curves** are the solutions of $\dot{x}^i(t) = v^i(x(t))$. The **flow** of V is the family of maps $\alpha_t : X \to X$ that sends each point x to the point reached after time t along the integral curve starting at x. On a compact X [closed and bounded, so no integral curve can run off to infinity in finite time] the flow exists for all t, each α_t is a diffeomorphism, and $\alpha_t \bullet \alpha_s = \alpha_{t+s}$ [flowing for time s and then for time t is the same as flowing for time $t+s$]. We write, with Nash's (6.13),

$$V = \sum_i v^i \frac{\partial}{\partial x^i}, \qquad \alpha_t = \exp[tV]. \tag{6.14}$$

Operationally: to find $\alpha_t(x)$, solve the ODE $\dot{x} = v(x)$ with initial condition x and read off the solution at time t. To first order, $\alpha_t(x) = x + t\,v(x) + O(t^2)$.

Flows give a one-parameter subgroup $t \mapsto \alpha_t$ of $\mathrm{Diff}(X)$ for every vector field, and every smooth one-parameter subgroup arises this way (differentiate it at $t = 0$ to recover V). So the tangent space at the identity is the space of vector fields:

$$\text{Lie Diff}(X) = \text{vect}(X), \tag{6.15}$$

Nash's (6.12). The bracket is the Lie bracket of vector fields,

$$[V, W] = \sum_{i,j} \left(v^j \frac{\partial w^i}{\partial x^j} - w^j \frac{\partial v^i}{\partial x^j} \right) \frac{\partial}{\partial x^i}, \tag{6.16}$$

which is just the commutator $VW - WV$ of the two first-order differential operators (the second derivatives cancel), Chapter T1.

Worked example 6.8: flows on the line: translations, dilatations and a blow-up

We solve $\dot{x} = v(x)$ for three vector fields on $\mathbb{R}$.

$$V_{-1} = \frac{\partial}{\partial x} : \quad \dot{x} = 1 \qquad \Rightarrow \quad \alpha_t(x) \overset{(1)}{=} x + t,$$
$$V_0 = x\frac{\partial}{\partial x} : \quad \dot{x} = x \qquad \Rightarrow \quad \alpha_t(x) \overset{(2)}{=} e^t x,$$
$$V_1 = x^2\frac{\partial}{\partial x} : \quad \dot{x} = x^2 \qquad \Rightarrow \quad \alpha_t(x) \overset{(3)}{=} \frac{x}{1 - tx}.$$

(1) Constant velocity: a translation.

(2) Exponential growth: a dilatation.

(3) Separate variables: $\mathrm{d}x/x^2 = \mathrm{d}t$ gives $-1/x(t) + 1/x = t$, so $x(t) = x/(1 - tx)$. Check: $\frac{d}{dt}\frac{x}{1-tx} = \frac{x^2}{(1-tx)^2} = x(t)^2$.

The third flow sends the point $x = 1/t$ to infinity in finite time, so on $\mathbb{R}$ it is not a diffeomorphism for any $t \neq 0$: $\mathbb{R}$ is not compact, and the vector field $x^2\partial_x$ is not complete. If we add the point at infinity and work on the circle $\mathbb{RP}^1 = \mathbb{R} \cup \{\infty\}$, all three flows are perfectly good diffeomorphisms: they are the Möbius maps $x \mapsto (ax + b)/(cx + d)$. The

brackets are

$$[V_m, V_n] \overset{(4)}{=} x^{m+1}\,(n+1)x^n \frac{\partial}{\partial x} - x^{n+1}(m+1)x^m \frac{\partial}{\partial x}$$
$$= (n-m)\,V_{m+n}, \qquad m,n \in \{-1,0,1\},$$

(4) Formula (6.16) with $V_m = x^{m+1}\partial_x$.

a three-dimensional Lie algebra, which is $\mathfrak{sl}(2,\mathbb{R})$. We will meet exactly this subalgebra inside the Witt algebra below.

Watch out: the sign of the bracket on $\mathrm{vect}(X)$

The Lie algebra bracket obtained from the group commutator of $\mathrm{Diff}(X)$ is *minus* the vector field bracket (6.16). Let us see this in one dimension. Take flows $\alpha_t(x) = x + tv(x) + O(t^2)$ and $\beta_s(x) = x + sw(x) + O(s^2)$. Then, keeping terms up to order ts,

$$(\alpha_t \bullet \beta_s)(x) \overset{(1)}{=} \beta_s(x) + t\,v\big(\beta_s(x)\big) \overset{(2)}{=} x + sw + tv + ts\,v'w,$$
$$(\beta_s \bullet \alpha_t)(x) \overset{(3)}{=} x + tv + sw + ts\,w'v,$$
$$\big(\alpha_t \bullet \beta_s \bullet \alpha_t^{-1} \bullet \beta_s^{-1}\big)(x) \overset{(4)}{=} x + ts\,(v'w - w'v) + \cdots .$$

(1) Apply α_t to the point $\beta_s(x)$.
(2) Taylor expand $v(x+sw) = v + sv'w + \cdots$.
(3) The same with the roles exchanged.
(4) The group commutator is $(\alpha \bullet \beta) \bullet (\beta \bullet \alpha)^{-1}$, which moves points by the difference of the two displacements.

The displacement $v'w - w'v$ is the coefficient of $-[V,W]$ in (6.16). The sign is harmless: the map $V \mapsto -V$ converts one bracket into the other. It also disappears in the way physicists use $\mathrm{Diff}(X)$, namely through its action on functions $f \mapsto f \circ \alpha^{-1}$, whose generator is the differential operator $-V$, and then all commutators are commutators of differential operators. Like Nash, we always use the commutator of differential operators.

6.1.7 Diffeomorphisms of the circle and the Witt algebra

The case $X = S^1$ is to $\mathrm{Diff}(X)$ what the loop group is to $\mathrm{Map}(X,G)$: the simplest case, and the most important one for physics. A closed string at fixed time is a map from a circle into spacetime, and reparametrising the string, $\sigma \mapsto \alpha(\sigma)$, does not change the physical string. So $\mathrm{Diff}(S^1)$ is a gauge group of string theory.

Components

$\mathrm{Diff}(S^1)$ has two components: the orientation-preserving maps $\mathrm{Diff}^+(S^1)$, with $\alpha' > 0$, and the orientation-reversing ones, such as the reflection $\theta \mapsto -\theta$. $\mathrm{Diff}^+(S^1)$ is connected, and in fact it can be deformed onto its subgroup of rigid rotations: the path $\alpha_s(\theta) = (1-s)\alpha(\theta) + s(\theta + \alpha(0))$, $s \in [0,1]$, stays inside $\mathrm{Diff}^+(S^1)$ (a convex combination of two increasing functions is increasing) and ends at a rotation. So, as a space, $\mathrm{Diff}^+(S^1)$ looks like a circle times a contractible piece, and $\pi_1(\mathrm{Diff}^+(S^1)) = \mathbb{Z}$.

The Witt basis

A vector field on S^1 is $v(\theta)\,\frac{d}{d\theta}$ with v periodic. Fourier expand v. The complex basis vector fields, in Nash's normalisation (6.15), are

$$L_n := ie^{in\theta}\,\frac{d}{d\theta}, \qquad n \in \mathbb{Z}. \tag{6.17}$$

It is often convenient to use the complex coordinate $z = e^{i\theta}$ on the unit circle. Then

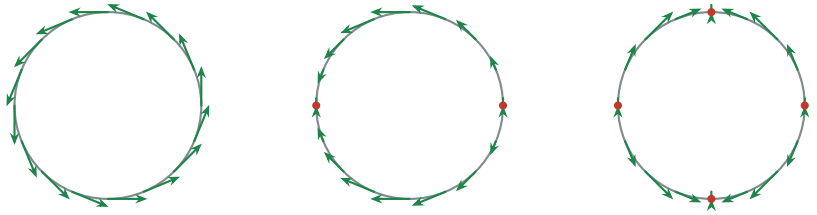
$$\begin{aligned} -z^{n+1}\frac{d}{dz} &\overset{(1)}{=} -z^{n+1}\,\frac{d\theta}{dz}\,\frac{d}{d\theta} \\ &\overset{(2)}{=} -e^{i(n+1)\theta}\,\frac{1}{ie^{i\theta}}\,\frac{d}{d\theta} \\ &\overset{(3)}{=} ie^{in\theta}\,\frac{d}{d\theta} = L_n. \end{aligned} \tag{6.18}$$

(1) Chain rule on the circle $|z| = 1$.
(2) $dz/d\theta = ie^{i\theta}$, so $d\theta/dz = 1/(ie^{i\theta})$.
(3) $-1/i = i$ and $e^{i(n+1)\theta}e^{-i\theta} = e^{in\theta}$.

So $L_n = -z^{n+1}\frac{d}{dz} = ie^{in\theta}\frac{d}{d\theta}$, which is Nash's (6.15). Real vector fields are real combinations:

$$\frac{d}{d\theta} = -iL_0, \qquad \cos n\theta\,\frac{d}{d\theta} = -\tfrac{i}{2}(L_n + L_{-n}), \qquad \sin n\theta\,\frac{d}{d\theta} = -\tfrac{1}{2}(L_n - L_{-n}). \tag{6.19}$$

(Check the last one: $L_n - L_{-n} = i(e^{in\theta} - e^{-in\theta})\frac{d}{d\theta} = i\cdot 2i\sin n\theta\,\frac{d}{d\theta} = -2\sin n\theta\,\frac{d}{d\theta}$.)



$\frac{d}{d\theta}$ $(n=0)$ $\qquad$ $\sin\theta\,\frac{d}{d\theta}$ $(n=1)$ $\qquad$ $\sin 2\theta\,\frac{d}{d\theta}$ $(n=2)$

Figure 6.5: Real vector fields on the circle, drawn as tangent arrows (green) whose length is $v(\theta)$. Left: the rotation generator $-iL_0$, the same at every point. Middle: $\sin\theta\,\frac{d}{d\theta} = -\frac{1}{2}(L_1 - L_{-1})$, which vanishes at $\theta = 0, \pi$ (red) and pushes points from 0 towards π. Right: the $n = 2$ field has four zeros and pushes points towards two of them. In general the mode-n field has $2n$ zeros.

Worked example 6.9: the flow of $\sin\theta\,\frac{d}{d\theta}$ is a Möbius map

Solve $\dot\theta = \sin\theta$. Separate variables:

$$t \overset{(1)}{=} \int_{\theta_0}^{\theta(t)} \frac{\mathrm{d}\theta}{\sin\theta} \overset{(2)}{=} \Big[\ln\tan\tfrac{\theta}{2}\Big]_{\theta_0}^{\theta(t)},$$

$$\tan\frac{\theta(t)}{2} \overset{(3)}{=} e^{t}\,\tan\frac{\theta_0}{2}.$$

(1) $\mathrm{d}t = \mathrm{d}\theta/\sin\theta$ along an integral curve.
(2) $\frac{d}{d\theta}\ln\tan\frac{\theta}{2} = \frac{1}{2}\frac{\sec^2(\theta/2)}{\tan(\theta/2)} = \frac{1}{2\sin(\theta/2)\cos(\theta/2)} = \frac{1}{\sin\theta}$.
(3) Exponentiate.

In the variable $u = \tan(\theta/2)$ (the stereographic coordinate of the circle) the flow is the

dilatation $u \mapsto e^t u$. Points with $0 < \theta_0 < \pi$ approach $\theta = \pi$ ($u \to \infty$) as $t \to \infty$, and the fixed points $\theta = 0, \pi$ stay put, exactly as in Figure 6.5. In the variable $z = e^{i\theta}$ the same map is

$$z \;\mapsto\; \frac{\cosh\frac{t}{2}\, z - \sinh\frac{t}{2}}{-\sinh\frac{t}{2}\, z + \cosh\frac{t}{2}},$$

a Möbius transformation preserving the unit circle. (Check: $u = \tan\frac{\theta}{2} = -i\frac{z-1}{z+1}$, and a dilatation of u is a fractional linear map of z.) This is the circle version of the flow of V_0 in Example 6.8.

The Witt algebra

Now compute the bracket of two basis fields. Using the form $L_n = -z^{n+1}\frac{d}{dz}$ and acting on a test function $f(z)$,

$$\begin{aligned}
[L_m, L_n]f &\overset{(1)}{=} z^{m+1}\frac{d}{dz}\Big(z^{n+1}f'\Big) - z^{n+1}\frac{d}{dz}\Big(z^{m+1}f'\Big) \\
&\overset{(2)}{=} (n+1)z^{m+n+1}f' + z^{m+n+2}f'' - (m+1)z^{m+n+1}f' - z^{m+n+2}f'' \\
&\overset{(3)}{=} (n-m)\,z^{m+n+1}f' \overset{(4)}{=} (m-n)\,L_{m+n}f\,.
\end{aligned} \tag{6.20}$$

(1) $L_mL_nf = (-z^{m+1}\partial_z)(-z^{n+1}\partial_z f)$: the two minus signs cancel.
(2) Product rule.
(3) The second-derivative terms cancel; this is why the commutator of two vector fields is again a vector field.
(4) $z^{m+n+1}f' = -L_{m+n}f$.

So

$$\boxed{[L_m, L_n] = (m-n)\,L_{m+n}}$$

This is the **Witt algebra**, the complexified Lie algebra of $\mathrm{Diff}^+(S^1)$. It is the Virasoro algebra with $c = 0$. Like the loop algebra, it is graded by the mode number: brackets add mode numbers.

Worked example 6.10: the same bracket in the angle variable

It is reassuring to redo (6.20) with $L_n = ie^{in\theta}\partial_\theta$, since this is the form used in the Sugawara construction later. For a test function $f(\theta)$,

$$\begin{aligned}
L_mL_nf &\overset{(1)}{=} ie^{im\theta}\partial_\theta\big(ie^{in\theta}f'\big) \overset{(2)}{=} -e^{im\theta}\big(ine^{in\theta}f' + e^{in\theta}f''\big), \\
[L_m, L_n]f &\overset{(3)}{=} -in\,e^{i(m+n)\theta}f' + im\,e^{i(m+n)\theta}f' \overset{(4)}{=} (m-n)\,ie^{i(m+n)\theta}f' = (m-n)L_{m+n}f\,.
\end{aligned}$$

(1) Definition (6.17), applied twice.
(2) Product rule and $i \cdot i = -1$.
(3) Subtract the same expression with $m \leftrightarrow n$; the f'' terms cancel.
(4) Collect: $-in + im = i(m-n)$.

Key idea: the $\mathfrak{sl}(2)$ inside the Witt algebra

The three generators L_{-1}, L_0, L_1 close among themselves:

$$[L_1, L_{-1}] = 2L_0, \qquad [L_0, L_1] = -L_1, \qquad [L_0, L_{-1}] = L_{-1}.$$

They are the only L_n that extend to holomorphic vector fields on the whole Riemann sphere $\mathbb{C} \cup \{\infty\}$ (see below), and they generate the Möbius transformations $z \mapsto (az+b)/(cz+d)$. In a quantum theory they are the symmetries that even the vacuum respects. Keep an eye on them: the central term of the Virasoro algebra will vanish exactly on this subalgebra.

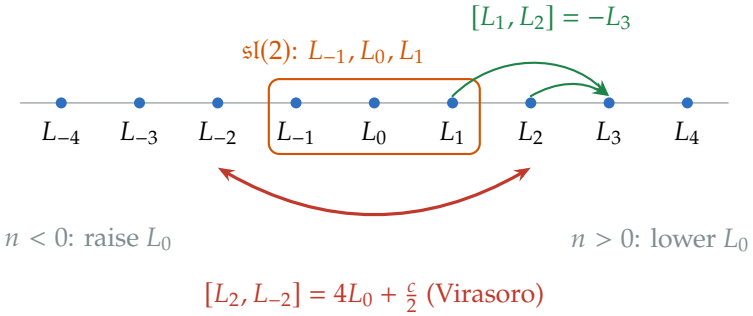


Figure 6.6: The Witt/Virasoro generators as points on a line labelled by mode number. Brackets add mode numbers (green). The orange box is the $\mathfrak{sl}(2)$ of Möbius transformations. The red arc is a bracket of L_n with L_{-n}, which in the Virasoro algebra picks up the c-number $\frac{c}{12}(n^3 - n)$; for $n = 2$ this is $\frac{c}{2}$, and for $n = 0, \pm 1$ it is zero, so the orange box is unaffected. In a quantum theory, $[L_0, L_n] = -nL_n$ says that L_n with $n > 0$ lowers the eigenvalue of L_0 by n and L_{-n} raises it by n.

6.1.8 The Virasoro algebra (first look)

Nash now states the second central actor of the chapter. In string theory the group that matters is not $\mathrm{Diff}(S^1)$ itself but a central extension $\widehat{\mathrm{Diff}(S^1)}$, the **Virasoro group**. Its Lie algebra, the **Virasoro algebra**, has generators L_n ($n \in \mathbb{Z}$) and one extra central element c, with

$$\boxed{[L_m, L_n] = (m-n)L_{m+n} + \frac{c}{12}\, m(m^2-1)\,\delta_{m,-n}, \qquad [L_m, c] = 0, \qquad m, n \in \mathbb{Z}} \tag{6.21}$$

Nash's (6.14). The number c (its value in an irreducible representation) is the **central charge**. Compare the Kac–Moody algebra (6.12): again the new term appears only when the mode numbers cancel, and again it is a c-number. Three quick checks:

- *Antisymmetry.* Swapping $m \leftrightarrow n$ with $n = -m$ turns $m(m^2-1)$ into $(-m)(m^2-1)$: the term is odd, as a commutator must be.
- *The $\mathfrak{sl}(2)$ is untouched.* $m(m^2-1) = (m-1)m(m+1)$ vanishes for $m = 0, \pm 1$.
- *Why the $\frac{1}{12}$ and why $m^3 - m$.* Both are forced. In Section 6.2.8 we show that the Jacobi identity allows only $am^3 + bm$, that the bm part can be absorbed in a shift of L_0, and in Section 6.2.9 that one free boson gives exactly $\frac{1}{12}(m^3 - m)$, so that with this normalisation c counts free bosons.

Physically c is a measure of the number of degrees of freedom: a free boson has $c = 1$, a free Majorana fermion $c = \frac{1}{2}$, and D free bosons (a string in D dimensions) have $c = D$.

6.1.9 The conformal group in two dimensions

Why is $\mathrm{Diff}(S^1)$ so closely tied to conformal invariance? In d dimensions a **conformal transformation** is a map that preserves angles; equivalently, it rescales the metric by a positive function, $g_{\mu\nu} \mapsto \Omega(x)^2 g_{\mu\nu}$. For $d \geq 3$ the conformal transformations of flat space form a finite-dimensional group. In $d = 2$ they form an infinite-dimensional one. Here is why.

The conformal Killing equation

An infinitesimal map $x^\mu \mapsto x^\mu + \varepsilon v^\mu(x)$ changes the flat metric $\delta_{\mu\nu}$ by $\varepsilon(\partial_\mu v_\nu + \partial_\nu v_\mu)$. It is conformal if this change is proportional to the metric. Taking the trace fixes the proportionality factor, and we get the **conformal Killing equation**

$$\partial_\mu v_\nu + \partial_\nu v_\mu = \frac{2}{d}\,(\partial\cdot v)\,\delta_{\mu\nu}. \tag{6.22}$$

In $d = 2$ with coordinates (x, y) and $v = (v_1, v_2)$ write out the three independent components:

$$\begin{aligned}
(\mu\nu) = (11): \quad & 2\partial_x v_1 \overset{(1)}{=} \partial_x v_1 + \partial_y v_2 && \Longrightarrow \quad \partial_x v_1 = \partial_y v_2,\\
(\mu\nu) = (22): \quad & 2\partial_y v_2 \overset{(2)}{=} \partial_x v_1 + \partial_y v_2 && \Longrightarrow \quad \text{the same equation},\\
(\mu\nu) = (12): \quad & \partial_x v_2 + \partial_y v_1 \overset{(3)}{=} 0 && \Longrightarrow \quad \partial_y v_1 = -\partial_x v_2.
\end{aligned}$$

(1) Equation (6.22) with $d = 2$ and $\partial\cdot v = \partial_x v_1 + \partial_y v_2$.
(2) The same, for the other diagonal component.
(3) The off-diagonal component: $\delta_{12} = 0$.

These are the *Cauchy–Riemann equations* for the complex function $v(z) = v_1 + iv_2$ of $z = x + iy$. So the conformal vector fields in two dimensions are exactly the holomorphic functions $v(z)$, and the transformations are $z \mapsto z + \varepsilon v(z)$. Finite conformal transformations are the holomorphic maps $z \mapsto w(z)$ with $w' \neq 0$ (and the antiholomorphic ones, which reverse orientation). There are infinitely many of them: $w(z) = z + \varepsilon z^{n+1}$ for every n. In $d \geq 3$, by contrast, (6.22) has only finitely many solutions: translations, rotations, the dilatation and the special conformal transformations, forming $\mathrm{SO}(d+1,1)$ of dimension $\frac{1}{2}(d+1)(d+2)$ (in $d = 4$: $4 + 6 + 1 + 4 = 15$).

Worked example 6.11: a holomorphic map preserves angles

Take $w = z^2$. The coordinate lines $x = \text{const}$ and $y = \text{const}$ meet at right angles, and so do their images. Indeed, with $w = u + iv$, $u = x^2 - y^2$, $v = 2xy$; a small displacement $\mathrm{d}z$ is mapped to

$$\mathrm{d}w \overset{(1)}{=} w'(z)\,\mathrm{d}z \overset{(2)}{=} 2z\,\mathrm{d}z = |2z|\; e^{i\arg(2z)}\,\mathrm{d}z.$$

(1) Holomorphic functions are complex differentiable: the derivative is a single complex number, independent of the direction of $\mathrm{d}z$.
(2) Write the complex number $2z$ in polar form.

Every small displacement at z is stretched by the same factor $|2z|$ and rotated by the same angle $\arg(2z)$. Two displacements are therefore rotated together, and the angle between them is preserved. The metric is rescaled: $|\mathrm{d}w|^2 = |w'|^2\,|\mathrm{d}z|^2$, i.e. $\Omega^2 = |w'(z)|^2$. (At $z = 0$, where $w' = 0$, angles are doubled; the map is not conformal there.)

Two commuting Witt algebras

The complexified conformal vector fields are $v(z)\partial_z$ (holomorphic) and $\bar v(\bar z)\partial_{\bar z}$ (antiholomorphic). Expanding $v(z)$ in a Laurent series gives the basis

$$\ell_n = -z^{n+1}\partial_z, \qquad \bar\ell_n = -\bar z^{n+1}\partial_{\bar z}, \qquad n \in \mathbb{Z}. \tag{6.23}$$

Exactly as in (6.20),

$$[\ell_m, \ell_n] = (m-n)\ell_{m+n}, \qquad [\bar\ell_m, \bar\ell_n] = (m-n)\bar\ell_{m+n}, \qquad [\ell_m, \bar\ell_n] = 0,$$

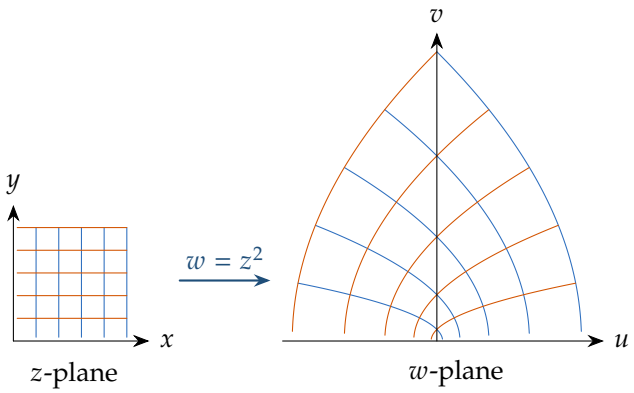


Figure 6.7: A holomorphic map is conformal. The square grid of the z-plane (left) is mapped by $w = z^2$ to a grid of parabolas (right). The cells are curved and of different sizes, but at every crossing the blue and orange curves still meet at a right angle. There is one such map for every holomorphic function, which is why the two-dimensional conformal group is infinite-dimensional.

the last because $\partial_z \bar{z} = 0$ and $\partial_{\bar{z}} z = 0$. So the local conformal algebra in two dimensions is two commuting copies of the Witt algebra: two Virasoro algebras with $c = 0$, as Nash says. In the quantum theory each copy acquires its own central charge, c and $\bar{c}$. For a closed string the two copies are the left-moving and right-moving oscillations.

Boundary values on the circle

Nash's remark that holomorphic maps are "determined by their boundary values on the unit circle" is the reason the conformal algebra reduces to vector fields on S^1. A holomorphic function on an annulus around $|z| = 1$ is determined by its Laurent series $\sum_n v_n z^n$, and the Laurent coefficients are just the Fourier coefficients of its restriction to the circle, $v(e^{i\theta}) = \sum_n v_n e^{in\theta}$. On the unit circle the field ℓ_n becomes exactly $L_n = ie^{in\theta}\partial_\theta$ by (6.18). So the Fourier expansion on S^1 "is" the Laurent expansion in the plane. In radial quantisation, circles $|z| = e^\tau$ are equal-time slices, τ is Euclidean time, and the dilatation $-(\ell_0 + \bar{\ell}_0) = z\partial_z + \bar{z}\partial_{\bar{z}} = \partial_\tau$ generates time translation: $L_0 + \bar{L}_0$ is the Hamiltonian (Chapter 11).

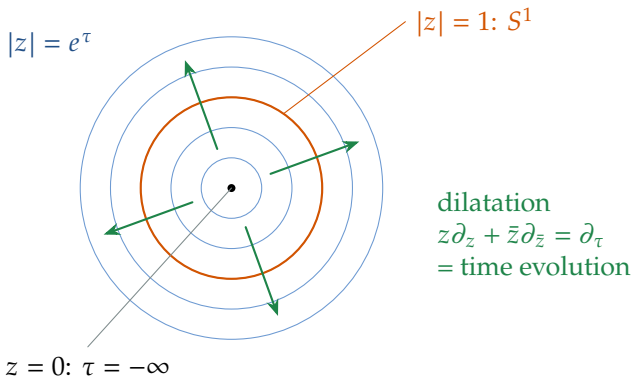


Figure 6.8: Radial quantisation. Map the cylinder (space S^1, Euclidean time τ) to the plane by $z = e^{\tau + i\theta}$. Equal-time circles become concentric circles (blue); the far past is the origin. The generators $L_n = -z^{n+1}\partial_z$ restricted to the unit circle (orange) are the vector fields on S^1, and the radial dilatation (green) is the Hamiltonian.

Watch out: local versus global conformal transformations

The vector fields ℓ_n are holomorphic on $\mathbb{C}\setminus\{0\}$ but most of them are singular somewhere on the Riemann sphere. $\ell_n = -z^{n+1}\partial_z$ is regular at $z=0$ only for $n \geq -1$. Near $z=\infty$ use $w = 1/z$: $\partial_z = -w^2\partial_w$, so $\ell_n = w^{1-n}\partial_w$, regular at $w=0$ only for $n \leq 1$. Only $n = -1, 0, 1$ survive both tests. So the *global* conformal group of the sphere is the finite-dimensional Möbius group $\mathrm{PSL}(2,\mathbb{C})$; the infinite-dimensional Witt algebra consists of *local* conformal transformations, defined on an annulus. For a physicist, local is what matters: Ward identities are local.

6.1.10 The group of gauge transformations

Physicists write a gauge transformation as a function $g(x)$ with values in G, and so it looks like an element of $\mathrm{Map}(X, G)$. That is right locally, but not globally. If the gauge field is a connection on a *nontrivial* principal bundle P (a monopole, an instanton), then there is no global choice of gauge, and a gauge transformation cannot be described by a single function on all of X. We need a description that uses the bundle itself.

Reminder: principal bundles

A **principal G-bundle** $\pi : P \to X$ is a space P with a free right action of G, $p \mapsto pg$, whose orbits are the fibres $P_x = \pi^{-1}(x)$, each a copy of G. Over a small open set $U \subset X$ we can choose a **local section** $s : U \to P$ and then write every point of $\pi^{-1}(U)$ uniquely as $p = s(x)h$; in these *local coordinates* $p = (x, h)$ and the right action is $(x,h)g = (x, hg)$. On an overlap of two patches the sections differ by a **transition function**: $s_j(x) = s_i(x)\, t_{ij}(x)$ with $t_{ij} : U_i \cap U_j \to G$. A connection is a $\mathfrak{g}$-valued 1-form A_i on each patch, glued by $A_j = t_{ij}^{-1} A_i t_{ij} + t_{ij}^{-1}\mathrm{d}t_{ij}$ (Chapter T3).

Dictionary: anti-Hermitian versus Hermitian gauge fields

In this subsection we use the mathematicians' convention of Chapter T3: A is anti-Hermitian, $F = \mathrm{d}A + A\wedge A$, and a gauge transformation acts as $A \mapsto g^{-1}Ag + g^{-1}\mathrm{d}g$. The physicists' Hermitian field is A^a_{phys} with $A_{\text{math}} = -i\, g_{\text{YM}} A^a_{\text{phys}} T^a$ and Hermitian T^a. For $G = \mathrm{U}(1)$ and $g = e^{i\lambda}$, $A_{\text{math}} = -iea$ transforms as $-iea \mapsto -iea + i\mathrm{d}\lambda$, i.e. $a \mapsto a - \frac{1}{e}\mathrm{d}\lambda$, the textbook gauge transformation of electromagnetism.

Bundle automorphisms

A bundle has the same physical content after any transformation that preserves its structure. So the symmetries of a gauge theory built on P are the maps $P \to P$ that respect the G-action.

Working definition 6.6 (bundle automorphism)

A **bundle automorphism** is a diffeomorphism $\alpha : P \to P$ that commutes with the action of G [moving p to pg and then applying α lands at the same point as applying α first and then multiplying by g]. Nash writes the action on the left, $g\cdot p := pg^{-1}$ (his (6.17)), so the condition reads $\alpha(g\cdot p) = g\cdot\alpha(p)$ (his (6.16)); replacing g by g^{-1} this is the same as

$$\alpha(pg) = \alpha(p)\, g \qquad \text{for all } p \in P,\ g \in G, \tag{6.24}$$

his (6.18). The automorphisms form a group $\mathrm{Aut}(P)$ under composition, $(\alpha\bullet\beta)(p) = \alpha(\beta(p))$. *Operationally:* α may move fibres around, but it must map each fibre [$P_x = \pi^{-1}(x)$, the

copy of G over the point x] to a fibre, and it must do so "by left multiplication", so that it commutes with right multiplication.

Why must an automorphism map fibres to fibres? The fibre through p is the orbit $\{pg\}$. By (6.24) its image is $\{\alpha(p)g\}$, the orbit of $\alpha(p)$, which is again a fibre. Hence α induces a map $h(\alpha) : X \to X$ of the base, $h(\alpha)(\pi(p)) := \pi(\alpha(p))$. It is a diffeomorphism, and $h(\alpha \bullet \beta) = h(\alpha) \circ h(\beta)$. So h is a group homomorphism

$$h : \mathrm{Aut}(P) \longrightarrow \mathrm{Diff}(X).$$

Working definition 6.7 (gauge group)

The **group of gauge transformations** is the kernel of h:

$$\bar{\mathcal{G}} := \mathrm{Ker}\, h = \{\alpha \in \mathrm{Aut}(P) : \ \pi(\alpha(p)) = \pi(p) \text{ for all } p\},$$

the automorphisms that map every fibre *to itself*. The **based gauge group** is the subgroup

$$\mathcal{G} := \{\alpha \in \bar{\mathcal{G}} : \ \alpha(p_0) = p_0\},$$

for a fixed point $p_0 \in P$ over a base point $x_0 = \pi(p_0)$. By (6.24), $\alpha(p_0 g) = p_0 g$ for all g, so an element of $\mathcal{G}$ acts as the identity on the whole fibre P_{x_0}.

The based gauge group is used in Chapter 8: unlike $\bar{\mathcal{G}}$, it acts *freely* on the space of connections, so the quotient is a manifold. The quotient $\bar{\mathcal{G}}/\mathcal{G}$ is a copy of G (the value of the gauge transformation at the base point), exactly as in the splitting $\mathrm{Map}(X, G) \cong G \times \mathrm{Map}_0(X, G)$ of (6.4).

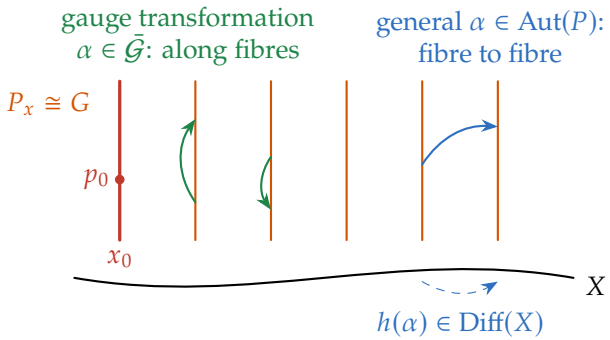


Figure 6.9: The bundle P drawn as fibres (orange) over the base X. A gauge transformation moves points only up and down their own fibre (green). A general automorphism may carry one fibre onto another (blue), and so covers a diffeomorphism of the base (dashed). Based gauge transformations are, in addition, the identity on the fibre over the base point (red).

Key idea: $\mathrm{Aut}(P)$ contains both kinds of invariance

The sequence

$$1 \longrightarrow \bar{\mathcal{G}} \longrightarrow \mathrm{Aut}(P) \xrightarrow{h} \mathrm{Diff}(X) \tag{6.25}$$

is exact: $\bar{\mathcal{G}}$ is exactly what h kills. The group $\mathrm{Aut}(P)$ therefore contains the two invariances of a gauge theory coupled to gravity: *internal* gauge invariance, which transforms the fibres ($\bar{\mathcal{G}}$), and *Einstein's* general coordinate invariance, which transforms the base ($\mathrm{Diff}(X)$). For a trivial bundle $P = X \times G$ every diffeomorphism lifts, $\alpha(x, g) = (\phi(x), g)$, so h is onto and $\mathrm{Aut}(P)$ is the semidirect product $\mathrm{Map}(X, G) \rtimes \mathrm{Diff}(X)$. This is our first example of a *group*

extension, the subject of §6.2.

Gauge transformations as sections of $\mathrm{Ad}P$

Now we describe $\bar{\mathcal{G}}$ concretely. Given $\alpha \in \bar{\mathcal{G}}$, the points p and $\alpha(p)$ lie in the same fibre, so they differ by a unique group element. Define $\hat{\alpha} : P \to G$ by

$$\alpha(p) = p\,\hat{\alpha}(p), \tag{6.26}$$

Nash's (6.19). The equivariance (6.24) becomes a condition on $\hat{\alpha}$:

$$\begin{aligned} (pg)\,\hat{\alpha}(pg) &\overset{(1)}{=} \alpha(pg) \overset{(2)}{=} \alpha(p)\,g \overset{(3)}{=} p\,\hat{\alpha}(p)\,g, \\ \Longrightarrow \quad \hat{\alpha}(pg) &\overset{(4)}{=} g^{-1}\,\hat{\alpha}(p)\,g. \end{aligned} \tag{6.27}$$

(1) Definition (6.26) at the point pg.
(2) Equivariance (6.24).
(3) Definition (6.26) at the point p.
(4) Cancel p (the action is free, so $pa = pb$ implies $a = b$) and multiply by g^{-1} on the left.

This is Nash's (6.20). In local coordinates $p = (x, h)$, i.e. $p = s(x)h$, define $\hat{\alpha}_X(x) := \hat{\alpha}(s(x))$. Then by (6.27) with $p = s(x)$ and $g = h$,

$$\alpha(x,h) \overset{(1)}{=} (x,h)\,\hat{\alpha}(x,h) \overset{(2)}{=} \big(x,\ h\,h^{-1}\hat{\alpha}_X(x)h\big) \overset{(3)}{=} \big(x,\ \hat{\alpha}_X(x)\,h\big). \tag{6.28}$$

(1) Definition (6.26).
(2) Equivariance (6.27) of $\hat{\alpha}$, and the right action $(x,h)g = (x,hg)$.
(3) $hh^{-1} = 1$.

This is Nash's (6.21): *in a local trivialisation, a gauge transformation is left multiplication of the fibre coordinate by a G-valued function $\hat{\alpha}_X(x)$.* That is the physicist's $g(x)$.

What happens when we change patch? Let $s_j = s_i t_{ij}$ on an overlap and write $\phi_i = \hat{\alpha}(s_i)$, $\phi_j = \hat{\alpha}(s_j)$. Then

$$\phi_j(x) \overset{(1)}{=} \hat{\alpha}\big(s_i(x)\,t_{ij}(x)\big) \overset{(2)}{=} t_{ij}(x)^{-1}\,\phi_i(x)\,t_{ij}(x). \tag{6.29}$$

(1) Definition of ϕ_j and $s_j = s_i t_{ij}$.
(2) Equation (6.27) with $g = t_{ij}$.

So the local functions $\phi_i : U_i \to G$ do *not* glue to a single function $X \to G$. They glue by *conjugation* with the transition functions. An object that is locally a G-valued function and glues this way is a section of the **adjoint bundle**

$$\mathrm{Ad}P := P \times_{\mathrm{Ad}} G,$$

the bundle with fibre G whose transition functions act on the fibre by conjugation, $\phi \mapsto t^{-1}\phi t$, rather than by multiplication.

Working definition 6.8 (gauge group as sections)

$$\bar{\mathcal{G}} = \Gamma(X, \mathrm{Ad}P), \qquad \mathrm{Ad}P = P \times_{\mathrm{Ad}} G.$$

[Γ denotes sections, a smooth choice of one point $\phi(x)$ in the fibre over each $x \in X$]

Operationally: a gauge transformation is a collection of functions $\phi_i : U_i \to G$, one per patch, related on overlaps by $\phi_j = t_{ij}^{-1}\phi_i t_{ij}$ [t_{ij} is the transition function, $s_j = s_i t_{ij}$ on the overlap $U_i \cap U_j$]. If G is abelian, conjugation does nothing, $\mathrm{Ad}P = X \times G$, and $\bar{\mathcal{G}} = \mathrm{Map}(X, G)$ whatever P is. If P is trivial [$P = X \times G$, so a single section covers all of X], one patch suffices and again $\bar{\mathcal{G}} = \mathrm{Map}(X, G)$.

The product is still pointwise

Nash checks that composition of automorphisms becomes pointwise multiplication of sections. For $\alpha, \beta \in \bar{\mathcal{G}}$,

$$\alpha\big(\beta(p)\big) \overset{(1)}{=} \alpha\big(p\,\hat{\beta}(p)\big) \overset{(2)}{=} \alpha(p)\,\hat{\beta}(p) \overset{(3)}{=} p\,\hat{\alpha}(p)\,\hat{\beta}(p). \tag{6.30}$$

(1) Definition (6.26) for β.
(2) Equivariance (6.24) of α with $g = \hat{\beta}(p)$.
(3) Definition (6.26) for α.

So $\widehat{\alpha \bullet \beta} = \hat{\alpha}\,\hat{\beta}$, the pointwise product; in local coordinates $\hat{\gamma}_X(x) = \hat{\alpha}_X(x)\hat{\beta}_X(x)$, which is Nash's (6.22)–(6.23). (Nash's version inserts $\hat{\beta}_X\hat{\beta}_X^{-1}$ in the middle; the result is the same.) Differentiating at the identity, the Lie algebra of $\bar{\mathcal{G}}$ is the space of sections of the **adjoint Lie algebra bundle** $\mathrm{ad}P = P \times_{\mathrm{Ad}} \mathfrak{g}$, whose fibre is $\mathfrak{g}$ and whose transition functions act by $\xi \mapsto t^{-1}\xi t$. Summarising (Nash's (6.24)):

$$\boxed{\bar{\mathcal{G}} = \Gamma(X, \mathrm{Ad}P), \quad \mathrm{Ad}P = P \times_{\mathrm{Ad}} G; \qquad \mathrm{Lie}\,\bar{\mathcal{G}} = \Gamma(X, \mathrm{ad}P), \quad \mathrm{ad}P = P \times_{\mathrm{Ad}} \mathfrak{g}.} \tag{6.31}$$

Worked example 6.12: gauge transformations on the instanton bundle

Let P be the SU(2) bundle over S^4 with instanton number 1. Cover S^4 by a northern and a southern patch overlapping near the equator S^3, with transition function $t(x) = x_4 + i\vec{x}\cdot\vec{\sigma}$, the degree-one map of Example 6.5. A gauge transformation is a pair ϕ_N, ϕ_S of SU(2)-valued functions with $\phi_S = t^{-1}\phi_N t$ on the equator.

- The constant pair $\phi_N = \phi_S = -1$ is allowed, because -1 is central and $t^{-1}(-1)t = -1$.
- Take $\phi_N = u = e^{i\lambda\sigma_3}$, constant on the northern patch. On the overlap the rule forces $\phi_S = t^{-1}ut$, which is *not* constant: as x runs over the equator, $t^{-1}ut$ runs over the whole conjugacy class of u, a 2-sphere inside SU(2). This map $S^3 \to \mathrm{SU}(2)$ misses most of SU(2), so it can be contracted and hence extended smoothly over the southern patch. The result is a legitimate gauge transformation, constant in the north and not in the south, and no single function on S^4 describes it.
- The pair $\phi_N = \phi_S = u$ with u a constant non-central element is *not* allowed, because $t^{-1}ut \neq u$ at most points of the equator. So "global" SU(2) rotations are not gauge transformations of this bundle: only the centre $\{\pm 1\}$ survives as constants.

The last point is physical: in an instanton background there is no global colour rotation, only the local gauge transformations. It is also why $\bar{\mathcal{G}}$ is not $\mathrm{Map}(S^4, \mathrm{SU}(2))$ here.

Worked example 6.13: the monopole bundle: abelian, so nothing changes

Let P be the U(1) bundle over S^2 of the Dirac monopole of charge n, with transition function

$t = e^{in\varphi}$ on the equator. A gauge transformation is a pair $\phi_N = e^{i\lambda_N}$, $\phi_S = e^{i\lambda_S}$ with

$$\phi_S \overset{(1)}{=} t^{-1}\phi_N t \overset{(2)}{=} \phi_N .$$

(1) The gluing rule (6.29).
(2) U(1) is abelian.

So the two local functions agree on the overlap and define one function $S^2 \to \mathrm{U}(1)$: $\bar{\mathcal{G}} = \mathrm{Map}(S^2, \mathrm{U}(1))$, although P is nontrivial. Acting on the potential $A_N = -\frac{in}{2}(1-\cos\vartheta)\mathrm{d}\varphi$ (anti-Hermitian convention) it gives $A_N \mapsto A_N + i\,\mathrm{d}\lambda$, the familiar gauge transformation, the same in both patches.

Watch out: large gauge transformations and nontrivial bundles are different things

Two different topological issues are easily confused. (i) $\bar{\mathcal{G}}$ may have many components even for a trivial bundle: $\pi_0(\mathrm{Map}(S^3, \mathrm{SU}(2))) = \mathbb{Z}$ (Example 6.5); these are the large gauge transformations behind the θ-vacuum. (ii) For a nontrivial bundle, $\bar{\mathcal{G}}$ is not a group of maps at all, but a group of sections of $\mathrm{Ad}P$. Both matter in Chapter 8.

6.1.11 Current algebra

The first time an infinite-dimensional Lie algebra appeared in physics was the current algebra of the strong interactions (Adler and Dashen; Gell-Mann and Ne'eman). The idea: even if we cannot solve the dynamics, the *equal-time commutators* of the conserved currents $j^a_\mu(x)$ of an internal symmetry are fixed by the symmetry algebra, and they lead to sum rules and low-energy theorems. The charges $Q^a = \int j^a_0\,\mathrm{d}^3x$ generate the finite group G; the local densities $j^a_0(x)$ generate its gauge group $\mathrm{Map}(\mathbb{R}^3, G)$.

Currents on a circle

Nash's concrete example is two-dimensional spacetime with periodic space, a circle with angle θ. Package the Kac–Moody generators of (6.12) (now thought of as operators on a Hilbert space) into a current density:

$$j^a(\theta) \coloneqq \frac{1}{2\pi}\sum_{n\in\mathbb{Z}} T^a_n\, e^{in\theta}, \qquad T^a_n = \int_0^{2\pi} j^a(\theta)\, e^{-in\theta}\,\mathrm{d}\theta. \tag{6.32}$$

This is Nash's (6.25), where we have put in the factor $\frac{1}{2\pi}$ that makes T^a_n the Fourier coefficients (Nash leaves it out; see below). If $(T^a_n)^\dagger = T^a_{-n}$, which is the case in a unitary representation, then $j^a(\theta)$ is a Hermitian operator at each θ. The total charge is $T^a_0 = \int j^a\,\mathrm{d}\theta$.

To compute the commutator of two currents we need the Fourier series of the delta function on the circle,

$$\sum_{m\in\mathbb{Z}} e^{im(\theta-\varphi)} = 2\pi\,\delta(\theta-\varphi), \qquad \sum_{m\in\mathbb{Z}} m\, e^{im(\theta-\varphi)} = -2\pi i\,\partial_\theta\delta(\theta-\varphi). \tag{6.33}$$

(The first is the completeness of the Fourier basis; the second follows by applying $-i\partial_\theta$ to the

first.) Now

$$
\begin{aligned}
[j^a(\theta), j^b(\varphi)] &\overset{(1)}{=} \frac{1}{4\pi^2}\sum_{m,n} e^{im\theta}e^{in\varphi}\,[T^a_m, T^b_n] \\
&\overset{(2)}{=} \frac{1}{4\pi^2}\sum_{m,n} e^{im\theta+in\varphi}\Big(i f^{abc}T^c_{m+n} + k\,m\,\delta^{ab}\delta_{m+n,0}\Big) \\
&\overset{(3)}{=} \frac{i f^{abc}}{4\pi^2}\sum_p T^c_p e^{ip\varphi}\sum_m e^{im(\theta-\varphi)} + \frac{k\,\delta^{ab}}{4\pi^2}\sum_m m\,e^{im(\theta-\varphi)} \\
&\overset{(4)}{=} i f^{abc}\,j^c(\varphi)\,\delta(\theta-\varphi) - \frac{ik}{2\pi}\,\delta^{ab}\,\partial_\theta\delta(\theta-\varphi). \qquad (6.34)
\end{aligned}
$$

(1) Definition (6.32), twice; the commutator is bilinear.
(2) The Kac–Moody relations (6.12).
(3) In the first term set $p = m + n$, so $e^{im\theta+in\varphi} = e^{ip\varphi}e^{im(\theta-\varphi)}$; in the second the Kronecker delta sets $n = -m$.
(4) The two sums (6.33), and $\frac{1}{2\pi}\sum_p T^c_p e^{ip\varphi} = j^c(\varphi)$.

This is Nash's (6.26), $[j^a(\theta), j^b(\varphi)] = i f^{abc} j^c(\theta)\delta(\theta-\varphi) + ic\,\delta^{ab}\delta'(\theta-\varphi)$, with his constant c equal to $-k/2\pi$ in our normalisation (and $j^c(\theta)\delta(\theta-\varphi) = j^c(\varphi)\delta(\theta-\varphi)$).

Key idea: the central term is the Schwinger term

The first term of (6.34) is the classical, pointwise bracket (6.3) of the gauge algebra: currents at different points commute, and at the same point they rotate into each other. The second term, proportional to the *derivative* of the delta function and to the unit operator, is the **Schwinger term**. It is the central extension of the loop algebra, seen in position space.

Worked example 6.14: smeared currents and the cocycle

Delta functions are awkward; smear them. For real test functions f, g on the circle define $Q^a(f) = \int_0^{2\pi} f(\theta)j^a(\theta)\,\mathrm{d}\theta$. Then

$$
\begin{aligned}
[Q^a(f), Q^b(g)] &\overset{(1)}{=} \int\!\!\int f(\theta)g(\varphi)\,[j^a(\theta), j^b(\varphi)]\,\mathrm{d}\theta\,\mathrm{d}\varphi \\
&\overset{(2)}{=} i f^{abc}\int f g\, j^c\,\mathrm{d}\varphi - \frac{ik}{2\pi}\delta^{ab}\int\!\!\int f(\theta)g(\varphi)\,\partial_\theta\delta(\theta-\varphi)\,\mathrm{d}\theta\,\mathrm{d}\varphi \\
&\overset{(3)}{=} i f^{abc}\,Q^c(fg) + \frac{ik}{2\pi}\,\delta^{ab}\int_0^{2\pi} f'(\varphi)\,g(\varphi)\,\mathrm{d}\varphi.
\end{aligned}
$$

(1) Bilinearity.
(2) Equation (6.34); the delta function does the θ integral in the first term.
(3) Integrate by parts in θ (no boundary on a circle): $\int f(\theta)\partial_\theta\delta(\theta-\varphi)\mathrm{d}\theta = -f'(\varphi)$.

The first term is the loop algebra: the smeared charges multiply their test functions pointwise. The second is a number that depends on f and g only, antisymmetric because $\int f'g = -\int fg'$. It vanishes if f or g is constant, so the global charges $Q^a(1) = T^a_0$ still form $\mathfrak{g}$. This number is the 2-*cocycle* of the extension; we will study exactly this bilinear form in Section 6.2.4. As a check, take $f = e^{im\theta}$, $g = e^{in\theta}$ (complex test functions). Then $Q^a(e^{im\theta}) = T^a_{-m}$ and the central term is $\frac{ik}{2\pi}\,(im)\,2\pi\,\delta_{m+n,0} = -km\,\delta_{m+n,0}$, which is the central term of $[T^a_{-m}, T^b_{-n}]$ in (6.12). ✓

Why the Schwinger term cannot vanish

Schwinger's 1959 argument shows that a term like this is forced on us in any relativistic quantum field theory, in any dimension. Let j_μ be a conserved current, $\partial_0 j_0 = -\partial_1 j_1$ (one space dimension for simplicity), and suppose naively that $[j_0(x), j_1(y)] = 0$ at equal times. Consider the smeared charge density $j_0(f) = \int f(x) j_0(x)\,\mathrm{d}x$ and the vacuum expectation value of a double commutator with the Hamiltonian H, normalised so that $H|0\rangle = 0$:

$$\langle 0|[j_0(f),[H,j_0(f)]]|0\rangle \overset{(1)}{=} 2\sum_n E_n\ |\langle n|j_0(f)|0\rangle|^2\ >\ 0,$$

$$[H,j_0(f)] \overset{(2)}{=} -i\int f\,\partial_0 j_0\,\mathrm{d}x \overset{(3)}{=} i\int f\,\partial_1 j_1\,\mathrm{d}x \overset{(4)}{=} -i\int f'\,j_1\,\mathrm{d}x.$$

(1) Insert a complete set of energy eigenstates; $j_0(f)$ is Hermitian and $E_n \geq 0$. The sum is strictly positive unless $j_0(f)|0\rangle = 0$, which would make the charge density annihilate the vacuum; for any interacting or free theory with charged states it does not.
(2) Heisenberg equation $\partial_0 O = i[H, O]$.
(3) Current conservation.
(4) Integrate by parts.

So $\langle 0|[j_0(f), j_1(f')]|0\rangle \neq 0$, contradicting the naive assumption. The equal-time commutator of j_0 and j_1 must contain a term with a derivative of a delta function, a Schwinger term. In two dimensions, with chiral currents $j = j_0 \pm j_1$, this is precisely the central term of (6.34). We will compute it explicitly for free fermions in Section 6.2.9, where it comes out of the filled Dirac sea.

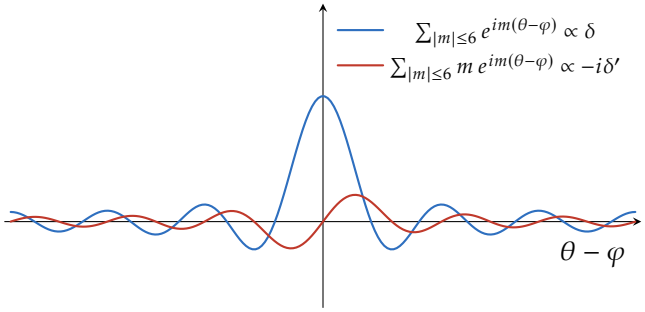


Figure 6.10: Truncated Fourier sums (rescaled for display). The sum of the modes $e^{im(\theta-\varphi)}$ (blue) peaks at $\theta = \varphi$ and tends to $2\pi\delta(\theta-\varphi)$: this gives the pointwise term of the current algebra. Weighting each mode by m, as the central term $km\delta_{m+n,0}$ does, gives an odd function (red; its imaginary part is shown) that tends to $-2\pi i\delta'$: the Schwinger term.

6.1.12 Analyticity, and why $\mathrm{Diff}(X)$ is strange

All the groups above are smooth infinite-dimensional Lie groups. Nash ends the section with a warning: beyond smoothness, the two families behave very differently.

$\mathrm{Map}(X, G)$ is analytic

A finite-dimensional Lie group G is automatically *analytic*: in exponential coordinates $g = g_0 \exp(\xi)$ the group multiplication is given by convergent power series (the Baker–Campbell–Hausdorff series). The group $\mathrm{Map}(X, G)$ inherits this, pointwise. Topologise $\mathrm{Map}(X, G)$ by *uniform convergence of all derivatives*: $\alpha_n \to \alpha$ if $\alpha_n \to \alpha$ uniformly on X and so do all partial

derivatives of all orders. Then charts

$$\begin{aligned}\mathrm{Map}(X,\mathfrak{g}) \supset \{\xi \text{ small}\} &\longrightarrow \mathrm{Map}(X,G),\\ \xi &\longmapsto \alpha_0 \exp(\xi), \qquad (\alpha_0 \exp \xi)(x) = \alpha_0(x)\exp(\xi(x))\end{aligned}$$

cover a neighbourhood of α_0, and the changes of chart and the group law are given, point by point, by the same convergent power series as in G. So $\mathrm{Map}(X,G)$ has an atlas of analytic charts ("Taylor series with no remainder"), and its exponential map $\mathrm{Map}(X,\mathfrak{g}) \to \mathrm{Map}(X,G)$ is a local diffeomorphism near 0, just as for G.

$\mathrm{Diff}(X)$ is not

For $\mathrm{Diff}(X)$ none of this holds (Milnor). The composition $\alpha \circ \beta$ involves α evaluated at a moved point, and Taylor expanding $\alpha(\beta(x))$ in β costs derivatives of α; the power series do not close on any function space. The most striking symptom concerns the exponential map $\exp : \mathrm{vect}(X) \to \mathrm{Diff}(X)$, $V \mapsto \alpha_1$ (the time-one flow).

Key idea: the exponential map of $\mathrm{Diff}(S^1)$ is not a local homeomorphism

There are diffeomorphisms of the circle arbitrarily close to the identity that lie on *no* one-parameter subgroup; and there are diffeomorphisms arbitrarily close to the identity that lie on *several* one-parameter subgroups. So near the identity, exp is neither onto nor one-to-one, and the flows of vector fields do not provide canonical coordinates on $\mathrm{Diff}(X)$. In finite dimensions the inverse function theorem forbids this; in the Fréchet space $\mathrm{vect}(S^1)$ the inverse function theorem fails.

Both halves can be seen with explicit examples. We work on S^1 with angle θ.

Worked example 6.15: close to the identity, but on no one-parameter subgroup

For $n \geq 1$ and a small $\varepsilon > 0$, let

$$f_n(\theta) = \theta + \frac{\pi}{n} + \varepsilon \sin^2(n\theta).$$

As $n \to \infty$ with, say, $\varepsilon = e^{-n}$, f_n and all its derivatives converge uniformly to those of the identity (the k-th derivative of the ε term is at most $\varepsilon(2n)^k \to 0$). So $f_n \to \mathrm{id}$ in $\mathrm{Diff}(S^1)$.
Step 1: f_n has a periodic orbit. Let $\theta_j = j\pi/n$. Then $\sin^2(n\theta_j) = \sin^2(j\pi) = 0$, so

$$f_n(\theta_j) \overset{(1)}{=} \theta_j + \frac{\pi}{n} = \theta_{j+1}, \qquad f_n^{2n}(\theta_0) \overset{(2)}{=} \theta_{2n} = 2\pi \equiv \theta_0.$$

(1) The ε term vanishes at the points θ_j.
(2) Apply this $2n$ times.

Step 2: $f_n^{2n} \neq \mathrm{id}$. Take θ not of the form $j\pi/n$. Since $\varepsilon \sin^2 \geq 0$, and f_n is increasing,

$$f_n^{2n}(\theta) \overset{(1)}{=} f_n^{2n-1}\big(f_n(\theta)\big) \overset{(2)}{\geq} f_n(\theta) + (2n-1)\frac{\pi}{n} \overset{(3)}{>} \theta + 2\pi.$$

(1) Split off the first step.
(2) Each application of f_n adds at least π/n.
(3) $f_n(\theta) = \theta + \pi/n + \varepsilon \sin^2(n\theta) > \theta + \pi/n$ because $\sin(n\theta) \neq 0$.

Step 3: f_n is not $\exp(V)$ for any V. Suppose f_n is the time-one flow of $V = v(\theta)\partial_\theta$. If $v(\theta_*) = 0$ somewhere, θ_* would be fixed by the flow and hence by f_n; but $f_n(\theta) > \theta$ everywhere. So v never vanishes. A nowhere-vanishing vector field on a circle can be straightened: in

the new coordinate $s(\theta) = \lambda \int_0^\theta d\theta'/v(\theta')$, with λ chosen so that s runs from 0 to 2π, the field is the constant $\lambda\,\partial_s$, and its time-one flow is the rigid rotation $s \mapsto s + \lambda$. A rigid rotation that has a periodic orbit of period $2n$ going once round the circle is the rotation by π/n, and then *every* orbit has period $2n$, i.e. its $2n$-th power is the identity. Changing coordinates does not affect this, so $f_n^{2n} = \mathrm{id}$, contradicting Step 2.
So f_n lies on no one-parameter subgroup, however close to the identity it is.

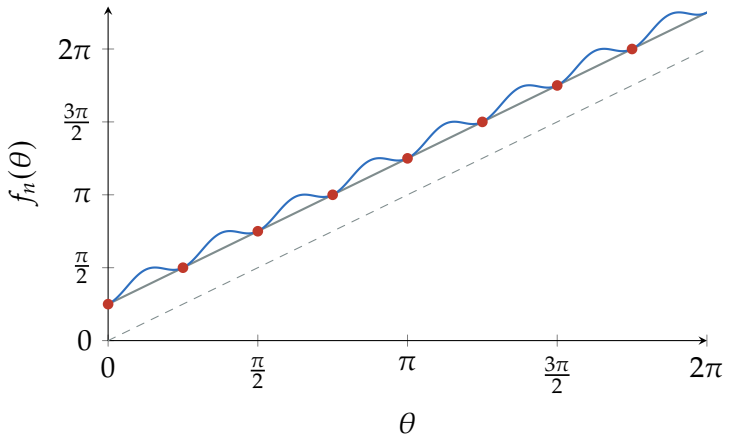


Figure 6.11: Milnor's example with $n = 4$ (and ε exaggerated). The blue graph $f_4(\theta) = \theta + \frac{\pi}{4} + \varepsilon \sin^2 4\theta$ touches the graph of the rigid rotation by $\pi/4$ (grey) exactly at the red points $\theta = j\pi/4$, which form a periodic orbit of period 8. Everywhere else f_4 moves points strictly further, so $f_4^8 \neq \mathrm{id}$. A time-one flow with a period-8 orbit would have $f^8 = \mathrm{id}$; hence f_4 is not the exponential of any vector field. Letting $n \to \infty$ and $\varepsilon \to 0$ gives such maps arbitrarily close to the identity (dashed).

Worked example 6.16: close to the identity, on many one-parameter subgroups

Let R be the rotation $\theta \mapsto \theta + \frac{2\pi}{n}$, the time-one flow of the constant field $V = \frac{2\pi}{n}\partial_\theta$. Let ϕ be any diffeomorphism that commutes with R, for instance $\phi(\theta) = \theta + \delta \sin(n\theta)$ with $|\delta| < 1/n$. Then

$$\phi \circ \exp(tV) \circ \phi^{-1} \overset{(1)}{=} \exp\big(t\,\phi_* V\big),$$
$$\phi \circ R \circ \phi^{-1} \overset{(2)}{=} R.$$

(1) Conjugating a flow by a diffeomorphism gives the flow of the transported vector field $\phi_* V$.
(2) ϕ commutes with R: $\phi(\theta + 2\pi/n) = \theta + 2\pi/n + \delta \sin(n\theta + 2\pi) = \phi(\theta) + 2\pi/n$.

So $R = \exp(V) = \exp(\phi_* V)$ at $t = 1$. The field $\phi_* V = \frac{2\pi}{n}\phi'(\phi^{-1}(\theta))\partial_\theta$ is not constant (it is $\frac{2\pi}{n}(1 + n\delta\cos(n\phi^{-1}(\theta)))\partial_\theta$), so R lies on two different one-parameter subgroups (in fact on a whole family, one for each δ). As $n \to \infty$ both R and the vector fields tend to the identity and to zero. So exp is not one-to-one on any neighbourhood of 0.

Watch out: what this does and does not mean for physics

Nothing here stops us from using $\mathrm{vect}(S^1)$ and its central extension, the Virasoro algebra, in quantum theory: the Lie algebra is perfectly well defined, and its representations are what we construct. What fails is the naive idea that every group element near the identity is $e^{i\sum \varepsilon_n L_n}$. For $\mathrm{Map}(X, G)$ that idea is fine.

Reading Nash (pp. 174–178, (6.11)–(6.26))

Nash's Diff(X) example is (6.11)–(6.13), stated without the flows; our Examples 6.8 and 6.9 and the sign pitfall fill in the details. His Virasoro relations (6.14) are quoted with the reference to Gel'fand–Fuks; we derive them in §6.2. The formula (6.15), $L_n = -z^{n+1}\frac{d}{dz} = e^{in\theta} i\frac{d}{d\theta}$, is our (6.18); the "two commuting Virasoro algebras (with $c = 0$)" are the $\ell_n, \bar{\ell}_n$ of (6.23). For the gauge group, his (6.16)–(6.24) are our (6.24)–(6.31); we have added the gluing rule (6.29), which is the real reason the gauge group is $\Gamma(X, \mathrm{Ad}P)$. In (6.25)–(6.26) Nash drops the factor $\frac{1}{2\pi}$ and leaves the constant in front of δ' unnamed; with our normalisation it is $-k/2\pi$. His analyticity paragraph cites Milnor for the failure of exp; Example 6.15 is the standard example.

Recap: Section 6.1

- Map(X, G): pointwise product; Lie algebra Map($X, \mathfrak{g}$) with pointwise bracket; $\pi_0 = \pi_0(G) \times [X, G]_0$, e.g. $\pi_3(\mathrm{SU}(2)) = \mathbb{Z}$ (degree), and $\tilde{K}^{-1}(X)$ in general.
- $LG = \mathrm{Map}(S^1, G)$; basis $T^a_n = e^{in\theta}T^a$ with $[T^a_m, T^b_n] = if^{abc}T^c_{m+n}$; the quantum version adds the central term $km\delta^{ab}\delta_{m+n,0}$ (affine Kac–Moody).
- Diff(X): composition; Lie algebra = vector fields (bracket up to a sign convention); flows $\alpha_t = \exp(tV)$.
- Diff(S^1): $L_n = ie^{in\theta}\partial_\theta = -z^{n+1}\partial_z$, Witt algebra $[L_m, L_n] = (m-n)L_{m+n}$; Virasoro adds $\frac{c}{12}m(m^2-1)\delta_{m+n,0}$; the $\mathfrak{sl}(2)$ of $L_{0,\pm1}$ is untouched.
- 2d conformal group: Cauchy–Riemann, two commuting Witt algebras.
- Gauge group $\bar{\mathcal{G}} = \mathrm{Ker}(\mathrm{Aut}(P) \to \mathrm{Diff}(X)) = \Gamma(X, \mathrm{Ad}P)$, pointwise product; locally $g(x)$ acting by left multiplication.
- Current algebra: $j^a(\theta)$ with Schwinger term $\propto \delta'$ = Kac–Moody central term.
- Map(X, G) is analytic; Diff(S^1) is not, and its exp is neither locally onto nor locally one-to-one.

6.2 Group extensions

In §6.1 we quoted two Lie algebras, the affine Kac–Moody algebra (6.12) and the Virasoro algebra (6.21), in a form with an extra c-number in the commutators. The classical groups LG and Diff(S^1) have no such term: their Lie algebras are $[T^a_m, T^b_n] = if^{abc}T^c_{m+n}$ and $[L_m, L_n] = (m-n)L_{m+n}$. This section answers three questions.

(a) *Where do c-numbers in commutators come from?* From the fact that quantum mechanics represents symmetries on rays, so that operators need only multiply up to a phase. The mathematical name is a *central extension*.

(b) *When can they be removed by redefining the generators?* This is a cohomology question: the c-number is a *2-cocycle*, and it can be removed exactly when it is a *coboundary*. For finite-dimensional semisimple groups it always can. For loop groups and Diff(S^1) it cannot.

(c) *How are the numbers k and c fixed in practice?* By the consistency (Jacobi) condition, which pins down the *form* of the central terms, and by explicit quantum field theories (free

bosons, free fermions, the Sugawara construction), which produce definite values.

6.2.1 Extensions of groups

We start with the group-theoretic language. It is the language of short exact sequences, which we met in Chapter 1.

Working definition 6.9 (group extension)

An **extension** of a group G by a group H is a group $\bar{G}$ together with homomorphisms [maps that respect products, $\alpha(h_1 h_2) = \alpha(h_1)\alpha(h_2)$]

$$1 \longrightarrow H \xrightarrow{\alpha} \bar{G} \xrightarrow{\beta} G \longrightarrow 1 \tag{6.35}$$

forming an **exact sequence** (Nash's (6.27)): at each place the image of the incoming map equals the kernel [the elements sent to the identity 1] of the outgoing one. Concretely:

- α is injective [distinct elements of H have distinct images] (the kernel of $H \to \bar{G}$ is the image of 1), so H sits inside $\bar{G}$ as a subgroup;
- β is surjective (the image of $\bar{G} \to G$ is the kernel of $G \to 1$, which is everything);
- $\alpha(H) = \operatorname{Ker}\beta$: the elements of $\bar{G}$ that "do nothing" in G are exactly those of H.

Operationally: $\bar{G}$ is a bigger group that projects onto G, and the extra information lost in the projection is an element of H.

Three consequences follow at once.

- H is a **normal subgroup** of $\bar{G}$, because a kernel always is: if $\beta(h) = 1$ then $\beta(ghg^{-1}) = \beta(g)\beta(g)^{-1} = 1$.
- $\bar{G}/H \cong G$: the fibres of β are the cosets of H.
- As a *set*, $\bar{G}$ looks like $H \times G$: each element of G has an H's worth of preimages. The group law of $\bar{G}$, however, need not be the product law.

Two special cases have names.

Working definition 6.10 (split and central extensions)

The extension (6.35) **splits** if β has a right inverse that is a homomorphism [$s(g_1 g_2) = s(g_1)s(g_2)$, not just any choice of one preimage of each g], $s : G \to \bar{G}$ with $\beta \circ s = \mathrm{id}$. Then $\bar{G}$ is the **semidirect product** $H \rtimes G$: every element is uniquely $h\, s(g)$, and

$$\big(h_1 s(g_1)\big)\big(h_2 s(g_2)\big) = \big(h_1\, s(g_1) h_2 s(g_1)^{-1}\big)\, s(g_1 g_2),$$

so G acts on H by conjugation. The extension is **central** if $\alpha(H)$ lies in the centre of $\bar{G}$: every element of H commutes with every element of $\bar{G}$. Then H is necessarily abelian.

For quantum mechanics the important case is a central extension by $H = \mathrm{U}(1)$ (or its Lie algebra $\mathbb{R}$): the extra information is a phase.

Worked example 6.17: $\mathrm{SU}(2)$ is a non-split central extension of $\mathrm{SO}(3)$

Every $U \in \mathrm{SU}(2)$ rotates the Pauli matrices, $U\sigma_j U^\dagger = \sum_i R_{ij}(U)\,\sigma_i$, and the rotation matrix is $R_{ij}(U) = \frac{1}{2}\operatorname{tr}(\sigma_i U \sigma_j U^\dagger)$. This defines $\beta : \mathrm{SU}(2) \to \mathrm{SO}(3)$, $U \mapsto R(U)$. Its kernel consists of the U commuting with all three Pauli matrices, i.e. the multiples of the identity in $\mathrm{SU}(2)$, $U = \pm 1$. So

$$1 \longrightarrow \mathbb{Z}_2 \longrightarrow \mathrm{SU}(2) \xrightarrow{\beta} \mathrm{SO}(3) \longrightarrow 1,$$

and the extension is central (± 1 commute with everything). Does it split? Suppose $s : \mathrm{SO}(3) \to \mathrm{SU}(2)$ were a homomorphism with $\beta \circ s = \mathrm{id}$. Let R_π be the rotation by π about the z-axis. Then

$$s(R_\pi)^2 \overset{(1)}{=} s(R_\pi^2) \overset{(2)}{=} s(\mathrm{id}) \overset{(3)}{=} 1,$$
$$s(R_\pi) \overset{(4)}{=} \pm 1, \qquad \text{so} \qquad R_\pi \overset{(5)}{=} \beta\big(s(R_\pi)\big) = \beta(\pm 1) = \mathrm{id}.$$

(1) s is a homomorphism.
(2) Rotating twice by π is rotating by 2π, the identity.
(3) A homomorphism sends the identity to the identity.
(4) If $V \in \mathrm{SU}(2)$ and $V^2 = 1$, the eigenvalues of V are ± 1 with product $\det V = 1$, so both are $+1$ or both are -1; a diagonalisable matrix with equal eigenvalues is a multiple of 1.
(5) $\beta \circ s = \mathrm{id}$, and ± 1 lie in the kernel.

This contradicts $R_\pi \neq \mathrm{id}$. So there is no way to choose the sign of the spin-$\frac{1}{2}$ rotation operators consistently: the physical statement is that a rotation by 2π acts on a spinor as -1.

Worked example 6.18: the Euclidean group is split but not central

The rigid motions of the plane, $x \mapsto Rx + a$ with $R \in \mathrm{SO}(2)$ and $a \in \mathbb{R}^2$, compose as

$$(a, R)\,(b, S) : \; x \overset{(1)}{\longmapsto} R(Sx + b) + a \overset{(2)}{=} (RS)x + (a + Rb).$$

(1) First apply (b, S), then (a, R).
(2) Expand.

So $(a, R)(b, S) = (a + Rb, RS)$. The translations $H = \mathbb{R}^2$ form a normal subgroup, the quotient is $G = \mathrm{SO}(2)$, and $R \mapsto (0, R)$ is a homomorphic section: $E(2) = \mathbb{R}^2 \rtimes \mathrm{SO}(2)$. It is not central, because rotations do not commute with translations. (The Poincaré group $\mathbb{R}^4 \rtimes \mathrm{SO}(3,1)$ and the group $\mathrm{Aut}(X \times G) = \mathrm{Map}(X, G) \rtimes \mathrm{Diff}(X)$ of Section 6.1.10 are of the same kind.)

Worked example 6.19: the Heisenberg group: the simplest central extension

Consider the real upper-triangular matrices

$$M(a, b, c) = \begin{pmatrix} 1 & a & c \\ 0 & 1 & b \\ 0 & 0 & 1 \end{pmatrix}, \qquad a, b, c \in \mathbb{R}.$$

Multiply two of them:

$$M(a,b,c)\,M(a',b',c') \overset{(1)}{=} \begin{pmatrix} 1 & a'+a & c'+ab'+c \\ 0 & 1 & b'+b \\ 0 & 0 & 1 \end{pmatrix} \overset{(2)}{=} M\big(a+a',\, b+b',\, c+c'+ab'\big).$$

(1) Row times column; the only non-obvious entry is the top right one, $1 \cdot c' + a \cdot b' + c \cdot 1$.
(2) Read off the entries.

The elements $M(0,0,c)$ commute with everything (the ab' term only involves a and b'), so $H = \{M(0,0,c)\} \cong \mathbb{R}$ is central. The quotient, forgetting c, is $G = \mathbb{R}^2$ with $(a,b)+(a',b')$: an abelian group. So we have a central extension

$$0 \longrightarrow \mathbb{R} \longrightarrow \mathrm{Heis} \longrightarrow \mathbb{R}^2 \longrightarrow 0.$$

It is not trivial, because Heis is not abelian:

$$M(a,0,0)M(0,b,0) \overset{(3)}{=} M(a,b,ab), \qquad M(0,b,0)M(a,0,0) \overset{(4)}{=} M(a,b,0),$$
$$\Rightarrow \quad M(a,0,0)\,M(0,b,0) \overset{(5)}{=} M(0,0,ab)\;M(0,b,0)\,M(a,0,0).$$

(3) The product rule with $a' = 0$, $b = 0$ renamed: the extra term is $a \cdot b$.
(4) Now the first factor has $a = 0$, so the extra term vanishes.
(5) $M(0,0,ab)$ is central and adds ab to the corner entry.

Two commuting translations of $\mathbb{R}^2$ have lifts that commute only up to a central element. That is what a central extension does. In quantum mechanics this is the canonical commutation relation, as the next subsection shows.

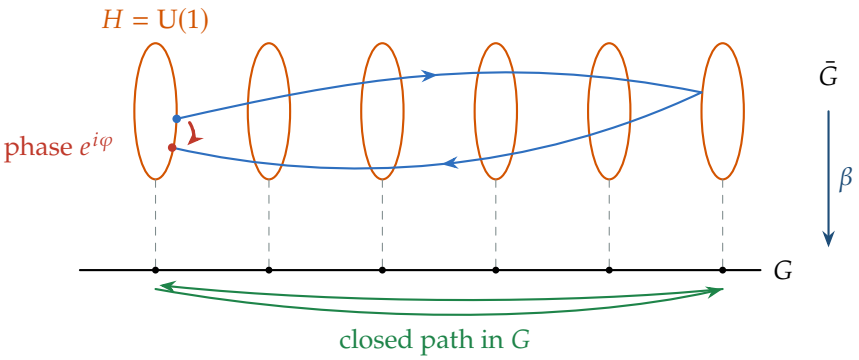


Figure 6.12: A central extension by U(1) pictured as a circle bundle: over each point of G sits a circle of elements of $\bar{G}$ (orange), and β projects each circle to its base point. A closed path in G (green) lifts to a path in $\bar{G}$ (blue) that need not close: it can return to the same fibre rotated by a phase (red). In quantum mechanics this phase is the projective multiplier of the symmetry operators.

6.2.2 Projective representations in quantum mechanics

Why should a physicist expect central extensions? Because of the way symmetries act on quantum states (Wigner). A physical state is not a vector ψ but a *ray*, $\{e^{i\varphi}\psi\}$: all measurable quantities, the transition probabilities $\left|\langle\phi|\psi\rangle\right|^2$, are blind to the phase. A symmetry g is therefore represented by a unitary operator T_g that is determined only up to a phase. When we compose two symmetries, the operators need only agree up to a phase:

Working definition 6.11 (projective representation)

A **projective representation** of a group G is an assignment $g \mapsto T_g$ of unitary operators with

$$T_g\,T_h = c(g,h)\,T_{gh}, \qquad |c(g,h)| = 1. \tag{6.36}$$

This is Nash's (6.56). The function $c : G \times G \to \mathrm{U}(1)$ [U(1) is the set of phases $e^{i\varphi}$] is the **multiplier**. *Operationally:* multiply the operators for g and h and compare with the operator for gh; the ratio is a phase, and that phase is $c(g,h)$.

The multiplier is not arbitrary. Operator multiplication is associative, so the two ways of computing $T_g T_h T_k$ must agree:

$$\begin{aligned} (T_g T_h)\,T_k &\overset{(1)}{=} c(g,h)\,T_{gh}T_k \overset{(2)}{=} c(g,h)\,c(gh,k)\,T_{ghk}, \\ T_g\,(T_h T_k) &\overset{(3)}{=} c(h,k)\,T_g T_{hk} \overset{(4)}{=} c(h,k)\,c(g,hk)\,T_{ghk}, \\ \Longrightarrow \quad c(g,h)\,c(gh,k) &\overset{(5)}{=} c(h,k)\,c(g,hk). \end{aligned} \tag{6.37}$$

(1) Equation (6.36) for the pair (g,h).
(2) Equation (6.36) for the pair (gh,k).
(3) Equation (6.36) for the pair (h,k).
(4) Equation (6.36) for the pair (g,hk).
(5) Associativity of operator products and of the group law; T_{ghk} is invertible, so the coefficients must match.

Equation (6.37) is the **cocycle condition**, and a function obeying it is a **group 2-cocycle** (QFT notes, (2.7.5); there the phases are written $e^{i\phi}$ and the condition is additive modulo 2π).

Some multipliers are fake. If we rephase the operators, $T'_g = \lambda(g)T_g$ with $|\lambda(g)| = 1$, then

$$T'_g T'_h \overset{(1)}{=} \lambda(g)\lambda(h)\,c(g,h)\,T_{gh} \overset{(2)}{=} \frac{\lambda(g)\lambda(h)}{\lambda(gh)}\,c(g,h)\,T'_{gh}. \tag{6.38}$$

(1) Definition of T' and (6.36).
(2) $T_{gh} = \lambda(gh)^{-1}T'_{gh}$.

So c and $c'(g,h) = c(g,h)\lambda(g)\lambda(h)/\lambda(gh)$ describe the same physics. A multiplier of the form $\lambda(g)\lambda(h)/\lambda(gh)$ is a **coboundary**; it can be removed by rephasing. The genuinely projective representations are classified by cocycles modulo coboundaries, the second cohomology group $H^2(G;\mathrm{U}(1))$.

Key idea: a projective representation is a true representation of a central extension

Given a multiplier c, define the group $\bar{G} = \mathrm{U}(1) \times G$ with product

$$(\lambda, g)\cdot(\mu, h) \coloneqq \big(\lambda\mu\,c(g,h),\ gh\big). \tag{6.39}$$

It is associative precisely because of the cocycle condition (6.37), and $\mathrm{U}(1) = \{(\lambda, e)\}$ is central. The operators $\bar{T}_{(\lambda,g)} \coloneqq \lambda T_g$ then obey $\bar{T}_{(\lambda,g)}\bar{T}_{(\mu,h)} = \lambda\mu\,c(g,h)T_{gh} = \bar{T}_{(\lambda,g)(\mu,h)}$: a true representation of $\bar{G}$. So *projective representations of G are the same as true representations of central extensions of G by* U(1). This is the sentence "Of course a projective representation of G is a true representation of the extension $\bar{G}$" of Nash, p. 186.

Let us check associativity of (6.39) once, since everything rests on it:

$$\begin{aligned} \big((\lambda,g)(\mu,h)\big)(\nu,k) &\overset{(1)}{=} \big(\lambda\mu\nu\,c(g,h)\,c(gh,k),\ ghk\big), \\ (\lambda,g)\big((\mu,h)(\nu,k)\big) &\overset{(2)}{=} \big(\lambda\mu\nu\,c(h,k)\,c(g,hk),\ ghk\big), \end{aligned}$$

(1) Apply (6.39) twice, left pair first.

(2) Apply (6.39) twice, right pair first.

and the two agree by (6.37).

There are exactly two ways in which a representation can be intrinsically projective (QFT notes, §2.7): *topologically*, because G is not simply connected, so that a phase can build up around a non-contractible loop; and *algebraically*, because the Lie algebra has central terms that cannot be removed. The three examples below show both.

Worked example 6.20: spin $\frac{1}{2}$: a topological phase

The spin-$\frac{1}{2}$ operators $U(\hat{n},\phi) = \exp(-i\phi\,\hat{n}\cdot\vec{\sigma}/2)$ represent rotations. For the rotation by $\phi = 2\pi$ about any axis,

$$U(\hat{n},2\pi) \overset{(1)}{=} \cos\pi\;\mathbb{1} - i\sin\pi\;\hat{n}\cdot\vec{\sigma} \overset{(2)}{=} -\mathbb{1},$$

(1) $(\hat{n}\cdot\vec{\sigma})^2 = \mathbb{1}$, so the exponential series splits into cosine and sine.
(2) $\cos\pi = -1$, $\sin\pi = 0$.

while in SO(3) the rotation by 2π is the identity. So, as operators on spinors, $T_{R_1}T_{R_2} = \pm T_{R_1R_2}$, with the sign depending on how the path from 1 to R_1R_2 winds around SO(3): $c(g,h) \in \{\pm 1\}$. The phase is removed by passing to the extension SU(2) of Example 6.17, the universal cover. At the Lie algebra level nothing happens: $\mathfrak{su}(2) = \mathfrak{so}(3)$ has no central terms (see Example 6.24 below).

Worked example 6.21: the Galilean group: mass as a central charge

In non-relativistic quantum mechanics a translation by $\vec{A}$ is $e^{-i\vec{P}\cdot\vec{A}}$ and a boost by $\vec{v}$ is $e^{-i\vec{K}\cdot\vec{v}}$. Classically boosts and translations commute (both just shift $\vec{x}$). Quantum mechanically the Galilean algebra contains (QFT notes, (2.4.34))

$$[K_i, P_j] = -iM\,\delta_{ij},$$

where M is the mass operator, which commutes with everything. For operators X, Y whose commutator is central, the Baker–Campbell–Hausdorff formula stops after one term, $e^X e^Y = e^{X+Y+\frac{1}{2}[X,Y]}$. With $X = -i\vec{K}\cdot\vec{v}$ and $Y = -i\vec{P}\cdot\vec{A}$,

$$[X,Y] \overset{(1)}{=} -v_iA_j\,[K_i,P_j] \overset{(2)}{=} -v_iA_j(-iM\delta_{ij}) = iM\,\vec{v}\cdot\vec{A},$$
$$e^{-i\vec{K}\cdot\vec{v}}\,e^{-i\vec{P}\cdot\vec{A}} \overset{(3)}{=} e^{\frac{i}{2}M\vec{v}\cdot\vec{A}}\;e^{-i(\vec{K}\cdot\vec{v}+\vec{P}\cdot\vec{A})}.$$

(1) Bilinearity of the commutator; $(-i)^2 = -1$.
(2) The Galilean commutator.
(3) BCH with a central commutator (QFT notes, (2.4.35)).

The product of a boost and a translation equals the "combined" transformation only up to the phase $e^{\frac{i}{2}M\vec{v}\cdot\vec{A}}$. This cannot be removed by shifting $\vec{K}$ or $\vec{P}$ by constants, because in the Galilean algebra $[K_i,P_j]$ would otherwise be zero, and a shift of generators cannot create a nonzero commutator from a vanishing one. So the Galilean group acts projectively, and the mass is a central charge. (The physical consequence is Bargmann's superselection rule: one cannot superpose states of different mass.)

Worked example 6.22: the canonical commutation relation

Let x and $p = -i\partial_x$ act on wavefunctions $\psi(x)$ (with $\hbar = 1$), so $[x, p] = i$. The translations $U(a) = e^{-iap}$, $U(a)\psi(x) = \psi(x-a)$, and the momentum kicks $V(b) = e^{ibx}$ are two commuting copies of $\mathbb{R}$ as far as their classical action on phase space is concerned. As operators,

$$V(b)U(a)\psi(x) \overset{(1)}{=} e^{ibx}\psi(x-a),$$
$$U(a)V(b)\psi(x) \overset{(2)}{=} e^{ib(x-a)}\psi(x-a) \implies V(b)U(a) = e^{iab}\,U(a)V(b).$$

(1) First translate, then multiply by e^{ibx}.
(2) First multiply, then translate: the phase factor is also shifted.

So $T(a,b) := V(b)U(a)$ is a projective representation of the abelian group $\mathbb{R}^2$, with multiplier $T(a,b)T(a',b') = e^{-iab'}T(a+a', b+b')$ (move $U(a)$ past $V(b')$). The operators $\rho(a,b,c) = e^{-ic}V(b)U(a)$ form a *true* representation of the Heisenberg group of Example 6.19: the phase $e^{-iab'}$ supplies the extra ab' in its product law. The Schrödinger representation is the Heisenberg group in disguise.

Dictionary: projective representations

Mathematics	Physics
ray, projective Hilbert space	physical state up to phase
multiplier $c(g,h)$, 2-cocycle	phase in $T_gT_h = c\,T_{gh}$
coboundary $\lambda(g)\lambda(h)/\lambda(gh)$	phase removable by rephasing T_g
$H^2(G;\mathrm{U}(1)) \neq 0$	intrinsically projective symmetry
central extension $\bar{G}$	enlarged symmetry group acting linearly
central element of $\bar{\mathfrak{g}}$	c-number in a commutator: M, $\hbar$, k, c

6.2.3 Lie algebra extensions and 2-cocycles

Near the identity, a central extension of a Lie group by U(1) becomes a central extension of its Lie algebra by $\mathbb{R}$, and the multiplier becomes a c-number in the commutators. Let us set this up directly for Lie algebras, with a general abelian $\mathfrak{h}$ in place of $\mathbb{R}$.

As a *vector space* the extended algebra is a direct sum, $\bar{\mathfrak{g}} = \mathfrak{g} \oplus \mathfrak{h}$. We write its elements as pairs (U, V) with $U \in \mathfrak{g}$, $V \in \mathfrak{h}$. All the content lies in the bracket. For a central extension, $\mathfrak{h}$ commutes with everything, and the bracket of two elements of $\mathfrak{g}$ is the old bracket plus something in $\mathfrak{h}$:

$$\big[(U_1,V_1),(U_2,V_2)\big] = \big([U_1,U_2],\ c(U_1,U_2)\big), \tag{6.40}$$

Nash's (6.28), with $c : \mathfrak{g}\times\mathfrak{g} \to \mathfrak{h}$ bilinear. Two conditions make this a Lie bracket.

Antisymmetry. The left side changes sign when we swap the two entries, so $c(U_2,U_1) = -c(U_1,U_2)$.

Jacobi identity. Compute a double bracket:

$$\big[(U_1,V_1),[(U_2,V_2),(U_3,V_3)]\big] \overset{(1)}{=} \big[(U_1,V_1),\big([U_2,U_3],c(U_2,U_3)\big)\big]$$
$$\overset{(2)}{=} \Big([U_1,[U_2,U_3]],\ c\big(U_1,[U_2,U_3]\big)\Big).$$

(1) The bracket (6.40) of the inner pair.
(2) The bracket (6.40) again; the $\mathfrak{h}$ entries V_1 and $c(U_2,U_3)$ never appear on the right, because $\mathfrak{h}$ is central.

Summing over cyclic permutations, the $\mathfrak{g}$ entries add up to zero by the Jacobi identity of $\mathfrak{g}$. The $\mathfrak{h}$ entries must also add up to zero:

$$\boxed{c\big([U,V],W\big)+c\big([V,W],U\big)+c\big([W,U],V\big)=0} \tag{6.41}$$

(Nash's (6.29); we used antisymmetry to write $c(U_1,[U_2,U_3])=-c([U_2,U_3],U_1)$.)

Working definition 6.12 (Lie algebra 2-cocycle)

A **2-cocycle** on $\mathfrak{g}$ with values in $\mathfrak{h}$ [the abelian algebra of the central terms. For an extension by U(1), $\mathfrak{h}=\mathbb{R}$ and $c(U,V)$ is a number] is a bilinear [linear in each argument separately], antisymmetric map $c:\mathfrak{g}\times\mathfrak{g}\to\mathfrak{h}$ satisfying the cyclic condition (6.41). *Operationally:* it is the c-number you are allowed to add to the commutators of $\mathfrak{g}$ without spoiling the Jacobi identity.

Watch out: what is and is not required of c

Nash calls c "a G-invariant anti-symmetric form". Antisymmetry and (6.41) are the requirements. Invariance under the adjoint action of G is an extra property that the cocycles we care about happen to have, and that can always be arranged for compact G by averaging; it is not part of the definition.

Trivial cocycles

Some cocycles only look like extensions. Suppose we choose new representatives of $\mathfrak{g}$ inside $\bar{\mathfrak{g}}$, shifting each U by a c-number, $\tilde{U}:=(U,\lambda(U))$ with $\lambda:\mathfrak{g}\to\mathfrak{h}$ linear. Then

$$\begin{aligned}
[\tilde{U}_1,\tilde{U}_2] &\overset{(1)}{=} \big([U_1,U_2],\ c(U_1,U_2)\big)\\
&\overset{(2)}{=} \big([U_1,U_2],\ \lambda([U_1,U_2])\big)+\big(0,\ c(U_1,U_2)-\lambda([U_1,U_2])\big)\\
&\overset{(3)}{=} \widetilde{[U_1,U_2]}+\big(0,\ c'(U_1,U_2)\big), \qquad c'(U_1,U_2):=c(U_1,U_2)-\lambda\big([U_1,U_2]\big).
\end{aligned} \tag{6.42}$$

(1) The bracket (6.40); the shifts $\lambda(U_i)$ are central and drop out.
(2) Add and subtract $\lambda([U_1,U_2])$ in the $\mathfrak{h}$ slot.
(3) Definition of $\widetilde{\ }$ applied to $[U_1,U_2]$.

In the new basis the cocycle is $c'=c-\lambda\circ[\,,\,]$. A cocycle of the form $c(U_1,U_2)=\lambda([U_1,U_2])$ is a **coboundary**: it can be removed completely by shifting generators by constants. This is the Lie-algebra version of rephasing (6.38); in the QFT notes it is (2.7.22)–(2.7.24).

Key idea: extensions are classified by $H^2(\mathfrak{g};\mathfrak{h})$

$$H^2(\mathfrak{g};\mathfrak{h})=\frac{\{\text{2-cocycles}\}}{\{\text{coboundaries } \lambda\circ[\,,\,]\}}.$$

Nontrivial central extensions of $\mathfrak{g}$ by $\mathfrak{h}$ correspond to nonzero classes. A class is a c-number term in the commutators that survives every redefinition of the generators.

This is a genuine cohomology, with a differential, as in de Rham theory (Chapter T2). The **Chevalley–Eilenberg** differential sends a p-cochain (an antisymmetric p-linear map) to a $(p+1)$-cochain; in the lowest degrees, with trivial action on $\mathfrak{h}$,

$$(\delta\lambda)(U,V)=-\lambda([U,V]), \qquad (\delta c)(U,V,W)=-c([U,V],W)-c([V,W],U)-c([W,U],V).$$

So "cocycle" means $\delta c = 0$, which is (6.41), and "coboundary" means $c = \delta(-\lambda)$. Every coboundary is a cocycle:

$$\delta(\delta\lambda)(U,V,W) \overset{(1)}{=} \lambda([[U,V],W]) + \lambda([[V,W],U]) + \lambda([[W,U],V]) \overset{(2)}{=} \lambda(0) = 0.$$

(1) Insert $\delta\lambda$ into the formula for δc; the two minus signs cancel.
(2) λ is linear, and the three double brackets add to zero by the Jacobi identity of $\mathfrak{g}$.

So $\delta^2 = 0$, exactly like $\mathrm{d}^2 = 0$ for forms, and H^2 makes sense.

Worked example 6.23: the Heisenberg algebra: every cocycle is nontrivial

Let $\mathfrak{g} = \mathbb{R}^2$ with basis X, P and the zero bracket (the Lie algebra of the abelian group $\mathbb{R}^2$). Since $[\,,\,] = 0$, the coboundaries $\lambda([U,V])$ all vanish, and (6.41) holds trivially. So every antisymmetric bilinear form is a nontrivial cocycle. On $\mathbb{R}^2$ there is one up to scale, $c(X,P) = -c(P,X) = \hbar$, and it gives

$$[X,P] = \hbar Z, \qquad [X,Z] = [P,Z] = 0,$$

with Z the central generator. In a representation where Z acts as i this is $[x,p] = i\hbar$: *the canonical commutation relation is a central extension of the abelian algebra of phase-space translations*, and $H^2(\mathbb{R}^2;\mathbb{R}) = \mathbb{R}$. The Lie group is the Heisenberg group of Example 6.19. With n degrees of freedom one gets the $(2n+1)$-dimensional Heisenberg algebra; the free boson of Section 6.2.9 is the case $n = \infty$.

Worked example 6.24: no central extensions of $\mathfrak{su}(2)$

Take the physicists' basis $T^a = \sigma^a/2$ with $[T^a, T^b] = i\varepsilon^{abc}T^c$, and try to add central terms: $[T^a,T^b] = i\varepsilon^{abc}T^c + iC^{ab}\mathbb{1}$ with $C^{ab} = -C^{ba}$ real. Every antisymmetric 3×3 matrix can be written $C^{ab} = \varepsilon^{abc}\mu^c$ for a vector μ (namely $\mu^c = \frac{1}{2}\varepsilon^{abc}C^{ab}$). Now shift the generators, $\tilde{T}^a = T^a + \mu^a\mathbb{1}$:

$$[\tilde{T}^a,\tilde{T}^b] \overset{(1)}{=} [T^a,T^b] \overset{(2)}{=} i\varepsilon^{abc}T^c + i\varepsilon^{abc}\mu^c \overset{(3)}{=} i\varepsilon^{abc}\big(\tilde{T}^c - \mu^c\big) + i\varepsilon^{abc}\mu^c = i\varepsilon^{abc}\tilde{T}^c.$$

(1) Multiples of $\mathbb{1}$ commute with everything.
(2) The assumed extended bracket with $C^{ab} = \varepsilon^{abc}\mu^c$.
(3) $T^c = \tilde{T}^c - \mu^c$.

The central term is gone. We did not even need the cocycle condition: for $\mathfrak{su}(2)$ every antisymmetric C is a coboundary, because $[\,,\,]$ is onto. So $H^2(\mathfrak{su}(2);\mathbb{R}) = 0$, and angular momentum admits no c-number in its commutators.

The same is true for every semisimple $\mathfrak{g}$ (**Whitehead's lemma**: $H^2(\mathfrak{g};\mathbb{R}) = 0$). The argument is short enough to give. Let K be the Killing form, which is nondegenerate and invariant, $K([A,B],C) = K(A,[B,C])$. Given a cocycle c, nondegeneracy lets us write $c(U,V) = K(DU,V)$ for a unique linear map $D : \mathfrak{g} \to \mathfrak{g}$, and antisymmetry of c says $K(DU,V) = -K(U,DV)$. The cocycle condition then says that D is a *derivation*:

$$\begin{aligned} 0 &\overset{(1)}{=} K\big(D[U,V],W\big) + K\big(D[V,W],U\big) + K\big(D[W,U],V\big) \\ &\overset{(2)}{=} K\big(D[U,V],W\big) - K\big([DU,V],W\big) - K\big([U,DV],W\big). \end{aligned}$$

(1) The cocycle condition (6.41) with $c(A,B) = K(DA,B)$.

(2) Second term: $K(D[V,W],U) = -K([V,W],DU) = -K(DU,[V,W]) = -K([DU,V],W)$, using antisymmetry of D, symmetry of K and invariance. Third term: $K(D[W,U],V) = -K(DV,[W,U]) = K(DV,[U,W]) = K([DV,U],W) = -K([U,DV],W)$ in the same way.

Since W is arbitrary, $D[U,V] = [DU,V] + [U,DV]$. For semisimple $\mathfrak{g}$ every derivation is inner, $D = \mathrm{ad}_X$ for some $X \in \mathfrak{g}$. Then

$$c(U,V) = K([X,U],V) = K(X,[U,V]) = \lambda([U,V]), \qquad \lambda := K(X,\cdot),$$

a coboundary. *Finite-dimensional semisimple Lie algebras have no central extensions*: every projective representation of a semisimple Lie group comes from topology (the universal cover), never from the algebra. The loop algebra escapes this conclusion, because the Killing form of $\mathrm{Map}(S^1,\mathfrak{g})$ is not a nondegenerate form of the right kind and there are outer derivations, the most important one being $\frac{d}{d\theta}$. We now exhibit its cocycle.

6.2.4 The Kac–Moody cocycle

The loop algebra $\mathrm{Map}(S^1,\mathfrak{g})$ has elements $\xi(\theta) = \xi_a(\theta)T^a$ (Hermitian T^a, real periodic ξ_a, or complex ones in the complexification), with the pointwise bracket $[\xi,\eta](\theta) = [\xi(\theta),\eta(\theta)]$. We now have an outer derivation to play with, the derivative $\xi \mapsto \xi' = \frac{d\xi}{d\theta}$, and an invariant trace. Combine them.

Working definition 6.13 (Kac–Moody cocycle)

For $\xi,\eta \in \mathrm{Map}(S^1,\mathfrak{g})$ put

$$\omega_k(\xi,\eta) := \frac{k}{\pi i}\int_0^{2\pi} \mathrm{tr}\big(\xi'(\theta)\,\eta(\theta)\big)\,\mathrm{d}\theta = \frac{k}{2\pi i}\int_0^{2\pi} \xi_a'(\theta)\,\eta_a(\theta)\,\mathrm{d}\theta, \tag{6.43}$$

where the second form uses $\mathrm{tr}(T^aT^b) = \frac{1}{2}\delta^{ab}$. The Kac–Moody algebra (6.12) is the extension (6.40) [$\omega_k(\xi,\eta)$ is added to the bracket as a c-number that commutes with everything] of the loop algebra by this cocycle [the constant k is the level appearing in (6.12)]: $[\xi,\eta]_{\mathrm{KM}} = [\xi,\eta] + \omega_k(\xi,\eta)$. *Operationally:* to find the central term of a commutator, differentiate the first loop, multiply by the second, take the trace and integrate round the circle.

We check four things: that ω_k reproduces the central term of (6.12), that it is antisymmetric, that it is a cocycle, and that it is not a coboundary.

(1) The central term. For $\xi = T^a_m = e^{im\theta}T^a$ and $\eta = T^b_n = e^{in\theta}T^b$ we have $\xi_c = \delta_{ac}e^{im\theta}$, so

$$\omega_k(T^a_m,T^b_n) \overset{(1)}{=} \frac{k}{2\pi i}\int_0^{2\pi} (im)\,e^{im\theta}\,e^{in\theta}\,\delta^{ab}\,\mathrm{d}\theta \overset{(2)}{=} \frac{k}{2\pi i}\,(im)\,2\pi\,\delta_{m+n,0}\,\delta^{ab} = k\,m\,\delta^{ab}\,\delta_{m+n,0}. \tag{6.44}$$

(1) The second form of (6.43); $\frac{d}{d\theta}e^{im\theta} = ime^{im\theta}$.

(2) $\int_0^{2\pi} e^{i(m+n)\theta}\mathrm{d}\theta = 2\pi\delta_{m+n,0}$.

This is exactly the central term of Nash's (6.9).

(2) Antisymmetry. Integrate by parts on the circle, where there is no boundary term:

$$\omega_k(\eta,\xi) \overset{(1)}{=} \frac{k}{\pi i}\int_0^{2\pi} \mathrm{tr}(\eta'\xi)\,\mathrm{d}\theta \overset{(2)}{=} -\frac{k}{\pi i}\int_0^{2\pi} \mathrm{tr}(\eta\,\xi')\,\mathrm{d}\theta \overset{(3)}{=} -\omega_k(\xi,\eta).$$

(1) Definition.

(2) $\int_0^{2\pi} \frac{d}{d\theta}\operatorname{tr}(\eta\xi)\mathrm{d}\theta = 0$ by periodicity, and $\frac{d}{d\theta}\operatorname{tr}(\eta\xi) = \operatorname{tr}(\eta'\xi) + \operatorname{tr}(\eta\xi')$.
(3) Cyclicity of the trace, $\operatorname{tr}(\eta\xi') = \operatorname{tr}(\xi'\eta)$.

(3) The cocycle condition. We need the cyclic sum of $\omega_k([\xi,\eta],\zeta)$ to vanish. Two facts about traces are used repeatedly: $\operatorname{tr}([A,B]C) = \operatorname{tr}(A[B,C])$ (expand and use cyclicity), and the product rule $[\xi,\eta]' = [\xi',\eta] + [\xi,\eta']$. Then

$$\begin{aligned}\operatorname{tr}\big([\xi,\eta]'\,\zeta\big) &\overset{(1)}{=} \operatorname{tr}\big([\xi',\eta]\zeta\big) + \operatorname{tr}\big([\xi,\eta']\zeta\big)\\ &\overset{(2)}{=} \operatorname{tr}\big(\xi'[\eta,\zeta]\big) + \operatorname{tr}\big(\eta'[\zeta,\xi]\big).\end{aligned}$$

(1) Product rule for the pointwise bracket.
(2) First term: $\operatorname{tr}([A,B]C) = \operatorname{tr}(A[B,C])$ with $A = \xi'$. Second term: use $[\xi,\eta'] = -[\eta',\xi]$, then $\operatorname{tr}([A,B]C) = \operatorname{tr}(A[B,C])$ again with $A = \eta'$, then antisymmetry of the bracket once more, to get $\operatorname{tr}(\eta'[\zeta,\xi])$.

Now sum over the three cyclic permutations $(\xi,\eta,\zeta) \to (\eta,\zeta,\xi) \to (\zeta,\xi,\eta)$. The second term on the right is the first term with (ξ,η,ζ) shifted cyclically, so both give the same cyclic sum $S := \operatorname{tr}(\xi'[\eta,\zeta]) + \operatorname{tr}(\eta'[\zeta,\xi]) + \operatorname{tr}(\zeta'[\xi,\eta])$, and

$$\sum_{\text{cyc}} \omega_k\big([\xi,\eta],\zeta\big) \overset{(1)}{=} \frac{k}{\pi i}\int_0^{2\pi} 2S\,\mathrm{d}\theta \overset{(2)}{=} \frac{2k}{\pi i}\int_0^{2\pi} \frac{d}{d\theta}\operatorname{tr}\big(\xi[\eta,\zeta]\big)\,\mathrm{d}\theta \overset{(3)}{=} 0. \tag{6.45}$$

(1) Definition (6.43) and the computation above.
(2) Product rule: $\frac{d}{d\theta}\operatorname{tr}(\xi[\eta,\zeta]) = \operatorname{tr}(\xi'[\eta,\zeta]) + \operatorname{tr}(\xi[\eta',\zeta]) + \operatorname{tr}(\xi[\eta,\zeta'])$, and by $\operatorname{tr}(A[B,C]) = \operatorname{tr}(B[C,A]) = \operatorname{tr}(C[A,B])$ the last two terms are $\operatorname{tr}(\eta'[\zeta,\xi])$ and $\operatorname{tr}(\zeta'[\xi,\eta])$. So the derivative is exactly S.
(3) The integral of a derivative of a periodic function over a period: the fundamental theorem of calculus on a circle, i.e. Stokes' theorem with $\partial S^1 = \varnothing$.

So the Jacobi identity of the Kac–Moody algebra is, at bottom, Stokes' theorem on the circle.

(4) Nontriviality. Suppose $\omega_k = \lambda \circ [\,,\,]$ for some linear λ. Take $\xi = T^3_m$ and $\eta = T^3_{-m}$ with $m \neq 0$. These commute in the loop algebra, $[T^3_m, T^3_{-m}] = if^{33c}T^c_0 = 0$, so $\lambda([\xi,\eta]) = 0$. But $\omega_k(\xi,\eta) = km \neq 0$ by (6.44). Contradiction. So for $k \neq 0$ the extension is nontrivial: no redefinition of generators removes the Kac–Moody central term.

Worked example 6.25: the Jacobi identity in mode form

It is instructive to see (3) directly in the mode basis. Take three generators T^a_l, T^b_m, T^c_n. In $[T^a_l,[T^b_m,T^c_n]]$ the central term comes from the bracket of T^a_l with the $if^{bcd}T^d_{m+n}$ part of the inner bracket:

$$\big[T^a_l,[T^b_m,T^c_n]\big]\Big|_{\text{central}} \overset{(1)}{=} if^{bcd}\,\big[T^a_l,T^d_{m+n}\big]\Big|_{\text{central}} \overset{(2)}{=} if^{bcd}\,k\,l\,\delta^{ad}\,\delta_{l+m+n,0} = ik\,l\,f^{bca}\,\delta_{l+m+n,0}.$$

(1) The inner bracket is $if^{bcd}T^d_{m+n}$ + (central), and a central element commutes with T^a_l.
(2) The central term of (6.12).

Adding the two cyclic permutations, and using $f^{bca} = f^{cab} = f^{abc}$,

$$ik\,\delta_{l+m+n,0}\big(l\,f^{bca} + m\,f^{cab} + n\,f^{abc}\big) = ik\,f^{abc}\,(l+m+n)\,\delta_{l+m+n,0} = 0.$$

The central terms cancel because the mode numbers add up to zero whenever the Kronecker delta is nonzero. (The non-central parts cancel by the Jacobi identity of $\mathfrak{g}$.)

Worked example 6.26: $G = \mathrm{U}(1)$: infinitely many oscillators

For $G = \mathrm{U}(1)$ there is one generator, $T_n = e^{in\theta}$, and $f = 0$. The loop algebra is abelian, and the Kac–Moody algebra is

$$[T_m, T_n] = k\, m\, \delta_{m+n,0}.$$

In a unitary representation with $T_n^\dagger = T_{-n}$ and $k > 0$, set $a_n := T_n/\sqrt{kn}$ for $n > 0$. Then

$$[a_m, a_n^\dagger] \overset{(1)}{=} \frac{[T_m, T_{-n}]}{k\sqrt{mn}} \overset{(2)}{=} \frac{k\, m\, \delta_{m,n}}{k\, m} = \delta_{mn}, \qquad [a_m, a_n] = 0 \quad (m, n > 0).$$

(1) Definition of a_n and $a_n^\dagger = T_{-n}/\sqrt{kn}$.
(2) The central term, with $\delta_{m-n,0}$ forcing $n = m$.

One independent harmonic oscillator for each $n = 1, 2, 3, \ldots$: the U(1) Kac–Moody algebra is the infinite-dimensional Heisenberg algebra of Example 6.23, i.e. the mode algebra of a chiral free boson. We return to it in Section 6.2.9.

Two further properties of ω_k are worth recording. It is invariant under constant gauge rotations, $\omega_k(g\xi g^{-1}, g\eta g^{-1}) = \omega_k(\xi, \eta)$, by cyclicity of the trace; this is the invariance Nash mentions. And it is invariant under rigid rotations of the circle, $\theta \mapsto \theta + a$, because $\mathrm{d}\theta$ is. It is *not* invariant under general diffeomorphisms of the circle, and that is the germ of the Virasoro algebra inside the Kac–Moody algebra (Section 6.2.10).

6.2.5 Cohomology of groups and of their Lie algebras

Nash next relates the Lie algebra cohomology $H^*(\mathfrak{g}; \mathbb{R})$ to the de Rham cohomology of the group. The bridge is the space of left-invariant forms.

Left-invariant forms

A p-form ω on G is **left invariant** if $L_h^* \omega = \omega$ for every left translation $L_h : g \mapsto hg$. Such a form is determined by its value at the identity, which is an antisymmetric p-linear form on $T_1 G = \mathfrak{g}$. So

$$\{\text{left-invariant } p\text{-forms on } G\} \;\cong\; \{\text{antisymmetric } p\text{-linear forms on } \mathfrak{g}\}.$$

The basic example is the **Maurer–Cartan form** $\theta = g^{-1}\mathrm{d}g$, a $\mathfrak{g}$-valued left-invariant 1-form: $(hg)^{-1}\mathrm{d}(hg) = g^{-1}\mathrm{d}g$. It satisfies the Maurer–Cartan equation

$$\mathrm{d}\theta \overset{(1)}{=} \mathrm{d}(g^{-1}) \wedge \mathrm{d}g \overset{(2)}{=} -g^{-1}\mathrm{d}g\, g^{-1} \wedge \mathrm{d}g = -\theta \wedge \theta,$$

(1) $\mathrm{d}(g^{-1}\mathrm{d}g) = \mathrm{d}(g^{-1}) \wedge \mathrm{d}g + g^{-1}\mathrm{d}\mathrm{d}g$ and $\mathrm{d}\mathrm{d}g = 0$.
(2) Differentiate $g^{-1}g = 1$: $\mathrm{d}(g^{-1}) = -g^{-1}\mathrm{d}g\, g^{-1}$.

Under the identification with forms on $\mathfrak{g}$, the exterior derivative d becomes the Chevalley–Eilenberg differential δ; the Maurer–Cartan equation is the statement that d of a left-invariant 1-form λ evaluates on U, V to $-\lambda([U, V])$, i.e. $\mathrm{d}\lambda = \delta\lambda$ (up to the convention-dependent factor in front). So left-invariant closed forms modulo left-invariant exact forms is exactly $H^*(\mathfrak{g}; \mathbb{R})$.

Key idea: for compact G, $H^*(G; \mathbb{R}) = H^*(\mathfrak{g}; \mathbb{R})$

For compact connected G every de Rham class contains a left-invariant representative: average. If ω is closed, each $L_h^* \omega$ is cohomologous to ω (since L_h can be deformed to

the identity through the path from h to 1, and homotopic maps pull back closed forms to cohomologous ones), so the average $\bar{\omega} = \int_G L_h^* \omega \, \mathrm{d}h$ (normalised Haar measure) is left-invariant and cohomologous to ω. A similar averaging shows that a left-invariant form that is exact is the derivative of a left-invariant form. Hence $H^*(G;\mathbb{R}) = H^*(\mathfrak{g};\mathbb{R})$.

Worked example 6.27: the generator of $H^3(\mathrm{SU}(2))$

The 3-form $\Omega = \mathrm{tr}(g^{-1}\mathrm{d}g)^3$ on SU(2) is left-invariant (it is built from θ) and closed (it is a top form on a 3-manifold). It is not exact, because its integral over the closed manifold $\mathrm{SU}(2) = S^3$ is nonzero: by Example 6.5, $\int_{S^3} \Omega = 24\pi^2$, and an exact form would integrate to zero by Stokes. At the identity it is the 3-cocycle on $\mathfrak{su}(2)$

$$\Omega(U,V,W) \propto \mathrm{tr}\big(U[V,W]\big),$$

the unique invariant antisymmetric trilinear form. In contrast $H^1 = H^2 = 0$: a left-invariant 1-form λ is closed iff $\lambda([U,V]) = 0$ for all U, V, but $[\mathfrak{g},\mathfrak{g}] = \mathfrak{g}$ for semisimple $\mathfrak{g}$, so $\lambda = 0$; and $H^2 = 0$ is Whitehead's lemma (Example 6.24). These agree with $H^*(S^3)$: nonzero only in degrees 0 and 3.

For LG the same statement survives, even though LG is not compact: $H^*(LG;\mathbb{R}) = H^*(\mathrm{Map}(S^1,\mathfrak{g});\mathbb{R})$ (Pressley and Segal). In particular the Kac–Moody cocycle ω_k extends to a left-invariant 2-form on LG, which we also call ω_k, and it is closed because ω_k is a cocycle. For $\mathrm{Map}(X,G)$ with $\dim X > 1$ the statement fails: there are de Rham classes on $\mathrm{Map}(X,G)$ that cannot be made left invariant, so $H^*(\mathrm{Map}(X,G);\mathbb{R})$ is smaller than $H^*(\mathrm{Map}(X,\mathfrak{g});\mathbb{R})$. This is one reason why the extensions of $\mathrm{Map}(X,G)$ in Section 6.2.12 are harder.

Watch out: topology of LG

For simply connected G, $LG \cong G \times \Omega G$ as spaces (constants times based loops, (6.4)), and $\pi_k(\Omega G) = \pi_{k+1}(G)$. So $\pi_1(LG) = \pi_1(G) \oplus \pi_2(G) = 0$ and $\pi_2(LG) = \pi_2(G) \oplus \pi_3(G) = \mathbb{Z}$ for simple G. Hence $H^2(LG;\mathbb{Z}) = \mathbb{Z}$: there is room for exactly one integer's worth of circle bundles over LG. That integer will be the level.

6.2.6 From the Lie algebra to the group: integrality

In finite dimensions every Lie algebra extension integrates to an extension of the simply connected group. In infinite dimensions this fails, and the reason is one that every physicist knows from the Dirac monopole.

A warm-up: magnetic translations and flux quantisation

A charged particle ($e = \hbar = 1$) in a uniform magnetic field B in the plane has *magnetic translation* operators $t(\vec{a}) = e^{-i\vec{a}\cdot\vec{K}}$ that commute with the Hamiltonian; their generators satisfy $[K_1, K_2] = iB$ (up to a sign convention). This is a Heisenberg algebra with central element B. By the BCH formula for central commutators,

$$\begin{aligned} t(\vec{a})\,t(\vec{b}) &\overset{(1)}{=} e^{-i(\vec{a}+\vec{b})\cdot\vec{K}}\, e^{\frac{1}{2}[-i\vec{a}\cdot\vec{K},\,-i\vec{b}\cdot\vec{K}]} \\ &\overset{(2)}{=} e^{-\frac{1}{2}a_i b_j [K_i,K_j]}\, t(\vec{a}+\vec{b}) \overset{(3)}{=} e^{-\frac{i}{2}B(a_1 b_2 - a_2 b_1)}\, t(\vec{a}+\vec{b}), \\ t(\vec{a})\,t(\vec{b}) &\overset{(4)}{=} e^{-iB(\vec{a}\times\vec{b})}\, t(\vec{b})\,t(\vec{a}). \end{aligned} \tag{6.46}$$

(1) $e^X e^Y = e^{X+Y} e^{\frac{1}{2}[X,Y]}$ for central $[X, Y]$.
(2) $(-i)^2 = -1$ and bilinearity.
(3) $a_i b_j [K_i, K_j] = (a_1 b_2 - a_2 b_1)\, iB$.
(4) Write the same for $t(\vec{b})t(\vec{a})$ and divide: the phases differ by $e^{-iB(a_1 b_2 - a_2 b_1)}$.

The phase is $e^{-i\Phi}$, with $\Phi = B\,(\vec{a} \times \vec{b})$ the magnetic flux through the parallelogram spanned by $\vec{a}$ and $\vec{b}$: the Aharonov–Bohm phase of the loop "go along $\vec{a}$, then $\vec{b}$, then back". So far this is just a Lie algebra extension, and B can be anything.

Now put the particle on a torus of side L. Translating by a full period is a symmetry that does nothing to the torus, so $t(L\hat{x})$ and $t(L\hat{y})$ must commute (each acts on the Hilbert space as a phase at most). By (6.46) this requires

$$e^{-iBL^2} = 1 \qquad \Longleftrightarrow \qquad \frac{BL^2}{2\pi} = \frac{\Phi_{\text{total}}}{2\pi} \in \mathbb{Z}.$$

The Lie algebra extension of $\mathbb{R}^2$ by B exists for all B, but it descends to a group extension of the *compact* group T^2 only if the total flux is an integer multiple of 2π: Dirac's quantisation condition.

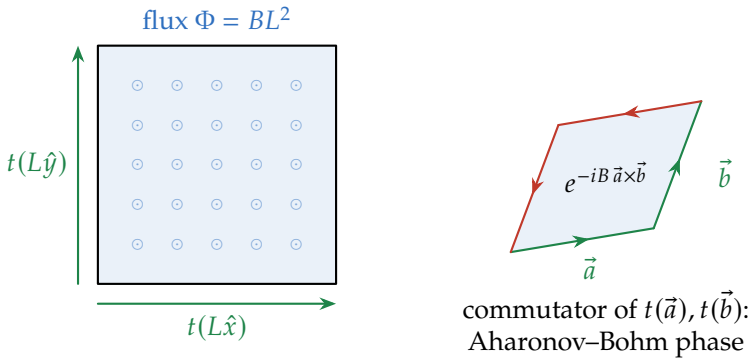


Figure 6.13: Right: two magnetic translations commute only up to the Aharonov–Bohm phase of the flux through the parallelogram they span; this is a Lie algebra cocycle (the 2-form $B\,\mathrm{d}x \wedge \mathrm{d}y$). Left: on a torus the translations by a full period must commute, so the flux through the whole torus must be $2\pi\times$ an integer. Integrating the cocycle 2-form over a closed surface must give $2\pi\mathbb{Z}$: this is the integrality condition for central extensions of groups.

The general statement

The same mechanism controls every central extension by U(1). Topologically, $\bar{G}$ is a principal U(1)-bundle over G (Figure 6.12). The Lie algebra cocycle c defines a left-invariant 2-form ω on G, and ω is the curvature of a natural connection on this bundle: the phase acquired by lifting a small closed loop in G is $e^{i\int \omega}$ over the enclosed surface, exactly as in (6.46). A U(1)-bundle with curvature ω exists if and only if the flux through every closed surface is quantised, i.e. if and only if

$$\left[\frac{\omega}{2\pi}\right] \in H^2(G;\mathbb{Z}) \quad \text{(an integral class)}, \qquad \text{i.e.} \qquad \int_\Sigma \frac{\omega}{2\pi} \in \mathbb{Z} \text{ for every closed surface } \Sigma \subset G. \tag{6.47}$$

The class $[\omega/2\pi]$ is then the first Chern class of the bundle $\bar{G} \to G$, which characterises it completely. The factor 2π is the usual normalisation of characteristic classes as differential forms (Nash's footnote 2; compare $c_1 = \frac{i}{2\pi}\operatorname{tr} F$ in Chapter T2).

Key idea: Lie algebras without groups

The Lie algebra $\mathrm{Map}(S^1,\mathfrak{g})\oplus\mathbb{R}$ with bracket $[\xi,\eta]+\omega_k(\xi,\eta)$ exists for every real k. The group extension $1\to \mathrm{U}(1)\to \widetilde{LG}\to LG\to 1$ exists only when $\omega_k/2\pi$ is an integral class on LG. Since $H^2(LG;\mathbb{Z})=\mathbb{Z}$ for simple simply connected G (the pitfall above), this is one integer condition on k. With our normalisation ($\mathrm{tr}\,T^aT^b=\frac{1}{2}\delta^{ab}$, long roots of length 1 for $\mathrm{SU}(N)$) it reads

$$N=2k\in\mathbb{Z},$$

which is the same quantisation of the level that we will find from representation theory in Section 6.3.9 (Nash's (6.62)). For non-integral N, $\mathrm{Map}(S^1,\mathfrak{g})\oplus\mathbb{R}$ is a Lie algebra that is the Lie algebra of no group.

Watch out: a class, not a form

Integrality is a condition on the cohomology class of $\omega/2\pi$, not on the form itself. Changing c by a coboundary changes ω by an exact form and does not affect (6.47). This is why the integer (the level, the Chern class, the monopole charge) is robust: it cannot change under small deformations, and it cannot be redefined away.

6.2.7 Where the c-numbers come from: counting powers of $\hbar$

So far the central terms have been an algebraic possibility. Nash now explains why a physicist should *expect* them, by keeping track of Planck's constant. In this subsection only, $\hbar\neq 1$.

Classical symmetries and Poisson brackets

In classical mechanics a symmetry group G acts on phase space, and each generator T^a of its Lie algebra is represented by a function on phase space (the conserved charge). The group structure is encoded in the Poisson brackets of these functions,

$$\{T^a,T^b\}=f^{abc}\,T^c, \tag{6.48}$$

Nash's (6.30). Here $\{A,B\}=\sum_i\left(\frac{\partial A}{\partial x_i}\frac{\partial B}{\partial p_i}-\frac{\partial A}{\partial p_i}\frac{\partial B}{\partial x_i}\right)$.

Worked example 6.28: angular momentum as Poisson brackets

Take $L_i=\varepsilon_{ijk}x_jp_k$, so $L_1=x_2p_3-x_3p_2$ and $L_2=x_3p_1-x_1p_3$. The only nonzero basic brackets are $\{x_i,p_j\}=\delta_{ij}$, and the bracket is a derivation in each slot, so

$$\begin{aligned}\{L_1,L_2\}&\overset{(1)}{=}\{x_2p_3,x_3p_1\}-\{x_2p_3,x_1p_3\}-\{x_3p_2,x_3p_1\}+\{x_3p_2,x_1p_3\}\\&\overset{(2)}{=}x_2p_1\{p_3,x_3\}-0-0+p_2x_1\{x_3,p_3\}\\&\overset{(3)}{=}-x_2p_1+x_1p_2\overset{(4)}{=}L_3.\end{aligned}$$

(1) Bilinearity.
(2) Leibniz rule; the middle two brackets contain no conjugate pair (p_3 with no x_3, and x_3 with no p_3).
(3) $\{p_3,x_3\}=-1$, $\{x_3,p_3\}=1$.
(4) $L_3=x_1p_2-x_2p_1$.

So $\{L_i,L_j\}=\varepsilon_{ijk}L_k$: (6.48) with $f^{abc}=\varepsilon^{abc}$.

Quantisation and the first correction

Dirac's rule replaces Poisson brackets by commutators, $\{A,B\} \to \frac{1}{i\hbar}[\hat{A},\hat{B}]$. Applied to (6.48),

$$[T^a, T^b] = i\hbar\, f^{abc}\, T^c, \tag{6.49}$$

which is Nash's (6.31). (Nash phrases the rule as "replace T^a by $i\hbar T^a$"; the effect is the same.) The rule is exact for angular momentum. It is not exact in general: the quantum operator for a product of x's and p's depends on how we order them, and the commutator of two quantised functions differs from $i\hbar$ times the quantised Poisson bracket by terms of higher order in $\hbar$ (Groenewold and van Hove showed that no ordering rule can remove them all). So in general

$$[T^a, T^b] = i\hbar\, f^{abc}\, T^c + O(\hbar^2), \tag{6.50}$$

Nash's (6.32). The simplest possible correction is a c-number:

$$[T^a, T^b] = i\hbar\, f^{abc}\, T^c + i\hbar^2\, c^{ab}\, \mathbb{1}, \tag{6.51}$$

which is Nash's (6.33) (he writes $c\,\delta^{ab}$; for $\mathfrak{g}$ itself c^{ab} must be antisymmetric, and in the loop algebra it becomes the $\delta^{ab} m\, \delta_{m+n,0}$ structure). This is precisely a central extension (6.40), with $\hbar^2 c^{ab}$ as the cocycle.

Worked example 6.29: an $\hbar^2$ c-number from normal ordering

Take one degree of freedom and the complex coordinate $z = (x + ip)/\sqrt{2}$, for which $\{z,\bar z\} = \frac{1}{2}\{x+ip, x-ip\} = \frac{1}{2}(-i-i) = -i$. The three quadratic functions

$$K_+ = \tfrac{1}{2}\bar z^2, \qquad K_- = \tfrac{1}{2} z^2, \qquad K_0 = \tfrac{1}{2}\bar z z$$

generate the symplectic group $\mathrm{Sp}(2,\mathbb{R})$ (squeezing and rotating phase space). Classically

$$\{K_-, K_+\} \overset{(1)}{=} \tfrac{1}{4}\,(2z)(2\bar z)\,\{z,\bar z\} \overset{(2)}{=} -i\, z\bar z = -2i\, K_0.$$

(1) Chain rule: $\{f(z), g(\bar z)\} = f'(z) g'(\bar z)\{z,\bar z\}$.
(2) $\{z,\bar z\} = -i$.

Now quantise with $\hat z = \sqrt{\hbar}\, a$, $\hat{\bar z} = \sqrt{\hbar}\, a^\dagger$, $[a, a^\dagger] = 1$ (check: $[\hat z, \hat{\bar z}] = \hbar = i\hbar\{z,\bar z\}$). Choose the normal-ordered operators $\hat K_+ = \frac{\hbar}{2} a^{\dagger 2}$, $\hat K_- = \frac{\hbar}{2} a^2$, $\hat K_0 = \frac{\hbar}{2} a^\dagger a$. Then

$$[\hat K_-, \hat K_+] \overset{(3)}{=} \tfrac{\hbar^2}{4}\,[a^2, a^{\dagger 2}] \overset{(4)}{=} \tfrac{\hbar^2}{4}\,(4a^\dagger a + 2) \overset{(5)}{=} i\hbar\,(-2i\hat K_0) + \tfrac{\hbar^2}{2}.$$

(3) Definitions.
(4) $[a^2, a^{\dagger 2}] = a[a, a^{\dagger 2}] + [a, a^{\dagger 2}]a = a\,2a^\dagger + 2a^\dagger a = 4a^\dagger a + 2$, using $[a, a^{\dagger 2}] = 2a^\dagger$ and $aa^\dagger = a^\dagger a + 1$.
(5) $\hbar^2 a^\dagger a = 2\hbar \hat K_0$ and $i\hbar(-2i) = 2\hbar$.

So the commutator is $i\hbar$ times the quantised Poisson bracket *plus the c-number* $\hbar^2/2$: exactly the pattern (6.51). Here the c-number is a coboundary: redefining $\hat K_0 \to \hat K_0 + \frac{\hbar}{4}$ (the symmetric ordering $\frac{\hbar}{4}(a^\dagger a + a a^\dagger)$) removes it, as Whitehead's lemma says it must for the semisimple $\mathfrak{sp}(2)$. The point of the example is where it comes from: the reordering needed to make an operator well defined.

In finite dimensions such c-numbers can always be absorbed. The same mechanism in field theory, where the normal-ordering constant is an *infinite* sum and cannot be absorbed, gives the loop-algebra version, Nash's (6.34):

$$[T^a_m, T^b_n] = i\hbar\, f^{abc}\, T^c_{m+n} + c\, m\, \hbar^2\, \delta^{ab}\, \delta_{m,-n}. \tag{6.52}$$

With $\hbar = 1$ and $c = k$ this is the Kac–Moody algebra (6.12).

Physical picture: the central term is a one-loop effect

The power $\hbar^2$ (one more than the classical term) is the signature of a quantum (one-loop) effect. In the free-fermion realisation below, the classical currents Poisson-commute up to the f^{abc} term; the central term comes entirely from filling the Dirac sea, which is a statement about the ground state and has no classical analogue. The same is true of the Virasoro central charge. This is why Nash calls the central term an *anomaly*: a classical symmetry that survives quantisation only projectively.

Watch out: the conformal anomaly and the critical dimension

In string theory $\mathrm{Diff}(S^1)$ (two copies of it, one for each chirality) is a *gauge* symmetry: reparametrising the string does not change the physics. Gauge symmetries must act as true representations on physical states, so the total central charge must vanish. D free bosons (the string coordinates) contribute $c = D$ (Section 6.2.9), and the Faddeev–Popov ghosts of the reparametrisation gauge contribute $c = -26$. So $c_{\text{total}} = D - 26$ vanishes only in the **critical dimension** $D = 26$ (and $D = 10$ for the superstring). Nash returns to this in Chapter 9 and Chapter 10.

6.2.8 The Virasoro cocycle from the Jacobi identity

We now derive the Virasoro algebra (6.21) instead of quoting it. The question is: *what c-numbers can be added to the Witt algebra* $[L_m, L_n] = (m-n)L_{m+n}$ *without spoiling the Jacobi identity, and which of them are genuine?* The answer comes in four steps. We add one central element Z and write the most general central extension,

$$[L_m, L_n] = (m-n)\,L_{m+n} + c(m,n)\,Z, \qquad c(n,m) = -c(m,n). \tag{6.53}$$

Step 1: the cocycle condition

For three generators the cocycle condition (6.41) reads $c([L_k,L_m],L_n) + c([L_m,L_n],L_k) + c([L_n,L_k],L_m) = 0$, where $c([L_k,L_m],L_n)$ means $(k-m)\,c(k+m,n)$. So

$$(k-m)\,c(k+m,n) + (m-n)\,c(m+n,k) + (n-k)\,c(n+k,m) = 0 \qquad \text{for all } k,m,n. \tag{6.54}$$

Step 2: only $m+n=0$ survives

Put $k = 0$ in (6.54):

$$\begin{aligned} 0 &\overset{(1)}{=} -m\,c(m,n) + (m-n)\,c(m+n,0) + n\,c(n,m) \\ &\overset{(2)}{=} -(m+n)\,c(m,n) + (m-n)\,c(m+n,0). \end{aligned} \tag{6.55}$$

(1) $(k-m)c(k+m,n) \to -m\,c(m,n)$; $(m-n)c(m+n,k) \to (m-n)c(m+n,0)$; $(n-k)c(n+k,m) \to n\,c(n,m)$.

(2) Antisymmetry, $c(n,m) = -c(m,n)$.

Now use the freedom to redefine generators. Shifting $L_p \to L'_p = L_p + \lambda_p Z$ changes the cocycle by the coboundary (6.42):

$$c'(m,n) = c(m,n) - (m-n)\,\lambda_{m+n}. \tag{6.56}$$

Choose $\lambda_p = c(p,0)/p$ for every $p \neq 0$. Then $c'(p,0) = c(p,0) - p\lambda_p = 0$ for $p \neq 0$, and (6.55) (which holds for c' as well, since c' is again a cocycle) says $(m+n)\,c'(m,n) = 0$. Hence $c'(m,n) = 0$ unless $m+n = 0$. Dropping the prime,

$$c(m,n) = h(m)\,\delta_{m+n,0}, \qquad h(-m) = -h(m), \quad h(0) = 0. \tag{6.57}$$

We have not yet used λ_0: it remains available.

Step 3: the functional equation

Put $n = -k-m$ in (6.54), so that all three terms are of the form (6.57):

$$\begin{aligned} 0 &\overset{(1)}{=} (k-m)\,h(k+m) + (2m+k)\,c(-k,k) + (-2k-m)\,c(-m,m) \\ &\overset{(2)}{=} (k-m)\,h(k+m) - (k+2m)\,h(k) + (2k+m)\,h(m). \end{aligned} \tag{6.58}$$

(1) $c(k+m,n) = c(k+m,-(k+m)) = h(k+m)$; $m-n = 2m+k$ and $m+n = -k$; $n-k = -2k-m$ and $n+k = -m$.
(2) $c(-k,k) = h(-k) = -h(k)$ and $c(-m,m) = -h(m)$.

This is a linear equation for the function h. Setting $k = 1$ gives a recursion,

$$h(m+1) = \frac{(m+2)\,h(m) - (2m+1)\,h(1)}{m-1} \qquad (m \geq 2),$$

so h is completely determined by the two numbers $h(1)$ and $h(2)$ (for $m = 1$ the equation reads $0 = 0$ and leaves $h(2)$ free). The solution space is therefore at most two-dimensional. Two solutions are easy to verify:

$$\begin{aligned} h(m) = m: &\quad (k-m)(k+m) - (k+2m)k + (2k+m)m \overset{(1)}{=} k^2 - m^2 - k^2 - 2km + 2km + m^2 = 0, \\ h(m) = m^3: &\quad (k-m)(k+m)^3 - (k+2m)k^3 + (2k+m)m^3 \\ &\overset{(2)}{=} \left(k^4 + 2k^3m - 2km^3 - m^4\right) - k^4 - 2k^3m + 2km^3 + m^4 = 0. \end{aligned}$$

(1) Expand.
(2) $(k-m)(k+m)^3 = (k-m)(k^3 + 3k^2m + 3km^2 + m^3) = k^4 + 2k^3m - 2km^3 - m^4$.

They are independent, so the general solution is

$$h(m) = a\,m^3 + b\,m. \tag{6.59}$$

Step 4: which part is genuine

The remaining freedom is λ_0, i.e. $L_0 \to L_0 + \lambda_0 Z$. By (6.56) with $n = -m$ it changes $h(m) \to h(m) - 2m\lambda_0$: it shifts b by $-2\lambda_0$ and leaves a alone. So:

- the linear term bm is a coboundary: it is an artefact of where we put the zero of L_0;
- the cubic term am^3 is *not* a coboundary: coboundaries $\lambda([L_m, L_{-m}]) = 2m\lambda_0$ are all linear in m.

Hence $H^2(\mathrm{Witt};\mathbb{C})$ is one-dimensional: the Witt algebra has exactly one central extension up to equivalence, and it is the Virasoro algebra. The conventional choice of λ_0 is $b = -a$, so

that $h(m) = a(m^3 - m)$ vanishes for $m = 0, \pm 1$ and the $\mathfrak{sl}(2)$ subalgebra of Section 6.1.7 is not extended. Calling $a = c/12$ (and letting Z act as 1), we arrive at

$$[L_m, L_n] = (m-n)L_{m+n} + \frac{c}{12}(m^3 - m)\,\delta_{m+n,0},$$

which is (6.21). The factor $\frac{1}{12}$ is a convention, chosen so that a free boson has $c = 1$; we check this in the next subsection.

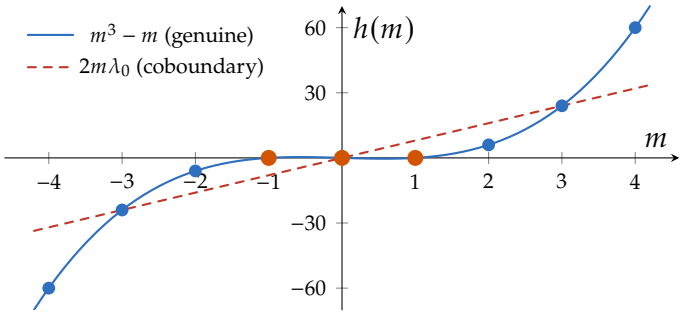


Figure 6.14: The Virasoro cocycle $h(m)$, the c-number in $[L_m, L_{-m}]$. Every central extension of the Witt algebra has $h(m) = am^3 + bm$. The linear part (red, dashed) can be changed at will by shifting L_0 by a constant, so it carries no information. The cubic part cannot. The standard choice $h = \frac{c}{12}(m^3 - m)$ (blue) vanishes at $m = 0, \pm 1$ (orange): the Möbius subalgebra $\mathfrak{sl}(2)$ is left unextended.

Worked example 6.30: the cocycle as an integral: Gel'fand–Fuks

As with the Kac–Moody cocycle (6.43), the Virasoro cocycle can be written for general vector fields $f(\theta)\partial_\theta$, $g(\theta)\partial_\theta$:

$$\omega(f\partial_\theta, g\partial_\theta) = \frac{c}{24\pi i}\int_0^{2\pi} \big(f'''(\theta) + f'(\theta)\big)\, g(\theta)\,\mathrm{d}\theta. \tag{6.60}$$

Check it on $L_m = f_m\partial_\theta$, $f_m = ie^{im\theta}$:

$$f_m''' + f_m' \overset{(1)}{=} i\big((im)^3 + im\big)e^{im\theta} \overset{(2)}{=} (m^3 - m)\,e^{im\theta},$$
$$\omega(L_m, L_n) \overset{(3)}{=} \frac{c}{24\pi i}(m^3 - m)\, i\int_0^{2\pi} e^{i(m+n)\theta}\,\mathrm{d}\theta \overset{(4)}{=} \frac{c}{12}(m^3 - m)\,\delta_{m+n,0}.$$

(1) Each derivative brings down im.
(2) $i(im)^3 = i(-i)m^3 = m^3$ and $i \cdot im = -m$.
(3) $g = f_n = ie^{in\theta}$.
(4) $\int_0^{2\pi} e^{i(m+n)\theta}\mathrm{d}\theta = 2\pi\delta_{m+n,0}$, and $\frac{c\cdot 2\pi i}{24\pi i} = \frac{c}{12}$.

Antisymmetry follows by integrating by parts: odd numbers of derivatives change sign when moved from f to g. And the operator $f \mapsto f''' + f'$ annihilates exactly $1, \cos\theta, \sin\theta$: the three real vector fields of $\mathfrak{sl}(2)$. Compare the Kac–Moody cocycle, which has one derivative and annihilates the constants: in both cases the cocycle measures how much a field fails to be "rigid".

6.2.9 Central terms from normal ordering: free fields

The cocycle condition fixes the *form* of the central terms. Their *values* are fixed by a concrete quantum theory. We compute three: the Virasoro central charge of a free boson ($c = 1$), of a

free Majorana fermion ($c = \frac{1}{2}$), and the Kac–Moody level of the charge current of a free Dirac fermion ($k = 1$). In each case the c-number comes from a single source: the ground state is not annihilated by everything, and the operators must be normal ordered.

A lemma we use three times

Suppose operators A_m on a Hilbert space satisfy: (i) the space is spanned by polynomials in some "building block" operators acting on a vacuum $|0\rangle$ (a *cyclic* vacuum); (ii) an operator D commutes with all the building blocks; (iii) there is a grading operator $L_0 \geq 0$ with $L_0|0\rangle = 0$, whose eigenvalue-0 space is spanned by $|0\rangle$, and $[L_0, D] = -q\,D$. Then $D = \langle 0|D|0\rangle\,\mathbb{1}$ if $q = 0$, and $D|0\rangle = 0$ if $q > 0$. Indeed $D\,P|0\rangle = P\,D|0\rangle$ for any polynomial P in the building blocks, so D is fixed by $D|0\rangle$; and $D|0\rangle$ has L_0-eigenvalue $-q$, which forces $D|0\rangle = 0$ for $q > 0$ and $D|0\rangle \propto |0\rangle$ for $q = 0$. We use this to reduce "compute a commutator" to "compute one vacuum expectation value".

The free boson: $c = 1$

A chiral free boson on a circle has modes α_n, $n \in \mathbb{Z}$, with

$$[\alpha_m, \alpha_n] = m\,\delta_{m+n,0}, \qquad \alpha_n^\dagger = \alpha_{-n}, \qquad \alpha_n|0\rangle = 0\ (n \geq 0). \tag{6.61}$$

This is the U(1) Kac–Moody algebra of Example 6.26 with $k = 1$ (the modes of the current $\partial\phi$). The states $\alpha_{-n_1}\cdots\alpha_{-n_r}|0\rangle$ span the Fock space. Classically the stress tensor is $\frac{1}{2}(\partial\phi)^2$, whose Fourier modes are $\frac{1}{2}\sum_m \alpha_{n-m}\alpha_m$. Quantum mechanically this is ambiguous for $n = 0$:

$$\langle 0|\tfrac{1}{2}\sum_m \alpha_{-m}\alpha_m|0\rangle \overset{(1)}{=} \tfrac{1}{2}\sum_{m>0}\langle 0|\alpha_m\alpha_{-m}|0\rangle \overset{(2)}{=} \tfrac{1}{2}\sum_{m>0} m = \infty. \tag{6.62}$$

(1) For $m \geq 0$ the right factor α_m annihilates $|0\rangle$, so those terms vanish. For $m < 0$ write $m = -m'$ with $m' > 0$: the term is $\alpha_{m'}\alpha_{-m'}$, the annihilator standing on the left. Renaming $m' \to m$ gives the sum over $m > 0$.

(2) $\alpha_m\alpha_{-m}|0\rangle = [\alpha_m, \alpha_{-m}]|0\rangle = m|0\rangle$.

This is the zero-point energy $\frac{1}{2}\hbar\omega$ of infinitely many oscillators with frequencies $\omega = m$. The remedy is **normal ordering**: $:\alpha_m\alpha_n:$ puts the annihilator (α_m with $m \geq 0$) on the right. So we define

$$L_n := \frac{1}{2}\sum_{m\in\mathbb{Z}} :\alpha_{n-m}\alpha_m:, \qquad L_0 = \frac{1}{2}\alpha_0^2 + \sum_{m>0}\alpha_{-m}\alpha_m. \tag{6.63}$$

For $n \neq 0$ normal ordering changes nothing, since α_{n-m} and α_m commute unless $n = 0$. On any Fock state only finitely many terms of the sum are nonzero, so L_n is well defined.

First commutator. For each fixed p, $[\alpha_{n-p}\alpha_p, \alpha_m] = \alpha_{n-p}\,p\,\delta_{p+m,0} + (n-p)\,\delta_{n-p+m,0}\,\alpha_p$. Summing over p and multiplying by $\frac{1}{2}$,

$$[L_n, \alpha_m] \overset{(1)}{=} \tfrac{1}{2}\big((-m)\,\alpha_{n+m} + (-m)\,\alpha_{n+m}\big) \overset{(2)}{=} -m\,\alpha_{m+n}. \tag{6.64}$$

(1) The first delta sets $p = -m$; the second sets $p = n + m$, and then $n - p = -m$. The c-number from normal ordering commutes with α_m and drops out.

(2) Add.

So L_n acts on the modes exactly as the vector field L_n acts on the function $e^{im\theta}$ (compare Section 6.2.10 below).

Second commutator. Let $D_{mn} \coloneqq [L_m, L_n] - (m-n)L_{m+n}$. Using (6.64) twice and the Jacobi identity,

$$\begin{aligned}[D_{mn}, \alpha_p] &\overset{(1)}{=} [L_m, [L_n, \alpha_p]] - [L_n, [L_m, \alpha_p]] - (m-n)[L_{m+n}, \alpha_p] \\ &\overset{(2)}{=} p(n+p)\,\alpha_{m+n+p} - p(m+p)\,\alpha_{m+n+p} + (m-n)p\,\alpha_{m+n+p} \overset{(3)}{=} 0. \end{aligned} \tag{6.65}$$

(1) Jacobi: $[[L_m, L_n], X] = [L_m, [L_n, X]] - [L_n, [L_m, X]]$.
(2) $[L_n, \alpha_p] = -p\alpha_{n+p}$, then $[L_m, -p\alpha_{n+p}] = p(n+p)\alpha_{m+n+p}$; similarly for the second term; the third is $-(m-n)(-p)\alpha_{m+n+p}$.
(3) $p(n+p-m-p+m-n) = 0$.

Also $[L_0, D_{mn}] = -(m+n)D_{mn}$, since D_{mn} is built from modes of total index $m+n$ and $[L_0, \alpha_p] = -p\alpha_p$. The lemma now says that D_{mn} is the c-number $\langle 0|D_{mn}|0\rangle$, which vanishes unless $m+n=0$ (for $m+n<0$ take adjoints: $D_{mn}^\dagger = -D_{-m,-n}$). For $n=-m$, $m>0$:

$$\langle 0|D_{m,-m}|0\rangle \overset{(1)}{=} \langle 0|L_m L_{-m}|0\rangle - \langle 0|L_{-m}L_m|0\rangle - 2m\langle 0|L_0|0\rangle \overset{(2)}{=} \| L_{-m}|0\rangle\|^2 . \tag{6.66}$$

(1) Definition of D.
(2) $L_m|0\rangle = 0$ for $m>0$ (every term of L_m has an annihilator on the right, since the two indices add to $m>0$), $L_0|0\rangle = 0$, and $L_m = L_{-m}^\dagger$.

In $L_{-m}|0\rangle = \frac{1}{2}\sum_p :\alpha_{-m-p}\alpha_p: |0\rangle$ only the terms with both indices negative survive: $-m < p < 0$. Writing $p=-q$,

$$L_{-m}|0\rangle = \tfrac{1}{2}\sum_{q=1}^{m-1} \alpha_{-(m-q)}\,\alpha_{-q}|0\rangle :$$

one term for each way of splitting m into two positive parts (Figure 6.15). Its norm is

$$\begin{aligned}\| L_{-m}|0\rangle\|^2 &\overset{(1)}{=} \tfrac{1}{4}\sum_{q,q'=1}^{m-1} \langle 0|\alpha_{q'}\alpha_{m-q'}\,\alpha_{-(m-q)}\alpha_{-q}|0\rangle \\ &\overset{(2)}{=} \tfrac{1}{4}\sum_{q,q'=1}^{m-1} q(m-q)\big(\delta_{q'q} + \delta_{q',m-q}\big) \overset{(3)}{=} \tfrac{1}{2}\sum_{q=1}^{m-1} q(m-q) \\ &\overset{(4)}{=} \tfrac{1}{2}\Big(m\,\frac{m(m-1)}{2} - \frac{(m-1)m(2m-1)}{6}\Big) \overset{(5)}{=} \frac{m^3-m}{12}. \end{aligned} \tag{6.67}$$

(1) Adjoint of $L_{-m}|0\rangle$, using $\alpha_n^\dagger = \alpha_{-n}$.
(2) Move the annihilators to the right (Wick's theorem): $\alpha_{m-q'}$ contracts either with $\alpha_{-(m-q)}$ (giving $(m-q)\delta_{q'q}$) or with α_{-q} (giving $q\,\delta_{m-q',q}$), and then $\alpha_{q'}$ contracts with the remaining creator.
(3) Each Kronecker delta contributes the same sum.
(4) $\sum_{q=1}^{m-1} q = \frac{m(m-1)}{2}$ and $\sum_{q=1}^{m-1} q^2 = \frac{(m-1)m(2m-1)}{6}$.
(5) $\frac{m(m-1)}{12}\big(3m-(2m-1)\big) = \frac{m(m-1)(m+1)}{12}$.

Comparing with $\frac{c}{12}(m^3-m)$:

Key idea: one free boson has $c=1$

The operators (6.63) satisfy the Virasoro algebra (6.21) with $c=1$. For D independent bosons α_n^μ, $\mu = 1,\dots,D$, the L_n add and so do the central charges: $c=D$. The whole c-number comes from the norm of $L_{-m}|0\rangle$: a classical stress tensor would annihilate the

classical vacuum, but the quantum one creates pairs of quanta out of it.

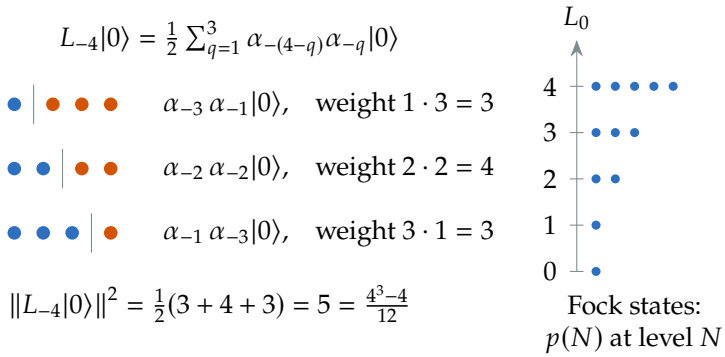


Figure 6.15: Left: $L_{-m}|0\rangle$ is a sum over the ways of splitting m into two positive integers, shown for $m = 4$ (blue dots: q; orange dots: $m-q$). Each splitting contributes $q(m-q)$ to the squared norm, and the total is $\frac{1}{12}(m^3 - m)$: this is the Virasoro central charge $c = 1$. Right: the Fock space of one boson has $p(N)$ states at $L_0 = N$ (p = number of partitions: $1, 1, 2, 3, 5, \ldots$); L_{-m} raises the level by m.

Watch out: the $-\frac{1}{24}$

Dropping the infinite constant (6.62) is not the end of the story. If one regularises it instead, with the zeta function $\sum_{m>0} m \to \zeta(-1) = -\frac{1}{12}$, the zero-point energy is $\frac{1}{2}\zeta(-1) = -\frac{1}{24} = -\frac{c}{24}$. This is the Casimir energy of a boson on a circle, and on a cylinder it is the statement that the Hamiltonian is $L_0 - \frac{c}{24}$. It is also the "bm" freedom of (6.59) in disguise: a shift of L_0 by a constant.

The free Majorana fermion: $c = \frac{1}{2}$

A real (Majorana) chiral fermion with antiperiodic boundary conditions (the Neveu–Schwarz sector) has modes b_r, $r \in \mathbb{Z} + \frac{1}{2}$, with

$$\{b_r, b_s\} = \delta_{r+s,0}, \qquad b_r^\dagger = b_{-r}, \qquad b_r|0\rangle = 0 \ (r > 0).$$

Its Virasoro generators are

$$L_m = \frac{1}{2}\sum_{r\in\mathbb{Z}+\frac{1}{2}} \left(r + \frac{m}{2}\right) :b_{-r}b_{m+r}: . \tag{6.68}$$

The weights $r + \frac{m}{2}$ are fixed by the requirement that the L_m act on the modes as vector fields act on a field of conformal weight $\frac{1}{2}$:

$$[b_{-r}b_{m+r}, b_s] \overset{(1)}{=} b_{-r}\{b_{m+r}, b_s\} - \{b_{-r}, b_s\}b_{m+r} = b_{-r}\,\delta_{m+r+s,0} - \delta_{s-r,0}\,b_{m+r},$$
$$[L_m, b_s] \overset{(2)}{=} \tfrac{1}{2}\left(-s - \tfrac{m}{2}\right)b_{m+s} - \tfrac{1}{2}\left(s + \tfrac{m}{2}\right)b_{m+s} = -\left(s + \frac{m}{2}\right)b_{m+s}.$$

(1) $[AB, C] = A\{B, C\} - \{A, C\}B$ for fermionic A, B, C.
(2) The first delta sets $r = -m - s$, with weight $r + \frac{m}{2} = -s - \frac{m}{2}$; the second sets $r = s$.

The same argument as for the boson shows that $[L_m, L_n] - (m-n)L_{m+n}$ is the c-number $\|L_{-m}|0\rangle\|^2\,\delta_{m+n,0}$. Now $L_{-m}|0\rangle = \frac{1}{2}\sum_r (r - \frac{m}{2})\,:b_{-r}b_{r-m}:\,|0\rangle$, and only $0 < r < m$ survives (both

operators must be creators). The terms r and $m-r$ are equal, because $(m-r-\frac{m}{2})\,b_{-(m-r)}b_{-r} = (\frac{m}{2}-r)(-b_{-r}b_{-(m-r)}) = (r-\frac{m}{2})\,b_{-r}b_{-(m-r)}$. So

$$L_{-m}|0\rangle \overset{(1)}{=} \sum_{0<r<m/2} \left(r-\frac{m}{2}\right) b_{-r}b_{-(m-r)}|0\rangle,$$

$$\|L_{-m}|0\rangle\|^2 \overset{(2)}{=} \sum_{0<r<m/2} \left(\frac{m}{2}-r\right)^2 \overset{(3)}{=} \frac{1}{2}\sum_{i=0}^{m-1}\left(\frac{m-1-2i}{2}\right)^2 \overset{(4)}{=} \frac{1}{8}\cdot\frac{m(m^2-1)}{3} = \frac{1}{2}\cdot\frac{m^3-m}{12}.$$

(1) Pair the terms r and $m-r$; the term $r=\frac{m}{2}$ (if present) has coefficient zero.
(2) The two-fermion states $b_{-r}b_{-s}|0\rangle$ with $r\neq s$ are orthonormal.
(3) Write $r = i+\frac{1}{2}$, $i = 0,\dots,m-1$, and sum over *all* $0<r<m$, which counts each pair twice.
(4) $\sum_{i=0}^{m-1}(m-1-2i)^2 = \frac{m(m^2-1)}{3}$ (the sum of the squares of m equally spaced numbers symmetric about 0; check $m=3$: $4+0+4=8$).

So *a Majorana fermion has* $c=\frac{1}{2}$. (Check at $m=2$: $L_{-2}|0\rangle = -\frac{1}{2}b_{-1/2}b_{-3/2}|0\rangle$, of norm $\frac{1}{4}=\frac{c}{2}$.) This is the central charge of the critical Ising model, which we will meet again as the first member of the unitary discrete series in Section 6.3.6. A complex (Dirac) fermion is two Majoranas and has $c=1$.

The Dirac sea and the Schwinger term: $k=1$

Finally, the Kac–Moody central term itself. Take a complex chiral fermion on a circle with single-particle momenta $p\in\mathbb{Z}+\frac{1}{2}$ and creation and annihilation operators $c_p^\dagger, c_p$, $\{c_p, c_q^\dagger\}=\delta_{pq}$. The ground state $|\Omega\rangle$ is the **Dirac sea**: every level with $p<0$ filled, every level with $p>0$ empty. The Fourier modes of the charge density $\psi^\dagger\psi$ are

$$\rho_n = \sum_p :c_p^\dagger c_{p+n}:\,, \qquad \rho_n^\dagger = \rho_{-n}. \tag{6.69}$$

ρ_n moves one particle from momentum $p+n$ to momentum p, lowering the total momentum by n. Two facts:

- $\rho_n|\Omega\rangle = 0$ for $n>0$: we would need an occupied level $p+n<0$ and an empty level $p>0$ with $p<p+n$, which is impossible.
- Naively $[\rho_m,\rho_n]=0$. Using $[c_a^\dagger c_b, c_c^\dagger c_d] = \delta_{bc}\,c_a^\dagger c_d - \delta_{ad}\,c_c^\dagger c_b$, the double sum gives $\sum_p c_p^\dagger c_{p+m+n} - \sum_q c_q^\dagger c_{q+m+n}$, two copies of the same divergent sum with opposite signs. Shifting the summation variable in an infinite sum of operators that do not annihilate the sea is exactly the illegal step.

The lemma replaces the illegal step by a vacuum expectation value. The commutator $[\rho_m,\rho_n]$ commutes with every c_p (the operator parts do cancel term by term when acting on fermion operators, since each c_p appears in finitely many terms), so it is a c-number, zero unless $m+n=0$. To evaluate it, look at what ρ_{-m} ($m>0$) does to the sea. $\rho_{-m}=\sum_p :c_p^\dagger c_{p-m}:$ moves a particle from $p-m$ up to p. Acting on the sea, it needs $p-m$ occupied ($p-m<0$) and p empty ($p>0$), i.e. $0<p<m$. For $p\in\mathbb{Z}+\frac{1}{2}$ there are exactly m such values, $p=\frac{1}{2},\frac{3}{2},\dots,m-\frac{1}{2}$, and each gives a different particle–hole state of norm 1. So

$$[\rho_m,\rho_n] \overset{(1)}{=} \langle\Omega|[\rho_m,\rho_{-m}]|\Omega\rangle\;\delta_{m+n,0} \overset{(2)}{=} \|\rho_{-m}|\Omega\rangle\|^2\;\delta_{m+n,0} \overset{(3)}{=} m\,\delta_{m+n,0}. \tag{6.70}$$

(1) The commutator is a c-number, nonzero only for $m+n=0$ (the lemma).

(2) $\rho_m|\Omega\rangle = 0$ for $m > 0$ and $\rho_m = \rho_{-m}^\dagger$.
(3) The m orthonormal particle–hole states just counted.

This is the U(1) Kac–Moody algebra at level $k = 1$: the *Schwinger term is the number of ways of lifting a particle out of the Dirac sea by m units of momentum* (Figure 6.16). In position space it is the δ' term of (6.34).

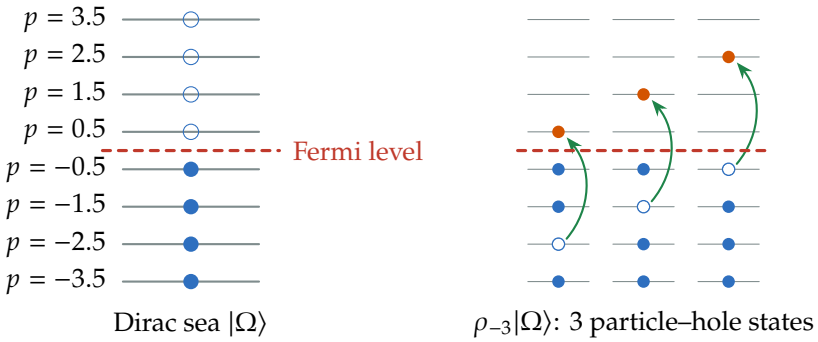


Figure 6.16: The Schwinger term counted. Left: the Dirac sea, all negative momentum levels filled. Right: ρ_{-m} lifts one particle by m units of momentum across the Fermi level; for $m = 3$ there are exactly three ways (green arrows, holes open circles, particles orange). So $\|\rho_{-m}|\Omega\rangle\|^2 = m$ and $[\rho_m, \rho_{-m}] = m$: the U(1) Kac–Moody algebra at level 1. With no Dirac sea (a classical charge density) this number would be 0.

Physical picture: bosons from fermions

Since $[\rho_m, \rho_n] = m\delta_{m+n,0}$ is exactly the boson algebra (6.61), the particle–hole excitations of the Dirac sea *are* the quanta of a chiral boson. This is bosonisation in two dimensions. As a check, the counting matches: the number of states of the Dirac sea (at fixed charge) with momentum N is $p(N)$, the same as the boson Fock space in Figure 6.15. With N_f species of fermion transforming in the fundamental of $\mathrm{U}(N_f)$, the currents $\sum_p :c_p^\dagger T^a c_{p+n}:$ give the $\mathrm{SU}(N_f)$ Kac–Moody algebra at level 1 in the physics normalisation.

6.2.10 The Sugawara construction

Nash's next point is that every Kac–Moody algebra comes with a Virasoro algebra built out of it. The reason is geometric.

Why a Virasoro algebra must be there

$\mathrm{Diff}^+(S^1)$ acts on the loop group by reparametrisation: a diffeomorphism ϕ sends a loop $\alpha(\theta)$ to $\alpha(\phi^{-1}(\theta))$. This is an automorphism of LG (the pointwise product is respected), so the symmetry group of a theory with LG symmetry is the semidirect product $LG \rtimes \mathrm{Diff}^+(S^1)$. Infinitesimally, the vector field $L_n = ie^{in\theta}\partial_\theta$ acts on the loop algebra element $T_m^a = e^{im\theta}T^a$ by differentiation:

$$L_n \cdot T_m^a \overset{(1)}{=} ie^{in\theta}\frac{d}{d\theta}\big(e^{im\theta}\big)T^a \overset{(2)}{=} ie^{in\theta}\,(im)\,e^{im\theta}\,T^a \overset{(3)}{=} -m\,T_{m+n}^a. \tag{6.71}$$

(1) The vector field differentiates the function part of the loop.
(2) $\frac{d}{d\theta}e^{im\theta} = ime^{im\theta}$.
(3) $i \cdot im = -m$, and $e^{i(m+n)\theta}T^a = T_{m+n}^a$.

In a quantum theory we want operators L_n that implement this action on the Hilbert space, i.e. with $[L_n, T^a_m] = -m\,T^a_{m+n}$. Since the Kac–Moody algebra is only a projective representation of LG, we should expect the L_n to form only a projective representation of $\mathrm{Diff}^+(S^1)$: a Virasoro algebra. The surprise is that the L_n can be built from the T^a_n themselves.

The stress tensor as a square of currents

Sugawara (1968) proposed, in four dimensions, that the energy–momentum tensor of a theory of currents be

$$T_{\mu\nu} = j^a_\mu j^a_\nu - \frac{g_{\mu\nu}}{2}\, j^a_\alpha j^{a\alpha}, \tag{6.72}$$

Nash's (6.35). In two dimensions this tensor is automatically traceless, the hallmark of conformal invariance:

$$g^{\mu\nu}T_{\mu\nu} \overset{(1)}{=} j^a_\mu j^{a\mu} - \frac{g^{\mu\nu}g_{\mu\nu}}{2}\, j^a_\alpha j^{a\alpha} \overset{(2)}{=} j^2 - \frac{2}{2}\, j^2 = 0.$$

(1) Contract with the inverse metric.
(2) $g^{\mu\nu}g_{\mu\nu} = \delta^\mu_\mu = 2$ in two dimensions.

In light-cone coordinates $x^\pm = t \pm x$ the only nonzero components are $T_{++} \propto j^a_+ j^a_+$ and $T_{--} \propto j^a_- j^a_-$, each a square of one chiral current. The Fourier modes of T_{++} on a circle are then bilinear in the Kac–Moody generators: $L_n \propto \sum_m T^a_m T^a_{n-m}$. A Lagrangian field theory in which all this happens is the two-dimensional sigma model with values in G and a Wess–Zumino term (Witten), the **WZW model**; the Wess–Zumino term is built from the 3-form $\mathrm{tr}(g^{-1}\mathrm{d}g)^3$ whose integrality we met in Section 6.2.5, and its coefficient is the level.

Why normal ordering is unavoidable

Nash gives the physical reason for normal ordering: products of operators at the same point, $A(x)B(y)$ as $y \to x$, are singular. Let us see it for the current. With $j^a(\theta) = \frac{1}{2\pi}\sum_n T^a_n e^{in\theta}$ (6.32) in a representation with vacuum $|0\rangle$ ($T^a_n|0\rangle = 0$ for $n \geq 0$), the vacuum two-point function of the current, for $u = \theta - \varphi$ with a small negative imaginary part, is

$$\begin{aligned}\langle 0|j^a(\theta)j^b(\varphi)|0\rangle &\overset{(1)}{=} \frac{1}{4\pi^2}\sum_{m>0}\langle 0|T^a_m T^b_{-m}|0\rangle\, e^{im(\theta-\varphi)} \overset{(2)}{=} \frac{k\,\delta^{ab}}{4\pi^2}\sum_{m>0} m\, e^{imu} \\ &\overset{(3)}{=} \frac{k\,\delta^{ab}}{4\pi^2}\,\frac{e^{iu}}{(1-e^{iu})^2} \overset{(4)}{=} -\frac{k\,\delta^{ab}}{16\pi^2\sin^2(u/2)} \xrightarrow{u\to 0} -\frac{k\,\delta^{ab}}{4\pi^2 u^2}.\end{aligned} \tag{6.73}$$

(1) Only $m > 0$, $n = -m$ survives: $T^b_n|0\rangle = 0$ for $n \geq 0$, and $\langle 0|T^a_m = 0$ for $m \leq 0$.
(2) $\langle 0|T^a_m T^b_{-m}|0\rangle = \langle 0|[T^a_m, T^b_{-m}]|0\rangle = km\delta^{ab}$ (the $f^{abc}T^c_0$ term has zero vacuum expectation value).
(3) $\sum_{m>0} m x^m = x\frac{d}{dx}\frac{1}{1-x} = \frac{x}{(1-x)^2}$ for $|x| < 1$.
(4) $1 - e^{iu} = -2ie^{iu/2}\sin\frac{u}{2}$, so $(1-e^{iu})^2 = -4e^{iu}\sin^2\frac{u}{2}$.

So $j^a(\theta)j^a(\theta)$ is infinite: the operator product has a double pole as the points collide, with coefficient proportional to the level. The cure is *point splitting*: define $:j^a j^a:(\theta) = \lim_{\varphi\to\theta}\big(j^a(\theta)j^a(\varphi) - \langle 0|j^a(\theta)j^a(\varphi)|0\rangle\big)$. In modes this subtraction is exactly the normal ordering prescription (6.76) below.

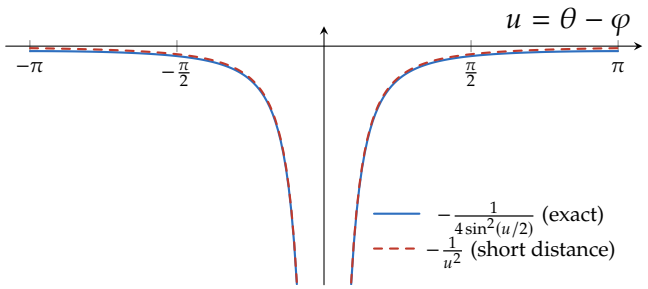


Figure 6.17: The vacuum correlator $\langle 0|j(\theta)j(\varphi)|0\rangle$ of a Kac–Moody current (in units of $k/4\pi^2$), from (6.73). It blows up like $1/u^2$ as the two points collide, so the square $j(\theta)^2$ of the current at one point is not defined. Subtracting this singular c-number is point splitting; in Fourier modes it is normal ordering.

The Sugawara operators

Let C (Nash's $c(G)$) be the value of the quadratic Casimir in the adjoint representation,

$$f^{acd}f^{bcd} = C\,\delta^{ab}, \qquad C = \frac{f^{abc}f^{abc}}{\dim G}. \tag{6.74}$$

(The adjoint generators are $(T^a_{\rm adj})_{bc} = -if^{abc}$, and $(T^a_{\rm adj}T^a_{\rm adj})_{bd} = -f^{abc}f^{acd} = f^{bca}f^{dca} = C\delta^{bd}$.) For SU(2) with $f^{abc} = \varepsilon^{abc}$, $\varepsilon^{acd}\varepsilon^{bcd} = 2\delta^{ab}$, so $C = 2$; for SU(N) in our normalisation $C = N$. Define

$$\boxed{L_n = \frac{1}{2k + C}\sum_{m\in\mathbb{Z}} :T^a_m T^a_{n-m}:} \tag{6.75}$$

(sum over a understood), Nash's (6.36), with the normal ordering of (6.37):

$$:T^a_m T^a_n: = \begin{cases} T^a_n T^a_m, & m \geq 0, \\ T^a_m T^a_n, & m < 0. \end{cases} \tag{6.76}$$

The annihilators ($m \geq 0$) are placed on the right. As for the boson, normal ordering matters only for $n = 0$: for $n \neq 0$, $T^a_m T^a_{n-m} - T^a_{n-m}T^a_m = if^{aac}T^c_n + k\,m\,\delta^{aa}\delta_{n,0} = 0$. On any state of a highest-weight representation (where T^a_m, $m > 0$, lowers L_0 and states have L_0 bounded below) only finitely many terms are nonzero.

Why $2k + C$ and not $2k$? Classically, a Poisson-bracket version of (6.75) with $\frac{1}{2k}$ would work. The extra C is a quantum correction, a shift of the level by (half) the dual Coxeter number, and it comes from the same illegal-shift phenomenon as the Dirac sea. We now compute it.

The key commutator

Let $S_n := \sum_p :T^a_p T^a_{n-p}:$, so that $L_n = S_n/(2k + C)$. We compute $[T^b_m, S_n]$ for $m > 0$ ($m < 0$ follows by taking adjoints, and $m = 0$ gives 0 because S_n is $\mathfrak{g}$-invariant). Because infinite sums must not be reshuffled, we cut the sum off at $|p| \leq M$ and let $M \to \infty$ only at the end, acting on a fixed state $|\psi\rangle$. Normal ordering changes each term by a c-number, which commutes with T^b_m, so we

may drop the colons inside the commutator. Put $q = m + n$.

$$\Big[T^b_m, \sum_{|p|\le M} T^a_p T^a_{n-p}\Big] \overset{(1)}{=} \sum_{|p|\le M} \Big((if^{bac}T^c_{m+p} + km\delta^{ab}\delta_{m+p,0})T^a_{n-p} + T^a_p\big(if^{bac}T^c_{q-p} + km\delta^{ab}\delta_{q-p,0}\big)\Big)$$
$$\overset{(2)}{=} 2km\,T^b_q + if^{bac}\Big(\sum_{p=m-M}^{m+M} T^c_p T^a_{q-p} - \sum_{p=-M}^{M} T^c_p T^a_{q-p}\Big)$$
$$\overset{(3)}{=} 2km\,T^b_q + if^{bac}\Big(\sum_{p=M+1}^{M+m} T^c_p T^a_{q-p} - \sum_{p=-M}^{-M+m-1} T^c_p T^a_{q-p}\Big). \tag{6.77}$$

(1) $[X, YZ] = [X, Y]Z + Y[X, Z]$ and the Kac–Moody relations (6.12).
(2) The two central terms each give kmT^b_q (from $p = -m$ and $p = q$, both inside the cutoff for large M). In the first f-term shift $p \to p - m$. In the second rename $a \leftrightarrow c$ and use $f^{bca} = -f^{bac}$.
(3) The two sums have the same summand and overlap on $m - M \le p \le M$; what is left are m terms at each end.

The naive answer would have been zero for the f-terms. The "end" terms do not vanish, because the operators are not normal ordered there (Figure 6.18). Act on $|\psi\rangle$ and let $M \to \infty$:

- *Lower end*, $p \approx -M$: then $q - p \approx M$ is large and positive, so $T^a_{q-p}|\psi\rangle = 0$ (it would lower L_0 below its minimum). These terms vanish.
- *Upper end*, $p \approx M$: now T^c_p is the annihilator but it stands on the left. Commute it through:

$$if^{bac}\,T^c_p T^a_{q-p}|\psi\rangle \overset{(1)}{=} if^{bac}\big(T^a_{q-p}T^c_p + [T^c_p, T^a_{q-p}]\big)|\psi\rangle \overset{(2)}{=} if^{bac}\,\big(if^{cad}T^d_q + kp\,\delta^{ca}\delta_{q,0}\big)|\psi\rangle$$
$$\overset{(3)}{=} -f^{bac}f^{cad}\,T^d_q|\psi\rangle \overset{(4)}{=} C\,T^b_q|\psi\rangle. \tag{6.78}$$

(1) Reorder.
(2) $T^c_p|\psi\rangle = 0$ for large p; the Kac–Moody relations.
(3) The δ^{ca} term dies against $f^{baa} = 0$.
(4) $f^{bac} = -f^{bca}$ and $f^{cad} = f^{dca}$, so $-f^{bac}f^{cad} = f^{bca}f^{dca} = C\delta^{bd}$ by (6.74).

There are m such terms, so altogether

$$[T^b_m, S_n] = (2k + C)\,m\,T^b_{m+n}, \qquad \text{i.e.} \qquad \boxed{[L_n, T^b_m] = -m\,T^b_{m+n}} \tag{6.79}$$

which is exactly the geometric action (6.71). The normalisation $\frac{1}{2k+C}$ in (6.75) is *forced* by demanding this.

The Virasoro algebra and its central charge

Now the argument we used for the free boson goes through word for word. Let $D_{mn} = [L_m, L_n] - (m - n)L_{m+n}$. By the Jacobi computation (6.65), with α_p replaced by T^b_p and (6.64) by (6.79), D_{mn} commutes with every T^b_p. Work in the vacuum representation: the states are polynomials in the T^a_{-n} acting on a vacuum with $T^a_n|0\rangle = 0$ for $n \ge 0$. By the lemma of Section 6.2.9, $D_{mn} = d_m\,\delta_{m+n,0}\,\mathbb{1}$. The numbers d_m form a cocycle of the Witt algebra (the Jacobi identity of the operators L_n forces (6.54)), so by (6.59) $d_m = am^3 + bm$. Two vacuum expectation values fix a and b.

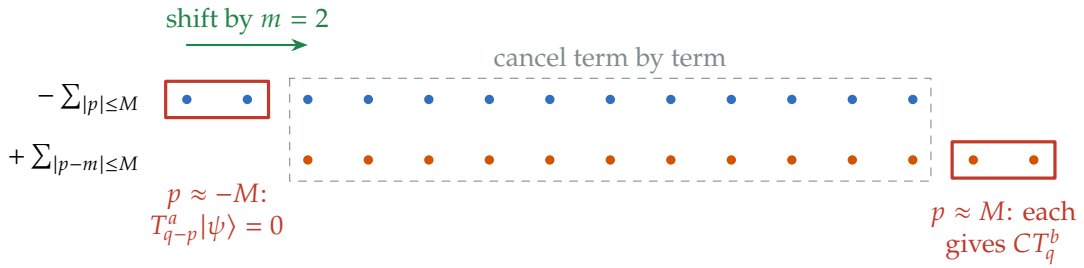


Figure 6.18: Why the Sugawara commutator is not the naive one. The two f^{abc}-sums in (6.77) have the same summand, but one is shifted by m (here $m = 2$). They cancel on the overlap (dashed box), leaving m terms at each end (red boxes). On any state the lower end vanishes, but at the upper end the annihilator stands on the wrong side, and each term leaves behind $C\,T^b_{m+n}$. This is the origin of the C in $2k + C$, and it is the same mechanism as the Dirac-sea counting of Figure 6.16.

$m = 1$. In $L_{-1}|0\rangle = \frac{1}{2k+C}\sum_p :T^a_p T^a_{-1-p}: |0\rangle$ a term survives only if both indices are negative, $p < 0$ and $-1-p < 0$, which is impossible; the borderline terms $p = 0, -1$ contain T^a_0 on the right, which kills the vacuum. So $L_{-1}|0\rangle = 0$, and $d_1 = \langle 0|[L_1, L_{-1}]|0\rangle - 2\langle 0|L_0|0\rangle = 0$: $a + b = 0$.

$m = 2$. Similarly $L_{-2}|0\rangle = \frac{1}{2k+C}\,T^a_{-1}T^a_{-1}|0\rangle$ (only $p = -1$ survives). Write $N_k := 2k + C$. First

$$\begin{aligned} T^b_1\,T^a_{-1}T^a_{-1}|0\rangle &\overset{(1)}{=} [T^b_1, T^a_{-1}]\,T^a_{-1}|0\rangle + T^a_{-1}[T^b_1, T^a_{-1}]|0\rangle \\ &\overset{(2)}{=} \big(if^{bac}T^c_0 + k\delta^{ab}\big)T^a_{-1}|0\rangle + T^a_{-1}\big(if^{bac}T^c_0 + k\delta^{ab}\big)|0\rangle \\ &\overset{(3)}{=} if^{bac}\,if^{cad}\,T^d_{-1}|0\rangle + 2k\,T^b_{-1}|0\rangle \overset{(4)}{=} (C + 2k)\,T^b_{-1}|0\rangle. \end{aligned} \tag{6.80}$$

(1) $T^b_1|0\rangle = 0$, so only the commutators contribute.
(2) Kac–Moody relations with $m = 1$.
(3) $T^c_0|0\rangle = 0$, and $T^c_0T^a_{-1}|0\rangle = [T^c_0, T^a_{-1}]|0\rangle = if^{cad}T^d_{-1}|0\rangle$.
(4) $-f^{bac}f^{cad} = C\delta^{bd}$, as in (6.78).

Hence

$$d_2 \overset{(1)}{=} \||L_{-2}|0\rangle\|^2 \overset{(2)}{=} \frac{1}{N_k^2}\,\langle 0|T^b_1T^b_1\,T^a_{-1}T^a_{-1}|0\rangle \overset{(3)}{=} \frac{1}{N_k}\,\langle 0|T^b_1T^b_{-1}|0\rangle \overset{(4)}{=} \frac{k\dim G}{N_k}.$$

(1) As in (6.66).
(2) $(T^a_{-1}T^a_{-1}|0\rangle)^\dagger = \langle 0|T^a_1T^a_1$.
(3) (6.80).
(4) $\langle 0|T^b_1T^b_{-1}|0\rangle = \langle 0|[T^b_1, T^b_{-1}]|0\rangle = k\delta^{bb} = k\dim G$.

With $d_2 = 8a + 2b = 6a$ we find $a = \frac{k\dim G}{6N_k}$, and writing $a = c/12$:

Key idea: the Sugawara central charge

The operators (6.75) satisfy

$$[L_m, L_n] = (m-n)L_{m+n} + \frac{2k\dim G}{12(2k+C)}\,m(m^2-1)\,\delta_{m,-n}, \qquad c = \frac{2k\dim G}{2k+C}, \tag{6.81}$$

Nash's (6.38). Every Kac–Moody algebra carries a Virasoro algebra, whose central charge is determined by the level.

Worked example 6.31: central charges of some Sugawara theories

In terms of the integer level $N = 2k$ of Section 6.3.9 and the dual Coxeter number $h^\vee$ (for our normalisation $C = h^\vee$ for SU(n)), the formula reads $c = \frac{N \dim G}{N + h^\vee}$.

G	$\dim G$	$h^\vee$	c at level N	$N = 1, 2, 3$
U(1)	1	0	1	1, 1, 1
SU(2)	3	2	$\frac{3N}{N+2}$	1, $\frac{3}{2}$, $\frac{9}{5}$
SU(3)	8	3	$\frac{8N}{N+3}$	2, $\frac{16}{5}$, 4
E_8	248	30	$\frac{248N}{N+30}$	8, $\frac{31}{2}$, $\frac{744}{33}$

Two checks. For U(1) ($f = 0$, $C = 0$), (6.75) is $L_n = \frac{1}{2k} \sum :T_m T_{n-m}:$, which with $\alpha_n = T_n/\sqrt{k}$ is the free boson (6.63), and $c = \frac{2k}{2k} = 1$ as it must be. As $N \to \infty$, $c \to \dim G$: at large level the currents behave like $\dim G$ free bosons (the classical limit). At level 1, c equals the rank (1, 2, 8): these theories are free bosons on the maximal torus in disguise (Frenkel–Kac).

6.2.11 The Casimir operator and the true L_0

Nash gives a second, more mathematical, way to see why normal ordering and the combination $2k + C$ appear (following Pressley and Segal): try to write down the Casimir operator of the loop algebra.

The finite-dimensional Casimir

For a finite-dimensional $\mathfrak{g}$ with orthonormal basis T^a the **Casimir operator** is

$$\Delta = T^a (T^a)^* = T^a T^a, \tag{6.82}$$

Nash's (6.39) (T^* is the adjoint, and our T^a are Hermitian). It commutes with every generator:

$$[T^b, \Delta] \overset{(1)}{=} [T^b, T^a] T^a + T^a [T^b, T^a] \overset{(2)}{=} i f^{bac} \left(T^c T^a + T^a T^c \right) \overset{(3)}{=} 0.$$

(1) Leibniz rule for commutators.
(2) $[T^b, T^a] = i f^{bac} T^c$.
(3) f^{bac} is antisymmetric in a, c, while $T^c T^a + T^a T^c$ is symmetric.

So Δ is a number on each irreducible representation: $j(j+1)$ for spin j of SU(2), C on the adjoint.

The loop version diverges

For the loop algebra the obvious analogue, with $(T^a_n)^* = T^a_{-n}$ in a unitary representation, is

$$\begin{aligned} \tilde{\Delta} &\overset{(1)}{=} \sum_{n \in \mathbb{Z}} T^a_n (T^a_n)^* \overset{(2)}{=} \sum_{n>0} \left(T^a_n T^a_{-n} + T^a_{-n} T^a_n \right) + T^a_0 T^a_0 \\ &\overset{(3)}{=} \sum_{n>0} 2\, T^a_{-n} T^a_n + T^a_0 T^a_0 + k \dim G \sum_{n>0} n. \end{aligned} \tag{6.83}$$

(1) Definition (Nash's (6.40)).
(2) Split into $n > 0$, $n < 0$ and $n = 0$; this is Nash's (6.41).
(3) $T^a_n T^a_{-n} = T^a_{-n} T^a_n + [T^a_n, T^a_{-n}] = T^a_{-n} T^a_n + i f^{aac} T^c_0 + kn\, \delta^{aa} = T^a_{-n} T^a_n + kn \dim G$ (Nash's (6.42)).

The last term is the divergent sum $\sum n$, the same zero-point energy as in (6.62). Dropping it defines

$$\tilde{\Delta}_0 := \sum_{n>0} 2\,T^a_{-n}T^a_n + T^a_0 T^a_0 = \sum_p :T^a_p T^a_{-p}: = S_0, \tag{6.84}$$

Nash's (6.43): it is exactly the normal-ordered operator, and $\tilde{\Delta}_0 = (2k + C)\,L_0$.

Fixing up the Casimir

$\tilde{\Delta}_0$ is not a Casimir operator: by (6.79) with $n = 0$,

$$[T^a_n, \tilde{\Delta}_0] = n\,(2k + C)\,T^a_n, \tag{6.85}$$

Nash's (6.44). (Nash prints a minus sign here; it is inconsistent with his own (6.47) at $n = 0$, which agrees with us.) The failure is proportional to n, which is what a rotation of the circle does. Indeed, for the geometric rotation generator $i\frac{d}{d\theta}$ acting on loops,

$$\left[T^a_n,\, i\frac{d}{d\theta}\right] f \overset{(1)}{=} e^{in\theta}T^a\, if' - i\frac{d}{d\theta}\left(e^{in\theta}T^a f\right) \overset{(2)}{=} -i(in)\,e^{in\theta}T^a f = n\,T^a_n f, \tag{6.86}$$

(1) Act on a test function $f(\theta)$.
(2) Product rule; the if' terms cancel.

which is Nash's (6.45). Subtracting,

$$\left[T^a_n,\ \tilde{\Delta}_0 - (2k + C)\,i\frac{d}{d\theta}\right] = n(2k + C)T^a_n - (2k + C)nT^a_n = 0, \tag{6.87}$$

so the true Casimir is $\tilde{\Delta}_0 - (2k + C)\,i\frac{d}{d\theta}$ (Nash's (6.46), where the sign of the second term follows his sign in (6.44)).

Key idea: the Casimir says "Sugawara L_0 = rotation generator"

To make room for $i\frac{d}{d\theta}$ we must enlarge the group from $\widetilde{LG}$ to the semidirect product $S^1 \ltimes \widetilde{LG}$, where $S^1 \subset \mathrm{Diff}^+(S^1)$ is the group of rigid rotations of the loop. Its Lie algebra is $\mathbb{R} \oplus \mathbb{R} \oplus \mathrm{Map}(S^1, \mathfrak{g})$ (rotation generator, central element, loops). In an irreducible representation the Casimir is a constant, so

$$(2k + C)\left(L_0 - i\frac{d}{d\theta}\right) = \text{const}:$$

up to a constant, the operator L_0 built from the currents *is* the generator of rotations of the circle. Pressley and Segal observe that a "regularised" Casimir must have the form $\tilde{\Delta}_0 + \alpha\, i\frac{d}{d\theta}$ by this argument; the calculation (6.85) fixes $\alpha = -(2k + C)$.

For $n \neq 0$ define $\tilde{\Delta}_n := \sum_m T^a_m T^a_{n-m} = S_n$ (no ordering is needed). Then (6.79) reads

$$[T^a_l, \tilde{\Delta}_n] = l\,(2k + C)\,T^a_{l+n}, \tag{6.88}$$

Nash's (6.47), and the operators

$$L_n = \frac{\tilde{\Delta}_n}{2k + C} \quad (n \neq 0), \qquad L_0 = \frac{1}{2k + C}\left(\sum_{n>0} 2\,T^a_{-n}T^a_n + T^a_0 T^a_0\right) \tag{6.89}$$

(Nash's (6.48)) are the Sugawara operators (6.75). They obey the Virasoro algebra (6.81) (Nash's (6.49)).

Worked example 6.32: the conformal weight of a highest-weight state

Let $|\lambda\rangle$ be a state annihilated by all T^a_n, $n > 0$, and transforming in an irreducible representation V_λ of $\mathfrak{g}$ under the T^a_0. Then only the $T^a_0 T^a_0$ term of L_0 acts:

$$L_0|\lambda\rangle \overset{(1)}{=} \frac{1}{2k+C}\, T^a_0 T^a_0 |\lambda\rangle \overset{(2)}{=} \frac{C_\lambda}{2k+C}\,|\lambda\rangle,$$

(1) The terms $T^a_{-n}T^a_n$ with $n > 0$ annihilate $|\lambda\rangle$.
(2) $T^a_0 T^a_0$ is the Casimir (6.82) of V_λ, with eigenvalue C_λ.

For SU(2) at level $N = 2k$ and spin j: $h_j = \frac{j(j+1)}{N+2}$. At level 1, the spin-$\frac{1}{2}$ state has $h = \frac{3/4}{3} = \frac{1}{4}$: the conformal weight of the operator $e^{i\phi/\sqrt{2}}$ of a free boson, as bosonisation requires.

Reading Nash (pp. 180–183, (6.30)–(6.49))

Nash's (6.30)–(6.34) are our (6.48)–(6.52); the Sp(2) example is ours, added to show a c-number arising at order $\hbar^2$ from ordering. Nash quotes the Virasoro algebra (6.14) and says its central term "gives rise to a conformal anomaly unless the space-time has a certain critical dimension"; we derived the cocycle in Section 6.2.8 and computed c for free fields in Section 6.2.9. His Sugawara formulae (6.35)–(6.38) are (6.72)–(6.81); he states that "these L_n satisfy the Virasoro algebra" without proof, and our (6.77)–(6.80) is the proof. His Casimir discussion (6.39)–(6.49) is Section 6.2.11. Two typographical points: in (6.42)–(6.43) the normal-ordered product should read $2T^a_{-n}T^a_n$ (annihilator on the right), and the sign in (6.44) and (6.46) should be flipped to agree with (6.47). His $c(G)$ is our C.

6.2.12 Extensions of $\mathrm{Map}(X, G)$ when $\dim X > 1$

So far the space on which the maps live has been a circle. Physics in $3+1$ dimensions needs more: a gauge transformation at fixed time is a map $g : X \to G$ from a *three-dimensional* space X. The equal-time commutators of the currents, or of the Gauss law generators, of such a theory may again pick up c-number terms, the higher-dimensional Schwinger terms. So we ask the question of this section once more, now for $\mathrm{Map}(X, G)$ with $\dim X > 1$:

What extensions does $\mathrm{Map}(X, \mathfrak{g})$ *have, and where do they come from?*

The answer comes in two parts. The *central* extensions are all inherited from the loop group by restricting maps to loops in X. But the extensions that actually occur in gauge theory are *not* central: the "c-number" in the commutator turns out to depend on the gauge field. We take the two parts in turn.

Central extensions: restrict to a loop

Nash's recipe is short: pick a loop in X and restrict. Let

$$f : S^1 \longrightarrow X \tag{6.90}$$

be a smooth loop (Nash's (6.50)). A map on X can be restricted to the loop by composing:

$$\hat{f} : \mathrm{Map}(X, \mathfrak{g}) \longrightarrow \mathrm{Map}(S^1, \mathfrak{g}), \qquad \alpha \longmapsto \alpha \circ f, \tag{6.91}$$

Nash's (6.51). In words: $\hat{f}(\alpha)(\theta) = \alpha(f(\theta))$ is the value of α at the point of the loop with parameter θ (Figure 6.19). The same formula defines a group homomorphism $\mathrm{Map}(X, G) \to LG$.

$\hat f$ respects brackets. The bracket in both algebras is pointwise, so at every θ

$$\hat f\big([\alpha,\beta]\big)(\theta) \overset{(1)}{=} [\alpha,\beta]\big(f(\theta)\big) \overset{(2)}{=} \big[\alpha(f(\theta)),\beta(f(\theta))\big] \overset{(3)}{=} \big[\hat f(\alpha),\hat f(\beta)\big](\theta).$$

(1) Definition (6.91) of $\hat f$.
(2) The pointwise bracket (6.3) on $\mathrm{Map}(X,\mathfrak g)$.
(3) Definition of $\hat f$ again, and the pointwise bracket on $\mathrm{Map}(S^1,\mathfrak g)$.

So $\hat f$ is a Lie algebra homomorphism.

Pulling back the cocycle. Any 2-cocycle c on $\mathrm{Map}(S^1,\mathfrak g)$ now gives one on $\mathrm{Map}(X,\mathfrak g)$:

$$(\hat f^{*}c)(\alpha,\beta) \coloneqq c\big(\hat f(\alpha),\hat f(\beta)\big). \tag{6.92}$$

It is antisymmetric because c is. It satisfies the cocycle condition (6.41) because

$$\begin{aligned}(\hat f^{*}c)\big([\alpha,\beta],\gamma\big) + \text{cyclic} &\overset{(1)}{=} c\big(\hat f([\alpha,\beta]),\hat f(\gamma)\big) + \text{cyclic}\\ &\overset{(2)}{=} c\big([\hat f\alpha,\hat f\beta],\hat f\gamma\big) + \text{cyclic} \overset{(3)}{=} 0.\end{aligned}$$

(1) Definition (6.92).
(2) $\hat f$ is a homomorphism (the chain above).
(3) c is a cocycle on $\mathrm{Map}(S^1,\mathfrak g)$, applied to the three loops $\hat f\alpha,\hat f\beta,\hat f\gamma$.

This is a general principle: *a homomorphism pulls cocycles back to cocycles*. With Nash we write $\hat f^{*}c \in H^2(\mathrm{Map}(X,\mathfrak g);\mathbb Z)$ when c is integral; since $\hat f$ is also a homomorphism of groups, the circle bundle $\widetilde{LG} \to LG$ pulls back along $\hat f : \mathrm{Map}(X,G) \to LG$ to a circle bundle over $\mathrm{Map}(X,G)$, and the integrality survives. So the central extension of the group exists too.

What the cocycle looks like. Take $c = \omega_k$, the Kac–Moody cocycle (6.43). The chain rule turns the θ-derivative of the restricted map into the differential of α along the loop:

$$\begin{aligned}(\hat f^{*}\omega_k)(\alpha,\beta) &\overset{(1)}{=} \frac{k}{\pi i}\int_0^{2\pi} \mathrm{tr}\Big(\frac{d}{d\theta}\big(\alpha(f(\theta))\big)\,\beta(f(\theta))\Big)\mathrm d\theta\\ &\overset{(2)}{=} \frac{k}{\pi i}\int_0^{2\pi} \mathrm{tr}\Big(\partial_\mu\alpha\big(f(\theta)\big)\,\frac{df^\mu}{d\theta}\,\beta\big(f(\theta)\big)\Big)\mathrm d\theta \overset{(3)}{=} \frac{k}{\pi i}\oint_f \mathrm{tr}\big(\mathrm d\alpha\;\beta\big).\end{aligned} \tag{6.93}$$

(1) Definitions (6.92) and (6.43).
(2) Chain rule, in local coordinates x^μ on X.
(3) $\partial_\mu\alpha\,\frac{df^\mu}{d\theta}\mathrm d\theta$ is the pull-back $f^{*}(\mathrm d\alpha)$ of the $\mathfrak g$-valued 1-form $\mathrm d\alpha = \partial_\mu\alpha\,\mathrm dx^\mu$, so the integral is the line integral of the 1-form $\mathrm{tr}(\mathrm d\alpha\,\beta)$ round the loop.

So the pulled-back cocycle is a *line integral* over the loop f of the 1-form $\mathrm{tr}(\mathrm d\alpha\,\beta)$ on X. We can now see directly why it is a cocycle: antisymmetry is $\oint_f \mathrm{tr}(\mathrm d\alpha\,\beta + \alpha\,\mathrm d\beta) = \oint_f \mathrm d\,\mathrm{tr}(\alpha\beta) = 0$, and the cocycle condition reduces, as in (6.45), to $\oint_f \mathrm d\,\mathrm{tr}(\alpha[\beta,\gamma]) = 0$. Both are Stokes' theorem on the closed curve $f(S^1)$, which has no boundary.

Worked example 6.33: the two-torus: two levels and a toroidal algebra

Let $X = T^2$ with angles (θ,φ). A basis of the complexified $\mathrm{Map}(T^2,\mathfrak g)$ is

$$T^a_{(m,n)} \coloneqq e^{i(m\theta+n\varphi)}\,T^a, \qquad m,n\in\mathbb Z,$$

with the pointwise bracket $[T^a_{(m,n)},T^b_{(m',n')}] = i f^{abc}T^c_{(m+m',n+n')}$ (multiply the exponentials).

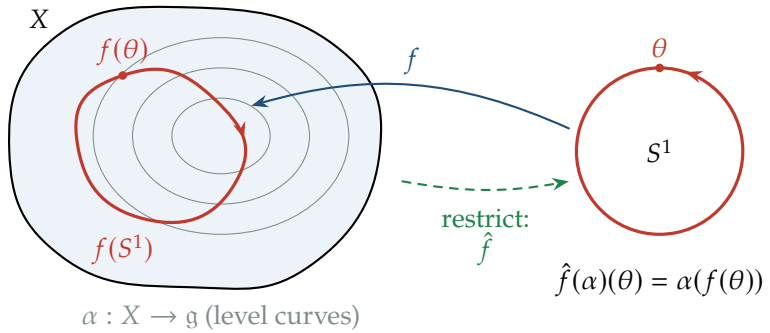


Figure 6.19: Restricting a map α on X (faint level curves) to a loop $f : S^1 \to X$ (red) produces a loop in $\mathfrak{g}$. Because the bracket is pointwise, restriction is a homomorphism $\mathrm{Map}(X,\mathfrak{g}) \to \mathrm{Map}(S^1,\mathfrak{g})$, and any cocycle of the loop algebra pulls back. The pulled-back Kac–Moody cocycle is the line integral $\frac{k}{\pi i}\oint_f \mathrm{tr}(\mathrm{d}\alpha\,\beta)$ along the red curve.

Take the loop $f_{\varphi_0}(\theta) = (\theta, \varphi_0)$, a circle of fixed φ. Then

$$\begin{aligned}\big(\hat{f}^*_{\varphi_0}\omega_k\big)\big(T^a_{(m,n)}, T^b_{(m',n')}\big) &\overset{(1)}{=} \omega_k\big(e^{in\varphi_0}\,T^a_m,\ e^{in'\varphi_0}\,T^b_{m'}\big)\\ &\overset{(2)}{=} e^{i(n+n')\varphi_0}\ k\,m\,\delta^{ab}\,\delta_{m+m',0}.\end{aligned}$$

(1) On the loop, $T^a_{(m,n)}$ restricts to the loop-algebra element $e^{in\varphi_0}e^{im\theta}T^a = e^{in\varphi_0}T^a_m$.
(2) Bilinearity and (6.44).

Each circle gives a cocycle, and so does any superposition of circles. The average over all the circles $\varphi_0 \in [0, 2\pi)$ is

$$c_1\big(T^a_{(m,n)}, T^b_{(m',n')}\big) \overset{(1)}{=} \frac{1}{2\pi}\int_0^{2\pi} e^{i(n+n')\varphi_0}\mathrm{d}\varphi_0\ k\,m\,\delta^{ab}\delta_{m+m',0} \overset{(2)}{=} k\,m\,\delta^{ab}\,\delta_{m+m',0}\,\delta_{n+n',0}.$$

(1) Average the previous result over φ_0.
(2) Orthogonality of Fourier modes.

Doing the same with the circles of fixed θ gives $c_2 = k'\,n\,\delta^{ab}\delta_{m+m',0}\delta_{n+n',0}$. The general combination $c = c_1 + c_2$ defines the extended algebra

$$[T^a_{(m,n)}, T^b_{(m',n')}] = i f^{abc}\,T^c_{(m+m',n+n')} + (k\,m + k'\,n)\,\delta^{ab}\,\delta_{m+m',0}\,\delta_{n+n',0},$$

a **toroidal Lie algebra** with *two* levels, one for each independent cycle of the torus. Restricted to the modes with $n = 0$ it is the ordinary Kac–Moody algebra at level k.

The example shows the general pattern. A single loop, or any superposition of loops, gives a central extension. The linear functionals on 1-forms of the type $\omega \mapsto \sum_i \lambda_i \oint_{f_i}\omega$, and their limits, are what de Rham called *closed 1-currents*. Nash's statement that *all* central extensions of $\mathrm{Map}(X,\mathfrak{g})$ (for simple $\mathfrak{g}$) are obtained in this fashion means: every continuous central 2-cocycle has the form $(\alpha,\beta) \mapsto T\big(\mathrm{tr}(\mathrm{d}\alpha\,\beta)\big)$ for a closed 1-current T on X (Pressley and Segal, §4.2). Because the line integral is sensitive to the 1-form $\mathrm{tr}(\mathrm{d}\alpha\,\beta)$ itself, and not just to its de Rham class, even a small contractible loop gives a nontrivial extension.

Watch out: central extensions are not where the higher-dimensional physics is

A loop in X is a one-dimensional object, and the extensions just found see only what happens along it. Local field theories in $3+1$ dimensions produce Schwinger terms that are local densities over all of space. These do not come from loops. They come from the second construction below, and they are *not central*.

Transgression: from forms on G to forms on $\mathrm{Map}(X,G)$

Nash's second construction produces a 2-form on the group $\mathrm{Map}(X,G)$ from a higher-degree form on G. It is called **transgression**, and it uses two operations we already know well: pull-back and integration.

Let $\dim X = n$, with X compact, oriented and without boundary, and let ω be a closed $(n+2)$-form on G. The **evaluation map**

$$\mathrm{Ev} : X \times \mathrm{Map}(X,G) \longrightarrow G, \qquad (x,\alpha) \longmapsto \alpha(x) \tag{6.94}$$

(Nash's (6.52)) just reads off the value of the map α at the point x. Pull ω back along it and integrate out the X directions:

$$c \coloneqq \int_X \mathrm{Ev}^*\omega, \qquad [c] \in H^2\big(\mathrm{Map}(X,G);\mathbb{R}\big), \tag{6.95}$$

Nash's (6.53). The degree count is: $\mathrm{Ev}^*\omega$ is an $(n+2)$-form on the product; integrating over the n-dimensional fibre X uses up n of its slots and leaves a 2-form on $\mathrm{Map}(X,G)$ (Figure 6.20).

Working definition 6.14 (integration over the fibre)

To evaluate $c = \int_X \mathrm{Ev}^*\omega$ at a point $\alpha \in \mathrm{Map}(X,G)$ on two tangent vectors $\delta_1\alpha, \delta_2\alpha$ (infinitesimal changes of the map), do this at each $x \in X$: feed ω at the point $\alpha(x) \in G$ the two vectors $\delta_1\alpha(x), \delta_2\alpha(x)$ and the n vectors $\partial_1\alpha(x), \dots, \partial_n\alpha(x)$ (the derivatives of α along coordinate directions of X). That gives a density on X [a number at each x times $\mathrm{d}^n x$, ready to be integrated over X]; integrate it:

$$c_\alpha(\delta_1\alpha, \delta_2\alpha) = \int_X \omega_{\alpha(x)}\big(\delta_1\alpha, \delta_2\alpha, \partial_1\alpha, \dots, \partial_n\alpha\big)\, \mathrm{d}^n x.$$

Why c is closed. Exterior derivatives commute with pull-backs, and integration over a fibre without boundary commutes with d (the extra term would be an integral over $\partial X = \varnothing$). So

$$\mathrm{d}c \overset{(1)}{=} \int_X \mathrm{d}\,\mathrm{Ev}^*\omega \overset{(2)}{=} \int_X \mathrm{Ev}^*\mathrm{d}\omega \overset{(3)}{=} 0.$$

(1) $\mathrm{d}\int_X = \int_X \mathrm{d}$ for a closed fibre X: Stokes' theorem on X, with $\partial X = \varnothing$.
(2) d commutes with pull-back (Chapter T1).
(3) ω is closed.

If $\omega = \mathrm{d}\eta$ is exact, the same steps give $c = \mathrm{d}\int_X \mathrm{Ev}^*\eta$, also exact. So transgression is a well-defined map $H^{n+2}(G;\mathbb{R}) \to H^2(\mathrm{Map}(X,G);\mathbb{R})$.

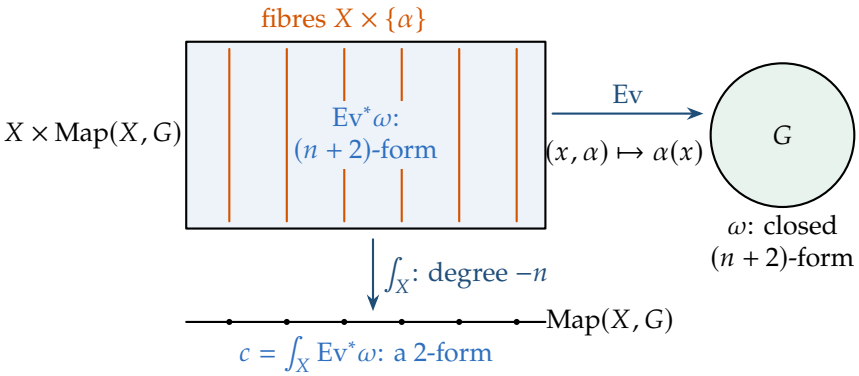


Figure 6.20: Transgression. Pull the $(n+2)$-form ω back from G to the product $X \times \mathrm{Map}(X, G)$ along the evaluation map, then integrate along the fibres X (orange). What is left is a 2-form c on $\mathrm{Map}(X, G)$, closed because X has no boundary.

Worked example 6.34: the case $n = 1$: transgression produces the level

Take $X = S^1$ and G compact, simple and simply connected, for instance SU(2). The group G carries the closed 3-form $\Omega = \mathrm{tr}(g^{-1}\mathrm{d}g)^3$ that generates $H^3(G;\mathbb{R})$ (for SU(2) we met it as the degree density, with $\int_{\mathrm{SU}(2)} \Omega = 24\pi^2$). Its transgression $c = \int_{S^1} \mathrm{Ev}^*\Omega$ is a closed 2-form on LG. Its value on a 2-sphere in LG can be computed without evaluating any form. A map $\Sigma : S^2 \to LG$ is a family of loops, i.e. a map $\hat{\Sigma} : S^1 \times S^2 \to G$, $\hat{\Sigma}(\theta, s) = \Sigma(s)(\theta)$, and

$$\int_{S^2} \Sigma^* c \overset{(1)}{=} \int_{S^1 \times S^2} \hat{\Sigma}^*\Omega \overset{(2)}{=} \deg(\hat{\Sigma}) \int_G \Omega \overset{(3)}{=} 24\pi^2 \deg(\hat{\Sigma}).$$

(1) The definition of c as a fibre integral: integrating first over the circle and then over the sphere is integrating over the product (Fubini), and $\mathrm{Ev} \circ (\mathrm{id} \times \Sigma) = \hat{\Sigma}$.
(2) For a map between closed oriented 3-manifolds, the pull-back of a top form integrates to the degree times the integral of the form (Example 6.5).
(3) The value of $\int_{\mathrm{SU}(2)} \Omega$.

So $c/24\pi^2$ integrates to an integer over every 2-sphere in LG, and since $\pi_2(LG) = \pi_3(G) = \mathbb{Z}$ (the pitfall in Section 6.2.5) there are spheres with degree 1. Therefore $\frac{c}{24\pi^2}$ represents a generator of $H^2(LG;\mathbb{Z}) = \mathbb{Z}$. On the other hand, $H^2(LG;\mathbb{R}) = H^2(\mathrm{Map}(S^1, \mathfrak{g});\mathbb{R})$ is one-dimensional and spanned by the left-invariant Kac–Moody form ω_k. Hence ω_k is cohomologous to a multiple of c, and the integrality condition (6.47) becomes an integrality condition on k. This is the geometric source of the quantised level: *the level of a loop group extension is the degree of a map into G*, just as the winding of the Wess–Zumino term is (Chapter 9).
Notice what did *not* happen: c itself is not left invariant. At the constant loop $\alpha \equiv 1$ the derivative $\partial_\theta \alpha$ vanishes, so by the working definition $c_1 = 0$ there, whereas $\omega_k \neq 0$. For $X = S^1$ this is harmless, because every class on LG has a left-invariant representative. For $\dim X > 1$ it is not harmless, and it is the reason for what follows.

Non-central extensions: cocycles with values in functions

For $\dim X = n > 1$ the 2-form c of (6.95) cannot in general be made left invariant (this is the failure of $H^*(\mathrm{Map}(X, G)) = H^*(\mathrm{Map}(X, \mathfrak{g}))$ noted in Section 6.2.5). Nash's move is to *keep* the non-invariance. Every $\alpha \in \mathrm{Map}(X, \mathfrak{g})$ defines a left-invariant vector field α^L on the group

$\mathrm{Map}(X,G)$, and the value of the 2-form c on two of them,

$$\hat{c}(\alpha,\beta) := c(\alpha^L,\beta^L), \tag{6.96}$$

is not a number but a *function* on $\mathrm{Map}(X,G)$ (Nash's (6.54)). So we get a map

$$\hat{c} : \mathrm{Map}(X,\mathfrak{g}) \times \mathrm{Map}(X,\mathfrak{g}) \longrightarrow \mathcal{H}, \qquad \mathcal{H} = \Omega^0\big(\mathrm{Map}(X,G)\big)$$

into the space $\mathcal{H}$ of functions on the group. (Nash writes $\Omega^0(\mathrm{Map}(X,\mathfrak{g}))$; since the function is on the group, we write G.) What kind of extension does such a thing define?

Working definition 6.15 (extension by a module; cocycle with coefficients)

Let $\mathfrak{g}$ act linearly on an abelian Lie algebra $\mathcal{H}$ [a vector space whose bracket is zero, here the functions on the group]: for each $U \in \mathfrak{g}$ an operator $\rho(U)$ on $\mathcal{H}$ with $[\rho(U),\rho(V)] = \rho([U,V])$. A bilinear antisymmetric $\hat{c} : \mathfrak{g}\times\mathfrak{g} \to \mathcal{H}$ defines a bracket on $\mathfrak{g}\oplus\mathcal{H}$:

$$\big[(U,F),(V,G)\big] = \big([U,V],\ \rho(U)G - \rho(V)F + \hat{c}(U,V)\big). \tag{6.97}$$

It satisfies the Jacobi identity if and only if

$$\sum_{\mathrm{cyc}(U,V,W)} \Big(\rho(U)\,\hat{c}(V,W) - \hat{c}\big([U,V],W\big)\Big) = 0. \tag{6.98}$$

Test for triviality: $\hat{c}$ can be removed by a redefinition $U \mapsto (U,\lambda(U))$ exactly when $\hat{c}(U,V) = \rho(U)\lambda(V) - \rho(V)\lambda(U) - \lambda([U,V])$ for some linear $\lambda : \mathfrak{g} \to \mathcal{H}$. The classes form $H^2(\mathfrak{g};\mathcal{H})$ [cocycles, with two counted as the same when they differ by such a removable term]. When $\rho = 0$ everything reduces to the central case (6.41), (6.42).

The ideal $\mathcal{H}$ is abelian but *not central*: the bracket of $(U,0)$ with $(0,G)$ is $(0,\rho(U)G)$, not zero. This is the price of letting the "c-number" be a function.

Why the cocycle condition is the Jacobi identity. Take three elements $(U,0),(V,0),(W,0)$. Using (6.97) twice,

$$\begin{aligned}\big[[(U,0),(V,0)],(W,0)\big] &\overset{(1)}{=} \big[([U,V],\hat{c}(U,V)),(W,0)\big] \\ &\overset{(2)}{=} \Big([[U,V],W],\ -\rho(W)\hat{c}(U,V) + \hat{c}([U,V],W)\Big).\end{aligned}$$

(1) The inner bracket.
(2) The outer bracket, with $F = \hat{c}(U,V)$ and $G = 0$.

The cyclic sum of the first entries vanishes by the Jacobi identity of $\mathfrak{g}$. The cyclic sum of the second entries is minus the left side of (6.98) (write $\rho(W)\hat{c}(U,V)$ as the cyclic partner of $\rho(U)\hat{c}(V,W)$). General elements (U,F) add only terms that cancel by $[\rho(U),\rho(V)] = \rho([U,V])$.

Why a closed 2-form gives such a cocycle. For a 2-form c and vector fields X,Y,Z on any manifold,

$$\mathrm{d}c(X,Y,Z) = \sum_{\mathrm{cyc}} \Big(X\big(c(Y,Z)\big) - c\big([X,Y],Z\big)\Big).$$

On left-invariant vector fields, $[\alpha^L,\beta^L] = [\alpha,\beta]^L$, and a vector field acts on functions by differentiation, $\rho(\alpha)F = \alpha^L F$. So $\mathrm{d}c = 0$ is literally the cocycle condition (6.98) for $\hat{c}$ with values in the module $\mathcal{H}$ of functions. Nash summarises: the Lie algebra $\mathrm{Map}(X,\mathfrak{g})$ has been extended by $\mathcal{H}$, and $\hat{c} \in H^2(\mathrm{Map}(X,\mathfrak{g});\mathcal{H})$.

The Mickelsson–Faddeev extension

Nash's concrete example is $G = \mathrm{SU}(N)$ with $N \geq 3$ and $X = S^3$. The physics setting is a non-abelian gauge theory with chiral fermions on the spatial 3-sphere. Its states are wave functionals $\Psi[A]$ of the spatial gauge field A, i.e. functions on the space $\mathcal{A}$ of connections on the (trivial) bundle over S^3. The gauge algebra acts on $\mathcal{A}$ by the infinitesimal gauge transformation, so the natural module is $\mathcal{E} = \mathrm{Map}(\mathcal{A}, \mathbb{R})$, not functions on the group.

Conventions for this subsection. We use real components: a basis $e_a = -iT^a$ of anti-Hermitian matrices, so that $[e_a, e_b] = f^{abc} e_c$; the gauge field is $A = A^a\, e_a$ with $A^a = A^a_\mu \, \mathrm{d}x^\mu$ real 1-forms (this is the dictionary $A_{\text{math}} = -iA^a_{\text{phys}} T^a$ with the coupling absorbed), and a gauge parameter is $\alpha = \alpha^a e_a$ with real functions α^a. The infinitesimal gauge transformation $A \mapsto A + \varepsilon\, D_A\alpha$ has components

$$(D_A\alpha)^c = \mathrm{d}\alpha^c + f^{abc} A^a \alpha^b, \tag{6.99}$$

from $\delta A = \mathrm{d}\alpha + [A, \alpha]$. It acts on functionals by the directional derivative $(\rho(\alpha)\Psi)[A] = \frac{d}{d\varepsilon}\Psi[A + \varepsilon D_A\alpha]\big|_{\varepsilon=0}$. This is a Lie algebra action: the vector fields $V_\alpha(A) = D_A\alpha$ on $\mathcal{A}$ obey $[V_\alpha, V_\beta] = V_{[\alpha,\beta]}$ (a two-line check using the Jacobi identity of $\mathfrak{g}$). The only group-theory input we need beyond f^{abc} is the totally symmetric tensor

$$d^{abc} := 2\,\mathrm{tr}\big(T^a\{T^b, T^c\}\big), \tag{6.100}$$

real because the T^a are Hermitian, and *invariant*: $f^{eax} d^{xbc} + f^{ebx} d^{axc} + f^{ecx} d^{abx} = 0$. (Proof: the trace of a commutator vanishes, so $\mathrm{tr}[T^e, T^a\{T^b, T^c\}] = 0$; expanding the commutator by the Leibniz rule and using $[T^e, T^a] = i f^{eax} T^x$ gives the three terms.)

Working definition 6.16 (Mickelsson–Faddeev cocycle)

For $\alpha, \beta \in \mathrm{Map}(S^3, \mathfrak{su}(N))$ and a connection A [the spatial gauge field $A = A^a e_a$ on S^3],

$$\hat{c}(\alpha,\beta)[A] := C \int_{S^3} d^{abc}\, A^c \wedge \mathrm{d}\alpha^a \wedge \mathrm{d}\beta^b = C \int_{S^3} d^{abc}\, \varepsilon^{ijk}\, A^c_k\, \partial_i\alpha^a\, \partial_j\beta^b\, \mathrm{d}^3x, \tag{6.101}$$

a function on $\mathcal{A}$ for each pair (α, β). Operationally: a Schwinger term [an extra term in the commutator of two gauge generators] that is linear in the gauge field and has one derivative on each gauge parameter, glued together with the d-symbol.

Relation to Nash's formula. Nash writes (6.55) as $c(\alpha,\beta) = C\int_{S^3} \mathrm{tr}(\mathrm{d}\alpha \wedge \mathrm{d}\beta \wedge A)$ with matrix-valued forms. A 2-cocycle must be antisymmetric, so this is to be read antisymmetrised in α, β. Then

$$\begin{aligned}
\mathrm{tr}(\mathrm{d}\alpha \wedge \mathrm{d}\beta \wedge A) - \mathrm{tr}(\mathrm{d}\beta \wedge \mathrm{d}\alpha \wedge A) &\overset{(1)}{=} \mathrm{d}\alpha^a \wedge \mathrm{d}\beta^b \wedge A^c \left(\mathrm{tr}(e_a e_b e_c) + \mathrm{tr}(e_b e_a e_c)\right) \\
&\overset{(2)}{=} (-i)^3\, \mathrm{tr}\big(\{T^a, T^b\}T^c\big)\, \mathrm{d}\alpha^a \wedge \mathrm{d}\beta^b \wedge A^c \\
&\overset{(3)}{=} \frac{i}{2}\, d^{abc}\, A^c \wedge \mathrm{d}\alpha^a \wedge \mathrm{d}\beta^b .
\end{aligned}$$

(1) Expand in components; in the second term swap the two real 1-forms, $\mathrm{d}\beta^b \wedge \mathrm{d}\alpha^a = -\mathrm{d}\alpha^a \wedge \mathrm{d}\beta^b$, and rename $a \leftrightarrow b$ so both terms carry the same forms.

(2) $e_a = -iT^a$.

(3) $(-i)^3 = i$; $\mathrm{tr}(\{T^a, T^b\}T^c) = \mathrm{tr}(T^c\{T^a, T^b\}) = \frac{1}{2} d^{cab} = \frac{1}{2} d^{abc}$; a 1-form moves past a 2-form with no sign.

So Nash's formula and (6.101) agree up to the constant, and only the totally symmetric part of the trace, the d-symbol, survives the antisymmetrisation.

Check 1: antisymmetry. Swap $\alpha \leftrightarrow \beta$ in (6.101), rename the summed indices $a \leftrightarrow b$, use $d^{bac} = d^{abc}$ and $\mathrm{d}\beta^a \wedge \mathrm{d}\alpha^b \to \mathrm{d}\beta^b \wedge \mathrm{d}\alpha^a = -\mathrm{d}\alpha^a \wedge \mathrm{d}\beta^b$: the integrand changes sign.

Check 2: the cocycle condition. We must show that

$$\delta\hat{c}(\alpha,\beta,\gamma) \coloneqq \rho(\alpha)\hat{c}(\beta,\gamma) - \rho(\beta)\hat{c}(\alpha,\gamma) + \rho(\gamma)\hat{c}(\alpha,\beta) - \hat{c}([\alpha,\beta],\gamma) + \hat{c}([\alpha,\gamma],\beta) - \hat{c}([\beta,\gamma],\alpha)$$

vanishes (this is (6.98), written out). Each term is an integral of either (i) $\mathrm{d}\alpha\,\mathrm{d}\beta\,\mathrm{d}\gamma$, with no A and no undifferentiated parameter, or (ii) A times two differentiated parameters times one undifferentiated one.

Terms of type (i) come from the $\mathrm{d}\gamma^c$ part of $D_A\gamma$ in $\rho(\gamma)\hat{c}(\alpha,\beta) = C\int d^{abc}(D_A\gamma)^c \wedge \mathrm{d}\alpha^a \wedge \mathrm{d}\beta^b$ and its two partners. Each is the integral of an exact form, e.g. $d^{abc}\mathrm{d}\gamma^c \wedge \mathrm{d}\alpha^a \wedge \mathrm{d}\beta^b = \mathrm{d}\big(d^{abc}\gamma^c\,\mathrm{d}\alpha^a \wedge \mathrm{d}\beta^b\big)$, over the closed manifold S^3: zero by Stokes.

Terms of type (ii) with α undifferentiated come from three places: the $f^{xyc}A^x\alpha^y$ part of $D_A\alpha$ in $\rho(\alpha)\hat{c}(\beta,\gamma)$; the $f^{uvx}\alpha^u\mathrm{d}\beta^v$ part of $\mathrm{d}[\alpha,\beta]^x$ in $-\hat{c}([\alpha,\beta],\gamma)$; and the $f^{uvx}\alpha^u\mathrm{d}\gamma^v$ part of $\mathrm{d}[\alpha,\gamma]^x$ in $+\hat{c}([\alpha,\gamma],\beta)$. Writing all three in the form $C\int(\cdots)A^d\,\alpha^e\,\mathrm{d}\beta^b \wedge \mathrm{d}\gamma^c$,

$$\begin{aligned}\text{coefficient} &\overset{(1)}{=} d^{bcx}f^{dex} \;-\; d^{xcd}f^{ebx} \;-\; d^{xbd}f^{ecx}\\ &\overset{(2)}{=} -\big(f^{edx}d^{bcx} + f^{ebx}d^{xcd} + f^{ecx}d^{bxd}\big) \overset{(3)}{=} 0.\end{aligned}$$

(1) The three sources in order. In the first, $(D_A\alpha)^x \supset f^{dex}A^d\alpha^e$ fills the A slot of $d^{bcx}A^x \wedge \mathrm{d}\beta^b \wedge \mathrm{d}\gamma^c$. In the second, $[\alpha,\beta]^x = f^{uvx}\alpha^u\beta^v$, so $\mathrm{d}[\alpha,\beta]^x \supset f^{ebx}\alpha^e\mathrm{d}\beta^b$ sits in $-d^{xcd}A^d \wedge \mathrm{d}[\alpha,\beta]^x \wedge \mathrm{d}\gamma^c$. In the third, $\mathrm{d}[\alpha,\gamma]^x \supset f^{ecx}\alpha^e\mathrm{d}\gamma^c$ sits in $d^{xbd}A^d \wedge \mathrm{d}[\alpha,\gamma]^x \wedge \mathrm{d}\beta^b$, and $\mathrm{d}\gamma^c \wedge \mathrm{d}\beta^b = -\mathrm{d}\beta^b \wedge \mathrm{d}\gamma^c$ gives the sign.
(2) $f^{dex} = -f^{edx}$, and the d's are totally symmetric.
(3) Invariance of d^{abc} under the adjoint action of T^e.

Since $\delta\hat{c}$ is totally antisymmetric in (α,β,γ), the terms with β or γ undifferentiated cancel in the same way, with the letters permuted. Hence $\delta\hat{c} = 0$: (6.101) is a cocycle, and the Jacobi identity of the extended algebra reduces to two things we trust, *Stokes' theorem on S^3 and the gauge invariance of d^{abc}*.

Key idea: the Mickelsson–Faddeev extension (Nash, pp. 184–185)

The gauge algebra $\mathrm{Map}(S^3,\mathfrak{su}(N))$, $N \geq 3$, has a nontrivial extension by the module $\mathcal{E} = \mathrm{Map}(\mathcal{A},\mathbb{R})$ of functions of the gauge field:

$$[(\alpha,F),(\beta,G)] = \big([\alpha,\beta],\ \rho(\alpha)G - \rho(\beta)F + \hat{c}(\alpha,\beta)\big),$$

with $\hat{c}$ the cocycle (6.101). Since $\mathcal{E}$ is the Lie algebra of the abelian group $\mathrm{Map}(\mathcal{A},S^1)$ (the Lie algebra of S^1 is $\mathbb{R}$), the corresponding group is an extension

$$1 \longrightarrow \mathrm{Map}(\mathcal{A},S^1) \longrightarrow \widehat{\mathrm{Map}(S^3,G)} \longrightarrow \mathrm{Map}(S^3,G) \longrightarrow 1,$$

the **Mickelsson–Faddeev extension**, in which $\mathrm{Map}(S^3,G)$ acts on $\mathrm{Map}(\mathcal{A},S^1)$ through its action on connections. It is the $3+1$ dimensional analogue of the Kac–Moody extension, and it is what chiral fermions produce as a Schwinger term.

Worked example 6.35: why $N \geq 3$: the d-symbol of SU(2) and SU(3)

For SU(2), $T^a = \frac{1}{2}\sigma^a$, and the Pauli matrices anticommute to the identity. So

$$d^{abc} \overset{(1)}{=} 2\,\mathrm{tr}\big(T^a\{T^b, T^c\}\big) \overset{(2)}{=} \tfrac{1}{4}\,\mathrm{tr}\big(\sigma^a\{\sigma^b, \sigma^c\}\big) \overset{(3)}{=} \tfrac{1}{4}\cdot 2\delta^{bc}\,\mathrm{tr}\,\sigma^a \overset{(4)}{=} 0.$$

(1) Definition (6.100).
(2) $T^a = \sigma^a/2$ contributes $(1/2)^3$, times the 2 in front.
(3) $\{\sigma^b, \sigma^c\} = 2\delta^{bc}\mathbb{1}$.
(4) The Pauli matrices are traceless.

So for SU(2) the cocycle (6.101) vanishes identically. For SU(3) it does not: with the Gell-Mann matrices, $T^1 = \frac{1}{2}\lambda^1$ and $T^8 = \frac{1}{2\sqrt{3}}\mathrm{diag}(1,1,-2)$,

$$d^{811} \overset{(1)}{=} 2\,\mathrm{tr}\big(T^8\cdot 2(T^1)^2\big) \overset{(2)}{=} 2\,\mathrm{tr}\Big(\tfrac{1}{2\sqrt{3}}\mathrm{diag}(1,1,-2)\cdot\tfrac{1}{2}\mathrm{diag}(1,1,0)\Big) \overset{(3)}{=} \frac{2}{4\sqrt{3}}\,(1+1) = \frac{1}{\sqrt{3}}.$$

(1) $\{T^1, T^1\} = 2(T^1)^2$.
(2) $\lambda^1 = \begin{pmatrix}0&1&0\\1&0&0\\0&0&0\end{pmatrix}$ squares to $\mathrm{diag}(1,1,0)$, so $(T^1)^2 = \frac{1}{4}\mathrm{diag}(1,1,0)$ and $2(T^1)^2 = \frac{1}{2}\mathrm{diag}(1,1,0)$.
(3) Multiply the diagonals and add.

This is the standard value $d_{118} = 1/\sqrt{3}$. The same dichotomy is topological: the cocycle is the transgression of the 5-form $\mathrm{tr}(g^{-1}\mathrm{d}g)^5$ on G over $X = S^3$ ($n+2 = 5$), and this form represents a nonzero class in $H^5(\mathrm{SU}(N);\mathbb{R})$ exactly when $N \geq 3$. For $\mathrm{SU}(2) = S^3$ there is nothing in degree 5. In homotopy language, $\pi_5(\mathrm{SU}(N)) = \mathbb{Z}$ for $N \geq 3$, while $\pi_5(\mathrm{SU}(2)) = \mathbb{Z}_2$ is torsion and invisible to real cohomology. The d-symbol is also exactly what appears in the triangle anomaly $\propto d^{abc}$ of chiral gauge theories (Chapter 10).

Worked example 6.36: the abelian case: a Schwinger term that can be removed

For $G = \mathrm{U}(1)$ drop all indices ($f = 0$, and the "d-symbol" is 1): $\hat{c}(\alpha,\beta)[A] = \int_{S^3} A\wedge\mathrm{d}\alpha\wedge\mathrm{d}\beta$, with $\rho(\alpha)$ the derivative along $\delta A = \mathrm{d}\alpha$. Try the functional $\lambda(\beta)[A] := \int_{S^3}\beta\,A\wedge\mathrm{d}A$, a gauge parameter times the Chern–Simons density. Its coboundary is

$$(\delta\lambda)(\alpha,\beta) \overset{(1)}{=} \rho(\alpha)\lambda(\beta) - \rho(\beta)\lambda(\alpha) \overset{(2)}{=} \int \beta\,\mathrm{d}\alpha\wedge\mathrm{d}A - \int \alpha\,\mathrm{d}\beta\wedge\mathrm{d}A$$
$$\overset{(3)}{=} -2\int \alpha\,\mathrm{d}\beta\wedge\mathrm{d}A \overset{(4)}{=} -2\int \mathrm{d}\alpha\wedge\mathrm{d}\beta\wedge A \overset{(5)}{=} -2\,\hat{c}(\alpha,\beta).$$

(1) The coboundary formula of the working definition; $[\alpha,\beta] = 0$.
(2) $\rho(\alpha)\int\beta A\,\mathrm{d}A = \int\beta\,(\mathrm{d}\alpha\wedge\mathrm{d}A + A\wedge\mathrm{d}\mathrm{d}\alpha)$ and $\mathrm{d}\mathrm{d}\alpha = 0$.
(3) Stokes: $0 = \int\mathrm{d}(\alpha\beta\,\mathrm{d}A) = \int\beta\,\mathrm{d}\alpha\wedge\mathrm{d}A + \int\alpha\,\mathrm{d}\beta\wedge\mathrm{d}A$.
(4) Stokes again: $\mathrm{d}(\alpha\,\mathrm{d}\beta\wedge A) = \mathrm{d}\alpha\wedge\mathrm{d}\beta\wedge A - \alpha\,\mathrm{d}\beta\wedge\mathrm{d}A$ integrates to zero.
(5) A 1-form moves past a 2-form without sign.

So $\hat{c} = \delta(-\frac{1}{2}\lambda)$ is a coboundary. Physically, the abelian Schwinger term is removed by redefining the Gauss law generator, $G_\alpha \to G_\alpha + \frac{1}{2}\int\alpha\,A\wedge\mathrm{d}A$. In the non-abelian case with $d^{abc} \neq 0$ no such redefinition exists (Mickelsson, Faddeev), and the extension is genuine.

Physical picture: why a non-central Schwinger term signals an anomaly

In a gauge theory in the Hamiltonian formalism, physical states must be gauge invariant: $G_\alpha|\psi\rangle = 0$ for the Gauss law generators G_α (one for each gauge parameter $\alpha(x)$). If the G_α represent the Lie algebra of $\mathrm{Map}(X, G)$ this is consistent. If instead

$$[G_\alpha, G_\beta] = G_{[\alpha,\beta]} + i\,\hat{c}(\alpha,\beta)[A],$$

then applying both sides to a physical state gives $\hat{c}(\alpha,\beta)[A]|\psi\rangle = 0$, which is impossible for a non-vanishing multiplication operator. The Gauss law cannot be imposed; the gauge symmetry is anomalous. This is the Hamiltonian face of the chiral anomaly, and the Mickelsson–Faddeev cocycle is related to the four-dimensional anomaly $\propto d^{abc}$ by the descent equations of Chapter 10. In $1+1$ dimensions the analogous obstruction is the central Kac–Moody term itself.

Reading Nash (pp. 178–180 and 183–185, (6.27)–(6.29), (6.50)–(6.55))

The opening of Nash's §2, the exact sequence (6.27), split and central extensions, the bracket (6.28) and the cocycle condition (6.29), is our Sections 6.2.1 and 6.2.3; the projective representations and the examples (Heisenberg, Galilean, $\mathrm{SU}(2) \to \mathrm{SO}(3)$) are added background. His comparison of $H^*(G)$ with $H^*(\mathfrak{g})$ and the integrality discussion (footnote 2) are Sections 6.2.5 and 6.2.6. On pp. 183–185, (6.50)–(6.51) are our (6.90)–(6.91) and (6.52)–(6.53) are (6.94)–(6.95). Nash writes the coefficient module of (6.54) as $\Omega^0(\mathrm{Map}(X,\mathfrak{g}))$; the functions live on the group $\mathrm{Map}(X, G)$ (or, in the example, on $\mathcal{A}$). His (6.55) must be read antisymmetrised in α, β, which leaves the d-symbol form (6.101). The transgression example for $X = S^1$, the module-cocycle formalism, the verification of the cocycle condition, and the SU(2) and U(1) examples are added.

Recap: Section 6.2

- An extension $1 \to H \to \bar{G} \to G \to 1$ is central if H commutes with everything. Quantum symmetries are projective, $U(g)U(h) = e^{i\phi(g,h)}U(gh)$, so they are true representations of a central extension.
- Lie algebra version: $[U, V] \mapsto ([U, V], c(U, V))$ with c an antisymmetric 2-cocycle, $\sum_{\mathrm{cyc}} c([U, V], W) = 0$; removable iff $c = \lambda \circ [\,,\,]$. Classes: $H^2(\mathfrak{g};\mathfrak{h})$. Semisimple finite-dimensional $\mathfrak{g}$: $H^2 = 0$ (Whitehead).
- Loop algebra: $\omega_k(\xi,\eta) = \frac{k}{\pi i}\oint \mathrm{tr}(\xi'\eta)$ gives $km\delta^{ab}\delta_{m+n,0}$; the cocycle identity is Stokes on S^1. Group extension exists iff $[\omega/2\pi]$ is integral: the level is quantised, $N = 2k \in \mathbb{Z}$.
- Witt algebra: the Jacobi identity allows only $am^3 + bm$; bm is a shift of L_0; $c/12\,(m^3 - m)$ remains. Free boson $c = 1$, Majorana fermion $c = \frac{1}{2}$, Dirac sea gives $k = 1$ (Schwinger term).
- Sugawara: $L_n = \frac{1}{2k+C}\sum :T^a_m T^a_{n-m}:$ obeys Virasoro with $c = \frac{2k\dim G}{2k+C}$; L_0 is the regularised Casimir, equal to rotation of the circle up to a constant.
- $\mathrm{Map}(X, G)$, $\dim X > 1$: central extensions come from loops in X (line integrals of $\mathrm{tr}(\mathrm{d}\alpha\,\beta)$). Transgression $c = \int_X \mathrm{Ev}^*\omega$ gives non-central extensions with values in functions; for $X = S^3$, $G = \mathrm{SU}(N \geq 3)$ this is the Mickelsson–Faddeev cocycle $\int d^{abc}A^c\,\mathrm{d}\alpha^a\,\mathrm{d}\beta^b$, the higher-dimensional Schwinger term.

6.3 Representations

We now have the algebras. The Kac–Moody algebra (6.12) and the Virasoro algebra (6.21) are the Lie algebras of central extensions of LG and $\mathrm{Diff}(S^1)$. The question of this section is the one a physicist asks of any symmetry:

On what Hilbert spaces can these algebras act, and what does the space of states look like?

For SU(2) the answer is the list of spins $j = 0, \frac{1}{2}, 1, \ldots$ and the multiplets $|j, m\rangle$. We want the analogue for the infinite-dimensional algebras.

Nash begins with a warning. The representation theory of infinite-dimensional groups is "still under construction"; it is nothing like as complete as that of compact Lie groups. Two cases are well understood: the loop groups LG and $\mathrm{Diff}(S^1)$. Much less is known for $\mathrm{Map}(X, G)$ with $\dim X > 1$ (Pressley and Segal summarise what is known). Also, as in physics practice, we mostly represent the *Lie algebra* and not the group: we look for operators L_n and T^a_n on a vector space obeying the commutation relations, and worry about exponentiating them only at the end (Section 6.3.10).

The section runs as follows. The single physical requirement that the energy be bounded below leads to *highest-weight representations* (Section 6.3.1). We construct the biggest one, the *Verma module* (Section 6.3.2), find the *null states* inside it and remove them (Sections 6.3.3 and 6.3.4), and see how the inner product decides which states are null (Section 6.3.5). Unitarity then fixes the allowed values of the central charge (Section 6.3.6) and forces the representation to be projective (Section 6.3.7). The same story for the Kac–Moody algebra gives the quantised level (Sections 6.3.8 and 6.3.9). Finally the Borel–Weil theorem shows where the representations live geometrically (Section 6.3.10).

6.3.1 Positive energy and highest weights

Why L_0 is an energy

Recall that on the circle $L_n = ie^{in\theta}\frac{d}{d\theta}$ (6.17), so $L_0 = i\frac{d}{d\theta}$ generates rotations of the circle. In a quantum field theory on a spatial circle, the generator of rotations is the total momentum P (up to sign). Now think of a *chiral* field, a left mover $\phi(t + \theta)$. For such a field, a shift in time and a shift in angle are the same thing, so its momentum and its energy coincide: L_0 *is* the Hamiltonian of the left movers, up to an additive constant. For a closed string or a two-dimensional conformal field theory with both chiralities, $H = L_0 + \bar{L}_0 - \frac{c+\bar{c}}{24}$ and $P = L_0 - \bar{L}_0$ (radial quantisation, Figure 6.8 and Chapter 11). Either way:

Key idea: the positive-energy condition

In every physical application, the spectrum of L_0 is the spectrum of an energy, and so must be *bounded below*. This one condition selects the representations that matter.

Nash makes a point here that we will prove at the end of this section (Section 6.3.7): once L_0 is bounded below, every nontrivial irreducible unitary representation of $\mathrm{vect}(S^1)$ is a representation of the *central extension*, i.e. it has $c \neq 0$. In group language, the representations of $\mathrm{Diff}(S^1)$ we want are *projective*: recall from (6.36) that unitary operators T_g form a projective representation if

$$T_g T_h = c(g, h)\, T_{gh}, \qquad |c(g, h)| = 1, \tag{6.102}$$

Nash's (6.56), and that a projective representation of G is a true representation of the extension $\bar{G}$ (Section 6.2.2). So this is the physics reason to attach more importance to the Virasoro algebra

than to vect(S^1) itself.

Unitarity: $L_n^\dagger = L_{-n}$

A unitary representation of a group represents a *real* Lie algebra element ξ by an anti-Hermitian operator $\rho(\xi)$, because $e^{t\rho(\xi)}$ must be unitary. The real vector fields on the circle are $f(\theta)\frac{d}{d\theta}$ with real f. The L_n are complex combinations of them, and the complex-linear extension of ρ fixes their adjoints:

$$\begin{aligned}\rho(L_n)^\dagger &\overset{(1)}{=} \Big(i\,\rho\big(\cos n\theta\,\tfrac{d}{d\theta}\big) - \rho\big(\sin n\theta\,\tfrac{d}{d\theta}\big)\Big)^\dagger \\ &\overset{(2)}{=} -i\,\rho\big(\cos n\theta\,\tfrac{d}{d\theta}\big)^\dagger - \rho\big(\sin n\theta\,\tfrac{d}{d\theta}\big)^\dagger \\ &\overset{(3)}{=} i\,\rho\big(\cos n\theta\,\tfrac{d}{d\theta}\big) + \rho\big(\sin n\theta\,\tfrac{d}{d\theta}\big) \overset{(4)}{=} \rho(L_{-n}).\end{aligned}$$

(1) $L_n = ie^{in\theta}\frac{d}{d\theta} = i\cos n\theta\,\frac{d}{d\theta} - \sin n\theta\,\frac{d}{d\theta}$, and ρ is complex linear.
(2) The adjoint is antilinear: $(iA)^\dagger = -iA^\dagger$.
(3) The real vector fields are represented by anti-Hermitian operators.
(4) $L_{-n} = ie^{-in\theta}\frac{d}{d\theta} = i\cos n\theta\,\frac{d}{d\theta} + \sin n\theta\,\frac{d}{d\theta}$.

So in a unitary representation

$$L_n^\dagger = L_{-n}, \qquad L_0^\dagger = L_0, \qquad c \in \mathbb{R}. \tag{6.103}$$

(The last because c is a real multiple of the identity; take the adjoint of (6.21).) The free boson (6.63) satisfies this, since $\alpha_n^\dagger = \alpha_{-n}$. This is the Virasoro analogue of $J_\pm^\dagger = J_\mp$ for angular momentum.

Raising and lowering

The commutator with L_0 is the whole story of the spectrum. Setting $m = 0$ in (6.21) gives $[L_0, L_n] = -nL_n$ (the central term vanishes for $m = 0$). Let $|\psi\rangle$ be an eigenvector, $L_0|\psi\rangle = E|\psi\rangle$. Then

$$L_0\big(L_n|\psi\rangle\big) \overset{(1)}{=} [L_0, L_n]|\psi\rangle + L_nL_0|\psi\rangle \overset{(2)}{=} -nL_n|\psi\rangle + E\,L_n|\psi\rangle = (E-n)\,L_n|\psi\rangle. \tag{6.104}$$

(1) Add and subtract $L_nL_0|\psi\rangle$.
(2) $[L_0, L_n] = -nL_n$ and $L_0|\psi\rangle = E|\psi\rangle$.

So L_n with $n > 0$ *lowers* the energy by n, and L_{-n} *raises* it by n, exactly as J_- and J_+ lower and raise J_3 (Figure 6.21).

Working definition 6.17 (highest-weight representation)

A representation is a **highest-weight representation** with **highest weight** (c, h) if the central element [the extra generator of (6.21) that commutes with every L_n] acts as the number c and there is a state $|h\rangle$, the **vacuum** or **highest-weight state**, with

$$L_0|h\rangle = h|h\rangle, \qquad L_n|h\rangle = 0 \quad (n \geq 1), \tag{6.105}$$

Nash's (6.57), such that the whole space is obtained by acting on $|h\rangle$ with the L_{-n} ($|h\rangle$ is **cyclic**). Operationally: find the state of lowest energy; check that every L_n with $n > 0$ kills it; then h is its energy and c is the central charge.

Why the lowest-energy state is killed by the lowering operators. This is Nash's argument. Suppose h is the lowest eigenvalue of L_0 and $L_n|h\rangle \neq 0$ for some $n \geq 1$. By (6.104) with $E = h$, $L_n|h\rangle$ is an eigenvector of L_0 with eigenvalue $h - n < h$. That contradicts the assumption that h is the lowest eigenvalue. So $L_n|h\rangle = 0$ for all $n \geq 1$. The positive-energy condition *forces* (6.105).

The raising operators do not kill it. Nash states that $L_{-n}|h\rangle \neq 0$ "as may be seen by applying the commutation relations". Here is the calculation. In a unitary representation, with $\langle h|h\rangle = 1$,

$$\|L_{-n}|h\rangle\|^2 \overset{(1)}{=} \langle h|L_n L_{-n}|h\rangle \overset{(2)}{=} \langle h|[L_n, L_{-n}]|h\rangle \overset{(3)}{=} \langle h|\Big(2nL_0 + \frac{c}{12}n(n^2-1)\Big)|h\rangle$$
$$\overset{(4)}{=} 2nh + \frac{c}{12}\,n(n^2-1). \tag{6.106}$$

(1) $(L_{-n}|h\rangle)^\dagger = \langle h|L_{-n}^\dagger = \langle h|L_n$ by (6.103).
(2) $L_n|h\rangle = 0$ for $n \geq 1$, so $\langle h|L_{-n}L_n|h\rangle = 0$ may be subtracted.
(3) The Virasoro algebra (6.21) with $m = n$, $n \to -n$: $(n-(-n))L_0 = 2nL_0$ and $\delta_{n,n} = 1$.
(4) $L_0|h\rangle = h|h\rangle$ and $\langle h|h\rangle = 1$.

For $n = 1$ this is $2h$; for $n \geq 2$ the central term enters. Unless h and c are both zero, the right side is nonzero for large n, so $L_{-n}|h\rangle$ cannot vanish. Two immediate consequences, which are the first unitarity conditions:

- $n = 1$: $\|L_{-1}|h\rangle\|^2 = 2h \geq 0$, so $h \geq 0$.
- $n \to \infty$: the cubic term dominates, so $c \geq 0$.

This is why Nash says the highest weight is a pair of *non-negative* numbers (c, h).

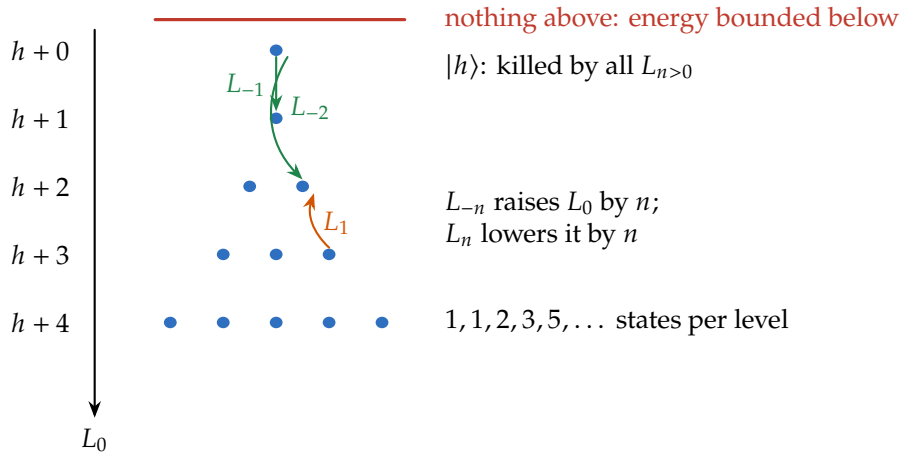


Figure 6.21: The spectrum of L_0 in a highest-weight representation. Each row is an eigenspace of L_0 (drawn downwards, so that the bounded-below end is at the top). The raising operators L_{-n} (green) move down n rows; the lowering operators L_n (orange) move up. The positive-energy condition says there is a top row, and its states are then killed by every L_n with $n > 0$. The number of dots per row is counted in Section 6.3.2.

Watch out: highest weight or lowest weight?

Nash's footnote 4: the name "highest weight" is not ideal, since $|h\rangle$ has the *lowest* eigenvalue of L_0 in the representation. The terms *lowest weight*, *anti-dominant weight* and *positive energy representation* are also used for the same thing. The name comes from Lie theory, where the mode operators $L_{n>0}$ play the role of the positive root vectors (like J_+) and $|h\rangle$ is killed by them, like the state $|j, j\rangle$ of highest J_3. We follow Nash and say "highest weight".

6.3.2 The Verma module: everything a highest-weight state generates

Given (c, h), which representation do we get? We answer this in the way a physicist builds a Fock space: start with a vacuum and apply creation operators. The result is called the Verma module. It is not yet the answer, because some of the states we build can turn out to be "null" and must be thrown away. The whole construction fits in one line:

$$|h\rangle \xrightarrow[\text{descendants}]{\text{generate}} V(c,h) = U(\text{vir}_-)|h\rangle \xrightarrow[\text{submodule } N]{\text{quotient by null}} L(c,h) = V(c,h)/N \tag{6.107}$$

Figure 6.22 shows the three objects. We take the steps one at a time, and at each step we say why the step is legitimate.

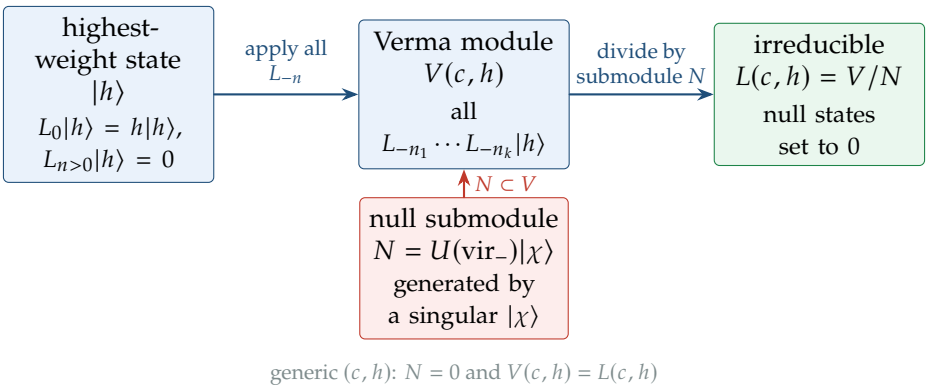


Figure 6.22: The pipeline (6.107). The Verma module (centre) is the largest highest-weight representation with given (c, h). At special values of (c, h) it contains a proper submodule N (red), generated by a singular vector; dividing it out gives the irreducible representation (green), which is the space of states of the physical theory.

A vector space, or a module?

First a word about language, because it is the main source of confusion for a physicist reading the mathematics. We are used to starting with a Hilbert space $\mathcal{H}$ and then asking how the symmetry operators act on it. The highest-weight construction goes the other way: we start from the algebra, and the construction produces the vector space *together with* the action of the algebra on it. Such a pair has a name.

Working definition 6.18 (module)

A **module** over a Lie algebra $\mathfrak{g}$ (a $\mathfrak{g}$-module) is a vector space V together with a linear action $X \mapsto \rho(X)$ of $\mathfrak{g}$ on V that respects brackets:

$$\rho([X,Y]) = \rho(X)\rho(Y) - \rho(Y)\rho(X).$$

Operationally: a module is a *representation space*. "V is a vir-module" means exactly "the L_n act on V as operators obeying the Virasoro algebra". A **submodule** is a subspace $W \subset V$ with $\rho(X)W \subset W$ for all X: an invariant subspace.

Dictionary: modules

Algebra	Physics
module V	representation space; space of states
submodule $W \subset V$	invariant subspace, closed under all L_n
irreducible module	irreducible representation, no invariant subspaces
cyclic vector	a state from which all others are obtained
quotient module V/W	states of V, with those in W declared to be zero
Shapovalov (Kac) form	the inner product $\langle \cdot \mid \cdot \rangle$ on states

Notice what is *not* in the definition of a module: an inner product. The Verma module is built from the algebra alone. The Hilbert space structure enters later, through the rule $L_n^\dagger = L_{-n}$ (Section 6.3.5), and it is the inner product that tells us which states are "null" and must be removed before we have a Hilbert space. Keeping these two things apart is the key to the whole construction.

Step 1: the descendants

The **descendants** of $|h\rangle$ are all the states obtained by acting with the raising operators:

$$|h\rangle, \quad L_{-1}|h\rangle, \quad L_{-2}|h\rangle, \quad L_{-1}^2|h\rangle, \quad L_{-3}|h\rangle, \quad L_{-2}L_{-1}|h\rangle, \quad L_{-1}L_{-2}|h\rangle, \ \ldots$$

Not all orderings are needed. A product in the "wrong" order can be rewritten in the standard order using the commutators, and the commutators of raising operators are again raising operators. For example

$$L_{-1}L_{-2}|h\rangle \overset{(1)}{=} L_{-2}L_{-1}|h\rangle + [L_{-1}, L_{-2}]|h\rangle \overset{(2)}{=} L_{-2}L_{-1}|h\rangle + L_{-3}|h\rangle.$$

(1) Add and subtract $L_{-2}L_{-1}|h\rangle$.
(2) $[L_{-1}, L_{-2}] = (-1+2)L_{-3} = L_{-3}$; no central term since $-1-2 \neq 0$.

In general, repeating this move puts every product in the form

$$L_{-n_1}L_{-n_2}\cdots L_{-n_k}|h\rangle, \qquad n_1 \geq n_2 \geq \cdots \geq n_k \geq 1, \tag{6.108}$$

plus shorter products. (This is the Poincaré–Birkhoff–Witt theorem; for a physicist it is the statement that we may normal order.) The states (6.108) are the analogue of the Fock states $a_{n_1}^\dagger \cdots a_{n_k}^\dagger|0\rangle$.

By (6.104), the state (6.108) has $L_0 = h + N$ with $N = n_1 + \cdots + n_k$. We call N the **level**. The number of states (6.108) at level N is the number of ways of writing N as an unordered sum of positive integers: the **partition number** $p(N)$.

level N	basis of $V(c,h)$ at level N	$p(N)$
0	$\lvert h\rangle$	1
1	$L_{-1}\lvert h\rangle$	1
2	$L_{-2}\lvert h\rangle,\ L_{-1}^2\lvert h\rangle$	2
3	$L_{-3}\lvert h\rangle,\ L_{-2}L_{-1}\lvert h\rangle,\ L_{-1}^3\lvert h\rangle$	3
4	$L_{-4},\ L_{-3}L_{-1},\ L_{-2}^2,\ L_{-2}L_{-1}^2,\ L_{-1}^4$ (on $\lvert h\rangle$)	5
5	$\ldots$	7

All the levels are packaged in the generating function

$$\sum_{N\geq 0} p(N)\, q^N \overset{(1)}{=} \prod_{n\geq 1}\left(1+q^n+q^{2n}+\cdots\right) \overset{(2)}{=} \prod_{n\geq 1}\frac{1}{1-q^n} = 1+q+2q^2+3q^3+5q^4+7q^5+\cdots \tag{6.109}$$

(1) A state (6.108) is fixed by how many times each L_{-n} appears; j copies of L_{-n} contribute q^{jn}.
(2) Geometric series.

The **character** of the Verma module, the trace of q^{L_0}, is then $\chi_{V(c,h)}(q) = q^h \prod_n (1-q^n)^{-1}$. (It is the partition function of one free boson without zero mode, a fact we use below.)

Working definition 6.19 (Verma module)

The **Verma module** $V(c,h)$ is the vector space with basis (6.108) [the states $L_{-n_1}\cdots L_{-n_k}|h\rangle$ with $n_1 \geq \cdots \geq n_k \geq 1$, and $|h\rangle$ itself], with the L_n acting by the rule: *write L_n in front, and use the commutation relations to move it to the right until it hits $|h\rangle$, where $L_{n>0}|h\rangle = 0$, $L_0|h\rangle = h|h\rangle$ and c acts as the number c.* In symbols $V(c,h) = U(\mathrm{vir}_-)|h\rangle$, where $U(\mathrm{vir}_-)$ is the space of all polynomials in the raising operators $L_{-1}, L_{-2}, \ldots$ (the **universal enveloping algebra** of the subalgebra vir_- they span). Formally, $V(c,h) = U(\mathrm{vir}) \otimes_{U(\mathrm{vir}_{\geq 0})} \mathbb{C}_{c,h}$ [$\mathbb{C}_{c,h}$ is the line spanned by $|h\rangle$, on which $L_{n>0}$ act as 0, L_0 as h and the central element as c].

The Verma module contains everything the Virasoro algebra can generate from $|h\rangle$, and it imposes *no relations* beyond the defining ones (6.105). That makes it *universal*: every highest-weight representation with the same (c,h) is a quotient of $V(c,h)$, because every state in it is some polynomial in the L_{-n} acting on its vacuum.

Worked example 6.37: the action of L_1 and L_2 at level 2

We compute the lowering operators on the two states at level 2. These four numbers are all we need for the null vectors below. First L_1:

$$\begin{aligned} L_1L_{-2}|h\rangle &\overset{(1)}{=} [L_1,L_{-2}]|h\rangle \overset{(2)}{=} 3L_{-1}|h\rangle, \\ L_1L_{-1}^2|h\rangle &\overset{(3)}{=} [L_1,L_{-1}]L_{-1}|h\rangle + L_{-1}[L_1,L_{-1}]|h\rangle \overset{(4)}{=} 2L_0L_{-1}|h\rangle + L_{-1}\,2L_0|h\rangle \\ &\overset{(5)}{=} 2(h+1)L_{-1}|h\rangle + 2h\,L_{-1}|h\rangle = (4h+2)\,L_{-1}|h\rangle. \end{aligned}$$

(1) $L_1|h\rangle = 0$, so $L_1L_{-2}|h\rangle = [L_1,L_{-2}]|h\rangle$.
(2) $[L_1,L_{-2}] = (1+2)L_{-1} = 3L_{-1}$; no central term since $1-2\neq 0$.
(3) Move L_1 through the two factors; the term with L_1 on $|h\rangle$ vanishes.
(4) $[L_1,L_{-1}] = 2L_0$; the central term $\frac{c}{12}\cdot 1\cdot 0$ vanishes.
(5) $L_{-1}|h\rangle$ has $L_0 = h+1$.

Then L_2:

$$\begin{aligned} L_2L_{-2}|h\rangle &\overset{(1)}{=} [L_2,L_{-2}]|h\rangle \overset{(2)}{=} \left(4L_0 + \frac{c}{12}\cdot 2\cdot 3\right)|h\rangle = \left(4h+\frac{c}{2}\right)|h\rangle, \\ L_2L_{-1}^2|h\rangle &\overset{(3)}{=} [L_2,L_{-1}]L_{-1}|h\rangle + L_{-1}[L_2,L_{-1}]|h\rangle \overset{(4)}{=} 3L_1L_{-1}|h\rangle + 3L_{-1}L_1|h\rangle \overset{(5)}{=} 6h|h\rangle. \end{aligned}$$

(1) $L_2|h\rangle = 0$.
(2) $[L_2,L_{-2}] = 4L_0 + \frac{c}{12}\,2(2^2-1)$.
(3) Move L_2 through.
(4) $[L_2,L_{-1}] = 3L_1$.
(5) $L_1L_{-1}|h\rangle = [L_1,L_{-1}]|h\rangle = 2h|h\rangle$, and $L_1|h\rangle = 0$.

In the bases $(L_{-2}|h\rangle, L_{-1}^2|h\rangle)$ at level 2, $L_{-1}|h\rangle$ at level 1 and $|h\rangle$ at level 0, the lowering operators are the matrices

$$L_1\big|_{2\to1} = \begin{pmatrix} 3 & 4h+2 \end{pmatrix}, \qquad L_2\big|_{2\to0} = \begin{pmatrix} 4h+\frac{c}{2} & 6h \end{pmatrix}.$$

This is a representation made completely explicit: every matrix element of every L_n on $V(c,h)$ can be computed this way.

Try it: level 3

Show that $L_1L_{-3}|h\rangle = 4L_{-2}|h\rangle$, $L_1L_{-2}L_{-1}|h\rangle = 2h\,L_{-2}|h\rangle + 3L_{-1}^2|h\rangle$ and $L_1L_{-1}^3|h\rangle = (6h+6)L_{-1}^2|h\rangle$. *Answer:* use $[L_1, L_{-n}] = (n+1)L_{1-n}$ and the level-2 results; for the last, $L_1L_{-1}^3 = \sum_{j=0}^{2} L_{-1}^{j}\,2L_0\,L_{-1}^{2-j}$ on $|h\rangle$ gives $2\big((h+2)+(h+1)+h\big) = 6h+6$.

6.3.3 Singular vectors and the null submodule

Step 2: a singular vector can appear

At special values of (c,h) something new happens: a descendant is *itself* killed by all the lowering operators. It then behaves like a second highest-weight state sitting inside $V(c,h)$.

Working definition 6.20 (singular vector)

A state $|\chi\rangle \neq 0$ in $V(c,h)$ at level $N>0$ [$L_0|\chi\rangle = (h+N)|\chi\rangle$] is **singular** if

$$L_n|\chi\rangle = 0 \quad \text{for all } n \geq 1.$$

It is then a highest-weight vector of weight $(c, h+N)$. *Test:* it is enough to check $L_1|\chi\rangle = 0$ and $L_2|\chi\rangle = 0$, because every L_n with $n \geq 3$ is a commutator of these: $[L_2, L_1] = L_3$ and $[L_n, L_1] = (n-1)L_{n+1}$, so $L_1|\chi\rangle = L_2|\chi\rangle = 0$ implies $L_3|\chi\rangle = L_2L_1|\chi\rangle - L_1L_2|\chi\rangle = 0$, and so on by induction.

Level 1. The only candidate is $|\chi\rangle = L_{-1}|h\rangle$. Since $[L_1, L_{-1}] = 2L_0$, we have $L_1L_{-1}|h\rangle = 2h|h\rangle$; and $L_2L_{-1}|h\rangle = 3L_1|h\rangle = 0$ automatically. So $L_{-1}|h\rangle$ is singular exactly when $h=0$. This is the familiar case: the vacuum of a conformal field theory is invariant under L_{-1} (translations), and $L_{-1}|0\rangle$ is the first null vector.

Level 2. The general candidate, normalised so that the coefficient of L_{-2} is 1, is

$$|\chi\rangle = \big(L_{-2} + \alpha\, L_{-1}^2\big)|h\rangle. \tag{6.110}$$

(If the coefficient of L_{-2} vanished, $|\chi\rangle \propto L_{-1}^2|h\rangle$, and $L_1L_{-1}^2|h\rangle = (4h+2)L_{-1}|h\rangle$ would force $h = -\frac{1}{2}$, but then $L_2L_{-1}^2|h\rangle = 6h|h\rangle \neq 0$; so this case never occurs.) Now impose the two conditions, using Example 6.37:

$$L_1|\chi\rangle \overset{(1)}{=} L_1L_{-2}|h\rangle + \alpha\, L_1L_{-1}^2|h\rangle \overset{(2)}{=} \big(3 + \alpha(4h+2)\big)L_{-1}|h\rangle,$$
$$L_2|\chi\rangle \overset{(3)}{=} L_2L_{-2}|h\rangle + \alpha\, L_2L_{-1}^2|h\rangle \overset{(4)}{=} \Big(4h + \frac{c}{2} + 6h\alpha\Big)|h\rangle.$$

(1) Linearity of L_1.
(2) $L_1L_{-2}|h\rangle = 3L_{-1}|h\rangle$ and $L_1L_{-1}^2|h\rangle = (4h+2)L_{-1}|h\rangle$.
(3) Linearity of L_2.

(4) $L_2L_{-2}|h\rangle = (4h + \frac{c}{2})|h\rangle$ and $L_2L_{-1}^2|h\rangle = 6h|h\rangle$.

The first condition fixes α and the second then fixes c in terms of h:

$$L_1|\chi\rangle = 0 \iff \alpha = -\frac{3}{2(2h+1)}, \tag{6.111}$$

and, with this α, $L_2|\chi\rangle = 0$ says

$$\frac{c}{2} \overset{(1)}{=} -4h - 6h\alpha \overset{(2)}{=} -4h + \frac{9h}{2h+1} \overset{(3)}{=} \frac{h(5-8h)}{2h+1}, \qquad \text{i.e.} \quad c = \frac{2h(5-8h)}{2h+1}. \tag{6.112}$$

(1) Set the coefficient $4h + \frac{c}{2} + 6h\alpha$ to zero.
(2) Substitute (6.111): $-6h\alpha = \frac{18h}{2(2h+1)} = \frac{9h}{2h+1}$.
(3) Common denominator: $-4h(2h+1) + 9h = 5h - 8h^2$.

Clearing denominators, (6.112) is the quadratic

$$16h^2 + 2(c-5)h + c = 0, \qquad h_\pm(c) = \frac{5 - c \pm \sqrt{(1-c)(25-c)}}{16}. \tag{6.113}$$

(Check: $c(2h+1) = 10h - 16h^2$ rearranges to $16h^2 + (2c-10)h + c = 0$; the discriminant is $4(c-5)^2 - 64c = 4(c^2 - 26c + 25) = 4(c-1)(c-25)$.) So a level-2 singular vector exists exactly on the two curves $h = h_\pm(c)$ in the (c, h) plane.

Worked example 6.38: the Ising model: $c = \frac{1}{2}$

At $c = \frac{1}{2}$, (6.113) gives

$$h_\pm\Big(\frac{1}{2}\Big) \overset{(1)}{=} \frac{\frac{9}{2} \pm \sqrt{\frac{1}{2}\cdot\frac{49}{2}}}{16} \overset{(2)}{=} \frac{\frac{9}{2} \pm \frac{7}{2}}{16} = \frac{1}{2} \text{ or } \frac{1}{16}.$$

(1) $5 - c = \frac{9}{2}$, $(1-c)(25-c) = \frac{1}{2}\cdot\frac{49}{2}$.
(2) $\sqrt{49/4} = \frac{7}{2}$.

These are the conformal weights of the energy operator ε ($h = \frac{1}{2}$) and the spin operator σ ($h = \frac{1}{16}$) of the two-dimensional Ising model at its critical point (Chapter 11). The singular vectors are, by (6.111),

$$|\chi_\varepsilon\rangle = \Big(L_{-2} - \frac{3}{4}L_{-1}^2\Big)|\tfrac{1}{2}\rangle, \qquad |\chi_\sigma\rangle = \Big(L_{-2} - \frac{4}{3}L_{-1}^2\Big)|\tfrac{1}{16}\rangle.$$

(For $h = \frac{1}{2}$: $\alpha = -\frac{3}{2\cdot 2}$; for $h = \frac{1}{16}$: $\alpha = -\frac{3}{2\cdot\frac{9}{8}} = -\frac{4}{3}$.) In the field theory, setting these states to zero becomes a second-order differential equation for the correlation functions of σ and ε. This is how the Ising correlators are actually computed (Belavin, Polyakov and Zamolodchikov). At $c = 0$ the curves give $h = 0$ and $h = \frac{5}{8}$; at $c = 1$ they meet at $h = \frac{1}{4}$.

A singular vector generates a whole submodule

Here is the central point, and the one most often misunderstood. A single singular vector does not come alone. Since $|\chi\rangle$ is a highest-weight vector, we can apply the raising operators to it, exactly as we did to $|h\rangle$:

$$N \coloneqq U(\mathrm{vir}_-)|\chi\rangle = \mathrm{span}\big\{|\chi\rangle,\ L_{-1}|\chi\rangle,\ L_{-2}|\chi\rangle,\ L_{-1}^2|\chi\rangle,\ L_{-3}|\chi\rangle, \dots\big\}. \tag{6.114}$$

This is a Verma module in its own right, $N \cong V(c, h+2)$, sitting inside $V(c,h)$ with its top at level 2. So N has $p(N'-2)$ states at level N': $1,1,2,3,\ldots$ at levels $2,3,4,5,\ldots$ (Figure 6.23).

N is a submodule. We must check that every L_n maps N into itself. For $n<0$ this is the definition of N. For $n \geq 0$, move L_n to the right through the raising operators. Each commutator produces another Virasoro generator, and when a lowering operator finally reaches $|\chi\rangle$ it gives zero, while L_0 and c give numbers. Two instances:

$$L_1\,L_{-1}|\chi\rangle \overset{(1)}{=} [L_1,L_{-1}]|\chi\rangle + L_{-1}L_1|\chi\rangle \overset{(2)}{=} 2L_0|\chi\rangle \overset{(3)}{=} 2(h+2)|\chi\rangle \;\in N,$$
$$L_2\,L_{-1}|\chi\rangle \overset{(4)}{=} [L_2,L_{-1}]|\chi\rangle + L_{-1}L_2|\chi\rangle \overset{(5)}{=} 3L_1|\chi\rangle \overset{(6)}{=} 0 \;\in N.$$

(1) Add and subtract $L_{-1}L_1|\chi\rangle$.
(2) $[L_1,L_{-1}] = 2L_0$ and $L_1|\chi\rangle = 0$ ($|\chi\rangle$ is singular).
(3) $|\chi\rangle$ is at level 2, so $L_0|\chi\rangle = (h+2)|\chi\rangle$.
(4) As in (1).
(5) $[L_2,L_{-1}] = 3L_1$ and $L_2|\chi\rangle = 0$.
(6) $L_1|\chi\rangle = 0$.

The general case is the same bookkeeping: $L_n\,U(\mathrm{vir}_-)|\chi\rangle$ is a combination of terms $U(\mathrm{vir}_-)\,X|\chi\rangle$ with $X \in \{L_{m>0}, L_0, c\}$, and each $X|\chi\rangle$ is 0 or a multiple of $|\chi\rangle$.

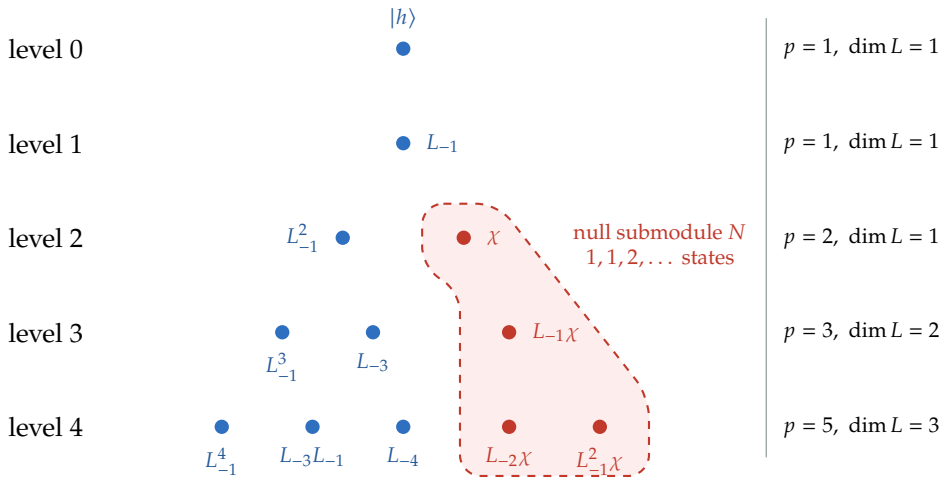


Figure 6.23: The Verma module $V(c,h)$ level by level, with $p(N) = 1,1,2,3,5$ states at levels $0,\ldots,4$, drawn for a point (c,h) on one of the curves (6.113). The red states span the null submodule N generated by the singular vector $|\chi\rangle$ at level 2: $1,1,2$ states at levels $2,3,4$. The blue states are representatives of the quotient $L(c,h) = V/N$, which therefore has $1,1,1,2,3$ states. (The blue labels abbreviate $L_{-1}^2|h\rangle$ etc.; any set of blue states that together with the red ones forms a basis would do.)

We should check that the blue states in Figure 6.23 really complement the red ones. At level 3 the red state is $L_{-1}|\chi\rangle = \big(L_{-1}L_{-2} + \alpha L_{-1}^3\big)|h\rangle = \big(L_{-2}L_{-1} + L_{-3} + \alpha L_{-1}^3\big)|h\rangle$, which contains the basis vector $L_{-2}L_{-1}|h\rangle$ that is not among the blue ones; so $\{L_{-1}^3, L_{-3}, L_{-1}\chi\}$ is a basis. At level 4 a short calculation gives $L_{-1}^2L_{-2} = L_{-2}L_{-1}^2 + 2L_{-3}L_{-1} + 2L_{-4}$, so the red states $L_{-2}|\chi\rangle = (L_{-2}^2 + \alpha L_{-2}L_{-1}^2)|h\rangle$ and $L_{-1}^2|\chi\rangle = (L_{-2}L_{-1}^2 + \alpha L_{-1}^4 + 2L_{-3}L_{-1} + 2L_{-4})|h\rangle$ supply the two missing directions L_{-2}^2 and $L_{-2}L_{-1}^2$.

6.3.4 The quotient: the irreducible representation

Step 3: why quotient?

When $V(c,h)$ contains the submodule N, the representation $V(c,h)$ is *reducible*: N is an invariant subspace. In a physical theory the state $|\chi\rangle$ is not an independent state. We will see in Section 6.3.5 that it has zero norm and is orthogonal to every state, so no measurement can detect it: it is zero in the Hilbert space. And because the algebra acts on it, everything in the sector it generates, $L_{-1}|\chi\rangle$, $L_{-2}|\chi\rangle$, and so on, is equally redundant. So we remove the whole null submodule.

Working definition 6.21 (quotient module)

Given a submodule $N \subset V$ [a subspace that every L_n maps into itself, here all states built on $|\chi\rangle$], declare two states equivalent if they differ by an element of N:

$$v_1 \sim v_2 \iff v_1 - v_2 \in N.$$

The equivalence class of v is $[v] = v + N$. The **quotient module** is the set of classes, V/N, with the action

$$L_n[v] \coloneqq [L_n v]. \tag{6.115}$$

Operationally: compute with the representatives as usual, and at the end set every element of N to zero. In particular $|\chi\rangle \sim 0$, $L_{-1}|\chi\rangle \sim 0$, $L_{-2}|\chi\rangle \sim 0, \ldots$. For the Virasoro algebra, the result is written

$$L(c,h) \coloneqq V(c,h)/N.$$

Step 4: why the quotient is still a Virasoro module

The definition (6.115) uses a representative v of the class. It makes sense only if a different representative gives the same answer. Suppose $[v] = [v']$. Then

$$[L_n v] - [L_n v'] \overset{(1)}{=} [L_n v - L_n v'] \overset{(2)}{=} [L_n(v - v')] \overset{(3)}{=} [0].$$

(1) Classes add like their representatives: $[a] - [b] = [a - b]$.
(2) L_n is linear.
(3) $v - v' \in N$ because $[v] = [v']$, and $L_n(v - v') \in N$ because N is a *submodule*; an element of N has class $[0]$.

So the action on classes is well defined. It also still obeys the algebra:

$$\big(L_m L_n - L_n L_m\big)[v] \overset{(1)}{=} [L_m L_n v - L_n L_m v] \overset{(2)}{=} \Big[(m-n)L_{m+n}v + \tfrac{c}{12}m(m^2-1)\delta_{m,-n}v\Big]$$
$$\overset{(3)}{=} (m-n)L_{m+n}[v] + \tfrac{c}{12}m(m^2-1)\delta_{m,-n}[v].$$

(1) Apply (6.115) twice.
(2) The Virasoro relations hold on V.
(3) (6.115) and linearity of the class map.

Step (3) of the first chain is the only place where anything could go wrong, and it is *exactly* the statement that N is closed under the action. That is why we quotient by a submodule and not by an arbitrary subspace.

Watch out: do not quotient by the single vector

Suppose we tried to remove only $|\chi\rangle$ itself, i.e. to divide by the one-dimensional subspace $W = \mathbb{C}|\chi\rangle$. Then $[|\chi\rangle] = [0]$, and (6.115) would require

$$[0] \overset{(1)}{=} L_{-1}[0] \overset{(2)}{=} L_{-1}[|\chi\rangle] \overset{(3)}{=} [L_{-1}|\chi\rangle] \overset{(4)}{\neq} [0],$$

(1) L_{-1} is linear, so it maps the zero class to the zero class.
(2) $[|\chi\rangle] = [0]$ in V/W.
(3) The definition (6.115).
(4) $L_{-1}|\chi\rangle$ is a nonzero state at level 3, not a multiple of $|\chi\rangle$, so it is not in W.

a contradiction. The action is *not* well defined on V/W: W is not invariant. The smallest invariant subspace containing $|\chi\rangle$ is the whole of $N = U(\mathrm{vir}_-)|\chi\rangle$. Removing a null state means removing its entire tower of descendants.

Step 5: irreducibility

$L(c,h)$ is irreducible when N is the *maximal* proper submodule of $V(c,h)$. At some special points there are several singular vectors (for instance at every level rs in the minimal models below), and then N must be taken to be the sum of all the submodules they generate. A clean characterisation of the maximal proper submodule comes from the inner product (Section 6.3.5): it consists of *all* the null states.

	what it contains	**states per level**
Verma module $V(c,h)$	all descendants $L_{-n_1}\cdots L_{-n_k}\|h\rangle$, no relations beyond (6.105)	$p(N)$: $1,1,2,3,5,7,\ldots$
null submodule N	all descendants of the singular (null) vectors; zero-norm states orthogonal to everything	e.g. level-2 singular vector: $p(N-2)$: $0,0,1,1,2,3,\ldots$
irreducible $L(c,h)$	V/N: the Verma module with the null states set to zero; nondegenerate inner product	$p(N) - \dim N_N$: $1,1,1,2,3,4,\ldots$

Step 6: generically nothing happens

A Verma module need not contain any null state. The level-1 and level-2 conditions are single equations in the (c,h) plane ($h = 0$; $h = h_\pm(c)$), and the conditions at higher levels are more curves of the same kind ((6.119) below). For *generic* (c,h), off all these curves, $N = \{0\}$ and

$$V(c,h) = L(c,h)\,:$$

the Verma module is already irreducible. The Verma module is the universal highest-weight representation built from (c,h); the physical theory's irreducible representation is what remains after imposing the null-state relations, and generically there are none to impose.

6.3.5 Where the inner product enters: the Kac form

So far we have a module but no Hilbert space. The physicist's Hilbert space needs an inner product, and in a unitary representation the inner product must satisfy $L_n^\dagger = L_{-n}$ (6.103). This single rule determines the inner product on the whole Verma module.

Working definition 6.22 (Shapovalov or Kac form)

On $V(c,h)$ define $\langle\cdot|\cdot\rangle$ by $\langle h|h\rangle = 1$ and $L_n^\dagger = L_{-n}$. To compute $\langle v|w\rangle$ for two basis states (6.108): take the adjoint of v, which reverses the order and flips the signs of the mode numbers, $\big(L_{-n_1}\cdots L_{-n_k}|h\rangle\big)^\dagger = \langle h|L_{n_k}\cdots L_{n_1}$; then move the lowering operators [the L_n with $n>0$, each lowers L_0 by n] to the right with the commutation relations until they hit $|h\rangle$. The **Gram matrix** $M_N(c,h)$ at level N is the $p(N)\times p(N)$ matrix of inner products of the basis states at level N. Its determinant is the **Kac determinant** [it vanishes exactly when some nonzero state at level N is orthogonal to every state at that level].

Different levels are orthogonal. If v and w have levels $N \neq N'$, then, since L_0 is Hermitian,

$$(h+N)\langle v|w\rangle \overset{(1)}{=} \langle v|L_0^\dagger|w\rangle \overset{(2)}{=} \langle v|L_0|w\rangle \overset{(3)}{=} (h+N')\langle v|w\rangle,$$

(1) $L_0|v\rangle = (h+N)|v\rangle$ with h real, and $\langle v|L_0^\dagger = (L_0|v\rangle)^\dagger$.
(2) $L_0^\dagger = L_0$.
(3) $L_0|w\rangle = (h+N')|w\rangle$.

so $\langle v|w\rangle = 0$. The form is therefore block diagonal, one block M_N per level, and we only need the Gram matrices.

Worked example 6.39: the Gram matrices at levels 1 and 2

At level 1, $M_1 = \|L_{-1}|h\rangle\|^2 = 2h$ by (6.106). At level 2, in the basis $(L_{-2}|h\rangle, L_{-1}^2|h\rangle)$,

$$\langle h|L_2L_{-2}|h\rangle \overset{(1)}{=} 4h+\tfrac{c}{2}, \qquad \langle h|L_2L_{-1}^2|h\rangle \overset{(2)}{=} 6h,$$
$$\langle h|L_1L_1L_{-1}^2|h\rangle \overset{(3)}{=} (4h+2)\langle h|L_1L_{-1}|h\rangle \overset{(4)}{=} (4h+2)\,2h.$$

(1) Example 6.37.
(2) Example 6.37; the adjoint of $L_{-2}|h\rangle$ is $\langle h|L_2$.
(3) $L_1L_{-1}^2|h\rangle = (4h+2)L_{-1}|h\rangle$ (Example 6.37); the adjoint of $L_{-1}^2|h\rangle$ is $\langle h|L_1L_1$.
(4) $\langle h|L_1L_{-1}|h\rangle = 2h$.

So

$$M_2(c,h) = \begin{pmatrix} 4h+\frac{c}{2} & 6h \\ 6h & 8h^2+4h \end{pmatrix},$$

and its determinant factorises:

$$\det M_2 \overset{(1)}{=} \Big(4h+\frac{c}{2}\Big)(8h^2+4h) - 36h^2 \overset{(2)}{=} 32h^3 - 20h^2 + 4ch^2 + 2ch$$
$$\overset{(3)}{=} 2h\big(16h^2+2(c-5)h+c\big) \overset{(4)}{=} 32\,h\,\big(h-h_+(c)\big)\big(h-h_-(c)\big). \tag{6.116}$$

(1) Determinant of a 2×2 matrix.
(2) Expand: $32h^3+16h^2+4ch^2+2ch-36h^2$.
(3) Take out $2h$.
(4) The roots of the quadratic are $h_\pm(c)$ of (6.113), and its leading coefficient is 16.

The determinant vanishes exactly where there is a singular vector: at $h=0$ (where $L_{-1}|h\rangle$ is singular, and its descendant $L_{-1}^2|h\rangle$ is in the null submodule) and on the curves $h_\pm(c)$ found in (6.113).

Null states are the zero-norm states

A singular vector has zero norm, and is orthogonal to *every* state. For $|\chi\rangle$ at level 2:

$$\langle\chi|\chi\rangle \overset{(1)}{=} \langle h|\big(L_2 + \bar{\alpha}L_1L_1\big)|\chi\rangle \overset{(2)}{=} \langle h|L_2|\chi\rangle + \bar{\alpha}\langle h|L_1\big(L_1|\chi\rangle\big) \overset{(3)}{=} 0.$$

(1) $\langle\chi| = \big((L_{-2} + \alpha L_{-1}^2)|h\rangle\big)^\dagger$ and $L_n^\dagger = L_{-n}$.
(2) Linearity.
(3) $L_2|\chi\rangle = 0$ and $L_1|\chi\rangle = 0$.

The same computation shows $\langle v|\chi\rangle = 0$ for every v at level 2, since $\langle v|$ is $\langle h|$ times a polynomial in lowering operators. Then for any descendant $L_{-m}|\chi\rangle \in N$ and any $v \in V$,

$$\langle v|L_{-m}\chi\rangle \overset{(1)}{=} \langle L_m v|\chi\rangle \overset{(2)}{=} 0,$$

(1) $L_{-m}^\dagger = L_m$.
(2) $L_m v$ is again a state of V, and $|\chi\rangle$ is orthogonal to all of them (both have level 2 or the levels differ).

and by induction every state of N is orthogonal to every state of V. So N lies in the **radical** of the form,

$$\mathrm{Rad} = \{w \in V : \ \langle v|w\rangle = 0 \text{ for all } v \in V\},$$

the null states. Conversely, *every* proper submodule $W \subset V$ consists of null states. Such a W contains no multiple of $|h\rangle$ (otherwise it would contain all descendants and not be proper). For $w \in W$ at level N and a basis state v at level N,

$$\langle v|w\rangle \overset{(1)}{=} \langle h|\, L_{n_k}\cdots L_{n_1}|w\rangle \overset{(2)}{=} 0,$$

(1) The adjoint of $v = L_{-n_1}\cdots L_{-n_k}|h\rangle$.
(2) $L_{n_k}\cdots L_{n_1}|w\rangle$ lies in W (a submodule) and at level 0; but W has nothing at level 0.

So the radical is the *maximal* proper submodule, and the irreducible module is

$$L(c,h) = V(c,h)/\mathrm{Rad}, \tag{6.117}$$

on which the form is nondegenerate. In particular $V(c,h)$ is irreducible if and only if $\det M_N(c,h) \neq 0$ for every level N.

Key idea: why unitarity needs the null states removed

A Hilbert space has a positive definite inner product: the only vector of zero length is 0. A null state $|\chi\rangle \neq 0$ has $\langle\chi|\chi\rangle = 0$ and zero overlap with every state, so all transition amplitudes involving it vanish. It is physically indistinguishable from 0. Keeping it would give a space with a degenerate inner product, which is not a Hilbert space. Removing the whole null submodule, $L = V/\mathrm{Rad}$, is exactly what turns the module into a pre-Hilbert space. Whether the form on L is then *positive* is a further question, the unitarity question.

The Kac determinant formula

The level-2 result generalises. Parametrise the central charge by t:

$$c = 13 - 6\Big(t + \frac{1}{t}\Big), \qquad h_{r,s}(t) = \frac{(r^2-1)\,t}{4} - \frac{rs-1}{2} + \frac{(s^2-1)}{4t}, \quad r,s = 1,2,3,\ldots \tag{6.118}$$

The result, found by Kac and proved by Feigin and Fuchs, is

$$\det M_N(c,h) = \text{const} \times \prod_{\substack{r,s\geq 1\\ rs\leq N}} \big(h - h_{r,s}(c)\big)^{p(N-rs)}, \tag{6.119}$$

with a positive constant. The factor $h - h_{r,s}$ first appears at level rs: there is a singular vector at level rs when $h = h_{r,s}$, and its submodule contributes to all higher levels, with multiplicity $p(N-rs)$.

Check at level 2. The pairs with $rs \leq 2$ are $(1,1)$, $(1,2)$, $(2,1)$. We have $h_{1,1} = 0$, with multiplicity $p(1) = 1$; and

$$h_{1,2} + h_{2,1} \overset{(1)}{=} \left(-\frac{1}{2} + \frac{3}{4t}\right) + \left(\frac{3t}{4} - \frac{1}{2}\right) \overset{(2)}{=} \frac{3}{4}\cdot\frac{13-c}{6} - 1 = \frac{5-c}{8},$$
$$h_{1,2}\,h_{2,1} \overset{(3)}{=} -\frac{3t}{8} + \frac{1}{4} + \frac{9}{16} - \frac{3}{8t} \overset{(4)}{=} \frac{13}{16} - \frac{3}{8}\cdot\frac{13-c}{6} = \frac{c}{16}.$$

(1) (6.118) with $(r,s) = (1,2)$ and $(2,1)$.
(2) $t + \frac{1}{t} = \frac{13-c}{6}$ from (6.118).
(3) Multiply out.
(4) Same substitution.

These are the sum $\frac{5-c}{8}$ and product $\frac{c}{16}$ of the roots of $h^2 + \frac{c-5}{8}h + \frac{c}{16}$, i.e. of (6.113) divided by 16. So $\{h_{1,2}, h_{2,1}\} = \{h_+, h_-\}$ and (6.119) reproduces (6.116).

Worked example 6.40: the Ising Gram matrix at $h = \frac{1}{16}$

At $c = \frac{1}{2}$, $h = \frac{1}{16}$:

$$M_2 = \begin{pmatrix} \frac{1}{4}+\frac{1}{4} & \frac{6}{16} \\ \frac{6}{16} & \frac{8}{256}+\frac{4}{16} \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & \frac{3}{8} \\ \frac{3}{8} & \frac{9}{32} \end{pmatrix}, \qquad \det M_2 = \frac{9}{64} - \frac{9}{64} = 0.$$

The null direction is the singular vector of Example 6.38:

$$M_2 \begin{pmatrix} 1 \\ -\frac{4}{3} \end{pmatrix} \overset{(1)}{=} \begin{pmatrix} \frac{1}{2} - \frac{4}{3}\cdot\frac{3}{8} \\ \frac{3}{8} - \frac{4}{3}\cdot\frac{9}{32} \end{pmatrix} \overset{(2)}{=} \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

(1) Matrix times vector; the vector is the components of $|\chi_\sigma\rangle = (L_{-2} - \frac{4}{3}L_{-1}^2)|\frac{1}{16}\rangle$.
(2) $\frac{4}{3}\cdot\frac{3}{8} = \frac{1}{2}$ and $\frac{4}{3}\cdot\frac{9}{32} = \frac{3}{8}$.

The other eigenvalue is $\operatorname{tr} M_2 = \frac{1}{2} + \frac{9}{32} = \frac{25}{32} > 0$. So the form on level 2 is positive *semi*-definite; after removing the null vector, $L(\frac{1}{2}, \frac{1}{16})$ has one state at level 2, and it has positive norm. This is consistent with the Ising model being a unitary theory.

Worked example 6.41: the free boson: a module where the null states are already zero

Take one free boson (6.61)–(6.63) and let $|p\rangle$ be the Fock vacuum with momentum p: $\alpha_0|p\rangle = p|p\rangle$, $\alpha_n|p\rangle = 0$ for $n > 0$. Then $c = 1$ and $L_0|p\rangle = \frac{1}{2}p^2|p\rangle$, $L_n|p\rangle = 0$ $(n > 0)$: $|p\rangle$ is a highest-weight state with $h = \frac{p^2}{2}$. The Fock space F_p, spanned by $\alpha_{-n_1}\cdots\alpha_{-n_k}|p\rangle$, has $p(N)$ states at level N — the same count as $V(1, \frac{p^2}{2})$. Compare the two. We need $[L_n, \alpha_m] = -m\alpha_{m+n}$ (6.64). At level 1,

$$L_{-1}|p\rangle \overset{(1)}{=} \Big(\alpha_{-1}\alpha_0 + \sum_{m\neq 0,-1} \tfrac{1}{2}\alpha_{-1-m}\alpha_m\Big)|p\rangle \overset{(2)}{=} p\,\alpha_{-1}|p\rangle.$$

(1) In $L_{-1} = \frac{1}{2}\sum_m \alpha_{-1-m}\alpha_m$ the terms $m = 0$ and $m = -1$ are equal and together give $\alpha_{-1}\alpha_0$.
(2) For $m \geq 1$ the annihilator α_m stands on the right and kills $|p\rangle$. For $m \leq -2$ the left factor α_{-1-m} is an annihilator ($-1-m \geq 1$); it commutes with α_m because the mode numbers add to $-1 \neq 0$, so it may be moved to the right, where it kills $|p\rangle$.

At $p = 0$ ($h = 0$) this vanishes: the level-1 null vector of $V(1,0)$ is *literally the zero vector* in the Fock space. At level 2, with $L_{-2}|p\rangle = (p\,\alpha_{-2} + \frac{1}{2}\alpha_{-1}^2)|p\rangle$ (same argument) and

$$L_{-1}^2|p\rangle \overset{(1)}{=} p\,L_{-1}\alpha_{-1}|p\rangle \overset{(2)}{=} p\big([L_{-1},\alpha_{-1}] + \alpha_{-1}L_{-1}\big)|p\rangle \overset{(3)}{=} \big(p\,\alpha_{-2} + p^2\alpha_{-1}^2\big)|p\rangle,$$

(1) The level-1 result.
(2) Add and subtract.
(3) $[L_{-1},\alpha_{-1}] = \alpha_{-2}$ and $L_{-1}|p\rangle = p\alpha_{-1}|p\rangle$.

the singular vector of $V(1,h)$ at $h = h_{\pm}(1) = \frac{1}{4}$ (where $\alpha = -1$ by (6.111)) becomes

$$\big(L_{-2} - L_{-1}^2\big)|p\rangle = \Big(\frac{1}{2} - p^2\Big)\alpha_{-1}^2|p\rangle = 0 \qquad \text{at } p^2 = \tfrac{1}{2}.$$

Again the null vector is zero in the Fock space. This is the general situation in a physical theory: the Hilbert space is built with a positive inner product, so the null states are *automatically* zero there. The Virasoro descendants of $|p\rangle$ span only one state at level 2, whereas F_p has two. The second, orthogonal, one is $|v\rangle = (\alpha_{-2} - 2p\,\alpha_{-1}^2)|p\rangle$, and one checks $L_1|v\rangle = (2-4p^2)\alpha_{-1}|p\rangle = 0$ and $L_2|v\rangle = (2p-2p)|p\rangle = 0$: it is a *new* highest-weight vector, with $h = \frac{1}{4} + 2 = \frac{9}{4}$. So at $p^2 = \frac{1}{2}$ the Fock space is a sum of several irreducible Virasoro modules, $F_p = L(1,\frac{1}{4}) \oplus L(1,\frac{9}{4}) \oplus \cdots$, not a single Verma module. For generic p there are no null vectors, and $F_p \cong V(1,\frac{p^2}{2}) = L(1,\frac{p^2}{2})$.

6.3.6 Unitary representations: the Friedan–Qiu–Shenker series

Now we ask which $L(c,h)$ are *unitary*: when is the Kac form on $L(c,h)$ positive definite? Equivalently, when are all the Gram matrices $M_N(c,h)$ positive semi-definite (with the zero eigenvalues exactly the null states)? Nash quotes the answer, due to Friedan, Qiu and Shenker (necessity) and Goddard, Kent and Olive (existence).

Key idea: the FQS classification (Nash's (6.58))

The irreducible highest-weight representation $L(c,h)$ is unitary if and only if either

$$c \geq 1, \quad h \geq 0,$$

or (c,h) belongs to the **discrete series**

$$c = 1 - \frac{6}{(m+1)(m+2)}, \qquad h = h_{p,q} = \frac{\big((m+2)p - (m+1)q\big)^2 - 1}{4(m+1)(m+2)}, \tag{6.120}$$

with $m = 1,2,3,\ldots$ and $1 \leq q \leq p \leq m$.

How the conditions arise

Levels 1 and ∞. We saw after (6.106) that $h \geq 0$ and $c \geq 0$ are necessary.

Level 2. By (6.116), $\det M_2 = 32h(h-h_+)(h-h_-)$. For $0 < c < 1$ both roots $h_\pm(c)$ are real and positive, with $h_- < h_+$. If $h_-(c) < h < h_+(c)$ then $\det M_2 < 0$: the 2×2 matrix M_2 has one positive and one negative eigenvalue, so some state at level 2 has negative norm. That strip is excluded (Figure 6.24). For $c \geq 1$ the roots coincide at $h = \frac{1}{4}$ ($c = 1$), are complex ($1 < c < 25$) or are real and negative ($c \geq 25$), so $\det M_2 \geq 0$ for all $h \geq 0$; since the diagonal entries of M_2 are then positive, M_2 is positive semi-definite.

Higher levels. Each level excludes more of the region $0 < c < 1$, $h > 0$, along the curves $h = h_{r,s}(c)$ of (6.119). FQS showed that what survives, below $c = 1$, is exactly the set of points where many such curves cross, and these are the points (6.120). For $1 < c < 25$ the values $h_{r,s}$ with $r \neq s$ are complex, and $h_{r,r} = \frac{(r^2-1)(1-c)}{24} \leq 0$ (use $t + \frac{1}{t} = \frac{13-c}{6}$); so no Kac determinant vanishes in the region $h > 0$. The determinants are positive for large h (the diagonal terms h^{power} dominate), hence positive throughout $h > 0$, and everything with $h \geq 0$ survives. (For $c \geq 25$ the same conclusion holds.)

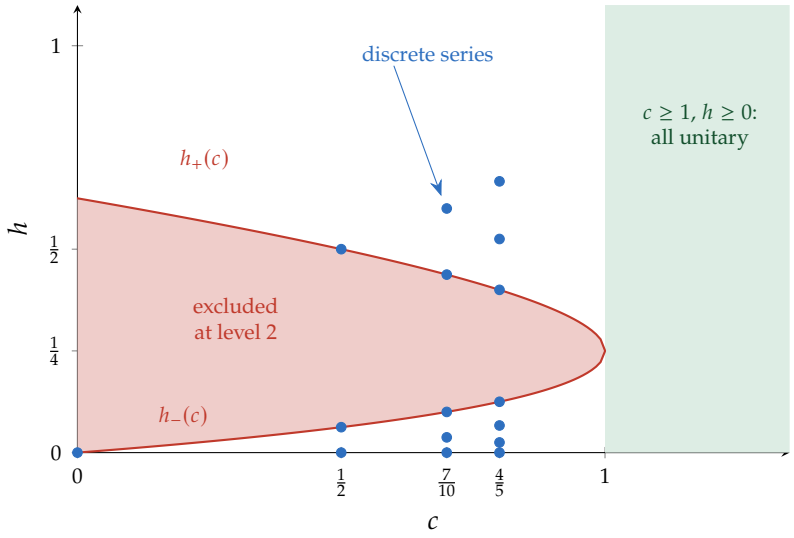


Figure 6.24: The unitarity question in the (c, h) plane. The red strip between the level-2 curves $h_\pm(c)$ of (6.113) is excluded because $\det M_2 < 0$ there. For $c \geq 1$ (green) every $h \geq 0$ is unitary. Below $c = 1$ higher levels exclude everything except the discrete series (blue dots, shown for $m = 1, 2, 3, 4$, i.e. $c = 0, \frac{1}{2}, \frac{7}{10}, \frac{4}{5}$; values $h > 1.1$ are off the picture). Notice that some of the blue dots lie *on* the red curves: these representations have a null vector at level 2, which is removed.

Checking (6.120) *against the Kac formula.* Take $t = \frac{m+2}{m+1}$ in (6.118), and write $a = m+1$, $b = m+2$, so $t = b/a$ and $b - a = 1$. Then

$$c \overset{(1)}{=} 13 - 6\,\frac{a^2+b^2}{ab} \overset{(2)}{=} \frac{13ab - 6(b-a)^2 - 12ab}{ab} \overset{(3)}{=} 1 - \frac{6}{(m+1)(m+2)},$$

$$h_{p,q} \overset{(4)}{=} \frac{(p^2-1)b^2 - 2(pq-1)ab + (q^2-1)a^2}{4ab} \overset{(5)}{=} \frac{(pb-qa)^2 - (b-a)^2}{4ab}$$

$$\overset{(6)}{=} \frac{\big((m+2)p - (m+1)q\big)^2 - 1}{4(m+1)(m+2)}.$$

(1) $t + \frac{1}{t} = \frac{b}{a} + \frac{a}{b} = \frac{a^2+b^2}{ab}$.
(2) $a^2 + b^2 = (b-a)^2 + 2ab$.
(3) $b - a = 1$ and $ab = (m+1)(m+2)$.
(4) (6.118) with $(r, s) = (p, q)$, over the common denominator $4ab$ ($t = b/a$).
(5) Complete the square: $(pb - qa)^2 = p^2b^2 - 2pqab + q^2a^2$, and $-b^2 + 2ab - a^2 = -(b-a)^2$.
(6) $b - a = 1$.

So the discrete series is the set of Kac-table values $h_{p,q}$ at the special central charges (6.120): every unitary representation with $c < 1$ has null vectors, in fact infinitely many.

Worked example 6.42: the first members of the discrete series

Evaluate (6.120) for small m. For $m = 2$, $(m+1)(m+2) = 12$:

$$h_{2,1} \overset{(1)}{=} \frac{(4\cdot 2 - 3\cdot 1)^2 - 1}{48} = \frac{24}{48} = \frac{1}{2}, \qquad h_{2,2} \overset{(2)}{=} \frac{(8-6)^2 - 1}{48} = \frac{1}{16}, \qquad h_{1,1} = 0.$$

(1) $m+2 = 4$, $m+1 = 3$, $(p,q) = (2,1)$.
(2) $(p,q) = (2,2)$.

The full table:

m	c	model	allowed h
1	0	trivial	0
2	$\frac{1}{2}$	Ising	$0,\ \frac{1}{16},\ \frac{1}{2}$
3	$\frac{7}{10}$	tricritical Ising	$0,\ \frac{3}{80},\ \frac{1}{10},\ \frac{7}{16},\ \frac{3}{5},\ \frac{3}{2}$
4	$\frac{4}{5}$	tetracritical Ising / 3-state Potts	$0, \frac{1}{40}, \frac{1}{15}, \frac{1}{8}, \frac{2}{5}, \frac{21}{40}, \frac{2}{3}, \frac{7}{5}, \frac{13}{8}, 3$

At each m there are $\frac{m(m+1)}{2}$ values (one for each $1 \le q \le p \le m$). For $m = 1$ only $h = 0$ survives, and $c = 0$: the trivial representation, in agreement with Section 6.3.7. The Ising values match Example 6.38: both $h = \frac{1}{16}$ and $h = \frac{1}{2}$ lie on the level-2 curves at $c = \frac{1}{2}$, which is why they have level-2 null vectors. The same is true of $\frac{1}{10}, \frac{7}{16}$ at $c = \frac{7}{10}$ and of $\frac{1}{8}, \frac{2}{5}$ at $c = \frac{4}{5}$: check with (6.113), e.g. $h_{\pm}(\frac{4}{5}) = \frac{4.2 \pm \sqrt{0.2\cdot 24.2}}{16} = \frac{4.2 \pm 2.2}{16}$. These discrete-series theories are the **minimal models**; they describe the critical points of two-dimensional statistical models (Chapter 11).

Constructing the discrete series: the GKO coset

Nash says that "representations for all these allowed values can be constructed" (Goddard, Kent and Olive). FQS only showed that nothing *else* can be unitary. The construction uses the Sugawara operators of Section 6.2.10 twice.

Let $\mathfrak{g} \supset \mathfrak{h}$ be a Lie algebra and a subalgebra, and suppose we have a unitary highest-weight representation of the Kac–Moody algebra of $\mathfrak{g}$. The currents of $\mathfrak{h}$ are among those of $\mathfrak{g}$, so we have two Sugawara Virasoro algebras, $L^{\mathfrak{g}}_n$ and $L^{\mathfrak{h}}_n$, with central charges $c_{\mathfrak{g}}$ and $c_{\mathfrak{h}}$. Define

$$K_n := L^{\mathfrak{g}}_n - L^{\mathfrak{h}}_n. \tag{6.121}$$

Step 1: K_n commutes with the $\mathfrak{h}$ currents. Both Sugawara operators act on any current J_m of $\mathfrak{h}$ as the vector field does, $[L^{\mathfrak{g}}_n, J_m] = -mJ_{m+n} = [L^{\mathfrak{h}}_n, J_m]$ (this is Nash's $[L_n, T^a_m] = -mT^a_{n+m}$ in (6.123) below). Subtracting, $[K_n, J_m] = 0$. Since $L^{\mathfrak{h}}_m$ is built from the J's, $[K_n, L^{\mathfrak{h}}_m] = 0$ as well.

Step 2: K_n is a Virasoro algebra.

$$\begin{aligned}
[K_m, K_n] &\overset{(1)}{=} [K_m, L^{\mathfrak{g}}_n] \overset{(2)}{=} [L^{\mathfrak{g}}_m, L^{\mathfrak{g}}_n] - [L^{\mathfrak{h}}_m, L^{\mathfrak{g}}_n] \overset{(3)}{=} [L^{\mathfrak{g}}_m, L^{\mathfrak{g}}_n] - [L^{\mathfrak{h}}_m, L^{\mathfrak{h}}_n] \\
&\overset{(4)}{=} (m-n)\big(L^{\mathfrak{g}}_{m+n} - L^{\mathfrak{h}}_{m+n}\big) + \frac{c_{\mathfrak{g}} - c_{\mathfrak{h}}}{12}\, m(m^2-1)\,\delta_{m,-n} \\
&= (m-n)K_{m+n} + \frac{c_{\mathfrak{g}} - c_{\mathfrak{h}}}{12}\, m(m^2-1)\,\delta_{m,-n}.
\end{aligned}$$

(1) $K_n = L^{\mathfrak{g}}_n - L^{\mathfrak{h}}_n$ and $[K_m, L^{\mathfrak{h}}_n] = 0$ (Step 1).

(2) Expand K_m.
(3) $L^{\mathfrak{g}}_n = K_n + L^{\mathfrak{h}}_n$ and $[L^{\mathfrak{h}}_m, K_n] = -[K_n, L^{\mathfrak{h}}_m] = 0$.
(4) Both are Virasoro algebras, with central charges $c_{\mathfrak{g}}$ and $c_{\mathfrak{h}}$.

So the K_n form a Virasoro algebra with $c = c_{\mathfrak{g}} - c_{\mathfrak{h}}$, the **coset construction**. It is unitary: $K_n^\dagger = K_{-n}$ on the unitary Kac–Moody representation space, whose inner product is positive. One checks that K_0 is also bounded below. Decomposing the space under the K_n therefore gives unitary highest-weight representations of Virasoro with $c = c_{\mathfrak{g}} - c_{\mathfrak{h}}$.

Worked example 6.43: the discrete series from SU(2)

Take $\mathfrak{g} = \mathfrak{su}(2) \oplus \mathfrak{su}(2)$ at levels $(\ell, 1)$ and $\mathfrak{h}$ the diagonal $\mathfrak{su}(2)$, whose currents $J^a_n = T^{a(1)}_n + T^{a(2)}_n$ have level $\ell + 1$ (levels add, because the central terms add). With $c(\mathfrak{su}(2)_\ell) = \frac{3\ell}{\ell+2}$ (Example 6.31),

$$c \overset{(1)}{=} \frac{3\ell}{\ell+2} + 1 - \frac{3(\ell+1)}{\ell+3} \overset{(2)}{=} \frac{3\ell(\ell+3) + (\ell+2)(\ell+3) - 3(\ell+1)(\ell+2)}{(\ell+2)(\ell+3)}$$
$$\overset{(3)}{=} \frac{\ell^2 + 5\ell}{(\ell+2)(\ell+3)} \overset{(4)}{=} 1 - \frac{6}{(\ell+2)(\ell+3)}.$$

(1) $c_{\mathfrak{g}} = c(\ell) + c(1)$ and $c_{\mathfrak{h}} = c(\ell+1)$; $c(1) = \frac{3}{3} = 1$.
(2) Common denominator.
(3) Expand: $3\ell^2 + 9\ell + \ell^2 + 5\ell + 6 - 3\ell^2 - 9\ell - 6$.
(4) $(\ell+2)(\ell+3) - 6 = \ell^2 + 5\ell$.

With $m = \ell + 1$ this is exactly the discrete series (6.120). For $\ell = 1$ it is $c = \frac{1}{2}$: the Ising model is $\frac{\mathrm{SU}(2)_1 \times \mathrm{SU}(2)_1}{\mathrm{SU}(2)_2}$. Every value $c < 1$ allowed by FQS therefore does occur.

The other side: Sugawara gives $c \geq 1$

Nash contrasts the discrete series with the Sugawara construction, which gives unitary representations "of the opposite type". By (6.81) the Sugawara central charge is $c = \frac{2k \dim G}{2k + C}$, and on a unitary Kac–Moody representation (integer level $N = 2k \geq 1$, Section 6.3.9) this is never less than one. For SU(2),

$$c - 1 \overset{(1)}{=} \frac{3N}{N+2} - 1 \overset{(2)}{=} \frac{2(N-1)}{N+2} \overset{(3)}{\geq} 0,$$

(1) $\dim \mathrm{SU}(2) = 3$, $C = 2$, $2k = N$.
(2) Common denominator.
(3) $N \geq 1$.

with equality at level one, where the theory is a single free boson (the bosonisation of Example 6.32). In general Nash notes the bound $c \geq \operatorname{rank} G \geq 1$: at level one for simply laced G the Sugawara theory is rank G free bosons on the root lattice (the Frenkel–Kac construction), and c increases with the level towards $\dim G$.

6.3.7 True representations are trivial (Goddard–Olive)

Finally we prove Nash's claim that the central extension is unavoidable. Suppose the representation is a *true* representation of $\mathrm{vect}(S^1)$ (or of $\mathrm{Diff}(S^1)$), not a projective one. Then the central term is absent: $c = 0$. Assume also that it is unitary, with L_0 bounded below, so that it has a highest-weight vector $|h\rangle$.

Step 1: $c = 0$ forces $h = 0$. This is the result of Goddard and Olive. Consider the two states $L_{-2n}|h\rangle$ and $L_{-n}^2|h\rangle$ at level $2n$. With $c = 0$ the computations of Example 6.39, repeated with n in place of 1, give

$$\langle h|L_{2n}L_{-2n}|h\rangle \overset{(1)}{=} 4nh, \qquad \langle h|L_{2n}L_{-n}^2|h\rangle \overset{(2)}{=} 3n\langle h|L_nL_{-n}|h\rangle = 6n^2h,$$
$$\langle h|L_n^2L_{-n}^2|h\rangle \overset{(3)}{=} \big(2n(h+n)+2nh\big)\langle h|L_nL_{-n}|h\rangle = (4nh+2n^2)\,2nh.$$

(1) (6.106) with $2n$ in place of n and $c = 0$.
(2) $L_{2n}L_{-n}^2|h\rangle = [L_{2n}, L_{-n}]L_{-n}|h\rangle + L_{-n}[L_{2n}, L_{-n}]|h\rangle = 3nL_nL_{-n}|h\rangle + 3nL_{-n}L_n|h\rangle$, and $L_n|h\rangle = 0$.
(3) $L_nL_{-n}^2|h\rangle = [L_n, L_{-n}]L_{-n}|h\rangle + L_{-n}[L_n, L_{-n}]|h\rangle = 2nL_0L_{-n}|h\rangle + 2nh\,L_{-n}|h\rangle$ with $c = 0$, and $L_{-n}|h\rangle$ has $L_0 = h + n$.

The Gram matrix of the two states therefore has determinant

$$\det\begin{pmatrix} 4nh & 6n^2h \\ 6n^2h & 2nh(4nh+2n^2)\end{pmatrix} \overset{(1)}{=} 8n^2h^2(4nh+2n^2) - 36n^4h^2 \overset{(2)}{=} 4n^3h^2\,(8h-5n).$$

(1) 2×2 determinant.
(2) $32n^3h^3 + 16n^4h^2 - 36n^4h^2 = 32n^3h^3 - 20n^4h^2$.

If $h > 0$, choose $n > \frac{8h}{5}$: the determinant is negative, so one of the states has negative norm, contradicting unitarity. Hence $h = 0$.

Step 2: then every $L_{-n}|h\rangle$ vanishes. Nash's (6.59) is (6.106) with $c = h = 0$:

$$\|L_{-n}|h\rangle\|^2 \overset{(1)}{=} \langle h|L_nL_{-n}|h\rangle \overset{(2)}{=} \langle h|[L_n, L_{-n}]|h\rangle \overset{(3)}{=} 0. \tag{6.122}$$

(1) $L_{-n}^\dagger = L_n$.
(2) $L_n|h\rangle = 0$ for $n \geq 1$.
(3) $[L_n, L_{-n}] = 2nL_0 + \frac{c}{12}n(n^2-1)$ and $c = h = 0$.

In a Hilbert space a vector of zero norm is zero, so $L_{-n}|h\rangle = 0$ for every $n \geq 1$.

Step 3: the representation is trivial. The vacuum is cyclic: the representation space is spanned by the vectors $L_{-1}^{a_1}L_{-2}^{a_2}\cdots L_{-s}^{a_s}|h\rangle$. All of these with some $a_i > 0$ vanish by Step 2 (the rightmost raising operator already kills $|h\rangle$). So the space is the single line $\mathbb{C}|h\rangle$, on which every L_n acts as zero.

Key idea: nontrivial positive-energy representations are projective (Nash, p. 187)

A unitary representation of $\mathrm{vect}(S^1)$ with L_0 bounded below and $c = 0$ is trivial. Therefore every nontrivial unitary highest-weight representation has $c > 0$: it is a representation of the Virasoro algebra, and a projective representation of $\mathrm{Diff}(S^1)$. The central charge is not an optional extra; positive energy forces it.

Physical picture: what the argument says physically

In a theory with $c = 0$ and a unique vacuum, $\|L_{-2}|0\rangle\|^2 = \frac{c}{2} = 0$: the stress tensor acting on the vacuum creates nothing. But the stress tensor of any theory with local degrees of freedom creates states from the vacuum, just as the electromagnetic field operator creates photons. So $c > 0$ measures the presence of local degrees of freedom ($c = 1$ for one boson, $c = \frac{1}{2}$ for one Majorana fermion), and $c = 0$ with unitarity means there are none. Non-unitary theories with $c = 0$ do exist (percolation, polymers), and they escape the

argument only by having negative-norm states.

Reading Nash (pp. 185–187, (6.56)–(6.59))

Nash defines a highest-weight representation as a pair of *non-negative* numbers (c, h), his (6.56)–(6.57), with $L_0|h\rangle = h|h\rangle$ and $L_n|h\rangle = 0$ for $n \geq 1$ forced by boundedness below. He states, and we have proved above, that $L_{-n}|h\rangle \neq 0$ "as may be seen by applying the commutation relations" – this is our computation of $\|L_{-n}|h\rangle\|^2$ in (6.106). He then quotes the Friedan–Qiu–Shenker classification (6.58) of unitary (c, h) without derivation; Section 6.3.6 supplies the level-2 half of the argument in full and sketches the rest. Finally he shows, in the computation we reproduce as (6.59), that a true (non-projective) representation of $\mathrm{Diff}(S^1)$ with L_0 bounded below is necessarily trivial – our Section 6.3.7. Nash does not discuss Verma modules, singular vectors or the Kac form by name; Sections 6.3.2 to 6.3.5 make explicit the construction that his brief remarks presuppose.

6.3.8 Representations of the Kac–Moody algebra

We now repeat the story of the Virasoro algebra for the loop group LG. Nash's strategy is the same as before: represent the Lie algebra rather than the group, impose a boundedness condition on the energy, and see what is left. The answer has the same shape, highest-weight vector → big module → quotient by null states, but two new features appear. The lowest energy states now form a whole *multiplet* of the finite-dimensional group G, rather than a single vector. And unitarity quantises the central term k not into a discrete series of fractions but into integers, the *level*.

The full algebra

A physical theory with a loop group symmetry has two sets of operators. It has the currents T^a_m, the Fourier modes of $j^a(\theta)$ (Section 6.1.11). It also has a Hamiltonian and a momentum, which generate rotations of the circle and, in a conformal theory, the whole Virasoro algebra. In the Sugawara construction (Section 6.2.10) the second set is built from the first, and we computed all the commutators there. Collecting them, Nash's (6.60) reads

$$\begin{aligned} [L_m, L_n] &= (m-n)\,L_{m+n} + \frac{2k\dim G}{12(2k+C)}\,m(m^2-1)\,\delta_{m,-n}, \\ [T^a_m, T^b_n] &= i f^{abc}\,T^c_{m+n} + k\,m\,\delta^{ab}\,\delta_{m,-n}, \\ [L_n, T^a_m] &= -m\,T^a_{m+n}, \qquad [L_n, k] = [T^a_m, k] = 0. \end{aligned} \tag{6.123}$$

The first line is the Sugawara Virasoro algebra (6.81). The second is the Kac–Moody algebra (6.12). The third is (6.79), which says that the modes of the current transform under reparametrisations of the circle exactly as the vector fields $L_n = ie^{in\theta}\frac{d}{d\theta}$ act on the functions $e^{im\theta}$ (6.71). The central element k commutes with everything, so in an irreducible representation it is a number (Schur's lemma). Here C is the adjoint Casimir (6.74), Nash's $c(G)$.

Even without Sugawara, the third line holds whenever the L_n implement the action of $\mathrm{Diff}^+(S^1)$ on the loop group. In CFT language it says that $j^a(\theta)$ is a **primary field of conformal weight one**.

Watch out: adjoints

Throughout, the T^a are Hermitian (the physicists' convention of the pitfall at the start of the chapter). A real loop $\xi(\theta) = \sum_n \xi_{a,n} e^{in\theta} T^a$ has $\xi_{a,-n} = \overline{\xi_{a,n}}$. So in a unitary representation

$$(T^a_n)^\dagger = T^a_{-n}, \qquad L_n^\dagger = L_{-n}, \qquad k \in \mathbb{R}, \tag{6.124}$$

exactly as for the Virasoro generators (6.103). We use these rules to compute every norm below.

The boundedness condition

Setting $n = 0$ in the third line of (6.123) gives $[L_0, T^a_m] = -m\, T^a_m$. So T^a_m shifts the energy by $-m$, just as L_m does. Let $L_0|\psi\rangle = E|\psi\rangle$. Then

$$L_0\big(T^a_m|\psi\rangle\big) \overset{(1)}{=} [L_0, T^a_m]|\psi\rangle + T^a_m L_0|\psi\rangle \overset{(2)}{=} -m\, T^a_m|\psi\rangle + E\, T^a_m|\psi\rangle = (E - m)\, T^a_m|\psi\rangle. \tag{6.125}$$

(1) Write $L_0 T^a_m = [L_0, T^a_m] + T^a_m L_0$.
(2) The third line of (6.123) with $n = 0$, and $L_0|\psi\rangle = E|\psi\rangle$.

For $m > 0$ the mode T^a_m *lowers* the energy by m; for $m < 0$ it raises it; and the zero modes T^a_0 do not change the energy at all.

Now impose Nash's criterion: *the spectrum of L_0 is bounded below*. Let h be the lowest eigenvalue and $|h\rangle$ a state with that energy. The state $T^a_n|h\rangle$, $n \geq 1$, would have energy $h - n < h$, which is impossible unless it vanishes. The same holds for $L_n|h\rangle$. Hence Nash's (6.61):

$$\boxed{L_0|h\rangle = h|h\rangle \implies L_n|h\rangle = 0, \quad T^a_n|h\rangle = 0, \qquad n \geq 1.} \tag{6.126}$$

This defines a **highest-weight representation** of the Kac–Moody algebra. (As in Section 6.3.1, "highest" is the mathematicians' word; for a physicist $|h\rangle$ is the state of *lowest* energy.)

The ground states form a multiplet of G

There is a genuinely new feature here. The zero modes T^a_0 commute with L_0:

$$[L_0, T^a_0] \overset{(1)}{=} -0 \cdot T^a_0 = 0, \qquad [T^a_0, T^b_0] \overset{(2)}{=} i f^{abc}\, T^c_0.$$

(1) The third line of (6.123) with $n = m = 0$.
(2) The second line with $m = n = 0$; the central term carries a factor m and vanishes.

So the T^a_0 form an ordinary copy of the finite-dimensional Lie algebra $\mathfrak{g}$ (these are the constant loops), and they map the lowest-energy eigenspace

$$V_0 := \{|\psi\rangle : L_0|\psi\rangle = h|\psi\rangle\}$$

into itself. V_0 is therefore a representation of $\mathfrak{g}$, the **ground-state multiplet**. In a quantum field theory with a global G symmetry this is just the statement that the lowest-energy states come in G-multiplets: the proton and the neutron, the pions, the spin states of an atom. Every state of V_0 satisfies (6.126).

Key idea: the ground-state multiplet is irreducible (Nash, p. 188)

If the Kac–Moody representation is irreducible, then V_0 is an irreducible representation V_λ of $\mathfrak{g}$, labelled by a highest weight λ. The whole representation is determined by two pieces of data: the central term k and the weight λ. Nash writes the pair as (k, λ).

Why is V_0 irreducible? Suppose $V_0 = V' \oplus V''$ with V' a nonzero proper $\mathfrak{g}$-invariant subspace. Let $\mathcal{A} := U(\hat{\mathfrak{g}})$ be all polynomials in the generators. Every such polynomial can be reordered (by the commutation relations) into a sum of products

$$(\text{raising modes } T^a_{-n},\ n \geq 1) \times (\text{zero modes } T^a_0) \times (\text{lowering modes } T^a_n,\ n \geq 1),$$

the Poincaré–Birkhoff–Witt ordering. Acting on V', the lowering modes give zero by (6.126), the zero modes keep us in V', and the raising modes strictly increase the energy. So the energy-h part of $\mathcal{A}V'$ is V' itself, and $\mathcal{A}V'$ is an invariant subspace not containing V'': the representation would be reducible. Hence V_0 is irreducible.

The energy of the ground states is also fixed. With the Sugawara L_0, the Casimir of the finite group sets the energy:

Worked example 6.44: the ground-state energy is a Casimir

By (6.75) and the normal ordering (6.76),

$$L_0 = \frac{1}{2k + C}\Big(T^a_0 T^a_0 + 2\sum_{m>0} T^a_{-m} T^a_m\Big).$$

(The $m = 0$ term of $\sum_m :T^a_m T^a_{-m}:$ is $T^a_0 T^a_0$; the terms m and $-m$ both become $T^a_{-|m|} T^a_{|m|}$.) On $|\psi\rangle \in V_\lambda$,

$$L_0|\psi\rangle \overset{(1)}{=} \frac{1}{2k + C}\, T^a_0 T^a_0 |\psi\rangle \overset{(2)}{=} \frac{C_\lambda}{2k + C}\,|\psi\rangle, \qquad \text{so} \qquad h_\lambda = \frac{C_\lambda}{2k + C}.$$

(1) $T^a_m|\psi\rangle = 0$ for $m > 0$ by (6.126).

(2) $T^a_0 T^a_0$ is the quadratic Casimir of $\mathfrak{g}$; on the irreducible V_λ it is a number C_λ (Schur's lemma).

For SU(2), $T^a_0 T^a_0 = \vec{J}^2 = j(j+1)$ on spin j, $C = 2$ and $2k = N$, so

$$h_j = \frac{j(j+1)}{N+2}. \tag{6.127}$$

At level $N = 1$ and $j = \frac{1}{2}$ this is $h = \frac{3/4}{3} = \frac{1}{4}$, the dimension of the vertex operator $e^{i\phi/\sqrt{2}}$ of a free boson at the self-dual radius (Example 6.31 said that $\mathrm{SU}(2)_1$ is such a boson). At $N = 2$ the values are $h = 0, \frac{3}{16}, \frac{1}{2}$: three Majorana fermions ($c = \frac{3}{2}$), with $\frac{1}{2}$ the fermion itself and $\frac{3}{16} = 3 \times \frac{1}{16}$ the product of three Ising spin fields.

The Cartan–Weyl basis

To go further we need the root decomposition of $\mathfrak{g}$ (a refresher; the Lie theory is in Chapter T3).

Working definition 6.23 (roots and step operators)

Pick a maximal set of commuting Hermitian generators $H^1, \ldots, H^r$ (the **Cartan subalgebra** $\mathfrak{t}$; r is the rank). The rest of $\mathfrak{g}_{\mathbb{C}}$ [complex linear combinations of the generators] splits into

simultaneous eigenvectors E^α of the H^i,

$$[H^i, E^\alpha] = \alpha^i E^\alpha, \qquad (E^\alpha)^\dagger = E^{-\alpha}, \qquad [E^\alpha, E^{-\alpha}] = \alpha \cdot H,$$

labelled by **roots** $\alpha = (\alpha^1, \dots, \alpha^r) \in \mathbb{R}^r$. We normalise with the same inner product that appears in the central term, $(T^a, T^b) = \delta^{ab}$, so that $(H^i, H^j) = \delta^{ij}$ and $(E^\alpha, E^{-\beta}) = \delta_{\alpha\beta}$. A choice of "positive" direction [a vector in $\mathbb{R}^r$ perpendicular to no root. A root is positive when its dot product with this vector is positive] splits the roots into positive (E^α raise the weight) and negative ($E^{-\alpha}$ lower it). In a representation, a **weight** μ is a simultaneous eigenvalue $H^i|\mu\rangle = \mu^i|\mu\rangle$, and the **highest weight** λ is the one killed by all positive E^α. *How to compute:* diagonalise ad H^i [the map $X \mapsto [H^i, X]$] on $\mathfrak{g}_{\mathbb{C}}$.

For SU(2) with $T^a = \sigma^a/2$ the rank is 1, $H = T^3$, and $E^\pm = \frac{1}{\sqrt{2}}(T^1 \pm iT^2)$, i.e. $E^\pm = J^\pm/\sqrt{2}$ in terms of the angular momentum ladder operators $J^\pm = T^1 \pm iT^2$. Then

$$[E^+, E^-] \overset{(1)}{=} \tfrac{1}{2}\big(-i[T^1, T^2] + i[T^2, T^1]\big) \overset{(2)}{=} \tfrac{1}{2}\big(-i \cdot iT^3 + i \cdot (-i)T^3\big) = T^3,$$

(1) Expand $[T^1 + iT^2, T^1 - iT^2]$; the $[T^1, T^1]$ and $[T^2, T^2]$ terms vanish.
(2) $[T^1, T^2] = iT^3$.

so the roots are $\alpha = \pm 1$ and $|\alpha|^2 = 1$. This is the statement "long roots have length 1" in our normalisation, which we used in Section 6.2.6. On spin j the weights are $m = -j, \dots, j$ and $\lambda = j$.

The one fact we need about the finite-dimensional theory is this. For every root α the three operators

$$e_\alpha = \frac{\sqrt{2}}{|\alpha|} E^\alpha, \qquad f_\alpha = \frac{\sqrt{2}}{|\alpha|} E^{-\alpha}, \qquad h_\alpha = \frac{2}{|\alpha|^2}\, \alpha \cdot H \tag{6.128}$$

satisfy the $\mathfrak{sl}(2)$ relations $[h, e] = 2e$, $[h, f] = -2f$, $[e, f] = h$ (check: $[h_\alpha, e_\alpha] = \frac{2}{|\alpha|^2}\, \alpha \cdot \alpha\, e_\alpha = 2e_\alpha$ and $[e_\alpha, f_\alpha] = \frac{2}{|\alpha|^2}[E^\alpha, E^{-\alpha}] = h_\alpha$). For SU(2) they are J^+, J^-, $2J^3$. Everything about levels follows from the representation theory of this $\mathfrak{sl}(2)$.

The $\mathfrak{sl}(2)$ fact we need

Let e, f, h act unitarily ($e^\dagger = f$, $h^\dagger = h$), and let v be a vector with $e\,v = 0$ and $h\,v = n\,v$. First, by induction on j,

$$\begin{aligned} e\, f^j v &\overset{(1)}{=} [e, f]\, f^{j-1} v + f\, e\, f^{j-1} v \overset{(2)}{=} h\, f^{j-1} v + f\,(j-1)(n-j+2)\, f^{j-2} v \\ &\overset{(3)}{=} \big(n - 2(j-1)\big) f^{j-1} v + (j-1)(n-j+2)\, f^{j-1} v \overset{(4)}{=} j\,(n-j+1)\, f^{j-1} v. \end{aligned} \tag{6.129}$$

(1) $ef = [e, f] + fe$.
(2) $[e, f] = h$; the induction hypothesis for $j-1$ (for $j = 1$ the second term is absent because $ev = 0$).
(3) f lowers the h-eigenvalue by 2 ($[h, f] = -2f$), so $hf^{j-1}v = (n - 2j + 2)f^{j-1}v$.
(4) $(n - 2j + 2) + (j-1)(n-j+2) = j(n-j+1)$, as expanding both sides shows.

Hence the norms of the states in the "multiplet" built on v are

$$\big\|f^j v\big\|^2 \overset{(1)}{=} \langle v|\, e^j f^j |v\rangle \overset{(2)}{=} j(n-j+1)\, \big\|f^{j-1} v\big\|^2 \overset{(3)}{=} j!\; n(n-1)\cdots(n-j+1)\; \|v\|^2 . \tag{6.130}$$

(1) $(f^j)^\dagger = e^j$.
(2) Peel off one e with (6.129).
(3) Repeat j times.

Key idea: unitarity quantises n

If every $\left\|f^j v\right\|^2$ is ≥ 0, then n is a non-negative integer, and $f^{n+1}v = 0$.

Why: if $n < 0$ then $\|fv\|^2 = n\,\|v\|^2 < 0$. If $n > 0$ is not an integer, let $j = \lfloor n \rfloor + 2$. The factors $n, n-1, \ldots, n-j+2$ are positive but the last one, $n - j + 1 = n - \lfloor n \rfloor - 1$, is negative, so $\left\|f^j v\right\|^2 < 0$. If n is a non-negative integer, the factor $n - j + 1$ vanishes at $j = n + 1$, so $\left\|f^{n+1}v\right\| = 0$ and, in a Hilbert space, $f^{n+1}v = 0$. This is the familiar statement that $2j$ is an integer and the multiplet stops at $m = -j$.

No representations at $k = 0$

Nash says we "can easily verify that there are no non-trivial highest weight representations of the algebra $\mathrm{Map}(S^1, \mathfrak{g})$ itself". This is the Kac–Moody analogue of Goddard–Olive (Section 6.3.7), and it is a two-line computation. In the loop algebra with the central term, the step-operator modes $E^\alpha_m = e^{im\theta}E^\alpha$ obey

$$[E^\alpha_m, E^{-\alpha}_n] \overset{(1)}{=} \left([E^\alpha, E^{-\alpha}]\right)_{m+n} + k\,m\,(E^\alpha, E^{-\alpha})\,\delta_{m,-n} \overset{(2)}{=} \alpha \cdot H_{m+n} + k\,m\,\delta_{m,-n}. \tag{6.131}$$

(1) The second line of (6.123) is bilinear in the generators: for any $X, Y \in \mathfrak{g}_{\mathbb{C}}$, $[X_m, Y_n] = [X, Y]_{m+n} + km(X, Y)\delta_{m,-n}$.
(2) The Cartan–Weyl relations and the normalisation $(E^\alpha, E^{-\alpha}) = 1$.

Now let $|\lambda\rangle$ be the highest-weight state of the ground multiplet, $H^i_0|\lambda\rangle = \lambda^i|\lambda\rangle$, and let α be a positive root. Two norms:

$$\left\|E^{-\alpha}_0|\lambda\rangle\right\|^2 \overset{(1)}{=} \langle\lambda|[E^\alpha_0, E^{-\alpha}_0]|\lambda\rangle \overset{(2)}{=} \alpha \cdot \lambda, \tag{6.132}$$

$$\left\|E^{\alpha}_{-1}|\lambda\rangle\right\|^2 \overset{(3)}{=} \langle\lambda|[E^{-\alpha}_1, E^{\alpha}_{-1}]|\lambda\rangle \overset{(4)}{=} k - \alpha \cdot \lambda. \tag{6.133}$$

(1) $(E^{-\alpha}_0)^\dagger = E^\alpha_0$, and $E^\alpha_0|\lambda\rangle = 0$ because λ is the highest weight.
(2) (6.131) with $m = n = 0$.
(3) $(E^\alpha_{-1})^\dagger = E^{-\alpha}_1$, and $E^{-\alpha}_1|\lambda\rangle = 0$ by (6.126).
(4) (6.131) with $-\alpha$ in place of α and $m = 1$, $n = -1$: $[E^{-\alpha}_1, E^\alpha_{-1}] = -\alpha \cdot H_0 + k$.

Unitarity needs both to be non-negative:

$$0 \leq \alpha \cdot \lambda \leq k \qquad \text{for every positive root } \alpha. \tag{6.134}$$

Set $k = 0$. Then $\alpha \cdot \lambda = 0$ for all positive roots, so $\lambda = 0$ (the roots span the dual of $\mathfrak{t}$): the ground multiplet is the trivial one-dimensional representation, $V_0 = \mathbb{C}|0\rangle$ with $T^a_0|0\rangle = 0$. Next,

$$\|T^a_{-n}|0\rangle\|^2 \overset{(1)}{=} \langle 0|[T^a_n, T^a_{-n}]|0\rangle \overset{(2)}{=} \langle 0|\left(if^{aac}T^c_0 + kn\right)|0\rangle \overset{(3)}{=} kn = 0 \qquad \text{(no sum over } a\text{)},$$

(1) $(T^a_{-n})^\dagger = T^a_n$ and $T^a_n|0\rangle = 0$ for $n \geq 1$.
(2) The Kac–Moody relation with $b = a$.
(3) $f^{aac} = 0$ by antisymmetry, and $k = 0$.

So every $T^a_{-n}|0\rangle$ vanishes, and since $|0\rangle$ is cyclic the whole representation is the one-dimensional trivial one. Every nontrivial positive-energy unitary representation of the loop algebra is therefore a representation of the *central extension* with $k > 0$ ($k < 0$ is ruled out by $\big\|T^a_{-1}|0\rangle\big\|^2 = k$, or by (6.134)). This is why Nash says "we only consider the centrally extended algebra". As for $\mathrm{Diff}(S^1)$, the central extension is forced by positive energy.

Physical picture: why positive energy needs the Schwinger term

The inequality (6.134) says that the charge $\alpha \cdot \lambda$ carried by the ground state is bounded by the central term. With $k = 0$ no state can carry charge, and the current operators, which should create particle–hole pairs from the Dirac sea, create nothing. It is the Schwinger term that gives the current a nonzero vacuum two-point function (6.73), and hence the ability to create states.

The module: pipeline for the Kac–Moody algebra

We now build the representation space, following the same four steps as for the Virasoro algebra (Sections 6.3.2 to 6.3.4).

$$\underbrace{V_\lambda}_{\text{ground multiplet}} \xrightarrow{\text{apply } T^a_{-n},\ n\geq 1} \underbrace{M(k,\lambda) = U(\hat{\mathfrak{g}}_-)\, V_\lambda}_{\text{generalised Verma module}} \xrightarrow{\text{quotient by the null submodule } N} \underbrace{L(k,\lambda) = M(k,\lambda)/N}_{\text{irreducible}} \tag{6.135}$$

Step 1: the ground multiplet. Start from the irreducible representation V_λ of $\mathfrak{g}$, extended to the non-negative modes by $T^a_n V_\lambda = 0$ for $n \geq 1$ and $k =$ a number. This is a module for the subalgebra $\hat{\mathfrak{g}}_{\geq 0}$ spanned by k and T^a_n, $n \geq 0$.

Step 2: generate descendants. Let $\hat{\mathfrak{g}}_-$ be spanned by the raising modes T^a_{-n}, $n \geq 1$. The **generalised Verma module** $M(k,\lambda)$ is spanned by the ordered monomials

$$T^{a_1}_{-n_1} T^{a_2}_{-n_2} \cdots T^{a_s}_{-n_s} |\lambda, \mu\rangle, \qquad n_1 \geq n_2 \geq \cdots \geq n_s \geq 1, \tag{6.136}$$

(with $a_1 \leq a_2 \leq \ldots$ among equal n's), where $|\lambda,\mu\rangle$ runs over a basis of V_λ. These vectors are declared linearly independent; the algebra acts by commuting generators to the right until they hit V_λ. This is Nash's "we can construct the representation space by application of polynomials in the generators to the vacuum". The word "generalised" means that we start from a whole irreducible multiplet V_λ, rather than from one vector: the null vectors of the finite-dimensional algebra (such as $(J^-)^{2j+1}|j,j\rangle = 0$) have already been removed.

A monomial (6.136) has energy $h_\lambda + \sum n_i$; the integer $\sum n_i$ is its **grade**. Counting monomials grade by grade is counting partitions in which each part comes in $\dim G$ colours:

$$\sum_{\text{grades } N} \dim M(k,\lambda)_N \; q^N = \dim V_\lambda \prod_{n\geq 1} \frac{1}{(1-q^n)^{\dim G}}. \tag{6.137}$$

For SU(2) and $\lambda = 0$ this is $1 + 3q + 9q^2 + 22q^3 + 51q^4 + \cdots$. At grade 2, for instance, there are 3 states $T^a_{-2}|0\rangle$ and 6 symmetric products $T^a_{-1}T^b_{-1}|0\rangle$ (the antisymmetric products are not new: $[T^a_{-1}, T^b_{-1}] = if^{abc}T^c_{-2}$).

Step 3: find the null submodule. The inner product on $M(k,\lambda)$ is fixed by (6.124), $\||\lambda\rangle\| = 1$ and the commutation relations; it is the Kac–Moody version of the Kac form of Section 6.3.5.

A **singular vector** is a state $|\chi\rangle$ of positive grade killed by all lowering operators. As for the Virasoro algebra, $|\chi\rangle$ is orthogonal to every state and has zero norm, and so is everything it generates. The whole generated submodule N is the null submodule. For SU(2) the singular vector, when it exists, is

$$|\chi\rangle = (J_{-1}^{+})^{N-2j+1}|j,j\rangle, \qquad J_n^{\pm} := T_n^1 \pm iT_n^2, \tag{6.138}$$

as we show in Section 6.3.9.

Step 4: quotient. The irreducible representation is $L(k,\lambda) = M(k,\lambda)/N$. The action on classes, $T_n^a[v] = [T_n^a v]$, is well defined for the same reason as in (6.115): if $v - v' \in N$ then $T_n^a v - T_n^a v' = T_n^a(v-v') \in N$, because N is a submodule. The unitary representations are exactly the quotients for which the induced inner product on $L(k,\lambda)$ is positive definite.

	$M(k,\lambda)$	null submodule N	$L(k,\lambda) = M/N$
built from	V_λ by all T_{-n}^a	$\|\chi\rangle$ by all T_{-n}^a	classes $v+N$
grade-N count	(6.137)	starts at the grade of $\|\chi\rangle$	difference
inner product	degenerate if $N \neq 0$	zero norm, $\perp$ everything	nondegenerate
when $N=0$	generic level k	—	$L = M$
unitary?	never for $k>0$ (degenerate or indefinite)	—	iff $N = 2k \in \mathbb{Z}$, $2j \le N$

For generic values of k (in particular non-rational ones) there are no singular vectors, and $M(k,\lambda) = L(k,\lambda)$ (Kac and Kazhdan). But then, as we now show, the inner product is not positive: the unitary representations live exactly at the special values of k where null states appear and are removed.

6.3.9 The level is an integer

Nash states the result without derivation: "just as in the $c<1$ Virasoro case, the central term k can only take certain discrete values":

$$\boxed{N \in \mathbb{Z}, \qquad N = \frac{2k}{l^2}} \tag{6.139}$$

where l is the length of a long root of $\mathfrak{g}$. The integer N is the **level** of the representation. The mechanism is beautifully simple: inside the Kac–Moody algebra there is, for every root, a *second* copy of $\mathfrak{su}(2)$, and its "spin" involves k. Unitarity forces the spin of every $\mathfrak{su}(2)$ to be a half-integer.

The hidden $\mathfrak{su}(2)$

For a root α define, by analogy with (6.128) but using a raising mode of *one* root and a lowering mode of the *opposite* root,

$$\hat{e}_\alpha = \frac{\sqrt{2}}{|\alpha|}E_1^{-\alpha}, \qquad \hat{f}_\alpha = \frac{\sqrt{2}}{|\alpha|}E_{-1}^{\alpha}, \qquad \hat{h}_\alpha = \frac{2k}{|\alpha|^2} - \frac{2}{|\alpha|^2}\,\alpha\cdot H_0. \tag{6.140}$$

Let us verify the $\mathfrak{sl}(2)$ relations:

$$[\hat{e}_\alpha,\hat{f}_\alpha] \overset{(1)}{=} \frac{2}{|\alpha|^2}[E_1^{-\alpha},E_{-1}^{\alpha}] \overset{(2)}{=} \frac{2}{|\alpha|^2}(-\alpha\cdot H_0 + k) = \hat{h}_\alpha,$$

$$[\hat{h}_\alpha,\hat{e}_\alpha] \overset{(3)}{=} -\frac{2}{|\alpha|^2}\cdot\frac{\sqrt{2}}{|\alpha|}[\alpha\cdot H_0, E_1^{-\alpha}] \overset{(4)}{=} -\frac{2}{|\alpha|^2}\cdot\frac{\sqrt{2}}{|\alpha|}(-\alpha\cdot\alpha)\,E_1^{-\alpha} = 2\,\hat{e}_\alpha,$$

(1) Definition (6.140).
(2) (6.131) with $-\alpha$ in place of α, $m = 1$, $n = -1$.
(3) k is central, so only the H_0 term contributes.
(4) $[H_0^i, E_1^{-\alpha}] = -\alpha^i E_1^{-\alpha}$: the zero modes of the Cartan generators act on any mode of $E^{-\alpha}$ as on $E^{-\alpha}$.

and similarly $[\hat{h}_\alpha, \hat{f}_\alpha] = -2\hat{f}_\alpha$. Also $\hat{e}_\alpha^\dagger = \hat{f}_\alpha$ by (6.124). So $(\hat{e}_\alpha, \hat{f}_\alpha, \hat{h}_\alpha)$ is a unitary $\mathfrak{su}(2)$, different from the finite one (6.128): its $\hat{h}$ contains the central term. In the geometric picture of Figure 6.25, $\hat{e}_\alpha$ moves one step in the loop direction and one step down in weight at the same time.

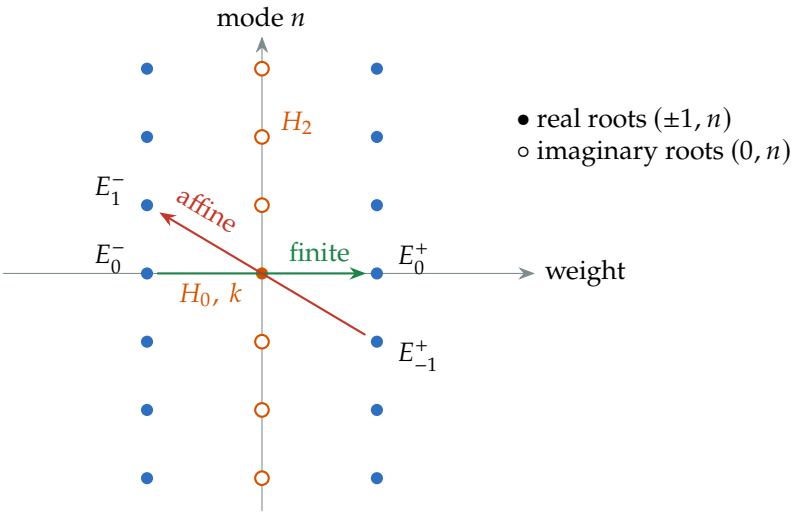


Figure 6.25: The roots of the affine algebra $\widehat{\mathfrak{su}}(2)$. Each generator $E_n^\pm$, H_n is drawn at (its T^3-weight, its mode number). The finite $\mathfrak{su}(2)$ (green) is the horizontal pair $E_0^\pm$. The hidden $\mathfrak{su}(2)$ (red) pairs E_1^- with E_{-1}^+: it is another line through the origin, and its Cartan element $\hat{h} = 2k - 2H_0$ picks up the central term. Every pair $(\pm 1, n)$, $(\mp 1, -n)$ gives an $\mathfrak{su}(2)$ of this kind; the two drawn generate the rest. The points $(0, n)$ are the modes of the Cartan generator (the "imaginary roots").

The quantisation

Apply the $\mathfrak{sl}(2)$ fact to the highest-weight state $v = |\lambda\rangle$ of the ground multiplet and a positive root α. It is killed by $\hat{e}_\alpha \propto E_1^{-\alpha}$ (a lowering mode, (6.126)), and it is an eigenvector of $\hat{h}_\alpha$:

$$\hat{h}_\alpha|\lambda\rangle \overset{(1)}{=} \left(\frac{2k}{|\alpha|^2} - \frac{2\,\alpha\cdot\lambda}{|\alpha|^2}\right)|\lambda\rangle \overset{(2)}{=} \left(\frac{2k}{|\alpha|^2} - \langle\lambda, \alpha^\vee\rangle\right)|\lambda\rangle. \tag{6.141}$$

(1) $H_0^i|\lambda\rangle = \lambda^i|\lambda\rangle$.
(2) Notation: $\langle\lambda, \alpha^\vee\rangle := \frac{2\,\alpha\cdot\lambda}{|\alpha|^2}$ is the eigenvalue of the finite h_α (6.128) on $|\lambda\rangle$.

By the keyidea on unitarity, this eigenvalue is a non-negative integer. The finite $\mathfrak{sl}(2)$ already forces $\langle\lambda, \alpha^\vee\rangle$ to be a non-negative integer (it is "$2j$" of the finite multiplet). Therefore

$$\frac{2k}{|\alpha|^2} \in \mathbb{Z} \qquad \text{and} \qquad \langle\lambda, \alpha^\vee\rangle \le \frac{2k}{|\alpha|^2} \qquad \text{for every positive root } \alpha. \tag{6.142}$$

For a long root $|\alpha| = l$ and the first condition is Nash's (6.62), $N = 2k/l^2 \in \mathbb{Z}$. For a short root $|\alpha|^2 = l^2/p$ with $p = 2$ or 3 (in the non-simply-laced algebras), and $2k/|\alpha|^2 = pN$ is then automatically an integer: the long roots give the strongest condition. The second condition,

applied to the *highest root* θ (which is long), is the strongest of the inequalities:

$$\boxed{\langle \lambda, \theta^\vee \rangle \le N} \tag{6.143}$$

Representations satisfying (6.139) and (6.143) are called **integrable**. They are exactly the unitary highest-weight representations, and exactly the ones that exponentiate to (projective) representations of the loop group (Pressley and Segal). There are finitely many at each level.

Key idea: the level (Nash's (6.62))

In a unitary highest-weight representation of the affine Kac–Moody algebra, $N = 2k/l^2$ is a non-negative integer, and the highest weight of the ground multiplet satisfies $\langle \lambda, \theta^\vee \rangle \le N$. For SU(2) in our normalisation ($l^2 = 1$): $N = 2k \in \mathbb{Z}_{\ge 0}$ and the ground-state spin satisfies

$$j \in \{0, \tfrac{1}{2}, 1, \dots, \tfrac{N}{2}\}.$$

When $\langle \lambda, \theta^\vee \rangle \le N$ the null vector $\hat{f}_\theta^{\,N - \langle \lambda, \theta^\vee \rangle + 1} |\lambda\rangle$ must be quotiented out, which for SU(2) is (6.138).

This is the same integer we met topologically in Section 6.2.6: there the central extension of the *group* LG existed only for $N \in \mathbb{Z}$ (the cocycle had to be an integral class, like the flux of a monopole). Here unitarity of the representation of the *algebra* gives the same condition. The two must agree, because a unitary highest-weight representation of the algebra integrates to a projective representation of the group, i.e. a true representation of the extension.

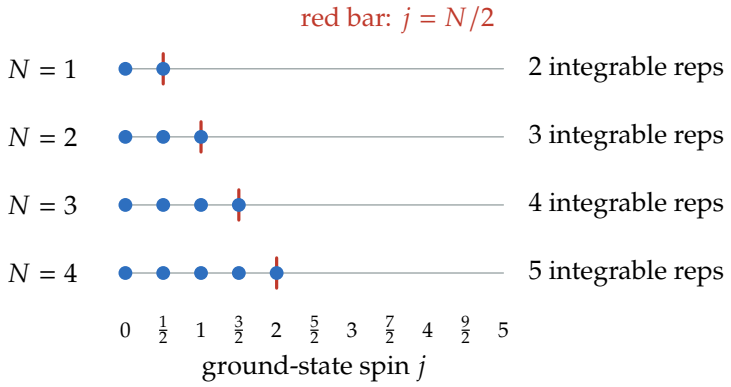


Figure 6.26: The integrable representations of $\widehat{\mathfrak{su}}(2)$ at levels $N = 1, \dots, 4$. Each dot is an allowed spin j of the ground multiplet; the red bar is the bound $2j \le N$ of (6.143). At level N there are $N + 1$ representations, with lowest energies $h_j = j(j+1)/(N+2)$ (6.127). As $N \to \infty$ the bound disappears and every spin of SU(2) is allowed: the classical limit.

Watch out: normalisations of the level

Many CFT books write the Kac–Moody algebra as $[J^a_m, J^b_n] = i f^{abc} J^c_{m+n} + \kappa\, m\, \delta^{ab} \delta_{m+n,0}$ with the generators normalised so that long roots have length^2 2; then κ itself is the integer level. Nash's form $N = 2k/l^2$ is normalisation-independent: rescaling $T^a \to sT^a$ multiplies k by s^2 and l^2 by s^2. In our normalisation ($T^a = \sigma^a/2$ for SU(2), $l^2 = 1$) the integer is $N = 2k$, which is why $c = \frac{3N}{N+2}$ in Example 6.31.

The SU(2) case in detail

For SU(2) the hidden $\mathfrak{su}(2)$ is built from $J_n^{\pm} = T_n^1 \pm iT_n^2$ ($= \sqrt{2}E_n^{\pm}$). First the basic commutator:

$$[J_m^+, J_n^-] \overset{(1)}{=} [T_m^1, T_n^1] + [T_m^2, T_n^2] - i[T_m^1, T_n^2] + i[T_m^2, T_n^1]$$
$$\overset{(2)}{=} 2km\,\delta_{m,-n} - i\big(iT_{m+n}^3\big) + i\big(-iT_{m+n}^3\big) = 2T_{m+n}^3 + 2km\,\delta_{m,-n}. \tag{6.144}$$

(1) Expand $[T_m^1 + iT_m^2,\ T_n^1 - iT_n^2]$ by bilinearity.
(2) The Kac–Moody relations with $f^{abc} = \varepsilon^{abc}$: $[T_m^1, T_n^1] = km\delta_{m,-n}$ (the same for T^2), $[T_m^1, T_n^2] = iT_{m+n}^3$ and $[T_m^2, T_n^1] = -iT_{m+n}^3$ (their central terms vanish, since $\delta^{12} = 0$).

The hidden $\mathfrak{su}(2)$ is $\hat{e} = J_1^-$, $\hat{f} = J_{-1}^+$ and

$$\hat{h} \overset{(1)}{=} [J_1^-, J_{-1}^+] \overset{(2)}{=} -\big(2T_0^3 + 2k\cdot(-1)\big) = 2k - 2T_0^3,$$

(1) The $\mathfrak{sl}(2)$ relation $[\hat{e}, \hat{f}] = \hat{h}$.
(2) $[J_1^-, J_{-1}^+] = -[J_{-1}^+, J_1^-]$ and (6.144) with $m = -1$, $n = 1$.

in agreement with (6.140) for $|\alpha|^2 = 1$. On the top state $|j, j\rangle$ of the ground multiplet, $\hat{h} = 2k - 2j = N - 2j$. So the condition "$\hat{h}$ has non-negative integer eigenvalue" reads $N - 2j \in \mathbb{Z}_{\geq 0}$, and the singular vector is $\hat{f}^{\,N-2j+1}|j, j\rangle = (J_{-1}^+)^{N-2j+1}|j, j\rangle$, which is (6.138). It is killed by all the lowering operators: by $\hat{e} = J_1^-$ because of (6.129) with $j = n + 1$, and by J_0^+ because $[J_0^+, J_{-1}^+] = 0$ and $J_0^+|j, j\rangle = 0$; these two generate all the lowering operators under commutators (for example $[J_0^+, J_1^-] = 2T_1^3$).

Worked example 6.45: the vacuum module of SU(2) at level 1

Take $j = 0$. We compute the lowest grades of $M(k, 0)$ and see where unitarity fails and where null states appear.

Grade 1. The three states $T_{-1}^a|0\rangle$ have norms $\big\|T_{-1}^a|0\rangle\big\|^2 = \langle 0|[T_1^a, T_{-1}^a]|0\rangle = k$ (no sum), positive for $k > 0$.

Grade 2. Consider the top-weight state $(J_{-1}^+)^2|0\rangle$, with $T_0^3 = 2$. Using $\hat{e} = J_1^-$,

$$J_1^-\,(J_{-1}^+)^2|0\rangle \overset{(1)}{=} [J_1^-, J_{-1}^+]\,J_{-1}^+|0\rangle + J_{-1}^+\,[J_1^-, J_{-1}^+]|0\rangle$$
$$\overset{(2)}{=} (2k - 2T_0^3)\,J_{-1}^+|0\rangle + J_{-1}^+\,(2k - 2T_0^3)|0\rangle$$
$$\overset{(3)}{=} (2k-2)\,J_{-1}^+|0\rangle + 2k\,J_{-1}^+|0\rangle = (4k-2)\,J_{-1}^+|0\rangle,$$

(1) $J_1^-|0\rangle = 0$, so only commutators survive.
(2) $[J_1^-, J_{-1}^+] = \hat{h} = 2k - 2T_0^3$.
(3) $J_{-1}^+|0\rangle$ has $T_0^3 = 1$ (because $[T_0^3, J_{-1}^+] = J_{-1}^+$), and $T_0^3|0\rangle = 0$.

so, with $\big\|J_{-1}^+|0\rangle\big\|^2 = \langle 0|[J_1^-, J_{-1}^+]|0\rangle = 2k$,

$$\big\|(J_{-1}^+)^2|0\rangle\big\|^2 \overset{(1)}{=} \langle 0|J_1^-\,J_1^-(J_{-1}^+)^2|0\rangle \overset{(2)}{=} (4k-2)\,\big\|J_{-1}^+|0\rangle\big\|^2 \overset{(3)}{=} 4k\,(2k-1).$$

(1) $\big((J_{-1}^+)^2\big)^\dagger = (J_1^-)^2$.
(2) The previous chain.
(3) $\big\|J_{-1}^+|0\rangle\big\|^2 = 2k$.

This agrees with (6.130): $j = 2$, $n = 2k$, giving $2\cdot 2k(2k-1)$. Read the answer as a function of the level:

- $0 < 2k < 1$: the norm is *negative*. There is no unitary representation. This is level quantisation happening at grade two.
- $2k = 1$: the norm is zero. The state $(J^+_{-1})^2|0\rangle$ is the singular vector (6.138) with $N = 1$, $j = 0$.
- $2k > 1$ but not an integer: positive here, but a higher power $(J^+_{-1})^{\lfloor 2k \rfloor + 2}|0\rangle$ has negative norm.

Now set $N = 1$ and quotient. The null vector has $T^3_0 = 2$, and it is the top state of a spin-2 multiplet under the zero modes T^a_0 (it is killed by J^+_0); all five states $(J^-_0)^r(J^+_{-1})^2|0\rangle$, $r = 0, \dots, 4$, lie in N, since N is a submodule. Under SU(2) the six symmetric products $T^a_{-1}T^b_{-1}|0\rangle$ decompose as $\mathrm{Sym}^2(\mathbf{3}) = \mathbf{5} \oplus \mathbf{1}$, and the $\mathbf{5}$ is exactly this null multiplet. What survives at grade 2 is

$$\underbrace{T^a_{-2}|0\rangle}_{\mathbf{3}} \;\oplus\; \underbrace{T^a_{-1}T^a_{-1}|0\rangle = (2k + C)\,L_{-2}|0\rangle}_{\mathbf{1}\text{ (Sugawara stress tensor)}},$$

four states instead of nine.

The counting can be checked against the free boson. By Example 6.31, $L(1/2, 0)$ is the Fock space of a chiral boson at the self-dual radius, with T^3_0 the momentum $n \in \mathbb{Z}$ and grade n^2 + (oscillator level). So the number of states with $T^3_0 = n$ at grade g is $p(g - n^2)$, with p the partition function:

$$\sum_g \dim L(\tfrac{1}{2}, 0)_g \, q^g = \frac{\sum_{n\in\mathbb{Z}} q^{n^2}}{\prod_{m\geq 1}(1 - q^m)} = 1 + 3q + 4q^2 + 7q^3 + 13q^4 + \cdots . \tag{6.145}$$

Compare the generalised Verma module $1 + 3q + 9q^2 + 22q^3 + 51q^4 + \cdots$: the quotient throws away most of the states. Figure 6.27 shows where the survivors sit.

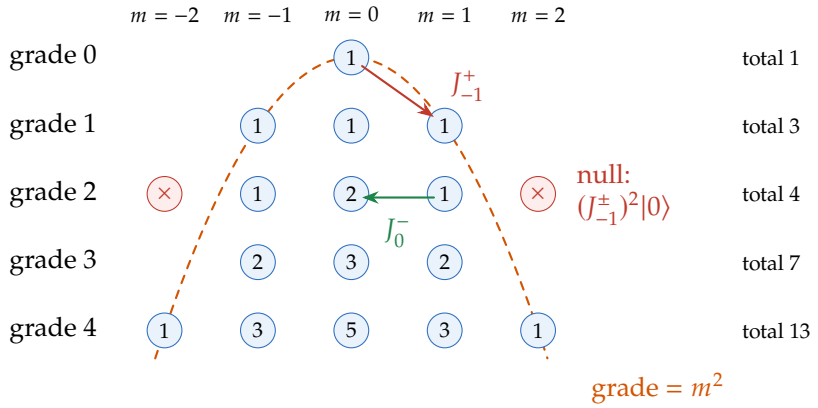


Figure 6.27: The irreducible vacuum module $L(\frac{1}{2}, 0)$ of $\widehat{\mathfrak{su}}(2)$ at level $N = 1$. Horizontal: the T^3_0 weight m; vertical (downwards): the grade. The number in each circle is the multiplicity $p(g - m^2)$. No state lies above the dashed parabola $g = m^2$: to reach weight m one must spend grade at least m^2. The red crosses are the null states $(J^\pm_{-1})^2|0\rangle$, which would have $|m| = 2$ at grade 2 in the generalised Verma module; they, and the whole multiplet and submodule they generate, are removed. The red arrow is the hidden $\mathfrak{su}(2)$ lowering operator $\hat{f} = J^+_{-1}$, the green one a finite $\mathfrak{su}(2)$ step. The total numbers are (6.145).

Physical picture: the parabola

Why must the states lie inside a parabola? The single current $T^3(\theta)$ is itself a U(1) Kac–Moody algebra with central term $k = N/2$, and it has its own Sugawara operator $L_0^{\mathrm{U}(1)} = \frac{1}{2k}(T_0^3)^2 + (\text{oscillators})$, whose lowest value in the sector of charge m is m^2/N. The difference $L_0 - L_0^{\mathrm{U}(1)}$ is a coset Virasoro generator (6.121), and on a unitary representation it is non-negative. So every state of charge m has energy at least m^2/N; for $N = 1$ and $h = 0$ that is grade $\geq m^2$. Physically: to put charge m on a circle you pay an electrostatic energy $m^2/2k$, quadratic in the charge, with the level playing the role of the capacitance.

Try it: level 1, spin one-half

In $L(\frac{1}{2}, \frac{1}{2})$, show that $J_{-1}^+|\frac{1}{2}, \frac{1}{2}\rangle$ is null. *Answer:* $\left\|J_{-1}^+|\frac{1}{2}, \frac{1}{2}\rangle\right\|^2 = \langle\frac{1}{2}, \frac{1}{2}|(2k - 2T_0^3)|\frac{1}{2}, \frac{1}{2}\rangle = 2k - 1 = 0$ at $N = 1$. This is (6.138) with $N - 2j + 1 = 1$. The lowest states of charge $\frac{3}{2}$ therefore sit at grade 2, not 1.

Reading Nash (pp. 187–188, (6.60)–(6.62))

Nash lists the three families of commutators (6.123), states the boundedness criterion (6.126), and asserts that $k = 0$ admits no nontrivial highest-weight representations (we derived it from the norms (6.132)–(6.133)). He notes that the T_0^a act irreducibly on the ground states, so a representation is labelled by (k, λ), and quotes the quantisation $N = 2k/l^2 \in \mathbb{Z}$ with l the length of a long root, referring to Kac and to Pressley–Segal for proofs. The proof is the hidden $\mathfrak{su}(2)$ of (6.140). Nash writes $\delta_{m,-n}$ where we sometimes write $\delta_{m+n,0}$; they are the same.

6.3.10 The Borel–Weil construction

So far we have built representations *algebraically*: start with a highest-weight vector, act with raising operators, throw away the null states. Nash ends the chapter with a second, *geometric* way of producing the same representations. It emphasises the group rather than the algebra, and it looks exactly like quantum mechanics: the representation space is a space of wavefunctions on a manifold.

Here is the question it answers. A physicist already knows one way to manufacture representations of SO(3) out of geometry: take functions on the sphere $S^2 = \mathrm{SO}(3)/\mathrm{SO}(2)$ and let rotations act on them by $(R \cdot \psi)(\vec{x}) = \psi(R^{-1}\vec{x})$. The result is

$$L^2(S^2) = \bigoplus_{l=0}^{\infty} \operatorname{span}\{Y_{lm}\}_{m=-l}^{l},$$

which contains every integer spin, once each. That is too much: we want *one* irreducible representation at a time, and we want the half-integer spins as well. The Borel–Weil construction fixes both problems with two changes.

(a) Replace functions by *sections of a line bundle*: wavefunctions of a particle that carries a charge n under the circle T that we quotiented by. (Physically: put a magnetic monopole inside the sphere.) This lets the half-integer spins in.

(b) Keep only the *holomorphic* sections. This cuts the infinite sum down to a single irreducible representation. (Physically: keep only the lowest Landau level.)

The rest of this subsection makes each step concrete, first for a general compact group in words, then for SU(2) with every formula written out.

$$\underbrace{\lambda : T \to \mathrm{U}(1)}_{\text{a weight}} \xrightarrow{\text{associate}} \underbrace{L_\lambda = G \times_T \mathbb{C} \to G/T}_{\text{a line bundle}} \xrightarrow{G/T = G_{\mathbb{C}}/B} \underbrace{L_\lambda \text{ holomorphic}}_{\text{unique structure}}$$
$$\xrightarrow{\text{take } \Gamma_{\text{hol}}} \underbrace{\Gamma(G/T, \mathcal{O}(L_\lambda)) \cong V_\lambda}_{\text{the irreducible representation}} \tag{6.146}$$

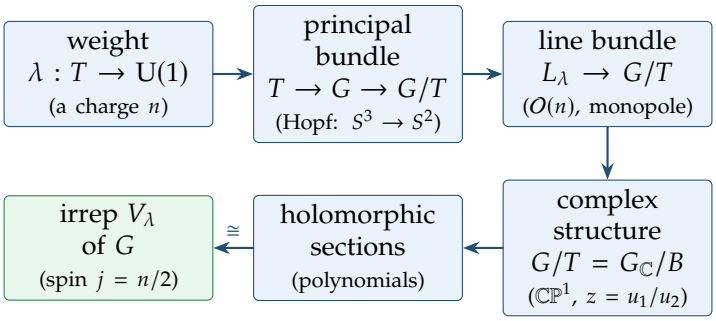


Figure 6.28: The Borel–Weil pipeline (6.146). Read it like the Verma pipeline of Figure 6.22: a small piece of data (here a weight, there a highest weight) is fed into a construction that produces a big space (here all sections, there the Verma module), and a condition (here holomorphicity, there the removal of null states) cuts it down to exactly one irreducible representation. The small type follows the SU(2) example that we work out in full.

Step 1: weights are characters of the torus

Let G be a compact connected Lie group, such as SU(2) or SU(3). A **maximal torus** $T \subset G$ is a maximal connected abelian subgroup. For SU(N) it is the diagonal matrices. Every T is a product of circles, $T \cong \mathrm{U}(1)^r$, and r is the rank.

In a unitary representation of G, the elements of T commute with each other, so they can be diagonalised simultaneously. On the common eigenvectors v_i, each $t \in T$ acts by a phase: $t\, v_i = \lambda_i(t)\, v_i$ with $|\lambda_i(t)| = 1$. Since $(tt')v_i = t(t'v_i)$, the map $\lambda_i : T \to \mathrm{U}(1)$ is a group homomorphism.

Working definition 6.24 (weight as a character)

A **weight** of a representation is a homomorphism $\lambda : T \to \mathrm{U}(1)$ such that some nonzero vector obeys $t\, v = \lambda(t)\, v$ for all $t \in T$. *How to compute:* diagonalise the torus. For $T = \mathrm{U}(1)^r$ every continuous homomorphism is $\lambda(e^{i\phi_1}, \dots, e^{i\phi_r}) = e^{i(n_1\phi_1 + \cdots + n_r\phi_r)}$ with integers n_i, so a weight is a vector of integers. Its derivative $\mathrm{d}\lambda$ at the identity is the "weight" μ of Section 6.3.8 (an eigenvalue of the Cartan generators H^i [the commuting Hermitian generators whose exponentials make up T]).

Why must the exponents be integers? A continuous homomorphism $\lambda : \mathrm{U}(1) \to \mathrm{U}(1)$ has the form $\lambda(e^{i\phi}) = e^{i\kappa\phi}$ for some real κ (it is a one-parameter group of phases). Then

$$1 \overset{(1)}{=} \lambda(1) \overset{(2)}{=} \lambda(e^{2\pi i}) \overset{(3)}{=} e^{2\pi i\kappa} \qquad \Longrightarrow \qquad \kappa \in \mathbb{Z}.$$

(1) A homomorphism sends the identity to the identity.
(2) $e^{2\pi i} = 1$ in U(1).
(3) The formula for λ with $\phi = 2\pi$.

This is the same argument that quantises angular momentum about an axis: a rotation by 2π must act trivially. The lattice of weights is where all the integers of representation theory come from.

Worked example 6.46: the weights of SU(2)

For SU(2) take

$$T = \left\{ t_\phi = \begin{pmatrix} e^{i\phi} & 0 \\ 0 & e^{-i\phi} \end{pmatrix} \right\} \cong \mathrm{U}(1), \qquad \lambda_n(t_\phi) := e^{in\phi}, \quad n \in \mathbb{Z}. \tag{6.147}$$

We find the weights of the spin-j representation. In the fundamental representation $J^3 = \sigma^3/2$, so $t_\phi = e^{i\phi\sigma^3} = e^{2i\phi J^3}$. This identity between group elements holds in every representation, so on the basis $|j, m\rangle$

$$t_\phi |j,m\rangle \overset{(1)}{=} e^{2i\phi J^3}|j,m\rangle \overset{(2)}{=} e^{2im\phi}|j,m\rangle \overset{(3)}{=} \lambda_{2m}(t_\phi)|j,m\rangle.$$

(1) $t_\phi = \exp(2i\phi J^3)$.
(2) $J^3|j,m\rangle = m|j,m\rangle$.
(3) Definition (6.147) with $n = 2m$.

So the weights of spin j are $\lambda_{2j}, \lambda_{2j-2}, \ldots, \lambda_{-2j}$, and the **highest weight** is λ_{2j}. The integer $n = 2m$ is always an integer even when m is a half-integer: this is why the torus of SU(2) (rather than of SO(3), where t would be a rotation by ϕ and not 2ϕ) sees the spinors. The irreducible representation with highest weight λ_n, $n \geq 0$, is spin $j = n/2$.

Watch out: which weight labels the representation

An irreducible representation contains many weights, but only one highest weight (for a fixed choice of which roots are "positive"). Nash writes V_λ for the irreducible representation whose highest weight is λ. For SU(2), V_{λ_n} is spin $n/2$ and has dimension $n+1$. The highest weights that actually occur are the **dominant** ones; for SU(2) this means $n \geq 0$.

Step 2: the principal T-bundle $G \to G/T$

The torus acts on G by right multiplication, $g \mapsto gt$. The orbits are copies of T, and the space of orbits is the homogeneous space G/T. So G is a principal T-bundle over G/T (Nash's (6.63)):

$$T \longrightarrow G \xrightarrow{\pi} G/T. \tag{6.148}$$

(Refresher: a principal T-bundle is a space with a free right T-action whose orbits are the fibres; see Chapters 1 and T3.) The dimension of the base is $\dim G - \dim T$: $3 - 1 = 2$ for SU(2), $8 - 2 = 6$ for SU(3).

Worked example 6.47: $\mathrm{SU}(2) \to S^2$ is the Hopf fibration

Write a general element of SU(2) in terms of its *second* column $u = (u_1, u_2)$:

$$g = \begin{pmatrix} \bar{u}_2 & u_1 \\ -\bar{u}_1 & u_2 \end{pmatrix}, \qquad |u_1|^2 + |u_2|^2 = 1. \tag{6.149}$$

(Check: $\det g = |u_2|^2 + |u_1|^2 = 1$, and the two columns are orthonormal.) So $\mathrm{SU}(2) \cong S^3$, the unit sphere in $\mathbb{C}^2$. Now multiply on the right by the torus:

$$g\,t_\phi \overset{(1)}{=} \begin{pmatrix} \bar{u}_2 & u_1 \\ -\bar{u}_1 & u_2 \end{pmatrix} \begin{pmatrix} e^{i\phi} & 0 \\ 0 & e^{-i\phi} \end{pmatrix} \overset{(2)}{=} \begin{pmatrix} \bar{u}_2 e^{i\phi} & u_1 e^{-i\phi} \\ -\bar{u}_1 e^{i\phi} & u_2 e^{-i\phi} \end{pmatrix}.$$

(1) Definitions (6.147) and (6.149).
(2) Right multiplication by a diagonal matrix scales the columns.

The second column goes to $e^{-i\phi}u$: right multiplication by T is the overall phase rotation of $u \in S^3$. Two elements are in the same orbit iff their second columns are proportional, so the orbit is labelled by the ratio

$$\pi(g) = [u_1 : u_2] \in \mathbb{CP}^1, \qquad z := \frac{u_1}{u_2} \in \mathbb{C} \cup \{\infty\}.$$

Thus $\mathrm{SU}(2)/T = \mathbb{CP}^1 = S^2$ and (6.148) is the Hopf fibration $S^1 \to S^3 \to S^2$ (Figure 6.29). The point z is also a direction in $\mathbb{R}^3$: $\pi(g)$ can be identified with the unit vector $\hat{n} = u^\dagger \vec{\sigma} u$, the direction of the spin of the state u.

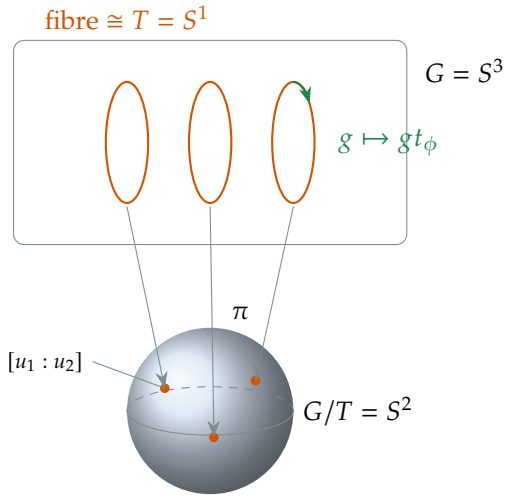


Figure 6.29: The principal T-bundle $G \to G/T$ for $G = \mathrm{SU}(2)$. Each point $[u_1 : u_2]$ of the base sphere has a whole circle of group elements above it, the phases $e^{-i\phi}u$ of the second column; the torus moves points round these circles. This is the Hopf fibration, and the associated line bundles are the monopole bundles over S^2.

Step 3: the associated line bundle L_λ

A weight $\lambda : T \to \mathrm{U}(1)$ lets T act on the complex line $\mathbb{C}$. From a principal bundle and an action on a vector space we can always build an associated vector bundle (Chapter T3): here a complex line bundle.

Working definition 6.25 (the line bundle L_λ and its sections)

$L_\lambda = G \times_T \mathbb{C}$ is the set of pairs $[g, w]$, $g \in G$, $w \in \mathbb{C}$, with the identification $[gt, w] = [g, \lambda(t)w]$. The projection is $[g, w] \mapsto \pi(g)$ [the coset gT, the point of G/T lying below g].
How to compute with it: a section of L_λ [a smooth choice of one point of L_λ over each point

of G/T] is the same thing as a function $f : G \to \mathbb{C}$ obeying

$$f(gt) = \lambda(t)^{-1} f(g) \qquad \text{for all } g \in G,\ t \in T. \tag{6.150}$$

The group acts on sections by left translation, $(h \cdot f)(g) = f(h^{-1}g)$.

The dictionary between sections s and functions f is $s(\pi(g)) = [g, f(g)]$. For this to be well defined, the answer must not depend on which g above the point we used. Replacing g by gt,

$$[gt, f(gt)] \overset{(1)}{=} [g, \lambda(t)\, f(gt)] \overset{(2)}{=} [g, f(g)].$$

(1) The identification rule $[gt, w] = [g, \lambda(t)w]$ with $w = f(gt)$.
(2) The equivariance (6.150): $\lambda(t)f(gt) = f(g)$.

So equivariant functions give sections, and conversely every section gives an equivariant function. Left translation preserves the condition (6.150), because left and right multiplication commute: $(h \cdot f)(gt) = f(h^{-1}gt) = \lambda(t)^{-1} f(h^{-1}g) = \lambda(t)^{-1}(h \cdot f)(g)$. Hence the space of all sections $\Gamma(G/T, L_\lambda)$ is a representation of G. It is infinite-dimensional; the next two steps cut it down.

Physical picture: sections are wavefunctions of a charged particle

A function obeying (6.150) is a wavefunction on the big space G that picks up a definite phase when you move along the fibre. This is exactly the "wavefunction in a fixed gauge" of a particle of charge n in a magnetic field: choosing a representative g over each point is choosing a gauge, and changing $g \to gt$ is a gauge transformation, under which the wavefunction changes by the phase $\lambda(t)^{-1}$. The line bundle L_λ is the bundle whose sections are these wavefunctions, and its first Chern class is the magnetic flux (Example 6.48).

Worked example 6.48: the sections of L_{λ_n} over S^2

Take $G = \mathrm{SU}(2)$ and $\lambda = \lambda_n$. By (6.149) a function on SU(2) is a function $f(u, \bar{u})$ on S^3, and by the example above $g\, t_\phi$ has second column $e^{-i\phi}u$. The condition (6.150) becomes

$$f\left(e^{-i\phi}u,\ e^{i\phi}\bar{u}\right) = e^{-in\phi}\, f(u, \bar{u}). \tag{6.151}$$

The monomial $u_1^p u_2^q \bar{u}_1^r \bar{u}_2^s$ gets the factor $e^{-i(p+q-r-s)\phi}$, so the smooth sections are (limits of) polynomials with $p + q - r - s = n$: "charge" n = (number of u's) − (number of $\bar{u}$'s). For $n = 0$ these are the functions on S^2, i.e. all spherical harmonics. In every case the space is infinite-dimensional.

The Chern class. Choose a representative in each of two patches: $\sigma_N(z)$ with $u_2 > 0$ real (defined where $u_2 \neq 0$, i.e. away from $z = \infty$) and $\sigma_S(z)$ with $u_1 > 0$ real (away from $z = 0$). On the overlap they differ by a torus element: $\sigma_S = \sigma_N\, t_{\phi(z)}$. To find ϕ, compare second columns:

$$e^{-i\phi}\left(\frac{u_1\,\bar{u}_2}{|u_2|},\ |u_2|\right) \overset{(1)}{=} \left(|u_1|,\ \frac{u_2\,\bar{u}_1}{|u_1|}\right) \qquad \Longrightarrow \qquad e^{i\phi} \overset{(2)}{=} \frac{u_1\bar{u}_2}{|u_1|\,|u_2|} \overset{(3)}{=} \frac{z}{|z|}.$$

(1) The second column of σ_N is u rephased to make u_2 positive; that of σ_S is u rephased to make u_1 positive; right multiplication by t_ϕ multiplies the second column by $e^{-i\phi}$.
(2) Solve the first component for $e^{i\phi}$.
(3) $z = u_1/u_2 = u_1\bar{u}_2/|u_2|^2$, so $z/|z| = u_1\bar{u}_2/(|u_1|\,|u_2|)$.

So $\phi(z) = \arg z$. The local wavefunctions $f_N = f \circ \sigma_N$ and $f_S = f \circ \sigma_S$ are then related by

$$f_S = f(\sigma_N t_\phi) = \lambda_n(t_\phi)^{-1} f_N = e^{-in\arg z}\, f_N.$$

The transition function $e^{-in\arg z}$ winds n times (up to orientation) round the equator, which is the definition of a line bundle of Chern number n (Chapter T2): L_{λ_n} is the bundle of a Dirac monopole with $eg = n/2$ (Dirac's condition is $2eg \in \mathbb{Z}$). In algebraic geometry it is called $\mathcal{O}(n)$ (Chapter 5).

Step 4: G/T is a complex manifold

To speak of holomorphic sections, G/T must be a complex manifold. It is, because of a second description (Nash's (6.64)):

$$G/T \;=\; G_{\mathbb{C}}/B. \tag{6.152}$$

Here $G_{\mathbb{C}}$ is the complexification of G ($\mathrm{SL}(2,\mathbb{C})$ for $\mathrm{SU}(2)$, $\mathrm{SL}(N,\mathbb{C})$ for $\mathrm{SU}(N)$), and B is a **Borel subgroup**: for $\mathrm{SL}(N,\mathbb{C})$, the lower-triangular matrices of determinant one. The right-hand side is a quotient of a complex group by a complex subgroup, so it is automatically a complex manifold, and (6.152) transports the complex structure to G/T.

Worked example 6.49: $\mathrm{SL}(2,\mathbb{C})/B = \mathbb{CP}^1 = \mathrm{SU}(2)/T$ by Gram–Schmidt

$\mathrm{SL}(2,\mathbb{C})$ acts on lines in $\mathbb{C}^2$. Let it act on the line $\mathbb{C}e_2$ spanned by $e_2 = (0,1)^T$; then $g\cdot\mathbb{C}e_2$ is the line through the second column of g. The stabiliser of $\mathbb{C}e_2$ is

$$B = \left\{ \begin{pmatrix} \tau^{-1} & 0 \\ s & \tau \end{pmatrix} :\; \tau\in\mathbb{C}^\times,\; s\in\mathbb{C} \right\},$$

the lower-triangular matrices, because $ge_2 = (g_{12}, g_{22})$ lies in $\mathbb{C}e_2$ iff $g_{12} = 0$. Every line is reached, so $\mathrm{SL}(2,\mathbb{C})/B = \mathbb{CP}^1$, the space of lines, with the holomorphic coordinate $z = g_{12}/g_{22}$.
Why is this the same as $\mathrm{SU}(2)/T$? Two facts.

(i) $\mathrm{SU}(2)$ alone already acts transitively on lines: any unit vector u is the second column of the element (6.149). So every $g\in\mathrm{SL}(2,\mathbb{C})$ can be written $g = kb$ with $k\in\mathrm{SU}(2)$, $b\in B$. This is Gram–Schmidt run from the second column: normalise ge_2 to get u, take k as in (6.149); then $k^{-1}g$ maps e_2 into $\mathbb{C}e_2$, i.e. is lower triangular.

(ii) The elements of $\mathrm{SU}(2)$ that fix the line $\mathbb{C}e_2$ are $\mathrm{SU}(2)\cap B$, the unitary lower-triangular matrices. A unitary triangular matrix is diagonal (its columns are orthonormal), so $\mathrm{SU}(2)\cap B = T$.

Together: $\mathrm{SU}(2)/T \to \mathrm{SL}(2,\mathbb{C})/B$, $kT\mapsto kB$, is well defined, onto by (i) and one-to-one by (ii). In coordinates it is $[u_1:u_2]\mapsto z = u_1/u_2$, the coordinate of Example 6.47.

The same argument works in general: Gram–Schmidt (the Iwasawa decomposition) gives $G_{\mathbb{C}} = G\cdot B$, and $G\cap B = T$. For $\mathrm{SU}(3)$, $G_{\mathbb{C}}/B$ is the space of **flags** (a line inside a plane inside $\mathbb{C}^3$), a complex manifold of complex dimension 3.

Key idea: G/T is Kähler

G/T carries a G-invariant Kähler metric: a Riemannian metric compatible with the complex structure whose 2-form ω is closed. For $\mathbb{CP}^1$ it is the round metric, with

$$\omega = \frac{i\,\mathrm{d}z\wedge\mathrm{d}\bar z}{(1+|z|^2)^2}, \qquad \frac{1}{2\pi}\int_{\mathbb{CP}^1}\omega = 1. \tag{6.153}$$

The integral is a two-line check. With $z = re^{i\theta}$, $i\,\mathrm{d}z \wedge \mathrm{d}\bar{z} = 2\,\mathrm{d}x \wedge \mathrm{d}y = 2r\,\mathrm{d}r \wedge \mathrm{d}\theta$, and

$$\frac{1}{2\pi}\int_{\mathbb{CP}^1}\omega \overset{(1)}{=} \frac{1}{2\pi}\int_0^{2\pi}\mathrm{d}\theta\int_0^{\infty}\frac{2r\,\mathrm{d}r}{(1+r^2)^2} \overset{(2)}{=} \int_0^{\infty}\frac{\mathrm{d}s}{(1+s)^2} = 1.$$

(1) The point $z = \infty$ has measure zero, so the chart $\mathbb{C}$ suffices.
(2) The θ integral gives 2π; substitute $s = r^2$.

So $\omega/2\pi$ is an integral class, the Chern class of L_{λ_1}. In the language of classical mechanics, $n\,\omega$ is the symplectic form on the phase space of a classical spin of length $n/2$, and the whole Borel–Weil construction is its quantisation (see the physical picture below).

Step 5: the holomorphic structure is unique

A complex line bundle over a complex manifold may carry many inequivalent holomorphic structures, or none. Nash needs to know that L_λ carries exactly one. The tool is the sheaf-cohomology comparison of Chapter 5 (and Chapter 3): the exponential sequence of sheaves

$$0 \longrightarrow \mathbb{Z} \longrightarrow \mathcal{O} \xrightarrow{\exp(2\pi i\,\cdot)} \mathcal{O}^* \longrightarrow 0$$

gives the exact sequence

$$H^1(M;\mathcal{O}) \longrightarrow H^1(M;\mathcal{O}^*) \xrightarrow{c_1} H^2(M;\mathbb{Z}) \longrightarrow H^2(M;\mathcal{O}). \tag{6.154}$$

Here $H^1(M;\mathcal{O}^*)$ is the group of holomorphic line bundles (a Čech 1-cocycle of nonvanishing holomorphic functions is a set of holomorphic transition functions), and c_1 sends a bundle to its Chern class, which classifies the underlying *smooth* bundle. So the two outer groups measure the difference between holomorphic and smooth line bundles. If both vanish, c_1 is an isomorphism: every smooth line bundle has one and only one holomorphic structure.

For G/T they do vanish. This is Nash's (6.65),

$$H^i(G/T;\mathcal{O}) = 0, \qquad i \geq 1, \tag{6.155}$$

a consequence of the Kodaira vanishing theorem (G/T is Kähler with positive curvature in the right sense). For $\mathbb{CP}^1$ we can see it by hand.

Worked example 6.50: $H^1(\mathbb{CP}^1;\mathcal{O}) = 0$ by Laurent series

Cover $\mathbb{CP}^1$ by $U_0 = \{z \neq \infty\}$ and $U_1 = \{z \neq 0\}$, with $U_0 \cap U_1 = \mathbb{C}^\times$. A Čech 1-cochain is a holomorphic function h on $\mathbb{C}^\times$, i.e. a Laurent series. It is a coboundary if $h = f_0 - f_1$ with f_0 holomorphic on U_0 (a power series in z) and f_1 holomorphic on U_1 (a power series in $w = 1/z$, i.e. in non-positive powers of z). Every Laurent series splits this way:

$$h(z) \overset{(1)}{=} \sum_{k\in\mathbb{Z}} h_k z^k \overset{(2)}{=} \underbrace{\sum_{k\geq 0} h_k z^k}_{f_0,\ \text{hol. on } U_0} \; - \; \underbrace{\Big(-\sum_{k\geq 1} h_{-k} w^k\Big)}_{f_1,\ \text{hol. on } U_1}.$$

(1) Laurent expansion on the annulus $\mathbb{C}^\times$.
(2) Split into non-negative and negative powers; $z^{-k} = w^k$.

So every 1-cochain is a coboundary and $H^1(\mathbb{CP}^1;\mathcal{O}) = 0$. With only two open sets there are no 2-cochains, so $H^2(\mathbb{CP}^1;\mathcal{O}) = 0$ too. By (6.154), $H^1(\mathbb{CP}^1;\mathcal{O}^*) \cong H^2(\mathbb{CP}^1;\mathbb{Z}) = \mathbb{Z}$: holomorphic line bundles on $\mathbb{CP}^1$ are classified by one integer, the degree. The holomorphic

structure on L_{λ_n} is unique, and the bundle is $\mathcal{O}(n)$.

The theorem

Key idea: the Borel–Weil theorem (Nash, p. 189)

Let G be compact and connected, T a maximal torus, and $\lambda : T \to \mathrm{U}(1)$ a weight. The space $\Gamma(G/T, \mathcal{O}(L_\lambda))$ of holomorphic sections of the line bundle L_λ is nonzero if and only if λ is the highest weight of an irreducible representation (i.e. λ is dominant). In that case

$$\Gamma(G/T, \mathcal{O}(L_\lambda)) \;\cong\; V_\lambda$$

as representations of G (acting by left translation). Every irreducible representation of G arises this way, exactly once.

The rest of this subsection makes the theorem plausible in two ways: by an explicit computation for SU(2), and by a general argument (Nash's) that uses the Lie algebra and the Peter–Weyl theorem.

The theorem for SU(2), completely explicit

The holomorphicity condition as a differential equation. By Example 6.48, a section of L_{λ_n} is an equivariant function f on S^3. On the region $u_2 \neq 0$ define

$$F := \frac{f}{u_2^n}.$$

Under $u \to e^{-i\phi}u$ both f and u_2^n get the factor $e^{-in\phi}$, so F is invariant: it is a function on the base, $F = F(z, \bar{z})$. Therefore

$$f(u, \bar{u}) = u_2^n \, F(z, \bar{z}), \qquad z = \frac{u_1}{u_2}. \tag{6.156}$$

The section is **holomorphic** when F is a holomorphic function of the complex coordinate z: $\partial_{\bar{z}} F = 0$. This is the Cauchy–Riemann equation. (Replacing u_2^n by any other nowhere-zero holomorphic local section gives the same condition, because the ratio of two such is holomorphic.)

The holomorphic sections are polynomials. Let f be holomorphic. On $u_2 \neq 0$, $F(z)$ is holomorphic in all of $\mathbb{C}$, so it has a convergent power series $F(z) = \sum_{k \geq 0} a_k z^k$. On the other patch $u_1 \neq 0$ use $w = u_2/u_1 = 1/z$ and the local section u_1^n:

$$f \overset{(1)}{=} u_2^n F(z) \overset{(2)}{=} u_1^n \, w^n F(1/w) \overset{(3)}{=} u_1^n \sum_{k \geq 0} a_k \, w^{\,n-k}.$$

(1) (6.156).
(2) $u_2 = u_1 w$ and $z = 1/w$.
(3) The power series of F.

Holomorphy at $w = 0$ (the point $z = \infty$) requires every exponent $n - k$ to be ≥ 0: $a_k = 0$ for $k > n$. Hence

$$f = u_2^n \sum_{k=0}^{n} a_k \left(\frac{u_1}{u_2}\right)^k = \sum_{k=0}^{n} a_k \, u_1^k u_2^{n-k}, \qquad \dim \Gamma(\mathbb{CP}^1, \mathcal{O}(n)) = \begin{cases} n+1, & n \geq 0, \\ 0, & n < 0. \end{cases} \tag{6.157}$$

The holomorphic sections of L_{λ_n} are exactly the homogeneous polynomials of degree n in (u_1, u_2). For $n < 0$ the conditions leave nothing: there are no holomorphic sections, in agreement with the theorem (λ_n is dominant iff $n \geq 0$).

The group action. We now show that these $n+1$ polynomials carry spin $j = n/2$. Let G act by $(U(h)f)(g) = f(h^{-1}g)$; the second column of $h^{-1}g$ is $h^{-1}u$. With the physicists' convention $U(e^{-i\theta\sigma^a/2}) = e^{-i\theta J^a}$, the generators are

$$J^a f := i\,\frac{\mathrm{d}}{\mathrm{d}\theta}\, f\big(e^{i\theta\sigma^a/2}g\big)\Big|_{\theta=0}.$$

For $a = 3$, $e^{i\theta\sigma^3/2}$ multiplies u_1 by $e^{i\theta/2}$ and u_2 by $e^{-i\theta/2}$:

$$J^3 f \overset{(1)}{=} i\,\frac{\mathrm{d}}{\mathrm{d}\theta}\, f\big(e^{i\theta/2}u_1,\, e^{-i\theta/2}u_2\big)\Big|_0 \overset{(2)}{=} i\cdot\frac{i}{2}\big(u_1\partial_1 - u_2\partial_2\big)f = \tfrac{1}{2}\big(u_2\partial_2 - u_1\partial_1\big)f,$$

(1) The second column of $e^{i\theta\sigma^3/2}g$.
(2) Chain rule, with $\partial_i = \partial/\partial u_i$ (f is holomorphic, so it does not depend on $\bar u$).

For the ladder operators combine $J^\pm = J^1 \pm iJ^2$. Since J^a is linear in σ^a, J^+ is obtained by the same formula with $\sigma^a/2$ replaced by $\sigma^+ = \tfrac{1}{2}(\sigma^1 + i\sigma^2) = \left(\begin{smallmatrix}0&1\\0&0\end{smallmatrix}\right)$ (legitimate on holomorphic f, which extend to $\mathrm{SL}(2,\mathbb{C})$):

$$J^+ f \overset{(1)}{=} i\,\frac{\mathrm{d}}{\mathrm{d}s}\, f\big(u_1 + is\,u_2,\, u_2\big)\Big|_0 \overset{(2)}{=} i\cdot i\,u_2\,\partial_1 f = -u_2\,\partial_1 f,$$

(1) $e^{is\sigma^+} = \mathbb{1} + is\sigma^+$ adds is times the second row to the first, so $u_1 \to u_1 + isu_2$.
(2) Chain rule.

and in the same way (σ^- adds is times the first row to the second) $J^- f = -u_1\,\partial_2 f$. Collecting:

$$J^3 = \tfrac{1}{2}\big(u_2\partial_2 - u_1\partial_1\big), \qquad J^+ = -u_2\,\partial_1, \qquad J^- = -u_1\,\partial_2. \tag{6.158}$$

These are the Schwinger–Jordan ("two-oscillator") realisation of angular momentum. Check one relation:

$$[J^+, J^-] \overset{(1)}{=} [u_2\partial_1,\, u_1\partial_2] \overset{(2)}{=} u_2\partial_2 - u_1\partial_1 \overset{(3)}{=} 2J^3.$$

(1) The two minus signs cancel.
(2) $[u_2\partial_1, u_1\partial_2] = u_2[\partial_1, u_1]\partial_2 + u_1[u_2,\partial_2]\partial_1 = u_2\partial_2 - u_1\partial_1$.
(3) (6.158).

On a monomial,

$$J^3\, u_1^{j-m}u_2^{j+m} = \tfrac{1}{2}\big((j+m) - (j-m)\big)u_1^{j-m}u_2^{j+m} = m\,u_1^{j-m}u_2^{j+m},$$

so with $n = 2j$ the monomials $u_1^{j-m}u_2^{j+m}$, $m = -j, \dots, j$, are the J^3 eigenstates. The top one, u_2^n, is killed by $J^+ = -u_2\partial_1$: it is the highest-weight vector, with weight λ_n. The space (6.157) is V_{λ_n}, spin $n/2$. This is the Borel–Weil theorem for SU(2).

The inner product. The representation is unitary for the L^2 inner product on $\mathrm{SU}(2) = S^3$ with the normalised round (Haar) measure $\mathrm{d}\mu$, which is invariant under left and right multiplication. The norms of the monomials follow from a Gaussian trick. On $\mathbb{C}^2$, with d^4u the Lebesgue measure,

$$\begin{aligned} p!\,q! &\overset{(1)}{=} \int_{\mathbb{C}^2} |u_1|^{2p}\,|u_2|^{2q}\; e^{-|u|^2}\,\frac{\mathrm{d}^4u}{\pi^2} \\ &\overset{(2)}{=} \frac{2\pi^2}{\pi^2}\int_0^\infty r^{2(p+q)+3}e^{-r^2}\,\mathrm{d}r \int_{S^3} |u_1|^{2p}\,|u_2|^{2q}\,\mathrm{d}\mu \\ &\overset{(3)}{=} (p+q+1)!\int_{S^3} |u_1|^{2p}\,|u_2|^{2q}\,\mathrm{d}\mu. \end{aligned} \tag{6.159}$$

(1) The integral factorises into two copies of $\int_{\mathbb{C}} |w|^{2p}\, e^{-|w|^2} \frac{\mathrm{d}^2 w}{\pi} = \int_0^\infty \rho^{2p} e^{-\rho^2}\, 2\rho\, \mathrm{d}\rho = p!$.
(2) Polar coordinates $u = r\hat{u}$ in $\mathbb{R}^4$: $\mathrm{d}^4 u = r^3 \mathrm{d}r\, \mathrm{d}\Omega_3$ with $\int \mathrm{d}\Omega_3 = 2\pi^2$, so $\mathrm{d}\Omega_3 = 2\pi^2\, \mathrm{d}\mu$.
(3) $\int_0^\infty r^{2N+3} e^{-r^2} \mathrm{d}r = \frac{1}{2}\Gamma(N+2) = \frac{1}{2}(N+1)!$ with $N = p + q$.

So $\left\| u_1^p u_2^q \right\|^2 = \frac{p!\, q!}{(p+q+1)!}$, and distinct monomials are orthogonal (integrate over the phases of u_1 and u_2). The orthonormal basis, with the Condon–Shortley phases, is

$$|j, m\rangle \longleftrightarrow e_m \coloneqq (-1)^{j-m} \sqrt{\frac{(2j+1)!}{(j-m)!\,(j+m)!}}\; u_1^{j-m} u_2^{j+m}, \tag{6.160}$$

and (6.158) gives the textbook matrix elements, e.g. $J^+ e_m = \sqrt{(j-m)(j+m+1)}\, e_{m+1}$. (The sign $(-1)^{j-m}$ absorbs the minus sign in $J^+ = -u_2 \partial_1$.)

Worked example 6.51: spin one as quadratic polynomials

Take $n = 2$, $j = 1$. By (6.160)

$$e_1 = \sqrt{3}\, u_2^2, \qquad e_0 = -\sqrt{6}\, u_1 u_2, \qquad e_{-1} = \sqrt{3}\, u_1^2.$$

Their norms are $3 \cdot \frac{0!\,2!}{3!} = 1$, $6 \cdot \frac{1!\,1!}{3!} = 1$, and $3 \cdot \frac{2!\,0!}{3!} = 1$. Apply $J^+ = -u_2 \partial_1$:

$$J^+ e_{-1} \overset{(1)}{=} \sqrt{3}\,(-u_2)(2u_1) = -2\sqrt{3}\, u_1 u_2 \overset{(2)}{=} \sqrt{2}\, e_0,$$
$$J^+ e_0 \overset{(3)}{=} -\sqrt{6}\,(-u_2)\, u_2 = \sqrt{6}\, u_2^2 \overset{(4)}{=} \sqrt{2}\, e_1, \qquad J^+ e_1 \overset{(5)}{=} 0.$$

(1) $\partial_1 u_1^2 = 2u_1$.
(2) $\sqrt{2}\, e_0 = -\sqrt{12}\, u_1 u_2 = -2\sqrt{3}\, u_1 u_2$.
(3) $\partial_1 (u_1 u_2) = u_2$.
(4) $\sqrt{2}\, e_1 = \sqrt{6}\, u_2^2$.
(5) u_2^2 does not depend on u_1.

In the basis (e_1, e_0, e_{-1}),

$$J^3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix}, \qquad J^+ = \begin{pmatrix} 0 & \sqrt{2} & 0 \\ 0 & 0 & \sqrt{2} \\ 0 & 0 & 0 \end{pmatrix}, \qquad J^- = (J^+)^T,$$

the standard spin-one matrices, and $[J^+, J^-] = \mathrm{diag}(2, 0, -2) = 2J^3$. Three holomorphic sections of the monopole bundle of charge 1 are a vector of SO(3). For $n = 1$ the two sections $\sqrt{2}\, u_2$ and $-\sqrt{2}\, u_1$ are a spinor. The half-integer spins appear because the element $-\mathbb{1} = t_\pi \in T$ acts on sections of L_{λ_n} by $f(-u) = (-1)^n f(u)$: for odd n the rotation by 2π is -1, so only SU(2), not SO(3), acts.

Where the states live. A section does not have a value in $\mathbb{C}$ at a point, but its length $|f(u)|^2$ is a function on S^2 (it is invariant under $u \to e^{-i\phi} u$). For the basis (6.160), parametrise the point by $z = \tan\frac{\vartheta}{2}\, e^{i\varphi}$, so that $|u_1| = \sin\frac{\vartheta}{2}$ and $|u_2| = \cos\frac{\vartheta}{2}$. Then

$$\sum_{m=-j}^{j} |e_m|^2 \overset{(1)}{=} (2j+1) \sum_m \binom{2j}{j-m} \left(\sin^2\tfrac{\vartheta}{2}\right)^{j-m} \left(\cos^2\tfrac{\vartheta}{2}\right)^{j+m} \overset{(2)}{=} (2j+1) \left(\sin^2\tfrac{\vartheta}{2} + \cos^2\tfrac{\vartheta}{2}\right)^{2j} = 2j+1. \tag{6.161}$$

(1) (6.160): $\frac{(2j+1)!}{(j-m)!(j+m)!} = (2j+1)\binom{2j}{j-m}$.

(2) The binomial theorem.

The total density is uniform over the sphere, as rotational invariance demands (the analogue of Unsöld's theorem $\sum_m |Y_{lm}|^2 = \frac{2l+1}{4\pi}$). The individual states are rings: $|e_m|^2 \propto \sin^{2(j-m)}\frac{\vartheta}{2}\cos^{2(j+m)}\frac{\vartheta}{2}$ peaks where $\cos\vartheta = m/j$, the latitude at which a classical spin of length j has z-component m (Figure 6.30).

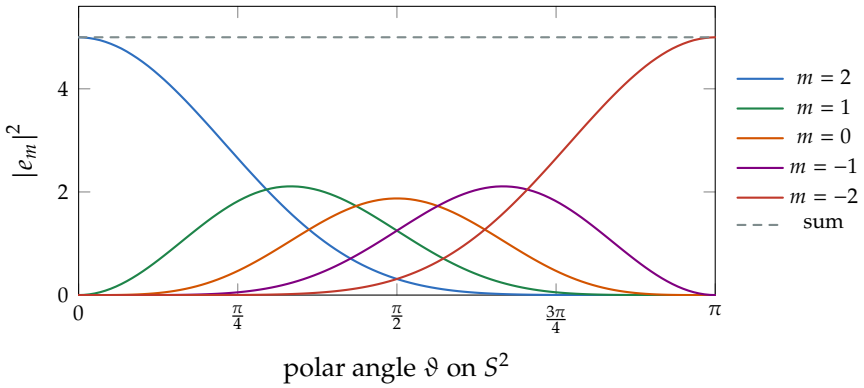


Figure 6.30: The holomorphic sections of $O(4)$ over S^2, the spin-2 representation, drawn as densities $|e_m|^2$ against the polar angle. Each state is a ring around the z-axis, peaked at $\cos\vartheta = m/j$, where a classical spin of length $j = 2$ has $J^3 = m$; the highest-weight state $e_2 = \sqrt{5}\,u_2^4$ sits at the north pole $z = 0$. The sum (dashed) is the constant $2j + 1 = 5$ (6.161).

Physical picture: the lowest Landau level and the quantisation of a spin

Put a charged particle on a sphere around a magnetic monopole with n flux quanta. Its wavefunctions are sections of L_{λ_n} (Example 6.48), and its Hamiltonian is the covariant Laplacian. The states of lowest energy (the **lowest Landau level**) are exactly the holomorphic sections, just as in the plane the lowest Landau level consists of $\psi = F(z)\,e^{-|z|^2/4\ell_B^2}$ with F holomorphic. So:

- the degeneracy of the lowest Landau level is $n + 1$ = (number of flux quanta) +1, which is (6.157) and also the Riemann–Roch / index theorem index $\bar\partial$ = deg +1 on $\mathbb{CP}^1$ (Chapter 4);
- the rotations act on it with spin $j = n/2$, which is the known result that a charge in a monopole field has total angular momentum at least $|eg| = n/2$;
- as $n \to \infty$ at fixed radius the states become sharply localised rings (Figure 6.30), i.e. classical.

Read backwards: the sphere with the area form $n\,\omega$ is the classical phase space of a spin, $\vec S \in \mathbb{R}^3$ with $\left|\vec S\right| = n/2$. The Borel–Weil space is its quantisation, with "holomorphic" playing the role of "functions of q only" and the number of states equal to $\frac{1}{2\pi}\int n\,\omega + 1$ (phase-space area over $2\pi\hbar$, plus a quantum correction). This is **geometric quantisation**, and Borel–Weil is its founding example.

Worked example 6.52: $SU(3)$ **and the flag manifold**

For $G = \mathrm{SU}(3)$, $G/T = G_{\mathbb{C}}/B$ is the space of flags $\ell \subset P \subset \mathbb{C}^3$, a line inside a plane. Record the line by a nonzero vector $x \in \mathbb{C}^3$ (up to scale) and the plane by its normal covector p (up

to scale), with $P = \{v : \ p \cdot v = 0\}$. The condition $\ell \subset P$ is

$$p \cdot x = p_1 x^1 + p_2 x^2 + p_3 x^3 = 0, \tag{6.162}$$

so the flag manifold is the hypersurface (6.162) in $\mathbb{CP}^2 \times \mathbb{CP}^{2*}$, of complex dimension $2 + 2 - 1 = 3$ (real dimension $6 = 8 - 2$, as it should be).
The weights are now pairs of integers (a, b), and the holomorphic sections of the line bundle $L_{(a,b)}$ are the polynomials of degree a in x and b in p, modulo the relation (6.162). Count them. There are $\binom{a+2}{2}$ monomials of degree a in three variables. Multiples of $p \cdot x$ in bidegree (a, b) are $p \cdot x$ times an arbitrary polynomial of bidegree $(a - 1, b - 1)$. So

$$\dim \Gamma(L_{(a,b)}) \overset{(1)}{=} \binom{a+2}{2}\binom{b+2}{2} - \binom{a+1}{2}\binom{b+1}{2} \overset{(2)}{=} \tfrac{1}{2}(a+1)(b+1)(a+b+2).$$

(1) All bihomogeneous polynomials, minus the multiples of $p \cdot x$ (multiplication by $p \cdot x$ is injective on polynomials).
(2) Algebra: $\frac{1}{4}(a+1)(b+1)\big[(a+2)(b+2) - ab\big] = \frac{1}{4}(a+1)(b+1)(2a+2b+4)$.

This is the Weyl dimension formula for SU(3). Check: $(1, 0)$ gives 3 and $(0, 1)$ gives 3 (the coordinates x^i and p_i: the two three-dimensional representations **3** and $\bar{\mathbf{3}}$), $(1, 1)$ gives $9 - 1 = 8$ (all $x^i p_j$ minus the trace $p \cdot x$: the octet), $(2, 0)$ gives 6 and $(3, 0)$ gives 10 (the decuplet). The quark model is a list of holomorphic sections over the flag manifold. For $a < 0$ or $b < 0$ there are no polynomials and no sections, which is the "only if" half of the theorem.

6.3.11 The complex structure of G/T from the Lie algebra

Step 4 gave G/T its complex structure through the identification $G/T = G_{\mathbb{C}}/B$. That is quick, but it hides the answer to a simple question: standing at a point of G/T, which tangent directions are the "holomorphic" ones, and which are their complex conjugates? Nash answers it directly with the Lie algebra (his (6.66)–(6.68)). The answer is the same one a physicist uses every day for angular momentum: split the generators into Cartan generators, raising operators and lowering operators.

The tangent space of G/T

Let $o = eT$ be the base point of G/T. A curve through o is $s \mapsto e^{sX}T$ with $X \in \mathfrak{g}$. If $X \in \mathfrak{t}$, then $e^{sX} \in T$ and the curve does not move. So the tangent space is

$$T_o(G/T) = \mathfrak{g}/\mathfrak{t}.$$

It is convenient to pick a complement of $\mathfrak{t}$ inside $\mathfrak{g}$. Nash's (6.66) is

$$\mathfrak{g} = \mathfrak{t} \oplus \mathfrak{h}, \tag{6.163}$$

where $\mathfrak{h}$ is the orthogonal complement of $\mathfrak{t}$ for the inner product $\langle X, Y\rangle = -\operatorname{tr}(XY)$. Then $T_o(G/T) \cong \mathfrak{h}$ and $\dim G/T = \dim \mathfrak{h} = \dim G - \operatorname{rank} G$. (Here and below $\mathfrak{g}$ means the real Lie algebra of *anti-Hermitian* matrices, the mathematicians' convention of Chapter T3; a physicist's Hermitian generator J corresponds to $iJ \in \mathfrak{g}$.)

Worked example 6.53: the splitting for SU(2) and SU(3)

For SU(2), $\mathfrak{g} = \mathrm{span}_{\mathbb{R}}\{i\sigma^1, i\sigma^2, i\sigma^3\}$, the torus (6.147) has $\mathfrak{t} = \mathrm{span}_{\mathbb{R}}\{i\sigma^3\}$, and $\mathfrak{h} = \mathrm{span}_{\mathbb{R}}\{i\sigma^1, i\sigma^2\}$: two directions, the tangent plane of S^2.
For SU(3), $\mathfrak{t}$ is the traceless diagonal matrices $\mathrm{diag}(i\theta_1, i\theta_2, i\theta_3)$ with $\theta_1 + \theta_2 + \theta_3 = 0$ (two real parameters), and $\mathfrak{h}$ is the off-diagonal anti-Hermitian matrices, fixed by three complex numbers a_{12}, a_{13}, a_{23} above the diagonal (six real parameters). So $\dim G/T = 6$, the flag manifold of Example 6.52.

The torus rotates $\mathfrak{h}$: root spaces

Why does Nash call $\mathfrak{h}$ "the sum of the root spaces"? Because the torus acts on $\mathfrak{g}$ by conjugation, $\mathrm{Ad}(t)X = tXt^{-1}$, and this action fixes $\mathfrak{t}$ pointwise and *rotates* $\mathfrak{h}$. For SU(2),

$$t_\phi\,\sigma^1\,t_\phi^{-1} \overset{(1)}{=} \begin{pmatrix} 0 & e^{2i\phi} \\ e^{-2i\phi} & 0 \end{pmatrix} \overset{(2)}{=} \cos 2\phi\;\sigma^1 - \sin 2\phi\;\sigma^2,$$
$$t_\phi\,\sigma^2\,t_\phi^{-1} \overset{(3)}{=} \begin{pmatrix} 0 & -ie^{2i\phi} \\ ie^{-2i\phi} & 0 \end{pmatrix} \overset{(4)}{=} \sin 2\phi\;\sigma^1 + \cos 2\phi\;\sigma^2.$$

(1) Conjugating by $\mathrm{diag}(e^{i\phi}, e^{-i\phi})$ multiplies the $(1,2)$ entry by $e^{i\phi}\cdot e^{i\phi}$ and the $(2,1)$ entry by $e^{-i\phi}\cdot e^{-i\phi}$.
(2) Compare entries: $\cos 2\phi\,\sigma^1 - \sin 2\phi\,\sigma^2$ has $(1,2)$ entry $\cos 2\phi + i\sin 2\phi = e^{2i\phi}$.
(3) The same rule applied to σ^2, whose $(1,2)$ entry is $-i$.
(4) $(1,2)$ entry of the right side: $\sin 2\phi - i\cos 2\phi = -ie^{2i\phi}$.

So $\mathrm{Ad}(t_\phi)$ turns the plane $\mathfrak{h}$ by the angle -2ϕ. A real rotation has no real eigenvectors, but it has complex ones. That is why Nash complexifies, (6.67):

$$\mathfrak{g}_{\mathbb{C}} = \mathfrak{t}_{\mathbb{C}} \oplus \mathfrak{h}_{\mathbb{C}}. \tag{6.164}$$

In $\mathfrak{h}_{\mathbb{C}}$ the eigenvectors of the rotation are the familiar step operators

$$\sigma^{\pm} = \tfrac{1}{2}(\sigma^1 \pm i\sigma^2), \qquad \sigma^+ = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad \sigma^- = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix},$$

and by the rule in item (1),

$$\mathrm{Ad}(t_\phi)\,\sigma^{\pm} = e^{\pm 2i\phi}\,\sigma^{\pm} = \lambda_{\pm 2}(t_\phi)\,\sigma^{\pm}. \tag{6.165}$$

The eigenvalues are weights, in the sense of Step 1, of the adjoint representation: they are $\lambda_{\pm 2}$, the two nonzero weights of spin 1. These weights of the adjoint representation are the **roots**. Nash's (6.68) splits $\mathfrak{h}_{\mathbb{C}}$ into positive and negative root spaces:

$$\mathfrak{g}_{\mathbb{C}} = \mathfrak{t}_{\mathbb{C}} \oplus \mathfrak{h}^+ \oplus \mathfrak{h}^-, \qquad \mathfrak{h}^+ = \mathbb{C}\sigma^+, \quad \mathfrak{h}^- = \mathbb{C}\sigma^-. \tag{6.166}$$

Working definition 6.26 (roots and a choice of positive roots)

Diagonalise $\mathrm{Ad}(T)$ [conjugation $X \mapsto tXt^{-1}$ by each element t of the torus] on $\mathfrak{g}_{\mathbb{C}}$. The eigenvectors outside $\mathfrak{t}_{\mathbb{C}}$ are the **step operators** E_α, with $\mathrm{Ad}(t)E_\alpha = \alpha(t)E_\alpha$; the characters α [homomorphisms $T \to \mathrm{U}(1)$, exactly as for weights in Step 1] are the **roots**. They come in pairs $\pm\alpha$ (complex conjugation swaps E_α and $E_{-\alpha}$ up to sign). *How to choose the positive ones:* draw a line (a hyperplane) through the origin of the weight space [$\mathbb{R}^r$, with each weight recorded by its vector of integers $(n_1, \ldots, n_r)$] that contains no root, and call positive the

roots on one side of it. The choice is a **Weyl chamber**. Then $\mathfrak{h}^{\pm}$ is spanned by the $E_{\pm\alpha}$ with $\alpha > 0$. For SU(N) with the usual choice, $\mathfrak{h}^{+}$ is the strictly upper-triangular and $\mathfrak{h}^{-}$ the strictly lower-triangular matrices.

For SU(3) the step operators are the matrix units E_{ij}, $i \neq j$ (a 1 in row i, column j). Conjugation by $t = \operatorname{diag}(e^{i\theta_1}, e^{i\theta_2}, e^{i\theta_3})$ gives $\operatorname{Ad}(t)E_{ij} = e^{i(\theta_i-\theta_j)}E_{ij}$, so the roots are $\theta_i - \theta_j$. Name the two **simple roots** $\alpha_1 = \theta_1 - \theta_2$ and $\alpha_2 = \theta_2 - \theta_3$; then $\theta_1 - \theta_3 = \alpha_1 + \alpha_2$. The six roots form a hexagon (Figure 6.31), and the upper-triangular choice is

$$\mathfrak{h}^{+} = \operatorname{span}\{E_{12}, E_{23}, E_{13}\} \quad \longleftrightarrow \quad \{\alpha_1,\ \alpha_2,\ \alpha_1 + \alpha_2\}.$$

Notice that $\mathfrak{t}_{\mathbb{C}} \oplus \mathfrak{h}^{-}$ is the lower-triangular matrices: the Lie algebra $\mathfrak{b}$ of the Borel subgroup B of Step 4. This is the link between Nash's two descriptions of the complex structure.

The holomorphic directions

Now Nash's prescription: *declare the directions in $\mathfrak{h}^{+}$ to be the holomorphic ones.* More precisely, the complexified tangent space at o is $\mathfrak{h}_{\mathbb{C}} = \mathfrak{h}^{+} \oplus \mathfrak{h}^{-}$, and we declare

$$T_o^{1,0}(G/T) = \mathfrak{h}^{+}, \qquad T_o^{0,1}(G/T) = \mathfrak{h}^{-}.$$

Then move this choice to every other point gT by the left translation $g : G/T \to G/T$. The result is well defined because the only ambiguity in the translating element is $g \to gt$, and $\operatorname{Ad}(t)$ maps $\mathfrak{h}^{+}$ to itself by (6.165). A split of the complexified tangent spaces of this kind is the same thing as an **almost complex structure** J, a linear map on each real tangent space with $J^2 = -1$: the $(1,0)$ vectors are $v - iJv$ and the $(0,1)$ vectors are $v + iJv$.

Worked example 6.54: the rule reproduces $z = u_1/u_2$

Take SU(2) and the coordinate $z = g_{12}/g_{22}$ of Example 6.49, which is 0 at o. Move away from o along the two real directions of $\mathfrak{h}$:

$$e^{s\,i\sigma^2} \overset{(1)}{=} \begin{pmatrix} \cos s & \sin s \\ -\sin s & \cos s \end{pmatrix} \qquad \Longrightarrow \quad z = \tan s \approx s,$$

$$e^{s\,i\sigma^1} \overset{(2)}{=} \begin{pmatrix} \cos s & i\sin s \\ i\sin s & \cos s \end{pmatrix} \qquad \Longrightarrow \quad z = i\tan s \approx is.$$

(1) $(i\sigma^2)^2 = -\mathbb{1}$, so $e^{s\,i\sigma^2} = \cos s\,\mathbb{1} + \sin s\,i\sigma^2$, and $i\sigma^2 = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$.

(2) The same with $i\sigma^1 = \left(\begin{smallmatrix} 0 & i \\ i & 0 \end{smallmatrix}\right)$.

So $X_2 = i\sigma^2$ points along the real z-axis and $X_1 = i\sigma^1$ along the imaginary axis. Multiplication by i turns the first into the second, so the complex structure is $JX_2 = X_1$ (and $JX_1 = -X_2$). The $(1,0)$ vector built from X_2 is

$$X_2 - iJX_2 = i\sigma^2 - i\,(i\sigma^1) = \sigma^1 + i\sigma^2 = 2\sigma^{+} \in \mathfrak{h}^{+},$$

exactly as Nash prescribes. Conversely, moving along the complex direction $\sigma^{\pm}$ with a complex parameter s:

$$z\big(e^{s\sigma^{+}}\big) \overset{(1)}{=} \frac{s}{1} = s, \qquad z\big(e^{s\sigma^{-}}\big) \overset{(2)}{=} \frac{0}{1} = 0.$$

(1) $e^{s\sigma^{+}} = \left(\begin{smallmatrix} 1 & s \\ 0 & 1 \end{smallmatrix}\right)$, whose second column is $(s, 1)$.

(2) $e^{s\sigma^{-}} = \left(\begin{smallmatrix} 1 & 0 \\ s & 1 \end{smallmatrix}\right) \in B$ does not move the point o at all.

The step operator σ^{+} moves z holomorphically ($z = s$); the step operator σ^{-} lies in $\mathfrak{b}$ and does nothing to the point.

Integrability: why $\mathfrak{b}^+$ must be closed under brackets

An almost complex structure need not come from complex coordinates. It does exactly when the $(0,1)$ vector fields are closed under the Lie bracket (the **Newlander–Nirenberg theorem**). The necessity of this condition is a one-line physicist's argument. A holomorphic function f is one that is killed by every $(0,1)$ vector field, $\bar{Z}f = 0$. If $\bar{Z}$ and $\bar{W}$ kill f, so does their commutator:

$$[\bar{Z}, \bar{W}]f = \bar{Z}(\bar{W}f) - \bar{W}(\bar{Z}f) = 0.$$

If $[\bar{Z}, \bar{W}]$ had a $(1,0)$ part, the holomorphic functions would be killed by some holomorphic derivative as well, and there would not be enough of them to serve as local coordinates. (The sufficiency is the hard half of the theorem, and we take it on trust.)

For the invariant structure on G/T the $(0,1)$ directions are generated by $\mathfrak{b}^-$ (together with $\mathfrak{t}_{\mathbb{C}}$, which does not move the point), so the condition is that $\mathfrak{b} = \mathfrak{t}_{\mathbb{C}} \oplus \mathfrak{b}^-$ be closed under brackets. Complex conjugation swaps $\mathfrak{b}^+$ and $\mathfrak{b}^-$, so this is the same as Nash's statement that $\mathfrak{b}^+$ is closed. It holds because of how roots add:

$$\mathrm{Ad}(t)\,[E_\alpha, E_\beta] \overset{(1)}{=} [\mathrm{Ad}(t)E_\alpha,\ \mathrm{Ad}(t)E_\beta] \overset{(2)}{=} \alpha(t)\,\beta(t)\,[E_\alpha, E_\beta]. \tag{6.167}$$

(1) Conjugation is a Lie algebra automorphism: $t[X,Y]t^{-1} = [tXt^{-1}, tYt^{-1}]$.
(2) Each E is an eigenvector.

So $[E_\alpha, E_\beta]$ has root $\alpha + \beta$ (in additive notation), or vanishes if $\alpha + \beta$ is not a root. If α and β lie on the positive side of the chosen line, so does $\alpha + \beta$. Hence $[\mathfrak{b}^+, \mathfrak{b}^+] \subset \mathfrak{b}^+$. For SU(3) the only nonzero bracket among the positive step operators is

$$[E_{12}, E_{23}] = E_{12}E_{23} - E_{23}E_{12} = E_{13} - 0 = E_{13},$$

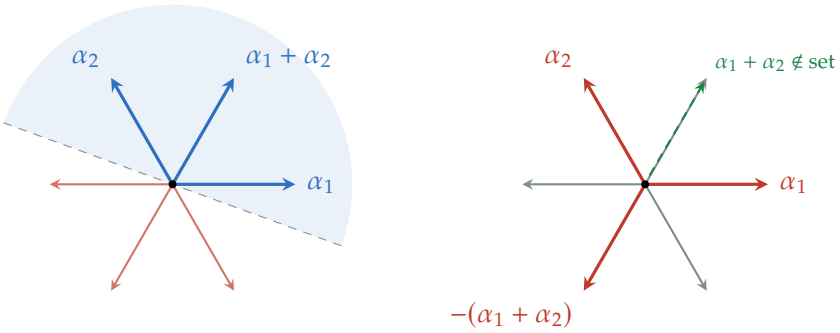
which is root addition $\alpha_1 + \alpha_2$.

Worked example 6.55: an almost complex structure that is not integrable

For SU(3) there are three pairs of roots, so there are $2^3 = 8$ ways to pick one root from each pair and call its step operator "holomorphic". Each choice gives a G-invariant almost complex structure on the flag manifold. Which are integrable? Write the choice as $\{\epsilon_1\alpha_1,\ \epsilon_2\alpha_2,\ \epsilon_3(\alpha_1+\alpha_2)\}$ with signs $\epsilon_i = \pm$. The set must contain the sum of any two of its members whenever that sum is a root. Checking the three possible sums:

- $\epsilon_1\alpha_1 + \epsilon_2\alpha_2$ is a root only if $\epsilon_1 = \epsilon_2$, and then it is $\epsilon_1(\alpha_1+\alpha_2)$: we need $\epsilon_3 = \epsilon_1$;
- $\epsilon_1\alpha_1 + \epsilon_3(\alpha_1+\alpha_2)$ is a root only if $\epsilon_3 = -\epsilon_1$, and then it is $\epsilon_3\alpha_2$: we need $\epsilon_2 = \epsilon_3$;
- $\epsilon_2\alpha_2 + \epsilon_3(\alpha_1+\alpha_2)$ is a root only if $\epsilon_3 = -\epsilon_2$, and then it is $\epsilon_3\alpha_1$: we need $\epsilon_1 = \epsilon_3$.

Run through the eight sign patterns: the only failures are $(+,+,-)$ and $(-,-,+)$, i.e. the choices $\{\alpha_1, \alpha_2, -(\alpha_1+\alpha_2)\}$ and its negative. The other six are the six Weyl chambers (the Weyl group of SU(3) is the permutation group S_3, of order 6), and they all give integrable structures, isomorphic to each other by a permutation of the coordinates of $\mathbb{C}^3$. For the bad choice, $[E_{12}, E_{23}] = E_{13}$ is *not* in $\mathrm{span}\{E_{12}, E_{23}, E_{31}\}$: the almost complex structure is not integrable, and there are no holomorphic coordinates adapted to it. Geometrically (Figure 6.31): the three chosen roots sum to zero, so they cannot all lie on one side of a line.



a Weyl chamber:
$\mathfrak{b}^+$ closed, integrable

three roots summing to zero:
not closed, not integrable

Figure 6.31: The six roots of SU(3). Left: the upper-triangular choice $\mathfrak{b}^+ = \{\alpha_1, \alpha_2, \alpha_1 + \alpha_2\}$ (blue) lies on one side of the dashed line; sums of blue roots stay blue, so $\mathfrak{b}^+$ is closed under brackets and the complex structure on the flag manifold is integrable. Right: the choice $\{\alpha_1, \alpha_2, -(\alpha_1 + \alpha_2)\}$ (red) cannot be put on one side of any line; the bracket of the first two produces $\alpha_1 + \alpha_2$ (green), which is not in the set, so this almost complex structure has no holomorphic coordinates.

Watch out: two meanings of "$\mathfrak{b}$"

Nash's $\mathfrak{b}$ in (6.66) is the complement of the Cartan subalgebra, not the Cartan subalgebra itself (which many physics texts call $\mathfrak{b}$). Here the Cartan subalgebra is $\mathfrak{t}$, and $\mathfrak{b}$ is the span of the step operators.

6.3.12 The Peter–Weyl theorem: Fourier analysis on a group

To prove the Borel–Weil theorem Nash uses a second big theorem, which answers the question: *how do we Fourier-analyse a function on a compact group?* For the circle the answer is Fourier series. The Peter–Weyl theorem is the same statement for any compact G, with the exponentials $e^{in\theta}$ replaced by matrix elements of irreducible representations, the Wigner D-functions of angular momentum theory.

Left and right translations

The group acts on functions on itself in two commuting ways:

$$(L_h f)(g) := f(h^{-1}g), \qquad (R_h f)(g) := f(gh). \tag{6.168}$$

Both are representations: $L_h L_{h'} = L_{hh'}$ (this is why h^{-1} appears) and $(R_h R_{h'} f)(g) = (R_{h'} f)(gh) = f(ghh') = (R_{hh'} f)(g)$. Both are unitary on $L^2(G)$ with the Haar measure $\mathrm{d}\mu$, which is invariant under left and right multiplication (for SU(2) $= S^3$ it is the round measure). The infinitesimal version of R is the derivative along a Lie algebra element,

$$(R_X f)(g) := \frac{\mathrm{d}}{\mathrm{d}s} f\left(g\, e^{sX}\right)\Big|_{s=0}, \tag{6.169}$$

extended complex-linearly to $X \in \mathfrak{g}_{\mathbb{C}}$, $R_{X+iY} := R_X + iR_Y$. Nash's (6.69) is the decomposition of $L^2(G)$ under R into irreducibles,

$$L^2(G) = \bigoplus_{\lambda} c_\lambda\, V_\lambda,$$

with multiplicities c_λ. The theorem says what they are.

Matrix coefficients

Given a finite-dimensional representation V, a vector $v \in V$ and a covector $\xi \in V^*$ (a linear function on V), define the **matrix coefficient**

$$f_{\xi,v}(g) := \xi\big(g^{-1}v\big). \tag{6.170}$$

In a basis, with ξ the i-th dual basis vector and v the j-th basis vector, $f_{\xi,v}(g) = D(g^{-1})_{ij}$: one entry of the representation matrix. Watch what the two translations do to it:

$$(L_h f_{\xi,v})(g) \overset{(1)}{=} \xi\big((h^{-1}g)^{-1}v\big) \overset{(2)}{=} \xi\big(g^{-1}(hv)\big) = f_{\xi,hv}(g), \tag{6.171}$$

$$(R_h f_{\xi,v})(g) \overset{(3)}{=} \xi\big(h^{-1}g^{-1}v\big) \overset{(4)}{=} (h\cdot\xi)\big(g^{-1}v\big) = f_{h\cdot\xi,v}(g). \tag{6.172}$$

(1) Definitions (6.168) and (6.170).
(2) $(h^{-1}g)^{-1} = g^{-1}h$.
(3) $(gh)^{-1} = h^{-1}g^{-1}$.
(4) The **dual representation** on V^* is $(h\cdot\xi)(w) := \xi(h^{-1}w)$.

So left translation moves v and right translation moves ξ. The matrix coefficients of V span a copy of $V \otimes V^*$ inside the functions on G, with $G \times G$ acting by (left, right). This is Nash's (6.70) in the making.

Key idea: the Peter–Weyl theorem (Nash's (6.69), (6.70))

For compact G, the matrix coefficients of the irreducible unitary representations span a dense subspace of $L^2(G)$, and

$$L^2(G) = \widehat{\bigoplus_{\lambda}} V_\lambda \otimes V_\lambda^* \tag{6.173}$$

(the hat means completion) as a representation of $G \times G$, the first factor acting by left and the second by right translation. Each irreducible V_λ occurs exactly once in the pair, so under right translation alone it occurs $c_\lambda = \dim V_\lambda$ times, and the block of V_λ contributes $(\dim V_\lambda)^2$ independent functions. Different irreducibles, and different entries of one, are orthogonal:

$$\int_G \overline{D^\lambda_{ij}(g)}\, D^\mu_{kl}(g)\, \mathrm{d}\mu(g) = \frac{\delta_{\lambda\mu}\,\delta_{ik}\,\delta_{jl}}{\dim V_\lambda} \qquad \Big(\textstyle\int_G \mathrm{d}\mu = 1\Big). \tag{6.174}$$

Why is it true? The orthogonality (6.174) is Schur's lemma: averaging a matrix over the group produces an intertwiner, which is zero between inequivalent irreducibles and a multiple of $\mathbb{1}$ otherwise. Completeness (that nothing is missing) has a physicist's proof via the Laplacian. The Laplacian Δ of a bi-invariant metric on G is a self-adjoint elliptic operator on a compact manifold, so $L^2(G)$ is the orthogonal sum of its eigenspaces, and each eigenspace is finite-dimensional (Chapter 2). Since the metric is bi-invariant, Δ commutes with L_h and R_h, so each eigenspace is a finite-dimensional representation of $G \times G$ inside the functions; a finite-dimensional space of functions invariant under right translations is spanned by matrix coefficients (evaluate at $g = e$). Hence every function is a limit of sums of matrix coefficients. In fact Δ is the quadratic Casimir, so its eigenvalue on $V_\lambda \otimes V_\lambda^*$ is the Casimir of λ.

Worked example 6.56: the circle: Fourier series

For $G = \mathrm{U}(1)$ the irreducible representations are $V_n = \mathbb{C}$ with $e^{i\theta} \mapsto e^{in\theta}$, and the only matrix coefficient is $f(\theta) = e^{-in\theta}$ (from g^{-1}). So (6.173) reads $L^2(S^1) = \widehat{\bigoplus}_{n\in\mathbb{Z}} \mathbb{C}\, e^{in\theta}$: Fourier series. Each irreducible is one-dimensional and occurs once, $(\dim V_n)^2 = 1$; (6.174) is

$\frac{1}{2\pi}\int e^{-im\theta}e^{in\theta}\mathrm{d}\theta = \delta_{mn}$; and the Laplacian $-\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}$ has eigenvalue n^2, the Casimir of U(1).

Worked example 6.57: SU(2) = S^3: hyperspherical harmonics

For SU(2) the irreducibles are the spins $j = 0, \frac{1}{2}, 1, \dots$, so (6.173) predicts that $L^2(S^3)$ contains $(2j+1)^2 = (n+1)^2$ functions (with $n = 2j$) for every $n \geq 0$. Compare with the harmonic analysis of $S^3 \subset \mathbb{R}^4$: the eigenfunctions of the Laplacian are the restrictions of the harmonic homogeneous polynomials on $\mathbb{R}^4$. The polynomials of degree n in 4 variables number $\binom{n+3}{3}$, and the harmonic ones are the kernel of ∇^2, which maps them onto the polynomials of degree $n-2$. So their number is

$$\binom{n+3}{3} - \binom{n+1}{3} \overset{(1)}{=} \frac{(n+3)(n+2)(n+1) - (n+1)n(n-1)}{6} \overset{(2)}{=} \frac{(n+1)\,(6n+6)}{6} = (n+1)^2.$$

(1) Definition of the binomial coefficients.
(2) $(n+3)(n+2) - n(n-1) = n^2 + 5n + 6 - n^2 + n = 6n + 6$.

Exactly as predicted. The eigenvalue matches too: on the unit S^3 the degree-n harmonics have $-\nabla^2_{S^3} = n(n+2)$, and with the unit-length left-invariant vector fields ξ_a (brackets $[\xi_a, \xi_b] = 2\epsilon_{abc}\xi_c$) and $J_a = \frac{i}{2}\xi_a$ (which obey $[J_a, J_b] = i\epsilon_{abc}J_c$),

$$-\nabla^2_{S^3} = -\sum_a \xi_a^2 = 4\sum_a J_a^2 = 4j(j+1) = 2j(2j+2) = n(n+2).$$

The Laplacian on the three-sphere is four times the angular-momentum Casimir. The normalisation (6.174) can also be checked with the moments (6.159): for $j = \frac{1}{2}$ the matrix entries of g are $\pm u_i, \pm\bar{u}_i$, and $\int_{S^3} |u_1|^2\,\mathrm{d}\mu = \frac{1!\,0!}{2!} = \frac{1}{2} = \frac{1}{\dim V_{1/2}}$.

What sits inside $L^2(G)$: functions on G/T, sections, holomorphic sections

The Peter–Weyl theorem now sorts the three spaces of the Borel–Weil story. In each case the condition is a condition on the *right* behaviour of a function, so by (6.172) it is a condition on the covector ξ alone, and v is free.

Sections of L_λ. A section is a function with $f(gt) = \lambda(t)^{-1}f(g)$ (6.150). On a matrix coefficient,

$$f_{\xi,v}(gt) \overset{(1)}{=} f_{t\cdot\xi,v}(g) \overset{(2)}{=} \lambda(t)^{-1}f_{\xi,v}(g) \quad \text{for all } t \in T \iff t\cdot\xi = \lambda(t)^{-1}\xi.$$

(1) (6.172) with $h = t$.
(2) The equivariance condition, which we want to hold.

So ξ must be a weight vector of V^* with weight λ^{-1} (in additive language, $-\lambda$). If e_i is a basis of weight vectors of V with weights μ_i, the dual basis vector e^i has weight μ_i^{-1}, because $(t\cdot e^i)(e_j) = e^i(t^{-1}e_j) = \mu_j(t)^{-1}\delta_{ij}$. Hence

$$\Gamma(G/T, L_\lambda) = \widehat{\bigoplus_\mu} V_\mu \otimes (\text{weight-}\lambda \text{ space of } V_\mu), \tag{6.175}$$

and V_μ appears as often as λ appears among its weights. For λ trivial these are the functions on G/T.

Holomorphic sections. A section is holomorphic when it is killed by the $(0,1)$ directions, i.e. by R_W for every $W \in \mathfrak{h}^-$ (Nash: "it is annihilated by all $W \in \mathfrak{h}^-_{\mathbb{C}}$"). Differentiating (6.172),

$$R_W f_{\xi,v} = f_{W\cdot\xi,v}, \qquad (W\cdot\xi)(w) = -\xi(Ww),$$

so the condition is $W \cdot \xi = 0$ for all $W \in \mathfrak{h}^-$: ξ is killed by all lowering operators. In an irreducible V_μ^* such a vector is unique up to scale. It is the **lowest-weight vector** of V_μ^* (the dual of the highest-weight vector of V_μ), and its weight is μ^{-1}. It satisfies the equivariance condition only if $\mu^{-1} = \lambda^{-1}$. Therefore

$$\begin{aligned}\Gamma\big(G/T, \mathcal{O}(L_\lambda)\big) &= \bigoplus_\mu V_\mu \otimes \big\{\xi \in V_\mu^* : \ \mathfrak{h}^- \cdot \xi = 0, \ \text{weight } \lambda^{-1}\big\} \\ &= \begin{cases} V_\lambda \otimes \mathbb{C}\,\xi_\lambda & \lambda \text{ a highest weight,} \\ 0 & \text{otherwise.} \end{cases}\end{aligned} \tag{6.176}$$

This is the Borel–Weil theorem. Nash's phrasing is that the pulled-back section $\tilde{f}$ is "a lowest weight vector for an irreducible representation": it is a lowest-weight vector for the *right* translations (which act on the V^* factor), and it carries the irreducible representation V_λ under the left translations.

Worked example 6.58: the SU(2) sections are one row of the D-matrix

For spin j, the highest-weight vector of V_j is $|j,j\rangle$ and the covector $\xi_\lambda = \langle j,j|$ is the lowest-weight vector of V_j^* (it is killed by $\langle j,j|J^-$, as $\langle j,j|J^- = (J^+|j,j\rangle)^\dagger = 0$). By (6.176) the holomorphic sections are

$$f_m(g) = \langle j,j|\, g^{-1}\, |j,m\rangle = D^j_{jm}(g^{-1}), \qquad m = -j, \dots, j :$$

one row of the Wigner matrix of g^{-1}. For $j = \frac{1}{2}$ and g as in (6.149), $g^{-1} = g^\dagger = \left(\begin{smallmatrix} u_2 & -u_1 \\ \bar{u}_1 & \bar{u}_2 \end{smallmatrix}\right)$, whose first row is $(u_2, -u_1)$. These are the two holomorphic sections u_2 and $-u_1$ of $\mathcal{O}(1)$ found in (6.157), with the sign of (6.160). By contrast the sections of L_{λ_n} without the holomorphy condition are, by (6.175), all $V_{j'}$ with $j' \geq \frac{|n|}{2}$ and $j' - \frac{n}{2} \in \mathbb{Z}$, each once: these are the **monopole harmonics** of Wu and Yang. For $n = 0$ they are the integer spins $j' = 0, 1, 2, \dots$: the spherical harmonics with which this subsection began.

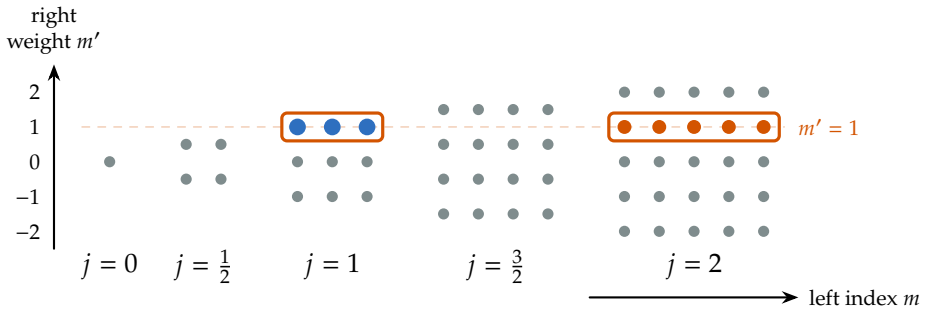


Figure 6.32: Peter–Weyl for SU(2), and the Borel–Weil funnel. Each block of dots is one $V_j \otimes V_j^*$ inside $L^2(S^3)$: $(2j+1)^2$ matrix coefficients, with the left index m running along a row and the height of the row giving the weight m' of the basis vector dual to the covector ξ. The sections of L_{λ_2} are the rows with $m' = 1$, the value fixed by the equivariance (orange frames on the dashed line): one row in each integer spin $j \geq 1$, infinitely many in all (monopole harmonics). The holomorphic sections are the single row (blue) that is *extreme* in its block, killed by all lowering operators: this happens only at $j = 1$, and that row is V_1, spin one.

Borel–Weil and the Verma module

The algebraic construction of Section 6.3.2 and the geometric one are two views of one object, and the SU(2) case shows how they fit. Work on the big cell $u_2 \neq 0$ and write a holomorphic section as $f = u_2^n F(z)$ with $z = u_1/u_2$, as in (6.156). The operators (6.158) become differential operators in z:

$$\begin{aligned} J^+ f &\overset{(1)}{=} -u_2\,\partial_1\big(u_2^n F\big) = -u_2^n\,F'(z), \\ J^- f &\overset{(2)}{=} -u_1\,\partial_2\big(u_2^n F\big) = -u_1\Big(n\,u_2^{n-1}F - u_2^n F'\,\frac{u_1}{u_2^2}\Big) = u_2^n\big(z^2F' - nzF\big), \\ J^3 f &\overset{(3)}{=} \tfrac12\big(u_2\partial_2 - u_1\partial_1\big)\big(u_2^n F\big) = u_2^n\Big(\frac{n}{2}F - zF'\Big). \end{aligned} \tag{6.177}$$

(1) $\partial_1 z = 1/u_2$ and u_2^n does not depend on u_1.
(2) Product rule; $\partial_2 z = -u_1/u_2^2$; then $u_1/u_2 = z$.
(3) $u_2\partial_2(u_2^n F) = u_2^n(nF - zF')$ and $u_1\partial_1(u_2^n F) = u_2^n zF'$.

So on F: $J^+ = -\frac{\mathrm{d}}{\mathrm{d}z}$, $J^- = z^2\frac{\mathrm{d}}{\mathrm{d}z} - nz$, $J^3 = \frac{n}{2} - z\frac{\mathrm{d}}{\mathrm{d}z}$. (Check: $[-\partial_z,\, z^2\partial_z - nz] = -(2z\partial_z - n) = 2J^3$.) On monomials,

$$J^3 z^p = \Big(\frac{n}{2} - p\Big)z^p, \qquad J^+ z^p = -p\,z^{p-1}, \qquad J^- z^p = (p-n)\,z^{p+1}.$$

Read this as a module for the Lie algebra alone, allowing *every* polynomial F, i.e. every function holomorphic on the big cell:

- $F = 1$ has $J^3 = \frac{n}{2}$ and $J^+1 = 0$: a highest-weight vector of weight $j = \frac{n}{2}$. Applying J^- repeatedly gives $z, z^2, z^3, \ldots$ (up to factors): an infinite ladder, one state at each weight $j, j-1, j-2, \ldots$, which is the weight diagram of the Verma module $V(j)$ of $\mathfrak{sl}(2)$.
- If n is a non-negative integer, $J^- z^n = 0$: the ladder is cut, and $\mathrm{span}\{1, z, \ldots, z^n\}$ is a *submodule*, the irreducible spin j.
- The condition that singles out this submodule is holomorphy at $z = \infty$ (the other chart), which by (6.157) means $\deg F \leq n$.

The polynomial module $\mathbb{C}[z]$ is the **dual** (contragredient) Verma module: it has the same weights as $V(j)$, but the arrows are reversed. In the Verma module the irreducible is a *quotient* $L(j) = V(j)/N$, and the null submodule N is generated by the singular vector $(J^-)^{n+1}|j\rangle$. In $\mathbb{C}[z]$ the irreducible is a *submodule*, and $z^{n+1}, z^{n+2}, \ldots$ survive only in the quotient $\mathbb{C}[z]/\mathrm{span}\{1, \ldots, z^n\}$, whose top state $[z^{n+1}]$ has weight $j - (n+1) = -j - 1$, the weight of the singular vector. Global holomorphy on the compact $\mathbb{CP}^1$ does the job that "remove the null states" does in the Verma module.

Watch out: what happens for λ not dominant: Borel–Weil–Bott

For $n < 0$ there are no holomorphic sections, but the representation does not disappear: it moves into higher cohomology. For $\mathbb{CP}^1$, $H^1(\mathbb{CP}^1; \mathcal{O}(n))$ has dimension $-n-1$ for $n \leq -2$ (by the same Laurent-series argument as Example 6.50, now with the transition function z^n), and it carries spin $\frac{-n-2}{2}$; for $n = -1$ everything vanishes. The general statement is the Borel–Weil–Bott theorem, and the shift by -2 (the "ρ-shift") is the same shift that relates the Verma module $V(j)$ to its singular submodule $V(-j-1)$.

6.3.13 Loop groups: holomorphic sections over LG/T

Nash closes the chapter with one sentence: for the loop group, G/T is replaced by LG/T, and the highest-weight representations of $\widetilde{LG}$ are realised as spaces of holomorphic sections of line bundles over LG/T. We now unpack that sentence. The question is: *can the Fock spaces and Kac–Moody modules of Sections 6.2.9 and 6.3.8 be seen as "lowest Landau levels" on an infinite-dimensional manifold?* They can, and for the free boson the whole construction fits on a page.

The homogeneous space LG/T

Here $T \subset G \subset LG$ is the maximal torus of G, embedded as constant loops. LG/T is the space of loops modulo right multiplication by a constant torus element. Evaluating a loop at $\theta = 0$ gives a map $LG/T \to G/T$, $\gamma T \mapsto \gamma(0)T$, whose fibre is the group ΩG of based loops (Section 6.1.4):

$$\Omega G \longrightarrow LG/T \longrightarrow G/T.$$

So LG/T is the finite-dimensional flag manifold G/T with an infinite-dimensional space of loops attached at each point.

The complex structure. Write $z = e^{i\theta}$. The complexified loop algebra has the Fourier basis Xz^n, $X \in \mathfrak{g}_{\mathbb{C}}$, $n \in \mathbb{Z}$ (Section 6.1.5), and the tangent space at the base point is $L\mathfrak{g}_{\mathbb{C}}/\mathfrak{t}_{\mathbb{C}}$. Split it, as in (6.166), into

$$\mathfrak{p}^+ = \mathfrak{h}^+ \oplus \bigoplus_{n>0} \mathfrak{g}_{\mathbb{C}}\, z^n, \qquad \mathfrak{p}^- = \mathfrak{h}^- \oplus \bigoplus_{n<0} \mathfrak{g}_{\mathbb{C}}\, z^n, \tag{6.178}$$

where for $n \neq 0$ the whole of $\mathfrak{g}_{\mathbb{C}}$ (Cartan part included) appears. The positive directions $\mathfrak{p}^+$ are exactly the operators that annihilate a Kac–Moody highest-weight state (6.126): the positive step operators at mode 0 and every mode T^a_n with $n > 0$. Declare $\mathfrak{p}^+$ holomorphic. The integrability condition of Section 6.3.11 holds for the same reason as before: brackets add mode numbers ($[Xz^m, Yz^n] = [X,Y]z^{m+n}$), negative plus negative is negative, and at mode 0 we are back to $[\mathfrak{h}^-, \mathfrak{h}^-] \subset \mathfrak{h}^-$. So $\mathfrak{b}_{\mathrm{loop}} = \mathfrak{t}_{\mathbb{C}} \oplus \mathfrak{p}^-$ is a subalgebra. Its group B_{loop} consists of the loops $\sum_{n\le 0} b_n z^n$ that extend holomorphically to the outside of the unit disc, including $z = \infty$, with value at ∞ in B. Pressley and Segal show that

$$LG/T = LG_{\mathbb{C}}/B_{\mathrm{loop}},$$

the loop analogue of (6.152), so LG/T is a complex manifold. (Pressley and Segal use the mirror convention, with loops holomorphic *inside* the disc; the two differ by $z \to 1/z$.)

The free boson as a lowest Landau level

Take $G = \mathrm{U}(1)$ and restrict to the identity component of the based loops, $\gamma(\theta) = e^{i\varphi(\theta)}$ with φ real, periodic and of zero mean. (The other components are labelled by the winding number, and T is the constant loops; we do not need them.) Expand

$$\varphi(\theta) = \sum_{m\neq 0} \varphi_m\, e^{im\theta}, \qquad \varphi_{-m} = \bar{\varphi}_m. \tag{6.179}$$

The independent coordinates are the complex numbers $\varphi_1, \varphi_2, \dots$. Which are holomorphic? The complexified loop is e^{ψ} with $\psi = \sum \psi_n z^n$ complex, and B_{loop} consists of the e^{ψ} with only $n \le 0$ modes. The group is abelian, so the class of e^{ψ} in $LG_{\mathbb{C}}/B_{\mathrm{loop}}$ is fixed by the positive modes $\psi_1, \psi_2, \dots$ alone. For a real loop $\psi = i\varphi$, so the holomorphic coordinates are φ_m with $m > 0$, and the $\bar{\varphi}_m$ are antiholomorphic.

The line bundle. Its curvature should be the Kac–Moody cocycle, now read as a real 2-form. For U(1) the cocycle (6.43) on real loops is $\omega_k(\xi,\eta) = \frac{k}{2\pi i}\int_0^{2\pi} \xi'\eta\,\mathrm{d}\theta$, which is imaginary; the real 2-form is $\Omega := i\,\omega_k$. In modes, with $\xi = \sum \xi_m e^{im\theta}$ and $\eta = \sum \eta_m e^{im\theta}$ two tangent vectors (variations of φ),

$$\Omega(\xi,\eta) \overset{(1)}{=} \frac{k}{2\pi}\int_0^{2\pi} \xi'(\theta)\,\eta(\theta)\,\mathrm{d}\theta \overset{(2)}{=} k\sum_{m\neq 0}(im)\,\xi_m\,\eta_{-m}$$
$$\overset{(3)}{=} k\sum_{m>0} i\,m\,\big(\xi_m\bar\eta_m - \bar\xi_m\eta_m\big) \overset{(4)}{=} \Big(i\,k\sum_{m>0} m\;\mathrm{d}\varphi_m\wedge\mathrm{d}\bar\varphi_m\Big)(\xi,\eta). \tag{6.180}$$

(1) $\Omega = i\omega_k$.
(2) $\xi' = \sum im\,\xi_m e^{im\theta}$ and $\frac{1}{2\pi}\int e^{i(m+n)\theta}\mathrm{d}\theta = \delta_{m+n,0}$.
(3) Pair the terms m and $-m$, using $\xi_{-m} = \bar\xi_m$ and $\eta_{-m} = \bar\eta_m$ (real loops).
(4) $(\mathrm{d}\varphi_m\wedge\mathrm{d}\bar\varphi_m)(\xi,\eta) = \xi_m\bar\eta_m - \bar\xi_m\eta_m$.

With $\varphi_m = x_m + iy_m$ we have $i\,\mathrm{d}\varphi_m\wedge\mathrm{d}\bar\varphi_m = 2\,\mathrm{d}x_m\wedge\mathrm{d}y_m$, so $\Omega = \sum_{m>0} 2km\,\mathrm{d}x_m\wedge\mathrm{d}y_m$: *each Fourier mode is a plane threaded by a uniform magnetic field, and the field in the m-th plane is proportional to km.* The space of loops is an infinite stack of Landau problems.

A holomorphic line bundle over $\mathbb{C}^\infty$ with curvature $\Omega = i\partial\bar\partial K$ is described by a Kähler potential K; a holomorphic section is a holomorphic function F, and its norm is $\int |F|^2\,e^{-K}$. (This is the same convention as on $\mathbb{CP}^1$, where $\omega = i\partial\bar\partial\log(1+|z|^2)$ and sections of $\mathcal{O}(n)$ have $\|F\|^2 = \int |F|^2\,(1+|z|^2)^{-n}\omega$.) Here

$$K = k\sum_{m>0} m\;|\varphi_m|^2\,,\qquad \|F\|^2 = \int |F(\varphi_1,\varphi_2,\dots)|^2\prod_{m>0} e^{-km|\varphi_m|^2}\,\frac{km\,\mathrm{d}^2\varphi_m}{\pi}, \tag{6.181}$$

since $i\partial\bar\partial\,(km\,|\varphi_m|^2) = ikm\,\mathrm{d}\varphi_m\wedge\mathrm{d}\bar\varphi_m$, which is (6.180). The measure is normalised so that $\|1\| = 1$.

Key idea: holomorphic sections over $LU(1)/U(1)$ are the Fock space

The normalisable holomorphic sections are the polynomials in $\varphi_1,\varphi_2,\varphi_3,\dots$. With

$$T_{-m} := km\,\varphi_m \text{ (multiplication)},\qquad T_m := \frac{\partial}{\partial\varphi_m},\qquad m>0, \tag{6.182}$$

they form the U(1) Kac–Moody module of Example 6.26: the vacuum is $F=1$, and $T_m^\dagger = T_{-m}$ in the inner product (6.181).

Let us check the three claims, one mode at a time. Write $w = \varphi_m$ and $\beta = km$.

The algebra. On any polynomial F,

$$[T_m,T_{-m}]F \overset{(1)}{=} \partial_w\big(\beta w F\big) - \beta w\,\partial_w F \overset{(2)}{=} \beta F = km\,F,$$

(1) Definitions (6.182).
(2) Product rule.

which is $[T_m,T_n] = km\,\delta_{m+n,0}$. Different modes commute because they involve different variables. And $T_m 1 = \partial_w 1 = 0$: the constant function is the vacuum.

The norms and the adjoint. The Gaussian integral

$$\int_{\mathbb{C}} |w|^{2p}\,e^{-\beta|w|^2}\,\frac{\beta\,\mathrm{d}^2w}{\pi} \overset{(1)}{=} \int_0^\infty r^{2p}e^{-\beta r^2}\,2\beta r\,\mathrm{d}r \overset{(2)}{=} \frac{1}{\beta^p}\int_0^\infty s^p e^{-s}\mathrm{d}s = \frac{p!}{\beta^p} \tag{6.183}$$

(1) Polar coordinates, $\mathrm{d}^2w = r\,\mathrm{d}r\,\mathrm{d}\theta$, and the angular integral gives 2π.
(2) Substitute $s = \beta r^2$.

gives $\|w^p\|^2 = p!/\beta^p$, and distinct powers are orthogonal (the angular integral of $e^{i(p-q)\theta}$). Now test the adjoint on monomials:

$$\langle w^{p+1},\, T_{-m}w^p\rangle = \beta\left\|w^{p+1}\right\|^2 = \frac{(p+1)!}{\beta^p}, \qquad \langle T_m w^{p+1},\, w^p\rangle = (p+1)\,\|w^p\|^2 = \frac{(p+1)!}{\beta^p}.$$

They agree, so $T_{-m}^\dagger = T_m$. Finally $\|(T_{-m})^p 1\|^2 = \beta^{2p}\,p!/\beta^p = p!\,(km)^p$, which is the norm of $(a^\dagger)^p|0\rangle$ for an oscillator with $[a, a^\dagger] = km$. Everything matches Example 6.26. Polynomials are exactly the normalisable holomorphic functions (the Bargmann–Fock theorem), so nothing else appears.

Worked example 6.59: the Weyl operators are magnetic translations

Where does the *group* act, and why is it the central extension? A real loop $e^{i\delta\varphi}$ with $\delta\varphi = ae^{im\theta} + \bar{a}e^{-im\theta}$ shifts the m-th coordinate, $w \to w + a$. On sections it acts by the **Weyl operator**

$$(U_a F)(w) := e^{\beta\bar{a}w - \frac{\beta}{2}|a|^2}\,F(w-a). \tag{6.184}$$

The prefactor is needed for unitarity: substituting $w \to w + a$,

$$|(U_aF)(w+a)|^2\,e^{-\beta|w+a|^2} \overset{(1)}{=} |F(w)|^2\;e^{2\beta\,\mathrm{Re}(\bar{a}(w+a)) - \beta|a|^2 - \beta|w+a|^2} \overset{(2)}{=} |F(w)|^2\;e^{-\beta|w|^2},$$

(1) $\left|e^X\right|^2 = e^{2\,\mathrm{Re}X}$.
(2) $|w+a|^2 = |w|^2 + 2\,\mathrm{Re}(\bar{a}w) + |a|^2$, so the exponent is $2\beta\mathrm{Re}(\bar{a}w) + 2\beta\,|a|^2 - \beta\,|a|^2 - \beta\,|w|^2 - 2\beta\mathrm{Re}(\bar{a}w) - \beta\,|a|^2 = -\beta\,|w|^2$.

so $\|U_aF\| = \|F\|$. Now compose two of them:

$$\begin{aligned}(U_aU_bF)(w) &\overset{(1)}{=} e^{\beta\bar{a}w - \frac{\beta}{2}|a|^2}\,e^{\beta\bar{b}(w-a) - \frac{\beta}{2}|b|^2}\,F(w-a-b)\\ &\overset{(2)}{=} e^{\frac{\beta}{2}(\bar{a}b - a\bar{b})}\,e^{\beta(\bar{a}+\bar{b})w - \frac{\beta}{2}|a+b|^2}\,F(w-a-b) \overset{(3)}{=} e^{i\beta\,\mathrm{Im}(\bar{a}b)}\,(U_{a+b}F)(w).\end{aligned}$$

(1) Apply (6.184) twice: U_b first, evaluated at $w - a$.
(2) $-\beta\bar{b}a - \frac{\beta}{2}|a|^2 - \frac{\beta}{2}|b|^2 = -\frac{\beta}{2}|a+b|^2 + \frac{\beta}{2}(\bar{a}b - a\bar{b})$.
(3) $\bar{a}b - a\bar{b} = 2i\,\mathrm{Im}(\bar{a}b)$, and the definition of U_{a+b}.

The translations of the m-th plane are represented only up to a phase: this is the Heisenberg group of Example 6.19, and the phase is the magnetic flux β times the area spanned by a and b, as for the magnetic translations (6.46). Thus the loop group acts on the holomorphic sections through its central extension $\widehat{L\mathrm{U}(1)}$. The infinitesimal Weyl operators along $a = \epsilon$ and $a = i\epsilon$ are $D_1 = \beta w - \partial_w$ and $D_2 = -i(\beta w + \partial_w)$, and their complex combinations are $\frac{1}{2}(D_1 + iD_2) = \beta w = T_{-m}$ and $\frac{1}{2}(D_1 - iD_2) = -\partial_w = -T_m$: the Kac–Moody generators (6.182), up to the sign conventions for the generators.

Energy and the partition function. The rotation $\theta \to \theta + \alpha$ of the circle acts on loops by $\varphi_m \to e^{-im\alpha}\varphi_m$, and on functions its generator is (up to a factor $-i$) the degree operator

$$L_0 = \sum_{m>0} m\,\varphi_m\frac{\partial}{\partial\varphi_m} = \frac{1}{k}\sum_{m>0} T_{-m}T_m,$$

which is the Sugawara form of L_0 for U(1) ((6.75) with $C = 0$ and $T_0 = 0$). A monomial $\varphi_1^{p_1}\varphi_2^{p_2}\cdots$

has $L_0 = \sum m p_m$, so the number of states at energy N is the number of partitions $p(N)$, and $\mathrm{Tr}\, q^{L_0} = \prod_{m\geq 1}(1-q^m)^{-1}$, the same count as (6.109).

Physical picture: the vacuum wave functional of a free field

In the Schrödinger picture the vacuum of a free field is a Gaussian functional of the field configuration, with width $\propto \omega^{-1/2}$ in each normal mode. Here the vacuum is the section $F = 1$, whose density $|F|^2 e^{-K} = \prod_m e^{-km|\varphi_m|^2}$ is precisely such a Gaussian: the loop $\varphi(\theta)$ is the field on the circle, and the m-th mode, with frequency $\propto m$, is squeezed with width $1/\sqrt{km}$ (Figure 6.33). Borel–Weil for $LU(1)$ is the Schrödinger (holomorphic, or Bargmann) representation of the free chiral boson: excited states are the vacuum Gaussian times polynomials in the positive-frequency modes, exactly like the lowest Landau level of each mode plane.

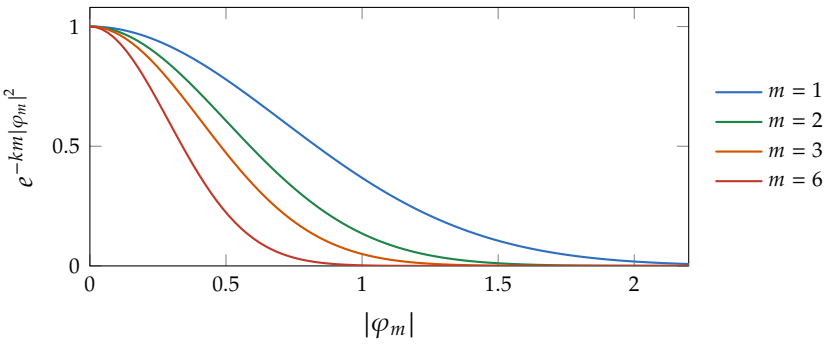


Figure 6.33: The vacuum section of the U(1) loop-group line bundle at level $k = 1$, drawn as the density $e^{-km|\varphi_m|^2}$ in each Fourier-mode plane. Each plane is a Landau problem with field $\propto km$; higher modes feel a stronger field and are squeezed harder. The product over m is the Gaussian vacuum wave functional of the free chiral boson.

Nonabelian G: the level as a Chern class, integrability as positivity

For simple, simply connected G the same picture holds, with two new features.

(1) The line bundles. A homogeneous line bundle over LG/T is now labelled by a weight of G *and* an integer. The weight comes from G/T, as before. The integer is the Chern class along the ΩG fibres: the curvature there is the Kac–Moody 2-form $\Omega = i\omega_k$, and by the integrality condition (6.47) (with $H^2(\Omega G;\mathbb{Z}) = \mathbb{Z}$) its class is the level N. The loop group acts on the bundle only through its central extension: lifting the action of LG on LG/T to the fibres of the line bundle produces phases, exactly as in Example 6.59, and the group of lifts is $\widetilde{LG}$. So the pair (weight, level) labelling a Kac–Moody highest-weight representation (Nash, p. 188) is literally the first Chern class of the bundle.

(2) Which bundles have sections. In the finite case the answer was "λ dominant", i.e. $\langle \lambda, \alpha^\vee\rangle \geq 0$ for the positive roots. For the loop group there is one more $\mathfrak{su}(2)$ to test, the hidden one of (6.140), built from $E_1^{-\alpha}$, E_{-1}^{α} and $\hat{h}_\alpha = \frac{2}{|\alpha|^2}(k - \alpha\cdot H_0)$. It generates a subgroup $\mathrm{SU}(2)_{\hat{\alpha}} \subset \widetilde{LG}$ whose orbit through the base point is a sphere $\mathbb{CP}^1 \subset LG/T$. Restrict the line bundle to that sphere. By Step 3 and (6.157) for this SU(2), the restriction is $\mathcal{O}(d)$, with d the eigenvalue of $\hat{h}_\alpha$ on the highest-weight line:

$$d = \frac{2k}{|\alpha|^2} - \langle \lambda, \alpha^\vee\rangle$$

(this is (6.141)). The highest-weight section does not vanish at the base point (at $g = e$ it equals

$\xi_\lambda(v_\lambda) \neq 0$), so its restriction is a nonzero holomorphic section of $\mathcal{O}(d)$, which requires $d \geq 0$. For the highest root θ this is

$$\langle \lambda, \theta^\vee \rangle \leq N,$$

the integrability condition (6.143), now read as *positivity of a line bundle on a sphere*. For SU(2) at level N the bundles with sections are those with spin $0 \leq j \leq \frac{N}{2}$, the $N+1$ integrable representations of Section 6.3.9. (That these conditions are also sufficient is the Borel–Weil theorem for loop groups, proved by Pressley and Segal.)

	compact G	loop group LG
homogeneous space	$G/T = G_{\mathbb{C}}/B$, compact, $\dim_{\mathbb{C}} = \#\{\alpha > 0\}$	$LG/T = LG_{\mathbb{C}}/B_{\text{loop}}$, infinite-dimensional
holomorphic directions	$\mathfrak{b}^+$ (positive step operators)	$\mathfrak{p}^+$: $\mathfrak{b}^+$ and all modes z^n, $n > 0$
line bundles labelled by	a weight λ (c_1 on G/T)	(λ, N): weight and level (c_1 on ΩG)
group acting on sections	G	the central extension $\widetilde{LG}$
sections exist iff	λ dominant	λ dominant and $\langle \lambda, \theta^\vee \rangle \leq N$ (integrable)
the representation	V_λ, finite-dimensional	integrable highest-weight module, finite-dimensional at each L_0
SU(2) / U(1) example	polynomials of degree n in u_1, u_2 (spin $\frac{n}{2}$)	polynomials in $\varphi_1, \varphi_2, \ldots$ (Fock space)
physics	lowest Landau level on a sphere around a monopole	Gaussian vacuum functional of a chiral field times polynomials

Table 6.1: Borel–Weil for compact groups and for loop groups, side by side.

Reading Nash (pp. 185–191, (6.56)–(6.70))

Nash's §3 begins by noting that the representation theory of infinite-dimensional groups is far from complete: substantial for LG and $\mathrm{Diff}(S^1)$, thin for $\mathrm{Map}(X, G)$ with $\dim X > 1$.

- *Virasoro* (pp. 185–187). The projective representation (6.56) is our (6.102); the highest-weight conditions (6.57) are (6.105), with the ladder argument of Section 6.3.1 supplying Nash's "since L_0 is bounded below". The FQS–GKO classification (6.58) is Section 6.3.6; Nash's $h = \frac{[(m+2)p-(m+1)q]^2-1}{4(m+1)(m+2)}$ is our (6.120). The Verma module, singular vectors, the quotient and the Kac determinant (Sections 6.3.2 to 6.3.5) are background that Nash's statement of (6.58) presupposes. The Goddard–Olive argument (6.59) is Section 6.3.7.

- *Kac–Moody* (pp. 187–188). See the box at the end of Section 6.3.9.

- *Borel–Weil* (pp. 188–191). (6.63) is (6.148), (6.64) is (6.152), (6.65) is (6.155), and the uniqueness of the holomorphic structure via $H^1(\mathcal{O}^*) \cong H^2(\mathbb{Z})$ is Step 5. (6.66)–(6.68) are (6.163), (6.164) and (6.166); (6.69)–(6.70) are (6.173). Nash's closing argument is (6.176). Two small points of translation. Nash says that $\tilde{f}$ is multiplied by $\lambda(Y)$ under "left translation" by $Y \in \mathfrak{t}$; the torus acts on the *right* (it is the fibre action of $G \to G/T$), and with our equivariance convention the factor is $\lambda(t)^{-1}$. And his "lowest weight vector" refers to the right action on V_λ^*; under the left action the same functions carry V_λ with its highest weight λ. The final paragraph on LG/T is Section 6.3.13.

Everything explicit (the SU(2) and SU(3) computations, the Landau level picture, monopole

harmonics, the Verma comparison, Borel–Weil–Bott, and the Fock space over $L\mathrm{U}(1)/\mathrm{U}(1)$) is added.

Recap: Section 6.3

- Positive energy (L_0 bounded below) forces highest-weight representations: $L_0|h\rangle = h|h\rangle$, $L_n|h\rangle = 0$ $(n > 0)$. Unitarity gives $\|L_{-n}|h\rangle\|^2 = 2nh + \frac{c}{12}(n^3 - n) \geq 0$, so $c \geq 0$, $h \geq 0$; a true (non-projective) representation has $c = 0$ and is trivial (Goddard–Olive).
- Pipeline: $|h\rangle \to V(c,h) = U(\mathrm{vir}_-)|h\rangle \to L(c,h) = V(c,h)/N$. V is a module (a representation space), graded by level with $p(N)$ states; N is the submodule generated by singular vectors, the zero-norm states of the Kac form; the quotient is still a module because N is invariant. Generic (c,h): $N = 0$.
- Kac determinant and unitarity: for $c \geq 1$ all $h \geq 0$; for $c < 1$ only $c = 1 - \frac{6}{(m+1)(m+2)}$ with the Kac values $h_{p,q}$ (FQS), all realised by the GKO coset construction.
- Kac–Moody: highest-weight states form an irreducible G-multiplet λ; the hidden $\mathfrak{su}(2)$ built from $E_1^{-\alpha}$ and E_{-1}^{α} forces $N = 2k/l^2 \in \mathbb{Z}_{\geq 0}$ and $\langle \lambda, \theta^\vee \rangle \leq N$ (integrable representations; for SU(2), $j \leq N/2$).
- Borel–Weil: a weight λ gives a line bundle $L_\lambda \to G/T$; $G/T = G_{\mathbb{C}}/B$ is complex with holomorphic directions $\mathfrak{h}^+$; holomorphic sections form V_λ (nonzero iff λ dominant). Proof: Peter–Weyl, $L^2(G) = \bigoplus V_\lambda \otimes V_\lambda^*$, and holomorphy selects the lowest-weight covector. SU(2): polynomials of degree n in u_1, u_2, the lowest Landau level on a sphere with n flux quanta.
- Loop groups: LG/T is complex with holomorphic directions the positive modes; line bundles are labelled by (weight, level); sections carry $\widetilde{LG}$; for U(1) they are the Fock space of the free boson in the Bargmann representation.

Chapter summary

The chapter asked what happens to symmetry when the symmetry group has a parameter at every point: a gauge transformation $g(x)$, a reparametrisation $x \mapsto x'(x)$, a conformal map $z \mapsto w(z)$. The answer came in three layers. First, the groups themselves ($\mathrm{Map}(X,G)$, LG, $\mathrm{Diff}(X)$, the gauge group) are perfectly good infinite-dimensional Lie groups with computable Lie algebras and components. Second, quantum mechanics sees them only projectively, and for these groups the phases cannot be removed: the Lie algebras acquire central terms, the level k of the Kac–Moody algebra and the central charge c of the Virasoro algebra, which appear physically as Schwinger terms and normal-ordering constants. Third, positivity of the energy and of the norm turns these central terms into quantum numbers: the level is an integer, $c < 1$ is quantised, and the representations can be built either algebraically (Verma modules modulo null states) or geometrically (holomorphic sections, Borel–Weil). The tables list every concept, what to do with it, and where it returns.

Section 6.1: the groups

concept	operational meaning	used later
$\mathrm{Map}(X, G)$, Lie algebra $\mathrm{Map}(X, \mathfrak{g})$	multiply maps pointwise; the Lie algebra is $\mathfrak{g}$-valued functions with pointwise bracket	gauge groups (Chapters 8 and 10)
based maps, π_0	$\mathrm{Map} \cong G \times \mathrm{Map}_0$; $\pi_0(\mathrm{Map}_0(S^m, G)) = \pi_m(G)$, e.g. $\pi_3(\mathrm{SU}(n)) = \mathbb{Z}$ counted by the winding integral; general X: $KU^{-1}(X)$	large gauge transformations, θ-vacua (Chapter 8); global anomalies (Chapter 10)
loop group LG	$\mathrm{Map}(S^1, G)$; basis $T^a_n = e^{in\theta}T^a$; connected for simply connected G	WZW models, CFT (Chapter 11)
affine Kac–Moody algebra	$[T^a_m, T^b_n] = if^{abc}T^c_{m+n} + km\delta^{ab}\delta_{m+n,0}$; T^a_0 span $\mathfrak{g}$	currents in CFT (Chapter 11); Chern–Simons, knot invariants (Chapter 12)
$\mathrm{Diff}(X)$, $\mathrm{vect}(X)$	compose maps; the Lie algebra is vector fields, $\alpha_t = \exp(tV)$ the flow, bracket = Lie bracket (up to sign)	reparametrisations of strings (Chapter 9); gravity
Witt algebra	$L_n = -z^{n+1}\partial_z = e^{in\theta}i\partial_\theta$, $[L_m, L_n] = (m-n)L_{m+n}$	conformal symmetry (Chapters 9 and 11)
Virasoro algebra	Witt plus $\frac{c}{12}m(m^2-1)\delta_{m+n,0}$; c = central charge	critical dimension (Chapter 9); minimal models (Chapter 11)
2d conformal group	holomorphic maps; boundary values on S^1; two commuting Witt/Virasoro copies	CFT (Chapter 11)
$\mathrm{Aut}(P)$ and gauge group $\bar{\mathcal{G}}$	$\alpha(pg) = \alpha(p)g$; kernel of $\mathrm{Aut}(P) \to \mathrm{Diff}(X)$; $\mathcal{G} = \Gamma(X, \mathrm{Ad}P)$, $\mathrm{Lie}\,\mathcal{G} = \Gamma(X, \mathrm{ad}P)$; based gauge group acts freely	orbit space $\mathcal{A}/\mathcal{G}$, moduli spaces (Chapter 8); anomalies (Chapter 10)
current algebra	$j^a(\theta) = \sum T^a_n e^{in\theta}$; $[j^a, j^b]$ has the Schwinger term $\propto k\,\delta'(\theta - \phi)$	anomalies (Chapter 10)
analyticity	$\mathrm{Map}(X, G)$ is analytic; $\mathrm{Diff}(X)$ is not, and $\exp : \mathrm{vect} \to \mathrm{Diff}$ is not locally onto	caution in all uses of $\mathrm{Diff}(S^1)$

Section 6.2: extensions

concept	operational meaning	used later
extension $1 \to H \to \bar{G} \to G \to 1$	H normal; split = semidirect product; central if H commutes with all of $\bar{G}$	throughout; $SU(2) \to SO(3)$, Heisenberg, Galilean mass
projective representation	$U(g)U(h) = e^{i\phi(g,h)}U(gh)$; phases form a group 2-cocycle; removable iff a coboundary	spin; anomalies as obstructions (Chapter 10)
Lie algebra 2-cocycle	c antisymmetric with $\sum_{\text{cyc}} c([U,V],W) = 0$; trivial iff $c = \lambda([\,,\,])$; classes $H^2(\mathfrak{g};\mathfrak{h})$	consistency conditions for anomalies (Chapter 10)
Whitehead lemma	semisimple finite-dimensional $\mathfrak{g}$: $H^2 = 0$, so no central charges	why only infinite-dimensional groups have them
Kac–Moody cocycle	$\omega_k(\xi,\eta) = \frac{k}{\pi i}\oint \mathrm{tr}(\xi'\eta)$; cocycle identity is $\oint(\dots)' = 0$	WZW, Chern–Simons (Chapters 11 and 12)
integrality	a group extension by U(1) exists iff $[\omega/2\pi]$ is integral; c_1 of $\widetilde{LG} \to LG$; $N = 2k \in \mathbb{Z}$	quantised levels (Chapter 12)
$\hbar$ expansion	$\{\,,\,\} \to [\,,\,]/i\hbar$; central terms enter at order $\hbar^2$ as c-numbers from reordering	anomalies (Chapter 10)
Virasoro cocycle	Jacobi forces $c(m) = am^3 + bm$; bm is a shift of L_0; Gel'fand–Fuks $\oint u'''v$	strings (Chapter 9)
free fields	boson $c = 1$, Majorana fermion $c = \frac{1}{2}$ via $\|\|L_{-2}\|0\rangle\|\|^2$; Dirac sea gives $k = 1$	$c = 26$ counting (Chapter 9); Ising (Chapter 11)
Sugawara construction	$L_n = \frac{1}{2k+C}\sum_m :T^a_m T^a_{n-m}:$, $c = \frac{2k\dim G}{2k+C}$	WZW, cosets (Chapter 11)
Casimir	$\sum_n T^a_{-n}T^a_n$ diverges by $k\dim G\sum n$; regularised, it is $(2k+C)L_0$	conformal weights (Chapter 11)
extensions of $\mathrm{Map}(X,G)$	central ones pulled back along loops $f: S^1 \to X$; transgression $c = \int_X \mathrm{Ev}^*\omega$; Mickelsson–Faddeev $\propto \int d^{abc}A^c\,\mathrm{d}\alpha^a\,\mathrm{d}\beta^b$	Schwinger terms in 3 + 1 dimensions (Chapter 10)

Section 6.3: representations

concept	operational meaning	used later
positive energy, highest weight	L_0 bounded below $\Rightarrow$ a state with $L_n\|h\rangle = 0$, $n > 0$; labels (c, h)	primary fields (Chapter 11)
Verma module $V(c, h)$	all $L_{-n_1} \cdots L_{-n_r}\|h\rangle$; $p(N)$ states at level N; a module (representation space)	null-vector equations (Chapter 11)
singular vector, null submodule N	$L_1\chi = L_2\chi = 0$; e.g. $(L_{-2} - \frac{3}{2(2h+1)}L_{-1}^2)\|h\rangle$ when $c = \frac{2h(5-8h)}{2h+1}$; N = all its descendants	minimal models (Chapter 11)
irreducible $L(c, h) = V/N$	quotient is a module because N is invariant; generic (c, h): $V = L$	
Kac form and determinant	Gram matrix of level-N states; zeros at $h_{p,q}(c)$; null states have zero norm	unitarity analysis (Chapter 11)
FQS series, GKO	unitary $c < 1$ only at $1 - \frac{6}{(m+1)(m+2)}$, $h = h_{p,q}$; realised as cosets	critical statistical models (Chapter 11)
Goddard–Olive	$c = 0$ unitary $\Rightarrow$ trivial; nontrivial representations are projective	
Kac–Moody representations	ground states form a G-multiplet λ; label (k, λ); $k = 0$ trivial	WZW primaries (Chapter 11)
level, integrability	hidden $\mathfrak{su}(2)$ from $E_1^{-\alpha}, E_{-1}^{\alpha}$: $N = 2k/l^2 \in \mathbb{Z}$, $\langle\lambda, \theta^\vee\rangle \le N$	Verlinde formula, Chern–Simons Hilbert spaces (Chapter 12)
Borel–Weil	weight $\lambda \to L_\lambda \to G/T = G_{\mathbb{C}}/B$; holomorphic sections = V_λ; SU(2): degree-n polynomials	geometric quantisation, Chern–Simons (Chapter 12); Riemann–Roch (Chapter 4)
root complex structure	holomorphic directions $\mathfrak{b}^+$; integrable because positive roots add to positive roots	flag manifolds
Peter–Weyl	$L^2(G) = \bigoplus V_\lambda \otimes V_\lambda^*$; matrix coefficients are the Fourier modes of G	harmonic analysis on groups, lattice gauge theory
LG/T	holomorphic directions = positive modes; line bundles labelled by (weight, level); sections carry $\widetilde{LG}$; U(1): Fock space	conformal blocks (Chapters 11 and 12)

Two threads run through all three tables. The first is the *central term as topology*: the level k is the Chern class of the circle bundle $\widetilde{LG} \to LG$, and the magnetic-translation picture (flux through a plaquette) explains both its integrality and its role as the obstruction to a true representation. The second is *representations as spaces of states built from a vacuum*: either algebraically, by applying lowering operators and removing zero-norm states, or geometrically, as holomorphic wavefunctions in a lowest Landau level. In the chapters on strings, anomalies and conformal and topological field theory the same two ideas reappear as the central charge that must cancel, the Schwinger term that signals an anomaly, and the finite-dimensional spaces of conformal blocks.

Chapter 7

Morse Theory

Where this chapter is going

Every physicist has drawn a potential-energy landscape. There are valleys where a ball comes to rest, peaks where it balances precariously, and mountain passes in between. Now suppose the landscape lives not on a flat plane but on a closed surface such as a sphere or a doughnut. A surprising thing happens: the *shape* of the surface forces the landscape to have a certain minimum number of valleys, passes and peaks, whatever the potential. On a sphere you cannot avoid having at least one peak and one valley. On a doughnut you cannot avoid mountain passes. Turn this around and you get a way to read off the topology of a space by counting the stationary points of one well-chosen function. That two-way street is **Morse theory**.

Physics is full of stationary points: the vacua of a field theory, the classical paths of the action principle, geodesics, instantons, the solutions of the Yang–Mills equations. Morse theory tells us how many there must be, and how their "instability counts" (the number of directions in which they roll downhill) are tied to the topology of the space of all configurations. It is also the mathematics behind the semiclassical approximation: expanding the energy to second order around a stationary point and counting its negative modes is exactly the Hessian and its index, which is where the chapter starts.

The chapter has four sections in Nash. §1 introduces non-degenerate critical points and their index, and the *Morse inequalities* that bound the number of critical points from below by the Betti numbers. §2 allows the critical points to come in whole families (critical submanifolds), which is what symmetric problems always produce. §3 adds a symmetry group G acting on the space and replaces ordinary cohomology by *equivariant cohomology*; this section also looks at the infinite-dimensional case, where the configuration space is a space of paths and the critical points are geodesics. §4 (the next part of these notes) is Witten's reformulation: the Morse inequalities come out of the ground states of a supersymmetric quantum mechanics, and the gradient flow lines between critical points are its instantons. That picture feeds directly into the instanton chapter and into Floer homology in the chapter on topological quantum field theory.

This first part covers Nash's §1–§3. Nash's treatment here is a sequence of statements with very few calculations, so we fill in the geometry he takes for granted: the Hessian and why its index is well defined, the Morse lemma, gradient flows and how a manifold is built by attaching handles, why the inequalities hold, the cohomology of the examples, classifying spaces, and the calculus of variations for geodesics.

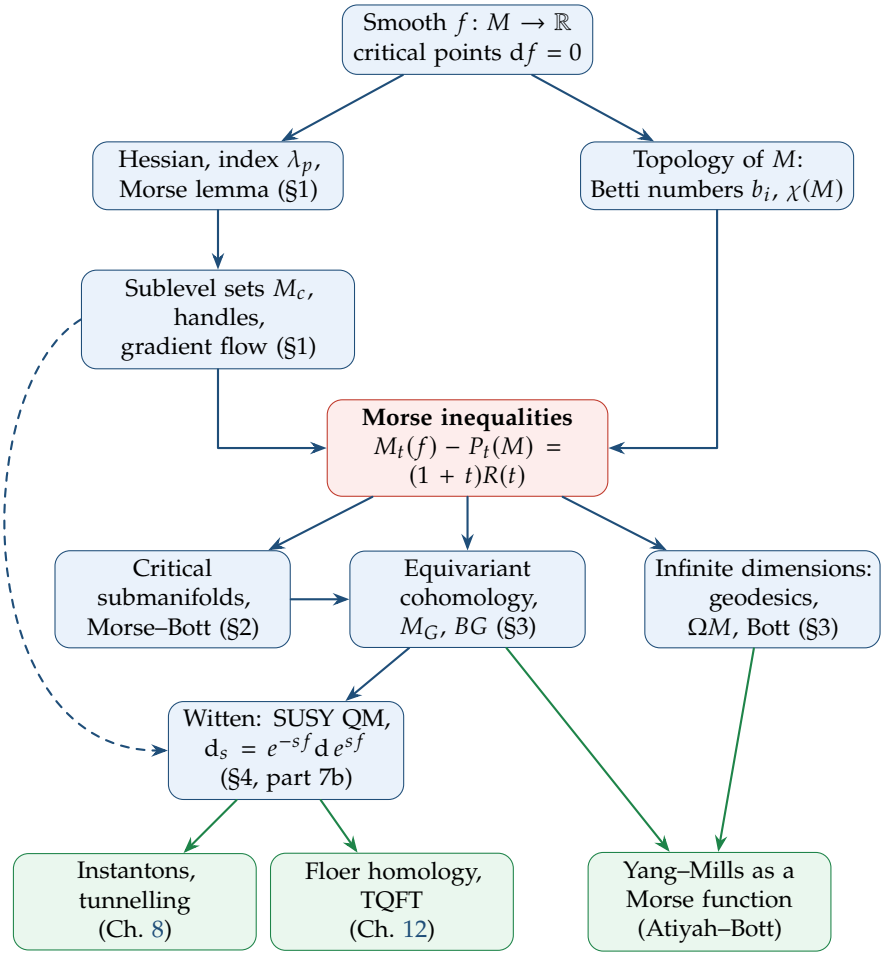


Figure 7.1: The logic of Chapter 7. A single function f on M is examined in two ways: locally, through the Hessian at each critical point (left), and globally, through the cohomology of M (right). The Morse inequalities (red) are where the two meet. The three generalisations in the middle row are all needed for physics: symmetric problems have critical *families*, gauge theories have symmetry groups, and field theories live in infinite dimensions. Witten's supersymmetric picture (dashed arrow: the gradient flow lines become instantons) leads into the physics chapters (green).

7.1 Critical points, the Hessian and the index

Here is the question that starts the subject. We are given a smooth real-valued function on a manifold,

$$f : M \longrightarrow \mathbb{R}, \tag{7.1}$$

where M is compact and (unless we say otherwise) has no boundary [Nash p. 192, (7.1)]. Where is f stationary, and what does f look like near such a point? The answer to the second question turns out to be astonishingly simple: near a "generic" stationary point, f looks like a sum of squares with some plus signs and some minus signs, and the only information is *how many minus signs*. That number, the index, is the atom of Morse theory.

Throughout, $\dim M = n$ and $x^1, \dots, x^n$ are local coordinates. We write $\partial_i = \partial/\partial x^i$.

7.1.1 Critical points

Working definition 7.1 (critical point)

A point $p \in M$ is a **critical point** of f if

$$\mathrm{d}f(p) = 0, \qquad \text{i.e.} \qquad \partial_1 f(p) = \partial_2 f(p) = \cdots = \partial_n f(p) = 0 \tag{7.2}$$

in any (hence every) coordinate chart around p [if all n first partials vanish in one chart, they vanish in every other chart] [Nash p. 193, (7.2)]. The value $f(p)$ is then a **critical value**. *Operational test:* write f in coordinates and solve the n equations $\partial_i f = 0$.

Why "in every chart"? The differential $\mathrm{d}f = \partial_i f \, \mathrm{d}x^i$ is a 1-form, a coordinate-free object (see Chapter T1). In new coordinates y^a its components are

$$\frac{\partial f}{\partial y^a} \overset{(1)}{=} \frac{\partial x^i}{\partial y^a}\,\frac{\partial f}{\partial x^i},$$

(1) Chain rule; the repeated index i is summed.

so the components vanish in one chart exactly when they vanish in the other: the Jacobian matrix $\partial x^i/\partial y^a$ is invertible. Being a critical point is a geometric property of f at p, not of the coordinates.

Physical picture: Equilibria of a potential

If $f = V$ is the potential energy of a particle moving on M, the force is $-\nabla V$. Critical points are exactly the *equilibria*, the points where a particle placed at rest stays at rest. A minimum is a stable equilibrium, a maximum is maximally unstable, and a saddle is stable in some directions and unstable in others. The number of unstable directions is going to be the index.

7.1.2 The Hessian

To see what f looks like near a critical point we Taylor expand to second order. The first-order term vanishes because $\mathrm{d}f(p) = 0$. What is left is

$$f(x) = f(p) + \tfrac{1}{2}\, H_{ij}\, x^i x^j + O(|x|^3), \qquad H_{ij} = \left.\frac{\partial^2 f}{\partial x^i \partial x^j}\right|_p, \tag{7.3}$$

with coordinates centred at p (so p has coordinates $x = 0$).

Working definition 7.2 (Hessian)

The **Hessian** of f at a critical point p is the symmetric $n \times n$ matrix

$$Hf(p) = \left[\partial^2 f/\partial x^i \partial x^j\big|_p\right]_{n\times n} \tag{7.4}$$

[Nash p. 193, (7.3)]. Think of it as a quadratic form [a rule giving one number for each vector v, homogeneous of degree two in v] $v \mapsto H_{ij}v^i v^j$ on the tangent space T_pM [the n-dimensional space of velocity vectors v of curves through p]: it gives the second derivative of f along a curve through p with velocity v.

A matrix of second partial derivatives in one chart does not have to mean anything in another chart. At a critical point it does. Change coordinates $x = x(y)$ with $y = 0$ at p. Then

$$\begin{aligned}\frac{\partial^2 f}{\partial y^a \partial y^b} &\overset{(1)}{=} \frac{\partial}{\partial y^a}\left(\frac{\partial x^j}{\partial y^b}\frac{\partial f}{\partial x^j}\right)\\ &\overset{(2)}{=} \frac{\partial x^i}{\partial y^a}\frac{\partial x^j}{\partial y^b}\frac{\partial^2 f}{\partial x^i \partial x^j} + \frac{\partial^2 x^j}{\partial y^a \partial y^b}\frac{\partial f}{\partial x^j}\\ &\overset{(3)}{=} \frac{\partial x^i}{\partial y^a}\frac{\partial x^j}{\partial y^b}\frac{\partial^2 f}{\partial x^i \partial x^j} \qquad \text{at } p.\end{aligned} \tag{7.5}$$

(1) Chain rule for the first derivative.
(2) Product rule, and the chain rule once more on $\partial f/\partial x^j$.
(3) At a critical point $\partial f/\partial x^j = 0$, so the second term drops out.

In matrix language, with $J = [\partial x^i/\partial y^a]$ the Jacobian,

$$H^{(y)} = J^{\mathsf{T}} H^{(x)} J. \tag{7.6}$$

This is the transformation law of a symmetric bilinear form (a $(0,2)$-tensor), not of a linear map. So at a critical point the Hessian is a genuine geometric object: a quadratic form on T_pM.

Watch out: The Hessian is only tensorial at critical points

Away from a critical point the second term in (7.5) is present, and the matrix of second partials depends on the chart. Example: $f = x$ on $\mathbb{R}$ has zero second derivative, but in the coordinate y with $x = y + y^2$ we get $\mathrm{d}^2 f/\mathrm{d}y^2 = 2$. Nothing is wrong: x has no critical point, so its "Hessian" has no meaning. (With a Riemannian metric one can define a covariant Hessian $\nabla\nabla f$ everywhere; at critical points it agrees with ours.)

7.1.3 Non-degeneracy, index and nullity

The law (7.6) tells us which features of the matrix H are meaningful. Its eigenvalues are *not* (congruence $J^{\mathsf{T}}HJ$ rescales them). What survives is captured by Sylvester's law of inertia.

Key idea: Sylvester's law of inertia

Any real symmetric matrix H can be brought by a congruence $H \mapsto J^{\mathsf{T}}HJ$ (J invertible) to the form $\mathrm{diag}(-1,\dots,-1,+1,\dots,+1,0,\dots,0)$. The numbers of -1's, $+1$'s and 0's do not depend on the choice of J. They equal the numbers of negative, positive and zero eigenvalues of H.

Why is the count of negative signs invariant? It has a basis-free meaning: it is the largest dimension of a subspace on which the quadratic form is negative definite. A subspace on which $H(v, v) < 0$ for all $v \neq 0$ stays such a subspace after any change of basis, so its maximal dimension cannot change.

Working definition 7.3 (non-degenerate, index, nullity)

Let p be a critical point of f with Hessian $H = Hf(p)$.

- p is **non-degenerate** if $\det H \neq 0$ [Nash (7.3)]. Test: compute the determinant of the matrix of second partials in any chart.
- The **index** λ_p is the number of negative eigenvalues of H. Equivalently, it is the dimension of the largest subspace of T_pM on which f decreases to second order [$H_{ij}v^iv^j < 0$ for every nonzero v in the subspace]: the number of independent "downhill directions".
- The **nullity** is $\dim M - \operatorname{rank} H$ [$\operatorname{rank} H$ is the number of independent rows of the matrix H], the number of zero eigenvalues. Non-degenerate means nullity 0.

A function all of whose critical points are non-degenerate is a **Morse function**.

For $n = 2$ there are exactly three kinds of non-degenerate critical point: index 0 (a local minimum), index 1 (a saddle) and index 2 (a local maximum). In general index 0 is a local minimum, index n a local maximum and everything in between is a saddle of some type.

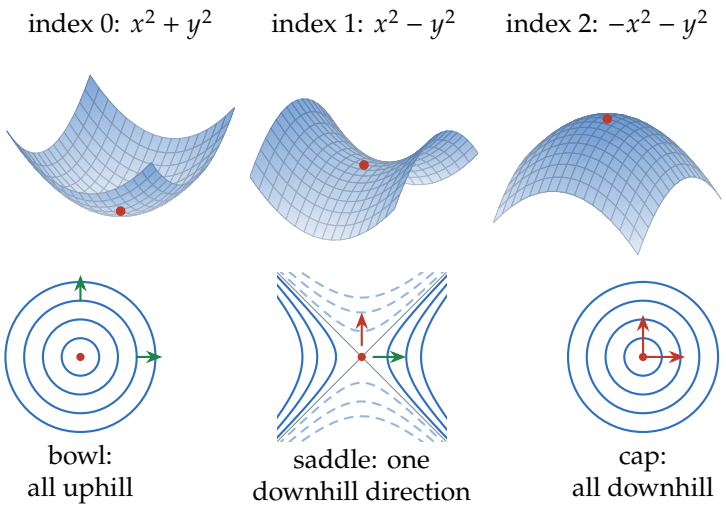


Figure 7.2: The three non-degenerate critical points in two dimensions. Top: graphs of the normal forms. Bottom: their level curves $f = \text{const}$. For the saddle, solid curves are $f > 0$ and dashed curves $f < 0$; the two faint diagonals are the level set $f = 0$ through the critical point. Green arrows point uphill, red arrows downhill. The index is the number of independent red (downhill) directions.

Worked example 7.1: A cubic on the line

Let $f(x) = x^3 - 3x$ on $\mathbb{R}$. Then $f'(x) = 3x^2 - 3$, so the critical points are $x = \pm 1$. The "Hessian" is the 1×1 matrix $f''(x) = 6x$:

$$f''(1) = 6 > 0 \;\Rightarrow\; \text{non-degenerate, index 0 (minimum)},$$
$$f''(-1) = -6 < 0 \;\Rightarrow\; \text{non-degenerate, index 1 (maximum)}.$$

Compare $g(x) = x^3$: the only critical point is $x = 0$, with $g''(0) = 0$. It is degenerate, with nullity 1. It is neither a minimum nor a maximum: an inflection point. We will see in a moment that x^3 is what you get when the two critical points of $x^3 - 3\varepsilon x$ collide as $\varepsilon \to 0$.

Worked example 7.2: Four critical points on the flat torus

Let $T^2 = \mathbb{R}^2/(2\pi\mathbb{Z})^2$, with angles (x, y), and $f(x, y) = \cos x + \cos y$. This is the potential of a particle on a square lattice (a tight-binding band, if you like). Then

$$\partial_x f = -\sin x, \qquad \partial_y f = -\sin y,$$

so critical points need $\sin x = \sin y = 0$: the four points $(x, y) \in \{0, \pi\}^2$. The Hessian is

$$Hf(x, y) \overset{(1)}{=} \begin{pmatrix} \partial_x^2 f & \partial_x\partial_y f \\ \partial_y\partial_x f & \partial_y^2 f \end{pmatrix} \overset{(2)}{=} \begin{pmatrix} -\cos x & 0 \\ 0 & -\cos y \end{pmatrix}.$$

(1) Definition (7.4) with $n = 2$.
(2) $\partial_x^2 \cos x = -\cos x$, and the mixed derivative vanishes because f is a sum of a function of x and a function of y.

Evaluating:

point	Hf	f	index
$(0, 0)$	$\mathrm{diag}(-1, -1)$	2	2 (max)
$(\pi, 0)$	$\mathrm{diag}(+1, -1)$	0	1 (saddle)
$(0, \pi)$	$\mathrm{diag}(-1, +1)$	0	1 (saddle)
(π, π)	$\mathrm{diag}(+1, +1)$	-2	0 (min)

All four are non-degenerate, so f is a Morse function. It has one minimum, two saddles and one maximum. Remember this pattern $(1, 2, 1)$: we will see that the Betti numbers of T^2 are $(b_0, b_1, b_2) = (1, 2, 1)$ and that this is no coincidence.

Worked example 7.3: Degenerate critical points

Three functions on $\mathbb{R}^2$ with a critical point at the origin, all degenerate:

- $f = x^2 + y^3$: $Hf(0) = \mathrm{diag}(2, 0)$, nullity 1. Along the y-axis it behaves like the inflection point y^3.
- $f = x^3 - 3xy^2 = \mathrm{Re}(x + iy)^3$, the **monkey saddle**: all second derivatives vanish at 0, so $Hf(0) = 0$ and the nullity is 2. It has three "valleys" going down (room for a monkey's two legs and tail) instead of the ordinary saddle's two.
- $f = x^2$ (viewed on $\mathbb{R}^2$): $Hf = \mathrm{diag}(2, 0)$ at *every* point of the y-axis. The critical points are not isolated; they fill a whole line. This is the situation of Nash's §2 (Section 7.5), and it is degenerate in a controlled way: non-degenerate in the directions normal to the critical line.

7.1.4 The Morse lemma: the normal form near a critical point

The Taylor expansion (7.3) says f is approximately quadratic near p. The Morse lemma says that for a non-degenerate critical point we can do better: we can choose coordinates in which f is *exactly* quadratic, and after rescaling, exactly a sum of ± squares.

Key idea: The Morse lemma (Nash (7.4))

If p is a non-degenerate critical point of index $\lambda = \lambda_p$, there are coordinates $x^1, \dots, x^n$ centred at p in which

$$f(x) = f(p) \underbrace{- x_1^2 - x_2^2 - \cdots - x_\lambda^2}_{\lambda \text{ terms}} \underbrace{+ x_{\lambda+1}^2 + \cdots + x_n^2}_{n-\lambda \text{ terms}} \tag{7.7}$$

exactly, in a whole neighbourhood of p [Nash p. 193, (7.4)]. (We lower the indices on the coordinates here, as Nash does, to reduce clutter.)

This is a striking statement. It says a non-degenerate critical point has *no local invariants* other than the integer λ: the cubic and higher terms of the Taylor series can all be absorbed by a change of coordinates. The proof is a calculation worth seeing once, because it is just "completing the square" with coefficients that depend on position.

Step 1: f is exactly a quadratic form with variable coefficients. Put p at the origin and subtract the constant so that $f(0) = 0$. For fixed x, consider $\phi(t) = f(tx)$ for $0 \le t \le 1$. Then

$$\begin{aligned} f(x) &\overset{(1)}{=} \phi(1) - \phi(0) \\ &\overset{(2)}{=} \phi'(0) + \int_0^1 (1-t)\,\phi''(t)\,\mathrm{d}t \\ &\overset{(3)}{=} x^i\,\partial_i f(0) + \int_0^1 (1-t)\,x^i x^j\,\partial_i\partial_j f(tx)\,\mathrm{d}t \\ &\overset{(4)}{=} x^i x^j\,h_{ij}(x), \qquad h_{ij}(x) := \int_0^1 (1-t)\,\partial_i\partial_j f(tx)\,\mathrm{d}t. \end{aligned} \tag{7.8}$$

(1) $\phi(1) = f(x)$ and $\phi(0) = f(0) = 0$.

(2) Taylor's theorem with integral remainder, to first order: $\phi(1) = \phi(0) + \phi'(0) + \int_0^1 (1-t)\phi''(t)\,\mathrm{d}t$ (integrate the remainder by parts to check it).

(3) Chain rule: $\phi'(t) = x^i \partial_i f(tx)$ and $\phi''(t) = x^i x^j \partial_i\partial_j f(tx)$.

(4) $\partial_i f(0) = 0$ because 0 is critical; the $x^i x^j$ come out of the t-integral.

The functions $h_{ij}(x)$ are smooth and symmetric, and $h_{ij}(0) = \int_0^1 (1-t)\,\mathrm{d}t\;\partial_i\partial_j f(0) = \frac{1}{2}H_{ij}$. So f is exactly a quadratic form whose coefficient matrix varies smoothly with position and equals half the Hessian at the origin.

Step 2: complete the squares one at a time. Since H is invertible we may (after a linear change of coordinates) assume $h_{11}(0) \neq 0$, hence $h_{11}(x) \neq 0$ near 0. Collect all terms containing x^1:

$$\begin{aligned} x^i x^j h_{ij} &\overset{(1)}{=} h_{11}(x^1)^2 + 2x^1 \sum_{k>1} h_{1k} x^k + \sum_{k,l>1} h_{kl} x^k x^l \\ &\overset{(2)}{=} h_{11}\Big(x^1 + \sum_{k>1} \frac{h_{1k}}{h_{11}} x^k\Big)^2 + \sum_{k,l>1} \Big(h_{kl} - \frac{h_{1k}h_{1l}}{h_{11}}\Big) x^k x^l \\ &\overset{(3)}{=} \pm\,(y^1)^2 + \sum_{k,l>1} \tilde{h}_{kl}(x)\,x^k x^l, \end{aligned}$$

(1) Split the double sum into the $(1,1)$ term, the cross terms (twice, by symmetry) and the rest.

(2) Complete the square. Expanding the square reproduces the first two terms, plus the extra piece $\sum_{k,l>1} h_{1k}h_{1l}\,x^k x^l/h_{11}$, which we then subtract again.
(3) Define the new coordinate

$$y^1 = \sqrt{|h_{11}(x)|}\left(x^1 + \sum_{k>1} \frac{h_{1k}(x)}{h_{11}(x)}\,x^k\right),$$

take the sign $\pm$ to be that of $h_{11}(0)$, and call the remaining coefficients $\tilde{h}_{kl}$.

The map $(x^1,\dots,x^n) \mapsto (y^1, x^2,\dots,x^n)$ has invertible Jacobian at 0 (its $(1,1)$ entry is $\sqrt{|h_{11}(0)|} \neq 0$), so by the inverse function theorem it is a legitimate coordinate change near 0. The leftover form $\tilde{h}_{kl}$ in the $n-1$ variables $x^2,\dots,x^n$ is still non-degenerate at 0 (it is the Schur complement of an invertible matrix). Repeat n times. The result is $f = \pm(y^1)^2 \pm \cdots \pm (y^n)^2$; reorder so the minus signs come first. By Sylvester's law the number of minus signs is the index λ.

Worked example 7.4: Straightening a cubic correction

Let $f(x,y) = x^2 - y^2 + x^3$ near the origin. The Hessian at 0 is $\mathrm{diag}(2,-2)$: non-degenerate, index 1. The Morse lemma promises coordinates with $f = u^2 - v^2$ exactly. Here we can find them by inspection:

$$\begin{aligned} f &\overset{(1)}{=} x^2(1+x) - y^2 \\ &\overset{(2)}{=} \left(x\sqrt{1+x}\right)^2 - y^2 \overset{(3)}{=} u^2 - v^2. \end{aligned}$$

(1) Factor x^2 out of the x-terms.
(2) For $x > -1$ the square root is smooth and positive.
(3) Set $u = x\sqrt{1+x}$ and $v = y$. Since $\partial u/\partial x = 1$ at $x = 0$, this is a valid coordinate change near the origin.

Far from the origin the new coordinates break down (at $x = -1$), and indeed f has a second critical point at $(x,y) = (-\frac{2}{3}, 0)$ which $u^2 - v^2$ knows nothing about. The Morse lemma is a strictly local statement.

Worked example 7.5: Rotating a cross term away

Let $f(x,y) = xy$. The Hessian at 0 is $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, with eigenvalues ± 1: index 1. The coordinates $u = (x+y)/2$, $v = (x-y)/2$ give

$$u^2 - v^2 \overset{(1)}{=} \tfrac{1}{4}\left[(x+y)^2 - (x-y)^2\right] \overset{(2)}{=} xy.$$

(1) Definition of u, v.
(2) $(x+y)^2 - (x-y)^2 = 4xy$.

This is the linear-algebra part of the Morse lemma (diagonalising a constant symmetric matrix) with no higher-order terms to absorb.

7.1.5 Consequences: isolation and finiteness

Two consequences of the normal form (7.7) are used constantly.

Non-degenerate critical points are isolated. In the Morse coordinates, $\partial f/\partial x_i = \pm 2x_i$, which vanishes only at $x = 0$. So a neighbourhood of p contains no other critical point.

A Morse function on a compact manifold has finitely many critical points. Suppose there were

infinitely many, $p_1, p_2, \ldots$. By compactness a subsequence converges to some $q \in M$. By continuity of $\mathrm{d}f$,

$$\mathrm{d}f(q) \overset{(1)}{=} \lim_{k\to\infty} \mathrm{d}f(p_k) \overset{(2)}{=} 0,$$

(1) $\mathrm{d}f$ is continuous and $p_k \to q$ along the subsequence.
(2) Each p_k is critical.

so q is itself a critical point, and every neighbourhood of q contains other critical points p_k. That contradicts isolation. This is the reason Nash gives for the Morse series (7.11) below being a finite sum [Nash p. 193].

7.1.6 Morse functions are everywhere

Is being Morse a strong restriction? No: almost every function is Morse, and any function can be made Morse by an arbitrarily small tilt.

Key idea: Genericity of Morse functions

Let $M \subset \mathbb{R}^N$ be an embedded compact manifold and f any smooth function on it. For almost every vector $a \in \mathbb{R}^N$ (all a outside a set of measure zero) the tilted function

$$f_a(x) = f(x) + a \cdot x$$

is a Morse function. In particular (take $f = 0$) the "height function" $x \mapsto a \cdot x$ in almost every direction a is Morse.

The mechanism is Sard's theorem: the set of critical values of a smooth map has measure zero, and a makes f_a degenerate somewhere exactly when $-a$ is a critical value of a certain smooth map built from $\mathrm{d}f$. At the physicist's level: degeneracy is a coincidence (a determinant vanishing on top of a gradient vanishing), and a random small perturbation destroys coincidences.

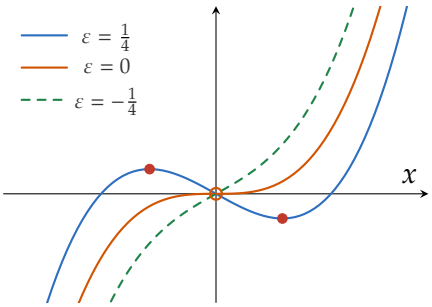


Figure 7.3: The family $f_\varepsilon(x) = x^3 - 3\varepsilon x$. For $\varepsilon > 0$ (blue) there are two non-degenerate critical points (red dots): a maximum of index 1 and a minimum of index 0. As $\varepsilon \to 0$ they approach each other and merge into the degenerate inflection point of x^3 (orange circle). For $\varepsilon < 0$ (green, dashed) they have disappeared. A small perturbation always removes the degenerate case, but it can add or remove critical points *in pairs of adjacent index*. This pair creation and annihilation will explain the factor $(1 + t)$ in the Morse inequalities.

Worked example 7.6: Height functions on the sphere

Let $S^n = \{x \in \mathbb{R}^{n+1} : |x| = 1\}$ and $f_a(x) = a \cdot x$ for a unit vector a. Critical points of the restriction to S^n are points where the gradient of $a \cdot x$ in $\mathbb{R}^{n+1}$, which is the constant vector a, is normal to the sphere, i.e. parallel to x. So $x = \pm a$: the "north and south poles" in

the direction a. Take $a = e_{n+1}$ and use $u = (x_1, \dots, x_n)$ as coordinates near the north pole, where $x_{n+1} = \sqrt{1 - |u|^2}$. Then

$$f \overset{(1)}{=} \sqrt{1 - |u|^2} \overset{(2)}{=} 1 - \tfrac{1}{2}|u|^2 - \tfrac{1}{8}|u|^4 - \dots,$$

(1) On the upper hemisphere the last coordinate is the positive root.
(2) Binomial series $\sqrt{1-s} = 1 - \frac{1}{2}s - \frac{1}{8}s^2 - \dots$ with $s = |u|^2$.

so the Hessian at the north pole is $-\mathbb{1}_n$: non-degenerate, index n (a maximum). Near the south pole $f = -\sqrt{1 - |u|^2}$ and the Hessian is $+\mathbb{1}_n$: index 0 (a minimum). The height function on S^n is a Morse function with exactly two critical points, of index 0 and n.

7.1.7 The height function on an upright torus

This is *the* example of Morse theory, and Nash uses it on p. 196. We set it up explicitly now and use it throughout the part. Take a torus with tube radius r and central-circle radius $R > r$, standing on its end like a tyre on the road, so that its axis of symmetry is horizontal. Coordinates: θ goes around the tube, ϕ around the central circle. The embedding in $\mathbb{R}^3$ is

$$(\theta, \phi) \mapsto \big((R + r\cos\theta)\cos\phi,\ r\sin\theta,\ (R + r\cos\theta)\sin\phi\big), \tag{7.9}$$

with the third coordinate the height. (The central circle lies in the vertical xz-plane, and the axis of symmetry is the horizontal y-axis.) The height function is therefore

$$f(\theta, \phi) = (R + r\cos\theta)\sin\phi. \tag{7.10}$$

Worked example 7.7: Critical points and indices on the upright torus

The first derivatives are

$$\partial_\theta f = -r\sin\theta\,\sin\phi, \qquad \partial_\phi f = (R + r\cos\theta)\cos\phi.$$

Since $R > r$, the factor $R + r\cos\theta$ is never zero, so $\partial_\phi f = 0$ forces $\cos\phi = 0$, i.e. $\phi = \pm\pi/2$. Then $\sin\phi = \pm 1 \neq 0$ and $\partial_\theta f = 0$ forces $\sin\theta = 0$, i.e. $\theta \in \{0, \pi\}$. Four critical points. The second derivatives are

$$\partial_\theta^2 f = -r\cos\theta\sin\phi, \qquad \partial_\phi^2 f = -(R + r\cos\theta)\sin\phi, \qquad \partial_\theta\partial_\phi f = -r\sin\theta\cos\phi,$$

and the mixed derivative vanishes at all four points (where $\sin\theta = 0$). So the Hessians are diagonal:

(θ, ϕ)	where on the tyre	f	Hf	index
$(0, \pi/2)$	top of outer rim	$R + r$	$\mathrm{diag}(-r, -(R + r))$	2
$(\pi, \pi/2)$	top of inner rim	$R - r$	$\mathrm{diag}(+r, -(R - r))$	1
$(\pi, -\pi/2)$	bottom of inner rim	$-(R - r)$	$\mathrm{diag}(-r, +(R - r))$	1
$(0, -\pi/2)$	bottom of outer rim	$-(R + r)$	$\mathrm{diag}(+r, +(R + r))$	0

All four are non-degenerate: the height function on the upright torus is a Morse function with one minimum, two saddles and one maximum, the same $(1, 2, 1)$ pattern as Example 7.2. The two saddles are Nash's "two distinguished points on the inner circle". At the top of the hole, moving along the inner rim (ϕ direction) goes down, while moving around the tube (θ direction, towards the outside) goes up.

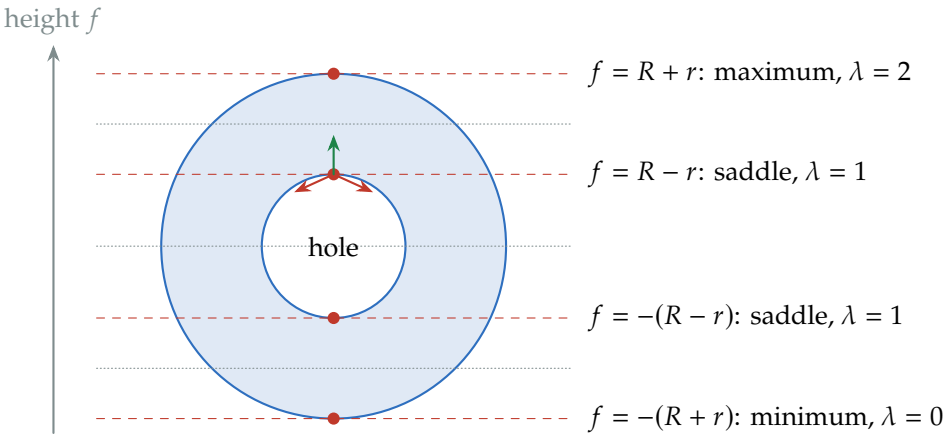


Figure 7.4: The torus standing on its end, seen along its horizontal axis (so it looks like an annulus). The height function has four critical points, all on the vertical line through the centre. Dashed red lines are the four critical levels; dotted lines are ordinary levels. At the upper saddle (top of the hole) the two red arrows, along the inner rim, point downhill; the green arrow, up over the tube, points uphill. One downhill direction: index 1.

Reading Nash (pp. 192–193, (7.1)–(7.4))

Nash states the set-up in a few lines: $f\colon M \to \mathbb{R}$ [Nash (7.1)], critical points $\mathrm{d}f = 0$ [Nash (7.2)], non-degeneracy $\det Hf(p) \neq 0$ [Nash (7.3)], the index λ_p as the number of negative eigenvalues, the normal form [Nash (7.4)] "for suitable coordinates", the nullity $\dim M - \operatorname{rank} Hf(p)$, and Morse functions. What we added: why the Hessian is coordinate-independent at a critical point (7.6), why the index is well defined (Sylvester), a proof of the normal form (the Morse lemma), why non-degenerate critical points are isolated and finitely many, and why Morse functions are generic. Nash's list of applications (infinitely many geodesics on S^n, Bott periodicity, Milnor's exotic 7-spheres, Smale's proof of the Poincaré conjecture in dimensions ≥ 5) we return to in Section 7.8. Milnor's *Morse Theory* is the classic reference Nash recommends.

Recap

- A critical point of f is where $\mathrm{d}f = 0$: an equilibrium of the potential f.
- At a critical point the Hessian $\partial_i \partial_j f$ is a well-defined quadratic form on T_pM (it transforms as $J^{\mathsf{T}} H J$).
- Non-degenerate: $\det H \neq 0$. Index λ_p: the number of negative eigenvalues, i.e. of independent downhill directions.
- Morse lemma: near a non-degenerate critical point, $f = f(p) - x_1^2 - \cdots - x_\lambda^2 + x_{\lambda+1}^2 + \cdots + x_n^2$ exactly.
- Non-degenerate critical points are isolated, so a Morse function on a compact M has finitely many. Morse functions are generic.
- Examples: S^n height (indices $0, n$); torus height and $\cos x + \cos y$ (indices $0, 1, 1, 2$).

7.2 Counting critical points: the Morse inequalities

We now have a local invariant attached to each critical point: its index. The question of this section is global. *Given a compact manifold M, how few critical points of each index can a Morse function on M have?* The answer is controlled by the Betti numbers of M, the numbers of independent "holes" of each dimension. It comes as a list of inequalities, the Morse inequalities, which Nash packages elegantly into a single statement about two polynomials. We first build the two polynomials, then state the inequalities and squeeze every consequence out of them. *Why* they are true is the subject of Section 7.3.

7.2.1 The Morse series: book-keeping for critical points

Let f be a Morse function on a compact M of dimension n. Let

$$m_i := \#\{\text{critical points of } f \text{ of index } i\}, \qquad i = 0, 1, \dots, n.$$

Rather than carry around the list $(m_0, \dots, m_n)$ we put the numbers into a polynomial in a dummy variable t, with the power of t recording the index.

Working definition 7.4 (Morse series)

The **Morse series** (really a polynomial) of f is

$$M_t(f) \;=\; \sum_{\text{critical } p} t^{\lambda_p} \;=\; \sum_{i=0}^{n} m_i\, t^i \tag{7.11}$$

[λ_p is the index of p, its number of independent downhill directions] [Nash p. 193, (7.5)]. *Operational recipe:* find all critical points, compute each index, and write down one monomial t^{λ_p} per critical point.

The two ways of writing (7.11) agree because grouping the monomials t^{λ_p} by their power collects exactly m_i copies of t^i. The sum is finite because a Morse function on a compact manifold has finitely many critical points (Section 7.1.5).

Worked example 7.8: Morse series of the examples so far

Reading off the indices from Section 7.1:

- Height on S^n (Example 7.6): one point of index 0, one of index n, so $M_t = 1 + t^n$.
- $\cos x + \cos y$ on the flat torus (Example 7.2) and the height on the upright torus (Example 7.7): indices $0, 1, 1, 2$, so $M_t = 1 + 2t + t^2$.
- $f(\theta) = \cos\theta$ on the circle S^1: $f' = -\sin\theta$ vanishes at $\theta = 0$ ($f'' = -1$, index 1) and $\theta = \pi$ ($f'' = +1$, index 0), so $M_t = 1 + t$.
- $f(\theta) = \cos 2\theta$ on S^1: critical points at $\theta = 0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}$, with $f'' = -4\cos 2\theta$ equal to $-4, +4, -4, +4$. Two maxima and two minima: $M_t = 2 + 2t$.

The last two examples live on the same manifold but have different Morse series. The Morse series depends on f. What survives every choice of f is what the Morse inequalities are about.

7.2.2 What the topology supplies: Betti numbers and the Poincaré series

Nash writes $P_t(M)$ down in one line. Let us recall what goes into it, since everything in this chapter is measured against it. (For the full story see Chapter 1 and Chapter T2.)

De Rham cohomology in one paragraph. A p-form α is **closed** if $\mathrm{d}\alpha = 0$ and **exact** if $\alpha = \mathrm{d}\beta$ for a globally defined $(p-1)$-form β. Since $\mathrm{d}^2 = 0$, exact forms are closed. The p-th de Rham cohomology is the quotient

$$H^p_{\mathrm{dR}}(M) = \frac{\{\text{closed } p\text{-forms}\}}{\{\text{exact } p\text{-forms}\}},$$

the closed forms that are not derivatives, counted modulo derivatives. By Stokes' theorem an exact form integrates to zero over every closed p-dimensional submanifold, so a closed form with a nonzero integral over some closed p-surface is a nonzero class. For compact M each $H^p_{\mathrm{dR}}(M)$ is finite-dimensional.

Working definition 7.5 (Betti numbers, Poincaré series)

The i-th **Betti number** is $b_i = \dim H^i(M;\mathbb{R})$: the number of independent closed, non-exact i-forms, or equivalently the number of independent i-dimensional "holes" (closed i-surfaces that do not bound [are not the boundary of any $(i+1)$-dimensional region inside M]). The **Poincaré series** (a polynomial for compact M) is

$$P_t(M) = \sum_{i=0}^{n} \dim H^i(M;\mathbb{R})\, t^i = \sum_{i=0}^{n} b_i\, t^i \tag{7.12}$$

[Nash p. 193, (7.6)]. *Operational recipe:* b_0 is the number of connected components; $b_n = 1$ for a connected, closed, orientable M; the rest you compute from a decomposition of M or from the rules below.

Three rules let us compute almost every Poincaré series we need.

- *Poincaré duality.* For M compact, orientable, of dimension n: $b_i = b_{n-i}$. The polynomial P_t is palindromic.
- *Künneth formula.* $P_t(M \times N) = P_t(M)\,P_t(N)$.
- *Euler characteristic.* $\chi(M) = \sum_i (-1)^i b_i = P_{-1}(M)$.

Worked example 7.9: The Betti numbers of the torus, by hand

On $T^2 = \mathbb{R}^2/(2\pi\mathbb{Z})^2$ with angles (x, y), the 1-forms $\mathrm{d}x$ and $\mathrm{d}y$ are globally defined (the angles jump by 2π, their differentials do not) and closed. They are not exact. Let a be the loop $y = 0$, $0 \le x \le 2\pi$, and b the loop $x = 0$. Then

$$\oint_a \mathrm{d}x \overset{(1)}{=} \int_0^{2\pi} \mathrm{d}x = 2\pi \neq 0, \qquad \oint_b \mathrm{d}x \overset{(2)}{=} 0,$$

(1) Along a we can use x itself as parameter.
(2) Along b, x is constant, so $\mathrm{d}x$ restricts to zero.

and similarly $\oint_a \mathrm{d}y = 0$, $\oint_b \mathrm{d}y = 2\pi$. If $\mathrm{d}x = \mathrm{d}g$ for a function g on T^2, Stokes (the fundamental theorem of calculus) would give $\oint_a \mathrm{d}g = 0$. So $[\mathrm{d}x] \neq 0$, and the matrix of periods $\mathrm{diag}(2\pi, 2\pi)$ shows that no combination $c_1\mathrm{d}x + c_2\mathrm{d}y$ with $(c_1, c_2) \neq 0$ is exact either. Hence $b_1 \ge 2$, and in fact $b_1 = 2$. With $b_0 = 1$ (connected) and $b_2 = 1$ (the area form

$\mathrm{d}x \wedge \mathrm{d}y$ has integral $4\pi^2 \neq 0$), we find $P_t(T^2) = 1 + 2t + t^2$. The Künneth formula gives the same answer in one line: $P_t(S^1 \times S^1) = (1 + t)(1 + t)$.

M	$P_t(M)$	$\chi(M) = P_{-1}(M)$
point	1	1
S^1	$1 + t$	0
S^n	$1 + t^n$	$1 + (-1)^n$
$T^2 = S^1 \times S^1$	$1 + 2t + t^2$	0
T^n	$(1 + t)^n$	0
Σ_p (genus p surface)	$1 + 2pt + t^2$	$2 - 2p$
$\mathbb{CP}^n$	$1 + t^2 + t^4 + \cdots + t^{2n}$	$n + 1$
$S^2 \times S^2$	$1 + 2t^2 + t^4$	4
$\mathbb{RP}^2$ (real coefficients)	1	1

The $\mathbb{CP}^n$ line we shall *derive* by Morse theory in Section 7.4. The last line is the first hint that the choice of coefficients matters.

7.2.3 Other coefficient fields, and torsion

Nash immediately generalises to cohomology "with coefficients in a general field $\mathbf{F}$" [Nash p. 194, (7.8)]:

$$P_t(M; \mathbf{F}) = \sum_{i=0}^{n} \dim_{\mathbf{F}} H^i(M; \mathbf{F})\, t^i. \tag{7.13}$$

De Rham theory only produces real coefficients. To get others we describe cohomology combinatorially. Chop M into cells, let C_i be the vector space over $\mathbf{F}$ with one basis vector per i-cell, and let $\partial\colon C_i \to C_{i-1}$ record, with signs and multiplicities, how the boundary of each cell wraps onto the cells one dimension down. Then $\partial^2 = 0$, and $H_i = \operatorname{Ker}\partial/\operatorname{Im}\partial$. (Cohomology is the dual construction; over a field it has the same dimensions.) The only place the field enters is in the *integers* that appear in ∂: a multiplicity 2 is invertible in $\mathbb{R}$ but is zero in $\mathbb{Z}_2$.

Worked example 7.10: The real projective plane with $\mathbb{R}$ and with $\mathbb{Z}_2$ coefficients

$\mathbb{RP}^2$ is the sphere S^2 with antipodal points identified, or equivalently a disc whose boundary circle has antipodal points identified. It has a cell structure with one cell in each dimension: a point e_0, a circle's worth of edge e_1 (half the boundary of the disc; the other half is identified with it) and the open disc e_2. Going around the boundary of the disc we traverse e_1 *twice in the same direction*, and e_1 is a closed loop starting and ending at e_0. So the chain complex is

$$0 \longrightarrow C_2 \xrightarrow{\partial_2 = 2} C_1 \xrightarrow{\partial_1 = 0} C_0 \longrightarrow 0, \qquad C_0 \cong C_1 \cong C_2 \cong \mathbf{F}.$$

Over $\mathbf{F} = \mathbb{R}$: multiplication by 2 is injective, so $\operatorname{Ker}\partial_2 = 0$ and $\operatorname{Im}\partial_2 = C_1$. Hence

$$b_2 \overset{(1)}{=} \dim \operatorname{Ker}\partial_2 = 0, \qquad b_1 \overset{(2)}{=} \dim \operatorname{Ker}\partial_1 - \dim \operatorname{Im}\partial_2 = 1 - 1 = 0, \qquad b_0 = 1.$$

(1) Nothing maps into C_2 from above, so $H_2 = \operatorname{Ker}\partial_2$.
(2) $\partial_1 = 0$ so all of C_1 is closed, but it is all hit by ∂_2.

Over $\mathbf{F} = \mathbb{Z}_2$: now $2 = 0$, so $\partial_2 = 0$ as well, and every C_i contributes: $b_0 = b_1 = b_2 = 1$. So

$$P_t(\mathbb{RP}^2; \mathbb{R}) = 1, \qquad P_t(\mathbb{RP}^2; \mathbb{Z}_2) = 1 + t + t^2. \tag{7.14}$$

Over the integers $H_1(\mathbb{RP}^2; \mathbb{Z}) = \mathbb{Z}/2\mathbb{Z}$: a loop which is not a boundary, but twice which is. This finite piece of the integral cohomology is called **torsion**. Real coefficients are blind to it; $\mathbb{Z}_2$ coefficients see it. Notice that $\chi = 1$ in both cases: the Euler characteristic is the alternating count of cells, $1 - 1 + 1$, whatever the field.

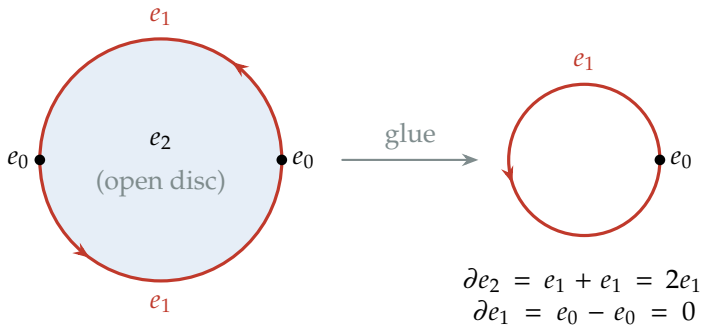


Figure 7.5: $\mathbb{RP}^2$ as a disc with antipodal boundary points identified. The upper and lower red arcs are the same edge e_1, and the arrows show they are glued with the same orientation, so the boundary of the disc runs over e_1 twice: $\partial e_2 = 2e_1$. With real coefficients e_1 is "half a boundary" and dies; with $\mathbb{Z}_2$ coefficients $2e_1 = 0$ and it survives.

7.2.4 The weak Morse inequalities

Here is the fundamental result.

Key idea: The weak Morse inequalities (Nash (7.7)–(7.9))

For any Morse function f on a compact manifold M and any coefficient field $\mathbf{F}$,

$$M_t(f) \;\geq\; P_t(M; \mathbf{F}), \qquad \text{i.e.} \qquad m_i \;\geq\; b_i(\mathbf{F}) \quad \text{for every } i = 0, 1, \ldots, n. \tag{7.15}$$

An inequality between polynomials means: coefficient by coefficient. *A Morse function has at least as many critical points of index i as M has independent i-dimensional holes.*

The second form follows from the first by comparing coefficients of t^i [Nash (7.9)]. The mechanism is the subject of Section 7.3: each critical point of index i attaches one i-dimensional cell to M, and each independent i-dimensional hole must be created by some i-cell. You cannot make a hole of dimension i without an i-cell to span it.

Worked example 7.11: Lower bounds from the weak inequalities

- *Any Morse function on T^2 has at least 4 critical points*, at least one of each index 0 and 2 and at least two saddles, since $P_t(T^2) = 1 + 2t + t^2$. The upright torus height and $\cos x + \cos y$ are optimal.
- *Any Morse function on Σ_p has at least $2p + 2$ critical points*, of which at least $2p$ are saddles.
- *Any Morse function on T^n has at least 2^n critical points*, since the sum of the coefficients

of $(1+t)^n$ is 2^n. The function $\sum_{k=1}^{n} \cos x_k$ has exactly 2^n: the points with every $x_k \in \{0, \pi\}$, and the index is the number of x_k equal to 0.

- *Any Morse function on $\mathbb{RP}^2$ has at least 3 critical points.* Here real coefficients only give $m_0 \geq 1$. Coefficients in $\mathbb{Z}_2$ give $m_0, m_1, m_2 \geq 1$. This is why Nash insists on allowing any field: different fields can give different, stronger bounds.

7.2.5 The strong Morse inequalities

Morse found much more. Look at the difference of the two polynomials.

Key idea: The strong Morse inequalities (Nash (7.10))

For any Morse function f and any field $\mathbf{F}$ there is a polynomial $R(t)$ with *non-negative* integer coefficients such that

$$M_t(f) - P_t(M;\mathbf{F}) = (1+t)\,R(t), \qquad R(t) = \sum_i r_i t^i, \quad r_i \geq 0. \tag{7.16}$$

The surplus critical points (those not forced by topology) always come in pairs of adjacent index, $t^i + t^{i+1} = t^i(1+t)$.

This is the precise form of what Figure 7.3 suggested. When a deformation of f creates critical points, it creates them in pairs of indices $(i, i+1)$; when it destroys them, it destroys them in such pairs. A pair $t^i(1+t)$ can be created or destroyed freely, so it carries no topological information. What cannot be removed is $P_t(M)$. We will see in Section 7.3.8 where the factor $(1+t)$ comes from in the proof: it is the exactness of a long exact sequence.

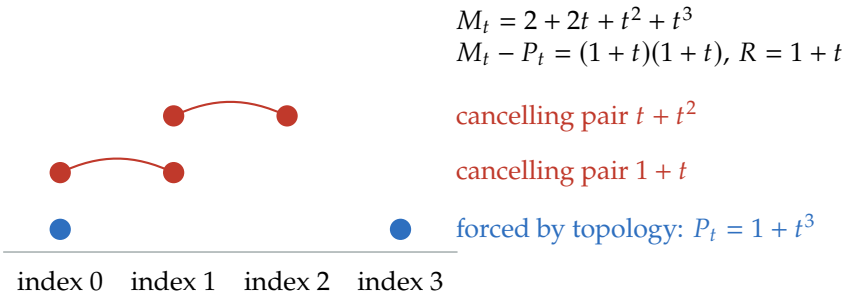


Figure 7.6: The ledger picture of the strong Morse inequalities, for a hypothetical Morse function on S^3 with critical points of indices $0, 0, 1, 1, 2, 3$. Every critical point is a dot in the column of its index. The blue dots are the minimum that the topology demands, $P_t(S^3) = 1 + t^3$. The red surplus can always be paired off, each pair linking *adjacent* columns; each red arc is one factor $t^i(1+t)$ in $(1+t)R(t)$. A surplus that cannot be paired this way is impossible.

Unpacking (7.16) into inequalities. Nash divides by $(1+t)$ and expands the geometric series. Write $\Delta(t) = M_t - P_t$ with coefficients $\Delta_i = m_i - b_i$ (and $\Delta_i = 0$ for $i < 0$ or $i > n$). Then

$$R(t) \overset{(1)}{=} \frac{\Delta(t)}{1+t} \overset{(2)}{=} \Big(\sum_{i=0}^{n} \Delta_i t^i\Big)\Big(\sum_{k=0}^{\infty} (-1)^k t^k\Big)$$

$$\overset{(3)}{=} \sum_{j=0}^{\infty} \Big(\sum_{i=0}^{j} (-1)^{j-i} \Delta_i \Big) t^j. \tag{7.17}$$

(1) Definition (7.16).
(2) $(1+t)^{-1} = 1 - t + t^2 - \dots$ as a formal power series.
(3) Collect the terms with total power $j = i + k$; then $k = j - i$ and i runs from 0 to j.

The coefficient of t^j in R is $r_j \geq 0$. Reading this off:

$$r_j \;=\; \sum_{i=0}^{j} (-1)^{j-i} (m_i - b_i) \;\geq\; 0 \quad\Longleftrightarrow\quad m_j - m_{j-1} + \cdots \pm m_0 \;\geq\; b_j - b_{j-1} + \cdots \pm b_0 \tag{7.18}$$

for every j [Nash p. 194, (7.11)]. Written out for the first few j:

$$\begin{aligned} j = 0: &\quad m_0 \geq b_0, \\ j = 1: &\quad m_1 - m_0 \geq b_1 - b_0, \\ j = 2: &\quad m_2 - m_1 + m_0 \geq b_2 - b_1 + b_0, \\ &\;\;\vdots \\ j = n: &\quad \sum_{i=0}^{n} (-1)^{n-i} m_i = \sum_{i=0}^{n} (-1)^{n-i} b_i. \end{aligned}$$

The last line is an *equality*. Why? $R(t)$ is a polynomial of degree at most $n-1$, because $(1+t)R(t)$ has degree at most n. So $r_n = 0$. The alternating sum in (7.18) for $j = n$ is $\pm$ the Euler characteristic; we return to it in a moment.

Watch out: The strong inequalities imply the weak ones, not conversely

Adding the strong inequalities for j and $j-1$ gives $r_j + r_{j-1} = m_j - b_j \geq 0$: the weak inequality. The strong ones say more, as the next example shows.

Worked example 7.12: Two valleys force a mountain pass

Can a Morse function on S^3 have two minima, two maxima and nothing else? The weak inequalities do not object: $m = (2,0,0,2)$ and $b = (1,0,0,1)$ satisfy $m_i \geq b_i$, and even the Euler characteristic matches, $2 - 0 + 0 - 2 = 0 = \chi(S^3)$. But the strong inequality with $j = 1$ says

$$m_1 - m_0 = 0 - 2 = -2 \;\overset{(1)}{\geq}\; b_1 - b_0 = 0 - 1 = -1,$$

(1) This is (7.18) for $j = 1$; it is violated.

which is false. Equivalently, $M_t - P_t = 1 + t^3 = (1+t)(1 - t + t^2)$, and $R = 1 - t + t^2$ has a negative coefficient. So the function cannot exist. In general, on a connected manifold ($b_0 = 1$, $b_1 = 0$ here), $m_1 \geq m_0 - 1$: k separate valleys need at least $k-1$ passes to connect them. Every hiker knows this.

Worked example 7.13: A non-perfect function on S^2: the two-humped sphere

Deform the round sphere so that it has two peaks, like a camel's back (Figure 7.7). The height function now has two maxima (index 2), one saddle between them (index 1) and

one minimum (index 0):

$$M_t - P_t(S^2) \overset{(1)}{=} (1 + t + 2t^2) - (1 + t^2)$$
$$= t + t^2 \overset{(2)}{=} (1 + t)\, t.$$

(1) Morse series of the humped sphere; Poincaré series of S^2.
(2) Factor out t.

So $R(t) = t$: one cancelling pair of indices $(1, 2)$, namely the extra peak and the saddle. Pushing the extra peak down until it merges with the saddle (a birth–death event run backwards) removes both and gives back the round sphere.

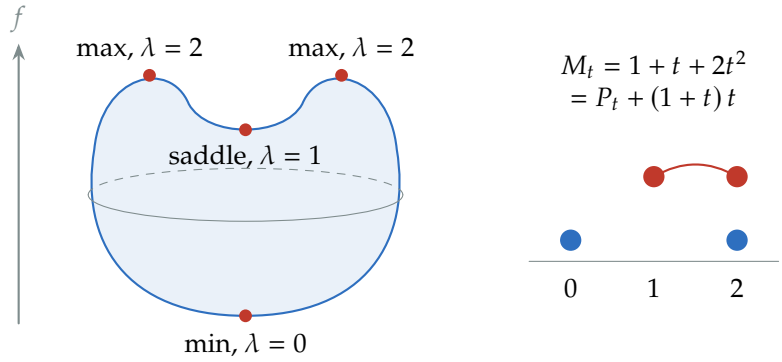


Figure 7.7: A sphere with two humps. Its height function has an extra maximum and an extra saddle compared with the round sphere. Right: the ledger. The blue dots are forced by $P_t(S^2) = 1 + t^2$; the red pair (indices 1 and 2) is the cancelling pair $t(1 + t)$. Nash's example $f = z^2$ on S^2 in Section 7.5 has the same Morse series for a different reason.

7.2.6 Two special values of t: counting and the Euler characteristic

Since $M_t(f)$ and $P_t(M)$ are polynomials, we may evaluate them at numbers. Two values are informative [Nash pp. 194–195, (7.12)–(7.13)].

At $t = 1$ every monomial becomes 1:

$$M_1(f) = \sum_i m_i = \text{total number of critical points of } f. \tag{7.19}$$

This depends strongly on f (2 or 4 on S^1 in Example 7.8), and the weak inequalities bound it below by $P_1(M) = \sum_i b_i$.

At $t = -1$ the factor $(1 + t)$ kills the surplus:

$$M_{-1}(f) \overset{(1)}{=} P_{-1}(M) + (1 + (-1))\, R(-1)$$
$$\overset{(2)}{=} P_{-1}(M) \overset{(3)}{=} \sum_{i=0}^{n} (-1)^i b_i \overset{(4)}{=} \chi(M). \tag{7.20}$$

(1) Set $t = -1$ in the strong inequalities (7.16).
(2) $1 + (-1) = 0$; $R(-1)$ is a finite number because R is a polynomial.
(3) Definition of the Poincaré series at $t = -1$.
(4) Definition of the Euler(–Poincaré) characteristic.

So

$$\sum_{\text{critical } p} (-1)^{\lambda_p} = \chi(M) \qquad \text{for every Morse function } f. \tag{7.21}$$

The left side is computed from f; the right side knows nothing about f. This is the cleanest piece of evidence for the power of the strong inequalities: the weak inequalities alone could never produce an equality.

Worked example 7.14: Checking the Euler count

upright torus: $1-2+1=0=\chi(T^2)$, two-humped S^2: $1-1+2=2=\chi(S^2)$,
$\cos 2\theta$ on S^1: $2-2=0=\chi(S^1)$, height on S^n: $1+(-1)^n=\chi(S^n)$.

The index of a critical point sets the sign $(-1)^{\lambda_p}$, so a saddle on a surface counts -1. A surface of genus p needs $2p$ more saddles than extrema to get $\chi = 2-2p$.

Physical picture: Poincaré–Hopf and Gauss–Bonnet

Equip M with a metric and consider the gradient vector field ∇f. Its zeros are the critical points. Near a critical point of index λ the field is $\nabla f = (-2x_1, \dots, -2x_\lambda, 2x_{\lambda+1}, \dots, 2x_n)$ in Morse coordinates, whose Jacobian has determinant sign $(-1)^\lambda$: this is the *index of the vector field* at that zero. So (7.21) is the Poincaré–Hopf theorem: the sum of the indices of the zeros of any vector field is $\chi(M)$. For surfaces, Gauss–Bonnet gives the same number as $\frac{1}{2\pi}\int_M K\,dA$ (Chapter T2). "You can't comb a hairy ball" is the statement $\chi(S^2) = 2 \neq 0$.

7.2.7 The lacunary principle: when counting is enough

The strong inequalities have a remarkable consequence. If the indices of f never occur in adjacent degrees, there is no room for cancelling pairs, so f must have exactly the minimum number of critical points, and its Morse series *is* the Poincaré series.

Key idea: Morse's lacunary principle (Nash p. 195, (7.14))

If $M_t(f)$ contains no two consecutive powers of t (whenever $m_i \neq 0$, both $m_{i-1} = 0$ and $m_{i+1} = 0$), then for every coefficient field $\mathbf{F}$, $R(t) \equiv 0$ and

$$M_t(f) = P_t(M; \mathbf{F}). \tag{7.22}$$

("Lacunary" means "full of gaps".)

Why. Suppose $R \neq 0$ and let $r_j t^j$ be its *lowest* nonzero term, so $r_j > 0$ and $r_i = 0$ for $i < j$. Look at two coefficients of $M_t = P_t + (1+t)R$:

$$m_j \overset{(1)}{=} b_j + r_j + r_{j-1} \overset{(2)}{=} b_j + r_j \overset{(3)}{>} 0,$$
$$m_{j+1} \overset{(4)}{=} b_{j+1} + r_{j+1} + r_j \overset{(5)}{\geq} r_j > 0.$$

(1) The coefficient of t^j in $(1+t)R(t) = \sum_i r_i(t^i + t^{i+1})$ is $r_j + r_{j-1}$.
(2) $r_{j-1} = 0$ because r_j is the lowest nonzero coefficient.
(3) $b_j \geq 0$ and $r_j > 0$.
(4) The coefficient of t^{j+1} in $(1+t)R$ is $r_{j+1} + r_j$.

(5) $b_{j+1} \geq 0$ and $r_{j+1} \geq 0$.

So t^j and t^{j+1} both occur in $M_t(f)$: two consecutive powers, contradicting the assumption. Hence $R = 0$. The argument used nothing about $\mathbf{F}$ except the strong inequalities, so it holds for every field. (Nash phrases the same argument using the first nonzero power of M_t; the essential point in both versions is that $(1+t)R$ always contains a pair of neighbouring powers.)

Worked example 7.15: Recovering the cohomology of spheres

The height function on S^n has $M_t = 1 + t^n$. For $n \geq 2$ the powers 0 and n are not consecutive, so by the lacunary principle

$$P_t(S^n; \mathbf{F}) = 1 + t^n \qquad \text{for every field } \mathbf{F}.$$

We have computed the cohomology of every sphere $S^{n\geq 2}$ by finding the critical points of one function. For S^1 the powers 0 and 1 are adjacent and the principle is silent. (The answer $1 + t$ is still right; the lacunary principle is sufficient, not necessary.) The torus is another example where it is silent: $1 + 2t + t^2$ has every power, yet the upright height function turns out to be perfect.

7.2.8 Perfect Morse functions and torsion

Working definition 7.6 (perfect Morse function)

A Morse function is **perfect** for the field $\mathbf{F}$ [the numbers used as coefficients in the cohomology, such as $\mathbb{R}$ or $\mathbb{Z}_q$] if $M_t(f) = P_t(M; \mathbf{F})$, i.e. $R \equiv 0$ [$R(t)$ is the polynomial with nonnegative coefficients in the Morse inequalities $M_t - P_t = (1+t)R(t)$]: it has exactly as many critical points of each index as the topology requires. It is *perfect* (with no qualifier) if this holds for every field [Nash p. 195]. *Test:* compare the Morse series with the Poincaré series; if they agree over $\mathbb{R}$ and over every $\mathbb{Z}_q$ (q prime), f is perfect.

Perfect Morse functions are the most economical landscapes a space allows. The lacunary principle says that a function with gaps in its indices is automatically perfect. Nash draws a sharp consequence. Perfection for all fields forces the dimensions $\dim H^i(M; \mathbf{F})$ to be independent of $\mathbf{F}$, and that happens exactly when the integral cohomology $H^*(M; \mathbb{Z})$ has no torsion (torsion like $\mathbb{Z}/q\mathbb{Z}$ is invisible over $\mathbb{R}$ and visible over $\mathbb{Z}_q$, as in Example 7.10). So:

Key idea: Torsion obstructs perfection

If f is perfect for every field then $H^*(M; \mathbb{Z})$ is torsion-free. On a manifold whose integral cohomology has torsion, such as $\mathbb{RP}^n$ ($n \geq 2$), lens spaces or the Klein bottle, *no* Morse function is perfect for every field.

Worked example 7.16: The minimal Morse function on $\mathbb{RP}^2$

On $S^2 \subset \mathbb{R}^3$ take $f = ax^2 + by^2 + cz^2$ with $0 < a < b < c$. Since $f(-x) = f(x)$ it descends to $\mathbb{RP}^2$. Critical points on S^2: points where the gradient $2(ax, by, cz)$ in $\mathbb{R}^3$ is parallel to the position vector (x, y, z). Because a, b, c are distinct this forces two of the coordinates to vanish: the six points $\pm e_1, \pm e_2, \pm e_3$. Near $e_1 = (1, 0, 0)$ use y, z as coordinates, with $x^2 = 1 - y^2 - z^2$:

$$f \overset{(1)}{=} a(1 - y^2 - z^2) + by^2 + cz^2$$

$$\overset{(2)}{=} a + (b-a)\,y^2 + (c-a)\,z^2,$$

(1) Substitute $x^2 = 1 - y^2 - z^2$ on the sphere.
(2) Collect terms.

which is *exactly* in Morse normal form: both coefficients positive, so index 0. The same computation near $\pm e_2$ gives $b + (a-b)x^2 + (c-b)z^2$, index 1, and near $\pm e_3$ gives $c + (a-c)x^2 + (b-c)y^2$, index 2. On S^2:

$$M_t^{S^2} - P_t(S^2) \overset{(1)}{=} (2 + 2t + 2t^2) - (1 + t^2) = 1 + 2t + t^2 \overset{(2)}{=} (1+t)(1+t).$$

(1) Two critical points of each index; $P_t(S^2) = 1 + t^2$.
(2) The surplus factorises with $R = 1 + t \geq 0$, as it must.

On $\mathbb{RP}^2$ the antipodal pairs become single points, so $M_t^{\mathbb{RP}^2} = 1 + t + t^2$. Comparing with (7.14):

$$\begin{aligned} \text{over } \mathbb{Z}_2: &\quad M_t - P_t = 0 \quad \text{(perfect)}, \\ \text{over } \mathbb{R}: &\quad M_t - P_t = t + t^2 = (1+t)\,t \quad \text{(not perfect)}. \end{aligned}$$

Three is the minimum number of critical points on $\mathbb{RP}^2$ (the $\mathbb{Z}_2$ bound), and the function achieving it is perfect over $\mathbb{Z}_2$ but not over $\mathbb{R}$. Keep this function in mind: its complex version is exactly Nash's function on $\mathbb{CP}^n$ in Section 7.4.

Reading Nash (pp. 193–195, (7.5)–(7.14))

Nash defines the Morse series [Nash (7.5)] and Poincaré series [Nash (7.6)], states $M_t(f) \geq P_t(M)$ as "the fundamental result of Morse theory" [Nash (7.7)], generalises to any field [Nash (7.8)], extracts $m_i \geq b_i$ [Nash (7.9)], and gives Morse's strong form $M_t - P_t = (1+t)R(t)$, $r_i \geq 0$ [Nash (7.10)], unpacking it by expanding $(1+t)^{-1}$ [Nash (7.11)]. Then $t = 1$ counts critical points [Nash (7.12)] and $t = -1$ gives $\chi(M)$ [Nash (7.13)]. The lacunary principle is stated as a theorem with a short proof [Nash (7.14)], followed by the definition of perfect Morse functions and the torsion remark. He gives no example in this stretch; we added the S^1, S^3, two-humped S^2 and $\mathbb{RP}^2$ examples, the cellular computation of $\mathbb{RP}^2$ with two coefficient fields, and the Poincaré–Hopf reading of (7.21). Nash does not prove the inequalities; he refers to Nash and Sen. We give the mechanism in Section 7.3.8.

Recap

- Morse series $M_t(f) = \sum_p t^{\lambda_p} = \sum m_i t^i$ counts critical points by index; Poincaré series $P_t(M;\mathbf{F}) = \sum b_i t^i$ counts holes by dimension.
- Weak inequalities: $m_i \geq b_i$. Strong inequalities: $M_t - P_t = (1+t)R(t)$ with $R \geq 0$, equivalently the alternating sums (7.18).
- Surplus critical points come in cancelling pairs of adjacent index.
- $t = -1$: $\sum_p (-1)^{\lambda_p} = \chi(M)$ for every Morse f (Poincaré–Hopf).
- Lacunary principle: no adjacent indices $\Rightarrow M_t = P_t$ for every field. Spheres: $P_t(S^n) = 1 + t^n$.
- Perfect for every field $\Rightarrow$ no torsion. $\mathbb{RP}^2$: three critical points, perfect over $\mathbb{Z}_2$ only.

7.3 How the topology changes at a critical point: sublevel sets, handles and gradient flow

The inequalities of Section 7.2 relate a local count (critical points and their indices) to a global count (Betti numbers). There has to be a mechanism connecting them. Here it is. Imagine flooding the manifold: pour water into the landscape f and watch the region under water, $\{f \le c\}$, as the water level c rises. Nothing interesting happens to the *shape* of the flooded region until the water reaches a critical point. There it changes in a completely standard way: a cell of dimension λ is glued on, where λ is the index. Building M out of cells, one per critical point, is what forces the Morse inequalities.

7.3.1 Level sets and sublevel sets

Working definition 7.7 (level and sublevel sets)

For a real number c, the **level set** of f at c is $f^{-1}(c) = \{x \in M : f(x) = c\}$ and the **sublevel set** (Nash calls it a "half space") is

$$M_c = \{x \in M : f(x) \le c\} \tag{7.23}$$

[Nash p. 195, (7.15)]. If c is not a critical value [$c \neq f(p)$ for every critical point p], $f^{-1}(c)$ is a smooth hypersurface [a submanifold of dimension $n-1$] in M and it is the boundary of M_c.

Why is a level set smooth at a non-critical value? Near a point x with $f(x) = c$ and $\mathrm{d}f(x) \neq 0$, some $\partial_i f(x) \neq 0$, and the implicit function theorem lets us solve $f = c$ for x^i as a smooth function of the other $n-1$ coordinates. So $f^{-1}(c)$ is locally a graph, an $(n-1)$-dimensional submanifold. At a critical point this fails: the level set through a saddle of a surface is a figure eight, with a crossing at the saddle.

The level sets are the higher-dimensional analogue of the contour lines on a map. On the upright torus (Figure 7.4), a level between the minimum and the lower saddle is one small circle around the bottom of the tube; between the two saddles it is two circles, one on each side of the hole; above the upper saddle it is again one circle; and at each saddle it is a figure eight.

7.3.2 Gradient flow

To move points from one level to another we need a direction field. Choose a Riemannian metric g_{ij} on M (any one will do). The **gradient** of f is the vector field

$$(\nabla f)^i = g^{ij}\,\partial_j f, \tag{7.24}$$

the vector "dual" to the 1-form $\mathrm{d}f$. It vanishes exactly at the critical points. The **gradient flow** is the system of ODEs

$$\frac{\mathrm{d}x^i}{\mathrm{d}t} = -(\nabla f)^i(x(t)), \tag{7.25}$$

"roll downhill as steeply as possible" (overdamped motion in the potential f; physically, a particle in a very viscous fluid). Along a solution,

$$\frac{\mathrm{d}}{\mathrm{d}t} f(x(t)) \overset{(1)}{=} \partial_i f\,\frac{\mathrm{d}x^i}{\mathrm{d}t} \overset{(2)}{=} -\partial_i f\, g^{ij} \partial_j f \overset{(3)}{=} -\,|\nabla f|^2 \;\le\; 0. \tag{7.26}$$

(1) Chain rule.

(2) The flow equation (7.25) and the definition (7.24).
(3) $g^{ij}\partial_i f \partial_j f = g_{ij}(\nabla f)^i(\nabla f)^j$ is the squared length of ∇f.

So f strictly decreases along every flow line except at the critical points, where the flow is stationary. On a compact M every flow line therefore comes from a critical point as $t \to -\infty$ and goes to a critical point as $t \to +\infty$. (It cannot decrease f forever, since f is bounded, and it can only slow down where ∇f is small, i.e. near critical points.)

Worked example 7.17: Gradient flow near a critical point of index λ

In Morse coordinates, with the flat metric $g_{ij} = \delta_{ij}$ (we may choose the metric this way near p), $f = f(p) - \sum_{i\le\lambda} x_i^2 + \sum_{i>\lambda} x_i^2$. The flow (7.25) is

$$\dot{x}_i \overset{(1)}{=} -\partial_i f = +2x_i \quad (i \le \lambda), \qquad \dot{x}_i \overset{(2)}{=} -\partial_i f = -2x_i \quad (i > \lambda),$$

(1) $\partial_i(-x_i^2) = -2x_i$ for the downhill coordinates.
(2) $\partial_i(+x_i^2) = +2x_i$ for the uphill coordinates.

with solution $x_i(t) = x_i(0)\,e^{2t}$ for $i \le \lambda$ and $x_i(t) = x_i(0)\,e^{-2t}$ for $i > \lambda$. The points that flow *into* p as $t \to +\infty$ are those with $x_1 = \cdots = x_\lambda = 0$: an $(n-\lambda)$-dimensional disc, the **stable** (ascending) manifold. The points that flow *out of* p (reach p as $t \to -\infty$) have $x_{\lambda+1} = \cdots = x_n = 0$: a λ-dimensional disc, the **unstable** (descending) manifold $W^-(p)$. *The index is the dimension of the set of points that roll away from p.* This is the linear stability analysis of an equilibrium, exactly as in mechanics: λ unstable modes, $n-\lambda$ stable ones.

7.3.3 Nothing happens between critical values

Key idea: Bott–Morse–Smale theorem, part 1 (Nash p. 196)

If $f^{-1}[a,b]$ is compact and contains no critical point, then M_a and M_b are diffeomorphic. In fact M_a is a deformation retract of M_b: M_b can be continuously squeezed onto M_a.

The proof is a flow. On the region $a \le f \le b$ the gradient is nonzero, so we may rescale it to the vector field

$$X = -\frac{\nabla f}{|\nabla f|^2}$$

(cut off smoothly to zero away from $f^{-1}[a,b]$). Along a flow line of X,

$$\frac{\mathrm{d}}{\mathrm{d}s} f(x(s)) \overset{(1)}{=} \partial_i f\, X^i \overset{(2)}{=} -\frac{\partial_i f\, g^{ij}\partial_j f}{|\nabla f|^2} \overset{(3)}{=} -1.$$

(1) Chain rule.
(2) Definition of X and of the gradient.
(3) The numerator is $|\nabla f|^2$.

So flowing along X for a time s lowers f by exactly s. Flowing for time $b-a$ carries the level $f^{-1}(b)$ onto $f^{-1}(a)$ and the whole of M_b into M_a, and the flow for time $-(b-a)$ is the inverse map. Hence $M_b \cong M_a$. Morally: if there are no critical points in the slab, the slab is just a product $f^{-1}(a) \times [a,b]$, a collar that can be absorbed.

7.3.4 Passing a critical point attaches a cell

Key idea: Bott–Morse–Smale theorem, part 2 (Nash p. 196)

If (a, b) contains exactly one critical value, belonging to a single non-degenerate critical point p of index λ, then

$$M_b \simeq M_a \cup e_\lambda, \tag{7.27}$$

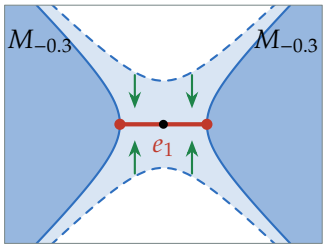
where e_λ is a λ-dimensional cell (a closed disc D^λ) glued to M_a along its boundary sphere $S^{\lambda-1}$, and $\simeq$ means "has the same homotopy type as" (can be deformed into, allowing thickness to be squashed).

We can see this in the local model. Shift so that $f(p) = 0$ and take $a = -\varepsilon$, $b = +\varepsilon$. Split the Morse coordinates as $u = (x_1, \dots, x_\lambda)$ (downhill) and $v = (x_{\lambda+1}, \dots, x_n)$ (uphill), so near p

$$f = -|u|^2 + |v|^2 .$$

- $M_{-\varepsilon}$ near p is $\{|u|^2 \geq |v|^2 + \varepsilon\}$: points with a large downhill displacement. It does not contain p.
- $M_{+\varepsilon}$ near p additionally contains the region $\{-\varepsilon < -|u|^2 + |v|^2 \leq \varepsilon\}$, which includes p.
- The **core** of the new region is the λ-disc $e_\lambda = \{v = 0,\ |u|^2 \leq \varepsilon\}$, a piece of the unstable manifold of p. Its boundary $\{v = 0,\ |u|^2 = \varepsilon\}$ is a sphere $S^{\lambda-1}$ that lies on the level set $f = -\varepsilon$, i.e. on the boundary of $M_{-\varepsilon}$.

The rest of the new region retracts onto the core: slide each point along the v directions towards $v = 0$ (this lowers f, since $|v|^2$ decreases) until it either reaches the core or hits $M_{-\varepsilon}$. So up to homotopy, $M_{+\varepsilon}$ is $M_{-\varepsilon}$ with a λ-disc glued on along its boundary sphere. The same squeeze, done without collapsing the v directions, shows that $M_{+\varepsilon}$ is *diffeomorphic* to $M_{-\varepsilon}$ with a *handle* attached, as follows.



solid: $f = -0.3$
dashed: $f = +0.3$
dark: $M_{-0.3}$
light: added to make $M_{+0.3}$
red: the core cell e_1

u (downhill) horizontal, v (uphill) vertical

Figure 7.8: Crossing a saddle ($n = 2$, $\lambda = 1$, $f = -u^2 + v^2$). Below the critical level, the sublevel set (dark) near p is two separate pieces. Just above it (dark plus light) they are joined through a neck. The green arrows squeeze the neck vertically onto the red segment e_1, whose two endpoints (S^0) lie on the old boundary. So up to homotopy, crossing the saddle glues a 1-cell onto $M_{-\varepsilon}$ along its two ends.

7.3.5 Handles

Working definition 7.8 (λ-handle)

In dimension n, a **λ-handle** is a copy of $D^\lambda \times D^{n-\lambda}$ (a thickened λ-disc) [D^k is the solid ball of dimension k, with boundary sphere S^{k-1}] glued to the boundary of a manifold along $S^{\lambda-1} \times D^{n-\lambda}$ (a thickened copy of its boundary sphere). Its **core** is $D^\lambda \times \{0\}$. Crossing a critical point of index λ attaches exactly one λ-handle: $M_{+\varepsilon} \cong M_{-\varepsilon} \cup (D^\lambda \times D^{n-\lambda})$ [$M_{\pm\varepsilon}$ are the sublevel sets just above and just below the critical value $f(p) = 0$. The symbol $\cong$ means diffeomorphic]. Shrinking the thickening gives (7.27).

In dimension 2 there are three kinds, pictured in Figure 7.9:

- a *0-handle* $D^0 \times D^2$ is a disc glued along $S^{-1} \times D^2 = \varnothing$, i.e. not glued at all: a new component is born (a new lake appears at a minimum);
- a *1-handle* $D^1 \times D^1$ is a strip glued along its two ends $S^0 \times D^1$: a bridge (two lakes merge at a pass, or one lake wraps around and meets itself);
- a *2-handle* $D^2 \times D^0$ is a disc glued along its whole boundary circle: a cap (a hole in the flooded region is plugged at a peak).

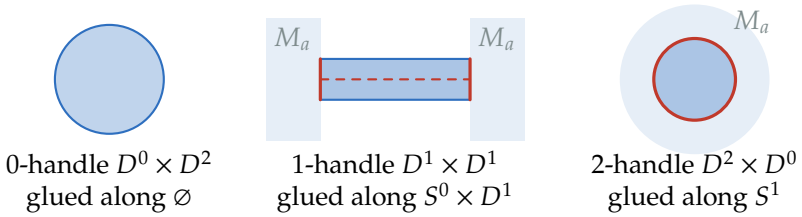


Figure 7.9: The three kinds of handle on a surface. Dark blue: the new handle; light blue: the old sublevel set M_a; red: the attaching region $S^{\lambda-1} \times D^{n-\lambda}$; dashed red: the core of the 1-handle, which is the cell e_1. A minimum starts a new piece, a saddle builds a bridge, a maximum puts on a lid.

7.3.6 The torus, built one critical point at a time

Now we watch the upright torus of Example 7.7 fill with water. The four critical values are $-(R+r) < -(R-r) < R-r < R+r$.

Worked example 7.18: Sublevel sets of the upright torus

1. $c < -(R+r)$: $M_c = \varnothing$.
2. $-(R+r) < c < -(R-r)$: we have passed the minimum (index 0). M_c is a small cap at the bottom of the tyre: a *disc*. A 0-handle.
3. $-(R-r) < c < R-r$: we have passed the lower saddle (index 1). The water now climbs both sides of the hole; M_c is the bottom half of the tyre, a U-shaped tube: a *cylinder*, i.e. a disc with a strip attached. Up to homotopy it is a circle: $e_0 \cup e_1$.
4. $R-r < c < R+r$: we have passed the upper saddle (index 1). The two arms of the U have joined over the top of the hole. M_c is the torus with a small disc around the top removed: a *punctured torus*. Up to homotopy it is a figure eight of two circles, $e_0 \cup e_1 \cup e_1$.

5. $c > R + r$: we have passed the maximum (index 2), which caps off the remaining hole. $M_c = T^2 = e_0 \cup e_1 \cup e_1 \cup e_2$.

One cell per critical point, with dimension equal to the index.

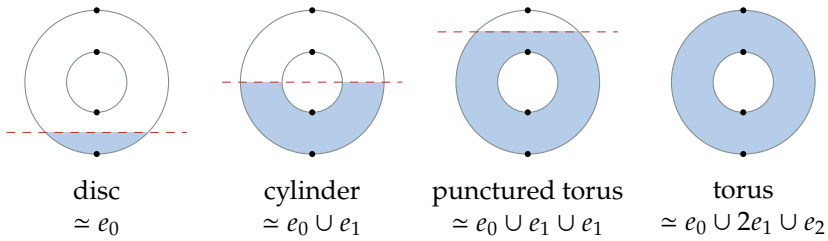


Figure 7.10: Flooding the upright torus (seen along its axis, as in Figure 7.4). Shaded: the sublevel set M_c, for c just above the minimum, between the saddles, above the upper saddle, and above the maximum. The red dashed line is the water level. Each time the water passes a critical point (black dot) one cell is added, of dimension equal to the index: first a disc, then a bridge around the left or right of the hole giving a cylinder, then a second bridge over the top giving a punctured torus, then a lid.

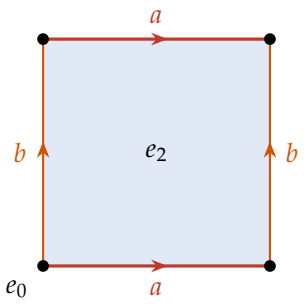
Reading off the Morse series and the Euler characteristic [Nash p. 196, (7.17)–(7.18)]:

$$M_t(f) \overset{(1)}{=} t^0 + t^1 + t^1 + t^2 = 1 + 2t + t^2, \tag{7.28}$$

$$M_{-1}(f) \overset{(2)}{=} 1 - 2 + 1 \overset{(3)}{=} 0 = \chi(T^2). \tag{7.29}$$

(1) One monomial per critical point, indices $0, 1, 1, 2$ (Example 7.7).
(2) Set $t = -1$.
(3) The Euler characteristic of the torus is $b_0 - b_1 + b_2 = 1 - 2 + 1$, or, counting cells, 1 vertex −2 edges +1 face.

Since also $P_t(T^2) = 1 + 2t + t^2$, the height function on the upright torus is a perfect Morse function (over every field: T^2 has no torsion).



one 0-cell: the four corners
two 1-cells: a and b
one 2-cell: the open square
$\partial e_2 = a + b - a - b = 0, \quad \partial a = \partial b = 0$

Figure 7.11: The cell structure $T^2 = e_0 \cup e_1 \cup e_1 \cup e_2$ produced by the height function, drawn as the usual square with opposite sides identified. All boundary maps vanish, so every cell carries a cohomology class and $P_t = 1 + 2t + t^2$ equals the Morse series: the function is perfect. Compare $\mathbb{RP}^2$ in Figure 7.5, where $\partial e_2 = 2e_1 \neq 0$.

7.3.7 Cell decompositions (Nash's "stratification")

Applying (7.27) at every critical value, from the bottom up, builds M up to homotopy as a union of cells, one per critical point:

$$M \simeq \bigcup_{\text{critical } p} e_{\lambda_p} \tag{7.30}$$

[Nash p. 196, (7.16)]. A space built by successively gluing discs along their boundary spheres is a **CW complex**. So a Morse function hands us a CW structure with m_i cells of dimension i. Examples:

$$\begin{aligned} S^n &\simeq e_0 \cup e_n && \text{(height function: } M_t = 1 + t^n\text{)},\\ T^2 &\simeq e_0 \cup e_1 \cup e_1 \cup e_2 && \text{(upright height: } 1 + 2t + t^2\text{)},\\ \mathbb{RP}^2 &\simeq e_0 \cup e_1 \cup e_2 && \text{(Example 7.16: } 1 + t + t^2\text{)},\\ \mathbb{CP}^n &\simeq e_0 \cup e_2 \cup e_4 \cup \cdots \cup e_{2n} && \text{(Section 7.4: } 1 + t^2 + \cdots + t^{2n}\text{)}. \end{aligned}$$

For S^n this is the familiar picture of a sphere as a disc with its whole boundary collapsed to a point.

Watch out: "Stratification"

Nash calls (7.30) a *stratification* of M. In most modern texts that word means a decomposition into smooth pieces (strata) of various dimensions, such as an orbit space broken up by stabiliser type. What Morse theory gives is a *cell decomposition up to homotopy*: the equality in (7.30) is $\simeq$, not $=$. That is enough for computing cohomology, which is all we need.

7.3.8 Why the Morse inequalities are true

We can now prove the strong inequalities (7.16). The idea: each time a λ-cell is attached, *either* it creates a new independent λ-dimensional class (its boundary sphere was already a boundary in M_a, so the cell closes up into a new cycle), *or* it kills an existing $(\lambda-1)$-dimensional class (its boundary sphere was a nontrivial cycle, and now it bounds). In the first case b_λ goes up by one, in the second $b_{\lambda-1}$ goes down by one. The first kind of cell is "needed by the topology"; the second kind comes paired with an earlier $(\lambda-1)$-cell, and the pair contributes $t^{\lambda-1} + t^\lambda = t^{\lambda-1}(1+t)$ to the surplus.

To make this precise we use the cohomology of a pair.

The exact sequence of a pair. For $Y \subset X$ there are relative cohomology groups $H^i(X, Y)$ (forms on X that vanish on Y, roughly) and a **long exact sequence**

$$\cdots \to H^{i-1}(Y) \xrightarrow{\delta} H^i(X, Y) \xrightarrow{j} H^i(X) \xrightarrow{r} H^i(Y) \xrightarrow{\delta} H^{i+1}(X, Y) \to \cdots$$

"Exact" means the image of each map is the kernel of the next. Here r is restriction to Y and δ is the connecting map. For a single cell, $H^i(M_b, M_a) \cong H^i(D^\lambda, S^{\lambda-1})$ (everything outside the new cell cancels; this is called excision), and this is $\mathbf{F}$ for $i = \lambda$ and 0 otherwise. So

$$P_t(M_b, M_a) = t^\lambda. \tag{7.31}$$

Counting ranks. Write $q_i = \operatorname{rank}\big(\delta\colon H^i(Y) \to H^{i+1}(X,Y)\big)$. Exactness at each of the three kinds of spot says that each space splits into the kernel of the outgoing map (which is the image of the incoming map) and a complement mapped isomorphically onto the image:

$$\begin{aligned} \dim H^i(X,Y) &\overset{(1)}{=} q_{i-1} + \operatorname{rank} j_i, \\ \dim H^i(X) &\overset{(2)}{=} \operatorname{rank} j_i + \operatorname{rank} r_i, \\ \dim H^i(Y) &\overset{(3)}{=} \operatorname{rank} r_i + q_i. \end{aligned}$$

(1) Rank–nullity for j_i, with $\operatorname{Ker} j_i = \operatorname{Im} \delta_{i-1}$.
(2) Rank–nullity for r_i, with $\operatorname{Ker} r_i = \operatorname{Im} j_i$.
(3) Rank–nullity for δ_i, with $\operatorname{Ker} \delta_i = \operatorname{Im} r_i$.

Taking the first minus the second plus the third, then multiplying by t^i and summing,

$$\begin{aligned} P_t(X,Y) - P_t(X) + P_t(Y) &\overset{(1)}{=} \sum_i (q_{i-1} + q_i)\, t^i \\ &\overset{(2)}{=} \sum_i q_i\, t^{i+1} + \sum_i q_i\, t^i \overset{(3)}{=} (1+t)\, Q(t), \qquad Q(t) = \sum_i q_i t^i. \end{aligned} \tag{7.32}$$

(1) The rank j_i and rank r_i terms cancel.
(2) Shift the summation index in the first sum.
(3) Factor; Q has non-negative coefficients because ranks are non-negative.

This is where the factor $(1+t)$ comes from: every nonzero connecting map δ links degree i with degree $i+1$.

Summing over the critical points. Order the critical values $c_1 < c_2 < \cdots < c_N$ (perturb f slightly so each level has one critical point) and choose $a_0 < c_1 < a_1 < c_2 < \cdots < c_N < a_N$. Then $M_{a_0} = \varnothing$, $M_{a_N} = M$, and $(M_{a_k}, M_{a_{k-1}})$ contains exactly one critical point, of index λ_k. Apply (7.32) with (7.31) to each step and add:

$$\begin{aligned} &\sum_{k=1}^{N} \big[t^{\lambda_k} - P_t(M_{a_k}) + P_t(M_{a_{k-1}})\big] \overset{(1)}{=} (1+t) \sum_k Q_k(t) \\ &\quad\Longrightarrow\quad M_t(f) - P_t(M) + P_t(\varnothing) \overset{(2)}{=} (1+t)\, R(t), \qquad R := \textstyle\sum_k Q_k. \end{aligned} \tag{7.33}$$

(1) Relation (7.32) for $(X,Y) = (M_{a_k}, M_{a_{k-1}})$, summed over k.
(2) $\sum_k t^{\lambda_k} = M_t(f)$, and the $P_t(M_{a_k})$ terms telescope, leaving $-P_t(M_{a_N}) + P_t(M_{a_0})$.

Since $P_t(\varnothing) = 0$, this is precisely $M_t(f) - P_t(M) = (1+t)R(t)$ with $R \geq 0$: the strong Morse inequalities. Nothing depended on the field $\mathbf{F}$, which is why they hold for every field.

Worked example 7.19: Which cells are paired on the two-humped sphere?

Flood the camel-back sphere of Example 7.13. Order of critical values: minimum, saddle, lower peak, higher peak.

- Minimum ($\lambda = 0$): a disc appears; it creates $b_0 = 1$. No pairing.
- Saddle ($\lambda = 1$): the water reaches the pass. The flooded region was a disc covering the lower part of the sphere; now the strip across the pass turns it into an *annulus* whose two boundary circles surround the two peaks. A new 1-cycle is created (a

loop around one peak): $b_1(M_c) = 1$.

- Lower peak ($\lambda = 2$): the lid over the lower peak closes one boundary circle. The loop around that peak now bounds the lid, so the 1-cycle is *killed*: $\delta \neq 0$, and the peak is paired with the saddle. The flooded region is a disc again.
- Higher peak ($\lambda = 2$): the last lid closes the sphere and creates $b_2 = 1$.

So $R = Q_{\text{lower peak}} = t$, matching $M_t - P_t = (1 + t)\,t$. The pairing is concrete: the saddle and the peak that "undoes" it.

7.3.9 A preview: counting flow lines (the Morse complex)

The proof above uses only the *number* of cells in each dimension. The cells themselves come with more information: the way the boundary of each cell wraps around the lower cells. For a Morse function this information is carried by the gradient flow lines, and it gives a complete way to compute the cohomology from f. This goes beyond Nash's §1–§3, but it is exactly what Witten's §4 (part 7b) reconstructs from supersymmetric quantum mechanics, so we preview it.

Working definition 7.9 (Morse complex, over $\mathbb{Z}_2$)

Let C_i be the $\mathbb{Z}_2$-vector space with one basis vector $\langle p\rangle$ for each critical point of index i. For generic metric [all Riemannian metrics except a thin set, the metric being needed to define ∇f], the flow lines of $-\nabla f$ [curves $\gamma(s)$ with $\dot\gamma = -\nabla f$, leaving p as $s \to -\infty$ and arriving at q as $s \to +\infty$] from a critical point p of index i to a critical point q of index $i-1$ are finite in number. Define

$$\partial\langle p\rangle = \sum_{\lambda_q = i-1} \big(\#\text{flow lines } p \to q \bmod 2\big)\,\langle q\rangle.$$

Then $\partial^2 = 0$, and $\operatorname{Ker}\partial/\operatorname{Im}\partial$ [combinations c with $\partial c = 0$, counted as equal when they differ by some ∂b] is the homology of M with $\mathbb{Z}_2$ coefficients. (Over $\mathbb{Z}$ or $\mathbb{R}$ each flow line carries a sign fixed by orientations.)

Worked example 7.20: The Morse complex of $\cos x + \cos y$ on the torus

For $f = \cos x + \cos y$ with the flat metric, $-\nabla f = (\sin x, \sin y)$. The critical points are the maximum $P = (0,0)$, the saddles $S_1 = (\pi, 0)$ and $S_2 = (0, \pi)$, and the minimum $Q = (\pi, \pi)$ (Example 7.2). The line $y = 0$ is invariant under the flow (since $\sin 0 = 0$), and on it $\dot x = \sin x > 0$ for $0 < x < \pi$ and < 0 for $\pi < x < 2\pi$. So there are exactly two flow lines from P to S_1: rightwards from $x = 0$ and leftwards from $x = 2\pi$ (the same point P). Likewise two from P to S_2 (along $x = 0$), two from S_1 to Q (along $x = \pi$) and two from S_2 to Q (along $y = \pi$). Hence, over $\mathbb{Z}_2$,

$$\partial\langle P\rangle \overset{(1)}{=} 2\langle S_1\rangle + 2\langle S_2\rangle = 0, \qquad \partial\langle S_k\rangle \overset{(2)}{=} 2\langle Q\rangle = 0.$$

(1) Two flow lines from P to each saddle.
(2) Two flow lines from each saddle to the minimum.

All boundary maps vanish, so every critical point gives a homology class: $b_0 = 1$, $b_1 = 2$, $b_2 = 1$. (With signs the two flow lines cancel too, since the torus is orientable; for $\mathbb{RP}^2$ with Example 7.16 the two flow lines from the maximum to the saddle add up to 2, reproducing $\partial e_2 = 2e_1$ of Example 7.10.)

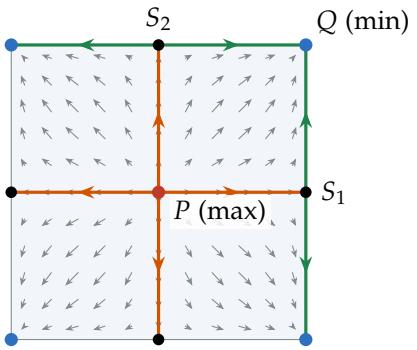


Figure 7.12: The torus as a square $[-\pi,\pi]^2$ with opposite sides identified, centred on the maximum P of $f=\cos x+\cos y$. Grey arrows: the downhill field $-\nabla f=(\sin x,\sin y)$. Orange: the four flow lines leaving P, two to each saddle (the left and right black dots are the same point S_1, the top and bottom ones are S_2). Green: the flow lines from the saddles to the minimum Q (the four corners are one point). Every critical point is joined to each lower neighbour by an *even* number of flow lines, so the Morse differential vanishes and the function is perfect.

Physical picture: Flow lines are tunnelling paths

In Witten's picture (part 7b), critical points are classical vacua of a supersymmetric particle moving on M with potential built from f. The perturbative ground states sit at the critical points, one per critical point, with fermion number equal to the index: that is $M_t(f)$. Instantons (tunnelling paths) are the gradient flow lines between critical points of adjacent index. They lift pairs of would-be ground states to nonzero energy, exactly the cancelling pairs $(1+t)R(t)$. The states that survive are the true supersymmetric ground states, counted by $P_t(M)$. The Morse inequalities become "the true ground states are a subset of the perturbative ones".

Reading Nash (pp. 195–196, (7.15)–(7.18))

Nash introduces the level surfaces $\{f=c\}$ and the "half spaces" M_c [Nash (7.15)], and states the Bott–Morse–Smale theorem in two parts: no critical value in $[a,b]$ gives $M_a\cong M_b$; one critical point of index λ gives $M_b\simeq M_a\cup e_\lambda$. He calls the resulting cell decomposition (7.30) a stratification [Nash (7.16)]. He deliberately does not prove the Morse inequalities ("rather than give another proof we shall just explain the part that [level surfaces] play"). The upright torus example with $M_t=1+2t+t^2$ and $\chi=0$ is [Nash (7.17)–(7.18)]. We have added the gradient flow and the stable/unstable manifolds, the proof of both parts of the theorem in the local model, handles, the flooding picture of the torus, the exact-sequence proof of the strong inequalities, and the Morse complex.

Recap

- $M_c=\{f\le c\}$ keeps its shape until c crosses a critical value (flow along $-\nabla f/|\nabla f|^2$ lowers f at unit rate).
- Crossing a critical point of index λ attaches a λ-handle $D^\lambda\times D^{n-\lambda}$; up to homotopy, a λ-cell whose core is the unstable manifold.
- So $M\simeq\bigcup_p e_{\lambda_p}$: a CW complex with m_i cells of dimension i.
- Each cell either creates a class in degree λ or kills one in degree $\lambda-1$; the exact sequence of a pair turns this into $M_t-P_t=(1+t)R$, $R\ge 0$.

- Upright torus: disc → cylinder → punctured torus → torus; $M_t = 1 + 2t + t^2 = P_t$, $\chi = 0$.
- Flow lines between critical points of adjacent index compute the boundary maps (Morse complex); they are Witten's instantons.

7.4 Three classic calculations: $\mathbb{CP}^n$, surfaces of genus p, and degenerate critical points

So far we have mostly used topology to predict critical points. Now we run the machine the other way. Nash's first showpiece is the complex projective space $\mathbb{CP}^n$: we shall find its complete cohomology by locating the critical points of one quadratic function and counting minus signs, with no cohomology theory at all beyond the lacunary principle. The second example, the surface of genus p, shows the $(1, 2p, 1)$ pattern of saddles. The third asks what survives when we give up non-degeneracy: then critical points can merge into "monkey saddles", the Morse inequalities fall silent, and a different and cruder lower bound, the Lusternik–Schnirelmann category, takes over.

7.4.1 $\mathbb{CP}^n$ as a quotient of an odd sphere

Working definition 7.10 (complex projective space)

$\mathbb{CP}^n$ is the set of complex lines through the origin in $\mathbb{C}^{n+1}$ [each line is $\{\mu z : \mu \in \mathbb{C}\}$ for one nonzero vector z, a copy of $\mathbb{C}$]. A point is written in **homogeneous coordinates** $[z_1 : z_2 : \cdots : z_{n+1}]$, not all z_i zero, with the rule

$$[z_1 : \cdots : z_{n+1}] = [\mu z_1 : \cdots : \mu z_{n+1}] \qquad \text{for every } \mu \in \mathbb{C}^\times = \mathbb{C} \setminus \{0\}.$$

Operational version: use $|\mu|$ to normalise $\sum_i |z_i|^2 = 1$. What is left of μ is a phase $e^{i\theta}$. So

$$\mathbb{CP}^n = S^{2n+1}/\mathrm{U}(1), \qquad S^{2n+1} = \Big\{ z \in \mathbb{C}^{n+1} : \sum_{i=1}^{n+1} z_i \bar{z}_i = 1 \Big\}, \tag{7.34}$$

[points of S^{2n+1} that differ by an overall phase $e^{i\theta}$ count as one point of $\mathbb{CP}^n$] where $\mathrm{U}(1) = S^1$ acts by $e^{i\theta} \cdot (z_1, \ldots, z_{n+1}) = (e^{i\theta} z_1, \ldots, e^{i\theta} z_{n+1})$ [Nash p. 197, (7.19)–(7.20)].

The U(1) action is **free**: if $e^{i\theta} z = z$ with $z \neq 0$ then $e^{i\theta} = 1$. Each orbit is a great circle in S^{2n+1}, and the orbits are exactly the points of $\mathbb{CP}^n$. This free-quotient picture will reappear in Section 7.6, where we compare free and non-free actions.

Charts. On the open set $U_k = \{z_k \neq 0\}$ we may use $\mu = 1/z_k$ to set $z_k = 1$. The remaining n ratios

$$w_j = z_j / z_k \qquad (j \neq k)$$

are arbitrary complex numbers, so $U_k \cong \mathbb{C}^n \cong \mathbb{R}^{2n}$. The $n+1$ charts $U_1, \ldots, U_{n+1}$ cover $\mathbb{CP}^n$ (some z_k is nonzero), so $\mathbb{CP}^n$ is a compact manifold of real dimension $2n$. The complement of U_{n+1} is $\{z_{n+1} = 0\}$, which is a copy of $\mathbb{CP}^{n-1}$. So

$$\mathbb{CP}^n = \mathbb{C}^n \sqcup \mathbb{CP}^{n-1} = \mathbb{C}^n \sqcup \mathbb{C}^{n-1} \sqcup \cdots \sqcup \mathbb{C}^1 \sqcup \mathbb{C}^0. \tag{7.35}$$

Morse theory will tell us that this decomposition is a cell structure with one cell in each even dimension, and hence what the cohomology is.

Worked example 7.21: $\mathbb{CP}^1$ is the two-sphere

The map $\pi\colon S^3 \to \mathbb{R}^3$,

$$\pi(z_1, z_2) = \big(2\,\mathrm{Re}\, z_1\bar{z}_2,\ 2\,\mathrm{Im}\, z_1\bar{z}_2,\ |z_1|^2 - |z_2|^2\big),$$

is unchanged by $z \mapsto e^{i\theta}z$ (each entry is a product of a z and a $\bar{z}$), so it is really a map from $\mathbb{CP}^1$. It lands on the unit sphere:

$$|\pi(z)|^2 \overset{(1)}{=} 4\,|z_1\bar{z}_2|^2 + \big(|z_1|^2 - |z_2|^2\big)^2 \overset{(2)}{=} \big(|z_1|^2 + |z_2|^2\big)^2 \overset{(3)}{=} 1.$$

(1) $(2\,\mathrm{Re}\, c)^2 + (2\,\mathrm{Im}\, c)^2 = 4\,|c|^2$ with $c = z_1\bar{z}_2$.
(2) $4\,|z_1|^2\,|z_2|^2 + (A - B)^2 = (A + B)^2$ with $A = |z_1|^2$, $B = |z_2|^2$.
(3) We are on S^3.

It is a bijection $\mathbb{CP}^1 \to S^2$ (the inverse sends a point of S^2 to its stereographic coordinate $w = z_1/z_2$, the Riemann sphere). This π is the **Hopf map** $S^3 \to S^2$, whose fibres are the U(1) orbits.

7.4.2 The Morse function on $\mathbb{CP}^n$

Pick $n+1$ real numbers $a_1 < a_2 < \cdots < a_{n+1}$ and set

$$\tilde{f}(z) = a_1 z_1\bar{z}_1 + a_2 z_2\bar{z}_2 + \cdots + a_{n+1} z_{n+1}\bar{z}_{n+1} \qquad \text{on } S^{2n+1} \tag{7.36}$$

[Nash p. 197, (7.21)]. Since $\left|e^{i\theta}z_i\right| = |z_i|$ it is constant on orbits, so it defines a function f on $\mathbb{CP}^n$ with $\tilde{f} = f \circ \pi$. In homogeneous coordinates, without normalising,

$$f([z]) = \frac{\sum_i a_i\,|z_i|^2}{\sum_i |z_i|^2}. \tag{7.37}$$

Physically, f is the expectation value $\langle z|A|z\rangle/\langle z|z\rangle$ of the Hermitian matrix $A = \mathrm{diag}(a_1, \dots, a_{n+1})$ in the state $|z\rangle$. Here $\mathbb{CP}^n$ plays the role of the space of pure states of an $(n+1)$-level quantum system (the ray space), and the stationary points of f are exactly the eigenstates of A – this is the Rayleigh–Ritz variational principle.

Where are the critical points? Since π is a submersion, $\mathrm{d}\tilde{f} = \mathrm{d}f \circ \mathrm{d}\pi$ vanishes at z exactly when $\mathrm{d}f$ vanishes at $\pi(z)$. So we find the critical points of $\tilde{f}$ on the sphere, by a Lagrange multiplier μ for the constraint $\sum |z_i|^2 = 1$. Writing $z_i = x_i + iy_i$ and differentiating in x_i and y_i,

$$\mathrm{d}\tilde{f} = \mu\,\mathrm{d}\Big(\sum_i |z_i|^2\Big) \overset{(1)}{=} \iff 2a_i x_i = 2\mu x_i \ \text{ and } \ 2a_i y_i = 2\mu y_i \quad \text{for all } i$$

$$\overset{(2)}{=} \iff (a_i - \mu)\,z_i = 0 \quad \text{for all } i.$$

(1) $\partial_{x_i}(a_i(x_i^2 + y_i^2)) = 2a_i x_i$, and likewise for y_i and for the constraint.
(2) Combine the real and imaginary parts.

Because the a_i are distinct, at most one z_i can be nonzero, and on the sphere exactly one is. So the critical points of f are the $n+1$ *coordinate points*

$$P_k = [0 : \cdots : 0 : \underset{k}{1} : 0 : \cdots : 0], \qquad k = 1, \dots, n+1,$$

with critical values $f(P_k) = a_k$. In the quantum language: the eigenstates of A. Nash: "one for each direction z_i".

The Hessian at P_k. Use the chart U_k, with $z_k = 1$ and $w_j = z_j$ for $j \neq k$. Write $A(w) = \sum_{j\neq k} a_j |w_j|^2$ and $S(w) = \sum_{j\neq k} |w_j|^2$, both quadratic in w. Then

$$\begin{aligned} f(w) &\overset{(1)}{=} \frac{a_k + A(w)}{1 + S(w)} \\ &\overset{(2)}{=} (a_k + A)(1 - S + S^2 - \dots) \\ &\overset{(3)}{=} a_k + A - a_k S + O(|w|^4) \\ &\overset{(4)}{=} a_k + \sum_{j\neq k} (a_j - a_k)\,|w_j|^2 + O(|w|^4). \end{aligned} \tag{7.38}$$

(1) Formula (7.37) with $z_k = 1$, $z_j = w_j$.
(2) Geometric series $1/(1+S) = 1 - S + \dots$, valid for small w.
(3) Keep the terms of order ≤ 2; AS and S^2 are quartic.
(4) Combine $A - a_k S = \sum_{j\neq k}(a_j - a_k)|w_j|^2$.

Now the crucial observation Nash makes. With $w_j = u_j + iv_j$, $|w_j|^2 = u_j^2 + v_j^2$, so each complex direction w_j contributes *two* equal real eigenvalues $2(a_j - a_k)$ to the $2n \times 2n$ real Hessian:

$$Hf(P_k) = \operatorname{diag}\big(\underbrace{2(a_1 - a_k),\, 2(a_1 - a_k)}_{w_1},\ \dots,\ \underbrace{2(a_{n+1} - a_k),\, 2(a_{n+1} - a_k)}_{w_{n+1}}\big) \quad (j \neq k). \tag{7.39}$$

No eigenvalue is zero (the a's are distinct), so f is a Morse function. The negative eigenvalues come from the j with $a_j < a_k$, i.e. $j < k$. There are $k-1$ of them, each counted twice:

$$\lambda_{P_k} = 2(k-1), \qquad k = 1, \dots, n+1. \tag{7.40}$$

Key idea: The cohomology of $\mathbb{CP}^n$ (Nash (7.22)–(7.23))

The Morse series of f is

$$M_t(f) = 1 + t^2 + t^4 + \cdots + t^{2n}. \tag{7.41}$$

Only even powers occur, so no two are consecutive, and the lacunary principle gives, for every coefficient field $\mathbf{F}$,

$$P_t(\mathbb{CP}^n; \mathbf{F}) = 1 + t^2 + t^4 + \cdots + t^{2n} = \frac{1 - t^{2n+2}}{1 - t^2}. \tag{7.42}$$

So $b_{2i}(\mathbb{CP}^n) = 1$ for $0 \le i \le n$ and all odd Betti numbers vanish; $H^*(\mathbb{CP}^n; \mathbb{Z})$ has no torsion; $\chi(\mathbb{CP}^n) = n+1$; and f is perfect.

Worked example 7.22: $\mathbb{CP}^2$ with $a = (0, 1, 2)$

Take $f = (|z_2|^2 + 2|z_3|^2)/(|z_1|^2 + |z_2|^2 + |z_3|^2)$. Apply (7.38) at each coordinate point:

point	chart	f near the point	$f(P_k)$	real Hessian	index
$[1:0:0]$	$z_1 = 1$	$0 + \lvert w_2\rvert^2 + 2\lvert w_3\rvert^2$	0	$\mathrm{diag}(2,2,4,4)$	0
$[0:1:0]$	$z_2 = 1$	$1 - \lvert w_1\rvert^2 + \lvert w_3\rvert^2$	1	$\mathrm{diag}(-2,-2,2,2)$	2
$[0:0:1]$	$z_3 = 1$	$2 - 2\lvert w_1\rvert^2 - \lvert w_2\rvert^2$	2	$\mathrm{diag}(-4,-4,-2,-2)$	4

(Quartic terms omitted.) So $M_t = 1+t^2+t^4$, and by the lacunary principle $P_t(\mathbb{CP}^2) = 1+t^2+t^4$: $\mathbb{CP}^2$ has no 1-dimensional holes ($b_1 = 0$), one 2-cycle (the line $\mathbb{CP}^1 \subset \mathbb{CP}^2$) and a fundamental class in degree 4. The Euler characteristic is 3.

Worked example 7.23: $\mathbb{CP}^1$: the function is the height on S^2

For $n = 1$, on S^3 we have $|z_1|^2 + |z_2|^2 = 1$ and, with x_3 the third component of the Hopf map of Example 7.21, $|z_1|^2 - |z_2|^2 = x_3$. So $|z_1|^2 = (1 + x_3)/2$ and $|z_2|^2 = (1 - x_3)/2$. Then

$$f \overset{(1)}{=} a_1 \frac{1 + x_3}{2} + a_2 \frac{1 - x_3}{2} \overset{(2)}{=} \frac{a_1 + a_2}{2} - \frac{a_2 - a_1}{2} x_3.$$

(1) Substitute into $f = a_1 |z_1|^2 + a_2 |z_2|^2$.
(2) Collect the constant and the x_3 terms.

Since $a_2 > a_1$ this is a positive multiple of *minus* the height, plus a constant. Its minimum is at $x_3 = +1$, which is $P_1 = [1:0]$ (index 0), and its maximum at $x_3 = -1$, which is $P_2 = [0:1]$ (index 2). We recover Example 7.6 for $n = 2$: $M_t = 1 + t^2 = P_t(S^2)$.

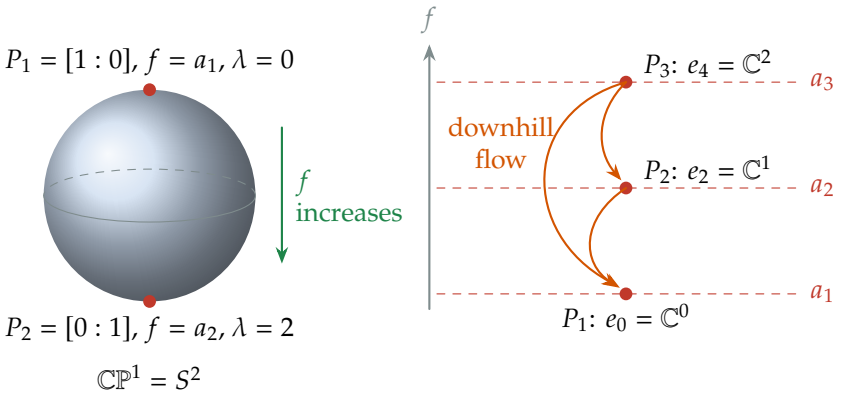


Figure 7.13: Left: on $\mathbb{CP}^1 = S^2$ the function $a_1|z_1|^2 + a_2|z_2|^2$ is (minus) the height, with one minimum and one maximum. Right: the three critical levels of f on $\mathbb{CP}^2$. Each P_k has index $2(k-1)$; the orange downhill flow lines leaving P_k sweep out a copy of $\mathbb{C}^{k-1}$, the cell $e_{2(k-1)}$. There are no odd cells, so there is nothing for a boundary map to connect: every cell is a cohomology class.

The cells, explicitly. We can see the cell decomposition (7.35) emerging from the flow. For the Fubini–Study metric (the natural round metric on $\mathbb{CP}^n$) the downhill gradient flow of f is, up to a positive constant c in the exponent fixed by the normalisation of the metric,

$$[z_1 : \cdots : z_{n+1}] \longmapsto [e^{-ca_1 t} z_1 : \cdots : e^{-ca_{n+1} t} z_{n+1}].$$

As $t \to -\infty$ the component with the largest a_j among the nonzero z_j dominates. So the points that flow out of P_k (the unstable manifold of P_k) are exactly those with $z_k \neq 0$ and $z_j = 0$ for $j > k$:

$$W^-(P_k) = \{[z_1 : \cdots : z_k : 0 : \cdots : 0],\ z_k \neq 0\} \cong \mathbb{C}^{k-1},$$

an open cell of real dimension $2(k-1) = \lambda_{P_k}$, as the general theory of Example 7.17 demands. Together they give $\mathbb{CP}^n = e_0 \cup e_2 \cup \cdots \cup e_{2n}$, and the sublevel set just below the top critical value is (up to homotopy) $\mathbb{CP}^{n-1}$.

Watch out: Nash's labelling of the index

Nash says "the i-th critical point has index $2i$". With his labelling $i = 1, \ldots, n+1$ that would give indices $2, 4, \ldots, 2n+2$, which is off by one step (the maximum index on a $2n$-manifold is $2n$). The correct statement is (7.40): the point where $f = a_k$ has index $2(k-1)$, i.e. index $2i$ if we count $i = 0, 1, \ldots, n$. His formula (7.22) is correct.

Physical picture: Why complex manifolds give even indices

Nash remarks that this "illustrates the point that we can use Morse theory on complex manifolds even though f is a real valued function". There is no choice: a holomorphic function on a compact complex manifold is constant (it attains a maximum of $|F|$, so by the maximum principle it is constant), so Morse functions are always real. What the complex structure does is *pair up the real directions*, $w = u + iv$. Our f is invariant under the phase rotations $z_j \mapsto e^{i\theta_j} z_j$, and a function invariant under a rotation of the (u_j, v_j) plane has equal Hessian eigenvalues in that plane. Hence every index is even, and the lacunary principle applies. (This is the simplest case of a general fact: on a compact symplectic manifold with a circle action, the "moment map" is a Morse–Bott function with even indices and is always perfect. Our f is the moment map of the phase rotations.) By contrast, the real part of a holomorphic function has critical points of index *exactly* half the real dimension: $\mathrm{Re}(z^2) = x^2 - y^2$ has index 1 on $\mathbb{C}$. We use this in Section 7.4.4.

7.4.3 The height function on a surface of genus p

A **surface of genus** p, Σ_p, is a closed orientable surface with p holes: a sphere with p handles attached, or p tori glued in a row. Every closed orientable surface is one of these. Stand Σ_p on its end, with the holes stacked vertically, and let f be the height [Nash p. 197].

Worked example 7.24: Critical points of the upright Σ_p

Reading off from Figure 7.14: a minimum at the bottom (index 0), a maximum at the top (index 2), and for each hole two saddles (index 1), one at the bottom of the hole and one at its top. Each saddle is exactly like those of the upright torus, Example 7.7. So

$$M_t(f) = 1 + 2p\,t + t^2 \tag{7.43}$$

[Nash (7.24)], and

$$\chi(\Sigma_p) \overset{(1)}{=} M_{-1}(f) \overset{(2)}{=} 1 - 2p + 1 = 2 - 2p. \tag{7.44}$$

(1) Relation (7.21): the Euler characteristic is the signed count of critical points.
(2) Set $t = -1$ in (7.43).

Since $P_t(\Sigma_p) = 1 + 2pt + t^2$ (the $2p$ independent loops $a_1, b_1, \ldots, a_p, b_p$ around and through each hole), this height function is perfect. Gauss–Bonnet gives a completely independent

check, $\frac{1}{2\pi}\int_{\Sigma_p} K\,\mathrm{d}A = 2-2p$ (Chapter T2): the saddles are where the surface is negatively curved.

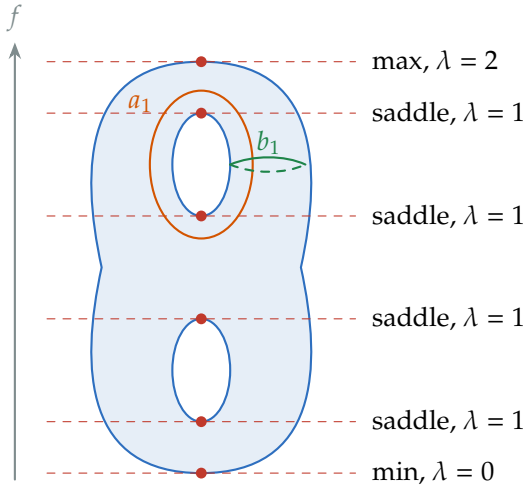


Figure 7.14: A surface of genus 2 standing upright. The height has $2p+2=6$ critical points: a minimum, a maximum, and a saddle at the bottom and the top of each hole. The dashed red lines are the six critical levels. The orange loop a_1 goes around the upper hole and the green loop b_1 goes through the tube beside it; the $2p$ loops a_i, b_i are the $2p$ independent 1-dimensional holes that force the $2p$ saddles.

Try it: Sublevel sets of Σ_2

Flood the surface of Figure 7.14. Describe M_c just above each critical level.
Answer: disc; annulus (cylinder); annulus with its two ends joined over the lower hole, i.e. a punctured torus ($\simeq$ wedge of 2 circles); then the lower saddle of the upper hole adds a strip, giving a torus with two punctures ($\simeq$ wedge of 3 circles); the top saddle of the upper hole gives a punctured genus-2 surface ($\simeq$ wedge of 4 circles); the maximum caps it. Cells: $e_0 \cup 4e_1 \cup e_2$.

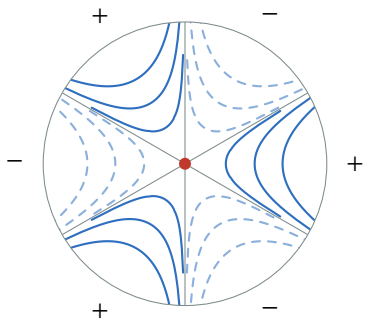
7.4.4 Squashing the saddles together: degenerate critical points

The Morse inequalities say that a Morse function on Σ_p needs at least $2p+2$ critical points. Nash points out that if we drop non-degeneracy we can do much better: *four* critical points suffice for every p, a maximum, a minimum and two degenerate saddles. Deform the upright Σ_p so that the p lower saddles slide together into one point, and likewise the p upper ones. Near the merged saddle, in suitable coordinates (x,y) centred on it,

$$f = \mathrm{Re}(x+iy)^{p+1} = \mathrm{Re}\, z^{p+1} = r^{p+1}\cos\big((p+1)\varphi\big) \tag{7.45}$$

[Nash p. 198], where $z = x+iy = re^{i\varphi}$.

What this looks like. As we go once around the point, $\cos((p+1)\varphi)$ changes sign $2(p+1)$ times: there are $p+1$ "valleys" ($f<0$) and $p+1$ "ridges" ($f>0$) alternating around it. For $p=1$ this is the ordinary saddle x^2-y^2 (two valleys). For $p=2$ it is the monkey saddle x^3-3xy^2 of Example 7.3 (three valleys). All second derivatives of $r^{p+1}\cos((p+1)\varphi)$ vanish at 0 when $p+1\geq 3$, so the Hessian is the zero matrix: the point is as degenerate as possible and has no Morse index.



$\operatorname{Re} z^3$: three ridges (+),
three valleys (−)

Figure 7.15: Level curves of the monkey saddle $f = \operatorname{Re} z^3 = r^3 \cos 3\varphi$ ($p = 2$ in (7.45)). Solid: $f > 0$; dashed: $f < 0$; faint rays: the level $f = 0$, six half-lines through the critical point. Compare the ordinary saddle in Figure 7.2, which has two ridges and two valleys. A monkey saddle is two ordinary saddles squashed into one point.

Counting a degenerate critical point: the gradient winding number. The Morse index is lost, but the Euler count (7.21) must still hold, since χ is topological. The replacement for $(-1)^\lambda$ is the **index of the vector field** ∇f at the zero: the number of times the direction of ∇f turns around as we go once counter-clockwise around a small circle about the point. For a holomorphic $F = f + ig$ the Cauchy–Riemann equations give $\partial_x f = \operatorname{Re} F'$ and $\partial_y f = -\operatorname{Im} F'$. So

$$\begin{aligned} \partial_x f + i\,\partial_y f &\overset{(1)}{=} \operatorname{Re} F'(z) - i \operatorname{Im} F'(z) \\ &\overset{(2)}{=} \overline{F'(z)} \overset{(3)}{=} (p+1)\,\bar{z}^{\,p} \overset{(4)}{=} (p+1)\,r^p e^{-ip\varphi}. \end{aligned} \tag{7.46}$$

(1) The Cauchy–Riemann equations just quoted.
(2) $\operatorname{Re} w - i \operatorname{Im} w = \bar{w}$.
(3) $F = z^{p+1}$, so $F' = (p+1)z^p$.
(4) $z = re^{i\varphi}$.

As φ runs from 0 to 2π the direction $e^{-ip\varphi}$ of ∇f turns p times *clockwise*: the vector-field index is $-p$. Check $p = 1$: an ordinary saddle has index $-1 = (-1)^1$. Now the Euler count on the squashed Σ_p:

$$\chi(\Sigma_p) \overset{(1)}{=} \underbrace{(+1)}_{\max} + \underbrace{(+1)}_{\min} + \underbrace{2 \times (-p)}_{\text{two degenerate saddles}} = 2 - 2p.$$

(1) Poincaré–Hopf: the sum of the indices of the zeros of ∇f is χ. A non-degenerate max or min has vector-field index +1.

It matches. The two degenerate saddles carry all the topology of the $2p$ original saddles.

Unfolding the degeneracy. Nash notes that a small perturbation removes the degeneracy: replace

$$z^{p+1} \longrightarrow F_\varepsilon(z) = (z - \varepsilon_1)(z - \varepsilon_2) \cdots (z - \varepsilon_{p+1}) \tag{7.47}$$

with distinct small ε_i. The critical points of $f_\varepsilon = \operatorname{Re} F_\varepsilon$ are, by (7.46) with F_ε in place of z^{p+1}, the zeros of F'_ε, a polynomial of degree p. For generic ε_i its p roots $\zeta_1, \ldots, \zeta_p$ are distinct and

$F''_\varepsilon(\zeta_m) \neq 0$. Near each root,

$$f_\varepsilon \overset{(1)}{=} \operatorname{Re}\Big[F_\varepsilon(\zeta_m) + \tfrac{1}{2}F''_\varepsilon(\zeta_m)(z-\zeta_m)^2 + \dots\Big]$$
$$\overset{(2)}{=} \text{const} + \tfrac{1}{2}\,|F''_\varepsilon(\zeta_m)|\,\operatorname{Re}(\xi^2) + \dots \overset{(3)}{=} \text{const} + \tfrac{1}{2}\,|F''_\varepsilon(\zeta_m)|\,(\xi_1^2 - \xi_2^2) + \dots$$

(1) Taylor expansion of the holomorphic F_ε; the linear term vanishes because $F'_\varepsilon(\zeta_m) = 0$.
(2) Write $F''_\varepsilon(\zeta_m) = |F''_\varepsilon|\,e^{i\alpha}$ and rotate, $\xi = e^{i\alpha/2}(z-\zeta_m)$.
(3) $\operatorname{Re}\xi^2 = \xi_1^2 - \xi_2^2$ for $\xi = \xi_1 + i\xi_2$.

So each ζ_m is a non-degenerate saddle of index 1. The single degenerate point has *bifurcated* into p ordinary saddles, and the total vector-field index is conserved: $p \times (-1) = -p$. Doing this to both degenerate points restores a Morse function with $2p+2$ critical points.

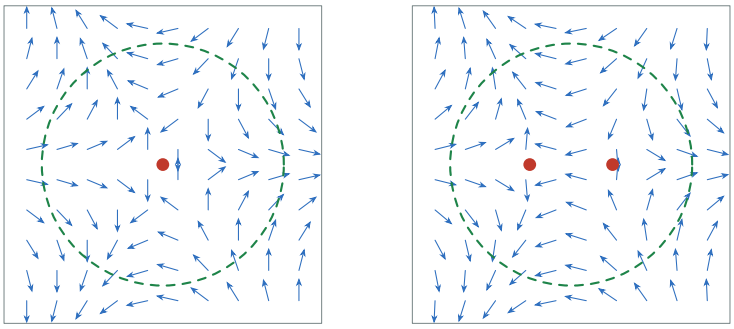
Worked example 7.25: Splitting a monkey saddle

Take $p = 2$ and $F_\varepsilon = z^3 - 3\varepsilon z$ with $\varepsilon > 0$ (this is $(z-\varepsilon_1)(z-\varepsilon_2)(z-\varepsilon_3)$ with $\varepsilon_1+\varepsilon_2+\varepsilon_3 = 0$, a harmless normalisation). Then

$$F'_\varepsilon(z) \overset{(1)}{=} 3(z^2-\varepsilon) \quad\Longrightarrow\quad \zeta_\pm = \pm\sqrt{\varepsilon}, \qquad F''_\varepsilon(\zeta_\pm) \overset{(2)}{=} \pm 6\sqrt{\varepsilon} \neq 0.$$

(1) Differentiate.
(2) $F''_\varepsilon = 6z$.

In real terms $f_\varepsilon = x^3 - 3xy^2 - 3\varepsilon x$, and at $(\pm\sqrt{\varepsilon}, 0)$ the Hessian is $\operatorname{diag}(\pm 6\sqrt{\varepsilon}, \mp 6\sqrt{\varepsilon})$: two non-degenerate saddles. As $\varepsilon \to 0$ they approach each other and merge into the monkey saddle. This is the two-dimensional cousin of the birth–death picture Figure 7.3, except that here two critical points of the *same* index merge, which a Morse function could never do generically.



$\varepsilon = 0$: monkey saddle, index -2 | $\varepsilon = 0.3$: two saddles, index -1 each

Figure 7.16: The direction of ∇f (unit arrows) for $f_\varepsilon = x^3 - 3xy^2 - 3\varepsilon x$. Follow the arrows around the green dashed circle: in both pictures they turn twice clockwise, so the total index inside the circle is -2. On the left it is concentrated in one degenerate point; on the right it is shared by two ordinary saddles (red dots). The perturbation moves the zeros but cannot change the winding number on a circle it does not touch.

Watch out: Degenerate functions have no Morse series

The Morse series (7.11) is only defined for Morse functions: a degenerate critical point has no index to put in the exponent. The Morse inequalities therefore say *nothing* about degenerate functions, and indeed the four-point function on Σ_p violates $M_1(f) \geq 2p+2$.

What survives degeneration is only the Euler count, with vector-field indices in place of $(-1)^\lambda$.

7.4.5 A lower bound for every function: the Lusternik–Schnirelmann category

Since degenerate functions can have fewer critical points, Nash asks: is there any topological lower bound on the number of critical points of an *arbitrary* smooth function? Yes.

Working definition 7.11 (Lusternik–Schnirelmann category)

The **category** $\operatorname{cat}(M)$ is the smallest number k of closed subsets $M_1, \dots, M_k$ with $M = M_1 \cup \cdots \cup M_k$ such that each M_i is contractible in M (it can be shrunk to a point through M) [Nash p. 198]. *Operational use:* upper bounds come from exhibiting a cover [an explicit choice of k such sets, which shows $\operatorname{cat}(M) \le k$]; lower bounds come from cohomology (see below). (Many modern texts subtract one, so that cat of a contractible space is 0; we follow Nash, for whom it is 1.)

Key idea: The Lusternik–Schnirelmann theorem

Every smooth function on a compact manifold M has at least $\operatorname{cat}(M)$ distinct critical points. $\operatorname{cat}(M)$ is a topological invariant.

Why: the minimax principle. Let $\mathcal{A}_j$ be the family of closed sets $A \subset M$ that cannot be covered by fewer than j closed sets contractible in M. Define

$$c_j = \inf_{A \in \mathcal{A}_j} \max_{x \in A} f(x), \qquad j = 1, \dots, \operatorname{cat}(M).$$

$\mathcal{A}_1$ contains single points, so $c_1 = \min f$, and $\mathcal{A}_{\operatorname{cat}(M)}$ contains M, so $c_{\operatorname{cat} M} \le \max f$. The families shrink as j grows, so $c_1 \le c_2 \le \dots$. Each c_j is a critical value: if it were not, the downhill flow of Section 7.3.3 would push the whole of a near-optimal A below c_j, and a deformation cannot lower the category of A, so we would have found a set in $\mathcal{A}_j$ with maximum below c_j, contradicting the definition. If all the c_j are different we have $\operatorname{cat}(M)$ critical values. If two coincide, $c_j = c_{j+1}$, one shows that the critical set at that level cannot be contractible in M, so it is infinite. Either way there are at least $\operatorname{cat}(M)$ critical points. (This is the physicist's familiar "mountain pass" or minimax argument: the lowest pass over which you must climb to get from one valley to another is a saddle point.)

How to compute it. Two estimates bracket $\operatorname{cat}(M)$ for a connected n-manifold:

$$(\text{cup length of } M) + 1 \;\le\; \operatorname{cat}(M) \;\le\; n + 1. \tag{7.48}$$

The **cup length** is the largest number of positive-degree cohomology classes whose wedge (cup) product is nonzero. The lower bound holds because a class of positive degree restricts to zero on a set contractible in M, and a product of classes each vanishing on one of k sets covering M vanishes on their union. The upper bound comes from a fine triangulation.

Worked example 7.26: Categories of the standard examples

- S^n: the two closed hemispheres (slightly enlarged) are contractible, so cat ≤ 2; S^n is not contractible, so cat $\neq 1$. Hence $\operatorname{cat}(S^n) = 2$ [Nash (7.25)], and the height function (two critical points) achieves the bound.

- T^2: $[\mathrm{d}x] \wedge [\mathrm{d}y] = [\mathrm{d}x \wedge \mathrm{d}y] \neq 0$, so the cup length is 2 and cat ≥ 3; also cat ≤ 3 by (7.48). So $\mathrm{cat}(T^2) = 3$.
- Σ_p, $p \geq 1$: the classes dual to a_1 and b_1 have nonzero product (the area class), so again $\mathrm{cat}(\Sigma_p) = 3$ [Nash (7.25)].
- T^n: $[\mathrm{d}x_1] \wedge \cdots \wedge [\mathrm{d}x_n] \neq 0$, so $\mathrm{cat}(T^n) = n + 1$. Compare the Morse bound 2^n.
- $\mathbb{CP}^n$: the generator $u \in H^2$ has $u^n \neq 0$ (it integrates to 1 over $\mathbb{CP}^n$, suitably normalised), so the cup length is n and $\mathrm{cat}(\mathbb{CP}^n) = n + 1$. Here the LS bound equals the Morse bound, and our f achieves it.

Worked example 7.27: A function on the torus with only three critical points

Let $f(x, y) = \sin x + \sin y + \sin(x + y)$ on $T^2 = \mathbb{R}^2/(2\pi\mathbb{Z})^2$. Critical points need

$$\partial_x f = \cos x + \cos(x + y) = 0, \qquad \partial_y f = \cos y + \cos(x + y) = 0.$$

Subtracting, $\cos x = \cos y$, so $y = x$ or $y = -x \pmod{2\pi}$.

- If $y = -x$: $\cos x + \cos 0 = 0$ forces $\cos x = -1$, so $(x, y) = (\pi, \pi)$.
- If $y = x$: $\cos x + \cos 2x = 0$, i.e. $2\cos^2 x + \cos x - 1 = 0$, i.e. $(2\cos x - 1)(\cos x + 1) = 0$. So $x = \pm\pi/3$ or $x = \pi$.

Exactly three critical points: $(\frac{\pi}{3}, \frac{\pi}{3})$ with $f = \frac{3\sqrt{3}}{2}$ (the maximum), $(-\frac{\pi}{3}, -\frac{\pi}{3})$ with $f = -\frac{3\sqrt{3}}{2}$ (the minimum), and (π, π) with $f = 0$. At the last one,

$$\partial_x^2 f \overset{(1)}{=} -\sin x - \sin(x + y) = 0, \quad \partial_y^2 f = 0, \quad \partial_x \partial_y f \overset{(2)}{=} -\sin(x + y) = 0,$$

(1) Differentiate $\partial_x f$ once more; $\sin\pi = \sin 2\pi = 0$.
(2) Differentiate $\partial_x f$ with respect to y.

The Hessian vanishes: it is a monkey saddle, with vector-field index -2, and $1 + 1 - 2 = 0 = \chi(T^2)$. So the LS bound $\mathrm{cat}(T^2) = 3$ is attained. No Morse function could do this; the Morse bound is 4.

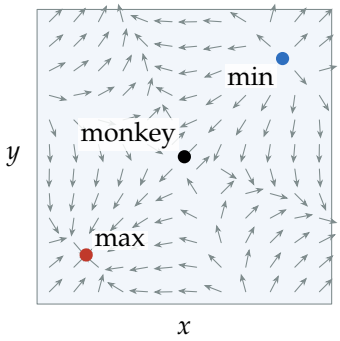


Figure 7.17: The direction of ∇f for $f = \sin x + \sin y + \sin(x + y)$ on the torus $[0, 2\pi]^2$ (opposite sides identified). The arrows flow out of the minimum (blue) and into the maximum (red); around the black point at (π, π) they turn twice clockwise, the signature of a monkey saddle. Three critical points in all, the minimum allowed by the Lusternik–Schnirelmann category.

Reading Nash (pp. 196–198, (7.19)–(7.25))

Nash's $\mathbb{CP}^n$ example: S^{2n+1} [Nash (7.19)], the S^1 action [Nash (7.20)], $\tilde{f} = \sum a_i z_i \bar{z}_i$ [Nash (7.21)], "a straightforward piece of calculus" for the indices (we did it in (7.38)), $M_t = 1+t^2+\cdots+t^{2n}$ [Nash (7.22)] and the lacunary conclusion [Nash (7.23)]. The genus-p height function gives [Nash (7.24)] and $\chi = 2-2p$. Nash then describes the four-point degenerate function, the local form $\mathrm{Re}(x+iy)^{p+1}$, the Lusternik–Schnirelmann category with $\mathrm{cat}(S^2) = 2$, $\mathrm{cat}(\Sigma_p) = 3$ [Nash (7.25)], and the unfolding $z^{p+1} \to \prod(z-\varepsilon_i)$. He also remarks that three critical points can be attained on Σ_p, "but not by any embedding of Σ_p in three dimensions" (i.e. not as a height function); we have not checked this claim and do not use it. What we added: charts and the Hopf map, the Lagrange multiplier and Hessian computations, the explicit cells, the quantum-state reading of f, the gradient winding number and Poincaré–Hopf check, the explicit bifurcation, the minimax mechanism and cup-length bounds, and the three-point function on T^2.

Recap

- $\mathbb{CP}^n = S^{2n+1}/\mathrm{U}(1)$; $f = \sum a_i |z_i|^2 / \sum |z_i|^2$ has $n+1$ critical points (the eigenstates of $\mathrm{diag}(a_i)$) of indices $0, 2, \ldots, 2n$; complex directions double every eigenvalue.
- Lacunary $\Rightarrow P_t(\mathbb{CP}^n) = 1 + t^2 + \cdots + t^{2n}$ over every field; $\mathbb{CP}^n = e_0 \cup e_2 \cup \cdots \cup e_{2n}$.
- Upright Σ_p: $M_t = 1 + 2pt + t^2$, perfect, $\chi = 2 - 2p$.
- Degenerate points: $\mathrm{Re}\, z^{p+1}$ has vector-field index $-p$; perturbing it splits it into p ordinary saddles.
- Any smooth f has $\geq \mathrm{cat}(M)$ critical points; $\mathrm{cat}(S^n) = 2$, $\mathrm{cat}(\Sigma_p) = 3$, $\mathrm{cat}(\mathbb{CP}^n) = n+1$; $\sin x + \sin y + \sin(x+y)$ attains 3 on T^2.

7.5 Critical submanifolds: Morse–Bott theory

Symmetry and isolated critical points do not get on. If a function is invariant under rotations about an axis, then rotating a critical point gives another critical point, and unless the point sits on the axis, it sweeps out a whole circle of them. Physics is full of such families. The minimum of a Mexican-hat potential is a circle of vacua; the classical solutions of a rotationally invariant problem come in orbits; every closed geodesic can be started at any point along itself. A Morse function cannot have any of this: its critical points are isolated. Bott's generalisation, which Nash describes in §2, keeps the whole machine running when the critical points come in smooth families, provided the function is non-degenerate *across* each family. The key new fact is that a critical family C contributes not a monomial t^λ but a whole polynomial $t^{\lambda_C} P_t(C)$.

Physical picture: Vacuum manifolds and Goldstone modes

Take the Mexican hat $V(\phi) = (|\phi|^2 - v^2)^2$ for a complex field $\phi = \phi_1 + i\phi_2$ (constant in space). Its minima form the circle $|\phi| = v$. At a point of the circle, the Hessian has one positive eigenvalue (the radial "Higgs" direction, $\mathrm{mass}^2 = 8v^2$) and one zero eigenvalue, along the circle (the Goldstone mode). The zero mode is not an accident to be perturbed away: it is forced by the U(1) symmetry. Morse–Bott theory is the statement that zero modes of this kind, *exactly tangent to the critical family*, are harmless.

7.5.1 The torus on a dinner plate

Nash's motivating example [Nash p. 199]: the height function on the torus of Section 7.1.7, but now lying flat, as it would on a dinner plate. The embedding is

$$(\theta, \phi) \mapsto \big((R + r\cos\theta)\cos\phi,\ (R + r\cos\theta)\sin\phi,\ r\sin\theta\big), \tag{7.49}$$

with θ around the tube and ϕ around the vertical axis. The height is $f(\theta, \phi) = r\sin\theta$, independent of ϕ.

Worked example 7.28: Critical circles of the flat torus

The derivatives are

$$\partial_\theta f = r\cos\theta, \qquad \partial_\phi f = 0.$$

So $\mathrm{d}f = 0$ exactly when $\cos\theta = 0$: on the two circles

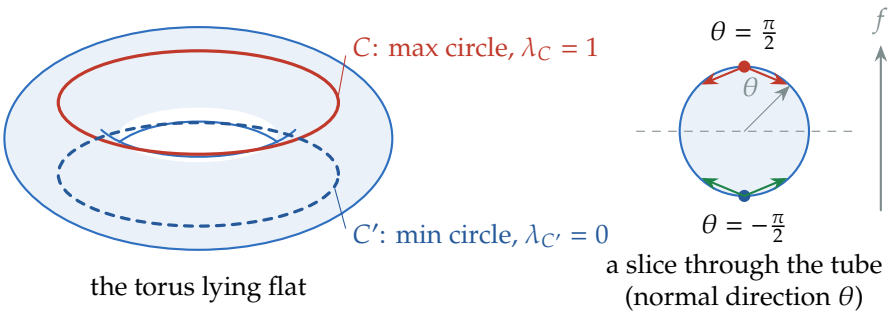
$$C = \{\theta = \tfrac{\pi}{2}\}\ (\text{top}, f = r), \qquad C' = \{\theta = -\tfrac{\pi}{2}\}\ (\text{bottom}, f = -r).$$

The Hessian in the coordinates (θ, ϕ) is

$$Hf \overset{(1)}{=} \begin{pmatrix} -r\sin\theta & 0 \\ 0 & 0 \end{pmatrix} \overset{(2)}{=} \begin{cases} \mathrm{diag}(-r, 0) & \text{on } C, \\ \mathrm{diag}(+r, 0) & \text{on } C'. \end{cases}$$

(1) $\partial_\theta^2(r\sin\theta) = -r\sin\theta$; all derivatives involving ϕ vanish because f does not depend on ϕ.
(2) $\sin(\pm\pi/2) = \pm 1$.

Every critical point is degenerate ($\det Hf = 0$), so f is not a Morse function and the theory of Section 7.2 does not apply. But the degeneracy has a very particular form. The zero eigenvector ∂_ϕ is tangent to the critical circle, and in the one direction *normal* to the circle, ∂_θ, the Hessian is non-degenerate: negative on the top circle (moving off it, over the tube, goes down), positive on the bottom one.



the torus lying flat

a slice through the tube (normal direction θ)

Figure 7.18: Left: the torus lying flat. The height is constant along the top circle C (red) and along the hidden bottom circle C' (blue, dashed), and these are its only critical points. Right: a vertical slice through the tube. In the one normal direction θ, the top of the slice is a maximum (red arrows go downhill) and the bottom a minimum (green arrows go uphill). Along the circles themselves nothing changes: that is the zero eigenvalue.

7.5.2 Non-degenerate critical submanifolds

First a basic fact: along a critical family the tangent directions are *always* in the kernel of the Hessian. Let C be a submanifold on which $\mathrm{d}f = 0$, and let $\gamma(s)$ be a curve in C. For each j, the function $s \mapsto \partial_j f(\gamma(s))$ is identically zero, so

$$0 \overset{(1)}{=} \frac{\mathrm{d}}{\mathrm{d}s}\, \partial_j f(\gamma(s)) \overset{(2)}{=} \partial_i \partial_j f\ \dot\gamma^i \overset{(3)}{=} H_{ji}\, \dot\gamma^i. \tag{7.50}$$

(1) $\partial_j f = 0$ at every point of C, in particular along γ.
(2) Chain rule.
(3) Definition of the Hessian.

So $H\dot\gamma = 0$: every tangent vector to C is a zero eigenvector of the Hessian. The Hessian therefore *must* have at least $\dim C$ zero eigenvalues. Bott's condition asks that it have no more.

Working definition 7.12 (non-degenerate critical submanifold (Bott))

A connected submanifold $C \subset M$ is a **non-degenerate critical manifold** of f if

1. $\mathrm{d}f = 0$ at every point of C, and
2. at every $p \in C$ the kernel of $Hf(p)$ [the vectors v with $H_{ij}v^j = 0$] is exactly T_pC [the velocities of curves that stay inside C].

Equivalently, the Hessian restricted to the directions normal to C [directions leaving C, complementary to T_pC inside T_pM] is non-degenerate [Nash p. 199]. *Operational test:* at each point of C, the nullity $\dim M - \operatorname{rank} Hf(p)$ equals $\dim C$. A function all of whose critical points lie on non-degenerate critical manifolds is a **Morse–Bott function**. (An isolated non-degenerate critical point is the case $\dim C = 0$.)

For the flat torus, $\dim C = 1$ and the nullity is 1 everywhere on C and C': both are non-degenerate critical circles.

The normal bundle and its splitting. At each $p \in C$ the **normal space** $N_p(C)$ is the orthogonal complement of T_pC in T_pM (for some metric; without a metric, take the quotient T_pM/T_pC). Together they form the **normal bundle** $N(C) \to C$, a vector bundle of rank $n - \dim C$. The Hessian is a non-degenerate symmetric form on each fibre N_p, so its eigenvectors split each fibre into a positive and a negative part:

$$N(C) = N^+(C) \oplus N^-(C) \tag{7.51}$$

[Nash p. 199, (7.26)], where N^- is spanned by the eigenvectors of the Hessian with negative eigenvalues and N^+ by those with positive eigenvalues.

Working definition 7.13 (index of a critical manifold)

The **index** of C is $\lambda_C = \operatorname{rank} N^-(C)$, the number of negative eigenvalues of the Hessian at any point of C [this number is the same at every $p \in C$, as shown below]: the number of independent directions in which one can roll off C downhill.

Why does λ_C not depend on the point $p \in C$? The eigenvalues of $Hf(p)$ vary continuously with p. On the normal directions none of them is ever zero, so none can change sign as p moves along the connected manifold C. The number of negative ones is locally constant, hence constant.

Key idea: The Morse–Bott lemma

Near a non-degenerate critical manifold C of index λ_C there are coordinates (c, u, v), with c coordinates along C, $u \in \mathbb{R}^{\lambda_C}$ along the fibres of N^- and v along the fibres of N^+, in which

$$f = f(C) - |u|^2 + |v|^2 . \tag{7.52}$$

The function does not depend on c at all: near C, f is "the Morse lemma in the normal directions, repeated at every point of C".

The proof is the Morse lemma of Section 7.1.4 done with the coordinates c carried along as parameters: Step 1 (writing f as a quadratic form with variable coefficients) only uses the normal coordinates, since f and $\mathrm{d}f$ vanish identically on C, and Step 2 (completing squares) works with c-dependent coefficients because the normal Hessian is invertible everywhere on C.

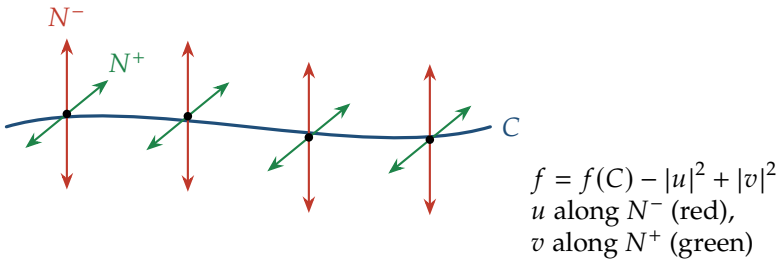


Figure 7.19: The local picture near a critical manifold C. At every point of C the normal space splits into a downhill part N^- (red) and an uphill part N^+ (green); along C itself f is constant. The index λ_C is the number of red directions. As we move along C the fibres twist, which is why N^- is a *bundle*, possibly non-trivial.

7.5.3 What a critical manifold contributes to the Morse series

Key idea: Bott's contribution (Nash (7.27)–(7.28))

A non-degenerate critical manifold C of index λ_C contributes

$$t^{\lambda_C} P_t(C) \tag{7.53}$$

to the Morse series, where $P_t(C)$ is the Poincaré polynomial of C. The **Morse–Bott series** is

$$M_t(f) = \sum_{\text{critical manifolds } C} t^{\lambda_C} P_t(C). \tag{7.54}$$

When C is a point, $P_t(C) = 1$ and we get back t^{λ_p}.

There are two ways to see why the whole Poincaré polynomial of C appears. Both are worth knowing.

Mechanism 1: tilt the function. Let g be a Morse function on C, extended to a small neighbourhood of C so that it does not depend on the normal coordinates. Consider $f_\varepsilon = f + \varepsilon g$. In the Morse–Bott coordinates,

$$f_\varepsilon = f(C) - |u|^2 + |v|^2 + \varepsilon\, g(c).$$

Its critical points near C need $u = v = 0$ (from ∂_u, ∂_v) and $\mathrm{d}g(c) = 0$ (from ∂_c). So they are exactly the critical points q of g on C, and at such a point the Hessian is block diagonal,

$$Hf_\varepsilon(q) = \operatorname{diag}\big(\underbrace{-2,\dots,-2}_{\lambda_C},\ \underbrace{+2,\dots,+2}_{\operatorname{rank} N^+},\ \varepsilon\, Hg(q)\big), \tag{7.55}$$

so the index of q as a critical point of f_ε is $\lambda_C + \lambda_q(g)$. The tilted function is Morse, and C has been replaced by the critical points of g, each shifted up in index by λ_C. Its contribution to $M_t(f_\varepsilon)$ is $t^{\lambda_C} M_t(g)$. If we choose g perfect on C, this is $t^{\lambda_C} P_t(C)$.

Worked example 7.29: Tilting the dinner plate

On the flat torus take $g = x = (R + r\cos\theta)\cos\phi$ (the tilt you get by lifting one side of the plate slightly). On the top circle C, $g = R\cos\phi$, which has a maximum at $\phi = 0$ (index 1 on the circle) and a minimum at $\phi = \pi$ (index 0). Likewise on C'. So the tilted function has four critical points:

circle	λ_C	point on circle	$\lambda_q(g\vert_C)$	index in f_ε
C (top)	1	$\phi = 0$	1	2
C (top)	1	$\phi = \pi$	0	1
C' (bottom)	0	$\phi = 0$	1	1
C' (bottom)	0	$\phi = \pi$	0	0

The indices are $0, 1, 1, 2$, as for the upright torus. That is no surprise: tilting the plate further and further ends with the torus standing on its end. The top circle contributed $t^1 + t^2 = t(1 + t) = t\,P_t(S^1)$ and the bottom one $1 + t = P_t(S^1)$.

Mechanism 2: attach a disc bundle (the Thom isomorphism). Redo the handle argument of Section 7.3.4 with the Morse–Bott normal form. As the level rises through $f(C)$, what gets attached to $M_{f(C)-\varepsilon}$ is no longer a single disc D^λ but a disc D^{λ_C} in the fibre of N^- over *every* point of C: the **disc bundle** $D(N^-)$, glued along its **sphere bundle** $S(N^-)$ (the boundary spheres of the fibre discs). The relative cohomology of this pair is given by the **Thom isomorphism**: if N^- is orientable,

$$H^k\big(D(N^-), S(N^-)\big) \cong H^{k-\lambda_C}(C), \qquad \text{i.e.} \qquad P_t\big(D(N^-), S(N^-)\big) = t^{\lambda_C} P_t(C). \tag{7.56}$$

Each class on C gets "lifted" by λ_C degrees by wedging it with a form on the fibre that integrates to 1 over each disc and vanishes near the boundary (the Thom form). When C is a point this is (7.31). Now substitute (7.56) for (7.31) in the telescoping argument (7.33). Nothing else changes, and out come the Morse–Bott inequalities.

Worked example 7.30: The Thom isomorphism for a circle with a trivial line bundle

Let $C = S^1$ and $N^- = S^1 \times \mathbb{R}$ (rank 1, trivial). Then $X = D(N^-) = S^1 \times [-1, 1]$ is an annulus and $Y = S(N^-) = S^1 \times \{\pm 1\}$ is its two boundary circles. The exact sequence of the pair (Section 7.3.8) reads, with real coefficients,

$$0 \to H^0(X,Y) \to \underbrace{H^0(X)}_{\mathbb{R}} \xrightarrow{r_0} \underbrace{H^0(Y)}_{\mathbb{R}^2} \to H^1(X,Y) \to \underbrace{H^1(X)}_{\mathbb{R}} \xrightarrow{r_1} \underbrace{H^1(Y)}_{\mathbb{R}^2} \to H^2(X,Y) \to 0.$$

Here $r_0(a) = (a, a)$ (a constant restricts to the same constant on both circles) and $r_1(b) = (b, b)$ (the loop around the annulus is homotopic to each boundary circle). Both are injective

with rank 1. So

$$\dim H^0(X,Y) \overset{(1)}{=} 0, \qquad \dim H^1(X,Y) \overset{(2)}{=} (2-1)+0 = 1, \qquad \dim H^2(X,Y) \overset{(3)}{=} 2-1 = 1.$$

(1) $H^0(X,Y)$ is the kernel of r_0, which is injective.
(2) $H^1(X,Y)$ receives the cokernel of r_0 (dimension $2-1$) and maps onto the kernel of r_1 (dimension 0).
(3) $H^2(X,Y)$ is the cokernel of r_1.

Hence $P_t(X,Y) = t + t^2 = t(1+t) = t^1 P_t(S^1)$, exactly (7.56) with $\lambda_C = 1$.

Watch out: The negative bundle must be orientable, or use $\mathbb{Z}_2$

Nash says C is a "connected oriented" submanifold. What matters for (7.56) is really the orientability of N^-. Take instead N^- to be the Möbius line bundle over S^1. Then $X = D(N^-)$ is a Möbius band and $Y = S(N^-)$ is its boundary, a *single* circle that runs twice around the core. Now $r_0 : \mathbb{R} \to \mathbb{R}$ is an isomorphism, and $r_1 : \mathbb{R} \to \mathbb{R}$ is multiplication by 2 (the boundary wraps twice), also an isomorphism over $\mathbb{R}$. Every relative group vanishes: $P_t(X,Y;\mathbb{R}) = 0 \neq t(1+t)$. Over $\mathbb{Z}_2$, multiplication by 2 is zero, so $H^1(X,Y;\mathbb{Z}_2) \cong \mathbb{Z}_2$ and $H^2(X,Y;\mathbb{Z}_2) \cong \mathbb{Z}_2$: $P_t(X,Y;\mathbb{Z}_2) = t + t^2$, as (7.56) predicts. So: with $\mathbb{Z}_2$ coefficients the Morse–Bott formula always holds; with real coefficients it needs N^- orientable (or else $P_t(C)$ must be replaced by the cohomology of C twisted by the orientation of N^-).

7.5.4 The Morse–Bott inequalities and the examples

Key idea: Morse–Bott inequalities (Nash (7.30))

For a Morse–Bott function f on a compact M, and any field $\mathbf{F}$ (provided every $N^-(C)$ is orientable when $\mathbf{F}$ does not have characteristic 2),

$$M_t(f) = \sum_C t^{\lambda_C} P_t(C;\mathbf{F}) \;\geq\; P_t(M;\mathbf{F}), \qquad M_t(f) - P_t(M;\mathbf{F}) = (1+t)R(t),\ R \geq 0. \tag{7.57}$$

Setting $t = -1$: $\sum_C (-1)^{\lambda_C} \chi(C) = \chi(M)$.

Worked example 7.31: The dinner-plate torus is perfect (Nash (7.29))

With $P_t(C) = P_t(C') = 1 + t$, $\lambda_C = 1$, $\lambda_{C'} = 0$:

$$M_t(f) \overset{(1)}{=} t^{\lambda_{C'}} P_t(C') + t^{\lambda_C} P_t(C) \overset{(2)}{=} (1+t) + t(1+t) = 1 + 2t + t^2 \overset{(3)}{=} P_t(T^2).$$

(1) The Morse–Bott series (7.54) with the two critical circles.
(2) $P_t(S^1) = 1 + t$ and the indices from Example 7.28.
(3) Example 7.9.

The degenerate height function is, in Bott's sense, perfect: "the height function is restored to perfection".

Worked example 7.32: The square of the height on S^2 (Nash (7.31)–(7.33))

Let $f = z^2$ on $S^2 = \{x^2 + y^2 + z^2 = 1\}$. Use the cylindrical coordinates (z, ϕ), valid away from the poles. There $\partial_\phi f = 0$ and $\partial_z f = 2z$, so the critical set away from the poles is the

equator $z = 0$. At the poles use (x, y): near $z = \pm 1$, $z^2 = 1 - x^2 - y^2$ exactly, so the poles are critical with Hessian $\mathrm{diag}(-2, -2)$.

- The poles $z = \pm 1$: maxima, index 2, each contributing $t^2 P_t(\mathrm{pt}) = t^2$.
- The equator $E = \{z = 0\}$: in the coordinates (z, ϕ) the Hessian is $\mathrm{diag}(2, 0)$. The zero eigenvector ∂_ϕ is tangent to E and the normal one is positive, so E is a non-degenerate critical circle of index 0 (a circle of minima, $f = 0$), contributing $t^0 P_t(S^1) = 1 + t$.

Therefore

$$M_t(f) \overset{(1)}{=} (1 + t) + t^2 + t^2 = 1 + t + 2t^2, \tag{7.58}$$

$$M_t(f) - P_t(S^2) \overset{(2)}{=} (1 + t + 2t^2) - (1 + t^2) = t + t^2 \overset{(3)}{=} (1 + t)\, t. \tag{7.59}$$

(1) Morse–Bott series: equator plus two poles.
(2) $P_t(S^2) = 1 + t^2$.
(3) Factor.

So $R(t) = t$: the function is *not* perfect, and the surplus is one cancelling pair of indices $(1, 2)$, exactly as the strong inequalities demand. Euler check: $\chi(S^1) + \chi(\mathrm{pt}) + \chi(\mathrm{pt}) = 0 + 1 + 1 = 2 = \chi(S^2)$.

Where is the cancelling pair? Flood the sphere. Just above $f = 0$ the sublevel set $\{z^2 \le c\}$ is a band $|z| \le \sqrt{c}$ around the equator, an annulus: it has $b_0 = 1$ and $b_1 = 1$, the "$1 + t$" of the equator. At $f = 1$ both polar caps are attached at once. The first cap fills in one end of the band and *kills* the 1-cycle (the equator now bounds the cap): that cap is paired with the t from the equator. The second cap closes the sphere and creates $b_2 = 1$. The pair "t from the equator, t^2 from one pole" is the $(1 + t)\, t$ of (7.59). Compare the two-humped sphere (Example 7.13), which had the same Morse series for a different reason.

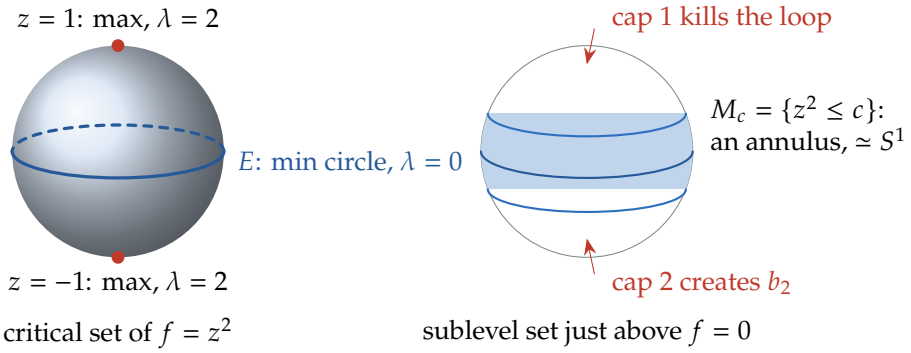


Figure 7.20: The function $f = z^2$ on S^2. Left: its critical set is the equator (a circle of minima, blue) and the two poles (maxima, red). Right: the sublevel set just above the minimum is a band around the equator, which already has a 1-cycle. The two caps added at $f = 1$ play different roles: one kills the 1-cycle (the cancelling pair $t \cdot (1 + t)$), the other creates the 2-cycle.

Worked example 7.33: A degenerate $\mathbb{CP}^2$ function: a critical $\mathbb{CP}^1$

Return to $\mathbb{CP}^2$ and let two of the a's coincide: $f = |z_3|^2 / (|z_1|^2 + |z_2|^2 + |z_3|^2)$, i.e. $a = (0, 0, 1)$. The Lagrange condition $(a_i - \mu) z_i = 0$ of Section 7.4.2 now allows z_1, z_2 both nonzero (with

$\mu = 0$, $z_3 = 0$), or z_3 alone nonzero ($\mu = 1$). So the critical set is

$$C = \{z_3 = 0\} = \{[z_1 : z_2 : 0]\} \cong \mathbb{CP}^1, \qquad f = 0, \qquad \text{and} \qquad P_3 = [0 : 0 : 1], \qquad f = 1.$$

At a point of C, say $[1 : 0 : 0]$, the expansion (7.38) gives $f \approx 0 \cdot |w_2|^2 + 1 \cdot |w_3|^2$: zero along w_2 (tangent to C), positive along w_3 (normal). So C is a non-degenerate critical manifold of index 0 (by the U(2) symmetry mixing z_1, z_2, every point of C looks the same). At P_3 the expansion gives $1 - |w_1|^2 - |w_2|^2$: index 4.

$$M_t(f) \overset{(1)}{=} t^0 P_t(\mathbb{CP}^1) + t^4 \overset{(2)}{=} 1 + t^2 + t^4 = P_t(\mathbb{CP}^2).$$

(1) The Morse–Bott series: the critical $\mathbb{CP}^1$ and the point P_3.
(2) $P_t(\mathbb{CP}^1) = P_t(S^2) = 1 + t^2$.

Perfect. Letting eigenvalues of $A = \operatorname{diag}(a_i)$ coincide turns isolated critical points into critical projective spaces (the eigenspaces), and the Morse–Bott series does not notice.

Worked example 7.34: $f = z^2$ on $\mathbb{RP}^2$: two fields, two answers

$f = z^2$ is even under $x \mapsto -x$, so it descends to $\mathbb{RP}^2 = S^2/\pm$. The two poles become one point (index 2). The equator becomes the equator modulo antipodes, which is again a circle (going halfway round the equator closes up), of index 0. The negative bundle of the circle has rank 0, so orientability is no issue. Then

$$M_t(f) \overset{(1)}{=} (1 + t) + t^2,$$
$$\text{over } \mathbb{Z}_2: \; M_t - P_t \overset{(2)}{=} (1 + t + t^2) - (1 + t + t^2) = 0,$$
$$\text{over } \mathbb{R}: \; M_t - P_t \overset{(3)}{=} (1 + t + t^2) - 1 = t(1 + t).$$

(1) One critical circle of index 0 and one point of index 2.
(2) $P_t(\mathbb{RP}^2; \mathbb{Z}_2) = 1 + t + t^2$ from (7.14).
(3) $P_t(\mathbb{RP}^2; \mathbb{R}) = 1$.

Perfect over $\mathbb{Z}_2$, not over $\mathbb{R}$: the same pattern as the Morse function of Example 7.16, as it must be (torsion obstructs perfection).

Try it: Quadratic functions on S^n

On $S^n \subset \mathbb{R}^{n+1}$ let $f = x_1^2 + \cdots + x_k^2$ with $1 \le k \le n$. Find the critical manifolds, their indices and $R(t)$.
Answer: the minimum set $\{x_1 = \cdots = x_k = 0\} \cong S^{n-k}$ ($f = 0$, index 0) and the maximum set $\{x_{k+1} = \cdots = x_{n+1} = 0\} \cong S^{k-1}$ ($f = 1$, index $n - k + 1$, the codimension). $M_t = (1 + t^{n-k}) + t^{n-k+1}(1 + t^{k-1}) = 1 + t^{n-k} + t^{n-k+1} + t^n$, so $M_t - P_t(S^n) = t^{n-k}(1 + t)$ and $R = t^{n-k}$. For $n = 2$, $k = 1$ this is x^2 on S^2, which is the z^2 example rotated. (For $k = 1$ the maximum set S^0 is two points.)

Worked example 7.35: The extreme case: a constant function

If f is constant, the whole of M is critical, $C = M$, the normal bundle has rank 0, the Hessian is (vacuously) non-degenerate on it, and $\lambda_M = 0$. So $M_t(f) = P_t(M)$: a constant is a perfect Morse–Bott function. This is a useless function, but it is a good check that the

formula (7.54) interpolates correctly between "all critical points isolated" and "everything critical".

Reading Nash (pp. 199–201, (7.26)–(7.33))

Nash motivates §2 with the torus on a dinner plate, defines a non-degenerate critical manifold ("$\mathrm{d}f = 0$ everywhere on C and the Hessian is non-degenerate in the directions normal to C"), splits the normal bundle $N(C) = N^+ \oplus N^-$ [Nash (7.26)] (his text says $N^{\mp}$ is spanned by the positive and negative eigenvectors; the sign labels should be read the other way round, N^- being the negative part, since $\lambda_C = \operatorname{rank} N^-$), and quotes Bott's contribution $t^{\lambda_C} P_t(C)$ [Nash (7.27)] and the series [Nash (7.28)]. The dinner plate gives [Nash (7.29)], the inequalities are [Nash (7.30)], and $f = z^2$ on S^2 is [Nash (7.31)–(7.33)]. We added: why tangent directions are always null (7.50), the constancy of λ_C, the Morse–Bott lemma, both mechanisms for $t^{\lambda_C} P_t(C)$ (tilting and the Thom isomorphism, with the annulus and Möbius computations), the flooding picture of z^2, and the $\mathbb{CP}^2$, $\mathbb{RP}^2$ and S^n examples.

Recap

- Symmetric functions have critical families. Tangent directions to a critical family are always zero modes of the Hessian.
- Non-degenerate critical manifold: the Hessian's kernel is exactly TC. Normal bundle $N = N^+ \oplus N^-$; index $\lambda_C = \operatorname{rank} N^-$.
- Morse–Bott lemma: $f = f(C) - |u|^2 + |v|^2$ near C.
- C contributes $t^{\lambda_C} P_t(C)$: tilt by a perfect g on C, or attach the disc bundle $D(N^-)$ (Thom isomorphism; needs N^- orientable over $\mathbb{R}$).
- Dinner-plate torus: $(1+t) + t(1+t) = P_t(T^2)$, perfect. z^2 on S^2: $1 + t + 2t^2$, $R = t$, not perfect.

7.6 Equivariant cohomology: keeping track of a symmetry

The previous section showed that a symmetry produces critical *families*. There is a second, more ambitious thing one wants to do with a symmetry: *divide by it*. In physics we do this all the time. The configuration space of a gauge theory is the space of gauge potentials modulo gauge transformations; the configuration space of a rigid rotor is a group modulo a subgroup; the phase space of a system with a conserved angular momentum can be reduced by the rotations it generates. If f is invariant under a group G acting on M, then f really lives on the space of orbits M/G, and it is natural to hope that Morse theory on M/G is simpler.

Nash's §3 opens with an example in which this hope fails in the simplest possible way, and then explains the cure, which is due to Borel and was made into a tool for Morse theory by Atiyah and Bott. The cure is to never divide by a group that does not act freely. Instead one first replaces M by a space that has the same cohomology but on which the action *is* free, and divides that. The resulting cohomology is **equivariant cohomology**. This section builds it from scratch: group actions, universal bundles and classifying spaces, the Borel construction, and two complete computations for the rotating sphere. Section 7.7 then puts it to work in Morse theory.

7.6.1 The rotating sphere: a quotient that forgets too much

Take the height function on the unit sphere [Nash p. 201, (7.34)],

$$M = S^2 = \{(x,y,z) : x^2+y^2+z^2 = 1\}, \qquad f(x,y,z) = z. \tag{7.60}$$

We know everything about it (Example 7.6). It has two critical points, the north pole $z=1$ (index 2) and the south pole $z=-1$ (index 0), so $M_t(f) = 1+t^2 = P_t(S^2)$: it is perfect.

It is also symmetric. Let $\mathrm{U}(1) = \{e^{i\alpha}\}$ act by rotation about the z-axis,

$$e^{i\alpha}\cdot\begin{pmatrix}x\\y\\z\end{pmatrix} = \underbrace{\begin{pmatrix}\cos\alpha & -\sin\alpha & 0\\ \sin\alpha & \cos\alpha & 0\\ 0 & 0 & 1\end{pmatrix}}_{R_\alpha}\begin{pmatrix}x\\y\\z\end{pmatrix}. \tag{7.61}$$

Writing $w = x+iy$, this is simply $(w,z)\mapsto(e^{i\alpha}w, z)$. The height is invariant, because the third row of R_α is $(0,0,1)$:

$$f\big(R_\alpha(x,y,z)\big) \overset{(1)}{=} 0\cdot x + 0\cdot y + 1\cdot z = f(x,y,z).$$

(1) The third component of $R_\alpha(x,y,z)^{\mathsf{T}}$ is the third row of R_α applied to (x,y,z).

So f is constant along each orbit, and ought to define a function on the space of orbits. What is that space? Two points lie on the same orbit exactly when they have the same height z (the rotations can move (x,y) anywhere on the circle $x^2+y^2 = 1-z^2$ but cannot change z). So the map

$$q\colon S^2 \to [-1,1], \qquad (x,y,z)\mapsto z, \tag{7.62}$$

identifies the orbit space:

- over $-1<z<1$ the orbit is the latitude circle of radius $\sqrt{1-z^2}$;
- over $z=\pm1$ the orbit is a single point, the pole.

Hence

$$S^2/\mathrm{U}(1) \;\cong\; [-1,1] \;\cong\; [0,1] \tag{7.63}$$

[Nash (7.35)]. An interval is contractible, so

$$P_t\big(S^2/\mathrm{U}(1)\big) = 1. \tag{7.64}$$

The hope is dashed. The quotient has trivial cohomology, and the function $f_G(z) = z$ on $[-1,1]$ has no critical points at all: it is monotonic, and its extreme values sit on the boundary. Morse theory on the quotient knows nothing about the sphere.

Worked example 7.36: Where did the critical points go?

Look near the north pole. Use $w = x+iy$ as a coordinate there, so that $z = \sqrt{1-|w|^2}$. The invariant combination of w and $\bar w$ is $\rho = |w|^2 = w\bar w$; it is a good coordinate on the quotient near the endpoint, with $\rho\ge 0$ and $\rho = 0$ at the pole. Then

$$f \overset{(1)}{=} \sqrt{1-|w|^2} \overset{(2)}{=} 1 - \tfrac12|w|^2 - \tfrac18|w|^4 - \dots \overset{(3)}{=} 1 - \tfrac12\rho - \tfrac18\rho^2 - \dots$$

(1) On the upper hemisphere $z = +\sqrt{1-x^2-y^2}$.

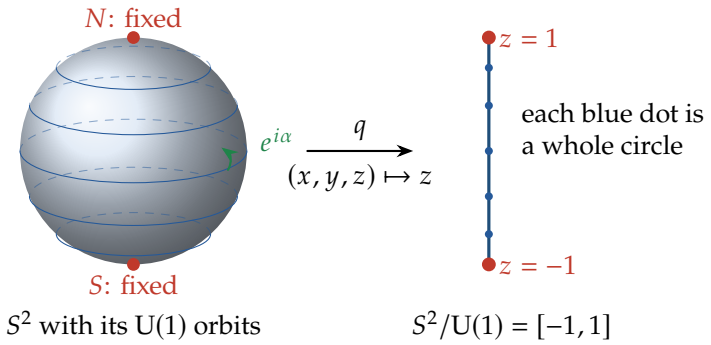


Figure 7.21: Dividing the sphere by rotations. Each latitude circle (an orbit) is collapsed to one point of the interval; the poles, which the rotations do not move, go to the two endpoints. The quotient is contractible, so all the topology of S^2 (its 2-cycle) has been lost. The culprits are the two fixed points.

(2) Taylor series $\sqrt{1-s} = 1 - \frac{1}{2}s - \frac{1}{8}s^2 - \dots$ with $s = |w|^2$.
(3) Substitute the invariant $\rho = |w|^2$.

On S^2 the pole is a non-degenerate maximum: in terms of $w = x + iy$, $f \approx 1 - \frac{1}{2}(x^2 + y^2)$, two negative eigenvalues. On the quotient it is the boundary point $\rho = 0$ of the half-line $\rho \geq 0$, and there $\mathrm{d}f_G/\mathrm{d}\rho = -\frac{1}{2} \neq 0$. *A non-degenerate critical point has turned into a boundary point with a nonzero slope.* The quadratic function $-|w|^2$ is linear in the quotient coordinate $|w|^2$, so the information that it had two negative directions has disappeared.

Physical picture: Gauge fixing near a symmetric configuration

This is the finite-dimensional cartoon of a familiar problem in gauge theory. Gauge-fixing works well near a configuration that the gauge group moves freely. Near a configuration with a nontrivial stabiliser (in gauge theory: a *reducible* connection, such as the trivial connection $A = 0$, which is fixed by all constant gauge transformations), the gauge orbits change dimension, the space of orbits develops a singularity, and naive quotients lose information. The equivariant cohomology built below is the topologists' version of "keep the gauge symmetry and add ghosts" instead of "fix the gauge".

7.6.2 Group actions, orbits and stabilisers

We need a little vocabulary. Throughout, G is a compact Lie group (usually U(1) or a finite group) acting smoothly on M.

Working definition 7.14 (action, orbit, stabiliser, free action)

A (left) [h acts first and then g, as the rule $g \cdot (h \cdot x) = (gh) \cdot x$ below says] **action** of G on M is a smooth map $G \times M \to M$, $(g, x) \mapsto g \cdot x$, with $1 \cdot x = x$ and $g \cdot (h \cdot x) = (gh) \cdot x$. In practice: a group of symmetry transformations, written as explicit matrices or maps.

- The **orbit** of x is $G \cdot x = \{g \cdot x : g \in G\}$: everywhere the symmetry can move x.
- The **stabiliser** (isotropy group) of x is $G_x = \{g : g \cdot x = x\}$: the symmetries that do not move x.
- x is a **fixed point** if $G_x = G$.

- The action is **free** if every stabiliser is trivial, $G_x = \{1\}$: no group element other than 1 fixes any point.
- The **orbit space** M/G is the set of orbits, with the topology in which a set of orbits is open when the union of those orbits is open in M [so orbits passing close to each other in M are close in M/G].

Test for freeness: solve $g \cdot x = x$ for g; if the only solution is $g = 1$ for every x, the action is free.

The map $g \mapsto g \cdot x$ from G onto the orbit takes the same value on g and gh for every $h \in G_x$. So the orbit is a copy of the coset space:

$$G \cdot x \;\cong\; G/G_x, \qquad \dim(G \cdot x) = \dim G - \dim G_x. \tag{7.65}$$

This **orbit–stabiliser relation** explains Figure 7.21: at a generic point of S^2 the stabiliser is trivial and the orbit is a circle ($\mathrm{U}(1)/\{1\}$); at a pole the stabiliser is all of $\mathrm{U}(1)$ and the orbit is a point ($\mathrm{U}(1)/\mathrm{U}(1)$). When the orbit dimension jumps, the quotient cannot be a manifold of one fixed dimension near the jump.

Worked example 7.37: A catalogue of actions

G	acting on	by	stabilisers, quotient
$\mathrm{U}(1)$	S^2	rotation (7.61)	$\mathrm{U}(1)$ at poles; $[-1,1]$
$\mathrm{U}(1)$	$\mathbb{C}$	$z \mapsto e^{i\alpha}z$	$\mathrm{U}(1)$ at 0; $[0,\infty)$ (a cone)
$\mathbb{Z}_2$	$\mathbb{R}$	$x \mapsto -x$	$\mathbb{Z}_2$ at 0; $[0,\infty)$
$\mathbb{Z}_2$	S^n	$x \mapsto -x$	free; $\mathbb{RP}^n$
$\mathrm{U}(1)$	$S^{2n+1} \subset \mathbb{C}^{n+1}$	$z \mapsto e^{i\alpha}z$	free; $\mathbb{CP}^n$
$\mathrm{U}(1)$	S^1	$e^{i\beta} \mapsto e^{i(\alpha+\beta)}$	free; a point
$\mathrm{U}(1)$	$S^3 \subset \mathbb{C}^2$	$(z_1,z_2) \mapsto (e^{i\alpha}z_1, e^{2i\alpha}z_2)$	$\mathbb{Z}_2$ on $z_1 = 0$; a "teardrop"

Two of these deserve a calculation.
The antipodal map is free. If $-x = x$ then $2x = 0$, so $x = 0$, which is not on the sphere. Hence no point is fixed by the non-trivial element.
The weighted action on S^3. Solve $(e^{i\alpha}z_1, e^{2i\alpha}z_2) = (z_1, z_2)$. If $z_1 \neq 0$ we need $e^{i\alpha} = 1$, so the stabiliser is trivial. If $z_1 = 0$ then $|z_2| = 1$ and we only need $e^{2i\alpha} = 1$, i.e. $e^{i\alpha} = \pm 1$: stabiliser $\mathbb{Z}_2$. The quotient is still a topological sphere, but with one conical point (the image of the circle $z_1 = 0$). Finite stabilisers produce mild "orbifold" singularities; positive-dimensional stabilisers, as at the poles of S^2, produce boundaries and cones of lower dimension.

Key idea: Free actions have smooth quotients (cf. Nash Ch. 1 and Ch. 5)

If a compact Lie group G acts smoothly and freely on M, then M/G is a smooth manifold of dimension $\dim M - \dim G$, and the projection $M \to M/G$ is a principal G-bundle: every fibre is a copy of G. If the action has fixed points or other non-trivial stabilisers, M/G is in general singular (boundaries, cones, orbifold points), as the table shows.

For a free action there is no problem in doing Morse theory downstairs [Nash p. 201]. A G-invariant f defines $f_G\colon M/G \to \mathbb{R}$ by $f_G([x]) = f(x)$, which is well defined because f is constant on orbits. In a local trivialisation $M \supset U \cong V \times G$ (with $V \subset M/G$), $f(v,g) = f_G(v)$

does not depend on g. So the critical points of f_G are the orbits on which $\mathrm{d}f = 0$, and each such orbit is a critical manifold of f of dimension $\dim G$, non-degenerate (in Bott's sense) exactly when the critical point of f_G is non-degenerate, and of the same index.

Worked example 7.38: Upstairs on S^{2n+1}, downstairs on $\mathbb{CP}^n$

Take Nash's function on $\mathbb{CP}^n$ (7.36) and read it upstairs, as a U(1)-invariant function on S^{2n+1}:

$$f(z) = \sum_{i=1}^{n+1} a_i \left|z_i\right|^2, \qquad a_1 < a_2 < \cdots < a_{n+1}, \qquad \sum |z_i|^2 = 1.$$

Downstairs we found $n+1$ critical points of index $0, 2, \ldots, 2n$ (Section 7.4.2). Upstairs each of them is a whole orbit: the circle $C_i = \{|z_i| = 1,\ z_j = 0\ (j \neq i)\}$. Near C_i the normal coordinates are z_j, $j \neq i$ ($2n$ real coordinates; the circle direction is tangent and the radial direction in z_i is not on the sphere), and

$$f \overset{(1)}{=} a_i\Big(1 - \sum_{j\neq i} \left|z_j\right|^2\Big) + \sum_{j\neq i} a_j \left|z_j\right|^2 \overset{(2)}{=} a_i + \sum_{j\neq i} (a_j - a_i)\, \left|z_j\right|^2.$$

(1) On the sphere $|z_i|^2 = 1 - \sum_{j\neq i} \left|z_j\right|^2$.
(2) Collect terms: this is already the Morse–Bott normal form (7.52), with no dependence on the phase of z_i (the coordinate along C_i).

The coefficient $a_j - a_i$ is negative for the $i-1$ values $j < i$, each giving two real negative directions, so $\lambda_{C_i} = 2(i-1)$. The Morse–Bott series on S^{2n+1} is

$$M_t(f) \overset{(1)}{=} \sum_{i=1}^{n+1} t^{2(i-1)} P_t(S^1) = (1+t)(1+t^2+\cdots+t^{2n}) = 1 + t + t^2 + \cdots + t^{2n+1},$$

$$M_t(f) - P_t(S^{2n+1}) \overset{(2)}{=} t + t^2 + \cdots + t^{2n} \overset{(3)}{=} (1+t)\,(t + t^3 + \cdots + t^{2n-1}).$$

(1) Each critical circle contributes $t^{\lambda_{C_i}} P_t(S^1)$ (7.54).
(2) $P_t(S^{2n+1}) = 1 + t^{2n+1}$ removes the first and last terms.
(3) Group the terms in pairs $t^{2k-1} + t^{2k} = (1+t)t^{2k-1}$, $k = 1, \ldots, n$.

So upstairs f is far from perfect ($R = t + t^3 + \cdots + t^{2n-1}$), while downstairs f_G is perfect, $M_t(f_G) = 1 + t^2 + \cdots + t^{2n} = P_t(\mathbb{CP}^n)$. Dividing by a *free* action threw away the redundant circle directions and, with them, all the cancelling pairs. This is what we would like to do for non-free actions too.

7.6.3 Universal bundles and classifying spaces

Borel's idea is simple to state. Find a space EG that is

1. *topologically trivial*: contractible, so that $M \times EG$ has the same cohomology as M; and
2. *a free G-space*: G acts on it freely.

Then G acts freely on $M \times EG$ (we check this below), so we can divide $M \times EG$ by G without any singularity, even if the action on M was far from free. The price is that EG is necessarily infinite-dimensional, but it is an entirely explicit space.

Working definition 7.15 (universal bundle, classifying space)

A **universal G-bundle** is a contractible space EG with a free G-action. Its quotient $BG = EG/G$ is the **classifying space** of G, and $EG \to BG$ is a principal G-bundle [the fibre over each point of BG is one orbit, a copy of G] [Nash p. 202]. *In practice*: find a sequence of free G-spaces $E_1 \subset E_2 \subset \cdots$ whose cohomology vanishes in degrees $1, \dots, N$ for E_N, with $N \to \infty$; take EG to be their union.

(In Nash's Chapter 1, BG appears as the space that classifies principal G-bundles: every principal G-bundle $P \to X$ is the pull-back of $EG \to BG$ along some map $X \to BG$, unique up to homotopy. We shall come back to this meaning below; for equivariant cohomology the two properties above are what matter.)

A finite-dimensional sphere is not contractible, but it is "contractible up to its dimension": S^m has no cohomology in degrees $1, \dots, m-1$. So we can use spheres of ever larger dimension and take the limit

$$S^\infty = \bigcup_m S^m = \Big\{(x_1, x_2, \dots) : \text{ finitely many } x_i \neq 0,\ \textstyle\sum x_i^2 = 1\Big\}.$$

Worked example 7.39: The infinite sphere is contractible

In finite dimensions you cannot shrink a sphere to a point within the sphere. In infinite dimensions you can, because there is always a spare direction to move into. Let $S(x_1, x_2, \dots) = (0, x_1, x_2, \dots)$ be the shift, and $e_1 = (1, 0, 0, \dots)$. Consider the two straight-line homotopies, each divided by its length to stay on the sphere,

$$h_s(x) = \frac{(1-s)\,x + s\,Sx}{\|(1-s)\,x + s\,Sx\|}, \qquad k_s(x) = \frac{(1-s)\,Sx + s\,e_1}{\|(1-s)\,Sx + s\,e_1\|}, \qquad 0 \le s \le 1.$$

The first deforms the identity into the shift, the second deforms the shift into the constant map to e_1. Together they contract S^∞ to the point e_1. The only thing to check is that neither numerator ever vanishes.

- For h_s: let x_k be the first nonzero coordinate of x. The k-th coordinate of Sx is $x_{k-1} = 0$, so the k-th coordinate of the numerator is $(1-s)x_k$, nonzero for $s < 1$; at $s = 1$ the numerator is $Sx \neq 0$.
- For k_s: the first coordinate of Sx is 0, so the first coordinate of the numerator is s, and the rest is $(1-s)Sx$. Both vanish only if $s = 0$ and $s = 1$ at once, which is impossible.

In a finite-dimensional sphere the shift is not available (the last coordinate would fall off the end), and indeed S^m is not contractible.

Now let the groups act.

Worked example 7.40: Three classifying spaces

(a) $G = \mathbb{Z}$. $\mathbb{Z}$ acts on $\mathbb{R}$ by translations, $n \cdot x = x + n$. This is free ($x + n = x$ forces $n = 0$) and $\mathbb{R}$ is contractible. So $E\mathbb{Z} = \mathbb{R}$ and $B\mathbb{Z} = \mathbb{R}/\mathbb{Z} = S^1$.
(b) $G = \mathbb{Z}_2$. $\mathbb{Z}_2 = \{\pm 1\}$ acts on S^∞ by $x \mapsto -x$, freely (Example 7.37). So $E\mathbb{Z}_2 = S^\infty$ and $B\mathbb{Z}_2 = S^\infty/\pm = \mathbb{RP}^\infty = \bigcup_n \mathbb{RP}^n$. From the cell structure of $\mathbb{RP}^n$ (one cell in each dimension, Example 7.10),

$$P_t(\mathbb{RP}^\infty; \mathbb{Z}_2) = 1 + t + t^2 + \cdots = \frac{1}{1-t}, \qquad P_t(\mathbb{RP}^\infty; \mathbb{R}) = 1.$$

(c) $G = \mathrm{U}(1)$**.** Regard S^∞ as the unit sphere in $\mathbb{C}^\infty = \bigcup_n \mathbb{C}^{n+1}$, i.e. as the union of the odd spheres $S^1 \subset S^3 \subset S^5 \subset \cdots$. The action $z \mapsto e^{i\alpha}z$ is free on each S^{2n+1} (Example 7.37), hence on the union, and the quotients are the projective spaces of Section 7.4.1. So

$$E\mathrm{U}(1) = S^\infty, \qquad B\mathrm{U}(1) = S^\infty/\mathrm{U}(1) = \mathbb{CP}^\infty = \bigcup_n \mathbb{CP}^n.$$

The Poincaré series of $\mathbb{CP}^\infty$ is the limit of those of $\mathbb{CP}^n$ [Nash p. 204, (7.48)]:

$$P_t(\mathbb{CP}^\infty) \overset{(1)}{=} \lim_{n\to\infty} P_t(\mathbb{CP}^n) \overset{(2)}{=} \lim_{n\to\infty} \left(1 + t^2 + t^4 + \cdots + t^{2n}\right) \overset{(3)}{=} \frac{1}{1-t^2}. \tag{7.66}$$

(1) $\mathbb{CP}^\infty$ is built from the cells of all the $\mathbb{CP}^n$ (one in each even dimension); $H^k(\mathbb{CP}^\infty) = H^k(\mathbb{CP}^n)$ as soon as $2n > k$, since adding cells of dimension $> k+1$ cannot change H^k.
(2) The perfect Morse function on $\mathbb{CP}^n$, (7.42).
(3) Geometric series, read as a formal power series in t.

And the numerators of the same limit tell us that S^∞ really is a universal bundle, as Nash argues [Nash (7.49)]: S^{2n+1} has no cohomology in degrees $1, \ldots, 2n$, and as $n \to \infty$ the first non-zero degree runs away to infinity. Hence

$$B\mathrm{U}(1) = \mathbb{CP}^\infty, \qquad H^*(B\mathrm{U}(1); \mathbb{R}) = \mathbb{R}[u], \quad \deg u = 2, \qquad P_t(B\mathrm{U}(1)) = \frac{1}{1-t^2}. \tag{7.67}$$

Here u is a generator of $H^2(\mathbb{CP}^\infty)$, and u^k generates H^{2k}. (The ring statement, that the powers of u are all nonzero, is the fact that on $\mathbb{CP}^n$ the n-th power of the Fubini–Study Kähler form is a volume form.)

Worked example 7.41: More classifying spaces

(d) $G = \mathrm{SU}(2)$**.** $\mathrm{SU}(2)$ is the group of unit quaternions, $S^3 \subset \mathbb{H}$ (Chapter T3). It acts freely on the unit sphere $S^{4n+3} \subset \mathbb{H}^{n+1}$ by right multiplication, with quotient the quaternionic projective space $\mathbb{HP}^n$, which has one cell in each dimension divisible by 4. As $n \to \infty$:

$$B\mathrm{SU}(2) = \mathbb{HP}^\infty, \qquad P_t(B\mathrm{SU}(2)) = 1 + t^4 + t^8 + \cdots = \frac{1}{1-t^4}.$$

The generator in degree 4 is the second Chern class c_2 of the universal $\mathrm{SU}(2)$-bundle: the characteristic class that counts instantons (Chapter 8).
(e) $G = T^k = \mathrm{U}(1)^k$**.** Take products: $ET^k = (S^\infty)^k$, $BT^k = (\mathbb{CP}^\infty)^k$, $P_t(BT^k) = (1-t^2)^{-k}$.
(f) $G = \mathrm{U}(n)$**.** $B\mathrm{U}(n)$ is the Grassmannian of complex n-planes in $\mathbb{C}^\infty$; $H^*(B\mathrm{U}(n); \mathbb{R}) = \mathbb{R}[c_1, \ldots, c_n]$ with $\deg c_i = 2i$, the universal Chern classes, and $P_t(B\mathrm{U}(n)) = \prod_{i=1}^n (1-t^{2i})^{-1}$.

Why "classifying". Let $P \to X$ be a principal G-bundle. Choose a G-equivariant map $\Phi\colon P \to EG$ (it exists because EG is contractible, and is unique up to homotopy for the same reason). It descends to $\phi\colon X \to BG$, and P is recovered as the pull-back ϕ^*EG. Characteristic classes of P are then pull-backs $\phi^*(c)$ of classes $c \in H^*(BG)$.

Worked example 7.42: The monopole bundle inside the universal bundle

The Hopf bundle $S^3 \to S^2 = \mathbb{CP}^1$ (Example 7.21), the bundle of the Dirac monopole of unit charge, is literally the first stage $S^3 \to \mathbb{CP}^1$ of the universal bundle $S^\infty \to \mathbb{CP}^\infty$. Its classifying map is the inclusion $\phi\colon \mathbb{CP}^1 \hookrightarrow \mathbb{CP}^\infty$. The generator $u \in H^2(\mathbb{CP}^\infty)$ restricts to a

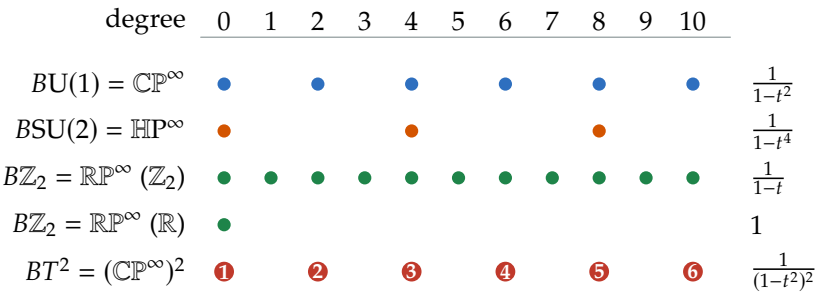


Figure 7.22: Betti numbers of some classifying spaces, degree by degree (a dot means $b_k = 1$; for BT^2 the number in the dot is b_k). Unlike an ordinary compact manifold, BG has cohomology in infinitely many degrees: this is what makes the equivariant cohomology of a single point so rich. Note that $B\mathbb{Z}_2$ is invisible with real coefficients: finite groups have no real equivariant cohomology to speak of.

generator of $H^2(\mathbb{CP}^1) = H^2(S^2)$, i.e. (up to the sign fixed by orientation conventions) to the first Chern class of the monopole bundle:

$$\int_{S^2} \phi^* u = \pm 1.$$

A monopole of charge k corresponds to a map $S^2 \to \mathbb{CP}^\infty$ of "degree" k. So the whole family of monopole bundles over S^2 is organised by the single space $\mathbb{CP}^\infty$.

Watch out: EG is not unique, but BG is unique up to homotopy

Any two contractible free G-spaces give homotopy-equivalent classifying spaces, which is all that cohomology sees. We could equally use a model of EG in which S^∞ is replaced by the unit sphere of a separable Hilbert space. When a statement about BG is proved using the finite approximations $S^{2n+1} \to \mathbb{CP}^n$, it is true in degrees $< 2n+1$ at stage n, and hence in every degree in the limit.

7.6.4 The Borel construction: the homotopy quotient

Now combine M with EG. Let G act on the product by the **diagonal action** [Nash (7.36)]

$$G \times (M \times EG) \to M \times EG, \qquad g \cdot (x, e) = (g \cdot x,\ g \cdot e). \tag{7.68}$$

This action is free, however badly G acts on M:

$$g \cdot (x, e) = (x, e) \overset{(1)}{\Longrightarrow} g \cdot e = e \overset{(2)}{\Longrightarrow} g = 1.$$

(1) Compare second components.
(2) G acts freely on EG.

So the quotient is well behaved:

Working definition 7.16 (homotopy quotient, Borel construction)

The **homotopy quotient** of M by G is

$$M_G = (M \times EG)/G \ = \ M \times_G EG \tag{7.69}$$

[EG is a contractible space on which G acts freely, such as S^∞ for U(1)] [Nash (7.37)], the set of pairs $[x, e]$ with the identification $[g \cdot x, g \cdot e] = [x, e]$. *Recipe:* to compute with M_G, replace EG by a finite approximation (for U(1): S^{2N+1}), which is honest in all degrees below $2N+1$ [it gives the same cohomology of M_G as EG does in those degrees], and let $N \to \infty$.

The space M_G has two natural projections, and each tells us something.

Projection to BG: M_G is an M-bundle. The map $\pi[x, e] = [e] \in EG/G = BG$ is well defined (it forgets x), and the fibre over $[e]$ is $\{[x, e] : x \in M\}$, which is a copy of M (for fixed e, different x give different points, since only $g = 1$ fixes e). So there is a fibre bundle [Nash (7.38)]

$$M \longrightarrow M_G \xrightarrow{\pi} BG. \tag{7.70}$$

When the action on M is trivial, $M_G = M \times BG$. In general M_G is a twisted product, the twisting being the action.

Projection to M/G: the fibres are classifying spaces of the stabilisers. The map $p[x, e] = [x] \in M/G$ is also well defined. Fix a point x in a given orbit. Every point of M_G over that orbit can be written $[x, e]$ for some e (if $[x', e']$ has $x' = g \cdot x$, then $[x', e'] = [x, g^{-1}e']$), and

$$[x, e] = [x, e'] \overset{(1)}{\iff} (x, e') = (h \cdot x, h \cdot e) \text{ for some } h \in G \overset{(2)}{\iff} e' = h \cdot e \text{ for some } h \in G_x.$$

(1) Definition of the quotient by the diagonal action.
(2) $h \cdot x = x$ says exactly that h is in the stabiliser G_x.

So the fibre of p over $[x]$ is EG/G_x. Since G_x acts freely on the contractible EG, this is a model of BG_x:

$$p^{-1}([x]) \;=\; EG/G_x \;\simeq\; BG_x. \tag{7.71}$$

Key idea: What the homotopy quotient remembers

Over a point of M/G where the action is free, M_G has a contractible fibre EG: nothing is added. Over a fixed point, M_G has the fibre BG, with all its cohomology. So

- if G acts freely, $M_G \simeq M/G$ [Nash p. 202];
- if G acts trivially, $M_G = M \times BG$;
- in general, M_G is M/G with the singular orbits "thickened" by the classifying spaces of their stabilisers.

Working definition 7.17 (equivariant cohomology)

The **equivariant cohomology** of a G-space M [a space with a smooth action of G, fixed points allowed] is the ordinary cohomology of its homotopy quotient [Nash (7.39)],

$$H^*_G(M; \mathbf{F}) := H^*(M_G; \mathbf{F}), \tag{7.72}$$

and its **equivariant Poincaré series** is [Nash p. 203, (7.42)]

$$P^G_t(M) = \sum_i \dim H^i_G(M; \mathbf{F})\, t^i = \sum_i \dim H^i(M_G; \mathbf{F})\, t^i = P_t(M_G). \tag{7.73}$$

It is a formal power series [a list of coefficients written as $\sum_i a_i t^i$, with t only a bookkeeping variable, so no convergence is needed], usually with infinitely many terms.

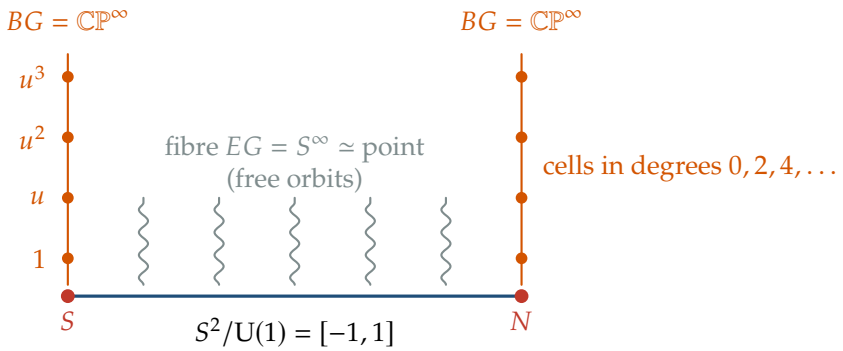


Figure 7.23: The homotopy quotient $S^2_{\mathrm{U}(1)}$ drawn over the naive quotient $[-1, 1]$. Over interior points the fibre is EG, which is contractible and adds nothing. Over each fixed point the fibre is a whole $BG = \mathbb{CP}^\infty$, with a class u^k in every even degree. The equivariant cohomology is built from the two orange columns, glued along the contractible middle: $1 + 2t^2 + 2t^4 + \cdots$.

Worked example 7.43: Four quick computations

A point [Nash (7.40)–(7.41)]. The action on $\{p\}$ is trivial, so

$$\{p\}_G \overset{(1)}{=} (\{p\} \times EG)/G \overset{(2)}{=} EG/G = BG, \qquad H^*_G(\{p\}) = H^*(BG).$$

(1) Definition (7.69).
(2) $\{p\} \times EG$ is just EG, with its own action.

For $G = \mathrm{U}(1)$: $H^*_{\mathrm{U}(1)}(\mathrm{pt}) = \mathbb{R}[u]$, $P_t^{\mathrm{U}(1)}(\mathrm{pt}) = 1/(1 - t^2)$. A point is as topologically trivial as a space can be, yet its equivariant cohomology is infinite. It remembers the symmetry group.

A free orbit. Let $\mathrm{U}(1)$ act on S^1 by rotation. The action is free, so $S^1_{\mathrm{U}(1)} \simeq S^1/\mathrm{U}(1) = \mathrm{pt}$ and $P_t^{\mathrm{U}(1)}(S^1) = 1$. More generally G acting on itself by left multiplication has $G_G = G \times_G EG \cong EG \simeq \mathrm{pt}$.

A free action. $H^*_{\mathrm{U}(1)}(S^{2n+1}) = H^*(\mathbb{CP}^n)$, so $P_t^{\mathrm{U}(1)}(S^{2n+1}) = 1 + t^2 + \cdots + t^{2n}$.

A trivial action. If G acts trivially on M, $M_G = M \times BG$ and by the Künneth formula $P_t^G(M) = P_t(M)\, P_t(BG)$. For instance, $\mathrm{U}(1)$ acting trivially on S^2 gives $(1 + t^2)/(1 - t^2)$.

The equivariant parameter. The projection $\pi\colon M_G \to BG$ pulls back classes from BG, so $H^*_G(M)$ is a module over $H^*(BG)$. For $\mathrm{U}(1)$ this means we can multiply equivariant classes by u (raising the degree by 2). How u acts is a good detector of fixed points. For a free action, $H^*_G(M) = H^*(M/G)$ vanishes in high degrees, so some power of u acts as zero (on $\mathbb{CP}^n$, $u^{n+1} = 0$). For a fixed point, u acts on $H^*_G(\mathrm{pt}) = \mathbb{R}[u]$ with no power ever zero. This is the germ of the **localisation theorem** of Atiyah and Bott: after inverting u, the equivariant cohomology of M only sees the fixed-point set.

7.6.5 The equivariant cohomology of the rotating sphere, twice

We now compute Nash's quantity [Nash (7.45)]

$$P_t^{\mathrm{U}(1)}(S^2) = P_t\big(S^2_{\mathrm{U}(1)}\big), \qquad G = \mathrm{U}(1), \tag{7.74}$$

first by Nash's route (a sphere bundle with vanishing Euler class) and then by cutting the sphere in two (Mayer–Vietoris), which makes Figure 7.23 precise.

Route 1: a sphere bundle over BG

By (7.70), S^2_G is an S^2-bundle over BG. It is the unit-sphere bundle of a vector bundle of rank 3. Write $\mathbb{R}^3 = \mathbb{C} \oplus \mathbb{R}$ with coordinates (w, z); $\mathrm{U}(1)$ acts on $\mathbb{C}$ with weight one and trivially on $\mathbb{R}$. Form

$$V = EG \times_G \mathbb{R}^3 = \underbrace{EG \times_G \mathbb{C}}_{L} \oplus \underbrace{EG \times_G \mathbb{R}}_{\mathbb{R}}, \qquad S^2_G = S(V), \tag{7.75}$$

where $L \to \mathbb{CP}^\infty$ is the universal complex line bundle (a real rank-2 bundle) and the second summand is the trivial line bundle, because G does not act on the z-axis. So S^2_G is the sphere bundle of $V = L \oplus \mathbb{R}$, which has rank 3 [Nash p. 203].

Two standard tools compute the cohomology of a sphere bundle.

Key idea: Euler class, Gysin sequence, Leray–Hirsch

Let $V \to B$ be an oriented real vector bundle of rank m, and $S^{m-1} \to S(V) \xrightarrow{\pi} B$ its unit-sphere bundle.

- Its **Euler class** $e(V) \in H^m(B)$ is the obstruction to a nowhere-vanishing section: if V has a section with no zeros, $e(V) = 0$. (For the tangent bundle of a surface it is the Gauss–Bonnet form, $\int e = \chi$.)
- The **Gysin sequence** is exact:

$$\cdots \to H^{k-m}(B) \xrightarrow{\cup e} H^k(B) \xrightarrow{\pi^*} H^k(S(V)) \longrightarrow H^{k-m+1}(B) \xrightarrow{\cup e} H^{k+1}(B) \to \cdots$$

- **Leray–Hirsch**: if there are classes on the total space E of a fibre bundle $F \to E \to B$ that restrict to a basis of $H^*(F)$ on every fibre, then $H^*(E) \cong H^*(B) \otimes H^*(F)$ as vector spaces, so $P_t(E) = P_t(B)\,P_t(F)$ [Nash (7.46)].

For our $V = L \oplus \mathbb{R}$ the Euler class vanishes, and there are two reasons. The direct one: V has the nowhere-zero section $[e] \mapsto [e, (0,1)]$ (the unit vector along the fixed z-axis, which G does not move). The general one, which Nash quotes: any bundle V_{2k+1} of *odd* rank has $2e(V) = 0$, because the fibrewise map $v \mapsto -v$ is a bundle automorphism that reverses the orientation of odd-dimensional fibres and so sends e to $-e$; with real coefficients, $e = 0$.

With $e = 0$ every map $\cup e$ in the Gysin sequence is zero, and the sequence breaks into short exact pieces:

$$\begin{aligned} \dim H^k(S(V)) &\overset{(1)}{=} \dim H^k(B) + \dim H^{k-m+1}(B), \\ P_t(S(V)) &\overset{(2)}{=} P_t(B) + t^{m-1}P_t(B) = (1 + t^{m-1})\,P_t(B) = P_t(S^{m-1})\,P_t(B). \end{aligned} \tag{7.76}$$

(1) With $\cup e = 0$ the sequence reads $0 \to H^k(B) \to H^k(S(V)) \to H^{k-m+1}(B) \to 0$, and in a short exact sequence of vector spaces the middle dimension is the sum of the other two.
(2) Multiply by t^k and sum over k; the second sum is shifted by $m-1$ degrees.

This is the Leray–Hirsch factorisation [Nash (7.46)]. With $m = 3$, $B = B\mathrm{U}(1)$ and (7.67) we get [Nash p. 204, (7.47)]

$$P_t^{\mathrm{U}(1)}(S^2) \overset{(1)}{=} P_t(S^2)\,P_t(B\mathrm{U}(1)) \overset{(2)}{=} (1+t^2)(1+t^2+t^4+\cdots) \overset{(3)}{=} \frac{1+t^2}{1-t^2} = 1 + 2t^2 + 2t^4 + 2t^6 + \cdots . \tag{7.77}$$

(1) (7.76) with $m = 3$: $S(V) = S^2_G$ and $P_t(S^2) = 1 + t^2$.

(2) (7.67).
(3) Sum the geometric series.

Worked example 7.44: When the Euler class does not vanish: the free circle

For contrast, let U(1) act on the circle $S^1 \subset \mathbb{C}$ by rotation. Then $S^1_G = S(L)$ is the circle bundle of the rank-2 bundle L over $\mathbb{CP}^\infty$, whose Euler class is the generator u (up to sign), and $\cup u \colon H^j(\mathbb{CP}^\infty) \to H^{j+2}(\mathbb{CP}^\infty)$ is an isomorphism for every $j \geq 0$ (it sends u^i to u^{i+1}). The Gysin sequence with $m = 2$ reads

$$H^{k-2}(B) \xrightarrow{\cup u} H^k(B) \to H^k(S(L)) \to H^{k-1}(B) \xrightarrow{\cup u} H^{k+1}(B),$$

so

$$\begin{aligned} H^k(S(L)) &\overset{(1)}{=} \operatorname{Coker}\big(\cup u \colon H^{k-2}(B) \to H^k(B)\big) \oplus \operatorname{Ker}\big(\cup u \colon H^{k-1}(B) \to H^{k+1}(B)\big) \\ &\overset{(2)}{=} \begin{cases} \mathbb{R}, & k = 0, \\ 0, & k \geq 1. \end{cases} \end{aligned}$$

(1) Exactness: $H^k(S(L))$ receives the cokernel of the left map and maps onto the kernel of the right map.
(2) $\cup u$ is injective, so every kernel vanishes; it is onto H^k for $k \geq 2$ and for $k = 1$ ($H^1(B) = 0$), so the only cokernel is $H^0(B) = \mathbb{R}$ (there is no H^{-2}).

So $S^1_G \simeq$ pt, as it must be for a free action (Example 7.43). Indeed $S(L) = S^\infty$ itself. A non-zero Euler class is how the sphere bundle "untwists" the BG direction.

Route 2: cut the sphere at the equator

Let $U = S^2 \setminus \{S\}$ and $V = S^2 \setminus \{N\}$ be the complements of the two poles. Both are U(1)-invariant, $U \cup V = S^2$, and $U \cap V$ is the open band $-1 < z < 1$. Now look at them equivariantly.

- U shrinks onto the north pole along the meridians, and this contraction commutes with the rotations. So $U_G \simeq \{N\}_G = BG$. Likewise $V_G \simeq BG$.
- $U \cap V$ shrinks onto the equator, on which G acts freely. So $(U \cap V)_G \simeq S^1_G \simeq$ pt.

The Mayer–Vietoris sequence of $S^2_G = U_G \cup V_G$ is

$$\cdots \to H^{k-1}\big((U \cap V)_G\big) \to H^k(S^2_G) \to H^k(U_G) \oplus H^k(V_G) \xrightarrow{r} H^k\big((U \cap V)_G\big) \to \cdots$$

with $r(a, b) = a|_{(U\cap V)_G} - b|_{(U\cap V)_G}$ (the difference of the restrictions). In degree 0, $r \colon \mathbb{R} \oplus \mathbb{R} \to \mathbb{R}$, $(a, b) \mapsto a - b$, is onto with one-dimensional kernel; in every other degree $H^k((U \cap V)_G) = 0$. Therefore

$$\begin{aligned} H^0(S^2_G) &\overset{(1)}{=} \operatorname{Ker} r = \{(a, a)\} \cong \mathbb{R}, \\ H^1(S^2_G) &\overset{(2)}{=} 0, \\ H^{2k}(S^2_G) &\overset{(3)}{=} H^{2k}(BG) \oplus H^{2k}(BG) \cong \mathbb{R}^2 \quad (k \geq 1), \qquad H^{\mathrm{odd}}(S^2_G) = 0. \end{aligned}$$

(1) Exactness at H^0: $H^0(S^2_G)$ injects into $H^0(U_G) \oplus H^0(V_G)$ with image $\operatorname{Ker} r$.
(2) $H^1(S^2_G)$ sits between $\operatorname{Coker}(r)$ in degree 0, which is zero because r is onto, and $H^1(U_G) \oplus H^1(V_G) = H^1(BG)^{\oplus 2}$, which is zero because $H^1(\mathbb{CP}^\infty) = 0$.

(3) For $k \geq 1$ both neighbours $H^{k-1}((U\cap V)_G)$ and $H^k((U\cap V)_G)$ vanish, so $H^k(S^2_G) \cong H^k(BG)^{\oplus 2}$, which is $\mathbb{R}^2$ in even degrees and 0 in odd ones.

So

$$P_t^{\mathrm{U}(1)}(S^2) = 1 + 2t^2 + 2t^4 + \cdots \overset{(1)}{=} 1 + \frac{2t^2}{1-t^2} = \frac{1+t^2}{1-t^2},$$

(1) $2t^2 + 2t^4 + \cdots = 2t^2(1 + t^2 + t^4 + \dots)$, a geometric series.

in agreement with (7.77). This route shows *where* the classes live: in positive degrees there is one copy of $H^*(BG)$ at each fixed point, exactly the two orange columns of Figure 7.23, and in degree 0 the two constants are glued through the contractible free part.

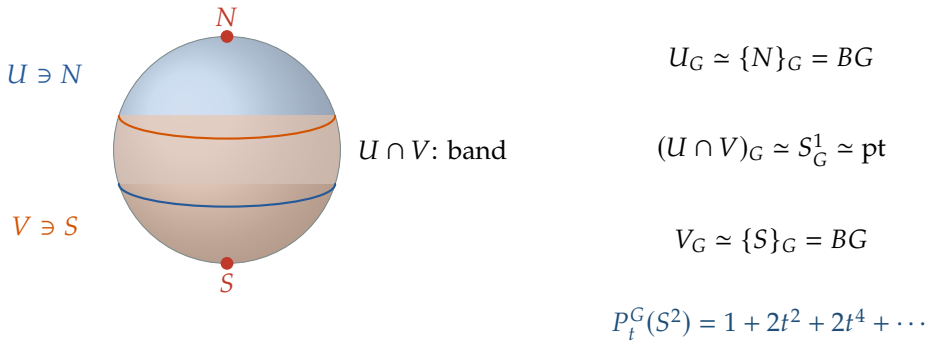


Figure 7.24: The Mayer–Vietoris cover used in Route 2. Each cap contracts equivariantly onto its pole, a fixed point, so each contributes a full copy of $H^*(BG)$. The band contracts onto the equator, where the action is free, so it contributes only a point. The two copies of $H^*(BG)$ are glued only in degree 0.

Watch out: Equivariant cohomology is not the cohomology of the quotient

Compare the three answers for U(1) acting on S^2:

$$P_t(S^2) = 1 + t^2, \qquad P_t(S^2/\mathrm{U}(1)) = 1, \qquad P_t^{\mathrm{U}(1)}(S^2) = \frac{1+t^2}{1-t^2}.$$

The equivariant answer is neither of the others. It equals $P_t(S^2)P_t(BG)$ here (the sphere is **equivariantly formal**), but that is a property of this example, not a general rule: for the free action on S^3, $P_t^{\mathrm{U}(1)}(S^3) = 1 + t^2$, whereas $P_t(S^3)P_t(BG) = (1 + t^3)/(1 - t^2)$.

7.6.6 Equivariant cohomology with differential forms: the Cartan model

The Borel construction is conceptually clean but uses an infinite-dimensional space. For $G = \mathrm{U}(1)$ and real coefficients there is a model built only from forms on M itself. It is the version physicists meet in localisation computations and in topological field theory.

Let X be the vector field that generates the action (for the rotating sphere, $X = \partial_\phi$, i.e. $X = -y\,\partial_x + x\,\partial_y$). Consider U(1)-invariant forms with coefficients that are polynomials in a formal variable u of degree 2:

$$\Omega_G(M) = \big(\Omega^*(M)\big)^{\mathrm{U}(1)} \otimes \mathbb{R}[u], \qquad \mathrm{d}_G = \mathrm{d} - u\,\iota_X, \tag{7.78}$$

where ι_X is the interior product. Both d and $u\,\iota_X$ raise the total degree by one (ι_X lowers the

form degree by 1 and u adds 2). The operator squares to zero on invariant forms:

$$d_G^2 \overset{(1)}{=} d^2 - u\,(d\iota_X + \iota_X d) + u^2\iota_X^2 \overset{(2)}{=} -u\,\mathcal{L}_X \overset{(3)}{=} 0.$$

(1) Expand $(d - u\iota_X)^2$; u is a number-like (even) variable, so it commutes with everything.
(2) $d^2 = 0$, $\iota_X^2 = 0$, and Cartan's formula $\mathcal{L}_X = d\iota_X + \iota_X d$ (Chapter T1).
(3) Invariant forms satisfy $\mathcal{L}_X\alpha = 0$.

Key idea: Cartan's theorem

For a compact connected group acting on M, the cohomology of $(\Omega_G(M), d_G)$ is $H^*_G(M;\mathbb{R})$. For U(1), u corresponds to the generator of $H^2(BU(1))$.

Check it on a point: there are no forms of positive degree and $X = 0$, so $\Omega_G = \mathbb{R}[u]$ with $d_G = 0$, and the cohomology is $\mathbb{R}[u]$, in agreement with (7.67).

Worked example 7.45: The equivariant area form of the sphere and its moment map

On S^2 use z and ϕ (cylindrical coordinates). The area form is

$$\omega = \sin\theta\, d\theta \wedge d\phi = d\phi \wedge dz, \tag{7.79}$$

because $z = \cos\theta$ gives $dz = -\sin\theta\, d\theta$, so $d\phi \wedge dz = -\sin\theta\, d\phi \wedge d\theta = \sin\theta\, d\theta \wedge d\phi$. It is closed but not d_G-closed. Look for a correction $u\,h$ with h a function:

$$\begin{aligned} d_G(\omega + u\,h) &\overset{(1)}{=} d\omega + u\,dh - u\,\iota_X\omega - u^2\iota_X h \\ &\overset{(2)}{=} u\,(dh - \iota_X\omega) \overset{(3)}{=} u\,(dh - dz). \end{aligned}$$

(1) Definition (7.78).
(2) $d\omega = 0$ (a top form) and $\iota_X h = 0$ (a function has no slot for a vector).
(3) $\iota_{\partial_\phi}(d\phi \wedge dz) = dz$.

So $h = z$ works, and

$$\omega_G = \omega + u\,z \tag{7.80}$$

is an equivariantly closed extension of the area form. The function that appears is the *height function*: the Morse function of this whole section. The equation $dz = \iota_X\omega$ says that the rotation is the Hamiltonian flow of z for the symplectic form ω (in classical mechanics: z is the angular momentum L_z that generates rotations about the z-axis). Such an h is called a **moment map**.

The restrictions to the fixed points make the Mayer–Vietoris picture concrete. At a point, forms of positive degree restrict to 0, so

$$\omega_G\big|_N = u \cdot z(N) = u, \qquad \omega_G\big|_S = u \cdot z(S) = -u, \qquad 1\big|_N = 1\big|_S = 1.$$

In degree 2 the classes $u \cdot 1 \mapsto (u, u)$ and $\omega_G \mapsto (u, -u)$ are independent, matching $\dim H^2_G(S^2) = 2$ from (7.77).

Dictionary: Equivariant cohomology in physics language

Mathematics	Physics
G acting on M	a symmetry, or a gauge group acting on configurations
orbit space M/G	physical (gauge-inequivalent) configurations
non-trivial stabiliser	reducible connection; symmetric vacuum
EG, contractible free G-space	auxiliary "gauge-fixing" degrees of freedom that remove the singularities
equivariant parameter $u \in H^2(BG)$	the ghost-for-ghost / Cartan generator ϕ of topological field theory
$\mathrm{d}_G = \mathrm{d} - u\,\iota_X$	equivariant BRST operator; $\mathrm{d}_G^2 = -u\mathcal{L}_X$ is a gauge transformation
moment map h with $\mathrm{d}h = \iota_X\omega$	the conserved charge that generates the symmetry

Watch out: Finite stabilisers are invisible over $\mathbb{R}$

If every stabiliser is finite, then with real coefficients $H^*_G(M;\mathbb{R}) \cong H^*(M/G;\mathbb{R})$, because $H^*(B\Gamma;\mathbb{R}) = \mathbb{R}$ for a finite group Γ (for instance $P_t(\mathbb{RP}^\infty;\mathbb{R}) = 1$). Over $\mathbb{Z}_p$ this is false. The serious trouble that equivariant cohomology repairs comes from positive-dimensional stabilisers, such as the $\mathrm{U}(1)$ fixing the poles of S^2.

Reading Nash (pp. 201–204, (7.34)–(7.42), (7.45)–(7.49))

Nash's §3 starts with the height function on S^2 [Nash (7.34)] and the disappointing quotient $S^2/\mathrm{U}(1) \simeq [0,1]$ [Nash (7.35)], blames the fixed points, recalls that free actions have smooth quotients on which f_G lives, and then, following Atiyah and Bott, defines the diagonal action [Nash (7.36)], the homotopy quotient [Nash (7.37)], the fibration over BG [Nash (7.38)], equivariant cohomology [Nash (7.39)], the point [Nash (7.40)–(7.41)] and the equivariant Poincaré series [Nash (7.42)]. Two misprints: in the diagram [Nash (7.38)] the base of the fibration should be BG, not M; and in [Nash (7.46)] "$H^*(M)$" means the cohomology of the base BG. The isomorphism [Nash (7.46)] is one of vector spaces (the Leray–Hirsch theorem), which is all that Poincaré series need. He computes $P_t^G(S^2)$ by the sphere-bundle argument [Nash (7.45)–(7.47)] and justifies $P_t(B\mathrm{U}(1)) = 1/(1-t^2)$ by a digression on $\mathbb{CP}^\infty$ and S^∞ [Nash (7.48)–(7.49)]. We added: orbits and stabilisers, the catalogue of actions, the pole-to-endpoint computation, the $S^{2n+1} \to \mathbb{CP}^n$ comparison, the contraction of S^∞, the classifying spaces of $\mathbb{Z}$, $\mathbb{Z}_2$, $\mathrm{SU}(2)$, T^k, $\mathrm{U}(n)$, the monopole bundle as a piece of the universal bundle, the two projections of M_G and the fibre BG_x, the Gysin computation for the free circle, the Mayer–Vietoris computation, and the Cartan model with the moment map of the sphere.

Recap

- Dividing by a non-free action loses information: $S^2/\mathrm{U}(1) = [0,1]$, and the critical points become boundary points.
- Orbit $\cong G/G_x$; free actions have smooth quotients; fixed points make cones and

boundaries.

- EG: contractible with a free action; $BG = EG/G$. $B\mathrm{U}(1) = \mathbb{CP}^\infty$, $P_t = 1/(1-t^2)$; $B\mathrm{SU}(2) = \mathbb{HP}^\infty$, $P_t = 1/(1-t^4)$.
- Borel: $M_G = (M \times EG)/G$, an M-bundle over BG, with fibre BG_x over the orbit of x. $H^*_G(M) \coloneqq H^*(M_G)$.
- Point: $H^*(BG)$. Free: $H^*(M/G)$. Trivial action: $H^*(M) \otimes H^*(BG)$.
- Rotating sphere: $P_t^{\mathrm{U}(1)}(S^2) = (1+t^2)/(1-t^2)$, by Gysin (Euler class zero) or by Mayer–Vietoris (one BG per pole).
- Cartan model: $\mathrm{d}_G = \mathrm{d} - u\,\iota_X$; the area form extends to $\omega + u\,z$, and the height is the moment map.

7.7 Equivariant Morse theory

We now have both halves of an equivariant Morse theory. The topological half is the equivariant Poincaré series $P_t^G(M)$ of (7.73). The question of this section is: *what is the equivariant Morse series, and does it satisfy Morse inequalities?* The answer is that one simply does Morse–Bott theory for the function that f induces on the homotopy quotient M_G. Each critical manifold N then contributes $P_t^G(N)\,t^{\lambda_N}$ instead of $P_t(N)\,t^{\lambda_N}$. Fixed points contribute a whole $P_t(BG)$; free orbits contribute only 1. We shall see that this repairs the height function on the rotating sphere, turns the imperfect function z^2 into a perfect one, computes equivariant cohomology by counting fixed points, and leads, via the Cartan model, to an exact stationary-phase formula, the Duistermaat–Heckman formula, which is the simplest example of localisation in physics.

7.7.1 The induced function on the homotopy quotient

Let $f\colon M \to \mathbb{R}$ be G-invariant, $f(g\cdot x) = f(x)$. Define

$$f_G\colon M_G \to \mathbb{R}, \qquad f_G[x,e] = f(x). \tag{7.81}$$

This is well defined, because the two representatives of the same point give the same value:

$$f_G[g\cdot x, g\cdot e] \overset{(1)}{=} f(g\cdot x) \overset{(2)}{=} f(x) = f_G[x,e].$$

(1) Definition (7.81) applied to the representative $(g\cdot x, g\cdot e)$.
(2) f is G-invariant.

On each fibre of $\pi\colon M_G \to BG$ (a copy of M) the function f_G is just f.

To do calculus we use a finite approximation E_N of EG (for $\mathrm{U}(1)$, $E_N = S^{2N+1}$), so that $M \times_G E_N$ is a compact manifold, and let $N \to \infty$ at the end. First pull f back to the product: $F(x,e) = f(x)$ on $M \times E_N$. In product coordinates (x^i, e^a),

$$\mathrm{d}F \overset{(1)}{=} \partial_i f\,\mathrm{d}x^i + 0\cdot \mathrm{d}e^a, \qquad HF(x,e) \overset{(2)}{=} \begin{pmatrix} Hf(x) & 0 \\ 0 & 0 \end{pmatrix}. \tag{7.82}$$

(1) F does not depend on e.
(2) All second derivatives involving e vanish for the same reason.

So the critical set of F is $\mathrm{Crit}(f) \times E_N$, and if N is a non-degenerate critical manifold of f, then $N \times E_N$ is a non-degenerate critical manifold of F. Its tangent directions are $TN \oplus TE_N$, which is exactly the kernel of HF, and the Hessian on the normal directions is that of f. In particular the index is still λ_N.

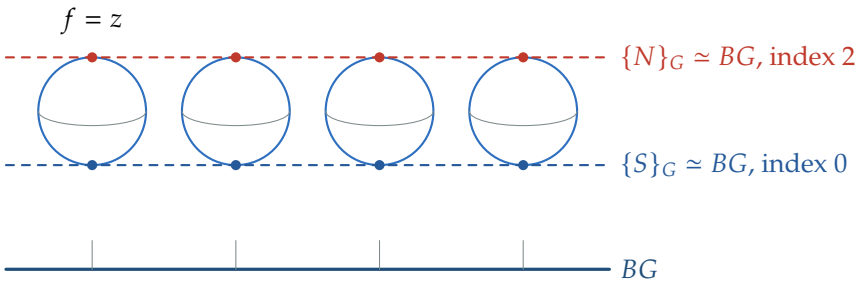
Next divide by G. When G is connected, each critical manifold N of f is G-invariant: $g \cdot N$ is again a critical manifold (since f is invariant), it can be joined to N through the path of group elements from 1 to g, and distinct critical manifolds are disjoint. So $N \times E_N$ is invariant under the free diagonal action, and it descends to

$$N_G = (N \times EG)/G \;\subset\; M_G, \tag{7.83}$$

the homotopy quotient of N by G. Dividing by a free action does not change the normal directions (the group directions are tangent to $N \times E_N$), so:

Key idea: Critical manifolds of f_G (Nash p. 203)

The critical manifolds of f_G on M_G are the homotopy quotients N_G of the critical manifolds N of f. The index of f_G on N_G is the *same* as the index λ_N of f on N.



each fibre is a copy of S^2 carrying f; the critical set of f_G is swept out by the poles

Figure 7.25: The induced function f_G on S^2_G, drawn as a bundle of spheres over BG. On every fibre f_G is the height $f = z$, with its two poles. As the fibre moves over BG the poles sweep out the critical manifolds $\{N\}_G$ and $\{S\}_G$, each a copy of BG (dashed). The normal directions to them are the directions inside a fibre, so the indices are 2 and 0 as before.

7.7.2 The equivariant Morse series and inequalities

Applying the Morse–Bott series (7.54) to f_G gives:

Working definition 7.18 (equivariant Morse series)

For a G-invariant [$f(g \cdot x) = f(x)$ for every $g \in G$] Morse–Bott function f with critical manifolds N,

$$M_t^G(f) = \sum_N P_t(N_G)\, t^{\lambda_N} = \sum_N P_t^G(N)\, t^{\lambda_N} \tag{7.84}$$

[$N_G = (N \times EG)/G$ is the homotopy quotient of N, and λ_N is the number of independent downhill directions normal to N] [Nash (7.43)]. *Recipe*: list the critical manifolds of f with their indices; for each, work out its equivariant Poincaré series (a fixed point: $P_t(BG)$; a free orbit: 1; a manifold N on which G acts freely: $P_t(N/G)$); multiply by t^{λ_N} and add.

Key idea: Equivariant Morse inequalities (Nash (7.44))

For G-invariant Morse–Bott functions,

$$M_t^G(f) - P_t^G(M) = (1+t)\,R(t), \tag{7.85}$$

where $R(t)$ is a formal power series with non-negative coefficients.

Why. Apply the Morse–Bott inequalities (7.57) to f_G on the compact manifold $M \times_G E_N$. Its cohomology agrees with that of M_G in degrees below $2N+1$ (for U(1)), and so do the Poincaré series of the critical manifolds $N \times_G E_N$. Therefore (7.85) holds coefficient by coefficient up to t^{2N}. Since N is arbitrary, it holds for all coefficients. (For real coefficients we also need the negative bundles N_G^- to be orientable; they are in all our examples, where they are complex vector bundles.)

The algebra of Section 7.2 carries over to power series with one exception.

- The weak inequalities $m_i^G \geq b_i^G$ and the alternating forms still hold.
- The *lacunary principle still holds*, with the same proof: if $R \neq 0$, look at its lowest term $r_j t^j$; then $(1+t)R$ contributes to both t^j and t^{j+1} in M_t^G. So if $M_t^G(f)$ has no two consecutive powers, $M_t^G(f) = P_t^G(M)$.
- The value $t = -1$ is *not* available: the series are infinite and need not converge there. There is no equivariant Euler characteristic in this naive sense.

Working definition 7.19 (equivariantly perfect)

f is **equivariantly perfect** if $M_t^G(f) = P_t^G(M)$, i.e. $R \equiv 0$ [$R(t)$ is the series in $M_t^G - P_t^G = (1+t)R(t)$. It vanishes exactly when the two series agree term by term].

7.7.3 The height function on the rotating sphere

Return to $f = z$ on S^2 with the rotation action. The critical points are the two poles, fixed by U(1), with indices 2 (N) and 0 (S). Each is a point on which G acts trivially, so its homotopy quotient is $\{p\}_G = BG$ (7.67). The equivariant Morse series is [Nash p. 204, (7.50)]

$$M_t^G(f) \overset{(1)}{=} P_t^G(\{S\})\,t^0 + P_t^G(\{N\})\,t^2 \overset{(2)}{=} (1+t^2)\,P_t(BG) \overset{(3)}{=} \frac{1+t^2}{1-t^2} \overset{(4)}{=} P_t^G(S^2). \tag{7.86}$$

(1) (7.84) with the two poles.
(2) Each pole is a fixed point, so $P_t^G(\{p\}) = P_t(BG)$.
(3) $P_t(BU(1)) = 1/(1-t^2)$.
(4) (7.77).

The height function is **equivariantly perfect**. (It had to be: the series $(1+t^2)(1+t^2+t^4+\cdots)$ contains only even powers, so the equivariant lacunary principle applies. This gives a third derivation of $P_t^G(S^2)$.) The naive quotient $[0,1]$ lost the poles; the homotopy quotient has kept each of them, fattened into a copy of BG.

7.7.4 The square of the height: equivariantly perfect, not perfect

Now take $f = z^2$ (7.58), which is also U(1)-invariant. The Poincaré side is unchanged, $P_t^G(S^2) = (1+t^2)/(1-t^2)$; only the Morse side must be recomputed [Nash p. 205]. Its critical manifolds are the two poles (index 2) and the equator $E = \{z = 0\}$ (index 0).

- The poles are fixed points. Together they contribute [Nash (7.51)]

$$2\,t^2\,P_t(BU(1)) = \frac{2t^2}{1-t^2}.$$

- The equator is a single free orbit. Since G acts freely, $E_G \simeq E/G$ [Nash (7.52)]:

$$S^1_{U(1)} \simeq S^1/U(1) = \mathrm{pt}, \qquad P_t^G(E) = 1,$$

  so the equator contributes $t^0 \cdot 1 = 1$.

Hence [Nash (7.53)]

$$M_t^G(z^2) \overset{(1)}{=} 1 + \frac{2t^2}{1-t^2} \overset{(2)}{=} \frac{1-t^2+2t^2}{1-t^2} = \frac{1+t^2}{1-t^2} \overset{(3)}{=} P_t^G(S^2). \tag{7.87}$$

(1) Equator plus the two poles.
(2) Common denominator.
(3) (7.77).

So z^2 is *equivariantly perfect* even though it is not perfect (Example 7.32, where $R = t$). Equivariant perfection does not imply ordinary perfection.

What happened to the cancelling pair? In the ordinary theory the surplus was "the t from the equator, cancelled by one of the t^2's from the poles": the equator's 1-cycle was created at $f = 0$ and killed when the first cap was attached (Figure 7.20). Equivariantly, the equator's 1-cycle is precisely the orbit direction, which the homotopy quotient divides out ($E_G \simeq \mathrm{pt}$). So there is nothing for the cap to kill, and each pole instead contributes a fresh copy of $H^*(BG)$ starting in degree 2. The ledger below shows the coefficients.

	t^0	t^1	t^2	t^3	t^4	t^6	
$P_t(S^2)$	1	0	1	0	0	0	
$M_t(z)$	1	0	1	0	0	0	perfect
$M_t(z^2)$	1	1	2	0	0	0	$R = t$
$P_t^G(S^2)$	1	0	2	0	2	2	
$M_t^G(z)$: $\{S\}_G, \{N\}_G$	1	0	$1+1$	0	$1+1$	$1+1$	perfect
$M_t^G(z^2)$: $E_G, 2\{N,S\}_G$	1	0	$0+2$	0	$0+2$	$0+2$	perfect

Figure 7.26: The ordinary and equivariant ledgers for U(1) acting on S^2. Top: ordinary Morse theory, where z^2 has the surplus pair shown in red. Bottom: equivariant Morse theory. For z each pole contributes $1, 1, 1, \ldots$ in degrees $\lambda, \lambda+2, \ldots$ (a copy of $P_t(BG)$). For z^2 the equator contributes only its t^0, the poles contribute 2 in each even degree ≥ 2, and the surplus has gone.

7.7.5 More examples: free actions and fixed points

Worked example 7.46: A free action: equivariant Morse theory is Morse theory on the quotient

Take $f = \sum a_i\,|z_i|^2$ on S^{2n+1} with the free U(1) action, as in Example 7.38. Its critical manifolds are the circles C_i of index $2(i-1)$, each a single free orbit, so $P_t^G(C_i) =$

$P_t(C_i/\mathrm{U}(1)) = 1$:

$$M_t^G(f) \overset{(1)}{=} \sum_{i=1}^{n+1} t^{2(i-1)} \cdot 1 = 1 + t^2 + \cdots + t^{2n} \overset{(2)}{=} P_t(\mathbb{CP}^n) \overset{(3)}{=} P_t^G(S^{2n+1}).$$

(1) (7.84); each critical circle is a free orbit and its homotopy quotient is a point.
(2) (7.42).
(3) Free action: $H_G^*(S^{2n+1}) = H^*(S^{2n+1}/\mathrm{U}(1)) = H^*(\mathbb{CP}^n)$ (Example 7.43).

Upstairs f had $R = t + t^3 + \cdots + t^{2n-1}$; equivariantly $R = 0$. For a free action the equivariant Morse theory of f *is* the ordinary Morse theory of f_G on M/G [Nash p. 201].

Worked example 7.47: Counting fixed points computes $H^*_{\mathrm{U}(1)}(\mathbb{CP}^2)$

Let $\mathrm{U}(1)$ act on $\mathbb{CP}^2$ by

$$e^{i\alpha} \cdot [z_1 : z_2 : z_3] = [z_1 : e^{i\alpha} z_2 : e^{2i\alpha} z_3].$$

Fixed points. $[z]$ is fixed when $(z_1, e^{i\alpha} z_2, e^{2i\alpha} z_3) = \mu(\alpha)\,(z_1, z_2, z_3)$ for every α, i.e. when z is an eigenvector of $\mathrm{diag}(1, e^{i\alpha}, e^{2i\alpha})$. For generic α the three eigenvalues are distinct, so the eigenvectors are the coordinate axes. The fixed points are $P_1 = [1:0:0]$, $P_2 = [0:1:0]$, $P_3 = [0:0:1]$.
The function. Nash's $f = \sum a_i |z_i|^2 / \sum |z_j|^2$ with $a_1 < a_2 < a_3$ (7.36) depends only on the moduli $|z_i|$, so it is invariant. Its critical points are P_1, P_2, P_3 with indices $0, 2, 4$ (Section 7.4.2): exactly the fixed points.
Equivariant Morse series. Each critical point is a fixed point:

$$M_t^G(f) \overset{(1)}{=} (1 + t^2 + t^4)\, P_t(BG) \overset{(2)}{=} \frac{1 + t^2 + t^4}{1 - t^2} = 1 + 2t^2 + 3t^4 + 3t^6 + 3t^8 + \cdots .$$

(1) Three fixed points of indices $0, 2, 4$, each contributing $t^\lambda P_t(BG)$.
(2) $P_t(B\mathrm{U}(1)) = 1/(1 - t^2)$.

Only even powers occur, so by the equivariant lacunary principle

$$P_t^{\mathrm{U}(1)}(\mathbb{CP}^2) = \frac{1 + t^2 + t^4}{1 - t^2} = P_t(\mathbb{CP}^2)\, P_t(BG).$$

We have computed an equivariant cohomology without ever constructing $\mathbb{CP}^2_{\mathrm{U}(1)}$: counting fixed points was enough.
A remark on the choice of f. With $a = (0, 1, 2)$, the weights of the action, f is the moment map of the rotation for the Fubini–Study form, just as z was for S^2 (Example 7.45). Atiyah and Bott, and later Kirwan, showed that this is a general phenomenon: the norm-square of a moment map (and, for a circle, the moment map itself) is an equivariantly perfect Morse–Bott function.

Try it: Two rotating spheres

Let $\mathrm{U}(1)$ rotate both factors of $S^2 \times S^2$ about their z-axes, and let $f = z_1 + 2z_2$. Find the critical points, their indices and $M_t^G(f)$, and deduce $P_t^G(S^2 \times S^2)$.
Answer: the critical points are the four pairs of poles (all fixed): (S, S) index 0; (N, S) and (S, N) index 2; (N, N) index 4. So $M_t^G = (1 + 2t^2 + t^4)/(1 - t^2) = (1 + t^2)^2/(1 - t^2)$. Only

even powers, hence this is $P_t^G(S^2 \times S^2)$, equal to $P_t(S^2 \times S^2)P_t(BG)$.

Watch out: Non-free orbits with finite stabiliser

If a critical orbit has a finite stabiliser Γ, e.g. the circle $z_1 = 0$ of the weighted action on S^3 (Example 7.37), then its homotopy quotient is $B\Gamma$, which is a point for real coefficients but not for $\mathbb{Z}_p$. This matters for closed geodesics (Section 7.8.10): a geodesic traversed k times is fixed by the rotations by $2\pi/k$.

7.7.6 Localisation: an exact stationary-phase formula

Equivariant cohomology has a striking application that every physicist recognises: in favourable cases the *stationary-phase approximation is exact*. Consider the oscillatory integral over the sphere

$$Z(s) = \int_{S^2} e^{isz}\,\omega, \tag{7.88}$$

a toy "path integral" in which z plays the role of the action and s of $1/\hbar$.

Exact evaluation. By (7.79), $\omega = \mathrm{d}\phi \wedge \mathrm{d}z$, so

$$Z(s) \overset{(1)}{=} \int_0^{2\pi} \mathrm{d}\phi \int_{-1}^{1} e^{isz}\,\mathrm{d}z \overset{(2)}{=} 2\pi\,\frac{e^{is} - e^{-is}}{is} \overset{(3)}{=} \frac{4\pi \sin s}{s}.$$

(1) The coordinates (ϕ, z) cover the sphere minus the poles with the standard orientation, and the area form is $\mathrm{d}\phi\,\mathrm{d}z$.

(2) $\int_{-1}^{1} e^{isz}\mathrm{d}z = [e^{isz}/(is)]_{-1}^{1}$.

(3) $e^{is} - e^{-is} = 2i\sin s$.

The first step is Archimedes' theorem: the area of the band of the sphere between two heights depends only on the difference of the heights (Figure 7.27).

Stationary phase. For large s, the integral is dominated by the critical points of the phase z, the poles. Near N, use (x, y): then $z \approx 1 - \tfrac{1}{2}(x^2 + y^2)$ and $\omega \approx \mathrm{d}x \wedge \mathrm{d}y$, so

$$Z_N(s) \overset{(1)}{=} e^{is} \int_{\mathbb{R}^2} e^{-\frac{is}{2}(x^2+y^2)}\,\mathrm{d}x\,\mathrm{d}y \overset{(2)}{=} e^{is}\left(\sqrt{\tfrac{2\pi}{is}}\right)^2 = \frac{2\pi\, e^{is}}{is}.$$

(1) Keep the Hessian (quadratic) term of the phase at the pole and the leading term of the area form; extend the integral to the whole plane.

(2) Fresnel integral $\int e^{-ax^2/2}\mathrm{d}x = \sqrt{2\pi/a}$ with $a = is$ (interpreted by continuation from $\operatorname{Re} a > 0$).

Near S, $z \approx -1 + \tfrac{1}{2}(x^2 + y^2)$, and the same computation with $a = -is$ gives $Z_S(s) = 2\pi e^{-is}/(-is)$. Adding:

$$Z_N(s) + Z_S(s) = \frac{2\pi\,(e^{is} - e^{-is})}{is} \overset{(1)}{=} Z(s) \qquad \text{exactly.}$$

(1) Compare with the exact evaluation above.

The Gaussian approximation at the two critical points has no corrections at all. This is the **Duistermaat–Heckman formula**: if h is the moment map of a circle action on a compact symplectic manifold, then $\int e^{ish}\,\omega^n/n!$ equals its stationary-phase approximation.

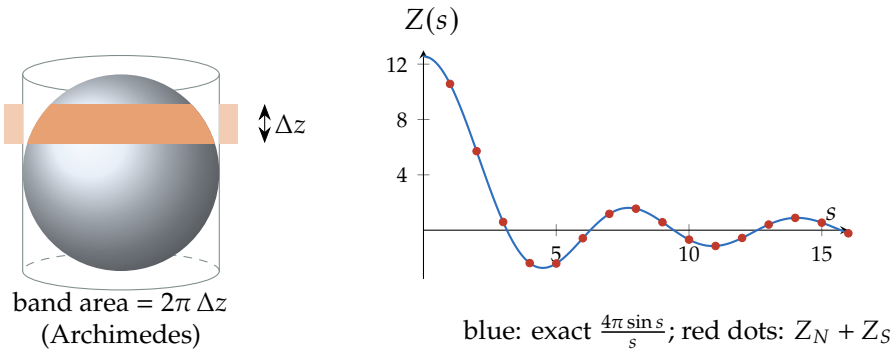


Figure 7.27: Left: Archimedes' theorem. The band of the unit sphere between two heights has the same area as the matching band of the circumscribing cylinder, $2\pi\,\Delta z$. So the pushforward of the area measure to the z-axis is flat, and $Z(s)$ is the Fourier transform of a box. Right: the exact integral $Z(s) = 4\pi \sin s/s$ and the sum of the two stationary-phase contributions from the poles, which agree for every s, not just asymptotically.

Why it is exact: the equivariant mechanism. The integrand is the top-degree part of an equivariantly closed form. With the Cartan model of Section 7.6.6 and $\omega_G = \omega + u\,z$ (7.80),

$$e^{\omega_G} = e^{uz}\big(1+\omega\big), \qquad \int_{S^2} e^{\omega_G} \;:=\; \int_{S^2} \big[e^{\omega_G}\big]_{\text{top}} = \int_{S^2} e^{uz}\,\omega,$$

which is $Z(s)$ at $u = is$. (The series stops at ω since $\omega^2 = 0$ on a surface.) Now the key fact: the integral of an equivariantly closed form α ($\mathrm{d}_G\alpha = 0$) does not change if we multiply it by $e^{\tau\,\mathrm{d}_G\theta}$ for an invariant 1-form θ. Indeed

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\int_M \alpha\, e^{\tau\,\mathrm{d}_G\theta} \overset{(1)}{=} \int_M (\mathrm{d}_G\theta)\,\alpha\, e^{\tau\,\mathrm{d}_G\theta} \overset{(2)}{=} \int_M \mathrm{d}_G\big(\theta\,\alpha\, e^{\tau\,\mathrm{d}_G\theta}\big) \overset{(3)}{=} \int_M \mathrm{d}\big[\theta\,\alpha\, e^{\tau\,\mathrm{d}_G\theta}\big]_{\text{top}-1} \overset{(4)}{=} 0.$$

(1) Differentiate the exponential; $\mathrm{d}_G\theta$ has even degree, so everything commutes.
(2) d_G obeys the Leibniz rule, $\mathrm{d}_G\alpha = 0$, and $\mathrm{d}_G(\mathrm{d}_G\theta) = -u\,\mathcal{L}_X\theta = 0$ because θ is invariant.
(3) Only the top-degree part of the form is integrated; the top part of $\mathrm{d}_G\beta$ is d of the part of β one degree lower, since $u\,\iota_X$ lowers the form degree.
(4) Stokes' theorem on the closed manifold M.

Choose $\theta = g(X,\cdot)$, the metric dual of the generating vector field. Then $\mathrm{d}_G\theta = \mathrm{d}\theta - u\,\iota_X\theta = \mathrm{d}\theta - u\,|X|^2$, so the factor $e^{\tau\mathrm{d}_G\theta}$ contains $e^{-\tau u|X|^2}$. For $\tau \to \infty$ (with u treated as a positive real number, then continued) this suppresses everything except a neighbourhood of the zeros of X: the fixed points, which here are the poles, the critical points of the moment map. There the Gaussian integral is exact. This is the same logic as supersymmetric localisation: add a "Q-exact" term $\tau\,Q\theta$ to the action, note that nothing depends on τ, and send $\tau \to \infty$. Witten's approach to Morse theory (§4, part 7b) is the supersymmetric version of the same idea.

Physical picture: Atiyah–Bott: Yang–Mills as an equivariant Morse function

Nash's reference for equivariant Morse theory is Atiyah and Bott's paper on the Yang–Mills equations over Riemann surfaces. There M is the space $\mathcal{A}$ of all connections on a bundle over a surface Σ, the group is the gauge group $\mathcal{G}$, and the function is the Yang–Mills action $\mathrm{YM}(A) = \int_\Sigma |F_A|^2$, which is gauge invariant. The action of $\mathcal{G}$ is not free: reducible connections, and in particular every connection (via the constant central gauge transformations), have stabilisers. So ordinary Morse theory on $\mathcal{A}/\mathcal{G}$ is hopeless and equivariant Morse theory is the natural tool. Because $\mathcal{A}$ is an affine space, hence

contractible, its equivariant cohomology is simply $H^*(B\mathcal{G})$, which can be computed. The critical points of YM are the Yang–Mills connections; the minimum is the set of flat connections (the moduli space the physicist is interested in), and the higher critical sets are built from reducible connections. Atiyah and Bott showed that YM is equivariantly perfect, so

$$P_t(B\mathcal{G}) = P_t^{\mathcal{G}}(\text{flat connections}) + \sum_{\text{higher } C_\mu} t^{\lambda_\mu} P_t^{\mathcal{G}}(C_\mu),$$

from which they computed the cohomology of the moduli space of flat connections (equivalently, of stable holomorphic bundles). The infinite-dimensional setting is the subject of the next section, and Yang–Mills theory returns in Chapter 8.

Reading Nash (pp. 203–205, (7.42)–(7.44), (7.50)–(7.53))

Nash defines the equivariant Poincaré series [Nash (7.42)] and the equivariant Morse series as the ordinary Morse(–Bott) series of f_G on M_G: a critical manifold N of f gives a critical manifold N_G of f_G of the same index, and [Nash (7.43)] follows; the inequalities are [Nash (7.44)]. After computing $P_t^G(S^2)$ (previous section) he shows that the height function is equivariantly perfect [Nash (7.50)], and that z^2, with the poles contributing $2t^2/(1-t^2)$ [Nash (7.51)] and the free equator only 1 [Nash (7.52)], is equivariantly perfect too [Nash (7.53)]: "equivariant perfection does not imply perfection in the ordinary sense". We added: why f_G is well defined and why N_G has the same index (7.82), the finite-dimensional approximation behind (7.85), the equivariant lacunary principle, the ledger of cancelling pairs, the free-action and $\mathbb{CP}^2$ examples, the Duistermaat–Heckman formula with its localisation argument, and the Atiyah–Bott application that Nash alludes to.

Recap

- A G-invariant f induces $f_G[x,e] = f(x)$ on M_G; its critical manifolds are the N_G, with the same indices.
- $M_t^G(f) = \sum_N P_t^G(N)\, t^{\lambda_N}$: a fixed point gives $P_t(BG)$, a free orbit gives 1.
- $M_t^G - P_t^G = (1+t)R$, $R \geq 0$; the lacunary principle survives, $t = -1$ does not.
- On the rotating S^2: z and z^2 are both equivariantly perfect, $M_t^G = (1+t^2)/(1-t^2)$; z^2 is not perfect.
- Counting fixed points of a circle action computes H_G^* (e.g. $\mathbb{CP}^2$).
- Localisation: $\int_{S^2} e^{isz}\omega = 4\pi \sin s/s$ equals its stationary-phase approximation exactly (Duistermaat–Heckman).

7.8 Infinite dimensions: geodesics, loop spaces and Bott periodicity

Every example so far has had a finite-dimensional M. The problem that created the subject was not of that kind. Morse wanted to count the *geodesics* joining two points of a Riemannian manifold, and for that the "manifold" on which the function lives is the space of all paths from one point to the other. That space is infinite dimensional [Nash p. 205].

A physicist has met this situation many times. The action principle says that classical trajectories are the stationary points of a functional on the space of all paths. The free particle

moving on a curved space M has action $S = \frac{1}{2}\int g_{ij}\dot{x}^i\dot{x}^j\,dt$, whose stationary points are geodesics. The semiclassical expansion of the path integral expands S to second order around each classical path. The resulting quadratic form, the "fluctuation operator", has a finite number of negative modes, and that number enters the answer as a phase. This is exactly the index of Morse theory, now in infinite dimensions.

In this section we set up the calculus of variations for the energy functional, compute its Hessian, and find the index of every geodesic on a round sphere. Then we do Morse theory on the path space of S^n, which proves Morse's theorem that any two points of a sphere, whatever the metric, are joined by infinitely many geodesics. We see how to reduce everything to finite dimensions with broken geodesics (the time-slicing of the path integral), turn to closed geodesics and the free loop space where the circle action of the previous sections reappears, and finish with Bott's use of all this to prove his periodicity theorem, plus the other triumphs that Nash lists at the start of the chapter.

7.8.1 The space of paths and the energy functional

Let (M, g) be a connected Riemannian manifold, and fix two points $p, q \in M$. The **path space** is

$$PM(p,q) = \big\{\gamma\colon [0,1] \to M \;\big|\; \gamma \text{ smooth (or piecewise smooth)},\ \gamma(0) = p,\ \gamma(1) = q\big\}. \tag{7.89}$$

A "point" of $PM(p,q)$ is a whole path. On it Morse puts the **energy functional** [Nash (7.54)]

$$E(\gamma) = \int_0^1 \left|\frac{d\gamma}{dt}\right|^2 dt = \int_0^1 g_{ij}\big(\gamma(t)\big)\,\frac{d\gamma^i}{dt}\frac{d\gamma^j}{dt}\,dt, \tag{7.90}$$

a real-valued function on the path space [Nash (7.55)]:

$$E\colon PM(p,q) \longrightarrow \mathbb{R}, \qquad \gamma \longmapsto E(\gamma). \tag{7.91}$$

Up to a factor 2, E is the action of a free non-relativistic particle of unit mass moving on M for a unit of time. Here $\gamma^i(t)$ are the coordinates of $\gamma(t)$ in a chart and we sum over repeated indices. The metric g and the endpoints are fixed; only the path varies.

Tangent vectors to the path space. To do calculus on $PM(p,q)$ we need to know what a tangent vector is. Take a one-parameter family of paths $\gamma_\varepsilon \in PM(p,q)$ with $\gamma_0 = \gamma$, that is, a map $(t,\varepsilon) \mapsto \gamma_\varepsilon(t)$. Its velocity at $\varepsilon = 0$ is

$$V(t) = \left.\frac{\partial \gamma_\varepsilon(t)}{\partial \varepsilon}\right|_{\varepsilon=0} \in T_{\gamma(t)}M, \qquad V(0) = V(1) = 0. \tag{7.92}$$

The endpoint conditions hold because $\gamma_\varepsilon(0) = p$ and $\gamma_\varepsilon(1) = q$ for all ε. So:

Working definition 7.20 (tangent space of the path space)

A tangent vector to $PM(p,q)$ at γ is a **variation vector field**: a vector field $V(t)$ along γ [one tangent vector $V(t) \in T_{\gamma(t)}M$ at each point of the path] that vanishes at both endpoints. Every such V comes from a variation, for example $\gamma_\varepsilon(t) = \exp_{\gamma(t)}(\varepsilon V(t))$ [the point reached by following the geodesic that leaves $\gamma(t)$ with velocity $\varepsilon V(t)$ for unit time]. *In practice*: to differentiate a functional [a function whose argument is a whole path, such as E] at γ, replace γ by $\gamma + \varepsilon V$ (in a chart) and differentiate in ε at $\varepsilon = 0$.

The tangent space is infinite dimensional. In flat $\mathbb{R}^n$ a basis is given by Fourier sine modes:

$$V(t) = \sum_{k=1}^{\infty} a_k \sin(k\pi t), \qquad a_k \in \mathbb{R}^n, \tag{7.93}$$

since $\sin(k\pi t)$ vanishes at $t = 0, 1$. The coordinates $(a_1, a_2, \dots)$ run over an infinite sequence of vectors. This is the same function space as for a quantum particle in a box.

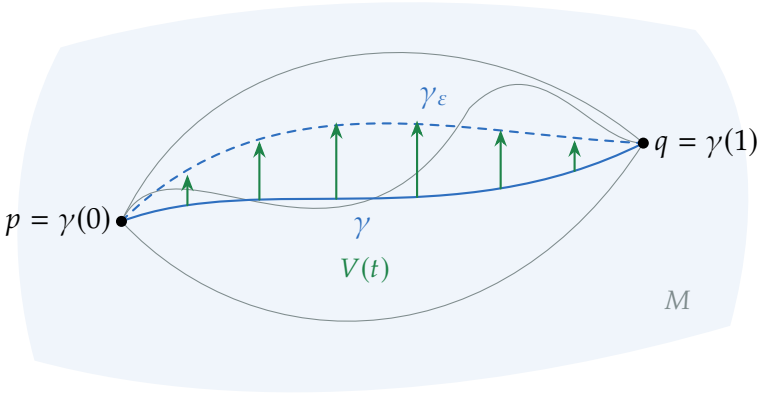


Figure 7.28: The path space $PM(p, q)$. Each curve from p to q (grey, blue) is a single point of this infinite-dimensional space. A tangent vector at γ is a vector field V along γ (green) which vanishes at the two ends, because every nearby path γ_ε (dashed) has the same endpoints. The energy E assigns a number to each curve; its critical points are the geodesics.

Worked example 7.48: Flat space: one critical point, index 0

Let $M = \mathbb{R}^n$ with $g_{ij} = \delta_{ij}$. Write any path from p to q as the straight line plus a deviation that vanishes at the ends: $\gamma(t) = p + t(q - p) + \eta(t)$, $\eta(0) = \eta(1) = 0$. Then

$$\begin{aligned} E(\gamma) &\overset{(1)}{=} \int_0^1 |(q - p) + \dot\eta(t)|^2 \, \mathrm{d}t \\ &\overset{(2)}{=} |q - p|^2 + 2\,(q - p) \cdot \int_0^1 \dot\eta \, \mathrm{d}t + \int_0^1 |\dot\eta|^2 \, \mathrm{d}t \\ &\overset{(3)}{=} |q - p|^2 + \int_0^1 |\dot\eta|^2 \, \mathrm{d}t . \end{aligned}$$

(1) $\dot\gamma = (q - p) + \dot\eta$, and g is the Euclidean dot product.

(2) Expand the square and integrate term by term; $\int_0^1 \mathrm{d}t = 1$.

(3) $\int_0^1 \dot\eta \, \mathrm{d}t = \eta(1) - \eta(0) = 0$.

So E is *exactly* quadratic around the straight line $\eta = 0$. It has a single critical point, the straight line, which is the global minimum with $E = |q - p|^2$. In the sine modes (7.93), $\eta = \sum a_k \sin k\pi t$,

$$\int_0^1 |\dot\eta|^2 \, \mathrm{d}t \overset{(1)}{=} \sum_{k,l} (k\pi)(l\pi)\, a_k \cdot a_l \int_0^1 \cos k\pi t \, \cos l\pi t \, \mathrm{d}t \overset{(2)}{=} \sum_{k \geq 1} \frac{(k\pi)^2}{2} \, |a_k|^2 .$$

(1) Differentiate the series term by term.

(2) $\int_0^1 \cos k\pi t \cos l\pi t \, \mathrm{d}t = \tfrac{1}{2}\delta_{kl}$ for $k, l \geq 1$.

The Hessian is diagonal in the sine modes, with infinitely many eigenvalues, all positive. The index is 0. The Morse series is $M_t(E) = 1$, which is the Poincaré series of a point. This is right: $P\mathbb{R}^n(p, q)$ is contractible (slide every path to the straight line along $\gamma_s = (1 - s)\gamma + s\,\gamma_{\text{line}}$).

7.8.2 First variation: the critical points are geodesics

Nash says that the extrema of E "are easily seen to be the geodesics" and writes the geodesic equation [Nash (7.56)]. Let us see it. Take a variation γ_ε with variation field V, and write $\dot{} = \mathrm{d}/\mathrm{d}t$. In a chart,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\varepsilon} E(\gamma_\varepsilon)\Big|_{\varepsilon=0} &\overset{(1)}{=} \int_0^1 \Big[\partial_k g_{ij}\, V^k \dot\gamma^i \dot\gamma^j + 2\, g_{ij}\, \dot\gamma^i \dot V^j \Big] \mathrm{d}t \\
&\overset{(2)}{=} \int_0^1 \Big[\partial_k g_{ij}\, \dot\gamma^i \dot\gamma^j - 2\, \frac{\mathrm{d}}{\mathrm{d}t}\big(g_{kj}\dot\gamma^j\big) \Big] V^k \, \mathrm{d}t \\
&\overset{(3)}{=} \int_0^1 \Big[\partial_k g_{ij}\, \dot\gamma^i \dot\gamma^j - 2\, \partial_i g_{kj}\, \dot\gamma^i \dot\gamma^j - 2\, g_{kj}\, \ddot\gamma^j \Big] V^k \, \mathrm{d}t \\
&\overset{(4)}{=} -2 \int_0^1 \Big[g_{kj}\, \ddot\gamma^j + \Gamma_{kij}\, \dot\gamma^i \dot\gamma^j \Big] V^k \, \mathrm{d}t \\
&\overset{(5)}{=} -2 \int_0^1 g_{kl}\, V^k \Big[\ddot\gamma^l + \Gamma^l_{ij}\, \dot\gamma^i \dot\gamma^j \Big] \mathrm{d}t .
\end{aligned} \tag{7.94}$$

(1) Differentiate under the integral. The metric depends on ε through its argument $\gamma_\varepsilon(t)$, giving $\partial_k g_{ij} V^k$; each velocity gives $\partial_\varepsilon \dot\gamma^i_\varepsilon = \dot V^i$, and the two such terms are equal because g_{ij} is symmetric.

(2) Integrate the second term by parts. The boundary term $2\big[g_{kj}\dot\gamma^j V^k\big]_0^1$ vanishes because $V(0) = V(1) = 0$.

(3) Chain rule: $\frac{\mathrm{d}}{\mathrm{d}t} g_{kj}(\gamma(t)) = \partial_i g_{kj}\, \dot\gamma^i$.

(4) Because $\dot\gamma^i\dot\gamma^j$ is symmetric, $2\partial_i g_{kj}\dot\gamma^i\dot\gamma^j = (\partial_i g_{kj} + \partial_j g_{ki})\dot\gamma^i\dot\gamma^j$. Then the bracket is $-2\Gamma_{kij}\dot\gamma^i\dot\gamma^j$ with the Christoffel symbols of the first kind $\Gamma_{kij} = \tfrac12(\partial_i g_{kj} + \partial_j g_{ki} - \partial_k g_{ij})$.

(5) Raise the index: $\Gamma^l_{ij} = g^{lk}\Gamma_{kij}$, so $\Gamma_{kij} = g_{kl}\Gamma^l_{ij}$.

If this vanishes for *every* variation field V, the bracket must vanish at every t. (If it were non-zero somewhere, choose V to be the bracket times a bump function around that t; then the integrand is ≥ 0 and not identically zero.) So the critical points of E are the solutions of

$$\frac{\mathrm{d}^2\gamma^i}{\mathrm{d}t^2} + \Gamma^i_{jk}\, \frac{\mathrm{d}\gamma^j}{\mathrm{d}t}\frac{\mathrm{d}\gamma^k}{\mathrm{d}t} = 0, \tag{7.95}$$

the **geodesic equation** [Nash (7.56)]. It is the Euler–Lagrange equation of the Lagrangian $g_{ij}\dot x^i \dot x^j$: Newton's law with no force, written in curvilinear coordinates. In invariant language (7.94) reads

$$\mathrm{d}E_\gamma(V) = -2 \int_0^1 \Big\langle V,\ \frac{D\dot\gamma}{\mathrm{d}t} \Big\rangle \mathrm{d}t, \tag{7.96}$$

where $D/\mathrm{d}t$ is the covariant derivative along γ (Levi-Civita connection, Chapter T2). So the "gradient" of E is -2 times the covariant acceleration, and a critical point is an unaccelerated path.

Geodesics have constant speed. Along a solution of (7.95), the speed is constant:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\big(g_{ij}\dot\gamma^i\dot\gamma^j\big) &\overset{(1)}{=} \partial_k g_{ij}\, \dot\gamma^k\dot\gamma^i\dot\gamma^j + 2\, g_{ij}\, \ddot\gamma^i\dot\gamma^j \\
&\overset{(2)}{=} 2\,\Gamma_{jki}\, \dot\gamma^j\dot\gamma^k\dot\gamma^i + 2\, g_{ij}\, \ddot\gamma^i\dot\gamma^j \\
&\overset{(3)}{=} 2\, g_{ij}\, \dot\gamma^j \big(\ddot\gamma^i + \Gamma^i_{kl}\dot\gamma^k\dot\gamma^l\big) \overset{(4)}{=} 0.
\end{aligned}$$

(1) Chain rule and product rule; the two velocity terms are equal by symmetry of g.
(2) Fully contracted with three velocities, the three terms of $2\Gamma_{jki} = \partial_k g_{ji} + \partial_i g_{jk} - \partial_j g_{ki}$ are all equal to $\partial_k g_{ij}\dot\gamma^k\dot\gamma^i\dot\gamma^j$, so $2\Gamma_{jki}\dot\gamma^j\dot\gamma^k\dot\gamma^i = (1+1-1)\,\partial_k g_{ij}\dot\gamma^k\dot\gamma^i\dot\gamma^j$.
(3) $\Gamma_{jkl} = g_{ji}\Gamma^i_{kl}$; relabel dummy indices.
(4) The geodesic equation (7.95).

This is energy conservation for the free particle. It will matter in a moment.

Worked example 7.49: Great circles on the round sphere

On the unit sphere, $g = \mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2$, so $g_{\theta\theta} = 1$, $g_{\phi\phi} = \sin^2\theta$, $g_{\theta\phi} = 0$. The only non-constant component is $g_{\phi\phi}$, and it depends only on θ. The non-zero Christoffel symbols are

$$\Gamma^\theta_{\phi\phi} \overset{(1)}{=} \tfrac12 g^{\theta\theta}\big(2\partial_\phi g_{\theta\phi} - \partial_\theta g_{\phi\phi}\big) \overset{(2)}{=} -\tfrac12\,\partial_\theta(\sin^2\theta) = -\sin\theta\cos\theta,$$
$$\Gamma^\phi_{\theta\phi} = \Gamma^\phi_{\phi\theta} \overset{(3)}{=} \tfrac12 g^{\phi\phi}\,\partial_\theta g_{\phi\phi} \overset{(4)}{=} \frac{2\sin\theta\cos\theta}{2\sin^2\theta} = \cot\theta.$$

(1) $\Gamma^i_{jk} = \tfrac12 g^{il}(\partial_j g_{lk} + \partial_k g_{lj} - \partial_l g_{jk})$ with $i=\theta$, $j=k=\phi$; the metric is diagonal so only $l=\theta$ survives.
(2) $g^{\theta\theta} = 1$ and $g_{\theta\phi} = 0$.
(3) Same formula with $i=\phi$, $j=\theta$, $k=\phi$: only the term $\partial_\theta g_{\phi\phi}$ survives.
(4) $g^{\phi\phi} = 1/\sin^2\theta$.

The geodesic equations (7.95) are therefore

$$\ddot\theta - \sin\theta\cos\theta\,\dot\phi^2 = 0, \qquad \ddot\phi + 2\cot\theta\,\dot\theta\,\dot\phi = 0.$$

- A *meridian* $\phi = \phi_0$, $\theta = \theta_0 + ct$: $\ddot\theta = 0$ and $\dot\phi = 0$, so both equations hold. It is a geodesic.
- The *equator* $\theta = \pi/2$, $\phi = ct$: $\ddot\theta = 0$ and $\sin\theta\cos\theta = 0$; $\ddot\phi = 0$ and $\dot\theta = 0$. It is a geodesic.
- A *latitude circle* $\theta = \theta_0 \neq \pi/2$, $\phi = ct$: the first equation gives $-\sin\theta_0\cos\theta_0\,c^2 \neq 0$. It is *not* a geodesic. To keep a particle on it you must push it towards the nearer pole (Figure 7.29).

Since rotations are isometries, rotating a meridian gives all the great circles. These are all the geodesics of the round sphere.

7.8.3 Energy versus length

The obvious functional for geodesics is the *length*

$$L(\gamma) = \int_0^1 \left|\frac{\mathrm{d}\gamma}{\mathrm{d}t}\right| \mathrm{d}t. \tag{7.97}$$

Nash records the relation between the two [Nash (7.57), (7.58)]:

$$E(\gamma) \;\geq\; L(\gamma)^2 \quad \text{for every path,} \qquad E(\gamma) = L(\gamma)^2 \quad \text{if } \gamma \text{ is a geodesic.} \tag{7.98}$$

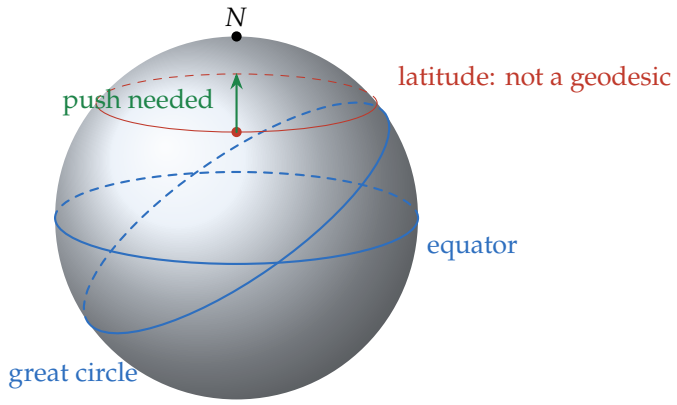


Figure 7.29: Critical points of E on the round sphere. Great circles (blue), such as the equator or a tilted circle, satisfy the geodesic equation. A latitude circle (red) does not: a particle released along it drifts towards the equator, and to keep it on the latitude one must push it towards the pole (green), which is the term $-\sin\theta_0\cos\theta_0\,\dot\phi^2 \neq 0$.

The inequality is the Cauchy–Schwarz inequality for the functions $|\dot\gamma|$ and 1 on $[0,1]$:

$$L(\gamma)^2 \overset{(1)}{=} \left(\int_0^1 |\dot\gamma|\cdot 1\,\mathrm{d}t\right)^2 \overset{(2)}{\leq} \left(\int_0^1 |\dot\gamma|^2\,\mathrm{d}t\right)\left(\int_0^1 1^2\,\mathrm{d}t\right) \overset{(3)}{=} E(\gamma).$$

(1) Definition (7.97), with a factor 1 inserted.
(2) Cauchy–Schwarz, $(\int fh)^2 \leq \int f^2 \int h^2$, with equality if and only if f is a constant multiple of h, i.e. $|\dot\gamma|$ is constant.
(3) Definition (7.90) and $\int_0^1 1\,\mathrm{d}t = 1$.

A geodesic has constant speed, so equality holds: $|\dot\gamma| = L$ and $E = \int_0^1 L^2\mathrm{d}t = L^2$. Consequently the minimum of E on $PM(p,q)$ is $d(p,q)^2$, attained exactly at the shortest geodesics traversed at constant speed.

Why use E and not L? Because L has too much symmetry. If $\varphi\colon[0,1]\to[0,1]$ is an increasing diffeomorphism, then $L(\gamma\circ\varphi) = L(\gamma)$. So every critical point of L comes with an infinite-dimensional family of reparametrised copies, all critical. The Hessian of L has an infinite-dimensional kernel. This is as degenerate as a function can be. E is not reparametrisation invariant: it picks out the constant-speed representative and penalises the others.

Worked example 7.50: Reparametrising a straight line

In $\mathbb{R}^n$ let $\gamma(t) = p + t(q-p)$ and $\tilde\gamma(t) = \gamma(t^2) = p + t^2(q-p)$. Both trace the same segment, so $L(\tilde\gamma) = L(\gamma) = |q-p|$. But

$$E(\tilde\gamma) \overset{(1)}{=} \int_0^1 |2t\,(q-p)|^2\,\mathrm{d}t \overset{(2)}{=} \tfrac{4}{3}\,|q-p|^2 \;>\; |q-p|^2 = E(\gamma).$$

(1) $\dot{\tilde\gamma} = 2t\,(q-p)$.
(2) $\int_0^1 4t^2\mathrm{d}t = \tfrac{4}{3}$.

The slow-then-fast parametrisation costs extra energy, so it is not a critical point of E even though it is one of L.

Dictionary: Length and energy in physics

$L(\gamma)=\int\lvert\dot\gamma\rvert\,\mathrm{d}t$	relativistic point particle, $S=-m\int\mathrm{d}s$: reparametrisation invariant, i.e. a gauge symmetry (worldline diffeomorphisms)
$E(\gamma)=\int\lvert\dot\gamma\rvert^2\,\mathrm{d}t$	the "einbein" or Polyakov form $\frac12\int e^{-1}\dot x^2\mathrm{d}t$ in the gauge $e=1$: the gauge is fixed, only constant-speed paths are critical
$E\geq L^2$, equality on geodesics	eliminating e by its equation of motion returns the square-root action

The same pair (Nambu–Goto area versus Polyakov energy) reappears for strings in Chapter 9.

7.8.4 Second variation: the Hessian of the energy

To do Morse theory we need the index of each critical point. So we need the Hessian of E at a geodesic γ. Nash does not compute it; here it is.

Conventions. $D/\mathrm{d}t$ and $D/\mathrm{d}\varepsilon$ are covariant derivatives along the two-parameter family $\gamma_\varepsilon(t)$ (Levi-Civita connection). The Riemann tensor is $R(X,Y)Z=\nabla_X\nabla_Y Z-\nabla_Y\nabla_X Z-\nabla_{[X,Y]}Z$. We use three facts from Riemannian geometry (Chapter T2):

(a) metric compatibility: $\partial_\varepsilon\langle A,B\rangle=\langle\frac{D}{\mathrm{d}\varepsilon}A,B\rangle+\langle A,\frac{D}{\mathrm{d}\varepsilon}B\rangle$, and the same for t;

(b) no torsion: $\frac{D}{\mathrm{d}\varepsilon}\partial_t\gamma=\frac{D}{\mathrm{d}t}\partial_\varepsilon\gamma$;

(c) curvature measures the failure of the two covariant derivatives to commute: $\frac{D}{\mathrm{d}\varepsilon}\frac{D}{\mathrm{d}t}W-\frac{D}{\mathrm{d}t}\frac{D}{\mathrm{d}\varepsilon}W=R(\partial_\varepsilon\gamma,\partial_t\gamma)W$.

In a two-parameter variation the ε and t coordinate lines commute, which is why no bracket term appears in (c).

Let γ be a geodesic and γ_ε a variation with variation field $V=\partial_\varepsilon\gamma_\varepsilon|_{\varepsilon=0}$, fixed endpoints. Write $\dot\gamma_\varepsilon=\partial_t\gamma_\varepsilon$.

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\varepsilon}E(\gamma_\varepsilon)&\overset{(1)}{=}2\int_0^1\Big\langle\frac{D}{\mathrm{d}\varepsilon}\dot\gamma_\varepsilon,\dot\gamma_\varepsilon\Big\rangle\mathrm{d}t\overset{(2)}{=}2\int_0^1\Big\langle\frac{D}{\mathrm{d}t}\partial_\varepsilon\gamma_\varepsilon,\dot\gamma_\varepsilon\Big\rangle\mathrm{d}t,\\
\frac{\mathrm{d}^2}{\mathrm{d}\varepsilon^2}E(\gamma_\varepsilon)&\overset{(3)}{=}2\int_0^1\Big[\Big\langle\frac{D}{\mathrm{d}\varepsilon}\frac{D}{\mathrm{d}t}\partial_\varepsilon\gamma_\varepsilon,\dot\gamma_\varepsilon\Big\rangle+\Big\langle\frac{D}{\mathrm{d}t}\partial_\varepsilon\gamma_\varepsilon,\frac{D}{\mathrm{d}t}\partial_\varepsilon\gamma_\varepsilon\Big\rangle\Big]\mathrm{d}t\\
&\overset{(4)}{=}2\int_0^1\Big[\Big\langle\frac{D}{\mathrm{d}t}\frac{D}{\mathrm{d}\varepsilon}\partial_\varepsilon\gamma_\varepsilon+R(\partial_\varepsilon\gamma_\varepsilon,\dot\gamma_\varepsilon)\partial_\varepsilon\gamma_\varepsilon,\dot\gamma_\varepsilon\Big\rangle+\Big|\frac{D}{\mathrm{d}t}\partial_\varepsilon\gamma_\varepsilon\Big|^2\Big]\mathrm{d}t.
\end{aligned}\tag{7.99}$$

(1) Differentiate $\langle\dot\gamma_\varepsilon,\dot\gamma_\varepsilon\rangle$ using (a).
(2) No torsion (b).
(3) Differentiate again using (a); the second factor gives $\frac{D}{\mathrm{d}\varepsilon}\dot\gamma_\varepsilon=\frac{D}{\mathrm{d}t}\partial_\varepsilon\gamma_\varepsilon$ by (b) once more.
(4) Commute the covariant derivatives with (c), $W=\partial_\varepsilon\gamma_\varepsilon$.

Now set $\varepsilon=0$. The first term is a total derivative plus something that vanishes on a geodesic:

$$\int_0^1\Big\langle\frac{D}{\mathrm{d}t}A,\dot\gamma\Big\rangle\mathrm{d}t\overset{(1)}{=}\big[\langle A,\dot\gamma\rangle\big]_0^1-\int_0^1\Big\langle A,\frac{D\dot\gamma}{\mathrm{d}t}\Big\rangle\mathrm{d}t\overset{(2)}{=}0,\qquad A:=\frac{D}{\mathrm{d}\varepsilon}\partial_\varepsilon\gamma_\varepsilon\Big|_{\varepsilon=0}.$$

(1) Integration by parts with metric compatibility (a).
(2) At the endpoints $\gamma_\varepsilon(0) = p$ for all ε, so $\partial_\varepsilon \gamma_\varepsilon$ and its ε-derivative vanish there: $A(0) = A(1) = 0$. And $D\dot\gamma/\mathrm{d}t = 0$ because γ is a geodesic.

The curvature term is rewritten using the antisymmetry of $\langle R(X,Y)Z, W\rangle$ in its last two slots,

$$\big\langle R(V,\dot\gamma)V,\dot\gamma\big\rangle = -\big\langle R(V,\dot\gamma)\dot\gamma, V\big\rangle.$$

Writing $V' = DV/\mathrm{d}t$, we arrive at the **second variation formula**:

$$\boxed{\frac{\mathrm{d}^2}{\mathrm{d}\varepsilon^2}E(\gamma_\varepsilon)\Big|_{\varepsilon=0} = 2\int_0^1 \Big(|V'|^2 - \big\langle R(V,\dot\gamma)\dot\gamma, V\big\rangle\Big)\mathrm{d}t} \tag{7.100}$$

The first term is the flat-space answer of Example 7.48: bending a path costs energy. The second term is the effect of curvature. For positive curvature, $\langle R(V,\dot\gamma)\dot\gamma, V\rangle = K\,|V_\perp|^2\,|\dot\gamma|^2 > 0$ (with K the sectional curvature and $V_\perp$ the part of V normal to $\dot\gamma$). Positive curvature *lowers* the energy of a deformation: nearby geodesics are focused together, as meridians are focused at the south pole.

The Hessian as an operator. Integrating the first term by parts once more (boundary terms vanish as $V(0) = V(1) = 0$) gives

$$\frac{\mathrm{d}^2E}{\mathrm{d}\varepsilon^2} = 2\int_0^1 \big\langle V,\ \mathcal{J}V\big\rangle\mathrm{d}t, \qquad \mathcal{J}V := -\frac{D^2V}{\mathrm{d}t^2} - R(V,\dot\gamma)\dot\gamma. \tag{7.101}$$

$\mathcal{J}$ is the **Jacobi operator**. In a parallel orthonormal frame along γ it is the matrix Schrödinger operator $-\mathrm{d}^2/\mathrm{d}t^2 + U(t)$, with "potential" $U_{ab}(t) = -\langle R(e_a,\dot\gamma)\dot\gamma, e_b\rangle$, acting on vector-valued functions on $[0,1]$ with Dirichlet boundary conditions.

Key idea: The index of a geodesic is finite

The Hessian of E at a geodesic γ is the quadratic form of the Jacobi operator $\mathcal{J}$ with Dirichlet conditions. Its eigenvalues, like those of any Schrödinger operator on a bounded interval, form a discrete list tending to $+\infty$. Only finitely many are negative. So although $PM(p,q)$ is infinite dimensional:

- the **index** λ_γ (number of negative eigenvalues, with multiplicity) is *finite*;
- the **nullity** (dimension of the kernel) is finite, at most $n-1$.

Physical picture: The fluctuation operator of the path integral

In the semiclassical evaluation of the propagator $\langle q|e^{-iHT/\hbar}|p\rangle$ of a particle on M, one expands the action to second order around each classical path. The quadratic part is exactly (7.101), and the Gaussian integral over fluctuations gives $(\det\mathcal{J})^{-1/2}$. Each negative eigenvalue of $\mathcal{J}$ contributes a phase $e^{-i\pi/2}$, so a path of index λ_γ carries the phase $e^{-i\pi\lambda_\gamma/2}$. This is the **Maslov phase** of the Van Vleck–Gutzwiller formula. The Morse index of a geodesic is something quantum mechanics measures.

7.8.5 Jacobi fields, conjugate points and the Morse index theorem

The kernel of $\mathcal{J}$ has a direct geometric meaning.

Working definition 7.21 (Jacobi field, conjugate point)

A **Jacobi field** along a geodesic γ is a solution of $\mathcal{J}V = 0$, i.e.

$$\frac{D^2V}{\mathrm{d}t^2} + R(V, \dot\gamma)\dot\gamma = 0. \tag{7.102}$$

[$D/\mathrm{d}t$ is the covariant derivative along γ. The curvature term $R(V, \dot\gamma)\dot\gamma$ says how nearby geodesics bend towards or away from γ] It is the variation field of a family of *geodesics* (linearise the geodesic equation). The point $\gamma(t_0)$ is **conjugate** to $p = \gamma(0)$ along γ if there is a non-zero Jacobi field with $V(0) = 0$ and $V(t_0) = 0$. The number of independent such fields is the **multiplicity** of the conjugate point. *Test*: send out the geodesics from p with slightly different initial directions; $\gamma(t_0)$ is conjugate if, to first order, they come back together there.

So the nullity of E at γ is the multiplicity of $q = \gamma(1)$ as a conjugate point, and γ is a non-degenerate critical point exactly when q is not conjugate to p along γ. The index is counted by the conjugate points *before* q:

Key idea: Morse index theorem

The index of the geodesic $\gamma\colon [0,1] \to M$ as a critical point of E is

$$\lambda_\gamma = \sum_{0<t_0<1} \big(\text{multiplicity of } \gamma(t_0) \text{ as a point conjugate to } \gamma(0)\big). \tag{7.103}$$

The sum is finite.

Why. Restrict the geodesic to $[0, T]$ and follow the Dirichlet eigenvalues $\mu_k(T)$ of the Jacobi operator on $[0, T]$ as T grows from 0 to 1.

- For very small T the kinetic term wins: $\mu_k(T) \approx (k\pi/T)^2 > 0$. The index starts at 0.
- The eigenvalues decrease as T increases. (Extending an eigenfunction on $[0, T]$ by zero gives a trial function on $[0, T']$; the min–max principle then pushes every eigenvalue down.)
- An eigenvalue is *zero* at time T exactly when there is a Jacobi field vanishing at 0 and T, i.e. when $\gamma(T)$ is conjugate to $\gamma(0)$, and the number of zero eigenvalues is the multiplicity.

So each time T passes a conjugate time, that many eigenvalues cross from positive to negative, and none cross back. Counting crossings up to $T = 1$ gives (7.103). Figure 7.31 shows the crossings on the sphere. (The same reasoning in quantum mechanics is the Sturm oscillation theorem: the kth eigenfunction has $k - 1$ interior nodes.)

Worked example 7.51: Every geodesic of the round sphere S^n

Let γ be a great-circle arc of length L on the unit sphere S^n, starting at p. Use the arc-length parameter $s = Lt \in [0, L]$ and write $'$ for $D/\mathrm{d}s$. (Changing from t to s multiplies the Hessian by a positive constant, which changes no signs.) Choose a parallel orthonormal frame along γ: $e_1 = \gamma'$, and $e_2, \dots, e_n$ normal to it. Write $V = \sum_a v^a(s)\, e_a$; then $V'' = \sum_a v^a{}'' e_a$, because the frame is parallel. The unit sphere has constant curvature

1: $R(X,Y)Z = \langle Y,Z\rangle X - \langle X,Z\rangle Y$. Hence

$$R(V,\gamma')\gamma' \overset{(1)}{=} \langle\gamma',\gamma'\rangle V - \langle V,\gamma'\rangle\gamma' \overset{(2)}{=} V - v^1 e_1 = \sum_{a=2}^{n} v^a e_a,$$

$$\mathcal{J}V \overset{(3)}{=} -v^1{}'' e_1 + \sum_{a=2}^{n}\bigl(-v^a{}'' - v^a\bigr)e_a.$$

(1) Constant curvature formula with $X = V, Y = Z = \gamma'$.
(2) Unit speed, $|\gamma'| = 1$; and $\langle V,\gamma'\rangle = v^1$.
(3) $\mathcal{J}V = -V'' - R(V,\gamma')\gamma'$ (7.101).

The operator splits into n decoupled Sturm–Liouville problems on $[0,L]$ with $v(0) = v(L) = 0$. The eigenfunctions are $\sin(k\pi s/L)$, $k = 1,2,\dots$, with eigenvalues

$$\text{tangential } (a=1): \quad \mu_k = \left(\frac{k\pi}{L}\right)^2 > 0, \qquad \text{normal } (a = 2,\dots,n): \quad \mu_k = \left(\frac{k\pi}{L}\right)^2 - 1.$$

A normal eigenvalue is negative when $k\pi < L$, and each value of k occurs $n-1$ times. Therefore

$$\lambda_\gamma = (n-1)\cdot\#\{k \geq 1 :\ k\pi < L\}, \qquad \text{nullity} = \begin{cases} n-1, & L \in \pi\mathbb{Z}, \\ 0, & \text{otherwise.} \end{cases} \tag{7.104}$$

Check against the index theorem. The Jacobi fields vanishing at $s = 0$ are $V = \sin s\, e_a$ ($a \geq 2$), the variation fields of the great circles through p. They vanish again at $s = \pi, 2\pi, \dots$: the antipode $-p$ and p itself, each conjugate to p with multiplicity $n-1$. All geodesics leaving p refocus there (Figure 7.30). The number of conjugate points strictly inside $(0,L)$, counted with multiplicity, is $(n-1)\cdot\#\{k : k\pi < L\}$, in agreement with (7.104).
For example, on S^2 an arc shorter than a half circle has index 0 (it is the shortest path); an arc of length between π and 2π has index 1: it can be shortened by sliding it off the great circle, over the antipodal point.

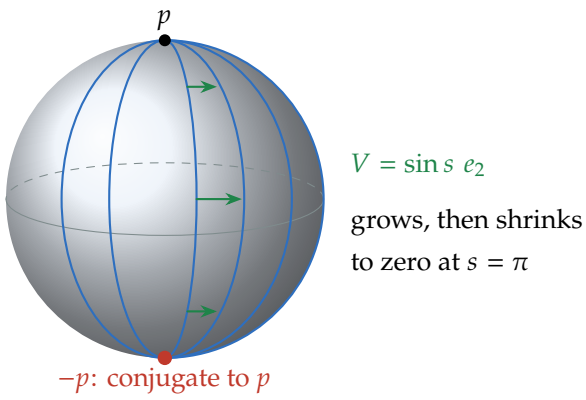


Figure 7.30: Conjugate points on the sphere. The great circles leaving p (blue) spread out and then refocus at the antipode $-p$. The Jacobi field $V = \sin s\, e_2$ (green) measures their separation: it vanishes at $s = 0$ and again at $s = \pi$, so $-p$ is conjugate to p, with multiplicity $n-1$ on S^n (one for each normal direction). A geodesic that passes through $-p$ picks up $n-1$ negative directions.

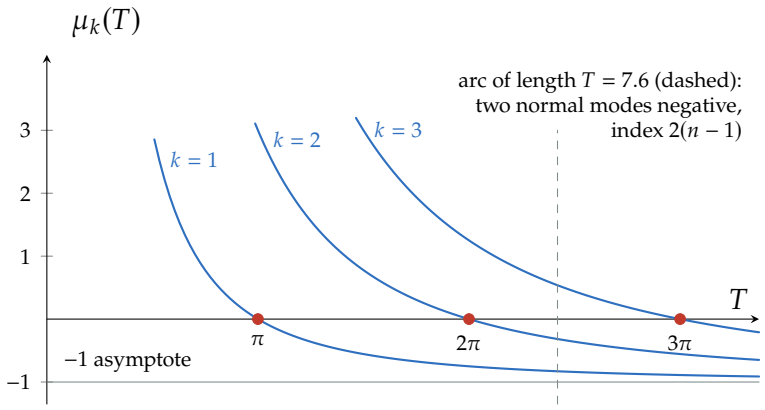


Figure 7.31: Eigenvalues of the normal Jacobi operator $-\mathrm{d}^2/\mathrm{d}s^2 - 1$ on $[0, T]$ for great-circle arcs of the unit sphere, $\mu_k(T) = (k\pi/T)^2 - 1$. As the arc gets longer every eigenvalue decreases; the kth one crosses zero (red dot) exactly when $T = k\pi$, the conjugate times. The index of an arc of length T is the number of curves below the axis (times $n-1$), which is the number of red dots to the left of T: the Morse index theorem.

Worked example 7.52: Flat and negatively curved spaces: every geodesic is a local minimum

If the sectional curvature is ≤ 0, then $\langle R(V,\dot\gamma)\dot\gamma, V\rangle \le 0$, and (7.100) gives

$$\frac{\mathrm{d}^2E}{\mathrm{d}\varepsilon^2} \overset{(1)}{\ge} 2\int_0^1 |V'|^2\,\mathrm{d}t \overset{(2)}{>} 0 \qquad (V \neq 0).$$

(1) The curvature term is non-negative after the minus sign.
(2) If $V' \equiv 0$ then V is parallel, and $V(0) = 0$ forces $V \equiv 0$.

So every geodesic has index 0 and nullity 0: no conjugate points at all. On the flat torus $T^n = \mathbb{R}^n/2\pi\mathbb{Z}^n$ the geodesics from p to q are the straight lines from $\tilde p$ to the lattice translates $\tilde q + 2\pi k$, $k \in \mathbb{Z}^n$: infinitely many geodesics, all local minima, one in each homotopy class of paths.

Try it: A geodesic on the cylinder

On the cylinder $S^1 \times \mathbb{R}$ with the flat metric $\mathrm{d}\phi^2 + \mathrm{d}z^2$, how many geodesics join $(0,0)$ to $(\pi/2, 1)$, and what are their indices?
Answer: lift to the plane: the straight lines to $(\pi/2 + 2\pi m, 1)$, $m \in \mathbb{Z}$. Infinitely many, one for each winding number m, all of index 0 (the cylinder is flat). Their energies are $(\pi/2 + 2\pi m)^2 + 1$.

7.8.6 Loop spaces and path spaces

We now know the critical points of E on $PM(p,q)$ (geodesics) and how to compute their indices (conjugate points). To run the Morse inequalities we also need the other side of the equation: the Betti numbers of the space $PM(p,q)$ itself. This subsection collects what we need about that space.

Working definition 7.22 (based loop space)

Fix a base point $p \in M$. The **based loop space** is

$$\Omega M = \{\gamma\colon [0,1] \to M \mid \gamma(0) = \gamma(1) = p\} = PM(p,p). \tag{7.105}$$

A point of ΩM is a loop that starts and ends at p. Two loops can be multiplied by running first one and then the other at double speed. The *connected components* of ΩM are the homotopy classes of loops [two loops count as one class when one can be deformed continuously into the other, keeping p fixed], i.e. the elements of $\pi_1(M,p)$.

$PM(p,q)$ and ΩM have the same topology. Choose once and for all a path β from q back to p. Then

$$\Phi\colon PM(p,q) \longrightarrow \Omega M, \qquad \gamma \longmapsto \gamma * \beta \quad (\text{first } \gamma,\ \text{then } \beta), \tag{7.106}$$

turns every path from p to q into a loop at p. Composing with the reverse path $\bar\beta$ goes back: $\ell \mapsto \ell * \bar\beta$. Both composites are homotopic to the identity (a path followed by $\beta * \bar\beta$ can be shrunk back to itself by pulling in the "there and back" tail). So

$$PM(p,q) \;\simeq\; \Omega M \tag{7.107}$$

for any p, q in a connected M. In particular they have the same Betti numbers. The advantage of $PM(p,q)$ is that for a well-chosen q the energy is a Morse function; on ΩM itself the constant loop at p is a critical manifold, and p is conjugate to itself along closed geodesics on symmetric spaces, which makes things degenerate.

The homotopy groups shift down by one. A map of a sphere S^k into ΩM is a k-parameter family of loops at p; sweeping out the loop parameter as well gives a map $S^k \times [0,1] \to M$ that sends the ends $S^k \times \{0,1\}$ to p. Collapsing those ends, and the k-parameter family's base point, turns $S^k \times [0,1]$ into S^{k+1}. This is the **suspension** correspondence, and it gives

$$\pi_k(\Omega M) \;\cong\; \pi_{k+1}(M). \tag{7.108}$$

(The more systematic reason: the space $PM(p,\cdot)$ of all paths starting at p is contractible, because every path can be reeled back to p along itself, and the endpoint map $PM(p,\cdot) \to M$ has fibre ΩM. The long exact sequence of this fibration, with a contractible total space, gives (7.108).) This is the reason Morse theory on ΩM can tell us about the homotopy groups of M, which is how Bott will use it.

Worked example 7.53: The loop space of a circle

Let $M = S^1 = \mathbb{R}/2\pi\mathbb{Z}$ with the flat metric, $p = 0$. A loop at 0 lifts to a path in $\mathbb{R}$ from 0 to $2\pi m$ for a unique integer m, the winding number. So

$$\Omega S^1 = \bigsqcup_{m\in\mathbb{Z}} \Omega_m S^1, \qquad \Omega_m S^1 \;\simeq\; P\mathbb{R}(0,2\pi m) \;\simeq\; \text{point}.$$

Each component is contractible by Example 7.48. The energy on $\Omega_m S^1$ has exactly one critical point, the uniform winding $\gamma_m(t) = 2\pi m t$, of index 0 and energy $E = (2\pi m)^2$. The Morse series is $\sum_{m\in\mathbb{Z}} t^0$, an infinite sum of 1's, and $b_0(\Omega S^1) = \infty$: one component for each $m \in \pi_1(S^1) = \mathbb{Z}$. Consistent with (7.108): $\pi_0(\Omega S^1) = \pi_1(S^1) = \mathbb{Z}$ and $\pi_k(\Omega S^1) = \pi_{k+1}(S^1) = 0$ for $k \geq 1$. All the geodesics here are explained by the fundamental group; the interesting case is a simply connected M, where ΩM is connected and the geodesics must be explained by higher Betti numbers.

7.8.7 Morse theory on the path space of the sphere

Take the unit sphere S^n ($n \geq 2$) with the round metric, and two points p, q that are neither equal nor antipodal. Let $\theta = d(p,q) \in (0,\pi)$ be the angle between them. We list all the critical

points of E on $PS^n(p,q)$, their indices, and read off the Morse series.

The critical points. A geodesic from p to q is a great-circle arc (Example 7.49). Because $q \neq \pm p$, exactly one great circle passes through both, the intersection of S^n with the plane spanned by p and q. A geodesic may leave p along it in either direction and may wind around it any number of times before stopping at q:

- towards q, winding k times first: length $L = 2\pi k + \theta$;
- away from q (through $-p$), winding k times first: length $L = 2\pi(k+1) - \theta$.

Order them by length and call them $\gamma_0, \gamma_1, \gamma_2, \ldots$:

$$L(\gamma_{2k}) = 2\pi k + \theta, \qquad L(\gamma_{2k+1}) = 2\pi(k+1) - \theta, \qquad k = 0, 1, 2, \ldots \tag{7.109}$$

Since $0 < \theta < \pi$, these lengths strictly increase with j and none is a multiple of π.

The indices. By (7.104) the index of an arc of length L is $(n-1)$ times the number of positive integers m with $m\pi < L$.

$$\lambda(\gamma_{2k}) \overset{(1)}{=} (n-1)\,\#\{m \geq 1 :\ m\pi < 2\pi k + \theta\} \overset{(2)}{=} (n-1)\cdot 2k,$$
$$\lambda(\gamma_{2k+1}) \overset{(3)}{=} (n-1)\,\#\{m \geq 1 :\ m\pi < 2\pi k + 2\pi - \theta\} \overset{(4)}{=} (n-1)(2k+1).$$

(1) The index formula (7.104) with $L = 2\pi k + \theta$.
(2) $2k\pi < 2\pi k + \theta < (2k+1)\pi$ because $0 < \theta < \pi$, so $m = 1, \ldots, 2k$ qualify and $m = 2k+1$ does not.
(3) The same formula with $L = 2\pi(k+1) - \theta$.
(4) $(2k+1)\pi < 2\pi k + 2\pi - \theta < (2k+2)\pi$, so $m = 1, \ldots, 2k+1$ qualify.

So, in one line,

$$\lambda(\gamma_j) = j\,(n-1), \qquad j = 0, 1, 2, \ldots, \tag{7.110}$$

and every γ_j is non-degenerate, because its length is never a multiple of π, so q is never conjugate to p along it. Geometrically: each time the arc passes through $-p$ or through p itself (the conjugate points, Figure 7.30), it picks up $n-1$ more directions in which it can be shortened. Figure 7.32 shows the first few.

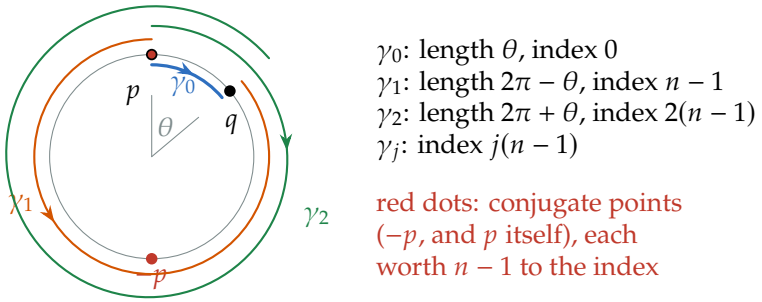


Figure 7.32: All the geodesics from p to q on the round S^n, drawn in the plane of the unique great circle through p and q (the curves are pushed apart radially only so that they can be seen). γ_0 goes the short way, γ_1 the long way through $-p$, γ_2 once around and then the short way, and so on. Every passage through a red conjugate point adds $n-1$ to the index, which gives $\lambda(\gamma_j) = j(n-1)$.

The Morse series. One critical point of each index $j(n-1)$, so

$$M_t(E) = \sum_{j=0}^{\infty} t^{\,j(n-1)} = 1 + t^{n-1} + t^{2(n-1)} + \cdots = \frac{1}{1-t^{n-1}}. \tag{7.111}$$

It is an infinite series, since there are infinitely many critical points, but each coefficient is finite (0 or 1). That is all the Morse inequalities need.

The lacunary principle does the rest. For $n \geq 3$ the gaps between the exponents in (7.111) are $n-1 \geq 2$: no two consecutive powers of t appear. The lacunary principle (Section 7.2.7) works degree by degree, so it applies to an infinite series just as well. Suppose $R(t) = \sum r_i t^i \neq 0$ and let r_i be its first non-zero coefficient. Then $(1+t)R(t)$ has the positive coefficient r_i in degree i and $r_i + r_{i+1} > 0$ in degree $i+1$. So $m_i - b_i > 0$ and $m_{i+1} - b_{i+1} > 0$, which needs $m_i \neq 0$ and $m_{i+1} \neq 0$: two consecutive powers in M_t, which do not exist. Hence $R = 0$, E is perfect, and

$$\boxed{P_t(\Omega S^n) = P_t\big(PS^n(p,q)\big) = \frac{1}{1-t^{n-1}}}, \qquad H^i(\Omega S^n;\mathbb{R}) = \begin{cases} \mathbb{R}, & i = 0,\ n-1,\ 2(n-1), \ldots \\ 0, & \text{otherwise.} \end{cases} \tag{7.112}$$

This is a calculation of the cohomology of an infinite-dimensional space from counting great circles. The same argument works with any coefficient field, so there is no torsion.

The Bott–Morse–Smale picture (Section 7.3.4) tells us more. The energy levels $E(\gamma_j) = L_j^2$ increase with j, so as c increases the sublevel set $\{E \leq c\}$ acquires one cell at a time:

$$\Omega S^n \;\simeq\; e^0 \cup e^{n-1} \cup e^{2(n-1)} \cup e^{3(n-1)} \cup \cdots . \tag{7.113}$$

The first two cells already form a sphere $S^{n-1} = e^0 \cup e^{n-1}$. It has a direct meaning: the loops at p that run once over the sphere through $-p$ along a great circle are labelled by their direction at $-p$, a point of the unit sphere S^{n-1} of directions. (Figure 7.33.)

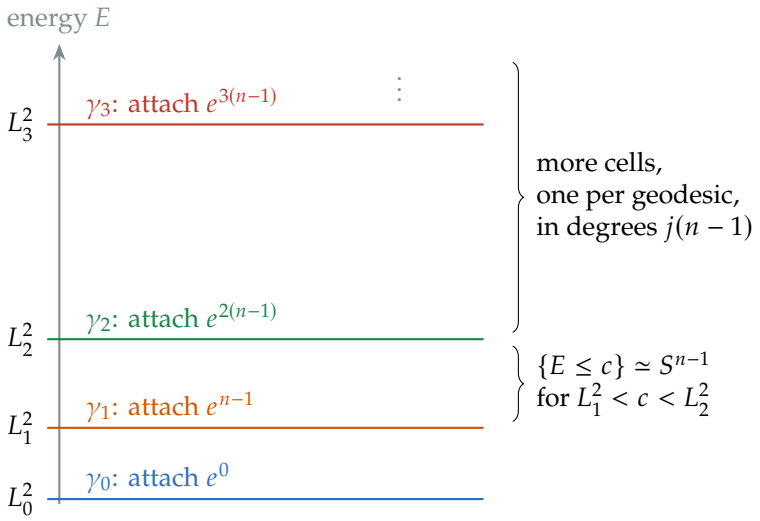


Figure 7.33: The path space of the round S^n, built by raising the energy. Each geodesic γ_j from p to q is a critical level (drawn to scale for $\theta = 1.2$), and crossing it attaches one cell of dimension $\lambda(\gamma_j) = j(n-1)$. Between L_1^2 and L_2^2 the sublevel set is homotopy equivalent to $S^{n-1} = e^0 \cup e^{n-1}$; in the limit one obtains a cell complex with exactly one cell in each degree $j(n-1)$, which is why $P_t(\Omega S^n) = 1/(1-t^{n-1})$.

Worked example 7.54: Checking the answer for S^3 and S^2

S^3. (7.112) gives $P_t(\Omega S^3) = 1 + t^2 + t^4 + \dots$. The lowest non-trivial class is in degree 2. By (7.108) and the Hurewicz theorem (for the simply connected space ΩS^3, the first non-zero homotopy group equals the first non-zero homology group),

$$H_2(\Omega S^3) \overset{(1)}{\cong} \pi_2(\Omega S^3) \overset{(2)}{=} \pi_3(S^3) \overset{(3)}{=} \mathbb{Z}.$$

(1) Hurewicz: ΩS^3 is connected and $\pi_1(\Omega S^3) = \pi_2(S^3) = 0$.
(2) The shift (7.108).
(3) The degree of a map $S^3 \to S^3$; this is the group of instanton numbers of SU(2) gauge theory (Chapter 8).

So $b_2(\Omega S^3) = 1$, as the Morse series says.
S^2. Here $n - 1 = 1$ and (7.111) is $M_t = 1 + t + t^2 + t^3 + \dots$: consecutive powers, so the lacunary principle is silent. An independent calculation settles it. The Hopf fibration $S^1 \to S^3 \to S^2$ gives a homotopy equivalence $\Omega S^2 \simeq S^1 \times \Omega S^3$. (Lift a loop in S^2 to a path in S^3 starting at a fixed point over p; its end lies in the fibre circle S^1 over p. This defines $\Omega S^2 \to S^1$, whose fibre is the space of loops that lift to closed loops, i.e. ΩS^3. The resulting fibration $\Omega S^3 \to \Omega S^2 \to S^1$ is trivial up to homotopy, because a loop of loops mapping once around S^1 exists, and loops in S^2 can be multiplied.) Hence

$$P_t(\Omega S^2) \overset{(1)}{=} P_t(S^1)\, P_t(\Omega S^3) \overset{(2)}{=} \frac{1+t}{1-t^2} \overset{(3)}{=} \frac{1}{1-t} = M_t(E).$$

(1) The Künneth formula for a product.
(2) $P_t(S^1) = 1 + t$ and the S^3 result above.
(3) $1 - t^2 = (1 - t)(1 + t)$.

So the energy is perfect on $PS^2(p, q)$ too, even though counting alone could not show it.

Morse's theorem: infinitely many geodesics for every metric

Now forget the round metric. Put *any* Riemannian metric g on S^n and pick p, q. The Betti numbers of $PS^n(p, q)$ do not know about g: they are those of ΩS^n, computed above using the round metric. For all but a measure-zero set of q, the point q is not conjugate to p along any geodesic (Sard's theorem applied to the exponential map at p: conjugate points are its critical values). Then E is a Morse function on $PS^n(p, q)$ and the weak Morse inequalities give

$$m_{j(n-1)} \;\geq\; b_{j(n-1)}(\Omega S^n) = 1 \qquad \text{for every } j = 0, 1, 2, \dots \tag{7.114}$$

So there is at least one geodesic of each index $j(n - 1)$, and in particular infinitely many.

Key idea: Morse's theorem (Nash p. 192)

For any Riemannian metric on S^n, and any two points p, q that are not conjugate along any geodesic, there are infinitely many geodesics from p to q; at least one has index $j(n - 1)$, for each $j \geq 0$. (With the Lusternik–Schnirelmann version of the argument, Section 7.4.5, the conclusion "infinitely many" survives for conjugate p, q too.)

Picture a potato-shaped S^2. Most of the geodesics we are guaranteed are long curves that wind around it several times; the theorem says that however the potato is shaped, those windings cannot all be undone. The reason is not geometric but topological: the cohomology classes of ΩS^n in degrees $j(n - 1)$ each have to be "carried" by a critical point.

Serre's generalisation. Serre proved in 1951 that for *every* compact simply connected manifold M of positive dimension, infinitely many Betti numbers of ΩM are non-zero (with coefficients in some field). The same inequality then shows: on any compact Riemannian manifold, any two non-conjugate points are joined by infinitely many geodesics. (If M is not simply connected and $\pi_1(M)$ is infinite, the components of ΩM already give infinitely many, as for S^1 in Example 7.53; if π_1 is finite, pass to the universal cover.)

7.8.8 Why Morse theory survives in infinite dimensions

We have used the Morse inequalities on an infinite-dimensional space without comment. Two things could go wrong.

Infinite index. A critical point with infinitely many negative directions would attach an "infinite-dimensional cell", and infinite-dimensional discs can be contracted onto their boundary spheres, so nothing would be detected. The second-variation formula rules this out: the Hessian of E is a Schrödinger operator with finitely many negative eigenvalues (Section 7.8.4). The energy is "bounded below in all but finitely many directions". The action of a *relativistic* string, or the Dirac action, is not of this form; it is unbounded in both directions, and for those problems a different construction (Floer's, Chapter 12) is needed.

Critical points escaping to infinity. This can go wrong even in one dimension, once the space is not compact.

Worked example 7.55: A minimum that runs away

Let $f(x) = e^x$ on $M = \mathbb{R}$. Then $f'(x) = e^x \neq 0$ everywhere: no critical points, $M_t(f) = 0$. But $\mathbb{R}$ is connected, $P_t(\mathbb{R}) = 1$, and the weak inequality $m_0 \geq b_0$ fails. What happened is visible in Figure 7.34: f is bounded below by 0 and wants a minimum, but the minimum has escaped to $x = -\infty$. Along the sequence $x_k = -k$,

$$f(x_k) = e^{-k} \text{ is bounded}, \qquad f'(x_k) = e^{-k} \to 0,$$

yet x_k has no convergent subsequence. It is a "critical point at infinity".

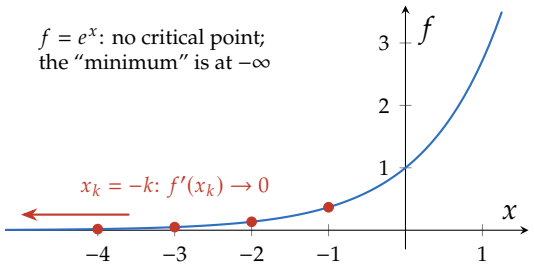


Figure 7.34: The failure mode that the Palais–Smale condition excludes. The function e^x is bounded below and becomes flatter and flatter to the left, but never has a critical point. A sequence along which f stays bounded and $f' \to 0$ (red dots) runs off to infinity instead of converging. On such a space the Morse inequalities fail: $\mathbb{R}$ has $b_0 = 1$ but f has no minimum.

Working definition 7.23 (Palais–Smale condition)

A function f on a (possibly infinite-dimensional, complete [every Cauchy sequence of points converges, so no points are missing]) Riemannian manifold satisfies the **Palais–Smale condition** if every sequence x_k along which $f(x_k)$ is bounded and $|\nabla f(x_k)| \to 0$ has a convergent subsequence. *In words*: approximate critical points at bounded height converge to genuine ones. On a compact finite-dimensional M it holds automatically.

Palais and Smale showed (1964) that with this condition, finite indices and the gradient flow on a complete Hilbert manifold, the whole of finite-dimensional Morse theory goes through: nothing happens between critical values, crossing a critical point of index λ attaches a λ-cell, and the Morse inequalities hold. For the energy one works with paths of finite energy (the Sobolev space H^1 of paths whose velocity is square integrable), which is a complete Hilbert manifold. Then E satisfies Palais–Smale whenever M is compact: a sequence of paths of bounded energy is equicontinuous (by the chain (7.115) below, short time intervals give short arcs), so by Arzelà–Ascoli it has a uniformly convergent subsequence, and $\nabla E \to 0$ upgrades this to convergence in H^1. Physically: a particle of bounded kinetic energy on a compact space cannot run away.

7.8.9 Broken geodesics: a finite-dimensional model (time slicing)

Morse himself avoided infinite-dimensional analysis. His device, which Nash describes for closed loops [Nash p. 206], is to replace paths by *broken geodesics*: chains of short geodesic segments. Any physicist will recognise it as the time-slicing of the path integral, where a path is replaced by its positions $x_1, \dots, x_{k-1}$ at intermediate times.

Bounded energy means short segments. Cut $[0,1]$ into k equal intervals $[t_{i-1}, t_i]$, $t_i = i/k$. Let L_i be the length of the ith piece of a path γ. Then

$$L_i^2 \overset{(1)}{=} \left(\int_{t_{i-1}}^{t_i} |\dot\gamma| \cdot 1 \, \mathrm{d}t \right)^2 \overset{(2)}{\le} (t_i - t_{i-1}) \int_{t_{i-1}}^{t_i} |\dot\gamma|^2 \, \mathrm{d}t \overset{(3)}{\le} \frac{E(\gamma)}{k}. \tag{7.115}$$

(1) Definition of length, with a factor 1 inserted.
(2) Cauchy–Schwarz on the interval $[t_{i-1}, t_i]$, as in (7.98).
(3) $t_i - t_{i-1} = 1/k$, and the integral over one piece is at most the total energy.

Now fix an energy ceiling c. Every compact Riemannian manifold has an $\varepsilon > 0$ (for instance any number below the injectivity radius) such that two points at distance $< \varepsilon$ are joined by a *unique* shortest geodesic, which depends smoothly on its endpoints. Choose $k > c/\varepsilon^2$. Then every path with $E < c$ has all its pieces of length $< \varepsilon$, by (7.115).

Working definition 7.24 (space of broken geodesics)

With k and c as above, a **broken geodesic** is a path that is the unique short geodesic [the one shortest geodesic joining $\gamma(t_{i-1})$ to $\gamma(t_i)$, which exists since their distance is below ε] on each interval $[t_{i-1}, t_i]$. It is determined by its break points $x_i = \gamma(t_i)$, $i = 1, \dots, k-1$ (with $x_0 = p$, $x_k = q$). So the space of broken geodesics with $E < c$ is the finite-dimensional open set

$$B_k(c) = \big\{ (x_1, \dots, x_{k-1}) \in M^{k-1} : \; E < c \big\}, \qquad \dim B_k(c) = (k-1)\, n,$$

on which the energy is the explicit function

$$E_k(x_1,\dots,x_{k-1}) = k\sum_{i=1}^{k} d(x_{i-1},x_i)^2. \tag{7.116}$$

[$d(x,y)$ is the Riemannian distance, the length of the shortest path from x to y].

The formula (7.116) holds because a geodesic segment of length d traversed in time $1/k$ has speed kd, hence energy $\int_0^{1/k}(kd)^2\mathrm{d}t = kd^2$.

Key idea: Broken geodesics carry everything below the ceiling

- Replacing each piece of a path with $E < c$ by the short geodesic with the same endpoints lowers its energy (a geodesic minimises energy among paths with its endpoints and duration). Doing this gradually is a deformation retraction of $\{E < c\} \subset PM(p,q)$ onto $B_k(c)$. So the two have the same topology.
- The critical points of E_k on $B_k(c)$ are exactly the unbroken geodesics from p to q with $E < c$, with the same index and the same nullity as in the full path space.

Thus Morse theory on $PM(p,q)$ below any energy is honest finite-dimensional Morse theory on $B_k(c)$. Letting $c \to \infty$ (and $k \to \infty$ with it) recovers the whole path space.

Why is a critical point of E_k unbroken? Differentiating (7.116) in x_i only affects the two pieces meeting at x_i. The gradient of $d(x_{i-1},x)^2$ at $x = x_i$ is $2d$ times the unit vector pointing away from x_{i-1} along the segment, i.e. $2/k$ times the arriving velocity; that of $d(x,x_{i+1})^2$ is minus $2/k$ times the departing velocity. So

$$\frac{\partial E_k}{\partial x_i} = 2\big(\dot\gamma(t_i^-) - \dot\gamma(t_i^+)\big), \tag{7.117}$$

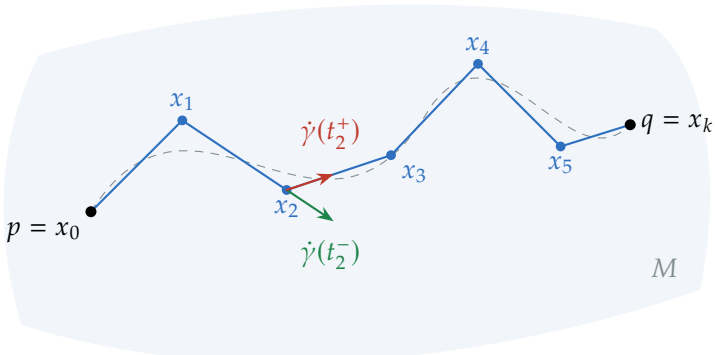
and a critical point has no kink at any x_i: the incoming and outgoing velocities agree, and the broken geodesic is a smooth geodesic. (This is also the first-variation formula (7.96) for a piecewise geodesic: the only contributions are the jumps of velocity at the corners.) Figure 7.35 shows the picture.

Worked example 7.56: Time slicing in flat space

Take $M = \mathbb{R}^n$, where $d(x,y) = |x-y|$ and every k is allowed. Put $x_i = p + \frac{i}{k}(q-p) + \eta_i$ with $\eta_0 = \eta_k = 0$. The same steps as in Example 7.48 give

$$\begin{aligned} E_k &\overset{(1)}{=} k\sum_{i=1}^{k}\Big|\tfrac{1}{k}(q-p)+\eta_i-\eta_{i-1}\Big|^2 \\ &\overset{(2)}{=} |q-p|^2 + 2(q-p)\cdot\sum_{i=1}^{k}(\eta_i-\eta_{i-1}) + k\sum_{i=1}^{k}|\eta_i-\eta_{i-1}|^2 \\ &\overset{(3)}{=} |q-p|^2 + k\sum_{i,j=1}^{k-1} T_{ij}\,\eta_i\cdot\eta_j, \qquad T = \begin{pmatrix} 2 & -1 & & \\ -1 & 2 & -1 & \\ & \ddots & \ddots & \ddots \\ & & -1 & 2 \end{pmatrix}. \end{aligned}$$

(1) (7.116) with $x_i - x_{i-1} = \frac{1}{k}(q-p) + \eta_i - \eta_{i-1}$.



each piece: the unique short geodesic, length $< \varepsilon$, duration $1/k$;
$\partial E_k/\partial x_i = 2(\dot\gamma(t_i^-) - \dot\gamma(t_i^+))$: kinks cost energy

Figure 7.35: A broken geodesic with $k = 6$ pieces. It is fixed by the $k-1$ break points $x_1, \dots, x_5$, so the space of broken geodesics is finite dimensional. Any path of energy $< c$ (dashed) can be pulled tight piece by piece onto a broken geodesic without raising its energy. The gradient of E_k at a break point is the jump in velocity there (green versus red arrows at x_2), so the critical points are the broken geodesics with no kinks: the true geodesics.

(2) Expand the square; the k copies of $k \cdot \frac{1}{k^2}|q-p|^2$ add up to $|q-p|^2$.

(3) The middle sum telescopes to $\eta_k - \eta_0 = 0$; expanding $\sum_i |\eta_i - \eta_{i-1}|^2$ gives $2|\eta_i|^2$ on the diagonal and $-\eta_i \cdot \eta_{i\pm1}$ off it (the discrete Laplacian).

The Hessian is $2kT \otimes \mathbb{1}_n$. The eigenvectors of T are the discrete sine modes $(\eta_i) = (\sin \frac{m\pi i}{k})$, with eigenvalues $4\sin^2 \frac{m\pi}{2k}$, $m = 1, \dots, k-1$ (substitute and use $\sin(a+b) + \sin(a-b) = 2\sin a \cos b$). All are positive, so the index is 0, as in the continuum. For fixed m and large k, $2k \cdot 4\sin^2 \frac{m\pi}{2k} \approx \frac{2(m\pi)^2}{k}$; since $\sum_i |\eta_i|^2 \approx \frac{k}{2}|a_m|^2$ for $\eta_i = a_m \sin \frac{m\pi i}{k}$, the quadratic energy of one mode tends to $\frac{1}{2} \cdot \frac{2(m\pi)^2}{k} \cdot \frac{k}{2}|a_m|^2 = \frac{(m\pi)^2}{2}|a_m|^2$, exactly the continuum value of Example 7.48. The time-sliced action converges mode by mode to the continuum action, which is the statement physicists use when they define the path integral by slicing.

Worked example 7.57: Two-piece broken geodesics on S^2

On the unit S^2 with $d(p,q) = \theta \in (0,\pi)$, take $k = 2$: a single break point x, and

$$E_2(x) = 2\big(d(p,x)^2 + d(x,q)^2\big), \qquad x \in S^2 \setminus \{-p, -q\}$$

(each piece must be shorter than π to be the unique shortest arc). We find its critical points and their Hessians directly, and recover γ_0 and γ_1 with indices 0 and 1.

Coordinates near a point m of the great circle C through p and q. Let s be the displacement along C and y the (small) displacement perpendicular to C. For a point a_0 on C at distance a from m, the spherical theorem of Pythagoras (right angle at the foot of the perpendicular) gives, to second order,

$$\cos d(a_0,x) \overset{(1)}{=} \cos(a \mp s)\cos y \quad\Longrightarrow\quad d(a_0,x) \overset{(2)}{=} a \mp s + \tfrac{1}{2}y^2 \cot a + O(3).$$

(1) Spherical Pythagoras $\cos(\text{hypotenuse}) = \cos(\text{leg}_1)\cos(\text{leg}_2)$, with legs $a \mp s$ along C and y across it (the sign depends on which side of m the point a_0 lies).

(2) Write $d = (a \mp s) + \delta$. Then $\cos(a \mp s) - \delta \sin(a \mp s) \approx \cos(a \mp s)(1 - \frac{1}{2}y^2)$, so $\delta = \frac{1}{2}y^2 \cot(a \mp s) = \frac{1}{2}y^2 \cot a + O(3)$.

Short arc. Let m be the midpoint of the short arc from p to q, so p and q lie on C at distance $a = \theta/2$ from m, on opposite sides. With $Y = \frac{1}{2}y^2 \cot a$,

$$E_2 \overset{(1)}{=} 2\big[(a - s + Y)^2 + (a + s + Y)^2\big] \overset{(2)}{=} 4a^2 + 4s^2 + 8aY + O(3)$$
$$\overset{(3)}{=} \theta^2 + 4s^2 + 2\theta \cot \tfrac{\theta}{2}\; y^2 + O(3).$$

(1) The expansion above for $d(p, x)$ and $d(x, q)$.
(2) Expand, keeping terms of order two (Y is already second order).
(3) $a = \theta/2$ and $8aY = 4a \cot a\; y^2$.

Both coefficients are positive since $\cot \frac{\theta}{2} > 0$: a minimum, index 0, with $E_2 = \theta^2 = L(\gamma_0)^2$.
Long arc. Let m be the midpoint of the long arc (through $-p$ and $-q$). Now $-p$ and $-q$ lie at distance $a = \theta/2$ from m, and $d(p, x) = \pi - d(-p, x)$, $d(q, x) = \pi - d(-q, x)$. So

$$E_2 \overset{(1)}{=} 2\big[(\pi - a - s - Y)^2 + (\pi - a + s - Y)^2\big]$$
$$\overset{(2)}{=} 4(\pi - a)^2 + 4s^2 - 8(\pi - a)Y + O(3) \overset{(3)}{=} (2\pi - \theta)^2 + 4s^2 - 2(2\pi - \theta) \cot \tfrac{\theta}{2}\; y^2 + O(3).$$

(1) Antipodal distances on the unit sphere, then the expansion for $d(-p, x)$ and $d(-q, x)$.
(2) Expand to second order.
(3) $a = \theta/2$ and $8(\pi - a)Y = 4(\pi - a) \cot a\; y^2$.

The y^2 coefficient is negative: one downhill direction, index $1 = n - 1$, with $E_2 = (2\pi - \theta)^2 = L(\gamma_1)^2$. Moving the break point off the great circle shortens both pieces. The reason is that each piece is longer than a quarter circle, and beyond a quarter circle the circles of constant distance from p are shrinking again (they close up at $-p$), so a sideways step moves x to a smaller circle, i.e. closer to p; this is the $-\cot$ sign above. On S^n the same computation has $n - 1$ perpendicular directions y, and gives index $n - 1$.
Longer geodesics are invisible to E_2: γ_2 has length $> 2\pi$ and cannot be built from two pieces shorter than π. To see γ_j one needs $k > L_j/\pi$ break points, which is why the ceiling c and k must grow together.

Try it: An equilateral triangle on a great circle

Nash's version for closed loops uses *geodesic polygons* with n vertices $p_1, \dots, p_n$, with energy $E_n = n \sum_{i=1}^{n} d(p_i, p_{i+1})^2$ ($p_{n+1} = p_1$). Show that three equally spaced points on a great circle of S^2 form a critical point of E_3, with the energy of a once-traversed great circle. Why does $n = 2$ not work?
Answer: each side has length $2\pi/3 < \pi$, so it is the unique short arc; at each vertex the incoming and outgoing directions are both tangent to the great circle and equal, so by (7.117) the gradient vanishes. $E_3 = 3 \cdot 3 \cdot (2\pi/3)^2 = 4\pi^2 = (2\pi)^2$. With $n = 2$ the two vertices would have to be antipodal, at distance π, where the short arc is not unique and d^2 is not smooth.

7.8.10 Closed geodesics and the free loop space

A **closed geodesic** is a geodesic that comes back to its starting point with its starting velocity, so that it can go round again: a periodic orbit of the free particle. To find them, Nash replaces paths with fixed ends by all loops [Nash p. 206, (7.59)]:

$$E\colon LM := \operatorname{Map}(S^1, M) \longrightarrow \mathbb{R}, \qquad E(\gamma) = \int_0^1 |\dot\gamma(t)|^2 \, \mathrm{d}t, \tag{7.118}$$

where a point of the **free loop space** $LM = \mathrm{Map}(S^1, M)$ is a map of the circle $S^1 = [0,1]/(0 \sim 1)$ into M, with no base point imposed.

First variation for loops. The computation (7.94) goes through with one change: $V(0) = V(1)$ need not vanish, only agree. The boundary term from the integration by parts is $2\big[\langle V, \dot\gamma\rangle\big]_0^1 = 2\langle V(0), \dot\gamma(1) - \dot\gamma(0)\rangle$. So the critical points of E on LM satisfy the geodesic equation *and* $\dot\gamma(1) = \dot\gamma(0)$: they are exactly the closed geodesics, together with the constant loops (geodesics of speed zero), which form the absolute minimum $E = 0$.

The relation to the based loop space. Evaluating a loop at $t = 0$ gives a map $LM \to M$, $\gamma \mapsto \gamma(0)$. The loops with $\gamma(0) = p$ are the based loops, so the fibre over p is ΩM:

$$\Omega M \longrightarrow LM \xrightarrow{\gamma \mapsto \gamma(0)} M. \tag{7.119}$$

The constant loops give a section $M \to LM$. This fibration is what Nash means by "the topology of $\mathrm{Map}(S^1, M)$ is closely related to that of ΩM" [Nash p. 206]. It implies the inequality $P_t(LM) \le P_t(M)\,P_t(\Omega M)$ coefficientwise (with field coefficients and M simply connected), with equality in favourable cases. The inequality is the statement that the cohomology of the total space is at most that of "base times fibre"; equality means that, as for a product, every class of the fibre extends over the total space (the Leray–Hirsch situation of Section 7.6.5).

The circle acts, so critical points come in circles. The group U(1) of rotations of the domain circle acts on LM,

$$(e^{2\pi i s} \cdot \gamma)(t) = \gamma(t + s), \tag{7.120}$$

and E is invariant: shifting the time origin does not change the kinetic energy. So E is a U(1)-invariant function, and the equivariant theory of Section 7.7 applies, as Nash remarks. The consequence for the ordinary theory is immediate: a closed geodesic γ is never an isolated critical point, because every rotated copy $\gamma(\cdot + s)$ is also critical, with the same energy. The critical sets are at least circles, and E is at best a Morse–Bott function (Section 7.5). The stabiliser of a closed geodesic traversed k times ($\gamma(t) = \gamma_1(kt)$) is the cyclic group $\mathbb{Z}_k$ of rotations by multiples of $1/k$; its orbit is $S^1/\mathbb{Z}_k$, and for real coefficients the equivariant contribution of such an orbit is that of a point (Section 7.7).

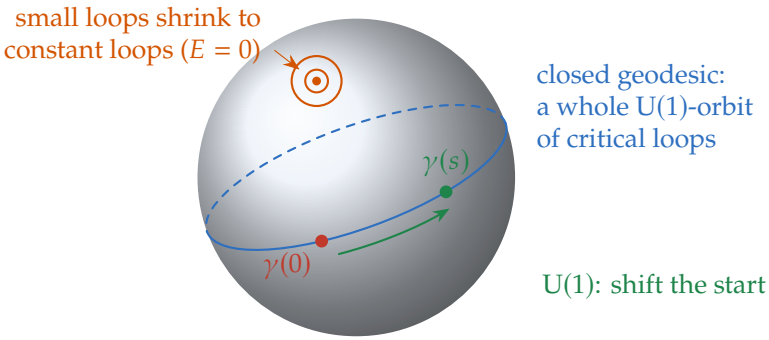


Figure 7.36: Critical points of the energy on the free loop space LS^2. A great circle (blue) is a closed geodesic; but the loop is a parametrised map, and moving the starting point from $\gamma(0)$ (red) to $\gamma(s)$ (green) gives a different point of LS^2 with the same energy. So each closed geodesic is a circle of critical points. Tilting the great circle gives more: the whole family is the unit tangent bundle T^1S^2. At the bottom of the landscape lie the constant loops (orange), a copy of S^2 at $E = 0$.

Worked example 7.58: The flat torus: critical tori

On $T^n = \mathbb{R}^n/2\pi\mathbb{Z}^n$ with the flat metric, a closed loop lifts to a path from $\tilde{x}$ to $\tilde{x} + 2\pi m$, $m \in \mathbb{Z}^n$. The closed geodesics in the class m are the straight loops

$$\gamma_{x,m}(t) = x + 2\pi m\, t, \qquad x \in T^n,$$

one for each starting point x: a critical manifold $C_m \cong T^n$, at energy $E = 4\pi^2 |m|^2$ (for $m = 0$ these are the constant loops). Each has index 0 by Example 7.52. Is C_m non-degenerate in the Morse–Bott sense? The Jacobi operator is $-\mathrm{d}^2/\mathrm{d}t^2$ on periodic vector fields; its kernel is the constant fields $V \equiv v \in \mathbb{R}^n$, which is n-dimensional, exactly the tangent space of C_m (moving x). So yes. The Morse–Bott series is

$$M_t(E) \overset{(1)}{=} \sum_{m\in\mathbb{Z}^n} t^0 P_t(T^n) \overset{(2)}{=} \sum_{m\in\mathbb{Z}^n} (1+t)^n,$$

(1) (7.54) with one critical torus of index 0 in each class m.
(2) $P_t(T^n) = (1+t)^n$ by Künneth.

which is $P_t(LT^n)$: the component of LT^n of class m is homotopy equivalent to T^n (slide the loop to the straight one). The energy is perfect.

Worked example 7.59: The round sphere: iterated great circles

On the unit S^n the closed geodesics are the great circles traversed k times, $k = 1, 2, \ldots$:

$$\gamma(t) = \cos(2\pi k t)\, x + \sin(2\pi k t)\, v, \qquad x, v \in S^n \subset \mathbb{R}^{n+1}, \quad x \perp v,$$

of length $2\pi k$ and energy $(2\pi k)^2$. The pair (x, v), a point and a unit tangent vector there, runs over the unit tangent bundle, so each critical manifold is

$$C_k \cong T^1 S^n, \qquad \dim C_k = n + (n-1) = 2n - 1.$$

Index and nullity. Use arc length $s \in [0, 2\pi k]$ with periodic boundary conditions and the parallel frame of Example 7.51. The Jacobi operator is $-\mathrm{d}^2/\mathrm{d}s^2$ on the tangential component and $-\mathrm{d}^2/\mathrm{d}s^2 - 1$ on each of the $n-1$ normal components. Periodic eigenfunctions are 1, $\cos\frac{ms}{k}$, $\sin\frac{ms}{k}$ ($m \geq 1$), with eigenvalues $(m/k)^2$ and $(m/k)^2 - 1$ respectively. Counting:

$$\lambda_{C_k} \overset{(1)}{=} (n-1)\Big[1 + 2\,\#\{m :\ 1 \leq m \leq k-1\}\Big] \overset{(2)}{=} (2k-1)(n-1),$$

$$\text{nullity} \overset{(3)}{=} 1 + (n-1)\cdot 2 \overset{(4)}{=} 2n - 1 = \dim C_k.$$

(1) Negative normal eigenvalues: $m = 0$ (one function) and $1 \leq m \leq k-1$ (two functions each), since $(m/k)^2 < 1$ exactly for $m < k$. Tangential eigenvalues are never negative.
(2) $1 + 2(k-1) = 2k - 1$.
(3) Zero eigenvalues: the tangential constant (rotating the start point) and the two normal modes with $m = k$, in each of $n-1$ normal directions (tilting the great circle).
(4) So the kernel of the Hessian is exactly the tangent space of C_k: the critical manifold is non-degenerate.

The $m = 0$ normal mode is the familiar one: push the whole great circle towards a latitude circle, which is shorter. The constant loops, $C_0 \cong S^n$, have index 0.
With $\mathbb{Z}_2$ coefficients (so that no orientation of the negative bundles is needed) and $n = 2$,

where $T^1S^2 \cong \mathrm{SO}(3) \cong \mathbb{RP}^3$ has $P_t(\mathbb{RP}^3;\mathbb{Z}_2) = 1 + t + t^2 + t^3$, the Morse–Bott series is

$$M_t(E) \overset{(1)}{=} (1+t^2) + \sum_{k\geq 1} t^{2k-1}(1+t+t^2+t^3)$$
$$\overset{(2)}{=} (1+t^2) + (1+t)(1+t^2)\,\frac{t}{1-t^2} \overset{(3)}{=} (1+t^2)\Big(1+\frac{t}{1-t}\Big) = \frac{1+t^2}{1-t}.$$

(1) Constant loops ($P_t(S^2)$, index 0) plus C_k with index $2k-1$.
(2) $1+t+t^2+t^3 = (1+t)(1+t^2)$, and $\sum_{k\geq 1} t^{2k-1} = t/(1-t^2)$.
(3) $1-t^2 = (1-t)(1+t)$.

Compare with the fibration (7.119): $P_t(S^2)P_t(\Omega S^2) = (1+t^2)/(1-t)$ by Example 7.54. The two agree; this equality is Ziller's theorem that for the round sphere (indeed for every compact symmetric space) E is a perfect Morse–Bott function with $\mathbb{Z}_2$ coefficients.

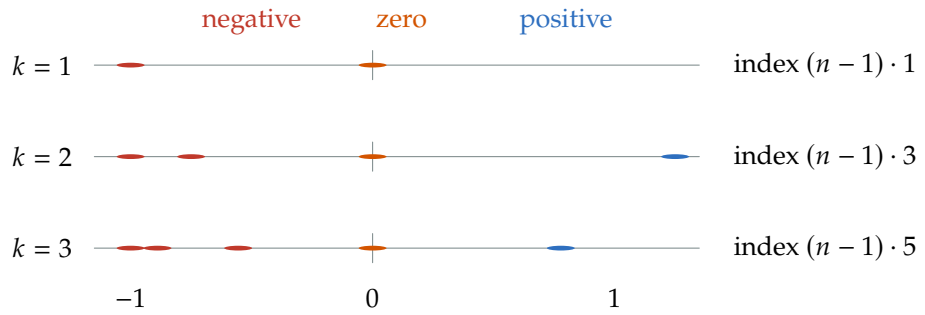


Figure 7.37: The spectrum of the normal Jacobi operator $-\mathrm{d}^2/\mathrm{d}s^2 - 1$ for a great circle traversed k times, one dot per eigenvalue $(m/k)^2 - 1$, $m = 0, 1, 2, \ldots$ (each $m \geq 1$ counts twice: cos and sin). Going round more times packs the eigenvalues more densely, so more of them fall below zero (red): $1, 3, 5, \ldots$ per normal direction. The orange zero modes, $m = k$, tilt the circle, and are tangent to the critical manifold.

How many closed geodesics are there? The trouble with counting closed geodesics by Morse theory is that each one produces infinitely many critical manifolds: all its iterates γ^k. A single closed geodesic, iterated, could in principle carry all the Betti numbers of LM. To conclude that there are infinitely many *geometrically distinct* closed geodesics one must control how the index grows under iteration. The known results:

- *Lusternik–Fet* (1951): every compact Riemannian manifold has at least one non-constant closed geodesic. (Take a non-trivial homotopy class of maps $S^j \to M$, view it as a family of loops, and minimise the maximum energy over the family: a minimax, as in Section 7.4.5.)
- *Lusternik–Schnirelmann*: every metric on S^2 has at least three simple closed geodesics (the three principal "equators" of an ellipsoid are the model).
- *Gromoll–Meyer* (1969): if M is simply connected and the Betti numbers of LM (rational coefficients) are unbounded, there are infinitely many geometrically distinct closed geodesics. The hypothesis fails for spheres, whose loop spaces have bounded Betti numbers (as we just saw).
- *Bangert and Franks* (1992–93): every metric on S^2 has infinitely many closed geodesics. For S^n with $n \geq 3$ and a general metric the question is harder; Nash's sentence about "the existence of an infinite number of closed geodesics on S^k" refers to Klingenberg's programme, and it is best read as a statement about the method.

Physical picture: Periodic orbits in physics

Closed geodesics are periodic orbits of the free particle on M. They are the input of the Gutzwiller trace formula, which expresses the density of quantum energy levels as a sum over classical periodic orbits, each weighted by a phase built from its Morse (Maslov) index. The index growth $\lambda(\gamma^k) \approx k \cdot (\text{const})$ we saw for great circles is the "mean index" that enters there. The U(1) action by time translation is the reason each orbit appears as a whole circle of critical points, and in the trace formula it produces the factor of the primitive period.
A second connection is deeper. Atiyah observed that for the supersymmetric particle on M, the path integral over the free loop space is formally an equivariantly closed integral for the rotation (7.120). Its fixed points are the constant loops, i.e. M itself, and applying the Duistermaat–Heckman localisation of Section 7.7.6 formally to LM gives an integral over M of the $\hat{A}$ genus: the index of the Dirac operator, Chapter 4. The circle action on $\mathrm{Map}(S^1, M)$ that Nash mentions in passing is the reason the index theorem has a one-line path-integral proof.

7.8.11 Loop groups and Bott periodicity

Nash closes his discussion of geodesics with one sentence [Nash p. 207]: replace the sphere S^k by a Lie group G and use Morse theory on ΩG; "this strategy was very successfully employed by Bott in the proof of his periodicity theorems". This subsection unpacks that sentence. It is worth the effort, because the answer is a table that every gauge theorist uses.

The question. Physicists meet the homotopy groups of the unitary groups all the time.

- $\pi_1(\mathrm{U}(1)) = \mathbb{Z}$: flux quantisation, vortices, winding of a phase.
- $\pi_3(\mathrm{SU}(N)) = \mathbb{Z}$ for every $N \geq 2$: the instanton number of SU(N) Yang–Mills theory (Chapter 8), and the baryon number of Skyrmions.
- $\pi_4(\mathrm{SU}(2)) = \mathbb{Z}_2$: Witten's global SU(2) anomaly (Chapter 10).
- $\pi_5(\mathrm{SU}(3)) = \mathbb{Z}$: the quantisation of the Wess–Zumino–Witten term in the chiral Lagrangian of QCD with three flavours.

Two things stand out. Some of these groups do not depend on N once N is large enough: every SU(N) with $N \geq 2$ has the same π_3. And the groups look simple: $\mathbb{Z}$ or 0, apart from a few small-N oddities like $\pi_4(\mathrm{SU}(2))$. Bott's theorem explains both, and it does so by Morse theory on a loop space.

Step 0: stability

Working definition 7.25 (stable homotopy groups of the unitary group)

The group $\mathrm{U}(n+1)$ acts on the unit sphere $S^{2n+1} \subset \mathbb{C}^{n+1}$ by matrix multiplication, transitively [every unit vector is ge_{n+1} for some $g \in \mathrm{U}(n+1)$]. The matrices that fix the last basis vector e_{n+1} form a copy of $\mathrm{U}(n)$ (upper-left block). Hence

$$\mathrm{U}(n) \longrightarrow \mathrm{U}(n+1) \xrightarrow{\;g \mapsto ge_{n+1}\;} S^{2n+1} \tag{7.121}$$

is a fibre bundle: the last column of a unitary matrix is a unit vector, and the other columns can be anything unitary orthogonal to it. For $j \leq 2n-1$ the inclusion $\mathrm{U}(n) \hookrightarrow$

$\mathrm{U}(n+1)$ induces an isomorphism on π_j [maps $S^j \to \mathrm{U}(n)$, up to continuous deformation, correspond one to one with maps $S^j \to \mathrm{U}(n+1)$]. So for fixed j the groups $\pi_j(\mathrm{U}(n))$ stop changing once $2n > j$. The common value is the **stable** homotopy group, written $\pi_j(\mathrm{U})$, where $\mathrm{U} = \bigcup_n \mathrm{U}(n)$ is the infinite unitary group.

Why stability holds. Every fibre bundle $F \to E \to B$ has a long exact sequence of homotopy groups, $\cdots \to \pi_{j+1}(B) \to \pi_j(F) \to \pi_j(E) \to \pi_j(B) \to \dots$ (a sphere in E projects to a sphere in B; a sphere in B lifts to a disc in E whose boundary lies in one fibre). Apply it to (7.121):

$$\underbrace{\pi_{j+1}(S^{2n+1})}_{0} \longrightarrow \pi_j(\mathrm{U}(n)) \longrightarrow \pi_j(\mathrm{U}(n+1)) \longrightarrow \underbrace{\pi_j(S^{2n+1})}_{0} \qquad (j+1 \le 2n).$$

The outer groups vanish because a sphere S^{2n+1} has no homotopy below its dimension: a map $S^i \to S^{2n+1}$ with $i < 2n+1$ misses a point (after a small smoothing) and can then be contracted through the complement, which is a copy of $\mathbb{R}^{2n+1}$. Exactness then forces the middle arrow to be an isomorphism. With $\mathrm{U}(1) = S^1$, $\mathrm{U}(2) \simeq S^1 \times S^3$ (as spaces, $\mathrm{U}(n) = \mathrm{SU}(n) \times \mathrm{U}(1)$ via $g \mapsto (g\,\mathrm{diag}(\det g^{-1}, 1, \dots), \det g)$) and a few known homotopy groups of spheres, the pattern is visible in a table.

	π_1	π_2	π_3	π_4	π_5	π_6
$\mathrm{U}(1)$	$\mathbb{Z}$	0	0	0	0	0
$\mathrm{U}(2)$	$\mathbb{Z}$	0	$\mathbb{Z}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_{12}$
$\mathrm{U}(3)$	$\mathbb{Z}$	0	$\mathbb{Z}$	0	$\mathbb{Z}$	$\mathbb{Z}_6$
$\mathrm{U}(4)$	$\mathbb{Z}$	0	$\mathbb{Z}$	0	$\mathbb{Z}$	0
U (stable)	$\mathbb{Z}$	0	$\mathbb{Z}$	0	$\mathbb{Z}$	0

Table 7.1: Homotopy groups of the unitary groups. An entry $\pi_j(\mathrm{U}(n))$ is in the stable range when $j \le 2n-1$; the red entries are outside it and are "accidents" of small n (for example $\pi_4(\mathrm{U}(2)) = \pi_4(S^3) = \mathbb{Z}_2$). Read the columns downwards: each one settles to the bottom row. Read the bottom row across: it alternates $\mathbb{Z}, 0, \mathbb{Z}, 0, \dots$. Bott's theorem is the statement that this alternation goes on for ever.

Key idea: Bott periodicity, complex case (Nash p. 207)

$$\pi_j(\mathrm{U}) \;\cong\; \pi_{j+2}(\mathrm{U}) \quad \text{for all } j \ge 0, \qquad \text{so} \qquad \pi_j(\mathrm{U}) = \begin{cases} 0, & j \text{ even}, \\ \mathbb{Z}, & j \text{ odd}. \end{cases}$$

Equivalently, in the stable range $\pi_j(\mathrm{U}(n)) = \pi_j(\mathrm{SU}(n))$ for $j \ge 2$ is $\mathbb{Z}$ in odd degrees and 0 in even degrees.

The rest of this subsection is Bott's proof. The plan is shown in Figure 7.38. The only new ingredients are the geometry of a Lie group with an invariant metric (Step 1) and one index computation (Step 3). Everything else is the Morse theory we already have.

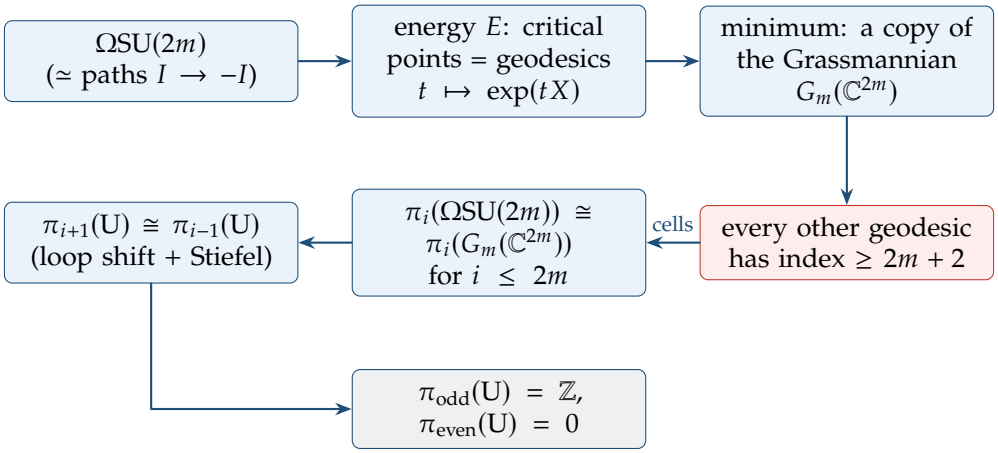


Figure 7.38: The logic of Bott's proof. Morse theory on the loop space of $\mathrm{SU}(2m)$ says that, up to dimension $2m$, the loop space looks like its own minimum set, a finite-dimensional Grassmannian. Two exact sequences then turn this into a statement comparing homotopy groups of the unitary group two degrees apart.

Step 1: geodesics in a Lie group

Working definition 7.26 (bi-invariant metric on $\mathrm{SU}(n)$)

The Lie algebra $\mathfrak{su}(n)$ [the tangent space at the identity I, the velocities of curves in $\mathrm{SU}(n)$ through I] is the space of traceless anti-Hermitian $n \times n$ matrices. On it put

$$\langle X, Y\rangle = -\operatorname{tr}(XY), \qquad |X|^2 = -\operatorname{tr}(X^2) = \sum_{j,l} |X_{jl}|^2 \geq 0. \tag{7.122}$$

Transport it to every tangent space $T_g\mathrm{SU}(n) = g\,\mathfrak{su}(n)$ by left multiplication [declare $\langle gX, gY\rangle := \langle X, Y\rangle$ at the point g]. Because $\operatorname{tr}\big((hXh^{-1})(hYh^{-1})\big) = \operatorname{tr}(XY)$, the result is also invariant under right multiplication: it is a **bi-invariant** metric. Infinitesimally, conjugation invariance says that $\mathrm{ad}_Z = [Z, \cdot\,]$ is skew [its matrix in an orthonormal basis of $\mathfrak{su}(n)$ is antisymmetric]:

$$\langle [Z, X], Y\rangle = -\langle X, [Z, Y]\rangle, \qquad \text{equivalently} \qquad \langle [X, Y], Z\rangle = \langle X, [Y, Z]\rangle. \tag{7.123}$$

Check (7.123) directly: $-\operatorname{tr}([X,Y]Z) = -\operatorname{tr}(XYZ) + \operatorname{tr}(YXZ) = -\operatorname{tr}(XYZ) + \operatorname{tr}(XZY) = -\operatorname{tr}(X[Y,Z])$, using the cyclicity of the trace in the middle step.

The Levi-Civita connection. For left-invariant vector fields X, Y, Z (each determined by its value in $\mathfrak{su}(n)$) all inner products are constant. The Koszul formula for the Levi-Civita connection, $2\langle \nabla_X Y, Z\rangle = X\langle Y,Z\rangle + Y\langle Z,X\rangle - Z\langle X,Y\rangle + \langle [X,Y],Z\rangle - \langle [Y,Z],X\rangle + \langle [Z,X],Y\rangle$, then gives

$$\begin{aligned} 2\langle \nabla_X Y, Z\rangle &\overset{(1)}{=} \langle [X,Y],Z\rangle - \langle [Y,Z],X\rangle + \langle [Z,X],Y\rangle \\ &\overset{(2)}{=} \langle [X,Y],Z\rangle - \langle [X,Y],Z\rangle + \langle [X,Y],Z\rangle \\ &= \langle [X,Y],Z\rangle, \qquad \text{so} \qquad \nabla_X Y = \tfrac{1}{2}[X,Y]. \end{aligned}$$

(1) The three derivative terms vanish because the inner products of left-invariant fields are constant functions on the group.

(2) Cyclic moves with (7.123): $\langle [Y,Z],X\rangle = \langle Y,[Z,X]\rangle = \langle [Z,X],Y\rangle$, and $\langle [Z,X],Y\rangle = \langle Z,[X,Y]\rangle = \langle [X,Y],Z\rangle$.

In particular $\nabla_X X = \tfrac{1}{2}[X,X] = 0$. The integral curves of a left-invariant field are $t \mapsto g\exp(tX)$, so

Key idea: geodesics of a bi-invariant metric

The geodesics through the identity are exactly the one-parameter subgroups $\gamma(t) = \exp(tX)$, $X \in \mathfrak{su}(n)$. The geodesic with $\gamma(0) = I$ and $\gamma(1) = g$ is $\exp(tX)$ with $\exp(X) = g$; it has constant speed $|X|$ and energy $E(\gamma) = \int_0^1 |X|^2\,\mathrm{d}t = |X|^2$ (Nash's normalisation [Nash (7.54)]).

A physicist knows this as the statement that a free symmetric top has constant angular velocity in the body frame: the geodesic equation of the bi-invariant metric is Euler's equation with all moments of inertia equal.

The curvature. For left-invariant fields, $R(X,Y)Z = \nabla_X\nabla_Y Z - \nabla_Y\nabla_X Z - \nabla_{[X,Y]}Z$ (the convention of Section 7.8.4, for which the unit sphere has $R(X,Y)Z = \langle Y,Z\rangle X - \langle X,Z\rangle Y$). Then

$$R(X,Y)Z \overset{(1)}{=} \tfrac{1}{4}[X,[Y,Z]] - \tfrac{1}{4}[Y,[X,Z]] - \tfrac{1}{2}[[X,Y],Z]$$
$$\overset{(2)}{=} \tfrac{1}{4}[[X,Y],Z] - \tfrac{1}{2}[[X,Y],Z] = -\tfrac{1}{4}[[X,Y],Z].$$

(1) Apply $\nabla_X Y = \tfrac{1}{2}[X,Y]$ twice in each of the first two terms, and once in the third.
(2) The Jacobi identity: $[X,[Y,Z]] - [Y,[X,Z]] = [[X,Y],Z]$.

The operator that enters the Jacobi equation along the geodesic $\gamma(t) = \exp(tX)$ is $K_X(V) = R(V,X)X$ (with $\gamma' = X$ in the left-invariant frame, as in the operator $\mathcal{J}$ of (7.101)):

$$K_X(V) = R(V,X)X \overset{(1)}{=} -\tfrac{1}{4}[[V,X],X] \overset{(2)}{=} -\tfrac{1}{4}\,\mathrm{ad}_X^2(V). \tag{7.124}$$

(1) The curvature formula just derived, with $(X,Y,Z) \to (V,X,X)$.
(2) $[[V,X],X] = [X,[X,V]] = \mathrm{ad}_X(\mathrm{ad}_X V)$, using antisymmetry twice.

Since ad_X is skew (7.123), its eigenvalues are imaginary, $\pm i\theta$, and ad_X^2 has eigenvalues $-\theta^2 \le 0$. So K_X is positive semi-definite with eigenvalues $\theta^2/4$: a compact group with a bi-invariant metric has non-negative sectional curvature, and geodesics refocus, as on a sphere.

Conjugate points. A bi-invariant metric makes G a symmetric space, for which the curvature tensor is parallel. So the eigenvectors of K_X can be chosen parallel along γ, and the Jacobi equation $V'' + K_X V = 0$ (7.102) splits into scalar equations

$$v'' + \frac{\theta^2}{4}v = 0, \qquad v(0) = 0 \quad\Longrightarrow\quad v(t) = \sin\frac{\theta t}{2}, \tag{7.125}$$

one for each eigenvector of ad_X^2 with eigenvalue $-\theta^2 < 0$. This solution vanishes again at $t = 2\pi k/\theta$, $k = 1,2,\dots$. The directions with $\theta = 0$ (the centraliser of X) give $v = t$, which never vanishes again. By the Morse index theorem (7.103):

Working definition 7.27 (index of a geodesic in a compact Lie group)

For the geodesic $\gamma(t) = \exp(tX)$, $0 \le t \le 1$:

1. find the eigenvalues $\pm i\theta$ of ad_X on the Lie algebra, with their real multiplicities [the

multiplicity of θ is the real dimension of the eigenspace of ad_X^2 with eigenvalue $-\theta^2$];

2. each $\theta > 0$ contributes $\#\{k \geq 1 : \ 2\pi k/\theta < 1\}$ conjugate points inside $(0, 1)$, each counted with the multiplicity of θ;

3. the index $\lambda(\gamma)$ is the total; the nullity [the number of independent Jacobi fields vanishing at both $t = 0$ and $t = 1$] is the number counted with $2\pi k/\theta = 1$ exactly.

Step 2: a warm-up, $\mathrm{SU}(2) = S^3$

Worked example 7.60: Geodesics from I to $-I$ in $\mathrm{SU}(2)$

A basis of $\mathfrak{su}(2)$ is $i\sigma_1, i\sigma_2, i\sigma_3$, with the Pauli matrices

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \qquad \langle i\sigma_a, i\sigma_b \rangle = -\operatorname{tr}(i\sigma_a\, i\sigma_b) = 2\delta_{ab}.$$

Write $X = i\theta\, \hat{n} \cdot \vec{\sigma}$ with $\hat{n}$ a unit vector and $\theta \geq 0$. Since $(\hat{n} \cdot \vec{\sigma})^2 = I$,

$$\exp(X) \overset{(1)}{=} \sum_{r \geq 0} \frac{(i\theta)^r (\hat{n} \cdot \vec{\sigma})^r}{r!} \overset{(2)}{=} \cos\theta\, I + i \sin\theta\, \hat{n} \cdot \vec{\sigma}.$$

(1) The exponential series.
(2) Even powers of $\hat{n} \cdot \vec{\sigma}$ are I and odd powers are $\hat{n} \cdot \vec{\sigma}$; the even and odd terms sum to $\cos\theta$ and $i \sin\theta$.

So $\exp(X) = -I$ exactly when $\theta = (2l+1)\pi$, $l = 0, 1, 2, \ldots$, with $\hat{n}$ arbitrary. The geodesics from I to $-I$ therefore come in families

$$N_l = \left\{ t \mapsto \exp\bigl(i(2l+1)\pi t\, \hat{n} \cdot \vec{\sigma}\bigr) \;:\; \hat{n} \in S^2 \right\} \;\cong\; S^2, \qquad E(N_l) = |X|^2 = 2(2l+1)^2\pi^2.$$

Index. Take $\hat{n} = \hat{z}$, so $X = i\theta\sigma_3$. From $[\sigma_3, \sigma_\pm] = \pm 2\sigma_\pm$ (with $\sigma_\pm = \frac{1}{2}(\sigma_1 \pm i\sigma_2)$) we get $\mathrm{ad}_X\, \sigma_\pm = \pm 2i\theta\, \sigma_\pm$ and $\mathrm{ad}_X\, \sigma_3 = 0$. So ad_X has the eigenvalue pair $\pm i \cdot 2\theta$ on the real 2-plane spanned by $i\sigma_1, i\sigma_2$, and 0 on $i\sigma_3$. With $2\theta = 2(2l+1)\pi$ the conjugate points sit at

$$t = \frac{2\pi k}{2(2l+1)\pi} = \frac{k}{2l+1}, \qquad k = 1, 2, \ldots,$$

and those inside $(0, 1)$ are $k = 1, \ldots, 2l$, each of multiplicity 2. Hence

$$\lambda(N_l) = 2 \cdot 2l = 4l, \qquad \text{nullity} = 2\ (k = 2l+1) = \dim N_l.$$

The nullity equals the dimension of the critical family: each N_l is a non-degenerate critical manifold in the sense of Section 7.5.

Morse–Bott series. By (7.54),

$$M_t(E) \overset{(1)}{=} \sum_{l \geq 0} t^{\lambda(N_l)} P_t(N_l) \overset{(2)}{=} (1 + t^2) \sum_{l \geq 0} t^{4l} \overset{(3)}{=} \frac{1 + t^2}{1 - t^4} \overset{(4)}{=} \frac{1}{1 - t^2}.$$

(1) Each critical manifold contributes its Poincaré polynomial shifted by its index.
(2) $\lambda(N_l) = 4l$ and $P_t(S^2) = 1 + t^2$.
(3) Geometric series in t^4.
(4) $1 - t^4 = (1 - t^2)(1 + t^2)$.

This is exactly $P_t(\Omega S^3)$ from (7.112) with $n = 3$. The energy is a perfect Morse–Bott function (all exponents are even, so the lacunary argument applies to the expanded series $1 + t^2 + t^4 + \cdots$).

Try it: Closed loops at the identity

Do the same count for geodesic *loops* at I: $\exp(X) = I$. Show that the critical families are the constant loop ($X = 0$, index 0) and 2-spheres N_l' with $\theta = 2l\pi$, $l \geq 1$, of index $4l - 2$, and that $1 + (1 + t^2) \sum_{l \geq 1} t^{4l-2} = 1/(1 - t^2)$ again.

Answer. For $\theta = 2l\pi$ the conjugate points are at $t = k/(2l)$, $k = 1, \ldots, 2l-1$ inside $(0, 1)$, each of multiplicity 2: index $2(2l-1) = 4l-2$. Then $1 + (1+t^2)(t^2 + t^6 + t^{10} + \cdots) = 1 + t^2 + t^4 + t^6 + \cdots$.

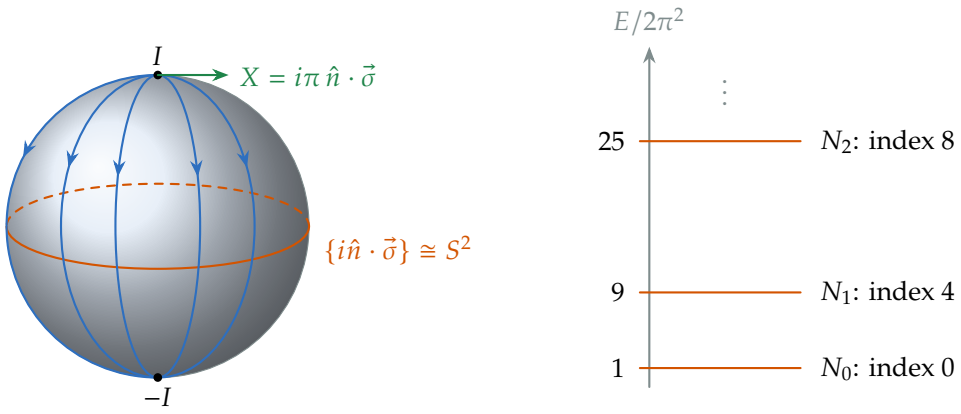


Figure 7.39: Left: SU(2) is a 3-sphere (drawn one dimension down). The shortest geodesics from I to $-I$ are the half great circles $\exp(i\pi t\,\hat{n}\cdot\vec{\sigma})$, one for each direction $\hat{n}$: a whole 2-sphere of minima (orange). Right: the energy levels of the critical families N_l (not to scale). Each is a 2-sphere; the index jumps by 4 from one to the next, because a geodesic that winds $2l+1$ half-turns passes $2l$ conjugate points, each of multiplicity 2.

Step 3: geodesics from I to $-I$ in $\mathrm{SU}(2m)$

Now take $G = \mathrm{SU}(2m)$ and consider the space of paths from I to $-I$. It has the homotopy type of $\Omega\mathrm{SU}(2m)$ by (7.107). A geodesic is $\exp(tX)$ with $\exp(X) = -I$. Diagonalise the anti-Hermitian matrix X by a unitary U:

$$X = U\,\mathrm{diag}\big(i\pi k_1, \dots, i\pi k_{2m}\big)\,U^{-1}, \qquad k_j \text{ odd integers}, \quad \sum_j k_j = 0. \tag{7.126}$$

The k_j must be odd because $\exp(i\pi k_j) = -1$, and they must sum to zero because $\mathrm{tr}\,X = 0$. Their energy is

$$E = |X|^2 = -\,\mathrm{tr}\,X^2 = \pi^2 \sum_{j=1}^{2m} k_j^2. \tag{7.127}$$

The index. The matrix units E_{jl} (a 1 in slot (j,l)) diagonalise the adjoint action of the diagonal matrix $D = \mathrm{diag}(i\pi k_j)$: $[D, E_{jl}] = i\pi(k_j - k_l)E_{jl}$. For each pair $j < l$ the real 2-plane in $\mathfrak{su}(2m)$ spanned by $E_{jl} - E_{lj}$ and $i(E_{jl} + E_{lj})$ carries the eigenvalue pair $\pm i\theta_{jl}$ with $\theta_{jl} = \pi\,|k_j - k_l|$. Using the recipe of Step 1:

$$\lambda(\gamma) \overset{(1)}{=} \sum_{j<l} 2 \cdot \#\Big\{k \geq 1 : \frac{2\pi k}{\pi\,|k_j - k_l|} < 1\Big\} \overset{(2)}{=} \sum_{j<l} 2\Big(\frac{|k_j - k_l|}{2} - 1\Big)_+ \overset{(3)}{=} \sum_{k_j > k_l} (k_j - k_l - 2). \tag{7.128}$$

(1) Each pair $j < l$ is a real 2-dimensional eigenspace; count its conjugate points inside $(0,1)$.
(2) $k_j - k_l$ is even (both are odd), so $2k < |k_j - k_l|$ holds for $k = 1, \dots, |k_j - k_l|/2 - 1$; $(\cdot)_+$ means "or 0 if negative" (for $k_j = k_l$).
(3) Rewrite the sum over unordered pairs with $k_j \neq k_l$ as a sum over ordered pairs with $k_j > k_l$; pairs with $|k_j - k_l| = 2$ contribute 0.

The index depends only on the list (k_j), not on U: the geodesics with a given list form one critical manifold, the orbit of D under conjugation.

The minimum. The energy (7.127) is smallest when every $|k_j| = 1$. Then $\sum k_j = 0$ forces m entries $+1$ and m entries -1:

$$E_{\min} = 2m\pi^2, \qquad \lambda = 0 \ \text{(every difference is 0 or 2)}.$$

Such an X is $i\pi$ on an m-dimensional complex subspace $V = \ker(X - i\pi)$ of $\mathbb{C}^{2m}$ and $-i\pi$ on its orthogonal complement. Conversely every m-dimensional V gives one such X. So

$$\{\text{minimal geodesics } I \to -I\} \;\cong\; G_m(\mathbb{C}^{2m}), \qquad \dim_{\mathbb{R}} G_m(\mathbb{C}^{2m}) = 2m^2, \tag{7.129}$$

the **Grassmannian** of complex m-planes in $\mathbb{C}^{2m}$. As a check of non-degeneracy: the nullity is the number of conjugate points at $t = 1$, which is 2 for each of the $m \cdot m$ pairs with $k_j = +1$, $k_l = -1$, i.e. $2m^2$, the dimension of the Grassmannian. For $m = 1$ this is $G_1(\mathbb{C}^2) = \mathbb{CP}^1 = S^2$, the sphere N_0 of Example 7.60.

The gap. Every non-minimal geodesic has index at least $2m + 2$. Suppose some $|k_j| \geq 3$; by the symmetry $X \to -X$ we may take $k_1 \geq 3$. Let p and q be the numbers of positive and negative entries, $p + q = 2m$.

- If $q \geq m + 1$: each negative $k_l \leq -1$ pairs with k_1 to give $k_1 - k_l - 2 \geq 3 + 1 - 2 = 2$ in (7.128), so $\lambda \geq 2q \geq 2m + 2$.
- If $q \leq m$, then $p \geq m \geq q$. The positive entries add up to at least $3 + (p - 1) = p + 2$, so the negative ones add up to at most $-(p + 2)$; with only $q \leq p$ of them, some $k_l \leq -3$. Each of the p positive entries pairs with this k_l to give at least $1 + 3 - 2 = 2$, and the pair (k_1, k_l) gives at least $3 + 3 - 2 = 4$. So $\lambda \geq 2p + 2 \geq 2m + 2$.

The bound is attained, for instance by $k = (3, \underbrace{1, \dots, 1}_{m-2}, \underbrace{-1, \dots, -1}_{m+1})$ (sum $3 + (m-2) - (m+1) = 0$): only the $m + 1$ pairs $(3, -1)$ contribute, each 2, giving $2m + 2$. See Figure 7.40.

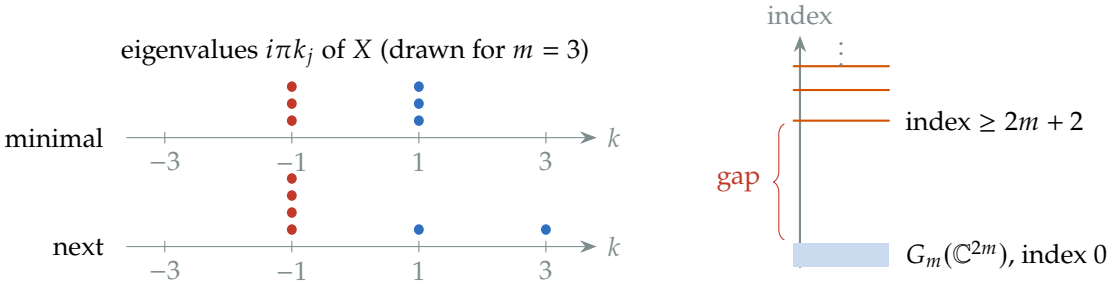


Figure 7.40: Left: the eigenvalue patterns of X for geodesics from I to $-I$ in $\mathrm{SU}(2m)$. The minimal ones have m eigenvalues $+i\pi$ (blue) and m eigenvalues $-i\pi$ (red); the lowest non-minimal ones move one eigenvalue out to $3i\pi$. Right: the resulting Morse–Bott ladder. The bottom rung is the Grassmannian; every other rung sits at index $2m + 2$ or higher, so up to dimension $2m$ the loop space cannot tell itself from the Grassmannian.

Step 4: from the gap to periodicity

What the gap buys. By the Bott–Morse–Smale picture (Section 7.3.4), extended to critical manifolds in Section 7.5.3, raising the energy past a critical manifold of index λ attaches cells of dimension $\geq \lambda$. So

$$\Omega\mathrm{SU}(2m) \;\simeq\; G_m(\mathbb{C}^{2m}) \;\cup\; (\text{cells of dimension } \geq 2m + 2).$$

Attaching a cell e^d cannot change π_i for $i \leq d - 2$: a map $S^i \to X \cup e^d$ with $i < d$ can be pushed off the centre of the cell and then radially out of it into X, and the same works for a homotopy $S^i \times [0, 1]$ between two maps as long as $i + 1 < d$. With $d \geq 2m + 2$:

$$\pi_i\big(\Omega\mathrm{SU}(2m)\big) \;\cong\; \pi_i\big(G_m(\mathbb{C}^{2m})\big) \qquad (i \leq 2m). \tag{7.130}$$

The Grassmannian's homotopy. Let $V_m(\mathbb{C}^{2m})$ be the **Stiefel manifold** of orthonormal m-frames $(v_1, \dots, v_m)$ in $\mathbb{C}^{2m}$. Forgetting the frame but keeping the plane it spans is a bundle

$$\mathrm{U}(m) \longrightarrow V_m(\mathbb{C}^{2m}) \longrightarrow G_m(\mathbb{C}^{2m}), \tag{7.131}$$

whose fibre is the set of orthonormal frames of a fixed m-plane, a copy of $\mathrm{U}(m)$. The Stiefel manifold is highly connected: choosing $v_1 \in S^{4m-1}$, then v_2 in the unit sphere S^{4m-3} of $v_1^{\perp}$, and so on down to $v_m \in S^{2m+1}$, builds it out of spheres of dimension $\geq 2m+1$, so $\pi_i(V_m(\mathbb{C}^{2m})) = 0$ for $i \leq 2m$. The exact sequence of (7.131),

$$\underbrace{\pi_i(V_m)}_{0} \to \pi_i(G_m) \to \pi_{i-1}(\mathrm{U}(m)) \to \underbrace{\pi_{i-1}(V_m)}_{0} \qquad (i \leq 2m),$$

gives $\pi_i(G_m(\mathbb{C}^{2m})) \cong \pi_{i-1}(\mathrm{U}(m))$.

The chain. For $1 \leq i \leq 2m$,

$$\begin{aligned}\pi_{i-1}(\mathrm{U}) &\overset{(1)}{=} \pi_{i-1}(\mathrm{U}(m)) \overset{(2)}{=} \pi_i\big(G_m(\mathbb{C}^{2m})\big) \overset{(3)}{=} \pi_i\big(\Omega\mathrm{SU}(2m)\big) \\ &\overset{(4)}{=} \pi_{i+1}\big(\mathrm{SU}(2m)\big) \overset{(5)}{=} \pi_{i+1}\big(\mathrm{U}(2m)\big) \overset{(6)}{=} \pi_{i+1}(\mathrm{U}).\end{aligned} \tag{7.132}$$

(Every "=" here is an isomorphism of groups.)

(1) Stability: $i-1 \leq 2m-1$ is in the stable range of $\mathrm{U}(m)$.
(2) The Stiefel bundle (7.131).
(3) The gap, (7.130).
(4) The loop shift $\pi_i(\Omega M) \cong \pi_{i+1}(M)$ (7.108).
(5) $\mathrm{U}(2m) \cong \mathrm{SU}(2m) \times S^1$ as a space, and $\pi_{i+1}(S^1) = 0$ for $i+1 \geq 2$.
(6) Stability again: $i+1 \leq 2m+1 \leq 2(2m)-1$.

Taking m as large as we like, $\pi_j(\mathrm{U}) \cong \pi_{j+2}(\mathrm{U})$ for every $j \geq 0$. Two starting values finish the job: $\pi_0(\mathrm{U}) = 0$ ($\mathrm{U}(n)$ is connected: every unitary matrix is exp of an anti-Hermitian one) and $\pi_1(\mathrm{U}) = \mathbb{Z}$ (the winding number of $\det g$, since $\pi_1(\mathrm{SU}(n)) = 0$). Hence $\pi_{\mathrm{even}}(\mathrm{U}) = 0$ and $\pi_{\mathrm{odd}}(\mathrm{U}) = \mathbb{Z}$, the bottom row of Table 7.1.

Worked example 7.61: Checking the chain in low degrees

$m = 1$**.** The gap statement (7.130) reads $\pi_i(\Omega S^3) \cong \pi_i(S^2)$ for $i \leq 2$. Directly: $\pi_1(\Omega S^3) = \pi_2(S^3) = 0 = \pi_1(S^2)$ and $\pi_2(\Omega S^3) = \pi_3(S^3) = \mathbb{Z} = \pi_2(S^2)$. The second is realised by the inclusion of the minimal geodesics: the map $S^2 \to \Omega S^3$ sending a direction $\hat{n}$ to the half great circle $\exp(i\pi t\, \hat{n} \cdot \vec{\sigma})$ (the orange sphere of Figure 7.39) generates $\pi_2(\Omega S^3) = \pi_3(S^3)$. This is the first case of the Freudenthal suspension theorem, $\pi_2(S^2) \cong \pi_3(S^3)$, proved here by counting geodesics.
$m = 2$**,** $i = 2$**.** The chain gives

$$\pi_3(\mathrm{SU}(4)) \overset{(1)}{=} \pi_2(\Omega\mathrm{SU}(4)) \overset{(2)}{=} \pi_2\big(G_2(\mathbb{C}^4)\big) \overset{(3)}{=} \pi_1(\mathrm{U}(2)) \overset{(4)}{=} \mathbb{Z}.$$

(1) Loop shift.
(2) The gap: $i = 2 \leq 2m = 4$.
(3) Stiefel bundle.
(4) $\det\colon \mathrm{U}(2) \to S^1$, with simply connected fibre $\mathrm{SU}(2) = S^3$.

So the instanton number of SU(4) gauge theory is "the same" as the winding number of a U(2) phase: an integer. The class in $\pi_2(G_2(\mathbb{C}^4))$ is the first Chern class of the tautological

2-plane bundle over a 2-sphere in the Grassmannian; this is where the monopole charge of Chapter 1 enters.
$m = 2$, $i = 3$ **and** $i = 4$. $\pi_4(\mathrm{SU}(4)) \cong \pi_2(\mathrm{U}(2)) = 0$ and $\pi_5(\mathrm{SU}(4)) \cong \pi_3(\mathrm{U}(2)) = \pi_3(S^3) = \mathbb{Z}$, matching Table 7.1.

Watch out: Periodicity is a stable statement

Bott's theorem says nothing about $\pi_j(\mathrm{U}(n))$ outside the range $j \le 2n-1$. That is where the physically interesting torsion lives: $\pi_4(\mathrm{SU}(2)) = \mathbb{Z}_2$ (the Witten anomaly), $\pi_6(\mathrm{SU}(3)) = \mathbb{Z}_6$, $\pi_6(\mathrm{SU}(2)) = \mathbb{Z}_{12}$. A theory with gauge group SU(2) in four dimensions is sensitive to $\pi_4(\mathrm{SU}(2))$ precisely because $4 > 2\cdot 2 - 1$: the group is too small to be stable in that degree.

The loop space of a compact group, and the real case

The whole loop space. Bott's analysis gives more than homotopy groups. For a compact simply connected Lie group G with a bi-invariant metric, the closed geodesics at e are $\exp(tX)$ with $\exp(X) = e$; the critical manifolds are conjugacy orbits $G/Z(X)$ ($Z(X)$ the centraliser), which have cohomology only in even degrees; and all indices are even (for $\mathrm{SU}(n)$ this is the $\mathbb{Z}$-valued version of (7.128), where each conjugate point has multiplicity 2). The lacunary principle (Section 7.2.7), in its Morse–Bott form, then says the energy is perfect and $H^{\mathrm{odd}}(\Omega G) = 0$. For $\mathrm{SU}(n)$ the result is

$$P_t\big(\Omega\mathrm{SU}(n)\big) = \prod_{k=1}^{n-1} \frac{1}{1-t^{2k}}, \tag{7.133}$$

which for $n = 2$ is the $1/(1-t^2)$ of Example 7.60. The loop group ΩG, with pointwise multiplication $(\gamma\delta)(t) = \gamma(t)\delta(t)$, is the infinite-dimensional group of Chapter 6; the Morse–Bott structure found here is the starting point of the Atiyah–Pressley and Atiyah–Bott analyses of loop groups and of Yang–Mills theory on Riemann surfaces (the Atiyah–Bott discussion at the end of Section 7.7.6).

The real (orthogonal) case. Bott ran the same argument through a chain of symmetric spaces for the orthogonal group, $\mathrm{O} \to \mathrm{O}/\mathrm{U} \to \mathrm{U}/\mathrm{Sp} \to \cdots$, each the space of minimal geodesics in the loop space of the previous one. After eight steps it returns to O:

$j \bmod 8$	0	1	2	3	4	5	6	7
$\pi_j(\mathrm{O})$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	0	$\mathbb{Z}$	0	0	0	$\mathbb{Z}$

Here $\pi_0(\mathrm{O}) = \mathbb{Z}_2$ counts the two components ($\det = \pm 1$), $\pi_1(\mathrm{SO}(n)) = \mathbb{Z}_2$ for $n \ge 3$ is the reason spinors exist (the double cover $\mathrm{Spin}(n) \to \mathrm{SO}(n)$, Chapter T4), and $\pi_3(\mathrm{O}) = \mathbb{Z}$ is again the instanton number.

Dictionary: Bott periodicity in physics

mathematics	physics
$\pi_1(\mathrm{U}) = \mathbb{Z}$	flux quantisation, vortex winding
$\pi_3(\mathrm{U}) = \mathbb{Z}$, $\pi_3(\mathrm{O}) = \mathbb{Z}$	instanton number, Skyrmion baryon number
$\pi_5(\mathrm{U}) = \mathbb{Z}$	quantised WZW level
$\pi_1(\mathrm{SO}(n)) = \mathbb{Z}_2$	spin: 2π rotation $= -1$ on spinors
period 2 (U) and period 8 (O)	the "periodic table" of topological insulators and superconductors (Kitaev)
$\mathrm{Cl}_{n+8} \cong \mathrm{Cl}_n \otimes \mathbb{R}(16)$	period-8 pattern of real spinor types (Chapter T4)
$\widetilde{K}(S^{2k}) = \pi_{2k-1}(\mathrm{U}) = \mathbb{Z}$	topological charges of vector bundles on spheres; the index theorem (Chapter 4)

Physical picture: Why a loop space "remembers" a Grassmannian

Think of a geodesic $\exp(i\pi t P)$ from I to $-I$, where P is $+1$ on a plane V and -1 on $V^\perp$. It rotates the phase of vectors in V by $+\pi$ and of vectors in $V^\perp$ by $-\pi$. A loop of unitary matrices is, after a gauge choice, a way of splitting $\mathbb{C}^{2m}$ into "positive" and "negative frequency" parts; this is the same splitting that defines a Dirac sea. The statement $\Omega\mathrm{U} \simeq \mathbb{Z} \times B\mathrm{U}$ (the infinite Grassmannian, with the integer counting the index) is the mathematical form of the physicists' observation that the Dirac sea has a charge and that families of Dirac seas are classified by vector bundles. It reappears in the treatment of anomalies (Chapter 10) and of the infinite-dimensional Grassmannians of Chapter 6.

7.8.12 Three more triumphs: Reeb, Milnor and Smale

Nash's opening list of the successes of Morse theory [Nash p. 192] has four entries: Morse's infinitely many geodesics on S^n (Section 7.8.7), Bott periodicity (just done), Milnor's exotic spheres, and Smale's proof of the Poincaré conjecture in dimensions ≥ 5. The last two are about finite-dimensional manifolds, and both run on the machinery of Section 7.3: a manifold is built from handles, one per critical point, and the handles can be counted, rearranged and cancelled. We explain the mechanism of each, at the level of "what is the Morse function, and what does it tell us". A common thread is a simple question.

If a closed manifold carries a function with only two critical points, what is it?

Reeb's theorem: two critical points means a sphere

The height function on the round S^n has two critical points, a minimum and a maximum (Example 7.6). The converse is the first result.

Key idea: Reeb's theorem

Let M be a closed n-manifold and $f\colon M \to \mathbb{R}$ a Morse function with exactly two critical points. Then M is *homeomorphic* to the sphere S^n.

First, the homology. This is a one-line use of the weak Morse inequalities.

Worked example 7.62: Two critical points force the homology of a sphere

A Morse function on a closed connected M has at least one minimum and at least one maximum (compactness), so the two critical points are a minimum (index 0) and a maximum (index n). For any coefficient field F and $0 < i < n$,

$$b_i(M;F) \overset{(1)}{\leq} m_i(f) \overset{(2)}{=} 0,$$

(1) The weak Morse inequality (7.15).
(2) The only critical points have index 0 and n.

and $b_0 = b_n = 1$ because M is connected and (by the $t = -1$ check $\chi = 1 + (-1)^n$ of (7.21)) the top Betti number is not lost. So $P_t(M) = 1 + t^n$: the Poincaré series of S^n. In the other direction, the torus can never carry such a function, since $m_1 \geq b_1(T^2) = 2$ forces at least four critical points, as for the upright torus of Section 7.1.7.

Homology alone does not identify a sphere (there are "homology spheres" with the same Betti numbers but a non-trivial fundamental group). Reeb's theorem is stronger, and the reason is geometric: sublevel sets.

Why it is true. Let p be the minimum and q the maximum, and rescale so $f(p) = 0$, $f(q) = 1$.

1. *Near the bottom.* By the Morse lemma (7.7), near p we have $f = x_1^2 + \cdots + x_n^2$. So for small ε the sublevel set $M_\varepsilon = \{f \leq \varepsilon\}$ is the round ball $\{|x|^2 \leq \varepsilon\}$: a closed disc D^n.

2. *The middle.* There are no critical values in $[\varepsilon, 1-\varepsilon]$. By the first half of the Bott–Morse–Smale theorem (Section 7.3.3), the flow of $\nabla f/|\nabla f|^2$ carries M_ε diffeomorphically onto $M_{1-\varepsilon}$. So $M_{1-\varepsilon}$ is also a disc.

3. *Near the top.* By the Morse lemma at q, $f = 1 - |y|^2$ nearby, and $\{f \geq 1-\varepsilon\}$ is again a disc $\{|y|^2 \leq \varepsilon\}$.

So M is two closed discs glued along their boundary spheres:

$$M \;\cong\; D^n \cup_\varphi D^n, \qquad \varphi\colon S^{n-1} \to S^{n-1} \text{ a diffeomorphism.} \tag{7.134}$$

Here φ records how the boundary of the lower disc is matched with the boundary of the upper one; it is produced by the flow, and we have no control over it.

Any such gluing is a topological sphere (the Alexander trick). For $\varphi = \mathrm{id}$, (7.134) is the northern and southern hemispheres: S^n. For a general φ, extend it to the whole upper disc by coning it off,

$$\Phi\colon D^n \to D^n, \qquad \Phi(r\,x) = r\,\varphi(x) \qquad (0 \leq r \leq 1,\ x \in S^{n-1}). \tag{7.135}$$

This is a bijection with inverse $r\,x \mapsto r\,\varphi^{-1}(x)$, and it is continuous with continuous inverse (at $r = 0$ both send small balls to small balls). Using id on the lower disc and Φ on the upper disc defines a homeomorphism $D^n \cup_{\mathrm{id}} D^n \to D^n \cup_\varphi D^n$, i.e. $S^n \to M$.

Watch out: Homeomorphic, not diffeomorphic

The cone (7.135) is smooth everywhere except at the centre $r = 0$. There its derivative would have to be "the derivative of φ at radius zero", which does not exist unless φ is linear. So Reeb's theorem gives a homeomorphism with at most one bad point. That one point is exactly the room in which exotic spheres live: a manifold can be $D^n \cup_\varphi D^n$, hence a topological sphere, and still fail to be *diffeomorphic* to S^n when φ is not isotopic to a rotation. These are Milnor's spheres.

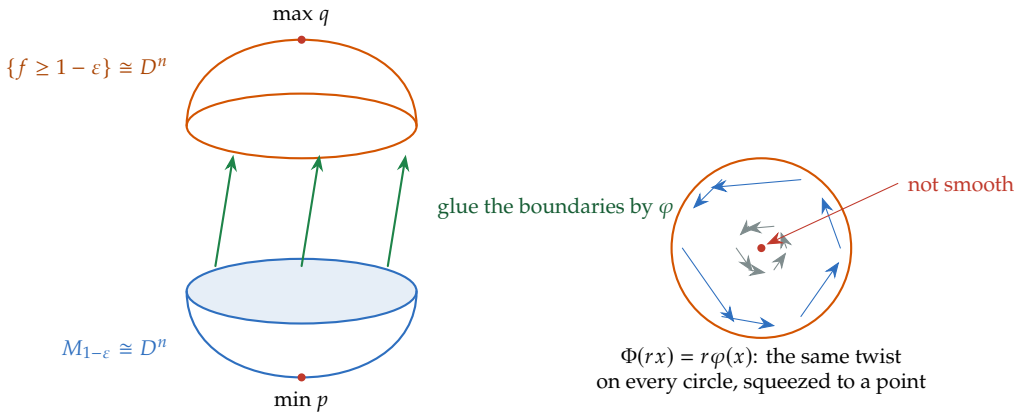


Figure 7.41: Reeb's theorem. Left: the sublevel set below any regular level is a disc (blue), the region above it is another disc (orange), and M is the two glued along their boundary spheres by some map φ. Right: the Alexander trick extends φ from the boundary circle to the whole disc by applying the same twist on every concentric sphere. The result is continuous, but at the centre (red) it is not differentiable. So M is a sphere topologically; smoothly it may not be.

Milnor's exotic 7-spheres

The question. Is there a smooth manifold homeomorphic to S^7 but not diffeomorphic to it? Milnor (1956) said yes, and his proof is a Morse function with two critical points plus a characteristic-class calculation. We give each step, using quaternions, with the physicist's reminder that quaternions are 2×2 matrices.

Working definition 7.28 (Quaternions as SU(2))

A quaternion is $u = u_0 + u_1\mathbf{i} + u_2\mathbf{j} + u_3\mathbf{k}$, $u_\mu \in \mathbb{R}$, with $\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = \mathbf{ijk} = -1$. Represent it as the matrix

$$u \;\longleftrightarrow\; u_0\,\mathbb{1} - i(u_1\sigma_1 + u_2\sigma_2 + u_3\sigma_3) = \begin{pmatrix} u_0 - iu_3 & -iu_1 - u_2 \\ -iu_1 + u_2 & u_0 + iu_3 \end{pmatrix},$$

with σ_a the Pauli matrices [the three 2×2 matrices of spin $\frac{1}{2}$, written out in Example 7.60]. Then $|u|^2 = u_0^2 + \cdots + u_3^2 = \det u$, the real part is $\operatorname{Re} u = u_0 = \frac{1}{2}\operatorname{tr} u$, and the unit quaternions $|u| = 1$ are exactly $\mathrm{SU}(2) = S^3$. The space of all quaternions is $\mathbb{H} \cong \mathbb{R}^4$ [as a real vector space, with coordinates (u_0, u_1, u_2, u_3)].

Step 1: sphere bundles over S^4. The sphere S^4 is $\mathbb{H} \cup \{\infty\}$, covered by two charts $\mathbb{R}^4$ that overlap in $\mathbb{H} \setminus \{0\}$, with transition $u \mapsto u^{-1}$ (the quaternionic version of the Riemann sphere $\mathbb{CP}^1$). A bundle with fibre S^3 over S^4 is two trivial pieces $\mathbb{R}^4 \times S^3$ glued over the overlap by a

transition function (Chapter 1). For integers h, j Milnor takes

$$(u, v) \sim (u', v') = \Big(\frac{u}{|u|^2},\ \frac{u^h\, v\, u^j}{|u|^{h+j}} \Big), \qquad u \in \mathbb{H} \setminus \{0\},\ v \in S^3. \tag{7.136}$$

The fibre map $v \mapsto u^h v u^j / |u|^{h+j}$ is left and right multiplication by unit quaternions, so it preserves $|v| = 1$. Call the total space $M_{h,j}$: a closed smooth 7-manifold. The simplest case $(h, j) = (1, 0)$ is the quaternionic Hopf fibration $S^3 \to S^7 \to S^4$, the SU(2) instanton bundle of Chapter 8.

Step 2: when is $M_{h,j}$ a homotopy sphere? The bundle has two integer invariants read off from the clutching map: the Euler number $h + j$ and the first Pontryagin number $\pm 2(h - j)$. The Euler class controls the cohomology via the Gysin sequence (the same tool as in Section 7.6.5): $H^4(M_{h,j}) = \mathbb{Z}/(h + j)$, and all other cohomology in degrees 1 to 6 vanishes. So set

$$h + j = 1, \qquad h - j = k \quad (k \text{ odd}), \qquad \text{i.e.} \quad h = \tfrac{1+k}{2},\ j = \tfrac{1-k}{2}, \tag{7.137}$$

and write M_k. Then M_k has the cohomology of S^7.

Step 3: a Morse function with two critical points. On the first chart, Milnor takes

$$f(u, v) = \frac{\operatorname{Re} v}{\sqrt{1 + |u|^2}}. \tag{7.138}$$

When $h + j = 1$, a short quaternion computation shows that in the second chart the same function reads $f = \operatorname{Re} u'' / \sqrt{1 + |u''|^2}$, where $u'' = u'(v')^{-1}$ is another quaternion coordinate. That expression has no critical points at all: its derivative in the real direction of u'' is $(1 + |u''|^2 - (\operatorname{Re} u'')^2)/(1 + |u''|^2)^{3/2} > 0$. So all critical points lie in the first chart, where $\partial f / \partial u = 0$ forces $u = 0$ and then v must be critical for $\operatorname{Re} v$ on S^3, i.e. $v = \pm 1$.

Worked example 7.63: The two critical points of Milnor's function are non-degenerate

Near $(u, v) = (0, 1)$ write $v = \sqrt{1 - |w|^2} + w$ with w a small imaginary quaternion ($w \in \mathbb{R}^3$ is the coordinate on S^3 near 1). Then

$$\begin{aligned} f &\overset{(1)}{=} \sqrt{1 - |w|^2}\,\big(1 + |u|^2\big)^{-1/2} \\ &\overset{(2)}{=} \big(1 - \tfrac{1}{2}|w|^2 + \dots\big)\big(1 - \tfrac{1}{2}|u|^2 + \dots\big) \\ &\overset{(3)}{=} 1 - \tfrac{1}{2}|u|^2 - \tfrac{1}{2}|w|^2 + O(4). \end{aligned}$$

(1) $\operatorname{Re} v = \sqrt{1 - |w|^2}$, since w is purely imaginary.
(2) Binomial series for both square roots, as in Example 7.6.
(3) Multiply out and keep the quadratic terms.

The Hessian in the seven coordinates $(u, w) \in \mathbb{R}^4 \times \mathbb{R}^3$ is $-\mathbb{1}_7$: a non-degenerate maximum of index 7. At $(0, -1)$ the same computation with an overall minus sign gives $+\mathbb{1}_7$: a minimum of index 0. So f is a Morse function on M_k with exactly two critical points, and by Reeb's theorem M_k is homeomorphic to S^7 for every odd k.

Step 4: a smooth invariant that tells them apart. Suppose M_k were diffeomorphic to S^7. M_k is the boundary of the associated disc bundle W_k (fibre D^4), an 8-manifold with boundary.

Glue an 8-disc to W_k along $\partial W_k = M_k \cong S^7$: the result $\overline{W}_k$ is a closed smooth 8-manifold. Its topology is simple: $H^4(\overline{W}_k) = \mathbb{Z}$, generated by a class e dual to the zero section S^4, with $e^2[\overline{W}_k] = h + j = 1$ (the zero section meets a small push-off of itself in $h + j$ points; the Euler number). So the intersection form is the 1×1 matrix (1) and the signature is $\sigma = 1$. The first Pontryagin class is $p_1 = 2k\,e$ (sign fixed by orientation). Hirzebruch's signature theorem for an 8-manifold (the L-genus of Chapter 4) reads $\sigma = \frac{1}{45}(7p_2 - p_1^2)[\overline{W}_k]$.

Worked example 7.64: The integrality test

Solve the signature theorem for the Pontryagin number $p_2[\overline{W}_k]$:

$$p_2[\overline{W}_k] \overset{(1)}{=} \frac{45\,\sigma + p_1^2[\overline{W}_k]}{7} \overset{(2)}{=} \frac{45 + (2k)^2\,e^2[\overline{W}_k]}{7} \overset{(3)}{=} \frac{45 + 4k^2}{7}.$$

(1) Rearrange $45\sigma = 7p_2 - p_1^2$.
(2) $\sigma = 1$ and $p_1 = 2k\,e$.
(3) $e^2[\overline{W}_k] = 1$.

A Pontryagin number of a closed smooth manifold is an integer. Check:

- $k = 1$: $(45 + 4)/7 = 7$. An integer, as it must be, since $M_1 = S^7$ (the Hopf fibration) and $\overline{W}_1 = \mathbb{HP}^2$, the quaternionic projective plane.
- $k = 3$: $(45 + 36)/7 = 81/7$. **Not an integer.**

So the assumption was false: M_3 is homeomorphic to S^7 (by Step 3) but not diffeomorphic to it. It is an **exotic sphere**.

Try it

For which odd k does the test fail to detect anything? *Answer:* $(45 + 4k^2)/7 \in \mathbb{Z}$ iff $4k^2 \equiv -45 \equiv 4 \pmod 7$ iff $k^2 \equiv 1 \pmod 7$ iff $k \equiv \pm 1 \pmod 7$. So $k = 1, 13, 15, \ldots$ pass the test, and $k = 3, 5, 7, 9, 11$ are all exotic. (Milnor's finer invariant, $\lambda(M_k) \equiv k^2 - 1 \pmod 7$, is the same calculation packaged as an invariant of M_k alone.)

Kervaire and Milnor later showed that there are exactly 28 smooth structures on S^7 up to orientation-preserving diffeomorphism (Chapter 1 discusses the count, and Chapter 10 meets the same group through $\pi_0\,\mathrm{Diff}(S^n)$). The role of Morse theory was decisive: without Reeb's theorem there was no way to know that the manifolds M_k are spheres at all.

Physical picture: Why the physicist should care

An exotic sphere is a spacetime that no topological measurement can distinguish from a sphere, but on which "smooth" means something different: the space of smooth metrics, of smooth gauge fields, and of solutions of field equations can change. The gravitational instantons and global gravitational anomalies of Chapter 10 are sensitive to exactly this kind of difference, and in four dimensions (exotic $\mathbb{R}^4$, Chapter 1) gauge theory itself is the tool that detects it (Donaldson theory, Chapter 12).

Smale: cancelling handles and the Poincaré conjecture

The generalised Poincaré conjecture says: a closed n-manifold with the homotopy type of S^n is homeomorphic to S^n. Smale (1961) proved it for $n \geq 5$. By Reeb's theorem it would be enough to find a Morse function with only two critical points. Smale's idea is to start with *any* Morse function and remove critical points in pairs until only two are left.

The basic move: cancelling a pair. We have already seen pairs of critical points born or annihilated in the family $x^3 - 3\varepsilon x$ (Figure 7.3). The move is to run that family backwards, on purpose.

Worked example 7.65: A cancelling pair in the plane

Take $f_\varepsilon(x, y) = x^3 - 3\varepsilon x + y^2$ on $\mathbb{R}^2$, with $\varepsilon > 0$. The gradient is $(3x^2 - 3\varepsilon,\, 2y)$, which vanishes at $y = 0$, $x = \pm\sqrt{\varepsilon}$. The Hessians are

$$Hf_\varepsilon(\pm\sqrt{\varepsilon}, 0) \overset{(1)}{=} \begin{pmatrix} 6x & 0 \\ 0 & 2 \end{pmatrix}_{x=\pm\sqrt{\varepsilon}} = \begin{pmatrix} \pm 6\sqrt{\varepsilon} & 0 \\ 0 & 2 \end{pmatrix},$$

(1) Second derivatives $\partial_x^2 f = 6x$, $\partial_y^2 f = 2$, mixed derivative 0.

so $(+\sqrt{\varepsilon}, 0)$ is a minimum (index 0) and $(-\sqrt{\varepsilon}, 0)$ is a saddle (index 1). The critical values are

$$f_\varepsilon(\pm\sqrt{\varepsilon}, 0) \overset{(1)}{=} \pm\varepsilon^{3/2} \mp 3\varepsilon^{3/2} \overset{(2)}{=} \mp 2\varepsilon^{3/2}.$$

(1) Substitute $x = \pm\sqrt{\varepsilon}$, $y = 0$.
(2) Collect terms.

The segment $-\sqrt{\varepsilon} < x < \sqrt{\varepsilon}$, $y = 0$ is a flow line of $-\nabla f$ (since $\partial_y f = 0$ on $y = 0$ and $-\partial_x f = 3(\varepsilon - x^2) > 0$ there): *exactly one* flow line runs from the saddle down to the minimum. As $\varepsilon \to 0$ the two points and their flow line shrink to the origin, and for $\varepsilon < 0$ the gradient $(3x^2 + 3|\varepsilon|, 2y)$ never vanishes. Two critical points, of adjacent index, have disappeared without changing f far away.

Key idea: Handle cancellation (Morse, Smale)

Let p and q be critical points of index $\lambda + 1$ and λ, with no critical values between $f(q)$ and $f(p)$. If exactly one gradient flow line runs from p to q, then f can be modified in a neighbourhood of that flow line to a function with the same critical points except that p and q are gone. In handle language: the $(\lambda + 1)$-handle is attached across the λ-handle exactly once, and together they form a ball glued along a ball, which changes nothing.

This is where the Morse complex of Section 7.3.9 becomes a geometric tool. There, a flow-line count $\partial\langle p\rangle = \pm\langle q\rangle$ was an algebraic statement. Here, "exactly one flow line" is a geometric statement, and the two differ whenever several flow lines with opposite signs cancel algebraically.

The proof strategy, step by step. Let M^n be a closed simply connected manifold with the homology of S^n, $n \geq 6$ (the case $n = 5$ needs an extra argument).

1. Remove two small open discs from M. What is left, W, is a compact manifold with two boundary spheres $\partial_0 W \cong \partial_1 W \cong S^{n-1}$, and the homology of W relative to $\partial_0 W$

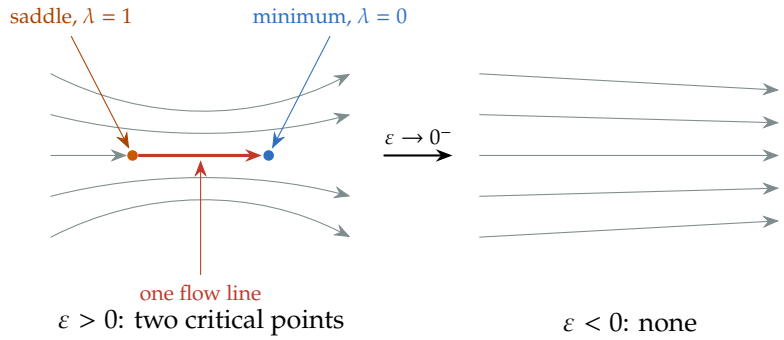


Figure 7.42: Cancelling a pair of critical points, Example 7.65, drawn as the flow of $-\nabla f$ (grey). Left: a saddle (orange) and a minimum (blue) joined by exactly one flow line (red). Right: after the move the flow runs straight through; no critical points are left. The move is local: far from the red segment the flow is unchanged.

vanishes: W looks like a cylinder $S^{n-1} \times [0,1]$ as far as homology can tell. Such a W is an **h-cobordism**.

2. Choose a Morse function $f : W \to [0,1]$ with $f = 0$ on $\partial_0 W$ and $f = 1$ on $\partial_1 W$. Rearrange it so that critical points of lower index have lower critical values (this is always possible: flow lines between critical points of the "wrong" order can be pushed apart).

3. Get rid of indices 0 and 1 (and dually n and $n-1$) using the simple connectivity: each index-0 point cancels against an index-1 point, and extra index-1 points are traded for index-3 points.

4. The Morse complex computes $H_*(W, \partial_0 W) = 0$, so the boundary maps are isomorphisms, and after a change of basis each $\lambda + 1$ critical point is paired with one λ critical point by an algebraic count ± 1.

5. **The Whitney trick.** In a level set between the two, the descending sphere of p (dimension λ) and the ascending sphere of q (dimension $n-1-\lambda$) meet in the points counted by the flow lines. When the algebraic count is ± 1 but the geometric count is larger, pairs of intersection points of opposite sign are removed by pushing one sphere across an embedded 2-disc (a "Whitney disc"). The two spheres have complementary dimensions, $\lambda + (n-1-\lambda) = n-1$, so they meet in isolated points. A generic 2-disc in a d-dimensional manifold is embedded when $2+2 < d$ and misses a k-dimensional submanifold when $2 + k < d$; in the level set, $d = n-1$, and once the low indices are gone both conditions hold for $n - 1 \geq 5$. This is where the dimension enters (the case $n = 5$ needs extra care).

6. Now each pair has exactly one flow line and cancels (Figure 7.42). When nothing is left, f has no critical points, and by Section 7.3.3 $W \cong S^{n-1} \times [0,1]$ (the **h-cobordism theorem**). Putting back the two discs, $M \cong D^n \cup (S^{n-1} \times [0,1]) \cup D^n$: a Morse function with two critical points, and Reeb's theorem finishes the job.

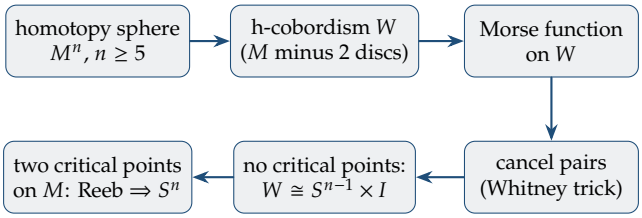


Figure 7.43: The logic of Smale's proof of the Poincaré conjecture in dimensions $n \geq 5$. Every arrow is Morse theory: sublevel sets, handles, the Morse complex, cancellation, and finally Reeb's two-critical-point theorem.

Dictionary: The three triumphs

result	Morse-theory mechanism
Reeb	two critical points $\Rightarrow$ two discs glued $\Rightarrow$ topological sphere
Milnor	explicit f with two critical points on an S^3-bundle over S^4 + a smooth invariant (Pontryagin numbers) $\Rightarrow$ exotic S^7
Smale	cancel handles in pairs (one flow line each) until two critical points remain $\Rightarrow$ Poincaré conjecture, $n \geq 5$
(physics)	Witten's tunnelling picture of §4 (part 7b) and Floer theory (Chapter 12) are the quantum version of counting flow lines

Watch out: Dimensions 3 and 4

The Whitney trick fails in dimension 4 (two 2-discs in a 4-manifold generically meet in points), and this is why low dimensions are different. The 4-dimensional topological Poincaré conjecture was proved by Freedman (1982) by other means; the smooth 4-dimensional one is still open. The 3-dimensional conjecture was proved by Perelman (2003) with Ricci flow, a gradient flow of a different kind. Nash's statement "for dim $M \geq 5$" [Nash p. 192] is the range where Morse theory alone does the job.

Reading Nash (pp. 205–207, (7.54)–(7.59); and p. 192)

Nash's discussion of infinite dimensions is two pages. He defines the energy $E(\gamma) = \int_0^1 g_{ij}\,\dot\gamma^i\dot\gamma^j\,dt$ [Nash (7.54)] on the path space $PM(p,q)$ [Nash (7.55)] (our (7.90), with the same normalisation), says that its extrema are the geodesics [Nash (7.56)] (our first variation, (7.95)), that at a geodesic $E = L^2$ [Nash (7.57)] and in general $E \geq L^2$ [Nash (7.58)] (Section 7.8.3, via Cauchy–Schwarz). For closed geodesics he replaces $PM(p,q)$ by $\mathrm{Map}(S^1, M)$ [Nash (7.59)] and remarks that U(1) acts by rotating the circle, so equivariant Morse theory should be relevant (Section 7.8.10). He sketches Morse's approximation of loops by geodesic polygons, $\mathrm{Map}_n(S^1, S^k)$, and the passage $n \to \infty$ (Section 7.8.9), with Klingenberg as the reference. He closes by saying that replacing S^k by a Lie group G and studying ΩG is Bott's route to periodicity (Section 7.8.11).

What we added: the second variation and the Hessian of E (7.100), Jacobi fields, conjugate points and the Morse index theorem (7.103), the explicit index count on S^n and the Poincaré series of ΩS^n (7.112), the Palais–Smale condition that makes infinite-dimensional Morse theory legitimate, the explicit time-slicing model, the free loop space fibration and the index growth of iterated closed geodesics, the full Morse-theoretic proof of complex Bott

periodicity, and the Reeb–Milnor–Smale applications that Nash lists on p. 192 but does not discuss.
Two remarks on Nash's wording. His sentence that one "eventually deduces ... the existence of an infinite number of closed geodesics on S^k" should be read as describing the programme; the precise known results are the ones listed in Section 7.8.10. And his claim that E is perfect holds for the round metric: on the based path space $PS^k(p,q)$ by the lacunary principle (Section 7.8.7), and on the free loop space as a Morse–Bott function with $\mathbb{Z}_2$ coefficients (Section 7.8.10). For a general metric only the inequalities survive.

Recap

- The energy $E(\gamma) = \int_0^1 |\dot\gamma|^2 \mathrm{d}t$ on the path space $PM(p,q)$ is a Morse function in infinitely many variables. Its critical points are the geodesics from p to q, and $E \geq L^2$ with equality for constant-speed paths, in particular at geodesics.
- Its Hessian is the Jacobi operator $-\nabla_t^2 - R(\cdot,\dot\gamma)\dot\gamma$ with Dirichlet conditions; the index is the number of conjugate points, counted with multiplicity (Morse index theorem). On the round S^n the k-th geodesic has index $k(n-1)$.
- Palais–Smale (or time slicing by broken geodesics) makes Morse theory work in infinite dimensions. Result: $P_t(\Omega S^n) = 1/(1-t^{n-1})$, and every metric on S^n has infinitely many geodesics between two points.
- Closed geodesics are critical points of E on the free loop space $\mathrm{Map}(S^1, M)$, with a U(1) symmetry by rotation: each closed geodesic is a critical circle, a Morse–Bott and equivariant problem.
- Bott: on a compact Lie group, the minimal geodesics from I to $-I$ form a Grassmannian, and all other geodesics have large index. Hence $\Omega\mathrm{SU}(2m) \simeq G_m(\mathbb{C}^{2m})$ (the Grassmannian) up to degree $2m$, and $\pi_j(\mathrm{U}) = \pi_{j+2}(\mathrm{U})$: $\mathbb{Z}$ in odd degrees, 0 in even degrees (period 8 for O).
- Reeb: two critical points $\Rightarrow$ topological sphere. Milnor: such a function on an S^3-bundle over S^4 plus an integrality test gives exotic 7-spheres. Smale: cancelling handles gives the Poincaré conjecture for $n \geq 5$.

7.9 Supersymmetric quantum mechanics and Morse theory: the plan

The first half of this chapter was pure topology. We took a function $f : M \to \mathbb{R}$ on a compact manifold, looked at its critical points, and found that they are forced to be at least as numerous as the "holes" of M. That is the content of the Morse inequalities.

In 1982 Edward Witten gave a completely different explanation of the same inequalities [Nash p. 207]. His explanation uses nothing but quantum mechanics: a Hamiltonian, its ground states, perturbation theory about classical minima, and tunnelling. The question of this section is simple: *what does a quantum-mechanical particle on M have to do with the critical points of a function on M?* We answer it in outline here, and then spend the rest of the chapter filling in every step.

7.9.1 Three facts we will string together

Witten's argument has three ingredients. Each one is a fact a physicist already believes, used in an unfamiliar setting.

1. **Ground states count topology.** There is a quantum system whose wavefunctions are differential forms on M and whose Hamiltonian is the Hodge Laplacian $\Delta = \mathrm{d}\mathrm{d}^* + \mathrm{d}^*\mathrm{d}$. By Hodge theory its zero-energy states in the sector of p-forms are in one-to-one correspondence with the de Rham cohomology $H^p_{\mathrm{dR}}(M)$. So
$$b_p = \#\{\text{zero-energy states made of } p\text{-forms}\}.$$
2. **A potential can be switched on without changing the ground-state count.** Witten deforms the system by a real parameter t, using the Morse function f. The deformation is a similarity transformation of d, so the number of zero-energy states cannot change. But the Hamiltonian now contains a potential $t^2\,|\nabla f|^2$, which is huge except near the critical points of f.
3. **At large t everything localises.** When the potential is huge, low-energy wavefunctions must crowd into its minima, i.e. near the critical points of f. Near each critical point the system is a set of harmonic oscillators. Solving them shows that each critical point of index p supplies exactly one candidate zero-energy p-form. Tunnelling between critical points can lift some candidates away from zero, but it can never create new ones.

Putting the three together:

$$\begin{aligned} b_p &\overset{(1)}{=} \#\{\text{exact zero modes of } H \text{ in degree } p\} \\ &\overset{(2)}{=} \#\{\text{exact zero modes of } H_t \text{ in degree } p\} \\ &\overset{(3)}{\leq} \#\{\text{approximate zero modes of } H_t \text{ in degree } p \text{ at large } t\} \\ &\overset{(4)}{=} \#\{\text{critical points of index } p\} = m_p. \end{aligned} \tag{7.139}$$

(1) Hodge theory: harmonic p-forms represent de Rham cohomology one-to-one (Section 7.11).
(2) Witten's deformation $\mathrm{d}_t = e^{-tf}\mathrm{d}e^{tf}$ is conjugation by an invertible operator, which preserves cohomology (Section 7.12).
(3) Every exact zero mode is in particular a low-lying state, but tunnelling can push some low-lying states up to small nonzero energies (Section 7.15).
(4) The harmonic-oscillator analysis gives exactly one low-lying state per critical point, sitting in the degree equal to the index (Section 7.14).

The inequality $b_p \leq m_p$ is the *weak* Morse inequality. Witten goes further. The operator d_t acts on the low-lying states and squares to zero. Its matrix elements at large t are tunnelling amplitudes, computed by instantons, and those instantons turn out to be the gradient flow lines of f. So the low-lying states form a small cochain complex, one basis vector per critical point, with a differential that counts flow lines. Its cohomology is $H^*(M)$. That is the **Morse–Witten complex**, and it gives the *strong* Morse inequalities as a piece of linear algebra (Section 7.16).

7.9.2 Reminder: the Morse data we need

We collect the facts from the first half of the chapter (Chapter 7) that this half uses. Let M be a compact manifold without boundary, of dimension n, and $f : M \to \mathbb{R}$ smooth.

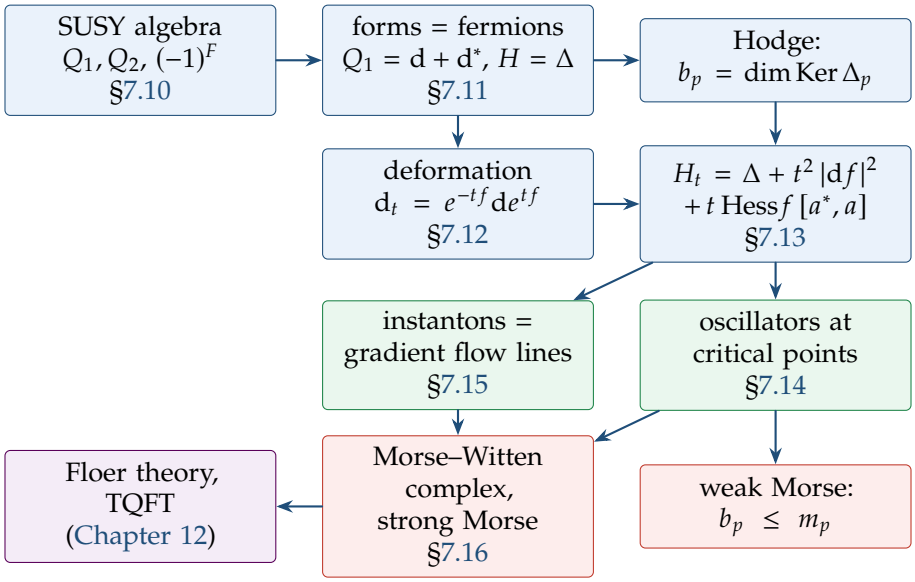


Figure 7.44: The logic of Witten's argument and of this half of the chapter. Blue boxes are algebra and geometry that hold exactly; green boxes are the large-t semiclassical analysis; red boxes are the two conclusions. The instanton box feeds only the *strong* inequalities: the weak ones need no tunnelling at all.

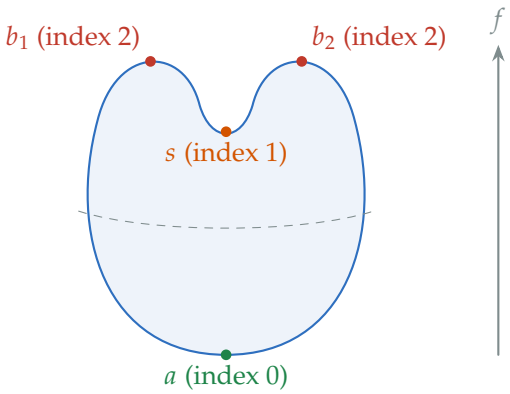
- A **critical point** is a point p where $df = 0$, i.e. all partial derivatives $\partial f/\partial x^i$ vanish.
- At a critical point the **Hessian** is the symmetric matrix $H_{ij} = \partial^2 f/\partial x^i \partial x^j$. The point is **non-degenerate** if $\det H_{ij} \neq 0$. Its **index** λ_p is the number of negative eigenvalues of H_{ij}: the number of independent directions in which f goes *down*.
- f is a **Morse function** if all its critical points are non-degenerate. Then there are finitely many of them. We write m_p for the number of critical points of index p.
- The **Morse lemma**: near a non-degenerate critical point of index λ there are coordinates with $f = f(p) - x_1^2 - \cdots - x_\lambda^2 + x_{\lambda+1}^2 + \cdots + x_n^2$.
- $b_p = \dim H^p_{\mathrm{dR}}(M)$ is the p-th Betti number. The Morse inequalities, written with a dummy variable u, are

$$\sum_p m_p u^p - \sum_p b_p u^p = (1+u)\, R(u), \qquad R(u) = \sum_p r_p u^p,\ r_p \ge 0. \tag{7.140}$$

  Setting $u = -1$ gives $\sum_p (-1)^p m_p = \sum_p (-1)^p b_p = \chi(M)$. Comparing coefficients of u^p gives $m_p - b_p = r_p + r_{p-1} \ge 0$, so in particular the weak inequalities $m_p \ge b_p$.

Watch out: Which letter is t?

Nash uses the letter t twice: for the variable of the Morse and Poincaré polynomials in §§1–3 ($M_t(f)$, $P_t(M)$), and for Witten's deformation parameter in §4. In this half of the chapter t *always* means Witten's parameter, a large positive number. We write the polynomial variable as u, so the Poincaré polynomial is $P_u(M) = \sum_p b_p u^p$ and the Morse polynomial is $M_u(f) = \sum_p m_p u^p$. The letter s will be Euclidean time along an instanton.



degree p	0	1	2
m_p (critical points)	1	1	2
b_p (S^2)	1	0	1
$m_p - b_p$	0	1	1

The saddle and one peak form a "cancelling pair": $m_u - P_u = u + u^2 = (1+u)\,u$, so $R(u) = u$.

Figure 7.45: A two-humped sphere and its height function: one minimum, one saddle, two maxima. Witten's large-t analysis will produce four low-lying states (one per dot, in degree equal to the index). Only two of them are true zero-energy states, matching $b_0 = b_2 = 1$. The saddle state and a combination of the two peak states will be paired up by tunnelling along the flow line joining them (Example 7.93).

7.9.3 What is new compared with the classical proof

Nash stresses one point [Nash p. 207]. The classical proof of the Morse inequalities (attaching a cell each time we pass a critical value) bounds the Betti numbers but does not hand us the cohomology in a usable form. Witten's argument constructs an explicit cochain complex whose generators are the critical points. Its differential is computed by counting gradient flow lines with signs. So the cohomology of M is *built* from the critical points of f and the flows between them.

This idea turned out to be enormously influential. When the manifold M is replaced by an infinite-dimensional space (the space of connections on a 3-manifold, or the loop space of a symplectic manifold), there are no Betti numbers to compare with directly, and ordinary Morse theory breaks down because indices are infinite. But the Morse–Witten complex still makes sense, since it only needs *differences* of indices and counts of flow lines. That is Floer theory, which Nash takes up in Chapter 12.

Dictionary: the whole of this chapter in one table

Mathematics	Physics
differential form on M	wavefunction of a particle on M with spin degrees of freedom
degree p of the form	fermion number F
d, d^*	supercharges
Hodge Laplacian Δ	Hamiltonian
harmonic forms, b_p	supersymmetric ground states
Euler characteristic $\chi(M)$	Witten index $\mathrm{Tr}(-1)^F$
Morse function f	superpotential
critical points of f	classical vacua
index of a critical point	fermion number of the perturbative vacuum
gradient flow line	instanton (tunnelling path)
Morse–Witten differential δ	instanton-corrected supercharge

Reading Nash (p. 207, opening of §4)

Nash introduces the section by saying Witten's paper "provides a new way of looking at Morse theory", that it is the "point of departure for the Floer theory discussed in chapter 12", and that "an important extra feature" of the quantum proof is that "the cohomology of M is also explicitly constructed". Everything in the flowchart of Figure 7.44 is compressed into the next eight pages of Nash (pp. 207–215, equations (7.60)–(7.95)). We take them one step at a time.

Recap

- Witten's proof has three moves: ground states of the Hodge Laplacian count Betti numbers; a deformation by f keeps the count but adds a potential $t^2\,|\nabla f|^2$; at large t the low-lying states sit one per critical point in degree equal to the index.
- Weak inequalities follow from the oscillators alone. Strong inequalities and the cohomology itself follow from tunnelling, i.e. from counting gradient flow lines.
- In this half of the chapter t is Witten's parameter and u is the polynomial variable.

7.10 The supersymmetry algebra

A symmetry of a quantum system is an operator that commutes with the Hamiltonian. It maps energy eigenstates to energy eigenstates of the same energy, so it organises the spectrum into multiplets. Rotations give the $(2l+1)$-fold degeneracy of hydrogen levels; isospin puts the proton and neutron together.

Supersymmetry is a symmetry that maps bosons into fermions and back [Nash p. 207]. It organises the spectrum into boson–fermion pairs. The question of this section is: *what algebra do such symmetries obey, and what can we learn just from that algebra?* The answer will be that the energy is never negative, that every state of positive energy has a partner of the opposite statistics, and that the difference

$$\text{(number of bosonic zero-energy states)} - \text{(number of fermionic zero-energy states)}$$

is unusually robust. It is the Witten index, and in Section 7.11 it will turn out to be the Euler characteristic.

7.10.1 Anticommutators and fermions: a reminder

For two operators A, B we write

$$[A,B] = AB - BA \quad \text{(commutator)}, \qquad \{A,B\} = AB + BA \quad \text{(anticommutator)}.$$

Bosonic creation and annihilation operators obey commutation relations, $[c, c^\dagger] = 1$. Fermionic ones obey anticommutation relations.

Working definition 7.29 (a single fermionic mode)

A **fermionic mode** is a pair of operators $b, b^\dagger$ with

$$\{b, b^\dagger\} = 1, \qquad \{b, b\} = 2b^2 = 0, \qquad (b^\dagger)^2 = 0. \tag{7.141}$$

[$(b^\dagger)^2 = 0$ is the Pauli principle, the mode cannot be filled twice] The **fermion number** is $F = b^\dagger b$ [it has eigenvalue 0 on the empty state and 1 on the occupied state]. *Operational test:* if you can find two operators with these relations, you have a two-state system, empty and occupied.

The relations are realised by 2×2 matrices. Take the basis $|0\rangle = \binom{1}{0}$ (empty) and $|1\rangle = \binom{0}{1}$ (occupied), and

$$b^\dagger = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \qquad b = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

Then

$$\{b, b^\dagger\} \overset{(1)}{=} \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$
$$\overset{(2)}{=} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} = \mathbf{1}, \qquad F = b^\dagger b = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

(1) Definition of the anticommutator, with the two matrices inserted.
(2) Multiply out: $bb^\dagger$ projects onto $|0\rangle$, $b^\dagger b$ projects onto $|1\rangle$.

And $(b^\dagger)^2 = 0$ because the matrix is strictly lower triangular. That is the Pauli exclusion principle: you cannot put two fermions in the same mode. The operator

$$(-1)^F = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = \sigma_z$$

is $+1$ on the empty state and -1 on the occupied one. It anticommutes with b and $b^\dagger$, since each of them changes F by one: $\sigma_z b = -b\,\sigma_z$, as one checks by multiplying the matrices.

With n independent modes $b_1, \dots, b_n$ (all anticommuting, $\{b_i, b_j^\dagger\} = \delta_{ij}$) we get 2^n states, labelled by which modes are occupied. We will meet exactly this Fock space in Section 7.11, where it will be the exterior algebra.

7.10.2 Supersymmetry generators

A supersymmetry generator (or **supercharge**) is a Hermitian operator Q that changes a boson into a fermion and vice versa. So it is itself "fermionic", and the natural algebraic relations for it are anticommutators. With N supercharges $Q_1, \dots, Q_N$, the first relation is [Nash p. 207, (7.60)]

$$\{Q_i, Q_j\} = Q_i Q_j + Q_j Q_i = 0 \qquad (i \neq j). \tag{7.142}$$

This says nothing about $Q_i^2 = \tfrac{1}{2}\{Q_i, Q_i\}$. The squares are fixed by the physics. In the case Nash considers, a theory in one space and one time dimension with $N = 2$, the squares are [Nash p. 207, (7.61)]

$$Q_1^2 = H + P, \qquad Q_2^2 = H - P, \tag{7.143}$$

where H is the Hamiltonian and P the momentum. This is the defining feature of supersymmetry: *a supercharge is a square root of the energy–momentum*. It is the fermionic analogue of the Dirac operator being a square root of the Laplacian.

Key idea: consequences of the algebra, Nash (7.62)

From (7.142) and (7.143),

$$[Q_i, H] = 0, \qquad H = \tfrac{1}{2}\big(Q_1^2 + Q_2^2\big), \qquad H \geq 0. \tag{7.144}$$

So the supercharges are genuine symmetries, and the energy is never negative.

Here is why. Adding the two equations in (7.143), $Q_1^2 + Q_2^2 = 2H$, which is the middle statement. For the first, take $i = 1$:

$$\begin{aligned}
[Q_1, H] &\overset{(1)}{=} \tfrac{1}{2}\big[Q_1,\, Q_1^2 + Q_2^2\big] \\
&\overset{(2)}{=} \tfrac{1}{2}\big(Q_1Q_2Q_2 - Q_2Q_2Q_1\big) \\
&\overset{(3)}{=} \tfrac{1}{2}\big(-Q_2Q_1Q_2 - Q_2Q_2Q_1\big) \\
&\overset{(4)}{=} \tfrac{1}{2}\big(-Q_2Q_1Q_2 + Q_2Q_1Q_2\big) = 0.
\end{aligned}$$

(1) Insert $H = \frac{1}{2}(Q_1^2 + Q_2^2)$.
(2) Q_1 commutes with its own square, so only the Q_2^2 term survives.
(3) Move Q_1 one place to the right in the first term using $Q_1Q_2 = -Q_2Q_1$ from (7.142).
(4) Do the same in the second term: $Q_2Q_1 = -Q_1Q_2$, so $-Q_2Q_2Q_1 = +Q_2Q_1Q_2$.

The same computation with $1 \leftrightarrow 2$ gives $[Q_2, H] = 0$. For positivity, take any state $|\psi\rangle$ and use $Q_i^\dagger = Q_i$:

$$\begin{aligned}
\langle\psi|H|\psi\rangle &\overset{(1)}{=} \tfrac{1}{2}\big(\langle\psi|Q_1Q_1|\psi\rangle + \langle\psi|Q_2Q_2|\psi\rangle\big) \\
&\overset{(2)}{=} \tfrac{1}{2}\big(\|Q_1|\psi\rangle\|^2 + \|Q_2|\psi\rangle\|^2\big) \overset{(3)}{\geq} 0.
\end{aligned}$$

(1) The middle statement of (7.144).
(2) $\langle\psi|Q_iQ_i|\psi\rangle = \big(Q_i|\psi\rangle\big)^\dagger\big(Q_i|\psi\rangle\big)$ because Q_i is Hermitian.
(3) A squared length is non-negative.

The last line tells us more than positivity. *A state has zero energy if and only if every supercharge annihilates it:* $H|\psi\rangle = 0 \iff Q_1|\psi\rangle = Q_2|\psi\rangle = 0$. We will use this over and over. The same argument applied to $Q_1^2 = H + P$ and $Q_2^2 = H - P$ separately gives $H \pm P \geq 0$, i.e. $H \geq |P|$: supersymmetry forbids anything from moving faster than light, energetically speaking.

7.10.3 Quantum mechanics: $P = 0$ and a nilpotent supercharge

In quantum mechanics (one time dimension and no space) there is no momentum generator, and $P = 0$ [Nash p. 209]. Then $Q_1^2 = Q_2^2 = H$. It is very convenient to combine the two Hermitian supercharges into one complex one:

$$Q = \tfrac{1}{2}(Q_1 - \mathrm{i}Q_2), \qquad Q^\dagger = \tfrac{1}{2}(Q_1 + \mathrm{i}Q_2), \qquad \text{equivalently} \quad Q_1 = Q + Q^\dagger,\ \ Q_2 = \mathrm{i}(Q - Q^\dagger). \tag{7.145}$$

Then

$$\begin{aligned}
Q^2 &\overset{(1)}{=} \tfrac{1}{4}\big(Q_1^2 - \mathrm{i}(Q_1Q_2 + Q_2Q_1) - Q_2^2\big) \\
&\overset{(2)}{=} \tfrac{1}{4}\big((H + P) - 0 - (H - P)\big) = \tfrac{1}{2}P, \\
\{Q, Q^\dagger\} &\overset{(3)}{=} \tfrac{1}{4}\big(2Q_1^2 + 2Q_2^2\big) = H.
\end{aligned} \tag{7.146}$$

(1) Expand $(Q_1 - iQ_2)^2$ keeping the order of the factors.
(2) Use (7.143) and (7.142).
(3) Expand $(Q_1 - iQ_2)(Q_1 + iQ_2) + (Q_1 + iQ_2)(Q_1 - iQ_2)$: the cross terms cancel in the sum, leaving $2Q_1^2 + 2Q_2^2$, which is $4H$.

So Q *squares to zero exactly when* $P = 0$. A nilpotent operator Q with $Q^2 = 0$ whose anticommutator with its adjoint is the Hamiltonian: that is exactly the structure of the exterior derivative d, its adjoint d^*, and the Laplacian $dd^* + d^*d$. We make this precise in Section 7.11.

7.10.4 Bosons, fermions and the block structure

To tell bosons from fermions, introduce the operator $(-1)^F$ ("-1 to the fermion number") [Nash p. 207, (7.63)]:

$$(-1)^F|P\rangle = \begin{cases} -|P\rangle & \text{if } P \text{ is a fermion,} \\ +|P\rangle & \text{if } P \text{ is a boson.} \end{cases} \tag{7.147}$$

It squares to one, and its two eigenspaces split the one-particle Hilbert space [Nash p. 208, (7.64)]:

$$\mathcal{H}_1 = \mathcal{H}_1^B \oplus \mathcal{H}_1^F, \qquad (-1)^F = \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix} \tag{7.148}$$

in a basis adapted to the splitting [Nash p. 208, (7.65)]. A space split like this is called a **$\mathbb{Z}_2$-graded** or **super** vector space.

Supercharges turn bosons into fermions. Algebraically, this means they *anticommute* with $(-1)^F$: $\{Q_i, (-1)^F\} = 0$. Write Q_i in blocks and impose this:

$$\begin{aligned} 0 = \{Q_i, (-1)^F\} &\overset{(1)}{=} \begin{pmatrix} A & B \\ C & D \end{pmatrix}\begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix} + \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix}\begin{pmatrix} A & B \\ C & D \end{pmatrix} \\ &\overset{(2)}{=} \begin{pmatrix} A & -B \\ C & -D \end{pmatrix} + \begin{pmatrix} A & B \\ -C & -D \end{pmatrix} = \begin{pmatrix} 2A & 0 \\ 0 & -2D \end{pmatrix}. \end{aligned}$$

(1) A general operator in block form, with $A\colon \mathcal{H}^B \to \mathcal{H}^B$, $B\colon \mathcal{H}^F \to \mathcal{H}^B$, $C\colon \mathcal{H}^B \to \mathcal{H}^F$, $D\colon \mathcal{H}^F \to \mathcal{H}^F$.
(2) Right multiplication by $(-1)^F$ flips the sign of the second column; left multiplication flips the sign of the second row.

So $A = D = 0$: the supercharges are block off-diagonal [Nash p. 208, (7.66)–(7.67)]:

$$Q_i = \begin{pmatrix} 0 & Q_i^{BF} \\ Q_i^{FB} & 0 \end{pmatrix}, \qquad Q_i^{BF}\colon \mathcal{H}_1^F \to \mathcal{H}_1^B, \quad Q_i^{FB}\colon \mathcal{H}_1^B \to \mathcal{H}_1^F. \tag{7.149}$$

Hermiticity of Q_i says $Q_i^{BF} = (Q_i^{FB})^\dagger$. The Hamiltonian, being a sum of squares of off-diagonal operators, is block *diagonal*: $H = \operatorname{diag}(Q_i^{BF}Q_i^{FB},\ Q_i^{FB}Q_i^{BF})$ for either i (in the $P = 0$ case). Bosons stay bosons under time evolution, as they must.

Worked example 7.66: the smallest supersymmetric system

Take a single boson state and a single fermion state, both of energy $E > 0$, i.e. $\mathcal{H} = \mathbb{C}^2$ with $(-1)^F = \sigma_z$ and $H = E\,\mathbf{1}$. Let $Q = \sqrt{E}\,b^\dagger$ (it creates the fermion out of the boson). Then

$$Q_1 \overset{(1)}{=} \sqrt{E}\,(b^\dagger + b) = \sqrt{E}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \sqrt{E}\,\sigma_x,$$

$$Q_2 \overset{(2)}{=} \mathrm{i}\sqrt{E}\,(b^\dagger - b) = \sqrt{E}\begin{pmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{pmatrix} = \sqrt{E}\,\sigma_y .$$

(1) $Q_1 = Q + Q^\dagger$ from (7.145), with the matrices of Section 7.10.1.
(2) $Q_2 = \mathrm{i}(Q - Q^\dagger)$; the factor i turns the real antisymmetric matrix into σ_y.

All relations can be read off from the Pauli algebra: $\{\sigma_x, \sigma_y\} = 0$ gives (7.142); $\sigma_x^2 = \sigma_y^2 = \mathbf{1}$ gives $Q_1^2 = Q_2^2 = E = H$; and σ_z anticommutes with σ_x, σ_y, so both supercharges are off-diagonal. The minimal "supermultiplet" is a spin-$\frac{1}{2}$ system in disguise: $Q_1, Q_2, (-1)^F$ are $\sqrt{E}\sigma_x, \sqrt{E}\sigma_y, \sigma_z$. Note that this construction needs $E > 0$: at $E = 0$ we would have $Q_1 = Q_2 = 0$ and nothing pairs the two states.

7.10.5 Supersymmetric quantum mechanics on a line

Here is the example that the whole chapter generalises. A particle on the real line with one fermionic mode. States are pairs of wavefunctions,

$$\Psi = \begin{pmatrix} \varphi_0(x) \\ \varphi_1(x) \end{pmatrix}, \qquad \varphi_0 \in \mathcal{H}^B,\ \varphi_1 \in \mathcal{H}^F,$$

the upper component with the fermion mode empty, the lower one with it occupied. Choose a real function $W(x)$, the **superpotential**, and define

$$A = \frac{\mathrm{d}}{\mathrm{d}x} + W'(x), \qquad A^\dagger = -\frac{\mathrm{d}}{\mathrm{d}x} + W'(x), \qquad Q = \begin{pmatrix} 0 & 0 \\ A & 0 \end{pmatrix}, \quad Q^\dagger = \begin{pmatrix} 0 & A^\dagger \\ 0 & 0 \end{pmatrix}. \tag{7.150}$$

$A^\dagger$ really is the adjoint of A on $L^2(\mathbb{R})$: integrating by parts, $\int \bar\chi\,(\varphi' + W'\varphi)\,\mathrm{d}x = \int \overline{(-\chi' + W'\chi)}\,\varphi\,\mathrm{d}x$ for functions that vanish at infinity. Clearly $Q^2 = 0$ (a strictly lower triangular block matrix). The Hamiltonian is

$$H = \{Q, Q^\dagger\} = \begin{pmatrix} A^\dagger A & 0 \\ 0 & AA^\dagger \end{pmatrix} = \begin{pmatrix} H_B & 0 \\ 0 & H_F \end{pmatrix}. \tag{7.151}$$

Let us compute the two blocks. Acting on a test function φ,

$$\begin{aligned} A^\dagger A\,\varphi &\overset{(1)}{=} \Big(-\frac{\mathrm{d}}{\mathrm{d}x} + W'\Big)\big(\varphi' + W'\varphi\big) \\ &\overset{(2)}{=} -\varphi'' - W''\varphi - W'\varphi' + W'\varphi' + W'^2\varphi \\ &= \Big(-\frac{\mathrm{d}^2}{\mathrm{d}x^2} + W'^2 - W''\Big)\varphi, \end{aligned} \tag{7.152}$$

$$AA^\dagger\,\varphi \overset{(3)}{=} \Big(-\frac{\mathrm{d}^2}{\mathrm{d}x^2} + W'^2 + W''\Big)\varphi. \tag{7.153}$$

(1) Definitions (7.150).
(2) Product rule on $(W'\varphi)'$; the two first-derivative terms $\mp W'\varphi'$ cancel.
(3) The same computation with the signs of the derivative terms swapped; now $+(W'\varphi)' - W'\varphi' = +W''\varphi$.

So the boson and the fermion feel *different* potentials,

$$V_B = W'^2 - W'', \qquad V_F = W'^2 + W'',$$

differing by the "spin–orbit" term $\pm W''$. Supersymmetry guarantees that, apart from zero-energy states, the two Schrödinger operators have exactly the same spectrum. The Hermitian supercharges are $Q_1 = \left(\begin{smallmatrix} 0 & A^\dagger \\ A & 0 \end{smallmatrix}\right)$ and $Q_2 = \left(\begin{smallmatrix} 0 & -\mathrm{i}A^\dagger \\ \mathrm{i}A & 0 \end{smallmatrix}\right)$, so in Nash's notation $Q_1^{FB} = A$ and $Q_1^{BF} = A^\dagger$.

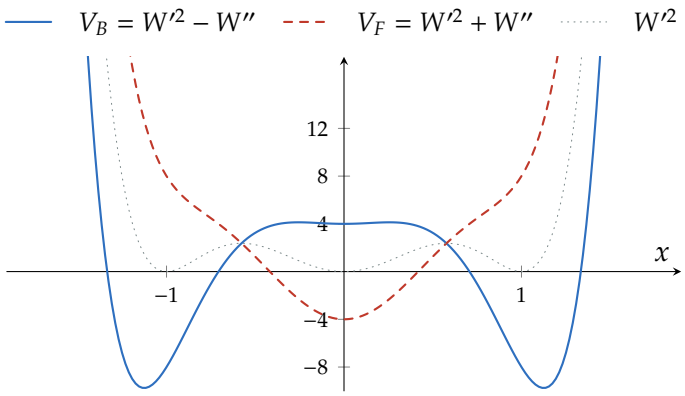


Figure 7.46: The two potentials of supersymmetric quantum mechanics for the double-well superpotential $W = t\left(\frac{1}{4}x^4 - \frac{1}{2}x^2\right)$ at $t = 4$. The boson (solid) sits in deep wells at the minima $x = \pm 1$ of W, where $-W'' < 0$; the fermion (dashed) has its deepest well at the maximum $x = 0$, where $+W'' < 0$. The wells of both are the zeros of W'^2 (dotted), i.e. the critical points of W. This is the one-dimensional picture of the whole chapter.

Worked example 7.67: the supersymmetric harmonic oscillator

Take $W = \frac{1}{2}\omega x^2$ with $\omega > 0$. Then $W' = \omega x$ and $W'' = \omega$, so

$$H_B = -\frac{\mathrm{d}^2}{\mathrm{d}x^2} + \omega^2 x^2 - \omega, \qquad H_F = -\frac{\mathrm{d}^2}{\mathrm{d}x^2} + \omega^2 x^2 + \omega.$$

The operator $-\mathrm{d}^2/\mathrm{d}x^2 + \omega^2 x^2$ is a harmonic oscillator (in units with $\hbar = 1$, $m = \frac{1}{2}$) with levels $\omega(2n+1)$, $n = 0, 1, 2, \ldots$ (we derive this in Section 7.14.1). Hence

$$\operatorname{spec} H_B = \{\omega(2n+1) - \omega\} = \{0,\, 2\omega,\, 4\omega, \ldots\},$$
$$\operatorname{spec} H_F = \{\omega(2n+1) + \omega\} = \{2\omega,\, 4\omega,\, 6\omega, \ldots\}.$$

Every level $2k\omega$ with $k \geq 1$ contains one boson and one fermion; the level 0 contains a single boson, $\varphi_0 \propto e^{-\omega x^2/2}$. The zero-point energy $+\omega$ of the bosonic oscillator is exactly cancelled by the fermionic contribution $-\omega$: the famous cancellation of vacuum energies in supersymmetric theories, in its simplest form.

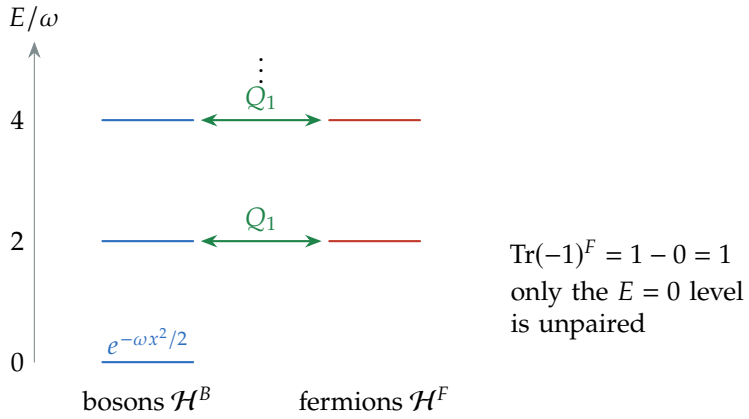


Figure 7.47: Spectrum of the supersymmetric harmonic oscillator. Every positive level is a boson–fermion pair, exchanged by the supercharge Q_1. Only the zero-energy level can be unpaired, and here it is a single boson.

Zero-energy states on the line. A zero-energy state must satisfy $Q\Psi = Q^\dagger\Psi = 0$, i.e. $A\varphi_0 = 0$ and $A^\dagger\varphi_1 = 0$. These are first-order equations:

$$\varphi_0' + W'\varphi_0 = 0 \implies \varphi_0 = C\,e^{-W(x)},$$
$$-\varphi_1' + W'\varphi_1 = 0 \implies \varphi_1 = C'\,e^{+W(x)}.$$

Whether these are physical depends on normalisability, i.e. on how W behaves at $\pm\infty$:

behaviour of W	zero-energy states	$n_B^0 - n_F^0$
$W \to +\infty$ at both ends (e.g. x^2, x^4)	boson e^{-W}	$+1$
$W \to -\infty$ at both ends (e.g. $-x^2$)	fermion e^{+W}	-1
$W \to \pm\infty$ at opposite ends (e.g. x^3)	none	0

In the last case there is no state of zero energy at all: the ground state has $E > 0$, and supersymmetry is **spontaneously broken**. Notice that the answer depends only on the behaviour of W far away, not on its wiggles. That is our first glimpse of a topological invariant.

7.10.6 The Witten index

The last example suggests counting zero-energy states with signs. Define $n_B(E)$ and $n_F(E)$ as the numbers of bosonic and fermionic states of energy E.

Working definition 7.30 (Witten index)

The **Witten index** of a supersymmetric system is

$$\operatorname{Tr}(-1)^F := n_B(0) - n_F(0). \tag{7.154}$$

[$(-1)^F$ is $+1$ on bosonic states and -1 on fermionic states] *Operational recipe:* find all zero-energy states (they are the states killed by every supercharge [the operators Q that turn bosonic states into fermionic ones and back, with H built from them]), and count bosons minus fermions. Equivalently, compute $\operatorname{Tr}\big((-1)^F e^{-\beta H}\big)$ [a sum over all states, of every energy, each weighted by $\pm e^{-\beta E}$] for any $\beta > 0$.

The "equivalently" is the key fact: *all positive-energy states cancel in pairs*. Here is the pairing argument. Let $E > 0$ and let $\mathcal{H}_E = \mathcal{H}_E^B \oplus \mathcal{H}_E^F$ be the eigenspace of H with eigenvalue E (in quantum mechanics, $P = 0$, so $Q_1^2 = H$).

- Q_1 maps $\mathcal{H}_E$ into itself, because $[Q_1, H] = 0$: if $H\psi = E\psi$ then $H(Q_1\psi) = Q_1 H\psi = E\,Q_1\psi$.
- Q_1 maps $\mathcal{H}_E^B$ into $\mathcal{H}_E^F$ and back, because it is off-diagonal (7.149).
- On $\mathcal{H}_E$, $Q_1^2 = H = E$, so Q_1 is invertible there, with inverse Q_1/E.

An invertible linear map between two vector spaces forces them to have the same dimension. Hence $n_B(E) = n_F(E)$ for every $E > 0$, and

$$\operatorname{Tr}\big((-1)^F e^{-\beta H}\big) \overset{(1)}{=} \sum_E e^{-\beta E}\big(n_B(E) - n_F(E)\big)$$
$$\overset{(2)}{=} n_B(0) - n_F(0) = \operatorname{Tr}(-1)^F. \tag{7.155}$$

(1) Evaluate the trace in an eigenbasis of H adapted to the splitting into bosons and fermions: each boson of energy E contributes $+e^{-\beta E}$, each fermion $-e^{-\beta E}$.
(2) Every $E > 0$ term vanishes by the pairing argument; only $E = 0$ is left, where $e^{-\beta\cdot 0} = 1$.

So the regulated trace does not depend on β. One can also see this directly by differentiating:

$$
\begin{aligned}
-\frac{\mathrm{d}}{\mathrm{d}\beta}\operatorname{Tr}\big((-1)^F e^{-\beta H}\big) &\overset{(1)}{=} \operatorname{Tr}\big((-1)^F Q_1 Q_1 e^{-\beta H}\big) \\
&\overset{(2)}{=} -\operatorname{Tr}\big(Q_1 (-1)^F Q_1 e^{-\beta H}\big) \\
&\overset{(3)}{=} -\operatorname{Tr}\big((-1)^F Q_1 e^{-\beta H} Q_1\big) \overset{(4)}{=} -\operatorname{Tr}\big((-1)^F Q_1 Q_1 e^{-\beta H}\big).
\end{aligned}
$$

(1) Differentiate the exponential and write $H = Q_1^2$.
(2) Q_1 anticommutes with $(-1)^F$.
(3) Cyclicity of the trace: move the first Q_1 to the end.
(4) Q_1 commutes with H, hence with $e^{-\beta H}$.

The derivative equals minus itself, so it is zero.

Key idea: why the Witten index is robust

Deform the system continuously while keeping supersymmetry (change the potential, the metric, a coupling constant). Energy levels move, but each positive level carries equal numbers of bosons and fermions, and a state can reach or leave $E = 0$ only together with its partner. So $n_B(0)$ and $n_F(0)$ can each jump, but only by the same amount: $n_B(0) - n_F(0)$ *cannot change*. The only escape is if states come in from or leave to infinity, which on a compact configuration space cannot happen.

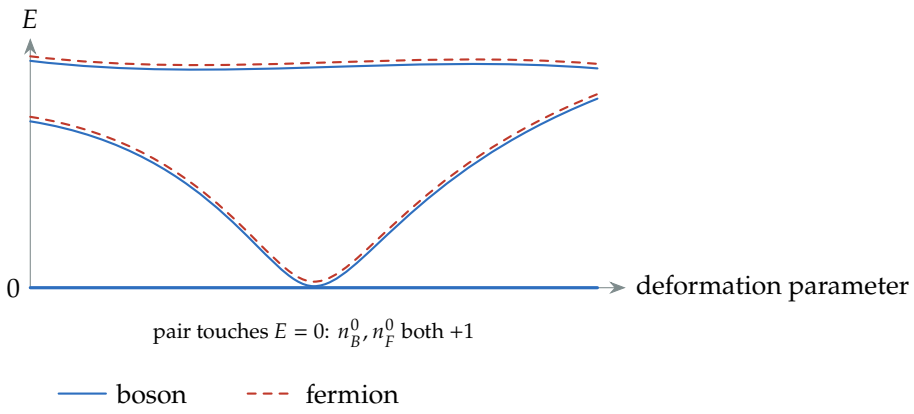


Figure 7.48: Levels of a supersymmetric system under a deformation. Positive levels always carry a boson (solid) and a fermion (dashed) together, so when a level comes down to $E = 0$ it adds one boson *and* one fermion to the zero-energy count. The zero-energy boson on the bottom line never moves. The difference $n_B(0) - n_F(0)$ stays equal to 1 throughout.

Physical picture: the path integral view

$\operatorname{Tr}\big((-1)^F e^{-\beta H}\big)$ is a partition function in Euclidean time β with an insertion of $(-1)^F$. In a path integral, inserting $(-1)^F$ changes the boundary condition of the fermions around the time circle from antiperiodic (the ordinary thermal trace) to *periodic*. With periodic boundary conditions the boson and fermion fluctuation determinants cancel mode by mode, which is the path-integral version of the pairing argument. As $\beta \to 0$ the path

integral reduces to an ordinary integral over the configuration space, and one can compute the index there; this is the route to the index theorems of Chapter 4.

Watch out: not every supersymmetric system has an index

On the line with $W = x^3$ the index is 0 and there is no zero-energy state. If we now deform $W = x^3$ into $W = x^3 - ax$ with $a > 0$, the answer does not change (the behaviour at infinity is the same). But now W has two critical points, a minimum at $x = +\sqrt{a/3}$ and a maximum at $x = -\sqrt{a/3}$, and V_B, V_F each have a deep well at one of them. Perturbation theory around each well predicts a zero-energy boson (at the minimum of W) and a zero-energy fermion (at the maximum), and the prediction is wrong: tunnelling between the wells lifts both to a small positive energy, as a boson–fermion pair. The index is robust; the naive semiclassical count of ground states is not. This is exactly the gap between the weak inequality $b_p \le m_p$ and the equality we will get from the Morse–Witten complex.

Reading Nash (pp. 207–208, (7.60)–(7.67))

Nash states the algebra (7.142)–(7.143) for a $(1+1)$-dimensional theory and deduces (7.144) "showing ... that the Q_i are bona fide symmetry operators, and ... that the Hamiltonian is non-negative"; we filled in both computations. His (7.63)–(7.67) are our (7.147)–(7.149), with the block-off-diagonal form derived from $\{Q_i, (-1)^F\} = 0$. Nash does not name the Witten index or give the pairing argument at this point; we included them because they are the reason the construction computes topology. Nash's "one particle states" $\mathcal{H}_1$ will be the space of all differential forms on M.

Recap

- Supercharges are Hermitian, anticommute with each other and with $(-1)^F$, and square to $H \pm P$. In quantum mechanics $P = 0$, and $Q = \frac{1}{2}(Q_1 - \mathrm{i}Q_2)$ is nilpotent with $\{Q, Q^\dagger\} = H$.
- Consequences: $[Q_i, H] = 0$, $H \ge 0$, and $H\psi = 0$ iff every supercharge kills ψ.
- Positive-energy states come in boson–fermion pairs, so the Witten index $\mathrm{Tr}(-1)^F = n_B(0) - n_F(0) = \mathrm{Tr}(-1)^F e^{-\beta H}$ is independent of β and of continuous deformations.
- On a line, $V_{B,F} = W'^2 \mp W''$; zero modes are $e^{\mp W}$; the index is fixed by the behaviour of W at infinity.

7.11 Supersymmetry from differential forms

The algebra of Section 7.10 needs a Hilbert space with a $\mathbb{Z}_2$-grading and a nilpotent operator Q whose anticommutator with its adjoint is the Hamiltonian. The question of this section is: *is there such a system attached to every Riemannian manifold, built from nothing but its geometry?* There is, and it is so natural that we have met all of its pieces before: the exterior derivative d is Q, the space of all differential forms is the Hilbert space, and the Hodge Laplacian is the Hamiltonian [Nash p. 208]. Along the way we find the first half of Witten's argument: the zero-energy states of this system are the harmonic forms, so they count the Betti numbers.

7.11.1 The Hilbert space of forms and the adjoint d^*

Let M be compact, oriented, without boundary, with a Riemannian metric g. Write $\Omega^p(M)$ for the p-forms. The metric gives the Hodge star $\star\colon \Omega^p \to \Omega^{n-p}$ and with it an inner product (Chapter T1):

$$\langle \alpha, \beta \rangle = \int_M \alpha \wedge \star\beta = \int_M \frac{1}{p!}\, \alpha_{i_1 \dots i_p}\, \beta^{i_1 \dots i_p}\, \sqrt{g}\, \mathrm{d}^n x \qquad (\alpha, \beta \in \Omega^p), \tag{7.156}$$

and forms of different degrees are orthogonal. In an orthonormal frame this is just "sum of products of components, integrated over M". For functions it is $\int \alpha\beta$ vol; for 1-forms $A = A_i \mathrm{d}x^i$ it is $\int A_i B^i$ vol, the L^2 product of the corresponding vector fields. (For complex forms put a complex conjugate on α.) Completing gives the Hilbert space

$$\mathcal{H} = \bigoplus_{p=0}^{n} L^2\Omega^p(M).$$

Working definition 7.31 (the adjoint d^*)

$\mathrm{d}^*\colon \Omega^p \to \Omega^{p-1}$ is defined by $\langle \mathrm{d}^*\alpha, \beta \rangle = \langle \alpha, \mathrm{d}\beta \rangle$ for all β [every $(p-1)$-form β on M]. *Operational formula:*

$$\mathrm{d}^* = (-1)^{n(p+1)+1} \star \mathrm{d} \star \quad \text{on } p\text{-forms}, \qquad (\mathrm{d}^*\alpha)_{i_2 \dots i_p} = -\nabla^i \alpha_{i\, i_2 \dots i_p}. \tag{7.157}$$

[∇ is the Levi-Civita covariant derivative, and the index on ∇^i is raised with g] It is minus a divergence. For a 1-form A, $\mathrm{d}^*A = -\nabla_i A^i = -\operatorname{div} A$.

The minus sign is the one familiar from integration by parts. For $\alpha = A$ a 1-form and $\beta = \phi$ a function on flat $\mathbb{R}^n$,

$$\langle A, \mathrm{d}\phi \rangle \overset{(1)}{=} \int A_i\, \partial_i \phi\, \mathrm{d}^n x \overset{(2)}{=} -\int (\partial_i A_i)\, \phi\, \mathrm{d}^n x \overset{(3)}{=} \langle -\operatorname{div} A, \phi \rangle,$$

(1) The inner product (7.156) on 1-forms, with $\mathrm{d}\phi = \partial_i \phi\, \mathrm{d}x^i$.
(2) Integrate by parts; there is no boundary term on a closed manifold (or for functions that decay at infinity).
(3) Read off the function paired with ϕ.

So $\mathrm{d}^*A = -\operatorname{div} A$: d is the gradient and d^* is minus the divergence, which is the familiar statement that $-\operatorname{div}$ is the adjoint of grad.

7.11.2 Nash's construction

Now define [Nash p. 208, (7.68)]

$$Q_1 = \mathrm{d} + \mathrm{d}^*, \qquad Q_2 = \mathrm{i}(\mathrm{d} - \mathrm{d}^*), \qquad \mathcal{H}_1^B = \bigoplus_{p \geq 0} \Omega^{2p}(M), \qquad \mathcal{H}_1^F = \bigoplus_{p \geq 0} \Omega^{2p+1}(M). \tag{7.158}$$

A particle is a differential form, and the parity of its degree is its statistics: even forms are bosons, odd forms are fermions. So $(-1)^F$ acts on a p-form as $(-1)^p$. Both d and d^* change the degree by one, so both change the parity: Q_1, Q_2 are off-diagonal as they should be. Comparing with (7.145), the complex supercharge is

$$Q = \tfrac{1}{2}(Q_1 - \mathrm{i}Q_2) = \tfrac{1}{2}\big(\mathrm{d} + \mathrm{d}^* + \mathrm{d} - \mathrm{d}^*\big) = \mathrm{d}, \qquad Q^\dagger = \mathrm{d}^*.$$

The supersymmetry algebra follows from a single fact, $\mathrm{d}^2 = 0$, together with its adjoint $(\mathrm{d}^*)^2 = 0$. Here is the anticommutator [Nash p. 208, (7.69)]:

$$\begin{aligned} Q_1 Q_2 &\overset{(1)}{=} \mathrm{i}(\mathrm{d}+\mathrm{d}^*)(\mathrm{d}-\mathrm{d}^*) = \mathrm{i}\big(\mathrm{d}^2 + \mathrm{d}^*\mathrm{d} - \mathrm{d}\mathrm{d}^* - \mathrm{d}^{*2}\big) \\ &\overset{(2)}{=} \mathrm{i}(\mathrm{d}^*\mathrm{d} - \mathrm{d}\mathrm{d}^*), \\ Q_2 Q_1 &\overset{(3)}{=} \mathrm{i}(\mathrm{d}-\mathrm{d}^*)(\mathrm{d}+\mathrm{d}^*) = \mathrm{i}\big(\mathrm{d}^2 - \mathrm{d}^*\mathrm{d} + \mathrm{d}\mathrm{d}^* - \mathrm{d}^{*2}\big) = -\mathrm{i}(\mathrm{d}^*\mathrm{d} - \mathrm{d}\mathrm{d}^*) \\ &\overset{(4)}{=} -Q_1 Q_2. \end{aligned} \tag{7.159}$$

(1) Multiply out, keeping the order of the operators.
(2) $\mathrm{d}^2 = 0$ and $(\mathrm{d}^*)^2 = 0$ (the adjoint of $\mathrm{d}^2 = 0$).
(3) The same with the signs of d^* swapped.
(4) Compare with the second line.

So $\{Q_1, Q_2\} = 0$. The squares:

$$\begin{aligned} Q_1^2 &\overset{(1)}{=} \mathrm{d}^2 + \mathrm{d}\mathrm{d}^* + \mathrm{d}^*\mathrm{d} + \mathrm{d}^{*2} = \mathrm{d}\mathrm{d}^* + \mathrm{d}^*\mathrm{d}, \\ Q_2^2 &\overset{(2)}{=} -\big(\mathrm{d}^2 - \mathrm{d}\mathrm{d}^* - \mathrm{d}^*\mathrm{d} + \mathrm{d}^{*2}\big) = \mathrm{d}\mathrm{d}^* + \mathrm{d}^*\mathrm{d}, \\ H &\overset{(3)}{=} \tfrac{1}{2}(Q_1^2 + Q_2^2) = \mathrm{d}\mathrm{d}^* + \mathrm{d}^*\mathrm{d} = \bigoplus_{p \geq 0} \Delta_p. \end{aligned} \tag{7.160}$$

(1) Multiply out $(\mathrm{d}+\mathrm{d}^*)^2$ and drop $\mathrm{d}^2, \mathrm{d}^{*2}$.
(2) $\mathrm{i}^2 = -1$, then the same.
(3) Nash's (7.70): average the two; Δ_p is the Laplacian acting on p-forms. It preserves the degree because $\mathrm{d}\mathrm{d}^*$ and $\mathrm{d}^*\mathrm{d}$ each lower and raise it once.

Since $Q_1^2 = Q_2^2$, comparison with (7.143) gives $P = 0$ [Nash p. 209]: this is supersymmetric quantum mechanics, not a field theory. In Nash's words, a "non-relativistic model".

Working definition 7.32 (Hodge Laplacian)

$\Delta = \mathrm{d}\mathrm{d}^* + \mathrm{d}^*\mathrm{d}$ on forms. *Operational formulas.* On functions, $\Delta\phi = \mathrm{d}^*\mathrm{d}\phi = -\nabla^i\nabla_i\phi$, the *positive* Laplacian (minus the usual ∇^2). On forms in flat $\mathbb{R}^n$ with Cartesian coordinates, Δ acts on each component as $-\sum_i \partial_i^2$. On a curved manifold, $\Delta = \nabla^*\nabla + (\text{curvature})$ [$\nabla^*\nabla = -\nabla^i\nabla_i$ acting on every component index] (the Weitzenböck formula); on 1-forms the curvature term is the Ricci tensor [it acts as $A_i \mapsto R_i{}^j A_j$].

Worked example 7.68: Δ on 1-forms in the plane, by hand

Take $A = A_x\mathrm{d}x + A_y\mathrm{d}y$ on $\mathbb{R}^2$. Then

$$\mathrm{d}^*\mathrm{d}A \overset{(1)}{=} \mathrm{d}^*\big((\partial_x A_y - \partial_y A_x)\,\mathrm{d}x \wedge \mathrm{d}y\big) \overset{(2)}{=} -\partial_x B\,\mathrm{d}y + \partial_y B\,\mathrm{d}x, \qquad B := \partial_x A_y - \partial_y A_x,$$

$$\mathrm{d}\mathrm{d}^*A \overset{(3)}{=} \mathrm{d}\big(-\partial_x A_x - \partial_y A_y\big) = -\partial_x(\operatorname{div} A)\,\mathrm{d}x - \partial_y(\operatorname{div} A)\,\mathrm{d}y.$$

(1) $\mathrm{d}A$ is the "magnetic field" B times the area form.
(2) $\mathrm{d}^* = -\star\,\mathrm{d}\star$ in two dimensions; $\star(\mathrm{d}x \wedge \mathrm{d}y) = 1$, $\star\mathrm{d}x = \mathrm{d}y$, $\star\mathrm{d}y = -\mathrm{d}x$.
(3) $\mathrm{d}^*A = -\operatorname{div} A$ from (7.157).

Adding, the $\mathrm{d}x$ component is $\partial_y B - \partial_x \operatorname{div} A = \partial_y\partial_x A_y - \partial_y^2 A_x - \partial_x^2 A_x - \partial_x\partial_y A_y = -(\partial_x^2 + \partial_y^2)A_x$, and similarly for $\mathrm{d}y$. So $\Delta A = -(\nabla^2 A_x)\mathrm{d}x - (\nabla^2 A_y)\mathrm{d}y$: this is the vector identity $\operatorname{curl}\operatorname{curl} - \operatorname{grad}\operatorname{div} = -\nabla^2$ in disguise.

7.11.3 Zero-energy states are harmonic forms

By the positivity argument of Section 7.10.2, a form has zero energy exactly when both supercharges kill it. For forms this reads:

$$\begin{aligned}\langle \alpha, \Delta\alpha\rangle &\overset{(1)}{=} \langle \alpha, \mathrm{d}\mathrm{d}^*\alpha\rangle + \langle \alpha, \mathrm{d}^*\mathrm{d}\alpha\rangle \\ &\overset{(2)}{=} \langle \mathrm{d}^*\alpha, \mathrm{d}^*\alpha\rangle + \langle \mathrm{d}\alpha, \mathrm{d}\alpha\rangle = \|\mathrm{d}^*\alpha\|^2 + \|\mathrm{d}\alpha\|^2 .\end{aligned} \tag{7.161}$$

(1) Definition of Δ.
(2) Move one d or d^* to the left side of each inner product using the definition of the adjoint.

A sum of two squared norms vanishes only if both do. So

$$\Delta\alpha = 0 \iff \mathrm{d}\alpha = 0 \text{ and } \mathrm{d}^*\alpha = 0.$$

Such forms are called **harmonic**. The key theorem of Hodge relates them to cohomology.

Key idea: Hodge theorem, Nash (7.74)

On a compact oriented Riemannian manifold without boundary, every de Rham cohomology class contains exactly one harmonic form. Hence

$$b_p = \dim H^p_{\mathrm{dR}}(M) = \dim \operatorname{Ker} \Delta_p . \tag{7.162}$$

More precisely, $\Omega^p = \operatorname{Ker}\Delta_p \oplus \operatorname{Im}\mathrm{d} \oplus \operatorname{Im}\mathrm{d}^*$ (the **Hodge decomposition**), and the three pieces are orthogonal.

Here is the physicist's reason. Fix a closed form α and look at all forms $\alpha + \mathrm{d}\beta$ in its class. Among them, pick the one of *smallest norm* (an elliptic-PDE fact guarantees a minimiser exists; see Chapter 2). At the minimiser, varying $\beta \to \beta + \epsilon\gamma$ must not change the norm to first order:

$$0 = \frac{\mathrm{d}}{\mathrm{d}\epsilon}\Big|_{\epsilon=0} \|\alpha + \mathrm{d}\beta + \epsilon\,\mathrm{d}\gamma\|^2 \overset{(1)}{=} 2\,\langle \mathrm{d}\gamma,\ \alpha + \mathrm{d}\beta\rangle \overset{(2)}{=} 2\,\langle \gamma,\ \mathrm{d}^*(\alpha + \mathrm{d}\beta)\rangle .$$

(1) Expand the square to first order in ϵ.
(2) Definition of the adjoint.

Since γ is arbitrary, $\mathrm{d}^*(\alpha + \mathrm{d}\beta) = 0$. The minimiser is closed (it is in the class of a closed form) and co-closed, i.e. harmonic. It is unique: if ω and $\omega + \mathrm{d}\beta$ are both harmonic, then $\mathrm{d}\beta$ is harmonic, and $\|\mathrm{d}\beta\|^2 = \langle \beta, \mathrm{d}^*\mathrm{d}\beta\rangle = 0$. So: *the harmonic representative is the minimum-energy representative of the cohomology class*. For the Maxwell field this is familiar: among all vector potentials describing a given field, the one satisfying the Lorenz gauge condition $\mathrm{d}^*A = 0$ is picked out.

Key idea: the Witten index of forms is the Euler characteristic

With $(-1)^F = (-1)^p$ on p-forms,

$$\operatorname{Tr}(-1)^F \overset{(1)}{=} \sum_p (-1)^p \dim \operatorname{Ker}\Delta_p \overset{(2)}{=} \sum_p (-1)^p b_p \overset{(3)}{=} \chi(M). \tag{7.163}$$

(1) Definition (7.154): zero-energy states are harmonic forms, weighted by the parity of their degree.
(2) Hodge theorem (7.162).
(3) Definition of the Euler characteristic.

This is our first "physics computes topology" result. The Witten index is independent of the metric (a continuous deformation), just as the pairing argument predicts. In the language of Chapter 4, it is the index of the elliptic operator $\mathrm{d}+\mathrm{d}^*\colon \Omega^{\text{even}} \to \Omega^{\text{odd}}$.

The pairing, concretely. For forms the boson–fermion pairing of Section 7.10.6 is visible directly. Let α be an eigenform, $\Delta\alpha = \lambda\alpha$ with $\lambda > 0$, and suppose it is co-exact ($\alpha = \mathrm{d}^*\gamma$). Then $\mathrm{d}\alpha \neq 0$ (otherwise α would be harmonic), $\mathrm{d}\alpha$ has one degree more, and $\Delta\mathrm{d}\alpha = \mathrm{d}\Delta\alpha = \lambda\,\mathrm{d}\alpha$ because d commutes with Δ. So every positive eigenvalue appears in degrees p and $p+1$ together. The following examples show this in action.

Worked example 7.69: the spectrum of Δ on the circle

On S^1 with angle $\theta \in [0, 2\pi)$ and metric $\mathrm{d}\theta^2$, a 0-form is a periodic function and a 1-form is $\varphi(\theta)\mathrm{d}\theta$. Δ acts as $-\mathrm{d}^2/\mathrm{d}\theta^2$ on the coefficient in both degrees. Eigenforms are $e^{\mathrm{i}k\theta}$ and $e^{\mathrm{i}k\theta}\mathrm{d}\theta$, $k \in \mathbb{Z}$, with eigenvalue k^2. For $k \neq 0$,

$$\mathrm{d}\big(e^{\mathrm{i}k\theta}\big) = \mathrm{i}k\, e^{\mathrm{i}k\theta}\mathrm{d}\theta,$$

so d maps the bosonic state at level k^2 onto the fermionic one. At $k = 0$ we get 1 and $\mathrm{d}\theta$, both harmonic and unpaired: $b_0 = b_1 = 1$, and $\mathrm{Tr}(-1)^F = 1 - 1 = 0 = \chi(S^1)$.

Worked example 7.70: the spectrum on the flat torus

On $T^2 = \mathbb{R}^2/(2\pi\mathbb{Z})^2$ with coordinates (x, y), the eigenforms are $e^{\mathrm{i}(kx+ly)}$ times one of $1,\ \mathrm{d}x,\ \mathrm{d}y,\ \mathrm{d}x \wedge \mathrm{d}y$, with eigenvalue $k^2 + l^2$ (Δ acts on each Cartesian component as $-\partial_x^2 - \partial_y^2$, by Example 7.68). At each level with $(k, l) \neq (0, 0)$ there are 1 zero-form, 2 one-forms and 1 two-form per Fourier mode, so bosons minus fermions is $1 - 2 + 1 = 0$. At $(k, l) = (0, 0)$ we get the harmonic forms $1, \mathrm{d}x, \mathrm{d}y, \mathrm{d}x \wedge \mathrm{d}y$: $b = (1, 2, 1)$, $\chi(T^2) = 0$.

Worked example 7.71: the spectrum on the round sphere

On the unit S^2, the eigenfunctions of Δ on 0-forms are the spherical harmonics Y_{lm}, with eigenvalue $l(l+1)$ and multiplicity $2l+1$. Since Δ commutes with d and with $\star$, we also get

- 2-forms $Y_{lm}\,\mathrm{vol}$ (apply $\star$), eigenvalue $l(l+1)$, multiplicity $2l+1$;
- 1-forms $\mathrm{d}Y_{lm}$ (exact) and $\star\mathrm{d}Y_{lm}$ (co-exact), $l \geq 1$, eigenvalue $l(l+1)$, multiplicity $2(2l+1)$.

These exhaust the spectrum (Hodge decomposition, and $b_1(S^2) = 0$). Count at level $l(l+1)$ with $l \geq 1$:

$$n_B - n_F \overset{(1)}{=} \big[(2l+1) + (2l+1)\big] - 2(2l+1) = 0,$$

(1) Bosons are the 0-forms and 2-forms, fermions the 1-forms.

At $l = 0$: the constant function and vol, both bosons, both harmonic. So $b = (1, 0, 1)$ and $\mathrm{Tr}(-1)^F = 2 = \chi(S^2)$.

7.11.4 Forms are fermions: the exterior algebra as a Fock space

Nash says that there is a "formal equivalence between exterior and interior multiplication and Fermion creation and annihilation" [Nash p. 210]. We now make this precise, with matrices,

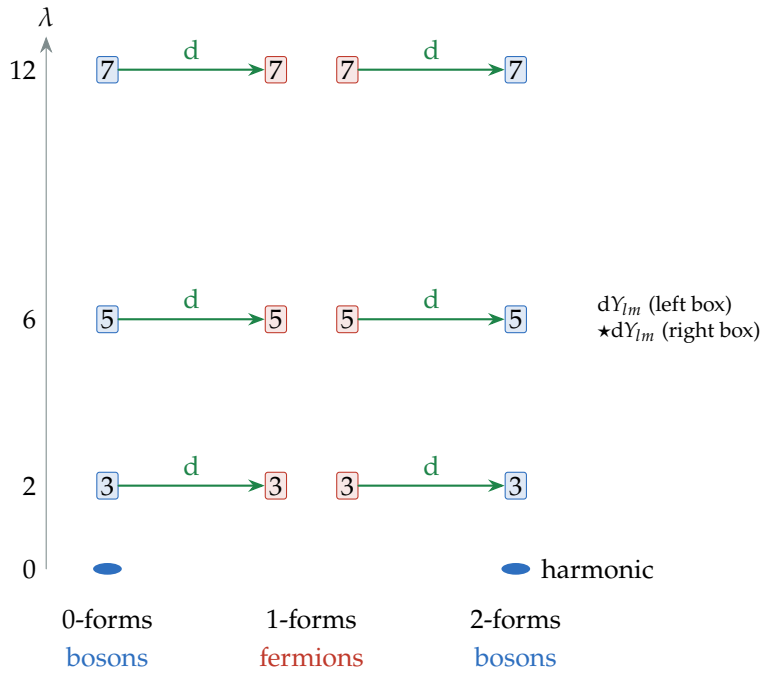


Figure 7.49: The spectrum of the Hodge Laplacian on the round S^2, sorted by degree. Each box is a level $l(l+1)$ with its multiplicity. The exterior derivative (green) maps 0-forms onto exact 1-forms and co-exact 1-forms onto 2-forms, preserving the eigenvalue. Every positive level cancels in $\mathrm{Tr}(-1)^F$; only the two harmonic forms at $\lambda = 0$ survive, and $\chi(S^2) = 2$.

because the Hamiltonian of Section 7.13 is written in this language.

At one point $x \in M$, pick an orthonormal basis $e_1, \dots, e_n$ of T_xM and the dual basis $e^1, \dots, e^n$ of T_x^*M (in flat space, $e_i = \partial_i$ and $e^i = \mathrm{d}x^i$). Nash writes e_i^* for e^i. Define two operators on forms at x:

$$a^{i*}\alpha := e^i \wedge \alpha \quad \text{(exterior multiplication)}, \qquad a^i\alpha := \iota_{e_i}\alpha \quad \text{(interior multiplication)}. \tag{7.164}$$

Here ι_v is contraction with v in the first slot, so $\iota_{e_i}e^j = \delta_i^j$ and $\iota_v(\alpha\wedge\beta) = (\iota_v\alpha)\wedge\beta + (-1)^p\alpha\wedge\iota_v\beta$ for α a p-form.

Key idea: canonical anticommutation relations

$$\{a^i, a^{j*}\} = \delta^{ij}, \qquad \{a^i, a^j\} = 0, \qquad \{a^{i*}, a^{j*}\} = 0. \tag{7.165}$$

These are the relations of n independent fermionic modes. The "vacuum" (no fermions) is the constant function 1; the state with fermions in modes $i_1 < \cdots < i_p$ is $e^{i_1}\wedge\cdots\wedge e^{i_p}$. The fermion number operator $N = \sum_i a^{i*}a^i$ is the *degree*: $N\alpha = p\,\alpha$ on p-forms.

The last two relations say $e^i\wedge e^j = -e^j\wedge e^i$ and $\iota_{e_i}\iota_{e_j} = -\iota_{e_j}\iota_{e_i}$. For the first, use the Leibniz rule for ι with e^j a 1-form:

$$a^i a^{j*}\alpha \overset{(1)}{=} \iota_{e_i}(e^j\wedge\alpha) \overset{(2)}{=} (\iota_{e_i}e^j)\,\alpha - e^j\wedge\iota_{e_i}\alpha \overset{(3)}{=} \delta_i^j\,\alpha - a^{j*}a^i\alpha.$$

(1) Definitions (7.164).
(2) The Leibniz rule for ι, with sign $(-1)^1$ because e^j has degree one.
(3) $\iota_{e_i}e^j = \delta_i^j$ for dual bases.

Moving the last term to the left gives $\{a^i, a^{j*}\} = \delta^{ij}$. And $N = \sum_i e^i\wedge\iota_{e_i}$ applied to $e^{i_1}\wedge\cdots\wedge e^{i_p}$ removes each of the p factors in turn and puts it back in front, returning p copies of the form.

Worked example 7.72: the Fock space of $\Lambda(\mathbb{R}^2)$ as 4×4 matrices

In the plane, order the basis of forms as $(1,\ \mathrm{d}x,\ \mathrm{d}y,\ \mathrm{d}x\wedge\mathrm{d}y)$. Reading off the images of the basis vectors column by column:

$$a^{1*}=\begin{pmatrix}0&0&0&0\\1&0&0&0\\0&0&0&0\\0&0&1&0\end{pmatrix},\qquad a^{2*}=\begin{pmatrix}0&0&0&0\\0&0&0&0\\1&0&0&0\\0&-1&0&0\end{pmatrix},\qquad a^1=\left(a^{1*}\right)^T,\quad a^2=\left(a^{2*}\right)^T.$$

For example, the second column of a^{2*} is $-\mathrm{d}x\wedge\mathrm{d}y$ because $\mathrm{d}y\wedge\mathrm{d}x=-\mathrm{d}x\wedge\mathrm{d}y$; this minus sign is the whole of fermionic statistics. The fermion number and the grading are

$$N=a^{1*}a^1+a^{2*}a^2=\mathrm{diag}(0,1,1,2),\qquad (-1)^F=\mathrm{diag}(1,-1,-1,1).$$

As a check of (7.165): a^1a^{2*} sends $\mathrm{d}x\mapsto\iota_x(-\mathrm{d}x\wedge\mathrm{d}y)=-\mathrm{d}y$ and $a^{2*}a^1$ sends $\mathrm{d}x\mapsto\mathrm{d}y\wedge1=\mathrm{d}y$; all other basis vectors are killed by both, so $\{a^1,a^{2*}\}=0$. Similarly $a^1a^{1*}+a^{1*}a^1=\mathrm{diag}(1,0,1,0)+\mathrm{diag}(0,1,0,1)=\mathbf{1}$.

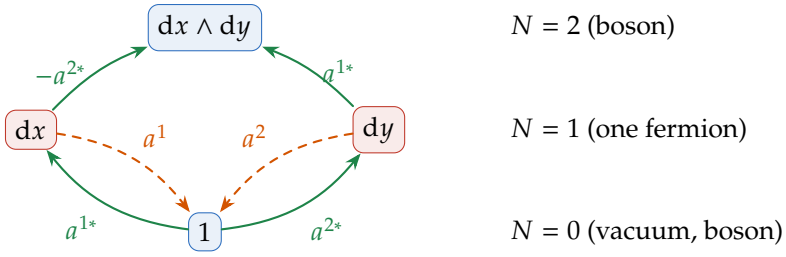


Figure 7.50: The exterior algebra of $\mathbb{R}^2$ as a fermionic Fock space. Exterior multiplication by $\mathrm{d}x^i$ (green) creates a fermion in mode i; interior multiplication by ∂_i (orange, dashed) destroys one. The degree is the fermion number. The minus sign on one path to $\mathrm{d}x\wedge\mathrm{d}y$ is the antisymmetry of fermionic states.

d and d^* as fermion bilinears. In flat space with Cartesian coordinates,

$$\mathrm{d}=\sum_i a^{i*}\,\partial_i,\qquad \mathrm{d}^*=-\sum_i a^i\,\partial_i. \tag{7.166}$$

The first is the definition $\mathrm{d}\alpha=\mathrm{d}x^i\wedge\partial_i\alpha$. The second is its adjoint: $(a^{i*})^\dagger=a^i$ (the matrices above are transposes) and $(\partial_i)^\dagger=-\partial_i$, and a^i commutes with ∂_i. On a curved manifold the same formulas hold with an orthonormal frame and ∂_i replaced by the Levi-Civita covariant derivative ∇_{e_i}. Compare with (7.157): on a 1-form, $-a^i\partial_i(A_j\mathrm{d}x^j)=-\partial_iA_i$, minus the divergence, as it should be.

Clifford operators and the Dirac operator. Two combinations of a and a^* are especially useful:

$$c^i:=a^{i*}-a^i,\qquad \hat{c}^{\,i}:=a^{i*}+a^i. \tag{7.167}$$

From (7.165),

$$\begin{aligned}\{c^i,c^j\}&\overset{(1)}{=}\{a^{i*},a^{j*}\}-\{a^{i*},a^j\}-\{a^i,a^{j*}\}+\{a^i,a^j\}=-2\delta^{ij},\\ \{\hat{c}^{\,i},\hat{c}^{\,j}\}&=+2\delta^{ij},\qquad \{c^i,\hat{c}^{\,j}\}\overset{(2)}{=}-\{a^{i*},a^j\}+\{a^i,a^{j*}\}=0.\end{aligned} \tag{7.168}$$

(1) Expand bilinearly; only the mixed anticommutators are nonzero, each equal to δ^{ij}.
(2) The two mixed terms enter with opposite signs and cancel.

So the c^i generate a Clifford algebra with $(c^i)^2 = -1$, the $\hat{c}^{\,i}$ generate a second Clifford algebra with $(\hat{c}^{\,i})^2 = +1$, and the two anticommute (Chapter T4). The forms $\Lambda(\mathbb{R}^n)$, of dimension 2^n, carry both at once. In flat space

$$Q_1 = \mathrm{d} + \mathrm{d}^* = \sum_i (a^{i*} - a^i)\,\partial_i = \sum_i c^i \partial_i, \tag{7.169}$$

a Dirac operator built with the "gamma matrices" c^i. Squaring it,

$$Q_1^2 = \sum_{i,j} c^i c^j \partial_i \partial_j = \tfrac{1}{2} \sum_{i,j} \{c^i, c^j\} \partial_i \partial_j = -\sum_i \partial_i^2 = \Delta,$$

because $\partial_i \partial_j$ is symmetric. This is the same mechanism by which the Dirac operator squares to the Laplacian. The operator $\mathrm{d} + \mathrm{d}^*$ is sometimes called the **signature** or **Euler** operator, or the Dirac–Kähler operator; in Chapter 4 its index is $\chi(M)$.

Dictionary: forms and fermions

Forms	Fermions
1 (a function)	Fock vacuum $\lvert 0\rangle$
$e^i\wedge$ (exterior mult.)	creation operator $a^{i*} = \psi^{i\dagger}$
ι_{e_i} (interior mult.)	annihilation operator $a^i = \psi^i$
degree p	fermion number F
$(-1)^p$	$(-1)^F$
$e^{i_1} \wedge \cdots \wedge e^{i_p}$	p-fermion state $\psi^{i_1\dagger} \cdots \psi^{i_p\dagger}\lvert 0\rangle$
d, d^*	supercharges Q, $Q^\dagger$
$\mathrm{d} + \mathrm{d}^*$	Dirac operator $c^i \nabla_i$

Reading Nash (pp. 208–209, (7.68)–(7.70), (7.74))

Nash writes the definitions (7.158) and remarks that "a particle is a p-form and the parity of p determines whether it is a Boson or a Fermion". His (7.69) is our (7.159); his (7.70), $H = \mathrm{dd}^* + \mathrm{d}^*\mathrm{d} = \bigoplus_p \Delta_p$, is (7.160); and his (7.74), $b_p = \dim \mathrm{Ker}\,\Delta_p$, is the Hodge theorem (7.162), which he quotes without comment. The fermionic interpretation of $e^i\wedge$ and ι_{e_i}, which Nash mentions on p. 210 as notation borrowed from Witten, is Section 7.11.4.

Recap

- Every compact Riemannian manifold carries a supersymmetric quantum mechanics: $\mathcal{H}$ = forms, $(-1)^F = (-1)^{\text{degree}}$, $Q = \mathrm{d}$, $Q^\dagger = \mathrm{d}^*$, $H = \Delta$.
- Zero-energy states = harmonic forms = cohomology classes (Hodge): $b_p = \dim \mathrm{Ker}\,\Delta_p$. The Witten index is $\chi(M)$.
- Exterior and interior multiplication are fermion creation and annihilation operators; $c^i = a^{i*} - a^i$ are Clifford generators, and $\mathrm{d} + \mathrm{d}^* = c^i \nabla_i$ is a Dirac operator.
- Nonzero eigenvalues of Δ are paired between adjacent degrees by d (circle, torus, sphere examples).

7.12 Witten's deformation of the exterior derivative

So far the manifold M carries a metric, and nothing else. The supersymmetric quantum mechanics of Section 7.11 knows the Betti numbers, but it does not know about any function f, so it cannot know about critical points. The question of this section is: *how do we put a Morse function into the system without spoiling the supersymmetry algebra, and without changing the number of zero-energy states?*

Witten's answer [Nash p. 209, (7.71)] is to conjugate d by the function e^{tf}. It is a one-line idea, but everything later rests on it, so we look at it from several sides: as a formula, as a change of basis that preserves cohomology, as a quantum-mechanical system, and through its exact zero-energy states.

7.12.1 The deformed differential

Let $f : M \to \mathbb{R}$ be smooth and $t \in \mathbb{R}$ a real parameter (later large and positive). Multiplication by the positive function e^{tf} is an invertible operator on forms, with inverse multiplication by e^{-tf}. Define [Nash p. 209, (7.71)]

$$\mathrm{d}_t := e^{-tf}\,\mathrm{d}\,e^{tf}. \tag{7.170}$$

Read it from right to left: multiply the form by e^{tf}, take d, multiply back by e^{-tf}. The result is again a first-order operator raising the degree by one. To see which one, apply it to a p-form α:

$$\begin{aligned} \mathrm{d}_t\alpha &\overset{(1)}{=} e^{-tf}\,\mathrm{d}\big(e^{tf}\alpha\big) \\ &\overset{(2)}{=} e^{-tf}\big(\mathrm{d}(e^{tf})\wedge\alpha + e^{tf}\,\mathrm{d}\alpha\big) \\ &\overset{(3)}{=} e^{-tf}\big(t\,e^{tf}\,\mathrm{d}f\wedge\alpha + e^{tf}\,\mathrm{d}\alpha\big) \\ &= \mathrm{d}\alpha + t\,\mathrm{d}f\wedge\alpha. \end{aligned} \tag{7.171}$$

(1) Definition (7.170).
(2) Leibniz rule for d of a function times a form: $\mathrm{d}(g\alpha) = \mathrm{d}g\wedge\alpha + g\,\mathrm{d}\alpha$.
(3) Chain rule, $\mathrm{d}(e^{tf}) = t\,e^{tf}\,\mathrm{d}f$.

So

$$\boxed{\mathrm{d}_t = \mathrm{d} + t\,\mathrm{d}f\wedge \;=\; \mathrm{d} + t\,\frac{\partial f}{\partial x^i}\,a^{i*}} \tag{7.172}$$

in the fermion notation of Section 7.11.4. The deformation adds a zeroth-order term: exterior multiplication by the 1-form $t\,\mathrm{d}f$.

Physical picture: an imaginary gauge field

On functions, $\mathrm{d}_t\phi = (\partial_i\phi + t\,\partial_i f\,\phi)\,\mathrm{d}x^i$. Compare the covariant derivative of a charged field, $D_i\phi = (\partial_i - \mathrm{i}eA_i)\phi$. The deformation is a covariant derivative with the "gauge field" $A = \mathrm{i}\,t\,\mathrm{d}f/e$: a pure gauge (it is exact), but *imaginary*. A real gauge transformation $e^{\mathrm{i}e\chi}$ is unitary and changes nothing physical. Our "gauge transformation" e^{tf} is real and not unitary, so it will change the spectrum, although, being pure gauge, it cannot change the topology. This is the whole trick in one sentence.

It still squares to zero. The conjugation formula makes this immediate:

$$\mathrm{d}_t^2 \overset{(1)}{=} e^{-tf}\,\mathrm{d}\,e^{tf}\,e^{-tf}\,\mathrm{d}\,e^{tf} \overset{(2)}{=} e^{-tf}\,\mathrm{d}^2\,e^{tf} \overset{(3)}{=} 0.$$

(1) Definition, twice.
(2) The inner factors $e^{tf}e^{-tf}$ cancel.
(3) $\mathrm{d}^2 = 0$.

It is instructive to check it from (7.171) as well, since there the cancellation is less obvious: $\mathrm{d}_t^2\alpha = \mathrm{d}(\mathrm{d}\alpha + t\,\mathrm{d}f\wedge\alpha) + t\,\mathrm{d}f\wedge(\mathrm{d}\alpha + t\,\mathrm{d}f\wedge\alpha)$. The first bracket gives $\mathrm{d}^2\alpha + t\,\mathrm{d}^2 f\wedge\alpha - t\,\mathrm{d}f\wedge\mathrm{d}\alpha$ (the sign from moving d past the 1-form $\mathrm{d}f$), the second gives $t\,\mathrm{d}f\wedge\mathrm{d}\alpha + t^2\,\mathrm{d}f\wedge\mathrm{d}f\wedge\alpha$. Everything cancels: $\mathrm{d}^2 = 0$, $\mathrm{d}^2 f = 0$, the two $\mathrm{d}f\wedge\mathrm{d}\alpha$ terms have opposite signs, and $\mathrm{d}f\wedge\mathrm{d}f = 0$.

7.12.2 The adjoint and the new supercharges

The adjoint of a product is the product of the adjoints in reverse order. Multiplication by a real function is self-adjoint for the L^2 inner product (7.156). Hence

$$\mathrm{d}_t^* \overset{(1)}{=} \left(e^{-tf}\,\mathrm{d}\,e^{tf}\right)^* \overset{(2)}{=} (e^{tf})^*\,\mathrm{d}^*\,(e^{-tf})^* \overset{(3)}{=} e^{tf}\,\mathrm{d}^*\,e^{-tf}, \tag{7.173}$$

(1) Definition (7.170).
(2) $(ABC)^* = C^*B^*A^*$.
(3) $e^{\pm tf}$ is real, so multiplication by it is self-adjoint.

which is Nash's formula for d_t^* [Nash p. 209]. Note the exponents have swapped sign. Explicitly, the adjoint of exterior multiplication by a 1-form ω is interior multiplication by the dual vector $\omega^\sharp$ (in an orthonormal frame, $(a^{i*})^\dagger = a^i$), so

$$\mathrm{d}_t^* = \mathrm{d}^* + t\,\iota_{\nabla f} = \mathrm{d}^* + t\,\frac{\partial f}{\partial x^i}\,a^i, \tag{7.174}$$

where $\nabla f = g^{ij}\partial_j f\,\partial_i$ is the gradient vector field. (One can also derive this from (7.173) using $\mathrm{d}^* = -\iota_{e_i}\nabla_{e_i}$, exactly as we derived (7.171): the derivative hits e^{-tf} and produces $-(-t\,\partial_i f)\,a^i$.)

Now copy (7.158) with d replaced by d_t [Nash p. 209, (7.72)]:

$$Q_1(t) = \mathrm{d}_t + \mathrm{d}_t^*, \qquad Q_2(t) = \mathrm{i}\big(\mathrm{d}_t - \mathrm{d}_t^*\big). \tag{7.175}$$

Nash calls it "a routine matter to verify" that the algebra is unchanged. The verification in (7.159)–(7.160) used *only* two facts: $\mathrm{d}^2 = 0$, and d^* is the adjoint of d. Both hold for d_t: we just showed $\mathrm{d}_t^2 = 0$, and d_t^* is its adjoint by construction, so $(\mathrm{d}_t^*)^2 = (\mathrm{d}_t^2)^* = 0$. Therefore, with no further work,

$$\{Q_1(t), Q_2(t)\} = 0, \qquad Q_1(t)^2 = Q_2(t)^2 = \mathrm{d}_t\mathrm{d}_t^* + \mathrm{d}_t^*\mathrm{d}_t := H_t = \bigoplus_p \Delta_p(t) \tag{7.176}$$

[Nash p. 209, (7.73)]. Each $Q_i(t)$ is Hermitian, anticommutes with $(-1)^F$ (it changes the degree by ± 1) and commutes with H_t. So for every value of t we have a supersymmetric quantum mechanics with complex supercharge $Q = \mathrm{d}_t$, and all the conclusions of Section 7.10 apply to it: $H_t \geq 0$; $H_t\alpha = 0$ iff $\mathrm{d}_t\alpha = \mathrm{d}_t^*\alpha = 0$; positive energy levels pair up; the Witten index is independent of t.

7.12.3 Why the zero-energy states are still counted by the Betti numbers

Nash says that "conjugation by e^{ft} means that the spectrum of H_t depends on t", but "the null spaces of H_t and H clearly coincide", so that [Nash p. 209, (7.75)]

$$b_p = \dim\operatorname{Ker}\Delta_p(t) \qquad \text{for every } t. \tag{7.177}$$

Let us see exactly why, because the word "coincide" hides a subtlety.

Step 1: e^{tf} is a chain isomorphism. Rewrite the definition (7.170) as

$$\mathrm{d} \circ e^{tf} = e^{tf} \circ \mathrm{d}_t. \tag{7.178}$$

So multiplication by e^{tf} carries d_t into d. It maps d_t-closed forms to d-closed forms ($\mathrm{d}_t\alpha = 0 \Rightarrow \mathrm{d}(e^{tf}\alpha) = e^{tf}\mathrm{d}_t\alpha = 0$) and d_t-exact forms to d-exact forms ($\alpha = \mathrm{d}_t\beta \Rightarrow e^{tf}\alpha = \mathrm{d}(e^{tf}\beta)$), and its inverse e^{-tf} does the reverse. Hence it induces an isomorphism

$$H^p(\Omega^*(M), \mathrm{d}_t) = \frac{\operatorname{Ker} \mathrm{d}_t}{\operatorname{Im} \mathrm{d}_t} \xrightarrow[\cong]{e^{tf}} \frac{\operatorname{Ker} \mathrm{d}}{\operatorname{Im} \mathrm{d}} = H^p_{\mathrm{dR}}(M). \tag{7.179}$$

Step 2: Hodge theory for d_t. The argument of Section 7.11.3 used only $\mathrm{d}^2 = 0$, the adjoint, and ellipticity of the Laplacian (which is unaffected by adding lower-order terms). So it works verbatim for d_t: every d_t-cohomology class contains exactly one form killed by $\Delta(t)$, the one of minimal norm. Hence

$$\dim \operatorname{Ker} \Delta_p(t) \overset{(1)}{=} \dim H^p(\Omega^*, \mathrm{d}_t) \overset{(2)}{=} \dim H^p_{\mathrm{dR}}(M) \overset{(3)}{=} b_p.$$

(1) Hodge theorem for the complex (Ω^*, d_t).
(2) The chain isomorphism (7.179).
(3) Definition of the Betti number.

$$\begin{array}{ccccccccc}
\cdots & \xrightarrow{\mathrm{d}_t} & \Omega^{p-1} & \xrightarrow{\mathrm{d}_t} & \Omega^{p} & \xrightarrow{\mathrm{d}_t} & \Omega^{p+1} & \xrightarrow{\mathrm{d}_t} & \cdots \\
 & & \downarrow{\scriptstyle e^{tf}\cong} & & \downarrow{\scriptstyle e^{tf}\cong} & & \downarrow{\scriptstyle e^{tf}\cong} & & \\
\cdots & \xrightarrow{\mathrm{d}} & \Omega^{p-1} & \xrightarrow{\mathrm{d}} & \Omega^{p} & \xrightarrow{\mathrm{d}} & \Omega^{p+1} & \xrightarrow{\mathrm{d}} & \cdots
\end{array}$$

Figure 7.51: Multiplication by e^{tf} (vertical arrows) intertwines the deformed complex (top row) with the de Rham complex (bottom row): every square commutes, by (7.178). So the two rows have the same cohomology, degree by degree, for every t. The vertical maps are not unitary, so the Laplacians built from the two rows have different spectra.

Watch out: "the null spaces coincide" means "have the same dimension"

The harmonic forms of Δ and of $\Delta(t)$ are *different forms*. For example, on functions $\operatorname{Ker} \Delta_0$ is spanned by the constant 1, whereas $\operatorname{Ker} \Delta_0(t)$ is spanned by e^{-tf} (see below). They are related by e^{tf}, but not literally equal, and H_t is not unitarily equivalent to H: *its positive eigenvalues move with t*. That is the point of the deformation. Only the number of zero eigenvalues in each degree is frozen.

7.12.4 The exact zero modes in degree 0 and n

Two zero-energy states of H_t can be written down in closed form, for every f and every t. They are our first look at localisation.

Degree 0. On functions $\mathrm{d}_t^* = 0$ (there are no (-1)-forms), so $H_t\phi = \mathrm{d}_t^*\mathrm{d}_t\phi$, and ϕ is a zero mode iff $\mathrm{d}_t\phi = 0$. Now

$$\mathrm{d}_t\left(e^{-tf}\right) \overset{(1)}{=} e^{-tf}\,\mathrm{d}\left(e^{tf}e^{-tf}\right) \overset{(2)}{=} e^{-tf}\,\mathrm{d}(1) \overset{(3)}{=} 0.$$

(1) Definition (7.170).

(2) The exponentials cancel.
(3) A constant has zero derivative.

So $\phi_0 = e^{-tf}$ spans $\operatorname{Ker}\Delta_0(t)$ on a connected M (it is the image of $1 \in \operatorname{Ker}\Delta_0$ under e^{-tf}).

Degree n. On top forms $\mathrm{d}_t = 0$, so a zero mode needs only $\mathrm{d}_t^*\omega = 0$. By (7.173), $\mathrm{d}_t^*(e^{tf}\,\mathrm{vol}) = e^{tf}\,\mathrm{d}^*(e^{-tf}e^{tf}\mathrm{vol}) = e^{tf}\,\mathrm{d}^*\mathrm{vol} = 0$, because the volume form is harmonic. So $\omega_n = e^{tf}\,\mathrm{vol}$ spans $\operatorname{Ker}\Delta_n(t)$ (for M connected and oriented).

Key idea: localisation of the exact zero modes

As $t \to +\infty$,

$$|\phi_0|^2 = e^{-2tf} \text{ piles up at the global minimum of } f,$$
$$|\omega_n|^2 = e^{2tf} \text{ piles up at the global maximum of } f.$$

A minimum has index 0 and a maximum has index n. So already the two explicit zero modes live at critical points whose index equals their degree. This is the pattern we will establish in general in Section 7.14.

Worked example 7.73: the circle with $f = \cos\theta$

On S^1 with $f = \cos\theta$ there is a maximum at $\theta = 0$ ($f = 1$) and a minimum at $\theta = \pi$ ($f = -1$). A state is a pair $(\phi(\theta), \psi(\theta)\,\mathrm{d}\theta)$. From (7.171) and (7.174), with $f' = -\sin\theta$,

$$\mathrm{d}_t\phi \overset{(1)}{=} (\phi' - t\sin\theta\,\phi)\,\mathrm{d}\theta,$$
$$\mathrm{d}_t^*(\psi\,\mathrm{d}\theta) \overset{(2)}{=} -\psi' - t\sin\theta\,\psi.$$

(1) $\mathrm{d}\phi = \phi'\mathrm{d}\theta$ and $t\,\mathrm{d}f\,\phi = -t\sin\theta\,\phi\,\mathrm{d}\theta$.
(2) $\mathrm{d}^*(\psi\mathrm{d}\theta) = -\psi'$ (minus the divergence), and $t\,\iota_{\nabla f}(\psi\mathrm{d}\theta) = tf'\psi = -t\sin\theta\,\psi$.

The zero modes solve first-order ODEs: $\phi' = t\sin\theta\,\phi$ gives $\phi_0 = e^{-t\cos\theta}$, and $\psi' = -t\sin\theta\,\psi$ gives $\psi_1 = e^{t\cos\theta}$. Both are smooth and periodic, so $\dim\operatorname{Ker}\Delta_0(t) = \dim\operatorname{Ker}\Delta_1(t) = 1 = b_0 = b_1$, as (7.177) demands. At $t = 0$ they are the constants 1 and $\mathrm{d}\theta$ of Example 7.69. As t grows the 0-form concentrates at the minimum $\theta = \pi$ and the 1-form at the maximum $\theta = 0$ (Figure 7.52). The normalised density is

$$\frac{e^{-2t\cos\theta}}{\int_0^{2\pi} e^{-2t\cos\theta}\,\mathrm{d}\theta} = \frac{e^{-2t\cos\theta}}{2\pi I_0(2t)} \approx \sqrt{\frac{t}{\pi}}\,e^{-t(\theta-\pi)^2} \quad (t \gg 1),$$

where I_0 is a modified Bessel function; the last form uses $\cos\theta \approx -1 + \tfrac{1}{2}(\theta-\pi)^2$ and $\int e^{-tu^2}\mathrm{d}u = \sqrt{\pi/t}$. The width of the peak is $1/\sqrt{t}$: this $t^{-1/2}$ localisation length will be a constant companion.

7.12.5 The line: we have seen this system before

Take $M = \mathbb{R}$ with coordinate x (not compact, but it is the local model of everything). A state is $(\phi(x), \psi(x)\,\mathrm{d}x)$ and

$$\mathrm{d}_t\phi = (\phi' + tf'\phi)\mathrm{d}x, \qquad \mathrm{d}_t^*(\psi\,\mathrm{d}x) = -\psi' + tf'\psi.$$

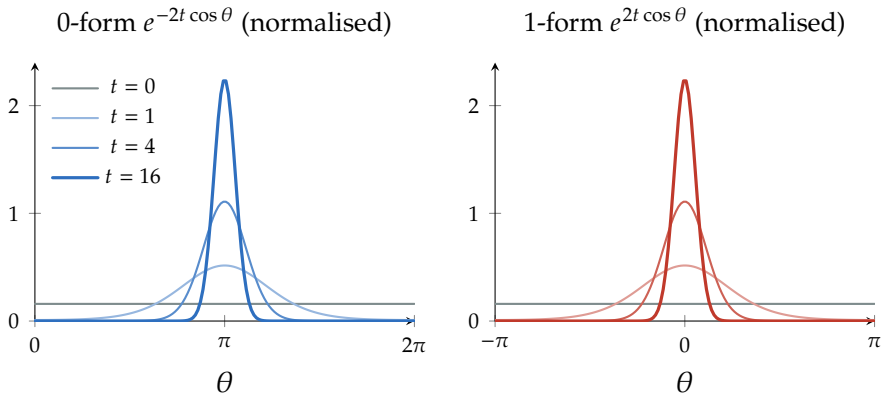


Figure 7.52: The two exact zero-energy states of H_t on the circle with $f = \cos\theta$, as probability densities, for $t = 0, 1, 4, 16$. The bosonic 0-form (left) concentrates at the minimum $\theta = \pi$ of f; the fermionic 1-form (right, plotted on $[-\pi, \pi]$) at the maximum $\theta = 0$. The peak height grows like $\sqrt{t/\pi}$ and the width shrinks like $t^{-1/2}$.

Compare with (7.150): these are exactly $A\phi$ and $A^\dagger\psi$ with superpotential

$$W = t\,f. \tag{7.180}$$

So the supersymmetric quantum mechanics on a line of Section 7.10.5 *is* Witten's deformed de Rham complex of the real line. In particular, reading off (7.152)–(7.153),

$$H_t = \begin{pmatrix} -\dfrac{\mathrm{d}^2}{\mathrm{d}x^2} + t^2 f'^2 - t f'' & 0 \\ 0 & -\dfrac{\mathrm{d}^2}{\mathrm{d}x^2} + t^2 f'^2 + t f'' \end{pmatrix} \quad \text{on} \quad \begin{pmatrix} \text{0-forms} \\ \text{1-forms} \end{pmatrix}. \tag{7.181}$$

The potential $t^2 f'^2$ is large except at the critical points; the "spin" term $\mp t f''$ decides which critical points are good for 0-forms (minima, where $f'' > 0$) and which for 1-forms (maxima). The general version of this formula, on any M, is the subject of Section 7.13.

Worked example 7.74: the double well on the line

Take $f = \frac{1}{4}(x^2 - 1)^2$, with minima at $x = \pm 1$ ($f = 0$, $f'' = 2$) and a maximum at $x = 0$ ($f = \frac{1}{4}$, $f'' = -1$). The zero-mode equations give

$$\phi_0 = e^{-tf} = e^{-t(x^2-1)^2/4} \quad \text{(normalisable)}, \qquad \psi_1 = e^{+tf} \quad \text{(not normalisable)}.$$

So there is one bosonic zero-energy state and no fermionic one. The boson has two bumps, one at each minimum. Naive counting ("one state per critical point, in degree equal to the index") predicts *three* low-energy states: two 0-forms at $x = \pm 1$ and one 1-form at $x = 0$. The exact answer has one. The other two form a boson–fermion pair at a small positive energy: numerically, at $t = 16$ the three lowest 0-form levels are 0, 4.6×10^{-3}, 27.9, and the 1-form sector has a matching level 4.6×10^{-3}. We will compute that small energy from tunnelling in Section 7.15.

7.12.6 Reversing the function and Poincaré duality

Replacing f by $-f$ swaps minima and maxima, and replaces the index λ of each critical point by $n - \lambda$. On the quantum side, the Hodge star swaps d_t with d^*_{-t}, up to sign.

Try it: the Hodge star intertwines H_t and H_{-t}

Using $\mathrm{d}^* = \pm \star \mathrm{d} \star$ and the fact that $\star$ commutes with multiplication by functions, show that $\star \mathrm{d}_t^* = \pm \mathrm{d}_{-t} \star$ and hence $\star \Delta_p(t) = \Delta_{n-p}(-t) \star$. Conclude that $\dim \operatorname{Ker} \Delta_p(t) = \dim \operatorname{Ker} \Delta_{n-p}(-t)$, i.e. $b_p = b_{n-p}$.
Answer. $\star \mathrm{d}_t^* = \star e^{tf} \mathrm{d}^* e^{-tf} = \pm e^{tf} \star \star \mathrm{d} \star e^{-tf} = \pm e^{tf} \mathrm{d} e^{-tf} \star = \pm \mathrm{d}_{-t} \star$ (using $\star\star = \pm 1$). Taking adjoints gives the companion relation, and $\Delta(t) = \mathrm{d}_t \mathrm{d}_t^* + \mathrm{d}_t^* \mathrm{d}_t$ goes over to $\mathrm{d}_{-t}^* \mathrm{d}_{-t} + \mathrm{d}_{-t} \mathrm{d}_{-t}^*$. This is Poincaré duality, seen by Witten's Hamiltonian. In the circle example it maps $e^{-t\cos\theta}$ to $e^{-t\cos\theta} \mathrm{d}\theta$, the 1-form zero mode for $-f$.

Reading Nash (p. 209, (7.71)–(7.75))

Nash introduces $\mathrm{d}_t = e^{-ft} \mathrm{d} e^{ft}$, states $\mathrm{d}_t^* = e^{ft} \mathrm{d}^* e^{-ft}$, defines $Q_1(t)$, $Q_2(t)$ and H_t, and says the algebra is unchanged "as a routine matter"; the routine is that $\mathrm{d}_t^2 = 0$ and d_t^* is its adjoint. His (7.75), $b_p = \dim \operatorname{Ker} \Delta_p(t)$, rests on the chain isomorphism (7.179) plus Hodge theory for d_t. "Coincide" should be read as "are isomorphic via e^{tf}". Nash writes e^{-ft} where we write e^{-tf}; it is the same thing.

Recap

- $\mathrm{d}_t = e^{-tf} \mathrm{d} e^{tf} = \mathrm{d} + t\, \mathrm{d}f\wedge$ and $\mathrm{d}_t^* = e^{tf} \mathrm{d}^* e^{-tf} = \mathrm{d}^* + t\, \iota_{\nabla f}$. Still $\mathrm{d}_t^2 = 0$, so the supersymmetry algebra holds for every t, with $H_t = \mathrm{d}_t \mathrm{d}_t^* + \mathrm{d}_t^* \mathrm{d}_t$.
- e^{tf} is a chain isomorphism, so $\dim \operatorname{Ker} \Delta_p(t) = b_p$ for all t; but e^{tf} is not unitary, so the positive spectrum moves.
- Exact zero modes e^{-tf} (degree 0) and $e^{tf}\,\mathrm{vol}$ (degree n) localise at the minimum and maximum of f, with width $t^{-1/2}$.
- On the line, d_t is supersymmetric quantum mechanics with superpotential $W = tf$.

7.13 The deformed Laplacian: a Schrödinger operator with spin

The deformation leaves the number of zero modes alone, but it changes the Hamiltonian. The question of this section is: *what does H_t look like as a differential operator?* The answer [Nash p. 210, (7.76)] is that it is the ordinary Laplacian plus a potential. The potential has a bosonic part $t^2 |\mathrm{d}f|^2$, which is large except at the critical points, and a "spin" part, linear in t, built from the Hessian of f and the fermion operators. Everything in the rest of the chapter is read off from this formula.

7.13.1 Expanding the product

Write $\varepsilon := \mathrm{d}f\wedge$ (exterior multiplication by $\mathrm{d}f$) and $\iota := \iota_{\nabla f}$ (its adjoint). Then $\mathrm{d}_t = \mathrm{d} + t\varepsilon$ and $\mathrm{d}_t^* = \mathrm{d}^* + t\iota$ by (7.172)–(7.174), and

$$\begin{aligned} H_t &\overset{(1)}{=} (\mathrm{d} + t\varepsilon)(\mathrm{d}^* + t\iota) + (\mathrm{d}^* + t\iota)(\mathrm{d} + t\varepsilon) \\ &\overset{(2)}{=} \big(\mathrm{d}\mathrm{d}^* + \mathrm{d}^*\mathrm{d}\big) + t\big(\{\mathrm{d}, \iota\} + \{\varepsilon, \mathrm{d}^*\}\big) + t^2\{\varepsilon, \iota\} \\ &\overset{(3)}{=} \Delta + t\big(\mathcal{L}_{\nabla f} + \mathcal{L}_{\nabla f}^*\big) + t^2 |\mathrm{d}f|^2 . \end{aligned} \tag{7.182}$$

(1) Definition (7.176) with the explicit forms of d_t and d_t^*.

(2) Multiply out and group by powers of t. The expansion stops at t^2 because d_t is only linear in t.
(3) Three facts, derived below: Cartan's formula $\{\mathrm{d}, \iota_X\} = \mathcal{L}_X$; its adjoint $\{\varepsilon, \mathrm{d}^*\} = \{\iota, \mathrm{d}\}^* = \mathcal{L}_X^*$; and $\{\varepsilon, \iota\} = |\mathrm{d}f|^2$.

The t^2 term: a potential. In an orthonormal frame $\varepsilon = \partial_i f\, a^{i*}$ and $\iota = \partial_j f\, a^j$, so

$$\{\varepsilon, \iota\} \overset{(1)}{=} \partial_i f\, \partial_j f\, \{a^{i*}, a^j\} \overset{(2)}{=} \partial_i f\, \partial_j f\, \delta^{ij} = |\mathrm{d}f|^2 = g^{ij}\partial_i f\, \partial_j f.$$

(1) The coefficients are functions, which commute with the constant matrices a^{i*}, a^j; pull them out.
(2) Canonical anticommutation relations (7.165).

It is multiplication by a function, the same in every degree. This is Nash's $t^2 g^{ij}\, \partial_i f\, \partial_j f$, the "(grad f)^2" of (7.77).

The t term: first derivatives cancel. A priori $\{\mathrm{d}, \iota\}$ is a first-order differential operator: Cartan's formula $\mathrm{d}\iota_X + \iota_X \mathrm{d} = \mathcal{L}_X$ says it is the Lie derivative along the gradient flow, which differentiates. So is its adjoint. But the derivative parts cancel in the sum. On functions, where everything is explicit,

$$\begin{aligned}(\mathcal{L}_X + \mathcal{L}_X^*)\phi &\overset{(1)}{=} X^i\partial_i\phi - \frac{1}{\sqrt{g}}\partial_i\big(\sqrt{g}\, X^i\phi\big)\\ &\overset{(2)}{=} X^i\partial_i\phi - X^i\partial_i\phi - (\operatorname{div} X)\,\phi\\ &\overset{(3)}{=} -(\nabla^2 f)\,\phi.\end{aligned} \tag{7.183}$$

(1) $\mathcal{L}_X\phi = X(\phi)$; the adjoint of $X^i\partial_i$ for the measure $\sqrt{g}\,\mathrm{d}^n x$ is $-\frac{1}{\sqrt{g}}\partial_i\sqrt{g}X^i$ (integrate by parts).
(2) Product rule; $\operatorname{div} X = \frac{1}{\sqrt{g}}\partial_i(\sqrt{g}X^i)$.
(3) $X = \nabla f$, so $\operatorname{div} X = \nabla^2 f$, the ordinary Laplacian of f.

A pure multiplication operator: the derivatives cancelled. This is no accident. A first derivative $X^i\partial_i$ is anti-self-adjoint up to terms without derivatives (think of ∂_x, whose adjoint is $-\partial_x$), so adding it to its own adjoint always kills the derivative part. What survives is built from the first derivatives of X, and for $X = \nabla f$ these are the second derivatives of f.

In every degree the cleanest route is the fermion calculus. In flat space with Cartesian coordinates, using $\mathrm{d} = a^{k*}\partial_k$ and $\mathrm{d}^* = -a^k\partial_k$ from (7.166), and writing $f_j = \partial_j f$, $f_{kj} = \partial_k\partial_j f$:

$$\begin{aligned}\{\mathrm{d}, \iota\} &\overset{(1)}{=} a^{k*}\partial_k\big(f_j a^j\,\cdot\big) + f_j a^j\, a^{k*}\partial_k\\ &\overset{(2)}{=} f_{kj}\, a^{k*}a^j + f_j\big(a^{k*}a^j + a^j a^{k*}\big)\partial_k \overset{(3)}{=} f_{kj}\, a^{k*}a^j + f_k\,\partial_k,\\ \{\varepsilon, \mathrm{d}^*\} &\overset{(4)}{=} -f_i a^{i*}a^k\partial_k - a^k\partial_k\big(f_i a^{i*}\,\cdot\big)\\ &\overset{(5)}{=} -f_i\big(a^{i*}a^k + a^k a^{i*}\big)\partial_k - f_{ki}\, a^k a^{i*} \overset{(6)}{=} -f_k\,\partial_k - f_{kj}\, a^j a^{k*}.\end{aligned} \tag{7.184}$$

(1) Write out the anticommutator; the dot marks where the form being acted on goes.
(2) Product rule: ∂_k hits either the coefficient f_j (giving f_{kj}) or the form.
(3) $\{a^{k*}, a^j\} = \delta^{kj}$.
(4) The same for $\{\varepsilon, \mathrm{d}^*\}$ with $\mathrm{d}^* = -a^k\partial_k$.

(5) Product rule again.
(6) $\{a^{i*}, a^k\} = \delta^{ik}$, and rename $i \to j$ in the last term using $f_{kj} = f_{jk}$.

Adding, the first-derivative terms $\pm f_k \partial_k$ cancel, and

$$\{\mathrm{d}, \iota\} + \{\varepsilon, \mathrm{d}^*\} = f_{kj}\big(a^{k*}a^j - a^j a^{k*}\big) = \frac{\partial^2 f}{\partial x^k \partial x^j}\,\big[a^{k*}, a^j\big]. \tag{7.185}$$

On functions $[a^{k*}, a^j]1 = -a^j a^{k*}1 = -\delta^{jk}$, so this reproduces $-\nabla^2 f$, in agreement with (7.183). On a curved manifold the same computation, done in an orthonormal frame with covariant derivatives, replaces f_{kj} by the **covariant Hessian** $\nabla_k \nabla_j f = \partial_k \partial_j f - \Gamma^l_{kj} \partial_l f$, which Nash writes $D^2 f / Dx^k Dx^j$. (At a critical point the Christoffel term vanishes, so the covariant Hessian is the ordinary one of Chapter 7.)

Key idea: Witten's Hamiltonian, Nash (7.76)

$$H_t = \mathrm{d}_t \mathrm{d}_t^* + \mathrm{d}_t^* \mathrm{d}_t = \Delta \;+\; t^2\, g^{ij} \frac{\partial f}{\partial x^i}\frac{\partial f}{\partial x^j} \;+\; t\, \frac{D^2 f}{Dx^i Dx^j}\,\big[a^{i*}, a^j\big] \;\coloneqq\; \Delta + V. \tag{7.186}$$

It is a Schrödinger operator (Laplacian plus potential) acting on multi-component wavefunctions. The potential V is a matrix in the fermion indices: its t^2 part is a scalar, the same for every degree; its t part couples the Hessian of f to the fermions, like a spin–orbit or Zeeman term.

Watch out: H_t is not a conjugate of Δ

Nash's (7.76) begins "$H_t = \mathrm{d}_t \mathrm{d}_t^* + \mathrm{d}_t^* \mathrm{d}_t = e^{-ft}(\mathrm{dd}^* + \mathrm{d}^*\mathrm{d})e^{ft}$". The second equality is not correct as written. We have $\mathrm{d}_t \mathrm{d}_t^* = e^{-tf}\,\mathrm{d}\,e^{tf}\,e^{tf}\mathrm{d}^* e^{-tf} = e^{-tf}\,\mathrm{d}\,e^{2tf}\,\mathrm{d}^* e^{-tf}$: the two conjugations point in opposite directions, so the exponentials do not cancel in the middle. Were H_t a conjugate of Δ, it would have the same spectrum as Δ, and the whole method would be empty. The correct statement is that the complex (Ω^*, d_t) is conjugate to (Ω^*, d); the Laplacians are not. What *is* true is the rest of Nash's line, the explicit formula (7.186).

7.13.2 The same Hamiltonian in Clifford form

The derivation is shorter, and more illuminating, with the Clifford operators of (7.167): $c^i = a^{i*} - a^i$ and $\hat{c}^{\,i} = a^{i*} + a^i$, with $\{c^i, c^j\} = -2\delta^{ij}$, $\{\hat{c}^{\,i}, \hat{c}^{\,j}\} = 2\delta^{ij}$, $\{c^i, \hat{c}^{\,j}\} = 0$. The supercharge is

$$Q_1(t) \overset{(1)}{=} \big(a^{i*} - a^i\big)\partial_i + t\, f_i \big(a^{i*} + a^i\big) \overset{(2)}{=} c^i \partial_i + t\, f_i\, \hat{c}^{\,i}. \tag{7.187}$$

(1) $\mathrm{d}_t + \mathrm{d}_t^*$ with $\mathrm{d} = a^{i*}\partial_i$, $\mathrm{d}^* = -a^i \partial_i$, $\varepsilon = f_i a^{i*}$, $\iota = f_i a^i$.
(2) Definitions of c^i and $\hat{c}^{\,i}$.

So the undeformed part is a Dirac operator built with the c's, and the deformation is a "mass term" $t\,\mathrm{d}f$ built with the *other* Clifford algebra, the $\hat{c}$'s. Squaring:

$$\begin{aligned} Q_1(t)^2 &\overset{(1)}{=} \big(c^i \partial_i\big)^2 + t^2 \big(f_i \hat{c}^{\,i}\big)^2 + t\big\{c^i \partial_i,\; f_j \hat{c}^{\,j}\big\} \\ &\overset{(2)}{=} -\partial_i \partial_i + t^2 f_i f_i + t\big(c^i \hat{c}^{\,j} f_{ij} + f_j \{c^i, \hat{c}^{\,j}\}\partial_i\big) \\ &\overset{(3)}{=} \Delta + t^2\,|\mathrm{d}f|^2 + t\, f_{ij}\, c^i \hat{c}^{\,j}. \end{aligned} \tag{7.188}$$

(1) $(A + B)^2 = A^2 + B^2 + \{A, B\}$.

(2) $(c^i\partial_i)^2 = \frac{1}{2}\{c^i, c^j\}\partial_i\partial_j = -\partial_i\partial_i$ and $(f_i\hat{c}^{\,i})^2 = \frac{1}{2}\{\hat{c}^{\,i}, \hat{c}^{\,j}\}f_if_j = f_if_i$ (symmetric coefficients pick out the anticommutator). In the cross term, ∂_i either differentiates f_j or passes through, and $c^i\hat{c}^{\,j} + \hat{c}^{\,j}c^i = \{c^i, \hat{c}^{\,j}\}$.

(3) The two Clifford algebras anticommute, so the first-derivative term vanishes identically.

Here the cancellation of first derivatives is one line: *it is the anticommutation of c with $\hat{c}$*. To compare with (7.186), expand for a symmetric matrix f_{ij}:

$$f_{ij}\,c^i\hat{c}^{\,j} \overset{(1)}{=} f_{ij}\big(a^{i*}a^{j*} + a^{i*}a^j - a^ia^{j*} - a^ia^j\big) \overset{(2)}{=} f_{ij}\big(a^{i*}a^j - a^ja^{i*}\big) = f_{ij}\big[a^{i*}, a^j\big].$$

(1) Multiply out $(a^{i*} - a^i)(a^{j*} + a^j)$.

(2) $f_{ij}a^{i*}a^{j*} = 0$ and $f_{ij}a^ia^j = 0$ (symmetric times antisymmetric); in the third term swap the dummy labels $i \leftrightarrow j$.

The two forms agree.

Dictionary: the three ways to write the Hessian term

fermion form	$t\,f_{ij}\,[a^{i*}, a^j] = t\,f_{ij}\,(2a^{i*}a^j - \delta^{ij})$
Clifford form	$t\,f_{ij}\,c^i\hat{c}^{\,j}$
on p-forms, diagonal Hessian	$t\big(\sum_{i\in I}\lambda_i - \sum_{i\notin I}\lambda_i\big)$ on e^I
path integral	$t\,D_iD_jf\;\bar{\psi}^i\psi^j$ (Nash (7.88))

The first line uses $[a^{i*}, a^j] = 2a^{i*}a^j - \{a^j, a^{i*}\} = 2a^{i*}a^j - \delta^{ij}$. The third line follows from it: if $f_{ij} = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$, then $a^{i*}a^i$ is 1 on a basis form $e^I = e^{i_1} \wedge \cdots \wedge e^{i_p}$ if $i \in I$ and 0 otherwise, so $\lambda_i(2a^{i*}a^i - 1)$ is $+\lambda_i$ or $-\lambda_i$.

Worked example 7.75: the Hessian term as a 4×4 matrix in the plane

In the basis $(1, \mathrm{d}x, \mathrm{d}y, \mathrm{d}x \wedge \mathrm{d}y)$ of Example 7.72, take a general symmetric Hessian $\left(\begin{smallmatrix} h_{11} & h_{12} \\ h_{12} & h_{22} \end{smallmatrix}\right)$. From $[a^{i*}, a^j] = 2a^{i*}a^j - \delta^{ij}$ and the matrices of that example, $a^{1*}a^1 = \mathrm{diag}(0,1,0,1)$, $a^{2*}a^2 = \mathrm{diag}(0,0,1,1)$, and $a^{1*}a^2$ sends $\mathrm{d}y \mapsto \mathrm{d}x$ (and kills everything else), $a^{2*}a^1$ sends $\mathrm{d}x \mapsto \mathrm{d}y$. Therefore

$$h_{ij}[a^{i*}, a^j] = \begin{pmatrix} -h_{11} - h_{22} & 0 & 0 & 0 \\ 0 & h_{11} - h_{22} & 2h_{12} & 0 \\ 0 & 2h_{12} & h_{22} - h_{11} & 0 \\ 0 & 0 & 0 & h_{11} + h_{22} \end{pmatrix}.$$

Read it degree by degree.

- On functions: $-\operatorname{tr} h = -\nabla^2 f$.
- On 1-forms: $2h - (\operatorname{tr} h)\mathbf{1}$. Its eigenvalues are $\pm(\lambda_1 - \lambda_2)$, where $\lambda_{1,2}$ are the eigenvalues of h.
- On 2-forms: $+\operatorname{tr} h$.

At a saddle, $\lambda_1 = -1$, $\lambda_2 = +1$ say, the four diagonal entries in an eigenbasis are $0, -2, +2, 0$: the most negative "spin energy" is found on the 1-form $\mathrm{d}x$ along the *downhill* direction. At a minimum ($\lambda_1 = \lambda_2 = 1$) the most negative entry, -2, is on the 0-form; at a maximum it is on the 2-form. This is the index appearing already, before we have solved anything.

Physical picture: Pauli Hamiltonian

In one dimension $c = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ and $\hat{c} = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ in the basis $(1, \mathrm{d}x)$, so $c\hat{c} = \mathrm{diag}(-1, 1) = -\sigma_z$ and

$$H_t = -\frac{\mathrm{d}^2}{\mathrm{d}x^2} + t^2 f'^2 - t f'' \, \sigma_z ,$$

which is (7.181). This is exactly the Pauli Hamiltonian of a spin-$\frac{1}{2}$ particle in a magnetic field $\propto f''$ along z: the spin-up component (0-forms) feels $-t f''$, the spin-down component (1-forms) feels $+t f''$. In n dimensions the "spin" is the 2^n-dimensional fermion Fock space and the "magnetic field" is the Hessian matrix.

7.13.3 Large t: the potential wins

Nash's next observation [Nash p. 210, (7.77)] is that for large t the potential is dominated by its t^2 part,

$$V \longrightarrow t^2 \left|\mathrm{grad}\, f\right|^2 . \tag{7.189}$$

Think of $H_t\psi = E\psi$ as a Schrödinger equation. Wherever $|\mathrm{d}f| \neq 0$, the potential energy is of order t^2, which is enormous. A state of moderate energy cannot afford to be there. The only places where V is not huge are the zeros of $\mathrm{d}f$, the critical points. So as $t \to \infty$ the low-energy eigenfunctions are forced to concentrate near the critical points of f.

Let us make "moderate" quantitative. Suppose f is Morse, and remove a small ball B_r of radius r around each critical point. Outside the balls $|\mathrm{d}f| \geq c > 0$ (a continuous nonvanishing function on a compact set is bounded below), and the Hessian term is bounded by $t\,C$ for some constant C. So outside the balls

$$V \;\geq\; t^2 c^2 - tC \;\xrightarrow{\;t\to\infty\;}\; \infty \quad \text{like } t^2 .$$

Inside a ball, near a critical point x_0, we have $\mathrm{d}f \approx h \cdot (x - x_0)$ and all three pieces of H_t are comparable once we zoom in by the right amount. Set $x - x_0 = y/\sqrt{t}$:

$$\begin{aligned} H_t &\overset{(1)}{=} -t\,\partial_y^2 + t^2\,\frac{(hy)^2}{t} + t\,h_{ij}[a^{i*}, a^j] + O(t^{1/2}) \\ &\overset{(2)}{=} t\Big(-\partial_y^2 + (hy)^2 + h_{ij}[a^{i*}, a^j]\Big) + O(t^{1/2}). \end{aligned} \tag{7.190}$$

(1) $\partial_x = \sqrt{t}\,\partial_y$ and $|\mathrm{d}f|^2 = |h(x - x_0)|^2 + O(|x - x_0|^3)$; the cubic correction becomes $t^2 \cdot t^{-3/2}\,|y|^3 = O(t^{1/2})$. The Hessian is evaluated at x_0 up to a correction of order $|x - x_0| \sim t^{-1/2}$, also giving $O(t^{1/2})$.

(2) Factor out t.

Key idea: the two scales of Witten's analysis

Near each critical point, H_t is t times a *fixed*, t-independent operator, the bracket in (7.190), plus corrections smaller by $t^{-1/2}$. That operator is a set of harmonic oscillators with spin (Section 7.14). So low-lying states have energies of order t (or smaller), and they live in balls of radius of order $t^{-1/2}$ around the critical points. Everything else costs order t^2.

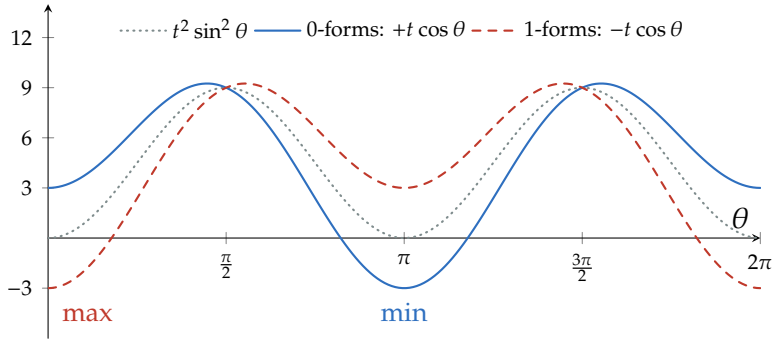


Figure 7.53: Witten's potential on the circle with $f = \cos\theta$, at $t = 3$. The common part $t^2\,|\mathrm{d}f|^2 = t^2\sin^2\theta$ (dotted) vanishes at both critical points. The Hessian term $\mp t f''$ lowers the 0-form potential (solid) at the minimum $\theta = \pi$, to $-t$, and raises it at the maximum. For 1-forms (dashed) it is the other way round. Each sector has a single deep well, at the critical point whose index is its degree; the $-t$ at the bottom is exactly what is needed to cancel the zero-point energy $+t$ of the oscillator there.

Worked example 7.76: the height function on the round S^2

Take the unit sphere and $f = z = \cos\theta$, with its minimum at the south pole and maximum at the north pole. We need $|\mathrm{d}f|^2$ and the covariant Hessian of z.

- $\mathrm{d}z = -\sin\theta\,\mathrm{d}\theta$ and $|\mathrm{d}\theta| = 1$ in the round metric $\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2$, so $|\mathrm{d}f|^2 = \sin^2\theta = 1 - z^2$.
- The coordinate functions restricted to the unit sphere satisfy $\nabla_i\nabla_j z = -z\,g_{ij}$. (Check at the north pole: in the tangent-plane coordinates $z = \sqrt{1-x^2-y^2} \approx 1 - \frac{1}{2}(x^2+y^2)$, so the Hessian is $-\mathbf{1} = -z\,g$ there. The general case follows by rotational symmetry, or from the Gauss formula for the second fundamental form of the sphere.)

Insert $h_{ij} = -z\,\delta_{ij}$ (orthonormal frame) into $h_{ij}[a^{i*}, a^j] = 2h_{ij}a^{i*}a^j - \operatorname{tr} h$:

$$t\,h_{ij}[a^{i*}, a^j] \overset{(1)}{=} t\left(-2z\,a^{i*}a^i + 2z\right) \overset{(2)}{=} 2tz\,(1-N),$$

(1) $h_{ij}a^{i*}a^j = -z\,a^{i*}a^i$ and $\operatorname{tr} h = -2z$ in two dimensions.
(2) $N = a^{i*}a^i$ is the degree.

So, degree by degree,

$$H_t\big|_{\Omega^0} = \Delta + t^2\sin^2\theta + 2t\cos\theta, \qquad H_t\big|_{\Omega^1} = \Delta + t^2\sin^2\theta, \qquad H_t\big|_{\Omega^2} = \Delta + t^2\sin^2\theta - 2t\cos\theta.$$

The 0-forms feel an extra $-2t$ at the south pole ($\cos\theta = -1$), the 2-forms an extra $-2t$ at the north pole, and the 1-forms nothing. Near the south pole the 0-form operator is a two-dimensional oscillator $-\nabla^2 + t^2\rho^2 - 2t$ (with $\rho = \pi - \theta$ small), whose zero-point energy $2\cdot t$ cancels the $-2t$ exactly. The exact zero modes of Section 7.12.4, $e^{-t\cos\theta}$ and $e^{t\cos\theta}\mathrm{vol}$, sit precisely there.

Worked example 7.77: the flat torus: two circles at once

On $T^2 = \mathbb{R}^2/(2\pi\mathbb{Z})^2$ take $f = \cos x + \cos y$. The metric is flat, so the covariant Hessian is

the ordinary one, $h = \operatorname{diag}(-\cos x, -\cos y)$, and $|\mathrm{d}f|^2 = \sin^2 x + \sin^2 y$. Then

$$H_t \overset{(1)}{=} \left(-\partial_x^2 + t^2 \sin^2 x - t \cos x\,[a^{1*}, a^1]\right) + \left(-\partial_y^2 + t^2 \sin^2 y - t \cos y\,[a^{2*}, a^2]\right)$$
$$\overset{(2)}{=} H_t^{(x)} \otimes \mathbf{1} + \mathbf{1} \otimes H_t^{(y)}.$$

(1) (7.186) with a diagonal, coordinate-separated Hessian.
(2) Forms on T^2 are tensor products of forms in x and forms in y ($\mathrm{d}x \wedge \mathrm{d}y = \mathrm{d}x \otimes \mathrm{d}y$), and each bracket acts on one factor only.

The Hamiltonian is the sum of two copies of the circle Hamiltonian of Example 7.73. Its eigenstates are products and its eigenvalues are sums. In particular its exact zero modes are the four products of the circle zero modes,

$$e^{-t\cos x} e^{-t\cos y}, \quad e^{t\cos x} e^{-t\cos y}\,\mathrm{d}x, \quad e^{-t\cos x} e^{t\cos y}\,\mathrm{d}y, \quad e^{t\cos x} e^{t\cos y}\,\mathrm{d}x \wedge \mathrm{d}y,$$

localised at the minimum (π, π), the two saddles $(0, \pi)$ and $(\pi, 0)$, and the maximum $(0, 0)$ respectively: $b = (1, 2, 1)$. This is the Künneth formula $H^*(T^2) = H^*(S^1) \otimes H^*(S^1)$ at the level of supersymmetric ground states.

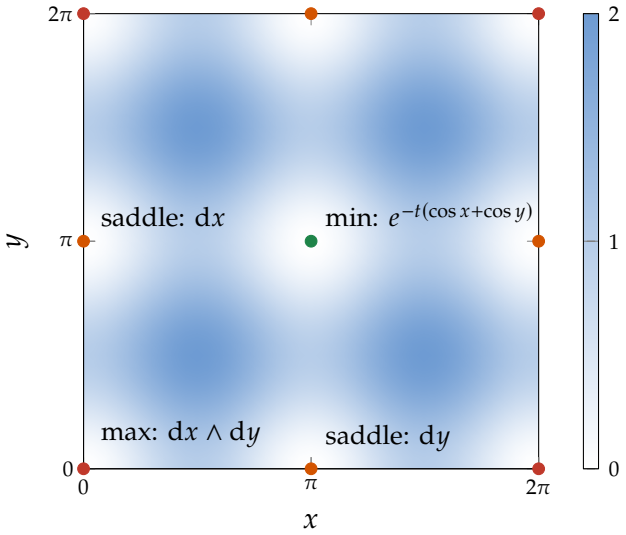


Figure 7.54: The bosonic potential $|\mathrm{d}f|^2 = \sin^2 x + \sin^2 y$ on the flat torus (the square with opposite edges identified; the four red corners are one point). It has four zeros, one per critical point of $f = \cos x + \cos y$: the maximum (red), two saddles (orange) and the minimum (green). At large t the potential $t^2\,|\mathrm{d}f|^2$ confines the low-lying states to these four wells, and the Hessian term selects the degree of the state in each well, which equals the index. Here the four states are exact zero modes, so $b = (1, 2, 1)$.

Reading Nash (p. 210, (7.76)–(7.77))

Nash states (7.186) with "the expansion terminates at the t^2 term" because $\mathrm{d}\mathrm{d}^* + \mathrm{d}^*\mathrm{d}$ is second order; more precisely, it terminates because d_t is linear in t. We derived every term: the t^2 term from the anticommutation relations, the t term from the product rule, and the disappearance of first derivatives from $\{c, \hat{c}\} = 0$. Nash writes $[a_i^*, a_j]$ with lower indices; in an orthonormal frame the position of indices does not matter. His remark that "this enormous growth of the potential energy ... will force the eigenfunctions of H_t to be zero"

away from critical points is the estimate $V \geq t^2c^2 - tC$ above.

Recap

- $H_t = \Delta + t^2\,|\mathrm{d}f|^2 + t\,\nabla_i\nabla_j f\,[a^{i*}, a^j]$: a Laplacian, a scalar potential, and a Hessian–fermion coupling.
- In Clifford form $Q_1(t) = c^i\nabla_i + t\,\partial_i f\,\hat{c}^{\,i}$, and $H_t = Q_1(t)^2 = \Delta + t^2\,|\mathrm{d}f|^2 + t\,f_{ij}c^i\hat{c}^{\,j}$; first derivatives cancel because $\{c^i, \hat{c}^{\,j}\} = 0$.
- On a basis form e^I, a diagonal Hessian contributes $t(\sum_{i\in I}\lambda_i - \sum_{i\notin I}\lambda_i)$: most negative when I is the set of downhill directions.
- Large t: states of energy $\ll t^2$ live within $t^{-1/2}$ of the critical points, where $H_t \approx t\times$(oscillators with spin).

7.14 Harmonic oscillators at the critical points: the weak Morse inequalities

At large t the low-lying states of H_t live in tiny balls around the critical points of f, and inside each ball H_t is t times a fixed operator (Section 7.13.3). The question of this section is: *what is the spectrum of that local operator, and in particular how many states of (almost) zero energy does each critical point contribute, and in which degree?* The answer [Nash pp. 210–211, (7.78)–(7.80)] is the heart of Witten's argument: *each critical point of index p supplies exactly one low-lying state, and it is a p-form.* Comparing with the exact count b_p of zero-energy p-forms gives the weak Morse inequalities $b_p \leq m_p$.

The only piece of physics needed is the harmonic oscillator. We first review it in the normalisation used here, then solve the local model, and then explain how the local answers are glued into a global statement.

7.14.1 The harmonic oscillator, by ladder operators

Our oscillators have the form

$$h_\omega = -\frac{\mathrm{d}^2}{\mathrm{d}x^2} + \omega^2x^2, \qquad \omega > 0. \tag{7.191}$$

This is $p^2 + \omega^2x^2$ with $p = -\mathrm{i}\,\mathrm{d}/\mathrm{d}x$: the usual $\frac{p^2}{2m} + \frac{1}{2}m\Omega^2x^2$ with $\hbar = 1$, $m = \frac{1}{2}$ and $\Omega = 2\omega$. So we expect levels spaced by $\hbar\Omega = 2\omega$. Let us derive them without remembering anything. Introduce the first-order operators

$$A = \frac{\mathrm{d}}{\mathrm{d}x} + \omega x, \qquad A^\dagger = -\frac{\mathrm{d}}{\mathrm{d}x} + \omega x.$$

They are the supercharge blocks of (7.150) with superpotential $W = \frac{1}{2}\omega x^2$. Acting on a test function φ,

$$\begin{aligned} A^\dagger A\,\varphi &\overset{(1)}{=} \Big(-\frac{\mathrm{d}}{\mathrm{d}x} + \omega x\Big)\big(\varphi' + \omega x\varphi\big) \\ &\overset{(2)}{=} -\varphi'' - \omega\varphi - \omega x\varphi' + \omega x\varphi' + \omega^2x^2\varphi = (h_\omega - \omega)\,\varphi, \\ AA^\dagger\,\varphi &\overset{(3)}{=} -\varphi'' + \omega\varphi + \omega^2x^2\varphi = (h_\omega + \omega)\,\varphi. \end{aligned} \tag{7.192}$$

(1) Definitions of A and $A^\dagger$.
(2) Product rule on $(\omega x\varphi)' = \omega\varphi + \omega x\varphi'$; the two first-derivative terms cancel.
(3) The same computation with the sign of $\mathrm{d}/\mathrm{d}x$ reversed in both factors; now the product rule gives $+\omega\varphi$.

Subtracting, $[A, A^\dagger] = 2\omega$, and

$$h_\omega = A^\dagger A + \omega, \qquad [h_\omega, A^\dagger] = 2\omega\, A^\dagger, \qquad [h_\omega, A] = -2\omega\, A. \tag{7.193}$$

(For the middle one: $[A^\dagger A, A^\dagger] = A^\dagger[A, A^\dagger] = 2\omega A^\dagger$.) So $A^\dagger$ raises the energy by 2ω and A lowers it by 2ω. Lowering cannot go on for ever, because $\langle\varphi, h_\omega\varphi\rangle = \|A\varphi\|^2 + \omega\,\|\varphi\|^2 \geq \omega\,\|\varphi\|^2$. The ladder must stop at a state with $A\varphi_0 = 0$, a first-order equation:

$$\varphi_0' + \omega x\varphi_0 = 0 \implies \varphi_0(x) = \left(\frac{\omega}{\pi}\right)^{1/4} e^{-\omega x^2/2}, \qquad h_\omega\varphi_0 = \omega\,\varphi_0. \tag{7.194}$$

The normalisation uses $\int_{-\infty}^{\infty} e^{-\omega x^2}\mathrm{d}x = \sqrt{\pi/\omega}$.

Key idea: the oscillator in one line

The spectrum of $h_\omega = -\mathrm{d}^2/\mathrm{d}x^2 + \omega^2 x^2$ is

$$\omega\,(2n+1), \qquad n = 0, 1, 2, \ldots, \tag{7.195}$$

each level non-degenerate, with eigenfunctions $\varphi_n \propto (A^\dagger)^n\varphi_0$. The ground state is a Gaussian of width $\langle x^2\rangle^{1/2} = (2\omega)^{-1/2}$. The *zero-point energy* ω is the price of confining a particle to that width.

For example $\varphi_1 \propto A^\dagger\varphi_0 = \left(-\frac{\mathrm{d}}{\mathrm{d}x} + \omega x\right)e^{-\omega x^2/2} = 2\omega x\, e^{-\omega x^2/2}$, the familiar odd first excited state, with energy 3ω. In n dimensions, $\sum_i\left(-\partial_i^2 + \omega_i^2 x_i^2\right)$ separates into n independent oscillators, with eigenvalues $\sum_i \omega_i(2n_i+1)$ and eigenfunctions $\prod_i \varphi_{n_i}(x_i)$.

7.14.2 Zooming in on one critical point

Fix a non-degenerate critical point x_0 of f, of index k. We want coordinates in which both the metric and f are as simple as possible.

- *The metric.* Riemann normal coordinates centred at x_0 give $g_{ij}(x) = \delta_{ij} + O(|x|^2)$ and $\Gamma^i_{jk}(0) = 0$ (the metric is "flat to second order", as Nash puts it [Nash p. 210]). Rotating these coordinates we may also diagonalise the Hessian at x_0, since a symmetric matrix can be diagonalised by an orthogonal change of basis.

- *The function.* Taylor's theorem then gives

$$f(x) = f(x_0) + \tfrac{1}{2}\left(\lambda_1 x_1^2 + \cdots + \lambda_n x_n^2\right) + O(|x|^3), \tag{7.196}$$

  where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of the Hessian, all nonzero. Exactly k of them are negative. We order them so that $\lambda_1, \ldots, \lambda_k < 0 < \lambda_{k+1}, \ldots, \lambda_n$.

Nash writes (7.196) with $\lambda_i = \mp 1$ [Nash p. 210, (7.78)] and remarks that the factor $\frac{1}{2}$ is "just chosen for convenience". That is the Morse lemma normal form (Section 7.9.2) rescaled by $\frac{1}{2}$.

Watch out: can we really have $\lambda_i = \pm 1$ *and* a flat metric?

The Morse lemma coordinates, in which f is *exactly* quadratic with coefficients $\pm\frac{1}{2}$, are in general not orthonormal for a given metric. With the metric fixed, the honest statement is (7.196) with arbitrary nonzero λ_i. There are two ways out, and both are fine. (i) Keep general λ_i; we do this below, and only their signs will matter. (ii) Use the freedom to choose the metric: the Betti numbers do not depend on it, so we may take g to be the Euclidean metric of the Morse coordinates near each critical point, glued to anything elsewhere. Then $\lambda_i = \pm 1$ exactly and the cubic remainder vanishes. This is what Nash's (7.78) implicitly does.

7.14.3 The local Hamiltonian separates

Now insert (7.196) into Witten's Hamiltonian (7.186), dropping everything that is small after the rescaling $x = y/\sqrt{t}$ of (7.190). To this order $\partial_i f = \lambda_i x_i$ (no sum), the covariant Hessian is $\operatorname{diag}(\lambda_1, \dots, \lambda_n)$, and the Laplacian is the flat one acting on each component of the form. So

$$\begin{aligned} \widetilde{H}_t &\overset{(1)}{=} -\sum_i \partial_i^2 + t^2 \sum_i (\lambda_i x_i)^2 + t \sum_{i,j} \lambda_i \delta_{ij} \, [a^{i*}, a^j] \\ &\overset{(2)}{=} \sum_{i=1}^{n} \Big\{ -\frac{\partial^2}{\partial x_i^2} + t^2 \lambda_i^2 x_i^2 + t \lambda_i \, [a^{i*}, a^i] \Big\} \;:=\; \sum_{i=1}^{n} h^{(i)}. \end{aligned} \tag{7.197}$$

(1) The three terms of (7.186): $\Delta = -\sum_i \partial_i^2$ on each component in flat Cartesian coordinates (Example 7.68), $g^{ij}\partial_i f \, \partial_j f = \sum_i \lambda_i^2 x_i^2$, and $D^2 f / Dx^i Dx^j = \lambda_i \delta_{ij}$.
(2) Collect the terms belonging to each coordinate direction.

This is Nash's (7.79) [Nash p. 211], with $\lambda_i^2 = 1$ there. It is a sum of n commuting pieces, one per direction. Each piece is a *supersymmetric harmonic oscillator* (Example 7.67) with frequency $\omega_i = t\,|\lambda_i|$: a bosonic oscillator in x_i plus a fermionic "spin" in mode i.

Why do the pieces commute? Different $h^{(i)}$ involve different coordinates, and their fermionic parts are the even operators $[a^{i*}, a^i]$, which commute with each other for different i (moving a pair of fermion operators past another pair produces two minus signs). Within a single $h^{(i)}$, the oscillator part acts only on the coefficient functions and the spin part only on the fermion index, so they commute too. That is Nash's remark that "the harmonic oscillator contribution to H_t commutes with the correction terms".

The spin part has eigenvalues ± 1. For one mode,

$$[a^{i*}, a^i] \overset{(1)}{=} a^{i*}a^i - a^i a^{i*} \overset{(2)}{=} a^{i*}a^i - \big(1 - a^{i*}a^i\big) = 2\, a^{i*}a^i - 1.$$

(1) Definition of the commutator.
(2) $\{a^i, a^{i*}\} = 1$ from (7.165).

The occupation number $a^{i*}a^i$ is 1 on a basis form $e^I = \mathrm{d}x^{i_1} \wedge \cdots \wedge \mathrm{d}x^{i_p}$ if $i \in I$ and 0 if $i \notin I$. So

$$[a^{i*}, a^i]\, e^I = \varepsilon_i \, e^I, \qquad \varepsilon_i = \begin{cases} +1 & i \in I \ (\mathrm{d}x^i \text{ present}), \\ -1 & i \notin I \ (\mathrm{d}x^i \text{ absent}). \end{cases} \tag{7.198}$$

These are Nash's $\varepsilon_i = \mp 1$ [Nash p. 211]: "the simple spectrum ∓ 1".

7.14.4 The spectrum and the unique zero mode

A complete set of eigenstates of (7.197) is

$$\Psi_{n,I} = \varphi_{n_1}(x_1)\cdots\varphi_{n_n}(x_n)\, e^I, \qquad n = (n_1,\dots,n_n)\in\mathbb{N}^n, \quad I\subseteq\{1,\dots,n\}, \tag{7.199}$$

where φ_{n_i} is the n_i-th eigenfunction of the oscillator with $\omega_i = t\,|\lambda_i|$. Each $h^{(i)}$ acts on $\Psi_{n,I}$ by a number, so

$$\begin{aligned}\widetilde{H}_t\,\Psi_{n,I} &\overset{(1)}{=} \sum_i \Big(t\,|\lambda_i|\,(2n_i+1) + t\lambda_i\varepsilon_i\Big)\Psi_{n,I}\\ &\overset{(2)}{=} t\sum_i \Big\{|\lambda_i|\,(1+2n_i) + \lambda_i\varepsilon_i\Big\}\Psi_{n,I}.\end{aligned} \tag{7.200}$$

(1) The oscillator part of $h^{(i)}$ gives $\omega_i(2n_i+1)$ by (7.195); the spin part gives $t\lambda_i\varepsilon_i$ by (7.198).
(2) Factor out t. With $|\lambda_i| = 1$ this is Nash's (7.80), $t\sum_i\{(1+2n_i)+\lambda_i\varepsilon_i\}$.

The degree of $\Psi_{n,I}$ is $p = |I|$, the number of ε_i equal to $+1$. To find the zero eigenvalues in degree p, look at one term of the sum at a time:

$$|\lambda_i|\,(1+2n_i) + \lambda_i\varepsilon_i \overset{(1)}{\geq} |\lambda_i| + \lambda_i\varepsilon_i \overset{(2)}{\geq} |\lambda_i| - |\lambda_i| = 0. \tag{7.201}$$

(1) $n_i \geq 0$, with equality only for $n_i = 0$.
(2) $\lambda_i\varepsilon_i = \pm|\lambda_i|$, with the minus sign exactly when $\varepsilon_i = -\operatorname{sgn}\lambda_i$.

Every term is non-negative, so the total vanishes only if every term does, i.e. only if for every i

$$n_i = 0 \qquad\text{and}\qquad \varepsilon_i = \begin{cases} +1 & \text{if } \lambda_i < 0 \text{ (downhill direction)},\\ -1 & \text{if } \lambda_i > 0 \text{ (uphill direction)}.\end{cases}$$

This fixes n and I completely: $I = \{1,\dots,k\}$, the set of downhill directions, of size k = index.

Key idea: one zero mode per critical point, in degree = index

The local Hamiltonian (7.197) at a critical point of index k has exactly one state of zero energy,

$$\Psi_0 = \prod_{i=1}^n \Big(\frac{t\,|\lambda_i|}{\pi}\Big)^{1/4} \exp\Big(-\frac{t}{2}\sum_{i=1}^n |\lambda_i|\,x_i^2\Big)\,\mathrm{d}x^1\wedge\cdots\wedge\mathrm{d}x^k, \tag{7.202}$$

a Gaussian of width $\sim t^{-1/2}$ times the wedge of the downhill directions. It is a k-form. Every other state has energy at least $2t\,\min_i|\lambda_i|$.

The gap statement follows from (7.201): to move away from the zero mode we must either raise some n_i to 1 (cost $2t\,|\lambda_i|$) or flip some ε_i (cost $2t\,|\lambda_i|$ as well).

A direct check. The zero mode must be annihilated by both supercharges of the local model, $\mathrm{d}_t = \mathrm{d} + t\,\mathrm{d}f\wedge$ and $\mathrm{d}_t^* = \mathrm{d}^* + t\,\iota_{\nabla f}$. Write $\Psi_0 = G\,\mathrm{d}x^1\wedge\cdots\wedge\mathrm{d}x^k$ with $G = \exp(-\frac{t}{2}\sum_j|\lambda_j|\,x_j^2)$, so $\partial_j G = -t\,|\lambda_j|\,x_j G$, and $\partial_j f = \lambda_j x_j$. Then

$$\mathrm{d}_t\Psi_0 \overset{(1)}{=} \sum_{j>k}\big(\partial_j G + t\lambda_j x_j G\big)\,\mathrm{d}x^j\wedge\mathrm{d}x^1\wedge\cdots\wedge\mathrm{d}x^k \overset{(2)}{=} \sum_{j>k}\big(-t\lambda_j x_j + t\lambda_j x_j\big)G\,(\cdots) = 0,$$

$$\mathrm{d}_t^*\Psi_0 \overset{(3)}{=} \sum_{j\leq k}\big(-\partial_j G + t\lambda_j x_j G\big)\,\iota_{\partial_j}\big(\mathrm{d}x^1\wedge\cdots\wedge\mathrm{d}x^k\big) \overset{(4)}{=} \sum_{j\leq k}\big(t\,|\lambda_j| + t\lambda_j\big)x_j G\,(\cdots) = 0.$$

(1) $\mathrm{d}(G\,e^I) = \partial_j G\,\mathrm{d}x^j \wedge e^I$ and $t\,\mathrm{d}f \wedge e^I = t\lambda_j x_j\,\mathrm{d}x^j \wedge e^I$; the terms with $j \le k$ vanish because $\mathrm{d}x^j$ already appears in e^I.
(2) For $j > k$ (uphill), $|\lambda_j| = \lambda_j$, so $\partial_j G = -t\lambda_j x_j G$ cancels the second term.
(3) $\mathrm{d}^* = -\sum_j a^j \partial_j$ from (7.166) and $\iota_{\nabla f} = \sum_j \lambda_j x_j\, a^j$; only $j \le k$ survive, because ι_{∂_j} kills e^I when $\mathrm{d}x^j$ is absent.
(4) For $j \le k$ (downhill), $|\lambda_j| = -\lambda_j$, so the bracket vanishes.

Two first-order equations, both solved. Notice what the Gaussian is: along the uphill directions it is $e^{-t(f-f(x_0))}$, the local version of the 0-form zero mode e^{-tf}; along the downhill directions it is $e^{+t(f-f(x_0))}$ times $\mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^k$, the local version of the top-form zero mode $e^{tf}\,\mathrm{vol}$ of Section 7.12.4. Near a critical point of index k, M looks like $\mathbb{R}^k_{\text{down}} \times \mathbb{R}^{n-k}_{\text{up}}$, and the zero mode is the product of the "top form on the first factor" and the "function on the second".

Worked example 7.78: one dimension: a minimum and a maximum

Take $n = 1$ and $f = f(x_0) + \frac{1}{2}\lambda x^2$. The two sectors are $\varepsilon = -1$ (0-forms) and $\varepsilon = +1$ (1-forms). From (7.200):

	0-forms ($\varepsilon = -1$)	1-forms ($\varepsilon = +1$)
minimum, $\lambda > 0$	$t\lambda(2n+1) - t\lambda = 2t\lambda\, n$	$t\lambda(2n+1) + t\lambda = 2t\lambda(n+1)$
maximum, $\lambda < 0$	$t\,\lvert\lambda\rvert\,(2n+1) + t\,\lvert\lambda\rvert = 2t\,\lvert\lambda\rvert\,(n+1)$	$2t\,\lvert\lambda\rvert\; n$

At a minimum the 0-forms have the levels $0, 2t\lambda, 4t\lambda, \ldots$ and the 1-forms $2t\lambda, 4t\lambda, \ldots$: exactly the supersymmetric oscillator of Example 7.67 with $\omega = t\lambda$, whose superpotential is $W = tf$ as in (7.180). The unique zero mode is the 0-form $e^{-t\lambda x^2/2}$. At a maximum the roles are swapped, and the zero mode is the 1-form $e^{-t|\lambda|x^2/2}\mathrm{d}x$. In the language of Figure 7.53: the deep well of the 0-form potential is at the minimum, that of the 1-form potential at the maximum, and in each well the "spin" energy $-t\,|\lambda|$ cancels the zero-point energy $+t\,|\lambda|$.

Worked example 7.79: two dimensions: minimum, saddle, maximum

Take $n = 2$ and $|\lambda_1| = |\lambda_2| = 1$. The four fermion states are $1, \mathrm{d}x, \mathrm{d}y, \mathrm{d}x \wedge \mathrm{d}y$ with $(\varepsilon_1, \varepsilon_2) = (-,-), (+,-), (-,+), (+,+)$. The term $\sum_i \lambda_i \varepsilon_i$ shifts the oscillator energy $t(2 + 2n_1 + 2n_2)$. The shifts, in units of t:

	1	$\mathrm{d}x$	$\mathrm{d}y$	$\mathrm{d}x \wedge \mathrm{d}y$
minimum $(\lambda_1, \lambda_2) = (+1, +1)$, index 0	-2	0	0	$+2$
saddle $(\lambda_1, \lambda_2) = (-1, +1)$, index 1	0	-2	$+2$	0
maximum $(\lambda_1, \lambda_2) = (-1, -1)$, index 2	$+2$	0	0	-2

The lowest level in each sector is $t(2 + \text{shift})$: zero only for the entry -2, and that entry sits in degree 0, 1, 2 respectively. The saddle's zero mode is $e^{-t(x^2+y^2)/2}\,\mathrm{d}x$: a 1-form pointing along the downhill direction x. We met exactly these numbers as the diagonal of the 4×4 Hessian matrix in Example 7.75. The full low-lying spectrum is drawn in Figure 7.55; at every level above zero, bosons (1, $\mathrm{d}x \wedge \mathrm{d}y$) and fermions ($\mathrm{d}x$, $\mathrm{d}y$) are equal in number, as supersymmetry requires.

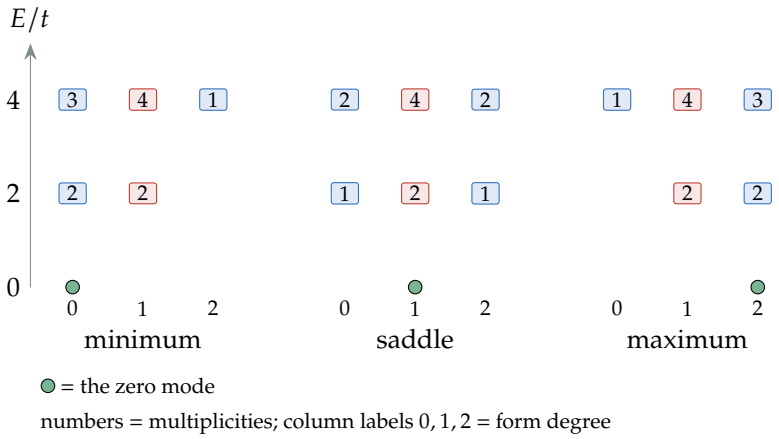


Figure 7.55: Spectrum of the local model $\widetilde{H}_t$ at the three kinds of non-degenerate critical point in two dimensions ($|\lambda_i| = 1$), up to $E = 4t$, sorted by form degree. Blue boxes are bosons (even forms), red boxes fermions (odd forms). Each critical point has a single zero-energy state (green dot), in degree equal to its index; everything else is at least $2t$ above it, and every positive level is boson–fermion balanced.

Worked example 7.80: general eigenvalues, and the round sphere

With general λ_i, the levels near a critical point are $t\sum_i\{|\lambda_i|\,(1+2n_i)+\lambda_i\varepsilon_i\}$. For the height function $f=\cos\theta$ on the unit S^2 at the south pole, $f=-1+\frac{1}{2}\rho^2+O(\rho^4)$ in geodesic polar coordinates, so $\lambda_1=\lambda_2=+1$; at the north pole $\lambda_1=\lambda_2=-1$. The local zero modes are

$$\Psi_{\mathrm{S}}=\sqrt{\tfrac{t}{\pi}}\,e^{-t\rho^2/2},\qquad \Psi_{\mathrm{N}}=\sqrt{\tfrac{t}{\pi}}\,e^{-t\rho'^2/2}\,\mathrm{vol},$$

with ρ, ρ' the geodesic distances from the two poles. Compare the *exact* zero modes $e^{-t\cos\theta}$ and $e^{t\cos\theta}\mathrm{vol}$ of Example 7.76: near the south pole $e^{-t\cos\theta}=e^{t}e^{-t(1-\cos\rho)}\approx e^{t}e^{-t\rho^2/2}$, the same Gaussian up to normalisation. For the sphere the local approximation is therefore accurate: both local zero modes survive as exact ones, matching $b_0=b_2=1$, $b_1=0$. The first excited states at each pole are at $2t$, confirming the gap $2t\min|\lambda_i|$.

7.14.5 From local to global: the weak Morse inequalities

We now have one approximate zero mode per critical point, in degree equal to the index. How do we turn this into a statement about the global operator $\Delta_p(t)$ on M? Define the counting function

$$N_p(t;E):=\#\{\text{eigenvalues of }\Delta_p(t)\text{ in }[0,E]\},\tag{7.203}$$

counted with multiplicity. The precise claim, due to Witten and proved by Simon and by Helffer and Sjöstrand, is:

Key idea: the low-lying spectrum at large t

Fix any constant c with $0<c<2\min|\lambda|$ (the minimum over all Hessian eigenvalues at all critical points). Then for all sufficiently large t,

$$N_p(t;\,ct)=m_p=\#\{\text{critical points of index }p\}.\tag{7.204}$$

In words: at large t the spectrum of $\Delta_p(t)$ splits into m_p **low-lying** eigenvalues that are $o(t)$ and the rest, which are at least of order t. The low-lying eigenfunctions are close to the

local zero modes (7.202), one at each critical point of index p.

The physicist believes this at once: the potential confines the particle to the wells, each well is an oscillator, and the oscillator levels are what we computed. It is worth seeing how the two halves of (7.204) are actually established, because the tools are standard quantum mechanics.

The min–max principle. For a self-adjoint $H \geq 0$ with discrete spectrum $E_1 \leq E_2 \leq \dots$, the k-th eigenvalue is

$$E_k = \min_{\dim V = k} \; \max_{\psi \in V,\, \psi \neq 0} \frac{\langle \psi, H\psi \rangle}{\langle \psi, \psi \rangle}.$$

Two consequences are all we need.

(a) If there is a k-dimensional space of trial states on which the Rayleigh quotient is $\leq E$, then H has at least k eigenvalues $\leq E$.

(b) If $\langle \psi, H\psi \rangle \geq E \, \|\psi\|^2$ for every ψ orthogonal to some k-dimensional space, then H has at most k eigenvalues $< E$.

(For a 2×2 matrix, (a) says that if some vector has Rayleigh quotient $\leq E$ then the smaller eigenvalue is $\leq E$; (b) says that if the quotient is $\geq E$ on the orthogonal complement of one vector, then the larger eigenvalue is $\geq E$.)

At least m_p (trial states). For each critical point x_0 of index p take the local zero mode (7.202), written in normal coordinates around x_0, and multiply it by a smooth cut-off function J equal to 1 in a ball of radius r and 0 outside radius $2r$. The m_p trial states have disjoint supports, so they are orthogonal and H_t, a differential operator, has no matrix elements between them. The Rayleigh quotient of any combination is therefore at most the largest of the individual quotients. Each individual quotient is small: the local zero mode has $\widetilde{H}_t$-energy zero, and the errors come from (i) the cut-off, which only acts where the Gaussian is $e^{-t|\lambda| r^2/2}$, exponentially small, and (ii) the difference $H_t - \widetilde{H}_t$, which on a Gaussian of width $t^{-1/2}$ is $o(t)$ by the scaling (7.190). So by (a), $N_p(t; ct) \geq m_p$ for large t.

At most m_p (localisation). Choose a partition of unity $\sum_k J_k^2 = 1$ on M: one function J_{x_0} supported in a ball $B_{2r}(x_0)$ around each critical point, and one function J_∞ supported away from all critical points. For a Schrödinger operator $H = -\nabla^2 + V$ there is an exact identity (the **IMS localisation formula**)

$$H = \sum_k J_k \, H \, J_k - \sum_k |\nabla J_k|^2 . \tag{7.205}$$

It is proved by expanding a double commutator:

$$\sum_k J_k H J_k \overset{(1)}{=} \sum_k \Big(\tfrac{1}{2}\big(J_k^2 H + H J_k^2\big) - \tfrac{1}{2}\big[J_k, [J_k, H]\big] \Big)$$
$$\overset{(2)}{=} H - \tfrac{1}{2} \sum_k \big[J_k, [J_k, -\nabla^2]\big] \overset{(3)}{=} H + \sum_k |\nabla J_k|^2 .$$

(1) Expand $[J,[J,H]] = J^2 H - 2JHJ + HJ^2$ and solve for JHJ.
(2) $\sum_k J_k^2 = 1$; the potential (a multiplication operator, even a matrix one) commutes with J_k and drops out of the commutator.
(3) $[J, -\nabla^2]\varphi = (\nabla^2 J)\varphi + 2\nabla J \cdot \nabla \varphi$, and commuting once more with J leaves $-2\,|\nabla J|^2\,\varphi$.

Now estimate each piece of (7.205) with $r = t^{-2/5}$ (small, but much larger than the width $t^{-1/2}$ of the Gaussians):

- The localisation error $|\nabla J_k|^2 \sim r^{-2} = t^{4/5}$ is $o(t)$.
- On the support of J_∞, $|\mathrm{d}f| \gtrsim r$, so $V \geq t^2 r^2 c_1 - tC = c_1 t^{6/5} - tC$, which is $\gg t$.
- Inside $B_{2r}(x_0)$, H_t differs from the local model $\widetilde{H}_t$ by terms of size at most $t^2 r^3 + tr = O(t^{4/5})$, again $o(t)$. And $\widetilde{H}_t$ is $\geq 2t \min|\lambda|$ on the orthogonal complement of its single zero mode, in the one degree where it has a zero mode, and $\geq 2t \min|\lambda|$ everywhere in the other degrees.

Putting these into (7.205): on p-forms orthogonal to the m_p cut-off local zero modes, $\langle \psi, H_t \psi \rangle \geq (2 \min|\lambda| \; t - o(t)) \, \|\psi\|^2$. By (b), $N_p(t; ct) \leq m_p$ for large t. Together with the first half this is (7.204).

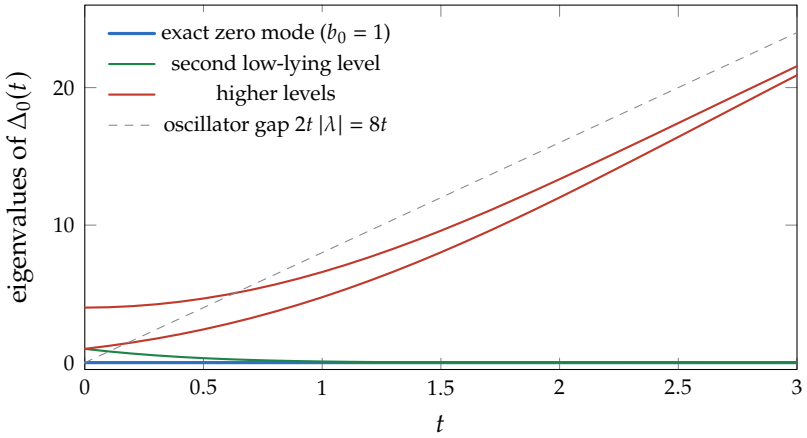


Figure 7.56: The lowest eigenvalues of $\Delta_0(t)$ on the circle for $f = -\cos 2\theta$ (two minima, two maxima, $|\lambda| = 4$), computed numerically in a Fourier basis. At $t = 0$ they are the free levels $0, 1, 1, 4, 4$. As t grows, exactly $m_0 = 2$ levels stay low: the exact zero mode (blue) and a second level (green) that falls to zero exponentially fast. All other levels (red; the upper curve is doubly degenerate) climb with slope close to the oscillator gap $8t$ (dashed). The weak inequality $b_0 = 1 \leq m_0 = 2$ is visible directly; the green level is lifted from zero only by tunnelling (Section 7.15).

Key idea: the weak Morse inequalities, Nash p. 211

For every p,

$$b_p \overset{(1)}{=} \dim \operatorname{Ker} \Delta_p(t) \overset{(2)}{\leq} N_p(t;\, ct) \overset{(3)}{=} m_p. \tag{7.206}$$

(1) Nash's (7.75), (7.177): the number of exact zero modes does not depend on t.
(2) An exact zero mode is in particular an eigenvalue in $[0, ct]$.
(3) (7.204), for t large enough.

Nash's comment is exact: "if our approximation were exact we would know that the number of these zero p-forms is equal to the Betti number b_p. Taking into account the approximation means that some of the zero p forms may disappear in an exact calculation." The inequality in step (2) is where information is lost: the m_p low-lying states include the b_p exact zero modes and possibly some others, at energies that are small but not zero.

7.14.6 The Euler characteristic comes for free

Supersymmetry adds one more piece of information. The low-lying levels that are *not* exactly zero come in boson–fermion pairs, like every positive level (Section 7.10.6). The pairing is done by $Q_1(t)$, which commutes with H_t and changes the degree by one, so it maps the low-lying space into itself. Hence

$$\begin{aligned}\sum_p (-1)^p m_p &\overset{(1)}{=} \sum_p (-1)^p N_p(t;ct)\\ &\overset{(2)}{=} \sum_p (-1)^p \dim\operatorname{Ker}\Delta_p(t) + \sum_{0<E\le ct}\big(n_B(E)-n_F(E)\big)\\ &\overset{(3)}{=} \sum_p (-1)^p b_p = \chi(M).\end{aligned} \tag{7.207}$$

(1) (7.204).
(2) Split the low-lying levels into $E = 0$ and $E > 0$; even degrees are bosons, odd degrees fermions.
(3) The pairing argument kills every $E > 0$ term; the $E = 0$ term is the Witten index (7.163).

This is the Morse-theoretic formula $\chi(M) = \sum_p (-1)^p m_p$ of Chapter 7, now obtained as a statement about ground states: the Witten index computed exactly (Hodge theory) equals the Witten index computed semiclassically (one state per critical point, with sign $(-1)^{\text{index}}$). Physically, $(-1)^{\text{index}}$ is the fermion number parity of the perturbative vacuum at that critical point. For a gradient vector field ∇f it is also the local index of the zero of the vector field, so (7.207) is the Poincaré–Hopf theorem for gradient fields.

Worked example 7.81: counting on familiar examples

The weak inequalities, and the Euler characteristic check, on the examples of this chapter:

M	f	(m_0, m_1, m_2)	(b_0, b_1, b_2)	$\sum(-1)^p m_p = \chi$
S^1	$\cos\theta$	$(1,1)$	$(1,1)$	0
S^1	$-\cos 2\theta$	$(2,2)$	$(1,1)$	0
S^2	height z	$(1,0,1)$	$(1,0,1)$	2
S^2	two-humped height	$(1,1,2)$	$(1,0,1)$	2
T^2	$\cos x + \cos y$	$(1,2,1)$	$(1,2,1)$	0
T^2	upright height	$(1,2,1)$	$(1,2,1)$	0
$\mathbb{RP}^2$	$\frac{x^2+2y^2+3z^2}{x^2+y^2+z^2}$	$(1,1,1)$	$(1,0,0)$	1

When $m_p = b_p$ for all p (a **perfect** Morse function), all the local zero modes must survive as exact zero modes: there is nothing for tunnelling to do. When $m_p > b_p$, some low-lying states must be lifted. On S^1 with $-\cos 2\theta$ one bosonic and one fermionic state are lifted, together; on the two-humped sphere a 1-form and a 2-form. In each case $m_p - b_p$ summed with alternating signs is zero, as (7.207) requires. For $\mathbb{RP}^2$ we use real coefficients, $b = (1,0,0)$; the lifted pairs are in degrees $(1,2)$. We return to $\mathbb{RP}^2$ with integer coefficients in Section 7.16.

Watch out: perturbation theory cannot see the lifting

On the line with the double well $f = \frac{1}{4}(x^2-1)^2$ (Example 7.74) the local analysis gives three approximate zero modes: 0-forms at $x = \pm 1$ and a 1-form at $x = 0$. Only one exact zero mode exists. The other two sit at an energy of about 5×10^{-3} at $t = 16$, which is $\sim e^{-t/2}$. This is smaller than any power of $1/t$, so no order of perturbation theory around the wells can produce it: the anharmonic corrections are power series in $t^{-1/2}$ and never lift a supersymmetric zero mode. The lifting is *non-perturbative*. It comes from tunnelling, and it needs a different tool, the instanton.

Physical picture: why tunnelling must pair a p-state with a $(p+1)$-state

A single approximate zero mode cannot be lifted on its own: positive-energy states come in pairs related by d_t or d_t^*, which change the degree by exactly one. So lifting always takes one low-lying state of degree p and one of degree $p+1$ together, located at critical points whose indices differ by one. The amplitude for this involves the matrix element $\langle b|\mathrm{d}_t|a\rangle$ between the two local states, and that is a tunnelling amplitude from one critical point to the other. The next two sections compute it.

Reading Nash (pp. 210–211, (7.78)–(7.80))

Nash expands the metric "to make it flat to order two in the coordinates" and writes f in the form (7.78) with $\lambda_i = \mp 1$. His (7.79) is our (7.197) and his (7.80), $t\sum_i\{(1+2n_i)+\lambda_i\varepsilon_i\}$, is (7.200). He then argues exactly as we did in (7.201): a zero eigenvalue needs all $n_i = 0$ and "$\varepsilon_i = +1$ precisely when $\lambda_i = -1$", giving "a single zero eigenvalue corresponding to a p-form where p is the index of the critical point". His conclusion $b_p \le m_p$ is (7.206); the passage from the local oscillators to the global spectrum, which he leaves implicit, is Section 7.14.5. Nash ends with "this is as far as one gets with the present asymptotic analysis in t"; the strong inequalities need the complex and the tunnelling of the next two sections.

Recap

- $h_\omega = -\partial_x^2 + \omega^2 x^2$ has levels $\omega(2n+1)$ and Gaussian ground state $e^{-\omega x^2/2}$ of width $\sim \omega^{-1/2}$.
- Near a critical point with Hessian eigenvalues λ_i, $H_t \approx \sum_i\{-\partial_i^2 + t^2\lambda_i^2 x_i^2 + t\lambda_i[a^{i*}, a^i]\}$, with spectrum $t\sum_i\{|\lambda_i|\,(1+2n_i)+\lambda_i\varepsilon_i\}$.
- The unique zero mode is a Gaussian times the wedge of the downhill directions: degree = index. The gap above it is $2t\min|\lambda|$.
- Globally, $\Delta_p(t)$ has exactly m_p eigenvalues below ct at large t (trial states for $\ge$, IMS localisation for $\le$). Hence $b_p \le m_p$, and supersymmetric pairing gives $\sum(-1)^p m_p = \chi(M)$.
- The discrepancy $m_p - b_p$ is due to non-perturbative tunnelling, which lifts states in pairs of adjacent degree.

7.15 Instantons: tunnelling between critical points

The harmonic-oscillator analysis gave us m_p low-lying p-forms, one per critical point of index p. Only b_p of them are true zero modes. The others have been pushed up to small positive energies, by an amount that no order of perturbation theory can see (Section 7.14.6). The question of this section is: *what pushes them up, and by how much?*

The answer is quantum tunnelling. A particle sitting at one critical point can leak through the potential barrier $t^2\,|\nabla f|^2$ to a neighbouring critical point. The amplitude for this is exponentially small in t, and we will compute it twice: first by a direct and elementary calculation in one dimension, then by Witten's route, the Euclidean path integral, where the tunnelling paths turn out to be the gradient flow lines of f [Nash pp. 212–214, (7.83), (7.87)–(7.93)]. The second route is the one that generalises, and it is the physical origin of the Morse–Witten differential of Section 7.16.

7.15.1 What has to be computed

d_t preserves energy. The deformed exterior derivative commutes with the Hamiltonian:

$$H_t\,\mathrm{d}_t \overset{(1)}{=} \big(\mathrm{d}_t\mathrm{d}_t^* + \mathrm{d}_t^*\mathrm{d}_t\big)\mathrm{d}_t \overset{(2)}{=} \mathrm{d}_t\mathrm{d}_t^*\mathrm{d}_t \overset{(3)}{=} \mathrm{d}_t\big(\mathrm{d}_t^*\mathrm{d}_t + \mathrm{d}_t\mathrm{d}_t^*\big) = \mathrm{d}_t\,H_t. \tag{7.208}$$

(1) Definition (7.176) of H_t.
(2) $\mathrm{d}_t^2 = 0$ kills the second term.
(3) Add $\mathrm{d}_t\mathrm{d}_t\mathrm{d}_t^* = 0$, again because $\mathrm{d}_t^2 = 0$.

So if $H_t\psi = E\psi$ then $H_t(\mathrm{d}_t\psi) = E\,(\mathrm{d}_t\psi)$: d_t maps each eigenspace of H_t into itself, raising the degree by one. In particular it maps the **low-lying space**

$$V_p := \mathrm{span}\{\text{eigenforms of } \Delta_p(t) \text{ with eigenvalue} \le ct\}, \qquad \dim V_p = m_p, \tag{7.209}$$

of (7.204) into V_{p+1}. The whole question of which low-lying states are lifted is therefore a question about a *finite* matrix: the matrix of $\mathrm{d}_t : V_p \to V_{p+1}$.

The matrix elements. To leading order, V_p is spanned by the local zero modes $|a\rangle$ of (7.202), one for each critical point a of index p. So we need

$$\langle b|\,\mathrm{d}_t\,|a\rangle, \qquad \mathrm{index}(a) = p, \quad \mathrm{index}(b) = p+1, \tag{7.210}$$

which is Nash's (7.87) [Nash p. 213]. Only critical points whose indices differ by exactly one can appear: $|a\rangle$ is a p-form, $\mathrm{d}_t|a\rangle$ is a $(p+1)$-form, and it can only overlap with local states of degree $p+1$.

Once we know these numbers, the lifted energies follow. Write D_p for the $m_{p+1}\times m_p$ matrix $\langle b|\mathrm{d}_t|a\rangle$. On V_p the Hamiltonian is $H_t = \mathrm{d}_t^*\mathrm{d}_t + \mathrm{d}_t\mathrm{d}_t^*$, so in the (approximately orthonormal) basis of local states

$$H_t\big|_{V_p} \approx D_p^{\mathsf{T}}D_p + D_{p-1}D_{p-1}^{\mathsf{T}}. \tag{7.211}$$

The small eigenvalues are the eigenvalues of this matrix.

Why perturbation theory gives zero. Each local state is a Gaussian of width $t^{-1/2}$ centred on its critical point. d_t is a differential operator, so it does not move a wavefunction anywhere: $\mathrm{d}_t|a\rangle$ is concentrated where $|a\rangle$ is. The overlap $\langle b|\mathrm{d}_t|a\rangle$ therefore comes entirely from the *tails* of the two wavefunctions, far from both critical points. Tails of Gaussians are exponentially

small, so the answer is of the form $e^{-t\,(\text{something})}$, and every power series in $1/t$ misses it. This is exactly the structure of the tunnelling splitting in a double-well potential, which is the standard example of a non-perturbative effect in quantum mechanics.

Physical picture: tunnelling lifts would-be ground states

Think of the symmetric double well of elementary quantum mechanics. Each well alone has a ground state. Tunnelling mixes the two, and the combinations (even and odd) are split by $\Delta E \propto e^{-S_0/\hbar}$, where S_0 is the action of the classical path through the inverted potential. Witten's situation is the same, with one twist: supersymmetry forces the lower of the two combinations to stay at *exactly* zero when it is a true cohomology class, and the upper one moves up. What decides which combinations survive is the topology of M, via the signs of the tunnelling amplitudes.

7.15.2 A one-dimensional calculation without path integrals

In one dimension the amplitude (7.210) can be computed in a few lines, using only the exact zero-mode equations of Section 7.12.4. This calculation already contains every feature of the general answer: the exponential, the prefactor and the sign.

Let M be the line or the circle, with coordinate x and metric $\mathrm{d}x^2$. Let a be a local minimum of f (index 0) and b a neighbouring local maximum (index 1), with no critical point between them. Write $f''(a) > 0$ and $f''(b) < 0$.

Step 1: WKB states. Rather than the Gaussians (7.202), which are only approximate solutions, we use functions that solve the zero-mode equations *exactly* on a large region. Recall from Section 7.12.4 that e^{-tf} solves $\mathrm{d}_t\phi = 0$ and $e^{tf}\,\mathrm{d}x$ solves $\mathrm{d}_t^*\omega = 0$. So we set

$$\psi_a = N_a\,\chi_a(x)\,e^{-t(f(x)-f(a))}, \qquad \psi_b = N_b\,\chi_b(x)\,e^{+t(f(x)-f(b))}\,\mathrm{d}x. \tag{7.212}$$

Here χ_a is a smooth cut-off that equals 1 on the basin of a except near its edges and falls to 0 just before the neighbouring maxima; likewise χ_b equals 1 on the interval between the two minima adjacent to b, except near its ends. (Without cut-offs the functions would blow up away from their own critical point, since $e^{-t(f-f(a))}$ grows where $f < f(a)$.) Both states are sharply peaked: ψ_a at a (where $f - f(a)$ is smallest) and ψ_b at b (where $f - f(b)$ is largest, namely 0). See Figure 7.57.

Step 2: normalisation. Near a, $f(x) - f(a) \approx \tfrac{1}{2}f''(a)(x-a)^2$, so

$$1 = \|\psi_a\|^2 \overset{(1)}{=} N_a^2 \int e^{-2t(f(x)-f(a))}\,\mathrm{d}x \overset{(2)}{=} N_a^2 \int_{-\infty}^{\infty} e^{-tf''(a)(x-a)^2}\,\mathrm{d}x \overset{(3)}{=} N_a^2 \sqrt{\frac{\pi}{tf''(a)}},$$

$$\Longrightarrow \quad N_a = \left(\frac{tf''(a)}{\pi}\right)^{1/4}, \qquad \text{and likewise} \quad N_b = \left(\frac{t\,|f''(b)|}{\pi}\right)^{1/4}. \tag{7.213}$$

(1) Definition (7.212); the cut-off is 1 wherever the integrand is not exponentially small.
(2) Laplace's method: at large t only a neighbourhood of width $t^{-1/2}$ around a contributes, and there f is quadratic (errors are $O(1/t)$ relative).
(3) The Gaussian integral $\int e^{-\alpha y^2}\mathrm{d}y = \sqrt{\pi/\alpha}$ with $\alpha = tf''(a)$.

These are the same normalisations as the oscillator ground states (7.194) with $\omega = t\,|f''|$, as they must be.

Step 3: apply d_t. The exponential in ψ_a is killed by d_t exactly; only the cut-off is differentiated:

$$\mathrm{d}_t\psi_a \overset{(1)}{=} N_a\, e^{-tf}\, \mathrm{d}\big(e^{tf}\,\chi_a e^{-tf}\, e^{tf(a)}\big) \overset{(2)}{=} N_a\, e^{tf(a)}\, e^{-tf}\, \mathrm{d}\chi_a \overset{(3)}{=} N_a\, \chi_a'(x)\, e^{-t(f(x)-f(a))}\, \mathrm{d}x. \tag{7.214}$$

(1) $\mathrm{d}_t = e^{-tf}\,\mathrm{d}\,e^{tf}$, (7.170).
(2) The factors $e^{tf}\,e^{-tf}$ cancel inside; $e^{tf(a)}$ is a constant.
(3) $\mathrm{d}\chi_a = \chi_a'\mathrm{d}x$, and recombine the exponentials.

So $\mathrm{d}_t\psi_a$ is supported only where the cut-off varies: near the maxima that bound the basin of a. This is the precise sense in which tunnelling is a boundary (tail) effect.

Step 4: the overlap. Now take the inner product with ψ_b. The inner product of two 1-forms $u\,\mathrm{d}x, v\,\mathrm{d}x$ is $\int uv\,\mathrm{d}x$.

$$\begin{aligned}\langle b|\mathrm{d}_t|a\rangle &\overset{(1)}{=} N_aN_b \int \chi_b\,\chi_a'\, e^{t(f-f(b))}\, e^{-t(f-f(a))}\,\mathrm{d}x \overset{(2)}{=} N_aN_b\, e^{-t(f(b)-f(a))} \int \chi_b\,\chi_a'\,\mathrm{d}x \\ &\overset{(3)}{=} N_aN_b\, e^{-t(f(b)-f(a))}\,\big[\chi_a\big]_{\text{from } a \text{ to } b} \overset{(4)}{=} \varepsilon_{ba}\, N_aN_b\, e^{-t(f(b)-f(a))}.\end{aligned} \tag{7.215}$$

(1) (7.212) and (7.214); the states are real.
(2) The x-dependent exponentials cancel *exactly*: $e^{tf}e^{-tf} = 1$. What remains is the constant $e^{-tf(b)}e^{tf(a)}$.
(3) $\chi_b = 1$ on the whole region between a and b where χ_a falls from 1 to 0 (and $\chi_b\chi_a' = 0$ elsewhere), so the integral is the total change of χ_a along the segment from a towards b, oriented by $\mathrm{d}x$.
(4) If b lies to the right of a, going from a to b means increasing x and χ_a drops from 1 to 0: the integral is -1. If b lies to the left, the same drop is traversed in the direction of decreasing x, and $\int \chi_a'\mathrm{d}x = +1$. So $\varepsilon_{ba} = -1$ if $b > a$ and $+1$ if $b < a$.

The sign has a clean description. The downhill flow from b to a moves in the direction $+\partial_x$ when a is to the right of b ($\varepsilon = +1$) and $-\partial_x$ when a is to the left ($\varepsilon = -1$). In other words

$$\varepsilon_{ba} = \iota_{\dot\gamma}(\mathrm{d}x)\,/\,|\dot\gamma|\,,$$

where $\dot\gamma$ is the velocity of the descending flow line leaving b and $\mathrm{d}x$ is the orientation we chose for the local 1-form state at b. This is exactly Nash's orientation rule [Nash p. 213], which we develop in general in Section 7.16.4.

Key idea: the one-dimensional tunnelling amplitude

For a minimum a and an adjacent maximum b,

$$\langle b|\,\mathrm{d}_t\,|a\rangle = \varepsilon_{ba}\sqrt{\frac{t}{\pi}}\,\big(f''(a)\;|f''(b)|\big)^{1/4}\, e^{-t\,(f(b)-f(a))}\big(1+O(t^{-1})\big), \qquad \varepsilon_{ba} = \pm 1. \tag{7.216}$$

The exponent is t times the height of the barrier between the two critical points. The sign records the direction of the flow line from b down to a.

Two remarks. First, the computation never needed to know the shape of the cut-offs: all that mattered was that χ_a drops from 1 to 0 somewhere between a and b. The integral of χ_a' is "topological", a boundary term. Second, the product $e^{t(f-f(b))}e^{-t(f-f(a))}$ is constant along the whole segment: the two exponentials are *perfectly* matched, and this is why the answer is so clean. In the path-integral language of the next subsections, the constant $f(b) - f(a)$ will reappear as the instanton action divided by t.

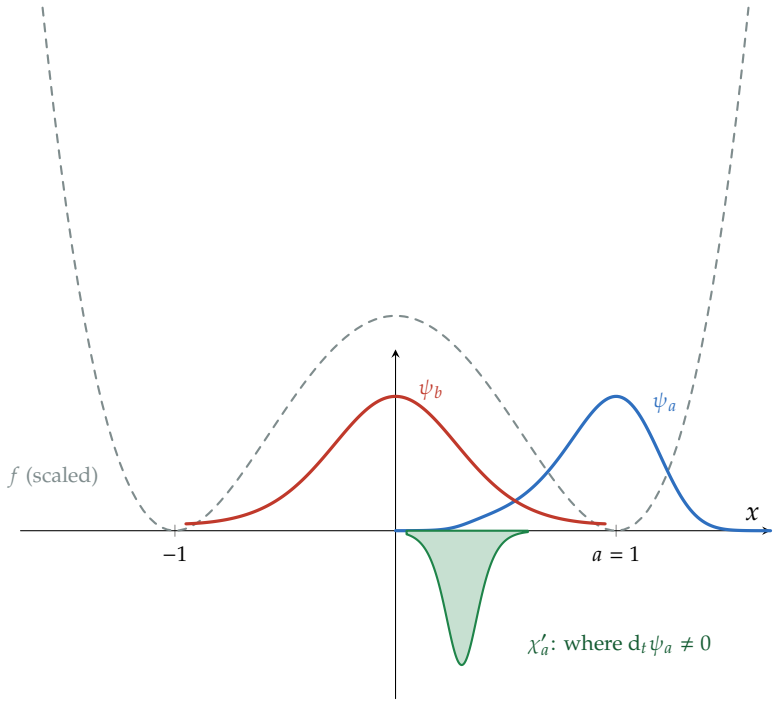


Figure 7.57: The one-dimensional overlap calculation for the double well $f = \frac{1}{4}(x^2-1)^2$ (only the right half of the story is drawn). The WKB state $\psi_a \propto e^{-t(f-f(a))}$ (blue) peaks at the minimum $a = 1$ and is cut off before the maximum; $\psi_b \propto e^{t(f-f(b))}$ (red) peaks at the maximum $b = 0$. $\mathrm{d}_t\psi_a$ is nonzero only where the cut-off varies (green). There ψ_b is not small, and the product of the two exponentials is the constant $e^{-t(f(b)-f(a))}$: the tunnelling amplitude is this constant times the total drop of the cut-off.

Worked example 7.82: the double well on the line: the splitting

Take $f = \frac{1}{4}(x^2-1)^2$ as in Example 7.74: minima $a_\pm = \pm 1$ with $f'' = 2$, $f = 0$; maximum $b = 0$ with $f'' = -1$, $f = \frac{1}{4}$. The low-lying space is $V_0 = \mathrm{span}(\psi_{a_-}, \psi_{a_+})$ and $V_1 = \mathrm{span}(\psi_b)$. By (7.216), with

$$A := \sqrt{\tfrac{t}{\pi}}\,(2\cdot 1)^{1/4} e^{-t/4},$$

the flow line from b to a_+ moves right ($\varepsilon = +1$) and the one from b to a_- moves left ($\varepsilon = -1$). So the 1×2 matrix of d_t and the Hamiltonian on V_0 are

$$D_0 = A\begin{pmatrix}-1 & 1\end{pmatrix}, \qquad H_t\big|_{V_0} \overset{(1)}{=} D_0^{\mathsf{T}} D_0 \overset{(2)}{=} A^2\begin{pmatrix}1 & -1\\ -1 & 1\end{pmatrix}, \qquad H_t\big|_{V_1} \overset{(3)}{=} D_0 D_0^{\mathsf{T}} = 2A^2.$$

(1) (7.211) with $p = 0$ (there is no D_{-1}).
(2) Multiply the column $(-1,1)^{\mathsf{T}}$ by the row $(-1,1)$.
(3) (7.211) with $p = 1$ (there is no D_1 on the line).

The eigenvalues on V_0 are 0, with eigenvector $\psi_{a_-} + \psi_{a_+}$ (the exact zero mode e^{-tf}, which has a bump at each minimum), and

$$E_{\text{lift}} = 2A^2 = 2\sqrt{2}\,\frac{t}{\pi}\,e^{-t/2}, \tag{7.217}$$

with eigenvector $\psi_{a_+} - \psi_{a_-}$. On V_1 the single level is the same $2A^2$: the lifted boson and the lifted fermion are superpartners, as they must be. Compare with a direct numerical diagonalisation of (7.181) (finite differences on $[-2.4, 2.4]$):

t	lifted 0-form level	lifted 1-form level	$2\sqrt{2}\,(t/\pi)e^{-t/2}$	ratio
8	1.191×10^{-1}	1.191×10^{-1}	1.319×10^{-1}	0.903
12	2.485×10^{-2}	2.485×10^{-2}	2.678×10^{-2}	0.928
16	4.574×10^{-3}	4.575×10^{-3}	4.832×10^{-3}	0.947
20	7.821×10^{-4}	7.833×10^{-4}	8.175×10^{-4}	0.957

The ratio tends to 1 like $1-O(1/t)$, the expected size of the corrections to Laplace's method. The next level up (not shown) is at $\approx 2t$ (for example 27.9 at $t=16$), the oscillator gap of the minima.

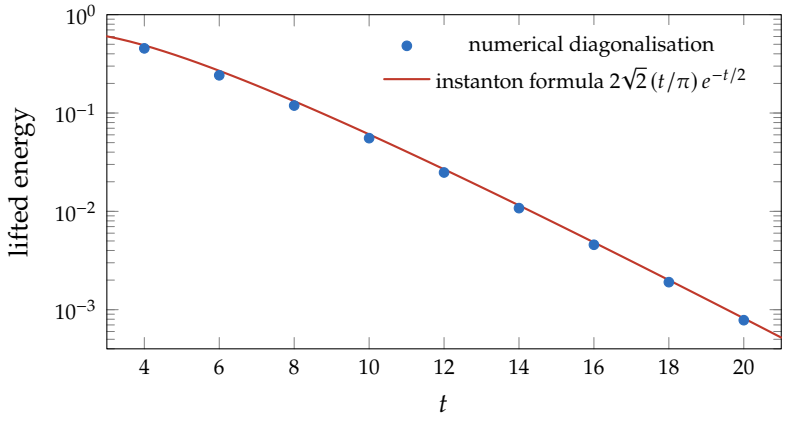


Figure 7.58: The tunnelling splitting of the line double well on a logarithmic scale. The straight-line behaviour is the factor $e^{-t/2}=e^{-2S_0}$ with $S_0=t\,(f(b)-f(a))=t/4$: the energy is the *square* of an amplitude e^{-S_0}. The prefactor t/π is what the Gaussian normalisations contribute. Agreement improves steadily with t.

Worked example 7.83: the circle with two wells, $f=-\cos 2\theta$

On S^1 with $f=-\cos 2\theta$ there are minima $a_1=0$, $a_2=\pi$ ($f=-1$, $f''=4$) and maxima $b_1=\frac{\pi}{2}$, $b_2=\frac{3\pi}{2}$ ($f=1$, $f''=-4$). The barrier height is $f(b)-f(a)=2$, so each amplitude has modulus

$$A=\sqrt{\tfrac{t}{\pi}}\,(4\cdot4)^{1/4}e^{-2t}=2\sqrt{\tfrac{t}{\pi}}\,e^{-2t}.$$

Each minimum is joined to both maxima. Reading off the directions of the descending flow lines (from $b_1=\frac{\pi}{2}$ to $a_1=0$: leftwards, $\varepsilon=-1$; from $b_2=-\frac{\pi}{2}$ to a_1: rightwards, $\varepsilon=+1$; and the reverse for a_2):

$$D_0=A\begin{pmatrix}-1 & +1\\ +1 & -1\end{pmatrix}\begin{matrix}\leftarrow b_1\\ \leftarrow b_2,\end{matrix}\qquad H_t\big|_{V_0}\overset{(1)}{=}D_0^{\mathsf T}D_0\overset{(2)}{=}2A^2\begin{pmatrix}1 & -1\\ -1 & 1\end{pmatrix}.$$

(1) (7.211), $p=0$; the columns of D_0 are labelled by a_1,a_2.
(2) Matrix multiplication: each column has squared length $2A^2$ and the two columns have inner product $-2A^2$.

The eigenvalues are 0 (eigenvector a_1+a_2, the exact zero mode $e^{t\cos 2\theta}$) and

$$E_{\text{lift}}=4A^2=\frac{16\,t}{\pi}\,e^{-4t}.$$

At $t = 2$ this is 3.42×10^{-3}; the numerical value from Figure 7.56 is 3.18×10^{-3}. At $t = 3$ it is 9.39×10^{-5} against 8.97×10^{-5}. The 1-form sector $D_0 D_0^{\mathsf{T}}$ has the same two eigenvalues 0 and $4A^2$: its zero mode $b_1 + b_2$ is the exact harmonic 1-form $e^{-t\cos 2\theta}\mathrm{d}\theta$ at large t, and $b_1 - b_2$ is the superpartner of the lifted boson.

Worked example 7.84: the circle with $f = \cos\theta$: cancelling instantons

Now take $f = \cos\theta$: one maximum $b = 0$ and one minimum $a = \pi$. There are *two* descending flow lines from b to a: one going through $\theta \in (0, \pi)$, moving in the direction $+\partial_\theta$ ($\varepsilon = +1$), and one through $\theta \in (-\pi, 0)$, moving in the direction $-\partial_\theta$ ($\varepsilon = -1$). In the overlap calculation, χ_a now drops from 1 to 0 twice, once near $\theta = 0^+$ and once near $\theta = 0^- = 2\pi^-$, in opposite directions. So

$$\langle b|\mathrm{d}_t|a\rangle = (+1 - 1)\sqrt{\tfrac{t}{\pi}}\, e^{-2t} = 0.$$

Nothing is lifted, in agreement with Example 7.73, where we found exact zero modes $e^{-t\cos\theta}$ and $e^{t\cos\theta}\mathrm{d}\theta$. The two tunnelling paths interfere destructively. This is the simplest example of the cancellations that make the Morse–Witten complex compute cohomology: $b_0 = b_1 = 1$ because the two ways round the circle carry opposite signs.

7.15.3 The Euclidean path integral, briefly

In higher dimensions the local states are not simple exponentials of f, and the direct calculation becomes hard. Witten computes $\langle b|\mathrm{d}_t|a\rangle$ instead by the semiclassical evaluation of a path integral. We recall what that means.

Imaginary time. For a particle with Hamiltonian $H = \tfrac{1}{2}p^2 + U(x)$, the amplitude to go from x_i to x_f in imaginary time T (with $\hbar = 1$) is

$$\langle x_f|e^{-TH}|x_i\rangle = \int_{x(-T/2)=x_i}^{x(T/2)=x_f} \mathcal{D}x\; e^{-S_E[x]}, \qquad S_E[x] = \int_{-T/2}^{T/2} \Big(\tfrac{1}{2}\dot{x}^2 + U(x)\Big)\mathrm{d}s. \tag{7.218}$$

The derivation is the same as in real time. Slice T into N steps $\epsilon = T/N$ and insert complete sets of position states. For one slice,

$$\langle x'|e^{-\epsilon H}|x\rangle \overset{(1)}{=} \langle x'|e^{-\epsilon p^2/2}\, e^{-\epsilon U(\hat{x})}|x\rangle\big(1 + O(\epsilon^2)\big) \overset{(2)}{=} \int \frac{\mathrm{d}p}{2\pi}\, e^{\mathrm{i}p(x'-x)}\, e^{-\epsilon p^2/2}\, e^{-\epsilon U(x)}$$
$$\overset{(3)}{=} \frac{1}{\sqrt{2\pi\epsilon}}\, \exp\!\Big[-\epsilon\Big(\tfrac{1}{2}\Big(\frac{x'-x}{\epsilon}\Big)^2 + U(x)\Big)\Big].$$

(1) Split the exponential (the commutator is higher order in ϵ).
(2) $U(\hat{x})$ acts on $|x\rangle$ by the number $U(x)$; insert $\int \mathrm{d}p\, |p\rangle\langle p|$ with $\langle x'|p\rangle = e^{\mathrm{i}px'}/\sqrt{2\pi}$.
(3) Gaussian integral over p, completing the square.

The exponent is $-\epsilon$ times the Euclidean Lagrangian $\tfrac{1}{2}\dot{x}^2 + U$ at one time step; multiplying the slices and letting $\epsilon \to 0$ gives (7.218). The sign in front of U is $+$, unlike the real-time Lagrangian $\tfrac{1}{2}\dot{x}^2 - U$: going to imaginary time flips the sign of the kinetic term relative to the potential.

Semiclassics. When the action is large, the integral is dominated by the paths that minimise S_E, and

$$\langle x_f | e^{-TH} | x_i \rangle \approx \sum_{\text{minima } x_{\mathrm{cl}}} (\text{fluctuation determinant}) \times e^{-S_E[x_{\mathrm{cl}}]}.$$

A minimum of S_E satisfies the Euler–Lagrange equation

$$\ddot{x} = U'(x), \tag{7.219}$$

which is Newton's law in the *inverted* potential $-U$. It conserves the "Euclidean energy" $\frac{1}{2}\dot{x}^2 - U(x)$ (differentiate: $\dot{x}\ddot{x} - U'\dot{x} = 0$). A path that starts at a zero of U at $s = -\infty$ and ends at another zero at $s = +\infty$ has Euclidean energy 0, so along it

$$\tfrac{1}{2}\dot{x}^2 = U(x). \tag{7.220}$$

Such a finite-action solution, localised in Euclidean time, is an **instanton**. In a double well it is the path that rolls off one hill of $-U$ and comes to rest on the other.

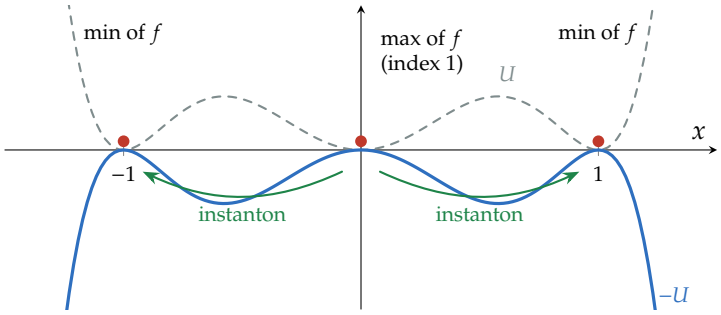


Figure 7.59: The bosonic potential $U = \frac{1}{2}t^2 f'^2$ of the double well (dashed) vanishes at *every* critical point of f, maxima included: all critical points are classical vacua. In Euclidean time the particle moves in the inverted potential $-U$ (blue), whose hilltops are the critical points. A zero-energy trajectory leaves one hilltop at $s = -\infty$ and arrives on the next at $s = +\infty$: that is an instanton. Here the instantons join the maximum of f to each minimum.

7.15.4 Witten's supersymmetric action

The Hamiltonian H_t acts on forms, i.e. on wavefunctions with fermionic degrees of freedom. Its path integral therefore needs fermionic (Grassmann) variables as well as the position $x^i(s)$. Nash quotes the Euclidean action [Nash p. 213, (7.88)]:

$$S = \frac{1}{2}\int \left\{ g_{ij}\frac{\mathrm{d}x^i}{\mathrm{d}s}\frac{\mathrm{d}x^j}{\mathrm{d}s} + t^2 g^{ij}\frac{\partial f}{\partial x^i}\frac{\partial f}{\partial x^j} + \mathrm{i}\,\bar{\psi}^i \frac{D\psi^j}{Ds} g_{ij} + \tfrac{1}{4}R_{ijkl}\bar{\psi}^i\psi^k\bar{\psi}^j\psi^l + t\,\frac{D^2 f}{Dx^i Dx^j}\left[\bar{\psi}^i, \psi^j\right] \right\} \mathrm{d}s. \tag{7.221}$$

Term by term:

- $\frac{1}{2}g_{ij}\dot{x}^i\dot{x}^j$: the kinetic energy of a particle on the Riemannian manifold M. The $x^i(s)$ are the bosonic fields (Nash: "x^i now counting as a Boson field").
- $\frac{1}{2}t^2\,|\nabla f|^2$: the bosonic potential U, the t^2 term of H_t in (7.186).
- $\bar{\psi}^i D\psi^j/Ds$: the fermion kinetic term. The $\psi^i(s)$ and $\bar{\psi}^i(s)$ are anticommuting variables carrying a tangent index, and D/Ds is the covariant derivative along the path, $D\psi^j/Ds = \dot{\psi}^j + \Gamma^j_{kl}\dot{x}^k\psi^l$. It is first order in s, as fermion kinetic terms always are. (The factor i depends on the convention used to continue the fermions to imaginary time; it drops out of everything below.)

- $\frac{1}{4}R_{ijkl}\bar\psi^i\psi^k\bar\psi^j\psi^l$: a four-fermion term built from the Riemann tensor of g. On quantisation it is the curvature part of the Hodge Laplacian on forms (the Weitzenböck term): Δ on p-forms is not just $-\nabla^2$ when M is curved.
- $t\,(D^2f/Dx^iDx^j)[\bar\psi^i,\psi^j]$: the Hessian coupled to a fermion bilinear. On quantisation it becomes the "spin" term $t\,(D^2f/Dx^iDx^j)[a^{i*},a^j]$ of (7.186).

Watch out: two misprints in (7.88)

Nash's last term reads $t\,D^2h/Dx^iDx^j$ with no fermions. The h should be f, the Morse function, and the Hessian multiplies the fermion bilinear $[\bar\psi^i,\psi^j]$ (Witten writes it this way); a Hessian term without fermions would be a t-dependent change of the bosonic potential, and could not produce the $\mp t\lambda_i$ spin energies of (7.200).

Check in one dimension. On $M=\mathbb{R}$ with flat metric the curvature term vanishes and the action becomes

$$S=\int\Big\{\tfrac12\dot x^2+\tfrac12 t^2f'(x)^2+\bar\psi\dot\psi+\tfrac12 t\,f''(x)\,[\bar\psi,\psi]\Big\}\mathrm{d}s. \tag{7.222}$$

(We have absorbed the i and the $\frac12$ of the fermion kinetic term into the normalisation of ψ.) Let us quantise it and recover H_t. A first-order term $\bar\psi\dot\psi$ means that $\bar\psi$ is the momentum conjugate to ψ; for Grassmann variables canonical quantisation imposes an *anti*commutator,

$$\{\psi,\bar\psi\}=1,\qquad \psi^2=\bar\psi^2=0,$$

realised by $\psi\mapsto a$, $\bar\psi\mapsto a^*$ on the two-dimensional fermion space $\{1,\mathrm{d}x\}$ of Section 7.11.4. The Hamiltonian is the Legendre transform: the fermion kinetic term contributes nothing (it is already of the form $p\dot q$), so

$$H\overset{(1)}{=}\tfrac12p^2+\tfrac12t^2f'^2+\tfrac12t\,f''[a^*,a]\overset{(2)}{=}\tfrac12\Big(-\frac{\mathrm{d}^2}{\mathrm{d}x^2}+t^2f'^2+t\,f''[a^*,a]\Big)\overset{(3)}{=}\tfrac12\,H_t. \tag{7.223}$$

(1) Bosonic Legendre transform of $\frac12\dot x^2$ plus the potential terms, with $[\bar\psi,\psi]\mapsto[a^*,a]$.
(2) $p=-\mathrm{i}\,\mathrm{d}/\mathrm{d}x$.
(3) Compare (7.181): $[a^*,a]=-1$ on 0-forms and $+1$ on 1-forms ((7.198)), giving the potentials $t^2f'^2\mp tf''$.

So the action (7.221) generates the evolution $e^{-sH_t/2}$. The factor $\frac12$ is a harmless rescaling of Euclidean time. The dictionary is

path integral	operators on forms
$x^i(s)$	position on M, coefficient functions of forms
$\psi^i,\bar\psi^i$	$a^i=\iota_{e_i}$, $a^{i*}=e^i\wedge$
$\{\psi^i,\bar\psi^j\}=g^{ij}$	canonical anticommutation relations (7.165)
$\bar\psi^i$ insertion	raises form degree by one
fermion number of a path	change of form degree

7.15.5 Bogomolny's trick: instantons are gradient flow lines

The instantons are the minima of the action with the fermions set to zero, because the classical fermion fields vanish. So we minimise the bosonic part [Nash p. 214, (7.89)]. Write $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ for

the metric inner product integrated over s, and ∇f for the gradient vector field ($(\nabla f)^i = g^{ij}\partial_j f$). Then

$$\begin{aligned} S_B &\overset{(1)}{=} \frac{1}{2}\left\|\frac{\mathrm{d}x}{\mathrm{d}s}\right\|^2 + \frac{t^2}{2}\,\|\nabla f\|^2 \\ &\overset{(2)}{=} \frac{1}{2}\left\|\frac{\mathrm{d}x}{\mathrm{d}s} + t\,\nabla f\right\|^2 - t\left\langle \frac{\mathrm{d}x}{\mathrm{d}s}, \nabla f\right\rangle \\ &\overset{(3)}{=} \frac{1}{2}\left\|\frac{\mathrm{d}x}{\mathrm{d}s} + t\,\nabla f\right\|^2 - t\int_{-\infty}^{\infty} \frac{\mathrm{d}}{\mathrm{d}s} f\big(x(s)\big)\,\mathrm{d}s \\ &\overset{(4)}{=} \frac{1}{2}\left\|\frac{\mathrm{d}x}{\mathrm{d}s} + t\,\nabla f\right\|^2 + t\big(f(x(-\infty)) - f(x(+\infty))\big). \end{aligned} \tag{7.224}$$

(1) The bosonic terms of (7.221), with $g^{ij}\partial_i f\,\partial_j f = |\nabla f|^2$.
(2) Complete the square: $\frac{1}{2}|v + tw|^2 = \frac{1}{2}|v|^2 + \frac{t^2}{2}|w|^2 + t\langle v, w\rangle$.
(3) The chain rule: $\langle \dot{x}, \nabla f\rangle = g_{ij}\dot{x}^i g^{jk}\partial_k f = \dot{x}^k \partial_k f = \mathrm{d}f(x(s))/\mathrm{d}s$.
(4) The fundamental theorem of calculus.

The last term depends only on the endpoints. For a path from $x(-\infty) = b$ to $x(+\infty) = a$ it is $t\,(f(b) - f(a))$, and the first term is non-negative. Hence

$$S_B \;\geq\; t\,\big(f(b) - f(a)\big), \qquad \text{with equality iff} \qquad \frac{\mathrm{d}x}{\mathrm{d}s} = -t\,\nabla f(x). \tag{7.225}$$

This is Nash's (7.90)–(7.91). Absorbing t into s (i.e. writing $\sigma = ts$), the condition becomes $\mathrm{d}x/\mathrm{d}\sigma = -\nabla f$, the equation of steepest descent (7.83), $\mathrm{d}\gamma/\mathrm{d}\sigma = -\operatorname{grad} f(\gamma)$.

Which way round? The bound (7.225) is useful only if $f(b) > f(a)$: the path must go *downhill*. For a path going uphill, from a to b, complete the square the other way, $\frac{1}{2}|\dot{x} - t\nabla f|^2 + t\langle \dot{x}, \nabla f\rangle$, and get $S_B \geq t\,(f(b) - f(a))$ again, with equality iff $\dot{x} = +t\nabla f$: an ascending flow line. This is Nash's remark "by changing f to $-f$, if necessary, we can assume that the inner product $-\langle \mathrm{d}x/\mathrm{d}s, \partial f/\partial x\rangle$ is positive". Either way,

Key idea: instantons = gradient flow lines

The paths of least Euclidean action joining two critical points a and b are the gradient flow lines of f between them, traversed downhill ($\dot{x} = -t\nabla f$) or uphill ($\dot{x} = +t\nabla f$) according to the direction of travel. Every such path has the same action,

$$S_{\text{inst}} = t\;|f(b) - f(a)|\,, \tag{7.226}$$

t times the height difference. It depends on nothing but the critical values.

Two checks. (i) The first-order equation $\dot{x} = \mp t\nabla f$ implies the zero-energy condition (7.220): $\frac{1}{2}|\dot{x}|^2 = \frac{1}{2}t^2|\nabla f|^2 = U$. So the flow lines are indeed the zero-energy solutions of Newton's equation in $-U$. (ii) On a flow line the kinetic and potential terms are equal, so $S_B = \int |\dot{x}|^2\,\mathrm{d}s = \int \dot{x}\cdot(-t\nabla f)\,\mathrm{d}s = -t\int \mathrm{d}f = t(f(b) - f(a))$, as it should.

Watch out: the second-order equation has more solutions

Newton's equation in $-U$ is second order and has many solutions joining the two critical points only if we allow nonzero Euclidean energy or infinite time. Bogomolny's trick selects, among all paths with the given endpoints at $s = \pm\infty$, those satisfying the *first-order*

flow equation. This is the finite-dimensional ancestor of the (anti-)self-duality equations for Yang–Mills instantons, which arise from exactly the same completion of a square (Chapter 8).

Reading Nash (pp. 212–214, (7.83), (7.89)–(7.91))

Nash first introduces the paths of steepest descent (7.83) as a piece of mathematics, needed to define δ, and only later shows (7.90) that they are the instantons. He calls $\mathrm{d}\gamma/\mathrm{d}s = -\operatorname{grad} f$ the "descending gradient flow" and notes that the paths begin and end at critical points because $\operatorname{grad} f$ vanishes there. In (7.89) he writes $\partial f/\partial x$ for the gradient; with a non-flat metric it is $\nabla f = g^{ij}\partial_j f\,\partial_i$. His (7.91) is the action (7.226).

7.15.6 Explicit instantons

The flow equation is first order, so in simple cases we can solve it in closed form. Seeing the solutions makes the abstract statements concrete.

Worked example 7.85: the kink of the double well

For $f = \tfrac{1}{4}(x^2-1)^2$ on $\mathbb{R}$, $f' = x^3 - x$, the downhill flow from the maximum $b = 0$ to the minimum $a = 1$ is $\dot{x} = -t(x^3 - x) = t\,x(1-x^2)$. Put $y = x^2$:

$$\dot{y} \overset{(1)}{=} 2x\dot{x} \overset{(2)}{=} 2t\,x^2(1-x^2) = 2t\,y(1-y) \quad\Longrightarrow\quad y(s) \overset{(3)}{=} \frac{1}{1+e^{-2t(s-s_0)}}.$$

(1) Chain rule.
(2) The flow equation.
(3) The logistic equation $\dot{y} = ky(1-y)$ has solution $1/(1+e^{-k(s-s_0)})$; here $k = 2t$, and s_0 is a free constant.

So

$$x_{\mathrm{cl}}(s) = \left(1 + e^{-2t(s-s_0)}\right)^{-1/2}, \tag{7.227}$$

which goes from 0 at $s = -\infty$ to 1 at $s = +\infty$ (Figure 7.60). The transition takes a Euclidean time of order $1/t$, centred at s_0. Its action is $t\,(f(0) - f(1)) = t/4$, and the splitting (7.217) is $\propto e^{-2\cdot t/4}$: the *square* of the one-instanton factor $e^{-S_{\mathrm{inst}}}$, because an energy is the square of an amplitude ((7.211)). Let us verify the action directly:

$$S_B \overset{(1)}{=} \int \dot{x}_{\mathrm{cl}}^2\,\mathrm{d}s \overset{(2)}{=} \int_0^1 \dot{x}\,\mathrm{d}x \overset{(3)}{=} \int_0^1 t\,x(1-x^2)\,\mathrm{d}x = t\left(\tfrac{1}{2} - \tfrac{1}{4}\right) = \frac{t}{4}.$$

(1) On a flow line the kinetic and potential terms are equal.
(2) Change variable from s to x, $\mathrm{d}x = \dot{x}\,\mathrm{d}s$.
(3) The flow equation expresses $\dot{x}$ as a function of x.

The mirror-image kink $-x_{\mathrm{cl}}(s)$ joins $b = 0$ to $a_- = -1$.

Worked example 7.86: the circle and the sphere: $\theta(s) = 2\arctan e^{ts}$

On the unit circle take $f = \cos\theta$, so $f' = -\sin\theta$ and the downhill flow is $\dot{\theta} = t\sin\theta$. Then

$$\frac{\mathrm{d}}{\mathrm{d}s}\tan\frac{\theta}{2} \overset{(1)}{=} \frac{\dot{\theta}}{2\cos^2(\theta/2)} \overset{(2)}{=} \frac{t\cdot 2\sin(\theta/2)\cos(\theta/2)}{2\cos^2(\theta/2)} = t\tan\frac{\theta}{2},$$

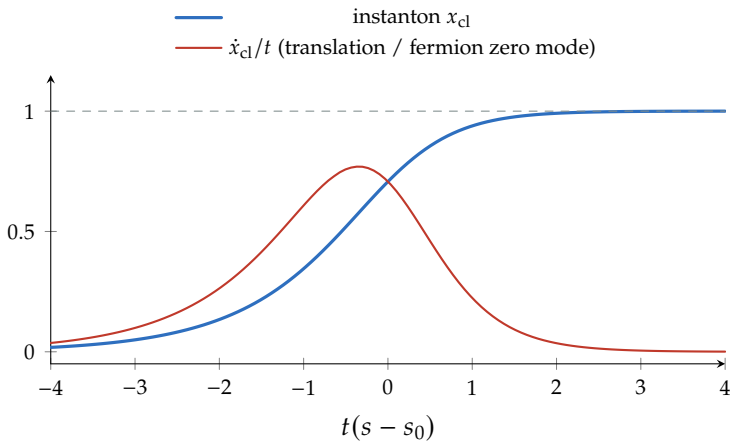


Figure 7.60: The double-well instanton (7.227) from the maximum $x = 0$ to the minimum $x = 1$ (blue), as a function of Euclidean time in units of $1/t$. It spends almost all its time sitting on one critical point or the other and switches over in a time $\sim 1/t$. Its velocity (red, scaled up by 2 for visibility) is a normalisable bump: it is the zero mode of both the bosonic and the fermionic fluctuation operators (Section 7.15.7).

(1) Chain rule, $(\tan u)' = \sec^2 u$.
(2) The flow equation and $\sin\theta = 2\sin\frac{\theta}{2}\cos\frac{\theta}{2}$.

so $\tan(\theta/2) = \pm e^{t(s-s_0)}$ and

$$\theta_\pm(s) = \pm 2\arctan e^{t(s-s_0)}.$$

Both start at the maximum $\theta = 0$ at $s = -\infty$; θ_+ arrives at $\theta = \pi$, θ_- at $\theta = -\pi$, the same point, the minimum. These are the two instantons of Example 7.84, each of action $t\,(f(0) - f(\pi)) = 2t$, moving in opposite directions.

On the round unit sphere take the height $f = \cos\theta$ (polar angle θ). The metric is $\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2$, so $\nabla f = -\sin\theta\,\partial_\theta$, and the downhill flow is $\dot\theta = t\sin\theta$, $\dot\phi = 0$: the same equation. Its solutions are $\theta(s) = 2\arctan e^{t(s-s_0)}$ at *every* fixed longitude ϕ_0. So the instantons from the north pole (index 2) to the south pole (index 0) form a family with two parameters, (s_0, ϕ_0). Up to the time translation s_0 they form a circle S^1 of meridians. The dimension count "2 = index(N) − index(S)" is general (Section 7.16.3): the flow lines from b to a form a family of dimension index(b) − index(a), including the time translation. Because the indices differ by 2, these instantons do not contribute to $\langle b|\mathrm{d}_t|a\rangle$ (which needs a difference of 1); they would contribute to amplitudes of operators that change the degree by two.

Worked example 7.87: the flat torus

On $T^2 = \mathbb{R}^2/2\pi\mathbb{Z}^2$ with $f = \cos x + \cos y$: maximum $P = (0,0)$, saddles $S_1 = (\pi, 0)$ and $S_2 = (0, \pi)$, minimum $Q = (\pi, \pi)$. The downhill flow is $(\dot x, \dot y) = t(\sin x, \sin y)$, which separates. Each coordinate obeys the circle equation of Example 7.86. The instantons between critical points of adjacent index are the ones in which only one coordinate moves:

- $P \to S_1$: $y \equiv 0$, $x = \pm 2\arctan e^{t(s-s_0)}$ (two instantons);
- $P \to S_2$: $x \equiv 0$, $y = \pm 2\arctan e^{t(s-s_0)}$ (two);
- $S_1 \to Q$: $x \equiv \pi$, $y = \pm 2\arctan e^{t(s-s_0)}$ (two);
- $S_2 \to Q$: $y \equiv \pi$, $x = \pm 2\arctan e^{t(s-s_0)}$ (two).

Each has action $2t$. When both coordinates move we get the flow lines from P to Q (index difference 2), a two-parameter family $(x, y) = \big(\pm 2 \arctan e^{t(s-s_1)}, \pm 2 \arctan e^{t(s-s_2)}\big)$, of action $4t$. We will see in Example 7.91 that the two instantons in each pair carry opposite signs.

7.15.7 Fluctuations: why the determinants cancel up to a sign

The semiclassical approximation expands the path integral to second order about each instanton. For a path $x = x_{\mathrm{cl}} + \xi$ and fermions $\psi, \bar\psi$ (which are already "small"), the quadratic action is

$$S \approx S_{\mathrm{inst}} + \tfrac{1}{2}\int \xi\, L_B\, \xi \,\mathrm{d}s + \int \bar\psi\, D_F\, \psi \,\mathrm{d}s,$$

and Gaussian integration gives, schematically,

$$\text{contribution of one instanton} = \frac{\det D_F}{\sqrt{\det L_B}}\, e^{-S_{\mathrm{inst}}}. \tag{7.228}$$

This is Nash's $\det F_{ab}\, \exp[-t(f(b)-f(a))]$ [Nash p. 214, (7.92)]. A Grassmann Gaussian integral gives the determinant ($\int \mathrm{d}\bar\psi \mathrm{d}\psi\, e^{-\bar\psi D \psi} = \det D$), a bosonic one gives the inverse square root of the determinant. Let us compute the two operators in one dimension and see why the ratio is only a sign.

The two operators. From (7.222), the second variation of the bosonic action about x_{cl} is

$$L_B = -\frac{\mathrm{d}^2}{\mathrm{d}s^2} + U''(x_{\mathrm{cl}}(s)), \qquad U = \tfrac{1}{2}t^2 f'^2,$$

and the fermion operator is read off from $\bar\psi\dot\psi + t f''\bar\psi\psi$:

$$D_F = \frac{\mathrm{d}}{\mathrm{d}s} + W(s), \qquad W(s) := t\, f''(x_{\mathrm{cl}}(s)).$$

Now compute $D_F^\dagger D_F$, with $D_F^\dagger = -\mathrm{d}/\mathrm{d}s + W$:

$$\begin{aligned} D_F^\dagger D_F\, \eta &\overset{(1)}{=} \Big(-\frac{\mathrm{d}}{\mathrm{d}s} + W\Big)\big(\dot\eta + W\eta\big) \overset{(2)}{=} -\ddot\eta - \dot W\eta - W\dot\eta + W\dot\eta + W^2\eta \\ &\overset{(3)}{=} -\ddot\eta + \big(t^2 f''^2 - t f'''\, \dot x_{\mathrm{cl}}\big)\eta \overset{(4)}{=} -\ddot\eta + t^2\big(f''^2 + f' f'''\big)\eta \\ &\overset{(5)}{=} -\ddot\eta + U''(x_{\mathrm{cl}})\,\eta = L_B\,\eta. \end{aligned} \tag{7.229}$$

(1) Definitions.
(2) Product rule on $(W\eta)^{\cdot}$; the $W\dot\eta$ terms cancel.
(3) $W^2 = t^2 f''^2$ and $\dot W = t f''' \dot x_{\mathrm{cl}}$ by the chain rule.
(4) On the instanton $\dot x_{\mathrm{cl}} = -t f'$, the flow equation.
(5) $U'' = \tfrac{1}{2}t^2 (f'^2)'' = t^2(f''^2 + f' f''')$.

So *the bosonic fluctuation operator is the square of the fermionic one*: $L_B = D_F^\dagger D_F$. This is supersymmetry at work, in the same way that $H = Q^\dagger Q$ in Section 7.10. Formally $\det L_B = \det D_F^\dagger \det D_F = |\det D_F|^2$, and the ratio (7.228) is

$$\frac{\det D_F}{|\det D_F|} = \pm 1.$$

All the magnitudes cancel. What survives is a sign: Nash's "all that is left is the sign of $\det F_{ab}$".

Zero modes: one boson, one fermion. The instanton has a free parameter, its centre s_0. Differentiating the family $x_{\mathrm{cl}}(s - s_0)$ with respect to s_0 shows that $\dot{x}_{\mathrm{cl}}$ is a zero mode of L_B (moving the centre costs no action). It is also a zero mode of D_F:

$$D_F\,\dot{x}_{\mathrm{cl}} = \ddot{x}_{\mathrm{cl}} + t f''\dot{x}_{\mathrm{cl}} = \frac{\mathrm{d}}{\mathrm{d}s}\big(-t f'(x_{\mathrm{cl}})\big) + t f''\dot{x}_{\mathrm{cl}} = -t f''\dot{x}_{\mathrm{cl}} + t f''\dot{x}_{\mathrm{cl}} = 0.$$

And $D_F^\dagger$ has no normalisable zero mode: $D_F^\dagger \eta = 0$ means $\dot{\eta} = W\eta$, whose solution is $1/\dot{x}_{\mathrm{cl}}$ (check: $(1/\dot{x})^{\cdot} = -\ddot{x}/\dot{x}^2 = t f''/\dot{x}$), and this blows up at both ends. So

$$\dim \operatorname{Ker} D_F - \dim \operatorname{Ker} D_F^\dagger = 1 = \operatorname{index}(b) - \operatorname{index}(a). \tag{7.230}$$

In n dimensions the same holds with $D_F = \nabla_s + t\,\mathrm{Hess} f$ acting on vector fields along the instanton: its index is $\operatorname{index}(b) - \operatorname{index}(a)$, the number of directions that are unstable at b but stable at a.

The zero modes are handled as usual. The bosonic zero mode is traded for an integral over the collective coordinate s_0, which gives a positive Jacobian ($\sqrt{S_{\mathrm{inst}}}$ times constants). The fermionic zero mode is more interesting. A Grassmann integral over a mode that does not appear in the action vanishes, $\int \mathrm{d}\eta_0\, 1 = 0$, unless the integrand contains exactly one factor of that mode. So the instanton contributes only to amplitudes with *one* fermion insertion, i.e. to matrix elements of an operator that changes the fermion number by one. That operator is d_t (or d_t^*), and this is why instantons between critical points whose indices differ by one are precisely what $\langle b|\mathrm{d}_t|a\rangle$ sees.

Physical picture: fermion zero modes as selection rules

In gauge theory a BPST instanton carries fermion zero modes, and they make it violate a chiral charge by a definite amount: the 't Hooft vertex. Here the flow line from b to a carries $\operatorname{index}(b) - \operatorname{index}(a)$ fermion zero modes, so it changes the form degree by exactly that amount. The Morse index plays the role of the chiral charge, and "index difference $= 1$" is the selection rule for the supercharge d_t. The same mechanism (an index theorem for the linearised operator) will count dimensions of instanton moduli spaces in Chapter 8 and gradings in Floer theory (Chapter 12).

The sign. The sign of $\det D_F$ (suitably regularised and with the zero mode removed) is a "relative orientation": it compares the orientation of the fermion zero-mode space with the orientations chosen for the states $|a\rangle$ and $|b\rangle$. In one dimension we computed it directly in (7.215): $\varepsilon = \pm 1$ according to the direction of the flow. In general it is the sign $\varepsilon_{\gamma_{ab}}$ defined by Nash's orientation rule, which we work out in Section 7.16.4.

7.15.8 Assembling the amplitude

Summing over all instantons from b to a:

$$\langle b|\,\mathrm{d}_t\,|a\rangle \;\approx\; \sum_{\gamma_{ab}} \varepsilon_{\gamma_{ab}}\, e^{-t(f(b)-f(a))} \qquad (t \to \infty), \tag{7.231}$$

which is Nash's (7.93), up to a positive prefactor that we suppress (in one dimension it is $\sqrt{t/\pi}\,(f''(a)\,|f''(b)|)^{1/4}$, and with Nash's normalisation $|\lambda_i| = 1$ it is the universal $\sqrt{t/\pi}$). The sum is finite for a generic metric (Section 7.16.3). In terms of the action of d_t on the local states,

$$\mathrm{d}_t|a\rangle \approx \sum_b \Big(\sum_{\gamma_{ab}} \varepsilon_{\gamma_{ab}}\Big) e^{-t(f(b)-f(a))}|b\rangle, \tag{7.232}$$

Nash's (7.94), the sum running over critical points b of index $\operatorname{index}(a) + 1$.

The exponentials are a change of basis. Rescale each local state by a constant, $|\tilde{a}\rangle := e^{-tf(a)}|a\rangle$. Then

$$\mathrm{d}_t|\tilde{a}\rangle \overset{(1)}{=} e^{-tf(a)} \sum_b \varepsilon(a,b)\, e^{-tf(b)} e^{tf(a)}|b\rangle \overset{(2)}{=} \sum_b \varepsilon(a,b)\, e^{-tf(b)}|b\rangle \overset{(3)}{=} \sum_b \varepsilon(a,b)\,|\tilde{b}\rangle, \tag{7.233}$$

(1) (7.232), with $\varepsilon(a,b) := \sum_{\gamma_{ab}} \varepsilon_{\gamma_{ab}}$.
(2) The factors $e^{\pm tf(a)}$ cancel.
(3) Definition of $|\tilde{b}\rangle$.

In the rescaled basis, the matrix of d_t has the *integer* entries $\varepsilon(a,b)$. A change of basis does not change kernels, images or ranks, so the exponentials carry no information about cohomology. This is Witten's observation that "the exponential factors are inessential and can be discarded" [Nash p. 215]. It leads to the definition

$$\delta(a) = \sum_b \varepsilon(a,b)\, b, \tag{7.234}$$

Nash's (7.95), which is the same as the purely mathematical definition (7.86) that we give in Section 7.16. And since $\mathrm{d}_t^2 = 0$ exactly, its shadow on the low-lying space squares to zero too: $\delta^2 = 0$ "is really a consequence of $\mathrm{d}_t^2 = 0$".

Watch out: what "inessential" does and does not mean

The rescaling (7.233) removes the exponentials exactly, and a common positive prefactor (such as $\sqrt{t/\pi}$) is a harmless overall scalar. But the new basis is not orthonormal, so the rescaled matrix is not the matrix of an operator we can square with its adjoint: the lifted *energies* (Example 7.82) do depend on the exponentials. What is preserved is everything linear-algebraic: ranks, kernels, images, and hence cohomology. The step "to leading order, the matrix of d_t on the low-lying space is δ up to change of basis" is made rigorous by the semiclassical analysis of Helffer and Sjöstrand (for a metric satisfying the Morse–Smale condition of Section 7.16.3): for large t the complex $(V_\bullet, \mathrm{d}_t)$ is isomorphic to the Morse–Witten complex with real coefficients.

Reading Nash (pp. 213–215, (7.87)–(7.95))

Nash's physics argument runs: compute the amplitude (7.87); evaluate it at large t by summing over instantons, found from the action (7.88); complete the square (7.89) to identify instantons with steepest-descent paths (7.90), of action (7.91); quote WKB for the contribution (7.92), where fermionic and bosonic determinants "mutually cancel" leaving the sign $\varepsilon_{\gamma_{ab}}$; sum to get (7.93) and the action (7.94) of d_t; discard exponentials to get (7.95). We have filled in the one-dimensional computation of the amplitude with its prefactor (7.216), the canonical quantisation check (7.223), the identity $L_B = D_F^\dagger D_F$ behind the cancellation of determinants, the role of the fermion zero mode, and the explanation of why dropping the exponentials is a change of basis.

Recap

- d_t commutes with H_t, so it maps the m_p-dimensional low-lying space V_p to V_{p+1}. The lifted energies are the eigenvalues of $D^\mathsf{T}D + DD^\mathsf{T}$, where $D = \langle b|\mathrm{d}_t|a\rangle$.
- In one dimension, WKB states make the calculation exact up to Laplace-method corrections: $\langle b|\mathrm{d}_t|a\rangle = \varepsilon\sqrt{t/\pi}\,(f''(a)\,|f''(b)|)^{1/4} e^{-t(f(b)-f(a))}$. The double-well splitting

$2\sqrt{2}(t/\pi)e^{-t/2}$ matches numerics.

- Witten's action has bosons x^i, fermions $\psi^i, \bar\psi^i$ and quantises to $\frac{1}{2}H_t$. Completing the square shows its instantons are gradient flow lines, of action $t\,|f(b)-f(a)|$.
- $L_B = D_F^\dagger D_F$: the determinants cancel up to sign. The instanton has index$(b)-$index(a) fermion zero modes, so only index difference one contributes to $\langle b|\mathrm{d}_t|a\rangle$.
- $\mathrm{d}_t|a\rangle \approx \sum \varepsilon(a,b)e^{-t(f(b)-f(a))}|b\rangle$. Rescaling the basis by e^{-tf} removes the exponentials and leaves the integer matrix $\varepsilon(a,b)$: the Morse–Witten differential δ. Opposite signs cancel (the circle with $\cos\theta$), same signs add.

7.16 The Morse–Witten complex

The last section ended with a finite matrix of integers. For every pair of critical points a, b with index(b) = index$(a)+1$ we have a number $\varepsilon(a,b)$, a signed count of the gradient flow lines running down from b to a, and at large t the operator d_t acts on the low-lying states as

$$\mathrm{d}_t|\tilde a\rangle \approx \sum_b \varepsilon(a,b)\,|\tilde b\rangle$$

((7.233)). This section answers three questions about this matrix.

1. *What does it compute?* It is the differential of a cochain complex, one generator per critical point, and the cohomology of that complex is the cohomology of M. The strong Morse inequalities then follow from a line of linear algebra.
2. *How do we compute it without quantum mechanics?* By looking at the geometry of the gradient flow: stable and unstable manifolds, and Nash's orientation rule for the signs [Nash pp. 212–213, (7.84)–(7.86)].
3. *Why does it square to zero?* Physically because $\mathrm{d}_t^2 = 0$ [Nash p. 215]; geometrically because flow lines between critical points of index difference two break into pairs of flow lines of index difference one.

Along the way we compute the complex for every example of this chapter, and meet one example ($\mathbb{RP}^2$) where the complex sees more than the Betti numbers: it sees torsion.

7.16.1 Cochain complexes and the algebra of the strong inequalities

Nash says that "the additional mathematics is a result which says that if one can form a cohomology complex from the set of critical points of f then this gives the cohomology of M and also implies the strong form of the Morse inequalities" [Nash p. 211]. We first make the second half of this sentence completely explicit, because it is pure linear algebra and it is the reason the whole construction is worth doing.

Working definition 7.33 (cochain complex and its cohomology)

A (finite) **cochain complex** is a list of finite-dimensional vector spaces and linear maps

$$0 \longrightarrow C^0 \xrightarrow{\delta_0} C^1 \xrightarrow{\delta_1} C^2 \xrightarrow{\delta_2} \cdots \xrightarrow{\delta_{n-1}} C^n \longrightarrow 0$$

with $\delta_{p+1} \circ \delta_p = 0$ for every p (written $\delta^2 = 0$). To compute with it, choose a basis of each C^p and write δ_p as a matrix D_p of size $\dim C^{p+1} \times \dim C^p$; the condition is $D_{p+1} D_p = 0$. Its **cohomology** is

$$H^p(C) := \frac{\operatorname{Ker} \delta_p}{\operatorname{Im} \delta_{p-1}}, \qquad \dim H^p(C) = \dim \operatorname{Ker} D_p - \operatorname{rank} D_{p-1}.$$

[vectors $x \in C^p$ with $\delta_p x = 0$, counted as equal when they differ by some $\delta_{p-1} y$. The rank of D_{p-1} is its number of independent columns] Formally, $H^\bullet(C)$ is a sequence of quotient spaces; in practice it is "closed modulo exact", computed by Gaussian elimination.

The quotient makes sense precisely because of $\delta^2 = 0$. For $\operatorname{Ker} \delta_p / \operatorname{Im} \delta_{p-1}$ to be a quotient of a space by a subspace we need $\operatorname{Im} \delta_{p-1} \subseteq \operatorname{Ker} \delta_p$, i.e. $\delta_p(\delta_{p-1} x) = 0$ for all x. That is literally $\delta^2 = 0$. The de Rham complex $(\Omega^\bullet, \mathrm{d})$ is the infinite-dimensional example; the deformed complex $(\Omega^\bullet, \mathrm{d}_t)$ is another (Section 7.12.3); the complexes of this section are finite.

The counting identity. Suppose $\dim C^p = m_p$ and $\dim H^p(C) = b_p$. Write $r_p := \operatorname{rank} \delta_p \geq 0$ (with $r_{-1} = r_n = 0$). Then

$$b_p \overset{(1)}{=} \dim \operatorname{Ker} \delta_p - \dim \operatorname{Im} \delta_{p-1} \overset{(2)}{=} (m_p - r_p) - r_{p-1}, \tag{7.235}$$

(1) Definition of the cohomology as a quotient: the dimension of a quotient is the difference of dimensions.
(2) Rank–nullity for $\delta_p \colon C^p \to C^{p+1}$: $\dim \operatorname{Ker} \delta_p = \dim C^p - \operatorname{rank} \delta_p$; and $\dim \operatorname{Im} \delta_{p-1} = r_{p-1}$ by definition.

so $m_p - b_p = r_p + r_{p-1}$. Multiply by u^p and sum:

$$\begin{aligned} \sum_p m_p u^p - \sum_p b_p u^p &\overset{(1)}{=} \sum_p r_p u^p + \sum_p r_{p-1} u^p \overset{(2)}{=} \sum_p r_p u^p + u \sum_q r_q u^q \\ &\overset{(3)}{=} (1+u)\, R(u), \qquad R(u) := \sum_p r_p u^p. \end{aligned} \tag{7.236}$$

(1) (7.235), multiplied by u^p and summed over p.
(2) Rename $q = p - 1$ in the second sum; $u^p = u \cdot u^q$.
(3) Factor out $\sum_p r_p u^p$.

Since every r_p is a rank, $R(u)$ has non-negative coefficients. This is exactly the form (7.140) of the strong Morse inequalities.

Key idea: the strong inequalities are rank–nullity

Any cochain complex with $\dim C^p = m_p$ whose cohomology has dimensions b_p satisfies

$$M_u - P_u = (1+u) R(u), \qquad R(u) = \sum_p \operatorname{rank}(\delta_p)\, u^p,$$

with all coefficients of R non-negative. So the strong Morse inequalities follow as soon as we find *any* complex with one generator per critical point (in degree equal to the index) and cohomology $H^\bullet(M)$. The coefficient r_p counts how many index-p critical points are "cancelled" against index-$(p+1)$ critical points by the differential.

Worked example 7.88: a two-term complex: the circle with two wells

Take $C^0 = \mathbb{R}^2$, $C^1 = \mathbb{R}^2$ and $D_0 = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}$ (the matrix of Example 7.83 with the factor A removed). There is only one map, so $\delta^2 = 0$ is automatic. Then

$$\operatorname{rank} D_0 \overset{(1)}{=} 1, \qquad b_0 \overset{(2)}{=} 2 - 1 = 1, \qquad b_1 \overset{(3)}{=} 2 - 1 = 1.$$

(1) The two columns are negatives of each other and nonzero.
(2) $\operatorname{Ker} D_0$ is spanned by $(1,1)^{\mathsf{T}}$: $b_0 = m_0 - r_0$.
(3) $b_1 = m_1 - r_0$ (there is no δ_1 on a 1-manifold).

These are the Betti numbers of S^1. The identity (7.236) reads $(2+2u) - (1+u) = (1+u)\cdot 1$, so $R(u) = 1$: one cancelling pair, in degrees $(0,1)$.

7.16.2 Witten's low-lying complex

Is there such a complex? Quantum mechanics hands us one immediately, before any instanton is computed. Recall the low-lying space V_p of (7.209): the span of the eigenforms of $\Delta_p(t)$ with eigenvalue at most ct. At large t, $\dim V_p = m_p$ ((7.204)), and d_t maps V_p into V_{p+1} ((7.208)). So

$$0 \longrightarrow V_0 \xrightarrow{\mathrm{d}_t} V_1 \xrightarrow{\mathrm{d}_t} \cdots \xrightarrow{\mathrm{d}_t} V_n \longrightarrow 0 \tag{7.237}$$

is a finite cochain complex, since $\mathrm{d}_t^2 = 0$. Its dimensions are the m_p. Its cohomology is the cohomology of M. Here is why.

Positive-energy eigenspaces carry no cohomology. Let E_λ be the space of eigenforms of H_t with eigenvalue λ (all degrees). Both d_t and d_t^* commute with H_t, so each maps E_λ to itself, and $(E_\lambda, \mathrm{d}_t)$ is a complex by itself. If $\lambda > 0$ and $\alpha \in E_\lambda$ is d_t-closed, then

$$\alpha \overset{(1)}{=} \frac{1}{\lambda} H_t\, \alpha \overset{(2)}{=} \frac{1}{\lambda}\left(\mathrm{d}_t \mathrm{d}_t^* \alpha + \mathrm{d}_t^* \mathrm{d}_t \alpha\right) \overset{(3)}{=} \mathrm{d}_t\left(\frac{1}{\lambda}\, \mathrm{d}_t^* \alpha\right), \tag{7.238}$$

(1) $H_t \alpha = \lambda \alpha$ and $\lambda \neq 0$.
(2) Definition (7.176) of H_t.
(3) $\mathrm{d}_t \alpha = 0$ kills the second term.

and $\frac{1}{\lambda}\mathrm{d}_t^*\alpha$ lies in E_λ again. So every closed element of E_λ is exact *within* E_λ: the complex $(E_\lambda, \mathrm{d}_t)$ has zero cohomology. (The operator $K = \mathrm{d}_t^*/\lambda$ is a **contracting homotopy**: $\mathrm{d}_t K + K \mathrm{d}_t = \mathrm{id}$ on E_λ.) Physically this is the statement that positive-energy states come in boson–fermion pairs $(\beta, \mathrm{d}_t\beta)$ and cannot carry a supersymmetric ground state.

Putting the eigenspaces together. The low-lying complex is the direct sum of the eigenspace complexes with $\lambda \leq ct$:

$$H^p\left(V_\bullet, \mathrm{d}_t\right) \overset{(1)}{=} \bigoplus_{\lambda \leq ct} H^p\left(E_\lambda, \mathrm{d}_t\right) \overset{(2)}{=} H^p\left(E_0, \mathrm{d}_t\right) \overset{(3)}{=} \operatorname{Ker} \Delta_p(t) \overset{(4)}{\cong} H^p_{\mathrm{dR}}(M). \tag{7.239}$$

(1) d_t preserves each E_λ, so kernels and images split eigenspace by eigenspace.
(2) (7.238): only $\lambda = 0$ contributes.
(3) On E_0 we have $\mathrm{d}_t = 0$ (a zero mode of H_t satisfies $\mathrm{d}_t\alpha = \mathrm{d}_t^*\alpha = 0$, Section 7.12.3), so every element is closed and none but 0 is exact.
(4) Hodge theory for d_t, (7.177).

Key idea: Witten's first proof of the strong inequalities

At large t the low-lying states of H_t form a cochain complex $(V_\bullet, \mathrm{d}_t)$ with $\dim V_p = m_p$ and cohomology $H^\bullet_{\mathrm{dR}}(M)$. By (7.236),

$$\sum_p m_p u^p - \sum_p b_p u^p = (1+u)R(u), \qquad r_p = \operatorname{rank}\bigl(\mathrm{d}_t : V_p \to V_{p+1}\bigr) \geq 0. \tag{7.240}$$

No instanton calculation is needed for this. The instantons are needed to compute the matrix of d_t on $V_\bullet$, i.e. to *construct* the cohomology of M from f.

The physical meaning of r_p. Decompose V_p orthogonally as

$$V_p = \underbrace{\operatorname{Ker}\Delta_p(t)}_{b_p \text{ exact ground states}} \oplus \underbrace{\mathrm{d}_t V_{p-1}}_{r_{p-1} \text{ lifted}} \oplus \underbrace{\mathrm{d}_t^* V_{p+1}}_{r_p \text{ lifted}}.$$

This is the Hodge decomposition of the finite complex. A lifted state in $\mathrm{d}_t^* V_{p+1}$ is a p-form β with $\mathrm{d}_t\beta \neq 0$; its partner $\mathrm{d}_t\beta$ is a lifted $(p+1)$-form with the same energy. So r_p is the number of boson–fermion pairs that tunnelling lifts *between degrees p and $p+1$*, and the factor $(1+u)$ in the Morse inequalities is the supersymmetric pairing $u^p + u^{p+1}$ of these states. See Figure 7.61.

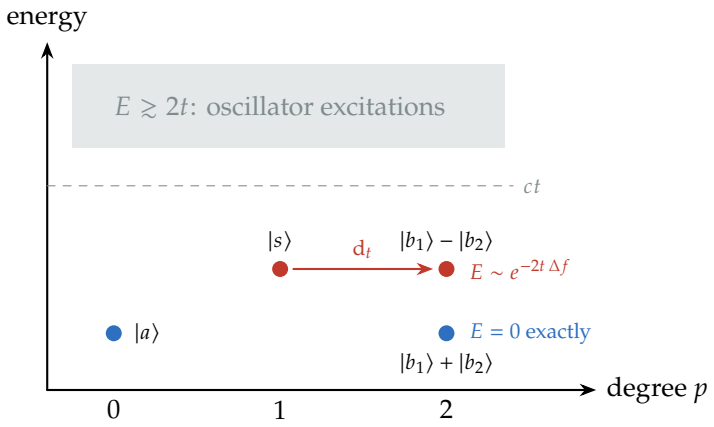


Figure 7.61: The low-lying spectrum of the two-humped sphere of Figure 7.45 at large t (schematic; energies not to scale). The four low-lying states sit one per critical point in degree equal to the index. Two (blue) are exact zero modes, $b_0 = b_2 = 1$. The saddle state and the antisymmetric combination of the two peak states (red) are superpartners, lifted together by tunnelling and joined by d_t. They are the single cancelling pair $r_1 = 1$ of $M_u - P_u = (1+u)\,u$. Everything else is at energies of order t.

Worked example 7.89: the low-lying complex on the circle with two wells

For $f = -\cos 2\theta$ on S^1 (Example 7.83) the low-lying complex is $V_0 \to V_1$ with $\dim V_0 = \dim V_1 = 2$. We computed the matrix of d_t in the basis of local states, $D_0 = A\begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}$ with $A = 2\sqrt{t/\pi}\,e^{-2t}$. Its rank is 1 for every t (the factor A is never zero), so $r_0 = 1$, $b_0 = b_1 = 1$, $R(u) = 1$, exactly as in Example 7.88. The exponentially small A does not affect the rank: this is the statement that the exponential factors are "inessential" [Nash p. 215].

Watch out: rank is robust, but only if the matrix is exactly right

The rank of a matrix is not a continuous function of its entries: a tiny perturbation can raise it. In (7.237) this causes no trouble, because d_t on $V_\bullet$ is an *exact* complex whose cohomology we know exactly. But when we replace the true matrix by its semiclassical approximation ((7.232)), we must make sure that approximation errors cannot change ranks. They cannot, because the approximate matrix is itself (after rescaling) an integer matrix with $\delta^2 = 0$ and the corrections are relatively small by powers of $1/t$; Helffer and Sjöstrand showed that for large t the two complexes are isomorphic. The upshot is that we may compute the ranks from the integer matrix $\varepsilon(a,b)$.

7.16.3 Stable and unstable manifolds, and spaces of flow lines

We now turn to the geometry that computes $\varepsilon(a,b)$ without any quantum mechanics. Nash introduces the **descending gradient flow** [Nash p. 212, (7.83)]

$$\frac{\mathrm{d}\gamma}{\mathrm{d}\sigma} = -\nabla f\big(\gamma(\sigma)\big), \qquad (\nabla f)^i = g^{ij}\partial_j f. \tag{7.241}$$

(We write σ for the flow parameter; the instanton of Section 7.15.5 is $\gamma(ts)$.) Two facts make it useful.

f decreases along the flow. Along a solution,

$$\frac{\mathrm{d}}{\mathrm{d}\sigma} f\big(\gamma(\sigma)\big) \overset{(1)}{=} \partial_i f\,\frac{\mathrm{d}\gamma^i}{\mathrm{d}\sigma} \overset{(2)}{=} -\partial_i f\, g^{ij}\partial_j f \overset{(3)}{=} -|\nabla f|^2 \;\le\; 0, \tag{7.242}$$

(1) Chain rule.
(2) The flow equation (7.241).
(3) g^{ij} is positive definite; equality only at critical points.

So f strictly decreases except at critical points. On a compact M the flow exists for all $\sigma \in \mathbb{R}$, and $f(\gamma(\sigma))$ is bounded and monotone, so it has limits as $\sigma \to \pm\infty$. Integrating (7.242) shows $\int |\nabla f|^2\,\mathrm{d}\sigma < \infty$, which forces $\gamma(\sigma)$ to approach points where $\nabla f = 0$. For a Morse function the critical points are isolated, and *every flow line starts at a critical point and ends at a critical point* (Nash: "when grad f vanishes we have a critical point and so there are critical points at the ends of the paths").

The local picture. Near a critical point, in the Morse coordinates of (7.196) with a flat metric, $f = f(c) + \tfrac{1}{2}\sum_i \lambda_i x_i^2$ and the flow is

$$\dot{x}_i = -\lambda_i x_i \quad\Longrightarrow\quad x_i(\sigma) = x_i(0)\,e^{-\lambda_i \sigma}. \tag{7.243}$$

Coordinates with $\lambda_i > 0$ decay as $\sigma \to +\infty$; those with $\lambda_i < 0$ grow. A point flows *into* c as $\sigma \to +\infty$ iff its $\lambda_i < 0$ coordinates vanish, and it came *out of* c at $\sigma = -\infty$ iff its $\lambda_i > 0$ coordinates vanish.

Working definition 7.34 (stable and unstable manifolds)

For a critical point c,

$$W^u(c) := \{x \in M :\ \text{the flow line through } x \text{ tends to } c \text{ as } \sigma \to -\infty\} \quad (\textbf{unstable}),$$
$$W^s(c) := \{x \in M :\ \text{the flow line through } x \text{ tends to } c \text{ as } \sigma \to +\infty\} \quad (\textbf{stable}).$$

[the flow line through x is the solution $\gamma(\sigma)$ of (7.241) with $\gamma(0) = x$] These are the **unstable manifold** and the **stable manifold** of c. Operationally: $W^u(c)$ is everything that can be reached by flowing *downhill* out of c, $W^s(c)$ everything that flows downhill *into* c. By (7.243), $W^u(c)$ is a disc of dimension index(c) [the number of independent downhill directions at c], tangent at c to $T_c^- M$, the span of the Hessian's negative eigenvectors; $W^s(c)$ is a disc of dimension $n - \text{index}(c)$, tangent to $T_c^+ M$.

Nash's $T_b^- M$ in (7.84), "the $(p+1)$-dimensional space corresponding to the $(p+1)$ negative eigenvalues of the Hessian Hf at b", is $T_b\, W^u(b)$. It is also the space of downhill directions whose wedge product is the local zero mode $|b\rangle$ (7.202). Every flow line leaving b lies in $W^u(b)$ and leaves b in a direction of $T_b^- M$.

Flow lines from b to a. A point lies on a flow line from b down to a iff it lies in $W^u(b) \cap W^s(a)$. So the space of *parametrised* flow lines from b to a is

$$\mathcal{M}(b,a) := W^u(b) \cap W^s(a),$$

(each point of the intersection is the value $\gamma(0)$ of one flow line). The flow parameter can be shifted, $\gamma(\sigma) \mapsto \gamma(\sigma + c)$, so the space of *unparametrised* flow lines is $\widehat{\mathcal{M}}(b,a) = \mathcal{M}(b,a)/\mathbb{R}$. Its dimension, when the intersection is transversal, is

$$\begin{aligned} \dim \widehat{\mathcal{M}}(b,a) &\overset{(1)}{=} \dim W^u(b) + \dim W^s(a) - n - 1 \\ &\overset{(2)}{=} \text{index}(b) + \big(n - \text{index}(a)\big) - n - 1 = \text{index}(b) - \text{index}(a) - 1. \end{aligned} \tag{7.244}$$

(1) Two submanifolds of dimensions k and l meeting transversally in an n-manifold meet in a submanifold of dimension $k + l - n$ (count equations: $W^s(a)$ is cut out by $n - l$ equations, and they stay independent on $W^u(b)$). Then subtract 1 for the time translation.

(2) $\dim W^u(b) = \text{index}(b)$, $\dim W^s(a) = n - \text{index}(a)$.

Key idea: finitely many flow lines between adjacent indices

If $\text{index}(b) = \text{index}(a) + 1$ and the intersection is transversal, the unparametrised flow lines from b to a form a 0-dimensional space. It is also compact (a sequence of such flow lines can only degenerate by breaking at an intermediate critical point, which would need a critical point of index strictly between index(a) and index(b)), so it is a *finite* set of points. This is why Nash's sum over "all these paths of steepest descent from b to a" in (7.85) is finite.

Working definition 7.35 (Morse–Smale condition)

The pair (f, g) [the Morse function and the Riemannian metric, both needed to define the flow of $-\nabla f$] is **Morse–Smale** if for all critical points a, b the manifolds $W^u(b)$ and $W^s(a)$ intersect transversally: at every common point their tangent spaces together span $T_x M$. Test: at a point of $W^u(b) \cap W^s(a)$, add up the dimensions of the two tangent spaces and subtract the dimension of their intersection; the result must be n. Smale's theorem: for any Morse function the condition can be achieved by an arbitrarily small change of the metric (or of f).

A consequence worth remembering: under Morse–Smale, flow lines only go *down* in index. Since $\dim \widehat{\mathcal{M}}(b,a) \geq 0$ whenever it is non-empty, a flow line from b to a forces $\text{index}(b) \geq \text{index}(a) + 1$. In particular there are no flow lines between two critical points of the same index.

Worked example 7.90: the round sphere and the flat torus

Sphere, height $f = z$. $W^u(N) = S^2 \setminus \{S\}$ (everything flows out of the north pole), $W^s(S) = S^2 \setminus \{N\}$. Their intersection is $S^2 \setminus \{N, S\}$, of dimension $2 = \text{index}(N) - \text{index}(S)$, and the unparametrised flow lines form the circle of meridians, of dimension $1 = 2 - 0 - 1$, in agreement with Example 7.86. The intersection is transversal because both are open subsets of S^2.
Flat torus, $f = \cos x + \cos y$. At the saddle $S_1 = (\pi, 0)$, $f \approx -(x-\pi)^2/2 \cdot (-1) + \dots$; precisely, $\cos x \approx -1 + \frac{1}{2}(x-\pi)^2$ and $\cos y \approx 1 - \frac{1}{2}y^2$, so the downhill direction is y and the uphill direction is x. Hence $W^u(S_1)$ is the circle $\{x = \pi\}$ minus the point Q, and $W^s(S_1)$ is the circle $\{y = 0\}$ minus P. At S_1 they cross at right angles. The intersection $W^u(P) \cap W^s(S_1) = W^s(S_1) \setminus \{S_1\}$ consists of two open arcs: the two flow lines from P to S_1 of Example 7.87, and $\dim \widehat{\mathcal{M}}(P, S_1) = 0 = 2 - 1 - 1$.

When Morse–Smale fails: the upright torus. Stand the torus on its end, like a tyre on a road, and let f be the height. Its four critical points are the top P (index 2), the top of the hole S_u and the bottom of the hole S_l (index 1), and the bottom Q (index 0). At S_u the downhill directions run along the rim of the hole, and following them we slide down the inside of the hole, left or right, and arrive at S_l. So there are two flow lines from S_u to S_l, *between critical points of the same index*. Here $W^u(S_u)$ and $W^s(S_l)$ are the same curve (the rim of the hole), and two coinciding curves in a surface are not transversal: their tangent spaces span only a line. The count (7.244) predicts $\dim \widehat{\mathcal{M}}(S_u, S_l) = 1 - 1 - 1 = -1$, i.e. generically no such flow lines at all.

Tilting the torus slightly towards us breaks the reflection symmetry that keeps the rim of the hole invariant. The two flow lines leaving S_u now miss S_l, one passing in front and one behind, and continue to the minimum Q; and S_l now receives two flow lines from the top P. Figure 7.62 shows both pictures, computed numerically on the torus of radii 2 and 1 with its induced metric. After the tilt the flow lines between adjacent indices are

$$P \to S_u \ (2), \qquad P \to S_l \ (2), \qquad S_u \to Q \ (2), \qquad S_l \to Q \ (2),$$

exactly the pattern of the flat torus. Saddle-to-saddle connections are destroyed by any small generic perturbation, which is Smale's theorem in action. This is why Nash's upright-torus example [Nash p. 196] is fine for counting critical points but must be tilted before we count flow lines.

Watch out: the upright torus is not a counter-example

For the upright torus the naive Morse–Witten count happens to give the right answer (all the relevant signed counts vanish). That is luck. In general, saddle-to-saddle connections make the counts ill-defined: a small change of the metric changes which flow lines exist. The correct statement is that the complex is defined, and computes $H^\bullet(M)$, for Morse–Smale pairs (f, g); different Morse–Smale pairs give different complexes with the same cohomology. Nash does not mention the condition; it is implicit in his statement that the sum (7.85) is finite.

7.16.4 The signs: Nash's orientation rule

Over $\mathbb{Z}_2$ one could simply count flow lines mod 2 (Chapter 7 previews this). Over $\mathbb{Z}$ or $\mathbb{R}$ each flow line needs a sign, and the examples show that the sign is essential: on the circle with $f = \cos\theta$ the two flow lines from the maximum to the minimum must *cancel* (Example 7.84), because $b_0 = b_1 = 1$ requires $\delta = 0$. Nash's rule [Nash pp. 212–213, (7.84)–(7.85)] produces these signs. We state it, explain each step, and then compute it in examples.

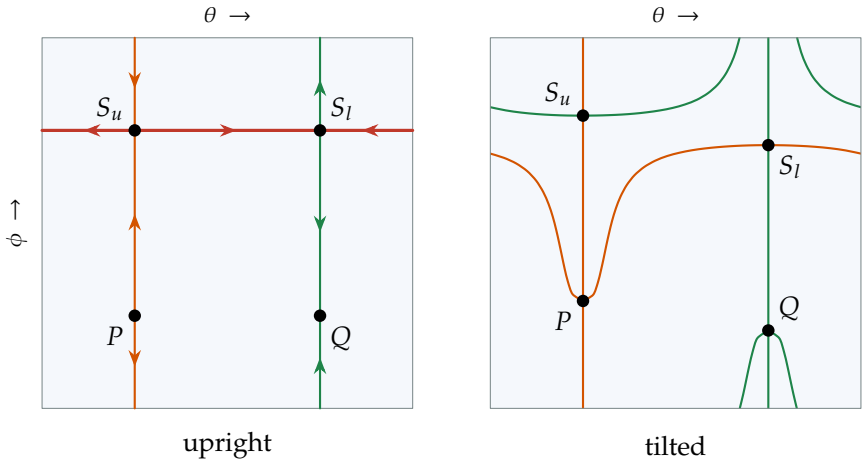


Figure 7.62: Flow lines of the height function on a torus of radii 2 and 1, drawn on the square of angles (θ, ϕ) (opposite sides identified; θ runs around the big circle, ϕ around the tube). Orange: flow lines from the top P into the saddles; green: flow lines from the saddles to the bottom Q. *Left:* the upright torus. The two red flow lines run from the saddle S_u to the saddle S_l, between points of the same index: Morse–Smale fails. *Right:* the same torus tilted by 0.25 radians (numerical solution of $\dot\gamma = -\nabla f$ with the induced metric). The red connection has broken open: the lines leaving S_u now curve away and end at Q, and S_l receives two new lines from P. Every saddle now has exactly two lines in and two lines out.

Orientations are signs of states. The local zero mode at a critical point b of index $p+1$ is ((7.202))

$$|b\rangle \propto e^{-\frac{t}{2}\sum|\lambda_i|x_i^2}\, \mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^{p+1},$$

the wedge of the downhill coordinate differentials. The state is defined only up to a sign: reordering two of the $\mathrm{d}x^i$ flips it. Choosing the sign of $|b\rangle$ is the same as choosing an **orientation** of the $(p+1)$-dimensional space $T_b^- M$: the ordered bases $(v_1, \dots, v_{p+1})$ with $(\mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^{p+1})(v_1, \dots, v_{p+1}) > 0$ are declared positive. That is Nash's phrase "using the critical point b thought of as a $(p+1)$-form to orient $T_b^- M$". At a minimum ($p = -1$, i.e. index 0) the local state is a positive Gaussian function, and the orientation of the zero-dimensional space $T_a^- M = \{0\}$ is the number +1.

Working definition 7.36 (the sign ε_γ of a flow line (Nash's rule))

Let γ be a flow line from b (index $p+1$) down to a (index p), and let ω_b, ω_a be the chosen orientation forms on $T_b^- M$ and $T_a^- M$ [top-degree forms, such as $\mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^{p+1}$ at b, that fix the signs of $|b\rangle$ and $|a\rangle$].

1. Let $v \in T_b^- M$ be the direction in which γ leaves b (pointing downhill, along γ). Split $T_b^- M = \mathbb{R}v \oplus O_b$ with $O_b \perp v$: Nash's (7.84), $T_b^- M = T_b \gamma_{ab} \oplus O_b$.
2. Orient the p-dimensional O_b by the p-form $\iota_v \omega_b = \omega_b(v, \cdot, \dots, \cdot)$: a basis $(w_1, \dots, w_p)$ of O_b is positive iff $\omega_b(v, w_1, \dots, w_p) > 0$.
3. Transport $(w_1, \dots, w_p)$ along γ with the linearised flow [the derivative of the flow map, carrying a tangent vector at one point of γ to a tangent vector at a later point]. As the flow line approaches a, the transported vectors, projected onto $T_a^- M$, form a basis $(w_1', \dots, w_p')$ of $T_a^- M$.
4. $\varepsilon_\gamma := \operatorname{sgn} \omega_a(w_1', \dots, w_p') = \pm 1$: +1 if the orientations agree, −1 if they disagree.

Then $\varepsilon(a,b) := \sum_{\gamma:\, b \to a} \varepsilon_\gamma$ (Nash's (7.85)).

Why step 3 works. Two things need checking: that the transported vectors end up in (or project isomorphically onto) $T_a^- M$, and that the answer does not depend on how close to a we look. Near a, write a tangent vector as $w = (w_-, w_+)$ with $w_- \in T_a^- M$ (the p directions with $\lambda_i < 0$) and $w_+ \in T_a^+ M$. The linearised flow of (7.243) acts by

$$w_i(\sigma) = e^{-\lambda_i \sigma}\, w_i(0): \qquad w_- \text{ grows like } e^{|\lambda|\sigma}, \qquad w_+ \text{ decays like } e^{-\lambda\sigma}. \tag{7.245}$$

The transported O_b stays inside the tangent spaces of $W^u(b)$, and the Morse–Smale condition says that these meet $T\,W^s(a) \approx T_a^+ M$ only in the direction of γ itself. Hence the transported O_b is complementary to $T_a^+ M$, its projection onto $T_a^- M$ is an isomorphism, and by (7.245) the transported subspace converges to $T_a^- M$ as $\sigma \to \infty$ (the w_+ parts die, the w_- parts are stretched). Projection is continuous and never degenerates along the way, so the sign in step 4 is the same wherever it is evaluated. (This is the "λ-lemma" of dynamical systems in its simplest form.)

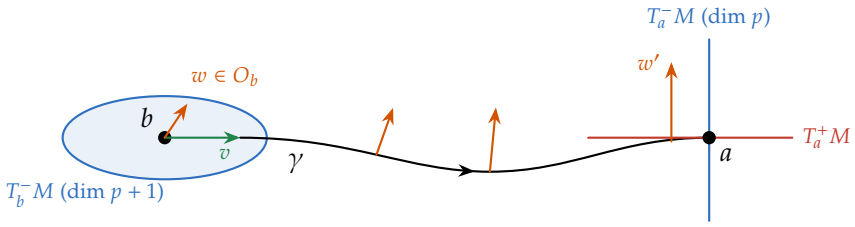


Figure 7.63: Nash's orientation rule. At b the unstable space $T_b^- M$ (blue) contains the exit direction v of the flow line γ and its complement O_b. Interior multiplication by v turns the orientation of $T_b^- M$ into an orientation of O_b. Carried along γ by the linearised flow, the frame of O_b is stretched in the unstable directions of a and squeezed in its stable directions (red), so it lines up with $T_a^- M$ (blue). Comparing with the chosen orientation of $T_a^- M$ gives $\varepsilon_\gamma = \pm 1$.

Check: one dimension. For $p = 0$, b is a local maximum of a function on a line or circle with coordinate x, $\omega_b = \mathrm{d}x$, and $T_b^- M$ is the whole tangent line. Then $O_b = \{0\}$ and step 2 gives the number $\iota_v \mathrm{d}x = \mathrm{d}x(v)$, whose sign is $+1$ if the flow line leaves b in the direction of increasing x and -1 otherwise. At the minimum a, $\omega_a = +1$. So

$$\varepsilon_\gamma = \operatorname{sgn} \mathrm{d}x(v),$$

which is exactly the sign we found by direct computation of $\langle b|\mathrm{d}_t|a\rangle$ in (7.215). Nash's rule is the generalisation of that one-dimensional computation to every dimension.

Worked example 7.91: the flat torus: all pairs cancel

Take $f = \cos x + \cos y$ on the flat torus with orientation $\mathrm{d}x \wedge \mathrm{d}y$. Choose

$$\omega_P = \mathrm{d}x \wedge \mathrm{d}y, \qquad \omega_{S_1} = \mathrm{d}y, \qquad \omega_{S_2} = \mathrm{d}x, \qquad \omega_Q = +1,$$

using the downhill directions found in Example 7.90 (y at $S_1 = (\pi, 0)$; by the same computation x at $S_2 = (0, \pi)$).

From P to S_1. The two flow lines run along $y = 0$, leaving P in the directions $v = \pm\partial_x$. Then

$$\iota_{\pm\partial_x}(\mathrm{d}x \wedge \mathrm{d}y) \overset{(1)}{=} \pm\big(\mathrm{d}x(\partial_x)\,\mathrm{d}y - \mathrm{d}y(\partial_x)\,\mathrm{d}x\big) \overset{(2)}{=} \pm\mathrm{d}y.$$

(1) The interior product of a 2-form, $\iota_v(\alpha \wedge \beta) = \alpha(v)\beta - \beta(v)\alpha$.
(2) $\mathrm{d}x(\partial_x) = 1$, $\mathrm{d}y(\partial_x) = 0$.

So $O_P = \mathbb{R}\partial_y$ is oriented by $\pm\mathrm{d}y$. Along the invariant line $y = 0$ the linearised flow maps ∂_y to a positive multiple of ∂_y (the y-component obeys $\dot{w}_y = \cos(0)\, w_y$ along the line for

the flow $\dot y = \sin y$). At S_1, $T^-_{S_1}M = \mathbb{R}\partial_y$ with $\omega_{S_1} = \mathrm{d}y$. Comparing: the rightward line has $\varepsilon = +1$, the leftward line $\varepsilon = -1$, and $\varepsilon(S_1, P) = 0$.
From P to S_2. Now $v = \pm\partial_y$ and $\iota_{\pm\partial_y}(\mathrm{d}x \wedge \mathrm{d}y) = \mp\mathrm{d}x$: the upward line has $\varepsilon = -1$, the downward line $\varepsilon = +1$, and $\varepsilon(S_2, P) = 0$.
From S_1 and S_2 to Q. From S_1 the lines leave along $v = \pm\partial_y$, and $\iota_{\pm\partial_y}\mathrm{d}y = \pm 1$, compared with $\omega_Q = +1$: $\varepsilon = \pm 1$, total 0. Likewise from S_2 along $\pm\partial_x$: total 0.
So every $\varepsilon(a, b)$ vanishes and $\delta = 0$:

$$H^0 = \mathbb{Z}\,Q, \qquad H^1 = \mathbb{Z}\,S_1 \oplus \mathbb{Z}\,S_2, \qquad H^2 = \mathbb{Z}\,P,$$

the cohomology of T^2. The same signs hold for the tilted torus of Figure 7.62, by continuity of the picture. Physically: every critical point carries an exact supersymmetric ground state, because for each pair of adjacent critical points the two tunnelling paths interfere destructively.

Watch out: conventions for the sign

Authors differ on details of the rule: some contract ω_b with v in the last slot rather than the first, some use the incoming direction at a instead of the outgoing one at b, some work with homology and flow upwards. Each change multiplies all $\varepsilon(a, b)$ with $\mathrm{index}(a) = p$ by a factor depending only on p, or reverses some basis vectors. Both are changes of basis of the C^p, so they do not affect $\delta^2 = 0$ or the cohomology. Likewise, reversing the chosen orientation at a single critical point c replaces the basis vector c by $-c$: a change of basis. Only the combination of choices used in one computation must be consistent. Ours is fixed by agreement with the d_t computation (7.215).

7.16.5 Why $\delta^2 = 0$: broken flow lines

Nash defines δ by [Nash p. 213, (7.86)]

$$\delta(a) = \sum_{b \in C^{p+1}} \varepsilon(a, b)\, b,$$

and then remarks that "a direct calculation [of $\delta^2 = 0$] with the present combinatorial definition is difficult". He sidesteps it with physics: δ is d_t in the rescaled basis (7.233), and $\mathrm{d}_t^2 = 0$. That argument is complete as far as it goes. But it is worth seeing what the "difficult" combinatorial statement actually says, because it is a beautiful piece of geometry, and it is exactly the argument that survives in Floer theory, where there is no d_t to fall back on.

What has to vanish. Apply δ twice to a critical point a of index p:

$$\delta^2(a) \overset{(1)}{=} \delta\Big(\sum_b \varepsilon(a, b)\, b\Big) \overset{(2)}{=} \sum_b \varepsilon(a, b) \sum_c \varepsilon(b, c)\, c \overset{(3)}{=} \sum_c \Big(\sum_b \varepsilon(a, b)\, \varepsilon(b, c)\Big) c. \tag{7.246}$$

(1) Definition (7.86); b runs over critical points of index $p + 1$.
(2) δ is linear; apply (7.86) again to each b, with c running over index $p + 2$.
(3) Exchange the two finite sums.

So $\delta^2 = 0$ says: for every pair (a, c) with $\mathrm{index}(c) = \mathrm{index}(a) + 2$,

$$N(a, c) := \sum_b \varepsilon(a, b)\, \varepsilon(b, c) = \sum_{\substack{\text{broken lines} \\ c \to b \to a}} \varepsilon_{\gamma_{bc}}\, \varepsilon_{\gamma_{ab}} = 0. \tag{7.247}$$

The middle expression expands each ε as a sum over flow lines (7.244): $N(a,c)$ is a signed count of **broken flow lines**, pairs consisting of a flow line from c down to some b followed by a flow line from that b down to a.

Where broken flow lines come from. By the dimension formula (7.244), the unparametrised flow lines from c to a form a space $\widehat{\mathcal{M}}(c,a)$ of dimension $\mathrm{index}(c) - \mathrm{index}(a) - 1 = 1$: a union of curves. A 1-manifold is a union of circles and open intervals. An open interval has two ends, and we can ask what the flow lines do as we run off an end. Three facts, which we take from the analysis of the gradient flow, answer this.

1. *Compactness up to breaking.* A sequence of flow lines from c to a that leaves every compact part of $\widehat{\mathcal{M}}(c,a)$ has a subsequence converging to a broken flow line $c \to b \to a$. (The flow line is spending longer and longer near b. Energy $f(c) - f(a)$ is conserved in the limit; it splits as $(f(c) - f(b)) + (f(b) - f(a))$.)
2. *Gluing.* Conversely, every broken flow line $c \to b \to a$ is the limit of exactly one end of $\widehat{\mathcal{M}}(c,a)$. (Start slightly off the first line, on the side that leads to a; the flow carries you past b and down to a.)
3. *Signs at the ends.* The orientation data that define the ε's also orient each component of $\widehat{\mathcal{M}}(c,a)$, and at the two ends of an interval the products $\varepsilon_{\gamma_{bc}}\varepsilon_{\gamma_{ab}}$ are *opposite*. (One end is "incoming", the other "outgoing", as for the boundary of an oriented interval $\partial[0,1] = \{1\} - \{0\}$.)

Key idea: $\delta^2 = 0$ is Stokes' theorem for an interval

The broken flow lines $c \to b \to a$ are exactly the ends of the one-dimensional space $\widehat{\mathcal{M}}(c,a)$, and each open-interval component contributes two ends with opposite signs. Circles contribute no ends. So

$$N(a,c) = \sum_{\text{intervals}} (+1 - 1) = 0,$$

which is (7.247). This is the same principle as "$\int_{\partial I} 1 = 1 - 1 = 0$": a compact oriented 1-manifold has as many positively as negatively oriented boundary points.

Worked example 7.92: the torus family $P \to Q$ and its broken ends

On the flat torus with $f = \cos x + \cos y$ (Example 7.87), take the flow parameter σ with $t = 1$. In the quadrant $0 < x, y < \pi$ the flow lines from $P = (0,0)$ to $Q = (\pi,\pi)$ are

$$x(\sigma) = 2\arctan e^{\sigma}, \qquad y(\sigma) = 2\arctan e^{\sigma - \Delta}, \qquad \Delta \in \mathbb{R},$$

where we used the time translation to put the x-kink at $\sigma = 0$. The unparametrised flow lines in this quadrant are labelled by the single number Δ, the delay of the y-kink behind the x-kink: an open interval, as the dimension count $2 - 0 - 1 = 1$ predicts. Its two ends are:

- $\Delta \to +\infty$: x jumps from 0 to π long before y moves. The curve hugs $y = 0$ from P to $S_1 = (\pi, 0)$, waits there, then climbs $x = \pi$ to Q: the broken line $P \to S_1 \to Q$.
- $\Delta \to -\infty$: y moves first: the broken line $P \to S_2 \to Q$.

The signs from Example 7.91 are: $P \to S_1$ rightwards $+1$, then $S_1 \to Q$ upwards $+1$ (product $+1$); $P \to S_2$ upwards -1, then $S_2 \to Q$ rightwards $+1$ (product -1). Opposite, as promised. The other three quadrants are the same up to reflections:

quadrant	via S_1	product	via S_2	product
$x>0,\ y>0$	$(+1)(+1)$	$+1$	$(-1)(+1)$	-1
$x<0,\ y>0$	$(-1)(+1)$	-1	$(-1)(-1)$	$+1$
$x<0,\ y<0$	$(-1)(-1)$	$+1$	$(+1)(-1)$	-1
$x>0,\ y<0$	$(+1)(-1)$	-1	$(+1)(+1)$	$+1$

Here "via S_1" lists $\varepsilon_{\gamma_{S_1P}}\,\varepsilon_{\gamma_{QS_1}}$ for the two segments of the broken line. There are $2\times 2=4$ broken lines through S_1 and 4 through S_2; the 8 broken lines are the 8 ends of the 4 intervals, and each interval has one end of each sign. So $N(Q,P)=0$. On the torus this is no surprise, because every single ε already vanishes; but the pairing of ends holds interval by interval, which is the mechanism in general.

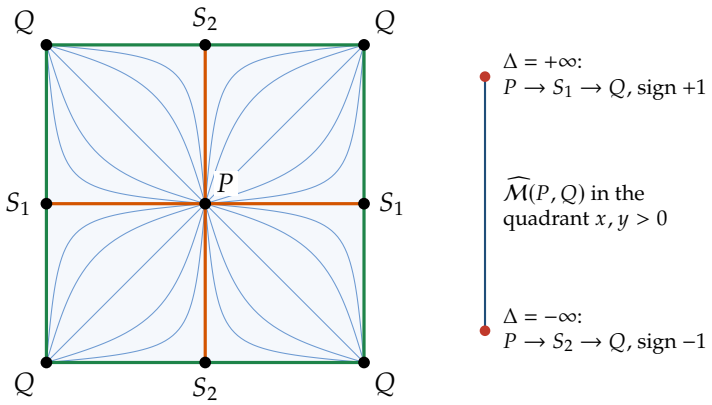


Figure 7.64: Why $\delta^2=0$ on the flat torus (drawn on $[-\pi,\pi]^2$ with $P=(0,0)$ at the centre; the sides are identified, so the four corners are the one minimum Q and the midpoints of opposite sides are the same saddle). Blue: the one-parameter families of flow lines from P to Q, one family per quadrant, labelled by the delay Δ. As $\Delta\to\pm\infty$ a blue curve degenerates to a broken line made of an orange segment ($P\to$ saddle) followed by a green one (saddle $\to Q$). Each family is an open interval (right) whose two ends are broken lines with opposite signs, so the signed count of broken lines, which is the coefficient of P in $\delta^2(Q)$, is zero.

Two proofs, one fact. The physics proof ($\mathrm{d}_t^2=0$ and (7.233)) and the geometric proof (ends of $\widehat{\mathcal{M}}(c,a)$) are the same statement seen from two sides. In the rescaled basis, the matrix element $\langle\tilde{c}|\mathrm{d}_t\mathrm{d}_t|\tilde{a}\rangle$ is a sum over intermediate low-lying states $|\tilde{b}\rangle$, and each term is a product of two single-instanton amplitudes: a broken flow line $c\to b\to a$. That the sum vanishes is guaranteed by $\mathrm{d}_t^2=0$. On the geometric side the same cancellation appears as the pairing of the ends of the one-parameter family of unbroken flow lines from c to a, which are the tunnelling histories that "interpolate" between the broken ones. Stokes' theorem on the moduli space is how the path integral knows that $\mathrm{d}_t^2=0$.

Watch out: breaking needs Morse–Smale

On the upright torus (Figure 7.62, left) the family of flow lines from P to Q is not a union of intervals whose ends are broken lines between adjacent indices: some flow lines leave P, pass through S_u and S_l in turn (two breaks), because of the saddle-to-saddle connection. Transversality is what guarantees that a 1-dimensional family can only break once. This is the second reason (after finiteness of the counts) for insisting on a Morse–Smale pair (f,g).

7.16.6 Worked Morse–Witten complexes

We now have a complete recipe. It is worth stating it as a pipeline, since every example below follows it step by step.

Working definition 7.37 (computing the Morse–Witten complex)

1. *Critical points.* List the critical points of f and their indices. C^p is the free abelian group [all combinations $\sum_c n_c\, c$ with integers n_c, and no relations] (or vector space) with one basis vector for each critical point of index p (Nash's (7.81); "included in C_p are integral linear combinations").
2. *Orientations.* At each critical point c choose an orientation ω_c of $T_c^- M$ (the downhill directions). At a minimum choose +1.
3. *Flow lines.* For each pair (a, b) with $\mathrm{index}(b) = \mathrm{index}(a) + 1$, find the flow lines of $-\nabla f$ from b to a (check Morse–Smale; if it fails, perturb the metric).
4. *Signs.* Give each flow line its sign ε_γ by Nash's rule, and add: $\varepsilon(a, b) = \sum_\gamma \varepsilon_\gamma$.
5. *Differential.* $\delta(a) = \sum_b \varepsilon(a, b)\, b$; write the matrices D_p and check $D_{p+1} D_p = 0$.
6. *Cohomology.* $H^p = \mathrm{Ker}\, D_p / \mathrm{Im}\, D_{p-1}$, by Gaussian elimination (over $\mathbb{Z}$: by the Smith normal form [integer row and column operations that make D_p diagonal. Diagonal entries other than 0 and ±1 give the torsion], which also finds torsion).

Why is step 1 legitimate with integer coefficients? The entries $\varepsilon(a, b)$ are integers, so δ maps integer combinations to integer combinations, and all of the construction makes sense over $\mathbb{Z}$. Witten's d_t acts on real (or complex) forms and sees the complex tensored with $\mathbb{R}$ [the same complex with real instead of integer coefficients]; the integer complex is a refinement that physics does not directly provide but that the geometry of flow lines does.

Worked example 7.93: the two-humped sphere

Take the two-humped sphere of Figure 7.45: a minimum a at the bottom, two peaks b_1 (left) and b_2 (right), and a saddle s on the ridge between them. Near the top use coordinates (x, y) with x running along the ridge from left to right and y running from back to front, and assume the surface is symmetric under $y \mapsto -y$.

Steps 1–2. $C^0 = \mathbb{Z}a$, $C^1 = \mathbb{Z}s$, $C^2 = \mathbb{Z}b_1 \oplus \mathbb{Z}b_2$. Orientations: $\omega_{b_1} = \omega_{b_2} = \mathrm{d}x \wedge \mathrm{d}y$ (the same orientation of the sphere near both peaks), $\omega_s = \mathrm{d}y$ (at the saddle the downhill direction is across the ridge), $\omega_a = +1$.

Step 3. From b_1 exactly one flow line reaches s: the crest of the ridge, moving to the right. From b_2 one flow line, moving left. From s two flow lines descend to a, one down the front and one down the back. Every other flow line from a peak goes straight to a. The pair is Morse–Smale: in two dimensions, any pair (a, b) with adjacent indices involves a maximum or a minimum, whose unstable or stable manifold is open, and an open set meets anything transversally.

Step 4. From s to a: the lines leave s along $v = \pm\partial_y$, and $\iota_{\pm\partial_y}\mathrm{d}y = \pm 1$. So $\varepsilon(a, s) = (+1) + (-1) = 0$. From b_1 to s: $v = +\partial_x$, and

$$\iota_{\partial_x}(\mathrm{d}x \wedge \mathrm{d}y) \overset{(1)}{=} \mathrm{d}x(\partial_x)\,\mathrm{d}y - \mathrm{d}y(\partial_x)\,\mathrm{d}x \overset{(2)}{=} \mathrm{d}y,$$

(1) The interior product of a wedge of two 1-forms.
(2) $\mathrm{d}x(\partial_x) = 1$, $\mathrm{d}y(\partial_x) = 0$.

so $O_{b_1} = \mathbb{R}\partial_y$ is oriented by ∂_y. The reflection $y \mapsto -y$ fixes the crest pointwise and commutes with the flow, so the linearised flow carries ∂_y to a positive multiple of ∂_y. At s we compare with $\omega_s = \mathrm{d}y$: $\varepsilon(s, b_1) = +1$. From b_2 the line leaves along $v = -\partial_x$, $\iota_{-\partial_x}(\mathrm{d}x \wedge \mathrm{d}y) = -\mathrm{d}y$, and $\varepsilon(s, b_2) = -1$.
Steps 5–6.

$$\delta(a) = 0, \qquad \delta(s) = b_1 - b_2, \qquad \delta(b_i) = 0; \qquad D_0 = \begin{pmatrix} \end{pmatrix}, \quad D_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

Then

$$H^0 \overset{(1)}{=} \operatorname{Ker} D_0 = \mathbb{Z}a \cong \mathbb{Z},$$
$$H^1 \overset{(2)}{=} \operatorname{Ker} D_1 / \operatorname{Im} D_0 = 0,$$
$$H^2 \overset{(3)}{=} \frac{\mathbb{Z}b_1 \oplus \mathbb{Z}b_2}{\mathbb{Z}(b_1 - b_2)} \cong \mathbb{Z}.$$

(1) Nothing maps into C^0, and $D_0 = 0$.
(2) D_1 is injective: $D_1(ns) = n(b_1 - b_2) \neq 0$ for $n \neq 0$.
(3) The map $mb_1 + nb_2 \mapsto m + n$ is onto $\mathbb{Z}$ with kernel exactly $\mathbb{Z}(b_1 - b_2)$; so the quotient is $\mathbb{Z}$, generated by the class of b_1 (equal to the class of b_2).

This is $H^\bullet(S^2; \mathbb{Z}) = (\mathbb{Z}, 0, \mathbb{Z})$. The single cancelling pair is $(s,\ b_1 - b_2)$, in degrees $(1, 2)$: $R(u) = u$, as announced in Figure 7.45.
The physics. Back in the unrescaled basis (7.232),

$$\mathrm{d}_t|s\rangle \approx A_1|b_1\rangle - A_2|b_2\rangle, \qquad A_i \propto e^{-t(f(b_i) - f(s))}.$$

The exact 2-form ground state must be orthogonal to $\operatorname{Im} \mathrm{d}_t$, so it is proportional to $A_2|b_1\rangle + A_1|b_2\rangle$. If the peaks have equal height it is the symmetric combination $|b_1\rangle + |b_2\rangle$ of Figure 7.61. If b_1 is higher, then $A_1 \ll A_2$ and the ground state sits almost entirely on b_1: it concentrates on the *highest* peak. That agrees with the exact zero mode e^{tf} vol of Section 7.12, which for large t is dominated by the global maximum of f. The lifted 2-form is the orthogonal combination $A_1|b_1\rangle - A_2|b_2\rangle$, the superpartner of $|s\rangle$.

Worked example 7.94: perfect functions: the round sphere and $\mathbb{CP}^n$

For the height on the round S^2 there are two critical points, of index 0 and 2. $C^1 = 0$, so both maps $\delta_0 \colon C^0 \to C^1$ and $\delta_1 \colon C^1 \to C^2$ vanish, whatever the flow lines are. The complex is $\mathbb{Z} \to 0 \to \mathbb{Z}$ and $H^\bullet = (\mathbb{Z}, 0, \mathbb{Z})$. The circle of meridians from N to S (Example 7.90) plays no role: it connects indices that differ by 2.
The same argument works whenever no two critical indices are adjacent, because then every C^p next to a non-zero $C^{p\pm1}$ is zero. For $f = \sum_i a_i |z_i|^2$ on $\mathbb{CP}^n$ (Chapter 7, the classic example of Nash's §1) the indices are $0, 2, \ldots, 2n$. So $\delta = 0$ identically and $H^{2k}(\mathbb{CP}^n; \mathbb{Z}) = \mathbb{Z}$ for $0 \leq k \leq n$, all other groups zero. This is the Morse–Witten version of the **lacunary principle**: a Morse function whose indices have no two consecutive values is perfect, over every coefficient ring. Physically, every perturbative ground state is exact, because there is no state of adjacent fermion number for it to pair with.

Worked example 7.95: six critical points on the round sphere

Let $f(x,y,z) = x^2 + 2y^2 + 3z^2$ restricted to the unit sphere, with the round metric. Its critical points are the six points $\pm e_1, \pm e_2, \pm e_3$. To avoid writing "$-e_2$" for a basis vector (it would look like the negative of e_2), we name them

$$a_\pm = (\pm 1, 0, 0), \qquad s_\pm = (0, \pm 1, 0), \qquad b_\pm = (0, 0, \pm 1).$$

Indices. Near a_+ use (y,z) as coordinates; then $x^2 = 1 - y^2 - z^2$ and $f = 1 + y^2 + 2z^2$: index 0. Near s_+, with coordinates (x,z), $f = 2 - x^2 + z^2$: index 1, downhill direction ∂_x. Near b_+, with (x,y), $f = 3 - 2x^2 - y^2$: index 2. The same holds at the antipodes. So $m = (2,2,2)$ while $b(S^2) = (1,0,1)$: two cancelling pairs, $M_u - P_u = 1 + 2u + u^2 = (1+u)(1+u)$ and $R(u) = 1 + u$.

The flow, exactly. Write $A = \operatorname{diag}(1,2,3)$, so $f(x) = x \cdot Ax$. The gradient on the unit sphere is the ambient gradient $2Ax$ minus its normal part, $\nabla f = 2\big(Ax - (x \cdot Ax)\, x\big)$. We claim that

$$x(\sigma) = \frac{y(\sigma)}{|y(\sigma)|}, \qquad y(\sigma) = e^{-2A\sigma} x_0,$$

solves $\dot{x} = -\nabla f$. Indeed, with $\dot{y} = -2Ay$,

$$\dot{x} \overset{(1)}{=} \frac{\dot{y}}{|y|} - \frac{y\,(y \cdot \dot{y})}{|y|^3} \overset{(2)}{=} -2Ax + 2x\,\frac{y \cdot Ay}{|y|^2} \overset{(3)}{=} -2\big(Ax - (x \cdot Ax)\, x\big) = -\nabla f.$$

(1) Quotient rule, with $\frac{\mathrm{d}}{\mathrm{d}\sigma}|y| = (y \cdot \dot{y})/|y|$.
(2) $\dot{y} = -2Ay$ and $y/|y| = x$.
(3) $(y \cdot Ay)/|y|^2 = x \cdot Ax$.

So the flow line through $x_0 = (x_0, y_0, z_0)$ is the direction of $(x_0 e^{-2\sigma},\ y_0 e^{-4\sigma},\ z_0 e^{-6\sigma})$. As $\sigma \to +\infty$ the slowest-decaying non-zero component wins: the line ends at $a_\pm$ if $x_0 \neq 0$ (sign of x_0), and at $s_\pm$ if $x_0 = 0$, $y_0 \neq 0$. As $\sigma \to -\infty$ the fastest-growing non-zero component wins: the line starts at $b_\pm$ if $z_0 \neq 0$. Hence the flow lines between adjacent indices are quarter great circles:

- $b_\pm \to s_\pm$ along the great circle $x = 0$ (four lines, one for each pair of signs);
- $s_\pm \to a_\pm$ along the equator $z = 0$ (four lines).

Morse–Smale holds for the same reason as in the previous example.

Orientations. In the ambient coordinates, choose $\omega_{b_\pm} = \mathrm{d}x \wedge \mathrm{d}y$ (the tangent planes at $b_\pm$ are both horizontal), $\omega_{s_\pm} = \mathrm{d}x$, $\omega_{a_\pm} = +1$.

Signs, $s \to a$. The line $s_+ \to a_+$ leaves $s_+ = (0,1,0)$ towards $(1,0,0)$, i.e. along $v = +\partial_x$: $\varepsilon = \mathrm{d}x(\partial_x) = +1$. The line $s_+ \to a_-$ leaves along $-\partial_x$: $\varepsilon = -1$. The same for s_-.

Signs, $b \to s$. The line $b_+ \to s_+$ leaves b_+ along $v = +\partial_y$. Then $\iota_{\partial_y}(\mathrm{d}x \wedge \mathrm{d}y) = -\mathrm{d}x$, so $O_{b_+} = \mathbb{R}\partial_x$ is positively spanned by $w = -\partial_x$. Along the great circle $x = 0$ the vector ∂_x is tangent to the sphere, and the reflection $x \mapsto -x$ (a symmetry of f and g fixing that circle) forces the linearised flow to carry ∂_x to a positive multiple of itself. So w arrives at s_+ as a positive multiple of $-\partial_x$, and $\omega_{s_+}(-\partial_x) = -1$: $\varepsilon = -1$. The line $b_+ \to s_-$ leaves along $-\partial_y$, $\iota_{-\partial_y}(\mathrm{d}x \wedge \mathrm{d}y) = +\mathrm{d}x$, $w = +\partial_x$, $\varepsilon = +1$. The lines from b_- leave in the same ambient directions ($\pm\partial_y$) with the same ω, so they have the same signs.

The complex.

$$\delta(a_+) = s_+ + s_-, \qquad \delta(a_-) = -s_+ - s_-,$$
$$\delta(s_+) = -b_+ - b_-, \qquad \delta(s_-) = b_+ + b_-,$$

$$D_0 = \begin{pmatrix} 1 & -1 \\ 1 & -1 \end{pmatrix} \begin{matrix} \leftarrow s_+ \\ \leftarrow s_- \end{matrix}, \qquad D_1 = \begin{pmatrix} -1 & 1 \\ -1 & 1 \end{pmatrix} \begin{matrix} \leftarrow b_+ \\ \leftarrow b_- \end{matrix},$$

columns labelled (a_+, a_-) and (s_+, s_-) respectively. Check:

$$D_1 D_0 \overset{(1)}{=} \begin{pmatrix} -1\cdot 1 + 1\cdot 1 & -1\cdot(-1) + 1\cdot(-1) \\ -1\cdot 1 + 1\cdot 1 & -1\cdot(-1) + 1\cdot(-1) \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

(1) Row times column.

Cohomology: $\operatorname{Ker} D_0 = \mathbb{Z}(a_+ + a_-)$, so $H^0 = \mathbb{Z}$. $\operatorname{Ker} D_1 = \mathbb{Z}(s_+ + s_-) = \operatorname{Im} D_0$, so $H^1 = 0$. $H^2 = (\mathbb{Z}b_+ \oplus \mathbb{Z}b_-)/\mathbb{Z}(b_+ + b_-) \cong \mathbb{Z}$, generated by the class of b_+. Again $H^\bullet(S^2;\mathbb{Z})$, from a different complex. The two cancelling pairs are $(a_+ - a_-,\ s_+ + s_-)$ in degrees $(0,1)$ and $(s_+ - s_-,\ b_+ + b_-)$ in degrees $(1,2)$, matching $R(u) = 1 + u$.

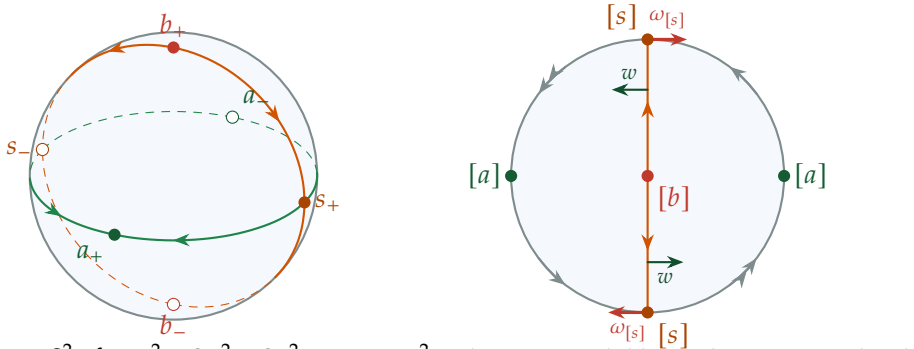


Figure 7.65: *Left:* the six critical points of $f = x^2 + 2y^2 + 3z^2$ on the round sphere (open circles are on the hidden back side, dashed arcs are hidden). Orange quarter circles run from the maxima $b_\pm$ to the saddles $s_\pm$; green ones from the saddles to the minima $a_\pm$. *Right:* the same function on $\mathbb{RP}^2$, drawn as the upper hemisphere seen from above with antipodal boundary points identified. There is one critical point of each index. The two flow lines from $[b]$ reach $[s]$ from opposite sides. The orientation of $[s]$ (red) looks reversed at the two lifts because the gluing map is $v \mapsto -v$. The transported vectors w (green) disagree with it on *both* lines, so the two signs add: $\varepsilon([s],[b]) = -2$.

Worked example 7.96: $\mathbb{RP}^2$: the complex sees torsion

The antipodal map $\tau(x) = -x$ preserves $f = x^2 + 2y^2 + 3z^2$ and the round metric, so f descends to $\mathbb{RP}^2 = S^2/\{\pm 1\}$ (the function in the table of Example 7.81). It has one critical point of each index: $[a] = \{a_\pm\}$, $[s] = \{s_\pm\}$, $[b] = \{b_\pm\}$, and $m = (1,1,1)$.

Orientations. $\mathbb{RP}^2$ is not orientable, but Nash's rule never needs an orientation of M; it needs orientations of the small spaces $T_c^- M$. Choose them at the lifts a_+, s_+, b_+ as before: $+1$, $\mathrm{d}x$, $\mathrm{d}x \wedge \mathrm{d}y$. To compare at the other lift, push forward by τ, whose derivative is $v \mapsto -v$ on ambient vectors:

- at s_- the orientation of $[s]$ is carried by $\tau_* \partial_x = -\partial_x$: it is $-\mathrm{d}x$, opposite to the ω_{s_-} of Example 7.95;
- at b_- it is carried by $(\tau_* \partial_x, \tau_* \partial_y) = (-\partial_x, -\partial_y)$, and $(\mathrm{d}x \wedge \mathrm{d}y)(-\partial_x, -\partial_y) = +1$: it is $+\mathrm{d}x \wedge \mathrm{d}y$, the same as ω_{b_-}.

Flow lines and signs. Each flow line on $\mathbb{RP}^2$ has two lifts; use the one starting at the chosen lift of its upper end.

- $[s] \to [a]$: the lines $s_+ \to a_+$ (sign +1) and $s_+ \to a_-$ (sign −1, compared with +1 at $[a]$). So $\varepsilon([a],[s]) = 0$.
- $[b] \to [s]$: the line $b_+ \to s_+$ has sign −1 (computed above). The line $b_+ \to s_-$ had sign +1 relative to $\omega_{s_-} = \mathrm{d}x$; relative to the orientation $-\mathrm{d}x$ of $[s]$ at s_- it is −1. So $\varepsilon([s],[b]) = -1-1 = -2$.

The complex and its cohomology.

$$0 \to \mathbb{Z}[a] \xrightarrow{0} \mathbb{Z}[s] \xrightarrow{\times(-2)} \mathbb{Z}[b] \to 0.$$

$$H^0 \overset{(1)}{=} \mathbb{Z}, \qquad H^1 \overset{(2)}{=} 0, \qquad H^2 \overset{(3)}{=} \mathbb{Z}/2\mathbb{Z}.$$

(1) $\delta_0 = 0$, so all of C^0 is closed.
(2) Multiplication by −2 on $\mathbb{Z}$ is injective.
(3) $\operatorname{Im}\delta_1 = 2\mathbb{Z}[b]$, and $\mathbb{Z}/2\mathbb{Z}$ is what is left.

This is $H^\bullet(\mathbb{RP}^2;\mathbb{Z}) = (\mathbb{Z},0,\mathbb{Z}_2)$. Change the coefficients:

coefficients	δ_1	(H^0,H^1,H^2)	comment
$\mathbb{Z}$	$\times(-2)$	$(\mathbb{Z},0,\mathbb{Z}_2)$	torsion in H^2
$\mathbb{R}$ (Witten's d_t)	$\times(-2)$, invertible	$(\mathbb{R},0,0)$	$b = (1,0,0)$
$\mathbb{Z}_2$	0	$(\mathbb{Z}_2,\mathbb{Z}_2,\mathbb{Z}_2)$	perfect mod 2
$\mathbb{Z}_3$	$\times 1$, invertible	$(\mathbb{Z}_3,0,0)$	like $\mathbb{R}$

The table reproduces what Chapter 7 says about $\mathbb{RP}^2$: f is perfect over $\mathbb{Z}_2$ and not over $\mathbb{R}$, and no Morse function on $\mathbb{RP}^2$ can be perfect over $\mathbb{Z}$, because the integer cohomology has torsion. The number 2 is visible only to the integer complex; real coefficients, and so the real forms of Witten's quantum mechanics, only see whether it is zero.

The physics of $\mathbb{RP}^2$. Wavefunctions on $\mathbb{RP}^2$ are forms on S^2 invariant under τ. The exact zero modes on S^2 are e^{-tf} and $e^{tf}\mathrm{vol}$; the first is invariant, but $\tau^*\mathrm{vol} = -\mathrm{vol}$ (the antipodal map reverses the orientation of S^2), so the second is not. Hence $\mathbb{RP}^2$ has a single supersymmetric ground state, a bosonic one, and Witten index $1 = \chi(\mathbb{RP}^2)$. In the low-lying complex, the invariant combinations are

$$[a] \leftrightarrow a_+ + a_-, \qquad [s] \leftrightarrow s_+ - s_-, \qquad [b] \leftrightarrow b_+ + b_-,$$

(the relative sign in $[s]$ is the orientation flip found above), and indeed, using Example 7.95,

$$\delta(s_+ - s_-) \overset{(1)}{=} (-b_+ - b_-) - (b_+ + b_-) \overset{(2)}{=} -2(b_+ + b_-).$$

(1) $\delta(s_+) = -b_+ - b_-$ and $\delta(s_-) = b_+ + b_-$.
(2) Collect terms.

The two tunnelling paths from $[b]$ to $[s]$ interfere *constructively*, and the pair $([s],[b])$ is lifted. On the circle with $\cos\theta$ they interfered destructively and nothing was lifted. The τ-odd combinations $a_+ - a_-$, $s_+ + s_-$, $b_+ - b_-$ form the other half of the S^2 complex: $\delta(a_+ - a_-) = 2(s_+ + s_-)$, $\delta(s_+ + s_-) = 0$, and $b_+ - b_-$ carries the odd ground state $e^{tf}\mathrm{vol}$.

Try it

Take $M = S^1 \times S^1$ with the product metric and $f(x, y) = -\cos 2x + \cos y$. List the critical points and indices, find M_u, and guess δ from Examples 7.83 and 7.84. *Answer:* critical points are (minimum or maximum of $-\cos 2x$) $\times$ (maximum or minimum of $\cos y$): $m = (2, 4, 2)$, $M_u = (2 + 2u)(1 + u)$. The differential is built from the circle complexes of the two factors (next subsection): δ_x has rank 1 and $\delta_y = 0$. So $M_u - P_u = (2+2u)(1+u) - (1+u)^2 = (1+u)^2$ and $R(u) = 1 + u$: one pair cancelled in degrees $(0, 1)$ and one in degrees $(1, 2)$.

7.16.7 Products, and why the answer does not depend on f

Two questions remain about the complex as a computational tool. How do we build it for a product manifold from the complexes of the factors? And since it is built from a particular function f and a particular metric g, why is its cohomology independent of both? Both questions have a quantum mechanical answer and a geometric one, and it is instructive to see them agree.

Critical points of a sum. Let $M = M_1 \times M_2$ with the product metric, and let $f(x_1, x_2) = f_1(x_1) + f_2(x_2)$. Then

$$\mathrm{d}f = 0 \overset{(1)}{=} \mathrm{d}f_1 \oplus \mathrm{d}f_2 = 0 \iff \mathrm{d}f_1(x_1) = 0 \text{ and } \mathrm{d}f_2(x_2) = 0,$$
$$Hf(c_1, c_2) \overset{(2)}{=} \begin{pmatrix} Hf_1(c_1) & 0 \\ 0 & Hf_2(c_2) \end{pmatrix}, \qquad \operatorname{index}(c_1, c_2) \overset{(3)}{=} \operatorname{index} c_1 + \operatorname{index} c_2.$$

(1) f_1 depends only on the coordinates of M_1, f_2 only on those of M_2, so the gradient splits into two blocks.
(2) Mixed second derivatives $\partial^2 f / \partial x_1 \partial x_2$ vanish for the same reason.
(3) The negative eigenvalues of a block-diagonal matrix are those of the blocks.

So the critical points of f are the pairs of critical points, and the Morse polynomial multiplies:

$$M_u(f) = \sum_{c_1, c_2} u^{\operatorname{index} c_1 + \operatorname{index} c_2} = \Big(\sum_{c_1} u^{\operatorname{index} c_1}\Big)\Big(\sum_{c_2} u^{\operatorname{index} c_2}\Big) = M_u(f_1)\, M_u(f_2). \tag{7.248}$$

The Poincaré polynomial multiplies too (Künneth formula, over a field): $P_u(M_1 \times M_2) = P_u(M_1) P_u(M_2)$. We now see the complex behind both statements.

The quantum mechanics of a product. A form on $M_1 \times M_2$ is a sum of products $\alpha \wedge \beta$ of forms pulled back from the two factors. Write $\alpha \otimes \beta$ for $\alpha \wedge \beta$ and $(-1)^{F_1}$ for the operator that multiplies α by $(-1)^{\deg \alpha}$. The Leibniz rule $\mathrm{d}(\alpha \wedge \beta) = \mathrm{d}\alpha \wedge \beta + (-1)^{\deg \alpha} \alpha \wedge \mathrm{d}\beta$ says

$$\mathrm{d} = \mathrm{d}^{(1)} \otimes 1 + (-1)^{F_1} \otimes \mathrm{d}^{(2)}.$$

Because $e^{tf} = e^{tf_1} e^{tf_2}$ is a product of functions of the separate factors, conjugating by it conjugates each term separately, and because the L^2 inner product of a product metric factorises (and $(-1)^{F_1}$ is self-adjoint),

$$\mathrm{d}_t = \mathrm{d}_t^{(1)} \otimes 1 + (-1)^{F_1} \otimes \mathrm{d}_t^{(2)}, \qquad \mathrm{d}_t^* = \mathrm{d}_t^{(1)*} \otimes 1 + (-1)^{F_1} \otimes \mathrm{d}_t^{(2)*}. \tag{7.249}$$

Add them: $Q_1(t) = Q^{(1)} \otimes 1 + (-1)^{F_1} \otimes Q^{(2)}$ with $Q^{(i)} = \mathrm{d}_t^{(i)} + \mathrm{d}_t^{(i)*}$. Its square is the Hamiltonian:

$$\begin{aligned} H_t = Q_1(t)^2 &\overset{(1)}{=} \big(Q^{(1)}\big)^2 \otimes 1 + \big((-1)^{F_1}\big)^2 \otimes \big(Q^{(2)}\big)^2 + \big(Q^{(1)}(-1)^{F_1} + (-1)^{F_1} Q^{(1)}\big) \otimes Q^{(2)} \\ &\overset{(2)}{=} H_t^{(1)} \otimes 1 + 1 \otimes H_t^{(2)}. \end{aligned} \tag{7.250}$$

(1) Multiply out $(A\otimes 1 + B\otimes C)^2 = A^2\otimes 1 + B^2\otimes C^2 + (AB+BA)\otimes C$, keeping the order of operators in each factor.
(2) $((-1)^{F_1})^2 = 1$; and $Q^{(1)}$ changes the degree by ± 1, so it *anticommutes* with $(-1)^{F_1}$ and the cross term vanishes. Finally $(Q^{(i)})^2 = H_t^{(i)}$ by (7.73).

This is the supersymmetric version of "the Hamiltonian of two decoupled systems is the sum". The sign $(-1)^{F_1}$ in the supercharge, which looks like a nuisance, is exactly what kills the cross term: without it $Q_1(t)^2$ would contain $2Q^{(1)}\otimes Q^{(2)}$ and would not be a sum of two commuting Hamiltonians. (The same sign is the Jordan–Wigner string that makes fermions on two sites anticommute.)

Consequences, read off from (7.250):

- The spectrum of H_t consists of the sums $E_1 + E_2$. Both are ≥ 0, so $E = 0$ iff $E_1 = E_2 = 0$: the ground states of the product are the products of ground states, and $\ker H_t = \ker H_t^{(1)} \otimes \ker H_t^{(2)}$. That is the Künneth formula for harmonic forms.
- Since $F = F_1 + F_2$, $(-1)^F = (-1)^{F_1}\otimes(-1)^{F_2}$ and the Witten index multiplies: $\chi(M_1\times M_2) = \chi(M_1)\,\chi(M_2)$.
- A low-lying state has $E_1 + E_2 \leq ct$, and since the next levels in each factor are of order t, both factors must be low-lying. So $V(M) = V(M_1)\otimes V(M_2)$ and, by (7.249), the low-lying complex of the product is the tensor product of the low-lying complexes, with the sign $(-1)^{F_1}$.

Worked example 7.97: the four ground states of the flat torus

On the circle with $f = \cos\theta$ the two exact ground states are $e^{-t\cos\theta}$ (a 0-form) and $e^{t\cos\theta}\mathrm{d}\theta$ (a 1-form) (Example 7.73). On T^2 with $f = \cos x + \cos y$ the products are

$$e^{-t\cos x}\,e^{-t\cos y}, \qquad e^{t\cos x}\,e^{-t\cos y}\,\mathrm{d}x, \qquad e^{-t\cos x}\,e^{t\cos y}\,\mathrm{d}y, \qquad e^{t\cos x}\,e^{t\cos y}\,\mathrm{d}x\wedge\mathrm{d}y.$$

Their degrees are $0,1,1,2$, and $b_\bullet(T^2) = (1,2,1)$. We can check one directly with (7.172), $\mathrm{d}_t = \mathrm{d} + t\,\mathrm{d}f\wedge$:

$$\begin{aligned}\mathrm{d}_t\big(e^{t\cos x - t\cos y}\mathrm{d}x\big) &\overset{(1)}{=} \mathrm{d}\big(e^{t\cos x - t\cos y}\big)\wedge\mathrm{d}x + t\,(-\sin x\,\mathrm{d}x - \sin y\,\mathrm{d}y)\wedge e^{t\cos x - t\cos y}\mathrm{d}x\\ &\overset{(2)}{=} e^{t\cos x - t\cos y}\big(t\sin y\,\mathrm{d}y\wedge\mathrm{d}x - t\sin y\,\mathrm{d}y\wedge\mathrm{d}x\big) \overset{(3)}{=} 0.\end{aligned}$$

(1) Definition of d_t with $\mathrm{d}f = -\sin x\,\mathrm{d}x - \sin y\,\mathrm{d}y$.
(2) $\mathrm{d}(e^{t\cos x - t\cos y}) = e^{t\cos x - t\cos y}(-t\sin x\,\mathrm{d}x + t\sin y\,\mathrm{d}y)$, and $\mathrm{d}x\wedge\mathrm{d}x = 0$ removes the x-parts.
(3) The two terms cancel.

A similar computation with $\mathrm{d}_t^* = \mathrm{d}^* + t\,\iota_{\nabla f}$ shows it is also annihilated by d_t^*. Each ground state is localised on one critical point: the first on the minimum Q (where $\cos x = \cos y = -1$), the second on the saddle $(0,\pi)$, the third on the saddle $(\pi,0)$, the last on the maximum P. Every critical point carries an exact ground state, just as the vanishing of the whole complex in Example 7.91 predicted.

The same product rule, from the flow lines. Geometrically, the gradient flow of $f_1 + f_2$ for the product metric is the product of the two flows, run with the same time s: a flow line is a pair $(\gamma_1(s), \gamma_2(s))$, each component a flow line or a constant. The unparametrised flow lines from (b_1, b_2) to (a_1, a_2) form a space of dimension $\Delta_1 + \Delta_2 - 1$, with $\Delta_i = \text{index}\,b_i - \text{index}\,a_i \geq 0$

((7.244), applied to each factor, minus one for the common time shift). For an index difference of one, one of the Δ_i is zero, so *only one factor moves*. The flow lines that count are (γ_1, c_2) and (c_1, γ_2).

Their signs follow from Nash's rule with the product orientation $\omega_{(c_1,c_2)} = \omega_{c_1} \wedge \omega_{c_2}$. If the second factor moves, the exit vector v lies in the second factor and

$$\iota_v\big(\omega_{c_1} \wedge \omega_{b_2}\big) \overset{(1)}{=} (\iota_v \omega_{c_1}) \wedge \omega_{b_2} + (-1)^{\operatorname{index} c_1}\, \omega_{c_1} \wedge \iota_v \omega_{b_2} \overset{(2)}{=} (-1)^{\operatorname{index} c_1}\, \omega_{c_1} \wedge \iota_v \omega_{b_2}.$$

(1) The graded Leibniz rule for the interior product.
(2) ω_{c_1} involves only covectors of the first factor, which vanish on v.

The first-factor vectors sit at the fixed point c_1 and are only stretched by positive factors, so the comparison at the lower end gives $\varepsilon = (-1)^{\operatorname{index} c_1} \varepsilon_{\gamma_2}$. If the first factor moves, ι_v hits ω_{b_1} first and there is no sign. Altogether

$$\delta(c_1 \times c_2) = \delta^{(1)} c_1 \times c_2 + (-1)^{\operatorname{index} c_1}\, c_1 \times \delta^{(2)} c_2, \tag{7.251}$$

which is exactly the tensor product of complexes predicted by (7.249): the geometric orientation sign and the fermionic sign $(-1)^{F_1}$ are the same sign. Its square vanishes for the same reason as the cross term in (7.250):

$$\begin{aligned}\delta^2(c_1 \times c_2) &\overset{(1)}{=} (\delta^{(1)})^2 c_1 \times c_2 + (-1)^{\operatorname{index} c_1 + 1} \delta^{(1)} c_1 \times \delta^{(2)} c_2 \\ &\quad + (-1)^{\operatorname{index} c_1} \delta^{(1)} c_1 \times \delta^{(2)} c_2 + c_1 \times (\delta^{(2)})^2 c_2 \overset{(2)}{=} 0.\end{aligned}$$

(1) Apply (7.251) twice; $\delta^{(1)} c_1$ has index index $c_1 + 1$, and $(-1)^{\operatorname{index} c_1}(-1)^{\operatorname{index} c_1} = 1$.
(2) $(\delta^{(i)})^2 = 0$ in each factor, and the two middle terms have opposite signs.

Worked example 7.98: the torus with $f = -\cos 2x + \cos y$, from the two circles

Let a_1, a_2 (index 0) and b_1, b_2 (index 1) be the critical points of $-\cos 2x$, with $\delta^{(1)} a_1 = -b_1 + b_2$, $\delta^{(1)} a_2 = b_1 - b_2$ (Example 7.83), and let Q' (the minimum $y = \pi$) and P' (the maximum $y = 0$) be those of $\cos y$, with $\delta^{(2)} = 0$ (Example 7.84). The product complex has

$$C^0 = \langle a_1 Q', a_2 Q' \rangle, \qquad C^1 = \langle b_1 Q', b_2 Q', a_1 P', a_2 P' \rangle, \qquad C^2 = \langle b_1 P', b_2 P' \rangle,$$

so $m = (2, 4, 2)$. By (7.251), with $\delta^{(2)} = 0$, $\delta(c_1 c_2) = (\delta^{(1)} c_1) c_2$, and in these bases

$$D_0 = \begin{pmatrix} -1 & 1 \\ 1 & -1 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}, \qquad D_1 = \begin{pmatrix} 0 & 0 & -1 & 1 \\ 0 & 0 & 1 & -1 \end{pmatrix}.$$

Check $D_1 D_0 = 0$: the only nonzero columns of D_1 are the last two, and the last two rows of D_0 vanish. Then

$$b_0 \overset{(1)}{=} 2 - 1 = 1, \qquad b_1 \overset{(2)}{=} (4 - 1) - 1 = 2, \qquad b_2 \overset{(3)}{=} 2 - 1 = 1.$$

(1) $\operatorname{rank} D_0 = 1$; the kernel is spanned by $a_1 Q' + a_2 Q'$.
(2) $\operatorname{rank} D_1 = 1$, so $\dim \operatorname{Ker} D_1 = 3$; subtract $\operatorname{rank} D_0 = 1$.
(3) $\operatorname{Im} D_1$ is spanned by $b_1 P' - b_2 P'$.

Representatives of H^1 are $b_1 Q'$ (equivalently $b_2 Q'$, since $b_1 Q' - b_2 Q' = -\delta(a_1 Q')$) and $a_1 P' + a_2 P'$: the classes $[\mathrm{d}x]$ and $[\mathrm{d}y]$ of the torus, i.e. $H^1(S^1) \otimes H^0(S^1) \oplus H^0(S^1) \otimes H^1(S^1)$. The strong inequality reads $M_u - P_u = (2 + 4u + 2u^2) - (1 + 2u + u^2) = (1 + u)(1 + u)$, so $R(u) = 1 + u$, as found in the exercise above. Physically: each factor-circle has one lifted pair, and the lifted pairs of the torus are (lifted pair)×(ground state of the other factor).

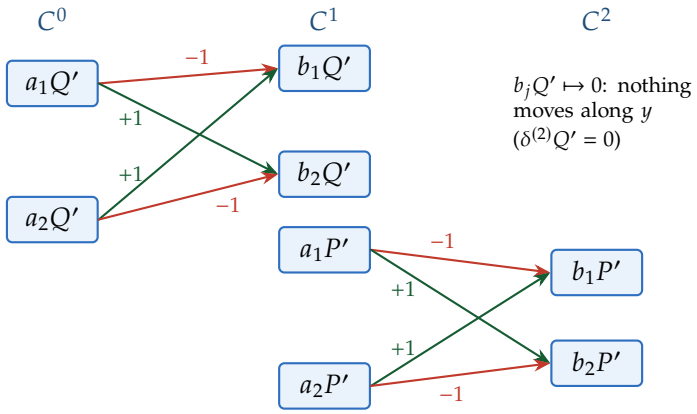


Figure 7.66: The Morse–Witten complex of $f = -\cos 2x + \cos y$ on the torus, as a product. Each arrow is a flow line in which only the x-factor moves; green arrows have $\varepsilon = +1$, red ones $\varepsilon = -1$. The top block is the circle complex of $-\cos 2x$ times the minimum Q' of $\cos y$; the bottom block is the same complex times the maximum P'. The flow lines in the y-direction come in cancelling pairs and are not drawn. The two blocks each kill one pair, leaving $b_\bullet = (1, 2, 1)$.

Why the cohomology does not depend on f or g. The complex $(C^\bullet, \delta)$ certainly depends on the choices: $f = \cos\theta$ and $f = -\cos 2\theta$ on the circle give complexes of different sizes. There are two ways to see that the cohomology does not.

Quantum mechanics. At large t the cohomology of the low-lying complex is the space of exact ground states (7.239), and those have dimension $b_p = \dim H^p_{\mathrm{dR}}(M)$ for *every* t (Section 7.12.3), by the chain isomorphism $e^{tf}\colon (\Omega^\bullet, \mathrm{d}_t) \to (\Omega^\bullet, \mathrm{d})$. The de Rham cohomology knows nothing about f or g. So the complex may change, but only by adding or removing pairs of generators joined by an invertible entry of δ: states that are lifted together.

Geometry: birth and death of pairs. Deform f through a family f_s. For generic families only one thing can happen to the critical points, and it happens in a one-dimensional slice. The local model is

$$f_s(x) = \tfrac{1}{3}x^3 - s\,x, \qquad f_s'(x) = x^2 - s. \tag{7.252}$$

For $s < 0$ there are no critical points. At $s = 0$ there is one degenerate critical point (an inflection). For $s > 0$ there are two:

$$x_\pm = \pm\sqrt{s}, \qquad f_s''(x_\pm) \overset{(1)}{=} \pm 2\sqrt{s}, \qquad f_s(x_-) - f_s(x_+) \overset{(2)}{=} \tfrac{4}{3}\,s^{3/2}.$$

(1) $f_s'' = 2x$. So x_+ is a minimum (index 0) and x_- a maximum (index 1).
(2) $f_s(\pm\sqrt{s}) = \pm\tfrac{1}{3}s^{3/2} \mp s^{3/2} = \mp\tfrac{2}{3}s^{3/2}$.

Exactly one flow line runs from x_- down to x_+, in the direction $+\partial_x$, so $\varepsilon(x_+, x_-) = +1$: $\delta x_+ = x_- + \dots$, an invertible entry. (In n dimensions one adds $\sum \pm y_j^2$ in the other directions; the two new critical points have indices k and $k+1$.) The new pair of generators cancels in cohomology. This is the **birth–death** (or cancellation) of critical points, the geometric fact behind the invariance: Morse theory "with handles" (Chapter 7) says that a cancelling pair of handles can be removed without changing the manifold.

Worked example 7.99: the birth of a pair, in quantum mechanics

On the line, with $f = f_s$ of (7.252), the 0-form Hamiltonian is $H_0 = \mathrm{d}_t^*\mathrm{d}_t = -\partial_x^2 + t^2 f'^2 - t f''$ ((7.181)). Rescale $x = t^{-1/3}y$:

$$H_0 \overset{(1)}{=} t^{2/3}(-\partial_y^2) + t^2\left(t^{-2/3}y^2 - s\right)^2 - 2t\,t^{-1/3}y \overset{(2)}{=} t^{2/3}\left[-\partial_y^2 + (y^2 - \sigma)^2 - 2y\right], \qquad \sigma := s\,t^{2/3}.$$

(1) $\partial_x = t^{1/3}\partial_y$, $f' = x^2 - s$, $f'' = 2x$.
(2) Factor out $t^{2/3}$: $t^2(t^{-2/3}y^2 - s)^2 = t^{2/3}(y^2 - st^{2/3})^2$.

So the whole family depends on the single combination σ, and all energies are $t^{2/3}g(\sigma)$. For $\sigma \gg 1$ the barrier between the two critical points is high, and (7.216) gives the lifted energy of the pair,

$$E \overset{(1)}{=} |\langle x_-|\mathrm{d}_t|x_+\rangle|^2 \overset{(2)}{=} \frac{t}{\pi}\,(2\sqrt{s}\cdot 2\sqrt{s})^{1/2}\,e^{-\frac{8}{3}ts^{3/2}} \overset{(3)}{=} t^{2/3}\cdot\frac{2}{\pi}\sqrt{\sigma}\,e^{-\frac{8}{3}\sigma^{3/2}}.$$

(1) (7.211) with a 1×1 matrix D_0.
(2) (7.216) with $f''(x_+)\,|f''(x_-)| = 4s$ and barrier $\frac{4}{3}s^{3/2}$, squared.
(3) $t\sqrt{s} = t^{2/3}\sqrt{\sigma}$ and $ts^{3/2} = \sigma^{3/2}$.

The numerically computed lowest eigenvalue of the bracket (Figure 7.67) agrees with this for $\sigma \gtrsim 1$. As σ decreases the barrier shrinks, the energy of the pair grows, and for $\sigma < 0$ it has merged with the rest of the spectrum: there is nothing low-lying left. The same energy appears in the 1-form sector (supersymmetric pairing), so the pair never contributes to the Witten index or to cohomology. The birth of two critical points is, in quantum mechanics, a boson–fermion pair descending from the high part of the spectrum.

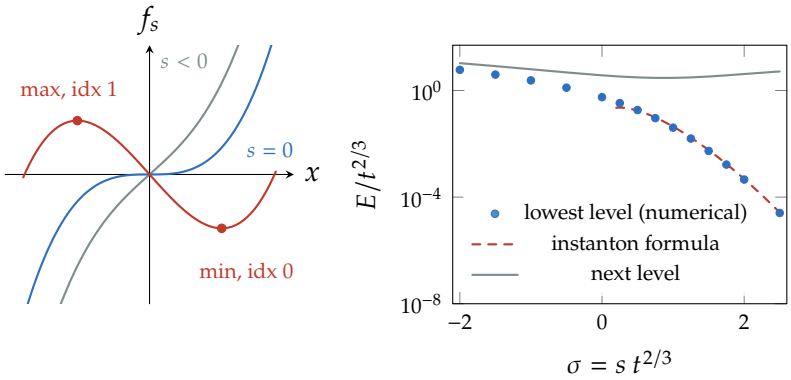


Figure 7.67: Birth of a pair of critical points. *Left:* $f_s = \frac{1}{3}x^3 - sx$ for $s < 0$ (grey, no critical point), $s = 0$ (blue, one degenerate point) and $s > 0$ (red, a maximum and a minimum joined by one flow line). *Right:* the lowest eigenvalue of the rescaled Hamiltonian $-\partial_y^2 + (y^2 - \sigma)^2 - 2y$ (dots), which becomes exponentially small as the pair is born, compared with the instanton formula (dashed). The next level (grey) stays of order one. The pair is always lifted, so the Witten index and the cohomology never notice it.

Continuation maps. There is a third, more constructive argument, which is the one that generalises to Floer theory. Given two Morse–Smale pairs (f_0, g_0) and (f_1, g_1), interpolate by a family f_s, g_s with $f_s = f_0$ for $s \le 0$ and $f_s = f_1$ for $s \ge 1$, and count solutions of the *time-dependent* flow $\dot\gamma(s) = -\nabla_{g_s} f_s(\gamma(s))$ running from a critical point of f_0 (at $s = -\infty$) to one of f_1 of the same index (at $s = +\infty$). Such solutions are isolated (no time-shift symmetry any more), and their signed count is a matrix $\Phi\colon C^\bullet(f_0) \to C^\bullet(f_1)$. Looking at the ends of the one-dimensional families, exactly as in Section 7.16.5, gives $\Phi\delta_0 = \delta_1\Phi$: Φ is a chain map. Running the interpolation backwards gives a chain map the other way, and composing the two is chain-homotopic to the identity. So Φ induces an isomorphism of cohomology. We do not need this in finite dimensions, where de Rham cohomology is available, but in Floer theory it is the only proof that the answer is an invariant.

7.16.8 Where the construction goes next: Floer's infinite-dimensional version

Nash closes the section with the remark that "the instanton ideas employed for this purpose will be useful again in the somewhat analogous infinite dimensional Floer theory" [Nash p. 215]. It is worth seeing now what survives and what breaks when M is replaced by an infinite-dimensional space, because it explains why the flow-line definition of δ, and not the spectral one, is the one that generalises.

The ingredients of the construction, listed. Everything above used four things:

1. a space X and a function $f\colon X \to \mathbb{R}$ whose critical points are isolated and non-degenerate;
2. a grading of the critical points by the index;
3. the spaces of gradient flow lines between critical points, of dimension $\operatorname{index}(b) - \operatorname{index}(a) - 1$, with signs;
4. compactness: families of flow lines can only degenerate by breaking (Section 7.16.5).

Witten's spectral route needs, in addition, the Laplacian Δ_t on forms on X, and so a manifold on which differential forms and a Hodge theory make sense. In infinite dimensions that is exactly what is lost. Items 1–4, on the other hand, can survive.

Floer's example: the Chern–Simons function. Let Y be a closed oriented 3-manifold, $\mathcal{A}$ the space of SU(2) connections on Y (anti-Hermitian $\mathfrak{su}(2)$-valued 1-forms, $F_A = \mathrm{d}A + A \wedge A$), and $\mathcal{G}$ the gauge group. The **Chern–Simons function**

$$\mathrm{CS}(A) = \frac{1}{8\pi^2} \int_Y \operatorname{tr}\Big(A \wedge \mathrm{d}A + \tfrac{2}{3} A \wedge A \wedge A \Big) \tag{7.253}$$

is gauge invariant up to integers, so it is a function (mod $\mathbb{Z}$) on $X = \mathcal{A}/\mathcal{G}$. Its first variation is

$$\left.\frac{\mathrm{d}}{\mathrm{d}\eta}\right|_{\eta=0} \mathrm{CS}(A + \eta\, a) \overset{(1)}{=} \frac{1}{8\pi^2} \int_Y \operatorname{tr}\big(a \wedge \mathrm{d}A + A \wedge \mathrm{d}a + 2\, a \wedge A \wedge A \big) \overset{(2)}{=} \frac{1}{8\pi^2} \int_Y \operatorname{tr}\big(2\, a \wedge \mathrm{d}A + 2\, a \wedge A \wedge A \big) \overset{(3)}{=} \frac{1}{4\pi^2} \int_Y \operatorname{tr}\big(a \wedge F_A \big).$$

(1) Differentiate each term; the cubic term gives three equal terms by cyclicity of the trace (moving a 1-form past a 2-form costs no sign).
(2) Integrate by parts: $\mathrm{d}\operatorname{tr}(A \wedge a) = \operatorname{tr}(\mathrm{d}A \wedge a) - \operatorname{tr}(A \wedge \mathrm{d}a)$ and $\partial Y = \varnothing$, so $\int \operatorname{tr}(A \wedge \mathrm{d}a) = \int \operatorname{tr}(\mathrm{d}A \wedge a) = \int \operatorname{tr}(a \wedge \mathrm{d}A)$.
(3) $F_A = \mathrm{d}A + A \wedge A$.

So the critical points of CS are the **flat connections**, $F_A = 0$, and with the inner product $\langle a, b \rangle = -\int_Y \operatorname{tr}(a \wedge \star b)$ the gradient is $\nabla \mathrm{CS} = \mp \frac{1}{4\pi^2} \star F_A$ (the sign depends on the orientation convention). Two things are new:

- *The Hessian has infinitely many negative and infinitely many positive eigenvalues.* It is the operator $a \mapsto \star \mathrm{d}_A a$ on 1-forms (plus gauge fixing), a first-order elliptic operator like the Dirac operator, whose spectrum is unbounded in both directions. The Morse index is infinite. What is finite is the *difference* of indices between two critical points: the number of eigenvalues that cross zero along a path from one to the other, the **spectral flow**. Floer grades the complex by this relative index, which is all that items 2 and 3 need. Because

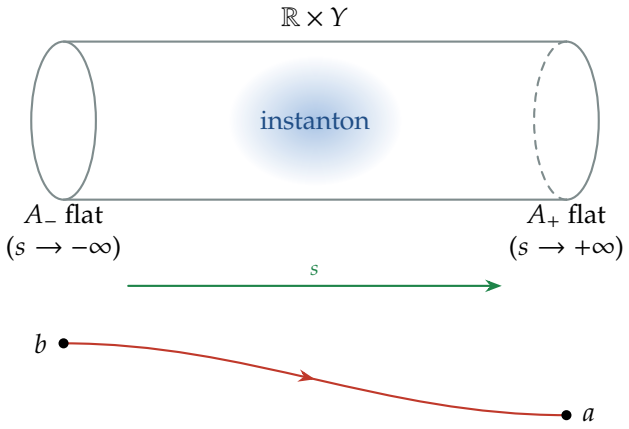
CS is only defined modulo integers on $\mathcal{A}/\mathcal{G}$, a loop in $\mathcal{A}/\mathcal{G}$ can change the relative index; for SU(2) the change is a multiple of 8 (an index-theorem computation), so the grading is by $\mathbb{Z}_8$ and there are eight Floer groups.

- *The flow lines are instantons.* A path $A(s)$ of connections on Y is a connection on the cylinder $\mathbb{R}\times Y$ (in the gauge with no $\mathrm{d}s$ component), with curvature $F = \mathrm{d}s\wedge\partial_s A + F_{A(s)}$. With the orientation $\mathrm{d}s\wedge\mathrm{vol}_Y$, the 4D Hodge star acts as $\star_4(\mathrm{d}s\wedge\alpha) = \star_3\alpha$ and $\star_4\beta = \mathrm{d}s\wedge\star_3\beta$ for a 1-form α and a 2-form β on Y. So

$$F = \mp\star_4 F \iff \partial_s A = \mp\star_3 F_{A(s)}:$$

the (anti-)self-duality equations on the cylinder are the gradient flow of CS (up to a positive constant). The Yang–Mills action of such a flow line is, by the Bogomolny argument of Section 7.15.5, $8\pi^2$ times the drop of CS between its ends: the analogue of (7.91), $S = t(f(b) - f(a))$.

Floer's differential counts these instantons on $\mathbb{R}\times Y$ between flat connections whose relative index is one, and $\delta^2 = 0$ is proved by the broken-flow-line argument of Section 7.16.5, now requiring the compactness theory of instantons (Uhlenbeck) of Chapter 8. The result, **instanton Floer homology** $HF(Y)$, is an invariant of the 3-manifold. It is the space of states of a topological field theory in the sense of Chapter 12: in Witten's language it is the space of supersymmetric ground states of a quantum field theory on $\mathbb{R}\times Y$, exactly as $H^\bullet(M)$ is the space of supersymmetric ground states of the quantum mechanics of this chapter.



gradient flow line of f on M: a quantum-mechanical instanton

Figure 7.68: Floer's dictionary. *Top:* a flow line of the Chern–Simons function is a path of connections $A(s)$ on Y, i.e. a connection on the cylinder $\mathbb{R}\times Y$; the flow equation is the (anti-)self-duality equation, so the flow line is a Yang–Mills instanton, concentrated where $A(s)$ changes, interpolating between flat connections $A_\pm$. *Bottom:* the finite-dimensional template of this section, a tunnelling path from b to a whose action is $t(f(b) - f(a))$.

Dictionary: Witten's quantum mechanics and Floer theory

Morse–Witten (this chapter)	instanton Floer theory (Chapter 12)
compact manifold M	$\mathcal{A}/\mathcal{G}$ for a 3-manifold Y
Morse function f	Chern–Simons function CS
critical points	flat connections
index index(b)	relative index: spectral flow of $\star \mathrm{d}_A$
gradient flow $\dot\gamma = -\nabla f$	anti-self-dual connection on $\mathbb{R}\times Y$
action $t\,(f(b)-f(a))$	$8\pi^2\ \lvert\Delta\mathrm{CS}\rvert$, the instanton number
$\delta^2 = 0$ from broken flow lines	$\delta^2 = 0$ from Uhlenbeck compactness
$H^\bullet(M)$ = SUSY ground states of QM	$HF(Y)$ = SUSY ground states of QFT on $\mathbb{R}\times Y$

The symplectic version. On a symplectic manifold (P,ω) with a time-periodic Hamiltonian H_τ, the **action functional** of classical mechanics, $\mathcal{A}(x) = -\int_D u^*\omega + \int_0^1 H_\tau(x(\tau))\,\mathrm{d}\tau$ on contractible loops x (with u a disc bounding x), has as critical points the 1-periodic orbits of the Hamiltonian flow: this is Hamilton's principle. Its flow lines are maps $u(s,\tau)$ of a cylinder solving a perturbed Cauchy–Riemann equation $\partial_s u + J(\partial_\tau u - X_H) = 0$. Counting them gives symplectic Floer homology, which turns out to be $H^\bullet(P)$. The strong Morse inequalities then say that a non-degenerate Hamiltonian has at least $\sum_p b_p(P)$ periodic orbits: the **Arnold conjecture**, proved by Floer in this way. It is the Morse inequality of this chapter, applied on the loop space of P, and it is the most famous application of the instanton picture of Morse theory.

Reading Nash (pp. 211–213 and 215, (7.81)–(7.86), (7.95))

Nash states the "additional mathematics" in one sentence: if a coboundary $\delta\colon C_p \to C_{p+1}$ with $\delta^2 = 0$ can be built on the integral combinations of critical points (7.81)–(7.82), its cohomology is that of M and the Morse inequalities follow. We made the second half explicit as rank–nullity (7.236). Nash then defines δ geometrically: steepest-descent paths (7.83), the splitting $T_b^- M = T_b\gamma_{ab} \oplus O_b$ (7.84), the sign $\varepsilon_{\gamma_{ab}}$ from orienting O_b by interior multiplication and flowing it to a, the total $\varepsilon(a,b)$ (7.85) and $\delta(a) = \sum_b \varepsilon(a,b)b$ (7.86). He remarks that checking $\delta^2 = 0$ "with the present combinatorial definition is difficult" and appeals to $\mathrm{d}_t^2 = 0$ instead; we gave both that argument (the low-lying complex) and the direct one (broken flow lines). Finally (7.95) recovers (7.86) from the instanton calculation. Nash writes the index as a subscript, C_p, but his δ raises the degree, so it is a cochain complex; and his statement that the complex gives the cohomology of M is true over $\mathbb{Z}$ (for a Morse–Smale metric), which is why the $\mathbb{RP}^2$ example sees the torsion $\mathbb{Z}_2$ that Witten's real forms cannot.

Recap

- Any cochain complex with m_p generators in degree p and cohomology of dimension b_p satisfies $M_u - P_u = (1+u)R(u)$ with $R \geq 0$ (rank–nullity). So the strong Morse inequalities follow from the existence of one such complex.
- Witten's low-lying complex $(V_\bullet, \mathrm{d}_t)$ is such a complex over $\mathbb{R}$. At large t, in the rescaled basis, its matrix is the integer matrix $\varepsilon(a,b)$.
- $\varepsilon(a,b)$ is a signed count of gradient flow lines from b to a (index difference one). The flow lines form a space of dimension index b − index a − 1 for a Morse–Smale metric; the signs come from Nash's orientation rule (interior multiplication by the

exit direction, then transport).

- $\delta^2 = 0$ physically because $\mathrm{d}_t^2 = 0$, geometrically because the ends of one-dimensional families of flow lines are broken flow lines, which cancel in pairs.
- Worked complexes: circles, the flat and tilted torus ($\delta = 0$), the two-humped and six-point spheres (pairs cancel), $\mathbb{RP}^2$ ($\delta_1 = \times(-2)$: torsion over $\mathbb{Z}$, invisible over $\mathbb{R}$).
- Products: $\delta = \delta^{(1)} \otimes 1 + (-1)^{F_1} \otimes \delta^{(2)}$, the same sign as in the supercharge. Changing f creates or annihilates cancelling pairs (birth–death), a boson–fermion pair in the spectrum; the cohomology never changes.
- The flow-line definition generalises to infinite dimensions, where Hodge theory is lost: Floer theory, with Chern–Simons flow lines = instantons on $\mathbb{R} \times Y$.

Chapter summary

The chapter asked one question: *what does the landscape of a single function $f : M \to \mathbb{R}$ know about the shape of M?* The answer is that the critical points of f, counted with their indices, bound the Betti numbers from above, and that the gradient flow lines between them, counted with signs, reconstruct the cohomology exactly.

The first half of the chapter (Nash §§1–3) is classical. The index of a non-degenerate critical point is the number of downhill directions, and passing a critical value of index λ attaches a λ-cell to the sublevel set. Counting cells gives the weak and strong Morse inequalities and the Euler characteristic. Allowing whole critical submanifolds gives the Morse–Bott version; keeping track of a symmetry gives the equivariant version; and allowing M to be a space of paths gives Morse's theory of geodesics and Bott's periodicity theorem.

The second half (Nash §4, Section 7.9 onwards) is Witten's: the same results, and more, from supersymmetric quantum mechanics. The differential forms on M are the states of a particle with fermionic spin, $\mathrm{d} + \mathrm{d}^*$ is the supercharge, the Laplacian is the Hamiltonian, and the zero-energy states are the cohomology. Deforming by e^{tf} localises the low-energy states at the critical points, where they are harmonic oscillators; tunnelling between critical points, computed by instantons, is the Morse–Witten differential.

The two proofs, each as one chain

The whole of Witten's argument fits into one labelled chain. Every step is one section of the second half.

$$b_p \overset{(1)}{=} \dim \operatorname{Ker} \Delta_p \overset{(2)}{=} \dim \operatorname{Ker} \Delta_p(t) \overset{(3)}{=} \dim H^p\left(V_\bullet, \mathrm{d}_t\right)$$
$$\overset{(4)}{=} \dim H^p\left(C^\bullet, \delta\right) \overset{(5)}{=} m_p - r_p - r_{p-1} \overset{(6)}{\leq} m_p. \tag{7.254}$$

(1) The Hodge theorem: every de Rham class has exactly one harmonic representative, and the harmonic forms are the supersymmetric ground states of $H = \mathrm{dd}^* + \mathrm{d}^*\mathrm{d}$ (Section 7.11, Nash (7.74)).

(2) Multiplication by e^{tf} is a chain isomorphism from $(\Omega^\bullet, \mathrm{d}_t)$ to $(\Omega^\bullet, \mathrm{d})$, so the number of zero-energy states does not change with t (Section 7.12, Nash (7.75)).

(3) At large t the spectrum splits into m_p low-lying levels per degree, one per critical point of index p, and levels of order t. d_t preserves the low-lying space $V_\bullet$, and every exact ground state lies in it (Sections 7.14 and 7.16.2).

(4) On $V_\bullet$ the matrix of d_t is a sum over instantons, $\varepsilon(a,b)\,e^{-t(f(b)-f(a))}$, and rescaling the basis removes the exponentials, leaving the integer matrix of δ (Section 7.15, Nash (7.93)–(7.95)).
(5) Rank–nullity, with $r_p = \operatorname{rank}\delta_p$ ((7.235)).
(6) Ranks are non-negative: the weak Morse inequality. Summing $(-1)^p u^p$-weighted versions of step 5 gives the strong form $M_u - P_u = (1+u)R(u)$ ((7.236)).

The classical proof reaches the same inequalities from the other end: the sublevel sets $M_c = \{f \le c\}$ change only at critical values, each change is the attachment of a cell of dimension equal to the index, and the exact sequence of the pair $(M_{c+\varepsilon}, M_{c-\varepsilon})$ changes each Betti number by at most one per cell, in a way that makes the alternating sums subadditive.

The Euler characteristic is the special value where all the cancellations become irrelevant:

$$\chi(M) \overset{(1)}{=} \operatorname{Tr}(-1)^F \overset{(2)}{=} \sum_p (-1)^p b_p \overset{(3)}{=} \sum_p (-1)^p m_p \overset{(4)}{=} M_{-1}(f). \tag{7.255}$$

(1) The Witten index of the de Rham supersymmetric quantum mechanics ((7.163)); $(-1)^F$ is $(-1)^{\text{degree}}$.
(2) Only zero-energy states contribute to the Witten index: states with $E > 0$ come in boson–fermion pairs (Section 7.10.6).
(3) Deformation invariance: at large t each critical point carries one low-lying state of degree equal to its index, and lifted states still pair (Section 7.14.6).
(4) The definition of the Morse polynomial at $u = -1$.

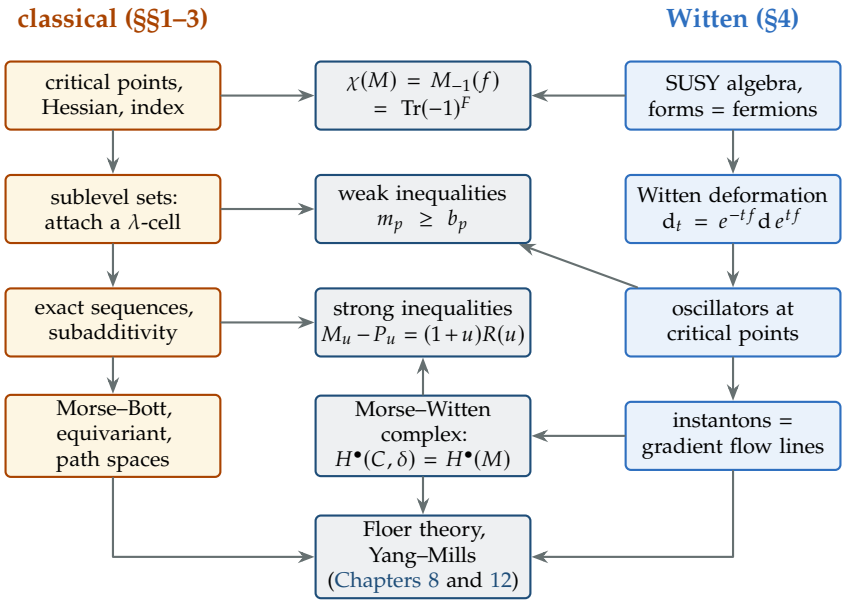


Figure 7.69: The logic of the chapter. The classical route (orange, left) and Witten's quantum route (blue, right) arrive at the same results (middle). The quantum route passes through the Morse–Witten complex, which proves the strong inequalities by linear algebra, and it is the version of the argument that survives in the infinite-dimensional settings of later chapters.

Every concept, how to use it, and where it returns

The tables follow Nash's four sections. The middle column is the *operational* meaning: what we actually do to compute or test the thing. The right column says where the idea is developed further or used again. In this chapter the variable of the Morse and Poincaré polynomials is written t in §§1–3, as in Nash, and u in §4, where t is Witten's deformation parameter.

Nash §1: the topology of critical points

Concept	Operational meaning	Returns in
critical point	a point where $\mathrm{d}f = 0$: solve $\partial_i f = 0$ in a chart	everywhere; Chapter 8 (Yang–Mills solutions)
Hessian, non-degenerate	the matrix $\partial_i \partial_j f$ at a critical point; non-degenerate iff $\det \neq 0$	Section 7.14.2 (oscillator frequencies)
index λ	number of negative eigenvalues of the Hessian (Sylvester: chart independent); number of downhill directions	Section 7.14.4 (degree of the zero mode)
nullity, Morse function	nullity $= n - \operatorname{rank} Hf$; Morse = all critical points non-degenerate; generic: $f + a \cdot x$ is Morse for almost every a	Section 7.16.7 (birth–death)
Morse lemma	near a non-degenerate critical point $f = f(p) - x_1^2 - \cdots - x_\lambda^2 + x_{\lambda+1}^2 + \cdots + x_n^2$	Section 7.14.2, Nash (7.78)
Morse series $M_t(f)$	$\sum_p t^{\lambda_p} = \sum_i m_i t^i$: count critical points by index	Section 7.16.1
Poincaré series $P_t(M)$	$\sum_i b_i t^i$ with $b_i = \dim H^i(M; F)$; depends on the coefficient field F when there is torsion	Example 7.96
weak inequalities	$m_i \geq b_i$ for every i	Section 7.14.5
strong inequalities	$M_t - P_t = (1+t)R(t)$ with $R \geq 0$; equivalently the alternating partial sums $m_j - m_{j-1} + \cdots \geq b_j - b_{j-1} + \ldots$	Section 7.16.1 (rank–nullity)
Euler characteristic	$\chi = M_{-1}(f) = P_{-1}(M)$, the same for every f	Section 7.14.6, Chapter 4 (Gauss–Bonnet, Hopf)
lacunary principle	no two consecutive powers in $M_t \Rightarrow R = 0$ and $M_t = P_t$	Example 7.94
perfect Morse function	$M_t = P_t$; impossible over $\mathbb{Z}$ if $H^\bullet(M; \mathbb{Z})$ has torsion	Example 7.96
sublevel set M_c	$\{f \leq c\}$; changes topology only at critical values	Section 7.16.3
cell attachment, handles	crossing a critical value of index λ attaches a λ-cell ($D^\lambda \times D^{n-\lambda}$ handle); Nash's stratification $M = \bigcup e_\lambda$	Section 7.16.5, Chapter 12 (cobordisms)
gradient flow	$\dot\gamma = -\nabla f$; f decreases along it, and it connects critical points	Section 7.15.5 (instantons)
$\mathbb{CP}^n$, Σ_p, upright torus	$\sum a_i \lvert z_i\rvert^2$ has indices $0, 2, \ldots, 2n$ (lacunary); the height on Σ_p has $M_t = 1 + 2pt + t^2$	Chapter 5, Chapter 9
degenerate critical points	e.g. the monkey saddle $\operatorname{Re} z^3$; perturb to split into non-degenerate ones	Example 7.99
LS category	least number of contractible closed sets covering M; a lower bound for the number of critical points of *any* f	–

Nash §2: critical submanifolds

Concept	Operational meaning	Returns in
non-degenerate critical submanifold	a connected C with $\mathrm{d}f\|_C = 0$ whose Hessian is non-degenerate on the normal bundle $N(C)$	Chapter 8 (critical sets of Yang–Mills)
$N = N^+ \oplus N^-$, index λ_C	split the normal bundle by the sign of the Hessian; $\lambda_C = \operatorname{rank} N^-$	–
Morse–Bott series	$M_t(f) = \sum_C t^{\lambda_C} P_t(C)$ (orientable N^-, or $\mathbb{Z}_2$ coefficients)	Chapter 8 (Atiyah–Bott)
dinner-plate torus, z^2 on S^2	two critical circles give $(1+t)+t(1+t) = P_t(T^2)$ (perfect); z^2 gives $M_t - P_t = t(1+t)$ (not perfect)	–
perturbing to Morse	add a small Morse function on C: each critical point of it of index μ becomes one of index $\lambda_C + \mu$	Section 7.16.7

Nash §3: equivariant Morse theory, and infinite dimensions

Concept	Operational meaning	Returns in
group action, stabiliser	G moves points of M; free iff all stabilisers are trivial; then M/G is a manifold	Chapter 6, Chapter 8 (gauge group)
naive quotient fails	$S^2/\mathrm{U}(1) = [0,1]$ has $P_t = 1$: fixed points are forgotten	–
universal bundle $EG \to BG$	EG contractible with free G-action; $E\mathrm{U}(1) = S^\infty$, $B\mathrm{U}(1) = \mathbb{CP}^\infty$, $B\mathbb{Z}_2 = \mathbb{RP}^\infty$	Chapter 10 (classifying spaces)
homotopy quotient M_G	$(M \times EG)/G$; equals M/G up to homotopy if the action is free	–
equivariant cohomology	$H_G(M) = H(M_G)$; $H_G(\mathrm{pt}) = H(BG)$, $P_t(B\mathrm{U}(1)) = 1/(1-t^2)$	Chapter 12 (localisation)
equivariant Morse series	$\sum_N P_t(N_G)\, t^{\lambda_N}$ over critical manifolds N; inequalities $M_t^G - P_t^G = (1+t)R$	Chapter 8 (Yang–Mills on a Riemann surface is equivariantly perfect)
equivariantly perfect	e.g. z^2 on the rotating S^2: $(1+t^2)/(1-t^2)$ on both sides, though not perfect	–
Cartan model, localisation	equivariantly closed forms $\mathrm{d}\alpha = \iota_X \alpha$-type condition; Duistermaat–Heckman: stationary phase is exact	Chapter 12
energy functional	$E(\gamma) = \int \lvert\dot\gamma\rvert^2$ on paths $p \to q$; critical points are geodesics; $E \geq L^2$	Chapter 9 (Polyakov vs. Nambu–Goto)
second variation, Jacobi fields	Hessian of E is $-\nabla_s^2 - R(\cdot, \dot\gamma)\dot\gamma$; its kernel are the Jacobi fields	–
Morse index theorem	index of a geodesic = number of conjugate points along it, with multiplicity	Chapter 10 (spectral flow)
loop space ΩS^n	$P_t = 1/(1-t^{n-1})$: infinitely many geodesics between two points, for every metric	Chapter 6 (loop groups)
Palais–Smale	the compactness condition that lets Morse theory work in infinite dimensions	Section 7.16.8
closed geodesics	critical points of E on $\mathrm{Map}(S^1, M)$; U(1) rotates the loop, so the theory is equivariant	–
Bott periodicity	minimal geodesics in $\mathrm{SU}(2m)$ are Grassmannians; $\Omega^2\mathrm{U} \simeq \mathbb{Z} \times B\mathrm{U}$ (period 2), period 8 for O	Chapter 10, Chapter 4 (K-theory)
Reeb, Milnor, Smale	two critical points $\Rightarrow$ sphere; exotic 7-spheres; handle cancellation $\Rightarrow$ Poincaré conjecture in dim ≥ 5	Chapter 12

Nash §4: supersymmetric quantum mechanics (Witten)

Concept	Operational meaning	Returns in
supersymmetry algebra	$\{Q_i, Q_j\} = 0$ $(i \neq j)$, $Q_1^2 = H + P$, $Q_2^2 = H - P$; so $H \geq 0$	Section 7.10, Chapter 9, Chapter 12
fermion number $(-1)^F$	$+1$ on bosons, -1 on fermions; the Q_i are off-diagonal in the split $\mathcal{H}^B \oplus \mathcal{H}^F$	Section 7.10.4
nilpotent supercharge	for $P = 0$, $Q = \frac{1}{2}(Q_1 - iQ_2)$ has $Q^2 = 0$; in the forms model $Q = \mathrm{d}$ (or d_t)	Section 7.10.3
SUSY QM on a line	$Q = \partial + W'$ on functions; partner Hamiltonians $-\partial^2 + W'^2 \mp W''$; zero modes $e^{\mp W}$	Section 7.10.5, Section 7.12.5
Witten index	$\mathrm{Tr}(-1)^F = \mathrm{Tr}(-1)^F e^{-\beta H}$ counts ground states with sign; $E > 0$ levels pair; deformation invariant	Section 7.10.6, Chapter 4 (heat-kernel proof), Chapter 10
forms model	$\mathcal{H} = \Omega^\bullet(M)$, $Q_1 = \mathrm{d} + \mathrm{d}^*$, $Q_2 = i(\mathrm{d} - \mathrm{d}^*)$, $H = \Delta$; ground states = harmonic forms	Section 7.11.2, Section 7.11.3
forms as fermions	$a^{i*} = \mathrm{d}x^i \wedge$, $a^i = \iota_{\partial_i}$ obey $\{a^i, a^{j*}\} = \delta^{ij}$; degree = fermion number	Section 7.11.4
Clifford operators	$c^i = a^{i*} - a^i$, $\hat{c}^i = a^{i*} + a^i$; $\mathrm{d} + \mathrm{d}^*$ is a Dirac-type operator	Section 7.13.2, Chapter T4, Chapter 4
Witten deformation	$\mathrm{d}_t = e^{-tf}\,\mathrm{d}\,e^{tf} = \mathrm{d} + t\,\mathrm{d}f\wedge$, $\mathrm{d}_t^* = \mathrm{d}^* + t\,\iota_{\nabla f}$; same Betti numbers	Section 7.12
deformed Laplacian	$H_t = \Delta + t^2\,\lvert\mathrm{d}f\rvert^2 + t\,(D^2 f/Dx^i Dx^j)[a^{i*}, a^j]$: a Schrödinger operator with a spin–Hessian coupling	Section 7.13
local oscillators	at a critical point $H_t \approx \sum_i (-\partial_i^2 + t^2\lambda_i^2 x_i^2 + t\lambda_i[a^{i*}, a^i])$; spectrum $t\sum(\lvert\lambda_i\rvert\,(1 + 2n_i) + \lambda_i\varepsilon_i)$	Section 7.14.4
one zero mode per critical point	Gaussian times $\mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^\lambda$ (downhill directions): degree = index	Section 7.14.5
low-lying complex	V_p = span of eigenforms with $E \leq ct$; $\dim V_p = m_p$; $\mathrm{d}_t : V_p \to V_{p+1}$	Section 7.16.2
tunnelling amplitude	$\langle b\rvert\mathrm{d}_t\lvert a\rangle \approx \varepsilon\,\sqrt{t/\pi}\,(f''(a)\,\lvert f''(b)\rvert)^{1/4} e^{-t(f(b)-f(a))}$ in 1D	Section 7.15, Chapter 8
Bogomolny bound	for a path from b down to a, $S_B = \frac{1}{2}\,\lVert\dot{x} + t\nabla f\rVert^2 + t\,(f(b) - f(a)) \geq t\,(f(b) - f(a))$; minimisers are gradient flow lines	Section 7.15.5, Chapter 8 (BPST)
determinant cancellation	bosonic and fermionic fluctuation determinants cancel up to the sign ε_γ; fermion zero modes select index difference one	Chapter 10 (zero modes)
stable/unstable manifolds, Morse–Smale	$W^u(b)$, $W^s(a)$ of dimensions index b, $n -$ index a; transverse $\Rightarrow$ flow lines form a space of dimension index $b -$ index $a - 1$	Section 7.16.3
orientation rule	orient O_b by $\iota_v\omega_b$, transport to a, compare with ω_a: $\varepsilon_\gamma = \pm 1$	Section 7.16.4
Morse–Witten complex	$\delta(a) = \sum_b \varepsilon(a,b)\,b$; $\delta^2 = 0$ by broken flow lines; $H(C,\delta) = H(M;\mathbb{Z})$	Section 7.16, Chapter 12
products, birth–death	$\delta = \delta^{(1)} \otimes 1 + (-1)^{F_1} \otimes \delta^{(2)}$; pairs of critical points appear and disappear together	Section 7.16.7
Floer theory	the same complex for CS on $\mathcal{A}/\mathcal{G}$: relative index by spectral flow, flow lines = instantons on $\mathbb{R} \times Y$	Section 7.16.8, Chapter 12

The examples, side by side

Every example of the chapter is one row of Table 7.2. Reading down the last column is a quick way to recall what each one teaches. (We write the Morse and Poincaré polynomials in u; $R(u) = (M_u - P_u)/(1+u)$ counts the cancelling pairs, and real coefficients are meant unless stated.)

M, f	$M_u(f)$	$P_u(M)$	$R(u)$	what it shows
S^1, $\cos\theta$	$1+u$	$1+u$	0	two instantons with opposite signs cancel (Example 7.84)
S^1, $-\cos 2\theta$	$2+2u$	$1+u$	1	one pair lifted, $E = \frac{16t}{\pi}e^{-4t}$ (Example 7.83)
$\mathbb{R}$, double well	$2+u$	1	1	tunnelling splitting $2\sqrt{2}\frac{t}{\pi}e^{-t/2}$ (Example 7.82)
$\mathbb{R}$, $\frac{1}{3}x^3 - sx$	$1+u$ or 0	1	–	birth–death of a pair (Example 7.99; not compact)
S^2, height	$1+u^2$	$1+u^2$	0	lacunary, perfect; exact zero modes e^{-tf}, $e^{tf}\,\mathrm{vol}$
S^2, two humps	$1+u+2u^2$	$1+u^2$	u	the saddle cancels one peak (Example 7.93)
S^2, $x^2+2y^2+3z^2$	$2+2u+2u^2$	$1+u^2$	$1+u$	two cancelling pairs (Example 7.95)
$\mathbb{RP}^2$, same f	$1+u+u^2$	1 (over $\mathbb{R}$)	u	$\delta_1 = \times(-2)$: $H^2(\mathbb{RP}^2;\mathbb{Z}) = \mathbb{Z}_2$ (Example 7.96)
T^2, $\cos x + \cos y$	$1+2u+u^2$	$1+2u+u^2$	0	$\delta = 0$, four exact ground states (Examples 7.91 and 7.97)
T^2, upright height	$1+2u+u^2$	$1+2u+u^2$	0	perfect, but not Morse–Smale until tilted (Figure 7.62)
T^2, $-\cos 2x + \cos y$	$2+4u+2u^2$	$1+2u+u^2$	$1+u$	product of complexes (Example 7.98)
$\mathbb{CP}^n$, $\sum a_i\lvert z_i\rvert^2$	$1+u^2+\cdots+u^{2n}$	same	0	lacunary: counting is enough
Σ_p, height	$1+2pu+u^2$	same	0	perfect; $\chi = 2-2p$
S^2, z^2 (Morse–Bott)	$1+u+2u^2$	$1+u^2$	u	a critical circle contributes $P_u(S^1)$
$S^2_{\mathrm{U}(1)}$, z^2	$\frac{1+u^2}{1-u^2}$	same	0	equivariantly perfect, though not perfect
ΩS^n, energy	$\frac{1}{1-u^{n-1}}$	same	0	perfect in infinite dimensions; infinitely many geodesics

Table 7.2: Every example of Chapter VII in one table: Morse polynomial, Poincaré polynomial, the cancelling-pair polynomial $R(u)$, and the lesson.

Dictionary: Morse theory and supersymmetric quantum mechanics

The whole of §4 can be compressed into the translation table Table 7.3. Read left to right it says what each piece of Morse theory *is* physically; read right to left it says what each piece of the quantum mechanics computes topologically.

Where Chapter VII goes next

- *Yang–Mills theory* (Chapter 8). The Yang–Mills functional on the space of connections is a Morse function in infinite dimensions. Its minima are the instantons; on a Riemann

Morse theory	supersymmetric quantum mechanics
differential forms on M	wavefunctions of a particle with n fermionic modes
degree p of a form	fermion number
d, d^*	supercharges Q, $Q^\dagger$
Hodge Laplacian Δ	Hamiltonian $H = \{Q, Q^\dagger\}$
harmonic forms, b_p	supersymmetric ground states in the p-fermion sector
Euler characteristic $\chi(M)$	Witten index $\mathrm{Tr}(-1)^F$
Morse function f, deformation e^{tf}	superpotential tf; $t \sim 1/\hbar$
critical point of index λ	classical vacuum; oscillator with λ fermions filled
weak Morse inequalities	perturbative ground states $\geq$ exact ground states
gradient flow line	instanton (tunnelling path)
$\varepsilon(a, b)$	sign of the fermion determinant
Morse–Witten differential δ	d_t restricted to the low-lying states
cancelling pair of critical points	boson–fermion pair lifted by tunnelling
$\delta^2 = 0$	$Q^2 = 0$
birth–death of critical points	a supermultiplet entering or leaving the low-energy spectrum

Table 7.3: The dictionary between Morse theory and supersymmetric quantum mechanics that Chapter VII §4 establishes.

surface Atiyah and Bott showed it is equivariantly perfect with respect to the gauge group, the equivariant Morse theory of Nash §3 in its most famous application. The Bogomolny bound $S \geq 8\pi^2 |k|$ is the gauge-theory version of the completed square of Section 7.15.5.

- *Index theory and anomalies* (Chapters 4 and 10). The Witten index is the index of the Dirac-type operator $\mathrm{d} + \mathrm{d}^*$ from even to odd forms. The same supersymmetric trick, $\mathrm{Tr}(-1)^F e^{-\beta H}$ independent of β, gives the heat-kernel proof of the index theorem. The Morse index theorem and the relative Morse index are instances of spectral flow, which computes anomalies.
- *Loop groups and strings* (Chapters 6 and 9). The energy functional on loop spaces and Bott periodicity return with the loop groups ΩG and with the string action.
- *Topological field theory* (Chapter 12). Floer homology is the Morse–Witten complex of the Chern–Simons function; Witten's interpretation of Donaldson theory treats it as the space of supersymmetric ground states of a twisted $N = 2$ gauge theory, exactly as $H^\bullet(M)$ was the space of ground states here.

Self-test

Try it

1. A Morse function on a closed surface has 3 minima and 2 maxima. How many saddles does it have, if the surface is a torus? a sphere?
2. Why can a Morse function on $\mathbb{RP}^2$ never be perfect over $\mathbb{Z}$, but can be over $\mathbb{Z}_2$?
3. At large t, what is the energy (to leading order) of the local state $\Psi_{n,I}$ at a critical point with Hessian eigenvalues $\lambda = (-1, 2)$, with $n = (0, 0)$ and the fermion in the x_2

direction occupied?

4. Why does $\langle b|\mathrm{d}_t|a\rangle$ vanish unless index b = index a + 1?

5. Why is the cross term in $Q_1(t)^2$ for a product manifold zero?

Answers. (1) $\chi = m_0 - m_1 + m_2$: on T^2, $0 = 3 - m_1 + 2$, so 5 saddles; on S^2, $2 = 3 - m_1 + 2$, so 3 saddles. (2) $H^2(\mathbb{RP}^2;\mathbb{Z}) = \mathbb{Z}_2$ is torsion, and a perfect function would give a complex with $\delta = 0$ and free cohomology; over $\mathbb{Z}_2$ the function $x^2 + 2y^2 + 3z^2$ has $\delta = 0$ mod 2 (Example 7.96). (3) By (7.200), $E = t\{(1\cdot 1 + (-1)(-1)) + (2\cdot 1 + 2\cdot(+1))\} = t(2+4) = 6t$, since the x_1-fermion is empty ($\varepsilon_1 = -1$) and the x_2-fermion is filled ($\varepsilon_2 = +1$). The zero mode instead fills the downhill direction x_1. (4) d_t raises the degree by one and the local state at b has degree index b; in the path integral, the instanton has index b − index a fermion zero modes, and d_t supplies exactly one fermion. (5) The supercharge of the second factor carries a factor $(-1)^{F_1}$, and $Q^{(1)}$ anticommutes with it ((7.250)).

Chapter 8

Instantons and Monopoles

Where this chapter is going

Chapter 8 is about the two most famous classical solutions of non-Abelian gauge theory: *instantons* (finite-action solutions of the Euclidean Yang–Mills equations in four dimensions) and *monopoles* (finite-energy static solutions of Yang–Mills–Higgs theory in three space dimensions). Nash is not mainly interested in writing the solutions down. He wants the *topology*: of the fields, of the gauge group, of the space of all fields modulo gauge transformations, and of the spaces of solutions (the moduli spaces).

The chapter starts from one simple fact: the space $\mathcal{A}$ of all gauge potentials is a vector space in disguise, so it has no topology at all. All the interesting topology comes from dividing by the gauge group $\mathcal{G}$. The quotient $\mathcal{A}/\mathcal{G}$ is the true configuration space of a gauge theory, and it is topologically rich: it is the classifying space $B\mathcal{G}$. On a sphere S^n its homotopy groups are homotopy groups of the Lie group G, shifted in degree. This is where instanton numbers, large gauge transformations, the θ-vacua and the Gribov ambiguity come from. Section 2 then builds explicit differential forms for the cohomology of $\mathcal{G}$ and $\mathcal{A}/\mathcal{G}$ out of the curvature, by the Chern–Simons (transgression) construction. The same Chern–Simons forms give the topologically massive gauge theory in three dimensions, and in Chapter 10 they give the anomalies. Sections 3–6 (the later parts of this chapter) turn to the instanton and monopole moduli spaces and to Morse theory on $\mathcal{A}/\mathcal{G}$.

Reading Nash (pp. 216–227)

Nash's §1 "The topology of gauge fields" runs from p. 216 to the middle of p. 223, equations (8.1)–(8.38). His §2 "Secondary characteristic classes" runs to the top of p. 227, equations (8.39)–(8.62). That is only eleven pages, and nearly every line hides a construction from algebraic topology or a calculation. In these notes we unpack all of them. We also add the physics Nash only alludes to: the winding number of $g^{-1}\mathrm{d}g$ at infinity, large gauge transformations and the θ-vacua.

8.1 The topology of gauge fields

What is the space of all gauge fields, and what does it look like? A physicist usually meets gauge fields one at a time, as solutions of field equations or as integration variables in a path integral. Here we step back and look at the *set* of all of them. There are three reasons to care.

- The path integral $Z = \int \mathcal{D}A\, e^{-S[A]}$ is an integral over this space. If the space has holes or twists, the integral knows about them.

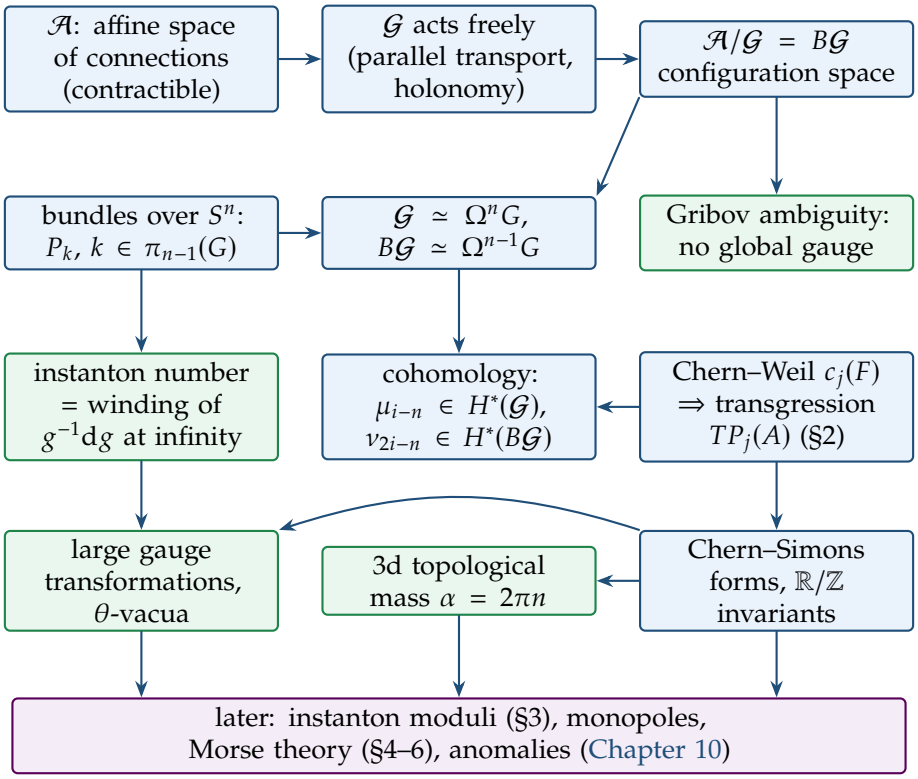


Figure 8.1: The logic of this part (Nash §§1–2) and where it leads. Blue boxes are mathematical constructions, green boxes are their physical consequences. Read the top row first: contractible $\mathcal{A}$, free action of $\mathcal{G}$, and the quotient $B\mathcal{G}$. Everything below it computes the topology of that quotient and gives it a physical meaning.

- Gauge-equivalent potentials describe the same physics. So the physical configuration space is the space of potentials divided by gauge transformations. Dividing can create topology where there was none.

- Topological invariants of gauge fields (instanton number, monopole charge) are exactly the labels of the "pieces" of these spaces.

This section answers four questions in turn. What is the space of connections $\mathcal{A}$? (An affine space, so topologically trivial.) How does the gauge group $\mathcal{G}$ act on it? (Freely, if we use the based gauge group.) What is the quotient $\mathcal{A}/\mathcal{G}$? (The classifying space $B\mathcal{G}$; the bundle $\mathcal{A} \to \mathcal{A}/\mathcal{G}$ is twisted, which is the Gribov ambiguity.) And what are the homotopy and cohomology of $\mathcal{G}$ and $B\mathcal{G}$ when the base is a sphere? (They are those of G, shifted.)

8.1.1 Conventions and a reminder about principal bundles

Throughout this part M is a smooth, connected, oriented manifold of dimension n (in the examples usually S^n or $\mathbb{R}^n$), and G is a compact Lie group, usually $\mathrm{U}(N)$ or $\mathrm{SU}(N)$, realised as a group of $N \times N$ matrices. Its Lie algebra $\mathfrak{g}$ consists of *anti-Hermitian* matrices.

Principal bundles in one paragraph

A **principal G-bundle** $\pi : P \to M$ is a space P with a free right action of G, $p \mapsto pg$, whose orbits are the fibres $P_x = \pi^{-1}(x)$.[1] Each fibre is a copy of G (but without a preferred identity element). Over a small open set $U \subset M$ we choose a **local section** $s : U \to P$. Every point over U can then be written uniquely as $p = s(x)g$, and the pair (x, g) are **local coordinates** on P. Nash writes exactly these coordinates in (8.3). If U_i, U_j are two patches with sections s_i, s_j, then on the overlap

$$s_j(x) = s_i(x)\, t_{ij}(x), \qquad t_{ij} : U_i \cap U_j \to G, \tag{8.1}$$

and the t_{ij} are the **transition functions**. They obey the cocycle condition $t_{ij}t_{jk} = t_{ik}$ on triple overlaps. A bundle is **trivial**, $P \cong M \times G$, exactly when one section can be chosen on all of M. (Many books write g_{ij} or $g_{\alpha\beta}$ for the transition functions; we keep the letter g for gauge transformations.) Figure 8.2 shows every ingredient of this paragraph at once.

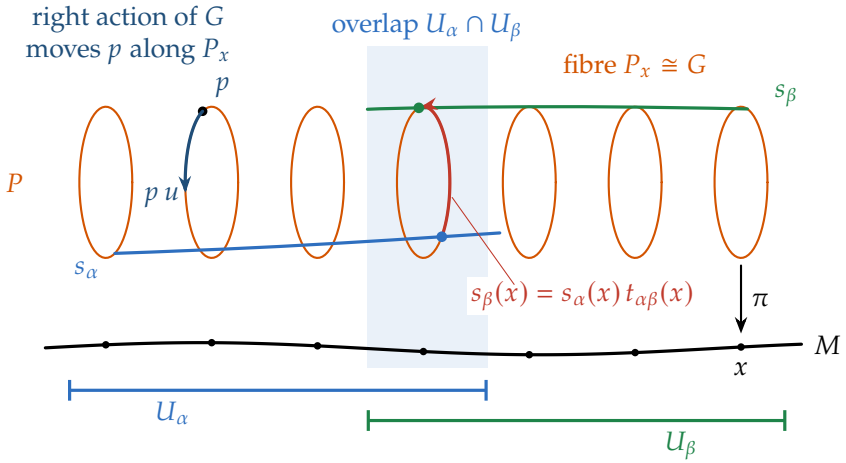


Figure 8.2: A principal bundle $\pi : P \to M$ for $G = \mathrm{U}(1)$, so that every fibre is a circle. What to see: each fibre $P_x = \pi^{-1}(x)$ (orange) is a copy of G sitting over one point x; π squashes the whole fibre to x. The right action of $u \in G$ slides a point p around its own fibre and never leaves it (dark blue arrow). A local section $s_\alpha : U_\alpha \to P$ (blue) picks one point in every fibre over the patch U_α, and so gives the trivialisation $p = s_\alpha(x)u \leftrightarrow (x, u)$; likewise s_β (green) over U_β. On the overlap (shaded) both sections exist, and they differ by the group element $t_{\alpha\beta}(x)$, the transition function (red arrow, often written $g_{\alpha\beta}$). The bundle is twisted exactly when no choice of sections makes all the transition functions equal to $\mathbb{1}$.

Connections: global on P, local on M

A connection can be described in two equivalent ways, and we will need both.

- *On P:* a $\mathfrak{g}$-valued 1-form ω on the total space P which, on vertical vectors (tangent to a fibre), returns the Lie algebra element that generates them, and which transforms as $R_g^*\omega = g^{-1}\omega g$ under the right action. This ω is defined *globally* on P, even when P is twisted.

[1] Vocabulary of fibre bundles (defined in Chapter 1, all drawn in Figure 8.2). The *total space* is P (for a general bundle often called E); the *base* is M; the *projection* $\pi : P \to M$ sends each point to the point of M it sits over. The *fibre over* x is $P_x = \pi^{-1}(x)$, and every fibre is a copy of one *typical fibre* F; for a principal bundle $F = G$. A *local trivialisation* is an identification $\pi^{-1}(U) \cong U \times F$ that respects π. A *(local) section* is a map $s : U \to P$ with $\pi \circ s = \mathrm{id}_U$: a choice of one point in each fibre over U. The *transition functions* compare two trivialisations on an overlap $U_i \cap U_j$; for principal bundles they are the t_{ij} of (8.1).

- *On M:* the pull-backs $A_i = s_i^*\omega$ by the local sections. These are the physicist's gauge potentials. They live on patches and are glued by

$$A_j = t_{ij}^{-1} A_i \, t_{ij} + t_{ij}^{-1} \mathrm{d} t_{ij}. \tag{8.2}$$

In the local coordinates $p = (x, g)$ the two descriptions are related by

$$\omega_{(x,g)} = g^{-1} A_i(x)\, g + g^{-1}\mathrm{d}g. \tag{8.3}$$

The curvature is $F = \mathrm{d}A + A \wedge A$ on each patch, and it transforms covariantly, $F_j = t_{ij}^{-1} F_i t_{ij}$. That is why $\operatorname{tr} F^k$ and all other invariant polynomials in F are globally defined forms on M, even though A is not. We will use this over and over in Section 8.2. For more background see Chapter T3.

Covariant derivatives: the transformation law dictates the formula

We will differentiate several kinds of objects: matter fields ψ, gauge parameters ε, the curvature F. The covariant derivative $D\psi = \mathrm{d}\psi + A\psi$, the transformation law $A \mapsto A^g = g^{-1}Ag + g^{-1}\mathrm{d}g$, its infinitesimal form $\delta A = \mathrm{d}\varepsilon + [A, \varepsilon]$, the curvature as D^2 and the Bianchi identity $DF = 0$ were developed in Section T3.8.3 and the rest of Section T3.8, which also has the dictionary to the Hermitian conventions of physics texts (Section T3.8.9). One rule organises all the formulas: *the covariant derivative of an object must transform exactly as the object does.* The bare d spoils this, because it also differentiates g, so we add terms built from A that cancel the damage. How many terms, and with which signs, depends only on how many factors of g the object carries and on which side.

Key idea: One-sided versus two-sided transformation laws

With $A \mapsto g^{-1}Ag + g^{-1}\mathrm{d}g$,

$$\psi \mapsto g^{-1}\psi \quad\Longrightarrow\quad D\psi = \mathrm{d}\psi + A\psi, \tag{8.4}$$

$$\Phi \mapsto g^{-1}\Phi\, g \quad\Longrightarrow\quad D\Phi = \mathrm{d}\Phi + A\wedge\Phi - (-1)^p\,\Phi\wedge A =: \mathrm{d}\Phi + [A,\Phi]. \tag{8.5}$$

The first line is for a **fundamental** (one-sided) object, such as a quark field. The second is for an **adjoint** (two-sided) p-form, such as F ($p = 2$) or an infinitesimal gauge parameter ε ($p = 0$). The term $+A\wedge\Phi$ repairs the derivative of the *left* factor g^{-1}; the term $-(-1)^p\Phi\wedge A$ repairs the derivative of the *right* factor g.

The fundamental case is checked in Section T3.8.3. Here is the check of the two-sided case, which shows each correction term doing its job. Nothing is needed beyond the product rule and $\mathrm{d}(g^{-1}) = -g^{-1}\mathrm{d}g\, g^{-1}$. Put $\Phi' = g^{-1}\Phi g$ and $A' = A^g$. Moving d past the p-form Φ costs $(-1)^p$, so $\mathrm{d}(\Phi g) = \mathrm{d}\Phi\, g + (-1)^p\,\Phi\wedge\mathrm{d}g$, and

$$\begin{aligned}
\mathrm{d}\Phi' &\overset{(1)}{=} -g^{-1}\mathrm{d}g\, g^{-1}\wedge\Phi g + g^{-1}\mathrm{d}\Phi\, g + (-1)^p\, g^{-1}\Phi\wedge\mathrm{d}g,\\
A'\wedge\Phi' &\overset{(2)}{=} g^{-1}A\wedge\Phi\, g + g^{-1}\mathrm{d}g\, g^{-1}\wedge\Phi g,\\
\Phi'\wedge A' &\overset{(3)}{=} g^{-1}\Phi\wedge A\, g + g^{-1}\Phi\wedge\mathrm{d}g,\\
\Longrightarrow\quad \mathrm{d}\Phi' + A'\wedge\Phi' - (-1)^p\,\Phi'\wedge A' &\overset{(4)}{=} g^{-1}\big(\mathrm{d}\Phi + A\wedge\Phi - (-1)^p\,\Phi\wedge A\big)g.
\end{aligned} \tag{8.6}$$

(1) Product rule on $g^{-1}\cdot\Phi\cdot g$, with $\mathrm{d}(g^{-1}) = -g^{-1}\mathrm{d}g\, g^{-1}$ and the sign rule just stated.
(2) Multiply $A' = g^{-1}Ag + g^{-1}\mathrm{d}g$ into $g^{-1}\Phi g$ from the left; the inner gg^{-1} cancel.

(3) Multiply from the right: $g^{-1}\Phi g \wedge g^{-1}Ag = g^{-1}\Phi \wedge Ag$ and $g^{-1}\Phi g \wedge g^{-1}\mathrm{d}g = g^{-1}\Phi \wedge \mathrm{d}g$.
(4) Add the first two lines and subtract $(-1)^p$ times the third. The term $g^{-1}\mathrm{d}g\, g^{-1} \wedge \Phi g$ (from differentiating the left g^{-1}) cancels against the one in $+A' \wedge \Phi'$; the term $(-1)^p g^{-1}\Phi \wedge \mathrm{d}g$ (from differentiating the right g) cancels against the one in $-(-1)^p\Phi' \wedge A'$.

So $D\Phi \mapsto g^{-1}(D\Phi)g$, as it must. For $\Phi = \varepsilon$ ($p = 0$) this is $D\varepsilon = \mathrm{d}\varepsilon + A\varepsilon - \varepsilon A$, the combination in the infinitesimal gauge transformation $A^{e^\varepsilon} \approx A + \mathrm{d}\varepsilon + [A, \varepsilon]$ (Section 8.1.10). For $\Phi = F$ ($p = 2$) it is $DF = \mathrm{d}F + A \wedge F - F \wedge A$, and its vanishing is the Bianchi identity (T3.61); we use it in Section 8.2 as the reason $\mathrm{d}\,\mathrm{tr}\, F^k = 0$. (This way of explaining the signs follows a Physics Stack Exchange answer, question 483197, answer 483214.)

The graded commutator. For a matrix-valued p-form α and q-form β the **graded commutator** is

$$[\alpha, \beta] := \alpha \wedge \beta - (-1)^{pq}\, \beta \wedge \alpha. \tag{8.7}$$

For two 0-forms it is the matrix commutator, and for $\alpha = A$ ($p = 1$) it is exactly the bracket in (8.5). Unlike the ordinary commutator, it does not vanish when $\alpha = \beta$ has odd degree. For the gauge potential, $A = A_\mu \mathrm{d}x^\mu$,

$$\begin{aligned}[A, A] &\overset{(1)}{=} A \wedge A - (-1)^{1\cdot 1} A \wedge A = 2\, A \wedge A \\ &\overset{(2)}{=} 2\, A_\mu A_\nu\, \mathrm{d}x^\mu \wedge \mathrm{d}x^\nu \\ &\overset{(3)}{=} [A_\mu, A_\nu]\, \mathrm{d}x^\mu \wedge \mathrm{d}x^\nu.\end{aligned}$$

(1) Definition (8.7) with $p = q = 1$.
(2) Expand both factors; the matrices A_μ stay in order, the $\mathrm{d}x$'s are wedged.
(3) Split $A_\mu A_\nu$ into symmetric and antisymmetric parts in $\mu\nu$; the symmetric part dies against the antisymmetric $\mathrm{d}x^\mu \wedge \mathrm{d}x^\nu$, leaving $\frac{1}{2}[A_\mu, A_\nu]$, times 2.

So $A \wedge A$ is zero for U(1) and nonzero as soon as the components fail to commute. For SU(2) with $e_k = -i\sigma_k$ (so $e_1 e_2 = -e_2 e_1 = e_3$), the potential $A = e_1\, \mathrm{d}x + e_2\, \mathrm{d}y$ has $A \wedge A = (e_1 e_2 - e_2 e_1)\, \mathrm{d}x \wedge \mathrm{d}y = 2e_3\, \mathrm{d}x \wedge \mathrm{d}y$. In this notation the curvature is

$$F = \mathrm{d}A + A \wedge A = \mathrm{d}A + \tfrac{1}{2}[A, A]. \tag{8.8}$$

Watch out: A is not covariant, and F is not "$D_A A$"

If A were an adjoint 1-form, (8.5) would give its covariant derivative as $\mathrm{d}A + [A, A] = \mathrm{d}A + 2A \wedge A$, which is *not* F. The factor $\frac{1}{2}$ in (8.8) is the sign that something different is going on: A transforms inhomogeneously, $A \mapsto g^{-1}Ag + g^{-1}\mathrm{d}g$, so it is not a two-sided object and the rule of (8.5) does not apply to it; A has no covariant derivative at all. One sometimes sees "$\mathrm{d}_A A = \mathrm{d}A + A \wedge A = F$"; this is only a notation for F, not a covariant derivative of A. What is remarkable is that F, built from the non-covariant A, does transform covariantly, $F \mapsto g^{-1}Fg$: the inhomogeneous pieces cancel between $\mathrm{d}A$ and $A \wedge A$. Differences of connections, $A - A'$, are two-sided (Section 8.1.2), and for them $D(A - A')$ makes perfect sense.

Dictionary: anti-Hermitian (Nash) versus Hermitian (physics) fields

We use Nash's convention: A is anti-Hermitian, $F = \mathrm{d}A + A \wedge A$, and a gauge transformation acts as

$$A \longmapsto g^{-1}Ag + g^{-1}\mathrm{d}g.$$

The physicists' field is Hermitian: $A_{\text{math}} = -i\, g_{\text{YM}}\, A^a_{\text{phys}} T^a$ with Hermitian generators T^a, normalised by $\operatorname{tr}(T^a T^b) = \frac{1}{2}\delta^{ab}$ for SU(N). Then $F_{\text{math}} = -i\, g_{\text{YM}} F^a_{\text{phys}} T^a$ and

$$-\operatorname{tr}\big(F_{\text{math}} \wedge \star F_{\text{math}}\big) = \tfrac{1}{2}\, g_{\text{YM}}^2\, F^a_{\text{phys}} \wedge \star F^a_{\text{phys}}.$$

For $G = \mathrm{U}(1)$ and $g = e^{i\lambda}$ the rule becomes $A \mapsto A + i\,\mathrm{d}\lambda$, i.e. $a \mapsto a - \mathrm{d}\lambda/e$ for the real potential a with $A = -iea$.

Working definition 8.1 (quaternion notation for SU(2))

We use the conventions of the rest of this chapter for SU(2) [the 2×2 complex matrices g with $g^\dagger g = \mathbb{1}$ and $\det g = 1$]. Write $e_k = -i\sigma_k$ ($k = 1,2,3$) with the Pauli matrices

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Then $e_1 e_2 = e_3$ (and cyclically) and $e_k^2 = -1$: the e_k are the quaternion units $\mathbf{i}, \mathbf{j}, \mathbf{k}$. A point $x = (x^1, x^2, x^3, x^4) \in \mathbb{R}^4$ is identified with the 2×2 matrix

$$x = x^4\,\mathbb{1} + x^k e_k = \begin{pmatrix} x^4 - ix^3 & -x^2 - ix^1 \\ x^2 - ix^1 & x^4 + ix^3 \end{pmatrix}, \qquad \det x = |x|^2 . \tag{8.9}$$

Its conjugate is $\bar{x} = x^\dagger = x^4\mathbb{1} - x^k e_k$, and $\bar{x}x = |x|^2\,\mathbb{1}$. So the unit sphere $S^3 = \{|x| = 1\}$ *is* the group SU(2). The orientation of $\mathbb{R}^4$ [the choice of which 4-form counts as positive, which fixes the sign of every integral over $\mathbb{R}^4$] is $\mathrm{d}x^1 \wedge \mathrm{d}x^2 \wedge \mathrm{d}x^3 \wedge \mathrm{d}x^4$.

Watch out: Signs of c_2 and of the instanton number

For SU(N) with anti-Hermitian F, the second Chern class is $c_2(F) = \frac{1}{8\pi^2}\operatorname{tr}(F \wedge F)$. For a self-dual field this is *negative*, because $\operatorname{tr}(e_a e_b) = -2\delta_{ab}$. We follow the rest of this chapter and Nash's formula (8.61) and define the **instanton number**

$$k := -\frac{1}{8\pi^2}\int_M \operatorname{tr}(F \wedge F) = -c_2, \tag{8.10}$$

which is $+1$ for the standard (BPST) instanton. Nash's polynomial $P_2(F) = -\frac{1}{8\pi^2}\operatorname{tr}(F\wedge F)$ in (8.60) is therefore $-c_2$, not c_2. This does not matter for anything topological (a sign never changes whether a class vanishes), but it matters when we compare numbers.

8.1.2 The space of all connections is an affine space

Let $\mathcal{A}$ denote the set of all connections on a fixed bundle $P \to M$. What kind of space is it?

A first guess is "a vector space": gauge potentials are 1-forms, and 1-forms can be added. But this is wrong in an instructive way. If A and A' are connections, then $A + A'$ is *not* a connection. Look at the gluing rule (8.2):

$$\begin{aligned}(A + A')_j &\overset{(1)}{=} \big(t^{-1}A_i t + t^{-1}\mathrm{d}t\big) + \big(t^{-1}A'_i t + t^{-1}\mathrm{d}t\big) \\ &= t^{-1}(A + A')_i\, t + 2\,t^{-1}\mathrm{d}t.\end{aligned}$$

(1) Each of A and A' obeys the gluing rule (8.2) with the same transition function $t = t_{ij}$.

The inhomogeneous term $t^{-1}\mathrm{d}t$ appears twice, so $A + A'$ glues with the wrong rule. Likewise $2A$ is not a connection, and neither is 0 on a nontrivial bundle (the zero potential in one patch is not zero in another).

What *does* work is the difference. Subtracting,

$$\begin{aligned}(A - A')_j &\overset{(1)}{=} \left(t^{-1}A_i t + t^{-1}\mathrm{d}t\right) - \left(t^{-1}A'_i t + t^{-1}\mathrm{d}t\right)\\ &= t^{-1}(A - A')_i\, t.\end{aligned} \tag{8.11}$$

(1) Gluing rule (8.2) for both connections.

The inhomogeneous terms cancel. So $a = A - A'$ transforms like the curvature does, by conjugation only. Such an object is a globally defined 1-form with values in the bundle of Lie algebras $\operatorname{ad} P = P \times_G \mathfrak{g}$ (the adjoint bundle): an element of $\Omega^1(M; \operatorname{ad} P)$. This is an honest vector space.

Working definition 8.2 (affine space)

A set $\mathcal{A}$ is an **affine space** modelled on a vector space V [here $V = \Omega^1(M; \operatorname{ad} P)$, the differences $a = A - A'$ of two connections] if you can subtract two points to get a vector, $A - A' \in V$, and add a vector to a point to get a point, $A + a \in \mathcal{A}$, with the usual rules. *In practice:* an affine space is a vector space in which nobody has told you where the origin is. Choose any point A_0 as an origin; then every point is $A = A_0 + a$ for a unique vector a.

Key idea: The space of connections

The space $\mathcal{A}$ of connections on P is an affine space modelled on $\Omega^1(M; \operatorname{ad} P)$. Choosing a reference connection A_0,

$$\mathcal{A} = \left\{A_0 + a \;:\; a \in \Omega^1(M; \operatorname{ad} P)\right\}. \tag{8.12}$$

This is Nash's first statement. The "origin" A_0 is arbitrary. On a trivial bundle the obvious choice is $A_0 = 0$, and then $\mathcal{A}$ really is the vector space of $\mathfrak{g}$-valued 1-forms. On a nontrivial bundle (a monopole or instanton bundle) there is no such canonical choice.

Straight lines and contractibility

In an affine space we cannot add points, but we can take *weighted averages* whose weights sum to one. Nash's equation (8.1) is exactly this. For two connections A_1, A_2 and $t \in \mathbb{R}$, set

$$A_t := (1 - t)A_1 + tA_2 = A_1 + t\,(A_2 - A_1). \tag{8.13}$$

Is A_t a connection? Check the gluing rule:

$$\begin{aligned}(A_t)_j &\overset{(1)}{=} (1-t)\left(t_{ij}^{-1}A_{1,i}t_{ij} + t_{ij}^{-1}\mathrm{d}t_{ij}\right) + t\left(t_{ij}^{-1}A_{2,i}t_{ij} + t_{ij}^{-1}\mathrm{d}t_{ij}\right)\\ &\overset{(2)}{=} t_{ij}^{-1}\left((1-t)A_{1,i} + tA_{2,i}\right)t_{ij} + \left((1-t)+t\right)t_{ij}^{-1}\mathrm{d}t_{ij}\\ &\overset{(3)}{=} t_{ij}^{-1}(A_t)_i\, t_{ij} + t_{ij}^{-1}\mathrm{d}t_{ij}.\end{aligned}$$

(1) Each A_a glues by (8.2).
(2) Collect the conjugated terms and the inhomogeneous terms separately.
(3) The weights sum to 1, so the inhomogeneous term appears exactly once, which is the correct gluing rule.

So $A_t \in \mathcal{A}$ for every real t: there is a whole straight line through any two connections.

Now fix any $A_0 \in \mathcal{A}$ and define

$$H : \mathcal{A} \times [0,1] \to \mathcal{A}, \qquad H(A,s) = (1-s)A + sA_0. \tag{8.14}$$

At $s = 0$ this is the identity map of $\mathcal{A}$, and at $s = 1$ it sends everything to the single point A_0. It is continuous (it is affine in both arguments). A space with such a map is **contractible**: it can be shrunk continuously to a point. Consequently every loop, every sphere, every closed surface drawn in $\mathcal{A}$ can be shrunk to a point, and

$$\pi_j(\mathcal{A}) = 0 \quad \text{for all } j \geq 0, \qquad H^j(\mathcal{A}) = 0 \quad \text{for all } j \geq 1. \tag{8.15}$$

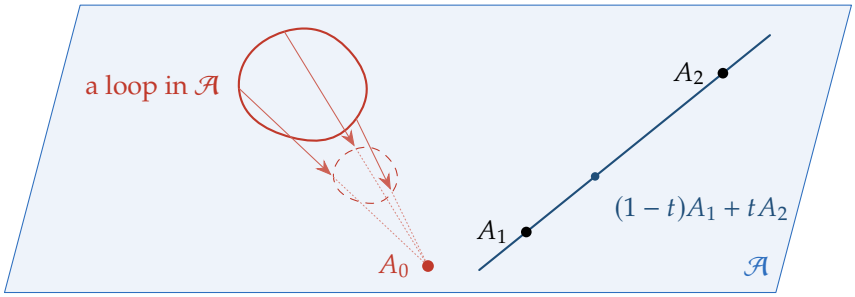


Figure 8.3: The space of connections is a (very large) flat space without an origin. Any two points A_1, A_2 are joined by the straight line (8.1). A loop (red) is shrunk to the point A_0 along straight lines (thin arrows, which end on the dashed loop: the halfway stage $s = \frac{1}{2}$ of (8.14)), so $\mathcal{A}$ has no topology. All the topology in this chapter comes from the gauge group, not from $\mathcal{A}$.

Worked example 8.1: Connections on the trivial U(1) bundle over S^1

Let $M = S^1$ with angle $\phi \in [0, 2\pi)$ and $G = \mathrm{U}(1)$. Every U(1) bundle over S^1 is trivial (a transition function would be a map from a disconnected overlap into the connected group U(1), which can be deformed to 1). A connection is then a single global 1-form

$$A = i\,a(\phi)\,\mathrm{d}\phi, \qquad a : S^1 \to \mathbb{R} \text{ smooth},$$

anti-Hermitian because it is i times a real form. So $\mathcal{A}$ is literally the vector space $C^\infty(S^1;\mathbb{R})$ of periodic real functions, with origin $A_0 = 0$. A straight line is $a_t(\phi) = (1-t)a_1(\phi) + t a_2(\phi)$. Contractibility is obvious: $a \mapsto (1-s)a$ shrinks everything to 0.
This tiny example will reappear several times. It is the simplest place to see a nontrivial quotient $\mathcal{A}/\mathcal{G}$ (Example 8.10).

Worked example 8.2: Why 0 is not a connection on the monopole bundle

Cover S^2 by the northern and southern patches $U_N = S^2 \setminus \{\mathrm{S}\}$ and $U_S = S^2 \setminus \{\mathrm{N}\}$, glued along the equator by $t_{NS}(\phi) = e^{im\phi}$ with $m \in \mathbb{Z}$. The standard Dirac monopole connection is

$$A_N = -\tfrac{im}{2}(1-\cos\theta)\,\mathrm{d}\phi, \qquad A_S = \tfrac{im}{2}(1+\cos\theta)\,\mathrm{d}\phi.$$

Check the gluing rule at the equator:

$$\begin{aligned} t_{NS}^{-1} A_N t_{NS} + t_{NS}^{-1}\mathrm{d}t_{NS} &\overset{(1)}{=} A_N + im\,\mathrm{d}\phi \\ &\overset{(2)}{=} \Big(-\tfrac{im}{2}(1-\cos\theta) + im\Big)\mathrm{d}\phi \\ &= \tfrac{im}{2}(1+\cos\theta)\,\mathrm{d}\phi = A_S. \end{aligned}$$

(1) U(1) is Abelian, so conjugation does nothing, and $e^{-im\phi}\mathrm{d}(e^{im\phi}) = im\,\mathrm{d}\phi$.
(2) Substitute A_N.

Now try the "zero connection" $A_N = 0$, $A_S = 0$. The gluing rule would demand $A_S = 0 + im\,\mathrm{d}\phi$, which is not zero if $m \neq 0$. So on the monopole bundle, zero is not a point of $\mathcal{A}$. Any other connection has the form $A = A_{\text{Dirac}} + a$ with $a = i\,\alpha$ for a *global* real 1-form α on S^2. The curvature is $F = \mathrm{d}A_N = -\frac{im}{2}\sin\theta\,\mathrm{d}\theta\wedge\mathrm{d}\phi + i\,\mathrm{d}\alpha$, and

$$\int_{S^2}\frac{i}{2\pi}F \overset{(1)}{=} \frac{m}{4\pi}\int_{S^2}\sin\theta\,\mathrm{d}\theta\wedge\mathrm{d}\phi - \frac{1}{2\pi}\int_{S^2}\mathrm{d}\alpha \overset{(2)}{=} m.$$

(1) Multiply out $\frac{i}{2\pi}\cdot(-\frac{im}{2}) = \frac{m}{4\pi}$ and $\frac{i}{2\pi}\cdot i = -\frac{1}{2\pi}$.
(2) The area of the unit sphere is 4π; the exact form $\mathrm{d}\alpha$ integrates to zero by Stokes, because S^2 has no boundary.

So every point of $\mathcal{A}$ has the same first Chern number m. Moving around in $\mathcal{A}$ never changes the topology of the bundle; the bundle is fixed once and for all when we define $\mathcal{A}$.

Physical picture: Why a physicist should not be surprised

In a path integral we integrate over "all field configurations". For a scalar field that is a vector space, and nobody expects topology there. The statement that $\mathcal{A}$ is affine says the same is true of gauge potentials, *as long as we do not divide by gauge transformations*. Perturbation theory lives happily in $\mathcal{A}$: expand around a background A_0 and integrate over the fluctuation a. The topological effects (instantons, θ-vacua, Gribov copies) appear only once we insist that gauge-equivalent fields be counted once.

Watch out: Infinite dimensions

$\mathcal{A}$ is infinite-dimensional. To make "continuous" and "smooth" precise one completes it in a Sobolev norm (see Chapter 6). For the topology none of this matters: straight-line contractions work in any topological vector space.

8.1.3 Gauge transformations and the two gauge groups

The space $\mathcal{A}$ has no topology of its own. The next ingredient is the group that acts on it. What exactly is a gauge transformation on a bundle that may be twisted?

On a single patch the physicist's answer is familiar: a gauge transformation is a function $g : U \to G$, and it acts on the potential by

$$A \longmapsto A^g := g^{-1}Ag + g^{-1}\mathrm{d}g. \tag{8.16}$$

On a nontrivial bundle we have one such function g_i on each patch U_i. They cannot be chosen independently. The transformed potentials $A_i^{g_i}$ must still glue by the rule (8.2), and that forces

$$g_j = t_{ij}^{-1}\,g_i\,t_{ij} \qquad \text{on } U_i\cap U_j. \tag{8.17}$$

Let us check that (8.17) is enough. Write $t = t_{ij}$ and use $\mathrm{d}(t^{-1}) = -t^{-1}\mathrm{d}t\,t^{-1}$:

$$\begin{aligned} t^{-1}A_i^{g_i}t + t^{-1}\mathrm{d}t &\overset{(1)}{=} t^{-1}g_i^{-1}A_ig_it + t^{-1}g_i^{-1}\mathrm{d}g_i\,t + t^{-1}\mathrm{d}t \\ &\overset{(2)}{=} g_j^{-1}\big(t^{-1}A_it\big)g_j + g_j^{-1}t^{-1}\mathrm{d}g_i\,t + t^{-1}\mathrm{d}t \\ &\overset{(3)}{=} g_j^{-1}\big(A_j - t^{-1}\mathrm{d}t\big)g_j + g_j^{-1}t^{-1}\mathrm{d}g_i\,t + t^{-1}\mathrm{d}t \end{aligned}$$

$$\overset{(4)}{=} g_j^{-1} A_j g_j + g_j^{-1} \mathrm{d} g_j = A_j^{g_j}.$$

(1) Definition (8.16) of $A_i^{g_i}$.
(2) Insert $\mathbb{1} = tt^{-1}$ twice and use $t^{-1} g_i^{\pm 1} t = g_j^{\pm 1}$ from (8.17); in the middle term write $t^{-1} g_i^{-1} = g_j^{-1} t^{-1}$.
(3) The gluing rule (8.2) says $t^{-1} A_i t = A_j - t^{-1} \mathrm{d} t$.
(4) Differentiate $g_j = t^{-1} g_i t$: $\mathrm{d} g_j = -t^{-1} \mathrm{d} t \, g_j + t^{-1} \mathrm{d} g_i \, t + g_j t^{-1} \mathrm{d} t$. Multiply on the left by g_j^{-1}; the three terms are exactly $-g_j^{-1} t^{-1} \mathrm{d} t \, g_j + g_j^{-1} t^{-1} \mathrm{d} g_i \, t + t^{-1} \mathrm{d} t$, which is what stands on the previous line after $g_j^{-1} A_j g_j$.

So the transformed potentials glue correctly. A collection $\{g_i\}$ obeying (8.17) is a section of the bundle $\mathrm{Ad}\, P$, whose fibre is a copy of G on which the transition functions act by *conjugation* (compare P itself, where they act by left multiplication). On a trivial bundle it is simply a map $M \to G$.

Working definition 8.3 (gauge transformation)

A **gauge transformation** of P is a collection of maps $g_i : U_i \to G$ that glue by conjugation [on each overlap $g_j = t_{ij}^{-1} g_i t_{ij}$, with the same transition functions t_{ij} as P], (8.17); equivalently a section of $\mathrm{Ad}\, P = P \times_{\mathrm{Ad}} G$. It acts on connections by (8.16), patch by patch. It acts on the bundle P itself, in the local coordinates $p = (x, u)$, by

$$\alpha(x, u) = \big(x, \; g_i(x)^{-1} u\big). \tag{8.18}$$

Formally: a bundle automorphism $\alpha : P \to P$ with $\alpha(pu) = \alpha(p)u$ that covers the identity map of M [α maps each fibre P_x to itself, so no point of M is moved] (Nash, Ch. 6, p. 176, and Chapter 6).

The rule (8.18) is patch-independent. A point p has coordinates u_i and u_j in the two patches, with $u_i = t_{ij} u_j$ (because $p = s_j u_j = s_i t_{ij} u_j$). Transforming in patch j and converting back,

$$t_{ij}\big(g_j^{-1} u_j\big) \overset{(1)}{=} t_{ij} t_{ij}^{-1} g_i^{-1} t_{ij} u_j \overset{(2)}{=} g_i^{-1} u_i,$$

(1) Invert (8.17): $g_j^{-1} = t_{ij}^{-1} g_i^{-1} t_{ij}$.
(2) Cancel $t_{ij} t_{ij}^{-1}$ and use $t_{ij} u_j = u_i$.

which is the transformation computed in patch i. The action commutes with the right action of G on P (left and right multiplications commute), as a bundle automorphism must. The inverse in (8.18) is chosen so that the action on connections comes out as (8.16); we will see why in Section 8.1.5.

Two gauge groups

Gauge transformations can be multiplied pointwise, $(gh)_i(x) = g_i(x) h_i(x)$, and they form an infinite-dimensional group. Nash uses two versions of it.

Working definition 8.4 (the full and the based gauge groups)

- $\hat{\mathcal{G}}$ is the group of *all* gauge transformations (sections of $\mathrm{Ad}\, P$). For a trivial bundle, $\hat{\mathcal{G}} = \mathrm{Map}(M, G)$.

- Fix a base point $m_0 \in M$. The **based gauge group** $\mathcal{G}$ [the group Nash divides by, because the only element of it that leaves a connection unchanged is $\mathbb{1}$] is the subgroup of gauge transformations that are the identity on the fibre over m_0:

$$\mathcal{G} = \{ g \in \hat{\mathcal{G}} \ : \ g(m_0) = \mathbb{1} \}. \tag{8.19}$$

Evaluating at the base point is a group homomorphism $\mathrm{ev}_{m_0} : \hat{\mathcal{G}} \to G$, $g \mapsto g(m_0)$. It is onto (take constant maps near m_0, cut off smoothly), and its kernel is $\mathcal{G}$. So

$$1 \longrightarrow \mathcal{G} \longrightarrow \hat{\mathcal{G}} \xrightarrow{\mathrm{ev}_{m_0}} G \longrightarrow 1, \qquad \hat{\mathcal{G}}/\mathcal{G} \cong G. \tag{8.20}$$

The based group loses only a finite-dimensional copy of G: the "global" (constant) gauge transformations at one point.

Why bother? Because $\hat{\mathcal{G}}$ does *not* act freely on $\mathcal{A}$, and $\mathcal{G}$ does. An action is **free** if the only group element that fixes any point is the identity. For $\hat{\mathcal{G}}$ this fails in the most blatant way.

Worked example 8.3: Constant gauge transformations fix connections

Let c be a constant element of the centre $Z(G)$ of G, for example $c = -\mathbb{1} \in \mathrm{SU}(2)$ or any $c = e^{i\lambda} \in \mathrm{U}(1)$. Then for every connection A,

$$A^c \overset{(1)}{=} c^{-1} A c + c^{-1} \mathrm{d} c \overset{(2)}{=} A.$$

(1) Definition (8.16).

(2) c commutes with every matrix in $\mathfrak{g}$, so $c^{-1}Ac = A$; and $\mathrm{d}c = 0$ because c is constant.

So *every* point of $\mathcal{A}$ has a nontrivial stabiliser in $\hat{\mathcal{G}}$. For special connections the stabiliser is larger. Take the trivial SU(2) bundle and $A = 0$. Every constant $c \in \mathrm{SU}(2)$ fixes it, so its stabiliser is the whole of SU(2). Take instead the diagonal potential

$$A = \begin{pmatrix} i\,a & 0 \\ 0 & -i\,a \end{pmatrix}, \qquad a \text{ a real 1-form.}$$

It is fixed by the constant diagonal matrices $c = \mathrm{diag}(e^{i\lambda}, e^{-i\lambda})$ (which commute with A), a copy of U(1). None of these c is in $\mathcal{G}$ unless $c = \mathbb{1}$, because $c(m_0) = c$.

The quotient of a space by a group that does not act freely is singular: the orbits have different dimensions, and the quotient is not a manifold near the special points. Passing to the based group removes the problem in one stroke. Nash proves this by an argument with parallel transport. We first need parallel transport in a form that is easy to compute with.

8.1.4 Parallel transport in local coordinates

Parallel transport is the answer to the question: how do we compare "internal directions" at two different points of spacetime? A connection is precisely the rule that tells us. For the toolkit version see Chapter T3; here we set it up the way Nash uses it.

Take two points $a, b \in M$ and a smooth path between them, Nash's (8.2):

$$\gamma_{ab} : [0,1] \to M, \qquad \gamma_{ab}(0) = a, \quad \gamma_{ab}(1) = b. \tag{8.21}$$

A **lift** of γ_{ab} is a path $\hat{\gamma}_{ab}(t)$ in P lying over it, $\pi(\hat{\gamma}_{ab}(t)) = \gamma_{ab}(t)$. In local coordinates it is Nash's (8.3):

$$\hat{\gamma}_{ab}(t) = \big(\gamma_{ab}(t),\, g(t)\big), \qquad g(t) \in G. \tag{8.22}$$

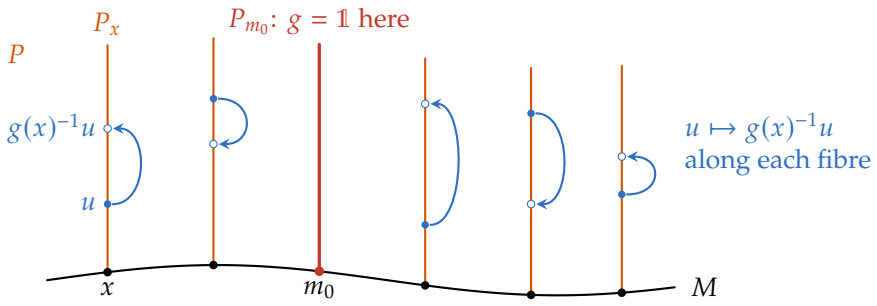


Figure 8.4: A gauge transformation moves each fibre P_x into itself, by left multiplication with $g(x)^{-1}$ in local coordinates, independently at different points (blue arrows). A *based* gauge transformation does nothing on one chosen fibre P_{m_0} (red). This single condition is enough to stop the group from fixing any connection.

There are many lifts: $g(t)$ could be anything. The connection picks out the **horizontal** ones, those whose velocity is annihilated by the connection form ω on P. In local coordinates (8.3) says $\omega = g^{-1}Ag + g^{-1}\mathrm{d}g$. The velocity of the lift is $(\dot\gamma, \dot g)$, so

$$\begin{aligned} 0 &\overset{(1)}{=} \omega\big(\dot\gamma_{ab}\big) \\ &\overset{(2)}{=} g^{-1}A(\dot\gamma_{ab})\,g + g^{-1}\dot g \\ \Longrightarrow \qquad \frac{\mathrm{d}g(t)}{\mathrm{d}t} &\overset{(3)}{=} -A(\dot\gamma_{ab})\,g(t). \end{aligned} \tag{8.23}$$

(1) Definition of a horizontal lift.
(2) Evaluate $\omega = g^{-1}Ag + g^{-1}\mathrm{d}g$ on the vector $(\dot\gamma, \dot g)$: the 1-form A on M eats $\dot\gamma$, and $\mathrm{d}g$ eats $\dot g$.
(3) Multiply on the left by g.

This is Nash's parallel transport equation (8.4), $\dot g + A\dot\gamma_{ab}\,g = 0$. In components, with $A = A_\mu \mathrm{d}x^\mu$, it reads $\dot g = -A_\mu(\gamma(t))\,\dot\gamma^\mu(t)\,g$.

Physical picture: What the equation says

Think of $g(t)$ as the "internal frame" carried along the path. The equation says that the frame changes only as much as the gauge field forces it to. If $A = 0$ the frame stays constant. A physicist meets exactly this equation as the Schrödinger equation of a charged particle's internal state moving along a prescribed path: in the Hermitian convention $A = -iA^a_{\text{phys}}T^a$ it reads $\dot g = iA^a_\mu T^a \dot x^\mu g$, a "time-dependent Hamiltonian" $-A^a_\mu T^a \dot x^\mu$ with t as time.

Solving the equation

Equation (8.23) is a linear, first-order ODE for the matrix $g(t)$. So once $g(0)$ is chosen, $g(t)$ is determined for all t. Moreover, because the equation is linear and g multiplies from the right, the solution has the form

$$g(t) = U(t)\,g(0), \qquad \dot U = -A(\dot\gamma)\,U, \quad U(0) = \mathbb{1}, \tag{8.24}$$

with a single matrix $U(t)$ that does not depend on the initial point. Nash's (8.5)–(8.6) are the endpoints of the lift:

$$\hat\gamma_{ab}(0) = \big(a, g(0)\big), \qquad \hat\gamma_{ab}(1) = \big(b, g(1)\big) = \big(b, U(1)g(0)\big). \tag{8.25}$$

In the *Abelian* case the equation is solved by an exponential. For $G = \mathrm{U}(1)$ write $A = i\,a_\mu \mathrm{d}x^\mu$ with real a_μ. Then

$$\frac{\mathrm{d}}{\mathrm{d}t}\ln U(t) \overset{(1)}{=} -i\,a_\mu(\gamma(t))\dot\gamma^\mu(t)$$
$$\Longrightarrow \quad U(1) \overset{(2)}{=} \exp\!\Big(-i\int_0^1 a_\mu(\gamma(t))\dot\gamma^\mu(t)\,\mathrm{d}t\Big) \overset{(3)}{=} \exp\!\Big(-i\int_{\gamma_{ab}} a\Big).$$

(1) For a 1×1 matrix, $\dot U/U = \frac{\mathrm{d}}{\mathrm{d}t}\ln U$; divide (8.24) by U.
(2) Integrate from 0 to 1 with $U(0) = 1$.
(3) The integral of a 1-form along a path is defined by pulling back to the parameter interval (Chapter T2).

This is the Aharonov–Bohm phase: with $A = -iea$ in terms of the physical potential a and charge e, $U(1) = \exp(ie\int_\gamma a)$.

In the *non-Abelian* case the matrices $A(\dot\gamma(t))$ at different times do not commute and the exponential of the integral is wrong. The fix is to iterate. Integrating (8.24),

$$\begin{aligned} U(t) &\overset{(1)}{=} \mathbb{1} - \int_0^t \mathrm{d}t_1\,\mathsf{A}(t_1)U(t_1) \\ &\overset{(2)}{=} \mathbb{1} - \int_0^t \mathrm{d}t_1\,\mathsf{A}(t_1) + \int_0^t \mathrm{d}t_1 \int_0^{t_1} \mathrm{d}t_2\,\mathsf{A}(t_1)\mathsf{A}(t_2) - \cdots \\ &\overset{(3)}{=} \mathcal{P}\exp\!\Big(-\int_0^t \mathsf{A}(s)\,\mathrm{d}s\Big), \end{aligned} \tag{8.26}$$

(1) Integrate $\dot U = -\mathsf{A}U$ with $U(0) = \mathbb{1}$, where $\mathsf{A}(t) := A_\mu(\gamma(t))\dot\gamma^\mu(t)$ is a matrix-valued function of t.
(2) Substitute the equation into itself repeatedly. The later times always stand to the *left*: $t \geq t_1 \geq t_2 \geq \cdots$.
(3) This is the definition of the path-ordered exponential $\mathcal{P}$ exp: expand the exponential and order each product with later times to the left. It is the same object as the Dyson series of time-dependent perturbation theory.

The series converges for every t (the n-th term is bounded by $(\int\|\mathsf{A}\|)^n/n!$). Its practical meaning is "chop the path into tiny pieces and multiply the tiny exponentials in order":

$$U(1) = \lim_{N\to\infty} e^{-\mathsf{A}(t_N)\Delta t}\cdots e^{-\mathsf{A}(t_2)\Delta t}\,e^{-\mathsf{A}(t_1)\Delta t}, \qquad \Delta t = 1/N. \tag{8.27}$$

Working definition 8.5 (parallel transport map)

The **parallel transport** of the connection A along γ_{ab} is the map from the fibre over a [the copy P_a of G sitting above the point a] to the fibre over b obtained by following horizontal lifts [curves $g(t)$ in P above γ_{ab} that solve $\dot g = -A(\dot\gamma_{ab})\,g$, i.e. (8.23)] (Nash's (8.7)):

$$PT_{\gamma_{ab}}(A) : P_a \longrightarrow P_b, \qquad g(0) \longmapsto g(1) = U_{\gamma_{ab}}(A)\,g(0), \tag{8.28}$$

with $U_{\gamma_{ab}}(A) = \mathcal{P}\exp(-\int_{\gamma_{ab}} A)$. *To compute it:* solve (8.23), or multiply the short-step exponentials (8.27). It commutes with the right action of G [$p \mapsto pu$, multiplying the fibre coordinate by a fixed $u \in G$ from the right]: $PT(pu) = PT(p)u$, because U multiplies from the left.

Two basic properties follow at once from the ODE. Running a path backwards inverts the map, $PT_{\gamma^{-1}} = PT_\gamma^{-1}$. Following γ_1 from a to b and then γ_2 from b to c composes the maps,

$$PT_{\gamma_2 * \gamma_1} = PT_{\gamma_2} \circ PT_{\gamma_1}, \qquad U_{\gamma_2 * \gamma_1} = U_{\gamma_2} U_{\gamma_1}. \tag{8.29}$$

Both are clear from (8.27): the product for the combined path is the product for γ_2 times the product for γ_1, and reversing a path reverses the order of the factors and flips the sign of each exponent.

Worked example 8.4: Holonomy of a small square: curvature appears

Let $M = \mathbb{R}^2$ and take a *constant* non-Abelian potential $A = X\,\mathrm{d}x + Y\,\mathrm{d}y$, where $X, Y \in \mathfrak{g}$ are fixed anti-Hermitian matrices that do not commute. Its curvature is

$$F \overset{(1)}{=} \mathrm{d}A + A \wedge A \overset{(2)}{=} XY\,\mathrm{d}x \wedge \mathrm{d}y + YX\,\mathrm{d}y \wedge \mathrm{d}x \overset{(3)}{=} [X, Y]\,\mathrm{d}x \wedge \mathrm{d}y.$$

(1) Definition of the curvature.
(2) $\mathrm{d}A = 0$ since X, Y are constant; in $A \wedge A$ only the mixed terms survive because $\mathrm{d}x \wedge \mathrm{d}x = 0$.
(3) $\mathrm{d}y \wedge \mathrm{d}x = -\mathrm{d}x \wedge \mathrm{d}y$.

Transport around the square of side ε with corners $(0,0)$, $(\varepsilon,0)$, $(\varepsilon,\varepsilon)$, $(0,\varepsilon)$, traversed anticlockwise. On each leg A is constant, so each leg contributes an honest exponential: $e^{-\varepsilon X}$ for the first leg ($\dot\gamma = \varepsilon\partial_x$), $e^{-\varepsilon Y}$ for the second, $e^{+\varepsilon X}$ and $e^{+\varepsilon Y}$ for the two return legs. By (8.29),

$$\begin{aligned} U_\square &\overset{(1)}{=} e^{\varepsilon Y} e^{\varepsilon X} e^{-\varepsilon Y} e^{-\varepsilon X} \\ &\overset{(2)}{=} \mathbb{1} + \varepsilon^2 [Y, X] + O(\varepsilon^3) \\ &\overset{(3)}{=} \mathbb{1} - \varepsilon^2 F_{xy} + O(\varepsilon^3). \end{aligned}$$

(1) Later legs stand to the left.
(2) Expand each exponential to second order and multiply out: all terms of first order cancel, the squares cancel, and the cross terms leave $\varepsilon^2(YX - XY)$. (This is the group commutator formula $e^a e^b e^{-a} e^{-b} = \mathbb{1} + [a, b] + \cdots$.)
(3) $F_{xy} = [X, Y]$ from the previous calculation.

The transport around a small closed loop differs from the identity by the curvature times the enclosed area. This is the most concrete meaning of F. In the Abelian case X and Y commute, the four exponentials cancel and $U_\square = \mathbb{1}$ exactly; consistently, a constant Abelian potential has $F = 0$. A small loop detects the curvature, not the potential.

8.1.5 Gauge transformations commute with parallel transport

Now we let a gauge transformation α, with local functions g_i, act on everything at once: on the bundle by (8.18) and on the connection by (8.16). Write $\alpha \cdot A$ for A^g, as Nash does. Nash's (8.8) says that transporting and then transforming is the same as transforming and then transporting with the transformed connection. Denote by $\alpha_a : P_a \to P_a$ the restriction of α to a fibre.

Key idea: Gauge covariance of parallel transport, Nash (8.8)

$$\alpha_b \circ PT_{\gamma_{ab}}(A) = PT_{\gamma_{ab}}(\alpha \cdot A) \circ \alpha_a, \qquad U_{\gamma_{ab}}(A^g) = g(b)^{-1}\, U_{\gamma_{ab}}(A)\, g(a). \tag{8.30}$$

Here is the whole argument, which is one calculation. Let $u(t)$ be a horizontal lift for A, so $\dot{u} = -\mathsf{A}\,u$. Transform it pointwise, $h(t) \coloneqq g(\gamma(t))^{-1}u(t)$, and write $g(t)$ for $g(\gamma(t))$. Then

$$\begin{aligned}\dot{h} &\overset{(1)}{=} -g^{-1}\dot{g}\,g^{-1}u + g^{-1}\dot{u}\\ &\overset{(2)}{=} -g^{-1}\dot{g}\,h - g^{-1}\mathsf{A}\,u\\ &\overset{(3)}{=} -\big(g^{-1}\mathsf{A}\,g + g^{-1}\dot{g}\big)h\\ &\overset{(4)}{=} -\big(A^g\big)(\dot{\gamma})\,h.\end{aligned}$$

(1) Product rule, with $\frac{\mathrm{d}}{\mathrm{d}t}g^{-1} = -g^{-1}\dot{g}\,g^{-1}$.
(2) Recognise $g^{-1}u = h$ in the first term; use the transport equation for u in the second.
(3) Write $u = gh$ in the second term.
(4) $g^{-1}\mathsf{A}g + g^{-1}\dot{g}$ is the transformed potential (8.16) evaluated on $\dot{\gamma}$ (since $\dot{g} = \mathrm{d}g(\dot{\gamma})$).

So $h(t)$ is a horizontal lift for the transformed connection A^g. It starts at $h(0) = g(a)^{-1}u(0) = \alpha_a(u(0))$ and ends at $h(1) = g(b)^{-1}u(1) = \alpha_b(u(1))$. Reading this as maps of fibres gives (8.30). (This is also why the inverse appears in (8.18): it makes the action on P match the familiar rule $A \mapsto g^{-1}Ag + g^{-1}\mathrm{d}g$.)

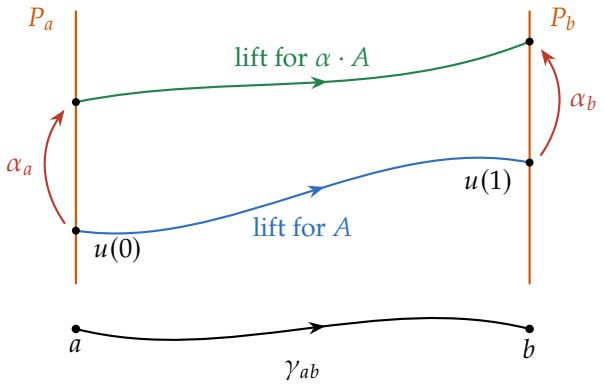


Figure 8.5: Nash's (8.8) as a picture. The blue curve is the horizontal lift for A from $u(0)$ to $u(1)$. The gauge transformation moves every point of it vertically (red arrows at the two ends). The moved curve (green) is again horizontal, for the transformed connection $\alpha \cdot A$. Going round the square either way gives the same point of P_b.

Worked example 8.5: Checking (8.8) for U(1)

For $G = \mathrm{U}(1)$, $g = e^{i\lambda(x)}$ and $A = ia$, the transformed connection is $A^g = ia + i\,\mathrm{d}\lambda$ and

$$\begin{aligned}U_\gamma(A^g) &\overset{(1)}{=} \exp\Big(-i\int_\gamma (a + \mathrm{d}\lambda)\Big) \overset{(2)}{=} e^{-i(\lambda(b)-\lambda(a))}\exp\Big(-i\int_\gamma a\Big)\\ &\overset{(3)}{=} g(b)^{-1}\,U_\gamma(A)\,g(a).\end{aligned}$$

(1) The Abelian solution of the transport equation, applied to A^g.
(2) Fundamental theorem of calculus: $\int_\gamma \mathrm{d}\lambda = \lambda(b) - \lambda(a)$.
(3) $e^{-i\lambda(b)} = g(b)^{-1}$ and $e^{i\lambda(a)} = g(a)$.

This is the statement, familiar from quantum mechanics, that the Aharonov–Bohm phase along an open path is not gauge invariant: it changes by the gauge phases at its two ends. Around a *closed* path the two ends coincide and the phase is invariant.

8.1.6 The based gauge group acts freely

Now the promised result.

Key idea: Nash: $\mathcal{G}$ acts freely on $\mathcal{A}$

If a based gauge transformation $\alpha \in \mathcal{G}$ fixes some connection, $\alpha \cdot A = A$, then α is the identity. Hence $\mathcal{G}$ acts on $\mathcal{A}$ without fixed points.

Why. Take $a = m_0$, the base point. Nash's (8.9) combines the two facts

$$\alpha \cdot A = A \qquad \text{and} \qquad \alpha_{m_0} = I.$$

Put them into (8.30), for any path γ from m_0 to a point b:

$$\alpha_b \circ PT_\gamma(A) \overset{(1)}{=} PT_\gamma(\alpha \cdot A) \circ \alpha_{m_0} \overset{(2)}{=} PT_\gamma(A) \circ I = PT_\gamma(A).$$

(1) Gauge covariance of parallel transport (8.30).
(2) α fixes A, and α is the identity on the base fibre.

Parallel transport is a bijection $P_{m_0} \longrightarrow P_b$ (its inverse is transport backwards), so we can cancel it on the right: $\alpha_b = I$. If M is path connected, every b can be reached from m_0 by a path, so α is the identity everywhere.

The same thing can be seen directly from the equation $A^g = A$. Multiply $g^{-1}Ag + g^{-1}\mathrm{d}g = A$ on the left by g:

$$\mathrm{d}g = gA - Ag, \qquad \text{i.e.} \qquad \mathrm{d}g + [A, g] = 0. \tag{8.31}$$

Along any path this is a linear first-order ODE for $g(\gamma(t))$: $\dot{g} = g\mathsf{A} - \mathsf{A}g$. At $t = 0$ we are at m_0, so $g = \mathbb{1}$. The constant $g \equiv \mathbb{1}$ solves the ODE (since $\mathbb{1}\mathsf{A} - \mathsf{A}\mathbb{1} = 0$) with the same initial value. By uniqueness of solutions, $g \equiv \mathbb{1}$ along the whole path.

Physical picture: Covariantly constant gauge transformations

Equation (8.31) says that a gauge transformation fixing A is *covariantly constant* ($\mathrm{d}_A g = 0$, with g viewed as a section of $\mathrm{Ad}\,P$). A covariantly constant object is determined by its value at one point, just as a solution of $\mathrm{d}f = 0$ is. In $\hat{\mathcal{G}}$ that value can be any element of G that commutes with the holonomy (see below), so there may be many such g. In $\mathcal{G}$ it is pinned to $\mathbb{1}$, so there is only one.

Watch out: Free action needs the base point, not the centre

One might try to cure the non-free action of $\hat{\mathcal{G}}$ by dividing out only its centre (the constant central elements). That removes the stabiliser of a *generic* (irreducible) connection, but not the larger stabilisers of reducible ones such as $A = 0$ in Example 8.3. The based group removes all of them at once, at the price of the leftover global symmetry $G = \hat{\mathcal{G}}/\mathcal{G}$, which still acts on $\mathcal{A}/\mathcal{G}$. Section 3 of this chapter deals with that residual action on instanton moduli spaces.

8.1.7 The holonomy group

Closely related is the **holonomy** of a connection. Let the path close up into a loop γ_a that starts and ends at a. Parallel transport is now a map of one fibre to itself, Nash's (8.10):

$$PT_{\gamma_a}(A) : P_a \longrightarrow P_a, \qquad u \longmapsto U_{\gamma_a}(A)\, u. \tag{8.32}$$

It is generally not the identity: transport around a loop can rotate the internal frame. This is the Aharonov–Bohm effect in the Abelian case and the Wilson loop in the non-Abelian one.

Working definition 8.6 (holonomy group; reducible and irreducible connections)

The **holonomy group** $\mathrm{Hol}_a(A)$ is the set of all matrices $U_{\gamma_a}(A)$ as γ_a runs over all loops based at a. *To compute it:* find the transport matrices of enough loops and see what subgroup of G they generate [all finite products of these matrices and their inverses]. Nash calls A **irreducible** if $\mathrm{Hol}_a(A) = G$ and **reducible** if the holonomy group is a proper subgroup [a subgroup strictly smaller than G, such as a U(1) inside SU(2)].

It is a group because loops can be composed (giving the product, by (8.29)), reversed (giving the inverse) and taken constant (giving $\mathbb{1}$). Changing the base point from a to b conjugates it: if σ is a path from a to b, every loop at b is $\sigma * \gamma_a * \sigma^{-1}$ for a loop γ_a at a, and

$$U_{\sigma*\gamma_a*\sigma^{-1}} = U_\sigma \, U_{\gamma_a} \, U_\sigma^{-1}, \qquad \mathrm{Hol}_b(A) = U_\sigma \, \mathrm{Hol}_a(A) \, U_\sigma^{-1}. \tag{8.33}$$

So on a connected M the holonomy groups at different points are isomorphic, as Nash says. By construction the holonomy group is a subgroup of the structure group G. Under a gauge transformation it is conjugated by $g(a)$, by (8.30): $U_{\gamma_a}(A^g) = g(a)^{-1} U_{\gamma_a}(A) g(a)$. The trace of a holonomy matrix, the **Wilson loop** $W_\gamma(A) = \operatorname{tr} U_\gamma(A)$, is therefore gauge invariant.

The holonomy tells us the stabiliser in the full gauge group $\hat{\mathcal{G}}$. A solution of (8.31) is determined by its value $c = g(a)$. Going around a loop and coming back must reproduce c; by (8.30) with $A^g = A$, this means $U_{\gamma_a} = c^{-1} U_{\gamma_a} c$. So

$$\mathrm{Stab}_{\hat{\mathcal{G}}}(A) \cong \{ c \in G : \; cU = Uc \text{ for all } U \in \mathrm{Hol}_a(A) \}, \tag{8.34}$$

the **centraliser** of the holonomy group. The bigger the holonomy, the smaller the stabiliser. If $\mathrm{Hol}(A) = G$ the stabiliser is the centre $Z(G)$, the minimum possible. (Section 3 of this chapter returns to this.)

Worked example 8.6: Flat U(1) connections on the circle

On S^1 take the constant potential $A = i\,a\,\mathrm{d}\phi$, $a \in \mathbb{R}$. It is flat ($F = \mathrm{d}A = 0$: there are no 2-forms on a circle). The loop going once around has

$$U = \exp\!\left(-i \int_0^{2\pi} a \, \mathrm{d}\phi \right) = e^{-2\pi i a}.$$

Going around n times gives $e^{-2\pi i n a}$. A loop that goes out and comes back without going around gives $\mathbb{1}$. So $\mathrm{Hol}(A) = \{ e^{-2\pi i n a} : n \in \mathbb{Z} \}$.

- $a \in \mathbb{Z}$: Hol $= \{1\}$. The connection is gauge equivalent to $A = 0$ (by $g = e^{-i a \phi}$, a legitimate map $S^1 \to \mathrm{U}(1)$ when a is an integer).
- $a = p/q$ in lowest terms: Hol $= \mathbb{Z}_q$, the q-th roots of unity.
- a irrational: Hol is infinite cyclic and *dense* in U(1), but not all of it.

All of these are reducible in Nash's sense except that none reaches all of U(1). (For U(1) the notion is not very useful; it matters for non-Abelian G.) Note also that $\mathrm{Hol}(A)$ need not be a closed subgroup.

Worked example 8.7: The Dirac monopole: holonomy equals flux

Take the monopole connection of Example 8.2 and the latitude circle at polar angle θ_0, based at a point on it and traversed in the direction of increasing ϕ. Using the northern potential,

$$U \overset{(1)}{=} \exp\Big(-\oint A_N\Big) \overset{(2)}{=} \exp\Big(\frac{im}{2}(1-\cos\theta_0)\cdot 2\pi\Big) \overset{(3)}{=} \exp\Big(\frac{im}{2}\,\Omega(\theta_0)\Big),$$

(1) Abelian transport around a closed loop.
(2) $A_N = -\frac{im}{2}(1-\cos\theta)\mathrm{d}\phi$ is constant on the circle, and $\oint \mathrm{d}\phi = 2\pi$.
(3) $\Omega(\theta_0) = 2\pi(1-\cos\theta_0)$ is the solid angle of the polar cap enclosed by the circle.

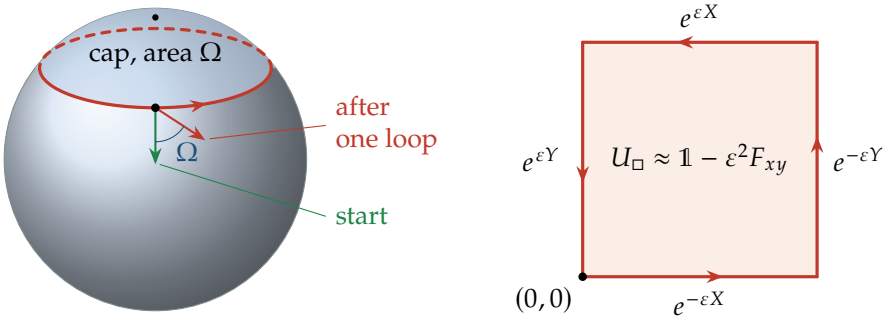
By Stokes, $-\oint A_N = -\int_{\text{cap}} F$: the holonomy is the magnetic flux through the cap, in the form $e^{ie\Phi}$. As θ_0 runs from 0 to π the phase runs over all of U(1) (for $m \neq 0$), so Hol = U(1). As $\theta_0 \to \pi$ the loop shrinks to the south pole and $U \to e^{2\pi i m} = 1$: the Dirac quantisation condition, seen as the statement that a vanishing loop must have trivial holonomy.

Worked example 8.8: Levi-Civita connection on S^2

Parallel transport of tangent vectors on the unit sphere is an SO(2) connection on the frame bundle. Carry a tangent vector around the latitude circle $\theta = \theta_0$. It comes back rotated by the angle $2\pi(1-\cos\theta_0)$, the enclosed area (the curvature is 1, so this is Gauss–Bonnet for the cap, Chapter T2). All angles occur, so Hol = SO(2). For the flat plane the same experiment gives no rotation: Hol = $\{1\}$. See Figure 8.6.

Worked example 8.9: A reducible SU(2) connection

Embed a U(1) connection ia in SU(2) diagonally, $A = \mathrm{diag}(ia, -ia)$. Every transport matrix is $U = \mathrm{diag}(e^{-i\int a}, e^{i\int a})$: the holonomy lies in the diagonal U(1) $\subset$ SU(2), a proper subgroup. So A is reducible. The centraliser of the diagonal U(1) in SU(2) is the diagonal U(1) itself, which matches the stabiliser found in Example 8.3. Conversely, an SU(2) connection whose holonomy contains two non-commuting matrices (say, $e^{\varepsilon e_1}$ and $e^{\varepsilon e_2}$) has centraliser $\{\pm\mathbb{1}\} = Z(\mathrm{SU}(2))$. In the square example Example 8.4 this happens as soon as $[X, Y] \neq 0$.



Levi-Civita on S^2 constant $A = X\mathrm{d}x + Y\mathrm{d}y$, side ε

Figure 8.6: Two ways to see holonomy. Left: a tangent vector carried around a latitude circle on the sphere returns rotated by the enclosed area Ω, so the holonomy group of the round sphere is all of SO(2). Right: the four legs of a small square contribute four matrix exponentials, whose ordered product differs from $\mathbb{1}$ by the curvature times the area (Example 8.4).

Watch out: Two definitions of "irreducible"

Nash defines irreducible as "holonomy group equal to G". Most of the modern literature (and Section 3 of this chapter) uses the weaker condition "stabiliser equal to the centre $Z(G)$", i.e. the centraliser of the holonomy is as small as possible. For SU(2) the two agree in practice: the only closed connected proper subgroups of SU(2) are $\{1\}$ and the circles U(1) (up to conjugation), whose centralisers are SU(2) and U(1), while any subgroup containing two non-commuting elements has centraliser $\{\pm\mathbb{1}\}$. For larger groups, a subgroup can be proper and still have trivial centraliser (for example $\mathrm{SO}(3) \subset \mathrm{SU}(3)$), so the two notions differ.

Reading Nash (pp. 216–218, (8.2)–(8.10))

Nash's (8.2)–(8.7) set up parallel transport in local coordinates (x, g): the ODE (8.4) $\dot{g} + A\dot{\gamma}_{ab}\, g = 0$ is our (8.23), and $PT_{\gamma_{ab}}(A)$ is (8.28). His (8.8) is our (8.30); he states it as "the commutativity of gauge transformations and parallel transport" and does not prove it. The one-line computation of $\dot{h}$ above is the proof. His (8.9) is the free-action argument. The "$\alpha \cdot A$" in Nash is our A^g. Nash's reference to "chapter 6 p. 176" is to the definition of $\mathcal{G}$ as the based subgroup of $\hat{\mathcal{G}} = \ker(\mathrm{Aut}\, P \to \mathrm{Diff}\, M)$.

Recap

- A gauge transformation is a section of $\mathrm{Ad}\, P$: local maps g_i glued by conjugation. $\hat{\mathcal{G}}$ is the group of all of them, $\mathcal{G}$ the subgroup with $g(m_0) = \mathbb{1}$; $\hat{\mathcal{G}}/\mathcal{G} = G$.
- Parallel transport solves $\dot{g} = -A(\dot{\gamma})g$; the solution is a path-ordered exponential. Around a small loop it is $\mathbb{1} - F \cdot (\text{area})$.
- Gauge transformations commute with parallel transport: $U(A^g) = g(b)^{-1} U(A) g(a)$.
- Hence the based group $\mathcal{G}$ acts freely on $\mathcal{A}$; the full group $\hat{\mathcal{G}}$ has stabilisers equal to the centraliser of the holonomy.
- Irreducible: holonomy $= G$ (Nash), or stabiliser $= Z(G)$ (modern).

8.1.8 The quotient $\mathcal{A}/\mathcal{G}$

Since $\mathcal{G}$ acts freely on $\mathcal{A}$, we can divide. The **gauge orbit** of a connection A is the set of everything gauge equivalent to it, $\mathcal{G} \cdot A = \{A^g : g \in \mathcal{G}\}$. The quotient $\mathcal{A}/\mathcal{G}$ is the set of orbits. A point of $\mathcal{A}/\mathcal{G}$ is a gauge field "up to gauge transformations": what a physicist would call a physical field configuration.

Because the action is free, every orbit is a faithful copy of $\mathcal{G}$ (the map $g \mapsto A^g$ is one-to-one). So $\mathcal{A}$ is fibred over $\mathcal{A}/\mathcal{G}$ with every fibre a copy of $\mathcal{G}$. This is Nash's (8.11):

$$\begin{array}{ccc} \mathcal{G} & \longrightarrow & \mathcal{A} \\ & & \big\downarrow \pi \\ & & \mathcal{A}/\mathcal{G} \end{array} \tag{8.35}$$

It is a principal $\mathcal{G}$-bundle: $\mathcal{G}$ acts on the total space $\mathcal{A}$, freely, with the fibres as orbits. (Local triviality needs a local "slice" through each orbit; the Coulomb gauge condition provides one, as Section 3 of this chapter explains. We take it for granted here.)

Worked example 8.10: U(1) on the circle: $\mathcal{A}/\mathcal{G}$ is a circle

Return to Example 8.1: $M = S^1$, $G = \mathrm{U}(1)$, $A = i\,a(\phi)\,\mathrm{d}\phi$ with a a periodic real function. A based gauge transformation is $g = e^{i\lambda(\phi)}$ with $g(0) = 1$. Because g must be single valued, λ need not be periodic: it may wind,

$$\lambda(0) = 0, \qquad \lambda(2\pi) = 2\pi n, \quad n \in \mathbb{Z}.$$

The action is $a \mapsto a + \lambda'$, since $A^g = A + g^{-1}\mathrm{d}g = i(a + \lambda')\mathrm{d}\phi$. When are a and $\tilde{a}$ in the same orbit? We need $\tilde{a} - a = \lambda'$ for an allowed λ. Integrating,

$$\oint (\tilde{a} - a)\,\mathrm{d}\phi \overset{(1)}{=} \int_0^{2\pi} \lambda'(\phi)\,\mathrm{d}\phi \overset{(2)}{=} \lambda(2\pi) - \lambda(0) \overset{(3)}{=} 2\pi n.$$

(1) Assume $\tilde{a} = a + \lambda'$.
(2) Fundamental theorem of calculus.
(3) Boundary conditions on λ.

Conversely, if $\oint(\tilde{a} - a) = 2\pi n$, then $\lambda(\phi) := \int_0^{\phi}(\tilde{a} - a)$ is allowed and does the job. So the only gauge invariant is the total flux modulo 2π, i.e. the holonomy $e^{-i\oint a}$, and

$$\mathcal{A}/\mathcal{G} \;\cong\; \mathbb{R}/2\pi\mathbb{Z} \;\cong\; S^1, \qquad [a] \longmapsto \oint a\,\mathrm{d}\phi \bmod 2\pi. \tag{8.36}$$

The total space $\mathcal{A}$ is contractible, the quotient is a circle, and the fibre $\mathcal{G}$ has infinitely many components labelled by the winding n. Walking once around the circle $\mathcal{A}/\mathcal{G}$ (increasing $\oint a$ from 0 to 2π) and lifting the walk to $\mathcal{A}$, we arrive at a connection that differs from the starting one by a gauge transformation of winding $n = 1$. This is the first appearance of a *large gauge transformation* (Section 8.1.17).

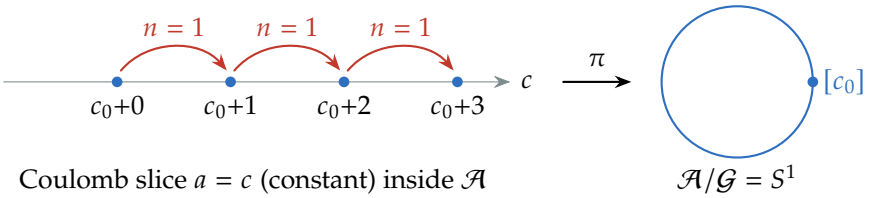


Figure 8.7: The U(1) theory on a circle. The constant potentials $a = c$ form a line inside $\mathcal{A}$ (coordinate $c = \frac{1}{2\pi}\oint a\,\mathrm{d}\phi$) that meets every gauge orbit, but it meets each orbit infinitely often, at $c_0 + n$ (blue dots), related by the winding gauge transformations $e^{in\phi}$ (red arrows). Under the projection π the line wraps infinitely often around the quotient circle. Choosing one representative per orbit continuously is impossible: this is the Gribov problem in miniature.

8.1.9 Universal bundles and classifying spaces

Nash's next sentence, (8.12), is loaded: "since $\mathcal{A}$ is contractible it follows that $\mathcal{A}$ is the total space of a universal $\mathcal{G}$-bundle; hence $\mathcal{A}/\mathcal{G}$ is the classifying space for $\mathcal{G}$ bundles." Let us unpack it.

Working definition 8.7 (universal bundle, classifying space)

Let H be a topological group [here $H = \mathcal{G}$, the based gauge group]. A principal H-bundle $EH \to BH$ [H acts freely on EH and $BH = EH/H$, as $\mathcal{G}$ does on $\mathcal{A}$] whose total space EH is *contractible* [it can be shrunk continuously to one point, as $\mathcal{A}$ can along straight lines to A_0] is called a **universal H-bundle**, and its base BH is a **classifying space** for H. *In practice:* find any contractible space on which H acts freely; the quotient is BH. It is unique up to homotopy equivalence [two answers are related by maps both ways whose composites can be deformed continuously into the identity].

Why the name? Two facts, both standard in algebraic topology.

1. *Classification.* For any (reasonable) space X, principal H-bundles over X correspond one-to-one to homotopy classes of maps $f : X \to BH$. The bundle belonging to f is the pull-back f^*EH. So BH "classifies" H-bundles.
2. *Homotopy.* The homotopy groups of BH are those of H, shifted up by one:

$$\pi_j(BH) \cong \pi_{j-1}(H). \tag{8.37}$$

The second fact is the one we use all the time, and it follows from the long exact sequence of a fibration, which we now recall.

The long exact homotopy sequence

For any fibre bundle $F \to E \to B$ (with B connected) there is a sequence of group homomorphisms

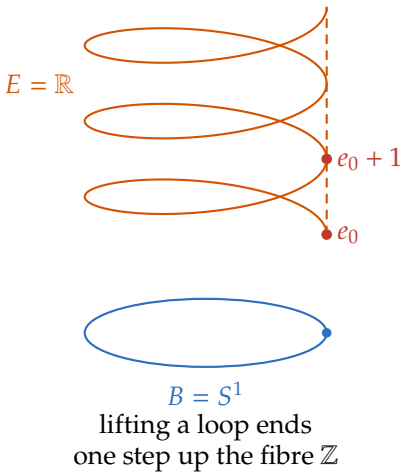
$$\cdots \to \pi_j(F) \xrightarrow{i_*} \pi_j(E) \xrightarrow{\pi_*} \pi_j(B) \xrightarrow{\partial} \pi_{j-1}(F) \xrightarrow{i_*} \pi_{j-1}(E) \to \cdots \to \pi_0(E) \to 0, \tag{8.38}$$

which is **exact**: at each place, the image of the incoming map equals the kernel of the outgoing one. The maps i_* and π_* are induced by the inclusion of a fibre and by the projection. The interesting one is the **boundary map** ∂. Here is how it works for $j = 1$. Take a loop ℓ in B at the base point. Lift it to a path $\hat{\ell}$ in E starting at a chosen point e_0 of the fibre (always possible for a bundle). The lift need not close up: it ends somewhere in the same fibre F. The class of that end point in $\pi_0(F)$ is $\partial[\ell]$. For higher j the same idea applies to a j-dimensional ball whose boundary sphere is squashed to the base point: lift the ball, and look at where its boundary lands, a $(j-1)$-sphere in F.

If E is contractible, all $\pi_j(E) = 0$, and exactness at the two places around ∂ gives $0 \to \pi_j(B) \xrightarrow{\partial} \pi_{j-1}(F) \to 0$: so ∂ is an isomorphism. This is (8.37).

Worked example 8.11: A gallery of classifying spaces

- $H = \mathbb{Z}$ acting on $\mathbb{R}$ by translations. $\mathbb{R}$ is contractible, so $B\mathbb{Z} = \mathbb{R}/\mathbb{Z} = S^1$. Check: $\pi_1(S^1) = \mathbb{Z} = \pi_0(\mathbb{Z})$.
- $H = \mathbb{Z}_2$ acting on the infinite sphere $S^\infty = \bigcup_n S^n$ by $v \mapsto -v$. S^∞ is contractible (every sphere S^n can be shrunk inside S^{n+1}), so $B\mathbb{Z}_2 = \mathbb{RP}^\infty$. Check: $\pi_1(\mathbb{RP}^\infty) = \mathbb{Z}_2$.
- $H = \mathrm{U}(1)$ acting on the unit sphere $S^\infty \subset \mathbb{C}^\infty$ by $v \mapsto e^{i\lambda}v$. So $B\mathrm{U}(1) = \mathbb{CP}^\infty$. Check: $\pi_2(\mathbb{CP}^\infty) = \pi_1(\mathrm{U}(1)) = \mathbb{Z}$, and all other homotopy groups vanish. The finite stage $S^3 \to \mathbb{CP}^1 = S^2$ is the Hopf fibration.



E contractible:

$$\underbrace{\pi_j(E)}_{0} \to \pi_j(B) \xrightarrow{\partial} \pi_{j-1}(F) \to \underbrace{\pi_{j-1}(E)}_{0}$$

$$\Longrightarrow \quad \pi_j(B) \cong \pi_{j-1}(F)$$

here: $\pi_1(S^1) = \mathbb{Z} = \pi_0(\mathbb{Z})$

Figure 8.8: The boundary map of the long exact sequence, in the simplest universal bundle $\mathbb{Z} \to \mathbb{R} \to S^1$ ($x \mapsto e^{2\pi i x}$). A loop going once around the base lifts to a path in $E = \mathbb{R}$ that climbs from e_0 to $e_0 + 1$. The mismatch lives in the fibre. When E is contractible, this mismatch is all the information there is: $\pi_j(B) = \pi_{j-1}(F)$.

- $H = \mathrm{U}(N)$: $B\mathrm{U}(N)$ is the Grassmannian of N-planes in $\mathbb{C}^\infty$. Its cohomology is generated by the universal Chern classes c_i; we use this in Section 8.1.25.

The classification statement in the U(1) case: U(1) bundles over S^2 correspond to $[S^2, \mathbb{CP}^\infty] = \pi_2(\mathbb{CP}^\infty) = \mathbb{Z}$. The integer is the monopole charge m of Example 8.2.

Key idea: Nash (8.12): the configuration space is a classifying space

$\mathcal{A}$ is contractible and $\mathcal{G}$ acts freely on it. Therefore

$$\mathcal{A}/\mathcal{G} = B\mathcal{G}, \qquad \pi_j(\mathcal{A}/\mathcal{G}) \cong \pi_{j-1}(\mathcal{G}) \quad (j \geq 1). \tag{8.39}$$

The whole topology of the configuration space of a gauge theory is the topology of the gauge group, shifted by one degree.

Check it on Example 8.10: $\mathcal{A}/\mathcal{G} = S^1$, so $\pi_1(\mathcal{A}/\mathcal{G}) = \mathbb{Z}$; and $\pi_0(\mathcal{G}) = \mathbb{Z}$, the winding number of λ. The loop in $\mathcal{A}/\mathcal{G}$ that increases $\oint a$ by 2π corresponds, via ∂, to the gauge transformation of winding one, exactly as the figure showed.

8.1.10 Why physicists care: the path integral lives on $\mathcal{A}/\mathcal{G}$

The partition function and the correlation functions of a gauge theory are functional integrals over $\mathcal{A}$,

$$Z = \int_{\mathcal{A}} \mathcal{D}A \; e^{-S[A]}, \qquad \langle O \rangle = \frac{1}{Z} \int_{\mathcal{A}} \mathcal{D}A \; O[A] \, e^{-S[A]}. \tag{8.40}$$

If S and O are gauge invariant, the integrand is constant along each orbit. The integral over the orbit then just produces the (infinite) volume of $\mathcal{G}$, and all the physics is in the integral over the space of orbits $\mathcal{A}/\mathcal{G}$. (Nash's caveat, "in the absence of anomalies", means: as long as the *measure* $\mathcal{D}A$ combined with the fermion determinant is also gauge invariant. When it is not, the integrand is not constant along orbits and the projection fails; that is Chapter 10.)

To perform the projection in practice one **fixes a gauge**: pick a condition $\chi(A) = 0$ that selects one point on each orbit, and integrate only over those points, with the right Jacobian. The

Jacobian is the **Faddeev–Popov determinant**. The procedure is clearest in a finite-dimensional toy model, which also shows its weak spot.

Worked example 8.12: Faddeev–Popov for rotations of the plane

Take the "field space" $\mathbb{R}^2$ with the "gauge group" SO(2) of rotations R_θ, and a rotation-invariant integrand $f(|x|)$. We know the answer: $Z = \int \mathrm{d}^2x\, f = 2\pi \int_0^\infty r\, f(r)\, \mathrm{d}r$. Try the gauge condition $\chi(x) = x_2 = 0$. Along the orbit of $x = (r\cos\phi, r\sin\phi)$,

$$\chi(R_\theta x) = r\sin(\phi+\theta), \qquad \left|\frac{\partial\chi(R_\theta x)}{\partial\theta}\right| = r\left|\cos(\phi+\theta)\right|.$$

The condition has *two* solutions on each orbit, $\theta = -\phi$ and $\theta = \pi - \phi$ (the positive and negative x_1 axis). With the Faddeev–Popov factor $\Delta(x) = r$ (the value of $|\partial_\theta\chi|$ at the solutions),

$$\int_0^{2\pi} \mathrm{d}\theta\; \Delta(x)\,\delta\big(\chi(R_\theta x)\big) \overset{(1)}{=} \sum_{\text{zeros } \theta_*} \frac{\Delta(x)}{|\partial_\theta\chi(R_\theta x)|_{\theta_*}} \overset{(2)}{=} 1 + 1 = 2.$$

(1) $\delta(h(\theta)) = \sum_{\theta_*} \delta(\theta-\theta_*)/|h'(\theta_*)|$ over the zeros of h.
(2) At both zeros $|\cos| = 1$, so each contributes $r/r = 1$.

So "$1 = \int \mathrm{d}\theta\,\Delta\,\delta(\chi)$", the identity that Faddeev–Popov insert into the path integral, is *false* here: the right-hand side is 2. Inserting $1 = \frac{1}{2}\int \mathrm{d}\theta\,\Delta\,\delta(\chi)$ instead,

$$\begin{aligned} Z &\overset{(1)}{=} \tfrac{1}{2}\int_0^{2\pi}\mathrm{d}\theta \int \mathrm{d}^2x\; r\,\delta\big(\chi(R_\theta x)\big) f(|x|) \\ &\overset{(2)}{=} \tfrac{1}{2}\cdot 2\pi \int \mathrm{d}^2x\; r\,\delta(x_2)\, f(|x|) \overset{(3)}{=} \pi\int_{-\infty}^{\infty}\mathrm{d}x_1\; |x_1|\, f(|x_1|) \overset{(4)}{=} 2\pi\int_0^\infty r\, f(r)\,\mathrm{d}r. \end{aligned}$$

(1) Insert the corrected identity and exchange the integrals.
(2) Change variables $x \to R_{-\theta}x$; d^2x, r and f are invariant, so the θ integral gives the group volume 2π.
(3) The δ sets $x_2 = 0$, so $r = |x_1|$.
(4) The integrand is even in x_1.

The right answer, but only after dividing by the number of **Gribov copies** (here 2). In this toy we could have avoided the problem by the better gauge "$x_2 = 0$, $x_1 > 0$", which meets each orbit exactly once. That works because the bundle $\mathbb{R}^2\setminus\{0\} \to (0,\infty)$ of orbits is trivial. In Yang–Mills theory there is no such escape, as we now show.

For Yang–Mills the steps are the same. Near the identity a gauge transformation is $g = e^\varepsilon \approx \mathbb{1} + \varepsilon$ with ε a $\mathfrak{g}$-valued function, and

$$\begin{aligned} A^g &\overset{(1)}{=} (\mathbb{1}-\varepsilon)A(\mathbb{1}+\varepsilon) + (\mathbb{1}-\varepsilon)\,\mathrm{d}\varepsilon + O(\varepsilon^2) \\ &\overset{(2)}{=} A + \mathrm{d}\varepsilon + [A,\varepsilon] + O(\varepsilon^2) \overset{(3)}{=} A + \mathrm{d}_A\varepsilon + O(\varepsilon^2). \end{aligned} \tag{8.41}$$

(1) Definition (8.16) with $g^{-1} \approx \mathbb{1} - \varepsilon$.
(2) Multiply out and drop second-order terms.
(3) $\mathrm{d}_A\varepsilon := \mathrm{d}\varepsilon + [A,\varepsilon]$ is the covariant derivative in the adjoint representation.

For the Lorenz (Landau) gauge $\chi(A) = \partial^\mu A_\mu$ the change of the gauge condition is $\delta\chi = \partial^\mu D_\mu\varepsilon$, so the Faddeev–Popov factor is $\Delta_{\mathrm{FP}}[A] = \left|\det(\partial^\mu D_\mu)\right|$. Writing the determinant with

anticommuting ghost fields $c, \bar{c}$,

$$Z = \int_{\mathcal{A}} \mathcal{D}A\, \mathcal{D}c\, \mathcal{D}\bar{c}\; \delta\big(\partial^\mu A_\mu\big)\, e^{-S[A] - \int \mathrm{tr}(\bar{c}\, \partial^\mu D_\mu c)} \quad (\text{“per Gribov copy”}). \tag{8.42}$$

This is “the familiar Faddeev–Popov ghost Jacobian” of Nash’s text. The derivation assumes that $\partial^\mu A_\mu = 0$ meets each orbit exactly once. Gribov noticed that for large fields it does not: the Coulomb or Lorenz condition has several solutions on one orbit, and the operator $\partial^\mu D_\mu$ acquires zero modes where two copies merge (the **Gribov horizon**). Perturbation theory never notices, since it only explores a neighbourhood of $A = 0$. The question is whether a cleverer gauge condition could do better. Topology says no.

8.1.11 The bundle $\mathcal{A} \to \mathcal{A}/\mathcal{G}$ is not trivial: the Gribov ambiguity

A *global gauge condition* is a rule that picks one connection on each orbit, continuously. In bundle language it is a continuous **section** $s : \mathcal{A}/\mathcal{G} \to \mathcal{A}$ of the bundle (8.35).

Key idea: Sections of principal bundles

A principal bundle has a global section if and only if it is trivial. Given a section s, the map

$$\mathcal{G} \times (\mathcal{A}/\mathcal{G}) \longrightarrow \mathcal{A}, \qquad (g, [A]) \longmapsto s([A])^g \tag{8.43}$$

is a bijection, and it respects the fibration: every point of $\mathcal{A}$ is reached exactly once, since each orbit contains exactly one $s([A])$ and the action is free.

So a global gauge condition exists only if $\mathcal{A} \cong \mathcal{G} \times (\mathcal{A}/\mathcal{G})$, Nash’s (8.13). Nash shows that this cannot be. If the product decomposition held, homotopy groups of a product being products,

$$\pi_j(\mathcal{A}) = \pi_j(\mathcal{G}) \oplus \pi_j(\mathcal{A}/\mathcal{G}), \tag{8.44}$$

which is Nash’s (8.14). Now compare the two sides.

- The left side is 0 for every j, because $\mathcal{A}$ is contractible (8.15).
- The right side contains $\pi_j(\mathcal{G})$, which is *not* zero for many j. For $M = S^4$ we will find $\pi_j(\mathcal{G}) = \pi_{j+4}(G)$ (8.56); for $G = \mathrm{SU}(2)$ this gives $\pi_0(\mathcal{G}) = \pi_4(S^3) = \mathbb{Z}_2$, $\pi_1(\mathcal{G}) = \pi_5(S^3) = \mathbb{Z}_2$, $\pi_2(\mathcal{G}) = \pi_6(S^3) = \mathbb{Z}_{12}$, and so on.

A group of the form $0 = \pi_j(\mathcal{G}) \oplus (\dots)$ with $\pi_j(\mathcal{G}) \neq 0$ is impossible. So the bundle is nontrivial and has no global section.

Key idea: The Gribov ambiguity (Gribov, Singer, Mitter–Viallet)

There is no continuous gauge-fixing condition that selects exactly one representative from every gauge orbit. Any gauge condition either misses some orbits or meets some orbits more than once (Gribov copies), or is discontinuous.

Worked example 8.13: A nontrivial bundle with no global gauge: the Hopf fibration

The Hopf bundle $\mathrm{U}(1) \to S^3 \to S^2$ is the finite-dimensional model of the Gribov problem. Write $S^3 = \{(z_1, z_2) \in \mathbb{C}^2 : |z_1|^2 + |z_2|^2 = 1\}$, with $\mathrm{U}(1)$ acting by $(z_1, z_2) \mapsto (e^{i\lambda} z_1, e^{i\lambda} z_2)$. The orbit space is $\mathbb{CP}^1 = S^2$. A natural gauge condition is “z_2 real and positive”, i.e.

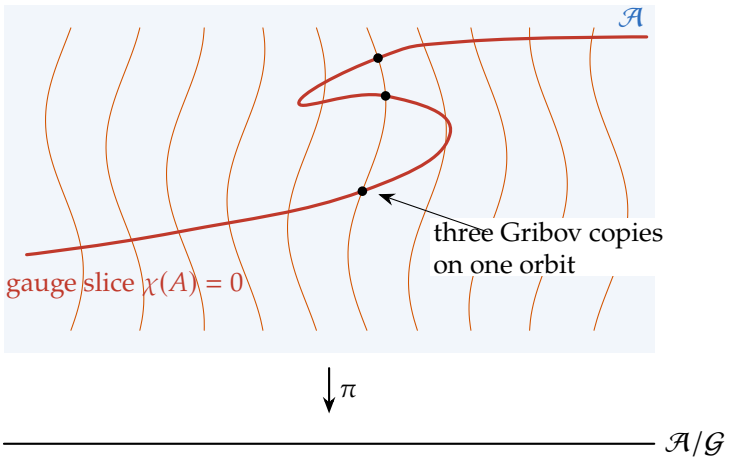


Figure 8.9: Gauge orbits (orange) fibre the space of connections. A gauge condition is a surface (red) that should cut every orbit once. Near $A = 0$ it does (left). For large fields the surface folds back and cuts some orbits several times; where two cuts merge, the Faddeev–Popov operator has a zero mode. The nontriviality of $\mathcal{A} \to \mathcal{A}/\mathcal{G}$ guarantees that *every* continuous choice of surface misbehaves somewhere.

$\arg z_2 = 0$. It picks exactly one point on each orbit with $z_2 \neq 0$, namely

$$s(z_1 : z_2) = \left(\frac{z_1 \bar{z}_2}{|z_2|},\ |z_2| \right).$$

But on the orbit $z_2 = 0$ (one point of S^2, the "north pole") the condition cannot be imposed at all, and near it s is discontinuous: approaching the pole from different directions gives different limits $z_1\bar{z}_2/|z_2|$. Any other choice fails somewhere else. The reason is the same as Nash's, with π_1 in place of π_j: if the bundle were trivial we would have $S^3 \cong \mathrm{U}(1) \times S^2$ and hence $0 = \pi_1(S^3) = \pi_1(\mathrm{U}(1)) \oplus \pi_1(S^2) = \mathbb{Z}$, which is absurd. So the Hopf bundle is nontrivial, and no global gauge exists. Physically this is the Dirac monopole of charge one: no single potential works on all of S^2.

Watch out: What the theorem does and does not say

Singer's theorem does not say that gauge fixing is useless. Locally, near any connection, the Coulomb gauge works perfectly, and perturbation theory is local in this sense. It says the local slices cannot be patched into one global slice, so any exact non-perturbative treatment must either use several patches on $\mathcal{A}/\mathcal{G}$ (as we use several patches on a monopole bundle) or live with Gribov copies. The precise statement needs a compact spacetime (or fields with fixed behaviour at infinity), such as S^4 or S^3; on $\mathbb{R}^n$ with no condition at infinity, $\mathcal{G}$ can be contractible.

Reading Nash (pp. 218–219, (8.11)–(8.14))

Nash gives the argument in four lines: (8.11) the bundle, (8.12) $\mathcal{A}/\mathcal{G} = B\mathcal{G}$ because $\mathcal{A}$ is contractible, (8.13)–(8.14) the product decomposition and its homotopy groups. The non-vanishing of $\pi_j(\mathcal{G})$ is postponed to (8.23), which we reach in Section 8.1.15. His references are Gribov (the Coulomb gauge copies), Singer (the topological theorem) and Mitter–Viallet (the geometry of $\mathcal{A} \to \mathcal{A}/\mathcal{G}$ as a Riemannian submersion).

Recap

- $\mathcal{A} \to \mathcal{A}/\mathcal{G}$ is a principal $\mathcal{G}$-bundle with contractible total space, so $\mathcal{A}/\mathcal{G} = B\mathcal{G}$ and $\pi_j(\mathcal{A}/\mathcal{G}) = \pi_{j-1}(\mathcal{G})$.
- $\mathcal{A}/\mathcal{G}$ is the physical configuration space; the path integral projects onto it (no anomalies), via gauge fixing and the Faddeev–Popov determinant.
- A global gauge condition is a section; it exists only if the bundle is trivial; triviality would force $\pi_j(\mathcal{G}) = 0$; that is false. Hence Gribov copies are unavoidable.
- Toy models: U(1) on S^1 ($\mathcal{A}/\mathcal{G} = S^1$, Coulomb slice meets each orbit infinitely often), rotations of the plane (two copies), the Hopf bundle (no global section).

8.1.12 Bundles over spheres: the clutching construction

To compute anything we now specialise to $M = S^n$. This is the natural setting for instantons: a finite-action field on $\mathbb{R}^4$ extends to the one-point compactification S^4 (Section 3 of this chapter and Section 8.1.16 below). Nash states that a bundle over S^n is determined by "the degree k of a homotopy class in $\pi_{n-1}(G)$". Here is where that comes from.

Cover S^n by two patches, the northern and southern "hemispheres" D_+ and D_-, slightly enlarged so that they overlap in a thin band around the equator. The band retracts onto the equator, which is a sphere S^{n-1}. Each hemisphere is a ball, hence contractible, so any bundle over it is trivial. All the information is therefore in the single transition function on the overlap,

$$t : S^{n-1} \longrightarrow G. \tag{8.45}$$

This way of building a bundle is the **clutching construction**: take two trivial pieces $D_\pm \times G$ and glue them along $S^{n-1} \times G$ by $(x, u) \mapsto (x, t(x)u)$.

Working definition 8.8 (bundles over S^n)

Principal G-bundles over S^n (G connected) correspond to homotopy classes of clutching maps $t : S^{n-1} \to G$ [two clutching maps count as the same when one can be deformed continuously into the other], i.e. to elements of $\pi_{n-1}(G)$ [these classes form a group, for example $\pi_3(\mathrm{SU}(2)) = \mathbb{Z}$]. *In practice:* write down one transition function on the equator, and compute its degree or winding number [the integer k counting how many times t wraps S^{n-1} around G]. Deforming t continuously gives an isomorphic bundle, because the deformation can be absorbed into a change of trivialisation [multiplying the fibre coordinate on that hemisphere by a map from it to G] on one hemisphere.

For n even the integer k can also be computed from the curvature, as Nash says, by integrating the Chern class $c_{n/2}$ over S^n. For $n = 2$ we did this in Example 8.2: $\int_{S^2} \frac{i}{2\pi} F = m$, the winding of $t = e^{im\phi}$. For $n = 4$ and $G = \mathrm{SU}(2)$, the analogous statement is that $-\frac{1}{8\pi^2}\int \mathrm{tr}\, F \wedge F$ equals the winding number of $t : S^3 \to \mathrm{SU}(2)$; we derive this in Section 8.1.17. The label k therefore does double duty: a topological class (the homotopy class of t) and an integral of the curvature (the Chern number).

Nash's (8.15) displays all the objects that depend on k:

$$G \to P_k \to S^n, \qquad \mathcal{G}_k \to \mathcal{A}_k \to B\mathcal{G}_k, \qquad c_{n/2}(P_k) = k, \qquad \mathcal{G}_k \subset \mathrm{Aut}(P_k). \tag{8.46}$$

For each k there is a bundle P_k, its space of connections $\mathcal{A}_k$, its (based) gauge group $\mathcal{G}_k$ of automorphisms of P_k, and the quotient $\mathcal{A}_k/\mathcal{G}_k = B\mathcal{G}_k$. (We keep the sign convention of (8.10) for $n = 4$, so strictly $k = -c_2(P_k)$ there; this changes nothing below.)

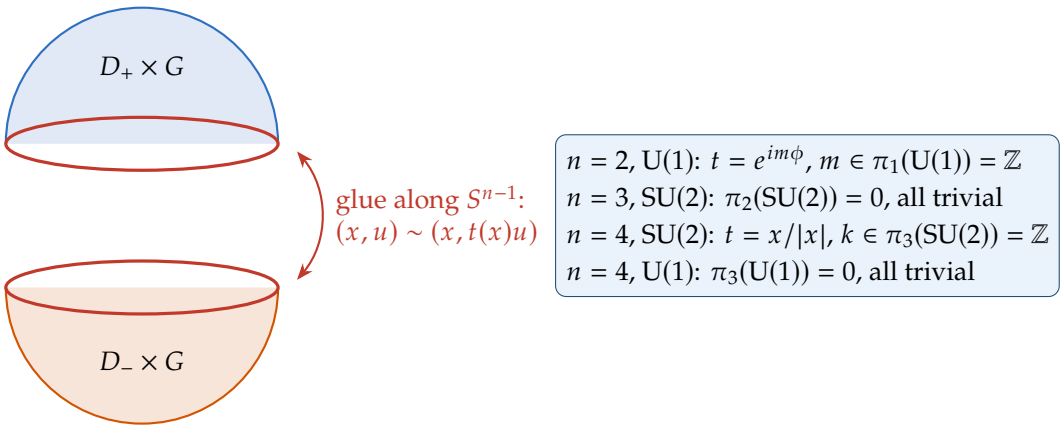


Figure 8.10: The clutching construction. A bundle over S^n is two trivial bundles over the two hemispheres, glued along the equator S^{n-1} (red) by a transition function $t : S^{n-1} \to G$. Only the homotopy class of t matters, so bundles are counted by $\pi_{n-1}(G)$ (examples on the right).

8.1.13 The gauge group is an iterated loop space

The trivial bundle

Start with $k = 0$, the trivial bundle $P_0 = S^n \times G$. A gauge transformation is then just a map $g : S^n \to G$, and a based one sends the base point m_0 to $\mathbb{1}$. This is Nash's (8.16):

$$P_0 = G \times S^n \quad \Longrightarrow \quad \mathcal{G}_0 = \mathrm{Map}_0(S^n, G) \equiv \Omega^n G. \tag{8.47}$$

Here $\mathrm{Map}_0(X, Y)$ denotes base-point-preserving continuous maps, and $\Omega^n G$ is the **n-fold loop space** of G. Let us see why the two are the same.

Working definition 8.9 (loop space)

The **loop space** ΩX of a space X with base point x_0 [a chosen point of X, for gauge groups $x_0 = \mathbb{1} \in G$] is the space of loops $\ell : [0,1] \to X$ with $\ell(0) = \ell(1) = x_0$, i.e. based maps $S^1 \to X$. The iterated loop space is $\Omega^n X = \Omega(\Omega^{n-1} X)$ [loops in the space of loops. Its base point is the constant loop at x_0]. *In practice:* a point of $\Omega^n X$ is a map from the cube $[0,1]^n$ to X that sends the whole boundary of the cube to x_0. Since a cube with its boundary collapsed to a point is a sphere S^n, this is a based map $S^n \to X$:

$$\Omega^n X = \mathrm{Map}_0(S^n, X). \tag{8.48}$$

For $n = 2$: a loop of loops is a one-parameter family $s \mapsto \ell_s$ of loops, each starting and ending at x_0, with ℓ_0 and ℓ_1 the constant loop. Write it as a function of two variables, $f(s,t) = \ell_s(t)$. It is x_0 on all four edges of the square. Collapsing the edges turns the square into a sphere, so f is a based map $S^2 \to X$ (Figure 8.11).

The homotopy groups of loop spaces are shifted:

$$\pi_j(\Omega X) \cong \pi_{j+1}(X), \qquad \pi_j(\Omega^n X) \cong \pi_{j+n}(X). \tag{8.49}$$

The reason is the same picture: a j-sphere's worth of loops is a map of $S^j \times S^1$ to X with enough of it sent to x_0 that it becomes a map of S^{j+1}. In particular $\pi_0(\Omega^n X) = \pi_n(X)$: *the components of $\Omega^n X$ are the homotopy classes of maps $S^n \to X$.*

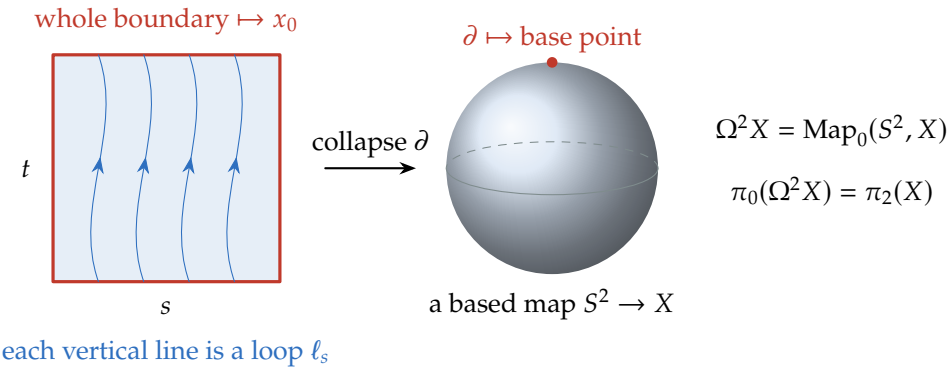


Figure 8.11: A loop of loops is a map from a square whose boundary goes to the base point. Collapsing the boundary turns the square into a sphere, so $\Omega^2 X = \mathrm{Map}_0(S^2, X)$. The same with an n-cube gives $\Omega^n X = \mathrm{Map}_0(S^n, X)$: the based gauge group of the trivial bundle over S^n is $\Omega^n G$.

Independence of k

For $k \neq 0$ the bundle P_k is twisted, and $\mathcal{G}_k$ is the group of sections of the twisted bundle $\mathrm{Ad}\, P_k$. Nash quotes Atiyah and Jones: the homotopy type does not depend on k. So up to homotopy we have Nash's (8.17) and (8.18):

$$\mathcal{G}_k \simeq \mathcal{G}_{k-1} \simeq \cdots \simeq \Omega^n G, \qquad \text{so we write} \quad \mathcal{G} = \Omega^n G. \tag{8.50}$$

Why should this be true? A clean way to see it uses the classifying space and is one line once we have (8.53) below: $B\mathcal{G}_k$ is one connected component of the space $\mathrm{Map}_0(S^n, BG)$, labelled by the bundle type k, and that mapping space is itself a loop space (it equals $\Omega^n BG$). The components of a loop space are all homotopy equivalent to each other: concatenating with a fixed loop of type $-k$ maps the k-component onto the 0-component, with inverse given by concatenating with a loop of type $+k$. So $B\mathcal{G}_k \simeq B\mathcal{G}_0$ for every k, and therefore $\mathcal{G}_k \simeq \Omega B\mathcal{G}_k \simeq \Omega B\mathcal{G}_0 \simeq \mathcal{G}_0$.

Watch out: Components: $B\mathcal{G}_k$ versus $\Omega^{n-1}G$

$\mathcal{A}_k$ is contractible, hence connected, so $B\mathcal{G}_k = \mathcal{A}_k/\mathcal{G}_k$ is *connected*. But $\Omega^{n-1}G$ has components labelled by $\pi_{n-1}(G)$, which is $\mathbb{Z}$ for $n = 4$. The correct statement is that $B\mathcal{G}_k$ is homotopy equivalent to *one* component of $\Omega^{n-1}G$ (all components being alike). Nash's equality $B\mathcal{G} \simeq \Omega^{n-1}G$ in (8.19) should be read this way; equivalently, the union over all k of the $B\mathcal{G}_k$ is $\Omega^{n-1}G$. The formula $\pi_m(B\mathcal{G}) = \pi_{m+n-1}(G)$ below is correct for $m \geq 1$; for $m = 0$ it counts the bundles P_k, not components of a single $B\mathcal{G}_k$.

8.1.14 Ω and B are inverse to each other

Nash now uses "the inverse relationship between the operands Ω and B" (his p. 73). The statement is

$$\Omega BH \simeq H, \qquad B\Omega X \simeq X \quad (X \text{ connected}). \tag{8.51}$$

Why $\Omega BH \simeq H$. Take the universal bundle $EH \to BH$ and a base point $e_0 \in EH$ over $b_0 \in BH$. Given a loop ℓ at b_0, lift it to a path in EH starting at e_0. It ends at a point $e_0 h$ of the same fibre, for a unique $h \in H$. The map $\ell \mapsto h$ is well defined up to homotopy (because EH is contractible, any two lifts are homotopic), and it is a homotopy equivalence $\Omega BH \to H$. This is the boundary map ∂ of Section 8.1.9 "before taking π_0".

Worked example 8.14: Two checks of $\Omega BH \simeq H$

- $H = \mathbb{Z}$, $BH = S^1$. The loops in S^1 fall into components labelled by the winding number $n \in \mathbb{Z}$, and each component is contractible (straight-line homotopy of the lifts in $\mathbb{R}$). So $\Omega S^1 \simeq \mathbb{Z}$.
- $H = \mathrm{U}(1)$, $BH = \mathbb{CP}^\infty$. Then $\pi_j(\Omega\mathbb{CP}^\infty) = \pi_{j+1}(\mathbb{CP}^\infty)$, which is $\mathbb{Z}$ for $j = 1$ and 0 otherwise: the same as the homotopy groups of the circle $\mathrm{U}(1)$.

Why $B\Omega X \simeq X$. Consider the **path fibration** $\Omega X \to PX \to X$, where PX is the space of paths starting at x_0 and the projection takes a path to its end point. The fibre over x_0 is the set of paths that end at x_0, i.e. ΩX. The space PX is contractible (shrink each path back along itself to x_0). Up to homotopy this is a universal ΩX-bundle (loops act on paths by concatenation), so its base X is $B\Omega X$.

Now Nash's chain (8.19) is immediate. With $\mathcal{G} = \Omega^n G = \Omega(\Omega^{n-1}G)$,

$$B\mathcal{G} \overset{(1)}{=} B\Omega^n G \overset{(2)}{=} B\,\Omega\big(\Omega^{n-1}G\big) \overset{(3)}{\simeq} \Omega^{n-1}G. \tag{8.52}$$

(1) The gauge group is $\Omega^n G$, (8.50).
(2) Peel off one Ω.
(3) $B\Omega X \simeq X$ for $X = \Omega^{n-1}G$ (component-wise; see the pitfall above).

And in the other direction, Nash's (8.20), using $G \simeq \Omega BG$:

$$\Omega^{n-1}G \overset{(1)}{\simeq} \Omega^{n-1}\Omega BG \overset{(2)}{=} \Omega^n BG \overset{(3)}{=} \mathrm{Map}_0(S^n, BG) \quad\Longrightarrow\quad B\mathcal{G} \simeq \mathrm{Map}_0(S^n, BG). \tag{8.53}$$

(1) $G \simeq \Omega BG$ from (8.51), applied inside Ω^{n-1}.
(2) $\Omega^{n-1}\Omega = \Omega^n$.
(3) (8.48) with $X = BG$.

This has a pleasing interpretation. Points of $\mathrm{Map}_0(S^n, BG)$ are, by the classification property of BG, G-bundles over S^n (with a trivialisation at the base point). A *connection* modulo gauge transformations is a point of $B\mathcal{G}$. So up to homotopy, "a connection on P modulo based gauge transformations" is the same kind of object as "a map $S^n \to BG$ in the class of P". Indeed, a connection does give such a map, by holonomy; this is the idea behind many constructions in Chapter 10.

Families of bundles, Nash's (8.21)

Iterating once more,

$$\begin{aligned} \Omega^{m+n}BG &\overset{(1)}{=} \Omega^m\Omega^n BG \overset{(2)}{\simeq} \Omega^m B\mathcal{G} \overset{(3)}{=} \mathrm{Map}_0(S^m, B\mathcal{G}), \\ \Longrightarrow\quad \mathrm{Map}_0\big(S^m, \mathrm{Map}_0(S^n, BG)\big) &\simeq \mathrm{Map}_0(S^m, B\mathcal{G}). \end{aligned} \tag{8.54}$$

(1) Split the loop space.
(2) $\Omega^n BG \simeq B\mathcal{G}$ by (8.53).
(3) (8.48) with $X = B\mathcal{G}$.

(Nash writes the left-hand side of (8.21) with the roles of m and n swapped; since $S^m \wedge S^n = S^{m+n}$ is symmetric, both orders give $\Omega^{m+n}BG$.) Read the two sides with the classification property. A map $S^m \to B\mathcal{G}$ is a $\mathcal{G}$-bundle over S^m. A map from S^m into $\mathrm{Map}_0(S^n, BG)$ is an S^m-parametrised family of G-bundles over S^n. So: *a bundle over a parameter sphere S^m with fibre the gauge group is*

the same thing as a family of gauge bundles over spacetime, parametrised by that sphere. This is how families of Dirac operators, parametrised by $\mathcal{A}/\mathcal{G}$, enter the theory of anomalies (Chapter 10, §6).

8.1.15 Homotopy groups of $\mathcal{G}$ and $B\mathcal{G}$

Collecting, Nash's (8.22) is

$$\mathcal{G} \simeq \Omega^n G, \qquad B\mathcal{G} \simeq \Omega^{n-1} G. \tag{8.55}$$

The homotopy groups follow at once from (8.49). This is Nash's (8.23):

$$\pi_m(\mathcal{G}) \overset{(1)}{=} \pi_m(\Omega^n G) \overset{(2)}{=} \pi_{m+n}(G), \tag{8.56}$$

$$\pi_m(B\mathcal{G}) \overset{(3)}{=} \pi_m(\Omega^{n-1} G) \overset{(2)}{=} \pi_{m+n-1}(G) \overset{(4)}{=} \pi_{m-1}(\mathcal{G}). \tag{8.57}$$

(1) $\mathcal{G} \simeq \Omega^n G$.
(2) Homotopy groups of loop spaces are shifted, (8.49).
(3) $B\mathcal{G} \simeq \Omega^{n-1} G$ (for $m \geq 1$).
(4) Compare with (8.56) at $m-1$; this is also (8.39), a good consistency check.

Key idea: Topology of the gauge group on a sphere

On S^n, the m-th homotopy group of the gauge group is the $(m+n)$-th homotopy group of the Lie group: $\pi_m(\mathcal{G}) = \pi_{m+n}(G)$, $\pi_m(\mathcal{A}/\mathcal{G}) = \pi_{m+n-1}(G)$. In particular $\pi_0(\mathcal{G}) = \pi_n(G)$ counts the disconnected pieces of the gauge group (*large* gauge transformations), and $\pi_1(\mathcal{A}/\mathcal{G}) = \pi_n(G)$ counts non-contractible loops of physical field configurations.

To use this we need the homotopy groups of the classical groups. We collect what is needed.

	π_1	π_2	π_3	π_4	π_5	π_6	π_7
$\mathrm{U}(1) = S^1$	$\mathbb{Z}$	0	0	0	0	0	0
$\mathrm{SU}(2) = S^3$	0	0	$\mathbb{Z}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_{12}$	$\mathbb{Z}_2$
$\mathrm{U}(2)$	$\mathbb{Z}$	0	$\mathbb{Z}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_{12}$	$\mathbb{Z}_2$
$\mathrm{SU}(3)$	0	0	$\mathbb{Z}$	0	$\mathbb{Z}$	$\mathbb{Z}_6$	0
$\mathrm{U}(N)$, $N \geq 3$ (stable, $j < 2N$)	$\mathbb{Z}$	0	$\mathbb{Z}$	0	$\mathbb{Z}$	0	$\mathbb{Z}$

Table 8.1: Homotopy groups of unitary groups. Every Lie group has $\pi_2 = 0$, and every compact simple one has $\pi_3 = \mathbb{Z}$. In the *stable range* $j < 2N$, Bott periodicity gives $\pi_j(\mathrm{U}(N)) = \mathbb{Z}$ for j odd and 0 for j even; the row for $\mathrm{U}(N)$ is valid for the entries with $j < 2N$ (so $\pi_6(\mathrm{U}(3)) = \mathbb{Z}_6$, not 0). The entries for SU(2) are the homotopy groups of the sphere S^3.

A few words on where the table comes from.

- $\mathrm{U}(N) \cong \mathrm{U}(1) \times \mathrm{SU}(N)$ *as a space* (via $u \mapsto (\det u,\ \mathrm{diag}(\det u^{-1}, 1, \ldots, 1)\, u)$), so $\pi_j(\mathrm{U}(N)) = \pi_j(\mathrm{U}(1)) \oplus \pi_j(\mathrm{SU}(N))$, and for $j \geq 2$ this is just $\pi_j(\mathrm{SU}(N))$.
- SU(2) is the 3-sphere (8.9). Its $\pi_3 = \mathbb{Z}$ is the degree of a map $S^3 \to S^3$. $\pi_4(S^3) = \mathbb{Z}_2$ is generated by the suspension of the Hopf map; the "twice" of it can be undone.
- $\mathrm{SU}(N)$ acts transitively on the unit sphere $S^{2N-1} \subset \mathbb{C}^N$ with stabiliser $\mathrm{SU}(N-1)$, giving a fibration $\mathrm{SU}(N-1) \to \mathrm{SU}(N) \to S^{2N-1}$. Its long exact sequence shows $\pi_j(\mathrm{SU}(N-1)) \to \pi_j(\mathrm{SU}(N))$ is an isomorphism for $j < 2N-2$. This is why the groups "stabilise" for large N.

Now Nash's example (8.24): $M = S^4$ and $G = \mathrm{U}(N)$. With $n = 4$,

$$\pi_1(B\mathcal{G}) \overset{(1)}{=} \pi_4(G) = \begin{cases} 0, & N = 1 \\ \mathbb{Z}_2, & N = 2 \\ 0, & N \geq 3 \end{cases} \qquad \pi_2(B\mathcal{G}) \overset{(1)}{=} \pi_5(G) = \begin{cases} 0, & N = 1 \\ \mathbb{Z}_2, & N = 2 \\ \mathbb{Z}, & N \geq 3 \end{cases} \tag{8.58}$$

(1) (8.57) with $n = 4$, then read off Table 8.1: π_4, π_5 of U(1) vanish; those of U(2) are those of S^3; for $N \geq 3$ we are in the stable range $(4, 5 < 2N)$.

By (8.56) the same groups are $\pi_0(\mathcal{G})$ and $\pi_1(\mathcal{G})$.

Worked example 8.15: Checks and interpretations in low dimensions

- $n = 1$, $G = \mathrm{U}(1)$. $\mathcal{G} = \Omega\mathrm{U}(1)$ and $\pi_0(\mathcal{G}) = \pi_1(\mathrm{U}(1)) = \mathbb{Z}$: the winding of λ in Example 8.10. $B\mathcal{G} \simeq \Omega^0\mathrm{U}(1) = \mathrm{U}(1)$: the circle $\mathcal{A}/\mathcal{G}$ we found by hand.
- $n = 2$, $G = \mathrm{U}(1)$. $\pi_m(\mathcal{G}) = \pi_{m+2}(\mathrm{U}(1)) = 0$ for all m, and each $B\mathcal{G}_k$ is contractible. So an Abelian theory on S^2 has no Gribov problem. Directly: the orbit of a is $a + \mathrm{d}\lambda$, and by Hodge theory on S^2 (no harmonic 1-forms) each orbit contains exactly one a with $\mathrm{d}^*a = 0$. The Coulomb gauge works globally, as the topology says it must.
- $n = 4$, $G = \mathrm{SU}(2)$. $\pi_0(\mathcal{G}) = \pi_4(S^3) = \mathbb{Z}_2$: the gauge group on S^4 has two components. A gauge transformation in the other component cannot be deformed to the identity. This $\mathbb{Z}_2$ is the origin of Witten's global SU(2) anomaly (Chapter 10): a theory with an odd number of Weyl doublets has a fermion determinant that changes sign under such a transformation.
- $n = 4$, $G = \mathrm{SU}(3)$. $\pi_2(\mathcal{A}/\mathcal{G}) = \pi_5(\mathrm{SU}(3)) = \mathbb{Z}$: the configuration space has a non-contractible 2-sphere. The integer is detected by the class ν_2 constructed in Section 8.1.25; it is what makes the non-Abelian anomaly possible.
- $n = 3$, G *simple*. $\pi_0(\mathcal{G}) = \pi_3(G) = \mathbb{Z}$. Gauge transformations on S^3 (or on $\mathbb{R}^3$ with $g \to \mathbb{1}$ at infinity) carry an integer winding number. This is the subject of the next section.

Reading Nash (pp. 219–220, (8.15)–(8.24))

Nash compresses the whole of this section into two pages. The "degree k of a homotopy class in $\pi_{n-1}(G)$" is the clutching function (8.45). The Atiyah–Jones independence of k is quoted without reason; we gave the loop-space argument. Nash's p. 73 is where he explains Ω and B. In (8.19)–(8.22) all equalities are homotopy equivalences, and $B\mathcal{G}$ should be read component-wise (pitfall above). The table (8.24) is (8.58); Nash's "with the same results applying to $\pi_0(\mathcal{G})$ and $\pi_1(\mathcal{G})$" is (8.57).

Recap

- Bundles over S^n are clutched by $t : S^{n-1} \to G$ and counted by $\pi_{n-1}(G)$; for n even the class is a Chern number.
- Based gauge group of the trivial bundle: $\mathcal{G} = \mathrm{Map}_0(S^n, G) = \Omega^n G$; up to homotopy the same for every k.
- $\Omega BH \simeq H$ and $B\Omega X \simeq X$ give $B\mathcal{G} \simeq \Omega^{n-1}G \simeq \mathrm{Map}_0(S^n, BG)$.
- $\pi_m(\mathcal{G}) = \pi_{m+n}(G)$, $\pi_m(B\mathcal{G}) = \pi_{m+n-1}(G)$; on S^4 with SU(2): $\pi_0(\mathcal{G}) = \mathbb{Z}_2$; with SU(3):

$\pi_2(B\mathcal{G}) = \mathbb{Z}$.

8.1.16 Finite action and the behaviour at infinity

Nash works on the sphere S^4 and labels bundles by $k \in \pi_3(G)$. A physicist works on flat Euclidean space $\mathbb{R}^4$, where every bundle is trivial. Where does the integer come from on $\mathbb{R}^4$? From the requirement that the action be finite. Finiteness forces the field to become "pure gauge" at infinity, and the gauge transformation it becomes is a map from the 3-sphere at infinity to G. Its winding number is the instanton number.

Finite action forces a flat connection at infinity

The Euclidean Yang–Mills action is (Chapter T3, and Section 3 of this chapter)

$$S[A] = -\int_{\mathbb{R}^4} \operatorname{tr}(F \wedge \star F) = \frac{1}{4}\int_{\mathbb{R}^4} \mathrm{d}^4x \, \sum_{\mu,\nu} \bigl(-\operatorname{tr} F_{\mu\nu}F_{\mu\nu}\bigr), \tag{8.59}$$

and $-\operatorname{tr} F_{\mu\nu}F_{\mu\nu} \geq 0$ because $F_{\mu\nu}$ is anti-Hermitian. In polar coordinates $\mathrm{d}^4x = r^3\,\mathrm{d}r\,\mathrm{d}\Omega_3$. If the curvature fell off only like $|F| \sim 1/r^2$, the integrand would behave like $r^{-4}\cdot r^3 = r^{-1}$ and the r-integral would diverge logarithmically. So a finite-action field must have

$$r^2\,|F| \longrightarrow 0 \qquad (r \to \infty). \tag{8.60}$$

The potential, on the other hand, is allowed to fall off only like $1/r$. Indeed it *must* be allowed to, because the leading $1/r$ piece can have zero curvature. A connection with $F = 0$ is called **flat**.

Key idea: Flat means pure gauge on a simply connected region

If $F = 0$ on a simply connected region U, then there is a map $g : U \to G$ with

$$A = g^{-1}\mathrm{d}g \qquad \text{on } U. \tag{8.61}$$

Equivalently, A is the gauge transform of the zero connection by g: $A = 0^g$.

Why. Fix a point $x_0 \in U$ and, for every $x \in U$, let $U(x)$ be the transport matrix (8.24) along some path from x_0 to x. Two different paths differ by a loop, and a loop in a simply connected region bounds a disc that can be tiled by small squares. For each small square the transport is $\mathbb{1} - \varepsilon^2 F + \cdots = \mathbb{1}$ to second order by Example 8.4; composing the squares, the transport around the whole loop is trivial. So $U(x)$ does not depend on the path. Moving the end point by $\mathrm{d}x$ gives the transport equation $\mathrm{d}U = -AU$. Put $g = U^{-1}$:

$$g^{-1}\mathrm{d}g \overset{(1)}{=} U\,\mathrm{d}\bigl(U^{-1}\bigr) \overset{(2)}{=} U\bigl(-U^{-1}\mathrm{d}U\,U^{-1}\bigr) \overset{(3)}{=} -\mathrm{d}U\,U^{-1} \overset{(4)}{=} A.$$

(1) Substitute $g = U^{-1}$.
(2) $\mathrm{d}(U^{-1}) = -U^{-1}\mathrm{d}U\,U^{-1}$, from differentiating $UU^{-1} = \mathbb{1}$.
(3) $UU^{-1} = \mathbb{1}$.
(4) The transport equation $\mathrm{d}U = -AU$ gives $-\mathrm{d}U\,U^{-1} = A$.

The region $r > R$ of $\mathbb{R}^4$ is $S^3 \times (R, \infty)$, which is simply connected. So at large r a finite-action potential has the form

$$A \;=\; g_\infty^{-1}\mathrm{d}g_\infty + O\Bigl(r^{-1-\delta}\Bigr), \qquad g_\infty : S^3_\infty \longrightarrow G, \tag{8.62}$$

where g_∞ depends only on the angles. (A function of the angles only has a gradient of size $1/r$, which is why the pure-gauge term is $O(1/r)$ while F is smaller.) The map g_∞ is the whole topological content of the field at infinity.

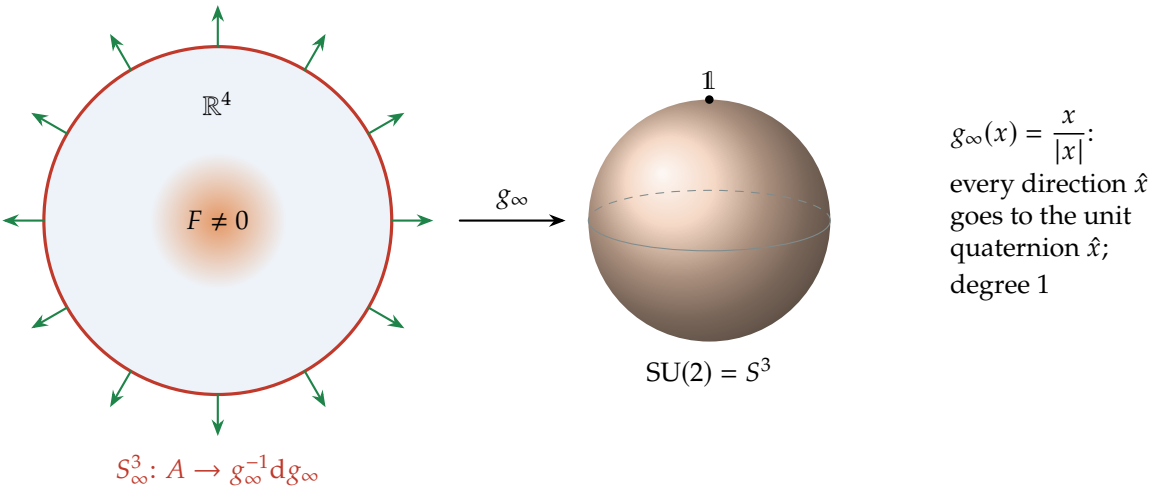


Figure 8.12: A finite-action field on $\mathbb{R}^4$. Its curvature is concentrated in a finite region (orange). Far away $F \to 0$, and the potential becomes a pure gauge $g_\infty^{-1}\mathrm{d}g_\infty$, where g_∞ assigns a group element to each direction (green arrows). For SU(2) the group is itself a 3-sphere, and g_∞ is a map $S^3 \to S^3$; the number of times it covers the target is the instanton number.

From $\mathbb{R}^4$ to a bundle over S^4

Add the point at infinity. $\mathbb{R}^4 \cup \{\infty\} = S^4$ is covered by two patches: the big ball $D_+ = \{|x| < 2R\}$ and the neighbourhood of infinity $D_- = \{|x| > R\} \cup \{\infty\}$. On D_+ use the potential A. On D_- use the gauge-transformed potential

$$A_- := A^{g_\infty^{-1}} = g_\infty A\, g_\infty^{-1} + g_\infty\, \mathrm{d}\big(g_\infty^{-1}\big) \xrightarrow{r\to\infty} 0, \tag{8.63}$$

which tends to zero at infinity (substitute (8.62): the two pure-gauge terms cancel) and so extends smoothly over the added point, after the conformal change of coordinates $y = x/|x|^2$. The two potentials glue by the gauge transformation g_∞ on the overlap, which retracts onto the equatorial S^3. In the language of Section 8.1.12:

Key idea: Finite action on $\mathbb{R}^4$ = connection on a bundle over S^4

A finite-action gauge field on $\mathbb{R}^4$ is a connection on the bundle P_k over S^4 whose clutching function is the asymptotic gauge transformation $g_\infty : S^3 \to G$. Its class $k \in \pi_3(G)$ is the homotopy class of g_∞.

This is why Nash can work on S^4 throughout: the physicist's boundary condition at infinity is the mathematician's choice of bundle.

Worked example 8.16: The two-dimensional version: vortices and $\pi_1(\mathrm{U}(1))$

On $\mathbb{R}^2$ with $G = \mathrm{U}(1)$, finite action again forces $A \to g_\infty^{-1}\mathrm{d}g_\infty$ with $g_\infty : S^1_\infty \to \mathrm{U}(1)$. Take $g_\infty = e^{im\phi}$, $m \in \mathbb{Z}$, and the smooth interpolating potential

$$A = f(r)\, g_\infty^{-1}\mathrm{d}g_\infty = i\, m\, f(r)\, \mathrm{d}\phi, \qquad f(0) = 0, \quad f(\infty) = 1,$$

with $f \sim r^2$ near the origin so that A is smooth there. Then $F = \mathrm{d}A = i\, m\, f'(r)\, \mathrm{d}r \wedge \mathrm{d}\phi$,

and the first Chern number is

$$c_1 = \frac{i}{2\pi}\int_{\mathbb{R}^2} F \overset{(1)}{=} \frac{i}{2\pi}\cdot i\,m\int_0^{2\pi}\mathrm{d}\phi\int_0^{\infty} f'(r)\,\mathrm{d}r$$
$$\overset{(2)}{=} -\frac{m}{2\pi}\cdot 2\pi\,\big(f(\infty)-f(0)\big) \overset{(3)}{=} -m.$$

(1) Substitute F; the orientation $\mathrm{d}x\wedge\mathrm{d}y = r\,\mathrm{d}r\wedge\mathrm{d}\phi$ makes $\mathrm{d}r\wedge\mathrm{d}\phi$ positive.
(2) $i\cdot i=-1$; the ϕ-integral gives 2π; the r-integral is the fundamental theorem of calculus.
(3) Boundary values of f.

The shape of f is irrelevant: only the winding of g_∞ survives. By Stokes the same number is $\frac{i}{2\pi}\oint_{S^1_\infty} A = \frac{i}{2\pi}\oint g_\infty^{-1}\mathrm{d}g_\infty$. The flux of a vortex of the Abelian Higgs model is quantised for exactly this reason. (The minus sign is the orientation bookkeeping of (8.10): with our conventions, the Chern number is minus the winding of g_∞, in two and in four dimensions.)

8.1.17 The winding number of a map $S^3 \to \mathrm{SU}(2)$

How do we compute the homotopy class of $g : S^3 \to G$? For $G = \mathrm{SU}(2) = S^3$ it is the **degree**: the number of times the map covers the target sphere, counted with orientation. Pick a generic target point $h \in \mathrm{SU}(2)$, look at its preimages, and add $+1$ for each preimage where g preserves orientation and -1 where it reverses it. There is also an integral formula, which is the one that connects to curvature.

Working definition 8.10 (winding number of a map into a Lie group)

For $g : S^3 \to G$ (with $G \subset \mathrm{U}(N)$ a matrix group) set $\theta := g^{-1}\mathrm{d}g$, the pull-back of the **Maurer–Cartan form** [the matrix-valued 1-form $h^{-1}\mathrm{d}h$ on G itself, evaluated along $h = g(\sigma)$]. The **winding number** is

$$w(g) := \frac{1}{24\pi^2}\int_{S^3}\mathrm{tr}\big(g^{-1}\mathrm{d}g\big)^3 = \frac{1}{24\pi^2}\int_{S^3}\mathrm{tr}(\theta\wedge\theta\wedge\theta), \tag{8.64}$$

where S^3 carries its orientation as the boundary of the unit ball in $\mathbb{R}^4$ (outward normal first) [a frame (v_1, v_2, v_3) on S^3 is positive when (outward normal, v_1, v_2, v_3) is positive in $\mathbb{R}^4$]. *To compute it:* in coordinates $\sigma^1, \sigma^2, \sigma^3$ on S^3, $\mathrm{tr}\,\theta^3 = \epsilon^{abc}\,\mathrm{tr}(\theta_a\theta_b\theta_c)\,\mathrm{d}\sigma^1\wedge\mathrm{d}\sigma^2\wedge\mathrm{d}\sigma^3$ with $\theta_a = g^{-1}\partial_a g$. It is an integer, equal to the degree for $G = \mathrm{SU}(2)$.

The basic example: the identity map

Let $g(x) = x$ for $x \in S^3$, in the quaternion notation (8.9): $g = x^4\mathbb{1} + x^k e_k$ with $|x| = 1$. It covers SU(2) exactly once, so we expect $w = 1$. Three observations make the integral easy.

(a) $\mathrm{tr}\,\theta^3$ *is left-invariant.* Replacing g by hg with a constant h leaves $\theta = g^{-1}h^{-1}h\,\mathrm{d}g = g^{-1}\mathrm{d}g$ unchanged. The round volume form of S^3 is also invariant under left multiplication (it is a rotation of $\mathbb{R}^4$: $|hx| = |x|$). On a 3-dimensional group, two left-invariant 3-forms are proportional. So $\mathrm{tr}\,\theta^3 = c\cdot\mathrm{vol}_{S^3}$ for a constant c, which we may compute at the identity.

(b) At the identity $x = (0,0,0,1)$, $g = \mathbb{1}$, the tangent space of S^3 is spanned by the x^1, x^2, x^3 directions ($\mathrm{d}x^4 = 0$ there), and $\theta = \mathrm{d}g = e_1\mathrm{d}x^1 + e_2\mathrm{d}x^2 + e_3\mathrm{d}x^3$. Then

$$\theta\wedge\theta\wedge\theta \overset{(1)}{=} \sum_{i,j,k} e_i e_j e_k\;\mathrm{d}x^i\wedge\mathrm{d}x^j\wedge\mathrm{d}x^k$$

$$\overset{(2)}{=} \Big(\sum_{\sigma\in S_3} \operatorname{sgn}(\sigma)\, e_{\sigma(1)}e_{\sigma(2)}e_{\sigma(3)}\Big)\mathrm{d}x^1\wedge\mathrm{d}x^2\wedge\mathrm{d}x^3$$

$$\overset{(3)}{=} 6\, e_1e_2e_3\ \mathrm{d}x^1\wedge\mathrm{d}x^2\wedge\mathrm{d}x^3 \overset{(4)}{=} -6\cdot\mathbb{1}\ \mathrm{d}x^1\wedge\mathrm{d}x^2\wedge\mathrm{d}x^3,$$

$$\operatorname{tr}\theta^3\big|_{\mathbb{1}} \overset{(5)}{=} -12\ \mathrm{d}x^1\wedge\mathrm{d}x^2\wedge\mathrm{d}x^3. \tag{8.65}$$

(1) Multiply out; the matrices do not commute, so keep their order.
(2) Only distinct i,j,k survive, and $\mathrm{d}x^{\sigma(1)}\wedge\mathrm{d}x^{\sigma(2)}\wedge\mathrm{d}x^{\sigma(3)} = \operatorname{sgn}(\sigma)\,\mathrm{d}x^1\wedge\mathrm{d}x^2\wedge\mathrm{d}x^3$.
(3) Quaternion relations: $e_ie_j = -e_je_i$ for $i\neq j$, so reordering $e_{\sigma(1)}e_{\sigma(2)}e_{\sigma(3)}$ to $e_1e_2e_3$ costs exactly $\operatorname{sgn}(\sigma)$; the six terms are equal.
(4) $e_1e_2e_3 = e_3e_3 = -\mathbb{1}$.
(5) $\operatorname{tr}\mathbb{1} = 2$.

(c) Orientation. At the point $(0,0,0,1)$ the outward normal is ∂_4. The boundary orientation declares (v_1,v_2,v_3) positive when (∂_4,v_1,v_2,v_3) is positive in $\mathbb{R}^4$. Since $(\partial_4,\partial_1,\partial_2,\partial_3)$ is an odd permutation of $(\partial_1,\dots,\partial_4)$, the positive volume form of S^3 at this point is $\mathrm{vol}_{S^3} = -\mathrm{d}x^1\wedge\mathrm{d}x^2\wedge\mathrm{d}x^3$.

Putting (a)–(c) together,

$$w(\mathrm{id}) \overset{(1)}{=} \frac{1}{24\pi^2}\int_{S^3} 12\,\mathrm{vol}_{S^3} \overset{(2)}{=} \frac{12\cdot 2\pi^2}{24\pi^2} = 1. \tag{8.66}$$

(1) By (a), $\operatorname{tr}\theta^3 = c\,\mathrm{vol}_{S^3}$ everywhere, and by (b), (c), $c = 12$: $-12\,\mathrm{d}x^{123} = 12\,(-\mathrm{d}x^{123})$.
(2) The volume of the unit 3-sphere is $2\pi^2$.

The mysterious $24\pi^2$ in (8.64) is just $12\times 2\pi^2$, chosen to make this answer 1.

Watch out: Orientation signs

Had we oriented S^3 by $\mathrm{d}x^1\wedge\mathrm{d}x^2\wedge\mathrm{d}x^3$ at the north pole we would have found $w(\mathrm{id}) = -1$. Many books absorb this sign in different places (Hermitian versus anti-Hermitian θ, x^0 first versus x^4 last). Any consistent choice works; ours makes the BPST instanton, whose potential tends to $g^{-1}\mathrm{d}g$ with $g = x/|x|$, have $k = +1$.

Why the winding number is a topological invariant

Deform g slightly, $g\mapsto g(\mathbb{1}+\xi)$, with ξ a small $\mathfrak{g}$-valued function on S^3. Then $\delta\theta = \mathrm{d}\xi + [\theta,\xi]$, and

$$\delta\operatorname{tr}\theta^3 \overset{(1)}{=} 3\operatorname{tr}\big(\theta^2\,\delta\theta\big) \overset{(2)}{=} 3\operatorname{tr}\big(\theta^2\mathrm{d}\xi\big) + 3\operatorname{tr}\big(\theta^3\xi - \theta^2\xi\theta\big)$$

$$\overset{(3)}{=} 3\operatorname{tr}\big(\theta^2\mathrm{d}\xi\big) \overset{(4)}{=} 3\,\mathrm{d}\operatorname{tr}\big(\theta^2\xi\big) - 3\operatorname{tr}\big(\mathrm{d}(\theta^2)\,\xi\big) \overset{(5)}{=} 3\,\mathrm{d}\operatorname{tr}\big(\theta^2\xi\big). \tag{8.67}$$

(1) Product rule on $\operatorname{tr}(\theta\theta\theta)$; each of the three terms can be cycled to the form $\operatorname{tr}(\theta^2\delta\theta)$, because moving a 1-form past a 2-form under the trace costs the sign $(-1)^{1\cdot 2} = +1$.
(2) Insert $\delta\theta = \mathrm{d}\xi + \theta\xi - \xi\theta$.
(3) Cycle the 1-form θ from the end of $\operatorname{tr}(\theta^2\xi\theta)$ to the front (past the 2-form $\theta^2\xi$, no sign): it equals $\operatorname{tr}(\theta^3\xi)$, so the bracket vanishes.
(4) Product rule: $\mathrm{d}\operatorname{tr}(\theta^2\xi) = \operatorname{tr}(\mathrm{d}(\theta^2)\xi) + \operatorname{tr}(\theta^2\mathrm{d}\xi)$ (moving d past the 2-form θ^2 costs no sign).
(5) The Maurer–Cartan equation $\mathrm{d}\theta = -\theta\wedge\theta$ (differentiate $g^{-1}\mathrm{d}g$) gives $\mathrm{d}(\theta^2) = \mathrm{d}\theta\,\theta - \theta\,\mathrm{d}\theta = -\theta^3 + \theta^3 = 0$.

The change of the integrand is exact, so by Stokes on the closed manifold S^3 the integral does not change. A continuous deformation is a succession of small ones, so $w(g)$ depends only on the homotopy class of g. Being also an integer (it is the degree), it cannot change continuously at all.

Winding numbers add

For two maps $g, h : S^3 \to G$, the pointwise product has

$$w(gh) = w(g) + w(h). \tag{8.68}$$

Write $a = g^{-1}\mathrm{d}g$ and $b = \mathrm{d}h\, h^{-1}$. Then $(gh)^{-1}\mathrm{d}(gh) = h^{-1}(a+b)h$, and

$$\begin{aligned}\operatorname{tr}\big((gh)^{-1}\mathrm{d}(gh)\big)^3 &\overset{(1)}{=} \operatorname{tr}(a+b)^3 \overset{(2)}{=} \operatorname{tr} a^3 + 3\operatorname{tr}(a^2 b) + 3\operatorname{tr}(ab^2) + \operatorname{tr} b^3 \\ &\overset{(3)}{=} \operatorname{tr}\big(g^{-1}\mathrm{d}g\big)^3 + \operatorname{tr}\big(h^{-1}\mathrm{d}h\big)^3 - 3\,\mathrm{d}\operatorname{tr}(ab).\end{aligned}$$

(1) Conjugation by h drops out of the trace.
(2) Expand the eight terms and cycle each to one of these four forms (moving a 1-form past a 2-form under the trace is free).
(3) $\operatorname{tr} b^3 = \operatorname{tr}(h\,h^{-1}\mathrm{d}h\,h^{-1})^3 = \operatorname{tr}(h^{-1}\mathrm{d}h)^3$; and $\mathrm{d}\operatorname{tr}(ab) = \operatorname{tr}(\mathrm{d}a\,b - a\,\mathrm{d}b) = -\operatorname{tr}(a^2 b) - \operatorname{tr}(ab^2)$ using $\mathrm{d}a = -a^2$ and $\mathrm{d}b = \mathrm{d}h\,\mathrm{d}(h^{-1})\cdot(-1)\cdot(-1) = b^2$.

The exact term integrates to zero. So w is a group homomorphism from $\pi_3(G)$ (with pointwise multiplication) to $\mathbb{Z}$.

Worked example 8.17: Maps of every degree

- $g_n(x) = x^n$ (quaternion power, $|x| = 1$). By (8.68) applied $n-1$ times, $w(g_n) = n\,w(\mathrm{id}) = n$. Likewise $w(\bar{x}) = -1$, because $\bar{x} = x^{-1}$ and $w(x\,x^{-1}) = w(\mathbb{1}) = 0$.
- A constant map has $\theta = 0$ and $w = 0$.
- A map into a U(1) subgroup, $g = \exp(\phi(x)e_3)$, has $\theta = e_3\,\mathrm{d}\phi$ and $\theta\wedge\theta = e_3^2\,\mathrm{d}\phi\wedge\mathrm{d}\phi = 0$. So $w = 0$: an Abelian map is never topologically non-trivial in three dimensions, matching $\pi_3(\mathrm{U}(1)) = 0$.
- For $G = \mathrm{SU}(N)$, embed an SU(2) map in the upper left 2×2 block. The trace in (8.64) is unchanged, so $w = 1$ is still attained, and $\pi_3(\mathrm{SU}(N)) = \mathbb{Z}$ is generated by this embedding.

8.1.18 The instanton number is the winding at infinity

Now connect curvature to winding. The key identity is that $\operatorname{tr} F\wedge F$ is exact, *locally*:

$$\operatorname{tr}(F\wedge F) = \mathrm{d}\,Q_3(A), \qquad Q_3(A) := \operatorname{tr}\Big(A\wedge\mathrm{d}A + \tfrac{2}{3}A\wedge A\wedge A\Big). \tag{8.69}$$

Q_3 is the **Chern–Simons 3-form**. We will meet it again as Nash's $TP_2(A)$ in Section 8.2. The check:

$$\begin{aligned}\mathrm{d}Q_3 &\overset{(1)}{=} \operatorname{tr}\big(\mathrm{d}A\wedge\mathrm{d}A\big) + \tfrac{2}{3}\operatorname{tr}\big(\mathrm{d}A\,A\,A - A\,\mathrm{d}A\,A + A\,A\,\mathrm{d}A\big) \\ &\overset{(2)}{=} \operatorname{tr}\big(\mathrm{d}A\wedge\mathrm{d}A\big) + 2\operatorname{tr}\big(\mathrm{d}A\wedge A\wedge A\big) \\ &\overset{(3)}{=} \operatorname{tr}\big(\mathrm{d}A\,\mathrm{d}A + \mathrm{d}A\,A^2 + A^2\,\mathrm{d}A + A^2A^2\big) \overset{(4)}{=} \operatorname{tr}\big((\mathrm{d}A + A^2)\wedge(\mathrm{d}A + A^2)\big) = \operatorname{tr} F\wedge F.\end{aligned}$$

(1) $\mathrm{d}(A \wedge \mathrm{d}A) = \mathrm{d}A \wedge \mathrm{d}A$ (since $\mathrm{d}\mathrm{d}A = 0$), and the product rule on $A\,A\,A$ with a sign each time d passes a 1-form.
(2) Cycle under the trace: $\mathrm{tr}(A\,\mathrm{d}A\,A)$ moves the last A past the 3-form $A\,\mathrm{d}A$, costing $(-1)^{1\cdot 3} = -1$, so $-\mathrm{tr}(A\,\mathrm{d}A\,A) = +\mathrm{tr}(A^2\mathrm{d}A)$; and $\mathrm{tr}(A^2\mathrm{d}A) = \mathrm{tr}(\mathrm{d}A\,A^2)$ (2-forms commute under the trace). The three terms add to $3\,\mathrm{tr}(\mathrm{d}A\,A^2)$.
(3) Split $2\,\mathrm{tr}(\mathrm{d}A\,A^2) = \mathrm{tr}(\mathrm{d}A\,A^2) + \mathrm{tr}(A^2\mathrm{d}A)$, and add $\mathrm{tr}(A^4) = 0$ (moving one 1-form past a 3-form under the trace gives $\mathrm{tr}(A^4) = -\mathrm{tr}(A^4)$).
(4) Factorise; $F = \mathrm{d}A + A \wedge A$.

Integrate over a large ball B_R and use Stokes. With the instanton number normalised as in (8.10),

$$k \overset{(1)}{=} -\frac{1}{8\pi^2}\lim_{R\to\infty}\int_{B_R}\mathrm{tr}\,F\wedge F \overset{(2)}{=} -\frac{1}{8\pi^2}\lim_{R\to\infty}\int_{S^3_R} Q_3(A)$$
$$\overset{(3)}{=} -\frac{1}{8\pi^2}\int_{S^3} Q_3\big(g_\infty^{-1}\mathrm{d}g_\infty\big) \overset{(4)}{=} -\frac{1}{8\pi^2}\int_{S^3}\Big(-\tfrac{1}{3}\Big)\,\mathrm{tr}\,\theta^3 \overset{(5)}{=} w(g_\infty). \tag{8.70}$$

(1) Definition, with $\mathbb{R}^4$ exhausted by balls.
(2) $\mathrm{tr}\,F^2 = \mathrm{d}Q_3$ on $\mathbb{R}^4$, where A is a single smooth potential; Stokes' theorem with the boundary orientation.
(3) At large R, $A = g_\infty^{-1}\mathrm{d}g_\infty + O(r^{-1-\delta})$; the correction contributes $O(R^{-\delta}) \to 0$ to an integral over a sphere of area $2\pi^2R^3$ of a form of size $R^{-3-\delta}$.
(4) For a pure gauge $A = \theta$: $\mathrm{d}\theta = -\theta^2$, so $Q_3(\theta) = \mathrm{tr}(-\theta^3 + \frac{2}{3}\theta^3) = -\frac{1}{3}\,\mathrm{tr}\,\theta^3$.
(5) $\frac{1}{8\pi^2}\cdot\frac{1}{3} = \frac{1}{24\pi^2}$, and (8.64).

Key idea: Instanton number = winding number

For a finite-action field on $\mathbb{R}^4$,

$$k = -\frac{1}{8\pi^2}\int_{\mathbb{R}^4}\mathrm{tr}\,F\wedge F = \frac{1}{24\pi^2}\int_{S^3_\infty}\mathrm{tr}\big(g_\infty^{-1}\mathrm{d}g_\infty\big)^3 = w(g_\infty) \in \mathbb{Z}.$$

The integral of a local density over all of space is an integer, because it only sees how the field wraps the group at infinity. This is the concrete content of Nash's "$c_{n/2}(P_k) = k$" in (8.15), for $n = 4$.

Worked example 8.18: Where the BPST instanton keeps its charge

The BPST instanton of size ρ centred at the origin (Section 3 of this chapter) is, in our conventions,

$$A = f(r)\,g^{-1}\mathrm{d}g, \qquad g = \frac{x}{|x|}, \qquad f(r) = \frac{r^2}{r^2+\rho^2}.$$

It tends to $g^{-1}\mathrm{d}g$ at infinity, so $k = w(x/|x|) = 1$. How much of this charge lies inside the ball of radius R? Apply the same Stokes argument on B_R. On the sphere S^3_R the function f is constant, so there $\mathrm{d}(f\theta) = f\,\mathrm{d}\theta = -f\theta^2$ and

$$k(R) \overset{(1)}{=} -\frac{1}{8\pi^2}\int_{S^3_R} Q_3(f\theta) \overset{(2)}{=} -\frac{1}{8\pi^2}\int_{S^3_R}\Big(-f^2+\tfrac{2}{3}f^3\Big)\,\mathrm{tr}\,\theta^3$$
$$\overset{(3)}{=} -\frac{24\pi^2}{8\pi^2}\Big(-f^2+\tfrac{2}{3}f^3\Big) \overset{(4)}{=} 3f^2 - 2f^3 = \frac{R^4(R^2+3\rho^2)}{(R^2+\rho^2)^3}, \qquad f = f(R).$$

(1) Stokes on B_R, with A smooth inside (it vanishes like r^2 at the origin, where g itself is undefined).

(2) $Q_3(f\theta) = \operatorname{tr}(f\theta \wedge (-f\theta^2) + \frac{2}{3}f^3\theta^3)$ restricted to the sphere.
(3) $\int_{S^3_R} \operatorname{tr}\theta^3 = 24\pi^2$ for every R (the map $x \mapsto x/|x|$ has winding one on every sphere).
(4) Algebra: $f^2(3-2f)$ with $f = R^2/(R^2+\rho^2)$.

At $R = \rho$, $k = 3/4 - 2/8 = 1/2$: half the charge lies within one instanton radius, and the density falls like ρ^4/r^8 outside. (We checked the density $\operatorname{tr} F \wedge F = -48\rho^4/(r^2+\rho^2)^4\,\mathrm{d}^4x$ symbolically; it integrates to $k = 1$.) See Figure 8.13.

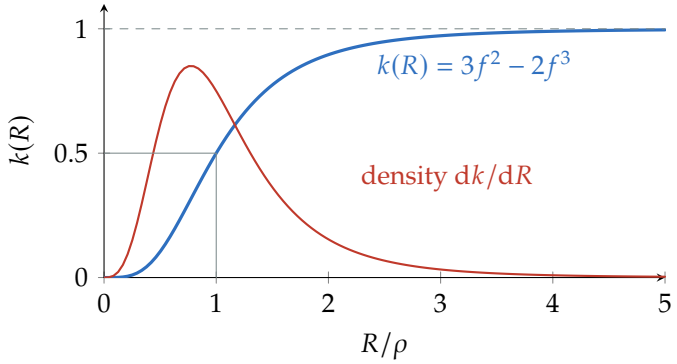


Figure 8.13: The instanton number inside a ball of radius R for the BPST instanton of size ρ (blue), and its radial density (red). Half the charge lies within $R = \rho$, and the total is exactly 1, whatever ρ: the size is a modulus, the charge is topological.

Watch out: The winding can move, but not disappear

Apply the gauge transformation g^{-1} (with $g = x/|x|$) to the BPST potential. Far away the new potential tends to 0; near the origin it has the singular piece $g\,\mathrm{d}(g^{-1})$. This *singular gauge* is popular for multi-instanton ansätze ('t Hooft). The winding has moved from infinity to a small sphere around the origin, where g is singular. The total, computed as the integral of the gauge-invariant density $\operatorname{tr} F \wedge F$, is still 1. What cannot happen is a *smooth* gauge transformation on all of $\mathbb{R}^4$ that changes g_∞'s class, because w of the restriction of a smooth map on the ball to its boundary is zero (the map extends over the ball, so it is null-homotopic).

8.1.19 Large gauge transformations and the Chern–Simons number

Now switch from four-dimensional Euclidean spacetime to the Hamiltonian picture: three-dimensional space, evolving in time. Here $\pi_3(G)$ appears again, this time as π_0 of the gauge group of *space*.

The gauge group of space is disconnected

In the temporal gauge $A_4 = 0$ (Euclidean) or $A_0 = 0$ (Minkowski), the potential is a time-dependent family $A(t) = A_i(t,\vec{x})\,\mathrm{d}x^i$ of connections on space $\mathbb{R}^3$. The gauge transformations that preserve $A_0 = 0$ are the time-independent ones, $g(\vec{x})$. For the Gauss law and the charges to make sense we require $g \to \mathbb{1}$ at spatial infinity. So g is a based map from $\mathbb{R}^3 \cup \{\infty\} = S^3$ to G, and the residual gauge group is

$$\mathcal{G}_{\text{space}} = \operatorname{Map}_0(S^3, G) = \Omega^3 G, \qquad \pi_0(\mathcal{G}_{\text{space}}) = \pi_3(G) = \mathbb{Z}. \tag{8.71}$$

This is Nash's $\mathcal{G}$ with $n = 3$, and (8.56) with $m = 0$. Gauge transformations with $n(g) = 0$ are connected to the identity: **small** gauge transformations. The others are **large**.

Worked example 8.19: An explicit large gauge transformation

Inverse stereographic projection maps $\mathbb{R}^3 \cup \{\infty\}$ onto the unit quaternions. With a length scale $\lambda > 0$ define

$$g_1(\vec{x}) = \frac{(r^2 - \lambda^2)\,\mathbb{1} - 2\lambda\, x^k e_k}{r^2 + \lambda^2}, \qquad r = |\vec{x}|. \tag{8.72}$$

It is unitary: $|g_1|^2 = \big((r^2-\lambda^2)^2 + 4\lambda^2 r^2\big)/(r^2+\lambda^2)^2 = 1$. It tends to $\mathbb{1}$ as $r \to \infty$ and equals $-\mathbb{1}$ at the origin. Every unit quaternion is hit exactly once. Symbolic computation gives

$$\operatorname{tr}\big(g_1^{-1}\mathrm{d}g_1\big)^3 = -\frac{96\,\lambda^3}{(r^2+\lambda^2)^3}\,\mathrm{d}x^1 \wedge \mathrm{d}x^2 \wedge \mathrm{d}x^3.$$

At the origin this is easy to check by hand: there $g_1 = -\mathbb{1}$ and $\mathrm{d}g_1 = -(2/\lambda)\,e_k \mathrm{d}x^k$, so $\theta = (2/\lambda) e_k \mathrm{d}x^k$ and, by (8.65), $\operatorname{tr}\theta^3 = (2/\lambda)^3(-12)\,\mathrm{d}x^{123} = -96\lambda^{-3}\mathrm{d}x^{123}$. Integrating,

$$\int_{\mathbb{R}^3} \operatorname{tr}\theta^3 \overset{(1)}{=} -96\lambda^3 \cdot 4\pi \int_0^\infty \frac{r^2\,\mathrm{d}r}{(r^2+\lambda^2)^3} \overset{(2)}{=} -96\lambda^3 \cdot 4\pi \cdot \frac{\pi}{16\lambda^3} = -24\pi^2.$$

(1) The density depends only on r; $\mathrm{d}^3x = 4\pi r^2 \mathrm{d}r$ after the angular integral.

(2) $\int_0^\infty r^2(r^2+\lambda^2)^{-3}\mathrm{d}r = \pi/(16\lambda^3)$ (substitute $r = \lambda\tan u$: the integral becomes $\lambda^{-3}\int_0^{\pi/2}\sin^2 u\cos^2 u\,\mathrm{d}u = \lambda^{-3}\pi/16$).

Working definition 8.11 (winding of a gauge transformation of space)

For $g : \mathbb{R}^3 \to G$ with $g \to \mathbb{1}$ at infinity [so g extends to a map $S^3 \to G$, with the point at infinity sent to $\mathbb{1}$],

$$n(g) := -\frac{1}{24\pi^2}\int_{\mathbb{R}^3} \operatorname{tr}\big(g^{-1}\mathrm{d}g\big)^3, \tag{8.73}$$

with $\mathbb{R}^3$ oriented by $\mathrm{d}x^1 \wedge \mathrm{d}x^2 \wedge \mathrm{d}x^3$. So $n(g_1) = +1$. The minus sign is orientation bookkeeping: a slice $\{x^4 = \text{const}\}$, seen as the *top* face of a slab in $\mathbb{R}^4$ oriented by $\mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^4$, carries the boundary orientation [outward normal first. On the top face the outward normal is ∂_4, and moving it past three vectors gives a sign -1] $-\mathrm{d}x^1 \wedge \mathrm{d}x^2 \wedge \mathrm{d}x^3$ (the same computation as in (c) above). With that orientation (8.73) is exactly (8.64).

The Chern–Simons number and how it changes

Define the **Chern–Simons number** of a spatial configuration,

$$N_{\mathrm{CS}}(A) := \frac{1}{8\pi^2}\int_{\mathbb{R}^3} Q_3(A) = \frac{1}{8\pi^2}\int_{\mathbb{R}^3} \operatorname{tr}\Big(A \wedge \mathrm{d}A + \tfrac{2}{3}A^3\Big). \tag{8.74}$$

It is *not* gauge invariant. Its change under a gauge transformation is the most important formula of this section. Write $b := \mathrm{d}g\,g^{-1}$, so that $A^g = g^{-1}(A+b)g$ and $\theta = g^{-1}\mathrm{d}g = g^{-1}bg$. It is convenient to write $Q_3 = \operatorname{tr}(AF) - \frac{1}{3}\operatorname{tr}A^3$ (substitute $\mathrm{d}A = F - A^2$). Then

$$\begin{aligned} Q_3(A^g) &\overset{(1)}{=} \operatorname{tr}\big((A+b)F\big) - \tfrac{1}{3}\operatorname{tr}(A+b)^3 \\ &\overset{(2)}{=} Q_3(A) + \operatorname{tr}(bF) - \operatorname{tr}(A^2 b) - \operatorname{tr}(Ab^2) - \tfrac{1}{3}\operatorname{tr}b^3 \end{aligned}$$

$$\overset{(3)}{=} Q_3(A) + \operatorname{tr}(b\,\mathrm{d}A) - \operatorname{tr}(Ab^2) - \tfrac{1}{3}\operatorname{tr}\theta^3$$
$$\overset{(4)}{=} Q_3(A) - \mathrm{d}\operatorname{tr}(bA) - \tfrac{1}{3}\operatorname{tr}\theta^3. \tag{8.75}$$

(1) $A^g = g^{-1}(A+b)g$ and $F^g = g^{-1}Fg$; the conjugations drop out of the traces.
(2) Expand $\operatorname{tr}(A+b)^3 = \operatorname{tr}A^3 + 3\operatorname{tr}(A^2b) + 3\operatorname{tr}(Ab^2) + \operatorname{tr}b^3$ as in the proof of (8.68).
(3) $\operatorname{tr}(bF) = \operatorname{tr}(b\,\mathrm{d}A) + \operatorname{tr}(bA^2)$ and $\operatorname{tr}(bA^2) = \operatorname{tr}(A^2b)$ (a 1-form past a 2-form); $\operatorname{tr}b^3 = \operatorname{tr}\theta^3$ by conjugation.
(4) $\mathrm{d}b = \mathrm{d}g\wedge\mathrm{d}g^{-1}\cdot(-1) = b^2$, so $\mathrm{d}\operatorname{tr}(bA) = \operatorname{tr}(b^2A) - \operatorname{tr}(b\,\mathrm{d}A)$, and $\operatorname{tr}(b^2A) = \operatorname{tr}(Ab^2)$.

(We also checked (8.75) symbolically for random non-Abelian A and g.) Integrate over $\mathbb{R}^3$. The exact term gives a surface integral at infinity, which vanishes because $b\to 0$ there. The last term gives the winding:

$$N_{\mathrm{CS}}(A^g) = N_{\mathrm{CS}}(A) - \frac{1}{24\pi^2}\int_{\mathbb{R}^3}\operatorname{tr}\theta^3 = N_{\mathrm{CS}}(A) + n(g). \tag{8.76}$$

Key idea: N_{CS} is gauge invariant modulo integers

Small gauge transformations leave N_{CS} unchanged; a large one of winding n shifts it by the integer n. So N_{CS} mod 1, equivalently $e^{2\pi i N_{\mathrm{CS}}}$, is gauge invariant. This is our first example of a quantity that is only defined in $\mathbb{R}/\mathbb{Z}$: a *secondary* invariant (Section 8.2.7).

Classical vacua and the instanton as a tunnelling path

A classical vacuum of the Hamiltonian theory has zero energy, so zero field strength on space: $F=0$ on $\mathbb{R}^3$. By (8.61) it is a pure gauge, $A = g^{-1}\mathrm{d}g$. Its Chern–Simons number is

$$N_{\mathrm{CS}}\big(g^{-1}\mathrm{d}g\big) \overset{(1)}{=} N_{\mathrm{CS}}(0^g) \overset{(2)}{=} N_{\mathrm{CS}}(0) + n(g) = n(g).$$

(1) A pure gauge is the gauge transform of the zero connection.
(2) (8.76); $Q_3(0)=0$.

So the classical vacua fall into classes $|n\rangle$, $n\in\mathbb{Z}$, labelled by their Chern–Simons number. Vacua in the same class are related by small gauge transformations: they are physically the same state. Vacua in different classes are related only by large gauge transformations. To go from one class to another *continuously* one must pass through configurations with $F\neq 0$, i.e. over an energy barrier (its top is the **sphaleron**).

In Euclidean time the passage is an instanton. Take the slab $[-T,T]\times\mathbb{R}^3$ in temporal gauge, with $A\to\mathbb{1}$-based pure gauges at spatial infinity and at $x^4=\pm T$. Stokes' theorem on the slab gives

$$k \overset{(1)}{=} -\frac{1}{8\pi^2}\int_{\mathrm{slab}}\mathrm{d}Q_3 \overset{(2)}{=} -\frac{1}{8\pi^2}\Big(-\int_{\mathbb{R}^3}Q_3\big|_{x^4=T} + \int_{\mathbb{R}^3}Q_3\big|_{x^4=-T}\Big)$$
$$\overset{(3)}{=} N_{\mathrm{CS}}\big(A(T)\big) - N_{\mathrm{CS}}\big(A(-T)\big) \xrightarrow{T\to\infty} n_+ - n_-. \tag{8.77}$$

(1) $\operatorname{tr}F\wedge F = \mathrm{d}Q_3$ and (8.10).
(2) The boundary of the slab is the top face with orientation $-\mathrm{d}x^{123}$, the bottom face with $+\mathrm{d}x^{123}$, and the sides at spatial infinity, which contribute nothing in temporal gauge ($A_4=0$ and the Q_3 components along $\mathrm{d}x^4$ vanish there).
(3) Definition (8.74); at $x^4=\pm T\to\pm\infty$ the field is a vacuum of class $n_\pm$.

An instanton of charge k is a Euclidean path that tunnels from the vacuum $|n\rangle$ to the vacuum $|n+k\rangle$.

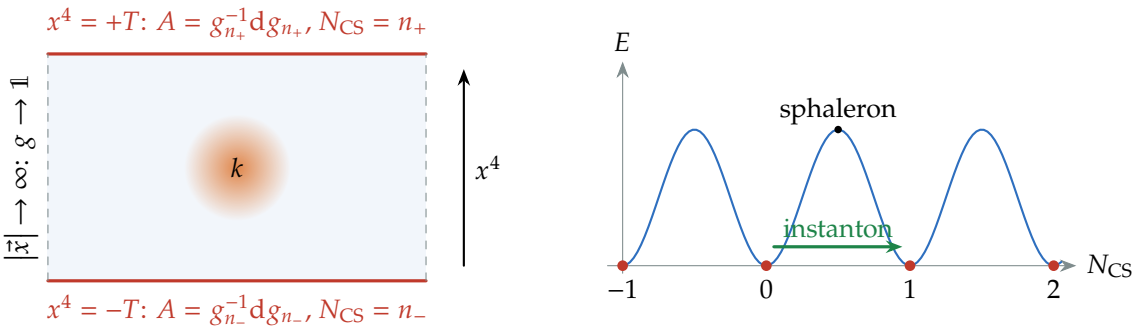


Figure 8.14: Left: an instanton in temporal gauge on the Euclidean slab. The field starts and ends as a pure gauge, and $k = N_{\mathrm{CS}}(+T) - N_{\mathrm{CS}}(-T)$. Right: the potential energy of Yang–Mills theory along a path in configuration space, plotted against the Chern–Simons number. The classical vacua sit at the integers and are related by large gauge transformations. The instanton (green) tunnels under the barrier, whose top is the sphaleron.

Physical picture: A non-contractible loop in configuration space

Walk in $\mathcal{A}$ from $A = 0$ to $A = g_1^{-1}\mathrm{d}g_1$ along any path, for example the straight line. In $\mathcal{A}$ this is an open path. In $\mathcal{A}/\mathcal{G}_{\mathrm{space}}$ its two ends are the same point, because they differ by the gauge transformation g_1. So it is a *loop* in the physical configuration space, and it cannot be contracted, because N_{CS} increases by one along any lift. This is precisely $\pi_1(\mathcal{A}/\mathcal{G}) = \pi_0(\mathcal{G}) = \pi_3(G) = \mathbb{Z}$ for space S^3, from (8.39) and (8.56). The U(1) circle of Example 8.10 is the same phenomenon one dimension down.

8.1.20 θ-vacua

Quantum mechanics on a configuration space with a non-contractible loop has a freedom that classical mechanics does not: a wave function may pick up a phase when it is carried around the loop. For Yang–Mills this phase is the vacuum angle θ. It is cleanest to see it first in the smallest possible example.

Worked example 8.20: A particle on a circle: the θ-angle in quantum mechanics

A particle of moment of inertia I moves on a circle, angle ϕ. Its configuration space is S^1, with $\pi_1(S^1) = \mathbb{Z}$. Work on the covering space $\mathbb{R} \ni \phi$ (the analogue of $\mathcal{A}$) where ϕ and $\phi + 2\pi$ are "gauge equivalent"; the translation $T : \phi \mapsto \phi + 2\pi$ is the analogue of a large gauge transformation. Physical states must be eigenstates of T (it commutes with H), with eigenvalue a phase:

$$\psi(\phi + 2\pi) = e^{i\theta}\,\psi(\phi).$$

The energy eigenstates of $H = -\frac{1}{2I}\partial_\phi^2$ with this twisted periodicity are $\psi_m = e^{i(m+\theta/2\pi)\phi}$, $m \in \mathbb{Z}$, and

$$E_m(\theta) \overset{(1)}{=} \frac{1}{2I}\left(m + \frac{\theta}{2\pi}\right)^2.$$

(1) $-\partial_\phi^2 e^{i\kappa\phi} = \kappa^2 e^{i\kappa\phi}$ with $\kappa = m + \theta/2\pi$; the twisted condition requires $e^{2\pi i\kappa} = e^{i\theta}$.

The spectrum depends on θ and is periodic in it: $\theta \to \theta + 2\pi$ just relabels $m \to m + 1$. Now undo the twist: $\psi = e^{i\theta\phi/2\pi}\chi$ with χ periodic. Then $H\chi = \frac{1}{2I}(-i\partial_\phi + \theta/2\pi)^2\chi$. This

is the Hamiltonian of the Lagrangian

$$L = \frac{I}{2}\dot{\phi}^2 - \frac{\theta}{2\pi}\dot{\phi}.$$

The extra term is a total derivative: it does not change the classical equations of motion. But its integral over a closed Euclidean history is $\theta \times \frac{1}{2\pi}\oint \dot{\phi}\,\mathrm{d}t = \theta \times$ (winding number of the history). So in the path integral histories are weighted by $e^{i\theta\cdot\text{winding}}$. Physically, θ is an Aharonov–Bohm flux threading the ring.

Yang–Mills: large gauge transformations on the Hilbert space

In the quantum theory the Gauss law $D_iE_i = 0$ generates small gauge transformations, and physical states are invariant under them. The large ones are not generated by the Gauss law. Let $\mathcal{T}$ be the unitary operator that implements g_1. It commutes with the Hamiltonian (which is gauge invariant), and on the perturbative vacua

$$\mathcal{T}|n\rangle = |n+1\rangle. \tag{8.78}$$

A physical vacuum must be an eigenstate of $\mathcal{T}$, with a unimodular eigenvalue $e^{i\theta}$. The solution is the Bloch-wave superposition, the **θ-vacuum**

$$|\theta\rangle := \sum_{n\in\mathbb{Z}} e^{-in\theta}|n\rangle. \tag{8.79}$$

Check:

$$\mathcal{T}|\theta\rangle \overset{(1)}{=} \sum_n e^{-in\theta}|n+1\rangle \overset{(2)}{=} \sum_{n'} e^{-i(n'-1)\theta}|n'\rangle \overset{(3)}{=} e^{i\theta}|\theta\rangle.$$

(1) (8.78), term by term.
(2) Relabel $n' = n+1$.
(3) Pull out $e^{i\theta}$.

Gauge-invariant operators commute with $\mathcal{T}$, so they cannot connect different θ: each value of $\theta \in [0, 2\pi)$ labels a separate **superselection sector**, a separate version of the theory.

The θ-term in the path integral

The Euclidean amplitude between perturbative vacua is a path integral over fields that tunnel from n to n', i.e. over the sector with instanton number $k = n' - n$ by (8.77): $\langle n'|e^{-HT}|n\rangle = Z_{n'-n}(T)$, depending only on the difference. Then

$$\begin{aligned}\langle\theta'|e^{-HT}|\theta\rangle &\overset{(1)}{=} \sum_{n,n'} e^{in'\theta'}e^{-in\theta}\,Z_{n'-n}(T) \overset{(2)}{=} \sum_n e^{in(\theta'-\theta)}\sum_k e^{ik\theta'}Z_k(T)\\ &\overset{(3)}{=} 2\pi\,\delta(\theta'-\theta)\sum_{k\in\mathbb{Z}} e^{ik\theta}Z_k(T) \overset{(4)}{=} 2\pi\,\delta(\theta'-\theta)\int_{\mathcal{A}}\mathcal{D}A\; e^{-S[A]+i\theta k[A]}.\end{aligned} \tag{8.80}$$

(1) Expand both θ-states by (8.79); the bra carries the complex conjugate phase.
(2) Substitute $n' = n + k$.
(3) $\sum_n e^{in\alpha} = 2\pi\delta(\alpha)$ on $[0, 2\pi)$; set $\theta' = \theta$ in the remaining sum.
(4) Z_k is the path integral over the sector k; summing sectors with the weight $e^{ik\theta}$ is the same as integrating over all fields with the extra factor $e^{i\theta k[A]}$.

So the θ-vacuum is equivalent to adding a **topological term** to the action:

$$S_\theta[A] = S[A] - i\theta\, k[A] = S[A] + \frac{i\theta}{8\pi^2}\int \operatorname{tr} F\wedge F. \tag{8.81}$$

It is a total derivative ($\operatorname{tr} F\wedge F = \mathrm{d}Q_3$), so it never shows up in perturbation theory or in the classical equations. Yet it changes the physics, because it weights the topologically distinct sectors differently. And θ is an angle: since $k\in\mathbb{Z}$, $\theta\mapsto\theta+2\pi$ does nothing.

Dictionary: the θ-term in physics normalisation

With $A = -ig_{\mathrm{YM}}A^aT^a$, $\operatorname{tr}(T^aT^b)=\frac12\delta^{ab}$, and $\tilde F^{a\mu\nu}=\frac12\epsilon^{\mu\nu\rho\sigma}F^a_{\rho\sigma}$ ($\epsilon^{1234}=1$):

$$k \overset{(1)}{=} -\frac{1}{8\pi^2}\int\left(-g_{\mathrm{YM}}^2\right)\tfrac12 F^a\wedge F^a \overset{(2)}{=} \frac{g_{\mathrm{YM}}^2}{16\pi^2}\int \tfrac14\epsilon^{\mu\nu\rho\sigma}F^a_{\mu\nu}F^a_{\rho\sigma}\,\mathrm{d}^4x \overset{(3)}{=} \frac{g_{\mathrm{YM}}^2}{32\pi^2}\int F^a_{\mu\nu}\tilde F^{a\mu\nu}\mathrm{d}^4x.$$

(1) $\operatorname{tr}(F\wedge F) = (-ig_{\mathrm{YM}})^2F^a\wedge F^b\,\operatorname{tr}(T^aT^b)$.
(2) $F^a=\frac12F^a_{\mu\nu}\mathrm{d}x^\mu\wedge\mathrm{d}x^\nu$, and $\mathrm{d}x^\mu\wedge\mathrm{d}x^\nu\wedge\mathrm{d}x^\rho\wedge\mathrm{d}x^\sigma=\epsilon^{\mu\nu\rho\sigma}\mathrm{d}^4x$.
(3) Definition of $\tilde F$.

So the familiar $\mathcal{L}_\theta=\theta\frac{g^2}{32\pi^2}F\tilde F$ is θ times the instanton density. In QCD it violates P and CP; the neutron electric dipole moment bounds $|\theta|\lesssim 10^{-10}$ (the strong CP problem).

The dilute instanton gas: the vacuum energy depends on θ

How big are the Z_k? The Bogomolny bound (Section 3 of this chapter) says $S\geq 8\pi^2|k|/g_{\mathrm{YM}}^2$ in physics normalisation, with equality for (anti-)instantons. The QFT notes [QFT notes §4.13.4, (4.13.13)–(4.13.14)] make the point that such a sector contributes a factor $e^{-8\pi^2|k|/g^2}$, which is invisible at every order of perturbation theory in g. If instantons are rare and far apart (dilute), a configuration with N instantons and $\bar N$ anti-instantons at arbitrary positions and times contributes $(KVTe^{-S_0})^{N+\bar N}/(N!\bar N!)$, with $S_0=8\pi^2/g^2$, V the volume of space, and K a determinant factor from the small fluctuations. Its instanton number is $k=N-\bar N$. Then

$$\begin{aligned}\sum_k e^{ik\theta}Z_k &\overset{(1)}{=} Z_{\mathrm{pert}}\sum_{N,\bar N\geq 0}\frac{\left(KVTe^{-S_0}e^{i\theta}\right)^N}{N!}\frac{\left(KVTe^{-S_0}e^{-i\theta}\right)^{\bar N}}{\bar N!}\\ &\overset{(2)}{=} Z_{\mathrm{pert}}\exp\left(2KVTe^{-S_0}\cos\theta\right)\overset{(3)}{=} e^{-E(\theta)T},\qquad \frac{E(\theta)}{V}=\frac{E_0}{V}-2Ke^{-8\pi^2/g^2}\cos\theta.\end{aligned} \tag{8.82}$$

(1) Each configuration with N instantons and $\bar N$ anti-instantons has $k=N-\bar N$, so $e^{ik\theta}=(e^{i\theta})^N(e^{-i\theta})^{\bar N}$; the $1/N!\bar N!$ removes the relabellings of identical objects.
(2) Two exponential series; $e^{i\theta}+e^{-i\theta}=2\cos\theta$.
(3) At large T, $\langle\theta|e^{-HT}|\theta\rangle\propto e^{-E(\theta)T}$ defines the vacuum energy; $Z_{\mathrm{pert}}=e^{-E_0T}$.

The vacuum energy density is lowest at $\theta=0$ and depends on θ through a non-perturbatively small term. This is the tight-binding band of a particle in a periodic potential (Figure 8.14), with θ as crystal momentum.

Watch out: No instantons in Abelian theories on $\mathbb{R}^4$

The QFT notes [QFT notes §4.13.4] illustrate instantons with the Abelian action $\frac{1}{g^2}\int\operatorname{tr}F\wedge\star F$ and write $\int\operatorname{tr}F\wedge F\in 8\pi^2\mathbb{Z}$. Read that schematically. For $G=\mathrm{U}(1)$ on $\mathbb{R}^4$ (or S^4) every bundle is trivial, because $\pi_3(\mathrm{U}(1))=0$ (Example 8.17); a finite-action Abelian field has

$\int F \wedge F = 0$. Abelian instantons exist only on 4-manifolds with non-zero H^2, such as T^4 or $\mathbb{CP}^2$. On $\mathbb{R}^4$ one needs a non-Abelian group, and SU(2) is the smallest.

Reading Nash (p. 219, (8.15); p. 220, (8.23) with $n = 3$; p. 227, (8.59))

Nash never writes the winding integral. He states that $P_k \to S^4$ has $c_2(P_k) = k$ and that k is the degree of a class in $\pi_3(G)$; the chain (8.70) is the bridge between the two statements. His formula $\pi_0(\mathcal{G}) = \pi_n(G)$ with $n = 3$ is the statement that large gauge transformations of space exist; he uses it himself in (8.59) for the three-dimensional Chern–Simons theory. The formula (8.75) is the simplest case of the $\mathbb{R}/\mathbb{Z}$ property of secondary classes in his (8.45).

Recap

- Finite action on $\mathbb{R}^4 \Rightarrow A \to g_\infty^{-1}\mathrm{d}g_\infty$, with $g_\infty : S^3_\infty \to G$ the clutching function of a bundle over S^4.
- Winding number $w(g) = \frac{1}{24\pi^2}\int \mathrm{tr}(g^{-1}\mathrm{d}g)^3$: an integer, homotopy invariant (its variation is exact), additive; $w(x/|x|) = 1$.
- $\mathrm{tr}\, F \wedge F = \mathrm{d}\,\mathrm{tr}(A\mathrm{d}A + \frac{2}{3}A^3) \Rightarrow k = -\frac{1}{8\pi^2}\int \mathrm{tr}\, F \wedge F = w(g_\infty)$.
- On space S^3, $\pi_0(\mathcal{G}) = \pi_3(G) = \mathbb{Z}$: large gauge transformations. N_{CS} shifts by $n(g)$ under them; vacua $|n\rangle$; instantons tunnel $n \to n + k$.
- θ-vacua $|\theta\rangle = \sum e^{-in\theta}|n\rangle \Leftrightarrow$ the topological term $e^{i\theta k}$ in the path integral; $E(\theta) \propto -e^{-8\pi^2/g^2}\cos\theta$.

8.1.21 From homotopy to cohomology: why and how

Homotopy groups tell us *that* the gauge group and the configuration space have holes. They do not hand us anything we can integrate. In physics we want more: a differential form on $\mathcal{G}$ or on $\mathcal{A}/\mathcal{G}$ that we can write down, evaluate on a family of fields, and put into a path integral or an effective action. That is what cohomology gives. A de Rham class on $\mathcal{G}$ is a closed form μ; its integral over a closed family of gauge transformations (a loop, a sphere) is a number that does not change when the family is deformed.

The prime physical example is the *non-Abelian anomaly* (Chapter 10). The phase of a chiral fermion determinant changes under an infinitesimal gauge transformation ξ by $\delta_\xi \log\det = i\,\mathcal{G}(\xi)$. Here $\mathcal{G}$ is a 1-form on the gauge group, and the Wess–Zumino consistency condition says exactly that it is closed. If it is exact, $\mathcal{G} = \delta\Gamma$, a local counterterm Γ removes it; if not, it is a genuine anomaly. So "is there an anomaly?" becomes "what is $H^1(\mathcal{G})$?". Nash computes $H^*(\mathcal{G})$ and $H^*(B\mathcal{G})$ in this subsection and gives explicit representatives in Section 8.2.

Free part and torsion

Integer cohomology $H^p(X;\mathbb{Z})$ is a finitely generated Abelian group (for the spaces we meet), so it splits as

$$H^p(X;\mathbb{Z}) \cong \mathbb{Z}^{b_p} \oplus (\text{finite group}).$$

The $\mathbb{Z}^{b_p}$ is the **free part**; b_p is the p-th Betti number. The finite group is the **torsion**. De Rham cohomology sees only the free part: $H^p_{\mathrm{dR}}(X) \cong \mathbb{R}^{b_p}$. The reason is simple. If a class x has $2x = 0$,

then every integral of a representing form over a cycle c satisfies $2\int_c \omega = 0$, so the integral is 0; the class is invisible to integration.

Worked example 8.21: Torsion that forms cannot see

Take $X = \mathbb{RP}^3 = \mathrm{SO}(3)$. Its integral homology is $H_1 = \mathbb{Z}_2$ (a loop that goes once round is not a boundary, twice round is), $H_2 = 0$, $H_3 = \mathbb{Z}$. So $H^1_{\mathrm{dR}}(\mathrm{SO}(3)) = 0$: every closed 1-form on SO(3) is exact, although SO(3) has a non-contractible loop. Concretely, a closed 1-form ω integrates to some number c around the generating loop γ; then γ^2 bounds, so

$$2c \overset{(1)}{=} \oint_{\gamma^2} \omega \overset{(2)}{=} \int_\Sigma \mathrm{d}\omega \overset{(3)}{=} 0,$$

(1) Going round γ twice doubles the integral.
(2) $\gamma^2 = \partial\Sigma$ for a surface Σ, and Stokes.
(3) ω is closed.

so $c = 0$. The physical cousin: in Example 8.15 we found $\pi_0(\mathcal{G}) = \mathbb{Z}_2$ for SU(2) on S^4. No de Rham class on $\mathcal{G}$ can detect the two components as being different in a way stable under products, and indeed Witten's SU(2) anomaly is a torsion ("global") effect, not a local one.

Nash states the division of labour clearly: in this chapter only the free part of $H^*(\mathcal{G})$ is computed, by de Rham forms. The torsion, which needs finer tools, is left to the anomaly chapter.

Rational homotopy

The free part is governed by homotopy groups *up to torsion*. For a finitely generated Abelian group π, the tensor product $\pi \otimes \mathbb{Q}$ keeps the free part and kills the torsion: $\mathbb{Z}\otimes\mathbb{Q} = \mathbb{Q}$, $\mathbb{Z}_k \otimes \mathbb{Q} = 0$ (because $1 \otimes q = 1 \otimes k \cdot (q/k) = k \cdot 1 \otimes q/k = 0$). The groups $\pi_i(X) \otimes \mathbb{Q}$ are the **rational homotopy groups**. Two spaces are **rationally homotopy equivalent**, written $X \sim Y$, if there is a map between them (or a chain of maps) inducing isomorphisms $\pi_i \otimes \mathbb{Q}$ for all i (for the simply connected or H-space case we need). Rationally equivalent spaces have the same cohomology with $\mathbb{Q}$ or $\mathbb{R}$ coefficients, so the same de Rham cohomology.

Working definition 8.12 (rational homotopy groups of spheres)

Serre's theorem: the only non-zero rational homotopy groups of spheres are

$$\pi_i(S^{2k-1}) \otimes \mathbb{Q} = \begin{cases} \mathbb{Q}, & i = 2k-1 \\ 0, & \text{otherwise,} \end{cases} \qquad \pi_i(S^{2k}) \otimes \mathbb{Q} = \begin{cases} \mathbb{Q}, & i = 2k \text{ or } i = 4k-1 \\ 0, & \text{otherwise.} \end{cases}$$

How to use it: every entry of a table of homotopy groups of spheres is finite [a finite group, which becomes 0 after $\otimes\mathbb{Q}$] except $\pi_{2k-1}(S^{2k-1}) = \mathbb{Z}$, $\pi_{2k}(S^{2k}) = \mathbb{Z}$ and the extra $\mathbb{Z}$ in $\pi_{4k-1}(S^{2k})$ (for S^2 this is $\pi_3(S^2) = \mathbb{Z}$, generated by the Hopf map [$S^3 \to S^2$, sending a unit vector $(z_1, z_2) \in \mathbb{C}^2$ to the ratio $z_1/z_2 \in \mathbb{C} \cup \{\infty\}$]). Rationally, an odd sphere looks like it has a single hole, in its own dimension.

Check against Table 8.1: the row of $\mathrm{SU}(2) = S^3$ reads $0, 0, \mathbb{Z}, \mathbb{Z}_2, \mathbb{Z}_2, \mathbb{Z}_{12}, \mathbb{Z}_2$, and only π_3 survives $\otimes\mathbb{Q}$.

8.1.22 The unitary group is rationally a product of odd spheres

Nash's (8.25) is

$$\mathrm{U}(N) \sim S^1 \times S^3 \times S^5 \times \cdots \times S^{2N-1}. \tag{8.83}$$

It is not true as an equivalence of spaces: $\mathrm{U}(2) \cong S^1 \times S^3$ is true, but $\mathrm{SU}(3)$ is not $S^3 \times S^5$ (it is a non-trivial S^3 bundle over S^5; compare $\pi_4(\mathrm{SU}(3)) = 0$ with $\pi_4(S^3 \times S^5) = \mathbb{Z}_2$). It is true after $\otimes\mathbb{Q}$. Here is where it comes from.

The fibration tower

$\mathrm{U}(N)$ acts on $\mathbb{C}^N$ preserving lengths, so it acts on the unit sphere $S^{2N-1} \subset \mathbb{C}^N$. The action is transitive: any unit vector can be rotated to $e_N = (0, \ldots, 0, 1)$. The stabiliser of e_N is the set of unitary matrices of the form $\left(\begin{smallmatrix} u & 0 \\ 0 & 1 \end{smallmatrix}\right)$ with $u \in \mathrm{U}(N-1)$. So the map "last column"

$$\mathrm{U}(N-1) \longrightarrow \mathrm{U}(N) \xrightarrow{u \mapsto u e_N} S^{2N-1} \tag{8.84}$$

is a fibre bundle. For $N = 1$ it is $\mathrm{U}(1) = S^1$. For $N = 2$: a matrix in $\mathrm{U}(2)$ is determined by its last column $(b, d)^T$ (a unit vector in $\mathbb{C}^2$, i.e. a point of S^3) and its determinant $e^{i\phi}$:

$$u = \begin{pmatrix} e^{i\phi}\bar{d} & b \\ -e^{i\phi}\bar{b} & d \end{pmatrix}, \qquad \det u = e^{i\phi}(|d|^2 + |b|^2) = e^{i\phi}.$$

So $\mathrm{U}(2) \cong S^1 \times S^3$ exactly.

For larger N the long exact sequence of (8.84) (Eq. (8.38)) tensored with $\mathbb{Q}$ reads, in the only interesting spot,

$$\cdots \to \pi_i(\mathrm{U}(N-1)) \otimes \mathbb{Q} \to \pi_i(\mathrm{U}(N)) \otimes \mathbb{Q} \to \pi_i(S^{2N-1}) \otimes \mathbb{Q} \to \pi_{i-1}(\mathrm{U}(N-1)) \otimes \mathbb{Q} \to \cdots$$

By induction $\pi_*(\mathrm{U}(N-1)) \otimes \mathbb{Q}$ is $\mathbb{Q}$ in degrees $1, 3, \ldots, 2N-3$ and zero otherwise. The sphere contributes one $\mathbb{Q}$ in degree $2N-1$. The connecting map $\pi_{2N-1}(S^{2N-1}) \otimes \mathbb{Q} \to \pi_{2N-2}(\mathrm{U}(N-1)) \otimes \mathbb{Q}$ goes into an even degree, where the group is zero; so it vanishes and the sequence splits. Hence

$$\pi_i(\mathrm{U}(N)) \otimes \mathbb{Q} = \begin{cases} \mathbb{Q}, & i = 1, 3, 5, \ldots, 2N-1 \\ 0, & \text{otherwise,} \end{cases} \tag{8.85}$$

which is also the rational homotopy of $S^1 \times S^3 \times \cdots \times S^{2N-1}$ (homotopy groups of products are products). That is (8.83). For $\mathrm{SU}(N)$ remove the S^1: $\mathrm{SU}(N) \sim S^3 \times \cdots \times S^{2N-1}$.

The de Rham cohomology of $\mathrm{U}(N)$

Rationally equivalent spaces have the same real cohomology, and the Künneth formula says the cohomology of a product is the tensor product. So (Nash's (8.26))

$$H^*_{\mathrm{dR}}(\mathrm{U}(N)) \cong H^*_{\mathrm{dR}}(S^1) \otimes H^*_{\mathrm{dR}}(S^3) \otimes \cdots \otimes H^*_{\mathrm{dR}}(S^{2N-1}) = \Lambda(\omega_1, \omega_3, \ldots, \omega_{2N-1}). \tag{8.86}$$

Each $H^*_{\mathrm{dR}}(S^{2k-1})$ is $\mathbb{R}$ in degrees 0 and $2k-1$, spanned by 1 and a volume form. So the whole ring is the **exterior algebra** on one generator ω_{2k-1} in each odd degree: products of distinct generators, with $\omega_i \wedge \omega_i = 0$ and $\omega_i \wedge \omega_j = -\omega_j \wedge \omega_i$ (both odd). Its dimension is 2^N and its Poincaré polynomial is

$$P_{\mathrm{U}(N)}(t) := \sum_p \dim H^p_{\mathrm{dR}}\, t^p = (1+t)(1+t^3)(1+t^5)\cdots(1+t^{2N-1}).$$

For $\mathrm{U}(2)$: $1 + t + t^3 + t^4$, i.e. $b_0 = b_1 = b_3 = b_4 = 1$, which matches $S^1 \times S^3$. Setting $t = 1$ gives 2^N; setting $t = -1$ gives 0, the Euler characteristic of a compact Lie group (it has a nowhere-vanishing vector field).

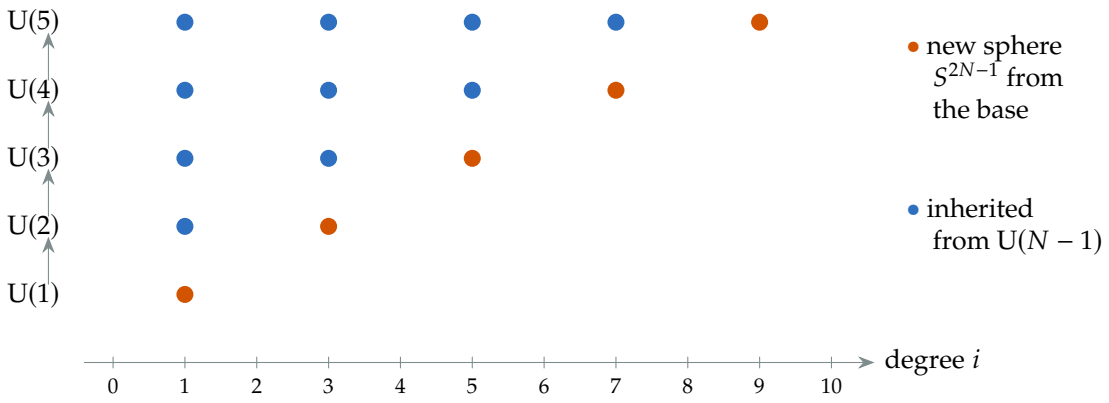


Figure 8.15: Rational homotopy of U(N). Each row lists the degrees i with $\pi_i(\mathrm{U}(N)) \otimes \mathbb{Q} = \mathbb{Q}$. Going up one row (from U($N-1$) to U(N)) the fibration $\mathrm{U}(N-1) \to \mathrm{U}(N) \to S^{2N-1}$ adds exactly one new odd degree, $2N-1$ (orange). All generators sit in odd degrees; that is why nothing can cancel and U(N) looks rationally like $S^1 \times S^3 \times \cdots \times S^{2N-1}$.

Working definition 8.13 (explicit generators ω_{2k-1} on a matrix group)

Let $\theta = g^{-1}\mathrm{d}g$ be the Maurer–Cartan form on $G \subset \mathrm{U}(N)$. The forms

$$\omega_{2k-1} := C_k \operatorname{tr}\!\left(\theta^{2k-1}\right), \qquad k = 1, \dots, N, \tag{8.87}$$

are closed, bi-invariant [unchanged under $g \mapsto hg$ and under $g \mapsto gh$ for any fixed $h \in G$], and represent the generators of (8.86). The constants C_k fix the normalisation; the natural choice, which comes out of the transgression in Section 8.2.5, is

$$C_1 = \frac{i}{2\pi}, \qquad C_2 = \frac{1}{24\pi^2}, \qquad C_3 = -\frac{i}{240\pi^3}.$$

With it, ω_3 is the winding density (8.64). *To test a map* $f : S^{2k-1} \to G$ for non-triviality in $\pi_{2k-1}(G) \otimes \mathbb{Q}$, compute $\int_{S^{2k-1}} f^*\omega_{2k-1}$: it is non-zero exactly when f is rationally non-trivial [no multiple $m[f]$ with $m \neq 0$ is the trivial class, i.e. $[f]$ has infinite order].

Why are the ω_{2k-1} closed? The Maurer–Cartan equation $\mathrm{d}\theta = -\theta^2$ gives

$$\mathrm{d}\operatorname{tr}\theta^m \overset{(1)}{=} \sum_{r=0}^{m-1} (-1)^r \operatorname{tr}\!\left(\theta^r\, \mathrm{d}\theta\, \theta^{m-1-r}\right) \overset{(2)}{=} -\sum_{r=0}^{m-1} (-1)^r \operatorname{tr}\!\left(\theta^{m+1}\right) \overset{(3)}{=} \begin{cases} -\operatorname{tr}\theta^{m+1} = 0, & m \text{ odd} \\ 0, & m \text{ even.} \end{cases}$$

(1) Graded product rule: passing d across r 1-forms costs $(-1)^r$.
(2) $\mathrm{d}\theta = -\theta^2$.
(3) For m odd the alternating sum of m terms is 1, and $\operatorname{tr}\theta^{m+1}$ with $m+1$ even vanishes: cycling one 1-form past the other m (odd degree) costs $(-1)^m = -1$, so $\operatorname{tr}\theta^{m+1} = -\operatorname{tr}\theta^{m+1}$. For m even the alternating sum is 0.

The same cycling argument shows that $\operatorname{tr}\theta^{2k} = 0$: even powers are useless, which is another way to see that only odd generators appear.

Worked example 8.22: The generators of U(1) and U(2) explicitly

U(1). With $g = e^{i\phi}$, $\theta = i\,\mathrm{d}\phi$ and $\omega_1 = \frac{i}{2\pi}\, i\,\mathrm{d}\phi = -\frac{1}{2\pi}\mathrm{d}\phi$. Its integral over the circle is -1. (The sign is the price of the convention $c_1 = \frac{i}{2\pi}F$; we could use $-\omega_1$ as the generator equally well.)

U(2). Write $u = e^{i\phi}h$ with h in SU(2) is not quite possible globally (the pair (ϕ, h) and $(\phi + \pi, -h)$ give the same u), but locally it is fine and $\theta_u = i\,\mathrm{d}\phi\,\mathbb{1} + \theta_h$. Then

$$\omega_1(u) \overset{(1)}{=} \frac{i}{2\pi}\big(2i\,\mathrm{d}\phi + \operatorname{tr}\theta_h\big) \overset{(2)}{=} -\frac{1}{\pi}\mathrm{d}\phi, \qquad \omega_3(u) \overset{(3)}{=} \frac{1}{24\pi^2}\operatorname{tr}\theta_h^3.$$

(1) $\operatorname{tr}\mathbb{1} = 2$.
(2) $\operatorname{tr}\theta_h = \operatorname{tr}(h^{-1}\mathrm{d}h) = \mathrm{d}\log\det h = 0$ for $h \in \mathrm{SU}(2)$.
(3) Expand $(i\,\mathrm{d}\phi\,\mathbb{1} + \theta_h)^3$; every term with two or more $\mathrm{d}\phi$ vanishes, and the terms with one $\mathrm{d}\phi$ give $3i\,\mathrm{d}\phi\operatorname{tr}\theta_h^2 = 0$ (tr of an even power).

So ω_1 measures the phase winding and ω_3 the SU(2) winding, and the ring is $\Lambda(\omega_1, \omega_3)$ with top class $\omega_1 \wedge \omega_3$ in degree $4 = \dim \mathrm{U}(2)$.

8.1.23 The evaluation map and integration over S^n

We now transport the generators ω_i of $H^*_{\mathrm{dR}}(G)$ to the gauge group $\mathcal{G}$. Recall that here $\mathcal{G}$ is the based gauge group of a bundle over $M = S^n$, which is homotopy equivalent to $\mathrm{Map}_0(S^n, G) = \Omega^n G$ (Section 8.1.13), and that Nash takes n even. The tool is the **evaluation map** (Nash's (8.27)):

$$\mathrm{Ev} : S^n \times \mathcal{G} \longrightarrow G, \qquad (x, \alpha) \longmapsto \alpha(x). \tag{8.88}$$

It is the most natural map there is: a point of $\mathcal{G}$ is a map $\alpha : S^n \to G$, and we evaluate it. Pull the generator back (Nash's (8.28)): $\mathrm{Ev}^*\omega_i$ is a closed i-form on the product $S^n \times \mathcal{G}$. Then "divide out" the S^n by integrating over it (Nash's (8.29)):

$$\mu_{i-n} := \int_{S^n} \mathrm{Ev}^*\omega_i \;\in\; H^{i-n}_{\mathrm{dR}}(\mathcal{G}). \tag{8.89}$$

Integration over the fibre

What does $\int_{S^n}$ of a form on $S^n \times Y$ mean? Use coordinates x on S^n and y on Y. Every form on the product is a sum of pieces of **bidegree** (p, q): p legs along S^n, q along Y. We always write the S^n legs first:

$$\beta = \sum f_{IJ}(x, y)\,\mathrm{d}x^I \wedge \mathrm{d}y^J, \qquad |I| = p,\ |J| = q.$$

Working definition 8.14 (integration over the fibre S^n)

For β of total degree r [$p + q = r$ for every piece] on $S^n \times Y$, keep only the part with $p = n$ and integrate out x:

$$\int_{S^n} \beta := \sum_{|J| = r-n} \Big(\int_{S^n} f_{IJ}(x, y)\,\mathrm{d}x^I\Big)\,\mathrm{d}y^J \qquad (|I| = n).$$

The result is an $(r - n)$-form on Y. The pieces with $p < n$ are thrown away (they cannot be integrated over an n-dimensional space).

The one property we need: *integration over a closed fibre commutes with* d (up to a sign), so it maps closed forms to closed forms and exact to exact, and so is well defined on cohomology. Write $\mathrm{d} = \mathrm{d}_x + \mathrm{d}_y$. Then

$$\mathrm{d}_Y \int_{S^n} \beta \overset{(1)}{=} \int_{S^n} (-1)^n\,\mathrm{d}_y\beta \overset{(2)}{=} (-1)^n \int_{S^n} (\mathrm{d}\beta - \mathrm{d}_x\beta) \overset{(3)}{=} (-1)^n \int_{S^n} \mathrm{d}\beta. \tag{8.90}$$

(1) d_y acts on the coefficient functions and on the $\mathrm{d}y^J$; to bring a new $\mathrm{d}y^k$ to the right of the n legs $\mathrm{d}x^I$ costs $(-1)^n$. Differentiation under the integral sign in x is harmless.
(2) $\mathrm{d}_y = \mathrm{d} - \mathrm{d}_x$.
(3) The $(n, *)$ part of $\mathrm{d}_x\beta$ comes from the $(n-1, *)$ part of β and is exact along S^n; its integral over the closed S^n is zero by Stokes ($\partial S^n = \varnothing$).

So $\mathrm{d}\beta = 0$ implies $\mathrm{d}\int_{S^n}\beta = 0$, and $\beta = \mathrm{d}\gamma$ implies $\int\beta = \pm\mathrm{d}\int\gamma$.

Worked example 8.23: Integration over the fibre on a torus

Let $S^1 \times S^1$ have angles x (the fibre) and y. For $\beta = \cos^2 x\,\mathrm{d}x \wedge \mathrm{d}y + \sin y\,\mathrm{d}y$,

$$\int_{S^1_x}\beta \overset{(1)}{=} \Big(\int_0^{2\pi}\cos^2 x\,\mathrm{d}x\Big)\mathrm{d}y \overset{(2)}{=} \pi\,\mathrm{d}y.$$

(1) Only the $(1,1)$ piece has a full fibre leg; the $(0,1)$ piece $\sin y\,\mathrm{d}y$ is discarded.
(2) $\int_0^{2\pi}\cos^2 x\,\mathrm{d}x = \pi$.

β is closed (a top form plus $\mathrm{d}(-\cos y)$), and the result $\pi\,\mathrm{d}y$ is closed too, as (8.90) promises.

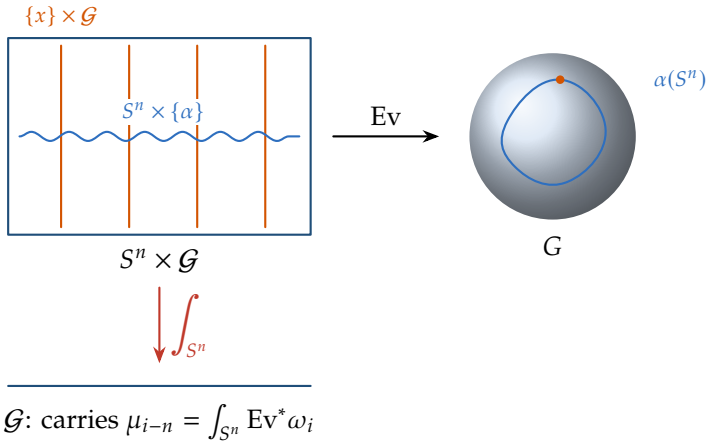


Figure 8.16: The evaluation map. A horizontal slice $S^n \times \{\alpha\}$ of the product is mapped by Ev onto the image $\alpha(S^n)$ of the gauge transformation α; a vertical line $\{x\} \times \mathcal{G}$ records the values of all gauge transformations at one point x. Pulling back an i-form on G and integrating out the n horizontal directions leaves an $(i-n)$-form on $\mathcal{G}$.

What μ_{i-n} measures

Why should μ_{i-n} be non-trivial? Take a family of gauge transformations parametrised by a sphere, $f : S^m \to \mathcal{G}$, $s \mapsto \alpha_s$, with $m = i - n$. By the adjunction of Section 8.1.13 it is the same thing as a single map $F : S^m \times S^n \to G$, $F(s,x) = \alpha_s(x)$, which sends $S^m \times \{x_0\} \cup \{s_0\} \times S^n$ to $\mathbb{1}$ and hence factors through $S^m \wedge S^n = S^{m+n}$. Then

$$\int_{S^m} f^*\mu_m \overset{(1)}{=} \int_{S^m}\int_{S^n}(\mathrm{id}\times f)^*\mathrm{Ev}^*\omega_{m+n} \overset{(2)}{=} \int_{S^n\times S^m} F^*\omega_{m+n} \overset{(3)}{=} \int_{S^{m+n}}\bar{F}^*\omega_{m+n}. \tag{8.91}$$

(1) Definition (8.89); pulling back along f commutes with integrating over the fibre S^n (a change of variables in y only).
(2) Fubini: integrating first over S^n and then over S^m is integrating over the product; $\mathrm{Ev} \circ (\mathrm{id} \times f) = F$.

(3) F factors through the quotient map $S^n \times S^m \to S^{m+n}$, which has degree one, as $F = \bar{F} \circ q$.

The right-hand side is the test of Eq. (8.87) for the element $[\bar{F}] \in \pi_{m+n}(G) = \pi_m(\mathcal{G})$. So μ_m pairs with the rational homotopy of $\mathcal{G}$ exactly as ω_{m+n} pairs with that of G. *Integration over S^n is the cohomological shadow of the shift $\pi_m(\mathcal{G}) = \pi_{m+n}(G)$.*

Worked example 8.24: $n = 1$ and $n = 3$: μ_0 counts windings

Although Nash takes n even, odd n is the best warm-up.
$n = 1$, $G = \mathrm{U}(1)$. $\mathcal{G} = \Omega\mathrm{U}(1)$, the based loops $\alpha : S^1 \to \mathrm{U}(1)$. The only generator is ω_1, so the only μ is

$$\mu_0(\alpha) \overset{(1)}{=} \int_{S^1} \alpha^* \omega_1 \overset{(2)}{=} \frac{i}{2\pi} \oint \alpha^{-1} \mathrm{d}\alpha \overset{(3)}{=} -\frac{1}{2\pi} \oint \mathrm{d}\lambda = -\operatorname{wind}(\alpha),$$

(1) A 0-form on $\mathcal{G}$ is a function; at a point α of $\mathcal{G}$, the slice $S^1 \times \{\alpha\}$ is mapped by α itself.
(2) (8.87) with $C_1 = i/2\pi$.
(3) Write $\alpha = e^{i\lambda(x)}$ locally; $\alpha^{-1}\mathrm{d}\alpha = i\,\mathrm{d}\lambda$.

A closed 0-form is a locally constant function: μ_0 labels the components of $\mathcal{G}$ by the winding number, and $H^0_{\mathrm{dR}}(\mathcal{G})$ is the space of functions on $\pi_0(\mathcal{G}) = \mathbb{Z}$.
$n = 3$, $G = \mathrm{SU}(2)$. Likewise $\mu_0 = \int_{S^3} \alpha^* \omega_3 = w(\alpha)$, the winding number (8.64). This is the label of large gauge transformations in Section 8.1.19, now seen as the simplest of Nash's cohomology classes.

8.1.24 H-spaces, Hopf algebras, and the cohomology of $\mathcal{G}$

Nash next argues that the μ's *generate* the whole cohomology ring, with no relations beyond those of an exterior algebra. The argument has two ingredients: a structure theorem for the cohomology of group-like spaces, and the rational homotopy of $\mathcal{G}$.

H-spaces and Hopf algebras

An **H-space** is a space X with a continuous multiplication $X \times X \to X$ that has a unit up to homotopy. Every topological group is an H-space. So is $\mathcal{G}$ (pointwise multiplication of gauge transformations), and so is every loop space ΩY (concatenation of loops; the constant loop is a unit up to reparametrisation). The multiplication $m : X \times X \to X$ gives, on cohomology, a **coproduct**

$$m^* : H^*(X) \to H^*(X \times X) = H^*(X) \otimes H^*(X).$$

Together with the cup (wedge) product this makes $H^*(X; \mathbb{Q})$ a **Hopf algebra**. A class x is **primitive** if $m^* x = x \otimes 1 + 1 \otimes x$: it behaves additively under multiplication, like a logarithm.

Worked example 8.25: The winding forms are primitive

The winding number is additive, $w(gh) = w(g) + w(h)$ (8.68). At the level of forms, the computation in that proof says $m^*\omega_3 = \omega_3 \otimes 1 + 1 \otimes \omega_3 + \mathrm{d}(\ldots)$. So ω_3 is primitive. The same holds for every ω_{2k-1} and for the μ's. The products $\omega_1 \wedge \omega_3$ and so on are not primitive.

Key idea: Hopf's theorem (as Nash quotes it, with Whitehead)

For a connected H-space X of finite type, the rational cohomology is the free graded-commutative algebra on the dual of the rational homotopy:

$$H^*(X;\mathbb{Q}) \cong S(A) \otimes \Lambda(A), \qquad A^i := (\pi_i(X) \otimes \mathbb{Q})^*, \tag{8.92}$$

where $S(A) = \bigotimes_i S(A^{2i})$ is the polynomial algebra on the even-degree generators and $\Lambda(A) = \bigotimes_i \Lambda(A^{2i-1})$ is the exterior algebra on the odd-degree ones.

Why is it plausible? For an H-space the rational Hurewicz map is injective onto the primitives, so each rational homotopy class of degree i gives one primitive cohomology generator of degree i. Generators of odd degree anticommute with themselves ($x \wedge x = -x \wedge x$), so their squares vanish: exterior. Generators of even degree commute with themselves, and in an H-space nothing forces their powers to vanish: polynomial. The theorem says there are no other relations.

Worked example 8.26: An exterior and a polynomial case

- $X = \mathrm{U}(N)$: $\pi_* \otimes \mathbb{Q}$ in degrees $1, 3, \dots, 2N-1$, all odd, so $H^* = \Lambda(\omega_1, \dots, \omega_{2N-1})$, as in (8.86). The ring is finite dimensional, as it must be for a compact manifold.
- $X = \Omega S^3$ (based loops in SU(2)): $\pi_i(\Omega S^3) = \pi_{i+1}(S^3)$, so the only rational class is in degree 2. Hence $H^*(\Omega S^3;\mathbb{Q}) = \mathbb{Q}[y]$ with $\deg y = 2$: one class in *every* even degree, $1, y, y^2, \dots$. This infinite ladder is exactly what the Morse theory of geodesics on S^3 produces (Chapter 7): the k-th family of closed geodesics has index $2k$.
- $X = \mathbb{CP}^\infty = B\mathrm{U}(1)$: $\pi_2 = \mathbb{Z}$ only; $H^* = \mathbb{Q}[c_1]$, the polynomial ring on the first Chern class. (Every $K(\mathbb{Z},2)$ is an H-space.)

Applying Hopf to $\mathcal{G}$

Now Nash's (8.31)–(8.33). We have $\mathcal{G} \simeq \Omega^n G$ and $G \sim S^1 \times S^3 \times \cdots \times S^{2N-1}$. Loop spaces of products are products of loop spaces, so

$$\pi_i(\mathcal{G}) \otimes \mathbb{Q} \overset{(1)}{=} \pi_{n+i}(G) \otimes \mathbb{Q} \overset{(2)}{=} \bigoplus_{j=1}^{N} \pi_{n+i}(S^{2j-1}) \otimes \mathbb{Q} \overset{(3)}{=} \bigoplus_{j=1}^{N} H^{n+i}(S^{2j-1};\mathbb{Q}). \tag{8.93}$$

(1) $\pi_i(\Omega^n G) = \pi_{n+i}(G)$, (8.56).
(2) (8.83): rationally G is the product of odd spheres, and π_* of a product is the direct sum.
(3) Serre: $\pi_k(S^{2j-1}) \otimes \mathbb{Q}$ is $\mathbb{Q}$ exactly when $k = 2j-1$, which is also exactly when $H^k(S^{2j-1};\mathbb{Q}) = \mathbb{Q}$ ($k > 0$).

So $\pi_i(\mathcal{G}) \otimes \mathbb{Q} \neq 0$ only when $n + i = 2j - 1$ for some $j \leq N$:

$$i = 2j - n - 1, \qquad j = 1, \dots, N, \qquad i \geq 0. \tag{8.94}$$

For n even, i is *odd*. There are no even-degree generators, so the symmetric factor $S(A)$ in (8.92) is trivial and

$$H^*_{\mathrm{dR}}(\mathcal{G}) \cong \Lambda\big(\mu_{2j-1-n} \,:\, 2j-1 \geq n+1,\ j \leq N\big). \tag{8.95}$$

By (8.91) each μ_{2j-1-n} pairs non-trivially with the rational generator of $\pi_{2j-1-n}(\mathcal{G})$, so the μ's built by the evaluation map are precisely a set of generators. This is Nash's conclusion below (8.33).

Watch out: Components of $\mathcal{G}$

Hopf's theorem is for a *connected* H-space, but $\mathcal{G}$ has $\pi_0(\mathcal{G}) = \pi_n(G)$ components. For n even, $\pi_n(G) \otimes \mathbb{Q} = 0$, so there are only finitely many components, all homotopy equivalent (multiply by a fixed element of each component). Apply the theorem to the identity component and take one copy per component; in degree 0 this gives the functions on $\pi_0(\mathcal{G})$. For odd n the case $i = 0$ in (8.94) occurs, and μ_0 counts the infinitely many components (Example 8.24).

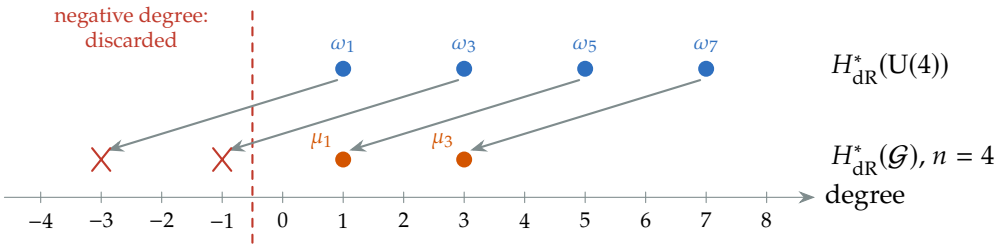


Figure 8.17: Degree shift by integration over S^4, for $G = \mathrm{U}(4)$. Each generator ω_i of the group (blue, degree i) produces μ_{i-4} on the gauge group (orange). Those that would land in negative degree (ω_1, ω_3) give nothing. What survives, μ_1 and μ_3, generates $H^*_{\mathrm{dR}}(\mathcal{G}) = \Lambda(\mu_1, \mu_3)$: Nash's table (8.38).

8.1.25 The cohomology of the configuration space $B\mathcal{G}$

The same method works for $B\mathcal{G} = \mathcal{A}/\mathcal{G}$. Now we use the homotopy equivalence (Nash's (8.34), our (8.53))

$$B\mathcal{G} \simeq \mathrm{Map}_0(S^n, BG),$$

a point of which is (the classifying map of) a G-bundle on S^n together with a connection, up to gauge. The evaluation map is now (Nash's (8.35))

$$\mathrm{Ev} : S^n \times \mathrm{Map}_0(S^n, BG) \longrightarrow BG, \qquad (x, \alpha) \mapsto \alpha(x). \tag{8.96}$$

The cohomology of BG is well known. For $G = \mathrm{U}(N)$ it is the polynomial ring on the universal Chern classes,

$$H^*(B\mathrm{U}(N); \mathbb{Z}) = \mathbb{Z}[c_1, c_2, \ldots, c_N], \qquad c_i \in H^{2i}(B\mathrm{U}(N)).$$

(For $N = 1$ this is $H^*(\mathbb{CP}^\infty) = \mathbb{Z}[c_1]$ of Example 8.26.) "Universal" means: for a bundle $P \to X$ with classifying map $f : X \to BG$, $c_i(P) = f^* c_i$. So we set (Nash's (8.36))

$$\nu_{2i-n} := \int_{S^n} \mathrm{Ev}^* c_i \ \in \ H^{2i-n}_{\mathrm{dR}}(B\mathcal{G}). \tag{8.97}$$

With $n = 2m$ (Nash's (8.37)),

$$\mu_{i-2m} \in H^{i-2m}_{\mathrm{dR}}(\mathcal{G}) \ (i \text{ odd}), \qquad \nu_{2i-2m} \in H^{2i-2m}_{\mathrm{dR}}(B\mathcal{G}). \tag{8.98}$$

The generators of $H^*_{\mathrm{dR}}(\mathcal{G})$ are all of odd degree, those of $H^*_{\mathrm{dR}}(B\mathcal{G})$ all of even degree. This matches the rational homotopy: $\pi_k(B\mathcal{G}) \otimes \mathbb{Q} = \pi_{k-1}(\mathcal{G}) \otimes \mathbb{Q}$ is non-zero only for $k - 1$ odd. For $n \geq 2$, $B\mathcal{G} \simeq \Omega^{n-1} G$ is itself an H-space, so Hopf's theorem applies and gives a *polynomial* ring:

$$H^*_{\mathrm{dR}}(B\mathcal{G}) \cong \mathbb{R}\big[\nu_{2i-n} \ : \ 2i > n, \ i \leq N\big] \quad \text{(on each component)}. \tag{8.99}$$

Worked example 8.27: ν_0 is the instanton number

Take $n = 4$, $G = \mathrm{SU}(N)$ and $i = 2$. Then $\nu_0 = \int_{S^4} \mathrm{Ev}^* c_2$ is a function on $\mathrm{Map}(S^4, BG)$. At a point α (a bundle P with connection A),

$$\nu_0(\alpha) \overset{(1)}{=} \int_{S^4} \alpha^* c_2 \overset{(2)}{=} \int_{S^4} c_2(P) \overset{(3)}{=} \frac{1}{8\pi^2} \int_{S^4} \operatorname{tr} F \wedge F \overset{(4)}{=} -k.$$

(1) On the slice $S^4 \times \{\alpha\}$ the evaluation map is α.
(2) Universality: $\alpha^* c_2 = c_2$ of the bundle classified by α.
(3) Chern–Weil representative of c_2 for $\mathrm{SU}(N)$ (Eq. (8.10)).
(4) Definition of the instanton number.

So ν_0 is locally constant and labels the components $B\mathcal{G}_k$ of the space of all bundles-with-connection by their instanton number. In the text we mostly work on one component, where ν_0 is a constant. (Also $\nu_{-2} = \int_{S^4} c_1 = 0$: an S^4 has no 2-cycles.)

Worked example 8.28: $n = 2$, $G = \mathrm{SU}(2)$: a polynomial ring

Two-dimensional Yang–Mills on S^2. Here $B\mathcal{G} \simeq \Omega S^3$, and the only Chern class is c_2, giving $\nu_{4-2} = \nu_2 = \int_{S^2} \mathrm{Ev}^* c_2 \in H^2(B\mathcal{G})$. By Example 8.26,

$$H^*_{\mathrm{dR}}(\mathcal{A}/\mathcal{G}) = \mathbb{R}[\nu_2] : \quad \mathbb{R} \text{ in every even degree},$$

spanned by $1, \nu_2, \nu_2^2, \ldots$ This is the infinite-dimensional analogue of $\mathbb{CP}^\infty$. (Atiyah and Bott used exactly this ring, via Morse theory of the Yang–Mills functional, to study moduli of flat connections on Riemann surfaces.) Meanwhile $\mathcal{G} \simeq \Omega^2 S^3$ has $\pi_1(\mathcal{G}) = \pi_3(S^3) = \mathbb{Z}$ and $H^*_{\mathrm{dR}}(\mathcal{G}) = \Lambda(\mu_1)$ with $\mu_1 = \int_{S^2} \mathrm{Ev}^* \omega_3$.

Nash's table (8.38): how the answer grows with N

Fix $n = 4$ ($M = S^4$). The group $\mathrm{U}(N)$ has generators $\omega_1, \omega_3, \ldots, \omega_{2N-1}$, giving μ_{i-4} for $i \geq 5$, i.e. $\mu_1, \mu_3, \ldots, \mu_{2N-5}$. The Chern classes $c_1, \ldots, c_N$ give ν_{2i-4} for $i \geq 3$, i.e. $\nu_2, \nu_4, \ldots, \nu_{2N-4}$.

G	generators of $H^*_{\mathrm{dR}}(G)$	generators of $H^*_{\mathrm{dR}}(\mathcal{G})$	$P_{\mathcal{G}}(t)$	generators of $H^*_{\mathrm{dR}}(B\mathcal{G})$
$\mathrm{U}(1)$	ω_1	none	1	none
$\mathrm{U}(2)$	ω_1, ω_3	none	1	none
$\mathrm{U}(3)$	$\omega_1, \omega_3, \omega_5$	μ_1	$1 + t$	ν_2
$\mathrm{U}(4)$	$\omega_1, \ldots, \omega_7$	μ_1, μ_3	$(1 + t)(1 + t^3)$	ν_2, ν_4
$\mathrm{U}(N)$	$\omega_1, \ldots, \omega_{2N-1}$	$\mu_1, \ldots, \mu_{2N-5}$	$\prod_{j=3}^{N}(1 + t^{2j-5})$	$\nu_2, \ldots, \nu_{2N-4}$

Table 8.2: Real cohomology of the (based) gauge group and of the configuration space for $M = S^4$ (Nash's (8.38) completed). U(3) is the first group with any free cohomology in positive degree. Nothing changes if $\mathrm{U}(N)$ is replaced by $\mathrm{SU}(N)$, because the lost generator ω_1 gives nothing on S^4 anyway.

Watch out: "has just" in Nash's table

Nash writes that U(4) "has just $H^1(\mathcal{G}) \neq 0$, $H^3(\mathcal{G}) \neq 0$". He means *generators*. The ring also contains the product $\mu_1 \wedge \mu_3$, so $H^4(\mathcal{G}) \neq 0$ as well. For $B\mathcal{G}$ there are all monomials $\nu_2^a \nu_4^b$, so $H^{2k}(B\mathcal{G}) \neq 0$ for every k once $N \geq 3$.

Three physical remarks.

- SU(2) *on S^4 has no free cohomology at all* (beyond degree 0). The only topology of its gauge group visible to de Rham forms is nothing; the $\mathbb{Z}_2 = \pi_0(\mathcal{G})$ of Example 8.15 is pure torsion. So the perturbative (local) non-Abelian anomaly cannot occur for SU(2); only the global Witten anomaly can. This matches the group theory: the anomaly coefficient $\operatorname{tr} T^a\{T^b, T^c\}$ vanishes for SU(2) (all its representations are pseudo-real).

- SU(3) *(colour) is the first case with $H^1(\mathcal{G}) \neq 0$.* The class μ_1 is the one that represents the four-dimensional non-Abelian anomaly; its partner $\nu_2 \in H^2(\mathcal{A}/\mathcal{G})$ is the first Chern class of the determinant line bundle of the chiral Dirac operator (Chapter 10).

- *The integral in* (8.91) *is an integer multiple of a fixed number.* For SU(3) and the normalisation of Eq. (8.87), a generator of $\pi_5(\mathrm{SU}(3)) = \mathbb{Z}$ gives $\int_{S^5} \omega_5 = \pm 2$, not ± 1 (Bott: the c_3-normalised form is divisible by $(3-1)!$ on spheres). The rational statements do not care.

Reading Nash (pp. 220–223, (8.25)–(8.38))

Nash's "$G \sim S^1 \times \cdots \times S^{2N-1}$" is a rational homotopy equivalence, not a homeomorphism; we derived it from the fibrations (8.84). His (8.27)–(8.29) are our (8.88) and (8.89); the reason the μ's generate is the pairing (8.91), which he leaves implicit. The Hopf–Whitehead structure theorem (8.30) is our (8.92); he writes it for a group $\hat{\mathcal{G}}$ and then applies it to $\mathcal{G}$. His (8.32) contains the isomorphism $\pi_{n+i}(S^{2j-1}) \otimes \mathbb{Q} \cong H^{n+i}(S^{2j-1}; \mathbb{Q})$, which is Serre's theorem for odd spheres. The last sentence of §1 ("no sensitivity to the difference between U(N) and SU(N)") is the remark under Table 8.2.

Recap: cohomology of $\mathcal{G}$ and $B\mathcal{G}$

- De Rham forms see the free part of cohomology, i.e. rational homotopy; torsion (such as Witten's $\mathbb{Z}_2$) needs other tools.
- $\mathrm{U}(N) \sim S^1 \times S^3 \times \cdots \times S^{2N-1}$, so $H^*_{\mathrm{dR}}(\mathrm{U}(N)) = \Lambda(\omega_1, \ldots, \omega_{2N-1})$ with $\omega_{2k-1} \propto \operatorname{tr}(g^{-1}\mathrm{d}g)^{2k-1}$.
- $\mu_{i-n} = \int_{S^n} \mathrm{Ev}^* \omega_i$ and $\nu_{2i-n} = \int_{S^n} \mathrm{Ev}^* c_i$; integration over S^n lowers degrees by n, mirroring $\pi_m(\mathcal{G}) = \pi_{m+n}(G)$.
- For n even: $H^*_{\mathrm{dR}}(\mathcal{G})$ is exterior on odd μ's, $H^*_{\mathrm{dR}}(B\mathcal{G})$ polynomial on even ν's. On S^4: μ_1, ν_2 appear first for U(3).

Summary of Section 8.1

Recap: the topology of gauge fields

- $\mathcal{A}$ is affine, hence contractible; the based gauge group $\mathcal{G}$ acts on it freely (parallel transport commutes with gauge transformations); the holonomy group measures the stabiliser of the full group.
- $\mathcal{G} \to \mathcal{A} \to \mathcal{A}/\mathcal{G}$ is a principal bundle with contractible total space, so $\mathcal{A}/\mathcal{G} = B\mathcal{G}$ and $\pi_k(\mathcal{A}/\mathcal{G}) = \pi_{k-1}(\mathcal{G})$. It is not trivial: no global gauge fixing exists (Gribov).
- On S^n: bundles are clutched by $S^{n-1} \to G$; $\mathcal{G} \simeq \Omega^n G$, $B\mathcal{G} \simeq \Omega^{n-1} G \simeq \mathrm{Map}_0(S^n, BG)$; homotopy groups shift by n.
- On $\mathbb{R}^4$: finite action $\Rightarrow A \to g^{-1}\mathrm{d}g$ at infinity, and the instanton number is the winding $\frac{1}{24\pi^2}\int \mathrm{tr}(g^{-1}\mathrm{d}g)^3$. On $\mathbb{R}^3$: large gauge transformations, N_{CS} shifts by integers, θ-vacua.
- Cohomology: $H^*_{\mathrm{dR}}(\mathcal{G}) = \Lambda(\mu)$, $H^*_{\mathrm{dR}}(B\mathcal{G}) = \mathbb{R}[\nu]$, built by evaluation and integration over S^n.

8.2 Secondary characteristic classes

At the end of Section 8.1 we had the cohomology of the gauge group $\mathcal{G}$ and of the configuration space $B\mathcal{G} = \mathcal{A}/\mathcal{G}$, described by generators μ_{i-n} and ν_{2i-n}. They were built by an abstract recipe: evaluate, pull back a class from G or BG, integrate over S^n. A physicist wants more. Which *differential form*, built from the gauge potential, represents μ_1? How would we write it in a Lagrangian, or recognise it in a Feynman diagram calculation? Nash's Section 2 answers this. The answer is the **Chern–Simons form**, and the construction behind it is Chern's theory of **secondary characteristic classes**.

The central idea fits in one sentence.

Key idea: The principle of secondary classes

When a characteristic class $c_j(F)$ *vanishes*, its vanishing is not the end of the story: the form that makes it vanish (the β with $c_j = \mathrm{d}\beta$) carries new, finer information. That β is a *secondary* class. It is only defined up to integers, so it lives in $\mathbb{R}/\mathbb{Z}$ rather than in $\mathbb{Z}$.

We have met this already twice without naming it. The Chern–Simons number N_{CS} of Section 8.1.19 was gauge invariant only modulo integers. And the phase of an Aharonov–Bohm holonomy is a real number defined modulo 2π. In this section we see that both are instances of one construction. The plan is:

1. a warm-up in one dimension, where everything is elementary;
2. a reminder of invariant polynomials and Chern–Weil theory, the machine that produces $c_j(F)$;
3. the *transgression formula* (Nash's (8.42)): on the total space P of the bundle, every $c_j(F)$ becomes exact, $\pi^* c_j = \mathrm{d}TP_j$;
4. the Chern–Simons forms $TP_j(A)$, written out for $j = 1, 2, 3$;
5. what TP_j does on a single fibre (it gives the generators ω_{2j-1} of $H^*_{\mathrm{dR}}(G)$ from Section 8.1.25);

6. the secondary class and its $\mathbb{R}/\mathbb{Z}$ nature (Nash's (8.43)–(8.45));
7. the formulae for μ and ν (Nash's (8.46)–(8.58)), and the three-dimensional Chern–Simons gauge theory (Nash's (8.59)–(8.62)).

8.2.1 Warm-up: the holonomy of a flat $\mathrm{U}(1)$ connection

Take the simplest possible case: $G = \mathrm{U}(1)$ and $M = S^1$, a circle of circumference L with coordinate $x \in [0, L)$. A connection is $A = i\,a(x)\,\mathrm{d}x$ with a real (anti-Hermitian convention; a is the physicist's vector potential times the charge). Its curvature is a 2-form on a 1-manifold, so $F = 0$: *every* connection on a circle is flat, and the first Chern class $c_1(F) = \frac{i}{2\pi}F$ vanishes identically, as a form.

Does the connection then carry no information? It does carry some: the holonomy around the circle. Parallel transport solves $\mathrm{d}\psi/\mathrm{d}x + a_x\, i\,\psi = 0$ (Section 8.1.7 with $A(\dot\gamma) = ia$), so

$$\mathrm{Hol}(A) = \exp\Big(-\oint_{S^1} A\Big) = \exp\Big(-i\oint a\,\mathrm{d}x\Big), \tag{8.100}$$

the Aharonov–Bohm phase. Now form the first-order expression that Chern–Weil theory would suggest,

$$TP_1(A) \coloneqq \frac{i}{2\pi}A, \qquad \int_{S^1} TP_1(A) = \frac{i}{2\pi}\oint i\,a\,\mathrm{d}x = -\frac{1}{2\pi}\oint a\,\mathrm{d}x. \tag{8.101}$$

It is a *potential* for c_1, in the sense that $\mathrm{d}TP_1(A) = \frac{i}{2\pi}\mathrm{d}A = c_1(F)$. Here $c_1(F) = 0$, so $TP_1(A)$ is closed (every 1-form on a circle is), and its integral is a real number.

How does this number change under a gauge transformation $g = e^{i\lambda(x)}$? The function λ need not be periodic: only $e^{i\lambda}$ must be, so $\lambda(L) - \lambda(0) = 2\pi m$ with $m \in \mathbb{Z}$ the winding number of g. Then $A^g = A + g^{-1}\mathrm{d}g = A + i\,\mathrm{d}\lambda$ and

$$\begin{aligned}\int_{S^1} TP_1(A^g) &\overset{(1)}{=} \frac{i}{2\pi}\oint \big(A + i\,\mathrm{d}\lambda\big) \overset{(2)}{=} \int_{S^1} TP_1(A) - \frac{1}{2\pi}\big(\lambda(L) - \lambda(0)\big)\\ &\overset{(3)}{=} \int_{S^1} TP_1(A) - m.\end{aligned} \tag{8.102}$$

(1) The U(1) transformation law $A \mapsto A + g^{-1}\mathrm{d}g$ with $g^{-1}\mathrm{d}g = i\,\mathrm{d}\lambda$.
(2) $\frac{i}{2\pi}\cdot i = -\frac{1}{2\pi}$, and the fundamental theorem of calculus for $\oint \mathrm{d}\lambda$ on the interval $[0, L]$.
(3) λ winds by $2\pi m$ around the circle.

Worked example 8.29: The number that survives

Take the constant potential $a = \Phi/L$, so $\oint a\,\mathrm{d}x = \Phi$ is the flux threading the ring (in units where the charge is 1). Then $\int TP_1(A) = -\Phi/2\pi$. The large gauge transformation $g = e^{2\pi i m x/L}$ shifts $\Phi \mapsto \Phi + 2\pi m$ and $\int TP_1 \mapsto \int TP_1 - m$. So:

- the real number $\int TP_1(A)$ is *not* gauge invariant;
- its class modulo integers, $\int TP_1(A) \bmod 1 \in \mathbb{R}/\mathbb{Z}$, *is*;
- and $\exp\big(2\pi i \int TP_1(A)\big) = e^{-i\Phi} = \mathrm{Hol}(A)$, the Aharonov–Bohm phase of (8.100).

The rotor of Example 8.20, with its energy levels $\frac{1}{2I}(m + \theta/2\pi)^2$, sees exactly this $\mathbb{R}/\mathbb{Z}$ number: $\theta = -\Phi$ modulo 2π.

This is the whole pattern in miniature. A primary class (c_1) vanishes; the form that trivialises it (TP_1) has a well-defined integral; that integral is gauge invariant only modulo the integer periods of $g^{-1}\mathrm{d}g$; the exponential of $2\pi i$ times it is a physical phase.

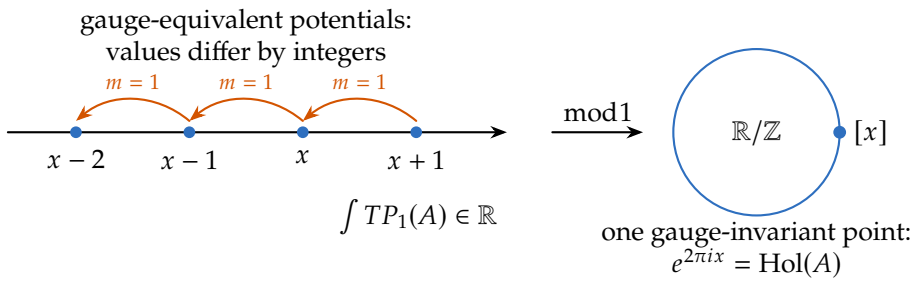


Figure 8.18: Why secondary invariants live in $\mathbb{R}/\mathbb{Z}$. The integral of the Chern–Simons form $TP_1(A) = \frac{i}{2\pi}A$ over a circle takes values that jump by integers under large gauge transformations (orange arrows). Only the image on the circle $\mathbb{R}/\mathbb{Z}$ (right) is physical; its exponential is the Aharonov–Bohm holonomy.

8.2.2 Invariant polynomials and Chern–Weil theory: a reminder

Nash writes (his (8.40))

$$c_j(F) = P_j(\underbrace{F, F, \dots, F}_{j \text{ entries}}),$$

and uses the multilinear object $P_j(\,\cdot\,,\dots,\cdot\,)$ with *different* arguments in the transgression formula. So we need the polynomial in its "polarised" form. This subsection collects what we need; the Chern classes themselves were introduced in Chapter 1.

Working definition 8.15 (invariant polynomial)

An **invariant polynomial** of degree j on the Lie algebra $\mathfrak{g}$ [here anti-Hermitian $N \times N$ matrices, traceless for $\mathrm{SU}(N)$] is a function $P(X_1, \dots, X_j)$ of j matrices $X_i \in \mathfrak{g}$ which is

1. *multilinear* (linear in each slot),
2. *symmetric* (unchanged when two slots are swapped),
3. Ad-*invariant*: $P(hX_1h^{-1}, \dots, hX_jh^{-1}) = P(X_1, \dots, X_j)$ for all $h \in G$.

To build one: take a conjugation-invariant polynomial function $p(X)$ of one matrix, such as $\operatorname{tr} X^j$ or a coefficient of $\det(\mathbb{1} + X)$, and *polarise* it: $P(X_1, \dots, X_j)$ is the coefficient of $t_1 t_2 \cdots t_j$ in $p(t_1X_1 + \cdots + t_jX_j)$, divided by $j!$. For $p = \operatorname{tr} X^j$ this gives the **symmetrised trace**

$$P(X_1, \dots, X_j) = \frac{1}{j!} \sum_{\sigma \in S_j} \operatorname{tr}\big(X_{\sigma(1)} X_{\sigma(2)} \cdots X_{\sigma(j)}\big). \tag{8.103}$$

Inserting forms

The arguments we insert are Lie-algebra-valued *forms*, such as F (a 2-form) or A (a 1-form). Then the entries of the matrices are multiplied with $\wedge$, and the symmetry property picks up the usual sign when two odd forms are swapped. For the arguments that actually occur, one

1-form and several 2-forms, the symmetrisation collapses to a plain trace:

$$P(A,F_1,\dots,F_{j-1}) = \frac{1}{(j-1)!}\sum_{\sigma}\operatorname{tr}\big(A\,F_{\sigma(1)}\cdots F_{\sigma(j-1)}\big), \qquad P(A,F,\dots,F) = \operatorname{tr}\big(A\,F^{j-1}\big)\cdot\kappa_j, \tag{8.104}$$

where κ_j is the overall normalisation of the polynomial. The reason: under the trace, a 2-form can be moved cyclically past anything without a sign, so all orderings in which A sits in a given slot can be cycled to bring A to the front.

The polynomials we use

For $\mathrm{U}(N)$ the Chern classes are defined by the **Chern polynomial**

$$\det\Big(\mathbb{1} + \frac{i}{2\pi}F\Big) = 1 + c_1(F) + c_2(F) + \cdots . \tag{8.105}$$

Expanding the determinant with $X = \frac{i}{2\pi}F$ and $\det(\mathbb{1}+X) = \exp\operatorname{tr}\log(\mathbb{1}+X) = \exp\big(\operatorname{tr}X - \frac{1}{2}\operatorname{tr}X^2 + \frac{1}{3}\operatorname{tr}X^3 - \cdots\big)$,

$$\begin{aligned} c_1 &= \operatorname{tr}X = \frac{i}{2\pi}\operatorname{tr}F, \\ c_2 &= \tfrac{1}{2}\big((\operatorname{tr}X)^2 - \operatorname{tr}X^2\big) = \frac{1}{8\pi^2}\big(\operatorname{tr}F\wedge F - \operatorname{tr}F\wedge\operatorname{tr}F\big), \\ c_3 &= \tfrac{1}{6}(\operatorname{tr}X)^3 - \tfrac{1}{2}\operatorname{tr}X\operatorname{tr}X^2 + \tfrac{1}{3}\operatorname{tr}X^3 \xrightarrow{\operatorname{tr}F=0} \frac{1}{3}\Big(\frac{i}{2\pi}\Big)^3\operatorname{tr}F^3 = -\frac{i}{24\pi^3}\operatorname{tr}(F\wedge F\wedge F). \end{aligned} \tag{8.106}$$

For $\mathrm{SU}(N)$ ($\operatorname{tr}F = 0$) only the single-trace terms survive. Table 8.3 lists them together with Nash's normalisations.

j	$P_j(F,\dots,F)$ used by Nash	relation to c_j	where Nash uses it
1	$\frac{i}{2\pi}\operatorname{tr}F$	$= c_1$	(U(1) holonomy, ω_1)
2	$-\frac{1}{8\pi^2}\operatorname{tr}(F\wedge F)$	$= -c_2$ for $\mathrm{SU}(N)$	(8.60)–(8.61)
3	$-\frac{i}{24\pi^3}\operatorname{tr}(F\wedge F\wedge F)$	$= c_3$ for $\mathrm{SU}(N)$	(8.57)

Table 8.3: The invariant polynomials of this section, for $\mathrm{SU}(N)$ (and U(1) for $j=1$). They are real-valued on anti-Hermitian F: for example $\operatorname{tr}F^2$ is real because $F^\dagger = -F$. Nash's P_2 is $-c_2$, so that $\int_{S^4}P_2(F)$ is the instanton number k of (8.10).

Worked example 8.30: The polynomials on $\mathfrak{su}(2)$

Write $X = X^a e_a$ with the quaternion units $e_a = -i\sigma_a$ of (8.9), which satisfy $e_a e_b = -\delta_{ab}\mathbb{1} + \epsilon_{abc}e_c$. Then

$$\operatorname{tr}(XY) \overset{(1)}{=} X^aY^b\operatorname{tr}(e_ae_b) \overset{(2)}{=} -2\,X^aY^a,$$
$$\operatorname{tr}\big(X\{Y,Z\}\big) \overset{(3)}{=} X^aY^bZ^c\operatorname{tr}\big(e_a(-2\delta_{bc}\mathbb{1})\big) \overset{(4)}{=} 0.$$

(1) Linearity of the trace.
(2) $\operatorname{tr}(e_ae_b) = -\delta_{ab}\operatorname{tr}\mathbb{1} + \epsilon_{abc}\operatorname{tr}e_c = -2\delta_{ab}$ (the e_c are traceless).
(3) $\{e_b,e_c\} = e_be_c + e_ce_b = -2\delta_{bc}\mathbb{1}$: the ϵ terms cancel.
(4) $\operatorname{tr}e_a = 0$.

So P_2 on $\mathfrak{su}(2)$ is (minus) the Euclidean dot product, $P_2(X,Y) \propto X^a Y^a$, while the symmetrised cubic trace vanishes identically: $P_3 \equiv 0$ *for* SU(2). This is the Lie-algebra fact behind the entries "all zero" for U(2) in Nash's table (8.38), and behind the absence of a perturbative SU(2) gauge anomaly. For SU(3) the analogue $\mathrm{tr}(T^a\{T^b,T^c\}) = \frac{1}{2}d^{abc}$ is non-zero.

Two facts about invariant polynomials of forms

(i) d *can be replaced by* D. Differentiate the invariance condition $P(hX_1h^{-1},\dots) = P(X_1,\dots)$ at $h = \mathbb{1}$ along $h = \mathbb{1} + \epsilon Y$: the result is $\sum_i P(X_1,\dots,[Y,X_i],\dots,X_j) = 0$. With forms, and $Y = A$ a 1-form, the same holds with graded commutators and the signs of the graded Leibniz rule. Adding this zero to the ordinary Leibniz rule gives

$$\mathrm{d}P(X_1,\dots,X_j) = \sum_i (-1)^{|X_1|+\cdots+|X_{i-1}|} P(X_1,\dots,DX_i,\dots,X_j), \qquad DX = \mathrm{d}X + [A,X], \quad (8.107)$$

where $|X|$ is the form degree and $[A,X] = AX - (-1)^{|X|}XA$ is the graded commutator of (8.7). (This is the "two-sided" covariant derivative of Section 8.1.1.)

(ii) Chern–Weil: $P_j(F)$ *is closed, and its class does not depend on* A. Closedness is (8.107) with every $X_i = F$ plus the Bianchi identity $DF = 0$. For independence, take two connections A_0, A_1 on the same bundle. By Nash's (8.1) the straight line $A_s = A_0 + s\,a$, $a = A_1 - A_0$, is a path of connections, and a is a globally defined $\mathfrak{g}$-valued 1-form (it transforms two-sidedly). Then $\mathrm{d}F_s/\mathrm{d}s = D_s a$ and

$$\begin{aligned} P_j(F_1) - P_j(F_0) &\overset{(1)}{=} \int_0^1 \frac{\mathrm{d}}{\mathrm{d}s} P_j(F_s,\dots,F_s)\,\mathrm{d}s \overset{(2)}{=} j\int_0^1 P_j(D_s a, F_s,\dots,F_s)\,\mathrm{d}s \\ &\overset{(3)}{=} \mathrm{d}\Big(j\int_0^1 P_j(a,F_s,\dots,F_s)\,\mathrm{d}s \Big). \end{aligned} \quad (8.108)$$

(1) Fundamental theorem of calculus in the parameter s.
(2) Multilinearity and symmetry: differentiating each of the j slots gives the same term; $\frac{\mathrm{d}}{\mathrm{d}s}F_s = \mathrm{d}a + [A_s,a] = D_s a$ (substitute $F_s = F_0 + s\,D_0 a + s^2 a^2$).
(3) (8.107) with $X_1 = a$ and the other slots F_s, whose covariant derivatives vanish by Bianchi.

The form in brackets is built from a and F_s, both tensorial, so it is a *global* form on M: the two Chern–Weil forms differ by an exact form, and define the same class $c_j(P) \in H^{2j}_{\mathrm{dR}}(M)$. This *relative* transgression is the prototype of everything that follows. The trick of Nash's (8.41) is to use it with $A_0 = 0$, which is not a connection on M but is perfectly good on P.

8.2.3 Pulling back to P: the transgression formula

We now follow Nash's (8.39)–(8.42). Let $\pi : P \to M$ be a principal G-bundle (his (8.39)) with a connection.

Why go up to P?

On M a connection is a collection of local potentials A_i on patches U_i, glued by $A_j = g_{ij}^{-1}A_i g_{ij} + g_{ij}^{-1}\mathrm{d}g_{ij}$. There is no global potential unless the bundle is trivial, and so no global way to write $c_j(F) = \mathrm{d}(\text{something built from } A)$. That is as it must be: if $c_j(F)$ were globally exact on M, its integral over closed $2j$-cycles would vanish, and the monopole and the instanton would have zero charge.

On the total space P the situation is different. The connection is a *single global* 1-form ω on P (Section 8.1.1); in a local trivialisation $\pi^{-1}(U) \cong U \times G$, (x, h), it is

$$\omega = h^{-1}A\,h + h^{-1}\mathrm{d}h, \qquad F_\omega := \mathrm{d}\omega + \omega \wedge \omega = h^{-1}F\,h, \tag{8.109}$$

and the gluing rule for the A_i is exactly what makes these local expressions agree on overlaps. Because of Ad-invariance, $P_j(F_\omega, \dots, F_\omega) = P_j(h^{-1}Fh, \dots) = \pi^* P_j(F, \dots, F)$: the Chern–Weil form of F_ω on P *is* the pull-back of $c_j(F)$.

Working definition 8.16 (the family $\omega_t = t\omega$ (Nash's (8.41)))

For $t \in [0, 1]$ put

$$\omega_t := t\,\omega, \qquad F_t := \mathrm{d}\omega_t + \omega_t \wedge \omega_t = t\,\mathrm{d}\omega + t^2\,\omega \wedge \omega. \tag{8.110}$$

At $t = 1$ this is the curvature F_ω; at $t = 0$ it is zero. For $0 < t < 1$, $t\omega$ is *not* a connection (on a fibre [with the point $x \in M$ held fixed, so the $h^{-1}Ah$ term drops out] it restricts to $t\,h^{-1}\mathrm{d}h$, not to the Maurer–Cartan form), but F_t is a perfectly good $\mathfrak{g}$-valued 2-form on P, and that is all we need.

Two algebraic identities hold for *any* $\mathfrak{g}$-valued 1-form, so in particular for ω_t. With $D_t X := \mathrm{d}X + [\omega_t, X]$:

$$\frac{\mathrm{d}F_t}{\mathrm{d}t} \overset{(1)}{=} \mathrm{d}\omega + 2t\,\omega \wedge \omega \overset{(2)}{=} \mathrm{d}\omega + [\omega_t, \omega] = D_t\omega, \tag{8.111}$$

$$D_t F_t \overset{(3)}{=} t^2(\mathrm{d}\omega\,\omega - \omega\,\mathrm{d}\omega) + t^2(\omega\,\mathrm{d}\omega - \mathrm{d}\omega\,\omega) + t^3(\omega^3 - \omega^3) \overset{(4)}{=} 0. \tag{8.112}$$

(1) Differentiate (8.110) in t.

(2) For two 1-forms the graded commutator is an anticommutator: $[t\omega, \omega] = t(\omega\omega + \omega\omega) = 2t\,\omega^2$.

(3) $\mathrm{d}F_t = t^2\mathrm{d}(\omega\omega) = t^2(\mathrm{d}\omega\,\omega - \omega\,\mathrm{d}\omega)$ and $[\omega_t, F_t] = \omega_t F_t - F_t\omega_t$ (a 2-form: commutator); expand.

(4) Everything cancels. This is the Bianchi identity, and it never used that ω_t is a connection.

Key idea: The transgression formula (Nash's (8.42))

On the total space P, every Chern–Weil form becomes exact:

$$\pi^* c_j(F) = \mathrm{d}\,TP_j(\omega), \qquad TP_j(\omega) := j\int_0^1 \mathrm{d}t\; P_j(\omega, F_t, \dots, F_t) \;\in\; \Omega^{2j-1}(P). \tag{8.113}$$

T stands for **transgression**. The $(2j-1)$-form $TP_j(\omega)$ is globally defined on P (it is built from the global ω).

Derivation.

$$\begin{aligned}
\pi^* c_j(F) &\overset{(1)}{=} P_j(F_1, \dots, F_1) - P_j(F_0, \dots, F_0) \overset{(2)}{=} \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t} P_j(F_t, \dots, F_t)\,\mathrm{d}t \\
&\overset{(3)}{=} j\int_0^1 P_j\big(D_t\omega, F_t, \dots, F_t\big)\mathrm{d}t \overset{(4)}{=} j\int_0^1 \mathrm{d}P_j\big(\omega, F_t, \dots, F_t\big)\mathrm{d}t \\
&\overset{(5)}{=} \mathrm{d}\Big(j\int_0^1 P_j(\omega, F_t, \dots, F_t)\,\mathrm{d}t\Big) = \mathrm{d}\,TP_j(\omega).
\end{aligned} \tag{8.114}$$

(1) $F_1 = F_\omega$ and $P_j(F_\omega) = \pi^* c_j(F)$; $F_0 = 0$ gives $P_j(0) = 0$.

(2) Fundamental theorem of calculus in t.
(3) Differentiate each of the j slots (symmetry makes the j terms equal) and use (8.111).
(4) (8.107) for ω_t: $\mathrm{d}P_j(\omega, F_t, \ldots) = P_j(D_t\omega, F_t, \ldots) - \sum P_j(\omega, \ldots, D_tF_t, \ldots)$, and the sum vanishes by (8.112).
(5) d acts on P and commutes with the integral over the parameter t.

This is the same calculation as (8.108) with $A_0 = 0$, $A_1 = \omega$. On M it could not be done, because 0 and A are not two connections on the same bundle; on P it can, because we never needed ω_t to be a connection, only the two identities (8.111)–(8.112).

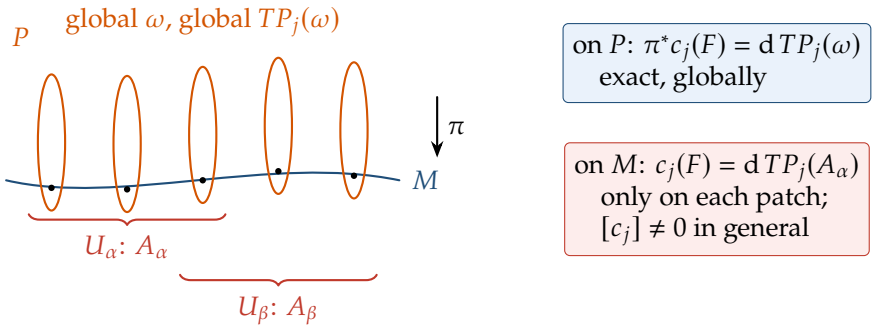


Figure 8.19: Transgression. Upstairs, on the total space P, the connection is one global form ω, and every characteristic form becomes the derivative of the global form $TP_j(\omega)$. Downstairs, on M, only local potentials A_α exist (red braces), so $c_j(F)$ is exact only patch by patch, and its class can be non-zero. The fibres (orange) absorb the difference.

Watch out: Exact on P does not mean trivial

π^*c_j being exact on P says that the class c_j *dies* when pulled back to P. This is a statement about P, not about M: π^* can kill classes. Example: the Hopf bundle $S^3 \to S^2$ below, where $H^2_{\mathrm{dR}}(S^3) = 0$ kills the generator of $H^2_{\mathrm{dR}}(S^2) = \mathbb{R}$. Nash cites Nash and Sen for the general statement.

Worked example 8.31: The Hopf bundle: transgression by hand

$P = S^3 = \{(z_1, z_2) \in \mathbb{C}^2 : |z_1|^2 + |z_2|^2 = 1\}$, $G = \mathrm{U}(1)$ acting by $(z_1, z_2) \cdot \lambda = (z_1\lambda, z_2\lambda)$, $M = S^2 = \mathbb{CP}^1$ with the coordinate $u = z_1/z_2$ away from one point.

- *The connection.* $\omega := \bar{z}_1\mathrm{d}z_1 + \bar{z}_2\mathrm{d}z_2$ restricted to S^3. It is imaginary, as a $\mathfrak{u}(1)$-valued form should be: differentiating $\bar{z}_kz_k = 1$ gives $\bar{z}_k\mathrm{d}z_k = -\overline{\bar{z}_k\mathrm{d}z_k}$. On the fibre direction $\frac{\mathrm{d}}{\mathrm{d}\phi}(ze^{i\phi}) = iz$ it gives $\omega(iz) = i\,|z|^2 = i$, the generator: it restricts to the Maurer–Cartan form $i\,\mathrm{d}\phi$ of $\mathrm{U}(1)$.
- *The curvature.* $F_\omega = \mathrm{d}\omega = \mathrm{d}\bar{z}_1 \wedge \mathrm{d}z_1 + \mathrm{d}\bar{z}_2 \wedge \mathrm{d}z_2$ (Abelian, so $\omega \wedge \omega = 0$).
- *Transgression.* For $j = 1$, $F_t = t\,\mathrm{d}\omega$ and $TP_1(\omega) = 1 \cdot \int_0^1 \mathrm{d}t\,\frac{i}{2\pi}\omega = \frac{i}{2\pi}\omega$. Indeed $\mathrm{d}TP_1(\omega) = \frac{i}{2\pi}F_\omega = \pi^*c_1$: exact on S^3.
- *Down on S^2.* The local section $s(u) = (u, 1)/\sqrt{1 + |u|^2}$ pulls ω back to

$$A = s^*\omega \overset{(1)}{=} \frac{\bar{u}\,\mathrm{d}u - u\,\mathrm{d}\bar{u}}{2(1 + |u|^2)} \overset{(2)}{=} \frac{i\,r^2\,\mathrm{d}\phi}{1 + r^2}, \qquad u = re^{i\phi}.$$

(1) With $\rho = \sqrt{1+|u|^2}$: $\bar z_1 \mathrm{d}z_1 + \bar z_2 \mathrm{d}z_2 = \bar u\,\mathrm{d}u/\rho^2 - (|u|^2+1)\mathrm{d}\rho/\rho^3 = \bar u\,\mathrm{d}u/\rho^2 - \mathrm{d}\rho/\rho$, and $\mathrm{d}\rho/\rho = (\bar u\,\mathrm{d}u + u\,\mathrm{d}\bar u)/2\rho^2$.
(2) $\bar u\,\mathrm{d}u - u\,\mathrm{d}\bar u = 2i\,r^2\,\mathrm{d}\phi$.

Now compute the Chern number by Stokes, using the section on the disc $|u| < R$:

$$\int_{S^2} c_1(F) \overset{(1)}{=} \lim_{R\to\infty} \int_{|u|<R} \mathrm{d}\,TP_1(A) \overset{(2)}{=} \lim_{R\to\infty} \frac{i}{2\pi} \oint_{|u|=R} \frac{i\,R^2\,\mathrm{d}\phi}{1+R^2} \overset{(3)}{=} -1.$$

(1) The disc exhausts S^2 minus the point $u=\infty$ (measure zero); on the disc $c_1 = \mathrm{d}(s^*TP_1(\omega))$.
(2) Stokes, with $\oint \mathrm{d}\phi = 2\pi$.
(3) $\frac{i\cdot i}{2\pi}\cdot 2\pi\cdot 1 = -1$.

Look at what the boundary circle is. As $R\to\infty$, $s(Re^{i\phi}) \to (e^{i\phi}, 0)$: the image of the boundary loop is *one whole fibre*, the fibre over $u=\infty$. So the Chern number equals the integral of $TP_1(\omega)$ over a fibre, $\int_{\mathrm{U}(1)} \frac{i}{2\pi}\, i\,\mathrm{d}\phi = -1$. (The value -1 is the standard answer for the tautological bundle over $\mathbb{CP}^1$ with its complex orientation.) The restriction of the transgression form to a fibre is the subject of Section 8.2.5.

8.2.4 Chern–Simons forms

To get back down to M, pull the transgression formula back with a local section $s: U \to P$. Since $s^*\omega = A$ is the local potential and pull-back commutes with d and $\wedge$,

$$c_j(F) = \mathrm{d}\,TP_j(A) \quad \text{on } U, \qquad TP_j(A) = j\int_0^1 \mathrm{d}t\; P_j(A, F_t, \dots, F_t), \quad F_t = t\,\mathrm{d}A + t^2 A\wedge A. \tag{8.115}$$

This is Nash's (8.42) read on M, and it is how he uses it in (8.56) and (8.61). The local forms $TP_j(A)$ are the **Chern–Simons forms**. The t-integrals are elementary; we do them for $j = 1, 2, 3$.

Worked example 8.32: TP_1, TP_2, TP_3 explicitly

Use the normalisations of Table 8.3 and (8.104).
$j=1$. $TP_1(A) = \int_0^1 \mathrm{d}t\,\frac{i}{2\pi}\operatorname{tr} A = \frac{i}{2\pi}\operatorname{tr} A$.
$j=2$ *(Nash's (8.61)).*

$$\begin{aligned} TP_2(A) &\overset{(1)}{=} 2\int_0^1 \mathrm{d}t\left(-\frac{1}{8\pi^2}\right)\operatorname{tr}\big(A\wedge(t\,\mathrm{d}A + t^2 A\wedge A)\big) \\ &\overset{(2)}{=} -\frac{2}{8\pi^2}\left(\frac{1}{2}\operatorname{tr}(A\wedge\mathrm{d}A) + \frac{1}{3}\operatorname{tr}(A\wedge A\wedge A)\right) \overset{(3)}{=} -\frac{1}{8\pi^2}\operatorname{tr}\left(A\wedge\mathrm{d}A + \frac{2}{3}A\wedge A\wedge A\right). \end{aligned} \tag{8.116}$$

(1) (8.115) with $P_2(X,Y) = -\frac{1}{8\pi^2}\operatorname{tr}(XY)$.
(2) $\int_0^1 t\,\mathrm{d}t = \frac{1}{2}$, $\int_0^1 t^2\mathrm{d}t = \frac{1}{3}$.
(3) Multiply out.

So $TP_2(A) = -\frac{1}{8\pi^2}Q_3(A)$ with Q_3 the Chern–Simons 3-form of (8.69), and $\mathrm{d}TP_2 = -\frac{1}{8\pi^2}\operatorname{tr}F^2 = P_2(F)$, which we verified line by line in Section 8.1.18. (Nash writes $\operatorname{tr}(\mathrm{d}A\wedge A)$ in (8.62); it equals $\operatorname{tr}(A\wedge\mathrm{d}A)$, because moving the 2-form $\mathrm{d}A$ past the 1-form A under the trace costs no sign.)

$j = 3$ *(Nash's (8.56)).* Write $\kappa_3 = -\frac{i}{24\pi^3}$. Then

$$\begin{aligned} TP_3(A) &\overset{(1)}{=} 3\kappa_3 \int_0^1 \mathrm{d}t\ \mathrm{tr}\big(A(t\,\mathrm{d}A + t^2A^2)(t\,\mathrm{d}A + t^2A^2)\big) \\ &\overset{(2)}{=} 3\kappa_3 \int_0^1 \mathrm{d}t\,\Big(t^2\,\mathrm{tr}(A\,\mathrm{d}A\,\mathrm{d}A) + 2t^3\,\mathrm{tr}(A^3\mathrm{d}A) + t^4\,\mathrm{tr}\,A^5\Big) \\ &\overset{(3)}{=} \kappa_3\ \mathrm{tr}\Big(A(\mathrm{d}A)^2 + \frac{3}{2}A^3\,\mathrm{d}A + \frac{3}{5}A^5\Big). \end{aligned} \tag{8.117}$$

(1) (8.115) and (8.104); the factor $j = 3$.
(2) Expand. The two cross terms $\mathrm{tr}(A\,\mathrm{d}A\,A^2)$ and $\mathrm{tr}(A\,A^2\,\mathrm{d}A)$ are equal: cycle the 2-form A^2 from the back of the first to the front.
(3) $\int t^2 = \frac{1}{3}$, $\int t^3 = \frac{1}{4}$, $\int t^4 = \frac{1}{5}$; so $3(\frac{1}{3}, \frac{2}{4}, \frac{1}{5}) = (1, \frac{3}{2}, \frac{3}{5})$.

One can check $\mathrm{d}TP_3 = \kappa_3\,\mathrm{tr}\,F^3$ by the same cycling rules as in Section 8.1.18; it is the five-dimensional Chern–Simons form that governs the four-dimensional anomaly (Section 8.2.8).

Dictionary: Chern–Simons forms in physics notation

With $A = -igA^aT^a$ (Hermitian T^a, $\mathrm{tr}\,T^aT^b = \frac{1}{2}\delta^{ab}$, $[T^a, T^b] = if^{abc}T^c$), the 3-form becomes

$$\mathrm{tr}\Big(A\,\mathrm{d}A + \tfrac{2}{3}A^3\Big) = -\frac{g^2}{2}\Big(A^a \wedge \mathrm{d}A^a + \frac{g}{3}f^{abc}A^a \wedge A^b \wedge A^c\Big),$$

and in components $A^a \wedge \mathrm{d}A^a = \epsilon^{\mu\nu\rho}A^a_\mu\partial_\nu A^a_\rho\,\mathrm{d}^3x$. The Abelian form $\epsilon^{\mu\nu\rho}a_\mu\partial_\nu a_\rho$ is the Chern–Simons term of the quantum Hall effect.

8.2.5 On a fibre: the transgression of the Maurer–Cartan generators

What does $TP_j(\omega)$ look like on a single fibre $\pi^{-1}(x) \cong G$? There A contributes nothing (the fibre is a point of M), and ω restricts to the Maurer–Cartan form $\theta = h^{-1}\mathrm{d}h$. Since pull-back to the fibre commutes with d and $\wedge$, $TP_j(\omega)|_{\text{fibre}} = TP_j(\theta)$, which we can compute. The Maurer–Cartan equation $\mathrm{d}\theta = -\theta^2$ gives

$$F_t\big|_{\text{fibre}} = t\,\mathrm{d}\theta + t^2\theta^2 = (t^2 - t)\,\theta \wedge \theta, \tag{8.118}$$

which vanishes at $t = 0$ and $t = 1$, as it must: the genuine connection on a fibre is flat. With $P_j(X, \ldots, X) = \kappa_j\,\mathrm{tr}\,X^j$,

$$TP_j(\theta) \overset{(1)}{=} j\,\kappa_j \int_0^1 \mathrm{d}t\,(t^2 - t)^{j-1}\,\mathrm{tr}\big(\theta\,(\theta^2)^{j-1}\big) \overset{(2)}{=} j\,\kappa_j\,(-1)^{j-1}\frac{((j-1)!)^2}{(2j-1)!}\,\mathrm{tr}\,\theta^{2j-1}. \tag{8.119}$$

(1) (8.113) with $\omega \to \theta$ and (8.118); by (8.104) the polynomial is a plain trace.
(2) $\int_0^1 (t^2-t)^m \mathrm{d}t = (-1)^m \int_0^1 t^m(1-t)^m \mathrm{d}t = (-1)^m \frac{(m!)^2}{(2m+1)!}$, the Euler Beta integral $B(m+1, m+1)$, with $m = j - 1$.

Explicitly, with the normalisations of Table 8.3:

$$\begin{aligned} &TP_1(\theta) = \frac{i}{2\pi}\,\mathrm{tr}\,\theta, \qquad TP_2(\theta) \overset{(1)}{=} 2\Big(-\frac{1}{8\pi^2}\Big)\Big(-\frac{1}{6}\Big)\,\mathrm{tr}\,\theta^3 = \frac{1}{24\pi^2}\,\mathrm{tr}\,\theta^3, \\ &TP_3(\theta) \overset{(2)}{=} 3\Big(-\frac{i}{24\pi^3}\Big)\frac{1}{30}\,\mathrm{tr}\,\theta^5 = -\frac{i}{240\pi^3}\,\mathrm{tr}\,\theta^5. \end{aligned} \tag{8.120}$$

(1) $j = 2$: the Beta integral is $-\frac{1}{6}$.

(2) $j = 3$: the Beta integral is $\frac{(2!)^2}{5!} = \frac{1}{30}$.

These are exactly the forms $\omega_{2j-1} = C_j \operatorname{tr} \theta^{2j-1}$ of (8.87), with the constants C_1, C_2, C_3 quoted there. In particular $TP_2(\theta)$ is the winding density (8.64).

Key idea: Transgression links $H^*(G)$ and $H^*(BG)$

For each j there is one form $TP_j(\omega)$ on P with

$$\underbrace{TP_j(\omega)\big|_{\text{fibre}} = \omega_{2j-1}}_{\text{generator of } H^{2j-1}_{\mathrm{dR}}(G)} \qquad \text{and} \qquad \underbrace{\mathrm{d}\,TP_j(\omega) = \pi^* c_j}_{\text{generator of } H^{2j}(BG)}.$$

Topologists say that the class $[\omega_{2j-1}]$ of the fibre **transgresses** to the class c_j of the base; this is where Nash's name "T" comes from. For the universal bundle $EG \to BG$ it matches the two rings of Section 8.1 generator by generator: the exterior algebra $\Lambda(\omega_1, \omega_3, \dots, \omega_{2N-1})$ of $\mathrm{U}(N)$ and the polynomial algebra $\mathbb{R}[c_1, c_2, \dots, c_N]$ of $B\mathrm{U}(N)$.

$$G \xrightarrow{\ \iota\ } P \xrightarrow{\ \pi\ } M$$

$$\omega_{2j-1} = C_j \operatorname{tr} \theta^{2j-1} \xleftarrow{\ \iota^*\ } TP_j(\omega) \xmapsto{\ \mathrm{d}\ } \mathrm{d}TP_j(\omega) = \pi^* c_j(F)$$

Figure 8.20: The transgression in one diagram. The single form $TP_j(\omega)$ on the total space restricts on each fibre G to the Maurer–Cartan generator ω_{2j-1} (left), and its derivative is the pull-back of the Chern–Weil form $c_j(F)$ of the base (right). Read left to right, a class of the group "moves" to a class of the base, raising the degree by one.

Worked example 8.33: Chern number = fibre integral, again

The Hopf computation of Example 8.31 was not a coincidence. Let $P \to S^{2j}$ be clutched by $g : S^{2j-1} \to G$ along the equator (Section 8.1.12). Over the northern hemisphere D_+ use a section s_+, over the southern D_- a section s_-, with $s_- = s_+ g$ on the equator. Then

$$\int_{S^{2j}} c_j(F) \overset{(1)}{=} \int_{D_+} \mathrm{d}\, s_+^* TP_j(\omega) + \int_{D_-} \mathrm{d}\, s_-^* TP_j(\omega) \overset{(2)}{=} \int_{S^{2j-1}} (s_+^* - s_-^*) TP_j(\omega)$$
$$\overset{(3)}{=} -\int_{S^{2j-1}} \left(TP_j(A_+^g) - TP_j(A_+) \right) \overset{(4)}{=} -\int_{S^{2j-1}} g^* \omega_{2j-1}.$$

(1) On each hemisphere $c_j = \mathrm{d}TP_j$ of the local potential.

(2) Stokes. The equator is the boundary of D_+ with one orientation and of D_- with the other.

(3) $s_-^* \omega = A_- = A_+^g$, the gauge transform by the clutching function.

(4) The gauge-variation formula (8.122) of the next subsection: the difference is $g^* \omega_{2j-1}$ plus an exact form, which integrates to zero on the closed equator.

So the Chern number is (minus) the Maurer–Cartan period of the clutching function. For $j = 2$, with Nash's $P_2 = -c_2$, this is $k = \int_{S^4} P_2 = \frac{1}{24\pi^2} \int \operatorname{tr}(g^{-1}\mathrm{d}g)^3 = w(g)$, the result (8.70) of Section 8.1.18 (the overall sign depends on which hemisphere is called D_+).

8.2.6 When $c_j(F)$ vanishes: the secondary class

Nash now asks (his (8.43)–(8.44)): what if the form $c_j(F)$ is identically zero? Then the transgression formula reads $\mathrm{d}TP_j(\omega) = 0$. The form $TP_j(\omega)$ is *closed* on P and defines a de Rham class

$$[TP_j(\omega)] \in H^{2j-1}_{\mathrm{dR}}(P) = H^{2j-1}(P;\mathbb{R}), \tag{8.121}$$

of *odd* degree. This is the **secondary characteristic class**, or **Chern–Simons class**, of the connection.

Working definition 8.17 (secondary (Chern–Simons) class)

Given a connection ω with $P_j(F) = 0$ *as a form*, the secondary class is the cohomology class of $TP_j(\omega)$ [the transgression form (8.113), with $\mathrm{d}TP_j(\omega) = P_j(F_\omega)$, so it is closed here] on P. *To compute it:* pull back by a local section [a map $s : U \to P$ with $\pi \circ s = \mathrm{id}$, such as $h = \mathbb{1}$, along which $s^*\omega = A$], $TP_j(A)$, and integrate over closed $(2j-1)$-cycles (for $j = 2$: over closed 3-manifolds in M). *Formally:* the class in $H^{2j-1}(P;\mathbb{R})$ of the transgression form.

When does $c_j(F)$ vanish as a form? Three situations cover the applications.

1. *For dimensional reasons*: $c_j(F)$ is a $2j$-form, so it is zero on any manifold of dimension less than $2j$. On a 3-manifold, $c_2(F) = 0$ for every connection. This is the case of Chern–Simons gauge theory (Section 8.2.10), and of Nash's (8.50) below.
2. *For flat connections*: $F = 0$ kills every $c_j(F)$ with $j \geq 1$. Flat connections are classified by their holonomy, and the secondary classes are then invariants of the holonomy representation, such as the $\mathbb{R}/\mathbb{Z}$ phases of Section 8.2.1.
3. *For algebraic reasons*: $c_j \equiv 0$ for $j > N$ on a rank-N bundle, and $P_3 \equiv 0$ on $\mathfrak{su}(2)$ (Example 8.30).

Watch out: A real class, not an integral one

The primary class $c_j(F)$ is *integral*: its integrals over closed $2j$-cycles are integers. The secondary class TP_j is only *real*: its integrals are arbitrary real numbers (the flux $-\Phi/2\pi$ of Example 8.29 can be anything). That is what makes it interesting. It depends on the connection, not just on the bundle; it is a *geometric* invariant, not merely a topological one. Only its reduction modulo integers is independent of the choice of gauge, as we now show.

8.2.7 Gauge dependence and the $\mathbb{R}/\mathbb{Z}$ class

To integrate the secondary class over a cycle in M we must bring it down from P, and that needs a section. Different sections differ by gauge transformations. How does TP_j change?

Key idea: Gauge variation of the Chern–Simons form

For a gauge transformation $g : U \to G$ and $A^g = g^{-1}Ag + g^{-1}\mathrm{d}g$,

$$TP_j(A^g) = TP_j(A) + g^*\omega_{2j-1} + \mathrm{d}\beta_j(A,g), \qquad g^*\omega_{2j-1} = TP_j(g^{-1}\mathrm{d}g) = C_j\,\mathrm{tr}\big(g^{-1}\mathrm{d}g\big)^{2j-1}, \tag{8.122}$$

for some $(2j-2)$-form β_j built from A and g. For $j = 2$ (Nash's normalisation):

$$TP_2(A^g) = TP_2(A) + \frac{1}{24\pi^2}\,\mathrm{tr}\big(g^{-1}\mathrm{d}g\big)^3 + \frac{1}{8\pi^2}\,\mathrm{d}\,\mathrm{tr}\big(\mathrm{d}g\,g^{-1}\wedge A\big). \tag{8.123}$$

The case $j = 2$. This is (8.75) multiplied by $-\frac{1}{8\pi^2}$:

$$TP_2(A^g) \overset{(1)}{=} -\frac{1}{8\pi^2} Q_3(A^g) \overset{(2)}{=} -\frac{1}{8\pi^2}\Big(Q_3(A) - \mathrm{d}\,\mathrm{tr}(bA) - \tfrac{1}{3}\,\mathrm{tr}\,\theta^3\Big)$$
$$\overset{(3)}{=} TP_2(A) + \frac{1}{8\pi^2}\mathrm{d}\,\mathrm{tr}(bA) + \frac{1}{24\pi^2}\,\mathrm{tr}\,\theta^3,$$

(1) (8.116).
(2) (8.75), with $b = \mathrm{d}g\, g^{-1}$ and $\theta = g^{-1}\mathrm{d}g$.
(3) Distribute the prefactor.

Why the general formula holds. Let $\Delta(A) := TP_j(A^g) - TP_j(A)$ for fixed g.

- $\Delta(A)$ is closed: $\mathrm{d}TP_j(A^g) = P_j(F^g) = P_j(g^{-1}Fg) = P_j(F) = \mathrm{d}TP_j(A)$ by Ad-invariance.
- Δ changes by exact forms when A is deformed. The variation of a Chern–Simons form under $A \to A + \delta A$ is

$$\delta\, TP_j(A) = j\, P_j(\delta A, F, \dots, F) + \mathrm{d}(\dots). \tag{8.124}$$

  For $j = 2$ this is a two-line check: $\delta\,\mathrm{tr}(A\,\mathrm{d}A + \frac{2}{3}A^3) = \mathrm{tr}(\delta A\,\mathrm{d}A + A\,\mathrm{d}\delta A + 2\,\delta A\,A^2) = 2\,\mathrm{tr}(\delta A\,F) - \mathrm{d}\,\mathrm{tr}(A\,\delta A)$, using $\mathrm{tr}(A\,\mathrm{d}\delta A) = \mathrm{tr}(\mathrm{d}A\,\delta A) - \mathrm{d}\,\mathrm{tr}(A\,\delta A)$. For general j the same integration by parts in t works. Now $\delta A^g = g^{-1}\delta A\, g$ and $F^g = g^{-1}Fg$, so the P_j terms in $\delta\Delta$ cancel by Ad-invariance, leaving an exact form.
- Deform A to 0 along sA, $s \in [0,1]$ (fine on a patch). Then $\Delta(A) = \Delta(0) + \mathrm{d}(\dots) = TP_j(0^g) - TP_j(0) + \mathrm{d}(\dots) = TP_j(g^{-1}\mathrm{d}g) + \mathrm{d}(\dots)$, and $TP_j(g^{-1}\mathrm{d}g) = g^*TP_j(\theta) = g^*\omega_{2j-1}$ by (8.120).

The periods of $g^*\omega_{2j-1}$ are integers

Integrate (8.122) over a closed $(2j-1)$-cycle $\Sigma \subset M$ (on which A and A^g are both defined). The exact term drops out by Stokes, and

$$\int_\Sigma TP_j(A^g) - \int_\Sigma TP_j(A) = \int_\Sigma g^*\omega_{2j-1} \in \mathbb{Z}. \tag{8.125}$$

Why an integer? The class $[\omega_{2j-1}]$ is the transgression of the *integral* class c_j (Example 8.33 shows that its periods on spheres are Chern numbers), so it is itself integral: its integral over the image of any closed cycle is an integer. For $j = 1$ this is the winding number of $e^{i\lambda}$ around a circle, (8.102); for $j = 2$ and $\Sigma = S^3$ it is the winding number (8.64); for $j = 2$ and a general closed 3-manifold and $G = \mathrm{SU}(2)$ it is the degree of $g : \Sigma \to S^3$.

Key idea: Nash's (8.45): three classes, three coefficient groups

$$c_j(F) \in H^{2j}(M;\mathbb{Z}), \qquad TP_j(\omega) \in H^{2j-1}(P;\mathbb{R}), \qquad \widehat{TP}_j(A) \in H^{2j-1}(M;\mathbb{R}/\mathbb{Z}).$$

The primary class is integral. The secondary class lives upstairs as a real class. Brought down to M with any local section, it is defined only up to the integer periods (8.125), so on M it is a class with $\mathbb{R}/\mathbb{Z}$ coefficients. Concretely: for every closed $(2j-1)$-cycle Σ, $\int_\Sigma TP_j(A)$ mod 1 is gauge invariant, and so is $\exp\big(2\pi i \int_\Sigma TP_j(A)\big)$.

Worked example 8.34: The Chern–Simons invariant of a flat connection on $\mathbb{RP}^3$

Here is a genuinely fractional value. Identify S^3 with SU(2) as in (8.9), and let $\mathbb{RP}^3 = S^3/\{x \sim -x\}$. On S^3 take the pure gauge potential

$$A = x^{-1}\mathrm{d}x \qquad \text{(the Maurer–Cartan form, with } g(x) = x\text{)}.$$

- *It descends to* $\mathbb{RP}^3$. $(-x)^{-1}\mathrm{d}(-x) = x^{-1}\mathrm{d}x$, so A is a well-defined $\mathfrak{su}(2)$-valued 1-form on $\mathbb{RP}^3$, i.e. a connection on the trivial SU(2) bundle over $\mathbb{RP}^3$.
- *It is flat*, $F = \mathrm{d}\theta + \theta^2 = 0$, so $c_2(F) = 0$ and $TP_2(A)$ is closed.
- *It is not gauge-trivial.* A covariantly constant ψ solves $\mathrm{d}\psi + A\psi = 0$, i.e. $\psi = x^{-1}\psi_0$. Along a path from x to $-x$, which is a closed loop in $\mathbb{RP}^3$, ψ returns as $-\psi$: the holonomy is $-\mathbb{1}$.

Now compute its Chern–Simons invariant. The double cover $S^3 \to \mathbb{RP}^3$ covers each point twice, so

$$\int_{\mathbb{RP}^3} TP_2(A) \overset{(1)}{=} \frac{1}{2}\int_{S^3} TP_2\big(x^{-1}\mathrm{d}x\big) \overset{(2)}{=} \frac{1}{2}\cdot\frac{1}{24\pi^2}\int_{S^3}\operatorname{tr}\theta^3 \overset{(3)}{=} \frac{1}{2}\,w(\mathrm{id}) = \frac{1}{2}.$$

(1) The integral of a pulled-back form over a two-sheeted cover is twice the integral downstairs (orient $\mathbb{RP}^3$ by the orientation of S^3; the antipodal map of S^3 preserves it).
(2) (8.120): for a pure gauge $TP_2(\theta) = \frac{1}{24\pi^2}\operatorname{tr}\theta^3$.
(3) (8.64) and (8.66).

Any gauge transformation $h : \mathbb{RP}^3 \to \mathrm{SU}(2)$ shifts this by the integer $\int_{\mathbb{RP}^3} h^*\omega_3$. So the class $\frac{1}{2}$ mod 1 cannot be removed: the $\mathbb{R}/\mathbb{Z}$ invariant *detects* that this flat connection is not the trivial one, which has invariant 0. (Reversing the orientation gives $-\frac{1}{2}$, which is the same point of $\mathbb{R}/\mathbb{Z}$.)

Defining the invariant with a bounding 4-manifold

There is a second way to see why only the fractional part is meaningful, which is the one used in Chern–Simons theory. Every closed oriented 3-manifold M is the boundary of some compact oriented 4-manifold W. Suppose the bundle and the connection extend over W (for SU(N) this can always be arranged). If the bundle over W is trivial with global potential A, Stokes gives

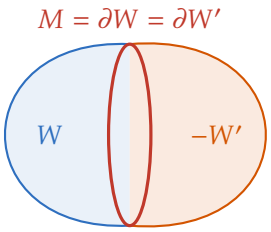
$$\int_M TP_2(A) \overset{(1)}{=} \int_W \mathrm{d}\,TP_2(A) \overset{(2)}{=} \int_W P_2(F), \tag{8.126}$$

(1) Stokes on W with $\partial W = M$.
(2) The transgression formula (8.115).

and one can take the right-hand side as the *definition* of the left in general. Another choice W' gives a closed 4-manifold $X = W \cup_M (-W')$ by gluing along M, and

$$\int_W P_2(F) - \int_{W'} P_2(F') = \int_X P_2 = k(X) \in \mathbb{Z}, \tag{8.127}$$

the instanton number of the glued bundle on the closed 4-manifold X. The Chern–Simons invariant of M is therefore well defined modulo $\mathbb{Z}$: it is "a fraction of an instanton", and the fraction is what the boundary determines. See Figure 8.21.



$$\int_M TP_2(A) = \int_W P_2(F) \pmod{\mathbb{Z}}$$
$$\int_W P_2 - \int_{W'} P_2 = k(W \cup_M -W') \in \mathbb{Z}$$

Figure 8.21: The Chern–Simons invariant of a 3-manifold M (red) computed from a 4-manifold W that it bounds. A different filling W' changes the answer by the instanton number of the closed manifold obtained by gluing W to W' reversed, an integer. So only $\int_M TP_2$ mod 1 is intrinsic to M.

Worked example 8.35: The Chern–Simons number of a pure gauge on S^3

For $M = S^3$ (space compactified, as in Section 8.1.19) and $A = g^{-1}\mathrm{d}g$ with g of winding n, fill S^3 by the ball $W = B^4$. On B^4 take $A_W = f(r)\, g^{-1}\mathrm{d}g$ with $f(0) = 0$, $f(1) = 1$ and g extended radially (singular at the centre, but f tames it, as for the BPST potential in Example 8.18). Then $\int_{S^3} TP_2(g^{-1}\mathrm{d}g) = \int_{B^4} P_2(F_W) = n$: the vacuum of class n has Chern–Simons number n, an integer, and so it is 0 in $\mathbb{R}/\mathbb{Z}$. The $\mathbb{R}/\mathbb{Z}$ invariant cannot tell the classical vacua $|n\rangle$ apart; only the real number can, and the real number is not gauge invariant. This is the $\mathbb{R}/\mathbb{Z}$ way of saying that the $|n\rangle$ are related by large gauge transformations. (N_{CS} of (8.74) is $-\int TP_2$ with the opposite orientation convention, $-\mathrm{d}x^{123}$ versus $\mathrm{d}x^{123}$.)

Watch out: Non-trivial bundles and differential characters

When the bundle over M is not trivial there is no global section, and the recipe "pull TP_j back by a section" has to be carried out patch by patch, with correction terms on overlaps. The clean framework is that of Cheeger–Simons *differential characters*: a secondary invariant assigns to every closed $(2j-1)$-cycle an element of $\mathbb{R}/\mathbb{Z}$, compatibly with $c_j(F)$. In this chapter we only need trivial bundles over 3-manifolds (every SU(N) bundle over a 3-manifold is trivial, because $\pi_2(\mathrm{SU}(N)) = 0$ and $\pi_1(\mathrm{SU}(N)) = 0$), so the naive recipe suffices. Nash's remark that the phenomenon reappears "in chapter 12" refers to Witten's Chern–Simons TQFT (Chapter 12).

Reading Nash (pp. 223–224, (8.39)–(8.45))

Nash states the transgression formula (8.42) without derivation; our (8.114) supplies it, and (8.111)–(8.112) are the two identities it rests on. His "A" in (8.41)–(8.42) is the connection form on P (our ω); the same formula then holds for the local potentials by pull-back. His (8.43)–(8.44) are (8.121). His (8.45) is the keyidea above, and his remark that "the two classes will differ but their difference will be an integral class" is (8.125). The example he promises "in 8.62" is the three-dimensional Chern–Simons action of Section 8.2.10.

Recap: transgression and secondary classes

- Invariant polynomials P_j: symmetric, multilinear, Ad-invariant; $c_j(F) = P_j(F, \dots, F)$ is closed and its class does not depend on A (straight-line homotopy).
- On P: $\pi^* c_j(F) = \mathrm{d}TP_j(\omega)$ with $TP_j = j\int_0^1 P_j(\omega, F_t, \dots, F_t)\,\mathrm{d}t$, $F_t = t\,\mathrm{d}\omega + t^2\omega^2$.
- Chern–Simons forms: $TP_1 = \frac{i}{2\pi}\operatorname{tr} A$, $TP_2 = -\frac{1}{8\pi^2}\operatorname{tr}(A\,\mathrm{d}A + \frac{2}{3}A^3)$, $TP_3 =$

$-\frac{i}{24\pi^3}\operatorname{tr}(A(\mathrm{d}A)^2+\frac{3}{2}A^3\mathrm{d}A+\frac{3}{5}A^5)$.

- On a fibre, TP_j restricts to the generator $\omega_{2j-1}=C_j\operatorname{tr}\theta^{2j-1}$ of $H^*_{\mathrm{dR}}(G)$: transgression.
- If $c_j(F)=0$ as a form, TP_j is closed: a real secondary class on P; on M it is defined modulo the integer periods of $g^*\omega_{2j-1}$, i.e. in $\mathbb{R}/\mathbb{Z}$. Examples: Aharonov–Bohm phase; $\frac{1}{2}$ for the flat connection with holonomy $-\mathbb{1}$ on $\mathbb{RP}^3$.

8.2.8 Secondary classes on the gauge group: an explicit formula for μ

In Section 8.1.23 we built the classes $\mu_{i-n}\in H^{i-n}_{\mathrm{dR}}(\mathcal{G})$ by evaluating gauge transformations, pulling back a generator ω_i of $H^*_{\mathrm{dR}}(G)$ and integrating over S^n. That recipe never mentions the gauge field. But in physics the class μ_1 turns up in a very specific place: as the variation of the effective action of chiral fermions under a gauge transformation (the non-Abelian anomaly, Chapter 10). There it comes out of a Feynman diagram as an explicit polynomial in A and $\mathrm{d}A$. We want the same kind of formula here. Nash's idea is to build μ from the Chern–Simons forms TP_j instead of from ω_{2j-1}. Since TP_j restricts to ω_{2j-1} on each fibre (Section 8.2.5), this should give the same class; and since TP_j is built from the connection, the result will be a formula in A.

Step 1: the evaluation map on $\mathcal{G}\times P$

A gauge transformation $\alpha\in\mathcal{G}$ is a map $\alpha:P\to P$ that moves each point along its own fibre and commutes with the right G-action (a refresher: Section 8.1.3). So there is again an evaluation map, this time with values in P (Nash's (8.46)):

$$\mathrm{Ev}:\mathcal{G}\times P\longrightarrow P,\qquad(\alpha,p)\longmapsto\alpha(p),\qquad\pi\circ\mathrm{Ev}(\alpha,p)=\pi(p).\tag{8.128}$$

The last equation says that Ev does not move the base point, because α covers the identity of $M=S^n$. Pull the Chern–Simons form back (Nash's (8.47)): $\mathrm{Ev}^*TP_j(\omega)$ is a $(2j-1)$-form on the product $\mathcal{G}\times P$.

We do not want forms on $\mathcal{G}\times P$; we want forms on $\mathcal{G}$. The fibres of P are in the way, so we cut them out with a section. Let $s:S^n\to P$ be a section (for the moment assume the bundle is trivial, so that a global one exists; the general case is discussed in the pitfall at the end of this subsection). Define

$$\widehat{\mathrm{Ev}}:S^n\times\mathcal{G}\longrightarrow P,\qquad(x,\alpha)\longmapsto\alpha\big(s(x)\big).\tag{8.129}$$

This is Nash's "restriction of Ev^*TP_j to $\mathcal{G}\times S^n$": we use the section to view S^n as the subset $s(S^n)$ of P. Then (Nash's (8.48))

$$\boxed{\mu_{2j-1-n}:=\int_{S^n}\widehat{\mathrm{Ev}}^*TP_j(\omega)\ \in\ H^{2j-1-n}_{\mathrm{dR}}(\mathcal{G}).}\tag{8.130}$$

Step 2: what TP_j looks like on $S^n\times\mathcal{G}$

Everything is now a computation on the product manifold $S^n\times\mathcal{G}$. We need some bookkeeping for forms on a product.

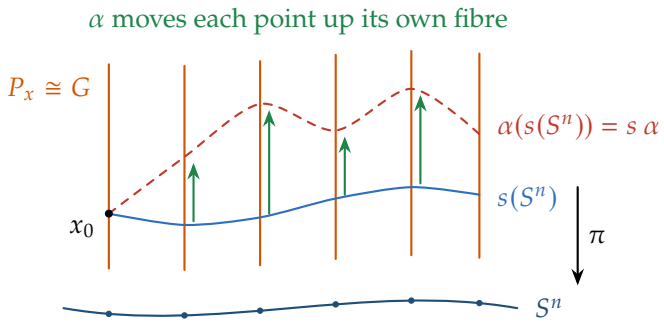


Figure 8.22: The map $\widehat{\mathrm{Ev}}$ of (8.129). A section s (blue) picks one point in each fibre. A gauge transformation α slides every point along its own fibre (green arrows); the image of the section is the new section $s\alpha$ (red, dashed). For the based gauge group α fixes the fibre over x_0. Letting α range over all of $\mathcal{G}$ sweeps out a copy of $S^n \times \mathcal{G}$ inside P (or rather mapped into P), and that is where we pull back TP_j.

Working definition 8.18 (forms on $S^n \times \mathcal{G}$: bidegree and the two derivatives)

- A form on $S^n \times \mathcal{G}$ has legs along S^n (the $\mathrm{d}x^\mu$) and legs along $\mathcal{G}$ [slots that take tangent vectors to $\mathcal{G}$, i.e. infinitesimal gauge transformations]. A form with p legs of the first kind and q of the second has **bidegree** (p, q), and total degree $p + q$.
- The exterior derivative splits as $\mathrm{d}_{\mathrm{tot}} = \mathrm{d} + \delta$: d differentiates along S^n and adds a $\mathrm{d}x$-leg; δ differentiates along $\mathcal{G}$ and adds a $\mathcal{G}$-leg. From $\mathrm{d}_{\mathrm{tot}}^2 = 0$ and the bidegree count, $\mathrm{d}^2 = 0$, $\delta^2 = 0$ and $\mathrm{d}\delta = -\delta\mathrm{d}$.
- Signs follow the *total* degree: a $\mathcal{G}$-leg anticommutes with a $\mathrm{d}x$-leg, and $\delta(\alpha \wedge \beta) = \delta\alpha \wedge \beta + (-1)^{|\alpha|}\alpha \wedge \delta\beta$ with $|\alpha|$ the total degree.
- δ is the familiar "variation". On the space of connections $\mathcal{A}$ the exterior derivative of a functional $\Phi[A]$ is its first variation, $\delta\Phi = \int \frac{\delta\Phi}{\delta A}\,\delta A$; the δA of (8.124) was exactly this.

The key object is a $\mathfrak{g}$-valued 1-form on $\mathcal{G}$, the Maurer–Cartan form of the gauge group:

$$\vartheta := \alpha^{-1}\delta\alpha, \qquad \vartheta_\alpha\Big(\frac{\mathrm{d}}{\mathrm{d}\tau}\alpha_\tau\Big|_{\tau=0}\Big) = \alpha^{-1}(x)\,\frac{\mathrm{d}\alpha_\tau(x)}{\mathrm{d}\tau}\Big|_{\tau=0} := \xi(x). \tag{8.131}$$

In words: feed ϑ a tangent vector to $\mathcal{G}$ at α (a curve of gauge transformations α_τ through α) and it returns the *infinitesimal gauge transformation* $\xi : S^n \to \mathfrak{g}$ that generates the curve, $\alpha_\tau \approx \alpha(\mathbb{1} + \tau\xi)$. So ϑ has bidegree $(0, 1)$: it is a function of x (no $\mathrm{d}x$-legs) and a 1-form on $\mathcal{G}$.

The pulled-back connection. Work in a trivialisation in which the section is $s(x) = (x, \mathbb{1})$ and a gauge transformation acts by $\alpha(s(x)) = s(x)\,\alpha(x)$ for a G-valued function also called $\alpha(x)$. (With the convention (8.18) this function is $g(x)^{-1}$; only the name changes.) Then

$$\begin{aligned}\mathbb{A} := \widehat{\mathrm{Ev}}^*\omega &\overset{(1)}{=} h^{-1}A\,h + h^{-1}\mathrm{d}_{\mathrm{tot}}h\Big|_{h(x,\alpha)=\alpha(x)} \\ &\overset{(2)}{=} \underbrace{\alpha^{-1}A\,\alpha + \alpha^{-1}\mathrm{d}\alpha}_{A^\alpha\text{: bidegree (1,0)}} + \underbrace{\alpha^{-1}\delta\alpha}_{\vartheta\text{: bidegree (0,1)}} = A^\alpha + \vartheta.\end{aligned} \tag{8.132}$$

(1) $\widehat{\mathrm{Ev}}(x, \alpha) = s(x)\,h(x, \alpha)$ with $h = \alpha(x)$, a G-valued function on $S^n \times \mathcal{G}$. By (8.3) the connection form at the point $s(x)h$ is $h^{-1}Ah + h^{-1}\mathrm{d}h$, and pulling back along a map from $S^n \times \mathcal{G}$ replaces d by the exterior derivative of $S^n \times \mathcal{G}$, which is $\mathrm{d}_{\mathrm{tot}}$.

(2) $\mathrm{d}_{\mathrm{tot}} h = \mathrm{d}\alpha + \delta\alpha$; sort by bidegree.

So $\mathbb{A}$ is the gauge transform of A by the function $\alpha(x)$, *with the total derivative*. Its $(1,0)$ part is the ordinary gauge transform A^α; its $(0,1)$ part is ϑ.

Its curvature. The rule $F^g = g^{-1}Fg$ is an algebraic consequence of the Leibniz rule and of $\mathrm{d}(g^{-1}) = -g^{-1}(\mathrm{d}g)g^{-1}$ (Chapter T3), and both hold for $\mathrm{d}_{\mathrm{tot}}$. Hence

$$\mathbb{F} \coloneqq \mathrm{d}_{\mathrm{tot}}\mathbb{A} + \mathbb{A}\wedge\mathbb{A} \overset{(1)}{=} \alpha^{-1}\big(\mathrm{d}_{\mathrm{tot}}A + A\wedge A\big)\alpha \overset{(2)}{=} \alpha^{-1}\big(\mathrm{d}A + A\wedge A\big)\alpha = \alpha^{-1}F\alpha = F^\alpha. \tag{8.133}$$

(1) Gauge covariance of the curvature, applied on $S^n\times\mathcal{G}$ to the gauge transformation $h=\alpha(x)$ with derivative $\mathrm{d}_{\mathrm{tot}}$.
(2) A lives on S^n and does not depend on α, so $\delta A = 0$.

This is the whole secret of the construction. *The curvature $\mathbb{F}$ has no legs along $\mathcal{G}$*: it is purely of bidegree $(2,0)$. The gauge group directions are "flat" directions, because moving along $\mathcal{G}$ does not change the physics.

Physical picture

Think of $S^n\times\mathcal{G}$ as a family of copies of S^n, labelled by α. On the copy labelled α we put the gauge-transformed field A^α. All these fields are gauge equivalent, so they have "the same" curvature, $F^\alpha = \alpha^{-1}F\alpha$. The extra piece ϑ of $\mathbb{A}$ is exactly what is needed to make the family into a single connection on the big space whose curvature does not see the $\mathcal{G}$ directions at all. A family of equivalent fields is a flat family.

Step 3: the forms are closed, and the degree count

Now the two conditions in Nash's (8.49)–(8.50) come out at once.

$$\mathrm{d}_{\mathrm{tot}}\big(\widehat{\mathrm{Ev}}^*TP_j\big) \overset{(1)}{=} \widehat{\mathrm{Ev}}^*\big(\mathrm{d}TP_j\big) \overset{(2)}{=} \widehat{\mathrm{Ev}}^*\pi^*c_j(F) \overset{(3)}{=} \mathrm{pr}^*c_j(F) \overset{(4)}{=} 0 \qquad \text{if } 2j>n. \tag{8.134}$$

(1) Pull-back commutes with the exterior derivative.
(2) The transgression formula (8.113): $\mathrm{d}TP_j(\omega) = \pi^*c_j(F)$ on P.
(3) $\pi\circ\widehat{\mathrm{Ev}}(x,\alpha) = x$, so $\pi\circ\widehat{\mathrm{Ev}}$ is the projection $\mathrm{pr}: S^n\times\mathcal{G}\to S^n$. (Equivalently: the pulled-back c_j is $P_j(\mathbb{F},\dots,\mathbb{F}) = P_j(F^\alpha,\dots) = P_j(F,\dots)$ by (8.133) and Ad-invariance.)
(4) $c_j(F)$ is a $2j$-form on the n-dimensional manifold S^n; if $2j>n$ it vanishes identically.

This is Nash's remark that "we have not yet checked that the characteristic class $c_j(F)$ vanishes but it turns out that this is automatic". The secondary class exists on $S^n\times\mathcal{G}$ precisely because the primary class has no room to live on S^n.

By (8.90), integration over the closed fibre S^n turns the closed $(2j-1)$-form $\widehat{\mathrm{Ev}}^*TP_j$ into a closed form of degree $2j-1-n$ on $\mathcal{G}$. For a non-trivial answer the degree must be non-negative:

$$2j-1-n \;\geq\; 0 \quad\Longleftrightarrow\quad 2j \;\geq\; n+1. \tag{8.135}$$

Since n is even (Nash's standing assumption), $2j\geq n+1$ is the same as $2j>n$, so *whenever μ_{2j-1-n} has a chance of being non-zero, the closedness* (8.134) *is guaranteed*. Figure 8.23 shows the count for $n=2$ and $n=4$.

Step 4: it is the same class as before

Is (8.130) the class (8.89) of Section 8.1.23, or something new? Use the gauge-variation formula (8.122). It was derived on an arbitrary manifold, so it holds on $S^n\times\mathcal{G}$ with $\mathrm{d}\to\mathrm{d}_{\mathrm{tot}}$ and gauge

	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	$j=6$
$n=2$	-1	μ_1	μ_3	μ_5	μ_7	μ_9
$n=4$	-3	-1	μ_1	μ_3	μ_5	μ_7

generators ω_{2j-1} exist for SU(N) when $2 \le j \le N$

Figure 8.23: The degree $2j-1-n$ of the form μ_{2j-1-n} built from TP_j. Grey boxes (negative degree) give nothing; blue boxes give a class. The red line is the condition $2j \ge n+1$ of (8.135). For $n=4$ the first class is μ_1 from $j=3$, so it needs a group with an independent ω_5, i.e. SU(N) with $N \ge 3$: this is Nash's table (8.38) again.

function $h(x,\alpha)=\alpha(x)$. Note that h, as a map $S^n \times \mathcal{G} \to G$, *is* the evaluation map Ev of (8.88). So

$$\widehat{\mathrm{Ev}}^* TP_j(\omega) \overset{(1)}{=} TP_j\big((\mathrm{pr}^* A)^h\big) \overset{(2)}{=} \mathrm{pr}^* TP_j(A) + h^*\omega_{2j-1} + \mathrm{d}_{\mathrm{tot}}\beta_j \overset{(3)}{=} \mathrm{Ev}^*\omega_{2j-1} + \mathrm{d}_{\mathrm{tot}}\beta_j. \tag{8.136}$$

(1) (8.132): $\mathbb{A}$ is the (total) gauge transform of the pulled-back potential pr^*A by h; and pull-back commutes with building TP_j out of the connection.
(2) (8.122) on the manifold $S^n \times \mathcal{G}$.
(3) $\mathrm{pr}^* TP_j(A)$ has only S^n-legs, and there are $2j-1 > n$ of them: it is zero. And $h = \mathrm{Ev}$.

Integrating over S^n and using (8.90) (with n even, so the sign is +),

$$\int_{S^n} \widehat{\mathrm{Ev}}^* TP_j = \int_{S^n} \mathrm{Ev}^*\omega_{2j-1} + \delta\Big(\int_{S^n} \beta_j\Big).$$

The new μ and the old μ differ by an exact form: *they are the same cohomology class*, as promised. The new one is better for physics: it is an explicit functional of the gauge field.

Step 5: the descent expansion

To read off the form μ_{2j-1-n} explicitly, fix a point (x,α) and abbreviate $B := A^\alpha$ (a $(1,0)$ form) and ϑ (a $(0,1)$ form). By (8.132)–(8.133), $\mathbb{A} = B + \vartheta$ and $\mathbb{F} = F_B$, the curvature of B. The transgression formula needs $\mathbb{F}_t = t\,\mathrm{d}_{\mathrm{tot}}\mathbb{A} + t^2\mathbb{A}^2$, and we rewrite it using $\mathrm{d}_{\mathrm{tot}}\mathbb{A} = \mathbb{F} - \mathbb{A}^2$:

$$\mathbb{F}_t = t\,\mathbb{F} + (t^2 - t)\,\mathbb{A}^2 = t\,F_B + (t^2-t)\big(B^2 + B\vartheta + \vartheta B + \vartheta^2\big). \tag{8.137}$$

This is pointwise algebra: no derivatives of ϑ appear. Now expand the transgression form $TP_j(\mathbb{A}) = j\int_0^1 P_j(\mathbb{A}, \mathbb{F}_t, \dots)\,\mathrm{d}t$ in powers of ϑ:

$$TP_j(\mathbb{A}) = \sum_{k=0}^{2j-1} \Omega^{(k)}_{2j-1-k}, \qquad \Omega^{(k)}_{2j-1-k} \text{ of bidegree } (2j-1-k,\ k). \tag{8.138}$$

The superscript counts the factors of ϑ (the legs along $\mathcal{G}$), the subscript the legs along S^n. Integration over S^n keeps only the terms with exactly n legs along S^n, i.e. $k = 2j-1-n$:

$$\boxed{\mu_{2j-1-n}\Big|_\alpha = \int_{S^n} \Omega_n^{(2j-1-n)}\big(A^\alpha, \vartheta\big).} \tag{8.139}$$

The $k=0$ term is just $TP_j(A^\alpha)$, the Chern–Simons form of the transformed field.

Dictionary: the gauge group Maurer–Cartan form is the ghost

Compute how the two pieces of $\mathbb{A}$ change along $\mathcal{G}$. With $\delta\alpha = \alpha\vartheta$, $\delta(\alpha^{-1}) = -\vartheta\alpha^{-1}$ and $\delta\mathrm{d} = -\mathrm{d}\delta$:

$$\delta\vartheta \overset{(1)}{=} \delta(\alpha^{-1})\,\delta\alpha \overset{(2)}{=} -\vartheta\,\alpha^{-1}\alpha\,\vartheta = -\vartheta^2,$$
$$\delta(\alpha^{-1}\mathrm{d}\alpha) \overset{(3)}{=} -\vartheta\,\alpha^{-1}\mathrm{d}\alpha - \alpha^{-1}\mathrm{d}(\alpha\vartheta) \overset{(4)}{=} -\vartheta\,\theta - \theta\,\vartheta - \mathrm{d}\vartheta, \qquad \theta := \alpha^{-1}\mathrm{d}\alpha,$$
$$\delta(\alpha^{-1}A\alpha) \overset{(5)}{=} -\vartheta\,\alpha^{-1}A\alpha - \alpha^{-1}A\,\delta\alpha = -\vartheta\,(\alpha^{-1}A\alpha) - (\alpha^{-1}A\alpha)\,\vartheta.$$

(1) Leibniz rule; $\delta^2\alpha = 0$.
(2) $\delta(\alpha^{-1}) = -\alpha^{-1}(\delta\alpha)\alpha^{-1}$ and $\delta\alpha = \alpha\vartheta$.
(3) Leibniz rule for δ; $\delta\mathrm{d}\alpha = -\mathrm{d}\delta\alpha = -\mathrm{d}(\alpha\vartheta)$.
(4) $\mathrm{d}(\alpha\vartheta) = \mathrm{d}\alpha\,\vartheta + \alpha\,\mathrm{d}\vartheta$ (α has degree 0).
(5) $\delta A = 0$; moving δ past the 1-form A costs a sign; then insert $\alpha\alpha^{-1}$.

Adding the last two lines, with $B = A^\alpha$ and the graded commutator $[B,\vartheta] = B\vartheta + \vartheta B$ of (8.7),

$$\delta B = -\big(\mathrm{d}\vartheta + [B,\vartheta]\big) = -D_B\vartheta, \qquad \delta\vartheta = -\vartheta^2.$$

These are the **BRST transformations** of a gauge field and of the Faddeev–Popov ghost (in one common sign convention), with $\delta \leftrightarrow s$ (the BRST operator) and $\vartheta \leftrightarrow c$ (the ghost). The ghosts that Section 8.1.10 introduced as a trick to exponentiate a Jacobian are, geometrically, the Maurer–Cartan form of the gauge group. The expansion (8.138) in powers of ϑ is what physicists call the expansion in **ghost number**, and the associated identities are the **Stora–Zumino descent equations** (Chapter 10).

Worked example 8.36: $n = 2$, $G = \mathrm{SU}(2)$: the form μ_1 explicitly

On S^2 the first class is μ_{2j-3} with $j = 2$, i.e. μ_1 built from TP_2 (Figure 8.23). We need the part of $TP_2(\mathbb{A})$ with one ϑ and two $\mathrm{d}x$-legs. Write $P_2(X,Y) = \kappa\ \mathrm{tr}(XY)$ with $\kappa = -\frac{1}{8\pi^2}$ (Table 8.3); for X odd and Y even the symmetrised trace is just $\mathrm{tr}(XY)$, since $\mathrm{tr}(YX) = \mathrm{tr}(XY)$.

$$TP_2(\mathbb{A}) \overset{(1)}{=} 2\kappa \int_0^1 \mathrm{d}t\ \mathrm{tr}\Big((B+\vartheta)\big(tF_B + (t^2-t)(B^2 + B\vartheta + \vartheta B + \vartheta^2)\big)\Big)$$
$$\Omega_2^{(1)} \overset{(2)}{=} 2\kappa \int_0^1 \mathrm{d}t\,\Big[t\ \mathrm{tr}(\vartheta F_B) + (t^2-t)\big(\mathrm{tr}(\vartheta B^2) + \mathrm{tr}(B^2\vartheta) + \mathrm{tr}(B\vartheta B)\big)\Big]$$
$$\overset{(3)}{=} 2\kappa \int_0^1 \mathrm{d}t\,\Big[t\ \mathrm{tr}(\vartheta F_B) + 3(t^2-t)\ \mathrm{tr}(\vartheta B^2)\Big]$$
$$\overset{(4)}{=} 2\kappa\Big[\tfrac{1}{2}\,\mathrm{tr}(\vartheta F_B) - \tfrac{1}{2}\,\mathrm{tr}(\vartheta B^2)\Big] \overset{(5)}{=} \kappa\ \mathrm{tr}(\vartheta\,\mathrm{d}B).$$

(1) Transgression formula (8.113) with (8.137) for $j = 2$.
(2) Keep terms with exactly one ϑ: ϑ from the first slot times $tF_B + (t^2-t)B^2$, or B from the first slot times $(t^2-t)(B\vartheta + \vartheta B)$.
(3) Cyclicity of the graded trace: moving a factor of degree a from the end to the front past a product of degree b costs $(-1)^{ab}$. Moving ϑ past B^2 (degree 2) and B past $B\vartheta$ (degree 2) costs nothing, so all three traces equal $\mathrm{tr}(\vartheta B^2)$.
(4) $\int_0^1 t\,\mathrm{d}t = \frac{1}{2}$ and $\int_0^1 (t^2-t)\,\mathrm{d}t = -\frac{1}{6}$.
(5) $F_B - B^2 = \mathrm{d}B$.

So, feeding in a tangent vector ξ at $\alpha \in \mathcal{G}$ (with $\vartheta(\xi) = \xi$, and putting the $\mathrm{d}x$-legs first, which costs no sign since $\mathrm{d}B$ is even),

$$\mu_1\big|_\alpha(\xi) = -\frac{1}{8\pi^2}\int_{S^2} \mathrm{tr}\big(\xi\, \mathrm{d}A^\alpha\big). \tag{8.140}$$

Check against (8.89). Take $A = 0$, so $B = \alpha^{-1}\mathrm{d}\alpha = \theta$ and $\mathrm{d}B = -\theta^2$ (Maurer–Cartan). Then $\Omega_2^{(1)} = -\kappa\,\mathrm{tr}(\vartheta\theta^2) = \frac{1}{8\pi^2}\,\mathrm{tr}(\vartheta\theta^2)$. On the other hand the Maurer–Cartan form of G pulled back by Ev is $\mathrm{Ev}^*(g^{-1}\mathrm{d}g) = \theta + \vartheta$, and the $(2,1)$ part of $\omega_3 = \frac{1}{24\pi^2}\,\mathrm{tr}(\theta+\vartheta)^3$ is $\frac{3}{24\pi^2}\,\mathrm{tr}(\vartheta\theta^2) = \frac{1}{8\pi^2}\,\mathrm{tr}(\vartheta\theta^2)$ (three placements of ϑ, all equal by cyclicity). The two agree on the nose, as (8.136) predicts (for $A = 0$ the correction β_j vanishes).

Worked example 8.37: A loop of gauge transformations on S^2 with $\oint \mu_1 = \pm 1$

For $M = S^2$ and $G = \mathrm{SU}(2)$, $\pi_1(\mathcal{G}) = \pi_{1+2}(\mathrm{SU}(2)) = \pi_3(S^3) = \mathbb{Z}$ (Eq. (8.56)). Let us exhibit a generator and evaluate μ_1 on it. Identify SU(2) with the unit quaternions, with imaginary units $e_a = -i\sigma_a$, and write a point of S^2 as a unit vector $\hat{n}$, with $\hat{n}\cdot e := n^a e_a$, so that $(\hat{n}\cdot e)^2 = -\mathbb{1}$. For any angle ϕ, $\exp(\phi\,\hat{n}\cdot e) = \cos\phi\,\mathbb{1} + \sin\phi\,\hat{n}\cdot e$. Take the base point $\hat{n}_0$ to be the north pole and define, for $\tau \in [0,1]$,

$$\alpha_\tau(\hat{n}) = \exp\big(\pi\tau\,\hat{n}\cdot e\big)\,\exp\big(-\pi\tau\,\hat{n}_0\cdot e\big).$$

- *Based:* at $\hat{n} = \hat{n}_0$ the two factors cancel, $\alpha_\tau(\hat{n}_0) = \mathbb{1}$.
- *A loop:* $\alpha_0 = \mathbb{1}$, and at $\tau = 1$ each factor is $\cos\pi = -\mathbb{1}$, so $\alpha_1 = (-\mathbb{1})(-\mathbb{1}) = \mathbb{1}$.
- *Degree one:* by (8.91), $\oint \mu_1 = \int_{S^2\times S^1} F^*\omega_3$ with $F(\hat{n},\tau) = \alpha_\tau(\hat{n})$, which is the degree of F as a map to S^3. Count preimages of the point $-\mathbb{1}$: $\alpha_\tau(\hat{n}) = -\mathbb{1}$ means $\exp(\pi\tau\,\hat{n}\cdot e) = -\exp(\pi\tau\,\hat{n}_0\cdot e) = \exp\big(\pi(1-\tau)(-\hat{n}_0)\cdot e\big)$. For $\tau \in (0,1)$ both angles lie in $(0,\pi)$, where the polar form $\cos\phi + \sin\phi\,\hat{m}\cdot e$ is unique; so $\tau = 1-\tau$ and $\hat{n} = -\hat{n}_0$. There is exactly one preimage, $(\hat{n},\tau) = (\text{south pole}, \frac{1}{2})$, so the degree is ± 1.

A direct numerical evaluation of $\frac{1}{24\pi^2}\int \mathrm{tr}\,\theta^3$ with the orientation (polar angle, azimuth, τ) of $S^2\times S^1$ gives -1.000. For comparison, the unbased loop $\tau \mapsto \exp(2\pi\tau\,\hat{n}\cdot e)$ covers S^3 twice (once for $\tau < \frac{1}{2}$, once for $\tau > \frac{1}{2}$ with $\hat{n} \to -\hat{n}$), and gives -2. Physically: in two-dimensional SU(2) gauge theory on a sphere there are closed loops of gauge transformations that cannot be shrunk, and μ_1, the explicit functional (8.140), counts how many times a loop winds.

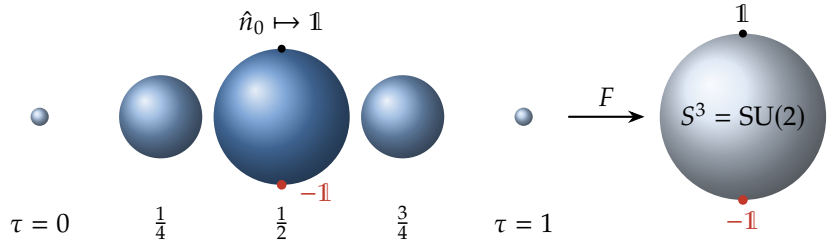


Figure 8.24: The loop of Example 8.37 as a map $F : S^2\times S^1 \to S^3$. Each slice τ is a sphere S^2 mapped into SU(2); the image sphere grows from the point $\mathbb{1}$ at $\tau = 0$, sweeps across S^3, and shrinks back to $\mathbb{1}$ at $\tau = 1$, while the base point $\hat{n}_0$ is always sent to $\mathbb{1}$. Only one point, the south pole at $\tau = \frac{1}{2}$ (red), goes to $-\mathbb{1}$: the map has degree ± 1, so $\oint \mu_1 = \pm 1$ and the loop generates $\pi_1(\mathcal{G}) = \mathbb{Z}$.

Step 6: Nash's example, μ_1 on S^4

Nash's own example (his (8.54)–(8.58)) is the physically important case: four dimensions, $n = 4$, $G = \mathrm{SU}(N)$. By (8.135) the class μ_1 comes from

$$2j - 1 - n = 1 \quad \Longrightarrow \quad j = \frac{2+n}{2} = 3, \tag{8.141}$$

that is, from the third Chern class and its Chern–Simons form (Nash's (8.56))

$$TP_3(A) = 3\int_0^1 \mathrm{d}t\, P_3(A, F_t, F_t).$$

The invariant polynomial (Nash's (8.57)) is the $\mathrm{SU}(N)$ Chern class c_3 of (8.106):

$$P_3(F,F,F) = \frac{1}{3}\Big(\frac{i}{2\pi}\Big)^3 \operatorname{tr} F^3 \overset{(1)}{=} \frac{1}{3}\cdot\frac{-i}{8\pi^3}\operatorname{tr} F^3 = -\frac{i}{24\pi^3}\operatorname{tr}(F\wedge F\wedge F), \tag{8.142}$$

(1) $(i/2\pi)^3 = i^3/(8\pi^3) = -i/(8\pi^3)$.

and its polarisation is $P_3(X,Y,Z) = c\,\mathrm{STr}(X,Y,Z)$ with $c := -\frac{i}{24\pi^3}$, where STr is the graded symmetrised trace (Section 8.2.2). Putting the pieces together (Nash's (8.58)):

$$\mu_1 = 3c\int_0^1 \mathrm{d}t \int_{S^4} \operatorname{tr}\big(\mathbb{A}\wedge\mathbb{F}_t\wedge\mathbb{F}_t\big) = -\frac{i}{(2\pi)^3}\int_0^1 \mathrm{d}t\int_{S^4}\operatorname{tr}\big(\mathbb{A}\wedge\mathbb{F}_t\wedge\mathbb{F}_t\big), \tag{8.143}$$

since $3c = -\frac{i}{8\pi^3} = -\frac{i}{(2\pi)^3}$. Here the first step uses that in $P_3(\mathbb{A},\mathbb{F}_t,\mathbb{F}_t)$, with $\mathbb{F}_t$ of even degree, all orderings are cyclically equal to $\operatorname{tr}(\mathbb{A}\mathbb{F}_t\mathbb{F}_t)$.

Watch out: How to read Nash's (8.58)

Nash writes (8.58) with "A and F the connection and curvature on S^4". Read literally this is zero: $\operatorname{tr}(A\wedge F_t\wedge F_t)$ would be a 5-form on a 4-manifold. The A in (8.58) must be the pulled-back connection $\mathbb{A} = A^\alpha + \vartheta$ on $S^4\times\mathcal{G}$ of (8.132), and F_t is built from it by (8.137). The integral over S^4 then picks out the part with one ϑ, which is a 1-form on $\mathcal{G}$. That is what we now compute.

The computation is the $n = 2$ one of Example 8.36, one step longer. Write $Y := tF_B + s\,B^2$ and $C := s\,(B\vartheta + \vartheta B)$ with $s := t^2 - t$, so that $\mathbb{F}_t = Y + C + s\,\vartheta^2$. The terms of $P_3(\mathbb{A},\mathbb{F}_t,\mathbb{F}_t)$ with one ϑ are

$$P_3(\vartheta,Y,Y) + 2P_3(B,Y,C) = c\Big[\operatorname{tr}(\vartheta Y^2) + \operatorname{tr}(BYC) + \operatorname{tr}(BCY)\Big].$$

(For $P_3(B,Y,C)$: of the six orderings, three are cyclically equal to $\operatorname{tr}(BYC)$ and three to $\operatorname{tr}(BCY)$, all with sign + because Y and C are even; the $\frac{1}{6}$ of the symmetrisation and the factor 2 combine to 1.) Every trace can be brought to the form $\operatorname{tr}(\vartheta W)$ with W a 4-form built from B and F_B: moving a block of degree a from the back to the front past a block of degree b costs $(-1)^{ab}$, and one checks that every move needed here has ab even. Collecting by monomials:

monomial	from $\operatorname{tr}(\vartheta Y^2)$	from $\operatorname{tr}(BYC)+\operatorname{tr}(BCY)$	$3\int_0^1(\cdot)\,\mathrm{d}t$
$\operatorname{tr}(\vartheta F_B^2)$	t^2	–	1
$\operatorname{tr}(\vartheta F_B B^2)$	ts	ts	$-\frac{1}{2}$
$\operatorname{tr}(\vartheta B^2 F_B)$	ts	ts	$-\frac{1}{2}$
$\operatorname{tr}(\vartheta B F_B B)$	–	$2ts$	$-\frac{1}{2}$
$\operatorname{tr}(\vartheta B^4)$	s^2	$4s^2$	$\frac{1}{2}$

using $\int_0^1 t^2\,\mathrm{d}t = \frac{1}{3}$, $\int_0^1 ts\,\mathrm{d}t = \frac{1}{4} - \frac{1}{3} = -\frac{1}{12}$ and $\int_0^1 s^2\,\mathrm{d}t = \frac{1}{5} - \frac{1}{2} + \frac{1}{3} = \frac{1}{30}$. Hence

$$\begin{aligned}\frac{1}{c}\,\Omega_4^{(1)} &\overset{(1)}{=} \operatorname{tr}\Big(\vartheta\big[F_B^2 - \tfrac{1}{2}F_B B^2 - \tfrac{1}{2}B^2F_B - \tfrac{1}{2}BF_BB + \tfrac{1}{2}B^4\big]\Big)\\ &\overset{(2)}{=} \operatorname{tr}\Big(\vartheta\big[(\mathrm{d}B)^2 + \tfrac{1}{2}\big(\mathrm{d}B\,B^2 - B\,\mathrm{d}B\,B + B^2\,\mathrm{d}B\big) + \big(1 - \tfrac{1}{2} - \tfrac{1}{2} - \tfrac{1}{2} + \tfrac{1}{2}\big)B^4\big]\Big)\\ &\overset{(3)}{=} \operatorname{tr}\Big(\vartheta\,\mathrm{d}\big(B\,\mathrm{d}B + \tfrac{1}{2}B^3\big)\Big).\end{aligned} \tag{8.144}$$

(1) The table.
(2) Substitute $F_B = \mathrm{d}B + B^2$: $F_B^2 = (\mathrm{d}B)^2 + \mathrm{d}B\,B^2 + B^2\mathrm{d}B + B^4$, $F_BB^2 = \mathrm{d}B\,B^2 + B^4$, $B^2F_B = B^2\mathrm{d}B + B^4$, $BF_BB = B\,\mathrm{d}B\,B + B^4$, and collect.
(3) $\mathrm{d}(B\,\mathrm{d}B) = (\mathrm{d}B)^2$ and $\mathrm{d}(B^3) = \mathrm{d}B\,B^2 - B\,\mathrm{d}B\,B + B^2\,\mathrm{d}B$ (graded Leibniz rule); the B^4 terms cancel.

(A computer-algebra expansion of $3\int_0^1 P_3(\mathbb{A}, \mathbb{F}_t, \mathbb{F}_t)\,\mathrm{d}t$ in a free graded algebra confirms (8.144), and also confirms that the ϑ-free part is $c\ \operatorname{tr}(B(\mathrm{d}B)^2 + \frac{3}{2}B^3\mathrm{d}B + \frac{3}{5}B^5)$, the Chern–Simons 5-form (8.117).)

Key idea: The explicit μ_1 on S^4

For $G = \mathrm{SU}(N)$ on S^4, at the point $\alpha \in \mathcal{G}$ and on the infinitesimal gauge transformation ξ,

$$\mu_1\big|_\alpha(\xi) = -\frac{i}{24\pi^3}\int_{S^4} \operatorname{tr}\Big(\xi\,\mathrm{d}\big(A^\alpha\,\mathrm{d}A^\alpha + \tfrac{1}{2}(A^\alpha)^3\big)\Big). \tag{8.145}$$

It is linear in ξ, as a 1-form must be, and it has the famous structure $\operatorname{tr}\,\xi\,\mathrm{d}(A\,\mathrm{d}A + \frac{1}{2}A^3)$ of the *consistent non-Abelian anomaly* of a chiral fermion in four dimensions. In Chapter 10 the anomaly is found from Feynman diagrams and is identified, up to a factor involving 2π and the normalisation of the generators, with this class.

Two checks.

1. *Pure gauge, against* (8.89). For $A = 0$, $B = \theta$ and $\mathrm{d}\theta = -\theta^2$:

$$\mathrm{d}\big(\theta\,\mathrm{d}\theta + \tfrac{1}{2}\theta^3\big) \overset{(1)}{=} \mathrm{d}\big(-\theta^3 + \tfrac{1}{2}\theta^3\big) = -\tfrac{1}{2}\,\mathrm{d}(\theta^3) \overset{(2)}{=} -\tfrac{1}{2}\big(\mathrm{d}\theta\,\theta^2 - \theta\,\mathrm{d}\theta\,\theta + \theta^2\mathrm{d}\theta\big) \overset{(3)}{=} \tfrac{1}{2}\,\theta^4.$$

   (1) Maurer–Cartan, $\mathrm{d}\theta = -\theta^2$.
   (2) Graded Leibniz rule for the three factors of θ^3.
   (3) Each $\mathrm{d}\theta$ becomes $-\theta^2$: $-\frac{1}{2}(-\theta^4 + \theta^4 - \theta^4) = \frac{1}{2}\theta^4$.

   So $\Omega_4^{(1)} = \frac{c}{2}\operatorname{tr}(\vartheta\theta^4) = -\frac{i}{48\pi^3}\operatorname{tr}(\vartheta\theta^4)$. The $(4,1)$ part of $\mathrm{Ev}^*\omega_5 = C_3\operatorname{tr}(\theta+\vartheta)^5$ is $5C_3\operatorname{tr}(\vartheta\theta^4) = -\frac{5i}{240\pi^3}\operatorname{tr}(\vartheta\theta^4) = -\frac{i}{48\pi^3}\operatorname{tr}(\vartheta\theta^4)$, with C_3 from (8.87). They agree exactly.

2. *Abelian shadow.* If all matrices commuted (a U(1) factor), $B^2 = 0$, $\mathbb{F}_t = tF$ and $TP_3 = c\,(B + \vartheta)F^2$, whose one-$\vartheta$ part is $c\,\vartheta(\mathrm{d}B)^2 = c\,\vartheta\,\mathrm{d}(B\,\mathrm{d}B)$: the non-Abelian formula with the B^3 term dropped.

Try it

Show that μ_1 vanishes identically for $G = \mathrm{SU}(2)$ on S^4, in agreement with $\pi_1(\mathcal{G}) \otimes \mathbb{Q} = \pi_5(\mathrm{SU}(2)) \otimes \mathbb{Q} = \mathbb{Z}_2 \otimes \mathbb{Q} = 0$.
Answer. P_3 is built from the totally symmetric tensor $d_{abc} \propto \operatorname{tr}(e_a\{e_b, e_c\})$. For $\mathfrak{su}(2)$,

$\{e_b, e_c\} = -2\delta_{bc}\mathbb{1}$, so $d_{abc} \propto \delta_{bc}\,\mathrm{tr}\,e_a = 0$. Every symmetrised trace of three $\mathfrak{su}(2)$-valued forms vanishes, hence $TP_3 = 0$. The torsion class of $\pi_5(\mathrm{SU}(2)) = \mathbb{Z}_2$ is invisible to de Rham cohomology (Example 8.21); it resurfaces as the global SU(2) anomaly (Chapter 10).

Watch out: Non-trivial bundles

For the instanton bundles P_k over S^4 with $k \neq 0$ there is no global section, so (8.129) is only defined over each hemisphere. The recipe still works: on the overlap the two local versions of $\Omega_4^{(1)}$ differ by an exact form plus terms that cancel between the two hemispheres, and (8.145) holds on each patch with the local potential. Since the class μ_1 is defined without sections by (8.89) and by the homotopy equivalence $\mathcal{G}_k \simeq \mathcal{G}_0$ (Eq. (8.50)), nothing depends on these choices.

8.2.9 The classes ν on $\mathcal{A}/\mathcal{G}$ and the double transgression

The ν's live on the configuration space $B\mathcal{G} = \mathcal{A}/\mathcal{G}$. Nash relates them to the μ's through the principal bundle (Nash's (8.51), our (8.35))

$$\mathcal{G} \longrightarrow \mathcal{A} \xrightarrow{\pi} \mathcal{A}/\mathcal{G} = B\mathcal{G}.$$

This is the same shape as the universal bundle $G \to EG \to BG$: a group, a contractible total space, a quotient. For $G \to P \to M$ we saw (Section 8.2.5) that a class on the base pulled up to the total space becomes exact, and that its primitive restricted to a fibre is a class on the group. The same happens here, one level up.

Step 1: pull back to $\mathcal{A}$, where everything is exact

Pull $\nu_{2j} \in H^{2j}_{\mathrm{dR}}(B\mathcal{G})$ back to $\mathcal{A}$. The space $\mathcal{A}$ is affine, hence contractible ((8.14)), and on a contractible space every closed form of positive degree is exact. (Proof of the Poincaré lemma on an affine space: the straight-line homotopy $A_s = A_0 + s(A - A_0)$ gives an explicit primitive, exactly as in the derivation of the transgression formula.) So (Nash's (8.52))

$$\pi^*\nu_{2j} = \delta f_j \qquad \text{for some } f_j \in \Omega^{2j-1}(\mathcal{A}). \tag{8.146}$$

Restrict f_j to one fibre of π, i.e. to one gauge orbit $\mathcal{G}\cdot A$. On the orbit π is constant, so $\pi^*\nu_{2j}$ restricts to zero (the pull-back of a form of positive degree by a constant map vanishes). Hence

$$\delta\big(f_j|_{\mathcal{G}\cdot A}\big) = \big(\pi^*\nu_{2j}\big)\big|_{\mathcal{G}\cdot A} = 0:$$

f_j restricted to an orbit is a closed $(2j-1)$-form on $\mathcal{G}$. Nash's claim (8.53) is that this is μ_{2j-1}. To see it, we make f_j explicit.

Step 2: the universal connection on $S^n \times \mathcal{A}$

Over $S^n \times \mathcal{A}$ there is a tautological bundle with a tautological connection: over the point (x, A) put the fibre P_x, and let the connection in the S^n directions be A itself. What connection should we use in the $\mathcal{A}$ directions? The answer is dictated by the requirement that everything descend to $S^n \times \mathcal{A}/\mathcal{G}$: moving *along* a gauge orbit should be undone by the corresponding gauge transformation. That requires a way of splitting a tangent vector $a \in T_A\mathcal{A} = \Omega^1(M;\mathfrak{g})$ into a part along the orbit and a part across it, i.e. a connection on the principal bundle $\mathcal{A} \to \mathcal{A}/\mathcal{G}$.

Working definition 8.19 (the Coulomb connection on $\mathcal{A} \to \mathcal{A}/\mathcal{G}$)

At $A \in \mathcal{A}$ the orbit directions are the infinitesimal gauge transformations $a = D_A\xi$ (Eq. (8.41)). Declare the *horizontal* directions to be those orthogonal to the orbit in the L^2 inner product [the integral over M of the pointwise inner product of two $\mathfrak{g}$-valued 1-forms], i.e. those with $D_A^* a = 0$ [D_A^* is the adjoint of D_A for this inner product]: the *background Coulomb gauge* condition. The connection 1-form is the Lie($\mathcal{G}$)-valued [its value on a is an infinitesimal gauge transformation ξ] 1-form on $\mathcal{A}$

$$\Theta_A(a) := \left(D_A^* D_A\right)^{-1} D_A^*\, a. \tag{8.147}$$

To use it: to split a variation a of the gauge field, solve the Poisson-type equation [a Laplace equation for ξ. It has exactly one solution when A is irreducible, since then $D_A\xi = 0$ only for $\xi = 0$] $D_A^* D_A \xi = D_A^* a$ for ξ; then $a = D_A\xi + (a - D_A\xi)$ is the vertical plus horizontal decomposition, and $\Theta(a) = \xi$.

It reproduces the generators of the orbit, which is the defining property of a connection form:

$$\Theta_A(D_A\xi) \overset{(1)}{=} \left(D_A^* D_A\right)^{-1}\left(D_A^* D_A\right)\xi \overset{(2)}{=} \xi.$$

(1) Definition (8.147) with $a = D_A\xi$.
(2) $D_A^* D_A$ is invertible when A is irreducible (no covariantly constant ξ; Example 8.3), which is where the gauge group acts freely.

The horizontal subspace $\{D_A^* a = 0\}$ is the tangent space of the local Coulomb slice of Section 8.1.11. That the slice does not extend globally (Gribov) is the statement that the bundle $\mathcal{A} \to \mathcal{A}/\mathcal{G}$ is non-trivial; the connection Θ exists globally all the same, and its *curvature* is what measures the twisting. Figure 8.25 shows the geometry.

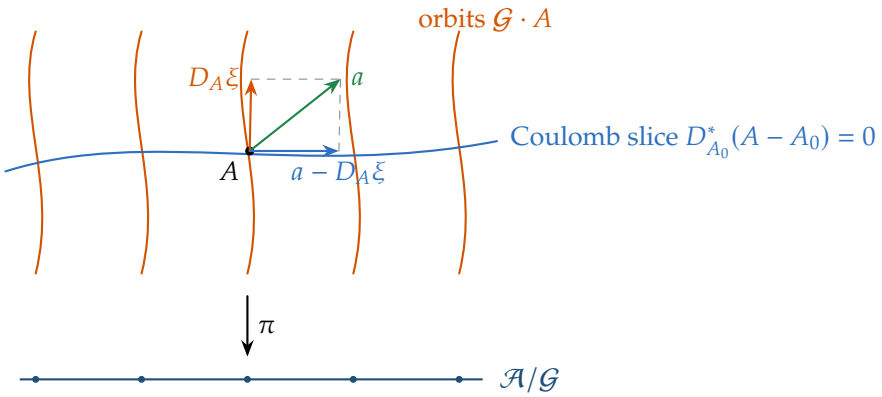


Figure 8.25: The Coulomb connection on $\mathcal{A} \to \mathcal{A}/\mathcal{G}$. Each orbit (orange) projects to a single point of $\mathcal{A}/\mathcal{G}$. A variation a of the gauge field at A (green) splits into a piece $D_A\xi$ along the orbit, which is a gauge transformation and changes nothing physical, and a horizontal piece orthogonal to the orbit (blue). The connection form Θ returns ξ. Horizontal directions are tangent to the Coulomb slice.

The universal connection and its curvature. On $S^n \times \mathcal{A}$ define

$$\mathbb{A}_{\text{univ}} := A + \Theta, \tag{8.148}$$

where A is the tautological $(1,0)$ form (at the point (x, A) it is the value of A at x) and Θ is the $(0,1)$ form (8.147), a $\mathfrak{g}$-valued function of x for each tangent vector. Its curvature, sorted by

bidegree, is

$$\begin{aligned}\mathbb{F}_{\text{univ}} &\overset{(1)}{=} (\mathrm{d}+\delta)(A+\Theta)+(A+\Theta)^2 \\ &\overset{(2)}{=} \underbrace{\mathrm{d}A+A^2}_{(2,0):\,F} + \underbrace{\delta A+\mathrm{d}\Theta+A\Theta+\Theta A}_{(1,1):\,\delta A+D_A\Theta} + \underbrace{\delta\Theta+\Theta^2}_{(0,2):\,\Omega_\Theta}.\end{aligned} \tag{8.149}$$

(1) Definition of curvature with $\mathrm{d}_{\text{tot}} = \mathrm{d}+\delta$.
(2) Expand and sort: $\mathrm{d}A$ and A^2 have two x-legs; δA, $\mathrm{d}\Theta$ and $A\Theta+\Theta A$ have one of each; $\delta\Theta$ and Θ^2 have two $\mathcal{A}$-legs. $D_A\Theta = \mathrm{d}\Theta + [A,\Theta]$ with the graded bracket.

Each piece has a meaning. The $(2,0)$ part is the field strength. The $(0,2)$ part Ω_Θ is the curvature of the Coulomb connection on $\mathcal{A}\to\mathcal{A}/\mathcal{G}$. The $(1,1)$ part, evaluated on a tangent vector a, is (up to the sign conventions of the bigrading) the *horizontal projection* $a - D_A\Theta(a)$. In particular it vanishes on vertical vectors. To check this, restrict to an orbit with the orbit map $o_A:\mathcal{G}\to\mathcal{A}$, $\alpha\mapsto A^\alpha$. By the dictionary box of Section 8.2.8, $o_A^*(\delta A) = \delta(A^\alpha) = -D_{A^\alpha}\vartheta$, and by the defining property of Θ, $o_A^*\Theta = \vartheta$ (the tangent vector to the orbit generated by ξ is $D_{A^\alpha}\xi$, and Θ returns ξ). So

$$\begin{aligned}(\mathrm{id}\times o_A)^*\,\mathbb{A}_{\text{univ}} &\overset{(1)}{=} A^\alpha+\vartheta \overset{(2)}{=} \mathbb{A}, \\ (\mathrm{id}\times o_A)^*\big(\delta A + D_A\Theta\big) &\overset{(3)}{=} -D_{A^\alpha}\vartheta + D_{A^\alpha}\vartheta = 0.\end{aligned} \tag{8.150}$$

(1) The tautological A pulls back to A^α and Θ to ϑ.
(2) This is precisely the connection $\widehat{\mathrm{Ev}}^*\,\omega$ of (8.132).
(3) $o_A^*\delta A = -D_{A^\alpha}\vartheta$ from the dictionary box, and $o_A^*(D_A\Theta) = D_{A^\alpha}\vartheta$.

On an orbit the universal connection is the connection we used to define μ, and its curvature is purely $(2,0)$, in agreement with (8.133).

Step 3: the explicit f_j and Nash's (8.53)

The universal curvature is basic (it descends to $S^n\times\mathcal{A}/\mathcal{G}$, because Θ is a genuine connection), so the Chern–Weil form of the universal bundle over $S^n\times B\mathcal{G}$ pulled back to $S^n\times\mathcal{A}$ is $P_J(\mathbb{F}_{\text{univ}})$. With $J := j+\frac{n}{2}$, so that $2J-n = 2j$, the class ν_{2j} of (8.97) is represented on $\mathcal{A}$ by

$$\pi^*\nu_{2j} \overset{(1)}{=} \int_{S^n} P_J(\mathbb{F}_{\text{univ}},\dots,\mathbb{F}_{\text{univ}}) \overset{(2)}{=} \int_{S^n} \mathrm{d}_{\text{tot}} TP_J(\mathbb{A}_{\text{univ}}) \overset{(3)}{=} \delta\int_{S^n} TP_J(\mathbb{A}_{\text{univ}}). \tag{8.151}$$

(1) Chern–Weil for the universal bundle: c_J is represented by $P_J(\mathbb{F})$, and $\nu_{2j} = \int_{S^n}\mathrm{Ev}^*c_J$ by definition.
(2) The bundle over $S^n\times\mathcal{A}$ is pulled back from S^n (it is pr^*P), so the transgression formula applies there with a section: a section s of P over S^n (trivial bundle, as in Section 8.2.8) is also one over $S^n\times\mathcal{A}$, and pulled back by it, $P_J(\mathbb{F}) = \mathrm{d}_{\text{tot}}TP_J(\mathbb{A})$ as in (8.115).
(3) (8.90) with n even.

So an explicit primitive is

$$f_j = \int_{S^n} TP_{j+n/2}\big(A+\Theta\big), \tag{8.152}$$

and restricted to an orbit, by (8.150),

$$\boxed{f_j\big|_{\mathcal{G}\cdot A} = \int_{S^n} TP_{j+n/2}\big(A^\alpha+\vartheta\big) = \int_{S^n}\widehat{\mathrm{Ev}}^*TP_{j+n/2} = \mu_{2j-1}, \qquad \pi^*\nu_{2j} = \delta f_j.} \tag{8.153}$$

This is Nash's (8.53), with the "adjusted indices": the class $\mu_{2(j+n/2)-1-n}$ of (8.130) is μ_{2j-1}. With the Coulomb connection the coincidence is exact, not merely in cohomology. Any other primitive f_j' of $\pi^*\nu_{2j}$ differs from f_j by a closed, hence exact, form on the contractible $\mathcal{A}$, and restricts to a cohomologous form on the orbit; this is Nash's "coincides cohomologically".

$$\begin{array}{ccccccc} H^{2j-1}_{\mathrm{dR}}(\mathcal{G}) \ni \mu_{2j-1} & \xleftarrow{\text{orbit}} & f_j \in \Omega^{2j-1}(\mathcal{A}) & \xrightarrow{\delta} & \pi^*\nu_{2j} \in \Omega^{2j}(\mathcal{A}) & \xleftarrow{\pi^*} & \nu_{2j} \in H^{2j}_{\mathrm{dR}}(\mathcal{A}/\mathcal{G}) \\ \int_{S^n}\mathrm{Ev}^* \uparrow & & \int_{S^n}\uparrow & & & & \uparrow \int_{S^n}\mathrm{Ev}^* \\ \omega_{2J-1} \in H^{2J-1}_{\mathrm{dR}}(G) & \xleftarrow{\text{fibre}} & TP_J \in \Omega^{2J-1}(P) & \xrightarrow{\mathrm{d}} & \pi^*c_J \in \Omega^{2J}(P) & \xleftarrow{\pi^*} & c_J \in H^{2J}(BG) \end{array}$$

Figure 8.26: The double transgression, with $J = j + \frac{n}{2}$. Bottom row: the transgression for $G \to P \to M$ (Section 8.2.5); the Chern–Simons form TP_J has $\mathrm{d}TP_J = \pi^*c_J$ and restricts to the generator ω_{2J-1} on a fibre. Top row: the same pattern for $\mathcal{G} \to \mathcal{A} \to \mathcal{A}/\mathcal{G}$. Horizontal arrows to the left are restriction to an orbit or a fibre. The vertical arrows are "evaluate and integrate over S^n", which lowers every degree by n. Read the diagram as: *integration over S^n intertwines the two transgressions.*

Step 4: exactness of μ and of ν go together

Nash ends with: if μ_1 is exact the same is true of ν_2, and conversely. One direction is a short chain. Suppose $\nu_{2j} = \delta h$ on $\mathcal{A}/\mathcal{G}$. Then

$$\delta\big(f_j - \pi^*h\big) \overset{(1)}{=} \pi^*\nu_{2j} - \pi^*\nu_{2j} = 0, \qquad f_j - \pi^*h \overset{(2)}{=} \delta k,$$
$$\mu_{2j-1} \overset{(3)}{=} f_j\big|_{\mathcal{G}\cdot A} \overset{(4)}{=} \big(\pi^*h + \delta k\big)\big|_{\mathcal{G}\cdot A} \overset{(5)}{=} \delta\big(k|_{\mathcal{G}\cdot A}\big).$$

(1) (8.146), and δ commutes with pull-back.
(2) A closed form of positive degree on the contractible $\mathcal{A}$ is exact.
(3) (8.153).
(4) The previous line.
(5) π is constant on an orbit, so π^*h restricts to zero.

So ν_{2j} exact $\Rightarrow$ μ_{2j-1} exact. The converse is the statement that for a principal bundle with *contractible* total space the transgression $\mu_{2j-1} \mapsto \nu_{2j}$ is an isomorphism on these generators, exactly as for $G \to EG \to BG$, where $\omega_{2J-1} \mapsto c_J$. In homotopy language: the long exact sequence (8.38) with $\pi_*(\mathcal{A}) = 0$ gives $\pi_k(B\mathcal{G}) \cong \pi_{k-1}(\mathcal{G})$, and rationally the generators correspond one to one; a generator cannot die on one side without dying on the other.

Worked example 8.38: $n = 4$, SU(3): the pair (μ_1, ν_2)

For SU(3) on S^4 the lowest classes are $\mu_1 \in H^1_{\mathrm{dR}}(\mathcal{G})$ and $\nu_2 \in H^2_{\mathrm{dR}}(\mathcal{A}/\mathcal{G})$, both from $J = 3$ (i.e. $j = 1$ in (8.153)):

$$\pi^*\nu_2 = \int_{S^4} P_3(\mathbb{F}_{\mathrm{univ}}) = \delta f_1, \qquad f_1 = \int_{S^4} TP_3(A + \Theta), \qquad f_1\big|_{\mathcal{G}\cdot A} = \mu_1.$$

What does ν_2 look like at a connection A? We need the $(4,2)$ part of $P_3(\mathbb{F}_{\mathrm{univ}}) = c\,\mathrm{tr}\,\mathbb{F}^3$, with $\mathbb{F} = F + \psi + \Omega_\Theta$ split as in (8.149) (ψ the $(1,1)$ part). The terms with four x-legs and two $\mathcal{A}$-legs are F, F, Ω_Θ and F, ψ, ψ; all three placements of each are cyclically equal (all pieces are even), so

$$\pi^*\nu_2 = -\frac{i}{8\pi^3}\int_{S^4} \mathrm{tr}\big(F^2\,\Omega_\Theta + F\,\psi^2\big).$$

On two variations a, b of the gauge field it is a bilinear, antisymmetric expression built from $\mathrm{tr}(F \wedge a_{\mathrm{h}} \wedge b_{\mathrm{h}})$ (with $a_{\mathrm{h}}, b_{\mathrm{h}}$ the horizontal parts) and from $\mathrm{tr}(F \wedge F\, \Omega_\Theta(a,b))$. It is a 2-form on the space of physical configurations. Its non-triviality, $\nu_2 \neq 0$ in $H^2_{\mathrm{dR}}(\mathcal{A}/\mathcal{G})$, is equivalent by Step 4 to $\mu_1 \neq 0$, i.e. to the existence of the anomaly 1-form (8.145); the pair is the "curvature" and the "connection" of one line bundle over $\mathcal{A}/\mathcal{G}$, whose first Chern class is ν_2. In Chapter 10 this line bundle is the determinant line bundle of the chiral Dirac operator.

Reading Nash (pp. 225–226, (8.46)–(8.58))

Nash's (8.46)–(8.48) are our (8.128)–(8.130); the "restriction to $\mathcal{G} \times S^n$" requires the section of (8.129). His (8.49)–(8.50) are (8.135) and (8.134). He writes $\mathcal{G} \times P$ and $\mathcal{G} \times S^n$ with $\mathcal{G}$ first; we put S^n first so that integration over S^n has the sign convention of (8.90). His (8.52) is (8.146); the claim that the restricted f_j "one can easily check coincides cohomologically with μ_{2j-1}" is our (8.150)–(8.153), which uses the Coulomb connection (8.147) that Nash does not write down. His (8.55)–(8.57) are (8.141)–(8.142); his (8.58) is (8.143), and its evaluation (8.145) is not in Nash. His closing sentence ("if μ_1 is exact the same is true for ν_2") is Step 4.

8.2.10 A secondary class in the action: three-dimensional Chern–Simons gauge theory

Up to now secondary classes were cohomology classes: on P, on the gauge group $\mathcal{G}$, on the configuration space $\mathcal{A}/\mathcal{G}$. Nash's last example of the section puts one straight into a Lagrangian. In three dimensions the Chern–Simons form $TP_2(A)$ is a 3-form, so it can be integrated over spacetime and added to the action. This was done by Deser, Jackiw and Templeton (1982) and by Schonfeld (1981). Two things happen, and both are consequences of the $\mathbb{R}/\mathbb{Z}$ nature of the secondary class that we found in Section 8.2.7:

1. the coefficient of the Chern–Simons term must be quantised;
2. the gauge field acquires a mass, without any Higgs field.

The three-dimensional gauge group is disconnected

Take spacetime to be a compact oriented 3-manifold M; the model case is $M = S^3$, Euclidean $\mathbb{R}^3$ with the point at infinity added (fields approach a pure gauge at infinity, as in Section 8.1.16). The based gauge group of the trivial bundle is $\mathcal{G} \simeq \mathrm{Map}_0(S^3, G) = \Omega^3 G$ (Section 8.1.13), and so (Nash's (8.59))

$$\pi_0(\mathcal{G}) \overset{(1)}{=} \pi_0\big(\Omega^3 G\big) \overset{(2)}{=} \big[S^0, \Omega^3 G\big]_* \overset{(3)}{=} \big[S^3, G\big]_* \overset{(4)}{=} \pi_3(G) \overset{(5)}{=} \mathbb{Z}. \tag{8.154}$$

(1) $\mathcal{G} \simeq \Omega^3 G$.
(2) The components of a space X are the based homotopy classes of maps from the two-point space S^0 (one point goes to the base point, the other picks the component).
(3) The adjunction $[\Sigma X, Y]_* = [X, \Omega Y]_*$ used three times, with $\Sigma^3 S^0 = S^3$ (Section 8.1.13).
(4) Definition of π_3.
(5) For every compact, connected, simple Lie group, $\pi_3(G) = \mathbb{Z}$ (Bott). For $G = \mathrm{SU}(2) = S^3$ the integer is the degree; for $\mathrm{SU}(N)$ a generator is the SU(2) identity map placed in the top-left 2×2 block.

So the gauge transformations of a three-dimensional theory fall into infinitely many classes, labelled by the winding number $w(g)$ of (8.64). These are the "large" gauge transformations of Section 8.1.19; there they acted on the spatial S^3 of a four-dimensional theory, here they act on the whole three-dimensional spacetime.

The action

The action is the Yang–Mills action plus a multiple of the Chern–Simons invariant (Nash's (8.60)):

$$S = \|F\|^2 + \alpha \int_M TP_2(A), \qquad \|F\|^2 := -\int_M \operatorname{tr}\big(F \wedge \star F\big). \tag{8.155}$$

The sign of $\|F\|^2$. With anti-Hermitian F the minus sign makes the Yang–Mills term positive. Writing $F = \frac{1}{2}F_{\mu\nu}\mathrm{d}x^\mu \wedge \mathrm{d}x^\nu$ on Euclidean M,

$$-\operatorname{tr}\big(F \wedge \star F\big) \overset{(1)}{=} -\tfrac{1}{2}\operatorname{tr}\big(F_{\mu\nu}F_{\mu\nu}\big)\,\mathrm{vol} \overset{(2)}{=} \tfrac{1}{2}\operatorname{tr}\big(F^\dagger_{\mu\nu}F_{\mu\nu}\big)\,\mathrm{vol} \overset{(3)}{=} \tfrac{1}{2}\sum_{\mu,\nu}\sum_{a,b}\big|(F_{\mu\nu})_{ab}\big|^2\,\mathrm{vol} \;\geq\; 0.$$

(1) For 2-forms, $\beta \wedge \star\gamma = \frac{1}{2}\beta_{\mu\nu}\gamma_{\mu\nu}\,\mathrm{vol}$ (Chapter T1); the trace just rides along.
(2) Anti-Hermitian: $F_{\mu\nu} = -F^\dagger_{\mu\nu}$.
(3) $\operatorname{tr}(X^\dagger X)$ is the sum of the squared moduli of the entries of X.

The Chern–Simons 3-form (Nash's (8.61)). With $P_2(X,Y) = -\frac{1}{8\pi^2}\operatorname{tr}(XY)$ and $F_t = t\,\mathrm{d}A + t^2 A\wedge A$,

$$\begin{aligned} TP_2(A) &\overset{(1)}{=} 2\int_0^1 \mathrm{d}t\; P_2(A,F_t) \overset{(2)}{=} -\frac{2}{8\pi^2}\int_0^1 \mathrm{d}t\;\operatorname{tr}\big(A\wedge(t\,\mathrm{d}A + t^2 A\wedge A)\big) \\ &\overset{(3)}{=} -\frac{2}{8\pi^2}\Big(\tfrac{1}{2}\operatorname{tr}(A\wedge\mathrm{d}A) + \tfrac{1}{3}\operatorname{tr}(A\wedge A\wedge A)\Big) \overset{(4)}{=} -\frac{1}{8\pi^2}\operatorname{tr}\Big(A\wedge\mathrm{d}A + \tfrac{2}{3}A\wedge A\wedge A\Big). \end{aligned} \tag{8.156}$$

(1) The transgression formula (8.115) with $j = 2$.
(2) The polynomial P_2 of Table 8.3, polarised; one slot is A, the other F_t.
(3) $\int_0^1 t\,\mathrm{d}t = \frac{1}{2}$, $\int_0^1 t^2\,\mathrm{d}t = \frac{1}{3}$.
(4) Pull out $\frac{1}{2}$.

This is (8.116) again. (The middle line of Nash's (8.61) reads $\operatorname{tr}\{A\wedge(tA\wedge\mathrm{d}A + t^2A\wedge A\wedge A)\}$; the inner bracket should be $F_t = t\,\mathrm{d}A + t^2A\wedge A$, as the last line of (8.61) confirms.)

Putting it into (8.155) gives Nash's (8.62):

$$\begin{aligned} S &\overset{(1)}{=} -\int_M \operatorname{tr}(F\wedge\star F) - \frac{\alpha}{8\pi^2}\int_M \operatorname{tr}\Big(A\wedge\mathrm{d}A + \tfrac{2}{3}A^3\Big) \\ &\overset{(2)}{=} -\operatorname{tr}\int_M \Big\{F\wedge\star F + \frac{\alpha}{8\pi^2}\Big(\mathrm{d}A\wedge A + \tfrac{2}{3}A\wedge A\wedge A\Big)\Big\}. \end{aligned} \tag{8.157}$$

(1) (8.155) and (8.156).
(2) Nash writes $\mathrm{d}A\wedge A$ where we have $A\wedge\mathrm{d}A$. Under the trace they are equal: moving the 2-form $\mathrm{d}A$ past the 1-form A costs $(-1)^{2\cdot1} = +1$, and the trace is cyclic.

In components. The Chern–Simons term uses no metric at all: there is no $\star$ in it. Indeed, with $A = A_\mu\mathrm{d}x^\mu$ and $\mathrm{d}A = \partial_\nu A_\rho\,\mathrm{d}x^\nu\wedge\mathrm{d}x^\rho$,

$$\begin{aligned} \operatorname{tr}\Big(A\,\mathrm{d}A + \tfrac{2}{3}A^3\Big) &\overset{(1)}{=} \operatorname{tr}\Big(A_\mu\partial_\nu A_\rho + \tfrac{2}{3}A_\mu A_\nu A_\rho\Big)\mathrm{d}x^\mu\wedge\mathrm{d}x^\nu\wedge\mathrm{d}x^\rho \\ &\overset{(2)}{=} \epsilon^{\mu\nu\rho}\operatorname{tr}\Big(A_\mu\partial_\nu A_\rho + \tfrac{2}{3}A_\mu A_\nu A_\rho\Big)\mathrm{d}^3x. \end{aligned}$$

(1) Multiply out the forms, keeping the order of the matrix factors.
(2) In three dimensions $\mathrm{d}x^\mu \wedge \mathrm{d}x^\nu \wedge \mathrm{d}x^\rho = \epsilon^{\mu\nu\rho}\,\mathrm{d}^3x$ with $\epsilon^{123} = +1$.

A term that does not see the metric is called **topological**. It does not contribute to the energy–momentum tensor ($T_{\mu\nu} = \frac{2}{\sqrt{g}}\frac{\delta S}{\delta g^{\mu\nu}}$ gets nothing from it), and it is odd under orientation reversal: for an orientation-reversing diffeomorphism ϕ (a reflection $x^1 \to -x^1$), $\int_M \phi^* TP_2 = -\int_M TP_2$. So the Chern–Simons term violates parity.

Gauge variation and the quantisation of α

The Yang–Mills term is gauge invariant: $F^g = g^{-1}Fg$ and $\star(g^{-1}Fg) = g^{-1}(\star F)g$, so the trace does not change. The Chern–Simons term is not. For a gauge transformation $g : M \to G$,

$$\begin{aligned}\int_M TP_2(A^g) &\overset{(1)}{=} \int_M TP_2(A) + \frac{1}{24\pi^2}\int_M \mathrm{tr}\big(g^{-1}\mathrm{d}g\big)^3 + \frac{1}{8\pi^2}\int_M \mathrm{d}\,\mathrm{tr}\big(\mathrm{d}g\, g^{-1}\wedge A\big)\\ &\overset{(2)}{=} \int_M TP_2(A) + w(g).\end{aligned} \tag{8.158}$$

(1) The gauge variation formula (8.123), integrated over M.
(2) Stokes: M is closed, so the exact term integrates to zero. The middle term is the winding number (8.64) (for $M = S^3$; for a general closed M it is the integer of (8.125)).

Hence $S(A^g) = S(A) + \alpha\, w(g)$. *The action itself is not gauge invariant.* It does not have to be: what enters the path integral is the weight e^{iS}. Then

$$e^{iS(A^g)} \overset{(1)}{=} e^{iS(A)}\, e^{i\alpha w(g)}, \qquad e^{i\alpha w} = 1 \ \forall\, w \in \mathbb{Z} \overset{(2)}{\iff} e^{i\alpha} = 1 \iff \alpha \in 2\pi\mathbb{Z}. \tag{8.159}$$

(1) (8.158); the exponential turns a sum into a product.
(2) The case $w = 1$ exists (the generator of (8.154)), and if $e^{i\alpha} = 1$ then $e^{i\alpha w} = (e^{i\alpha})^w = 1$ for all w.

Key idea: Nash's quantisation condition

The weight e^{iS} of the Deser–Jackiw–Templeton theory is gauge invariant if and only if

$$\alpha = 2\pi k, \qquad k \in \mathbb{Z}.$$

The integer k is the **level**. With it the Chern–Simons term takes its standard form

$$\alpha\int_M TP_2(A) = 2\pi k\cdot\Big(-\frac{1}{8\pi^2}\Big)\int_M \mathrm{tr}\Big(A\,\mathrm{d}A + \tfrac{2}{3}A^3\Big) = -\frac{k}{4\pi}\int_M \mathrm{tr}\Big(A\,\mathrm{d}A + \tfrac{2}{3}A^3\Big),$$

and $-\mathrm{tr}$ is positive definite on anti-Hermitian matrices, so this is the familiar "$\frac{k}{4\pi}\int \mathrm{tr}(A\,\mathrm{d}A + \frac{2}{3}A^3)$ with a positive trace". This is the secondary class at work: $\int_M TP_2$ is only defined in $\mathbb{R}/\mathbb{Z}$ as a gauge invariant, and $e^{2\pi i k\int TP_2}$ is exactly the kind of function of an $\mathbb{R}/\mathbb{Z}$-valued quantity that makes sense.

Worked example 8.39: Small gauge transformations cost nothing

For an infinitesimal gauge transformation, $\delta A = D_A\xi$, the change of the Chern–Simons term should vanish (it is a gauge transformation with $w = 0$). Directly, using the variation

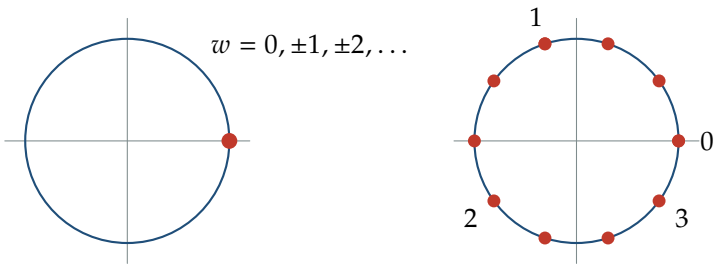


Figure 8.27: The phases $e^{i\alpha w(g)}$ picked up by e^{iS} under large gauge transformations of winding w. For $\alpha \in 2\pi\mathbb{Z}$ (left) they are all 1: the weight is gauge invariant. For any other α (right, $\alpha/2\pi = 0.3$, points labelled by w) different windings give different phases, so gauge-equivalent fields would get different weights and the theory would be inconsistent.

formula (8.124) and writing $\kappa = -\frac{1}{8\pi^2}$:

$$\delta \int_M TP_2(A) \overset{(1)}{=} 2\kappa \int_M \operatorname{tr}\big(\delta A \wedge F\big) + \int_M \mathrm{d}(\ldots) \overset{(2)}{=} 2\kappa \int_M \operatorname{tr}\big(D_A\xi \wedge F\big)$$
$$\overset{(3)}{=} 2\kappa \int_M \Big(\mathrm{d}\operatorname{tr}(\xi F) - \operatorname{tr}(\xi\, D_A F)\Big) \overset{(4)}{=} 2\kappa \int_M \mathrm{d}\operatorname{tr}(\xi F) \overset{(5)}{=} 0.$$

(1) (8.124) with $j = 2$ and $P_2(X, Y) = \kappa \operatorname{tr}(XY)$.
(2) $\delta A = D_A\xi$; the total derivative integrates to zero on closed M.
(3) Covariant Leibniz rule under the trace: $\mathrm{d}\operatorname{tr}(\xi F) = \operatorname{tr}(D_A\xi \wedge F) + \operatorname{tr}(\xi\, D_A F)$ (the commutator terms cancel under the trace; Section 8.1.1).
(4) Bianchi identity $D_A F = 0$.
(5) Stokes, $\partial M = \varnothing$.

So only the *topology* of g can change $\int_M TP_2$: its value mod $\mathbb{Z}$ is a gauge invariant of A. (On a manifold with boundary the last step fails and the Chern–Simons term is gauge variant already under small transformations; this is the origin of the edge modes of the quantum Hall effect.)

Worked example 8.40: The one-dimensional analogue: charge quantisation

Drop two dimensions. On a circle $M = S^1$ of length T (Euclidean time, say) with $G = \mathrm{U}(1)$, write $A = i\,a(t)\,\mathrm{d}t$. The one-dimensional Chern–Simons form is $TP_1(A) = \frac{i}{2\pi}A = -\frac{1}{2\pi}a\,\mathrm{d}t$ (Section 8.2.1). The term $\alpha \int_{S^1} TP_1 = -\frac{\alpha}{2\pi}\oint a\,\mathrm{d}t$ is the coupling of a point charge $q := \alpha/2\pi$ to the gauge field along its world line: a Wilson loop. A large gauge transformation $g(t) = e^{2\pi i n t/T}$ has

$$\int_{S^1} TP_1(A^g) - \int_{S^1} TP_1(A) \overset{(1)}{=} \frac{i}{2\pi}\oint g^{-1}\mathrm{d}g \overset{(2)}{=} \frac{i}{2\pi}\cdot\frac{2\pi i n}{T}\cdot T = -n,$$

(1) For $\mathrm{U}(1)$, $A^g = A + g^{-1}\mathrm{d}g$.
(2) $g^{-1}\mathrm{d}g = \frac{2\pi i n}{T}\mathrm{d}t$.

so $e^{i\alpha\int TP_1}$ changes by $e^{-i\alpha n}$ and is invariant iff $\alpha \in 2\pi\mathbb{Z}$, i.e. $q \in \mathbb{Z}$. *Quantisation of the level in one dimension is quantisation of electric charge*: a Wilson loop $e^{iq\oint a}$ is gauge invariant under all gauge transformations on a circle only for integer q. The three-dimensional statement is the same mechanism one rung up the ladder of Chern–Simons forms.

The topological mass

Why is the Chern–Simons term interesting physically? Because it gives the gauge field a mass while keeping gauge invariance (of e^{iS}). The effect is already visible in the Abelian theory, and the non-Abelian theory reduces to it at quadratic order.

Worked example 8.41: Maxwell–Chern–Simons theory: a massive photon

Work in Minkowski space $\mathbb{R}^{1,2}$ with signature $(+,-,-)$, $\epsilon^{012}=+1$, a real (Hermitian-convention) potential a_μ, $f_{\mu\nu}=\partial_\mu a_\nu-\partial_\nu a_\mu$, and

$$\mathcal{L}=-\frac{1}{4e^2}f_{\mu\nu}f^{\mu\nu}+\frac{\kappa}{2}\,\epsilon^{\mu\nu\rho}a_\mu\partial_\nu a_\rho .$$

Define the dual field strength $\tilde f^\sigma:=\frac12\epsilon^{\sigma\nu\rho}f_{\nu\rho}=\epsilon^{\sigma\nu\rho}\partial_\nu a_\rho$ (in three dimensions a 2-form is dual to a vector: $\tilde f^0$ is the magnetic field, $\tilde f^i$ are the components of the electric field up to signs). It satisfies $\partial_\sigma\tilde f^\sigma=0$ (Bianchi) and $f_{\mu\nu}=\epsilon_{\mu\nu\lambda}\tilde f^\lambda$.

Equation of motion. The Euler–Lagrange equation $\partial_\nu\frac{\partial\mathcal{L}}{\partial(\partial_\nu a_\sigma)}=\frac{\partial\mathcal{L}}{\partial a_\sigma}$ reads

$$-\frac{1}{e^2}\partial_\mu f^{\mu\sigma}+\frac{\kappa}{2}\epsilon^{\mu\nu\sigma}\partial_\nu a_\mu \overset{(1)}{=} \frac{\kappa}{2}\epsilon^{\sigma\nu\rho}\partial_\nu a_\rho \quad\overset{(2)}{\Longrightarrow}\quad \partial_\mu f^{\mu\sigma}=-e^2\kappa\,\tilde f^\sigma .$$

(1) $\partial\mathcal{L}/\partial(\partial_\nu a_\sigma)=-\frac{1}{e^2}f^{\nu\sigma}+\frac{\kappa}{2}\epsilon^{\mu\nu\sigma}a_\mu$ and $\partial\mathcal{L}/\partial a_\sigma=\frac{\kappa}{2}\epsilon^{\sigma\nu\rho}\partial_\nu a_\rho$.

(2) $\epsilon^{\mu\nu\sigma}\partial_\nu a_\mu=\epsilon^{\sigma\mu\nu}\partial_\nu a_\mu=-\epsilon^{\sigma\nu\mu}\partial_\nu a_\mu=-\tilde f^\sigma$; collect the two κ terms and multiply by $-e^2$.

Write $m:=e^2\kappa$. With $f_{\mu\sigma}=\epsilon_{\mu\sigma\lambda}\tilde f^\lambda$ the equation is $\epsilon^{\mu\sigma\lambda}\partial_\mu\tilde f_\lambda=-m\tilde f^\sigma$, a first-order equation for $\tilde f$.

Squaring it. Lowering all indices, the first-order equation reads $\epsilon_{\alpha\beta\sigma}\partial^\alpha\tilde f^\sigma=-m\tilde f_\beta$, equivalently $\epsilon_{\beta\alpha\sigma}\partial^\alpha\tilde f^\sigma=+m\tilde f_\beta$. Apply $\epsilon_{\beta\alpha\sigma}\partial^\alpha$ to $\epsilon^{\mu\sigma\lambda}\partial_\mu\tilde f_\lambda=-m\tilde f^\sigma$:

$$\text{left side:}\quad \epsilon_{\beta\alpha\sigma}\epsilon^{\mu\sigma\lambda}\partial^\alpha\partial_\mu\tilde f_\lambda \overset{(1)}{=} \big(\delta^\lambda_\beta\delta^\mu_\alpha-\delta^\mu_\beta\delta^\lambda_\alpha\big)\partial^\alpha\partial_\mu\tilde f_\lambda=\Box\tilde f_\beta-\partial_\beta(\partial\cdot\tilde f)\overset{(2)}{=}\Box\tilde f_\beta,$$

$$\text{right side:}\quad -\,m\,\epsilon_{\beta\alpha\sigma}\partial^\alpha\tilde f^\sigma \overset{(3)}{=} -m\big(m\tilde f_\beta\big)=-m^2\tilde f_\beta .$$

(1) The contraction identity for the three-dimensional Levi-Civita symbol, $\epsilon_{\beta\alpha\sigma}\epsilon^{\mu\sigma\lambda}=\delta^\lambda_\beta\delta^\mu_\alpha-\delta^\mu_\beta\delta^\lambda_\alpha$ (with $\epsilon_{012}=\det\eta\cdot\epsilon^{012}=+1$; checked by writing out the cases). $\Box=\partial_\alpha\partial^\alpha=\partial_t^2-\nabla^2$.

(2) Bianchi, $\partial\cdot\tilde f=0$.

(3) The equation of motion in its lowered form, quoted above.

Equating the two lines,

$$\big(\Box+m^2\big)\tilde f_\beta=0,\qquad m=e^2\kappa. \tag{8.160}$$

Every component of the field strength obeys the Klein–Gordon equation with mass $|m|$: plane waves have $\omega^2=\vec p^{\,2}+m^2$. Pure Maxwell theory in $2+1$ dimensions has one massless polarisation; with the Chern–Simons term it becomes one massive particle, of spin $+1$ or -1 according to the sign of κ (the first-order equation picks one helicity; that is the parity violation). Nothing was broken: no Higgs field, no gauge-non-invariant $a_\mu a^\mu$ term. The mass is **topological**.

The non-Abelian theory at quadratic order. Restore the coupling in Minkowski signature: $\mathcal{L}=\frac{1}{2g^2}\operatorname{tr}F_{\mu\nu}F^{\mu\nu}+$ (CS term of (8.157)). Expand $A=A^aT_a$ with anti-Hermitian generators

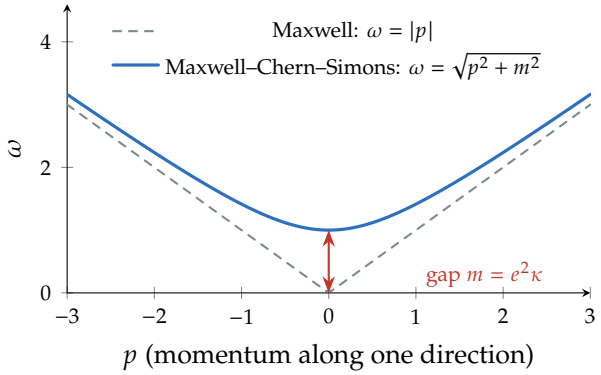


Figure 8.28: Dispersion relations in 2 + 1 dimensions. The Chern–Simons term opens a gap (red): the photon becomes massive, with $m = e^2\kappa$ (here $m = 1$). In the non-Abelian theory $m = g^2\alpha/8\pi^2 = g^2 k/4\pi$ is quantised in units of $g^2/4\pi$.

normalised by $\operatorname{tr}(T_a T_b) = -\tfrac{1}{2}\delta_{ab}$ (for SU(2), $T_a = -\tfrac{i}{2}\sigma_a$). Then

$$\frac{1}{2g^2}\operatorname{tr} F_{\mu\nu}F^{\mu\nu} \overset{(1)}{=} -\frac{1}{4g^2}F^a_{\mu\nu}F^{a\,\mu\nu},$$
$$-\frac{\alpha}{8\pi^2}\operatorname{tr}(A\,\mathrm{d}A) \overset{(2)}{=} \frac{\alpha}{16\pi^2}A^a \wedge \mathrm{d}A^a \overset{(3)}{=} \frac{1}{2}\cdot\frac{\alpha}{8\pi^2}\epsilon^{\mu\nu\rho}A^a_\mu \partial_\nu A^a_\rho \,\mathrm{d}^3x.$$

(1) $\operatorname{tr}(T_a T_b) = -\tfrac{1}{2}\delta_{ab}$.
(2) The same normalisation.
(3) The component formula above.

To quadratic order each colour component is a copy of Example 8.41 with $e^2 \to g^2$ and $\kappa \to \alpha/8\pi^2$, so

$$m = \frac{g^2\alpha}{8\pi^2} = \frac{g^2 k}{4\pi} \qquad (\alpha = 2\pi k). \tag{8.161}$$

In three dimensions g^2 has the dimension of a mass (from $[A] = 1$ and a dimensionless action, $\int \mathrm{d}^3x\, F^2/g^2$ forces $[g^2] = 1$), so m is a mass, and it comes in integer multiples of $g^2/4\pi$. Nash writes the action with $g = 1$, i.e. with the coupling absorbed into the field; then the coefficient of the Chern–Simons term itself carries the dimension of mass, which is his remark that "α has the dimension of a mass: we obtain a 3-dimensional gauge theory with a quantised mass".

Watch out: e^{iS}, Euclidean and Minkowski

Nash writes the weight as $\exp[iS]$ with S containing both $\|F\|^2$ and the Chern–Simons term. The two terms behave differently under Wick rotation $t = -i\tau$. The Yang–Mills term turns into the positive Euclidean action, giving $e^{-\|F\|^2}$. The Chern–Simons term contains exactly one time index ($\epsilon^{\mu\nu\rho}$ with one index 0) and no metric, so it picks up a single factor of i and stays a *phase*: the Euclidean weight is $e^{-\|F\|^2 + i\alpha\int TP_2}$ (up to the orientation sign). Either way, the quantisation condition concerns the coefficient of $\int TP_2$ in the phase, and it is $\alpha \in 2\pi\mathbb{Z}$. Dropping $\|F\|^2$ altogether leaves pure Chern–Simons theory, Witten's topological quantum field theory of knots (Chapter 12).

Higher odd dimensions

Nash closes with the remark that $\mathcal{G}$ is disconnected "for all the larger odd values of dim M". For $M = S^{2j-1}$ and $G = \mathrm{SU}(N)$, $\pi_0(\mathcal{G}) = \pi_{2j-1}(\mathrm{SU}(N)) = \mathbb{Z}$ as soon as $N \geq j$ (the stable range of Bott

periodicity, Table 8.1), and the corresponding Chern–Simons form TP_j is a $(2j-1)$-form that can be put into a $(2j-1)$-dimensional action. The pattern is always the same: the coefficient must be quantised, because $\int TP_j$ changes by the integer $\int g^*\omega_{2j-1}$ under a large gauge transformation, by (8.125).

dim M	Chern–Simons form	$\pi_0(\mathcal{G})$	physics
1	$TP_1 = \frac{i}{2\pi}\operatorname{tr} A$	$\pi_1(\mathrm{U}(1)) = \mathbb{Z}$	charge quantisation (Example 8.40)
3	$TP_2 = -\frac{1}{8\pi^2}\operatorname{tr}(A\,\mathrm{d}A + \frac{2}{3}A^3)$	$\pi_3(\mathrm{SU}(N)) = \mathbb{Z}$	topological mass, level k
5	TP_3 of (8.117)	$\pi_5(\mathrm{SU}(N)) = \mathbb{Z}$, $N \geq 3$	anomaly inflow to 4d
$2j-1$	TP_j	$\pi_{2j-1}(\mathrm{SU}(N)) = \mathbb{Z}$, $N \geq j$	quantised coefficient

The five-dimensional row connects back to Section 8.2.8. The gauge variation of the Chern–Simons 5-form contains, by the descent expansion (8.138), the 4-form $\Omega_4^{(1)}$ of (8.144). On a five-manifold with boundary, the gauge variation of $\int_{M_5} TP_3$ is therefore a boundary term of the form $\int_{\partial M_5} \operatorname{tr}\,\xi\,\mathrm{d}(A\,\mathrm{d}A + \frac{1}{2}A^3)$: exactly the four-dimensional anomaly μ_1. A five-dimensional Chern–Simons term can cancel the anomaly of chiral fermions living on its boundary (**anomaly inflow**, Chapter 10).

Reading Nash (p. 227, (8.59)–(8.62))

Nash's (8.59) is (8.154). His action (8.60) is (8.155), with $\|F\|^2 = -\int \operatorname{tr} F \wedge \star F$. His (8.61) is (8.156); its middle line has a misprint ($tA \wedge \mathrm{d}A$ should be $t\,\mathrm{d}A$). His (8.62) is (8.157), where $\mathrm{d}A \wedge A$ may be read as $A \wedge \mathrm{d}A$ under the trace. His argument that "$TP_2(A)$ changes by an integer under a gauge transformation" is (8.158), and the quantisation $\alpha = 2\pi n$ is (8.159). The "quantised mass" is (8.161), derived in Example 8.41; Nash does not derive it.

Recap: secondary characteristic classes in one page

- *Secondary classes.* On P, $\pi^*c_j(F) = \mathrm{d}TP_j$ with $TP_j = j\int_0^1 P_j(\omega, F_t, \ldots)\,\mathrm{d}t$. When $c_j(F) = 0$ the Chern–Simons form is closed and defines a real class on P, and an $\mathbb{R}/\mathbb{Z}$ class on M: two gauges change it by the integer periods of $g^*\omega_{2j-1}$.
- *μ from Chern–Simons forms.* On $S^n \times \mathcal{G}$ the connection $\mathbb{A} = A^\alpha + \vartheta$ ($\vartheta = \alpha^{-1}\delta\alpha$, the ghost) has curvature F^α with no $\mathcal{G}$-legs. So $\widehat{\mathrm{Ev}}^* TP_j$ is closed when $2j > n$, and $\mu_{2j-1-n} = \int_{S^n} \widehat{\mathrm{Ev}}^* TP_j$ is the old μ plus an exact form.
- *Explicit μ_1.* On S^2: $\mu_1(\xi) = -\frac{1}{8\pi^2}\int \operatorname{tr}(\xi\,\mathrm{d}A)$. On S^4 for SU(N): $\mu_1(\xi) = -\frac{i}{24\pi^3}\int \operatorname{tr}\big(\xi\,\mathrm{d}(A\,\mathrm{d}A + \frac{1}{2}A^3)\big)$, the shape of the non-Abelian anomaly.
- *ν and the double transgression.* With the Coulomb connection Θ on $\mathcal{A} \to \mathcal{A}/\mathcal{G}$, $\mathbb{A}_{\mathrm{univ}} = A + \Theta$; then $\pi^*\nu_{2j} = \delta f_j$ with $f_j = \int_{S^n} TP_{j+n/2}(A + \Theta)$, and f_j restricted to an orbit is μ_{2j-1}. μ_{2j-1} and ν_{2j} are exact or non-exact together.
- *Chern–Simons theory in 3d.* $S = \|F\|^2 + \alpha\int TP_2$; large gauge transformations shift $\int TP_2$ by the winding number, so e^{iS} is invariant iff $\alpha = 2\pi k$; the gauge field gets the quantised topological mass $m = g^2k/4\pi$.

8.3 Instantons: the Yang–Mills action and its critical points

In Euclidean quantum field theory every gauge field configuration A comes with a weight $e^{-S(A)/g^2}$. When the coupling g is small, the path integral is dominated by the configurations where S is stationary. The trivial field $A = 0$ is one such point, and perturbation theory expands around it. But there are others. They sit in different topological sectors, they have finite action, and they contribute effects of order $e^{-8\pi^2/g^2}$, which no Feynman diagram can see. These are the **instantons**.

Nash's Section 3 asks a very concrete question about them: *how many instantons are there?* More precisely, instantons come in continuous families, and we want the number of parameters in a family: the dimension of the **moduli space**. A physicist knows this number by another name. It is the number of *collective coordinates*, or equivalently the number of normalisable bosonic zero modes of the fluctuation operator around the instanton. Each zero mode must be traded for an integral over a collective coordinate in the path integral, so the number matters for every instanton calculation.

The road is long, so here is the plan of this part.

1. Find the equations of motion of the Yang–Mills action (this section).
2. Show that the self-dual fields solve them and are the absolute minima of the action in each topological sector (Section 8.4).
3. Write down the simplest instanton, the BPST solution, completely explicitly (Section 8.5).
4. Understand why gauge invariance spoils ellipticity, and fix a gauge (Section 8.6).
5. Define the moduli space carefully (Section 8.7).
6. Linearise the self-duality equation and package the zero modes into an elliptic complex (Section 8.8).
7. Count the zero modes with the index theorem: $\dim \mathcal{M}_k = 8k - 3$ for SU(2) on S^4 (Section 8.9).
8. Check the count by hand for $k = 1$, and see the ADHM construction (Section 8.10).
9. Look at the global structure and Donaldson's theorem (Section 8.11).

8.3.1 The setting and our conventions

The stage is a **closed** (compact, no boundary), oriented, Riemannian 4-manifold M. The main example is the round sphere S^4. Flat $\mathbb{R}^4$ with fields that fall off at infinity will turn out to be the same thing as S^4, because the action is conformally invariant (Section 8.4.1). Over M we have a principal G-bundle P, with G a compact simple Lie group; our workhorse is $G = \mathrm{SU}(2)$.

A connection on P is, in a local trivialisation, a Lie-algebra-valued 1-form

$$A = A_\mu \, \mathrm{d}x^\mu, \qquad A_\mu(x) \in \mathfrak{g}.$$

We use the mathematicians' (and Nash's) convention: for $G = \mathrm{SU}(N)$ the matrices A_μ are *anti-Hermitian* and traceless. The curvature is

$$F = \mathrm{d}A + A \wedge A, \qquad F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu + [A_\mu, A_\nu], \tag{8.162}$$

where $F = \frac{1}{2} F_{\mu\nu} \, \mathrm{d}x^\mu \wedge \mathrm{d}x^\nu$. Under a gauge transformation $g : M \to G$ (locally),

$$A \mapsto A_g = g^{-1} A g + g^{-1} \mathrm{d} g, \qquad F \mapsto g^{-1} F g. \tag{8.163}$$

This is reviewed in Chapter T3; we only need the formulas.

Dictionary: anti-Hermitian versus Hermitian gauge fields

Physics books write $A_\mu^{\text{phys}} = A_\mu^a T^a$ with Hermitian generators T^a, normalised by $\operatorname{tr}(T^a T^b) = \frac{1}{2}\delta^{ab}$, and put the coupling into the covariant derivative $D_\mu = \partial_\mu - igA_\mu^{\text{phys}}$. The translation is

$$A_\mu^{\text{math}} = -ig\,A_\mu^a T^a, \qquad F_{\mu\nu}^{\text{math}} = -ig\,F_{\mu\nu}^a T^a, \qquad F_{\mu\nu}^a = \partial_\mu A_\nu^a - \partial_\nu A_\mu^a + g f^{abc} A_\mu^b A_\nu^c.$$

For SU(2), $T^a = \sigma_a/2$. We shall also use the anti-Hermitian basis $e_a = -i\sigma_a$ (so $T^a = \frac{i}{2}e_a$ and $-igA^aT^a = \frac{g}{2}A^a e_a$). With these rules the physicist's action is $\frac{1}{4}\int F_{\mu\nu}^a F_{\mu\nu}^a \,\mathrm{d}^4x = \left\|F^{\text{math}}\right\|^2/g^2$; we check this in the worked example just below.

Lie-algebra-valued forms and the bundle $\operatorname{ad} P$

A **Lie-algebra-valued p-form** is a p-form whose components are matrices in $\mathfrak{g}$. The curvature F is one. So is the difference $a = A' - A$ of two connections. Both transform by conjugation, $a \mapsto g^{-1}ag$, without the inhomogeneous $g^{-1}\mathrm{d}g$ term (the inhomogeneous terms cancel in the difference). Objects that transform by conjugation are sections of the **adjoint bundle** $\operatorname{ad} P$: the vector bundle whose fibre is $\mathfrak{g}$ and whose transition functions act by $X \mapsto g^{-1}Xg$. We write

$$\Omega^p(M, \operatorname{ad} P) = \{\mathfrak{g}\text{-valued } p\text{-forms that transform by conjugation}\}.$$

For SU(2) the fibre $\mathfrak{su}(2)$ is 3-dimensional; $\operatorname{ad} P$ is a real rank-3 vector bundle. It carries the adjoint representation, which is the spin-1 representation, so $\pm g$ act identically on it.

Working definition 8.20 (covariant exterior derivative)

On $\omega \in \Omega^p(M, \operatorname{ad} P)$,

$$\mathrm{d}_A\omega = \mathrm{d}\omega + A \wedge \omega - (-1)^p\,\omega \wedge A. \tag{8.164}$$

The two most used cases are

$$\mathrm{d}_A f = \mathrm{d}f + [A, f] \;\; (p=0), \qquad \mathrm{d}_A a = \mathrm{d}a + A \wedge a + a \wedge A \;\; (p=1).$$

Test: $\mathrm{d}_A\omega$ transforms by conjugation whenever ω does [the $g^{-1}\mathrm{d}g$ terms coming from $\mathrm{d}\omega$ and from the transformed A cancel]. In components, $(\mathrm{d}_A f)_\mu = D_\mu f = \partial_\mu f + [A_\mu, f]$.

The sign $-(-1)^p$ is what makes d_A covariant: moving A past a p-form costs $(-1)^p$, and the combination in (8.164) is the graded commutator $[A, \omega]$. Nash states this formula in a footnote on [Nash p. 235].

Worked example 8.42: The Bianchi identity $\mathrm{d}_A F = 0$

We apply (8.164) with $p = 2$ to $F = \mathrm{d}A + A \wedge A$:

$$\begin{aligned}
\mathrm{d}_A F &\overset{(1)}{=} \mathrm{d}F + A \wedge F - F \wedge A \\
&\overset{(2)}{=} \big(\mathrm{d}A \wedge A - A \wedge \mathrm{d}A\big) + \big(A \wedge \mathrm{d}A + A \wedge A \wedge A \wedge A\big) - \big(\mathrm{d}A \wedge A + A \wedge A \wedge A\big) \\
&\overset{(3)}{=} 0.
\end{aligned}$$

(1) Definition (8.164) with $p = 2$, so $-(-1)^2 = -1$.
(2) $\mathrm{d}(\mathrm{d}A) = 0$, and the graded Leibniz rule gives $\mathrm{d}(A \wedge A) = \mathrm{d}A \wedge A - A \wedge \mathrm{d}A$ (the minus because A is a 1-form); then substitute $F = \mathrm{d}A + A \wedge A$ in the other two terms.

(3) Every term appears once with each sign.

The Bianchi identity is an *identity*: it holds for every connection, not just for solutions. For $G = \mathrm{U}(1)$ it is $\mathrm{d}F = 0$, i.e. the two homogeneous Maxwell equations $\nabla\cdot\vec{B} = 0$, $\nabla\times\vec{E} + \partial_t\vec{B} = 0$.

The inner product and the action

To write an action we need a positive inner product on $\Omega^p(M, \mathrm{ad}\,P)$. Two ingredients combine. The Hodge star $\star$ (Chapter T1) gives, for ordinary forms, $\alpha\wedge\star\beta = \frac{1}{p!}\alpha_{\mu_1\dots\mu_p}\beta_{\mu_1\dots\mu_p}\,\mathrm{vol}$. The trace gives a positive pairing on anti-Hermitian matrices, once we insert a minus sign: $-\operatorname{tr}(X^2) = \operatorname{tr}(X^\dagger X) = \sum_{ij}|X_{ij}|^2 \geq 0$.

Working definition 8.21 (the L^2 inner product)

For $\alpha, \beta \in \Omega^p(M, \mathrm{ad}\,P)$,

$$\langle\alpha,\beta\rangle = -\int_M \operatorname{tr}(\alpha\wedge\star\beta) = -\frac{1}{p!}\int_M \operatorname{tr}(\alpha_{\mu_1\dots\mu_p}\beta_{\mu_1\dots\mu_p})\,\mathrm{d}^4x\sqrt{\det g}, \qquad \|\alpha\|^2 = \langle\alpha,\alpha\rangle \geq 0. \tag{8.165}$$

It is gauge invariant [unchanged when α,β are replaced by $g^{-1}\alpha g$, $g^{-1}\beta g$], because $\operatorname{tr}(g^{-1}Xg\,g^{-1}Yg) = \operatorname{tr}(XY)$.

Key idea: The Yang–Mills action (Nash (8.63))

$$S(A) = \|F\|^2 = -\int_M \operatorname{tr}(F\wedge\star F) = -\frac{1}{2}\int_M \operatorname{tr}(F_{\mu\nu}F_{\mu\nu})\,\mathrm{d}^4x. \tag{8.166}$$

It is non-negative, gauge invariant, and it vanishes exactly when $F = 0$. Instantons are critical points of S with $F \neq 0$.

Worked example 8.43: From Nash's action to the physicist's action

Take $G = \mathrm{SU}(2)$ and $F = \frac{g}{2}F^a e_a$ from the dictionary, with $e_a = -i\sigma_a$. First the trace of two basis elements:

$$\begin{aligned}\operatorname{tr}(e_a e_b) &\overset{(1)}{=} -\operatorname{tr}(\sigma_a\sigma_b)\\ &\overset{(2)}{=} -\operatorname{tr}(\delta_{ab}\mathbb{1} + i\varepsilon_{abc}\sigma_c)\\ &\overset{(3)}{=} -2\delta_{ab}.\end{aligned}$$

(1) $(-i)^2 = -1$.
(2) The Pauli product rule $\sigma_a\sigma_b = \delta_{ab}\mathbb{1} + i\varepsilon_{abc}\sigma_c$.
(3) $\operatorname{tr}\mathbb{1} = 2$ and the Pauli matrices are traceless.

Now the action density:

$$\begin{aligned}-\tfrac{1}{2}\operatorname{tr}(F_{\mu\nu}F_{\mu\nu}) &\overset{(1)}{=} -\tfrac{1}{2}\cdot\tfrac{g^2}{4}F^a_{\mu\nu}F^b_{\mu\nu}\operatorname{tr}(e_a e_b)\\ &\overset{(2)}{=} \tfrac{g^2}{4}F^a_{\mu\nu}F^a_{\mu\nu}.\end{aligned}$$

(1) Substitute $F_{\mu\nu} = \frac{g}{2}F^a_{\mu\nu}e_a$; the trace is linear.
(2) Use $\operatorname{tr}(e_a e_b) = -2\delta_{ab}$.

So $\|F\|^2 = g^2 \cdot \frac{1}{4}\int F^a_{\mu\nu}F^a_{\mu\nu}\,\mathrm{d}^4x = g^2 S_{\text{phys}}$. The Boltzmann weight is $e^{-S_{\text{phys}}} = e^{-\|F\|^2/g^2}$. Nash sets $g = 1$; we restore it when we quote physical numbers such as $e^{-8\pi^2/g^2}$.

8.3.2 Varying the action: the Yang–Mills equations

The space $\mathcal{A}$ of all connections on P is an *affine* space: pick one connection A, and every other one is $A + a$ with $a \in \Omega^1(M, \operatorname{ad} P)$. So we can draw straight lines through A (Figure 8.29):

$$A_t = A + ta, \qquad t \in \mathbb{R}. \tag{8.167}$$

A is a critical point of S if the slope of $S(A_t)$ at $t = 0$ vanishes for every direction a. That is ordinary calculus on a line.

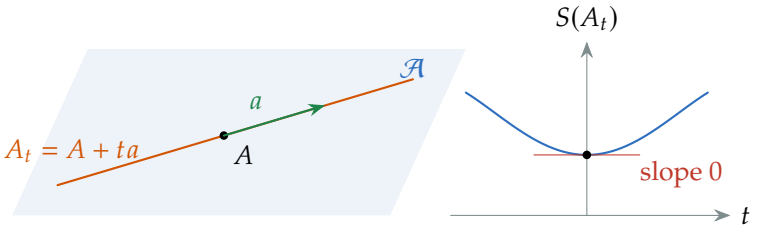


Figure 8.29: Left: the space $\mathcal{A}$ of connections is affine, so through any connection A there is a straight line $A + ta$ in every direction a. Right: A is a critical point when S restricted to every such line has zero slope at $t = 0$. The Yang–Mills equation is this condition, written for all directions a at once.

The curvature along the line

$$\begin{aligned} F(A_t) &\overset{(1)}{=} \mathrm{d}(A + ta) + (A + ta) \wedge (A + ta) \\ &\overset{(2)}{=} \mathrm{d}A + A \wedge A + t\,(\mathrm{d}a + A \wedge a + a \wedge A) + t^2\, a \wedge a \\ &\overset{(3)}{=} F(A) + t\,\mathrm{d}_A a + t^2\, a \wedge a. \end{aligned} \tag{8.168}$$

(1) Definition (8.162) applied to A_t.
(2) Expand and collect powers of t. Keep the order of the wedge products: $A \wedge a$ and $a \wedge A$ are different matrices.
(3) The t^0 terms are $F(A)$; the t^1 terms are $\mathrm{d}_A a$ by (8.164) with $p = 1$.

This is exact, not just to first order: the curvature is quadratic in the connection.

The action along the line

$$\begin{aligned} S(A_t) &\overset{(1)}{=} \langle F + t\,\mathrm{d}_A a + t^2 a \wedge a,\ F + t\,\mathrm{d}_A a + t^2 a \wedge a\rangle \\ &\overset{(2)}{=} \langle F, F\rangle + t\big(\langle \mathrm{d}_A a, F\rangle + \langle F, \mathrm{d}_A a\rangle\big) + O(t^2) \\ &\overset{(3)}{=} S(A) + 2t\,\langle F, \mathrm{d}_A a\rangle + O(t^2). \end{aligned} \tag{8.169}$$

(1) $S = \|F\|^2$ and (8.168).

(2) The inner product is bilinear; collect the terms linear in t.
(3) The inner product is symmetric: $\operatorname{tr}(\alpha \wedge \star\beta) = \operatorname{tr}(\beta \wedge \star\alpha)$ for two 2-forms, since $\alpha \wedge \star\beta$ is $\tfrac{1}{2}\alpha_{\mu\nu}\beta_{\mu\nu}\mathrm{vol}$.

So A is a critical point iff

$$\left.\frac{\mathrm{d}S(A_t)}{\mathrm{d}t}\right|_{t=0} = 2\langle F, \mathrm{d}_A a\rangle = 0 \quad \text{for all } a \in \Omega^1(M, \operatorname{ad} P). \tag{8.170}$$

Moving d_A across: the adjoint d_A^*

To turn (8.170) into a differential equation for A we move the derivative off a, exactly as one integrates by parts to derive the Euler–Lagrange equations. The operator that appears is the **formal adjoint** d_A^*, defined by

$$\langle \mathrm{d}_A \omega, \eta\rangle = \langle \omega, \mathrm{d}_A^* \eta\rangle \quad \text{for all } \omega \in \Omega^{p-1},\ \eta \in \Omega^p. \tag{8.171}$$

With it, (8.170) reads $\langle \mathrm{d}_A^* F, a\rangle = 0$ for all a. Choosing $a = \mathrm{d}_A^* F$ gives $\left\|\mathrm{d}_A^* F\right\|^2 = 0$. Hence:

Key idea: The Yang–Mills equations (Nash (8.69)–(8.70))

A is a critical point of S if and only if

$$\mathrm{d}_A^* F = 0. \tag{8.172}$$

Together with the Bianchi identity, the curvature of a Yang–Mills connection satisfies the pair

$$\mathrm{d}_A F = 0, \qquad \mathrm{d}_A^* F = 0. \tag{8.173}$$

In components, $\mathrm{d}_A^* F = 0$ is $D_\mu F_{\mu\nu} = \partial_\mu F_{\mu\nu} + [A_\mu, F_{\mu\nu}] = 0$ (see the worked example in Section 8.3.3), the Euclidean version of $D_\mu F^{\mu\nu} = j^\nu$ with no sources.

8.3.3 A formula for d_A^* in four dimensions

Nash next proves that on 2-forms in dimension four,

$$\mathrm{d}_A^* = -\star\, \mathrm{d}_A\, \star\,. \tag{8.174}$$

The proof is Stokes' theorem on a well-chosen $(n-1)$-form, the same trick that proves $\int u\, v' = -\int u'\, v$ on a circle. Two small facts about the trace go in first.

Working definition 8.22 (graded trace identities)

For $\mathfrak{g}$-valued forms [forms whose components are matrices in $\mathfrak{g}$] α (degree p) and β (degree q):

(i) $\operatorname{tr}(\alpha \wedge \beta) = (-1)^{pq} \operatorname{tr}(\beta \wedge \alpha)$;
(ii) $\mathrm{d}\, \operatorname{tr}(\alpha \wedge \beta) = \operatorname{tr}(\mathrm{d}_A\alpha \wedge \beta) + (-1)^p \operatorname{tr}(\alpha \wedge \mathrm{d}_A\beta)$.

(i) is cyclicity of the matrix trace [$\operatorname{tr}(XY) = \operatorname{tr}(YX)$ for matrices X, Y], plus the sign from reordering the 1-forms. (ii) says that inside a trace, d_A obeys the same Leibniz rule as d: the extra A-terms cancel.

Let us check (ii), since everything rests on it.

$$\begin{aligned}
&\operatorname{tr}(\mathrm{d}_A\alpha\wedge\beta)+(-1)^p\operatorname{tr}(\alpha\wedge\mathrm{d}_A\beta)\\
&\overset{(1)}{=}\operatorname{tr}(\mathrm{d}\alpha\wedge\beta)+(-1)^p\operatorname{tr}(\alpha\wedge\mathrm{d}\beta)+\operatorname{tr}(A\wedge\alpha\wedge\beta)-(-1)^p\operatorname{tr}(\alpha\wedge A\wedge\beta)\\
&\qquad+(-1)^p\operatorname{tr}(\alpha\wedge A\wedge\beta)-(-1)^{p+q}\operatorname{tr}(\alpha\wedge\beta\wedge A)\\
&\overset{(2)}{=}\mathrm{d}\operatorname{tr}(\alpha\wedge\beta)+\operatorname{tr}(A\wedge\alpha\wedge\beta)-(-1)^{p+q}(-1)^{p+q}\operatorname{tr}(A\wedge\alpha\wedge\beta)\\
&\overset{(3)}{=}\mathrm{d}\operatorname{tr}(\alpha\wedge\beta).
\end{aligned}$$

(1) Expand both d_A's using (8.164), with degree p for α and degree q for β.
(2) The ordinary Leibniz rule for d (the trace commutes with d); the two middle terms cancel; for the last term use (i) with the 1-form A and the $(p+q)$-form $\alpha\wedge\beta$.
(3) $(-1)^{2(p+q)}=1$, so the A-terms cancel.

Now the adjoint. Take $\omega\in\Omega^{p-1}(M,\operatorname{ad}P)$ and $\eta\in\Omega^p(M,\operatorname{ad}P)$ on a closed n-manifold. Then $\omega\wedge\star\eta$ is an $(n-1)$-form, and

$$\begin{aligned}
0&\overset{(1)}{=}\int_M\mathrm{d}\operatorname{tr}(\omega\wedge\star\eta)\\
&\overset{(2)}{=}\int_M\operatorname{tr}(\mathrm{d}_A\omega\wedge\star\eta)+(-1)^{p-1}\int_M\operatorname{tr}(\omega\wedge\mathrm{d}_A\star\eta)\\
&\overset{(3)}{=}-\langle\mathrm{d}_A\omega,\eta\rangle+(-1)^{p-1}(-1)^{(p-1)(n-p+1)}\int_M\operatorname{tr}\big(\omega\wedge\star\star\,\mathrm{d}_A\star\eta\big)\\
&\overset{(4)}{=}-\langle\mathrm{d}_A\omega,\eta\rangle-(-1)^{p-1}(-1)^{(p-1)(n-p+1)}\big\langle\omega,\star\mathrm{d}_A\star\eta\big\rangle.
\end{aligned}\tag{8.175}$$

(1) Stokes: M has no boundary, so the integral of an exact form is zero.
(2) Identity (ii) with $\alpha=\omega$ (degree $p-1$) and $\beta=\star\eta$.
(3) Definition (8.165) for the first term. In the second term, $\mathrm{d}_A\star\eta$ is an $(n-p+1)$-form, and on k-forms $\star\star=(-1)^{k(n-k)}$ in Riemannian signature; with $k=n-p+1$ this is $(-1)^{(p-1)(n-p+1)}$, so we may insert $\star\star$ at the cost of that sign.
(4) Definition (8.165) again, now with the $(p-1)$-forms ω and $\star\mathrm{d}_A\star\eta$.

Comparing with (8.171) we read off, on p-forms,

$$\mathrm{d}_A^{*}=(-1)^p(-1)^{(p-1)(n-p+1)}\ \star\,\mathrm{d}_A\star=(-1)^{n(p+1)+1}\ \star\,\mathrm{d}_A\,\star\,.\tag{8.176}$$

(The second form follows because $p+(p-1)(n-p+1)\equiv n(p+1)+1 \pmod 2$, which you can check by expanding.) In $n=4$ the sign is -1 for every p, and in particular on 2-forms we get (8.174). The Yang–Mills equation (8.172) becomes, after applying $\star$ (which is invertible),

$$\mathrm{d}_A\star F=0.\tag{8.177}$$

This is Nash's (8.75).

Worked example 8.44: d^* in components on $\mathbb{R}^4$

Take the trivial connection and the 2-form $\omega=f\,\mathrm{d}x^1\wedge\mathrm{d}x^2$, i.e. $\omega_{12}=-\omega_{21}=f$ and all other components zero. The orientation is $\mathrm{vol}=\mathrm{d}x^1\wedge\mathrm{d}x^2\wedge\mathrm{d}x^3\wedge\mathrm{d}x^4$. Then

$$\begin{aligned}
-\star\,\mathrm{d}\star\omega&\overset{(1)}{=}-\star\,\mathrm{d}\big(f\,\mathrm{d}x^3\wedge\mathrm{d}x^4\big)\\
&\overset{(2)}{=}-\star\,\big(\partial_1 f\,\mathrm{d}x^1\wedge\mathrm{d}x^3\wedge\mathrm{d}x^4+\partial_2 f\,\mathrm{d}x^2\wedge\mathrm{d}x^3\wedge\mathrm{d}x^4\big)
\end{aligned}$$

$$\overset{(3)}{=} -(\partial_1 f\, \mathrm{d}x^2 - \partial_2 f\, \mathrm{d}x^1)$$
$$= \partial_2 f\, \mathrm{d}x^1 - \partial_1 f\, \mathrm{d}x^2.$$

(1) $\star(\mathrm{d}x^1 \wedge \mathrm{d}x^2) = \mathrm{d}x^3 \wedge \mathrm{d}x^4$, because $\mathrm{d}x^1 \wedge \mathrm{d}x^2 \wedge \mathrm{d}x^3 \wedge \mathrm{d}x^4 = \mathrm{vol}$.
(2) Only $\partial_1 f$ and $\partial_2 f$ survive the wedge with $\mathrm{d}x^3 \wedge \mathrm{d}x^4$.
(3) $\star(\mathrm{d}x^1 \wedge \mathrm{d}x^3 \wedge \mathrm{d}x^4) = \mathrm{d}x^2$ since $\mathrm{d}x^1 \wedge \mathrm{d}x^3 \wedge \mathrm{d}x^4 \wedge \mathrm{d}x^2 = +\mathrm{vol}$ (two transpositions), and $\star(\mathrm{d}x^2 \wedge \mathrm{d}x^3 \wedge \mathrm{d}x^4) = -\mathrm{d}x^1$ since $\mathrm{d}x^2 \wedge \mathrm{d}x^3 \wedge \mathrm{d}x^4 \wedge \mathrm{d}x^1 = -\mathrm{vol}$ (three transpositions).

Compare with $(\mathrm{d}^*\omega)_\nu = -\partial_\mu \omega_{\mu\nu}$: for $\nu = 1$ it gives $-\partial_2\omega_{21} = \partial_2 f$, and for $\nu = 2$ it gives $-\partial_1\omega_{12} = -\partial_1 f$. They agree. So d^* on 2-forms is "minus the divergence on the first index", and with a connection, $(\mathrm{d}_A^* F)_\nu = -D_\mu F_{\mu\nu}$.

8.3.4 A non-linear version of Hodge theory

Compare the pair (8.173) with the definition of a harmonic form (Nash (2.50); see Chapter 2):

$$\mathrm{d}\omega = 0, \qquad \mathrm{d}^*\omega = 0. \tag{8.178}$$

For $G = \mathrm{U}(1)$ the connection is $A = i\mathcal{A}$ with $\mathcal{A}$ a real 1-form, all commutators vanish, $\mathrm{d}_A = \mathrm{d}$, and (8.173) is literally (8.178) for F: the source-free Maxwell equations say the field strength is a harmonic 2-form. Hodge theory then tells us that in each de Rham class there is exactly one harmonic representative. The class of $F/2\pi i$ is fixed by the bundle (it is c_1, see Chapter T2), so on a closed M each U(1) bundle carries exactly one Maxwell field strength, up to gauge. The Maxwell problem is completely linear.

For non-Abelian G, the equations look the same, but d_A itself depends on the unknown A. So (8.173) is a *non-linear* harmonic condition. Two consequences, both visible below: solutions are no longer unique in their topological class, and they come in finite-dimensional families.

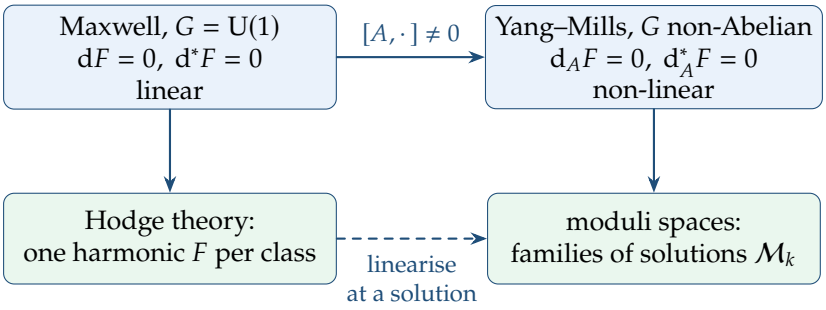


Figure 8.30: The Yang–Mills equations are the non-linear cousin of the harmonic-form equations. In the linear (Maxwell) case each cohomology class has one solution. In the non-linear case we get families of solutions. The tangent space to a family is computed by a *linear* Hodge-type problem at one solution (the dashed arrow), which is how Nash will compute $\dim \mathcal{M}_k$.

Worked example 8.45: Constant non-Abelian fields are not solutions

On the flat torus T^4 (or on $\mathbb{R}^4$), try the simplest non-Abelian ansatz: constant matrices $A_1 = \alpha e_1$, $A_2 = \beta e_2$, $A_3 = A_4 = 0$, with real constants α, β. For a Maxwell field any constant F solves the equations. Here:

$$F_{12} \overset{(1)}{=} [A_1, A_2] = \alpha\beta\,[e_1, e_2]$$
$$\overset{(2)}{=} 2\alpha\beta\, e_3,$$

(1) The fields are constant, so only the commutator term of (8.162) survives.
(2) $e_1e_2 = (-i)^2\sigma_1\sigma_2 = -i\sigma_3 = e_3$ and $e_2e_1 = -e_3$, so $[e_1, e_2] = 2e_3$. (These are the quaternion rules $ij = k = -ji$.)

and every other $F_{\mu\nu}$ with $\mu < \nu$ vanishes. The Yang–Mills equation for $\nu = 2$ is

$$\begin{aligned} -(\mathrm{d}_A^* F)_2 = D_\mu F_{\mu 2} &\overset{(1)}{=} [A_1, F_{12}] \\ &\overset{(2)}{=} 2\alpha^2\beta\,[e_1, e_3] \\ &\overset{(3)}{=} -4\alpha^2\beta\, e_2. \end{aligned}$$

(1) $\partial_\mu F_{\mu 2} = 0$ for constant fields, and only $\mu = 1$ has $F_{\mu 2} \neq 0$.
(2) Substitute A_1 and F_{12}.
(3) $[e_1, e_3] = 2e_1e_3 = -2e_2$ (quaternions: $ik = -j$).

This vanishes only if $\alpha\beta = 0$, i.e. only if the field is Abelian. The self-interaction $[A_\mu, F_{\mu\nu}]$ acts like a source: the gluon field is charged under itself. This is the non-linearity in action.

Try it: The Yang–Mills equations are gauge covariant

Show that if A solves $\mathrm{d}_A^* F = 0$ then so does A_g. *Answer:* $F_{A_g} = g^{-1}Fg$ and $\mathrm{d}_{A_g}(g^{-1}\omega g) = g^{-1}(\mathrm{d}_A\omega)g$, while $\star$ commutes with conjugation. So $\star\mathrm{d}_{A_g} \star F_{A_g} = g^{-1}(\star\mathrm{d}_A \star F)g = 0$. Solutions come in gauge orbits; this is why we will divide by the gauge group.

Reading Nash (pp. 227–229, (8.63)–(8.74))

Nash defines $S = \|F\|^2 = -\int \mathrm{tr}(F \wedge \star F)$ (8.63), varies along $A_t = A + ta$ (8.64)–(8.67), and gets $\mathrm{d}_A^* F = 0$ (8.69), paired with Bianchi (8.70). His (8.72)–(8.74) is the Stokes argument of (8.175). He writes the pairing as $\int \mathrm{tr}(\omega \wedge \star\eta)$ without the minus sign of our (8.165); the sign is an overall factor and does not affect $\mathrm{d}_A^* = -\star\,\mathrm{d}_A\star$. His general sign $(-1)^{1+p+(4-p)p}$ is our (8.176) with $n = 4$ (he labels by the degree of the form d_A^* acts on); it equals -1 for every p. The remark that Yang–Mills is "a kind of non-linear harmonic condition" is Figure 8.30.

Recap

- Conventions: anti-Hermitian A, $F = \mathrm{d}A + A \wedge A$, $\langle\alpha, \beta\rangle = -\int \mathrm{tr}(\alpha \wedge \star\beta)$, $S = \|F\|^2 = g^2 S_{\text{phys}}$.
- Along a line $A + ta$: $F \mapsto F + t\,\mathrm{d}_A a + t^2 a \wedge a$ exactly; $\dot{S}(0) = 2\langle F, \mathrm{d}_A a\rangle$.
- Critical points: $\mathrm{d}_A^* F = 0$; always $\mathrm{d}_A F = 0$. In 4D, $\mathrm{d}_A^* = -\star\,\mathrm{d}_A\star$, so Yang–Mills is $\mathrm{d}_A \star F = 0$.
- Abelian case: F harmonic, one solution per class. Non-Abelian: a non-linear harmonic condition, with families of solutions.

8.4 Self-duality and the Bogomolny bound

The Yang–Mills equation $\mathrm{d}_A \star F = 0$ is a second-order, non-linear partial differential equation for A. Nobody can solve that in general. But Nash observes (just after his (8.75)) that there is a trick. The Bianchi identity $\mathrm{d}_A F = 0$ holds for *every* connection. So if we can arrange $\star F = \pm F$, then $\mathrm{d}_A \star F = \pm \mathrm{d}_A F = 0$ comes for free. The second-order equation has been replaced by a first-order one. Even better, these first-order solutions turn out to be the *absolute minima* of the

action in their topological sector. This section explains both statements. The key player is the Hodge star on 2-forms in four dimensions, which has a special property that holds in no other dimension: it maps 2-forms to 2-forms.

8.4.1 The Hodge star on 2-forms in four dimensions

We work first on $\mathbb{R}^4$ with the flat metric and orientation

$$\mathrm{vol} = \mathrm{d}x^1 \wedge \mathrm{d}x^2 \wedge \mathrm{d}x^3 \wedge \mathrm{d}x^4.$$

We abbreviate $\mathrm{d}x^{\mu\nu} = \mathrm{d}x^\mu \wedge \mathrm{d}x^\nu$. Recall (Chapter T1) that $\star$ is defined by $\alpha \wedge \star\beta = \langle \alpha, \beta \rangle \,\mathrm{vol}$. For a basis 2-form this means: $\star \mathrm{d}x^{\mu\nu} = \pm \mathrm{d}x^{\rho\sigma}$, where $\{\rho, \sigma\}$ is the complementary pair, and the sign is chosen so that $\mathrm{d}x^{\mu\nu} \wedge \star \mathrm{d}x^{\mu\nu} = +\mathrm{vol}$.

Worked example 8.46: The matrix of $\star$ on $\Lambda^2\mathbb{R}^4$

There are $\binom{4}{2} = 6$ basis 2-forms. Take $\mathrm{d}x^{13}$. Its complement is $\{2, 4\}$, and

$$\begin{aligned} \mathrm{d}x^{13} \wedge \mathrm{d}x^{24} &\overset{(1)}{=} \mathrm{d}x^1 \wedge \mathrm{d}x^3 \wedge \mathrm{d}x^2 \wedge \mathrm{d}x^4 \\ &\overset{(2)}{=} -\mathrm{d}x^1 \wedge \mathrm{d}x^2 \wedge \mathrm{d}x^3 \wedge \mathrm{d}x^4 = -\mathrm{vol}, \end{aligned}$$

(1) Definition of the abbreviation.
(2) Swap the adjacent 1-forms $\mathrm{d}x^3$ and $\mathrm{d}x^2$; one transposition costs one minus sign.

so $\star \mathrm{d}x^{13} = -\mathrm{d}x^{24}$. Doing the same for all six:

$$\begin{aligned} &\star \mathrm{d}x^{12} = \mathrm{d}x^{34}, \quad \star \mathrm{d}x^{13} = -\mathrm{d}x^{24}, \quad \star \mathrm{d}x^{14} = \mathrm{d}x^{23}, \\ &\star \mathrm{d}x^{23} = \mathrm{d}x^{14}, \quad \star \mathrm{d}x^{24} = -\mathrm{d}x^{13}, \quad \star \mathrm{d}x^{34} = \mathrm{d}x^{12}. \end{aligned}$$

In the ordered basis $(\mathrm{d}x^{12}, \mathrm{d}x^{13}, \mathrm{d}x^{14}, \mathrm{d}x^{23}, \mathrm{d}x^{24}, \mathrm{d}x^{34})$, the star is the 6×6 matrix

$$\star = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \qquad \star^2 = \mathbb{1}_6.$$

The square is the identity: each basis form is sent to its partner and back, with the same sign both times. This agrees with the general rule $\star\star = (-1)^{p(n-p)} = (-1)^4 = +1$ for $p = 2$, $n = 4$.

Since $\star^2 = 1$, the eigenvalues of $\star$ are ± 1, and the 6-dimensional space of 2-forms splits into two 3-dimensional eigenspaces:

$$\Lambda^2 = \Lambda^2_+ \oplus \Lambda^2_-, \qquad \Lambda^2_\pm = \{\omega : \star\omega = \pm\omega\}. \tag{8.179}$$

Forms in Λ^2_+ are called **self-dual** (SD), and forms in Λ^2_- **anti-self-dual** (ASD).

Working definition 8.23 (self-dual and anti-self-dual parts)

Any 2-form [at each point, an element of the 6-dimensional space Λ^2 above] splits as

$\omega = \omega^+ + \omega^-$ with

$$\omega^+ = \pi_+\omega = \tfrac{1}{2}(\omega + \star\omega), \qquad \omega^- = \pi_-\omega = \tfrac{1}{2}(\omega - \star\omega). \tag{8.180}$$

Test: $\star\omega^\pm = \frac{1}{2}(\star\omega \pm \omega) = \pm\omega^\pm$, using $\star^2 = 1$. The maps $\pi_\pm$ are projectors [each keeps its own eigenspace $\Lambda^2_\pm$ unchanged and sends the other one to 0]: $\pi_\pm^2 = \pi_\pm$, $\pi_+\pi_- = 0$, $\pi_+ + \pi_- = 1$. Nash's (8.95) is π_-.

An explicit basis: the 't Hooft symbols

Pairing each $\mathrm{d}x^{a4}$ ($a = 1, 2, 3$) with its complement gives bases of the two eigenspaces:

$$\begin{aligned} \omega_1^+ &= \mathrm{d}x^{14} + \mathrm{d}x^{23}, & \omega_2^+ &= \mathrm{d}x^{24} + \mathrm{d}x^{31}, & \omega_3^+ &= \mathrm{d}x^{34} + \mathrm{d}x^{12}, \\ \omega_1^- &= -\mathrm{d}x^{14} + \mathrm{d}x^{23}, & \omega_2^- &= -\mathrm{d}x^{24} + \mathrm{d}x^{31}, & \omega_3^- &= -\mathrm{d}x^{34} + \mathrm{d}x^{12}. \end{aligned} \tag{8.181}$$

(The table above shows $\star\mathrm{d}x^{14} = \mathrm{d}x^{23}$, and $\star\mathrm{d}x^{24} = -\mathrm{d}x^{13} = \mathrm{d}x^{31}$, so these are indeed eigenvectors.) Physicists write the coefficients as the **'t Hooft symbols**:

$$\omega_a^+ = \tfrac{1}{2}\eta^a_{\mu\nu}\,\mathrm{d}x^{\mu\nu}, \qquad \omega_a^- = \tfrac{1}{2}\bar\eta^a_{\mu\nu}\,\mathrm{d}x^{\mu\nu}, \tag{8.182}$$

$$\eta^a_{jk} = \bar\eta^a_{jk} = \varepsilon_{ajk}, \qquad \eta^a_{j4} = -\eta^a_{4j} = \delta_{aj}, \qquad \bar\eta^a_{j4} = -\bar\eta^a_{4j} = -\delta_{aj} \qquad (j,k = 1,2,3).$$

Each η^a is an antisymmetric 4×4 matrix. For example

$$\eta^1 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{pmatrix}, \qquad \bar\eta^1 = \begin{pmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

The component form of self-duality is $\frac{1}{2}\varepsilon_{\mu\nu\rho\sigma}\eta^a_{\rho\sigma} = \eta^a_{\mu\nu}$, and similarly with a minus sign for $\bar\eta$. We will use two identities that follow by direct checking of a few components:

$$\eta^a_{\mu\nu}\eta^b_{\mu\nu} = 4\delta^{ab}, \qquad \eta^a_{\mu\nu}\bar\eta^b_{\mu\nu} = 0, \qquad \eta^a_{\mu\nu}\eta^b_{\mu\rho} = \delta^{ab}\delta_{\nu\rho} + \varepsilon^{abc}\eta^c_{\nu\rho}. \tag{8.183}$$

The first says each η^a has four non-zero entries of modulus one; the second is the orthogonality of Λ^2_+ and Λ^2_-; the third says the η^a multiply like quaternions (we will see why in Section 8.5).

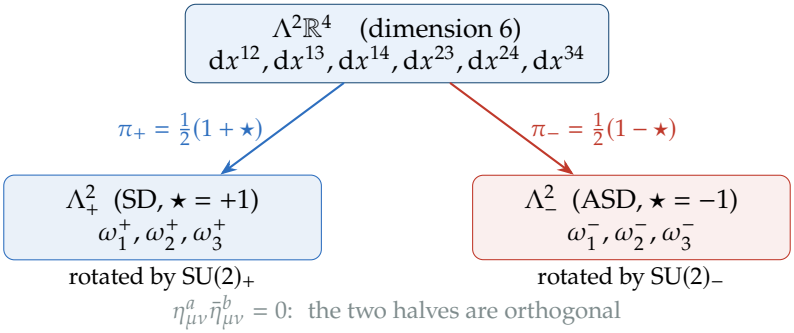


Figure 8.31: The six-dimensional space of 2-forms on $\mathbb{R}^4$ splits into two three-dimensional halves. The rotation group $\mathrm{SO}(4) = (\mathrm{SU}(2)_+ \times \mathrm{SU}(2)_-)/\mathbb{Z}_2$ acts on each half separately, as SO(3); this is the statement $\Lambda^2_\pm \cong$ adjoint of $\mathrm{SU}(2)_\pm$. A self-dual field strength has only three independent components per colour, not six.

Physical picture: self-duality is $\vec{E} = \vec{B}$

Label the components of a Euclidean field strength like the electric and magnetic fields: $E_j = F_{j4}$ and $B_j = \frac{1}{2}\varepsilon_{jkl}F_{kl}$ (so $B_1 = F_{23}$). The basis (8.181) shows

$$F = \sum_j E_j \, \mathrm{d}x^{j4} + \sum_j B_j \, \tfrac{1}{2}\varepsilon_{jkl}\mathrm{d}x^{kl}, \qquad \star F = \sum_j B_j \, \mathrm{d}x^{j4} + \sum_j E_j \, \tfrac{1}{2}\varepsilon_{jkl}\mathrm{d}x^{kl}.$$

So $\star$ swaps $\vec{E}$ and $\vec{B}$, and

$$F = \star F \iff \vec{E} = \vec{B}, \qquad F = -\star F \iff \vec{E} = -\vec{B}.$$

For a non-Abelian field this holds colour by colour: $\vec{E}^a = \pm\vec{B}^a$.

Watch out: There are no real instantons in Minkowski space

In Lorentzian signature $(-,+,+,+)$ the rule is $\star\star = -(-1)^{p(n-p)} = -1$ on 2-forms, so the eigenvalues of $\star$ are $\pm i$ and a real F can satisfy $\star F = \lambda F$ only if $F = 0$. In the $\vec{E}, \vec{B}$ language, the Lorentzian star sends $\vec{E} \mapsto \vec{B}$ and $\vec{B} \mapsto -\vec{E}$, and "self-duality" is $\vec{E} = i\vec{B}$. That is why instantons live in Euclidean signature: they describe tunnelling (imaginary time), not real-time classical motion.

Conformal invariance

On a general Riemannian 4-manifold the star depends on the metric. But on 2-forms in four dimensions it depends only on the *conformal class* of the metric: the metrics g and $\Omega^2 g$ (with $\Omega(x) > 0$ any function) give the same $\star$ on 2-forms.

$$\begin{aligned}(\star_{\Omega^2 g}\,\omega)_{\mu\nu} &\overset{(1)}{=} \tfrac{1}{2}\sqrt{\det(\Omega^2 g)}\;\varepsilon_{\mu\nu\rho\sigma}\,(\Omega^{-2}g^{\rho\alpha})(\Omega^{-2}g^{\sigma\beta})\,\omega_{\alpha\beta}\\ &\overset{(2)}{=} \tfrac{1}{2}\,\Omega^4\sqrt{\det g}\;\Omega^{-4}\,\varepsilon_{\mu\nu\rho\sigma}g^{\rho\alpha}g^{\sigma\beta}\,\omega_{\alpha\beta}\\ &\overset{(3)}{=} (\star_g\,\omega)_{\mu\nu}.\end{aligned}$$

(1) The coordinate formula for the star of a 2-form in four dimensions: raise both indices with the inverse metric, contract with the Levi-Civita symbol, multiply by $\sqrt{\det g}$. The inverse of $\Omega^2 g$ is $\Omega^{-2}g^{-1}$.

(2) $\det(\Omega^2 g) = \Omega^8 \det g$ for a 4×4 matrix, so the square root gives Ω^4; two inverse metrics give Ω^{-4}.

(3) The powers of Ω cancel exactly. In n dimensions on p-forms the leftover power would be Ω^{n-2p}; it vanishes only for $n = 2p$.

The same count shows that the action density $\mathrm{tr}(F \wedge \star F)$ is conformally invariant, so the Yang–Mills action has no scale in four dimensions: *classical Yang–Mills theory in 4D is conformally invariant*.

This matters because the **stereographic projection** $S^4 \setminus \{\text{north pole}\} \to \mathbb{R}^4$ is conformal (Figure 8.32):

$$g_{S^4} = \frac{4}{(1+|x|^2)^2}\, g_{\mathbb{R}^4} \tag{8.184}$$

for the unit sphere. So a self-dual field on $\mathbb{R}^4$ is the same thing as a self-dual field on S^4 minus a point, with the same action. If the action is finite, a theorem of Uhlenbeck (the **removable**

singularity theorem) says the connection extends smoothly, after a gauge transformation, over the missing point. Finite-action instantons on $\mathbb{R}^4$ are therefore exactly instantons on S^4. This is how the compact manifold S^4 of the mathematicians and the flat $\mathbb{R}^4$ of the physicists become the same problem.

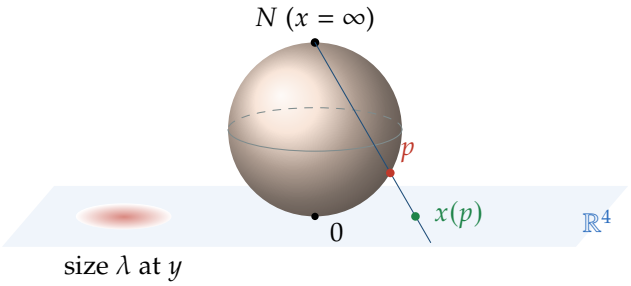


Figure 8.32: Stereographic projection from the north pole N identifies $S^4 \setminus \{N\}$ with $\mathbb{R}^4$. It is conformal, and on 2-forms in 4D the Hodge star sees only the conformal class, so self-duality and the action are the same on both sides. The point at infinity of $\mathbb{R}^4$ is the single point N; finite action is what lets the field extend smoothly over it. An instanton of size λ centred at $y \in \mathbb{R}^4$ is a lump of action density that looks the same, in this sense, on the plane and on the sphere.

8.4.2 Self-dual fields solve the Yang–Mills equations

Everything above is pointwise linear algebra, so it carries over to $\mathfrak{g}$-valued 2-forms on any oriented Riemannian 4-manifold: $\star$ acts on the form part and leaves the matrix part alone. In particular every curvature splits as

$$F = F^+ + F^-, \qquad F^\pm = \tfrac{1}{2}(F \pm \star F). \tag{8.185}$$

Key idea: First order implies second order (Nash, after (8.75))

If $F = \pm \star F$, then A solves the Yang–Mills equations:

$$\begin{aligned} \mathrm{d}_A \star F &\overset{(1)}{=} \mathrm{d}_A(\pm F) \\ &\overset{(2)}{=} \pm\, \mathrm{d}_A F \\ &\overset{(3)}{=} 0. \end{aligned}$$

(1) Apply $\star$ to $F = \pm \star F$ and use $\star^2 = 1$ on 2-forms: $\star F = \pm F$.
(2) d_A is linear.
(3) The Bianchi identity, proved in the worked example of Section 8.3.1.

We have solved a non-linear second-order equation, $\mathrm{d}_A \star F = 0$, by solving a non-linear *first-order* one, $F = \pm \star F$.

The converse is false in general: there are Yang–Mills connections on S^4 that are neither SD nor ASD (Sibner, Sibner and Uhlenbeck, 1989). But the self-dual ones are the most important, because of the energy bound we derive next.

Worked example 8.47: Self-duality in components

Write $F = \frac{1}{2}F_{\mu\nu}\mathrm{d}x^{\mu\nu}$ on $\mathbb{R}^4$. Using the table of the first worked example of Section 8.4.1,

$F = \star F$ means

$$F_{12}\,\mathrm{d}x^{12} + F_{34}\,\mathrm{d}x^{34} + \ldots \overset{(1)}{=} \star\big(F_{12}\,\mathrm{d}x^{12} + F_{34}\,\mathrm{d}x^{34} + \ldots\big) \overset{(2)}{=} F_{12}\,\mathrm{d}x^{34} + F_{34}\,\mathrm{d}x^{12} + \ldots,$$

(1) $F = \sum_{\mu<\nu} F_{\mu\nu}\mathrm{d}x^{\mu\nu}$, written out for the pair $\{12, 34\}$; the other pairs work the same way.
(2) $\star\mathrm{d}x^{12} = \mathrm{d}x^{34}$ and $\star\mathrm{d}x^{34} = \mathrm{d}x^{12}$.

so comparing coefficients, and doing the same for the other two pairs,

$$F_{12} = F_{34}, \qquad F_{13} = -F_{24}, \qquad F_{14} = F_{23}. \tag{8.186}$$

Three matrix equations: they say that the components of F along the three ASD basis forms ω_a^- vanish, i.e. $F^- = 0$. For SU(2) that is $3\times 3 = 9$ real equations on $4\times 3 = 12$ real functions A_μ^a, and 3 of the 12 are gauge. So the count is "9 equations for 9 unknowns": a determined system, as it should be for an elliptic problem. We make this precise in Section 8.8.

8.4.3 The instanton number

Topology enters through the second Chern class. For an SU(N) bundle E the total Chern class is $c(E) = \det\big(1 + \frac{i}{2\pi}F\big)$ (see Chapter T2 and Chapter 4). We need its 4-form part.

$$\begin{aligned}\det\Big(1 + \tfrac{i}{2\pi}F\Big) &\overset{(1)}{=} \exp\operatorname{tr}\log\Big(1 + \tfrac{i}{2\pi}F\Big)\\ &\overset{(2)}{=} \exp\Big(\tfrac{i}{2\pi}\operatorname{tr}F - \tfrac{1}{2}\big(\tfrac{i}{2\pi}\big)^2\operatorname{tr}(F\wedge F) + \ldots\Big)\\ &\overset{(3)}{=} \exp\Big(\tfrac{1}{8\pi^2}\operatorname{tr}(F\wedge F) + \ldots\Big)\\ &\overset{(4)}{=} 1 + \tfrac{1}{8\pi^2}\operatorname{tr}(F\wedge F) + (\text{6-forms and higher}).\end{aligned}$$

(1) The matrix identity det = exp tr log; it is fine for matrices of even forms, which commute with each other.
(2) Expand $\log(1+X) = X - \frac{1}{2}X^2 + \ldots$.
(3) $\operatorname{tr}F = 0$ for $\mathfrak{su}(N)$, and $-\frac{1}{2}(i/2\pi)^2 = \frac{1}{8\pi^2}$.
(4) Expand the exponential; on a 4-manifold nothing beyond 4-forms survives.

So $c_2(E) = \frac{1}{8\pi^2}\operatorname{tr}(F\wedge F)$, and we define the **instanton number**

$$k = -c_2(E)[M] = -\frac{1}{8\pi^2}\int_M \operatorname{tr}(F\wedge F). \tag{8.187}$$

This is an integer, since c_2 is an integral class, and it does not change under continuous deformations of A, since the integrand changes by an exact form (Chapter T2; part 8a derived the transgression formula $\operatorname{tr}(F\wedge F) = \mathrm{d}\,\operatorname{tr}(A\wedge\mathrm{d}A + \frac{2}{3}A\wedge A\wedge A)$). On S^4 the integer k classifies SU(2) bundles completely.

Watch out: The sign of k

Nash's (8.77) writes $k = \frac{1}{8\pi^2}\int\operatorname{tr}(F\wedge F)$ in its first line, then $(\|F^+\|^2 - \|F^-\|^2)/8\pi^2$ in its last. With anti-Hermitian F these two differ by a sign (we show this below), and the first

line is missing a minus sign; his words "k is minus $c_2(F)$" agree with (8.187). With our conventions, orientation dx^{1234} and $k = -c_2$, the BPST instanton of Section 8.5 is self-dual and has $k = +1$.

8.4.4 The action splits; the Bogomolny bound

Two facts do all the work.

Fact 1: SD and ASD forms are orthogonal. Let $\alpha \in \Lambda^2_+$ and $\beta \in \Lambda^2_-$ be $\mathfrak{g}$-valued. Then

$$\begin{aligned}\operatorname{tr}(\alpha \wedge \star\beta) &\overset{(1)}{=} -\operatorname{tr}(\alpha \wedge \beta)\\ &\overset{(2)}{=} -\operatorname{tr}(\beta \wedge \alpha)\\ &\overset{(3)}{=} -\operatorname{tr}(\beta \wedge \star\alpha)\\ &\overset{(4)}{=} -\operatorname{tr}(\alpha \wedge \star\beta),\end{aligned}$$

(1) $\star\beta = -\beta$.
(2) Graded cyclicity of the trace with two 2-forms: sign $(-1)^{2\cdot 2} = +1$.
(3) $\alpha = \star\alpha$.
(4) Symmetry of the pointwise inner product, $\operatorname{tr}(\beta\wedge\star\alpha) = \operatorname{tr}(\alpha\wedge\star\beta)$ (both are $\frac{1}{2}\operatorname{tr}(\alpha_{\mu\nu}\beta_{\mu\nu})\mathrm{vol}$).

A quantity equal to its own negative is zero. So $\langle\alpha,\beta\rangle = 0$, pointwise and after integration.

Fact 2: $\operatorname{tr}(F\wedge F)$ *sees the difference.* Using $F\wedge F = F\wedge\star\star F$ and splitting:

$$\begin{aligned}-\int_M \operatorname{tr}(F\wedge F) &\overset{(1)}{=} -\int_M \operatorname{tr}\big((F^+ + F^-)\wedge(F^+ + F^-)\big)\\ &\overset{(2)}{=} -\int_M \operatorname{tr}(F^+\wedge F^+) - \int_M \operatorname{tr}(F^-\wedge F^-)\\ &\overset{(3)}{=} -\int_M \operatorname{tr}(F^+\wedge\star F^+) + \int_M \operatorname{tr}(F^-\wedge\star F^-)\\ &\overset{(4)}{=} \left\|F^+\right\|^2 - \|F^-\|^2 .\end{aligned} \tag{8.188}$$

(1) The splitting (8.185).
(2) The cross terms are $\operatorname{tr}(F^+\wedge F^-) = -\operatorname{tr}(F^+\wedge\star F^-)$, which vanishes pointwise by Fact 1 (and the same for $F^-\wedge F^+$).
(3) $F^+ = \star F^+$ and $F^- = -\star F^-$.
(4) The definition (8.165) of the norm.

Meanwhile the action itself is a sum:

$$\begin{aligned}S = \|F\|^2 &\overset{(1)}{=} \left\|F^+\right\|^2 + \|F^-\|^2 + 2\langle F^+, F^-\rangle\\ &\overset{(2)}{=} \left\|F^+\right\|^2 + \|F^-\|^2 .\end{aligned} \tag{8.189}$$

(1) Expand the norm of a sum using bilinearity and symmetry.
(2) Fact 1.

This is Nash's (8.76). Combining with (8.187) and (8.188), Nash's (8.77) (with the corrected sign) reads

$$8\pi^2 k = \left\|F^+\right\|^2 - \|F^-\|^2 , \qquad S = \left\|F^+\right\|^2 + \|F^-\|^2 . \tag{8.190}$$

Key idea: The Bogomolny bound (Nash (8.78)–(8.79))

For every connection with instanton number k,

$$S(A) \;\geq\; 8\pi^2|k|, \tag{8.191}$$

with equality if and only if $F^- = 0$ (self-dual, $F = \star F$, possible when $k \geq 0$) or $F^+ = 0$ (anti-self-dual, $F = -\star F$, possible when $k \leq 0$). *Why:* for real numbers $a^2 + b^2 \geq |a^2 - b^2|$, with equality iff $a = 0$ or $b = 0$. Apply it to $a = \|F^+\|$, $b = \|F^-\|$ in (8.190).

So in each topological sector the SD (or ASD) connections are the absolute minima of the action. They are stable: no small deformation lowers the action. In physical units the minimal action is $S_{\text{phys}} = 8\pi^2|k|/g^2$, and each instanton contributes the famous suppression factor $e^{-8\pi^2/g^2}$ to the Euclidean path integral.

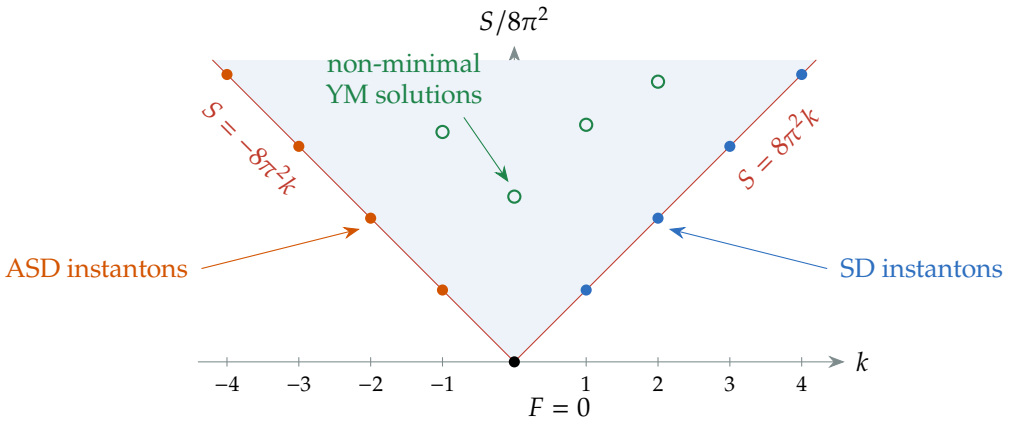


Figure 8.33: The Bogomolny bound. Every connection lies in the shaded wedge $S \geq 8\pi^2|k|$. The solid dots on the edges are the absolute minima: SD instantons for $k > 0$, ASD for $k < 0$, the flat connection at $k = 0$. Hollow circles are Yang–Mills critical points that are neither: they are saddle points of S, not minima. Reversing the orientation of M reflects the picture in the vertical axis.

Worked example 8.48: The Bogomolny trick: completing the square

Physicists usually derive (8.191) in one line, by completing a square. Since $\star$ is an isometry on 2-forms ($\|\star F\| = \|F\|$),

$$\begin{aligned} 0 \overset{(1)}{\leq} \|F - \star F\|^2 &\overset{(2)}{=} \|F\|^2 + \|\star F\|^2 - 2\langle F, \star F\rangle \\ &\overset{(3)}{=} 2S + 2\int_M \operatorname{tr}(F \wedge \star\star F) \\ &\overset{(4)}{=} 2S + 2\int_M \operatorname{tr}(F \wedge F) \\ &\overset{(5)}{=} 2S - 16\pi^2 k. \end{aligned}$$

(1) A norm squared is non-negative; it is zero iff $F = \star F$.
(2) Expand the square.
(3) $\|\star F\|^2 = \|F\|^2 = S$, and $\langle F, \star F\rangle = -\int \operatorname{tr}(F \wedge \star\star F)$ by (8.165).
(4) $\star\star = 1$ on 2-forms.
(5) Definition (8.187) of k.

So $S \geq 8\pi^2 k$, with equality iff F is SD. Replacing $F - \star F$ by $F + \star F$ gives $S \geq -8\pi^2 k$, with

equality iff F is ASD. Together: $S \geq 8\pi^2|k|$.

Worked example 8.49: A $\mathrm{U}(1)$ check with numbers: constant fields on T^4

Let $M = T^4 = \mathbb{R}^4/(L\mathbb{Z})^4$, a flat torus of side L, and $G = \mathrm{U}(1)$. Then $F = if$ with f a real closed 2-form, and there is no trace. Take the constant field

$$f = \frac{2\pi}{L^2}\left(n\,\mathrm{d}x^{12} + m\,\mathrm{d}x^{34}\right), \qquad n, m \in \mathbb{Z}.$$

The integers are forced: the flux of f through the 2-torus in the 12 plane is $\int f = \frac{2\pi}{L^2} n\, L^2 = 2\pi n$, and Dirac quantisation (the first Chern class $[f/2\pi]$ is integral) requires $n \in \mathbb{Z}$; similarly for m. Now compute both sides of the bound. The action is

$$\begin{aligned} S = \int_{T^4} f \wedge \star f &\overset{(1)}{=} \left(\frac{2\pi}{L^2}\right)^2 \int_{T^4} (n\,\mathrm{d}x^{12} + m\,\mathrm{d}x^{34}) \wedge (n\,\mathrm{d}x^{34} + m\,\mathrm{d}x^{12}) \\ &\overset{(2)}{=} \left(\frac{2\pi}{L^2}\right)^2 (n^2 + m^2)\, L^4 \\ &= 4\pi^2(n^2 + m^2). \end{aligned}$$

(1) For U(1), $-\operatorname{tr}(F \wedge \star F) = -(if) \wedge \star(if) = f \wedge \star f$; use $\star \mathrm{d}x^{12} = \mathrm{d}x^{34}$ and $\star \mathrm{d}x^{34} = \mathrm{d}x^{12}$.
(2) $\mathrm{d}x^{12} \wedge \mathrm{d}x^{34} = \mathrm{d}x^{34} \wedge \mathrm{d}x^{12} = \mathrm{vol}$, the other products vanish, and $\int \mathrm{vol} = L^4$.

The topological number is

$$\begin{aligned} k = -\frac{1}{8\pi^2}\int_{T^4} F \wedge F &\overset{(1)}{=} \frac{1}{8\pi^2}\int_{T^4} f \wedge f \\ &\overset{(2)}{=} \frac{1}{8\pi^2}\left(\frac{2\pi}{L^2}\right)^2 2nm\, L^4 = nm. \end{aligned}$$

(1) $F \wedge F = (if) \wedge (if) = -f \wedge f$.
(2) $(n\,\mathrm{d}x^{12} + m\,\mathrm{d}x^{34})^{\wedge 2} = 2nm\,\mathrm{vol}$ (two cross terms).

So $S - 8\pi^2|k| = 4\pi^2(n^2 + m^2 - 2|nm|) = 4\pi^2(|n| - |m|)^2 \geq 0$, with equality exactly when $m = \pm n$, i.e. when $f = \frac{2\pi n}{L^2}(\mathrm{d}x^{12} \pm \mathrm{d}x^{34})$ is SD or ASD. The Bogomolny bound in this case is the inequality $n^2 + m^2 \geq 2|nm|$ between integers.

Worked example 8.50: Self-dual fields carry no energy–momentum

The Euclidean stress tensor of a Yang–Mills field is (physicists' units) $T_{\mu\nu} = F^a_{\mu\rho}F^a_{\nu\rho} - \frac{1}{4}\delta_{\mu\nu}F^a_{\rho\sigma}F^a_{\rho\sigma}$. In terms of $\vec{E}^a$, $\vec{B}^a$ (intuition box above), $F^a_{\rho\sigma}F^a_{\rho\sigma} = 2(|\vec{E}^a|^2 + |\vec{B}^a|^2)$, summed over a. Then

$$\begin{aligned} T_{44} &\overset{(1)}{=} F^a_{4j}F^a_{4j} - \tfrac{1}{2}\left(|\vec{E}|^2 + |\vec{B}|^2\right) \\ &\overset{(2)}{=} \tfrac{1}{2}\left(|\vec{E}|^2 - |\vec{B}|^2\right), \\ T_{j4} &\overset{(3)}{=} F^a_{jk}F^a_{4k} \overset{(4)}{=} -(\vec{E}^a \times \vec{B}^a)_j, \end{aligned}$$

(1) Definition of T with $\mu = \nu = 4$; $F_{44} = 0$.
(2) $F_{4j} = -E_j$, so $F^a_{4j}F^a_{4j} = |\vec{E}|^2$ (sums over a understood).
(3) Definition with $\mu = j$, $\nu = 4$; only $\rho = k$ spatial contributes.
(4) $F_{jk} = \varepsilon_{jkl}B_l$ and $F_{4k} = -E_k$.

If $\vec{E}^a = \pm\vec{B}^a$ for every a then $T_{44} = 0$ and $T_{j4} = 0$, and a similar two-line check gives $T_{jk} = 0$. An instanton exerts no stress. One consequence: an instanton on a curved 4-manifold does not back-react on the metric in Euclidean Einstein–Yang–Mills theory.

Physical picture: instantons as tunnelling

In the temporal gauge, the finite-action configurations on $\mathbb{R}^4$ interpolate between classical vacua $A = g^{-1}\mathrm{d}g$ at $x^4 = -\infty$ and $x^4 = +\infty$ whose winding numbers differ by k (part 8a). The instanton is the dominant tunnelling path between neighbouring vacua, and its amplitude is $e^{-8\pi^2/g^2}$. Adding the topological term $i\theta k$ to the Euclidean action weights the sectors by $e^{i\theta k}$: the θ-vacua. Nash does not discuss this physics, but it is why the moduli spaces we are about to study matter to a physicist: each point of $\mathcal{M}_k$ is a tunnelling path, and integrating over $\mathcal{M}_k$ is summing over them.

Watch out: SD versus ASD is a convention

Reversing the orientation of M changes the sign of $\star$ (the volume form changes sign), swapping Λ^2_+ and Λ^2_-, and also flips the sign of k. Nothing physical changes. Nash, like us, "deems" the minima with $k > 0$ to be self-dual. Much of the mathematical literature (Donaldson, Atiyah–Hitchin–Singer) uses ASD connections with $k > 0$ instead, because it fits the complex geometry of twistor space better. When comparing formulas, swap $+ \leftrightarrow -$ everywhere: for example b_2^- in Nash becomes b_2^+ in Donaldson–Kronheimer.

Reading Nash (pp. 229–230, (8.75)–(8.79); p. 233, (8.95))

After (8.75), Nash notes that $F = \mp\star F$ plus Bianchi solves the Yang–Mills equation, "solving a non-linear second order equation by solving a non-linear first order equation". He splits $F = F^+ + F^-$ (8.76) and asserts that "the crossed terms in the norm contribute zero": this is our Fact 1. (8.77) expresses k as $(\|F^+\|^2 - \|F^-\|^2)/8\pi^2$; see the sign pitfall above. (8.78)–(8.79) is the bound $S \geq 8\pi^2|k|$ from $a^2 + b^2 \geq |a^2 - b^2|$. The projector $\pi_-\omega = \frac{1}{2}(\omega - \star\omega)$ of (8.95) is our (8.180); Nash introduces it only later, to write the SD equation as $\pi_- F = 0$. Nash's wording "$F = \mp \star F$ corresponds to $F^\mp = 0$" uses the opposite pairing of signs to ours in the first formula; the content is the same: the minimum with $k > 0$ has $F^- = 0$.

Recap

- In 4D, $\star$ maps 2-forms to 2-forms with $\star^2 = 1$: $\Lambda^2 = \Lambda^2_+ \oplus \Lambda^2_-$, three dimensions each, bases given by the 't Hooft symbols η^a, $\bar{\eta}^a$.
- On 2-forms in 4D the star (hence the SD equation and the action) is conformally invariant; $\mathbb{R}^4$ and S^4 give the same instantons.
- SD or ASD plus Bianchi implies Yang–Mills.
- $k = -c_2 = -\frac{1}{8\pi^2}\int \mathrm{tr}\, F \wedge F = (\|F^+\|^2 - \|F^-\|^2)/8\pi^2$ and $S = \|F^+\|^2 + \|F^-\|^2$, so $S \geq 8\pi^2|k|$ with equality exactly for SD ($k > 0$) or ASD ($k < 0$) fields.
- Physically: $\vec{E}^a = \pm\vec{B}^a$, zero stress tensor, tunnelling amplitude $e^{-8\pi^2|k|/g^2}$.

8.5 The BPST instanton and the 't Hooft ansatz

The Bogomolny bound tells us where to look: in the sector with instanton number k, the minima of the action are the self-dual connections. But does the sector $k = 1$ contain any self-dual connection at all? Nash takes the answer for granted; he goes straight to the moduli

space $\mathcal{M}_k$ and counts its dimension. Before we follow him, we want an instanton in our hands, as a formula we can differentiate. This section constructs it.

The solution was found by Belavin, Polyakov, Schwarz and Tyupkin in 1975 and is called the **BPST instanton**. The cleanest way to write it uses quaternions, because $\mathbb{R}^4 = \mathbb{H}$ and $\mathrm{SU}(2) = \{\text{unit quaternions}\}$: the spacetime and the gauge group are made of the same stuff. We will

1. set up quaternions as 2×2 matrices, and see the 't Hooft symbols of Section 8.4 fall out of quaternion multiplication;
2. understand why a finite-action field must be "pure gauge at infinity", and how the winding of that gauge transformation is k;
3. write down the BPST connection, compute its curvature line by line, check that it is self-dual, and compute its action $8\pi^2$ (that is, $8\pi^2/g^2$ in physical units) and its charge $k = 1$;
4. move it to the *singular gauge*, and generalise it to k instantons with the *'t Hooft ansatz*.

At the end we count parameters. The BPST instanton has 5 of them (a centre in $\mathbb{R}^4$ and a size), and this is the number $8k-3$ at $k=1$ that Nash computes with the index theorem in Section 8.9.

8.5.1 Quaternions as 2×2 matrices

A quaternion is a number $q = q^4 + q^1\mathbf{i} + q^2\mathbf{j} + q^3\mathbf{k}$ with real q^μ and the rules $\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = \mathbf{ijk} = -1$. We never need the abstract definition, because the quaternions have a concrete matrix representation built from the Pauli matrices.

Working definition 8.24 (quaternions as matrices)

Let

$$e_1 = -i\sigma_1 = \begin{pmatrix} 0 & -i \\ -i & 0 \end{pmatrix}, \quad e_2 = -i\sigma_2 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad e_3 = -i\sigma_3 = \begin{pmatrix} -i & 0 \\ 0 & i \end{pmatrix}, \quad e_4 = \mathbb{1}.$$

A point $x = (x^1, x^2, x^3, x^4) \in \mathbb{R}^4$ is identified with the matrix

$$x = x^4\mathbb{1} + x^k e_k = \begin{pmatrix} x^4 - ix^3 & -x^2 - ix^1 \\ x^2 - ix^1 & x^4 + ix^3 \end{pmatrix}. \tag{8.192}$$

The **conjugate** is $\bar{x} = x^4\mathbb{1} - x^k e_k = x^\dagger$ (the Hermitian adjoint of the matrix [its complex-conjugate transpose]). The **real part** and **imaginary part** are

$$\operatorname{Re} x = \tfrac{1}{2}(x + \bar{x}) = x^4\mathbb{1}, \qquad \operatorname{Im} x = \tfrac{1}{2}(x - \bar{x}) = x^k e_k.$$

Test: a 2×2 complex matrix is a quaternion iff it has the form $\begin{pmatrix} \alpha & -\bar{\beta} \\ \beta & \bar{\alpha} \end{pmatrix}$. Imaginary quaternions are exactly the matrices of $\mathfrak{su}(2)$: anti-Hermitian and traceless.

Worked example 8.51: The quaternion rules, by matrix multiplication

We check $e_1 e_2 = e_3$:

$$e_1 e_2 \overset{(1)}{=} (-i)^2 \sigma_1\sigma_2$$
$$\overset{(2)}{=} -i\sigma_3$$

$$\overset{(3)}{=} e_3.$$

(1) Definition $e_k = -i\sigma_k$.
(2) Pauli rule $\sigma_1\sigma_2 = i\sigma_3$, and $(-i)^2 = -1$.
(3) Definition of e_3.

In the same way $e_2e_3 = e_1$, $e_3e_1 = e_2$, $e_k^2 = -\sigma_k^2 = -\mathbb{1}$, and $e_je_k = -e_ke_j$ for $j \neq k$. In one line:

$$e_je_k = -\delta_{jk}\mathbb{1} + \varepsilon_{jkl}e_l, \qquad [e_j, e_k] = 2\varepsilon_{jkl}e_l. \tag{8.193}$$

So e_1, e_2, e_3 are $\mathbf{i}, \mathbf{j}, \mathbf{k}$. The commutator form says that $\{e_a\}$ is a basis of $\mathfrak{su}(2)$ with structure constants 2ε.

Two facts make the matrix picture powerful.

Length is a determinant. Multiply out:

$$\begin{aligned}\bar{x}x &\overset{(1)}{=} (x^4 - x^je_j)(x^4 + x^ke_k)\\ &\overset{(2)}{=} (x^4)^2\mathbb{1} - x^jx^ke_je_k + x^4x^k(e_k - e_k)\\ &\overset{(3)}{=} \big((x^4)^2 + x^jx^k\delta_{jk}\big)\mathbb{1} - x^jx^k\varepsilon_{jkl}e_l\\ &\overset{(4)}{=} |x|^2\,\mathbb{1}.\end{aligned}$$

(1) Definitions of x and $\bar{x}$.
(2) Expand the product; the two cross terms $x^4x^ke_k$ and $-x^jx^4e_j$ cancel.
(3) The product rule (8.193).
(4) x^jx^k is symmetric and ε_{jkl} antisymmetric in jk, so the last term vanishes.

Likewise $x\bar{x} = |x|^2\mathbb{1}$, and $\det x = |x|^2$ (check it on (8.192): $(x^4)^2 + (x^3)^2 + (x^1)^2 + (x^2)^2$). So $x^{-1} = \bar{x}/|x|^2$ and $|xy| = |x||y|$.

Unit quaternions are SU(2). If $|x| = 1$ then $x^\dagger x = \bar{x}x = \mathbb{1}$ and $\det x = 1$, so $x \in \mathrm{SU}(2)$. Conversely every SU(2) matrix has the form $\begin{pmatrix}\alpha & -\bar{\beta}\\ \beta & \bar{\alpha}\end{pmatrix}$ with $|\alpha|^2 + |\beta|^2 = 1$, which is a unit quaternion. Hence

$$\mathrm{SU}(2) = \{x \in \mathbb{H} : |x| = 1\} = S^3. \tag{8.194}$$

This is the coincidence at the root of instantons: the sphere at infinity of $\mathbb{R}^4$ is a copy of the gauge group SU(2).

The 't Hooft symbols come from quaternion multiplication

Define the matrix-valued 1-forms $\mathrm{d}x = e_\mu \mathrm{d}x^\mu$ and $\mathrm{d}\bar{x} = \bar{e}_\mu \mathrm{d}x^\mu$ (with $\bar{e}_k = -e_k$, $\bar{e}_4 = \mathbb{1}$). Products of two of them produce the self-dual and anti-self-dual 2-forms.

Worked example 8.52: $\mathrm{d}\bar{x} \wedge \mathrm{d}x$ is self-dual, $\mathrm{d}x \wedge \mathrm{d}\bar{x}$ is anti-self-dual

Since $\mathrm{d}x^\mu \wedge \mathrm{d}x^\nu$ is antisymmetric,

$$\begin{aligned}\mathrm{d}\bar{x} \wedge \mathrm{d}x &\overset{(1)}{=} \bar{e}_\mu e_\nu\, \mathrm{d}x^\mu \wedge \mathrm{d}x^\nu\\ &\overset{(2)}{=} \sum_{\mu<\nu} (\bar{e}_\mu e_\nu - \bar{e}_\nu e_\mu)\, \mathrm{d}x^{\mu\nu}\end{aligned}$$

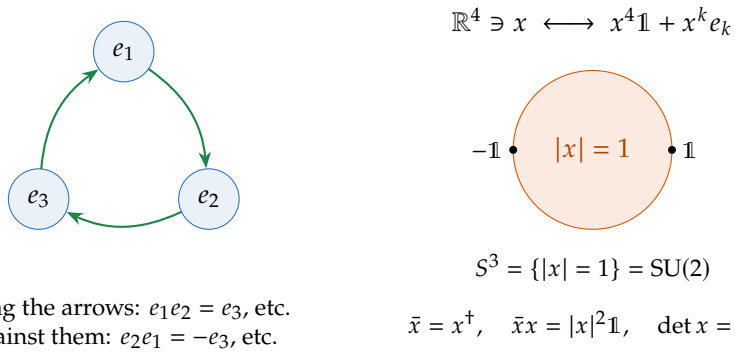


Figure 8.34: Left: the quaternion units multiply cyclically, like the cross product of $\hat{x}, \hat{y}, \hat{z}$; going against the arrow costs a sign. Right: identifying $\mathbb{R}^4$ with 2×2 matrices, the unit sphere S^3 becomes the group SU(2). Every instanton formula below is a statement about multiplying such matrices.

$$\overset{(3)}{=} \sum_k (-e_k - e_k)\,\mathrm{d}x^{k4} + \sum_{j<k} (-e_j e_k + e_k e_j)\,\mathrm{d}x^{jk}$$
$$\overset{(4)}{=} -2\sum_l e_l \left(\mathrm{d}x^{l4} + \tfrac{1}{2}\varepsilon_{ljk}\mathrm{d}x^{jk}\right)$$
$$\overset{(5)}{=} -2\, e_a\, \omega_a^+.$$

(1) Definitions of $\mathrm{d}\bar{x}$ and $\mathrm{d}x$.
(2) Pair the terms (μ,ν) and (ν,μ); $\mathrm{d}x^{\nu\mu} = -\mathrm{d}x^{\mu\nu}$.
(3) For $(\mu,\nu) = (k,4)$: $\bar{e}_k e_4 - \bar{e}_4 e_k = -e_k - e_k$. For spatial (j,k): $\bar{e}_j = -e_j$, $\bar{e}_k = -e_k$.
(4) $-e_j e_k + e_k e_j = -[e_j, e_k] = -2\varepsilon_{jkl}e_l$ by (8.193); then collect the coefficient of each e_l (the sum over $j<k$ equals half the sum over all j,k).
(5) The bracket is ω_l^+ of (8.181): for $l=1$ it is $\mathrm{d}x^{14} + \mathrm{d}x^{23}$.

Since each ω_a^+ is self-dual, so is $\mathrm{d}\bar{x}\wedge\mathrm{d}x$. The same steps with the order reversed give

$$\mathrm{d}\bar{x}\wedge\mathrm{d}x = -2\,e_a\,\omega_a^+ \ \text{(SD)}, \qquad \mathrm{d}x\wedge\mathrm{d}\bar{x} = -2\,e_a\,\omega_a^- \ \text{(ASD)}. \tag{8.195}$$

For the second, the $(k,4)$ coefficient is $e_k\bar{e}_4 - e_4\bar{e}_k = +2e_k$ and the spatial one is again $-2\varepsilon_{jkl}e_l$, and $2e_l\mathrm{d}x^{l4} - 2e_l\frac{1}{2}\varepsilon_{ljk}\mathrm{d}x^{jk} = -2e_l\omega_l^-$.

Reading off components gives a formula for the 't Hooft symbols that we use repeatedly:

$$\mathrm{Im}(\bar{e}_\nu e_\mu) = \eta^a_{\mu\nu}\,e_a, \qquad \mathrm{Im}(e_\mu\bar{e}_\nu) = -\bar{\eta}^a_{\mu\nu}\,e_a. \tag{8.196}$$

Check two entries of the first: $\mathrm{Im}(\bar{e}_4 e_1) = e_1$ and $\eta^1_{14} = 1$; $\mathrm{Im}(\bar{e}_2 e_1) = \mathrm{Im}(-e_2 e_1) = e_3$ and $\eta^3_{12} = \varepsilon_{312} = 1$. Of the second: $\mathrm{Im}(e_1\bar{e}_4) = e_1$ and $-\bar{\eta}^1_{14} = 1$; $\mathrm{Im}(e_1\bar{e}_2) = -e_1e_2 = -e_3$ and $-\bar{\eta}^3_{12} = -1$. The product identities (8.183) are now just associativity of quaternion multiplication, which is why they looked like "η's multiply like quaternions".

Physical picture: why SD 2-forms rotate under only half of SO(4)

Every rotation of $\mathbb{R}^4$ has the form $x \mapsto pxq^{-1}$ with unit quaternions p, q (this is $\mathrm{SO}(4) = (\mathrm{SU}(2)\times\mathrm{SU}(2))/\mathbb{Z}_2$; the pair $(-p,-q)$ gives the same rotation). Then $\mathrm{d}x \mapsto p\,\mathrm{d}x\,q^{-1}$ and $\mathrm{d}\bar{x} \mapsto q\,\mathrm{d}\bar{x}\,p^{-1}$, so

$$\mathrm{d}\bar{x}\wedge\mathrm{d}x \ \mapsto\ q\,(\mathrm{d}\bar{x}\wedge\mathrm{d}x)\,q^{-1}, \qquad \mathrm{d}x\wedge\mathrm{d}\bar{x} \ \mapsto\ p\,(\mathrm{d}x\wedge\mathrm{d}\bar{x})\,p^{-1}.$$

The SD forms feel only q and the ASD forms feel only p: this is the splitting $\mathrm{SU}(2)_+ \times \mathrm{SU}(2)_-$ drawn in Figure 8.31, made explicit.

8.5.2 Pure gauge at infinity and the winding number

We want a connection on $\mathbb{R}^4$ with $k = 1$ and finite action. Finite action constrains its behaviour at large $|x| = r$.

Finite action forces $F \to 0$. In polar coordinates $\mathrm{d}^4x = r^3\,\mathrm{d}r\,\mathrm{d}\Omega_3$, so $\int |F|^2 \mathrm{d}^4x < \infty$ needs $|F|^2$ to decay faster than r^{-4}, that is $|F| = o(r^{-2})$. So at large r the connection is approximately flat. A flat connection is locally pure gauge, $A = g^{-1}\mathrm{d}g$. Therefore

$$A \longrightarrow g^{-1}\mathrm{d}g \quad \text{as } r \to \infty, \qquad g : S^3_\infty \to \mathrm{SU}(2), \tag{8.197}$$

where S^3_∞ is a large sphere $|x| = R$. It is a map from S^3 to $\mathrm{SU}(2) = S^3$, and such maps have a **degree** (winding number): how many times the image wraps around S^3, counted with orientation.

Key idea: The instanton number is a winding number at infinity

For a finite-action $\mathrm{SU}(2)$ connection on $\mathbb{R}^4$ with $A \to g^{-1}\mathrm{d}g$,

$$k = \frac{1}{24\pi^2}\int_{S^3_\infty} \mathrm{tr}\big((g^{-1}\mathrm{d}g)^3\big) = \deg g. \tag{8.198}$$

Here is where it comes from. Part 8a showed that $\mathrm{tr}(F \wedge F) = \mathrm{d}\,\mathrm{CS}(A)$ with $\mathrm{CS}(A) = \mathrm{tr}(A \wedge \mathrm{d}A + \frac{2}{3}A \wedge A \wedge A)$. Then

$$\begin{aligned} k &\overset{(1)}{=} -\frac{1}{8\pi^2}\int_{\mathbb{R}^4} \mathrm{tr}(F \wedge F) \\ &\overset{(2)}{=} -\frac{1}{8\pi^2}\lim_{R\to\infty}\int_{S^3_R} \mathrm{CS}(A) \\ &\overset{(3)}{=} -\frac{1}{8\pi^2}\int_{S^3_\infty} \mathrm{tr}\big(-A^3 + \tfrac{2}{3}A^3\big) \\ &\overset{(4)}{=} \frac{1}{24\pi^2}\int_{S^3_\infty} \mathrm{tr}\big((g^{-1}\mathrm{d}g)^3\big). \end{aligned}$$

(1) Definition (8.187).
(2) $\mathrm{tr}\,F \wedge F = \mathrm{d}\,\mathrm{CS}$ and Stokes on the ball $|x| \le R$ (Chapter T2); the sphere carries the outward orientation.
(3) At infinity $F = \mathrm{d}A + A^2 \to 0$, so $\mathrm{d}A \to -A^2$. We write $A^3 = A \wedge A \wedge A$.
(4) $-1 + \frac{2}{3} = -\frac{1}{3}$, and $-\frac{1}{8\pi^2}\cdot(-\frac{1}{3}) = \frac{1}{24\pi^2}$; insert $A = g^{-1}\mathrm{d}g$.

The normalisation $\frac{1}{24\pi^2}$ is such that the result is an integer, the degree of g. We check it on the most important map.

Worked example 8.53: The identity map $g(x) = x/|x|$ has winding number one

Take $g = x/r$, which sends the sphere $|x| = r$ to $\mathrm{SU}(2)$ by radial projection: it is the identity map $S^3 \to S^3$, so its degree should be 1. First a formula for $g^{-1}\mathrm{d}g$ that we will need again:

$$g^{-1}\mathrm{d}g \overset{(1)}{=} \frac{\bar{x}}{r}\Big(\frac{\mathrm{d}x}{r} - \frac{x\,\mathrm{d}r}{r^2}\Big)$$

$$\begin{aligned} &\overset{(2)}{=} \frac{\bar{x}\,\mathrm{d}x}{r^2} - \frac{\mathrm{d}r}{r}\,\mathbb{1} \\ &\overset{(3)}{=} \frac{\mathrm{Im}(\bar{x}\,\mathrm{d}x)}{r^2}. \end{aligned} \tag{8.199}$$

(1) $g^{-1} = \bar{g} = \bar{x}/r$ for a unit quaternion; product rule for $\mathrm{d}(x/r)$.
(2) $\bar{x}x = r^2\mathbb{1}$.
(3) $\mathrm{Re}(\bar{x}\,\mathrm{d}x) = \frac{1}{2}(\bar{x}\,\mathrm{d}x + \mathrm{d}\bar{x}\,x) = \frac{1}{2}\mathrm{d}(\bar{x}x) = \frac{1}{2}\mathrm{d}(r^2) = r\,\mathrm{d}r$, so subtracting $\mathrm{d}r/r$ removes exactly the real part.

Now evaluate the 3-form $\mathrm{tr}(g^{-1}\mathrm{d}g)^3$ at the point $x = (0,0,0,R)$, where $g = \mathbb{1}$ and the sphere is tangent to the $x^1x^2x^3$ directions:

$$\begin{aligned} \mathrm{tr}\big((g^{-1}\mathrm{d}g)^3\big)\big|_{x=(0,0,0,R)} &\overset{(1)}{=} R^{-3}\,\mathrm{tr}(e_ie_je_k)\,\mathrm{d}x^i\wedge\mathrm{d}x^j\wedge\mathrm{d}x^k \\ &\overset{(2)}{=} R^{-3}\,6\;\mathrm{tr}(e_1e_2e_3)\,\mathrm{d}x^{123} \\ &\overset{(3)}{=} -12\,R^{-3}\,\mathrm{d}x^{123} \\ &\overset{(4)}{=} +12\,\mathrm{vol}_{S^3_R}. \end{aligned}$$

(1) At this point $g^{-1} = \mathbb{1}$ and, along the sphere, $\mathrm{d}g = \mathrm{d}x/R = e_k\mathrm{d}x^k/R$ ($\mathrm{d}x^4$ vanishes on the tangent space).
(2) The six orderings of $(1,2,3)$ each give sign $\times\, e_{\pi(1)}e_{\pi(2)}e_{\pi(3)} = e_1e_2e_3$, because distinct e's anticommute.
(3) $e_1e_2e_3 = e_3e_3 = -\mathbb{1}$, and $\mathrm{tr}\,\mathbb{1} = 2$.
(4) The outward normal at this point is ∂_4, and the boundary orientation is $\iota_{\partial_4}(\mathrm{d}x^{1234}) = -\mathrm{d}x^{123}$. The tangent space is spanned by the orthonormal $\partial_1,\partial_2,\partial_3$, so the positively oriented Riemannian volume form of S^3_R at this point is $\mathrm{vol}_{S^3_R} = -\mathrm{d}x^{123}$.

The form $\mathrm{tr}(g^{-1}\mathrm{d}g)^3$ is invariant under $g \mapsto hg$ for constant $h \in \mathrm{SU}(2)$ (the h's cancel in $g^{-1}\mathrm{d}g$), and left multiplication moves any point of S^3 to any other. So the value $12R^{-3}\mathrm{vol}$ holds everywhere, and

$$\frac{1}{24\pi^2}\int_{S^3_R} 12R^{-3}\,\mathrm{vol}_{S^3_R} = \frac{12\cdot 2\pi^2R^3}{24\pi^2R^3} = 1,$$

using $\mathrm{Vol}(S^3_R) = 2\pi^2R^3$. The identity map has degree one, as it must.

Try it: The winding of g^n

Show that $g(x) = (x/r)^{-1} = \bar{x}/r$ has winding number -1. *Answer:* $g^{-1}\mathrm{d}g = x\,\mathrm{d}\bar{x}/r^2 - \mathrm{d}r/r$, and at $(0,0,0,R)$ it equals $-e_k\mathrm{d}x^k/R$; the cube changes sign. More generally $(x/r)^n$ has winding n: the winding number is additive under pointwise multiplication of maps.

8.5.3 The BPST solution

Now we can guess the instanton. We want A to be smooth at the origin and to approach $g^{-1}\mathrm{d}g$ with $g = x/r$ at infinity. The simplest interpolation multiplies the pure gauge by a function that rises from 0 to 1:

$$A = f(r)\,g^{-1}\mathrm{d}g, \qquad f(0) = 0,\ f(\infty) = 1.$$

Imposing self-duality gives an ordinary differential equation for f, whose solution is $f = r^2/(r^2+\lambda^2)$ for an arbitrary constant $\lambda > 0$. Rather than solve the ODE, we take the answer and *check* it; the check is the useful calculation.

Working definition 8.25 (the BPST instanton)

With $\phi(x) = |x|^2 + \lambda^2$,

$$A = \frac{|x|^2}{|x|^2+\lambda^2}\, g^{-1}\mathrm{d}g = \frac{\operatorname{Im}(\bar{x}\,\mathrm{d}x)}{|x|^2+\lambda^2} = \frac{\eta^a_{\mu\nu}x^\nu}{\phi}\, e_a\,\mathrm{d}x^\mu. \tag{8.200}$$

The second form follows from (8.199) [$g^{-1}\mathrm{d}g = \operatorname{Im}(\bar{x}\,\mathrm{d}x)/|x|^2$ for $g = x/|x|$, the pure gauge that A approaches at infinity]; the third from $\operatorname{Im}(\bar{x}e_\mu) = x^\nu \operatorname{Im}(\bar{e}_\nu e_\mu) = \eta^a_{\mu\nu}x^\nu e_a$ by (8.196) [$\eta^a_{\mu\nu}$ are the 't Hooft symbols, the components of the self-dual basis 2-forms ω^+_a]. The constant λ is the **size** of the instanton, and the origin is its **centre**. Translating, $x \mapsto x - y$, gives an instanton centred at $y \in \mathbb{R}^4$.

In the physicists' language of the dictionary in Section 8.3.1, $A = \frac{g}{2}A^a_\mu e_a \mathrm{d}x^\mu$, so

$$gA^a_\mu = \frac{2\,\eta^a_{\mu\nu}x^\nu}{x^2+\lambda^2},$$

which is the formula printed in every textbook on instantons. As an explicit matrix, the first component is

$$A_1 = \frac{x^\nu \eta^a_{1\nu} e_a}{\phi} = \frac{x^4 e_1 + x^2 e_3 - x^3 e_2}{\phi} = \frac{1}{\phi}\begin{pmatrix} -ix^2 & x^3 - ix^4 \\ -x^3 - ix^4 & ix^2 \end{pmatrix},$$

using $\eta^1_{14} = 1$, $\eta^3_{12} = 1$, $\eta^2_{13} = \varepsilon_{213} = -1$. It is anti-Hermitian and traceless, as a $\mathfrak{su}(2)$ connection must be.

The curvature, line by line

We compute $F = \mathrm{d}A + A \wedge A$ with $A = \phi^{-1}\operatorname{Im}(\bar{x}\,\mathrm{d}x)$. Three small facts first. Since $\phi = \bar{x}x + \lambda^2$,

$$\mathrm{d}\phi = \bar{x}\,\mathrm{d}x + \mathrm{d}\bar{x}\,x, \qquad \operatorname{Im}(\bar{x}\,\mathrm{d}x) = \bar{x}\,\mathrm{d}x - \tfrac{1}{2}\mathrm{d}\phi, \qquad \mathrm{d}\operatorname{Im}(\bar{x}\,\mathrm{d}x) = \mathrm{d}\bar{x}\wedge\mathrm{d}x. \tag{8.201}$$

The first is the product rule. The second says $\operatorname{Re}(\bar{x}\,\mathrm{d}x) = \frac{1}{2}(\bar{x}\,\mathrm{d}x + \mathrm{d}\bar{x}\,x) = \frac{1}{2}\mathrm{d}\phi$. The third: $\mathrm{d}(\bar{x}\,\mathrm{d}x) = \mathrm{d}\bar{x}\wedge\mathrm{d}x$ and $\mathrm{d}(\mathrm{d}\bar{x}\,x) = -\mathrm{d}\bar{x}\wedge\mathrm{d}x$ (moving d past the 1-form $\mathrm{d}\bar{x}$ costs a sign), so $\mathrm{d}\,\frac{1}{2}(\bar{x}\,\mathrm{d}x - \mathrm{d}\bar{x}\,x) = \mathrm{d}\bar{x}\wedge\mathrm{d}x$. Note that $\mathrm{d}\phi$ is an ordinary (scalar) 1-form, so it commutes with every matrix and anticommutes with every matrix-valued 1-form.

Step 1: the quadratic term.

$$\begin{aligned}\phi^2 A\wedge A &\overset{(1)}{=} \left(\bar{x}\,\mathrm{d}x - \tfrac{1}{2}\mathrm{d}\phi\right)\wedge\left(\bar{x}\,\mathrm{d}x - \tfrac{1}{2}\mathrm{d}\phi\right)\\ &\overset{(2)}{=} \bar{x}\,\mathrm{d}x\wedge\bar{x}\,\mathrm{d}x - \tfrac{1}{2}\,\bar{x}\,\mathrm{d}x\wedge\mathrm{d}\phi - \tfrac{1}{2}\,\mathrm{d}\phi\wedge\bar{x}\,\mathrm{d}x + \tfrac{1}{4}\,\mathrm{d}\phi\wedge\mathrm{d}\phi\\ &\overset{(3)}{=} \bar{x}\,\mathrm{d}x\wedge\bar{x}\,\mathrm{d}x.\end{aligned}$$

(1) Second fact of (8.201).
(2) Expand.
(3) $\mathrm{d}\phi\wedge\bar{x}\,\mathrm{d}x = -\bar{x}\,\mathrm{d}x\wedge\mathrm{d}\phi$ (scalar 1-form past a matrix 1-form), so the middle terms cancel; and $\mathrm{d}\phi\wedge\mathrm{d}\phi = 0$.

Step 2: the derivative term.

$$\begin{aligned} \mathrm{d}A &\overset{(1)}{=} \mathrm{d}(\phi^{-1}) \wedge \mathrm{Im}(\bar{x}\,\mathrm{d}x) + \phi^{-1}\,\mathrm{d}\ \mathrm{Im}(\bar{x}\,\mathrm{d}x) \\ &\overset{(2)}{=} -\phi^{-2}\,\mathrm{d}\phi \wedge \left(\bar{x}\,\mathrm{d}x - \tfrac{1}{2}\mathrm{d}\phi\right) + \phi^{-1}\,\mathrm{d}\bar{x} \wedge \mathrm{d}x \\ &\overset{(3)}{=} -\phi^{-2}\,(\bar{x}\,\mathrm{d}x + \mathrm{d}\bar{x}\,x) \wedge \bar{x}\,\mathrm{d}x + \phi^{-1}\,\mathrm{d}\bar{x} \wedge \mathrm{d}x \\ &\overset{(4)}{=} -\phi^{-2}\big(\bar{x}\,\mathrm{d}x \wedge \bar{x}\,\mathrm{d}x + |x|^2\,\mathrm{d}\bar{x} \wedge \mathrm{d}x\big) + \phi^{-1}\,\mathrm{d}\bar{x} \wedge \mathrm{d}x. \end{aligned}$$

(1) Leibniz rule for a function times a 1-form.
(2) $\mathrm{d}(\phi^{-1}) = -\phi^{-2}\mathrm{d}\phi$, and the second and third facts of (8.201).
(3) $\mathrm{d}\phi \wedge \mathrm{d}\phi = 0$, then the first fact.
(4) $\mathrm{d}\bar{x}\,x \wedge \bar{x}\,\mathrm{d}x = \mathrm{d}\bar{x}\,(x\bar{x}) \wedge \mathrm{d}x$ and $x\bar{x} = |x|^2\mathbb{1}$ is a scalar function.

Step 3: add.

$$\begin{aligned} F = \mathrm{d}A + A \wedge A &\overset{(1)}{=} \frac{-\bar{x}\,\mathrm{d}x \wedge \bar{x}\,\mathrm{d}x - |x|^2\,\mathrm{d}\bar{x} \wedge \mathrm{d}x + \phi\,\mathrm{d}\bar{x} \wedge \mathrm{d}x + \bar{x}\,\mathrm{d}x \wedge \bar{x}\,\mathrm{d}x}{\phi^2} \\ &\overset{(2)}{=} \frac{(\phi - |x|^2)\,\mathrm{d}\bar{x} \wedge \mathrm{d}x}{\phi^2} \\ &\overset{(3)}{=} \frac{\lambda^2\,\mathrm{d}\bar{x} \wedge \mathrm{d}x}{(|x|^2+\lambda^2)^2}. \end{aligned} \tag{8.202}$$

(1) Steps 1 and 2, over the common denominator ϕ^2.
(2) The quartic terms $\bar{x}\,\mathrm{d}x \wedge \bar{x}\,\mathrm{d}x$ cancel.
(3) $\phi - |x|^2 = \lambda^2$.

Everything non-linear has cancelled, and what is left is a scalar function times the fixed 2-form $\mathrm{d}\bar{x} \wedge \mathrm{d}x$. By (8.195) that 2-form is self-dual. Hence

Key idea: BPST is self-dual

$$F = \frac{\lambda^2\,\mathrm{d}\bar{x} \wedge \mathrm{d}x}{(|x|^2+\lambda^2)^2} = -\frac{2\lambda^2}{(|x|^2+\lambda^2)^2}\,e_a\,\omega_a^+, \qquad \star F = F.$$

In components $F_{\mu\nu} = -2\lambda^2\eta^a_{\mu\nu}e_a/\phi^2$; in physicists' normalisation $gF^a_{\mu\nu} = -4\lambda^2\eta^a_{\mu\nu}/(x^2+\lambda^2)^2$.

As matrices, for instance,

$$F_{12} = F_{34} = -\frac{2\lambda^2}{\phi^2}e_3 = \frac{2\lambda^2}{\phi^2}\begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}, \qquad F_{14} = F_{23} = -\frac{2\lambda^2}{\phi^2}e_1 = \frac{2\lambda^2}{\phi^2}\begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix},$$

and one sees the self-duality conditions (8.186) directly. The field strength is largest at the centre, where $|F_{\mu\nu}| \sim 2/\lambda^2$, and it decays like λ^2/r^4, safely faster than the r^{-2} required for finite action. By (8.198) and Example 8.53, $k = 1$, since $A \to g^{-1}\mathrm{d}g$ with $g = x/r$ (the prefactor $r^2/\phi \to 1$). We now confirm this, and find the action, by direct integration.

The action and the charge

Worked example 8.54: $S = 8\pi^2$ and $k = 1$ by direct integration

The action density, from (8.166):

$$\begin{aligned} -\tfrac{1}{2}\operatorname{tr}(F_{\mu\nu}F_{\mu\nu}) &\overset{(1)}{=} -\tfrac{1}{2}\cdot\frac{4\lambda^4}{\phi^4}\,\eta^a_{\mu\nu}\eta^b_{\mu\nu}\operatorname{tr}(e_a e_b) \\ &\overset{(2)}{=} -\tfrac{1}{2}\cdot\frac{4\lambda^4}{\phi^4}\cdot 4\delta^{ab}\cdot(-2\delta_{ab}) \\ &\overset{(3)}{=} \frac{48\lambda^4}{(|x|^2+\lambda^2)^4}. \end{aligned}$$

(1) Substitute $F_{\mu\nu} = -2\lambda^2\eta^a_{\mu\nu}e_a/\phi^2$.
(2) $\eta^a_{\mu\nu}\eta^b_{\mu\nu} = 4\delta^{ab}$ (8.183), $\operatorname{tr}(e_a e_b) = -2\delta_{ab}$ (Example 8.43).
(3) $\delta^{ab}\delta_{ab} = 3$, and $\frac{1}{2}\cdot 4\cdot 4\cdot 2\cdot 3 = 48$.

It depends only on $r = |x|$, so we integrate in polar coordinates:

$$\begin{aligned} S &\overset{(1)}{=} \int_0^\infty \frac{48\lambda^4}{(r^2+\lambda^2)^4}\,2\pi^2 r^3\,\mathrm{d}r \\ &\overset{(2)}{=} 48\pi^2\lambda^4\int_0^\infty \frac{u\,\mathrm{d}u}{(u+\lambda^2)^4} \\ &\overset{(3)}{=} 48\pi^2\lambda^4\cdot\frac{1}{6\lambda^4} \\ &= 8\pi^2. \end{aligned}$$

(1) $\mathrm{d}^4x = r^3\mathrm{d}r\,\mathrm{d}\Omega_3$ and $\int\mathrm{d}\Omega_3 = \operatorname{Vol}(S^3) = 2\pi^2$.
(2) Substitute $u = r^2$, $\mathrm{d}u = 2r\,\mathrm{d}r$, so $r^3\mathrm{d}r = \frac{1}{2}u\,\mathrm{d}u$.
(3) Write $u = (u+a) - a$ with $a = \lambda^2$: $\int_0^\infty\left[(u+a)^{-3} - a(u+a)^{-4}\right]\mathrm{d}u = \frac{1}{2a^2} - \frac{a}{3a^3} = \frac{1}{6a^2}$.

Since F is self-dual, $\operatorname{tr}(F\wedge F) = \operatorname{tr}(F\wedge\star F)$, so $-\int\operatorname{tr}(F\wedge F) = S = 8\pi^2$ and

$$k = -\frac{1}{8\pi^2}\int\operatorname{tr}(F\wedge F) = \frac{8\pi^2}{8\pi^2} = 1.$$

The Bogomolny bound $S \geq 8\pi^2|k|$ is saturated. Restoring the coupling (Example 8.43), $S_{\text{phys}} = 8\pi^2/g^2$: the tunnelling amplitude is $e^{-8\pi^2/g^2}$. Notice that λ has dropped out. This is the conformal (in particular scale) invariance of Section 8.4.1: a big instanton and a small one cost the same.

Worked example 8.55: How concentrated is an instanton?

Redo the radial integral only up to $r = R$. With $a = \lambda^2$ and $U = R^2$, the antiderivative of $u(u+a)^{-4}$ is $-\frac{1}{2}(u+a)^{-2} + \frac{a}{3}(u+a)^{-3}$, so

$$\begin{aligned} \frac{S(R)}{8\pi^2} &\overset{(1)}{=} 6a^2\left[-\frac{1}{2(u+a)^2} + \frac{a}{3(u+a)^3}\right]_0^U \\ &\overset{(2)}{=} 6a^2\left[\frac{1}{6a^2} - \frac{1}{2(U+a)^2} + \frac{a}{3(U+a)^3}\right] \end{aligned}$$

$$\overset{(3)}{=} 1 - \frac{3\lambda^4}{(R^2+\lambda^2)^2} + \frac{2\lambda^6}{(R^2+\lambda^2)^3}.$$

(1) The previous example, stopped at $u = U$, divided by $8\pi^2$: $48\pi^2 a^2/8\pi^2 = 6a^2$.
(2) At $u = 0$ the bracket is $-\frac{1}{2a^2} + \frac{1}{3a^2} = -\frac{1}{6a^2}$.
(3) Multiply out and substitute back.

At $R = \lambda$: $1 - \frac{3}{4} + \frac{2}{8} = \frac{1}{2}$. *Half the action lies inside the ball of radius* λ. At $R = 3\lambda$ the fraction is $1 - 0.03 + 0.002 \approx 0.97$. So λ really is the size of a lump of action (Figure 8.35).

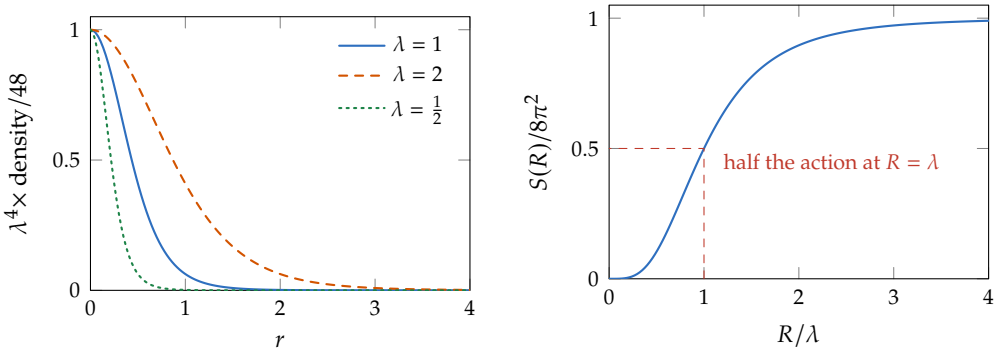


Figure 8.35: Left: the BPST action density $48\lambda^4/(r^2+\lambda^2)^4$, rescaled by $\lambda^4/48$ so that the peak height is 1 for every λ; what changes with λ is only the width. (Without the rescaling a small instanton is a tall narrow spike and a large one a low wide hill, always with the same area $8\pi^2$.) Right: the fraction of the action inside radius R, a universal function of R/λ; half of it sits inside $R = \lambda$.

The anti-instanton and the parameters

Swapping the roles of x and $\bar{x}$ gives the **anti-instanton**

$$\bar{A} = \frac{\operatorname{Im}(x\,\mathrm{d}\bar{x})}{|x|^2+\lambda^2}, \qquad \bar{F} = \frac{\lambda^2\,\mathrm{d}x\wedge\mathrm{d}\bar{x}}{(|x|^2+\lambda^2)^2} = -\frac{2\lambda^2}{\phi^2}\,e_a\,\omega_a^-.$$

The same three steps prove the formula for $\bar{F}$, and (8.195) shows it is ASD. Its pure gauge at infinity is $g = \bar{x}/r$, of winding -1 (the Try-it box above), so $k = -1$ and $S = 8\pi^2$.

How many parameters does the BPST family have? The formula (8.200) was written with centre 0 and a particular orientation in the gauge group. We can

- translate: $x \mapsto x - y$, $y \in \mathbb{R}^4$ (4 parameters);
- rescale: $\lambda > 0$ (1 parameter);
- rotate in colour space by a constant $h \in \mathrm{SU}(2)$: $A \mapsto h^{-1}Ah$ (3 parameters, but h and $-h$ act the same way on $\mathfrak{su}(2)$, so it is really SO(3)).

Rotating $\mathbb{R}^4$ gives nothing new: under $x \mapsto pxq^{-1}$ the BPST connection changes by a constant gauge rotation (by q, from the intuition box above), so spacetime rotations are already among the colour rotations. The constant colour rotations are gauge transformations, so they do not change the physical configuration; but they are the "framing" degrees of freedom that Nash keeps track of in Section 8.7. So:

$$\underbrace{4}_{\text{centre } y} + \underbrace{1}_{\text{size } \lambda} = 5 \text{ gauge-inequivalent parameters}, \qquad 5 + \underbrace{3}_{\text{colour}} = 8 \text{ framed parameters}.$$

The index theorem calculation of Section 8.9 will give $\dim \mathcal{M}_1 = 8k - 3 = 5$, and Section 8.10 shows that the BPST family is all of $\mathcal{M}_1$, so $\mathcal{M}_1 = \mathbb{R}^4 \times \mathbb{R}^+$.

Physical picture: the size modulus and the instanton gas

Because the action does not depend on λ, the classical theory has no preference for any instanton size. In the quantum theory, the one-loop fluctuation determinant around the instanton turns the tunnelling amplitude into an integral over the moduli, schematically $\int \mathrm{d}^4y\,\mathrm{d}\lambda\,\lambda^{-5}\,e^{-8\pi^2/g^2(1/\lambda)}$, where the coupling now runs with the scale $1/\lambda$ (the λ^{-5} is fixed by dimensional analysis: the measure $\mathrm{d}^4y\,\mathrm{d}\lambda$ has length dimension 5). In an asymptotically free theory $g(1/\lambda)$ grows with λ, the suppression $e^{-8\pi^2/g^2}$ weakens, and the integral diverges at large λ. This "infrared problem" of instanton physics is a direct consequence of the non-compact direction $\lambda\to\infty$ of $\mathcal{M}_1$; the other non-compact end $\lambda\to 0$ is the one Donaldson uses (Section 8.11).

8.5.4 The singular gauge

In the form (8.200) the instanton is smooth everywhere, but it falls off only like $1/r$ at infinity (it tends to the pure gauge $g^{-1}\mathrm{d}g$, which is of order $1/r$). This is the **regular gauge**. For superposing several instantons it is better to use a gauge in which A falls off fast at infinity; the price is a singularity at the centre. We apply the gauge transformation $h=g^{-1}=\bar{x}/r$, which is smooth except at $x=0$.

$$\begin{aligned} A_h = h^{-1}Ah + h^{-1}\mathrm{d}h &\overset{(1)}{=} gAg^{-1} + g\,\mathrm{d}(g^{-1}) \\ &\overset{(2)}{=} \frac{r^2}{\phi}\,g(g^{-1}\mathrm{d}g)g^{-1} - \mathrm{d}g\,g^{-1} \\ &\overset{(3)}{=} \Big(\frac{r^2}{\phi}-1\Big)\mathrm{d}g\,g^{-1} \\ &\overset{(4)}{=} -\frac{\lambda^2}{r^2+\lambda^2}\cdot\frac{\mathrm{Im}(\mathrm{d}x\,\bar{x})}{r^2}. \end{aligned} \tag{8.203}$$

(1) $h^{-1}=g$ and $h=g^{-1}$; substitute into (8.163).
(2) First form of (8.200); and $\mathrm{d}(gg^{-1})=0$ gives $g\,\mathrm{d}(g^{-1})=-\mathrm{d}g\,g^{-1}$.
(3) $g\,g^{-1}\mathrm{d}g\,g^{-1}=\mathrm{d}g\,g^{-1}$.
(4) $\frac{r^2}{\phi}-1=-\frac{\lambda^2}{\phi}$, and the same computation as (8.199) with the factors in the other order gives $\mathrm{d}g\,g^{-1}=\mathrm{Im}(\mathrm{d}x\,\bar{x})/r^2$.

In components, $\mathrm{Im}(e_\mu\bar{x})=x^\nu\,\mathrm{Im}(e_\mu\bar{e}_\nu)=-\bar{\eta}^a_{\mu\nu}x^\nu e_a$ by (8.196), so

$$A^{\mathrm{sing}}_\mu = \frac{\lambda^2\,\bar{\eta}^a_{\mu\nu}x^\nu}{r^2(r^2+\lambda^2)}\,e_a. \tag{8.204}$$

Three things to see:

- At large r, $A^{\mathrm{sing}}\sim\lambda^2/r^3$: it falls off fast, and the gauge transformation at infinity is now trivial.
- Near $r=0$, $A^{\mathrm{sing}}\approx -\mathrm{d}g\,g^{-1}=h^{-1}\mathrm{d}h$ with $h=\bar{x}/r$: pure gauge, but singular, with winding. The topology has been moved from the sphere at infinity to an infinitesimal sphere around the centre.
- The curvature is $F_h=h^{-1}Fh=\lambda^2\,x\,(\mathrm{d}\bar{x}\wedge\mathrm{d}x)\,\bar{x}/(r^2\phi^2)$. Conjugation acts on the matrix part only, so F_h is still self-dual, and $\mathrm{tr}\,F_h\wedge F_h=\mathrm{tr}\,F\wedge F$ is the same smooth density. The singularity is in A, not in anything gauge invariant: it is a **gauge singularity**.

Curiously, the self-dual instanton in singular gauge is written with the *anti*-self-dual symbols $\bar\eta$. There is no contradiction: the self-duality of F is a property of the 2-form, not of the symbols used for the 1-form A.

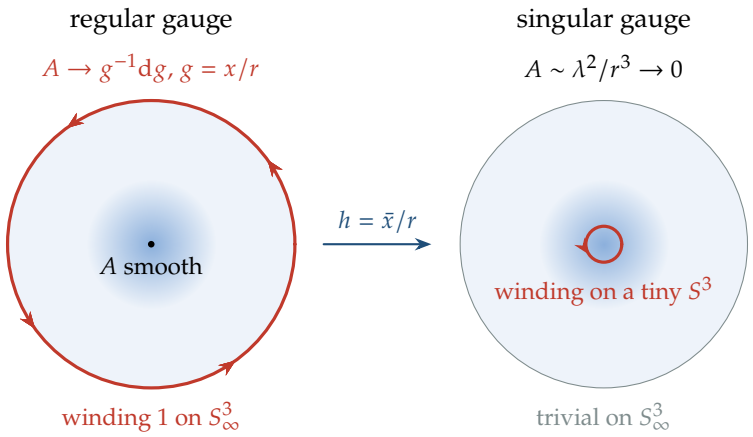


Figure 8.36: The same instanton in two gauges. The blue shading is the action density $48\lambda^4/(r^2+\lambda^2)^4$, identical in both, since it is gauge invariant. The red loop marks where the non-trivial winding of the gauge potential sits: on the sphere at infinity in the regular gauge, on an infinitesimal sphere around the centre in the singular gauge. The gauge transformation h that relates them is itself singular at the origin, which is why it can move the winding.

8.5.5 The 't Hooft ansatz: many instantons at once

Write $\sigma = \ln\rho$ for the function $\rho = 1 + \lambda^2/r^2$. Then $\partial_\nu\sigma = \partial_\nu\rho/\rho = -2\lambda^2 x_\nu/(r^2(r^2+\lambda^2))$, and (8.204) becomes

$$A_\mu = -\tfrac{1}{2}\,\bar\eta^a_{\mu\nu}\,\partial_\nu \ln\rho\; e_a. \tag{8.205}$$

't Hooft's idea was to take (8.205) as an **ansatz** for an arbitrary positive function ρ, and ask what ρ must satisfy for F to be self-dual. The answer is strikingly simple.

Key idea: The 't Hooft ansatz

For the connection (8.205),

$$F^- = \frac{\Box\rho}{4\rho}\; e_a\,\omega^-_a, \qquad \Box = \partial_\mu\partial_\mu. \tag{8.206}$$

So A is self-dual exactly where ρ is harmonic, $\Box\rho = 0$. The self-dual Yang–Mills equations, non-linear in A, have become the *linear* Laplace equation for ρ.

Worked example 8.56: Deriving (8.206)

Write $A_\mu = a^c_\mu e_c$ with $a^c_\mu = -\frac{1}{2}\bar\eta^c_{\mu\lambda}\sigma_\lambda$, where $\sigma_\lambda = \partial_\lambda\sigma$ and $\sigma_{\mu\lambda} = \partial_\mu\partial_\lambda\sigma$. By (8.193), $[A_\mu, A_\nu] = 2\varepsilon_{abc}a^a_\mu a^b_\nu e_c$, so the e_c-component of $F_{\mu\nu}$ is

$$\begin{aligned} F^c_{\mu\nu} &\overset{(1)}{=} \partial_\mu a^c_\nu - \partial_\nu a^c_\mu + 2\varepsilon_{abc}a^a_\mu a^b_\nu \\ &\overset{(2)}{=} -\tfrac{1}{2}\big(\bar\eta^c_{\nu\lambda}\sigma_{\mu\lambda} - \bar\eta^c_{\mu\lambda}\sigma_{\nu\lambda}\big) + \tfrac{1}{2}\varepsilon_{abc}\bar\eta^a_{\mu\lambda}\bar\eta^b_{\nu\kappa}\sigma_\lambda\sigma_\kappa. \end{aligned}$$

(1) $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu + [A_\mu, A_\nu]$.
(2) Substitute a^c_μ; $2\cdot(-\tfrac{1}{2})^2 = \tfrac{1}{2}$.

The ASD part of F is the part along the ω_d^- $= \frac{1}{2}\bar\eta^d_{\mu\nu}\mathrm{d}x^{\mu\nu}$. Since $\bar\eta^d_{\mu\nu}\bar\eta^{d'}_{\mu\nu} = 4\delta^{dd'}$ and $\bar\eta^d_{\mu\nu}\eta^{d'}_{\mu\nu} = 0$, the coefficient is $f^{cd} = \frac{1}{4}\bar\eta^d_{\mu\nu}F^c_{\mu\nu}$, i.e. $F^- = f^{cd}e_c\,\omega_d^-$. We need two identities, checked by multiplying out (they hold for η and $\bar\eta$ alike):

$$\text{(I)}\quad \bar\eta^a_{\mu\nu}\bar\eta^b_{\mu\rho} = \delta^{ab}\delta_{\nu\rho} + \varepsilon^{abc}\bar\eta^c_{\nu\rho},$$
$$\text{(II)}\quad \varepsilon_{abc}\bar\eta^a_{\mu\lambda}\bar\eta^b_{\nu\kappa} = \delta_{\mu\nu}\bar\eta^c_{\lambda\kappa} - \delta_{\mu\kappa}\bar\eta^c_{\lambda\nu} - \delta_{\lambda\nu}\bar\eta^c_{\mu\kappa} + \delta_{\lambda\kappa}\bar\eta^c_{\mu\nu}.$$

Derivative terms.

$$-\tfrac{1}{8}\bar\eta^d_{\mu\nu}\big(\bar\eta^c_{\nu\lambda}\sigma_{\mu\lambda} - \bar\eta^c_{\mu\lambda}\sigma_{\nu\lambda}\big) \overset{(1)}{=} -\tfrac{1}{8}\big(-(\delta^{dc}\delta_{\mu\lambda} + \varepsilon^{dce}\bar\eta^e_{\mu\lambda})\sigma_{\mu\lambda} - (\delta^{dc}\delta_{\nu\lambda} + \varepsilon^{dce}\bar\eta^e_{\nu\lambda})\sigma_{\nu\lambda}\big)$$
$$\overset{(2)}{=} \tfrac{1}{4}\,\delta^{cd}\,\Box\sigma.$$

(1) In the first term write $\bar\eta^d_{\mu\nu}\bar\eta^c_{\nu\lambda} = -\bar\eta^d_{\nu\mu}\bar\eta^c_{\nu\lambda}$ and use (I) with summed index ν; the second term is (I) directly with summed index μ.

(2) $\sigma_{\mu\lambda}$ is symmetric and $\bar\eta^e$ antisymmetric, so the ε terms drop; $\delta_{\mu\lambda}\sigma_{\mu\lambda} = \Box\sigma$.

Quadratic terms. First contract (II) with $\sigma_\lambda\sigma_\kappa$:

$$\tfrac{1}{2}\varepsilon_{abc}\bar\eta^a_{\mu\lambda}\bar\eta^b_{\nu\kappa}\sigma_\lambda\sigma_\kappa \overset{(1)}{=} \tfrac{1}{2}\big(0 - \bar\eta^c_{\lambda\nu}\sigma_\lambda\sigma_\mu - \bar\eta^c_{\mu\kappa}\sigma_\nu\sigma_\kappa + \bar\eta^c_{\mu\nu}\,\sigma_\lambda\sigma_\lambda\big),$$

(1) The four terms of (II): $\bar\eta^c_{\lambda\kappa}\sigma_\lambda\sigma_\kappa = 0$ by antisymmetry; the Kronecker deltas set $\kappa = \mu$, $\lambda = \nu$, $\lambda = \kappa$ respectively.

then project with $\frac{1}{4}\bar\eta^d_{\mu\nu}$:

$$\tfrac{1}{8}\bar\eta^d_{\mu\nu}\big(-\bar\eta^c_{\lambda\nu}\sigma_\lambda\sigma_\mu - \bar\eta^c_{\mu\kappa}\sigma_\nu\sigma_\kappa + \bar\eta^c_{\mu\nu}\sigma^2\big) \overset{(1)}{=} \tfrac{1}{8}\big(-\delta^{dc}\sigma^2 - \delta^{dc}\sigma^2 + 4\delta^{dc}\sigma^2\big)$$
$$= \tfrac{1}{4}\,\delta^{cd}\,\sigma_\lambda\sigma_\lambda,$$

(1) Here $\sigma^2 = \sigma_\lambda\sigma_\lambda$. The first term: $\bar\eta^d_{\mu\nu}\bar\eta^c_{\lambda\nu} = \bar\eta^d_{\nu\mu}\bar\eta^c_{\nu\lambda} = \delta^{dc}\delta_{\mu\lambda} + \varepsilon\bar\eta_{\mu\lambda}$ by (I), and the antisymmetric part dies against $\sigma_\lambda\sigma_\mu$. The second term likewise by (I) with summed μ. The third is $\bar\eta^d\cdot\bar\eta^c = 4\delta^{dc}$.

Total.

$$f^{cd} \overset{(1)}{=} \tfrac{1}{4}\delta^{cd}\big(\Box\sigma + \partial_\lambda\sigma\,\partial_\lambda\sigma\big)$$
$$\overset{(2)}{=} \tfrac{1}{4}\delta^{cd}\Big(\frac{\Box\rho}{\rho} - \frac{(\partial\rho)^2}{\rho^2} + \frac{(\partial\rho)^2}{\rho^2}\Big) = \frac{\Box\rho}{4\rho}\,\delta^{cd}.$$

(1) Sum of the two pieces.

(2) $\sigma = \ln\rho$: $\partial\sigma = \partial\rho/\rho$ and $\Box\sigma = \Box\rho/\rho - (\partial\rho)^2/\rho^2$.

So $F^- = \frac{\Box\rho}{4\rho}\sum_c e_c\omega_c^-$, which is (8.206). The magic is the cancellation in the last line: the non-linear term $(\partial\sigma)^2$ from $[A,A]$ is exactly what turns $\Box\ln\rho$ into $\Box\rho/\rho$.

Harmonic functions with poles

Positive harmonic functions on all of $\mathbb{R}^4$ are constant, and a constant ρ gives $A = 0$. So we allow isolated poles. In four dimensions the basic harmonic function with a pole is $1/|x|^2$ (the analogue of $1/r$ in three dimensions):

$$\Box\frac{1}{r^2} \overset{(1)}{=} \frac{1}{r^3}\frac{\mathrm{d}}{\mathrm{d}r}\Big(r^3\frac{\mathrm{d}}{\mathrm{d}r}r^{-2}\Big)$$

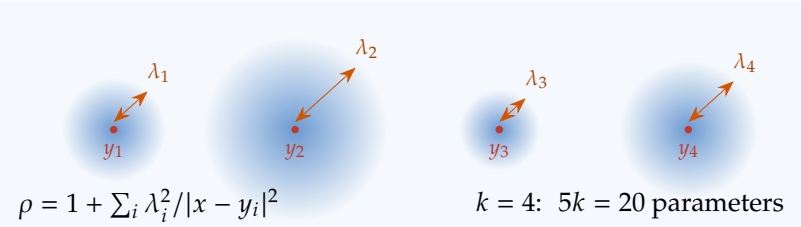
$$\overset{(2)}{=} \frac{1}{r^3}\frac{\mathrm{d}}{\mathrm{d}r}(-2) = 0 \qquad (r \neq 0).$$

(1) The Laplacian of a radial function in four dimensions, $\Box f = r^{-3}(r^3 f')'$ (the r^3 is the area factor of S^3_r).
(2) $r^3 \cdot (-2r^{-3}) = -2$, a constant.

(In the sense of distributions, $\Box(1/r^2) = -4\pi^2\delta^4(x)$, the 4D Coulomb law.) Hence

$$\rho(x) = 1 + \sum_{i=1}^{k} \frac{\lambda_i^2}{|x - y_i|^2} \tag{8.207}$$

is harmonic away from the k points y_i, and (8.205) is self-dual there. Near each y_i, $\rho \approx \lambda_i^2/|x-y_i|^2$ and A looks like a singular-gauge BPST instanton centred at y_i: a gauge singularity, removable by a local gauge transformation. So (8.207) describes k instantons of sizes λ_i at positions y_i, which may be anywhere, even overlapping. For $k = 1$ it is exactly the BPST instanton in singular gauge.



	positions y_i	sizes λ_i	relative colour orientations
count	$4k$	k	$3(k-1)$
in the 't Hooft ansatz?	yes	yes	no (all aligned)

Figure 8.37: A 't Hooft multi-instanton: k lumps of action, each with a centre y_i (red: the gauge singularity) and a size λ_i. The ansatz has $5k$ parameters. The full moduli space has dimension $8k - 3 = 5k + 3(k-1)$: the missing $3(k-1)$ are the orientations of the instantons in colour space relative to the first one, which the 't Hooft ansatz freezes. They are supplied by the ADHM construction of Section 8.10.

The action from ρ alone

For the 't Hooft ansatz there is a neat formula (one more page of index gymnastics of the same kind as Example 8.56, which we omit; it can be checked by computer algebra at any point):

$$-\tfrac{1}{2}\operatorname{tr}(F_{\mu\nu}F_{\mu\nu}) = -\tfrac{1}{2}\,\Box\Box \ln\rho \qquad \text{(away from the } y_i\text{)}. \tag{8.208}$$

Worked example 8.57: Checking (8.208) on the BPST instanton

Take $\rho = 1 + \lambda^2/r^2$, so $\ln\rho = \ln(r^2+\lambda^2) - \ln r^2$. The second term is harmonic after one Laplacian: $\Box \ln r^2 = 2\Box \ln r = 4/r^2$ (for $f = \ln r$, $r^{-3}(r^3 \cdot r^{-1})' = 2/r^2$), and $\Box(4/r^2) = 0$. So only the first term matters. With $u = r^2 + \lambda^2$ and the radial formula $\Box f = f'' + \frac{3}{r}f'$:

$$\Box \ln u \overset{(1)}{=} \left(\frac{2}{u} - \frac{4r^2}{u^2}\right) + \frac{3}{r}\cdot\frac{2r}{u}$$
$$\overset{(2)}{=} \frac{8}{u} - \frac{4(u-\lambda^2)}{u^2} = \frac{4}{u} + \frac{4\lambda^2}{u^2},$$

(1) $f = \ln u$ has $f' = 2r/u$ and $f'' = 2/u - 4r^2/u^2$.
(2) $r^2 = u - \lambda^2$.

and applying $\Box$ once more to $h = 4u^{-1} + 4\lambda^2 u^{-2}$:

$$\begin{aligned}\Box h &\overset{(1)}{=} \Big(-\frac{8}{u^2} + \frac{32r^2}{u^3} - \frac{16\lambda^2}{u^3} + \frac{96\lambda^2 r^2}{u^4}\Big) + \frac{3}{r}\Big(-\frac{8r}{u^2} - \frac{16\lambda^2 r}{u^3}\Big)\\ &\overset{(2)}{=} -\frac{32}{u^2} + \frac{32r^2}{u^3} - \frac{64\lambda^2}{u^3} + \frac{96\lambda^2 r^2}{u^4}\\ &\overset{(3)}{=} -\frac{96\lambda^4}{u^4}.\end{aligned}$$

(1) $h' = -8r/u^2 - 16\lambda^2 r/u^3$; differentiate again with $u' = 2r$.
(2) Collect: $-8 - 24 = -32$ and $-16 - 48 = -64$.
(3) Substitute $r^2 = u - \lambda^2$: the u^{-2} terms give $-32 + 32 = 0$, the $\lambda^2 u^{-3}$ terms $-32 - 64 + 96 = 0$, leaving $-96\lambda^4/u^4$.

So $-\frac{1}{2}\Box\Box \ln \rho = 48\lambda^4/(r^2 + \lambda^2)^4$, exactly the density of Example 8.54.

Worked example 8.58: Each pole carries one unit of action

Integrate (8.208) over $\mathbb{R}^4$ minus small balls $B_\varepsilon(y_i)$. Let $u = \Box \ln \rho$.

$$\begin{aligned}S &\overset{(1)}{=} -\tfrac{1}{2}\int_{\mathbb{R}^4 \setminus \cup B_\varepsilon(y_i)} \Box u \, \mathrm{d}^4x\\ &\overset{(2)}{=} -\tfrac{1}{2}\Big(\oint_{S^3_\infty} \partial_r u \, \mathrm{d}S - \sum_i \oint_{S^3_\varepsilon(y_i)} \partial_{r_i} u \, \mathrm{d}S\Big)\\ &\overset{(3)}{=} -\tfrac{1}{2}\Big(0 - \sum_{i=1}^{k} \frac{8}{\varepsilon^3} \cdot 2\pi^2\varepsilon^3\Big)\\ &= 8\pi^2 k.\end{aligned}$$

(1) (8.208). The true density is smooth at y_i (the pole is a gauge artefact), so removing tiny balls changes S by an amount that vanishes as $\varepsilon \to 0$.
(2) Gauss's theorem (Chapter T2). The boundary of the region is the sphere at infinity (outward normal ∂_r) and the small spheres, whose outward normal from the region points *into* the balls, i.e. along $-\partial_{r_i}$ with $r_i = |x - y_i|$.
(3) At infinity $\ln \rho = \sum \lambda_i^2/|x - y_i|^2 + O(r^{-4})$; the first part is harmonic, so $u = O(r^{-6})$ and the flux $r^3 \partial_r u = O(r^{-4}) \to 0$. Near y_i, $\ln \rho = -2 \ln r_i + (\text{smooth})$, so $u = -4/r_i^2 + (\text{bounded})$ and $\partial_{r_i} u = 8/r_i^3 + (\text{bounded})$; the sphere has area $2\pi^2\varepsilon^3$.

Since F is self-dual, k in the Bogomolny sense is also k: $8\pi^2 k = \|F^+\|^2 = S$. Each pole of ρ is one instanton.

Watch out: The 't Hooft ansatz is not the general solution

The ansatz gives $5k$ parameters, but the moduli space has dimension $8k - 3$ (Section 8.9). For $k = 1$ these agree ($5 = 5$). For $k = 2$ the ansatz gives 10 against 13. Jackiw, Nohl and Rebbi noticed that one may drop the "1" and use $\rho = \sum_{i=0}^{k} \lambda_i^2/|x - y_i|^2$, with $k + 1$ poles, which still has instanton number k (the extra pole is a pure gauge artefact); for $k = 2$ this reaches all 13 dimensions, but for $k \geq 3$ it falls short. The complete answer for every k is

the ADHM construction, outlined in Section 8.10. The missing parameters are the relative colour orientations of the lumps (Figure 8.37), which the 't Hooft ansatz fixes to be "all aligned", since every lump uses the same basis e_a.

Try it: Anti-instantons by the same trick

Show that $A_\mu = -\frac{1}{2}\eta^a_{\mu\nu}\partial_\nu \ln\rho\, e_a$ (with η instead of $\bar\eta$) has $F^+ = \frac{\Box\rho}{4\rho} e_a \omega^+_a$, so it is ASD when $\Box\rho = 0$. *Answer:* every step of Example 8.56 goes through with $\eta \leftrightarrow \bar\eta$ and $\omega^- \leftrightarrow \omega^+$, because identities (I) and (II) hold for both kinds of symbol.

Reading Nash (pp. 228–241)

Nash never writes an instanton down in his §3. After the Bogomolny bound (8.78)–(8.79) he says that the minima "are to be deemed self-dual" and moves straight to the moduli space. The explicit solution enters only implicitly, on [Nash p. 241], when he says that for SU(2) on S^4 and $k = 1$ "$\mathcal{M}_1$ is five-dimensional" and has the shape of hyperbolic 5-space. This section supplies what lies behind that sentence: the BPST connection (8.200), its self-dual curvature (8.202), its action $8\pi^2$, and its five parameters (centre and size). Our orientation $\mathrm{d}x^{1234}$ and $k = -c_2$ make it self-dual with $k = +1$, matching Nash's choice that the minima with $k > 0$ are self-dual.

Recap

- Quaternions are the matrices $x^4\mathbb{1} + x^k e_k$, $e_k = -i\sigma_k$; $\bar x x = |x|^2$, unit quaternions form $\mathrm{SU}(2) = S^3$. $\mathrm{d}\bar x \wedge \mathrm{d}x = -2e_a\omega^+_a$ is SD and $\mathrm{d}x \wedge \mathrm{d}\bar x = -2e_a\omega^-_a$ is ASD; the 't Hooft symbols are $\mathrm{Im}(\bar e_\nu e_\mu) = \eta^a_{\mu\nu} e_a$.
- Finite action forces $A \to g^{-1}\mathrm{d}g$ at infinity, and $k = \frac{1}{24\pi^2}\int_{S^3_\infty} \mathrm{tr}(g^{-1}\mathrm{d}g)^3 = \deg g$; the map $g = x/r$ has degree 1.
- BPST: $A = \mathrm{Im}(\bar x\,\mathrm{d}x)/(|x|^2+\lambda^2)$, $F = \lambda^2 \mathrm{d}\bar x \wedge \mathrm{d}x/(|x|^2+\lambda^2)^2$, self-dual, density $48\lambda^4/(r^2+\lambda^2)^4$, action $8\pi^2$ ($8\pi^2/g^2$ physically), $k = 1$, half the action within $r < \lambda$.
- 5 parameters up to gauge (centre, size), 8 with colour orientation.
- Singular gauge: $A \sim \lambda^2/r^3$ at infinity, winding moved to the centre. 't Hooft ansatz $A_\mu = -\frac{1}{2}\bar\eta^a_{\mu\nu}\partial_\nu \ln\rho\, e_a$: $F^- = \frac{\Box\rho}{4\rho} e_a \omega^-_a$, so harmonic $\rho = 1 + \sum \lambda_i^2/|x - y_i|^2$ gives k instantons with $5k$ of the $8k - 3$ moduli.

8.6 Gauge invariance, gauge fixing and ellipticity

We now have one explicit instanton and a family of others. Nash's aim is to describe *all* of them: the space of self-dual connections modulo gauge, and in particular its dimension. The strategy (following Atiyah, Hitchin and Singer) is to linearise the equations around a solution and count the solutions of the linearised problem with the index theorem. The index theorem (Chapter 4) applies to *elliptic* operators. So the first question is: are the Yang–Mills equations elliptic?

They are not, and the reason is gauge invariance. This section explains the failure, cures it by gauge fixing, and in doing so builds the geometric picture (orbits, slices, tangent spaces) that the rest of the part uses. A physicist has met every idea here already, in disguise: the photon propagator does not exist until one fixes a gauge.

8.6.1 A reminder: ellipticity and the symbol

A linear differential operator P of order m acting on vector-valued functions has a **principal symbol** $\sigma_P(x,\xi)$: a matrix-valued function of position x and a covector ξ ("momentum").

Working definition 8.26 (principal symbol and ellipticity)

To compute $\sigma_P(x,\xi)$: keep only the terms of P with the maximal number m of derivatives, and replace each ∂_μ by $i\xi_\mu$. The result is a matrix whose entries are homogeneous polynomials of degree m in ξ [every term contains exactly m factors of ξ]. The operator is **elliptic** if $\sigma_P(x,\xi)$ is an invertible matrix [no nonzero vector v has $\sigma_P(x,\xi)v=0$] for every x and every $\xi\neq 0$. (Formally: $\sigma_P(x,\xi)$ is a linear map between the fibres of the bundles P maps between [the vector spaces at x where the input and the output of P take their values]; see Chapter 2.)

The symbol is what a plane wave $e^{i\xi\cdot x}$ of very high frequency sees: for large ξ, $P(e^{i\xi\cdot x}v)\approx \sigma_P(x,\xi)v\,e^{i\xi\cdot x}$. Ellipticity means that no short-wavelength mode is annihilated. That is what gives the two properties we need:

- *Regularity.* Solutions of $Pu=0$ (and of non-linear elliptic equations) are smooth: no rough high-frequency component can hide in the kernel.
- *Fredholm property.* On a compact manifold $\operatorname{Ker}P$ and $\operatorname{Coker}P$ are finite dimensional, so $\operatorname{index}P=\dim\operatorname{Ker}P-\dim\operatorname{Coker}P$ makes sense and is computed by the index theorem.

Worked example 8.59: Symbols of d, d^* and the Laplacian on $\mathbb{R}^4$

On functions, $\mathrm{d}f=\partial_\mu f\,\mathrm{d}x^\mu$, so $\sigma_{\mathrm{d}}(\xi)f=i\xi_\mu f\,\mathrm{d}x^\mu=i\xi f$. On 1-forms in four dimensions $\mathrm{d}^*=-\star\,\mathrm{d}\star$, which gives $\mathrm{d}^*a=-\partial_\mu a_\mu$, so $\sigma_{\mathrm{d}^*}(\xi)a=-i\,\xi_\mu a_\mu=-i\,\xi\cdot a$. Composing,

$$\begin{aligned}\sigma_{\mathrm{d}^*\mathrm{d}}(\xi)f &\overset{(1)}{=} \sigma_{\mathrm{d}^*}(\xi)\,\sigma_{\mathrm{d}}(\xi)f\\ &\overset{(2)}{=} (-i)(i)\,\xi_\mu\xi_\mu\,f\\ &\overset{(3)}{=} |\xi|^2 f.\end{aligned}$$

(1) The symbol of a composition is the product of the symbols (only the terms where all derivatives hit the exponential survive at leading order).
(2) The two symbols above.
(3) $(-i)(i)=1$.

Indeed $\mathrm{d}^*\mathrm{d}f=-\Box f$. The symbol $|\xi|^2$ is invertible for $\xi\neq 0$: the Laplacian is the model elliptic operator. The covariant version $\mathrm{d}_A^*\mathrm{d}_A f=-D_\mu D_\mu f$ has the same symbol, because A enters without derivatives on f and only affects lower-order terms.

8.6.2 Why the Yang–Mills operator is not elliptic

The Yang–Mills equation $\mathrm{d}_A^*F(A)=0$ is non-linear. For a non-linear equation, ellipticity means ellipticity of its linearisation. Nash writes the operator loosely as "$\mathrm{d}_A^*\mathrm{d}_A$" acting on A; precisely, we perturb $A\to A+ta$ and differentiate at $t=0$.

$$\begin{aligned}\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0}\mathrm{d}^*_{A+ta}F(A+ta) &\overset{(1)}{=} \left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0}\Big(-\star\,\mathrm{d}_{A+ta}\star\big(F+t\,\mathrm{d}_A a+t^2 a\wedge a\big)\Big)\\ &\overset{(2)}{=} -\star\,\mathrm{d}_A\star\,\mathrm{d}_A a-\star[a,\star F]\end{aligned}$$

$$\overset{(3)}{=} \mathrm{d}_A^* \mathrm{d}_A a - \star[a, \star F].$$

(1) $\mathrm{d}_A^* = -\star \mathrm{d}_A \star$ on 2-forms (Section 8.3.3), and the curvature along the line, (8.168).
(2) $\mathrm{d}_{A+ta}\omega = \mathrm{d}_A\omega + t[a,\omega]$, with $[a,\omega]$ the graded commutator $a\wedge\omega - \omega\wedge a$ for a 2-form ω; keep the terms linear in t.
(3) $-\star\mathrm{d}_A\star = \mathrm{d}_A^*$ again.

The second term has no derivatives of a, so the principal symbol comes from $\mathrm{d}_A^*\mathrm{d}_A$ alone. In components on $\mathbb{R}^4$, $(\mathrm{d}^*\mathrm{d}a)_\nu = -\partial_\mu(\partial_\mu a_\nu - \partial_\nu a_\mu) = -\Box a_\nu + \partial_\nu(\partial\cdot a)$, so

$$\sigma_{\mathrm{d}_A^*\mathrm{d}_A}(\xi)_{\nu\mu} = |\xi|^2\delta_{\nu\mu} - \xi_\nu\xi_\mu. \tag{8.209}$$

Worked example 8.60: The symbol as a 4×4 matrix

Take $\xi = (0,0,0,1)$, a wave travelling in the x^4 direction. Then

$$\sigma(\xi) = |\xi|^2\mathbb{1}_4 - \xi\xi^T = \begin{pmatrix} 1&0&0&0\\ 0&1&0&0\\ 0&0&1&0\\ 0&0&0&0\end{pmatrix}.$$

Its kernel is spanned by $a = (0,0,0,1) = \xi$, the *longitudinal* polarisation. For general ξ, $\sigma(\xi) = |\xi|^2(\mathbb{1} - \xi\xi^T/|\xi|^2)$ is $|\xi|^2$ times the projector onto polarisations transverse to ξ. It kills exactly $a \propto \xi$. And $a = \xi f$ is precisely the symbol of a gauge direction: $\sigma_{\mathrm{d}_A}(\xi)f = i\xi f$. So the non-invertibility of the symbol is gauge invariance, seen at high frequency. For SU(2) the full symbol is this matrix tensored with $\mathbb{1}_3$ on the colour index, with a 3-dimensional kernel $\{\xi\otimes f : f\in\mathfrak{su}(2)\}$.

The mechanism is general. If $S(A_g) = S(A)$ for all gauge transformations g, then the gradient of S (the Yang–Mills operator) is constant along gauge orbits, so its linearisation annihilates the tangent vectors to the orbits. Those tangent vectors are $\mathrm{d}_A f$, a first-order expression in an arbitrary function f, so the kernel is infinite-dimensional and contains arbitrarily high frequencies: exactly what ellipticity forbids.

Dictionary: the photon propagator

For $G = \mathrm{U}(1)$ and $A = 0$, (8.209) in momentum space is the kinetic operator of the photon, $k^2\delta_{\mu\nu} - k_\mu k_\nu$. Every QFT course notes that this matrix has no inverse, because it annihilates k_μ, and that one cannot write the propagator until one adds a gauge-fixing term $\frac{1}{\xi}(\partial\cdot A)^2$ to the Lagrangian. In Feynman gauge ($\xi = 1$) the operator becomes $k^2\delta_{\mu\nu}$ with inverse $\delta_{\mu\nu}/k^2$. "The YM operator is not elliptic" is the mathematician's way of saying "the photon propagator does not exist without gauge fixing"; the cure below is the same.

Nash adds a second, more analytic, remark. If A is a smooth solution, so is $A_g = g^{-1}Ag + g^{-1}\mathrm{d}g$ for every gauge transformation g. If g is only, say, continuous, A_g still solves the equation in a weak sense but is not smooth. So without gauge fixing, "solutions are smooth" is simply false: regularity can only hold up to gauge.

8.6.3 Infinitesimal gauge transformations and the orbit

To fix the gauge near a connection A we need to know which directions in the space $\mathcal{A}$ of connections are gauge directions.

Worked example 8.61: Nash (8.80): the tangent to a gauge orbit

Take $g = \exp(tf)$ with $f \in \Omega^0(M, \operatorname{ad} P)$ and t small, so $g \approx \mathbb{1} + tf$ and $g^{-1} \approx \mathbb{1} - tf$.

$$\begin{aligned} A_g &\overset{(1)}{=} g^{-1}Ag + g^{-1}\mathrm{d}g \\ &\overset{(2)}{=} (\mathbb{1} - tf)A(\mathbb{1} + tf) + (\mathbb{1} - tf)\,\mathrm{d}(\mathbb{1} + tf) + O(t^2) \\ &\overset{(3)}{=} A + t(Af - fA) + t\,\mathrm{d}f + O(t^2) \\ &\overset{(4)}{=} A + t\,\mathrm{d}_A f + O(t^2). \end{aligned}$$

(1) The gauge transformation law (8.163).
(2) Expand $\exp(\pm tf)$ to first order.
(3) Multiply out and drop t^2 terms; $\mathrm{d}\mathbb{1} = 0$.
(4) $\mathrm{d}f + [A, f] = \mathrm{d}_A f$ by (8.164) with $p = 0$.

So the velocity of the curve $t \mapsto A_{\exp(tf)}$ at $t = 0$ is $\mathrm{d}_A f$.

Working definition 8.27 (gauge orbit and its tangent space)

The **gauge orbit** of A is $\mathcal{G} \cdot A = \{A_g : g \in \mathcal{G}\}$, all connections gauge equivalent to A. Its tangent space at A is

$$T_A(\mathcal{G} \cdot A) = \operatorname{Im} \mathrm{d}_A = \{\mathrm{d}_A f : f \in \Omega^0(M, \operatorname{ad} P)\} \subset \Omega^1(M, \operatorname{ad} P) = T_A\mathcal{A}. \tag{8.210}$$

Test: a perturbation a is "pure gauge to first order" [a is the velocity at $t = 0$ of a curve $A_{g(t)}$ of gauge transforms of A, with $g(0) = \mathbb{1}$] iff $a = \mathrm{d}_A f$ for some f. This is Nash's (8.83).

Recall that $\mathcal{A}$ is an affine space: any two connections differ by an element of $\Omega^1(M, \operatorname{ad} P)$, so every tangent space $T_A\mathcal{A}$ is $\Omega^1(M, \operatorname{ad} P)$ itself (Figure 8.29).

8.6.4 The slice: $T_A\mathcal{A} = \operatorname{Im} \mathrm{d}_A \oplus \operatorname{Ker} \mathrm{d}_A^*$

Physical (non-gauge) directions should be the directions orthogonal to the orbit. Orthogonality is measured by the inner product (8.165).

$$\begin{aligned} a \perp \operatorname{Im} \mathrm{d}_A &\overset{(1)}{=} \Longleftrightarrow \ \langle \mathrm{d}_A f, a \rangle = 0 \ \text{ for all } f \\ &\overset{(2)}{=} \Longleftrightarrow \ \langle f, \mathrm{d}_A^* a \rangle = 0 \ \text{ for all } f \\ &\overset{(3)}{=} \Longleftrightarrow \ \mathrm{d}_A^* a = 0. \end{aligned}$$

(1) Definition of the orthogonal complement.
(2) Definition (8.171) of the adjoint d_A^*.
(3) Take $f = \mathrm{d}_A^* a$: then $\left\|\mathrm{d}_A^* a\right\|^2 = 0$.

So the orthogonal complement of the orbit directions is $\operatorname{Ker} \mathrm{d}_A^*$. That the two pieces span everything is a statement of Hodge-theory type; we can see it by constructing the splitting explicitly.

Key idea: The gauge-slice decomposition (Nash (8.81), (8.84))

Every $a \in \Omega^1(M, \operatorname{ad} P)$ splits uniquely and orthogonally as

$$a = \underbrace{\mathrm{d}_A f}_{\text{gauge}} + \underbrace{a^\perp}_{\mathrm{d}_A^* a^\perp = 0}, \qquad T_A\mathcal{A} = \operatorname{Im} \mathrm{d}_A \oplus \operatorname{Ker} \mathrm{d}_A^* = T_A(\mathcal{G}\cdot A) \oplus \operatorname{Ker} \mathrm{d}_A^*. \tag{8.211}$$

Recipe: solve the elliptic equation $\mathrm{d}_A^* \mathrm{d}_A f = \mathrm{d}_A^* a$ for f, then set $a^\perp = a - \mathrm{d}_A f$.

Why the recipe works, one step at a time:

1. $\Delta_A = \mathrm{d}_A^* \mathrm{d}_A$ on $\Omega^0(\operatorname{ad} P)$ is the covariant Laplacian, elliptic by Example 8.59, and self-adjoint. On a compact M the equation $\Delta_A f = h$ has a solution iff h is orthogonal to $\operatorname{Ker} \Delta_A$ (the Fredholm alternative).
2. $\operatorname{Ker} \Delta_A = \operatorname{Ker} \mathrm{d}_A$, since $\langle \Delta_A f, f\rangle = \|\mathrm{d}_A f\|^2$.
3. $h = \mathrm{d}_A^* a$ is orthogonal to $\operatorname{Ker} \mathrm{d}_A$: for $\mathrm{d}_A f_0 = 0$, $\langle \mathrm{d}_A^* a, f_0\rangle = \langle a, \mathrm{d}_A f_0\rangle = 0$. So a solution f exists (unique up to adding elements of $\operatorname{Ker} \mathrm{d}_A$, which do not change $\mathrm{d}_A f$).
4. Then $\mathrm{d}_A^* a^\perp = \mathrm{d}_A^* a - \mathrm{d}_A^* \mathrm{d}_A f = 0$, and the two pieces are orthogonal by the chain above. Uniqueness: an element of both pieces is orthogonal to itself.

Worked example 8.62: The splitting in Fourier space: longitudinal and transverse

Take $A = 0$ and $G = \mathrm{U}(1)$ on $\mathbb{R}^4$ (or a torus), so $\mathrm{d}_A = \mathrm{d}$. For a plane wave $a_\mu(x) = \hat{a}_\mu e^{i\xi\cdot x}$ the recipe reads

$$|\xi|^2 \hat{f} \overset{(1)}{=} -i\,\xi\cdot\hat{a}, \qquad \text{so} \qquad \hat{f} = \frac{-i\,\xi\cdot\hat{a}}{|\xi|^2},$$

$$\hat{a}^\perp_\mu \overset{(2)}{=} \hat{a}_\mu - i\xi_\mu \hat{f} \overset{(3)}{=} \hat{a}_\mu - \frac{\xi_\mu\,\xi\cdot\hat{a}}{|\xi|^2}.$$

(1) $\mathrm{d}^*\mathrm{d} f = -\Box f$ becomes $|\xi|^2\hat{f}$ and $\mathrm{d}^* a = -\partial\cdot a$ becomes $-i\xi\cdot\hat{a}$.
(2) $a^\perp = a - \mathrm{d}f$ and $\mathrm{d}f \to i\xi\hat{f}$.
(3) Substitute $\hat{f}$.

This is the familiar split of a vector field into its longitudinal part $\xi(\xi\cdot\hat{a})/|\xi|^2$ (pure gauge) and its transverse part, which satisfies $\xi\cdot\hat{a}^\perp = 0$, the Lorenz (Euclidean: Landau) gauge condition. With numbers: $\xi = (1,0,0,0)$, $\hat{a} = (3,1,0,2)$ gives gauge part $(3,0,0,0)$ and transverse part $(0,1,0,2)$.

The background gauge

The decomposition suggests the gauge condition Nash adopts. For connections $A + a$ near A, impose

$$\mathrm{d}_A^* a = 0 \qquad \big(\text{in components } -D_\mu a_\mu = -(\partial_\mu a_\mu + [A_\mu, a_\mu]) = 0\big). \tag{8.212}$$

This is Nash's (8.82). Physicists call it the **background field gauge** (with A the background); for $A = 0$ it is the ordinary Lorenz condition $\partial\cdot a = 0$. It picks, in each nearby gauge orbit, the representative closest to A (Example 8.63 below). Unlike the temporal gauge $A_4 = 0$ or the axial gauge $A_3 = 0$, it is covariant, which is what makes the resulting equations elliptic.

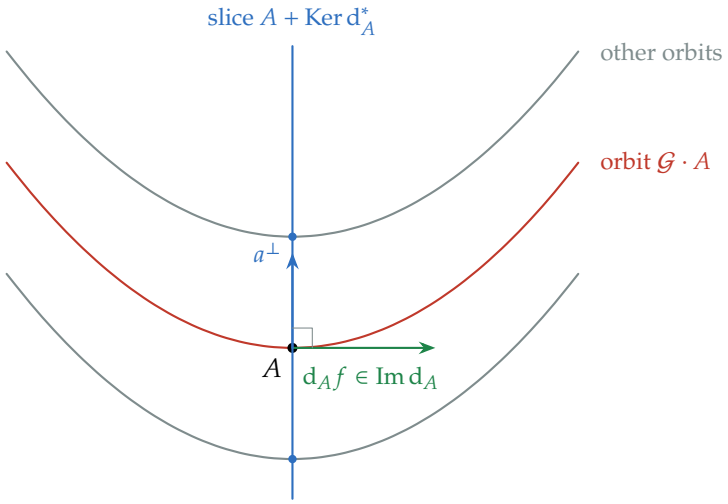


Figure 8.38: The geometry of gauge fixing, drawn in the (infinite-dimensional) space $\mathcal{A}$ of connections. Gauge orbits are the curves: every connection on one curve describes the same physics. At A the tangent to the orbit is $\operatorname{Im} \mathrm{d}_A$ (green). The slice (blue) passes through A orthogonally to the orbit; its tangent directions satisfy the background gauge condition $\mathrm{d}_A^* a = 0$. Near A, each nearby orbit meets the slice exactly once, so points of the slice are local coordinates on $\mathcal{A}/\mathcal{G}$.

Worked example 8.63: The background gauge picks the closest point of the orbit

Fix A and a nearby connection $A + a$. Move $A + a$ along its orbit by an infinitesimal gauge transformation $\exp(t f)$: to first order it becomes $A + a + t\, \mathrm{d}_{A+a} f \approx A + a + t\, \mathrm{d}_A f$ (dropping $t a f$ terms, which are second order in small quantities). The squared distance to A changes as

$$\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} \|a + t\, \mathrm{d}_A f\|^2 \overset{(1)}{=} 2\langle a, \mathrm{d}_A f\rangle \overset{(2)}{=} 2\langle \mathrm{d}_A^* a, f\rangle.$$

(1) Expand the norm squared; bilinearity and symmetry.
(2) Adjoint.

The distance is stationary under all gauge motions iff $\mathrm{d}_A^* a = 0$. So the slice is the set of points of each orbit that are (locally) nearest to A.

8.6.5 The gauge-fixed equations are elliptic

Adding the gauge condition to the Yang–Mills equation gives the system (Nash's (8.85))

$$\mathrm{d}_{A+a}^* F(A + a) = 0, \qquad \mathrm{d}_A^* a = 0.$$

A standard trick packages it as one equation: since $\mathrm{d}_A \mathrm{d}_A^* a$ vanishes when $\mathrm{d}_A^* a = 0$, we may add it. The linearised operator becomes

$$a \longmapsto \mathrm{d}_A^* \mathrm{d}_A a + \mathrm{d}_A \mathrm{d}_A^* a + (\text{zeroth order}),$$

whose principal symbol is

$$\sigma(\xi)_{\nu\mu} \overset{(1)}{=} \left(|\xi|^2 \delta_{\nu\mu} - \xi_\nu \xi_\mu\right) + \xi_\nu \xi_\mu \overset{(2)}{=} |\xi|^2\, \delta_{\nu\mu}.$$

(1) (8.209), plus the symbol of dd^*: $(\mathrm{dd}^* a)_\nu = -\partial_\nu \partial_\mu a_\mu$, so $\sigma = -(i\xi_\nu)(i\xi_\mu) = \xi_\nu \xi_\mu$.
(2) The longitudinal pieces cancel.

This is invertible for $\xi \neq 0$: the gauge-fixed operator is the Hodge Laplacian on 1-forms (plus lower-order terms), the model elliptic operator. In the 4×4 matrix of Example 8.60, the missing 1 in the bottom right corner has been restored. Elliptic regularity then gives Nash's conclusion: the gauge-fixed Yang–Mills equations are an elliptic system, and their solutions are smooth.

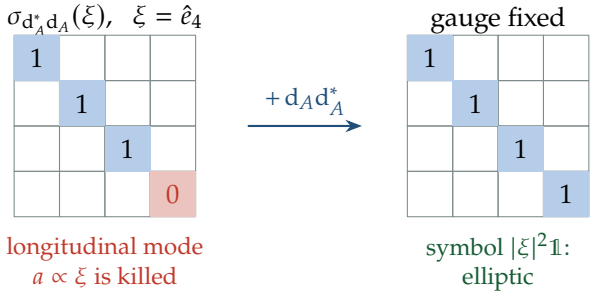


Figure 8.39: The symbol of the linearised Yang–Mills operator for a wave along x^4, before and after gauge fixing. Before: the longitudinal polarisation (the gauge direction) has zero symbol, so the operator is not elliptic. After adding $\mathrm{d}_A \mathrm{d}_A^*$ (the gauge-fixing term, Feynman gauge in physics language) every polarisation has symbol $|\xi|^2$.

Watch out: Gauge fixing is local: the Gribov problem

The slice meets each *nearby* orbit exactly once. Far from A an orbit can cross the slice several times, or be tangent to it; then the condition $\mathrm{d}_A^* a = 0$ does not pick a unique representative. Physicists know this as the **Gribov ambiguity** of the Coulomb and Landau gauges. Singer (1978) showed it is unavoidable for non-Abelian G on S^4: there is no continuous global gauge-fixing section of $\mathcal{A} \to \mathcal{A}/\mathcal{G}$, because the bundle $\mathcal{A} \to \mathcal{A}/\mathcal{G}$ is topologically non-trivial. Nash's word "(local) gauge" is exactly this caveat. For the moduli space it is harmless: we only ever need coordinates near one self-dual connection at a time.

Reading Nash (pp. 230–231, (8.80)–(8.85))

Nash motivates ellipticity by the analogy of Yang–Mills with (non-linear) Hodge theory (Section 8.3.4). He then states that gauge invariance "gives the Yang–Mills operator $\mathrm{d}_A^* \mathrm{d}_A$ a non-invertible symbol, whose kernel is parametrised by the space of gauge transformations": (8.209) and Example 8.60. In his "$\mathrm{d}_A^* \mathrm{d}_A A \equiv \mathrm{d}_A^* F(A)$" the operator notation is loose: the equation is non-linear, and the operator with a symbol is its linearisation. (8.80) is our Example 8.61; (8.81) and (8.84) are (8.211), which he calls "standard linear algebra" (in infinite dimensions it needs the elliptic argument we gave); (8.82) is the background gauge (8.212); (8.83) is (8.210). (8.85) is the elliptic gauge-fixed system. (Nash's "orthogonal compliment" is a typo for complement.)

Recap

- Elliptic means the principal symbol is invertible for $\xi \neq 0$; it gives smooth solutions and finite-dimensional kernels, hence indices.
- The linearised YM operator has symbol $|\xi|^2 \delta - \xi\xi$, which kills the longitudinal (gauge) polarisation: YM is not elliptic, because of gauge invariance. Physics: no photon propagator without gauge fixing.
- Infinitesimal gauge transformations move A by $\mathrm{d}_A f$; the orbit's tangent space is

$\operatorname{Im} \mathrm{d}_A$.

- $T_A\mathcal{A} = \operatorname{Im} \mathrm{d}_A \oplus \operatorname{Ker} \mathrm{d}_A^*$, constructed by solving $\mathrm{d}_A^* \mathrm{d}_A f = \mathrm{d}_A^* a$. The slice $\mathrm{d}_A^* a = 0$ (background gauge) picks the orbit point nearest A.
- With the gauge condition the operator becomes the Hodge Laplacian, symbol $|\xi|^2$: elliptic. Gauge fixing is only local (Gribov, Singer).

8.7 The instanton moduli space and reducible connections

We have an explicit family of instantons: the BPST solution with its centre $y \in \mathbb{R}^4$ and size $\lambda > 0$, and the 't Hooft solutions with $5k$ parameters. Are there others? How many parameters does the most general self-dual connection of charge k have? And in what sense are two instantons "the same"? Physically this is the question of the **collective coordinates** of the instanton. In the semiclassical evaluation of the path integral one integrates over them exactly, and each one comes with a zero mode of the fluctuation operator (Section 8.10). Mathematically it is the question of the **moduli space**.

This section defines the moduli space carefully. The one real subtlety is that the gauge group does not always act freely: some connections are fixed by some gauge transformations. Those connections are called *reducible*, and they produce singularities in the moduli space. We explain what reducibility is, how to test for it, why it causes singularities, and how to avoid it (Nash's two devices: divide by Ad G and restrict to irreducible connections, or keep a framing).

8.7.1 Definition of $\mathcal{M}_k$

Fix the bundle, that is, fix the instanton number k. Write $\mathcal{A}_k$ for the space of all connections on this bundle and $\mathcal{G}_k$ for the group of gauge transformations (locally: maps $g : M \to G$, acting by (8.163)). Two connections related by a gauge transformation describe the same physics, so the *configuration space* of the theory is the quotient $\mathcal{A}_k/\mathcal{G}_k$: the set of gauge orbits.

Working definition 8.28 (instanton moduli space)

The **moduli space** of charge-k instantons is

$$\mathcal{M}_k = \frac{\{A \in \mathcal{A}_k : \star F(A) = F(A)\}}{\mathcal{G}_k} \subset \frac{\mathcal{A}_k}{\mathcal{G}_k}, \tag{8.213}$$

the set of gauge orbits of self-dual connections (Nash's (8.86)–(8.87)). *To compute with it:* pick a self-dual A; nearby points of $\mathcal{M}_k$ are self-dual $A + a$ with a in the slice $\mathrm{d}_A^* a = 0$ [perturbations a that are L^2-orthogonal to the gauge orbit through A] of Section 8.6.4. *To count its dimension:* count the independent solutions a of the linearised self-duality equation [the part of $\star F(A + a) = F(A + a)$ linear in a, which is $(\mathrm{d}_A a)^- = 0$] that are orthogonal to the gauge orbit (Section 8.8).

The whole point of Nash's section is that $\mathcal{M}_k$ is *finite* dimensional, although it sits inside the infinite-dimensional space $\mathcal{A}_k/\mathcal{G}_k$. The self-duality equation is a system of three equations per colour for four unknowns per colour, and the fourth "missing" equation is supplied by gauge invariance; so the system is "determined" and we expect a discrete set of solutions plus finitely many parameters.

Worked example 8.64: Counting the BPST parameters

The BPST connection centred at y with size λ is $A_{y,\lambda} = \operatorname{Im}\big((\bar{x}-\bar{y})\,\mathrm{d}x\big)/(|x-y|^2+\lambda^2)$ (Section 8.5.3). Conjugating by a constant $q \in \mathrm{SU}(2)$ gives another self-dual connection $q^{-1}A_{y,\lambda}\,q$. So in $\mathcal{A}_1$ we see at least $4+1+3=8$ parameters (y,λ,q). How many survive in $\mathcal{M}_1$?

- The constant conjugation *is* a gauge transformation ((8.163) with $\mathrm{d}q=0$). So $q^{-1}Aq$ and A are the same point of $\mathcal{A}_1/\mathcal{G}_1$, and q does not survive.
- Different (y,λ) give genuinely different points: the gauge-invariant action density $s(x) = -\frac{1}{2}\operatorname{tr}F_{\mu\nu}F_{\mu\nu} = 48\lambda^4/(|x-y|^2+\lambda^2)^4$ peaks at $x=y$ with height $48/\lambda^4$, so it remembers both y and λ.

So the BPST family contributes a 5-dimensional piece of $\mathcal{M}_1$, the set of pairs $(y,\lambda)\in\mathbb{R}^4\times\mathbb{R}^+$. The index computation of Section 8.9 will show that $\dim\mathcal{M}_1=5$ exactly, so this piece is (an open subset of, and in fact all of) the moduli space.

Worked example 8.65: A constant rotation of colour is a rotation of space

The last example has a pretty twist. Take $y=0$ and a unit quaternion q (so $q^{-1}=\bar{q}$). Then

$$q^{-1}A\,q \overset{(1)}{=} \frac{\bar{q}\,\operatorname{Im}(\bar{x}\,\mathrm{d}x)\,q}{\phi} \overset{(2)}{=} \frac{\operatorname{Im}\big(\bar{q}\,\bar{x}\,\mathrm{d}x\,q\big)}{\phi} \overset{(3)}{=} \frac{\operatorname{Im}\big(\overline{(xq)}\,\mathrm{d}(xq)\big)}{|xq|^2+\lambda^2}.$$

(1) The BPST formula (8.200), with $\phi=|x|^2+\lambda^2$.

(2) Conjugation by a unit quaternion preserves real parts ($\bar{q}\,1\,q=1$), hence commutes with taking the imaginary part.

(3) $\overline{xq}=\bar{q}\bar{x}$, $\mathrm{d}(xq)=\mathrm{d}x\,q$ because q is constant, and $|xq|=|x|$.

The right-hand side is the BPST connection pulled back by the map $x\mapsto xq$, which is a rotation of $\mathbb{R}^4$ (it preserves $|x|$). So for the instanton a global colour rotation is the same as a spatial rotation. The BPST instanton is invariant under the combined transformation "rotate space by $x\mapsto xq$, then rotate colour back by q". This is the "spin–isospin locking" of the instanton. It is the reason the instanton has only 5 moduli rather than $5+6$ (rotations of $\mathbb{R}^4$ form the 6-dimensional group SO(4)): the rotations either leave it fixed or act on it as gauge transformations.

8.7.2 The residual gauge symmetry: base points and $\mathrm{Ad}\,G$

Nash now points out a remaining invariance. To see it, recall how the gauge group is organised. A gauge transformation is a section g of the bundle of groups $\mathrm{Ad}\,P$ (locally a map $g: M\to G$). Choose a **base point** $p_0\in M$ (on S^4, the point at infinity of $\mathbb{R}^4$ is the natural choice) and evaluate:

$$\mathrm{ev}_{p_0}:\mathcal{G}\to G, \qquad g\mapsto g(p_0).$$

The kernel is the **based gauge group** $\mathcal{G}_0=\{g\in\mathcal{G}:\ g(p_0)=\mathbb{1}\}$. Every g is a based gauge transformation times a constant (on a trivialised neighbourhood of p_0), and $\mathcal{G}/\mathcal{G}_0\cong G$. So the quotient can be taken in two stages:

$$\frac{\mathcal{A}_k}{\mathcal{G}_k} = \frac{\mathcal{A}_k/\mathcal{G}_0}{G}. \tag{8.214}$$

Nash's phrasing is that the action "does not depend on our choice of base point $p_0 \in P_a$": a point of the fibre P_{p_0} is a choice of frame at p_0, the frames at p_0 are permuted by G, and this G still acts on $\mathcal{A}_k/\mathcal{G}_0$. (Here $\mathcal{G}_k$ in Nash's notation is the based group; the distinction matters only in this paragraph.)

Working definition 8.29 (framed connection, framed moduli space)

A **framed connection** is a connection together with a choice of frame (a basis, or equivalently a point of P) at the base point p_0. Gauge transformations in $\mathcal{G}_0$ do not move the frame; constant gauge transformations [equal to one fixed $\alpha \in G$ near p_0, in the trivialisation used there] do. The **framed moduli space** is

$$\mathcal{M}'_k = \frac{\{A \in \mathcal{A}_k : \star F = F\}}{\mathcal{G}_0}, \qquad \text{with} \qquad \mathcal{M}_k = \mathcal{M}'_k/G. \tag{8.215}$$

Physically: $\mathcal{M}'_k$ remembers the orientation of the instanton in colour space [the internal space $\mathbb{C}^2$ on which SU(2) acts] relative to a fixed reference frame at infinity; $\mathcal{M}_k$ forgets it.

Worked example 8.66: The framed $k = 1$ moduli space has dimension 8

For the BPST family, the constant conjugation $q^{-1}Aq$ changes the frame at infinity (it is a gauge transformation with $g(\infty) = q \neq \mathbb{1}$), so in $\mathcal{M}'_1$ the colour orientation q *is* a parameter. The count of Example 8.64 becomes $4 + 1 + 3 = 8$. But the element $q = -\mathbb{1}$ acts trivially ($(-1)^{-1}A(-1) = A$), so the orientation lives in $\mathrm{SU}(2)/\{\pm\mathbb{1}\} = \mathrm{SO}(3)$, and

$$\mathcal{M}'_1 \cong \mathbb{R}^4 \times \mathbb{R}^+ \times \mathrm{SO}(3) \cong \mathbb{R}^4 \times \big(\mathbb{R}^4 \setminus \{0\}\big)/\mathbb{Z}_2.$$

In the last form the size and the orientation are combined into one non-zero quaternion λq (modulus λ, direction q), defined up to sign. We will meet exactly this parameter in the ADHM construction (Section 8.10.6).

The centre acts trivially

Does the residual G act freely on $\mathcal{A}_k/\mathcal{G}_0$? Nash observes that it never does, for a simple reason. G is simple, so its centre $Z(G)$ (the elements commuting with everything) is a finite group. For $G = \mathrm{SU}(N)$, $Z(G) = \{e^{2\pi i m/N}\mathbb{1} : m = 0, \dots, N-1\} \cong \mathbb{Z}_N$; for SU(2) it is $\{\pm\mathbb{1}\}$.

$$\begin{aligned} z \in Z(G) \text{ constant}: \qquad A_z &\overset{(1)}{=} z^{-1}Az + z^{-1}\mathrm{d}z \\ &\overset{(2)}{=} z^{-1}Az \\ &\overset{(3)}{=} A. \end{aligned}$$

(1) The gauge transformation law (8.163).
(2) z is constant, so $\mathrm{d}z = 0$.
(3) z commutes with every matrix in $\mathfrak{g}$ (it is a multiple of the identity for $\mathrm{SU}(N)$).

So the centre fixes every connection. It is harmless: we simply divide it out from the start and let the group

$$\mathrm{Ad}\, G = G/Z(G)$$

act instead. For SU(2), $\operatorname{Ad}\mathrm{SU}(2) = \mathrm{SU}(2)/\mathbb{Z}_2 = \mathrm{SO}(3)$, the group of rotations of $\mathfrak{su}(2) \cong \mathbb{R}^3$ by $X \mapsto g^{-1}Xg$. Nash's new configuration space is (8.88):

$$\left(\mathcal{A}_k^{\mathrm{irred}}/\mathcal{G}_k\right) / \operatorname{Ad} G. \tag{8.216}$$

The superscript "irred" restricts to *irreducible* connections, which are defined next. The reason is that some connections are fixed by more than the centre, and those are the ones to exclude.

8.7.3 Stabilisers, holonomy and reducibility

Working definition 8.30 (stabiliser)

The **stabiliser** (or isotropy group) of a connection is

$$\Gamma_A = \{g \in \mathcal{G} : \; A_g = A\}.$$

Test: $A_g = A$ means $g^{-1}Ag + g^{-1}\mathrm{d}g = A$, i.e.

$$\mathrm{d}g + [A, g] = 0, \qquad \text{i.e.} \qquad \mathrm{d}_A g = 0 : \tag{8.217}$$

g is *covariantly constant* [unchanged by parallel transport along every path], as a section of $\operatorname{Ad} P$. (Multiply $A_g = A$ on the left by g: $Ag + \mathrm{d}g = gA$.)

A covariantly constant object is determined by its value at one point, because we can parallel transport that value everywhere. That is why stabilisers are controlled by *holonomy*. Let us make this precise; it is Nash's (8.89)–(8.90).

Parallel transport and how it changes under gauge transformations

Along a curve $\gamma : [0,1] \to M$ from $a = \gamma(0)$ to $b = \gamma(1)$, the **parallel transport** $U_\gamma(A)$ is the matrix obtained by solving

$$\frac{\mathrm{d}\psi}{\mathrm{d}t} = -A(\dot\gamma)\,\psi, \qquad \psi(0) = \psi_0, \qquad U_\gamma(A)\psi_0 \coloneqq \psi(1),$$

i.e. the path-ordered exponential $U_\gamma(A) = \mathcal{P}\exp\bigl(-\int_\gamma A\bigr)$. Physicists know it as the **Wilson line** (see Chapter T3). For a closed loop based at a, $U_\gamma(A)$ is the **holonomy** around γ; the set of all holonomies around loops at a forms the **holonomy group** $\mathrm{Hol}_a(A) \subset G$.

How does U_γ change under $A \mapsto A_g$? Try $\tilde\psi = g^{-1}\psi$, where ψ solves the transport equation for A:

$$\begin{aligned}
\frac{\mathrm{d}\tilde\psi}{\mathrm{d}t} &\overset{(1)}{=} \frac{\mathrm{d}g^{-1}}{\mathrm{d}t}\psi + g^{-1}\frac{\mathrm{d}\psi}{\mathrm{d}t} \\
&\overset{(2)}{=} -g^{-1}\dot g\, g^{-1}\psi - g^{-1}A(\dot\gamma)\psi \\
&\overset{(3)}{=} -\bigl(g^{-1}\dot g + g^{-1}A(\dot\gamma)g\bigr)\, g^{-1}\psi \\
&\overset{(4)}{=} -A_g(\dot\gamma)\,\tilde\psi.
\end{aligned}$$

(1) Product rule; g is evaluated along γ.
(2) $\mathrm{d}(g^{-1}) = -g^{-1}\mathrm{d}g\, g^{-1}$ (differentiate $g^{-1}g = \mathbb{1}$), and the transport equation for ψ.
(3) Insert $\mathbb{1} = gg^{-1}$ in the second term and factor.
(4) $A_g = g^{-1}Ag + g^{-1}\mathrm{d}g$ evaluated on $\dot\gamma$; $g^{-1}\dot g = (g^{-1}\mathrm{d}g)(\dot\gamma)$.

So $\tilde{\psi}$ is transported by A_g, and comparing endpoints,

$$U_\gamma(A_g) = g(b)^{-1}\, U_\gamma(A)\, g(a). \tag{8.218}$$

This is Nash's (8.89), $\alpha_b \circ PT_{\gamma_{ab}}(A) = PT_{\gamma_{ab}}(\alpha \cdot A) \circ \alpha_a$, written with our conventions (α_a denotes the value of the gauge transformation at a).

Stabilisers commute with holonomy

Now take a loop γ based at a and a stabiliser element $g \in \Gamma_A$, with value $\alpha = g(a)$ at the base point.

$$\begin{aligned} U_\gamma(A) &\overset{(1)}{=} U_\gamma(A_g) \\ &\overset{(2)}{=} \alpha^{-1}\, U_\gamma(A)\, \alpha, \end{aligned}$$

(1) $A_g = A$ because g is in the stabiliser.
(2) (8.218) with $a = b$, so $g(a) = g(b) = \alpha$.

hence $\alpha\, U_\gamma(A) = U_\gamma(A)\, \alpha$ for every loop: this is (8.90). Conversely, given α commuting with the holonomy group, define $g(x) = U_{\gamma_x}(A)\, \alpha\, U_{\gamma_x}(A)^{-1}$ by transporting along any path γ_x from a to x; the result does not depend on the path (two paths differ by a loop, whose holonomy commutes with α), and it is covariantly constant. So:

Key idea: Stabiliser = centraliser of the holonomy group

Evaluation at the base point identifies

$$\Gamma_A \;\cong\; C_G\big(\mathrm{Hol}(A)\big) = \{\alpha \in G :\ \alpha h = h\alpha \text{ for all } h \in \mathrm{Hol}(A)\}. \tag{8.219}$$

A stabiliser element is fixed by its value at one point. In particular a based gauge transformation ($g(p_0) = \mathbb{1}$) that fixes A is the identity: $\mathcal{G}_0$ *acts freely on all connections.*

Working definition 8.31 (irreducible connection)

A connection is **irreducible** if its stabiliser is as small as possible, namely the centre [the elements of G that commute with all of G, $\pm\mathbb{1}$ for SU(2)]: $\Gamma_A = Z(G)$. Otherwise it is **reducible**. (Nash's formulation: A is irreducible if its holonomy group is all of G; then the centraliser is $Z(G)$. The name comes from representation theory: A is reducible if the holonomy preserves a proper subspace [a nonzero subspace smaller than the whole fibre, such as a line in $\mathbb{C}^2$], i.e. the bundle splits into parallel sub-bundles.) *Infinitesimal test* (Nash's (8.91)): A is reducible iff there is a non-zero $f \in \Omega^0(M, \mathrm{ad}\, P)$ with

$$\mathrm{d}_A f = 0. \tag{8.220}$$

Why the infinitesimal test works: a covariantly constant Lie-algebra element f exponentiates to a one-parameter family of stabiliser elements $g_t = \exp(tf)$, and conversely the stabiliser is a closed subgroup of G (a Lie group), so if it is bigger than the finite group $Z(G)$ it has a non-trivial Lie algebra, whose elements satisfy (8.220). Nash shows the first half by expanding A_g in powers of t, (8.92). We can do better and show that A_{g_t} is exactly constant in t:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} A_{g_t} &\overset{(1)}{=} \frac{\mathrm{d}}{\mathrm{d}t}\Big(e^{-tf} A\, e^{tf} + e^{-tf}\, \mathrm{d}\, e^{tf}\Big) \\ &\overset{(2)}{=} e^{-tf}(Af - fA)e^{tf} + \Big(-f\, e^{-tf}\, \mathrm{d}e^{tf} + e^{-tf}\, \mathrm{d}\big(f\, e^{tf}\big)\Big) \end{aligned}$$

$$\overset{(3)}{=} e^{-tf}[A,f]e^{tf} + e^{-tf}(\mathrm{d}f)\,e^{tf}$$
$$\overset{(4)}{=} e^{-tf}\,(\mathrm{d}_A f)\,e^{tf}$$
$$\overset{(5)}{=} 0.$$

(1) The gauge transformation law with $g_t = e^{tf}$.
(2) $\frac{\mathrm{d}}{\mathrm{d}t}e^{\pm tf} = \pm f e^{\pm tf}$, and f commutes with $e^{\pm tf}$.
(3) Leibniz: $\mathrm{d}(fe^{tf}) = (\mathrm{d}f)e^{tf} + f\,\mathrm{d}e^{tf}$; the terms $\mp f e^{-tf}\mathrm{d}e^{tf}$ cancel (again $fe^{-tf} = e^{-tf}f$).
(4) $\mathrm{d}_A f = \mathrm{d}f + [A,f]$.
(5) The assumption (8.220).

So $A_{g_t} = A_{g_0} = A$ for all t. Expanding $e^{-tf}Xe^{tf} = X + t[X,f] + \frac{t^2}{2}[[X,f],f] + \dots$ in the fourth line and integrating reproduces Nash's series $A_g = A + t\,\mathrm{d}_A f + \frac{t^2}{2}[\mathrm{d}_A f, f] + \dots$ term by term.

Worked example 8.67: Three connections and their stabilisers, for $G = \mathrm{SU}(2)$

(a) The trivial connection $A = 0$. Then $\mathrm{d}_A f = \mathrm{d}f$, so every constant $f \in \mathfrak{su}(2)$ is covariantly constant: the space of solutions of (8.220) is 3-dimensional. The holonomy is trivial, its centraliser is all of SU(2), and $\Gamma_A = \mathrm{SU}(2)$ (the constant gauge transformations). This is the "most reducible" connection.
(b) A U(1) *connection.* Let $\alpha = i\alpha_\mu \mathrm{d}x^\mu$ be a U(1) connection (α_μ real) and put

$$A = \begin{pmatrix} \alpha & 0 \\ 0 & -\alpha \end{pmatrix} = i\alpha_\mu\,\sigma_3\,\mathrm{d}x^\mu = -\alpha_\mu\,e_3\,\mathrm{d}x^\mu.$$

It takes values in the diagonal (Cartan) subalgebra. Take $f = i\sigma_3 = -e_3$:

$$\mathrm{d}_A f \overset{(1)}{=} \mathrm{d}f + [A,f]$$
$$\overset{(2)}{=} 0 + \alpha_\mu\,\mathrm{d}x^\mu\,[i\sigma_3, i\sigma_3]$$
$$\overset{(3)}{=} 0.$$

(1) Definition of d_A on 0-forms.
(2) f is constant; $A = \alpha_\mu \mathrm{d}x^\mu\, i\sigma_3$.
(3) A matrix commutes with itself.

So A is reducible; the solutions of (8.220) are the multiples of $i\sigma_3$ (if the curvature of α is non-zero somewhere, $[F,f] = 0$ forces $f \propto i\sigma_3$; see (c)). The stabiliser is the diagonal circle $\{\mathrm{diag}(e^{i\theta}, e^{-i\theta})\} \cong \mathrm{U}(1)$. The bundle splits as $E = L \oplus L^{-1}$, the eigenlines of σ_3, each preserved by parallel transport.
(c) The BPST instanton. Suppose $\mathrm{d}_A f = 0$. Apply d_A again:

$$0 \overset{(1)}{=} \mathrm{d}_A \mathrm{d}_A f$$
$$\overset{(2)}{=} [F,f]$$
$$\overset{(3)}{=} \tfrac{1}{2}\,[F_{\mu\nu}, f]\,\mathrm{d}x^\mu \wedge \mathrm{d}x^\nu.$$

(1) Apply d_A to $\mathrm{d}_A f = 0$.
(2) $\mathrm{d}_A\mathrm{d}_A f = Ff - fF$ for any $f \in \Omega^0(\mathrm{ad}\,P)$; this is Nash's (8.108), which we verify in Section 8.8.4.
(3) Components.

So $f(x)$ must commute with every $F_{\mu\nu}(x)$. For BPST, $F_{\mu\nu} = -2\lambda^2\eta^a_{\mu\nu}e_a/\phi^2$ (Section 8.5.3); at any point, $F_{14} \propto e_1$, $F_{24} \propto e_2$, $F_{34} \propto e_3$. An element of $\mathfrak{su}(2)$ commuting with e_1, e_2 and e_3 is zero, by (8.193). Hence $f = 0$: the BPST instanton is irreducible, $\Gamma_A = \{\pm\mathbb{1}\} = Z(\mathrm{SU}(2))$.

The argument in (c) is a useful general test.

Key idea: A curvature test for irreducibility

If at some point the components $F_{\mu\nu}(x)$ of the curvature generate $\mathfrak{g}$ as a Lie algebra, then A is irreducible. For SU(2): if at some point two components $F_{\mu\nu}(x)$ are not parallel (as vectors in $\mathfrak{su}(2) \cong \mathbb{R}^3$), A is irreducible.

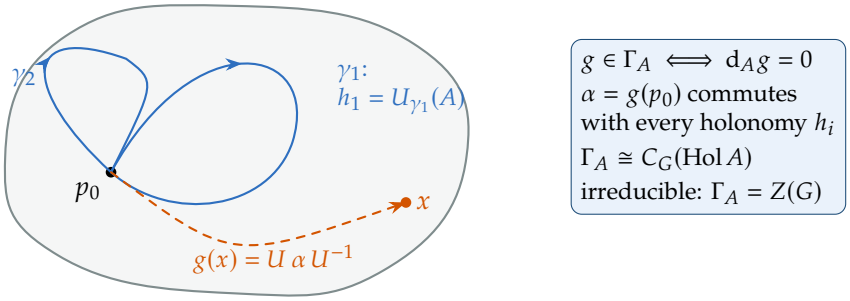


Figure 8.40: Why stabilisers are governed by holonomy. A gauge transformation g fixing A is covariantly constant, so it is determined by its value α at the base point: its value at x is obtained by parallel transport (orange). Transporting around a loop (blue) must bring α back to itself, so α commutes with every holonomy h_i. If the holonomies generate all of G, only central α survive.

With irreducibility in hand we can finish Nash's argument for (8.216). If A is irreducible, its stabiliser in $\mathcal{G}$ is $Z(G)$, and its image in $\mathrm{Ad}\,G = G/Z(G)$ is trivial. So $\mathrm{Ad}\,G$ acts *freely* on the irreducible part of $\mathcal{A}_k/\mathcal{G}_0$, and the quotient (8.216) has no fixed-point singularities. Nash's summary sentence, "a connection with a non-trivial stability group must be reducible", is the contrapositive.

8.7.4 Why reducible connections make singularities

What goes wrong at a reducible connection? The quotient of a space by a group acting with fixed points is not a manifold near the fixed points. The simplest example shows the typical shape: a cone.

Worked example 8.68: The plane divided by rotations is a half-line

Let $\mathrm{U}(1)$ act on $\mathbb{C} = \mathbb{R}^2$ by $z \mapsto e^{i\theta}z$. Every $z \neq 0$ has trivial stabiliser and its orbit is a circle of radius $|z|$; the origin is fixed by the whole group. The orbit space is parametrised by $r = |z|$:

$$\mathbb{C}/\mathrm{U}(1) \cong [0, \infty).$$

Away from $r = 0$ it is a 1-manifold; at $r = 0$ it has a boundary point, which is where the fixed point went. In higher dimensions: $\mathrm{U}(1)$ acting on $\mathbb{C}^n$ by $z \mapsto e^{i\theta}z$ has orbit space

$$\mathbb{C}^n/\mathrm{U}(1) \cong \{t\,w : t \geq 0,\ [w] \in \mathbb{C}^n\setminus 0/\mathrm{U}(1)\} = \text{cone on } S^{2n-1}/\mathrm{U}(1) = \text{cone on } \mathbb{CP}^{n-1}.$$

The link of the singular point (the orbit space of a small sphere around it) is $\mathbb{CP}^{n-1}$ instead of a sphere S^{2n-2}; so the orbit space is not a manifold at the vertex unless $n = 1$.

The same thing happens in $\mathcal{A}/\mathcal{G}$. Near a reducible connection A with stabiliser Γ_A, the slice construction of Section 8.6.4 still works, but the slice is acted on by Γ_A (the gauge transformations that fix A also move points of the slice around). Locally the moduli space looks like

$$\mathcal{M}_k \;\approx\; \frac{\{\text{solutions in the slice through } A\}}{\Gamma_A}.$$

For a U(1)-reducible SU(2) instanton the solutions near A form (in good cases) a copy of $\mathbb{C}^n$ on which $\Gamma_A = \mathrm{U}(1)$ acts with weight one, and $\mathcal{M}_k$ has a cone singularity with link $\mathbb{CP}^{n-1}$. In the case used in Donaldson's theorem ($k = 1$, a 5-dimensional moduli space, $n = 3$), the neighbourhood of each reducible is a *cone on* $\mathbb{CP}^2$ (Section 8.11).

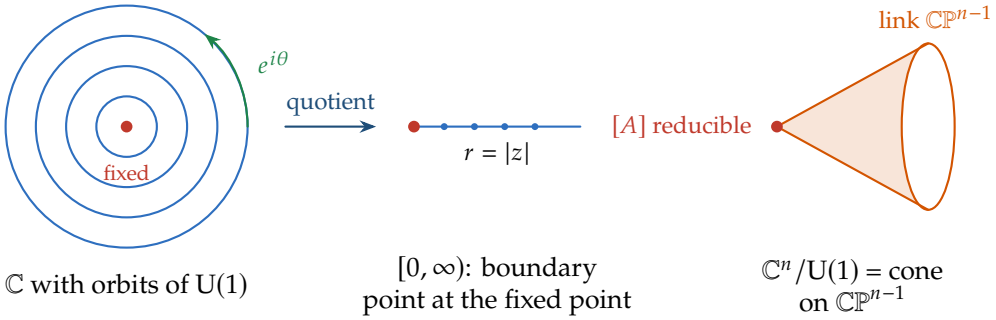


Figure 8.41: Fixed points of a group action produce singularities in the quotient. Left and centre: the rotations of the plane fix only the origin, and the orbit space is a half-line whose end point is the image of the fixed point. Right: in higher dimensions the quotient of $\mathbb{C}^n$ by U(1) is a cone whose cross-section is $\mathbb{CP}^{n-1}$. A reducible instanton, whose stabiliser is U(1), sits at the vertex of such a cone in $\mathcal{M}_k$.

There is also an infinitesimal symptom. The dimension count of the next section uses $h_0 = \dim \operatorname{Ker} \mathrm{d}_A$ on $\Omega^0(\mathrm{ad}\, P)$, the number of infinitesimal stabilisers. At an irreducible connection $h_0 = 0$. At a reducible one $h_0 > 0$, the index formula no longer gives the dimension of the moduli space, and the "expected" dimension is off by exactly the dimension of the stabiliser we have divided by (Example 8.84).

So Nash proceeds with two conventions:

1. $\mathcal{M}_k$ denotes the space of gauge equivalence classes of *irreducible* self-dual connections;
2. if we do not divide by Ad G, we obtain the framed space $\mathcal{M}'_k$, which is fibred over $\mathcal{M}_k$ with fibre Ad G (Figure 8.42).

The framed space has a bonus: since $\mathcal{G}_0$ acts freely on *all* connections (Section 8.7.3), $\mathcal{M}'_k$ has no fixed-point singularities even at reducible points. Over a reducible $[A]$ the fibre is not the full Ad G but Ad G/Γ_A.

8.7.5 Reducible SU(2) instantons and the intersection form

Nash returns to reducibility at the end of the dimension count (his (8.135)–(8.136)). We bring the discussion here, where it belongs logically. The question is: *on which 4-manifolds can an* SU(2) *instanton be reducible?*

For $G = \mathrm{SU}(2)$, a proper subgroup that can be a holonomy group of a non-flat connection is (up to conjugation) the circle $\mathrm{U}(1) = \{\mathrm{diag}(e^{i\theta}, e^{-i\theta})\}$; its centraliser is itself. Holonomy in this circle means that parallel transport preserves the two eigenlines, so the rank-2 bundle E splits:

$$E = L \oplus L^{-1}, \tag{8.221}$$

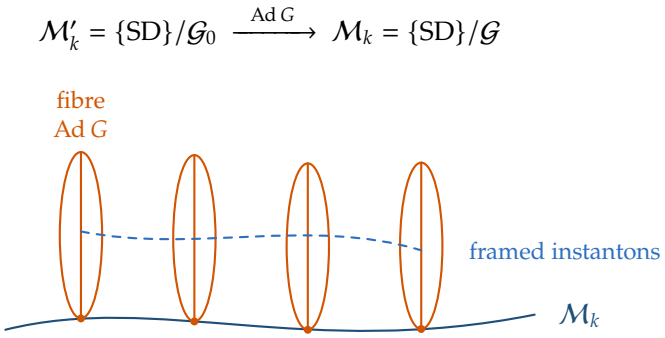


Figure 8.42: The framed moduli space $\mathcal{M}'_k$ is fibred over the moduli space $\mathcal{M}_k$. The fibre over an irreducible instanton is Ad G: all the ways of rotating the instanton in colour space relative to a fixed frame at the base point. For SU(2), $k = 1$ on S^4, the base is the 5-dimensional space of centres and sizes and the fibre is SO(3).

with L a complex line bundle carrying a U(1) connection α and L^{-1} carrying $-\alpha$ (this is Example 8.67(b)). The two lines must be inverse because $\det E$ is trivial for SU(2).

The topology of E is now determined by the topology of L. The total Chern class is multiplicative on direct sums (Chapter 4): writing $e = c_1(L) \in H^2(M;\mathbb{Z})$,

$$\begin{aligned} 1 + c_1(E) + c_2(E) &\overset{(1)}{=} c(L \oplus L^{-1}) \\ &\overset{(2)}{=} c(L)\, c(L^{-1}) \\ &\overset{(3)}{=} (1+e)(1-e) \\ &\overset{(4)}{=} 1 - e^2. \end{aligned}$$

(1) Definition of the total Chern class of E.
(2) Whitney product formula: $c(V \oplus W) = c(V)c(W)$.
(3) For a line bundle $c(L) = 1 + c_1(L)$, and $c_1(L^{-1}) = -c_1(L)$.
(4) Multiply out; $e \cdot e = e \cup e$ is a 4-dimensional class.

Comparing degrees: $c_1(E) = 0$ (as it must be for SU(2)) and $c_2(E) = -e^2$. So

$$k = -c_2(E)[M] = (e \cup e)[M] = q(e,e), \tag{8.222}$$

where q is the **intersection form** of M (Chapter 1): the symmetric bilinear form $q(a,b) = \int_M a \wedge b$ on $H^2(M;\mathbb{Z})$ (for de Rham representatives). Nash writes this $c_2(E) = -c_1(L) \cup c_1(L) = -q(c_1(L), c_1(L)) = -k$.

Key idea: Reducible SU(2) instantons live on classes with $q(e,e) = k$

A reducible charge-k SU(2) connection has $E = L \oplus L^{-1}$ with $e = c_1(L)$ satisfying $q(e,e) = k$. In particular $e \neq 0$, so we need $H^2(M;\mathbb{Z}) \neq 0$. If the connection is also self-dual, the curvature of L is a self-dual harmonic 2-form representing e (up to a factor $2\pi i$), so e lies in the positive part of the intersection form.

Worked example 8.69: No reducible instantons on S^4; some on $\mathbb{CP}^2$

S^4. $H^2(S^4;\mathbb{Z}) = 0$, so there is no non-trivial line bundle and $e = 0$, giving $k = 0$. Every instanton with $k \neq 0$ on S^4 is irreducible. (This is also visible directly: BPST is irreducible by Example 8.67(c).)

$\mathbb{CP}^2$. $H^2(\mathbb{CP}^2;\mathbb{Z}) = \mathbb{Z}$, generated by the hyperplane class h (the Poincaré dual of a line $\mathbb{CP}^1 \subset \mathbb{CP}^2$), with $q(h,h) = 1$: two lines meet in one point. A line bundle is $L = \mathcal{O}(n)$ with $e = nh$, and

$$k = q(nh, nh) = n^2.$$

So reducible SU(2) instantons can exist on $\mathbb{CP}^2$ exactly in the sectors $k = 1, 4, 9, \ldots$. For $k = 1$ the candidates are $n = \pm 1$, and these give the same reducible connection ($L \oplus L^{-1}$ and $L^{-1} \oplus L$ are exchanged by the Weyl reflection $\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \in \mathrm{SU}(2)$). The Fubini–Study metric has a self-dual harmonic 2-form (the Kähler form, with the complex orientation), which is the curvature of such a reducible instanton. So on $\mathbb{CP}^2$ the $k = 1$ moduli space contains exactly one reducible point. This is precisely the case Nash says "is not similarly excluded".

$S^2 \times S^2$. $H^2 = \mathbb{Z}^2$ with $q = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ (the two spheres meet once, and each has self-intersection 0). For $e = (m,n)$, $q(e,e) = 2mn$, so k must be even: reducibles appear only in the even sectors.

Try it: Counting reducibles through the intersection form

On the connected sum $\mathbb{CP}^2 \# \mathbb{CP}^2$, $H^2 = \mathbb{Z}^2$ with $q = \mathrm{diag}(1,1)$. How many reducible points (counted as pairs $\pm e$) does the $k = 1$ moduli space have? *Answer:* $q(e,e) = m^2 + n^2 = 1$ has solutions $\pm(1,0)$ and $\pm(0,1)$: two pairs, hence two reducible points. For the diagonal form $\mathrm{diag}(1,\ldots,1)$ of rank r there are exactly r pairs. This count is the engine of Donaldson's theorem (Section 8.11).

Watch out: Irreducible is not automatic for larger groups

For SU(2) the only way to be reducible is to split off a line bundle, and on S^4 this is impossible. For a big group G, a self-dual connection may have its holonomy in a proper subgroup $H \subset G$ for purely group theoretic reasons, even on S^4: e.g. the BPST instanton embedded in the upper-left 2×2 block of SU(3) has holonomy in $\mathrm{SU}(2) \subset \mathrm{SU}(3)$ and commutes with $\mathrm{diag}(e^{i\theta}, e^{i\theta}, e^{-2i\theta})$. This is why Nash's table of dimensions (8.137) comes with an "irreducibility restriction" column (Section 8.9.8).

Reading Nash (pp. 231–233 and 240, (8.86)–(8.92), (8.135)–(8.136))

Nash defines $\mathcal{M}_k \subset \mathcal{A}_k/\mathcal{G}_k$ ((8.86)–(8.87)) and then notes, citing Atiyah and Jones, that "there is a remaining invariance": change of base point $p_0 \in P_a$, i.e. the residual action of $G = \mathcal{G}/\mathcal{G}_0$ of (8.214). The centre acts trivially, so he uses $\mathrm{Ad}\, G = G/Z(G)$ and restricts to irreducible connections, getting (8.88), our (8.216). (8.89) is our (8.218); (8.90) is the statement that stabilisers commute with holonomy, our (8.219). (8.91)–(8.92) are the infinitesimal test (8.220) and the proof that $\exp(tf)$ then fixes A, which we did exactly rather than as a series. His $\mathcal{M}'_k$ "fibred over $\mathcal{M}_k$ with fibre Ad G" is (8.215). The remark on reducible SU(2) connections, (8.135)–(8.136), is Section 8.7.5.

Recap

- $\mathcal{M}_k$ = self-dual connections of charge k modulo gauge. The BPST family gives 5 parameters (centre, size); colour orientation is a gauge transformation, and for BPST it equals a spatial rotation.
- Gauge group $\mathcal{G}$ = based group $\mathcal{G}_0$ times constants G. The framed space $\mathcal{M}'_k = \{\text{SD}\}/\mathcal{G}_0$ keeps the colour frame at the base point; $\mathcal{M}_k = \mathcal{M}'_k/\operatorname{Ad} G$. For $k = 1$ on S^4: $\dim \mathcal{M}'_1 = 8$, $\dim \mathcal{M}_1 = 5$.
- Stabiliser Γ_A = covariantly constant gauge transformations $\cong$ centraliser of the holonomy group. The centre $Z(G)$ always fixes A; irreducible means nothing else does. Test: no non-zero f with $\mathrm{d}_A f = 0$; sufficient: the $F_{\mu\nu}$ at a point generate $\mathfrak{g}$.
- Reducible connections give cone singularities (link $\mathbb{CP}^{n-1}$ for a U(1) stabiliser). $\mathcal{G}_0$ acts freely, so $\mathcal{M}'_k$ avoids fixed points.
- A reducible SU(2) instanton is $L \oplus L^{-1}$ with $q(c_1 L, c_1 L) = k$: none on S^4; one for $k = 1$ on $\mathbb{CP}^2$; r for the diagonal form of rank r.

8.8 The deformation complex

How many parameters does the general instanton have? Nash answers this the way one computes the dimension of any solution space: work *infinitesimally*. The dimension of a smooth space equals the dimension of its tangent space at any point. So take one instanton A, perturb it to $A + ta$, and ask which perturbations a keep it self-dual (to first order) and are not merely gauge transformations. Those a form the tangent space $T_{[A]}\mathcal{M}_k$. The advantage, as Nash says, is that the dimension of this tangent space can be computed by the index theorem.

The same strategy computes the dimension of the moduli space $\mathcal{M}_p$ of Riemann surfaces of genus p (Chapter 5): there one linearises the condition of being a complex structure and quotients by diffeomorphisms, and Riemann–Roch gives $\dim_{\mathbb{C}} \mathcal{M}_p = 3p - 3$. We compare the two at the end of the section.

The physicist knows this calculation as the counting of **zero modes**. Expand the action around the instanton, $S(A + a) = S(A) + \frac{1}{2}\langle a, \mathcal{K}a\rangle + \dots$; after gauge fixing, the zero eigenvectors of the fluctuation operator $\mathcal{K}$ are exactly the tangent vectors to the moduli space, and their number is the number of collective coordinates.

8.8.1 Tangent vectors: the two requirements

Let $A_t = A + ta$ be a one-parameter family of connections through A. Its velocity $a = \frac{\mathrm{d}}{\mathrm{d}t}A_t|_{t=0}$ is a tangent vector to the space of connections,

$$a \in T_A\mathcal{A}_k = \Omega^1(M, \operatorname{ad} P) \qquad \text{(Nash's (8.93))}.$$

Write $[A]$ for the point of $\mathcal{M}_k$ represented by A. We want to produce from a an element of $T_{[A]}\mathcal{M}_k$ (Nash's (8.94)). Two things must be done:

1. the family must stay self-dual: a must solve the *linearised* self-duality equation;
2. we must throw away the a's that point along the gauge orbit, i.e. those in $T_A(\mathcal{G}\cdot A) = \operatorname{Im} \mathrm{d}_A$ ((8.210)).

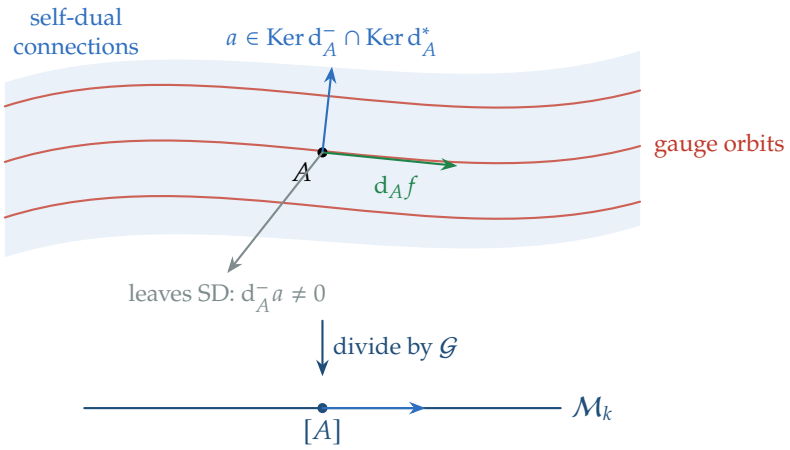


Figure 8.43: The tangent space to the moduli space. Inside the space of connections, the self-dual ones form a (blue) sheet, made of whole gauge orbits (red). A perturbation a of A must stay in the sheet to first order ($\mathrm{d}_A^- a = 0$) and should not point along the orbit. Taking the component orthogonal to the orbit ($\mathrm{d}_A^* a = 0$) gives one representative of each tangent vector of $\mathcal{M}_k$ at $[A]$.

8.8.2 Linearising the self-duality equation

Self-duality says the anti-self-dual part of F vanishes. With the projector $\pi_- \omega = \frac{1}{2}(\omega - \star\omega)$ of (8.180) (Nash's (8.95)), the equation is

$$\pi_- F(A) = 0 \qquad \text{(Nash's (8.96))}. \tag{8.223}$$

Substitute $A + ta$ (Nash's (8.97)) and expand the curvature:

$$\begin{aligned} F(A+ta) &\overset{(1)}{=} \mathrm{d}(A+ta) + (A+ta)\wedge(A+ta) \\ &\overset{(2)}{=} \mathrm{d}A + A\wedge A + t\big(\mathrm{d}a + A\wedge a + a\wedge A\big) + t^2\, a\wedge a \\ &\overset{(3)}{=} F(A) + t\,\mathrm{d}_A a + t^2\, a\wedge a. \end{aligned} \tag{8.224}$$

(1) Definition $F = \mathrm{d}A + A\wedge A$.
(2) Expand; the order of the matrix-valued forms in $A\wedge a$ and $a\wedge A$ must be kept.
(3) $\mathrm{d}_A a = \mathrm{d}a + A\wedge a + a\wedge A$ is (8.164) for $p = 1$.

This is Nash's (8.98) (and our (8.168)). If A and every $A + ta$ are self-dual,

$$\begin{aligned} 0 &\overset{(1)}{=} \pi_- F(A+ta) \\ &\overset{(2)}{=} \pi_- F(A) + t\,\pi_-(\mathrm{d}_A a) + t^2\,\pi_-(a\wedge a) \\ &\overset{(3)}{=} t\,\pi_-(\mathrm{d}_A a) + t^2\,\pi_-(a\wedge a). \end{aligned}$$

(1) $A + ta$ is self-dual.
(2) (8.224) and linearity of π_-.
(3) A is self-dual.

Divide by t and let $t \to 0$ (Nash's (8.99)–(8.100)):

Working definition 8.32 (the linearised self-duality operator)

$$\mathrm{d}_A^- := \pi_- \circ \mathrm{d}_A : \ \Omega^1(M, \mathrm{ad}\, P) \to \Omega^2_-(M, \mathrm{ad}\, P), \qquad \text{tangent vectors satisfy} \qquad \mathrm{d}_A^- a = 0. \tag{8.225}$$

(Nash's (8.101).) *To compute:* form the 2-form $\mathrm{d}_A a$, with components $(\mathrm{d}_A a)_{\mu\nu} = D_\mu a_\nu - D_\nu a_\mu$, and take its anti-self-dual part: three components per colour [for each of the $\mathfrak{su}(2)$ directions e_1, e_2, e_3 separately], along $\omega_1^-, \omega_2^-, \omega_3^-$.

Worked example 8.70: The linearised operator in $\vec{E}, \vec{B}$ language

Take $G = \mathrm{U}(1)$ (or one colour component) and $A = 0$ on $\mathbb{R}^4$, so $\mathrm{d}_A a = \mathrm{d}a = f$ with $f_{\mu\nu} = \partial_\mu a_\nu - \partial_\nu a_\mu$. With $E_j = f_{j4}$ and $B_j = \frac{1}{2}\varepsilon_{jkl} f_{kl}$ as in Section 8.4,

$$\begin{aligned} f &\overset{(1)}{=} \sum_j E_j \, \mathrm{d}x^{j4} + \sum_j B_j \, \tfrac{1}{2}\varepsilon_{jkl} \mathrm{d}x^{kl} \\ &\overset{(2)}{=} \sum_j \Big(E_j \, \tfrac{1}{2}(\omega_j^+ - \omega_j^-) + B_j \, \tfrac{1}{2}(\omega_j^+ + \omega_j^-) \Big) \\ &\overset{(3)}{=} \sum_j \tfrac{1}{2}(E_j + B_j)\, \omega_j^+ + \sum_j \tfrac{1}{2}(B_j - E_j)\, \omega_j^- . \end{aligned}$$

(1) The $\vec{E}, \vec{B}$ decomposition of a 2-form (Section 8.4).
(2) Invert the basis (8.181): $\omega_j^\pm = \pm \mathrm{d}x^{j4} + \frac{1}{2}\varepsilon_{jkl}\mathrm{d}x^{kl}$, so $\mathrm{d}x^{j4} = \frac{1}{2}(\omega_j^+ - \omega_j^-)$ and $\frac{1}{2}\varepsilon_{jkl}\mathrm{d}x^{kl} = \frac{1}{2}(\omega_j^+ + \omega_j^-)$.
(3) Collect.

So $\mathrm{d}^- a = \frac{1}{2}\sum_j (B_j - E_j)\, \omega_j^-$, and the linearised self-duality equation is $\vec{E} = \vec{B}$ for the perturbation: three first-order equations,

$$\partial_j a_4 - \partial_4 a_j = (\nabla \times \vec{a})_j, \qquad j = 1, 2, 3,$$

for four unknown functions $a_1, \ldots, a_4$. The deficit of one equation is made up by the gauge condition, as we shall see.

8.8.3 Removing the gauge directions

The second requirement is to discard perturbations of the form $a = \mathrm{d}_A f$. First, these *do* solve the linearised equation. This is forced by gauge invariance, and it is the key algebraic fact of the section:

$$\mathrm{d}_A^- \circ \mathrm{d}_A = 0 \qquad \text{on } \Omega^0(M, \mathrm{ad}\, P) \qquad \text{(Nash's (8.105))}. \tag{8.226}$$

We check it twice below (Section 8.8.4). Granting it, the space of genuine deformations is the quotient

$$T_{[A]}\mathcal{M}_k \ \cong \ \frac{\mathrm{Ker}\, \mathrm{d}_A^-}{\mathrm{Im}\, \mathrm{d}_A} \qquad \text{(Nash's (8.102))}: \tag{8.227}$$

solutions of the linearised equation, where two solutions are identified if they differ by an infinitesimal gauge transformation.

A quotient is awkward to compute with. The slice of Section 8.6.4 turns it into a subspace:

Key idea: Harmonic representatives of deformations

Each class in $\operatorname{Ker} \mathrm{d}_A^- / \operatorname{Im} \mathrm{d}_A$ has exactly one representative in the background gauge. Hence

$$T_{[A]}\mathcal{M}_k \;\cong\; \operatorname{Ker} \mathrm{d}_A^- \cap \operatorname{Ker} \mathrm{d}_A^* = \{a : \; \mathrm{d}_A^- a = 0, \; \mathrm{d}_A^* a = 0\}. \tag{8.228}$$

These are the *zero modes* of the instanton in background gauge.

Why: let $\mathrm{d}_A^- a = 0$. Split $a = \mathrm{d}_A f + a^\perp$ with $\mathrm{d}_A^* a^\perp = 0$ ((8.211)). Then

$$\begin{aligned} \mathrm{d}_A^- a^\perp &\overset{(1)}{=} \mathrm{d}_A^- a - \mathrm{d}_A^- \mathrm{d}_A f \\ &\overset{(2)}{=} 0 - 0, \end{aligned}$$

(1) Linearity of d_A^- and $a^\perp = a - \mathrm{d}_A f$.
(2) $a \in \operatorname{Ker} \mathrm{d}_A^-$, and (8.226).

so $a^\perp$ is in the intersection and represents the same class as a. It is unique: if $a_1^\perp, a_2^\perp$ are two such representatives, their difference is both in $\operatorname{Im} \mathrm{d}_A$ and in $\operatorname{Ker} \mathrm{d}_A^* = (\operatorname{Im} \mathrm{d}_A)^\perp$, so it is orthogonal to itself and vanishes.

8.8.4 The complex and the check $\mathrm{d}_A^- \circ \mathrm{d}_A = 0$

Let us name the spaces precisely. The perturbation a is a Lie-algebra-valued 1-form that transforms by conjugation, $a \mapsto g^{-1} a g$ (Section 8.3.1). So it is a section of the bundle

$$\operatorname{ad} P \otimes \Lambda^1 T^* M \qquad \text{(Nash's (8.103))} : \tag{8.229}$$

the first factor makes a transform correctly under gauge transformations, the second makes it a 1-form. Write $\Omega^i(M, \operatorname{ad} P) = \Gamma(M, \operatorname{ad} P \otimes \Lambda^i T^* M)$, and $\Omega^2_-(M, \operatorname{ad} P)$ for the anti-self-dual 2-forms. The three operators we have met fit into a sequence:

Key idea: The instanton deformation complex (Nash's (8.104))

$$0 \longrightarrow \Omega^0(M, \operatorname{ad} P) \xrightarrow{\mathrm{d}_A} \Omega^1(M, \operatorname{ad} P) \xrightarrow{\mathrm{d}_A^-} \Omega^2_-(M, \operatorname{ad} P) \longrightarrow 0. \tag{8.230}$$

Reading from left to right: infinitesimal gauge transformations f; perturbations a of the connection; perturbations of the ASD part of the curvature. It is a **complex**: two successive maps compose to zero, (8.226). It is **elliptic**: at the level of symbols, the sequence is exact (Example 8.71).

Check 1: the gauge argument (Nash's (8.107)). $\mathrm{d}_A f$ is the change in A under an infinitesimal gauge transformation e^{tf}, by (8.210). The curvature then changes by conjugation: $F \mapsto e^{-tf} F e^{tf} = F + t[F, f] + O(t^2)$. By (8.224) the same change is $t\, \mathrm{d}_A(\mathrm{d}_A f) + O(t^2)$. A gauge transform of a self-dual curvature is self-dual (because $\star$ acts on the form indices and conjugation on the matrix indices, so they commute). Hence

$$\begin{aligned} 0 &\overset{(1)}{=} \pi_-\left(e^{-tf} F e^{tf}\right) \\ &\overset{(2)}{=} \pi_- F + t\, \pi_-(\mathrm{d}_A \mathrm{d}_A f) + O(t^2) \\ &\overset{(3)}{=} t\, \mathrm{d}_A^- \mathrm{d}_A f + O(t^2), \end{aligned}$$

(1) The gauge-transformed field is still self-dual.

(2) The first-order change of F is $t\,\mathrm{d}_A\mathrm{d}_A f$.
(3) $\pi_- F = 0$, and the definition of d_A^-.

and the coefficient of t must vanish.

Check 2: explicitly (Nash's (8.108)). We use Nash's footnote formula $\mathrm{d}_A\omega = \mathrm{d}\omega + A\wedge\omega + (-1)^{p+1}\omega\wedge A$ for an $\mathrm{ad}\,P$-valued p-form, which is our (8.164).

$$\begin{aligned}
\mathrm{d}_A\mathrm{d}_A f &\overset{(1)}{=} \mathrm{d}(\mathrm{d}_A f) + A\wedge \mathrm{d}_A f + \mathrm{d}_A f\wedge A \\
&\overset{(2)}{=} \mathrm{d}(\mathrm{d}f + Af - fA) + A\wedge(\mathrm{d}f + Af - fA) + (\mathrm{d}f + Af - fA)\wedge A \\
&\overset{(3)}{=} \big(\mathrm{d}A\,f - A\wedge \mathrm{d}f - \mathrm{d}f\wedge A - f\,\mathrm{d}A\big) \\
&\qquad + \big(A\wedge \mathrm{d}f + A\wedge A\,f - A\,f\wedge A\big) + \big(\mathrm{d}f\wedge A + A\,f\wedge A - f\,A\wedge A\big) \\
&\overset{(4)}{=} (\mathrm{d}A + A\wedge A)\,f - f\,(\mathrm{d}A + A\wedge A) \\
&\overset{(5)}{=} Ff - fF = [F, f].
\end{aligned}$$

(1) The formula with $p=1$: $(-1)^{p+1} = +1$.
(2) $\mathrm{d}_A f = \mathrm{d}f + [A, f] = \mathrm{d}f + Af - fA$ (the case $p=0$).
(3) $\mathrm{d}\mathrm{d}f = 0$; Leibniz: $\mathrm{d}(Af) = \mathrm{d}A\,f - A\wedge\mathrm{d}f$ and $\mathrm{d}(fA) = \mathrm{d}f\wedge A + f\,\mathrm{d}A$ (a sign whenever d passes the 1-form A).
(4) Cancel in pairs: $\pm A\wedge\mathrm{d}f$, $\pm\mathrm{d}f\wedge A$, $\pm Af\wedge A$.
(5) Definition of F.

Now project. Since f is a 0-form, π_- acts only on the form part of F, so $\pi_-(Ff - fF) = (\pi_- F)f - f(\pi_- F)$, and

$$\mathrm{d}_A^-\mathrm{d}_A f = [\pi_- F, f] = 0 \qquad \text{because } \pi_- F = 0.$$

The identity $\mathrm{d}_A\mathrm{d}_A = [F, \cdot\,]$ is the covariant version of $\mathrm{d}\mathrm{d} = 0$: *the curvature is the obstruction to* $\mathrm{d}_A^2 = 0$. So the complex (8.230) closes precisely because $F^- = 0$. If we had used the full d_A into all 2-forms, $\mathrm{d}_A\mathrm{d}_A f = [F, f] \neq 0$ and we would not have a complex. This is why instantons (and flat connections) are special.

Worked example 8.71: Ellipticity of the complex: exactness of the symbol sequence

An elliptic complex is one whose symbol sequence is exact for every $\xi \neq 0$ (Chapter 2). Take $\xi = (0,0,0,1)$, i.e. the covector $\mathrm{d}x^4$, and one colour component. The symbols are obtained by replacing d with $i\xi\wedge$:

$$\sigma_0(\xi)f = i\,f\,\mathrm{d}x^4, \qquad \sigma_1(\xi)a = \pi_-\big(i\,\mathrm{d}x^4\wedge a\big).$$

Compute σ_1 on $a = a_\mu \mathrm{d}x^\mu$:

$$\begin{aligned}
\pi_-\big(\mathrm{d}x^4\wedge a\big) &\overset{(1)}{=} \pi_-\Big(\sum_{j=1}^{3} a_j\,\mathrm{d}x^4\wedge\mathrm{d}x^j\Big) \\
&\overset{(2)}{=} -\sum_j a_j\,\pi_-\big(\mathrm{d}x^{j4}\big) \\
&\overset{(3)}{=} -\sum_j a_j\,\pi_-\big(\tfrac12(\omega_j^+ - \omega_j^-)\big) \\
&\overset{(4)}{=} \tfrac12\sum_j a_j\,\omega_j^-.
\end{aligned}$$

(1) $\mathrm{d}x^4 \wedge \mathrm{d}x^4 = 0$ kills the a_4 term.
(2) $\mathrm{d}x^4 \wedge \mathrm{d}x^j = -\mathrm{d}x^{j4}$.
(3) $\mathrm{d}x^{j4} = \tfrac{1}{2}(\omega_j^+ - \omega_j^-)$, as in Example 8.70.
(4) $\pi_- \omega_j^+ = 0$, $\pi_- \omega_j^- = \omega_j^-$.

In the bases $\{1\}$, $\{\mathrm{d}x^1, \dots, \mathrm{d}x^4\}$, $\{\omega_1^-, \omega_2^-, \omega_3^-\}$ the symbol sequence is

$$0 \to \mathbb{C} \xrightarrow{\sigma_0 = i\begin{pmatrix}0\\0\\0\\1\end{pmatrix}} \mathbb{C}^4 \xrightarrow{\sigma_1 = \frac{i}{2}\begin{pmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\end{pmatrix}} \mathbb{C}^3 \to 0.$$

- $\sigma_1 \sigma_0 = 0$: the complex property at the symbol level.
- $\operatorname{Ker}\sigma_0 = 0$; $\operatorname{Ker}\sigma_1 = \operatorname{span}(\mathrm{d}x^4) = \operatorname{Im}\sigma_0$; $\operatorname{Im}\sigma_1 = \mathbb{C}^3$. So the sequence is exact everywhere.
- Dimension check: $1 - 4 + 3 = 0$, as it must be for an exact sequence.

Physically: of the four polarisations of a plane wave, one is longitudinal (pure gauge, $\operatorname{Im}\sigma_0$) and three are hit by the linearised equation. None is missed. That is the precise sense in which "self-duality plus gauge fixing" is a determined elliptic system. For SU(2) every space is tensored with $\mathfrak{su}(2) \cong \mathbb{R}^3$: ranks $3 \to 12 \to 9$.

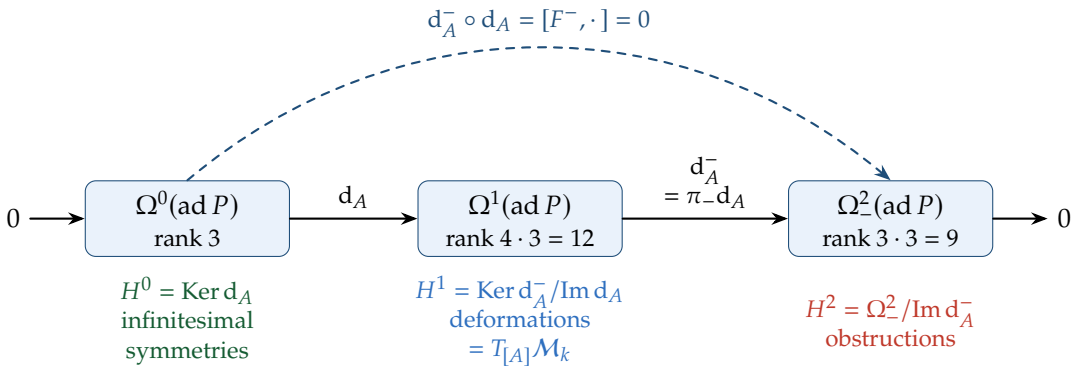


Figure 8.44: The instanton deformation complex for $G = \mathrm{SU}(2)$, with the ranks of the bundles (per point of M). Each cohomology group has a meaning: H^0 counts gauge transformations fixing A (zero when A is irreducible), H^1 is the tangent space to the moduli space, H^2 measures the failure of first-order deformations to extend (zero under the hypotheses of Atiyah–Hitchin–Singer). The composite of the two maps is the anti-self-dual curvature, so the sequence is a complex exactly when A is self-dual.

8.8.5 Cohomology of the complex and the index

Because two successive maps compose to zero, the image of each map lies in the kernel of the next, and we may form quotients. Denote the complex (8.230) by $\Omega_-(M, \mathrm{ad}\,P)$ and write $\mathrm{d}_A^{(p)}$ for d_A acting on $\mathrm{ad}\,P \otimes \Lambda^p T^*M$.

Working definition 8.33 (cohomology of the deformation complex (Nash's (8.109)))

$$H^0 = \operatorname{Ker}\mathrm{d}_A^{(0)}, \qquad h_0 = \dim H^0,$$

$$H^1 = \operatorname{Ker} \mathrm{d}_A^- / \operatorname{Im} \mathrm{d}_A^{(0)}, \qquad h_1 = \dim H^1,$$
$$H^2 = \Omega_-^2(M, \operatorname{ad} P) / \operatorname{Im} \mathrm{d}_A^-, \qquad h_2 = \dim H^2.$$

(Nash writes $H^2 = \operatorname{Ker} \pi_+ / \operatorname{Im} \mathrm{d}_A^-$: the next map in the complex is the zero map $\pi_+ : \Omega_-^2 \to 0$, whose kernel is all of Ω_-^2.) *To compute* each h_i: use harmonic representatives [in each class, the one element L^2-orthogonal to the image of the previous map]. $H^0 = \operatorname{Ker} \mathrm{d}_A$, $H^1 \cong \operatorname{Ker} \mathrm{d}_A^- \cap \operatorname{Ker} \mathrm{d}_A^*$ ((8.228)), and $H^2 \cong \operatorname{Ker}(\mathrm{d}_A^-)^*$ [$(\mathrm{d}_A^-)^*$ is the adjoint of d_A^- for the L^2 inner products] (the orthogonal complement of the image).

We want $h_1 = \dim T_{[A]}\mathcal{M}_k = \dim \mathcal{M}_k$ (Nash's (8.110)). The index theorem does not give h_1 directly; it gives the alternating sum

$$\text{index(complex)} = h_0 - h_1 + h_2 \qquad \text{(Nash's (8.111))}. \tag{8.231}$$

Why an alternating sum? Because the alternating sum is stable: it does not change when the operators are deformed, whereas the individual h_i can jump. The finite-dimensional version shows the mechanism.

Worked example 8.72: Euler characteristic of a finite-dimensional complex

Let $0 \to V_0 \xrightarrow{d_0} V_1 \xrightarrow{d_1} V_2 \to 0$ be finite-dimensional with $d_1 d_0 = 0$. Write $n_i = \dim V_i$, $r_0 = \operatorname{rank} d_0$, $r_1 = \operatorname{rank} d_1$.

$$\begin{aligned} h_0 - h_1 + h_2 &\overset{(1)}{=} (n_0 - r_0) - \big((n_1 - r_1) - r_0\big) + (n_2 - r_1) \\ &\overset{(2)}{=} n_0 - n_1 + n_2. \end{aligned}$$

(1) Rank–nullity: $h_0 = \dim \operatorname{Ker} d_0 = n_0 - r_0$; $h_1 = \dim \operatorname{Ker} d_1 - \dim \operatorname{Im} d_0 = (n_1 - r_1) - r_0$; $h_2 = n_2 - \dim \operatorname{Im} d_1 = n_2 - r_1$.

(2) The ranks cancel.

The alternating sum of cohomology dimensions depends only on the dimensions of the spaces, not on the maps. Concretely, with $V_0 = \mathbb{R}$, $V_1 = \mathbb{R}^3$, $V_2 = \mathbb{R}$, $d_0 = (1,0,0)^T$ and $d_1 = (0, s, 0)$: for $s \neq 0$ we get $(h_0, h_1, h_2) = (0, 1, 0)$; for $s = 0$ we get $(0, 2, 1)$. The individual numbers jump, but $h_0 - h_1 + h_2 = -1 = 1 - 3 + 1$ in both cases. In infinite dimensions the n_i are infinite, and the index theorem is the tool that computes the stable combination anyway.

The moral for instantons: if we can show $h_0 = 0$ and $h_2 = 0$, then $h_1 = -\text{index}$, and the index theorem gives $\dim \mathcal{M}_k$. We now show both vanish, under hypotheses.

Worked example 8.73: Why H^2 is called the obstruction space: a toy model

Consider solving $\Phi(x) = 0$ for a smooth map $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ near a solution x_0. The linearisation $L = \Phi'(x_0)$ plays the role of d_A^-. If L is onto ($\operatorname{Coker} L = 0$), the implicit function theorem says the solutions near x_0 form a manifold of dimension $\dim \operatorname{Ker} L = n - m$. If L is not onto, first-order solutions may fail to extend. The simplest case: $n = m = 1$, $\Phi(x) = x^2$ at $x_0 = 0$. Then $L = 0$, so $\operatorname{Ker} L = \mathbb{R}$ (one first-order deformation) and $\operatorname{Coker} L = \mathbb{R}$ (one obstruction), but the true solution set is the single point $\{0\}$. The second-order term x^2 is what the cokernel measures. Note that $\dim \operatorname{Ker} L - \dim \operatorname{Coker} L = 0 = n - m$ still gives the "expected" dimension correctly. For instantons the second-order term is $\pi_-(a \wedge a)$ in (8.224), and H^2 is where it lives.

8.8.6 The deformation operator as a quaternionic Cauchy–Riemann operator

It helps to see the deformation problem as a single square system. Combine the linearised equation and the gauge condition:

$$\mathcal{D}_A = \mathrm{d}_A^* \oplus \mathrm{d}_A^- : \ \Omega^1(M, \mathrm{ad}\, P) \to \Omega^0(M, \mathrm{ad}\, P) \oplus \Omega^2_-(M, \mathrm{ad}\, P), \qquad T_{[A]}\mathcal{M}_k \cong \operatorname{Ker} \mathcal{D}_A .$$

Four unknowns per colour, $1 + 3 = 4$ equations per colour. Its adjoint $\mathcal{D}_A^* = \mathrm{d}_A + (\mathrm{d}_A^-)^*$ is exactly the operator D that Nash introduces in (8.114) (Section 8.9.1). On flat $\mathbb{R}^4$ with $A = 0$ the operator $\mathcal{D}$ has a beautiful form in quaternions.

Worked example 8.74: The Fueter operator

Package the four components of $a = a_\mu \mathrm{d}x^\mu$ (one colour) into the quaternion $\bar{a} = a_4 \mathbb{1} - a_k e_k$, with the matrices e_k of Section 8.5.1, and define the **Fueter operator** $\mathfrak{D} = \partial_4 + e_j \partial_j$. Then

$$\begin{aligned} \mathfrak{D}\bar{a} &\overset{(1)}{=} (\partial_4 + e_j\partial_j)(a_4 - e_k a_k) \\ &\overset{(2)}{=} \partial_4 a_4 - e_k \partial_4 a_k + e_j \partial_j a_4 - e_j e_k\, \partial_j a_k \\ &\overset{(3)}{=} \partial_4 a_4 + \partial_k a_k + e_l\big(\partial_l a_4 - \partial_4 a_l - \varepsilon_{jkl}\partial_j a_k\big) \\ &\overset{(4)}{=} \partial_\mu a_\mu + e_l\big(E_l - B_l\big). \end{aligned}$$

(1) Definitions.
(2) Expand; the e's are constant.
(3) $e_j e_k = -\delta_{jk}\mathbb{1} + \varepsilon_{jkl} e_l$ by (8.193); rename indices.
(4) $E_l = f_{l4} = \partial_l a_4 - \partial_4 a_l$ and $B_l = \frac{1}{2}\varepsilon_{lmn} f_{mn} = \varepsilon_{lmn}\partial_m a_n = \varepsilon_{jkl}\partial_j a_k$.

The real part is $\partial_\mu a_\mu = -\mathrm{d}^* a$ (the gauge condition) and the imaginary part is $\vec{E} - \vec{B}$, i.e. $-2\times$ the components of $\mathrm{d}^- a$ along ω_l^- (Example 8.70). So

$$\mathcal{D}a = 0 \iff \mathfrak{D}\bar{a} = 0 :$$

the zero modes of a flat-space instanton problem are the solutions of a quaternionic Cauchy–Riemann equation. The analogy with complex analysis is exact: $\partial_{\bar z} = \frac{1}{2}(\partial_x + i\partial_y)$, and a function of z is holomorphic iff it is killed by $\partial_{\bar z}$.
The Fueter operator is elliptic: its symbol $i(\xi_4 + e_j\xi_j)$ is i times a quaternion of length $|\xi|$, invertible for $\xi \neq 0$. And it squares to the Laplacian. With $\bar{\mathfrak{D}} = \partial_4 - e_j\partial_j$,

$$\begin{aligned} \bar{\mathfrak{D}}\mathfrak{D} &\overset{(1)}{=} \partial_4^2 + (e_j - e_j)\partial_4\partial_j - e_j e_k\, \partial_j\partial_k \\ &\overset{(2)}{=} \partial_4^2 + \delta_{jk}\partial_j\partial_k - \varepsilon_{jkl} e_l\, \partial_j\partial_k \\ &\overset{(3)}{=} \Box. \end{aligned}$$

(1) Expand; the e_j are constants.
(2) The product rule (8.193).
(3) $\partial_j\partial_k$ is symmetric and ε_{jkl} antisymmetric.

This is the quaternionic analogue of $4\partial_z\partial_{\bar z} = \partial_x^2 + \partial_y^2$. It is also the Dirac operator in disguise: in the spinor language of Chapter T4, $\Lambda^1 \otimes \mathbb{C} \cong S^+ \otimes S^-$ and $\Lambda^0 \oplus \Lambda^2_- \cong S^- \otimes S^-$, and $\mathcal{D}$ is the Dirac operator acting on S^+ twisted by S^-. We shall use this identification to understand the index formula (Section 8.9.2).

8.8.7 The vanishing theorems: $h_0 = 0$ and $h_2 = 0$

$h_0 = 0$: irreducibility

$H^0 = \operatorname{Ker} d_A^{(0)}$ is the space of covariantly constant sections f of $\operatorname{ad} P$. By Section 8.7.3 a non-zero such f exists exactly when A is reducible. We restrict to irreducible connections, so $h_0 = 0$. On S^4 with $G = \mathrm{SU}(2)$ this costs nothing (Example 8.69).

$h_2 = 0$: a Weitzenböck positivity argument

H^2 is the orthogonal complement of $\operatorname{Im} d_A^-$ in Ω^2_-, i.e. $\operatorname{Ker}(d_A^-)^*$. To show it is zero, Nash follows Atiyah, Hitchin and Singer: show that the Laplacian on $\Omega^2_-(\operatorname{ad} P)$ built from the complex,

$$\Delta_A^{(2)} = d_A^-(d_A^-)^* + \pi_+\pi_+^* \qquad \text{(Nash's (8.112))}, \tag{8.232}$$

is positive definite. (The term $\pi_+\pi_+^*$ is zero: it is the Laplacian term from the zero map $\Omega^2_- \to 0$ at the end of the complex. It is written only to show the pattern "$dd^* + d^*d$".) If $\Delta_A^{(2)}$ is positive then $\operatorname{Ker}\Delta_A^{(2)} = 0$, and since $\langle \Delta_A^{(2)}\omega, \omega\rangle = \left\|(d_A^-)^*\omega\right\|^2$, also $\operatorname{Ker}(d_A^-)^* = 0$, i.e. $h_2 = 0$.

The tool is a **Weitzenböck formula**: an identity expressing a Laplacian built from d's as a "rough" Laplacian $\nabla^*\nabla$ plus curvature terms with no derivatives. The prototype is the Bochner formula on 1-forms, $(\mathrm{dd}^* + \mathrm{d}^*\mathrm{d})\omega = \nabla^*\nabla\omega + \mathrm{Ric}(\omega)$, which shows that a manifold of positive Ricci curvature has no harmonic 1-forms ($b_1 = 0$). For our operator the result is:

Key idea: Weitzenböck formula on $\Omega^2_-(\operatorname{ad} P)$ (Nash's (8.113))

For a self-dual connection A,

$$\Delta_A^{(2)} = d_A^-(d_A^-)^* = \tfrac{1}{2}\nabla_A^*\nabla_A + \frac{R}{6} - W_-. \tag{8.233}$$

Here ∇_A is the covariant derivative on $\operatorname{ad} P \otimes \Lambda^2_-$ built from A and the Levi-Civita connection, $\nabla_A^*\nabla_A = -D_\mu D_\mu$ is the rough Laplacian (a non-negative operator), R is the scalar curvature of M, and W_- is the anti-self-dual part of the Weyl tensor, acting on Λ^2_- as a symmetric traceless 3×3 matrix. A term $[F^-, \cdot\,]$ also appears in general; it vanishes because A is self-dual.

With it, positivity is one line: for $\omega \in \Omega^2_-(\operatorname{ad} P)$,

$$\begin{aligned}
\left\|(d_A^-)^*\omega\right\|^2 &\overset{(1)}{=} \langle \Delta_A^{(2)}\omega, \omega\rangle \\
&\overset{(2)}{=} \tfrac{1}{2}\left\|\nabla_A\omega\right\|^2 + \int_M \frac{R}{6}\,|\omega|^2 - \langle W_-\omega, \omega\rangle \\
&\overset{(3)}{=} \tfrac{1}{2}\left\|\nabla_A\omega\right\|^2 + \int_M \frac{R}{6}\,|\omega|^2 \\
&\overset{(4)}{>} 0 \qquad \text{unless } \omega = 0.
\end{aligned}$$

(1) Definition of the adjoint, $\langle d_A^-(d_A^-)^*\omega, \omega\rangle = \langle (d_A^-)^*\omega, (d_A^-)^*\omega\rangle$.
(2) The Weitzenböck formula, and $\langle \nabla^*\nabla\omega, \omega\rangle = \|\nabla\omega\|^2$.
(3) Assumption: $W_- = 0$ (M is **self-dual**).
(4) Assumption: $R > 0$ everywhere.

So under the two assumptions $W_- = 0$ and $R > 0$, $h_2 = 0$. Nash, following Atiyah–Hitchin–Singer, makes these assumptions from now on.

Worked example 8.75: The flat-space check of the Weitzenböck formula

On $\mathbb{R}^4$ with $A = 0$ we have $R = 0$, $W_- = 0$ and $\nabla^*\nabla = -\Box$, so (8.233) predicts $\mathrm{d}^-(\mathrm{d}^-)^* = -\frac{1}{2}\Box$ on ASD forms. Let ω be an ASD 2-form, $\star\omega = -\omega$. Then $(\mathrm{d}^-)^*\omega = \mathrm{d}^*\omega$ (since π_- is self-adjoint and $\pi_-\omega = \omega$), and

$$\begin{aligned} \mathrm{d}^*\mathrm{d}\,\omega &\overset{(1)}{=} -\star\,\mathrm{d}\star\,\mathrm{d}\omega \\ &\overset{(2)}{=} \star\,\mathrm{d}\,\mathrm{d}^*\star\,\omega \\ &\overset{(3)}{=} -\star\,(\mathrm{d}\mathrm{d}^*\omega), \end{aligned}$$

(1) $\mathrm{d}^* = -\star\,\mathrm{d}\star$ on forms of every degree in four dimensions ((8.176)).
(2) On the 2-form $\star\omega$: $\mathrm{d}^*\star\omega = -\star\mathrm{d}\star\star\omega = -\star\mathrm{d}\omega$, so $\star\mathrm{d}\omega = -\mathrm{d}^*\star\omega$ and $\mathrm{d}\star\mathrm{d}\omega = -\mathrm{d}\mathrm{d}^*\star\omega$.
(3) $\star\omega = -\omega$.

Apply π_-, using $\pi_-\star = -\pi_-$: $\pi_-(\mathrm{d}^*\mathrm{d}\omega) = \pi_-(\mathrm{d}\mathrm{d}^*\omega)$. Adding,

$$2\,\pi_-(\mathrm{d}\mathrm{d}^*\omega) = \pi_-\big((\mathrm{d}\mathrm{d}^* + \mathrm{d}^*\mathrm{d})\omega\big) = \pi_-(-\Box\omega) = -\Box\omega,$$

because the Hodge Laplacian on flat space is $-\Box$ acting on each component, and it preserves ASD forms. Hence $\mathrm{d}^-(\mathrm{d}^-)^*\omega = \pi_-\mathrm{d}\mathrm{d}^*\omega = -\frac{1}{2}\Box\omega$, as predicted. The factor $\frac{1}{2}$ in Nash's formula is exactly this halving: only half of the Hodge Laplacian is seen on one chirality.

Worked example 8.76: Which manifolds satisfy $W_- = 0$, $R > 0$?

- S^4, *round metric of radius* ρ. Constant sectional curvature $1/\rho^2$, so $R = n(n-1)/\rho^2 = 12/\rho^2 > 0$, and the whole Weyl tensor vanishes (it is conformally flat). Both conditions hold. On Λ^2_- the Weitzenböck potential is $R/6 = 2/\rho^2$.
- $\mathbb{CP}^2$ *with the Fubini–Study metric* and its complex orientation: $W_- = 0$ (it is self-dual, though $W_+ \neq 0$) and $R > 0$. This is the other main example; its intersection form is $(+1)$.
- *The flat torus* T^4: $R = 0$, $W = 0$. The argument gives only $\|(\mathrm{d}^-)^*\omega\|^2 = \frac{1}{2}\|\nabla\omega\|^2 \geq 0$, and indeed for the trivial connection the constant ASD forms ω_a^- are non-zero elements of H^2: the hypothesis $R > 0$ cannot be dropped.

Hitchin showed that among compact Einstein manifolds with $W_- = 0$ and $R > 0$ there are only these two, S^4 and $\mathbb{CP}^2$. For other manifolds one works with generic metrics instead (Freed–Uhlenbeck, Taubes), for which $h_2 = 0$ holds at irreducible instantons after a small perturbation of the metric.

8.8.8 Comparison with the moduli of Riemann surfaces

Nash says the technique is "similar to that used in chapter 5". The dictionary is worth writing out.

	Riemann surfaces of genus p	Instantons of charge k
object	complex structure on a surface Σ	self-dual connection on $P \to M$
equivalence	diffeomorphisms of Σ	gauge transformations
infinitesimal symmetries	$H^0(\Sigma, T\Sigma)$ (holomorphic vector fields)	$H^0 = \operatorname{Ker} \mathrm{d}_A$
deformations	$H^1(\Sigma, T\Sigma)$	$H^1 = \operatorname{Ker} \mathrm{d}_A^- / \operatorname{Im} \mathrm{d}_A$
elliptic operator	$\bar\partial$ on vector fields	$\mathrm{d}_A^* \oplus \mathrm{d}_A^-$ (Fueter type)
vanishing for large genus / irreducible	$H^0 = 0$ for $p \geq 2$	$H^0 = 0$ if irreducible
index theorem	Riemann–Roch: $h^0 - h^1 = 3 - 3p$	Atiyah–Singer: $h_0 - h_1 + h_2 = 3 - 8k$ on S^4
dimension	$\dim_{\mathbb{C}} \mathcal{M}_p = 3p - 3$	$\dim \mathcal{M}_k = 8k - 3$

The difference that matters is linearity. The $\bar\partial$ operator is linear, so the Riemann moduli problem is in a sense a linear one. The instanton equations are non-linear; only their linearisation is an elliptic complex, which is why the extra global conditions of Section 8.11 are needed.

Reading Nash (pp. 233–236, (8.93)–(8.113))

Nash's (8.93)–(8.94) set up $a \in T_A\mathcal{A}_k$ and the target $T_{[A]}\mathcal{M}_k$. (8.95)–(8.101) are Section 8.8.2: π_-, the expansion of $F(A + ta)$ and the operator $\mathrm{d}_A^- = \pi_- \circ \mathrm{d}_A$. (8.102) is the quotient (8.227); we added the harmonic representative (8.228). (8.103)–(8.104) are the bundle $\operatorname{ad} P \otimes \Lambda^1$ and the complex (8.230). The checks (8.105)–(8.108), with the footnote formula for d_A, are Section 8.8.4. (8.109)–(8.111) are Section 8.8.5. (8.112)–(8.113) are the Laplacian and the Weitzenböck formula (8.233), with the AHS hypotheses $W_- = 0$, $R > 0$. Nash does not show that the complex is elliptic; we did this at the symbol level in Example 8.71.

Recap

- Tangent vectors to $\mathcal{M}_k$: perturbations a with $\mathrm{d}_A^- a = \pi_- \mathrm{d}_A a = 0$, modulo gauge directions $\mathrm{d}_A f$; equivalently the zero modes $\mathrm{d}_A^- a = 0 = \mathrm{d}_A^* a$.
- $0 \to \Omega^0 \xrightarrow{\mathrm{d}_A} \Omega^1 \xrightarrow{\mathrm{d}_A^-} \Omega^2_- \to 0$ is a complex because $\mathrm{d}_A^- \mathrm{d}_A f = [F^-, f] = 0$; it is elliptic (symbol sequence $1 \to 4 \to 3$ exact, per colour).
- H^0 = symmetries, H^1 = deformations = $T_{[A]}\mathcal{M}_k$, H^2 = obstructions. The index gives $h_0 - h_1 + h_2$, stable under deformation.
- In flat space, gauge condition + linearised SD = the Fueter equation $(\partial_4 + e_j\partial_j)\bar a = 0$, a quaternionic Cauchy–Riemann (Dirac) equation that squares to $\Box$.
- $h_0 = 0$ for irreducible A; $h_2 = 0$ by the Weitzenböck formula $\mathrm{d}_A^-(\mathrm{d}_A^-)^* = \frac{1}{2}\nabla^*\nabla + R/6 - W_-$ when $W_- = 0$, $R > 0$ (S^4, $\mathbb{CP}^2$). Then $\dim \mathcal{M}_k = h_1 = -\operatorname{index}$.

8.9 Counting the moduli: the index theorem

We now have all the pieces of the zero-mode count. The tangent space to the moduli space at an irreducible instanton is H^1 of the deformation complex, $H^0 = 0$ because A is irreducible, and $H^2 = 0$ by the Weitzenböck argument when $W_- = 0$ and $R > 0$. What is left is arithmetic: compute $h_0 - h_1 + h_2$ with the Atiyah–Singer index theorem and read off h_1.

For a physicist this is a familiar move. One wants the number of zero modes of a fluctuation operator. Counting them directly means solving a partial differential equation. The index theorem replaces that by an integral of a local density (a polynomial in the curvature) over the manifold. The answer comes out in terms of two kinds of topological data:

- the *bundle*: how twisted the gauge field is, through the instanton number k;
- the *manifold*: how twisted spacetime is, through the Euler characteristic $\chi(M)$ and the signature $\tau(M)$.

The final formula for $G = \mathrm{SU}(2)$ on S^4 is $\dim \mathcal{M}_k = 8k - 3$. We derive it line by line, then do the same for other manifolds and other groups.

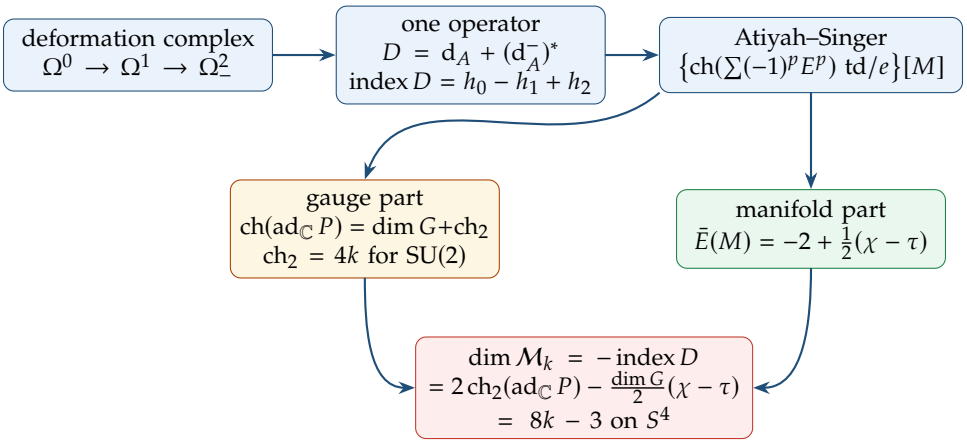


Figure 8.45: The logic of the dimension count. The complex is rolled up into a single elliptic operator D; its index is an integral given by the Atiyah–Singer formula; the integrand factorises into a gauge part (the Chern character of the adjoint bundle) and a manifold part (the index density of the untwisted complex). Only the four-form part of the product survives the integral. The red box is Nash's (8.124) and (8.132).

8.9.1 Rolling the complex up into one operator

The index theorem is a statement about one elliptic operator, not about a complex. The standard trick is to put the even terms of the complex on one side and the odd term on the other. Nash's (8.114) is

Working definition 8.34 (the rolled-up deformation operator)

$$D : \Omega^0(M, \mathrm{ad}\, P) \oplus \Omega^2_-(M, \mathrm{ad}\, P) \longrightarrow \Omega^1(M, \mathrm{ad}\, P), \qquad D(f, \omega) = \mathrm{d}_A f + (\mathrm{d}_A^-)^* \omega. \quad (8.234)$$

Per colour and per point, D maps $1 + 3 = 4$ components to 4 components, so it is a square first-order system [as many equations as unknowns, each with one derivative]. Its adjoint [the L^2 adjoint, with $\langle D(f,\omega), a\rangle = \langle (f,\omega), D^* a\rangle$] is $D^* a = (\mathrm{d}_A^* a,\ \mathrm{d}_A^- a)$, which is the operator $\mathcal{D}_A$ of Section 8.8.6.

The kernel of D. Suppose $D(f,\omega) = 0$. The two pieces of $D(f,\omega)$ are orthogonal, because of the complex property. Indeed,

$$\begin{aligned} \|D(f,\omega)\|^2 &\overset{(1)}{=} \|\mathrm{d}_A f\|^2 + 2\langle \mathrm{d}_A f, (\mathrm{d}_A^-)^*\omega\rangle + \left\|(\mathrm{d}_A^-)^*\omega\right\|^2 \\ &\overset{(2)}{=} \|\mathrm{d}_A f\|^2 + 2\langle \mathrm{d}_A^- \mathrm{d}_A f, \omega\rangle + \left\|(\mathrm{d}_A^-)^*\omega\right\|^2 \\ &\overset{(3)}{=} \|\mathrm{d}_A f\|^2 + \left\|(\mathrm{d}_A^-)^*\omega\right\|^2 . \end{aligned} \quad (8.235)$$

(1) Expand the square of a sum; the inner product (8.165) is real and symmetric.
(2) Move $(\mathrm{d}_A^-)^*$ to the other side of the inner product: this is the definition of the adjoint.
(3) $\mathrm{d}_A^- \mathrm{d}_A f = [F^-, f] = 0$ for a self-dual A ((8.226)).

A sum of two non-negative numbers vanishes only if both do. So $D(f,\omega)=0$ exactly when $\mathrm{d}_A f=0$ and $(\mathrm{d}_A^-)^*\omega=0$:

$$\operatorname{Ker} D=\operatorname{Ker}\mathrm{d}_A^{(0)}\oplus\operatorname{Ker}(\mathrm{d}_A^-)^*\cong H^0\oplus H^2,$$

using the harmonic description of H^2 in Section 8.8.5.

The cokernel of D. The cokernel is the orthogonal complement of the image, which is the kernel of the adjoint:

$$\operatorname{Coker} D\cong\operatorname{Ker} D^*=\{a:\ \mathrm{d}_A^*a=0,\ \mathrm{d}_A^-a=0\}\cong H^1$$

by (8.228). Therefore

$$\begin{aligned}&\operatorname{index} D=\dim\operatorname{Ker} D-\dim\operatorname{Coker} D=h_0+h_2-h_1,\\ &\text{and, when } h_0=h_2=0,\qquad \operatorname{index} D=-h_1=-\dim\mathcal{M}_k.\end{aligned}\tag{8.236}$$

This is Nash's (8.115). The index of the rolled-up operator is the alternating sum $h_0-h_1+h_2$ of the complex, as it must be.

Worked example 8.77: Rolling up the finite-dimensional toy complex

Take the complex of Example 8.72: $V_0=\mathbb{R}$, $V_1=\mathbb{R}^3$, $V_2=\mathbb{R}$, $d_0=(1,0,0)^T$, $d_1=(0,s,0)$. The rolled-up operator $D=d_0\oplus d_1^T: V_0\oplus V_2\to V_1$ is the 3×2 matrix

$$D=\begin{pmatrix}1&0\\0&s\\0&0\end{pmatrix}.$$

- $s\neq 0$: $\operatorname{Ker} D=0$, $\operatorname{Im} D$ is 2-dimensional, $\operatorname{Coker} D$ is 1-dimensional; $\operatorname{index} D=0-1=-1$.
- $s=0$: $\operatorname{Ker} D$ is spanned by $(0,1)$, so it is 1-dimensional; $\operatorname{Im} D$ is 1-dimensional, $\operatorname{Coker} D$ is 2-dimensional; $\operatorname{index} D=1-2=-1$.

In both cases $\operatorname{index} D=h_0+h_2-h_1$ with $(h_0,h_1,h_2)=(0,1,0)$ or $(0,2,1)$, and in both cases it equals $n_0-n_1+n_2=1-3+1$: the index of a map between spaces of dimensions 2 and 3 is always $2-3$. When $s\to 0$ the kernel and cokernel grow *together*. That is the finite model of what the index theorem exploits.

Watch out: Which way does D point?

Nash's D goes from the even terms $\Omega^0\oplus\Omega^2_-$ to the odd term Ω^1, so $\operatorname{index} D=h_0-h_1+h_2$ is *minus* the dimension of the moduli space. Its adjoint $\mathcal{D}_A=D^*$ has index $+h_1$. Nothing deep is involved, but the signs in Nash's (8.124) only make sense once one remembers that $\dim\mathcal{M}_k=-\operatorname{index} D$.

8.9.2 The index formula for an elliptic complex

The formula

In Chapter 4 we met the cohomological form of the index theorem for an elliptic complex $0\to\Gamma(E^0)\to\Gamma(E^1)\to\cdots\to\Gamma(E^m)\to 0$ on a compact oriented manifold of even dimension n (Nash's (4.48), repeated as his (8.116)):

$$\text{index}=(-1)^{n/2}\left\{\frac{\mathrm{ch}\big(\sum_p(-1)^pE^p\big)}{e(M)}\ \mathrm{td}(TM\otimes\mathbb{C})\right\}[M].\tag{8.237}$$

Let us recall what each symbol means and how to compute it.

- $\mathrm{ch}(E)$ is the **Chern character** of a complex vector bundle E with connection: $\mathrm{ch}(E) = \operatorname{tr}\exp(iF_E/2\pi)$, a sum of forms of degrees $0, 2, 4, \dots$. ch of a formal difference is the difference of the ch's.
- $e(M)$ is the **Euler class** of TM; its integral is $\chi(M)$ (Gauss–Bonnet, Chapter T2).
- td is the **Todd class**; for $TM \otimes \mathbb{C}$ it depends only on the Pontryagin classes of M.
- The quotient by $e(M)$ looks alarming, but for an elliptic complex the numerator is always divisible by $e(M)$ as a polynomial in the curvature (we shall see this happen explicitly).
- $\{\cdots\}[M]$ means: keep the part of degree n and integrate it over M.

In our case $n = 4$, so $(-1)^{n/2} = +1$, and only 4-forms matter.

Complexification

The index theorem is stated for complex bundles. Our operators are real: $\operatorname{ad} P$ is a real bundle with fibre $\mathfrak{g}$. We therefore complexify, replacing $\mathfrak{g}$ by $\mathfrak{g} \otimes \mathbb{C}$ and $\operatorname{ad} P$ by $\operatorname{ad}_{\mathbb{C}} P = \operatorname{ad} P \otimes \mathbb{C}$. This does not change the index: a real linear map and its complexification have the same matrix, so $\dim_{\mathbb{R}} \operatorname{Ker} = \dim_{\mathbb{C}} \operatorname{Ker}_{\mathbb{C}}$ and the same for the cokernel. In Nash's (8.117) the bundles of the complex become

$$E^0 = \operatorname{ad}_{\mathbb{C}} P \otimes \Lambda^0 T^*M_{\mathbb{C}}, \qquad E^1 = \operatorname{ad}_{\mathbb{C}} P \otimes \Lambda^1 T^*M_{\mathbb{C}}, \qquad E^2 = \operatorname{ad}_{\mathbb{C}} P \otimes \Lambda^2_- T^*M_{\mathbb{C}}. \tag{8.238}$$

Factorising: the Chern character is multiplicative

Every E^p is "gauge part $\otimes$ form part". The Chern character turns tensor products into products. Here is why, at the level of curvature. On a tensor product $V \otimes W$ the connection acts by the Leibniz rule, so the curvature is $F_{V\otimes W} = F_V \otimes \mathbb{1} + \mathbb{1} \otimes F_W$, and the two terms commute. Then

$$\begin{aligned} \mathrm{ch}(V \otimes W) &\overset{(1)}{=} \operatorname{tr}_{V\otimes W} \exp\Big(\tfrac{i}{2\pi}\big(F_V \otimes \mathbb{1} + \mathbb{1} \otimes F_W\big)\Big) \\ &\overset{(2)}{=} \operatorname{tr}_{V\otimes W}\Big(\exp\big(\tfrac{i}{2\pi}F_V\big) \otimes \exp\big(\tfrac{i}{2\pi}F_W\big)\Big) \\ &\overset{(3)}{=} \operatorname{tr}_V \exp\big(\tfrac{i}{2\pi}F_V\big)\ \operatorname{tr}_W \exp\big(\tfrac{i}{2\pi}F_W\big) = \mathrm{ch}(V)\,\mathrm{ch}(W). \end{aligned} \tag{8.239}$$

(1) Definition of ch and the Leibniz form of the curvature.
(2) $e^{X+Y} = e^X e^Y$ for commuting X, Y (the curvatures are even forms, so they commute as forms too), and $e^{X\otimes\mathbb{1}} = e^X \otimes \mathbb{1}$.
(3) $\operatorname{tr}(X \otimes Y) = \operatorname{tr} X\ \operatorname{tr} Y$ (a Kronecker-product identity).

(In the language of Chapter 4: the Chern roots of $V \otimes W$ are the pairwise sums of the roots of V and of W, and $e^{a+b} = e^a e^b$.) Applying this to each E^p and pulling out the common factor,

$$\operatorname{index} D = \big\{\mathrm{ch}(\operatorname{ad}_{\mathbb{C}} P)\,\bar{E}(M)\big\}[M], \qquad \bar{E}(M) = \frac{\mathrm{ch}\big(\sum_p (-1)^p \bar{E}^p\big)}{e(M)}\ \mathrm{td}(TM_{\mathbb{C}}), \tag{8.240}$$

with $\bar{E}^0 = \Lambda^0 T^*M_{\mathbb{C}}$, $\bar{E}^1 = \Lambda^1 T^*M_{\mathbb{C}}$, $\bar{E}^2 = \Lambda^2_- T^*M_{\mathbb{C}}$ (Nash's (8.118)). The class $\bar{E}(M)$ knows nothing about the gauge field. It is the index density of the *untwisted* complex $\Omega^0 \to \Omega^1 \to \Omega^2_-$, i.e. the same complex for the trivial connection on the trivial line bundle.

Because only the 4-form part of the product is integrated, we need $\mathrm{ch}(\operatorname{ad}_{\mathbb{C}} P)$ and $\bar{E}(M)$ each in degrees 0 and 4 (the degree-2 parts will turn out not to contribute). We compute the two factors in turn.

8.9.3 The gauge factor: $\mathrm{ch}(\mathrm{ad}_{\mathbb{C}}\, P)$

Chern character in terms of Chern classes

Diagonalise the curvature formally: $\frac{i}{2\pi}F_V$ has "eigenvalues" $x_1, \dots, x_r$ (the **Chern roots**, 2-forms), with $r = \mathrm{rk}\, V$. The total Chern class is $c(V) = \prod_i (1 + x_i)$, so $c_1 = \sum_i x_i$ and $c_2 = \sum_{i<j} x_i x_j$. Then

$$\begin{aligned} \mathrm{ch}(V) &\overset{(1)}{=} \sum_{i=1}^{r} e^{x_i} \\ &\overset{(2)}{=} r + \sum_i x_i + \tfrac{1}{2}\sum_i x_i^2 + \dots \\ &\overset{(3)}{=} \mathrm{rk}\, V + c_1(V) + \tfrac{1}{2}\big(c_1(V)^2 - 2c_2(V)\big) + \dots \end{aligned} \tag{8.241}$$

(1) The trace of the exponential of a diagonal matrix.
(2) Expand each exponential; on a 4-manifold, terms of degree > 4 (three or more x's) vanish.
(3) $\sum_i x_i^2 = (\sum_i x_i)^2 - 2\sum_{i<j} x_i x_j = c_1^2 - 2c_2$.

This is Nash's (8.119).

Real bundles: $c_1 = 0$ and the Pontryagin class

$\mathrm{ad}_{\mathbb{C}}\, P$ is the complexification of a real bundle. Two consequences.

- *The rank is* $\dim G$: the fibre is $\mathfrak{g} \otimes \mathbb{C}$.
- $c_1 = 0$. In a real representation the curvature is a real antisymmetric matrix (the adjoint representation of a compact group preserves the Killing form). The eigenvalues of a real antisymmetric matrix come in pairs $\pm i\mu$. So the Chern roots come in pairs $\pm x$, and $c_1 = \sum x_i = 0$. (Nash says this as: a complexified real bundle is *self-conjugate*, $V \cong \bar{V}$, while $c_1(\bar{V}) = -c_1(V)$; so $2c_1 = 0$, which kills c_1 in de Rham cohomology. All odd Chern classes vanish for the same reason.)

Hence

$$\mathrm{ch}(\mathrm{ad}_{\mathbb{C}}\, P) = \dim G + \mathrm{ch}_2(\mathrm{ad}_{\mathbb{C}}\, P), \qquad \mathrm{ch}_2(\mathrm{ad}_{\mathbb{C}}\, P) = -c_2(\mathrm{ad}_{\mathbb{C}}\, P). \tag{8.242}$$

The degree-2 part is zero, which is why degree-2 parts of $\bar{E}(M)$ never contribute.

Watch out: Nash's normalisation of p_1

The standard first Pontryagin class of a real bundle V is $p_1(V) = -c_2(V \otimes \mathbb{C})$. Nash writes $p_1(\mathrm{ad}_{\mathbb{C}}\, P) = -2c_2(\mathrm{ad}_{\mathbb{C}}\, P)$ and hence (8.120), $\mathrm{ch}(\mathrm{ad}_{\mathbb{C}}\, P) = \dim G + \frac{1}{2}p_1$. So *Nash's p_1 is twice the standard one.* The combination that enters the final answer, $-2c_2(\mathrm{ad}_{\mathbb{C}}\, P)$, is unambiguous. We shall write everything in terms of $c_2(\mathrm{ad}_{\mathbb{C}}\, P)$ and $\mathrm{ch}_2(\mathrm{ad}_{\mathbb{C}}\, P) = -c_2(\mathrm{ad}_{\mathbb{C}}\, P)$, and translate to Nash's p_1 at the end: $p_1^{\mathrm{Nash}} = 2\,\mathrm{ch}_2(\mathrm{ad}_{\mathbb{C}}\, P)$.

Worked example 8.78: The adjoint of $\mathrm{SU}(2)$ from the fundamental: $\mathrm{ch}_2(\mathrm{ad}) = 4k$

Let us compute $c_2(\mathrm{ad}_{\mathbb{C}}\, P)$ directly from the curvature, by comparing traces in the two representations. Take an element of $\mathfrak{su}(2)$ in its Cartan direction, $X = \theta e_3 = \mathrm{diag}(-i\theta, i\theta)$.

- Fundamental: eigenvalues $\mp i\theta$, so $\mathrm{tr}_{\mathbf{2}}\, X^2 = -2\theta^2$.

- Adjoint: $\operatorname{ad} X(Y) = [X, Y]$. On the basis e_1, e_2, e_3, using $[e_j, e_k] = 2\varepsilon_{jkl} e_l$ from (8.193),

$$\operatorname{ad} X = \theta \begin{pmatrix} 0 & -2 & 0 \\ 2 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad \text{eigenvalues } 0,\ \pm 2i\theta,$$

so $\operatorname{tr}_{\mathbf{3}}(\operatorname{ad} X)^2 = -8\theta^2$.

(Check the first column: $[e_3, e_1] = 2e_2$, so $\operatorname{ad} X(e_1) = 2\theta e_2$.) Every element of $\mathfrak{su}(2)$ is conjugate to such an X, and both traces are conjugation invariant, so $\operatorname{tr}_{\mathbf{3}}(\operatorname{ad} X)^2 = 4 \operatorname{tr}_{\mathbf{2}} X^2$ for all X. The same holds for the curvature 2-form, since $\operatorname{tr}(F \wedge F)$ is built from $\operatorname{tr}(X^2)$ by polarisation. With Chern–Weil, $c_2 = \frac{1}{8\pi^2} \operatorname{tr}(F \wedge F)$ for a bundle with traceless curvature (from $c(V) = \det(\mathbb{1} + \frac{i}{2\pi} F)$), and so

$$\begin{aligned} \operatorname{ch}_2(\operatorname{ad}_{\mathbb{C}} P)[M] &\overset{(1)}{=} -c_2(\operatorname{ad}_{\mathbb{C}} P)[M] \\ &\overset{(2)}{=} -\frac{1}{8\pi^2} \int_M \operatorname{tr}_{\mathbf{3}}\left(F_{\mathrm{ad}} \wedge F_{\mathrm{ad}}\right) \\ &\overset{(3)}{=} -\frac{4}{8\pi^2} \int_M \operatorname{tr}_{\mathbf{2}}(F \wedge F) \\ &\overset{(4)}{=} 4k. \end{aligned}$$

(1) (8.242).
(2) Chern–Weil for the traceless curvature of $\operatorname{ad}_{\mathbb{C}} P$.
(3) The trace identity $\operatorname{tr}_{\mathbf{3}} = 4 \operatorname{tr}_{\mathbf{2}}$.
(4) The instanton number $k = -\frac{1}{8\pi^2} \int \operatorname{tr}(F \wedge F) = -c_2(E)[M]$ ((8.187), Nash's (8.127)).

For the BPST instanton ($k = 1$) this says $\operatorname{ch}_2(\operatorname{ad}_{\mathbb{C}} P)[S^4] = 4$, i.e. Nash's $p_1(\operatorname{ad}_{\mathbb{C}} P) = 8$. Nash obtains the same number by decomposing $E \otimes E$ (Section 8.9.6); it is worth having both derivations.

8.9.4 The manifold factor: the truncated de Rham complex

Now $\bar{E}(M)$, the index density of $\Omega^0 \xrightarrow{\mathrm{d}} \Omega^1 \xrightarrow{\mathrm{d}^-} \Omega^2_-$ with no gauge field. Nash evaluates it without computing a single characteristic class, by identifying the cohomology of this complex with familiar Betti numbers.

The cohomology of the untwisted complex

Let $b_p = \dim H^p_{\mathrm{dR}}(M)$ be the Betti numbers. Recall (Chapter 1) that the harmonic 2-forms split into self-dual and anti-self-dual ones, of dimensions b_2^+ and b_2^-, and that the intersection form $q(\alpha, \beta) = \int_M \alpha \wedge \beta$ is positive on the first and negative on the second:

$$q(\omega, \omega) = \int_M \omega \wedge \omega = \pm \int_M \omega \wedge \star\omega = \pm \|\omega\|^2 \qquad \text{for } \star\,\omega = \pm\omega.$$

So b_2^+ (b_2^-) is the number of positive (negative) eigenvalues of the matrix of q, and the **signature** is $\tau(M) = b_2^+ - b_2^-$.

We compute the three cohomology groups of the untwisted complex, using harmonic representatives as in Section 8.8.5.

- $H^0 = \operatorname{Ker} \mathrm{d}$ on functions: the locally constant functions. $h_0 = b_0$ (= 1 for connected M).

- $H^1 = \operatorname{Ker} \mathrm{d}^- \cap \operatorname{Ker} \mathrm{d}^*$ on 1-forms. We claim that $\mathrm{d}^- a = 0$ forces $\mathrm{d}a = 0$ on a closed manifold. Split $\mathrm{d}a = (\mathrm{d}a)^+ + (\mathrm{d}a)^-$; then

$$\begin{aligned} \left\|(\mathrm{d}a)^+\right\|^2 - \|(\mathrm{d}a)^-\|^2 &\overset{(1)}{=} \int_M \mathrm{d}a \wedge \mathrm{d}a \\ &\overset{(2)}{=} \int_M \mathrm{d}(a \wedge \mathrm{d}a) \\ &\overset{(3)}{=} 0. \end{aligned}$$

  (1) The self-dual and anti-self-dual parts are orthogonal, and $\omega \wedge \omega = \pm|\omega|^2 \mathrm{vol}$ on each, as above.
  (2) $\mathrm{d}(a \wedge \mathrm{d}a) = \mathrm{d}a \wedge \mathrm{d}a - a \wedge \mathrm{d}\mathrm{d}a$ and $\mathrm{d}^2 = 0$.
  (3) Stokes; $\partial M = \varnothing$.

  So $(\mathrm{d}a)^- = 0$ implies $(\mathrm{d}a)^+ = 0$, i.e. $\mathrm{d}a = 0$. Together with $\mathrm{d}^* a = 0$, a is a harmonic 1-form. Hence $h_1 = b_1$.

- $H^2 = \Omega^2_- / \operatorname{Im} \mathrm{d}^- \cong \operatorname{Ker}(\mathrm{d}^-)^*$ on ASD 2-forms. For an ASD form, $(\mathrm{d}^-)^*\omega = \mathrm{d}^*\omega = -\star\mathrm{d}\star\omega = \star\mathrm{d}\omega$, so $(\mathrm{d}^-)^*\omega = 0$ iff $\mathrm{d}\omega = 0$; then also $\mathrm{d}^*\omega = 0$, and ω is an ASD harmonic form. Hence $h_2 = b_2^-$.

So the index of the untwisted complex is (Nash's (8.121))

$$\bar{E}(M)[M] = b_0 - b_1 + b_2^-.$$

Rewriting in terms of χ and τ

Nash explains that the complex is "the de Rham complex truncated in two ways": the last two terms Ω^3, Ω^4 are missing, and in the middle only Ω^2_- appears instead of all of Ω^2. The missing terms contribute the same as the first two by Poincaré duality, and the middle discrepancy is measured by the signature. In formulas (Nash's (8.122)–(8.123)):

$$\begin{aligned} \chi(M) &\overset{(1)}{=} b_0 - b_1 + b_2 - b_3 + b_4 \\ &\overset{(2)}{=} 2b_0 - 2b_1 + b_2 \\ &\overset{(3)}{=} 2b_0 - 2b_1 + 2b_2^- + (b_2^+ - b_2^-) \\ &\overset{(4)}{=} 2\big(b_0 - b_1 + b_2^-\big) + \tau(M), \end{aligned} \tag{8.243}$$

  (1) The Euler characteristic is the alternating sum of the Betti numbers.
  (2) Poincaré duality on a closed oriented 4-manifold: $b_4 = b_0$, $b_3 = b_1$ (the Hodge star maps harmonic p-forms to harmonic $(4-p)$-forms).
  (3) $b_2 = b_2^+ + b_2^-$, written as $2b_2^- + (b_2^+ - b_2^-)$.
  (4) Definition of τ.

and therefore

$$b_0 - b_1 + b_2^- = \tfrac{1}{2}\big(\chi(M) - \tau(M)\big). \tag{8.244}$$

Worked example 8.79: Checking (8.244) on five manifolds

M	b_0, b_1, b_2^+, b_2^-	χ	τ	$b_0 - b_1 + b_2^-$	$\frac{1}{2}(\chi - \tau)$
S^4	$1, 0, 0, 0$	2	0	1	1
$\mathbb{CP}^2$	$1, 0, 1, 0$	3	1	1	1
$S^2 \times S^2$	$1, 0, 1, 1$	4	0	2	2
T^4	$1, 4, 3, 3$	0	0	0	0
$K3$	$1, 0, 3, 19$	24	-16	20	20

For T^4: $b_1 = 4$ (the forms $\mathrm{d}x^\mu$), $b_2 = 6$ (the $\mathrm{d}x^{\mu\nu}$), split $3 + 3$ into the $\omega_a^\pm$ of Section 8.4, and $\chi = 1 - 4 + 6 - 4 + 1 = 0$. For $S^2 \times S^2$ the intersection form is $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ with eigenvalues ± 1, so $b_2^\pm = 1$. For $K3$ the form is $3H \oplus 2(-E_8)$ in one orientation (Chapter 1), giving $b_2^+ = 3$, $b_2^- = 19$.

The same answer from characteristic classes

It is reassuring, and it shows how the division by $e(M)$ works, to compute $\bar{E}(M)$ directly. By the splitting principle, treat TM as if it were a sum of two oriented planes, rotated with "angles" x_1, x_2 (2-forms). Then the Euler class is $e = x_1 x_2$, the first Pontryagin class is $p_1 = x_1^2 + x_2^2$, and the complexified bundles have the roots

$$\Lambda^0_{\mathbb{C}} : \{0\}, \qquad \Lambda^1_{\mathbb{C}} : \{\pm x_1, \pm x_2\}, \qquad \Lambda^2_{-,\mathbb{C}} : \{0,\ \pm(x_1 - x_2)\}$$

(the self-dual forms carry $\{0, \pm(x_1 + x_2)\}$; we checked in Section 8.4 that $\mathrm{d}x^{13} - \mathrm{d}x^{24}$, which rotates with angle $x_1 + x_2$, is self-dual). The numerator of $\bar{E}(M)$ is, with $\cosh y = 1 + \frac{y^2}{2} + \frac{y^4}{24} + \dots$,

$$\begin{aligned} \mathrm{ch}\big(\Lambda^0 - \Lambda^1 + \Lambda^2_-\big) &\overset{(1)}{=} 1 - \big(2\cosh x_1 + 2\cosh x_2\big) + \big(1 + 2\cosh(x_1 - x_2)\big) \\ &\overset{(2)}{=} \big[-x_1^2 - x_2^2 + (x_1 - x_2)^2\big] + \tfrac{1}{12}\big[(x_1 - x_2)^4 - x_1^4 - x_2^4\big] + \dots \\ &\overset{(3)}{=} -2x_1x_2 + x_1x_2\Big(-\tfrac{1}{3}x_1^2 + \tfrac{1}{2}x_1x_2 - \tfrac{1}{3}x_2^2\Big) + \dots \\ &\overset{(4)}{=} e\left(-2 - \tfrac{1}{3}p_1 + \tfrac{1}{2}e\right) + \dots \end{aligned}$$

(1) ch is the sum of e^{root}; pairs $\pm y$ give $2\cosh y$.
(2) Expand. The constant terms cancel: $1 - 4 + 3 = 0$ (the ranks of the complex, $1 - 4 + 3$).
(3) $(x_1 - x_2)^4 - x_1^4 - x_2^4 = -4x_1^3x_2 + 6x_1^2x_2^2 - 4x_1x_2^3$, and divide by 12.
(4) $e = x_1x_2$, $p_1 = x_1^2 + x_2^2$.

The numerator is divisible by e, as promised. The Todd class of $TM_{\mathbb{C}}$ (roots $\pm x_1, \pm x_2$) is

$$\mathrm{td}(TM_{\mathbb{C}}) = \prod_{i=1}^{2} \frac{x_i}{1 - e^{-x_i}} \cdot \frac{-x_i}{1 - e^{x_i}} = \prod_i \left(1 - \frac{x_i^2}{12} + \dots\right) = 1 - \frac{p_1}{12} + \dots$$

(multiply $1 + \frac{y}{2} + \frac{y^2}{12}$ by $1 - \frac{y}{2} + \frac{y^2}{12}$). Hence

$$\begin{aligned} \bar{E}(M) &\overset{(1)}{=} \left(-2 - \tfrac{1}{3}p_1 + \tfrac{1}{2}e\right)\left(1 - \tfrac{1}{12}p_1\right) \\ &\overset{(2)}{=} -2 + \left(\tfrac{1}{6} - \tfrac{1}{3}\right)p_1 + \tfrac{1}{2}e \\ &\overset{(3)}{=} -2 + \tfrac{1}{2}\left(e - \tfrac{p_1}{3}\right). \end{aligned} \tag{8.245}$$

(1) The two factors just computed.
(2) Keep degree ≤ 4: $(-2)(-\frac{1}{12}p_1) = \frac{1}{6}p_1$.
(3) Collect.

Integrating, $\int_M e = \chi(M)$ (Gauss–Bonnet) and $\int_M \frac{1}{3}p_1 = \tau(M)$ (the Hirzebruch signature theorem, Chapter 4). So the 4-form part integrates to $\frac{1}{2}(\chi - \tau)$, in agreement with (8.244). The new information is the *degree-0 part*, -2. It multiplies $\mathrm{ch}_2(\mathrm{ad}_{\mathbb{C}}\, P)$ in the final formula.

Physical picture: Why -2: the operator is a twisted Dirac operator

In Example 8.74 we saw that, in spinor language, $\Lambda^1_{\mathbb{C}} \cong S^+ \otimes S^-$ and $\Lambda^0_{\mathbb{C}} \oplus \Lambda^2_{-,\mathbb{C}} \cong S^- \otimes S^-$. So the adjoint $\mathcal{D}_A = D^* : \Omega^1 \to \Omega^0 \oplus \Omega^2_-$ is the chiral Dirac operator $S^+ \to S^-$ twisted by the bundle $V = S^- \otimes \mathrm{ad}_{\mathbb{C}}\, P$ of rank $2 \dim G$. By the twisted Dirac index theorem (Chapter 4), index $= \int \hat{A}(M)\,\mathrm{ch}(V)$, and the gauge field enters through $\mathrm{ch}(V) = \mathrm{ch}(S^-)\,\mathrm{ch}(\mathrm{ad}_{\mathbb{C}}\, P)$, whose degree-4 gauge part is $\mathrm{rk}(S^-)\ \mathrm{ch}_2(\mathrm{ad}_{\mathbb{C}}\, P) = 2\,\mathrm{ch}_2$. The "2" is the number of components of a Weyl spinor. It enters index $D = -\,\mathrm{index}\,\mathcal{D}_A$ with a minus sign: that is the -2 in $\bar{E}(M)$.

8.9.5 The dimension formula

Multiply the two factors and keep the 4-form part:

$$\begin{aligned} \mathrm{index}\, D &\overset{(1)}{=} \Big\{ \big(\dim G + \mathrm{ch}_2(\mathrm{ad}_{\mathbb{C}}\, P)\big)\Big(-2 + \tfrac{1}{2}\big(e - \tfrac{p_1}{3}\big)\Big)\Big\}[M] \\ &\overset{(2)}{=} \tfrac{\dim G}{2}\big(\chi(M) - \tau(M)\big) - 2\,\mathrm{ch}_2(\mathrm{ad}_{\mathbb{C}}\, P)[M], \end{aligned} \tag{8.246}$$

(1) (8.240) with (8.242) and (8.245).
(2) The 4-form part is (degree 0)×(degree 4) + (degree 4)×(degree 0); integrate using Gauss–Bonnet and the signature theorem.

and with $\dim \mathcal{M}_k = -\,\mathrm{index}\, D$ from (8.236):

Key idea: Dimension of the instanton moduli space (Nash's (8.124))

For G compact and simple, M compact, oriented, with $W_- = 0$ and $R > 0$, and A irreducible and self-dual,

$$\dim \mathcal{M}_k = 2\,\mathrm{ch}_2(\mathrm{ad}_{\mathbb{C}}\, P)[M] - \frac{\dim G}{2}\big(\chi(M) - \tau(M)\big) = p_1^{\mathrm{Nash}}(\mathrm{ad}_{\mathbb{C}}\, P) - \frac{\dim G}{2}(\chi - \tau). \tag{8.247}$$

Equivalently $\dim \mathcal{M}_k = -2c_2(\mathrm{ad}_{\mathbb{C}}\, P)[M] - \dim G\,(b_0 - b_1 + b_2^-)$. The first term counts the zero modes created by the gauge field; the second subtracts the symmetries and adds the obstructions that the untwisted complex would have.

Nash writes the product as $\{\dim G + \frac{1}{2}p_1\}\{2 - \frac{1}{2}(\chi - \tau)\}$: his second brace is $-\bar{E}(M)$, which absorbs the sign of $\dim \mathcal{M}_k = -\,\mathrm{index}\, D$. Taubes showed the formula holds more generally (for a class of non-self-dual manifolds, and for generic metrics), as Nash notes.

8.9.6 SU(2) instantons on S^4: $\dim \mathcal{M}_k = 8k - 3$

The manifold

S^4 with the round metric has $R > 0$ and $W = 0$ (Example 8.76), so the hypotheses hold. Its Betti numbers are $1, 0, 0, 0, 1$, so (Nash's (8.125))

$$\chi(S^4) = 2, \qquad \tau(S^4) = 0, \qquad \tfrac{1}{2}(\chi - \tau) = 1,$$

and with $\dim \mathrm{SU}(2) = 3$ the formula becomes (Nash's (8.126)) $\dim \mathcal{M}_k = 2\,\mathrm{ch}_2(\mathrm{ad}_{\mathbb{C}} P) - 3$.

The bundle: $E \otimes E = S^2E \oplus \Lambda^2 E$

Let E be the rank-2 complex vector bundle associated with P by the fundamental representation, so $k = -c_2(E)[S^4]$ (Nash's (8.127)). Nash finds $\mathrm{ad}_{\mathbb{C}} P$ inside $E \otimes E$. At the level of representations this is the most familiar decomposition in physics, "spin $\frac{1}{2}\otimes$ spin $\frac{1}{2}$ = spin $1\oplus$ spin 0":

$$E \otimes E = S^2E \oplus \Lambda^2 E, \qquad \mathbf{2} \otimes \mathbf{2} = \mathbf{3} \oplus \mathbf{1} \qquad \text{(Nash's (8.128))}. \tag{8.248}$$

Explicitly, with basis $u_1 = \uparrow$, $u_2 = \downarrow$ of the fibre $\mathbb{C}^2$,

$$S^2E:\ u_1 \otimes u_1,\ \tfrac{1}{\sqrt{2}}(u_1 \otimes u_2 + u_2 \otimes u_1),\ u_2 \otimes u_2; \qquad \Lambda^2 E:\ \tfrac{1}{\sqrt{2}}(u_1 \otimes u_2 - u_2 \otimes u_1).$$

Three facts identify the pieces.

1. *Ranks.* S^2E has rank $2 \cdot 3/2 = 3$ and $\Lambda^2 E$ has rank $\binom{2}{2} = 1$: it is a line bundle.
2. *$\Lambda^2 E$ is trivial.* $g \in \mathrm{SU}(2)$ acts on the singlet by $\det g = 1$. So $\Lambda^2 E = \det E$ is the trivial line bundle, and $\mathrm{ch}(\Lambda^2 E) = 1 + c_1(E) = 1$.
3. *S^2E is the adjoint.* The element $-\mathbb{1} \in \mathrm{SU}(2)$ acts on $E \otimes E$ by $(-1)(-1) = +1$. So $E \otimes E$ is really a representation of $\mathrm{SU}(2)/\mathbb{Z}_2 = \mathrm{SO}(3) = \mathrm{Ad}\,\mathrm{SU}(2)$, and its three-dimensional irreducible piece is the adjoint. Concretely, the weights match: $\mathrm{diag}(e^{-i\theta}, e^{i\theta})$ acts on $u_1 \otimes u_1$, on the symmetric middle vector and on $u_2 \otimes u_2$ by $e^{-2i\theta}, 1, e^{2i\theta}$, which are the eigenvalues of Ad found in Example 8.78 (exponentiated). Hence $S^2E \cong \mathrm{ad}_{\mathbb{C}} P$.

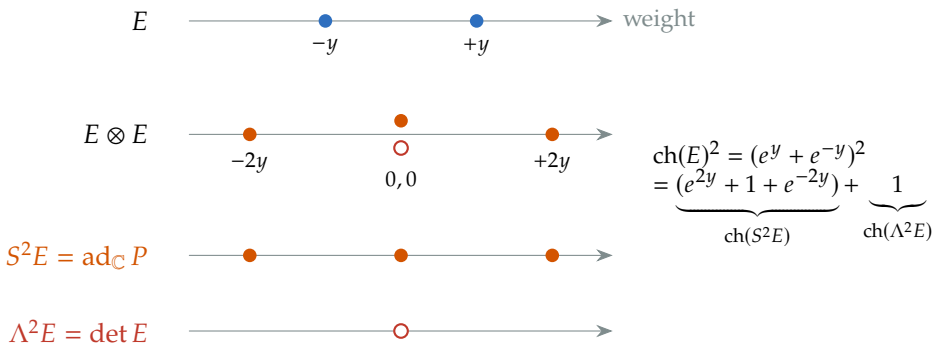


Figure 8.46: Weights (Chern roots) of E, $E \otimes E$ and its two pieces. The roots of a tensor product are the pairwise sums $\pm y \pm y$: this gives $-2y, 0, 0, +2y$. The symmetric part takes $\{-2y, 0, 2y\}$, the same weights as the adjoint representation; the antisymmetric part takes the remaining 0, a trivial line. Squaring the Chern character of E therefore splits into the two Chern characters, which is Nash's (8.129).

The characteristic class computation

Take the Chern character of (8.248). By multiplicativity (8.239), $\mathrm{ch}(E \otimes E) = \mathrm{ch}(E)^2$, so (Nash's (8.129)) $\mathrm{ch}(E)\ \mathrm{ch}(E) = \mathrm{ch}(\mathrm{ad}_{\mathbb{C}} P) + \mathrm{ch}(\Lambda^2 E)$. Expand both sides with (8.241) and keep degree

≤ 4 (Nash's (8.130)–(8.131)):

$$\Big(2 + c_1(E) + \tfrac{1}{2}\big(c_1(E)^2 - 2c_2(E)\big)\Big)^2 \overset{(1)}{=} 3 + \mathrm{ch}_2(\mathrm{ad}_{\mathbb{C}} P) + 1 + c_1(E)$$
$$\Longrightarrow \qquad 4 - 4c_2(E) \overset{(2)}{=} 4 + \mathrm{ch}_2(\mathrm{ad}_{\mathbb{C}} P)$$
$$\Longrightarrow \qquad \mathrm{ch}_2(\mathrm{ad}_{\mathbb{C}} P)[S^4] \overset{(3)}{=} -4c_2(E)[S^4] = 4k. \tag{8.249}$$

(1) Left: (8.241) for the rank-2 bundle E. Right: (8.242) with $\dim G = 3$, plus $\mathrm{ch}(\Lambda^2 E) = 1 + c_1(E)$.
(2) $c_1(E) = 0$ for an SU(2) bundle (traceless curvature), and on S^4 only degrees 0 and 4 exist; squaring $2 - c_2$ gives $4 - 4c_2$ up to degree 8.
(3) Solve, and use $k = -c_2(E)[S^4]$.

In Nash's normalisation, $p_1(\mathrm{ad}_{\mathbb{C}} P) = 2\,\mathrm{ch}_2 = -8c_2(E) = 8k$, his (8.131). The roots give the same thing in one line: if E has roots $\pm y$, then $c_2(E) = -y^2$, and S^2E has roots $\pm 2y, 0$, so $\mathrm{ch}_2(S^2E) = \frac{1}{2}(4y^2 + 4y^2) = 4y^2 = -4c_2(E)$. And it agrees with the trace computation of Example 8.78. Therefore:

Key idea: Nash's (8.132)

$$\dim \mathcal{M}_k = 2 \cdot 4k - 3 \cdot 1 = 8k - 3 \qquad (\mathrm{SU}(2) \text{ instantons on } S^4,\ k \geq 1). \tag{8.250}$$

The framed moduli space $\mathcal{M}'_k$ has dimension $8k$ (add back the 3 dimensions of $\mathrm{Ad}\,\mathrm{SU}(2) = \mathrm{SO}(3)$ acting freely, Section 8.7.2).

Worked example 8.80: Reading $8k - 3$ as collective coordinates

Imagine k BPST instantons so far apart that they do not interact (the **dilute gas**). Each has a position (4), a size (1) and an orientation in colour space (3): 8 parameters each, $8k$ in all. An overall constant colour rotation is a gauge transformation, which removes 3:

$$8k - 3 = \underbrace{4k}_{\text{positions}} + \underbrace{k}_{\text{sizes}} + \underbrace{3(k-1)}_{\text{relative orientations}}.$$

The index theorem says this naive count is exact, not only for far-separated instantons but everywhere in $\mathcal{M}_k$: overlapping instantons do not gain or lose parameters. The 't Hooft ansatz of Section 8.5.5 gives the first two groups, $5k$; it freezes the relative orientations. For $k = 1$ we get 5, the BPST family (y, λ), and for $k = 2$ we get 13.

Worked example 8.81: What the formula says at a reducible connection: $k = 0$

For $k = 0$ the formula gives $8 \cdot 0 - 3 = -3$, which cannot be a dimension. There is no contradiction: the only self-dual connection with $k = 0$ has $\|F\|^2 = 8\pi^2|k| = 0$ (Bogomolny), so it is flat, hence (on the simply connected S^4) the trivial connection. It is maximally reducible: every constant $f \in \mathfrak{su}(2)$ satisfies $\mathrm{d}_A f = \mathrm{d}f = 0$, so $h_0 = 3$. It has no deformations and no obstructions ($h_1 = 0$; $h_2 = 0$ by the Weitzenböck argument, which did not use irreducibility). So

$$h_0 - h_1 + h_2 = 3 - 0 + 0 = 3 = -(8 \cdot 0 - 3),$$

exactly what the index theorem predicts for $\mathrm{index}\, D$. The formula counts $h_1 - h_0 - h_2$, and only equals $\dim \mathcal{M}_k$ when $h_0 = h_2 = 0$. The moduli space $\mathcal{M}_0$ is a single point, of dimension $0 = -3 + h_0$.

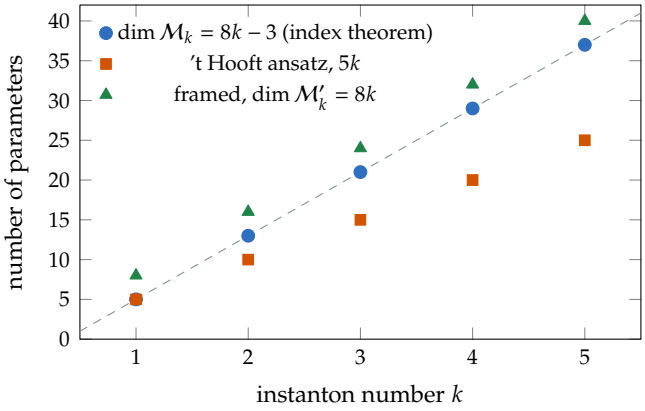


Figure 8.47: Parameter counts for SU(2) instantons on S^4. The index theorem (blue) agrees with the 't Hooft family (orange) only at $k = 1$; the gap $3(k-1)$ is the relative colour orientation of the lumps. The framed count (green) exceeds $\dim \mathcal{M}_k$ by the three global colour rotations.

8.9.7 Other manifolds: Donaldson's class

For $G = \mathrm{SU}(2)$ and a general M (satisfying the hypotheses, or with a generic metric), (8.247) and (8.244) give

$$\dim \mathcal{M}_k = 8k - 3\,(b_0 - b_1 + b_2^-) = 8k - 3\,(1 - b_1 + b_2^-) \tag{8.251}$$

for connected M (the computation of $\mathrm{ch}_2 = 4k$ in (8.249) did not use anything special about S^4). Nash's (8.133)–(8.134) is the special case:

$$\tfrac{1}{2}\big(\chi(M) - \tau(M)\big) \overset{(1)}{=} b_0 - b_1 + b_2^- \overset{(2)}{=} 1 - 0 + 0 = 1, \qquad \text{so} \qquad \dim \mathcal{M}_k = 8k - 3.$$

(1) (8.244).

(2) M connected: $b_0 = 1$. Simply connected: $\pi_1 = 0$, so $H_1(M;\mathbb{Z}) = 0$ and $b_1 = 0$. Positive definite intersection form: no negative eigenvalues, $b_2^- = 0$.

So every simply connected M with positive definite intersection form has the same instanton dimensions as S^4. By Chapter 1, this is exactly the class of manifolds to which Donaldson's theorem applies (Section 8.11).

Worked example 8.82: Expected dimensions on familiar manifolds

Using (8.251) and the Betti numbers of Example 8.79:

M	$1 - b_1 + b_2^-$	$\dim \mathcal{M}_k$	comment
S^4	1	$8k-3$	Nash's (8.132)
$\mathbb{CP}^2$	1	$8k-3$	simply connected, $q = (+1)$
$\mathbb{CP}^2\#\mathbb{CP}^2\#\mathbb{CP}^2$	1	$8k-3$	$q = \mathrm{diag}(1,1,1)$
$S^2 \times S^2$	2	$8k-6$	q indefinite
T^4	0	$8k$	$b_1 = 4$ adds flat directions
$K3$	20	$8k-60$	empty for $k \le 7$ (generic metric)

For $K3$ the "expected dimension" is negative for $k \le 7$: for a generic metric there are no irreducible instantons at all with those charges. The formula is an *expected* dimension; it is the true dimension only at points where $h_0 = h_2 = 0$.

8.9.8 Other gauge groups: Nash's table

For a general compact simple group G the only new input is $\mathrm{ch}_2(\mathrm{ad}_{\mathbb{C}}P)$. Normalise the instanton number so that the BPST instanton, embedded in an SU(2) subgroup of G in the standard way, has $k = 1$. Then the adjoint trace is related to the fundamental one by a number attached to the group, the **dual Coxeter number** $h^\vee$:

$$\mathrm{ch}_2(\mathrm{ad}_{\mathbb{C}}P)[M] = 2h^\vee k, \qquad \text{so on } S^4: \qquad \dim \mathcal{M}_k = 4h^\vee k - \dim G.$$

Worked example 8.83: The $\mathrm{SU}(N)$ row: $\mathrm{tr}_{\mathrm{ad}} = 2N\,\mathrm{tr}_{\mathbf{N}}$

Let X be an $N\times N$ matrix and let $\mathrm{ad}\,X$ act on all matrices Y by $[X,Y]$. Then $(\mathrm{ad}\,X)^2Y = X^2Y - 2XYX + YX^2$. Take the trace of this linear map on the basis E_{ij} (the matrix with a single 1 in place (i,j)):

$$\begin{aligned}\mathrm{tr}_{\mathfrak{gl}(N)}(\mathrm{ad}\,X)^2 &\overset{(1)}{=} \sum_{i,j}\big(X^2E_{ij} - 2XE_{ij}X + E_{ij}X^2\big)_{ij}\\ &\overset{(2)}{=} \sum_{i,j}\big((X^2)_{ii} - 2X_{ii}X_{jj} + (X^2)_{jj}\big)\\ &\overset{(3)}{=} 2N\,\mathrm{tr}(X^2) - 2(\mathrm{tr}\,X)^2.\end{aligned}$$

(1) The trace of a linear map is the sum of the diagonal coefficients $\langle E_{ij}, (\cdot)\rangle$ in the basis E_{ij}.
(2) $(AE_{ij}B)_{ij} = A_{ii}B_{jj}$.
(3) Sum over the free index.

For $X\in\mathfrak{su}(N)$, $\mathrm{tr}\,X = 0$. The identity matrix spans the extra direction of $\mathfrak{gl}(N)$ beyond $\mathfrak{sl}(N)$, and $\mathrm{ad}\,X(\mathbb{1}) = 0$, so it contributes nothing: $\mathrm{tr}_{\mathrm{ad}}(\mathrm{ad}\,X)^2 = 2N\,\mathrm{tr}_{\mathbf{N}}X^2$. For $N = 2$ this is the factor 4 of Example 8.78. Hence $\mathrm{ch}_2(\mathrm{ad}_{\mathbb{C}}P) = 2Nk$, $h^\vee = N$, and

$$\dim\mathcal{M}_k\big(\mathrm{SU}(N), S^4\big) = 4Nk - (N^2-1) = 4Nk - N^2 + 1,$$

the first row of Nash's table.

Doing the same with the other classical and exceptional groups gives Nash's table (8.137), taken from Bernard, Christ, Guth and Weinberg and from Atiyah–Hitchin–Singer. We add the two columns that generate it.

G	$\dim G$	$h^\vee$	$\dim\mathcal{M}_k = 4h^\vee k - \dim G$	irreducible only if
$\mathrm{SU}(N)$	N^2-1	N	$4Nk - N^2 + 1$	$k \ge N/2$
$\mathrm{Spin}(N)$	$N(N-1)/2$	$N-2$	$4(N-2)k - N(N-1)/2$	$k \ge N/4$, $N\ge 7$
$\mathrm{Sp}(N)$	$N(2N+1)$	$N+1$	$4(N+1)k - N(2N+1)$	$k \ge N$
E_6	78	12	$48k - 78$	$k\ge 3$
E_7	133	18	$72k - 133$	$k\ge 3$
E_8	248	30	$120k - 248$	$k\ge 3$
F_4	52	9	$36k - 52$	$k\ge 3$
G_2	14	4	$16k - 14$	$k\ge 2$

Checks: $\mathrm{Sp}(1) = \mathrm{SU}(2) = \mathrm{Spin}(3)$ gives $8k-3$ from the Sp row ($4\cdot 2k - 3$); SU(2) from the SU row gives $8k-3$ too. For $\mathrm{Spin}(N)$ the basic instanton sits in an SU(2) factor of $\mathrm{Spin}(4) = \mathrm{SU}(2)\times\mathrm{SU}(2)\subset\mathrm{Spin}(N)$.

The irreducibility column

Nash stresses that for large G "there may be many subgroups H to which a self-dual connection might reduce" (Section 8.7.3). The last column says when irreducible instantons exist. The mechanism is simple to state for SU(N): a charge-k instanton is built (for example by the ADHM construction of Section 8.10) from k "lumps", each using at most a $\mathbb{C}^2$ of colour space. So its bundle E splits off a trivial flat piece whenever $N > 2k$, and it is then reducible. Irreducible solutions need $N \le 2k$. When the formula is applied below the bound it gives the wrong number, off by h_0:

Worked example 8.84: SU(3) instantons with $k = 1$

The table gives $4 \cdot 3 \cdot 1 - 8 = 4$. But $k = 1 < 3/2$, so every $k = 1$ SU(3) instanton is reducible. In fact every such instanton is a BPST instanton in an SU(2) subgroup, say the upper-left block, and any two are conjugate by a constant element of SU(3). So the true moduli space is again $\{(y, \lambda)\} = \mathbb{R}^4 \times \mathbb{R}^+$, of dimension 5, not 4. The index theorem is not wrong; it computes $h_0 - h_1 + h_2$. At the embedded BPST instanton the stabiliser is the centraliser of the SU(2) block (Section 8.7.3):

$$\left\{\operatorname{diag}(e^{i\theta}, e^{i\theta}, e^{-2i\theta})\right\} \cong \mathrm{U}(1), \qquad h_0 = 1.$$

The obstruction space still vanishes, so

$$h_0 - h_1 + h_2 = 1 - 5 + 0 = -4 = \operatorname{index} D,$$

matching the table's 4 with the sign of (8.236). The same count for SU(4), $k = 1$: the table gives $16 - 15 = 1$; the centraliser of the block $\mathrm{SU}(2) \subset \mathrm{SU}(4)$ is $\{\operatorname{diag}(\mu, \mu, B) : \mu \in \mathrm{U}(1),\ B \in \mathrm{U}(2),\ \mu^2 \det B = 1\}$, of dimension $1 + 4 - 1 = 4$, and $4 - 5 + 0 = -1$. Again right. Framed, the count is uniform: $\dim \mathcal{M}_1' = 4N$ for SU(N) (for $N = 3$: the 5 parameters, plus the SU(3) orbit of the embedded SU(2) frame, $\dim \mathrm{SU}(3) - \dim \mathrm{U}(1) = 7$, total $12 = 4 \cdot 3$).

Reading Nash (pp. 237–241, (8.114)–(8.137))

Nash defines $D = \mathrm{d}_A^{(0)} + (\mathrm{d}_A^-)^*$ (8.114) and states $\operatorname{index} D = -h_1$ (8.115); we proved the kernel/cokernel identification (8.235). (8.116)–(8.118) are (8.237)–(8.240). (8.119)–(8.120) are (8.241)–(8.242), with the warning that Nash's p_1 is twice the usual one. (8.121)–(8.123) are Section 8.9.4; we added the direct characteristic-class evaluation (8.245), which exhibits the degree-0 part -2 that Nash's braces use implicitly. (8.124) is (8.247). (8.125)–(8.132) are Section 8.9.6. (8.133)–(8.134) are Section 8.9.7. (8.137) is the table above, to which we added the columns $\dim G$ and $h^\vee$ that generate it. Nash's reducible-connection remark (8.135)–(8.136) was treated in Section 8.7.5.

Recap

- Roll the complex into $D = \mathrm{d}_A + (\mathrm{d}_A^-)^*$: $\operatorname{Ker} D = H^0 \oplus H^2$, $\operatorname{Coker} D = H^1$, so $\operatorname{index} D = h_0 - h_1 + h_2 = -\dim \mathcal{M}_k$.
- Atiyah–Singer for the complex factorises: $\operatorname{index} D = \{\operatorname{ch}(\mathrm{ad}_{\mathbb{C}} P)\, \bar{E}(M)\}[M]$ with $\operatorname{ch}(\mathrm{ad}_{\mathbb{C}} P) = \dim G + \operatorname{ch}_2$ and $\bar{E}(M) = -2 + \frac{1}{2}(e - \frac{1}{3}p_1)$.
- The untwisted index is $b_0 - b_1 + b_2^- = \frac{1}{2}(\chi - \tau)$ by Poincaré duality.
- $\dim \mathcal{M}_k = 2\operatorname{ch}_2(\mathrm{ad}_{\mathbb{C}} P) - \frac{\dim G}{2}(\chi - \tau)$; $\operatorname{ch}_2 = 4k$ for SU(2) (from $\mathbf{2} \otimes \mathbf{2} = \mathbf{3} \oplus \mathbf{1}$ or from $\mathrm{tr}_3 = 4\,\mathrm{tr}_2$).
- SU(2) on S^4, or on any simply connected positive definite M: $\dim \mathcal{M}_k = 8k - 3 =$

positions + sizes + relative orientations.
- General G on S^4: $4h^\vee k - \dim G$, valid only above the irreducibility bound; below it the count is off by h_0.

8.10 The $k = 1$ moduli space and the ADHM construction

The index theorem told us that $\dim \mathcal{M}_k = 8k - 3$ for SU(2) on S^4. For $k = 1$ that is 5. Nash then makes three further claims in one sentence (p. 241): $\mathcal{M}_1$ is *non-compact*, it has *boundary* S^4, and it is in fact the five-dimensional *hyperbolic space* H^5. He gives no argument. This section supplies one, in four steps.

1. We write down, as explicit formulas, the five tangent vectors to $\mathcal{M}_1$ at the BPST instanton (eight for the framed space $\mathcal{M}_1'$), and check that they solve the linearised equations of Section 8.8. Physicists call them the **zero modes** of the instanton.
2. We find the one mechanism that produces all of them: each zero mode is the response of the instanton to a conformal symmetry of $\mathbb{R}^4$.
3. We compute their lengths. This gives the natural (L^2) metric on $\mathcal{M}_1$, shows that the boundary $\lambda = 0$ is at finite distance, and feeds the collective-coordinate measure of the instanton path integral.
4. We prove that there are no other $k = 1$ instantons, using the **ADHM construction**, which produces every instanton of every charge from finite-dimensional linear algebra.

Throughout, A is the BPST instanton (8.200), $A = \mathrm{Im}(\bar{x}\,\mathrm{d}x)/\phi$ with $\phi = |x|^2 + \lambda^2$, and $F = \lambda^2\,\mathrm{d}\bar{x} \wedge \mathrm{d}x/\phi^2$, so $F_{\mu\nu} = -2\lambda^2\eta^a_{\mu\nu}e_a/\phi^2$ (Section 8.5.3). We work on $\mathbb{R}^4$; by conformal invariance and Uhlenbeck's removable singularity theorem (Section 8.4.1) finite-action instantons on $\mathbb{R}^4$ are the same thing as instantons on S^4.

8.10.1 What a zero mode is

Recall the tangent space to the moduli space (Section 8.8). A tangent vector at $[A]$ is represented by a Lie-algebra-valued 1-form $a = a_\mu\,\mathrm{d}x^\mu$ that

- keeps $A + ta$ self-dual to first order: $\mathrm{d}_A^- a = 0$;
- is orthogonal to the gauge orbit: $\mathrm{d}_A^* a = 0$ (background gauge, (8.212)).

Working definition 8.35 (zero mode)

A **zero mode** of the instanton A is an L^2 [with $\|a\|^2$ finite, so it falls off fast enough at $|x| \to \infty$] 1-form a with values in $\mathfrak{su}(2)$ that satisfies, with $D_\mu = \partial_\mu + [A_\mu, \cdot\,]$,

$$\text{(i)}\quad f_{14} = f_{23},\quad f_{24} = f_{31},\quad f_{34} = f_{12},\qquad \text{where } f_{\mu\nu} = D_\mu a_\nu - D_\nu a_\mu;$$
$$\text{(ii)}\quad D_\mu a_\mu = 0.$$

Condition (i) says the anti-self-dual part of $\mathrm{d}_A a = \tfrac{1}{2} f_{\mu\nu}\,\mathrm{d}x^{\mu\nu}$ vanishes (it is the coefficient of ω_a^- in the basis (8.181)); condition (ii) is $\mathrm{d}_A^* a = -D_\mu a_\mu = 0$. *Test:* four equations (three from (i), one from (ii)) for the four components a_μ, for each of the three colour directions. The space of zero modes is H^1 of the deformation complex [solutions of $\mathrm{d}_A^- a = 0$, with two counted as the same when they differ by a pure gauge $\mathrm{d}_A f$], so its dimension is $h_1 = \dim \mathcal{M}_k$ when $h_0 = h_2 = 0$ (Section 8.8.5).

Why the name? Expand the action around the instanton, $S(A + a) = S(A) + \langle a, \mathcal{K}a\rangle + O(a^3)$,

where in background gauge the fluctuation operator $\mathcal{K}$ is a Laplace-type operator. A zero mode is an eigenvector of $\mathcal{K}$ with eigenvalue 0: a direction in which the action does not change, because it moves us along the valley of minima. In the semiclassical path integral each zero mode must be traded for an integral over a **collective coordinate** (Section 8.10.3).

Watch out: framed and unframed zero modes

For the unframed space $\mathcal{M}_1$ we quotient by all gauge transformations, including those that tend to a non-trivial constant at infinity. For the framed space $\mathcal{M}_1'$ (Section 8.7.2) we quotient only by *based* gauge transformations, $g(\infty) = \mathbb{1}$. So $a = \mathrm{d}_A\Lambda$ with $\Lambda \to 0$ at infinity is never a zero mode, but $a = \mathrm{d}_A\Lambda$ with $\Lambda \to C \neq 0$ at infinity (in a gauge where $A \to 0$ at infinity) is a genuine zero mode of the framed problem. We expect 5 unframed and 8 framed zero modes (Example 8.66).

8.10.2 Zero modes from symmetries

Where do we look for solutions of (i) and (ii)? The instanton is a point in a family; moving along the family is a zero mode. We know two ways to move an instanton: change its parameters, or act on it with a symmetry of the equations. The self-duality equation is invariant under the conformal transformations of $\mathbb{R}^4$ (Section 8.4.1). So every infinitesimal conformal transformation V gives a candidate.

The naive candidate is the Lie derivative $\mathcal{L}_V A$. It is not gauge covariant: under $A \mapsto g^{-1}Ag + g^{-1}\mathrm{d}g$ it does not transform nicely, because $\mathcal{L}_V$ also differentiates g. The fix is to subtract a gauge transformation.

Worked example 8.85: The gauge-covariant derivative of A along V

Let ι_V be the interior product (contraction with V). Then

$$\begin{aligned} \iota_V F &\overset{(1)}{=} \iota_V \mathrm{d}A + \iota_V(A \wedge A) \\ &\overset{(2)}{=} \iota_V \mathrm{d}A + (\iota_V A)\,A - A\,(\iota_V A) \\ &\overset{(3)}{=} \mathcal{L}_V A - \mathrm{d}(\iota_V A) - [A, \iota_V A] \\ &\overset{(4)}{=} \mathcal{L}_V A - \mathrm{d}_A(\iota_V A). \end{aligned} \tag{8.252}$$

(1) $F = \mathrm{d}A + A \wedge A$.
(2) ι_V is an antiderivation: $\iota_V(\alpha \wedge \beta) = (\iota_V\alpha)\beta - \alpha(\iota_V\beta)$ for 1-forms α, β; here $\iota_V A$ is a matrix-valued function.
(3) Cartan's formula $\mathcal{L}_V A = \iota_V \mathrm{d}A + \mathrm{d}\,\iota_V A$ (Chapter T1).
(4) On a matrix-valued function Λ, $\mathrm{d}_A\Lambda = \mathrm{d}\Lambda + [A, \Lambda]$.

So $\iota_V F$ is "move A along the flow of V, then undo the gauge transformation with parameter $\iota_V A$". It is manifestly gauge covariant, because F is. In components, $(\iota_V F)_\mu = V^\nu F_{\nu\mu}$.

Key idea: Which symmetries give zero modes

Let A be self-dual on $\mathbb{R}^4$ and V a conformal Killing vector field. Then $a = \iota_V F$ satisfies the linearised self-duality equation (i). It satisfies the gauge condition (ii) exactly when $F_{\nu\mu}\,\partial_\mu V^\nu = 0$; this holds in particular when the antisymmetric part of the matrix $\partial_\mu V^\nu$ has no self-dual component.

Why (i) holds. Differentiate:

$$\begin{aligned} \mathrm{d}_A(\iota_V F) &\overset{(1)}{=} \mathrm{d}_A(\iota_V F) + \iota_V(\mathrm{d}_A F) \\ &\overset{(2)}{=} \mathrm{d}\iota_V F + \iota_V \mathrm{d}F + [\iota_V A, F] \\ &\overset{(3)}{=} \mathcal{L}_V F + [\iota_V A, F]. \end{aligned}$$

(1) The Bianchi identity $\mathrm{d}_A F = 0$ (Example 8.42); we add zero.
(2) Expand $\mathrm{d}_A\beta = \mathrm{d}\beta + A \wedge \beta + \beta \wedge A$ on the 1-form $\beta = \iota_V F$ and $\mathrm{d}_A F = \mathrm{d}F + A \wedge F - F \wedge A$ on the 2-form F; using the antiderivation rule on $\iota_V(A \wedge F - F \wedge A)$, the terms $A \wedge \iota_V F + \iota_V F \wedge A$ cancel and what survives is $(\iota_V A)F - F(\iota_V A)$.
(3) Cartan's formula on the 2-form F.

Both terms are self-dual. The commutator $[\iota_V A, F]$ multiplies the self-dual 2-form F by a matrix, which does not touch the form part. And $\mathcal{L}_V$ commutes with $\star$ on 2-forms when V is conformal, because the flow of V changes the metric only by a scale factor, and $\star$ on 2-forms in four dimensions does not see that factor. So $\mathcal{L}_V F = \mathcal{L}_V \star F = \star \mathcal{L}_V F$. Hence $\mathrm{d}_A(\iota_V F)$ is self-dual, i.e. $\mathrm{d}_A^-(\iota_V F) = 0$.

When (ii) holds. Take the covariant divergence:

$$\begin{aligned} D_\mu(\iota_V F)_\mu &\overset{(1)}{=} D_\mu\big(V^\nu F_{\nu\mu}\big) \\ &\overset{(2)}{=} (\partial_\mu V^\nu) F_{\nu\mu} + V^\nu D_\mu F_{\nu\mu} \\ &\overset{(3)}{=} (\partial_\mu V^\nu) F_{\nu\mu}. \end{aligned}$$

(1) Components of $\iota_V F$.
(2) Leibniz rule; V^ν is an ordinary function, so D_μ acts on it as ∂_μ.
(3) The Yang–Mills equation $D_\mu F_{\nu\mu} = 0$, which holds for every self-dual field (Section 8.4.2).

Only the antisymmetric part of $\partial_\mu V^\nu$ survives the contraction with $F_{\nu\mu}$. Split it into self-dual and anti-self-dual parts; the anti-self-dual part contracts to zero with the self-dual F by the orthogonality $\eta^a_{\mu\nu}\bar\eta^b_{\mu\nu} = 0$ of (8.183). So (ii) holds when the self-dual part of the "rotation" $\partial_{[\mu}V_{\nu]}$ is orthogonal to F.

The fifteen conformal Killing fields

The conformal Killing fields of $\mathbb{R}^4$ form the Lie algebra $\mathfrak{so}(5,1)$, of dimension 15:

$$\underbrace{P_\nu = \partial_\nu}_{\text{4 translations}}, \qquad \underbrace{D = x^\mu \partial_\mu}_{\text{1 dilation}}, \qquad \underbrace{M_{\mu\nu}}_{\text{6 rotations}}, \qquad \underbrace{K_b = 2(b\cdot x)x^\mu\partial_\mu - |x|^2 b^\mu \partial_\mu}_{\text{4 special conformal}}.$$

Quaternions split the rotations into two halves. Every rotation of $\mathbb{R}^4$ is $x \mapsto p\,x\,\bar q$ with unit quaternions p, q (Section 8.5), so the six rotation generators are

$$V_C^{\mathrm{L}}(x) = Cx \quad \text{(left)}, \qquad V_C^{\mathrm{R}}(x) = xC \quad \text{(right)}, \qquad C \in \operatorname{Im}\mathbb{H}.$$

(Here a vector field on $\mathbb{R}^4 = \mathbb{H}$ is written as a quaternion-valued function.)

Worked example 8.86: Left rotations are anti-self-dual, right rotations self-dual

Take $C = e_1$. With $x = x^4 + x^1 e_1 + x^2 e_2 + x^3 e_3$ and the rules (8.193),

$$\begin{aligned} e_1 x &\overset{(1)}{=} x^4 e_1 + x^1 e_1^2 + x^2 e_1 e_2 + x^3 e_1 e_3 \\ &\overset{(2)}{=} -x^1 + x^4 e_1 - x^3 e_2 + x^2 e_3. \end{aligned}$$

(1) Distribute.
(2) $e_1^2 = -1$, $e_1 e_2 = e_3$, $e_1 e_3 = -e_3 e_1 = -e_2$.

So $V^{\mathrm{L}} = (V^1, V^2, V^3, V^4) = (x^4, -x^3, x^2, -x^1)$. The matrix $\partial_\mu V^\nu$ has entries $\partial_4 V^1 = 1$, $\partial_1 V^4 = -1$, $\partial_3 V^2 = -1$, $\partial_2 V^3 = 1$. As a 2-form $\frac{1}{2}\partial_\mu V_\nu \, \mathrm{d}x^\mu \wedge \mathrm{d}x^\nu$ this is $-\mathrm{d}x^{14} + \mathrm{d}x^{23} = \omega_1^-$: anti-self-dual. The same computation with $x e_1 = -x^1 + x^4 e_1 + x^3 e_2 - x^2 e_3$ gives $V^{\mathrm{R}} = (x^4, x^3, -x^2, -x^1)$ and the 2-form $-\mathrm{d}x^{14} - \mathrm{d}x^{23} = -\omega_1^+$: self-dual.

Now we run through the list with the BPST instanton.

Worked example 8.87: Translations: four zero modes

$V = P_\nu = \partial_\nu$ has $\partial_\mu V^\nu = 0$, so both (i) and (ii) hold. The mode is

$$a^{(\nu)} = \iota_{\partial_\nu} F = F_{\nu\mu} \, \mathrm{d}x^\mu = -\frac{2\lambda^2 \, \eta^a_{\nu\mu} \, e_a}{\phi^2} \, \mathrm{d}x^\mu.$$

By (8.252), $a^{(\nu)} = \partial_\nu A - \mathrm{d}_A(A_\nu)$. The first term is minus the derivative of the translated instanton $A(x - y)$ with respect to the centre y^ν; the second is an infinitesimal gauge transformation with parameter A_ν, which tends to zero at infinity ($A_\nu = O(1/r)$), so it is a *based* gauge transformation. Up to that correction, $a^{(\nu)}$ is "move the centre". It is in L^2 because $F = O(r^{-4})$. In fact it decays so fast that we can compute its norm at once (next subsection).

Worked example 8.88: The dilation: one zero mode

$V = D = x^\mu \partial_\mu$ has $\partial_\mu V^\nu = \delta^\nu_\mu$, which is symmetric, so (ii) holds. The mode is

$$\begin{aligned} \iota_D F &\overset{(1)}{=} x^\nu F_{\nu\mu} \, \mathrm{d}x^\mu \\ &\overset{(2)}{=} \frac{2\lambda^2}{\phi^2} \, \eta^a_{\mu\nu} x^\nu e_a \, \mathrm{d}x^\mu \\ &\overset{(3)}{=} \frac{2\lambda^2 \, \mathrm{Im}(\bar{x} \, \mathrm{d}x)}{\phi^2} \\ &\overset{(4)}{=} -\lambda \, \frac{\partial A}{\partial \lambda}. \end{aligned}$$

(1) Components of $\iota_V F$ with $V^\nu = x^\nu$.
(2) $F_{\nu\mu} = -2\lambda^2 \eta^a_{\nu\mu} e_a / \phi^2$ and $\eta^a_{\nu\mu} = -\eta^a_{\mu\nu}$.
(3) The third form of (8.200): $\mathrm{Im}(\bar{x} \, \mathrm{d}x) = \eta^a_{\mu\nu} x^\nu e_a \, \mathrm{d}x^\mu$.
(4) $A = \mathrm{Im}(\bar{x} \, \mathrm{d}x)/(|x|^2 + \lambda^2)$, so $\partial_\lambda A = -2\lambda \, \mathrm{Im}(\bar{x} \, \mathrm{d}x)/\phi^2$.

So the dilation mode is exactly the derivative with respect to the size, with no gauge correction. This is no accident: $x \mapsto sx$ maps the instanton of size λ to the one of size λ/s.

One can also check (ii) directly. With $a_\mu = \partial_\lambda A_\mu = h(r)\,\eta^a_{\mu\nu}x^\nu e_a$ and $h = -2\lambda/\phi^2$,

$$\partial_\mu a_\mu = h'(r)\frac{x^\mu}{r}\eta^a_{\mu\nu}x^\nu e_a + h\,\eta^a_{\mu\mu}e_a = 0, \qquad [A_\mu, a_\mu] = h\phi\,[A_\mu, A_\mu] = 0,$$

the first because $\eta^a_{\mu\nu}$ is antisymmetric and $x^\mu x^\nu$ symmetric, the second because $a_\mu = h\phi\,A_\mu$ is a multiple of A_μ for each μ.

Worked example 8.89: Left rotations: three framing modes

Take $V = V^{\mathrm{L}}_C = Cx$ with $C \in \operatorname{Im}\mathbb{H}$. By Example 8.86 the rotation part of V is anti-self-dual, so (ii) holds. To identify the mode, compute the two terms of (8.252). The flow of V is $x \mapsto p_t x$ with $p_t = e^{tC}$ a unit quaternion, and

$$\text{(pull-back of } A \text{ by } x \mapsto p_t x) \overset{(1)}{=} \frac{\operatorname{Im}\big(\overline{p_t x}\,\mathrm{d}(p_t x)\big)}{|p_t x|^2 + \lambda^2} \overset{(2)}{=} \frac{\operatorname{Im}(\bar{x}\,\bar{p}_t p_t\,\mathrm{d}x)}{|x|^2 + \lambda^2} = A.$$

(1) Substitute $p_t x$ for x in (8.200).
(2) $\overline{p_t x} = \bar{x}\bar{p}_t$, $\mathrm{d}(p_t x) = p_t\,\mathrm{d}x$ (p_t is constant), $|p_t x| = |x|$, and $\bar{p}_t p_t = 1$.

So $\mathcal{L}_V A = 0$: the BPST instanton is *literally* invariant under left rotations. Then (8.252) gives

$$\iota_V F = -\mathrm{d}_A \Lambda_C, \qquad \Lambda_C = \iota_V A = \frac{\operatorname{Im}(\bar{x}\,Cx)}{\phi} = \frac{\bar{x}\,C\,x}{|x|^2 + \lambda^2}$$

($\bar{x}Cx$ is already imaginary, since its conjugate is $\bar{x}\bar{C}x = -\bar{x}Cx$). So this mode is an infinitesimal gauge transformation. Is it a based one? At large $|x|$, $\Lambda_C \to \bar{x}Cx/|x|^2$, which is not zero; it is the constant C conjugated by $g = x/|x|$. Transform to the singular gauge (Section 8.5.4), in which $A \to 0$ at infinity; there the gauge parameter becomes

$$\Lambda^{\mathrm{sing}}_C = g\,\Lambda_C\,g^{-1} = \frac{x}{r}\frac{\bar{x}Cx}{\phi}\frac{\bar{x}}{r} = \frac{r^2}{r^2+\lambda^2}\,C \longrightarrow C.$$

So $\iota_V F$ is the infinitesimal gauge transformation that rotates the frame at infinity by C: a zero mode of the *framed* problem, and pure gauge in the unframed one. Three of them, one for each direction of C. Its gauge condition (ii) reads $D_\mu D_\mu \Lambda_C = 0$: Λ_C is a covariantly harmonic function that tends to the constant C at infinity.

Worked example 8.90: Right rotations: no zero modes

Take $V = V^{\mathrm{R}}_C = xC$, with flow $x \mapsto xq_t$, $q_t = e^{tC}$. Now

$$\text{(pull-back of } A) = \frac{\operatorname{Im}(\bar{q}_t\bar{x}\,\mathrm{d}x\,q_t)}{\phi} = \bar{q}_t A q_t,$$

because conjugation by a unit quaternion commutes with taking the imaginary part. Differentiating at $t = 0$, with $q_t \approx 1 + tC$:

$$\iota_V F \overset{(1)}{=} \mathcal{L}_V A - \mathrm{d}_A(\iota_V A) \overset{(2)}{=} [A, C] - \mathrm{d}_A\Big(\frac{\operatorname{Im}(\bar{x}\,xC)}{\phi}\Big)$$

$$\overset{(3)}{=} \mathrm{d}_A C - \mathrm{d}_A\Big(\frac{r^2 C}{\phi}\Big)$$
$$\overset{(4)}{=} \mathrm{d}_A\Big(\frac{\lambda^2}{r^2+\lambda^2}\, C\Big).$$

(1) (8.252).
(2) $\bar{q}_t A q_t \approx (1 - tC)A(1 + tC) = A + t[A, C]$; and $\iota_V A = \mathrm{Im}(\bar{x}V)/\phi$.
(3) C is constant, so $\mathrm{d}_A C = \mathrm{d}C + [A, C] = [A, C]$; and $\bar{x}x = r^2$.
(4) $1 - r^2/\phi = \lambda^2/\phi$.

The gauge parameter $\lambda^2 C/\phi$ vanishes at infinity, in either gauge. So $\iota_V F$ is a *based* gauge transformation: zero in H^1, not a zero mode. Consistently, it fails the gauge condition (ii), since its rotation part is self-dual (Example 8.86) and pairs non-trivially with F. This is the spin–isospin locking of Example 8.65 seen infinitesimally: a right rotation of space is undone by a colour rotation.

Special conformal transformations. For K_b the matrix $\partial_\mu V^\nu = 2b_\mu x^\nu - 2x_\mu b^\nu + 2(b\cdot x)\delta^\nu_\mu$ has an antisymmetric part $2(b_\mu x^\nu - x_\mu b^\nu)$ with both self-dual and anti-self-dual pieces, so (ii) fails. Modulo gauge, K_b moves the instanton the way a translation by $\lambda^2 b$ does; we will see this in Example 8.94. So $P_b - \lambda^{-2}K_b$ fixes the instanton, and the special conformal transformations add nothing new.

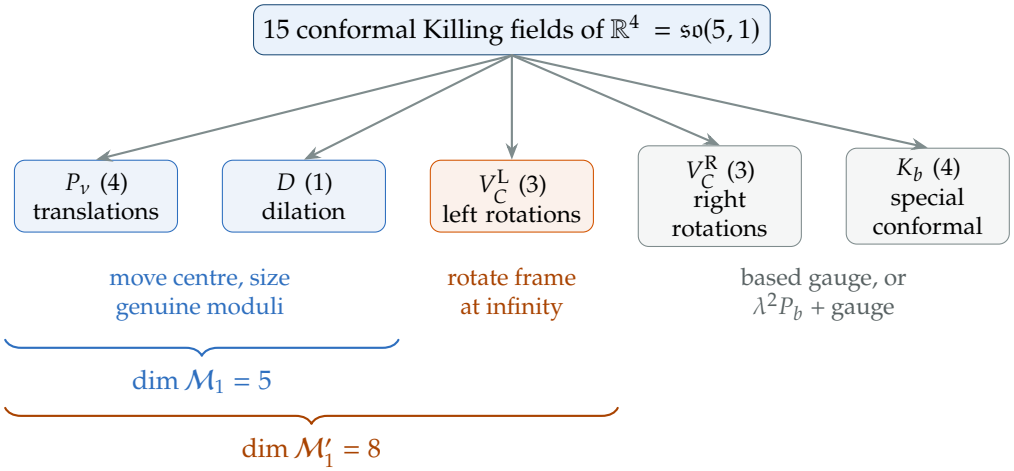


Figure 8.48: Zero-mode bookkeeping for the BPST instanton. Each conformal Killing field V gives the gauge-covariant variation $\iota_V F$. Translations and the dilation give the five genuine moduli; left rotations only turn the frame at infinity, so they count in the framed space $\mathcal{M}_1'$ but not in $\mathcal{M}_1$; the remaining seven generators fix the instanton up to gauge. The index theorem says there are no other zero modes.

Physical picture: the instanton is a highly symmetric object

Out of the 15-dimensional conformal group, a 10-dimensional subgroup (generated by the 6 rotations and the 4 combinations $P_b - \lambda^{-2}K_b$) fixes the unframed instanton. On the unit S^4 with $\lambda = 1$ this subgroup is the rotation group SO(5) of the sphere: the instanton is "round". The orbit of the instanton under the conformal group therefore has dimension $15 - 10 = 5$, which is the whole of $\mathcal{M}_1$. We return to this in Section 8.10.4.

8.10.3 The lengths of the zero modes: the metric on $\mathcal{M}_1$

The space of connections carries the flat L^2 metric $\langle a, b\rangle = -\int \mathrm{tr}(a \wedge \star b) = -\int \mathrm{tr}(a_\mu b_\mu)\,\mathrm{d}^4x$ ((8.165)). Restricting it to zero modes (which are orthogonal to the gauge orbits) gives a Riemannian metric on the moduli space, the L^2 **metric**. We now compute it for $\mathcal{M}_1$ and $\mathcal{M}_1'$.

Two tools. First, $\mathrm{tr}(e_a e_b) = -2\delta_{ab}$ (since $e_a = -i\sigma_a$). Second, contracting the third identity of (8.183) over $a = b$,

$$\eta^a_{\mu\nu}\eta^a_{\mu\rho} = 3\,\delta_{\nu\rho} \tag{8.253}$$

(the ε^{aac} term vanishes). And the radial integrals, with $u = r^2$ and $\mathrm{d}^4x = 2\pi^2 r^3\,\mathrm{d}r$ (the unit S^3 has volume $2\pi^2$):

$$\int_{\mathbb{R}^4} \frac{r^{2m}\,\mathrm{d}^4x}{(r^2+\lambda^2)^4} = \pi^2 \int_0^\infty \frac{u^{m+1}\,\mathrm{d}u}{(u+\lambda^2)^4} = \pi^2\lambda^{2m-4}\,\frac{(m+1)!\,(1-m)!}{3!} \qquad (m = 0, 1), \tag{8.254}$$

by the Beta integral $\int_0^\infty u^{p-1}(u+c)^{-p-q}\,\mathrm{d}u = c^{-q}\,\Gamma(p)\Gamma(q)/\Gamma(p+q)$ with $p = m+2$, $q = 2-m$. So the integral is $\pi^2/(6\lambda^4)$ for $m = 0$ and $\pi^2/(3\lambda^2)$ for $m = 1$.

Worked example 8.91: Norm of a translation mode

For fixed ν,

$$\begin{aligned}
\left\|a^{(\nu)}\right\|^2 &\overset{(1)}{=} -\int \mathrm{tr}\big(F_{\nu\mu}F_{\nu\mu}\big)\mathrm{d}^4x \\
&\overset{(2)}{=} -\int \frac{4\lambda^4}{\phi^4}\,\eta^a_{\nu\mu}\eta^b_{\nu\mu}\,\mathrm{tr}(e_a e_b)\,\mathrm{d}^4x \\
&\overset{(3)}{=} \int \frac{8\lambda^4}{\phi^4}\,\eta^a_{\nu\mu}\eta^a_{\nu\mu}\,\mathrm{d}^4x \\
&\overset{(4)}{=} 24\lambda^4 \int \frac{\mathrm{d}^4x}{(r^2+\lambda^2)^4} \\
&\overset{(5)}{=} 24\lambda^4\cdot\frac{\pi^2}{6\lambda^4} = 4\pi^2.
\end{aligned}$$

(1) Definition of the norm, with $(a^{(\nu)})_\mu = F_{\nu\mu}$ (sum over μ only).
(2) $F_{\nu\mu} = -2\lambda^2\eta^a_{\nu\mu}e_a/\phi^2$.
(3) $\mathrm{tr}(e_a e_b) = -2\delta_{ab}$.
(4) (8.253) with $\nu = \rho$ (no sum over ν): the sum over a, μ is 3.
(5) (8.254) with $m = 0$.

So $\left\|a^{(\nu)}\right\|^2 = 4\pi^2 = \frac{1}{2}S$, half the instanton action. (Sum over ν: $\sum_\nu \left\|a^{(\nu)}\right\|^2 = \|F\|^2\cdot 2 = 16\pi^2$, since $\sum_{\mu\nu} F_{\nu\mu}F_{\nu\mu}$ counts each component twice. A good check.) Different translation modes are orthogonal: $\eta^a_{\nu\mu}\eta^a_{\rho\mu} = 3\delta_{\nu\rho}$.

Worked example 8.92: Norm of the dilation mode

With $a = \partial_\lambda A$, $a_\mu = -2\lambda\,\eta^a_{\mu\nu}x^\nu e_a/\phi^2$:

$$\begin{aligned}
\|\partial_\lambda A\|^2 &\overset{(1)}{=} -\int \frac{4\lambda^2}{\phi^4}\,\eta^a_{\mu\nu}\eta^b_{\mu\rho}x^\nu x^\rho\ \mathrm{tr}(e_a e_b)\,\mathrm{d}^4x \\
&\overset{(2)}{=} \int \frac{8\lambda^2}{\phi^4}\cdot 3\,\delta_{\nu\rho}x^\nu x^\rho\,\mathrm{d}^4x
\end{aligned}$$

$$\overset{(3)}{=} 24\lambda^2 \int \frac{r^2\,\mathrm{d}^4x}{(r^2+\lambda^2)^4}$$
$$\overset{(4)}{=} 24\lambda^2 \cdot \frac{\pi^2}{3\lambda^2} = 8\pi^2.$$

(1) Definition of the norm and the formula for a_μ.
(2) $\operatorname{tr}(e_a e_b) = -2\delta_{ab}$, then (8.253).
(3) $\delta_{\nu\rho}x^\nu x^\rho = r^2$.
(4) (8.254) with $m = 1$.

The dilation mode is orthogonal to each translation mode: the integrand $-\operatorname{tr}(F_{\nu\mu}\partial_\lambda A_\mu)$ is odd under $x \mapsto -x$ (F is even, $\partial_\lambda A$ is odd), so it integrates to zero.

Worked example 8.93: Norm of a framing mode

For $V = Cx$ with $|C| = 1$, the mode is $a = \iota_V F$, $a_\mu = V^\nu F_{\nu\mu} = -2\lambda^2 V^\nu \eta^a_{\nu\mu} e_a/\phi^2$. Then pointwise

$$-\operatorname{tr}(a_\mu a_\mu) \overset{(1)}{=} \frac{8\lambda^4}{\phi^4}\,\eta^a_{\nu\mu}\eta^a_{\rho\mu}V^\nu V^\rho$$
$$\overset{(2)}{=} \frac{24\lambda^4}{\phi^4}\,|V|^2$$
$$\overset{(3)}{=} \frac{24\lambda^4\, r^2}{\phi^4},$$

(1) $\operatorname{tr}(e_a e_b) = -2\delta_{ab}$.
(2) (8.253) (with the antisymmetry of η used twice).
(3) $|V| = |Cx| = |C||x| = r$.

so, by (8.254) with $m = 1$, $\|\iota_V F\|^2 = 24\lambda^4 \cdot \pi^2/(3\lambda^2) = 8\pi^2\lambda^2$.
The same two steps work for *any* two vector fields V, W:

$$-\operatorname{tr}\big((\iota_V F)_\mu (\iota_W F)_\mu\big) = \frac{24\lambda^4}{\phi^4}\,V\cdot W. \tag{8.255}$$

All three norm computations are special cases: $V = W = \partial_\nu$ gives the translation density $24\lambda^4/\phi^4$, and $V = W = x$ (with $\partial_\lambda A = -\iota_x F/\lambda$) gives the dilation density $24\lambda^2 r^2/\phi^4$. It also settles orthogonality. A translation (V constant) against a dilation or framing mode (W linear in x): $V \cdot W$ is odd under $x \mapsto -x$, so the integral vanishes. Dilation against framing: $x \cdot Cx = \operatorname{Re}(\bar{x}\,Cx) = 0$ pointwise. Two framing modes: $Cx \cdot C'x = \operatorname{Re}(\overline{Cx}\,C'x) = r^2\,C\cdot C'$, zero for $C \perp C'$. So the eight modes with $C \in \{e_1, e_2, e_3\}$ are mutually orthogonal.

Key idea: The L^2 metric on the $k = 1$ moduli spaces of $\mathbb{R}^4$

In the coordinates (y, λ) of $\mathcal{M}_1 = \mathbb{R}^4 \times \mathbb{R}^+$, and with the framing $q \in \mathrm{SU}(2)/\{\pm 1\}$ for $\mathcal{M}_1'$,

$$\begin{aligned} \mathrm{d}s^2_{\mathcal{M}_1} &= 4\pi^2\,|\mathrm{d}y|^2 + 8\pi^2\,\mathrm{d}\lambda^2, \\ \mathrm{d}s^2_{\mathcal{M}_1'} &= 4\pi^2\,|\mathrm{d}y|^2 + 8\pi^2\big(\mathrm{d}\lambda^2 + \lambda^2|\mathrm{d}q|^2\big) = 4\pi^2\,|\mathrm{d}y|^2 + 8\pi^2\,|\mathrm{d}(\lambda q)|^2. \end{aligned} \tag{8.256}$$

Both are *flat*. The unframed space is a flat half-space; the framed space is $\mathbb{R}^4$ times the flat cone $\mathbb{H}^\times/\{\pm 1\} = (\mathbb{R}^4 \setminus 0)/\mathbb{Z}_2$, whose tip $\lambda q = 0$ is missing.

Here $|\mathrm{d}q|^2$ is the round metric of the unit $S^3 = \mathrm{SU}(2)$: moving q by $q\,e^{\theta C}$ has speed $|C| = 1$, matching the framing mode of norm $\sqrt{8}\,\pi\lambda$. The combination $\mathrm{d}\lambda^2 + \lambda^2|\mathrm{d}q|^2$ is flat $\mathbb{R}^4$ in polar coordinates, which is why the size and the framing merge into one quaternion λq (Example 8.66); we will meet the same quaternion as ADHM data in Section 8.10.6.

Physical picture: the boundary is at finite distance

In the metric (8.256) the distance from the instanton of size λ_0 to the "instanton of size 0" is $\sqrt{8}\,\pi\lambda_0$: finite. The moduli space is *incomplete*. A path in it can run off the edge in finite time, and what it runs into is a delta function of action density $8\pi^2\delta_y$ sitting at a point y of $\mathbb{R}^4$. This is the concrete meaning of Nash's "$\mathcal{M}_1$ is a non-compact space with boundary S^4". The other end, $\lambda \to \infty$, is at infinite distance.

Watch out: the metric depends on the metric of spacetime

The L^2 norm of a 2-form is conformally invariant in four dimensions, but that of a 1-form is not. So the metric (8.256) is the one induced by flat $\mathbb{R}^4$. On the round S^4 the L^2 metric on $\mathcal{M}_1$ is different (Groisser and Parker computed it): it is still invariant under the rotations SO(5) and still incomplete, with the boundary S^4 at finite distance, but it is neither flat nor the hyperbolic metric. What is canonical is the *action* of the conformal group, which we use next.

Dictionary: collective coordinates in the path integral

In the one-instanton contribution to $\int \mathcal{D}A\, e^{-S/g^2}$, the Gaussian integral over each zero mode would diverge. One changes variables from the zero-mode amplitudes to the moduli. If $\partial_i A$ are the zero modes with norms $\|\partial_i A\|$, the Jacobian is $\prod_i \|\partial_i A\| / (g\sqrt{2\pi})$ (the factor g because the action is S/g^2). For the framed $k = 1$ space, using the norms above,

$$\prod_{i=1}^{8} \frac{\|\partial_i A\|}{g\sqrt{2\pi}} \overset{(1)}{=} \frac{(2\pi)^4\,(\sqrt{8}\,\pi)\,(\sqrt{8}\,\pi\lambda)^3}{g^8(2\pi)^4} \overset{(2)}{=} \frac{64\pi^4\lambda^3}{g^8} = \frac{1}{64\pi^4}\Big(\frac{8\pi^2}{g^2}\Big)^4 \lambda^3.$$

(1) Four translation norms 2π, one dilation norm $\sqrt{8}\pi$, three framing norms $\sqrt{8}\pi\lambda$.
(2) $(\sqrt{8})^4 = 64$.

So the measure on $\mathcal{M}_1'$ is $\propto (8\pi^2/g^2)^4\,\lambda^3\,\mathrm{d}^4y\,\mathrm{d}\lambda\,\mathrm{d}^3q$: one power of $\sqrt{S_0}$ per zero mode, and $\lambda^3\mathrm{d}\lambda\,\mathrm{d}^3q$ is just the flat volume element $\mathrm{d}^4(\lambda q)$ of (8.256). Combined with the determinant of the non-zero modes this becomes 't Hooft's instanton density $\mathrm{d}^4y\,\mathrm{d}\lambda\,\lambda^{-5}\,(8\pi^2/g^2)^4\,e^{-8\pi^2/g^2(1/\lambda)}$, where the running coupling turns $e^{-8\pi^2/g^2}$ into $(\lambda\Lambda_{\mathrm{QCD}})^{22/3}$ for SU(2). The integral over λ then diverges at large λ: this is the infrared problem of Section 8.5.3, and it is the non-compact end $\lambda \to \infty$ of $\mathcal{M}_1$.

8.10.4 The global shape: $\mathcal{M}_1 = \mathbb{R}^4 \times \mathbb{R}^+ = H^5$

Suppose, as the ADHM construction will prove, that every $k = 1$ instanton on S^4 is a BPST instanton, determined up to gauge by its centre $y \in \mathbb{R}^4$ and size $\lambda > 0$ (Example 8.64). Then

$$\mathcal{M}_1 = \{(y, \lambda)\} = \mathbb{R}^4 \times \mathbb{R}^+,$$

the upper half-space of $\mathbb{R}^5$. What does Nash mean by calling it H^5?

Working definition 8.36 (hyperbolic space H^5)

The **upper half-space model** is $\{(y,\lambda)\in\mathbb{R}^4\times\mathbb{R}^+\}$ with the metric

$$\mathrm{d}s^2_{H^5}=\frac{|\mathrm{d}y|^2+\mathrm{d}\lambda^2}{\lambda^2}.$$

Its group of orientation-preserving isometries [maps of the half-space to itself that preserve $\mathrm{d}s^2_{H^5}$ and the orientation] is $\mathrm{SO}_0(5,1)$ (the identity component [the elements that can be joined to $\mathbb{1}$ by a continuous path] of $\mathrm{SO}(5,1)$), generated by translations and rotations of y, the scalings $(y,\lambda)\mapsto(sy,s\lambda)$, and the map $(y,\lambda)\mapsto(\bar y,\lambda)/(|y|^2+\lambda^2)$: the inversion in the unit sphere of $\mathbb{R}^5$ followed by the reflection $y\mapsto\bar y$. The **boundary at infinity** [the limit points, which lie at infinite distance in this metric] is the plane $\lambda=0$ together with one point at infinity: a copy of S^4. *Ball model:* the same space is the open unit ball B^5 with metric $4|\mathrm{d}X|^2/(1-|X|^2)^2$, whose boundary is the unit S^4.

The claim is that the conformal group of S^4 acts on $\mathcal{M}_1$ exactly as the isometry group of H^5 acts on the upper half-space. Translations of $\mathbb{R}^4$ move y. Rotations of $\mathbb{R}^4$ rotate y (and fix λ). The dilation $x\mapsto sx$ maps (y,λ) to $(sy,s\lambda)$. It remains to check an inversion, which together with these generates the conformal group. One care is needed: the inversion $I(x)=x/|x|^2$ *reverses* orientation (its differential is $|x|^{-2}$ times a reflection), so it turns self-dual fields into anti-self-dual ones. The orientation-preserving inversion is $x\mapsto x^{-1}=\bar x/|x|^2$, which is I followed by the reflection $x\mapsto\bar x$ (itself orientation reversing, as it flips three axes).

Worked example 8.94: The inversion acts on (y,λ) as inversion in the unit sphere of $\mathbb{R}^5$

Substitute $x=I(x')=x'/|x'|^2$. The instanton is determined up to gauge by its action density $s\,\mathrm{d}^4x$ with $s=48\lambda^4/(|x-y|^2+\lambda^2)^4$ (Example 8.64). This density is a conformally invariant measure (Section 8.4.1), so under the conformal map I it transforms as $s(x)\,\mathrm{d}^4x\mapsto s(I(x'))\,|\det\mathrm{d}I|\,\mathrm{d}^4x'$. Put $R^2=|y|^2+\lambda^2$, $y'=y/R^2$, $\lambda'=\lambda/R^2$. Then

$$\begin{aligned}|x-y|^2+\lambda^2 &\overset{(1)}{=}\frac{1}{|x'|^2}-\frac{2\,x'\cdot y}{|x'|^2}+R^2\\ &\overset{(2)}{=}\frac{R^2}{|x'|^2}\Big(|x'|^2-\frac{2\,x'\cdot y}{R^2}+\frac{1}{R^2}\Big)\\ &\overset{(3)}{=}\frac{R^2}{|x'|^2}\Big(|x'-y'|^2+\frac{1}{R^2}-\frac{|y|^2}{R^4}\Big)\\ &\overset{(4)}{=}\frac{R^2}{|x'|^2}\big(|x'-y'|^2+\lambda'^2\big).\end{aligned}$$

(1) Expand $|x-y|^2=|x|^2-2x\cdot y+|y|^2$ with $|x|^2=1/|x'|^2$ and $x\cdot y=x'\cdot y/|x'|^2$.
(2) Factor out $R^2/|x'|^2$.
(3) Complete the square: $|x'-y'|^2=|x'|^2-2x'\cdot y/R^2+|y|^2/R^4$.
(4) $1/R^2-|y|^2/R^4=\lambda^2/R^4=\lambda'^2$.

The inversion has $|\det\mathrm{d}I|=|x'|^{-8}$ (it is conformal with scale factor $|x'|^{-2}$, raised to the power 4). So

$$\begin{aligned}\frac{48\lambda^4\,\mathrm{d}^4x}{(|x-y|^2+\lambda^2)^4} &\overset{(1)}{=}\frac{48\lambda^4\,|x'|^8}{R^8\,(|x'-y'|^2+\lambda'^2)^4}\cdot\frac{\mathrm{d}^4x'}{|x'|^8}\\ &\overset{(2)}{=}\frac{48\lambda'^4\,\mathrm{d}^4x'}{(|x'-y'|^2+\lambda'^2)^4}.\end{aligned}$$

(1) The previous chain and the Jacobian.
(2) $\lambda^4/R^8 = (\lambda/R^2)^4 = \lambda'^4$.

So I carries the density of the instanton (y, λ) to the density of the (anti-)instanton with parameters

$$(y', \lambda') = \frac{(y, \lambda)}{|y|^2 + \lambda^2},$$

the inversion of the point $(y, \lambda) \in \mathbb{R}^5$ in the unit sphere. Following it by the reflection $x \mapsto \bar{x}$ restores self-duality and replaces y' by $\bar{y}'$. So the orientation-preserving inversion $x \mapsto x^{-1}$ acts on $\mathcal{M}_1$ by $(y, \lambda) \mapsto (\bar{y}, \lambda)/(|y|^2+\lambda^2)$, an isometry of H^5. For instance the instanton $(0, \lambda)$ goes to $(0, 1/\lambda)$: inversion exchanges small and large instantons, and fixes the instanton of size 1 centred at 0. The special conformal transformation is $K_b = I \circ P_b \circ I$ (two inversions, so orientation is preserved); on $(0, \lambda)$ it gives $(0, \lambda) \mapsto (0, 1/\lambda) \mapsto (\varepsilon b, 1/\lambda) \mapsto (\varepsilon\lambda^2 b, \lambda)+O(\varepsilon^2)$, a translation by $\lambda^2\varepsilon b$, as promised in Section 8.10.2.

Key idea: $\mathcal{M}_1$ is hyperbolic 5-space

The conformal group $\mathrm{SO}(5,1)$ of S^4 acts on $\mathcal{M}_1$ transitively, by the isometries of the hyperbolic metric $(|\mathrm{d}y|^2 + \mathrm{d}\lambda^2)/\lambda^2$. The stabiliser of the instanton $(0, 1)$ is $\mathrm{SO}(5)$. Hence

$$\mathcal{M}_1 \cong \mathrm{SO}(5,1)/\mathrm{SO}(5) = H^5, \qquad \partial\mathcal{M}_1 = S^4.$$

The hyperbolic metric is the unique (up to scale) metric invariant under the conformal group; it is *not* the L^2 metric (Section 8.10.3).

Worked example 8.95: The centre of H^5 is the round instanton

Take $y = 0$, $\lambda = 1$ and carry the instanton to the unit S^4 by stereographic projection, $g_{S^4} = 4(1 + r^2)^{-2} g_{\mathbb{R}^4}$ (8.184). The volume form of S^4 is then $\mathrm{vol}_{S^4} = 16(1 + r^2)^{-4}\mathrm{d}^4x$, so

$$s\,\mathrm{d}^4x \overset{(1)}{=} \frac{48\,\mathrm{d}^4x}{(1 + r^2)^4} \overset{(2)}{=} 3\,\mathrm{vol}_{S^4}.$$

(1) The action density with $\lambda = 1$, $y = 0$.
(2) $48 = 3 \cdot 16$.

The action density is *uniform* on the sphere. Check the total: $3\,\mathrm{Vol}(S^4) = 3 \cdot \frac{8\pi^2}{3} = 8\pi^2$. This is the $\mathrm{SO}(5)$-invariant instanton, the "centre" of H^5. Moving away from the centre towards a boundary point $p \in S^4$ concentrates the density near p.

Try it: Where is the instanton on the sphere?

Donaldson's **centre-of-mass map** sends an instanton on the unit $S^4 \subset \mathbb{R}^5$ to $c(A) = \frac{1}{8\pi^2}\int_{S^4} X\,|F|^2\mathrm{vol}$, where $X \in \mathbb{R}^5$ is the position vector and $|F|^2\mathrm{vol} = -\operatorname{tr}(F \wedge \star F)$. It is a point of the open ball B^5. For the instanton $(0, \lambda)$, only the height $X_5 = (r^2-1)/(r^2+1)$ contributes. Show that $c_5(0) = -1$ (south pole, the image of $x = 0$), $c_5 = 0$ at $\lambda = 1$, and $c_5 \to 1$ as $\lambda \to \infty$. *Answer:* the radial integral gives $c_5(\lambda) = \big(\lambda^8 - 8\lambda^6 + 24\lambda^4 \ln\lambda + 8\lambda^2 - 1\big)/(1 - \lambda^2)^4$, which has the three stated values (at $\lambda = 1$ the numerator vanishes to fifth order and the denominator to fourth, so the limit is 0, as it must be by the symmetry $\lambda \mapsto 1/\lambda$, which

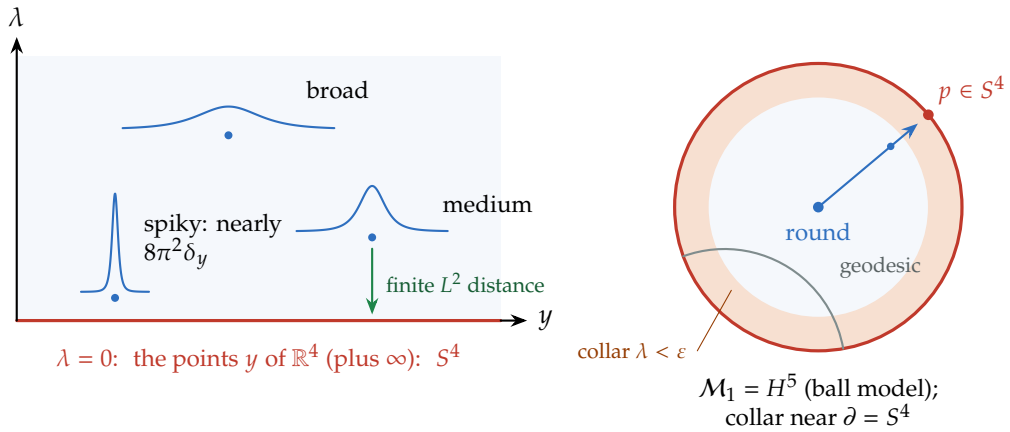


Figure 8.49: Two pictures of $\mathcal{M}_1$. *Left:* the upper half-space, one dimension of y shown. Each point (y, λ) is a BPST instanton, drawn as its action-density profile: broad high up, spiky near the boundary. As $\lambda \to 0$ the density tends to $8\pi^2\delta_y$, so the boundary is a copy of spacetime. *Right:* the ball model. The centre is the round instanton of Example 8.95; moving along a radius towards $p \in S^4$ concentrates the instanton at p. The shaded collar of small instantons is a copy of $S^4 \times (0, \varepsilon)$; it is the part of the picture that Donaldson generalises (Section 8.11).

maps $c_5 \mapsto -c_5$). The map $c : \mathcal{M}_1 \to B^5$ is a diffeomorphism: this is the ball model made concrete.

Watch out: which constant rotations are framings

In the regular gauge, conjugating A by a constant unit quaternion q gives the rotated instanton (Example 8.65), but a constant in the regular gauge is *not* constant near infinity: in the singular gauge it becomes $x\, q\, \bar{x}/r^2$, which depends on the direction of approach. The framing rotations are the gauge transformations that are constant *at infinity in the singular gauge*, like $\Lambda_C^{\text{sing}} \to C$ of Example 8.89. Both descriptions give the same conclusion, $\mathcal{M}_1' = \mathcal{M}_1 \times \mathrm{SO}(3)$, but only the second one is literally a statement about the frame at the base point.

8.10.5 The ADHM construction

We have assumed that every $k = 1$ instanton is BPST. More generally we would like to know *all* instantons. The 't Hooft ansatz of Section 8.5.5 gives $5k$ parameters, short of $8k - 3$ for $k \geq 3$. In 1978 Atiyah, Drinfeld, Hitchin and Manin found a construction that produces every instanton on S^4, for every k, from a few constant matrices. It is the **ADHM construction**. We give it for $\mathrm{SU}(2) = \mathrm{Sp}(1)$, using quaternions, and then say how it looks for $\mathrm{SU}(N)$.

Key idea: The ADHM pipeline

1. *Data.* Constant quaternionic matrices a, b of size $(k+1) \times k$.
2. *A linear family.* $\Delta(x) = a + b\,\bar{x}$, one matrix for each $x \in \mathbb{H} = \mathbb{R}^4$.
3. *ADHM condition.* $\Delta(x)^\dagger \Delta(x)$ is *real* (a real $k \times k$ matrix times the quaternion 1) and invertible, for every x. Write $f(x) = (\Delta^\dagger \Delta)^{-1}$.
4. *A unit vector.* $U(x) \in \mathbb{H}^{k+1}$ with $U^\dagger \Delta = 0$ and $U^\dagger U = 1$.
5. *The connection.* $A = U^\dagger \mathrm{d}U$. It is self-dual, with $F = U^\dagger b\, f\, (\mathrm{d}\bar{x} \wedge \mathrm{d}x)\, b^\dagger U$, and has charge k.

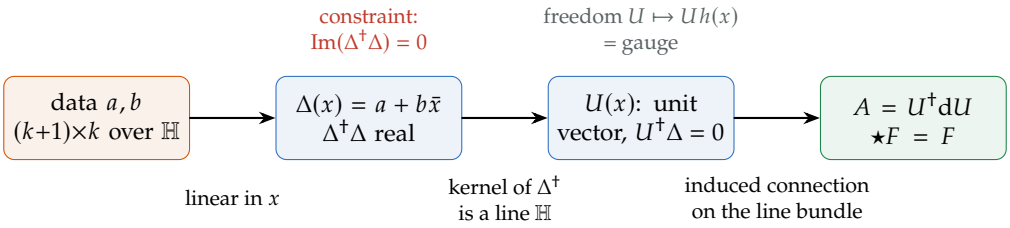


Figure 8.50: The ADHM construction as a pipeline. All the analysis is in the last box and is automatic; the only equation to solve is the algebraic constraint in the second box. The kernel of $\Delta(x)^\dagger$ is a quaternionic line in $\mathbb{H}^{k+1}$ that turns as x moves: this family of lines is the instanton bundle, and $A = U^\dagger \mathrm{d}U$ is the connection it inherits from the flat $\mathbb{H}^{k+1}$.

The idea behind step 5 is one we know from quantum mechanics: a **Berry connection**. If a unit vector $U(x)$ in a fixed vector space depends on parameters x, then $U^\dagger \mathrm{d}U$ is the connection that measures how U turns as x changes. Here the vector is quaternionic, so the connection takes values in $\mathrm{Im}\,\mathbb{H} = \mathfrak{su}(2)$. We now check each step and ask why it is legitimate.

Why U exists and is unique up to a phase. If $\Delta^\dagger\Delta$ is invertible, Δ has trivial kernel, so its image is a k-dimensional (quaternionic) subspace of $\mathbb{H}^{k+1}$. Its orthogonal complement, which is the kernel of $\Delta^\dagger$, is one-dimensional. A unit vector U in it is unique up to $U \mapsto Uh$ with $|h| = 1$, i.e. $h \in \mathrm{Sp}(1) = \mathrm{SU}(2)$.

Why A is $\mathfrak{su}(2)$-valued, and why h is a gauge transformation. Differentiating $U^\dagger U = 1$ gives $\mathrm{d}U^\dagger U + U^\dagger \mathrm{d}U = 0$, so $A = U^\dagger \mathrm{d}U$ satisfies $A^\dagger = -A$: an imaginary quaternion. If $U' = Uh(x)$ then

$$\begin{aligned} A' = U'^\dagger \mathrm{d}U' &\overset{(1)}{=} \bar{h}\, U^\dagger (\mathrm{d}U\, h + U\, \mathrm{d}h) \\ &\overset{(2)}{=} \bar{h} A h + \bar{h}\, \mathrm{d}h = h^{-1} A h + h^{-1} \mathrm{d}h, \end{aligned}$$

(1) Product rule; $(Uh)^\dagger = \bar{h} U^\dagger$.
(2) $U^\dagger U = 1$ and $\bar{h} = h^{-1}$.

which is (8.163).

The completeness relation. The two projectors $UU^\dagger$ (onto the kernel of $\Delta^\dagger$) and $\Delta f \Delta^\dagger$ (onto the image of Δ; check $(\Delta f \Delta^\dagger)^2 = \Delta f (\Delta^\dagger\Delta) f \Delta^\dagger = \Delta f \Delta^\dagger$) project onto complementary orthogonal subspaces. So

$$UU^\dagger + \Delta f \Delta^\dagger = \mathbb{1}_{k+1}. \tag{8.257}$$

Worked example 8.96: The ADHM curvature

$$\begin{aligned} F = \mathrm{d}A + A \wedge A &\overset{(1)}{=} \mathrm{d}U^\dagger \wedge \mathrm{d}U + U^\dagger \mathrm{d}U \wedge U^\dagger \mathrm{d}U \\ &\overset{(2)}{=} \mathrm{d}U^\dagger \wedge \mathrm{d}U - \mathrm{d}U^\dagger U \wedge U^\dagger \mathrm{d}U \\ &\overset{(3)}{=} \mathrm{d}U^\dagger (\mathbb{1} - UU^\dagger) \wedge \mathrm{d}U \\ &\overset{(4)}{=} \mathrm{d}U^\dagger \Delta f \Delta^\dagger \wedge \mathrm{d}U \\ &\overset{(5)}{=} U^\dagger\, \mathrm{d}\Delta\, f \wedge \mathrm{d}\Delta^\dagger\, U \\ &\overset{(6)}{=} U^\dagger b\, f\, (\mathrm{d}\bar{x} \wedge \mathrm{d}x)\, b^\dagger U. \end{aligned}$$

(1) $\mathrm{d}(U^\dagger \mathrm{d}U) = \mathrm{d}U^\dagger \wedge \mathrm{d}U$.
(2) $U^\dagger \mathrm{d}U = -\mathrm{d}U^\dagger U$ (from $U^\dagger U = 1$).

(3) Factor.
(4) The completeness relation (8.257).
(5) Differentiate $U^\dagger\Delta = 0$: $\mathrm{d}U^\dagger\Delta = -U^\dagger\mathrm{d}\Delta$, and similarly $\Delta^\dagger\mathrm{d}U = -\mathrm{d}\Delta^\dagger U$; the two minus signs cancel.
(6) $\mathrm{d}\Delta = b\,\mathrm{d}\bar{x}$, $\mathrm{d}\Delta^\dagger = \mathrm{d}x\,b^\dagger$; take b real (the canonical form below), so b and f commute with the quaternion $\mathrm{d}\bar{x}$, and f can be moved to the left of $\mathrm{d}\bar{x}$.

Now $\mathrm{d}\bar{x}\wedge\mathrm{d}x = -2e_a\omega_a^+$ is self-dual (8.195), and F is a sum of terms $p\,(\mathrm{d}\bar{x}\wedge\mathrm{d}x)\,q$ with quaternions p, q, each a combination of the ω_a^+. So $\star F = F$. The reality of f was essential in step 6: a non-real f would not commute with $\mathrm{d}\bar{x}$, and anti-self-dual pieces could appear.

The ADHM data in canonical form

The data has redundancies: $\Delta \mapsto Q\Delta R$ with $Q \in \mathrm{Sp}(k+1)$ and $R \in \mathrm{GL}(k,\mathbb{R})$ changes nothing essential (it rotates $\mathbb{H}^{k+1}$ and reparametrises the image). Using them one brings b to the form

$$b = \begin{pmatrix} 0 \\ \mathbb{1}_k \end{pmatrix}, \qquad a = \begin{pmatrix} L \\ M \end{pmatrix}, \qquad L \in \mathbb{H}^{1\times k}, \quad M \in \mathbb{H}^{k\times k}.$$

Then $\Delta(x) = \begin{pmatrix} L \\ M + \bar{x}\mathbb{1}_k \end{pmatrix}$.

Worked example 8.97: The ADHM constraint in canonical form

$$\begin{aligned}\Delta^\dagger\Delta &\overset{(1)}{=} L^\dagger L + (M^\dagger + x\mathbb{1})(M + \bar{x}\mathbb{1}) \\ &\overset{(2)}{=} L^\dagger L + M^\dagger M + xM + M^\dagger\bar{x} + |x|^2\mathbb{1}.\end{aligned}$$

(1) Block multiplication; $(\bar{x}\mathbb{1})^\dagger = x\mathbb{1}$.
(2) Expand; $x\bar{x} = |x|^2$.

For this to be real for *all* x, the part linear in x must be real for all x, and the constant part must be real. Entry (i,j) of the linear part is $xM_{ij} + \overline{M_{ji}}\,\bar{x}$. If M is symmetric, $M_{ji} = M_{ij}$, this is $xM_{ij} + \overline{xM_{ij}} = 2\,\mathrm{Re}(xM_{ij})$: real. (Conversely, reality for all x forces $M^T = M$.) So the conditions are

$$M^T = M, \qquad \mathrm{Im}\big(L^\dagger L + M^\dagger M\big) = 0, \tag{8.258}$$

plus non-degeneracy ($\Delta^\dagger\Delta$ invertible for every x). The remaining redundancy preserving the canonical form is $\Delta \mapsto \begin{pmatrix} p & 0 \\ 0 & R \end{pmatrix}\Delta R^{-1}$ with $p \in \mathrm{Sp}(1)$ and $R \in \mathrm{O}(k)$, i.e. $L \mapsto pLR^{-1}$, $M \mapsto RMR^{-1}$.

8.10.6 The case $k = 1$ in full

For $k = 1$, $L = \lambda q$ is a single quaternion and $M = -\bar{y}$ is a single quaternion (a 1×1 matrix is automatically symmetric). The constraint $\mathrm{Im}(|L|^2 + |M|^2) = 0$ is automatic. So the data is simply

$$\Delta(x) = \begin{pmatrix} \lambda q \\ \bar{x} - \bar{y} \end{pmatrix}, \qquad \lambda > 0,\ |q| = 1,\ y \in \mathbb{H},$$

and $\Delta^\dagger\Delta = \lambda^2 + |x-y|^2$ is real and never zero. No equations to solve at all.

Worked example 8.98: ADHM reproduces BPST

Take $q = 1$ and $y = 0$ first. We need $U = (u_1, u_2)^T$ with $U^\dagger \Delta = \bar{u}_1 \lambda + \bar{u}_2 \bar{x} = 0$. Try $u_2 = \lambda/\sqrt{\phi}$; then $\bar{u}_1 \lambda = -\lambda \bar{x}/\sqrt{\phi}$, so $u_1 = -x/\sqrt{\phi}$:

$$U = \frac{1}{\sqrt{\phi}} \begin{pmatrix} -x \\ \lambda \end{pmatrix}, \qquad U^\dagger U = \frac{|x|^2 + \lambda^2}{\phi} = 1.$$

Then

$$\begin{aligned} A = U^\dagger \mathrm{d}U &\overset{(1)}{=} \frac{1}{\sqrt{\phi}} (-\bar{x},\ \lambda) \Big[\frac{1}{\sqrt{\phi}} \begin{pmatrix} -\mathrm{d}x \\ 0 \end{pmatrix} + \mathrm{d}\big(\phi^{-1/2}\big) \begin{pmatrix} -x \\ \lambda \end{pmatrix} \Big] \\ &\overset{(2)}{=} \frac{\bar{x}\, \mathrm{d}x}{\phi} + \sqrt{\phi}\, \mathrm{d}\big(\phi^{-1/2}\big) \\ &\overset{(3)}{=} \frac{\bar{x}\, \mathrm{d}x - \frac{1}{2}\mathrm{d}\phi}{\phi} \\ &\overset{(4)}{=} \frac{\mathrm{Im}(\bar{x}\, \mathrm{d}x)}{|x|^2 + \lambda^2}. \end{aligned}$$

(1) Product rule on $U = \phi^{-1/2}(-x, \lambda)^T$; λ is constant.
(2) First term: $(-\bar{x})(-\mathrm{d}x)/\phi$. Second term: $\phi^{-1/2}(\bar{x}x + \lambda^2)\mathrm{d}(\phi^{-1/2}) = \phi^{1/2}\mathrm{d}(\phi^{-1/2})$.
(3) $\phi^{1/2}\mathrm{d}(\phi^{-1/2}) = -\frac{1}{2}\phi^{-1}\mathrm{d}\phi$.
(4) (8.201): $\bar{x}\, \mathrm{d}x - \frac{1}{2}\mathrm{d}\phi = \mathrm{Im}(\bar{x}\, \mathrm{d}x)$.

This is exactly the BPST instanton (8.200) in the regular gauge. The curvature formula of Example 8.96 gives it in one line: $f = 1/\phi$ and $b^\dagger U = u_2 = \lambda/\sqrt{\phi}$, so $F = \frac{\lambda}{\sqrt{\phi}} \cdot \frac{1}{\phi}\, \mathrm{d}\bar{x} \wedge \mathrm{d}x \cdot \frac{\lambda}{\sqrt{\phi}} = \lambda^2\, \mathrm{d}\bar{x} \wedge \mathrm{d}x/\phi^2$, which is (8.202). With $y \neq 0$, replace x by $x - y$ throughout: the centre is y.

Worked example 8.99: The quaternion λq is size and framing

Now restore q. The data $(\lambda q, \bar{x})$ is obtained from $(\lambda, \bar{x})$ by the canonical-form symmetry with $p = q$, $R = 1$: $\Delta_q = \mathrm{diag}(q, 1)\Delta_1$. Then $U_q = \mathrm{diag}(q, 1)U_1$ solves $U_q^\dagger \Delta_q = U_1^\dagger \Delta_1 = 0$, and

$$A_q = U_q^\dagger \mathrm{d}U_q = U_1^\dagger\, \mathrm{diag}(\bar{q}, 1)\, \mathrm{diag}(q, 1)\mathrm{d}U_1 = A_1.$$

The connection is literally the same. What changes is the identification of the fibre at infinity with $\mathbb{H}$: as $|x| \to \infty$, $U_q \to (-q\, x/|x|, 0)^T$, so the top component, which is how the fibre at infinity is matched to the fixed $\mathbb{H}$ in the first row, is rotated by q. So q is the framing at infinity. The remaining symmetry $R \in \mathrm{O}(1) = \{\pm 1\}$ sends $(\lambda q, \bar{x} - \bar{y}) \mapsto (-\lambda q, \bar{x} - \bar{y})$. Therefore

$$\mathcal{M}_1' = \big\{(y, \lambda q)\big\}/\pm = \mathbb{R}^4 \times \big(\mathbb{H} \setminus \{0\}\big)/\mathbb{Z}_2, \qquad \mathcal{M}_1 = \mathcal{M}_1'/\mathrm{Sp}(1) = \mathbb{R}^4 \times \mathbb{R}^+,$$

in agreement with Example 8.66 and with the flat metric (8.256), in which λq is a Cartesian coordinate.

Since the ADHM construction gives *every* instanton (a theorem we discuss below), and at $k = 1$ it gives only BPST instantons, we have finally proved:

Key idea: Every charge-one $\mathrm{SU}(2)$ instanton on S^4 is BPST

$\mathcal{M}_1 = \{(y,\lambda)\} = \mathbb{R}^4 \times \mathbb{R}^+ \cong H^5$, and $\mathcal{M}_1' = \mathbb{R}^4 \times (\mathbb{R}^4 \setminus 0)/\mathbb{Z}_2$.

8.10.7 General k: counting, 't Hooft, and $\mathrm{SU}(N)$

Worked example 8.100: The ADHM count gives $8k$ and $8k-3$

Count real parameters in the canonical form:

$$\begin{aligned}\dim \mathcal{M}_k' &\overset{(1)}{=} \underbrace{4k}_{L} + \underbrace{4\cdot\tfrac{k(k+1)}{2}}_{M=M^T} - \underbrace{3\cdot\tfrac{k(k-1)}{2}}_{\mathrm{Im}(L^\dagger L + M^\dagger M)=0} - \underbrace{\tfrac{k(k-1)}{2}}_{\mathrm{O}(k)} \\ &\overset{(2)}{=} 4k + 2k^2 + 2k - 2k^2 + 2k \\ &\overset{(3)}{=} 8k.\end{aligned}$$

(1) L has k quaternion entries; a symmetric $k\times k$ quaternionic matrix has $k(k+1)/2$ independent entries. The matrix $X = L^\dagger L + M^\dagger M$ is quaternionic Hermitian, $X_{ji} = \overline{X_{ij}}$, so its imaginary part is antisymmetric: $k(k-1)/2$ independent entries, each with three imaginary components. Finally divide by $\mathrm{O}(k)$, of dimension $k(k-1)/2$ (it acts freely on non-degenerate data).
(2) $3k(k-1)/2 + k(k-1)/2 = 2k(k-1) = 2k^2 - 2k$.
(3) Collect.

Quotienting by the framing $\mathrm{Sp}(1)$ removes 3 more: $\dim \mathcal{M}_k = 8k-3$. The purely algebraic count agrees with the index theorem of Section 8.9. For $k=2$: L gives 8, M (three quaternions) gives 12, the constraint is one imaginary quaternion (3), and $\mathrm{O}(2)$ has dimension 1: $8+12-3-1 = 16 = 8\cdot 2$.

Worked example 8.101: The 't Hooft solutions are ADHM with diagonal M

Take $L = (\lambda_1,\dots,\lambda_k)$ *real* and $M = -\operatorname{diag}(\bar y_1,\dots,\bar y_k)$. Then $M^T = M$, and $L^\dagger L$ (entries $\lambda_i\lambda_j$) and $M^\dagger M$ (entries $|y_i|^2\delta_{ij}$) are real: the constraint holds. Now $\Delta^\dagger\Delta = \operatorname{diag}(r_1^2,\dots,r_k^2) + \lambda\lambda^T$ with $r_i = |x-y_i|$, and the matrix determinant lemma $\det(D+vv^T) = \det D\,(1+v^TD^{-1}v)$ gives

$$\begin{aligned}\det(\Delta^\dagger\Delta) &\overset{(1)}{=} \Big(\prod_i r_i^2\Big)\Big(1+\sum_i \frac{\lambda_i^2}{r_i^2}\Big) \\ &\overset{(2)}{=} \Big(\prod_i r_i^2\Big)\,\rho(x),\end{aligned}$$

(1) Determinant lemma with $D = \operatorname{diag}(r_i^2)$ and $v = \lambda$.
(2) $\rho = 1 + \sum \lambda_i^2/|x-y_i|^2$ is the 't Hooft potential (8.207).

For every ADHM instanton there is a formula of Osborn for the action density,

$$-\tfrac{1}{2}\operatorname{tr}(F_{\mu\nu}F_{\mu\nu}) = -\tfrac{1}{2}\,\Box\Box \ln\det\big(\Delta^\dagger\Delta\big). \tag{8.259}$$

Check it at $k=1$: $\det\Delta^\dagger\Delta = \phi$, and a short radial computation (as in Example 8.57) gives $\Box\Box\ln\phi = -96\lambda^4/\phi^4$, so the right side is $48\lambda^4/\phi^4$, the BPST density. For the diagonal data, $\ln\det = \sum_i \ln r_i^2 + \ln\rho$, and $\Box\Box\ln r_i^2 = \Box(4/r_i^2) = 0$ away from y_i. So the density is

$-\frac{1}{2}\Box\Box \ln \rho$, which is exactly the 't Hooft density (8.208). (A numerical evaluation of both sides at random points for $k = 2$ confirms the agreement.) So the 't Hooft solutions are the ADHM instantons with real L and diagonal M.

The 't Hooft family has real λ_i: all lumps have the same colour orientation. Making the λ_i genuine quaternions $\lambda_i q_i$ breaks the constraint: the off-diagonal entry of $\mathrm{Im}(L^\dagger L)$ is $\mathrm{Im}(\lambda_i \lambda_j \bar{q}_i q_j)$, non-zero when $q_i \neq \pm q_j$. To restore it the off-diagonal entries of M must be switched on. So the relative colour orientations of the lumps, the $3(k-1)$ parameters missing from Figure 8.37, are tied to off-diagonal "positions". For well-separated lumps the off-diagonal entries are small and the y_i are the centres; when the lumps overlap, the notion of individual positions dissolves.

Try it: The complex ADHM data for $\mathrm{SU}(N)$

For $\mathrm{SU}(N)$ the ADHM data are complex matrices B_1, B_2 $(k \times k)$, I $(k \times N)$, J $(N \times k)$, subject to

$$[B_1, B_2] + IJ = 0 \quad \text{(complex)}, \qquad [B_1, B_1^\dagger] + [B_2, B_2^\dagger] + II^\dagger - J^\dagger J = 0 \quad \text{(real)},$$

modulo $\mathrm{U}(k)$. Count the framed dimension, and deduce the first row of Nash's table (8.137). *Answer:* $2(2k^2 + 2kN)$ real parameters, minus $2k^2$ (complex equation) minus k^2 (real equation) minus k^2 ($\mathrm{U}(k)$) gives $\dim \mathcal{M}'_k = 4kN$. Removing the framing $\mathrm{SU}(N)$ gives $4kN - (N^2 - 1) = 4Nk - N^2 + 1$, Nash's entry for $\mathrm{SU}(N)$. At $k = 1$, $N = 2$ it is 5. Moreover the gauge field lives in $\mathbb{C}^N$, but the data only "see" the span of the columns of $I^\dagger$ and J, a subspace of dimension at most $2k$. If $N > 2k$, its orthogonal complement is a trivial flat summand of the bundle, and the instanton is reducible: this is Nash's condition $k \geq N/2$.

Why every instanton arises

The completeness of the construction is a theorem (Atiyah, Drinfeld, Hitchin and Manin, 1978). The route is through **twistor theory**, and we only outline it. There is a fibration $\mathbb{CP}^3 \to S^4$ whose fibres are $\mathbb{CP}^1$'s. The Penrose–Ward correspondence turns a self-dual connection on S^4 into a *holomorphic* vector bundle on $\mathbb{CP}^3$: pull the bundle and its connection back to $\mathbb{CP}^3$ and use the $(0,1)$ part of the pulled-back connection as a $\bar{\partial}$-operator. Self-duality on S^4 is exactly the integrability condition $\bar{\partial}^2 = 0$ that makes this a holomorphic structure; the bundle is then holomorphically trivial on every fibre. Holomorphic bundles on $\mathbb{CP}^3$ of this type can be described by linear algebra (Horrocks' *monads*, which Barth applied to this problem), and unwinding the monad gives exactly the matrix $\Delta(x)$. This is the algebraic geometry of Chapter 5 solving a problem of non-linear partial differential equations: the self-duality equation, which looks hard, becomes a statement about holomorphic objects, which are rigid.

Physical picture: why linear algebra can solve a non-linear PDE

The ADHM ansatz $A = U^\dagger \mathrm{d}U$ builds the instanton as a Berry connection of a moving line inside a fixed flat space $\mathbb{H}^{k+1}$. Every connection can be obtained this way for *some* embedding (this is a theorem of Narasimhan and Ramanan). What ADHM finds is that for self-dual connections on S^4 the embedding can always be chosen *linear in* x, as the kernel of $\Delta(x)^\dagger$. The self-duality equation becomes the algebraic condition that $\Delta^\dagger \Delta$ is real.

Reading Nash (p. 241)

Nash states that for $M = S^4$ the SU(2) moduli space $\mathcal{M}_1$ is five-dimensional, non-compact, with boundary S^4, and "in fact $\mathcal{M}_1$ is the five dimensional hyperbolic space H^5". The dimension is his (8.132) at $k = 1$. The rest is Section 8.10.4: the parameters are (y, λ), the conformal group SO(5, 1) of S^4 acts on them as the isometries of H^5 (Example 8.94), and the boundary $\lambda = 0$ consists of point-like instantons, one for each point of S^4. That there are no other $k = 1$ instantons is the ADHM theorem (Example 8.98). Nash does not mention zero modes or ADHM; they are the standard physicist's and mathematician's ways of making his statement concrete.

Recap

- A zero mode solves $\mathrm{d}_A^- a = 0$ and $\mathrm{d}_A^* a = 0$; it is a tangent vector to the moduli space. For every conformal Killing field V, $\iota_V F = \mathcal{L}_V A - \mathrm{d}_A(\iota_V A)$ solves the first equation, and the second when the rotation part of V has no self-dual component.
- For BPST: 4 translations and 1 dilation give $\mathcal{M}_1$; 3 left rotations rotate the frame at infinity, giving $\mathcal{M}_1'$; right rotations and special conformal maps add nothing.
- Norms: $4\pi^2$, $8\pi^2$, $8\pi^2\lambda^2$. The L^2 metric on $\mathcal{M}_1'$ is flat, $4\pi^2|\mathrm{d}y|^2 + 8\pi^2|\mathrm{d}(\lambda q)|^2$; the boundary $\lambda = 0$ is at finite distance.
- The conformal group acts on (y, λ) as the isometries of the upper half-space: $\mathcal{M}_1 = \mathrm{SO}(5,1)/\mathrm{SO}(5) = H^5$, with boundary S^4 made of point-like instantons.
- ADHM: $\Delta(x) = a + b\bar{x}$ with $\Delta^\dagger\Delta$ real, $U \in \operatorname{Ker}\Delta^\dagger$, $A = U^\dagger \mathrm{d}U$, $F = U^\dagger b f\, \mathrm{d}\bar{x} \wedge \mathrm{d}x\, b^\dagger U$ self-dual. At $k = 1$ it gives BPST and nothing else; in general it gives $8k$ framed parameters, the 't Hooft solutions as diagonal data, and for SU(N) the dimension $4Nk - N^2 + 1$.

8.11 The moduli space as a global object

Nash ends the section with a warning and two outlooks (pp. 240–242). The warning: "the Atiyah–Singer index calculation is an infinitesimal one." It counts tangent vectors at a point $[A]$. It does not tell us that there is a manifold for these vectors to be tangent to. The outlooks: the global moduli space is a powerful tool of *topology* (Donaldson's theorem), and it approximates the *configuration space* of gauge theory (Atiyah–Jones). This section takes them in turn:

1. the three conditions that turn a dimension count into a manifold: existence, integrability, and the Hausdorff property;
2. Donaldson's theorem, previewed through the moduli space $\mathcal{M}_1$ we now know well;
3. Nash's analogy with the moduli of Riemann surfaces;
4. the inclusion $\mathcal{M}_k' \hookrightarrow \mathcal{A}/\mathcal{G}$ and what it says about the topology of the space of all gauge fields.

8.11.1 From a dimension count to a manifold

A finite-dimensional analogy shows what can go wrong. Take a smooth map $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ and its zero set $Z = \Phi^{-1}(0)$. At a point $p \in Z$ the linearisation $\mathrm{d}\Phi_p$ has kernel $\operatorname{Ker}\mathrm{d}\Phi_p$ (the *infinitesimal* solutions) and cokernel $\operatorname{Coker}\mathrm{d}\Phi_p$. The index $\dim\operatorname{Ker} - \dim\operatorname{Coker} = n - m$ is the same at every point. We call it the **expected dimension**. The implicit function theorem says: if $\mathrm{d}\Phi_p$ is *onto* (cokernel zero), then near p the set Z is a manifold of dimension $\dim\operatorname{Ker}\mathrm{d}\Phi_p = n - m$. If the cokernel is not zero, anything can happen.

Worked example 8.102: Three zero sets in the plane

Take $n = 2$, $m = 1$, so the expected dimension is 1, and $p = 0$.

- $\Phi = y - x^2$. $\mathrm{d}\Phi_0 = (0, 1)$ is onto. Z is the parabola, a smooth curve with tangent line $\operatorname{Ker} \mathrm{d}\Phi_0 = \{y = 0\}$.
- $\Phi = xy$. $\mathrm{d}\Phi_0 = (0, 0)$: every vector is an infinitesimal solution ($\operatorname{Ker} = \mathbb{R}^2$) and the cokernel is $\mathbb{R}$. Z is the union of the two axes: not a manifold at 0. The quadratic term xy *obstructs* most infinitesimal solutions: only those along the axes extend to actual curves in Z.
- $\Phi = x^2 + y^2$. Again $\operatorname{Ker} = \mathbb{R}^2$, $\operatorname{Coker} = \mathbb{R}$. Now $Z = \{0\}$: *no* infinitesimal solution extends, and Z has dimension 0, less than expected.

In each case $\dim \operatorname{Ker} - \dim \operatorname{Coker} = 1$, as the index says. What differs is how the cokernel (the **obstruction space**) interacts with the higher-order terms.

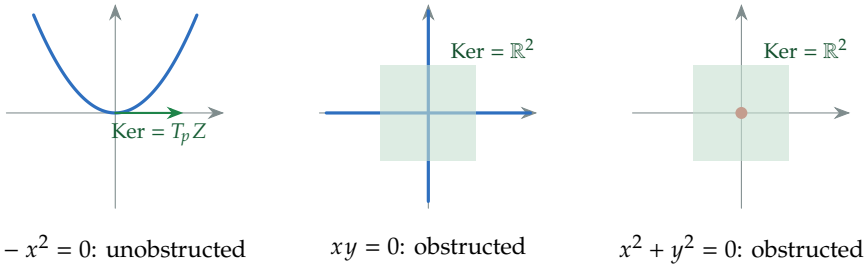


Figure 8.51: Three maps $\mathbb{R}^2 \to \mathbb{R}$ with the same index 1. When the derivative is onto (left) the zero set is a manifold whose tangent space is the kernel. When it is not (middle, right), the kernel (shaded) is larger than the zero set's true tangent directions, and the zero set may be singular or too small. For instantons, the cokernel is H^2 of the deformation complex.

Condition 1: existence

We must first have at least one instanton. On S^4 this is the BPST instanton and its relatives (Sections 8.5 and 8.10); as Nash says, "the first condition is easily met as many instantons are known". On a general M it is a theorem of Taubes (1982): if the intersection form of M has no negative part ($b_2^- = 0$ in our orientation, where instantons are self-dual), then for every $k \geq 1$ there are self-dual SU(2) connections of charge k. They are constructed by *grafting*: take k very small BPST instantons, cut them off far from their centres, glue them into the trivial connection on M at k points, and correct the result to an exact solution. The correction exists because the obstruction to it lies in a space built from self-dual-type harmonic forms on M, which vanishes when $b_2^- = 0$.

Condition 2: integrability, and the Kuranishi picture

Near an instanton A, write a nearby connection in background gauge, $A + a$ with $\mathrm{d}_A^* a = 0$. It is self-dual when

$$\Phi(a) := \pi_- F(A + a) = \mathrm{d}_A^- a + \pi_-(a \wedge a) = 0, \tag{8.260}$$

by the expansion (8.224). This is the infinite-dimensional version of the example above: $\Phi : \{a : \mathrm{d}_A^* a = 0\} \to \Omega^2_-(\mathrm{ad}\, P)$, with linearisation d_A^-, kernel H^1 and cokernel H^2.

Worked example 8.103: Extending a zero mode to second order

Given a zero mode $a_1 \in H^1$, look for a curve of instantons $A(t) = A + ta_1 + t^2 a_2 + O(t^3)$. Substituting into (8.260),

$$\begin{aligned}\pi_- F(A(t)) &\overset{(1)}{=} \pi_-\big(F + t\,\mathrm{d}_A a_1 + t^2\,\mathrm{d}_A a_2 + t^2\, a_1 \wedge a_1\big) + O(t^3)\\ &\overset{(2)}{=} t\,\mathrm{d}_A^- a_1 + t^2\big(\mathrm{d}_A^- a_2 + \pi_-(a_1 \wedge a_1)\big) + O(t^3)\\ &\overset{(3)}{=} t^2\big(\mathrm{d}_A^- a_2 + \pi_-(a_1 \wedge a_1)\big) + O(t^3).\end{aligned}$$

(1) $F(A+b) = F + \mathrm{d}_A b + b \wedge b$ with $b = ta_1 + t^2 a_2$; keep terms up to t^2.
(2) $\pi_- F = 0$ since A is self-dual; $\mathrm{d}_A^- = \pi_- \circ \mathrm{d}_A$.
(3) a_1 is a zero mode.

So we must solve $\mathrm{d}_A^- a_2 = -\pi_-(a_1 \wedge a_1)$. A linear equation $Lx = y$ with L Fredholm is solvable exactly when y is orthogonal to $\operatorname{Coker} L$. Here $\operatorname{Coker} \mathrm{d}_A^- = H^2$. So the *obstruction* to extending a_1 to second order is the H^2-component of $\pi_-(a_1 \wedge a_1)$. If $H^2 = 0$, there is never an obstruction, at this order or any other.

Key idea: The Kuranishi model

Near $[A]$, the moduli space looks like

$$\mathcal{M}_k \;\approx\; \kappa^{-1}(0)\,/\,\Gamma_A, \qquad \kappa : H^1 \to H^2,$$

where κ (the **Kuranishi map**) is a smooth map with $\kappa(0) = 0$ and $\mathrm{d}\kappa_0 = 0$, whose quadratic part is $a \mapsto$ (H^2-component of $\pi_-(a \wedge a)$), and Γ_A is the stabiliser of A (Section 8.7.3). Consequences:

- if $H^2 = 0$ and Γ_A is the centre, $\mathcal{M}_k$ is a manifold near $[A]$ with tangent space H^1: every infinitesimal deformation is tangent to a genuine family (*integrability*);
- if $H^2 = 0$ but Γ_A is bigger, $\mathcal{M}_k$ is locally H^1/Γ_A, e.g. the cone on $\mathbb{CP}^2$ at a reducible point (Section 8.7.4);
- if $H^2 \neq 0$, $\mathcal{M}_k$ can be singular or have the wrong dimension.

This is what Nash calls "an integrability question similar to those that arise in complex analysis when passing from an almost complex structure to a full complex structure": there too an infinitesimal (linear) condition must be promoted to a local (non-linear) one, and the obstruction is a curvature-like quadratic term (the Nijenhuis tensor). On S^4, $H^2 = 0$ at every instanton by the Weitzenböck argument of Section 8.8.7; this is how Atiyah, Hitchin and Singer show that every element of H^1 comes from a one-parameter family $A + ta + O(t^2)$. On a general M one uses a *generic* metric: Freed and Uhlenbeck proved that for almost every Riemannian metric, $H^2 = 0$ at every irreducible instanton, so $\mathcal{M}_k^{\mathrm{irr}}$ is a smooth manifold of the expected dimension.

Condition 3: the Hausdorff property

Local charts must fit into a global space that is *Hausdorff*: two distinct points have disjoint neighbourhoods. A quotient by a group action can fail this badly.

Worked example 8.104: A non-Hausdorff quotient

Let $\mathbb{R}$ act on $\mathbb{R}^2 \setminus \{0\}$ by $t \cdot (x, y) = (e^t x, e^{-t} y)$. The orbits are the branches of the hyperbolas $xy = c$ ($c \neq 0$) and the four open half-axes. Consider the orbits $O_+ = \{x > 0, y = 0\}$ and $O^+ = \{x = 0, y > 0\}$. The points $(1, \varepsilon) \in \{xy = \varepsilon\}$ converge to $(1, 0) \in O_+$, and the points $(\varepsilon, 1)$, which lie on the *same* hyperbola branch, converge to $(0, 1) \in O^+$. So the single sequence of orbits $\{xy = \varepsilon\}$, $\varepsilon \to 0$, converges to both O_+ and O^+ in the quotient. Any neighbourhoods of these two points of the quotient contain the orbit $\{xy = \varepsilon\}$ for small ε, so they intersect: the quotient is not Hausdorff.

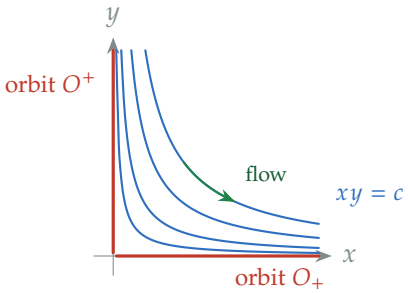


Figure 8.52: The orbits of $(x, y) \mapsto (e^t x, e^{-t} y)$ in the first quadrant. As $c \to 0$ the hyperbola $xy = c$ hugs *both* half-axes, so in the orbit space it converges to two different points at once. The gauge group acting on $\mathcal{A}$ does not behave like this: the estimate $|\mathrm{d}g| \leq |A| + |A_g|$ keeps gauge transformations from escaping to infinity.

What saves the gauge-theory quotient is that the gauge group acts *properly*: gauge transformations relating two nearby connections cannot run off to infinity. The mechanism is a one-line estimate.

Worked example 8.105: Why $\mathcal{A}/\mathcal{G}$ is Hausdorff

Suppose $A_i \to A$ and $g_i \cdot A_i \to B$ (uniformly, say). We want to show $B = g \cdot A$ for some gauge transformation g, so that $[A]$ and $[B]$ are the same point and no two distinct points are "glued" as in Example 8.104. From $A_g = g^{-1} A g + g^{-1} \mathrm{d}g$,

$$\mathrm{d}g_i \overset{(1)}{=} g_i \, (g_i \cdot A_i) - A_i \, g_i$$
$$\overset{(2)}{\Longrightarrow} \quad |\mathrm{d}g_i| \leq |g_i \cdot A_i| + |A_i|,$$

(1) Multiply $g^{-1}\mathrm{d}g = A_g - g^{-1} A g$ on the left by g.

(2) g_i is unitary at each point, so multiplying by it does not change norms; then the triangle inequality.

The right side is bounded uniformly in i, because both sequences converge. So the g_i are uniformly bounded (they are unitary) with uniformly bounded derivatives. By the Arzelà–Ascoli theorem a subsequence converges uniformly to some g, and passing to the limit in step (1) gives $\mathrm{d}g = gB - Ag$, i.e. $B = g \cdot A$. (The honest proof works in Sobolev spaces and bootstraps the regularity of g, but the mechanism is this estimate.)

Key idea: The global moduli space (Nash p. 241)

Existence, integrability and the Hausdorff property all hold for SU(2) instantons on S^4 (Atiyah–Hitchin–Singer): $\mathcal{M}_k$ is a smooth, Hausdorff manifold of dimension $8k - 3$, non-compact, with no singular points. On a simply connected M with positive definite

intersection form and a generic metric, $\mathcal{M}_k$ is a smooth manifold of dimension $8k-3$ away from finitely many cone points, one for each reducible instanton.

8.11.2 Donaldson's theorem: a preview

Nash's remark after (8.134) is the reason the whole section matters for topology: the manifolds for which $\dim \mathcal{M}_k = 8k-3$ by (8.134), simply connected with positive definite intersection form, are "precisely the class of M to which Donaldson's theorem applies". The theorem itself (Chapter 1) is:

Key idea: Donaldson's theorem (1983)

Let M be a smooth, closed, oriented, simply connected 4-manifold whose intersection form q on $H^2(M;\mathbb{Z})$ is positive definite. Then q is *diagonalisable over* $\mathbb{Z}$: there is a basis $e_1,\dots,e_r$ of $H^2(M;\mathbb{Z})$ with $q(e_i,e_j)=\delta_{ij}$.

By Freedman's work (Chapter 1), every unimodular form is the intersection form of some *topological* simply connected 4-manifold. So Donaldson's theorem says that most positive definite forms, for instance the form E_8 of rank 8, belong to topological manifolds that admit *no smooth structure*. A purely topological statement is proved with a moduli space of solutions of a physical field equation. We now describe the proof, using everything we have built. Nash notes (p. 241) that the S^4 space $\mathcal{M}_1$ "is the space used in the proof", and that "the boundary of $\mathcal{M}_1$ is M" for the manifolds of the theorem.

Step 1: the moduli space and its dimension

Take $G=\mathrm{SU}(2)$ and $k=1$. By (8.134), $\dim \mathcal{M}_1 = 8\cdot 1-3=5$. Choose a generic metric on M, so that (Freed–Uhlenbeck) $\mathcal{M}_1$ is a smooth 5-manifold away from the reducible points. Donaldson also proved that it is *orientable*; we need this for Step 5.

Step 2: the reducible points

From Section 8.7.5 (Nash (8.135)–(8.136)), a reducible SU(2) instanton has $E = L\oplus L^{-1}$ and $k=-c_2(E)=q(c_1(L),c_1(L))$. For $k=1$ we need a class $e=c_1(L)$ with $q(e,e)=1$. Conversely, each such e gives exactly one reducible instanton:

- Represent e by its harmonic 2-form ω. Since $b_2^- = 0$, every harmonic 2-form is self-dual, so the U(1) connection on L with curvature $-2\pi i\omega$ is self-dual, and so is the diagonal SU(2) connection on $L\oplus L^{-1}$.
- The U(1) connection with given curvature is unique up to gauge, because the connections differ by closed 1-forms and $H^1(M)=0$ (M is simply connected).
- e and $-e$ give the same reducible: $L\oplus L^{-1}\cong L^{-1}\oplus L$.

So the reducible points of $\mathcal{M}_1$ are in one-to-one correspondence with the pairs $\pm e$ with $q(e,e)=1$. Let m be the number of such pairs.

Step 3: the neighbourhood of a reducible point

At a reducible A the stabiliser is U(1) (Section 8.7.3), so $h_0=1$. With $h_2=0$ (generic metric), the index gives

$$h_0-h_1+h_2=-5 \quad\Longrightarrow\quad h_1=6.$$

The splitting $\operatorname{ad} P = \mathbb{R} \oplus L^2$ (the diagonal and off-diagonal parts of a traceless 2×2 matrix) makes H^1 a complex vector space: the real part contributes $H^1(M;\mathbb{R}) = 0$, and the L^2 part gives $H^1 \cong \mathbb{C}^3$, on which $u \in \mathrm{U}(1)$ acts by u^2. By the Kuranishi model, near $[A]$

$$\mathcal{M}_1 \approx \mathbb{C}^3/\mathrm{U}(1) = \text{cone on } \left(S^5/\mathrm{U}(1)\right) = \text{cone on } \mathbb{CP}^2,$$

exactly as in Section 8.7.4: the unit sphere $S^5 \subset \mathbb{C}^3$ modulo the circle is $\mathbb{CP}^2$ (the Hopf fibration $S^5 \to \mathbb{CP}^2$), and the radius is the cone coordinate.

Step 4: the end of the moduli space, and compactness

For $M = S^4$ we saw that $\mathcal{M}_1 = H^5$ has one end, the boundary $\lambda \to 0$ of point-like instantons, and that this end is a copy of $S^4 \times (0, \varepsilon)$ (Figure 8.49). Two theorems show that the same is true for any M of the theorem.

- *Taubes' gluing.* For each point $x \in M$ and each small λ, graft a BPST instanton of size λ at x (Condition 1). The gluing also involves an identification of the colour frames, an element of SO(3), but it is absorbed by the stabiliser SU(2) of the trivial connection we glue into. So the concentrated instantons form a collar $M \times (0, \varepsilon)$ inside $\mathcal{M}_1$.
- *Uhlenbeck's compactness.* Any sequence of charge-one instantons either has a subsequence converging (modulo gauge) to an instanton, or concentrates: its action density tends to $8\pi^2\delta_x$ at some point x, while away from x it converges to a charge-zero instanton, i.e. a flat connection, which is trivial since $\pi_1(M) = 0$. The concentrating sequences eventually enter the collar.

So, apart from the collar, $\mathcal{M}_1$ is compact. Adding the points $x \in M$ at $\lambda = 0$ closes up the collar with boundary M: this is Nash's "the boundary of $\mathcal{M}_1$ is M".

Step 5: the cobordism

Remove a small open cone neighbourhood of each reducible point, and add the boundary M at the end of the collar. The result is a compact, oriented, smooth 5-manifold W with

$$\partial W = M \sqcup \left(\pm\mathbb{CP}^2\right) \sqcup \cdots \sqcup \left(\pm\mathbb{CP}^2\right) \qquad (m \text{ copies}), \tag{8.261}$$

where the signs are the orientations induced from W. So M is **cobordant** to a disjoint union of m copies of $\pm\mathbb{CP}^2$.

Step 6: the signature

The signature of a boundary vanishes. Here is the Stokes-theorem argument. By the Hirzebruch signature theorem (Chapter 4), $\tau(X) = \frac{1}{3}\int_X p_1(TX)$ for a closed oriented 4-manifold, with p_1 in the standard normalisation (not Nash's, Section 8.9.3).

Worked example 8.106: $\tau(\partial W) = 0$ by Stokes

Let W be a compact oriented 5-manifold with a Riemannian metric that is a product near ∂W, and let P be the Pontryagin 4-form of TW built from its curvature. Then

$$3\,\tau(\partial W) \overset{(1)}{=} \int_{\partial W} p_1(T\partial W) \overset{(2)}{=} \int_{\partial W} P\big|_{\partial W}$$

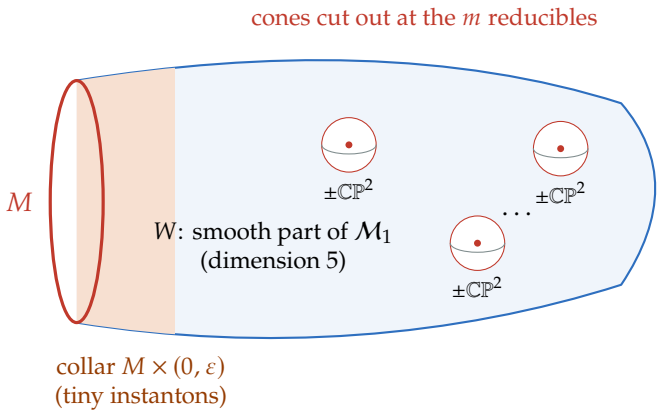


Figure 8.53: The Donaldson cobordism. The 5-dimensional moduli space $\mathcal{M}_1$ has one end, a collar of concentrated instantons whose boundary is a copy of M (left), and m cone points, one per pair $\pm e$ with $q(e,e) = 1$. Cutting out small cones leaves a compact W whose boundary is M together with m copies of $\pm\mathbb{CP}^2$ (the links of the cones). For $M = S^4$ there are no cones and W is the closed 5-ball, the compactified H^5.

$$\overset{(3)}{=} \int_W \mathrm{d}P$$
$$\overset{(4)}{=} 0.$$

(1) The signature theorem on the closed 4-manifold ∂W (its components added with their induced orientations).
(2) Near the boundary $TW = T\partial W \oplus \mathbb{R}$ with a product metric, so the curvature of TW restricted to ∂W is that of $T\partial W$, and the extra trivial line contributes nothing.
(3) Stokes' theorem.
(4) Characteristic forms are closed (Chern–Weil, Chapter T2).

Apply this to (8.261). The signature is additive over disjoint unions and changes sign with orientation, and $\tau(\mathbb{CP}^2) = 1$ (its form is (1)). So

$$0 = \tau(M) + \sum_{i=1}^{m} \varepsilon_i, \qquad \varepsilon_i = \pm 1 \quad \Longrightarrow \quad r = \tau(M) \le m,$$

where $r = \operatorname{rank} H^2(M)$ and $\tau(M) = r$ because q is positive definite.

Step 7: the algebra

Worked example 8.107: A positive definite form has at most r pairs of unit vectors, with equality only if diagonal

Let $e \neq \pm f$ with $q(e,e) = q(f,f) = 1$. Then

$$|q(e,f)| \overset{(1)}{<} \sqrt{q(e,e)\,q(f,f)}$$
$$\overset{(2)}{=} 1,$$

(1) Cauchy–Schwarz for the positive definite q, with strict inequality because e, f are not parallel (parallel integer vectors of equal length are equal up to sign).

(2) Both have length 1.

and $q(e,f)$ is an integer, so $q(e,f)=0$. Choosing one vector from each of the m pairs gives an *orthonormal* set, hence linearly independent: $m \le r$. If $m = r$, the orthonormal vectors span a sublattice Λ' of full rank with $\det(q|_{\Lambda'}) = 1$. Since $\det(q|_{\Lambda'}) = [\Lambda : \Lambda']^2 \det q$ and $\det q = 1$ (unimodularity, Poincaré duality), the index is 1: $\Lambda' = H^2(M;\mathbb{Z})$, and in this basis q is the identity.

Steps 6 and 7 together: $r \le m \le r$, so $m = r$ and q is diagonal. That is Donaldson's theorem.

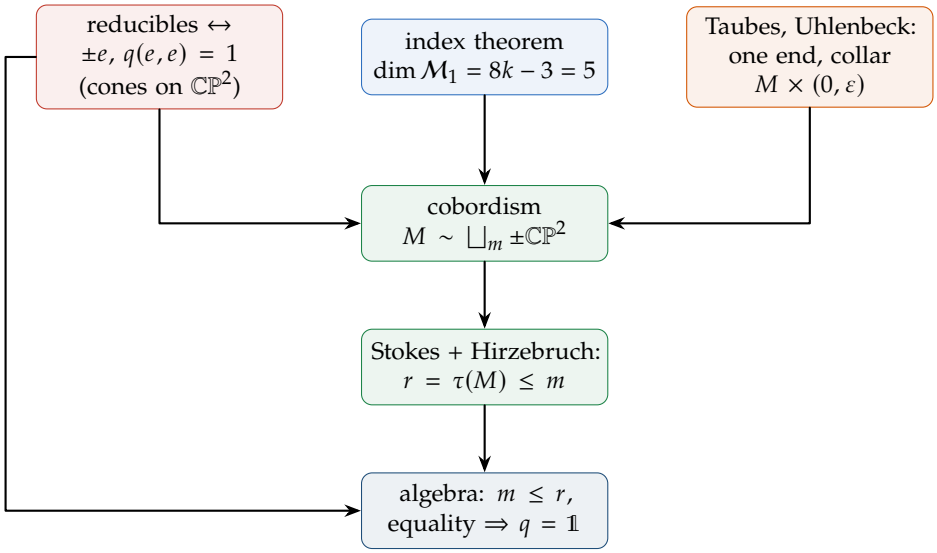


Figure 8.54: The logic of Donaldson's proof. Every ingredient of this section appears: the dimension from the index theorem (Section 8.9), the reducibles and their cones (Section 8.7), the collar of small instantons familiar from H^5 (Section 8.10), and the global structure of this section.

Worked example 8.108: The theorem at work

- $M = S^4$: $r = 0$, $m = 0$. W is the closed 5-ball with boundary S^4, and $\tau(S^4) = 0$. This is the compactified H^5 of Section 8.10.4.
- $M = \mathbb{CP}^2$: $q = (1)$, $m = 1$ (the pair $\pm h$ of Example 8.69). W is a cobordism from $\mathbb{CP}^2$ to one copy of $\mathbb{CP}^2$ (with the opposite induced orientation), and $1 - 1 = 0$. (For the Fubini–Study metric $\mathcal{M}_1$ is known explicitly: it is an open cone on $\mathbb{CP}^2$, with the reducible at the vertex and the collar at the open end.)
- $M = \mathbb{CP}^2 \# \mathbb{CP}^2$: $q = \mathrm{diag}(1,1)$, pairs $\pm(1,0)$, $\pm(0,1)$: $m = 2 = r$.
- A hypothetical smooth M with $q = E_8$: E_8 is even ($q(e,e)$ is even for all e), so it has no vectors with $q(e,e) = 1$, and $m = 0$. Then $\partial W = M$, so $\tau(M) = 0$; but $\tau(E_8) = 8$. Contradiction: the topological E_8 manifold of Freedman has no smooth structure.
- $q = (1) \oplus E_8$, rank 9: the only unit vectors are $\pm(1,0)$, so $m = 1 < 9$; excluded as well.

Physical picture: why a physical equation sees smooth structure

The self-duality equation needs a metric, and a metric needs a smooth structure. The moduli space is built from the equation, so it is a smooth invariant; its ends and singular points, however, are forced by topology (the collar is a copy of M, the cones are counted by q). The cobordism relates the two. A topological manifold that is not smoothable has no instantons to count, and the argument shows that for E_8 there could not be any consistent count.

8.11.3 An analogy: Riemann surfaces and Teichmüller theory

Nash compares Donaldson's use of $\mathcal{M}_k$ with the older theory of moduli of Riemann surfaces (Chapter 5): "In the latter the moduli of a linear elliptic operator—the $\bar\partial$-operator—are used to deduce topological information about Riemann surfaces, the main result being that dim $\mathcal{M}_p$ gives the genus". Let us make this concrete. A Riemann surface of genus p is a surface with a complex structure; its moduli space $\mathcal{M}_p$ parametrises complex structures up to diffeomorphism. The tangent space is computed by the same three-term pattern as for instantons, with the $\bar\partial$-operator on vector fields in place of d_A^-:

$$H^0(\Sigma, T\Sigma) = \{\text{holomorphic vector fields}\} \quad \text{(symmetries)},$$
$$H^1(\Sigma, T\Sigma) = \{\text{infinitesimal deformations of the complex structure}\}.$$

(By Serre duality $H^1(\Sigma, T\Sigma)$ is dual to the space of holomorphic quadratic differentials, which is how Teichmüller described it.)

Worked example 8.109: $\dim_{\mathbb{C}} \mathcal{M}_p = 3p-3$ from Riemann–Roch

For a holomorphic line bundle L on a surface of genus p, Riemann–Roch (the index theorem for $\bar\partial$, Chapter 4) says $h^0(L) - h^1(L) = \deg L + 1 - p$. The tangent bundle has degree $\deg T\Sigma = \chi(\Sigma) = 2 - 2p$ (Gauss–Bonnet or Poincaré–Hopf). So

$$\begin{aligned} h^0(T\Sigma) - h^1(T\Sigma) &\overset{(1)}{=} \deg T\Sigma + 1 - p \\ &\overset{(2)}{=} (2-2p) + 1 - p \\ &\overset{(3)}{=} 3 - 3p. \end{aligned}$$

(1) Riemann–Roch for $L = T\Sigma$.
(2) $\deg T\Sigma = 2 - 2p$.
(3) Collect.

For $p \geq 2$, $T\Sigma$ has negative degree, so it has no holomorphic sections: $h^0 = 0$ (no continuous symmetries), and $\dim_{\mathbb{C}} \mathcal{M}_p = h^1 = 3p - 3$, i.e. $6p - 6$ real dimensions. For $p = 1$: $h^0 = 1$ (translations of the torus), so $h^1 = 1$; the moduli space of tori is one complex dimensional. For $p = 0$: $h^0 = 3$ (the Möbius group $\mathrm{PSL}(2,\mathbb{C})$), $h^1 = 0$; the sphere is rigid. Inverting the first formula, $p = 1 + \dim_{\mathbb{R}} \mathcal{M}_p/6$: the dimension of the moduli space "gives the genus".

Worked example 8.110: The torus: another hyperbolic moduli space

A torus is $\mathbb{C}/(\mathbb{Z} + \tau\mathbb{Z})$ with $\operatorname{Im}\tau > 0$. Two values of τ give the same torus iff they are related by $\tau \mapsto (a\tau + b)/(c\tau + d)$ with $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}(2,\mathbb{Z})$, so $\mathcal{M}_1^{\text{tori}} = \mathbb{H}^2/\mathrm{SL}(2,\mathbb{Z})$, where $\mathbb{H}^2$ is the upper half-plane with its hyperbolic metric $(\mathrm{d}x^2 + \mathrm{d}y^2)/y^2$. Like the instanton moduli space $\mathcal{M}_1 = H^5$, it is a quotient of hyperbolic space, and it is non-compact: as $\tau \to i\infty$ the torus becomes infinitely long and thin, and in the limit it pinches to a sphere with two points identified. This degeneration plays the role of the instanton shrinking to a point.

Watch out: Nash's "$d_A^* d_A$"

Nash describes Donaldson's moduli as those of "the operator $d_A^* d_A$". He means the Yang–Mills operator $\mathrm{d}_A^* F_A$ (the equation $\mathrm{d}_A^* F = 0$, whose second-order linearisation is $\mathrm{d}_A^* \mathrm{d}_A$ on 1-forms) restricted to self-dual solutions. The operator whose index we actually

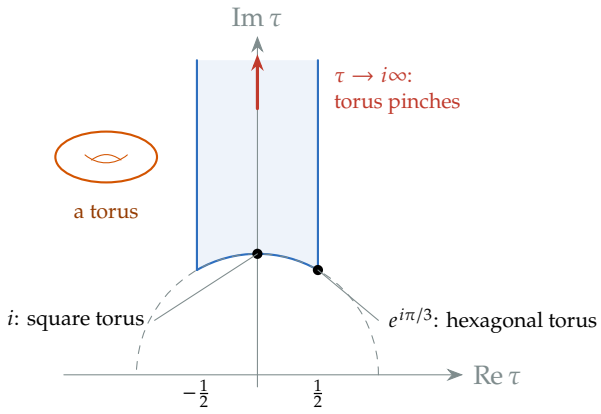


Figure 8.55: The moduli space of complex tori: the shaded fundamental domain of $\mathrm{SL}(2,\mathbb{Z})$ in the upper half-plane (with its sides identified). Each point is a torus $\mathbb{C}/(\mathbb{Z}+\tau\mathbb{Z})$. The space is a quotient of the hyperbolic plane and has one non-compact end, the cusp $\tau \to i\infty$, where the torus degenerates. Compare $\mathcal{M}_1 = H^5$ for instantons, whose end is made of instantons shrinking to points.

	Riemann surfaces	Instantons
object	complex structure on Σ_p	self-dual connection on M^4
operator	$\bar\partial$ on vector fields (linear)	d_A^-, d_A^* around A (linearised)
H^0	holomorphic vector fields	covariantly constant Λ (stabiliser)
H^1	deformations of complex structure	zero modes, $T_{[A]}\mathcal{M}_k$
H^2	none (complex dimension 1)	obstructions Ω^2_-
index	Riemann–Roch: $3-3p$	Atiyah–Singer: $3-8k$ (S^4)
dimension	$6p-6$ (real)	$8k-3$
non-compact end	pinching curves (cusps)	shrinking instantons (collar M)
use	$\dim \mathcal{M}_p$ determines p	$\partial W = M \sqcup m\,\mathbb{CP}^2$ constrains q

Table 8.4: Nash's analogy between Teichmüller theory and Donaldson theory. The same cohomological bookkeeping (H^0 symmetries, H^1 deformations, H^2 obstructions, index = expected dimension) organises both. The difference is linear versus non-linear: for instantons only the linearisation is elliptic, which is why the global conditions of Section 8.11.1 are needed.

computed is the first-order $D = \mathrm{d}_A + (\mathrm{d}_A^-)^*$ of Section 8.9.1.

8.11.4 Instantons and the topology of $\mathcal{A}/\mathcal{G}$: the Atiyah–Jones theorem

The last idea of Nash's section runs in the opposite direction. Instead of using the topology of M to learn about $\mathcal{M}_k$, we use $\mathcal{M}_k$ to learn about the **configuration space** $\mathcal{A}/\mathcal{G}$, the space of all gauge fields up to gauge, over which the Yang–Mills path integral is taken.

The homotopy type of $\mathcal{A}/\mathcal{G}_0$

Work with the based gauge group $\mathcal{G}_0$ (gauge transformations equal to $\mathbb{1}$ at the base point), since it acts freely (Section 8.7.2): if $g \cdot A = A$ then g is covariantly constant (Section 8.7.3), so it is determined by its value at the base point, which is $\mathbb{1}$. Two facts determine the topology.

$\mathcal{A}$ is contractible. It is an affine space: any A can be joined to a fixed A_0 by the straight line $A_0 + t(A - A_0)$, $t \in [0,1]$, and these lines depend continuously on A. So $\mathcal{A}$ deforms to the point A_0.

A contractible space with a free action. When a group $\mathcal{G}_0$ acts freely on a contractible space $\mathcal{A}$, the fibration $\mathcal{G}_0 \to \mathcal{A} \to \mathcal{A}/\mathcal{G}_0$ and the long exact sequence of homotopy groups give

$$\begin{aligned} \pi_i(\mathcal{A}/\mathcal{G}_0) &\overset{(1)}{=} \pi_{i-1}(\mathcal{G}_0) \\ &\overset{(2)}{=} \pi_{i-1}\big(\mathrm{Map}_*(S^4, S^3)\big) \\ &\overset{(3)}{=} \pi_{i+3}(S^3). \end{aligned} \tag{8.262}$$

(1) The exact sequence $\pi_i(\mathcal{A}) \to \pi_i(\mathcal{A}/\mathcal{G}_0) \to \pi_{i-1}(\mathcal{G}_0) \to \pi_{i-1}(\mathcal{A})$ with $\pi_\bullet(\mathcal{A}) = 0$.

(2) On S^4, a based gauge transformation is, up to homotopy, a based map $S^4 \to \mathrm{SU}(2) = S^3$ (trivialise the bundle away from a point; the twisting of P_k does not change the homotopy type of $\mathcal{G}_0$).

(3) A based map $S^j \to \mathrm{Map}_*(S^4, S^3)$ is the same as a based map $S^j \wedge S^4 = S^{j+4} \to S^3$, so $\pi_j(\mathrm{Map}_*(S^4, S^3)) = \pi_{j+4}(S^3)$; put $j = i - 1$.

In the language of loop spaces, $\mathcal{A}_k/\mathcal{G}_0 \simeq \Omega^3_k S^3$, a component of the space of maps $S^3 \to S^3$ (based), the component of degree k. The first few groups are

i	1	2	3	4	5	6
$\pi_i(\mathcal{A}/\mathcal{G}_0) = \pi_{i+3}(S^3)$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_{12}$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	$\mathbb{Z}_3$

The configuration space of SU(2) gauge theory on S^4 is topologically complicated. For instance $\pi_1 = \pi_4(S^3) = \mathbb{Z}_2$: there is a non-contractible loop of gauge fields. (This loop is behind Witten's global SU(2) anomaly, Chapter 10.) It is also why there is no continuous global gauge slice, the topological root of the Gribov ambiguity (Singer).

The inclusion of the instantons

Every framed instanton is a point of $\mathcal{A}_k/\mathcal{G}_0$, so there is an inclusion (Nash's (8.138))

$$i(k) : \ \mathcal{M}'_k \longrightarrow \mathcal{A}_k/\mathcal{G}_0. \tag{8.263}$$

$\mathcal{M}'_k$ has dimension $8k$ and $\mathcal{A}_k/\mathcal{G}_0$ is infinite dimensional; one cannot expect $i(k)$ to be a homotopy equivalence. But it can capture the topology in low dimensions.

Key idea: The Atiyah–Jones theorem and conjecture

Atiyah and Jones (1978) proved that $i(k)$ induces a surjection $H_q(\mathcal{M}'_k) \to H_q(\mathcal{A}_k/\mathcal{G}_0)$ in a range of dimensions $q < q(k)$ that grows with k. They conjectured that it is an isomorphism in such a range. This *Atiyah–Jones conjecture* was proved for SU(2) on S^4 by Boyer, Hurtubise, Mann and Milgram (1993), with a range growing linearly in k (roughly $q < k/2$). So as $k \to \infty$, "more and more of the homology of $\mathcal{A}/\mathcal{G}$ is captured by that of $\mathcal{M}'_k$": the finite-dimensional space of instantons approximates the infinite-dimensional space of all gauge fields.

Worked example 8.111: Low-dimensional check at $k = 1$

By Example 8.99, $\mathcal{M}'_1 = \mathbb{R}^4 \times (\mathbb{R}^4 \setminus 0)/\mathbb{Z}_2$. Shrinking the $\mathbb{R}^4$ factor to a point and the radial direction λ to $\lambda = 1$ is a deformation retraction, so

$$\mathcal{M}'_1 \simeq S^3/\mathbb{Z}_2 = \mathbb{RP}^3 = \mathrm{SO}(3), \qquad \pi_1(\mathcal{M}'_1) = H_1(\mathcal{M}'_1) = \mathbb{Z}_2.$$

On the other side, $\pi_1(\mathcal{A}_1/\mathcal{G}_0) = \pi_4(S^3) = \mathbb{Z}_2$ by (8.262). So both spaces have fundamental group $\mathbb{Z}_2$, and the candidate generator on the instanton side is the loop that rotates the framing of a single instanton through SO(3). Whether $i(1)$ carries one $\mathbb{Z}_2$ onto the other is exactly the kind of question the theorem answers for large k; at $k = 1$ it needs a separate argument.

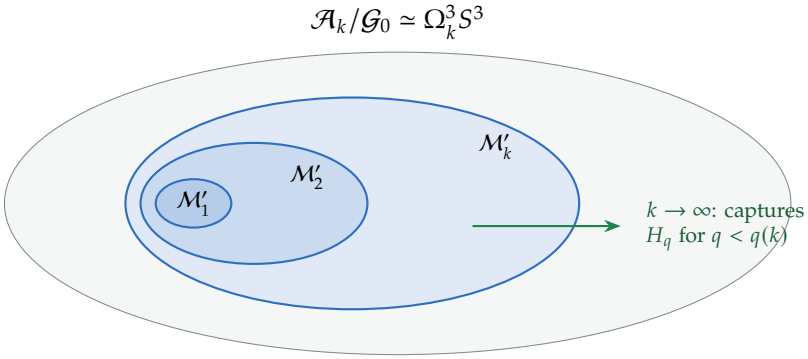


Figure 8.56: The Atiyah–Jones picture, schematically. The framed moduli spaces have dimension $8k$ and sit inside the infinite-dimensional configuration space. (Strictly, the spaces for different k lie in different components $\mathcal{A}_k/\mathcal{G}_0$, all homotopy equivalent to each other.) As k grows, the inclusion captures more and more of the homology of the configuration space: many-instanton configurations explore its topology.

Physical picture: a dilute instanton gas explores configuration space

Why should large k help? A charge-k instanton can be pulled apart into k widely separated, almost independent BPST lumps, each with its own position, size and framing (8 parameters each, matching $8k$). The topology of such configurations is that of k "particles" on S^4 carrying internal SO(3) labels, and configuration spaces of many particles are known to approximate loop spaces as the number of particles grows (this is the mechanism behind the proof). So the "instanton gas" of the semiclassical physicist is also a good model of the topology of the full space of gauge fields.

Reading Nash (pp. 240–242, (8.138))

Nash's warning that the index calculation "is an infinitesimal one" and his three conditions (existence; extending the infinitesimal coordinates at $[A]$ to a neighbourhood; fitting the local coordinates into a global Hausdorff space) are Section 8.11.1; the Kuranishi model and the estimate $|\mathrm{d}g| \le |A| + |A_g|$ are our additions. His references to Atiyah–Hitchin–Singer for S^4 correspond to the keyidea there. His remark that the S^4 space $\mathcal{M}_1$ "is the space used in the proof of Donaldson's theorem" and that "the boundary of $\mathcal{M}_1$ is M" for the other M is Section 8.11.2, where we sketched the whole argument. The Teichmüller analogy is Section 8.11.3 and Table 8.4. The inclusion $i(k) : \mathcal{M}'_k \to \mathcal{A}/\mathcal{G}$ (8.138) and the statement that for large k it "induces a projection of the homology of $\mathcal{M}'_k$ onto part of the homology of $\mathcal{A}/\mathcal{G}$" is the Atiyah–Jones theorem of Section 8.11.4; Nash's $\mathcal{A}/\mathcal{G}$ there means the quotient by the based gauge group, which is why the framed space $\mathcal{M}'_k$ appears.

Recap

- The index gives the *expected* dimension. It is the true dimension near $[A]$ when the obstruction space H^2 vanishes and the stabiliser is the centre; the Kuranishi model $\kappa^{-1}(0)/\Gamma_A$ describes the general case.
- The three global conditions: existence (BPST/ADHM on S^4, Taubes' grafting in general), integrability ($H^2 = 0$: Weitzenböck on S^4, generic metrics in general), Hausdorff (the gauge group acts properly, by $|\mathrm{d}g| \le |A| + |A_g|$).
- Donaldson: for M simply connected with $q > 0$, $\mathcal{M}_1$ is a 5-manifold with a collar end $\cong M \times (0, \varepsilon)$ and m cones on $\mathbb{CP}^2$, $m = \#\{\pm e : q(e,e) = 1\}$. Cobordism and $\tau(\partial W) = 0$ give $r \le m$; algebra gives $m \le r$ with equality iff q is diagonal.
- Teichmüller analogy: the same $H^0/H^1/H^2$ bookkeeping, linear for Riemann surfaces ($\dim_{\mathbb{C}} \mathcal{M}_p = 3p - 3$), non-linear for instantons.
- Atiyah–Jones: $\mathcal{A}_k/\mathcal{G}_0 \simeq \Omega^3_k S^3$ with $\pi_i = \pi_{i+3}(S^3)$, and $\mathcal{M}'_k \hookrightarrow \mathcal{A}_k/\mathcal{G}_0$ captures its homology in a range growing with k.

8.12 Monopoles: the Yang–Mills–Higgs system in three dimensions

So far in this chapter the stage has been four-dimensional and Euclidean, and the stars were instantons: finite-action solutions of $F = \pm \star F$. Now we drop one dimension. Nash opens his Section 4 by remarking that the Yang–Mills equations $\mathrm{d}_A^\star F = 0$ "change considerably" when the dimension of M changes. Here is the physical reason we care about dimension three.

A field configuration in $3 + 1$ dimensions that does not change in time is a field on *space*, which is three-dimensional. Its energy is an integral over $\mathbb{R}^3$. A **soliton** is a static configuration of finite energy that cannot decay, because some integer attached to it (a topological charge) cannot change continuously. Physically it is a particle made out of the fields themselves. The question of this section is:

> *Which gauge theories have solitons in three space dimensions, what integer protects them, and what equations do they satisfy?*

The answer is the **magnetic monopole** of 't Hooft and Polyakov. It needs a Higgs field as well as a gauge field. Its integer is a magnetic charge. The charge turns out to be a Chern number and also a degree of a map between two spheres. In the special "Prasad–Sommerfield" limit the monopoles solve a first-order equation, the **Bogomolny equation**, which is nothing but the self-duality equation of the previous section in disguise. That link, monopoles as symmetric instantons, is the thread that runs through Nash's Sections 4 and 5.

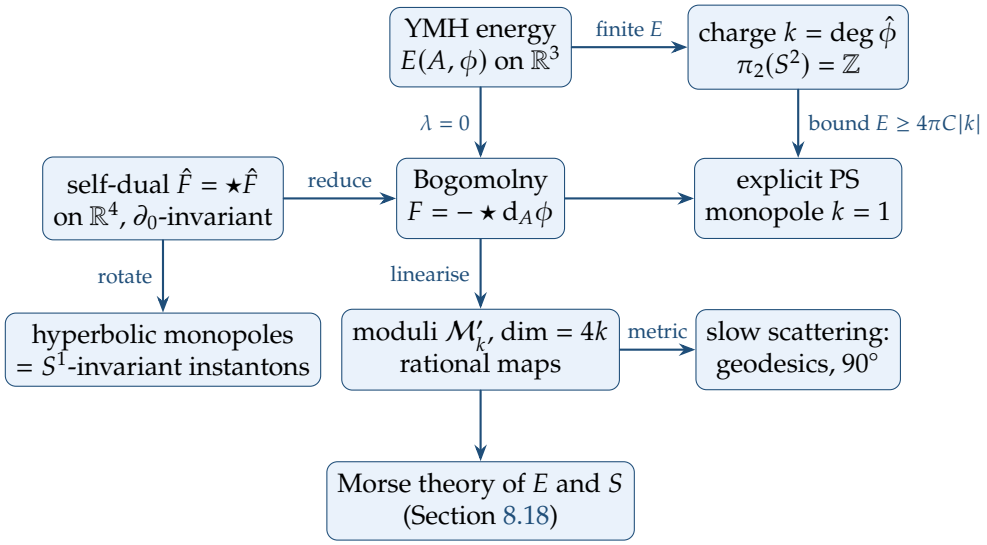


Figure 8.57: The logic of the last three sections of Nash's Chapter 8. Start at the top with the energy of a gauge field plus Higgs field in three dimensions. Follow the arrows: topology forces a charge, the charge bounds the energy, the bound is saturated by solutions of the Bogomolny equation, and those solutions form a finite-dimensional moduli space whose geometry governs how monopoles scatter. The left column shows the link to instantons.

8.12.1 Conventions for this part

We work with $G = \mathrm{SU}(2)$ throughout, as Nash does from p. 244 on. Its Lie algebra $\mathfrak{su}(2)$ is the space of 2×2 anti-Hermitian traceless matrices. We use the basis

$$T_a = -\frac{i}{2}\sigma_a, \qquad T_1 = \frac{1}{2}\begin{pmatrix} 0 & -i \\ -i & 0 \end{pmatrix}, \quad T_2 = \frac{1}{2}\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad T_3 = \frac{1}{2}\begin{pmatrix} -i & 0 \\ 0 & i \end{pmatrix}. \tag{8.264}$$

Everything we need about these matrices follows from the Pauli rule $\sigma_a\sigma_b = \delta_{ab}\mathbb{1} + i\varepsilon_{abc}\sigma_c$.

Worked example 8.112: The two facts about T_a we use constantly

First the commutator:

$$\begin{aligned} [T_a, T_b] &\overset{(1)}{=} \left(-\frac{i}{2}\right)^2 [\sigma_a, \sigma_b] \\ &\overset{(2)}{=} -\frac{1}{4}\cdot 2i\varepsilon_{abc}\sigma_c \\ &\overset{(3)}{=} \varepsilon_{abc}\, T_c. \end{aligned}$$

(1) Pull the scalar $-i/2$ out of both factors.
(2) The Pauli rule gives $\sigma_a\sigma_b - \sigma_b\sigma_a = 2i\varepsilon_{abc}\sigma_c$.
(3) $-\frac{i}{2}\sigma_c = T_c$.

Then the trace:

$$\mathrm{tr}(T_aT_b) \overset{(1)}{=} -\frac{1}{4}\,\mathrm{tr}(\sigma_a\sigma_b) \overset{(2)}{=} -\frac{1}{4}(2\delta_{ab}) = -\frac{1}{2}\delta_{ab}.$$

(1) $(-i/2)^2 = -1/4$.
(2) $\mathrm{tr}\,\mathbb{1} = 2$ and $\mathrm{tr}\,\sigma_c = 0$.

So if we write $X = X^aT_a$ and $Y = Y^aT_a$ with real components, then $[X,Y] = (\vec{X}\times\vec{Y})^aT_a$, and $-2\,\mathrm{tr}(XY) = \vec{X}\cdot\vec{Y}$. The Lie algebra $\mathfrak{su}(2)$ *is* $\mathbb{R}^3$ with the cross product.

Working definition 8.37 (pointwise and L^2 inner products in this part)

For $X, Y \in \mathfrak{su}(2)$ we set

$$(X,Y) = -2\,\mathrm{tr}(XY) = \vec{X}\cdot\vec{Y}, \qquad |X|^2 = -2\,\mathrm{tr}(X^2) = \sum_a (X^a)^2. \tag{8.265}$$

For $\mathfrak{su}(2)$-valued p-forms on a Riemannian manifold M [a metric is given, and the Hodge star $\star$ built from it makes $\alpha \wedge \star\beta$ the dot product of α and β times the volume form],

$$\langle \alpha, \beta\rangle = -2\int_M \mathrm{tr}(\alpha\wedge\star\beta) = \sum_a \int_M \alpha^a \wedge \star\beta^a, \qquad \|\alpha\|^2 = \langle\alpha,\alpha\rangle. \tag{8.266}$$

Operational meaning: expand in the basis T_a and use the ordinary dot product of components. $|X|$ is the length of the vector $\vec{X} \in \mathbb{R}^3$.

Watch out: A factor of 2 in Nash's traces

Nash defines $|\phi|^2 = -2\,\mathrm{tr}(\phi^2)$ (just after (8.139)), which agrees with (8.265). But in the energy (8.139), in (8.141) and in (8.148) he writes $-\,\mathrm{tr}$ without the 2. With his normalisation taken literally, the energy of the unit monopole would be 2π instead of the 4π he quotes in (8.191). Using $-2\,\mathrm{tr}$ everywhere, as we do, makes all his final numbers ($E = 4\pi k$, (8.155), $\dim \mathcal{M}'_k = 4k$) come out right. Earlier parts of these notes use $-\,\mathrm{tr}$; the difference is an overall factor of 2 in the definition of the action, and no equation of motion changes.

Gauge fields follow the conventions of the whole book: A is an $\mathfrak{su}(2)$-valued 1-form, $F = \mathrm{d}A + A\wedge A$, and under a gauge transformation $g : M \to \mathrm{SU}(2)$,

$$A \mapsto g^{-1}Ag + g^{-1}\mathrm{d}g, \qquad F \mapsto g^{-1}Fg.$$

The Higgs field ϕ is an $\mathfrak{su}(2)$-valued function (a section of $\mathrm{ad}\,P$) and transforms like F, $\phi \mapsto g^{-1}\phi g$. Its covariant derivative is the 1-form

$$\mathrm{d}_A\phi = \mathrm{d}\phi + [A,\phi], \qquad (D_i\phi)^a = \partial_i\phi^a + \varepsilon_{abc}A_i^b\phi^c. \tag{8.267}$$

The component form uses Example 8.112: $[A_i,\phi] = (\vec{A}_i \times \vec{\phi})^a T_a$.

Dictionary: mathematicians' fields versus physicists' fields

A physics text writes the gauge field as A_μ^a with Hermitian generators $\sigma_a/2$ and coupling e, and the Higgs triplet as ϕ^a_{phys} with vacuum value v. The translation is

$$A_\mu^{\text{math}} = -ie\,A_\mu^a\frac{\sigma_a}{2} = e\,A_\mu^a T_a, \qquad \phi^{\text{math}} = e\,\phi^a_{\text{phys}}T_a, \qquad C = e\,v.$$

Then $F^{\text{math}} = eF^aT_a$ with $F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu + e\,\varepsilon_{abc}A^b_\mu A^c_\nu$, and the physical energy is $E_{\text{phys}} = E^{\text{math}}/e^2$. So the monopole mass $E = 4\pi C$ that we shall find becomes $M = 4\pi v/e$, the textbook value. We set $e = 1$ from now on.

The Hodge star in three dimensions will appear everywhere. On $\mathbb{R}^3$ with the Euclidean metric and orientation $\mathrm{d}x^1\wedge\mathrm{d}x^2\wedge\mathrm{d}x^3$,

$$\star\mathrm{d}x^1 = \mathrm{d}x^2\wedge\mathrm{d}x^3, \quad \star\mathrm{d}x^2 = \mathrm{d}x^3\wedge\mathrm{d}x^1, \quad \star\mathrm{d}x^3 = \mathrm{d}x^1\wedge\mathrm{d}x^2, \qquad \star\star = 1. \tag{8.268}$$

In components, a 2-form $F = \tfrac12 F_{ij}\mathrm{d}x^i\wedge\mathrm{d}x^j$ has $\star F = B_i\,\mathrm{d}x^i$ with $B_i = \tfrac12\varepsilon_{ijk}F_{jk}$: the Hodge star turns the "field-strength tensor" into the "magnetic field vector". The formal adjoint of d_A is

$\mathrm{d}_A^\star = -\star\,\mathrm{d}_A\star$ on 1-forms and $\mathrm{d}_A^\star = +\star\,\mathrm{d}_A\star$ on 2-forms in three dimensions (the general rule is $\mathrm{d}^\star = (-1)^{n(p+1)+1}\star\,\mathrm{d}\star$ on p-forms in n dimensions, Chapter T1).

8.12.2 Why pure Yang–Mills theory has no solitons in three dimensions

Why does Nash need a Higgs field at all? Could a pure gauge field on $\mathbb{R}^3$ already form a lump of finite energy? A one-line scaling argument, due to Derrick, says no.

Worked example 8.113: Derrick's scaling argument

Take any static configuration (A, ϕ) on $\mathbb{R}^3$ and squeeze it by a factor $\mu > 0$:

$$A^{(\mu)}(x) = \mu\, A_i(\mu x)\,\mathrm{d}x^i, \qquad \phi^{(\mu)}(x) = \phi(\mu x).$$

(The extra μ in A makes $\mathrm{d} + A$ scale like a derivative.) Then $F_{ij}^{(\mu)}(x) = \mu^2 F_{ij}(\mu x)$ and $D_i^{(\mu)}\phi^{(\mu)}(x) = \mu\,(D_i\phi)(\mu x)$. Write $E = E_F + E_D + E_V$ for the magnetic, gradient and potential energies. Changing variables $y = \mu x$, $\mathrm{d}^3x = \mu^{-3}\mathrm{d}^3y$:

$$E_F^{(\mu)} = \frac{1}{2}\int |B^{(\mu)}(x)|^2\mathrm{d}^3x \overset{(1)}{=} \frac{1}{2}\int \mu^4|B(\mu x)|^2\mathrm{d}^3x \overset{(2)}{=} \mu\, E_F,$$
$$E_D^{(\mu)} = \frac{1}{2}\int |D^{(\mu)}\phi^{(\mu)}(x)|^2\mathrm{d}^3x \overset{(3)}{=} \mu^{-1}E_D, \qquad E_V^{(\mu)} \overset{(4)}{=} \mu^{-3}E_V.$$

(1) Each component of F scales as μ^2, so $|B|^2$ scales as μ^4.
(2) Substitute $y = \mu x$: $\mu^4 \cdot \mu^{-3} = \mu$.
(3) Similarly $\mu^2 \cdot \mu^{-3} = \mu^{-1}$ for the gradient term.
(4) The potential has no derivatives: $\mu^0 \cdot \mu^{-3}$.

A static solution is a critical point of E, so it must be stationary under this particular deformation: $\mathrm{d}E^{(\mu)}/\mathrm{d}\mu|_{\mu=1} = E_F - E_D - 3E_V = 0$. For *pure* Yang–Mills ($\phi = 0$) this says $E_F = 0$, i.e. $F = 0$: no lump. The field can always lower its energy by spreading out. With a Higgs field the gradient term pushes the other way, and a balance $E_F = E_D + 3E_V$ is possible.

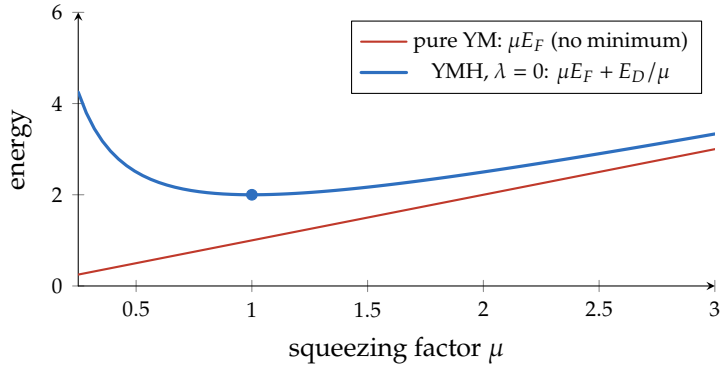


Figure 8.58: Derrick's argument in a picture, with $E_F = E_D = 1$. The pure gauge energy μE_F (red) decreases all the way to $\mu \to 0$: the lump spreads out and disappears. Adding a Higgs gradient term (blue) produces a minimum at $\mu = 1$, where $E_F = E_D$. That equality is exactly what the Bogomolny equation will give.

8.12.3 A warm-up with one gauge field: the Dirac monopole

Before adding the Higgs field, recall the oldest monopole, Dirac's, because the non-abelian monopole will look exactly like it from far away. Maxwell's equation $\nabla \cdot \vec{B} = 0$ forbids magnetic charges *if* $\vec{B}$ is the curl of a vector potential everywhere. Dirac asked what happens if we give up "everywhere".

Place a magnetic charge at the origin: $\vec{B} = g\,\hat{\vec{r}}/r^2$. As a 2-form on $\mathbb{R}^3 \setminus \{0\}$, using spherical coordinates,

$$F^{\text{Maxwell}} = \star(B_i \mathrm{d}x^i) = g\,\sin\theta\,\mathrm{d}\theta \wedge \mathrm{d}\varphi. \tag{8.269}$$

It is closed, but its integral over any sphere around the origin is $4\pi g \neq 0$, so by Stokes it is not exact: there is no global potential (this is the example in Chapter T2). Locally there are potentials. On the sphere with the south pole removed, and on the sphere with the north pole removed,

$$a_N = g(1-\cos\theta)\,\mathrm{d}\varphi, \qquad a_S = -g(1+\cos\theta)\,\mathrm{d}\varphi, \qquad \mathrm{d}a_N = \mathrm{d}a_S = F^{\text{Maxwell}}. \tag{8.270}$$

Each is singular where the other is fine: a_N blows up at $\theta = \pi$ (the coefficient of $\mathrm{d}\varphi$ does not vanish where the circle shrinks), and a_S at $\theta = 0$. The line where a potential is singular is the **Dirac string**. On the overlap the two differ by $a_N - a_S = 2g\,\mathrm{d}\varphi$, a gauge transformation.

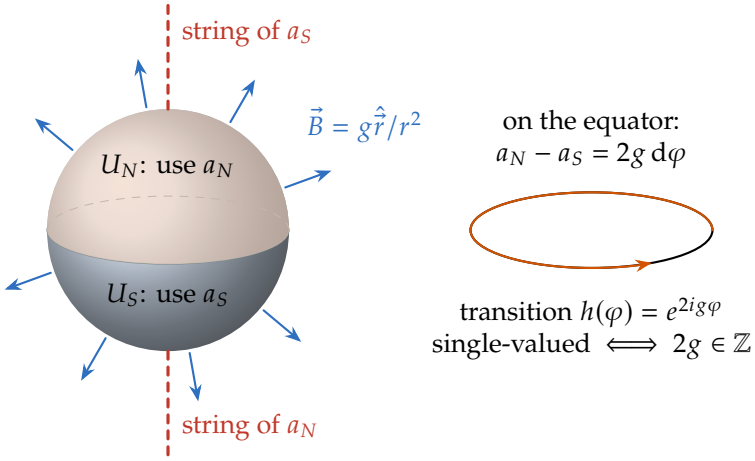


Figure 8.59: The Dirac monopole as two patches. The northern cap uses a_N, the southern cap a_S; each has a string of singularity where the other patch lives, so neither string is ever "seen". On the equatorial overlap the two potentials are related by the U(1) gauge transformation $h = e^{2ig\varphi}$, and this must be a genuine function on the circle.

Worked example 8.114: Dirac quantisation from single-valuedness

A charged wavefunction ψ of unit electric charge couples through $\mathrm{d} - ia$ (physicists' convention for U(1), with a real). If $a_N = a_S + \mathrm{d}\chi$, the wavefunctions in the two patches are related by $\psi_N = e^{i\chi}\psi_S$. Here $\chi = 2g\varphi$. Going once round the equator, $\varphi \to \varphi + 2\pi$, the relation must return to itself:

$$e^{i\chi(\varphi+2\pi)} \overset{(1)}{=} e^{i2g(\varphi+2\pi)} \overset{(2)}{=} e^{i\chi(\varphi)}\,e^{4\pi i g} \overset{(3)}{=} e^{i\chi(\varphi)} \qquad \Longleftrightarrow \qquad 2g = n \in \mathbb{Z}.$$

(1) Insert $\chi = 2g\varphi$.
(2) Split the exponential.
(3) We demand single-valuedness of ψ_N given ψ_S; this holds for every φ exactly when $e^{4\pi i g} = 1$.

The flux is then $\int_{S^2} F^{\text{Maxwell}} = 4\pi g = 2\pi n$. We can recover this from the patches alone, by

Stokes on each hemisphere:

$$\int_{S^2} F^{\text{Maxwell}} \overset{(1)}{=} \int_{U_N} \mathrm{d}a_N + \int_{U_S} \mathrm{d}a_S \overset{(2)}{=} \oint_{\text{eq}} a_N - \oint_{\text{eq}} a_S \overset{(3)}{=} \oint_{\text{eq}} \mathrm{d}\chi = 2\pi n.$$

(1) Split the sphere at the equator and use the regular potential in each half.
(2) Stokes; the equator is the boundary of U_N with the positive orientation and of U_S with the negative one.
(3) $a_N - a_S = \mathrm{d}\chi$; the winding of $\chi = n\varphi$ round the circle is $2\pi n$.

In the language of Chapter 1, $\frac{1}{2\pi}F^{\text{Maxwell}}$ represents the first Chern class of a U(1) bundle over S^2, and $n = c_1$ is an integer because it counts the winding of the transition function. For $n = 1$ the bundle is the Hopf bundle $S^3 \to S^2$.

Two lessons carry over. First, a magnetic charge is a *topological* number: it measures how the patches are glued, not the details of the field. Second, the Dirac monopole is singular at the origin, and its energy $\frac{1}{2}\int |\vec{B}|^2 \mathrm{d}^3x$ diverges there. The 't Hooft–Polyakov monopole cures the singularity by embedding U(1) in SU(2) and letting a Higgs field decide, point by point, which U(1) is "electromagnetism".

8.12.4 The Yang–Mills–Higgs energy

The system Nash studies is a pair (A, ϕ): an SU(2) connection A on a bundle $P \to M$ over a three-manifold M, and a Higgs field ϕ in the adjoint representation. Its energy is Nash's (8.139), written with our trace normalisation:

$$E(A,\phi) = \frac{1}{2}\int_M \Big\{ -2\,\mathrm{tr}(F \wedge \star F) - 2\,\mathrm{tr}(\mathrm{d}_A\phi \wedge \star \mathrm{d}_A\phi) + \lambda\big(|\phi|^2 - C^2\big)^2 \star 1 \Big\}. \tag{8.271}$$

On $\mathbb{R}^3$, in components, this is the familiar

$$E = \int_{\mathbb{R}^3} \Big\{ \frac{1}{2} B_i^a B_i^a + \frac{1}{2} D_i\phi^a \, D_i\phi^a + \frac{\lambda}{2}\big(\phi^a\phi^a - C^2\big)^2 \Big\} \mathrm{d}^3x, \tag{8.272}$$

because $-2\,\mathrm{tr}(F \wedge \star F) = \sum_a F^a \wedge \star F^a = B_i^a B_i^a \, \mathrm{d}^3x$ and similarly for the gradient term.

Read term by term:

- $\frac{1}{2}|B|^2$ is magnetic energy (the fields are static and we work in the gauge $A_0 = 0$, so there is no electric field).
- $\frac{1}{2}|D\phi|^2$ is the gradient energy of the Higgs field; the covariant derivative makes it gauge invariant.
- $\frac{\lambda}{2}(|\phi|^2 - C^2)^2$ is the Mexican-hat potential. It vanishes on the sphere $|\phi| = C$ in $\mathfrak{su}(2) \cong \mathbb{R}^3$.

Each term is non-negative. So the energy vanishes exactly when $F = 0$, $\mathrm{d}_A\phi = 0$ and $|\phi| = C$. On $\mathbb{R}^3$ this means that, up to a gauge transformation, $A = 0$ and ϕ is a constant of length C: the Higgs **vacuum**.

Working definition 8.38 (vacuum manifold)

The **vacuum manifold** is the set of Higgs values of zero potential energy,

$$\mathcal{V} = \{\phi \in \mathfrak{su}(2) : |\phi| = C\} \cong S^2.$$

Operational meaning: it is the set of directions in isospin space [$\mathfrak{su}(2) \cong \mathbb{R}^3$, where the Higgs field takes its values] in which the Higgs field can point at no cost. SU(2) acts on it by conjugation, which rotates $\mathbb{R}^3$. The stabiliser [the elements g with $g^{-1}\phi g = \phi$] of $\phi = CT_3$ is the circle $\{e^{\theta T_3}\} \cong \mathrm{U}(1)$, so $\mathcal{V} \cong \mathrm{SU}(2)/\mathrm{U}(1)$. This unbroken U(1) is electromagnetism.

Physical picture: Symmetry breaking in pictures

Think of $\phi(x)$ as a little arrow in an internal three-dimensional space, attached to each point x. In the vacuum all arrows have length C and are parallel (after a gauge transformation). Rotations about the common arrow do not disturb the vacuum; that is the unbroken U(1). The two gauge fields that rotate the arrow away from its direction get a mass (the $W^\pm$ bosons), and the photon, the gauge field of rotations about the arrow, stays massless.

Worked example 8.115: Masses in the vacuum

Put $\phi = C\,T_3$ and a small gauge field $A = A_i^a T_a \mathrm{d}x^i$. The gradient term is

$$\frac{1}{2}|D_i\phi|^2 \overset{(1)}{=} \frac{1}{2}\big|[A_i, CT_3]\big|^2 \overset{(2)}{=} \frac{C^2}{2}\big|\vec{A}_i \times \hat{\vec{e}}_3\big|^2 \overset{(3)}{=} \frac{C^2}{2}\big((A_i^1)^2 + (A_i^2)^2\big).$$

(1) ϕ is constant, so $\mathrm{d}\phi = 0$.
(2) Commutators are cross products (Example 8.112).
(3) $\vec{A} \times \hat{\vec{e}}_3 = (A^2, -A^1, 0)$.

This is a mass term $\frac{1}{2}m_W^2 A^2$ with $m_W = C$ for A^1, A^2 and no mass for A^3, which is the photon. In physical units $m_W = ev$.

8.12.5 The field equations

The static field equations are the Euler–Lagrange equations of E. We derive them by varying $A \to A + ta$ and $\phi \to \phi + t\eta$, with a, η of compact support, and demanding $\mathrm{d}E/\mathrm{d}t|_{t=0} = 0$. We need the first-order changes

$$F(A + ta) = F + t\,\mathrm{d}_A a + t^2 a \wedge a, \qquad \mathrm{d}_{A+ta}(\phi + t\eta) = \mathrm{d}_A\phi + t\big(\mathrm{d}_A\eta + [a, \phi]\big) + t^2[a, \eta]. \quad (8.273)$$

Worked example 8.116: Deriving Nash's (8.140)

Differentiate E at $t = 0$. Using the symmetry of $\langle\,,\rangle$,

$$\begin{aligned}\frac{\mathrm{d}E}{\mathrm{d}t}\Big|_0 &\overset{(1)}{=} \langle F, \mathrm{d}_A a\rangle + \langle \mathrm{d}_A\phi, \mathrm{d}_A\eta + [a,\phi]\rangle + 2\lambda \int (|\phi|^2 - C^2)(\phi, \eta) \star 1 \\ &\overset{(2)}{=} \langle \mathrm{d}_A^\star F, a\rangle + \langle \mathrm{d}_A^\star \mathrm{d}_A\phi, \eta\rangle + \langle \mathrm{d}_A\phi, [a,\phi]\rangle + \big\langle 2\lambda(|\phi|^2 - C^2)\phi, \eta\big\rangle \\ &\overset{(3)}{=} \big\langle \mathrm{d}_A^\star F - [\mathrm{d}_A\phi, \phi],\, a\big\rangle + \big\langle \mathrm{d}_A^\star \mathrm{d}_A\phi + 2\lambda(|\phi|^2 - C^2)\phi,\, \eta\big\rangle.\end{aligned}$$

(1) Differentiate each square: $\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}\|X + tY\|^2 = \langle X, Y\rangle$; for the potential, $\frac{\mathrm{d}}{\mathrm{d}t}|\phi + t\eta|^2 = 2(\phi, \eta)$, so $\frac{\mathrm{d}}{\mathrm{d}t}\frac{\lambda}{2}(|\phi|^2 - C^2)^2 = 2\lambda(|\phi|^2 - C^2)(\phi, \eta)$.
(2) Move d_A across the inner product as its adjoint $\mathrm{d}_A^\star$ (integration by parts; boundary terms vanish because a, η have compact support).
(3) Pointwise, $(\mathrm{d}_A\phi, [a,\phi]) = \sum_i D_i\vec{\phi} \cdot (\vec{a}_i \times \vec{\phi}) = \sum_i \vec{a}_i \cdot (\vec{\phi} \times D_i\vec{\phi}) = -(a, [\mathrm{d}_A\phi, \phi])$, by the cyclic property of the triple product.

The variation vanishes for all a and η only if

$$\mathrm{d}_A^\star F = [\mathrm{d}_A\phi, \phi], \qquad \mathrm{d}_A^\star \mathrm{d}_A\phi = -2\lambda\,\phi\,(|\phi|^2 - C^2). \tag{8.274}$$

The first equation is the non-abelian Ampère law: the magnetic field is sourced by the current $[\mathrm{d}_A\phi, \phi]$ carried by the charged components of the Higgs field. The second is a covariant nonlinear Klein–Gordon equation; on $\mathbb{R}^3$, since $\mathrm{d}_A^\star \mathrm{d}_A = -D_iD_i$, it reads $D_iD_i\phi = 2\lambda\phi(|\phi|^2 - C^2)$, i.e. "Laplacian of ϕ equals the derivative of the potential".

Watch out: Signs and factors in (8.140)

Nash's (8.140) reads $\mathrm{d}_A^\star \mathrm{d}_A\phi = 4\lambda\phi(|\phi|^2 - C^2)$. The 4 instead of our 2 comes from his use of $-\operatorname{tr}$ rather than $-2\operatorname{tr}$ in the inner product (the gradient term is then half as big relative to the potential). The sign is a slip: $\mathrm{d}_A^\star \mathrm{d}_A$ is a *positive* operator ($\langle\phi, \mathrm{d}_A^\star \mathrm{d}_A\phi\rangle = \|\mathrm{d}_A\phi\|^2 \geq 0$), and for the potential to push $|\phi|$ back towards C the right-hand side must carry a minus sign. Nothing later depends on this, because Nash immediately sets $\lambda = 0$.

These equations are coupled and nonlinear. Nash remarks that they are "difficult to solve explicitly" and refers to the book of Jaffe and Taubes. For $\lambda > 0$ no closed-form monopole is known; existence was proved by variational methods, and numerically the solutions look much like the $\lambda = 0$ solution we shall find in Section 8.13.

8.12.6 Finite energy forces a topological charge

Here is the step that produces the soliton. Finite energy on $\mathbb{R}^3$ requires each term of (8.272) to fall off at infinity. In particular $|\phi| \to C$ as $r \to \infty$ (when $\lambda > 0$ this is forced by the potential term; when $\lambda = 0$ we *impose* it, as Nash does in (8.152)). So at large r the Higgs field defines a map from the sphere at infinity to the vacuum manifold:

$$\hat\phi : S^2_\infty \to S^2 \subset \mathfrak{su}(2), \qquad \hat\phi(\hat{\vec{x}}) = \lim_{r\to\infty} \frac{\phi(r\hat{\vec{x}})}{|\phi(r\hat{\vec{x}})|}. \tag{8.275}$$

Maps between spheres of the same dimension have a **degree**: an integer counting how many times the image wraps round the target (Chapter 1). It cannot change under continuous deformations, and time evolution of finite-energy fields is continuous. Hence the degree is a conserved topological charge.

Working definition 8.39 (degree of a map $S^2 \to S^2$)

For a smooth map $\hat n : S^2 \to S^2 \subset \mathbb{R}^3$, $|\hat n| = 1$,

$$\deg \hat n = \frac{1}{4\pi}\int_{S^2} \hat n^*\omega = \frac{1}{8\pi}\int_{S^2} \varepsilon_{abc}\, n^a\, \mathrm{d}n^b \wedge \mathrm{d}n^c, \tag{8.276}$$

where ω is the area form of the unit sphere [the 2-form whose integral over any region is its area, 4π for all of S^2]. *How to compute:* either do the integral, or pick a generic point y [one at which the derivative of $\hat n$ is invertible at every preimage] of the target and count its preimages with signs (+1 where $\hat n$ preserves orientation, −1 where it reverses it).

The middle and right expressions agree because $\frac{1}{2}\varepsilon_{abc}n^a\mathrm{d}n^b \wedge \mathrm{d}n^c$ is the pullback of the area form: at the north pole $\vec n = (0,0,1)$ of the target it equals $\mathrm{d}n^1 \wedge \mathrm{d}n^2$, which is the area element there.

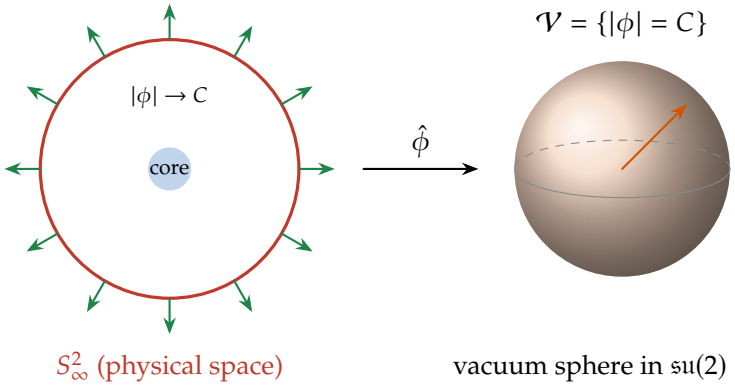


Figure 8.60: At large distance a finite-energy Higgs field lies on the vacuum sphere. Its direction defines a map $\hat{\phi}$ from the sphere at infinity (left) to the vacuum sphere (right). For the "hedgehog" drawn here the Higgs arrow points radially outward, so $\hat{\phi}$ is the identity map and has degree 1. A configuration of degree $k \neq 0$ cannot be deformed into the vacuum (degree 0) without passing through infinite energy.

Worked example 8.117: The degree of $\hat{\vec{n}}_m$ is m

Let $\vec{n}_m(\theta, \varphi) = (\sin\theta\cos m\varphi, \sin\theta\sin m\varphi, \cos\theta)$. It wraps the sphere m times round the vertical axis. The pullback of the area form $\omega = \sin\Theta\, \mathrm{d}\Theta \wedge \mathrm{d}\Phi$ under $\Theta = \theta$, $\Phi = m\varphi$ is

$$\frac{1}{4\pi}\int_{S^2} \vec{n}_m^{\,*}\omega \overset{(1)}{=} \frac{1}{4\pi}\int_{S^2} \sin\theta\, \mathrm{d}\theta \wedge (m\, \mathrm{d}\varphi) \overset{(2)}{=} \frac{m}{4\pi}\cdot 4\pi = m.$$

(1) $\Theta = \theta$ and $\mathrm{d}\Phi = m\, \mathrm{d}\varphi$.
(2) The remaining integral is the area of the unit sphere.

Preimage count: the point $\vec{n} = (1, 0, 0)$ has the m preimages $\theta = \pi/2$, $\varphi = 2\pi j/m$, each orientation-preserving. Both methods give m.

Key idea: The topological charge of a finite-energy configuration

Every finite-energy configuration on $\mathbb{R}^3$ carries an integer $k = \deg\hat{\phi} \in \pi_2(S^2) = \mathbb{Z}$. Configurations with different k lie in disconnected components of the space of finite-energy fields. The vacuum has $k = 0$; a configuration with $k \neq 0$ is a candidate monopole, and it cannot decay into the vacuum.

Two remarks from Nash's first paragraph fit here. First, if M is closed and compact there is no "sphere at infinity", and we shall see in Section 8.13.4 that monopoles on such M are trivial. Second, the charge lives entirely at infinity. The SU(2) bundle over $\mathbb{R}^3$ itself is trivial (Nash, p. 245), because $\mathbb{R}^3$ is contractible; the topology is in the Higgs field's asymptotic direction, not in the bundle.

8.12.7 The 't Hooft–Polyakov hedgehog

The simplest configuration of degree 1 points the Higgs field radially: $\hat{\phi}(\hat{\vec{x}}) = \hat{x}^a T_a$. For finite energy the gauge field must make $\mathrm{d}_A\phi$ vanish at infinity too, so that the gradient energy converges. This fixes the asymptotic gauge field. 't Hooft and Polyakov (1974) made the

spherically symmetric ansatz

$$\phi = h(r)\,\hat{x}^a T_a, \qquad A = \frac{1 - K(r)}{r}\,\varepsilon_{aij}\hat{x}^j\,T_a\,\mathrm{d}x^i, \tag{8.277}$$

with boundary conditions

$$h(0) = 0, \quad K(0) = 1 \qquad \text{(smooth at the origin)}, \qquad h(\infty) = C, \quad K(\infty) = 0 \qquad \text{(finite energy)}. \tag{8.278}$$

The name is visual: the Higgs arrow at $\vec{x}$ points along $\vec{x}$, like the spines of a hedgehog.

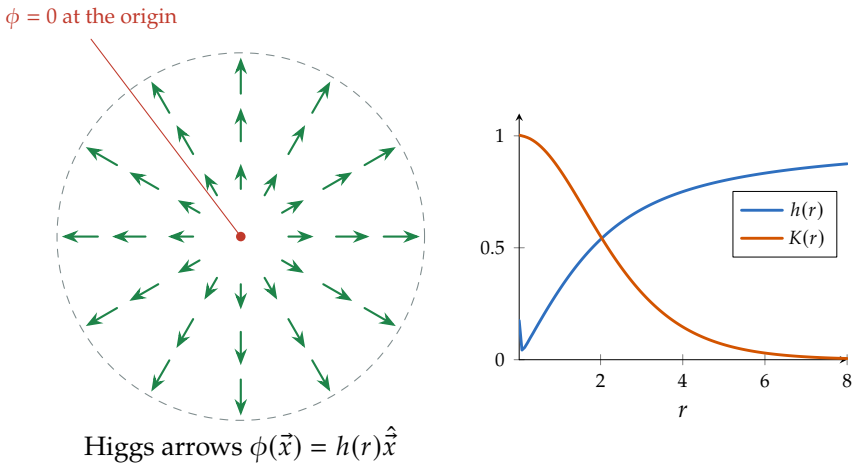


Figure 8.61: Left: the hedgehog. The Higgs field points radially in isospin space, its length grows from 0 at the centre to C far away, and it vanishes only at the origin, which is where the monopole "is". Right: the two profile functions for the Prasad–Sommerfield solution we shall derive in Example 8.123 (with $C = 1$). K controls the gauge field: $K = 1$ is "no gauge field" near the centre, $K = 0$ is the asymptotic monopole field.

Worked example 8.118: The hedgehog has charge one

At infinity $\hat{\phi}(\hat{\vec{x}}) = \hat{x}^a T_a$, whose component vector is $\vec{n} = \hat{\vec{x}}$. So $\hat{\phi}$ is the identity map of S^2, and every point of the target has exactly one orientation-preserving preimage: $\deg\hat{\phi} = 1$. With the integral (8.276), at $\hat{\vec{x}} = (0, 0, 1)$ the integrand $\frac{1}{2}\varepsilon_{abc}n^a\mathrm{d}n^b \wedge \mathrm{d}n^c = \mathrm{d}x^1 \wedge \mathrm{d}x^2$ is the area element, and integrating over the unit sphere gives $4\pi/4\pi = 1$.

To see why the gauge field in (8.277) is forced, compute its asymptotic form: with $K = 0$, $A_\infty = \frac{1}{r}\varepsilon_{aij}\hat{x}^j T_a \mathrm{d}x^i$. The covariant derivative of the unit Higgs field is then

$$\begin{aligned}(D_i\hat{\phi})^a &\overset{(1)}{=} \partial_i\hat{x}^a + \varepsilon_{abc}A_i^b\hat{x}^c \overset{(2)}{=} \frac{\delta_{ai} - \hat{x}_a\hat{x}_i}{r} + \frac{1}{r}\varepsilon_{abc}\varepsilon_{bij}\hat{x}^j\hat{x}^c \\ &\overset{(3)}{=} \frac{\delta_{ai} - \hat{x}_a\hat{x}_i}{r} + \frac{1}{r}\big(\delta_{ci}\delta_{aj} - \delta_{cj}\delta_{ai}\big)\hat{x}^j\hat{x}^c \overset{(4)}{=} \frac{\delta_{ai} - \hat{x}_a\hat{x}_i}{r} + \frac{\hat{x}_a\hat{x}_i - \delta_{ai}}{r} = 0.\end{aligned}$$

(1) Definition (8.267).
(2) $\partial_i(x^a/r) = (\delta_{ai} - \hat{x}_a\hat{x}_i)/r$; insert $A_i^b = \frac{1}{r}\varepsilon_{bij}\hat{x}^j$.
(3) $\varepsilon_{abc}\varepsilon_{bij} = \varepsilon_{bca}\varepsilon_{bij} = \delta_{ci}\delta_{aj} - \delta_{cj}\delta_{ai}$.
(4) Contract, using $\hat{x}^j\hat{x}^j = 1$.

So the asymptotic gauge field is exactly the one that makes the rotating Higgs direction covariantly constant. Its field strength does *not* vanish: it is the magnetic monopole field, as we shall see in Section 8.13.7.

Substituting the ansatz (8.277) into the energy reduces a problem about fields on $\mathbb{R}^3$ to a problem about two functions of r. The covariant derivatives decompose into a "transverse" part proportional to $P^T_{ai} = \delta_{ai} - \hat{x}_a\hat{x}_i$ and a "longitudinal" part proportional to $P^L_{ai} = \hat{x}_a\hat{x}_i$ (a direct computation like the one in Example 8.118, which we checked by computer algebra):

$$B^a_i = \frac{K'}{r}\,P^T_{ai} - \frac{1-K^2}{r^2}\,P^L_{ai}, \qquad D_i\phi^a = \frac{hK}{r}\,P^T_{ai} + h'\,P^L_{ai}. \tag{8.279}$$

Since P^T has trace 2, P^L has trace 1, and $P^TP^L = 0$, the squares are $|B|^2 = 2K'^2/r^2 + (1-K^2)^2/r^4$ and $|D\phi|^2 = 2h^2K^2/r^2 + h'^2$. With $\mathrm{d}^3x = 4\pi r^2\mathrm{d}r$ after the angular integral,

$$E = 4\pi\int_0^\infty \left\{K'^2 + \frac{(1-K^2)^2}{2r^2} + h^2K^2 + \frac{r^2h'^2}{2} + \frac{\lambda}{2}r^2\big(h^2-C^2\big)^2\right\}\mathrm{d}r. \tag{8.280}$$

Every term is non-negative, and the boundary conditions (8.278) make the integral converge at both ends (near $r=0$, $1-K^2 = O(r^2)$ and near $r=\infty$, $K\to 0$ exponentially). Minimising (8.280) over h, K is a two-function ordinary variational problem; for $\lambda > 0$ it is done numerically, and for $\lambda = 0$ we solve it in closed form in the next section.

Try it: The unit of energy

Show that for *any* functions obeying (8.278) the energy (8.280) with $\lambda = 0$ is at least $4\pi C$. *Hint:* complete two squares, $K'^2 + h^2K^2 = (K' + hK)^2 - 2hKK'$ and $\frac{(1-K^2)^2}{2r^2} + \frac{r^2h'^2}{2} = \frac{1}{2}\big(rh' - \frac{1-K^2}{r}\big)^2 + h'(1-K^2)$, and notice that the cross terms add up to $\frac{\mathrm{d}}{\mathrm{d}r}\big[h(1-K^2)\big]$. *Answer:* the total derivative integrates to $h(1-K^2)\big|_0^\infty = C$, so $E = 4\pi C + (\text{squares}) \geq 4\pi C$. We redo this properly in Example 8.123.

Reading Nash (pp. 242–243, (8.139)–(8.140))

Nash's Section 4 opens by listing what changes in three dimensions: monopoles are static, finite-energy critical points of a Yang–Mills–Higgs energy on a three-dimensional M; they carry an integer invariant k; they are trivial on closed compact M; and the usual choice is $M = \mathbb{R}^3$, whose non-compactness brings technical difficulties that "have not proved insurmountable". (8.139) is our (8.271) (with his $-$ tr replaced by -2 tr) and (8.140) is our (8.274). The topology at infinity, which Nash only introduces on p. 245, has been moved forward here because it is what makes k exist at all.

Recap

- In three dimensions pure Yang–Mills has no static lumps (Derrick); a Higgs field in the adjoint representation provides the balancing force.
- The YMH energy is magnetic + gradient + Mexican-hat potential; its vacuum manifold is $S^2 = \mathrm{SU}(2)/\mathrm{U}(1)$, with unbroken U(1) = photon.
- Finite energy forces $|\phi| \to C$ at infinity, so $\hat{\phi} : S^2_\infty \to S^2$ has a degree $k \in \pi_2(S^2) = \mathbb{Z}$: the conserved topological charge.
- The Dirac monopole shows why a magnetic charge is a Chern number; the 't Hooft–Polyakov hedgehog is a smooth, finite-energy version with $k = 1$.

8.13 The Bogomolny equation and the Prasad–Sommerfield monopole

The field equations (8.274) are second order and coupled. For instantons we met a trick that turns such a problem into a first-order one: complete the square in the action, notice that the cross term is topological, and conclude that the absolute minima in each topological sector solve a first-order equation. This section plays the same trick for monopoles. The questions are:

> *When does the monopole problem become first order? What is the first-order equation, what is the minimum energy, and can we solve it explicitly?*

The answers: when the Higgs potential is switched off ($\lambda = 0$); the equation is Bogomolny's $F = \mp \star \mathrm{d}_A\phi$; the minimum energy is $4\pi C|k|$; and for $k = 1$ the solution is the Prasad–Sommerfield monopole, which we write down in closed form.

8.13.1 The Prasad–Sommerfield limit

Nash's next step is to take the limit in which the potential term of (8.271) vanishes, $\lambda \to 0$. The energy becomes

$$E(A,\phi) = -\int_M \Big\{ \mathrm{tr}(F \wedge \star F) + \mathrm{tr}(\mathrm{d}_A\phi \wedge \star \mathrm{d}_A\phi) \Big\} = \frac{1}{2}\Big\{ \|F\|^2 + \big\|\mathrm{d}_A\phi\big\|^2 \Big\}, \tag{8.281}$$

with the norms of (8.266). Setting $\lambda = 0$ in (8.274) gives the field equations

$$\mathrm{d}_A^\star F = [\mathrm{d}_A\phi, \phi], \qquad \mathrm{d}_A^\star \mathrm{d}_A\phi = 0. \tag{8.282}$$

Physical picture: What survives of the potential

Setting $\lambda = 0$ does not mean "no symmetry breaking". We keep the *boundary condition* $|\phi| \to C$ at infinity, which is what the potential was there to enforce. Think of $\lambda \to 0^+$: the potential is infinitely weak, so it no longer shapes the core of the monopole, but it still selects the vacuum sphere at infinity. The price is that the Higgs field is now massless ($m_H^2 \propto \lambda C^2 \to 0$), and its long-range tail will matter when we discuss forces in Section 8.16.

8.13.2 The Bogomolny equation solves the field equations

Bogomolny observed that the first-order equation

$$F = \pm \star \mathrm{d}_A\phi \tag{8.283}$$

implies both equations (8.282). Nash writes the + sign in (8.143); the self-duality reduction below will produce the − sign, and so will the hedgehog. Both signs work, so we check them together. Two facts are needed.

Working definition 8.40 (two identities for d_A in the adjoint)

For an $\mathfrak{su}(2)$-valued function ϕ and an $\mathfrak{su}(2)$-valued 1-form ω,

$$\mathrm{d}_A\omega = \mathrm{d}\omega + A \wedge \omega + \omega \wedge A, \qquad \mathrm{d}_A\mathrm{d}_A\phi = [F,\phi], \qquad \mathrm{d}_A F = 0.$$

The second says "d_A^2 is curvature"; the third is the Bianchi identity [it holds for every connection, not only for solutions] (Chapter T3). *Operational meaning:* a covariant derivative applied twice does not vanish, it multiplies by F (in the adjoint [on $\mathfrak{su}(2)$-valued fields

such as ϕ, where A acts as $[A, \cdot\,]$, by commutator).

Worked example 8.119: Nash's (8.145): $\mathrm{d}_A \mathrm{d}_A \phi = [F, \phi]$ written out

Put $\omega = \mathrm{d}_A\phi = \mathrm{d}\phi + A\phi - \phi A$. Then

$$\begin{aligned}
\mathrm{d}_A\omega &\overset{(1)}{=} \mathrm{d}(\mathrm{d}\phi + A\phi - \phi A) + A \wedge (\mathrm{d}\phi + A\phi - \phi A) + (\mathrm{d}\phi + A\phi - \phi A) \wedge A \\
&\overset{(2)}{=} \big(\mathrm{d}A\,\phi - A \wedge \mathrm{d}\phi - \mathrm{d}\phi \wedge A - \phi\,\mathrm{d}A\big) + \big(A \wedge \mathrm{d}\phi + A \wedge A\,\phi - A\phi \wedge A\big) \\
&\qquad + \big(\mathrm{d}\phi \wedge A + A\phi \wedge A - \phi A \wedge A\big) \\
&\overset{(3)}{=} (\mathrm{d}A + A \wedge A)\phi - \phi(\mathrm{d}A + A \wedge A) \\
&\overset{(4)}{=} F\phi - \phi F = [F, \phi].
\end{aligned}$$

(1) Definition of d_A on an adjoint 1-form.
(2) $\mathrm{d}\mathrm{d}\phi = 0$; the Leibniz rule gives $\mathrm{d}(A\phi) = \mathrm{d}A\,\phi - A \wedge \mathrm{d}\phi$ and $\mathrm{d}(\phi A) = \mathrm{d}\phi \wedge A + \phi\,\mathrm{d}A$ (a 1-form passing d picks up a minus sign).
(3) The terms $\pm A \wedge \mathrm{d}\phi$, $\pm \mathrm{d}\phi \wedge A$ and $\pm A\phi \wedge A$ cancel in pairs.
(4) $F = \mathrm{d}A + A \wedge A$.

Now take $F = \pm \star \mathrm{d}_A\phi$, equivalently $\mathrm{d}_A\phi = \pm \star F$ (because $\star\star = 1$ in three Riemannian dimensions). Recall $\mathrm{d}_A^\star = - \star \mathrm{d}_A \star$ on 1-forms and $\mathrm{d}_A^\star = \star \mathrm{d}_A \star$ on 2-forms. The second field equation:

$$\mathrm{d}_A^\star \mathrm{d}_A \phi \overset{(1)}{=} - \star \mathrm{d}_A \star \mathrm{d}_A \phi \overset{(2)}{=} - \star \mathrm{d}_A \star (\pm \star F) \overset{(3)}{=} \mp \star \mathrm{d}_A F \overset{(4)}{=} 0.$$

(1) Adjoint of d_A on 1-forms in three dimensions.
(2) Bogomolny, in the form $\mathrm{d}_A\phi = \pm \star F$.
(3) $\star\star = 1$.
(4) Bianchi identity. This is Nash's remark that "the second equation just becomes the Bianchi identity" (8.144).

The first field equation:

$$\mathrm{d}_A^\star F \overset{(1)}{=} \star \mathrm{d}_A \star F \overset{(2)}{=} \star \mathrm{d}_A(\pm \mathrm{d}_A\phi) \overset{(3)}{=} \pm \star [F, \phi] \overset{(4)}{=} \big[\pm \star F, \phi\big] \overset{(5)}{=} [\mathrm{d}_A\phi, \phi].$$

(1) Adjoint of d_A on 2-forms in three dimensions.
(2) $\star F = \pm \star \star \mathrm{d}_A\phi = \pm \mathrm{d}_A\phi$.
(3) Example 8.119.
(4) $\star$ acts on the form part only and commutes with bracketing by the function ϕ.
(5) Bogomolny again.

So every solution of (8.283) solves (8.282). The converse is false in general, just as not every Yang–Mills field is self-dual; we return to the non-Bogomolny critical points in Section 8.18.

8.13.3 Completing the square: the Bogomolny bound

Nash notes that it is "much easier" to see why (8.283) is the right equation by completing the square in the energy. Since $\star$ is an isometry, $\big\|\star \mathrm{d}_A\phi\big\| = \big\|\mathrm{d}_A\phi\big\|$, and

$$E \overset{(1)}{=} \frac{1}{2}\Big\{\|F\|^2 + \big\|\star \mathrm{d}_A\phi\big\|^2\Big\}$$

$$\overset{(2)}{=} \frac{1}{2}\Big\{\big\|F \mp \star \mathrm{d}_A\phi\big\|^2 \pm 2\langle F, \star \mathrm{d}_A\phi\rangle\Big\}$$
$$\overset{(3)}{=} \frac{1}{2}\big\|F \mp \star \mathrm{d}_A\phi\big\|^2 \pm \langle F, \star \mathrm{d}_A\phi\rangle. \tag{8.284}$$

(1) Replace $\|\mathrm{d}_A\phi\|$ by $\|\star \mathrm{d}_A\phi\|$.
(2) Expand: $\|F \mp \star \mathrm{d}_A\phi\|^2 = \|F\|^2 + \|\star \mathrm{d}_A\phi\|^2 \mp 2\langle F, \star \mathrm{d}_A\phi\rangle$, so we add back the cross term. This is Nash's (8.146).
(3) Distribute the $\frac{1}{2}$.

The first term is non-negative and vanishes exactly when $F = \pm \star \mathrm{d}_A\phi$. The second term is the one Nash says "looks like a topological charge":

$$\langle F, \star \mathrm{d}_A\phi\rangle = -2\int_M \mathrm{tr}\big(F \wedge \star \star \mathrm{d}_A\phi\big) = -2\int_M \mathrm{tr}(F \wedge \mathrm{d}_A\phi). \tag{8.285}$$

It is indeed a total derivative:

$$\begin{aligned}
\mathrm{d}\ \mathrm{tr}(F\phi) &\overset{(1)}{=} \mathrm{tr}(\mathrm{d}F\,\phi) + \mathrm{tr}(F \wedge \mathrm{d}\phi) \\
&\overset{(2)}{=} \mathrm{tr}\big((\mathrm{d}_A F - A\wedge F + F\wedge A)\phi\big) + \mathrm{tr}\big(F \wedge (\mathrm{d}_A\phi - A\phi + \phi A)\big) \\
&\overset{(3)}{=} \mathrm{tr}(\mathrm{d}_A F\,\phi) + \mathrm{tr}(F \wedge \mathrm{d}_A\phi) + \mathrm{tr}\big(-A\wedge F\phi + F\phi \wedge A\big) \\
&\overset{(4)}{=} \mathrm{tr}(\mathrm{d}_A F\,\phi) + \mathrm{tr}(F \wedge \mathrm{d}_A\phi) \overset{(5)}{=} \mathrm{tr}(F \wedge \mathrm{d}_A\phi).
\end{aligned} \tag{8.286}$$

(1) Leibniz rule; F is a 2-form, so no sign.
(2) Solve $\mathrm{d}_A F = \mathrm{d}F + A\wedge F - F\wedge A$ and $\mathrm{d}_A\phi = \mathrm{d}\phi + A\phi - \phi A$ for $\mathrm{d}F$ and $\mathrm{d}\phi$.
(3) Multiply out; $-\mathrm{tr}(F\wedge A\phi)$ and $+\mathrm{tr}(F\wedge A\phi)$ cancel, leaving the two terms shown.
(4) Graded cyclicity $\mathrm{tr}(\alpha\wedge\beta) = (-1)^{pq}\,\mathrm{tr}(\beta\wedge\alpha)$ with $p=1, q=2$ gives $\mathrm{tr}(A\wedge F\phi) = \mathrm{tr}(F\phi\wedge A)$.
(5) Bianchi identity $\mathrm{d}_A F = 0$. This is Nash's (8.149).

Key idea: The Bogomolny bound

In the Prasad–Sommerfield limit,

$$E(A,\phi) = \tfrac{1}{2}\big\|F \mp \star \mathrm{d}_A\phi\big\|^2 \pm \langle F, \star \mathrm{d}_A\phi\rangle \ \geq\ \pm\langle F, \star \mathrm{d}_A\phi\rangle,$$

and the cross term depends only on boundary values, by (8.286) and Stokes. For fixed boundary data the energy is therefore smallest exactly on solutions of the **Bogomolny equation** $F = \pm \star \mathrm{d}_A\phi$ (Nash's (8.147)). Choose the sign that makes the cross term positive.

This is word for word the argument that produced the instanton bound $S \geq 8\pi^2|k|$ in the previous section: square plus topological term.

8.13.4 On a closed manifold every monopole is trivial

Nash now raises a warning. Suppose M is compact and has no boundary. Then Stokes kills the cross term:

$$\langle F, \star \mathrm{d}_A\phi\rangle \overset{(1)}{=} -2\int_M \mathrm{tr}(F\wedge \mathrm{d}_A\phi) \overset{(2)}{=} -2\int_M \mathrm{d}\ \mathrm{tr}(F\phi) \overset{(3)}{=} -2\int_{\partial M} \mathrm{tr}(F\phi) \overset{(4)}{=} 0. \tag{8.287}$$

(1) (8.285).
(2) (8.286).

(3) Stokes' theorem for the gauge-invariant 2-form $\mathrm{tr}(F\phi)$.
(4) $\partial M = \varnothing$ (Nash's (8.150)).

Hence on a Bogomolny solution $E = 0$, and since E is a sum of squares,

$$F = 0, \qquad \mathrm{d}_A\phi = 0. \tag{8.288}$$

This is Nash's (8.151). The connection is **flat** and the Higgs field is covariantly constant. A flat connection has no local content: near any point it is gauge equivalent to $A = 0$. What remains is global: the holonomy around non-contractible loops (Chapter T3). If M is simply connected there are no such loops, and the "monopole" is gauge equivalent to $A = 0$, $\phi =$ constant: the vacuum.

Worked example 8.120: A flat monopole on the three-torus

Let $M = T^3 = \mathbb{R}^3/(2\pi\mathbb{Z})^3$ with coordinates $x^i \sim x^i + 2\pi$. Take the constant connection and constant Higgs field

$$A = \theta_i\, T_3\, \mathrm{d}x^i, \qquad \phi = C\, T_3, \qquad \theta_i \in \mathbb{R}.$$

Then

$$F \overset{(1)}{=} \mathrm{d}A + A \wedge A \overset{(2)}{=} 0 + \theta_i\theta_j\, T_3T_3\, \mathrm{d}x^i \wedge \mathrm{d}x^j \overset{(3)}{=} 0,$$
$$\mathrm{d}_A\phi \overset{(4)}{=} \mathrm{d}\phi + [A, \phi] \overset{(5)}{=} 0 + C\theta_i[T_3, T_3]\mathrm{d}x^i = 0.$$

(1) Definition of curvature.
(2) The coefficients are constant, so $\mathrm{d}A = 0$.
(3) $\theta_i\theta_j$ is symmetric and $\mathrm{d}x^i \wedge \mathrm{d}x^j$ antisymmetric.
(4) Definition of d_A on the adjoint.
(5) ϕ is constant and T_3 commutes with itself.

So (A, ϕ) solves both Bogomolny equations, with $E = 0$. It is nevertheless *not* gauge equivalent to $A = 0$ when some $\theta_i \notin 4\pi\mathbb{Z}$: the holonomy round the i-th circle is

$$\mathrm{Hol}_i = \exp\Big(-\oint A\Big) = e^{-2\pi\theta_i T_3} = \begin{pmatrix} e^{i\pi\theta_i} & 0 \\ 0 & e^{-i\pi\theta_i} \end{pmatrix},$$

and conjugation cannot change its eigenvalues. This is the only way a closed M can support a nontrivial ground state: through its fundamental group, here $\pi_1(T^3) = \mathbb{Z}^3$.

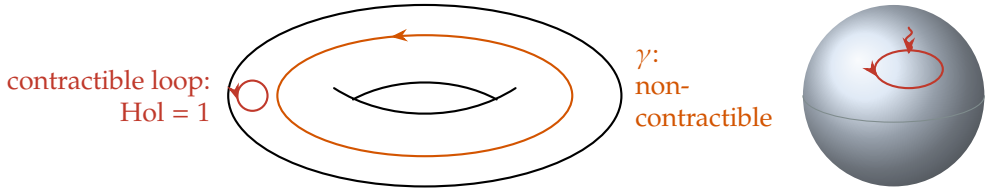


Figure 8.62: Why Bogomolny solutions on a closed M are dull. They are flat, so all their content is holonomy around loops. On the torus (left) a non-contractible loop can carry a holonomy $e^{-2\pi\theta T_3}$ that no gauge transformation removes. On a simply connected space such as S^3 (right, drawn as a ball) every loop can be shrunk to a point, the holonomy is trivial, and the configuration is gauge equivalent to the vacuum.

Key idea: Where non-trivial monopoles can live

Non-trivial monopoles need either a non-compact M or a boundary, so that the cross term $-2\int_{\partial M}\operatorname{tr}(F\phi)$ can be non-zero. Nash considers two cases: $M=\mathbb{R}^3$ with its flat metric (the rest of this section), and $M=H^3$, three-dimensional hyperbolic space (Section 8.14). In both the "boundary" is a sphere at infinity.

8.13.5 Monopoles on $\mathbb{R}^3$: boundary conditions and the magnetic charge

Take $M=\mathbb{R}^3$ with the Euclidean metric and $G=\mathrm{SU}(2)$. To make the energy integrals converge and the variational problem well posed we must specify how the fields behave as $r=|\vec{x}|\to\infty$. Nash's condition (8.152) is

$$|\phi|\to C\quad(r\to\infty),\tag{8.289}$$

together with enough decay of $\mathrm{d}_A\phi$ and F for the energy to be finite; on solutions of the Bogomolny equation the decay of F follows from that of $\mathrm{d}_A\phi$.

Watch out: The rate in (8.152)

Nash writes $|\phi|=C+O(r^{-2})$. That rate is right for a massive Higgs field ($\lambda>0$), whose deviation from C decays exponentially. It is *wrong* for the Bogomolny monopoles that are the subject of the section: the explicit solution below has $|\phi|=C-1/r+$ (exponentially small), and a charge-k monopole has $|\phi|=C-k/r+O(r^{-2})$. The $1/r$ tail is the Coulomb field of the massless Higgs, and it is *forced*, as we shall see: $\mathrm{d}_A\phi=\mp\star F$ ties the Higgs gradient to the magnetic Coulomb field $\sim k/r^2$. The correct condition is simply $|\phi|\to C$ with $|\mathrm{d}_A\phi|=O(r^{-2})$.

With these conditions the cross term becomes an integral over the sphere at infinity. Let S^2_R be the sphere of radius R, oriented by the outward normal. Then

$$-\langle F,\star\mathrm{d}_A\phi\rangle\overset{(1)}{=}2\lim_{R\to\infty}\int_{|x|\le R}\mathrm{d}\,\operatorname{tr}(F\phi)\overset{(2)}{=}2\lim_{R\to\infty}\oint_{S^2_R}\operatorname{tr}(F\phi)\overset{(3)}{=}-\oint_{S^2_\infty}(F,\phi),\tag{8.290}$$

(1) (8.285) and (8.286).
(2) Stokes on the ball.
(3) Pointwise inner product (8.265): $(F,\phi)=-2\operatorname{tr}(F\phi)=F^a\phi^a$, a real 2-form.

Define the **magnetic charge**

$$k=-\frac{1}{4\pi C}\langle F,\star\mathrm{d}_A\phi\rangle=-\frac{1}{4\pi C}\oint_{S^2_\infty}F^a\phi^a=\frac{1}{2\pi C}\oint_{S^2_\infty}\operatorname{tr}(F\phi).\tag{8.291}$$

(Nash's (8.153)–(8.154) have the opposite overall sign and his $-\operatorname{tr}$ normalisation; with our choices the hedgehog of Section 8.12.7 has $k=+1$ and satisfies $F=-\star\,\mathrm{d}_A\phi$, the sign that the self-duality reduction will produce.) In terms of k the Bogomolny bound reads

$$E=\tfrac12\left\|F+\star\mathrm{d}_A\phi\right\|^2+4\pi C\,k\;\ge\;4\pi C\,k,\qquad E=\tfrac12\left\|F-\star\mathrm{d}_A\phi\right\|^2-4\pi C\,k\;\ge\;-4\pi C\,k,\tag{8.292}$$

so $E\ge 4\pi C|k|$, with equality for $k>0$ iff $F=-\star\,\mathrm{d}_A\phi$ (monopoles) and for $k<0$ iff $F=+\star\,\mathrm{d}_A\phi$ (anti-monopoles).

Why is k an integer? Nash gives two interpretations; we make both concrete.

First reading: a Chern number of a $\mathrm{U}(1)$ bundle at infinity

At large r the Higgs field has length C and is covariantly constant to leading order. Its direction $\hat\phi = \phi/|\phi|$ picks out, at each point of S^2_∞, a line in $\mathfrak{su}(2)$ and hence an "unbroken" U(1): the rotations about $\hat\phi$. The component of F along $\hat\phi$,

$$f = (F, \hat\phi) = F^a \hat\phi^a,$$

is the field strength of electromagnetism, and (8.291) says $k = -\frac{1}{4\pi}\oint f$, i.e. k is (minus) the magnetic flux through the sphere at infinity in units of 4π. That this is an integer is the Dirac quantisation of Example 8.114: the U(1) that acts on the two-component (fundamental) fields with charges $\pm\frac12$ is a genuine circle bundle over S^2_∞, and its first Chern class is k. The next example makes this explicit by computing the gauge field at infinity.

Worked example 8.121: The gauge field at infinity is forced, and its curvature is the pulled-back area form

Write $\hat\phi = n^a T_a$ with $|\vec n| = 1$, and suppose $\mathrm{d}_A\hat\phi = 0$ at infinity. Claim: then

$$A = [\mathrm{d}\hat\phi, \hat\phi] + \alpha\,\hat\phi \tag{8.293}$$

for some real 1-form α, and the curvature is

$$F = \big(\mathrm{d}\alpha - \tfrac12 \varepsilon_{abc} n^a \mathrm{d}n^b \wedge \mathrm{d}n^c\big)\hat\phi. \tag{8.294}$$

First, (8.293) does make $\hat\phi$ covariantly constant. In vector notation ($[X,Y] \leftrightarrow \vec X \times \vec Y$),

$$\begin{aligned}\mathrm{d}_A\hat\phi &\overset{(1)}{=} \mathrm{d}\vec n + (\mathrm{d}\vec n \times \vec n)\times\vec n + \alpha\,\vec n\times\vec n \\ &\overset{(2)}{=} \mathrm{d}\vec n - \big(\mathrm{d}\vec n\,(\vec n\cdot\vec n) - \vec n\,(\vec n\cdot\mathrm{d}\vec n)\big) \overset{(3)}{=} \mathrm{d}\vec n - \mathrm{d}\vec n = 0.\end{aligned}$$

(1) $\mathrm{d}_A\hat\phi = \mathrm{d}\hat\phi + [A,\hat\phi]$ with (8.293).
(2) $\vec n\times\vec n = 0$ and $(\vec u\times\vec n)\times\vec n = -\vec n\times(\vec u\times\vec n)$, then "BAC–CAB".
(3) $\vec n\cdot\vec n = 1$, hence $\vec n\cdot\mathrm{d}\vec n = 0$.

Conversely, any A with $\mathrm{d}_A\hat\phi = 0$ differs from $[\mathrm{d}\hat\phi,\hat\phi]$ by a 1-form commuting with $\hat\phi$, i.e. a multiple of $\hat\phi$. So the gauge field at infinity is fixed by the Higgs direction, up to an abelian piece $\alpha\hat\phi$.

For the curvature, check (8.294) at one point, say where $\vec n = (0,0,1)$, using local coordinates (u,v) on S^2_∞ with $\vec n = (u, v, 1 - \frac12(u^2+v^2)) + \dots$ By rotational covariance this suffices. Put $\alpha = 0$ first. Then $A^a = \varepsilon_{abc}\mathrm{d}n^b\, n^c$ and to first order

$$A^1 = \mathrm{d}v + O(2), \qquad A^2 = -\mathrm{d}u + O(2), \qquad A^3 = v\,\mathrm{d}u - u\,\mathrm{d}v.$$

At the point $u = v = 0$:

$$\begin{aligned}F^3 &\overset{(1)}{=} \mathrm{d}A^3 + A^1\wedge A^2 \overset{(2)}{=} (\mathrm{d}v\wedge\mathrm{d}u - \mathrm{d}u\wedge\mathrm{d}v) + \mathrm{d}v\wedge(-\mathrm{d}u) \overset{(3)}{=} -\mathrm{d}u\wedge\mathrm{d}v, \\ F^1 &\overset{(4)}{=} \mathrm{d}A^1 + A^2\wedge A^3 = 0, \qquad F^2 = \mathrm{d}A^2 + A^3\wedge A^1 = 0.\end{aligned}$$

(1) $F^a = \mathrm{d}A^a + \frac12\varepsilon_{abc}A^b\wedge A^c$, the component form of $F = \mathrm{d}A + A\wedge A$ (Example 8.112).
(2) Insert the first-order expressions.
(3) Collect: $-2\,\mathrm{d}u\wedge\mathrm{d}v + \mathrm{d}u\wedge\mathrm{d}v$.
(4) $\mathrm{d}A^1, \mathrm{d}A^2$ vanish at the origin (their coefficients are $O(2)$), and A^3 vanishes there.

At this point $\frac12\varepsilon_{abc}n^a\mathrm{d}n^b\wedge\mathrm{d}n^c = \mathrm{d}n^1\wedge\mathrm{d}n^2 = \mathrm{d}u\wedge\mathrm{d}v$, so $F = -\mathrm{d}u\wedge\mathrm{d}v\,T_3$, which is (8.294). Adding $\alpha\hat\phi$ adds $\mathrm{d}\alpha\,\hat\phi$ (it commutes with everything else at leading order).

Now integrate. The abelian field is $f = (F, \hat{\phi}) = \mathrm{d}\alpha - \frac{1}{2}\varepsilon_{abc} n^a \mathrm{d}n^b \wedge \mathrm{d}n^c$, and

$$k \overset{(1)}{=} -\frac{1}{4\pi}\oint_{S^2_\infty} f \overset{(2)}{=} -\frac{1}{4\pi}\oint_{S^2_\infty} \mathrm{d}\alpha + \frac{1}{8\pi}\oint_{S^2_\infty} \varepsilon_{abc} n^a \mathrm{d}n^b \wedge \mathrm{d}n^c \overset{(3)}{=} \deg \hat{\phi}. \tag{8.295}$$

(1) (8.291) with $\phi = C\hat{\phi}$ at infinity.
(2) (8.294).
(3) $\mathrm{d}\alpha$ integrates to zero over a closed surface (if α is globally defined; if not, it contributes an integer which can be absorbed by a gauge transformation, as in the Dirac construction), and the last term is the degree (8.276).

Second reading: the degree of the Higgs map

This is Nash's (8.155): k is the winding number of $\hat{\phi} : S^2_\infty \to S^2_{\mathfrak{su}(2)}$. He writes it as a matrix integral. We check that his formula agrees with (8.276).

Worked example 8.122: Nash's winding formula in components

With $\hat{\phi} = n^a T_a$,

$$\begin{aligned} -\frac{1}{2\pi}\oint \operatorname{tr}\left(\hat{\phi}\,\mathrm{d}\hat{\phi}\wedge\mathrm{d}\hat{\phi}\right) &\overset{(1)}{=} -\frac{1}{2\pi}\oint n^a\,\mathrm{d}n^b\wedge\mathrm{d}n^c\,\operatorname{tr}(T_aT_bT_c) \\ &\overset{(2)}{=} -\frac{1}{2\pi}\oint n^a\,\mathrm{d}n^b\wedge\mathrm{d}n^c\,\tfrac{1}{2}\varepsilon_{bcd}\operatorname{tr}(T_aT_d) \\ &\overset{(3)}{=} -\frac{1}{2\pi}\oint n^a\,\mathrm{d}n^b\wedge\mathrm{d}n^c\,\tfrac{1}{2}\varepsilon_{bcd}\left(-\tfrac{1}{2}\delta_{ad}\right) \\ &\overset{(4)}{=} \frac{1}{8\pi}\oint \varepsilon_{abc}\,n^a\mathrm{d}n^b\wedge\mathrm{d}n^c \overset{(5)}{=} \deg\hat{\phi}. \end{aligned}$$

(1) Expand in the basis.
(2) $\mathrm{d}n^b \wedge \mathrm{d}n^c$ is antisymmetric in b, c, so only $\frac{1}{2}[T_b, T_c] = \frac{1}{2}\varepsilon_{bcd}T_d$ survives from T_bT_c.
(3) $\operatorname{tr}(T_aT_d) = -\frac{1}{2}\delta_{ad}$ (Example 8.112).
(4) Collect the constants: $(-\frac{1}{2\pi})(\frac{1}{2})(-\frac{1}{2}) = \frac{1}{8\pi}$.
(5) Definition (8.276).

So (8.155) holds exactly in our normalisation. For the hedgehog $\vec{n} = \hat{\vec{x}}$ it gives 1.

Key idea: The magnetic charge is the only topological invariant

For finite-energy SU(2) Yang–Mills–Higgs fields on $\mathbb{R}^3$,

$$k = \text{magnetic flux}/(-4\pi) = c_1(\text{unbroken U(1) bundle over } S^2_\infty) = \deg\left(\hat{\phi} : S^2_\infty \to S^2\right) \in \mathbb{Z}.$$

The SU(2) bundle over $\mathbb{R}^3$ is trivial ($\mathbb{R}^3$ is contractible), so all the topology is in the boundary behaviour of the Higgs field.

Dictionary: monopole charge

Mathematics	Physics
$\deg \hat{\phi} \in \pi_2(S^2)$	winding of the Higgs direction at infinity
c_1 of $L \to S^2_\infty$	quantised magnetic flux (Dirac)
$\langle F, \star \mathrm{d}_A \phi \rangle$	$\int \vec{B}^a \cdot \vec{D}\phi^a \, \mathrm{d}^3x$
Bogomolny bound $E \geq 4\pi C\lvert k\rvert$	BPS mass formula $M \geq 4\pi v\lvert k\rvert/e$

8.13.6 Solving the Bogomolny equation: the Prasad–Sommerfield monopole

Nash does not write down a solution. Here is the classic one, which we derive from scratch. Put the hedgehog ansatz (8.277) into $F = -\star\, \mathrm{d}_A\phi$, i.e. $B_i^a = -D_i\phi^a$, and use the decomposition (8.279) into transverse (P^T) and longitudinal (P^L) parts. Matching coefficients gives two ordinary differential equations.

Worked example 8.123: Deriving and solving the Bogomolny ODEs

From (8.279),

$$B_i^a + D_i\phi^a \overset{(1)}{=} \left(\frac{K'}{r} + \frac{hK}{r}\right)P^T_{ai} + \left(-\frac{1-K^2}{r^2} + h'\right)P^L_{ai} \overset{(2)}{=} 0$$

$$\Longleftrightarrow \qquad \boxed{K' = -hK, \qquad h' = \frac{1-K^2}{r^2}.}$$

(1) Add the two lines of (8.279).

(2) P^T and P^L are linearly independent (they project onto orthogonal subspaces), so each coefficient must vanish.

To solve, write $K = r/g(r)$ and eliminate h. From the first equation, $h = -K'/K = -(\ln K)' = g'/g - 1/r$. Substitute into the second:

$$h' \overset{(1)}{=} \frac{g''g - g'^2}{g^2} + \frac{1}{r^2} \overset{(2)}{=} \frac{1 - r^2/g^2}{r^2} \qquad \Longleftrightarrow \qquad g\,g'' - g'^2 = -1.$$

(1) Differentiate $h = g'/g - 1/r$.

(2) The second Bogomolny ODE with $K^2 = r^2/g^2$; then multiply by g^2 and cancel the $1/r^2$ terms.

The equation $gg'' - g'^2 = -1$ has the solutions $g = r$ (giving $K = 1$, $h = 0$: no monopole), $g = \sin(Cr)/C$ (giving a K that blows up at $r = \pi/C$), and

$$g = \frac{\sinh Cr}{C}: \qquad gg'' - g'^2 = \sinh^2 Cr - \cosh^2 Cr = -1 \;\checkmark. \tag{8.296}$$

Only the last satisfies $K(0) = 1$ and $K(\infty) = 0$. The result is the **Prasad–Sommerfield monopole**:

$$\boxed{h(r) = C\coth(Cr) - \frac{1}{r}, \qquad K(r) = \frac{Cr}{\sinh(Cr)}} \tag{8.297}$$

Check the boundary conditions: $h(r) = \frac{C^2 r}{3} + O(r^3) \to 0$ and $K \to 1$ as $r \to 0$; $h \to C - \frac{1}{r}$ and $K \approx 2Cr\, e^{-Cr} \to 0$ as $r \to \infty$. The profiles are plotted in Figure 8.61.

The asymptotics already tell us a lot. The gauge fields $W^\pm$ (the part of A proportional to K)

die off like e^{-Cr}: they have mass C, as in Example 8.115. The Higgs field approaches its vacuum value like $C - 1/r$: it is massless, with a Coulomb tail. And the magnetic field is $B_i^a \to -\hat{x}_a \hat{x}_i / r^2$: a Coulomb magnetic field along the Higgs direction.

Worked example 8.124: The energy of the PS monopole from its asymptotics

On any Bogomolny solution the energy density is a total Laplacian. With $\varepsilon = \frac{1}{2}(|B|^2 + |D\phi|^2)$ and $B = -D\phi$,

$$\varepsilon \overset{(1)}{=} |D\phi|^2 \overset{(2)}{=} \partial_i(\phi^a D_i \phi^a) - \phi^a D_i D_i \phi^a \overset{(3)}{=} \partial_i(\phi^a D_i \phi^a) \overset{(4)}{=} \tfrac{1}{2}\nabla^2 |\phi|^2.$$

(1) $|B| = |D\phi|$ on a solution.
(2) Leibniz rule for the gauge-invariant product $\partial_i(\vec{\phi} \cdot D_i\vec{\phi}) = D_i\vec{\phi} \cdot D_i\vec{\phi} + \vec{\phi} \cdot D_i D_i \vec{\phi}$.
(3) The second field equation (8.282): $D_i D_i \phi = 0$.
(4) $\phi^a D_i \phi^a = \phi^a \partial_i \phi^a = \frac{1}{2}\partial_i |\phi|^2$, since $\vec{\phi} \cdot (\vec{A}_i \times \vec{\phi}) = 0$.

Then Gauss' divergence theorem gives

$$E \overset{(1)}{=} \frac{1}{2} \lim_{R\to\infty} \oint_{S^2_R} \partial_r |\phi|^2 \, R^2 \mathrm{d}\Omega \overset{(2)}{=} \frac{1}{2} \lim_{R\to\infty} 4\pi R^2 \cdot 2\Big(C - \frac{1}{R}\Big)\frac{1}{R^2} \overset{(3)}{=} 4\pi C.$$

(1) Integrate $\frac{1}{2}\nabla^2|\phi|^2$ over a large ball.
(2) $|\phi| = h \approx C - 1/R$, so $\partial_r |\phi|^2 = 2|\phi|\partial_r|\phi| \approx 2(C - 1/R)/R^2$.
(3) Take the limit.

This is the Bogomolny bound saturated with $k = 1$. Note how the $1/r$ tail of the Higgs field carries the whole answer, which is why Nash's $O(r^{-2})$ cannot be right. At the centre, $h \approx C^2 r/3$ and $K \approx 1$, so $\varepsilon(0) = 2(C^2/3)^2 + (C^2/3)^2 = C^4/3$.

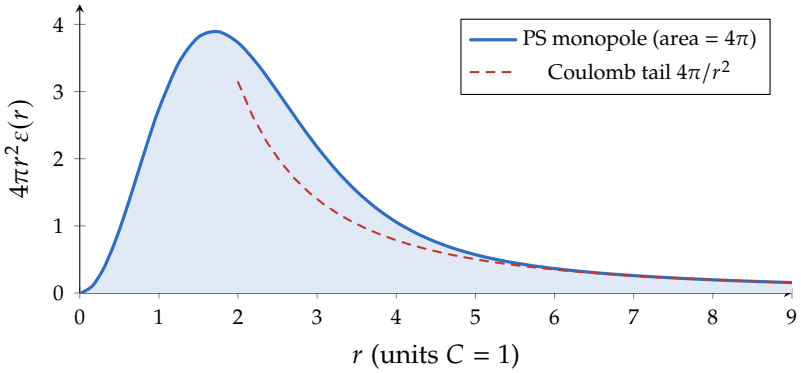


Figure 8.63: Radial energy density of the Prasad–Sommerfield monopole. The shaded area is the total energy $4\pi C$. The monopole has a core of size $\sim 1/C$ (the W Compton wavelength), but no sharp edge: the energy density (here $4\pi r^2 \varepsilon$) falls off only as a power, $4\pi/r^2$ (dashed), because the magnetic field and the massless Higgs field both have Coulomb tails.

Watch out: BPS is not the same as "λ small"

For $\lambda > 0$ the Higgs field is massive, the Higgs tail is exponential, and the energy is strictly above $4\pi C$ (numerically it grows from $4\pi C$ at $\lambda = 0$ to about $1.787 \times 4\pi C$ as $\lambda \to \infty$). The Bogomolny equation holds only at $\lambda = 0$. This distinction will matter for forces: BPS monopoles do not attract or repel at rest, whereas monopoles with $\lambda > 0$ do.

8.13.7 Unitary gauge and the Dirac string

Far from its core the PS monopole should look like a Dirac monopole. To see it, perform a gauge transformation that rotates the Higgs field to a *fixed* direction, $\phi \mapsto T_3$ up to its length. This is **unitary gauge**. In it the unbroken U(1) is the same at every point, and the abelian gauge field can be read off as the T_3 component.

Worked example 8.125: The hedgehog in unitary gauge

Use spherical angles (θ, φ) on S^2_∞ and the gauge transformation

$$g_N(\theta, \varphi) = e^{\varphi T_3}\, e^{\theta T_2}\, e^{-\varphi T_3}.$$

Rotation by $e^{\theta T_2}$ tilts the 3-axis by the angle θ towards the 1-axis, and conjugating by $e^{\varphi T_3}$ turns that tilt to azimuth φ. So $g_N T_3 g_N^{-1} = \hat{x}^a T_a$, i.e. $g_N^{-1}\hat{\phi}\, g_N = T_3$. At infinity the gauge field (8.293) (with $\alpha = 0$) becomes, by direct computation (we checked it by computer algebra),

$$A' = g_N^{-1} A\, g_N + g_N^{-1} \mathrm{d} g_N = -(1-\cos\theta)\,\mathrm{d}\varphi\; T_3, \qquad F' = \mathrm{d}A' = -\sin\theta\,\mathrm{d}\theta \wedge \mathrm{d}\varphi\; T_3. \quad (8.298)$$

Compare with the Dirac potential (8.270): $(A', T_3) = -(1-\cos\theta)\mathrm{d}\varphi = a_N$ with $g = -1$. The flux is $\oint (F', T_3) = -4\pi$, and (8.291) gives $k = +1$.
The gauge transformation g_N is smooth at the north pole ($g_N = 1$ at $\theta = 0$) but not at the south pole: at $\theta = \pi$, $g_N = e^{\varphi T_3}(2T_2)e^{-\varphi T_3}$ still depends on φ. So A' is singular along the negative x^3-axis: the Dirac string. It is a gauge artefact. In the original ("radial") gauge the fields are perfectly smooth; the string appears only because no smooth gauge transformation can rotate a degree-1 map $\hat{\phi}$ to a constant, which is the statement that $\deg \hat{\phi} \neq 0$.

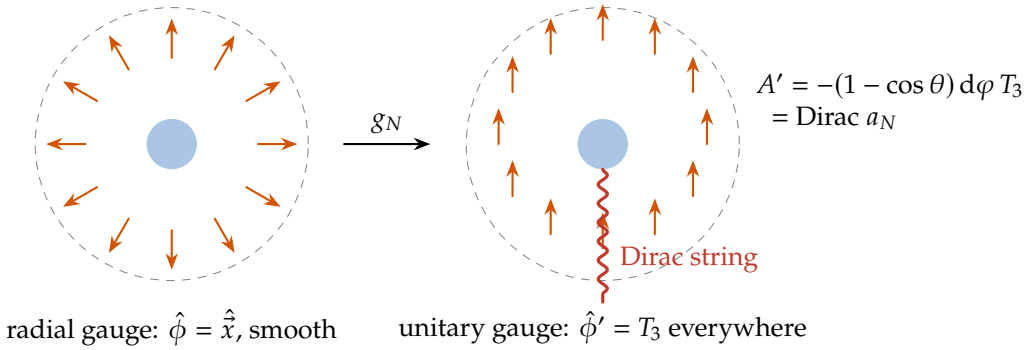


Figure 8.64: The same monopole in two gauges. Left: the Higgs direction (orange) points radially and every field is smooth. Right: after the gauge transformation g_N the Higgs field points "up" everywhere, the unbroken U(1) is the same at each point, and the T_3 part of A is exactly the Dirac potential. The price is a singular line, the Dirac string, where g_N is ill defined.

Try it: Dirac quantisation from the non-abelian side

In unitary gauge a field in the fundamental representation of SU(2) has T_3-eigenvalues $\mp\frac{i}{2}$, i.e. electric charges $q = \pm\frac{1}{2}$ under the unbroken U(1), and the W bosons have $q = \pm 1$. Check Dirac's condition $2qg \in \mathbb{Z}$ of Example 8.114 for the 't Hooft–Polyakov monopole. *Answer:* $g = -1$, so $2qg = \mp 1$ for the doublet and ∓ 2 for $W^\pm$; both are integers. The monopole carries the *minimal* Dirac charge allowed by the doublet, and twice the minimal charge

allowed by the W alone.

Reading Nash (pp. 243–245, (8.141)–(8.155))

Nash takes the Prasad–Sommerfield limit $\lambda = 0$ ((8.141)–(8.142)), states the Bogomolny equation $F = \star \mathrm{d}_A\phi$ (8.143) and checks it against both field equations by the direct calculation (8.144)–(8.145) that we reproduced in Example 8.119. He then completes the square (8.146), identifies the minima (8.147) and the topological cross term (8.148)–(8.149), and observes that on closed M the cross term vanishes (8.150), forcing $F = \mathrm{d}_A\phi = 0$ (8.151). On $\mathbb{R}^3$ he imposes (8.152) (see our pitfall about the rate), defines the magnetic charge (8.153)–(8.154) and rewrites it as a winding number (8.155). We fixed signs so that $k = \deg\hat{\phi}$ and so that the monopole (rather than the anti-monopole) obeys $F = -\star\,\mathrm{d}_A\phi$, which matches his later (8.162). The explicit solution, the energy computation and unitary gauge are not in Nash; they are the standard way to see that the abstract charge is a real magnetic monopole.

Recap

- At $\lambda = 0$ the energy is $\frac{1}{2}(\|F\|^2 + \|\mathrm{d}_A\phi\|^2)$. Completing the square gives $E = \frac{1}{2}\|F \pm \star\mathrm{d}_A\phi\|^2 + 4\pi C(\pm k)$, so $E \geq 4\pi C|k|$.
- The bound is saturated exactly by the first-order Bogomolny equation $F = \mp\star\,\mathrm{d}_A\phi$, which implies the second-order field equations.
- On a closed M the charge is zero and Bogomolny solutions are flat; they are non-trivial only through holonomy.
- On $\mathbb{R}^3$ the charge $k = \deg\hat{\phi} = c_1(L \to S^2_\infty)$ is an integer.
- The Prasad–Sommerfield monopole $h = C\coth Cr - 1/r$, $K = Cr/\sinh Cr$ has $E = 4\pi C$, a massive core of size $1/C$, and Coulomb tails; in unitary gauge it is a Dirac monopole with a gauge-artefact string.

8.14 Monopoles as symmetric instantons

We now have two first-order equations. On $\mathbb{R}^4$ there is the self-duality equation $\hat{F} = \star\hat{F}$ of the instantons. On $\mathbb{R}^3$ there is the Bogomolny equation $F = -\star\,\mathrm{d}_A\phi$ of the monopoles. They were found separately, but they look alike: both come from completing a square, both saturate a topological bound, and both say "one piece of the curvature equals the Hodge dual of another". This section explains the resemblance.

> *Is the Bogomolny equation a special case of the self-duality equation? If so, which symmetry of an instanton turns it into a monopole, and what happens to the topological charges?*

The answers, following Nash (who follows Atiyah):

- An instanton on $\mathbb{R}^4$ that does not depend on one coordinate x^0 is a monopole on $\mathbb{R}^3$. The Higgs field is the fourth component of the gauge field. This is **dimensional reduction**.
- Such an instanton has infinite action, because it is the same at every "time". A better symmetry is a *rotation* in a plane of $\mathbb{R}^4$. A rotation-invariant instanton is a monopole on

hyperbolic space H^3, and now the action is finite. The instanton number is then $c_2 = 2nk$, where k is the magnetic charge and n an integer that fixes the size of the Higgs field at infinity.

- The general mechanism is a conformal equivalence $\mathbb{R}^n \setminus \mathbb{R}^m \simeq S^{n-m-1} \times H^{m+1}$.

8.14.1 Dimensional reduction: time-independent instantons

Let $\hat{A} = \hat{A}_\mu \mathrm{d}x^\mu$ ($\mu = 0,1,2,3$) be an SU(2) connection on $\mathbb{R}^4$ with the Euclidean metric. This is Nash's (8.156). Its curvature has components

$$\hat{F}_{\mu\nu} = \partial_\mu \hat{A}_\nu - \partial_\nu \hat{A}_\mu + [\hat{A}_\mu, \hat{A}_\nu].$$

We use Nash's orientation $\mathrm{d}x^0 \wedge \mathrm{d}x^1 \wedge \mathrm{d}x^2 \wedge \mathrm{d}x^3$, i.e. $\varepsilon_{0123} = +1$. (Part 8b used $\mathrm{d}x^1 \wedge \mathrm{d}x^2 \wedge \mathrm{d}x^3 \wedge \mathrm{d}x^4$; if we call $x^4 = x^0$ the two orientations differ by a sign, so "self-dual" here is "anti-self-dual" there. Nothing below depends on this beyond an overall sign of the charge.) Self-duality in components is Nash's (8.157):

$$\hat{F}_{\mu\nu} = \tfrac{1}{2}\varepsilon_{\mu\nu\alpha\beta}\hat{F}_{\alpha\beta}. \tag{8.299}$$

Worked example 8.126: Splitting self-duality into space and time: Nash's (8.158)

Take $\mu\nu = ij$ with Latin indices $1,2,3$. In the sum over $\alpha\beta$ one index must be 0 and the other some k:

$$\hat{F}_{ij} \overset{(1)}{=} \tfrac{1}{2}\big(\varepsilon_{ij0k}\hat{F}_{0k} + \varepsilon_{ijk0}\hat{F}_{k0}\big) \overset{(2)}{=} \varepsilon_{ijk0}\hat{F}_{k0} \overset{(3)}{=} -\varepsilon_{0ijk}\hat{F}_{k0} \overset{(4)}{=} -\varepsilon_{ijk}\hat{F}_{k0}.$$

(1) $\varepsilon_{ij\alpha\beta}$ vanishes unless $\{\alpha,\beta\} = \{0,k\}$ for the third spatial index k.
(2) Both terms are equal, since swapping $0 \leftrightarrow k$ flips both ε and $\hat{F}$.
(3) Moving the index 0 from the last slot to the first takes three transpositions: a sign $(-1)^3$.
(4) $\varepsilon_{0ijk} = \varepsilon_{ijk}$ defines the three-dimensional symbol.

Now take $\mu\nu = i0$:

$$\hat{F}_{i0} \overset{(1)}{=} \tfrac{1}{2}\varepsilon_{i0jk}\hat{F}_{jk} \overset{(2)}{=} -\tfrac{1}{2}\varepsilon_{0ijk}\hat{F}_{jk} = -\tfrac{1}{2}\varepsilon_{ijk}\hat{F}_{jk}.$$

(1) Now both α,β must be spatial.
(2) One transposition, $\varepsilon_{i0jk} = -\varepsilon_{0ijk}$.

These are the two halves of (8.158). They say the same thing: contract the first with $\frac{1}{2}\varepsilon_{ijl}$ and use $\varepsilon_{ijl}\varepsilon_{ijk} = 2\delta_{lk}$ to get $\frac{1}{2}\varepsilon_{ijl}\hat{F}_{ij} = -\hat{F}_{l0}$, which is the second. Of the six components of (8.299) only three are independent, as it should be: Λ^2_+ is three-dimensional.

Now impose **x^0-independence**: $\partial_0 \hat{A}_\mu = 0$. Then

$$\hat{F}_{k0} = \partial_k \hat{A}_0 - \partial_0 \hat{A}_k + [\hat{A}_k, \hat{A}_0] = \partial_k \hat{A}_0 + [\hat{A}_k, \hat{A}_0], \tag{8.300}$$

which is Nash's (8.159). This is a covariant derivative of $\hat{A}_0$ with respect to the three-dimensional connection made of the other components. So define, as in Nash's (8.160), the pair on $\mathbb{R}^3$

$$A = \hat{A}_1 \mathrm{d}x^1 + \hat{A}_2 \mathrm{d}x^2 + \hat{A}_3 \mathrm{d}x^3, \qquad \phi = \hat{A}_0. \tag{8.301}$$

Under a gauge transformation $g(x^1,x^2,x^3)$ that is independent of x^0, $\hat{A}_0 \mapsto g^{-1}\hat{A}_0 g$: the fourth component of the gauge field transforms in the adjoint, exactly like a Higgs field. In these

variables $\hat{F}_{ij} = F_{ij}$, and (8.300) says $\hat{F}_{k0} = D_k\phi$. The first half of (8.158) becomes Nash's (8.161):

$$F_{ij} = -\varepsilon_{ijk}\big(\partial_k\phi + [A_k, \phi]\big) = -\varepsilon_{ijk}D_k\phi. \tag{8.302}$$

Apply the three-dimensional Hodge star: $\star F = B_i \mathrm{d}x^i$ with $B_i = \frac{1}{2}\varepsilon_{ijk}F_{jk} = -\frac{1}{2}\varepsilon_{ijk}\varepsilon_{jkl}D_l\phi = -D_i\phi$. So

$$\boxed{F = -\star \mathrm{d}_A\phi} \tag{8.303}$$

which is Nash's (8.162), and exactly the monopole sign we used in Section 8.13. *Self-dual connections on $\mathbb{R}^4$ that are invariant under translations in x^0 are the same thing as solutions of the Bogomolny equation on $\mathbb{R}^3$.*

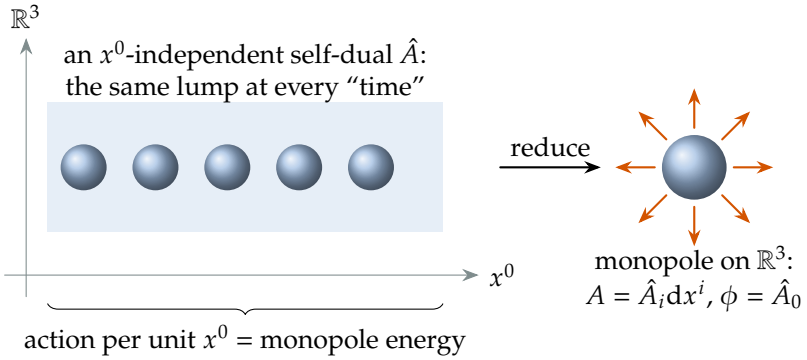


Figure 8.65: Dimensional reduction. A self-dual connection on $\mathbb{R}^4$ that does not depend on x^0 is a "tube" in the x^0 direction. Each slice $x^0 = \text{const}$ carries the same field on $\mathbb{R}^3$, and the component $\hat{A}_0$ becomes an adjoint scalar, the Higgs field. The self-duality equations become the Bogomolny equation, and the four-dimensional action per unit length in x^0 is the three-dimensional energy. The total action is infinite.

Worked example 8.127: What becomes of the action and the instanton number

Write the reduced curvature as a 2-form on $\mathbb{R}^4$. Since $\hat{F}_{k0}\mathrm{d}x^k \wedge \mathrm{d}x^0 = D_k\phi\, \mathrm{d}x^k \wedge \mathrm{d}x^0$,

$$\hat{F} = F + \mathrm{d}_A\phi \wedge \mathrm{d}x^0.$$

Action. With the norm of part 8b, $S = -\int \mathrm{tr}(\hat{F} \wedge \star\hat{F}) = \frac{1}{4}\int \hat{F}^a_{\mu\nu}\hat{F}^a_{\mu\nu}\,\mathrm{d}^4x$, and over a slab $0 \le x^0 \le T$,

$$S_{\text{slab}} \overset{(1)}{=} \frac{1}{2}\int_0^T \mathrm{d}x^0 \int_{\mathbb{R}^3}\Big(\sum_{i<j} F^a_{ij}F^a_{ij} + \sum_k D_k\phi^a D_k\phi^a\Big)\mathrm{d}^3x \overset{(2)}{=} T\cdot\frac{1}{2}\int_{\mathbb{R}^3}\big(|B|^2 + |D\phi|^2\big)\mathrm{d}^3x \overset{(3)}{=} T\,E.$$

(1) $\frac{1}{4}\sum_{\mu,\nu} = \frac{1}{2}\sum_{\mu<\nu}$, and the pairs $\mu < \nu$ are the spatial pairs ij and the pairs $0k$ with $\hat{F}_{0k} = -D_k\phi$.
(2) Nothing depends on x^0; and $\sum_{i<j}(F^a_{ij})^2 = |B|^2$.
(3) The energy (8.281) in components.

Instanton number. In three dimensions $F \wedge F = 0$ (a 4-form) and $\mathrm{d}x^0 \wedge \mathrm{d}x^0 = 0$, so only the cross terms survive:

$$\mathrm{tr}(\hat{F} \wedge \hat{F}) \overset{(1)}{=} 2\ \mathrm{tr}(F \wedge \mathrm{d}_A\phi) \wedge \mathrm{d}x^0 \overset{(2)}{=} -2\,\mathrm{d}x^0 \wedge \mathrm{tr}(F \wedge \mathrm{d}_A\phi) \overset{(3)}{=} -2\,\mathrm{d}x^0 \wedge \mathrm{d}\ \mathrm{tr}(F\phi).$$

(1) The two cross terms are equal by graded cyclicity (a 2-form and a 2-form commute under the trace).
(2) Move the 1-form $\mathrm{d}x^0$ past a 3-form: sign $(-1)^3$.

(3) The total-derivative identity (8.286).

Integrate over the slab with the orientation $\mathrm{d}x^0 \wedge \mathrm{d}^3x$ and use (8.291), $\int_{\mathbb{R}^3} \mathrm{d}\,\mathrm{tr}(F\phi) = \oint \mathrm{tr}(F\phi) = 2\pi C k$:

$$k_{\mathrm{inst}}(\mathrm{slab}) = -\frac{1}{8\pi^2}\int \mathrm{tr}(\hat{F}\wedge\hat{F}) = \frac{2T\cdot 2\pi C k}{8\pi^2} = \frac{C\,T}{2\pi}\,k.$$

So the Bogomolny bound in three dimensions is the instanton bound per unit time: $S_{\mathrm{slab}} \geq 8\pi^2 k_{\mathrm{inst}}$ reads $TE \geq 4\pi C k T$, i.e. $E \geq 4\pi C k$. Both sides are infinite as $T \to \infty$. This is Nash's remark that time-independent instantons are instantons "in a loose sense" only.

Physical picture: Periodic time and calorons

If instead of $\mathbb{R}^4$ we take $S^1_\beta \times \mathbb{R}^3$ with x^0 periodic of period β (finite temperature), x^0-independent configurations have finite action βE. Then the Higgs field is $\hat{A}_0$, whose holonomy $\exp(-\beta\hat{A}_0)$ round the thermal circle is gauge invariant, and the magnetic charge contributes $C\beta k/2\pi$ to the instanton number. Instantons on $S^1 \times \mathbb{R}^3$ (**calorons**) interpolate between instantons (small compared to β) and chains of monopoles (large compared to β). We shall see the same interpolation in the hyperbolic picture below.

8.14.2 Conformal invariance of self-duality

To get finite action we need a symmetry whose orbits are *compact*. The natural candidate is a rotation. The key tool is that self-duality in four dimensions does not see the size of the metric, only its angles.

Working definition 8.41 (conformal change of metric)

A **conformal change** replaces g by $g' = \Omega^2 g$ with Ω a positive function. It rescales lengths by Ω point by point and leaves angles unchanged. *Test:* two metrics are conformally related iff their matrices [the components $g_{\mu\nu}(x)$ in one coordinate system] are proportional at every point.

Worked example 8.128: The Hodge star on middle-degree forms is conformally invariant

Recall $(\star\alpha)_{\nu_1\ldots\nu_{n-p}} = \frac{1}{p!}\sqrt{\det g}\; g^{\mu_1\rho_1}\cdots g^{\mu_p\rho_p}\alpha_{\rho_1\ldots\rho_p}\varepsilon_{\mu_1\ldots\mu_p\nu_1\ldots\nu_{n-p}}$ (Chapter T1). Replace g by $\Omega^2 g$:

$$\star'\alpha \overset{(1)}{=} \left(\Omega^n\sqrt{\det g}\right)\left(\Omega^{-2}\right)^p(\ldots) \overset{(2)}{=} \Omega^{n-2p}\;\star\alpha.$$

(1) $\det(\Omega^2 g) = \Omega^{2n}\det g$, and each inverse metric carries Ω^{-2}; the index structure is unchanged.

(2) Collect the powers.

For $n = 4$ and $p = 2$ the power is $\Omega^0 = 1$. So $\star$ on 2-forms, and with it the equation $\hat{F} = \star\hat{F}$, is the same for g and $\Omega^2 g$. The Yang–Mills action density $\mathrm{tr}(\hat{F}\wedge\star\hat{F})$ is also unchanged, since $\hat{F}$ does not involve the metric at all. So in four dimensions both the equation *and* the action are conformally invariant.

8.14.3 Rotations: $\mathbb{R}^4 \setminus \mathbb{R}^2 \simeq S^1 \times H^3$

Following Nash, look for instantons invariant under rotations of the x^0–x^1 plane. Use polar coordinates (r, θ) in that plane, $x^0 = r\cos\theta$, $x^1 = r\sin\theta$. The flat metric becomes Nash's (8.163):

$$\mathrm{d}s^2 \overset{(1)}{=} \mathrm{d}r^2 + r^2\mathrm{d}\theta^2 + \mathrm{d}x_2^2 + \mathrm{d}x_3^2 \overset{(2)}{=} r^2\Big\{\mathrm{d}\theta^2 + \frac{\mathrm{d}r^2 + \mathrm{d}x_2^2 + \mathrm{d}x_3^2}{r^2}\Big\}. \tag{8.304}$$

(1) The flat metric of the x^0–x^1 plane in polar coordinates.
(2) Factor out r^2, which is legitimate where $r > 0$.

By Example 8.128 we may drop the factor r^2 without changing the self-duality equations. The new metric is Nash's (8.164):

$$\mathrm{d}s'^2 = \mathrm{d}\theta^2 + \frac{\mathrm{d}r^2 + \mathrm{d}x_2^2 + \mathrm{d}x_3^2}{r^2}, \qquad 0 < r < \infty, \quad x_2, x_3 \in \mathbb{R}, \quad \theta \in [0, 2\pi). \tag{8.305}$$

The first term is a circle of radius 1. The second term is the **upper half-space model** of three-dimensional hyperbolic space H^3, with the height coordinate called r.

Working definition 8.42 (hyperbolic space H^3, two models)

Half-space model: $\{(r, x_2, x_3) : r > 0\}$ with $\mathrm{d}s^2 = (\mathrm{d}r^2 + \mathrm{d}x_2^2 + \mathrm{d}x_3^2)/r^2$. *Ball model (radius 2):* $\{\vec{x} \in \mathbb{R}^3 : |\vec{x}| < 2\}$ with

$$\mathrm{d}s^2 = \frac{\mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2}{(1 - \rho^2/4)^2}, \qquad \rho = |\vec{x}|, \tag{8.306}$$

which is Nash's (8.167). Both have constant sectional curvature -1 [every geodesic triangle has angle sum equal to π minus its area]. The **boundary at infinity** ∂H^3 is the plane $r = 0$ plus one point (half-space), or the sphere $\rho = 2$ (ball); it is infinitely far away in the hyperbolic metric but is a perfectly good S^2. *How to recognise it:* a metric that is flat space divided by the square of (a multiple of) the distance to a boundary.

Worked example 8.129: The half-space metric is really at infinite distance from its boundary

Walk up the vertical line $x_2 = x_3 = 0$ from height r_1 to r_2:

$$\text{length} = \int_{r_1}^{r_2} \frac{\mathrm{d}r}{r} = \ln\frac{r_2}{r_1} \to \infty \quad \text{as } r_1 \to 0.$$

So the plane $r = 0$ is infinitely far away. The volume of the region $r > \varepsilon$ above a unit square is $\int_\varepsilon^\infty r^{-3}\mathrm{d}r = 1/(2\varepsilon^2)$, which also diverges at the boundary. This is why the metric (8.305) is "singular at $r = 0$", as Nash puts it: the conformal factor r^{-2} blows up there.

What have we removed from $\mathbb{R}^4$? The coordinates (r, θ) fail at $r = 0$, i.e. $x^0 = x^1 = 0$. That is the plane $\mathbb{R}^2$ spanned by x^2, x^3, the fixed set ("axis") of the rotation. Therefore we have used the conformal equivalence

$$\mathbb{R}^4 \setminus \mathbb{R}^2 \;\simeq\; S^1 \times H^3, \tag{8.307}$$

Nash's (8.165). The axis $\mathbb{R}^2$ has become the boundary at infinity of H^3 (the plane $r = 0$). Adding the point at infinity of $\mathbb{R}^4$, i.e. working on S^4, the axis becomes a 2-sphere and matches $\partial H^3 = S^2$.

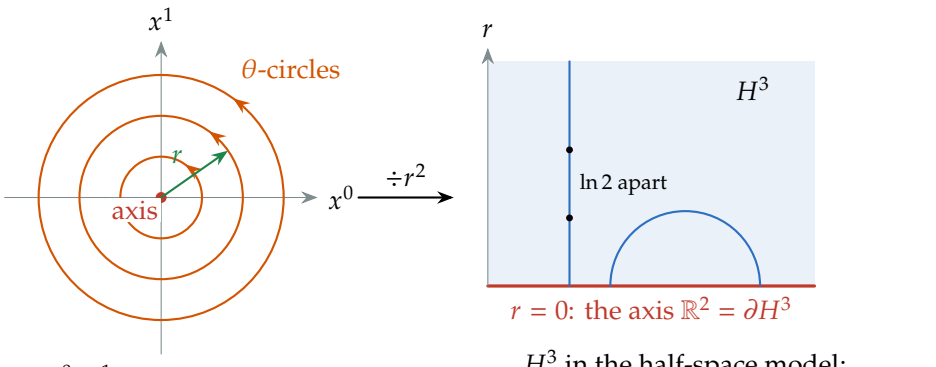


Figure 8.66: Nash's conformal equivalence $\mathbb{R}^4 \setminus \mathbb{R}^2 \simeq S^1 \times H^3$. Left: in the plane of rotation every point except the origin lies on a circle of radius r. Dividing the flat metric by r^2 gives each circle length 2π, and the remaining coordinates (r, x_2, x_3) form hyperbolic space. Right: the axis of rotation ($r = 0$) is pushed off to infinity and becomes the boundary plane of H^3.

8.14.4 Rotation-invariant instantons are hyperbolic monopoles

Now repeat Section 8.14.1 with θ in place of x^0. Take a self-dual $\hat{A}$ on $S^1 \times H^3$ that is invariant under rotations of the circle. In a gauge where nothing depends on θ, write

$$\hat{A} = A + \phi \, \mathrm{d}\theta, \tag{8.308}$$

where A is a connection on H^3 and ϕ an adjoint scalar on H^3. The metric (8.305) is a product with a circle of unit radius, so the computation of Example 8.126 goes through with $x^0 \to \theta$, flat $\mathbb{R}^3 \to H^3$ and the Hodge star of $\mathbb{R}^3$ replaced by that of the hyperbolic metric. The result is Nash's (8.166):

$$\mathrm{d}_A \phi = - \star_H F \qquad (\iff F = - \star_H \mathrm{d}_A \phi, \text{ since } \star_H \star_H = 1). \tag{8.309}$$

This is the Bogomolny equation on H^3, and its solutions are **hyperbolic monopoles**. The boundary condition replacing $|\phi| \to C$ is that $|\phi|$ tends to a constant on ∂H^3. Nash calls it $|\phi|_{\mathrm{as}}$ ("asymptotic").

Worked example 8.130: The hyperbolic Bogomolny equation in the ball, in components

In the ball model the metric is $g = f\,\delta$ with $f = (1 - \rho^2/4)^{-2}$. For a conformally flat metric in three dimensions, $\sqrt{\det g} = f^{3/2}$ and $g^{ij} = f^{-1}\delta^{ij}$, so for a 2-form $F = \frac{1}{2}F_{ij}\mathrm{d}x^i \wedge \mathrm{d}x^j$,

$$(\star_H F)_k \overset{(1)}{=} \tfrac{1}{2}\sqrt{\det g}\; g^{ii'} g^{jj'} F_{i'j'}\, \varepsilon_{ijl}\, g_{lk} \overset{(2)}{=} \tfrac{1}{2} f^{3/2} f^{-2} f\, \varepsilon_{ijk} F_{ij} = \tfrac{1}{2}\sqrt{f}\, \varepsilon_{ijk} F_{ij}.$$

(1) The general formula for $\star$ with the free index lowered.
(2) Insert $g = f\delta$.

So (8.309) reads

$$D_k \phi^a = -\tfrac{1}{2}\sqrt{f}\, \varepsilon_{ijk} F^a_{ij}, \qquad f = \left(1 - \rho^2/4\right)^{-2}. \tag{8.310}$$

Near the centre of the ball $f \approx 1$ and this is the flat Bogomolny equation. Near the boundary $\sqrt{f} \to \infty$: the curvature must fall off fast (in coordinate terms) to keep $\mathrm{d}_A\phi$ finite.

The isometry between the two models

Nash says the ball model is "completely equivalent" to the half-space model via an isometry $T = \alpha^{-1}\beta\alpha$ (his (8.168)), built from stereographic projection α and a quarter-turn β of S^3. We give T explicitly and check it.

Worked example 8.131: An explicit isometry from the ball to the half-space

Let $\vec{x}$ be in the ball $|\vec{x}| < 2$, let $\vec{e}_1 = (1,0,0)$ and $\vec{y} = \vec{x} + 2\vec{e}_1$. Define

$$\vec{w} = T(\vec{x}) = \frac{8\,\vec{y}}{|\vec{y}|^2} - 2\vec{e}_1, \qquad (w_1, w_2, w_3) = (r, x_2, x_3). \tag{8.311}$$

It is an inversion in the sphere of radius $2\sqrt{2}$ about $-2\vec{e}_1$, followed by a shift. First the height w_1:

$$w_1 \overset{(1)}{=} \frac{8(x_1+2) - 2|\vec{y}|^2}{|\vec{y}|^2} \overset{(2)}{=} \frac{8x_1 + 16 - 2\rho^2 - 8x_1 - 8}{|\vec{y}|^2} = \frac{2(4-\rho^2)}{|\vec{y}|^2}.$$

(1) Common denominator.
(2) $|\vec{y}|^2 = \rho^2 + 4x_1 + 4$.

So $w_1 > 0$ exactly when $\rho < 2$: the ball goes to the upper half-space, and the sphere $\rho = 2$ goes to the plane $w_1 = 0$. Now the metric. An inversion $\vec{y} \mapsto \vec{y}/|\vec{y}|^2$ multiplies lengths by $|\vec{y}|^{-2}$, so $|\mathrm{d}\vec{w}|^2 = 64|\mathrm{d}\vec{x}|^2/|\vec{y}|^4$, and

$$\frac{|\mathrm{d}\vec{w}|^2}{w_1^2} \overset{(1)}{=} \frac{64\,|\mathrm{d}\vec{x}|^2}{|\vec{y}|^4} \cdot \frac{|\vec{y}|^4}{4(4-\rho^2)^2} \overset{(2)}{=} \frac{16\,|\mathrm{d}\vec{x}|^2}{(4-\rho^2)^2} \overset{(3)}{=} \frac{|\mathrm{d}\vec{x}|^2}{(1-\rho^2/4)^2}.$$

(1) Insert the two results above.
(2) Cancel $|\vec{y}|^4$.
(3) Divide numerator and denominator by 16.

This is (8.306): T is an isometry. The centre of the ball goes to $T(0) = 8 \cdot 2\vec{e}_1/4 - 2\vec{e}_1 = 2\vec{e}_1$, the point $r = 2$, $x_2 = x_3 = 0$ of the half-space, as Nash states. (We checked the pull-back with computer algebra as well.)

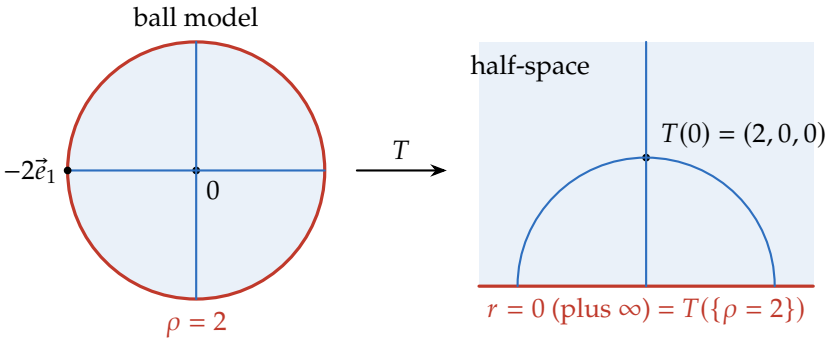


Figure 8.67: The isometry T of Example 8.131, drawn in a plane through the axis. The centre of the ball goes to the point at height 2 above the origin; the boundary sphere goes to the boundary plane (the point $-2\vec{e}_1$ of the sphere goes to infinity). Diameters of the ball (blue lines) become a vertical line and a semicircle, both geodesics.

8.14.5 Finite action, and $c_2 = 2nk$

Here is the difference from the time-independent case: a hyperbolic monopole gives an instanton of *finite* action. Integrating out θ costs only a factor 2π, and by conformal invariance the action on $S^1 \times H^3$ equals the action on $\mathbb{R}^4 \setminus \mathbb{R}^2$, hence on $\mathbb{R}^4$ (a plane has measure zero), and hence on S^4. So we expect an integer relation between three numbers: the instanton number $c_2(V)$ of the rank-2 bundle $V \to S^4$, the magnetic charge k, and the size $|\phi|_{\text{as}}$ of the Higgs field at infinity. Nash explains why the last one is quantised.

The rotation acts on the fibres over the axis. A rotation R_θ in the x^0–x^1 plane moves most points, but it fixes every point of the axis. For $\hat{A}$ to be invariant, the rotation must be lifted to the bundle V. Over a point p of the axis the lift maps the fibre $V_p \cong \mathbb{C}^2$ to itself by an SU(2) matrix M_θ, and $\theta \mapsto M_\theta$ is a group homomorphism $\mathrm{U}(1) \to \mathrm{SU}(2)$. Every such homomorphism can be conjugated into the diagonal torus, and $M_{2\pi} = M_0 = \mathbb{1}$ forces integer weights. So, as in Nash's (8.169),

$$M_\theta = \begin{pmatrix} e^{in\theta} & 0 \\ 0 & e^{-in\theta} \end{pmatrix}, \qquad n \in \mathbb{Z}. \tag{8.312}$$

The Higgs field is the generator of this action. The infinitesimal generator is $\mathrm{d}M_\theta/\mathrm{d}\theta|_0 = \mathrm{diag}(in, -in) = -2n\,T_3$. In the invariant gauge (8.308) the covariant derivative of an invariant section along ∂_θ is (derivative of the lifted action) $+\,\phi$. On the axis the vector field ∂_θ vanishes, so this covariant derivative is zero there, and ϕ must equal (minus) the generator of the fibre action: $\phi|_{\text{axis}}$ is conjugate to $\pm 2nT_3$. Its length in our norm is

$$|\phi|_{\text{as}} = |2nT_3| = 2n. \tag{8.313}$$

Nash, who normalises $|\phi|$ by its eigenvalues $\pm i|\phi|$ rather than by $-2\,\mathrm{tr}\,\phi^2$, writes this as $|\phi|_{\text{as}} = n$. Either way it is quantised.

The bundle over the axis splits. Restrict V to the axis-sphere $S^2 = \partial H^3$; call the restriction V'. The circle acts on each fibre of V' with weights $\pm n$, so V' splits into the two eigenline bundles,

$$V' = L \oplus L^{-1}, \tag{8.314}$$

Nash's (8.170). L is the eigenline of ϕ with eigenvalue $+in$ (in Nash's normalisation). Its Chern class is the degree of the map $\hat{\phi} : \partial H^3 \to S^2$, for exactly the reason of Example 8.121: $c_1(L) = k$, the magnetic charge.

Worked example 8.132: Deriving $c_2(V) = 2nk$

The hyperbolic Bogomolny bound is proved exactly as on $\mathbb{R}^3$ (Section 8.13.3): the energy is a square plus $-\langle F, \star_H \mathrm{d}_A\phi\rangle = 2\oint_{\partial H^3} \mathrm{tr}(F\phi)$, and on the boundary only $|\phi|_{\text{as}}$ and the degree of $\hat{\phi}$ enter. So on a solution, $E_H = 4\pi|\phi|_{\text{as}}\,k$. The instanton number (with the conventions of Example 8.127) is then

$$c_2(V) \overset{(1)}{=} \frac{1}{8\pi^2} S(\hat{A}) \overset{(2)}{=} \frac{1}{8\pi^2}\int_0^{2\pi} \mathrm{d}\theta\; E_H \overset{(3)}{=} \frac{2\pi\cdot 4\pi|\phi|_{\text{as}}k}{8\pi^2} = |\phi|_{\text{as}}\,k \overset{(4)}{=} 2nk.$$

(1) The instanton is self-dual, so it saturates $S = 8\pi^2|c_2|$ (part 8b); and S is conformally invariant (Example 8.128), so we may compute it on $S^1 \times H^3$.
(2) The reduction of Example 8.127, with $\mathrm{d}x^0$ replaced by $\mathrm{d}\theta$ and T by 2π: the action per unit θ is the hyperbolic energy.
(3) The hyperbolic Bogomolny bound, saturated.
(4) (8.313).

This is Nash's (8.171). The three integers k, n and $c_2(V)$ together hold all the topological

data, and they are tied by one relation.

Watch out: Why $|\phi|_{\text{as}}$ must be an integer, and what if it is not

The Bogomolny equation on H^3 makes sense for *any* constant $|\phi|_{\text{as}} > 0$, and hyperbolic monopoles exist for any value. But only for the quantised values of (8.313) does the corresponding θ-invariant connection on $\mathbb{R}^4 \setminus \mathbb{R}^2$ extend smoothly across the axis. For other values the holonomy round a small circle linking the axis, $\exp(-2\pi\phi)$, does not tend to $\mathbb{1}$ as the circle shrinks, so the would-be instanton is singular along the axis and its "instanton number" $|\phi|_{\text{as}}k$ is not an integer (Forgács, Horváth and Palla). So hyperbolic monopoles form a wider class than axially symmetric instantons.

8.14.6 The curvature of H^3 as a parameter, and the flat limit

Nash next inserts a length scale R. Write the flat metric as

$$\mathrm{d}s^2 = \frac{r^2}{R^2}\Big\{R^2\mathrm{d}\theta^2 + \frac{R^2}{r^2}\big(\mathrm{d}r^2 + \mathrm{d}x_2^2 + \mathrm{d}x_3^2\big)\Big\}.$$

Deleting the conformal factor r^2/R^2 leaves $S^1(R) \times H^3(R)$: a circle of radius R times hyperbolic space scaled up by R. Scaling a metric by R^2 divides curvatures by R^2, so $H^3(R)$ has sectional curvature $-1/R^2$ and scalar curvature $3 \cdot 2 \cdot (-1/R^2) = -6/R^2$. Its ball model has radius $2R$ and metric $f_R\,|\mathrm{d}\vec{x}|^2$ with $f_R = (1 - \rho^2/4R^2)^{-2}$ (substitute $\vec{x} \to \vec{x}/R$ in (8.306) and multiply by R^2).

Worked example 8.133: Bogomolny on $S^1(R) \times H^3(R)$

The unit 1-form along the circle is now $R\,\mathrm{d}\theta$. Write $\hat{A} = A + \phi\,\mathrm{d}\theta = A + (\phi/R)(R\,\mathrm{d}\theta)$: the Higgs field seen by an orthonormal frame is $\Phi = \phi/R$. The computation of Example 8.130 with $f \to f_R$ gives $D_k\Phi^a = -\frac{1}{2}\sqrt{f_R}\,\varepsilon_{ijk}F^a_{ij}$, that is

$$D_k\phi^a = -\frac{R}{2}\sqrt{f_R}\,\varepsilon_{ijk}F^a_{ij}, \qquad f_R = \big(1 - \rho^2/4R^2\big)^{-2}. \tag{8.315}$$

As $R \to \infty$ at fixed $\vec{x}$, $f_R \to 1$ and the ball of radius $2R$ fills all of $\mathbb{R}^3$: $H^3(R)$ becomes flat $\mathbb{R}^3$ and (8.315) becomes the flat Bogomolny equation $D_k\Phi = -B_k$ for $\Phi = \phi/R$.

Reading Nash (the factor in (8.172))

Nash prints $D_k\phi^a = -\frac{R}{2\sqrt{f_R}}\varepsilon_{ijk}F^a_{jk}$, with $\sqrt{f_R}$ in the denominator. A direct computation of the Hodge star for $g = f_R\,\delta$ (Example 8.130) puts $\sqrt{f_R}$ in the numerator, as in (8.315). The difference is invisible in the limit $R \to \infty$ that Nash cares about, where $f_R \to 1$.

So flat monopoles are limits of hyperbolic ones, and hyperbolic monopoles with quantised $|\phi|_{\text{as}}$ are instantons. This gives a picture of an $\mathbb{R}^3$ monopole as a limit of instantons (Chakrabarti; Nash).

8.14.7 Where is the monopole? Zeros of the Higgs field

A monopole has no sharp edge (Figure 8.63), so "its position" needs a definition. Nash's is the set of zeros of ϕ. Why must ϕ have zeros at all?

Key idea: A non-zero charge forces zeros of the Higgs field

Let $\hat{\phi} : \partial H^3 \to S^2$ (or $S^2_\infty \to S^2$ on $\mathbb{R}^3$) have degree k. The map $\hat{\phi} = \phi/|\phi|$ extends to a map from the whole closed ball to S^2 without singularity if and only if $k = 0$. Hence if $k \neq 0$, the field ϕ vanishes somewhere. Counted with signs (local degrees), the zeros add up to k.

Why. Suppose ϕ never vanishes on the closed ball B. Then $\hat{\phi}$ is defined on every sphere S_ρ of radius $\rho \leq \rho_{\max}$, and the family $\hat{\phi}|_{S_\rho}$ is a homotopy from the map on the boundary to the map on a tiny sphere around the centre. The latter is nearly constant (it is $\hat{\phi}(0)$ up to a small error), so it has degree 0; degrees are homotopy invariant; so $k = 0$. Conversely, if $k = 0$ the boundary map is homotopic to a constant ($\pi_2(S^2) = \mathbb{Z}$ is detected by the degree), and we can use the homotopy to fill in the ball radially. The signed count follows by applying the same argument to the region outside small spheres around each zero: the boundary degree equals the sum of the degrees on the small spheres.

Worked example 8.134: Local degrees of zeros

Think of $\vec{\phi}(\vec{x})$ as a vector field on $\mathbb{R}^3$ and use complex notation $w = x + iy$ in the xy-plane.

- $\vec{\phi} = (x, y, z)$: one zero at the origin, local degree +1 (the hedgehog).
- $\vec{\phi} = (x, -y, z)$: one zero, local degree −1 (a reflection). This is the Higgs field of an anti-monopole.
- $\vec{\phi} = (\operatorname{Re} w^2, \operatorname{Im} w^2, z)$: the planar part $w \mapsto w^2$ wraps the circle twice, so the boundary map has degree 2; there is a single, degenerate, zero at the origin of local degree 2.
- $\vec{\phi} = (\operatorname{Re}(w^2 - a^2), \operatorname{Im}(w^2 - a^2), z)$ with $a > 0$: two simple zeros at $w = \pm a$, $z = 0$, each of local degree +1 (near $w = a$, $w^2 - a^2 \approx 2a(w - a)$, an orientation-preserving map). The total is again 2.

The last two illustrate how a charge-2 monopole can be two separated unit monopoles or two coincident ones; we meet this again in Section 8.15.

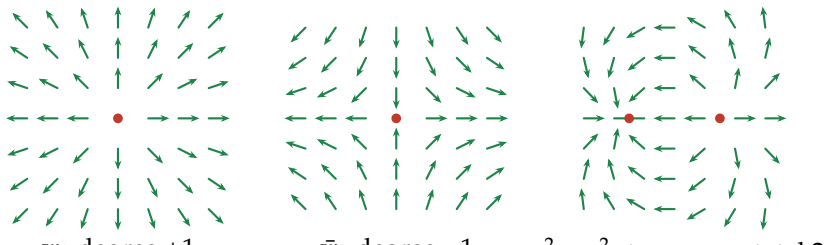


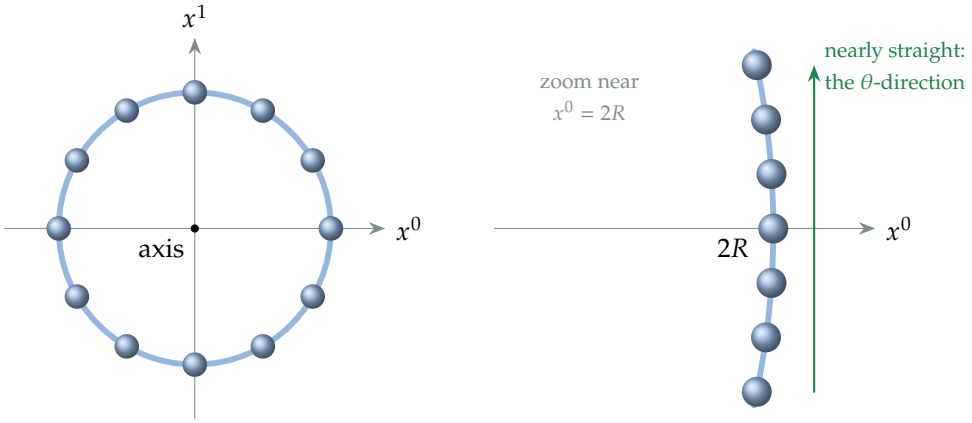
Figure 8.68: The planar part of three Higgs fields, drawn as unit vectors $\hat{\phi}$ in the xy-plane (the z-component is z and vanishes in this plane). Red dots are zeros. Walking once round a large circle, the arrow turns once (left), once backwards (middle) and twice (right). The count of turns at infinity equals the signed count of zeros inside: the monopoles "are" at the red dots.

The unit hyperbolic monopole as a ring of instanton

For the simplest hyperbolic monopole, $k = 1$, there is a single zero of ϕ, at the centre of the ball (Nash [1]). What is it in the instanton picture? Follow the zero through the maps:

$$\begin{aligned}\text{centre of the ball} &\xrightarrow{T} (r, x_2, x_3) = (2, 0, 0) \text{ in the half-space} \\ &\xrightarrow{\times S^1} \{x^0 = 2\cos\theta,\ x^1 = 2\sin\theta,\ x^2 = x^3 = 0\} \subset \mathbb{R}^4.\end{aligned}$$

The first arrow is Example 8.131. The second remembers that a point of H^3 stands for a whole θ-circle in $\mathbb{R}^4$, of radius r in the x^0–x^1 plane. So the "location" of the θ-invariant instanton is a circle of radius 2, centred at the origin in the x^0–x^1 plane: a ring of instanton.



$k = 1$, $R = 1$: the zero of ϕ is a circle of radius 2 (a ring of instanton)

$R \to \infty$: the ring becomes a line of instantons, i.e. a static $\mathbb{R}^3$ monopole

Figure 8.69: Instanton pictures of monopoles. Left: the unit hyperbolic monopole corresponds to an axially symmetric instanton "located" on a circle of radius 2 around the axis. Right: as the curvature radius R grows, the relevant part of the circle, of radius $2R$, straightens into a line. Along it (the θ-direction) the field is almost constant; that direction plays the role of time, so the picture is a static monopole. Chakrabarti made this limit precise.

In the limit $R \to \infty$, with larger k, the ring picture becomes a chain of instantons strung out along a line, and the result is an $\mathbb{R}^3$ monopole (Chakrabarti). This is the same interpolation as the calorons of the intuition box after Example 8.127.

8.14.8 The general conformal equivalence

Nash remarks that (8.307) is one case of

$$\mathbb{R}^n \setminus \mathbb{R}^m \ \simeq\ S^{n-m-1} \times H^{m+1}, \qquad m < n. \tag{8.316}$$

Worked example 8.135: Deriving the general equivalence: Nash's (8.174)

Ask for invariance under rotations of the first $l+1$ coordinates, i.e. under $\mathrm{SO}(l+1)$ acting on $(x_1, \dots, x_{l+1})$, with $l < n$. Let $r^2 = x_1^2 + \cdots + x_{l+1}^2$ and let $\mathrm{d}\Omega_l^2$ be the round metric of the unit S^l. Then

$$\begin{aligned}\mathrm{d}s^2 &\overset{(1)}{=} \mathrm{d}x_1^2 + \cdots + \mathrm{d}x_{l+1}^2 + \mathrm{d}x_{l+2}^2 + \cdots + \mathrm{d}x_n^2 \\ &\overset{(2)}{=} \mathrm{d}r^2 + r^2\mathrm{d}\Omega_l^2 + \mathrm{d}x_{l+2}^2 + \cdots + \mathrm{d}x_n^2\end{aligned}$$

$$\overset{(3)}{=} r^2\Big\{\mathrm{d}\Omega_l^2 + \frac{\mathrm{d}r^2 + \mathrm{d}x_{l+2}^2 + \cdots + \mathrm{d}x_n^2}{r^2}\Big\}.$$

(1) The flat metric.
(2) Spherical coordinates in the first $l+1$ variables.
(3) Factor out r^2 where $r > 0$.

The bracket is S^l times the half-space model of H^{n-l} (coordinates $r > 0$ and $x_{l+2}, \dots, x_n$). The removed set $r = 0$ is $\mathbb{R}^{n-l-1}$. So $\mathbb{R}^n \setminus \mathbb{R}^{n-l-1} \simeq S^l \times H^{n-l}$, and with $m = n - l - 1$ this is (8.316).

The cases that matter in four dimensions:

m	removed	result	symmetric instantons become
0	a point	$S^3 \times H^1 = S^3 \times \mathbb{R}$	$\mathbb{R} \ni \tau = \ln\lvert x\rvert$: instantons as tunnelling in "radial time"
1	a line	$S^2 \times H^2$	SO(3)-invariant instantons $\to$ abelian Higgs vortices on H^2
2	a plane	$S^1 \times H^3$	SO(2)-invariant instantons $\to$ hyperbolic monopoles
3	a 3-space	$S^0 \times H^4$	two copies of H^4: nothing is gained

The case $m = 1$ is Nash's (8.175), $\mathbb{R}^4 \setminus \mathbb{R}^1 \simeq S^2 \times H^2$, relevant for SO(3)-symmetric instantons. Here the reduction goes from four dimensions to two. Nash notes (citing Jaffe and Taubes) that SU(2) Yang–Mills reduced from S^4 to H^2 in this way is an *abelian* Yang–Mills–Higgs system with a complex Higgs field and a fixed value of the coupling λ.

Physical picture: Witten's ansatz: instantons as vortices

Why abelian? An SO(3)-invariant SU(2) field on $S^2 \times H^2$ can only use, at each point of H^2, the structures on S^2 that commute with rotations. What survives is a U(1) gauge field a on H^2 (the rotations about the radial direction in isospin space) and a complex scalar χ (the two components of the gauge field tangent to S^2, combined into one complex number). The self-duality equations become the first-order **vortex equations**

$$\bar\partial_a \chi = 0, \qquad \star f_a = \tfrac{1}{2}\big(1 - |\chi|^2\big)$$

(in a suitable normalisation) on H^2. They are the Bogomolny equations of the abelian Higgs model at the special coupling where vortices do not interact. A charge-k instanton becomes k vortices, located at the zeros of χ. This is Witten's 1977 construction of multi-instantons, and the dimension count works: k zeros in H^2 give $2k$ parameters.

Reading Nash (pp. 245–250, (8.156)–(8.175))

Nash shows that the Bogomolny equation is the static version of $F = \star F$ ((8.156)–(8.162)), then uses rotations and the conformal equivalence $\mathbb{R}^4 \setminus \mathbb{R}^2 \simeq S^1 \times H^3$ ((8.163)–(8.165)) to relate axially symmetric instantons to hyperbolic monopoles ((8.166)–(8.168)). He derives $c_2(V) = 2nk$ ((8.169)–(8.171)), introduces the curvature scale R and the limit to $\mathbb{R}^3$ (8.172), discusses zeros of ϕ, and ends with the general equivalence (8.173)–(8.175). We have supplied the component calculations, an explicit isometry T, the computation of the instanton number from the Bogomolny bound, and pictures. Two corrections: our (8.172)

has $\sqrt{f_R}$ in the numerator, and "$|\phi|_{\text{as}} = n$" is in Nash's eigenvalue normalisation (it is $2n$ in ours), which is what makes $c_2 = |\phi|_{\text{as}} k = 2nk$ come out.

Recap

- Dimensional reduction: x^0-independent self-dual fields on $\mathbb{R}^4$ are Bogomolny monopoles on $\mathbb{R}^3$, with $\phi = \hat{A}_0$. Their action per unit time is the monopole energy, so the total action is infinite.
- Self-duality and the action are conformally invariant in four dimensions. With it, $\mathbb{R}^4 \setminus \mathbb{R}^2 \simeq S^1 \times H^3$, and rotation-invariant instantons are monopoles on H^3, now with finite action.
- Over the axis the rotation acts on fibres with weights $\pm n$; this quantises $|\phi|_{\text{as}}$, splits $V|_{\text{axis}} = L \oplus L^{-1}$ with $c_1(L) = k$, and gives $c_2(V) = 2nk$.
- Flat monopoles are the $R \to \infty$ limit of hyperbolic ones. A monopole "is" at the zeros of ϕ, whose signed number is k. The unit hyperbolic monopole is a ring of instanton of radius 2.
- In general $\mathbb{R}^n \setminus \mathbb{R}^m \simeq S^{n-m-1} \times H^{m+1}$; SO(3)-symmetric instantons are vortices on H^2.

8.15 Monopole moduli

We found one explicit monopole, the Prasad–Sommerfield solution with $k = 1$. It sits at the origin. Translating it gives more solutions, so solutions come in families. For instantons (part 8b) we asked how big the family is, and the answer $\dim \mathcal{M}_k = 8k - 3$ came from an index theorem. Nash now asks the same question for monopoles:

> *How many parameters does a charge-k solution of the Bogomolny equation have? What are these parameters physically, and what does the space of all solutions look like as a manifold?*

The answers, which this section derives:

- The **monopole moduli space** $\mathcal{M}'_k$ has dimension $4k$. The count comes from a linearised Bogomolny equation, an elliptic complex, and an index computation, exactly as for instantons.
- Physically, $4k = k \times (3 + 1)$: each unit monopole has three position coordinates and one internal phase.
- Removing the overall phase gives $\mathcal{M}_k$ of dimension $4k - 1$; also fixing the centre gives $\mathcal{M}^0_k$ of dimension $4k - 4$.
- Topologically $\mathcal{M}'_k$ is the space of rational functions of degree k in one complex variable that vanish at infinity (Donaldson).

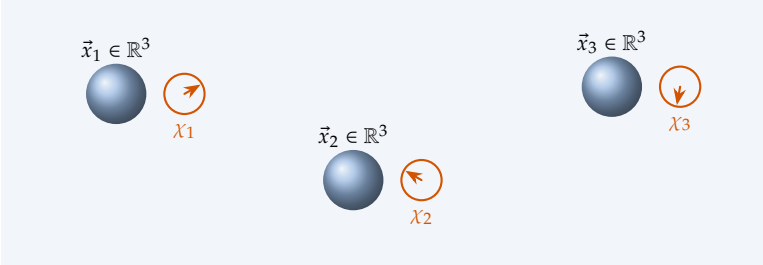
8.15.1 Counting parameters: why one expects $4k$

Before any analysis, here is the physicist's count, due to E. Weinberg. Suppose the k units of charge are far apart. Each then looks like a Prasad–Sommerfield monopole, and each can be moved independently. That gives $3k$ position parameters.

There is one more parameter per monopole. Take a single PS monopole and apply the gauge transformation

$$g_\chi = \exp\Big(\frac{\chi}{C}\,\phi\Big), \qquad \chi \in \mathbb{R}. \tag{8.317}$$

It commutes with ϕ, so ϕ is unchanged, and it changes A only by $g_\chi^{-1}\mathrm{d}g_\chi$, which is non-zero where $|\phi|$ is not constant. At infinity, g_χ is a rotation by the angle χ about the Higgs direction: it does *not* tend to the identity. Physically distinct configurations are those that differ by gauge transformations that are *not* the identity at infinity (Section 8.15.6 explains why). So χ is a genuine parameter, an internal **phase** of the monopole. When one monopole is alone, its phase can be absorbed into a global rotation; when there are several, their *relative* phases are physical. Total: $3k + k = 4k$.



k well-separated unit monopoles: 3 coordinates of position + 1 phase each $\Rightarrow \dim \mathcal{M}'_k = 4k$

Figure 8.70: Weinberg's parameter count. Far apart, a charge-k monopole is k Prasad–Sommerfield lumps (blue). Each has a position in $\mathbb{R}^3$ and an internal phase χ_i (orange dial), the angle of a gauge rotation about its Higgs direction. The index theorem below confirms that no parameters are lost or gained when the lumps come together.

This count is heuristic. When monopoles overlap they are no longer independent lumps. The index theorem makes the count rigorous: the dimension of the solution space cannot change as long as we deform continuously, so the value $4k$ for widely separated monopoles holds everywhere.

8.15.2 Linearising the Bogomolny equation

We follow the method of part 8b. Let (A, ϕ) solve the Bogomolny equation

$$F(A) = -\star\, \mathrm{d}_A\phi \tag{8.318}$$

(our sign, see (8.303); Nash's (8.176)–(8.179) use the opposite sign throughout, which changes a few signs below and nothing else). Deform it along a one-parameter family $(A + ta,\ \phi + t\eta)$, where a is an $\mathfrak{su}(2)$-valued 1-form and η an $\mathfrak{su}(2)$-valued function. This is Nash's (8.176).

Worked example 8.136: Expanding both sides in t

The curvature:

$$\begin{aligned} F(A+ta) &\overset{(1)}{=} \mathrm{d}(A+ta) + (A+ta)\wedge(A+ta) \\ &\overset{(2)}{=} \mathrm{d}A + A\wedge A + t\big(\mathrm{d}a + A\wedge a + a\wedge A\big) + t^2 a\wedge a \\ &\overset{(3)}{=} F(A) + t\,\mathrm{d}_A a + t^2 a\wedge a. \end{aligned}$$

(1) Definition $F = \mathrm{d}A + A\wedge A$.
(2) Multiply out and sort by powers of t.

(3) $\mathrm{d}_A a = \mathrm{d}a + A\wedge a + a\wedge A$ is the covariant derivative of an adjoint 1-form (Section 8.13.2).

The covariant derivative of the Higgs field:

$$\begin{aligned}\mathrm{d}_{A+ta}(\phi+t\eta) &\overset{(1)}{=} \mathrm{d}(\phi+t\eta) + [A+ta, \phi+t\eta]\\ &\overset{(2)}{=} \mathrm{d}_A\phi + t\big(\mathrm{d}_A\eta + [a,\phi]\big) + t^2[a,\eta].\end{aligned}$$

(1) Definition (8.267).
(2) Bilinearity of the bracket; $\mathrm{d}\phi + [A,\phi] = \mathrm{d}_A\phi$ and $\mathrm{d}\eta + [A,\eta] = \mathrm{d}_A\eta$.

Insert both into (8.318). The t^0 terms cancel because (A,ϕ) is a solution. The family is a family of solutions to first order in t iff the t^1 terms agree:

$$\mathrm{d}_A a = -\star\big(\mathrm{d}_A\eta + [a,\phi]\big). \tag{8.319}$$

This is Nash's (8.177) with our sign. In components ($\star$ of a 2-form gives $\frac{1}{2}\varepsilon_{ijk}(\cdot)_{jk}$, and $(\mathrm{d}_A a)_{jk} = D_j a_k - D_k a_j$),

$$\varepsilon_{ijk}D_j a_k + D_i\eta + [a_i,\phi] = 0, \qquad i = 1,2,3. \tag{8.320}$$

Infinitesimal gauge transformations

Some solutions of (8.319) are fake: they only move (A,ϕ) along its gauge orbit. Let $g = \exp(tf)$ with f an $\mathfrak{su}(2)$-valued function. To first order, $g = 1 + tf$ and $g^{-1} = 1 - tf$, so

$$\begin{aligned}g^{-1}Ag + g^{-1}\mathrm{d}g &\overset{(1)}{=} (1-tf)A(1+tf) + (1-tf)\,t\,\mathrm{d}f \overset{(2)}{=} A + t\big(\mathrm{d}f + [A,f]\big) + O(t^2)\\ &= A + t\,\mathrm{d}_A f + O(t^2),\\ g^{-1}\phi\, g &\overset{(3)}{=} \phi + t(\phi f - f\phi) + O(t^2) = \phi + t\,[\phi,f] + O(t^2).\end{aligned}$$

(1) The gauge transformation rule of Section 8.12.1.
(2) Multiply out: $-tfA + tAf = t[A,f]$.
(3) Same for the adjoint scalar.

So the infinitesimal gauge transformation generated by f is

$$(a,\eta) = \big(\mathrm{d}_A f,\ [\phi,f]\big), \tag{8.321}$$

Nash's (8.178).

Watch out: The sign of the Higgs part

Nash writes $(\mathrm{d}_A f, [f,\phi])$ in (8.178) and (8.179). With the gauge rule $\phi \mapsto g^{-1}\phi g$ that he (and we) use, the correct term is $[\phi,f]$. If we used $[f,\phi]$, the pair would not be tangent to the gauge orbit and the check $D_1 D_0 = 0$ below would fail. The slip is harmless for dimension counting.

8.15.3 The deformation complex

Package (8.321) and (8.319) as two operators, Nash's (8.179):

$$\begin{aligned}&D_0 : \Omega^0(\mathbb{R}^3, \mathrm{ad}\,P) \to \Omega^1(\mathbb{R}^3,\mathrm{ad}\,P)\oplus\Omega^0(\mathbb{R}^3,\mathrm{ad}\,P), \qquad f \mapsto \big(\mathrm{d}_A f,\ [\phi,f]\big),\\ &D_1 : \Omega^1(\mathbb{R}^3, \mathrm{ad}\,P)\oplus\Omega^0(\mathbb{R}^3,\mathrm{ad}\,P) \to \Omega^2(\mathbb{R}^3,\mathrm{ad}\,P), \quad (a,\eta)\mapsto \mathrm{d}_A a + \star\mathrm{d}_A\eta + \star[a,\phi].\end{aligned} \tag{8.322}$$

Here ad P is the bundle whose sections are $\mathfrak{su}(2)$-valued functions; over $\mathbb{R}^3$ it is trivial, $\mathbb{R}^3 \times \mathfrak{su}(2)$. The operator D_1 is "Bogomolny linearised": $D_1(a,\eta) = 0$ is (8.319).

Worked example 8.137: Gauge directions solve the linearised equation: $D_1 D_0 = 0$

$$\begin{aligned} D_1 D_0 f &\overset{(1)}{=} \mathrm{d}_A \mathrm{d}_A f + \star \mathrm{d}_A [\phi, f] + \star [\mathrm{d}_A f, \phi] \\ &\overset{(2)}{=} [F, f] + \star\big([\mathrm{d}_A \phi, f] + [\phi, \mathrm{d}_A f]\big) + \star [\mathrm{d}_A f, \phi] \\ &\overset{(3)}{=} [F, f] + \star [\mathrm{d}_A \phi, f] \overset{(4)}{=} \big[F + \star \mathrm{d}_A \phi,\ f\big] \overset{(5)}{=} 0. \end{aligned}$$

(1) Definitions (8.322).
(2) $\mathrm{d}_A \mathrm{d}_A f = [F, f]$ (Example 8.119), and the Leibniz rule for the covariant derivative of a bracket, $\mathrm{d}_A[\phi, f] = [\mathrm{d}_A \phi, f] + [\phi, \mathrm{d}_A f]$.
(3) $[\phi, \mathrm{d}_A f] + [\mathrm{d}_A f, \phi] = 0$.
(4) $\star$ acts only on the form part.
(5) The Bogomolny equation (8.318).

So the image of D_0 lies in the kernel of D_1, but *only* because (A, ϕ) is a solution. That is the typical feature of a deformation complex.

The physical deformations are the solutions of $D_1 = 0$ modulo the gauge directions. This is Nash's (8.180):

$$T_m \mathcal{M}'_k = \frac{\operatorname{Ker} D_1}{\operatorname{Im} D_0}, \tag{8.323}$$

where m is the gauge class of (A, ϕ). We have an **elliptic complex**, Nash's (8.181):

$$0 \longrightarrow \Omega^0(\mathrm{ad}\, P) \xrightarrow{D_0} \Omega^1(\mathrm{ad}\, P) \oplus \Omega^0(\mathrm{ad}\, P) \xrightarrow{D_1} \Omega^2(\mathrm{ad}\, P) \longrightarrow 0, \tag{8.324}$$

with cohomology groups (Nash's (8.182))

$$H^0 = \operatorname{Ker} D_0, \qquad H^1 = \frac{\operatorname{Ker} D_1}{\operatorname{Im} D_0}, \qquad H^2 = \frac{\Omega^2}{\operatorname{Im} D_1}, \qquad \hat{h}_i = \dim H^i. \tag{8.325}$$

Their meaning: H^0 = infinitesimal gauge symmetries of the solution (the stabiliser); H^1 = genuine deformations; H^2 = obstructions, i.e. directions in which the Bogomolny equation cannot be solved by a small deformation.

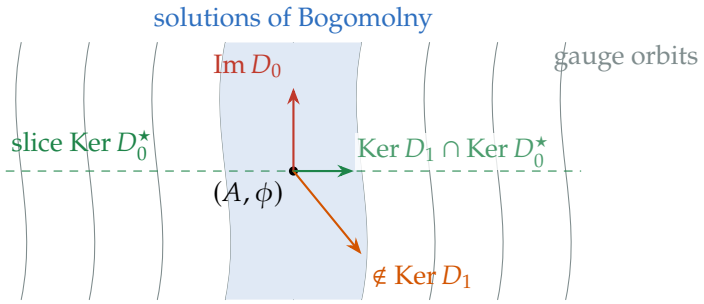


Figure 8.71: The geometry of (8.323), drawn in the infinite-dimensional space $\mathcal{A}_{\mathrm{YMH}}$ of pairs (A, ϕ). Grey curves are gauge orbits; the blue band is the set of solutions, a union of whole orbits. At a solution, tangent vectors along the orbit (Im D_0, red) are trivial; those leaving the band (orange) violate the linearised equation. The physical deformations are best represented by the vectors in Ker D_1 that are orthogonal to the orbit (green): this is the slice condition $D_0^\star(a, \eta) = 0$.

Working definition 8.43 (deformation complex, operationally)

Given a solution, (i) write the linearised equation $D_1 = 0$ [the part of the Bogomolny equation linear in the perturbation (a,η) of (A,ϕ)]; (ii) write the infinitesimal symmetries D_0 [the change of (A,ϕ) under an infinitesimal gauge transformation f]; (iii) check $D_1 D_0 = 0$. The number of parameters is $\hat{h}_1 = \dim(\operatorname{Ker} D_1/\operatorname{Im} D_0)$. To compute it, pick the representative orthogonal to the orbit [L^2-orthogonal to every $D_0 f$], $D_0^\star(a,\eta) = 0$, and count solutions of the *pair* of equations $D_1(a,\eta) = 0$, $D_0^\star(a,\eta) = 0$.

Counting components at a point

The complex (8.324) is elliptic in the usual sense: its symbol sequence is exact. A quick sanity check is that the fibre dimensions have alternating sum zero:

$$\underbrace{3}_{\Lambda^0\otimes\mathfrak{su}(2)} - \underbrace{(3\cdot 3+3)}_{(\Lambda^1\oplus\Lambda^0)\otimes\mathfrak{su}(2)} + \underbrace{3\cdot 3}_{\Lambda^2\otimes\mathfrak{su}(2)} = 3-12+9 = 0.$$

Rolling the complex up gives an operator between bundles of the same rank 12, which we now write down explicitly.

8.15.4 Rolling up: a quaternionic Dirac operator

We need the formal adjoint of D_0. Two facts suffice. First, $\mathrm{d}_A^\star a = -D_i a_i$ on 1-forms (the flat-space formula $\mathrm{d}^\star a = -\partial_i a_i$, made covariant). Second, the map $\mathrm{ad}_\phi = [\phi,\cdot\,]$ is antisymmetric for the inner product (8.265):

$$\big([\phi,X],Y\big) \overset{(1)}{=} -2\operatorname{tr}(\phi XY - X\phi Y) \overset{(2)}{=} -2\operatorname{tr}(XY\phi - X\phi Y) = -2\operatorname{tr}\big(X[Y,\phi]\big) = -\big(X,[\phi,Y]\big).$$

(1) Definition $(X,Y) = -2\operatorname{tr}(XY)$.
(2) Cyclicity of the trace, $\operatorname{tr}(\phi XY) = \operatorname{tr}(XY\phi)$.

(In vector language, $[\phi,X] \leftrightarrow \vec{\phi}\times\vec{X}$ and $(\vec{\phi}\times\vec{X})\cdot\vec{Y} = -\vec{X}\cdot(\vec{\phi}\times\vec{Y})$.) Hence

$$D_0^\star(a,\eta) = \mathrm{d}_A^\star a - [\phi,\eta] = -\big(D_i a_i + [\phi,\eta]\big). \tag{8.326}$$

The **background gauge condition** $D_0^\star(a,\eta) = 0$ is the monopole version of the Coulomb gauge $\mathrm{d}_A^\star a = 0$ of part 8b.

Now combine the four real equations: the gauge condition and the three components (8.320). Collect the unknowns into one quaternion-valued field. Let e_1, e_2, e_3 be the quaternion units $\mathbf{i},\mathbf{j},\mathbf{k}$, with

$$e_i e_j = -\delta_{ij} + \varepsilon_{ijk} e_k, \tag{8.327}$$

and set

$$q = \eta + a_1 e_1 + a_2 e_2 + a_3 e_3 \;\in\; \mathbb{H}\otimes\mathfrak{su}(2) \cong \mathbb{R}^{12}. \tag{8.328}$$

Worked example 8.138: The deformation equations are one quaternionic Dirac equation

Define

$$\mathcal{D}q = \sum_i e_i D_i q + [q,\phi]. \tag{8.329}$$

Expand, using (8.327):

$$\begin{aligned}\mathcal{D}q &\overset{(1)}{=} \sum_i e_i D_i\eta + \sum_{i,k} e_i e_k D_i a_k + [\eta,\phi] + \sum_k [a_k,\phi]e_k \\ &\overset{(2)}{=} \sum_l e_l D_l\eta - \sum_i D_i a_i + \sum_{i,k,l} \varepsilon_{ikl} e_l D_i a_k + [\eta,\phi] + \sum_l [a_l,\phi]e_l \\ &\overset{(3)}{=} -\big(D_i a_i + [\phi,\eta]\big) + \sum_l e_l\big(\varepsilon_{lik} D_i a_k + D_l\eta + [a_l,\phi]\big).\end{aligned}$$

(1) Insert (8.328); D_i acts on the $\mathfrak{su}(2)$ factor and commutes with the constant units e_k.
(2) $e_i e_k = -\delta_{ik} + \varepsilon_{ikl} e_l$.
(3) Sort into the real part and the e_l parts; $\varepsilon_{ikl} = \varepsilon_{lik}$ and $[\eta,\phi] = -[\phi,\eta]$.

The real part is $D_0^\star(a,\eta)$ from (8.326); the e_l part is the l-th component (8.320) of $D_1(a,\eta)$ (after $\star$). So

$$\mathcal{D}q = 0 \iff D_1(a,\eta) = 0 \text{ and } D_0^\star(a,\eta) = 0.$$

$\mathcal{D}$ is the rolled-up operator $D_0^\star + D_1$, and its kernel is H^1. Since $e_i^2 = -1$, the leading part $\sum e_i D_i$ squares to $-\nabla^2$ (plus lower order): it is a Dirac operator in three dimensions, with the quaternion units playing the role of the Pauli matrices $-i\sigma_i$.

Worked example 8.139: The Weitzenböck formula: $\mathcal{D}\mathcal{D}^\dagger = -D^2 - \mathrm{ad}_\phi^2$

Left multiplication by a unit imaginary quaternion is antisymmetric, and so is D_i (integration by parts); they commute, so $e_i D_i$ is symmetric. The map $q \mapsto [q,\phi]$ is antisymmetric. Hence $\mathcal{D}^\dagger q = \sum_i e_i D_i q - [q,\phi]$. Now compute:

$$\begin{aligned}\mathcal{D}\mathcal{D}^\dagger q &\overset{(1)}{=} \sum_{i,j} e_i e_j D_i D_j q - \sum_i e_i\big([D_i q,\phi] + [q, D_i\phi]\big) + \sum_j e_j [D_j q,\phi] - \big[[q,\phi],\phi\big] \\ &\overset{(2)}{=} -\sum_i D_i D_i q + \tfrac{1}{2}\sum_{i,j,k} \varepsilon_{ijk} e_k [F_{ij}, q] - \sum_i e_i [q, D_i\phi] - \big[[q,\phi],\phi\big] \\ &\overset{(3)}{=} -D^2 q + \sum_k e_k [B_k, q] + \sum_k e_k [D_k\phi, q] - \big[[q,\phi],\phi\big] \\ &\overset{(4)}{=} -D^2 q - \mathrm{ad}_\phi^2\, q.\end{aligned}$$

(1) Apply $\mathcal{D}$ to $\mathcal{D}^\dagger q$, using the Leibniz rule $D_i[q,\phi] = [D_i q,\phi] + [q, D_i\phi]$.
(2) Split $e_i e_j = -\delta_{ij} + \varepsilon_{ijk} e_k$; the antisymmetric part of $D_i D_j$ is $\frac{1}{2}[D_i, D_j] = \frac{1}{2}[F_{ij}, \cdot\,]$; the two $[D_i q,\phi]$ terms cancel.
(3) $B_k = \frac{1}{2}\varepsilon_{ijk} F_{ij}$, and $-[q, D_k\phi] = [D_k\phi, q]$.
(4) The Bogomolny equation $B_k = -D_k\phi$ cancels the middle terms; and $[[q,\phi],\phi] = [\phi,[\phi,q]] = \mathrm{ad}_\phi^2\, q$.

Both terms on the right are non-negative operators: $-D^2 = D_i^\dagger D_i$ and $-\mathrm{ad}_\phi^2 = \mathrm{ad}_\phi^\dagger\,\mathrm{ad}_\phi$. So if $\mathcal{D}^\dagger q = 0$ with q square-integrable,

$$0 = \langle q, \mathcal{D}\mathcal{D}^\dagger q\rangle = \sum_i \|D_i q\|^2 + \big\|[\phi,q]\big\|^2 \;\Rightarrow\; D_i q = 0 \;\Rightarrow\; |q| = \text{const} \;\Rightarrow\; q = 0.$$

Therefore $\operatorname{Ker}\mathcal{D}^\dagger = 0$.

Now read off the cohomology. $\operatorname{Ker} \mathcal{D}^\dagger = \operatorname{Ker}(D_0 + D_1^\star)$ consists of pairs $(f, \omega) \in \Omega^0 \oplus \Omega^2$ with $D_0 f + D_1^\star \omega = 0$; pairing with $D_0 f$ and using $D_1 D_0 = 0$ shows $D_0 f = 0$ and $D_1^\star \omega = 0$ separately. So $\operatorname{Ker} \mathcal{D}^\dagger \cong H^0 \oplus H^2$, and the Weitzenböck argument gives

$$\hat{h}_0 = \hat{h}_2 = 0.$$

Nash's (8.183) is then

$$\begin{aligned} \operatorname{index}(D_0 + D_1^\star) &\overset{(1)}{=} \dim \operatorname{Ker}(D_0 + D_1^\star) - \dim \operatorname{Ker}(D_0^\star + D_1) \\ &\overset{(2)}{=} \hat{h}_0 - \hat{h}_1 + \hat{h}_2 \overset{(3)}{=} -\hat{h}_1 = -\dim \mathcal{M}'_k. \end{aligned} \tag{8.330}$$

(1) Definition of the index of an operator: kernel minus cokernel, and the cokernel of $D_0 + D_1^\star$ is the kernel of its adjoint.

(2) $\operatorname{Ker}(D_0 + D_1^\star) = H^0 \oplus H^2$ as just shown, and $\operatorname{Ker}(D_0^\star + D_1) = \operatorname{Ker} \mathcal{D} = H^1$.

(3) $\hat{h}_0 = \hat{h}_2 = 0$ by the Weitzenböck argument; $\hat{h}_1 = \dim T_m \mathcal{M}'_k$.

Watch out: Which kernel vanishes in (8.184)

Nash writes the vanishing theorem as $\operatorname{Ker}(D_0^\star + D_1) = \phi$ (the empty set; he means 0). But $\operatorname{Ker}(D_0^\star + D_1) = H^1$ is the tangent space, which is $4k$-dimensional, not zero. What the Weitzenböck argument kills is the kernel of the *other* operator, $\operatorname{Ker}(D_0 + D_1^\star) = H^0 \oplus H^2$. With that correction, $\hat{h}_0 = \hat{h}_2 = 0$ follows as he says.

Key idea: Dimension of the monopole moduli space (Nash's (8.185))

Taubes (following Callias and E. Weinberg) showed $\operatorname{index}(D_0 + D_1^\star) = -4k$. With $\hat{h}_0 = \hat{h}_2 = 0$,

$$\dim \mathcal{M}'_k = \hat{h}_1 = 4k.$$

Why $4k$, at the physicist's level. The index is unchanged under continuous deformations that keep the operator Fredholm. Deform a charge-k solution into k widely separated unit monopoles. Each has its own zero modes, which we count explicitly below (four of them), and far apart they do not interfere. So the index is 4 per unit of charge. Callias' index theorem for Dirac operators coupled to a Higgs field on odd-dimensional spaces turns this into an exact formula: the index is an integral over the sphere at infinity of an expression in $\hat{\phi}$, proportional to $\deg \hat{\phi} = k$.

Watch out: Why $\mathbb{R}^3$ is harder than S^4

On a compact manifold elliptic operators are automatically Fredholm. On $\mathbb{R}^3$ they need not be: an operator can have continuous spectrum reaching down to 0. Here the term $-\operatorname{ad}_\phi^2$ in Example 8.139 gives the components of q that do *not* commute with ϕ a mass $\approx C$ at infinity, so they decay exponentially. The component along $\hat{\phi}$ (the unbroken U(1), the photon) is massless and decays only like a power. This is why Nash says the index calculation "requires considerable extra analysis": one has to choose function spaces (Taubes uses weighted Sobolev spaces) so that D_0 and D_1 are Fredholm and the cohomology can be built from square-integrable sections. The zero modes that result decay like $1/r^2$, which is square integrable in three dimensions.

8.15.5 The four zero modes of the Prasad–Sommerfield monopole

Let us see the four parameters of a unit monopole as explicit solutions of $\mathcal{D} q = 0$.

Worked example 8.140: Translation zero modes

Translating a solution in the direction x^j gives the tangent vector $(\partial_j A, \partial_j \phi)$. It solves the linearised Bogomolny equation (translations map solutions to solutions), but it is not orthogonal to the gauge orbit. Add the infinitesimal gauge transformation with $f = -A_j$:

$$a_k \overset{(1)}{=} \partial_j A_k - D_k A_j \overset{(2)}{=} \partial_j A_k - \partial_k A_j - [A_k, A_j] \overset{(3)}{=} F_{jk},$$
$$\eta \overset{(4)}{=} \partial_j \phi + [\phi, -A_j] \overset{(5)}{=} D_j \phi.$$

(1) Add $\mathrm{d}_A f$ with $f = -A_j$, from (8.321).
(2) $D_k A_j = \partial_k A_j + [A_k, A_j]$ (the adjoint covariant derivative of the function A_j).
(3) $F_{jk} = \partial_j A_k - \partial_k A_j + [A_j, A_k]$.
(4) Add $[\phi, f]$ with $f = -A_j$.
(5) $[\phi, -A_j] = [A_j, \phi]$ and $D_j \phi = \partial_j \phi + [A_j, \phi]$.

So the translation mode is $(a, \eta) = (\iota_{\partial_j} F,\ D_j \phi)$: gauge-covariant, as it should be. It still solves $D_1 = 0$ (we added a gauge direction, and $D_1 D_0 = 0$). Does it satisfy the gauge condition?

$$D_k a_k + [\phi, \eta] \overset{(1)}{=} D_k F_{jk} + [\phi, D_j \phi] \overset{(2)}{=} [D_j \phi, \phi] + [\phi, D_j \phi] = 0.$$

(1) Insert the mode.
(2) The first field equation (8.282), $\mathrm{d}_A^\star F = [\mathrm{d}_A \phi, \phi]$, in components: $(\mathrm{d}_A^\star F)_j = -D_k F_{kj} = D_k F_{jk}$.

So all three translation modes lie in the slice. (We also verified (8.320) for them numerically on the explicit PS solution.) Their length:

$$\left\| (\iota_{\partial_j} F, D_j \phi) \right\|^2 \overset{(1)}{=} \int \Big(\sum_k |F_{jk}|^2 + |D_j \phi|^2 \Big) \mathrm{d}^3 x \overset{(2)}{=} \frac{1}{3} \int \big(2|B|^2 + |D\phi|^2 \big) \mathrm{d}^3 x$$
$$\overset{(3)}{=} \int |B|^2 \mathrm{d}^3 x \overset{(4)}{=} 4\pi C.$$

(1) The norm (8.266) in components.
(2) Summed over j, $\sum_{j,k} |F_{jk}|^2 = 2|B|^2$ and $\sum_j |D_j \phi|^2 = |D\phi|^2$; for the spherically symmetric PS monopole each j contributes one third.
(3) $|B| = |D\phi|$ on a Bogomolny solution.
(4) $E = \frac{1}{2} \int (|B|^2 + |D\phi|^2) = \int |B|^2 = 4\pi C$.

The squared length of a translation mode is the *mass* of the monopole. We shall see in Section 8.16 that this is no accident.

Worked example 8.141: The phase zero mode

Now apply D_0 to $f = \phi$, the generator of (8.317):

$$D_0 \phi = \big(\mathrm{d}_A \phi, [\phi, \phi] \big) = (\mathrm{d}_A \phi, 0).$$

Check that $(a, \eta) = (\mathrm{d}_A \phi, 0)$ solves both equations:

$$\varepsilon_{ijk} D_j D_k \phi + [D_i \phi, \phi] \overset{(1)}{=} \tfrac{1}{2} \varepsilon_{ijk} [F_{jk}, \phi] + [D_i \phi, \phi] \overset{(2)}{=} [B_i + D_i \phi, \phi] \overset{(3)}{=} 0,$$
$$D_k D_k \phi + [\phi, 0] \overset{(4)}{=} 0.$$

(1) Antisymmetrise: $\varepsilon_{ijk}D_jD_k = \frac{1}{2}\varepsilon_{ijk}[D_j, D_k] = \frac{1}{2}\varepsilon_{ijk}[F_{jk}, \cdot\,]$.
(2) $B_i = \frac{1}{2}\varepsilon_{ijk}F_{jk}$.
(3) Bogomolny, $B_i = -D_i\phi$.
(4) The second field equation (8.282), $\mathrm{d}_A^\star \mathrm{d}_A\phi = -D_kD_k\phi = 0$.

Its length is $\|\mathrm{d}_A\phi\|^2 = \int |D\phi|^2 = 4\pi C$, finite because $|D\phi| \sim r^{-2}$.

The phase mode is peculiar. It has the form $D_0 f$, i.e. it looks like a gauge direction, yet it is a genuine element of H^1. The resolution is that $f = \phi$ is not square integrable ($|\phi| \to C$), so $\phi \notin \Omega^0$ in the L^2 sense of the complex, and $D_0\phi$ is not in the image of D_0 on L^2 functions. In the language of the previous parts: ϕ generates a gauge transformation that is not the identity at infinity, and such transformations are not quotiented out in $\mathcal{M}'_k$.

So for $k = 1$ we have found $3 + 1 = 4 = 4k$ independent zero modes, as the index theorem requires.

8.15.6 From $4k$ to $4k-1$ to $4k-4$

Why keep gauge transformations that are non-trivial at infinity? The situation is exactly that of the framed instanton moduli space of part 8b. The group $\mathcal{G}$ of gauge transformations tending to 1 at infinity acts freely, and dividing by it gives the **framed** (or extended) moduli space $\mathcal{M}'_k$. The leftover global rotations about the Higgs direction at infinity form a circle $S^1 = \{\exp(\chi\phi/C)\}$, and they act on $\mathcal{M}'_k$ as the phase rotation (8.317). Physically, the corresponding momentum is the *electric charge* (Section 8.16.4), so we want to keep this circle when we do dynamics.

Nash's (8.186)–(8.187) record the three spaces:

$$\begin{array}{ccc} S^1 & \to & \mathcal{M}'_k \\ & & \downarrow \\ & & \mathcal{M}_k \end{array} \qquad \mathcal{M}_k = \frac{\mathcal{M}'_k}{S^1}, \qquad \mathcal{M}^0_k = \frac{\mathcal{M}'_k}{S^1 \times \mathbb{R}^3}, \qquad \begin{array}{l} \dim \mathcal{M}'_k = 4k, \\ \dim \mathcal{M}_k = 4k-1, \\ \dim \mathcal{M}^0_k = 4k-4. \end{array} \tag{8.331}$$

Projecting out the one-dimensional subspace spanned by $(\mathrm{d}_A\phi, 0)$ is the passage from $\mathcal{M}'_k$ to $\mathcal{M}_k$. Dividing also by translations (fixing the "centre of mass") gives $\mathcal{M}^0_k$.

Worked example 8.142: $k = 1$: everything explicit

For $k = 1$ there is only one monopole up to translation and phase: the PS monopole, spherically symmetric, with the zero of ϕ at its centre. Its position $\vec{x} \in \mathbb{R}^3$ and phase $\chi \in \mathbb{R}/2\pi\mathbb{Z}$ are all the parameters. So

$$\mathcal{M}'_1 = \mathbb{R}^3 \times S^1, \qquad \mathcal{M}_1 = \mathbb{R}^3, \qquad \mathcal{M}^0_1 = \{\text{point}\}.$$

The phase is periodic with period 2π: $\exp(2\pi\hat{\phi})$, with $\hat{\phi}$ conjugate to $T_3 = \mathrm{diag}(-\frac{i}{2}, \frac{i}{2})$, is $-\mathbb{1}$, which acts trivially in the adjoint representation, where A and ϕ live. Dimensions: $4, 3, 0 = 4k, 4k-1, 4k-4$.

8.15.7 The topology of $\mathcal{M}'_k$: rational maps

Nash quotes a striking theorem of Donaldson.

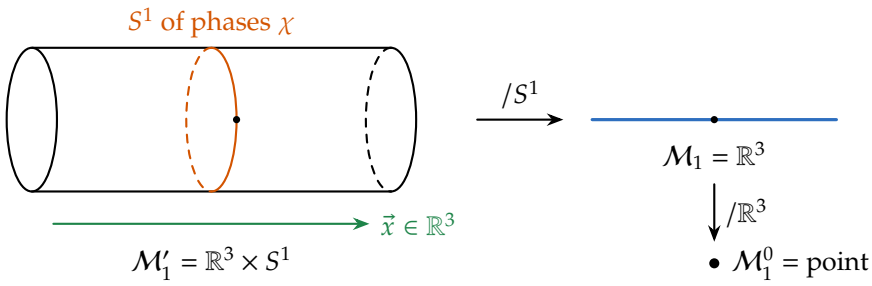


Figure 8.72: The moduli spaces for one monopole (the $\mathbb{R}^3$ drawn as a line). The framed space $\mathcal{M}_1'$ is a cylinder: position times phase. Forgetting the phase gives $\mathcal{M}_1 = \mathbb{R}^3$; forgetting the position as well leaves a single point, the PS monopole at the origin.

Working definition 8.44 (based rational map of degree k)

A **rational map** is a ratio of polynomials $f(z) = p(z)/q(z)$ in one complex variable, viewed as a map $\mathbb{CP}^1 \to \mathbb{CP}^1$ [$\mathbb{CP}^1 = \mathbb{C} \cup \{\infty\}$, a sphere S^2]. It is *based* if $f(\infty) = 0$, which means $\deg p < \deg q$. Its **degree** is $\deg q = k$ when p and q have no common root. Normalise q to be monic [leading coefficient 1]:

$$f(z) = \frac{a_{k-1}z^{k-1} + \cdots + a_1 z + a_0}{z^k + b_{k-1}z^{k-1} + \cdots + b_1 z + b_0}, \qquad a_i, b_i \in \mathbb{C}. \tag{8.332}$$

Operational test: write the fraction in lowest terms and count the poles with multiplicity [a double root of q counts twice]. The count is the degree, and it is also the topological degree of $f : S^2 \to S^2$.

Key idea: Donaldson's theorem (Nash's (8.188))

$\mathcal{M}_k'$ is diffeomorphic to the space Rat_k of based rational maps of degree k.

Parameter count. In (8.332) there are k complex numbers a_i and k complex numbers b_i: $2k$ complex = $4k$ real parameters. The only condition, "no common root", is an open condition, so Rat_k is an open subset of $\mathbb{C}^{2k}$ and has real dimension $4k = \dim \mathcal{M}_k'$. ✓

Where the rational map comes from. Donaldson's construction singles out a direction, say the x^3-axis, and writes $z = x^1 + ix^2$ for the position of a line parallel to it. Along each such line one solves an ordinary differential equation of the form $(D_3 - i\phi)s = 0$ for a section s of the $\mathbb{C}^2$-bundle. Generically the solution that decays as $x^3 \to +\infty$ grows as $x^3 \to -\infty$; the lines along which a solution decays at both ends are "spectral lines", and the scattering data along all lines assemble into a holomorphic function of z, namely f. For well-separated monopoles the spectral lines are the lines through the monopoles, so the poles of f are the positions of the monopoles projected to the z-plane.

Worked example 8.143: One monopole, Nash's (8.189)

For $k = 1$,

$$f(z) = \frac{a_0}{z + b_0}, \qquad a_0 \in \mathbb{C}^\times,\ b_0 \in \mathbb{C}.$$

The coprimality condition says only $a_0 \neq 0$. So $\mathrm{Rat}_1 = \mathbb{C}^\times \times \mathbb{C} \cong (\mathbb{R}_{>0} \times S^1) \times \mathbb{R}^2 \cong \mathbb{R}^3 \times S^1$, which is exactly $\mathcal{M}_1'$ from Example 8.142. The dictionary: the pole $z = -b_0$ is the position $x^1 + ix^2$ of the monopole; $|a_0|$ encodes the height x^3 (with the usual normalisation, $|a_0|$ grows exponentially with x^3, which is why $\mathbb{R}_{>0} \cong \mathbb{R}$ appears); $\arg a_0$ is the phase χ. Nash

notes that $a_0 = 1$, $b_0 = 0$ is the PS monopole at the origin.

Worked example 8.144: Two monopoles and the resultant

For $k = 2$,

$$f(z) = \frac{a_1 z + a_0}{z^2 + b_1 z + b_0}.$$

Numerator and denominator have a common root iff $q(-a_0/a_1) = 0$ (for $a_1 \neq 0$). Clearing the denominator,

$$a_1^2\, q\big(-a_0/a_1\big) \overset{(1)}{=} a_1^2\Big(\frac{a_0^2}{a_1^2} - b_1\frac{a_0}{a_1} + b_0\Big) \overset{(2)}{=} a_0^2 - a_0 a_1 b_1 + a_1^2 b_0 \;=:\; R(f).$$

(1) Evaluate $q(z) = z^2 + b_1 z + b_0$ at the root of the numerator.
(2) Multiply out.

The **resultant** $R(f)$ is also correct when $a_1 = 0$ (then $R = a_0^2$, which vanishes only if the numerator is identically zero). So $\mathrm{Rat}_2 = \{(a_0, a_1, b_0, b_1) \in \mathbb{C}^4 : R \neq 0\}$, the complement of a complex hypersurface in $\mathbb{C}^4$: real dimension $8 = 4k$. *Partial fractions*: if q has distinct roots $\hat{b}_1, \hat{b}_2$,

$$f(z) = \frac{\hat{a}_1}{z - \hat{b}_1} + \frac{\hat{a}_2}{z - \hat{b}_2},$$

the second form in Nash's (8.188). Each term looks like a unit monopole (Example 8.143): this is the approximate superposition of k unit monopoles. When $\hat{b}_1 \to \hat{b}_2$ the partial fraction form breaks down but f does not: coincident monopoles are perfectly regular points of $\mathcal{M}'_k$.

Worked example 8.145: The axially symmetric monopole, Nash's (8.190)

Take $f(z) = 1/z^k$. A rotation by α about the x^3-axis acts on the spectral data by $z \mapsto e^{i\alpha} z$, so $f \mapsto e^{-ik\alpha}/z^k$: the same map up to a phase. A phase can be undone by the S^1 action on $\mathcal{M}'_k$. So the monopole with rational map $1/z^k$ is invariant under rotations about the axis combined with a compensating phase rotation: it is **axially symmetric**. All k poles sit at $z = 0$. For $k = 2$ its energy density is a doughnut (a torus) around the x^3-axis; we will meet it again as the configuration in the middle of 90° scattering.

Finally, the unframed space: $\mathcal{M}_k = \mathcal{M}'_k/S^1$, and on rational maps the circle acts by $f \mapsto e^{i\theta} f$. So $\mathcal{M}_k$ is based rational maps modulo an overall phase of the numerator, as Nash says.

Try it: The topology of Rat_1 and a hint for Rat_2

Show that Rat_1 deformation-retracts onto a circle, so $\pi_1(\mathcal{M}'_1) = \mathbb{Z}$. For Rat_2, show that the map $f \mapsto R(f)/|R(f)| \in S^1$ is well defined and continuous. *Answer:* $\mathbb{C}^\times \times \mathbb{C}$ retracts onto $\{|a_0| = 1\} \times \{0\} = S^1$. For $k = 2$, $R \neq 0$ on Rat_2, so $R/|R|$ is continuous; Segal proved $\pi_1(\mathrm{Rat}_k) = \mathbb{Z}$ for all k, detected by this kind of phase.

Reading Nash (pp. 250–253, (8.176)–(8.190))

Nash linearises the Bogomolny equation ((8.176)–(8.177)), identifies the gauge directions (8.178), defines D_0 and D_1 (8.179), and obtains $T_m\mathcal{M}'_k = \mathrm{Ker}\, D_1/\mathrm{Im}\, D_0$ (8.180) from the elliptic complex (8.181)–(8.182). The index (8.183) is $-4k$ (Taubes) and a Weitzenböck

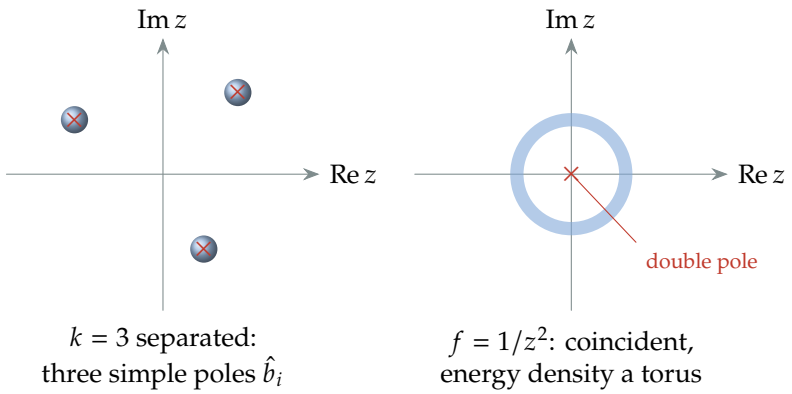


Figure 8.73: Rational maps and monopole positions, seen from above (the x^3-axis points out of the page). Left: for separated monopoles the poles of f (red crosses) sit under the monopoles (blue). Right: for $f = 1/z^2$ both poles coincide at the origin and the energy density is a ring around the axis, not two lumps: the location of the poles says where the monopoles are, not what shape the energy takes.

argument gives $\hat{h}_0 = \hat{h}_2 = 0$ ((8.184), with the kernel corrected as in our pitfall), so $\dim \mathcal{M}'_k = 4k$ (8.185). He explains $4k-1$ (8.186) by the non-L^2 gauge generator ϕ and lists the three spaces (8.187), then quotes Donaldson's rational maps (8.188)–(8.190). We added: the explicit quaternionic form of the operator, the full Weitzenböck computation, the explicit zero modes of the PS monopole and their norms, and the resultant for $k = 2$. Signs: our Bogomolny sign is opposite to Nash's here, and his $[f, \phi]$ should read $[\phi, f]$.

Recap

- Linearising $F = -\star \mathrm{d}_A\phi$ and quotienting by infinitesimal gauge transformations gives an elliptic complex; $T\mathcal{M}'_k = H^1$.
- In quaternionic form the deformation operator is $\mathcal{D}q = e_iD_iq + [q, \phi]$, and $\mathcal{D}\mathcal{D}^\dagger = -D^2 - \mathrm{ad}^2_\phi \geq 0$ forces $H^0 = H^2 = 0$. Taubes' index gives $\dim \mathcal{M}'_k = 4k$.
- For $k = 1$ the four zero modes are three translations $(\iota_j F, D_j\phi)$ and the phase mode $(\mathrm{d}_A\phi, 0)$, each of squared length $4\pi C$.
- $\mathcal{M}'_k \to \mathcal{M}_k = \mathcal{M}'_k/S^1 \to \mathcal{M}^0_k$, of dimensions $4k, 4k-1, 4k-4$; for $k = 1$: $\mathbb{R}^3 \times S^1, \mathbb{R}^3$, point.
- $\mathcal{M}'_k \cong \mathrm{Rat}_k$, based rational maps of degree k; poles $\approx$ monopole positions; $1/z^k$ is the axially symmetric monopole.

8.16 Monopole dynamics: forces and the geodesic approximation

So far every monopole in this part has been static. We found the charge-k solutions and we found that they form a $4k$-dimensional family $\mathcal{M}'_k$. Now we set them in motion. The questions of this section and the next are:

> *Do monopoles exert forces on each other? If two of them are sent towards each other, what comes out? And why does the answer need the geometry of $\mathcal{M}'_k$?*

Nash's answer (pp. 253–256) has three steps, and we follow them.

1. Static BPS monopoles exert *no* net force on each other: the energy $E = 4\pi C k$ does not depend on where they are.
2. Slowly moving monopoles therefore stay close to the bottom of the energy valley, which is $\mathcal{M}'_k$. Their motion is governed by the kinetic energy alone, and kinetic energy is a metric. The motion is a **geodesic** of that metric on $\mathcal{M}'_k$ (Manton).
3. The metric is **hyperkähler**. That property, together with rotational symmetry, fixes the metric for $k = 2$ (Atiyah–Hitchin), and its geodesics show two famous effects: head-on collisions scatter at 90°, and glancing collisions can turn monopoles into dyons.

This section does steps 1 and 2 and the general hyperkähler geometry; the next section does the $k = 2$ case.

8.16.1 The energy is linear in k: no static forces

Recall the Bogomolny bound (Section 8.13.3): with our normalisation, every finite-energy configuration of magnetic charge $k > 0$ obeys $E \geq 4\pi C k$, with equality exactly for solutions of the Bogomolny equation. So *every* point of $\mathcal{M}'_k$ has the same energy,

$$E\big|_{\mathcal{M}'_k} = 4\pi C k, \tag{8.333}$$

Nash's (8.191) (he sets $C = 1$). Two consequences:

- E is linear in k. A charge-k solution weighs exactly as much as k separate unit monopoles. There is no binding energy.
- On the family of k well-separated unit monopoles, E does not depend on the separations. A potential energy that does not depend on position gives no force: $-\partial E/\partial \vec{x}_i = 0$.

This is surprising. Like magnetic charges repel. So something must attract them with exactly the same strength. That something is the Higgs field.

Worked example 8.146: Magnetic repulsion versus Higgs attraction

Put two unit monopoles a large distance d apart. Far from its core, a unit monopole looks like a point source with two long-range fields (Section 8.13.6):

- a Coulomb *magnetic* field along the Higgs direction, $b_i = (B_i, \hat{\phi}) = -\hat{x}_i/r^2$, i.e. a point magnetic charge g with $b = g\,\hat{\vec{x}}/(4\pi r^2)$ (Heaviside–Lorentz units), $|g| = 4\pi$;
- a Coulomb *Higgs* tail, $|\phi| = C - 1/r$, because the Higgs field is massless in the Prasad–Sommerfield limit.

The magnetic interaction energy of two like point poles is the same as that of two like point charges:

$$\begin{aligned} U_{\text{mag}} &\overset{(1)}{=} \int_{\mathbb{R}^3} \vec{b}_1 \cdot \vec{b}_2 \, \mathrm{d}^3x \overset{(2)}{=} \frac{g^2}{16\pi^2} \int_{\mathbb{R}^3} \nabla\frac{1}{r_1} \cdot \nabla\frac{1}{r_2} \, \mathrm{d}^3x \\ &\overset{(3)}{=} \frac{g^2}{16\pi^2} \int_{\mathbb{R}^3} \frac{1}{r_1} \, 4\pi\delta^3(\vec{x} - \vec{x}_2) \, \mathrm{d}^3x \overset{(4)}{=} \frac{g^2}{4\pi d} = \frac{4\pi}{d}. \end{aligned}$$

(1) The field energy is $\frac{1}{2}\int |\vec{b}_1 + \vec{b}_2|^2$; the self-energies do not depend on d, and the cross term is $\int \vec{b}_1 \cdot \vec{b}_2$.
(2) $\vec{b}_j = -\frac{g}{4\pi}\nabla(1/r_j)$ with $r_j = |\vec{x} - \vec{x}_j|$.
(3) Integrate by parts (the boundary term vanishes like $1/R$) and use $-\nabla^2(1/r_2) = 4\pi\delta^3(\vec{x} - \vec{x}_2)$.
(4) Evaluate $1/r_1$ at $\vec{x}_2$, where $r_1 = d$; then $g^2 = 16\pi^2$.

Now the Higgs field. The mass of a monopole is $4\pi C$, and C is the length of the Higgs field at the monopole's location. Monopole 2 sits in the tail of monopole 1, where the Higgs field has length $C - 1/d$. So its mass, and hence the total energy, changes by

$$U_{\text{Higgs}} = 4\pi\left(C - \frac{1}{d}\right) - 4\pi C = -\frac{4\pi}{d}.$$

The total interaction energy is

$$U_{\text{mag}} + U_{\text{Higgs}} = \frac{4\pi}{d} - \frac{4\pi}{d} = 0.$$

In terms of forces: a magnetic repulsion $4\pi/d^2$ and a scalar attraction $4\pi/d^2$, which cancel exactly.

Worked example 8.147: A monopole and an antimonopole attract

Replace monopole 2 by an antimonopole (charge $k = -1$, obtained for instance by reflecting the hedgehog). Its magnetic charge is $-g$, so the magnetic force becomes an attraction of strength $4\pi/d^2$. The Higgs tail is still $|\phi| = C - 1/r$: both objects *lower* the length of the Higgs field near them, since $|\phi| = 0$ in both cores. So the scalar force is still attractive. Total:

$$F = -\frac{4\pi}{d^2} - \frac{4\pi}{d^2} = -\frac{8\pi}{d^2} \qquad \text{(attractive)}.$$

There is no static monopole–antimonopole solution of the Bogomolny equation (it would have $k = 0$ and hence $E = 0$). Whether some *non-minimal* static solution exists in the $k = 0$ sector is a question of critical point theory, and we return to it in Section 8.18.

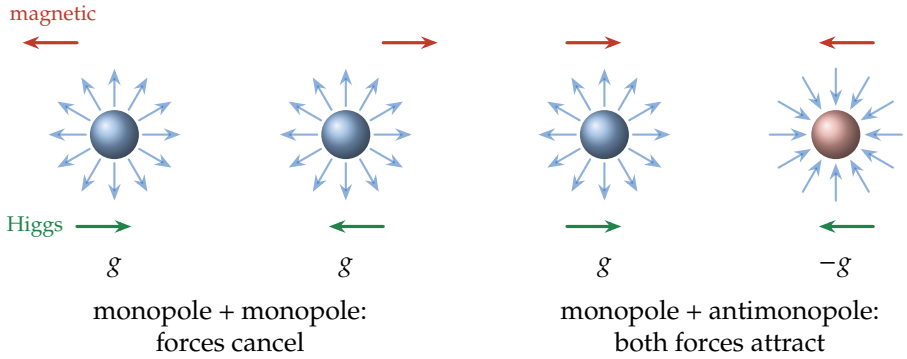


Figure 8.74: Forces between static BPS monopoles at large separation. Red arrows: the magnetic Coulomb force on each object; green arrows: the force from exchange of the massless Higgs field. For two monopoles (left) they have equal size and opposite direction, which is why the energy $4\pi Ck$ does not depend on the positions. For a monopole and an antimonopole (right) both attract.

Physical picture: Why the particle picture breaks down at short distance

The cancellation above is a statement about far-apart monopoles, each seen as a point. When they come within a distance of order $1/C$ (the size of the core, fixed by the W mass C) the cores overlap and there is no longer any sense in which there are "two objects". Nash stresses this: for $k \geq 2$ there are solutions whose charge is localised in one lump, for example the axially symmetric doughnut $f = 1/z^k$ of Example 8.145. Scattering is the process in which k separate lumps merge into such a configuration and then separate again. To describe it we need more than a force law; we need dynamics on the whole of $\mathcal{M}'_k$, including its core region.

8.16.2 Slow motion in a valley: the geodesic approximation

Here is the key idea, due to Manton, in a finite-dimensional setting. Think of a particle moving on a configuration space Q with Lagrangian

$$L = \tfrac{1}{2} G_{IJ}(X)\dot{X}^I \dot{X}^J - V(X),$$

where G_{IJ} is a metric on Q (the "mass matrix") and $V \geq V_0$ is a potential. Suppose the minimum of V is not a point but a whole submanifold $N \subset Q$, a flat-bottomed valley, and $V = V_0$ everywhere on N. Start the particle on N with a small velocity tangent to N. Then:

- it cannot climb the walls of the valley far, since it has little kinetic energy;
- along the bottom there is no force, since V is constant there;
- so it moves along N, driven by nothing except its own inertia.

"Moving by inertia" on a curved space means following a geodesic. We now make this precise.

Worked example 8.148: Kinetic energy on N gives the geodesic equation

Let z^r ($r = 1, \dots, \dim N$) be coordinates on N and $X(z)$ the embedding. A path $z(t)$ on N has velocity $\dot{X}^I = \partial_r X^I \, \dot{z}^r$, so the Lagrangian restricted to N is

$$L_N = \tfrac{1}{2} g_{rs}(z)\dot{z}^r \dot{z}^s - V_0, \qquad g_{rs} = G_{IJ}\, \partial_r X^I \, \partial_s X^J,$$

the **induced metric**. The Euler–Lagrange equation is

$$0 \overset{(1)}{=} \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L_N}{\partial \dot{z}^r} - \frac{\partial L_N}{\partial z^r} \overset{(2)}{=} \frac{\mathrm{d}}{\mathrm{d}t}\big(g_{rs}\dot{z}^s\big) - \tfrac{1}{2}\,\partial_r g_{st}\,\dot{z}^s\dot{z}^t$$
$$\overset{(3)}{=} g_{rs}\ddot{z}^s + \partial_t g_{rs}\,\dot{z}^t\dot{z}^s - \tfrac{1}{2}\,\partial_r g_{st}\,\dot{z}^s\dot{z}^t \overset{(4)}{=} g_{rs}\ddot{z}^s + \tfrac{1}{2}\big(\partial_t g_{rs} + \partial_s g_{rt} - \partial_r g_{st}\big)\dot{z}^s\dot{z}^t,$$

(1) Euler–Lagrange equation for the coordinate z^r.
(2) $\partial L_N/\partial \dot{z}^r = g_{rs}\dot{z}^s$ (symmetry of g); V_0 is constant and drops out.
(3) Product rule; $\dot{g}_{rs} = \partial_t g_{rs}\dot{z}^t$ (here $\partial_t = \partial/\partial z^t$).
(4) Write $\partial_t g_{rs}\dot{z}^t\dot{z}^s$ as half of itself plus half of the same term with $s \leftrightarrow t$.

Multiplying by the inverse metric g^{ur} gives the geodesic equation

$$\ddot{z}^u + \Gamma^u_{st}\,\dot{z}^s\dot{z}^t = 0, \qquad \Gamma^u_{st} = \tfrac{1}{2}g^{ur}\big(\partial_s g_{rt} + \partial_t g_{rs} - \partial_r g_{st}\big). \tag{8.334}$$

The potential has disappeared completely: only the metric matters.

Worked example 8.149: A ball in a Mexican hat: when is the approximation good?

Take a unit mass in the plane with $V = \frac{\lambda}{4}(|\vec{x}|^2 - 1)^2$. The valley N is the unit circle. Write $\vec{x} = (1+\rho)(\cos\theta, \sin\theta)$. Then

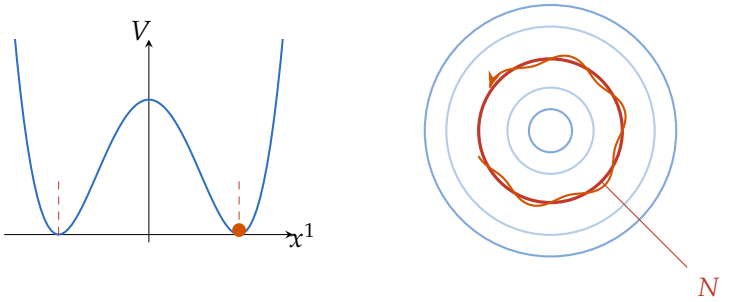
$$L \overset{(1)}{=} \tfrac{1}{2}\big(\dot\rho^2 + (1+\rho)^2\dot\theta^2\big) - \tfrac{\lambda}{4}\big((1+\rho)^2 - 1\big)^2$$
$$\overset{(2)}{=} \tfrac{1}{2}\big(\dot\rho^2 + (1+\rho)^2\dot\theta^2\big) - \lambda\rho^2\big(1 + \tfrac{\rho}{2}\big)^2.$$

(1) The flat metric in polar coordinates is $\mathrm{d}r^2 + r^2\mathrm{d}\theta^2$, with $r = 1+\rho$.
(2) $(1+\rho)^2 - 1 = 2\rho + \rho^2 = 2\rho(1+\rho/2)$.

The radial equation, to first order in ρ, is $\ddot\rho = (1+\rho)\dot\theta^2 - 2\lambda\rho + \dots$ For a uniform rotation $\dot\theta = \omega$ the radius settles at $\rho \approx \omega^2/(2\lambda)$: the centrifugal push is balanced by the stiffness of the wall. The motion on N ($\rho = 0$, $\theta = \omega t$, a geodesic of the circle) is a good approximation when

$$\omega^2 \ll \lambda,$$

i.e. when the speed is small compared with the "frequency" of oscillation across the valley. For monopoles the role of $\sqrt{\lambda}$ is played by the mass gap of the fields that are not zero modes: the W mass C (in units with $c = 1$). So the geodesic approximation should be good for speeds $v \ll 1$ (in units of the speed of light).



top view: orbit hugs the valley

Figure 8.75: The geodesic approximation in miniature. Left: the potential $V = \frac{\lambda}{4}(|\vec{x}|^2 - 1)^2$ along a line through the origin; its minimum set N (red) is a circle, not a point. Right: a slowly moving ball (orange) stays near the bottom of the valley, with small wiggles of amplitude $\omega^2/2\lambda$, and follows a geodesic of N. For monopoles, Q becomes the infinite-dimensional space $\mathcal{A}_{\mathrm{YMH}}/\mathcal{G}$ and N becomes $\mathcal{M}'_k$.

Dictionary: the geodesic approximation

configuration space Q	$\mathcal{A}_{\mathrm{YMH}}/\mathcal{G}$, pairs (A,ϕ) modulo gauge
potential V	static energy $E(A,\phi)$, Nash's (8.194)
valley bottom N, $V = V_0$	moduli space $\mathcal{M}'_k$, $E = 4\pi C k$
mass matrix G_{IJ}	the L^2 metric on field variations, Nash's (8.192)
induced metric g_{rs}	the moduli space metric
motion of the ball	slow monopole scattering

Reading Nash (p. 255, geodesics and the Morse energy functional)

Nash remarks that the equations of motion "are just those for the geodesics on $\mathcal{M}'_k$ . . . (compare with the geodesics giving the critical points of the Morse energy functional on

p. 206)". The link: a path $z(t)$, $0 \le t \le T$, is a geodesic iff it is a critical point of the energy functional $\mathcal{E}(z) = \frac{1}{2}\int_0^T g_{rs}\dot{z}^r\dot{z}^s\,\mathrm{d}t$ with fixed endpoints (the calculation in Example 8.148 is precisely $\delta\mathcal{E} = 0$). This functional is the one whose Morse theory computes the topology of the loop space in Chapter 7. The monopole's action on $\mathcal{M}'_k$ is $\int(L_N)\mathrm{d}t = \mathcal{E} - 4\pi Ck T$.

8.16.3 Kinetic energy on the space of fields and Gauss's law

To apply the idea to monopoles we need their kinetic energy. The static energy $E(A,\phi)$ came from a Lagrangian in $3+1$ dimensions. Write the time component of the gauge field as A_t, and let a dot denote $\partial/\partial t$. The electric field and the time covariant derivative of ϕ are

$$E_i \overset{(1)}{=} F_{ti} = \partial_t A_i - \partial_i A_t + [A_t, A_i] \overset{(2)}{=} \dot{A}_i - D_i A_t,$$
$$D_t\phi \overset{(3)}{=} \dot{\phi} + [A_t,\phi] = \dot{\phi} - [\phi, A_t].$$

(1) $F = \mathrm{d}A + A\wedge A$ in components: $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu + [A_\mu, A_\nu]$.
(2) $D_iA_t = \partial_i A_t + [A_i, A_t]$ and $[A_t, A_i] = -[A_i, A_t]$.
(3) Definition of the covariant derivative of an adjoint field.

The Yang–Mills–Higgs Lagrangian (in the Prasad–Sommerfield limit, with signature $(+,-,-,-)$ so that energies are positive) is

$$\mathcal{L} = \tfrac{1}{2}\left(\|E\|^2 + \left\|D_t\phi\right\|^2\right) - \tfrac{1}{2}\left(\|F\|^2 + \left\|\mathrm{d}_A\phi\right\|^2\right) = T - E(A,\phi), \tag{8.335}$$

with the norms of (8.266) taken over $\mathbb{R}^3$ at fixed time. The second bracket is the static energy of Nash's (8.141) and (8.194). The first is the kinetic energy. The two together say $\mathcal{L} = T - V$, as in the finite-dimensional model.

Now look at the formulas for E_i and $D_t\phi$ again. With $\dot{m} = (\dot{A}, \dot{\phi})$ the velocity of the configuration and $D_0 f = (\mathrm{d}_A f, [\phi, f])$ the infinitesimal gauge transformation (8.321),

$$(E,\ D_t\phi) = \dot{m} - D_0 A_t, \qquad T = \tfrac{1}{2}\left\|\dot{m} - D_0 A_t\right\|^2. \tag{8.336}$$

The field A_t has no time derivative in $\mathcal{L}$: it is not dynamical. Its equation of motion is a constraint, obtained by varying $A_t \mapsto A_t + sf$:

$$0 \overset{(1)}{=} \frac{\mathrm{d}}{\mathrm{d}s}\Big|_{s=0} \tfrac{1}{2}\left\|\dot{m} - D_0A_t - sD_0 f\right\|^2 \overset{(2)}{=} -\left\langle \dot{m} - D_0A_t,\ D_0 f\right\rangle \overset{(3)}{=} -\left\langle D_0^\star(\dot{m} - D_0A_t),\ f\right\rangle.$$

(1) Only T contains A_t, through (8.336); D_0 is linear.
(2) Differentiate the square of a norm.
(3) Definition of the adjoint operator $D_0^\star$.

This holds for all f, so $D_0^\star(E, D_t\phi) = 0$. By (8.326) this reads

$$D_iE_i + [\phi, D_t\phi] = 0, \tag{8.337}$$

which is **Gauss's law**: $\nabla\cdot\vec{E} = \rho$, with the charge density $\rho = -[\phi, D_t\phi]$ carried by the moving Higgs field.

Key idea: The meaning of Nash's projection

Gauss's law says that the physical velocity $(E, D_t\phi)$ is *orthogonal to the gauge orbit*. Since T is a positive quadratic function of A_t, the constraint picks the A_t that *minimises* T: $\dot{m} - D_0A_t$

is the part of $\dot m$ orthogonal to $\operatorname{Im} D_0$. So

$$T = \tfrac{1}{2}\,\|P\dot m\|^2\,, \qquad P = \text{orthogonal projection onto } (\operatorname{Im} D_0)^\perp = \operatorname{Ker} D_0^\star .$$

This is the precise meaning of Nash's sentence "this metric descends to the configuration space $\mathcal{A}_{\mathrm{YMH}}/\mathcal{G}$ if we project onto those (a,η) orthogonal to the orbit". The projection is not a choice; it is Gauss's law.

Working definition 8.45 (the moduli space metric, operationally)

Let z^r be coordinates on $\mathcal{M}'_k$ [the framed moduli space, Bogomolny solutions modulo based gauge transformations] and $m(z) = (A(z), \phi(z))$ a family of Bogomolny solutions. For each r:

1. differentiate: $\partial_r m = (\partial A/\partial z^r,\ \partial\phi/\partial z^r)$; it solves the linearised Bogomolny equation, $D_1(\partial_r m) = 0$;
2. add a gauge direction to put it in the slice [the vectors with $D_0^\star(a,\eta) = 0$, orthogonal to the gauge orbit]: $\delta_r m = \partial_r m - D_0\varepsilon_r$ with ε_r chosen so that $D_0^\star \delta_r m = 0$ (Gauss's law);
3. take inner products:

$$g_{rs} = \langle \delta_r m, \delta_s m\rangle = -2\int_{\mathbb{R}^3} \operatorname{tr}\big(\delta_r A \wedge \star\delta_s A\big) - 2\int_{\mathbb{R}^3} \operatorname{tr}\big(\delta_r\phi\,\delta_s\phi\big)\,\mathrm{d}^3x. \tag{8.338}$$

Then $T = \tfrac{1}{2} g_{rs}\dot z^r \dot z^s$. In practice one computes the zero modes of $\mathcal{D}$ in Example 8.138, which already satisfy the slice condition.

Key idea: Manton's geodesic approximation

Slowly moving BPS monopoles of charge k are described, to a good approximation, by a path $z(t)$ in $\mathcal{M}'_k$ with Lagrangian

$$L_{\mathrm{eff}} = \tfrac{1}{2} g_{rs}(z)\,\dot z^r \dot z^s - 4\pi C k.$$

The equations of motion are the geodesic equations (8.334) of the metric (8.338). Monopole scattering at low speeds *is* geodesic motion on $\mathcal{M}'_k$.

Nash's (8.192)–(8.194) write the same thing with coordinates X^I on $\mathcal{A}_{\mathrm{YMH}}$. The approximation was later proved to be valid for small speeds (Stuart): true solutions of the time-dependent field equations with initial data close to $\mathcal{M}'_k$ and small initial velocity stay close, for long times, to the geodesic.

Watch out: Factors in (8.192)–(8.194)

Nash writes $T = \langle \dot m, \dot m\rangle$ in (8.193) without the $\tfrac{1}{2}$ and the inner product with $-\operatorname{tr}$ instead of $-2\operatorname{tr}$. Neither factor affects the geodesics (an overall constant in a metric does not change its geodesics). We keep $T = \tfrac{1}{2}\|P\dot m\|^2$ with (8.266), which is what makes the metric coefficient of a translation equal to the mass of the monopole below.

8.16.4 One monopole: the metric on $\mathbb{R}^3 \times S^1$ and the dyon

We now compute the metric on $\mathcal{M}_1' = \mathbb{R}^3 \times S^1$, and part of it on every $\mathcal{M}_k'$. The zero modes of Section 8.15.5 are all we need.

Worked example 8.150: The translation metric is the total mass, for every k

Let (A, ϕ) be any Bogomolny solution of charge k. The translation zero modes are $\delta_j m = (\iota_{\partial_j} F,\ D_j\phi)$, i.e. $(\delta_j A)_k = F_{jk}$ and $\delta_j\phi = D_j\phi$ (Example 8.140). Then

$$\begin{aligned} g_{jl} &\overset{(1)}{=} \int_{\mathbb{R}^3} \Big(\sum_k (F_{jk}, F_{lk}) + (D_j\phi, D_l\phi) \Big) \mathrm{d}^3x \\ &\overset{(2)}{=} \int_{\mathbb{R}^3} \Big(\sum_k \varepsilon_{jkm}\varepsilon_{lkn}(B_m, B_n) + (B_j, B_l) \Big) \mathrm{d}^3x \\ &\overset{(3)}{=} \int_{\mathbb{R}^3} \Big(\delta_{jl}|B|^2 - (B_l, B_j) + (B_j, B_l) \Big) \mathrm{d}^3x \overset{(4)}{=} \delta_{jl} \int_{\mathbb{R}^3} |B|^2 \, \mathrm{d}^3x \overset{(5)}{=} 4\pi C k \, \delta_{jl}. \end{aligned}$$

(1) The metric (8.338) written with the pointwise product $(X, Y) = -2\,\mathrm{tr}(XY)$ and components.
(2) $F_{jk} = \varepsilon_{jkm} B_m$, and the Bogomolny equation $D_j\phi = -B_j$.
(3) $\varepsilon_{jkm}\varepsilon_{lkn} = \delta_{jl}\delta_{mn} - \delta_{jn}\delta_{ml}$.
(4) The last two terms cancel.
(5) On a Bogomolny solution $|B| = |D\phi|$, so $\int |B|^2 = \frac{1}{2}\int (|B|^2 + |D\phi|^2) = E = 4\pi C k$.

So $\frac{1}{2} g_{jl} \dot{X}^j \dot{X}^l = \frac{1}{2}(4\pi C k)|\dot{\vec{X}}|^2$: the centre of mass $\vec{X}$ moves like a free particle whose inertial mass equals the rest energy $4\pi C k$. This is the "no accident" promised in Example 8.140. It is Einstein's $E = Mc^2$ seen through the moduli space. Notice that we did not need spherical symmetry.

Worked example 8.151: The phase direction, and orthogonality

The phase rotation (8.317) is generated by $f = \phi/C$, so its tangent vector is $\delta_\chi m = D_0(\phi/C) = (\mathrm{d}_A\phi/C,\, 0)$. Its length:

$$g_{\chi\chi} \overset{(1)}{=} \frac{1}{C^2} \int_{\mathbb{R}^3} |D\phi|^2 \, \mathrm{d}^3x \overset{(2)}{=} \frac{4\pi C k}{C^2} = \frac{4\pi k}{C}.$$

(1) Only the A-component is non-zero.
(2) $\int |D\phi|^2 = \int |B|^2 = 4\pi C k$, as in the previous example.

It is orthogonal to the translations:

$$g_{j\chi} \overset{(1)}{=} \frac{1}{C} \int_{\mathbb{R}^3} \sum_k (F_{jk}, D_k\phi)\, \mathrm{d}^3x \overset{(2)}{=} -\frac{1}{C} \int_{\mathbb{R}^3} \varepsilon_{jkm}(D_m\phi, D_k\phi)\, \mathrm{d}^3x \overset{(3)}{=} 0.$$

(1) Pair $(\iota_{\partial_j} F, D_j\phi)$ with $(\mathrm{d}_A\phi/C, 0)$; only the 1-form parts meet.
(2) $F_{jk} = \varepsilon_{jkm} B_m = -\varepsilon_{jkm} D_m\phi$.
(3) ε_{jkm} is antisymmetric in k, m and $(D_m\phi, D_k\phi)$ is symmetric.

For $k = 1$ these four modes span $T\mathcal{M}_1'$, so

$$\boxed{\mathrm{d}s^2_{\mathcal{M}_1'} = 4\pi C\, |\mathrm{d}\vec{X}|^2 + \frac{4\pi}{C}\, \mathrm{d}\chi^2, \qquad \chi \sim \chi + 2\pi} \tag{8.339}$$

a flat metric on $\mathbb{R}^3 \times S^1$, the circle having radius $\sqrt{4\pi/C}$.

For $k \geq 2$ the same four directions (overall translation and overall phase) exist and have the metric $4\pi C k|\mathrm{d}\vec{X}|^2 + (4\pi k/C)\mathrm{d}\chi^2$. The remaining $4k-4$ directions are the *relative* motions, and one can show that they are orthogonal to these four. So, up to a finite covering,

$$\mathcal{M}'_k \cong \mathbb{R}^3 \times S^1 \times \mathcal{M}^0_k \quad \text{(as Riemannian manifolds, up to a } \mathbb{Z}_k \text{ quotient)},$$

with a flat metric on the first two factors. The centre of mass and the overall phase move freely, and all the interesting scattering happens on $\mathcal{M}^0_k$. This is why Nash says "it is the flow on this latter space which contains the interesting scattering information".

Worked example 8.152: Geodesics on $\mathbb{R}^3 \times S^1$: moving monopoles and dyons

Geodesics of the flat metric (8.339) are straight lines: $\vec{X}(t) = \vec{X}_0 + \vec{v}t$ and $\chi(t) = \chi_0 + \omega t$. The first is a monopole moving at constant velocity. What is the second?
Its conserved momentum is $p_\chi = \partial L_{\text{eff}}/\partial\dot{\chi} = (4\pi/C)\,\omega$. To interpret it, look at the fields. In the gauge $A_t = 0$ the velocity is $\dot{m} = \omega\,\delta_\chi m$, so the electric field is

$$E_i = \dot{A}_i = \frac{\omega}{C}\,D_i\phi.$$

It obeys Gauss's law (8.337): $D_iE_i + [\phi, D_t\phi] = \frac{\omega}{C}D_iD_i\phi + 0 = 0$ by the field equation $\mathrm{d}_A^\star \mathrm{d}_A\phi = 0$. Far away, the electric field along the unbroken U(1) is

$$e_i = (E_i, \hat{\phi}) \overset{(1)}{=} \frac{\omega}{C}\,(D_i\phi, \hat{\phi}) \overset{(2)}{=} \frac{\omega}{C}\,\partial_i|\phi| \overset{(3)}{=} \frac{\omega}{C}\,\frac{\hat{x}_i}{r^2}.$$

(1) Take the component of E_i along the unit Higgs direction $\hat{\phi} = \phi/|\phi|$.
(2) $(D_i\phi, \phi) = \frac{1}{2}\partial_i|\phi|^2$, since the bracket term $[A_i, \phi]$ is orthogonal to ϕ.
(3) The Higgs tail $|\phi| = C - 1/r$.

A Coulomb field $e = q\,\hat{\vec{x}}/(4\pi r^2)$ has charge $q = 4\pi\omega/C$. So

$$p_\chi = \frac{4\pi\omega}{C} = q :$$

the momentum conjugate to the phase is the electric charge. A monopole whose phase rotates is a **dyon**, carrying both magnetic and electric charge (Julia–Zee). Its energy is

$$E = 4\pi C + \tfrac{1}{2}\cdot\frac{4\pi}{C}\,\omega^2 = 4\pi C + \frac{C\,q^2}{8\pi}.$$

Compare the exact BPS dyon, whose mass is $C\sqrt{g^2+q^2}$ with $g = 4\pi$:

$$C\sqrt{16\pi^2 + q^2} = 4\pi C\left(1 + \frac{q^2}{16\pi^2}\right)^{1/2} = 4\pi C + \frac{Cq^2}{8\pi} + O(q^4).$$

The geodesic approximation reproduces the first correction exactly, as a non-relativistic expansion should. Finally, χ is periodic, so in quantum mechanics p_χ is an integer (with $\hbar = 1$): *electric charge is quantised because the phase lives on a circle.* The unit is the charge of the W boson.

Watch out: Dyons need the phase to be a genuine coordinate

If the phase rotation were a gauge transformation to be divided out, χ would not be a coordinate and there would be no electric charge. This is exactly why $\mathcal{M}'_k$ keeps the gauge

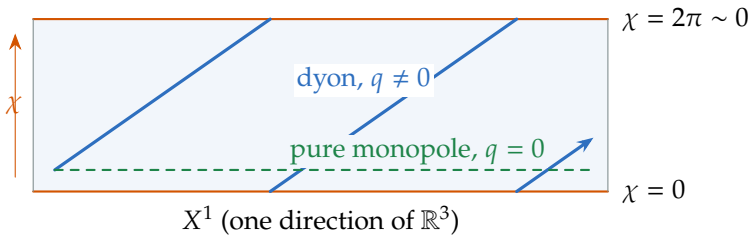


Figure 8.76: Geodesics on $\mathcal{M}_1' = \mathbb{R}^3 \times S^1$, drawn on the flat cylinder obtained by identifying the top and bottom edges. A horizontal line (green) is a monopole moving with constant velocity. A tilted line (blue) also winds around the phase circle: the phase rotates at a constant rate, and the rate is the electric charge. The slope is $\mathrm{d}\chi/\mathrm{d}X^1 = \omega/v = C^2 q/P$, where $P = 4\pi C v$ and q are the two conserved momenta.

transformations that are non-trivial at infinity (Section 8.15.6). Dividing by the circle gives $\mathcal{M}_k$ and forgets the charge; dividing also by $\mathbb{R}^3$ gives $\mathcal{M}_k^0$. For $k = 2$, $\mathcal{M}_2^0$ still contains the *relative* phase, and the relative electric charge will reappear as "spin" in Section 8.17.4.

8.16.5 Hyperkähler manifolds

To go beyond one monopole we need to know more about the metric. Nash's statement is that it is *hyperkähler*. We first recall Kähler metrics (Chapter 5), then build the quaternionic generalisation concretely on $\mathbb{R}^4$, then explain why monopole moduli spaces have it.

Kähler, recalled

A **complex structure** on a real vector space (or on each tangent space of a manifold) is a linear map J with $J^2 = -1$: "multiplication by i". A metric g is **Hermitian** if $g(JX, JY) = g(X, Y)$. Then the **Kähler form** $\omega(X, Y) = g(JX, Y)$ is antisymmetric:

$$\omega(Y, X) \overset{(1)}{=} g(JY, X) \overset{(2)}{=} g(J^2 Y, JX) \overset{(3)}{=} -g(Y, JX) \overset{(4)}{=} -g(JX, Y) = -\omega(X, Y).$$

(1) Definition of ω.
(2) g is Hermitian: apply J to both arguments.
(3) $J^2 = -1$.
(4) g is symmetric.

So ω is a 2-form. The metric is **Kähler** if $\mathrm{d}\omega = 0$ (and J comes from complex coordinates). In complex coordinates, $\omega = \frac{i}{2} g_{j\bar{k}} \mathrm{d}z^j \wedge \mathrm{d}\bar{z}^k$, as Nash writes.

Worked example 8.153: The Kähler form of $\mathbb{C}$

On $\mathbb{C}$ with $z = x + iy$ and $g = \mathrm{d}x^2 + \mathrm{d}y^2$, i.e. $g_{z\bar{z}} = 1$:

$$\tfrac{i}{2} \mathrm{d}z \wedge \mathrm{d}\bar{z} = \tfrac{i}{2} (\mathrm{d}x + i\mathrm{d}y) \wedge (\mathrm{d}x - i\mathrm{d}y) = \tfrac{i}{2}(-2i)\, \mathrm{d}x \wedge \mathrm{d}y = \mathrm{d}x \wedge \mathrm{d}y.$$

Directly: $J\partial_x = \partial_y$, so $\omega(\partial_x, \partial_y) = g(\partial_y, \partial_y) = 1$. The Kähler form is the area form. It is closed trivially (constant coefficients).

Three complex structures: the quaternions

Working definition 8.46 (hyperkähler manifold)

A Riemannian manifold (M, g) is **hyperkähler** if it carries three complex structures [maps I of each tangent space to itself with $I^2 = -1$, like multiplication by i] I, J, K satisfying the quaternion relations, Nash's (8.195),

$$I^2 = J^2 = K^2 = IJK = -1, \tag{8.340}$$

such that g is Hermitian for each [$g(IX, IY) = g(X, Y)$, and likewise for J and K], and the three Kähler forms, Nash's (8.196),

$$\omega_I(X, Y) = g(IX, Y), \qquad \omega_J(X, Y) = g(JX, Y), \qquad \omega_K(X, Y) = g(KX, Y) \tag{8.341}$$

are all closed. *How to test it:* exhibit the three 2-forms and check $\mathrm{d}\omega = 0$ for each. (Equivalently: the holonomy group of g [the rotations of a tangent space produced by parallel transport around loops] lies in $\mathrm{Sp}(n)$ [the length-preserving $n \times n$ quaternionic matrices, with $\dim M = 4n$], and such metrics are automatically Ricci-flat [Ric = 0, the vacuum Einstein equations].)

From (8.340) follow the familiar rules $IJ = K = -JI$, $JK = I = -KJ$, $KI = J = -IK$. For instance, multiplying $IJK = -1$ on the left by I gives $-JK = -I$, so $JK = I$; then $JI = J(JK) = J^2K = -K$. The rules imply that each unit combination is again a complex structure: for $a^2 + b^2 + c^2 = 1$,

$$\begin{aligned}(aI + bJ + cK)^2 &\overset{(1)}{=} a^2I^2 + b^2J^2 + c^2K^2 + ab(IJ + JI) + bc(JK + KJ) + ca(KI + IK) \\ &\overset{(2)}{=} -(a^2 + b^2 + c^2) \overset{(3)}{=} -1.\end{aligned}$$

(1) Multiply out, keeping the order of the non-commuting factors.
(2) $I^2 = J^2 = K^2 = -1$ and the anticommutators vanish.
(3) $a^2 + b^2 + c^2 = 1$.

So a hyperkähler manifold has a whole 2-sphere of complex structures, and is Kähler for each of them (Figure 8.77).

Worked example 8.154: $\mathbb{R}^4 = \mathbb{H}$: the complex structures as 4×4 matrices

Identify $(x^0, x^1, x^2, x^3) \in \mathbb{R}^4$ with the quaternion $q = x^0 + x^1\mathbf{i} + x^2\mathbf{j} + x^3\mathbf{k}$, where $\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = \mathbf{ijk} = -1$ (so $\mathbf{ij} = \mathbf{k}$, $\mathbf{jk} = \mathbf{i}$, $\mathbf{ki} = \mathbf{j}$). Let I, J, K be *left* multiplication by $\mathbf{i}, \mathbf{j}, \mathbf{k}$. Compute I:

$$\mathbf{i}\,q \overset{(1)}{=} x^0\mathbf{i} + x^1\mathbf{i}^2 + x^2\mathbf{ij} + x^3\mathbf{ik} \overset{(2)}{=} -x^1 + x^0\mathbf{i} - x^3\mathbf{j} + x^2\mathbf{k}.$$

(1) Distribute; real numbers commute with everything.
(2) $\mathbf{i}^2 = -1$, $\mathbf{ij} = \mathbf{k}$, $\mathbf{ik} = -\mathbf{ki} = -\mathbf{j}$; sort by units.

Doing the same for $\mathbf{j}$ and $\mathbf{k}$ (using $\mathbf{ji} = -\mathbf{k}$, $\mathbf{jk} = \mathbf{i}$, $\mathbf{ki} = \mathbf{j}$, $\mathbf{kj} = -\mathbf{i}$) gives

$$I = \begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \quad J = \begin{pmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix}, \quad K = \begin{pmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

Each is an orthogonal antisymmetric matrix, so $g(IX, IY) = g(X, Y)$ and $I^2 = -I^TI = -1$. Check $IJ = K$ on a basis vector: $Je_0 = e_2$ (third column of J) and $Ie_2 = e_3$, so $IJe_0 = e_3 = Ke_0$;

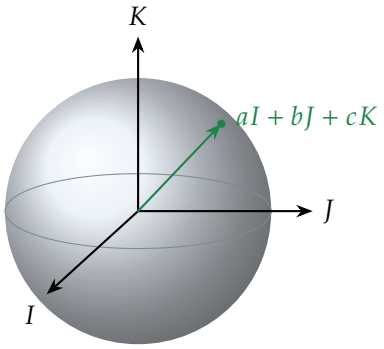
similarly on e_1, e_2, e_3. (Of course: $\mathbf{i}(\mathbf{j}q) = (\mathbf{ij})q = \mathbf{k}q$ by associativity.) The Kähler forms: $\omega_I(e_a, e_b) = g(Ie_a, e_b) = I_{ba}$. From $Ie_0 = e_1$ and $Ie_2 = e_3$,

$$\omega_I = \mathrm{d}x^0 \wedge \mathrm{d}x^1 + \mathrm{d}x^2 \wedge \mathrm{d}x^3,$$
$$\omega_J = \mathrm{d}x^0 \wedge \mathrm{d}x^2 + \mathrm{d}x^3 \wedge \mathrm{d}x^1,$$
$$\omega_K = \mathrm{d}x^0 \wedge \mathrm{d}x^3 + \mathrm{d}x^1 \wedge \mathrm{d}x^2.$$

All three are closed (constant coefficients), so $\mathbb{R}^4$ is hyperkähler. And all three are *self-dual*: $\star(\mathrm{d}x^0 \wedge \mathrm{d}x^1) = \mathrm{d}x^2 \wedge \mathrm{d}x^3$, etc. The three Kähler forms of $\mathbb{R}^4$ are a basis of Λ^2_+, the same space in which the curvature of an instanton lives.

Physical picture: Why hyperkähler means dimension $4n$

One complex structure makes each tangent space a complex vector space, so its real dimension is even. Three complex structures obeying the quaternion relations make it a vector space over the quaternions $\mathbb{H}$, so its real dimension is a multiple of 4: $\dim M = 4n$, as Nash says. This matches $\dim \mathcal{M}'_k = 4k$ and $\dim \mathcal{M}^0_k = 4k-4$. It does *not* fit $\mathcal{M}_k$ (dimension $4k-1$, odd), which is why $\mathcal{M}_k$ is not the space on which one does the geometry.



$a^2 + b^2 + c^2 = 1$: every point of the sphere is a complex structure, and g is Kähler for each.

spatial rotations of $\mathbb{R}^3$ rotate this sphere (Section 8.17.1)

Figure 8.77: The 2-sphere of complex structures of a hyperkähler manifold. The three axes are the quaternion units I, J, K; any unit combination squares to -1. For monopole moduli spaces the three axes correspond to the three components of the Bogomolny equation, so a rotation of physical space rotates the sphere.

Why monopole moduli spaces are hyperkähler

Two facts, one pointwise and one global.

Pointwise: each tangent space is quaternionic. In Example 8.138 we packaged a tangent vector (a, η) into the quaternion-valued field $q = \eta + a_i e_i$ and found that $T_m \mathcal{M}'_k = \operatorname{Ker} \mathcal{D}$ with $\mathcal{D}q = \sum_i e_i D_i q + [q, \phi]$. The operator uses *left* multiplication by the units e_i. Right multiplication by a quaternion commutes with left multiplication (associativity: $(uq)v = u(qv)$), and it commutes with D_i and with $[\,\cdot\,, \phi]$ (which act on the $\mathfrak{su}(2)$ factor). So if $q \in \operatorname{Ker} \mathcal{D}$ then $q\bar{e}_l \in \operatorname{Ker} \mathcal{D}$, where $\bar{e}_l = -e_l$ is the conjugate unit. Define

$$I\,q = q\,\bar{e}_1, \qquad J\,q = q\,\bar{e}_2, \qquad K\,q = q\,\bar{e}_3. \tag{8.342}$$

Then $IJq = q\,\bar{e}_2\bar{e}_1 = q\,e_2e_1 = -q\,e_3 = q\,\bar{e}_3 = Kq$, and $I^2q = q\,\bar{e}_1^2 = -q$: the quaternion relations hold. Right multiplication by a unit quaternion preserves the pointwise length $|q|^2$, so the L^2 metric is Hermitian for all three.

Worked example 8.155: The complex structures on $\mathcal{M}_1'$, explicitly

The phase mode is $(\mathrm{d}_A\phi, 0)$, i.e. $q_0 = \sum_i D_i\phi\, e_i$ (no real part). Apply I:

$$q_0\,\bar{e}_1 \overset{(1)}{=} -\sum_i D_i\phi\, e_i e_1 \overset{(2)}{=} -\sum_i D_i\phi\big(-\delta_{i1} + \varepsilon_{i1k}e_k\big) \overset{(3)}{=} D_1\phi + \sum_{i,k}\varepsilon_{1ik}D_i\phi\, e_k$$
$$\overset{(4)}{=} D_1\phi + \sum_{i,k}\varepsilon_{1ki}B_i\, e_k \overset{(5)}{=} D_1\phi + \sum_k F_{1k}\, e_k.$$

(1) Definition (8.342), $\bar{e}_1 = -e_1$; the real coefficients $D_i\phi$ commute with the units.
(2) Quaternion multiplication (8.327).
(3) The δ_{i1} term gives $D_1\phi$; and $-\varepsilon_{i1k} = \varepsilon_{1ik}$.
(4) Bogomolny, $D_i\phi = -B_i$, and $\varepsilon_{1ik} = -\varepsilon_{1ki}$.
(5) $F_{jk} = \varepsilon_{jkm}B_m$.

The result has real part $\eta = D_1\phi$ and imaginary part $a_k = F_{1k}$: it is exactly the translation mode $(\iota_{\partial_1}F, D_1\phi)$. So

$$I(\text{phase mode}) = \text{translation mode in the } x^1 \text{ direction},$$

and likewise J and K turn the phase mode into translations along x^2 and x^3. In the coordinates $(y, \vec{X})$ with $y = \chi/C$, the metric (8.339) is $4\pi C(\mathrm{d}y^2 + |\mathrm{d}\vec{X}|^2)$ and $I\partial_y = \partial_{X^1}$: this is $\mathbb{R}^4 = \mathbb{H}$ with y as the real coordinate, rolled up along y. So $\mathcal{M}_1'$ is hyperkähler, as the flat quotient $\mathbb{H}/\mathbb{Z}$.

Globally: the forms are closed. Pointwise quaternionic structure is not enough; the three Kähler forms must be closed. The reason is a general construction. Package the four fields into one quaternion-valued field $\mathbb{A} = \phi + A_1e_1 + A_2e_2 + A_3e_3$. The space of all such fields is a *flat* (infinite-dimensional) quaternionic vector space, so its three Kähler forms are constant, hence closed. The gauge group preserves them. For such a symmetry there are three **moment maps** (the "conserved charges" of the three symplectic forms, as angular momentum is for rotations), and here they are the three components of the Bogomolny expression:

$$\mu_i(A,\phi) = B_i + D_i\phi, \qquad i = 1,2,3.$$

So the Bogomolny equation says $\mu_1 = \mu_2 = \mu_3 = 0$.

Key idea: Hyperkähler quotient (Hitchin–Karlhede–Lindström–Roček)

If a group $\mathcal{G}$ acts on a hyperkähler manifold preserving g, I, J, K, with moment maps μ_1, μ_2, μ_3, then $\mu^{-1}(0)/\mathcal{G}$ is again hyperkähler. If the manifold is X, the quotient has dimension $\dim X - 4\dim\mathcal{G}$. Applied to the flat space of fields (A,ϕ) and the gauge group, it gives

$$\mathcal{M}_k' = \{B_i + D_i\phi = 0\}/\mathcal{G} \quad \text{is hyperkähler.}$$

Why dimension drops by 4 per generator: each gauge generator removes 3 dimensions through the three equations $\mu_i = 0$ and 1 more through the quotient. At the level of fields at a point: 4 real fields per Lie algebra direction (ϕ, A_1, A_2, A_3), minus 3 equations, minus 1 gauge freedom, leaves $4 - 3 - 1 = 0$ local degrees of freedom. That is why the solution space is finite-dimensional, and its dimension $4k$ is supplied by the index theorem.

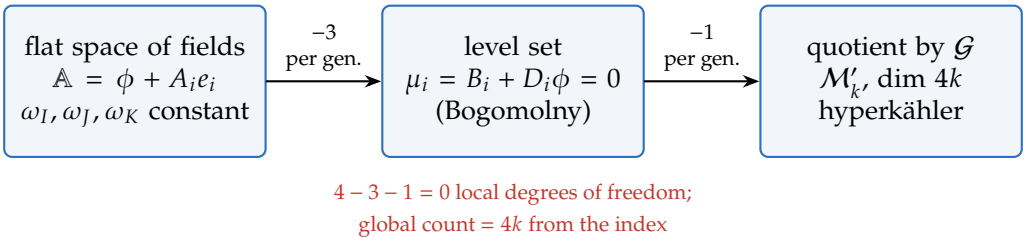


Figure 8.78: The hyperkähler quotient construction of the monopole moduli space. The three components of the Bogomolny equation are the three moment maps of the gauge action; imposing them and dividing by the gauge group preserves the hyperkähler structure.

Watch out: Which spaces are hyperkähler

$\mathcal{M}'_k$ (dimension $4k$) and $\mathcal{M}^0_k$ (dimension $4k-4$) are hyperkähler; $\mathcal{M}_k$ (dimension $4k-1$) cannot be. Nash's statement "the reduced moduli space $\mathcal{M}^0_k$ ... is also hyperkähler" uses the splitting $\mathcal{M}'_k \approx \mathbb{R}^3 \times S^1 \times \mathcal{M}^0_k$ above: the flat factor $\mathbb{R}^3 \times S^1$ is hyperkähler (Example 8.155), and the complement inherits the structure. Geodesics on $\mathcal{M}'_k$ project to geodesics on $\mathcal{M}^0_k$ because the flat factor splits off orthogonally.

Physical picture: Why hyperkähler makes the metric computable

A general metric on a 4-manifold with an SO(3) symmetry depends on several arbitrary functions of one variable. Demanding that it be hyperkähler turns those functions into solutions of a system of first-order ODEs, and demanding the right behaviour at large separation (point monopoles) and regularity picks a unique solution. So the metric on $\mathcal{M}^0_2$ can be found *without* solving the Bogomolny equation for two monopoles. This is what Nash means by "the hyperkählerity of the metric is a key factor in rendering it computable". We do it in the next section.

Reading Nash (pp. 253–255, (8.191)–(8.196))

Nash notes $E = 4\pi k$ (8.191) and deduces the absence of forces between well-separated monopoles, observes that the particle picture breaks down when they come close, and introduces Manton's idea: the L^2 metric (8.192) on $\mathcal{A}_{\mathrm{YMH}}$, projected orthogonally to gauge orbits, the kinetic energy (8.193) and potential energy (8.194), and the conclusion that slow motion is geodesic motion on $\mathcal{M}'_k$. He then recalls Kähler metrics and defines hyperkähler ones by (8.195)–(8.196). We added: the force balance behind (8.191), the derivation of the geodesic equation and the Mexican-hat model, the identification of the projection with Gauss's law, the exact translation and phase metrics, the dyon and charge quantisation, explicit quaternion matrices and Kähler forms, the quaternionic structure of the zero modes, and the hyperkähler quotient.

Recap

- $E = 4\pi C k$ on all of $\mathcal{M}'_k$: magnetic repulsion and Higgs attraction cancel, so static monopoles exert no net force.
- Slow motion stays near the valley bottom $\mathcal{M}'_k$ and follows geodesics of the induced (L^2) metric. The orthogonal projection off gauge orbits is Gauss's law.
- The translation metric is $4\pi C k\, \delta_{jl}$ (inertial mass = rest energy); the phase metric is

$4\pi k/C$. For $k = 1$: flat $\mathbb{R}^3 \times S^1$. Motion around the phase circle is a dyon, with electric charge = momentum of the phase.

- Hyperkähler: three complex structures with quaternion relations and three closed Kähler forms; dimension $4n$. On $\mathbb{R}^4$ they are left multiplication by $\mathbf{i}, \mathbf{j}, \mathbf{k}$ and their forms span Λ^2_+.
- $\mathcal{M}'_k$ is a hyperkähler quotient: the three Bogomolny components are moment maps for the gauge group.

8.17 Two monopoles: the Atiyah–Hitchin manifold and scattering

For $k = 2$ the relative moduli space $\mathcal{M}^0_2$ is four-dimensional. Atiyah and Hitchin found its metric, and its geodesics describe the slow scattering of two monopoles. Nash (pp. 255–256) describes the parameters and the two headline results in words. In this section we

- build the four parameters and the rotational symmetry explicitly;
- write the metric, check its asymptotic form, and plot it;
- derive the 90° scattering of head-on collisions in two independent ways;
- compute the scattering angle as a function of impact parameter;
- explain why glancing collisions produce dyons with spin.

Units: $C = 1$ throughout this section, so each monopole has mass 4π and core size of order 1.

8.17.1 The four parameters and the rotation group

Start with two well-separated monopoles, where the particle picture applies. $\mathcal{M}'_2$ has dimension $8 = 2 \times (3 + 1)$: two positions and two phases. Remove the centre of mass (3 parameters) and the overall phase (1 parameter). Four are left, Nash's description:

- the *relative position*: put the monopoles at $\pm\vec{x}$, so the separation is $2|\vec{x}|$ (3 parameters);
- the *relative phase*: encoded, as Nash says, by a unit vector $\vec{y}$ perpendicular to $\vec{x}$ (1 parameter: $\vec{y}$ lies on a circle in the plane orthogonal to $\vec{x}$).

Two identifications remove redundancy. The monopoles are identical, so $\vec{x}$ and $-\vec{x}$ describe the same configuration. And $\vec{y}$ and $-\vec{y}$ are identified (Nash says "for covering space reasons"): a relative phase $\chi_1 - \chi_2$ is recorded by $\vec{y}$ at *twice* its angle, so turning $\vec{y}$ by π is a full turn of the phase.

Worked example 8.156: The configurations form a space of frames

Given $(\vec{x}, \vec{y})$, set

$$\vec{e}_3 = \hat{\vec{x}}, \qquad \vec{e}_1 = \vec{y}, \qquad \vec{e}_2 = \vec{e}_3 \times \vec{e}_1.$$

This is an oriented orthonormal frame, i.e. a rotation matrix $R = (\vec{e}_1\, \vec{e}_2\, \vec{e}_3) \in \mathrm{SO}(3)$. Conversely a frame and a number $r = 2|\vec{x}|$ give back $(\vec{x}, \vec{y})$. So for well-separated monopoles

$$\mathcal{M}^0_2 \supset \{(r, R)\}/\sim \;=\; (r_{\min}, \infty) \times \mathrm{SO}(3)/D_2,$$

where the identifications $\vec{x} \to -\vec{x}$ (with $\vec{y}$ fixed, and hence $\vec{e}_2 \to -\vec{e}_2$) and $\vec{y} \to -\vec{y}$ (with $\vec{x}$ fixed) generate the group

$$D_2 = \{\mathbb{1},\ \mathrm{diag}(1,-1,-1),\ \mathrm{diag}(-1,1,-1),\ \mathrm{diag}(-1,-1,1)\}$$

acting on the right, i.e. on the frame vectors. Each non-identity element is a rotation by π about one frame axis. Dimension count: $1 + 3 = 4$. ✓

The rotation group SO(3) of physical space acts on $\mathcal{M}_2^0$ by rotating both monopoles: $R \mapsto R_0 R$. This is an isometry of the metric (the energy and the kinetic energy are rotation invariant). So the metric is determined by its values along a single curve transverse to the orbits, parametrised by r. This is the key simplification: *a metric with* SO(3) *symmetry on a 4-manifold with 3-dimensional orbits depends only on functions of one variable.*

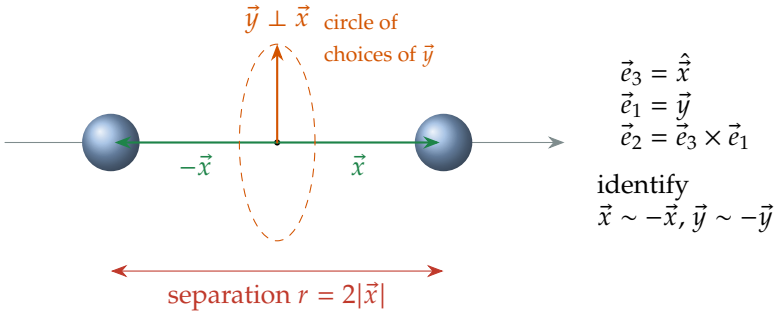


Figure 8.79: Nash's parameters for two well-separated monopoles: the half-separation vector $\vec{x}$ (3 numbers) and a unit vector $\vec{y}$ perpendicular to it (1 number, the relative phase). Together they define an orthonormal frame, so the configuration space is (separation)×SO(3), modulo the four-element group D_2 of sign flips.

The metric

Let $\sigma_1, \sigma_2, \sigma_3$ be the left-invariant 1-forms on SO(3): if $R(t)$ is a path of rotations, $R^{-1}\dot{R}$ is an antisymmetric matrix with entries $\varepsilon_{ijk}\,\omega_k$, and $\sigma_k(\dot{R}) = \omega_k$ is the angular velocity *about the body axis* $\vec{e}_k$. With the sign conventions of Atiyah and Hitchin they satisfy $\mathrm{d}\sigma_1 = \sigma_2 \wedge \sigma_3$ and cyclically. Rotation invariance of the metric under $R \mapsto R_0 R$ means that it is built from $\mathrm{d}r$ and the σ's with coefficients depending only on r. Invariance under the finite group D_2 (which flips the signs of two of the σ's) removes the cross terms. So

$$\mathrm{d}s^2 = f(r)^2\,\mathrm{d}r^2 + a(r)^2\,\sigma_1^2 + b(r)^2\,\sigma_2^2 + c(r)^2\,\sigma_3^2. \tag{8.343}$$

The coefficients a^2, b^2, c^2 are the *moments of inertia* for rotations about the three body axes: a about $\vec{y}$, b about $\vec{e}_2$, c about the line joining the monopoles.

Key idea: The Atiyah–Hitchin metric

The metric (8.343) is hyperkähler, with the SO(3) action rotating the 2-sphere of complex structures (Figure 8.77), if and only if

$$\frac{2bc}{f}\frac{\mathrm{d}a}{\mathrm{d}r} = (b-c)^2 - a^2, \qquad \frac{2ca}{f}\frac{\mathrm{d}b}{\mathrm{d}r} = (c-a)^2 - b^2, \qquad \frac{2ab}{f}\frac{\mathrm{d}c}{\mathrm{d}r} = (a-b)^2 - c^2. \tag{8.344}$$

The function f is a choice of radial coordinate; Atiyah and Hitchin take $f = -b/r$. The unique solution that is smooth and describes two monopoles at large r has

$$a \to 0, \quad b \to \pi, \quad c \to -\pi \qquad (r \to \pi), \tag{8.345}$$

and, up to exponentially small terms as $r \to \infty$,

$$a \approx b \approx r\sqrt{1 - \tfrac{2}{r}}, \qquad c \approx -\frac{2}{\sqrt{1 - 2/r}}. \tag{8.346}$$

Asymptotically r is the separation of the monopoles. The solution can be written with complete elliptic integrals: $r = 2K(\kappa)$, with κ running from 0 (where $K = \pi/2$, the value $r = \pi$) to 1 ($K \to \infty$).

Why the SO(3) *rotates the complex structures.* In the hyperkähler quotient of Section 8.16.5, the three complex structures correspond to the three components $\mu_i = B_i + D_i\phi$ of the Bogomolny equation, i.e. to the three directions of physical space. A rotation of space rotates (μ_1, μ_2, μ_3) as a vector, and so rotates (I, J, K). The three Kähler forms are therefore not individually SO(3)-invariant; they rotate into one another. Demanding that this rotating triple be closed is what produces the equations (8.344). (If the complex structures were fixed instead, one would get the simpler "Euler" system $2\,(bc)' = \dots$ that describes the Eguchi–Hanson and Taub–NUT metrics.)

Worked example 8.157: The asymptotic metric solves the equations exactly

Set $u = 1 - 2/r$, so $u' = 2/r^2$. Take the asymptotic forms as exact trial functions: $a = b = r\sqrt{u}$, $c = -2/\sqrt{u}$, and $f = -b/r = -\sqrt{u}$. The third equation:

$$\frac{2ab}{f}\,c' \overset{(1)}{=} \frac{2r^2u}{-\sqrt{u}} \cdot \frac{2}{r^2u^{3/2}} \overset{(2)}{=} -\frac{4}{u} \overset{(3)}{=} -c^2 \overset{(4)}{=} (a-b)^2 - c^2.$$

(1) $ab = r^2u$; and $c' = -2 \cdot (-\tfrac{1}{2})u^{-3/2}u' = u^{-3/2} \cdot \tfrac{2}{r^2}$.
(2) Collect powers: $u \cdot u^{-1/2} \cdot u^{-3/2} = u^{-1}$.
(3) $c^2 = 4/u$.
(4) $a = b$.

The first equation (the second is the same, since $a = b$):

$$\frac{2bc}{f}\,a' \overset{(1)}{=} \frac{2r\sqrt{u}\,(-2/\sqrt{u})}{-\sqrt{u}} \cdot \frac{r-1}{r\sqrt{u}} \overset{(2)}{=} \frac{4(r-1)}{u},$$

$$(b-c)^2 - a^2 \overset{(3)}{=} c^2 - 2ac \overset{(4)}{=} \frac{4}{u} + 4r \overset{(5)}{=} \frac{4 + 4r - 8}{u} = \frac{4(r-1)}{u}.$$

(1) $a' = \sqrt{u} + \dfrac{ru'}{2\sqrt{u}} = \dfrac{ru+1}{r\sqrt{u}} = \dfrac{r-1}{r\sqrt{u}}$, using $ru = r - 2$.
(2) $2r\sqrt{u} \cdot (-2/\sqrt{u})/(-\sqrt{u}) = 4r/\sqrt{u}$; multiply.
(3) Expand $(a-c)^2 - a^2$ with $b = a$.
(4) $c^2 = 4/u$ and $-2ac = -2\,r\sqrt{u}\,(-2/\sqrt{u}) = 4r$.
(5) $4r = 4ru/u = 4(r-2)/u$.

Both sides agree. So the asymptotic metric is an exact hyperkähler metric by itself:

$$\mathrm{d}s^2_{\mathrm{TN}} = \left(1 - \frac{2}{r}\right)\left(\mathrm{d}r^2 + r^2(\sigma_1^2 + \sigma_2^2)\right) + \frac{4}{1 - 2/r}\,\sigma_3^2, \tag{8.347}$$

the **Taub–NUT metric with negative mass**. It is singular at $r = 2$, which is why the true Atiyah–Hitchin metric, which departs from it at $r \lesssim 6$, is needed.

Physical picture: What the asymptotic metric says physically

At large r, $\sigma_1^2 + \sigma_2^2 = \mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\varphi^2$ in terms of the direction (θ, φ) of the separation vector, and $\sigma_3 = \mathrm{d}\psi + \cos\theta\,\mathrm{d}\varphi$ with ψ (half) the relative phase. So (8.347) is the flat metric $\mathrm{d}r^2 + r^2\mathrm{d}\Omega^2$ of the relative position, corrected by the factor $1 - 2/r$, plus a phase circle of radius about 2. The correction is a *velocity-dependent* interaction: the kinetic energy is $\frac{1}{2}(1 - 2/r)v^2$, so the force is of order v^2/r^2. Static monopoles do not interact, moving ones do. This is exactly what one gets by treating the two monopoles as point particles carrying magnetic, electric and scalar charge and keeping terms to order v^2 (Manton; Gibbons–Manton). The term $\cos\theta\,\mathrm{d}\varphi$ is a Dirac monopole potential on the sphere of directions: a moving dyon feels the magnetic field of the other monopole.

Worked example 8.158: Regularity at $r = \pi$: the bolt

Near $r = \pi$ put $\rho = r - \pi$. The first equation at $r = \pi$, with $b = \pi$, $c = -\pi$, $f = -b/r = -1$, $a = 0$:

$$\frac{2\pi(-\pi)}{-1}\,a'(\pi) \overset{(1)}{=} (\pi + \pi)^2 - 0 \quad\Longrightarrow\quad a'(\pi) \overset{(2)}{=} 2.$$

(1) Insert the boundary values (8.345) into (8.344).
(2) $2\pi^2 a'(\pi) = 4\pi^2$.

So $a \approx 2\rho$ and, near $r = \pi$, the metric along the rotations about $\vec{e}_1$ is

$$f^2\mathrm{d}r^2 + a^2\sigma_1^2 \approx \mathrm{d}\rho^2 + 4\rho^2\mathrm{d}\psi_1^2 = \mathrm{d}\rho^2 + \rho^2\mathrm{d}\Theta^2, \qquad \Theta = 2\psi_1,$$

where ψ_1 is the angle of rotation about $\vec{e}_1$ ($\sigma_1 = \mathrm{d}\psi_1$ along such rotations). These are polar coordinates on a plane, smooth at $\rho = 0$ *if and only if* Θ has period 2π, i.e. ψ_1 has period π. And indeed the rotation by π about $\vec{e}_1$ is in D_2. The circle of rotations about $\vec{e}_1$ shrinks to a point at $r = \pi$ (like the angular circle at the origin of polar coordinates), while the σ_2, σ_3 directions keep radius π. The set $r = \pi$ is a 2-dimensional surface, the **bolt**. Its points are configurations invariant under all rotations about $\vec{e}_1$: the axially symmetric doughnuts of Example 8.145, labelled by the direction of their axis $\vec{e}_1$ (up to sign).

8.17.2 Head-on collisions: scattering at right angles

Nash's first result: *when two monopoles collide head-on, they emerge along a line at right angles to the line of approach*. For point particles this would be absurd; a head-on collision of billiard balls or of charges sends them straight back. We give two explanations, one from the metric and one from the rational maps.

From the metric: passing through the bolt

A head-on collision has no angular momentum, and by symmetry the motion stays in the family where only r and the rotation angle ψ_1 about $\vec{e}_1 = \vec{y}$ vary: the separation axis $\vec{e}_3$ turns in the plane perpendicular to $\vec{y}$, and $\vec{y}$ stays fixed (Nash: "y is fixed"). On this family the metric is $f^2\mathrm{d}r^2 + a^2\mathrm{d}\psi_1^2$.

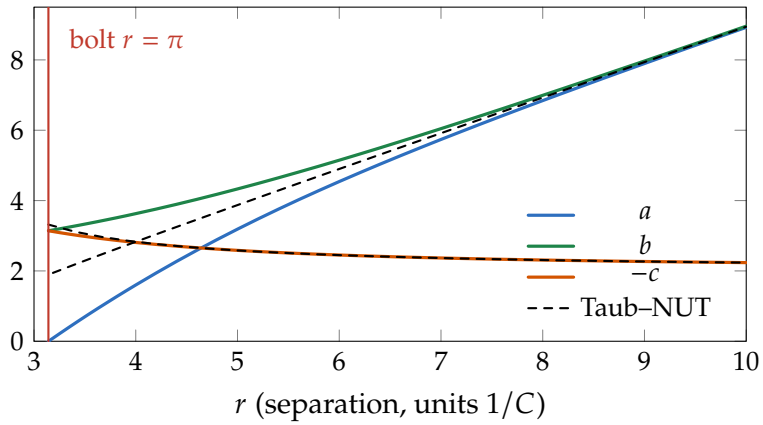


Figure 8.80: The Atiyah–Hitchin coefficients, from a numerical solution of (8.344) starting at the bolt (8.345). At $r = \pi$ the moment of inertia a^2 for rotations about $\vec{e}_1$ vanishes (the doughnut is symmetric about its axis) while $b = -c = \pi$. For $r \gtrsim 6$ the curves merge with the negative-mass Taub–NUT forms (8.346) (dashed): a and b both approach $r\sqrt{1-2/r}$, and $-c$ approaches $2/\sqrt{1-2/r} \to 2$. The differences decay like e^{-r} (at $r = 10$ they are about 0.02, at $r = 20$ about 3×10^{-6}).

Worked example 8.159: Why a radial geodesic turns by 90°

With zero angular momentum the geodesic is radial: ψ_1 is constant and r decreases to π. What happens next is seen in the smooth coordinates of Example 8.158:

$$\begin{aligned} \text{incoming ray: } \Theta = \Theta_0,\ \rho \downarrow 0 &\overset{(1)}{\Longrightarrow} \text{outgoing ray: } \Theta = \Theta_0 + \pi,\ \rho \uparrow \\ &\overset{(2)}{\Longrightarrow} \psi_1 \mapsto \psi_1 + \tfrac{\pi}{2} \\ &\overset{(3)}{\Longrightarrow} \vec{e}_3 \text{ rotated by } 90^\circ \text{ about } \vec{y}. \end{aligned}$$

(1) Near $\rho = 0$ the metric is the flat plane $\mathrm{d}\rho^2 + \rho^2 \mathrm{d}\Theta^2$; a straight line through the origin of a plane leaves along the opposite ray.
(2) $\Theta = 2\psi_1$.
(3) ψ_1 is the angle by which the frame is rotated about $\vec{e}_1$, and $\vec{e}_3$ is the direction of separation.

In the coordinate ψ_1 the geodesic looks as if it bent by 90°; in the smooth coordinate Θ it went straight. The "bending" is the halving $\psi_1 = \Theta/2$, i.e. the fact that the monopoles are identical and the configuration at $\psi_1 + \pi$ is the same as at ψ_1. The result does not depend on the incoming relative phase: every head-on geodesic reaches the bolt.

From rational maps: the poles rotate

The rational maps of Section 8.15.7 give a second, very concrete picture. Recall that the poles of the rational map mark (approximately, for well-separated monopoles) the positions of the monopoles projected on the (x^1, x^2)-plane, with $z = x^1 + ix^2$.

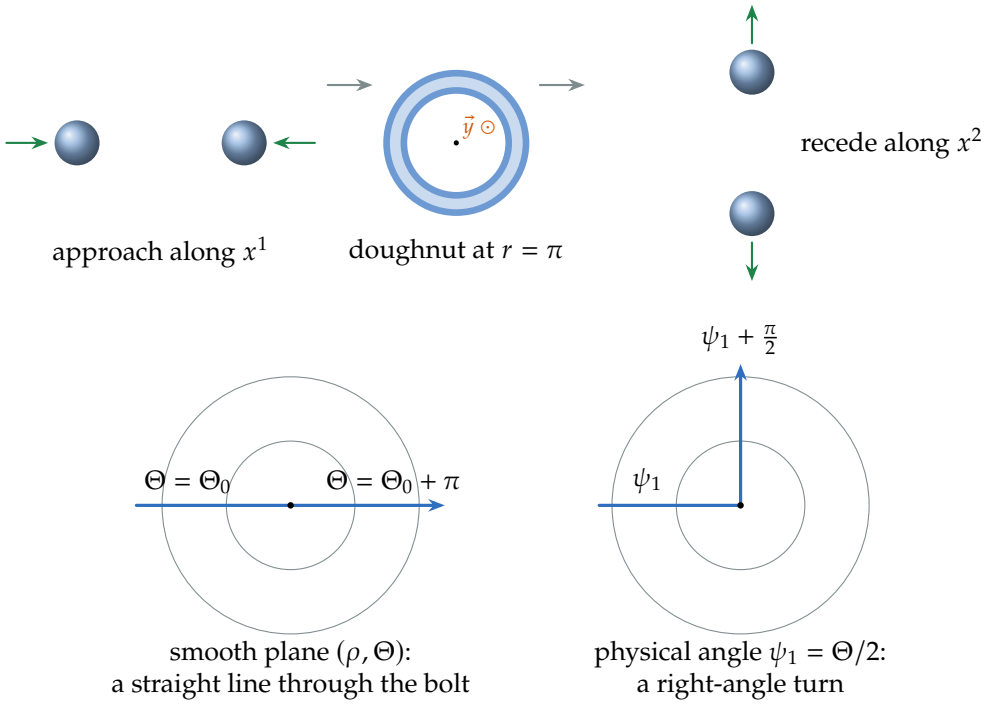


Figure 8.81: Right-angle scattering. Top: two monopoles approach head-on along the x^1-axis, merge into the axially symmetric doughnut (whose axis $\vec{y}$ points out of the page), and separate along the x^2-axis. Bottom: in the smooth coordinates (ρ, Θ) near the bolt the geodesic is a straight line through the origin (left); since the physical angle is $\psi_1 = \Theta/2$, the direction of separation turns by 90° (right).

Worked example 8.160: The family $f_\beta = 1/(z^2 - \beta)$

Consider the one-parameter family of based rational maps of degree 2

$$f_\beta(z) = \frac{1}{z^2 - \beta}, \qquad \beta \in \mathbb{R}.$$

Every member is a legitimate point of Rat_2: numerator 1 and denominator $z^2 - \beta$ have no common root. The poles are the roots of $z^2 = \beta$:

$$\begin{aligned} \beta = +\varepsilon > 0: &\quad z \overset{(1)}{=} \pm\sqrt{\varepsilon} && \text{(monopoles on the } x^1\text{-axis)}, \\ \beta = 0: &\quad z \overset{(2)}{=} 0 \text{ (double)} && \text{(the doughnut } 1/z^2\text{)}, \\ \beta = -\varepsilon < 0: &\quad z \overset{(3)}{=} \pm i\sqrt{\varepsilon} = e^{i\pi/2}\left(\pm\sqrt{\varepsilon}\right) && \text{(monopoles on the } x^2\text{-axis)}. \end{aligned}$$

(1) Real square roots.
(2) $z^2 = 0$.
(3) $\sqrt{-\varepsilon} = i\sqrt{\varepsilon}$, and multiplication by $i = e^{i\pi/2}$ is rotation by 90°.

As β decreases smoothly through 0, the pair of poles comes together along the real axis and leaves along the imaginary axis. A smooth path in the moduli space through the doughnut *is* a right-angle scattering. (The family is picked out by symmetry: it consists of the maps invariant under $z \mapsto -z$ and under complex conjugation, and a fixed set of isometries is totally geodesic, so it contains the head-on geodesic.) Compare with two point particles: their relative position $\pm\sqrt{\varepsilon}$ would have to pass through 0 and continue to $\mp\sqrt{\varepsilon}$, which is the *same* unordered pair, i.e. back-scattering or no scattering. The smooth coordinate on

$\mathcal{M}_2^0$ near coincidence is $\beta = z_1 z_2$-like, a symmetric function of the pole positions, not the positions themselves.

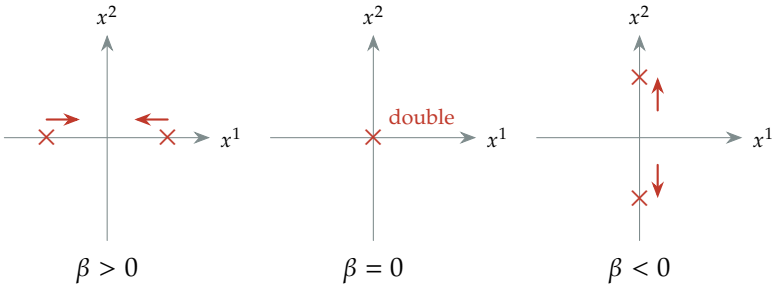


Figure 8.82: The poles of $f_\beta = 1/(z^2 - \beta)$ as β decreases through zero. They meet at the origin (the doughnut) and separate at right angles. The path in the space of rational maps is perfectly smooth: β is a good coordinate, the positions $\pm\sqrt{\beta}$ are not.

8.17.3 Scattering angle and impact parameter

Now give the collision an impact parameter L. We stay in the family of the previous subsection ($\vec{y}$ fixed, motion in the plane perpendicular to it), which is still consistent by symmetry. The metric on it is $f^2 \mathrm{d}r^2 + a^2 \mathrm{d}\psi_1^2$.

Worked example 8.161: The orbit equation

With unit speed and angular momentum ℓ, the Lagrangian $\frac{1}{2}(f^2\dot{r}^2 + a^2\dot{\psi}_1^2)$ gives

$$\ell = a^2\dot{\psi}_1, \qquad f^2\dot{r}^2 + a^2\dot{\psi}_1^2 = 1 \overset{(1)}{\Longrightarrow} \dot{r} = \pm\frac{1}{|f|}\sqrt{1 - \frac{\ell^2}{a^2}}$$
$$\overset{(2)}{\Longrightarrow} \frac{\mathrm{d}\psi_1}{\mathrm{d}r} = \frac{\dot{\psi}_1}{\dot{r}} = \pm\frac{\ell\,|f|}{a\sqrt{a^2 - \ell^2}}.$$

(1) ψ_1 is cyclic, so its momentum ℓ is conserved; substitute $\dot{\psi}_1 = \ell/a^2$ into the energy (speed) equation and solve for $\dot{r}$.

(2) Divide; $\sqrt{1 - \ell^2/a^2} = \sqrt{a^2 - \ell^2}/a$.

The monopoles come in from $r = \infty$ to the turning point r_0 with $a(r_0) = \ell$ and go out again, so the total rotation of the separation axis and the **scattering angle** are

$$\Delta\psi_1 = 2\int_{r_0}^{\infty} \frac{\ell\,|f|\,\mathrm{d}r}{a\sqrt{a^2 - \ell^2}}, \qquad \Theta = \pi - \Delta\psi_1. \tag{8.348}$$

With no interaction the separation vector turns by π (it points from one monopole to the other before and after) and $\Theta = 0$. At large r, $a \approx r$ and $|f| \approx 1$, so $\ell = r^2\dot{\psi}_1$ is the impact parameter times the speed: with unit speed, $\ell = L$.

Worked example 8.162: Taub–NUT scattering is Rutherford scattering

For the asymptotic metric (8.347) restricted to the plane, $(1 - 2/r)(\mathrm{d}r^2 + r^2\mathrm{d}\psi_1^2)$, the integral can be done. Let $u = 1/r$ and $u' = \mathrm{d}u/\mathrm{d}\psi_1$. The conserved quantities are $\ell = (1 - 2u)r^2\dot{\psi}_1$

and $(1-2u)(\dot r^2 + r^2\dot\psi_1^2) = 1$. Then

$$1 \overset{(1)}{=} (1-2u)\Big(\frac{\ell^2 u'^2}{(1-2u)^2} + \frac{\ell^2 u^2}{(1-2u)^2}\Big) \overset{(2)}{\Longrightarrow} \ell^2\big(u'^2+u^2\big) = 1-2u$$

$$\overset{(3)}{\Longrightarrow} u''+u = -\frac{1}{\ell^2} \overset{(4)}{\Longrightarrow} u = A\cos(\psi_1-\psi_0) - \frac{1}{\ell^2}, \qquad A = \frac{\sqrt{1+\ell^2}}{\ell^2}.$$

(1) $\dot\psi_1 = \ell u^2/(1-2u)$ and $\dot r = -u^{-2}u'\dot\psi_1 = -\ell u'/(1-2u)$; also $r^2\dot\psi_1^2 = \ell^2u^2/(1-2u)^2$.
(2) Multiply by $(1-2u)$.
(3) Differentiate with respect to ψ_1 and divide by $2\ell^2 u'$.
(4) General solution of the linear equation; A follows by inserting it back into step 2: $\ell^2A^2 - 2A\cos + \ell^{-2} = 1 - 2A\cos + 2\ell^{-2}$, so $\ell^2A^2 = 1+\ell^{-2}$.

This is a hyperbola: the orbit of a repulsive Coulomb problem. The asymptotes are at $u=0$, i.e. $\cos(\psi_1-\psi_0) = 1/(A\ell^2) = 1/\sqrt{1+\ell^2}$, so $\Delta\psi_1 = 2\arctan\ell$ and, with $\ell = L$,

$$\boxed{\tan\frac{\Theta_{\text{TN}}}{2} = \frac{1}{L}}$$

Rutherford's formula for a repulsive $1/r$ potential. Note that it is independent of the speed: the force is proportional to v^2, so slower monopoles are deflected by the same angle along the same orbit, only more slowly.

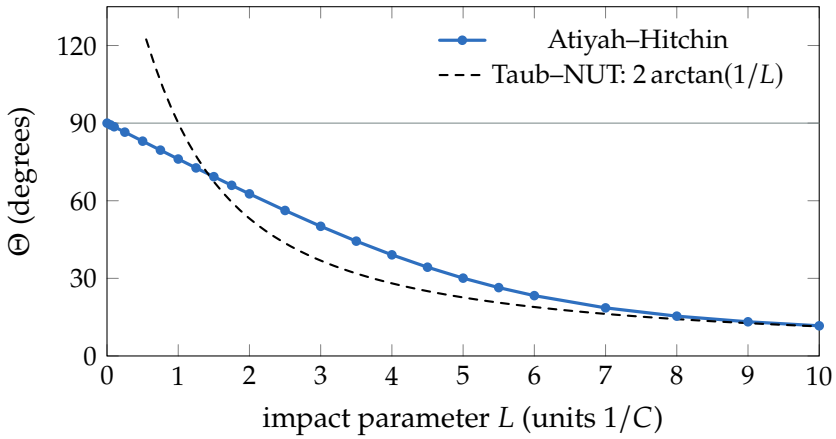


Figure 8.83: Scattering angle against impact parameter for geodesics in the plane perpendicular to $\vec y$, from numerical integration of (8.348) with the Atiyah–Hitchin coefficients. Head-on ($L\to 0$) collisions scatter by 90°. For large L the curve approaches the Rutherford formula of the asymptotic metric (dashed), which would instead give back-scattering (180°) at $L=0$: the 90° is entirely a core effect. Values: $\Theta = 76.1°$ at $L=1$, $62.7°$ at $L=2$, $30.1°$ at $L=5$, $11.7°$ at $L=10$.

8.17.4 Glancing collisions: dyons and spin

Nash's second result: *if the monopoles approach along parallel lines close enough together, the emerging monopoles have spin angular momentum; they are electrically charged, i.e. they have become dyons.* The planar family we used above has the orbital angular momentum along $\vec y = \vec e_1$. In general the incoming orbital angular momentum makes some angle α with $\vec y$ in the plane perpendicular to the line of approach, and then the motion is genuinely four-dimensional.

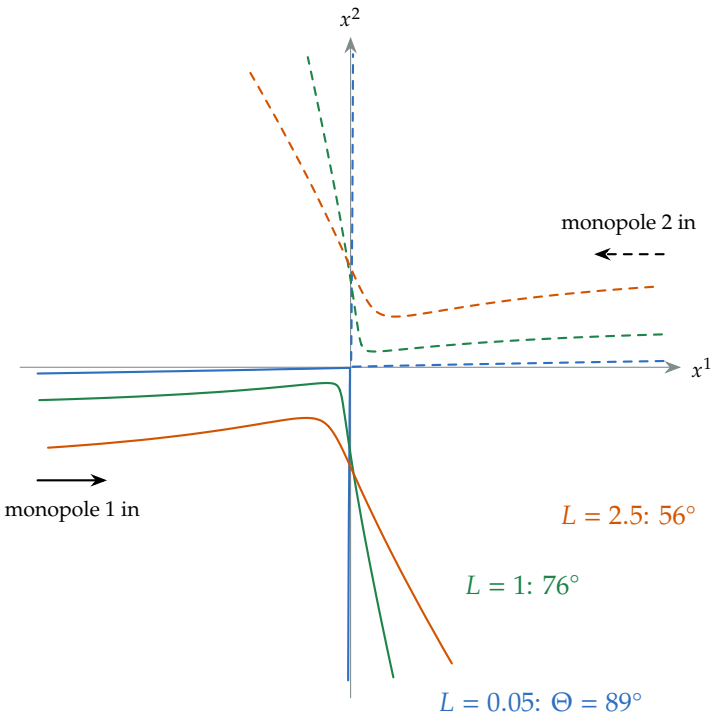


Figure 8.84: Paths of the two monopoles (solid: monopole 1, entering from the left; dashed: monopole 2, entering from the right) for three impact parameters, computed from the Atiyah–Hitchin geodesics. The plotted distance from the origin is $(r - \pi)/2$, a schematic "half-separation" that is exact at large r up to a shift and shrinks to zero at the doughnut; near the centre the monopoles are not separate objects. As $L \to 0$ the paths approach the right-angle cross.

Worked example 8.163: Euler's equations and the exchange of charge

Write the velocity in terms of the body angular velocities $\omega_i = \sigma_i(\dot{R})$. The Lagrangian is that of a rigid body whose moments of inertia $(I_1, I_2, I_3) = (a^2, b^2, c^2)$ depend on r:

$$L = \tfrac{1}{2} f^2 \dot{r}^2 + \tfrac{1}{2}\left(a^2\omega_1^2 + b^2\omega_2^2 + c^2\omega_3^2\right).$$

The body angular momenta $M_i = I_i\omega_i$ obey Euler's equations $\dot{\vec{M}} = \vec{M} \times \vec{\omega}$ (the time dependence of I_i enters only through $\omega_i = M_i/I_i$). In particular

$$\dot{M}_3 \overset{(1)}{=} M_1\omega_2 - M_2\omega_1 \overset{(2)}{=} M_1 M_2\left(\frac{1}{b^2} - \frac{1}{a^2}\right),$$
$$\frac{\mathrm{d}}{\mathrm{d}t}|\vec{M}|^2 \overset{(3)}{=} 2\vec{M}\cdot(\vec{M}\times\vec{\omega}) \overset{(4)}{=} 0.$$

(1) Third component of $\vec{M} \times \vec{\omega}$.
(2) $\omega_i = M_i/I_i$ with $I_1 = a^2$, $I_2 = b^2$.
(3) Differentiate $\vec{M}\cdot\vec{M}$ and use Euler's equation.
(4) $\vec{M} \times \vec{\omega}$ is perpendicular to $\vec{M}$.

Interpretation. M_3 is the angular momentum about the line joining the monopoles. Far apart, the σ_3 direction is the relative phase (Example 8.157), so M_3 is the momentum conjugate to the relative phase: by Example 8.152, the *relative electric charge*. $|\vec{M}|$ is the total angular momentum. Initially $M_3 = 0$ (uncharged monopoles, angular momentum purely orbital, perpendicular to the line joining them). Far away $a = b$ up to exponentially small

terms, so $\dot{M}_3 \approx 0$: charge is conserved while the monopoles are apart. In the core $a \neq b$ (Figure 8.80), and M_3 changes unless $M_1 M_2 = 0$, i.e. unless $\alpha = 0$ or 90° (the two planar families). So generic close encounters convert orbital angular momentum into relative charge.

Numerically integrating the full geodesic equations (Euler's equations plus the radial equation, with the Atiyah–Hitchin coefficients) from separation 30 gives the following fractions $|M_3|/|\vec{M}|$ of the total angular momentum that ends up as relative charge:

impact parameter L	$\alpha = 20^\circ$	$\alpha = 45^\circ$	$\alpha = 70^\circ$
1	0.32	0.68	0.93
2	0.26	0.58	0.88
3	0.18	0.40	0.61
4	0.10	0.21	0.22
6	0.02	0.04	0.03

The effect is large for close encounters and dies off rapidly (exponentially) with the impact parameter, because it is driven by $1/b^2 - 1/a^2$, which is non-zero only in the core. This is Nash's "close enough together".

Worked example 8.164: Why charge means spin: the angular momentum of a charge–monopole pair

Why does Nash call the outgoing angular momentum "spin"? Because a pair of dyons stores angular momentum in its electromagnetic field, even at rest. Take a point charge q at the origin and a point magnetic charge g at $\vec{d}$. The field angular momentum is $\vec{J} = \int \vec{x} \times (\vec{E} \times \vec{B})\, \mathrm{d}^3x$ with $\vec{E} = q\vec{x}/(4\pi r^3)$. Then

$$\vec{J} \overset{(1)}{=} \frac{q}{4\pi} \int \frac{1}{r^3} \big(\vec{x}(\vec{x}\cdot\vec{B}) - r^2\vec{B}\big)\mathrm{d}^3x \overset{(2)}{=} -\frac{q}{4\pi}\int (\vec{B}\cdot\nabla)\hat{\vec{x}}\,\mathrm{d}^3x$$
$$\overset{(3)}{=} \frac{q}{4\pi}\int (\nabla\cdot\vec{B})\,\hat{\vec{x}}\,\mathrm{d}^3x - \frac{q}{4\pi}\oint_\infty \hat{\vec{x}}\,(\vec{B}\cdot\hat{\vec{n}})\,\mathrm{d}S \overset{(4)}{=} \frac{qg}{4\pi}\,\hat{\vec{d}}.$$

(1) $\vec{x}\times(\vec{E}\times\vec{B}) = \vec{E}(\vec{x}\cdot\vec{B}) - \vec{B}(\vec{x}\cdot\vec{E})$ with $\vec{E} \propto \vec{x}$.

(2) $(\vec{B}\cdot\nabla)\hat{x}_i = B_j\partial_j(x_i/r) = (r^2 B_i - x_i\,\vec{x}\cdot\vec{B})/r^3$.

(3) Integrate by parts: $(\vec{B}\cdot\nabla)\hat{\vec{x}} = \nabla\cdot(\vec{B}\,\hat{\vec{x}}) - (\nabla\cdot\vec{B})\hat{\vec{x}}$, and use Gauss's theorem.

(4) $\nabla\cdot\vec{B} = g\,\delta^3(\vec{x}-\vec{d})$ gives $g\,\hat{\vec{d}}$; the surface term is $\frac{g}{4\pi}\oint \hat{\vec{x}}\,\mathrm{d}\Omega = 0$ since $\vec{B} \approx g\hat{\vec{x}}/(4\pi r^2)$ at infinity.

The answer does not depend on the distance $|\vec{d}|$ and points along the line joining the two. For two dyons with equal magnetic charge g and electric charges q_1, q_2, the two cross terms give $\vec{J} = \frac{g}{4\pi}(q_1 - q_2)\hat{\vec{n}}$ along the line joining them (a dyon's own $\vec{E}$ and $\vec{B}$ are parallel and carry no angular momentum). So *relative charge is spin about the line of centres*, exactly the M_3 of Example 8.163. Quantum mechanically this angular momentum must be a multiple of $\hbar/2$, which is Dirac's quantisation condition $qg/4\pi \in \frac{1}{2}\mathbb{Z}$ of Section 8.12.3, seen from the angular-momentum side.

Reading Nash (pp. 255–256)

Nash describes the $k = 2$ parameters (separation $2|x|$, a unit vector $y \perp x$ for the relative phase, $y \sim -y$) and states two results: head-on collisions in which "y is fixed" give scattering at right angles, and collisions along close parallel lines produce dyons with spin. His phrase "the scattering occurs in the x–y plane" is best read as the coordinate plane perpendicular to the fixed vector $\vec{y}$ (if $\vec{y}$ were in the plane of motion it could not stay perpendicular to $\vec{x}$). We supplied the frame description and the D_2 quotient, the metric ansatz and the Atiyah–Hitchin equations with their boundary conditions, the exact Taub–NUT check, the bolt and the 90° derivation, the rational-map picture, the scattering-angle formula with numbers and plots, Rutherford scattering for the asymptotic metric, Euler's equations for charge exchange with numbers, and the field angular momentum of a charge–monopole pair.

Recap

- $\mathcal{M}_2^0$ is 4-dimensional: separation plus an orthonormal frame modulo D_2. SO(3) acts isometrically, so the metric is $f^2\mathrm{d}r^2 + a^2\sigma_1^2 + b^2\sigma_2^2 + c^2\sigma_3^2$.
- Hyperkähler with rotating complex structures $\iff$ the Atiyah–Hitchin ODEs (8.344); smoothness gives the bolt at $r = \pi$ (the doughnuts), and at large r the metric is negative-mass Taub–NUT (velocity-dependent forces).
- Head-on collisions pass through the bolt and scatter by 90°; in rational-map language the poles $\pm\sqrt{\beta}$ rotate by 90° as β changes sign.
- The scattering angle falls from 90° at $L = 0$ to the Rutherford value $2\arctan(1/L)$ at large L.
- Off the planar families, Euler's equations $\dot{M}_3 = M_1 M_2(b^{-2} - a^{-2})$ convert orbital angular momentum into relative charge in the core: glancing collisions make dyons, whose field angular momentum $g(q_1 - q_2)/4\pi$ is the "spin".

8.18 Critical point theory and gauge theories

Everything we have solved in this chapter has been a *minimum*. Instantons are the absolute minima of the Yang–Mills action $S(A)$ in each topological sector, because $S \geq 8\pi^2|k|$ with equality exactly when $F = \pm \star F$. BPS monopoles are the absolute minima of the Yang–Mills–Higgs energy $E(A, \phi)$ in each sector, because $E \geq 4\pi C|k|$ with equality exactly on the Bogomolny equation (Section 8.13.3). In both cases the first-order equation was the *condition for saturating a bound*, and the bound came from completing a square.

But the field equations are second order, and they may have other solutions: saddle points of the action or the energy, which are not minima. In a mechanical system these are the *unstable* equilibria, the top of a pass between two valleys. In quantum field theory they matter a great deal. A saddle of the energy with one unstable direction is the top of the barrier that a tunnelling or thermally activated process must cross. The electroweak *sphaleron*, which mediates baryon-number violation in the early universe, is exactly such a saddle. So Nash asks:

> *Do gauge-theory functionals have non-minimal critical points? How many, of what type (index), and what topological mechanism forces them to exist?*

The tool is Morse theory (Chapter 7): the topology of a space forces a function on it to have

critical points, and the number and indices of those critical points are bounded below by the Betti numbers. The difficulty is that our spaces are infinite-dimensional and have an infinite-dimensional symmetry, the gauge group. We first recall the finite-dimensional theory, then see what breaks in infinite dimensions and how to repair it. Then we treat Nash's three cases in turn: Yang–Mills–Higgs on $\mathbb{R}^3$ (a good Morse function, by Taubes), Yang–Mills on a Riemann surface (an equivariantly perfect Morse function, by Atiyah and Bott), and Yang–Mills on S^4 (no Morse theory at the time Nash wrote, because of conformal invariance).

8.18.1 Morse theory in finite dimensions: a reminder

Let $f : M \to \mathbb{R}$ be a smooth function on a compact manifold of dimension n. A **critical point** is a point x where $\mathrm{d}f = 0$. There the second derivatives form a symmetric matrix, the **Hessian** $H_{ij} = \partial_i \partial_j f(x)$; it is independent of coordinates at a critical point (Chapter 7). The critical point is **non-degenerate** if $\det H \neq 0$, and its **index** $\lambda(x)$ is the number of negative eigenvalues of H: the number of independent directions in which f goes *down*.

Working definition 8.47 (Morse series, Poincaré series, perfect functions)

- Make a list of the critical points $x_1, x_2, \dots$ of f and their indices. The **Morse series** is $M_t(f) = \sum_i t^{\lambda(x_i)}$: the coefficient of t^λ counts the critical points of index λ.
- The **Poincaré series** of M is $P_t(M) = \sum_p b_p\, t^p$, with $b_p = \dim H^p(M)$ the Betti numbers.
- The **Morse inequalities** say $M_t(f) - P_t(M) = (1+t)Q_t$, where Q_t has non-negative coefficients. In particular (setting t to pick coefficients) the number of critical points of index p is at least b_p.
- f is **perfect** if $M_t(f) = P_t(M)$: it has the smallest number of critical points the topology allows.
- If the critical points come in non-degenerate critical manifolds [submanifolds made of critical points, with the Hessian invertible in the directions normal to them] N_j (Morse–Bott case), N_j of index λ_j contributes $t^{\lambda_j} P_t(N_j)$ to M_t.
- If a group G acts and f is invariant, the **equivariant** versions replace $P_t(M)$ by $P_t^G(M) = P_t(M_G)$, the Poincaré series of the homotopy quotient $M_G = M \times_G EG$ [$M \times EG$ divided by G acting on both factors, which G does freely because it acts freely on EG], and a critical orbit $G \cdot x$ with stabiliser [the elements g with $g \cdot x = x$] G_x contributes $t^\lambda P_t(BG_x)$ [BG_x is the classifying space of G_x].

The mechanism behind all of this is the picture of Chapter 7: the sublevel sets $M^a = \{f \leq a\}$ change topology only when a crosses a critical value, and crossing a critical point of index λ glues on a λ-dimensional cell. A space with a rich topology needs many cells, hence many critical points.

Worked example 8.165: Two perfect functions

The upright torus. The height function on a torus standing on its rim has four critical points: a minimum (index 0), two saddles (index 1) and a maximum (index 2). So

$$M_t(f) = 1 + 2t + t^2 = (1+t)^2 = P_t(T^2),$$

and the height function is perfect.

The rotating sphere. Let $U(1)$ rotate S^2 about the z-axis and let $f = z$. The critical points are the poles, a minimum and a maximum, each fixed by the whole $U(1)$. The equivariant Morse series is

$$M_t^{U(1)}(f) \overset{(1)}{=} t^0 P_t(BU(1)) + t^2 P_t(BU(1)) \overset{(2)}{=} \frac{1+t^2}{1-t^2} \overset{(3)}{=} P_t(S^2)\, P_t(BU(1)) \overset{(4)}{=} P_t^{U(1)}(S^2).$$

(1) Each pole is a critical orbit (a single point) with stabiliser $U(1)$ and index 0 or 2.
(2) $BU(1) = \mathbb{CP}^\infty$ has one cell in each even dimension, so $P_t(BU(1)) = 1 + t^2 + t^4 + \cdots = 1/(1-t^2)$.
(3) $P_t(S^2) = 1 + t^2$.
(4) The fibration $S^2 \to S^2_{U(1)} \to BU(1)$ has a spectral sequence that collapses (all cohomology is even), so the Poincaré series multiply (Chapter 7).

So f is *equivariantly* perfect. This little example is the model for the Atiyah–Bott theorem of Section 8.18.9: a function whose critical points are fixed by a symmetry group, each contributing a copy of the cohomology of a classifying space.

8.18.2 What changes in infinite dimensions

In gauge theory the "manifold" is a space of fields, such as the space $\mathcal{A}$ of connections, and the "function" is an action or energy. Three things can go wrong when we copy the finite-dimensional story.

1. The Hessian is a differential operator. In finite dimensions the Hessian is a matrix and its index is automatically finite. In field theory the second variation of an action is a differential operator, with infinitely many eigenvalues. The index (the number of negative eigenvalues) is finite only if the spectrum is discrete and bounded below with finitely many negative values. This is what *ellipticity* guarantees (Chapter 2): an elliptic second-order operator such as $-\nabla^2 + V$ on a compact space has discrete eigenvalues $\mu_1 \le \mu_2 \le \cdots \to +\infty$.

Worked example 8.166: The index of a path: the harmonic oscillator

Take the action of a harmonic oscillator on paths with fixed ends,

$$\mathcal{S}[x] = \frac{1}{2}\int_0^T \left(\dot{x}^2 - \omega^2 x^2\right)\mathrm{d}t, \qquad x(0) = x(T) = 0.$$

The path $x = 0$ is a critical point. Expanding $x = 0 + \varepsilon\,\xi$,

$$\mathcal{S}[\varepsilon\xi] \overset{(1)}{=} \frac{\varepsilon^2}{2}\int_0^T \left(\dot{\xi}^2 - \omega^2\xi^2\right)\mathrm{d}t \overset{(2)}{=} \frac{\varepsilon^2}{2}\int_0^T \xi\left(-\frac{\mathrm{d}^2}{\mathrm{d}t^2} - \omega^2\right)\xi\,\mathrm{d}t \overset{(3)}{=} \frac{\varepsilon^2}{2}\sum_{n\ge 1}\left(\frac{n^2\pi^2}{T^2} - \omega^2\right)c_n^2\,\frac{T}{2}.$$

(1) $\mathcal{S}$ is quadratic, so the expansion stops at order ε^2.
(2) Integrate $\dot{\xi}^2$ by parts; the boundary terms vanish because $\xi(0) = \xi(T) = 0$.
(3) Expand $\xi = \sum_n c_n \sin(n\pi t/T)$, the eigenfunctions of $-\mathrm{d}^2/\mathrm{d}t^2$ with these boundary conditions, and use $\int_0^T \sin^2(n\pi t/T)\mathrm{d}t = T/2$.

The Hessian $-\mathrm{d}^2/\mathrm{d}t^2 - \omega^2$ has eigenvalues $\mu_n = n^2\pi^2/T^2 - \omega^2$, infinitely many, but only finitely many are negative: the index is the number of $n \ge 1$ with $n\pi < \omega T$. Each time ωT passes a multiple of π (a *conjugate point* of the classical path, Chapter 7) the index jumps by one. In the path integral each negative mode contributes a factor i to the square root of the fluctuation determinant: the Maslov phase.



Figure 8.85: The spectrum of the Hessian $-\mathrm{d}^2/\mathrm{d}t^2 - \omega^2$ of Example 8.166 for $\omega T = 7$, in units of $\omega^2/4$. The eigenvalues march off to $+\infty$, so the operator has infinitely many directions but only finitely many (here two, red) negative ones. This is what ellipticity buys in general: a finite index.

2. Compactness can fail. On a compact manifold a function attains its minimum, and a sequence of points along which f becomes "more and more critical" has a convergent subsequence. On an infinite-dimensional or non-compact space this can fail, and critical points that the topology "ought" to force can escape to infinity.

Working definition 8.48 (The Palais–Smale condition)

A **Palais–Smale sequence** for f is a sequence x_n with $f(x_n)$ bounded and $|\nabla f(x_n)| \to 0$: the points look more and more like critical points. The function satisfies the **Palais–Smale condition** (Palais and Smale's "condition (C)") if every such sequence has a convergent subsequence [infinitely many of the x_n, taken in order, that converge to a point of the space]. *Test:* look for a way to send f to a finite value with vanishing gradient while the point runs off to infinity. If you find one, the condition fails.

The theorem of Palais and Smale (the "infinite-dimensional generalisation" Nash cites) is: *a smooth function on a complete Hilbert manifold that is bounded below and satisfies condition (C), whose critical points are non-degenerate with finite index, obeys the Morse inequalities.* The proof is the finite-dimensional one: the gradient flow deforms sublevel sets, and condition (C) is exactly what stops the flow from running off to infinity without reaching a critical point.

Worked example 8.167: Two ways to fail condition (C)

(a) A minimum at infinity. On $\mathbb{R}$ let $f(x) = e^{-x}$. It is bounded below by 0 but never reaches 0. The sequence $x_n = n$ has

$$f(x_n) = e^{-n} \to 0, \qquad f'(x_n) = -e^{-n} \to 0,$$

and no convergent subsequence. The "missing critical point" sits at $x = +\infty$.
(b) A mountain pass with no pass. On $\mathbb{R}^2$ let $f(x, y) = x^2 + (1 - x)^3 y^2$. Its critical points:

$$\partial_y f = 2(1 - x)^3 y = 0 \overset{(1)}{\Longrightarrow} x = 1 \text{ or } y = 0,$$
$$\partial_x f = 2x - 3(1 - x)^2 y^2 \overset{(2)}{\Longrightarrow} (x, y) = (0, 0) \text{ only}.$$

(1) A product vanishes when one factor does.
(2) At $x = 1$ we get $\partial_x f = 2 \neq 0$; at $y = 0$ we need $2x = 0$.

The origin is a strict local minimum ($f \approx x^2 + y^2$ there), and $f(4, 1) = 16 - 27 = -11 < 0$. So the landscape has a valley at the origin surrounded by a ring of positive values, and lower ground beyond it: exactly the geometry in which one expects a pass. Yet there is no second critical point. The would-be pass has escaped to infinity: along paths from $(0, 0)$ to $(4, 1)$ that go out to large $|y|$ near $x = 1$, the maximum height can be made to approach its infimum without any point where $\nabla f = 0$.

3. The symmetry group is infinite-dimensional. Gauge theory functionals are invariant under the gauge group $\mathcal{G}$. Every critical point sits on a whole orbit of critical points, so it is never non-degenerate: the Hessian has a zero eigenvalue for every infinitesimal gauge transformation. We deal with this in Section 8.18.3.

The mountain pass and min-max

The most useful tool for finding non-minimal critical points is the **min-max principle**. Suppose two points p, q lie in two different valleys. Among all paths γ from p to q, consider the highest point of each path and take the lowest such highest point:

$$c = \inf_{\gamma : p \to q} \max_{s \in [0,1]} f(\gamma(s)). \tag{8.349}$$

This is what a hiker does: choose the route whose highest point is as low as possible. The result (the **mountain pass theorem** of Ambrosetti and Rabinowitz): *if c is strictly larger than $f(p)$ and $f(q)$, and f satisfies condition (C), then c is a critical value*, attained at a critical point that is typically of index one. Why: if there were no critical point at height c, the downhill gradient flow would push every path whose maximum is slightly above c to a path whose maximum is below c, contradicting the definition of c. Condition (C) is what makes the flow push by a definite amount near height c; Example 8.167(b) is the counterexample when it fails.

More generally one takes the inf-max over any family of paths, loops or higher spheres that cannot be deformed away, for instance all loops in a fixed non-contractible class. This is Lusternik–Schnirelmann theory (Chapter 7), and it is how Taubes proved the existence of non-minimal monopole solutions.

Worked example 8.168: A mountain pass computed

Let $f(x, y) = (x^2 - 1)^2 + y^2$, with minima $p = (-1, 0)$, $q = (1, 0)$ where $f = 0$. Every path from p to q must cross the line $x = 0$, at some point $(0, y_*)$. Hence

$$\max_s f(\gamma(s)) \overset{(1)}{\geq} f(0, y_*) \overset{(2)}{=} 1 + y_*^2 \overset{(3)}{\geq} 1,$$

(1) The maximum over a path is at least the value at any one of its points.
(2) Substitute $x = 0$.
(3) $y_*^2 \geq 0$.

So $c \geq 1$. The straight path $y = 0$ has maximum $f(0, 0) = 1$, so $c = 1$. The min-max point is $(0, 0)$. Its Hessian is

$$H = \begin{pmatrix} \partial_x^2 f & \partial_x \partial_y f \\ \partial_y \partial_x f & \partial_y^2 f \end{pmatrix}_{(0,0)} = \begin{pmatrix} 12x^2 - 4 & 0 \\ 0 & 2 \end{pmatrix}_{x=0} = \begin{pmatrix} -4 & 0 \\ 0 & 2 \end{pmatrix},$$

with one negative eigenvalue: index one, as promised. The negative direction (x) points along the path, downhill towards the two valleys.

8.18.3 Dividing by the gauge group

A gauge-invariant functional f on a space X of fields has degenerate critical points: if x is critical, so is every $g \cdot x$. Differentiating $f(g_t \cdot x) = f(x)$ twice at a critical point shows that the infinitesimal gauge directions are zero modes of the Hessian. Nash's remark that "large gauge

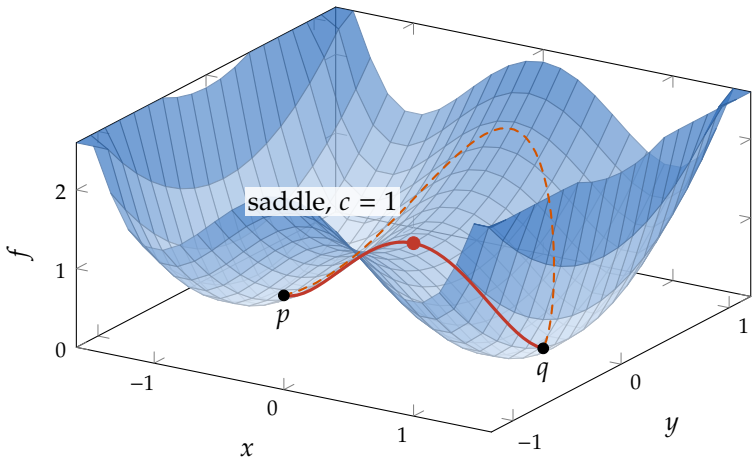


Figure 8.86: The mountain pass of Example 8.168, $f = (x^2 - 1)^2 + y^2$. Two valleys p, q. Every path from one to the other has a highest point; the dashed (orange) path's highest point is higher than that of the solid (red) path, whose highest point is the saddle at height $c = 1$. The saddle is the lowest possible highest point: the min-max critical point, of index one.

invariance . . . will have to be factored out (equivariantly or otherwise)" refers to two ways of dealing with this.

Key idea: Two ways to factor out the gauge group

1. *Quotient by a group that acts freely.* Replace X by $X/\mathcal{G}_0$, where $\mathcal{G}_0$ is a subgroup that acts without fixed points, typically the **based gauge group** of gauge transformations with $g(x_0) = \mathbb{1}$ at a base point x_0 (for us, the point at infinity). Then $X/\mathcal{G}_0$ is (morally) a manifold and f descends to it.

2. *Work equivariantly.* Keep the full group $\mathcal{G}$, including the constant gauge transformations that do have fixed points, and use the equivariant cohomology $H^*_{\mathcal{G}}(X) = H^*(X \times_{\mathcal{G}} E\mathcal{G})$ and the equivariant Morse series of Definition 8.47.

Worked example 8.169: Based gauge transformations act freely

Suppose a gauge transformation g fixes a connection: $g^{-1}Ag + g^{-1}\mathrm{d}g = A$. Then

$$g^{-1}Ag + g^{-1}\mathrm{d}g = A \overset{(1)}{\iff} \mathrm{d}g + Ag - gA = 0 \overset{(2)}{\iff} \mathrm{d}_A g = 0,$$

(1) Multiply on the left by g.
(2) Regard g as a section of the bundle of automorphisms, on which A acts by commutator: $\mathrm{d}_A g = \mathrm{d}g + [A, g]$.

So g is covariantly constant. A covariantly constant section is determined by its value at one point: along any path from x_0 it is parallel transported, $g(x) = \mathrm{hol}_\gamma\, g(x_0)\, \mathrm{hol}_\gamma^{-1}$. If g is based, $g(x_0) = \mathbb{1}$, then $g(x) = \mathbb{1}$ everywhere. So $\mathcal{G}_0$ acts freely. The *unbased* group always has a non-trivial stabiliser, at least the centre (for $U(n)$, the constant scalars $e^{i\theta}\mathbb{1}$), and more at reducible connections (for instance a direct sum of two line bundles is fixed by $\mathrm{U}(1) \times \mathrm{U}(1)$). Stabilisers are why the equivariant route is often more natural.

Watch out: $\mathcal{A}/\mathcal{G}$ versus $B\mathcal{G}$

The space $\mathcal{A}$ of all connections on a bundle is *affine*: given A_0, every other connection is $A_0 + a$ with a an $\mathrm{ad}\,P$-valued 1-form. So $\mathcal{A}$ is contractible (shrink a to zero). If a group acts freely on a contractible space, the quotient is a model of its classifying space. Thus $\mathcal{A}/\mathcal{G}_0 \simeq B\mathcal{G}_0$ exactly. For the full gauge group, which does not act freely, $\mathcal{A}/\mathcal{G}$ is *not* $B\mathcal{G}$; but its equivariant cohomology is, $H^*_{\mathcal{G}}(\mathcal{A}) = H^*(\mathcal{A} \times_{\mathcal{G}} E\mathcal{G}) = H^*(B\mathcal{G})$, because $\mathcal{A}$ is contractible. When Nash writes "$\mathcal{A}/\mathcal{G} = B\mathcal{G}$" he means one of these two statements.

8.18.4 The Yang–Mills–Higgs energy as a Morse function

We return to SU(2) Yang–Mills–Higgs theory on $\mathbb{R}^3$ in the Prasad–Sommerfield limit. The configuration space $\mathcal{A}_{YMH}$ is the space of pairs (A, ϕ) of an SU(2) connection and an adjoint Higgs field with finite energy and the boundary condition $|\phi| \to C$ at infinity (Section 8.13.5). On it lives the energy

$$E(A,\phi) = \frac{1}{2}\Big(\|F(A)\|^2 + \big\|\mathrm{d}_A\phi\big\|^2 \Big), \tag{8.350}$$

Nash's (8.197), with the norms (8.266). We have already found its minima in each sector. Now we treat E as a function on an infinite-dimensional space and ask for all its critical points.

Worked example 8.170: The critical points of E are the Yang–Mills–Higgs equations

Vary $A \to A + ta$ and $\phi \to \phi + t\psi$, with a an $\mathfrak{su}(2)$-valued 1-form and ψ an $\mathfrak{su}(2)$-valued function, both decaying at infinity. To first order in t,

$$F(A+ta) \overset{(1)}{=} F + t\,\mathrm{d}_A a + O(t^2), \qquad \mathrm{d}_{A+ta}(\phi + t\psi) \overset{(2)}{=} \mathrm{d}_A\phi + t\big(\mathrm{d}_A\psi + [a,\phi]\big) + O(t^2).$$

(1) $F(A+ta) = \mathrm{d}(A+ta) + (A+ta)\wedge(A+ta) = F + t(\mathrm{d}a + A\wedge a + a\wedge A) + t^2 a\wedge a$, and $\mathrm{d}a + A\wedge a + a\wedge A = \mathrm{d}_A a$ for an adjoint-valued 1-form.

(2) $\mathrm{d}_{A+ta}\chi = \mathrm{d}\chi + [A+ta,\chi]$; expand with $\chi = \phi + t\psi$.

Hence

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} E &\overset{(1)}{=} \langle F, \mathrm{d}_A a\rangle + \langle \mathrm{d}_A\phi, \mathrm{d}_A\psi + [a,\phi]\rangle \\ &\overset{(2)}{=} \langle \mathrm{d}_A^\star F, a\rangle + \langle \mathrm{d}_A^\star \mathrm{d}_A\phi, \psi\rangle + \langle \mathrm{d}_A\phi, [a,\phi]\rangle \\ &\overset{(3)}{=} \big\langle \mathrm{d}_A^\star F + [\phi, \mathrm{d}_A\phi],\, a\big\rangle + \langle \mathrm{d}_A^\star \mathrm{d}_A\phi, \psi\rangle. \end{aligned}$$

(1) $\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}\|X+tY\|^2 = \langle X, Y\rangle$ at $t=0$.

(2) $\mathrm{d}_A^\star$ is the formal adjoint of d_A; the boundary terms vanish because a, ψ decay.

(3) Pointwise, with $X = D_i\phi$: $-2\,\mathrm{tr}(X[a_i,\phi]) = -2\,\mathrm{tr}(a_i[\phi,X])$ by cyclicity of the trace, so $\langle \mathrm{d}_A\phi, [a,\phi]\rangle = \langle [\phi, \mathrm{d}_A\phi], a\rangle$.

This vanishes for all (a,ψ) if and only if

$$\mathrm{d}_A^\star F = [\mathrm{d}_A\phi, \phi], \qquad \mathrm{d}_A^\star \mathrm{d}_A\phi = 0, \tag{8.351}$$

which is Nash's (8.198), the $\lambda = 0$ case of (8.282).

The quotient and its components

E is gauge invariant, so it descends to the quotient $\mathcal{A}_{YMH}/\mathcal{G}$, where Nash takes $\mathcal{G}$ to be the *base point preserving* gauge transformations, $g \to \mathbb{1}$ at infinity. By Example 8.169 these act freely, so the quotient has no singular points. (They also leave the asymptotic Higgs field $\hat{\phi}_\infty : S^2_\infty \to S^2$ untouched, which matters for the topology below.)

The boundary condition splits the configuration space into pieces. Every finite-energy configuration has an integer magnetic charge $k = \deg \hat{\phi}_\infty$ (Section 8.13.5, (8.295)). So

$$\mathcal{A}_{YMH} = \bigcup_{k\in\mathbb{Z}} \mathcal{A}^k_{YMH}, \qquad \mathcal{A}^k_{YMH} = \left\{(A,\phi) : \; -\frac{1}{4\pi C}\langle F, \star \mathrm{d}_A\phi\rangle = k\right\}, \tag{8.352}$$

Nash's (8.199)–(8.200). Correspondingly the orbit space splits into $\mathcal{C}_k = \mathcal{A}^k_{YMH}/\mathcal{G}$, one for each charge.

Worked example 8.171: Why the pieces are disconnected

Suppose (A_s, ϕ_s), $0 \le s \le 1$, is a continuous path of finite-energy configurations. Then

$$k(s) \overset{(1)}{=} -\frac{1}{4\pi C}\langle F_s, \star \mathrm{d}_{A_s}\phi_s\rangle \overset{(2)}{\text{ is continuous in } s}, \qquad k(s) \overset{(3)}{\in} \mathbb{Z}.$$

(1) Definition (8.291).
(2) The pairing is continuous in the fields (in the finite-energy norm).
(3) k is the degree of $\hat{\phi}_\infty$, an integer (Section 8.13.5).

A continuous integer-valued function on $[0,1]$ is constant (it cannot jump without taking non-integer values in between). So $k(0) = k(1)$: no path joins $\mathcal{A}^k_{YMH}$ to $\mathcal{A}^{k'}_{YMH}$ with $k \neq k'$. Nash's normalisation "$\frac{1}{4\pi}\langle F, \star \mathrm{d}_A\phi\rangle = k$" is ours with $C = 1$ and the opposite sign convention for k (see Section 8.13.5).

The topology of the pieces

Why are the $\mathcal{C}_k$ topologically non-trivial, when each $\mathcal{A}^k_{YMH}$ is built out of a vector space of connections and Higgs fields? Because of the quotient. The based gauge group on $\mathbb{R}^3$ is the group of maps $g : \mathbb{R}^3 \cup \{\infty\} = S^3 \to \mathrm{SU}(2)$ with $g(\infty) = \mathbb{1}$. Such a map has a degree $\deg g \in \pi_3(S^3) = \mathbb{Z}$, and gauge transformations of different degrees cannot be deformed into each other. So $\mathcal{G}$ has infinitely many components, $\pi_0(\mathcal{G}) = \mathbb{Z}$.

Key idea: A non-contractible loop from a large gauge transformation

Let g_1 be a based gauge transformation of degree one, and $x \in \mathcal{A}^k_{YMH}$ any configuration. Join x to $g_1 \cdot x$ by any path γ in $\mathcal{A}^k_{YMH}$ (for example the straight line: the space of fields with given boundary behaviour is affine). In the quotient $\mathcal{C}_k$ the endpoints are the same point, so γ becomes a closed loop $[\gamma]$. *This loop cannot be contracted.* If it could, lifting the contraction to $\mathcal{A}^k_{YMH}$ would deform the endpoint $g_1 \cdot x$ back to x through points of the orbit $\mathcal{G} \cdot x$, i.e. it would give a continuous path in $\mathcal{G}$ from g_1 to $\mathbb{1}$; impossible, since their degrees differ.

In homotopy language: the free action gives a fibration $\mathcal{G} \to \mathcal{A}^k_{YMH} \to \mathcal{C}_k$, and its exact sequence contains $\pi_1(\mathcal{C}_k) \to \pi_0(\mathcal{G}) = \mathbb{Z} \to \pi_0(\mathcal{A}^k_{YMH})$. The loop above maps to the generator $1 \in \mathbb{Z}$. Continuing the exact sequence, the higher homotopy groups $\pi_n(\mathcal{G}) = \pi_{n+3}(S^3)$ (which

are non-zero for infinitely many n: $\mathbb{Z}_2, \mathbb{Z}_2, \mathbb{Z}_{12}, \mathbb{Z}_2, \mathbb{Z}_2, \dots$ for $n = 1, 2, 3, 4, 5$) feed into $\pi_{n+1}(\mathcal{C}_k)$. Together with the topology of the space of asymptotic Higgs fields (maps $S^2 \to S^2$ of degree k), this makes every $\mathcal{C}_k$ topologically rich. Taubes computed the homotopy type precisely; we only need that it is far from contractible.

Physical picture: Tunnelling between vacua

In the $k = 0$ sector take x to be the vacuum, $A = 0$, $\phi = CT_3$. Then $g_1 \cdot x$ is again a vacuum, of the same (zero) energy, but it is a *different point* of $\mathcal{A}^0_{YMH}$, and the only way between them in field space goes over a hill. In the quotient the two vacua are identified and the path becomes a non-contractible loop. The highest point of the best such loop is a saddle point of the energy: a static, unstable, non-minimal solution. This is precisely the logic of the electroweak sphaleron (Manton; Klinkhamer and Manton), whose energy, about 9 TeV, is the height of the barrier between topologically distinct vacua of the Standard Model; crossing it changes baryon number.

The minima as a Morse–Bott critical manifold

What does the Hessian of E look like at a BPS monopole? Use the completed square (8.292): on $\mathcal{A}^k_{YMH}$, with $k > 0$,

$$E(A, \phi) = \frac{1}{2} \left\| F + \star \mathrm{d}_A \phi \right\|^2 + 4\pi C k.$$

Expand around a solution (A, ϕ) of $F = - \star \mathrm{d}_A \phi$ in a direction (a, η):

$$E(A + ta, \phi + t\eta) \overset{(1)}{=} \frac{1}{2} \left\| F + \star \mathrm{d}_A \phi + t\, D_1(a, \eta) + O(t^2) \right\|^2 + 4\pi C k$$
$$\overset{(2)}{=} 4\pi C k + \frac{t^2}{2} \left\| D_1(a, \eta) \right\|^2 + O(t^3).$$

(1) The first-order change of $F + \star \mathrm{d}_A \phi$ is $\mathrm{d}_A a + \star(\mathrm{d}_A \eta + [a, \phi]) = D_1(a, \eta)$, the linearised Bogomolny operator (8.322); k does not change.

(2) At a solution $F + \star \mathrm{d}_A \phi = 0$, so only the square of the first-order term survives at order t^2.

So the Hessian at a monopole is the non-negative operator $D_1^\star D_1$. Its null space is $\ker D_1$; modulo gauge directions (the image of D_0) this is the tangent space to the moduli space, of dimension $4k$ (Section 8.15). In the language of Definition 8.47: the absolute minima of E on $\mathcal{C}_k$ form a critical manifold $\mathcal{M}'_k$ of index 0 and dimension $4k$. The Hessian has no negative directions, and its only normalisable zero modes (modulo gauge) are the $4k$ tangent directions; the Weitzenböck formula $DD^\dagger = -D^2 - \mathrm{ad}^2_\phi$ of Example 8.139 is what rules out any others. One technical caveat, typical of $\mathbb{R}^3$: the unbroken U(1) fields (photon and the massless Higgs mode) are massless, so the spectrum of $D_1^\star D_1$ is continuous down to 0 and there is no spectral gap above the zero modes. This is one of the "technical difficulties" of non-compact M that Nash mentions and that Taubes' weighted function spaces handle.

Taubes' theorems

Nash quotes two results of Taubes on this Morse theory.

Key idea: Taubes: E is a good Morse function

1. In each sector $\mathcal{C}_k$ there are critical points of E that are *not* minima: static solutions of (8.351) that do not satisfy the Bogomolny equation. (The first case Taubes settled was

$k = 0$, below.)

2. There are infinitely many of them, and their energies are unbounded: for every E_0 there is a non-minimal critical point with energy above E_0.

The mechanism is Lusternik–Schnirelmann min-max over the non-trivial topology of C_k. Each homotopy class of non-contractible spheres in C_k gives a min-max value; the rich homotopy described above gives infinitely many classes, and the min-max values of more and more complicated classes grow without bound. The hard analysis is to show that each min-max value is really attained, even though the Palais–Smale condition fails on $\mathbb{R}^3$. We now see concretely how it fails, and why it can be controlled.

8.18.5 Taubes' loop and the monopole–antimonopole saddle

Work in the $k = 0$ sector. Its minimum is the vacuum ($E = 0$). By the key idea above, the loop class $[\gamma]$ built from a degree-one gauge transformation is not contractible. Taubes drew an explicit, physical representative of it:

1. start at the vacuum;
2. create a monopole ($k = +1$) and an antimonopole ($k = -1$) at the origin and pull them apart to a large separation d along an axis;
3. keeping them apart, rotate the internal orientation of the monopole by 2π about that axis, relative to the antimonopole (a relative U(1) phase that winds once, the same phase that becomes the dyon angle in Section 8.16.4);
4. bring them together and annihilate them.

Taubes checks that the fields at the end of this process are not the original vacuum but $g_1 \cdot$ (vacuum) with $\deg g_1 = 1$. Heuristically, the 2π relative rotation leaves behind a gauge transformation that turns once around the axis joining the pair, and the creation and annihilation stages close this up into a map of S^3 onto SU(2) covering it once. So the loop is the non-contractible one of the key idea.

Now apply min-max. Let

$$c = \inf_{\gamma \in [\gamma]} \max_s E(\gamma(s)).$$

Two inequalities make this useful.

Lower bound, $c > 0$. The zero-energy configurations in C_0 are the vacua; up to based gauge transformations they are labelled by the constant asymptotic direction $\hat{\phi}_\infty \in S^2$, so they form a copy of S^2. A loop of very small energy stays close to this vacuum set, and loops near S^2 can be shrunk because $\pi_1(S^2) = 0$. A non-contractible loop must therefore climb to some definite height.

Upper bound, $c < 8\pi C$. Along Taubes' representative the energy is highest while the pair is separated. There the energy is approximately two monopole masses plus the (attractive) interaction energy.

Worked example 8.172: The energy of a separated monopole–antimonopole pair

For a well-separated pair at distance d (in units where the core size $1/C$ is small compared with d), treat each object as a point particle of mass $4\pi C$ carrying magnetic charge $\pm g$ with

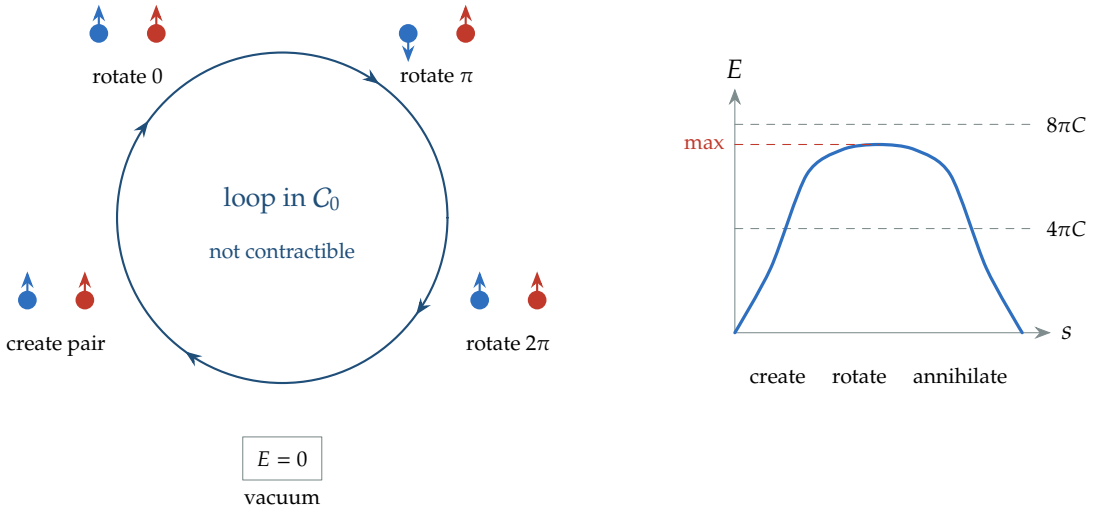


Figure 8.87: Taubes' non-contractible loop in the $k = 0$ configuration space. Left: a monopole (blue) and antimonopole (red) are created from the vacuum, the monopole's internal orientation (arrow) is turned by 2π relative to the antimonopole, and the pair annihilates. Right: the energy along such a loop stays below $8\pi C$, twice the monopole mass, because the pair attracts. Minimising the maximum over all loops in the class gives the monopole–antimonopole saddle point.

$g = 4\pi$ and a scalar charge of the same size (Examples 8.146 and 8.147). Then

$$E(d) \overset{(1)}{=} 4\pi C + 4\pi C + V_{\text{mag}}(d) + V_{\text{Higgs}}(d)$$
$$\overset{(2)}{=} 8\pi C - \frac{g^2}{4\pi d} - \frac{g^2}{4\pi d} \overset{(3)}{=} 8\pi C - \frac{8\pi}{d} + O(e^{-Cd}).$$

(1) Each isolated object has the BPS mass $4\pi C|k|$, $|k| = 1$; add the interaction energies.
(2) Opposite magnetic charges attract with the Coulomb energy $-g^2/(4\pi d)$; the massless Higgs exchange between a monopole and an antimonopole is also attractive with the same strength (Example 8.147).
(3) $g^2/(4\pi) = 4\pi$; the corrections from the massive fields fall off like e^{-Cd}.

The force is $-\mathrm{d}E/\mathrm{d}d = -8\pi/d^2$ (attractive), matching Example 8.147. So $E(d) < 8\pi C$ for every large d, and a rotation of the relative phase at fixed large separation costs only exponentially small energy. Hence loops in the class exist with maximum energy below $8\pi C$, and $c < 8\pi C$.

How the Palais–Smale condition fails, and why it does not matter

Example 8.172 also shows exactly how condition (C) fails for E on $\mathbb{R}^3$. Take the pair at separations $d_n \to \infty$. Then $E \to 8\pi C$ stays bounded, and the configuration becomes ever more nearly a solution: each constituent solves the field equations exactly except for the influence of the other, whose fields at distance d_n are of order $1/d_n$ with gradients of order $1/d_n^2$. So $|\nabla E| \to 0$. But the sequence has no limit: the monopole and antimonopole run off to infinity in opposite directions. It is Example 8.167(a) in field theory: a "critical point at infinity", namely an infinitely separated pair, at level $8\pi C$.

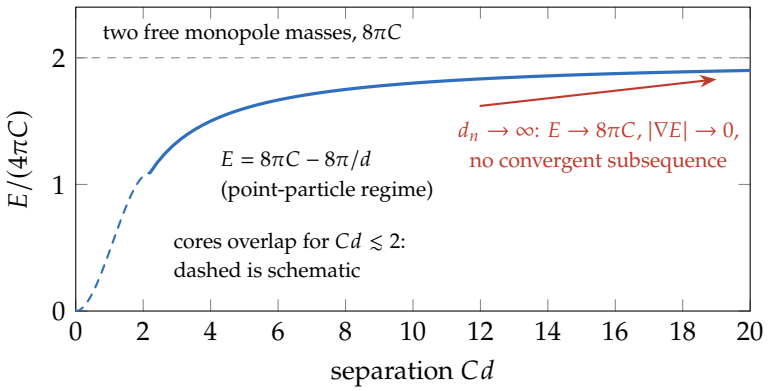


Figure 8.88: Energy of a monopole–antimonopole pair against separation (solid: the point-particle formula of Example 8.172; dashed: schematic behaviour when the cores overlap and the pair can annihilate). Sending the pair apart gives a Palais–Smale sequence at level $8\pi C$ with no limit: condition (C) fails on $\mathbb{R}^3$. Below the threshold $8\pi C$ this cannot happen.

Physical picture: Energy bookkeeping restores compactness

On $\mathbb{R}^3$ the only way a Palais–Smale sequence can fail to converge (up to gauge and translation) is by splitting into pieces that separate to infinity; this is the content of Taubes' analysis. Each piece that carries non-zero magnetic charge costs at least $4\pi C$ (the Bogomolny bound), and in the $k = 0$ sector the charges must add up to zero, so a splitting costs at least $4\pi C + 4\pi C = 8\pi C$ in the limit. The min-max level satisfies $0 < c < 8\pi C$. So a min-max sequence at level c has not enough energy to split, it converges, and its limit is a genuine critical point in C_0 with energy c: a non-minimal static solution of (8.351). Since it comes from a one-parameter family, it is expected to have index one, like the mountain pass.

The saddle point found by Taubes is the **monopole–antimonopole pair**. It has been computed numerically (Rüber; Kleihaus and Kunz): an axially symmetric static solution whose Higgs field has two zeros on the symmetry axis, a monopole and an antimonopole held apart by the balance between their attraction and the twisting of the gauge field between them, with an energy a little below $8\pi C$.

Dictionary: finite-dimensional Morse theory ↔ Yang–Mills–Higgs on $\mathbb{R}^3$

manifold M	orbit space $C_k = \mathcal{A}^k_{YMH}/\mathcal{G}$
function f	energy E of (8.350)
connected components	charge sectors $k \in \mathbb{Z}$
critical points	static solutions of (8.351)
minimum set (Morse–Bott)	BPS moduli space $\mathcal{M}'_k$, dim $= 4k$
index-one saddle	monopole–antimonopole pair (sphaleron-like)
non-contractible loop	path from vacuum to a degree-one gauge transform of it
failure of condition (C)	monopoles escaping to infinity
energy threshold	Bogomolny bound $4\pi C$ per constituent

Reading Nash (pp. 256–257, (8.197)–(8.200))

Nash recalls from Chapter 7 that non-degenerate critical points of a function are tied to the topology of its domain, notes that the gauge-theory examples need the infinite-dimensional generalisations of Palais and of Taubes, and warns that pure Yang–Mills in four dimensions behaves differently from Yang–Mills–Higgs, that two dimensions is special, and that the gauge symmetry must be factored out. For SU(2) Yang–Mills–Higgs on $\mathbb{R}^3$ he writes the energy (8.197), its critical points (8.198), the passage to the orbit space by base point preserving gauge transformations, the decomposition (8.199) by magnetic charge (8.200), and quotes Taubes: each C_k is topologically non-trivial and carries infinitely many non-minimal critical points of unbounded energy, so E has "the properties of a good Morse function". His (8.200), $\frac{1}{4\pi}\langle F, \star \mathrm{d}_A\phi\rangle = k$, is our (8.352) with $C = 1$, his $-\operatorname{tr}$ normalisation and the opposite sign convention for k. We added the finite-dimensional recap, the Palais–Smale condition and its failure (in calculus and for monopoles), the mountain pass, the free action of based gauge transformations, the first-variation calculation, the Morse–Bott structure of the minima, the non-contractible loop and Taubes' monopole–antimonopole saddle.

8.18.6 Yang–Mills theory on a Riemann surface

Nash's second example is pure Yang–Mills theory in *two* dimensions. It is the one gauge theory whose Morse theory is completely understood, and it is "striking" because everything can be computed. The setting: Σ is a closed oriented surface of genus g with a Riemannian metric of total area $\mathcal{A}_\Sigma$, and $P \to \Sigma$ is a principal $\mathrm{U}(n)$ bundle, equivalently a rank-n complex vector bundle E with a Hermitian metric. Its topological type is fixed by one integer, the degree $d = c_1(E) = \frac{i}{2\pi}\int_\Sigma \operatorname{tr} F$ (Chapter T2). The connections are $\mathfrak{u}(n)$-valued 1-forms, i.e. anti-Hermitian $n \times n$ matrices of 1-forms.

Conventions for this subsection. For $\mathrm{U}(n)$ we follow Nash's (8.201) literally and use $-\operatorname{tr}$ without the factor 2 of (8.266):

$$\langle \alpha, \beta\rangle = -\int_\Sigma \operatorname{tr}(\alpha \wedge \star\beta), \qquad S(A) = \|F\|^2 = -\int_\Sigma \operatorname{tr}(F \wedge \star F). \tag{8.353}$$

This is positive. For an anti-Hermitian matrix X ($X^\dagger = -X$),

$$-\operatorname{tr}(X^2) \overset{(1)}{=} \operatorname{tr}(X^\dagger X) \overset{(2)}{=} \sum_{i,j} |X_{ij}|^2 \;\geq\; 0.$$

(1) $X = -X^\dagger$.
(2) $(X^\dagger X)_{jj} = \sum_i \overline{X_{ij}} X_{ij}$.

The Hodge star on a surface, in a local oriented orthonormal coframe e^1, e^2 with area form $\mathrm{vol} = e^1 \wedge e^2$, is

$$\star 1 = \mathrm{vol}, \quad \star \mathrm{vol} = 1, \quad \star e^1 = e^2, \quad \star e^2 = -e^1, \qquad \star\star = -1 \text{ on 1-forms.}$$

The key point is the first two: *the dual of a 2-form is a function.* So $\star F$ is an $\mathrm{ad}\,P$-valued function, which we call $\Phi = \star F$; then $F = \Phi\,\mathrm{vol}$.

Worked example 8.173: The Yang–Mills equations in two dimensions

The critical points of S satisfy $\mathrm{d}_A^\star F = 0$ (the same first-variation calculation as in Example 8.170, without ϕ). In two dimensions the adjoint of d_A on p-forms is $\mathrm{d}_A^\star = (-1)^{n(p+1)+1} \star \mathrm{d}_A \star$ with $n = 2$, i.e. $\mathrm{d}_A^\star = - \star \mathrm{d}_A \star$ for every p. Hence

$$\mathrm{d}_A^\star F = 0 \overset{(1)}{\iff} \star\, \mathrm{d}_A(\star F) = 0 \overset{(2)}{\iff} \mathrm{d}_A(\star F) = 0 \overset{(3)}{\iff} \mathrm{d}\Phi + [A, \Phi] = 0.$$

(1) $\mathrm{d}_A^\star = - \star \mathrm{d}_A \star$; drop the sign.
(2) $\star$ is invertible on 1-forms ($\star\star = -1$).
(3) $\Phi = \star F$ is an adjoint-valued function, and d_A acts on it by commutator.

This is Nash's (8.202). The Yang–Mills equation says simply that the "field strength function" Φ is **covariantly constant**. (There is no Bianchi identity to add: $\mathrm{d}_A F$ is a 3-form, and there are none on a surface.)

Covariant constancy is a strong condition. It makes the eigenvalues of Φ constant, and it splits the bundle.

Worked example 8.174: Covariantly constant Φ: constant eigenvalues and a splitting

Eigenvalues. For each $m \geq 1$,

$$\mathrm{d}\ \mathrm{tr}(\Phi^m) \overset{(1)}{=} \mathrm{tr}\big(\mathrm{d}_A(\Phi^m)\big) \overset{(2)}{=} m\ \mathrm{tr}\big(\Phi^{m-1}\mathrm{d}_A\Phi\big) \overset{(3)}{=} 0.$$

(1) $\mathrm{d}_A(\Phi^m) = \mathrm{d}(\Phi^m) + [A, \Phi^m]$, and the trace of a commutator is zero.
(2) Leibniz rule for d_A, and cyclicity of the trace to collect the m terms.
(3) $\mathrm{d}_A\Phi = 0$.

So all the $\mathrm{tr}\,\Phi^m$ are constant on Σ, hence the characteristic polynomial of Φ is constant, hence its eigenvalues are constant. Since Φ is anti-Hermitian they are imaginary; write them $-i\lambda_1, \dots, -i\lambda_r$ with multiplicities $n_1, \dots, n_r$.
Splitting. Let E_j be the $-i\lambda_j$ eigenbundle. If s is a section of E_j, then

$$\Phi\,(\mathrm{d}_A s) \overset{(1)}{=} \mathrm{d}_A(\Phi s) - (\mathrm{d}_A\Phi)s \overset{(2)}{=} -i\lambda_j\, \mathrm{d}_A s,$$

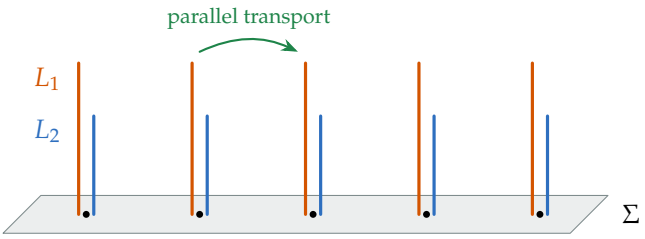
(1) Leibniz: $\mathrm{d}_A(\Phi s) = (\mathrm{d}_A\Phi)s + \Phi\, \mathrm{d}_A s$.
(2) $\Phi s = -i\lambda_j s$ with λ_j constant, and $\mathrm{d}_A\Phi = 0$.

so $\mathrm{d}_A s$ is again E_j-valued. The connection preserves each eigenbundle: $E = E_1 \oplus \cdots \oplus E_r$ and $A = A_1 \oplus \cdots \oplus A_r$ is *reducible*. On each summand the curvature is a constant multiple of the identity, $F_j = -i\lambda_j\, \mathbb{1}_{n_j}$ vol: a **projectively flat**, or constant central curvature, connection.
The constants. Integrate the first Chern class of E_j, of degree d_j and rank n_j:

$$d_j \overset{(1)}{=} \frac{i}{2\pi}\int_\Sigma \mathrm{tr}\, F_j \overset{(2)}{=} \frac{i}{2\pi}(-i\lambda_j)\, n_j\, \mathcal{A}_\Sigma \overset{(3)}{\implies} \lambda_j = \frac{2\pi d_j}{n_j \mathcal{A}_\Sigma}.$$

(1) $c_1 = \frac{i}{2\pi}\,\mathrm{tr}\, F$ (Chapter T2).
(2) $\mathrm{tr}\, F_j = -i\lambda_j n_j$ vol and $\int \mathrm{vol} = \mathcal{A}_\Sigma$.
(3) Solve.

So the eigenvalue λ_j is $2\pi/\mathcal{A}_\Sigma$ times the **slope** $\mu_j = d_j/n_j$ of the summand.



$$E = L_1 \oplus L_2, \quad \star F = \begin{pmatrix} -i\lambda_1 & 0 \\ 0 & -i\lambda_2 \end{pmatrix} \text{constant}, \quad \lambda_j = \frac{2\pi d_j}{\mathcal{A}_\Sigma}$$

Figure 8.89: A Yang–Mills connection on a rank-2 bundle over a surface. The covariantly constant field-strength function $\star F$ has constant eigenvalues, and its eigenlines L_1 (orange) and L_2 (blue) are carried into themselves by parallel transport: the bundle splits, and on each line the curvature is constant. When $\lambda_1 = \lambda_2$ there is no splitting and the connection is projectively flat: a minimum.

Key idea: The critical points of two-dimensional Yang–Mills

A connection on a $\mathrm{U}(n)$ bundle over Σ is Yang–Mills if and only if it is a direct sum of connections with constant central curvature, $E = \bigoplus_j E_j$, $\star F|_{E_j} = -\frac{2\pi i}{\mathcal{A}_\Sigma}\frac{d_j}{n_j}\mathbb{1}$. Its action is

$$S = -\int_\Sigma \operatorname{tr}\Phi^2 \operatorname{vol} = \sum_j n_j \lambda_j^2 \mathcal{A}_\Sigma = \frac{4\pi^2}{\mathcal{A}_\Sigma}\sum_j \frac{d_j^2}{n_j}. \tag{8.354}$$

The critical values are an explicit discrete set, labelled by the ways of writing (n, d) as a sum of (n_j, d_j).

Worked example 8.175: $\mathrm{U}(1)$ on the sphere is the Dirac monopole

For $n = 1$ there is only one summand, so every Yang–Mills connection has constant curvature $F = -i\frac{2\pi d}{\mathcal{A}_\Sigma}\mathrm{vol}$. On the unit sphere ($\mathcal{A}_\Sigma = 4\pi$, $\mathrm{vol} = \sin\theta\,\mathrm{d}\theta\wedge\mathrm{d}\varphi$) this is

$$F = -\frac{id}{2}\sin\theta\,\mathrm{d}\theta\wedge\mathrm{d}\varphi,$$

a uniform magnetic field through the sphere with total flux $\int iF = 2\pi d$: exactly the Dirac monopole of charge d of Section 8.12.3, restricted to a sphere around it. Its action is

$$S \overset{(1)}{=} -\int_{S^2}\left(-\frac{id}{2}\right)^2 \sin\theta\,\mathrm{d}\theta\,\mathrm{d}\varphi \overset{(2)}{=} \frac{d^2}{4}\cdot 4\pi = \pi d^2,$$

(1) $\Phi = \star F = -id/2$ is a constant function; $S = -\int \Phi^2 \mathrm{vol}$.

(2) $(-i)^2 = -1$ and $\int \mathrm{vol} = 4\pi$.

in agreement with (8.354): $4\pi^2 d^2/(4\pi) = \pi d^2$. In the $\mathrm{U}(1)$ theory there is exactly one critical value in each degree, and (as we shall see from the Hessian) it is the minimum.

Worked example 8.176: The minimum in each sector: completing the square once more

For any connection on a bundle of rank n and degree d, split $\Phi = \Phi_0 + \frac{1}{n}(\operatorname{tr}\Phi)\mathbb{1}$ into traceless

and central parts. Then

$$S \overset{(1)}{=} -\int \operatorname{tr}\Phi_0^2\,\mathrm{vol} - \frac{1}{n}\int (\operatorname{tr}\Phi)^2\mathrm{vol} \overset{(2)}{\geq} \frac{1}{n}\int |\operatorname{tr}\Phi|^2\mathrm{vol}$$
$$\overset{(3)}{\geq} \frac{1}{n\mathcal{A}_\Sigma}\left|\int \operatorname{tr}\Phi\,\mathrm{vol}\right|^2 \overset{(4)}{=} \frac{4\pi^2 d^2}{n\mathcal{A}_\Sigma}.$$

(1) Expand $\operatorname{tr}\Phi^2$; the cross term is $\frac{2}{n}\operatorname{tr}\Phi\operatorname{tr}\Phi_0 = 0$.
(2) $-\operatorname{tr}\Phi_0^2 \geq 0$ (anti-Hermitian), and $\operatorname{tr}\Phi$ is imaginary so $-(\operatorname{tr}\Phi)^2 = |\operatorname{tr}\Phi|^2$.
(3) Cauchy–Schwarz: $(\int |f|)^2 \leq \mathcal{A}_\Sigma \int |f|^2$.
(4) $\int \operatorname{tr}\Phi\,\mathrm{vol} = \int \operatorname{tr} F = -2\pi i d$.

Equality holds iff $\Phi_0 = 0$ and $\operatorname{tr}\Phi$ is constant: the projectively flat connections, with $r = 1$ summand in the key idea. This is the absolute minimum. Every other Yang–Mills connection (a genuine splitting with unequal slopes) has larger action; for example on a rank-2, degree-0 bundle the splitting into degrees $(m, -m)$ has $S = 8\pi^2 m^2/\mathcal{A}_\Sigma$, compared with the minimum 0.

Holomorphic methods

Nash says the equations "translate into a holomorphic problem". The translation is this. A Riemann surface is a complex curve, and on a complex vector bundle over a complex curve a unitary connection defines a **holomorphic structure**: split $\mathrm{d}_A = \partial_A + \bar\partial_A$ into its $(1,0)$ and $(0,1)$ parts (in a local coordinate z, $\bar\partial_A = \mathrm{d}\bar z\,(\partial_{\bar z} + A_{\bar z})$), and declare a section holomorphic when $\bar\partial_A s = 0$. In higher dimensions this requires $F^{0,2} = 0$; on a curve there are no $(0,2)$-forms, so *every* connection gives a holomorphic bundle. Gauge-equivalent connections give isomorphic holomorphic bundles, and conversely two connections giving isomorphic holomorphic bundles differ by a *complex* gauge transformation. Three classic facts then describe the Yang–Mills critical points in holomorphic terms:

- *Grothendieck.* On $\mathbb{CP}^1$ every holomorphic bundle is a sum of line bundles, $E \cong \mathcal{O}(a_1) \oplus \cdots \oplus \mathcal{O}(a_n)$, with $\mathcal{O}(a)$ the line bundle of degree a (so $\mathcal{O}(-1)$ is the tautological bundle of Chapter 5).

- *Narasimhan–Seshadri* (in the form Donaldson proved it). The minima, the projectively flat connections, correspond to the **polystable** holomorphic bundles, direct sums of stable bundles of the same slope. Here E is **stable** (**semistable**) if every holomorphic subbundle E' has slope $\deg E'/\operatorname{rank} E' < \deg E/\operatorname{rank} E$ ($\leq$). The semistable bundles form the open stratum of connections that flow down to the minimum.

- *Harder–Narasimhan.* Every bundle has a canonical filtration whose quotients are semistable with decreasing slopes; the Yang–Mills critical point that the downward flow from it reaches is (the polystable version of) the direct sum of those quotients. Its slopes (d_j/n_j) label a **stratum** of the space of connections.

So the critical set of S is a union of pieces, each the set of Yang–Mills connections of a given "type" $\mu = (d_1/n_1, \dots)$, and the space $\mathcal{A}$ of connections is stratified accordingly. This is the finite-dimensional-looking structure that makes a Morse theory possible.

8.18.7 The orbit space in two dimensions

Nash: "$S(A)$ is defined on the contractible space $\mathcal{A}$ but it descends to an orbit space whose topology is far from trivial." By the pitfall on $\mathcal{A}/\mathcal{G}$ and $B\mathcal{G}$, the topology of the orbit space is the topology of a classifying space. For the sphere it can be written down.

Worked example 8.177: Genus zero: $\mathcal{A}/\mathcal{G}_0 = \Omega\,\mathrm{SU}(n)$

Take $\Sigma = S^2$ and the based gauge group $\mathcal{G}_0$ (gauge transformations equal to $\mathbb{1}$ at a point). As a space it is the group of based maps $S^2 \to \mathrm{U}(n)$, the double loop space $\Omega^2\mathrm{U}(n)$. Then

$$\mathcal{A}/\mathcal{G}_0 \overset{(1)}{\simeq} B\mathcal{G}_0 \overset{(2)}{\simeq} B\big(\Omega(\Omega\mathrm{U}(n))\big) \overset{(3)}{\simeq} \Omega_0\mathrm{U}(n) \overset{(4)}{\simeq} \Omega\,\mathrm{SU}(n).$$

(1) $\mathcal{G}_0$ acts freely on the contractible space $\mathcal{A}$ (Example 8.169 and the pitfall after it).
(2) A based map of S^2 = (square with boundary collapsed) is a loop of loops.
(3) For any connected space Y, $B(\Omega Y) \simeq Y$: the path fibration $\Omega Y \to PY \to Y$ has contractible total space PY, so it is the universal ΩY-bundle. Here Y is the component $\Omega_0\mathrm{U}(n)$ of contractible loops (a loop of loops based at the constant loop stays in that component).
(4) $\mathrm{U}(n) = \mathrm{U}(1) \times \mathrm{SU}(n)$ as spaces, and the loops in U(1) that are contractible form a contractible space.

This is Nash's "$\mathcal{A}/\mathcal{G} = B\mathcal{G} = \Omega G$". For U(2) it is $\Omega\mathrm{SU}(2) = \Omega S^3$, the space of loops on the three-sphere, whose Morse theory we did in Chapter 7: the energy of paths has critical points (geodesics) of index $0, 2, 4, 6, \ldots$, one of each, and is perfect, so

$$P_t(\Omega S^3) = 1 + t^2 + t^4 + \cdots = \frac{1}{1-t^2}.$$

In general $P_t(\Omega\mathrm{SU}(n)) = \prod_{j=1}^{n-1}(1-t^{2j})^{-1}$ (Bott). A contractible space divided by a gauge group has become a space with infinitely many non-zero Betti numbers: Nash's "rich topology".

For the Atiyah–Bott theorem we need the full gauge group $\mathcal{G}$, whose classifying space is (Atiyah and Bott) the space of maps $\Sigma \to B\mathrm{U}(n)$ in the right homotopy class. For genus 0 and U(2) it has

$$P_t(B\mathcal{G}) \overset{(1)}{=} P_t(B\mathrm{U}(2))\,P_t(\Omega\mathrm{SU}(2)) \overset{(2)}{=} \frac{1}{(1-t^2)(1-t^4)} \cdot \frac{1}{1-t^2} = \frac{1}{(1-t^2)^2(1-t^4)}, \tag{8.355}$$

(1) Evaluation at the base point gives a fibration $\mathcal{G}_0 \to \mathcal{G} \to \mathrm{U}(2)$, hence $B\mathcal{G}_0 \to B\mathcal{G} \to B\mathrm{U}(2)$; all the cohomology is in even degrees, so the spectral sequence collapses and the Poincaré series multiply.
(2) $H^*(B\mathrm{U}(2))$ is polynomial on c_1 (degree 2) and c_2 (degree 4); Example 8.177.

and for genus g Atiyah and Bott found

$$P_t(B\mathcal{G}) = \frac{(1+t)^{2g}(1+t^3)^{2g}}{(1-t^2)^2(1-t^4)} \quad (\mathrm{U}(2)), \qquad P_t(B\mathcal{G}) = \frac{(1+t)^{2g}}{1-t^2} \quad (\mathrm{U}(1)). \tag{8.356}$$

The odd factors come from the $2g$ independent loops on Σ: a gauge transformation can wind around each of them.

8.18.8 The Hessian of the two-dimensional action

To know the *type* of each critical point we need the Hessian. Nash expands the action along a straight line $A_t = A + ta$ in the space of connections.

Worked example 8.178: Expanding the action: Nash's (8.204)–(8.206)

The curvature along the line is exactly quadratic in t:

$$F(A+ta) = \mathrm{d}(A+ta) + (A+ta)\wedge(A+ta) = F + t\,\mathrm{d}_A a + t^2\, a\wedge a.$$

Therefore

$$\begin{aligned} S(A_t) &\overset{(1)}{=} \langle F + t\,\mathrm{d}_A a + t^2 a\wedge a,\ F + t\,\mathrm{d}_A a + t^2 a\wedge a\rangle \\ &\overset{(2)}{=} S(A) + 2t\langle F, \mathrm{d}_A a\rangle + t^2\big\{\langle \mathrm{d}_A a, \mathrm{d}_A a\rangle + \langle a\wedge a, F\rangle + \langle F, a\wedge a\rangle\big\} + O(t^3) \\ &\overset{(3)}{=} S(A) + 2t\langle \mathrm{d}_A^\star F, a\rangle + t^2\big\{\|\mathrm{d}_A a\|^2 + 2\langle F, a\wedge a\rangle\big\} + O(t^3). \end{aligned}$$

(1) $S = \langle F, F\rangle$ with the pairing (8.353).
(2) Multiply out and collect powers of t; the t^3 and t^4 terms are $2t^3\langle \mathrm{d}_A a, a\wedge a\rangle + t^4\,\|a\wedge a\|^2$.
(3) Move d_A across the pairing; the real pairing is symmetric.

At a critical point $\mathrm{d}_A^\star F = 0$, so the t term vanishes, and the t^2 coefficient defines the Hessian. To write it as $\langle a, Ha\rangle$ we express $\langle F, a\wedge a\rangle$ through a. In an orthonormal coframe, $a = a_1 e^1 + a_2 e^2$, $a\wedge a = [a_1, a_2]\,\mathrm{vol}$, $F = \Phi\,\mathrm{vol}$, and $\star a = -a_2 e^1 + a_1 e^2$. Define $K(a) = [\Phi, \star a]$, whose components are $K_1 = -[\Phi, a_2]$, $K_2 = [\Phi, a_1]$. Then

$$\begin{aligned} \langle a, K(a)\rangle &\overset{(1)}{=} -\int \mathrm{tr}\big(-a_1[\Phi, a_2] + a_2[\Phi, a_1]\big)\mathrm{vol} \\ &\overset{(2)}{=} -\int \big(\mathrm{tr}(\Phi[a_1, a_2]) + \mathrm{tr}(\Phi[a_1, a_2])\big)\mathrm{vol} \overset{(3)}{=} 2\langle F, a\wedge a\rangle. \end{aligned}$$

(1) For 1-forms, $\alpha\wedge\star\beta = (\alpha_1\beta_1 + \alpha_2\beta_2)\mathrm{vol}$ (with matrix products in order).
(2) Cyclicity: $\mathrm{tr}(a_1[\Phi, a_2]) = \mathrm{tr}(\Phi a_2 a_1) - \mathrm{tr}(\Phi a_1 a_2) = -\mathrm{tr}(\Phi[a_1, a_2])$, and similarly $\mathrm{tr}(a_2[\Phi, a_1]) = \mathrm{tr}(\Phi[a_1, a_2])$.
(3) $\langle F, a\wedge a\rangle = -\int \mathrm{tr}(\Phi[a_1, a_2])\mathrm{vol}$, since $F\wedge\star(a\wedge a) = \Phi[a_1, a_2]\mathrm{vol}$.

Hence the second-order coefficient is

$$\langle a, Ha\rangle = \langle \mathrm{d}_A a, \mathrm{d}_A a\rangle + \langle a, [\star F, \star a]\rangle, \qquad H = \mathrm{d}_A^\star \mathrm{d}_A + [\star F, \star\,\cdot\,]. \tag{8.357}$$

Nash writes the zeroth-order term as "$\star(\star F(A)\,a)$": since $\star F$ is a function, $\star[\star F, a] = [\star F, \star a]$, and the product must be read as a commutator.

Worked example 8.179: Gauge directions are zero modes

Let ξ be an infinitesimal gauge transformation (an $\mathrm{ad}\,P$-valued function). The connection moves by $a = \mathrm{d}_A\xi$. At a critical point,

$$\begin{aligned} H(\mathrm{d}_A\xi) &\overset{(1)}{=} \mathrm{d}_A^\star\big([F, \xi]\big) + [\Phi, \star\mathrm{d}_A\xi] \overset{(2)}{=} -\star\,\mathrm{d}_A[\Phi, \xi] + [\Phi, \star\mathrm{d}_A\xi] \\ &\overset{(3)}{=} -\star\,\big([\mathrm{d}_A\Phi, \xi] + [\Phi, \mathrm{d}_A\xi]\big) + [\Phi, \star\mathrm{d}_A\xi] \overset{(4)}{=} 0. \end{aligned}$$

(1) $\mathrm{d}_A\mathrm{d}_A\xi = [F, \xi]$ (curvature is the square of the covariant derivative).
(2) $\mathrm{d}_A^\star = -\star\,\mathrm{d}_A\star$ and $\star[F, \xi] = [\Phi, \xi]$.

(3) Leibniz rule.
(4) $\mathrm{d}_A\Phi = 0$ at a critical point, and $\star$ commutes with $[\Phi, \cdot]$.

As expected, the whole tangent space to the gauge orbit, which is infinite-dimensional, lies in the kernel of H.

So, as Nash says, "we must also impose a gauge condition, which we take to be the standard one, $\mathrm{d}_A^\star a = 0$": the directions orthogonal to the orbit, since $\langle a, \mathrm{d}_A\xi\rangle = \langle \mathrm{d}_A^\star a, \xi\rangle$. The **index** of a critical point is then the dimension of a maximal subspace of $\ker \mathrm{d}_A^\star$ on which $\langle a, Ha\rangle < 0$.

Watch out: Why the index is finite

Nash says the index is finite "because ellipticity renders $\ker \mathrm{d}_A^\star$ finite dimensional". Taken literally that is not right: $\ker \mathrm{d}_A^\star$ on 1-forms is infinite-dimensional (it contains all the physical, transverse fluctuations). What is true is this. On $\ker \mathrm{d}_A^\star$, H agrees with

$$H + \mathrm{d}_A\mathrm{d}_A^\star = \left(\mathrm{d}_A^\star\mathrm{d}_A + \mathrm{d}_A\mathrm{d}_A^\star\right) + [\star F, \star\,\cdot\,],$$

a Laplacian plus a bounded zeroth-order term. That operator is elliptic on a compact surface, so its spectrum is discrete and goes to $+\infty$, as in Figure 8.85. Only finitely many eigenvalues are negative: the index is finite.

Worked example 8.180: Abelian critical points are minima

For U(1) all the fields are ordinary imaginary-valued forms and commutators vanish, so $K = 0$ and $H = \mathrm{d}^\star\mathrm{d}$, giving $\langle a, Ha\rangle = \|\mathrm{d}a\|^2 \geq 0$. The index is zero: every U(1) Yang–Mills connection is a minimum. The null directions in the gauge slice are the harmonic 1-forms, $\mathrm{d}a = \mathrm{d}^\star a = 0$, a space of dimension $b_1(\Sigma) = 2g$. They move the connection through the torus $H^1(\Sigma;\mathbb{R})/H^1(\Sigma;\mathbb{Z})$ of constant-curvature connections modulo gauge (the Jacobian of Σ). So in each degree the critical set of the U(1) theory is a torus T^{2g} of minima.

Worked example 8.181: Negative directions at a split U(2) connection

Take $E = L_1 \oplus L_2$ with degrees $d_1 > d_2$, so $\Phi = \mathrm{diag}(-i\lambda_1, -i\lambda_2)$ with $\lambda_1 > \lambda_2$. Write a fluctuation as

$$a = \begin{pmatrix} \alpha_1 & -\bar\beta \\ \beta & \alpha_2 \end{pmatrix},$$

with α_j imaginary 1-forms and β a complex 1-form with values in $\mathrm{Hom}(L_1, L_2)$ (the $(2,1)$ entry maps L_1 to L_2). The diagonal part commutes with Φ and feels only $\mathrm{d}^\star\mathrm{d} \geq 0$, exactly as in the abelian case. For the off-diagonal part, commuting with Φ multiplies the $(2,1)$ entry by $-i\lambda_2 + i\lambda_1$, so $K(a)_{21} = i(\lambda_1 - \lambda_2)\star\beta$. Using $\mathrm{tr}(XY) = -2\,\mathrm{Re}(\bar x y)$ for off-diagonal anti-Hermitian X, Y with lower entries x, y,

$$\langle a, K(a)\rangle_{\mathrm{off}} \overset{(1)}{=} 2(\lambda_1 - \lambda_2)\int \mathrm{Re}\big(i\,\bar\beta_k(\star\beta)_k\big)\mathrm{vol} \overset{(2)}{=} 4(\lambda_1 - \lambda_2)\int \mathrm{Im}(\bar\beta_1\beta_2)\,\mathrm{vol}$$
$$\overset{(3)}{=} 4(\lambda_1 - \lambda_2)\int \left(|\beta_z|^2 - |\beta_{\bar z}|^2\right)\mathrm{vol}.$$

(1) $\langle a, b\rangle = -\int \mathrm{tr}(a_k b_k)\mathrm{vol}$ and the trace formula above.
(2) $(\star\beta)_1 = -\beta_2$, $(\star\beta)_2 = \beta_1$, so $\bar\beta_k(\star\beta)_k = 2i\,\mathrm{Im}(\bar\beta_2\beta_1)$; multiply by i and take the real part.
(3) Write $\beta = \beta_z\mathrm{d}z + \beta_{\bar z}\mathrm{d}\bar z$ with $z = x + iy$, so $\beta_z = \tfrac12(\beta_1 - i\beta_2)$ and $\beta_{\bar z} = \tfrac12(\beta_1 + i\beta_2)$; expanding, $|\beta_z|^2 - |\beta_{\bar z}|^2 = \mathrm{Im}(\bar\beta_1\beta_2)$.

So the $(0,1)$-part of β gets a *negative* potential of strength proportional to $\lambda_1 - \lambda_2$, which is to say to $d_1 - d_2$. These $(0,1)$-forms are precisely the infinitesimal changes of the holomorphic structure $\bar\partial_A$ in the off-diagonal direction. Balancing the negative potential against the kinetic term $\|\mathrm{d}_A a\|^2$ (a Weitzenböck argument that we do not reproduce), the negative eigenspace of H on the gauge slice turns out to be exactly the space of holomorphic deformations $H^{0,1}(\Sigma, \mathrm{Hom}(L_1, L_2)) = H^1(\Sigma, L_1^* \otimes L_2)$.

Worked example 8.182: Counting with Riemann–Roch

Riemann–Roch for a holomorphic line bundle L of degree ℓ on a surface of genus g (Chapters 4 and 5) says $h^0(L) - h^1(L) = \ell + 1 - g$. Apply it to $L = L_1^* \otimes L_2$, of degree $\ell = d_2 - d_1 < 0$:

$$h^1(L_1^* L_2) \overset{(1)}{=} h^0(L_1^* L_2) - (d_2 - d_1) - 1 + g \overset{(2)}{=} d_1 - d_2 + g - 1.$$

(1) Riemann–Roch, solved for h^1.
(2) A line bundle of negative degree has no non-zero holomorphic sections (a section would have $\ell < 0$ zeros counted with multiplicity), so $h^0 = 0$.

Each complex dimension is two real ones, so

$$\mathrm{index}(L_1 \oplus L_2) = 2\,(d_1 - d_2 + g - 1) \qquad (d_1 > d_2). \tag{8.358}$$

On $\mathbb{CP}^1$ ($g = 0$) with $L_1^* L_2 = \mathcal{O}(-m)$, $m = d_1 - d_2$, we can see the $m - 1$ classes directly: by Serre duality $H^1(\mathcal{O}(-m)) \cong H^0(\mathcal{O}(m-2))^*$, the dual of the space of homogeneous polynomials of degree $m-2$ in two variables, which has dimension $m - 1$.
The simplest case is $\mathcal{O}(1) \oplus \mathcal{O}(-1)$ on $\mathbb{CP}^1$, of degree 0: $m = 2$, index 2, action $8\pi^2/\mathcal{A}_\Sigma$. Its single complex unstable direction deforms it into the trivial bundle through the non-split extension

$$0 \to \mathcal{O}(-1) \xrightarrow{s \mapsto (z_0 s,\, z_1 s)} \mathcal{O} \oplus \mathcal{O} \to \mathcal{O}(1) \to 0,$$

the tautological line sitting inside $\mathbb{C}^2$ (Chapter 5). Rolling downhill from $\mathcal{O}(1) \oplus \mathcal{O}(-1)$ along this direction ends at the flat connection on $\mathcal{O} \oplus \mathcal{O}$, with action 0.

Since $d_1 - d_2$ can be as large as we like, there are "critical points of arbitrarily large index", as Nash says; their actions $\frac{4\pi^2}{\mathcal{A}_\Sigma}(d_1^2 + d_2^2)$ grow without bound too.

8.18.9 Atiyah–Bott: the action is equivariantly perfect

We now have all the ingredients of the equivariant Morse series of S with respect to the full gauge group $\mathcal{G}$. Each critical set C_μ (the Yang–Mills connections of type μ) is a $\mathcal{G}$-invariant set of index λ_μ, contributing $t^{\lambda_\mu} P_t^{\mathcal{G}}(C_\mu)$. The theorem Nash records as (8.203) is

$$M_t^{\mathcal{G}}(S) \;:=\; \sum_\mu t^{\lambda_\mu} P_t^{\mathcal{G}}(C_\mu) \;=\; P_t^{\mathcal{G}}(\mathcal{A}) \;=\; P_t(B\mathcal{G}). \tag{8.359}$$

The second equality holds because $\mathcal{A}$ is contractible. The first is the content: the Morse inequalities are equalities.

Key idea: Why no cancellations can occur

In a Morse complex, cancellations between critical points of adjacent index are what make $M_t \neq P_t$. Atiyah and Bott showed they cannot happen here by a criterion that is visible in the rotating sphere of Example 8.165: at each critical set the stabiliser contains a circle that acts on the negative directions *without fixed vectors*. At $L_1 \oplus L_2$, the circle $\mathrm{diag}(e^{i\theta}, 1)$ fixes the connection but acts on the negative directions $\beta \in H^1(L_1^* L_2)$ by $\beta \mapsto e^{-i\theta}\beta$. Then the equivariant Euler class of the negative bundle is not a zero-divisor, each stratum's Thom–Gysin sequence splits into short exact sequences, and the Poincaré series simply add up. At the poles of the rotating sphere the rotation acts on the tangent plane with weight one: the same mechanism, in miniature.

Worked example 8.183: $\mathrm{U}(1)$: perfect in every genus

In each degree there is one critical set, the torus T^{2g} of minima (Example 8.180), with stabiliser the constant $\mathrm{U}(1)$. So

$$M_t^{\mathcal{G}}(S) \overset{(1)}{=} t^0\, P_t(T^{2g})\, P_t(B\mathrm{U}(1)) \overset{(2)}{=} \frac{(1+t)^{2g}}{1-t^2} \overset{(3)}{=} P_t(B\mathcal{G}).$$

(1) The constant $\mathrm{U}(1)$ acts trivially on the connection and $\mathcal{G}/\mathrm{U}(1)$ acts freely, so the equivariant cohomology of the critical set is $H^*(T^{2g}) \otimes H^*(B\mathrm{U}(1))$.
(2) $P_t(T^{2g}) = (1+t)^{2g}$ and $P_t(B\mathrm{U}(1)) = 1/(1-t^2)$.
(3) (8.356).

Worked example 8.184: $\mathrm{U}(2)$ on the sphere: the whole check

Take $\Sigma = S^2$ and a $\mathrm{U}(2)$ bundle of degree d. By Grothendieck and Example 8.174, the critical points are the splittings $\mathcal{O}(d_1) \oplus \mathcal{O}(d_2)$ with $d_1 \geq d_2$, $d_1 + d_2 = d$, each a single gauge orbit. The stabilisers and indices are:

- $d_1 = d_2$ (d even): stabiliser $\mathrm{U}(2)$, index 0, contribution $P_t(B\mathrm{U}(2)) = \dfrac{1}{(1-t^2)(1-t^4)}$;
- $d_1 > d_2$: stabiliser $\mathrm{U}(1) \times \mathrm{U}(1)$, index $2(d_1 - d_2 - 1)$ by (8.358), contribution $\dfrac{t^{2(d_1-d_2-1)}}{(1-t^2)^2}$.

Degree 0. The splittings are $(m, -m)$, $m \geq 0$, with $d_1 - d_2 = 2m$:

$$M_t^{\mathcal{G}} \overset{(1)}{=} \frac{1}{(1-t^2)(1-t^4)} + \sum_{m\geq 1} \frac{t^{2(2m-1)}}{(1-t^2)^2} \overset{(2)}{=} \frac{1}{(1-t^2)(1-t^4)} + \frac{t^2}{(1-t^2)^2(1-t^4)}$$
$$\overset{(3)}{=} \frac{(1-t^2)+t^2}{(1-t^2)^2(1-t^4)} = \frac{1}{(1-t^2)^2(1-t^4)} \overset{(4)}{=} P_t(B\mathcal{G}).$$

(1) The list above.
(2) Geometric series: $\sum_{m\geq 1} t^{4m-2} = t^2/(1-t^4)$.
(3) Common denominator.
(4) (8.355).

Expanding both sides term by term makes the bookkeeping visible:

degree $2j$	0	2	4	6	8
minimum $\mathcal{O}\oplus\mathcal{O}$ (index 0)	1	1	2	2	3
$\mathcal{O}(1)\oplus\mathcal{O}(-1)$ (index 2)	0	1	2	3	4
$\mathcal{O}(2)\oplus\mathcal{O}(-2)$ (index 6)	0	0	0	1	2
$\mathcal{O}(3)\oplus\mathcal{O}(-3)$ (index 10)	0	0	0	0	0
total $= b_{2j}(B\mathcal{G})$	1	2	4	6	9

Degree 1. The splittings are $(m+1,-m)$, $m\geq 0$, with $d_1-d_2=2m+1$ and index $4m$. There is no splitting with $d_1=d_2$, so the minimum is $\mathcal{O}(1)\oplus\mathcal{O}$ (index 0). Then

$$M_t^{\mathcal{G}}=\sum_{m\geq 0}\frac{t^{4m}}{(1-t^2)^2}=\frac{1}{(1-t^2)^2(1-t^4)}=P_t(B\mathcal{G}),$$

perfect again, with a completely different set of critical points.

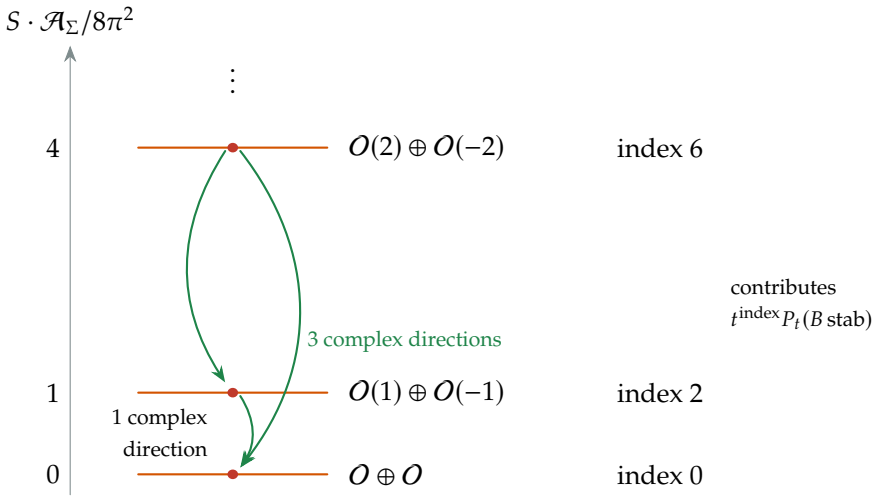


Figure 8.90: The critical points of S for U(2) bundles of degree 0 on the sphere: an infinite ladder of splittings $\mathcal{O}(m)\oplus\mathcal{O}(-m)$ with action $8\pi^2m^2/\mathcal{A}_\Sigma$ and index $2(2m-1)$. Arrows show downhill directions, which deform the splitting into a less unbalanced bundle. Each rung contributes its full equivariant cohomology; nothing cancels, and the ladder adds up to $P_t(B\mathcal{G})$ (Example 8.184).

Worked example 8.185: What perfection buys: the cohomology of a moduli space

Run the argument backwards. For a U(2) bundle of odd degree on a surface of genus g, the lowest stratum is the semistable one, $\mathcal{C}_{ss}$ (the projectively flat connections; for odd degree semistable means stable). The other critical sets are splittings $L_1\oplus L_2$ with $d_1-d_2=2m+1$, $m\geq 0$, of index $2(2m+g)$ by (8.358), each contributing $\left((1+t)^{2g}/(1-t^2)\right)^2$ (two copies of the U(1) answer of Example 8.183). Perfection gives

$$P_t^{\mathcal{G}}(\mathcal{C}_{ss})\overset{(1)}{=}P_t(B\mathcal{G})-\sum_{m\geq 0}t^{2(2m+g)}\frac{(1+t)^{4g}}{(1-t^2)^2}\overset{(2)}{=}\frac{(1+t)^{2g}\left[(1+t^3)^{2g}-t^{2g}(1+t)^{2g}\right]}{(1-t^2)^2(1-t^4)}.$$

(1) (8.359), solved for the bottom stratum.
(2) (8.356) and $\sum_m t^{4m}=1/(1-t^4)$.

The stabiliser on $\mathcal{C}_{ss}$ is just the constant U(1) (stable bundles have only scalar automorphisms), so dividing by $P_t(B\mathrm{U}(1))=1/(1-t^2)$ gives the moduli space of stable bundles,

and removing the Jacobian factor $(1+t)^{2g}$ (fixing the determinant) gives

$$P_t\big(\mathcal{N}_0(2,\text{odd})\big) = \frac{(1+t^3)^{2g} - t^{2g}(1+t)^{2g}}{(1-t^2)(1-t^4)}. \tag{8.360}$$

Checks: for $g=0$ the numerator is $1-1=0$, and indeed there are no stable rank-2 bundles on $\mathbb{CP}^1$ (Grothendieck). For $g=1$ the numerator is $1+2t^3+t^6-t^2-2t^3-t^4=(1-t^2)(1-t^4)$, so $P_t=1$: a single point, Atiyah's unique stable bundle on an elliptic curve. For $g=2$, long division gives $P_t = 1+t^2+4t^3+t^4+t^6$, the Betti numbers of a compact complex threefold (the intersection of two quadrics in $\mathbb{CP}^5$). None of these numbers was put in by hand.

Physical picture: Why two-dimensional Yang–Mills is special

Two-dimensional Yang–Mills has no propagating gluons: a gauge field in two dimensions has no transverse polarisations. Correspondingly its quantum theory is exactly solvable (Migdal, Witten): the partition function on a surface of genus g and area $\mathcal{A}_\Sigma$ is a sum over irreducible representations R of the gauge group, $Z=\sum_R(\dim R)^{2-2g}\exp\!\big(-\frac{e^2\mathcal{A}_\Sigma}{2}C_2(R)\big)$, with C_2 the quadratic Casimir. Witten showed that its weak-coupling limit localises on the Yang–Mills critical points we have just classified, and recovers the Atiyah–Bott results (and the intersection theory of the moduli space) from the path integral. The Morse theory is the classical shadow of an exactly solvable quantum field theory.

Try it: The torus

For U(1) on a flat torus of area $\mathcal{A}$, list the critical points of S, their action and index, and check (8.359).
Answer: in degree d the critical set is the torus T^2 of constant-curvature connections, $S=4\pi^2d^2/\mathcal{A}$, index 0; $M_t^{\mathcal{G}}=(1+t)^2/(1-t^2)=P_t(B\mathcal{G})$ from (8.356) with $g=1$.

8.18.10 Four dimensions: Yang–Mills on S^4

Finally Nash returns to where the chapter began: SU(2) Yang–Mills on S^4. Here the orbit space is again topologically rich, but (in 1983) no Morse theory was available and only the minima, the instantons, were known.

Worked example 8.186: The topology of $\mathcal{A}/\mathcal{G}_0$ on S^4

For a bundle of instanton number k over S^4, the based gauge group is (up to homotopy) the space of based maps $S^4\to \mathrm{SU}(2)=S^3$, i.e. Ω^4S^3. Then

$$\mathcal{A}_k/\mathcal{G}_0 \overset{(1)}{\simeq} B\mathcal{G}_0 \overset{(2)}{\simeq} \Omega^3_kS^3, \qquad \pi_i(\Omega^3S^3) \overset{(3)}{=} \pi_{i+3}(S^3).$$

(1) Free action on the contractible $\mathcal{A}_k$, as in two dimensions.
(2) $B(\Omega Y)\simeq Y$ with $Y=\Omega^3S^3$ (the component labelled by k).
(3) A based map of S^i into Ω^3S^3 is a based map of S^{i+3} into S^3.

The groups $\pi_{i+3}(S^3)$ are $\mathbb{Z}_2,\mathbb{Z}_2,\mathbb{Z}_{12},\mathbb{Z}_2,\mathbb{Z}_2,\mathbb{Z}_3,\mathbb{Z}_{15},\dots$ for $i=1,2,3,\dots$: the homotopy groups of spheres, all non-zero. So $\mathcal{A}_k/\mathcal{G}_0$ has "plenty of cohomology" (mostly torsion; with $\mathbb{Z}_p$ coefficients it is enormous). The Atiyah–Jones conjecture, later proved by Boyer, Hurtubise, Mann and Milgram, says that the instanton moduli space $\mathcal{M}_k$ captures more and more of this topology as k grows: the inclusion $\mathcal{M}_k\to\mathcal{A}_k/\mathcal{G}_0$ is an isomorphism on

homology up to a degree that grows linearly with k.

Yet Nash writes that "no Morse-type results hold". The reason is the symmetry he singles out: *conformal invariance*.

Worked example 8.187: Which actions are conformally invariant

Replace the metric g on an n-manifold by $g' = \Omega^2 g$ with $\Omega > 0$ a function. Inverse metrics scale as Ω^{-2} and the volume form as Ω^n. So

$$\int |F|^2_{g'}\,\mathrm{dvol}_{g'} \overset{(1)}{=} \int \tfrac{1}{2} g'^{\mu\alpha} g'^{\nu\beta} F_{\mu\nu} F_{\alpha\beta}\,\Omega^n \mathrm{dvol}_g \overset{(2)}{=} \int \Omega^{n-4}\,|F|^2_g\,\mathrm{dvol}_g,$$
$$\int |\mathrm{d}\phi|^2_{g'}\,\mathrm{dvol}_{g'} \overset{(3)}{=} \int \Omega^{n-2}\,|\mathrm{d}\phi|^2_g\,\mathrm{dvol}_g.$$

(1) The norm of a 2-form uses two inverse metrics; $\mathrm{dvol}_{g'} = \Omega^n \mathrm{dvol}_g$.
(2) Each inverse metric gives Ω^{-2}.
(3) A 1-form norm uses one inverse metric.

So the Yang–Mills action is conformally invariant exactly in $n = 4$, and the sigma-model energy $\int |\mathrm{d}\phi|^2$ exactly in $n = 2$. In $n = 3$ neither term is invariant, and they scale oppositely (Ω^{-1} and Ω^{+1}): this is Derrick's balance of Section 8.12.2, and the Higgs vacuum value C provides the length scale $1/C$ of a monopole. In $n = 2$ the Yang–Mills action scales as Ω^{-2} and the area $\mathcal{A}_\Sigma$ sets the scale, as in (8.354).

The conformal group of S^4 is non-compact (it contains the dilations $x \mapsto sx$ of $\mathbb{R}^4 = S^4 \setminus \{\infty\}$). Dilating an instanton does not change its action, so one can shrink it to a point. For the BPST instanton of size ρ the action density is $s(r) = 48\rho^4/(r^2+\rho^2)^4$ (normalised so that $\int_{\mathbb{R}^4} s\,\mathrm{d}^4x = 8\pi^2$), and

$$\int_{\mathbb{R}^4} s\,\mathrm{d}^4x \overset{(1)}{=} 2\pi^2 \int_0^\infty \frac{48\rho^4\, r^3}{(r^2+\rho^2)^4}\mathrm{d}r \overset{(2)}{=} 48\pi^2\rho^4 \int_0^\infty \frac{u\,\mathrm{d}u}{(u+\rho^2)^4} \overset{(3)}{=} 48\pi^2\rho^4 \cdot \frac{1}{6\rho^4} = 8\pi^2,$$

(1) Radial integration in $\mathbb{R}^4$: $\mathrm{d}^4x = 2\pi^2 r^3 \mathrm{d}r$.
(2) Substitute $u = r^2$, $\mathrm{d}u = 2r\,\mathrm{d}r$.
(3) $\int_0^\infty u(u+a)^{-4}\mathrm{d}u = a^{-2}\,\Gamma(2)\Gamma(2)/\Gamma(4) = 1/(6a^2)$ with $a = \rho^2$.

independently of ρ. As $\rho \to 0$ the density concentrates at the origin: a family of minima of constant action with no limit among smooth connections. It converges to a flat connection away from the origin plus a "bubble" of action $8\pi^2$ at the origin.

Key idea: Why four dimensions resisted Morse theory

A min-max sequence on $\mathcal{A}_k/\mathcal{G}_0$ over S^4 can lose compactness by bubbling off instantons (or anti-instantons), each carrying action at least $8\pi^2$. Unlike the monopole case, there is no energy threshold below which a non-trivial min-max level is guaranteed to sit: the non-trivial topology of $\Omega^3 S^3$ may be "explained" entirely by configurations near the boundary of the space, where instantons have shrunk to points. So the topology does not force non-minimal critical points in any obvious way, and none were known in 1983.

Two later results fill in the picture. Bourguignon, Lawson and Simons showed that on S^4 every *stable* SU(2) Yang–Mills connection (index zero) is self-dual or anti-self-dual: there are no local minima other than the instantons. And Sibner, Sibner and Uhlenbeck (1989) proved, by a min-max argument made compact by imposing a symmetry that prevents bubbling,

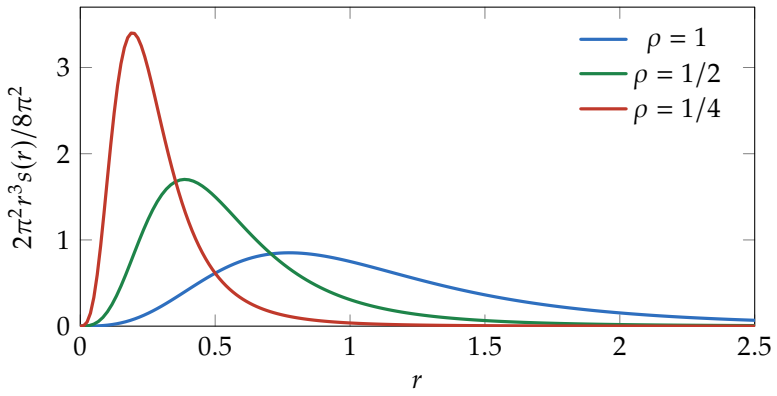


Figure 8.91: Bubbling. The radial distribution of the action of a BPST instanton for three sizes. Each curve has area one (total action $8\pi^2$), but as $\rho \to 0$ the action piles up at the origin. A sequence of shrinking instantons has constant action and zero gradient, yet no smooth limit: the Palais–Smale condition fails in four dimensions, by concentration rather than by escape to infinity.

that non-self-dual solutions of the Yang–Mills equations on S^4 do exist. They are necessarily unstable saddle points.

The two-dimensional sigma model

Nash's last remark is that the two-dimensional sigma model, also conformally invariant (Example 8.187), likewise "only has absolute minima". Take maps $\vec{n} : S^2 \to S^2 \subset \mathbb{R}^3$ with energy $E = \frac{1}{2}\int |\partial_1\vec{n}|^2 + |\partial_2\vec{n}|^2 \,\mathrm{d}^2x$ (the O(3) model; isothermal coordinates x^1, x^2 on the domain), and degree $N = \frac{1}{4\pi}\int \vec{n}\cdot(\partial_1\vec{n}\times\partial_2\vec{n})\,\mathrm{d}^2x$.

Worked example 8.188: The Bogomolny bound of the sigma model

$$0 \le \frac{1}{2}\int |\partial_1\vec{n} \pm \vec{n}\times\partial_2\vec{n}|^2 \mathrm{d}^2x \overset{(1)}{=} \frac{1}{2}\int \left(|\partial_1\vec{n}|^2 + |\partial_2\vec{n}|^2\right)\mathrm{d}^2x \pm \int \partial_1\vec{n}\cdot(\vec{n}\times\partial_2\vec{n})\,\mathrm{d}^2x$$
$$\overset{(2)}{=} E \mp \int \vec{n}\cdot(\partial_1\vec{n}\times\partial_2\vec{n})\,\mathrm{d}^2x \overset{(3)}{=} E \mp 4\pi N.$$

(1) Expand the square; $|\vec{n}\times\partial_2\vec{n}| = |\partial_2\vec{n}|$ because $\vec{n}\cdot\partial_2\vec{n} = 0$ (from $|\vec{n}| = 1$).
(2) Cyclic property of the triple product: $\partial_1\vec{n}\cdot(\vec{n}\times\partial_2\vec{n}) = \vec{n}\cdot(\partial_2\vec{n}\times\partial_1\vec{n}) = -\vec{n}\cdot(\partial_1\vec{n}\times\partial_2\vec{n})$.
(3) Definition of N.

So $E \ge 4\pi|N|$, with equality iff $\partial_1\vec{n} = \mp\vec{n}\times\partial_2\vec{n}$: the Cauchy–Riemann equations for $\vec{n}$ viewed (by stereographic projection) as a function $w(z)$. Holomorphic maps (rational functions of degree N) and antiholomorphic maps are the absolute minima: the exact analogue of instantons and of BPS monopoles.

Worked example 8.189: Every critical point is a minimum

Let $\vec{n}$ be any critical point (a harmonic map). Put $z = x^1 + ix^2$, $\vec{n}_z = \frac{1}{2}(\partial_1\vec{n} - i\partial_2\vec{n})$, and form $\Psi = \vec{n}_z\cdot\vec{n}_z$. Then

$$\partial_{\bar{z}}\Psi \overset{(1)}{=} 2\,\vec{n}_z\cdot\vec{n}_{z\bar{z}} \overset{(2)}{=} 2\,\vec{n}_z\cdot\left(-\tfrac{1}{4}|\nabla\vec{n}|^2\,\vec{n}\right) \overset{(3)}{=} 0.$$

(1) Product rule.

(2) $\vec{n}_{z\bar{z}} = \frac{1}{4}\nabla^2\vec{n}$, and the harmonic map equation for maps into the unit sphere is $\nabla^2\vec{n} = -|\nabla\vec{n}|^2\vec{n}$ (the Laplacian is normal to the sphere).
(3) $\vec{n}\cdot\vec{n}_z = 0$, since $|\vec{n}| = 1$.

So $\Psi\,\mathrm{d}z^2$ is a *holomorphic quadratic differential* on the domain sphere. On S^2 the only one is zero (it would be a holomorphic section of K^2, a line bundle of degree $-4 < 0$). Hence $\Psi = \frac{1}{4}\big(|\partial_1\vec{n}|^2 - |\partial_2\vec{n}|^2 - 2i\,\partial_1\vec{n}\cdot\partial_2\vec{n}\big) = 0$: the map is conformal, so it is holomorphic or antiholomorphic wherever $\mathrm{d}\vec{n} \neq 0$, and it saturates the bound of Example 8.188. *Every* harmonic map $S^2 \to S^2$ is an absolute minimum. (Eells and Wood; Lemaire.)

functional	scale	how condition (C) fails	non-minimal critical points
YMH, $\mathbb{R}^3$, $\lambda = 0$	$1/C$	monopoles escape to infinity	yes: infinitely many, unbounded energy (Taubes)
YM, surface Σ	area $\mathcal{A}_\Sigma$	it does not fail	yes: splittings, arbitrarily large index; equivariantly perfect (Atiyah–Bott)
YM, S^4	none (conformal)	instantons bubble	none known in 1983; exist (Sibner–Sibner–Uhlenbeck 1989)
sigma model, $S^2 \to S^2$	none (conformal)	harmonic spheres bubble	none: every critical point is a minimum

Reading Nash (pp. 257–258, (8.201)–(8.206))

For U(n) Yang–Mills on a Riemann surface Nash writes the action (8.201), the equations (8.202) with the observation that $\star F$ is a covariantly constant section of $\mathrm{ad}\,P$, mentions the holomorphic reformulation, identifies the orbit space for genus zero as $B\mathcal{G} = \Omega G$, and states the Atiyah–Bott equivariant perfection (8.203). He expands the action to second order (8.204), defines the Hessian (8.205) and writes it as $H = \mathrm{d}_A^\star\mathrm{d}_A + \star(\star F(A)\cdot)$ (8.206), which must be read as $\mathrm{d}_A^\star\mathrm{d}_A + [\star F, \star\,\cdot]$; he imposes the gauge $\mathrm{d}_A^\star a = 0$ and notes that the index is finite and can be arbitrarily large. The finiteness comes from ellipticity of H on the slice, not from finite dimensionality of $\ker \mathrm{d}_A^\star$ (see the pitfall above). For SU(2) on S^4 he says no Morse theory is known, only minima, and he links this to conformal invariance, which the two-dimensional sigma model shares. We added the splitting of Yang–Mills connections and their action values, the minimum bound, the holomorphic dictionary (Grothendieck, Narasimhan–Seshadri, Harder–Narasimhan), the loop-space identification, the explicit Hessian with its gauge zero modes and negative directions, the Riemann–Roch index formula, the full U(2) check and the moduli space Poincaré polynomial, the conformal weights, bubbling, the sigma-model bound and the harmonic-map theorem, and the later four-dimensional results.

Recap

- Morse theory: topology forces critical points; $M_t - P_t = (1+t)Q_t$. In infinite dimensions one needs an elliptic Hessian (finite index), the Palais–Smale condition (compactness), and a way of dividing out the gauge group (free action of $\mathcal{G}_0$, or equivariant cohomology).
- Min-max over non-contractible families (the mountain pass) produces saddles.
- YMH on $\mathbb{R}^3$: E descends to $C_k = \mathcal{A}^k_{YMH}/\mathcal{G}$, the sectors are labelled by the magnetic charge, the minima $\mathcal{M}'_k$ form a Morse–Bott critical manifold with Hessian $D_1^\star D_1$,

and $\pi_0(\mathcal{G}) = \pi_3(S^3) = \mathbb{Z}$ gives a non-contractible loop. Min-max over it gives Taubes' monopole–antimonopole saddle at a level below $8\pi C$, where escape to infinity is energetically forbidden. Taubes: infinitely many non-minimal critical points, of unbounded energy.

- YM on a surface: $\mathrm{d}_A(\star F) = 0$, so Yang–Mills connections are sums of constant-curvature pieces with $S = \frac{4\pi^2}{\mathcal{A}_\Sigma} \sum d_j^2/n_j$. The Hessian is $\mathrm{d}_A^\star \mathrm{d}_A + [\star F, \star\,\cdot\,]$; the index of $L_1 \oplus L_2$ is $2(d_1 - d_2 + g - 1)$ by Riemann–Roch. S is equivariantly perfect: $M_t^{\mathcal{G}}(S) = P_t(B\mathcal{G})$, checked in full for U(1) and for U(2) on S^2, and it computes the cohomology of moduli spaces of stable bundles.
- YM on S^4 and the sigma model on S^2 are conformally invariant; compactness fails by bubbling, and only minima were known (for the sigma model, only minima exist).

Chapter summary

This chapter asked one question of gauge theory in several guises: *what does the space of all gauge fields look like, and where on it do the solutions of the field equations sit?* The answers came in three layers.

1. *The arena* (Nash §§1–2). The space $\mathcal{A}$ of connections is affine and therefore contractible, but the gauge group $\mathcal{G}$ is topologically rich: it is an iterated loop space of the structure group, with homotopy groups made of homotopy groups of spheres. So the physical configuration space $\mathcal{A}/\mathcal{G}$ (the space the path integral is really over) is topologically rich too. The integers that label its components, the instanton number and the magnetic charge, are Chern numbers, and they are also winding numbers of gauge transformations or Higgs fields at infinity. Secondary classes (Chern–Simons forms) appear when a characteristic form is exact: they are the "half" of an instanton number that lives on the boundary, and the large gauge transformations that shift them produce θ-vacua.

2. *The special solutions* (Nash §§3–5). Completing a square gives a topological lower bound on the action or energy, saturated exactly by a first-order equation: self-duality $F = \pm \star F$ for instantons (bound $8\pi^2|k|$), the Bogomolny equation $F = -\star \mathrm{d}_A\phi$ for monopoles (bound $4\pi C|k|$). The two are the same equation, the second being the first with a translation (or rotation) symmetry imposed. Their solution spaces are finite-dimensional moduli spaces whose dimensions are indices of elliptic deformation complexes ($8k-3$ for SU(2) instantons on S^4, $4k$ for framed SU(2) monopoles), with explicit descriptions (ADHM for instantons, rational maps for monopoles). The monopole moduli spaces carry a natural hyperkähler metric whose geodesics are the slow motions of monopoles: 90° scattering in head-on collisions.

3. *All the critical points* (Nash §6). The topology of $\mathcal{A}/\mathcal{G}$ forces critical points beyond the minima, provided the functional behaves well enough for Morse theory to apply. For monopoles it does (Taubes), for two-dimensional Yang–Mills it does perfectly (Atiyah–Bott), and for four-dimensional Yang–Mills and the two-dimensional sigma model conformal invariance (bubbling) gets in the way.

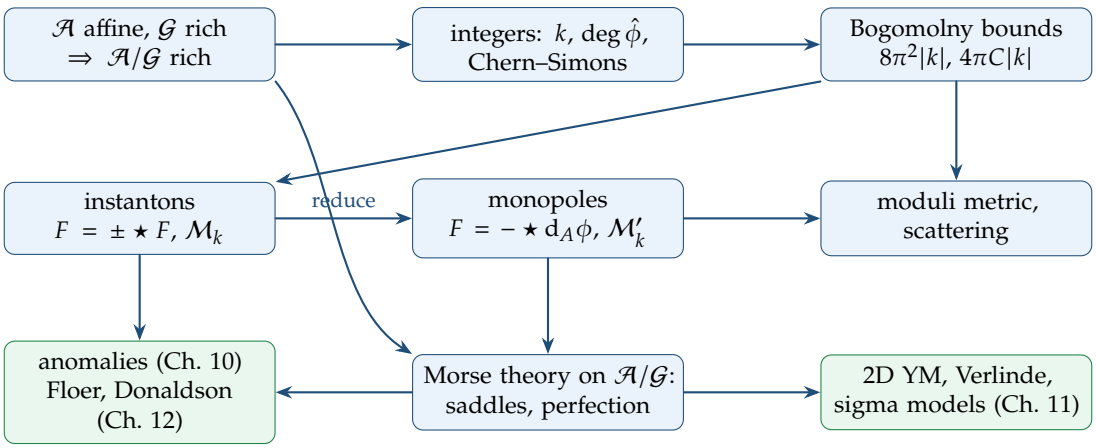


Figure 8.92: How the ideas of Chapter 8 fit together, and where they go next (green).

The tables below list every concept of the chapter, what it means operationally (how you compute or test for it), and where it is used again.

The topology of gauge fields and secondary classes

concept	operational meaning	used later
space of connections $\mathcal{A}$	every connection is $A_0 + a$, a an $\operatorname{ad} P$-valued 1-form; straight lines stay in $\mathcal{A}$, so $\mathcal{A}$ is contractible	Chapter 10 (determinant line over $\mathcal{A}/\mathcal{G}$), Chapter 12
gauge group $\mathcal{G}$, based $\mathcal{G}_0$	sections g of $\operatorname{Ad} P$; $\mathcal{G}_0$ has $g(x_0) = \mathbb{1}$ and acts freely (a fixed point is $\mathrm{d}_A g = 0$, hence $g \equiv \mathbb{1}$)	Chapter 10, Chapter 12
finite action at infinity	$F \to 0$ faster than r^{-2} forces $A \to g^{-1}\mathrm{d}g$: the field is pure gauge on the sphere at infinity	instantons, monopoles; Chapter 12
winding number of $g : S^3 \to \mathrm{SU}(2)$	$\frac{1}{24\pi^2}\int \operatorname{tr}(g^{-1}\mathrm{d}g)^3 \in \mathbb{Z}$; count preimages of a point with signs	Chapter 10 (global anomalies)
instanton number k	$\frac{1}{8\pi^2}\int \operatorname{tr} F \wedge F$ = winding of the gauge at infinity = c_2 (sign by orientation)	Chapter 10 (index of $\not{D}_A$), Chapter 12
large gauge transformation, Chern–Simons number	$\frac{1}{8\pi^2}\int_{S^3} \operatorname{tr}(A\mathrm{d}A + \frac{2}{3}A^3)$ jumps by the winding of g; tunnelling between vacua	Chapter 10, Chapter 12
θ-vacua	superpose the vacua $\lvert n\rangle$ with phases $e^{in\theta}$; adds θk to the action	Chapter 10
Chern–Weil and transgression	invariant polynomial $P(F)$ is closed; on P it is exact, $P(F) = \mathrm{d}\,TP(A)$	Chapter 10 (descent)
Chern–Simons form, secondary class	TP restricted where $P(F) = 0$; defines an $\mathbb{R}/\mathbb{Z}$ class; gauge change shifts it by an integer	Chapter 12 (Chern–Simons theory, Floer)
cohomology of $\mathcal{G}$ and $B\mathcal{G}$	evaluation map $\mathcal{G} \times X \to G$, pull back generators of $H^*(G)$, integrate over cycles of X	Chapter 10
$\mathcal{A}/\mathcal{G}_0 \simeq B\mathcal{G}_0$	free action on a contractible space; on S^4, $\mathcal{A}_k/\mathcal{G}_0 \simeq \Omega^3_k S^3$	Chapter 10, Chapter 12

Instantons and their moduli

concept	operational meaning	used later
Yang–Mills equations	$\mathrm{d}_A^\star F = 0$ from varying $S = \|F\|^2$; with Bianchi $\mathrm{d}_A F = 0$ a non-linear Hodge theory	Chapter 12
gauge fixing, slice	$T_A\mathcal{A} = \operatorname{Im}\mathrm{d}_A \oplus \operatorname{Ker}\mathrm{d}_A^\star$; the gauge-fixed operator is elliptic; Gribov: no global slice	Chapter 10 (Faddeev–Popov)
self-duality and bound	split $F = F^+ + F^-$; $S = \|F^\pm\|^2 \pm 8\pi^2 k \geq 8\pi^2\|k\|$, equality iff $F = \pm \star F$	Chapter 12 (Donaldson)
BPST instanton	$A = \frac{r^2}{r^2+\lambda^2} g^{-1}\mathrm{d}g$ with g the unit quaternion $x/\|x\|$ and size λ; 't Hooft ansatz for many	Chapter 10 (fermion zero modes)
deformation complex	$0 \to \Omega^0 \xrightarrow{\mathrm{d}_A} \Omega^1 \xrightarrow{\mathrm{d}_A^-} \Omega^2_- \to 0$; H^1 = tangent space to $\mathcal{M}_k$	Chapter 12 (Floer, Donaldson)
dimension of $\mathcal{M}_k$	index of the rolled-up complex by Atiyah–Singer: $8k-3$ for SU(2) on S^4	Chapter 12
reducible connections	stabiliser bigger than the centre; cone-like singularities of $\mathcal{M}_k$	Chapter 12 (Donaldson's theorem)
$\mathcal{M}_1$ and ADHM	$\mathcal{M}_1 = \mathbb{R}^4 \times \mathbb{R}^+ = H^5$ (centre and size); ADHM builds all instantons from matrices	Chapter 12

Monopoles and symmetric instantons

concept	operational meaning	used later
Derrick's argument	scale $x \to \lambda x$; pure YM energy in 3D scales as λ^{-1}, so no static solitons without a Higgs field	Section 8.18.10 (conformal weights), Chapter 11
Dirac monopole	U(1) bundle over S^2, patches glued by $e^{in\varphi}$; flux $2\pi n$, $c_1 = n$	Example 8.175; Chapter 10
YMH energy and equations	$E = \frac{1}{2}\int \|F\|^2 + \|\mathrm{d}_A\phi\|^2 + \lambda(\|\phi\|^2 - C^2)^2$; vary to get (8.274)	Section 8.18.4
magnetic charge k	$k = \deg(\hat\phi_\infty : S^2 \to S^2) = c_1$ of the unbroken U(1) at infinity	monopole moduli; Chapter 12
Bogomolny equation and bound	$F = -\star\,\mathrm{d}_A\phi$ (for $k > 0$); $E = \frac{1}{2}\left\|F + \star\mathrm{d}_A\phi\right\|^2 + 4\pi C k$	moduli, Morse theory
compact M is trivial	Stokes kills the boundary term: $k = 0$, $F = 0$, $\mathrm{d}_A\phi = 0$	–
Prasad–Sommerfield monopole	$\phi = \left(C\coth Cr - \frac{1}{r}\right)\hat x^a T_a$, A from $K = Cr/\sinh Cr$; $E = 4\pi C$	zero modes, scattering
unitary gauge, Dirac string	rotate $\hat\phi$ to T_3; the gauge field becomes a Dirac monopole with a string	Chapter 10
dimensional reduction	∂_0-invariant self-dual field on $\mathbb{R}^4$ with $A_0 = \phi$ is a Bogomolny solution	Chapter 12
hyperbolic monopoles	$\mathbb{R}^4 \setminus \mathbb{R}^2 \simeq S^1 \times H^3$ conformally; S^1-invariant instantons are monopoles on H^3, $c_2 = 2nk$	–

Monopole moduli and scattering

concept	operational meaning	used later
monopole deformation complex	$D_0 f = (\mathrm{d}_A f, [\phi, f])$, $D_1(a,\eta) = \mathrm{d}_A a + \star \mathrm{d}_A \eta + \star[a,\phi]$; tangent space $\operatorname{Ker} D_1 \cap \operatorname{Ker} D_0^\star$	Section 8.18.4 (Hessian $D_1^\star D_1$)
quaternionic Dirac operator	roll up $D_0^\star \oplus D_1$ into $\mathcal{D}$; $\mathcal{D}\mathcal{D}^\dagger = -D^2 - \mathrm{ad}_\phi^2 \geq 0$ kills the cokernel	Chapter 10 (Callias index)
$\dim \mathcal{M}_k' = 4k$, $\mathcal{M}_k$, $\mathcal{M}_k^0$	$4k$ framed; drop the phase $(4k-1)$; drop centre and phase $(4k-4)$	scattering
rational maps	Donaldson: $\mathcal{M}_k' \cong$ based rational maps $\mathbb{CP}^1 \to \mathbb{CP}^1$ of degree k	Chapter 12
no static forces	$E = 4\pi C k$ for every k-monopole: magnetic repulsion cancels Higgs attraction	Taubes' saddle
geodesic approximation	slow motion stays near $\mathcal{M}_k'$; kinetic energy restricted there is a metric; motion is geodesic	Chapter 7 (energy of paths)
$k=1$ metric, dyons	$\mathbb{R}^3 \times S^1$ with $4\pi C(\mathrm{d}\vec{x}^2 + C^{-2}\mathrm{d}\chi^2)$; motion in χ is electric charge	–
hyperkähler metric	three complex structures I, J, K with quaternion relations, all Kähler; the moduli metric is hyperkähler	Chapter 12
Atiyah–Hitchin manifold	$\mathcal{M}_2^0$ with metric $f^2\mathrm{d}r^2 + a^2\sigma_1^2 + b^2\sigma_2^2 + c^2\sigma_3^2$ solving the AH ODEs; bolt at $r=\pi$	–
90° scattering, dyon production	head-on geodesics pass through the bolt and turn by 90°; glancing ones convert orbital angular momentum into relative charge	–

Critical point theory

concept	operational meaning	used later
Palais–Smale condition	every sequence with f bounded and $\nabla f \to 0$ has a convergent subsequence; fails by escape or bubbling	Chapter 12 (Floer)
mountain pass, min-max	$c = \inf_\gamma \max f \circ \gamma$ over a non-contractible family is a critical value (index one)	Chapter 12
YMH energy as Morse function	sectors C_k; minima $\mathcal{M}_k'$ Morse–Bott with Hessian $D_1^\star D_1$; Taubes' saddles below $8\pi C$	sphalerons (electroweak)
2D Yang–Mills critical points	$\mathrm{d}_A(\star F) = 0$; sums of constant-curvature bundles, $S = \frac{4\pi^2}{\mathcal{A}_\Sigma}\sum d_j^2/n_j$	Chapter 11 (Verlinde), Chapter 12
Hessian of S	$H = \mathrm{d}_A^\star \mathrm{d}_A + [\star F, \star\,\cdot\,]$ on $\operatorname{Ker}\mathrm{d}_A^\star$; index of $L_1 \oplus L_2$ is $2(d_1 - d_2 + g - 1)$	Chapter 12 (relative Morse index)
Atiyah–Bott perfection	$\sum_\mu t^{\lambda_\mu} P_t^{\mathcal{G}}(C_\mu) = P_t(B\mathcal{G})$; gives the Betti numbers of moduli of stable bundles	Chapter 11, Chapter 12
conformal invariance, bubbling	$\int \lvert F\rvert^2$ invariant in 4D, $\int \lvert\mathrm{d}\phi\rvert^2$ in 2D; shrinking solutions have constant action	Chapter 11, Chapter 12

Key idea: The one sentence to remember from Chapter 8

The configuration space of a gauge theory is a quotient of a contractible space by a topologically rich group; its integers are Chern numbers and winding numbers, its minima are the solutions of first-order (self-duality or Bogomolny) equations with finite-dimensional moduli spaces counted by index theorems, and its topology, through Morse theory, forces further, unstable solutions whenever compactness can be controlled.

Chapter 9

The Elliptic Geometry of Strings

Where this chapter is going

A particle moving through space-time traces out a curve, its world line. A string traces out a surface, its **world sheet**. Quantising a particle by Feynman's recipe means summing e^{-S} over all curves. Quantising a string means summing e^{-S} over all surfaces. This chapter shows that the second sum, which looks hopelessly infinite-dimensional, collapses to an ordinary *finite-dimensional* integral. The collapse is driven entirely by the tools of the earlier chapters. Elliptic operators (Chapter 2) appear as the Laplacian acting on the string's position and as a first-order operator P_1 acting on vector fields. Their zero modes are counted by index theory and Riemann–Roch (Chapters 4 and 5). Their determinants are defined with zeta functions (Chapter 4). And the finite-dimensional space that is left over is the **moduli space** of Riemann surfaces from Chapter 5, now equipped with a natural metric, the **Weil–Petersson metric**.

The logic is the logic of gauge fixing, which a physicist already knows from Yang–Mills theory. The Polyakov action has a huge gauge symmetry: diffeomorphisms of the surface and local rescalings (Weyl transformations) of its metric. Three components of a two-dimensional metric, minus two functions' worth of diffeomorphisms, minus one function's worth of Weyl rescalings, leaves zero local degrees of freedom. So the metric is "pure gauge" up to a finite number of global parameters, the moduli. The Faddeev–Popov procedure turns the integral over metrics into an integral over moduli, weighted by determinants of elliptic operators. Figure 9.1 shows how the pieces fit together. The final formula, Nash's (9.36), is the genus-p contribution to the bosonic string's vacuum amplitude in $d = 26$ dimensions, written as an integral over the moduli space $\mathcal{M}_p$.

Conventions for this chapter.

- The world sheet Σ is a compact, connected, oriented surface without boundary, of genus p (Nash uses p for the genus so that g can denote a metric). Its Euler characteristic is $\chi = 2 - 2p$.
- Local coordinates on Σ are x^a, $a = 1, 2$ (Nash also writes (σ, τ)). The world-sheet metric is g_{ab}, positive definite (Euclidean signature). We write $\sqrt{g} = \sqrt{\det g_{ab}}$ and g^{ab} for the inverse matrix.
- Space-time (the **target space**) is flat Euclidean $\mathbb{R}^d$ with coordinates ϕ^μ, $\mu = 1, \dots, d$, and metric $\delta_{\mu\nu}$. We write $A \cdot B = \delta_{\mu\nu} A^\mu B^\nu$.
- The Laplacian on functions is the *positive* operator $\Delta_g = -\frac{1}{\sqrt{g}} \partial_a \big(\sqrt{g}\, g^{ab} \partial_b \big)$. On flat $\mathbb{R}^2$ it is $-(\partial_1^2 + \partial_2^2)$.

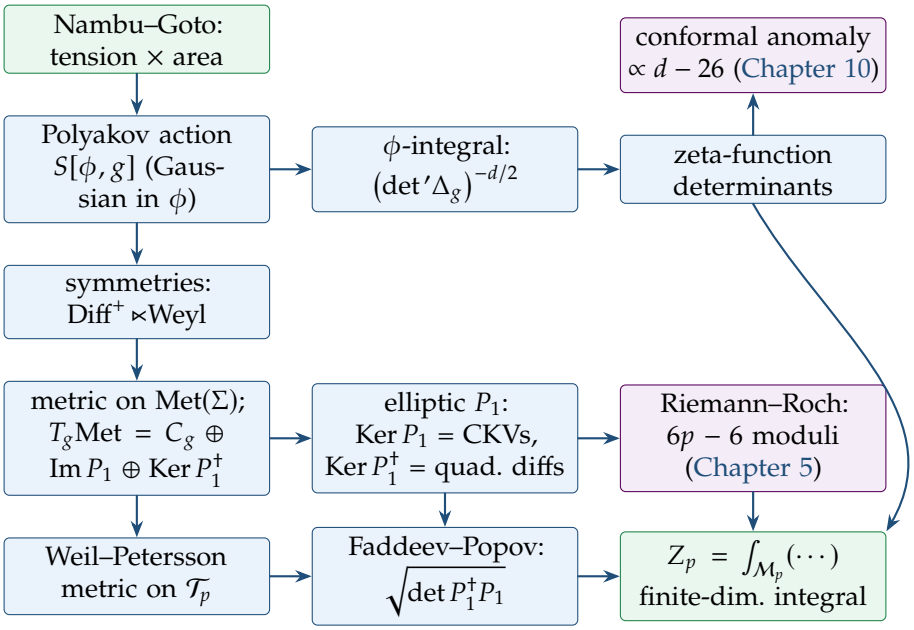


Figure 9.1: The logic of Chapter 9. Green boxes are physics, blue boxes are the machinery built in this chapter, purple boxes are results imported from other chapters. Follow the arrows from the Nambu–Goto area to the final formula for the genus-p partition function Z_p: every arrow is either a Gaussian integral, a symmetry, or a statement about an elliptic operator.

- Complex coordinates are $z = x^1 + ix^2$, $\bar{z} = x^1 - ix^2$, $\partial_z = \frac{1}{2}(\partial_1 - i\partial_2)$, $\partial_{\bar{z}} = \frac{1}{2}(\partial_1 + i\partial_2)$.
- det means a zeta-regularised determinant and det′ means that zero eigenvalues are left out (Section 9.1.10).

9.1 The bosonic string

What is the quantum amplitude for a closed string to appear out of the vacuum, propagate, and disappear again? In a point-particle theory the analogous question is answered by a path integral over closed world lines. For a string the world lines become closed surfaces, and the path integral becomes a sum over surfaces of every shape and every topology. This section sets up that sum. We start from the physics (the action of a relativistic string), rewrite the action in a form that is quadratic in the string's position (Polyakov's trick), identify its symmetries, and do the part of the path integral that is Gaussian. What is left is an integral over the metrics on the world sheet. That integral is the subject of the rest of the chapter.

9.1.1 From world lines to world sheets

Start with a relativistic particle, because every idea we need for the string already appears there in miniature. A particle in $\mathbb{R}^d$ traces out a curve $\tau \mapsto x^\mu(\tau)$. We work in Euclidean signature throughout (the Wick-rotated theory), so the natural action is the mass times the length of the curve:

$$S_{\text{particle}}[x] \;=\; m \int \mathrm{d}\tau\, \sqrt{\dot{x}\cdot\dot{x}}, \qquad \dot{x}^\mu = \frac{\mathrm{d}x^\mu}{\mathrm{d}\tau}. \tag{9.1}$$

Two properties are worth noticing. First, S does not care how we label the points of the curve: a new parameter $\tau' = f(\tau)$ with $f' > 0$ gives the same number. Second, its critical points are the curves of extremal length, the geodesics. In flat space these are straight lines.

Worked example 9.1: Critical points of the length are straight lines

Vary $x \to x + \delta x$ with δx vanishing at the end points. Write $\ell = \sqrt{\dot{x}\cdot\dot{x}}$. Then

$$\begin{aligned}\delta S_{\text{particle}} &\overset{(1)}{=} m\int \mathrm{d}\tau\,\frac{\dot{x}\cdot\delta\dot{x}}{\ell}\\ &\overset{(2)}{=} -m\int \mathrm{d}\tau\,\frac{\mathrm{d}}{\mathrm{d}\tau}\left(\frac{\dot{x}_\mu}{\ell}\right)\delta x^\mu\\ &\overset{(3)}{=} -m\int \mathrm{d}\tau\,\ddot{x}_\mu\,\delta x^\mu \qquad \text{(arc-length parameter)}.\end{aligned}$$

(1) $\delta\sqrt{u} = \delta u/(2\sqrt{u})$ with $u = \dot{x}\cdot\dot{x}$ and $\delta u = 2\dot{x}\cdot\delta\dot{x}$.
(2) Integrate by parts; the boundary term vanishes because $\delta x = 0$ at the end points.
(3) Use the freedom to reparametrise and choose τ to be arc length, so $\ell = 1$ everywhere and $\mathrm{d}/\mathrm{d}\tau$ acts only on $\dot{x}$.

So $\delta S = 0$ for all δx exactly when $\ddot{x} = 0$: a straight line traversed at constant speed.

The square root in (9.1) is awkward. It makes the action non-polynomial, and a path integral of e^{-S} is then not Gaussian. There is a standard trick for removing it, and it is worth seeing here because Polyakov's action is the same trick one dimension up. Introduce an extra positive function $e(\tau)$, called the **einbein**, and the action

$$S_e[x,e] \;=\; \frac{1}{2}\int \mathrm{d}\tau\left(\frac{\dot{x}\cdot\dot{x}}{e} + e\,m^2\right). \tag{9.2}$$

Here e plays the role of a metric on the world line: the invariant line element is $e\,\mathrm{d}\tau$.

Worked example 9.2: Eliminating the einbein gives back the length

The equation of motion of e is algebraic:

$$0 = \frac{\partial}{\partial e}\left(\frac{\dot{x}^2}{e} + e\,m^2\right) \overset{(1)}{=} -\frac{\dot{x}^2}{e^2} + m^2 \quad\Longrightarrow\quad e = \frac{\sqrt{\dot{x}^2}}{m}.$$

(1) Differentiate term by term; $\dot{x}^2$ does not depend on e.

We take the positive root because $e > 0$. Substitute back:

$$\begin{aligned}S_e\left[x,\,\sqrt{\dot{x}^2}/m\right] &\overset{(1)}{=} \frac{1}{2}\int \mathrm{d}\tau\left(m\sqrt{\dot{x}^2} + m\sqrt{\dot{x}^2}\right)\\ &= m\int \mathrm{d}\tau\,\sqrt{\dot{x}^2} \;=\; S_{\text{particle}}[x].\end{aligned}$$

(1) $\dot{x}^2/e = \dot{x}^2\cdot m/\sqrt{\dot{x}^2} = m\sqrt{\dot{x}^2}$, and $e\,m^2 = m\sqrt{\dot{x}^2}$.

At its critical point in e, the quadratic action equals the square-root action. Nash's sentence "at a critical point, the two expressions for the action coincide" (p. 260) is the string version of exactly this computation.

Physical picture: Why the einbein helps

In S_e the particle's position appears quadratically: for fixed e the integral over $x(\tau)$ is Gaussian, and Gaussian integrals are the one kind of infinite-dimensional integral we can do exactly. The price is an extra field e with a gauge symmetry (reparametrisations act on

e as $e\,d\tau = e'\,d\tau'$). The string will pay the same price, with a two-dimensional metric in place of e.

A string is a one-dimensional object. As time goes on it sweeps out a two-dimensional surface Σ in space-time, the world sheet (Figure 9.2). We restrict to **closed** strings (loops) and **oriented** strings (loops with a direction marked on them). A closed string sweeps out a tube; if the tube closes up on itself we get a closed surface. An orientation on the loop, together with the direction of time, orients the tube. So the world sheets we sum over are *closed orientable surfaces*: a sphere, a torus, a surface with p handles.

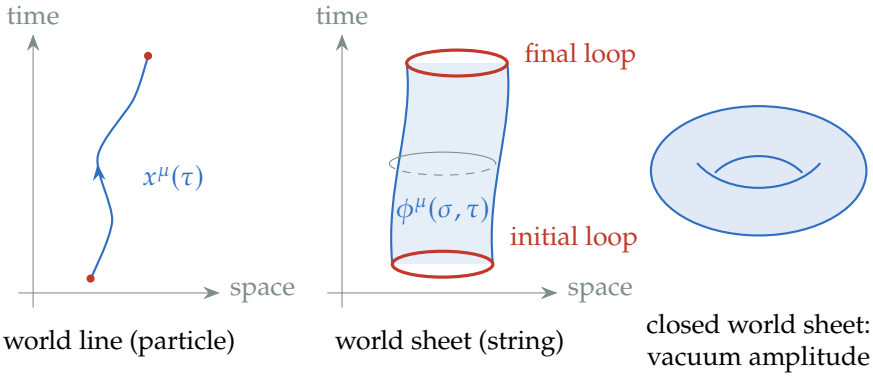


Figure 9.2: Left: a particle's history is a curve. Middle: a closed string's history is a tube $\phi^\mu(\sigma,\tau)$, with σ running around the loop and τ along the tube; its edges (red) are the initial and final loops. Right: if the final loop is glued back onto the initial one, the world sheet becomes a closed surface with no edge. Closed surfaces are what contribute to the vacuum amplitude studied in this chapter.

9.1.2 The Nambu–Goto action: tension times area

Label the points of the world sheet by two coordinates (σ,τ): σ runs around the string, τ runs along its history. The position of the string in space-time is a map

$$\phi : \Sigma \to \mathbb{R}^d, \qquad (\sigma,\tau) \mapsto \phi^\mu(\sigma,\tau), \quad \mu = 1,\dots,d.$$

The natural generalisation of "mass times length" is "tension times area". To compute the area of the image surface we need its **induced metric** (also called the pull-back metric). The two tangent vectors $\partial_\sigma\phi$ and $\partial_\tau\phi$ span a little parallelogram in $\mathbb{R}^d$ (Figure 9.3). Its squared lengths and angles are recorded by the 2×2 matrix

$$h_{ab} \;=\; \partial_a\phi\cdot\partial_b\phi \;=\; \begin{pmatrix} \partial_\sigma\phi\cdot\partial_\sigma\phi & \partial_\sigma\phi\cdot\partial_\tau\phi \\ \partial_\tau\phi\cdot\partial_\sigma\phi & \partial_\tau\phi\cdot\partial_\tau\phi \end{pmatrix}, \qquad a,b\in\{\sigma,\tau\}. \tag{9.3}$$

This is the metric that a creature living on the world sheet would measure with rulers borrowed from space-time.

Working definition 9.1 (Nambu–Goto action)

The **Nambu–Goto action** of an embedded surface $\phi : \Sigma \to \mathbb{R}^d$ is

$$S_{\mathrm{NG}}[\phi] \;=\; T\int_\Sigma d\sigma\,d\tau\,\sqrt{\det h} \;=\; T\int_\Sigma d\sigma\,d\tau\,\sqrt{(\partial_\sigma\phi\cdot\partial_\sigma\phi)(\partial_\tau\phi\cdot\partial_\tau\phi)-(\partial_\sigma\phi\cdot\partial_\tau\phi)^2}\,. \tag{9.4}$$

The constant T is the **string tension** (energy per unit length). To compute it: find the induced metric (9.3), take the square root of its determinant [the area of the little parallelogram spanned by $\partial_\sigma\phi$ and $\partial_\tau\phi$], integrate. The result is T times the area of the image.

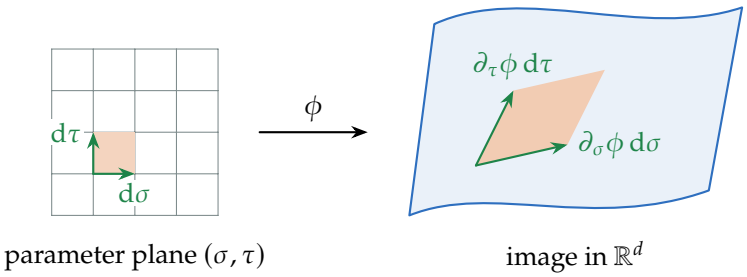


Figure 9.3: The map ϕ sends a small coordinate rectangle to a small parallelogram spanned by $\partial_\sigma\phi\,\mathrm{d}\sigma$ and $\partial_\tau\phi\,\mathrm{d}\tau$ (orange). Its area is $\sqrt{\det h}\,\mathrm{d}\sigma\,\mathrm{d}\tau$, where h is the Gram matrix of the two tangent vectors. Adding up these areas gives the Nambu–Goto action.

Why is $\sqrt{\det h}$ the area of the parallelogram? For two vectors A, B in $\mathbb{R}^3$ we know that the area is $|A \times B|$, and

$$\begin{aligned} |A\times B|^2 &\overset{(1)}{=} |A|^2\,|B|^2 \sin^2\theta \\ &\overset{(2)}{=} |A|^2\,|B|^2 - |A|^2\,|B|^2\cos^2\theta \\ &\overset{(3)}{=} (A\cdot A)(B\cdot B) - (A\cdot B)^2 \;=\; \det\begin{pmatrix} A\cdot A & A\cdot B \\ B\cdot A & B\cdot B\end{pmatrix}. \end{aligned}$$

(1) The magnitude of a cross product is the product of the lengths times the sine of the angle θ between the vectors.
(2) $\sin^2\theta = 1 - \cos^2\theta$.
(3) $A\cdot B = |A|\,|B|\cos\theta$.

The final expression, the **Gram determinant**, makes sense in any dimension d, which is why Nash writes the action (9.4) with dot products rather than cross products. With $A = \partial_\sigma\phi$ and $B = \partial_\tau\phi$ it is exactly $\det h$.

Worked example 9.3: A straight cylinder

Let a circular string of radius R move rigidly along the ϕ^3 axis in $\mathbb{R}^3$ for a Euclidean time $\tau \in [0, \mathcal{T}]$:

$$\phi(\sigma,\tau) = (R\cos\sigma,\; R\sin\sigma,\; \tau), \qquad \sigma \in [0, 2\pi).$$

Then $\partial_\sigma\phi = (-R\sin\sigma, R\cos\sigma, 0)$ and $\partial_\tau\phi = (0,0,1)$, so

$$\begin{aligned} S_{\mathrm{NG}} &\overset{(1)}{=} T\int_0^{\mathcal{T}}\mathrm{d}\tau\int_0^{2\pi}\mathrm{d}\sigma\,\sqrt{R^2\cdot 1 - 0^2} \\ &\overset{(2)}{=} T\cdot 2\pi R\cdot\mathcal{T}. \end{aligned}$$

(1) $\partial_\sigma\phi\cdot\partial_\sigma\phi = R^2(\sin^2\sigma + \cos^2\sigma) = R^2$, $\partial_\tau\phi\cdot\partial_\tau\phi = 1$, and the two tangent vectors are orthogonal.
(2) The integrand is the constant R.

This is tension × circumference × time, which is energy × time: the Euclidean action of a static object whose energy is T times its length. A physicist reads off from this that a string

loop wants to shrink: the action grows linearly with its size.

The critical points of the Nambu–Goto action are the surfaces whose area does not change to first order under small deformations: **minimal surfaces** (more precisely, extremal surfaces), such as soap films. They are the two-dimensional analogue of geodesics. So far the story parallels the particle exactly. And, as for the particle, the square root makes the quantum theory difficult. Polyakov's reformulation removes it.

9.1.3 The Polyakov action

Give the world sheet its own metric $g_{ab}(x)$, a positive-definite symmetric 2×2 matrix at each point, which is *independent* of the embedding. It is the two-dimensional analogue of the einbein. Rename the world-sheet coordinates x^a, $a = 1, 2$ (so $x^1 = \sigma$, $x^2 = \tau$).

Working definition 9.2 (Polyakov action)

The **Polyakov action** of a pair (embedding ϕ, world-sheet metric g) is

$$S[\phi, g] \;=\; \frac{T}{2}\int_\Sigma \mathrm{d}^2x\,\sqrt{g}\,g^{ab}\,\partial_a\phi^\mu\,\partial_b\phi_\mu, \tag{9.5}$$

where $\sqrt{g} = \sqrt{\det g_{ab}}$, g^{ab} is the inverse matrix of g_{ab}, and the index μ is summed with the flat metric of $\mathbb{R}^d$. For fixed g this is the action of d free massless scalar fields ϕ^μ [each ϕ^μ is a real function on Σ, and the action contains only its gradient squared, with no mass or interaction term] living on the curved surface (Σ, g). *How to use it:* for fixed g it is quadratic in ϕ, so every ϕ-integral is Gaussian [done like $\int \mathrm{d}x\, e^{-ax^2/2} = \sqrt{2\pi/a}$, with the number a replaced by an operator].

Watch out: The sign in front of Nash's (9.2)

Nash writes $S = -\frac{T}{2}\int \mathrm{d}^2x\sqrt{g}\,g^{ab}\partial_a\phi^\mu\partial_b\phi_\mu$ with a minus sign. That sign is a relic of Lorentzian world sheets, where g_{ab} has signature $(-,+)$ and $\sqrt{g}$ means $\sqrt{-\det g}$. In the Euclidean theory used everywhere else in the chapter (positive g_{ab}, weight e^{-S}), the action must be positive so that e^{-S} damps large fluctuations. Nash's own (9.3) writes $\exp[-\frac{T}{2}\int\cdots]$, which is the positive-action convention. We use (9.5) with a plus sign throughout.

Now the claim: *at its critical point in g, the Polyakov action equals the Nambu–Goto action.* To see this we need two formulae for how $\sqrt{g}$ and g^{ab} respond to a small change δg_{ab} of the metric.

Worked example 9.4: Varying the determinant and the inverse

For any invertible symmetric matrix g,

$$\begin{aligned}\delta\sqrt{\det g} &\overset{(1)}{=} \tfrac{1}{2}(\det g)^{-1/2}\,\delta\det g\\ &\overset{(2)}{=} \tfrac{1}{2}(\det g)^{-1/2}\,\det g\;\operatorname{tr}(g^{-1}\delta g)\\ &= \tfrac{1}{2}\sqrt{g}\;g^{ab}\,\delta g_{ab},\end{aligned}$$

(1) Chain rule for the square root.

(2) Jacobi's formula $\delta\det g = \det g\,\operatorname{tr}(g^{-1}\delta g)$, which follows from $\det(g+\delta g) = \det g\,\det(1+g^{-1}\delta g)$ and $\det(1+\epsilon) = 1+\operatorname{tr}\epsilon+O(\epsilon^2)$.

and, differentiating $g^{ac}g_{cb} = \delta^a{}_b$,

$$0 = \delta(g^{ac}g_{cb}) \overset{(1)}{=} \delta g^{ac}\, g_{cb} + g^{ac}\,\delta g_{cb} \quad\Longrightarrow\quad \delta g^{ab} = -g^{ac}\,g^{bd}\,\delta g_{cd}.$$

(1) Product rule; then multiply on the right by g^{bd} to isolate δg^{ad} and rename indices.

A quick check on a diagonal metric $g = \mathrm{diag}(\lambda_1, \lambda_2)$: $\sqrt{g} = \sqrt{\lambda_1\lambda_2}$, so $\delta\sqrt{g} = \frac{1}{2}\sqrt{\lambda_1\lambda_2}\,(\delta\lambda_1/\lambda_1 + \delta\lambda_2/\lambda_2)$, which is $\frac{1}{2}\sqrt{g}\,g^{ab}\delta g_{ab}$ as claimed.

Write $h_{ab} = \partial_a\phi\cdot\partial_b\phi$ for the induced metric (9.3), so the Polyakov Lagrangian is $\frac{T}{2}\sqrt{g}\,g^{ab}h_{ab}$. Vary g with ϕ fixed:

$$\begin{aligned}
\delta S &\overset{(1)}{=} \frac{T}{2}\int \mathrm{d}^2x\,\Big(\delta\sqrt{g}\;g^{ab}h_{ab} + \sqrt{g}\;\delta g^{ab}\;h_{ab}\Big)\\
&\overset{(2)}{=} \frac{T}{2}\int \mathrm{d}^2x\,\sqrt{g}\,\Big(\tfrac{1}{2}g^{cd}\,g^{ab}h_{ab} - g^{ca}g^{db}h_{ab}\Big)\delta g_{cd}\\
&\overset{(3)}{=} -\frac{T}{2}\int \mathrm{d}^2x\,\sqrt{g}\;g^{ca}g^{db}\,\Big(h_{ab} - \tfrac{1}{2}g_{ab}\,g^{ef}h_{ef}\Big)\,\delta g_{cd}.
\end{aligned} \tag{9.6}$$

(1) Product rule; h_{ab} does not depend on g.
(2) Insert the two formulae of Example 9.4.
(3) Factor out $g^{ca}g^{db}$, using $g^{cd} = g^{ca}g^{db}g_{ab}$.

The quantity in brackets is (up to normalisation) the world-sheet **energy–momentum tensor**:

$$T_{ab} \;\propto\; h_{ab} - \tfrac{1}{2}\,g_{ab}\,g^{cd}h_{cd} \;=\; \partial_a\phi\cdot\partial_b\phi - \tfrac{1}{2}\,g_{ab}\,g^{cd}\,\partial_c\phi\cdot\partial_d\phi. \tag{9.7}$$

The metric's equation of motion is $T_{ab} = 0$. Two things follow at once.

- T_{ab} is *traceless*: $g^{ab}T_{ab} \propto g^{ab}h_{ab} - \frac{1}{2}\cdot 2\cdot g^{cd}h_{cd} = 0$, using $g^{ab}g_{ab} = 2$ in two dimensions. We will see in a moment that tracelessness is the Noether identity of Weyl invariance.
- $T_{ab} = 0$ says $h_{ab} = \frac{1}{2}(g^{cd}h_{cd})\,g_{ab}$: the induced metric is a *positive multiple* of g. Equivalently $g_{ab} = \Omega(x)\,h_{ab}$ for some positive function Ω.

Worked example 9.5: At the critical metric, Polyakov = Nambu–Goto

Put $g_{ab} = \Omega\,h_{ab}$ into (9.5):

$$\begin{aligned}
S[\phi,\Omega h] &\overset{(1)}{=} \frac{T}{2}\int \mathrm{d}^2x\,\big(\Omega\sqrt{h}\big)\big(\Omega^{-1}h^{ab}\big)h_{ab}\\
&\overset{(2)}{=} \frac{T}{2}\int \mathrm{d}^2x\,\sqrt{h}\cdot 2\\
&\overset{(3)}{=} T\int \mathrm{d}^2x\,\sqrt{\det h} \;=\; S_{\mathrm{NG}}[\phi].
\end{aligned}$$

(1) For a 2×2 matrix, $\det(\Omega h) = \Omega^2\det h$, so $\sqrt{\det(\Omega h)} = \Omega\sqrt{h}$; and $(\Omega h)^{-1} = \Omega^{-1}h^{-1}$.
(2) The factors of Ω cancel, and $h^{ab}h_{ab} = \mathrm{tr}(h^{-1}h) = 2$.
(3) This is (9.4).

Notice that the undetermined function Ω dropped out completely. The equation of motion fixes g only up to a local rescaling. That is the first sign of the Weyl symmetry discussed next.

Worked example 9.6: A non-critical metric on the cylinder

Take the cylinder of Example 9.3, where $h = \operatorname{diag}(R^2, 1)$, and choose instead the flat world-sheet metric $g = \operatorname{diag}(1, 1)$. Then

$$S[\phi, g] \overset{(1)}{=} \frac{T}{2} \int_0^{\mathcal{T}} d\tau \int_0^{2\pi} d\sigma \, (h_{\sigma\sigma} + h_{\tau\tau})$$
$$\overset{(2)}{=} \frac{T}{2} \cdot 2\pi\mathcal{T} \, (R^2 + 1) \overset{(3)}{\geq} T \cdot 2\pi\mathcal{T} R = S_{\mathrm{NG}}.$$

(1) With $g = \mathbf{1}$: $\sqrt{g} = 1$ and $g^{ab} h_{ab} = h_{11} + h_{22}$.
(2) Both entries of h are constant.
(3) $\frac{1}{2}(R^2 + 1) \geq R$ is the inequality $(R - 1)^2 \geq 0$ between arithmetic and geometric means.

Equality holds exactly when $R = 1$, that is, exactly when $h \propto g$. In general, for a fixed embedding, $\frac{1}{2}\sqrt{g}\, g^{ab} h_{ab} \geq \sqrt{h}$ with equality iff $g \propto h$ (apply the same AM–GM inequality to the eigenvalues of $g^{-1}h$). So the Nambu–Goto area is the *minimum* of the Polyakov action over world-sheet metrics.

Dictionary: World-sheet and target-space language

String physics	Geometry
world sheet Σ	closed oriented surface of genus p
string position $\phi^\mu(x)$	a map $\Sigma \to \mathbb{R}^d$ (d scalar fields)
world-sheet metric g_{ab}	a Riemannian metric on Σ
induced metric h_{ab}	pull-back $\phi^* \delta$ of the flat metric
string tension T, Regge slope α'	$T = 1/(2\pi\alpha')$
Polyakov action	Dirichlet energy of ϕ w.r.t. g
classical string	harmonic, conformal map = minimal surface

9.1.4 The symmetries of the Polyakov action

The whole chapter is about what the symmetries of $S[\phi, g]$ do to the path integral, so we verify each of them by hand.

Weyl rescalings. Multiply the metric by a positive function,

$$g_{ab}(x) \longmapsto f(x)\, g_{ab}(x) = e^{2\omega(x)} g_{ab}(x), \qquad f > 0, \tag{9.8}$$

and leave ϕ alone. (Writing $f = e^{2\omega}$ is convenient for infinitesimal work: $\delta g_{ab} = 2\,\delta\omega\, g_{ab}$.) Such a change is called a **Weyl transformation** or **conformal rescaling**. It changes lengths but not angles: two vectors that are orthogonal for g are orthogonal for $f g$, and the ratio of their lengths is unchanged.

Worked example 9.7: Weyl invariance works only in two dimensions

Do the computation on an n-dimensional world volume, to see where $n = 2$ enters. Under $g \mapsto f g$,

$$\sqrt{\det(f g)}\, (f g)^{ab} \overset{(1)}{=} \left(f^{n/2}\sqrt{g}\right)\left(f^{-1} g^{ab}\right)$$
$$\overset{(2)}{=} f^{(n-2)/2} \sqrt{g}\, g^{ab}$$

$$\overset{(3)}{=} \sqrt{g}\, g^{ab} \qquad \text{if } n = 2.$$

(1) $\det(f\,g) = f^n \det g$ for an $n \times n$ matrix, and the inverse of $f\,g$ is $f^{-1}g^{-1}$.
(2) Collect the powers of f.
(3) $f^0 = 1$.

So for a world sheet ($n = 2$) the density $\sqrt{g}\, g^{ab}$ is Weyl invariant and $S[\phi, f\,g] = S[\phi, g]$. For a membrane ($n = 3$) the analogous action picks up $\sqrt{f}$, and there is no Weyl symmetry. This is one reason strings are special.

The infinitesimal version gives the tracelessness of T_{ab} found above. With $\delta g_{cd} = 2\,\delta\omega\, g_{cd}$ in (9.6),

$$\delta_{\text{Weyl}} S \overset{(1)}{=} -\frac{T}{2} \int \mathrm{d}^2x\, \sqrt{g}\, g^{ca} g^{db}\, T_{ab}\, 2\,\delta\omega\, g_{cd}$$
$$\overset{(2)}{=} -T \int \mathrm{d}^2x\, \sqrt{g}\, g^{ab} T_{ab}\, \delta\omega \overset{(3)}{=} 0,$$

(1) Insert the Weyl variation; here T_{ab} stands for the bracket in (9.6).
(2) $g^{ca} g^{db} g_{cd} = g^{ab}$.
(3) T_{ab} is traceless, identically in ϕ and g.

Classically, then, "Weyl invariance" and "$g^{ab}T_{ab} = 0$" are the same statement. The **conformal anomaly** is the discovery that quantum mechanically $\langle g^{ab}T_{ab}\rangle \neq 0$ unless $d = 26$ (Section 9.2.11 and Chapter 10).

Diffeomorphisms. Let $\alpha : \Sigma \to \Sigma$ be a diffeomorphism (a smooth invertible change of points, with smooth inverse). It acts on both fields by **pull-back**:

$$(\alpha^*\phi)(x) = \phi(\alpha(x)), \qquad (\alpha^* g)_{ab}(x) = \frac{\partial \alpha^c}{\partial x^a} \frac{\partial \alpha^d}{\partial x^b}\, g_{cd}(\alpha(x)). \tag{9.9}$$

The first formula says "the string at the new point x is where the string used to be at $\alpha(x)$". The second is the tensor transformation law: it is what makes the line element $g_{cd}\,\mathrm{d}y^c \mathrm{d}y^d$ agree with $(\alpha^* g)_{ab}\mathrm{d}x^a \mathrm{d}x^b$ when $y = \alpha(x)$. Nash abbreviates both to $(\phi(\alpha(x)), g(\alpha(x)))$, with the Jacobian factors in the metric understood.

Worked example 9.8: Diffeomorphism invariance is a change of variables

Write $J^c{}_a = \partial\alpha^c/\partial x^a$ for the Jacobian matrix and $y = \alpha(x)$. Evaluate the three ingredients of the Lagrangian for the transformed fields:

$$\sqrt{\det(\alpha^* g)}(x) \overset{(1)}{=} |\det J|\ \sqrt{g}\,(y),$$
$$(\alpha^* g)^{ab}(x) \overset{(2)}{=} (J^{-1})^a{}_c\, (J^{-1})^b{}_d\, g^{cd}(y),$$
$$\partial_a(\phi \circ \alpha)(x) \overset{(3)}{=} J^c{}_a\, (\partial_c \phi)(y).$$

(1) In matrix form $\alpha^* g = J^{\mathsf{T}}(g \circ \alpha)J$, and $\det(J^{\mathsf{T}} g J) = (\det J)^2 \det g$.
(2) Invert the matrix product: $(J^{\mathsf{T}} g J)^{-1} = J^{-1} g^{-1} J^{-\mathsf{T}}$.
(3) The chain rule.

Now multiply them together and integrate:

$$S[\alpha^*\phi, \alpha^* g] \overset{(1)}{=} \frac{T}{2}\int \mathrm{d}^2x\ |\det J|\ \sqrt{g}(y)\, g^{cd}(y)\, \partial_c\phi(y)\cdot\partial_d\phi(y)$$
$$\overset{(2)}{=} \frac{T}{2}\int \mathrm{d}^2y\ \sqrt{g}\, g^{cd}\, \partial_c\phi\cdot\partial_d\phi \ = \ S[\phi, g].$$

(1) The J^{-1} from the inverse metric cancel the J from the two derivatives: $(J^{-1})^a{}_c J^{c'}{}_a = \delta^{c'}_c$.
(2) Change the integration variable from x to $y = \alpha(x)$; $\mathrm{d}^2y = |\det J|\ \mathrm{d}^2x$, and α is a bijection of Σ.

The action is a coordinate-free object, so relabelling the points of Σ cannot change it.

Watch out: Why only orientation-preserving diffeomorphisms?

The computation above works for *every* diffeomorphism, since only $|\det J|$ appeared. Nash nevertheless restricts to the group $\mathrm{Diff}^+(\Sigma)$ of **orientation-preserving** diffeomorphisms ($\det J > 0$). The reason is not the action but the complex structure: a metric on an oriented surface determines a complex structure (Section 9.1.6), and an orientation-reversing map turns it into its complex conjugate. Since the world sheets are oriented, we identify only configurations related by $\mathrm{Diff}^+(\Sigma)$ (compare [Nash p. 150]).

Space-time symmetries. Finally, S only involves derivatives of ϕ, so it is unchanged by **translations** $\phi^\mu \mapsto \phi^\mu + a^\mu$, $a \in \mathbb{R}^d$. It is also unchanged by rotations $\phi \mapsto R\phi$, $R \in \mathrm{SO}(d)$, because it uses the dot product of $\mathbb{R}^d$. These are global symmetries acting on the target, not gauge symmetries of the world sheet, and their only role in this chapter is the zero mode of Section 9.1.9.

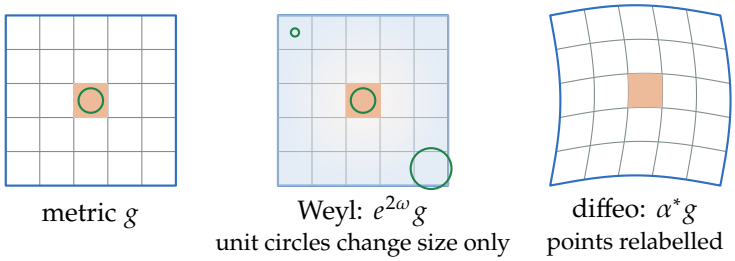


Figure 9.4: The two gauge symmetries of the Polyakov action. Middle: a Weyl rescaling multiplies all lengths near a point by the same factor e^ω, so small circles stay circles (green) and only their size varies from place to place. Right: a diffeomorphism slides the points of the surface around and drags the coordinate grid with it; it changes the *description* of the metric but not the geometry. Neither operation changes $S[\phi, g]$.

Physical picture: Two-dimensional gravity has nothing to propagate

Count functions. A symmetric 2×2 metric has three independent components g_{11}, g_{12}, g_{22} at each point. Diffeomorphisms of a surface are generated by vector fields, which have two components. Weyl rescalings are generated by one function ω. Three minus two minus one is zero: locally, *every* metric can be brought to the flat metric δ_{ab} by a diffeomorphism and a Weyl rescaling. (This is the existence of isothermal coordinates, Section 9.3.6.) The metric carries no local information at all, only a finite number of global parameters that cannot be removed. These are the moduli. The whole chapter is the careful version of this counting.

9.1.5 The functional integral and the vacuum amplitude

To quantise the string we follow Feynman: sum e^{-S} over all configurations of all the variables that appear in the action. For the Polyakov action there are two such variables, the embedding ϕ and the metric g. The basic object is therefore

$$Z_p \;=\; \int \mathcal{D}g\,\mathcal{D}\phi\;\exp\Big[-\frac{T}{2}\int_\Sigma \mathrm{d}^2x\,\sqrt{g}\,g^{ab}\,\partial_a\phi^\mu\,\partial_b\phi_\mu\Big], \tag{9.10}$$

where Σ is a closed oriented surface of genus p, $\mathcal{D}g$ is a (so far formal) integration over all metrics on Σ, and $\mathcal{D}\phi$ is a (so far formal) integration over all maps $\phi:\Sigma\to\mathbb{R}^d$. This is Nash's (9.3). We abbreviate ϕ^μ to ϕ and g_{ab} to g.

Watch out: Embeddings or maps?

Nash says ϕ "should be thought of as specifying an embedding of Σ in space-time". In the path integral we actually integrate over *all* smooth maps $\phi:\Sigma\to\mathbb{R}^d$, including ones that fold the surface over itself or crush it to a point (the constant map $\phi=$ const is the most important configuration of all, see Section 9.1.9). The word "embedding" describes the classical picture, not the domain of integration.

What does (9.10) compute? Nash calls it "a contribution to the partition function" and explains it as a vacuum-to-vacuum amplitude and as a trace. Both readings are familiar from quantum mechanics, so let us recall them there first.

Working definition 9.3 (Partition function as a sum over closed histories)

In quantum mechanics with Hamiltonian H, the **partition function**

$$Z(\beta)=\operatorname{Tr}e^{-\beta H}=\sum_n e^{-\beta E_n}$$

[E_n are the energy levels, the eigenvalues of H] is computed by a path integral over *periodic* paths, $x(0)=x(\beta)$. *How to see it:* the kernel $\langle y|e^{-\beta H}|x\rangle$ [the amplitude to go from x to y in Euclidean time β] is a path integral over paths from x to y; taking the trace sets $y=x$ and integrates over x, so the initial and final points are glued together and the world line closes into a circle.

Worked example 9.9: The free particle: a trace, and a volume factor

Take $H=p^2/2m$ on a line of length L (with $\hbar=1$, and L so large that we can ignore what happens at its ends). The Euclidean propagator is the heat kernel

$$\langle y|e^{-\beta H}|x\rangle=\sqrt{\frac{m}{2\pi\beta}}\;\exp\Big[-\frac{m(x-y)^2}{2\beta}\Big].$$

Then

$$\begin{aligned} Z(\beta) &\overset{(1)}{=} \int_0^L \mathrm{d}x\,\langle x|e^{-\beta H}|x\rangle \\ &\overset{(2)}{=} \int_0^L \mathrm{d}x\,\sqrt{\frac{m}{2\pi\beta}} \\ &= L\sqrt{\frac{m}{2\pi\beta}}. \end{aligned}$$

(1) The trace of an operator is the integral of its diagonal kernel.
(2) At $y = x$ the Gaussian factor is 1.

The factor L is the *volume of space*. It appears because the closed path can sit anywhere: translating a closed world line gives another closed world line with the same action. Exactly the same thing will happen for the string, where the volume of $\mathbb{R}^d$ appears because a closed world sheet can be translated anywhere in space-time. Nash removes it by a "compactness cutoff", which is our L.

Now the string. A state of a closed string at one instant is a loop in space. The amplitude to go from one loop configuration to another is a sum over tubes connecting them (Figure 9.2). The *trace* over string states identifies the final loop with the initial loop, gluing the two edge circles of the tube together. The result is a closed surface, with no edge at all. This is why the world sheets in (9.10) are closed. The simplest surface you get by gluing the two ends of a tube is a torus (Figure 9.5). Nash's other description, "a quantum object is emitted by the vacuum only to disappear by combining with itself", is the same picture read in time: a string pops out of the vacuum, propagates around, and annihilates with itself.

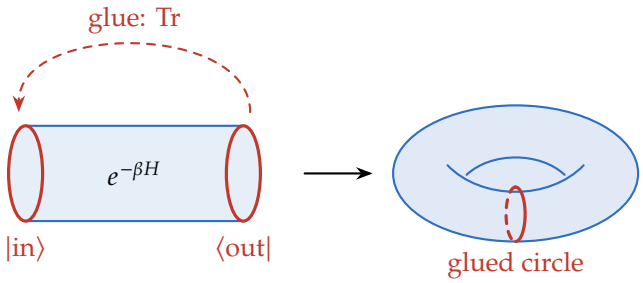


Figure 9.5: Why the world sheets are closed. A tube is the amplitude to evolve a string state for Euclidean time β. Taking the trace glues the outgoing circle onto the incoming one. The result is a torus, the genus-one term of the vacuum amplitude. The red circle on the torus is where the gluing happened; it is no longer a boundary.

The sum over topologies. A given closed surface has a definite genus p, but the string does not care about topology: every closed surface is a legitimate history. So the complete vacuum amplitude also sums over all topologies. The classification of surfaces says that every closed, connected, orientable surface is diffeomorphic to exactly one of the standard surfaces Σ_p, a sphere with p handles (Figure 9.6). Hence

$$Z \;=\; \sum_{\text{genera}} \int \mathcal{D}g\, \mathcal{D}\phi\; e^{-S[\phi,g]} \;=\; \sum_{p=0}^{\infty} Z_p, \tag{9.11}$$

Nash's (9.5) (with $T = 1$, see below).

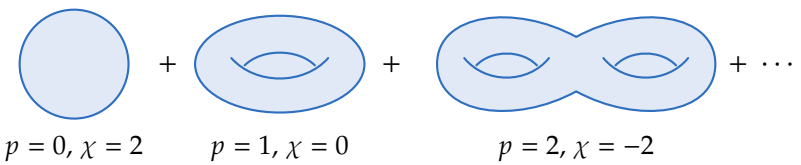


Figure 9.6: The sum over topologies in (9.11). Each term is a closed oriented surface, a sphere with p handles. In string perturbation theory each handle is a closed-string loop, so this is the string analogue of the loop expansion of Feynman diagrams, with one "diagram" per genus.

Physical picture: The genus expansion is the loop expansion

In field theory, the vacuum amplitude is a sum over Feynman diagrams, organised by the number of loops. In string theory there is exactly one world sheet per genus, and the genus plays the role of the number of loops. Each handle is a virtual closed string emitted and reabsorbed. A local vertex of Feynman diagrams is smeared out into smooth geometry, which is one reason string amplitudes are so much softer in the ultraviolet.

Worked example 9.10: Weighting the genera: the Euler characteristic is an action

Nash's sum (9.11) gives every genus the same weight. One may add to S a second term allowed by all the symmetries, the two-dimensional Einstein–Hilbert action

$$S_\lambda = \frac{\lambda}{4\pi}\int_\Sigma \mathrm{d}^2x\,\sqrt{g}\,R,$$

with $R = 2K$ the scalar curvature and K the Gauss curvature. It adds nothing to the dynamics, because it is a topological invariant: by Gauss–Bonnet (Chapter T2),

$$\frac{1}{4\pi}\int_\Sigma \mathrm{d}^2x\,\sqrt{g}\,R \overset{(1)}{=} \frac{1}{2\pi}\int_\Sigma K\,\mathrm{d}A \overset{(2)}{=} \chi(\Sigma) = 2-2p.$$

(1) $R = 2K$ in two dimensions and $\mathrm{d}A = \sqrt{g}\,\mathrm{d}^2x$.
(2) Gauss–Bonnet theorem.

Check it on a round sphere of radius r: $K = 1/r^2$ everywhere and the area is $4\pi r^2$, so $\frac{1}{2\pi}\cdot\frac{1}{r^2}\cdot 4\pi r^2 = 2 = \chi(S^2)$, independent of r. So $e^{-S_\lambda} = e^{-\lambda\chi} = (e^{\lambda})^{2p-2}$, and each extra handle costs a factor g_s^2 with $g_s = e^{\lambda}$, the **string coupling**. This term is Weyl and diffeomorphism invariant and changes none of the analysis of this chapter; it only reweights the Z_p.

9.1.6 A metric is a complex structure

A metric on an oriented surface does more than measure lengths. It singles out which coordinate systems are "holomorphic", that is, it gives the surface the structure of a **Riemann surface** (a one-dimensional complex manifold, Chapter 5). Two conformally related metrics, g and $e^{2\omega}g$, give the *same* complex structure. This is why Nash can pass freely between "metrics up to Weyl rescaling" and "complex structures".

The recipe. A complex structure at a point is the operation "rotate a tangent vector by +90°". To rotate by 90° we need to know angles (the metric, up to scale) and which way is positive (the orientation). So a conformal class of metrics plus an orientation gives a rotation J on every tangent plane, with $J^2 = -1$. Holomorphic functions are those whose differential commutes with J. Algebraically, the recipe is:

Working definition 9.4 (Beltrami coefficient of a metric)

Write a metric in real coordinates (x, y) and in terms of $z = x + iy$:

$$\mathrm{d}s^2 = A\,\mathrm{d}x^2 + 2B\,\mathrm{d}x\,\mathrm{d}y + C\,\mathrm{d}y^2 \;=\; \lambda\,|\mathrm{d}z + \mu\,\mathrm{d}\bar z|^2\,, \qquad \lambda>0,\ |\mu|<1. \tag{9.12}$$

The complex function μ is the **Beltrami coefficient**; it depends only on the conformal class [the metric up to multiplication by a positive function, $g \sim e^{2\omega}g$], while λ carries the scale.

A coordinate w is *holomorphic* for this metric [rotating by $+90°$ in this metric is multiplying $\mathrm{d}w$ by i] iff $\mathrm{d}w \propto \mathrm{d}z + \mu\,\mathrm{d}\bar{z}$, that is iff it solves the **Beltrami equation** $\partial_{\bar{z}} w = \mu\,\partial_z w$. In such a coordinate $\mathrm{d}s^2 = \lambda'\,|\mathrm{d}w|^2$.

Worked example 9.11: Solving for λ and μ

Expand the right side of (9.12) and compare coefficients. With $\mu = \mu_1 + i\mu_2$, and using the symmetric products $\mathrm{d}z\,\mathrm{d}\bar{z} = \mathrm{d}x^2 + \mathrm{d}y^2$ and $\mathrm{d}z^2 = \mathrm{d}x^2 - \mathrm{d}y^2 + 2i\,\mathrm{d}x\,\mathrm{d}y$:

$$\begin{aligned} |\mathrm{d}z + \mu\,\mathrm{d}\bar{z}|^2 &\overset{(1)}{=} (\mathrm{d}z + \mu\,\mathrm{d}\bar{z})(\mathrm{d}\bar{z} + \bar{\mu}\,\mathrm{d}z) \\ &\overset{(2)}{=} (1 + |\mu|^2)\,\mathrm{d}z\,\mathrm{d}\bar{z} + 2\,\mathrm{Re}\big(\bar{\mu}\,\mathrm{d}z^2\big) \\ &\overset{(3)}{=} (1 + |\mu|^2 + 2\mu_1)\,\mathrm{d}x^2 + 4\mu_2\,\mathrm{d}x\,\mathrm{d}y + (1 + |\mu|^2 - 2\mu_1)\,\mathrm{d}y^2. \end{aligned}$$

(1) $|w|^2 = w\bar{w}$, and the conjugate of $\mathrm{d}z$ is $\mathrm{d}\bar{z}$.
(2) Multiply out; the cross terms $\bar{\mu}\,\mathrm{d}z^2$ and $\mu\,\mathrm{d}\bar{z}^2$ are complex conjugates.
(3) $\mathrm{Re}\big((\mu_1 - i\mu_2)(\mathrm{d}x^2 - \mathrm{d}y^2 + 2i\,\mathrm{d}x\,\mathrm{d}y)\big) = \mu_1(\mathrm{d}x^2 - \mathrm{d}y^2) + 2\mu_2\,\mathrm{d}x\,\mathrm{d}y$.

Comparing with $A\,\mathrm{d}x^2 + 2B\,\mathrm{d}x\,\mathrm{d}y + C\,\mathrm{d}y^2$ gives $A + C = 2\lambda(1 + |\mu|^2)$, $A - C = 4\lambda\mu_1$, $B = 2\lambda\mu_2$. The determinant is $AC - B^2 = \lambda^2(1 - |\mu|^2)^2$. Solving,

$$\lambda = \frac{A + C + 2\sqrt{AC - B^2}}{4}, \qquad \mu = \frac{A - C + 2iB}{A + C + 2\sqrt{AC - B^2}}. \tag{9.13}$$

Two checks. (i) The Euclidean metric $A = C = 1$, $B = 0$ gives $\lambda = 1$, $\mu = 0$: z itself is holomorphic. (ii) Multiplying A, B, C by a positive f multiplies λ by f and leaves μ unchanged: a Weyl rescaling does not change the complex structure. For instance $\mathrm{d}s^2 = 4\,\mathrm{d}x^2 + \mathrm{d}y^2$ gives $\lambda = \frac{9}{4}$, $\mu = \frac{1}{3}$, and $w = \frac{3}{2}x + \frac{i}{2}\cdot\ldots$ is easiest found directly: $\mathrm{d}s^2 = |2\mathrm{d}x + i\,\mathrm{d}y|^2$, so $w = 2x + iy$ is a holomorphic coordinate, and indeed $\partial_{\bar{z}} w = \frac{1}{2}(2 - 1) = \frac{1}{2}$, $\partial_z w = \frac{1}{2}(2 + 1) = \frac{3}{2}$, ratio $\frac{1}{3} = \mu$.

The existence of a holomorphic coordinate w near every point (a local solution of the Beltrami equation) is a genuine theorem, the existence of **isothermal coordinates**. We use it freely: *every metric on a surface is locally of the form* $e^{2\omega}\,|\mathrm{d}w|^2$. We come back to it when we discuss the conformal gauge in Section 9.3.6. Nash records the same fact in [Nash p. 149, (5.35)].

The action in complex coordinates. In an isothermal coordinate $z = x^1 + ix^2$ write

$$\mathrm{d}s^2 = e^{2\omega}\big((\mathrm{d}x^1)^2 + (\mathrm{d}x^2)^2\big) = e^{2\omega}\,\mathrm{d}z\,\mathrm{d}\bar{z}, \qquad g_{ab} = e^{2\omega}\delta_{ab}. \tag{9.14}$$

In complex components the only non-zero metric entries are $g_{z\bar{z}} = g_{\bar{z}z} = \frac{1}{2}e^{2\omega}$, with inverse $g^{z\bar{z}} = g^{\bar{z}z} = 2e^{-2\omega}$. The Polyakov Lagrangian becomes very simple:

$$\begin{aligned} \sqrt{g}\,g^{ab}\,\partial_a\phi\cdot\partial_b\phi &\overset{(1)}{=} e^{2\omega}\,e^{-2\omega}\,\delta^{ab}\,\partial_a\phi\cdot\partial_b\phi \\ &\overset{(2)}{=} (\partial_1\phi)^2 + (\partial_2\phi)^2 \\ &\overset{(3)}{=} 4\,\partial_z\phi\cdot\partial_{\bar{z}}\phi. \end{aligned} \tag{9.15}$$

(1) $\sqrt{\det(e^{2\omega}\delta)} = e^{2\omega}$ and $g^{ab} = e^{-2\omega}\delta^{ab}$ (the Weyl invariance of Example 9.7 in action).
(2) Sum over a.

(3) $\partial_z\phi\,\partial_{\bar z}\phi = \tfrac14(\partial_1 - i\partial_2)\phi\,(\partial_1 + i\partial_2)\phi = \tfrac14\big((\partial_1\phi)^2 + (\partial_2\phi)^2\big)$ for real ϕ.

The conformal factor $e^{2\omega}$ has disappeared. Only the complex structure (which coordinates are isothermal) matters to the action.

The action as $\langle\phi, \Delta_g\phi\rangle$. Integrating by parts turns the action into a quadratic form of the Laplacian, which is the form we need for the Gaussian integral. Set $T = 1$ from now on (Nash does the same; it can be restored by rescaling $\phi \to \sqrt{T}\phi$). Then, on a closed surface,

$$\begin{aligned} S[\phi, g] &\overset{(1)}{=} \frac12 \int_\Sigma \mathrm{d}^2x\,\sqrt{g}\,g^{ab}\,\partial_a\phi\cdot\partial_b\phi \\ &\overset{(2)}{=} -\frac12 \int_\Sigma \mathrm{d}^2x\,\phi\cdot\partial_a\big(\sqrt{g}\,g^{ab}\partial_b\phi\big) \\ &\overset{(3)}{=} \frac12 \int_\Sigma \mathrm{d}^2x\,\sqrt{g}\,\phi\cdot\Delta_g\phi \;=\; \frac12\,\langle\phi, \Delta_g\phi\rangle_g, \end{aligned} \tag{9.16}$$

(1) The Polyakov action (9.5) with $T = 1$.
(2) Integrate by parts in x^a. There is no boundary term: Σ is closed (Stokes' theorem with $\partial\Sigma = \varnothing$, Chapter T2).
(3) Definition of the Laplacian, $\Delta_g = -\frac{1}{\sqrt{g}}\partial_a(\sqrt{g}\,g^{ab}\partial_b)$, and of the L^2 inner product $\langle u, v\rangle_g = \int \mathrm{d}^2x\sqrt{g}\,u\,v$.

In the isothermal coordinate (9.14) the Laplacian is

$$\Delta_g = -e^{-2\omega}\big(\partial_1^2 + \partial_2^2\big) = -4\,e^{-2\omega}\,\partial_z\partial_{\bar z}. \tag{9.17}$$

The same chain shows why Δ_g is a *positive* operator: $\langle\phi, \Delta_g\phi\rangle_g = \int\sqrt{g}\,\big|\nabla\phi\big|^2 \geq 0$, with equality only for constant ϕ.

Worked example 9.12: The Laplacian of the round sphere in a complex coordinate

Stereographic projection from the north pole gives the unit sphere the metric

$$\mathrm{d}s^2 = \frac{4\,\mathrm{d}z\,\mathrm{d}\bar z}{(1+|z|^2)^2}, \qquad e^{2\omega} = \frac{4}{(1+|z|^2)^2},$$

so by (9.17), $\Delta = -(1+|z|^2)^2\partial_z\partial_{\bar z}$. The height function of the sphere is $u = \cos\theta = (1-|z|^2)/(1+|z|^2)$. Write $r^2 = z\bar z$. Then

$$\begin{aligned} \partial_{\bar z}u &\overset{(1)}{=} \frac{-z(1+r^2) - (1-r^2)z}{(1+r^2)^2} = \frac{-2z}{(1+r^2)^2}, \\ \partial_z\partial_{\bar z}u &\overset{(2)}{=} \frac{-2(1+r^2)^2 + 2z\cdot 2(1+r^2)\bar z}{(1+r^2)^4} = \frac{-2(1+r^2)+4r^2}{(1+r^2)^3} = \frac{-2(1-r^2)}{(1+r^2)^3}, \\ \Delta u &\overset{(3)}{=} (1+r^2)^2\cdot\frac{2(1-r^2)}{(1+r^2)^3} = 2\,u. \end{aligned}$$

(1) Quotient rule, with $\partial_{\bar z}(z\bar z) = z$.
(2) Quotient rule again, with $\partial_z(1+z\bar z)^2 = 2(1+r^2)\bar z$.
(3) Multiply by $-(1+r^2)^2$.

So $\cos\theta$ is an eigenfunction with eigenvalue $2 = \ell(\ell+1)$ for $\ell = 1$, exactly as the spherical harmonics say. The complex-coordinate formula for Δ works, and the sign convention makes the eigenvalues positive.

Watch out: Nash's (9.4): signs and the symbol ∂

Nash writes $S = (-1/2)\int_\Sigma \langle \partial\phi, \partial\phi\rangle_g = (1/2)\int_\Sigma \langle \phi, \Delta_g \phi\rangle$ with $\langle \partial\phi, \partial\phi\rangle_g = g^{ab}\partial_a\phi^\mu \partial_b \phi_\mu \sqrt{g}\, \mathrm{d}z\, \mathrm{d}\bar{z}$ and $\Delta_g = \partial\partial^* + \partial^*\partial$. Three remarks.

(i) The two expressions differ by an integration by parts, which gives $+\int \langle \mathrm{d}\phi, \mathrm{d}\phi\rangle = +\int \langle \phi, \mathrm{d}^*\mathrm{d}\phi\rangle$ with no sign change; the minus sign in front of the first expression should be dropped (or absorbed into the orientation convention for $\mathrm{d}z\,\mathrm{d}\bar{z} = -2i\,\mathrm{d}x \wedge \mathrm{d}y$). Our (9.16) is the correct Euclidean statement.

(ii) In $\langle \partial\phi, \partial\phi\rangle_g$ the symbol ∂ is the full gradient d (the formula has $g^{ab}\partial_a\partial_b$ with both indices summed). With $\partial = \mathrm{d}$, $\Delta_g = \mathrm{d}\mathrm{d}^* + \mathrm{d}^*\mathrm{d}$ is the Hodge Laplacian (Chapter 2).

(iii) On functions the term $\mathrm{d}\mathrm{d}^*$ contributes nothing, as Nash notes: d^* lowers form degree, and there are no (-1)-forms, so $\mathrm{d}^*\phi = 0$ for a function ϕ. Hence $\Delta_g\phi = \mathrm{d}^*\mathrm{d}\phi$, which in coordinates is our $-\frac{1}{\sqrt{g}}\partial_a(\sqrt{g}g^{ab}\partial_b\phi)$. (If instead ∂ is read as the holomorphic Dolbeault operator, then on a Riemann surface $\partial^*\partial = \frac{1}{2}\mathrm{d}^*\mathrm{d}$ on functions; the factor $\frac{1}{2}$ only rescales determinants, see Example 9.26.)

9.1.7 The configuration space and its symmetry group

From now on space-time is flat Euclidean $\mathbb{R}^d$. The two variables of the Polyakov action live in two big spaces:

- $\mathrm{Map}(\Sigma, \mathbb{R}^d)$, the space of all smooth maps $\phi : \Sigma \to \mathbb{R}^d$. It is simply a vector space: d copies of the space $C^\infty(\Sigma)$ of smooth real functions.
- $\mathrm{Met}(\Sigma)$, the space of all Riemannian metrics on Σ.

Nash calls their product the **string configuration space**

$$B = \mathrm{Map}(\Sigma, \mathbb{R}^d) \times \mathrm{Met}(\Sigma), \tag{9.18}$$

and regards the action as a function $S : B \to \mathbb{R}$, $(\phi, g) \mapsto S(\phi, g)$. The path integral (9.10) is an integral over B.

What does $\mathrm{Met}(\Sigma)$ look like? At each point x a metric is a symmetric 2×2 matrix

$$g(x) = \begin{pmatrix} a & b \\ b & c \end{pmatrix}, \qquad a > 0, \quad ac - b^2 > 0.$$

The set of such matrices, call it $\mathrm{Pos}(2)$, is an open cone in the three-dimensional space $\{(a, b, c)\}$ of symmetric matrices: if g and g' are positive definite, so are tg for $t > 0$ and $g + g'$. So $\mathrm{Met}(\Sigma)$ is "one copy of this three-dimensional cone at each point of Σ". In particular it is an open convex subset of the vector space of symmetric tensor fields, and hence its tangent space at any metric g is that whole vector space:

$$T_g\mathrm{Met}(\Sigma) = \{\text{symmetric tensor fields } \delta g_{ab}(x) \text{ on } \Sigma\}. \tag{9.19}$$

A tangent vector to Met is "a small change of metric", which is a symmetric tensor field. This is the fact behind Nash's remark (p. 263) that tangent vectors to $\mathrm{Met}(\Sigma)$ "can be thought of both as tangents to $\mathrm{Met}(\Sigma)$ and as tensor fields on Σ".

The gauge group. We have found two groups that leave S invariant (Section 9.1.4).

- The **Weyl group** $C^\infty_+(\Sigma)$: the smooth *positive* functions on Σ, with pointwise multiplication as the group law. It is abelian. It acts by $g \mapsto f g$ and does not touch ϕ. Its Lie algebra is all of $C^\infty(\Sigma)$, via $f = e^{2\omega}$.
- The group $\mathrm{Diff}^+(\Sigma)$ of orientation-preserving diffeomorphisms, with composition as the group law. It acts on both variables by pull-back (9.9).

Nash writes the combined action as the map (his (9.6))

$$A : \big(\mathrm{Diff}^+(\Sigma) \ltimes C^\infty_+(\Sigma)\big) \times B \longrightarrow B, \qquad \big((\alpha, f), (\phi, g)\big) \longmapsto \big(\alpha^*\phi,\ f \cdot \alpha^* g\big), \tag{9.20}$$

that is, $(\phi(\alpha(x)),\ f(x)\, g(\alpha(x)))$ in Nash's shorthand, where $g(\alpha(x))$ is understood to include the Jacobian factors. The symbol $\ltimes$ says that the two groups do *not* simply commute. Let us see exactly how they fail to commute, by composing two such transformations.

Worked example 9.13: The group law of the semidirect product

Write $T_{(\alpha,f)}$ for the transformation (9.20). Apply $T_{(\alpha_2,f_2)}$ first and then $T_{(\alpha_1,f_1)}$:

$$\begin{aligned} T_{(\alpha_1,f_1)}\, T_{(\alpha_2,f_2)}(\phi, g) &\overset{(1)}{=} T_{(\alpha_1,f_1)}\big(\alpha_2^*\phi,\ f_2\, \alpha_2^* g\big) \\ &\overset{(2)}{=} \big(\alpha_1^*\alpha_2^*\phi,\ f_1\, \alpha_1^*\big(f_2\, \alpha_2^* g\big)\big) \\ &\overset{(3)}{=} \big((\alpha_2 \circ \alpha_1)^*\phi,\ f_1\, (f_2 \circ \alpha_1)\, (\alpha_2 \circ \alpha_1)^* g\big) \\ &\overset{(4)}{=} T_{(\alpha_2\circ\alpha_1,\ f_1\cdot(f_2\circ\alpha_1))}(\phi, g). \end{aligned}$$

(1) Definition of $T_{(\alpha_2,f_2)}$.
(2) Definition of $T_{(\alpha_1,f_1)}$, applied to the new pair.
(3) Pull-back reverses the order of composition, $\alpha_1^*\alpha_2^* = (\alpha_2 \circ \alpha_1)^*$ (check on a function: $\alpha_1^*(\alpha_2^*\phi)(x) = \phi(\alpha_2(\alpha_1(x)))$), and the pull-back of a product is the product of pull-backs, $\alpha_1^* f_2 = f_2 \circ \alpha_1$.
(4) Read off the result as a single transformation of the same type.

So the product in the gauge group is

$$(\alpha_1, f_1) \cdot (\alpha_2, f_2) = \big(\alpha_2 \circ \alpha_1,\ f_1 \cdot (f_2 \circ \alpha_1)\big). \tag{9.21}$$

The diffeomorphism α_1 acts on the Weyl factor f_2 by moving it around the surface. That twist is what the word "semidirect" means. The identity is $(\mathrm{id}, 1)$, and the inverse of (α, f) is $(\alpha^{-1},\ 1/(f \circ \alpha^{-1}))$: indeed $(\alpha, f) \cdot (\alpha^{-1}, h) = (\mathrm{id},\ f \cdot (h \circ \alpha))$, which is $(\mathrm{id}, 1)$ exactly when $h \circ \alpha = 1/f$.

Worked example 9.14: Weyl rescalings form a normal subgroup

Conjugate a pure Weyl rescaling (id, f) by a pure diffeomorphism $(\alpha, 1)$. Using (9.21) twice,

$$\begin{aligned} (\alpha, 1)^{-1} \cdot (\mathrm{id}, f) \cdot (\alpha, 1) &\overset{(1)}{=} (\alpha^{-1}, 1) \cdot (\mathrm{id}, f) \cdot (\alpha, 1) \\ &\overset{(2)}{=} (\alpha^{-1},\ f \circ \alpha^{-1}) \cdot (\alpha, 1) \\ &\overset{(3)}{=} (\mathrm{id},\ f \circ \alpha^{-1}). \end{aligned}$$

(1) The inverse of $(\alpha, 1)$ is $(\alpha^{-1}, 1)$ by the formula above.
(2) $(\alpha^{-1}, 1) \cdot (\mathrm{id}, f) = (\mathrm{id} \circ \alpha^{-1},\ 1 \cdot (f \circ \alpha^{-1}))$.
(3) $(\alpha^{-1}, f \circ \alpha^{-1}) \cdot (\alpha, 1) = (\alpha \circ \alpha^{-1},\ (f \circ \alpha^{-1}) \cdot 1)$.

A conjugated Weyl rescaling is again a Weyl rescaling, by the moved function $f \circ \alpha^{-1}$. So $C^\infty_+(\Sigma)$ is a **normal subgroup**, and the quotient by it is $\mathrm{Diff}^+(\Sigma)$. The diffeomorphisms, on the other hand, do not form a normal subgroup. (In $\ltimes$ the normal factor sits on the side of the open triangle; conventions for the symbol vary between books, so it is safest to remember "Weyl is normal".)

Worked example 9.15: A finite-dimensional semidirect product you already know

The affine maps of the line, $x \mapsto ax + b$ with $a > 0$, form a group. Represent (a, b) by the matrix $\begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix}$ acting on $\begin{pmatrix} x \\ 1 \end{pmatrix}$. Multiplying matrices,

$$\begin{pmatrix} a_1 & b_1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a_2 & b_2 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} a_1 a_2 & a_1 b_2 + b_1 \\ 0 & 1 \end{pmatrix}.$$

The dilation a_1 acts on the translation b_2 before the translations are added: the same twist as $f_2 \mapsto f_2 \circ \alpha_1$ above. Translations $(1, b)$ are normal (conjugating a translation by a dilation gives a translation by b/a or ab), dilations are not. The Euclidean group $\mathbb{R}^n \rtimes \mathrm{SO}(n)$ and the Poincaré group are built the same way.

Worked example 9.16: A diffeomorphism changes the components but not the geometry

On the square torus $\mathbb{R}^2/\mathbb{Z}^2$ with the flat metric $g = \mathbf{1}$, take the diffeomorphism $\alpha(x, y) = (x + \varepsilon \sin 2\pi y,\ y)$, which slides each horizontal circle sideways by an amount depending on its height. Its Jacobian and the pulled-back metric are

$$J = \begin{pmatrix} 1 & 2\pi\varepsilon \cos 2\pi y \\ 0 & 1 \end{pmatrix}, \qquad \alpha^* g \overset{(1)}{=} J^{\mathsf{T}} J \overset{(2)}{=} \begin{pmatrix} 1 & 2\pi\varepsilon \cos 2\pi y \\ 2\pi\varepsilon \cos 2\pi y & 1 + 4\pi^2\varepsilon^2 \cos^2 2\pi y \end{pmatrix}.$$

(1) $(\alpha^* g)_{ab} = J^c{}_a J^d{}_b\, g_{cd}$ with $g_{cd} = \delta_{cd}$.
(2) Multiply out the matrices.

The new metric looks quite different: it has an off-diagonal term and a position-dependent entry. But it is the old flat metric in disguise, since $\det(\alpha^* g) = 1$ and its curvature is zero. By (9.13) its Beltrami coefficient is $\mu = \frac{-\varepsilon'^2 + 2i\varepsilon'}{4 + \varepsilon'^2} \neq 0$ with $\varepsilon' = 2\pi\varepsilon \cos 2\pi y$: the diffeomorphism also changes which coordinates are holomorphic. The path integral must not count g and $\alpha^* g$ as two different configurations.

Physical picture: The orbits are infinite-dimensional, the slice almost is not

Each orbit is a copy of the gauge group, so it is infinite-dimensional. By the function counting of Section 9.1.4, the slice through $\mathrm{Met}(\Sigma)$ is *finite*-dimensional: it is parametrised by the moduli. The slice through $\mathrm{Map}(\Sigma, \mathbb{R}^d)$ is still infinite-dimensional (a string can wiggle in any way), but there the integral is Gaussian. That is the division of labour for the rest of the chapter: Gaussian integral over ϕ now, finite-dimensional integral over metrics

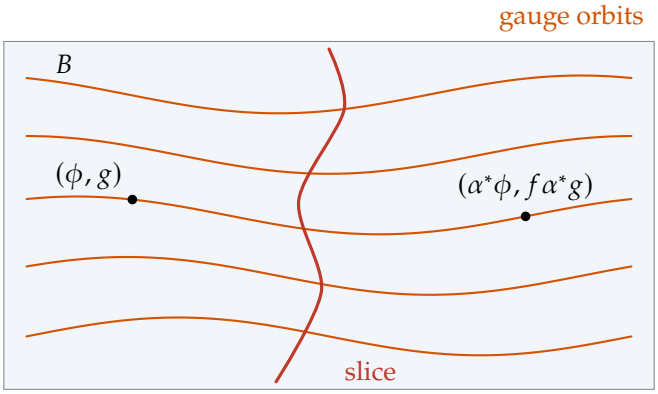


Figure 9.7: The configuration space $B = \mathrm{Map}(\Sigma, \mathbb{R}^d) \times \mathrm{Met}(\Sigma)$ is sliced into *orbits* (orange) of the gauge group $\mathrm{Diff}^+ \ltimes C^\infty_+$. All points on one orbit have the same action and describe the same physics. Integrating over all of B counts each physical configuration an infinite number of times (once per point of its orbit). The cure, carried out in Sections 9.2 and 9.3, is to integrate over a *slice* (red) that meets each orbit once, with a Jacobian that records the size of the orbit.

later.

9.1.8 The integral over the embeddings is Gaussian

For fixed g the action (9.16) is $S = \frac{1}{2}\langle\phi, \Delta_g \phi\rangle_g$, a quadratic form. So the ϕ-integral in (9.10) is an infinite-dimensional Gaussian integral. We first recall the finite-dimensional version in full, because every step of the infinite-dimensional one is a copy of it.

Key idea: The Gaussian integral

For a real symmetric positive-definite $n \times n$ matrix M,

$$\int_{\mathbb{R}^n} \mathrm{d}^n x \; e^{-\frac{1}{2}x^{\mathsf{T}} M x} = \frac{(2\pi)^{n/2}}{\sqrt{\det M}}. \tag{9.22}$$

To see it, rotate to the eigenbasis of M:

$$\begin{aligned}\int_{\mathbb{R}^n} \mathrm{d}^n x \; e^{-\frac{1}{2}x^{\mathsf{T}} M x} &\overset{(1)}{=} \int_{\mathbb{R}^n} \mathrm{d}^n c \; e^{-\frac{1}{2}\sum_k \lambda_k c_k^2} \\ &\overset{(2)}{=} \prod_{k=1}^{n} \int_{-\infty}^{\infty} \mathrm{d}c_k \; e^{-\frac{1}{2}\lambda_k c_k^2} \\ &\overset{(3)}{=} \prod_{k=1}^{n} \sqrt{\frac{2\pi}{\lambda_k}} \overset{(4)}{=} \frac{(2\pi)^{n/2}}{\sqrt{\det M}}.\end{aligned}$$

(1) Write $x = \sum_k c_k e_k$ with e_k an orthonormal eigenbasis, $M e_k = \lambda_k e_k$. The change of variables is a rotation, so $\mathrm{d}^n x = \mathrm{d}^n c$, and $x^{\mathsf{T}} M x = \sum \lambda_k c_k^2$.
(2) The exponential of a sum is a product of exponentials, and the integral factorises.
(3) The one-dimensional Gaussian $\int \mathrm{d}c \; e^{-\lambda c^2/2} = \sqrt{2\pi/\lambda}$.
(4) The determinant is the product of the eigenvalues.

Worked example 9.17: A 2×2 check

Let $M = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$, with eigenvalues 3 (eigenvector $(1,1)/\sqrt{2}$) and 1 (eigenvector $(1,-1)/\sqrt{2}$), so $\det M = 3$. The formula predicts $2\pi/\sqrt{3} \approx 3.628$. Directly, completing the square,

$$\int \mathrm{d}x\,\mathrm{d}y\; e^{-(x^2+xy+y^2)} \overset{(1)}{=} \int \mathrm{d}y\, e^{-\frac{3}{4}y^2} \int \mathrm{d}x\, e^{-(x+y/2)^2}$$
$$\overset{(2)}{=} \sqrt{\frac{4\pi}{3}} \cdot \sqrt{\pi} = \frac{2\pi}{\sqrt{3}}.$$

(1) $\frac{1}{2}x^{\mathsf{T}}Mx = x^2 + xy + y^2 = (x + \frac{y}{2})^2 + \frac{3}{4}y^2$.
(2) Shift x in the inner integral; then use $\int e^{-ac^2}\mathrm{d}c = \sqrt{\pi/a}$ twice.

The normalisation of the measure. The factors of 2π in (9.22) carry no physics, and in infinite dimensions their product would be meaningless. Nash's formula (9.7) has $\det(\Delta_g/2)$ instead of $\det \Delta_g$, which corresponds to normalising the measure so that each mode contributes $\int \frac{\mathrm{d}c}{\sqrt{\pi}}\, e^{-\lambda c^2/2} = (\lambda/2)^{-1/2}$. Equivalently, the measure $\mathcal{D}\phi$ is normalised by $\int \mathcal{D}\phi\, e^{-\langle\phi,\phi\rangle_g} = 1$. Either choice changes the answer by a constant, and we will see in Example 9.26 that for zeta determinants the constant depends only on the genus.

Where the metric g enters. To define $\mathcal{D}\phi$ we need orthonormal coordinates on the space of functions, and "orthonormal" requires an inner product. The natural one is

$$\langle u, v\rangle_g = \int_\Sigma \mathrm{d}^2x\, \sqrt{g}\, u(x)\, v(x). \tag{9.23}$$

So the measure $\mathcal{D}\phi$ itself depends on the metric. This matters: a Weyl rescaling $g \to e^{2\omega}g$ leaves the action invariant but changes $\sqrt{g}$ in (9.23), and hence changes the measure. That is the origin of the conformal anomaly (Section 9.2.11).

The infinite-dimensional Gaussian. Because Σ is compact and Δ_g is elliptic, Δ_g has a complete orthonormal set of eigenfunctions (Chapter 2),

$$\Delta_g e_n = \lambda_n e_n, \qquad \langle e_m, e_n\rangle_g = \delta_{mn}, \qquad 0 = \lambda_0 < \lambda_1 \le \lambda_2 \le \cdots \to \infty. \tag{9.24}$$

Ellipticity is what guarantees that the spectrum is a discrete list with finite multiplicities, growing without bound. The growth rate is given by **Weyl's law**: the number of eigenvalues below λ is $N(\lambda) \approx \frac{\mathrm{Area}}{4\pi}\lambda$ for large λ, so $\lambda_n \approx 4\pi n/\mathrm{Area}$.

Worked example 9.18: The spectrum of the round sphere and Weyl's law

On the unit sphere the eigenfunctions of Δ are the spherical harmonics $Y_{\ell m}$ with $\Delta Y_{\ell m} = \ell(\ell+1)Y_{\ell m}$, $m = -\ell, \dots, \ell$ (we checked $\ell = 1$ in Example 9.12). Count the eigenvalues up to $\Lambda = L(L+1)$:

$$N\big(L(L+1)\big) \overset{(1)}{=} \sum_{\ell=0}^{L} (2\ell+1)$$
$$\overset{(2)}{=} (L+1)^2 \;\approx\; L(L+1) = \frac{4\pi}{4\pi}\Lambda.$$

(1) Each ℓ contributes $2\ell + 1$ eigenfunctions.
(2) The sum of the first $L + 1$ odd numbers is $(L + 1)^2$; for large L this is $L(L + 1)$ up to lower-order terms.

With Area $= 4\pi$ this is exactly Weyl's law. Figure 9.8 shows the spectrum. The single eigenvalue 0 (the constants) is the zero mode.

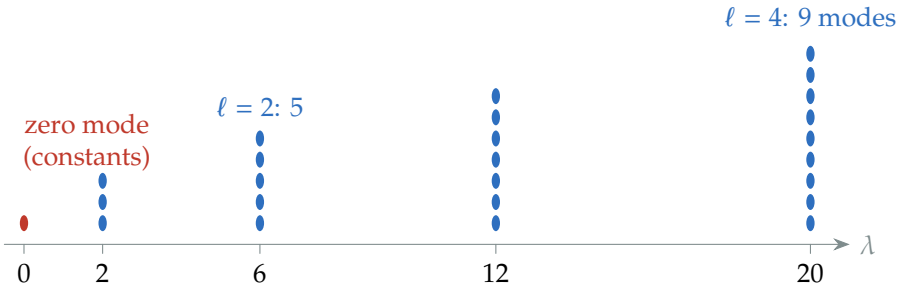


Figure 9.8: The spectrum of the Laplacian on the unit sphere: eigenvalue $\ell(\ell + 1)$ stacked with its multiplicity $2\ell + 1$. The eigenvalues grow without bound and the stacks grow linearly, so the total count grows like λ (Weyl's law). The red dot at $\lambda = 0$ is the constant function: the zero mode that must be treated separately in the Gaussian integral.

Expand each component of the embedding in the eigenfunctions, $\phi^\mu = \sum_n c_n^\mu e_n$. Then

$$S = \tfrac{1}{2}\langle \phi, \Delta_g \phi \rangle_g = \tfrac{1}{2} \sum_{\mu=1}^{d} \sum_{n \geq 0} \lambda_n \, (c_n^\mu)^2, \tag{9.25}$$

by orthonormality, and the measure is $\mathcal{D}\phi = \prod_{\mu,n} \mathrm{d}c_n^\mu / \sqrt{\pi}$. Each non-zero mode gives a factor $(\lambda_n/2)^{-1/2}$, and there are d components. Formally,

$$\int \mathcal{D}\phi \; e^{-S} \;=\; \underbrace{\Big(\int \prod_{\mu} \frac{\mathrm{d}c_0^\mu}{\sqrt{\pi}} \Big)}_{\text{zero mode}} \prod_{n \geq 1} \Big(\frac{\lambda_n}{2} \Big)^{-d/2} \;=\; (\text{zero-mode factor}) \times \Big(\det{}' \frac{\Delta_g}{2} \Big)^{-d/2}. \tag{9.26}$$

Two things need attention. The infinite product $\prod_n \lambda_n$ diverges, since $\lambda_n \to \infty$; it will be defined by a zeta function in Section 9.1.10. And the zero mode gives an infinite factor of a different kind, which we treat now.

9.1.9 The zero mode and the factor $(1, 1)_g$

The eigenvalue $\lambda_0 = 0$ belongs to the constant functions. A constant embedding $\phi^\mu(x) = \phi_0^\mu$ maps the whole world sheet to one point of space-time. It has zero action, and so does $\phi + \phi_0$ for any ϕ: this is translation invariance in $\mathbb{R}^d$. The Gaussian has a flat direction (Figure 9.9), and its integral is proportional to the volume of $\mathbb{R}^d$.

We need to be careful about one thing: the normalised zero mode is not the function 1 but $e_0 = 1/\sqrt{A_g}$, where

$$A_g \;=\; (1, 1)_g \;=\; \int_\Sigma \mathrm{d}^2 x \, \sqrt{g} \tag{9.27}$$

is the area of (Σ, g). (Nash writes $(1, 1)_g$ for the L^2 inner product of the constant function 1 with itself, which is the area.) So the coefficient c_0^μ of the zero mode is related to the physical position ϕ_0^μ of the string's "centre of mass" by $\phi_0^\mu = c_0^\mu / \sqrt{A_g}$.

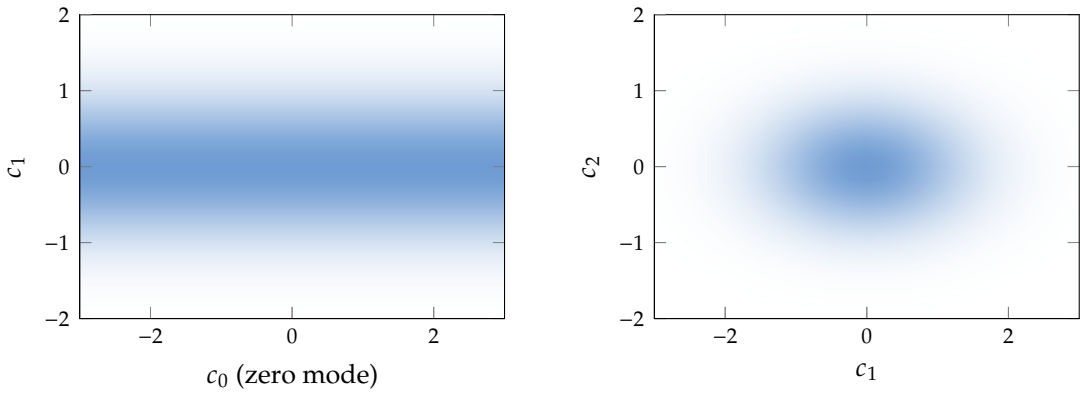


Figure 9.9: Shading shows e^{-S} on a two-dimensional slice of field space. Right: two non-zero modes; the Gaussian is a hill in every direction and its integral is finite, $\propto (\lambda_1\lambda_2)^{-1/2}$. Left: one direction is the zero mode; e^{-S} is constant along it, a ridge running off to infinity. The integral along the ridge is the length of the ridge, which for the string is the volume of space-time.

Worked example 9.19: Where the area comes from

Integrate over the zero mode of one component, with a box of side L in space-time as cutoff (Nash's "compactness cutoff"):

$$\int \frac{\mathrm{d}c_0}{\sqrt{\pi}} \overset{(1)}{=} \int_0^L \frac{\sqrt{A_g}\,\mathrm{d}\phi_0}{\sqrt{\pi}} \overset{(2)}{=} \frac{L}{\sqrt{\pi}}\,A_g^{1/2}.$$

(1) Change variables $c_0 = \sqrt{A_g}\,\phi_0$, and restrict the position ϕ_0 of the string to the box $[0, L]$.
(2) Nothing depends on ϕ_0.

With d components the zero-mode factor is $(L/\sqrt{\pi})^d\,A_g^{d/2} = (\text{volume of the box})\times\pi^{-d/2}\,A_g^{d/2}$. The volume L^d is the same for every metric, so it is an overall constant: we divide by it ("the cutoff may then be discarded"). But the factor $A_g^{d/2}$ *depends on the metric* and must be kept. It is not optional: without it the answer would not be invariant under constant rescalings of g, as we will verify in Example 9.26.

Worked example 9.20: A two-site toy model with a zero mode

Replace Σ by two points, and the Laplacian by the 2×2 matrix $M = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$, so that $S = \frac{1}{2}(x_1 - x_2)^2$ depends only on the difference. Its eigenvalues are 0 (eigenvector $(1,1)/\sqrt{2}$) and 2 (eigenvector $(1,-1)/\sqrt{2}$). The "area" is $\langle (1,1),(1,1)\rangle = 2$, the number of sites. Restrict the centre of mass $m = \frac{1}{2}(x_1 + x_2)$ to $[0, L]$:

$$\begin{aligned}\int \mathrm{d}x_1\,\mathrm{d}x_2\,e^{-\frac{1}{2}(x_1-x_2)^2} &\overset{(1)}{=} \int \mathrm{d}u\,\mathrm{d}v\,e^{-v^2} \\ &\overset{(2)}{=} \left(\sqrt{2}\int_0^L \mathrm{d}m\right)\sqrt{\pi} \\ &= \sqrt{2\pi}\,L\end{aligned}$$

$$\overset{(3)}{=} L \cdot 2^{1/2} \cdot (2\pi)^{1/2} \cdot 2^{-1/2}.$$

(1) Rotate to $u = (x_1+x_2)/\sqrt{2}$, $v = (x_1-x_2)/\sqrt{2}$; then $\frac{1}{2}(x_1-x_2)^2 = v^2$ and $\mathrm{d}x_1\mathrm{d}x_2 = \mathrm{d}u\,\mathrm{d}v$.
(2) $u = \sqrt{2}\,m$, so $\mathrm{d}u = \sqrt{2}\,\mathrm{d}m$; and $\int e^{-v^2}\mathrm{d}v = \sqrt{\pi}$.
(3) Rewrite $\sqrt{2\pi}\,L$ to display its structure: $L \times (\text{area})^{1/2} \times (2\pi)^{(n-1)/2} \times (\det{}'M)^{-1/2}$ with $n = 2$, area $= 2$, $\det{}'M = 2$.

This is (9.22) with the zero eigenvalue left out of the determinant and replaced by (volume)×(area)$^{1/2}$: exactly the structure of Nash's (9.7).

Worked example 9.21: The string on a lattice circle: $\det'$ by hand

Discretise a circle into n sites with the nearest-neighbour Laplacian $(Mx)_j = 2x_j - x_{j+1} - x_{j-1}$ (indices mod n). Its eigenvectors are the discrete Fourier modes $x_j = e^{2\pi i k j/n}$, with eigenvalues

$$\lambda_k \overset{(1)}{=} 2 - e^{2\pi i k/n} - e^{-2\pi i k/n}$$
$$\overset{(2)}{=} 2 - 2\cos\frac{2\pi k}{n} = 4\sin^2\frac{\pi k}{n}, \qquad k = 0, 1, \dots, n-1.$$

(1) Substitute the Fourier mode into $(Mx)_j$ and divide by x_j.
(2) $e^{i\theta} + e^{-i\theta} = 2\cos\theta$ and $1 - \cos\theta = 2\sin^2(\theta/2)$.

The zero mode is $k = 0$, the constant vector. The product of the non-zero eigenvalues is

$$\det{}'M = \prod_{k=1}^{n-1} 4\sin^2\frac{\pi k}{n} = n^2,$$

a classical trigonometric identity. Check: for $n = 3$ the eigenvalues are $3, 3$ and the product is 9; for $n = 4$ they are $2, 4, 2$ with product 16. (This is also Kirchhoff's matrix-tree theorem: $\det{}'M = n\times$(number of spanning trees of the n-cycle), and the n-cycle has n spanning trees.) Keep this n^2 in mind: the zeta-function determinant of the continuum circle will come out as L^2 in Example 9.23.

The result. Put the Gaussian (9.26) and the zero-mode factor of Example 9.19 together, and drop the constant $(L/\sqrt{\pi})^d$:

Key idea: The embedding integral, Nash's (9.7)

$$Z_p = \int_{\mathrm{Met}(\Sigma)} \mathcal{D}g \left(\frac{\det{}'(\Delta_g/2)}{(1,1)_g}\right)^{-d/2}. \tag{9.28}$$

Each of the d scalar fields ϕ^μ contributes $\left(\det{}'(\Delta_g/2)\right)^{-1/2}$ from its non-zero modes and $(1,1)_g^{1/2} = A_g^{1/2}$ from its zero mode.

Nash writes det rather than det′ and explains in words that the determinant is "restricted to the strictly positive subspace"; we write the prime to make that visible.

Watch out: Two different infinities

The zero mode and the infinite product are infinite for different reasons. The zero mode is an *infrared* infinity: it comes from the infinite volume of space-time, and it is physical (the vacuum amplitude really is proportional to the volume, as for the free particle in Example 9.9). The product $\prod_n \lambda_n$ is an *ultraviolet* infinity: it comes from arbitrarily short-wavelength wiggles of the string, and it is removed by regularisation. Nash's footnote handles the second with a zeta function.

9.1.10 Zeta-function determinants

We now have to give a meaning to $\det{}'\Delta_g = \prod_{n\geq 1}\lambda_n$, a product of infinitely many numbers that grow without bound. Nash's footnote (p. 261) does it in one line: "$\zeta_P(s)$ is always regular at the origin and this means that a regularised determinant of P can be defined as $\exp[-\zeta_P'(0)]$". Here is where that line comes from, and what it gives in examples.

The idea. For a finite list of positive numbers, the logarithm of the product is the sum of the logarithms, and a logarithm is a derivative of a power:

$$\log\lambda = -\frac{\mathrm{d}}{\mathrm{d}s}\lambda^{-s}\Big|_{s=0}.$$

So $\log\prod_n\lambda_n = -\frac{\mathrm{d}}{\mathrm{d}s}\sum_n\lambda_n^{-s}\big|_{s=0}$. The sum $\sum_n\lambda_n^{-s}$ still makes sense for an infinite list, provided s is large enough that the terms decay fast; and then we can try to continue it to $s=0$.

Working definition 9.5 (Zeta function and zeta-regularised determinant)

Let P be a positive elliptic operator (such as Δ_g) [its eigenvalues are real, ≥ 0 and discrete, and they grow without bound] on a compact manifold, with non-zero eigenvalues $0<\lambda_1\leq\lambda_2\leq\cdots$ listed with multiplicity [an eigenvalue with k independent eigenfunctions appears k times]. Its **zeta function** is

$$\zeta_P(s) = \sum_{n\geq 1}\lambda_n^{-s}. \tag{9.29}$$

The sum converges for $\operatorname{Re}s > \frac{\dim M}{\operatorname{ord}P}$ [$\operatorname{ord}P$ is the number of derivatives in P, 2 for Δ_g] (for Δ_g on a surface: $\operatorname{Re}s>1$, by Weyl's law $\lambda_n\approx 4\pi n/A$). It continues to a meromorphic function of s [there is exactly one function, analytic except at isolated poles, that equals the sum wherever the sum converges] that is regular at $s=0$. The **zeta-regularised determinant** is

$$\det{}'P = \exp\bigl[-\zeta_P'(0)\bigr]. \tag{9.30}$$

How to compute it: find the spectrum, write ζ_P in terms of a known zeta function (Riemann, Hurwitz, Epstein), and differentiate at $s=0$.

Worked example 9.22: For a matrix, nothing new

Let $P=\operatorname{diag}(2,3)$. Then $\zeta_P(s)=2^{-s}+3^{-s}$ and

$$\exp\bigl[-\zeta_P'(0)\bigr] \overset{(1)}{=} \exp\bigl[\log 2\cdot 2^{-0}+\log 3\cdot 3^{-0}\bigr]$$
$$\overset{(2)}{=} e^{\log 6} = 6 = \det P.$$

(1) $\frac{\mathrm{d}}{\mathrm{d}s}\lambda^{-s} = -\log\lambda\,\lambda^{-s}$.

(2) $\log 2 + \log 3 = \log 6$.

Note also $\zeta_P(0) = 2$, the number of eigenvalues. For a genuine operator $\zeta_P(0)$ is a finite "regularised number of eigenvalues", and it reappears in the scaling law of Example 9.26.

Two values of the Riemann zeta function. Everything on the circle and the torus reduces to the Riemann zeta function $\zeta(s) = \sum_{n\geq 1} n^{-s}$ (convergent for $\operatorname{Re} s > 1$), and we need its value and slope at $s = 0$:

$$\zeta(0) = -\tfrac{1}{2}, \qquad \zeta'(0) = -\tfrac{1}{2}\log(2\pi). \tag{9.31}$$

Here is a physicist's derivation. Use the alternating version $\eta(s) = \sum_{n\geq 1}(-1)^{n+1}n^{-s}$, which converges for $\operatorname{Re} s > 0$ and is better behaved. Subtracting twice the even terms from ζ,

$$\eta(s) \overset{(1)}{=} \sum_{n\geq 1} n^{-s} - 2\sum_{n\geq 1}(2n)^{-s} \overset{(2)}{=} (1 - 2^{1-s})\,\zeta(s).$$

(1) $1 - 2^{-s} + 3^{-s} - \cdots$ is the full sum minus twice the even terms.
(2) $(2n)^{-s} = 2^{-s}n^{-s}$.

This identity continues ζ to $\operatorname{Re} s > 0$ (except $s = 1$), and η itself continues further. At $s = 0$, $\eta(0) = 1 - 1 + 1 - \cdots = \frac{1}{2}$ in the sense of Abel summation ($\sum(-x)^n \to \frac{1}{1+x} \to \frac{1}{2}$), so $\zeta(0) = \eta(0)/(1-2) = -\frac{1}{2}$. Differentiating the identity at $s = 0$,

$$\eta'(0) \overset{(1)}{=} 2\log 2\;\zeta(0) + (1-2)\,\zeta'(0) \overset{(2)}{=} -\log 2 - \zeta'(0).$$

(1) Product rule, with $\frac{\mathrm{d}}{\mathrm{d}s}(1 - 2^{1-s}) = 2^{1-s}\log 2$, equal to $2\log 2$ at $s = 0$.
(2) $\zeta(0) = -\frac{1}{2}$.

Finally $\eta'(0) = \sum_{n\geq 1}(-1)^n \log n = \log\frac{2}{1} + \log\frac{4}{3} + \log\frac{6}{5} + \cdots$, and pairing the terms against Wallis' product $\prod_{k\geq 1}\frac{(2k)^2}{(2k-1)(2k+1)} = \frac{\pi}{2}$ gives $\eta'(0) = \frac{1}{2}\log\frac{\pi}{2}$. (The partial sums oscillate; their average converges to this value.) Hence $\zeta'(0) = -\log 2 - \frac{1}{2}\log\frac{\pi}{2} = -\frac{1}{2}\log(2\pi)$.

Watch out: This is not "$1 + 1 + 1 + \cdots = -\frac{1}{2}$"

$\zeta(0) = -\frac{1}{2}$ is a statement about the analytic continuation of a function, not about the divergent sum. What the zeta recipe does, in effect, is throw away terms that are *local* in the ultraviolet (terms that could be cancelled by counterterms in the action) and keep a finite remainder. The lattice comparison in Example 9.23 shows this explicitly.

Worked example 9.23: The Laplacian on a circle: $\det'\Delta = L^2$

Let S^1_L be a circle of circumference L and $\Delta = -\mathrm{d}^2/\mathrm{d}\theta^2$. The eigenfunctions are $e^{2\pi i n\theta/L}$, $n \in \mathbb{Z}$, with eigenvalue $(2\pi n/L)^2$. The value 0 ($n = 0$) is the zero mode, and every non-zero eigenvalue occurs twice ($\pm n$). So

$$\zeta_\Delta(s) \overset{(1)}{=} 2\sum_{n\geq 1}\left(\frac{2\pi n}{L}\right)^{-2s} \overset{(2)}{=} 2\left(\frac{L}{2\pi}\right)^{2s}\zeta(2s).$$

(1) Sum over $n \neq 0$; the terms for n and $-n$ are equal.

(2) Pull out $(2\pi/L)^{-2s}$ and recognise $\sum n^{-2s} = \zeta(2s)$.

Now differentiate and set $s = 0$:

$$\begin{aligned}\zeta'_\Delta(0) &\overset{(1)}{=} 2 \cdot 2\log\frac{L}{2\pi}\cdot\zeta(0) + 2\cdot 2\,\zeta'(0)\\ &\overset{(2)}{=} -2\log\frac{L}{2\pi} - 2\log(2\pi)\\ &= -2\log L.\end{aligned}$$

(1) Product rule: $\frac{\mathrm{d}}{\mathrm{d}s}(L/2\pi)^{2s} = 2\log(L/2\pi)(L/2\pi)^{2s}$ and $\frac{\mathrm{d}}{\mathrm{d}s}\zeta(2s) = 2\zeta'(2s)$; then set $s = 0$.
(2) The two Riemann values (9.31).

Therefore

$$\det{}'\Delta_{S^1_L} = e^{2\log L} = L^2, \qquad \zeta_\Delta(0) = 2\zeta(0) = -1. \tag{9.32}$$

Compare the lattice circle of Example 9.21, with n sites and spacing $a = L/n$. The lattice Laplacian with the correct units is M/a^2, so

$$\begin{aligned}\det{}'\Big(\frac{M}{a^2}\Big) &\overset{(1)}{=} a^{-2(n-1)}\,n^2\\ &\overset{(2)}{=} L^2\cdot a^{-2n} \;=\; L^2\cdot\exp\Big[-\frac{2\log a}{a}\,L\Big].\end{aligned}$$

(1) $n-1$ non-zero eigenvalues, each divided by a^2, and $\det{}'M = n^2$.
(2) $n^2a^{-2(n-1)} = (na)^2a^{-2n} = L^2a^{-2n}$, and $n = L/a$.

The lattice answer is the zeta answer L^2 times a factor that diverges as $a \to 0$. The divergent factor is the exponential of (a constant)×(the length L): the same form as the exponential of a local term $\mu\int \mathrm{d}s$ in the action (a "cosmological constant" on the world line), so it can be cancelled by a counterterm. The zeta function silently removes it. This is the precise sense of the pitfall above.

Worked example 9.24: The harmonic oscillator: $\det(\Delta+\omega^2) = 4\sinh^2(\omega\beta/2)$

A test with a known answer. The partition function of a harmonic oscillator of frequency ω at inverse temperature β is $\operatorname{Tr} e^{-\beta H} = \sum_{n\ge0} e^{-\beta\omega(n+1/2)} = \frac{1}{2\sinh(\beta\omega/2)}$. As a Euclidean path integral over periodic paths of period β (a circle of length $L = \beta$) it is $\big(\det(\Delta+\omega^2)\big)^{-1/2}$. There is no zero mode now. The eigenvalues are $(2\pi n/\beta)^2 + \omega^2$, $n \in \mathbb{Z}$. Compare with the massless case, where the ratio of products converges:

$$\begin{aligned}\frac{\det(\Delta+\omega^2)}{\det{}'\Delta} &\overset{(1)}{=} \omega^2\prod_{n\neq0}\frac{(2\pi n/\beta)^2+\omega^2}{(2\pi n/\beta)^2}\\ &\overset{(2)}{=} \omega^2\left[\prod_{n\ge1}\Big(1+\frac{(\omega\beta/2)^2}{\pi^2n^2}\Big)\right]^2\\ &\overset{(3)}{=} \omega^2\Big(\frac{\sinh(\omega\beta/2)}{\omega\beta/2}\Big)^2.\end{aligned}$$

(1) The $n = 0$ eigenvalue ω^2 has no partner in $\det{}'\Delta$; the others are divided pairwise. (For zeta determinants this multiplicativity holds here because the ratio product converges absolutely.)
(2) n and $-n$ give equal factors, and $\omega^2/(2\pi n/\beta)^2 = (\omega\beta/2)^2/(\pi n)^2$.

(3) Euler's product $\frac{\sinh x}{x} = \prod_{n\geq 1}\left(1+\frac{x^2}{\pi^2 n^2}\right)$.

With $\det'\Delta = \beta^2$ from (9.32), $\det(\Delta+\omega^2) = \omega^2\beta^2 \cdot \frac{4\sinh^2(\omega\beta/2)}{\omega^2\beta^2} = 4\sinh^2(\omega\beta/2)$, and so $\big(\det(\Delta+\omega^2)\big)^{-1/2} = \frac{1}{2\sinh(\omega\beta/2)}$, exactly the oscillator's partition function, *including the zero-point energy* $\frac{1}{2}\omega$ (visible as $e^{-\beta\omega/2}$ at large β). The zeta recipe gets the physically correct normalisation.

Why ζ_P is regular at $s=0$: the heat kernel. The analytic continuation, and the value $\zeta_P(0)$, come from the **heat trace** $K(t) = \operatorname{Tr} e^{-tP} = \sum_{n\geq 0} e^{-t\lambda_n}$ (Chapter 4). Its small-t behaviour is local and universal; on a surface, for the Laplacian,

$$K(t) = \frac{A_g}{4\pi t} + \frac{1}{4\pi}\int_\Sigma \mathrm{d}^2x\,\sqrt{g}\,\frac{R}{6} + O(t) = \frac{A_g}{4\pi t} + \frac{\chi(\Sigma)}{6} + O(t), \tag{9.33}$$

where the second form used Gauss–Bonnet, $\int\sqrt{g}\,R = 4\pi\chi$. The bridge between K and ζ is the integral $\lambda^{-s} = \frac{1}{\Gamma(s)}\int_0^\infty t^{s-1}e^{-t\lambda}\,\mathrm{d}t$, which gives

$$\zeta_P(s) = \frac{1}{\Gamma(s)}\int_0^\infty t^{s-1}\big(K(t)-n_0\big)\,\mathrm{d}t, \tag{9.34}$$

with $n_0 = \dim\operatorname{Ker} P$ (here $n_0 = 1$) subtracted so that only non-zero eigenvalues are counted.

Worked example 9.25: $\zeta_\Delta(0) = \frac{\chi}{6} - 1$ from the heat kernel

Split (9.34) at $t=1$. The part from 1 to ∞ converges for every s (the integrand decays like $e^{-\lambda_1 t}$). In the part from 0 to 1 insert $K(t) - n_0 = \frac{a_{-1}}{t} + (a_0 - n_0) + O(t)$:

$$\int_0^1 t^{s-1}\big(K(t)-n_0\big)\mathrm{d}t \overset{(1)}{=} \frac{a_{-1}}{s-1} + \frac{a_0-n_0}{s} + (\text{regular at } s=0).$$

(1) $\int_0^1 t^{s-2}\mathrm{d}t = \frac{1}{s-1}$ and $\int_0^1 t^{s-1}\mathrm{d}t = \frac{1}{s}$; the $O(t)$ terms give functions regular near $s=0$.

Near $s=0$, $\frac{1}{\Gamma(s)} = s + O(s^2)$. Multiplying, the pole $\frac{1}{s}$ is cancelled, ζ_P is regular at 0, and

$$\zeta_P(0) = a_0 - n_0.$$

For the Laplacian on a surface, $\zeta_\Delta(0) = \frac{\chi}{6} - 1$. Checks: the unit sphere has $K(t) = \sum_\ell (2\ell+1)e^{-t\ell(\ell+1)} = \frac{1}{t} + \frac{1}{3} + O(t)$ (Euler–Maclaurin, or a quick numerical sum at $t = 0.001$), so $\zeta(0) = \frac{1}{3} - 1 = -\frac{2}{3}$. For the circle there is no constant term ($K(t) \approx L/\sqrt{4\pi t}$ up to exponentially small corrections, Figure 9.10), so $\zeta(0) = 0 - 1 = -1$, agreeing with (9.32).

Worked example 9.26: Rescaling an operator: $\det'(cP) = c^{\zeta_P(0)}\det'P$

For a constant $c>0$ the eigenvalues of cP are $c\lambda_n$, so $\zeta_{cP}(s) = c^{-s}\zeta_P(s)$ and

$$\begin{aligned}\det'(cP) &\overset{(1)}{=} \exp\Big[-\frac{\mathrm{d}}{\mathrm{d}s}\big(c^{-s}\zeta_P(s)\big)\Big|_{s=0}\Big]\\ &\overset{(2)}{=} \exp\big[\log c\;\zeta_P(0) - \zeta_P'(0)\big]\\ &= c^{\zeta_P(0)}\,\det'P.\end{aligned}$$

(1) Definition (9.30) applied to cP.

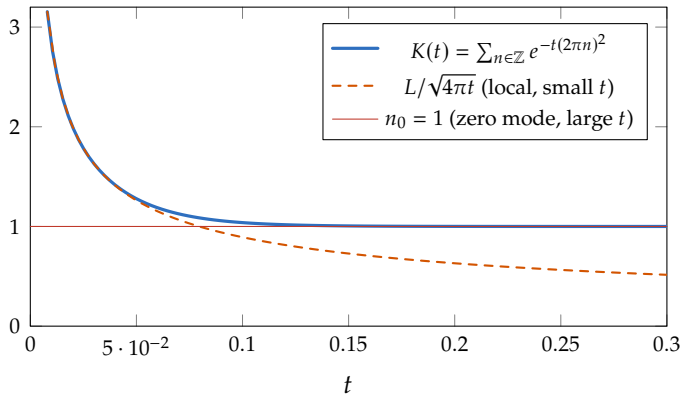


Figure 9.10: The heat trace of the circle of length $L = 1$. For small t it is indistinguishable from the local term $L/\sqrt{4\pi t}$: short times probe only short distances, where the circle looks like a line. For large t only the zero mode survives and $K \to 1$. The zeta function is the Mellin transform (9.34) of $K - 1$; the absence of a t^0 term at small t is why $\zeta_\Delta(0) = 0 - 1 = -1$.

(2) Product rule: $\frac{\mathrm{d}}{\mathrm{d}s} c^{-s} = -\log c\, c^{-s}$.

For a matrix $\zeta_P(0) = n$ and this is $\det(cP) = c^n \det P$. For an operator the "number of eigenvalues" is replaced by $\zeta_P(0)$. Two consequences.
(i) Nash's factor $\frac{1}{2}$. $\det'(\Delta_g/2) = 2^{-\zeta(0)} \det'\Delta_g = 2^{1-\chi/6} \det'\Delta_g$. The factor depends only on the genus, so it is a harmless constant in each Z_p.
(ii) Constant rescalings of the metric. Replace g by λg with $\lambda > 0$ constant. Then $\Delta_{\lambda g} = \lambda^{-1}\Delta_g$ and the area becomes λA_g, so

$$\frac{\det'\Delta_{\lambda g}}{A_{\lambda g}} \overset{(1)}{=} \frac{\lambda^{-\zeta(0)} \det'\Delta_g}{\lambda A_g} \overset{(2)}{=} \lambda^{-\chi/6} \frac{\det'\Delta_g}{A_g}.$$

(1) The scaling law with $c = \lambda^{-1}$.
(2) $-\zeta(0) - 1 = -(\chi/6 - 1) - 1 = -\chi/6$.

Without the area factor we would have had $\lambda^{1-\chi/6}$; the zero mode's A_g removes the "-1" coming from n_0. What remains, $\lambda^{-\chi/6}$, is the first appearance of the conformal anomaly: even a *constant* Weyl rescaling changes the matter determinant, except on the torus ($\chi = 0$). With $\lambda = e^{2\omega}$, the factor $(\det'\Delta/A)^{-d/2}$ in Z_p is multiplied by $e^{d\chi\omega/6}$. We return to this in Section 9.2.11.



Figure 9.11: Two circle spectra used in this subsection (horizontal positions schematic). Top: the massless Laplacian, with a zero mode (red) and doubly degenerate eigenvalues; $\det'\Delta = L^2$. Bottom: adding ω^2 shifts every eigenvalue, lifts the zero mode, and gives $\det(\Delta + \omega^2) = 4\sinh^2(\omega L/2)$, the harmonic oscillator.

Reading Nash (pp. 259–262, (9.1)–(9.7))

- (9.1) is the Nambu–Goto action (9.4); (9.2) the Polyakov action (9.5), whose minus sign we dropped (Euclidean world sheet). "At a critical point the two expressions coincide" is Example 9.5.
- (9.3) is (9.10); "vacuum-to-vacuum amplitude" and "trace" are Figure 9.5; (9.5) is the genus sum (9.11).
- (9.4) is (9.16), with the remarks on signs and on $\partial\partial^*$ in the pitfall after Example 9.12.
- (9.6), the map A, is (9.20); the group law of the semidirect product is (9.21).
- (9.7) is (9.28). "Restriction of $\Delta_g/2$ to its strictly positive subspace" is our det′; the "volume of $\mathbb{R}^d$" and the factor $(1,1)_g$ are Example 9.19; footnote 1 is Section 9.1.10.

Recap: Section

- The Polyakov action $S = \frac{1}{2}\int \sqrt{g}\, g^{ab}\partial_a\phi\cdot\partial_b\phi$ is quadratic in ϕ; eliminating g gives back the Nambu–Goto area.
- It is invariant under $\mathrm{Diff}^+(\Sigma) \ltimes C^\infty_+(\Sigma)$ (Weyl invariance is special to two dimensions) and under translations of $\mathbb{R}^d$.
- Closed world sheets compute the vacuum amplitude (a trace); summing over genus is the string loop expansion.
- A metric on an oriented surface is a complex structure plus a scale; in isothermal coordinates $S = 2\int \partial_z\phi\cdot\partial_{\bar z}\phi\,\mathrm{d}^2x$.
- The ϕ integral is Gaussian: $Z_p = \int \mathcal{D}g\,\left(\det{}'(\Delta_g/2)/A_g\right)^{-d/2}$, the area coming from the zero mode.
- $\det{}'P = e^{-\zeta_P'(0)}$; on the circle $\det{}'\Delta = L^2$; $\zeta_\Delta(0) = \chi/6 - 1$ on a surface; $\det{}'(cP) = c^{\zeta_P(0)}\det{}'P$.

9.2 The space of metrics

After the Gaussian integral, Nash's formula (9.28) still contains an integral over *all* metrics on Σ. How do we integrate over an infinite-dimensional space whose points are mostly related by gauge transformations? The answer is the one familiar from Yang–Mills theory: divide by the volume of the gauge group, pick a slice, and pay a Jacobian. In this section we set up the geometry needed to do that: a metric on the space of metrics, the infinitesimal gauge transformations, and an orthogonal decomposition of the tangent space $T_g\mathrm{Met}(\Sigma)$ into "Weyl", "diffeomorphism" and "moduli" directions. The moduli directions turn out to be the kernel of an elliptic operator, so there are finitely many of them.

9.2.1 What is left to do

Nash's plan (p. 262) has three steps.

1. Split the metrics into conformal classes (metrics related by Weyl rescalings). A conformal class on an oriented surface is the same thing as a complex structure (Section 9.1.6).

2. Identify complex structures related by diffeomorphisms. What is left is the moduli space $\mathcal{M}_p$ of Riemann surfaces of genus p from Chapter 5, which is finite-dimensional.

3. Express Z_p as an integral over $\mathcal{M}_p$, with the right Jacobian.

Watch out: Why $d = 26$?

Step 1 presupposes that the integrand does not depend on the Weyl factor. The classical action does not, but the quantum measure does (Example 9.26): the dependence cancels only when $d = 26$. This is the **conformal anomaly**, treated in Chapter 10 and previewed in Section 9.2.11. Following Nash we set $d = 26$ from now on, so that the Weyl group really is a symmetry of the quantum theory and can be divided out.

Dividing by the gauge group is expressed by Nash's formal definition (9.8):

$$\mathcal{D}g = \frac{\mathcal{D}\mathrm{Met}(\Sigma)}{\mathrm{vol}\big(C^\infty_+(\Sigma)\big)\ \mathrm{vol}\big(\mathrm{Diff}^+(\Sigma)\big)}. \tag{9.35}$$

Here $\mathcal{D}$Met is a (yet to be defined) measure on all of Met(Σ), and the denominator is the volume of a gauge orbit. Such a measure is "supported on the quotient": it integrates over one representative per orbit. Before doing this in infinite dimensions, let us do it once in two dimensions.

9.2.2 Gauge fixing in finite dimensions

Worked example 9.27: Dividing by rotations: polar coordinates as gauge fixing

Let $F(x)$ be a rotation-invariant function on $\mathbb{R}^2$, $F(x) = F(|x|)$. The "gauge group" is SO(2), the "gauge orbits" are the circles $|x| = r$, and a "slice" is the positive x^1-axis, which meets each orbit once (Figure 9.12). The generator of rotations moves the point $(r, 0)$ of the slice with velocity

$$A(\dot\theta) = \dot\theta\,(-x^2, x^1)\big|_{(r,0)} = \dot\theta\,(0, r),$$

a linear map $A : \mathfrak{so}(2) = \mathbb{R} \to T_{(r,0)}\mathbb{R}^2$. Now

$$\begin{aligned}\frac{1}{\mathrm{vol}\,\mathrm{SO}(2)}\int_{\mathbb{R}^2} \mathrm{d}^2x\, F(|x|) &\overset{(1)}{=} \frac{1}{2\pi}\int_0^\infty \mathrm{d}r \int_0^{2\pi} \mathrm{d}\theta\; r\, F(r)\\ &\overset{(2)}{=} \int_0^\infty \mathrm{d}r\, \sqrt{\det(A^\mathsf{T}A)}\, F(r).\end{aligned}$$

(1) Polar coordinates, with $\mathrm{vol}\,\mathrm{SO}(2) = \int_0^{2\pi}\mathrm{d}\theta = 2\pi$.
(2) The θ integral cancels the group volume. The Jacobian r is the length of the velocity $A(1) = (0, r)$, i.e. $\sqrt{A^\mathsf{T}A} = \sqrt{r^2}$.

Three features carry over verbatim to strings. (a) Dividing by the gauge volume leaves an integral over the slice. (b) The slice integral carries a Jacobian, $\sqrt{\det A^\mathsf{T}A}$, where A maps the Lie algebra of the gauge group to the directions along the orbit: this is the **Faddeev–Popov determinant**. (c) At $r = 0$ the orbit is a single point; the rotations do not move it at all ($A = 0$), the Jacobian vanishes, and the action of the group is not free there. For strings the analogue of $r = 0$ is a metric with conformal Killing vectors.

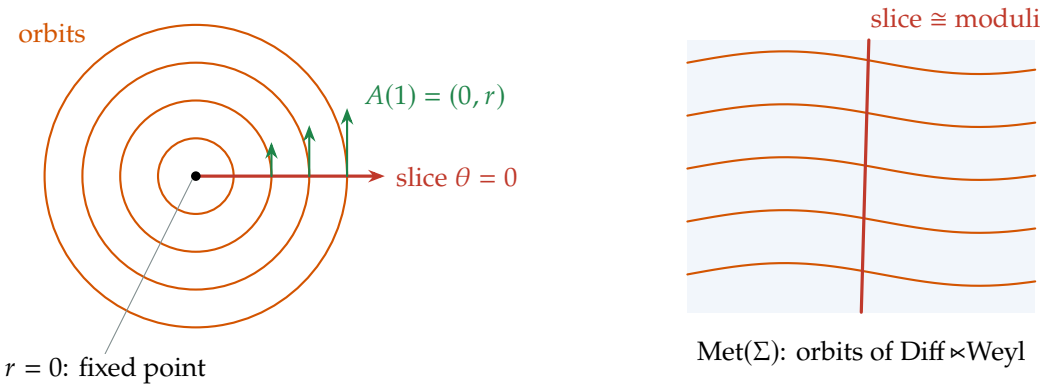


Figure 9.12: Left: gauge fixing for rotations of the plane. The orbits are circles; the slice is a ray; the Jacobian is the speed r with which the group moves the slice point, which vanishes at the fixed point $r = 0$. Right: the same picture for metrics. The orbits of $\mathrm{Diff}^+ \ltimes C^\infty_+$ are infinite-dimensional, the slice is finite-dimensional, and the Jacobian will be $\sqrt{\det{}' P_1^\dagger P_1}$.

9.2.3 Reminder: Teichmüller space and moduli space

The quotient we need was constructed in Chapter 5 (Nash §5.4–5.5, pp. 148–166). Here is what we use from it.

Three groups of diffeomorphisms. $\mathrm{Diff}^+(\Sigma)$ is the group of orientation-preserving diffeomorphisms. Its identity component $\mathrm{Diff}_0(\Sigma)$ consists of the diffeomorphisms that can be continuously deformed to the identity, for example flows $\exp[tV]$ of vector fields. The quotient

$$\Gamma_\Sigma = \mathrm{Diff}^+(\Sigma)/\mathrm{Diff}_0(\Sigma) \tag{9.36}$$

is the **mapping class group**, a discrete group. For the torus $\mathbb{R}^2/\mathbb{Z}^2$ it is $\mathrm{SL}(2,\mathbb{Z})$: the linear map $(x, y) \mapsto (x + y, y)$ (a **Dehn twist**) is a diffeomorphism of the torus, since it maps $\mathbb{Z}^2$ to itself, but it cannot be deformed to the identity, because it changes which loops are which in $\pi_1(T^2) = \mathbb{Z}^2$.

Uniformisation and constant-curvature metrics. Every metric on a closed surface is conformal to a metric of constant Gauss curvature K (Riemann's uniformisation theorem): $+1$ for the sphere, 0 for the torus, -1 for $p \geq 2$. The constant-curvature representative is unique for $p \neq 1$; for the torus we fix the scale by demanding area 1. Nash's (5.42) collects these representatives into

$$\mathrm{Met}_{\mathrm{const}}(\Sigma) = \begin{cases} \{g : K(g) = +1\}, & p = 0, \\ \{g : K(g) = 0,\ \mathrm{Area} = 1\}, & p = 1, \\ \{g : K(g) = -1\}, & p \geq 2. \end{cases} \tag{9.37}$$

So $\mathrm{Met}_{\mathrm{const}}$ is a slice for the Weyl group: every conformal class meets it once. (For $p \geq 2$ the area is then fixed by Gauss–Bonnet: $-\mathrm{Area} = \int K \, \mathrm{d}A = 2\pi\chi$, so $\mathrm{Area} = 4\pi(p-1)$.)

The quotients. The **Teichmüller space** and the **moduli space** are

$$\mathcal{T}_p = \frac{\mathrm{Met}_{\mathrm{const}}(\Sigma)}{\mathrm{Diff}_0(\Sigma)}, \qquad \mathcal{M}_p = \frac{\mathrm{Met}_{\mathrm{const}}(\Sigma)}{\mathrm{Diff}^+(\Sigma)} = \frac{\mathcal{T}_p}{\Gamma_\Sigma}. \tag{9.38}$$

$\mathcal{T}_p$ is a smooth manifold (a ball); $\mathcal{M}_p$ is an orbifold, because Γ_Σ has fixed points on $\mathcal{T}_p$. Their real dimensions are

$$\dim_{\mathbb{R}} \mathcal{M}_p = \begin{cases} 0, & p = 0, \\ 2, & p = 1, \\ 6p - 6, & p \geq 2, \end{cases} \tag{9.39}$$

(complex dimensions $0, 1, 3p-3$). For the torus, $\mathcal{T}_1$ is the upper half-plane $\{\tau : \operatorname{Im}\tau > 0\}$ and $\mathcal{M}_1 = \mathcal{T}_1/\mathrm{SL}(2,\mathbb{Z})$. We rederive (9.39) from the operator P_1 in Section 9.2.9.

$$\underset{\text{all metrics}}{\operatorname{Met}(\Sigma)} \xrightarrow{/\,C^\infty_+} \underset{\text{complex structures}}{\operatorname{Conf}(\Sigma) \cong \operatorname{Met}_{\text{const}}(\Sigma)} \xrightarrow{/\,\mathrm{Diff}_0} \underset{\text{Teichmüller, a ball}}{\mathcal{T}_p} \xrightarrow{/\,\Gamma_\Sigma} \underset{\text{moduli, an orbifold}}{\mathcal{M}_p}$$

$$\infty \text{ (3 fns)} \qquad \infty \text{ (2 fns)} \qquad 6p-6 \qquad 6p-6$$

Figure 9.13: The tower of quotients (Nash §5.4–5.5). Each arrow divides by a group; the bottom row is the real dimension ("3 fns" means three functions' worth). Dividing by the two-function-worth group Diff_0 takes an infinite-dimensional space to a finite-dimensional one: the moduli are what the counting "$3-1-2=0$" leaves behind globally.

Watch out: Why Nash switches from Met to $\mathrm{Met}_{\text{const}}$ and from Diff^+ to Diff_0

Two technical problems are hidden in (9.35). First, Diff^+ does not act freely: the discrete part Γ_Σ has fixed points, so one first divides by the connected group Diff_0 and deals with Γ_Σ at the end (it produces the factor $1/|\Gamma_\Sigma|$ in (9.35)). Second, for $p = 0, 1$ even Diff_0 does not act freely on conformal classes: there are diffeomorphisms (flows of conformal Killing vectors, Section 9.2.5) that map a metric to a conformally equivalent one. This is the $r = 0$ problem of Example 9.27, and it is why Nash quotients $\mathrm{Met}_{\text{const}}$ rather than Met directly.

9.2.4 A Riemannian metric on the space of metrics

To define a measure $\mathcal{D}\mathrm{Met}$ we copy the finite-dimensional recipe. On a Riemannian manifold with metric G, coordinates x^i and coordinate basis $e_i = \partial/\partial x^i$, the invariant volume element is

$$\mathrm{dvol} = \sqrt{\det\langle e_i, e_j\rangle_G}\, \mathrm{d}^n x = \sqrt{G}\, \mathrm{d}^n x. \tag{9.40}$$

For the unit sphere in coordinates (θ, φ), $\langle e_\theta, e_\theta\rangle = 1$, $\langle e_\varphi, e_\varphi\rangle = \sin^2\theta$, $\langle e_\theta, e_\varphi\rangle = 0$, so $\mathrm{dvol} = \sin\theta\, \mathrm{d}\theta\, \mathrm{d}\varphi$. The lesson: *a metric gives a measure*. So we look for a metric on $\mathrm{Met}(\Sigma)$.

A tangent vector at g is a symmetric tensor field (9.19). Nash writes tangent vectors in "coordinates" g^I on Met, where the index I runs over the pairs (ab) *and* the points x of Σ (his (9.9)):

$$X = X^I \frac{\partial}{\partial g^I} = X_{ab}(x) \frac{\partial}{\partial g_{ab}(x)}, \qquad Y = Y_{ab}(x) \frac{\partial}{\partial g_{ab}(x)},$$

with a sum over $a \leq b$ and an "integral" over x understood.

Working definition 9.6 (The natural metric on $\mathrm{Met}(\Sigma)$)

For $X, Y \in T_g\mathrm{Met}(\Sigma)$ [symmetric tensor fields $X_{ab}(x)$, the small changes $g \to g + \epsilon X$ of

the metric] (Nash's (9.10)–(9.11)),

$$\langle X, Y\rangle_{\mathrm{met}} = \int_\Sigma \mathrm{d}^2x\,\sqrt{g}\,\langle X, Y\rangle_{g(x)},$$
$$\langle X, Y\rangle_{g(x)} = g^{ac}(x)\,g^{bd}(x)\,X_{ab}(x)\,Y_{cd}(x) = \mathrm{tr}\big(g^{-1}Xg^{-1}Y\big). \tag{9.41}$$

How to use it: at each point, raise both indices of X with g, contract with Y (a matrix trace), and integrate with the Riemannian area element of g.

Worked example 9.28: The pointwise metric in components

At a point where $g = \mathbf{1}$, write $X = \begin{pmatrix} x_1 & x_2 \\ x_2 & x_3 \end{pmatrix}$. Then

$$\langle X, X\rangle_{\mathbf{1}} \overset{(1)}{=} \mathrm{tr}(X^2)$$
$$\overset{(2)}{=} x_1^2 + 2x_2^2 + x_3^2.$$

(1) $g^{-1} = \mathbf{1}$.
(2) $\mathrm{tr}(X^2) = \sum_{a,b} X_{ab}X_{ba}$; the off-diagonal entry appears twice.

Split X into its trace and traceless parts, $X = \frac{x_1+x_3}{2}\mathbf{1} + \begin{pmatrix} u & x_2 \\ x_2 & -u \end{pmatrix}$ with $u = \frac{x_1-x_3}{2}$. The two parts are orthogonal, $\mathrm{tr}\big(\mathbf{1}\cdot\left(\begin{smallmatrix} u & x_2 \\ x_2 & -u \end{smallmatrix}\right)\big) = 0$, and $\langle X, X\rangle = \frac{(x_1+x_3)^2}{2} + 2(u^2 + x_2^2)$. At a general point where $g = \mathrm{diag}(\lambda_1, \lambda_2)$, $\langle X, X\rangle_g = \frac{x_1^2}{\lambda_1^2} + \frac{2x_2^2}{\lambda_1\lambda_2} + \frac{x_3^2}{\lambda_2^2}$: a change of metric counts as large when it is large *relative to* g.

Watch out: Other choices are possible

The most general ultralocal (no derivatives) metric on Met that is invariant under diffeomorphisms is the **DeWitt metric** $\int \sqrt{g}\,\big(\mathrm{tr}(g^{-1}Xg^{-1}Y) + u\ \mathrm{tr}(g^{-1}X)\,\mathrm{tr}(g^{-1}Y)\big)$, with a constant $u > -\frac{1}{2}$. Nash takes $u = 0$. Since we will divide out the trace directions anyway (they are the Weyl directions), the choice of u only changes the normalisation of the Weyl-group volume, not the final integrand.

Worked example 9.29: Diffeomorphisms are isometries of Met

A diffeomorphism α acts on Met by $g \mapsto \alpha^*g$, and on tangent vectors by $X \mapsto \alpha^*X$ (a tangent vector is a symmetric tensor, and pulls back the same way). With J the Jacobian and $y = \alpha(x)$, in matrix notation $\alpha^*g = J^{\mathsf{T}}gJ$ and $\alpha^*X = J^{\mathsf{T}}XJ$ (evaluated at y). Then

$$\langle \alpha^*X, \alpha^*Y\rangle_{\alpha^*g} \overset{(1)}{=} \int \mathrm{d}^2x\ |\det J|\,\sqrt{g}\ \mathrm{tr}\big(J^{-1}g^{-1}J^{-\mathsf{T}}J^{\mathsf{T}}XJ\,J^{-1}g^{-1}J^{-\mathsf{T}}J^{\mathsf{T}}YJ\big)$$
$$\overset{(2)}{=} \int \mathrm{d}^2x\ |\det J|\,\sqrt{g}\ \mathrm{tr}\big(J^{-1}g^{-1}Xg^{-1}YJ\big)$$
$$\overset{(3)}{=} \int \mathrm{d}^2y\,\sqrt{g}\ \mathrm{tr}\big(g^{-1}Xg^{-1}Y\big) = \langle X, Y\rangle_g.$$

(1) Insert $\sqrt{\det(J^{\mathsf{T}}gJ)} = |\det J|\,\sqrt{g}$ and $(J^{\mathsf{T}}gJ)^{-1} = J^{-1}g^{-1}J^{-\mathsf{T}}$.
(2) $J^{-\mathsf{T}}J^{\mathsf{T}} = \mathbf{1}$ and $JJ^{-1} = \mathbf{1}$ inside the product.
(3) Cyclicity of the trace removes $J^{-1}\cdots J$; then change variables $\mathrm{d}^2y = |\det J|\,\mathrm{d}^2x$.

This is the property Nash uses in §3 (his footnote 3).
Weyl rescalings, by contrast, are *not* isometries: under $g \mapsto fg$, $X \mapsto fX$, the integrand picks up $f \cdot f^{-2} \cdot f^2 = f$ (from $\sqrt{g}$, the two inverse metrics, and X, Y). So the measure on Met "knows" about the conformal factor, one more source of the conformal anomaly.

A toy model: one point of Σ. To see the geometry of (9.41), forget the integral over Σ and look at the three-dimensional space Pos(2) of positive symmetric 2×2 matrices with the metric $ds^2 = \operatorname{tr}(g^{-1}dg\, g^{-1}dg)$. Every g can be written uniquely as a scale times a unimodular matrix, $g = \lambda\hat{g}$ with $\lambda = \sqrt{\det g}$ and $\det\hat{g} = 1$. This is the pointwise version of "conformal factor times conformal class" (Figure 9.14).

Worked example 9.30: $\mathrm{Pos}(2) = \mathbb{R}_+ \times$ (**hyperbolic plane**)

First, the scale separates:

$$\begin{aligned}\operatorname{tr}\big(g^{-1}dg\, g^{-1}dg\big) &\overset{(1)}{=} \operatorname{tr}\Big[\Big(\frac{d\lambda}{\lambda}\mathbf{1} + \hat{g}^{-1}d\hat{g}\Big)^2\Big]\\ &\overset{(2)}{=} 2\Big(\frac{d\lambda}{\lambda}\Big)^2 + 2\frac{d\lambda}{\lambda}\operatorname{tr}\big(\hat{g}^{-1}d\hat{g}\big) + \operatorname{tr}\big(\hat{g}^{-1}d\hat{g}\,\hat{g}^{-1}d\hat{g}\big)\\ &\overset{(3)}{=} 2\Big(\frac{d\lambda}{\lambda}\Big)^2 + \operatorname{tr}\big(\hat{g}^{-1}d\hat{g}\,\hat{g}^{-1}d\hat{g}\big).\end{aligned}$$

(1) $g^{-1}dg = \lambda^{-1}\hat{g}^{-1}(d\lambda\,\hat{g} + \lambda\, d\hat{g})$.
(2) Expand the square; $\operatorname{tr}\mathbf{1} = 2$.
(3) $\operatorname{tr}(\hat{g}^{-1}d\hat{g}) = d\log\det\hat{g} = 0$ (Jacobi's formula, $\det\hat{g} = 1$).

So the scale direction is orthogonal to the unimodular directions. Next, parametrise the unimodular matrices by a point $\tau = \tau_1 + i\tau_2$ of the upper half-plane:

$$\hat{g}_\tau = \frac{1}{\tau_2}\begin{pmatrix}1 & \tau_1\\ \tau_1 & |\tau|^2\end{pmatrix}, \qquad \det\hat{g}_\tau = \frac{|\tau|^2 - \tau_1^2}{\tau_2^2} = 1. \tag{9.42}$$

This is the matrix of the quadratic form $\frac{1}{\tau_2}|dx + \tau\, dy|^2$. A direct computation (differentiate the entries, multiply the 2×2 matrices, take traces) gives

$$\operatorname{tr}\big(\hat{g}^{-1}\partial_{\tau_i}\hat{g}\;\hat{g}^{-1}\partial_{\tau_j}\hat{g}\big) = \frac{2\,\delta_{ij}}{\tau_2^2}, \qquad \text{i.e.} \qquad \operatorname{tr}\big(\hat{g}^{-1}d\hat{g}\,\hat{g}^{-1}d\hat{g}\big) = \frac{2(d\tau_1^2 + d\tau_2^2)}{\tau_2^2}. \tag{9.43}$$

For instance, at $\tau = i$ we have $\hat{g} = \mathbf{1}$, $\partial_{\tau_1}\hat{g} = \left(\begin{smallmatrix}0&1\\1&0\end{smallmatrix}\right)$, $\partial_{\tau_2}\hat{g} = \left(\begin{smallmatrix}-1&0\\0&1\end{smallmatrix}\right)$, and both have $\operatorname{tr}(X^2) = 2$ while $\operatorname{tr}(XY) = 0$, as claimed. The unimodular part is the **hyperbolic plane** with (twice) its Poincaré metric. The same metric will reappear as the Weil–Petersson metric on the moduli space of the torus (Example 9.42).

A metric on vector fields. We will also need to measure the size of an infinitesimal diffeomorphism, that is, of a vector field. The space $\mathrm{vect}(\Sigma)$ of smooth vector fields gets the L^2 metric (Nash's (9.12))

$$\langle V, W\rangle_{\mathrm{vect}} = \int_\Sigma d^2x\,\sqrt{g}\,\langle V(x), W(x)\rangle_{g(x)} = \int_\Sigma d^2x\,\sqrt{g}\; g_{ab}V^aW^b. \tag{9.44}$$

It depends on g too; we suppress this in the notation.

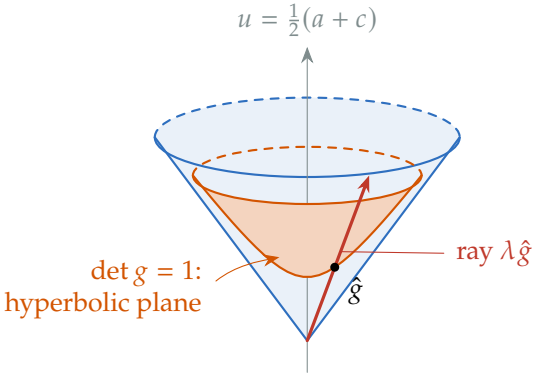


Figure 9.14: The space of positive 2×2 matrices $\begin{pmatrix} a & b \\ b & c \end{pmatrix}$ is an open cone (the horizontal directions are $\frac{1}{2}(a-c)$ and b). The rays through the origin (red) are the Weyl orbits, one point at a time. The surface $\det g = 1$ (orange) meets every ray once: it is the space of conformal structures at a point, and with the metric (9.41) it is a hyperbolic plane, orthogonal to the rays.

9.2.5 Infinitesimal diffeomorphisms: the Lie derivative of the metric

The Lie algebra of $\mathrm{Diff}_0(\Sigma)$ is $\mathrm{vect}(\Sigma)$. A vector field V generates the one-parameter family $\alpha_t = \exp[tV]$ of diffeomorphisms obtained by flowing along V for time t; for small t, $\alpha_t(x)^a = x^a + tV^a(x) + O(t^2)$. The metric moves along the orbit as $\alpha_t^* g = g + t\,\mathcal{L}_V g + O(t^2)$, which defines the **Lie derivative** $\mathcal{L}_V g \in T_g\mathrm{Met}$ (Nash's (9.13)). So $\mathcal{L}_V g$ is the tangent vector to the Diff_0 orbit through g in the direction V.

Worked example 9.31: Computing $\mathcal{L}_V g$ from the pull-back

Use (9.9) with α_t, whose Jacobian is $\partial_a \alpha_t^c = \delta_a^c + t\,\partial_a V^c$:

$$\begin{aligned}(\alpha_t^* g)_{ab}(x) &\overset{(1)}{=} (\delta_a^c + t\,\partial_a V^c)(\delta_b^d + t\,\partial_b V^d)\, g_{cd}(x+tV) \\ &\overset{(2)}{=} g_{ab} + t\big(V^c \partial_c g_{ab} + g_{cb}\,\partial_a V^c + g_{ac}\,\partial_b V^c\big) + O(t^2).\end{aligned}$$

(1) The pull-back formula with $\alpha_t(x) = x + tV + O(t^2)$.
(2) Taylor-expand $g_{cd}(x+tV) = g_{cd} + tV^e \partial_e g_{cd}$ and keep terms of order t.

Hence

$$(\mathcal{L}_V g)_{ab} = V^c \partial_c g_{ab} + g_{cb}\,\partial_a V^c + g_{ac}\,\partial_b V^c. \tag{9.45}$$

Nash's footnote 2 writes this covariantly. With the Levi-Civita connection, $\nabla_a V_b = \partial_a V_b - \Gamma^c_{ab} V_c$, $V_b = g_{bc}V^c$ and $\Gamma^c_{ab} g_{cd} = \frac{1}{2}(\partial_a g_{db} + \partial_b g_{da} - \partial_d g_{ab})$:

$$\begin{aligned}\nabla_a V_b + \nabla_b V_a &\overset{(1)}{=} \partial_a(g_{bc}V^c) + \partial_b(g_{ac}V^c) - 2\Gamma^d_{ab} g_{dc} V^c \\ &\overset{(2)}{=} g_{bc}\partial_a V^c + g_{ac}\partial_b V^c + V^c(\partial_a g_{bc} + \partial_b g_{ac}) - V^c(\partial_a g_{cb} + \partial_b g_{ca} - \partial_c g_{ab}) \\ &\overset{(3)}{=} g_{bc}\partial_a V^c + g_{ac}\partial_b V^c + V^c \partial_c g_{ab} = (\mathcal{L}_V g)_{ab}.\end{aligned}$$

(1) Definition of ∇; the Christoffel symbol is symmetric in a, b, so the two connection terms are equal.
(2) Product rule in the first two terms; the Christoffel formula in the last.
(3) The derivatives of g cancel in pairs except $\partial_c g_{ab}$; compare (9.45).

So $\mathcal{L}_V g_{ab} = \nabla_a V_b + \nabla_b V_a$, and its trace is $g^{ab}\mathcal{L}_V g_{ab} = 2\nabla_a V^a = 2\,\nabla \cdot V$.

Working definition 9.7 (Killing and conformal Killing vectors)

A vector field V is a **Killing vector** of g if $\mathcal{L}_V g = 0$ [$\mathcal{L}_V g$ is the first-order change of g when every point is moved a small distance along V]: its flow [the map that moves every point for a time t along V] is a symmetry (isometry) of g. It is a **conformal Killing vector** (CKV) if $\mathcal{L}_V g = f\, g$ for some function f (necessarily $f = \nabla \cdot V$): its flow changes g only by a Weyl rescaling [multiplication by a positive function, $g \to e^{2\omega} g$]. *Test:* compute (9.45) and check whether the traceless part $\mathcal{L}_V g - \frac{1}{2}\operatorname{tr}(\mathcal{L}_V g)\, g$ vanishes.

Worked example 9.32: Killing and conformal Killing vectors of the flat plane

For $g = \delta$, (9.45) is just $(\mathcal{L}_V g)_{ab} = \partial_a V_b + \partial_b V_a$.

- Translations $V = \partial_x$: $\mathcal{L}_V \delta = 0$. Killing.
- Rotation $V = -y\,\partial_x + x\,\partial_y$: $\partial_x V_x = 0$, $\partial_y V_y = 0$, $\partial_x V_y + \partial_y V_x = 1 - 1 = 0$. Killing.
- Dilation $V = x\,\partial_x + y\,\partial_y$: $\mathcal{L}_V \delta = 2\delta$. Conformal Killing with $f = 2 = \nabla \cdot V$, not Killing.
- Shear $V = y\,\partial_x$: $\mathcal{L}_V \delta = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, which is traceless and non-zero. Not conformal Killing.

The complex coordinate makes the pattern obvious. Write a vector field as $V = V^z \partial_z + V^{\bar z} \partial_{\bar z}$ with $V^{\bar z} = \overline{V^z}$ (for real $V = V^x \partial_x + V^y \partial_y$, $V^z = V^x + i V^y$), and a conformal metric as $g = \rho\, \mathrm{d}z\, \mathrm{d}\bar z$ (symmetric product, so $\rho\, \mathrm{d}z\, \mathrm{d}\bar z = \rho(\mathrm{d}x^2 + \mathrm{d}y^2)$).

Worked example 9.33: $\mathcal{L}_V g$ in complex coordinates

The Lie derivative obeys the Leibniz rule and commutes with d, and $\mathcal{L}_V \mathrm{d}z = \mathrm{d}(V^z)$:

$$\begin{aligned} \mathcal{L}_V\big(\rho\, \mathrm{d}z\, \mathrm{d}\bar z\big) &\overset{(1)}{=} V(\rho)\, \mathrm{d}z\, \mathrm{d}\bar z + \rho\, \mathrm{d}V^z\, \mathrm{d}\bar z + \rho\, \mathrm{d}z\, \mathrm{d}V^{\bar z} \\ &\overset{(2)}{=} \big(V(\rho) + \rho\, \partial_z V^z + \rho\, \partial_{\bar z} V^{\bar z}\big) \mathrm{d}z\, \mathrm{d}\bar z + \rho\, \partial_{\bar z} V^z\, \mathrm{d}\bar z^2 + \rho\, \partial_z V^{\bar z}\, \mathrm{d}z^2. \end{aligned}$$

(1) Leibniz rule; $\mathcal{L}_V \rho = V(\rho)$ for a function, and $\mathcal{L}_V \mathrm{d}z = \mathrm{d}\, \iota_V \mathrm{d}z = \mathrm{d}V^z$ (Cartan's formula, Chapter T1).
(2) Expand $\mathrm{d}V^z = \partial_z V^z\, \mathrm{d}z + \partial_{\bar z} V^z\, \mathrm{d}\bar z$ and similarly for $V^{\bar z}$, then collect terms.

The $\mathrm{d}z\, \mathrm{d}\bar z$ term is a multiple of g: the Weyl part. The $\mathrm{d}z^2$ and $\mathrm{d}\bar z^2$ terms are the traceless part. So

$$V \text{ is a conformal Killing vector} \iff \partial_{\bar z} V^z = 0 \iff V^z \text{ is holomorphic.}$$

In the flat case: ∂_x has $V^z = 1$, the rotation has $V^z = iz$, the dilation $V^z = z$ (all holomorphic), and the shear $V^z = y = \frac{z - \bar z}{2i}$ (not holomorphic).

Worked example 9.34: CKVs on the sphere, the torus, and higher genus

Whether V^z is holomorphic does not depend on the conformal factor ρ, so we may use any conformal coordinate.
Sphere. Cover $S^2 = \mathbb{C} \cup \{\infty\}$ by z and $w = 1/z$. A holomorphic $V^z(z)$ on $\mathbb{C}$ must also be holomorphic at ∞ in the coordinate w:

$$V^w \overset{(1)}{=} \frac{\mathrm{d}w}{\mathrm{d}z}\, V^z = -\frac{1}{z^2} V^z(z)$$

$$\overset{(2)}{=} -w^2\, V^z(1/w).$$

(1) Components of a vector transform with the Jacobian.
(2) Substitute $z = 1/w$.

If $V^z = \sum_k a_k z^k$, then $V^w = -\sum_k a_k w^{2-k}$, which is regular at $w = 0$ iff $a_k = 0$ for $k > 2$. So $V^z = a_0 + a_1 z + a_2 z^2$: *three complex, six real* CKVs. They generate the Möbius transformations $z \mapsto \frac{az+b}{cz+d}$, the group PSL(2, $\mathbb{C}$) of Nash's footnote (p. 165). Three of the six real ones are Killing vectors of the round metric (the rotations SO(3): $V^z = iz$ and $V^z = a - \bar{a}z^2$).

Torus. A holomorphic function on a compact surface is constant (Liouville), so V^z = const: *one complex, two real* CKVs, the translations of $\mathbb{C}/\Lambda$. They are Killing vectors of the flat metric.

Genus $p \geq 2$. There are none. We prove this in Example 9.39 with a curvature argument; in the holomorphic language, $V^z\partial_z$ is a holomorphic section of the tangent bundle K_Σ^{-1}, whose degree $2 - 2p$ is negative, and a line bundle of negative degree has no holomorphic sections (Chapter 5).

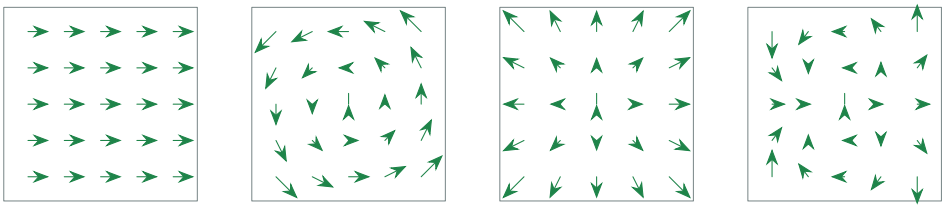


Figure 9.15: The four kinds of conformal Killing vector of the Riemann sphere, drawn in the plane $z = x + iy$ (arrows show the real vector field with $V^x + iV^y = V^z$). Real and imaginary multiples of these give all six. Translations and rotations are Killing for the flat metric; dilations and special conformal transformations only rescale it. On the round sphere the rotations iz and $1 - z^2$, $i(1 + z^2)$ are Killing. None of these fields exists on a surface of genus ≥ 2.

Physical picture: Why CKVs are a problem

A conformal Killing vector is a gauge transformation (a diffeomorphism) that acts on the metric *like another gauge transformation* (a Weyl rescaling). So the combined gauge group $\mathrm{Diff}_0 \ltimes C^\infty_+$ does not act freely: a one-parameter subgroup (the CKV flow combined with the compensating Weyl rescaling) leaves g fixed. In the Faddeev–Popov Jacobian this shows up as a zero mode, exactly like $r = 0$ in Example 9.27.

9.2.6 Splitting off the Weyl directions

We now have two kinds of gauge motion through a point $g \in \mathrm{Met}(\Sigma)$. A Weyl rescaling $g \mapsto e^{2\omega} g$ moves g in the direction $2\omega\, g$. A diffeomorphism generated by V moves g in the direction $\mathcal{L}_V g$. Both directions live in the same space $T_g\mathrm{Met}(\Sigma)$ of symmetric tensors, and we want to know how they sit relative to each other and what is left over. The tool is the metric $\langle\ ,\ \rangle_{\mathrm{met}}$: with a metric we can talk about orthogonal complements.

Working definition 9.8 (The conformal subspace C_g)

$C_g \subset T_g\mathrm{Met}(\Sigma)$ [the small changes $g \to g + \epsilon X$ of the metric, X a symmetric tensor field] is the set of tangent vectors of the form $\alpha\, g$, with α an arbitrary smooth function on Σ. It is the tangent space at g to the Weyl orbit $\{f g : f \in C^\infty_+(\Sigma)\}$ [f runs over the smooth functions with $f > 0$ everywhere]: differentiating $f_t = e^{t\alpha}$ at $t = 0$ gives $\frac{\mathrm{d}}{\mathrm{d}t}(f_t g) = \alpha g$. *Test:* a tangent vector X is in C_g iff at every point the matrix $g^{-1}X$ is a multiple of the identity.

The metric gives an orthogonal decomposition (Nash's (9.14))

$$T_g\mathrm{Met}(\Sigma) = C_g \oplus C_g^{\perp}. \tag{9.46}$$

What is $C_g^{\perp}$? A vector h is orthogonal to all of C_g iff $\langle \alpha g, h\rangle_{\mathrm{met}} = 0$ for every function α. We compute this inner product exactly as Nash does in (9.16):

$$\begin{aligned}
\langle \alpha g, h\rangle_{\mathrm{met}} &\overset{(1)}{=} \int_\Sigma \mathrm{d}^2x\, \sqrt{g}\; g^{a_1a_2} g^{b_1b_2}\, \alpha\, g_{a_1b_1}\, h_{a_2b_2} \\
&\overset{(2)}{=} \int_\Sigma \mathrm{d}^2x\, \sqrt{g}\; \alpha\; \delta^{a_2}_{b_1}\, g^{b_1b_2}\, h_{a_2b_2} \\
&\overset{(3)}{=} \int_\Sigma \mathrm{d}^2x\, \sqrt{g}\; \alpha\; g^{a_2b_2} h_{a_2b_2} \\
&\overset{(4)}{=} \int_\Sigma \mathrm{d}^2x\, \sqrt{g}\; \alpha\; \mathrm{tr}\, h.
\end{aligned} \tag{9.47}$$

(1) Definition (9.41) with $X = \alpha g$ and $Y = h$, written with Nash's index names.
(2) $g^{a_1a_2} g_{a_1b_1} = \delta^{a_2}_{b_1}$: the inverse metric undoes the metric.
(3) Sum over b_1 with the Kronecker delta.
(4) Define $\mathrm{tr}\, h = g^{ab} h_{ab}$, the trace of the matrix $g^{-1}h$.

Since α is arbitrary, (9.47) vanishes for all α iff $\mathrm{tr}\, h = 0$ at every point. (Take $\alpha = \mathrm{tr}\, h$: then the integral is $\int \sqrt{g}\,(\mathrm{tr}\, h)^2$, which is zero only if $\mathrm{tr}\, h \equiv 0$.) So

$$C_g^{\perp} = \{h \in T_g\mathrm{Met} :\ \mathrm{tr}\, h = g^{ab} h_{ab} = 0\} = \text{traceless symmetric tensors}. \tag{9.48}$$

In two dimensions a traceless symmetric 2×2 matrix has two independent entries. So, pointwise, C_g is one-dimensional and $C_g^{\perp}$ is two-dimensional: $1 + 2 = 3$ components of a metric. This is the infinitesimal version of the cone picture Figure 9.14: the ray direction is C_g, the hyperbolic-plane directions are $C_g^{\perp}$.

Splitting a Lie derivative. Apply this to the diffeomorphism direction $\mathcal{L}_V g$. Every symmetric tensor splits uniquely as "trace part plus traceless part":

$$\mathcal{L}_V g = \underbrace{\frac{\mathrm{tr}(\mathcal{L}_V g)}{2}\, g}_{\alpha g \in C_g} + \underbrace{\left\{\mathcal{L}_V g - \frac{\mathrm{tr}(\mathcal{L}_V g)}{2}\, g\right\}}_{h \in C_g^{\perp}} \tag{9.49}$$

(Nash's (9.17)). The first piece is in C_g by definition. The second is traceless because $\mathrm{tr}\, g = g^{ab} g_{ab} = \delta^a_a = 2$, so $\mathrm{tr}\, h = \mathrm{tr}(\mathcal{L}_V g) - \frac{\mathrm{tr}(\mathcal{L}_V g)}{2} \cdot 2 = 0$. This is Nash's (9.18):

$$h = \mathcal{L}_V g - \tfrac{1}{2}\,\mathrm{tr}(\mathcal{L}_V g)\, g, \qquad \alpha = \tfrac{1}{2}\,\mathrm{tr}(\mathcal{L}_V g). \tag{9.50}$$

With the covariant formula of Example 9.31, $\mathcal{L}_V g_{ab} = \nabla_a V_b + \nabla_b V_a$ and $\mathrm{tr}\, \mathcal{L}_V g = 2\nabla_c V^c$, so (Nash's footnote 2)

$$h_{ab} = \nabla_a V_b + \nabla_b V_a - (\nabla_c V^c)\, g_{ab}, \qquad \alpha = \nabla_c V^c. \tag{9.51}$$

Worked example 9.35: Splitting three flat-space vector fields

Take $g = \delta$ on $\mathbb{R}^2$ (or on a flat torus). Then $\nabla = \partial$ and (9.51) reads, in matrix form with $V = (u, v)$,

$$\alpha = u_x + v_y, \qquad h = \begin{pmatrix} u_x - v_y & u_y + v_x \\ u_y + v_x & -(u_x - v_y) \end{pmatrix}. \tag{9.52}$$

- Dilation $V = (x, y)$: $\alpha = 2$, $h = 0$. The flow only rescales the metric.
- Rotation $V = (-y, x)$: $\alpha = 0$, $h = 0$. The flow does not change the metric at all.
- Shear $V = (y, 0)$: $\alpha = 0$, $h = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$. The flow changes the *shape* of the metric, not its size.

The shear is the interesting one: it produces a genuinely non-conformal change. Whether such a change is "gauge" (produced by some diffeomorphism) or "physical" (a change of modulus) is exactly the question we answer next.

9.2.7 The operator P_1

The traceless part h of $\mathcal{L}_V g$ depends linearly on V, and it involves one derivative of V. So it is a first-order linear differential operator from vector fields to traceless symmetric tensors.

Working definition 9.9 (The operator $A_g = P_1$)

(Nash's (9.19).)

$$P_1 = A_g : \ \mathrm{vect}(\Sigma) \longrightarrow C_g^{\perp}, \qquad (P_1 V)_{ab} = \nabla_a V_b + \nabla_b V_a - (\nabla_c V^c)\, g_{ab}. \tag{9.53}$$

[vect(Σ) is the smooth vector fields V on Σ. $C_g^{\perp}$ is the traceless symmetric tensors, orthogonal to every αg] It measures how much the flow of V changes the *conformal class* of g [g up to multiplication by a positive function]. *How to compute it:* take $\mathcal{L}_V g$ from (9.45) and throw away its trace. Physicists call it P_1; Nash calls it A_g to stress that it depends on g. Its kernel [the V with $P_1 V = 0$] is the space of conformal Killing vectors.

P_1 is the Cauchy–Riemann operator. In flat coordinates, (9.52) says that $P_1 V = 0$ iff

$$u_x = v_y, \qquad u_y = -v_x, \tag{9.54}$$

the Cauchy–Riemann equations for $u + iv$. In the complex coordinate of Example 9.33, with $g = \rho\, \mathrm{d}z\, \mathrm{d}\bar{z}$ (so $g_{z\bar{z}} = \rho/2$), a traceless symmetric tensor has only the components h_{zz} and $h_{\bar{z}\bar{z}} = \overline{h_{zz}}$, because $h_{z\bar{z}} \propto \mathrm{tr}\, h = 0$. Reading off the coefficients of $\mathrm{d}z^2$ and $\mathrm{d}\bar{z}^2$ in Example 9.33,

$$(P_1 V)_{\bar{z}\bar{z}} = \rho\, \partial_{\bar{z}} V^z, \qquad (P_1 V)_{zz} = \rho\, \partial_z V^{\bar{z}} = \overline{(P_1 V)_{\bar{z}\bar{z}}}. \tag{9.55}$$

So, up to the harmless factor ρ, P_1 *is* the operator $\partial_{\bar{z}}$ acting on the component V^z. We can check that the two descriptions agree in the flat case $\rho = 1$:

$$\begin{aligned} (P_1 V)_{\bar{z}\bar{z}} &\overset{(1)}{=} \tfrac{1}{2}(\partial_x + i\partial_y)(u + iv) \\ &\overset{(2)}{=} \tfrac{1}{2}\big[(u_x - v_y) + i(u_y + v_x)\big] \\ &\overset{(3)}{=} \tfrac{1}{2}\big(h_{11} + i h_{12}\big) \overset{(4)}{=} h(\partial_{\bar{z}}, \partial_{\bar{z}}). \end{aligned}$$

(1) $\partial_{\bar{z}} = \frac{1}{2}(\partial_x + i\partial_y)$ and $V^z = u + iv$.

(2) Multiply out and collect real and imaginary parts.
(3) Compare with (9.52).
(4) $h_{\bar z\bar z} = h(\partial_{\bar z}, \partial_{\bar z}) = \frac{1}{4}(h_{11} + 2ih_{12} - h_{22})$, and $h_{22} = -h_{11}$.

P_1 is elliptic. The principal symbol of a first-order operator is obtained by replacing each derivative ∂_a by $i\xi_a$ and keeping the top-order terms (Chapter 2). For (9.52), as a map from (u, v) to the two independent entries (h_{11}, h_{12}) of a traceless tensor,

$$\sigma_{P_1}(\xi) = i\begin{pmatrix} \xi_1 & -\xi_2 \\ \xi_2 & \xi_1 \end{pmatrix}, \qquad \det \sigma_{P_1}(\xi) = i^2(\xi_1^2 + \xi_2^2) = -\,|\xi|^2\,. \tag{9.56}$$

The matrix is i times the matrix of multiplication by the complex number $\xi_1 + i\xi_2$, and it is invertible for every $\xi \neq 0$. That is what **elliptic** means: P_1 maps a rank-2 bundle (tangent vectors) to a rank-2 bundle (traceless symmetric tensors) and its symbol is an isomorphism off the zero section. (On a curved surface the lower-order Christoffel terms change nothing, because the symbol only sees the highest derivatives.)

Watch out: Why the trace must be removed first

The full map $V \mapsto \mathcal{L}_V g$ goes from a rank-2 bundle to the rank-3 bundle of all symmetric tensors. Its symbol $\xi \otimes V + V \otimes \xi$ is injective but cannot be surjective ($2 < 3$), so this map is *not* elliptic, and its "cokernel" is infinite-dimensional (it contains all of C_g). Removing the trace is what makes the counting work. This is special to two dimensions: in dimension n the traceless symmetric tensors have rank $\frac{n(n+1)}{2} - 1$, which equals n only for $n = 2$.

The adjoint. Nash's next step uses the adjoint $A_g^* = P_1^\dagger$, defined by $\langle h, P_1 V\rangle_{\text{met}} = \langle P_1^\dagger h, V\rangle_{\text{vect}}$ for all traceless h and all V. We compute it by integrating by parts.

$$\begin{aligned}
\langle h, P_1 V\rangle_{\text{met}} &\overset{(1)}{=} \int_\Sigma \mathrm{d}^2x\,\sqrt{g}\; h^{ab}\big(\nabla_a V_b + \nabla_b V_a - g_{ab}\nabla_c V^c\big) \\
&\overset{(2)}{=} 2\int_\Sigma \mathrm{d}^2x\,\sqrt{g}\; h^{ab}\,\nabla_a V_b \\
&\overset{(3)}{=} -2\int_\Sigma \mathrm{d}^2x\,\sqrt{g}\;(\nabla_a h^{ab})\,V_b \\
&\overset{(4)}{=} \Big\langle -2\nabla^a h_{a\,\cdot}\,,\, V\Big\rangle_{\text{vect}}.
\end{aligned} \tag{9.57}$$

(1) Definition (9.41); indices of h raised with g.
(2) h^{ab} is symmetric, so the first two terms are equal; h is traceless, so $h^{ab}g_{ab} = 0$ kills the third.
(3) Integrate by parts. $\sqrt{g}\,\nabla_a(h^{ab}V_b)$ is the divergence of a vector field, whose integral over the closed surface Σ vanishes by Stokes' theorem (Chapter T2).
(4) Definition (9.44) of $\langle\ ,\ \rangle_{\text{vect}}$.

So

$$(P_1^\dagger h)_b = -2\,\nabla^a h_{ab}, \qquad \text{and in complex form} \qquad (P_1^\dagger h)_z = -2g^{z\bar z}\,\partial_{\bar z} h_{zz} = -\frac{4}{\rho}\,\partial_{\bar z} h_{zz}. \tag{9.58}$$

For the complex form: $\nabla^a h_{az} = g^{z\bar z}\nabla_{\bar z} h_{zz} + g^{\bar z z}\nabla_z h_{\bar z z}$; the second term vanishes because $h_{\bar z z} = 0$ and the only non-zero Christoffel symbols of a conformal metric are $\Gamma^z_{zz} = \partial_z \log\rho$ and $\Gamma^{\bar z}_{\bar z\bar z} = \partial_{\bar z}\log\rho$; for the same reason $\nabla_{\bar z} h_{zz} = \partial_{\bar z} h_{zz}$.

Key idea: The kernel of $P_1^\dagger$: holomorphic quadratic differentials

A traceless h is in $\operatorname{Ker} P_1^\dagger$ iff it is divergence-free, $\nabla^a h_{ab} = 0$, iff its component h_{zz} is holomorphic, $\partial_{\bar z} h_{zz} = 0$. The object $\varphi = h_{zz}\,\mathrm{d}z^2$, which transforms under $z \mapsto w$ as $h_{ww} = h_{zz}(\mathrm{d}z/\mathrm{d}w)^2$, is a **holomorphic quadratic differential**: a holomorphic section of K^2, where K is the canonical bundle (Chapter 5). Nash: "it is easy to see that $\operatorname{Ker} A_g^*$ can be identified with the space of quadratic differentials" (p. 265). The identification is (9.58).

Dictionary: the operator P_1

Nash	physics / complex geometry
A_g	P_1, the "conformal Killing operator"
A_g^*	$P_1^\dagger = -2\,\mathrm{div}$; on h_{zz} it is $\propto \partial_{\bar z}$
$\operatorname{Ker} A_g$	conformal Killing vectors; holomorphic vector fields $V^z\partial_z$
$\operatorname{Ker} A_g^*$	holomorphic quadratic differentials $h_{zz}\mathrm{d}z^2$
$A_g^* A_g$	$-2(\nabla^2 + K)$ on vector fields; ghost kinetic operator
$\operatorname{Ker} A_g$, $\operatorname{Ker} A_g^*$	in the path integral: c-ghost and b-ghost zero modes

Worked example 9.36: No quadratic differentials on the sphere, one on the torus

Sphere. Use z on $\mathbb{C}$ and $w = 1/z$ near ∞. If $h_{zz}(z) = \sum_k b_k z^k$ is holomorphic on $\mathbb{C}$, then near $w = 0$

$$h_{ww} \overset{(1)}{=} h_{zz}\left(\frac{\mathrm{d}z}{\mathrm{d}w}\right)^2 \overset{(2)}{=} h_{zz}(1/w)\,\frac{1}{w^4} = \sum_{k\geq 0} b_k\, w^{-k-4}.$$

(1) A quadratic differential has two lower indices, each transforming with one factor of the Jacobian.
(2) $z = 1/w$, so $\mathrm{d}z/\mathrm{d}w = -1/w^2$, whose square is w^{-4}.

Every term has a negative power of w, so h_{ww} is regular at $w = 0$ only if all $b_k = 0$. The sphere has *no* holomorphic quadratic differentials, so $\operatorname{Ker} P_1^\dagger = 0$: no moduli, consistent with $\dim \mathcal{M}_0 = 0$. Compare the vector field count of Example 9.34: there the Jacobian came with the opposite power, w^{+2}, and three coefficients survived.
Torus $\mathbb{C}/\Lambda$. A holomorphic function on a compact surface is constant, so $h_{zz} = b$ and $\varphi = b\,\mathrm{d}z^2$: one complex, two real parameters, matching $\dim_{\mathbb{R}} \mathcal{M}_1 = 2$.

9.2.8 The orthogonal decomposition of $T_g\mathrm{Met}$

The last ingredient is a fact from linear algebra.

Key idea: Range plus cokernel (Nash's (9.20))

For a linear map $A : V \to W$ between inner product spaces, with adjoint A^*,

$$W = \operatorname{Range} A \;\oplus\; \operatorname{Ker} A^*, \qquad \text{orthogonally.} \tag{9.59}$$

Why: $w \perp \operatorname{Range} A$ iff $\langle w, Av\rangle = 0$ for all v iff $\langle A^* w, v\rangle = 0$ for all v iff $A^* w = 0$. So $\operatorname{Ker} A^*$ is exactly the orthogonal complement of the range.

Worked example 9.37: A 3×2 matrix

Let $A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 0 & 0 \end{pmatrix} : \mathbb{R}^2 \to \mathbb{R}^3$. Its range is the line spanned by $(1,1,0)$. Its adjoint is $A^{\mathsf{T}} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \end{pmatrix}$, whose kernel is $\{(a,b,c) : a+b=0\}$, the plane spanned by $(1,-1,0)$ and $(0,0,1)$. The line and the plane are orthogonal and together span $\mathbb{R}^3$: $1+2=3$. Note also that $\operatorname{Ker} A$ is the line spanned by $(1,-1)$: A is neither injective nor surjective, and both defects are visible separately. The **index** of A is $\dim \operatorname{Ker} A - \dim \operatorname{Ker} A^* = 1-2 = -1 = \dim V - \dim W$.

In infinite dimensions (9.59) can fail, because a range need not be closed. For elliptic operators on compact manifolds it holds: the range is closed and both kernels are finite-dimensional (Chapter 2). Applying it with $A = P_1$ and $W = C_g^{\perp}$, and combining with (9.46), gives Nash's (9.21):

Key idea: The decomposition of the tangent space to the space of metrics

$$T_g \mathrm{Met}(\Sigma) = \underbrace{C_g}_{\text{Weyl}} \oplus \underbrace{\operatorname{Range} P_1}_{\text{diffeomorphisms}} \oplus \underbrace{\operatorname{Ker} P_1^{\dagger}}_{\text{moduli}}. \tag{9.60}$$

Every infinitesimal change of metric is, uniquely, a Weyl rescaling plus the traceless part of an infinitesimal diffeomorphism plus a holomorphic quadratic differential. The first two are gauge; only the third changes the point of Teichmüller space.

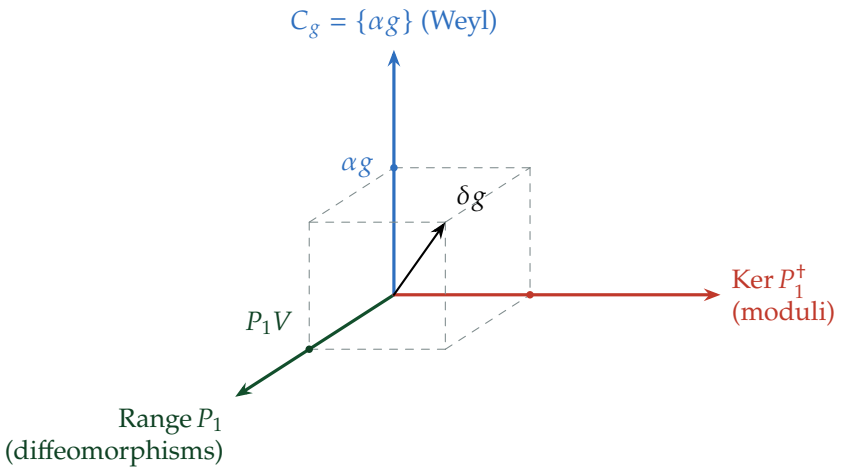


Figure 9.16: The three-way orthogonal splitting (9.60) of $T_g\mathrm{Met}(\Sigma)$. The picture is schematic: the blue and green axes each stand for an infinite-dimensional space, while the red direction $\operatorname{Ker} P_1^{\dagger}$ has finite dimension $6p-6$ (for $p \geq 2$). A general change of metric δg is resolved into a Weyl part (blue), a diffeomorphism part (green) and a modulus (red). Only the red component survives the quotient by the gauge group.

Worked example 9.38: The decomposition on the flat torus, mode by mode

Take $\Sigma = \mathbb{R}^2/\mathbb{Z}^2$ with $g = \delta$ ($\tau = i$, the square torus). Any traceless h is described by the complex function $h_{zz} = \frac{1}{2}(h_{11} + ih_{12})$, which we expand in Fourier modes $e^{2\pi i(nx+my)}$. By

(9.55), $(P_1V)_{\bar z\bar z} = \partial_{\bar z} V^z$, and on a Fourier mode

$$\partial_{\bar z}\, e^{2\pi i(nx+my)} \overset{(1)}{=} \tfrac{1}{2}(\partial_x + i\partial_y)\, e^{2\pi i(nx+my)} \overset{(2)}{=} \pi i\,(n+im)\, e^{2\pi i(nx+my)}.$$

(1) Definition of $\partial_{\bar z}$.
(2) $\partial_x \to 2\pi i n$, $\partial_y \to 2\pi i m$, and $\tfrac{1}{2}(2\pi i n + i\cdot 2\pi i m) = \pi i(n+im)$.

So P_1 is diagonal in Fourier space and multiplies the (n,m) mode by $\pi i(n+im)$. This number vanishes only for $(n,m)=(0,0)$. Hence:

- $\operatorname{Ker} P_1$: constant V^z, the two translations (the CKVs of the torus).
- $\operatorname{Range} P_1$: every traceless h with zero average. Given such an h, the vector field $V^z = \sum_{(n,m)\neq 0} \frac{\hat h(n,m)}{\pi i(n+im)} e^{2\pi i(nx+my)}$ solves $P_1 V = h$.
- $\operatorname{Ker} P_1^\dagger$: the constant traceless tensors, which are orthogonal to every mode with $(n,m)\neq 0$. Two real parameters, the moduli.

A concrete gauge deformation: $\delta g = \varepsilon\cos(2\pi y)\,(\mathrm{d}x^2 - \mathrm{d}y^2)$ has $h_{\bar z\bar z} = \frac{\varepsilon}{2}\cos 2\pi y$ and zero average, so it is pure gauge. Solving $\frac{i}{2}\partial_y V^z = \frac{\varepsilon}{2}\cos 2\pi y$ gives $V^z = -\frac{i\varepsilon}{2\pi}\sin 2\pi y$, i.e. $V = (0, v)$ with $v = -\frac{\varepsilon}{2\pi}\sin 2\pi y$. Then $\mathcal{L}_V\delta = 2v'\,\mathrm{d}y^2 = -2\varepsilon\cos(2\pi y)\,\mathrm{d}y^2$ and

$$\delta g = \mathcal{L}_V \delta + \varepsilon\cos(2\pi y)\,\delta,$$

a diffeomorphism plus a Weyl rescaling, as promised. By contrast the *constant* deformation $\delta g = \varepsilon(\mathrm{d}x^2 - \mathrm{d}y^2)$ would need $v' = -\varepsilon$, i.e. $v = -\varepsilon y$, which is not periodic: it is not gauge. It stretches the square torus into a rectangle, changing $\tau = i$ to $\tau \approx i(1-\varepsilon)$ (compare $\partial_{\tau_2}\hat g = \operatorname{diag}(-1,1)$ in Example 9.30).

9.2.9 How many conformal Killing vectors and moduli?

We have met $\operatorname{Ker} P_1$ (CKVs) and $\operatorname{Ker} P_1^\dagger$ (quadratic differentials) separately. Two further tools pin down their dimensions for every genus: a curvature identity that kills $\operatorname{Ker} P_1$ for $p \geq 2$, and the index theorem, which fixes the *difference* of the two dimensions.

Worked example 9.39: The Weitzenböck formula $P_1^\dagger P_1 = -2(\nabla^2 + K)$

Compose (9.53) and (9.58):

$$\begin{aligned}(P_1^\dagger P_1 V)_b &\overset{(1)}{=} -2\nabla^a\big(\nabla_a V_b + \nabla_b V_a - g_{ab}\nabla_c V^c\big)\\ &\overset{(2)}{=} -2\big(\nabla^2 V_b + \nabla^a\nabla_b V_a - \nabla_b\nabla_c V^c\big)\\ &\overset{(3)}{=} -2\big(\nabla^2 V_b + \nabla_b\nabla^a V_a + R_{bc}V^c - \nabla_b\nabla_c V^c\big)\\ &\overset{(4)}{=} -2\big(\nabla^2 V_b + K\,V_b\big).\end{aligned}$$

(1) Apply $P_1^\dagger = -2\nabla^a(\cdot)_{ab}$ to $h = P_1 V$.
(2) Distribute; $\nabla g = 0$, and $\nabla^2 = \nabla^a\nabla_a$.
(3) Commute the covariant derivatives: $[\nabla_a, \nabla_b]V^a = R_{bc}V^c$, the definition of the Ricci tensor.

(4) The divergence terms cancel, and in two dimensions the Ricci tensor is $R_{bc} = K\, g_{bc}$ with K the Gauss curvature (every 2D metric is an "Einstein metric" pointwise).

Now pair with V and integrate by parts once:

$$\|P_1 V\|^2_{\text{met}} \overset{(1)}{=} \langle V, P_1^\dagger P_1 V\rangle_{\text{vect}} \overset{(2)}{=} 2\int_\Sigma \sqrt{g}\,\nabla_a V_b\,\nabla^a V^b - 2\int_\Sigma \sqrt{g}\,K\,|V|^2 .$$

(1) Definition of the adjoint.

(2) $-\int \sqrt{g}\,V^b\nabla^2 V_b = +\int \sqrt{g}\,|\nabla V|^2$ by Stokes.

Three consequences.

- $p \geq 2$, $K = -1$: $\|P_1 V\|^2 = 2\,\|\nabla V\|^2 + 2\,\|V\|^2$, which is zero only for $V = 0$. *No CKVs* on a hyperbolic surface.
- $p = 1$, $K = 0$: $\|P_1 V\|^2 = 2\,\|\nabla V\|^2$, zero iff $\nabla V = 0$, i.e. V is covariantly constant: the two translations.
- $p = 0$, $K = +1$: the formula allows a kernel, and indeed the six CKVs of Example 9.34 satisfy $\nabla^2 V = -V$ on the unit sphere (for the three Killing vectors this is the standard identity $\nabla^2 V_b = -R_{bc}V^c$).

Note also that $P_1^\dagger P_1$ is minus twice a Laplacian plus a potential: a Laplace-type elliptic operator on vector fields, to which the zeta-function technology of Section 9.1.10 applies.

Key idea: The index of P_1 (Riemann–Roch)

$$\operatorname{index} P_1 = \dim_{\mathbb{R}} \operatorname{Ker} P_1 - \dim_{\mathbb{R}} \operatorname{Ker} P_1^\dagger = 3\chi(\Sigma) = 6 - 6p. \tag{9.61}$$

Why: by (9.55) and (9.58), P_1 is the operator $\bar\partial$ acting on sections $V^z\partial_z$ of the holomorphic tangent bundle $T^{1,0}\Sigma = K^{-1}$, and its cokernel is $H^0(K^2)$. The Riemann–Roch theorem (Chapters 4 and 5) says, for a holomorphic line bundle L,

$$\dim_{\mathbb{C}} H^0(L) - \dim_{\mathbb{C}} H^0(K\otimes L^{-1}) = \deg L - p + 1.$$

With $L = K^{-1}$, $\deg L = 2-2p$ and $K\otimes L^{-1} = K^2$: $\dim_{\mathbb{C}} H^0(K^{-1}) - \dim_{\mathbb{C}} H^0(K^2) = (2-2p) - p + 1 = 3 - 3p$. Doubling to real dimensions gives $6 - 6p = 3\chi$.

The index does not depend on the metric: it is a topological invariant, as the Atiyah–Singer theorem guarantees (Chapter 4). Combined with the vanishing results above, it fixes both dimensions for every genus. For $p \geq 2$ the number of real moduli is $-(6-6p) = 6p - 6$, recovering (9.39). The table and Figure 9.17 collect the answers.

genus	$\dim_{\mathbb{R}} \operatorname{Ker} P_1$ (CKVs)	$\dim_{\mathbb{R}} \operatorname{Ker} P_1^\dagger$ (moduli)	index	3χ
$p = 0$ (sphere)	6 ($\mathrm{PSL}(2,\mathbb{C})$)	0	6	6
$p = 1$ (torus)	2 (translations)	2 (τ)	0	0
$p \geq 2$	0	$6p - 6$	$6 - 6p$	$6 - 6p$

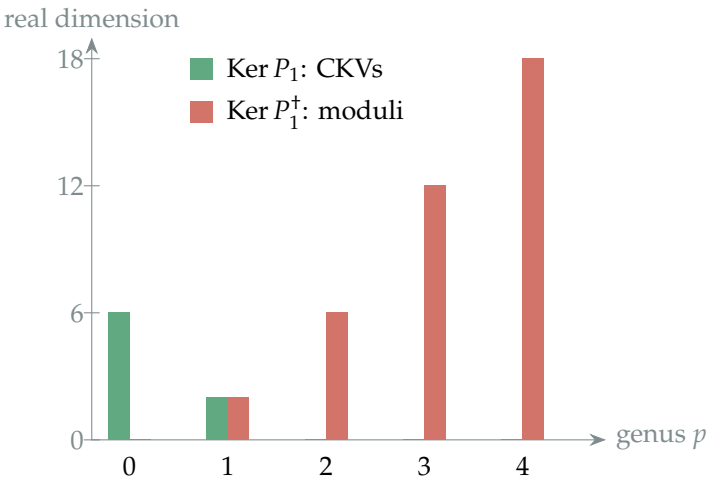


Figure 9.17: Kernel (green) and cokernel (red) of P_1 as the genus grows. The green bar minus the red bar is always $6-6p$, the index. For the sphere all the freedom is in the CKVs; for the torus the two exactly balance; from genus 2 onwards there are no CKVs and $6p-6$ moduli.

Physical picture: Counting degrees of freedom, globally

Locally a 2D metric has 3 functions, the Weyl group removes 1 and diffeomorphisms remove 2: $3-1-2=0$. Globally the count is corrected by finitely many modes on each side. Diffeomorphisms that do nothing to the conformal class (CKVs) mean we removed "too much", and deformations that no diffeomorphism can produce (moduli) mean we removed "too little". The index theorem says #moduli − #CKVs $= -3\chi$: a topological correction to naive counting, exactly like zero-mode counting for the Dirac operator in an instanton background (Chapter 8).

9.2.10 Teichmüller space from the decomposition

Now divide by the gauge groups. Dividing by the Weyl group removes C_g; dividing by Diff_0 removes Range P_1. What survives at the point $[g]$ of the quotient is Ker $P_1^\dagger$. Using the constant-curvature slice to divide by the Weyl group, this is Nash's (9.22),

$$\mathcal{T}_p = \mathrm{Met}_{\mathrm{const}}(\Sigma)/\mathrm{Diff}_0(\Sigma), \qquad T_{[g]}\mathcal{T}_p \;\cong\; \mathrm{Ker}\,P_1^\dagger \;=\; H^0(\Sigma, K^2) \text{ (as a real vector space).} \quad (9.62)$$

One point needs checking: the moduli directions $h \in \mathrm{Ker}\,P_1^\dagger$ really are tangent to the slice $\mathrm{Met}_{\mathrm{const}}$, so that moving along them keeps the curvature constant.

Worked example 9.40: Moduli deformations preserve constant curvature

The first-order change in the scalar curvature $R = 2K$ under $g \to g + h$ is the standard formula (derived by varying the Christoffel symbols; see Chapter T2 for R in terms of the connection)

$$\delta R = -R^{ab}h_{ab} + \nabla^a\nabla^b h_{ab} - \nabla^2\big(g^{ab}h_{ab}\big).$$

For $h \in \mathrm{Ker}\,P_1^\dagger$,

$$\delta R \overset{(1)}{=} -\tfrac{R}{2}\,g^{ab}h_{ab} + \nabla^a\big(\nabla^b h_{ab}\big) - \nabla^2(\mathrm{tr}\,h)$$
$$\overset{(2)}{=} 0 - 0 - 0 = 0.$$

(1) In two dimensions $R^{ab} = \frac{R}{2}g^{ab}$.

(2) h is traceless, so the first and third terms vanish; h is divergence-free, so the middle one does.

So the curvature, and with it the area (by Gauss–Bonnet), is unchanged to first order: $\operatorname{Ker} P_1^\dagger \subset T_g\mathrm{Met}_{\mathrm{const}}$. For the torus this is the statement that the deformation $\partial_\tau \hat g_\tau$ of (9.42) stays flat, which is obvious since all its entries are constant.

9.2.11 Preview: the conformal anomaly and the number 26

Nash's instruction "set $d = 26$" deserves a word of explanation now, even though the full story is Chapter 10. The division by $\operatorname{vol} C_+^\infty$ in (9.35) is legitimate only if the integrand does not depend on the Weyl factor. The classical action does not (Example 9.7), but the determinants do.

Key idea: The Weyl variation of the matter determinant

Under $g = e^{2\omega}\hat g$ (quoted from Chapter 10; it is the Polyakov–Alvarez formula),

$$\log \frac{\det' \Delta_g}{A_g} = \log \frac{\det' \Delta_{\hat g}}{A_{\hat g}} - \frac{1}{12\pi}\int_\Sigma \mathrm{d}^2x \sqrt{\hat g}\,\big(\hat g^{ab}\partial_a\omega\,\partial_b\omega + \hat R\,\omega\big). \tag{9.63}$$

For an infinitesimal $\delta\omega$ the linear term gives $\delta \log(\det'\Delta/A) = -\frac{1}{12\pi}\int \sqrt{g}\,R\,\delta\omega$.

We can test the formula against our own calculation. For *constant* ω the gradient term drops out and

$$\log \frac{\det' \Delta_{e^{2\omega}\hat g}}{A_{e^{2\omega}\hat g}} - \log \frac{\det' \Delta_{\hat g}}{A_{\hat g}} \overset{(1)}{=} -\frac{\omega}{12\pi}\int_\Sigma \sqrt{\hat g}\,\hat R \overset{(2)}{=} -\frac{\omega}{12\pi}\cdot 4\pi\chi = -\frac{\chi}{3}\,\omega,$$

(1) (9.63) with $\partial_a\omega = 0$.
(2) Gauss–Bonnet: $\int \sqrt{g}\,K = 2\pi\chi$ and $R = 2K$.

and $e^{-\chi\omega/3} = \lambda^{-\chi/6}$ with $\lambda = e^{2\omega}$: exactly the scaling found in Example 9.26(ii) from $\zeta_\Delta(0) = \chi/6 - 1$. The constant part of the anomaly is the heat-kernel coefficient; the local formula refines it point by point.

The d matter fields contribute $(\det'\Delta/A)^{-d/2}$, so their Weyl variation is $\delta \log Z_{\mathrm{matter}} = +\frac{d}{24\pi}\int \sqrt{g}\,R\,\delta\omega$. The Faddeev–Popov factor $\sqrt{\det' P_1^\dagger P_1}$ that we are about to derive has its own Weyl variation, with the *same* structure and the coefficient $-\frac{26}{24\pi}$ (Chapter 10). The total is

$$\delta_\omega \log\big(\text{integrand of } Z_p\big) = \frac{d-26}{24\pi}\int_\Sigma \mathrm{d}^2x\,\sqrt{g}\,R\,\delta\omega. \tag{9.64}$$

For $d = 26$ the integrand is Weyl invariant, the Weyl group decouples, and its volume can be divided out. For $d \neq 26$ the Weyl factor becomes a dynamical field (the Liouville field, with the action in (9.63)), and the reduction to a finite-dimensional integral fails. In conformal field theory language (Chapter 11), d is the central charge of the matter and -26 that of the ghosts. On a curved target with metric $G_{\mu\nu}$, Weyl invariance also demands $\alpha' R_{\mu\nu} = 0$ at one loop. That is derived in Section T5.7.1.

Watch out: The torus is anomaly-blind to constant rescalings, not to local ones

For $\chi = 0$ the constant part of (9.64) vanishes, so a *uniform* rescaling of a flat torus changes nothing in any dimension. The condition $d = 26$ is still needed on the torus, because a *non-constant* $\delta\omega$ sees the local curvature R of a non-flat representative metric.

Reading Nash (pp. 262–265, (9.8)–(9.22))

- "Z_p can be expressed as a finite dimensional integral over the moduli space" is the plan of Section 9.2.1; the aside on $d = 26$ is Section 9.2.11.
- (9.8) is (9.35). The remark that Diff^+ does not act freely, and the switch to Diff_0 and $\mathrm{Met}_{\mathrm{const}}$, are the pitfall after Figure 9.13.
- The finite-dimensional volume element $\sqrt{\det\langle e_i, e_j\rangle}\, \mathrm{d}^n x$ is (9.40); (9.9)–(9.11) are (9.41); (9.12) is (9.44); (9.13) is Example 9.31.
- (9.14)–(9.18) are (9.46)–(9.50); (9.16) is the chain (9.47); footnote 2 is (9.51). Nash's sentence "$\mathrm{tr}(h) = 0$ implies (9.15)" should be read as "is equivalent to".
- (9.19) is (9.53); (9.20) is (9.59); (9.21) is (9.60). "A_g can be seen to be elliptic" is (9.56); "it is easy to see that $\mathrm{Ker}\, A_g^*$ can be identified with the space of quadratic differentials" is (9.58) and the key idea after it.
- (9.22) is (9.62).

Recap: Section

- $\mathcal{D}g = \mathcal{D}\mathrm{Met}/(\mathrm{vol}\,\mathrm{Weyl} \cdot \mathrm{vol}\,\mathrm{Diff})$; gauge fixing means integrating over a slice with a Jacobian $\sqrt{\det A^{\mathsf{T}} A}$ (polar coordinates).
- $\mathrm{Met}(\Sigma)$ has the natural metric $\int \sqrt{g}\ \mathrm{tr}(g^{-1}Xg^{-1}Y)$, invariant under diffeomorphisms; at one point it is $\mathbb{R}_+\times$ (hyperbolic plane).
- $\mathcal{L}_V g = \nabla_a V_b + \nabla_b V_a$; CKVs are the V with traceless part zero, i.e. holomorphic V^z: six on S^2, two on T^2, none for $p \geq 2$.
- $T_g\mathrm{Met} = C_g \oplus \mathrm{Range}\, P_1 \oplus \mathrm{Ker}\, P_1^\dagger$, with $P_1 V$ = traceless part of $\mathcal{L}_V g$ (a Cauchy–Riemann operator, elliptic) and $P_1^\dagger = -2\,\mathrm{div}$.
- $\mathrm{Ker}\, P_1^\dagger$ = holomorphic quadratic differentials = $T\mathcal{T}_p$; $\mathrm{index}\, P_1 = 3\chi$ by Riemann–Roch; $P_1^\dagger P_1 = -2(\nabla^2 + K)$.
- Weyl invariance of the quantum integrand requires $d = 26$.

9.3 The Weil–Petersson metric

We now know that the moduli directions at g are the holomorphic quadratic differentials $\mathrm{Ker}\, P_1^\dagger$. To integrate over moduli space we need a *volume element* on it, and the cleanest way to get a volume element is from a metric ((9.40)). Nash's §3 builds that metric. The idea is simple: $\mathrm{Met}(\Sigma)$ already has a metric, the gauge group acts on it by isometries, and a quotient by isometries inherits a metric. The result is the **Weil–Petersson metric** on $\mathcal{T}_p$ and $\mathcal{M}_p$. We first see how quotient metrics work in a case we can draw.

9.3.1 Quotient metrics in finite dimensions

Worked example 9.41: The sphere divided by rotations about an axis

Let $M = S^2$ with the round metric $\mathrm{d}s^2 = \mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\varphi^2$, and let $G = \mathrm{SO}(2)$ act by $\varphi \mapsto \varphi + c$. These rotations are isometries. The orbits are the circles of latitude, and the quotient M/G is the interval $\theta \in [0, \pi]$.

To measure the length of a tangent vector $U = \dot\theta\,\partial_\theta$ to the quotient, lift it to a tangent vector X on M that projects to U. There are many lifts, $X = \dot\theta\,\partial_\theta + c\,\partial_\varphi$ for any c, because adding a vector along the orbit does not change the projection. The natural choice is the **horizontal lift**, the one orthogonal to the orbit: $c = 0$. Then

$$|U|^2_{M/G} \overset{(1)}{=} |X^{\mathrm{hor}}|^2_M \overset{(2)}{=} \dot\theta^2.$$

(1) Definition of the quotient metric: the length of the horizontal lift.
(2) $\langle \partial_\theta, \partial_\theta \rangle = 1$.

So the quotient metric is $\mathrm{d}\theta^2$: the length of the interval is π, the distance from pole to pole. If we had used a non-horizontal lift, we would have got $\dot\theta^2 + c^2\sin^2\theta$, which depends on the arbitrary c. Now compare volumes:

$$\mathrm{vol}(S^2) \overset{(1)}{=} \int_0^\pi \mathrm{d}\theta \int_0^{2\pi} \mathrm{d}\varphi\, \sin\theta \overset{(2)}{=} \int_0^\pi \underbrace{\mathrm{d}\theta}_{\text{quotient volume}} \underbrace{2\pi\sin\theta}_{\text{orbit length}} = 4\pi.$$

(1) The Riemannian volume element $\sqrt{\det G}\,\mathrm{d}\theta\,\mathrm{d}\varphi = \sin\theta\,\mathrm{d}\theta\,\mathrm{d}\varphi$.
(2) Do the φ integral first: it is the length of the orbit through θ.

The total volume is "quotient volume times orbit volume". The orbit volume is not constant: it is $2\pi\sin\theta = \mathrm{vol}(G) \times \sin\theta$, and the factor $\sin\theta$ is the length of the vector ∂_φ by which the Lie algebra of G moves the point. That factor is the Faddeev–Popov Jacobian of Example 9.27 again, now on a curved space.

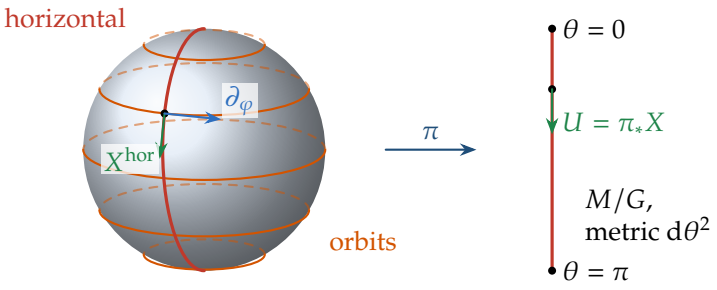


Figure 9.18: A quotient metric. The rotation group moves points along the orange orbits. A tangent vector U to the quotient has many lifts; the horizontal one (along a meridian, orthogonal to the orbits) is unique, and its length defines $|U|$. The Weil–Petersson metric is this construction applied to $\mathrm{Met}_{\mathrm{const}}(\Sigma) \to \mathcal{T}_p$, with the orbits of Diff_0 in place of the latitude circles.

Working definition 9.10 (Quotient (submersion) metric)

Let a group G act on a Riemannian manifold M by isometries [each group element preserves the length of every tangent vector], freely [no group element other than the identity fixes any point], with quotient map $\pi : M \to M/G$. For $U \in T_{\pi(m)}(M/G)$:

1. pick any point m in the fibre over $\pi(m)$ [the orbit $G \cdot m$, all the points that π sends to $\pi(m)$];
2. lift U to the unique $X \in T_m M$ with $\pi_* X = U$ [if X is the velocity of a curve in M, then $\pi_* X$ is the velocity of its image in M/G] that is orthogonal to the orbit $G \cdot m$;
3. set $|U|_{M/G} = |X|_M$.

Step 1 is harmless because G acts by isometries (it carries horizontal lifts at m to horizontal lifts at $g \cdot m$ without changing their length); step 2 is where the choice of lift is made canonical.

9.3.2 The metric on $\mathrm{Met}_{\mathrm{const}}$ and its isometries

$\mathrm{Met}_{\mathrm{const}}(\Sigma) \subset \mathrm{Met}(\Sigma)$ is a submanifold (the constant-curvature metrics (9.37)). Any submanifold of a Riemannian manifold inherits a metric by restriction: to measure a tangent vector to the submanifold, measure it as a vector in the big space. This is how the unit sphere in $\mathbb{R}^3$ gets its round metric from the Euclidean one. Nash denotes the restricted metric by

$$\langle X, Y \rangle_{\mathrm{const}} = \langle X, Y \rangle_{\mathrm{met}} \quad \text{for } X, Y \in T_g \mathrm{Met}_{\mathrm{const}}(\Sigma). \tag{9.65}$$

Diff_0 preserves $\mathrm{Met}_{\mathrm{const}}$ and acts on it by isometries. The pull-back of a metric of constant curvature K by a diffeomorphism again has constant curvature K (curvature is defined geometrically, so $K(\alpha^* g) = \alpha^* K(g) = K(g)$ when $K(g)$ is constant) and the same area. So α^* maps $\mathrm{Met}_{\mathrm{const}}$ to itself. That it preserves $\langle\ ,\ \rangle_{\mathrm{const}}$ is Example 9.29, restricted to tangent vectors of $\mathrm{Met}_{\mathrm{const}}$.

Watch out: Isometries of $\mathrm{Met}_{\mathrm{const}}$, not of Σ (Nash's footnote 3)

Two different statements are easy to confuse.

- α is an isometry *of the surface* (Σ, g) if $\alpha^* g = g$. Only special α do this (the Killing flows).
- α acts as an isometry *of the space of metrics*: $\langle \alpha^* X, \alpha^* Y \rangle_{\alpha^* g} = \langle X, Y \rangle_g$. Every diffeomorphism does this.

Nash means the second. Example: the Dehn twist $\alpha(x, y) = (x + y, y)$ on the square torus maps $g = \mathrm{d}x^2 + \mathrm{d}y^2$ to $\alpha^* g = (\mathrm{d}x + \mathrm{d}y)^2 + \mathrm{d}y^2 \neq g$, so it is not an isometry of the torus. But it maps the point g of $\mathrm{Met}_{\mathrm{const}}$ to the different point $\alpha^* g$, and the tangent spaces at the two points isometrically onto each other.

9.3.3 Definition of the Weil–Petersson metric

Following Nash, let

$$\pi : \mathrm{Met}_{\mathrm{const}}(\Sigma) \longrightarrow \mathrm{Met}_{\mathrm{const}}(\Sigma)/\mathrm{Diff}_0(\Sigma) = \mathcal{T}_p \tag{9.66}$$

be the projection (Nash's (9.23)), write $[g_{\mathrm{const}}]$ for the equivalence class of g_{const} (a point of $\mathcal{T}_p$), and $T_{[g_{\mathrm{const}}]}\mathcal{T}_p$ for the tangent space there (Nash's (9.24)).

Working definition 9.11 (Weil–Petersson metric (Nash's (9.25)))

For $U, V \in T_{[g]}\mathcal{T}_p$ choose a representative metric g and lifts $X, Y \in T_g\mathrm{Met}_{\mathrm{const}}$ with $\pi_* X = U$, $\pi_* Y = V$, horizontal (orthogonal to the Diff_0 orbit through g [all metrics $\alpha^* g$ with α a diffeomorphism that can be continuously deformed to the identity]). Then

$$\langle U, V\rangle_{\mathrm{wp}} = \langle X, Y\rangle_{\mathrm{const}} = \int_\Sigma \mathrm{d}^2x\sqrt{g}\ \mathrm{tr}\big(g^{-1}Xg^{-1}Y\big). \tag{9.67}$$

How to compute it: represent the tangent vectors to $\mathcal{T}_p$ by holomorphic quadratic differentials [tensors $q(z)\,\mathrm{d}z^2$ with q holomorphic in every local complex coordinate z] (they are the horizontal vectors, see below), and evaluate the L^2 inner product (9.69).

Two things must be checked: that the horizontal lifts are what we think they are, and that the answer does not depend on the representative g.

The horizontal vectors are the quadratic differentials. The tangent space to the Diff_0 orbit through g is $\{\mathcal{L}_V g\}$. A vector X is horizontal iff $\langle X, \mathcal{L}_V g\rangle_{\mathrm{met}} = 0$ for all V. The same integration by parts as in (9.57), now without assuming X traceless, gives $\langle X, \mathcal{L}_V g\rangle_{\mathrm{met}} = \langle -2\nabla^a X_{a\,\cdot}, V\rangle_{\mathrm{vect}}$, so horizontal means *divergence-free*, $\nabla^a X_{ab} = 0$. Take $p \geq 2$ and split $X = \beta g + h$ with h traceless. Two conditions constrain β:

$$\begin{aligned} 0 &\overset{(1)}{=} \nabla^a(\beta g_{ab} + h_{ab}) = \nabla_b\beta + \nabla^a h_{ab} \quad\Longrightarrow\quad \nabla^a\nabla^b h_{ab} = -\nabla^2\beta, \\ 0 &\overset{(2)}{=} \delta R = -\tfrac{R}{2}(2\beta) + \nabla^a\nabla^b(\beta g_{ab} + h_{ab}) - \nabla^2(2\beta) \\ &\overset{(3)}{=} -R\beta + \nabla^2\beta - \nabla^2\beta - 2\nabla^2\beta \overset{(4)}{=} 2\beta + 2\Delta\beta. \end{aligned}$$

(1) X is horizontal (divergence-free); take a further divergence.
(2) X is tangent to $\mathrm{Met}_{\mathrm{const}}$, so the curvature does not change; use the formula for δR of Example 9.40 with $\mathrm{tr}\, X = 2\beta$.
(3) $\nabla^a\nabla^b(\beta g_{ab}) = \nabla^2\beta$, and substitute the result of (1).
(4) $R = 2K = -2$ for $p \geq 2$, and $\Delta = -\nabla^2$ is the positive Laplacian.

Pairing $(1 + \Delta)\beta = 0$ with β gives $\|\beta\|^2 + \|\nabla\beta\|^2 = 0$, so $\beta = 0$. Hence a horizontal tangent vector to $\mathrm{Met}_{\mathrm{const}}$ is traceless and divergence-free: it lies in $\mathrm{Ker}\, P_1^\dagger$. Conversely every element of $\mathrm{Ker}\, P_1^\dagger$ is horizontal and tangent to $\mathrm{Met}_{\mathrm{const}}$ (Example 9.40). So

$$(\text{horizontal subspace at } g) = \mathrm{Ker}\, P_1^\dagger \cong T_{[g]}\mathcal{T}_p. \tag{9.68}$$

Independence of the representative (Nash's (9.26)–(9.29)). Suppose $[g'] = [g]$, so $g' = \alpha^* g$ for some $\alpha \in \mathrm{Diff}_0$ (Nash writes α_*). Let X, Y be the horizontal lifts of U, V at g. Then $\alpha^* X, \alpha^* Y$ are tangent to $\mathrm{Met}_{\mathrm{const}}$ at g', project to the same U, V (since $\pi \circ \alpha^* = \pi$), and are horizontal (since α^* is an isometry mapping orbits to orbits, it maps the orthogonal complement of one orbit's tangent space to that of the other). So they are *the* horizontal lifts at g', and

$$\begin{aligned} \langle U, V\rangle_{\mathrm{wp}}\big|_{\text{computed at } g'} &\overset{(1)}{=} \langle \alpha^* X, \alpha^* Y\rangle_{\mathrm{const},\, g'} \\ &\overset{(2)}{=} \langle X, Y\rangle_{\mathrm{const},\, g} \overset{(3)}{=} \langle U, V\rangle_{\mathrm{wp}}\big|_{\text{computed at } g}. \end{aligned}$$

(1) Definition (9.67) at g', with the lifts just found.
(2) α^* is an isometry of $\mathrm{Met}_{\mathrm{const}}$ (Example 9.29).
(3) Definition at g.

Watch out: Which lift?

Nash says that $X' = \alpha_* X$ follows "from the fibration". That is true for *horizontal* lifts only. A general lift is ambiguous up to adding a vertical vector $\mathcal{L}_W g$, and $\langle X + \mathcal{L}_W g, X + \mathcal{L}_W g\rangle \neq \langle X, X\rangle$ in general, exactly as $\dot\theta^2 + c^2 \sin^2\theta \neq \dot\theta^2$ in Example 9.41. An equivalent recipe that works with *any* lift is to first project orthogonally onto $\operatorname{Ker} P_1^\dagger$: $\langle U, V\rangle_{\mathrm{wp}} = \langle \Pi X, \Pi Y\rangle_{\mathrm{met}}$, where Π is the orthogonal projector of (9.60).

The Weil–Petersson metric in complex form. Write a horizontal vector as $h = \varphi\,\mathrm{d}z^2 + \bar\varphi\,\mathrm{d}\bar z^2$, with $\varphi = h_{zz}$ holomorphic, and the metric as $g = \rho\,\mathrm{d}z\,\mathrm{d}\bar z$ (so $g_{z\bar z} = \rho/2$, $g^{z\bar z} = 2/\rho$, $\sqrt{g} = \rho$ in the coordinates x, y). Then

$$\begin{aligned}\langle h, h'\rangle_{\mathrm{met}} &\overset{(1)}{=} \int \mathrm{d}^2x\,\rho\,\big(g^{z\bar z}g^{z\bar z}h_{zz}h'_{\bar z\bar z} + g^{\bar z z}g^{\bar z z}h_{\bar z\bar z}h'_{zz}\big)\\ &\overset{(2)}{=} \int \mathrm{d}^2x\,\rho\,\frac{4}{\rho^2}\big(\varphi\bar\varphi' + \bar\varphi\varphi'\big)\\ &\overset{(3)}{=} 8\,\mathrm{Re}\int_\Sigma \mathrm{d}^2x\,\frac{\varphi\,\bar\varphi'}{\rho}.\end{aligned} \tag{9.69}$$

(1) Evaluate $g^{ac}g^{bd}h_{ab}h'_{cd}$ in the complex basis; only $g^{z\bar z}$ is non-zero, and h has only zz and $\bar z\bar z$ components, so only two terms survive.
(2) $g^{z\bar z} = 2/\rho$ and $h'_{\bar z\bar z} = \bar\varphi'$.
(3) The two terms are complex conjugates.

The integral $\int \varphi\bar\varphi'\rho^{-1}\,\mathrm{d}^2x$ is the **Petersson inner product** of two quadratic differentials (the integrand is a well-defined 2-form because $\varphi\bar\varphi'$ has weight $(2,2)$ and ρ^{-1} has weight $(-1,-1)$). Weil's idea was to use it as a metric on Teichmüller space; hence the name.

Physical picture: What the Weil–Petersson distance measures

The WP length of a path in $\mathcal{T}_p$ is the L^2 size of the *shape change* of the surface along the path, after all size changes (Weyl) and all reparametrisations (diffeomorphisms) have been projected out. It is the unique "kinetic energy" for moving through the space of Riemann surfaces that is inherited from the obvious kinetic energy on metrics. Standard facts (quoted): it is Kähler, it is invariant under the mapping class group, so it descends to $\mathcal{M}_p$, and $\mathcal{M}_p$ has finite WP volume.

9.3.4 The torus: Weil–Petersson = hyperbolic

Worked example 9.42: The WP metric on $\mathcal{T}_1$

Every flat metric of area 1 on the torus $\mathbb{R}^2/\mathbb{Z}^2$ is, up to a diffeomorphism isotopic to the identity, one of

$$\hat g_\tau = \frac{1}{\tau_2}\,|\mathrm{d}x + \tau\,\mathrm{d}y|^2, \qquad \tau = \tau_1 + i\tau_2,\ \tau_2 > 0,$$

the metric (9.42) made constant over the torus. Its area is $\int \sqrt{\det \hat g_\tau}\,\mathrm{d}x\,\mathrm{d}y = 1$. So $\mathcal{T}_1$ is the upper half-plane, with coordinate τ. The coordinate vectors $\partial_{\tau_i}\hat g_\tau$ are constant in x, y, traceless (because $\det \hat g_\tau = 1$ and $\operatorname{tr}(\hat g^{-1}\partial\hat g) = \partial \log\det \hat g = 0$) and therefore divergence-free:

they are horizontal (Example 9.38). Hence

$$\langle \partial_{\tau_i}, \partial_{\tau_j} \rangle_{\mathrm{wp}} \overset{(1)}{=} \int_0^1 \int_0^1 \mathrm{d}x\,\mathrm{d}y\ \sqrt{\det \hat{g}_\tau}\ \operatorname{tr}\big(\hat{g}_\tau^{-1}\partial_{\tau_i}\hat{g}_\tau\ \hat{g}_\tau^{-1}\partial_{\tau_j}\hat{g}_\tau\big)$$
$$\overset{(2)}{=} 1 \cdot 1 \cdot \frac{2\,\delta_{ij}}{\tau_2^2},$$

(1) Definition (9.67); the lifts are horizontal.
(2) The integrand is constant; $\det \hat{g}_\tau = 1$; the trace is (9.43); the coordinate area of the unit square is 1.

so

$$\mathrm{d}s^2_{\mathrm{wp}} = \frac{2\,(\mathrm{d}\tau_1^2 + \mathrm{d}\tau_2^2)}{\tau_2^2} = \frac{2\,|\mathrm{d}\tau|^2}{(\operatorname{Im}\tau)^2}, \qquad \mathrm{d}\mu(\mathcal{T}_1) = \frac{2\,\mathrm{d}\tau_1\,\mathrm{d}\tau_2}{\tau_2^2}. \tag{9.70}$$

The Weil–Petersson metric of the torus is twice the Poincaré metric of the upper half-plane. *Check with the complex formula.* At $\tau = i$, $\hat{g} = \delta$, $\rho = 1$, and $\partial_{\tau_1}\hat{g} = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, i.e. $h = 2\,\mathrm{d}x\,\mathrm{d}y$. Its zz component is $\varphi = h(\partial_z, \partial_z) = \frac{1}{4}(h_{11} - 2ih_{12} - h_{22}) = -\frac{i}{2}$, and (9.69) gives $8 \cdot |\varphi|^2 \cdot 1 = 8 \cdot \frac{1}{4} = 2$, in agreement with $2/\tau_2^2$ at $\tau_2 = 1$.

The mapping class group acts by Möbius transformations. A matrix $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}(2,\mathbb{Z})$ defines a linear diffeomorphism of $\mathbb{R}^2/\mathbb{Z}^2$ (it maps $\mathbb{Z}^2$ to itself, and so does its inverse). Pull back $\hat{g}_\tau$ by $(x, y) = (dx' + by',\ cx' + ay')$:

$$\mathrm{d}x + \tau\,\mathrm{d}y \overset{(1)}{=} (d + c\tau)\,\mathrm{d}x' + (b + a\tau)\,\mathrm{d}y'$$
$$\overset{(2)}{=} (c\tau + d)\,\big(\mathrm{d}x' + \tau'\,\mathrm{d}y'\big), \qquad \tau' = \frac{a\tau + b}{c\tau + d},$$
$$\operatorname{Im}\tau' \overset{(3)}{=} \frac{\operatorname{Im}\big((a\tau + b)(c\bar{\tau} + d)\big)}{|c\tau + d|^2} \overset{(4)}{=} \frac{\tau_2}{|c\tau + d|^2}.$$

(1) Substitute $\mathrm{d}x = d\,\mathrm{d}x' + b\,\mathrm{d}y'$ and $\mathrm{d}y = c\,\mathrm{d}x' + a\,\mathrm{d}y'$.
(2) Factor out $(c\tau + d)$.
(3) Multiply numerator and denominator by $\overline{c\tau + d}$.
(4) $\operatorname{Im}\big(ad\,\tau + bc\,\bar{\tau}\big) = (ad - bc)\tau_2 = \tau_2$.

So $\hat{g}_\tau$ pulls back to $\frac{|c\tau+d|^2}{\tau_2}\,|\mathrm{d}x' + \tau'\mathrm{d}y'|^2 = \frac{1}{\tau_2'}\,|\mathrm{d}x' + \tau'\mathrm{d}y'|^2 = \hat{g}_{\tau'}$. The mapping class group $\Gamma_{T^2} = \mathrm{SL}(2,\mathbb{Z})$ acts on $\mathcal{T}_1$ by $\tau \mapsto \frac{a\tau+b}{c\tau+d}$. The matrix $-\mathbf{1}$ acts trivially (it is $(x, y) \mapsto (-x, -y)$, an isometry of every flat torus), so the effective group is $\mathrm{PSL}(2,\mathbb{Z})$. It is generated by

$$T:\ \tau \mapsto \tau + 1 \text{ (the Dehn twist)}, \qquad S:\ \tau \mapsto -\frac{1}{\tau} \text{ (exchange the two cycles)}.$$

These are WP isometries, as they must be:

$$\frac{|\mathrm{d}\tau'|^2}{(\operatorname{Im}\tau')^2} \overset{(1)}{=} \frac{|\mathrm{d}\tau|^2\,/|c\tau + d|^4}{\tau_2^2/|c\tau + d|^4} = \frac{|\mathrm{d}\tau|^2}{\tau_2^2}.$$

(1) $\mathrm{d}\tau' = \frac{a(c\tau+d)-c(a\tau+b)}{(c\tau+d)^2}\mathrm{d}\tau = \frac{\mathrm{d}\tau}{(c\tau+d)^2}$, and the formula for $\operatorname{Im}\tau'$ above.

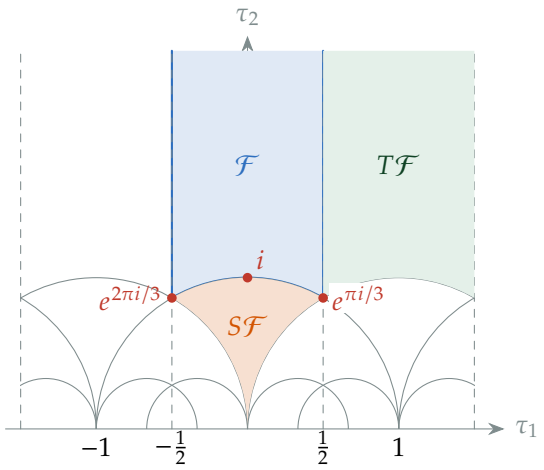


Figure 9.19: Teichmüller space $\mathcal{T}_1$ is the upper half-plane with the hyperbolic (Weil–Petersson) metric. The mapping class group $\mathrm{PSL}(2,\mathbb{Z})$ tiles it by copies of the fundamental domain $\mathcal{F} = \{|\tau_1| \le \frac{1}{2},\ |\tau| \ge 1\}$ (blue): T shifts it sideways (green) and S inverts it in the unit circle (orange). Each point of $\mathcal{M}_1 = \mathcal{T}_1/\mathrm{PSL}(2,\mathbb{Z})$ appears once in $\mathcal{F}$, up to boundary identifications. The red points are fixed by a finite subgroup ($\tau = i$ by S, $\tau = e^{2\pi i/3}$ by ST): the square and hexagonal tori have extra symmetry, which is why $\mathcal{M}_1$ is an orbifold rather than a manifold.

Worked example 9.43: The Weil–Petersson volume of $\mathcal{M}_1$

Integrate (9.70) over $\mathcal{F}$:

$$\begin{aligned}\mathrm{vol}_{\mathrm{wp}}(\mathcal{M}_1) &\overset{(1)}{=} \int_{-1/2}^{1/2} \mathrm{d}\tau_1 \int_{\sqrt{1-\tau_1^2}}^{\infty} \frac{2\,\mathrm{d}\tau_2}{\tau_2^2} \\ &\overset{(2)}{=} \int_{-1/2}^{1/2} \frac{2\,\mathrm{d}\tau_1}{\sqrt{1-\tau_1^2}} \\ &\overset{(3)}{=} 2\big[\arcsin\tau_1\big]_{-1/2}^{1/2} = 2\cdot\frac{\pi}{3} = \frac{2\pi}{3}.\end{aligned}$$

(1) The lower boundary of $\mathcal{F}$ is the unit circle $\tau_2 = \sqrt{1-\tau_1^2}$.

(2) $\int_a^\infty \mathrm{d}\tau_2/\tau_2^2 = 1/a$.

(3) $\int \mathrm{d}u/\sqrt{1-u^2} = \arcsin u$, and $\arcsin\frac{1}{2} = \frac{\pi}{6}$.

The moduli space of tori is non-compact (the cusp $\tau_2 \to \infty$, a very long thin torus) but has *finite* volume. Whether the string amplitude converges in the cusp is then a question about the integrand, which we meet in Example 9.56.

Watch out: Integrating over $\mathcal{T}_p$ versus $\mathcal{M}_p$

The integrand of Z_p is invariant under the whole Diff^+, so it is constant along Γ_Σ-orbits in $\mathcal{T}_p$. An integral over $\mathcal{T}_p$ therefore counts each point of $\mathcal{M}_p$ $|\Gamma_\Sigma|$ times, and $\frac{1}{|\Gamma_\Sigma|}\int_{\mathcal{T}_p} = \int_{\mathcal{M}_p}$. For $p \ge 1$ the group Γ_Σ is infinite, so this is a formal way of saying "integrate over a fundamental domain"; it is the form in which Nash writes (9.35).

9.3.5 The partition function as an integral over moduli

We now have every ingredient, and Nash's last step (pp. 267–268) assembles them. The question is: *what does the formal integral* $\int \mathcal{D}g\,(\cdots)$ *over all metrics become, once we use coordinates adapted to the gauge symmetry?* The answer is a finite-dimensional integral over Teichmüller space, with two Jacobian factors: the Weil–Petersson volume element, and the determinant $\sqrt{\det{}' P_1^\dagger P_1}$. The logic is exactly that of the polar-coordinate example Example 9.27, done once more in infinite dimensions. We go slowly, because every factor in the final formula (9.35) has a specific origin.

Watch out: Two meanings of the letter A

Nash writes A_g for the operator that we call P_1. Earlier in this chapter we used A_g for the area $(1,1)_g = \int_\Sigma \sqrt{g}$. From here on we write P_1 (and $P_1^\dagger = A_g^*$) for the operator and $(1,1)_g$ for the area, and we reserve the letter A for finite-dimensional matrices in toy models.

9.3.6 Coordinates adapted to the gauge symmetry: the conformal gauge

In Section 9.1.6 we used a local fact: near every point, every metric on a surface can be written $e^{2\omega}\,|\mathrm{d}w|^2$ in a suitable coordinate w (isothermal coordinates). The global version of that statement is what makes the gauge fixing work.

Key idea: Global conformal gauge

Choose, for each point $t = (t^1, \dots, t^n)$ of Teichmüller space $\mathcal{T}_p$ ($n = \dim_{\mathbb{R}} \mathcal{T}_p$), one constant-curvature metric $\hat{g}_t \in \mathrm{Met}_{\mathrm{const}}(\Sigma)$ in the class t, depending smoothly on t. Then every metric on Σ can be written as

$$g = \alpha^*\big(e^{\sigma}\,\hat{g}_t\big), \qquad \sigma \in C^\infty(\Sigma), \quad \alpha \in \mathrm{Diff}_0(\Sigma), \quad t \in \mathcal{T}_p. \tag{9.71}$$

Physicists call the choice of representatives $\hat{g}_t$ the **conformal gauge**: the metric is fixed to be "a Weyl factor times a reference metric", and the only gauge-invariant data left are the n numbers t^i.

Why is (9.71) *true?* Read the tower of quotients Figure 9.13 backwards. Given g, uniformisation gives a unique σ_1 with $e^{-\sigma_1} g \in \mathrm{Met}_{\mathrm{const}}$. Its class in $\mathcal{T}_p = \mathrm{Met}_{\mathrm{const}}/\mathrm{Diff}_0$ is some point t, and $\hat{g}_t$ is in the same class, so $e^{-\sigma_1} g = \alpha^* \hat{g}_t$ for some $\alpha \in \mathrm{Diff}_0$. Then $g = e^{\sigma_1}\alpha^*\hat{g}_t = \alpha^*\big(e^{\sigma}\hat{g}_t\big)$ with $\sigma = \sigma_1 \circ \alpha^{-1}$. A smooth choice $t \mapsto \hat{g}_t$ exists globally because $\mathcal{T}_p$ is a ball (a section of a bundle over a contractible base always exists). This is also why we work with $\mathcal{T}_p$ and Diff_0 first, and bring in the mapping class group only at the very end: over $\mathcal{M}_p$, which has orbifold points, no such smooth choice is possible.

Worked example 9.44: Putting a torus metric into conformal gauge

Let $\Sigma = \mathbb{R}^2/\mathbb{Z}^2$ and $g = \mathrm{d}x^2 + f(y)^2\,\mathrm{d}y^2$, where $f > 0$ is any smooth function of period 1. This metric is flat (it is $\mathrm{d}x^2 + \mathrm{d}u^2$ in the local coordinate $u = \int f\,\mathrm{d}y$) but it is not obviously one of the $\hat{g}_\tau = \frac{1}{\tau_2}\,|\mathrm{d}x + \tau\,\mathrm{d}y|^2$ of (9.42). Let

$$c = \int_0^1 f(y)\,\mathrm{d}y, \qquad Y(y) = \frac{1}{c}\int_0^y f(s)\,\mathrm{d}s.$$

Then $Y(y+1) = Y(y) + 1$ and $Y' = f/c > 0$, so $\alpha(x,y) = (x, Y(y))$ is a diffeomorphism of the torus. It lies in Diff_0: the family $Y_s = (1-s)\,y + s\,Y(y)$, $0 \le s \le 1$, consists of

diffeomorphisms (each has $Y_s' > 0$ and $Y_s(y+1) = Y_s(y) + 1$) joining the identity to α. Now

$$\begin{aligned} g &\overset{(1)}{=} \mathrm{d}x^2 + c^2\,\mathrm{d}Y^2 \\ &\overset{(2)}{=} c \cdot \frac{1}{c}\,|\mathrm{d}x + ic\,\mathrm{d}Y|^2 \\ &\overset{(3)}{=} \alpha^*\big(e^\sigma\,\hat{g}_\tau\big), \qquad e^\sigma = c, \quad \tau = ic. \end{aligned}$$

(1) $\mathrm{d}Y = \frac{f}{c}\,\mathrm{d}y$, so $f^2\mathrm{d}y^2 = c^2\mathrm{d}Y^2$.
(2) $|\mathrm{d}x + ic\,\mathrm{d}Y|^2 = \mathrm{d}x^2 + c^2\mathrm{d}Y^2$; multiply and divide by c.
(3) Compare with $\hat{g}_\tau = \frac{1}{\tau_2}\,|\mathrm{d}x + \tau\mathrm{d}y|^2$ at $\tau = ic$, written in the coordinates $(x, Y) = \alpha(x, y)$.

So the modulus is $\tau = i\int_0^1 f$, and the Weyl factor is the constant c, equal to the area of g. Two special cases make the point:

- $f = 1 + \varepsilon\cos 2\pi y$: then $c = 1$ and $\tau = i$. The wiggly metric is the square torus in disguise; the wiggle is pure gauge. (At first order this is the deformation that Example 9.38 showed to be "diffeomorphism plus Weyl".)
- $f = 2$: then $\tau = 2i$, the 1×2 rectangular torus. No choice of $\alpha \in \mathrm{Diff}_0$ and σ can turn this into the square torus.

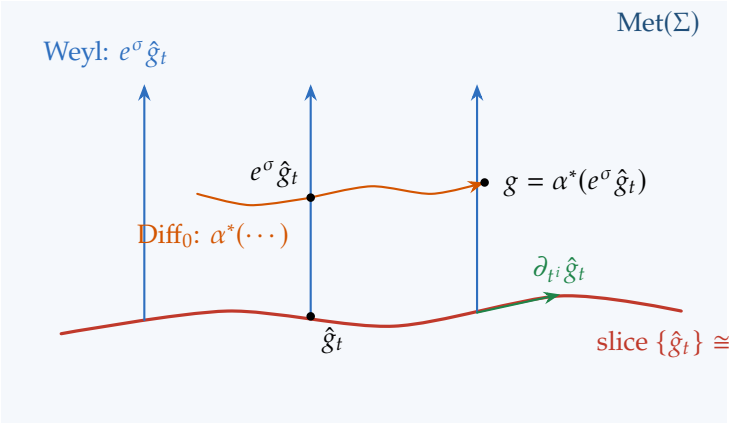


Figure 9.20: The coordinates (σ, α, t) of (9.71). Start on the finite-dimensional slice (red) at the reference metric $\hat{g}_t$; move up a Weyl line (blue) to $e^\sigma\,\hat{g}_t$; then move along a diffeomorphism orbit (orange) to g. Every metric is reached, and for $p \geq 2$ it is reached exactly once. The green arrow is a moduli direction: moving along the slice changes the conformal class.

Nash's (9.30)–(9.31). Infinitesimally, the three coordinates correspond to the three summands of the decomposition (9.60), which Nash repeats as (9.30):

$$T_g\mathrm{Met}(\Sigma) = C_g \oplus \mathrm{Range}\,P_1 \oplus \mathrm{Ker}\,P_1^\dagger.$$

Nash calls the coordinates $\{\sigma\}$ (with $f = e^\sigma \in C^\infty_+$), $\{P_1V\}$ (he writes $\{A_gV\}$) and $\{\mathcal{T}_p\}$, and writes the measure as

$$\mathcal{D}\mathrm{Met}(\Sigma) = N(f, P_1V, \mathcal{T}_p)\;\mathcal{D}\sigma\;\mathcal{D}(P_1V)\;\mathrm{d}\mathcal{T}_p, \tag{9.72}$$

with N "an appropriate normalising factor depending on the frames used". Two things need explaining: how $\mathcal{D}(P_1V)$ is related to a measure on vector fields (this produces the determinant), and what exactly $\mathrm{d}\mathcal{T}_p$ is (this produces the Weil–Petersson measure). Both come from one finite-dimensional fact about volumes.

9.3.7 The volume of the image of a linear map

Key idea: Image measure

Let $A : \mathbb{R}^k \to \mathbb{R}^n$ be linear and injective, $k \le n$. The image $A(\mathbb{R}^k)$ is a k-dimensional subspace, with the length measure inherited from $\mathbb{R}^n$. Then

$$\mathrm{d}^k(Av) = \sqrt{\det\bigl(A^{\mathsf{T}}A\bigr)}\,\mathrm{d}^k v\,: \tag{9.73}$$

the unit cube in $\mathbb{R}^k$ is mapped to a parallelepiped of k-volume $\sqrt{\det A^{\mathsf{T}}A}$.

Why: the image of the unit cube is spanned by the columns $a_1, \dots, a_k$ of A. The k-volume of a parallelepiped is the square root of the Gram determinant $\det\langle a_i, a_j\rangle$ (for $k = n$ this is $|\det A|$; for $k = 1$ it is the length $|a_1|$; in general it is what (9.40) says, for the induced metric on the image). And $\langle a_i, a_j\rangle = (A^{\mathsf{T}}A)_{ij}$.

Worked example 9.45: Two image volumes

(a) $A = \begin{pmatrix}1\\1\end{pmatrix} : \mathbb{R} \to \mathbb{R}^2$. The unit interval maps to the segment from $(0,0)$ to $(1,1)$, of length $\sqrt{2}$, and indeed $A^{\mathsf{T}}A = 2$.

(b) $A = \begin{pmatrix}1 & 0\\ 0 & 1\\ 1 & 1\end{pmatrix} : \mathbb{R}^2 \to \mathbb{R}^3$. The columns are $a_1 = (1,0,1)$ and $a_2 = (0,1,1)$. The area of the parallelogram they span is the length of the cross product,

$$|a_1 \times a_2| \overset{(1)}{=} |(0\cdot 1 - 1\cdot 1,\ 1\cdot 0 - 1\cdot 1,\ 1\cdot 1 - 0\cdot 0)| = |(-1,-1,1)| = \sqrt{3},$$

$$\sqrt{\det A^{\mathsf{T}}A} \overset{(2)}{=} \sqrt{\det\begin{pmatrix}2 & 1\\ 1 & 2\end{pmatrix}} = \sqrt{4-1} = \sqrt{3}.$$

(1) Cross product in components, $(a_1 \times a_2)_1 = a_{1,2}a_{2,3} - a_{1,3}a_{2,2}$ and cyclically.
(2) $\langle a_1, a_1\rangle = \langle a_2, a_2\rangle = 2$ and $\langle a_1, a_2\rangle = 1$.

The two answers agree, as (9.73) promised.

Nash's (9.32). Apply (9.73) with $A = P_1$, $\mathbb{R}^k = \mathrm{vect}(\Sigma)$ with the metric $\langle\ ,\ \rangle_{\mathrm{vect}}$, and $\mathbb{R}^n = C_g^{\perp}$ with $\langle\ ,\ \rangle_{\mathrm{met}}$:

$$\mathcal{D}(P_1 V) = \sqrt{\det\bigl(P_1^{\dagger}P_1\bigr)}\,\mathcal{D}V. \tag{9.74}$$

Here $P_1^{\dagger}$ is the adjoint computed in (9.58), the infinite-dimensional version of A^{T}, and the determinant is zeta-regularised: $P_1^{\dagger}P_1 = -2(\nabla^2 + K)$ is a positive Laplace-type operator (Example 9.39), so Section 9.1.10 applies to it. Formula (9.74) needs P_1 to be injective, i.e. $\operatorname{Ker} P_1 = 0$. By the Weitzenböck argument this holds for $p \ge 2$, and fails for the sphere and the torus (Section 9.3.10 below).

Watch out: What is being measured

$\mathcal{D}(P_1 V)$ is a measure on the *subspace* $\operatorname{Range} P_1$ of $C_g^{\perp}$. $\mathcal{D}V$ is a measure on $\mathrm{vect}(\Sigma)$, i.e. on the Lie algebra of Diff_0, and after exponentiation on Diff_0 itself. The determinant converts one into the other. It is *not* the Jacobian of a map from a space to itself, which is why it

comes as a square root of $\det P_1^\dagger P_1$ rather than as $\det P_1$ (the latter would not even make sense: P_1 maps between two different spaces).

9.3.8 The Jacobian is triangular

Now compute the Jacobian of the change of variables $(\sigma, \alpha, t) \mapsto g$ in (9.71). Because of the gauge invariance of the integrand, we only need it at $\alpha = \mathrm{id}$, $\sigma = 0$, i.e. at the point $g = \hat{g}_t$ of the slice. (At other points it is obtained by transporting with the gauge group, which acts by isometries of $\langle\ ,\ \rangle_{\mathrm{met}}$, Example 9.29; we come back to the σ-dependence in the normalisation N below.) First a finite-dimensional warm-up.

Worked example 9.46: A triangular Jacobian in $\mathbb{R}^3$

Let $e_1 = (1,0,0)$, $a = (c,r,0)$ and $b = (u,w,m)$ be three vectors in $\mathbb{R}^3$, and use (s,v,t) as coordinates via $x = s\,e_1 + v\,a + t\,b$. The Jacobian is

$$\left|\det \frac{\partial x}{\partial(s,v,t)}\right| \overset{(1)}{=} \left|\det \begin{pmatrix} 1 & c & u \\ 0 & r & w \\ 0 & 0 & m \end{pmatrix}\right| \overset{(2)}{=} |r\,m|\,.$$

(1) The columns are e_1, a and b.
(2) The determinant of a triangular matrix is the product of its diagonal entries.

The off-diagonal entries c, u, w do not matter. Geometrically, $1 = |e_1|$; r is the component of a orthogonal to e_1; and m is the component of b orthogonal to the plane of e_1 and a. So *the volume element is the product of the "new" orthogonal components*, taken in order.

The dictionary with strings: e_1 is the Weyl direction, a is a diffeomorphism direction (which has a Weyl component c, its trace), and b is a moduli direction (which may have both a Weyl component u and a diffeomorphism component w). Let us check that the infinite-dimensional Jacobian has exactly this triangular form.

At $g = \hat{g}_t$, a small change $(\delta\sigma, \delta V, \delta t)$ of the coordinates changes the metric by

$$\begin{aligned} \delta g &\overset{(1)}{=} \delta\sigma\, g + \mathcal{L}_{\delta V} g + \delta t^i\, \partial_i \hat{g}_t \\ &\overset{(2)}{=} \Big[\delta\sigma + \nabla_c \delta V^c + \tfrac{1}{2}\,\delta t^i\, \mathrm{tr}(\partial_i \hat{g}_t)\Big] g \;+\; P_1\big(\delta V + \delta t^i W_i\big) \;+\; \delta t^i\, \Pi\, \partial_i \hat{g}_t . \end{aligned} \tag{9.75}$$

(1) Differentiate $\alpha^*(e^\sigma \hat{g}_t)$: at $\sigma = 0$, $e^{\delta\sigma} = 1 + \delta\sigma$; the flow of δV changes g by $\mathcal{L}_{\delta V} g$ (Section 9.2.5); and $\partial_i = \partial/\partial t^i$.
(2) Split each term with (9.60). For the Lie derivative, (9.51) gives $\mathcal{L}_V g = (\nabla_c V^c)\, g + P_1 V$. For the moduli direction, the trace part is $\frac{1}{2}\,\mathrm{tr}(\partial_i \hat{g}_t)\, g$, the projection Π onto $\mathrm{Ker}\, P_1^\dagger$ gives the moduli part, and what is left is traceless and orthogonal to $\mathrm{Ker}\, P_1^\dagger$, hence in $\mathrm{Range}\, P_1$ by (9.59): it equals $P_1 W_i$ for some vector field W_i.

In block form, with rows labelled by the three orthogonal summands and columns by the three kinds of coordinate change,

$$\begin{pmatrix} [\delta g]_{C_g} \\ [\delta g]_{\mathrm{Range}\, P_1} \\ [\delta g]_{\mathrm{Ker}\, P_1^\dagger} \end{pmatrix} = \begin{pmatrix} \mathbf{1} & \nabla_c(\cdot)^c & \frac{1}{2}\,\mathrm{tr}\,\partial_i \hat{g} \\ 0 & P_1 & P_1 W_i \\ 0 & 0 & \Pi\,\partial_i \hat{g} \end{pmatrix} \begin{pmatrix} \delta\sigma \\ \delta V \\ \delta t^i \end{pmatrix}. \tag{9.76}$$

It is block upper triangular, exactly as in Example 9.46. So the volume element is the product of the three diagonal "volumes":

- the C_g block is the identity (times g): it gives $\mathcal{D}\sigma$, with the normalisation fixed by the metric $\langle \delta\sigma\, g, \delta\sigma\, g\rangle_{\text{met}} = 2\int \sqrt{g}\,(\delta\sigma)^2$;
- the Range P_1 block is P_1: by (9.74) it gives $\sqrt{\det P_1^\dagger P_1}\;\mathcal{D}V$;
- the $\operatorname{Ker} P_1^\dagger$ block is the finite matrix of vectors $\Pi\,\partial_i \hat{g}_t$, $i = 1, \dots, n$: by (9.73) it gives $\sqrt{\det\langle \Pi\partial_i \hat{g}_t, \Pi\partial_j \hat{g}_t\rangle_{\text{met}}}\;\mathrm{d}^n t$.

Key idea: The moduli volume element is the Weil–Petersson volume

By the pitfall "Which lift?" after the Weil–Petersson definition, $\langle \Pi X, \Pi Y\rangle_{\text{met}}$ is the Weil–Petersson inner product of the projected vectors. So the moduli block contributes

$$\mathrm{d}\mu(\mathcal{T}_p) = \sqrt{\det\big(G^{\text{wp}}_{ij}(t)\big)}\;\mathrm{d}^n t, \qquad G^{\text{wp}}_{ij} = \langle \partial_{t^i}, \partial_{t^j}\rangle_{\text{wp}} = \langle \Pi\partial_i \hat{g}_t, \Pi\partial_j \hat{g}_t\rangle_{\text{met}}, \tag{9.77}$$

the Riemannian volume element (9.40) of the Weil–Petersson metric. This is the measure Nash calls $\mathrm{d}\mu(\mathcal{T}_p)$ (he writes $\mathrm{d}\mathcal{T}_p$ in (9.31)). It does not depend on how the representatives $\hat{g}_t$ were chosen: changing the choice changes $\partial_i \hat{g}_t$ only by gauge directions, which Π kills.

Worked example 9.47: The moduli block for the torus

For $\hat{g}_\tau$ of (9.42), with $t = (\tau_1, \tau_2)$: the tangent vectors $\partial_{\tau_i}\hat{g}_\tau$ are constant on the torus and traceless, so $\operatorname{tr}(\partial_i \hat{g}) = 0$, $W_i = 0$ and $\Pi\partial_i \hat{g} = \partial_i \hat{g}$. The matrix (9.76) is then block *diagonal* in its last column, and

$$\sqrt{\det G^{\text{wp}}}\;\mathrm{d}\tau_1\mathrm{d}\tau_2 = \sqrt{\det\begin{pmatrix} 2/\tau_2^2 & 0 \\ 0 & 2/\tau_2^2 \end{pmatrix}}\,\mathrm{d}\tau_1\mathrm{d}\tau_2 = \frac{2\,\mathrm{d}\tau_1\,\mathrm{d}\tau_2}{\tau_2^2},$$

using (9.70). This is the hyperbolic area element (times 2).

The normalising factor N. The flat measures $\mathcal{D}\sigma$ and $\mathcal{D}V$ are defined by the metrics $2\int \sqrt{g}\,(\delta\sigma)^2$ and $\int \sqrt{g}\, g_{ab}\delta V^a \delta V^b$, and *both depend on the metric* g, through $\sqrt{g}$ and g_{ab}. When we move off the slice by a Weyl rescaling, these measures change. That change is what Nash hides in $N(f, P_1 V, \mathcal{T}_p)$. It is not innocent: it is one of the two sources (with the Weyl dependence of the determinants) of the conformal anomaly of Section 9.2.11. With $d = 26$ all the σ-dependence cancels, and N can be taken to be a constant, which we drop.

9.3.9 Dividing out the gauge volumes

We can now do the integral. Write the integrand of (9.28) with $d = 26$ as

$$F(g) = \left(\frac{\det'(\Delta_g/2)}{(1,1)_g}\right)^{-13}.$$

It is invariant under diffeomorphisms (the spectrum of Δ_g and the area do not change when g is pulled back by a diffeomorphism), and, for $d = 26$, the combination of F with the measure is invariant under Weyl rescalings. Take first $p \geq 2$, where $\operatorname{Ker} P_1 = 0$ and Diff_0 acts freely.

$$
\begin{aligned}
Z_p &\overset{(1)}{=} \frac{1}{\operatorname{vol}(C_+^\infty)\,\operatorname{vol}(\mathrm{Diff}^+)} \int_{\mathrm{Met}(\Sigma)} \mathcal{D}\mathrm{Met}\; F(g) \\
&\overset{(2)}{=} \frac{1}{\operatorname{vol}(C_+^\infty)\,\operatorname{vol}(\mathrm{Diff}^+)} \int \mathcal{D}\sigma \int_{\mathrm{Diff}_0} \mathcal{D}V \int_{\mathcal{T}_p} \mathrm{d}\mu(\mathcal{T}_p)\, \sqrt{\det P_1^\dagger P_1}\; F(\hat{g}_t) \\
&\overset{(3)}{=} \frac{\operatorname{vol}(C_+^\infty)\,\operatorname{vol}(\mathrm{Diff}_0)}{\operatorname{vol}(C_+^\infty)\,\operatorname{vol}(\mathrm{Diff}^+)} \int_{\mathcal{T}_p} \mathrm{d}\mu(\mathcal{T}_p)\, \sqrt{\det P_1^\dagger P_1}\; F(\hat{g}_t) \\
&\overset{(4)}{=} \frac{1}{|\Gamma_\Sigma|} \int_{\mathcal{T}_p} \mathrm{d}\mu(\mathcal{T}_p) \left(\frac{\det{}'(\Delta_g/2)}{(1,1)_g} \right)^{-13} \sqrt{\det P_1^\dagger P_1} \\
&\overset{(5)}{=} \int_{\mathcal{M}_p} \mathrm{d}\mu(\mathcal{M}_p) \left(\frac{\det{}'(\Delta_g/2)}{(1,1)_g} \right)^{-13} \sqrt{\det P_1^\dagger P_1}. \qquad (9.78)
\end{aligned}
$$

(1) Nash's definition (9.34), which repeats (9.8) = (9.35).
(2) Change variables to (σ, α, t) with the triangular Jacobian of Section 9.3.8; by gauge invariance the integrand and the Jacobian at $\alpha^*(e^\sigma \hat{g}_t)$ equal those at $\hat{g}_t$; the map $(\sigma, \alpha, t) \mapsto g$ is one-to-one for $p \geq 2$.
(3) Nothing depends on σ or α any more, so these integrals give the volumes of the Weyl group and of Diff_0.
(4) Diff^+ is the disjoint union of $|\Gamma_\Sigma|$ cosets of Diff_0, one for each element of the mapping class group (9.36), each with the same volume; so $\operatorname{vol}(\mathrm{Diff}^+) = |\Gamma_\Sigma| \operatorname{vol}(\mathrm{Diff}_0)$. Write g for $\hat{g}_t$.
(5) The integrand is Diff^+-invariant, hence constant on Γ_Σ-orbits in $\mathcal{T}_p$, so the integral over $\mathcal{T}_p$ is $|\Gamma_\Sigma|$ copies of the integral over a fundamental domain $\cong \mathcal{M}_p$ (the pitfall at the end of Section 9.3.4).

Every infinite volume has cancelled between numerator and denominator, and what is left is an integral over a space of dimension $6p - 6$.

Physical picture: Where each factor came from

Look at (9.78) factor by factor. $(\det{}'\Delta_g/(1,1)_g)^{-13}$ is the Gaussian integral over the 26 embedding fields, with their zero modes. $\sqrt{\det P_1^\dagger P_1}$ is the "speed" with which diffeomorphisms move metrics off the slice, the Faddeev–Popov determinant, the r of polar coordinates. $\mathrm{d}\mu(\mathcal{T}_p)$ is the volume of the slice itself, measured with the only metric available, and $1/|\Gamma_\Sigma|$ undoes the overcounting of metrics related by diffeomorphisms that cannot be deformed to the identity.

9.3.10 When P_1 has a kernel: conformal Killing vectors

For the sphere and the torus the argument breaks at one point: $\operatorname{Ker} P_1 \neq 0$. Nash puts it as follows. If $V \in \operatorname{Ker} P_1$ then

$$
P_1 V = 0 \quad \Longleftrightarrow \quad \mathcal{L}_V g = f\, g, \qquad f = \tfrac{1}{2} \operatorname{tr}(\mathcal{L}_V g) = \nabla_c V^c, \qquad (9.79)
$$

which is (9.49) with vanishing traceless part: the flow of V changes g only by a Weyl factor. Such V are the conformal Killing vectors (CKVs) of Section 9.2.5. They generate a finite-dimensional group, the **conformal Killing group** $\mathrm{CKG}(g)$: $\mathrm{PSL}(2, \mathbb{C})$ for the round sphere, the translations $U(1) \times U(1)$ for a flat torus.

Two things go wrong. First, $P_1^\dagger P_1$ has zero modes, so $\det P_1^\dagger P_1 = 0$ and (9.74) is useless as it stands. Second, the map $(\sigma, \alpha, t) \mapsto g$ is no longer one-to-one: if $\beta \in \mathrm{CKG}$, then $\beta^* \hat{g}_t = e^{\sigma_\beta} \hat{g}_t$

for some function σ_β, and (σ, α, t) and $(\sigma', \beta \circ \alpha, t)$ (with a suitably shifted σ') give the same metric. Each metric is hit once for every element of CKG.

Nash's cure (9.33) is to split the Lie algebra orthogonally,

$$\operatorname{vect}(\Sigma) = T_e \operatorname{Diff}_0(\Sigma) = \operatorname{Ker} P_1 \oplus \operatorname{Ker}^{\perp} P_1, \tag{9.80}$$

using $\langle\ ,\ \rangle_{\mathrm{vect}}$, and to treat the two pieces separately: on $\operatorname{Ker}^{\perp} P_1$ the operator is injective and (9.74) holds with the determinant restricted to non-zero modes; the $\operatorname{Ker} P_1$ directions are integrated separately and give the volume of CKG. The finite-dimensional model is a group acting with a stabiliser.

Worked example 9.48: Dividing by $\mathrm{SO}(3)$: a stabiliser appears

Let $F(x) = F(|x|)$ on $\mathbb{R}^3$ and divide by $G = \mathrm{SO}(3)$. Normalise the metric on G so that each rotation by angle θ about a fixed axis is a path of length θ; then $\operatorname{vol}\mathrm{SO}(3) = 8\pi^2$ and the stabiliser of a point, the rotations about one axis, has $\operatorname{vol}\mathrm{SO}(2) = 2\pi$. The slice is the positive x^3-axis, $x = (0, 0, r)$. The infinitesimal action is $A : \mathfrak{so}(3) = \mathbb{R}^3 \to \mathbb{R}^3$, $A(\omega) = \omega \times x$:

$$A(\omega) \overset{(1)}{=} (\omega_2\, r,\ -\omega_1\, r,\ 0), \qquad A^{\mathsf{T}}A = \begin{pmatrix} r^2 & 0 & 0 \\ 0 & r^2 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

$$\sqrt{\det{}' A^{\mathsf{T}}A} \overset{(2)}{=} \sqrt{r^2 \cdot r^2} = r^2.$$

(1) $\omega \times (0, 0, r) = (\omega_2 r - \omega_3 \cdot 0,\ \omega_3 \cdot 0 - \omega_1 r,\ 0)$. Rotations about the x^3-axis (ω_3) do not move the point: $\operatorname{Ker} A$ is the Lie algebra of the stabiliser.

(2) $\det'$ is the product of the non-zero eigenvalues.

Now compare the two sides of the gauge-fixed formula:

$$\begin{aligned} \frac{1}{\operatorname{vol}\mathrm{SO}(3)} \int_{\mathbb{R}^3} \mathrm{d}^3x\, F(|x|) &\overset{(1)}{=} \frac{1}{8\pi^2} \int_0^\infty 4\pi r^2 F(r)\, \mathrm{d}r \\ &\overset{(2)}{=} \frac{1}{2\pi} \int_0^\infty r^2 F(r)\, \mathrm{d}r \\ &\overset{(3)}{=} \frac{1}{\operatorname{vol}(\text{stabiliser})} \int_{\text{slice}} \mathrm{d}r\ \sqrt{\det{}' A^{\mathsf{T}}A}\ F(r). \end{aligned}$$

(1) Spherical coordinates: the orbit through $(0, 0, r)$ is a sphere of area $4\pi r^2$.

(2) $4\pi/8\pi^2 = 1/2\pi$.

(3) 2π is the volume of the stabiliser $\mathrm{SO}(2)$, and $r^2 = \sqrt{\det{}' A^{\mathsf{T}}A}$.

So when the group has a stabiliser, the gauge-fixed integral contains the *primed* Faddeev–Popov determinant and a factor $1/\operatorname{vol}(\text{stabiliser})$.

For strings the stabiliser is $\mathrm{CKG}(\hat{g}_t)$, its Lie algebra is $\operatorname{Ker} P_1$, and the same bookkeeping gives, for $p = 0, 1$,

$$Z_p = \frac{1}{|\Gamma_\Sigma|} \int_{\mathcal{T}_p} \mathrm{d}\mu(\mathcal{T}_p)\ \frac{1}{\operatorname{vol}\mathrm{CKG}(\hat{g}_t)} \left(\frac{\det'(\Delta_g/2)}{(1,1)_g} \right)^{-13} \sqrt{\det{}' P_1^{\dagger} P_1}, \tag{9.81}$$

with the volume of CKG computed with the metric $\langle\ ,\ \rangle_{\mathrm{vect}}$ restricted to $\operatorname{Ker} P_1$. Nash writes $\tilde{A}_g$ for "P_1 restricted to $\operatorname{Ker}^{\perp} P_1$", so that $\det(\tilde{A}_g^{*}\tilde{A}_g) = \det' P_1^{\dagger} P_1$, and absorbs the volume factor into the measure.

Worked example 9.49: The conformal Killing group of a flat torus has volume $(1,1)_g$

On $\mathbb{R}^2/\mathbb{Z}^2$ with the flat metric $\hat{g}_\tau$, the CKVs are the constant vector fields $V = v^1\partial_x + v^2\partial_y$ (Example 9.39), and they generate the translations $(x,y) \mapsto (x+v^1, y+v^2)$. Translating by $v \in \mathbb{Z}^2$ is the identity map, so the group is $\{v\}/\mathbb{Z}^2$, a torus of parameters $v \in [0,1)^2$. Its volume in the metric $\langle V, V\rangle_{\text{vect}}$ is

$$\begin{aligned}\operatorname{vol}\mathrm{CKG} &\overset{(1)}{=} \int_{[0,1)^2} \mathrm{d}^2v\, \sqrt{\det\big(\langle\partial_{v^a}, \partial_{v^b}\rangle_{\text{vect}}\big)} \\ &\overset{(2)}{=} \int_{[0,1)^2} \mathrm{d}^2v\, \sqrt{\det\Big((1,1)_{\hat{g}}\,\hat{g}_{ab}\Big)} \\ &\overset{(3)}{=} (1,1)_{\hat{g}}\,\sqrt{\det\hat{g}_{ab}} = 1.\end{aligned}$$

(1) Riemannian volume (9.40) on the group, with coordinates v^a.

(2) For constant V, $\langle V, W\rangle_{\text{vect}} = \int \sqrt{\hat{g}}\,\hat{g}_{ab}V^aW^b = (1,1)_{\hat{g}}\,\hat{g}_{ab}V^aW^b$, since $\hat{g}_{ab}$ is constant.

(3) det of the 2×2 matrix $(1,1)\hat{g}_{ab}$ is $(1,1)^2 \det\hat{g}$; and $(1,1)_{\hat{g}} = 1 = \det\hat{g}_{ab}$ for $\hat{g}_\tau$.

For a torus of area $(1,1)_g$ (i.e. $g = c\,\hat{g}_\tau$, $c = (1,1)_g$) the same computation gives $\operatorname{vol}\mathrm{CKG} = c\cdot c = (1,1)_g^2$ up to the normalisation of $\mathcal{D}V$. In the slice $\mathrm{Met}_{\text{const}}$ we have fixed the area to 1, so for the torus the CKG factor is simply 1.

Watch out: The sphere has an infinite conformal Killing group

For $p = 0$, CKG = PSL$(2,\mathbb{C})$, a non-compact six-dimensional group. Its volume is infinite, so (9.81) says $Z_0 = 0$ (or, more honestly, is undefined). This is not a mistake: the sphere with no insertions has no moduli, and the residual gauge symmetry cannot be fixed. Physically meaningful sphere amplitudes have at least three vertex operators, whose positions absorb the six real CKV parameters (Example 9.58). Nash's formula is written for all p, but it is the genus $p \geq 1$ terms that make sense as they stand.

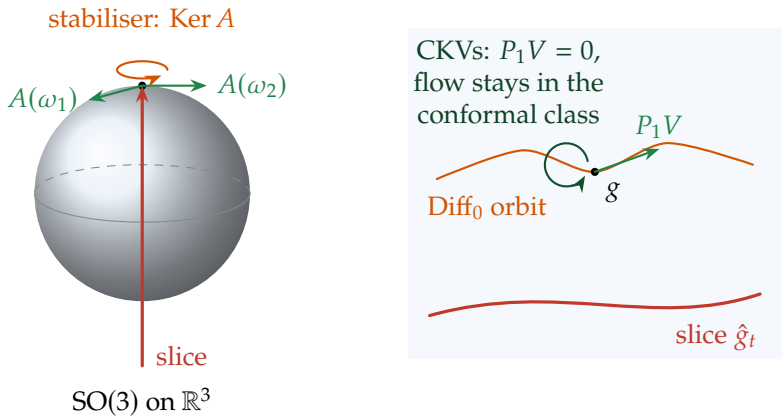


Figure 9.21: Stabilisers. Left: rotations about the vertical axis do not move the north pole; they span Ker A and must be integrated separately, giving $1/\mathrm{vol}(\mathrm{SO}(2))$. Right: for the sphere and the torus, the flows of conformal Killing vectors move g only inside its Weyl orbit (they are not seen by P_1); they span Ker P_1, and the path integral picks up $1/\operatorname{vol}\mathrm{CKG}$ and a primed determinant.

9.3.11 The Faddeev–Popov determinant as a ghost integral

Physicists rarely write $\sqrt{\det{}' P_1^\dagger P_1}$ as it stands. They write it as a Gaussian integral over anticommuting fields, the Faddeev–Popov **ghosts**. This makes the determinant look like one more field theory living on the world sheet, which is what allows it to be treated with the same tools (conformal field theory, Chapter 11) as the matter fields. Here is why it works.

Grassmann integration in one line. Anticommuting variables $\theta_i, \bar\theta_i$ satisfy $\theta_i\theta_j = -\theta_j\theta_i$ (so $\theta_i^2 = 0$), and the Berezin integral is defined by $\int \mathrm{d}\theta\, 1 = 0$, $\int \mathrm{d}\theta\, \theta = 1$, and likewise for $\bar\theta$. For one pair and a number λ,

$$\int \mathrm{d}\bar\theta\, \mathrm{d}\theta\; e^{-\bar\theta\lambda\theta} \overset{(1)}{=} \int \mathrm{d}\bar\theta\, \mathrm{d}\theta\; (1 - \lambda\,\bar\theta\theta) \overset{(2)}{=} \int \mathrm{d}\bar\theta\, \mathrm{d}\theta\; (1 + \lambda\,\theta\bar\theta) \overset{(3)}{=} \lambda.$$

(1) The power series of the exponential stops after the linear term, because $(\bar\theta\theta)^2 = \bar\theta\theta\bar\theta\theta = -\bar\theta\bar\theta\theta\theta = 0$.
(2) Anticommute: $\bar\theta\theta = -\theta\bar\theta$.
(3) $\int \mathrm{d}\theta\, \theta = 1$ does the inner integral, leaving $\int \mathrm{d}\bar\theta\, \bar\theta = 1$; the constant 1 integrates to zero.

Compare the ordinary Gaussian $\int \mathrm{d}\bar z\, \mathrm{d}z\, e^{-\bar z \lambda z} \propto \lambda^{-1}$: the Grassmann integral gives the determinant in the *numerator*.

Worked example 9.50: The 2×2 Grassmann Gaussian

Let M be a 2×2 matrix and $S = \bar\theta_i M_{ij} \theta_j$. Only the term of the exponential containing all four generators survives the integral, and it comes from $\frac{1}{2}S^2$. Its coefficient is

$$\begin{aligned} \tfrac{1}{2}S^2 &\overset{(1)}{=} \tfrac{1}{2} \sum_{i,j,k,l} M_{ij}M_{kl}\, \bar\theta_i\theta_j\bar\theta_k\theta_l \\ &\overset{(2)}{=} \left(M_{11}M_{22}\, \bar\theta_1\theta_1\bar\theta_2\theta_2 + M_{12}M_{21}\, \bar\theta_1\theta_2\bar\theta_2\theta_1\right) + \cdots \\ &\overset{(3)}{=} \left(M_{11}M_{22} - M_{12}M_{21}\right) \bar\theta_1\theta_1\bar\theta_2\theta_2 + \cdots = \det M\; \bar\theta_1\theta_1\bar\theta_2\theta_2 + \cdots . \end{aligned}$$

(1) Square the sum.
(2) Keep terms with $\{i,k\} = \{1,2\}$ and $\{j,l\} = \{1,2\}$. The terms with $i = 2, k = 1$ equal those with $i = 1, k = 2$ (the bilinears $\bar\theta_i\theta_j$ are even, so they commute), which cancels the $\frac{1}{2}$. The dots are terms with a repeated generator, which vanish.
(3) Move θ_1 two places to the left (two sign changes), then swap $\theta_2\bar\theta_2 = -\bar\theta_2\theta_2$: one net minus sign.

With $\int \mathrm{d}\bar\theta_1\mathrm{d}\theta_1\mathrm{d}\bar\theta_2\mathrm{d}\theta_2\; \bar\theta_1\theta_1\bar\theta_2\theta_2 = 1$ (the normalisation fixed by the one-pair case), the integral is $\det M$. The same bookkeeping gives $\det M$ for any $n\times n$ matrix.

Key idea: The b, c ghosts

Introduce an anticommuting vector field c^a (the **c-ghost**, living in the space $\mathrm{vect}(\Sigma)$ of the gauge parameters) and an anticommuting traceless symmetric tensor b_{ab} (the **b-ghost**,

living in $C_g^{\perp}$ where the gauge directions point). Then, formally,

$$\sqrt{\det{}' P_1^\dagger P_1} = \int \mathcal{D}b\,\mathcal{D}c\ \exp\Big[-\frac{1}{4\pi}\int_\Sigma \mathrm{d}^2x\,\sqrt{g}\ b^{ab}(P_1 c)_{ab}\Big] \quad \text{(non-zero modes only)}, \tag{9.82}$$

up to a constant. The exponent is $\langle b, P_1 c\rangle_{\text{met}}$: in an orthonormal basis of $\mathrm{Ker}^{\perp} P_1$ and a matching one of $\mathrm{Range}\, P_1$, P_1 is a square matrix whose determinant has absolute value $\sqrt{\det P_1^\dagger P_1}$. The normalisation $\frac{1}{4\pi}$ is the conventional one; it only rescales the result by a constant. In complex coordinates the ghost action is $\frac{1}{2\pi}\int \mathrm{d}^2z\,(b_{zz}\partial_{\bar z}c^z + \text{c.c.})$, which is the first-order "bc system" of conformal field theory (Chapter 11).

Physical picture: Ghost zero modes are CKVs and moduli

A Grassmann integral over a variable that does not appear in the exponent is zero: $\int \mathrm{d}\theta\, 1 = 0$. So if we did not remove the zero modes, the ghost integral would vanish. The zero modes of c are the solutions of $P_1 c = 0$, the CKVs; those of b are the solutions of $P_1^\dagger b = 0$, the holomorphic quadratic differentials, i.e. the moduli directions. In the ghost language, the c zero modes are soaked up by fixing the positions of vertex operators (Example 9.58), and the b zero modes are soaked up by the moduli integral: pairing b with the moduli deformations $\partial_i \hat g_t$ reproduces the Weil–Petersson (or rather the equivalent Gram-determinant) factor. This is the dictionary row "$\mathrm{Ker}\, A_g, \mathrm{Ker}\, A_g^* \leftrightarrow$ ghost zero modes" of Section 9.2.7, and it is the reason why the index $\dim\mathrm{Ker}\, P_1 - \dim\mathrm{Ker}\, P_1^\dagger = 3\chi$ counts the "ghost number anomaly".

9.3.12 Real versus complex determinants: Nash's (9.36)

Nash ends with a remark: P_1 ($= A_g$) is a *real* operator, but in terms of the $\bar\partial$-operators of Chapter 5,

$$\det\big(\tilde A_g^* \tilde A_g\big) = \big(\det(\bar\partial_{K_\Sigma}^* \bar\partial_{K_\Sigma^*})\big)^2, \tag{9.83}$$

so that $\sqrt{\det \tilde A_g^* \tilde A_g} = \det \bar\partial_{K^*}^* \bar\partial_{K^*}$. Here K_Σ is the canonical bundle (the holomorphic cotangent bundle, spanned locally by $\mathrm{d}z$), and $K_\Sigma^* = K_\Sigma^{-1}$ is its dual, the holomorphic tangent bundle, spanned by ∂_z. Two facts combine to give (9.83).

Fact 1: P_1 is complex-linear. Identify a real vector field $V = V^a\partial_a$ with its holomorphic component V^z, a section of K^{-1} (the real vector field is $V^z\partial_z + \overline{V^z}\partial_{\bar z}$). Identify a traceless tensor h with its component $h_{\bar z\bar z}$. By (9.55), $(P_1 V)_{\bar z\bar z} = \rho\,\partial_{\bar z} V^z$: in these complex variables P_1 is ρ times $\bar\partial_{K^{-1}}$, and multiplying V^z by i (rotating the vector field by 90°) multiplies $P_1 V$ by i. So P_1, and with it $P_1^\dagger P_1$, is a complex-linear operator on the complex line bundle K^{-1}.

Fact 2: a complex operator has real determinant $|\det|^2$.

Worked example 9.51: Real determinant of a complex matrix

One dimension. Multiplication by $\lambda = a + ib$ on $\mathbb{C} \cong \mathbb{R}^2$ sends $x + iy$ to $(ax - by) + i(bx + ay)$, so its real matrix and determinant are

$$\begin{pmatrix} a & -b \\ b & a \end{pmatrix}, \qquad \det_{\mathbb{R}} = a^2 + b^2 = |\lambda|^2 = |\det_{\mathbb{C}}\lambda|^2 .$$

n dimensions. Triangularise a complex $n \times n$ matrix L by a unitary change of basis (Schur): $L = UTU^{-1}$ with T upper triangular, diagonal entries $\lambda_1, \dots, \lambda_n$. As a real $2n \times 2n$ matrix, T is block upper triangular with the 2×2 blocks above on the diagonal, and U becomes a rotation of $\mathbb{R}^{2n}$ (determinant 1). Hence

$$\det_{\mathbb{R}} L \overset{(1)}{=} \det_{\mathbb{R}} T \overset{(2)}{=} \prod_{k=1}^{n} |\lambda_k|^2 \overset{(3)}{=} |\det_{\mathbb{C}} L|^2 .$$

(1) Similar matrices have equal determinants.
(2) A block triangular matrix has determinant the product of the determinants of its diagonal blocks, each $|\lambda_k|^2$.
(3) $\det_{\mathbb{C}} L = \prod_k \lambda_k$.

If L is Hermitian and positive, $\det_{\mathbb{C}} L > 0$ and $\det_{\mathbb{R}} L = (\det_{\mathbb{C}} L)^2$.

Apply this to the positive self-adjoint complex operator $P_1^\dagger P_1$ on $\operatorname{Ker}^\perp P_1$ (every eigenvalue appears twice as a real operator: V^z and iV^z are real-independent eigenvectors with the same eigenvalue). The zeta functions obey $\zeta^{\mathbb{R}}(s) = 2\zeta^{\mathbb{C}}(s)$, so $\det'_{\mathbb{R}} = \exp(-2\zeta^{\mathbb{C}\prime}(0)) = (\det'_{\mathbb{C}})^2$. That is Nash's statement, with $\bar\partial^*_{K^*}\bar\partial_{K^*}$ standing for the complex operator $P_1^\dagger P_1$ (up to the factor ρ and a constant, which rescale the determinant in a controlled way, Example 9.26).

Worked example 9.52: Checking on the flat square torus

Take $\hat g = \delta$ on $\mathbb{R}^2/\mathbb{Z}^2$ ($\rho = 1$). By Example 9.39, $P_1^\dagger P_1 = -2\nabla^2 = 2\Delta$ acting on each Cartesian component of V. On the Fourier mode $V^z = e^{2\pi i(nx+my)}$,

$$P_1^\dagger P_1 V^z \overset{(1)}{=} 2\Delta\, e^{2\pi i(nx+my)} \overset{(2)}{=} 8\pi^2(n^2+m^2)\, V^z .$$

(1) Flat metric, $K = 0$, so $P_1^\dagger P_1 = -2(\partial_x^2 + \partial_y^2)$ on components.
(2) $-\partial_x^2 - \partial_y^2$ multiplies the mode by $4\pi^2(n^2+m^2)$.

On the other hand, $\bar\partial^\dagger\bar\partial$ on functions, with $\langle f, f'\rangle = \int f\bar f'$: integrating by parts, $\int (\partial_{\bar z} f)\,\overline{f'} = -\int f\,\partial_{\bar z}\overline{f'} = -\int f\,\overline{\partial_z f'}$, so $\bar\partial^\dagger = -\partial_z$ and

$$\bar\partial^\dagger\bar\partial \overset{(1)}{=} -\partial_z\partial_{\bar z} \overset{(2)}{=} -\tfrac14(\partial_x^2+\partial_y^2) = \tfrac14\Delta.$$

(1) The adjoint just computed.
(2) $\partial_z\partial_{\bar z} = \frac14(\partial_x - i\partial_y)(\partial_x + i\partial_y) = \frac14(\partial_x^2+\partial_y^2)$.

So on the flat torus $P_1^\dagger P_1 = 8\,\bar\partial^\dagger\bar\partial$ as complex operators, and the complex eigenvalues $8\pi^2(n^2+m^2)$, $(n,m) \neq (0,0)$, each appear once in the complex count and twice in the real count. The excluded mode $(0,0)$ is the complex one-dimensional (real two-dimensional) space of CKVs.

Key idea: The final formula (Nash's (9.35)–(9.36))

With $d = 26$,

$$Z = \sum_p Z_p = \sum_p \int_{\mathcal{M}_p} \mathrm{d}\mu(\mathcal{M}_p) \left(\frac{\det'(\Delta_g/2)}{(1,1)_g} \right)^{-13} \det'\big(\bar\partial^*_{K^*_\Sigma} \bar\partial_{K^*_\Sigma}\big), \tag{9.84}$$

where g is the constant-curvature metric representing the point of $\mathcal{M}_p$, $\mathrm{d}\mu(\mathcal{M}_p)$ is the Weil–Petersson volume element, and for $p = 0, 1$ the determinant is restricted to the orthogonal complement of the conformal Killing vectors (with the factor $1/\mathrm{vol}\,\mathrm{CKG}$ understood). The integral over all surfaces and all embeddings has become, genus by genus, an integral over a finite-dimensional space of dimension $6p - 6$ (or 2 for the torus), of a function built from determinants of two elliptic operators.

Watch out: What (9.36) does and does not say

Nash writes det where we write det′: both determinants omit zero modes. The measure $\mathrm{d}\mu(\mathcal{M}_p)$ in (9.36) is the same as in (9.35); only the ghost factor has been rewritten. And the formula leaves out the string coupling: if the action includes the Euler characteristic term of Section 9.1.5, each Z_p is weighted by g_s^{2p-2}, and Z is the perturbation series of closed string theory. Finally, (9.36) is a *formal* result for $p \geq 2$: the integrals over $\mathcal{M}_p$ diverge near its boundary (where the surface degenerates), which is the string version of the infrared divergences caused by the tachyon, as the torus computation below shows.

9.3.13 The formula at work: the torus amplitude

Nash stops at (9.36). The formula looks forbidding: an integral over a moduli space of a ratio of zeta-regularised determinants. Is it ever a number? For the torus, yes, and computing it is the best way to see that every piece of the chapter does real work. The question for this subsection is: *what is Z_1, explicitly, as a function on the fundamental domain of Figure 9.19?* We will find

$$Z_1 = 2^{-13} \int_{\mathcal{F}} \frac{\mathrm{d}\tau_1\,\mathrm{d}\tau_2}{\tau_2^2}\, \tau_2^{-12}\, |\eta(\tau)|^{-48}, \tag{9.85}$$

where η is the Dedekind eta function defined below. On the way we meet the string's Casimir energy $-\frac{1}{24}$, the tachyon, and the reason string theory has no ultraviolet divergences.

The ingredients, genus by genus. Specialise (9.81) to $p = 1$ on the slice $\hat g_\tau$ of (9.42) (flat, area 1). We need four things:

(a) the measure $\mathrm{d}\mu(\mathcal{T}_1) = 2\,\mathrm{d}\tau_1\mathrm{d}\tau_2/\tau_2^2$ (Example 9.47);

(b) the conformal Killing group volume, which is 1 on the area-1 slice (Example 9.49);

(c) the matter factor $\big(\det'(\Delta_{\hat g_\tau}/2)/(1,1)_{\hat g_\tau}\big)^{-13}$;

(d) the ghost factor $\sqrt{\det' P_1^\dagger P_1}$.

The first two are done. The last two both reduce to a single number: the zeta-regularised determinant of the scalar Laplacian on a flat torus.

Step 1: the spectrum of the flat torus

Worked example 9.53: Eigenvalues of Δ on the torus $\hat{g}_\tau$

The torus is $\mathbb{R}^2/\mathbb{Z}^2$ with coordinates $(x, y) \sim (x+1, y) \sim (x, y+1)$ and the constant metric $\hat{g}_\tau$. The inverse of the matrix (9.42) is

$$\hat{g}_\tau^{-1} = \frac{1}{\tau_2}\begin{pmatrix} |\tau|^2 & -\tau_1 \\ -\tau_1 & 1 \end{pmatrix},$$

as one checks by multiplying: the product is $\tau_2^{-2}\,\mathrm{diag}(|\tau|^2-\tau_1^2, |\tau|^2-\tau_1^2) = \mathbf{1}$. Because the metric is constant, the Laplacian is $\Delta = -\hat{g}^{ab}\partial_a\partial_b$ (no Christoffel symbols), and the periodic functions $f_{n,m} = e^{2\pi i(nx+my)}$, $(n,m) \in \mathbb{Z}^2$, are eigenfunctions:

$$\begin{aligned} \Delta f_{n,m} &\overset{(1)}{=} -\hat{g}^{ab}\,(2\pi i k_a)(2\pi i k_b)\, f_{n,m}, \qquad k = (n,m), \\ &\overset{(2)}{=} \frac{4\pi^2}{\tau_2}\big(|\tau|^2 n^2 - 2\tau_1 nm + m^2\big) f_{n,m} \\ &\overset{(3)}{=} \frac{4\pi^2}{\tau_2}\,|m - n\tau|^2\, f_{n,m}. \end{aligned} \tag{9.86}$$

(1) Each derivative ∂_a brings down $2\pi i k_a$.
(2) Contract with the inverse metric displayed above.
(3) $|m-n\tau|^2 = (m-n\tau_1)^2 + n^2\tau_2^2 = m^2 - 2mn\tau_1 + n^2|\tau|^2$.

The Fourier modes are complete, so this is the whole spectrum. Relabelling $n \to -n$, the non-zero eigenvalues are

$$\lambda_{n,m} = \frac{4\pi^2}{\tau_2}\,|m + n\tau|^2\,, \qquad (n,m) \in \mathbb{Z}^2 \setminus \{(0,0)\}, \tag{9.87}$$

each with multiplicity one, and there is a single zero mode (the constants). *Sanity checks.* (i) At $\tau = i$ (square torus, $\hat{g} = \delta$), $\lambda = 4\pi^2(n^2+m^2)$, the familiar spectrum of $-\partial_x^2 - \partial_y^2$. (ii) Under $\tau \mapsto \tau + 1$ the set $\{m + n\tau\}$ is unchanged (replace m by $m-n$), and under $\tau \mapsto -1/\tau$ we get $|m - n/\tau|^2 / (\tau_2/|\tau|^2) = |m\tau - n|^2/\tau_2$, the same set again. The spectrum is modular invariant, as it must be: modular transformations are diffeomorphisms (Section 9.3.4), and isometric manifolds have the same spectrum.

Step 2: one row at a time, a twisted massive circle

Fix n. The row of eigenvalues is

$$\lambda_{n,m} = \frac{4\pi^2}{\tau_2}\Big[(m+b)^2 + c^2\Big], \qquad b = n\tau_1, \quad c = |n|\,\tau_2, \qquad m \in \mathbb{Z}.$$

For $n = 0$ this is ($m \neq 0$) the spectrum $(2\pi m/L)^2$ of the Laplacian on a circle of length $L = \sqrt{\tau_2}$. For $n \neq 0$ it is a circle Laplacian with a mass c and a shift b of the momentum. The shift is the Dehn twist τ_1 of the torus seen from the row: going once around the y-cycle, the x-coordinate is shifted by τ_1.

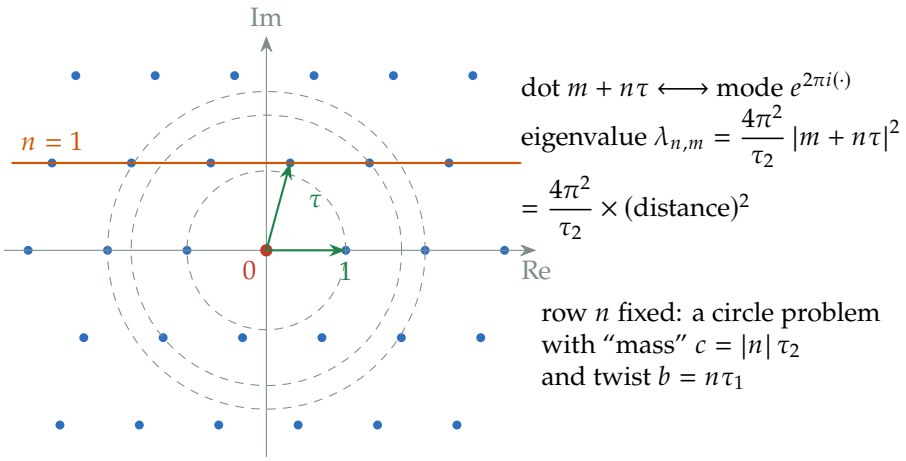


Figure 9.22: The spectrum of the flat torus is a lattice. Each dot $m + n\tau$ of the lattice $\mathbb{Z} + \tau\mathbb{Z}$ (blue) is one eigenfunction, and its eigenvalue is $4\pi^2/\tau_2$ times its squared distance from the origin; the dashed circles are eigenvalue shells. The red origin is the zero mode, dropped from det′. We compute the determinant row by row: each horizontal row (orange, fixed n) is the spectrum of a massive, twisted circle.

Worked example 9.54: The twisted massive circle: $\prod_m \left((m+b)^2 + c^2\right) = 4\,|\sin \pi(b + ic)|^2$

Let $D(b, c)$ be the zeta-regularised product of $(m + b)^2 + c^2$ over $m \in \mathbb{Z}$, with $c > 0$. We compute it in three moves.

Move 1: the untwisted case. For $b = 0$ this is the oscillator of Example 9.24 with $\beta = 2\pi$ and $\omega = c$: the eigenvalues $(2\pi m/\beta)^2 + \omega^2$ become $m^2 + c^2$, so $D(0, c) = 4\sinh^2(\omega\beta/2) = 4\sinh^2(\pi c)$.

Move 2: the twist as a convergent ratio. Write $w = b + ic$, so that $(m + b)^2 + c^2 = |m + w|^2$ and $m^2 + c^2 = |m + ic|^2$. The ratio of the two products converges once m and $-m$ are paired, so (as in Example 9.24)

$$\begin{aligned}\frac{D(b,c)}{D(0,c)} &\overset{(1)}{=} \frac{|w|^2}{c^2}\prod_{m\geq 1}\frac{|(w+m)(w-m)|^2}{|(ic+m)(ic-m)|^2}\\ &\overset{(2)}{=} \frac{|w|^2\prod_{m\geq 1}\left|1 - w^2/m^2\right|^2}{c^2\prod_{m\geq 1}\left(1 + c^2/m^2\right)^2}\\ &\overset{(3)}{=} \frac{|\sin \pi w|^2/\pi^2}{\sinh^2(\pi c)/\pi^2} = \frac{|\sin \pi w|^2}{\sinh^2(\pi c)}.\end{aligned}$$

(1) Separate the $m = 0$ factor $|w|^2/|ic|^2$, and pair m with $-m$ for $m \geq 1$.

(2) $(w + m)(w - m) = w^2 - m^2 = -m^2(1 - w^2/m^2)$ and $(ic + m)(ic - m) = -m^2(1 + c^2/m^2)$; the factors m^4 cancel between numerator and denominator.

(3) Euler's product $\sin \pi w = \pi w \prod_{m\geq 1}(1 - w^2/m^2)$, used at w and at $w = ic$, where $\sin(i\pi c) = i\sinh(\pi c)$.

Move 3: combine and rewrite. Multiplying, $D(b, c) = 4\,|\sin \pi w|^2$, and

$$\begin{aligned}D(b,c) &\overset{(1)}{=} \left|e^{i\pi w} - e^{-i\pi w}\right|^2\\ &\overset{(2)}{=} \left|e^{-i\pi w}\right|^2\,\left|1 - e^{2\pi i w}\right|^2\\ &\overset{(3)}{=} e^{2\pi c}\,\left|1 - e^{2\pi i(b+ic)}\right|^2.\end{aligned} \tag{9.88}$$

(1) $2i\sin\pi w = e^{i\pi w} - e^{-i\pi w}$, and $|2i|^2 = 4$.
(2) Factor out $e^{-i\pi w}$.
(3) $\left|e^{-i\pi(b+ic)}\right| = e^{\pi c}$.

The overall constant. The row eigenvalues carry the prefactor $a = 4\pi^2/\tau_2$. By the scaling law of Example 9.26, it contributes $a^{\zeta_{\mathrm{row}}(0)}$, and $\zeta_{\mathrm{row}}(0) = 0$ for a massive circle: the heat trace $\sum_m e^{-t((m+b)^2+c^2)} = \sqrt{\pi/t}\, e^{-tc^2} +$ (exponentially small) (Poisson summation) has no t^0 term, and by (9.34) the value $\zeta(0)$ is exactly the t^0 coefficient (minus the number of zero modes, here none). This is also why the oscillator answer did not depend on β through the prefactor $(2\pi/\beta)^2$. So a massive row contributes $D(b,c)$ with no extra constant.

Physical picture: What the twisted circle computes

$D(b,c)^{-1/2} = e^{-\pi c}\left|1 - e^{2\pi i(b+ic)}\right|^{-1}$ is the modulus of a *twisted* oscillator trace. For an oscillator of frequency ω with a phase $e^{i\theta}$ inserted per quantum,

$$\mathrm{Tr}\left(e^{-\beta H}e^{i\theta N}\right) = \sum_{k\geq 0} e^{-\beta\omega(k+\frac{1}{2})}e^{i\theta k} = \frac{e^{-\beta\omega/2}}{1 - e^{-\beta\omega+i\theta}},$$

and with $\beta\omega = 2\pi c$, $\theta = 2\pi b$ its modulus is exactly $D(b,c)^{-1/2}$. So the twist b of the torus becomes a phase per quantum. In the string, the rows n and $-n$ are the left- and right-moving oscillators of frequency n, and the Dehn twist τ_1 enters the Hamiltonian trace as the phase $e^{2\pi i\tau_1(L_0-\bar{L}_0)}$, a translation around the string.

Step 3: assemble the rows; the Dedekind eta function

Working definition 9.12 (Dedekind eta function)

For τ in the upper half-plane [$\tau = \tau_1 + i\tau_2$ with $\tau_2 > 0$, the shape of the flat torus] put $q = e^{2\pi i\tau}$ (so $|q| = e^{-2\pi\tau_2} < 1$) and

$$\eta(\tau) = q^{1/24}\prod_{n=1}^{\infty}(1-q^n), \qquad q^{1/24} := e^{2\pi i\tau/24}.$$

How to compute it: the product converges very fast; at $\tau = i$, $q = e^{-2\pi} \approx 0.00187$, and three factors already give $\eta(i) \approx 0.76823$. *Its two symmetry laws* [$\tau \to \tau + 1$ is the Dehn twist, $\tau \to -1/\tau$ exchanges the roles of the two basic loops of the torus], which we derive below from geometry: $\eta(\tau+1) = e^{i\pi/12}\eta(\tau)$ (read off from $q^{1/24}$), and $|\eta(-1/\tau)| = |\tau|^{1/2}\,|\eta(\tau)|$.

Worked example 9.55: $\det'\Delta_{\hat{g}_\tau} = \tau_2\,|\eta(\tau)|^4$

Multiply the rows of Figure 9.22. The row $n = 0$ (zero mode removed) is a circle of length $L = \sqrt{\tau_2}$, so it gives $L^2 = \tau_2$ by (9.32). For the row $n \geq 1$ we have $b + ic = n\tau_1 + in\tau_2 = n\tau$, so $e^{2\pi i(b+ic)} = q^n$; for the row $-n \leq -1$ we have $b + ic = -n\tau_1 + in\tau_2 = -n\bar{\tau}$, so $e^{2\pi i(b+ic)} = \bar{q}^{\,n}$. Hence

$$\det'\Delta_{\hat{g}_\tau} \overset{(1)}{=} \tau_2 \prod_{n\geq 1} D(n\tau_1, n\tau_2)\, D(-n\tau_1, n\tau_2)$$

$$\begin{aligned}
&\overset{(2)}{=} \tau_2 \prod_{n\geq 1} e^{2\pi n\tau_2}\,|1-q^n|^2 \cdot e^{2\pi n\tau_2}\,|1-\bar{q}^{\,n}|^2 \\
&\overset{(3)}{=} \tau_2\ \exp\Big(4\pi\tau_2 \sum_{n\geq 1} n\Big) \prod_{n\geq 1} |1-q^n|^4 \\
&\overset{(4)}{=} \tau_2\, e^{-\pi\tau_2/3} \prod_{n\geq 1} |1-q^n|^4 \\
&\overset{(5)}{=} \tau_2\ |\eta(\tau)|^4 \,. \qquad (9.89)
\end{aligned}$$

(1) The determinant is the product over rows; the $n=0$ row gives τ_2 and each massive row contributes $D(b,c)$ with no extra constant (Example 9.54).
(2) Formula (9.88) with $c = n\tau_2$ in both rows.
(3) $|1-\bar{q}^{\,n}| = |1-q^n|$, and collect the exponentials.
(4) Zeta-regularise the divergent sum: $\sum_{n\geq 1} n = \zeta_R(-1) = -\frac{1}{12}$ (the Riemann zeta value of (9.31)), so the exponent is $-\frac{4\pi\tau_2}{12}$.
(5) $|q|^{1/6} = e^{-2\pi\tau_2/6} = e^{-\pi\tau_2/3}$, and $|\eta|^4 = |q|^{4/24}\prod|1-q^n|^4$.

Checks. (i) The square torus: $\tau = i$ gives $\det'\Delta = \eta(i)^4 \approx 0.348$. (ii) Modular invariance: the determinant is a spectral invariant and the spectrum is modular invariant (Example 9.53), so $\tau_2\,|\eta(\tau)|^4$ must be too. Under $\tau \mapsto -1/\tau$, $\tau_2 \mapsto \tau_2/|\tau|^2$; invariance therefore *forces* $|\eta(-1/\tau)|^4 = |\tau|^2\,|\eta(\tau)|^4$, which is the second symmetry law in the working definition. A non-obvious identity of a q-series has been proved by the statement "isometric tori have the same spectrum".

Watch out: Iterated regularisation, and how to trust it

Step (4) regularised the product over rows separately from the products inside each row. Zeta regularisation is not guaranteed to commute with such rearrangements (a double sum regularised "row by row" can differ from the same double sum regularised all at once by a local polynomial term). Here it does not, and the proof is a classical formula. The torus zeta function is $\zeta_\Delta(s) = (4\pi^2)^{-s}E(\tau,s)$ with the **Epstein zeta function** $E(\tau,s) = \sum_{(n,m)\neq 0} \tau_2^s/|m+n\tau|^{2s}$, and **Kronecker's first limit formula** gives, near $s=1$, $E(\tau,s) = \frac{\pi}{s-1} + 2\pi\big(\gamma - \log 2 - \log(\sqrt{\tau_2}\,|\eta(\tau)|^2)\big) + O(s-1)$. The functional equation $\pi^{-s}\Gamma(s)E(\tau,s) = \pi^{s-1}\Gamma(1-s)E(\tau,1-s)$ carries this to $s=0$, where it says $E(\tau,0) = -1$ and $\partial_s E(\tau,0) = -\log\big(4\pi^2\tau_2\,|\eta|^4\big)$. Then

$$\begin{aligned}
-\zeta_\Delta'(0) &\overset{(1)}{=} \log(4\pi^2)\,E(\tau,0) - \partial_s E(\tau,0) \\
&\overset{(2)}{=} -\log(4\pi^2) + \log(4\pi^2) + \log\big(\tau_2\,|\eta|^4\big) = \log\big(\tau_2\,|\eta(\tau)|^4\big),
\end{aligned}$$

(1) Differentiate $(4\pi^2)^{-s}E(\tau,s)$ at $s=0$.
(2) The two values just quoted.

in agreement with (9.89). Note also $\zeta_\Delta(0) = E(\tau,0) = -1 = \chi(T^2)/6 - 1$, exactly as the heat kernel predicted in Example 9.25.

Physical picture: The $-\frac{1}{24}$ is a Casimir energy

Read the matter factor for one scalar field in Hamiltonian language. With $(1,1)=1$,

$$\left(\det{}'\Delta_{\hat{g}_\tau}\right)^{-1/2} = \tau_2^{-1/2}\,|q|^{-1/12}\prod_{n\geq 1}|1-q^n|^{-2} = \tau_2^{-1/2}\,q^{-1/24}\bar{q}^{-1/24}\prod_{n\geq 1}\frac{1}{1-q^n}\,\frac{1}{1-\bar{q}^n}.$$

The torus is a closed string propagating for Euclidean time $2\pi\tau_2$ and then glued back with a twist $2\pi\tau_1$ (Figure 9.5). Each factor $1/(1-q^n)=\sum_k q^{nk}$ is the trace over the Fock space of one oscillator of frequency n moving to the left, and $1/(1-\bar{q}^n)$ likewise for the right movers. The factor $\tau_2^{-1/2}$ is the integral over the centre-of-mass momentum. And $q^{-1/24}$ is $e^{-2\pi\tau_2 E_0}$ with the zero-point energy $E_0=\frac{1}{2}\sum_{n\geq 1} n=\frac{1}{2}\zeta_R(-1)=-\frac{1}{24}$ of the left movers: the same $\zeta_R(-1)$ as in step (4). The oscillator of Example 9.24 kept its $\frac{1}{2}\omega$; the string keeps its $-\frac{1}{24}$ per field.

Step 4: the ghost factor and the torus amplitude

Worked example 9.56: The torus amplitude Z_1

Ghosts. On the flat torus the Weitzenböck formula gives $P_1^\dagger P_1=-2(\nabla^2+K)=2\Delta$ acting on each of the two components V^1,V^2 (Example 9.39, with $K=0$ and no Christoffel symbols in the affine coordinates x,y). So the spectrum of $P_1^\dagger P_1$ is that of 2Δ with every eigenvalue doubled in multiplicity, and its kernel is the two constant vector fields (the CKVs). Therefore $\zeta_{P_1^\dagger P_1}=2\,\zeta_{2\Delta}$ and

$$\sqrt{\det{}'P_1^\dagger P_1}\overset{(1)}{=}\sqrt{\left(\det{}'(2\Delta)\right)^2}=\det{}'(2\Delta)$$
$$\overset{(2)}{=}2^{\zeta_\Delta(0)}\det{}'\Delta\overset{(3)}{=}\tfrac{1}{2}\,\tau_2\,|\eta(\tau)|^4\,.$$

(1) $\det'=\exp(-\zeta'(0))$ and $\zeta_{P_1^\dagger P_1}=2\zeta_{2\Delta}$ square the determinant.
(2) Scaling law of Example 9.26.
(3) $\zeta_\Delta(0)=-1$ and (9.89).

Matter. In the same way $\det'(\Delta/2)=2^{-\zeta_\Delta(0)}\det'\Delta=2\tau_2\,|\eta|^4$ (this is item (i) of Example 9.26 with $\chi=0$), and $(1,1)_{\hat{g}_\tau}=1$.

Assemble. Insert (a)–(d) into (9.81), with $\frac{1}{|\Gamma_\Sigma|}\int_{\mathcal{T}_1}=\int_{\mathcal{F}}$:

$$Z_1\overset{(1)}{=}\int_{\mathcal{F}}\frac{2\,\mathrm{d}\tau_1\mathrm{d}\tau_2}{\tau_2^2}\cdot 1\cdot\left(2\tau_2\,|\eta(\tau)|^4\right)^{-13}\cdot\tfrac{1}{2}\tau_2\,|\eta(\tau)|^4$$
$$\overset{(2)}{=}2^{-13}\int_{\mathcal{F}}\frac{\mathrm{d}\tau_1\mathrm{d}\tau_2}{\tau_2^2}\,\tau_2^{-12}\,|\eta(\tau)|^{-48}\,. \tag{9.90}$$

(1) Measure (a), CKG volume (b), matter factor (c), ghost factor (d).
(2) Powers of 2: $2\cdot 2^{-13}\cdot\frac{1}{2}=2^{-13}$. Powers of τ_2: $\tau_2^{-13}\cdot\tau_2=\tau_2^{-12}$. Powers of $|\eta|$: $-52+4=-48$.

Modular invariance. The measure $\mathrm{d}\tau_1\mathrm{d}\tau_2/\tau_2^2$ is $\mathrm{PSL}(2,\mathbb{Z})$-invariant (Section 9.3.4), and the integrand is $\tau_2^{-12}\,|\eta|^{-48}=\left(\tau_2\,|\eta|^4\right)^{-12}$, a power of the invariant of Example 9.55. So the integrand is a genuine function on $\mathcal{M}_1$, and integrating over $\mathcal{F}$ is legitimate. Notice how the exponent came out: 26 scalars give -13, the ghosts give $+1$, so the net power of the modular invariant $\tau_2\,|\eta|^4$ is $-13+1=-12=-(26-2)/2$. The $26-2=24$ is the number

of transverse directions of the string; the ghosts cancel two of the 26 fields. (Standard references write the same result as $\int_{\mathcal{F}} \frac{\mathrm{d}^2\tau}{\tau_2^2}(\sqrt{\tau_2}\,|\eta|^2)^{-24}$, times the space-time volume V_{26} from the zero mode of Section 9.1.9 and constants depending on α'.)

Watch out: The element -1 and other constants

$\mathrm{SL}(2,\mathbb{Z})$ contains $-\mathbf{1}$, which is the diffeomorphism $(x,y)\mapsto(-x,-y)$. It fixes every point of $\mathcal{T}_1$, and it is not a conformal Killing *flow* (it is not connected to the identity), so it is not divided out by $1/\mathrm{vol}\,\mathrm{CKG}$. Dividing by the whole symmetry group therefore gives an extra factor $\frac{1}{2}$ in front of (9.90). Such overall constants (together with V_{26}, the normalisation of $\mathcal{D}V$ and the N of (9.72)) are fixed in the end by unitarity; they do not affect the τ-dependence.

Step 5: what the answer means: the string spectrum and the tachyon

Expand the integrand in powers of q. Write $\prod_{n\geq1}(1-q^n)^{-24}=\sum_{N\geq0}d_N q^N$. The coefficients are

$$d_0=1,\quad d_1=24,\quad d_2=324,\quad d_3=3200,\quad d_4=25650,\ \ldots \tag{9.91}$$

and d_N counts the ways of distributing N units of excitation among 24 species of oscillator, one of each frequency $n\geq1$.

Worked example 9.57: Counting states: $d_1=24$, $d_2=324$

Expand each factor: $(1-q^n)^{-24}=\left(\sum_k q^{nk}\right)^{24}$. Up to q^2,

$$\begin{aligned}\prod_{n\geq1}(1-q^n)^{-24} &\overset{(1)}{=}(1-q)^{-24}(1-q^2)^{-24}+O(q^3)\\ &\overset{(2)}{=}\left(1+24q+\tbinom{25}{2}q^2\right)\left(1+24q^2\right)+O(q^3)\\ &\overset{(3)}{=}1+24q+(300+24)q^2+O(q^3).\end{aligned}$$

(1) Factors with $n\geq3$ start at q^3.
(2) $(1-q)^{-24}=\sum_k\binom{k+23}{23}q^k$ (choose k quanta among 24 oscillators α^μ_{-1}), so the q^2 coefficient is $\binom{25}{2}=300$; $(1-q^2)^{-24}=1+24q^2+\cdots$.
(3) Multiply out.

In oscillator language: level 1 is $\alpha^\mu_{-1}|0\rangle$ (24 states); level 2 is $\alpha^\mu_{-2}|0\rangle$ (24) or $\alpha^\mu_{-1}\alpha^\nu_{-1}|0\rangle$ (symmetric, 300).

With $|\eta|^{-48}=|q|^{-2}\sum_{N,\tilde N}d_N d_{\tilde N}q^N\bar q^{\tilde N}$ and $q^N\bar q^{\tilde N}=e^{2\pi i\tau_1(N-\tilde N)}e^{-2\pi\tau_2(N+\tilde N)}$, the part of $\mathcal{F}$ with $\tau_2>1$ (where $\mathcal{F}$ is simply the strip $|\tau_1|\leq\frac{1}{2}$) gives

$$\begin{aligned}Z_1\Big|_{\tau_2>1} &\overset{(1)}{=}2^{-13}\int_1^\infty \mathrm{d}\tau_2\,\tau_2^{-14}\sum_{N,\tilde N}d_N d_{\tilde N}\,e^{4\pi\tau_2}e^{-2\pi\tau_2(N+\tilde N)}\int_{-1/2}^{1/2}\mathrm{d}\tau_1\,e^{2\pi i\tau_1(N-\tilde N)}\\ &\overset{(2)}{=}2^{-13}\int_1^\infty \mathrm{d}\tau_2\,\tau_2^{-14}\sum_{N\geq0}d_N^2\,e^{-4\pi\tau_2(N-1)}.\end{aligned}\tag{9.92}$$

(1) Insert the expansion; $|q|^{-2}=e^{4\pi\tau_2}$; $\tau_2^{-2}\tau_2^{-12}=\tau_2^{-14}$.
(2) The τ_1 integral is $\delta_{N\tilde N}$: the integral over the twist projects onto states with equal left and right excitation (**level matching**, $L_0=\bar L_0$).

Now read the terms one by one.

- $N = 0$: $\int^{\infty} \mathrm{d}\tau_2\, \tau_2^{-14} e^{+4\pi\tau_2}$ *diverges*. The integrand grows exponentially in the cusp. This state has "negative mass squared": it is the **tachyon** of the bosonic string.

- $N = 1$: $d_1^2 = 576$ states with no exponential, $\int_1^{\infty} \tau_2^{-14} \mathrm{d}\tau_2 = \frac{1}{13}$, finite. These are the massless states $\alpha_{-1}^{\mu} \tilde{\alpha}_{-1}^{\nu} |0\rangle$: a 24×24 matrix, which splits into symmetric traceless (299, the graviton), antisymmetric (276, the Kalb–Ramond field $B_{\mu\nu}$) and trace (1, the dilaton); $299 + 276 + 1 = 576$.

- $N \geq 2$: exponentially convergent: massive states.

Physical picture: The torus is a sum of particle loops, without the ultraviolet

For one real scalar particle of mass m in $D = 26$ Euclidean dimensions, the one-loop vacuum amplitude is, by the proper-time (Schwinger) representation $\log \lambda = -\int_0^{\infty} \frac{\mathrm{d}t}{t} e^{-t\lambda} + \text{const}$,

$$-\tfrac{1}{2} \log \det\bigl(-\partial^2 + m^2\bigr) \overset{(1)}{=} \tfrac{1}{2} \int_0^{\infty} \frac{\mathrm{d}t}{t}\, \operatorname{Tr} e^{-t(-\partial^2 + m^2)}$$
$$\overset{(2)}{=} \tfrac{1}{2}\, V_{26} \int_0^{\infty} \frac{\mathrm{d}t}{t}\, (4\pi t)^{-13}\, e^{-tm^2}.$$

(1) Apply the proper-time formula to each eigenvalue and sum; the constant is dropped.
(2) $\operatorname{Tr} e^{t\partial^2} = V_{26} \int \frac{\mathrm{d}^{26}k}{(2\pi)^{26}} e^{-tk^2} = V_{26}(4\pi t)^{-13}$.

The integrand is $t^{-14} e^{-m^2 t}$, exactly the shape of (9.92) with $t \propto \tau_2$ and $m^2 \propto 4(N-1)$ (in conventional units, $t = \pi\alpha'\tau_2$ and $m^2 = 4(N-1)/\alpha'$). So in the cusp the string amplitude is a sum of particle vacuum loops, one for each level-matched state. The difference is the region of integration. For a particle, $t \to 0$ (short proper time, high momentum) gives the ultraviolet divergence $\int_0 \mathrm{d}t\, t^{-14}$. For the string, modular invariance restricts τ to $\mathcal{F}$, where $\tau_2 \geq \sqrt{3}/2$: the would-be ultraviolet region is not a separate region at all, it is a copy of the cusp under S. The only divergence left is the infrared one from the tachyon.

Try it: The integrand near $\tau = i$

Using $\eta(i) = \Gamma(\frac{1}{4})/(2\pi^{3/4}) \approx 0.76823$, evaluate the integrand of (9.90) (without the measure) at the square torus. *Answer:* $2^{-13}\,(\eta(i)^4)^{-12} = 2^{-13}\,\eta(i)^{-48} \approx 2^{-13} \times 3.1 \times 10^5 \approx 38$. Compare the cusp, where at $\tau_2 = 3$ the tachyon term alone is $2^{-13}\, 3^{-14} e^{12\pi} \approx 6 \times 10^5$ and growing: the integral is dominated by the cusp.

9.3.14 The sphere: fixing three points

For the sphere, (9.35) has no moduli integral ($\mathcal{M}_0$ is a point) but an infinite factor $1/\mathrm{vol}\,\mathrm{PSL}(2,\mathbb{C})$ (Section 9.3.10). The question is: *how does a physically meaningful sphere amplitude get rid of it?* Answer: scattering amplitudes contain vertex operators $\mathcal{V}_i$ inserted at points z_i of the world sheet, integrated over their positions. The conformal Killing group moves the points, and three points are exactly enough to use it up.

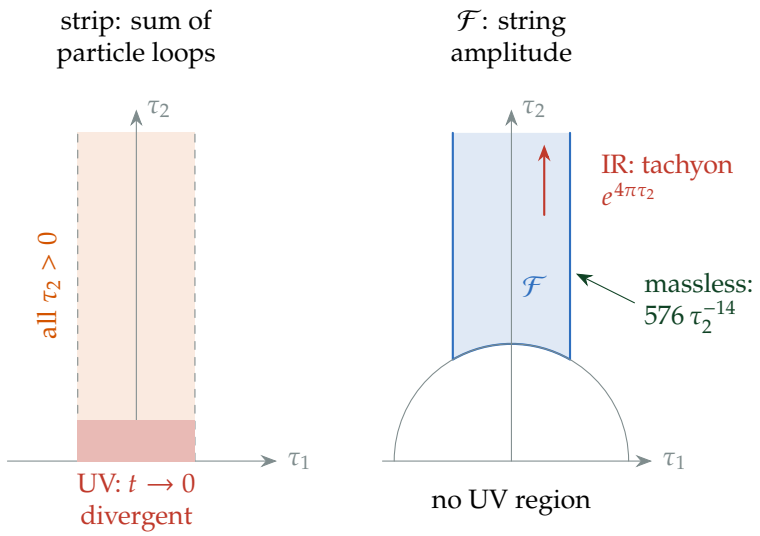


Figure 9.23: Where divergences live. Left: a sum of point-particle loops integrates the proper time $t \propto \tau_2$ over the whole strip, including $t \to 0$ (red), which is the ultraviolet divergence of field theory. Right: the torus amplitude integrates only over the fundamental domain $\mathcal{F}$; the region $\tau_2 \to 0$ has been mapped by $S : \tau \mapsto -1/\tau$ into the cusp and is not counted again. The remaining divergence is in the cusp $\tau_2 \to \infty$ (infrared), where the tachyon term grows like $e^{4\pi\tau_2}$; the massless term decays like τ_2^{-14} and is harmless.

Worked example 9.58: Three points absorb the conformal Killing group

A sphere amplitude with n insertions has the form

$$\mathcal{A}_n = \frac{1}{\operatorname{vol}\mathrm{PSL}(2,\mathbb{C})} \int \prod_{i=1}^{n} \mathrm{d}^2 z_i \; F(z_1, \ldots, z_n),$$

where F (the correlation function of the vertex operators, computed with the determinants of (9.35)) is such that $\prod \mathrm{d}^2 z_i \, F$ is invariant under Möbius maps. We trade the positions z_1, z_2, z_3 for the group. Near the identity, the CKV $V^z = a_0 + a_1 z + a_2 z^2$ of Example 9.34 moves the three points by $\delta z_i = a_0 + a_1 z_i + a_2 z_i^2$, i.e.

$$\begin{pmatrix} \delta z_1 \\ \delta z_2 \\ \delta z_3 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & z_1 & z_1^2 \\ 1 & z_2 & z_2^2 \\ 1 & z_3 & z_3^2 \end{pmatrix}}_{M} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix}.$$

This is a Vandermonde matrix. Its determinant, by row reduction:

$$\begin{aligned} \det M &\overset{(1)}{=} \det \begin{pmatrix} 1 & z_1 & z_1^2 \\ 0 & z_2 - z_1 & z_2^2 - z_1^2 \\ 0 & z_3 - z_1 & z_3^2 - z_1^2 \end{pmatrix} \\ &\overset{(2)}{=} (z_2 - z_1)(z_3 - z_1) \det \begin{pmatrix} 1 & z_2 + z_1 \\ 1 & z_3 + z_1 \end{pmatrix} \\ &\overset{(3)}{=} (z_2 - z_1)(z_3 - z_1)(z_3 - z_2). \end{aligned}$$

(1) Subtract row 1 from rows 2 and 3 (determinant unchanged).
(2) Expand along the first column, and pull the factors $z_2 - z_1$ and $z_3 - z_1$ out of rows 2 and 3, using $z_j^2 - z_1^2 = (z_j - z_1)(z_j + z_1)$.

(3) The 2×2 determinant is $(z_3+z_1)-(z_2+z_1)$.

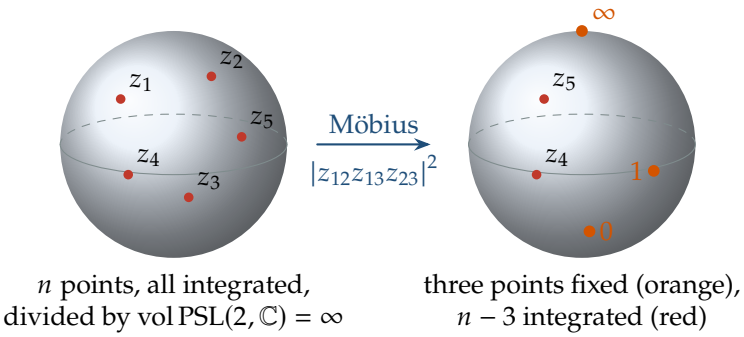
M is a complex-linear map $\mathbb{C}^3\to\mathbb{C}^3$; as a real map $\mathbb{R}^6\to\mathbb{R}^6$ its determinant is $|\det M|^2$ (Example 9.51). Hence, near the identity, $\mathrm{d}^2z_1\,\mathrm{d}^2z_2\,\mathrm{d}^2z_3 = |z_{12}z_{13}z_{23}|^2\,\mathrm{d}(\text{group})$ with $z_{ij}=z_i-z_j$ and $\mathrm{d}(\text{group})$ the (left-invariant) volume element on $\mathrm{PSL}(2,\mathbb{C})$ whose value at the identity is $\mathrm{d}^2a_0\,\mathrm{d}^2a_1\,\mathrm{d}^2a_2$. Since any three distinct points can be moved to any other three by exactly one Möbius map (for example to $0,1,\infty$ by $z\mapsto\frac{(z-z_1)(z_2-z_3)}{(z-z_3)(z_2-z_1)}$), the orbit of a triple is a copy of the group, and

$$\mathcal{A}_n=\int\prod_{i=4}^{n}\mathrm{d}^2z_i\;|z_{12}z_{13}z_{23}|^2\;F(z_1,\dots,z_n)\Big|_{z_1,z_2,z_3\text{ fixed}}. \tag{9.93}$$

The infinite volume has cancelled against the three integrals. The factor $|z_{12}z_{13}z_{23}|^2$ is the Faddeev–Popov determinant for this gauge fixing; in the ghost language of Section 9.3.11 it is the correlator $|\langle c(z_1)c(z_2)c(z_3)\rangle|^2$, the three c zero modes soaked up by the three fixed points.

Try it: The measure is Möbius invariant

Show that $\dfrac{\mathrm{d}z_1\,\mathrm{d}z_2\,\mathrm{d}z_3}{z_{12}\,z_{13}\,z_{23}}$ is invariant under $z\mapsto z'=\frac{az+b}{cz+d}$, $ad-bc=1$. *Answer:* $\mathrm{d}z_i'=\mathrm{d}z_i/(cz_i+d)^2$ and $z_i'-z_j'=(z_i-z_j)/\big((cz_i+d)(cz_j+d)\big)$. Each index i appears in two of the three differences, so the denominator acquires $\prod_i(cz_i+d)^{-2}$, which cancels the numerator's. Hence the invariant volume element on the group, written in terms of three points, is $|z_{12}z_{13}z_{23}|^{-2}\prod_{i=1}^3\mathrm{d}^2z_i$, consistent with (9.93).



n points, all integrated, divided by $\mathrm{vol}\,\mathrm{PSL}(2,\mathbb{C})=\infty$

three points fixed (orange), $n-3$ integrated (red)

Figure 9.24: Gauge fixing the conformal Killing group of the sphere. A Möbius map is fixed uniquely by where it sends three points, so the six real directions of $\mathrm{PSL}(2,\mathbb{C})$ can be traded for the positions of z_1,z_2,z_3, placed at $0,1,\infty$. The Jacobian of the trade is $|z_{12}z_{13}z_{23}|^2$; the remaining $n-3$ positions (and nothing else) are the moduli of the punctured sphere.

Punctures count as moduli. The same count works on every surface. Adding a puncture adds two real parameters (its position) and removes nothing, unless there are CKVs to absorb it. So the real dimension of the moduli space $\mathcal{M}_{p,n}$ of genus-p surfaces with n marked points is

$$\dim_{\mathbb{R}}\mathcal{M}_{p,n}=\underbrace{6p-6}_{-3\chi=-\,\mathrm{index}\,P_1}+2n, \tag{9.94}$$

valid whenever it is ≥ 0 and the CKVs are all used up, i.e. $2p-2+n>0$.

	$n=0$	$n=1$	$n=3$	$n=4$
$p=0$	(CKG dim 6 left)	(CKG dim 4 left)	0 (a point)	2 (cross-ratio)
$p=1$	2 (CKG dim 2 left)	2	6	8
$p=2$	6	8	12	14

For $p=0, n=4$ the single complex modulus is the cross-ratio $x=\frac{z_{12}z_{34}}{z_{13}z_{24}}$, the only Möbius invariant of four points; the four-tachyon amplitude becomes an integral $\int \mathrm{d}^2x\ |x|^A\,|1-x|^B$ over the sphere, the Virasoro–Shapiro amplitude. For $p=1$, $n=0$ the torus keeps its translations, as we saw in Example 9.49; with one puncture, the translations are used to put the puncture at $z=0$.

9.3.15 Higher genus: what the formula becomes

For $p\geq 2$ there are no CKVs (Example 9.34), and (9.35) holds as written: an integral over the $(6p-6)$-dimensional $\mathcal{M}_p$. Two developments, both quoted here without derivation, show how far the elliptic-operator form of the answer can be pushed. They are the "further reading" behind Nash's reference to D'Hoker and Phong.

Determinants as Selberg zeta values. A constant-curvature metric of genus $p\geq 2$ is hyperbolic: $\Sigma=\mathbb{H}/\Gamma$ with $\Gamma\subset \mathrm{PSL}(2,\mathbb{R})$ (Section 9.2.3). Its closed geodesics γ have lengths ℓ_γ, and the **Selberg zeta function** is $Z(s)=\prod_{\gamma\ \text{primitive}}\prod_{k\geq 0}\left(1-e^{-(s+k)\ell_\gamma}\right)$. The Selberg trace formula relates the spectrum of Δ to the lengths ℓ_γ, just as Poisson summation related the rows of Figure 9.22 to the lattice. D'Hoker and Phong showed that, with constants c_p, c_p' depending only on the genus,

$$\det{}'\Delta_g = c_p\, Z'(1), \qquad \det P_1^\dagger P_1 = c_p'\, Z(2). \tag{9.95}$$

So the integrand of (9.35) is a function of the length spectrum of the surface, $Z(2)^{1/2}\,Z'(1)^{-13}$ up to constants. The torus analogue is (9.89): there the "geodesics" are the lattice vectors, and the role of Z is played by $\prod(1-q^n)$.

Holomorphic factorisation. In the torus answer the integrand was $\tau_2^{-14}\left|\eta(\tau)^{24}\right|^{-2}$: a power of τ_2 times the squared modulus of a *holomorphic* function of τ (the cusp form η^{24}). Belavin and Knizhnik showed that the same happens in every genus, and that it happens *only* when $d=26$. Mathematically it is the Mumford isomorphism $\lambda_2\cong\lambda_1^{13}$ between two holomorphic line bundles over $\mathcal{M}_p$: λ_1 is the determinant line of the holomorphic 1-forms (from the matter determinant) and λ_2 that of the quadratic differentials (from the ghost determinant). The exponent 13 is the same $13=26/2$ as in (9.35). For $p=2$, in terms of the period matrix Ω (a 2×2 symmetric matrix with positive imaginary part, three complex moduli), the result is $\mathrm{d}\mu\propto(\det \mathrm{Im}\,\Omega)^{-13}\,|\chi_{10}(\Omega)|^{-2}\prod_{i\leq j}\mathrm{d}^2\Omega_{ij}$, with χ_{10} a Siegel cusp form of weight 10. The anomaly that decides $d=26$ (Section 9.2.11, Chapter 10) is the same obstruction that spoils holomorphic factorisation when $d\neq 26$.

Try it: Weight counting at genus 1 and 2

Under $\tau\mapsto\frac{a\tau+b}{c\tau+d}$: $\tau_2\mapsto\tau_2\,|c\tau+d|^{-2}$, $\mathrm{d}^2\tau\mapsto\mathrm{d}^2\tau\,|c\tau+d|^{-4}$ and $\eta^{24}\mapsto(c\tau+d)^{12}\eta^{24}$. Check that $\tau_2^{-14}\left|\eta^{24}\right|^{-2}\mathrm{d}^2\tau$ is invariant. At genus 2, with $\det\mathrm{Im}\,\Omega\mapsto|\det(C\Omega+D)|^{-2}\det\mathrm{Im}\,\Omega$, $\prod\mathrm{d}^2\Omega_{ij}\mapsto|\det(C\Omega+D)|^{-6}\prod\mathrm{d}^2\Omega_{ij}$ and $\chi_{10}\mapsto\det(C\Omega+D)^{10}\chi_{10}$, check the genus-2 measure. *Answer:* genus 1: $|c\tau+d|^{28-4-24}=1$. Genus 2: $|\det(C\Omega+D)|^{26-6-20}=1$. The extra power of τ_2 at genus 1 is the conformal Killing group of the torus.

Reading Nash (pp. 265–268, (9.23)–(9.36))

p. 265. $\mathrm{Met}_{\mathrm{const}}(\Sigma)$ is a Riemannian submanifold of $\mathrm{Met}(\Sigma)$, and Diff_0 acts on it by isometries (footnote 3: isometries of $\mathrm{Met}_{\mathrm{const}}$, not of Σ). This is Section 9.3.2. **pp. 265–267, (9.23)–(9.29).** The projection π, the tangent space at a class $[g_{\mathrm{const}}]$, the definition of $\langle\,,\,\rangle_{\mathrm{wp}}$ by lifting tangent vectors, and the check that the answer does not depend on the lift or on the representative: Section 9.3.3, where the lift is taken horizontal ($\operatorname{Ker} P_1^\dagger$). Nash does not compute an example; ours is the torus, $\mathrm{d}s^2_{\mathrm{wp}} = 2\,|\mathrm{d}\tau|^2/\tau_2^2$ (Example 9.42), with volume $2\pi/3$ for $\mathcal{M}_1$ (Example 9.43). **p. 267, (9.30)–(9.34).** The decomposition $T_g\mathrm{Met} = C_g \oplus \operatorname{Range} A_g \oplus \operatorname{Ker} A_g^*$, the coordinates $\{\sigma\}, \{A_g V\}, \{\mathcal{T}_p\}$, the factor N, the Jacobian $\sqrt{\det A_g^* A_g}$ and its failure when $\operatorname{Ker} A_g \neq 0$ (CKVs, $p = 0, 1$), and the split vect $= \operatorname{Ker} A_g \oplus \operatorname{Ker}^\perp A_g$. In our notation $A_g = P_1$: Sections 9.3.6 to 9.3.10. **pp. 267–268, (9.35)–(9.36).** The final formulae. Nash's det is our det′ (zero modes removed); $\tilde{A}_g$ is P_1 restricted to $\operatorname{Ker}^\perp P_1$; the $1/|\Gamma_\Sigma|$ converts $\int_{\mathcal{T}_p}$ into $\int_{\mathcal{M}_p}$; the real-versus-complex identity $\det \tilde{A}_g^* \tilde{A}_g = (\det \bar{\partial}_{K^*}^* \bar{\partial}_{K^*})^2$ is "easy to verify" in Nash, and verified in Section 9.3.12. Nash's chapter ends here; the torus computation Example 9.56 and the sphere gauge fixing Example 9.58 show the formula producing numbers.

Recap: Section

- Teichmüller space inherits a metric from $\mathrm{Met}_{\mathrm{const}}$ because Diff_0 acts by isometries: lift a tangent vector horizontally (to a holomorphic quadratic differential) and measure it with $\langle\,,\,\rangle_{\mathrm{met}}$. This is the Weil–Petersson metric.
- For the torus, $\mathcal{T}_1$ is the upper half-plane, WP is twice the hyperbolic metric, $\Gamma_{T^2} = \mathrm{PSL}(2,\mathbb{Z})$ acts by Möbius maps, and $\mathcal{M}_1 = \mathcal{F}$ has WP volume $2\pi/3$.
- In conformal-gauge coordinates $g = e^{2\sigma}\alpha^*\hat{g}_t$, the measure on Met factorises into Weyl, diffeomorphism and moduli pieces; the Jacobian is triangular, and its diagonal blocks are the Weyl factor, $\sqrt{\det' P_1^\dagger P_1}$ and the WP volume element.
- Dividing by the gauge volumes leaves Z_p as an integral over $\mathcal{M}_p$ of $(\det'(\Delta/2)/(1,1))^{-13}\sqrt{\det' P_1^\dagger P_1}$; conformal Killing vectors contribute $1/\mathrm{vol}\,\mathrm{CKG}$; ghosts rewrite the determinant as a Grassmann integral; as a complex operator the ghost determinant is $\det' \bar{\partial}_{K^*}^* \bar{\partial}_{K^*}$.
- For the torus everything is explicit: $\det'\Delta = \tau_2\,|\eta|^4$ and $Z_1 = 2^{-13}\int_{\mathcal{F}} \mathrm{d}^2\tau\, \tau_2^{-14}\,|\eta|^{-48}$, modular invariant, UV finite, divergent only through the tachyon in the cusp.
- For the sphere, three vertex-operator positions absorb $\mathrm{PSL}(2,\mathbb{C})$ with Jacobian $|z_{12}z_{13}z_{23}|^2$; each extra puncture adds one complex modulus.

Chapter summary

This chapter asked one question: *what does it mean to integrate over all surfaces?* The Polyakov path integral sums e^{-S} over pairs (embedding ϕ, world-sheet metric g). Both spaces are infinite-dimensional, and the symmetry group is infinite-dimensional too. The answer was that almost everything is gauge. Once the embeddings have been integrated out (a Gaussian) and the gauge group has been divided out (Faddeev–Popov), what is left is an *ordinary integral*

over a finite-dimensional space, the moduli space $\mathcal{M}_p$ of Riemann surfaces of genus p. The price of the reduction is a pair of determinants of elliptic operators, and the measure on $\mathcal{M}_p$ is the volume form of the Weil–Petersson metric.

The story, in ten sentences:

1. A string sweeps out a surface. Its natural action is tension times area (Nambu–Goto), which has a square root. Adding an independent world-sheet metric g removes the square root (Polyakov). The action becomes d free massless scalars on (Σ, g), and eliminating g by its equation of motion $T_{ab} = 0$ gives back the area.
2. The Polyakov action is invariant under $\mathrm{Diff}^+(\Sigma)$ (relabelling points) and, in two dimensions only, under the Weyl group $C^\infty_+(\Sigma)$ (rescaling lengths pointwise). Together they form the semidirect product $\mathrm{Diff}^+ \ltimes C^\infty_+$.
3. A metric on an oriented surface is the same as a complex structure plus a scale. So summing over metrics up to Weyl rescaling is summing over Riemann surfaces.
4. The ϕ integral is Gaussian. It gives $\bigl(\det{}'(\Delta_g/2)/(1,1)_g\bigr)^{-d/2}$, where the zero mode (the centre of mass) contributes the area $(1,1)_g$ and the volume of space-time.
5. Infinite products of eigenvalues are defined by zeta functions: $\det{}'P = e^{-\zeta'_P(0)}$. On a circle of length L this gives $\det{}'\Delta = L^2$, and the heat kernel gives $\zeta_\Delta(0) = \chi/6 - 1$ on a closed surface.
6. The space $\mathrm{Met}(\Sigma)$ carries a natural diffeomorphism-invariant metric, $\int \sqrt{g}\ \mathrm{tr}(g^{-1}Xg^{-1}Y)$. With it, each tangent space splits orthogonally into Weyl directions C_g, diffeomorphism directions $\mathrm{Range}\, P_1$ and moduli $\mathrm{Ker}\, P_1^\dagger$.
7. The operator $P_1 V =$ (traceless part of $\mathcal{L}_V g$) is a Cauchy–Riemann operator, $\partial_{\bar z}$ acting on V^z. So it is elliptic. Its kernel consists of the conformal Killing vectors, and its cokernel of the holomorphic quadratic differentials. Riemann–Roch gives $\mathrm{index}\, P_1 = 3\chi$.
8. Teichmüller space $\mathcal{T}_p = \mathrm{Met}_{\mathrm{const}}/\mathrm{Diff}_0$ has tangent space $\mathrm{Ker}\, P_1^\dagger$, and so real dimension $6p - 6$ for $p \geq 2$ (2 for the torus, 0 for the sphere). It inherits the Weil–Petersson metric by horizontal lifting; for the torus this is twice the Poincaré metric.
9. In the conformal-gauge coordinates $g = \alpha^*(e^\sigma \hat g_t)$ the Jacobian is triangular. Its diagonal blocks are a Weyl factor, the Faddeev–Popov determinant $\sqrt{\det{}' P_1^\dagger P_1}$ and the Weil–Petersson volume element. Dividing by the gauge volumes leaves an integral over $\mathcal{M}_p = \mathcal{T}_p/\Gamma_\Sigma$.
10. The division by the Weyl volume is legitimate only when the integrand is Weyl invariant. The matter and ghost anomalies are proportional to d and -26, so this happens exactly in $d = 26$. The result is Nash's (9.36), $Z_p = \int_{\mathcal{M}_p} \mathrm{d}\mu_{\mathrm{wp}}\, (\det{}'(\Delta/2)/(1,1))^{-13} \det{}' \bar\partial^*_{K^*} \bar\partial_{K^*}$.

One surface through every lens: the flat torus

Every construction of the chapter can be carried out on the torus $\mathbb{R}^2/\mathbb{Z}^2$ with the flat metric $\hat g_\tau = \frac{1}{\tau_2}\,|\mathrm{d}x + \tau\,\mathrm{d}y|^2$, and every number is explicit. Table 9.1 collects the facts. Four of the entries are worth deriving once more, because together they show the whole mechanism at work: a determinant, a count, a symmetry and an anomaly.

The determinant: the circle again. The torus determinant is built out of circle determinants (Example 9.54), and the circle is where the zeta recipe is easiest to see. For $\Delta = -\mathrm{d}^2/\mathrm{d}\theta^2$ on a

Concept	On the flat torus $(\mathbb{R}^2/\mathbb{Z}^2, \hat{g}_\tau)$
complex structure	$z = x + \tau y$; Beltrami coefficient 0 in this coordinate; $\hat{g}_\tau = \tau_2^{-1} \lvert \mathrm{d}z \rvert^2$
area $(1,1)_g$	$\sqrt{\det \hat{g}_\tau} = 1$, so the area is 1 for every τ
Laplacian spectrum	$\lambda_{n,m} = \frac{4\pi^2}{\tau_2} \lvert m + n\tau \rvert^2$, $(n,m) \neq (0,0)$; one zero mode
$\zeta_\Delta(0)$	$\chi/6 - 1 = -1$
$\det' \Delta$	$\tau_2 \lvert \eta(\tau) \rvert^4$ (Example 9.55); $\det'(\Delta/2) = 2\tau_2 \lvert \eta \rvert^4$
CKVs, $\operatorname{Ker} P_1$	the constant vector fields $v^1 \partial_x + v^2 \partial_y$: real dimension 2
conformal Killing group	translations $\mathbb{R}^2/\mathbb{Z}^2$, the torus itself; volume $(1,1)_g = 1$
quadratic differentials	$\operatorname{Ker} P_1^\dagger = \{c\, \mathrm{d}z^2\}$, $c \in \mathbb{C}$: real dimension 2
index of P_1	$2 - 2 = 0 = 3\chi(T^2)$
$P_1^\dagger P_1$	2Δ on each of the two components of V; $\sqrt{\det' P_1^\dagger P_1} = \frac{1}{2}\tau_2 \lvert \eta \rvert^4$
Teichmüller space	$\mathcal{T}_1 = \{\tau : \tau_2 > 0\}$, the upper half-plane
mapping class group	$\Gamma_{T^2} = \mathrm{SL}(2,\mathbb{Z})$ acting by $\tau \mapsto \frac{a\tau+b}{c\tau+d}$ (effectively $\mathrm{PSL}(2,\mathbb{Z})$)
moduli space	$\mathcal{M}_1 = \mathcal{F}$: $\lvert \tau_1 \rvert \leq \frac{1}{2}$, $\lvert \tau \rvert \geq 1$
Weil–Petersson metric	$\mathrm{d}s^2_{\mathrm{wp}} = 2 \lvert \mathrm{d}\tau \rvert^2 / \tau_2^2$, twice Poincaré; $\mathrm{vol}_{\mathrm{wp}} \mathcal{M}_1 = 2\pi/3$
amplitude	$Z_1 = 2^{-13} \int_{\mathcal{F}} \mathrm{d}^2\tau\, \tau_2^{-14} \lvert \eta(\tau) \rvert^{-48}$
spectrum of states	$\lvert \eta \rvert^{-48} = \lvert q \rvert^{-2} \sum_{N,\tilde{N}} d_N d_{\tilde{N}} q^N \bar{q}^{\tilde{N}}$, $d_1 = 24$: tachyon at $N = 0$, 576 massless states at $N = 1$

Table 9.1: The flat torus as a one-page model of Chapter 9. Every row is a concept of the chapter, evaluated explicitly. The only genuinely hard entry is $\det' \Delta$, which needs the product formula for sin and the value $\zeta_R(-1) = -\frac{1}{12}$.

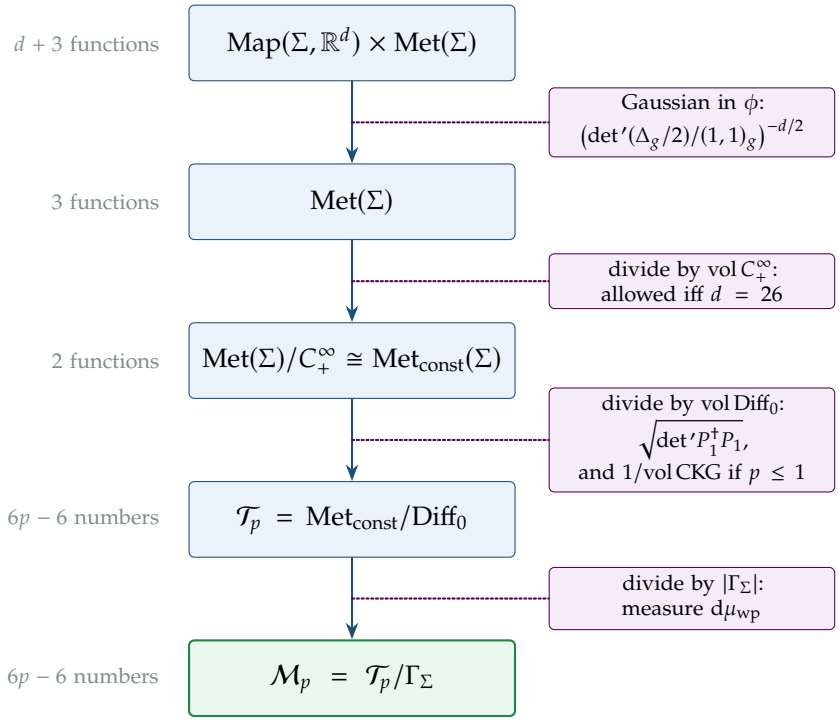


Figure 9.25: The chapter as a funnel. Each arrow integrates out or divides out one piece of the configuration space. The purple boxes show what the step leaves behind in the integrand. The grey counts on the left give the local and global size of each space: the three components of a two-dimensional metric are used up by one Weyl function and two diffeomorphism functions, and only the $6p-6$ moduli survive (2 for the torus, none for the sphere).

circle of length L, the eigenfunctions $e^{2\pi i n\theta/L}$ have eigenvalues $(2\pi n/L)^2$, and each non-zero value occurs twice ($\pm n$). So

$$\zeta_\Delta(s)\overset{(1)}{=}2\sum_{n\ge1}\Big(\frac{2\pi n}{L}\Big)^{-2s}\overset{(2)}{=}2\Big(\frac{L}{2\pi}\Big)^{2s}\zeta_R(2s),$$

$$\zeta'_\Delta(0)\overset{(3)}{=}4\log\Big(\frac{L}{2\pi}\Big)\zeta_R(0)+4\zeta'_R(0)\overset{(4)}{=}-2\log\Big(\frac{L}{2\pi}\Big)-2\log(2\pi)=-2\log L,$$

$$\det'\Delta\overset{(5)}{=}e^{-\zeta'_\Delta(0)}=L^2.$$

(1) Sum over the non-zero eigenvalues, with multiplicity 2.
(2) Pull out $(L/2\pi)^{2s}$ and recognise $\sum n^{-2s}=\zeta_R(2s)$.
(3) Product rule at $s=0$: $\frac{\mathrm{d}}{\mathrm{d}s}(L/2\pi)^{2s}=2\log(L/2\pi)(L/2\pi)^{2s}$ and $\frac{\mathrm{d}}{\mathrm{d}s}\zeta_R(2s)=2\zeta'_R(2s)$.
(4) The Riemann values $\zeta_R(0)=-\frac12$, $\zeta'_R(0)=-\frac12\log 2\pi$ (9.31).
(5) Definition (9.30) of the zeta-regularised determinant.

The same value $\zeta_\Delta(0)=2\zeta_R(0)=-1$ tells us how the determinant scales: replacing L by cL multiplies every eigenvalue by c^{-2} and the determinant by $(c^{-2})^{\zeta(0)}=c^2$, consistent with L^2. In the naive product, $\prod(\text{const})$ would be meaningless. The zeta function turns "infinitely many factors of c^{-2}" into the finite power $c^{-2\zeta(0)}$.

The count: moduli minus Killing vectors. Riemann–Roch (9.61) predicts $\dim\ker P_1-\dim\ker P_1^\dagger=3\chi$. On the three basic surfaces,

$$S^2:\quad \dim\ker P_1-\dim\ker P_1^\dagger\overset{(1)}{=}6-0=6=3\cdot 2,$$

$$T^2: \quad \dim\operatorname{Ker} P_1 - \dim\operatorname{Ker} P_1^\dagger \overset{(2)}{=} 2 - 2 = 0 = 3\cdot 0,$$
$$\Sigma_p,\ p\geq 2: \quad \dim\operatorname{Ker} P_1 - \dim\operatorname{Ker} P_1^\dagger \overset{(3)}{=} 0 - (6p-6) = 3(2-2p).$$

(1) On the sphere, $P_1 V = 0$ means V^z is holomorphic on $\mathbb{CP}^1$ including ∞, so $V^z = a_0 + a_1 z + a_2 z^2$ with $a_i \in \mathbb{C}$: the Lie algebra of $\mathrm{PSL}(2,\mathbb{C})$, six real dimensions. There are no holomorphic quadratic differentials on the sphere ($\mathrm{d}z^2$ has a pole of order 4 at ∞).
(2) On the torus, holomorphic V^z and holomorphic $\varphi\,\mathrm{d}z^2$ are both constant (a holomorphic function on a compact surface is constant), one complex dimension each.
(3) For $p \geq 2$ the curvature is negative, so the Weitzenböck formula $P_1^\dagger P_1 = -2(\nabla^2 + K)$ is a positive operator with no kernel (Example 9.39); the cokernel then has dimension $-3\chi = 6p - 6$ by Riemann–Roch.

So the index is the same topological number whichever way the individual dimensions happen to fall. Moduli and conformal Killing vectors are the two ways a surface can fail to be "pure gauge", and only their difference is topological.

The symmetry: modular invariance of Z_1. The torus amplitude must not depend on which point of $\mathcal{T}_1$ we choose to represent a conformal class, so its integrand must be invariant under $\mathrm{SL}(2,\mathbb{Z})$. The group is generated by $T:\tau\mapsto\tau+1$ and $S:\tau\mapsto -1/\tau$, and the Dedekind function obeys $\eta(\tau+1) = e^{i\pi/12}\eta(\tau)$ and $\eta(-1/\tau) = \sqrt{-i\tau}\,\eta(\tau)$. Write the integrand as $\frac{\mathrm{d}^2\tau}{\tau_2^2}\big(\tau_2\,|\eta|^4\big)^{-12}$. Under S, with $\tau' = -1/\tau$,

$$\tau_2' \overset{(1)}{=} \operatorname{Im}\!\left(-\frac{\bar\tau}{|\tau|^2}\right) = \frac{\tau_2}{|\tau|^2}, \qquad |\eta(\tau')|^4 \overset{(2)}{=} |-i\tau|^2\,|\eta(\tau)|^4 = |\tau|^2\,|\eta(\tau)|^4,$$
$$\tau_2'\,|\eta(\tau')|^4 \overset{(3)}{=} \tau_2\,|\eta(\tau)|^4,$$
$$\frac{\mathrm{d}^2\tau'}{\tau_2'^2} \overset{(4)}{=} \frac{|\tau|^{-4}\,\mathrm{d}^2\tau}{\tau_2^2\,|\tau|^{-4}} = \frac{\mathrm{d}^2\tau}{\tau_2^2}.$$

(1) Multiply numerator and denominator of $-1/\tau$ by $\bar\tau$; $\operatorname{Im}(-\bar\tau) = \tau_2$.
(2) The modular transformation of η, raised to the fourth power in absolute value: $\left|\sqrt{-i\tau}\right|^4 = |\tau|^2$.
(3) Multiply the two results of the first line: the factors $|\tau|^{\pm 2}$ cancel.
(4) $\mathrm{d}\tau'/\mathrm{d}\tau = \tau^{-2}$, so areas scale by $|\tau|^{-4}$; and $\tau_2'^2 = \tau_2^2\,|\tau|^{-4}$ by line 1.

Under T, τ_2 is unchanged and $|\eta|$ only picks up the phase $\left|e^{i\pi/12}\right| = 1$. So both factors are invariant separately: $\tau_2\,|\eta|^4$ is a modular invariant (it is $\det'\Delta$, which depends only on the metric, not on the choice of τ representing it), and $\mathrm{d}^2\tau/\tau_2^2$ is the invariant hyperbolic area. The integral over the fundamental domain $\mathcal{F}$ is therefore well defined: this is the $1/|\Gamma_\Sigma|$ of (9.78) at work.

The anomaly: a constant rescaling checks the coefficient. The Weyl variation formula (9.63) says $\delta\log(\det'\Delta/A) = -\frac{1}{12\pi}\int\sqrt{g}\,R\,\delta\omega$. We can check its coefficient with nothing more than $\zeta_\Delta(0)$, by taking ω constant. Then $g\mapsto e^{2\omega}g$ sends $\Delta\mapsto e^{-2\omega}\Delta$ and $A\mapsto e^{2\omega}A$, so

$$\log\frac{\det'(e^{-2\omega}\Delta)}{e^{2\omega}A} \overset{(1)}{=} -2\omega\,\zeta_\Delta(0) + \log\det'\Delta - 2\omega - \log A$$
$$\overset{(2)}{=} \log\frac{\det'\Delta}{A} - 2\omega\left(\frac{\chi}{6} - 1\right) - 2\omega = \log\frac{\det'\Delta}{A} - \frac{\chi}{3}\,\omega,$$

$$-\frac{1}{12\pi}\int_\Sigma \sqrt{g}\,R\,\omega \overset{(3)}{=} -\frac{\omega}{12\pi}\cdot 4\pi\chi = -\frac{\chi}{3}\,\omega.$$

(1) The scaling law $\det'(cP) = c^{\zeta_P(0)}\det' P$ of Example 9.26, with $c = e^{-2\omega}$.
(2) The heat-kernel value $\zeta_\Delta(0) = \chi/6 - 1$ (Example 9.25).
(3) Gauss–Bonnet $\int\sqrt{g}\,R = 4\pi\chi$ (with $R = 2K$).

The two sides agree. So the conformal anomaly of one scalar field is, at its crudest, the statement that $\zeta_\Delta(0)$ is not zero. For d scalars the matter factor $(\det'\Delta/A)^{-d/2}$ varies by $+\frac{d}{24\pi}\int\sqrt{g}\,R\,\delta\omega$. The ghost factor varies with $-\frac{26}{24\pi}$, and the total (9.64) vanishes only for $d = 26$. (On the torus $\chi = 0$, so constant rescalings are harmless in any dimension. And since $\tau_2\,|\eta|^4$ is modular invariant by itself, the torus integrand is modular invariant for every d. So on the torus the number 26 does not show up as a consistency condition. It shows up in the spectrum instead: the exponent $-48 = -2(d-2)$ of $|\eta|$ in Z_1 counts $d-2 = 24$ oscillator species, which is why $d_1 = 24$, and the leading $|q|^{-2} = \left|q^{-(d-2)/24}\right|^2$ is what puts the graviton exactly at the massless level $N = 1$.)

Every concept of the chapter

The tables below list every concept introduced in Chapter 9, one table per section of Nash, plus one for the explicit amplitudes we computed beyond Nash. The middle column is the *operational* meaning: how to compute it or test for it. The last column says where the concept is used later in the book (or where it came from, when it was imported).

Nash §1, The bosonic string (Nash pp. 259–262): actions and symmetries

Concept	Operational meaning	Used later
world line, einbein	$S_e = \frac{1}{2}\int(\dot{x}^2/e + em^2)\,\mathrm{d}\tau$; eliminating e gives $m\times$length	model for the Polyakov trick
world sheet	the surface $\phi:\Sigma\to\mathbb{R}^d$ swept by a string; closed and oriented here	Chapter 11
induced metric	$h_{ab} = \partial_a\phi\cdot\partial_b\phi$, lengths measured with space-time rulers	Nambu–Goto area
Nambu–Goto action	$T\int\sqrt{\det h}$: tension times area; Gram determinant	classical string, minimal surfaces
Polyakov action	$\frac{T}{2}\int\sqrt{g}\,g^{ab}\partial_a\phi\cdot\partial_b\phi$: d free scalars on (Σ, g)	all of Chapter 9, Chapter 10, Chapter 11
energy–momentum tensor	$T_{ab} = \partial_a\phi\cdot\partial_b\phi - \frac{1}{2}g_{ab}(\partial\phi)^2$; traceless; $T_{ab} = 0$ is the g equation	Virasoro algebra, c (Chapter 11)
classical equivalence	at $T_{ab} = 0$, $g \propto h$ and $S_{\mathrm{P}} = S_{\mathrm{NG}}$	
Weyl transformation	$g\mapsto e^{2\omega}g$; $\sqrt{g}\,g^{ab}$ unchanged in 2D only	conformal anomaly (Chapter 10)
diffeomorphism action	pull back: $\phi\mapsto\alpha^*\phi = \phi\circ\alpha$, $g\mapsto\alpha^*g$	gauge group of gravity
target translations	$\phi^\mu\mapsto\phi^\mu + a^\mu$; source of the zero mode	momentum conservation
configuration space B	$\mathrm{Map}(\Sigma,\mathbb{R}^d)\times\mathrm{Met}(\Sigma)$, Nash's domain of S	
gauge group	$\mathrm{Diff}^+\ltimes C^\infty_+$, law $(\alpha_1, f_1)(\alpha_2, f_2) = (\alpha_2\circ\alpha_1, f_1\,(f_2\circ\alpha_1))$	Faddeev–Popov (Section 9.2)
normal subgroup	C^∞_+ is normal: α conjugates a Weyl factor to $f\circ\alpha^{-1}$	quotient in two steps

Nash §1, The bosonic string (Nash pp. 259–262): path integral and determinants

Concept	Operational meaning	Used later
Z_p, vacuum amplitude	$\int \mathcal{D}g\,\mathcal{D}\phi\, e^{-S}$ over genus-p surfaces; closed surfaces compute a trace	Chapter 10, Chapter 12
genus expansion	$Z = \sum_p Z_p$; the Euler term weights each handle by g_s^2	string perturbation theory
Riemann surface	a metric up to scale = a complex structure; Beltrami μ from $\mathrm{d}s^2 = \lambda\,\lvert\mathrm{d}z + \mu\mathrm{d}\bar{z}\rvert^2$	Chapter 5, Chapter 11
isothermal coordinates	solve $\partial_{\bar{z}} w = \mu\,\partial_z w$; then $g = \rho\,\lvert\mathrm{d}w\rvert^2$	conformal gauge
action as $\langle\phi, \Delta\phi\rangle$	$S = \frac{1}{2}(\phi, \Delta_g\phi)$, $\Delta_g = -\frac{1}{\sqrt{g}}\partial_a\sqrt{g}g^{ab}\partial_b$; $= 2\int \partial\phi\cdot\bar{\partial}\phi$	Gaussian integral
Gaussian integral	$\int \mathrm{d}^n x\, e^{-x^{\mathsf{T}}Mx} = \pi^{n/2}(\det M)^{-1/2}$; mode by mode for Δ_g	every one-loop determinant
Weyl's law	$N(\lambda) \approx A\lambda/4\pi$; so $\sum \lambda_n^{-s}$ converges for $\mathrm{Re}\, s > 1$	heat kernel
zero mode, $(1,1)_g$	constants in $\mathrm{Ker}\,\Delta_g$; the centre-of-mass integral gives $\mathrm{vol}(\mathbb{R}^d)\cdot A_g^{d/2}$	momentum δ-function
det′	product of non-zero eigenvalues; lattice circle: $\det' M = n^2$	(9.7), (9.35)
zeta function	$\zeta_P(s) = \sum \lambda_n^{-s}$, continued to $s = 0$	Chapter 10, Chapter 12
zeta determinant	$\det' P = e^{-\zeta_P'(0)}$; circle: L^2; oscillator: $4\sinh^2(\beta\omega/2)$	Ray–Singer torsion (Chapter 12)
heat trace	$\mathrm{Tr}\, e^{-t\Delta} = \frac{A}{4\pi t} + \frac{\chi}{6} + O(t)$; Mellin transform gives ζ	index theorem (Chapter 4)
$\zeta_\Delta(0)$	$\chi/6 - 1$: the constant term of the heat trace minus the zero mode	scaling law, anomaly
scaling law	$\det'(cP) = c^{\zeta_P(0)}\det' P$	constant Weyl anomaly

Nash §2, The space of metrics (Nash pp. 262–265): quotients and metrics

Concept	Operational meaning	Used later
$\mathcal{D}g$, Nash's (9.8)	$\mathcal{D}\mathrm{Met}/(\mathrm{vol}\, C^\infty_+ \,\mathrm{vol}\,\mathrm{Diff}^+)$: one representative per orbit	(9.35)
Faddeev–Popov (finite)	polar coordinates: $\mathrm{d}^2x = r\,\mathrm{d}r\,\mathrm{d}\theta$; the Jacobian is the orbit's size	Yang–Mills gauge fixing (Chapter 8)
Diff_0, Γ_Σ	diffeos isotopic to the identity; mapping class group $\mathrm{Diff}^+/\mathrm{Diff}_0$ ($\mathrm{SL}(2,\mathbb{Z})$ for T^2)	global anomalies (Chapter 10)
$\mathrm{Met}_{\mathrm{const}}$	metrics of constant curvature $+1, 0, -1$ (area 1 for T^2): the Weyl slice	
Teichmüller, moduli	$\mathcal{T}_p = \mathrm{Met}_{\mathrm{const}}/\mathrm{Diff}_0$, $\mathcal{M}_p = \mathcal{T}_p/\Gamma_\Sigma$; $\dim_{\mathbb{R}} = 0, 2, 6p-6$	Chapter 5, Chapter 11
Riemannian volume	$\sqrt{\det\langle e_i, e_j\rangle}\,\mathrm{d}^n x$ in any frame	WP measure
metric on Met	$\langle X, Y\rangle_{\mathrm{met}} = \int \sqrt{g}\,\mathrm{tr}(g^{-1}Xg^{-1}Y)$ (Nash (9.10)–(9.11)); DeWitt family	Chapter 10
diffeo invariance	$\langle \alpha^* X, \alpha^* Y\rangle_{\alpha^* g} = \langle X, Y\rangle_g$: change of variables	WP well defined
Pos(2)	positive 2×2 matrices $= \mathbb{R}_+ \times$ hyperbolic plane under $\mathrm{tr}(g^{-1}\mathrm{d}g)^2$	torus WP metric
metric on vect	$\langle V, W\rangle_{\mathrm{vect}} = \int \sqrt{g}\, g_{ab} V^a W^b$ (Nash (9.12))	ghost measure
Lie derivative	$\mathcal{L}_V g = \nabla_a V_b + \nabla_b V_a$ (Nash (9.13)); infinitesimal diffeo	Chapter T1, Chapter 11
Killing, conformal Killing	$\mathcal{L}_V g = 0$; $\mathcal{L}_V g = f g$. CKVs: V^z holomorphic. 6, 2, 0 real for $p = 0, 1, \geq 2$	Virasoro L_{-1}, L_0, L_1 (Chapter 11)

Nash §2, The space of metrics (Nash pp. 262–265): the operator P_1

Concept	Operational meaning	Used later
C_g, $C_g^\perp$	$C_g = \{\alpha g\}$ (Weyl directions); $C_g^\perp$ = traceless tensors; test: $\mathrm{tr}(g^{-1}X)$ (Nash (9.14)–(9.16))	
split of $\mathcal{L}_V g$	$\alpha = \nabla_c V^c$ and $h = \mathcal{L}_V g - \alpha g$ (Nash (9.17)–(9.18))	
$P_1 = A_g$	$(P_1 V)_{ab} = \nabla_a V_b + \nabla_b V_a - g_{ab}\nabla_c V^c$ (Nash (9.19))	Chapter 10
Cauchy–Riemann form	$(P_1 V)_{\bar z\bar z} = \rho\,\partial_{\bar z} V^z$: P_1 is $\bar\partial$ on K^{-1}	determinant line bundles
ellipticity of P_1	principal symbol $\propto \xi_{\bar z}$, invertible for $\xi \neq 0$ (Chapter 2)	Fredholm, finite index
$P_1^\dagger$	$(P_1^\dagger h)_b = -2\nabla^a h_{ab}$, i.e. $-\frac{4}{\rho}\partial_{\bar z} h_{zz}$	
quadratic differentials	$\mathrm{Ker}\, P_1^\dagger = \{\varphi(z)\mathrm{d}z^2, \varphi \text{ holomorphic}\} = H^0(K^2)$	Beltrami differentials, b zero modes
Range $\oplus$ Ker	$W = \mathrm{Range}\, A \oplus \mathrm{Ker}\, A^*$ orthogonally (Nash (9.20))	Hodge-type decompositions
three-way split	$T_g\mathrm{Met} = C_g \oplus \mathrm{Range}\, P_1 \oplus \mathrm{Ker}\, P_1^\dagger$ (Nash (9.21))	conformal gauge
Weitzenböck	$P_1^\dagger P_1 = -2(\nabla^2 + K)$ on vector fields; no kernel if $K < 0$	ζ-determinant of ghosts
index of P_1	$\dim\mathrm{Ker}\, P_1 - \dim\mathrm{Ker}\, P_1^\dagger = 3\chi$ (Riemann–Roch for K^{-1})	Chapter 4, Chapter 10
$T\mathcal{T}_p$	$T_{[g]}\mathcal{T}_p \cong \mathrm{Ker}\, P_1^\dagger$ (Nash (9.22))	WP metric
conformal anomaly	$\delta_\omega \log(\text{integrand}) = \frac{d-26}{24\pi}\int \sqrt{g} R\,\delta\omega$	Chapter 10 ($d = 26$), Chapter 11

Nash §3, The Weil–Petersson metric (Nash pp. 265–268)

Concept	Operational meaning	Used later
quotient metric	length of $\pi_* X$:= length of the horizontal lift of it (e.g. $S^2/\mathrm{SO}(2)$)	
$\langle\ ,\ \rangle_{\mathrm{const}}$	restriction of $\langle\ ,\ \rangle_{\mathrm{met}}$ to $\mathrm{Met}_{\mathrm{const}}$; Diff_0 acts by isometries	
Weil–Petersson metric	lift $U, V \in T\mathcal{T}_p$ horizontally to $\mathrm{Ker}\, P_1^\dagger$ and take $\langle\ ,\ \rangle_{\mathrm{met}}$ (Nash (9.25))	measure on $\mathcal{M}_p$
independence	different lifts differ by $\mathrm{Range}\, P_1$; different representatives by an isometry (Nash (9.26)–(9.29))	
Petersson form	$8\,\mathrm{Re}\int \varphi\bar{\varphi}'/\rho\, \mathrm{d}^2x$ for $h = \varphi \mathrm{d}z^2 + \text{c.c.}$	
torus WP	$2\,\lvert \mathrm{d}\tau\rvert^2/\tau_2^2$ on the upper half-plane	Chapter 11 (modular forms)
modular group	$\tau \mapsto (a\tau + b)/(c\tau + d)$, generated by T, S; isometries of WP	Chapter 11, Chapter 10
fundamental domain	$\mathcal{F} = \{\lvert\tau_1\rvert \le \frac{1}{2}, \lvert\tau\rvert \ge 1\}$; WP volume $2\pi/3$	torus amplitudes
conformal gauge	$g = \alpha^*(e^\sigma \hat{g}_t)$ (Nash (9.30)–(9.31))	Chapter 11
image measure	$\mathrm{d}^k(Av) = \sqrt{\det A^{\mathsf{T}} A}\, \mathrm{d}^k v$ (Nash (9.32))	Faddeev–Popov
triangular Jacobian	only diagonal blocks count: Weyl, $\sqrt{\det P_1^\dagger P_1}$, WP	
Z_p, $p \ge 2$	$\int_{\mathcal{M}_p} \mathrm{d}\mu_{\mathrm{wp}} (\det'(\Delta/2)/(1,1))^{-13} \sqrt{\det P_1^\dagger P_1}$ (Nash (9.35))	Chapter 10
CKVs in Z_p	$\det \to \det'$ and a factor $1/\mathrm{vol}\,\mathrm{CKG}$ for $p = 0, 1$ (Nash (9.33)–(9.34))	sphere amplitudes
ghosts b, c	$\sqrt{\det' P_1^\dagger P_1} = \int \mathcal{D}b\, \mathcal{D}c\, e^{-\frac{1}{4\pi}\int b P_1 c}$	$c = -26$ (Chapter 11), BRST
real vs complex	$\det \tilde{A}_g^* \tilde{A}_g = (\det \bar{\partial}_{K^*}^* \bar{\partial}_{K^*})^2$ (Nash (9.36))	Quillen metric (Chapter 10)

Beyond Nash: the formula producing numbers

Concept	Operational meaning	Used later
twisted circle	$\prod_m \lvert m + w\rvert^2$ regularised $= 4\,\lvert\sin \pi w\rvert^2 = e^{2\pi c}\left\lvert 1 - e^{2\pi i(b+ic)}\right\rvert^2$	torus determinants
Dedekind η	$\eta(\tau) = q^{1/24}\prod_{n\ge1}(1-q^n)$, $q = e^{2\pi i\tau}$; weight $\frac{1}{2}$	Chapter 11
$\det'\Delta$ on T^2	$\tau_2\,\lvert\eta(\tau)\rvert^4$; equivalently Kronecker's first limit formula	Chapter 11
Casimir energy	$\frac{1}{2}\sum n = \frac{1}{2}\zeta_R(-1) = -\frac{1}{24}$ per field: the $q^{-1/24}$	$L_0 - \frac{c}{24}$ (Chapter 11)
Z_1	$2^{-13}\int_{\mathcal{F}} \mathrm{d}^2\tau\, \tau_2^{-14}\,\lvert\eta\rvert^{-48}$: modular invariant, UV finite	Chapter 11
level matching	$\int \mathrm{d}\tau_1$ projects onto $N = \tilde{N}$	closed-string spectrum
tachyon, massless states	$N = 0$: $e^{4\pi\tau_2}$ divergence; $N = 1$: $24^2 = 576 = 299 + 276 + 1$ (graviton, B, dilaton)	
sphere gauge fixing	fix z_1, z_2, z_3; Jacobian $\lvert z_{12}z_{13}z_{23}\rvert^2 = \lvert\langle ccc\rangle\rvert^2$	vertex operators (Chapter 11)
punctured moduli	$\dim_{\mathbb{R}} \mathcal{M}_{p,n} = 6p - 6 + 2n$ when $2p - 2 + n > 0$	
higher genus	$\det'\Delta \propto Z'(1)$, $\det P_1^\dagger P_1 \propto Z(2)$ (Selberg zeta); Belavin–Knizhnik	Chapter 10

Formulas to keep at your fingertips

F1. *Polyakov.* $S[\phi, g] = \frac{T}{2}\int_\Sigma \mathrm{d}^2x\sqrt{g}\, g^{ab}\partial_a\phi\cdot\partial_b\phi$; $T_{ab} = \partial_a\phi\cdot\partial_b\phi - \frac{1}{2}g_{ab}(\partial\phi)^2$, $g^{ab}T_{ab} = 0$.

F2. *Gauge group.* $(\alpha, f)\cdot(\phi, g) = (\alpha^*\phi, f\,\alpha^*g)$; $(\alpha_1, f_1)(\alpha_2, f_2) = (\alpha_2\circ\alpha_1, f_1\,(f_2\circ\alpha_1))$.

F3. *Embedding integral.* $\int \mathcal{D}\phi\, e^{-S} \propto \left(\det{}'(\Delta_g/2)/(1,1)_g\right)^{-d/2}$, with $(1,1)_g = \int\sqrt{g}$.

F4. *Zeta determinant.* $\zeta_P(s) = \sum\lambda_n^{-s}$, $\det{}'P = e^{-\zeta_P'(0)}$, $\det{}'(cP) = c^{\zeta_P(0)}\det{}'P$; circle: $\det{}'\Delta = L^2$.

F5. *Heat kernel.* $\operatorname{Tr} e^{-t\Delta} = \frac{A}{4\pi t} + \frac{\chi}{6} + O(t)$, so $\zeta_\Delta(0) = \frac{\chi}{6} - 1$.

F6. *Metrics.* $\langle X, Y\rangle_{\mathrm{met}} = \int\sqrt{g}\,\operatorname{tr}(g^{-1}Xg^{-1}Y)$, $\langle V, W\rangle_{\mathrm{vect}} = \int\sqrt{g}\,g_{ab}V^aW^b$.

F7. *The operator P_1.* $(P_1V)_{ab} = \nabla_aV_b + \nabla_bV_a - g_{ab}\nabla_cV^c$, $(P_1^\dagger h)_b = -2\nabla^a h_{ab}$, $P_1^\dagger P_1 = -2(\nabla^2 + K)$.

F8. *Decomposition.* $T_g\mathrm{Met} = C_g \oplus \operatorname{Range}P_1 \oplus \operatorname{Ker}P_1^\dagger$, and $T_{[g]}\mathcal{T}_p \cong \operatorname{Ker}P_1^\dagger = H^0(K^2)$.

F9. *Index.* $\dim_{\mathbb{R}}\operatorname{Ker}P_1 - \dim_{\mathbb{R}}\operatorname{Ker}P_1^\dagger = 3\chi = 6 - 6p$.

F10. *Weil–Petersson.* $\langle U, V\rangle_{\mathrm{wp}} = \langle X^{\mathrm{hor}}, Y^{\mathrm{hor}}\rangle_{\mathrm{met}}$; torus: $2\,|\mathrm{d}\tau|^2/\tau_2^2$.

F11. *Image measure.* $\mathrm{d}^k(Av) = \sqrt{\det A^{\mathsf{T}}A}\,\mathrm{d}^kv$, so $\mathcal{D}(P_1V) = \sqrt{\det P_1^\dagger P_1}\,\mathcal{D}V$.

F12. *The amplitude.* In $d = 26$, $Z_p = \int_{\mathcal{M}_p}\mathrm{d}\mu_{\mathrm{wp}}\left(\det{}'(\Delta/2)/(1,1)\right)^{-13}\det{}'\bar\partial_{K^*}^*\bar\partial_{K^*}$, with $1/\mathrm{vol}\,\mathrm{CKG}$ for $p \le 1$.

F13. *Anomaly.* $\delta_\omega\log(\text{integrand}) = \frac{d-26}{24\pi}\int\sqrt{g}\,R\,\delta\omega$.

F14. *Torus.* $\det{}'\Delta = \tau_2\,|\eta|^4$, $Z_1 = 2^{-13}\int_{\mathcal{F}}\mathrm{d}^2\tau\,\tau_2^{-14}\,|\eta|^{-48}$.

Where the chapter comes from and where it is used later

This chapter is a meeting point. It imports the analysis of elliptic operators, the index theorem and the geometry of Riemann surfaces, and it exports the conformal anomaly, the ghost system and the torus partition function. Figure 9.26 shows the main routes.

- **The critical dimension.** The division by $\mathrm{vol}\,C_+^\infty$ was allowed only in $d = 26$. Chapter 10 derives the number 26 in three ways: by the heat-kernel computation of the Weyl variation previewed in Section 9.2.11, by the first Chern class of the determinant line bundle of $\bar\partial_{K^*}$ over $\mathcal{M}_p$ (families index theorem), and by Mumford's isomorphism.
- **Global anomalies and large diffeomorphisms.** The mapping class group Γ_Σ, which we divided out to pass from $\mathcal{T}_p$ to $\mathcal{M}_p$, is the prototype of a group of large gauge transformations. Modular invariance of Z_1 is the statement that the theory has no global anomaly under it (Chapter 10).
- **Conformal field theory.** In conformal gauge the residual symmetry is generated by the traceless T_{ab}; its modes form the Virasoro algebra, and T_{zz} of the b, c ghosts has central charge -26. The torus partition function $\tau_2^{-1/2}\,|\eta|^{-2}$ of one free boson, the Casimir shift $-\frac{c}{24}$ and modular invariance are the starting point of Chapter 11.
- **Gluing and topological field theory.** The trace picture of the vacuum amplitude (cutting a torus along a circle gives $\operatorname{Tr}e^{-\beta H}$) is the germ of the axioms of Chapter 12, where

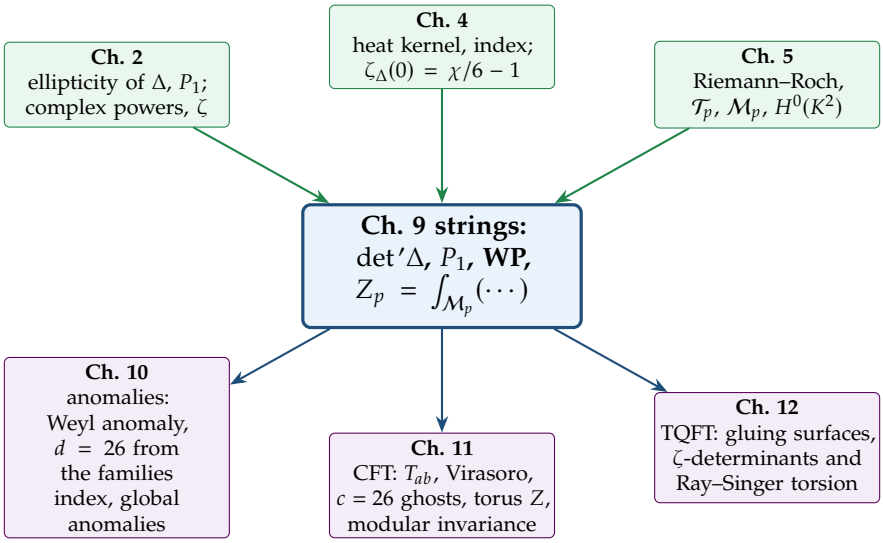


Figure 9.26: Where Chapter 9 comes from (green, above) and where it is used (purple, below). The three imports are exactly the three ingredients of Nash's formula: an elliptic operator theory to define the determinants, the heat kernel to evaluate them, and Riemann–Roch to count the moduli. The three exports are the three things the formula forces on us: the anomaly that fixes $d = 26$, the conformal field theory of the matter and ghost fields, and the idea of building amplitudes by gluing surfaces.

amplitudes are assigned to surfaces and glued along boundaries. The zeta-regularised determinants of this chapter reappear there as Ray–Singer torsion.

Watch out: The six confusions of this chapter

1. *The letter g.* Nash writes g for the metric and p for the genus. The Euler characteristic is $\chi = 2 - 2p$. In most physics texts g is the genus.
2. Diff$^+$ *versus* Diff$_0$. Dividing Met$_{\text{const}}$ by Diff$_0$ gives Teichmüller space, a smooth ball; dividing by all of Diff$^+$ gives the moduli space, which has orbifold points ($\tau = i$, $\tau = e^{i\pi/3}$). The difference is the discrete group Γ_Σ, and in the path integral it appears only as the factor $1/|\Gamma_\Sigma|$ that restricts $\int_{\mathcal{T}_p}$ to a fundamental domain.
3. det *versus* det$'$. Both Δ_g and (for $p \le 1$) P_1 have kernels. Their determinants are always primed, and the zero modes reappear elsewhere: the area $(1,1)_g$ and the volume of space-time for Δ, the factor $1/\operatorname{vol}\mathrm{CKG}$ for P_1. Forgetting either gives a wrong τ-dependence (for the torus) or an infinite answer (for the sphere).
4. *Classical versus quantum Weyl invariance.* The action is Weyl invariant in any d; the measure is not. Nash sets $d = 26$ at the point where he divides by $\operatorname{vol} C^\infty_+$. Everything after (9.8) assumes it.
5. *Real versus complex determinants.* P_1 is a real operator on real vector fields. As a complex operator it is $\bar\partial_{K^*}$ on one complex function V^z. Each complex eigenvalue appears twice among the real ones, so $\sqrt{\det P_1^\dagger P_1}$ (real) equals $\det \bar\partial^* \bar\partial$ (complex), up to the constant rescaling of the operator.
6. *Normalisations.* The factor 2 in the Weil–Petersson metric of the torus, the $\frac{1}{2}$ in $\Delta_g/2$ and the 2^{-13} in Z_1 are all conventions (the choice of $\langle\ ,\ \rangle_{\text{met}}$ and of T). They change overall constants, never the τ-dependence. The scaling law $\det'(cP) = c^{\zeta_P(0)} \det' P$ keeps track of them.

Try it: Test yourself on the chapter

(a) Why is Weyl invariance special to two dimensions?
Answer: under $g \mapsto fg$ in n dimensions, $\sqrt{g} \mapsto f^{n/2}\sqrt{g}$ and $g^{ab} \mapsto f^{-1}g^{ab}$, so $\sqrt{g}\,g^{ab} \mapsto f^{n/2-1}\sqrt{g}\,g^{ab}$, which is unchanged only for $n = 2$.

(b) How many real moduli does a genus-3 surface with 2 punctures have?
Answer: $6 \cdot 3 - 6 + 2 \cdot 2 = 16$.

(c) Compute $\det'(4\Delta)$ on a circle of length L.
Answer: $4^{\zeta_\Delta(0)}L^2 = 4^{-1}L^2 = L^2/4$, which is the determinant of Δ on a circle of length $L/2$, as it should be.

(d) Why is there no Faddeev–Popov factor $1/\mathrm{vol}\,\mathrm{CKG}$ for $p \geq 2$?
Answer: $K < 0$ makes $P_1^\dagger P_1 = -2(\nabla^2 + K)$ strictly positive, so $\mathrm{Ker}\,P_1 = 0$: there are no conformal Killing vectors, and Diff_0 acts freely on $\mathrm{Met}_{\mathrm{const}}$.

(e) Show that the hyperbolic area of $\mathcal{F}$ is $\pi/3$.
Answer: $\int_{-1/2}^{1/2} \mathrm{d}\tau_1 \int_{\sqrt{1-\tau_1^2}}^{\infty} \frac{\mathrm{d}\tau_2}{\tau_2^2} = \int_{-1/2}^{1/2} \frac{\mathrm{d}\tau_1}{\sqrt{1-\tau_1^2}} = 2\arcsin\frac{1}{2} = \frac{\pi}{3}$; the WP volume is twice this.

Chapter 10

Anomalies

Where this chapter is going

A symmetry of a classical action need not survive quantisation. When it does not, we say the theory has an **anomaly**. This chapter is about the anomalies of chiral fermions, and about the surprising fact that every one of them is a piece of topology. The chain of ideas is short. The fermion path integral produces a determinant, $\det \not{D}_A$. The symmetry we care about (gauge invariance, coordinate invariance, conformal invariance) acts on the background field A or on the metric ρ. The anomaly is the failure of the determinant to be invariant. The index theorem tells us when a determinant is forced to vanish; the *families* index theorem, applied to the whole family $\{\not{D}_A\}$, tells us when the determinant cannot even be defined as an invariant function. The answer is a cohomology class: a closed 1-form on the symmetry group $\mathcal{G}$, or equivalently a first Chern class of a line bundle over the orbit space $\mathcal{A}/\mathcal{G}$. Nash's §§1–3, which we expand into Sections 10.1 to 10.3 and 10.5 to 10.7, do this for gauge fields and for gravity. His §4 applies it to the string worldsheet and finds $d = 26$. His §5 treats *global* anomalies, which are torsion classes invisible to de Rham cohomology. His §6 retells the whole story in Hamiltonian language, where the anomaly becomes a failure of Gauss's law.

10.1 Introduction: what an anomaly is

Here is the puzzle that started it all. The neutral pion π^0 decays into two photons with a lifetime of about 8×10^{-17} s. In the late 1960s the best theory of pions, current algebra with an exactly conserved axial current, predicted that this decay should be *forbidden* for massless quarks (Sutherland and Veltman). Nature disagreed by many orders of magnitude. The resolution, due to Adler and to Bell and Jackiw, is that the axial current is conserved in the classical theory but *not* in the quantum theory. A single fermion loop, the triangle graph, spoils the conservation law, and the rate it predicts is the observed one.

That was the first **anomaly**. It is not a mistake in the calculation. It is a genuine feature of quantum field theory: the path-integral measure for chiral fermions cannot be made invariant under every classical symmetry at once. In this section we make that sentence precise, set up our conventions, and preview why the answer turns out to be topological.

Reading Nash (p. 269, §1)

Nash's introduction is one page. He says that the word "anomaly" "usually has to do with strange or even pathological behaviour of chiral fermions when coupled to gauge fields", names the Adler–Bell–Jackiw triangle anomaly for $G = \mathrm{U}(1)$, lists the later generalisations (Yang–Mills, gravity with $\mathrm{Diff}^+(M)$ in place of $\mathcal{G}$, the conformal anomaly

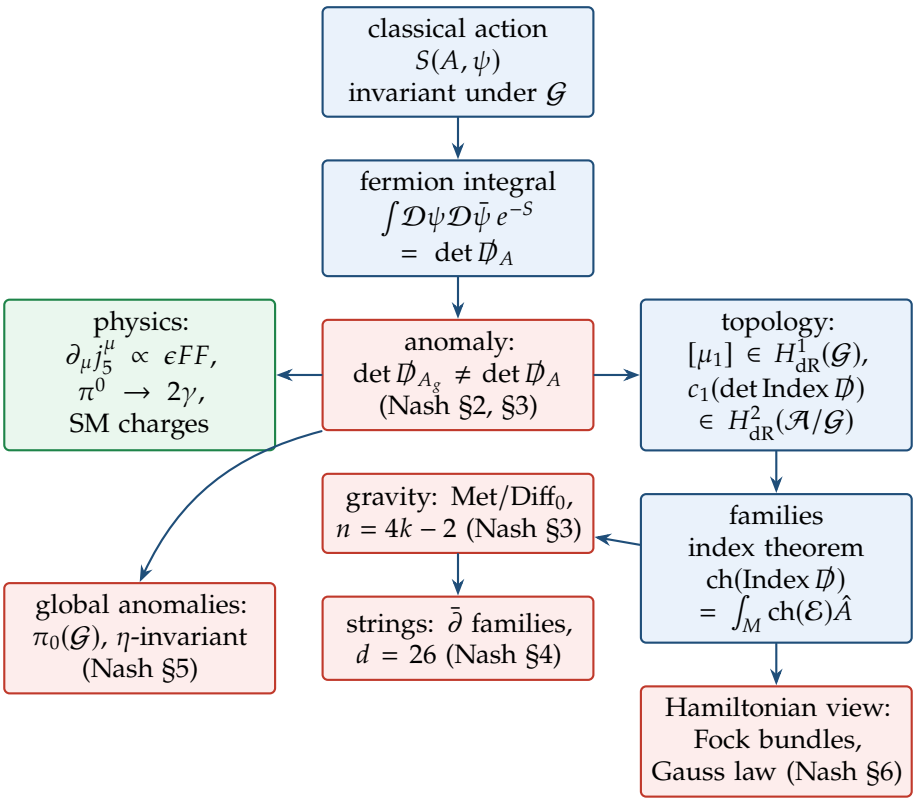


Figure 10.1: The logic of Chapter 10. Blue boxes are the mathematical spine (determinants, cohomology, the families index theorem); red boxes are the kinds of anomaly the chapter treats; the green box is where the anomaly touches experiment. Every red box is obtained by feeding a different family of elliptic operators into the same blue machine.

of the bosonic string, the Hamiltonian view), and then gives his definition: an anomaly is "some obstruction to the defining of a quantum theory with the same invariance groups as its underlying classical theory". The obstruction has a topological origin: it lives in the cohomology $H^*(\mathcal{G})$ of the invariance group. Everything in this part unpacks these statements.

10.1.1 Conventions: Euclidean spinors in two and four dimensions

Nash works on a compact Riemannian (Euclidean) manifold M of even dimension $n = 2m$. We do the same, and we fix the representation of every object now, because the sign of an anomaly depends on it.

Gamma matrices. In n Euclidean dimensions we use Hermitian matrices obeying

$$\gamma_\mu \gamma_\nu + \gamma_\nu \gamma_\mu = 2\delta_{\mu\nu}\,\mathbf{1}, \qquad \gamma_\mu^\dagger = \gamma_\mu, \tag{10.1}$$

so $\gamma_\mu^2 = \mathbf{1}$ for each μ (no sum). Indices are raised and lowered with $\delta_{\mu\nu}$, so $\gamma^\mu = \gamma_\mu$. The reminder of the general theory is in Chapter T4; here we only need two explicit cases.

- $n = 2$**.** Take $\gamma_1 = \sigma_1$, $\gamma_2 = \sigma_2$ (Pauli matrices).
- $n = 4$**.** Take, in 2×2 blocks,

$$\gamma_j = \begin{pmatrix} 0 & i\sigma_j \\ -i\sigma_j & 0 \end{pmatrix} \ (j = 1,2,3), \qquad \gamma_4 = \begin{pmatrix} 0 & \mathbf{1} \\ \mathbf{1} & 0 \end{pmatrix}. \tag{10.2}$$

Chirality. In $n = 2m$ dimensions the **chirality operator** is

$$\gamma_5 := (-i)^m\, \gamma_1 \gamma_2 \cdots \gamma_{2m}. \tag{10.3}$$

(Nash and most physicists write γ_5 in every even dimension; in $n = 2$ it is often called γ_3.) It squares to $\mathbf{1}$, is Hermitian, and anticommutes with every γ_μ. Its ± 1 eigenspaces are the positive and negative **chirality** spinors, the spaces $E^\pm$ of Nash's (10.2).

Worked example 10.1: The chirality matrix in two and four dimensions

For $n = 2$:

$$\begin{aligned} \gamma_5 &\overset{(1)}{=} -i\,\sigma_1\sigma_2 \\ &\overset{(2)}{=} -i\,(i\sigma_3) \\ &= \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \end{aligned}$$

(1) Definition (10.3) with $m = 1$.
(2) The Pauli identity $\sigma_1\sigma_2 = i\sigma_3$.

So positive chirality is the upper component. For $n = 4$, multiply the blocks of (10.2). A product of two block-off-diagonal matrices is block diagonal: $\left(\begin{smallmatrix} 0 & a \\ b & 0 \end{smallmatrix}\right)\left(\begin{smallmatrix} 0 & c \\ d & 0 \end{smallmatrix}\right) = \left(\begin{smallmatrix} ad & 0 \\ 0 & bc \end{smallmatrix}\right)$. Hence

$$\gamma_1\gamma_2\gamma_3\gamma_4 \overset{(1)}{=} \begin{pmatrix} (i\sigma_1)(-i\sigma_2) & 0 \\ 0 & (-i\sigma_1)(i\sigma_2) \end{pmatrix} \begin{pmatrix} (i\sigma_3)(\mathbf{1}) & 0 \\ 0 & (-i\sigma_3)(\mathbf{1}) \end{pmatrix}$$

$$\overset{(2)}{=} \begin{pmatrix} \sigma_1\sigma_2 \cdot i\sigma_3 & 0 \\ 0 & \sigma_1\sigma_2 \cdot (-i\sigma_3) \end{pmatrix}$$
$$\overset{(3)}{=} \begin{pmatrix} i\sigma_3 \cdot i\sigma_3 & 0 \\ 0 & i\sigma_3 \cdot (-i\sigma_3) \end{pmatrix} = \begin{pmatrix} -\mathbf{1} & 0 \\ 0 & \mathbf{1} \end{pmatrix}.$$

(1) Pair up $(\gamma_1\gamma_2)$ and $(\gamma_3\gamma_4)$ and use the block rule.
(2) $(i)(-i) = 1$ in both entries of the first factor.
(3) $\sigma_1\sigma_2 = i\sigma_3$ and $\sigma_3^2 = \mathbf{1}$.

Therefore $\gamma_5 = (-i)^2\gamma_1\gamma_2\gamma_3\gamma_4 = \operatorname{diag}(\mathbf{1}, -\mathbf{1})$: again positive chirality is the upper two components. A four-component spinor is a pair $\psi = (\psi_+, \psi_-)$ of two-component **Weyl spinors**.

Gauge fields. We follow Nash and the mathematicians: the gauge potential $A = A_\mu \, \mathrm{d}x^\mu$ is a 1-form with values in the Lie algebra $\mathfrak{g}$ of *anti-Hermitian* matrices, the covariant derivative is $D_\mu = \partial_\mu + A_\mu$, the curvature is $F = \mathrm{d}A + A \wedge A$ with components $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu + [A_\mu, A_\nu]$, and a gauge transformation $g : M \to G$ acts by

$$A \mapsto A_g = g^{-1}Ag + g^{-1}\mathrm{d}g, \qquad \psi \mapsto g^{-1}\psi, \qquad F \mapsto g^{-1}Fg. \tag{10.4}$$

The (Euclidean) Dirac operator is

$$\not{D}_A = \gamma^\mu D_\mu = \gamma^\mu(\partial_\mu + A_\mu). \tag{10.5}$$

Because γ^μ is Hermitian and D_μ is anti-self-adjoint, $\not{D}_A$ is *anti*-self-adjoint, so $i\not{D}_A$ is self-adjoint with real eigenvalues and $\not{D}_A^2 = -(i\not{D}_A)^2 \le 0$. Since γ_5 anticommutes with every γ^μ and commutes with A_μ, $\not{D}_A$ exchanges chiralities: it maps E^+ to E^- and back.

Dictionary: mathematicians' and physicists' gauge fields

Physicists write a Hermitian potential $\mathsf{A}_\mu = \mathsf{A}_\mu^a T^a$ and a coupling e (or g), with $D_\mu = \partial_\mu - ie\mathsf{A}_\mu$. The two conventions are related by

$$A_\mu^{\text{math}} = -ie\,\mathsf{A}_\mu^{\text{phys}}, \qquad F_{\mu\nu}^{\text{math}} = -ie\,\mathsf{F}_{\mu\nu}^{\text{phys}}.$$

For QED with a fermion of charge e, $A_\mu = -ie\,a_\mu$ with a_μ the real electromagnetic potential, and $F_{\mu\nu} = -ie\,f_{\mu\nu}$. Anything quadratic in F therefore picks up $(-ie)^2 = -e^2$: keep this in mind when comparing our formulas with a QFT textbook.

10.1.2 Classical symmetries of massless fermions

Take the Euclidean action of a fermion in a fixed background gauge field,

$$S_F = \int_M \mathrm{d}^n x \, \bar\psi \not{D}_A \psi. \tag{10.6}$$

(In Euclidean signature ψ and $\bar\psi$ are *independent* Grassmann variables; $\bar\psi$ is not $\psi^\dagger\gamma^0$.) This action has two global U(1) symmetries.

- The **vector** rotation $\psi \mapsto e^{i\alpha}\psi$, $\bar\psi \mapsto \bar\psi e^{-i\alpha}$. It rotates both chiralities by the same phase. Its current is $j^\mu = \bar\psi\gamma^\mu\psi$, and its charge is fermion number.
- The **axial** (chiral) rotation $\psi \mapsto e^{i\alpha\gamma_5}\psi$, $\bar\psi \mapsto \bar\psi e^{i\alpha\gamma_5}$. It rotates positive and negative chirality by opposite phases. Its current is $j_5^\mu = \bar\psi\gamma^\mu\gamma_5\psi$, and its charge counts $N_+ - N_-$.

Why the axial rotation is a symmetry. The Noether trick is to let the parameter depend on position and see what the action does. With $\alpha = \alpha(x)$ infinitesimal, $\delta\psi = i\alpha\gamma_5\psi$ and $\delta\bar\psi = i\alpha\bar\psi\gamma_5$, and

$$\begin{aligned}\delta S_F &\overset{(1)}{=} \int \mathrm{d}^n x \left[i\alpha\,\bar\psi\gamma_5 \not{D}_A\psi + \bar\psi\,\not{D}_A(i\alpha\gamma_5\psi)\right] \\ &\overset{(2)}{=} \int \mathrm{d}^n x \left[i\alpha\,\bar\psi\gamma_5 \not{D}_A\psi + i(\partial_\mu\alpha)\bar\psi\gamma^\mu\gamma_5\psi + i\alpha\,\bar\psi\,\not{D}_A\gamma_5\psi\right] \\ &\overset{(3)}{=} i\int \mathrm{d}^n x\,(\partial_\mu\alpha)\,j_5^\mu \overset{(4)}{=} -i\int \mathrm{d}^n x\,\alpha\,\partial_\mu j_5^\mu .\end{aligned} \tag{10.7}$$

(1) Vary $\bar\psi$ and ψ in (10.6).
(2) Leibniz: $\not{D}_A(\alpha\chi) = \gamma^\mu(\partial_\mu\alpha)\chi + \alpha\not{D}_A\chi$.
(3) $\not{D}_A\gamma_5 = -\gamma_5\not{D}_A$, so the first and third terms cancel.
(4) Integrate by parts; M is closed, so Stokes' theorem gives no boundary term.

For constant α the variation vanishes: that is the symmetry. On solutions of the equations of motion $\delta S_F = 0$ for *every* $\alpha(x)$, so classically

$$\partial_\mu j^\mu = 0, \qquad \partial_\mu j_5^\mu = 0. \tag{10.8}$$

The same computation with the vector rotation gives the first equation. Both would fail if we added a mass term $m\bar\psi\psi$: the vector rotation leaves it alone, but $\delta(m\bar\psi\psi) = 2im\alpha\,\bar\psi\gamma_5\psi$, so a mass breaks the axial symmetry *explicitly*. An anomaly is a breaking of a different kind. It is present even at $m = 0$, and it comes from quantisation.

Physical picture: Chirality as direction of motion in $1+1$ dimensions

Rotate back to $1+1$ Minkowski space for a moment (coordinates t, x). The massless Dirac equation splits into $(\partial_t + \partial_x)\psi_+ = 0$ and $(\partial_t - \partial_x)\psi_- = 0$ [QFT notes (22.2.10)]. So $\psi_+(t,x) = f(x-t)$ moves right at the speed of light and $\psi_-(t,x) = h(x+t)$ moves left. Classically the two species never talk to each other, so N_+ and N_- are separately conserved. The anomaly will say: in an electric field, right-movers are converted into left-movers, and $N_+ - N_-$ changes by an amount fixed by the flux.

10.1.3 The first anomaly: Adler, Bell and Jackiw

Couple a massless Dirac fermion to electromagnetism in four dimensions. Both currents in (10.8) are conserved classically. Quantum mechanically, the vector current stays conserved (it must: the photon couples to it, and gauge invariance needs $\partial_\mu j^\mu = 0$), but the axial current obeys

$$\partial_\mu j_5^\mu = \frac{e^2}{16\pi^2}\,\epsilon^{\mu\nu\rho\sigma} f_{\mu\nu} f_{\rho\sigma} = \frac{e^2}{2\pi^2}\,\mathbf{E}\cdot\mathbf{B} \qquad \text{(Euclidean; Minkowski signs differ, see Section 10.4.1).} \tag{10.9}$$

This is the **Adler–Bell–Jackiw (ABJ) anomaly**. We will derive it four ways in Section 10.3: from the vacuum polarisation in two dimensions, from a gauge-invariant point-split current, from the triangle graph, and from Fujikawa's analysis of the path-integral measure. The last route will show that the *integrated* anomaly is twice the index of the Dirac operator, which is where topology enters.

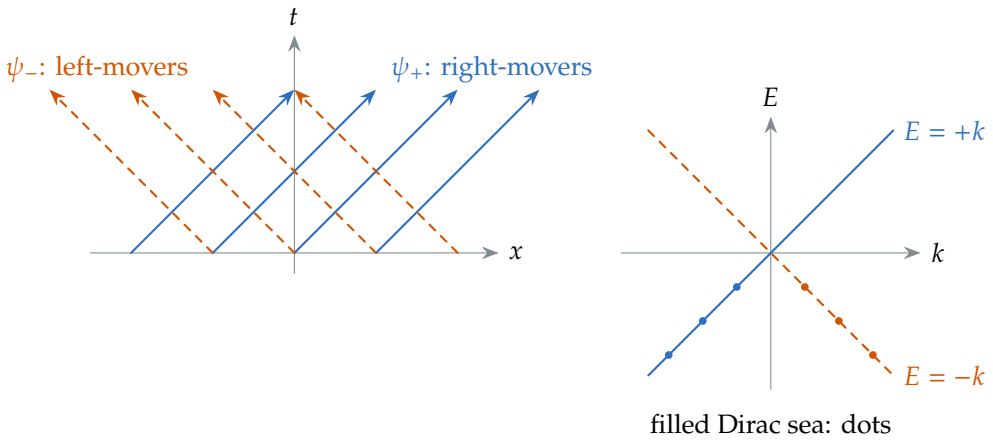


Figure 10.2: Left: in 1 + 1 dimensions a massless positive-chirality fermion moves right and a negative-chirality one moves left, along light rays. Right: their dispersion relations $E = \pm k$. The vacuum fills every negative-energy level of both branches (dots). The anomaly (Section 10.3.5) is the statement that an electric field slides the dots along the lines, pushing one branch's levels up through $E = 0$ and the other's down.

Physical picture: Why $\pi^0 \to \gamma\gamma$ needs the anomaly

The π^0 is (approximately) the Goldstone boson of the spontaneously broken axial isospin symmetry of the u and d quarks. Its field is proportional to the divergence of the corresponding axial current, $\partial_\mu j_5^{\mu 3} \approx f_\pi m_\pi^2 \pi^0$. If that current were conserved at the quantum level, the amplitude for $\pi^0 \to \gamma\gamma$ would vanish as $m_\pi \to 0$. The anomaly adds to $\partial_\mu j_5^{\mu 3}$ an $\epsilon f f$ term whose coefficient is $N_c\,(Q_u^2 - Q_d^2) = 3\left(\frac{4}{9} - \frac{1}{9}\right) = 1$. The resulting width,

$$\Gamma(\pi^0 \to \gamma\gamma) = \frac{\alpha^2 m_\pi^3}{64\pi^3 f_\pi^2} \approx 7.7\ \text{eV} \qquad (f_\pi \approx 92\ \text{MeV}),$$

agrees with experiment, and the factor $N_c = 3$ was one of the early pieces of evidence for colour [QFT notes §22.6.6].

10.1.4 Anomalies as obstructions

Nash's definition is more general than "a current that is not conserved".

Working definition 10.1 (anomaly)

Let a classical theory with fields Φ and action $S[\Phi]$ be invariant under a group Γ acting on the fields. The theory has an **anomaly** under Γ if the quantum theory cannot be defined so that its generating functional [the path integral Z as a function of the background fields, e.g. $Z_F[A]$ as a function of A] is Γ-invariant. *Operational test:* integrate out the fermions [do the $\psi, \bar\psi$ path integral with A held fixed] in a background field A to get $Z_F[A] = \det \not{D}_A$ (suitably regularised). Transform the background, $A \mapsto A_g$. If $Z_F[A_g] \neq Z_F[A]$, and *no* choice of local counterterm [an integral over M of a polynomial in A and its derivatives, added by hand. This is the freedom in choosing the regularisation] added to $\ln Z_F$ restores equality, there is an anomaly.

Two things about this definition matter.

1. The group Γ can be anything that acts on the fields. The cases in this chapter are listed in Table 10.1. The same mathematics handles all of them, because in each case the fermions see the symmetry through a family of Dirac-type operators.

2. The obstruction is *global on the group*. For gauge symmetry it is not enough to look at one gauge transformation; we must ask whether $\ln Z_F$ can be made into a function on the orbit space. This is a question about loops in $\mathcal{G}$ and about line bundles over $\mathcal{A}/\mathcal{G}$.

anomaly	group Γ	operators	where
chiral (ABJ)	global $\mathrm{U}(1)_A$	$\not{D}_A$, A fixed	Section 10.3; harmless, even useful
gauge (local)	$\mathcal{G} = \mathrm{Map}(M, G)$	$\{\not{D}_A\}_{A\in\mathcal{A}}$	Sections 10.5 and 10.6; must cancel
gravitational	$\mathrm{Diff}_0(M)$	$\{\not{D}_\rho\}_{\rho\in\mathrm{Met}(M)}$	Section 10.7; exists in $n = 4k-2$
Lorentz	$\mathcal{G}_L = \mathrm{Map}(M, \mathrm{SO}(n))$	vielbein rotations	Section 10.7.5; tradeable for the above
conformal	Weyl rescalings	$\{\bar\partial_g\}$ on the worldsheet	Nash §4; fixes $d = 26$
global	$\pi_0(\mathcal{G})$, $\pi_0(\mathrm{Diff}^+)$	S^1-families, η-invariant	Nash §5

Table 10.1: The kinds of anomaly in Chapter 10. The first two columns say what symmetry fails; the third says which family of operators the fermions see; the fourth says where it is treated and what physics does with it. Only anomalies in *gauged* symmetries (rows 2–6) are inconsistencies; an anomaly in a global symmetry (row 1) is just a prediction.

Watch out: Global symmetry anomalies are allowed; gauge anomalies are not

The ABJ anomaly breaks a *global* symmetry. That is fine: it just means the symmetry is not there, and the world ($\pi^0 \to \gamma\gamma$) shows it. An anomaly in a *gauge* symmetry is fatal. Gauge invariance is what removes the unphysical polarisations of the gauge field and makes the theory unitary and renormalisable; without it, the quantum theory is inconsistent. So for gauge anomalies the job is to arrange the fermion content so that they *cancel*. The Standard Model does this (Section 10.5.7).

10.1.5 Where the topology comes from: a preview

Why should topology have anything to do with this? Here is the idea in four steps. The rest of this part fills in each one.

1. **The fermions give a determinant.** Integrating out ψ and $\bar\psi$ gives $Z_F[A] = \det \not{D}_A$ (Section 10.2.5). For a chiral fermion $\not{D}_A : \Gamma(E^+) \to \Gamma(E^-)$ maps one space to a *different* one, so its determinant is not a number but an element of a line (Section 10.6.3).

2. **Determinants can vanish, and the index says when.** If $\ker \not{D}_A \neq 0$ the determinant is zero. The index $\dim\ker \not{D}_A - \dim \mathrm{Coker}\, \not{D}_A$ is topological, so for nonzero index the zero is forced (Section 10.2.4).

3. **The anomaly is a winding.** Along a loop $g(t)$ in the gauge group, $\det \not{D}_{A_{g(t)}}$ is a loop in $\mathbb{C}\setminus\{0\}$ (after removing zero modes). If it winds around the origin, $\ln\det$ cannot be a gauge invariant function. The winding is measured by a closed but not exact 1-form μ_1 on $\mathcal{G}$ (Section 10.6.2).

4. **The winding is a Chern class.** Equivalently, the determinants form a line bundle over $\mathcal{A}/\mathcal{G}$, and the anomaly is its first Chern class, which the families index theorem computes as $\int_M \mathrm{ch}(\mathcal{E})\hat{A}$ (Section 10.6.4).

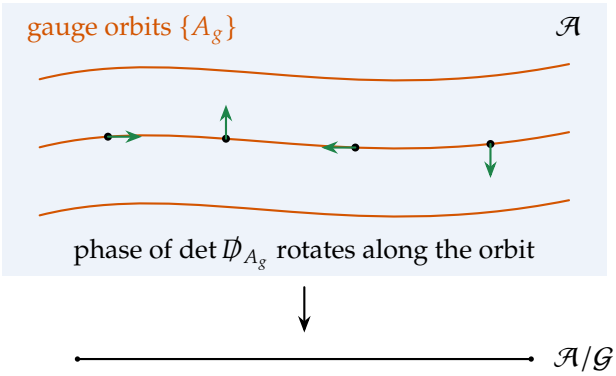


Figure 10.3: The picture behind every local anomaly. The space $\mathcal{A}$ of gauge fields is foliated by gauge orbits (orange). The classical action is constant along each orbit, so the path integral should descend to the quotient $\mathcal{A}/\mathcal{G}$ (bottom). The fermion determinant is not constant along the orbits: its phase (green arrows) rotates as we move along one. When the phase winds by a nonzero amount around a closed loop in the orbit, no redefinition can make it constant.

Recap

- An anomaly is a classical symmetry that the quantum theory cannot keep. For fermions it comes from the path-integral measure, i.e. from the determinant of the Dirac operator.
- Conventions: Hermitian Euclidean γ_μ, $\gamma_5 = (-i)^m \gamma_1 \cdots \gamma_{2m}$ (= σ_3 in 2D, diag($\mathbf{1}, -\mathbf{1}$) in 4D with (10.2)); anti-Hermitian A, $D = \mathrm{d} + A$, $\not{D}_A = \gamma^\mu D_\mu$ is anti-self-adjoint and flips chirality.
- Classically the vector and axial currents are both conserved (10.8). Quantum mechanically the axial current obeys the ABJ equation (10.9).
- Anomalies in global symmetries are predictions; anomalies in gauge symmetries must cancel.
- Topology enters through determinants, indices, windings of phases along loops in $\mathcal{G}$, and Chern classes of line bundles over $\mathcal{A}/\mathcal{G}$.

10.2 Anomalies and Yang–Mills theories I: chiral fermions, the index and the determinant

Nash's §2 is long, and it does three different jobs. It sets up a chiral gauge theory and its path integral; it shows that the fermion determinant is not gauge invariant, and explains what that means for currents and Feynman graphs; and it reinterprets the failure as a cohomology class. We split it into four sections. This one does the set-up. The question it answers is: *what exactly is the object whose gauge invariance fails?* The answer will be the fermion determinant, and we shall see already here that the index of the Dirac operator controls whether that determinant is zero.

Nash first separates two kinds of anomaly.

- A **local** (or **perturbative**) anomaly is visible in an *infinitesimal* gauge transformation $g = e^{tf}$ with $t \to 0$. It shows up in Feynman graphs, and in the non-conservation of a current. It is measured by a de Rham class, i.e. by derivatives.
- A **global** anomaly is visible only under gauge transformations that are *not* connected to

the identity. It is invisible to any derivative, hence to perturbation theory.

Until the global anomalies of Nash's §5 (treated in the second half of this chapter), "anomaly" means a local anomaly.

10.2.1 The action of a chiral gauge theory

We work on a compact Riemannian manifold M of even dimension $n = 2m$, with gauge group G (think U(1), SU(2) or SU(N)). Nash's action is

$$S(A,\psi) = \|F\|^2 + \langle \psi, \not{D}_A \psi\rangle = -\int_M \operatorname{tr}(F\wedge\star F) + \frac{1}{2}\int_M \bar\psi(x)\gamma_\mu\big(\partial^\mu + \Gamma^\mu + A^\mu\big)(1+\gamma_5)\psi(x), \quad (10.10)$$

[Nash p. 270, (10.1)]. Let us take it apart.

The Yang–Mills term is positive. With anti-Hermitian F, the trace $\operatorname{tr}(F_{\mu\nu}F_{\mu\nu})$ is *negative*, which is why Nash puts a minus sign in front. In components,

$$\begin{aligned} -\operatorname{tr}(F\wedge\star F) &\overset{(1)}{=} -\tfrac{1}{4}\operatorname{tr}(F_{\mu\nu}F_{\rho\sigma})\,\mathrm{d}x^\mu\wedge\mathrm{d}x^\nu\wedge\star(\mathrm{d}x^\rho\wedge\mathrm{d}x^\sigma) \\ &\overset{(2)}{=} -\tfrac{1}{2}\operatorname{tr}(F_{\mu\nu}F_{\mu\nu})\,\mathrm{dvol} \\ &\overset{(3)}{=} \tfrac{1}{2}\sum_{\mu,\nu}\operatorname{tr}\big(F^\dagger_{\mu\nu}F_{\mu\nu}\big)\,\mathrm{dvol} \;\geq\; 0. \end{aligned} \quad (10.11)$$

(1) Write $F = \frac{1}{2}F_{\mu\nu}\mathrm{d}x^\mu\wedge\mathrm{d}x^\nu$ and the same for $\star F$.
(2) For an orthonormal frame, $\mathrm{d}x^\mu\wedge\mathrm{d}x^\nu\wedge\star(\mathrm{d}x^\rho\wedge\mathrm{d}x^\sigma) = (\delta^{\mu\rho}\delta^{\nu\sigma} - \delta^{\mu\sigma}\delta^{\nu\rho})\,\mathrm{dvol}$, which is the defining property of the Hodge star (Chapter T1); the two terms give equal contributions, doubling $\frac{1}{4}$ to $\frac{1}{2}$.
(3) $F^\dagger_{\mu\nu} = -F_{\mu\nu}$, and $\operatorname{tr}(X^\dagger X) = \sum_{ij}|X_{ij}|^2 \geq 0$.

So $\|F\|^2 = \int \frac{1}{2}\sum \operatorname{tr}(F^\dagger_{\mu\nu}F_{\mu\nu})$ is the usual positive Euclidean Yang–Mills action. In the physicists' convention of the dictionary in Section 10.1.1, $F = -ig\,\mathsf{F}^a T^a$ with $\operatorname{tr} T^aT^b = \frac{1}{2}\delta^{ab}$ and it becomes $\frac{g^2}{4}\int \mathsf{F}^a_{\mu\nu}\mathsf{F}^a_{\mu\nu}$, the familiar $\frac{1}{4}F^2$ after rescaling $\mathsf{A}\to\mathsf{A}/g$.

The fermion term. The operator in (10.10) contains three connections at once:

- ∂^μ is the flat derivative;
- Γ^μ is the **spin connection**, built from the Levi-Civita connection of the metric. It tells a spinor how to parallel transport on a curved M (Chapter T4); on flat $\mathbb{R}^n$ or a flat torus it vanishes;
- A^μ is the Yang–Mills connection, acting on the colour index of ψ.

The factor $(1+\gamma_5)$ is (twice) the **chiral projector**:

$$P_\pm := \tfrac{1}{2}(1\pm\gamma_5), \qquad P_\pm^2 = P_\pm, \quad P_+P_- = 0, \quad P_+ + P_- = \mathbf{1}. \quad (10.12)$$

Since $\gamma_5^2 = \mathbf{1}$, these relations follow by multiplying out, e.g. $P_+^2 = \frac{1}{4}(1+2\gamma_5+\gamma_5^2) = \frac{1}{2}(1+\gamma_5)$. In the basis of Example 10.1, $P_+ = \operatorname{diag}(\mathbf{1},0)$ picks the upper (Weyl) components. So the fermion term is

$$\begin{aligned} \tfrac{1}{2}\bar\psi\,\not{D}_A(1+\gamma_5)\psi &\overset{(1)}{=} \bar\psi\,\not{D}_A P_+\psi \\ &\overset{(2)}{=} \bar\psi\,P_-\not{D}_A P_+\psi. \end{aligned} \quad (10.13)$$

(1) Definition (10.12).
(2) $\not{D}_A\gamma_5 = -\gamma_5\not{D}_A$ gives $\not{D}_A P_+ = P_-\not{D}_A$, and $P_-^2 = P_-$, so $\not{D}_A P_+ = P_-\not{D}_A P_+$.

Only the positive-chirality part $\psi_+ = P_+\psi$ of ψ and the negative-chirality part $\bar\psi P_-$ of $\bar\psi$ appear. That is what "chiral fermion" means: a Weyl fermion, one chirality only, coupled to the gauge field.

Watch out: Factors of two and the sign of γ_5

Nash writes $\frac{1}{2}\gamma_\mu(\dots)(1+\gamma_5)$, i.e. $\not{D}_A P_+$. Many QFT books use $\frac{1}{2}(1-\gamma_5)$ for left-handed fields, and Minkowski γ_5 conventions differ by factors of i. None of this changes the physics; it changes the overall sign of the anomaly. In this chapter the chiral fermion is always ψ_+ with $\gamma_5\psi_+ = +\psi_+$ and γ_5 as in (10.3).

Dictionary: Weyl fermions in Euclidean and Minkowski language

In Minkowski space a single left-handed Weyl fermion χ_L is a two-component spinor with action $i\chi_L^\dagger\bar\sigma^\mu D_\mu\chi_L$ [QFT notes §22.7.1]. Its Euclidean continuation is our $\bar\psi P_-\not{D}_A P_+\psi$: $\bar\psi P_-$ plays the role of $\chi_L^\dagger$, and the Euclidean $\bar\psi$ is an *independent* variable, so nothing forces $\bar\psi P_-$ to be the conjugate of $P_+\psi$. The choice ψ_+ versus ψ_- corresponds to left- versus right-handed, depending on conventions.

10.2.2 Why M must be a spin manifold

To write $\bar\psi\not{D}\psi$ on a curved M we need spinors, i.e. sections of a spinor bundle. Locally that is easy: pick an orthonormal frame (vielbein) on each coordinate patch and use constant gamma matrices in that frame. On an overlap the two frames differ by a rotation $h_{\alpha\beta}(x) \in \mathrm{SO}(n)$. A spinor must transform by a matrix $\tilde h_{\alpha\beta}(x) \in \mathrm{Spin}(n)$ that covers $h_{\alpha\beta}$ under the double cover $\mathrm{Spin}(n) \to \mathrm{SO}(n)$. The lifts are only determined up to sign, and the question is whether the signs can be chosen consistently on triple overlaps, $\tilde h_{\alpha\beta}\tilde h_{\beta\gamma}\tilde h_{\gamma\alpha} = +\mathbf{1}$.

Working definition 10.2 (spin manifold)

M is **spin** if it is orientable [the frames can be chosen with one handedness everywhere, so every $h_{\alpha\beta}$ lies in $\mathrm{SO}(n)$] and the frame transition functions can be lifted to $\mathrm{Spin}(n)$ with the cocycle condition [$\tilde h_{\alpha\beta}\tilde h_{\beta\gamma}\tilde h_{\gamma\alpha} = +\mathbf{1}$ on every triple overlap] intact. *Test:* compute the **Stiefel–Whitney classes** $w_1(TM), w_2(TM) \in H^*(M;\mathbb{Z}_2)$ [obstruction classes with coefficients in $\mathbb{Z}_2 = \{0,1\}$, where $1+1=0$. w_1 detects failure to give all frames one handedness, w_2 detects failure to fix the signs of the lifts consistently]. The manifold is orientable iff $w_1 = 0$, and an orientable manifold is spin iff also $w_2 = 0$. (Formally, w_2 is the $\mathbb{Z}_2$-valued Čech cocycle $\tilde h_{\alpha\beta}\tilde h_{\beta\gamma}\tilde h_{\gamma\alpha} = \pm\mathbf{1}$.)

Worked example 10.2: Which manifolds are spin

- S^n for $n \geq 3$: $H^1(S^n;\mathbb{Z}_2) = H^2(S^n;\mathbb{Z}_2) = 0$, so $w_1 = w_2 = 0$ automatically.
- S^2: now $H^2(S^2;\mathbb{Z}_2) = \mathbb{Z}_2$ is not zero, so we must compute. For an oriented surface w_2 is the Euler class mod 2, and

$$w_2(TS^2) \overset{(1)}{=} \chi(S^2) \bmod 2 \overset{(2)}{=} 2 \bmod 2 = 0.$$

(1) For an oriented rank-2 bundle, w_2 is the mod-2 reduction of the Euler class, whose integral over the surface is the Euler characteristic.

(2) $\chi(S^2) = 2$ (Gauss–Bonnet, Chapter T2).

Every oriented surface has even $\chi = 2 - 2g$, so every oriented surface is spin.

- $\mathbb{CP}^2$ is *not* spin: $w_2(\mathbb{CP}^2)$ is the generator of $H^2(\mathbb{CP}^2;\mathbb{Z}_2) = \mathbb{Z}_2$. (Its signature is 1, while Rokhlin's theorem forces the signature of a closed spin 4-manifold to be divisible by 16.)

So Nash's working example $M = S^{2m}$ is always allowed: "this is not too severe a requirement and is true for S^n".

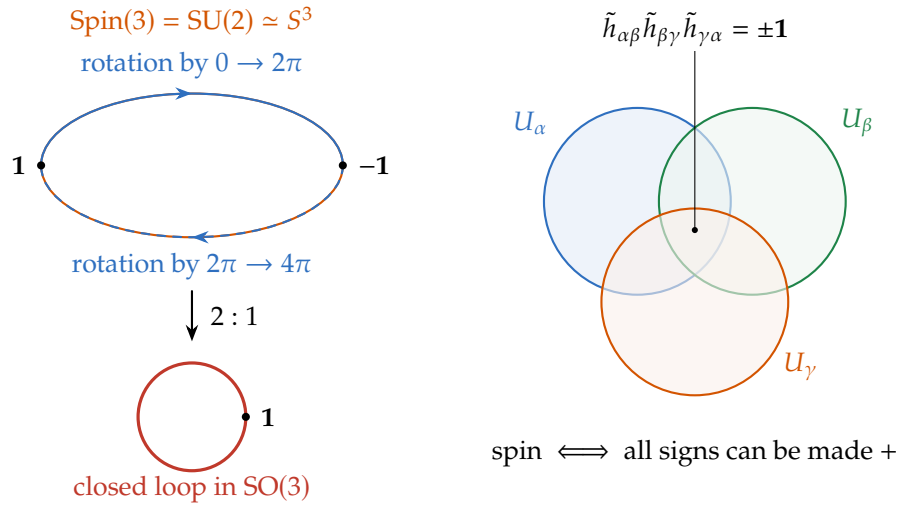


Figure 10.4: Left: a rotation through 2π is a closed loop in SO(3), but it lifts to a path from $\mathbf{1}$ to $-\mathbf{1}$ in Spin(3) = SU(2); only a rotation through 4π closes up. Spinors see this sign. Right: on a triple overlap of coordinate patches the lifted transition functions multiply to $\pm\mathbf{1}$. The manifold is spin when the signs can all be chosen to be +; the obstruction is the class w_2.

10.2.3 The chiral Dirac operator as a map between two spaces

Since $\not{D}_A$ anticommutes with γ_5, it swaps the two chiralities. If $E^{\pm}$ are the positive and negative spinor bundles and F the vector bundle on which the gauge field acts (the "colour" bundle), then

$$\not{D}_A : \Gamma(M, E^+ \otimes F) \longrightarrow \Gamma(M, E^- \otimes F), \tag{10.14}$$

[Nash p. 270, (10.2)]. We write D_A^+ for this map when we need to distinguish it from the full $\not{D}_A$. In a chiral basis the full Dirac operator is block off-diagonal,

$$\not{D}_A = \begin{pmatrix} 0 & D_A^- \\ D_A^+ & 0 \end{pmatrix}, \qquad D_A^+ : E^+ \to E^-, \quad D_A^- : E^- \to E^+, \qquad D_A^- = -(D_A^+)^\dagger. \tag{10.15}$$

The last relation is the anti-self-adjointness of $\not{D}_A$. Nash writes $\not{D}_A^*$ for the adjoint $(D_A^+)^\dagger : E^- \to E^+$.

Worked example 10.3: The blocks in two and four dimensions

Two dimensions. With $\gamma_1 = \sigma_1$, $\gamma_2 = \sigma_2$,

$$\not{D}_A = \sigma_1 D_1 + \sigma_2 D_2 \overset{(1)}{=} \begin{pmatrix} 0 & D_1 \\ D_1 & 0 \end{pmatrix} + \begin{pmatrix} 0 & -iD_2 \\ iD_2 & 0 \end{pmatrix}$$

$$\overset{(2)}{=} \begin{pmatrix} 0 & D_1 - iD_2 \\ D_1 + iD_2 & 0 \end{pmatrix}.$$

(1) $\sigma_1 = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, $\sigma_2 = \left(\begin{smallmatrix} 0 & -i \\ i & 0 \end{smallmatrix}\right)$.
(2) Add entrywise.

So $D_A^+ = D_1 + iD_2$. With $z = x_1 + ix_2$ and $\partial_{\bar z} = \frac{1}{2}(\partial_1 + i\partial_2)$, the free operator is $D_0^+ = 2\partial_{\bar z}$: *a positive-chirality zero mode is a holomorphic function*. This is the reason the two-dimensional Dirac operator is a Cauchy–Riemann operator, which is what Nash exploits for strings in §4.

Four dimensions. With (10.2),

$$\not{D}_A = \gamma_j D_j + \gamma_4 D_4 \overset{(1)}{=} \begin{pmatrix} 0 & i\sigma_j D_j + D_4 \\ -i\sigma_j D_j + D_4 & 0 \end{pmatrix},$$

(1) Insert the blocks of (10.2); σ_j acts on spin and D_j on space and colour, so they commute.

so $D_A^+ = D_4 - i\sigma_j D_j$ and $D_A^- = D_4 + i\sigma_j D_j$. Check the adjoint relation: since $D_\mu^\dagger = -D_\mu$ and $\sigma_j^\dagger = \sigma_j$, we have $(-i\sigma_j D_j)^\dagger = D_j^\dagger \sigma_j (+i) = -i\sigma_j D_j$, hence $(D_4 - i\sigma_j D_j)^\dagger = -D_4 - i\sigma_j D_j = -D_A^-$. These are the quaternionic "$\partial_4 \mp i\vec\sigma\cdot\vec\partial$" operators familiar from instanton calculus.

10.2.4 The index counts chiral zero modes

Let $n_+ = \dim \operatorname{Ker} D_A^+$ be the number of positive-chirality zero modes and $n_- = \dim \operatorname{Ker} D_A^- = \dim \operatorname{Ker}(D_A^+)^\dagger$ the number of negative-chirality ones. The index theorem for the twisted Dirac operator (Chapter 4) says

$$\begin{aligned} \operatorname{index} \not{D}_A &\overset{(1)}{=} n_+ - n_- \\ &\overset{(2)}{=} \int_M \hat A(M) \wedge \operatorname{ch}(F) \\ &\overset{(3)}{=} \int_{S^{2m}} \operatorname{ch}(F) \qquad \text{for } M = S^{2m} \\ &\overset{(4)}{=} k. \end{aligned} \tag{10.16}$$

(1) Definition: $\operatorname{index} D^+ = \dim \operatorname{Ker} D^+ - \dim \operatorname{Coker} D^+$, and $\operatorname{Coker} D^+ \cong \operatorname{Ker}(D^+)^\dagger$.
(2) The Atiyah–Singer formula for the Dirac operator twisted by F [Nash p. 270, (10.3); (4.122)]: only the top-degree ($2m$-form) part of the integrand contributes.
(3) $\hat A$ is a polynomial in the Pontrjagin classes, $\hat A = 1 - \frac{1}{24}p_1 + \frac{1}{5760}(7p_1^2 - 4p_2) + \cdots$, and every $p_i(S^n) = 0$, so $\hat A(S^{2m}) = 1$.
(4) Nash's name k for the resulting integer.

Why do the Pontrjagin classes of a sphere vanish? Because $TS^n \oplus \underline{\mathbb{R}}$ is the trivial bundle $S^n \times \mathbb{R}^{n+1}$ (split $\mathbb{R}^{n+1}$ at each point of S^n into the tangent plane and the normal line). Characteristic classes are unchanged by adding a trivial bundle, and all classes of a trivial bundle vanish.

Only the top form of $\operatorname{ch}(F) = \operatorname{tr}\exp(iF/2\pi)$ survives the integral, so

$$k = \int_{S^{2m}} \operatorname{ch}_m(F) = \frac{1}{m!}\left(\frac{i}{2\pi}\right)^m \int_{S^{2m}} \operatorname{tr} F^m. \tag{10.17}$$

For $m = 1$ this is the first Chern number $\int c_1(F)$, the monopole charge. For $m = 2$ and $G = \mathrm{SU}(N)$, where $c_1 = 0$, it is $\int \operatorname{ch}_2 = \int (\frac{1}{2}c_1^2 - c_2) = -\int c_2$, so $-k = \int c_2 = \frac{1}{8\pi^2}\int \operatorname{tr} F \wedge F$ is the

instanton number, as Nash says. On S^{2m} the class ch_m is (up to a nonzero integer factor) the only characteristic class there is, which is why Nash also calls k the "$c_m(F)$ which classifies bundles over S^{2m}".

Key idea: Chiral asymmetry

If $k \neq 0$, then $n_+ \neq n_-$: the massless chiral fermions have a forced excess of zero modes of one chirality. Since generic gauge fields on a sphere have $k \neq 0$ (only one bundle out of $\mathbb{Z}$ has $k = 0$), this is the typical situation. It is the first sign that chirality and topology are tied together.

Worked example 10.4: Zero modes in a magnetic field: Aharonov–Casher

Take $n = 2$, $G = \mathrm{U}(1)$, and a fermion of charge e in a magnetic field $B(x)$ on the plane, with total flux $\Phi = \int B\,\mathrm{d}^2x$. Choose the gauge $ea_1 = -\partial_2\phi$, $ea_2 = \partial_1\phi$ for a real function ϕ; then $eB = e(\partial_1 a_2 - \partial_2 a_1) = \nabla^2\phi$. In the mathematicians' convention $A_j = -iea_j$, and the positive-chirality equation is

$$\begin{aligned} 0 = D_A^+\psi_+ &\overset{(1)}{=} \big[(\partial_1 + i\partial_2) - ie(a_1 + ia_2)\big]\psi_+ \\ &\overset{(2)}{=} \big[2\partial_{\bar z} - i(-\partial_2\phi + i\partial_1\phi)\big]\psi_+ \\ &\overset{(3)}{=} \big[2\partial_{\bar z} + (\partial_1 + i\partial_2)\phi\big]\psi_+ \overset{(4)}{=} 2\big[\partial_{\bar z} + (\partial_{\bar z}\phi)\big]\psi_+. \end{aligned}$$

(1) $D_A^+ = D_1 + iD_2$ from Example 10.3, with $D_j = \partial_j - iea_j$.
(2) Insert the gauge choice.
(3) $-i\cdot(-\partial_2\phi) = i\partial_2\phi$ and $-i\cdot i\partial_1\phi = \partial_1\phi$.
(4) $\partial_1 + i\partial_2 = 2\partial_{\bar z}$.

The solutions are $\psi_+ = e^{-\phi}f(z)$ with f *holomorphic*: indeed $\partial_{\bar z}(e^{-\phi}f) = -(\partial_{\bar z}\phi)e^{-\phi}f$. Far away, Gauss's theorem on a large disc gives $2\pi r\,\phi'(r) \approx e\Phi$, so $\phi \approx \frac{e\Phi}{2\pi}\ln r$ and

$$|\psi_+| \approx r^{-e\Phi/2\pi}\,|f(z)|.$$

With $f = z^j$ this is square integrable ($\int |\psi_+|^2 r\,\mathrm{d}r < \infty$) iff $2j - e\Phi/\pi + 2 < 0$, i.e. $j < \frac{e\Phi}{2\pi} - 1$. The negative-chirality equation $D_A^-\psi_- = 0$ gives $\psi_- = e^{+\phi}g(\bar z)$, which grows and is never normalisable when $e\Phi > 0$. So for $e\Phi/2\pi = N$ a positive integer there are $N - 1$ modes on the plane and none of the other chirality. On the *sphere* (compactify the plane, the flux becoming a monopole of charge N) one extra mode becomes normalisable and the count is exactly

$$n_+ - n_- = N = \frac{e\Phi}{2\pi} = \int_{S^2}\frac{i}{2\pi}F = \int_{S^2}\mathrm{ch}_1(F),$$

in agreement with (10.16). (In the math convention $F = -ief$, so $\frac{i}{2\pi}F = \frac{e}{2\pi}f$.) All zero modes sit in the chirality that the field "prefers"; reversing the field reverses it.

Worked example 10.5: The instanton zero mode on S^4

For the SU(2) BPST instanton on S^4 (Chapter 8), the instanton number is $\int c_2 = 1$, so $k = \int \mathrm{ch}_2 = -1$ and $n_+ - n_- = -1$ for fermions in the fundamental (2-dimensional) representation. A vanishing theorem (the Weitzenböck formula $-\not{D}_A^2 = -D^\mu D_\mu + \frac{R}{4} - \frac{1}{2}\gamma^\mu\gamma^\nu F_{\mu\nu}$ with $R > 0$ on S^4 and F anti-self-dual or self-dual acting on only one chirality) kills the kernel of one chirality altogether. Hence there is exactly one zero mode, and it has

negative chirality with our conventions: this is 't Hooft's zero mode. For fermions in the adjoint (3-dimensional) representation, ch_2 of the adjoint bundle is 4 times as large, $k = -4$, and there are four zero modes (in supersymmetric Yang–Mills theory, where these are the gluino zero modes, two are generated by supersymmetry and two by superconformal transformations).

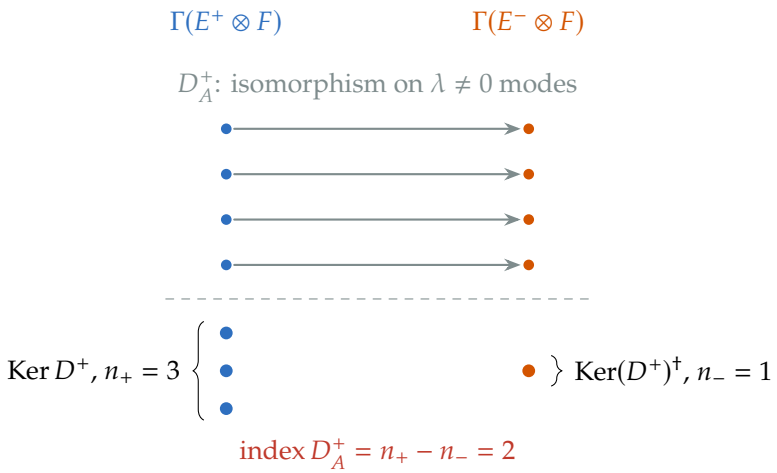


Figure 10.5: Why the index is robust. The non-zero modes of $-\not{D}_A^2$ come in pairs: if $(D^+)^\dagger D^+ u = \lambda u$ with $\lambda > 0$, then $v = D^+ u$ satisfies $D^+(D^+)^\dagger v = \lambda v$, and the grey arrows are bijections. Only the zero modes below the dashed line can be unpaired. When the gauge field is deformed, zero modes can appear or disappear only in $(+,-)$ pairs, so $n_+ - n_-$ never changes.

10.2.5 Integrating out the fermions: Grassmann integrals give determinants

The partition function of the theory is [Nash p. 271, (10.4)]

$$Z = \int \mathcal{D}A\, \mathcal{D}\psi\, \mathcal{D}\bar{\psi}\, \exp\!\left[- \|F\|^2 - \langle \psi, \not{D}_A \psi \rangle \right]. \tag{10.18}$$

The fermions appear quadratically, so their integral can be done exactly. We need the rules of Grassmann integration. Here they are, in finite dimensions.

Working definition 10.3 (Berezin integral)

Grassmann variables [the stand-ins for the fermion fields $\psi, \bar{\psi}$ in the path integral] anticommute: $\theta_i \theta_j = -\theta_j \theta_i$, so $\theta_i^2 = 0$ and every function is a polynomial of degree at most one in each variable. The integral [it picks out the coefficient of θ] is defined by

$$\int \mathrm{d}\theta\, 1 = 0, \qquad \int \mathrm{d}\theta\, \theta = 1,$$

with the $\mathrm{d}\theta_i$ also anticommuting among themselves and with the θ_j. *To compute:* expand the integrand, keep only the term containing every variable exactly once, and move the variables next to their differentials.

Worked example 10.6: Gaussian Grassmann integrals are determinants

One pair. Take independent variables $\theta, \bar\theta$ and a number a. First, $\int \mathrm{d}\bar\theta\,\mathrm{d}\theta\,\theta\bar\theta = \int \mathrm{d}\bar\theta\,\bar\theta = 1$ (inner integral first), so $\int \mathrm{d}\bar\theta\,\mathrm{d}\theta\,\bar\theta\theta = -1$. Then

$$\int \mathrm{d}\bar\theta\,\mathrm{d}\theta\, e^{-\bar\theta a\theta} \overset{(1)}{=} \int \mathrm{d}\bar\theta\,\mathrm{d}\theta\,(1 - a\bar\theta\theta) \overset{(2)}{=} -a\int \mathrm{d}\bar\theta\,\mathrm{d}\theta\,\bar\theta\theta \overset{(3)}{=} a.$$

(1) $(\bar\theta\theta)^2 = 0$, so the exponential series stops.
(2) $\int \mathrm{d}\bar\theta\,\mathrm{d}\theta\,1 = 0$.
(3) The value just computed.

Compare the bosonic $\int \mathrm{d}\bar z\,\mathrm{d}z\, e^{-\bar z a z} \propto a^{-1}$: fermions give the determinant upstairs, bosons downstairs.

Two pairs. With $\bar\theta M\theta = \sum_{ij}\bar\theta_i M_{ij}\theta_j$, only the second-order term of the exponential contains all four variables:

$$\begin{aligned}\int \prod_{i=1}^{2}\mathrm{d}\bar\theta_i\,\mathrm{d}\theta_i\; e^{-\bar\theta M\theta} &\overset{(1)}{=} \frac{1}{2}\int \prod_i \mathrm{d}\bar\theta_i\,\mathrm{d}\theta_i\,(\bar\theta M\theta)^2 \\ &\overset{(2)}{=} \int \prod_i \mathrm{d}\bar\theta_i\,\mathrm{d}\theta_i\,\left(M_{11}M_{22}\,\bar\theta_1\theta_1\bar\theta_2\theta_2 + M_{12}M_{21}\,\bar\theta_1\theta_2\bar\theta_2\theta_1\right) \\ &\overset{(3)}{=} M_{11}M_{22} - M_{12}M_{21} = \det M.\end{aligned}$$

(1) Terms of order 0 and 1 lack some variables and integrate to zero.
(2) Of the 16 terms in $(\bar\theta M\theta)^2$ only those using each of $\bar\theta_1, \bar\theta_2, \theta_1, \theta_2$ once survive; they come in two equal pairs (the order of the two bilinears is immaterial, as bilinears commute), cancelling the $\frac{1}{2}$.
(3) Moving θ_1 two places to the left (sign $+$) and then swapping $\theta_2\bar\theta_2$ (sign $-$) gives $\bar\theta_1\theta_2\bar\theta_2\theta_1 = -\bar\theta_1\theta_1\bar\theta_2\theta_2$. The bilinears $\bar\theta_i\theta_i$ and the pairs $\mathrm{d}\bar\theta_i\mathrm{d}\theta_i$ are even, so the integral factorises into two blocks $\int \mathrm{d}\bar\theta_i\mathrm{d}\theta_i\,\bar\theta_i\theta_i = -1$, whose product is $+1$.

In general $\int \prod \mathrm{d}\bar\theta_i\mathrm{d}\theta_i\, e^{-\bar\theta M\theta} = \det M$ for any $N\times N$ matrix M. The proof is the same bookkeeping: the terms that survive are indexed by permutations, with the sign of the permutation.

Applied formally to (10.18), with $M = \not{D}_A$ an operator on spinor fields, this says $\int \mathcal{D}\psi\mathcal{D}\bar\psi\, e^{-\langle\psi,\not{D}_A\psi\rangle} = \det\not{D}_A$. But Nash writes something slightly different [Nash p. 271, (10.5)]:

$$\int \mathcal{D}\psi\mathcal{D}\bar\psi\,\exp\left[-\langle\psi,\not{D}_A\psi\rangle\right] = \sqrt{\det(\not{D}_A^*\not{D}_A)}. \tag{10.19}$$

Where does the square root come from? From chirality. For a *chiral* fermion the operator in the action is $D_A^+ : E^+ \to E^-$. Its "determinant" is not a number: det of a map between two different spaces depends on the bases chosen in each (Section 10.6.3 makes this precise). What *is* a number is its modulus,

$$|\det D_A^+| = \sqrt{\det\big((D_A^+)^\dagger D_A^+\big)},$$

because $(D_A^+)^\dagger D_A^+$ maps E^+ to itself. That is Nash's (10.19) with $\not{D}_A^*\not{D}_A$ read as $(D_A^+)^\dagger D_A^+$. The next example checks, in finite dimensions, that a chiral fermion is "half" of a Dirac fermion.

Worked example 10.7: A Dirac fermion is two chiral fermions

Let L be a $k\times k$ matrix standing for D^+, and build the block operator $\mathcal{D} = \left(\begin{smallmatrix} 0 & -L^\dagger \\ L & 0 \end{smallmatrix}\right)$ modelling the full anti-self-adjoint $\not{D}$ of (10.15). Then

$$\det \mathcal{D} \overset{(1)}{=} (-1)^k \det(-L^\dagger) \det L \overset{(2)}{=} (-1)^k(-1)^k \overline{\det L} \det L \overset{(3)}{=} |\det L|^2 = \det(L^\dagger L).$$

(1) Moving the k lower rows above the k upper rows takes k^2 adjacent swaps, and $(-1)^{k^2} = (-1)^k$; the result is block diagonal.
(2) $\det(-X) = (-1)^k \det X$ and $\det L^\dagger = \overline{\det L}$.
(3) $(-1)^{2k} = 1$.

For $k = 1$: $\det\left(\begin{smallmatrix} 0 & -\bar{l} \\ l & 0 \end{smallmatrix}\right) = 0 - (-\bar{l})l = |l|^2$. So the Dirac determinant is $|\det D^+|^2$, and a single Weyl fermion contributes the square root of it, *times a phase*. The phase is the subtle part. It is exactly what (10.19) leaves out, and exactly where the anomaly will live.

With (10.19) the partition function becomes [Nash p. 271, (10.6)]

$$Z = \int \mathcal{D}A \sqrt{\det(\not{D}_A^* \not{D}_A)}\, \exp\left[-\|F\|^2\right]. \tag{10.20}$$

10.2.6 Gauge invariance of the integrand, and why it can fail

The next step in evaluating (10.20) would be to use gauge invariance. If the integrand were constant along each gauge orbit $\{A_g : g \in \mathcal{G}\}$, we could divide out the (infinite) volume of $\mathcal{G}$ and integrate only over the orbit space $\mathcal{A}/\mathcal{G}$, which is the classifying space $B\mathcal{G}$ when $\mathcal{G}$ acts freely (Chapter 6). This is the Faddeev–Popov procedure. The Yang–Mills term is manifestly invariant: $F_g = g^{-1}Fg$ and the trace is cyclic, so $\operatorname{tr}(F_g \wedge \star F_g) = \operatorname{tr}(F \wedge \star F)$. The trouble is the determinant. In general [Nash p. 271, (10.7)]

$$\det(\not{D}_{A_g}^* \not{D}_{A_g}) \neq \det(\not{D}_A^* \not{D}_A), \tag{10.21}$$

and when this happens we say an **anomaly** has arisen. It is surprising, because the operator itself transforms covariantly [Nash p. 271, (10.8)]:

$$\not{D}_{A_g} = g^{-1} \not{D}_A\, g. \tag{10.22}$$

Let us check this. For any spinor χ,

$$g^{-1}\not{D}_A(g\chi) \overset{(1)}{=} g^{-1}\gamma^\mu\big(\partial_\mu(g\chi) + A_\mu g\chi\big) \overset{(2)}{=} \gamma^\mu\big(g^{-1}(\partial_\mu g)\chi + \partial_\mu\chi + g^{-1}A_\mu g\,\chi\big) \overset{(3)}{=} \gamma^\mu\big(\partial_\mu + (A_g)_\mu\big)\chi = \not{D}_{A_g}\chi.$$

(1) Definition (10.5) of $\not{D}_A$.
(2) Leibniz rule; g acts on colour, γ^μ on spin, so they commute.
(3) $(A_g)_\mu = g^{-1}A_\mu g + g^{-1}\partial_\mu g$, from (10.4).

So $\not{D}_{A_g}$ and $\not{D}_A$ are conjugate, and conjugate matrices have equal determinants: $\det(g^{-1}Xg) = \det g^{-1} \det X \det g = \det X$. The catch is in the middle step, which uses $\det(XY) = \det X \det Y$.

In infinite dimensions a determinant must be *regularised*, and regularised determinants need not be multiplicative. There is a further catch for chiral fermions, explained in the pitfall below: in $D^+_{A_g} = g^{-1} D^+_A g$, the g on the right acts on E^+ and the g^{-1} on the left acts on E^-, two different spaces, and there is no reason for a regularisation to treat them the same way.

Watch out: The modulus is invariant; the phase is not

Read literally, (10.21) is too strong. The operator $\not{D}^*_{A_g}\not{D}_{A_g} = g^{-1}(\not{D}^*_A \not{D}_A)g$ is conjugate to $\not{D}^*_A\not{D}_A$ by a *unitary* map of a single space $E^+ \otimes F$ to itself, so the two have the same eigenvalues, the same zeta function, and the same zeta-regularised determinant. The modulus $|\det D^+_A|$ is gauge invariant. What is not gauge invariant is the *phase* of the chiral determinant, which $\sqrt{\det \not{D}^*\not{D}}$ does not see. That is why Nash (following Atiyah and Singer) soon replaces $\not{D}^*_{A_g}\not{D}_{A_g}$ by the "mixed" operator $T_g = \not{D}^*_A \not{D}_{A_g}$ (10.153): formally $\det T_g = \overline{\det D^+_A} \cdot \det D^+_{A_g}$, which does record the relative phase. Keep in mind throughout: "the determinant is not gauge invariant" means "its phase changes under gauge transformations in a way no local counterterm can undo".

10.2.7 Regularised determinants

Nash's footnote 1 mentions two ways to give an infinite determinant a meaning.

Operators close to the identity. If $X = \mathbf{1} + K$ with K of **trace class**, meaning that the sum of the singular values of K is finite, $\operatorname{tr}\sqrt{K^\dagger K} < \infty$, then the product $\prod_n (1+\kappa_n)$ over the eigenvalues κ_n of K converges absolutely. The determinant is then defined without any regularisation, and it is multiplicative: $\det(XY) = \det X \, \det Y$.

Worked example 10.8: A convergent infinite determinant

Let K be diagonal with eigenvalues $1/n^2$, $n = 1, 2, \dots$. Then $\sum_n 1/n^2 = \pi^2/6 < \infty$, so K is trace class, and

$$\det(\mathbf{1}+K) \overset{(1)}{=} \prod_{n=1}^{\infty}\Big(1+\frac{1}{n^2}\Big) \overset{(2)}{=} \frac{\sinh \pi}{\pi} \approx 3.676.$$

(1) The determinant of a diagonal operator is the product of its entries.
(2) Euler's product $\sinh(\pi x) = \pi x \prod_n (1 + x^2/n^2)$ at $x = 1$.

The operators in field theory are not of this kind. $\not{D}^*_A\not{D}_A$ is a second-order elliptic operator whose eigenvalues grow like $\lambda_j \sim j^{2/n}$ (Weyl's law), so $\prod \lambda_j$ diverges wildly. The standard cure is the zeta function (Chapter 2; Nash p. 261).

Working definition 10.4 (zeta-regularised determinant)

Let T be an elliptic operator [an operator such as $\not{D}^*_A\not{D}_A$ on the compact M, whose eigenvalues are discrete and grow like a power of j, so their product diverges] with eigenvalues λ_j, none zero, with positive real parts (or at least avoiding a ray from the origin, Section 10.6.1). Define

$$\zeta_T(s) := \sum_j \lambda_j^{-s}.$$

The sum converges for $\operatorname{Re} s > n/2$ (for T of second order in n dimensions), and continues

[there is exactly one function, analytic except at isolated poles, that equals the sum wherever the sum converges] to a function that is analytic at $s = 0$. Since formally $-\zeta_T'(0) = \sum_j \ln \lambda_j = \ln \prod_j \lambda_j$, define

$$\det T := \exp\bigl[-\zeta_T'(0)\bigr].$$

To compute: find the spectrum, express ζ_T in terms of known functions (Riemann, Hurwitz, Epstein zeta functions [$\sum_{n\geq 1} n^{-s}$, $\sum_{n\geq 0}(n+a)^{-s}$, and a sum of $|m|^{-2s}$ over the points $m \neq 0$ of a lattice]), and differentiate at $s = 0$.

Worked example 10.9: The Laplacian on a circle, and a multiplicative anomaly

On the circle $\theta \sim \theta + 2\pi$ take $T = -\mathrm{d}^2/\mathrm{d}\theta^2$, with eigenfunctions $e^{in\theta}$ and eigenvalues n^2. Drop the zero mode $n = 0$ (write det′). Each n^2 with $n \neq 0$ occurs twice ($\pm n$), so

$$\zeta_T(s) \overset{(1)}{=} 2\sum_{n=1}^{\infty} n^{-2s} \overset{(2)}{=} 2\zeta_R(2s),$$

$$\ln \det' T = -\zeta_T'(0) \overset{(3)}{=} -4\zeta_R'(0) \overset{(4)}{=} 2\ln(2\pi), \qquad \text{so} \quad \det' T = 4\pi^2.$$

(1) Count the degenerate pairs.
(2) Riemann's $\zeta_R(s) = \sum_{n\geq 1} n^{-s}$.
(3) Chain rule: $\frac{\mathrm{d}}{\mathrm{d}s} 2\zeta_R(2s) = 4\zeta_R'(2s)$.
(4) The classical value $\zeta_R'(0) = -\frac{1}{2}\ln(2\pi)$.

A divergent product $\prod_{n\neq 0} n^2$ has been given the finite value $4\pi^2$. Now rescale T by a constant $c > 0$. Naively $\det(cT) = c^{\infty} \det T$. The zeta function says:

$$\zeta_{cT}(s) \overset{(1)}{=} c^{-s}\zeta_T(s),$$

$$-\zeta_{cT}'(0) \overset{(2)}{=} \ln c\ \zeta_T(0) - \zeta_T'(0), \qquad \text{so} \quad \det'(cT) = c^{\zeta_T(0)} \det' T \overset{(3)}{=} c^{-1} \det' T.$$

(1) $(c\lambda)^{-s} = c^{-s}\lambda^{-s}$.
(2) Differentiate the product at $s = 0$: $\frac{\mathrm{d}}{\mathrm{d}s} c^{-s} = -\ln c\ c^{-s}$.
(3) $\zeta_T(0) = 2\zeta_R(0) = 2 \cdot (-\frac{1}{2}) = -1$.

But $cT = (c\mathbf{1})T$, and "$\det(c\mathbf{1})$" on an infinite-dimensional space is meaningless or, if regularised by its own zeta function $\sum_j c^{-s}$, not equal to c^{-1}. So $\det(XY) \neq \det X \det Y$ for zeta determinants in general. This **multiplicative anomaly** is the mechanism by which "conjugate operators have the same determinant" can fail once the two sides of the conjugation act on different spaces.

Physical picture: What regularisation does to a symmetry

Every regularisation of $\det D_A^+$ has to cut off high-frequency modes. "High frequency" must be measured by some operator, and for a chiral fermion there is no operator that is simultaneously gauge covariant and maps E^+ to E^+ built from D_A^+ alone: $(D_A^+)^\dagger D_A^+$ and $D_A^+(D_A^+)^\dagger$ live on different spaces. A cutoff that respects gauge invariance on one side breaks it on the other. The anomaly is the finite remainder of this mismatch that survives as the cutoff is removed. We will see this explicitly in the Fujikawa calculation of Section 10.4.1.

Reading Nash (pp. 270–271, (10.1)–(10.8))

Nash's Γ_μ and A_μ are the Levi-Civita (spin) and Yang–Mills connections. His requirement "$w_1(TM) = w_2(TM) = 0$" is the spin condition of Definition 10.2. In (10.3) he uses the index formula "4.122" of his Chapter 4 and the vanishing of the Pontrjagin classes of spheres "(p. 85)". His "$\mathcal{A}/\mathcal{G} = B\mathcal{G}$" is the identification of the orbit space with the classifying space of the gauge group (valid where $\mathcal{G}$ acts freely, e.g. after restricting to based gauge transformations). His footnote on regularisation is Section 10.2.7. Note that in (10.5) the square root of $\det(\not{D}_A^*\not{D}_A)$ is the modulus of the chiral determinant; its phase is where the anomaly lives (see the pitfall in Section 10.2.6).

Recap

- A chiral gauge theory couples only $\psi_+ = P_+\psi$ to A; the action is $\|F\|^2 + \int \bar\psi P_- \not{D}_A P_+ \psi$, with $\|F\|^2 \geq 0$ for anti-Hermitian F.
- M must be spin ($w_1 = w_2 = 0$); all spheres qualify, $\mathbb{CP}^2$ does not.
- $D_A^+ : \Gamma(E^+ \otimes F) \to \Gamma(E^- \otimes F)$; in 2D it is $2\partial_{\bar z} + \ldots$, in 4D $D_4 - i\vec\sigma\cdot\vec D$.
- $\operatorname{index} D_A^+ = n_+ - n_- = \int \hat A\, \operatorname{ch}(F) = \int_{S^{2m}} \operatorname{ch}_m(F) = k$; e.g. N zero modes for a charge-N monopole on S^2, one for the BPST instanton.
- Grassmann Gaussians give determinants; a chiral fermion gives $\det D_A^+$, whose modulus is $\sqrt{\det(\not{D}_A^*\not{D}_A)}$.
- $\not{D}_{A_g} = g^{-1}\not{D}_A g$ formally implies invariance, but regularised determinants are not multiplicative ($\det'(cT) = c^{\zeta_T(0)}\det' T$). The phase of the chiral determinant can change: that is the anomaly.

10.3 Anomalies and Yang–Mills theories II: the anomalous Ward identity, computed

In Section 10.2 we found the object whose gauge invariance is in question: the fermion determinant. Nash now does two things in quick succession [Nash pp. 272–273]. He differentiates the determinant along a gauge orbit and shows that the derivative is the vacuum expectation value of the divergence of a current. Then he says, in one sentence each, that "a direct perturbative way" to see that this is nonzero is the triangle graph, and that Fujikawa found the same thing in the path-integral measure. Those two sentences stand for a large amount of physics. This section and the next one do the calculations.

The question this section answers is: *how do we actually see, in a Feynman-graph calculation, that a classically conserved current is not conserved?* We do it first in two dimensions, where everything can be computed in a few lines and in three different ways, and then in four, where the triangle graph appears.

10.3.1 Varying the determinant along a gauge orbit

Take a one-parameter family of gauge transformations

$$g(t) = \exp(tf), \qquad f \in \operatorname{Lie}\mathcal{G}, \qquad f(x) = T^a f^a(x),$$

where T^a is a basis of anti-Hermitian matrices for the Lie algebra $\mathfrak{g}$ of G in the representation carried by the fermions, and the $f^a(x)$ are real functions on M. (Nash writes t^a for the generators; we use T^a to avoid a clash with the parameter t.) We want $\frac{\mathrm{d}}{\mathrm{d}t}$ of the fermion integral at $A_{g(t)}$, at $t = 0$.

Step 1: how A moves. To first order in t,

$$\begin{aligned} A_{g(t)} &\overset{(1)}{=} (1 - tf)A(1 + tf) + (1 - tf)\,\mathrm{d}(1 + tf) + O(t^2) \\ &\overset{(2)}{=} A + t\big(\mathrm{d}f + Af - fA\big) + O(t^2) \\ &\overset{(3)}{=} A + t\,D_A f + O(t^2), \qquad D_A f \coloneqq \mathrm{d}f + [A, f]. \end{aligned} \tag{10.23}$$

(1) The gauge transformation (10.4) with $g = 1 + tf + O(t^2)$ and $g^{-1} = 1 - tf + O(t^2)$.
(2) Multiply out and drop t^2.
(3) $D_A f$ is the covariant derivative of f in the adjoint representation; in components $(D_\mu f)^a = \partial_\mu f^a + f^{abc} A^b_\mu f^c$ when $[T^b, T^c] = f^{bca} T^a$. Nash writes this operator ∇^{ab}_μ.

So an infinitesimal gauge transformation moves A in the direction $D_A f$: this is the tangent vector to the gauge orbit through A.

Step 2: how the action moves. Only the fermion term depends on A, through $\not{D}_A P_+ = \gamma^\mu(\partial_\mu + A_\mu)P_+$. Define the **chiral (gauge) current**

$$J^{a\mu} \coloneqq \bar\psi\,\gamma^\mu T^a P_+ \psi = \tfrac{1}{2}\,\bar\psi\,\gamma^\mu(1 + \gamma_5)T^a\psi, \tag{10.24}$$

which is Nash's $j_5^{a\mu}$ [Nash p. 272, (10.9)]. Then, holding ψ and $\bar\psi$ fixed,

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} S_F[A_{g(t)}]\Big|_{t=0} &\overset{(1)}{=} \int_M \bar\psi\,\gamma^\mu (D_\mu f) P_+ \psi \\ &\overset{(2)}{=} \int_M (D_\mu f)^a\, J^{a\mu} \\ &\overset{(3)}{=} -\int_M f^a\,(D_\mu J^\mu)^a. \end{aligned} \tag{10.25}$$

(1) Only A_μ changes, by $t\,D_\mu f$ (10.23).
(2) Expand $D_\mu f = (D_\mu f)^a T^a$ and use the definition (10.24).
(3) Integrate by parts. The covariant derivative obeys a Leibniz rule for the invariant pairing $(D_\mu f)^a J^{a\mu} = \partial_\mu(f^a J^{a\mu}) - f^a (D_\mu J^\mu)^a$, and the total derivative integrates to zero on the closed manifold M by Stokes' theorem; this is the "Stokes' theorem" in Nash's sentence.

Step 3: how the fermion integral moves, if the measure does not. If the measure $\mathcal{D}\psi\mathcal{D}\bar\psi$ is held fixed, differentiating under the integral sign gives

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\int \mathcal{D}\psi\mathcal{D}\bar\psi\, e^{-S_F[A_{g(t)}]}\Big|_{t=0} &\overset{(1)}{=} -\int \mathcal{D}\psi\mathcal{D}\bar\psi\, e^{-S_F[A]}\,\frac{\mathrm{d}S_F}{\mathrm{d}t}\Big|_{t=0} \\ &\overset{(2)}{=} \int \mathcal{D}\psi\mathcal{D}\bar\psi\, e^{-S_F[A]} \int_M f^a(x)\,(D_\mu J^\mu)^a(x). \end{aligned} \tag{10.26}$$

(1) Chain rule on the exponential.
(2) Insert (10.25); the two minus signs cancel.

This is Nash's unnumbered equation on p. 272 (with his $\sqrt{\det}$ in place of the fermion integral). Dividing by the fermion integral itself (Nash's "standard normalisation factor", which removes vacuum graphs) turns the right side into an expectation value [Nash p. 272, (10.10)]:

$$\frac{\mathrm{d}}{\mathrm{d}t}\ln Z_F[A_{g(t)}]\Big|_{t=0} = \Big\langle \int_M f^a\,(D_\mu J^\mu)^a \Big\rangle_A, \qquad Z_F[A] := \int \mathcal{D}\psi\mathcal{D}\bar\psi\, e^{-S_F[A]}. \tag{10.27}$$

Key idea: Anomaly = non-conservation of the gauge current

The left side of (10.27) is the change of $\ln Z_F$ along the gauge orbit, in the direction f. The right side is the smeared divergence of the current the gauge field couples to. So

$$\ln Z_F \text{ gauge invariant} \iff \langle D_\mu J^{a\mu}(x)\rangle_A = 0 \text{ for all } x, a \qquad \text{[Nash (10.11)].}$$

The *classical* equations of motion give $D_\mu J^{a\mu} = 0$ (the same computation as (10.7)). The quantum statement fails exactly when an anomaly is present.

For $G = \mathrm{U}(1)$ the generator T is a number ($T = -ie$ for charge e), D_μ reduces to ∂_μ, and the statement is about the ordinary divergence of a U(1) current. Nash calls it the "singlet axial current" $\frac{1}{2}\bar\psi\gamma^\mu(1+\gamma_5)\psi$. It is worth untangling the names. With $j^\mu = \bar\psi\gamma^\mu\psi$ and $j_5^\mu = \bar\psi\gamma^\mu\gamma_5\psi$,

$$j_\pm^\mu := \bar\psi\gamma^\mu P_\pm\psi = \tfrac{1}{2}\big(j^\mu \pm j_5^\mu\big), \qquad j^\mu = j_+^\mu + j_-^\mu, \qquad j_5^\mu = j_+^\mu - j_-^\mu. \tag{10.28}$$

In a *vector-like* theory (QED) the photon couples to j^μ and the axial current j_5^μ is a global current; in a *chiral* theory the gauge field couples to j_+^μ alone. The anomaly calculations below are done for QED, where they are cleanest, and the translation to the chiral theory is made in Sections 10.4.1 and 10.5.

Watch out: Why the measure matters

Equation (10.26) assumed the measure does not change. But there is a second way to compute the left side: substitute $\psi \to g\psi$, $\bar\psi \to \bar\psi g^{-1}$ in the integral for $Z_F[A_g]$. By (10.22) this turns the action back into $S_F[A]$, so if the measure were invariant we would get $Z_F[A_g] = Z_F[A]$ exactly, and hence $\langle D_\mu J^\mu\rangle = 0$. Either the current is conserved or the measure is not invariant. Fujikawa's insight (Section 10.4.1) is that the second alternative is what happens, and that the Jacobian of the change of variables *is* the anomaly.

10.3.2 Tools for one-loop integrals

The explicit anomaly calculations below are one-loop Feynman integrals. They use five standard tools, and we collect them here once, with their derivations, so that every later step can point back to a formula on this page. The reader who knows them can skip to Section 10.3.3.

Notation. In this subsection and the next four we work in Minkowski space of dimension d, with metric $\eta = \mathrm{diag}(+1,-1,\dots,-1)$. For a vector v^μ we write $v^2 = \eta_{\mu\nu}v^\mu v^\nu$ and $v\cdot w = \eta_{\mu\nu}v^\mu w^\nu$. The gamma matrices obey the Clifford relation

$$\gamma^\mu\gamma^\nu + \gamma^\nu\gamma^\mu = 2\eta^{\mu\nu}\mathbf{1}, \tag{10.29}$$

and the **slash** of a vector is $\not{v} := \gamma^\mu v_\mu$. We write N_d for the size of the gamma matrices: $N_2 = 2$, $N_4 = 4$.

Tool 1: gamma-matrix traces

Everything follows from (10.29) and cyclicity of the trace, $\mathrm{tr}(XY) = \mathrm{tr}(YX)$.

(a) The slash of a vector squares to its length.

$$\not{v}\not{v} \overset{(1)}{=} v_\mu v_\nu \gamma^\mu \gamma^\nu \overset{(2)}{=} \tfrac{1}{2} v_\mu v_\nu \big(\gamma^\mu \gamma^\nu + \gamma^\nu \gamma^\mu\big) \overset{(3)}{=} v_\mu v_\nu \eta^{\mu\nu} \mathbf{1} = v^2\, \mathbf{1}. \tag{10.30}$$

(1) Definition of the slash.
(2) $v_\mu v_\nu$ is symmetric in μ, ν, so only the symmetric part of $\gamma^\mu \gamma^\nu$ survives.
(3) The Clifford relation (10.29).

So $\not{v}$ is invertible when $v^2 \neq 0$, with $\not{v}^{-1} = \not{v}/v^2$. This is why the massless fermion propagator can be written either as $1/\not{p}$ or as $\not{p}/p^2$.

(b) The trace of two.

$$\mathrm{tr}(\gamma^\mu \gamma^\nu) \overset{(1)}{=} \tfrac{1}{2}\,\mathrm{tr}\big(\gamma^\mu \gamma^\nu + \gamma^\nu \gamma^\mu\big) \overset{(2)}{=} \tfrac{1}{2}\,\mathrm{tr}\big(2\eta^{\mu\nu}\mathbf{1}\big) = N_d\, \eta^{\mu\nu}. \tag{10.31}$$

(1) Cyclicity: $\mathrm{tr}(\gamma^\mu \gamma^\nu) = \mathrm{tr}(\gamma^\nu \gamma^\mu)$, so the trace equals the average of the two.
(2) Clifford relation; $\mathrm{tr}\,\mathbf{1} = N_d$.

(c) The trace of four. Move γ^μ from the far left to the far right, anticommuting it past one matrix at a time; each step produces an η and a sign:

$$\begin{aligned}
\mathrm{tr}(\gamma^\mu \gamma^\alpha \gamma^\nu \gamma^\beta) &\overset{(1)}{=} 2\eta^{\mu\alpha}\,\mathrm{tr}(\gamma^\nu \gamma^\beta) - \mathrm{tr}(\gamma^\alpha \gamma^\mu \gamma^\nu \gamma^\beta) \\
&\overset{(2)}{=} 2\eta^{\mu\alpha}\,\mathrm{tr}(\gamma^\nu \gamma^\beta) - 2\eta^{\mu\nu}\,\mathrm{tr}(\gamma^\alpha \gamma^\beta) + \mathrm{tr}(\gamma^\alpha \gamma^\nu \gamma^\mu \gamma^\beta) \\
&\overset{(3)}{=} 2\eta^{\mu\alpha}\,\mathrm{tr}(\gamma^\nu \gamma^\beta) - 2\eta^{\mu\nu}\,\mathrm{tr}(\gamma^\alpha \gamma^\beta) + 2\eta^{\mu\beta}\,\mathrm{tr}(\gamma^\alpha \gamma^\nu) - \mathrm{tr}(\gamma^\alpha \gamma^\nu \gamma^\beta \gamma^\mu) \\
&\overset{(4)}{=} \eta^{\mu\alpha}\,\mathrm{tr}(\gamma^\nu \gamma^\beta) - \eta^{\mu\nu}\,\mathrm{tr}(\gamma^\alpha \gamma^\beta) + \eta^{\mu\beta}\,\mathrm{tr}(\gamma^\alpha \gamma^\nu) \\
&\overset{(5)}{=} N_d\big(\eta^{\mu\alpha}\eta^{\nu\beta} - \eta^{\mu\nu}\eta^{\alpha\beta} + \eta^{\mu\beta}\eta^{\alpha\nu}\big).
\end{aligned} \tag{10.32}$$

(1) $\gamma^\mu \gamma^\alpha = 2\eta^{\mu\alpha} - \gamma^\alpha \gamma^\mu$.
(2) $\gamma^\mu \gamma^\nu = 2\eta^{\mu\nu} - \gamma^\nu \gamma^\mu$ inside the last trace.
(3) $\gamma^\mu \gamma^\beta = 2\eta^{\mu\beta} - \gamma^\beta \gamma^\mu$ inside the last trace.
(4) By cyclicity the last trace equals the left-hand side; move it over and divide by 2.
(5) The trace of two, (10.31).

Contracting with vectors, $\mathrm{tr}(\not{a}\not{b}\not{c}\not{d}) = N_d\big[(a\cdot b)(c\cdot d) - (a\cdot c)(b\cdot d) + (a\cdot d)(b\cdot c)\big]$.

(d) Two contraction identities. With $\eta^\mu{}_\mu = d$,

$$\gamma^\rho \gamma^\mu \gamma_\rho \overset{(1)}{=} \big(2\eta^{\rho\mu} - \gamma^\mu \gamma^\rho\big)\gamma_\rho \overset{(2)}{=} 2\gamma^\mu - d\,\gamma^\mu = (2-d)\gamma^\mu, \tag{10.33}$$

$$\not{p}\gamma^\lambda \not{p} \overset{(3)}{=} \big(2p^\lambda - \gamma^\lambda \not{p}\big)\not{p} \overset{(4)}{=} 2p^\lambda \not{p} - p^2 \gamma^\lambda. \tag{10.34}$$

(1) Clifford relation for $\gamma^\rho \gamma^\mu$.
(2) $\eta^{\rho\mu}\gamma_\rho = \gamma^\mu$, and $\gamma^\rho \gamma_\rho = \tfrac{1}{2}\eta_{\rho\sigma}\{\gamma^\rho, \gamma^\sigma\} = \eta_{\rho\sigma}\eta^{\rho\sigma} = d$.
(3) $\not{p}\gamma^\lambda = p_\rho \gamma^\rho \gamma^\lambda = p_\rho(2\eta^{\rho\lambda} - \gamma^\lambda \gamma^\rho) = 2p^\lambda - \gamma^\lambda \not{p}$.
(4) $\not{p}\not{p} = p^2$ by (10.30).

(e) Odd traces vanish, and traces with γ^5. In even dimension there is a matrix γ^5 with $(\gamma^5)^2 = \mathbf{1}$ that anticommutes with every γ^μ (we write it down in each dimension below). For an odd

number $2r+1$ of gamma matrices,

$$\operatorname{tr}(\gamma^{\mu_1}\cdots\gamma^{\mu_{2r+1}}) \overset{(1)}{=} \operatorname{tr}(\gamma^5\gamma^5\gamma^{\mu_1}\cdots\gamma^{\mu_{2r+1}}) \overset{(2)}{=} (-1)^{2r+1}\operatorname{tr}(\gamma^5\gamma^{\mu_1}\cdots\gamma^{\mu_{2r+1}}\gamma^5)$$
$$\overset{(3)}{=} -\operatorname{tr}(\gamma^5\gamma^5\gamma^{\mu_1}\cdots\gamma^{\mu_{2r+1}}) = -\operatorname{tr}(\gamma^{\mu_1}\cdots\gamma^{\mu_{2r+1}}), \tag{10.35}$$

(1) Insert $\mathbf{1} = (\gamma^5)^2$.
(2) Move the second γ^5 to the far right, past $2r+1$ matrices, each giving a sign.
(3) Cyclicity moves the last γ^5 back to the front.

so the trace is zero. A similar trick shows $\operatorname{tr}\gamma^5 = \operatorname{tr}(\gamma^0\gamma^0\gamma^5) = -\operatorname{tr}(\gamma^0\gamma^5\gamma^0) = -\operatorname{tr}\gamma^5$, so $\operatorname{tr}\gamma^5 = 0$ (use $(\gamma^0)^2 = \mathbf{1}$, anticommute, then cycle); inserting instead $\gamma^\rho\gamma^\rho/\eta^{\rho\rho}$ with ρ different from α,β shows in the same way that $\operatorname{tr}(\gamma^5\gamma^\alpha\gamma^\beta) = 0$ in four dimensions. In four dimensions, with $\gamma^5 := i\gamma^0\gamma^1\gamma^2\gamma^3$ and $\epsilon^{0123} := +1$,

$$\operatorname{tr}(\gamma^5) = \operatorname{tr}(\gamma^5\gamma^\alpha\gamma^\beta) = 0, \qquad \operatorname{tr}(\gamma^5\gamma^\alpha\gamma^\beta\gamma^\gamma\gamma^\delta) = -4i\,\epsilon^{\alpha\beta\gamma\delta}. \tag{10.36}$$

Why: $\gamma^5\gamma^\alpha\gamma^\beta\gamma^\gamma\gamma^\delta$ can have nonzero trace only if the product contains each of $\gamma^0,\gamma^1,\gamma^2,\gamma^3$ an even number of times in total, so $\alpha\beta\gamma\delta$ must be a permutation of 0123; the trace is then antisymmetric, and one value fixes it:

$$\operatorname{tr}(\gamma^5\gamma^0\gamma^1\gamma^2\gamma^3) \overset{(1)}{=} \operatorname{tr}\big(\gamma^5\cdot(-i\gamma^5)\big) \overset{(2)}{=} -i\operatorname{tr}\mathbf{1} = -4i.$$

(1) $\gamma^0\gamma^1\gamma^2\gamma^3 = -i\gamma^5$, from the definition of γ^5 and $1/i = -i$.
(2) $(\gamma^5)^2 = \mathbf{1}$; this holds because $(\gamma^0\gamma^1\gamma^2\gamma^3)^2 = -\mathbf{1}$ (six anticommutations and $(\gamma^0)^2 = -(\gamma^j)^2 = \mathbf{1}$), so $(\gamma^5)^2 = i^2\cdot(-\mathbf{1}) = \mathbf{1}$.

Two consequences we use constantly: (i) $\operatorname{tr}(\gamma^5 \not{a}\not{b}\not{c}\not{d}) = -4i\,\epsilon^{\alpha\beta\gamma\delta}a_\alpha b_\beta c_\gamma d_\delta$ is *totally antisymmetric* in the four slots, so it vanishes if two slots hold the same vector; (ii) the four slots may be cyclically rotated only with a sign, because a cyclic shift of four objects is an odd permutation.

Tool 2: Feynman parameters

Products of propagator denominators are combined into a single power. The cleanest derivation uses the **Schwinger parameter** representation $\frac{1}{A} = \int_0^\infty \mathrm{d}s\, e^{-sA}$ (valid for $\operatorname{Re}A > 0$; the final identities are algebraic and continue to hold whenever the parameter integrals make sense, e.g. for denominators carrying Feynman's $+i0$). For two factors:

$$\frac{1}{AB} \overset{(1)}{=} \int_0^\infty \mathrm{d}s_1 \int_0^\infty \mathrm{d}s_2\, e^{-s_1A - s_2B}$$
$$\overset{(2)}{=} \int_0^\infty \mathrm{d}s\; s\int_0^1 \mathrm{d}x\; e^{-s[xA+(1-x)B]} \overset{(3)}{=} \int_0^1 \frac{\mathrm{d}x}{\big[xA+(1-x)B\big]^2}. \tag{10.37}$$

(1) One Schwinger integral for each factor.
(2) Change variables $s_1 = sx$, $s_2 = s(1-x)$ with $s\in[0,\infty)$, $x\in[0,1]$; the Jacobian is $\big|\partial(s_1,s_2)/\partial(s,x)\big| = \big|\det\left(\begin{smallmatrix} x & s \\ 1-x & -s\end{smallmatrix}\right)\big| = s$.
(3) $\int_0^\infty \mathrm{d}s\, s\, e^{-sX} = 1/X^2$ (integrate by parts once).

For three factors the same change of variables, $s_i = sx_i$ with $x_1+x_2+x_3 = 1$, has Jacobian s^2, and $\int_0^\infty \mathrm{d}s\, s^2 e^{-sX} = 2/X^3$, so

$$\frac{1}{ABC} = 2\int_0^1\mathrm{d}x\int_0^1\mathrm{d}y\int_0^1\mathrm{d}z\;\frac{\delta(1-x-y-z)}{\big[xA+yB+zC\big]^3}, \qquad \int \mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z\;\delta(1-x-y-z) = \tfrac{1}{2}. \tag{10.38}$$

(The last integral is the area of the triangle $x,y\ge 0$, $x+y\le 1$.) A quick test of (10.37): at $A = B$ it reads $1/A^2 = \int_0^1 \mathrm{d}x/A^2$, which is true.

Tool 3: Wick rotation

A Minkowski loop integral $\int \mathrm{d}^d\ell\, f(\ell^2)$ has its poles at $\ell^0 = \pm(\sqrt{\boldsymbol{\ell}^2+\Delta} - i0)$: just below the positive real axis and just above the negative one (Figure 10.6). The contour along the real ℓ^0 axis can therefore be rotated anticlockwise by 90° onto the imaginary axis without crossing a pole, and the arcs at infinity contribute nothing when f falls off fast enough. Set

$$\ell^0 = i\ell_E^0, \qquad \boldsymbol{\ell} = \boldsymbol{\ell}_E, \qquad \ell^2 = -\ell_E^2, \qquad \mathrm{d}^d\ell = i\,\mathrm{d}^d\ell_E, \tag{10.39}$$

where $\ell_E^2 = (\ell_E^0)^2 + \boldsymbol{\ell}^2$ is a Euclidean length.

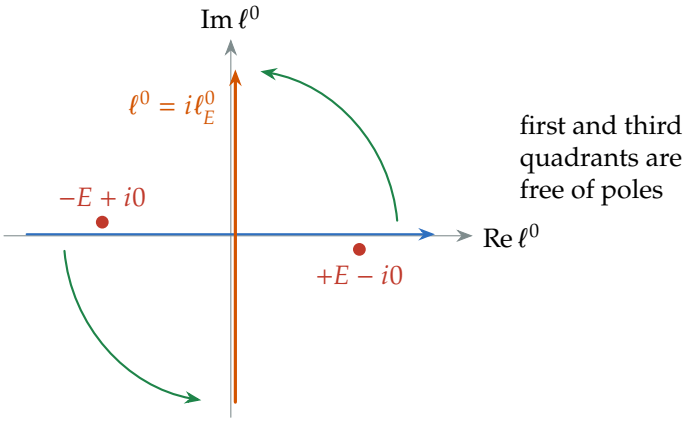


Figure 10.6: Wick rotation. The poles of a Feynman propagator (red, with $E = \sqrt{\boldsymbol{\ell}^2+\Delta}$) sit in the second and fourth quadrants. Rotating the integration contour anticlockwise (green arrows) from the real axis (blue) to the imaginary axis (orange) sweeps only the empty first and third quadrants, so the integral does not change. On the imaginary axis the Minkowski ℓ^2 becomes minus a Euclidean length.

Tool 4: the master integrals

With Wick rotation, a Schwinger parameter and a Gaussian integral we can do every integral we need. The Gaussian integral in d Euclidean dimensions is $\int \mathrm{d}^d\ell_E\, e^{-s\ell_E^2} = (\pi/s)^{d/2}$, the product of d one-dimensional ones. Then, for $\Delta > 0$,

$$\begin{aligned}
\int \frac{\mathrm{d}^d\ell}{(2\pi)^d}\frac{1}{(\ell^2-\Delta)^n} &\overset{(1)}{=} i\int \frac{\mathrm{d}^d\ell_E}{(2\pi)^d}\frac{(-1)^n}{(\ell_E^2+\Delta)^n} \\
&\overset{(2)}{=} \frac{i(-1)^n}{(2\pi)^d\,\Gamma(n)}\int_0^\infty \mathrm{d}s\; s^{n-1}e^{-s\Delta}\int \mathrm{d}^d\ell_E\, e^{-s\ell_E^2} \\
&\overset{(3)}{=} \frac{i(-1)^n\pi^{d/2}}{(2\pi)^d\,\Gamma(n)}\int_0^\infty \mathrm{d}s\; s^{n-1-d/2}e^{-s\Delta} \\
&\overset{(4)}{=} \frac{(-1)^n\, i}{(4\pi)^{d/2}}\,\frac{\Gamma(n-\frac{d}{2})}{\Gamma(n)}\,\Delta^{d/2-n}.
\end{aligned} \tag{10.40}$$

(1) Wick rotation (10.39): $\mathrm{d}^d\ell = i\,\mathrm{d}^d\ell_E$ and $\ell^2 - \Delta = -(\ell_E^2+\Delta)$.

(2) $X^{-n} = \frac{1}{\Gamma(n)}\int_0^\infty \mathrm{d}s\; s^{n-1}e^{-sX}$, the definition of the Gamma function after rescaling $s \to s/X$.

(3) The Gaussian integral.

(4) $\int_0^\infty \mathrm{d}s\; s^{m-1}e^{-s\Delta} = \Gamma(m)\Delta^{-m}$ with $m = n - \frac{d}{2}$, and $\pi^{d/2}/(2\pi)^d = (4\pi)^{-d/2}$.

The same steps with an extra $\ell^2 = -\ell_E^2$ in the numerator, using $\int \mathrm{d}^d\ell_E\, \ell_E^2 e^{-s\ell_E^2} = -\partial_s(\pi/s)^{d/2} = \frac{d}{2s}(\pi/s)^{d/2}$, give

$$\int \frac{\mathrm{d}^d\ell}{(2\pi)^d}\frac{\ell^2}{(\ell^2-\Delta)^n} = \frac{(-1)^{n-1}\,i}{(4\pi)^{d/2}}\,\frac{d}{2}\,\frac{\Gamma(n-\frac{d}{2}-1)}{\Gamma(n)}\,\Delta^{d/2-n+1}. \tag{10.41}$$

Finally, **symmetric integration**: inside an integral whose other factors depend on ℓ only through ℓ^2,

$$\ell^\mu \to 0, \qquad \ell^\mu\ell^\nu \to \frac{\ell^2}{d}\,\eta^{\mu\nu}, \qquad \ell^\mu\ell^\nu\ell^\rho \to 0. \tag{10.42}$$

Odd powers vanish because $\ell \to -\ell$ is a symmetry of the rest of the integrand. For the even one, Lorentz invariance forces $\int \ell^\mu\ell^\nu f(\ell^2) = c\,\eta^{\mu\nu}$; contracting with $\eta_{\mu\nu}$ gives $\int \ell^2 f = c\,d$, so $c = \frac{1}{d}\int \ell^2 f$.

Tool 5: dimensional regularisation and the Gamma function

When an integral diverges in $d = 2$ or $d = 4$, **dimensional regularisation** evaluates it with (10.40)–(10.42) for general complex d, where the right-hand sides make sense, and takes $d \to 2$ or 4 at the end. Divergences show up as poles of the Gamma function. The facts we need are

$$\Gamma(z+1) = z\,\Gamma(z), \qquad \Gamma(1) = 1, \qquad \Gamma(\epsilon) = \frac{1}{\epsilon} - \gamma_E + O(\epsilon), \tag{10.43}$$

with $\gamma_E \approx 0.577$ the Euler constant. Two features matter for anomalies:

- Shifting the loop momentum, $\ell \to \ell + a$, is always allowed, because the formulas above depend only on ℓ^2 after completing the square. This is what makes dimensional regularisation respect gauge invariance.
- Factors of d from $\eta^\mu{}_\mu = d$ and from (10.42) are kept as d until the end. A factor $(d-2)$ or $(d-4)$ multiplying a pole $\Gamma(1-\frac{d}{2})$ or $\Gamma(2-\frac{d}{2})$ leaves a *finite* remainder: for instance $(1-\frac{d}{2})\Gamma(1-\frac{d}{2}) = \Gamma(2-\frac{d}{2}) \to 1$ as $d \to 2$. This "$0\times\infty$" is how every anomaly below appears.

10.3.3 Two dimensions: QED$_2$ and the vacuum polarisation

Two-dimensional massless QED (the **Schwinger model**) is the simplest theory with an anomaly. We use Minkowski signature here, because the physical interpretation (Section 10.3.5) is Hamiltonian, and we follow the conventions of [QFT notes §22.2]:

$$\eta = \mathrm{diag}(+1,-1), \qquad \gamma^0 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \gamma^1 = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}, \quad \gamma^5 := \gamma^0\gamma^1 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$
$$\mathcal{L} = \bar\psi\, i\gamma^\mu(\partial_\mu + ieA_\mu)\psi - \tfrac{1}{4}F_{\mu\nu}F^{\mu\nu}, \qquad \epsilon^{01} = +1. \tag{10.44}$$

(One checks $(\gamma^0)^2 = \mathbf{1}$, $(\gamma^1)^2 = -\mathbf{1}$, $\gamma^0\gamma^1 = -\gamma^1\gamma^0$ by multiplying the matrices. The interaction term is $-eA_\mu j^\mu$.)

A two-dimensional miracle: the axial current is the dual of the vector current. In two dimensions,

$$\gamma^\mu\gamma^5 = -\epsilon^{\mu\nu}\gamma_\nu, \qquad \text{hence} \qquad j_5^\mu = -\epsilon^{\mu\nu}j_\nu. \tag{10.45}$$

We check the first identity component by component:

$$\gamma^0\gamma^5 \overset{(1)}{=} \gamma^0\gamma^0\gamma^1 \overset{(2)}{=} \gamma^1, \qquad -\epsilon^{0\nu}\gamma_\nu \overset{(3)}{=} -\gamma_1 \overset{(4)}{=} \gamma^1,$$
$$\gamma^1\gamma^5 \overset{(1)}{=} \gamma^1\gamma^0\gamma^1 \overset{(5)}{=} -\gamma^0(\gamma^1)^2 = \gamma^0, \qquad -\epsilon^{1\nu}\gamma_\nu \overset{(3)}{=} +\gamma_0 = \gamma^0.$$

(1) Definition $\gamma^5 = \gamma^0\gamma^1$.
(2) $(\gamma^0)^2 = \mathbf{1}$.
(3) Only one term survives, with $\epsilon^{01} = 1 = -\epsilon^{10}$.
(4) Lowering a spatial index flips its sign: $\gamma_1 = \eta_{11}\gamma^1 = -\gamma^1$.
(5) Anticommute γ^1 past γ^0, then $(\gamma^1)^2 = -\mathbf{1}$.

Sandwiching between $\bar\psi$ and ψ gives the second identity. It means that in two dimensions *any* statement about j^μ is also a statement about j_5^μ.

The vacuum polarisation. Put the fermions in a weak background field A_μ. To first order, the induced current is linear in A: $\langle j^\mu(q)\rangle = -\frac{1}{e}\Pi^{\mu\nu}(q)A_\nu(q)$, where $i\Pi^{\mu\nu}(q)$ is the one-loop photon self-energy [QFT notes (22.2.15)–(22.2.23)]. By Lorentz invariance

$$\Pi^{\mu\nu}(q) = \mathsf{a}\,\eta^{\mu\nu} - \mathsf{b}\,\frac{q^\mu q^\nu}{q^2} \tag{10.46}$$

for two numbers a, b (in two dimensions both are dimensionless, times e^2, so no other structure is possible). Now observe:

$$q_\mu\langle j^\mu\rangle \overset{(1)}{=} -\tfrac{1}{e}(\mathsf{a}-\mathsf{b})q^\nu A_\nu,$$
$$q_\mu\langle j_5^\mu\rangle \overset{(2)}{=} \tfrac{1}{e}\,\epsilon^{\mu\nu}q_\mu\Big(\mathsf{a}A_\nu - \mathsf{b}\frac{q_\nu\, q\cdot A}{q^2}\Big) \overset{(3)}{=} \tfrac{\mathsf{a}}{e}\,\epsilon^{\mu\nu}q_\mu A_\nu. \tag{10.47}$$

(1) Contract (10.46) with q_μ: $q_\mu\eta^{\mu\nu} = q^\nu$ and $q_\mu q^\mu q^\nu/q^2 = q^\nu$.
(2) Use $j_5^\mu = -\epsilon^{\mu\nu}j_\nu$ from (10.45).
(3) $\epsilon^{\mu\nu}q_\mu q_\nu = 0$ by antisymmetry.

So the vector current is conserved iff $\mathsf{a} = \mathsf{b}$, and the axial current is conserved iff $\mathsf{a} = 0$. *Both* are conserved only if $\mathsf{a} = \mathsf{b} = 0$. The whole question is therefore: is b zero?

Worked example 10.10: Computing b: it is finite and nonzero

The integral. The Feynman rules of QED (in the conventions (10.44)) are: a vertex $-ie\gamma^\mu$ for each photon, a propagator $i/\not\ell = i\not\ell/\ell^2$ for each fermion line (Tool 1(a)), a factor -1 and a trace for the closed fermion loop, and $\int \mathrm{d}^2\ell/(2\pi)^2$ over the undetermined loop momentum ℓ. The loop has one propagator with momentum ℓ and one with $\ell + q$. So

$$i\Pi^{\mu\nu}(q) \overset{(1)}{=} (-1)\int \frac{\mathrm{d}^2\ell}{(2\pi)^2}\,\mathrm{tr}\Big[(-ie\gamma^\mu)\frac{i\not\ell}{\ell^2}(-ie\gamma^\nu)\frac{i(\not\ell+\not q)}{(\ell+q)^2}\Big]$$
$$\overset{(2)}{=} -e^2\int \frac{\mathrm{d}^2\ell}{(2\pi)^2}\,\frac{\mathrm{tr}\big[\gamma^\mu\not\ell\,\gamma^\nu(\not\ell+\not q)\big]}{\ell^2(\ell+q)^2}. \tag{10.48}$$

(1) The Feynman rules just listed.

(2) Collect the constants: $(-1)\cdot(-ie)^2\cdot i^2 = (-1)(-e^2)(-1) = -e^2$.

The numerator. By the trace of four (10.32) with $N_2 = 2$, $a = \ell$ in the second slot and $b = \ell + q$ in the fourth,

$$\operatorname{tr}\big[\gamma^\mu \ell \gamma^\nu (\ell\!\!\!/ + q\!\!\!/)\big] = 2\big[\ell^\mu(\ell+q)^\nu + \ell^\nu(\ell+q)^\mu - \eta^{\mu\nu}\,\ell\cdot(\ell+q)\big]. \tag{10.49}$$

The denominator. Use (10.37) with $A = (\ell+q)^2$, $B = \ell^2$ and complete the square:

$$\begin{aligned} xA + (1-x)B &\overset{(1)}{=} x\big(\ell^2 + 2\ell\cdot q + q^2\big) + (1-x)\ell^2 \overset{(2)}{=} \ell^2 + 2x\,\ell\cdot q + xq^2 \\ &\overset{(3)}{=} (\ell + xq)^2 - x^2q^2 + xq^2 \overset{(4)}{=} r^2 - \Delta, \qquad r := \ell + xq, \quad \Delta := -x(1-x)q^2. \end{aligned} \tag{10.50}$$

(1) Expand $(\ell+q)^2$.
(2) The $x\ell^2$ terms combine with $(1-x)\ell^2$ to ℓ^2.
(3) $(\ell+xq)^2 = \ell^2 + 2x\ell\cdot q + x^2q^2$.
(4) $xq^2 - x^2q^2 = x(1-x)q^2 = -\Delta$.

(For spacelike q, $q^2 < 0$, so $\Delta > 0$ and (10.40) applies directly; other q follow by analytic continuation.) We now shift the integration variable from ℓ to r; the integral we are shifting will turn out to be at most logarithmically divergent in $d = 2$, so in dimensional regularisation the shift is certainly allowed.

The numerator in terms of r. With $\ell = r - xq$ and $\ell + q = r + (1-x)q$,

$$\begin{aligned} \ell^\mu(\ell+q)^\nu &\overset{(1)}{=} r^\mu r^\nu + (1-x)r^\mu q^\nu - x\,q^\mu r^\nu - x(1-x)q^\mu q^\nu, \\ \ell\cdot(\ell+q) &\overset{(2)}{=} r^2 + (1-2x)\,r\cdot q - x(1-x)q^2. \end{aligned} \tag{10.51}$$

(1) Multiply out $(r-xq)^\mu(r+(1-x)q)^\nu$.
(2) Contract the same product with $\eta_{\mu\nu}$.

The terms linear in r vanish after integration by (10.42). Keeping the even terms, (10.49) becomes

$$2\Big[\underbrace{2r^\mu r^\nu - \eta^{\mu\nu}r^2}_{\text{divergent}} \;\underbrace{-2x(1-x)\,q^\mu q^\nu}_{\text{gives b}} \;\underbrace{+x(1-x)q^2\,\eta^{\mu\nu}}_{\text{finite part of a}}\Big], \tag{10.52}$$

where we added (10.51) to itself with $\mu \leftrightarrow \nu$ for the first two terms of (10.49). Altogether

$$i\Pi^{\mu\nu}(q) = -e^2\int_0^1 \mathrm{d}x \int \frac{\mathrm{d}^d r}{(2\pi)^d}\,\frac{2\big[2r^\mu r^\nu - \eta^{\mu\nu}r^2 - 2x(1-x)q^\mu q^\nu + x(1-x)q^2\eta^{\mu\nu}\big]}{(r^2-\Delta)^2}, \tag{10.53}$$

with $d = 2$ for the convergent terms.

The $q^\mu q^\nu$ term.

$$\begin{aligned} i\Pi^{\mu\nu}\big|_{q^\mu q^\nu} &\overset{(1)}{=} -e^2\int_0^1 \mathrm{d}x\,\big(-4x(1-x)\big)q^\mu q^\nu \int \frac{\mathrm{d}^2 r}{(2\pi)^2}\frac{1}{(r^2-\Delta)^2} \\ &\overset{(2)}{=} -e^2\int_0^1 \mathrm{d}x\,\big(-4x(1-x)\big)q^\mu q^\nu\,\frac{i}{4\pi\Delta} \\ &\overset{(3)}{=} -e^2\int_0^1 \mathrm{d}x\,\frac{-4x(1-x)\,i}{4\pi\big(-x(1-x)q^2\big)}\,q^\mu q^\nu \overset{(4)}{=} -\frac{ie^2}{\pi}\,\frac{q^\mu q^\nu}{q^2}\int_0^1 \mathrm{d}x = -\frac{ie^2}{\pi}\frac{q^\mu q^\nu}{q^2}. \end{aligned}$$

(1) The $q^\mu q^\nu$ term of (10.53): $2\times(-2x(1-x)) = -4x(1-x)$.

(2) Master integral (10.40) at $d = 2$, $n = 2$: $(-1)^2\, i\, (4\pi)^{-1}\Gamma(1)/\Gamma(2)\, \Delta^{-1}$. It is finite.
(3) Insert $\Delta = -x(1-x)q^2$.
(4) The $x(1-x)$ cancels (a coincidence of $d = 2$), and $\frac{-4i}{-4\pi} = \frac{i}{\pi}$.

Comparing with (10.46), $-\mathsf{b} = -e^2/\pi$, i.e. $\mathsf{b} = e^2/\pi$. It came from a *convergent* integral, so no regularisation can change it.

The coefficient a is where regularisation enters. The $\eta^{\mu\nu}$ part of (10.53) has two pieces. The *finite piece* is the same in every scheme:

$$i\Pi^{\mu\nu}\big|_{\text{finite},\,\eta} \overset{(1)}{=} -e^2 \int_0^1 \mathrm{d}x\; 2x(1-x)q^2\,\eta^{\mu\nu}\,\frac{i}{4\pi\Delta} \overset{(2)}{=} -e^2\int_0^1 \mathrm{d}x\;\frac{2x(1-x)q^2\,i}{4\pi(-x(1-x)q^2)}\,\eta^{\mu\nu} \overset{(3)}{=} \frac{ie^2}{2\pi}\,\eta^{\mu\nu}. \tag{10.54}$$

(1) The last term of (10.53), and the master integral (10.40) at $d = 2$, $n = 2$.
(2) Insert Δ.
(3) Cancel $x(1-x)q^2$; $-e^2\cdot\frac{2i}{-4\pi} = \frac{ie^2}{2\pi}$; $\int_0^1 \mathrm{d}x = 1$.

The *divergent piece* $2r^\mu r^\nu - \eta^{\mu\nu}r^2$ has numerator $\sim r^2$ against denominator $\sim r^4$ in two dimensions: logarithmically divergent. Its value depends on how we define it.

Procedure 1: naive symmetric integration in exactly $d = 2$. Replace $r^\mu r^\nu \to \frac{1}{2}\eta^{\mu\nu}r^2$ by (10.42) with $d = 2$. Then $2r^\mu r^\nu - \eta^{\mu\nu}r^2 \to \eta^{\mu\nu}r^2 - \eta^{\mu\nu}r^2 = 0$ *before* anyone has asked whether the integral converges. Only (10.54) survives: $i\Pi|_\eta = \frac{ie^2}{2\pi}\eta^{\mu\nu}$, so $\mathsf{a} = e^2/2\pi = \frac{1}{2}\mathsf{b}$. By (10.47) neither current is conserved.

Procedure 2: dimensional regularisation. Keep d general until the end:

$$\begin{aligned} i\Pi^{\mu\nu}\big|_{\text{div}} &\overset{(1)}{=} -2e^2\int_0^1\mathrm{d}x\int\frac{\mathrm{d}^d r}{(2\pi)^d}\,\frac{(\frac{2}{d}-1)\eta^{\mu\nu}r^2}{(r^2-\Delta)^2} \overset{(2)}{=} -2e^2\,\frac{2-d}{d}\,\eta^{\mu\nu}\int_0^1\mathrm{d}x\int\frac{\mathrm{d}^d r}{(2\pi)^d}\,\frac{r^2}{(r^2-\Delta)^2} \\ &\overset{(3)}{=} -2e^2\,\frac{2-d}{d}\,\eta^{\mu\nu}\int_0^1\mathrm{d}x\,\frac{(-1)\,i}{(4\pi)^{d/2}}\,\frac{d}{2}\,\frac{\Gamma(1-\frac{d}{2})}{\Gamma(2)}\,\Delta^{d/2-1} \\ &\overset{(4)}{=} 2ie^2\,\eta^{\mu\nu}\int_0^1\mathrm{d}x\,\frac{(1-\frac{d}{2})\Gamma(1-\frac{d}{2})}{(4\pi)^{d/2}}\,\Delta^{d/2-1} \overset{(5)}{=} 2ie^2\,\eta^{\mu\nu}\int_0^1\mathrm{d}x\,\frac{\Gamma(2-\frac{d}{2})}{(4\pi)^{d/2}}\,\Delta^{d/2-1} \\ &\overset{(6)}{=} \frac{2ie^2}{4\pi}\,\eta^{\mu\nu} = \frac{ie^2}{2\pi}\,\eta^{\mu\nu} \qquad (d\to 2). \end{aligned} \tag{10.55}$$

(1) Symmetric integration in d dimensions, $r^\mu r^\nu \to \frac{1}{d}\eta^{\mu\nu}r^2$ (10.42), so $2r^\mu r^\nu - \eta^{\mu\nu}r^2 \to (\frac{2}{d}-1)\eta^{\mu\nu}r^2$; the overall 2 of (10.53) is pulled out.
(2) $\frac{2}{d} - 1 = \frac{2-d}{d}$.
(3) Master integral with r^2 in the numerator (10.41) at $n = 2$: $(-1)^{n-1} = -1$. At $d = 2$ it has a pole, $\Gamma(1-\frac{d}{2}) \to \infty$.
(4) Collect: $-2e^2\cdot\frac{2-d}{d}\cdot(-i)\cdot\frac{d}{2} = ie^2(2-d) = 2ie^2(1-\frac{d}{2})$.
(5) $z\Gamma(z) = \Gamma(z+1)$ with $z = 1-\frac{d}{2}$ (10.43). The vanishing factor has eaten the pole exactly.
(6) At $d = 2$: $\Gamma(1) = 1$, $(4\pi)^{d/2} = 4\pi$, $\Delta^0 = 1$, $\int_0^1 \mathrm{d}x = 1$.

The divergent piece is not zero: it is a finite "$0\times\infty$" remainder equal to the finite piece. Adding (10.54) and (10.55), $i\Pi|_\eta = \frac{ie^2}{\pi}\eta^{\mu\nu}$, so $\mathsf{a} = e^2/\pi = \mathsf{b}$. Dimensional regularisation respects gauge invariance, and it produces $\mathsf{a} = \mathsf{b}$: the vector current is conserved, and then

$$\Pi^{\mu\nu}(q) = \frac{e^2}{\pi}\Big(\eta^{\mu\nu} - \frac{q^\mu q^\nu}{q^2}\Big), \qquad \langle j^\mu\rangle = -\frac{e}{\pi}\Big(\eta^{\mu\nu} - \frac{\partial^\mu\partial^\nu}{\Box}\Big)A_\nu, \tag{10.56}$$

in agreement with [QFT notes (22.2.18), (22.2.23)]. To return to position space, write $A_\nu(x) = \int \frac{d^2q}{(2\pi)^2} e^{-iq\cdot x}\tilde{A}_\nu(q)$ and similarly for the current; then ∂_μ acting on $e^{-iq\cdot x}$ brings down $-iq_\mu$. Multiplying the second line of (10.47) by $-i$,

$$(-iq_\mu)\langle \tilde{j}_5^\mu\rangle \overset{(1)}{=} \frac{\mathsf{a}}{e}\,\epsilon^{\mu\nu}(-iq_\mu)\tilde{A}_\nu \quad\Longrightarrow\quad \partial_\mu\langle j_5^\mu\rangle \overset{(2)}{=} \frac{\mathsf{a}}{e}\,\epsilon^{\mu\nu}\partial_\mu A_\nu \overset{(3)}{=} \frac{e}{\pi}\,\epsilon^{\mu\nu}\partial_\mu A_\nu,$$

(1) Both sides contain exactly one factor of q, so one factor of $-i$ goes to each.
(2) Replace $-iq_\mu \to \partial_\mu$ under the Fourier integral.
(3) $\mathsf{a} = e^2/\pi$.

that is,

$$\boxed{\partial_\mu\langle j_5^\mu\rangle = \frac{e}{\pi}\,\epsilon^{\mu\nu}\partial_\mu A_\nu = \frac{e}{2\pi}\,\epsilon^{\mu\nu}F_{\mu\nu}} \qquad \text{[QFT notes (22.2.27)].} \tag{10.57}$$

The last step is $\epsilon^{\mu\nu}F_{\mu\nu} = \epsilon^{\mu\nu}(\partial_\mu A_\nu - \partial_\nu A_\mu) = 2\epsilon^{\mu\nu}\partial_\mu A_\nu$, by relabelling $\mu \leftrightarrow \nu$ in the second term.

Physical picture: The photon gets a mass

Put (10.56) into Maxwell's equation $\partial_\mu F^{\mu\nu} = e\langle j^\nu\rangle$. In Lorenz gauge $\partial\cdot A = 0$ it reads $\Box A^\nu = -\frac{e^2}{\pi}A^\nu$, i.e. $(\Box + m^2)A^\nu = 0$ with

$$m_\gamma^2 = \frac{e^2}{\pi}.$$

The photon of the Schwinger model is massive, and this one-loop result is exact [QFT notes (22.2.21)]. The pole $1/q^2$ in $\Pi(q^2) = e^2/\pi q^2$ is the reason: in two dimensions a massless fermion–antifermion pair moving in the same direction behaves like a single massless particle, which the photon eats. The anomaly and the photon mass are the same number, $\mathsf{b} = e^2/\pi$.

Watch out: Which current is "anomalous" is a choice

Equation (10.47) leaves a one-parameter family of possibilities. Setting $\mathsf{a} = 0$ instead would conserve j_5^μ and break j^μ: $q_\mu\langle j^\mu\rangle = (e/\pi)\,q\cdot A$ [QFT notes p. 1371]. What is *not* a choice is that $\mathsf{b} = e^2/\pi \neq 0$, so one of them must go. We keep the vector current because the photon couples to it: breaking it would make the physics depend on the gauge. Nash's phrase "the lack of gauge invariance must occur somewhere" is this statement.

10.3.4 Two dimensions again: the point-split current

The vacuum-polarisation argument shows *that* there is an anomaly. The operator argument shows *where* it comes from: the current $j_5^\mu(x)$ is a product of two fields at the same point, and such products are singular.

Working definition 10.5 (gauge-invariant point splitting)

Define the current as a limit of fields at two nearby points, joined by a Wilson line [the phase factor $\exp[-ie\int A]$ along the short segment between the two points. Under a gauge transformation it picks up the phases of both ends] so that the combination is gauge invariant [QFT notes (22.2.31)]:

$$j_5^\mu(x) := \operatorname*{s\text{-}lim}_{\varepsilon\to 0}\ \bar{\psi}\big(x+\tfrac{\varepsilon}{2}\big)\,\gamma^\mu\gamma^5\,\exp\Big[-ie\int_{x-\varepsilon/2}^{x+\varepsilon/2} dz^\sigma A_\sigma(z)\Big]\,\psi\big(x-\tfrac{\varepsilon}{2}\big).$$

The **symmetric limit** averages over the direction of ε: s-lim $\varepsilon^\mu/\varepsilon^2 = 0$ and s-lim $\varepsilon^\mu\varepsilon^\nu/\varepsilon^2 = \eta^{\mu\nu}/d$. *Test of the definition:* under $\psi \to e^{-ie\alpha}\psi$, $A \to A + \mathrm{d}\alpha$, the Wilson line picks up $e^{-ie[\alpha(x+\varepsilon/2)-\alpha(x-\varepsilon/2)]}$, which cancels the phases of the two fermion fields.

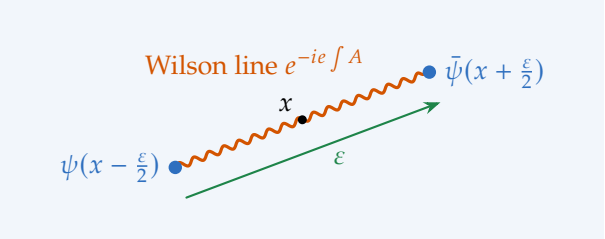


Figure 10.7: Point splitting. The two fermion fields sit a distance ε apart and are joined by a Wilson line to keep the product gauge invariant. Differentiating the Wilson line produces a factor $\varepsilon^\nu F_{\mu\nu}$, which vanishes as $\varepsilon \to 0$; but the fermion pair has a $1/\varepsilon$ singularity. The product of "zero" and "infinity" is finite and nonzero: that is the anomaly.

Differentiating the split current. Write $x_\pm := x \pm \frac{\varepsilon}{2}$ and $W := \exp\!\big[-ie\int_{x_-}^{x_+} \mathrm{d}z^\sigma A_\sigma(z)\big]$, and take ε along a fixed direction while we differentiate with respect to x. We need three ingredients.

(i) The equations of motion. The Dirac equation from (10.44) is $i\gamma^\mu(\partial_\mu + ieA_\mu)\psi = 0$, i.e. $\gamma^\mu\partial_\mu\psi = -ie\not{A}\psi$. Its Dirac conjugate (take the Hermitian conjugate, multiply by γ^0 on the right, and use $\gamma^0\gamma^{\mu\dagger}\gamma^0 = \gamma^\mu$) is $(\partial_\mu\bar\psi)\gamma^\mu = ie\,\bar\psi\not{A}$.

(ii) The derivative of the Wilson line. Parametrise the straight segment as $z = x + s\varepsilon$, $s \in [-\frac12, \frac12]$, so $\mathrm{d}z^\sigma = \varepsilon^\sigma \mathrm{d}s$. Then

$$\begin{aligned}
\partial_\mu W &\overset{(1)}{=} W\cdot(-ie)\,\partial_\mu \int_{-1/2}^{1/2} \mathrm{d}s\ \varepsilon^\sigma A_\sigma(x+s\varepsilon) \\
&\overset{(2)}{=} W\cdot(-ie)\int_{-1/2}^{1/2} \mathrm{d}s\ \varepsilon^\sigma\big[\partial_\mu A_\sigma(x) + s\,\varepsilon^\tau\partial_\tau\partial_\mu A_\sigma(x) + \cdots\big] \\
&\overset{(3)}{=} -ie\,\varepsilon^\sigma\partial_\mu A_\sigma(x)\,W + O(\varepsilon^3) \overset{(4)}{=} -ie\,\varepsilon^\sigma\partial_\mu A_\sigma(x) + O(\varepsilon^2).
\end{aligned} \tag{10.58}$$

(1) Chain rule on the exponential; the endpoints move with x, and this is taken into account by writing the integral in the variable s, whose range is fixed.
(2) Taylor expand A_σ about x.
(3) $\int_{-1/2}^{1/2}\mathrm{d}s = 1$ and $\int_{-1/2}^{1/2} s\,\mathrm{d}s = 0$; the next surviving term is $O(\varepsilon^3)$.
(4) $W = 1 + O(\varepsilon)$.

(iii) Moving γ^5. In two dimensions $\gamma^5 = \gamma^0\gamma^1$ anticommutes with γ^0 and γ^1 (check: $\gamma^0\gamma^5 = \gamma^1$ while $\gamma^5\gamma^0 = \gamma^0\gamma^1\gamma^0 = -\gamma^1$), hence with every $\not{\phi}$.

Now apply the product rule. W is a number (a phase), so it commutes with all matrices.

$$\begin{aligned}
\partial_\mu\big[\bar\psi(x_+)\gamma^\mu\gamma^5 W\psi(x_-)\big] &\overset{(1)}{=} (\partial_\mu\bar\psi)(x_+)\gamma^\mu\gamma^5 W\psi(x_-) + \bar\psi(x_+)\gamma^\mu\gamma^5(\partial_\mu W)\psi(x_-) \\
&\quad + \bar\psi(x_+)\gamma^\mu\gamma^5 W(\partial_\mu\psi)(x_-) \\
&\overset{(2)}{=} ie\,\bar\psi\not{A}(x_+)\gamma^5 W\psi + \bar\psi\gamma^\mu\gamma^5(\partial_\mu W)\psi - \bar\psi\gamma^5 W\gamma^\mu(\partial_\mu\psi)(x_-) \\
&\overset{(3)}{=} ie\,\bar\psi\not{A}(x_+)\gamma^5 W\psi + \bar\psi\gamma^\mu\gamma^5(\partial_\mu W)\psi + ie\,\bar\psi\gamma^5\not{A}(x_-)W\psi \\
&\overset{(4)}{=} ie\,\bar\psi\big[\not{A}(x_+) - \not{A}(x_-)\big]\gamma^5 W\psi + \bar\psi\gamma^\mu\gamma^5(\partial_\mu W)\psi
\end{aligned}$$

$$\overset{(5)}{=} ie\,\varepsilon^\nu \partial_\nu A_\mu\ \bar\psi\gamma^\mu\gamma^5\psi - ie\,\varepsilon^\nu\partial_\mu A_\nu\ \bar\psi\gamma^\mu\gamma^5\psi + O(\varepsilon^2)$$
$$\overset{(6)}{=} -ie\,\varepsilon^\nu F_{\mu\nu}\ \bar\psi(x_+)\gamma^\mu\gamma^5\psi(x_-) + O(\varepsilon^2). \tag{10.59}$$

(1) Product rule: ∂_μ hits $\bar\psi(x_+)$, W, or $\psi(x_-)$. (We abbreviate $\bar\psi = \bar\psi(x_+)$ and $\psi = \psi(x_-)$ from now on.)
(2) First term: equation of motion for $\bar\psi$ from (i). Third term: anticommute γ^μ past γ^5 by (iii), which costs a minus sign.
(3) Third term: $\gamma^\mu\partial_\mu\psi = -ie\not{A}\psi$ from (i); the two minus signs give $+ie$.
(4) Third term: $\gamma^5\not{A} = -\not{A}\gamma^5$ by (iii), and the first and third terms combine.
(5) First term: Taylor expand $A_\mu(x_+) - A_\mu(x_-) = \varepsilon^\nu\partial_\nu A_\mu(x) + O(\varepsilon^3)$ and set $W = 1$ since the prefactor is already $O(\varepsilon)$. Second term: (10.58).
(6) $\partial_\nu A_\mu - \partial_\mu A_\nu = -(\partial_\mu A_\nu - \partial_\nu A_\mu) = -F_{\mu\nu}$.

Classically we would now set $\varepsilon = 0$ and get zero. Quantum mechanically the fermion bilinear has a vacuum expectation value that blows up as $\varepsilon \to 0$.

The singular fermion bilinear. We need the free massless propagator in two dimensions. In momentum space it is $i\not{k}/(k^2 + i0)$. Write it as $i\not{\partial}$ acting on the scalar propagator $D(x) = \int \frac{d^2k}{(2\pi)^2}\frac{i}{k^2+i0}e^{-ik\cdot x}$, since $i\partial_\mu e^{-ik\cdot x} = k_\mu e^{-ik\cdot x}$. The two-dimensional massless scalar propagator is $D(x) = -\frac{1}{4\pi}\ln(-\mu^2x^2 + i0)$, with μ an infrared scale that drops out of derivatives. (Check: in Euclidean signature it is the familiar $-\frac{1}{2\pi}\ln|x|$, the Green's function of $-\nabla^2$ in the plane.) So

$$\langle\psi(x)\bar\psi(y)\rangle \overset{(1)}{=} i\gamma^\mu\partial_\mu D(x-y) \overset{(2)}{=} i\gamma^\mu\left(-\frac{1}{4\pi}\right)\frac{2(x-y)_\mu}{(x-y)^2} = -\frac{i}{2\pi}\,\frac{\gamma^\mu(x-y)_\mu}{(x-y)^2} \qquad \text{[QFT notes (22.2.39)]}. \tag{10.60}$$

(1) The Fourier representation just described.
(2) $\partial_\mu\ln(-x^2) = 2x_\mu/x^2$.

For any 2×2 matrix Γ,

$$\langle\bar\psi(x_+)\Gamma\psi(x_-)\rangle \overset{(1)}{=} \Gamma_{\alpha\beta}\,\langle\bar\psi_\alpha(x_+)\psi_\beta(x_-)\rangle \overset{(2)}{=} -\Gamma_{\alpha\beta}\,\langle\psi_\beta(x_-)\bar\psi_\alpha(x_+)\rangle$$
$$\overset{(3)}{=} -\Gamma_{\alpha\beta}\left(-\frac{i}{2\pi}\right)\frac{(\gamma^\sigma)_{\beta\alpha}(-\varepsilon_\sigma)}{\varepsilon^2} \overset{(4)}{=} -\frac{i}{2\pi}\,\frac{\mathrm{tr}(\Gamma\gamma^\sigma)\,\varepsilon_\sigma}{\varepsilon^2}. \tag{10.61}$$

(1) Write out the spinor indices; $\Gamma_{\alpha\beta}$ is a number.
(2) Fermion fields anticommute (inside the time-ordered product).
(3) The propagator (10.60) with $x - y = x_- - x_+ = -\varepsilon$ and $(x-y)^2 = \varepsilon^2$.
(4) $(-1)(-1) = 1$, and $\Gamma_{\alpha\beta}(\gamma^\sigma)_{\beta\alpha} = \mathrm{tr}(\Gamma\gamma^\sigma)$.

With $\Gamma = \gamma^\mu\gamma^5$ we need one trace:

$$\mathrm{tr}(\gamma^\mu\gamma^5\gamma^\sigma) \overset{(1)}{=} \mathrm{tr}(\gamma^\sigma\gamma^\mu\gamma^5) \overset{(2)}{=} -\epsilon^{\mu\nu}\,\mathrm{tr}(\gamma^\sigma\gamma_\nu) \overset{(3)}{=} -2\epsilon^{\mu\nu}\delta^\sigma_\nu = -2\epsilon^{\mu\sigma}.$$

(1) Cyclicity.
(2) $\gamma^\mu\gamma^5 = -\epsilon^{\mu\nu}\gamma_\nu$ (10.45).
(3) The trace of two (10.31) with $N_2 = 2$: $\mathrm{tr}(\gamma^\sigma\gamma_\nu) = 2\delta^\sigma_\nu$.

(Check at $\mu = 0$, $\sigma = 1$: $\mathrm{tr}(\gamma^0\gamma^0\gamma^1\gamma^1) = \mathrm{tr}(\mathbf{1}\cdot(-\mathbf{1})) = -2 = -2\epsilon^{01}$.)

The symmetric limit. Averaging $\varepsilon^\nu \varepsilon_\sigma/\varepsilon^2$ over the directions of ε gives a Lorentz-invariant tensor, hence $c\,\delta^\nu_\sigma$; taking the trace ($\nu = \sigma$, summed) gives $1 = c\,d$. So

$$\operatorname*{s-lim}_{\varepsilon\to 0} \frac{\varepsilon^\nu \varepsilon_\sigma}{\varepsilon^2} = \frac{\delta^\nu_\sigma}{d}, \qquad \operatorname*{s-lim}_{\varepsilon\to 0} \frac{\varepsilon^\nu}{\varepsilon^2} = 0, \tag{10.62}$$

the second because it is odd under $\varepsilon \to -\varepsilon$. This is the position-space twin of the symmetric integration rule (10.42).

Putting everything together,

$$\begin{aligned} \partial_\mu \langle j^\mu_5 \rangle &\overset{(1)}{=} \operatorname*{s-lim}\big(-ie\,\varepsilon^\nu F_{\mu\nu}\big)\Big(-\frac{i}{2\pi}\Big)\frac{(-2\epsilon^{\mu\sigma})\varepsilon_\sigma}{\varepsilon^2} \\ &\overset{(2)}{=} \frac{e}{\pi}\,\epsilon^{\mu\sigma}F_{\mu\nu}\operatorname*{s-lim}\frac{\varepsilon^\nu\varepsilon_\sigma}{\varepsilon^2} \overset{(3)}{=} \frac{e}{\pi}\,\epsilon^{\mu\sigma}F_{\mu\nu}\,\frac{\delta^\nu_\sigma}{2} \overset{(4)}{=} \frac{e}{2\pi}\,\epsilon^{\mu\nu}F_{\mu\nu}. \end{aligned} \tag{10.63}$$

(1) Take the expectation value of (10.59) and insert (10.61) with the trace just computed; F is a background field, so it comes out of the expectation value. The $O(\varepsilon^2)$ terms times $1/\varepsilon$ vanish in the limit.
(2) The constants: $(-ie)\cdot\big(-\frac{i}{2\pi}\big)\cdot(-2) = \frac{e}{\pi}\cdot(-i)(-i)(-1) = \frac{e}{\pi}\cdot(-1)(-1) = \frac{e}{\pi}$.
(3) The symmetric limit (10.62) with $d = 2$.
(4) δ^ν_σ sets $\sigma = \nu$.

This is (10.57) again [QFT notes (22.2.45)]. The Wilson line was essential. Without it (or with the opposite sign in its exponent, which breaks gauge invariance) the same steps give $\partial_\mu j^\mu_5 = 0$ and a non-conserved vector current.

10.3.5 Spectral flow: where the chiral charge goes

The anomaly equation (10.57) is a local statement. Here is its global meaning, which is the physical picture behind everything in this chapter [QFT notes §22.2.4].

Put the fermions on a circle of length L (periodic boundary conditions), with $A_0 = 0$ and a spatially constant $A^1 =: a(t)$ that changes slowly in time. The Dirac equation $i\gamma^\mu D_\mu\psi = 0$, multiplied by γ^0, is $i\partial_t\psi = -i\gamma^5(\partial_1 + ieA_1)\psi$. For $\psi_\pm \propto e^{ikx}$ with $\gamma^5\psi_\pm = \pm\psi_\pm$:

$$E_\pm(k) \overset{(1)}{=} \pm(-i)(ik + ieA_1) \overset{(2)}{=} \pm(k - e\,a), \qquad k = k_n = \frac{2\pi n}{L},\ n \in \mathbb{Z}. \tag{10.64}$$

(1) Replace $\partial_1 \to ik$ and $\gamma^5 \to \pm 1$.
(2) $A_1 = \eta_{11}A^1 = -a$.

So positive chirality fermions are right-movers with energy $k - ea$ (the kinetic momentum), negative chirality fermions left-movers with energy $-(k - ea)$. The Dirac vacuum fills every level with $E < 0$ in both towers.

Counting. Increase a by $\Delta a = 2\pi/(eL)$. The set of levels is the same as before (the Wilson loop $e^{ie\oint A^1 dx}$ has returned to its value), but each level has moved by one spacing, as in Figure 10.8. Levels move slowly, so their occupation does not change. One empty right-mover level has sunk below zero and one filled left-mover level has risen above zero. Measured from the new vacuum, $\Delta N_+ = -1$, $\Delta N_- = +1$, and $\Delta Q_5 = \Delta(N_+ - N_-) = -2$, while $\Delta Q = \Delta(N_+ + N_-) = 0$: charge is conserved, chirality is not.

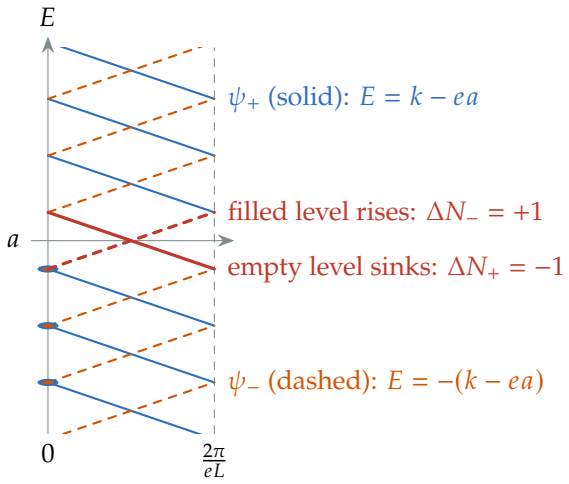


Figure 10.8: Spectral flow on a circle. As the constant potential a increases by $2\pi/eL$ the spectrum returns to itself, but every right-mover level (solid) has moved down one step and every left-mover level (dashed) up one step. The red levels cross $E = 0$. Occupations are preserved by the adiabatic theorem, so relative to the new vacuum one right-mover is missing and one extra left-mover is present: $\Delta(N_+ - N_-) = -2$, exactly what the anomaly equation predicts.

The anomaly predicts exactly this. Integrate (10.57) over the spacetime slab $[0,T]\times S^1$:

$$\Delta Q_5 \overset{(1)}{=} \int_0^T \mathrm{d}t \oint \mathrm{d}x\, \partial_\mu j_5^\mu \overset{(2)}{=} \frac{e}{2\pi}\int \mathrm{d}t\,\mathrm{d}x\, 2F_{01} \overset{(3)}{=} \frac{e}{\pi}\int \mathrm{d}t\,\mathrm{d}x\,\partial_0 A_1$$
$$\overset{(4)}{=} -\frac{e}{\pi}L\,\Delta a \overset{(5)}{=} -\frac{e}{\pi}L\,\frac{2\pi}{eL} = -2. \tag{10.65}$$

(1) $Q_5 = \oint j_5^0\,\mathrm{d}x$; the spatial divergence integrates to zero on the circle.
(2) $\epsilon^{\mu\nu}F_{\mu\nu} = \epsilon^{01}F_{01} + \epsilon^{10}F_{10} = 2F_{01}$.
(3) $F_{01} = \partial_0 A_1 - \partial_1 A_0 = \partial_0 A_1$ because $A_0 = 0$.
(4) $A_1 = -a$ is constant in x, so $\oint \mathrm{d}x = L$ and $\int \mathrm{d}t\,\partial_0 A_1 = -\Delta a$.
(5) The value of Δa that brings the spectrum back to itself.

The two counts agree.

Physical picture: An electric field pumps chirality

In terms of the electric field $E = -\partial_t a$, the anomaly says $\dot{Q}_5 = \frac{e}{\pi}L\,E$. Here is the elementary version. A force eE shifts every momentum at the rate $\dot{k} = eE$. The levels are spaced by $2\pi/L$, so each Fermi point sweeps through $eEL/2\pi$ levels per unit time. Right-movers gain particles at that rate and left-movers lose them: $\dot{N}_\pm = \pm eEL/2\pi$, so $\dot{Q}_5 = eEL/\pi$. The "creation from nothing" is really a rearrangement of the infinitely deep Dirac sea, which is why a regularisation (a way of saying what happens at $E \to -\infty$) is needed to see it, and why gauge invariance (charge conservation) decides which way the answer comes out.

Watch out: Levels crossing zero are an index

The number of levels that cross $E = 0$ in one direction minus the number crossing in the other is called the **spectral flow**. It is an integer and it is topological: small deformations of the path $a(t)$ cannot change it. In Euclidean language, the slab $[0,T]\times S^1$ with the ends identified (since the spectrum came back) is a torus carrying one unit of flux, $\frac{e}{2\pi}\int f = 1$, and the spectral flow of the one-dimensional Hamiltonian equals the index of the two-

dimensional Dirac operator on that torus. This is our first example of "integrated anomaly = index" (Section 10.4.5).

10.3.6 Four dimensions: the point-split current

We now repeat the point-splitting calculation in four dimensions (still Minkowski, $\eta = \mathrm{diag}(+,-,-,-)$, following [QFT notes §22.3]). The conventions are those of Tool 1(e): $\gamma^5 = i\gamma^0\gamma^1\gamma^2\gamma^3$, $\epsilon^{0123} = +1$ and $\mathrm{tr}(\gamma^5\gamma^\alpha\gamma^\beta\gamma^\gamma\gamma^\delta) = -4i\epsilon^{\alpha\beta\gamma\delta}$. The derivation of the operator identity (10.59) used only three things: the equations of motion, the derivative of the Wilson line, and the fact that γ^5 anticommutes with every γ^μ. All three hold in four dimensions (for the last, $\gamma^5\gamma^\mu = -\gamma^\mu\gamma^5$ because moving γ^μ through $\gamma^0\gamma^1\gamma^2\gamma^3$ costs three signs and one $(\gamma^\mu)^2$ that commutes back). So

$$\partial_\mu j_5^\mu = \operatorname*{s\text{-}lim}_{\varepsilon\to 0} \left(-ie\,\varepsilon^\nu F_{\mu\nu}\right) \bar\psi(x_+)\gamma^\mu\gamma^5\psi(x_-) \tag{10.66}$$

holds unchanged. What changes is the singular bilinear.

Step 1: the free propagator gives nothing. The four-dimensional scalar propagator is $D_4(x) = \int \frac{\mathrm{d}^4k}{(2\pi)^4}\frac{i\,e^{-ik\cdot x}}{k^2+i0} = -\frac{1}{4\pi^2(x^2-i0)}$ (in Euclidean signature this is $1/4\pi^2x_E^2$, the Green's function of $-\nabla^2$ in $\mathbb{R}^4$). Hence, exactly as in (10.60),

$$S(x-y) := \langle\psi(x)\bar\psi(y)\rangle_0 \overset{(1)}{=} i\gamma^\mu\partial_\mu D_4(x-y) \overset{(2)}{=} \frac{i}{2\pi^2}\,\frac{\gamma^\mu(x-y)_\mu}{(x-y)^4}. \tag{10.67}$$

(1) $i\not\partial e^{-ik\cdot x} = \not k\, e^{-ik\cdot x}$, so the Fourier transform of $i\not k/k^2$ is $i\not\partial D_4$.
(2) $\partial_\mu(-1/4\pi^2x^2) = 2x_\mu/4\pi^2x^4$.

By (10.61), whose derivation did not depend on the dimension, the free bilinear is

$$-\,\mathrm{tr}\big(\gamma^\mu\gamma^5 S(-\varepsilon)\big) \propto \mathrm{tr}(\gamma^5\gamma^\sigma\gamma^\mu)\,\frac{\varepsilon_\sigma}{\varepsilon^4},$$

which is *zero* by (10.36): a trace of γ^5 needs four other gamma matrices. The same is true with the Wilson line expanded to first order, since W is a number.

Step 2: the propagator in the background field, to first order. Let $G(x,y) = \langle\psi(x)\bar\psi(y)\rangle_A$ be the propagator in the background A. It obeys the Dirac equation with a source, $(i\not\partial_x - e\not A(x))G(x,y) = i\delta^4(x-y)$, and the free one obeys $i\not\partial_x S(x-y) = i\delta^4(x-y)$. Try $G = S + G_1$ with G_1 of first order in e:

$$\begin{aligned}(i\not\partial_x - e\not A)(S+G_1) &\overset{(1)}{=} i\delta^4(x-y) - e\not A(x)S(x-y) + i\not\partial_x G_1(x,y) + O(e^2)\\ \Longrightarrow\quad G_1(x,y) &\overset{(2)}{=} \int \mathrm{d}^4z\; S(x-z)\left(-ie\not A(z)\right)S(z-y).\end{aligned} \tag{10.68}$$

(1) Use the free equation; $e\not A G_1$ is second order.
(2) Setting the first-order terms to zero requires $i\not\partial_x G_1 = e\not A S$. Applying $i\not\partial_x$ to the proposed G_1 gives $\int \mathrm{d}^4z\, i\delta^4(x-z)(-ie\not A(z))S(z-y) = e\not A(x)S(x-y)$, as required.

This is the diagram of Figure 10.9: a free line, one interaction with the background (vertex $-ie\gamma^\alpha$), another free line.

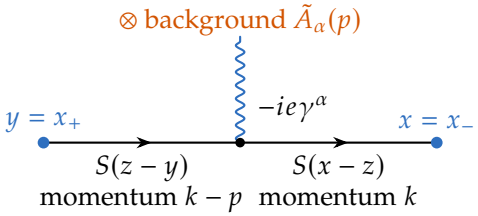


Figure 10.9: The first-order propagator G_1 of (10.68). The fermion line runs from $\bar\psi$ at x_+ to ψ at x_- and absorbs momentum p from the background field once. Compared with the free line it has one more propagator (less singular by one power of ε) but two more gamma matrices (γ^α and a second $\not k$), which is exactly what the trace with γ^5 needs.

Now go to momentum space: $A_\alpha(z) = \int_p e^{-ip\cdot z}\tilde A_\alpha(p)$ and $S(x) = \int_k \frac{i\not k}{k^2}e^{-ik\cdot x}$, with the shorthand $\int_k := \int \frac{\mathrm{d}^4k}{(2\pi)^4}$. Then

$$\begin{aligned} G_1(x,y) &\overset{(1)}{=} \int_p\int_k\int_{k'}\int \mathrm{d}^4z\; e^{-ik\cdot(x-z)}e^{-ip\cdot z}e^{-ik'\cdot(z-y)}\,\frac{i\not k}{k^2}\big(-ie\not{\tilde A}(p)\big)\frac{i\not k'}{k'^2} \\ &\overset{(2)}{=} \int_p\int_k e^{-ik\cdot x+i(k-p)\cdot y}\,\frac{i\not k\,(-ie\not{\tilde A}(p))\,i(\not k-\not p)}{k^2(k-p)^2} \\ &\overset{(3)}{=} ie\int_p\int_k e^{-ik\cdot x+i(k-p)\cdot y}\,\frac{\not k\,\not{\tilde A}(p)\,(\not k-\not p)}{k^2(k-p)^2}. \end{aligned} \tag{10.69}$$

(1) Insert the Fourier representations into (10.68).
(2) The z integral gives $(2\pi)^4\delta^4(k-p-k')$, which sets $k' = k-p$ (momentum conservation at the vertex) and cancels one $(2\pi)^4$.
(3) $i\cdot(-ie)\cdot i = -i^3e = ie$.

We need $G_1(x_-,x_+)$, i.e. $x = x_- = x - \frac{\varepsilon}{2}$ and $y = x_+ = x + \frac{\varepsilon}{2}$ in (10.69). The phase becomes

$$-ik\cdot\left(x-\tfrac{\varepsilon}{2}\right) + i(k-p)\cdot\left(x+\tfrac{\varepsilon}{2}\right) \overset{(1)}{=} -ip\cdot x + ik\cdot\varepsilon - \tfrac{i}{2}p\cdot\varepsilon.$$

(1) Multiply out: the $\pm ik\cdot x$ cancel, and $\frac{i}{2}k\cdot\varepsilon + \frac{i}{2}k\cdot\varepsilon = ik\cdot\varepsilon$.

Step 3: the trace. The bilinear is $B^\mu := \langle\bar\psi(x_+)\gamma^\mu\gamma^5\psi(x_-)\rangle = -\operatorname{tr}[\gamma^\mu\gamma^5 G(x_-,x_+)]$ by steps (1)–(2) of (10.61). The trace that appears is

$$\operatorname{tr}\big[\gamma^\mu\gamma^5\not k\gamma^\alpha(\not k-\not p)\big] \overset{(1)}{=} \operatorname{tr}\big[\gamma^5\not k\gamma^\alpha\not k\gamma^\mu\big] - \operatorname{tr}\big[\gamma^5\not k\gamma^\alpha\not p\gamma^\mu\big] \overset{(2)}{=} 0 + 4i\,\epsilon^{\kappa\alpha\pi\mu}k_\kappa p_\pi. \tag{10.70}$$

(1) Cyclicity moves γ^μ to the end; split $\not k - \not p$.
(2) (10.36): the first trace is $-4i\epsilon^{\kappa\alpha\lambda\mu}k_\kappa k_\lambda = 0$ (antisymmetric ϵ against symmetric $k_\kappa k_\lambda$); the second is $-4i\epsilon^{\kappa\alpha\pi\mu}k_\kappa p_\pi$.

Every term is therefore at least linear in p, i.e. in a derivative of A. As $\varepsilon\to 0$ the singular behaviour comes from large loop momenta, $|k|\sim 1/|\varepsilon| \gg |p|$. There we may set $(k-p)^2\to k^2$ and $e^{-\frac{i}{2}p\cdot\varepsilon}\to 1$: the corrections carry extra powers of $p\varepsilon$, i.e. more derivatives of A and more

powers of ε, and vanish in the limit. So

$$\begin{aligned} B^\mu &\overset{(1)}{=} -ie \int_p e^{-ip\cdot x}\tilde{A}_\alpha(p) \int_k e^{ik\cdot\varepsilon}\, \frac{4i\,\epsilon^{\kappa\alpha\pi\mu}k_\kappa p_\pi}{k^4} \\ &\overset{(2)}{=} 4e\,\epsilon^{\kappa\alpha\pi\mu}\Big[\int_p e^{-ip\cdot x}p_\pi\tilde{A}_\alpha(p)\Big] I_\kappa(\varepsilon), \qquad I_\kappa(\varepsilon) \coloneqq \int_k \frac{k_\kappa\, e^{ik\cdot\varepsilon}}{(k^2+i0)^2} \\ &\overset{(3)}{=} 4ie\,\epsilon^{\kappa\alpha\pi\mu}\,\partial_\pi A_\alpha(x)\, I_\kappa(\varepsilon) \overset{(4)}{=} 2ie\,\epsilon^{\kappa\alpha\pi\mu}F_{\pi\alpha}(x)\, I_\kappa(\varepsilon). \end{aligned} \tag{10.71}$$

(1) $B^\mu = -\operatorname{tr}[\gamma^\mu\gamma^5 G_1(x_-,x_+)]$ with (10.69), the phase, and the trace (10.70); the leading large-k form of the denominator is k^4.
(2) $(-ie)(4i) = 4e$; separate the p and k integrals.
(3) $p_\pi e^{-ip\cdot x} = i\partial_\pi e^{-ip\cdot x}$, so the bracket is $i\partial_\pi A_\alpha(x)$.
(4) $\epsilon^{\kappa\alpha\pi\mu}$ is antisymmetric in $\alpha\pi$, so only the antisymmetric part of $\partial_\pi A_\alpha$ survives: $\epsilon^{\kappa\alpha\pi\mu}\partial_\pi A_\alpha = \frac{1}{2}\epsilon^{\kappa\alpha\pi\mu}(\partial_\pi A_\alpha - \partial_\alpha A_\pi) = \frac{1}{2}\epsilon^{\kappa\alpha\pi\mu}F_{\pi\alpha}$.

The answer depends only on F, as gauge invariance demands.

Step 4: the Fourier integral. Write $I_\kappa = -i\,\partial_{\varepsilon^\kappa} J(\varepsilon)$ with $J(\varepsilon) \coloneqq \int_k e^{ik\cdot\varepsilon}/(k^2+i0)^2$, since $\partial_{\varepsilon^\kappa} e^{ik\cdot\varepsilon} = ik_\kappa e^{ik\cdot\varepsilon}$. Take ε spacelike (e.g. purely spatial), so that $k\cdot\varepsilon = -k_E\cdot\varepsilon_E$ after Wick rotation and $\varepsilon^2 = -\varepsilon_E^2$. Then

$$\begin{aligned} J(\varepsilon) &\overset{(1)}{=} i\int \frac{\mathrm{d}^4k_E}{(2\pi)^4}\,\frac{e^{-ik_E\cdot\varepsilon_E}}{k_E^4} \overset{(2)}{=} i\,G_E(\varepsilon_E), \qquad G_E(x) \coloneqq \int \frac{\mathrm{d}^4k_E}{(2\pi)^4}\,\frac{e^{ik_E\cdot x}}{k_E^4} \\ &\overset{(3)}{=} -\frac{i}{8\pi^2}\ln\big(\mu|\varepsilon_E|\big) \overset{(4)}{=} -\frac{i}{16\pi^2}\ln\big(-\mu^2\varepsilon^2\big). \end{aligned} \tag{10.72}$$

(1) Wick rotation (10.39): $\mathrm{d}^4k = i\,\mathrm{d}^4k_E$ and $(k^2)^2 = (-k_E^2)^2 = k_E^4$.
(2) $k_E \to -k_E$ flips the sign in the exponent.
(3) G_E solves $(\nabla^2)^2 G_E = \delta^4$ (apply $(\nabla^2)^2$ under the integral: $(\nabla^2)^2 e^{ik\cdot x} = k^4 e^{ik\cdot x}$). In four dimensions, with $r = |x|$, $\nabla^2\ln r = \frac{1}{r^3}\partial_r(r^3\cdot\frac{1}{r}) = \frac{2}{r^2}$ and $\nabla^2\frac{1}{r^2} = -4\pi^2\delta^4$ (Step 1). Hence $(\nabla^2)^2\big(-\frac{1}{8\pi^2}\ln r\big) = -\frac{2}{8\pi^2}(-4\pi^2)\delta^4 = \delta^4$. The constant μ is an infrared scale (the integral diverges logarithmically at small k); it drops out of derivatives.
(4) $\ln|\varepsilon_E| = \frac{1}{2}\ln\varepsilon_E^2$ and $\varepsilon_E^2 = -\varepsilon^2$.

Differentiating,

$$I_\kappa(\varepsilon) \overset{(1)}{=} -i\,\partial_{\varepsilon^\kappa}\Big(-\frac{i}{16\pi^2}\ln(-\mu^2\varepsilon^2)\Big) \overset{(2)}{=} -\frac{1}{16\pi^2}\,\frac{2\varepsilon_\kappa}{\varepsilon^2} = -\frac{1}{8\pi^2}\,\frac{\varepsilon_\kappa}{\varepsilon^2}. \tag{10.73}$$

(1) $I_\kappa = -i\partial_\kappa J$ with (10.72).
(2) $(-i)(-i) = -1$ and $\partial_\kappa\ln(-\varepsilon^2) = 2\varepsilon_\kappa/\varepsilon^2$.

Inserting (10.73) in (10.71),

$$B^\mu \overset{(1)}{=} -\frac{ie}{4\pi^2}\,\epsilon^{\kappa\alpha\pi\mu}F_{\pi\alpha}\,\frac{\varepsilon_\kappa}{\varepsilon^2} \overset{(2)}{=} -\frac{ie}{4\pi^2}\,\epsilon^{\mu\sigma\rho\tau}F_{\sigma\rho}\,\frac{\varepsilon_\tau}{\varepsilon^2} \qquad \text{[QFT notes (22.3.12)]}. \tag{10.74}$$

(1) $2ie\cdot(-\frac{1}{8\pi^2}) = -\frac{ie}{4\pi^2}$.
(2) Rename $\pi\to\sigma$, $\alpha\to\rho$, $\kappa\to\tau$; then $\epsilon^{\tau\rho\sigma\mu} = \epsilon^{\mu\sigma\rho\tau}$, because reversing the order of four indices is two transpositions (swap the outer pair and the inner pair).

It is singular like $1/\varepsilon$, one power less than the free bilinear, but it is not zero.

Step 5: the anomaly. Now

$$\partial_\mu\langle j_5^\mu\rangle \overset{(1)}{=} \text{s-lim}\big(-ie\,\varepsilon^\nu F_{\mu\nu}\big)\Big(-\frac{ie}{4\pi^2}\epsilon^{\mu\sigma\rho\tau}F_{\sigma\rho}\frac{\varepsilon_\tau}{\varepsilon^2}\Big) \overset{(2)}{=} -\frac{e^2}{4\pi^2}\,\epsilon^{\mu\sigma\rho\tau}F_{\mu\nu}F_{\sigma\rho}\,\text{s-lim}\,\frac{\varepsilon^\nu\varepsilon_\tau}{\varepsilon^2}$$
$$\overset{(3)}{=} -\frac{e^2}{4\pi^2}\,\epsilon^{\mu\sigma\rho\tau}F_{\mu\nu}F_{\sigma\rho}\,\frac{\delta^\nu_\tau}{4} \overset{(4)}{=} -\frac{e^2}{16\pi^2}\,\epsilon^{\mu\sigma\rho\nu}F_{\mu\nu}F_{\sigma\rho} \overset{(5)}{=} -\frac{e^2}{16\pi^2}\,\epsilon^{\mu\nu\sigma\rho}F_{\mu\nu}F_{\sigma\rho}. \tag{10.75}$$

(1) Expectation value of (10.66), with the bilinear (10.74).
(2) $(-ie)(-ie) = i^2e^2 = -e^2$.
(3) The symmetric limit (10.62) in $d = 4$.
(4) δ^ν_τ sets $\tau = \nu$.
(5) $(\sigma\rho\nu) \to (\nu\sigma\rho)$ is a cyclic permutation of three indices, which is even.

This is the **Adler–Bell–Jackiw anomaly** in Minkowski form [QFT notes (22.3.15)]. To see what it means, write it with $\mathbf{E}$ and $\mathbf{B}$. With $A^\mu = (\phi, \mathbf{A})$, $A_i = -A^i$, $\mathbf{E} = -\nabla\phi - \partial_t\mathbf{A}$ and $\mathbf{B} = \nabla\times\mathbf{A}$, the field strength is $F_{0i} = \partial_0 A_i - \partial_i A_0 = E_i$ and $F_{ij} = -(\partial_i A^j - \partial_j A^i) = -\epsilon_{ijk}B_k$. Then

$$\epsilon^{\mu\nu\sigma\rho}F_{\mu\nu}F_{\sigma\rho} \overset{(1)}{=} 4\,\epsilon^{0ijk}F_{0i}F_{jk} \overset{(2)}{=} 4\,\epsilon_{ijk}E_i\big(-\epsilon_{jkl}B_l\big) \overset{(3)}{=} -4\cdot 2\delta_{il}E_iB_l = -8\,\mathbf{E}\cdot\mathbf{B}, \tag{10.76}$$

(1) Exactly one of the four indices is 0. It can sit in either slot of either pair (4 choices), and each choice gives the same term after using the antisymmetry of both ϵ and F.
(2) $\epsilon^{0ijk} = \epsilon_{ijk}$ (with $\epsilon^{0123} = +1 = \epsilon_{123}$), and the field strengths above.
(3) $\epsilon_{ijk}\epsilon_{ljk} = 2\delta_{il}$.

so the anomaly equation reads

$$\partial_\mu\langle j_5^\mu\rangle = \frac{e^2}{2\pi^2}\,\mathbf{E}\cdot\mathbf{B}. \tag{10.77}$$

Parallel electric and magnetic fields create chirality. This is the four-dimensional version of Figure 10.8: the magnetic field organises the fermions into Landau levels, the lowest of which is a one-dimensional chiral fermion moving along $\mathbf{B}$ with $eB/2\pi$ states per unit transverse area, and the electric field then pumps it exactly as in two dimensions, $\dot n_5 = 2\cdot\frac{eE}{2\pi}\cdot\frac{eB}{2\pi} = \frac{e^2}{2\pi^2}EB$.

10.3.7 Four dimensions: the triangle graph

Nash's "direct perturbative way" is to expand (10.27) in powers of the coupling. At one loop, the matrix element of j_5^λ between the vacuum and a two-photon state is the sum of two **triangle graphs** (Figure 10.10). They are linearly divergent, and that is the whole story.

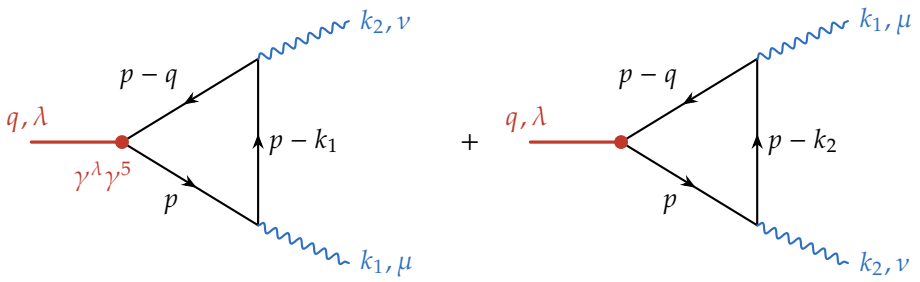


Figure 10.10: The two triangle graphs for $\langle k_1, k_2|j_5^\lambda|0\rangle$. The axial vertex (red) injects momentum $q = k_1 + k_2$; the two photons (blue) leave. The second graph is the first with the photons exchanged. Each graph is linearly divergent: three propagators $\sim 1/p^3$ against d^4p.

The amplitude. Let $|k_1,\epsilon_1;k_2,\epsilon_2\rangle$ be a state of two photons with momenta k_1,k_2 and polarisation vectors ϵ_1,ϵ_2, and write $\langle k_1,\epsilon_1;k_2,\epsilon_2|j_5^\lambda(0)|0\rangle = \epsilon_{1\mu}^*\epsilon_{2\nu}^*\,\mathcal{M}^{\mu\nu\lambda}$. The Feynman rules are as in Example 10.10: the operator $j_5^\lambda=\bar\psi\gamma^\lambda\gamma^5\psi$ contributes the matrix $\gamma^\lambda\gamma^5$ (no coupling constant: it is inserted, not exponentiated), each photon a vertex $-ie\gamma$, each fermion line a propagator $i/\not p$, and the loop a factor -1 and a trace. The fermion loop can run round the triangle in two orientations, which are the two graphs of Figure 10.10. Reading each trace against the fermion arrows,

$$\begin{aligned}\mathcal{M}^{\mu\nu\lambda} &\overset{(1)}{=} (-1)(-ie)^2\,i^3\int\frac{\mathrm{d}^4p}{(2\pi)^4}\Big\{\mathrm{tr}\big[S(p)\gamma^\lambda\gamma^5S(p-q)\gamma^\nu S(p-k_1)\gamma^\mu\big]+(k_1,\mu\leftrightarrow k_2,\nu)\Big\}\\ &\overset{(2)}{=} ie^2\,T^{\mu\nu\lambda}(k_1,k_2),\end{aligned}\tag{10.78}$$

(1) Here $S(p):=1/\not p=\not p/p^2$ (Tool 1(a)), so each propagator is iS. In the first graph the fermion with momentum p enters the μ vertex and leaves with $p-k_1$ (photon k_1 emitted), then leaves the ν vertex with $p-k_1-k_2=p-q$, and the axial vertex injects $q=k_1+k_2$ to bring it back to p.
(2) $(-1)(-e^2)(-i)=-ie^2$; we define T with the opposite sign, below.

where, with the couplings stripped off [QFT notes (22.5.2)],

$$\begin{aligned}T^{\mu\nu\lambda}(k_1,k_2):=-\int\frac{\mathrm{d}^4p}{(2\pi)^4}\Big\{&\mathrm{tr}\big[S(p)\gamma^\lambda\gamma^5S(p-q)\gamma^\nu S(p-k_1)\gamma^\mu\big]\\&+\mathrm{tr}\big[S(p)\gamma^\lambda\gamma^5S(p-q)\gamma^\mu S(p-k_2)\gamma^\nu\big]\Big\}.\end{aligned}\tag{10.79}$$

Since $\langle f|O(x)|0\rangle=e^{i(k_1+k_2)\cdot x}\langle f|O(0)|0\rangle$ (translate with $e^{iP\cdot x}$), a divergence ∂_λ becomes iq_λ. And gauge invariance of the photon couplings means that replacing a polarisation by its momentum, $\epsilon_1\to k_1$, gives zero. So the Ward identities we would like are

$$q_\lambda T^{\mu\nu\lambda}=0\quad\text{(axial)},\qquad k_{1\mu}T^{\mu\nu\lambda}=0=k_{2\nu}T^{\mu\nu\lambda}\quad\text{(vector)}.\tag{10.80}$$

Two algebraic identities. The first converts $\not q\gamma^5$ into a difference of inverse propagators:

$$\begin{aligned}S(p)\,\not q\gamma^5\,S(p-q)&\overset{(1)}{=}S(p)\big[\not p\gamma^5-(\not p-\not q)\gamma^5\big]S(p-q)\overset{(2)}{=}S(p)\big[\not p\gamma^5+\gamma^5(\not p-\not q)\big]S(p-q)\\&\overset{(3)}{=}\gamma^5S(p-q)+S(p)\gamma^5.\end{aligned}\tag{10.81}$$

(1) $\not q=\not p-(\not p-\not q)$.
(2) $\not\phi\gamma^5=-\gamma^5\not\phi$ for any vector v (Tool 1(e)).
(3) $S(p)\not p=\mathbf{1}$ and $(\not p-\not q)S(p-q)=\mathbf{1}$.

The second does the same for a photon vertex. From $S(u)\not u=\mathbf{1}=\not vS(v)$ we get $S(u)(\not v-\not u)S(v)=S(u)-S(v)$ for any u,v, and two instances of it are

$$S(p-k_1)\not k_1\,S(p)=S(p-k_1)-S(p),\qquad S(p-q)\not k_1\,S(p-k_2)=S(p-q)-S(p-k_2),\tag{10.82}$$

with $(u,v)=(p-k_1,p)$ in the first, where $v-u=k_1$, and $(u,v)=(p-q,p-k_2)$ in the second, where $v-u=q-k_2=k_1$.

The naive argument that both Ward identities hold. Contract (10.79) with q_λ and use (10.81) in each graph. Define

$$A_1(p):=\mathrm{tr}\big[\gamma^5S(p-q)\gamma^\nu S(p-k_1)\gamma^\mu\big],\qquad A_2(p):=\mathrm{tr}\big[\gamma^5S(p-q)\gamma^\mu S(p-k_2)\gamma^\nu\big].\tag{10.83}$$

Then

$$
\begin{aligned}
q_\lambda T^{\mu\nu\lambda} &\overset{(1)}{=} -\int \frac{\mathrm{d}^4p}{(2\pi)^4}\Big\{A_1(p) + \mathrm{tr}\big[S(p)\gamma^5\gamma^\nu S(p-k_1)\gamma^\mu\big] + A_2(p) + \mathrm{tr}\big[S(p)\gamma^5\gamma^\mu S(p-k_2)\gamma^\nu\big]\Big\} \\
&\overset{(2)}{=} -\int \frac{\mathrm{d}^4p}{(2\pi)^4}\Big\{A_1(p) - A_2(p+k_2) + A_2(p) - A_1(p+k_1)\Big\} \\
&\overset{(3)}{=} \int \frac{\mathrm{d}^4p}{(2\pi)^4}\big[A_1(p+k_1) - A_1(p)\big] + \int \frac{\mathrm{d}^4p}{(2\pi)^4}\big[A_2(p+k_2) - A_2(p)\big]. \qquad (10.84)
\end{aligned}
$$

(1) (10.81), with the $\gamma^5 S(p-q)$ term giving $A_{1,2}$ directly.
(2) For the fourth term: by cyclicity it is $\mathrm{tr}[\gamma^5\gamma^\mu S(p-k_2)\gamma^\nu S(p)]$, whose numerator is $-4i\epsilon^{\mu\alpha\nu\beta}(p-k_2)_\alpha p_\beta$ by (10.36). Also $A_1(p+k_1) = \mathrm{tr}[\gamma^5 S(p-k_2)\gamma^\nu S(p)\gamma^\mu]$ has numerator $-4i\epsilon^{\alpha\nu\beta\mu}(p-k_2)_\alpha p_\beta$ and the same denominator $(p-k_2)^2p^2$. Since $(\mu\alpha\nu\beta) \to (\alpha\nu\beta\mu)$ is a cyclic shift of four slots (odd), $\epsilon^{\mu\alpha\nu\beta} = -\epsilon^{\alpha\nu\beta\mu}$: the fourth term is $-A_1(p+k_1)$. The second term is $-A_2(p+k_2)$ in the same way.
(3) Regroup.

If we could shift $p \to p - k_1$ in $\int A_1(p+k_1)$, the first bracket would vanish, and likewise the second: $q_\lambda T = 0$. The vector identity works the same way:

$$
\begin{aligned}
k_{1\mu} T^{\mu\nu\lambda} &\overset{(1)}{=} -\int \frac{\mathrm{d}^4p}{(2\pi)^4}\Big\{\mathrm{tr}\big[\gamma^\lambda\gamma^5 S(p-q)\gamma^\nu S(p-k_1)\big] - \mathrm{tr}\big[\gamma^\lambda\gamma^5 S(p-q)\gamma^\nu S(p)\big] \\
&\qquad\qquad + \mathrm{tr}\big[S(p)\gamma^\lambda\gamma^5 S(p-q)\gamma^\nu\big] - \mathrm{tr}\big[S(p)\gamma^\lambda\gamma^5 S(p-k_2)\gamma^\nu\big]\Big\} \\
&\overset{(2)}{=} -\int \frac{\mathrm{d}^4p}{(2\pi)^4}\Big\{\mathrm{tr}\big[\gamma^\lambda\gamma^5 S(p-q)\gamma^\nu S(p-k_1)\big] - \mathrm{tr}\big[S(p)\gamma^\lambda\gamma^5 S(p-k_2)\gamma^\nu\big]\Big\} \\
&\overset{(3)}{=} -\int \frac{\mathrm{d}^4p}{(2\pi)^4}\big[B(p-k_1) - B(p)\big], \qquad B(p) \coloneqq \mathrm{tr}\big[S(p)\gamma^\lambda\gamma^5 S(p-k_2)\gamma^\nu\big]. \qquad (10.85)
\end{aligned}
$$

(1) First graph: cycle $S(p)$ to the end and use the first identity of (10.82). Second graph: the second identity of (10.82).
(2) The second and third terms are equal by cyclicity and cancel.
(3) $B(p-k_1) = \mathrm{tr}[S(p-k_1)\gamma^\lambda\gamma^5 S(p-q)\gamma^\nu]$, which is the first term after cycling $S(p-k_1)$ to the end.

Again a shift $p \to p + k_1$ would give zero. So: *if shifts were allowed, there would be no anomaly* [QFT notes (22.3.20)–(22.3.24)].

Shifting a linearly divergent integral. In one dimension, for an integrand with $f(\pm\infty)$ finite,

$$
\begin{aligned}
\int_{-\infty}^{\infty}\big[f(x+a) - f(x)\big]\mathrm{d}x &\overset{(1)}{=} \int_{-\infty}^{\infty}\big[af'(x) + \tfrac{1}{2}a^2 f''(x) + \cdots\big]\mathrm{d}x \\
&\overset{(2)}{=} a\big[f(+\infty) - f(-\infty)\big]. \qquad (10.86)
\end{aligned}
$$

(1) Taylor expand.
(2) The fundamental theorem of calculus; $f'(\pm\infty) = 0$ kills the higher terms.

If $\int f$ converges, $f(\pm\infty) = 0$ and the shift is harmless; if $\int f$ diverges linearly, $f(\pm\infty)$ can be nonzero and the shift costs a finite *surface term*. In four dimensions the same Taylor expansion plus Gauss's theorem on a large sphere give the formula we need [QFT notes (22.5.9)–(22.5.11)].

Suppose $f(p) \sim |p|^{-3}$ at large p (linear divergence). Then

$$\begin{aligned}\int \mathrm{d}^4p\,\big[f(p+a)-f(p)\big] &\overset{(1)}{=} \int \mathrm{d}^4p\,\big[a^\tau\partial_\tau f(p)+\tfrac{1}{2}a^\tau a^\sigma\partial_\tau\partial_\sigma f(p)+\cdots\big] \overset{(2)}{=} a^\tau\int \mathrm{d}^4p\,\partial_\tau f(p)\\ &\overset{(3)}{=} i\,a_E^\tau\int_{|p_E|<P}\mathrm{d}^4p_E\,\partial_\tau^E f \overset{(4)}{=} i\,a_E^\tau\oint_{|p_E|=P}\mathrm{d}S\,\frac{p_{E\tau}}{P}\,f \overset{(5)}{=} 2i\pi^2\,a_E^\tau\big\langle P^2 p_{E\tau} f\big\rangle_{|p_E|=P}\\ &\overset{(6)}{=} 2i\pi^2\lim_{P\to\infty}\big\langle p^2\,(a\cdot p)\,f(p)\big\rangle_{\text{angles}}. \end{aligned} \tag{10.87}$$

(1) Taylor expansion in a.
(2) The second-order term is a total derivative of $\partial f \sim P^{-4}$, whose surface integral $\sim P^3\cdot P^{-4}\to 0$; the higher ones fall off faster still.
(3) Wick rotate (10.39) ($\mathrm{d}^4p = i\,\mathrm{d}^4p_E$; the scalar $a^\tau\partial_\tau$ becomes $a_E^\tau\partial_\tau^E$), and cut the integral off at $|p_E| = P$, to be sent to infinity.
(4) Gauss's theorem in $\mathbb{R}^4$: the outward unit normal on the sphere is p_E/P.
(5) The 3-sphere of radius P has area $2\pi^2P^3$; $\langle\cdots\rangle$ is the average over the sphere.
(6) Back to Minkowski components: $a_E\cdot p_E = -a\cdot p$ and $P^2 = p_E^2 = -p^2$; the two signs cancel.

We also need the angular average $\langle p_\tau p_\alpha\rangle = \frac{1}{4}\eta_{\tau\alpha}p^2$, which is (10.42) at $d = 4$ (four equivalent directions). Combining, for any integrand whose leading term is $f(p)\to C^\alpha p_\alpha/p^4$,

$$\int\frac{\mathrm{d}^4p}{(2\pi)^4}\big[f(p+a)-f(p)\big] = \frac{2i\pi^2}{16\pi^4}\,a^\tau C^\alpha\left\langle\frac{p_\tau p_\alpha}{p^2}\right\rangle = \frac{i}{32\pi^2}\,a\cdot C. \tag{10.88}$$

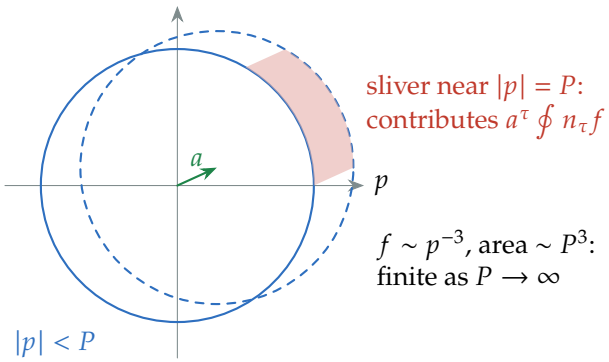


Figure 10.11: Why shifting a linearly divergent integral is not free. The integral over a ball of radius P and over the same ball shifted by a differ by the integral over a thin crescent near the surface. Its thickness is $\sim a$, its area $\sim P^3$, and the integrand $\sim P^{-3}$: the difference stays finite as $P\to\infty$. For a log-divergent or convergent integrand it would vanish.

Step 1: the axial Ward identity with the routing of (10.79). We need the leading large-p behaviour of A_1. Its denominator is $(p-q)^2(p-k_1)^2\to p^4$. Its numerator is

$$\begin{aligned}\mathrm{tr}\big[\gamma^5(\not p-\not q)\gamma^\nu(\not p-\not k_1)\gamma^\mu\big] &\overset{(1)}{=} -4i\,\epsilon^{\alpha\nu\beta\mu}(p-q)_\alpha(p-k_1)_\beta\\ &\overset{(2)}{=} -4i\,\epsilon^{\alpha\nu\beta\mu}\big[p_\alpha p_\beta - q_\alpha p_\beta - p_\alpha k_{1\beta} + q_\alpha k_{1\beta}\big]\\ &\overset{(3)}{=} -4i\,\epsilon^{\alpha\nu\beta\mu}\big[-q_\alpha p_\beta + k_{1\alpha}p_\beta\big] + (\text{const}) \overset{(4)}{=} 4i\,\epsilon^{\alpha\nu\beta\mu}k_{2\alpha}p_\beta + (\text{const}).\end{aligned} \tag{10.89}$$

(1) The γ^5 trace (10.36).

(2) Multiply out.
(3) $\epsilon^{\alpha\nu\beta\mu}p_\alpha p_\beta = 0$ (antisymmetric against symmetric). In the third term relabel $\alpha \leftrightarrow \beta$ and use $\epsilon^{\beta\nu\alpha\mu} = -\epsilon^{\alpha\nu\beta\mu}$. The constant term gives $A_1 \sim p^{-4}$, no surface term.
(4) $q - k_1 = k_2$.

So $A_1 \to C_1^\beta p_\beta/p^4$ with $C_1^\beta = 4i\epsilon^{\alpha\nu\beta\mu}k_{2\alpha}$. Then

$$\int \frac{\mathrm{d}^4p}{(2\pi)^4}\big[A_1(p+k_1) - A_1(p)\big] \overset{(1)}{=} \frac{i}{32\pi^2}\,k_{1\beta}\,4i\,\epsilon^{\alpha\nu\beta\mu}k_{2\alpha} \overset{(2)}{=} -\frac{1}{8\pi^2}\,\epsilon^{\alpha\nu\beta\mu}k_{2\alpha}k_{1\beta} \overset{(3)}{=} \frac{1}{8\pi^2}\,\epsilon^{\mu\nu\rho\sigma}k_{1\rho}k_{2\sigma}. \tag{10.90}$$

(1) The surface rule (10.88) with $a = k_1$.
(2) $\frac{i\cdot 4i}{32\pi^2} = -\frac{1}{8\pi^2}$.
(3) Rename $\alpha \to \sigma$, $\beta \to \rho$: $\epsilon^{\sigma\nu\rho\mu} = -\epsilon^{\mu\nu\rho\sigma}$ (swap the first and last slots).

The second graph is the first with $(k_1, \mu) \leftrightarrow (k_2, \nu)$, so its surface term is $-\frac{1}{8\pi^2}\epsilon^{\alpha\mu\beta\nu}k_{1\alpha}k_{2\beta} = -\frac{1}{8\pi^2}\epsilon^{\rho\mu\sigma\nu}k_{1\rho}k_{2\sigma} = +\frac{1}{8\pi^2}\epsilon^{\mu\nu\rho\sigma}k_{1\rho}k_{2\sigma}$ (three transpositions: $\rho\mu\sigma\nu \to \mu\rho\sigma\nu \to \mu\nu\sigma\rho \to \mu\nu\rho\sigma$). By (10.84),

$$q_\lambda T^{\mu\nu\lambda}\big|_{\text{routing of (10.79)}} = \frac{1}{8\pi^2}\epsilon^{\mu\nu\rho\sigma}k_{1\rho}k_{2\sigma} + \frac{1}{8\pi^2}\epsilon^{\mu\nu\rho\sigma}k_{1\rho}k_{2\sigma} = \frac{1}{4\pi^2}\,\epsilon^{\mu\nu\rho\sigma}k_{1\rho}k_{2\sigma}. \tag{10.91}$$

So with this routing the axial identity already fails. But we have not yet checked the vector identities, and the routing was arbitrary.

Step 2: the routing ambiguity. Since each graph is linearly divergent, the amplitude depends on how we label the loop momentum. Shift $p \to p + a$ in the first graph and, to preserve the symmetry between the photons, $p \to p + a'$ in the second, with a' obtained from a by $k_1 \leftrightarrow k_2$. A shift must be built from the external momenta, so the most general choice is $a = \alpha k_1 + (\alpha - \beta)k_2$ for two numbers α, β [QFT notes (22.5.12)]. Write $T(\beta)$ for the amplitude with the shifted routing. We need the leading large-p integrand of the first graph, including the minus sign in (10.79):

$$-\frac{\mathrm{tr}[\not p\gamma^\lambda\gamma^5\not p\gamma^\nu\not p\gamma^\mu]}{p^6} \overset{(1)}{=} \frac{\mathrm{tr}[\not p\gamma^\lambda\not p\,\gamma^5\gamma^\nu\not p\gamma^\mu]}{p^6} \overset{(2)}{=} \frac{2p^\lambda\,\mathrm{tr}[\not p\gamma^5\gamma^\nu\not p\gamma^\mu] - p^2\,\mathrm{tr}[\gamma^\lambda\gamma^5\gamma^\nu\not p\gamma^\mu]}{p^6}$$
$$\overset{(3)}{=} \frac{0 - p^2\,(-4i)\,\epsilon^{\nu\alpha\mu\lambda}p_\alpha}{p^6} \overset{(4)}{=} -4i\,\frac{\epsilon^{\lambda\nu\alpha\mu}p_\alpha}{p^4}. \tag{10.92}$$

(1) Move γ^5 to the right past $\not p$: one sign.
(2) The sandwich identity (10.34), $\not p\gamma^\lambda\not p = 2p^\lambda\not p - p^2\gamma^\lambda$.
(3) Cycle γ^5 to the front and use (10.36): the first trace is $-4i\epsilon^{\nu\alpha\mu\beta}p_\alpha p_\beta = 0$; the second is $\mathrm{tr}[\gamma^5\gamma^\nu\not p\gamma^\mu\gamma^\lambda] = -4i\epsilon^{\nu\alpha\mu\lambda}p_\alpha$.
(4) $\epsilon^{\nu\alpha\mu\lambda} = -\epsilon^{\lambda\nu\alpha\mu}$ (a cyclic shift of four slots is odd).

The second graph's leading integrand is the same with $\mu \leftrightarrow \nu$, i.e. $+4i\epsilon^{\lambda\nu\alpha\mu}p_\alpha/p^4$. Hence, by the surface rule (10.88),

$$T^{\mu\nu\lambda}(\beta) - T^{\mu\nu\lambda}(0) \overset{(1)}{=} \frac{i}{32\pi^2}(-4i)\,\epsilon^{\lambda\nu\alpha\mu}a_\alpha + \frac{i}{32\pi^2}(+4i)\,\epsilon^{\lambda\nu\alpha\mu}a'_\alpha \overset{(2)}{=} \frac{1}{8\pi^2}\,\epsilon^{\lambda\nu\alpha\mu}(a-a')_\alpha$$
$$\overset{(3)}{=} \frac{\beta}{8\pi^2}\,\epsilon^{\lambda\nu\alpha\mu}(k_1-k_2)_\alpha \overset{(4)}{=} \frac{\beta}{8\pi^2}\,\epsilon^{\mu\nu\lambda\alpha}(k_1-k_2)_\alpha. \tag{10.93}$$

(1) The surface rule for each graph, with $C^\alpha = \mp 4i\epsilon^{\lambda\nu\alpha\mu}$ and shifts a, a'.
(2) $\frac{i(-4i)}{32\pi^2} = \frac{1}{8\pi^2}$.

(3) $a - a' = \alpha k_1 + (\alpha - \beta)k_2 - \alpha k_2 - (\alpha - \beta)k_1 = \beta(k_1 - k_2)$: the α-dependence (a common shift of both graphs by αq) cancels.
(4) $(\lambda\nu\alpha\mu) \to (\mu\nu\alpha\lambda) \to (\mu\nu\lambda\alpha)$: two transpositions.

The triangle is defined only up to this one-parameter family: the momentum routing is not physics, so β must be fixed by a physical requirement.

Step 3: the vector Ward identity at $\beta = 0$. From (10.85) we need the leading behaviour of $B(p) = \mathrm{tr}[S(p)\gamma^\lambda\gamma^5 S(p - k_2)\gamma^\nu]$. Its numerator is

$$\mathrm{tr}\big[\not p\gamma^\lambda\gamma^5(\not p - \not k_2)\gamma^\nu\big] \overset{(1)}{=} \mathrm{tr}\big[\gamma^5(\not p - \not k_2)\gamma^\nu\not p\gamma^\lambda\big] \overset{(2)}{=} -4i\,\epsilon^{\alpha\nu\beta\lambda}(p - k_2)_\alpha p_\beta \overset{(3)}{=} 4i\,\epsilon^{\alpha\nu\beta\lambda}k_{2\alpha}p_\beta,$$

(1) Cyclicity, moving $\not p\gamma^\lambda$ to the end.
(2) (10.36).
(3) $\epsilon^{\alpha\nu\beta\lambda}p_\alpha p_\beta = 0$.

over $p^2(p - k_2)^2 \to p^4$. So $C^\beta = 4i\epsilon^{\alpha\nu\beta\lambda}k_{2\alpha}$ and the shift in (10.85) is $a = -k_1$:

$$k_{1\mu}T^{\mu\nu\lambda}(0) \overset{(1)}{=} -\frac{i}{32\pi^2}(-k_{1\beta})\,4i\,\epsilon^{\alpha\nu\beta\lambda}k_{2\alpha} \overset{(2)}{=} -\frac{1}{8\pi^2}\,\epsilon^{\alpha\nu\beta\lambda}k_{2\alpha}k_{1\beta} \overset{(3)}{=} -\frac{1}{8\pi^2}\,\epsilon^{\nu\lambda\rho\sigma}k_{1\rho}k_{2\sigma}. \tag{10.94}$$

(1) (10.85) (note its overall minus sign) with the surface rule.
(2) $-\frac{i\cdot(-1)\cdot 4i}{32\pi^2} = \frac{4i^2}{32\pi^2} = -\frac{1}{8\pi^2}$.
(3) Rename $\alpha \to \sigma$, $\beta \to \rho$; $(\sigma\nu\rho\lambda) \to (\nu\sigma\rho\lambda) \to (\nu\lambda\rho\sigma)$ is two transpositions, so $\epsilon^{\sigma\nu\rho\lambda} = \epsilon^{\nu\lambda\rho\sigma}$.

The vector identity also fails with the routing of (10.79).

Step 4: the anomaly budget. Contract the routing change (10.93) with q_λ and with $k_{1\mu}$:

$$\begin{aligned}
q_\lambda\big[T(\beta) - T(0)\big]^{\mu\nu\lambda} &\overset{(1)}{=} \frac{\beta}{8\pi^2}\,\epsilon^{\mu\nu\lambda\alpha}(k_1 + k_2)_\lambda(k_1 - k_2)_\alpha \\
&\overset{(2)}{=} \frac{\beta}{8\pi^2}\big(-\epsilon^{\mu\nu\lambda\alpha}k_{1\lambda}k_{2\alpha} + \epsilon^{\mu\nu\lambda\alpha}k_{2\lambda}k_{1\alpha}\big) \overset{(3)}{=} -\frac{\beta}{4\pi^2}\,\epsilon^{\mu\nu\rho\sigma}k_{1\rho}k_{2\sigma}, \\
k_{1\mu}\big[T(\beta) - T(0)\big]^{\mu\nu\lambda} &\overset{(4)}{=} \frac{\beta}{8\pi^2}\,\epsilon^{\mu\nu\lambda\alpha}k_{1\mu}(k_1 - k_2)_\alpha \\
&\overset{(5)}{=} -\frac{\beta}{8\pi^2}\,\epsilon^{\mu\nu\lambda\alpha}k_{1\mu}k_{2\alpha} \overset{(6)}{=} -\frac{\beta}{8\pi^2}\,\epsilon^{\nu\lambda\rho\sigma}k_{1\rho}k_{2\sigma}.
\end{aligned} \tag{10.95}$$

(1) $q = k_1 + k_2$.
(2) Multiply out; ϵkk with the same k twice vanishes.
(3) In the second term relabel $\lambda \leftrightarrow \alpha$ and use antisymmetry; it equals the first. Then rename $\lambda, \alpha \to \rho, \sigma$.
(4) $k_{1\mu}$ times (10.93).
(5) $\epsilon^{\mu\nu\lambda\alpha}k_{1\mu}k_{1\alpha} = 0$.
(6) Rename $\mu \to \rho$, $\alpha \to \sigma$: $\epsilon^{\rho\nu\lambda\sigma} = \epsilon^{\nu\lambda\rho\sigma}$ (a cyclic permutation of the first three slots, which is even).

Adding these to (10.91) and (10.94):

$$\begin{aligned}
q_\lambda T^{\mu\nu\lambda}(\beta) &= \frac{1 - \beta}{4\pi^2}\,\epsilon^{\mu\nu\rho\sigma}k_{1\rho}k_{2\sigma}, \\
k_{1\mu}T^{\mu\nu\lambda}(\beta) &= -\frac{1 + \beta}{8\pi^2}\,\epsilon^{\nu\lambda\rho\sigma}k_{1\rho}k_{2\sigma},
\end{aligned} \tag{10.96}$$

and $k_{2\nu}T^{\mu\nu\lambda}(\beta)$ follows from the second line by the photon exchange symmetry. The QFT notes obtain the same structure, $(1-\beta)$ and $(1+\beta)$, [QFT notes (22.5.21), (22.5.26)]. The axial identity needs $\beta=+1$; the vector identities need $\beta=-1$. *No routing satisfies both.*

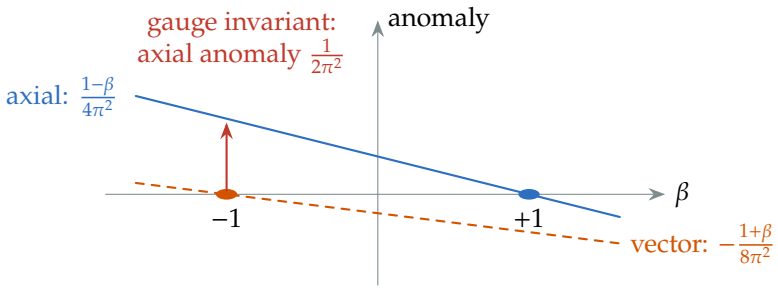


Figure 10.12: The anomaly budget of the triangle graph (in units of ϵkk). Changing the momentum routing β moves the anomaly between the axial current (solid) and the vector current (dashed). The two lines never vanish at the same β. Gauge invariance of electromagnetism forces $\beta=-1$ (orange dot), and the whole anomaly lands in the axial current.

At $\beta=-1$ the vector currents are conserved and

$$q_\lambda T^{\mu\nu\lambda}=\frac{1-(-1)}{4\pi^2}\,\epsilon^{\mu\nu\rho\sigma}k_{1\rho}k_{2\sigma}=\frac{1}{2\pi^2}\,\epsilon^{\mu\nu\rho\sigma}k_{1\rho}k_{2\sigma}. \tag{10.97}$$

Step 5: from the amplitude to the operator equation. Abbreviate $X\coloneqq\epsilon^{\mu\nu\rho\sigma}\epsilon^*_{1\mu}\epsilon^*_{2\nu}k_{1\rho}k_{2\sigma}$. The left side of the anomaly equation, between the two-photon state and the vacuum, is

$$\langle k_1,k_2|\partial_\lambda j^\lambda_5(0)|0\rangle\overset{(1)}{=}iq_\lambda\,\epsilon^*_{1\mu}\epsilon^*_{2\nu}\mathcal{M}^{\mu\nu\lambda}\overset{(2)}{=}i\cdot ie^2\,\epsilon^*_{1\mu}\epsilon^*_{2\nu}\,q_\lambda T^{\mu\nu\lambda}\overset{(3)}{=}-\frac{e^2}{2\pi^2}\,X. \tag{10.98}$$

(1) Translation invariance: $\partial_\lambda\to iq_\lambda$ (see below (10.79)).
(2) (10.78).
(3) $i^2=-1$ and (10.97).

For the right side we need the two-photon matrix element of ϵFF at lowest order. A free photon field obeys $\langle k,\epsilon|A_\beta(x)|0\rangle=\epsilon^*_\beta e^{ik\cdot x}$, so $\langle k,\epsilon|F_{\alpha\beta}(x)|0\rangle=i(k_\alpha\epsilon^*_\beta-k_\beta\epsilon^*_\alpha)e^{ik\cdot x}$. Then

$$\begin{aligned}\langle k_1,k_2|\epsilon^{\alpha\beta\gamma\delta}F_{\alpha\beta}F_{\gamma\delta}(0)|0\rangle&\overset{(1)}{=}2\,\epsilon^{\alpha\beta\gamma\delta}\,\langle k_1|F_{\alpha\beta}|0\rangle\langle k_2|F_{\gamma\delta}|0\rangle\overset{(2)}{=}2\,(2i)(2i)\,\epsilon^{\alpha\beta\gamma\delta}k_{1\alpha}\epsilon^*_{1\beta}k_{2\gamma}\epsilon^*_{2\delta}\\&\overset{(3)}{=}-8\,\epsilon^{\rho\mu\sigma\nu}k_{1\rho}\epsilon^*_{1\mu}k_{2\sigma}\epsilon^*_{2\nu}\overset{(4)}{=}+8X.\end{aligned} \tag{10.99}$$

(1) Either photon can be absorbed by either F (Wick's theorem for the free field), and the two assignments are equal because $\epsilon^{\alpha\beta\gamma\delta}=\epsilon^{\gamma\delta\alpha\beta}$.
(2) $\epsilon^{\alpha\beta\gamma\delta}(k_\alpha\epsilon^*_\beta-k_\beta\epsilon^*_\alpha)=2\epsilon^{\alpha\beta\gamma\delta}k_\alpha\epsilon^*_\beta$ by antisymmetry, for each factor.
(3) Rename $\alpha\beta\gamma\delta\to\rho\mu\sigma\nu$.
(4) $\epsilon^{\rho\mu\sigma\nu}=-\epsilon^{\mu\nu\rho\sigma}$ (three transpositions, as above).

Comparing, $-\frac{e^2}{2\pi^2}X=-\frac{e^2}{16\pi^2}\cdot 8X$, so the triangle says

$$\partial_\lambda j^\lambda_5=-\frac{e^2}{16\pi^2}\,\epsilon^{\alpha\beta\gamma\delta}F_{\alpha\beta}F_{\gamma\delta}\qquad\text{[QFT notes (22.3.46), (22.5.28)]}, \tag{10.100}$$

exactly the point-splitting result (10.75), sign and all. Two completely different calculations, one in position space with a Wilson line, one in momentum space with a surface term, agree. A fermion mass adds the classical term $2im\bar\psi\gamma^5\psi$ to the right side and changes nothing else.

Worked example 10.11: The same anomaly in dimensional regularisation

Dimensional regularisation fixes the routing problem automatically, because shifts are legal in d dimensions (Tool 5). The calculation (10.85) then gives $k_{1\mu}T^{\mu\nu\lambda} = 0$ with no surface term: the vector current is conserved. The price is paid by γ^5, which is intrinsically four-dimensional.

The 't Hooft–Veltman γ^5 [QFT notes (22.3.25)]. Keep $\gamma^5 := i\gamma^0\gamma^1\gamma^2\gamma^3$ in d dimensions. It anticommutes with $\gamma^0, \dots, \gamma^3$ as before, but *commutes* with the extra matrices $\gamma^4, \dots, \gamma^{d-1}$, since moving one of those through four anticommuting matrices gives $(-1)^4 = +1$. Split every loop vector into its four-dimensional part and the rest, $p = \bar{p} + p_\perp$; the external momenta k_1, k_2, q and indices μ, ν, λ are four-dimensional. Then $\not{p}\gamma^5 = -\gamma^5\not{\bar{p}} + \gamma^5\not{p}_\perp$, i.e.

$$\not{p}\,\gamma^5 = -\gamma^5\not{p} + 2\gamma^5\not{p}_\perp, \tag{10.101}$$

and the γ^5 traces (10.36) still hold with only the four-dimensional components contributing (an extra-dimensional γ^a must occur an even number of times for a nonzero trace).

Step (a): the modified identity. Redo (10.81):

$$\begin{aligned} S(p)\not{q}\gamma^5 S(p-q) &\overset{(1)}{=} S(p)\big[\not{p}\gamma^5 - (\not{p}-\not{q})\gamma^5\big]S(p-q) \overset{(2)}{=} S(p)\big[\not{p}\gamma^5 + \gamma^5(\not{p}-\not{q}) - 2\gamma^5\not{p}_\perp\big]S(p-q) \\ &\overset{(3)}{=} \gamma^5 S(p-q) + S(p)\gamma^5 - 2S(p)\gamma^5\not{p}_\perp S(p-q). \end{aligned} \tag{10.102}$$

(1) $\not{q} = \not{p} - (\not{p} - \not{q})$.
(2) (10.101) for the vector $p - q$, whose extra-dimensional part is $p_\perp$ since $q_\perp = 0$.
(3) $S(p)\not{p} = \mathbf{1} = (\not{p} - \not{q})S(p - q)$.

The first two terms give (10.84), whose derivation used only cyclicity and (10.36); now the shifts are legal, so they vanish. What remains, with the -1 of (10.79), is

$$\begin{aligned} q_\lambda T^{\mu\nu\lambda} &= 2\int \frac{\mathrm{d}^d p}{(2\pi)^d}\,\frac{N_1}{p^2(p-q)^2(p-k_1)^2} + (k_1, \mu \leftrightarrow k_2, \nu), \\ N_1 &:= \operatorname{tr}\big[\not{p}\gamma^5\not{p}_\perp(\not{p}-\not{q})\gamma^\nu(\not{p}-\not{k}_1)\gamma^\mu\big]. \end{aligned} \tag{10.103}$$

Step (b): the numerator. Only terms with an even number of $\perp$ vectors survive, and at most two, since the γ^5 trace needs four four-dimensional gamma matrices and only γ^μ, γ^ν are guaranteed. Using $\gamma^5\not{p}_\perp = \not{p}_\perp\gamma^5$ and $\not{p}\not{p}_\perp = \not{\bar{p}}\not{p}_\perp + p_\perp^2$ (by (10.30)),

$$\begin{aligned} N_1 &\overset{(1)}{=} p_\perp^2\ \operatorname{tr}\big[\gamma^5(\not{\bar{p}}-\not{q})\gamma^\nu(\not{\bar{p}}-\not{k}_1)\gamma^\mu\big] + \operatorname{tr}\big[\not{\bar{p}}\not{p}_\perp\gamma^5(\not{p}-\not{q})\gamma^\nu(\not{p}-\not{k}_1)\gamma^\mu\big] \\ &\overset{(2)}{=} p_\perp^2\ \operatorname{tr}\big[\gamma^5(\not{\bar{p}}-\not{q})\gamma^\nu(\not{\bar{p}}-\not{k}_1)\gamma^\mu\big] + \operatorname{tr}\big[\not{\bar{p}}\not{p}_\perp\gamma^5\not{p}_\perp\gamma^\nu(\not{\bar{p}}-\not{k}_1)\gamma^\mu\big] \\ &\qquad + \operatorname{tr}\big[\not{\bar{p}}\not{p}_\perp\gamma^5(\not{\bar{p}}-\not{q})\gamma^\nu\not{p}_\perp\gamma^\mu\big] + (\text{odd in } p_\perp) \\ &\overset{(3)}{=} p_\perp^2\Big\{\operatorname{tr}\big[\gamma^5(\not{\bar{p}}-\not{q})\gamma^\nu(\not{\bar{p}}-\not{k}_1)\gamma^\mu\big] + \operatorname{tr}\big[\not{\bar{p}}\gamma^5\gamma^\nu(\not{\bar{p}}-\not{k}_1)\gamma^\mu\big] + \operatorname{tr}\big[\not{\bar{p}}\gamma^5(\not{\bar{p}}-\not{q})\gamma^\nu\gamma^\mu\big]\Big\} + \cdots \\ &\overset{(4)}{=} -4i\,p_\perp^2\Big\{\epsilon\big(\bar{p}-q, \nu, \bar{p}-k_1, \mu\big) + \epsilon\big(\nu, \bar{p}-k_1, \mu, \bar{p}\big) + \epsilon\big(\bar{p}-q, \nu, \mu, \bar{p}\big)\Big\} + \cdots. \end{aligned} \tag{10.104}$$

(1) $\not{p}\gamma^5\not{p}_\perp = \not{p}\not{p}_\perp\gamma^5 = (\not{\bar{p}}\not{p}_\perp + p_\perp^2)\gamma^5$. In the $p_\perp^2$ term a further $\perp$ part would make four, so only $\bar{p}$ remains there.
(2) In the second trace the lone $\not{p}_\perp$ needs a partner: take the $\perp$ part of either $\not{p} - \not{q}$ or $\not{p} - \not{k}_1$ (both equal $\not{p}_\perp$). Taking both gives three: odd.
(3) Second trace: $\not{p}_\perp\gamma^5\not{p}_\perp = \not{p}_\perp\not{p}_\perp\gamma^5 = p_\perp^2\gamma^5$. Third trace: move the second $\not{p}_\perp$ left past γ^ν (-1), $\not{\bar{p}} - \not{q}$ (-1) and γ^5 ($+1$), then $\not{p}_\perp\not{p}_\perp = p_\perp^2$.

(4) Cycle γ^5 to the front of each trace and use (10.36); $\epsilon(a,b,c,d)$ denotes $\epsilon^{\alpha\beta\gamma\delta}$ contracted with the four slots.

Step (c): Feynman parameters and the shift. With (10.38), $A=(p-k_1)^2$, $B=(p-q)^2$, $C=p^2$:

$$xA+yB+zC \overset{(1)}{=} p^2-2p\cdot(xk_1+yq)+xk_1^2+yq^2 \overset{(2)}{=} (p-c)^2-\Delta,$$
$$c:=xk_1+yq, \qquad \Delta:=c^2-xk_1^2-yq^2.$$

(1) Expand, using $x+y+z=1$ for the p^2 terms.
(2) Complete the square.

Shift $p=\ell+c$ (legal). Since c is four-dimensional, $p_\perp=\ell_\perp$ and $\bar p=\bar\ell+c$. In the braces of (10.104), terms linear in $\bar\ell$ vanish by symmetric integration, and terms quadratic in $\bar\ell$ give $\epsilon(\bar\ell,\cdot,\bar\ell,\cdot)\to\eta\cdot\epsilon=0$. So we may replace $\bar p\to c$. Write $c=a\,k_1+b\,k_2$ with $a=x+y$, $b=y$, and note that for $u=u_1k_1+u_2k_2$, $v=v_1k_1+v_2k_2$, $\epsilon(u,\nu,v,\mu)=(u_1v_2-u_2v_1)\,E_0$ with $E_0:=\epsilon(k_1,\nu,k_2,\mu)$. Then

$$\begin{aligned}
\epsilon(c-q,\nu,c-k_1,\mu) &\overset{(1)}{=} \big[(a-1)b-(b-1)(a-1)\big]E_0=(a-1)E_0,\\
\epsilon(\nu,c-k_1,\mu,c) &\overset{(2)}{=} \epsilon(c-k_1,\nu,c,\mu)\\
&= \big[(a-1)b-ba\big]E_0=-b\,E_0,\\
\epsilon(c-q,\nu,\mu,c) &\overset{(3)}{=} -\epsilon(c-q,\nu,c,\mu)\\
&= -\big[(a-1)b-(b-1)a\big]E_0=-(a-b)E_0.
\end{aligned}$$

(1) $c-q=(a-1)k_1+(b-1)k_2$ and $c-k_1=(a-1)k_1+bk_2$.
(2) Swapping slots $1\leftrightarrow2$ and $3\leftrightarrow4$ is even; $c-k_1=(a-1,b)$, $c=(a,b)$.
(3) Swap the last two slots (odd).

The sum is $\big[(a-1)-b-(a-b)\big]E_0=-E_0$: all dependence on the Feynman parameters has cancelled. So after the shift the numerator is $-4i\,\ell_\perp^2\cdot(-E_0)=4i\,\ell_\perp^2E_0$.

Step (d): the integral.

$$\begin{aligned}
I:=\int\frac{\mathrm{d}^d\ell}{(2\pi)^d}\frac{\ell_\perp^2}{(\ell^2-\Delta)^3} &\overset{(1)}{=} \frac{d-4}{d}\int\frac{\mathrm{d}^d\ell}{(2\pi)^d}\frac{\ell^2}{(\ell^2-\Delta)^3}\\
&\overset{(2)}{=} \frac{d-4}{d}\cdot\frac{(-1)^2\,i}{(4\pi)^{d/2}}\cdot\frac{d}{2}\cdot\frac{\Gamma(2-\frac{d}{2})}{\Gamma(3)}\,\Delta^{d/2-2} \overset{(3)}{=} \frac{i\,(d-4)\,\Gamma(2-\frac{d}{2})}{4\,(4\pi)^{d/2}}\,\Delta^{d/2-2}\\
&\overset{(4)}{=} -\frac{i\,\Gamma(3-\frac{d}{2})}{2\,(4\pi)^{d/2}}\,\Delta^{d/2-2} \overset{(5)}{=} -\frac{i}{2(4\pi)^2}=-\frac{i}{32\pi^2}\quad(d\to4).
\end{aligned}$$

(1) Symmetric integration (10.42): $\ell_\perp^2=\sum_{a\ge4}\eta_{aa}\ell^a\ell^a\to\frac{\ell^2}{d}\sum_{a\ge4}\eta_{aa}\eta^{aa}=\frac{d-4}{d}\ell^2$.
(2) Master integral (10.41) with $n=3$.
(3) $\Gamma(3)=2$; collect.
(4) $(d-4)\Gamma(2-\frac{d}{2})=-2(2-\frac{d}{2})\Gamma(2-\frac{d}{2})=-2\Gamma(3-\frac{d}{2})$ by (10.43): the vanishing factor eats the pole.
(5) At $d=4$: $\Gamma(1)=1$ and $\Delta^0=1$, so the answer does not depend on Δ.

Step (e): assemble. For the first graph,

$$\begin{aligned}
2\int\frac{\mathrm{d}^dp}{(2\pi)^d}\frac{N_1}{p^2(p-q)^2(p-k_1)^2} &\overset{(1)}{=} 2\cdot2\int\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z\,\delta(1-x-y-z)\;4iE_0\,I\\
&\overset{(2)}{=} 2\cdot2\cdot\tfrac12\cdot4iE_0\cdot\Big(-\frac{i}{32\pi^2}\Big) \overset{(3)}{=} \frac{E_0}{4\pi^2} \overset{(4)}{=} \frac{1}{4\pi^2}\,\epsilon^{\mu\nu\rho\sigma}k_{1\rho}k_{2\sigma}.
\end{aligned}$$

(1) (10.38) and Steps (b)–(c).
(2) I is independent of the Feynman parameters, whose integral is $\frac{1}{2}$.
(3) $8i \cdot (-i)/32\pi^2 = 8/32\pi^2$.
(4) $E_0 = \epsilon^{\rho\nu\sigma\mu} k_{1\rho} k_{2\sigma}$ and $(\rho\nu\sigma\mu) \to (\mu\nu\sigma\rho) \to (\mu\nu\rho\sigma)$ is two transpositions.

The second graph gives the same (exchange $(k_1, \mu) \leftrightarrow (k_2, \nu)$: two sign flips of ϵ). Hence

$$q_\lambda T^{\mu\nu\lambda}\big|_{\text{dim reg}} = \frac{1}{2\pi^2}\, \epsilon^{\mu\nu\rho\sigma} k_{1\rho} k_{2\sigma},$$

exactly (10.97): dimensional regularisation automatically chooses the gauge-invariant routing $\beta = -1$ and puts the whole anomaly into the axial current [QFT notes (22.3.41)–(22.3.44)]. The pattern is the same as in the a coefficient of Section 10.3.3: a regulator that respects gauge invariance leaves behind "$0 \times \infty$ = finite", here $(d-4) \times \Gamma(2 - \frac{d}{2}) = -2$.

Physical picture: Why the anomaly is one-loop exact

The anomaly came from a surface term, i.e. from the fact that the one-loop triangle is only linearly divergent and depends on the routing. Graphs with more loops (photon lines dressing the triangle) have more propagators, converge better, and allow free shifts; they cannot contribute. This heuristic is the **Adler–Bardeen theorem**: the coefficient $e^2/16\pi^2$ receives no higher-order corrections [QFT notes §22.8]. It is the perturbative shadow of the fact, to be established in Section 10.4.5, that the integrated anomaly is an integer.

Reading Nash (pp. 272–273, (10.9)–(10.11))

Nash's "routine calculation" is our (10.23)–(10.26). His ∇_μ^{ab} is the adjoint covariant derivative D_μ, his $j_5^{b\mu}$ is our chiral current $J^{b\mu}$ of (10.24), and his (10.10) is our (10.27). The "celebrated divergent triangle graph" is Figure 10.10; "divergent" is the key word, since the anomaly is a surface term of a linearly divergent integral. Nash then cites Fujikawa [1] for the statement that the measure $\mathcal{D}\psi\mathcal{D}\bar{\psi}$ is not gauge invariant; we do that computation next.

Recap

- Along a gauge orbit, $\frac{\mathrm{d}}{\mathrm{d}t} \ln Z_F = \langle \int f^a D_\mu J^{a\mu} \rangle$: the determinant is gauge invariant iff the gauge current is conserved in the quantum theory.
- QED$_2$: $\Pi^{\mu\nu} = \mathsf{a}\eta^{\mu\nu} - \mathsf{b} q^\mu q^\nu/q^2$ with $\mathsf{b} = e^2/\pi$ finite. Vector conservation needs $\mathsf{a} = \mathsf{b}$, axial needs $\mathsf{a} = 0$. Gauge invariance wins: $\partial_\mu j_5^\mu = \frac{e}{2\pi}\epsilon^{\mu\nu}F_{\mu\nu}$, and $m_\gamma^2 = e^2/\pi$.
- Point splitting: $\varepsilon F \times \frac{1}{\varepsilon}$ is finite. Spectral flow on a circle: one flux quantum converts one right-mover into one left-mover, $\Delta Q_5 = -2$.
- QED$_4$: the triangle is linearly divergent; routing ambiguity $\frac{\beta}{8\pi^2}\epsilon(k_1 - k_2)$; axial WI needs $\beta = 1$, vector WI $\beta = -1$. With $\beta = -1$: $\partial_\mu j_5^\mu = -\frac{e^2}{16\pi^2}\epsilon^{\mu\nu\rho\sigma}F_{\mu\nu}F_{\rho\sigma}$ (Minkowski).
- Every method finds the anomaly as "$0 \times \infty$": a vanishing factor $(\varepsilon, d-4, 2-d)$ times a divergence, left behind by a gauge-invariant regulator.

10.4 Anomalies and Yang–Mills theories III: the measure, the heat kernel and the index

In Section 10.3 we found the anomaly three times, and each time it came out of a regularised product of fields at one point. Nash's sentence on p. 273 points to a fourth derivation, due to Fujikawa, which generalises to every even dimension and connects directly to topology: "the Fermionic 'measure' $\mathcal{D}\psi\mathcal{D}\bar\psi$ is not gauge invariant". The question this section answers is: *if the action is invariant and the anomaly is real, which part of the path integral changes, and by how much?* The answer is the measure. Its change is a Jacobian determinant, which we shall compute exactly. Integrated over spacetime, the same Jacobian turns out to be the index of the Dirac operator.

We work in Euclidean signature on a compact M of dimension $n = 2m$, with the conventions of Section 10.1.1: Hermitian γ_μ, $\gamma_5 = (-i)^m\gamma_1\cdots\gamma_{2m}$, anti-Hermitian A, $D_\mu = \partial_\mu + A_\mu$, and $\not{D}_A = \gamma^\mu D_\mu$ anti-self-adjoint. The orientation is fixed by $\epsilon_{12\cdots n} = +1$. For the local computations we may take M flat (a large torus): the answer is local, and curvature only adds terms that we treat in Section 10.7.

10.4.1 Fujikawa's idea: the Jacobian of a chiral rotation

The logic in one paragraph. Take a Dirac fermion in a fixed background A,

$$Z_F[A] = \int \mathcal{D}\psi\,\mathcal{D}\bar\psi\; e^{-S_F[\psi,\bar\psi]}, \qquad S_F = \int_M \mathrm{d}^n x\,\bar\psi\,\not{D}_A\psi. \tag{10.105}$$

The integration variables are dummies, so renaming them cannot change Z_F. Rename them by an infinitesimal *local* chiral rotation,

$$\psi' = (1 + i\alpha(x)\gamma_5)\psi, \qquad \bar\psi' = \bar\psi\,(1 + i\alpha(x)\gamma_5), \tag{10.106}$$

with $\alpha(x)$ a small real function. Two things happen: the action changes by the Noether term (10.7), and the measure changes by a Jacobian J. Since the total change is zero, the expectation value of the Noether term equals minus the log of the Jacobian. If J were 1 we would get $\langle\partial_\mu j^\mu_5\rangle = 0$. Fujikawa's discovery is that $J \neq 1$.

Step 1: a basis for the fields. To speak of a Jacobian we need coordinates on the space of fields. The operator $i\not{D}_A$ is self-adjoint (Section 10.1.1), so it has real eigenvalues λ_n and an orthonormal basis of eigenspinors,

$$i\not{D}_A\,\varphi_n = \lambda_n\varphi_n, \qquad \int_M \mathrm{d}^n x\,\varphi_n^\dagger(x)\varphi_k(x) = \delta_{nk}, \qquad \sum_n \varphi_n(x)\varphi_n^\dagger(y) = \delta^{(n)}(x-y)\,\mathbf{1}. \tag{10.107}$$

Here $\varphi_n(x)$ is a column vector (spin $\otimes$ colour) and $\mathbf{1}$ is the identity on spin $\otimes$ colour; the last relation is completeness. Expand

$$\psi(x) = \sum_n a_n\varphi_n(x), \qquad \bar\psi(x) = \sum_n b_n\varphi_n^\dagger(x), \tag{10.108}$$

with Grassmann coefficients a_n, b_n (independent, since in Euclidean signature $\bar\psi$ is not the conjugate of ψ), and *define* the measure as $\mathcal{D}\psi\mathcal{D}\bar\psi := \prod_n \mathrm{d}a_n\,\mathrm{d}b_n$ [QFT notes (22.3.54)]. In this basis the action is diagonal: $S_F = \sum_n b_n(-i\lambda_n)a_n$, because $\not{D}_A\varphi_n = -i\lambda_n\varphi_n$.

Step 2: how Grassmann measures change. This is the one place where fermions differ from bosons, and it is the source of the anomaly's sign.

Working definition 10.6 (change of Grassmann variables)

If $\theta_i' = \sum_j M_{ij}\theta_j$ for Grassmann variables $\theta_1, \dots, \theta_N$ [anticommuting variables, here the coefficients a_n, b_n of ψ and $\bar\psi$ in the eigenbasis], then

$$\prod_i \mathrm{d}\theta_i' = (\det M)^{-1} \prod_i \mathrm{d}\theta_i.$$

The Jacobian [the factor multiplying the measure $\prod_i \mathrm{d}\theta_i$] is the *inverse* of the bosonic one [for ordinary commuting variables $x' = Mx$, $\prod_i \mathrm{d}x_i' = \det M \prod_i \mathrm{d}x_i$].

Why: for one variable, $\theta' = c\,\theta$, and both $\int \mathrm{d}\theta'\,\theta' = 1$ and $\int \mathrm{d}\theta\,\theta = 1$ must hold. So

$$1 = \int \mathrm{d}\theta'\,\theta' \overset{(1)}{=} \int \mathrm{d}\theta'\,c\,\theta \overset{(2)}{=} c \cdot \big(\text{factor in } \mathrm{d}\theta' = \kappa\,\mathrm{d}\theta\big) \int \mathrm{d}\theta\,\theta = c\,\kappa,$$

(1) Substitute $\theta' = c\theta$.
(2) Write $\mathrm{d}\theta' = \kappa\,\mathrm{d}\theta$ and pull out the constants.

hence $\kappa = 1/c$. For N variables, antisymmetry gives $\theta_1'\theta_2'\cdots\theta_N' = (\det M)\,\theta_1\theta_2\cdots\theta_N$ (the same bookkeeping as in Example 10.6: only terms with distinct θ's survive, each with the sign of its permutation), and the same argument gives $(\det M)^{-1}$.

Step 3: the transformed coefficients. Insert (10.108) into (10.106) and project onto φ_n with the orthonormality (10.107):

$$\begin{aligned} a_n' &\overset{(1)}{=} \int \mathrm{d}^n x\,\varphi_n^\dagger\,\psi' \overset{(2)}{=} \sum_k \int \mathrm{d}^n x\,\varphi_n^\dagger(1 + i\alpha\gamma_5)\varphi_k\; a_k \overset{(3)}{=} \sum_k \big(\delta_{nk} + C_{nk}\big)a_k, \\ b_n' &\overset{(4)}{=} \int \mathrm{d}^n x\,\bar\psi'\varphi_n \overset{(5)}{=} \sum_k b_k \int \mathrm{d}^n x\,\varphi_k^\dagger(1 + i\alpha\gamma_5)\varphi_n = \sum_k b_k\big(\delta_{kn} + C_{kn}\big), \end{aligned} \tag{10.109}$$

with

$$C_{nk} := i\int_M \mathrm{d}^n x\,\alpha(x)\,\varphi_n^\dagger(x)\gamma_5\varphi_k(x). \tag{10.110}$$

(1) The coefficient of φ_n in any spinor χ is $\int \varphi_n^\dagger\chi$, by orthonormality.
(2) Insert $\psi' = (1 + i\alpha\gamma_5)\sum_k a_k\varphi_k$; the Grassmann a_k are constants and come out.
(3) $\int \varphi_n^\dagger\varphi_k = \delta_{nk}$; the rest is the definition of C_{nk}.
(4) The coefficient of $\varphi_n^\dagger$ in a row spinor $\bar\chi$ is $\int \bar\chi\varphi_n$.
(5) Insert $\bar\psi' = \sum_k b_k\varphi_k^\dagger(1 + i\alpha\gamma_5)$.

So $a' = (1 + C)a$ and $b' = (1 + C)^{\mathsf T}b$. By Definition 10.6, and $\det(1 + C)^{\mathsf T} = \det(1 + C)$,

$$\mathcal{D}\psi'\mathcal{D}\bar\psi' = \det(1 + C)^{-2}\,\mathcal{D}\psi\mathcal{D}\bar\psi. \tag{10.111}$$

Step 4: the determinant to first order. For an infinitesimal matrix C,

$$\begin{aligned} \det(1 + C)^{-2} &\overset{(1)}{=} \exp\big[-2\,\mathrm{Tr}\ln(1 + C)\big] \overset{(2)}{=} \exp\big[-2\,\mathrm{Tr}\,C + O(\alpha^2)\big] \\ &\overset{(3)}{=} \exp\Big[-2i\int_M \mathrm{d}^n x\,\alpha(x)\sum_n \varphi_n^\dagger(x)\gamma_5\varphi_n(x)\Big]. \end{aligned} \tag{10.112}$$

(1) $\det X = \exp\mathrm{Tr}\ln X$: true for diagonalisable X (product of eigenvalues = exponential of the sum of their logs), and by continuity for all X near 1.

(2) $\ln(1+C) = C - \frac{1}{2}C^2 + \cdots$, and C is first order in α.
(3) $\operatorname{Tr} C = \sum_n C_{nn}$ with (10.110).

The factor 2 counts ψ and $\bar\psi$, which rotate by the same $e^{i\alpha\gamma_5}$. For a *vector* rotation $\psi' = (1+i\alpha)\psi$, $\bar\psi' = \bar\psi(1-i\alpha)$ the two Jacobians are $\det(1+C_V)^{-1}$ and $\det(1-C_V)^{-1}$ with C_V the same matrix without γ_5, and to first order they cancel: *the measure is invariant under vector rotations*, so the vector current is conserved.

Step 5: the anomalous Ward identity. Define the **anomaly density**

$$\mathscr{A}(x) := \sum_n \varphi_n^\dagger(x)\gamma_5\varphi_n(x) \qquad \text{(to be regularised).} \tag{10.113}$$

Then

$$\begin{aligned} Z_F &\overset{(1)}{=} \int \mathcal{D}\psi'\mathcal{D}\bar\psi'\; e^{-S_F[\psi',\bar\psi']} \\ &\overset{(2)}{=} \int \mathcal{D}\psi\mathcal{D}\bar\psi\; e^{-2i\int\alpha\mathscr{A}}\; e^{-S_F[\psi,\bar\psi]+i\int\alpha\,\partial_\mu j_5^\mu} \\ &\overset{(3)}{=} Z_F + \int \mathcal{D}\psi\mathcal{D}\bar\psi\; e^{-S_F}\Big(-2i\int\alpha\mathscr{A} + i\int\alpha\,\partial_\mu j_5^\mu\Big) + O(\alpha^2). \end{aligned} \tag{10.114}$$

(1) Rename the dummy variables $\psi \to \psi'$, $\bar\psi \to \bar\psi'$.
(2) Express $\psi', \bar\psi'$ through $\psi, \bar\psi$: the measure gives (10.112), and the action gives $S_F[\psi'] = S_F[\psi] + \delta S_F$ with $\delta S_F = -i\int\alpha\,\partial_\mu j_5^\mu$ from (10.7) (the background A was already included there).
(3) Expand both exponentials to first order in α.

The first-order term must vanish for every $\alpha(x)$. Dividing by Z_F,

$$\boxed{\langle \partial_\mu j_5^\mu(x)\rangle_A = 2\,\mathscr{A}(x)} \tag{10.115}$$

The classical conservation law has become an equation whose right side is the Jacobian of the measure. Everything now depends on the value of $\mathscr{A}(x)$.

Watch out: $0\times\infty$ once more

Naively $\mathscr{A}(x) = \sum_n \varphi_n^\dagger\gamma_5\varphi_n = \operatorname{tr}\big[\gamma_5\sum_n \varphi_n(x)\varphi_n^\dagger(x)\big] = \operatorname{tr}(\gamma_5)\,\delta^{(n)}(0)$ by completeness. The first factor is 0 (γ_5 is traceless) and the second is ∞. The product is undefined and must be regularised. A regularisation cuts off large $|\lambda_n|$, and to keep gauge invariance the cutoff must be defined by the gauge-covariant operator $\not{D}_A$ itself, not by the free $\not{\partial}$.

10.4.2 The heat-kernel regulator, computed step by step

Following Fujikawa [QFT notes (22.3.62)], we define

$$\mathscr{A}(x) := \lim_{M\to\infty}\sum_n \varphi_n^\dagger(x)\gamma_5\, e^{-\lambda_n^2/M^2}\varphi_n(x). \tag{10.116}$$

The damping factor makes the sum converge (Weyl's law: the number of modes with $|\lambda| < \Lambda$ grows like Λ^n, while $e^{-\lambda^2/M^2}$ decays faster than any power). We now evaluate (10.116) in seven short steps.

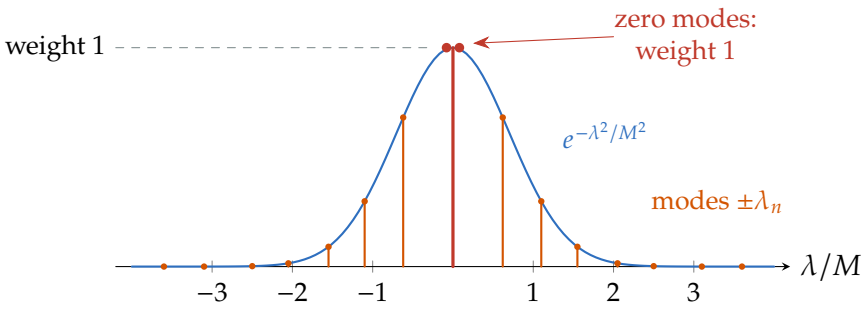


Figure 10.13: Fujikawa's regulator. Each eigenmode of $i\not{D}_A$ (orange sticks; they come in pairs $\pm\lambda_n$, see Section 10.4.5) is weighted by $e^{-\lambda_n^2/M^2}$, and each stick is drawn as tall as its weight. Modes with $|\lambda_n| \ll M$ keep weight ≈ 1; modes with $|\lambda_n| \gg M$ are switched off. Since the weight is a function of $\not{D}_A$, it is gauge covariant. Zero modes (red) always have weight 1: that is why the integrated anomaly will not depend on M.

Step 1: from modes to an operator kernel. Since $\not{D}_A\varphi_n = -i\lambda_n\varphi_n$, we have $\not{D}_A^2\varphi_n = -\lambda_n^2\varphi_n$ and so $e^{\not{D}_A^2/M^2}\varphi_n = e^{-\lambda_n^2/M^2}\varphi_n$. Hence

$$\sum_n \varphi_n^\dagger(x)\gamma_5 e^{-\lambda_n^2/M^2}\varphi_n(x) \overset{(1)}{=} \sum_n \operatorname{tr}\Big[\gamma_5\, e^{\not{D}_A^2/M^2}\varphi_n(x)\varphi_n^\dagger(y)\Big]_{y=x}$$
$$\overset{(2)}{=} \operatorname{tr}\Big[\gamma_5\, e^{\not{D}_A^2/M^2}\delta^{(n)}(x-y)\Big]_{y=x} =: \operatorname{tr}\,\langle x|\gamma_5\, e^{\not{D}_A^2/M^2}|x\rangle. \tag{10.117}$$

(1) $\chi^\dagger\gamma_5\chi = \operatorname{tr}(\gamma_5\chi\chi^\dagger)$ for a column vector χ; the operator $e^{\not{D}_A^2/M^2}$ acts on the x-dependence.
(2) Completeness (10.107). The trace is over spin and colour.

The object $\langle x|e^{t\not{D}_A^2}|y\rangle$ is the **heat kernel** of the operator $-\not{D}_A^2$ at "time" $t = 1/M^2$: it solves $\partial_t K = \not{D}_A^2 K$ with $K \to \delta$ as $t \to 0$. Fujikawa's regulator is the heat kernel at coincident points, at small time.

Step 2: the square of the Dirac operator. We need $\not{D}_A^2$ explicitly ([QFT notes (22.3.64)] in Minkowski form):

$$\not{D}_A^2 = \gamma^\mu\gamma^\nu D_\mu D_\nu \overset{(1)}{=} \tfrac{1}{2}\{\gamma^\mu,\gamma^\nu\}D_\mu D_\nu + \tfrac{1}{2}[\gamma^\mu,\gamma^\nu]D_\mu D_\nu$$
$$\overset{(2)}{=} D_\mu D_\mu + \tfrac{1}{4}[\gamma^\mu,\gamma^\nu][D_\mu,D_\nu]$$
$$\overset{(3)}{=} D_\mu D_\mu + \tfrac{1}{2}\gamma^\mu\gamma^\nu F_{\mu\nu}. \tag{10.118}$$

(1) Split $\gamma^\mu\gamma^\nu$ into its symmetric and antisymmetric parts.
(2) Clifford relation $\{\gamma^\mu,\gamma^\nu\} = 2\delta^{\mu\nu}$; in the second term only the part of $D_\mu D_\nu$ antisymmetric in $\mu\nu$ survives, which is $\frac{1}{2}[D_\mu,D_\nu]$.
(3) $[D_\mu,D_\nu] = \partial_\mu A_\nu - \partial_\nu A_\mu + [A_\mu,A_\nu] = F_{\mu\nu}$, and $\frac{1}{4}[\gamma^\mu,\gamma^\nu]F_{\mu\nu} = \frac{1}{2}\gamma^\mu\gamma^\nu F_{\mu\nu}$ since $F_{\mu\nu}$ is antisymmetric.

Write $\Phi := \frac{1}{2}\gamma^\mu\gamma^\nu F_{\mu\nu}$. It is a matrix acting on spin (through the γ's) and colour (through F), with no derivatives. It is the only place where γ-matrices appear in $\not{D}_A^2$.

Step 3: plane waves. On flat $\mathbb{R}^n$ (or a large torus) plane waves are complete: $\delta^{(n)}(x-y) = \int \frac{\mathrm{d}^n k}{(2\pi)^n} e^{ik\cdot(x-y)}$. For any operator O built from D_μ and multiplication operators,

$$\langle x|O|x\rangle \overset{(1)}{=} \int \frac{\mathrm{d}^n k}{(2\pi)^n}\, e^{-ik\cdot x}\, O\, e^{ik\cdot x} \overset{(2)}{=} \int \frac{\mathrm{d}^n k}{(2\pi)^n}\, O\big|_{D_\mu\to D_\mu+ik_\mu}\cdot\mathbf{1}. \tag{10.119}$$

(1) Act with $\mathcal{O}$ on $\delta(x-y)$ written as a Fourier integral, and set $y=x$.
(2) $e^{-ik\cdot x}D_\mu\big(e^{ik\cdot x}\chi\big)=(D_\mu+ik_\mu)\chi$ by the Leibniz rule; repeat for every factor of D_μ. The final "$\cdot\mathbf{1}$" means the operator acts on the constant function (with values the identity matrix).

Under $D\to D+ik$ the pieces of (10.118) become

$$(D_\mu+ik_\mu)(D_\mu+ik_\mu)=-k^2+2ik_\mu D_\mu+D_\mu D_\mu,\qquad [D_\mu+ik_\mu,D_\nu+ik_\nu]=F_{\mu\nu},\tag{10.120}$$

because the numbers k_μ commute with everything. So Φ is unchanged and

$$\mathscr{A}(x)=\lim_{M\to\infty}\int\frac{\mathrm{d}^nk}{(2\pi)^n}\ \mathrm{tr}\ \gamma_5\exp\Big[\frac{-k^2+2ik\cdot D+D^2+\Phi}{M^2}\Big]\cdot\mathbf{1}.\tag{10.121}$$

Step 4: rescale the momentum. Put $k=M\kappa$, so $\mathrm{d}^nk=M^n\mathrm{d}^n\kappa$:

$$\begin{aligned}\mathscr{A}(x)&\overset{(1)}{=}\lim_{M\to\infty}M^n\int\frac{\mathrm{d}^n\kappa}{(2\pi)^n}\ \mathrm{tr}\ \gamma_5\exp\Big[-\kappa^2+\frac{X_1}{M}+\frac{X_2}{M^2}\Big]\cdot\mathbf{1}\\&\overset{(2)}{=}\lim_{M\to\infty}M^n\int\frac{\mathrm{d}^n\kappa}{(2\pi)^n}\,e^{-\kappa^2}\ \mathrm{tr}\ \gamma_5\exp\Big[\frac{X_1}{M}+\frac{X_2}{M^2}\Big]\cdot\mathbf{1},\end{aligned}\tag{10.122}$$

with $X_1\coloneqq 2i\kappa_\mu D_\mu$ and $X_2\coloneqq D^2+\Phi$.

(1) Substitute $k=M\kappa$ in (10.121).
(2) $-\kappa^2$ is a number times the identity, so it commutes with X_1, X_2 and $e^{-\kappa^2+Y}=e^{-\kappa^2}e^{Y}$ exactly. (This is why no ordering subtlety arises.)

Step 5: power counting. Expand $e^{Y}=\sum_{j\ge0}Y^j/j!$ with $Y=X_1/M+X_2/M^2$. A product containing a factors of X_1 and b factors of X_2 carries M^{-a-2b}, so with the prefactor it is of order $M^{n-a-2b}=M^{2m-a-2b}$. Now use two facts:

- Only Φ contains γ-matrices, two per factor.
- $\mathrm{tr}(\gamma_5\times$ a product of fewer than $2m$ γ's$)=0$ (proved in Step 7).

So a nonzero trace needs at least m factors of Φ, hence $b\ge m$ and the order is at most $M^{2m-a-2b}\le M^0$. Equality needs $a=0$, $b=m$, and every one of the m factors of X_2 equal to Φ (a factor D^2 in place of a Φ would force $b\ge m+1$). Every other term vanishes as $M\to\infty$. The terms that would have *diverged*, $M^{2m}\,\mathrm{tr}\,\gamma_5$, $M^{2m-2}\,\mathrm{tr}(\gamma_5\Phi)$, …, all have too few γ's and vanish identically. Therefore

$$\mathscr{A}(x)=\int\frac{\mathrm{d}^n\kappa}{(2\pi)^n}\,e^{-\kappa^2}\,\frac{1}{m!}\,\mathrm{tr}\big(\gamma_5\,\Phi^m\big).\tag{10.123}$$

The coefficient $1/m!$ comes from the term $Y^m/m!$, the only one containing $(X_2)^m$ with no X_1.

Step 6: the Gaussian integral.

$$\int\frac{\mathrm{d}^n\kappa}{(2\pi)^n}\,e^{-\kappa^2}\overset{(1)}{=}\frac{1}{(2\pi)^n}\Big(\int_{-\infty}^{\infty}e^{-u^2}\mathrm{d}u\Big)^n\overset{(2)}{=}\frac{\pi^{n/2}}{(2\pi)^n}\overset{(3)}{=}\frac{1}{(4\pi)^m}.\tag{10.124}$$

(1) The integrand factorises over the n Cartesian components of κ.
(2) $\int e^{-u^2}\mathrm{d}u=\sqrt{\pi}$.
(3) $n=2m$: $\pi^m/(4^m\pi^{2m})=(4\pi)^{-m}$.

For $n=4$ this is $1/16\pi^2$; restoring M, it is the statement $\langle x|e^{\partial^2/M^2}|x\rangle=M^4/16\pi^2$ [QFT notes (22.3.69)].

Step 7: the γ_5 trace. We need

$$\operatorname{tr}\big(\gamma_5\gamma^{\mu_1}\gamma^{\mu_2}\cdots\gamma^{\mu_{2m}}\big) = (2i)^m\,\epsilon^{\mu_1\mu_2\cdots\mu_{2m}}, \tag{10.125}$$

and that the trace of γ_5 with fewer γ's vanishes. The argument has three parts.

1. *Products of distinct γ's are traceless.* Let P be a product of j distinct γ's, $1 \le j \le 2m$. If j is even, write $P = \gamma_a Q$ with Q the other $j-1$ factors; then $\gamma_a Q = (-1)^{j-1} Q\gamma_a = -Q\gamma_a$, and cyclicity gives $\operatorname{tr} P = \operatorname{tr}(\gamma_a Q) = \operatorname{tr}(Q\gamma_a) = -\operatorname{tr} P$. If j is odd and $j < 2m$, pick γ_c not in P; then $\gamma_c P \gamma_c = (-1)^j P = -P$ and $\operatorname{tr} P = \operatorname{tr}(\gamma_c\gamma_c P) = \operatorname{tr}(\gamma_c P\gamma_c) = -\operatorname{tr} P$. Either way $\operatorname{tr} P = 0$.

2. *Fewer than $2m$ γ's.* Use $\gamma_\mu\gamma_\mu = \mathbf{1}$ and anticommutation to remove repeated indices; γ_5 times what is left is $\pm$(a nonzero power of i) times a product of distinct γ's, which is traceless by part 1 unless it is the identity. It is the identity only if all $2m$ directions appeared, which needs at least $2m$ factors.

3. *Exactly $2m$ γ's.* If two indices coincide, part 2 gives zero; this matches $\epsilon = 0$. If all are distinct, reorder them to $\gamma_1\cdots\gamma_{2m}$, which costs the sign of the permutation, i.e. $\epsilon^{\mu_1\cdots\mu_{2m}}$. Finally $\gamma_1\cdots\gamma_{2m} = i^m\gamma_5$ (multiply (10.3) by i^m and use $(i)^m(-i)^m = 1$), so $\operatorname{tr}(\gamma_5\gamma_1\cdots\gamma_{2m}) = i^m\operatorname{tr}(\gamma_5^2) = i^m\,2^m$, since spinors have 2^m components.

Check: for $m=1$, $\operatorname{tr}(\gamma_5\gamma_1\gamma_2) = \operatorname{tr}(\sigma_3\sigma_1\sigma_2) = \operatorname{tr}(\sigma_3\cdot i\sigma_3) = 2i = (2i)^1$. For $m=2$, (10.125) gives $\operatorname{tr}(\gamma_5\gamma^\mu\gamma^\nu\gamma^\rho\gamma^\sigma) = -4\epsilon^{\mu\nu\rho\sigma}$; with $\gamma_1\gamma_2\gamma_3\gamma_4 = -\gamma_5$ from Example 10.1, indeed $\operatorname{tr}(\gamma_5\gamma_1\gamma_2\gamma_3\gamma_4) = -\operatorname{tr}\mathbf{1} = -4$.

The result. Put Steps 5–7 together:

$$\begin{aligned}\mathscr{A}(x) &\overset{(1)}{=} \frac{1}{(4\pi)^m}\,\frac{1}{m!}\,\frac{1}{2^m}\,\operatorname{tr}_{\text{spin}}\big(\gamma_5\gamma^{\mu_1}\gamma^{\mu_2}\cdots\gamma^{\mu_{2m}}\big)\,\operatorname{tr}_{\text{col}}\big(F_{\mu_1\mu_2}\cdots F_{\mu_{2m-1}\mu_{2m}}\big)\\ &\overset{(2)}{=} \frac{1}{m!}\Big(\frac{i}{4\pi}\Big)^m\,\epsilon^{\mu_1\cdots\mu_{2m}}\,\operatorname{tr}\big(F_{\mu_1\mu_2}\cdots F_{\mu_{2m-1}\mu_{2m}}\big).\end{aligned} \tag{10.126}$$

(1) (10.124) and $\Phi^m = 2^{-m}\gamma^{\mu_1}\gamma^{\mu_2}\cdots\gamma^{\mu_{2m}}F_{\mu_1\mu_2}\cdots F_{\mu_{2m-1}\mu_{2m}}$; the γ's and F's act on different indices, so the trace factorises.
(2) (10.125): $(2i)^m/(2^m(4\pi)^m) = (i/4\pi)^m$.

Key idea: The Fujikawa Jacobian is the Chern character

Compare with the Chern character form $\operatorname{ch}_m(F) = \frac{1}{m!}\big(\frac{i}{2\pi}\big)^m\operatorname{tr}F^m$. Since $F = \frac12 F_{\mu\nu}\mathrm{d}x^\mu\wedge\mathrm{d}x^\nu$,

$$\operatorname{tr}F^m = \frac{1}{2^m}\,\epsilon^{\mu_1\cdots\mu_{2m}}\operatorname{tr}\big(F_{\mu_1\mu_2}\cdots F_{\mu_{2m-1}\mu_{2m}}\big)\,\mathrm{d}^n x,$$

so $\operatorname{ch}_m(F) = \frac{1}{m!}\big(\frac{i}{4\pi}\big)^m\epsilon\operatorname{tr}(F\cdots F)\,\mathrm{d}^n x$, and

$$\mathscr{A}(x)\,\mathrm{d}^n x = \operatorname{ch}_m(F), \qquad \langle\partial_\mu j_5^\mu\rangle\,\mathrm{d}^n x = 2\,\operatorname{ch}_m(F). \tag{10.127}$$

The anomaly density is the top-degree Chern–Weil form of the gauge field.

Worked example 10.12: Four dimensions, written out

For $m=2$ the chain (10.126) reads

$$\mathscr{A} \overset{(1)}{=} \frac{1}{16\pi^2}\cdot\frac12\cdot\frac14\,\operatorname{tr}(\gamma_5\gamma^\mu\gamma^\nu\gamma^\rho\gamma^\sigma)\,\operatorname{tr}(F_{\mu\nu}F_{\rho\sigma}) \overset{(2)}{=} \frac{1}{128\pi^2}(-4)\,\epsilon^{\mu\nu\rho\sigma}\operatorname{tr}(F_{\mu\nu}F_{\rho\sigma})$$

$$= -\frac{1}{32\pi^2}\,\epsilon^{\mu\nu\rho\sigma}\,\mathrm{tr}(F_{\mu\nu}F_{\rho\sigma}).$$

(1) Gaussian $1/16\pi^2$, the $1/2!$ from $Y^2/2!$, and $\Phi^2 = \frac{1}{4}\gamma\gamma\gamma\gamma FF$.
(2) The four-dimensional trace $-4\epsilon^{\mu\nu\rho\sigma}$.

For QED, $A_\mu = -iea_\mu$ and $F_{\mu\nu} = -ief_{\mu\nu}$ (dictionary in Section 10.1.1), so $\mathrm{tr}(F_{\mu\nu}F_{\rho\sigma}) = -e^2 f_{\mu\nu}f_{\rho\sigma}$ and

$$\langle \partial_\mu j_5^\mu \rangle = 2\mathscr{A} = \frac{e^2}{16\pi^2}\,\epsilon^{\mu\nu\rho\sigma} f_{\mu\nu}f_{\rho\sigma} \qquad \text{(Euclidean)},$$

which is (10.9). The Minkowski computation of the QFT notes [QFT notes (22.3.70)–(22.3.74)] follows the same seven steps and gives $\partial_\mu j_5^\mu = -\frac{e^2}{16\pi^2}\epsilon^{\mu\nu\rho\sigma}F_{\mu\nu}F_{\rho\sigma}$ with $\epsilon^{0123} = +1$, the same result as the point-split and triangle computations (10.75).

Watch out: Signs between Euclidean and Minkowski

The two forms of the ABJ equation differ by a sign, and that is not a contradiction. In Minkowski signature with $\epsilon^{0123} = +1$, $F_{0i} = E_i$ and $F_{ij} = -\epsilon_{ijk}B_k$, the chain (10.76) gives $\epsilon^{\mu\nu\rho\sigma}F_{\mu\nu}F_{\rho\sigma} = -8\,\mathbf{E}\cdot\mathbf{B}$, so both statements say $\partial_\mu j_5^\mu = \frac{e^2}{2\pi^2}\mathbf{E}\cdot\mathbf{B}$. The Wick rotation $x^0 = -ix^4$ changes ϵ, the components of F, and the relation between j_5^0 and j_5^4, each by a factor of $\pm i$ or ± 1. When comparing with a book, compare $\mathbf{E}\cdot\mathbf{B}$ or the integrated (index) form, never the raw ϵFF.

10.4.3 Why the answer does not depend on the regulator

The Gaussian $e^{-\lambda^2/M^2}$ looks arbitrary. Replace it by $f(\lambda^2/M^2)$ with any smooth f such that $f(0) = 1$ and $f(u) \to 0$ fast as $u \to \infty$. Steps 1–5 go through with $e^{-\kappa^2+Y}$ replaced by $f(\kappa^2 - Y)$, where now $-\not{D}^2/M^2 \to \kappa^2 - Y$. (Here κ^2 commutes with Y, so Taylor expansion in Y is legitimate.) The surviving term is the one with Φ^m:

$$\begin{aligned}
\mathscr{A}_f &\overset{(1)}{=} \int \frac{\mathrm{d}^n\kappa}{(2\pi)^n}\,\frac{(-1)^m f^{(m)}(\kappa^2)}{m!}\,\mathrm{tr}(\gamma_5\Phi^m)\\
&\overset{(2)}{=} \frac{\pi^m}{(2\pi)^{2m}\,(m-1)!}\int_0^\infty \mathrm{d}u\; u^{m-1}(-1)^m f^{(m)}(u)\cdot\frac{\mathrm{tr}(\gamma_5\Phi^m)}{m!}\\
&\overset{(3)}{=} \frac{\pi^m}{(2\pi)^{2m}\,(m-1)!}\,(m-1)!\,\big[f(0)-f(\infty)\big]\frac{\mathrm{tr}(\gamma_5\Phi^m)}{m!} = \frac{1}{(4\pi)^m}\frac{\mathrm{tr}(\gamma_5\Phi^m)}{m!}.
\end{aligned} \tag{10.128}$$

(1) Taylor expand $f(\kappa^2 - Y)$ to order m in Y and keep $(-\Phi)^m/M^{2m}$, as in Step 5; the factor is $f^{(m)}(\kappa^2)(-1)^m/m!$.
(2) Polar coordinates in $n = 2m$ dimensions with $u = \kappa^2$: $\mathrm{d}^n\kappa = \frac{\pi^m}{(m-1)!}u^{m-1}\mathrm{d}u$ (the unit sphere S^{2m-1} has area $2\pi^m/(m-1)!$, and $\kappa^{2m-1}\mathrm{d}\kappa = \frac{1}{2}u^{m-1}\mathrm{d}u$).
(3) Integrate by parts $m-1$ times; each time the boundary terms vanish (u^j kills $u = 0$, decay kills $u = \infty$), giving $\int_0^\infty u^{m-1}(-1)^m f^{(m)}\mathrm{d}u = (m-1)!\int_0^\infty(-f')\mathrm{d}u = (m-1)!\,[f(0) - f(\infty)]$.

Only $f(0) = 1$ and $f(\infty) = 0$ matter: the anomaly is insensitive to how the cutoff is shaped. It only knows that low modes are kept and high modes removed. For $f(u) = e^{-u}$ we recover (10.123); a sharp cutoff $f = \theta(1-u)$ or the Pauli–Villars choice $f(u) = 1/(1+u)$ give the same.

Physical picture: Pauli–Villars in one sentence

A Pauli–Villars regulator adds a heavy fermion of mass M with the wrong statistics. Its mass term $M\bar\chi\chi$ breaks the chiral symmetry explicitly, by the term $2iM\bar\chi\gamma_5\chi$, and as $M\to\infty$ the expectation value of that term does not go to zero: it tends to a finite limit, which reproduces $2\mathscr{A}$. The anomaly is the ghost of the regulator's mass, which cannot be removed because a mass is the only gauge-invariant way to make a fermion heavy.

10.4.4 QED in $2k$ dimensions

The computation (10.126) was done in every even dimension at once. For QED, $F=-ief$ and tr is trivial:

$$\begin{aligned}\langle\partial_\mu j_5^\mu\rangle &\overset{(1)}{=} \frac{2}{m!}\Big(\frac{i}{4\pi}\Big)^m(-ie)^m\,\epsilon^{\mu_1\cdots\mu_{2m}}f_{\mu_1\mu_2}\cdots f_{\mu_{2m-1}\mu_{2m}}\\ &\overset{(2)}{=} \frac{2}{m!}\Big(\frac{e}{4\pi}\Big)^m\,\epsilon^{\mu_1\cdots\mu_{2m}}f_{\mu_1\mu_2}\cdots f_{\mu_{2m-1}\mu_{2m}}\qquad\text{(Euclidean).}\end{aligned}\tag{10.129}$$

(1) (10.115) with (10.126) and $F_{\mu\nu}=-ief_{\mu\nu}$.
(2) $i^m(-i)^m=1$.

This is [QFT notes (22.4.1)], which in Minkowski signature carries an extra $(-1)^{m+1}$ (and writes k for our m). Table 10.2 lists the first cases.

$n=2m$	$\mathscr{A}\,\mathrm{d}^nx=\mathrm{ch}_m(F)$	$\langle\partial_\mu j_5^\mu\rangle$ (QED, Euclidean)	needs
2	$\frac{i}{2\pi}\,\mathrm{tr}\,F$	$\frac{e}{2\pi}\,\epsilon^{\mu\nu}f_{\mu\nu}=\frac{e}{\pi}f_{12}$	$\mathrm{tr}(\gamma_5\gamma\gamma)$
4	$-\frac{1}{8\pi^2}\,\mathrm{tr}\,F^2$	$\frac{e^2}{16\pi^2}\,\epsilon^{\mu\nu\rho\sigma}f_{\mu\nu}f_{\rho\sigma}$	$\mathrm{tr}(\gamma_5\gamma\gamma\gamma\gamma)$
6	$-\frac{i}{48\pi^3}\,\mathrm{tr}\,F^3$	$\frac{e^3}{192\pi^3}\,\epsilon^{\mu\nu\rho\sigma\alpha\beta}f_{\mu\nu}f_{\rho\sigma}f_{\alpha\beta}$	six γ's
$2m$	$\frac{1}{m!}\big(\frac{i}{2\pi}\big)^m\,\mathrm{tr}\,F^m$	$\frac{2}{m!}\big(\frac{e}{4\pi}\big)^m\epsilon f\cdots f$	$2m$ γ's

Table 10.2: The chiral anomaly in $n=2m$ dimensions. The anomaly is always the top Chern character form; its tensor structure is forced by the fact that γ_5 needs $2m$ other γ's to have a nonzero trace, and each F supplies two.

Worked example 10.13: Two dimensions, and agreement with the vacuum polarisation

For $m=1$, (10.126) gives

$$\mathscr{A}\overset{(1)}{=}\frac{i}{4\pi}\,\epsilon^{\mu\nu}F_{\mu\nu}\overset{(2)}{=}\frac{i}{4\pi}\,(F_{12}-F_{21})=\frac{i}{2\pi}F_{12}\overset{(3)}{=}\frac{e}{2\pi}f_{12}.$$

(1) (10.126) with $m=1$ and abelian F.
(2) $\epsilon^{12}=1=-\epsilon^{21}$.
(3) $F_{12}=-ief_{12}$.

So $\langle\partial_\mu j_5^\mu\rangle=\frac{e}{\pi}f_{12}$. Directly from the modes: in a constant field $f_{12}=B>0$ on a torus of area L^2 the Landau levels of $-\not{D}^2$ are $2eB\ell$, $\ell=0,1,2,\dots$, each with degeneracy $eBL^2/2\pi$, and the lowest level ($\ell=0$) consists only of positive-chirality zero modes (Example 10.4). All higher levels contain both chiralities equally. The regulated sum (10.116) is therefore the zero-mode density $(eBL^2/2\pi)/L^2=eB/2\pi$, for every M. This agrees with $\mathscr{A}=\frac{e}{2\pi}f_{12}$. In Minkowski form this is (10.57).

Worked example 10.14: Six dimensions

For $m = 3$: $\mathscr{A} = \frac{1}{3!}\left(\frac{i}{4\pi}\right)^3 \epsilon^{\mu\nu\rho\sigma\alpha\beta}\,\mathrm{tr}(F_{\mu\nu}F_{\rho\sigma}F_{\alpha\beta}) = -\frac{i}{384\pi^3}\epsilon\,\mathrm{tr}(FFF)$, and in form language $\mathscr{A}\,\mathrm{d}^6x = \mathrm{ch}_3(F) = -\frac{i}{48\pi^3}\,\mathrm{tr}\,F^3$ (check: $\mathrm{tr}\,F^3 = \frac{1}{8}\epsilon\,\mathrm{tr}(FFF)\mathrm{d}^6x$ and $\frac{1}{48\cdot 8} = \frac{1}{384}$). For QED, $(-ie)^3\cdot(-i) = -e^3$, so $\langle\partial_\mu j_5^\mu\rangle = 2\mathscr{A} = \frac{e^3}{192\pi^3}\epsilon fff$, the entry in Table 10.2. The six-dimensional anomaly matters in Section 10.5: it is the polynomial whose descent gives the four-dimensional non-abelian anomaly.

10.4.5 The integrated anomaly is the index

Now integrate (10.116) over M *before* taking $M \to \infty$:

$$\int_M \mathrm{d}^n x \sum_n \varphi_n^\dagger \gamma_5 e^{-\lambda_n^2/M^2}\varphi_n = \sum_n e^{-\lambda_n^2/M^2}\,\langle\varphi_n, \gamma_5\varphi_n\rangle. \tag{10.130}$$

Two facts about the modes do the rest.

1. *Nonzero modes pair up and cancel.* If $i\not{D}_A\varphi_n = \lambda_n\varphi_n$ with $\lambda_n \neq 0$, then

$$i\not{D}_A(\gamma_5\varphi_n) \overset{(1)}{=} -\gamma_5\, i\not{D}_A\varphi_n \overset{(2)}{=} -\lambda_n\,\gamma_5\varphi_n.$$

   (1) γ_5 anticommutes with every γ^μ and commutes with D_μ.
   (2) The eigenvalue equation.

   So $\gamma_5\varphi_n$ is an eigenvector with the *different* eigenvalue $-\lambda_n$, hence orthogonal to φ_n (eigenvectors of a self-adjoint operator with different eigenvalues are orthogonal): $\langle\varphi_n, \gamma_5\varphi_n\rangle = 0$.

2. *Zero modes have definite chirality.* γ_5 maps $\ker\not{D}_A$ to itself, so we may choose a basis of the kernel consisting of n_+ spinors with $\gamma_5 = +1$ and n_- with $\gamma_5 = -1$. For them $e^{-0} = 1$ and $\langle\varphi, \gamma_5\varphi\rangle = \pm 1$.

Hence, for *every* M,

$$\int_M \mathscr{A}\,\mathrm{d}^n x = n_+ - n_- = \mathrm{index}\,D_A^+, \tag{10.131}$$

and combining with (10.127),

$$\boxed{\int_M \langle\partial_\mu j_5^\mu\rangle\,\mathrm{d}^n x = 2\,\mathrm{index}\,D_A^+ = 2\int_M \mathrm{ch}_m(F)} \tag{10.132}$$

This is Nash's (10.3) in integrated-anomaly form [Nash p. 270]: the left side is the total change of chiral charge, and the right side is twice the number of forced zero modes. (On a curved M, the $\hat{A}$-genus appears as well; Section 10.7.) The argument is the physicist's proof of the index theorem in this setting: the left side of (10.131) is an integer for every M, and as $M \to \infty$ it becomes a local integral that we computed. (It is the McKean–Singer formula $\mathrm{index}\,D^+ = \mathrm{Tr}\,\gamma_5 e^{t\not{D}^2}$ of Chapter 4, at $t = 1/M^2$.)

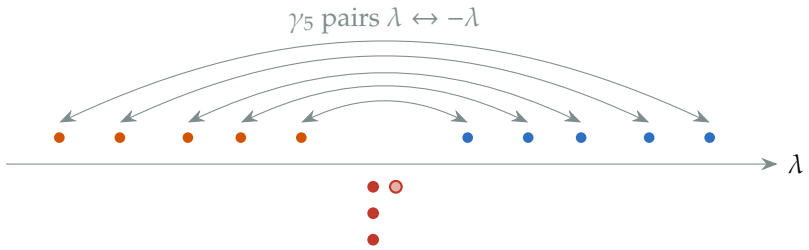


Figure 10.14: Why the integrated anomaly is an integer. The nonzero eigenvalues of $i\not{D}_A$ come in pairs $\pm\lambda$ related by γ_5 (grey arcs); each pair contributes $\langle\varphi, \gamma_5\varphi\rangle = 0$ to the regulated trace. Only the zero modes (red) survive, and they count $n_+ - n_-$. The local formula $\int \mathrm{ch}_m(F)$ must therefore give an integer.

Worked example 10.15: Flux on a torus: the anomaly counts Landau zero modes

On a two-torus with N units of flux, $\int \frac{e}{2\pi} f = N$. Then

$$\int \langle \partial_\mu j_5^\mu \rangle \, \mathrm{d}^2x \overset{(1)}{=} 2 \int \mathrm{ch}_1(F) \overset{(2)}{=} 2 \int \frac{i}{2\pi} F \overset{(3)}{=} 2 \int \frac{e}{2\pi} f = 2N.$$

(1) (10.132).
(2) $\mathrm{ch}_1 = \frac{i}{2\pi} \operatorname{tr} F$ for a line bundle.
(3) $F = -ief$.

On the other side, the lowest Landau level of Example 10.13 has N positive-chirality zero modes and nothing else, so index $D^+ = N$. Read in Minkowski time (the torus being the "periodic slab" of Section 10.3.5), $2N$ is the total change of Q_5 when N flux quanta are threaded: N steps of the spectral flow of Figure 10.8, each changing Q_5 by 2 in magnitude.

Worked example 10.16: The instanton flips chirality: 't Hooft's vertex

For the SU(2) BPST instanton on S^4 with one Dirac fermion in the fundamental representation, Example 10.5 gave index $D_A^+ = k = -1$: a single zero mode, of negative chirality. So

$$\int \langle \partial_\mu j_5^\mu \rangle \, \mathrm{d}^4x = 2k = -2.$$

In Minkowski language an instanton is a tunnelling event between gauge vacua, and in it the axial charge changes by $2|k|$ per Dirac flavour. With N_f flavours the path integral in the instanton background vanishes unless $2N_f$ fermion fields are inserted to soak up the zero modes: the **'t Hooft vertex**. In QCD this is why the axial U(1) is not a symmetry and the η' is heavy. In the electroweak theory, each of the 12 left-handed SU(2) doublets of the three generations (3×3 colours of quark doublets plus 3 lepton doublets) has one zero mode, so an electroweak instanton is a 12-fermion vertex with $\Delta B = \Delta L = 3$. Its amplitude carries the tunnelling factor $e^{-8\pi^2/g^2} = e^{-2\pi/\alpha_W} \approx 10^{-81}$, which is why protons do not decay this way.

Physical picture: Why the anomaly coefficient cannot be renormalised

The integrated anomaly is an integer for every background. A coefficient that depended on the coupling, $e^2/16\pi^2 \times (1 + c\,e^2 + \cdots)$, would make $\int \langle \partial j_5 \rangle$ non-integral for a unit instanton. So the one-loop coefficient is exact: this is the topological face of the Adler–Bardeen theorem mentioned at the end of Section 10.3.7.

10.4.6 The measure of a Weyl fermion and the covariant anomaly

So far the fermion was a Dirac fermion coupled *vectorially*, and the anomaly was in a global symmetry. Nash's concern is the chiral theory (10.13), in which only ψ_+ couples to A and the anomaly would be in the *gauge* symmetry itself. Here is how Fujikawa's method applies to it, and where it runs into trouble.

The difficulty. The operator in the action, $D_A^+ : \Gamma(E^+ \otimes F) \to \Gamma(E^- \otimes F)$, maps one space to another, so it has no eigenvalue problem, and there is no single basis in which to expand both ψ_+ and $\bar\psi_-$. We use two bases:

$$(D_A^+)^\dagger D_A^+ u_n = \mu_n u_n \text{ on } E^+, \qquad D_A^+ (D_A^+)^\dagger w_n = \mu_n w_n \text{ on } E^-, \tag{10.133}$$

orthonormal in their own spaces (for $\mu_n \neq 0$ one can take $w_n = D_A^+ u_n / \sqrt{\mu_n}$). Expand $\psi_+ = \sum a_n u_n$, $\bar\psi_- = \sum b_n w_n^\dagger$, and set $\mathcal{D}\psi_+ \mathcal{D}\bar\psi_- = \prod \mathrm{d}a_n \mathrm{d}b_n$.

The Jacobian of a gauge transformation. Let $g = e^{tv}$ with $v(x) \in \mathfrak{g}$. By (10.22), $Z_F[A_g] = \int \mathcal{D}\psi_+ \mathcal{D}\bar\psi_- \exp(-\int \bar\psi_- g^{-1} D_A^+ g \psi_+)$. Substitute $\chi = g\psi_+$, $\bar\chi = \bar\psi_- g^{-1}$; the action becomes $\int \bar\chi D_A^+ \chi$, and

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} \ln Z_F[A_{g(t)}]\Big|_{t=0} &\overset{(1)}{=} \frac{\mathrm{d}}{\mathrm{d}t} \ln \Big[\det(1 + tV^+)\, \det(1 - tV^-) \Big]_{t=0} \\ &\overset{(2)}{=} \operatorname{Tr} V^+ - \operatorname{Tr} V^- \overset{(3)}{=} \sum_n \int u_n^\dagger v\, u_n - \sum_n \int w_n^\dagger v\, w_n , \end{aligned} \tag{10.134}$$

where $V^+_{nk} = \int u_n^\dagger v u_k$ and $V^-_{nk} = \int w_n^\dagger v w_k$.

(1) As in (10.109): $\chi = (1+tv)\psi_+$ has coefficients $(1+tV^+)a$ and $\bar\chi = \bar\psi_-(1-tv)$ has coefficients $(1-tV^-)^{\mathsf{T}} b$; by Definition 10.6, $\mathcal{D}\psi_+\mathcal{D}\bar\psi_- = \det(1+tV^+)\det(1-tV^-)\,\mathcal{D}\chi\mathcal{D}\bar\chi$, and the χ-integral is $Z_F[A]$.
(2) $\frac{\mathrm{d}}{\mathrm{d}t} \ln\det(1+tV)|_0 = \operatorname{Tr} V$.
(3) Definitions of the matrix elements.

Regularise with the Dirac square. By (10.15), $\not{D}_A^2 = \operatorname{diag}\big(-(D_A^+)^\dagger D_A^+,\, -D_A^+ (D_A^+)^\dagger\big)$ on $E^+ \oplus E^-$, so cutting off each sum with $e^{-\mu_n/M^2}$ packages the difference of the two traces as a single trace with γ_5:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} \ln Z_F\Big|_{t=0} &\overset{(1)}{=} \lim_{M\to\infty} \int_M \mathrm{d}^n x\ \operatorname{tr} \langle x|\, v\, \gamma_5\, e^{\not{D}_A^2/M^2} |x\rangle \\ &\overset{(2)}{=} \int_M \frac{1}{m!} \Big(\frac{i}{4\pi}\Big)^m \epsilon^{\mu_1 \cdots \mu_{2m}} \operatorname{tr}\big(v\, F_{\mu_1\mu_2} \cdots F_{\mu_{2m-1}\mu_{2m}}\big) \mathrm{d}^n x = \int_M \operatorname{tr}\Big[v\, \frac{1}{m!} \Big(\frac{i}{2\pi}\Big)^m F^m \Big]. \end{aligned} \tag{10.135}$$

(1) $\gamma_5 = +1$ on E^+ and -1 on E^- gives the relative sign in (10.134); the colour matrix $v(x)$ is a multiplication operator and sits inside the trace.

(2) Steps 1–7 of Section 10.4.2 go through unchanged: v carries no γ's and no derivatives, so it just rides along inside the colour trace.

Comparing with (10.27) and writing $v = T^a f^a$,

$$\langle (D_\mu J^\mu)^a \rangle_{\text{cov}}\, \mathrm{d}^n x = \operatorname{tr}\Big[T^a \, \frac{1}{m!} \Big(\frac{i}{2\pi} \Big)^m F^m \Big]. \tag{10.136}$$

This is the **covariant anomaly**: it is built from F and transforms under gauge transformations like F does. For $G = \mathrm{U}(1)$ with charge e it is exactly half of the Dirac-fermion ABJ anomaly, as it should be: $j_+ = \frac{1}{2}(j + j_5)$.

Watch out: The covariant anomaly is not the variation of anything

The regulator in (10.135) uses $e^{-(D^+)^\dagger D^+/M^2}$ on ψ_+ and $e^{-D^+(D^+)^\dagger/M^2}$ on $\bar\psi_-$: two *different* A-dependent cutoffs on two different spaces (the intuition box of Section 10.2.7). Because the definition of the measure itself depends on A, the Jacobian computed this way is not the derivative of a single well-defined functional $\ln Z_F[A]$. We shall see in Section 10.5.2 that (10.135) fails the integrability condition that any true variation $\delta_v \ln Z_F$ must satisfy. The true variation, the **consistent anomaly**, has the same topological content but a different normalisation: for U(1) in $n = 2m$ dimensions it is $\frac{1}{m+1}$ of the covariant one.

Dictionary: three anomalies with similar names

- *ABJ (axial, singlet) anomaly*: $\partial_\mu j_5^\mu = 2\mathscr{A}$ for a Dirac fermion; a global symmetry fails; no inconsistency.
- *Covariant gauge anomaly*: $\langle (D \cdot J)^a \rangle$ of a chiral gauge current computed with covariant regulators, (10.136); gauge covariant, not integrable.
- *Consistent gauge anomaly*: $\delta_v \ln Z_F$ for a genuinely defined Z_F; satisfies the Wess–Zumino condition; computed by descent in Section 10.5.5. This is Nash's μ_1 (10.13) and (10.33).

Reading Nash (p. 273, Fujikawa; p. 270, (10.3))

Nash's one-line reference to Fujikawa is Sections 10.4.1 and 10.4.2: "the Fermionic 'measure' $\mathcal{D}\psi\mathcal{D}\bar\psi$ is not gauge invariant" is (10.111) with $C \neq 0$, and the non-invariance is computed by the heat kernel as the Chern character (10.127). Integrating gives back his index formula (10.3), now in the form (10.132). For the chiral theory the same Jacobian gives the covariant anomaly (10.136); Nash's cohomological μ_1 is the consistent version, which we derive next.

Recap

- Expand fields in eigenmodes of $i\not{D}_A$. A local chiral rotation changes the Grassmann measure by $\det(1 + C)^{-2} = \exp[-2i \int \alpha\, \mathscr{A}]$, so $\langle \partial_\mu j_5^\mu \rangle = 2\mathscr{A}$.
- With the gauge-covariant heat-kernel regulator, $\mathscr{A} = \lim \operatorname{tr}\langle x | \gamma_5 e^{\not{D}_A^2/M^2} | x \rangle = \frac{1}{m!}(\frac{i}{4\pi})^m \epsilon \operatorname{tr}(F \cdots F)$, i.e. $\mathscr{A}\, \mathrm{d}^n x = \mathrm{ch}_m(F)$. Only the Φ^m term survives, because $\operatorname{tr} \gamma_5$ needs $2m$ γ's.
- The result does not depend on the shape of the cutoff, only on $f(0) = 1$, $f(\infty) = 0$.
- QED_{2k}: $\langle \partial j_5 \rangle = \frac{2}{m!}(\frac{e}{4\pi})^m \epsilon f \cdots f$; $\frac{e}{\pi} f_{12}$ in 2D, $\frac{e^2}{16\pi^2} \epsilon f f$ in 4D.

- Integrated: $\int \mathscr{A} = n_+ - n_-$ for every M (paired modes cancel), so $\int \langle \partial j_5 \rangle = 2\,\mathrm{index}\, D_A^+ = 2\int \mathrm{ch}_m(F)$.
- A Weyl fermion with covariant regulators gives the covariant anomaly $\mathrm{tr}[T^a\,\mathrm{ch}_m(F)]$; the consistent anomaly differs and is the subject of the next section.

10.5 Anomalies and Yang–Mills theories IV: non-abelian anomalies, consistency and descent

So far the anomaly has been a statement about a U(1) current. Nash's set-up [Nash pp. 270–273] is a chiral fermion coupled to a general compact gauge group G, and for such a theory the anomaly is not just a nuisance: it is an inconsistency, because the gauge current is the current the gauge field couples to. The question this section answers is: *what exactly is* $\delta_v \ln Z_F[A]$ *for a non-abelian chiral gauge theory, and when does it vanish?* Three facts will do the work.

1. *Algebra.* Gauge transformations form a group, so the variations δ_v must close into a Lie algebra. This forces the anomaly to obey the **Wess–Zumino consistency condition**. The covariant anomaly of Section 10.4.6 fails it.
2. *Topology.* The solutions of the consistency condition are produced by the **descent equations** from the characteristic class $\mathrm{ch}_{m+1}(F)$ in two dimensions *more* than spacetime. This is the origin of the Chern–Simons forms and of Nash's formula (10.33).
3. *One number.* The overall coefficient is fixed by a single abelian graph, and Bose symmetry gives it: $\frac{1}{m+1}$ of the covariant anomaly.

The upshot is a purely group-theoretic cancellation condition, $\mathrm{Str}(T^aT^bT^c) = 0$ in four dimensions, which the Standard Model satisfies generation by generation.

10.5.1 Toolkit: Lie-algebra-valued forms and their traces

We use the conventions of Section 10.1.1 (and Chapter T3): $A = A_\mu \mathrm{d}x^\mu$ is a 1-form whose components are anti-Hermitian matrices in the representation carried by the fermions, $F = \mathrm{d}A + A \wedge A$, and a finite gauge transformation is $A \mapsto g^{-1}Ag + g^{-1}\mathrm{d}g$. We usually drop the wedge sign: $A^2 \coloneqq A \wedge A$, $A\mathrm{d}A \coloneqq A \wedge \mathrm{d}A$. A gauge parameter is a Lie-algebra-valued function $v(x) = T^a v^a(x)$, and the infinitesimal gauge transformation is (10.23),

$$\delta_v A = D_A v = \mathrm{d}v + [A, v] = \mathrm{d}v + Av - vA. \tag{10.137}$$

(The physicist's dictionary: $A = -ig\,A^a_{\mathrm{phys}} t^a$ with t^a Hermitian, $T^a = -it^a$; for U(1) with charge e, $A = -ie\,a$.)

Graded rules. A matrix-valued p-form times a matrix-valued q-form is a matrix of $(p+q)$-forms. Two rules are used constantly.

- *Graded commutator.* $[X, Y] \coloneqq XY - (-1)^{pq}YX$. For two 1-forms it is an *anti*commutator: $[A, A] = 2A^2$, which is why $A^2 \neq 0$ although $\mathrm{d}x^\mu \wedge \mathrm{d}x^\mu = 0$.
- *Graded cyclicity.* $\mathrm{tr}(XY) = (-1)^{pq}\,\mathrm{tr}(YX)$.

The second rule is worth one line of proof, for two 1-forms:

$$\operatorname{tr}(XY) \overset{(1)}{=} \operatorname{tr}(X_\mu Y_\nu)\,\mathrm{d}x^\mu \wedge \mathrm{d}x^\nu \overset{(2)}{=} \operatorname{tr}(Y_\nu X_\mu)\,\mathrm{d}x^\mu \wedge \mathrm{d}x^\nu \overset{(3)}{=} -\operatorname{tr}(Y_\nu X_\mu)\,\mathrm{d}x^\nu \wedge \mathrm{d}x^\mu = -\operatorname{tr}(YX).$$

(1) Write out the components; the matrices and the forms are multiplied separately.
(2) Ordinary cyclicity of the matrix trace.
(3) Swap the two 1-forms: $\mathrm{d}x^\mu \wedge \mathrm{d}x^\nu = -\mathrm{d}x^\nu \wedge \mathrm{d}x^\mu$.

In general the matrices contribute no sign and the forms contribute $(-1)^{pq}$. For example $\operatorname{tr}(A^4) = \operatorname{tr}(A \cdot A^3) = (-1)^{1\cdot 3}\operatorname{tr}(A^3 \cdot A) = -\operatorname{tr}(A^4)$, so $\operatorname{tr} A^{2k} = 0$ for every $k \geq 1$. The Leibniz rule is $\mathrm{d}(XY) = (\mathrm{d}X)Y + (-1)^p X\,\mathrm{d}Y$.

How F transforms, and the Bianchi identity. Two short chains that we use again and again:

$$\begin{aligned}
\delta_v F &\overset{(1)}{=} \mathrm{d}(\delta_v A) + (\delta_v A)A + A(\delta_v A) \\
&\overset{(2)}{=} \big(\mathrm{d}A\, v - A\,\mathrm{d}v - \mathrm{d}v\, A - v\,\mathrm{d}A\big) \\
&\qquad + \big(\mathrm{d}v\, A + AvA - vA^2\big) + \big(A\,\mathrm{d}v + A^2 v - AvA\big) \\
&\overset{(3)}{=} (\mathrm{d}A + A^2)v - v(\mathrm{d}A + A^2) = [F, v], \qquad (10.138)
\end{aligned}$$

$$\begin{aligned}
D_A F \coloneqq \mathrm{d}F + [A, F] &\overset{(4)}{=} \big(\mathrm{d}A\, A - A\,\mathrm{d}A\big) \\
&\qquad + \big(A\,\mathrm{d}A + A^3 - \mathrm{d}A\, A - A^3\big) = 0. \qquad (10.139)
\end{aligned}$$

(1) Vary $F = \mathrm{d}A + A^2$: $\delta(A^2) = (\delta A)A + A(\delta A)$ (no sign, as δ does not change form degree).
(2) Insert $\delta_v A = \mathrm{d}v + Av - vA$; use $\mathrm{d}(Av) = \mathrm{d}A\, v - A\,\mathrm{d}v$ and $\mathrm{d}(vA) = \mathrm{d}v\, A + v\,\mathrm{d}A$.
(3) The $\mathrm{d}v$ terms cancel in pairs and so do the AvA terms.
(4) $\mathrm{d}F = \mathrm{d}(A^2) = \mathrm{d}A\, A - A\,\mathrm{d}A$, and $[A, F] = AF - FA$ (a 1-form and a 2-form).

So under gauge transformations F rotates, $\delta_v F = [F, v]$, like the finite law $F \mapsto g^{-1}Fg$; it does not shift.

Invariant polynomials. The traces $\operatorname{tr} F^k$ are closed and gauge invariant:

$$\begin{aligned}
\delta_v \operatorname{tr} F^k &\overset{(1)}{=} \sum_j \operatorname{tr}\big(F^j [F, v] F^{k-1-j}\big) \overset{(2)}{=} \operatorname{tr}\big(F^k v - v F^k\big) \overset{(3)}{=} 0, \\
\mathrm{d} \operatorname{tr} F^k &\overset{(4)}{=} -\sum_j \operatorname{tr}\big(F^j [A, F] F^{k-1-j}\big) \overset{(5)}{=} -\operatorname{tr}\big(AF^k - F^k A\big) \overset{(6)}{=} 0. \qquad (10.140)
\end{aligned}$$

(1) Leibniz for δ_v and (10.138).
(2) The sum telescopes: $F^j(Fv - vF)F^{k-1-j} = F^{j+1}vF^{k-1-j} - F^j v F^{k-j}$.
(3) Cyclicity; v is a 0-form, so no sign.
(4) Bianchi (10.139): $\mathrm{d}F = -[A, F]$; F is even, so d passes it without sign.
(5) Telescoping again.
(6) Graded cyclicity: $\operatorname{tr}(F^k A) = (-1)^{2k\cdot 1}\operatorname{tr}(AF^k)$.

This is the Chern–Weil theorem in its computational form (Chapter T2): $\operatorname{tr} F^k$ defines a de Rham class, and (10.140) is why the Chern character $\operatorname{ch}(F) = \operatorname{tr}\exp(iF/2\pi)$ is built from these traces.

Worked example 10.17: SU(2) in the mathematicians' convention

Take $T^a = -\frac{i}{2}\sigma^a$. Then

$$[T^1,T^2] \overset{(1)}{=} \left(-\frac{i}{2}\right)^2 [\sigma^1,\sigma^2] \overset{(2)}{=} -\frac{1}{4}\cdot 2i\sigma^3 \overset{(3)}{=} -\frac{i}{2}\sigma^3 = T^3,$$

(1) Bilinearity of the commutator.
(2) $\sigma^1\sigma^2 = i\sigma^3 = -\sigma^2\sigma^1$, so $[\sigma^1,\sigma^2] = 2i\sigma^3$.
(3) Rewrite.

and cyclically, so $[T^a,T^b] = \epsilon^{abc}T^c$: the structure constants are $f^{abc} = \epsilon^{abc}$. With $A = A^aT^a$,

$$F = \mathrm{d}A^a\,T^a + A^bA^c\,T^bT^c = \left(\mathrm{d}A^a + \tfrac{1}{2}\epsilon^{abc}A^bA^c\right)T^a,$$

because A^bA^c is antisymmetric in b,c (two 1-forms), so only $\frac{1}{2}[T^b,T^c]$ survives. The symmetric invariant that will control the anomaly is

$$d^{abc} \coloneqq \operatorname{tr}\big(T^a\{T^b,T^c\}\big) \overset{(1)}{=} \left(-\frac{i}{2}\right)^3 \operatorname{tr}\big(\sigma^a\{\sigma^b,\sigma^c\}\big) \overset{(2)}{=} \frac{i}{8}\cdot 2\delta^{bc}\operatorname{tr}\sigma^a \overset{(3)}{=} 0.$$

(1) Factor out the constants.
(2) $\{\sigma^b,\sigma^c\} = 2\delta^{bc}\mathbf{1}$, and $(-i/2)^3 = i/8$.
(3) Pauli matrices are traceless.

We shall see that this one line is why SU(2) gauge theories never have perturbative anomalies [QFT notes (22.7.30)].

10.5.2 The consistent anomaly and the Wess–Zumino condition

Let $Z_F[A]$ be the chiral fermion integral of Section 10.2, however it is regularised, provided only that it is a genuine functional of A. Define the **consistent anomaly** as its gauge variation:

$$\mathcal{A}[v;A] \coloneqq \delta_v \ln Z_F[A] \coloneqq \frac{\mathrm{d}}{\mathrm{d}t}\ln Z_F[A+tD_Av]\Big|_{t=0} = \left\langle \int_M v^a (D_\mu J^\mu)^a \right\rangle_A, \tag{10.141}$$

the last equality being (10.27). It is linear in v, and we write $\mathcal{A}[v] = \int_M v^a\,G^a$ with $G^a(x) = \langle (D_\mu J^\mu)^a(x)\rangle$.

Step 1: gauge variations close. Think of δ_v as a vector field on the space $\mathcal{A}$ of connections, $X_v = \int (D_Av)^a_\mu\,\delta/\delta A^a_\mu$, acting on functionals $\Phi[A]$. Then for any Φ

$$\begin{aligned}
\delta_v\big(\delta_w\Phi\big) &\overset{(1)}{=} \frac{\mathrm{d}}{\mathrm{d}s}\left[\int (D_{A+sD_Av}w)\cdot\frac{\delta\Phi}{\delta A}\Big|_{A+sD_Av}\right]_{s=0} \\
&\overset{(2)}{=} \int [D_Av,w]\cdot\frac{\delta\Phi}{\delta A} + \iint (D_Aw)(D_Av)\cdot\frac{\delta^2\Phi}{\delta A\,\delta A}, \\
[\delta_v,\delta_w]\Phi &\overset{(3)}{=} \int \big([D_Av,w]-[D_Aw,v]\big)\cdot\frac{\delta\Phi}{\delta A} \overset{(4)}{=} \int D_A[v,w]\cdot\frac{\delta\Phi}{\delta A} = \delta_{[v,w]}\Phi.
\end{aligned} \tag{10.142}$$

(1) $\delta_w\Phi[A] = \int (D_Aw)\cdot\delta\Phi/\delta A$ is itself a functional of A; now move A in the direction D_Av.
(2) $D_{A+sD_Av}w = D_Aw + s[D_Av,w]$; the derivative of $\delta\Phi/\delta A$ gives the second functional derivative.
(3) Antisymmetrise in $v \leftrightarrow w$; the second-derivative term is symmetric and drops out.

(4) D_A is a derivation of the bracket: $D_A[v,w] = [D_A v, w] + [v, D_A w]$, and $[v, D_A w] = -[D_A w, v]$.

This is the infinitesimal form of the group law $(A_g)_h = A_{gh}$.

Step 2: the consistency condition. Apply (10.142) to $\Phi = \ln Z_F$:

$$\delta_v \mathcal{A}[w] - \delta_w \mathcal{A}[v] \overset{(1)}{=} \delta_v \delta_w \ln Z_F - \delta_w \delta_v \ln Z_F \overset{(2)}{=} \delta_{[v,w]} \ln Z_F \overset{(3)}{=} \mathcal{A}\big[[v,w]\big]. \tag{10.143}$$

(1) Definition (10.141), for parameters v, w that do not depend on A.
(2) (10.142).
(3) Definition again.

Equation (10.143) is the **Wess–Zumino consistency condition**. It uses nothing about fermions: *any* anomaly that is the variation of *some* functional must obey it.

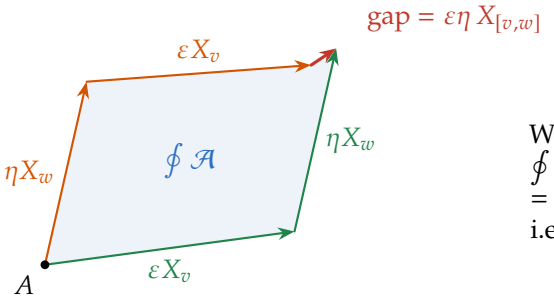
Working definition 10.7 (consistent anomaly)

A **consistent anomaly** is a local expression $\mathcal{A}[v;A] = \int_M v^a G^a(A)$ [a candidate for $\delta_v \ln Z_F$, the change of $\ln Z_F$ under the infinitesimal gauge transformation with parameter v], linear in v, built from A and its derivatives, that satisfies (10.143) [the Wess–Zumino condition $\delta_v \mathcal{A}[w] - \delta_w \mathcal{A}[v] = \mathcal{A}[[v,w]]$]. *To test a candidate:* compute $\delta_v \mathcal{A}[w] - \delta_w \mathcal{A}[v]$ using $\delta_v A = D_A v$ and $\delta_v F = [F, v]$, integrate by parts, and compare with $\mathcal{A}[[v,w]]$. It is **trivial** if $\mathcal{A}[v] = \delta_v \Phi_{\mathrm{loc}}[A]$ for a local functional Φ_{loc}; a trivial anomaly is removed by adding the counterterm [a local term added by hand, allowed because it only changes the regularisation] $-\Phi_{\mathrm{loc}}$ to $\ln Z_F$.

Geometric meaning: a closed 1-form. On a manifold, a 1-form μ and two vector fields obey $\mathrm{d}\mu(X,Y) = X\big(\mu(Y)\big) - Y\big(\mu(X)\big) - \mu([X,Y])$ (Chapter T1). Regard $\mathcal{A}$ as the 1-form on a gauge orbit whose value on the vector field X_v is $\mathcal{A}[v]$. Since $[X_v, X_w] = X_{[v,w]}$ by (10.142),

$$\mathrm{d}\mathcal{A}(X_v, X_w) = \delta_v \mathcal{A}[w] - \delta_w \mathcal{A}[v] - \mathcal{A}\big[[v,w]\big],$$

so *the Wess–Zumino condition says the anomaly is a closed* 1*-form on the gauge orbit*. A trivial anomaly is an *exact* one, $\mathcal{A} = \mathrm{d}\Phi_{\mathrm{loc}}$. This is precisely the cohomological statement of Nash [Nash p. 273]: the variation of $\ln Z_F$ "can be identified with an element of $H^1_{\mathrm{dR}}(\mathcal{G})$"; we return to it in Section 10.6.2.



Wess–Zumino:
$\oint \mathcal{A}$ round the small loop
$= \mathcal{A}$ along the gap,
i.e. $\mathrm{d}\mathcal{A} = 0$ on the orbit.

Figure 10.15: Two infinitesimal gauge transformations done in the two possible orders. The end points differ by the gauge transformation $[v,w]$ (red), because gauge transformations do not commute. The change of $\ln Z_F$ must be the same along both routes plus the gap; that bookkeeping is the Wess–Zumino condition (10.143).

Worked example 10.18: The covariant anomaly is not consistent

In two dimensions the covariant anomaly of (10.135) is $\mathcal{A}_{\mathrm{cov}}[v] = \frac{i}{2\pi}\int_M \mathrm{tr}(vF)$. Test it:

$$\delta_w \mathcal{A}_{\mathrm{cov}}[v] \overset{(1)}{=} \frac{i}{2\pi}\int \mathrm{tr}\big(v[F,w]\big) \overset{(2)}{=} \frac{i}{2\pi}\int \mathrm{tr}\big(vFw - vwF\big) \overset{(3)}{=} \frac{i}{2\pi}\int \mathrm{tr}\big([w,v]F\big),$$

$$\delta_v \mathcal{A}_{\mathrm{cov}}[w] - \delta_w \mathcal{A}_{\mathrm{cov}}[v] \overset{(4)}{=} \frac{i}{2\pi}\int \mathrm{tr}\big([v,w]F - [w,v]F\big)$$

$$\overset{(5)}{=} 2\cdot\frac{i}{2\pi}\int \mathrm{tr}\big([v,w]F\big) = 2\,\mathcal{A}_{\mathrm{cov}}\big[[v,w]\big].$$

(1) v does not depend on A; only F varies, by (10.138).
(2) Expand the commutator.
(3) Cyclicity: $\mathrm{tr}(vFw) = \mathrm{tr}(wvF)$.
(4) Use line 3 twice, with $v \leftrightarrow w$.
(5) $[w,v] = -[v,w]$.

The left side is *twice* the right side of (10.143). For G abelian both sides are zero and there is no contradiction; for non-abelian G the covariant anomaly cannot be the variation of any functional. It is not wrong: it is the divergence of a different current (the covariant current, which is not $\delta \ln Z_F/\delta A$).

Worked example 10.19: A consistent anomaly in two dimensions

Now test $\mathcal{A}_2[v] = \frac{i}{4\pi}\int_M \mathrm{tr}(v\,\mathrm{d}A)$. Since $\mathrm{d}\mathrm{d}w = 0$, $\delta_w(\mathrm{d}A) = \mathrm{d}[A,w]$. First, a useful integration by parts: for 0-forms v, w,

$$\int \mathrm{tr}\big(w\{A,\mathrm{d}v\}\big) \overset{(1)}{=} \int \big[\mathrm{tr}(\mathrm{d}w\,A\,v) - \mathrm{tr}(\mathrm{d}w\,v\,A)\big] + \int \big[\mathrm{tr}(w\,\mathrm{d}A\,v) - \mathrm{tr}(w\,v\,\mathrm{d}A)\big]$$

$$\overset{(2)}{=} \int \mathrm{tr}\big(v\{A,\mathrm{d}w\}\big) + \int \mathrm{tr}\big([v,w]\,\mathrm{d}A\big). \tag{10.144}$$

(1) $\{A,\mathrm{d}v\} = A\,\mathrm{d}v + \mathrm{d}v\,A$ (graded commutator of two 1-forms). Integrate each term by parts using the Leibniz rule: $\mathrm{d}\,\mathrm{tr}(wAv) = \mathrm{tr}(\mathrm{d}w\,Av) + \mathrm{tr}(w\,\mathrm{d}A\,v) - \mathrm{tr}(wA\,\mathrm{d}v)$ and $\mathrm{d}\,\mathrm{tr}(wvA) = \mathrm{tr}(\mathrm{d}w\,vA) + \mathrm{tr}(w\,\mathrm{d}v\,A) + \mathrm{tr}(wv\,\mathrm{d}A)$; total derivatives integrate to zero on closed M (Stokes).
(2) Cyclicity: $\mathrm{tr}(\mathrm{d}w\,Av) = \mathrm{tr}(v\,\mathrm{d}w\,A)$ and $\mathrm{tr}(\mathrm{d}w\,vA) = -\mathrm{tr}(A\,\mathrm{d}w\,v) = -\mathrm{tr}(vA\,\mathrm{d}w)$ (moving the 1-form A past the 1-form $\mathrm{d}w\,v$); and $\mathrm{tr}(w\,\mathrm{d}A\,v) - \mathrm{tr}(wv\,\mathrm{d}A) = \mathrm{tr}((vw-wv)\mathrm{d}A)$.

Next, expand the variation:

$$\mathrm{tr}\big(v\,\mathrm{d}[A,w]\big) \overset{(1)}{=} \mathrm{tr}\big(v(\mathrm{d}A\,w - A\,\mathrm{d}w - \mathrm{d}w\,A - w\,\mathrm{d}A)\big) \overset{(2)}{=} \mathrm{tr}\big([w,v]\mathrm{d}A\big) - \mathrm{tr}\big(v\{A,\mathrm{d}w\}\big).$$

(1) Leibniz: $\mathrm{d}(Aw) = \mathrm{d}A\,w - A\,\mathrm{d}w$ and $\mathrm{d}(wA) = \mathrm{d}w\,A + w\,\mathrm{d}A$.
(2) $\mathrm{tr}(v\,\mathrm{d}A\,w) = \mathrm{tr}(wv\,\mathrm{d}A)$ by cyclicity; collect.

So $\delta_w\mathcal{A}_2[v] = \frac{i}{4\pi}\int[\mathrm{tr}([w,v]\mathrm{d}A) - \mathrm{tr}(v\{A,\mathrm{d}w\})]$, and

$$\delta_v\mathcal{A}_2[w] - \delta_w\mathcal{A}_2[v] \overset{(1)}{=} \frac{i}{4\pi}\int\Big[2\,\mathrm{tr}([v,w]\mathrm{d}A) - \mathrm{tr}(w\{A,\mathrm{d}v\}) + \mathrm{tr}(v\{A,\mathrm{d}w\})\Big]$$

$$\overset{(2)}{=} \frac{i}{4\pi}\int\Big[2\,\mathrm{tr}([v,w]\mathrm{d}A) - \mathrm{tr}([v,w]\mathrm{d}A)\Big]$$

$$= \frac{i}{4\pi}\int \mathrm{tr}\big([v,w]\mathrm{d}A\big) = \mathcal{A}_2\big[[v,w]\big].$$

(1) The previous line with $v \leftrightarrow w$, subtracted; $[w,v] \to [v,w]$ gives the factor 2.
(2) (10.144).

It passes. Note what happened: $\mathcal{A}_2$ is *not* gauge covariant (dA is not F), and that is the price of consistency. In the abelian case $\mathcal{A}_2 = \frac{1}{2}\mathcal{A}_{\text{cov}}$.

10.5.3 Consistent versus covariant: Bose symmetry fixes the factor

Why should the consistent anomaly be *half* of the covariant one in two dimensions? Because in the consistent version every current in a correlation function is treated alike. We check this against explicit graphs in two and four dimensions.

Worked example 10.20: The chiral Schwinger model: one Weyl fermion in two dimensions

Use the Minkowski conventions of (10.44) and couple only the positive-chirality current $j_R^\mu := j_+^\mu = \frac{1}{2}(j^\mu + j_5^\mu)$ of (10.28) to A, with interaction $-eA_\mu j_R^\mu$. (Subscript R is the chirality; superscripts ± below are light-cone components.)
Light-cone components. For a vector V put $V^\pm = V^0 \pm V^1$ and $V_\pm = \frac{1}{2}(V_0 \pm V_1)$. Then

$$V\cdot W = V^0W^0 - V^1W^1 \overset{(1)}{=} \tfrac{1}{2}\big(V^+W^- + V^-W^+\big) \overset{(2)}{=} V_+W^+ + V_-W^-,$$
$$\epsilon^{\mu\nu}q_\mu A_\nu = q_0A_1 - q_1A_0 \overset{(3)}{=} 2\big(q_-A_+ - q_+A_-\big), \qquad q^\mu A_\mu \overset{(4)}{=} 2\big(q_-A_+ + q_+A_-\big).$$

(1) Expand $(V^0+V^1)(W^0-W^1) + (V^0-V^1)(W^0+W^1) = 2(V^0W^0 - V^1W^1)$.
(2) $V_\pm = \frac{1}{2}(V_0 \pm V_1) = \frac{1}{2}(V^0 \mp V^1) = \frac{1}{2}V^\mp$, using $V_1 = -V^1$.
(3) Substitute $q_0 = q_+ + q_-$, $q_1 = q_+ - q_-$ and similarly for A, and expand.
(4) Line (2) with $V = q$, $W = A$ read the other way round: $q^\mu A_\mu = q_+A^+ + q_-A^-$, and $A^\pm = 2A_\mp$.

In this basis $\eta^{++} = \eta^{--} = 0$, $\eta^{+-} = \eta^{-+} = 2$, and $q^2 = q^+q^-$.
Only one component couples. From $j_5^\mu = -\epsilon^{\mu\nu}j_\nu$ (10.45), $j_5^0 = -\epsilon^{01}j_1 = j^1$ and $j_5^1 = -\epsilon^{10}j_0 = j^0$, so

$$j_5^+ = j^1 + j^0 = j^+, \qquad j_5^- = j^1 - j^0 = -j^-, \qquad \text{hence} \qquad j_R^+ = j^+, \quad j_R^- = 0.$$

The chiral current has a single light-cone component, and the interaction is $-eA_\mu j_R^\mu = -eA_+j^+$. This is the interaction of an ordinary *Dirac* fermion in the background $\hat{A}$ with $\hat{A}_+ = A_+$, $\hat{A}_- = 0$: the negative-chirality fermions do not see $\hat{A}$ at all.
The induced current. So we may borrow the Dirac vacuum polarisation (10.46) with its unknown a:

$$\langle j_R^+\rangle \overset{(1)}{=} -\frac{1}{e}\,\Pi^{+\nu}\hat{A}_\nu \overset{(2)}{=} -\frac{1}{e}\,\Pi^{++}A_+ \overset{(3)}{=} -\frac{1}{e}\Big(\mathsf{a}\,\eta^{++} - \mathsf{b}\,\frac{q^+q^+}{q^+q^-}\Big)A_+ \overset{(4)}{=} \frac{\mathsf{b}}{e}\,\frac{q^+}{q^-}\,A_+.$$

(1) Linear response, as in Section 10.3.3.
(2) $\hat{A}_- = 0$, and $\Pi^{+\nu}\hat{A}_\nu = \Pi^{++}\hat{A}_+ + \Pi^{+-}\hat{A}_-$.
(3) (10.46) in light-cone components.
(4) $\eta^{++} = 0$: *the ambiguous coefficient* a *drops out*.

Because this current is the functional derivative of the (unambiguous) effective action of A_+, it is the consistent current. Its divergence is

$$q_\mu\langle j_R^\mu\rangle \overset{(1)}{=} q_+\langle j_R^+\rangle \overset{(2)}{=} \frac{q^-}{2}\,\frac{\mathsf{b}}{e}\,\frac{q^+}{q^-}A_+ \overset{(3)}{=} \frac{\mathsf{b}}{e}\,q_-A_+ \overset{(4)}{=} \frac{e}{4\pi}\big(\epsilon^{\mu\nu}q_\mu A_\nu + q^\mu A_\mu\big).$$

(1) $q \cdot j = q_+ j^+ + q_- j^-$ and $j_R^- = 0$.
(2) $q_+ = \frac{1}{2} q^-$.
(3) $q^+ = 2 q_-$.
(4) Add the two light-cone formulas: $\epsilon q A + q \cdot A = 4 q_- A_+$; and $\mathsf{b} = e^2/\pi$ (Example 10.10).

Compare with the covariant anomaly. The gauge-invariant Dirac scheme $\mathsf{a} = \mathsf{b}$ gives $q \cdot \langle j \rangle = 0$ and $q \cdot \langle j_5 \rangle = \frac{\mathsf{b}}{e} \epsilon q A$ (10.47), so $q \cdot \langle j_R \rangle_{\text{cov}} = \frac{1}{2}(0 + \frac{\mathsf{b}}{e} \epsilon q A) = \frac{e}{2\pi} \epsilon q A$. The two results differ in two ways.

- The term $\frac{e}{4\pi} q^\mu A_\mu$ is *trivial*: under $\delta A_\mu = \partial_\mu \lambda$ the local functional $\int A_\mu A^\mu$ changes by $2 \int A^\mu \partial_\mu \lambda = -2 \int \lambda \, \partial \cdot A$, so a counterterm $c \int A_\mu A^\mu$ with suitable real c removes it. (It uses A_-, which the fermions do not see but which is part of the gauge field.)
- The term $\frac{e}{4\pi} \epsilon^{\mu\nu} q_\mu A_\nu$ is *not* removable: the only local, Lorentz-invariant quadratic functionals of A of the right dimension are $\int A_\mu A^\mu$ and $\int \epsilon^{\mu\nu} A_\mu A_\nu = 0$.

So, after the best local counterterm,

$$\partial_\mu \langle j_R^\mu \rangle_{\text{consistent}} = \frac{1}{2} \partial_\mu \langle j_R^\mu \rangle_{\text{cov}} = \frac{e}{4\pi} \epsilon^{\mu\nu} \partial_\mu A_\nu .$$

The mechanism is visible in the graph: the vacuum polarisation has *two* identical chiral vertices, and the non-removable part of the anomaly is shared equally between them.

Worked example 10.21: The factor $\frac{1}{3}$ from the triangle budget

In four dimensions, couple a single Weyl fermion to a U(1) field through $\gamma^\mu P_+$ at every vertex. In a triangle,

$$\begin{aligned} \gamma^\lambda P_+ S(p_1) \gamma^\nu P_+ S(p_2) \gamma^\mu P_+ S(p_3) &\overset{(1)}{=} \gamma^\lambda S(p_1) P_- \gamma^\nu P_+ S(p_2) \gamma^\mu P_+ S(p_3) \\ &\overset{(2)}{=} \gamma^\lambda S(p_1) \gamma^\nu P_+ S(p_2) \gamma^\mu P_+ S(p_3) \\ &\overset{(3)}{=} \gamma^\lambda S(p_1) \gamma^\nu S(p_2) \gamma^\mu S(p_3) \, P_- \\ &\to \tfrac{1}{2}[\text{vector}] - \tfrac{1}{2}[\text{one-}\gamma^5]. \end{aligned}$$

(1) $S(p) = \not{p}/p^2$ contains one γ, so $P_+ S = S P_-$.
(2) $P_- \gamma^\nu = \gamma^\nu P_+$, and $P_+^2 = P_+$.
(3) Repeat to push all projectors to the right; they collapse to one $P_- = \frac{1}{2}(1 - \gamma^5)$ (a single $P_\pm$ in either case; which one depends on the position we push to).

So the part that can be anomalous is $\frac{1}{2}$ of the one-γ^5 triangle $T^{\mu\nu\lambda}$ of (10.79), with the γ^5 now at no preferred vertex. In (10.96) write each vertex's anomaly in a common form: with all momenta incoming (q at the axial vertex, $-k_1$ and $-k_2$ at the others), the divergence at vertex i is $c_i \, \epsilon^{(\cdot)(\cdot)\rho\sigma} p_{j\rho} p_{k\sigma}$ with (i, j, k) cyclic. Then

$$c_0 \overset{(1)}{=} \frac{1-\beta}{4\pi^2}, \qquad c_1 = c_2 \overset{(2)}{=} \frac{1+\beta}{8\pi^2}, \qquad c_0 + c_1 + c_2 \overset{(3)}{=} \frac{1}{2\pi^2} \quad \text{for every } \beta.$$

(1) The first line of (10.96).
(2) The second line of (10.96), times -1 for the incoming momentum $-k_1$, with $\epsilon^{\nu\lambda\rho\sigma}(-k_2)_\rho q_\sigma = \epsilon^{\nu\lambda\rho\sigma} k_{1\rho} k_{2\sigma}$; vertex 2 by the photon-exchange symmetry.
(3) $\frac{2(1-\beta)+(1+\beta)+(1+\beta)}{8\pi^2} = \frac{4}{8\pi^2}$.

The total is the *anomaly budget*; routing only redistributes it. The covariant (and ABJ) answer puts it all on one vertex ($\beta = -1$: $c_0 = \frac{1}{2\pi^2}$). With three identical chiral currents, Bose symmetry demands $c_0 = c_1 = c_2$, i.e. $\beta = \frac{1}{3}$, and each vertex gets $\frac{1}{6\pi^2}$:

$$\frac{\text{consistent}}{\text{covariant}} = \frac{c_0(\beta = \frac{1}{3})}{c_0(\beta = -1)} = \frac{1/6\pi^2}{1/2\pi^2} = \frac{1}{3}.$$

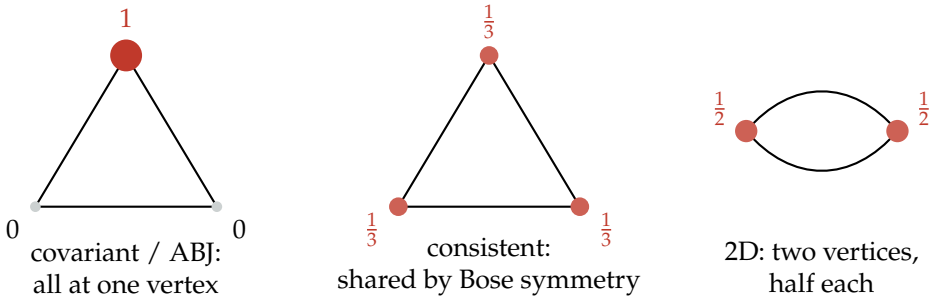


Figure 10.16: The anomaly budget of a one-loop polygon, drawn as the size of the red dot at each current vertex. The total is fixed and regulator independent. The covariant anomaly asks for all of it at the vertex we are looking at. A genuine effective action treats identical currents alike, so the $m+1$ vertices of the relevant polygon in $2m$ dimensions share it equally: the consistent anomaly is $\frac{1}{m+1}$ of the covariant one.

Key idea: Consistent = $\frac{1}{m+1}\times$ covariant, for U(1)

In $2m$ dimensions the anomaly of a U(1) Weyl fermion comes from the one-loop polygon with $m+1$ current vertices: the smallest loop with enough γ-matrices to produce $\epsilon^{\mu_1\cdots\mu_{2m}}$, with superficial degree of divergence $2m-(m+1) = m-1$ (logarithmic in two dimensions, linear in four). The covariant anomaly is the whole budget at one vertex; the consistent anomaly is the Bose-symmetric share, $\frac{1}{m+1}$ of it. For non-abelian G the two differ also in their A-dependence, as Examples 10.18 and 10.19 show, and the consistent one is found by descent.

10.5.4 Chern–Simons forms

The closed forms $\operatorname{tr} F^k$ of (10.140) are *locally exact*: on any region where A is a globally defined 1-form (for instance on all of $\mathbb{R}^n$, or on a ball) there is an explicit $(2k-1)$-form $Q_{2k-1}(A)$ with $\mathrm{d}Q_{2k-1} = \operatorname{tr} F^k$. It is found by switching the gauge field on slowly. Let

$$A_t := tA, \qquad F_t := F(A_t) = t\,\mathrm{d}A + t^2A^2, \qquad 0 \le t \le 1,$$

so $F_0 = 0$ and $F_1 = F$. Then

$$\begin{aligned}
\frac{\mathrm{d}F_t}{\mathrm{d}t} &\overset{(1)}{=} \mathrm{d}A + 2tA^2 \overset{(2)}{=} D_tA, \qquad D_t := D_{A_t} = \mathrm{d} + [tA, \cdot\,], \\
\frac{\mathrm{d}}{\mathrm{d}t}\operatorname{tr} F_t^k &\overset{(3)}{=} k\,\operatorname{tr}\Big(\frac{\mathrm{d}F_t}{\mathrm{d}t}F_t^{k-1}\Big) \overset{(4)}{=} k\,\operatorname{tr}\big((D_tA)F_t^{k-1}\big) \overset{(5)}{=} k\,\mathrm{d}\operatorname{tr}\big(AF_t^{k-1}\big), \\
\operatorname{tr} F^k &\overset{(6)}{=} \int_0^1 \mathrm{d}t\,\frac{\mathrm{d}}{\mathrm{d}t}\operatorname{tr} F_t^k \overset{(7)}{=} \mathrm{d}Q_{2k-1}(A), \qquad Q_{2k-1}(A) := k\int_0^1 \mathrm{d}t\,\operatorname{tr}\big(A\,F_t^{k-1}\big).
\end{aligned} \tag{10.145}$$

(1) Differentiate $t\,\mathrm{d}A + t^2A^2$.

(2) $[tA, A] = t(AA + AA) = 2tA^2$ (graded commutator of 1-forms).
(3) Leibniz, and cyclicity to bring each $\dot{F}_t$ to the front (F_t is even: no signs).
(4) Line 1.
(5) $\mathrm{d}\,\mathrm{tr}(X) = \mathrm{tr}(D_t X)$ for any matrix-valued form (the commutator part is traceless), then Leibniz for D_t: $D_t(AF_t^{k-1}) = (D_tA)F_t^{k-1} - A\,D_t(F_t^{k-1})$, and the Bianchi identity $D_tF_t = 0$.
(6) $F_0 = 0$ and $\mathrm{tr}\,F_0^k = 0$.
(7) Line 5 integrated; d commutes with $\int \mathrm{d}t$.

Q_{2k-1} is the **Chern–Simons form**. Two cases by hand.

Worked example 10.22: Q_3 and Q_5

$k = 2$.

$$Q_3 = 2\int_0^1 \mathrm{d}t\ \mathrm{tr}\big(A(t\,\mathrm{d}A + t^2A^2)\big) \overset{(1)}{=} 2\Big[\tfrac{1}{2}\mathrm{tr}(A\,\mathrm{d}A) + \tfrac{1}{3}\mathrm{tr}\,A^3\Big] = \mathrm{tr}\Big(A\,\mathrm{d}A + \tfrac{2}{3}A^3\Big).$$

(1) $\int_0^1 t\,\mathrm{d}t = \frac{1}{2}$, $\int_0^1 t^2\mathrm{d}t = \frac{1}{3}$.

Check: $\mathrm{d}\,\mathrm{tr}(A\,\mathrm{d}A) = \mathrm{tr}(\mathrm{d}A\,\mathrm{d}A)$, and $\mathrm{d}\,\mathrm{tr}\,A^3 = \mathrm{tr}(\mathrm{d}A\,A^2) - \mathrm{tr}(A\,\mathrm{d}A\,A) + \mathrm{tr}(A^2\mathrm{d}A) = 3\,\mathrm{tr}(\mathrm{d}A\,A^2)$ (move the lone A in the middle term to the end, sign $(-1)^{1\cdot3}$; move $\mathrm{d}A$ in the last to the front, sign +). So $\mathrm{d}Q_3 = \mathrm{tr}(\mathrm{d}A)^2 + 2\,\mathrm{tr}(\mathrm{d}A\,A^2)$. On the other hand $\mathrm{tr}\,F^2 = \mathrm{tr}(\mathrm{d}A)^2 + \mathrm{tr}(\mathrm{d}A\,A^2) + \mathrm{tr}(A^2\mathrm{d}A) + \mathrm{tr}\,A^4 = \mathrm{tr}(\mathrm{d}A)^2 + 2\,\mathrm{tr}(\mathrm{d}A\,A^2)$, since $\mathrm{tr}\,A^4 = 0$. They agree.

$k = 3$. Expand $F_t^2 = t^2(\mathrm{d}A)^2 + t^3(\mathrm{d}A\,A^2 + A^2\mathrm{d}A) + t^4A^4$. Then

$$\begin{aligned}\mathrm{tr}(AF_t^2) &\overset{(1)}{=} t^2\,\mathrm{tr}\big(A(\mathrm{d}A)^2\big) + t^3\big[\mathrm{tr}(A\,\mathrm{d}A\,A^2) + \mathrm{tr}(A^3\mathrm{d}A)\big] + t^4\,\mathrm{tr}\,A^5\\ &\overset{(2)}{=} t^2\,\mathrm{tr}\big(A(\mathrm{d}A)^2\big) + 2t^3\,\mathrm{tr}(A^3\mathrm{d}A) + t^4\,\mathrm{tr}\,A^5,\end{aligned}$$

$$Q_5 = 3\int_0^1 \mathrm{d}t\ \mathrm{tr}(AF_t^2) \overset{(3)}{=} \mathrm{tr}\Big(A(\mathrm{d}A)^2 + \tfrac{3}{2}A^3\mathrm{d}A + \tfrac{3}{5}A^5\Big).$$

(1) Multiply out.
(2) $\mathrm{tr}(A\,\mathrm{d}A\,A^2) = \mathrm{tr}(A^2\cdot A\,\mathrm{d}A)$: moving the even form A^2 from the back to the front costs no sign.
(3) $3\cdot\frac{1}{3} = 1$, $3\cdot2\cdot\frac{1}{4} = \frac{3}{2}$, $3\cdot\frac{1}{5} = \frac{3}{5}$.

Worked example 10.23: A pure gauge counts windings

If $A = g^{-1}\mathrm{d}g$ then $F = 0$, so $\mathrm{d}Q_3 = 0$, but Q_3 need not be exact. With $\mathrm{d}(g^{-1}\mathrm{d}g) = -g^{-1}\mathrm{d}g\,g^{-1}\mathrm{d}g = -A^2$ (from $\mathrm{d}(g^{-1}) = -g^{-1}\mathrm{d}g\,g^{-1}$),

$$Q_3(g^{-1}\mathrm{d}g) = \mathrm{tr}\big(A(-A^2) + \tfrac{2}{3}A^3\big) = -\tfrac{1}{3}\,\mathrm{tr}\big(g^{-1}\mathrm{d}g\big)^3.$$

For $g : S^3 \to \mathrm{SU}(2)$, $\frac{1}{24\pi^2}\int_{S^3}\mathrm{tr}(g^{-1}\mathrm{d}g)^3$ is (up to sign) the winding number, an integer (Chapter 8). So a Chern–Simons form evaluated on a gauge transformation measures $\pi_3(G)$. In the same way $Q_5(g^{-1}\mathrm{d}g) = \frac{1}{10}\,\mathrm{tr}(g^{-1}\mathrm{d}g)^5$ measures $\pi_5(G)$, the group that appears in Nash's (10.18). (For the last formula: with $A = g^{-1}\mathrm{d}g$, $\mathrm{d}A = -A^2$, so $Q_5 = \mathrm{tr}(A^5 - \frac{3}{2}A^5 + \frac{3}{5}A^5) = \frac{10-15+6}{10}\,\mathrm{tr}\,A^5 = \frac{1}{10}\,\mathrm{tr}\,A^5$.)

10.5.5 The Stora–Zumino descent equations

Here is the construction that turns a characteristic class in dimension $2m+2$ into an anomaly in dimension $2m$. Start with

$$\Omega_{2m+2} \coloneqq \mathrm{ch}_{m+1}(F) = \frac{1}{(m+1)!}\Big(\frac{i}{2\pi}\Big)^{m+1} \operatorname{tr} F^{m+1}, \tag{10.146}$$

a closed, gauge-invariant $(2m+2)$-form built from A. (It is a form on an imagined $(2m+2)$-dimensional space; on our $2m$-dimensional M it vanishes identically, but its *algebraic* structure is what matters.)

1. *Descent 1.* $\Omega_{2m+2} = \mathrm{d}\omega^0_{2m+1}$ with $\omega^0_{2m+1} = \frac{1}{(m+1)!}(\frac{i}{2\pi})^{m+1} Q_{2m+1}(A)$, by (10.145).

2. *Descent 2.* $\mathrm{d}(\delta_v \omega^0_{2m+1}) = \delta_v \Omega_{2m+2} = 0$ by (10.140) (and δ_v commutes with d). A closed form built polynomially from A, v and their derivatives is exact in the same class of forms (the algebraic Poincaré lemma), so

$$\delta_v \omega^0_{2m+1} = \mathrm{d}\omega^1_{2m}(v, A),$$

   with ω^1_{2m} a $2m$-form, linear in v.

3. *The anomaly.* $\int_M \omega^1_{2m}(v, A)$ is a consistent anomaly on the $2m$-dimensional M.

The superscript counts the number of gauge parameters; the subscript is the form degree. The pattern (one more δ, one less form degree) is the **descent**.

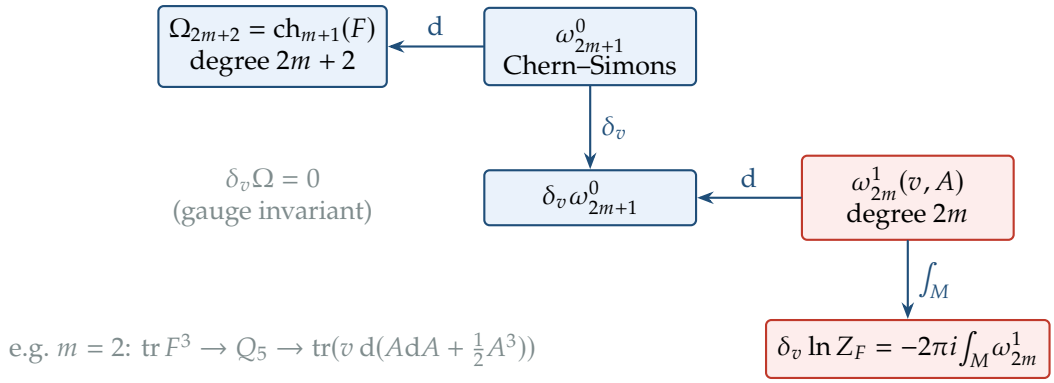


Figure 10.17: The descent ladder. Moving right lowers the form degree by one (d^{-1}); moving down applies one gauge variation. The anomaly in $2m$ dimensions is two steps from the characteristic class in $2m+2$ dimensions.

Step 1: how a Chern–Simons form varies under an arbitrary δA. We need a formula for δQ_{2k-1} valid for any variation, and then specialise to $\delta A = D_A v$. With $F_t = F(tA)$, $\delta F_t = D_t(t\,\delta A)$. We claim

$$\delta Q_{2k-1} = k\,\operatorname{tr}\big(\delta A\, F^{k-1}\big) - \mathrm{d}\Big[k \int_0^1 \mathrm{d}t\; t \sum_{j=0}^{k-2} \operatorname{tr}\big(A\, F_t^j\, \delta A\, F_t^{k-2-j}\big)\Big]. \tag{10.147}$$

Derivation: vary the integrand of (10.145),

$$\begin{aligned}
\delta \operatorname{tr}(A F_t^{k-1}) &\overset{(1)}{=} \operatorname{tr}(\delta A\, F_t^{k-1}) + t \sum_j \operatorname{tr}\big(A F_t^{j} (D_t \delta A) F_t^{k-2-j}\big) \\
&\overset{(2)}{=} \operatorname{tr}(\delta A\, F_t^{k-1}) + t \sum_j \operatorname{tr}\big((D_t A) F_t^{j}\, \delta A\, F_t^{k-2-j}\big) - t\, \mathrm{d} \sum_j \operatorname{tr}\big(A F_t^{j}\, \delta A\, F_t^{k-2-j}\big) \\
&\overset{(3)}{=} \operatorname{tr}(\delta A\, F_t^{k-1}) + t \sum_j \operatorname{tr}\big(\delta A\, F_t^{j}\, \dot F_t\, F_t^{k-2-j}\big) - t\, \mathrm{d} \sum_j \operatorname{tr}\big(A F_t^{j}\, \delta A\, F_t^{k-2-j}\big) \\
&\overset{(4)}{=} \frac{\mathrm{d}}{\mathrm{d}t}\Big[t\ \operatorname{tr}\big(\delta A\, F_t^{k-1}\big)\Big] - t\, \mathrm{d} \sum_j \operatorname{tr}\big(A F_t^{j}\, \delta A\, F_t^{k-2-j}\big).
\end{aligned}$$

(1) Leibniz; $\delta F_t = t\, D_t \delta A$.

(2) Integrate D_t by parts inside the trace: $\mathrm{d}\operatorname{tr}(A F_t^j \delta A F_t^{k-2-j}) = \operatorname{tr}((D_t A) F_t^j \delta A F_t^{k-2-j}) - \operatorname{tr}(A F_t^j (D_t \delta A) F_t^{k-2-j})$, using $D_t F_t = 0$; the sign is that of moving D_t past the 1-form A.

(3) $D_t A = \dot F_t$ from (10.145); then cyclically move the odd block $\delta A\, F_t^{k-2-j}$ to the front past the even block $\dot F_t F_t^j$ (no sign), and relabel $j \to k-2-j$.

(4) $\frac{\mathrm{d}}{\mathrm{d}t}[t \operatorname{tr}(\delta A F_t^{k-1})] = \operatorname{tr}(\delta A F_t^{k-1}) + t \sum_j \operatorname{tr}(\delta A\, F_t^j \dot F_t F_t^{k-2-j})$.

Multiplying by k and integrating over $t \in [0,1]$ (the boundary term is $1 \cdot \operatorname{tr}(\delta A F^{k-1}) - 0$) gives (10.147). For $k=2$ it reads $\delta Q_3 = 2\operatorname{tr}(\delta A\, F) - \mathrm{d}\operatorname{tr}(A\,\delta A)$, which one can confirm directly from $Q_3 = \operatorname{tr}(A\mathrm{d}A + \frac{2}{3}A^3)$.

Step 2: specialise to a gauge variation. For $\delta A = D_A v$ the first term is exact:

$$\operatorname{tr}\big((D_A v) F^{k-1}\big) \overset{(1)}{=} \operatorname{tr}\big(D_A (v F^{k-1})\big) \overset{(2)}{=} \mathrm{d}\operatorname{tr}\big(v F^{k-1}\big).$$

(1) Leibniz for D_A and Bianchi $D_A F = 0$.

(2) The trace of a commutator $[A, X]$ vanishes by graded cyclicity.

Hence $\delta_v Q_{2k-1} = \mathrm{d} Q^1_{2k-2}$ with

$$Q^1_{2k-2}(v, A) = k\ \operatorname{tr}\big(v F^{k-1}\big) - k \int_0^1 \mathrm{d}t\ t \sum_{j=0}^{k-2} \operatorname{tr}\big(A\, F_t^{j}\, (D_A v)\, F_t^{k-2-j}\big) \quad (\text{mod d-exact}). \tag{10.148}$$

Worked example 10.24: Descent in two dimensions: $\operatorname{tr} F^2 \to Q_3 \to \operatorname{tr}(v\,\mathrm{d}A)$

Put $k=2$ in (10.148):

$$\begin{aligned}
Q_2^1 &\overset{(1)}{=} 2\operatorname{tr}(vF) - 2\int_0^1 t\,\mathrm{d}t\ \ \operatorname{tr}(A\, D_A v) \overset{(2)}{=} 2\operatorname{tr}(v\,\mathrm{d}A) + 2\operatorname{tr}(vA^2) - \operatorname{tr}(A\,\mathrm{d}v) - \operatorname{tr}(A[A,v]) \\
&\overset{(3)}{=} 2\operatorname{tr}(v\,\mathrm{d}A) + 2\operatorname{tr}(vA^2) - \operatorname{tr}(A\,\mathrm{d}v) - 2\operatorname{tr}(vA^2) \overset{(4)}{=} 2\operatorname{tr}(v\,\mathrm{d}A) - \operatorname{tr}(v\,\mathrm{d}A) + \mathrm{d}\operatorname{tr}(vA) \\
&\equiv \operatorname{tr}(v\,\mathrm{d}A) \quad (\text{mod d-exact}).
\end{aligned}$$

(1) (10.148); the sum has the single term $j=0$.

(2) $F = \mathrm{d}A + A^2$, $\int_0^1 t\,\mathrm{d}t = \frac{1}{2}$, $D_A v = \mathrm{d}v + [A, v]$.

(3) $\operatorname{tr}(A[A,v]) = \operatorname{tr}(A^2 v) - \operatorname{tr}(AvA)$ and $\operatorname{tr}(AvA) = -\operatorname{tr}(A^2 v)$ (move the last A to the front past the 1-form Av), so $\operatorname{tr}(A[A,v]) = 2\operatorname{tr}(vA^2)$.

(4) $\mathrm{d}\,\mathrm{tr}(vA) = \mathrm{tr}(\mathrm{d}v\,A) + \mathrm{tr}(v\,\mathrm{d}A)$ and $\mathrm{tr}(A\,\mathrm{d}v) = -\mathrm{tr}(\mathrm{d}v\,A)$, so $\mathrm{tr}(A\,\mathrm{d}v) = \mathrm{tr}(v\,\mathrm{d}A) - \mathrm{d}\,\mathrm{tr}(vA)$.

Therefore

$$\omega_2^1 = \frac{1}{2}\Big(\frac{i}{2\pi}\Big)^2 \mathrm{tr}(v\,\mathrm{d}A) = -\frac{1}{8\pi^2}\,\mathrm{tr}(v\,\mathrm{d}A), \qquad -2\pi i\int_M \omega_2^1 = \frac{i}{4\pi}\int_M \mathrm{tr}(v\,\mathrm{d}A),$$

which is exactly the consistent anomaly $\mathcal{A}_2$ of Example 10.19. For U(1) ($A = -ie\,a$, $v = -ie\lambda$) this is $\frac{i}{4\pi}(-ie)^2\int \lambda\,\mathrm{d}a = -\frac{ie^2}{4\pi}\int \lambda\,\mathrm{d}a$, half of the covariant $\frac{i}{2\pi}\int \mathrm{tr}(vF) = -\frac{ie^2}{2\pi}\int \lambda\,\mathrm{d}a$, as the chiral Schwinger model demanded.

Worked example 10.25: Descent in four dimensions: $\mathrm{tr}\,F^3 \to Q_5 \to \mathrm{tr}(v\,\mathrm{d}(A\mathrm{d}A + \frac{1}{2}A^3))$

Put $k = 3$ in (10.148) and write $X := D_A v = \mathrm{d}v + Av - vA$. With $F_t = t\,\mathrm{d}A + t^2A^2$ and $\int_0^1 t^2\mathrm{d}t = \frac{1}{3}$, $\int_0^1 t^3\mathrm{d}t = \frac{1}{4}$,

$$Q_4^1 = \underbrace{3\,\mathrm{tr}(vF^2)}_{\text{(i)}} - \underbrace{\big[\mathrm{tr}(AX\mathrm{d}A) + \mathrm{tr}(A\,\mathrm{d}A\,X)\big]}_{\text{(ii)}} - \tfrac{3}{4}\underbrace{\big[\mathrm{tr}(AXA^2) + \mathrm{tr}(A^3X)\big]}_{\text{(iii)}}.$$

We bring every term to the form $\mathrm{tr}(v\cdots)$, working modulo exact forms, and use $\int \mathrm{tr}(\mathrm{d}v\,Y) = -\int \mathrm{tr}(v\,\mathrm{d}Y)$ for a 3-form Y (Leibniz and Stokes).

Term (i). $3\,\mathrm{tr}(vF^2) = 3\,\mathrm{tr}\big(v(\mathrm{d}A)^2\big) + 3\,\mathrm{tr}(v\,\mathrm{d}A\,A^2) + 3\,\mathrm{tr}(vA^2\mathrm{d}A) + 3\,\mathrm{tr}(vA^4)$.

Term (ii). Insert X:

$$\begin{aligned}
&\mathrm{tr}(A\,\mathrm{d}v\,\mathrm{d}A) + \mathrm{tr}(A\,\mathrm{d}A\,\mathrm{d}v) \overset{(1)}{=} -\mathrm{tr}\big(\mathrm{d}v\,(\mathrm{d}A\,A + A\,\mathrm{d}A)\big) \overset{(2)}{=} 2\,\mathrm{tr}\big(v(\mathrm{d}A)^2\big),\\
&\mathrm{tr}(AAv\,\mathrm{d}A) - \mathrm{tr}(AvA\,\mathrm{d}A)\\
&+\mathrm{tr}(A\,\mathrm{d}A\,Av) - \mathrm{tr}(A\,\mathrm{d}A\,vA)\\
&\qquad\qquad \overset{(3)}{=} \mathrm{tr}(v\,\mathrm{d}A\,A^2) + \mathrm{tr}(vA\,\mathrm{d}A\,A) + \mathrm{tr}(vA\,\mathrm{d}A\,A) + \mathrm{tr}(vA^2\mathrm{d}A).
\end{aligned}$$

(1) Move the leading A to the back past the 3-form $\mathrm{d}v\,\mathrm{d}A$ (sign $-$); move the trailing $\mathrm{d}v$ to the front past the 3-form $A\,\mathrm{d}A$ (sign $-$).

(2) Integrate by parts with $Y = \mathrm{d}A\,A + A\,\mathrm{d}A$, whose derivative is $\mathrm{d}Y = (\mathrm{d}A)^2 + (\mathrm{d}A)^2$.

(3) Term by term: $\mathrm{tr}(A^2\cdot v\mathrm{d}A) = \mathrm{tr}(v\,\mathrm{d}A\,A^2)$ (even block, no sign); $\mathrm{tr}(A\cdot vA\mathrm{d}A) = (-1)^{1\cdot 3}\,\mathrm{tr}(vA\,\mathrm{d}A\cdot A)$ (the 1-form A moved to the back past a 3-form), so $-\mathrm{tr}(AvA\,\mathrm{d}A) = +\mathrm{tr}(vA\,\mathrm{d}A\,A)$; $\mathrm{tr}(A\,\mathrm{d}A\,A\,v) = \mathrm{tr}(vA\,\mathrm{d}A\,A)$; $-\mathrm{tr}(A\mathrm{d}Av\cdot A) = +\mathrm{tr}(A\cdot A\mathrm{d}Av) = \mathrm{tr}(vA^2\mathrm{d}A)$.

So (ii) $\equiv 2\,\mathrm{tr}(v(\mathrm{d}A)^2) + \mathrm{tr}(v\,\mathrm{d}A\,A^2) + 2\,\mathrm{tr}(vA\,\mathrm{d}A\,A) + \mathrm{tr}(vA^2\mathrm{d}A)$.

Term (iii). Insert X:

$$\begin{aligned}
&\mathrm{tr}(A\,\mathrm{d}v\,A^2) + \mathrm{tr}(A^3\mathrm{d}v) \overset{(1)}{=} -2\,\mathrm{tr}(\mathrm{d}v\,A^3) \overset{(2)}{=} 2\,\mathrm{tr}\big(v\,\mathrm{d}(A^3)\big)\\
&\qquad\qquad \overset{(3)}{=} 2\,\mathrm{tr}(v\,\mathrm{d}A\,A^2) - 2\,\mathrm{tr}(vA\,\mathrm{d}A\,A) + 2\,\mathrm{tr}(vA^2\mathrm{d}A),\\
&\mathrm{tr}(A^2vA^2) - \mathrm{tr}(AvA^3)\\
&+\mathrm{tr}(A^4v) - \mathrm{tr}(A^3vA) \overset{(4)}{=} 4\,\mathrm{tr}(vA^4).
\end{aligned}$$

(1) Move A (front) to the back past the 3-form $\mathrm{d}vA^2$: sign $-$; move $\mathrm{d}v$ (back) to the front past A^3: sign $-$.

(2) Integration by parts.

(3) $\mathrm{d}(A^3) = \mathrm{d}A\,A^2 - A\,\mathrm{d}A\,A + A^2\mathrm{d}A$.
(4) Each term becomes $\pm\,\mathrm{tr}(vA^4)$ by cyclic moves, and the two minus signs are cancelled by the signs $(-1)^{1\cdot 3}$ of moving a single A past A^3.

Collect $Q_4^1 \equiv$ (i) $-$ (ii) $-\,\frac{3}{4}$(iii), coefficient by coefficient:

	(i)	−(ii)	$-\frac{3}{4}$(iii)	total
$\mathrm{tr}(v(\mathrm{d}A)^2)$	3	−2	0	1
$\mathrm{tr}(v\,\mathrm{d}A\,A^2)$	3	−1	$-\frac{3}{2}$	$\frac{1}{2}$
$\mathrm{tr}(vA\,\mathrm{d}A\,A)$	0	−2	$+\frac{3}{2}$	$-\frac{1}{2}$
$\mathrm{tr}(vA^2\mathrm{d}A)$	3	−1	$-\frac{3}{2}$	$\frac{1}{2}$
$\mathrm{tr}(vA^4)$	3	0	−3	0

The totals are $\mathrm{tr}(v(\mathrm{d}A)^2) + \frac{1}{2}\,\mathrm{tr}\big(v(\mathrm{d}A\,A^2 - A\,\mathrm{d}A\,A + A^2\mathrm{d}A)\big)$. Since $\mathrm{d}(A\,\mathrm{d}A) = (\mathrm{d}A)^2$ and $\mathrm{d}(A^3) = \mathrm{d}A\,A^2 - A\,\mathrm{d}A\,A + A^2\mathrm{d}A$,

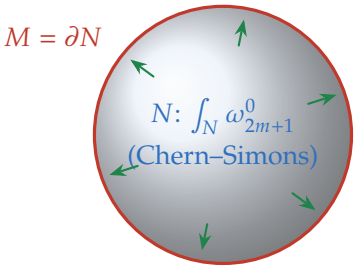
$$Q_4^1(v,A) \equiv \mathrm{tr}\Big(v\,\mathrm{d}\Big(A\,\mathrm{d}A + \tfrac{1}{2}A^3\Big)\Big), \qquad \omega_4^1 = \frac{1}{3!}\Big(\frac{i}{2\pi}\Big)^3 Q_4^1 = -\frac{i}{48\pi^3}\,\mathrm{tr}\Big(v\,\mathrm{d}\Big(A\,\mathrm{d}A + \tfrac{1}{2}A^3\Big)\Big). \tag{10.149}$$

Why $\int_M \omega^1$ is automatically consistent: anomaly inflow. Let N be a $(2m+1)$-manifold with boundary $\partial N = M$ (for $M = S^{2m}$ take N the ball B^{2m+1}), and extend A and v smoothly to $\tilde{A}$, $\tilde{v}$ on N. Consider the Chern–Simons action $\Gamma_N[\tilde{A}] := \int_N \omega^0_{2m+1}(\tilde{A})$. Then

$$\delta_{\tilde{v}}\Gamma_N \overset{(1)}{=} \int_N \delta_{\tilde{v}}\omega^0_{2m+1} \overset{(2)}{=} \int_N \mathrm{d}\omega^1_{2m}(\tilde{v},\tilde{A}) \overset{(3)}{=} \int_M \omega^1_{2m}(v,A), \tag{10.150}$$

(1) Vary under the integral.
(2) Descent 2.
(3) Stokes' theorem on N; on $\partial N = M$ the extensions restrict to A and v.

So the descent anomaly on M is literally the gauge variation of a functional (of the bulk extension), and (10.142) applied to Γ_N gives the Wess–Zumino condition (10.143) with no further computation. The right side depends only on boundary data, even though Γ_N does not.



$\delta_v\Gamma_N = \int_M \omega^1_{2m}$: the Chern–Simons bulk is not gauge invariant, and its failure is a pure boundary term. A chiral fermion on M with the *opposite* anomaly makes the total invariant. Charge flowing in from the bulk (green) is the anomalous non-conservation seen on the boundary.

Figure 10.18: Anomaly inflow. The descent equation $\delta_v\omega^0 = \mathrm{d}\omega^1$ says that the gauge variation of a Chern–Simons action on N lives entirely on ∂N. This is why the descent anomaly obeys the Wess–Zumino condition, and it is the mechanism behind the edge states of the quantum Hall effect.

Key idea: The consistent anomaly (Nash's μ_1), with its normalisation

For a Weyl fermion of positive chirality in a representation with anti-Hermitian generators,

on a $2m$-dimensional M,

$$\begin{aligned} \delta_v \ln Z_F[A] &= -2\pi i \int_M \omega^1_{2m}(v, A) \quad \text{(mod trivial anomalies)}, \\ \Omega_{2m+2} = \mathrm{ch}_{m+1}(F) &= \mathrm{d}\omega^0_{2m+1}, \qquad \delta_v \omega^0_{2m+1} = \mathrm{d}\omega^1_{2m}. \end{aligned} \tag{10.151}$$

Explicitly,

$$\begin{aligned} n = 2: &\quad \delta_v \ln Z_F = \frac{i}{4\pi} \int_M \mathrm{tr}(v\, \mathrm{d}A), \\ n = 4: &\quad \delta_v \ln Z_F = -\frac{1}{24\pi^2} \int_M \mathrm{tr}\Big(v\, \mathrm{d}\Big(A\, \mathrm{d}A + \tfrac{1}{2}A^3\Big)\Big). \end{aligned}$$

Why is this the right normalisation? The consistency condition fixes the *form* of the anomaly (the relevant cohomology is one-dimensional for each independent invariant polynomial), so only the overall number needs checking, and it suffices to check the abelian case against Section 10.5.3:

$$\begin{aligned} -2\pi i\, \omega^1_{2m}\big|_{\text{abelian}} &\overset{(1)}{=} -2\pi i\, \frac{1}{(m+1)!}\Big(\frac{i}{2\pi}\Big)^{m+1} v\, F^m \overset{(2)}{=} \frac{1}{(m+1)!}\Big(\frac{i}{2\pi}\Big)^{m} v\, F^m \\ &\overset{(3)}{=} \frac{1}{m+1}\ \mathrm{tr}\Big[v\, \frac{1}{m!}\Big(\frac{i}{2\pi}\Big)^{m} F^m\Big]. \end{aligned}$$

(1) For abelian A: $Q_{2m+1} = AF^m$ (all $F_t^m = t^m F^m$ commute, $(m+1)\int t^m \mathrm{d}t = 1$), and $\delta_v(AF^m) = \mathrm{d}v\, F^m = \mathrm{d}(vF^m)$ because $\mathrm{d}F = 0$; so $Q^1_{2m} = vF^m$.

(2) $-2\pi i \cdot \frac{i}{2\pi} = 1$.

(3) $\frac{1}{(m+1)!} = \frac{1}{m+1} \cdot \frac{1}{m!}$. The bracket is the covariant anomaly (10.135).

This is the $\frac{1}{m+1}$ of the key idea in Section 10.5.3: $\frac{1}{2}$ in two dimensions (Example 10.20) and $\frac{1}{3}$ in four (Example 10.21). The anomaly thus obeys the Wess–Zumino condition, is local, is not removable by local counterterms when $\mathrm{tr}\, F^{m+1} \not\equiv 0$, and is normalised by a single one-loop graph.

Reading Nash (p. 278, (10.33))

Nash quotes, for $M = S^4$ and $G = \mathrm{SU}(N)$, $\mu_1 = -\frac{i}{(2\pi)^3}\int_0^1 \mathrm{d}t \int_{S^4} \mathrm{tr}(A \wedge F_t \wedge F_t)$, $F_t = t\, \mathrm{d}A + t^2 A \wedge A$, "the expression for the anomaly found in Bardeen [1]". Compare with (10.145): the integrand $\int_0^1 \mathrm{d}t\, \mathrm{tr}(AF_t^2) = \frac{1}{3}Q_5$ is the five-dimensional Chern–Simons form. A 5-form cannot be integrated over S^4 as it stands; Nash's notation ("as used in chapter 8") is that A is the connection on $S^4 \times \mathcal{G}$ whose component along $\mathcal{G}$ is the Maurer–Cartan form $g^{-1}\delta g$. Integrating the part of Q_5 of degree 4 along S^4 and 1 along $\mathcal{G}$ over S^4 leaves a 1-form on $\mathcal{G}$, and that part is exactly the descent term ω^1_4 of (10.149), with v the value of $g^{-1}\delta g$ on the tangent vector. So (10.33) is our (10.151). Overall constants differ between references (Nash's prefactor is $2\omega^0_5$ in our normalisation, reflecting whether one tracks $\det T_g$ or the chiral determinant, and where one puts $2\pi i$); the structure, and in particular the vanishing criterion, do not. The combination $\mathrm{tr}(v\, \mathrm{d}(A\mathrm{d}A + \frac{1}{2}A^3))$ is the **Bardeen** (or minimal consistent) anomaly.

Dictionary: BRST language

Much of the literature writes the gauge parameter as a Grassmann-odd **ghost** $c(x) \in \mathfrak{g}$ and the gauge variation as a nilpotent operator s: $sA = -D_A c$, $sc = -c^2$, $s^2 = 0$ (with s anticommuting with d and with odd objects). The Wess–Zumino condition becomes $s\int_M \omega^1_{2m}(c, A) = 0$, and the descent equations collapse into the **Russian formula** $(\mathrm{d} + s)(A + c) + (A + c)^2 = F$, i.e. "the curvature of $A + c$ has no ghost components". Geometrically c is the Maurer–Cartan form $g^{-1}\delta g$ on $\mathcal{G}$, s is the exterior derivative on $\mathcal{G}$, $sc = -c^2$ is the Maurer–Cartan equation, and $s\mathcal{A} = 0$ is our statement $\mathrm{d}\mathcal{A} = 0$ (Figure 10.15). Nash's $H^1_{\mathrm{dR}}(\mathcal{G})$ is the cohomology of this s.

10.5.6 When does the anomaly vanish? The symmetric invariant

In four dimensions everything in ω^1_4 is controlled by one tensor. Write $A = A^a T^a$, $F_t = F_t^a T^a$. Then

$$\operatorname{tr}(AF_tF_t) \overset{(1)}{=} A^a F_t^b F_t^c \operatorname{tr}(T^a T^b T^c) \overset{(2)}{=} A^a F_t^b F_t^c \tfrac{1}{2}\operatorname{tr}\big(T^a\{T^b, T^c\}\big) \overset{(3)}{=} \tfrac{1}{2} d^{abc} A^a F_t^b F_t^c,$$
$$d^{abc} := \operatorname{tr}\big(T^a\{T^b, T^c\}\big) = 2\operatorname{Str}(T^a T^b T^c). \tag{10.152}$$

(1) Expand in components; the component forms are ordinary (matrix-free) forms.
(2) $F_t^b F_t^c$ is symmetric in $b \leftrightarrow c$ (2-forms commute), so only the symmetric part of $\operatorname{tr}(T^a T^b T^c)$ in b, c survives.
(3) Definition. d^{abc} is symmetric in all three indices: by cyclicity $\operatorname{tr}(T^b\{T^a, T^c\}) = \operatorname{tr}(T^b T^a T^c + T^b T^c T^a) = \operatorname{tr}(T^a T^c T^b + T^a T^b T^c) = d^{abc}$ [QFT notes (22.7.28)].

So $Q_5 = \frac{3}{2} d^{abc}\int_0^1 \mathrm{d}t\, A^a F_t^b F_t^c$, and Q_4^1 (derived from it) is proportional to d^{abc} too. Hence:

Key idea: The four-dimensional cancellation condition

A four-dimensional gauge theory of Weyl fermions is free of perturbative gauge anomalies iff

$$\sum_{\text{positive chirality}} d^{abc}(R) - \sum_{\text{negative chirality}} d^{abc}(R) = 0,$$

equivalently (writing every field as a left-handed Weyl fermion, below) $\sum_R d^{abc}(R) = 0$. For $G = \mathrm{U}(1)$ with charges q_i this is $\sum_i q_i^3 = 0$.

(The relative minus sign: the negative-chirality anomaly comes with $-\gamma_5$ in the Jacobian of Section 10.4.6.)

Conjugate and real representations. With anti-Hermitian generators the conjugate representation $\bar{R}$ has generators $\bar{T}^a = (T^a)^* = -(T^a)^{\mathsf{T}}$. Then

$$d^{abc}(\bar{R}) \overset{(1)}{=} -\operatorname{tr}\big((T^a)^{\mathsf{T}}\{(T^b)^{\mathsf{T}}, (T^c)^{\mathsf{T}}\}\big) \overset{(2)}{=} -\operatorname{tr}\big(\{T^c, T^b\}T^a\big)^{\mathsf{T}} \overset{(3)}{=} -d^{abc}(R).$$

(1) $(-1)^3 = -1$.
(2) $X^{\mathsf{T}}Y^{\mathsf{T}} = (YX)^{\mathsf{T}}$ for each product.
(3) $\operatorname{tr} Z^{\mathsf{T}} = \operatorname{tr} Z$, then cyclicity.

A *real* (or pseudoreal) representation is equivalent to its conjugate, $\bar{R} = URU^{-1}$, and d^{abc} is unchanged by conjugation, so $d^{abc}(R) = -d^{abc}(R) = 0$. In particular a Dirac fermion, whose two chiralities carry the same R, is never anomalous: this is why QED and QCD are safe

[QFT notes (22.7.34)]. Since every representation of SU(2) is real or pseudoreal, SU(2) is safe, as Example 10.17 found by hand.

Worked example 10.26: SU(3) triplets are anomalous; anomaly coefficients

For SU(N), $N \geq 3$, the fundamental has $d^{abc} \neq 0$. With the Hermitian Gell-Mann generator $t^8 = \frac{1}{2\sqrt{3}}\,\mathrm{diag}(1,1,-2)$,

$$\mathrm{tr}\big(t^8\{t^8,t^8\}\big) = 2\,\mathrm{tr}(t^8)^3 \overset{(1)}{=} 2\cdot\frac{1}{(2\sqrt{3})^3}\big(1+1-8\big) \overset{(2)}{=} -\frac{1}{2\sqrt{3}} \neq 0.$$

(1) The cube of a diagonal matrix is the cube of its entries.
(2) $(2\sqrt{3})^3 = 24\sqrt{3}$ and $2\cdot(-6)/(24\sqrt{3}) = -1/(2\sqrt{3})$.

Because SU(N) has only one symmetric invariant of this type, $d^{abc}(R) = A(R)\,d^{abc}(\mathbf{N})$ for a number $A(R)$, the **anomaly coefficient**: $A(\mathbf{N}) = 1$, $A(\bar{\mathbf{N}}) = -1$, $A(\text{adjoint}) = 0$. To compute $A(R)$ it is enough to compare $\mathrm{tr}_R\, X^3$ with $\mathrm{tr}_{\mathbf{N}}\, X^3$ for a single generic generator X. For the antisymmetric square $\Lambda^2\mathbf{N}$, use characters: $\mathrm{tr}_{\Lambda^2} e^{sX} = \frac{1}{2}[(\mathrm{tr}\, e^{sX})^2 - \mathrm{tr}\, e^{2sX}]$. Expanding to order s^3, with $\mathrm{tr}\, X = 0$,

$$\tfrac{1}{6}\,\mathrm{tr}_{\Lambda^2} X^3 \overset{(1)}{=} \tfrac{1}{2}\Big[2N\cdot\tfrac{1}{6}\,\mathrm{tr}\,X^3 - \tfrac{8}{6}\,\mathrm{tr}\,X^3\Big] \overset{(2)}{=} \frac{N-4}{6}\,\mathrm{tr}\,X^3, \qquad \text{so} \qquad A(\Lambda^2\mathbf{N}) = N-4.$$

(1) $(\mathrm{tr}\, e^{sX})^2 = (N + \frac{s^2}{2}\,\mathrm{tr}\,X^2 + \frac{s^3}{6}\,\mathrm{tr}\,X^3 + \dots)^2$ has s^3 coefficient $2N\cdot\frac{1}{6}\,\mathrm{tr}\,X^3$; and $\mathrm{tr}\, e^{2sX}$ has s^3 coefficient $\frac{8}{6}\,\mathrm{tr}\,X^3$.
(2) Simplify.

The same computation with a plus sign gives $A(\mathrm{Sym}^2\,\mathbf{N}) = N+4$ [QFT notes (22.7.39)]. For the SU(5) grand unified theory, one generation is $\bar{\mathbf{5}} \oplus \mathbf{10}$ (left-handed), with $\mathbf{10} = \Lambda^2\mathbf{5}$: $A = -1 + (5-4) = 0$. The generation is anomaly free, and this is part of the appeal of SU(5).

group	possible $d^{abc} \neq 0$?	reason
U(1)	yes: $\sum q^3$	abelian, $\mathrm{tr}\, F^3 = F^3$
SU(2), Sp(n), SO(n) ($n \neq 6$), $G_2, F_4, E_{6,7,8}$	no	no cubic Casimir (all representations real, or no symmetric invariant)
SU(N), $N \geq 3$	yes: $\sum A(R)$	cubic Casimir d^{abc}
SO(6) $\cong$ SU(4)	yes	same as SU(4)

(For product groups one also needs the mixed conditions below, and in the presence of gravity the linear condition $\sum q = 0$ for each U(1), Section 10.7.)

10.5.7 Anomaly cancellation in the Standard Model

The Standard Model is a chiral gauge theory with $G = \mathrm{SU}(3)\times\mathrm{SU}(2)\times\mathrm{U}(1)_Y$: left- and right-handed quarks and leptons carry different representations. Its consistency therefore rests on (10.152) summing to zero for every choice of three generators. We follow [QFT notes §23.4].

Writing everything as left-handed. A right-handed Weyl field ψ_R in a representation r is equivalent to a left-handed field $\psi'_L = \sigma^2\psi_R^*$ in the conjugate representation $\bar{r}$ [QFT notes (22.7.15)–(22.7.21)]; its hypercharge flips sign. We use the convention $Q = T_3 + Y$, with Hermitian

generators t^a normalised by $\mathrm{tr}(t^a t^b) = \frac{1}{2}\delta^{ab}$ in the fundamentals of SU(3) and SU(2). One generation is:

left-handed field	SU(3)	SU(2)	Y	number of Weyl components
$Q = (u_L, d_L)$	$\mathbf{3}$	$\mathbf{2}$	$+\frac{1}{6}$	6
u^c $(= \overline{u_R})$	$\bar{\mathbf{3}}$	$\mathbf{1}$	$-\frac{2}{3}$	3
d^c $(= \overline{d_R})$	$\bar{\mathbf{3}}$	$\mathbf{1}$	$+\frac{1}{3}$	3
$L = (\nu_L, e_L)$	$\mathbf{1}$	$\mathbf{2}$	$-\frac{1}{2}$	2
e^c $(= \overline{e_R})$	$\mathbf{1}$	$\mathbf{1}$	$+1$	1

(Check a charge: u_L has $T_3 = \frac{1}{2}$, so $Q = \frac{1}{2} + \frac{1}{6} = \frac{2}{3}$; u_R must have the same Q with $T_3 = 0$, so $Y(u_R) = \frac{2}{3}$ and $Y(u^c) = -\frac{2}{3}$.)

The list of triangles. Each of the three vertices carries a generator of one factor. A triangle containing exactly one SU(3) or exactly one SU(2) generator vanishes, because the other two generators act on different indices and leave a factor $\mathrm{tr}\, t^a = 0$. The remaining conditions are (Figure 10.19):

$$\mathrm{SU}(3)^3: \quad \sum A(r_3)\,\dim r_2 \overset{(1)}{=} \underbrace{1\cdot 2}_{Q} + \underbrace{(-1)\cdot 1}_{u^c} + \underbrace{(-1)\cdot 1}_{d^c} = 0,$$

$$\mathrm{SU}(2)^3: \quad d^{abc}(\mathrm{SU}(2)) = 0 \text{ identically (Example 10.17)},$$

$$\mathrm{SU}(3)^2\,\mathrm{U}(1): \quad \sum_{\text{coloured}} \tfrac{1}{2}\dim r_2\, Y \overset{(2)}{=} \tfrac{1}{2}\Big[2\cdot\tfrac{1}{6} + 1\cdot\big(-\tfrac{2}{3}\big) + 1\cdot\tfrac{1}{3}\Big] = \tfrac{1}{2}\Big[\tfrac{1}{3} - \tfrac{1}{3}\Big] = 0,$$

$$\mathrm{SU}(2)^2\,\mathrm{U}(1): \quad \sum_{\text{doublets}} \tfrac{1}{2}\dim r_3\, Y \overset{(3)}{=} \tfrac{1}{2}\Big[3\cdot\tfrac{1}{6} + 1\cdot\big(-\tfrac{1}{2}\big)\Big] = 0,$$

$$\mathrm{U}(1)^3: \quad \sum(\text{components})\, Y^3 \overset{(4)}{=} 6\cdot\tfrac{1}{216} + 3\cdot\big(-\tfrac{8}{27}\big) + 3\cdot\tfrac{1}{27} + 2\cdot\big(-\tfrac{1}{8}\big) + 1 = \tfrac{1-32+4-9+36}{36} = 0,$$

$$\mathrm{U}(1)\,(\mathrm{grav})^2: \quad \sum(\text{components})\, Y \overset{(5)}{=} 6\cdot\tfrac{1}{6} + 3\cdot\big(-\tfrac{2}{3}\big) + 3\cdot\tfrac{1}{3} + 2\cdot\big(-\tfrac{1}{2}\big) + 1 = 1 - 2 + 1 - 1 + 1 = 0.$$

(1) d^{abc} factorises: the SU(3) trace gives $A(r_3)d^{abc}(\mathbf{3})$ and the SU(2) indices give a factor $\dim r_2$. Q is two triplets; u^c, d^c are antitriplets with $A = -1$.

(2) $\mathrm{tr}_{r_3}(t^a t^b Y) = \frac{1}{2}\delta^{ab}\, Y \dim r_2$ for $r_3 = \mathbf{3}$ or $\bar{\mathbf{3}}$ (the $\bar{\mathbf{3}}$ generators $-t^{a\mathsf{T}}$ have the same $\mathrm{tr}(t^a t^b)$); the common $\frac{1}{2}\delta^{ab}$ is factored out. Colour singlets do not contribute.

(3) Only Q (three colours) and L are doublets.

(4) $d^{YYY} = 2\sum Y^3$ summed over all 15 Weyl components: $(\frac{1}{6})^3 = \frac{1}{216}$, $(-\frac{2}{3})^3 = -\frac{8}{27}$, $(\frac{1}{3})^3 = \frac{1}{27}$, $(-\frac{1}{2})^3 = -\frac{1}{8}$; over the common denominator 36: $\frac{6}{216} = \frac{1}{36}$, $-\frac{24}{27} = -\frac{32}{36}$, $\frac{3}{27} = \frac{4}{36}$, $-\frac{1}{4} = -\frac{9}{36}$.

(5) The mixed gauge–gravitational triangle is proportional to $\mathrm{tr}\, Y$ ([QFT notes (22.8.1)], Section 10.7).

Every condition holds. The QFT notes do the same sums with right-handed fields kept as they are and a factor -1 for left, $+1$ for right [QFT notes (23.4.2)–(23.4.5)]; since $Y(\psi^c) = -Y(\psi_R)$ and Y^3, Y are odd, the two bookkeepings agree term by term.

Worked example 10.27: Anomaly cancellation nearly fixes the hypercharges

Keep the SU(3) × SU(2) assignments of the table but treat the five hypercharges y_Q, y_u, y_d, y_L, y_e (of Q, u^c, d^c, L, e^c) as unknowns. The four conditions read

$$\mathrm{SU}(3)^2\mathrm{U}(1): 2y_Q + y_u + y_d = 0, \qquad \mathrm{SU}(2)^2\mathrm{U}(1): 3y_Q + y_L = 0,$$

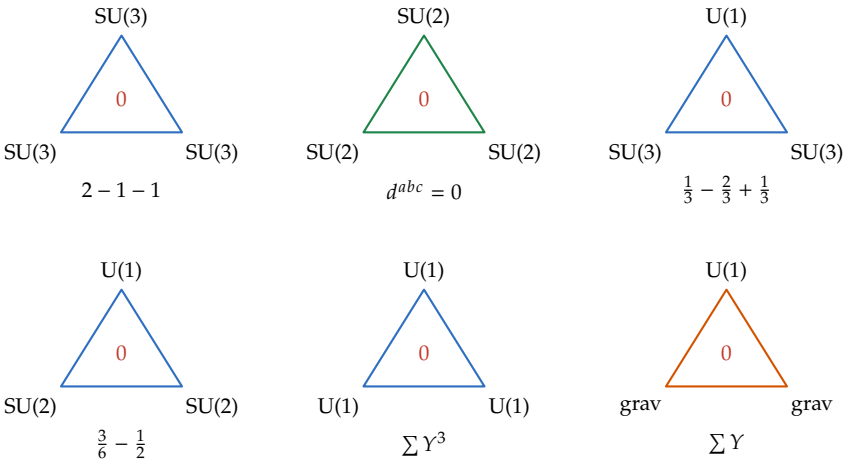


Figure 10.19: The six triangle conditions of the Standard Model that are not trivially zero. Blue: cancel between quarks and leptons, or between left and right, only because of the specific hypercharges and the three colours; green: automatic for SU(2); orange: the mixed gravitational condition. Triangles with a single SU(3) or single SU(2) vertex vanish because generators are traceless.

$$\text{grav}^2\text{U}(1) : 6y_Q + 3y_u + 3y_d + 2y_L + y_e = 0,$$
$$\text{U}(1)^3 : 6y_Q^3 + 3y_u^3 + 3y_d^3 + 2y_L^3 + y_e^3 = 0.$$

Solve them in turn:

$$y_L \overset{(1)}{=} -3y_Q, \qquad y_u + y_d \overset{(2)}{=} -2y_Q, \qquad y_e \overset{(3)}{=} -6y_Q - 3(-2y_Q) - 2(-3y_Q) = 6y_Q,$$
$$y_u^3 + y_d^3 \overset{(4)}{=} -\tfrac{1}{3}\big(6y_Q^3 + 2(-3y_Q)^3 + (6y_Q)^3\big) = -\tfrac{1}{3}\big(6 - 54 + 216\big)y_Q^3 = -56\,y_Q^3,$$
$$y_u y_d \overset{(5)}{=} \frac{(y_u + y_d)^3 - (y_u^3 + y_d^3)}{3(y_u + y_d)} = \frac{-8y_Q^3 + 56y_Q^3}{-6y_Q} = -8y_Q^2,$$
$$\{y_u, y_d\} \overset{(6)}{=} \text{roots of } z^2 + 2y_Q z - 8y_Q^2 = 0 = \{-4y_Q,\ 2y_Q\}.$$

(1) The $\text{SU}(2)^2\text{U}(1)$ condition.
(2) The $\text{SU}(3)^2\text{U}(1)$ condition.
(3) The gravitational condition with (1) and (2).
(4) The cubic condition with (1) and (3).
(5) The identity $a^3 + b^3 = (a + b)^3 - 3ab(a + b)$, assuming $y_Q \neq 0$.
(6) A pair with given sum $-2y_Q$ and product $-8y_Q^2$; the discriminant is $4y_Q^2 + 32y_Q^2 = 36y_Q^2$, so $z = -y_Q \pm 3y_Q$.

With $y_Q = \frac{1}{6}$ this gives $y_L = -\frac{1}{2}$, $y_e = 1$ and $\{y_u, y_d\} = \{-\frac{2}{3}, \frac{1}{3}\}$: exactly the table. The overall scale of Y is a convention, so *anomaly cancellation determines all the hypercharge ratios of a generation*, up to the label swap $u \leftrightarrow d$. In particular it forces $Q(e) = -Q(\nu) - 1$ and $Q(\text{proton}) = -Q(e)$: charge quantisation, from the consistency of the gauge theory.

Physical picture: Quarks and leptons need each other

The gravitational condition happens to cancel separately among leptons $(-1 + 1)$ and among quarks $(1 - 2 + 1)$, but the others do not. For $\text{SU}(2)^2\text{U}(1)$ the quarks give $3 \cdot \frac{1}{6} = +\frac{1}{2}$

and the leptons $-\frac{1}{2}$; for $\mathrm{U}(1)^3$ the quarks give $\frac{1-32+4}{36} = -\frac{3}{4}$ and the leptons $\frac{-9+36}{36} = +\frac{3}{4}$. So neither tower is consistent alone: cancellation needs a complete generation and exactly three colours. A theory with only one of the two towers, or with a different number of colours and the same charges, would be inconsistent: the gauge current would not be conserved and unphysical polarisations of the gauge bosons would not decouple. This is why, once the b quark and the τ lepton were found, a top quark completing the third generation was expected long before it was seen.

Watch out: Perturbative versus global

The triangle conditions are all *perturbative* (local) anomalies. There is one more, non-perturbative, condition: an SU(2) gauge theory must contain an even number of Weyl doublets (Witten's global anomaly, $\pi_4(\mathrm{SU}(2)) = \mathbb{Z}_2$), invisible to d^{abc} because SU(2) has none. Each Standard Model generation has four doublets (three coloured Q's and one L), so it passes. This is a torsion phenomenon in the topology of $\mathcal{G}$, treated in Nash's §5.

Reading Nash (pp. 272–273 and 278)

Nash states the anomaly as the non-vanishing of $\nabla^{ab}_{\mu} j_5^{b\mu}$ [Nash (10.11)] and later gives its cohomological form μ_1 [Nash (10.13), (10.33)]. Between these he does not write down the non-abelian anomaly explicitly, nor the consistency condition, nor its cancellation. This section is that missing middle: (10.143) is the statement that μ_1 is closed, (10.151) with (10.149) is his (10.33), the "trivial anomalies" $\delta_v \Phi_{\mathrm{loc}}$ are the exact forms whose subtlety he raises on p. 279 (local versus de Rham cohomology, Section 10.6.6), and the Standard Model table is the physical reason the question matters.

Recap

- The consistent anomaly $\mathcal{A}[v] = \delta_v \ln Z_F$ obeys the Wess–Zumino condition $\delta_v \mathcal{A}[w] - \delta_w \mathcal{A}[v] = \mathcal{A}[[v, w]]$, because $[\delta_v, \delta_w] = \delta_{[v,w]}$. Geometrically: $\mathcal{A}$ is a closed 1-form on the gauge orbit.
- The covariant anomaly $\mathrm{tr}[v\, \mathrm{ch}_m(F)]$ fails it (by a factor 2 in the commutator term, in 2D).
- Bose symmetry of identical currents: consistent $= \frac{1}{m+1} \times$ covariant for U(1) ($\frac{1}{2}$ from the chiral Schwinger model, $\frac{1}{3}$ from the triangle budget).
- Chern–Simons: $\mathrm{tr}\, F^k = \mathrm{d}Q_{2k-1}$, $Q_{2k-1} = k \int_0^1 \mathrm{tr}(A F_t^{k-1}) \mathrm{d}t$; $Q_3 = \mathrm{tr}(A\mathrm{d}A + \frac{2}{3}A^3)$, $Q_5 = \mathrm{tr}(A(\mathrm{d}A)^2 + \frac{3}{2}A^3 \mathrm{d}A + \frac{3}{5}A^5)$.
- Descent: $\mathrm{ch}_{m+1} = \mathrm{d}\omega^0$, $\delta_v \omega^0 = \mathrm{d}\omega^1$, $\delta_v \ln Z_F = -2\pi i \int \omega^1$: $\frac{i}{4\pi} \int \mathrm{tr}(v \mathrm{d}A)$ in 2D, $-\frac{1}{24\pi^2} \int \mathrm{tr}(v \mathrm{d}(A\mathrm{d}A + \frac{1}{2}A^3))$ in 4D. Inflow from a Chern–Simons bulk makes consistency automatic.
- 4D anomalies $\propto d^{abc} = \mathrm{tr}(T^a\{T^b, T^c\})$; real representations and SU(2) are safe; $A(\Lambda^2 \mathbf{N}) = N - 4$ makes $\bar{\mathbf{5}} \oplus \mathbf{10}$ safe.
- The Standard Model cancels all six nontrivial triangles within each generation, and the conditions fix the hypercharge ratios.

10.6 Anomalies and Yang–Mills theories V: the topology of the anomaly

We now have the anomaly as a formula. Nash's aim in the second half of §2 is to show that the formula is the shadow of a topological invariant [Nash p. 273]: "The rest of this section is devoted to a topological interpretation of the anomaly. This is made up of two parts." The question this section answers is: *what topological object does the anomaly measure, and how does the index theorem for families compute it?* The answer comes in two pictures that turn out to be the same.

1. *On the gauge group $\mathcal{G}$.* The variation of the (phase of the) determinant is a closed 1-form μ_1 on $\mathcal{G}$. It is exact precisely when there is no anomaly; a nontrivial class $[\mu_1] \in H^1_{\mathrm{dR}}(\mathcal{G})$ needs loops in $\mathcal{G}$, i.e. $\pi_1(\mathcal{G}) \neq 0$.
2. *On the orbit space $\mathcal{A}/\mathcal{G}$.* The determinant is not a function but a section of a line bundle, $\det \operatorname{Index} \not{D}$. There is no anomaly precisely when this bundle is trivial, which is detected by its first Chern class. The families index theorem computes that class.

10.6.1 The operator T_g and its determinant

Following Atiyah and Singer, Nash compares the chiral operator at A with the one at A_g through

$$T_g := \not{D}_A^{*}\, \not{D}_{A_g} \qquad \text{[Nash (10.12)]}, \tag{10.153}$$

where, as in Section 10.2.3, $\not{D}_A$ stands for the chiral operator $D_A^+ : \Gamma(E^+ \otimes F) \to \Gamma(E^- \otimes F)$ and $\not{D}_A^{*} = (D_A^+)^\dagger$ maps back. So $T_g : \Gamma(E^+ \otimes F) \to \Gamma(E^+ \otimes F)$ maps a space *to itself*, and its determinant is a number. Formally,

$$\det T_g \text{ "=" } \overline{\det D_A^+} \cdot \det D_{A_g}^+, \tag{10.154}$$

so as a function of g with A fixed it records how the chiral determinant changes along the orbit, the first factor being a fixed normalisation. For $g = 1$, $T_1 = (D_A^+)^\dagger D_A^+ \geq 0$ and $\det T_1 = |\det D_A^+|^2$.

T_g is elliptic [Nash (10.14)]. The leading symbol of a differential operator is obtained by keeping the highest derivatives and replacing $\partial_\mu \to i p_\mu$ (Chapter 2). For the chiral Dirac operator $D_A^+ = \gamma^\mu(\partial_\mu + A_\mu)P_+$ and its adjoint $-\gamma^\mu(\partial_\mu + A_\mu)P_-$ (recall $\partial_\mu^\dagger = -\partial_\mu$ and γ^μ Hermitian),

$$\sigma(T_g)(x,p) \overset{(1)}{=} \sigma(\not{D}_A^{*})\,\sigma(\not{D}_{A_g}) \overset{(2)}{=} \big(-i\gamma^\mu p_\mu\big)\big(g^{-1}\, i\gamma^\nu p_\nu\, g\big) \overset{(3)}{=} \gamma^\mu\gamma^\nu p_\mu p_\nu$$
$$\overset{(4)}{=} \tfrac{1}{2}\{\gamma^\mu,\gamma^\nu\}p_\mu p_\nu \overset{(5)}{=} g^{\mu\nu}p_\mu p_\nu\, \mathbf{1}.$$

(1) The leading symbol of a composition is the product of leading symbols.
(2) $\not{D}_{A_g} = g^{-1}\not{D}_A g$ (10.22); multiplication by g has symbol g (order 0), and A, A_g are lower order.
(3) $(-i)(i) = 1$; g acts on colour and $\gamma^\nu p_\nu$ on spin, so they commute and $g^{-1}g = 1$.
(4) $p_\mu p_\nu$ is symmetric, so only the symmetric part of $\gamma^\mu\gamma^\nu$ survives.
(5) The Clifford relation $\{\gamma^\mu,\gamma^\nu\} = 2g^{\mu\nu}$ (Euclidean, Riemannian metric).

This is Nash's "$g_{\mu\nu}p^\mu p^\nu I$". It is invertible for $p \neq 0$ because the metric is positive definite: T_g is elliptic, with the same leading symbol as a Laplacian.

But T_g need not be positive. The lower-order terms contain $g^{-1}\mathrm{d}g$, and they can push finitely many eigenvalues off the positive real axis. Elliptic theory (Seeley) guarantees that all but finitely many eigenvalues lie in a wedge around the positive real axis. The next example shows this happening.

Worked example 10.28: T_g on a circle: negative eigenvalues and a sign

On $M = S^1$ with coordinate $x \sim x + 2\pi$, take the "chiral" operator $D_a = \frac{\mathrm{d}}{\mathrm{d}x} + ia$ acting on functions, with a a real constant, $0 < a < 1$. Its adjoint is $D_a^\dagger = -\frac{\mathrm{d}}{\mathrm{d}x} - ia$. A gauge transformation $g(x) = e^{ikx}$, $k \in \mathbb{Z}$, shifts a:

$$g^{-1} D_a \, g \overset{(1)}{=} e^{-ikx}\Big(\frac{\mathrm{d}}{\mathrm{d}x} + ia\Big)e^{ikx} \overset{(2)}{=} \frac{\mathrm{d}}{\mathrm{d}x} + ik + ia = D_{a+k}.$$

(1) Definition of the conjugated operator, acting on a test function.
(2) Leibniz: $\frac{\mathrm{d}}{\mathrm{d}x}(e^{ikx} f) = e^{ikx}(f' + ikf)$.

On the Fourier modes e^{inx} both operators are diagonal:

$$\begin{aligned} T_g \, e^{inx} = D_a^\dagger D_{a+k} \, e^{inx} &\overset{(1)}{=} D_a^\dagger \, i(n+a+k) \, e^{inx} \overset{(2)}{=} (-i)(n+a)\, i(n+a+k)\, e^{inx} \\ &= (n+a)(n+a+k)\, e^{inx}. \end{aligned}$$

(1) $\frac{\mathrm{d}}{\mathrm{d}x} e^{inx} = in\, e^{inx}$.
(2) $D_a^\dagger e^{inx} = (-in - ia) e^{inx}$.

The eigenvalue $\lambda_n = (n+a)(n+a+k)$ is negative exactly when $n + a$ lies strictly between $-k$ and 0 (for $k > 0$), i.e. for $n = -1, -2, \dots, -k$: *exactly k negative eigenvalues*, all the rest positive and growing like n^2. This is Nash's "finite number of negative eigenvalues", and here it is g, not A, that causes it. The zeta-regularised determinant [Nash (10.15)] is

$$\det T_g = \big(\lambda_1^- \cdots \lambda_k^-\big) \, \exp\big[-\zeta'_{\hat{T}_g}(0)\big],$$

where $\hat{T}_g$ is T_g restricted to the positive eigenvalues, whose zeta function is defined as in Definition 10.4. Its sign is $(-1)^k$. Using the standard zeta-regularised product $|\prod_{n\in\mathbb{Z}}(n+a)| = 2|\sin \pi a|$ (from Lerch's formula $\prod_{n\geq 0}(n+a) = \sqrt{2\pi}/\Gamma(a)$ and $\Gamma(a)\Gamma(1-a) = \pi/\sin\pi a$) for each factor, the modulus is $4\sin^2 \pi a$, independent of k: the modulus is gauge invariant, as the pitfall of Section 10.2.6 said, and the gauge transformation shows up only in the sign,

$$\frac{\det D_{a+k}}{\det D_a} = \frac{\sin \pi(a+k)}{\sin \pi a} = (-1)^k.$$

A large gauge transformation (one that winds round U(1)) flips the sign of the chiral determinant when k is odd. This is a first glimpse of a *global* anomaly (Nash §5): invisible to any infinitesimal variation, since e^{ikx} with $k \neq 0$ is not connected to the identity.

Zero modes [Nash p. 274]. If T_g has a zero eigenvalue, $\det T_g = 0$ and its logarithm is useless. Nash distinguishes two cases.

- index $D_A^+ = 0$. Then zero modes are not forced. The index is unchanged by small perturbations (Chapter 4), and a generic small perturbation of A removes the kernel of D_A^+; since g is invertible, $\ker D_{A_g}^+ = g^{-1} \ker D_A^+$ is then empty too, and $\ker(D_A^+)^\dagger$ is empty because the index is zero. In the finite-dimensional model a square matrix is generically invertible. In Example 10.28, $a \in \mathbb{Z}$ gives a zero mode, and moving a off the integers removes it.

- index $D_A^+ \neq 0$. Then zero modes are forced: a $k \times (k+1)$ matrix always has a kernel. Say index $D_A^+ > 0$. We can perturb A so that $\ker(D_A^+)^\dagger = 0$, but $\ker D_{A_g}^+ \neq 0$ for every g. We

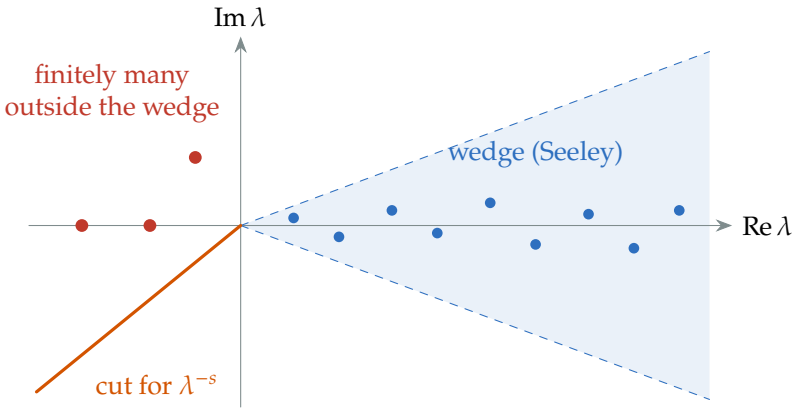


Figure 10.20: The spectrum of the non-self-adjoint elliptic operator T_g. Almost all eigenvalues (blue) lie in a wedge around the positive real axis, because the leading symbol is that of a Laplacian. Finitely many (red), produced by the lower-order terms containing $g^{-1}\mathrm{d}g$, may lie elsewhere, e.g. on the negative axis as in Example 10.28. To define $\lambda^{-s} = e^{-s\ln\lambda}$ one picks a branch cut (orange) outside the wedge; the finitely many stray eigenvalues are multiplied in by hand, as in (10.15).

restrict T_g to the orthogonal complement of $\ker D^+_{A_g}$, where it has no zero eigenvalue. As g varies, $\ker D^+_{A_g} = g^{-1}\ker D^+_A$ sweeps out a vector bundle over the orbit $\{A_g\}$, of rank index D^+_A.

Either way, $\det T_g$ is now a nowhere-vanishing function of $g \in \mathcal{G}$.

10.6.2 The anomaly as a 1-form on the gauge group

Define, with d the exterior derivative on the infinite-dimensional manifold $\mathcal{G}$,

$$\mu_1 := \frac{\mathrm{d}\,\det T_g}{\det T_g} \qquad \text{[Nash (10.13), (10.16)].} \tag{10.155}$$

Its value on the tangent vector at g generated by $v \in \mathfrak{g}$ (the curve $t \mapsto g e^{tv}$) is

$$\mu_1(X_v)\big|_g \overset{(1)}{=} \frac{\mathrm{d}}{\mathrm{d}t}\ln\det T_{ge^{tv}}\Big|_{t=0} \overset{(2)}{=} \frac{\mathrm{d}}{\mathrm{d}t}\ln\det D^+_{(A_g)_{e^{tv}}}\Big|_{t=0} \overset{(3)}{=} \delta_v \ln Z_F[A_g] \overset{(4)}{=} \mathcal{A}[v;A_g]. \tag{10.156}$$

(1) The derivative of $\ln f$ along a curve is $f^{-1}\,\mathrm{d}f/\mathrm{d}t$.
(2) (10.154): only the second factor depends on g; and $A_{ge^{tv}} = (A_g)_{e^{tv}}$.
(3) $Z_F = \det D^+$ and the definition (10.141) of δ_v at the point A_g.
(4) The consistent anomaly, computed in Section 10.5.5.

So μ_1 *is the consistent anomaly*, viewed as a 1-form on $\mathcal{G}$. Being $\mathrm{d}f/f$, it is closed wherever $f \neq 0$ (locally $\mu_1 = \mathrm{d}\ln f$ and $\mathrm{d}\mathrm{d} = 0$); closedness on the orbit is the Wess–Zumino condition of Section 10.5.2.

When is μ_1 exact? If a global logarithm $\ln\det T_g$ existed on $\mathcal{G}$, then $\mu_1 = \mathrm{d}\ln\det T_g$ would be exact and we could redefine the fermion determinant by a g-independent phase (in the local version, a counterterm) to make it invariant. A logarithm of a nowhere-zero complex function $f : W \to \mathbb{C}\setminus\{0\}$ exists iff no loop in W is mapped by f to a loop encircling 0. The standard example:

Worked example 10.29: The winding number obstructs the logarithm

On $W = S^1$, $f(\theta) = e^{in\theta}$ with $n \in \mathbb{Z}$. Then

$$\frac{\mathrm{d}f}{f} \overset{(1)}{=} in\,\mathrm{d}\theta, \qquad \oint_{S^1} \frac{\mathrm{d}f}{f} \overset{(2)}{=} 2\pi i\, n.$$

(1) $\mathrm{d}e^{in\theta} = in\, e^{in\theta}\mathrm{d}\theta$.
(2) $\oint \mathrm{d}\theta = 2\pi$.

If $\mathrm{d}f/f = \mathrm{d}h$ for a single-valued h, Stokes would give $\oint \mathrm{d}h = 0$ (a closed curve has no boundary). So for $n \neq 0$ there is no logarithm, even though $\mathrm{d}f/f$ is closed. The integer $\frac{1}{2\pi i}\oint \mathrm{d}f/f$ is the winding number of f around 0, and $[\mathrm{d}f/f] \in H^1_{\mathrm{dR}}(S^1) \cong \mathbb{R}$ is in times the generator. The same thing in finite-dimensional matrix language: $U(t) = \mathrm{diag}(e^{2\pi i t}, 1) \in \mathrm{U}(2)$, $t \in [0,1]$, is a loop with $\oint \mathrm{d}\ln\det U = 2\pi i$; the determinant detects the generator of $\pi_1(\mathrm{U}(2)) = \mathbb{Z}$.

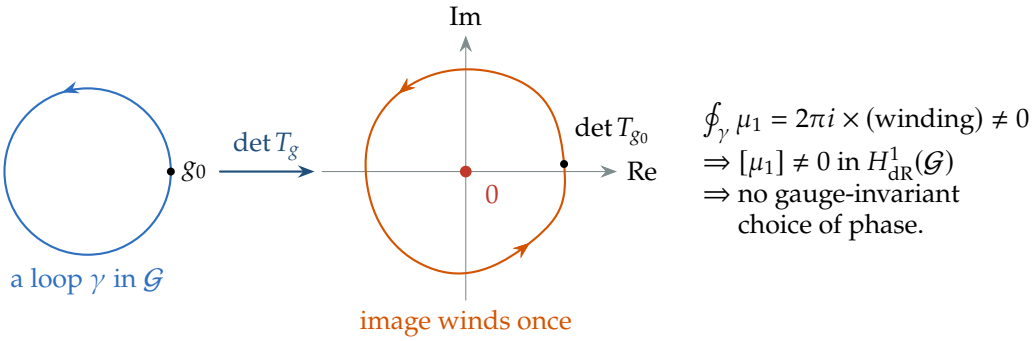


Figure 10.21: Nash's criterion for exactness of μ_1. A loop in the gauge group is mapped by the determinant to a loop in $\mathbb{C} \setminus \{0\}$. If the image winds round the origin, $\ln\det T_g$ cannot be defined continuously and μ_1 is not exact. Such loops exist only if $\pi_1(\mathcal{G}) \neq 0$.

Loops in $\mathcal{G}$ [Nash (10.17)–(10.18)]. If $\mathcal{G}$ is simply connected, every loop can be shrunk, the winding is zero, and μ_1 is exact. So topology enters through $\pi_1(\mathcal{G})$. Take $M = S^{2m}$ and (to make the action free) the *based* gauge group $\mathcal{G} = \mathrm{Map}_*(S^{2m}, G)$, maps sending a base point to 1. A loop in $\mathcal{G}$ is a family $g_t(x)$, $t \in S^1$, i.e. a map $S^1 \times S^{2m} \to G$ that is 1 whenever $t = 0$ or x is the base point. Collapsing $S^1 \times \{*\} \cup \{0\} \times S^{2m}$ to a point turns $S^1 \times S^{2m}$ into the sphere S^{2m+1} (the smash product $S^1 \wedge S^{2m}$). Hence

$$\pi_1(\mathcal{G}) = \pi_{2m+1}(G).$$

(Nash writes π_{2n+1}, with his n equal to our m, half the dimension of M.) For $M = S^4$ and $G = \mathrm{U}(N)$:

- $N = 1$: $\pi_5(\mathrm{U}(1)) = 0$, because $\mathrm{U}(1) = S^1$ has the contractible universal cover $\mathbb{R}$, so all its higher homotopy groups vanish.
- $N = 2$: $\pi_5(\mathrm{U}(2)) = \pi_5(\mathrm{SU}(2)) = \pi_5(S^3) = \mathbb{Z}_2$, a torsion group, which de Rham cohomology cannot see (H^1_{dR} is a real vector space).
- $N \geq 3$: $\pi_5(\mathrm{U}(N)) = \mathbb{Z}$ (Bott periodicity: $\pi_{2k-1}(\mathrm{U}(N)) = \mathbb{Z}$ once $N \geq k$). A generator is detected by the integral $\int_{S^5} \mathrm{tr}(g^{-1}\mathrm{d}g)^5$ (the Q_5 of a pure gauge, Example 10.23).

So for $N \geq 3$ there is room for $[\mu_1] \neq 0$ in $H^1_{\mathrm{dR}}(\mathcal{G})$, and indeed it is nonzero: integrating μ_1 round the generating loop gives $2\pi i$ times a nonzero integer proportional to the anomaly

coefficient d^{abc} of the fermion representation. This matches Section 10.5.6: SU($N \geq 3$) with a triplet-type representation is anomalous, SU(2) is not (perturbatively), and U(1) is a puzzle to which we return in Section 10.6.6.

10.6.3 Determinants in linear algebra: the determinant line

The second picture needs a precise notion of the determinant of a map between two *different* spaces. Let $O : V \to W$ be linear with $\dim V = \dim W = k$. The top exterior powers $\Lambda^k V$, $\Lambda^k W$ are one-dimensional and O induces a map between them,

$$\Lambda^k O : \; v_1 \wedge \cdots \wedge v_k \;\longmapsto\; Ov_1 \wedge \cdots \wedge Ov_k. \tag{10.157}$$

A linear map between one-dimensional spaces is "a number times an identification", i.e. an element of $\mathrm{Hom}(\Lambda^k V, \Lambda^k W) \cong (\Lambda^k V)^* \otimes \Lambda^k W$ [Nash (10.21)]. That element is $\det O$.

Worked example 10.30: The 2×2 determinant as a wedge product

Let $V = W = \mathbb{C}^2$ with basis e_1, e_2 and $O = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$, so $Oe_1 = ae_1 + ce_2$, $Oe_2 = be_1 + de_2$. Then

$$Oe_1 \wedge Oe_2 \overset{(1)}{=} ab\, e_1 \wedge e_1 + ad\, e_1 \wedge e_2 + cb\, e_2 \wedge e_1 + cd\, e_2 \wedge e_2 \overset{(2)}{=} (ad - bc)\, e_1 \wedge e_2.$$

(1) Bilinearity of $\wedge$.
(2) $e_i \wedge e_i = 0$ and $e_2 \wedge e_1 = -e_1 \wedge e_2$.

So $\Lambda^2 O = (ad - bc)\,(e_1 \wedge e_2)^* \otimes (e_1 \wedge e_2)$. If $V \neq W$ and we rescale the basis of W alone, $e_i' = 2e_i$, the *number* in front changes ($e_1 \wedge e_2 = \frac{1}{4} e_1' \wedge e_2'$), but the element of the line $(\Lambda^2 V)^* \otimes \Lambda^2 W$ does not. For a chiral Dirac operator $V = \Gamma(E^+)$ and $W = \Gamma(E^-)$ are different, so its determinant is canonically a vector in a line, not a number. This is the precise meaning of the remark in Section 10.2.5.

Unequal dimensions and kernels. If $\dim V \neq \dim W$, or O is not invertible, we split off the kernels. With inner products on V and W and the adjoint O^*,

$$V = \mathrm{Range}\, O^* \oplus \ker O, \qquad W = \mathrm{Range}\, O \oplus \ker O^* \quad \text{[Nash (10.22)]}. \tag{10.158}$$

The first holds because $\ker O$ is exactly the orthogonal complement of $\mathrm{Range}\, O^*$:

$$v \in \ker O \overset{(1)}{\iff} \langle Ov, w\rangle = 0 \;\; \forall w \in W \overset{(2)}{\iff} \langle v, O^* w\rangle = 0 \;\; \forall w \iff v \perp \mathrm{Range}\, O^*,$$

(1) A vector is zero iff it is orthogonal to everything.
(2) Definition of the adjoint.

and the second is the same statement for O^*. Next, top powers of direct sums factorise:

$$\Lambda^{\max}(X \oplus Y) \cong \Lambda^{\max} X \otimes \Lambda^{\max} Y, \qquad (x_1 \wedge \cdots \wedge x_p) \wedge (y_1 \wedge \cdots \wedge y_q) \leftrightarrow (x_1 \wedge \cdots \wedge x_p) \otimes (y_1 \wedge \cdots \wedge y_q), \tag{10.159}$$

since a basis of X followed by a basis of Y is a basis of $X \oplus Y$. Therefore [Nash (10.23)]

$$(\Lambda^{\max} V)^* \otimes \Lambda^{\max} W \overset{(1)}{=} \left(\Lambda^{\max} \mathrm{Range}\, O^*\right)^* \otimes (\Lambda^{\max} \ker O)^* \otimes \Lambda^{\max} \mathrm{Range}\, O \otimes \Lambda^{\max} \ker O^*.$$

(1) (10.158) and (10.159), with $(X \otimes Y)^* = X^* \otimes Y^*$ for lines.

On the ranges, O restricts to an *isomorphism* $\mathrm{Range}\, O^* \to \mathrm{Range}\, O$ (it is injective there, since the kernel has been removed, and the two ranges have the same dimension, the rank). So its

determinant is a *nonzero* element of $(\Lambda^{\max}\operatorname{Range} O^*)^* \otimes \Lambda^{\max}\operatorname{Range} O$, and we may use it to identify that line with $\mathbb{C}$. What remains is Nash's refined determinant line [Nash (10.24)]:

$$\det O \;\in\; (\Lambda^{\max}\ker O)^* \otimes (\Lambda^{\max}\ker O^*). \tag{10.160}$$

When $\ker O = 0 = \ker O^*$ both factors are $\Lambda^0 = \mathbb{C}$ canonically and $\det O$ is a nonzero number: this is the normalisation in Nash's footnote 2.

Worked example 10.31: Kernels that jump, a line that does not

Take $O_y : \mathbb{C}^2 \to \mathbb{C}^2$, $O_y = \left(\begin{smallmatrix} y & 0 \\ 0 & 0 \end{smallmatrix}\right)$, $y \in \mathbb{C}$.

- For $y \neq 0$: $\ker O_y = \ker O_y^* = \mathbb{C}e_2$, so (10.160) is the line $(\mathbb{C}e_2)^* \otimes \mathbb{C}e_2$, and the refined determinant is $y\, e_2^* \otimes e_2$: the determinant of the invertible part $y : \mathbb{C}e_1 \to \mathbb{C}e_1$ times the kernel line.
- For $y = 0$: $\ker O_0 = \ker O_0^* = \mathbb{C}^2$, and the line is $(\Lambda^2\mathbb{C}^2)^* \otimes \Lambda^2\mathbb{C}^2$.

The kernels jump from dimension 1 to 2, but both lines are canonically $\mathbb{C}$, and the determinant $y \mapsto y$ is a continuous (indeed holomorphic) section of a smooth line bundle over the y-plane, vanishing at $y = 0$. The general rule: if both kernels grow by the same space U, the line $(\Lambda^{\max}(K \oplus U))^* \otimes \Lambda^{\max}(K' \oplus U) \cong (\Lambda^{\max}K)^* \otimes (\Lambda^{\max}U)^* \otimes \Lambda^{\max}U \otimes \Lambda^{\max}K'$ is canonically $(\Lambda^{\max}K)^* \otimes \Lambda^{\max}K'$, because $(\Lambda^{\max}U)^* \otimes \Lambda^{\max}U = \mathbb{C}$ canonically (pair a vector with its dual). Kernels jump; the index (here $1 - 1 = 0$) and the determinant line do not.

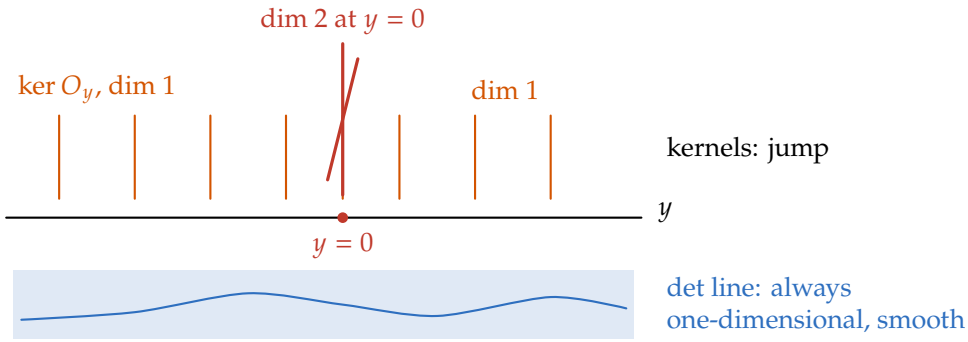


Figure 10.22: The kernel of a family of operators can jump in dimension at special parameter values (red), so $y \mapsto \ker O_y$ is not a vector bundle. The determinant line $(\Lambda^{\max}\ker O)^* \otimes \Lambda^{\max}\ker O^*$ is: extra kernel in $\ker O$ is matched by extra kernel in $\ker O^*$ (the index is constant), and the two cancel in the line.

10.6.4 Families of Dirac operators and the determinant line bundle

Now the whole partition function [Nash (10.19)] $Z = \int_{\mathcal{A}} \mathcal{D}A\, \sqrt{\det(\not{D}_A^*\not{D}_A)}\, e^{-\|F\|^2}$ involves not one Dirac operator but the whole family $\{D_A^+\}_{A\in\mathcal{A}}$. Chapter 4 (Chapter 4) attaches to a family of elliptic operators parametrised by a space Y an *index bundle* in the K-theory of Y:

$$\operatorname{Index}\not{D} = \{\ker D_A^+ : A \in \mathcal{A}\} - \{\ker(D_A^+)^\dagger : A \in \mathcal{A}\} = [\ker\not{D}] - [\ker\not{D}^*] \in K(\mathcal{A}) \qquad \text{[Nash (10.20)]}. \tag{10.161}$$

(Reminder: $K(Y)$ consists of formal differences $[V] - [W]$ of vector bundles over Y, with $[V]-[W] = [V']-[W']$ when $V \oplus W' \oplus U \cong V' \oplus W \oplus U$ for some U. The individual kernels are not bundles, since they jump, but their difference is well defined: add a fixed finite-dimensional space to the target to kill the cokernel locally, as in Example 10.31.) Because $D^+_{A_g} = g^{-1}D_A^+ g$

(10.22), the kernels at A and A_g are identified by g, and the index bundle descends to an element of $K(\mathcal{A}/\mathcal{G})$.

The determinant line of a K-theory class [Nash (10.25)–(10.28)]. To a vector bundle E of rank k associate its **determinant bundle** $\det E$, the line bundle with fibre $\Lambda^k E_y$ (Nash's footnote 3). To a K-theory element $[V]-[W]$ associate $(\det V)^* \otimes \det W$ [Nash (10.27)]. This does not depend on the representative:

$$\big(\det(V\oplus U)\big)^* \otimes \det(W\oplus U) \overset{(1)}{=} (\det V)^* \otimes (\det U)^* \otimes \det W \otimes \det U \overset{(2)}{=} (\det V)^* \otimes \det W,$$

(1) (10.159) fibre by fibre.
(2) $(\det U)^* \otimes \det U$ is canonically trivial.

and $[V\oplus U]-[W\oplus U]=[V]-[W]$ is the only freedom in the representative. Applied to the Dirac family,

$$\det \mathrm{Index}\, \not{D} = (\det\ker \not{D})^* \otimes (\det \ker \not{D}^*) \quad \text{over } \mathcal{A}/\mathcal{G} \qquad \text{[Nash (10.28)]}, \tag{10.162}$$

and (10.160) says that the chiral determinant $\det D_A^+$ is a *section* of this line bundle.

Working definition 10.8 (determinant line bundle)

The **determinant line bundle** of a family of Fredholm operators $\{O_y\}_{y\in Y}$ [operators that can have zero modes, i.e. nonzero solutions $\psi\neq 0$ of $O_y\psi=0$. How many independent zero modes there are can change as y varies] is the line bundle over Y [a copy of $\mathbb{C}$ attached to each point y] whose fibre at y is $(\Lambda^{\max}\ker O_y)^* \otimes \Lambda^{\max}\ker O_y^*$ [the line at y where $\det O_y$ lives. Every such line looks like $\mathbb{C}$, but only when O_y has no zero modes is there a natural way to read $\det O_y$ as an ordinary number], glued across jumps of the kernels as in Example 10.31. The determinant $y\mapsto \det O_y$ is a section [it picks one point from each of these copies], vanishing exactly where O_y has zero modes. *To test for an anomaly:* compute c_1 [an integer measuring how much the copies of $\mathbb{C}$ twist as one goes around Y] with the families index theorem below. If $c_1\neq 0$, no nowhere-vanishing section compatible with the gauge identifications exists, so the determinant cannot be made gauge invariant.

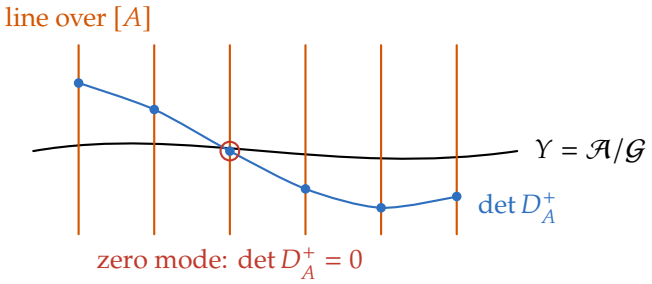


Figure 10.23: The determinant line bundle. Over each gauge field $[A]$ stands a line (orange). The chiral determinant picks a point on every line (blue), so it is a section. The base curve is where each line has its zero, and the section crosses it where D_A^+ has a zero mode (red). If the lines are glued with a twist ($c_1\neq 0$), every section must cross zero and no gauge-invariant choice of the determinant exists: the anomaly.

The Chern class of a determinant bundle. We need $c_1(\det E)=c_1(E)$. For a direct sum of line bundles $E=L_1\oplus L_2$ we have $\det E=L_1\otimes L_2$, and c_1 adds under tensor products, so

$c_1(\det E) = c_1(L_1) + c_1(L_2) = c_1(E)$. In curvature language (the general case, which the splitting principle reduces to this one): the induced connection on $\Lambda^k E$ has curvature $\operatorname{tr} F_E$, so

$$c_1(\det E) \overset{(1)}{=} \frac{i}{2\pi} F_{\det E} \overset{(2)}{=} \frac{i}{2\pi} \operatorname{tr} F_E \overset{(3)}{=} c_1(E),$$

(1) Definition of c_1 for a line bundle.
(2) On $e_1 \wedge \cdots \wedge e_k$ the curvature acts by the Leibniz rule, $\sum_i e_1 \wedge \cdots \wedge F_E e_i \wedge \cdots \wedge e_k = (\operatorname{tr} F_E)\, e_1 \wedge \cdots \wedge e_k$.
(3) Definition of $c_1(E)$ (Chapter T2).

and by linearity $c_1(\det \operatorname{Index} \not{D}) = c_1(\operatorname{Index} \not{D})$.

The families index theorem [Nash (10.29)–(10.30)]. Chapter 4's formula (Nash's (4.126)) for the Chern character of the index bundle, specialised to the spin complex on an even-dimensional M, reads

$$\operatorname{ch}(\operatorname{Index} \not{D}) = \int_M \operatorname{ch}(\mathcal{E})\, \hat{A}(M) \overset{M=S^n}{=} \int_{S^n} \operatorname{ch}(\mathcal{E}). \tag{10.163}$$

Here $\mathcal{E}$ is the **universal bundle** over $Z = M \times \mathcal{A}/\mathcal{G}$: over the point $(x, [A])$ its fibre is the fibre F_x of the colour bundle, and it carries a natural connection whose restriction to $M \times \{[A]\}$ is A. (Construct it as $(F \times \mathcal{A})/\mathcal{G}$; its curvature $\mathbb{F}$ has a $(2,0)$ part F_A along M, a $(1,1)$ part mixing M and $\mathcal{A}/\mathcal{G}$, and a $(0,2)$ part along $\mathcal{A}/\mathcal{G}$.) The integral $\int_M$ lowers the degree by n: the degree-$2j$ part of the left side comes from the degree-$(n+2j)$ part of $\operatorname{ch}(\mathcal{E})\hat{A}$. For c_1 we need $j = 1$:

$$c_1(\det \operatorname{Index} \not{D}) = \int_M \big[\operatorname{ch}(\mathcal{E})\hat{A}\big]_{(n+2)} \overset{M=S^{2m}}{=} \int_{S^{2m}} \operatorname{ch}_{m+1}(\mathbb{F}) \in H^2_{\mathrm{dR}}(\mathcal{A}/\mathcal{G}). \tag{10.164}$$

This is the moment to look back at the descent equations: ch_{m+1} is the Ω_{2m+2} of (10.146). The "imagined $(2m+2)$-dimensional space" on which it lived is $M \times \Sigma$ for a 2-cycle Σ of gauge orbits: the characteristic class that seeds the descent is exactly the one that measures the determinant line bundle.

Worked example 10.32: Flat connections on a torus: a determinant line bundle with $|c_1| = 1$

Let $M = T^2 = \mathbb{R}^2/\mathbb{Z}^2$ with coordinates $x_1, x_2 \in [0,1)$ and flat metric, and a single positive-chirality fermion of charge one. Consider the two-parameter family of flat connections

$$A_\theta = 2\pi i\,(\theta_1 \mathrm{d}x_1 + \theta_2 \mathrm{d}x_2), \qquad \theta = (\theta_1, \theta_2) \in \mathbb{R}^2.$$

The orbit space. The gauge transformation $g_n = e^{2\pi i(n_1 x_1 + n_2 x_2)}$, $n \in \mathbb{Z}^2$, is well defined on T^2 and gives $(A_\theta)_{g_n} = A_\theta + g_n^{-1}\mathrm{d}g_n = A_{\theta+n}$. So the flat connections modulo gauge form the dual torus $\hat{T}^2 = \mathbb{R}^2/\mathbb{Z}^2 \ni \theta$, a two-dimensional piece of $\mathcal{A}/\mathcal{G}$.
The spectrum. By Example 10.3, $D^+_\theta = D_1 + iD_2 = \partial_1 + i\partial_2 + 2\pi i(\theta_1 + i\theta_2)$. On the plane waves $e_k = e^{2\pi i(k_1 x_1 + k_2 x_2)}$, $k \in \mathbb{Z}^2$,

$$D^+_\theta e_k \overset{(1)}{=} \big[2\pi i k_1 + i \cdot 2\pi i k_2 + 2\pi i(\theta_1 + i\theta_2)\big] e_k \overset{(2)}{=} 2\pi i\,(k + \theta)\, e_k,$$

(1) $\partial_j e_k = 2\pi i k_j e_k$.
(2) Complex notation $k = k_1 + ik_2$, $\theta = \theta_1 + i\theta_2$.

So "$\det D^+_\theta = \prod_{k \in \mathbb{Z}[i]} 2\pi i(k+\theta)$" (regularised by the zeta function of $(D^+)^\dagger D^+$, with eigenvalues $4\pi^2|k+\theta|^2$). The kernels are nonzero only when $k + \theta = 0$ for some k, i.e. at the single point $\theta \equiv 0$ of $\hat{T}^2$, where $\ker D^+ = \ker(D^+)^\dagger = \mathbb{C} \cdot 1$: the index is $1 - 1 = 0$ everywhere, as it must be for a trivial bundle ($\int_{T^2} \operatorname{ch}_1 = 0$), but the kernels jump at one point.

The determinant is a section, not a function. Each factor $k+\theta$ depends holomorphically on θ, and the regularised determinant is a holomorphic object with a single simple zero on $\hat{T}^2$. It cannot be a function on $\hat{T}^2$: a holomorphic function on a compact torus is constant, and this one has a zero. It is a section of the determinant line bundle, periodic in θ only up to a multiplier ($\det D^+_{\theta+n} = \chi_n(\theta)\det D^+_\theta$), exactly like a Jacobi theta function. For a holomorphic section of a holomorphic line bundle on a compact Riemann surface, the number of zeros (with multiplicity) equals the degree c_1. Hence $|c_1(\det \mathrm{Index})| = 1$ and the bundle is not trivial.

The families index agrees. On $T^2 \times \hat{T}^2$ the universal connection is $\mathbb{A} = 2\pi i(\theta_1 \mathrm{d}x_1 + \theta_2 \mathrm{d}x_2)$, now with θ a coordinate too (the gluing by g_n makes the bundle over $T^2 \times \hat{T}^2$ the Poincaré bundle). Then

$$\mathbb{F} = \mathrm{d}\mathbb{A} \overset{(1)}{=} 2\pi i\,(\mathrm{d}\theta_1 \wedge \mathrm{d}x_1 + \mathrm{d}\theta_2 \wedge \mathrm{d}x_2),$$

$$\mathrm{ch}_2(\mathbb{F}) = \frac{1}{2}\Big(\frac{i}{2\pi}\mathbb{F}\Big)^2 \overset{(2)}{=} \frac{1}{2}\big(\mathrm{d}\theta_1\mathrm{d}x_1 + \mathrm{d}\theta_2\mathrm{d}x_2\big)^2 \overset{(3)}{=} \mathrm{d}\theta_1\mathrm{d}x_1\mathrm{d}\theta_2\mathrm{d}x_2 \overset{(4)}{=} -\mathrm{d}x_1\mathrm{d}x_2\,\mathrm{d}\theta_1\mathrm{d}\theta_2,$$

$$c_1(\det \mathrm{Index}) = \int_{T^2} \mathrm{ch}_2(\mathbb{F}) \overset{(5)}{=} -\mathrm{d}\theta_1 \wedge \mathrm{d}\theta_2, \qquad \int_{\hat{T}^2} c_1 \overset{(6)}{=} -1.$$

(1) d acts on θ as well: $\mathrm{d}(\theta_j \mathrm{d}x_j) = \mathrm{d}\theta_j \wedge \mathrm{d}x_j$.

(2) $\frac{i}{2\pi}\cdot 2\pi i = -1$, and the sign disappears on squaring.

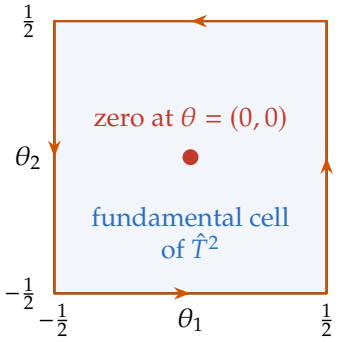
(3) The squares $(\mathrm{d}\theta_1\mathrm{d}x_1)^2$ vanish; the two cross terms are equal (2-forms commute), cancelling the $\frac{1}{2}$.

(4) Reordering $(\theta_1, x_1, \theta_2, x_2) \to (x_1, x_2, \theta_1, \theta_2)$ is an odd permutation (three transpositions).

(5) (10.164) with $\hat{A}(T^2) = 1$ (flat); integrate over $M = T^2$ with $\int \mathrm{d}x_1\mathrm{d}x_2 = 1$.

(6) $\int_{\hat{T}^2} \mathrm{d}\theta_1\mathrm{d}\theta_2 = 1$.

So $|c_1| = 1$, in agreement with the count of zeros. (The sign depends on orientation conventions and on whether one uses Nash's $(\det\ker)^* \otimes \det\ker^*$ or its dual; the magnitude does not.) The chiral U(1) theory on T^2 is anomalous, as the two-dimensional anomaly $\frac{i}{4\pi}\int \mathrm{tr}(v\,\mathrm{d}A) \neq 0$ predicted.



Going round the boundary of the cell, the phase of $\det D^+_\theta$ (measured against the multipliers that glue opposite sides) winds once: one zero inside, $|c_1(\det \mathrm{Index})| = 1$. A gauge-invariant, nowhere-vanishing choice of $\det D^+_\theta$ is impossible.

Figure 10.24: The determinant line bundle over the dual torus of flat U(1) connections on T^2 (Example 10.32). The kernel jumps only at $\theta \equiv 0$; that is where the section $\det D^+_\theta$ vanishes. Counting zeros of a holomorphic section computes c_1, and the families index theorem computes the same number from $\int_{T^2} \mathrm{ch}_2$ of the universal bundle.

10.6.5 The two pictures are one [Nash (10.31)–(10.34)]

We now have two classes [Nash (10.31)]:

$$[\mu_1] = \left[\frac{\mathrm{d}\,\det T_g}{\det T_g}\right] \in H^1_{\mathrm{dR}}(\mathcal{G}) \qquad \text{and} \qquad c_1(\det \operatorname{Index} \not{D}) \in H^2_{\mathrm{dR}}(\mathcal{A}/\mathcal{G}),$$

and Nash asserts (using his Chapter 8, Chapter 8) that [Nash (10.32)]

$$\mu_1 = 0 \iff c_1(\det \operatorname{Index} \not{D}) = 0.$$

Here is why. The space $\mathcal{A}$ of connections is an affine space (any two connections differ by a 1-form, and $A_0 + s(A_1 - A_0)$ is a connection for all s), hence contractible. So:

1. The pull-back of $\det \operatorname{Index} \not{D}$ to $\mathcal{A}$ is trivial (every bundle over a contractible space is), and the chiral determinant $A \mapsto \det D_A^+$ is simply a function on $\mathcal{A}$.
2. Along a gauge orbit this function is not invariant: $\det D^+_{A_g} = \chi(A, g)\, \det D_A^+$ for a nowhere-zero **multiplier** χ. This is how a line bundle over the quotient $\mathcal{A}/\mathcal{G}$ is described upstairs, exactly as theta functions describe line bundles over a torus by their multipliers under the lattice.
3. $\mathrm{d}_g \ln \chi$ is μ_1. Changing the trivialisation upstairs (multiplying $\det D_A^+$ by a nowhere-zero function $e^{\Phi(A)}$) changes μ_1 by the exact form $\mathrm{d}\Phi$. The bundle downstairs is trivial iff χ can be made $\equiv 1$, i.e. iff μ_1 is exact.

In topological language this is the isomorphism $\pi_1(\mathcal{G}) \cong \pi_2(\mathcal{A}/\mathcal{G})$ from the exact homotopy sequence of $\mathcal{G} \to \mathcal{A} \to \mathcal{A}/\mathcal{G}$ with $\mathcal{A}$ contractible (so $\mathcal{A}/\mathcal{G} = B\mathcal{G}$, the classifying space): a loop in $\mathcal{G}$ is the "equator" of a 2-sphere in $\mathcal{A}/\mathcal{G}$, and the winding of det round the loop equals c_1 on the sphere. (A familiar finite-dimensional instance: for the Hopf bundle $S^1 \to S^3 \to S^2$, the winding of a section's phase round the equator equals $c_1 = \pm 1$ on S^2.)

No anomaly $\Rightarrow$ a global section [Nash (10.34)]. Suppose there is no anomaly, so that $\det D_A^+$ can be chosen gauge invariant. Then the integrand of Z is constant on orbits and the integral over $\mathcal{A}$ reduces to an integral over $\mathcal{A}/\mathcal{G}$ (Faddeev–Popov):

$$\int_{\mathcal{A}} \mathcal{D}A\ \det(\not{D})\, e^{-\|F\|^2} = \int_{\mathcal{A}/\mathcal{G}} \mathcal{D}A_{\mathcal{G}}\ \mathrm{Jac}\ \det(\not{D})\, e^{-\|F\|^2},$$

with $\mathcal{D}A_{\mathcal{G}}$ the measure on orbits, Jac the Faddeev–Popov Jacobian (and the constant volume of $\mathcal{G}$ dropped). A finite-dimensional picture of the Jacobian:

Worked example 10.33: Integrating over orbits in the plane

Let U(1) act on $\mathbb{R}^2$ by rotations, and let $f(\vec{x}) = f(r)$ be invariant. The orbits are circles, the orbit space is $[0, \infty) \ni r$, and

$$\int_{\mathbb{R}^2} f\, \mathrm{d}^2x \overset{(1)}{=} \int_0^\infty \int_0^{2\pi} f(r)\, r\, \mathrm{d}\phi\, \mathrm{d}r \overset{(2)}{=} 2\pi \int_0^\infty \mathrm{d}r \underbrace{r}_{\mathrm{Jac}} f(r).$$

(1) Polar coordinates: $\mathrm{d}^2x = r\, \mathrm{d}r\, \mathrm{d}\phi$.
(2) The ϕ-integral is the volume of the group, 2π; the leftover r is the size of the orbit, the Jacobian.

The origin, a fixed point, is where the action is not free: the analogue of reducible connections (and of the isometries of Section 10.7), which is why Nash restricts to based gauge transformations.

In (10.162) language, a gauge-invariant $\det(\not{D})$ on $\mathcal{A}/\mathcal{G}$ is a *global* section of $\det \operatorname{Index} \not{D}$, nowhere vanishing (after the perturbations of Section 10.6.1); a line bundle with a nowhere-vanishing global section is trivial, so $c_1 = 0$. Contrapositive: $c_1(\det \operatorname{Index} \not{D}) \neq 0$ *implies an anomaly.* The index theorem for families thus "brings out very succinctly the anomaly as an obstacle to the definition of a gauge invariant determinant".

10.6.6 Local versus de Rham cohomology

There is a puzzle in (10.155) for $G = \mathrm{U}(1)$ on S^4: $\pi_1(\mathcal{G}) = \pi_5(\mathrm{U}(1)) = 0$, so $H^1_{\mathrm{dR}}(\mathcal{G}) = 0$ and μ_1 is exact, yet the triangle graph shows an anomaly that no counterterm removes. Nash's resolution [Nash p. 279]: $\mu_1 = \mathrm{d}\Phi$ with Φ *non-local*. The explicit counterterm makes this concrete.

Worked example 10.34: A non-local counterterm for the abelian anomaly

Take the four-dimensional U(1) consistent anomaly in the form $\mathcal{A}[\lambda] = c\int \lambda\, \epsilon^{\mu\nu\rho\sigma}\partial_\mu a_\nu \partial_\rho a_\sigma \,\mathrm{d}^4x$ (with $\delta a_\mu = \partial_\mu \lambda$ and a constant c), and let $\Box = \partial_\mu\partial_\mu$ with inverse $\Box^{-1}$ (convolution with the Green's function). Put

$$\Phi[a] := c\int (\partial_\mu a_\mu)\, \Box^{-1}\big(\epsilon^{\alpha\beta\gamma\delta}\partial_\alpha a_\beta \partial_\gamma a_\delta\big)\,\mathrm{d}^4x.$$

Then

$$\delta_\lambda \Phi \overset{(1)}{=} c\int (\partial_\mu\partial_\mu\lambda)\,\Box^{-1}\big(\epsilon\, \partial a\, \partial a\big) \overset{(2)}{=} c\int \lambda\, \Box\,\Box^{-1}\big(\epsilon\, \partial a\, \partial a\big) \overset{(3)}{=} c\int \lambda\, \epsilon\, \partial a\, \partial a = \mathcal{A}[\lambda].$$

(1) $\epsilon\, \partial a\, \partial a = \frac{1}{4}\epsilon f f$ is gauge invariant, so only $\partial \cdot a$ varies: $\delta(\partial_\mu a_\mu) = \Box\lambda$.
(2) Integrate $\Box$ by parts twice (it is symmetric).
(3) $\Box\Box^{-1} = 1$.

So $\ln Z_F - \Phi$ is gauge invariant: the anomaly is "d of something", as $H^1_{\mathrm{dR}}(\mathcal{G}) = 0$ demands. But Φ involves $\Box^{-1}$, i.e. it couples fields at distant points with a $1/|x-y|^2$ kernel. No *local* Φ (a polynomial in a and finitely many derivatives) works: the only parity-odd local 4-forms built from an abelian a are multiples of $a\,\mathrm{d}a\,\wedge(\cdot)$ with $a\wedge a = 0$, and $\int \mathrm{d}a\wedge \mathrm{d}a$ is gauge invariant. The same construction in two dimensions, $\Phi = c\int(\partial\cdot a)\Box^{-1}(\epsilon^{\mu\nu}\partial_\mu a_\nu)$, reproduces the chiral Schwinger model's non-local effective action.

Working definition 10.9 (local cohomology)

The **local cohomology** $H^1_{\mathrm{loc}}(\mathcal{G})$ is the space of local functionals $\mathcal{A}[v;A]$ (integrals of polynomials in v, A and finitely many derivatives) satisfying the Wess–Zumino condition [$\delta_v\mathcal{A}[w] - \delta_w\mathcal{A}[v] = \mathcal{A}[[v,w]]$], modulo those of the form $\delta_v\Phi_{\mathrm{loc}}$ with Φ_{loc} local [two anomalies count as the same if they differ by such a term, since a local counterterm removes it]. A physical anomaly is a nonzero class in H^1_{loc}. *To compute:* run the descent [write the polynomial as $\mathrm{d}\omega^0_{2m+1}$, then $\delta_v\omega^0_{2m+1} = \mathrm{d}\omega^1_{2m}$. The anomaly is $\int_M \omega^1_{2m}$] from each independent invariant polynomial of degree $m+1$ [a gauge-invariant polynomial in the curvature, such as $\operatorname{tr} F^{m+1}$]; each gives one class.

G on S^4	$\pi_1(\mathcal{G}) = \pi_5(G)$	$H^1_{\mathrm{dR}}(\mathcal{G})$	$H^1_{\mathrm{loc}}(\mathcal{G})$
U(1)	0	0	$\neq 0$: triangle anomaly $\propto \sum q^3$
SU(2)	$\mathbb{Z}_2$	0	0 (no d^{abc}); global anomaly in Nash §5
SU(N), $N \geq 3$	$\mathbb{Z}$	$\mathbb{R}$	$\neq 0$: $\propto d^{abc}$

The table shows that neither cohomology contains the other's information: de Rham cohomology misses the U(1) anomaly (which lives in the difference between "exact" and "exact with a local primitive"), and misses the SU(2) torsion. Nash cites Bonora and Cotta-Ramusino for the systematic theory; in physics language it is the BRST cohomology of local functionals, and the descent equations compute it.

Reading Nash (pp. 273–279, (10.12)–(10.34))

Nash's "first part" is Sections 10.6.1 and 10.6.2: the operator T_g (10.12), its elliptic symbol (10.14), the zeta determinant with finitely many negative eigenvalues (10.15), the 1-form μ_1 (10.13), (10.16), and the homotopy input (10.17)–(10.18). His "second part" is Sections 10.6.3 to 10.6.5: the index bundle (10.20), determinant lines (10.21)–(10.28), the families index theorem (10.29)–(10.30), and the equivalence (10.32). Two points he leaves implicit: μ_1 is exactly the consistent anomaly of Section 10.5.5 (10.156), and the ch_{m+1} that computes c_1 in (10.30) is exactly the polynomial that seeds the descent. His final paragraph (local cohomology) is Section 10.6.6. In (10.17) his n is half the dimension of M.

Recap

- $T_g = \not{D}^*_A \not{D}_{A_g}$ is elliptic (symbol $|p|^2\mathbf{1}$) but not positive; its zeta determinant multiplies in the finitely many stray eigenvalues. On S^1: $\lambda_n = (n+a)(n+a+k)$, k negative eigenvalues, sign $(-1)^k$.
- $\mu_1 = \mathrm{d}\ln\det T_g$ is the consistent anomaly as a closed 1-form on $\mathcal{G}$; it is exact iff det has no winding round loops of $\mathcal{G}$; $\pi_1(\mathcal{G}) = \pi_{2m+1}(G)$; $\pi_5(\mathrm{U}(N)) = 0, \mathbb{Z}_2, \mathbb{Z}$.
- The determinant of $O : V \to W$ lives in $(\Lambda^{\max} \ker O)^* \otimes \Lambda^{\max} \ker O^*$; kernels jump, the line does not.
- For the Dirac family: det Index $\not{D}$ over $\mathcal{A}/\mathcal{G}$, $c_1(\det \mathrm{Index}) = \int_M [\mathrm{ch}(\mathcal{E})\hat{A}]_{(n+2)}$; on T^2 with flat U(1) connections $|c_1| = 1$ (one zero of the determinant).
- $[\mu_1] \in H^1_{\mathrm{dR}}(\mathcal{G})$ and $c_1 \in H^2_{\mathrm{dR}}(\mathcal{A}/\mathcal{G})$ vanish together ($\mathcal{A}$ contractible); no anomaly $\Rightarrow$ global section $\Rightarrow c_1 = 0$.
- U(1): $H^1_{\mathrm{dR}} = 0$ yet anomalous, because the primitive is non-local; the physical object is local cohomology.

10.7 Gravitational anomalies

Gauge fields are not the only background a chiral fermion can feel. It also feels the metric, through the gamma matrices and the spin connection, and coordinate invariance plays the role of gauge invariance. The question this section answers is Nash's [Nash p. 279]: *can the fermion determinant, viewed as a function of the metric, fail to be invariant under diffeomorphisms, and if so in which dimensions?* The analysis is "analogous to that of the preceding section with $\mathcal{A}$ and $\mathcal{A}/\mathcal{G}$ replaced by $\mathrm{Met}(M)$ and $\mathrm{Met}(M)/\mathrm{Diff}_0(M)$", and the answer, due to Alvarez-Gaumé and

Witten, is: only in dimensions $n = 4k - 2$, i.e. $2, 6, 10, \ldots$

10.7.1 The Dirac operator depends on the metric

To concentrate on gravity we drop the Yang–Mills field. The chiral Dirac operator is then Nash's $\frac{1}{2}\gamma_\mu(\partial^\mu + \Gamma^\mu)(1+\gamma_5)$, where Γ_μ is the spin connection. Both ingredients depend on the metric $\rho \in \mathrm{Met}(M)$ (the space of all Riemannian metrics on M).

- *The gamma matrices.* The Clifford relation $\{\gamma^\mu, \gamma^\nu\} = 2g^{\mu\nu}$ involves the metric. In practice we fix constant matrices γ^a with $\{\gamma^a, \gamma^b\} = 2\delta^{ab}$ and an orthonormal frame (vielbein) $e_a = e_a{}^\mu \partial_\mu$, $g(e_a, e_b) = \delta_{ab}$, and put $\gamma^\mu = e_a{}^\mu \gamma^a$ (Chapter T4).
- *The spin connection.* With the dual coframe $e^a = e^a{}_\mu \mathrm{d}x^\mu$, the Levi-Civita connection in the frame is the antisymmetric matrix of 1-forms $\omega^a{}_b$ fixed by Cartan's first structure equation $\mathrm{d}e^a + \omega^a{}_b \wedge e^b = 0$ (Chapter T2). On spinors it acts through the spin representation: $\nabla_\mu = \partial_\mu + \frac{1}{4}\omega_\mu{}^{ab}\gamma_a\gamma_b$, i.e. Nash's $\Gamma_\mu = \frac{1}{4}\omega_\mu{}^{ab}\gamma_a\gamma_b$.

So $\not{D}_\rho = \gamma^a e_a{}^\mu \nabla_\mu$, and we write D^+_ρ for its chiral part.

Worked example 10.35: The Dirac operator of a conformally flat surface

Let M be two-dimensional with $g = e^{2\sigma(x)}(\mathrm{d}x_1^2 + \mathrm{d}x_2^2)$, and $\gamma^1 = \sigma_1$, $\gamma^2 = \sigma_2$ as in Example 10.3.

Frame and spin connection. $e^a = e^\sigma \mathrm{d}x^a$, $e_a = e^{-\sigma}\partial_a$. The only independent component is $\omega := \omega^1{}_2 = -\omega^2{}_1$, and we claim $\omega = \partial_2\sigma\,\mathrm{d}x_1 - \partial_1\sigma\,\mathrm{d}x_2$:

$$\begin{aligned}\mathrm{d}e^1 + \omega^1{}_2 \wedge e^2 &\overset{(1)}{=} \partial_2\sigma\, e^\sigma \mathrm{d}x_2 \wedge \mathrm{d}x_1 + (\partial_2\sigma\,\mathrm{d}x_1 - \partial_1\sigma\,\mathrm{d}x_2) \wedge e^\sigma \mathrm{d}x_2 \\ &\overset{(2)}{=} e^\sigma \partial_2\sigma\,(-1+1)\,\mathrm{d}x_1 \wedge \mathrm{d}x_2 = 0,\end{aligned}$$

$$\begin{aligned}\mathrm{d}e^2 + \omega^2{}_1 \wedge e^1 &\overset{(3)}{=} \partial_1\sigma\, e^\sigma \mathrm{d}x_1 \wedge \mathrm{d}x_2 - (\partial_2\sigma\,\mathrm{d}x_1 - \partial_1\sigma\,\mathrm{d}x_2) \wedge e^\sigma \mathrm{d}x_1 \\ &\overset{(4)}{=} e^\sigma \partial_1\sigma\,(1-1)\,\mathrm{d}x_1 \wedge \mathrm{d}x_2 = 0.\end{aligned}$$

(1) $\mathrm{d}(e^\sigma \mathrm{d}x_1) = e^\sigma \partial_2\sigma\,\mathrm{d}x_2 \wedge \mathrm{d}x_1$.
(2) $\mathrm{d}x_2 \wedge \mathrm{d}x_1 = -\mathrm{d}x_1 \wedge \mathrm{d}x_2$; $\mathrm{d}x_2 \wedge \mathrm{d}x_2 = 0$.
(3) $\mathrm{d}(e^\sigma \mathrm{d}x_2) = e^\sigma \partial_1\sigma\,\mathrm{d}x_1 \wedge \mathrm{d}x_2$ and $\omega^2{}_1 = -\omega$.
(4) $\mathrm{d}x_1 \wedge \mathrm{d}x_1 = 0$ and $-(-\partial_1\sigma)\mathrm{d}x_2 \wedge \mathrm{d}x_1 = -\partial_1\sigma\,\mathrm{d}x_1 \wedge \mathrm{d}x_2$.

The curvature is $R^1{}_2 = \mathrm{d}\omega = -(\partial_1^2 + \partial_2^2)\sigma\,\mathrm{d}x_1 \wedge \mathrm{d}x_2$; dividing by the area form $e^{2\sigma}\mathrm{d}x_1 \wedge \mathrm{d}x_2$ gives the Gaussian curvature $K = -e^{-2\sigma}\Delta\sigma$, the classical formula.

The Dirac operator. $\frac{1}{4}\omega_\mu{}^{ab}\gamma_a\gamma_b = \frac{1}{2}\omega_\mu\gamma_1\gamma_2 = \frac{i}{2}\omega_\mu\sigma_3$ (from $\sigma_1\sigma_2 = i\sigma_3$), so with $\omega_1 = \partial_2\sigma$, $\omega_2 = -\partial_1\sigma$,

$$\begin{aligned}\not{D}_g\psi &\overset{(1)}{=} e^{-\sigma}\Big[\sigma_a\partial_a\psi + \tfrac{i}{2}\big(\omega_1\sigma_1\sigma_3 + \omega_2\sigma_2\sigma_3\big)\psi\Big] \overset{(2)}{=} e^{-\sigma}\Big[\sigma_a\partial_a\psi + \tfrac{1}{2}\big(\partial_1\sigma\,\sigma_1 + \partial_2\sigma\,\sigma_2\big)\psi\Big] \\ &\overset{(3)}{=} e^{-3\sigma/2}\,\sigma_a\partial_a\big(e^{\sigma/2}\psi\big) = e^{-3\sigma/2}\not{D}_{\mathrm{flat}}\,e^{\sigma/2}\psi.\end{aligned}$$

(1) $\gamma^\mu = e^{-\sigma}\sigma_\mu$ and the spin connection term.
(2) $\sigma_1\sigma_3 = -i\sigma_2$ and $\sigma_2\sigma_3 = i\sigma_1$, so $\frac{i}{2}(\partial_2\sigma(-i\sigma_2) - \partial_1\sigma(i\sigma_1)) = \frac{1}{2}(\partial_2\sigma\,\sigma_2 + \partial_1\sigma\,\sigma_1)$.
(3) Leibniz: $\sigma_a\partial_a(e^{\sigma/2}\psi) = e^{\sigma/2}(\sigma_a\partial_a\psi + \frac{1}{2}\sigma_a\partial_a\sigma\,\psi)$.

So the Dirac operator of a conformally flat metric is the flat one, conjugated and rescaled. In particular zero modes are unchanged by σ (the index is a topological invariant), but the *spectrum* changes, and so does the regularised determinant.

Diffeomorphisms act covariantly [Nash (10.35)]. An orientation-preserving diffeomorphism $h \in \mathrm{Diff}^+(M)$ acts on metrics by pull-back, $h \cdot \rho = h^*\rho$, and (after lifting $\mathrm{d}h$ to the spin bundle, which is possible for h in the identity component) on spinor fields by pull-back. Since every ingredient of $\not{D}$ is constructed naturally from the metric,

$$\not{D}_{h\cdot\rho} = h^{-1}\,\not{D}_{\rho}\,h. \tag{10.165}$$

This is the analogue of $\not{D}_{A_g} = g^{-1}\not{D}_A g$: coordinate invariance says the operator at the transformed metric is conjugate to the original one. As before, the conjugation relates operators on *different* spaces (E^+ and E^- for the chiral operator), and the regularised determinant need not be invariant.

10.7.2 The space of metrics modulo diffeomorphisms

We want the determinant to descend from $\mathrm{Met}(M)$ to $\mathrm{Met}(M)/\mathrm{Diff}^+(M)$, the space of geometries. As in Chapter 6 (and Nash's Chapter 5), this quotient is singular where the action is not free: a metric with isometries (Killing vectors) is fixed by a subgroup of $\mathrm{Diff}^+(M)$. The cure is to restrict to the subgroup $\mathrm{Diff}_0(M)$ of diffeomorphisms in the identity component that fix a point $x_0 \in M$ and a frame at x_0. An isometry fixing a point and a frame there is the identity (it commutes with the exponential map, so it fixes every geodesic from x_0), so $\mathrm{Diff}_0(M)$ acts freely and $\mathrm{Met}(M)/\mathrm{Diff}_0(M)$ is well behaved.

Worked example 10.36: Metrics on a circle modulo diffeomorphisms

Every metric on $S^1 = [0,1]/(0 \sim 1)$ has the form $\rho = f(x)\,\mathrm{d}x^2$ with $f > 0$ periodic. Put $L = \int_0^1 \sqrt{f}\,\mathrm{d}x$ (the length) and $\phi(x) = \frac{1}{L}\int_0^x \sqrt{f(y)}\,\mathrm{d}y$. Then $\phi : S^1 \to S^1$ is a diffeomorphism fixing 0, and

$$\phi^*\big(L^2\mathrm{d}\phi^2\big) \overset{(1)}{=} L^2\phi'(x)^2\,\mathrm{d}x^2 \overset{(2)}{=} L^2\cdot\frac{f(x)}{L^2}\,\mathrm{d}x^2 = \rho.$$

(1) Pull-back of $\mathrm{d}\phi^2$ is $\phi'(x)^2\mathrm{d}x^2$.

(2) Fundamental theorem of calculus: $\phi'(x) = \sqrt{f(x)}/L$.

So every metric is equivalent to the constant metric of the same length: $\mathrm{Met}(S^1)/\mathrm{Diff}_0(S^1) \cong (0,\infty)$, one number. The rotations of the circle are isometries of $L^2\mathrm{d}\phi^2$; requiring the diffeomorphism to fix 0 removes them, which is Nash's reason for passing to Diff_0. (In two dimensions the same reasoning leaves the conformal factor and finitely many moduli: Nash's Chapter 9 and §4.)

The index bundle and its determinant [Nash (10.36)–(10.38)]. Exactly as for gauge fields,

$$\begin{aligned}\mathrm{Index}\,\not{D}_{\mathrm{grav}} &= \big\{\ker D^+_\rho : \rho \in \mathrm{Met}(M)\big\} - \big\{\ker (D^+_\rho)^\dagger : \rho \in \mathrm{Met}(M)\big\} \\ &= [\ker \not{D}_{\mathrm{grav}}] - [\ker \not{D}^*_{\mathrm{grav}}] \in K\big(\mathrm{Met}(M)/\mathrm{Diff}_0(M)\big), \qquad (10.166)\\ \det\mathrm{Index}\,\not{D}_{\mathrm{grav}} &= (\det\ker \not{D}_{\mathrm{grav}})^* \otimes (\det\ker \not{D}^*_{\mathrm{grav}}), \qquad (10.167)\end{aligned}$$

and the families index theorem, now with no gauge bundle ($\mathrm{ch}(\mathcal{E}) \to 1$) but with M fibred over the family, gives

$$\mathrm{ch}\big(\mathrm{Index}\,\not{D}_{\mathrm{grav}}\big) = \int_M \hat{A}. \tag{10.168}$$

Here $\hat{A}$ is the $\hat{A}$-genus of the *vertical tangent bundle* of the family $M \to Z \to \mathrm{Met}/\mathrm{Diff}_0$, a cohomology class on the total space Z. As always, $c_1(\det\mathrm{Index})$ is its degree-2 part.

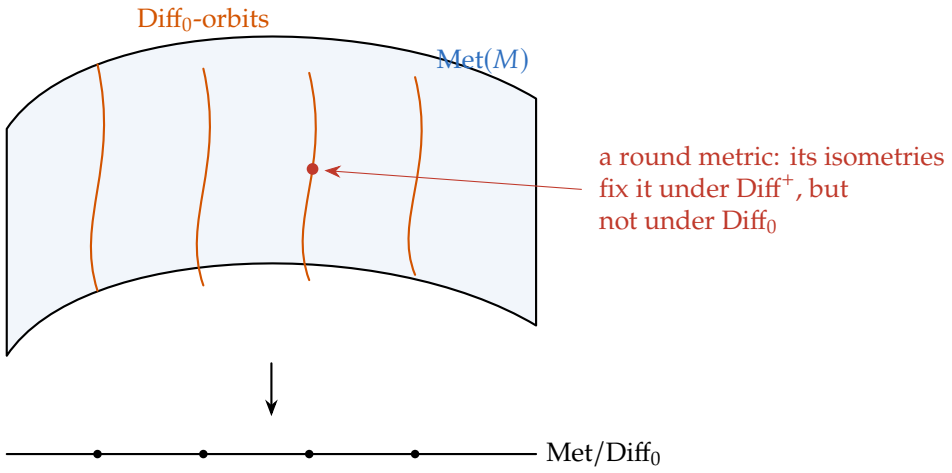


Figure 10.25: The gravitational analogue of Figure 10.3. Diffeomorphisms move a metric along an orbit (orange); the determinant of D_ρ^+ should be constant along each orbit if it is to descend to the space of geometries below. Metrics with isometries are fixed points of Diff$^+$; restricting to Diff$_0$ (fixing a point and a frame) makes the action free.

10.7.3 Why only $n = 4k - 2$

The $\hat{A}$-genus is a power series in the Pontrjagin classes, and p_j has degree $4j$. Let us see the first terms (Chapter 4). With formal Chern roots $\pm x_j$ of the tangent bundle, $\hat{A} = \prod_j \frac{x_j/2}{\sinh(x_j/2)}$ and $p_1 = \sum_j x_j^2$, $p_2 = \sum_{i<j} x_i^2 x_j^2$.

$$\frac{y}{\sinh y} \overset{(1)}{=} \left(1 + \frac{y^2}{6} + \frac{y^4}{120} + \cdots\right)^{-1} \overset{(2)}{=} 1 - \frac{y^2}{6} + \left(\frac{1}{36} - \frac{1}{120}\right)y^4 + \cdots = 1 - \frac{y^2}{6} + \frac{7y^4}{360} + \cdots,$$

$$\frac{x/2}{\sinh(x/2)} \overset{(3)}{=} 1 - \frac{x^2}{24} + \frac{7x^4}{5760} + \cdots,$$

$$\hat{A} \overset{(4)}{=} 1 - \frac{1}{24}\sum_j x_j^2 + \frac{7}{5760}\sum_j x_j^4 + \frac{1}{576}\sum_{i<j} x_i^2 x_j^2 + \cdots \overset{(5)}{=} 1 - \frac{p_1}{24} + \frac{7p_1^2 - 4p_2}{5760} + \cdots.$$

(1) Taylor series of $\sinh y / y$.

(2) $(1+u)^{-1} = 1 - u + u^2 - \cdots$ with $u = \frac{y^2}{6} + \frac{y^4}{120}$; the y^4 terms are $\frac{y^4}{36} - \frac{y^4}{120}$.

(3) $y = x/2$: $\frac{1}{6\cdot 4} = \frac{1}{24}$ and $\frac{7}{360\cdot 16} = \frac{7}{5760}$.

(4) Multiply the factors; the cross terms of degree 8 are $\frac{1}{24^2} x_i^2 x_j^2$.

(5) $\sum x_j^4 = p_1^2 - 2p_2$, so the degree-8 part is $\frac{7(p_1^2 - 2p_2) + 10p_2}{5760}$.

Only degrees $0, 4, 8, \ldots$ occur. Now, the integral over the n-dimensional fibre M lowers degree by n, so $\int_M \hat{A}$ has components only in degrees $4k - n$. For c_1 we need degree 2:

$$4k - n = 2 \iff n = 4k - 2 \qquad \text{[Nash (10.39)]}. \tag{10.169}$$

In $n = 4$ (our world) there is no purely gravitational perturbative anomaly for spin-$\frac{1}{2}$ fermions; in $n = 2, 6, 10$ there can be.

The cohomology of Diff$_0$**.** In the $H^1_{\mathrm{dR}}(\mathcal{G})$ picture, a gravitational anomaly is a nonzero class in $H^1_{\mathrm{dR}}(\mathrm{Diff}_0(M))$; for $M = S^{4k-2}$ Nash reports that $H^1_{\mathrm{dR}}(\mathrm{Diff}_0(S^{4k-2}))$ is known to vanish for $k = 1$ and $k > 6$ (Alvarez, Singer and Zumino) and is conjectured to vanish for all k. As

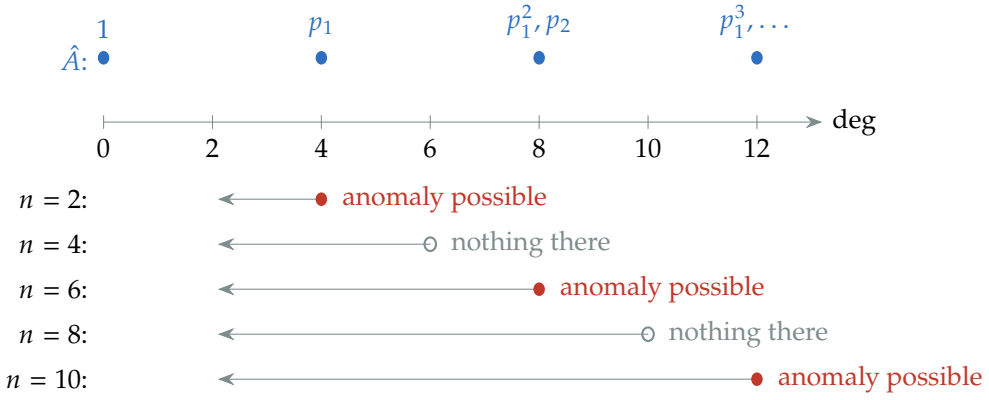


Figure 10.26: The degree count behind Nash's (10.39). A c_1 on the space of geometries needs the degree-$(n+2)$ part of $\hat{A}$ (red dot when it exists), which $\int_M$ lowers to degree 2 (grey arrow). Since $\hat{A}$ lives only in degrees divisible by 4, this happens exactly for $n = 2, 6, 10, \dots$

with U(1) in Section 10.6.6, this does not mean there is no anomaly: the relevant object is $H^1_{\mathrm{loc}}(\mathrm{Diff}_0(S^{4k-2}))$, which is not expected to vanish, and indeed the descent below produces nontrivial local anomalies in every dimension $4k-2$.

10.7.4 Anomaly polynomials and descent for gravity

The recipe of Section 10.5.5 carries over word for word, with the gauge field replaced by the spin connection ω (an $\mathfrak{so}(n)$-valued 1-form) and gauge transformations by local frame rotations $\theta(x) \in \mathfrak{so}(n)$: $\delta_\theta \omega = \mathrm{d}\theta + [\omega, \theta]$, $R = \mathrm{d}\omega + \omega^2$. The seed is the degree-$(n+2)$ part of the index density,

$$I_{n+2} = \left[\hat{A}(R)\ \mathrm{ch}(F)\right]_{n+2}, \tag{10.170}$$

which for a pure spin-$\frac{1}{2}$ field is $[\hat{A}]_{n+2}$. For $\mathfrak{so}(n)$ with real antisymmetric curvature matrices $p_1 = -\frac{1}{8\pi^2}\,\mathrm{tr}\,R^2$.

Worked example 10.37: The two-dimensional Lorentz anomaly by descent

For $n = 2$: $I_4 = -\frac{p_1}{24} = \frac{1}{192\pi^2}\,\mathrm{tr}\,R^2$. In two dimensions $\omega = w\,J$ with $J = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$ and $w = \omega^1{}_2$ a real 1-form, and a local Lorentz rotation is $\theta = \vartheta J$. Descent exactly as in Example 10.24, with $\mathrm{tr}(\cdot)$ now over the vector representation:

$$\omega_3^0 \overset{(1)}{=} \frac{1}{192\pi^2}\,\mathrm{tr}\left(\omega\,\mathrm{d}\omega + \tfrac{2}{3}\omega^3\right) \overset{(2)}{=} \frac{1}{192\pi^2}\,(-2)\,w\,\mathrm{d}w,$$

$$\omega_2^1 \overset{(3)}{=} \frac{1}{192\pi^2}\,\mathrm{tr}(\theta\,\mathrm{d}\omega) \overset{(4)}{=} -\frac{1}{96\pi^2}\,\vartheta\,\mathrm{d}w,$$

$$\delta_\vartheta \ln Z_F \overset{(5)}{=} -2\pi i \int_M \omega_2^1 \overset{(6)}{=} \frac{i}{48\pi}\int_M \vartheta\,\mathrm{d}w \overset{(7)}{=} \frac{i}{48\pi}\int_M \vartheta\,K\,\mathrm{d}A.$$

(1) Example 10.22 with $A \to \omega$, times the normalisation of I_4.
(2) $J^2 = -\mathbf{1}$ so $\mathrm{tr}\,J^2 = -2$; $\omega^3 = w \wedge w \wedge w\,J^3 = 0$.
(3) Example 10.24 with $v \to \theta$.
(4) $\mathrm{tr}(\vartheta J\,\mathrm{d}w\,J) = \vartheta\,\mathrm{d}w\,\mathrm{tr}\,J^2$.
(5) The normalisation (10.151).
(6) $-2\pi i \cdot (-\frac{1}{96\pi^2}) = \frac{i}{48\pi}$.
(7) $\mathrm{d}w = R^1{}_2 = K\,\mathrm{d}A$ with K the Gaussian curvature and $\mathrm{d}A$ the area form (Example 10.35).

A single Weyl fermion on a curved surface therefore has a Lorentz anomaly proportional to the curvature: rotating the frames by $\vartheta(x)$ changes the phase of the determinant by $\frac{1}{48\pi}\int \vartheta K \mathrm{d}A$. A Weyl fermion of the opposite chirality contributes the opposite sign, so a Dirac fermion is safe. In conformal-field-theory language the coefficient is proportional to $c_L - c_R$, the difference of the left and right central charges (Nash Ch. 11).

Other chiral fields (Alvarez-Gaumé and Witten). Chiral fermions are not the only chiral fields. A spin-$\frac{3}{2}$ field (gravitino) and an antisymmetric tensor field with self-dual field strength also have chiral determinants, and the same analysis gives their anomaly polynomials:

$$I_{1/2} = \big[\hat{A}(R)\big]_{n+2}, \qquad I_{3/2} = \big[\hat{A}(R)\big(\operatorname{tr} e^{iR/2\pi} - 1\big)\big]_{n+2}, \qquad I_A = -\tfrac{1}{8}\big[L(R)\big]_{n+2},$$

where $\operatorname{tr} e^{iR/2\pi}$ is over the vector representation (n-dimensional) and $L = 1 + \frac{p_1}{3} + \cdots$ is Hirzebruch's L-genus (whose integral is the signature). We quote these (they are derived from index theorems for the Rarita–Schwinger and signature operators) and check them in two dimensions, where $p_1 = x^2$ for the single Chern root x of TM:

$$\begin{aligned} I_{1/2} &\overset{(1)}{=} -\frac{x^2}{24}, \\ I_{3/2} &\overset{(2)}{=} \Big[\Big(1 - \frac{x^2}{24}\Big)\big(e^{x} + e^{-x} - 1\big)\Big]_4 \overset{(3)}{=} \Big[\Big(1 - \frac{x^2}{24}\Big)(1 + x^2)\Big]_4 = \frac{23}{24}x^2, \\ I_A &\overset{(4)}{=} -\frac{1}{8}\cdot\frac{x^2}{3} = -\frac{x^2}{24}. \end{aligned}$$

(1) The degree-4 part of $\hat{A}$.
(2) In the vector representation of $\mathfrak{so}(2)$ the matrix $\frac{i}{2\pi}R$ has eigenvalues $\pm x$ (this is what "Chern root" means), so $\operatorname{tr} e^{iR/2\pi} = e^{x} + e^{-x}$.
(3) $e^{x} + e^{-x} - 1 = 1 + x^2 + \cdots$.
(4) $L = 1 + \frac{p_1}{3}$.

Two lovely checks: $I_A = I_{1/2}$, i.e. a chiral boson has the same gravitational anomaly as a Weyl fermion, as bosonisation says it must; and $I_{3/2} = -23\, I_{1/2}$: a gravitino counts as -23 Weyl fermions. In ten dimensions Alvarez-Gaumé and Witten found that the chiral content of type IIB supergravity (a gravitino, a spin-$\frac{1}{2}$ field of the other chirality and a self-dual 4-form) has $I_{3/2} - I_{1/2} + I_A = 0$ at degree 12: an anomaly cancellation among fields of different spin. For type I and heterotic strings the degree-12 polynomial with gauge group SO(32) or $E_8 \times E_8$ factorises as $I_{12} \propto (\operatorname{tr} R^2 - \frac{1}{30}\operatorname{Tr} F^2)\, X_8$, which allows the Green–Schwarz counterterm to cancel it.

Mixed anomalies in four dimensions. In $n = 4$ the degree-6 part of $\hat{A}\operatorname{ch}(F)$ contains a mixed term,

$$\big[\hat{A}\ \operatorname{ch}(F)\big]_6 = \operatorname{ch}_3(F) + \operatorname{ch}_1(F)\Big(-\frac{p_1}{24}\Big) = \operatorname{ch}_3(F) + \frac{1}{192\pi^2}\,\frac{i}{2\pi}\operatorname{tr} F\ \operatorname{tr} R^2.$$

For non-abelian G, $\operatorname{tr} F = 0$. For each U(1) factor it is $\propto \sum q$: this is the "U(1)–graviton–graviton triangle" of Section 10.5.7, and the reason the Standard Model needs $\sum Y = 0$ [QFT notes (22.8.1)].

10.7.5 Lorentz anomalies and the automorphism group $\mathrm{Aut}(P)$

The Lorentz anomaly of Example 10.37 is a failure of invariance under frame rotations, not under diffeomorphisms. Nash's closing paragraph of §3 [Nash pp. 280–281] explains how the two fit together.

The larger symmetry group. Choosing a vielbein amounts to choosing a section of the frame bundle, a principal $\mathrm{SO}(n)$-bundle P (for spinors, its $\mathrm{Spin}(n)$ lift). The full symmetry of a theory written with vielbeins is the group $\mathrm{Aut}(P)$ of bundle automorphisms (Chapter 6): maps $P \to P$ commuting with the $\mathrm{SO}(n)$ action. Each covers a diffeomorphism of M, and those covering the identity are the vertical automorphisms, the local Lorentz transformations $\mathcal{G}_L = \mathrm{Map}(M, \mathrm{SO}(n))$. This is an exact sequence

$$1 \longrightarrow \mathcal{G}_L \longrightarrow \mathrm{Aut}(P) \longrightarrow \mathrm{Diff}_0(M) \longrightarrow 1,$$

and since the frame bundle of M has a canonical lift of diffeomorphisms (h acts on frames by $\mathrm{d}h$), it splits: $\mathrm{Aut}(P) = \mathcal{G}_L \rtimes \mathrm{Diff}_0(M)$, a semidirect product in which Diff_0 acts on $\mathcal{G}_L$ by transporting the local rotation, $(\lambda_1, h_1)(\lambda_2, h_2) = (\lambda_1 \cdot (h_1 \cdot \lambda_2),\ h_1 h_2)$ with $(h \cdot \lambda)(x) = \lambda(h^{-1}(x))$ (read in frames carried along by $\mathrm{d}h$). (For P Nash takes an $\mathrm{SO}(4k-2)$-bundle, since $n = 4k-2$.) The full anomaly is a class in $H^1(\mathrm{Aut}(P))$.

Trading one anomaly for the other. The polynomial $\operatorname{tr} R^2$ (more generally I_{n+2}) can be written with the spin connection ω, which transforms under $\mathcal{G}_L$ and is inert (a scalar) under diffeomorphisms, or with the Levi-Civita connection $\Gamma^\mu{}_\nu$ in a coordinate basis, a $\mathfrak{gl}(n)$-valued 1-form that transforms under diffeomorphisms and knows nothing of frames. The two are related by a $\mathrm{GL}(n)$ gauge transformation by the vielbein matrix $e = (e^a{}_\mu)$,

$$\Gamma = e^{-1}\omega\, e + e^{-1}\mathrm{d}e,$$

so their Chern–Simons forms differ by an exact form plus a *local* functional of e (Example 10.23 with $g \to e$). Descending with ω puts the anomaly entirely in $\mathcal{G}_L$ (a Lorentz anomaly, Einstein's diffeomorphism invariance intact); descending with Γ puts it entirely in Diff_0 (a gravitational anomaly, the stress tensor symmetric but not conserved). The local counterterm built from e moves it from one to the other (Alvarez-Gaumé and Ginsparg; Bardeen and Zumino). It can be moved, but not removed.

Key idea: Gravitational anomalies

A chiral field on M^n coupled to a metric has determinant $\det D_\rho^+$, a section of $\det \mathrm{Index}\, \not{D}_{\mathrm{grav}}$ over $\mathrm{Met}(M)/\mathrm{Diff}_0(M)$ with $\mathrm{ch} = \int_M \hat{A}$ (10.38). Because $\hat{A}$ has only degrees $4k$, c_1 can be nonzero only for $n = 4k-2$ (10.39). Locally the anomaly follows from descent of $I_{n+2} = [\hat{A}\,\mathrm{ch}\,F]_{n+2}$ with $A \to \omega$; it can be placed in local Lorentz transformations or in diffeomorphisms, but a nonzero I_{n+2} cannot be removed. Finally one may vary the Yang–Mills connection and the metric together; then the mixed terms of $\hat{A}\,\mathrm{ch}(F)$ (such as $\operatorname{tr} F\ \operatorname{tr} R^2$) must cancel as well.

Reading Nash (pp. 279–281, (10.35)–(10.39))

Nash's Dirac operator "$(1/2)\gamma_\mu(\partial^\mu + \Gamma^\mu)(1+\gamma_5)$" is our D_ρ^+ with $\Gamma_\mu = \frac{1}{4}\omega_\mu{}^{ab}\gamma_a\gamma_b$ (Example 10.35 computes it for a conformally flat surface). (10.35) is (10.165). His $\mathrm{Diff}_0(M)$ (identity component, fixing a frame at a point) is justified in Example 10.36 and Figure 10.25. (10.36)–(10.38) are (10.166)–(10.168); the degree count (10.39) is Figure 10.26. His remark that $H^1_{\mathrm{dR}}(\mathrm{Diff}_0(S^{4k-2}))$ vanishes for $k = 1$ and $k > 6$ but H^1_{local} is not expected to vanish is the gravitational version of Section 10.6.6; the explicit local anomaly in $n = 2$ is Example 10.37. The Lorentz anomaly, $\mathrm{Aut}(P) = \mathcal{G}_L \rtimes \mathrm{Diff}_0(M)$, and the trading between the two are Section 10.7.5.

Recap

- The Dirac operator depends on the metric through $\gamma^\mu = e_a{}^\mu \gamma^a$ and the spin connection; for $g = e^{2\sigma}\delta$ in 2D, $\not{D}_g = e^{-3\sigma/2}\not{D}_{\text{flat}}e^{\sigma/2}$. Diffeomorphisms act by conjugation, $\not{D}_{h\cdot\rho} = h^{-1}\not{D}_\rho h$.
- Metrics modulo Diff_0 (fixing a point and a frame, to kill isometries); on S^1 only the length survives.
- $\text{ch}(\det \text{Index}\, \not{D}_{\text{grav}}) = \int_M \hat{A}$ with $\hat{A} = 1 - \frac{p_1}{24} + \frac{7p_1^2 - 4p_2}{5760} - \cdots$; degree count gives $n = 4k - 2$.
- In 2D: $I_4 = \frac{1}{192\pi^2}\,\text{tr}\, R^2$, Lorentz anomaly $\frac{i}{48\pi}\int \vartheta K \mathrm{d}A$; chiral boson = Weyl fermion; gravitino = $-23\times$.
- In 4D no pure gravitational anomaly, but the mixed U(1)–gravity anomaly $\propto \sum q$.
- $\text{Aut}(P) = \mathcal{G}_L \rtimes \text{Diff}_0$: the anomaly can be shifted between Lorentz and diffeomorphisms by a local counterterm, not removed.

10.8 The critical dimension for strings

Why does the bosonic string want to live in exactly 26 dimensions? Of all the numbers in theoretical physics, this is one of the strangest. It is not put in by hand. Nobody chooses it for aesthetic reasons. It is *forced*, and the thing that forces it is an anomaly.

Here is the logic in one paragraph. The Polyakov action of a string is invariant under two groups acting on the worldsheet metric: coordinate changes $\text{Diff}^+(\Sigma)$ and local rescalings $g \mapsto e^{2\sigma}g$ (the **Weyl group** $C^\infty_+(\Sigma)$ of positive functions). To turn the path integral over all metrics into a finite-dimensional integral over the moduli space of Riemann surfaces (Chapter 9), we divide by the volumes of both groups. That division only makes sense if the integrand is constant along the orbits of both groups. The integrand is a product of two functional determinants: one from the d embedding coordinates X^μ, one from the Faddeev–Popov ghosts. Each determinant, separately, fails to be Weyl invariant. The failures are proportional to d and to -26 respectively. They cancel exactly when $d = 26$. That value of d is the **critical dimension**.

Nash does this in two pages with the families index theorem. We will do it three ways, because each teaches something different:

1. a direct heat-kernel computation of how a determinant changes under $g \mapsto e^{2\sigma}g$ (this subsection and the next): this is the physicist's *local* anomaly, $\langle T^a{}_a \rangle = -\frac{c}{12}R$;
2. Nash's computation: the first Chern class of a determinant line bundle, obtained from the families index theorem with Chern characters and Todd classes (Section 10.8.6);
3. two quick cross-checks physicists use every day: the central charge of the ghost system from an operator product, and the light-cone zero-point energy with $\zeta(-1) = -\frac{1}{12}$ (Section 10.8.8).

The three computations produce the same polynomial, $6j^2 - 6j + 1$, in three different guises. That coincidence is not a coincidence; it is the index theorem.

Reading Nash (p. 281, §4, (10.40))

Nash opens §4 with the sentence: "For the Bosonic string the critical dimension may be characterised as the only dimension for which the conformal anomaly vanishes. Instead of having a family of Dirac operators one has a family of Cauchy–Riemann operators." The Cauchy–Riemann operator is $\bar\partial = \partial/\partial\bar z$ acting on sections of a holomorphic line bundle over the worldsheet. The family is parametrised by the worldsheet metric. Everything below unpacks these two sentences.

10.8.1 The integrand we must keep invariant

We first recall the objects from Chapter 9, stating conventions as we go. The worldsheet Σ is a compact oriented surface of genus p without boundary, with a Riemannian metric g_{ab} and local coordinates x^a, $a = 1, 2$. The string's position in flat Euclidean space-time $\mathbb{R}^d$ is a map $\phi = (\phi^1, \dots, \phi^d) : \Sigma \to \mathbb{R}^d$. (Nash writes ϕ; physicists write X^μ.) The Polyakov action with unit tension is

$$S[\phi, g] = \frac{1}{2}\int_\Sigma \mathrm{d}^2x\, \sqrt{g}\, g^{ab}\, \partial_a\phi^\mu\, \partial_b\phi_\mu = \frac{1}{2}\int_\Sigma \mathrm{d}^2x \sqrt{g}\, \phi^\mu\, \Delta_g \phi_\mu, \tag{10.171}$$

where the scalar Laplacian is the *positive* operator

$$\Delta_g = -\frac{1}{\sqrt{g}}\,\partial_a\big(\sqrt{g}\, g^{ab}\partial_b\big). \tag{10.172}$$

The second form of (10.171) follows from the first by one integration by parts; there is no boundary term because $\partial\Sigma = \varnothing$.

The two symmetries. A diffeomorphism α moves both the metric and the fields, $(\phi, g) \mapsto (\phi \circ \alpha, \alpha^* g)$, and S is an integral of a scalar density, so it is unchanged. The more special symmetry is the Weyl rescaling. Let $f = e^{2\sigma}$ be a positive function and replace g_{ab} by $e^{2\sigma} g_{ab}$ without touching ϕ:

$$\begin{aligned}\sqrt{g}\, g^{ab} \;\mapsto\; \sqrt{\det(e^{2\sigma} g)}\, \big(e^{2\sigma} g\big)^{ab} &\overset{(1)}{=} \big(e^{2\sigma}\sqrt{g}\big)\big(e^{-2\sigma} g^{ab}\big) \\ &= \sqrt{g}\, g^{ab}.\end{aligned} \tag{10.173}$$

(1) In two dimensions $\det(e^{2\sigma} g) = e^{4\sigma} \det g$, so $\sqrt{\det}$ gains $e^{2\sigma}$; the inverse metric gains $e^{-2\sigma}$.

The combination that appears in S is untouched, so S is Weyl invariant. This works *only* in two dimensions: in D worldsheet dimensions $\sqrt{g}$ would gain $e^{D\sigma}$. Two consequences of (10.173) will be used repeatedly:

$$\Delta_{e^{2\sigma} g} = e^{-2\sigma}\Delta_g, \qquad \sqrt{g} \;\mapsto\; e^{2\sigma}\sqrt{g}. \tag{10.174}$$

The first holds because in (10.172) only the prefactor $1/\sqrt{g}$ feels the rescaling.

Doing the Gaussian integral over ϕ. For a positive symmetric $N \times N$ matrix M we have the familiar

$$\int_{\mathbb{R}^N} \mathrm{d}^N x\; e^{-\frac{1}{2}x^{\mathsf T} M x} = (2\pi)^{N/2} (\det M)^{-1/2}.$$

The field ϕ^μ is an infinite vector, expanded in the orthonormal eigenfunctions u_k of Δ_g, $\phi^\mu = \sum_k c_k^\mu u_k$, and $S = \frac{1}{2}\sum_k \lambda_k (c_k)^2$. There is one snag: $\lambda_0 = 0$ for the constant eigenfunction $u_0 = 1/\sqrt{A}$, where $A = \int_\Sigma \sqrt{g}\, \mathrm{d}^2x = \langle 1, 1\rangle_g$ is the area. The constant mode is the centre of

mass x_0^μ of the string, and it is not damped at all: its integral gives the volume of $\mathbb{R}^d$. Writing $c_0^\mu = \sqrt{A}\, x_0^\mu$ we find (dropping 2π's, which are metric independent)

$$\int \mathcal{D}\phi\; e^{-S[\phi,g]} = \mathrm{vol}(\mathbb{R}^d) \left(\frac{\det' \Delta_g}{A} \right)^{-d/2}, \tag{10.175}$$

where det′ means the product of the *non-zero* eigenvalues, defined with a ζ-function below. This is Nash's $\big(\det(\Delta_g/2)/\langle 1,1\rangle_g\big)^{-d/2}$ of (9.7), with $d = 26$ giving the exponent -13 in (9.35) and (10.40). The factor $1/2$ inside the determinant is harmless; see Section 10.8.3.

The ghost determinant. The division by the volume of $\mathrm{Diff}_0(\Sigma)$ is done with the Faddeev–Popov trick (Chapter 9, Nash (9.30)–(9.36)). The operator that appears is Nash's A_g (physicists call it P_1):

$$A_g : V \mapsto \mathcal{L}_V g - \tfrac{1}{2}\,\mathrm{tr}(\mathcal{L}_V g)\, g, \qquad (A_g V)_{ab} = \nabla_a V_b + \nabla_b V_a - g_{ab} \nabla_c V^c, \tag{10.176}$$

which sends a vector field to a symmetric *traceless* tensor. It is worth seeing what A_g is in complex coordinates, because this is how the Cauchy–Riemann operator enters. Choose a local complex coordinate $z = x^1 + i x^2$ with $g = e^{2\omega}\, \mathrm{d}z\, \mathrm{d}\bar{z}$, so $g_{z\bar{z}} = \frac{1}{2} e^{2\omega}$ and $g_{zz} = g_{\bar{z}\bar{z}} = 0$. A traceless symmetric tensor has only the components h_{zz} and $h_{\bar{z}\bar{z}}$ (the trace is $2 g^{z\bar{z}} h_{z\bar{z}}$). Then

$$\begin{aligned} (A_g V)_{\bar{z}\bar{z}} &\overset{(1)}{=} 2 \nabla_{\bar{z}} V_{\bar{z}} \\ &\overset{(2)}{=} 2\big(\partial_{\bar{z}} V_{\bar{z}} - \Gamma^{\bar{z}}_{\bar{z}\bar{z}} V_{\bar{z}} \big) \\ &\overset{(3)}{=} 2\big(\partial_{\bar{z}} (g_{\bar{z}z} V^z) - 2 \partial_{\bar{z}} \omega\; g_{\bar{z}z} V^z \big) \\ &\overset{(4)}{=} e^{2\omega}\, \partial_{\bar{z}} V^z . \end{aligned} \tag{10.177}$$

(1) Take the $\bar{z}\bar{z}$ component of (10.176); the trace term drops because $g_{\bar{z}\bar{z}} = 0$.
(2) Definition of the covariant derivative of a covector.
(3) Lower the index, $V_{\bar{z}} = g_{\bar{z}z} V^z$. For a conformal metric the only non-zero Christoffel symbols are $\Gamma^z_{zz} = \partial_z \ln g_{z\bar{z}} = 2\partial_z \omega$ and $\Gamma^{\bar{z}}_{\bar{z}\bar{z}} = 2\partial_{\bar{z}}\omega$.
(4) $g_{\bar{z}z} = \frac{1}{2} e^{2\omega}$, and $\partial_{\bar{z}}(e^{2\omega} V^z) = e^{2\omega}(\partial_{\bar{z}} V^z + 2\partial_{\bar{z}}\omega\, V^z)$: the second term cancels the Christoffel term exactly.

The component V^z is a section of the holomorphic tangent bundle, and the Christoffel symbol has dropped out: what is left is the bare Cauchy–Riemann derivative $\partial_{\bar{z}}$ (times a positive factor, which is where the metric enters the adjoint). So A_g is, up to conjugation and harmless factors, the Cauchy–Riemann operator $\bar{\partial}$ acting on sections of the holomorphic tangent bundle $T^{1,0}\Sigma$, which is the dual K_Σ^* of the canonical bundle $K_\Sigma = T^{*1,0}\Sigma$ (the line bundle of $(1,0)$-forms $f(z)\,\mathrm{d}z$). Nash records this as $\det(\tilde{A}_g^* \tilde{A}_g) = \big(\det \bar{\partial}^*_{K^*} \bar{\partial}_{K^*}\big)^2$, so the square root in the Faddeev–Popov factor gives exactly one power of $\det \bar{\partial}^*_{K^*} \bar{\partial}_{K^*}$.

Key idea: The integrand of the string partition function

Putting the matter and ghost factors together, the contribution of genus $p \geq 2$ to the string partition function is (Nash (10.40), with 26 replaced by a general d)

$$Z_p = \int_{\mathcal{M}_p} \mathrm{d}\mu(\mathcal{M}_p) \underbrace{\left(\frac{\det'(\Delta_g/2)}{\langle 1,1\rangle_g} \right)^{-d/2}}_{d \text{ free bosons}} \underbrace{\det\big(\bar{\partial}^*_{K^*} \bar{\partial}_{K^*}\big)}_{\text{ghosts}} . \tag{10.178}$$

The integral is over the finite-dimensional moduli space $\mathcal{M}_p$. For this to be legitimate, the integrand, which is computed from a metric g, must depend only on the *conformal class* of g, i.e. it must be unchanged under $g \mapsto e^{2\sigma} g$.

Reading Nash (p. 281, footnote 4)

Nash restricts to genus $p > 1$. For $p = 0$ (the sphere) and $p = 1$ (the torus) the Weyl and diffeomorphism groups do not act freely: there are conformal Killing vectors ($\ker A_g \neq 0$), six real ones on S^2 and two on T^2, and the quotient $\mathrm{Met}(\Sigma)/C^\infty_+(\Sigma)$ is singular. The anomaly computation below is local on Σ and does not care, but the clean statement "the determinant line bundle over the quotient" does.

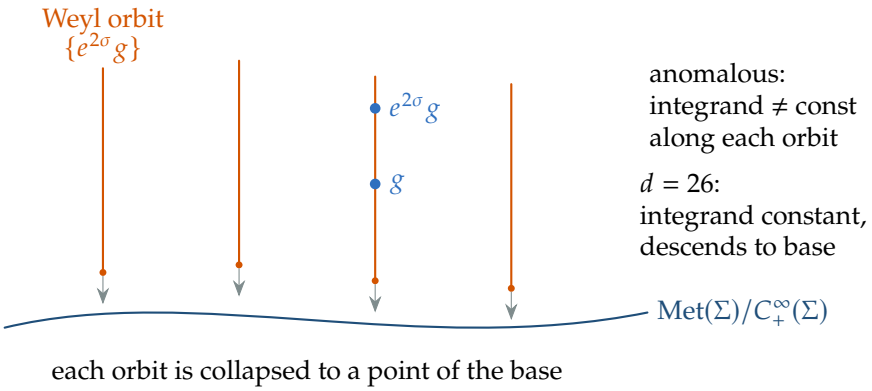


Figure 10.27: The space of metrics is fibred by Weyl orbits (orange). The string integrand is computed at a metric g in some fibre. Dividing by the volume of the Weyl group replaces each fibre by a point of the base; that is only meaningful if the integrand takes the same value all the way up the fibre. The conformal anomaly measures how much it changes.

10.8.2 Zeta-function determinants and the heat kernel

To see whether (10.178) is Weyl invariant we must know how a functional determinant responds to a change of metric. We first recall the definition (Nash p. 261 footnote, and Chapter 4).

Working definition 10.10 (ζ-regularised determinant)

Let P be a positive elliptic operator [here $\Delta_g/2$ or $\bar\partial^*_{K^*}\bar\partial_{K^*}$ from Z_p: built from the metric g, with eigenvalues growing without bound] with eigenvalues $0 < \lambda_1 \le \lambda_2 \le \cdots$ (zero modes discarded). Its **zeta function** is

$$\zeta_P(s) = \sum_k \lambda_k^{-s} = \frac{1}{\Gamma(s)} \int_0^\infty t^{s-1} \big(\mathrm{Tr}\, e^{-tP} - \dim \mathrm{Ker}\, P \big)\, \mathrm{d}t ,$$

convergent for $\mathrm{Re}\, s > \frac{\dim \Sigma}{\text{order}}$ and continued meromorphically [the function given by the sum where it converges extends to all s, with only isolated poles]; it is regular at $s = 0$. Since $\frac{\mathrm{d}}{\mathrm{d}s}\lambda^{-s}\big|_{s=0} = -\ln\lambda$, the formal $\ln \prod \lambda_k = \sum \ln \lambda_k$ equals $-\zeta_P'(0)$. So we *define*

$$\det{}' P := e^{-\zeta_P'(0)} .$$

How to compute it: know the spectrum (or the small-t behaviour of the heat trace [$\mathrm{Tr}\, e^{-tP} = \sum_k e^{-t\lambda_k}$]), build ζ_P, differentiate at 0.

The second equality in the definition is the Mellin transform $\lambda^{-s} = \frac{1}{\Gamma(s)}\int_0^\infty t^{s-1}e^{-t\lambda}\mathrm{d}t$, summed over eigenvalues. It makes the small-t behaviour of the heat trace $\operatorname{Tr} e^{-tP}$ the key input. We need a slightly more general object, the heat trace weighted by a function f on Σ:

$$\operatorname{Tr}\big(f\, e^{-t\Delta}\big) = \int_\Sigma \mathrm{d}^2x\sqrt{g}\; f(x)\, K_t(x,x),$$

where $K_t(x,y)$ is the heat kernel of Δ_g.

Key idea: Heat kernel on a surface, first two terms

For the scalar Laplacian on a surface, as $t \to 0^+$,

$$K_t(x,x) = \frac{1}{4\pi t}\Big(1 + \frac{t}{6}R(x) + O(t^2)\Big), \tag{10.179}$$

where R is the scalar curvature ($R = 2/r^2$ on a round sphere of radius r; $R = 2K$ with K the Gauss curvature).

Why these two terms? The leading term is the flat-space heat kernel $(4\pi t)^{-1}e^{-|x-y|^2/4t}$ at $x = y$: for small t the heat has spread only a distance $\sqrt{t}$, over which the surface looks flat. The correction is first order in the curvature and has dimension (length)$^{-2}$, so by dimensional analysis it must be $c\,tR$ for a pure number c. The number $c = \frac{1}{6}$ is the classic Minakshisundaram–Pleijel coefficient; we will *check* it on the round sphere in Example 10.38. Integrating (10.179) over Σ and using Gauss–Bonnet, $\int_\Sigma \sqrt{g}\,R = 4\pi\chi(\Sigma)$ (Chapter T2), the constant term of the heat trace is topological:

$$\operatorname{Tr} e^{-t\Delta} = \frac{A}{4\pi t} + \frac{\chi(\Sigma)}{6} + O(t). \tag{10.180}$$

Now the lemma that turns heat coefficients into ζ values.

Key idea: ζ at the origin reads off the t^0 heat coefficient

Let $\zeta(s;f) = \operatorname{Tr}'(f\Delta^{-s})$, the prime meaning that the zero mode is removed. Then $\zeta(s;f)$ is regular at $s = 0$ and

$$\zeta(0;f) = \frac{1}{24\pi}\int_\Sigma \sqrt{g}\,R\,f \;-\; \frac{1}{A}\int_\Sigma \sqrt{g}\,f. \tag{10.181}$$

Here is the calculation. Removing the zero mode $u_0 = 1/\sqrt{A}$ from the trace subtracts $\langle u_0, f u_0\rangle = \frac{1}{A}\int\sqrt{g}f =: P_f$. Split the Mellin integral at $t = 1$:

$$\begin{aligned}
\zeta(s;f) &\overset{(1)}{=} \frac{1}{\Gamma(s)}\int_0^1 t^{s-1}\Big(\frac{a_{-1}}{t} + a_0 - P_f + O(t)\Big)\mathrm{d}t + \frac{1}{\Gamma(s)}E(s)\\
&\overset{(2)}{=} \frac{1}{\Gamma(s)}\Big(\frac{a_{-1}}{s-1} + \frac{a_0 - P_f}{s} + (\text{regular at } 0)\Big) + \frac{E(s)}{\Gamma(s)}\\
&\overset{(3)}{=} \big(s + O(s^2)\big)\Big(\frac{a_0 - P_f}{s} + \text{regular}\Big)\\
&\xrightarrow{s\to 0} a_0 - P_f.
\end{aligned} \tag{10.182}$$

(1) Insert the small-t expansion $\operatorname{Tr}(fe^{-t\Delta}) = a_{-1}t^{-1} + a_0 + O(t)$ on $[0,1]$, with $a_{-1} = \frac{1}{4\pi}\int\sqrt{g}f$ and $a_0 = \frac{1}{24\pi}\int\sqrt{g}Rf$ from (10.179); $E(s)$ is the integral over $[1,\infty)$, which converges for every s because the non-zero spectrum is bounded away from 0 (exponential decay), so E is entire.

(2) $\int_0^1 t^{s-2}\mathrm{d}t = \frac{1}{s-1}$ and $\int_0^1 t^{s-1}\mathrm{d}t = \frac{1}{s}$.
(3) $1/\Gamma(s) = s + \gamma s^2 + \cdots$ has a simple zero at $s = 0$; it kills every regular term and $E(s)$ at $s = 0$.

This proves (10.181). With $f = 1$ it gives the useful special case

$$\zeta_\Delta(0) = \frac{\chi(\Sigma)}{6} - 1. \tag{10.183}$$

Worked example 10.38: $\zeta(0)$ on the round sphere: checking the coefficient $\frac{1}{6}$

On the unit sphere, Δ has eigenvalues $l(l+1)$ with multiplicity $2l+1$, $l = 0, 1, 2, \ldots$ (spherical harmonics Y_{lm}). The prediction (10.183) with $\chi(S^2) = 2$ is $\zeta(0) = \frac{2}{6} - 1 = -\frac{2}{3}$. Let us compute it directly. Write $\nu = l + \frac{1}{2}$, so $l(l+1) = \nu^2 - \frac{1}{4}$ and $2l + 1 = 2\nu$. Then

$$\begin{aligned}\zeta(s) = \sum_{l\geq 1} \frac{2l+1}{\big(l(l+1)\big)^s} &\overset{(1)}{=} \sum_{l\geq 1} 2\nu\, \nu^{-2s}\Big(1 - \frac{1}{4\nu^2}\Big)^{-s} \\ &\overset{(2)}{=} 2\sum_{l\geq 1} \nu^{1-2s} + \frac{s}{2}\sum_{l\geq 1} \nu^{-1-2s} + s\,G(s) \\ &\overset{(3)}{=} 2\,\zeta_H\big(2s-1, \tfrac{3}{2}\big) + \frac{s}{2}\,\zeta_H\big(2s+1, \tfrac{3}{2}\big) + s\,G(s).\end{aligned}$$

(1) Substitute $\nu = l + \frac{1}{2}$.
(2) Binomial expansion $(1-u)^{-s} = 1 + su + O(s\,u^2)$ with $u = 1/4\nu^2$; the remainder is s times a sum $G(s) = \sum O(\nu^{-3-2s})$ that converges near $s = 0$.
(3) Hurwitz zeta function $\zeta_H(s, a) = \sum_{n\geq 0}(n+a)^{-s}$; the sum over $l \geq 1$ starts at $\nu = \frac{3}{2}$.

Now use two standard facts: $\zeta_H(-1, a) = -\frac{1}{2}B_2(a) = -\frac{1}{2}(a^2 - a + \frac{1}{6})$, and $\zeta_H(1+2s, a) = \frac{1}{2s} + O(1)$. At $s = 0$:

$$\begin{aligned}\zeta(0) &\overset{(1)}{=} 2\cdot\Big(-\tfrac{1}{2}\Big)\Big(\tfrac{9}{4} - \tfrac{3}{2} + \tfrac{1}{6}\Big) + \frac{s}{2}\cdot\frac{1}{2s}\Big|_{s\to 0} \\ &= -\tfrac{11}{12} + \tfrac{1}{4} = -\tfrac{2}{3}.\end{aligned}$$

(1) The two facts just quoted with $a = \frac{3}{2}$; the $sG(s)$ term vanishes at $s = 0$.

The direct spectral computation agrees with the heat-kernel prediction. This checks the coefficient $\frac{1}{6}$ in (10.179): had it been c, we would have predicted $\zeta(0) = 2c - 1$.

10.8.3 The Weyl anomaly of d free bosons

We now vary the metric by an infinitesimal Weyl rescaling, $\delta g_{ab} = 2\delta\sigma\, g_{ab}$, and compute the change in the matter factor of (10.178). By (10.174), $\delta\Delta = -2\,\delta\sigma\,\Delta$.

$$\begin{aligned}\delta\zeta_\Delta(s) &\overset{(1)}{=} -s\ \mathrm{Tr}'\big(\delta\Delta\ \Delta^{-s-1}\big) \\ &\overset{(2)}{=} 2s\ \mathrm{Tr}'\big(\delta\sigma\ \Delta^{-s}\big) = 2s\,\zeta(s; \delta\sigma), \\ \delta\big(-\ln\det{}'\Delta\big) = \delta\zeta_\Delta'(0) &\overset{(3)}{=} 2\,\zeta(0; \delta\sigma) \\ &\overset{(4)}{=} \frac{1}{12\pi}\int_\Sigma \sqrt{g}\,R\,\delta\sigma - \frac{2}{A}\int_\Sigma \sqrt{g}\,\delta\sigma.\end{aligned} \tag{10.184}$$

(1) First-order perturbation theory for eigenvalues, $\delta(\lambda^{-s}) = -s\lambda^{-s-1}\delta\lambda$ with $\delta\lambda_k = \langle u_k, \delta\Delta\, u_k\rangle$; summing over k gives the trace.

(2) $\delta\Delta = -2\delta\sigma\Delta$ and cyclicity of the trace.
(3) Differentiate $2s\,\zeta(s;\delta\sigma)$ at $s=0$; since $\zeta(s;\delta\sigma)$ is regular there, only $2\zeta(0;\delta\sigma)$ survives.
(4) Formula (10.181) with $f=\delta\sigma$.

The area changes by $\delta A = \int \delta\sqrt{g} = 2\int\sqrt{g}\,\delta\sigma$, so $\delta\ln A = \frac{2}{A}\int\sqrt{g}\,\delta\sigma$, which exactly cancels the zero-mode term. Therefore

$$\boxed{\delta\ln\frac{\det'\Delta_g}{A} = -\frac{1}{12\pi}\int_\Sigma \sqrt{g}\,R\,\delta\sigma} \tag{10.185}$$

and for the matter factor of (10.178),

$$\delta\ln\left(\frac{\det'\Delta_g}{A}\right)^{-d/2} = +\frac{d}{24\pi}\int_\Sigma\sqrt{g}\,R\,\delta\sigma. \tag{10.186}$$

The right side is not zero unless $\delta\sigma$ is orthogonal to R. So the matter determinant is *not* Weyl invariant, even though the action is. This is the **conformal anomaly** (or Weyl anomaly, or trace anomaly).

Physical picture: Where the anomaly comes from

Classically nothing depends on σ. Quantum mechanically we must regularise the product of infinitely many eigenvalues, and every regulator brings in a length scale: here, the heat-kernel time t, which is a (length)^2 measured *with the metric* g. Rescaling g rescales the cutoff. The only memory the regulator keeps is the local term $R/6$ in the heat kernel: the anomaly is local, as every perturbative anomaly is.

Dictionary: Weyl anomaly ↔ trace anomaly ↔ central charge

Define the stress tensor by $T_{ab} = -\frac{4\pi}{\sqrt{g}}\frac{\delta S}{\delta g^{ab}}$ (Polchinski's normalisation). A Weyl rescaling $\delta g^{ab} = -2\delta\sigma g^{ab}$ changes $\ln Z$ by $-\langle\delta S\rangle = -\frac{1}{2\pi}\int\sqrt{g}\,\delta\sigma\,\langle T^a{}_a\rangle$. Comparing with (10.186):

$$\delta_\sigma \ln Z = \frac{c}{24\pi}\int\sqrt{g}\,R\,\delta\sigma \quad\Longleftrightarrow\quad \langle T^a{}_a\rangle = -\frac{c}{12}R,$$

with $c=d$ for d free bosons. The number c is the **central charge**: each free boson contributes 1. The classical tracelessness $T^a{}_a=0$ (the Noether statement of Weyl invariance) fails in the quantum theory by a c-number proportional to the curvature.

Watch out: The factor $1/2$ in $\det(\Delta_g/2)$

Nash writes $\det(\Delta_g/2)$ because (Nash (5.32)) $\Delta_g/2 = \bar\partial^*\bar\partial$ on functions. A constant factor c changes the ζ-determinant by $\det'(cP) = c^{\zeta_P(0)}\det' P$ (differentiate $c^{-s}\zeta_P(s)$ at 0). By (10.183), $\zeta_\Delta(0) = \chi/6 - 1$ is the same for every metric on Σ, so the factor is a constant: it can never contribute to the Weyl anomaly. It is a reminder, though, that ζ-determinants are *not* multiplicative: $\det(cP)\neq c^\infty\det P$.

Worked example 10.39: Constant rescaling: the anomaly is Gauss–Bonnet

Take $\delta\sigma=\sigma_0$ constant. The eigenvalues all scale by $e^{-2\sigma_0}$, so we can compute the change exactly, without (10.179). With $g'=e^{2\sigma_0}g$:

$$\ln\det{}'\Delta_{g'} - \ln\det{}'\Delta_g \overset{(1)}{=} \ln\det{}'\big(e^{-2\sigma_0}\Delta_g\big) - \ln\det{}'\Delta_g$$

$$\overset{(2)}{=} -2\sigma_0\,\zeta_\Delta(0)$$
$$\overset{(3)}{=} -2\sigma_0\Big(\frac{\chi}{6}-1\Big),$$
$$\ln A_{g'} - \ln A_g = 2\sigma_0,$$
$$\Rightarrow \quad \Delta \ln \frac{\det'\Delta}{A} = -\frac{\chi}{3}\,\sigma_0.$$

(1) (10.174) with constant σ.
(2) The scaling rule of Section 10.8.3 with $c = e^{-2\sigma_0}$.
(3) (10.183).

The general formula (10.185) gives $-\frac{1}{12\pi}\sigma_0\int\sqrt{g}R = -\frac{1}{12\pi}\sigma_0\cdot 4\pi\chi = -\frac{\chi}{3}\sigma_0$. They agree. On the sphere the partition function of a free boson depends on the *radius* $r = e^{\sigma_0}$ of the sphere: $\ln Z = -\frac{d}{2}\ln(\det'\Delta/A)$ changes by $+\frac{d}{2}\cdot\frac{\chi}{3}\sigma_0 = \frac{d}{3}\ln r$ (with $\chi = 2$), so $Z \propto r^{d/3}$. This is the famous statement that a CFT on a sphere knows its central charge through its size dependence, $Z \propto r^{c/3}$, with $c = d$ here.

10.8.4 The Liouville action: integrating the anomaly

The infinitesimal statement (10.185) can be integrated along the path $g_\tau = e^{2\tau\sigma}\hat{g}$, $\tau\in[0,1]$, to give the change under a *finite* rescaling. We need one formula from Riemannian geometry: for $g = e^{2\sigma}\hat{g}$ in two dimensions,

$$\sqrt{g}\,R = \sqrt{\hat{g}}\,\big(\hat{R} - 2\hat{\nabla}^2\sigma\big), \tag{10.187}$$

where $\hat{\nabla}^2 = -\hat{\Delta}$ is the ordinary (negative) Laplacian of $\hat{g}$.

Worked example 10.40: Checking (10.187): the sphere from the plane

Stereographic projection writes the unit sphere as the plane with $g = e^{2\sigma}(\mathrm{d}x^2+\mathrm{d}y^2)$, $e^{2\sigma} = 4/(1+r^2)^2$. Here $\hat{R}=0$ and $\sigma = \ln 2 - \ln(1+r^2)$. Then

$$\hat{\nabla}^2\ln(1+r^2) \overset{(1)}{=} \frac{1}{r}\frac{\mathrm{d}}{\mathrm{d}r}\Big(r\cdot\frac{2r}{1+r^2}\Big) = \frac{1}{r}\cdot\frac{4r}{(1+r^2)^2} = \frac{4}{(1+r^2)^2},$$
$$R \overset{(2)}{=} e^{-2\sigma}\big(0 - 2\hat{\nabla}^2\sigma\big) = \frac{(1+r^2)^2}{4}\cdot\frac{8}{(1+r^2)^2} = 2.$$

(1) Radial Laplacian $\frac{1}{r}\partial_r(r\partial_r)$.
(2) (10.187) divided by $\sqrt{g} = e^{2\sigma}\sqrt{\hat{g}}$.

$R = 2$ is correct for the unit sphere ($R = 2/r^2$, $r=1$). See Figure 10.28.

Now integrate (10.186) along g_τ, for which $\delta\sigma = \sigma\,\mathrm{d}\tau$ and $\sqrt{g_\tau}R_\tau = \sqrt{\hat{g}}(\hat{R} - 2\tau\hat{\nabla}^2\sigma)$:

$$\ln Z[e^{2\sigma}\hat{g}] - \ln Z[\hat{g}] \overset{(1)}{=} \frac{c}{24\pi}\int_0^1 \mathrm{d}\tau\int_\Sigma \sqrt{\hat{g}}\,\big(\hat{R} - 2\tau\hat{\nabla}^2\sigma\big)\sigma$$
$$\overset{(2)}{=} \frac{c}{24\pi}\int_\Sigma \sqrt{\hat{g}}\,\big(\hat{R}\sigma - \sigma\hat{\nabla}^2\sigma\big)$$
$$\overset{(3)}{=} \frac{c}{24\pi}\int_\Sigma \sqrt{\hat{g}}\,\big(\hat{g}^{ab}\partial_a\sigma\,\partial_b\sigma + \hat{R}\,\sigma\big). \tag{10.188}$$

(1) The infinitesimal anomaly $\frac{c}{24\pi}\int\sqrt{g}R\,\delta\sigma$ at each τ, with (10.187) for the metric $e^{2\tau\sigma}\hat{g}$.

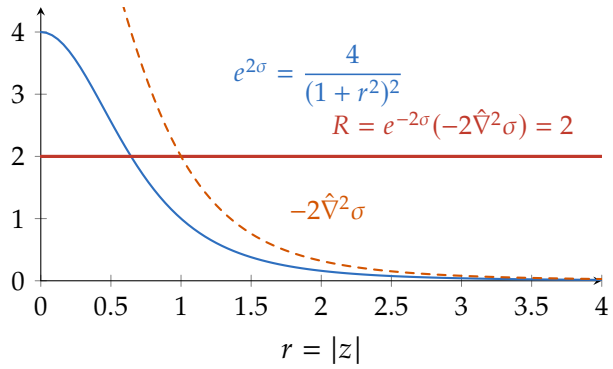


Figure 10.28: The round sphere as the flat plane with a conformal factor. The flat curvature is $\hat{R} = 0$; all the curvature comes from the Laplacian of σ (dashed). Dividing by $e^{2\sigma}$ (blue) gives the constant $R = 2$ (red). This is the mechanism by which the anomaly, which is linear in R, becomes a functional of σ.

(2) $\int_0^1 \mathrm{d}\tau = 1$ and $\int_0^1 2\tau\,\mathrm{d}\tau = 1$.
(3) Integrate $-\sigma\hat{\nabla}^2\sigma$ by parts on the closed surface.

The functional on the right is the **Liouville action**. If $c \neq 0$, the conformal factor σ does not drop out of the path integral: it becomes a dynamical field with this action (Polyakov's non-critical string). If $c = 0$ it decouples, and the integral over the Weyl orbit is just a volume we can divide out.

Recap: Weyl anomaly of the matter

- det′ is defined by ζ-functions; its Weyl variation is read off from the t^0 heat coefficient, $\frac{1}{24\pi}\int \sqrt{g}Rf$.
- d free bosons: $\delta \ln Z = \frac{d}{24\pi}\int \sqrt{g}R\,\delta\sigma$, i.e. central charge $c = d$, $\langle T^a{}_a \rangle = -\frac{c}{12}R$.
- Integrated form: the Liouville action (10.188).
- Checks: $\zeta(0) = -\frac{2}{3}$ on the round sphere; constant rescaling reproduces Gauss–Bonnet.
- Still to do: the ghost factor, which must supply $c = -26$.

10.8.5 Reminder: determinant line bundles and the families index

The heat-kernel computation of Section 10.8.3 is local and concrete. Nash's route is global and topological. It rests on two tools that were set up earlier in this chapter (Nash §2) and in Chapter 4. We recall both in the form we need.

The determinant line of a linear map. Let $O : V \to W$ be a linear map between finite-dimensional complex vector spaces. If $\dim V = \dim W = k$, the top exterior powers $\Lambda^k V$ and $\Lambda^k W$ are one-dimensional, and O induces a map between them, $v_1 \wedge \cdots \wedge v_k \mapsto Ov_1 \wedge \cdots \wedge Ov_k$. In bases, that map is multiplication by the number $\det O$. Without bases, $\det O$ is an element of the line $(\Lambda^k V)^* \otimes \Lambda^k W$. If the dimensions differ, or if O has a kernel, one keeps only what is left after cancelling the parts on which O is invertible:

$$\det O \;\in\; \mathcal{L}_O := \big(\Lambda^{\max} \operatorname{Ker} O\big)^* \otimes \Lambda^{\max} \operatorname{Coker} O. \tag{10.189}$$

(Nash (10.21)–(10.24); $\operatorname{Coker} O \cong \operatorname{Ker} O^*$.) When O is invertible both kernels vanish, $\Lambda^{\max}$ of the zero space is $\mathbb{C}$, and $\det O$ is an honest number.

The determinant line bundle. If O_y depends on a parameter $y \in Y$, the lines $\mathcal{L}_{O_y}$ fit together into a line bundle $\det \operatorname{Index} O_Y$ over Y, even though $\operatorname{Ker} O_y$ and $\operatorname{Coker} O_y$ can jump in dimension. The reason is that it depends only on the index bundle $\operatorname{Index} O_Y = [\operatorname{Ker} O] - [\operatorname{Coker} O] \in K(Y)$, and det turns sums into tensor products: $\det(E \oplus F) = \det E \otimes \det F$, so $\det([E] - [F]) = \det E \otimes (\det F)^*$. The function $y \mapsto \det O_y$ is a *section* of this bundle. It vanishes exactly where O_y has a kernel. If the bundle is non-trivial, no nowhere-zero section exists, and $\det O_y$ cannot be made into an ordinary function on Y.

Worked example 10.41: A family of 1×1 matrices over the sphere

Let $Y = \mathbb{CP}^1$ with homogeneous coordinates $[y_0 : y_1]$, and let O_y be "multiply by y_1/y_0" on the northern patch. As a function this blows up at $y_0 = 0$. But we may instead regard $O_y : \mathbb{C} \to L_y$ with $L = \mathcal{O}(1)$ the hyperplane bundle, $O_y(1) = y_1$ (a linear form in y, i.e. a section of $\mathcal{O}(1)$). Then $\det O_y = y_1$ is a perfectly good section of $\det \operatorname{Index} O = \mathcal{O}(1)$, with exactly one zero (at $y_1 = 0$), and

$$c_1(\det \operatorname{Index} O)[\mathbb{CP}^1] \overset{(1)}{=} \#\{\text{zeros of a generic section}\} \overset{(2)}{=} 1.$$

(1) The first Chern number of a line bundle over a closed surface counts the zeros (with multiplicity and sign) of a generic section; this is the line-bundle version of Poincaré–Hopf.
(2) y_1 vanishes once, simply.

So "the determinant is a section, not a function" is a statement with content: the zero cannot be removed by any choice of phase conventions, because its number is a Chern number.

The families index theorem for $\bar\partial$. For a family of Cauchy–Riemann operators $\bar\partial_E$ acting on a holomorphic vector bundle E over the fibres Σ of a family $Z \to Y$, the cohomological index formula (Nash (4.126), specialised to the Dolbeault complex; in algebraic geometry this is the Grothendieck–Riemann–Roch theorem) reads

$$\operatorname{ch}\big(\operatorname{Index} \bar\partial_E\big) = \int_\Sigma \operatorname{ch}(E)\ \operatorname{td}(T_{\mathrm{v}}), \tag{10.190}$$

where T_{v} is the bundle of tangent vectors along the fibres (the "vertical" tangent bundle) and $\int_\Sigma$ is integration over the fibre: it turns a form of degree k on Z into a form of degree $k-2$ on Y. Its degree-0 part is the ordinary Riemann–Roch theorem on one surface; its degree-2 part is

$$c_1\big(\det \operatorname{Index} \bar\partial_E\big) \overset{(1)}{=} c_1\big(\operatorname{Index} \bar\partial_E\big) \overset{(2)}{=} \Big[\int_\Sigma \operatorname{ch}(E)\operatorname{td}(T_{\mathrm{v}})\Big]_{(2)}. \tag{10.191}$$

(1) For any (virtual) bundle $c_1(\det E) = c_1(E)$: split E into line bundles L_i; then $\det E = \bigotimes L_i$ and both sides equal $\sum_i c_1(L_i)$.
(2) The degree-2 component of (10.190); the Chern character begins $\operatorname{ch} = \operatorname{rank} + c_1 + \cdots$.

So c_1 of the determinant bundle comes from the *degree*-4 part of the integrand $\operatorname{ch}(E)\operatorname{td}(T_{\mathrm{v}})$ on Z. On a single surface every 4-form vanishes; in a family it need not. This is the whole point.

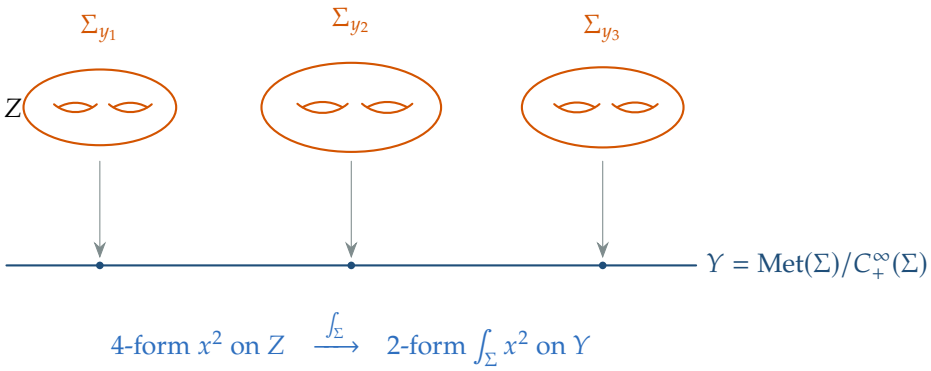


Figure 10.29: The universal family $Z \to Y$: over each point y of the base sits a surface Σ_y carrying the complex structure defined by y. Integrating a form over the fibres lowers its degree by two. The first Chern class of the determinant bundle lives on Y, so it comes from a *4-form* on Z, which is invisible on any single surface.

10.8.6 Nash's computation of 26

We now follow Nash's pp. 282–283 line by line, supplying each step.

Step 1: the scalar Laplacian is $2\bar\partial^*\bar\partial$. Nash quotes (5.32): $\Delta_g/2 = \bar\partial^*\bar\partial$ on functions. Let us check it for the flat metric $g = \mathrm{d}x^2 + \mathrm{d}y^2$ on a torus (the conformal factor goes through unchanged because both sides scale as $e^{-2\sigma}$). With $\partial_{\bar z} = \tfrac12(\partial_x + i\partial_y)$ we have $\bar\partial f = \partial_{\bar z} f\,\mathrm{d}\bar z$ and $|\mathrm{d}\bar z|^2 = |\mathrm{d}x|^2 + |\mathrm{d}y|^2 = 2$. For a complex function f,

$$\begin{aligned}
\langle f, \Delta f\rangle &\overset{(1)}{=} \int \left(|\partial_x f|^2 + |\partial_y f|^2\right)\mathrm{d}^2x \\
&\overset{(2)}{=} 2\int \left(|\partial_z f|^2 + |\partial_{\bar z} f|^2\right)\mathrm{d}^2x \\
&\overset{(3)}{=} 4\int |\partial_{\bar z} f|^2\,\mathrm{d}^2x \\
&\overset{(4)}{=} 2\,\langle \bar\partial f, \bar\partial f\rangle = \langle f, 2\bar\partial^*\bar\partial f\rangle .
\end{aligned}$$

(1) Integrate $\langle f, -\nabla^2 f\rangle$ by parts on the closed torus.
(2) Expand $\partial_x = \partial_z + \partial_{\bar z}$, $\partial_y = i(\partial_z - \partial_{\bar z})$; the cross terms cancel.
(3) $\int |\partial_z f|^2 = \int |\partial_{\bar z} f|^2$: integrate by parts twice, using $\partial_z\partial_{\bar z} = \partial_{\bar z}\partial_z$.
(4) $\|\partial_{\bar z} f\,\mathrm{d}\bar z\|^2 = \int 2|\partial_{\bar z} f|^2$.

So $\Delta = 2\bar\partial^*\bar\partial$, and the matter factor $\det(\Delta_g/2)^{-d/2} = \det(\bar\partial^*\bar\partial)^{-d/2}$ is built from a Cauchy–Riemann operator on the *trivial* bundle $I = \Sigma \times \mathbb{C}$, just as the ghost factor is built from one on K^*_Σ.

Step 2: a determinant is a section. For each metric g we have an operator $\bar\partial_g$. Nash says that "$\det(\bar\partial^*\bar\partial)$ can be thought of as a section of the determinant bundle $\det\mathrm{Index}\,\bar\partial$ over $\mathrm{Met}(\Sigma)/C^\infty_+(\Sigma)$". Strictly, the section is $\det\bar\partial$, and $\det(\bar\partial^*\bar\partial) = \|\det\bar\partial\|^2$ is its length squared in a metric on the line bundle (the Quillen metric of Section 10.8.9). The Weyl anomaly is the statement that this length is not constant along Weyl orbits; the topological statement is that the bundle on the *quotient* is not trivial.

Watch out: Why the base is Met/C^∞_+ and not Met

The space $\mathrm{Met}(\Sigma)$ of all metrics is convex, hence contractible, and every bundle over it is trivial. A trivial bundle has no Chern class. The anomaly is detected only after we ask the determinant to descend to the quotient by the symmetry we want to preserve. Exactly the same thing happened for gauge anomalies, where $\mathcal{A}$ is contractible but $\mathcal{A}/\mathcal{G}$ is not.

Step 3: raising to the power $-d/2$ is tensoring with $-(d/2)I$. If s is a section of a line bundle L, then s^m is a section of $L^{\otimes m}$, for positive or negative m (negative powers are sections of the dual). Now $L = \det \mathrm{Index}\,\bar\partial_I$, and det turns sums of index bundles into products, so

$$\left(\det \mathrm{Index}\,\bar\partial_I\right)^{\otimes m} = \det\left(m \cdot \mathrm{Index}\,\bar\partial_I\right) = \det \mathrm{Index}\,\bar\partial_{mI}. \tag{10.192}$$

For $m = -d/2$ the "bundle" mI has negative rank. It is not a vector bundle but an element of K-theory, a formal difference, and $\bar\partial$ with coefficients in it is Nash's $\bar\partial_{-(d/2)}$. Everything we need from it (its Chern character, its determinant line) is additive, so formal differences are harmless.

Watch out: Half-integer powers

If d is odd, $d/2$ is not an integer and $-(d/2)I$ is not even an element of $K(Y)$; it lives in $K(Y)\otimes\mathbb{Q}$. The first Chern class is still defined in rational cohomology, and the conclusion "$c_1 = 0$ iff $d = 26$" is unaffected. For d even the element is genuinely integral.

Step 4: one operator for the whole integrand. The ghost factor is $\det(\bar\partial^*\bar\partial)$ on K^*_Σ to the power $+1$, a section of $\det \mathrm{Index}\,\bar\partial_{K^*}$. Multiplying the two factors multiplies the line bundles, which by (10.192) adds the coefficient bundles. So the integrand of (10.178) is (the norm of) a section of the determinant bundle of a single family:

$$\bar\partial_Y := \bar\partial_{K^*_\Sigma - (d/2)I}, \qquad \det \mathrm{Index}\,\bar\partial_Y = \det \mathrm{Index}\,\bar\partial_{K^*} \otimes \left(\det \mathrm{Index}\,\bar\partial_I\right)^{-d/2}. \tag{10.193}$$

The integrand is Weyl invariant (can be made so by a local counterterm) exactly when this bundle is trivial on Y, which for a line bundle means $c_1 = 0$.

Step 5: the characteristic classes (Nash (10.41)–(10.42)). Let $x = c_1(K^*_\Sigma) = c_1(T_{\mathrm{v}})$ be the first Chern class of the vertical tangent line bundle on Z. For a line bundle with curvature F (anti-Hermitian, a $\mathfrak{u}(1) = i\mathbb{R}$-valued 2-form) the Chern–Weil representative is $x = \frac{i}{2\pi}F$ (Nash writes $\mathrm{tr}(iF/2\pi)$; the trace of a 1×1 matrix is the entry). The Todd class and Chern character of a line bundle are the power series

$$\begin{aligned}
\mathrm{td}(L) = \frac{x}{1-e^{-x}} &\overset{(1)}{=} \frac{x}{x - \frac{x^2}{2} + \frac{x^3}{6} - \cdots} = \left(1 - \frac{x}{2} + \frac{x^2}{6} - \cdots\right)^{-1} \\
&\overset{(2)}{=} 1 + \left(\frac{x}{2} - \frac{x^2}{6}\right) + \left(\frac{x}{2}\right)^2 + O(x^3) \\
&= 1 + \frac{x}{2} + \frac{x^2}{12} + O(x^3), \qquad\qquad (10.194)\\
\mathrm{ch}(L) = e^x &= 1 + x + \frac{x^2}{2} + O(x^3), \qquad \mathrm{ch}(mI) = m.
\end{aligned}$$

(1) Taylor-expand e^{-x} and divide numerator and denominator by x.

(2) $(1-u)^{-1} = 1 + u + u^2 + \cdots$ with $u = \frac{x}{2} - \frac{x^2}{6} + \cdots$, kept to second order: $u + u^2 = \frac{x}{2} - \frac{x^2}{6} + \frac{x^2}{4} + O(x^3)$.

For a surface, T_{v} restricted to each fibre is the line bundle K_Σ^* itself, so $\mathrm{td}(\Sigma) = \mathrm{td}(K_\Sigma^*)$; that is Nash's remark that the splitting principle is trivial here (there is only one Chern root). The Chern character is additive: $\mathrm{ch}(K^* - \frac{d}{2}I) = e^x - \frac{d}{2}$.

Step 6: multiply and integrate (Nash (10.43)).

$$\begin{aligned}\mathrm{ch}(\mathrm{Index}\,\bar\partial_Y) &\overset{(1)}{=} \int_\Sigma \Big(1 - \frac{d}{2} + x + \frac{x^2}{2} + \cdots\Big)\Big(1 + \frac{x}{2} + \frac{x^2}{12} + \cdots\Big) \\ &\overset{(2)}{=} \int_\Sigma \Big[\Big(1 - \frac{d}{2}\Big) + \Big(\frac{3}{2} - \frac{d}{4}\Big)x + \Big(\frac{1}{12}\Big(1 - \frac{d}{2}\Big) + \frac{1}{2} + \frac{1}{2}\Big)x^2 + \cdots\Big] \\ &\overset{(3)}{=} \Big(\frac{3}{2} - \frac{d}{4}\Big)\int_\Sigma x \;+\; \frac{26-d}{24}\int_\Sigma x^2 + \cdots \end{aligned} \tag{10.195}$$

(1) The families index theorem (10.190) with $E = K^* - \frac{d}{2}I$ and (10.194).

(2) Collect by degree. Degree 2: $x \cdot 1 + (1 - \frac{d}{2}) \cdot \frac{x}{2}$. Degree 4: $(1 - \frac{d}{2})\frac{x^2}{12} + x \cdot \frac{x}{2} + \frac{x^2}{2} \cdot 1$.

(3) The degree-0 term is a function on Z and integrates to zero over a 2-dimensional fibre. For the x^2 coefficient: $\frac{1}{12} - \frac{d}{24} + 1 = \frac{2-d+24}{24}$.

The first term is a number (a 0-form on Y): the virtual rank of the index bundle. The second is a 2-form on Y: the first Chern class. Hence

Key idea: Nash's (10.43): the critical dimension

$$c_1\big(\det \mathrm{Index}\,\bar\partial_Y\big) = \frac{26-d}{24}\int_\Sigma x^2 = \frac{26-d}{24}\,\kappa_1, \tag{10.196}$$

where $\kappa_1 := \int_\Sigma c_1(T_{\mathrm{v}})^2$ is a fixed, non-zero class on moduli space (the first Mumford–Morita–Miller class). The determinant bundle is trivial, and the conformal anomaly absent, precisely when $d = 26$.

Worked example 10.42: The degree-0 term is Riemann–Roch, and it counts moduli

The first term of (10.195) should be the virtual rank dim Ker − dim Coker of the family on one surface. Check it. Gauss–Bonnet says $\int_\Sigma x = \int_\Sigma c_1(T\Sigma) = \chi(\Sigma) = 2 - 2p$. So

$$\begin{aligned}\Big(\frac{3}{2} - \frac{d}{4}\Big)(2-2p) &\overset{(1)}{=} (3-3p) - \frac{d}{2}(1-p) \\ &\overset{(2)}{=} \mathrm{index}\,\bar\partial_{K^*} - \frac{d}{2}\,\mathrm{index}\,\bar\partial_I.\end{aligned}$$

(1) Arithmetic.

(2) Riemann–Roch for a line bundle L of degree $\deg L$ on a genus-p surface: $\mathrm{index}\,\bar\partial_L = h^0(L) - h^1(L) = \deg L + 1 - p$. For $L = I$, deg = 0, index $1 - p$; for $L = K^*$, deg $= 2 - 2p$, index $3 - 3p$.

Now interpret. On functions: $h^0(I) = 1$ (constants: the zero mode of Δ, i.e. the centre of mass that we treated separately in (10.175)) and $h^1(I) = p$ (the p holomorphic 1-forms). On vector fields, for $p \geq 2$: $h^0(K^*) = 0$ (a holomorphic vector field would be a conformal Killing vector, and there are none) and $h^1(K^*) = 3p - 3$. The cokernel of the ghost operator A_g is the space of *quadratic differentials*, the deformations of the complex structure not generated by diffeomorphisms or Weyl rescalings. So Riemann–Roch reproduces

$$\dim_{\mathbb{C}} \mathcal{M}_p = 3p - 3 \qquad (p \geq 2),$$

the complex dimension of the moduli space of Riemann surfaces. For $p = 2$: 3 complex

moduli; for $p = 3$: 6. This is why the ghost determinant and the measure $\mathrm{d}\mu(\mathcal{M}_p)$ in (10.178) come together: the zero modes of A_g^* are exactly the tangent directions of $\mathcal{M}_p$.

Worked example 10.43: x on the round sphere: Gauss–Bonnet as a Chern number

On the unit sphere with complex coordinate z (stereographic), $g = \frac{4\,\mathrm{d}z\,\mathrm{d}\bar z}{(1+|z|^2)^2}$, the holomorphic tangent bundle has fibre metric $h = g_{z\bar z} = \frac{2}{(1+|z|^2)^2}$ on the section ∂_z. The Chern connection of a holomorphic line bundle with metric h has curvature $F = \bar\partial\partial \ln h$. Then

$$\begin{aligned}\int_{S^2} x = \frac{i}{2\pi}\int_{S^2}\bar\partial\partial\ln h &\overset{(1)}{=} \frac{i}{2\pi}\int \partial_{\bar z}\partial_z\big(-2\ln(1+|z|^2)\big)\,\mathrm{d}\bar z\wedge\mathrm{d}z\\ &\overset{(2)}{=} \frac{i}{2\pi}\int\frac{-2}{(1+|z|^2)^2}\,\mathrm{d}\bar z\wedge\mathrm{d}z\\ &\overset{(3)}{=} \frac{i}{2\pi}\cdot(-2)\cdot 2i\int\frac{\mathrm{d}x\,\mathrm{d}y}{(1+r^2)^2}\\ &\overset{(4)}{=} \frac{2}{\pi}\cdot\pi = 2 = \chi(S^2).\end{aligned}$$

(1) The constant $\ln 2$ in $\ln h$ is killed by the derivatives.
(2) $\partial_z \ln(1+z\bar z) = \bar z/(1+z\bar z)$ and $\partial_{\bar z}\big(\bar z/(1+z\bar z)\big) = 1/(1+z\bar z)^2$.
(3) $\mathrm{d}\bar z\wedge\mathrm{d}z = 2i\,\mathrm{d}x\wedge\mathrm{d}y$.
(4) $\int_{\mathbb{R}^2}\frac{\mathrm{d}x\mathrm{d}y}{(1+r^2)^2} = 2\pi\int_0^\infty\frac{r\,\mathrm{d}r}{(1+r^2)^2} = \pi$.

So the class x really is the Euler class of the surface, and $\int_\Sigma x = \chi$. On one surface $x^2 = 0$; in a family, $\int_\Sigma x^2$ measures how the curvature of the fibres twists as we move in the base.

Reading Nash (pp. 282–283, (10.41)–(10.43))

Nash's (10.43) is written "ch(det Index $\bar\partial_Y$) = $\int_\Sigma(\cdots)(\cdots)$". Read the left side as ch(Index $\bar\partial_Y$); its degree-2 part is $c_1(\text{Index}) = c_1(\det\text{Index})$ by (10.191). The printed second factor "$(1+\frac{x}{2}+\frac{x^2}{12})$" is the Todd class and the first is ch$(K^*-\frac{d}{2}I)$, as in (10.42). His conclusion "the determinant bundle vanishes precisely when $d = 26$" means "is trivial" (has vanishing first Chern class), not that the bundle is zero.

10.8.7 The same polynomial for every weight: Mumford's formula

The computation generalises at once, and the generalisation explains the numbers. Let j be any integer (or half-integer, once a spin structure is chosen) and consider $\bar\partial$ on K^j, the bundle of **j-differentials** $f(z)\,(\mathrm{d}z)^j$. Physicists call a field with values in K^j a field of conformal weight j. Since $c_1(K^j) = -jx$,

$$\begin{aligned}c_1\big(\det\text{Index}\,\bar\partial_{K^j}\big) &\overset{(1)}{=} \Big[\int_\Sigma\Big(1-jx+\frac{j^2x^2}{2}\Big)\Big(1+\frac{x}{2}+\frac{x^2}{12}\Big)\Big]_{(2)}\\ &\overset{(2)}{=} \Big(\frac{j^2}{2}-\frac{j}{2}+\frac{1}{12}\Big)\int_\Sigma x^2 = \frac{6j^2-6j+1}{12}\,\kappa_1.\end{aligned}\tag{10.197}$$

(1) (10.191) with ch$(K^j) = e^{-jx}$.
(2) Degree-4 terms: $\frac{j^2x^2}{2} + (-jx)\frac{x}{2} + \frac{x^2}{12}$.

This is **Mumford's formula**. It contains every number of the bosonic and supersymmetric string.

field	weight j	statistics	power of $\det\bar\partial_{K^j}$	$\frac{6j^2-6j+1}{12}$	contribution to c_1/κ_1
X^μ, d real bosons	0	boson	$-d/2$	$\frac{1}{12}$	$-\frac{d}{24}$
b,c ghosts	$2,-1$	fermion	$+1$	$\frac{13}{12}$	$+\frac{26}{24}$
ψ^μ, d Majorana–Weyl	$\frac{1}{2}$	fermion	$+d/2$	$-\frac{1}{24}$	$-\frac{d}{48}$
β,γ superghosts	$\frac{3}{2},-\frac{1}{2}$	boson	-1	$\frac{11}{24}$	$-\frac{11}{24}$

Table 10.3: Mumford's coefficient for the fields of string theory. A complex fermionic pair of weights $(j,1-j)$ contributes one power of $\det\bar\partial_{K^{1-j}}$ (equivalently of K^j, by Serre duality); a real field contributes half a power; bosons contribute negative powers. The last column is always $-c/24$, where c is the central charge of the field.

Reading the table. *Bosonic string:* $-\frac{d}{24}+\frac{26}{24}=\frac{26-d}{24}$, the result (10.196) again. *Superstring:* add the worldsheet fermions and superghosts:

$$-\frac{d}{24}+\frac{26}{24}-\frac{d}{48}-\frac{11}{24}=\frac{-2d+52-d-22}{48}=\frac{30-3d}{48},$$

which vanishes at $d=10$. The critical dimension of the superstring comes out of the same one-line formula.

The symmetry $j\leftrightarrow 1-j$. The polynomial satisfies $6(1-j)^2-6(1-j)+1=6j^2-6j+1$. This is Serre duality: $H^1(K^j)\cong H^0(K^{1-j})^*$, so $\bar\partial_{K^j}$ and $\bar\partial_{K^{1-j}}$ have opposite index bundles, and a pair of fields of weights j and $1-j$ (like b and c) is really one system. The polynomial is smallest at $j=\frac{1}{2}$, the self-dual point, where the field is a spinor.

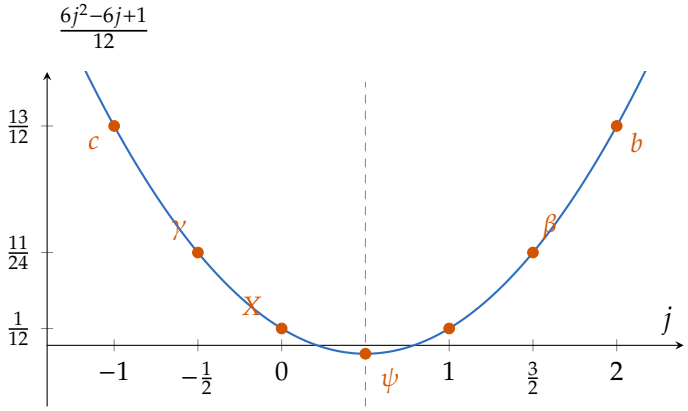


Figure 10.30: Mumford's polynomial is a parabola symmetric about the spinor point $j=\frac{1}{2}$ (dashed): Serre duality pairs weight j with $1-j$. The ghosts b,c sit high up at $j=2,-1$; the embedding coordinates at $j=0$. One ghost pair outweighs 26 half-units of X, which is where 26 comes from.

The local version. The Weyl-variation computation of Section 10.8.3 can be repeated for the Laplacian $\bar\partial^*_{K^j}\bar\partial_{K^j}$ on j-differentials, now with the twisted heat kernel (whose t^0 coefficient involves both $R/6$ and the curvature $-jR/2$ of K^j). The result (O. Alvarez 1983) is the local counterpart of (10.197):

$$\delta_\sigma \ln\frac{\det'\big(\bar\partial^*\bar\partial\big)_{K^j}}{\det\langle\text{zero modes}\rangle}=-\frac{6j^2-6j+1}{12\pi}\int_\Sigma\sqrt{g}\,R\,\delta\sigma. \tag{10.198}$$

For $j = 0$ this is exactly (10.185): the operator $\bar\partial^*\bar\partial = \Delta/2$ on complex functions has the same eigenvalues, with the same (complex) multiplicities, as $\Delta/2$ on real functions, and the zero-mode factor is the area. Assembling the partition function,

$$\begin{aligned}\delta_\sigma \ln Z_p &\overset{(1)}{=} \frac{d}{24\pi}\int \sqrt{g}R\,\delta\sigma - \frac{13}{12\pi}\int \sqrt{g}R\,\delta\sigma \\ &= \frac{d-26}{24\pi}\int_\Sigma \sqrt{g}\,R\,\delta\sigma .\end{aligned}$$

(1) Matter from (10.186); ghosts from (10.198) with $j = -1$ ($6 + 6 + 1 = 13$), with power $+1$ because the ghosts are fermions.

In the language of Section 10.8.3: the total central charge is $c_{\text{tot}} = d - 26$, and the Liouville field decouples iff $d = 26$.

10.8.8 Two cross-checks physicists use every day

(a) The ghost central charge from an operator product

In conformal field theory (Chapter 11) the central charge appears in the operator product of the stress tensor with itself,

$$T(z)T(w) \sim \frac{c/2}{(z-w)^4} + \frac{2T(w)}{(z-w)^2} + \frac{\partial T(w)}{z-w}. \tag{10.199}$$

For a fermionic pair b (weight λ) and c (weight $1-\lambda$) with $b(z)c(w) \sim \frac{1}{z-w}$, the stress tensor is

$$T = -\lambda\,{:}b\,\partial c{:} + (1-\lambda)\,{:}\partial b\,c{:}.$$

We compute the fourth-order pole of $T(z)T(w)$ directly.

Recall (Wick's theorem for operator products). The singular part of a product of two normal-ordered free-field expressions is the sum over all ways of contracting fields of the first with fields of the second, each contraction replaced by its two-point function and each fermion interchange giving a sign. A *double* contraction of two bilinears leaves no operator, and it is the only source of the highest pole.

The contractions we need. From $b(z)c(w) \sim \frac{1}{z-w}$ and the fermionic sign, $c(z)b(w) = -b(w)c(z) \sim -\frac{1}{w-z} = \frac{1}{z-w}$. Differentiating,

$$\begin{aligned}
\langle b(z)\,\partial c(w)\rangle &= \partial_w \tfrac{1}{z-w} = \tfrac{1}{(z-w)^2}, & \langle \partial b(z)\,c(w)\rangle &= \partial_z \tfrac{1}{z-w} = -\tfrac{1}{(z-w)^2},\\
\langle \partial c(z)\,b(w)\rangle &= -\tfrac{1}{(z-w)^2}, & \langle c(z)\,\partial b(w)\rangle &= \tfrac{1}{(z-w)^2},\\
\langle \partial b(z)\,\partial c(w)\rangle &= \partial_z \tfrac{1}{(z-w)^2} = -\tfrac{2}{(z-w)^3}, & \langle \partial c(z)\,\partial b(w)\rangle &= -\tfrac{2}{(z-w)^3}.
\end{aligned}$$

The double contraction of two bilinears. In $\beta_1\gamma_1(z)\,\beta_2\gamma_2(w)$ (with β's b-type and γ's c-type, all fermionic) contract γ_1 with β_2 and β_1 with γ_2. Bringing γ_2 next to β_1 moves it past the pair $\gamma_1\beta_2$, which is even, so there is no sign: the double contraction is $\langle\gamma_1(z)\beta_2(w)\rangle\langle\beta_1(z)\gamma_2(w)\rangle$. Now

expand $T(z)T(w)$ into its four products and doubly contract each:

$$\lambda^2 :b\partial c:(z)\,:b\partial c:(w) \overset{(1)}{=} \lambda^2\,\langle \partial c(z) b(w)\rangle\langle b(z)\partial c(w)\rangle = -\frac{\lambda^2}{(z-w)^4},$$

$$(1-\lambda)^2 :\partial b\, c:(z)\,:\partial b\, c:(w) \overset{(2)}{=} (1-\lambda)^2\langle c(z)\partial b(w)\rangle\langle \partial b(z) c(w)\rangle = -\frac{(1-\lambda)^2}{(z-w)^4},$$

$$-\lambda(1-\lambda) :b\partial c:(z)\,:\partial b\, c:(w) \overset{(3)}{=} -\lambda(1-\lambda)\langle \partial c(z)\partial b(w)\rangle\langle b(z) c(w)\rangle = \frac{2\lambda(1-\lambda)}{(z-w)^4},$$

$$-\lambda(1-\lambda) :\partial b\, c:(z)\,:b\partial c:(w) \overset{(4)}{=} -\lambda(1-\lambda)\langle c(z) b(w)\rangle\langle \partial b(z)\partial c(w)\rangle = \frac{2\lambda(1-\lambda)}{(z-w)^4}.$$

(1) The rule above with $\beta_1 = \beta_2 = b$, $\gamma_1 = \gamma_2 = \partial c$; the two contractions are $-\frac{1}{(z-w)^2}$ and $\frac{1}{(z-w)^2}$.
(2) $\beta = \partial b$, $\gamma = c$; contractions $\frac{1}{(z-w)^2}$ and $-\frac{1}{(z-w)^2}$.
(3) Contractions $-\frac{2}{(z-w)^3}$ and $\frac{1}{z-w}$; the coefficient of this cross term in $T\,T$ is $(-\lambda)(1-\lambda)$.
(4) Contractions $\frac{1}{z-w}$ and $-\frac{2}{(z-w)^3}$.

Adding the four lines and comparing with (10.199):

$$\frac{c_\lambda}{2} \overset{(1)}{=} -\lambda^2 - (1-\lambda)^2 + 4\lambda(1-\lambda) \overset{(2)}{=} -\lambda^2 - 1 + 2\lambda - \lambda^2 + 4\lambda - 4\lambda^2 = -(6\lambda^2 - 6\lambda + 1),$$
$$c_\lambda = -2\big(6\lambda^2 - 6\lambda + 1\big).$$

(1) The coefficient of $(z-w)^{-4}$ in the sum of the four lines.
(2) Expand $(1-\lambda)^2$ and $4\lambda(1-\lambda)$.

Check: $\lambda = \frac{1}{2}$ is a single complex (Dirac) fermion, $c = -2(\frac{3}{2} - 3 + 1) = 1$, as it should be. For the reparametrisation ghosts $\lambda = 2$: $c = -2(24 - 12 + 1) = -26$. For the superghosts (bosonic, so every contraction loop has the opposite sign and the overall sign flips) $\lambda = \frac{3}{2}$: $c = +2(\frac{27}{2} - 9 + 1) = 11$. And the polynomial $6\lambda^2 - 6\lambda + 1$ is Mumford's: $c = -24\times$ (coefficient in Table 10.3). The operator product and the Grothendieck–Riemann–Roch integral compute the same number.

(b) The light-cone zero-point energy

In light-cone gauge only the $d-2$ transverse oscillators survive. Each is a string of harmonic oscillators with frequencies $n = 1, 2, 3, \ldots$ (in units where the string length is fixed), each with zero-point energy $\frac{n}{2}$. The ground-state energy of the open string is therefore

$$a = \frac{d-2}{2}\sum_{n=1}^{\infty} n.$$

The sum diverges, but its divergence is a cutoff-dependent term that is absorbed in the (bare) string tension; the finite remainder is universal. Regulate with $e^{-\epsilon n}$:

$$\begin{aligned}\sum_{n=1}^{\infty} n\, e^{-\epsilon n} &\overset{(1)}{=} -\frac{\mathrm{d}}{\mathrm{d}\epsilon}\,\frac{1}{e^{\epsilon}-1} \\ &\overset{(2)}{=} -\frac{\mathrm{d}}{\mathrm{d}\epsilon}\Big(\frac{1}{\epsilon} - \frac{1}{2} + \frac{\epsilon}{12} + O(\epsilon^3)\Big) \\ &= \frac{1}{\epsilon^2} - \frac{1}{12} + O(\epsilon^2).\end{aligned}$$

(1) Geometric series $\sum_{n\geq 1} e^{-\epsilon n} = \frac{1}{e^{\epsilon}-1}$, and differentiate it term by term.
(2) $\frac{\epsilon}{e^{\epsilon}-1} = \sum_k B_k \frac{\epsilon^k}{k!} = 1 - \frac{\epsilon}{2} + \frac{\epsilon^2}{12} - \cdots$ (Bernoulli numbers $B_2 = \frac{1}{6}$).

The finite part is $-\frac{1}{12} = \zeta(-1)$; the same number the ζ-function gives directly. So $a = -\frac{d-2}{24}$. The first excited states $\alpha^i_{-1}|0\rangle$ form a vector of the transverse rotation group $\mathrm{SO}(d-2)$, with $d-2$ components. A massive vector in d dimensions needs $d-1$ components (a vector of the little group $\mathrm{SO}(d-1)$), so Lorentz invariance forces this state to be massless, $1 + a = 0$:

$$-\frac{d-2}{24} = -1 \quad \Longleftrightarrow \quad d = 26.$$

For the superstring (Neveu–Schwarz sector) each of the $d-2$ transverse fermions has modes $r = \frac{1}{2}, \frac{3}{2}, \ldots$ and zero-point energy $-\frac{r}{2}$ per mode (the opposite sign to a boson). Regulate as before:

$$\sum_{r=\frac{1}{2},\frac{3}{2},\ldots} r\, e^{-\epsilon r} \overset{(1)}{=} -\frac{\mathrm{d}}{\mathrm{d}\epsilon}\frac{e^{-\epsilon/2}}{1-e^{-\epsilon}} \overset{(2)}{=} -\frac{\mathrm{d}}{\mathrm{d}\epsilon}\frac{1}{2\sinh(\epsilon/2)} \overset{(3)}{=} -\frac{\mathrm{d}}{\mathrm{d}\epsilon}\left(\frac{1}{\epsilon} - \frac{\epsilon}{24} + O(\epsilon^3)\right) = \frac{1}{\epsilon^2} + \frac{1}{24} + O(\epsilon^2).$$

(1) Geometric series $\sum_{k\geq 0} e^{-\epsilon(k+\frac{1}{2})} = e^{-\epsilon/2}/(1-e^{-\epsilon})$, differentiated term by term.
(2) Multiply numerator and denominator by $e^{\epsilon/2}$.
(3) $2\sinh(\epsilon/2) = \epsilon + \epsilon^3/24 + \ldots$, so its inverse is $\epsilon^{-1}(1 - \epsilon^2/24 + \ldots)$.

The finite part is $+\frac{1}{24}$ (the Hurwitz value $\zeta_H(-1, \frac{1}{2})$), so each fermion contributes $-\frac{1}{2}\cdot\frac{1}{24} = -\frac{1}{48}$ and

$$a_{\mathrm{NS}} = (d-2)\left(-\frac{1}{24} - \frac{1}{48}\right) = -\frac{d-2}{16}.$$

The massless vector is now $\psi^i_{-1/2}|0\rangle$ at level $\frac{1}{2}$, so $a_{\mathrm{NS}} = -\frac{1}{2}$, i.e. $d = 10$.

Physical picture: Why a Casimir energy knows about an anomaly

The $-\frac{1}{12}$ in the light-cone count and the $\frac{1}{12}$ in the Todd class are the same Bernoulli number $B_2 = \frac{1}{6}$ (halved). In the heat kernel it is the coefficient $\frac{1}{6}$ of R. The Todd class $\frac{x}{1-e^{-x}}$ and the regulated sum $\frac{\epsilon}{e^{\epsilon}-1}$ are the same generating function. Casimir energy, conformal anomaly and index theory are three faces of one Bernoulli number.

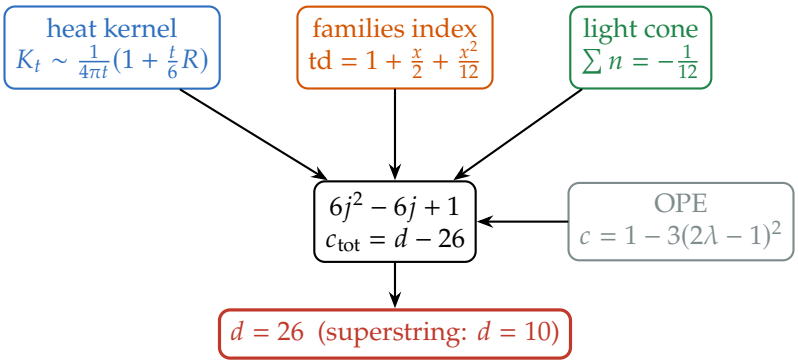


Figure 10.31: Four independent computations feed the same polynomial and the same critical dimension: the local Weyl anomaly (heat kernel), Nash's families index computation, the conformal-field-theory operator product, and the light-cone Casimir energy.

10.8.9 Quillen's metric, holomorphic factorisation and theta functions

Nash ends §4 with three results, each of which deserves a paragraph.

Quillen's metric. A line bundle has a first Chern class, but to write $\det(\bar\partial^*\bar\partial)$ as the *length* of the section $\det\bar\partial$ we need a metric on it. On a fibre where $\bar\partial$ has kernel H^0 and cokernel H^1, the line is $(\Lambda^{\max}H^0)^*\otimes\Lambda^{\max}H^1$ and inherits an L^2 metric $\|\cdot\|_{L^2}$ from the inner products of zero modes. That metric jumps when the kernel jumps. Quillen's fix is to multiply by the ζ-determinant of the Laplacian on the non-zero modes:

$$\|\cdot\|_{\mathrm{Q}}^2 := \|\cdot\|_{L^2}^2\cdot\det{}'\big(\bar\partial^*\bar\partial\big). \tag{10.200}$$

When an eigenvalue $\lambda\to 0$, the L^2 factor gains a zero mode and the $\det'$ factor loses the factor λ; the two changes compensate and $\|\cdot\|_{\mathrm{Q}}$ is smooth. Quillen proved that the curvature of the associated connection is the 2-form representing $c_1(\det\mathrm{Index})$ given by the families index formula at the level of forms. So the anomaly is literally the curvature of the Quillen metric: $d=26$ makes the bundle flat and (on simply connected pieces) trivial with a trivialising section of constant Quillen length.

Belavin–Knizhnik holomorphic factorisation. In the critical dimension the integrand of (10.178) is the Quillen norm squared of a section of a trivial bundle. Belavin and Knizhnik showed that then

$$Z_p=\int_{\mathcal{M}_p} f\,\bar f, \qquad f \text{ holomorphic on (the Teichmüller space covering) } \mathcal{M}_p. \tag{10.201}$$

The algebraic-geometry content is **Mumford's isomorphism**: in (10.197) the weight-2 bundle has coefficient $\frac{13}{12}$ and the weight-1 bundle (the Hodge bundle λ_1, $j=1$) has coefficient $\frac{1}{12}$, so $\lambda_2\cong\lambda_1^{\otimes 13}$; the holomorphic section implementing this isomorphism is f (times the holomorphic volume on moduli). The bosonic string measure is $|\text{Mumford form}|^2$.

Worked example 10.44: The one-loop torus, where everything is explicit

Take Σ a flat torus $\mathbb{C}/(\mathbb{Z}+\tau\mathbb{Z})$, $\tau=\tau_1+i\tau_2$ in the upper half plane. (Genus one is outside Nash's $p>1$, but the formulas are the cleanest illustration.) The non-zero eigenvalues of Δ are $\frac{4\pi^2}{\tau_2^2}|m+n\tau|^2$, and Kronecker's first limit formula evaluates the ζ-determinant:

$$\frac{\det'\Delta}{A}\ \propto\ \tau_2\,|\eta(\tau)|^4, \qquad \eta(\tau)=q^{1/24}\prod_{n\ge1}(1-q^n),\quad q=e^{2\pi i\tau},$$

with a τ-independent constant. The ghosts cancel two of the d bosons (the same $d-2$ as in light-cone gauge), and the measure on moduli is $\mathrm{d}^2\tau/\tau_2^2$. So

$$\begin{aligned} Z_1 &\overset{(1)}{=} \int_{\mathcal{F}}\frac{\mathrm{d}^2\tau}{\tau_2^2}\,\big(\tau_2|\eta(\tau)|^4\big)^{-(d-2)/2}\\ &\overset{(2)}{=} \int_{\mathcal{F}}\frac{\mathrm{d}^2\tau}{\tau_2^2}\,\tau_2^{-12}\,\eta(\tau)^{-24}\,\overline{\eta(\tau)^{-24}} \qquad (d=26), \end{aligned}$$

(1) (10.178) at genus one with the Kronecker formula; the integral is over a fundamental domain $\mathcal{F}$ of $\mathrm{SL}(2,\mathbb{Z})$.

(2) Put $d-2=24$.

The τ-dependence (apart from the powers of τ_2, which come from the zero modes and the non-compact momentum integral) has factorised into $f\bar f$ with $f=\eta^{-24}$ holomorphic. Expanding,

$$\eta(\tau)^{-24}=q^{-1}\prod_{n\ge1}(1-q^n)^{-24}=q^{-1}+24+324\,q+3200\,q^2+\cdots,$$

the coefficients count states: q^{-1} is the tachyon (the ground-state energy $a = -1$ again), and the 24 at order q^0 is the $d - 2 = 24$ oscillators α^i_{-1}, the massless vector. Modular invariance works because $\tau_2^{1/2}|\eta|^2$ is invariant under $\tau \mapsto -1/\tau$ ($\eta(-1/\tau) = \sqrt{-i\tau}\,\eta(\tau)$, $\tau_2 \mapsto \tau_2/|\tau|^2$).

Manin: theta functions. Manin showed that at higher genus the holomorphic f can be written with Riemann theta functions, $\vartheta[\delta](0,\Omega) = \sum_{n\in\mathbb{Z}^p} \exp\bigl(\pi i(n+a)^{\mathsf{T}}\Omega(n+a) + 2\pi i(n+a)^{\mathsf{T}}b\bigr)$, $\delta = (a,b)$ a characteristic, Ω the $p \times p$ period matrix. At genus one the relevant object is $\eta(\tau)^{24} \propto \vartheta_2\vartheta_3\vartheta_4(0,\tau)^8$ (Jacobi); at genus two it is the product of the ten even theta constants squared (the Siegel cusp form χ_{10}), and the genus-two bosonic string measure is proportional to $1/\chi_{10}$. Theta functions are the higher-genus generalisation of η.

Recap: Section 4

- The string integrand is a section of $\det \operatorname{Index} \bar\partial_Y$ with $\bar\partial_Y = \bar\partial_{K^*-(d/2)I}$ over Met/C^∞_+; powers of determinants are tensor powers, i.e. multiples in K-theory.
- Families index theorem with $\mathrm{td} = 1 + \frac{x}{2} + \frac{x^2}{12}$: $c_1 = \frac{26-d}{24}\kappa_1$. The anomaly vanishes iff $d = 26$.
- The degree-0 part of the same computation is Riemann–Roch and gives $\dim_{\mathbb{C}} \mathcal{M}_p = 3p - 3$.
- Mumford's polynomial $\frac{6j^2-6j+1}{12}$ gives the ghost central charge -26, the superghost $+11$ and the superstring dimension 10.
- Cross-checks: the TT operator product and the light-cone Casimir energy $\zeta(-1) = -\frac{1}{12}$.
- At $d = 26$: the Quillen connection is flat, $Z_p = \int f\bar f$ with f holomorphic (Belavin–Knizhnik), expressible by theta functions (Manin); at genus one $f = \eta^{-24}$.

10.9 Global anomalies

Everything so far in this chapter has been *infinitesimal*. We varied the gauge field by $\delta A = \nabla f$ with f in the Lie algebra, or the metric by $\delta g = 2\delta\sigma\, g$, and computed the first-order change of $\ln\det$. That is a derivative, and derivatives only see the part of the symmetry group that is connected to the identity. Suppose the group has several connected components. Then a function can be perfectly invariant under every infinitesimal transformation and still take different values on different components. The question of this section: *can the fermion determinant be invariant under all small gauge transformations and yet change sign under a large one?* Witten showed that the answer is yes, for the most innocent-looking theory imaginable: SU(2) Yang–Mills theory with a single left-handed fermion doublet. Such a change is a **global anomaly**.

Reading Nash (p. 281 top: the end of §3)

Before §4, Nash finishes the gravitational discussion. For a theory with fermions and gravity in $n = 4k - 2$ dimensions one works on the frame bundle P, an $\mathrm{SO}(4k-2)$-bundle, and the relevant symmetry group is $\mathrm{Aut}(P)$, the semi-direct product $\mathcal{G}_L \rtimes \mathrm{Diff}_0(M)$ of local Lorentz rotations and diffeomorphisms. The anomaly is then a class in $H^1(\mathrm{Aut}(P))$ mixing

the two. A local counterterm moves it back and forth between a Lorentz anomaly and a diffeomorphism anomaly (Alvarez-Gaumé and Ginsparg), much as a counterterm moves the U(1) anomaly between the vector and axial currents in QED_2 [QFT notes §22.2.2]. We will not need this below, except as a reminder that the "anomaly" is a property of the full symmetry group, not of one presentation of it.

10.9.1 Local versus global: invisible to derivatives

Here is the distinction in the smallest possible setting. Let G be a group and $\mathcal{Z} : G \to \mathbb{C}^\times$ a function (think: the determinant evaluated on gauge-transformed fields, divided by its value at the identity). A *local* anomaly is a non-zero derivative, $\mathrm{d}\ln\mathcal{Z} \neq 0$ at the identity. It is detected by a 1-form, and Nash's §2 packaged it as a de Rham class in $H^1_{\mathrm{dR}}(\mathcal{G})$. A *global* anomaly is the situation where $\mathrm{d}\ln\mathcal{Z} = 0$ everywhere, so $\mathcal{Z}$ is constant on each connected component, but the constants are different.

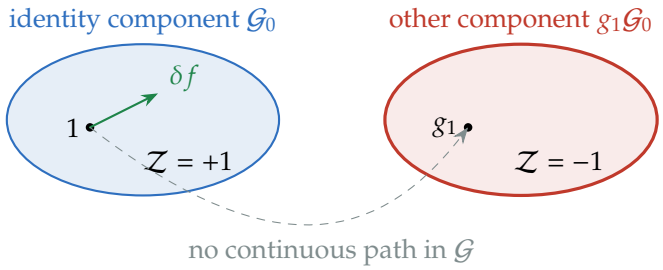


Figure 10.32: A global anomaly. The fermion determinant $\mathcal{Z}$ is invariant under every infinitesimal gauge transformation (green arrow), so it is constant on each component of $\mathcal{G}$. It is nevertheless not invariant: on the component that does not contain the identity it has the opposite sign. No derivative computed at 1 can see this.

The analogy with torsion. Nash compares this with torsion in homology. De Rham cohomology is built from differential forms, and a differential form can only measure things with real (or integer) values that vary smoothly. A torsion class, such as the generator of $H_1(\mathbb{RP}^2;\mathbb{Z}) = \mathbb{Z}_2$, is a loop that is not a boundary, but twice it is. No closed 1-form can integrate to something non-zero on it: if $\oint_\gamma \omega = c$ then $2c = \oint_{2\gamma}\omega = 0$, so $c = 0$. Forms are blind to torsion, and global anomalies turn out to be torsion classes (Section 10.9.6).

10.9.2 A global anomaly you can compute by hand: one fermion on a circle

Before Witten's four-dimensional example, here is the whole mechanism in $0+1$ dimensions, where every step is explicit. Take a single complex fermion $\psi(\tau)$ on a Euclidean time circle of length β, coupled to a U(1) gauge field $A = A(\tau)\,\mathrm{d}\tau$:

$$S = \int_0^\beta \mathrm{d}\tau\; \bar\psi\,(\partial_\tau - iA)\,\psi, \qquad \psi(\tau+\beta) = -\psi(\tau).$$

A gauge transformation $\psi \mapsto e^{i\alpha}\psi$, $A \mapsto A + \partial_\tau\alpha$ leaves S invariant. The function $e^{i\alpha(\tau)}$ must be single valued, so $\alpha(\beta) = \alpha(0) + 2\pi w$ with $w \in \mathbb{Z}$ the winding number. The gauge group has components labelled by w: $\pi_0(\mathcal{G}) = \mathbb{Z}$. Transformations with $w \neq 0$ are the **large** gauge transformations. The only gauge-invariant quantity built from A is the holonomy e^{ia}, $a = \oint A$; a large gauge transformation shifts $a \mapsto a + 2\pi w$.

Worked example 10.45: $Z = 2\cos(a/2)$ and its sign flip

Gauge-fix $A = a/\beta$ (constant). Antiperiodic eigenfunctions are $e^{i\omega_n\tau}$ with $\omega_n = \frac{2\pi}{\beta}(n+\frac{1}{2})$, and the operator $\partial_\tau - ia/\beta$ has eigenvalues $\lambda_n = \frac{i}{\beta}\big(2\pi(n+\frac{1}{2}) - a\big)$, $n \in \mathbb{Z}$. Pair the eigenvalue n with $-n-1$ and normalise by the value at $a = 0$:

$$\begin{aligned}\frac{Z(a)}{Z(0)} &\overset{(1)}{=} \prod_{n\geq 0} \frac{\lambda_n \lambda_{-n-1}\big|_a}{\lambda_n \lambda_{-n-1}\big|_0} = \prod_{n\geq 0} \frac{(2\pi(n+\frac{1}{2}))^2 - a^2}{(2\pi(n+\frac{1}{2}))^2} \\ &\overset{(2)}{=} \prod_{n\geq 0}\Big(1 - \frac{a^2}{(2n+1)^2\pi^2}\Big) \\ &\overset{(3)}{=} \cos\frac{a}{2}.\end{aligned}$$

(1) The Grassmann integral is $\det(\partial_\tau - iA) = \prod_n \lambda_n$; grouping n with $-n-1$ makes each factor $-\beta^{-2}\big((2\pi(n+\frac{1}{2}))^2 - a^2\big)$, and the a-independent prefactors cancel in the ratio. (This pairing is a choice of regularisation that respects charge conjugation $a \mapsto -a$.)

(2) Divide numerator and denominator by $(2\pi(n+\frac{1}{2}))^2 = (2n+1)^2\pi^2$.

(3) Euler's product $\cos x = \prod_{n\geq 0}\big(1 - \frac{4x^2}{(2n+1)^2\pi^2}\big)$ with $x = a/2$.

With $Z(0) = 2$ (the trace over the two states, empty and filled), $Z(a) = 2\cos(a/2)$. Now perform a large gauge transformation, $a \mapsto a + 2\pi$:

$$Z(a+2\pi) = 2\cos\Big(\frac{a}{2} + \pi\Big) = -Z(a).$$

The partition function changes sign under a gauge transformation, although it is perfectly invariant under small ones (it depends only on a). The Hamiltonian reading: the two states have charges $Q = \pm\frac{1}{2}$, and $Z = \operatorname{Tr} e^{iaQ} = e^{ia/2} + e^{-ia/2}$; half-integer charges are not single-valued under $a \mapsto a + 2\pi$.

Where the sign comes from: one eigenvalue crosses zero. As a increases from 0 to 2π, each eigenvalue $\lambda_n \propto (2\pi(n+\frac{1}{2}) - a)$ moves down by 2π, i.e. to the position of its neighbour λ_{n-1}. The spectrum as a set is unchanged (it had to be; a and $a + 2\pi$ are gauge equivalent), but exactly one eigenvalue, λ_0, has passed through zero, at $a = \pi$. In the paired product this is the factor $(\pi^2 - a^2)$ changing sign. This is the prototype of everything in this section: *a gauge-invariant spectrum, a gauge-non-invariant product, and the difference is a count of eigenvalues crossing zero.* Figure 10.33 shows it.

Try it: Two fermions, and a counterterm

(a) Show that with two identical fermions $Z = 4\cos^2(a/2)$ is invariant under $a \mapsto a + 2\pi$. (b) Show that multiplying Z by the "Chern–Simons" counterterm $e^{ia/2}$ gives a gauge-invariant $1 + e^{ia}$, but that this is no longer invariant under charge conjugation $a \mapsto -a$. *Answer:* (a) $\cos^2$ is π-periodic. (b) $e^{ia/2}\cdot 2\cos(a/2) = 1 + e^{ia}$; $1 + e^{-ia} \neq 1 + e^{ia}$. One can trade the global anomaly for the loss of a discrete symmetry, but one cannot keep both. An even number of fermions has no anomaly at all: the same pattern as Witten's "p even".

10.9.3 Counting components: $\pi_0(\mathcal{G}) = \pi_4(\mathrm{SU}(2)) = \mathbb{Z}_2$

Now the four-dimensional theory. Space-time is $M = S^4$ (or $\mathbb{R}^4$ with fields approaching a constant at infinity, which is the same thing by stereographic projection). A gauge transformation

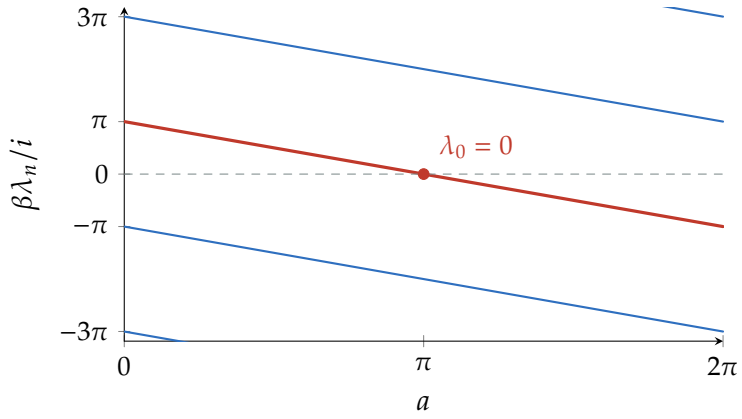


Figure 10.33: Spectral flow in the 0 + 1 toy model. As the holonomy a runs from 0 to 2π, every eigenvalue slides down one step; the spectrum at $a = 2\pi$ is the same set as at $a = 0$. One eigenvalue (red) crosses zero, so the product of eigenvalues, and with it Z, changes sign.

is a map $g : S^4 \to G$, and we fix $g(\infty) = 1$ (these are the transformations that act freely; the constant ones are global symmetries). The gauge group is then the based mapping space $\Omega^4 G$, and its components are homotopy classes of maps $S^4 \to G$ (Nash (10.44)):

$$\pi_0(\mathcal{G}) = \pi_0\big(\Omega^{2m} G\big) = \pi_{2m}(G), \qquad M = S^{2m}. \tag{10.202}$$

Two gauge transformations are in the same component iff one can be deformed into the other; that is exactly a homotopy of maps of spheres.

SU(2) **is the 3-sphere.** Every $g \in \mathrm{SU}(2)$ can be written

$$g = x_0\,\mathbb{1} + i\,x_k\sigma_k = \begin{pmatrix} x_0 + ix_3 & ix_1 + x_2 \\ ix_1 - x_2 & x_0 - ix_3 \end{pmatrix}, \qquad x_0^2 + x_1^2 + x_2^2 + x_3^2 = 1, \tag{10.203}$$

with σ_k the Pauli matrices. The determinant is $x_0^2 + x_3^2 + x_1^2 + x_2^2 = 1$, and unitarity follows from $\sigma_k^\dagger = \sigma_k$ and $\{\sigma_j, \sigma_k\} = 2\delta_{jk}$. So $\mathrm{SU}(2) \cong S^3$, and $\pi_4(\mathrm{SU}(2)) = \pi_4(S^3)$.

Key idea: Nash (10.45)

$$\pi_0(\mathcal{G}) = \pi_4(\mathrm{SU}(2)) = \pi_4(S^3) = \mathbb{Z}_2.$$

There is exactly one class of gauge transformations on S^4 that cannot be deformed to the identity, and doing it twice can.

Why $\mathbb{Z}_2$: framed circles and the belt trick. The Pontryagin–Thom construction turns maps $S^{n+k} \to S^n$ into geometry. Pick a regular value $y \in S^n$; its preimage is a k-dimensional submanifold of S^{n+k}, and the derivative of the map identifies the normal space at each preimage point with $T_y S^n$: a **framing** of the normal bundle. Homotopic maps give framed submanifolds that are framed-cobordant, and conversely. For $S^4 \to S^3$ ($n = 3$, $k = 1$) the preimage is a collection of circles in S^4, each carrying three normal vector fields. In S^4 any circle can be unknotted and shrunk, so only the framing matters. Going once round the circle, the normal frame traces a loop in SO(3), and $\pi_1(\mathrm{SO}(3)) = \mathbb{Z}_2$: a frame rotated by 2π is not deformable to the unrotated one, but a frame rotated by 4π is (Dirac's belt trick). Hence

$$\pi_4(S^3) \cong \{\text{framed circles}\}/\text{cobordism} \cong \pi_1(\mathrm{SO}(3))/(\text{shift}) \cong \mathbb{Z}_2.$$

(The "shift" is a bookkeeping subtlety: the framing induced by bounding a disc is taken as the zero; the result is the same $\mathbb{Z}_2$.) The generator is also the suspension $\Sigma\eta$ of the Hopf map $\eta : S^3 \to S^2$ (Chapter 1).

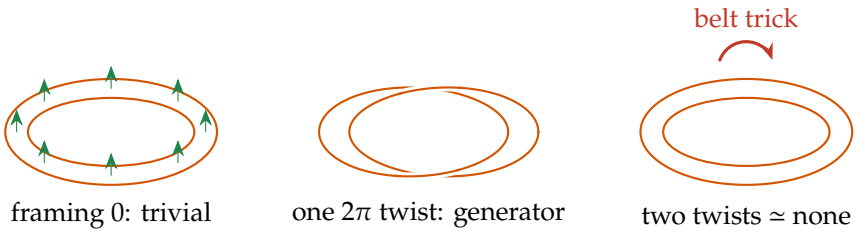


Figure 10.34: Maps $S^4 \to S^3$ as framed circles (Pontryagin–Thom). The only invariant is how many times the normal frame (drawn as a ribbon) twists going round, counted in $\pi_1(\mathrm{SO}(3)) = \mathbb{Z}_2$. One twist cannot be undone; two twists can, exactly as in Dirac's belt trick. This is why $\pi_4(\mathrm{SU}(2)) = \mathbb{Z}_2$.

Worked example 10.46: Which groups have a π_4?

From the tables of homotopy groups (Chapter 1):

$$\pi_4(\mathrm{SU}(2)) = \mathbb{Z}_2, \qquad \pi_4(\mathrm{SU}(N)) = 0\ (N \geq 3), \qquad \pi_4(\mathrm{Sp}(n)) = \mathbb{Z}_2, \qquad \pi_4(\mathrm{SO}(4)) = \mathbb{Z}_2 \oplus \mathbb{Z}_2.$$

Since $\mathrm{Sp}(1) = \mathrm{SU}(2)$, the symplectic groups inherit the anomaly; $\mathrm{SO}(4) = (\mathrm{SU}(2) \times \mathrm{SU}(2))/\mathbb{Z}_2$ has one copy for each factor. For SU(3) the large transformation of SU(2) becomes small once SU(2) is embedded in SU(3): there is now room to unwind it. We return to this below.

10.9.4 Why the determinant has a sign: pseudo-reality and Pfaffians

Nash writes that the fermionic integral produces $\sqrt{\det(\not{D}_A^\dagger \not{D}_A)}$ and that global gauge transformations "change the sign" of this square root. A positive square root of a positive number cannot change sign, so we must first understand which square root is meant. The answer is a Pfaffian, and it exists because the doublet of SU(2) is **pseudo-real**.

Pseudo-reality in one line. Let $\varepsilon = i\sigma_2 = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$. Then for the Pauli matrices

$$\begin{aligned} & \varepsilon\, \sigma_k^*\, \varepsilon^{-1} \overset{(1)}{=} -\sigma_k, \qquad k = 1,2,3, \\ \Rightarrow \quad & \varepsilon\, g^*\, \varepsilon^{-1} \overset{(2)}{=} g \qquad \text{for all } g \in \mathrm{SU}(2). \end{aligned} \tag{10.204}$$

(1) Check each: $\sigma_1^* = \sigma_1$, $\sigma_2^* = -\sigma_2$, $\sigma_3^* = \sigma_3$, and conjugation by ε flips the sign of σ_1, σ_3 and fixes σ_2; in all three cases the net effect is $-\sigma_k$.

(2) Write $g = x_0 + i x_k \sigma_k$ as in (10.203). Then $g^* = x_0 - i x_k \sigma_k^*$ and $\varepsilon g^* \varepsilon^{-1} = x_0 - i x_k(-\sigma_k) = g$.

So the doublet is equivalent to its complex conjugate, via a matrix ε with $\varepsilon\varepsilon^* = -1$. That minus sign is what "pseudo-real" (as opposed to real) means. The anti-linear map $J = \varepsilon K$ (K = complex conjugation) commutes with every g and satisfies $J^2 = \varepsilon\varepsilon^* = -1$.

The Weyl operator is real. The same structure exists on spinors: in four Euclidean dimensions the chiral spinors $S^\pm$ are also pseudo-real, with a charge-conjugation map C, $C^2 = -1$ on each chirality (Clifford algebra details in Chapter T4). On a left-handed doublet, $\psi \in \Gamma(S^+ \otimes \mathbb{C}^2)$, the combined anti-linear map $\mathcal{J} = C \otimes J$ squares to $(-1)(-1) = +1$ and commutes with the coupled Dirac operator. An anti-linear map squaring to $+1$ is a *real structure*: the fixed points of $\mathcal{J}$ form

a real vector space, and $\not{D}_A$ restricts to a real operator on it. This is Nash's statement that "$i\not{D}_A$ coupled to an SU(2) doublet can be taken to be a real skew-adjoint operator". In practice this means: a single Weyl doublet is like a Majorana fermion, and its path integral is a Pfaffian,

$$\int \mathcal{D}\psi \; e^{-\frac{1}{2}\int \psi^{\mathsf{T}} C \not{D}_A \psi} = \mathrm{Pf}(C\not{D}_A), \qquad \mathrm{Pf}(\cdot)^2 = \det(\cdot). \tag{10.205}$$

Working definition 10.11 (Pfaffian)

For a real antisymmetric $2N \times 2N$ matrix B, the **Pfaffian** is the polynomial $\mathrm{Pf}(B) = \frac{1}{2^N N!}\sum_\pi \mathrm{sgn}(\pi)\prod_{i=1}^{N} B_{\pi(2i-1)\pi(2i)}$ [a sum over all permutations π of $1,\dots,2N$. Each way of pairing up the indices appears $2^N N!$ times, hence the prefactor. For 2×2 it is just B_{12}], and it satisfies $\mathrm{Pf}(B)^2 = \det B$. *How to compute it:* bring B to block form $\bigoplus_k \left(\begin{smallmatrix} 0 & \lambda_k \\ -\lambda_k & 0 \end{smallmatrix}\right)$ by a rotation $R \in \mathrm{SO}(2N)$; then $\mathrm{Pf}(B) = \prod_k \lambda_k$. The Grassmann integral [keep only the term containing every θ_i exactly once] $\int \mathrm{d}^{2N}\theta \, e^{-\frac{1}{2}\theta^{\mathsf{T}} B \theta} = \mathrm{Pf}(B)$.

Worked example 10.47: 2×2 and 4×4 Pfaffians

For $B = \left(\begin{smallmatrix} 0 & b \\ -b & 0 \end{smallmatrix}\right)$, $\mathrm{Pf}\, B = b$ and $\det B = b^2$. For a 4×4 antisymmetric B,

$$\mathrm{Pf}\, B = B_{12}B_{34} - B_{13}B_{24} + B_{14}B_{23},$$

and one checks $\det B = (B_{12}B_{34} - B_{13}B_{24} + B_{14}B_{23})^2$ by expanding. The point: $\det B$ is a perfect square *as a polynomial*, and $\mathrm{Pf}\, B$ is its square root *with a definite sign*, a polynomial in the entries. Along a path $B(t)$ the Pfaffian is a smooth function that can pass through zero and change sign, while $\sqrt{\det}$ would have a kink. The sign of the Pfaffian is not a matter of convention once it is fixed at one point: it is determined by continuity.

Kramers pairs. The Pfaffian has a spectral description that makes the sign flip visible. The Hermitian operator $H = i\not{D}_A$ on the full (both chiralities) doublet space commutes with the anti-unitary J, $J^2 = -1$. By Kramers' theorem every eigenvalue is doubly degenerate: if $H\psi = \lambda\psi$ then $HJ\psi = \lambda J\psi$ (λ is real and J commutes with H), and $J\psi \perp \psi$. For the last claim use the defining property of an anti-unitary map, $\langle Ja, Jb\rangle = \langle b, a\rangle$:

$$\langle J\psi, \psi\rangle \overset{(1)}{=} -\langle J\psi, J^2\psi\rangle \overset{(2)}{=} -\langle J\psi, \psi\rangle \quad\Longrightarrow\quad \langle J\psi, \psi\rangle = 0.$$

(1) $J^2 = -1$, so $\psi = -J^2\psi$.
(2) Anti-unitarity with $a = \psi$, $b = J\psi$.

So each eigenvalue of $i\not{D}_A$ on the doublet comes as a Kramers pair $\{\psi, J\psi\}$. In addition γ_5 anticommutes with $\not{D}_A$, so the spectrum is symmetric: λ and $-\lambda$ both occur. Hence $\det(\not{D}_A^{*}\not{D}_A) = \prod_k \lambda_k^4$, the product running over the distinct positive eigenvalues λ_k (each counted once), while the Weyl-doublet integral is

$$\mathrm{Pf}(A) = \prod_k \lambda_k, \tag{10.206}$$

one factor for each Kramers pair in the positive half of the spectrum. Squaring gives $\prod_k \lambda_k^2$, which is $\sqrt{\det(\not{D}_A^{*}\not{D}_A)}$, as it should be. The expression $\prod_k \lambda_k$ with all $\lambda_k > 0$ is of course positive; the point is that "which half is the positive half" is only defined when no eigenvalue vanishes. Follow a continuous path of gauge fields: each factor is a continuous function $\lambda_k(t)$, and the product is defined by continuity from a reference configuration. If one Kramers pair crosses zero along the path, its factor changes sign and so does Pf.

Watch out: There is no perturbative $\mathrm{SU}(2)$ anomaly

The local anomaly of a Weyl fermion in representation R is proportional to $d^{abc} = \operatorname{tr}_R\big(T^a\{T^b,T^c\}\big)$. For the doublet, $T^a = \frac{1}{2}\sigma_a$ and $d^{abc} = \frac{1}{8}\operatorname{tr}\big(\sigma_a\{\sigma_b,\sigma_c\}\big) = \frac{1}{4}\delta_{bc}\operatorname{tr}\sigma_a = 0$. More generally $d^{abc} = 0$ for every pseudo-real or real representation. Nash expresses this topologically: the local anomaly lives in $H^1_{\mathrm{dR}}(\mathcal{G})$, and $\pi_1(\mathcal{G}) = \pi_5(\mathrm{SU}(2)) = \mathbb{Z}_2$ is pure torsion, so $H^1_{\mathrm{dR}}(\mathcal{G}) = 0$. The families argument of §2 therefore has nothing to say. Whatever is wrong with the theory must be global.

10.9.5 The mod-2 index and spectral flow

Real skew-adjoint operators have a $\mathbb{Z}_2$ index. Let D be a real skew-adjoint Fredholm operator, $D^{\mathsf{T}} = -D$. Its non-zero eigenvalues (over $\mathbb{C}$) come in pairs $\pm i\lambda$, and the real invariant subspaces are 2-planes on which D acts as $\left(\begin{smallmatrix}0 & \lambda\\ -\lambda & 0\end{smallmatrix}\right)$. The ordinary index is useless: $\operatorname{index} D = \dim\operatorname{Ker} D - \dim\operatorname{Ker} D^* = 0$ since $D^* = -D$. But the kernel can only change by two dimensions at a time: an eigenvalue $i\lambda$ can reach zero only together with its partner $-i\lambda$, and then a whole 2-plane joins the kernel.

Working definition 10.12 (mod-2 index)

For a real skew-adjoint Fredholm operator D [$D\psi = 0$ has only finitely many independent solutions $\psi \neq 0$], the **mod-2 index** is

$$\operatorname{index}_2 D := \dim\operatorname{Ker} D \pmod 2 \in \mathbb{Z}_2.$$

It is unchanged by continuous deformations [moving D along a path of such operators, e.g. $\not{D}_{A_t}$ as the gauge field moves. Independent zero modes then appear or disappear two at a time]. *How to compute it:* find the kernel in one convenient member of the deformation class [all operators reachable from D by such a path] and count.

Worked example 10.48: Finite matrices and the circle

(a) $D(t) = \left(\begin{smallmatrix}0 & t\\ -t & 0\end{smallmatrix}\right)$ on $\mathbb{R}^2$: $\dim\operatorname{Ker} = 2$ at $t = 0$, 0 otherwise; $\operatorname{index}_2 = 0$ throughout. A real skew 3×3 matrix always has a kernel of odd dimension (its non-zero spectrum is paired, and 3 is odd), so $\operatorname{index}_2 = 1$: in finite dimensions the mod-2 index is just the parity of the dimension. (b) Infinite dimensions, where it is interesting: $D = \frac{\mathrm{d}}{\mathrm{d}x}$ on real functions on the circle $x \in [0, 2\pi]$. It is real and skew (integrate by parts). With *periodic* boundary conditions the kernel is the constants, $\dim\operatorname{Ker} = 1$, so $\operatorname{index}_2 D = 1$. With *antiperiodic* boundary conditions ($f(x+2\pi) = -f(x)$) the Fourier modes are $\cos(n+\frac{1}{2})x$, $\sin(n+\frac{1}{2})x$ and none is constant: $\operatorname{index}_2 D = 0$. The two spin structures on S^1 are distinguished by the mod-2 index of their Dirac operators; no continuous deformation of the operator can change one into the other.

Families over a circle. Now the key idea (Nash p. 284). A loop of gauge fields $A_t = (1-t)A + tA^g$, $t \in [0,1]$, runs from A to its gauge transform A^g (this is a straight line in the affine space $\mathcal{A}$, but a closed loop in $\mathcal{A}/\mathcal{G}$). At the two ends the operators are conjugate, $\not{D}_{A^g} = g^{-1}\not{D}_A g$, so they have the same spectrum. The Pfaffian at the two ends differs by $(-1)^N$, with N the number of Kramers pairs that crossed zero along the way (Figure 10.35):

$$\operatorname{Pf}(A^g) = (-1)^N \operatorname{Pf}(A), \qquad N = \text{spectral flow of the family} \pmod 2. \tag{10.207}$$

Witten's step is to reinterpret the family as *one* operator in five dimensions. On $S^4 \times [0,1]$ with t the fifth coordinate, glue the ends using g (the **mapping torus**); the operator $\mathcal{D}_5 = \partial_t + \not{D}_{A_t}$ is a real skew-adjoint operator on this five-manifold, and its zero modes correspond, in the adiabatic limit of a slowly varying A_t, to the level crossings. So

$$N \equiv \operatorname{index}_2 \mathcal{D}_5 \pmod 2.$$

The right side is topological: it depends only on the component of g in $\pi_4(\mathrm{SU}(2))$. Witten computed it for the generator and found 1.

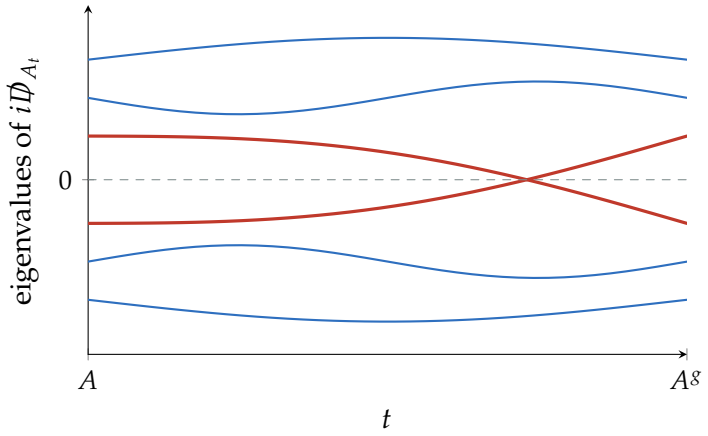


Figure 10.35: Witten's spectral flow. Each curve is a Kramers pair (doubly degenerate); the spectrum is symmetric about 0. Along the path from A to A^g the spectrum returns to itself, but one Kramers pair (red) has crossed from positive to negative while its mirror crossed the other way. The Pfaffian, the product over the positive branch, has lost one positive factor and gained one negative one: it has changed sign.

Key idea: Witten's SU(2) anomaly

For SU(2) gauge theory on S^4 with p left-handed Weyl doublets and g the generator of $\pi_4(\mathrm{SU}(2)) = \mathbb{Z}_2$,

$$Z_\psi(A^g) = (-1)^p\, Z_\psi(A), \qquad Z_\psi = \mathrm{Pf}(C\not{D}_A)^p .$$

For odd p the fermion integral is odd under a gauge transformation, so the integral over gauge fields vanishes identically, $\int \mathcal{D}A\, e^{-S_{\mathrm{YM}}} Z_\psi(A) = 0$ (the contributions of A and A^g cancel), and so do all gauge-invariant correlation functions. The theory is inconsistent. For even p it is fine.

Which representations are anomalous? The same argument applies to any pseudo-real representation R of SU(2), i.e. half-integer isospin j. The parity of the spectral flow turns out to equal the parity of the number of zero modes of a Weyl fermion in R in a one-instanton background, i.e. of the index $\operatorname{index} \not{D}_A = 2T(R)$, with $T(\frac{1}{2}) = \frac{1}{2}$:

$$2T(j) = \tfrac{2}{3}\, j(j+1)(2j+1): \qquad j = \tfrac{1}{2} \mapsto 1, \quad j = \tfrac{3}{2} \mapsto 10, \quad j = \tfrac{5}{2} \mapsto 35, \quad j = \tfrac{7}{2} \mapsto 84.$$

So isospin $\frac{1}{2}, \frac{5}{2}, \frac{9}{2}, \ldots$ are anomalous and $\frac{3}{2}, \frac{7}{2}, \ldots$ are not. Physically: in an instanton transition, fermion number changes by the index ('t Hooft). If that number is odd, the fermion parity $(-1)^F$, which is the gauge transformation $g = -1 \in \mathrm{SU}(2)$ acting on the fermions, cannot be a good symmetry of the gauge-invariant states, and the theory cannot be consistently quantised.

The $(-1)^F$ argument, step by step. This last argument uses only facts we have already computed, so we give it in full. Quantise on space S^3 with p Weyl doublets.

1. $(-1)^F$ *is a gauge transformation.* The constant gauge transformation $g = -\mathbb{1} \in \mathrm{SU}(2)$ acts on every doublet field by $\psi \mapsto -\psi$. On Fock states it therefore acts as $(-1)^F$, where F is the number of doublet quanta (particles and antiparticles each flip sign).

2. *It is a small gauge transformation.* The path $g_s = \exp(i\pi s\sigma_3) = \mathrm{diag}(e^{i\pi s}, e^{-i\pi s})$, $s \in [0,1]$, runs from $\mathbb{1}$ to $-\mathbb{1}$ through constant gauge transformations. So $-\mathbb{1}$ lies in the identity component of the gauge group on S^3, and Gauss's law (which imposes invariance under the identity component) requires

$$(-1)^F|\mathrm{phys}\rangle = |\mathrm{phys}\rangle$$

   for every physical state.

3. *Instantons change F by the index.* In a one-instanton background each doublet has index $\not{D}_A = 2T(\frac{1}{2}) = 1$ zero mode (Chapter 10). The Grassmann integral over a zero mode vanishes unless the mode appears in the integrand, so an amplitude in the one-instanton sector is non-zero only if it contains one extra ψ for each doublet. Hence tunnelling from an initial state $|i\rangle$ to a final state $|f\rangle$ obeys $F_f - F_i \equiv p \pmod 2$.

4. *Contradiction for odd p.* Combine 2 and 3:

$$(-1)^{F_f} \overset{(1)}{=} (-1)^{F_i+p} \overset{(2)}{=} (-1)^p\,(-1)^{F_i} \overset{(3)}{=} -(-1)^{F_i} \qquad (p \text{ odd}).$$

   (1) Step 3.
   (2) Arithmetic of signs.
   (3) p odd.

   So if $|i\rangle$ is physical ($(-1)^{F_i} = 1$), $|f\rangle$ is not. The gauge-invariant Hamiltonian would map physical states to unphysical ones, which is impossible. For even p no contradiction arises.

The Euclidean statement (the Pfaffian changes sign along the path from A to A^g) and this Hamiltonian statement are the same fact. The bridge is the Atiyah–Singer mod-2 index theorem, which identifies the parity of the spectral flow along a loop in the class of $g \in \pi_4(\mathrm{SU}(2))$ with the parity of the instanton index; we do not reproduce its proof.

Worked example 10.49: The Standard Model passes

The weak gauge group of the Standard Model is $\mathrm{SU}(2)_L$. Per generation, the left-handed doublets are the lepton doublet $(\nu_e, e)_L$ and the quark doublet $(u, d)_L$ in each of three colours:

$$p = 1 + 3 = 4 \qquad \text{per generation.}$$

Even, so there is no Witten anomaly, and this is automatic once the perturbative anomalies cancel (they require the same quark–lepton balance; [QFT notes §23.4, pp. 1458–1461]). A hypothetical fourth generation of leptons alone ($p = 1$) would be inconsistent; so would one extra colour of quark.

The SU(3) cross-check. Embed $\mathrm{SU}(2) \subset \mathrm{SU}(3)$ as the upper-left block. The triplet of SU(3) restricts to a doublet plus a singlet. Since $\pi_4(\mathrm{SU}(3)) = 0$, the large SU(2) transformation g is small in SU(3): there is a path g_s in $\mathcal{G}_{\mathrm{SU}(3)}$ from 1 to g. Along that path the SU(3) triplet has a *perturbative* anomaly (its $d^{abc} \neq 0$), and integrating the infinitesimal anomaly along the path

gives a phase $e^{i\pi} = -1$. Witten observed that this is the same sign: a single SU(2) doublet is anomalous because a single SU(3) triplet is. Anomalies are consistent under embedding, the global one of the subgroup being a remnant of the local one of the larger group.

10.9.6 The topology: a torsion class and a Möbius bundle

Nash reformulates Witten's result in the language of determinant bundles, which is what connects it to the rest of the chapter.

Pulling back along a loop. Let $Y = \mathcal{A}/\mathcal{G}$ and choose a map $\alpha : S^1 \to Y$. The family $\not{D}_Y$ over Y pulls back to a family over S^1, and so do its index and determinant bundles (Nash (10.46)):

$$\alpha^* \det \operatorname{Index} \not{D}_Y = \det \operatorname{Index} \not{D}_{S^1} = (\operatorname{Ker} \not{D}_{S^1})^* \otimes (\operatorname{Ker} \not{D}^*_{S^1}). \tag{10.208}$$

(For a determinant bundle the kernels mean their top exterior powers.) Up to isomorphism the pull-back depends only on the homotopy class of α, an element of $\pi_1(Y)$.

$\pi_1(\mathcal{A}/\mathcal{G})$. $\mathcal{A}$ is an affine space (any two connections differ by a Lie-algebra-valued 1-form), so it is contractible, and $\mathcal{G}$ acts freely on it (based gauge transformations). The fibration $\mathcal{G} \to \mathcal{A} \to \mathcal{A}/\mathcal{G}$ has the long exact homotopy sequence $\cdots \to \pi_1(\mathcal{A}) \to \pi_1(\mathcal{A}/\mathcal{G}) \to \pi_0(\mathcal{G}) \to \pi_0(\mathcal{A}) \to \cdots$, and with $\pi_k(\mathcal{A}) = 0$:

$$\pi_1(\mathcal{A}/\mathcal{G}) \overset{(1)}{=} \pi_0(\mathcal{G}) \overset{(2)}{=} \mathbb{Z}_2, \qquad H_1(\mathcal{A}/\mathcal{G};\mathbb{Z}) \overset{(3)}{=} \mathbb{Z}_2. \tag{10.209}$$

(1) Exactness with contractible $\mathcal{A}$. Concretely: a loop in $\mathcal{A}/\mathcal{G}$ lifts to a path in $\mathcal{A}$ from A to some A^g; the component of g is the invariant.
(2) (10.202) and Nash (10.45).
(3) Hurewicz: H_1 is the abelianisation of π_1, and $\mathbb{Z}_2$ is already abelian.

The universal coefficient theorem. To get a class for a line bundle we need H^2, not H_1. The universal coefficient theorem (Nash (10.48), Bott–Tu) relates them: write $H_i(M;\mathbb{Z}) = F_i \oplus T_i$ with F_i free and T_i torsion. Then

$$H^i(M;\mathbb{Z}) \cong F_i \oplus T_{i-1}. \tag{10.210}$$

Torsion moves *up one degree* from homology to cohomology.

Worked example 10.50: The universal coefficient theorem for $\mathbb{RP}^2$ and $\mathbb{RP}^3$

$\mathbb{RP}^2$ has $H_0 = \mathbb{Z}, H_1 = \mathbb{Z}_2, H_2 = 0$ (it is non-orientable). By (10.210): $H^0 = \mathbb{Z}, H^1 = F_1 \oplus T_0 = 0$, $H^2 = F_2 \oplus T_1 = \mathbb{Z}_2$. Check $H^1 = 0$ directly: $H^1(M;\mathbb{Z}) = \operatorname{Hom}(H_1, \mathbb{Z})$, and there is no non-zero homomorphism $\mathbb{Z}_2 \to \mathbb{Z}$. And $H^2(\mathbb{RP}^2;\mathbb{Z}) = \mathbb{Z}_2$ is generated by the first Chern class of the complexified tautological line bundle, a line bundle with $2c_1 = 0$ but $c_1 \neq 0$. For $\mathbb{RP}^3$: $H_* = (\mathbb{Z}, \mathbb{Z}_2, 0, \mathbb{Z})$ and $H^* = (\mathbb{Z}, 0, \mathbb{Z}_2, \mathbb{Z})$. The same shift.

For $\mathcal{A}/\mathcal{G}$, Nash's (10.49):

$$F_1(\mathcal{A}/\mathcal{G}) = 0, \quad F_2(\mathcal{A}/\mathcal{G}) = 0, \quad T_1(\mathcal{A}/\mathcal{G}) = \mathbb{Z}_2 \quad \Longrightarrow \quad H^2(\mathcal{A}/\mathcal{G};\mathbb{Z}) \supseteq T_1 = \mathbb{Z}_2. \tag{10.211}$$

($F_2 = 0$ because $H^2_{\mathrm{dR}}(\mathcal{A}/\mathcal{G}) = 0$ in this case: there is no local anomaly.) So every line bundle L on $\mathcal{A}/\mathcal{G}$ has $c_1(L)$ torsion: $2c_1(L) = 0$. No curvature form can detect it (its de Rham image is zero), which is the precise sense in which the global anomaly is invisible to the methods of §2.

The real determinant bundle is a Möbius band. Because $\not{D}_{S^1}$ is a family of *real* operators, its determinant bundle is a *real* line bundle. The structure group of a real line bundle (with a metric) is $\mathrm{O}(1) = \{\pm 1\} = \mathbb{Z}_2$. Over S^1 there are exactly two real line bundles, since $\pi_0(\mathrm{O}(1)) = \mathbb{Z}_2$ classifies the transition function across one cut: the cylinder (+1) and the Möbius band (−1).

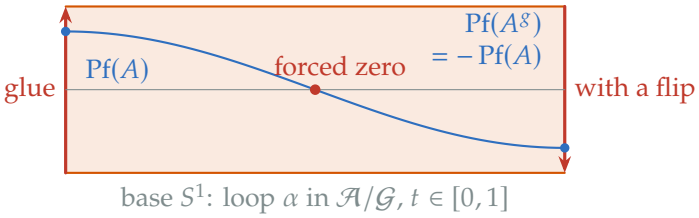


Figure 10.36: The determinant line bundle of the real family over the loop α is a Möbius band. A section (blue) that starts at $+\mathrm{Pf}(A)$ must arrive at the glued edge as $-\mathrm{Pf}(A)$ to be continuous across the twisted gluing, so it must vanish somewhere: that zero is the level crossing. A global section with no zero, i.e. a gauge-invariant non-vanishing fermion determinant, does not exist.

Now put the pieces together, as Nash does on p. 285. Try to define a trivialising section "$\det \not{D}$" of $\det \mathrm{Index}\, \not{D}_{S^1}$. Going once round S^1 it is not periodic: it comes back multiplied by the non-trivial element -1 of the structure group. So $\det \mathrm{Index}\, \not{D}_{S^1}$ is the Möbius bundle, and this reflects the non-trivial torsion class $c_1 \in H^2(\mathcal{A}/\mathcal{G};\mathbb{Z}) = \mathbb{Z}_2$ of the determinant bundle on $\mathcal{A}/\mathcal{G}$; which in turn is the sign change of the Pfaffian under the large gauge transformation. With p doublets the bundle is the p-th tensor power. The Möbius transition function is -1, and $(-1)^p = 1$ iff p is even: the tensor square of a Möbius band is a cylinder. Hence

$$\left(\det \mathrm{Index}\, \not{D}_{S^1}\right)^{\otimes p} \text{ trivial} \iff p \text{ even}.$$

Reading Nash (pp. 283–285, (10.44)–(10.49))

Nash's "$\sqrt{\det(\not{D}_A^* \not{D}_A)}$ changes sign" should be read as: the Pfaffian (10.205), a canonically defined square root up to one overall sign fixed by continuity, changes sign. His "$\not{D}_A$ can be taken to be a real skew-adjoint operator" refers to the real structure $\mathcal{J}$ built from the pseudo-reality of the doublet and of the chiral spinors. The five-dimensional operator and its mod-2 index are Witten's; Nash postpones spectral flow to §6 and adiabatic limits to the gravitational discussion. In (10.49) his $F_2(\mathcal{A}/\mathcal{G}) = 0$ is justified by $H^2_{\mathrm{dR}} = 0$, not by the displayed "$H^2_{\text{de Rham}}(\mathcal{A}/\mathcal{G}) = 0$" alone, but the two statements are the same.

Recap: Witten's SU(2) anomaly

- Global anomaly: invariance under the identity component of $\mathcal{G}$ but not under the other components; invisible to derivatives, to curvature and to de Rham cohomology.
- Toy model: one fermion in $0+1$ dimensions, $Z = 2\cos(a/2) \mapsto -Z$ under $a \mapsto a + 2\pi$; one eigenvalue crosses zero.
- $\pi_0(\mathcal{G}) = \pi_4(\mathrm{SU}(2)) = \mathbb{Z}_2$ (framed circles, the belt trick).
- The doublet is pseudo-real, the Weyl operator is real, the fermion integral is a Pfaffian with a sign fixed by continuity.
- Sign change $= (-1)^{\text{spectral flow}} = (-1)^{\mathrm{index}_2 \not{D}_5}$; it is -1 for one doublet. With p doublets: consistent iff p even. The Standard Model has $p = 4$ per generation.

- Topology: $\pi_1(\mathcal{A}/\mathcal{G}) = \mathbb{Z}_2$, $H^2(\mathcal{A}/\mathcal{G};\mathbb{Z}) \supseteq \mathbb{Z}_2$ by the universal coefficient theorem; the real determinant bundle over the loop is a Möbius band.

10.9.7 Global gravitational anomalies: large diffeomorphisms

Everything we just did for gauge transformations has a gravitational twin. In §3 the symmetry group was $\mathrm{Diff}^+(M)$, the orientation-preserving diffeomorphisms of space-time, acting on the space $\mathrm{Met}(M)$ of metrics. The local (perturbative) gravitational anomaly was a class in H^1_{dR} of (the identity component of) this group. The question now is the one we asked for SU(2): *is the fermion determinant invariant under diffeomorphisms that cannot be continuously deformed to the identity?* Such diffeomorphisms are called **large diffeomorphisms**, and they exist exactly when $\pi_0(\mathrm{Diff}^+(M)) \neq 0$. Nash (following Witten) takes the case $M = S^n$.

A large diffeomorphism you can see: the Dehn twist. Before spheres, the simplest example. On the torus $T^2 = \mathbb{R}^2/\mathbb{Z}^2$, the linear map

$$\tau = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} : \ (x, y) \mapsto (x + y, \ y)$$

maps the lattice $\mathbb{Z}^2$ to itself, so it descends to a diffeomorphism of T^2. Its effect on the two basic loops a (the x-circle) and b (the y-circle) is $a \mapsto a$, $b \mapsto a + b$ (Figure 10.37). A diffeomorphism connected to the identity acts trivially on $H_1(T^2;\mathbb{Z}) = \mathbb{Z}^2$, because homotopic maps induce the same map on homology. τ acts by a non-identity matrix, so it is not connected to the identity. In fact $\pi_0(\mathrm{Diff}^+(T^2)) = \mathrm{SL}(2,\mathbb{Z})$, the **modular group** of the torus, whose role in the string partition function was the subject of Example 10.44 (invariance under $\tau \mapsto \tau + 1$ and $\tau \mapsto -1/\tau$). There, modular invariance of the torus amplitude was a global gravitational anomaly-freedom condition in disguise.

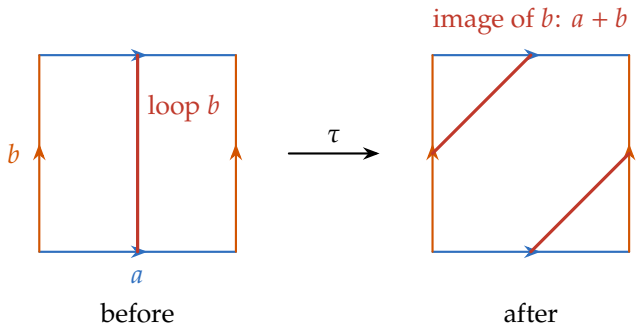


Figure 10.37: The Dehn twist $(x, y) \mapsto (x + y, y)$ on the torus (opposite sides of the square are identified). The vertical loop b (red, left) is carried to a loop that winds once around a as it goes once around b (red, right). Because it changes the homology class of a loop, the twist cannot be deformed to the identity: it is a large diffeomorphism, an element of $\pi_0(\mathrm{Diff}^+(T^2)) = \mathrm{SL}(2,\mathbb{Z})$.

For spheres there are no loops to twist ($H_1(S^n) = 0$ for $n \geq 2$), and it is much less obvious that large diffeomorphisms exist at all. They do, and they are intimately tied to exotic spheres.

10.9.8 Large diffeomorphisms of spheres and exotic spheres

Recall from Chapter 1 that an **exotic sphere** is a smooth manifold that is homeomorphic to S^n but not diffeomorphic to it: the same topological space, a different notion of which functions are smooth. Milnor found the first ones in $n = 7$.

The clutching construction. Here is how to build a candidate n-sphere from a diffeomorphism. Cut S^n along its equator S^{n-1} into the northern and southern hemispheres H^+ and H^-, each a closed n-ball with boundary S^{n-1}. Pick $f \in \mathrm{Diff}^+(S^{n-1})$. Glue H^+ back to H^- by identifying $x \in \partial H^-$ with $f(x) \in \partial H^+$:

$$S^n_f := H^- \cup_f H^+, \qquad \partial H^- \ni x \;\sim\; f(x) \in \partial H^+. \tag{10.212}$$

The result is always homeomorphic to S^n: a homeomorphism of S^{n-1} extends to a homeomorphism of the ball by "coning" ($rx \mapsto rf(x)$ for $0 \le r \le 1$), so we can undo the twist topologically. But the cone of a diffeomorphism is not smooth at the centre, $r = 0$. So the smooth structure of S^n_f can be exotic.

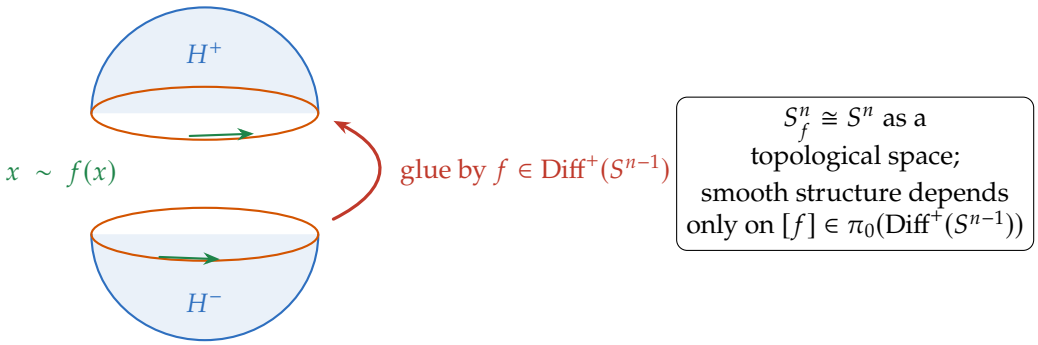


Figure 10.38: Building an n-sphere from two balls glued along their boundary spheres (orange) with a twist f. Topologically the twist can always be undone, so the result is a sphere. Smoothly it can be undone only if f extends to a diffeomorphism of the ball; if not, S^n_f is an exotic sphere. Deforming f within its component does not change S^n_f.

Key idea: Nash p. 285–286; Cerf

Two such spheres S^n_f and S^n_g are diffeomorphic precisely when f and g lie in the same component of $\mathrm{Diff}^+(S^{n-1})$. The components of $\mathrm{Diff}^+(S^{n-1})$ correspond exactly to the smooth structures on the n-sphere (for $n \ge 6$ or so, which is all we need). In symbols, with Θ_n the group of exotic n-spheres (under connected sum),

$$\pi_0\big(\mathrm{Diff}^+(S^{n-1})\big) \;\cong\; \Theta_n.$$

Why is "deforming f does not change the result" true? If f_s is a smooth path from f_0 to f_1, glue a collar $S^{n-1} \times [0,1]$ into the seam, using f_s on the slice $S^{n-1} \times \{s\}$; this gives a diffeomorphism $S^n_{f_0} \cong S^n_{f_1}$. The converse direction is Cerf's hard theorem ("pseudo-isotopy implies isotopy").

Worked example 10.51: The numbers: Kervaire–Milnor

Kervaire and Milnor computed Θ_n. The first few are

n	≤ 6	7	8	9	10	11	12	15
Θ_n	0	$\mathbb{Z}_{28}$	$\mathbb{Z}_2$	$\mathbb{Z}_2^3$	$\mathbb{Z}_6$	$\mathbb{Z}_{992}$	0	$\mathbb{Z}_{16256} \oplus \mathbb{Z}_2$

(for $n = 4$ the smooth question is open; "0" there means only that no exotic S^4 is known). Reading the table through Cerf's theorem:

- $\Theta_7 = \mathbb{Z}_{28}$, so $\pi_0(\mathrm{Diff}^+(S^6)) = \mathbb{Z}_{28}$: the first sphere with large diffeomorphisms, as Nash says, is S^6;

- $\Theta_{11} = \mathbb{Z}_{992}$, so $\pi_0(\mathrm{Diff}^+(S^{10})) = \mathbb{Z}_{992}$. Of these 992 smooth structures one is the standard sphere and 991 are exotic. This is Nash's (10.50).

The number 992 has a formula: it is the order of the subgroup bP_{12} of spheres that bound parallelisable 12-manifolds, $|bP_{4k}| = 2^{2k-2}(2^{2k-1}-1)\cdot \mathrm{num}(4B_{2k}/k)$, which for $k=3$ gives $2^4\cdot 31\cdot 2 = 992$ (with $B_6 = \frac{1}{42}$, $4B_6/3 = \frac{2}{63}$). That Bernoulli numbers appear is a first hint that these numbers are secretly $\hat{A}$-genus computations, which is what the anomaly calculation will be.

10.9.9 Superstrings in ten dimensions: what needs checking

Why should a physicist care about $\pi_0(\mathrm{Diff}^+(S^{10}))$? Because ten is the critical dimension of the superstring (the superstring analogue of 26; Section 10.8.7). Green and Schwarz discovered in 1984 that the $N=1$ supersymmetric string theories in $d=10$ are free of *local* gauge, gravitational and mixed anomalies provided the gauge group is one of the two 496-dimensional groups

$$E_8\times E_8 \quad (\dim = 248+248 = 496) \qquad \text{or} \qquad \mathrm{SO}(32) \quad \Big(\dim = \tfrac{32\cdot 31}{2} = 496\Big).$$

First, local gravitational anomalies really are possible here: §3 showed they need $n = 4k-2$, and $10 = 4\cdot 3 - 2$. Local cancellation leaves the *global* question open. Witten answered it; Nash describes how.

Global gauge anomalies are absent. By (10.202), on $M = S^{10}$ the gauge group has $\pi_0(\mathcal{G}) = \pi_{10}(G)$. For E_8, the homotopy groups vanish in degrees $1 \le i \le 14$ except $\pi_3 = \mathbb{Z}$, so $\pi_{10}(E_8) = 0$. For SO(32) we are in the stable range ($10 < 31$), and Bott periodicity gives $\pi_{10}(\mathrm{SO}) = \pi_2(\mathrm{SO}) = 0$ (the real Bott sequence $\pi_i(\mathrm{O}) = \mathbb{Z}_2, \mathbb{Z}_2, 0, \mathbb{Z}, 0, 0, 0, \mathbb{Z}$ for $i \equiv 0,\dots,7 \bmod 8$). So

$$\pi_0(\mathcal{G}) = \pi_{10}(G) = 0 \qquad \text{for } G = E_8\times E_8 \text{ or } \mathrm{SO}(32):$$

every gauge transformation is small, and there is nothing global to check. Only the gravitational case remains.

The topology of the gravitational parameter space. The parameter space is

$$Y = \mathrm{Met}(S^{10})/\mathrm{Diff}^+(S^{10}).$$

The group acts on the contractible space $\mathrm{Met}(S^{10})$ (a convex cone in a vector space: $(1-t)\rho_0 + t\rho_1$ is a metric if ρ_0, ρ_1 are), so we have the fibre homotopy exact sequence (Nash (10.52))

$$\cdots \to \pi_i\big(\mathrm{Diff}^+\big) \to \pi_i\big(\mathrm{Met}\big) \to \pi_i(Y) \to \pi_{i-1}\big(\mathrm{Diff}^+\big) \to \pi_{i-1}\big(\mathrm{Met}\big) \to \cdots$$

(here we pretend, as Nash does, that the action is free; the proper treatment uses the based diffeomorphisms of §3 and changes nothing below).

$$\pi_1(Y) \overset{(1)}{=} \pi_0\big(\mathrm{Diff}^+(S^{10})\big) \overset{(2)}{=} \mathbb{Z}_{992},$$
$$H_1(Y;\mathbb{Z}) \overset{(3)}{=} \mathbb{Z}_{992}, \qquad H^2(Y;\mathbb{Z}) \overset{(4)}{=} F_2 \oplus T_1 \overset{(5)}{=} \mathbb{Z}_{992}. \tag{10.213}$$

(1) Exactness, with $\pi_1(\mathrm{Met}) = \pi_0(\mathrm{Met}) = 0$ because Met is contractible.
(2) Cerf and Kervaire–Milnor, Nash (10.50).
(3) Hurewicz: H_1 is the abelianisation of π_1.

(4) The universal coefficient theorem (10.210).
(5) $F_2 = 0$ because there is no free part in H_2 (the local anomaly has cancelled, $H^2_{\mathrm{dR}}(Y) = 0$), and $T_1 = \mathbb{Z}_{992}$ by (3). This is Nash's (10.51).

Hence (Nash (10.53)) the determinant line bundle of the gravitational Dirac family has a first Chern class that is *pure torsion*:

$$c_1\big(\det \operatorname{Index} \not{D}_{\mathrm{grav}}\big) \in H^2\big(\mathrm{Met}(S^{10})/\mathrm{Diff}^+(S^{10}); \mathbb{Z}\big) = \mathbb{Z}_{992}.$$

It could be any of 992 values, and we must compute which. De Rham methods, curvature, and the families index density are all blind to it.

10.9.10 Witten's construction: the mapping torus

Let ρ be a metric on S^{10} and f a large diffeomorphism. The question is how $\det(\not{D}^*_\rho \not{D}_\rho)$ (really its square root, the chiral determinant) changes under $\rho \mapsto f \cdot \rho$, the pulled-back metric. We interpolate (Nash (10.54)):

$$\rho^t = (1-t)\rho + t\, f\cdot\rho, \qquad t \in [0,1]. \tag{10.214}$$

Each ρ^t is a metric (a convex combination of positive-definite forms is positive definite), $\rho^0 = \rho$ and $\rho^1 = f\cdot\rho$. In $\mathrm{Met}(S^{10})$ this is a path; in Y it is a closed loop, because its endpoints are related by a diffeomorphism. It is the gravitational twin of the path $A_t = (1-t)A + tA^g$ we used for SU(2).

One metric instead of a family. Witten's trick, which we already used in the SU(2) case, is to trade the one-parameter family of ten-dimensional metrics for a single metric on an eleven-dimensional manifold N, with line element (Nash (10.55))

$$ds^2 = dt^2 + \rho^t_{ij}(x)\, dx^i dx^j. \tag{10.215}$$

Which N? Start with the cylinder $S^{10}\times[0,1]$. Its boundary is $S^{10}\times\{0\} \cup S^{10}\times\{1\}$. Glue the two ends together, but with a twist by f:

$$N = S^{10}\times[0,1]\big/\big((x,1)\sim(f(x),0)\big). \tag{10.216}$$

The metric (10.215) is smooth across the seam precisely because $\rho^1 = f\cdot\rho = f^*\rho^0$: the metric at $t=1$ is the pull-back by the gluing map of the metric at $t=0$. N is a closed 11-manifold, a fibre bundle over the circle (the t direction) with fibre S^{10} and structure group $\mathrm{Diff}^+(S^{10})$. It is called the **mapping torus** of f. A bundle over S^1 is determined by its one transition function, up to deformation, so there are exactly 992 such N, one for each element of $\pi_0(\mathrm{Diff}^+(S^{10}))$.

N as a pull-back. Nash adds a second description. The families index theorem of Chapter 4 is formulated for a fibre bundle $Z \to Y$ with fibre M and structure group $\mathrm{Diff}^+(M)$ (the "universal" family: over the point $[\rho]\in Y$ sits M with the metric ρ). Given a loop $\alpha : S^1 \to Y$, the pull-back α^*Z is a bundle over S^1 with fibre M. Take $M = S^{10}$ and α the loop $t \mapsto [\rho^t]$: then $\alpha^*Z = N$. The mapping torus is the universal family restricted to our loop, exactly as the Möbius band of Figure 10.36 was the determinant bundle restricted to a loop.

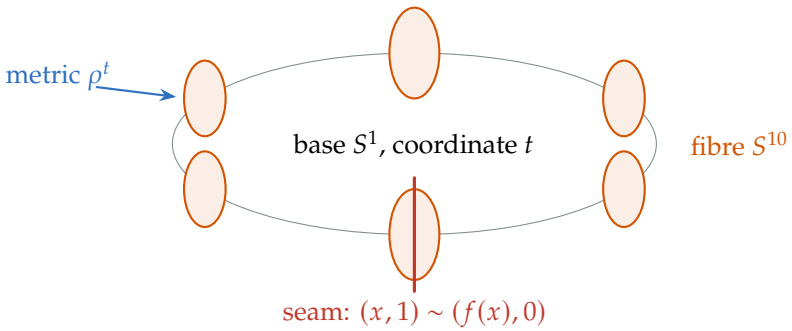


Figure 10.39: The mapping torus N of Nash (10.55). Over each point t of a circle sits a copy of S^{10} (orange) carrying the metric ρ^t. Going once around, the fibre is glued back to itself by the large diffeomorphism f (red seam). The whole family of ten-dimensional metrics has become one eleven-dimensional geometry, and f is now built into the topology of N.

10.9.11 Eta invariants: the phase of a determinant

On the odd-dimensional manifold N there is no chirality. The Weyl operator of ten dimensions is replaced by the full Dirac operator $\mathcal{D}$ (Nash's footnote 5), which is self-adjoint, so it has a real spectrum $\{\lambda_k\}$ that is *not* symmetric about zero. The spectral asymmetry is measured by the eta function.

Working definition 10.13 (eta invariant)

For a self-adjoint elliptic operator $\mathcal{D}$ [such as the Dirac operator on N: its eigenvalues are real, of both signs, and grow in size] with eigenvalues λ_k,

$$\eta_{\mathcal{D}}(s) \coloneqq \sum_{\lambda_k \neq 0} \frac{\operatorname{sgn} \lambda_k}{|\lambda_k|^s} = \sum_{\lambda_k > 0} \lambda_k^{-s} - \sum_{\lambda_k < 0} |\lambda_k|^{-s},$$

defined for $\Re s$ large and continued to $s = 0$ [there is exactly one function, analytic except at isolated poles, that equals the sum where the sum converges, and it has no pole at $s = 0$]. The number $\eta_{\mathcal{D}}(0)$ is "the number of positive eigenvalues minus the number of negative ones", made finite. Also let $h = \dim \operatorname{Ker} \mathcal{D}$. *How to compute it:* split the spectrum into its positive and negative parts, write each as a Hurwitz-type zeta function [a sum $\sum_{n \geq 0} (n + a)^{-s}$ over evenly spaced eigenvalues], and evaluate at $s = 0$. Note $\eta_{c\mathcal{D}}(0) = \eta_{\mathcal{D}}(0)$ for any $c > 0$: rescaling the operator (or the metric) does not change it.

Why the eta invariant is the phase. For a self-adjoint $\mathcal{D}$ with no zero modes, $\det \mathcal{D} = \prod_k \lambda_k = \prod_k |\lambda_k| \cdot \prod_{\lambda_k < 0} (-1)$. Writing $-1 = e^{i\pi}$ and regularising the count of negative eigenvalues with the same zeta-function recipe as the modulus:

$$\begin{aligned} \det \mathcal{D} &\overset{(1)}{=} \left|\det \mathcal{D}\right| \, \exp\!\left(i\pi N_-\right), \qquad N_- = \sum_{\lambda_k < 0} 1 \\ &\overset{(2)}{=} \left|\det \mathcal{D}\right| \, \exp\!\left(i\pi \lim_{s \to 0} \tfrac{1}{2}\Big[\sum_k |\lambda_k|^{-s} - \sum_k \operatorname{sgn}\lambda_k |\lambda_k|^{-s}\Big]\right) \\ &\overset{(3)}{=} \left|\det \mathcal{D}\right| \, \exp\!\left(\tfrac{i\pi}{2}\left[\zeta_{|\mathcal{D}|}(0) - \eta_{\mathcal{D}}(0)\right]\right). \end{aligned} \tag{10.217}$$

(1) Each negative eigenvalue contributes a factor $-1 = e^{i\pi}$ to the product (this choice of branch is a convention; see the pitfall below).

(2) For each k, $|\lambda_k|^{-s} - \operatorname{sgn}\lambda_k |\lambda_k|^{-s}$ is 0 if $\lambda_k > 0$ and $2|\lambda_k|^{-s}$ if $\lambda_k < 0$. So half the bracket counts negative eigenvalues, weighted by $|\lambda_k|^{-s}$, and the $s \to 0$ limit (by analytic continuation) is the regularised count.
(3) Definitions of $\zeta_{|\mathcal{D}|}(s) = \sum |\lambda_k|^{-s}$ and of $\eta_{\mathcal{D}}(s)$.

On a closed odd-dimensional manifold, $\zeta_{|\mathcal{D}|}(0)$ is given by a local heat-kernel coefficient, as in (10.181), and for Dirac operators in odd dimensions that coefficient vanishes, so $\zeta_{|\mathcal{D}|}(0) = -h$. The phase of the determinant is therefore $\exp(-\frac{i\pi}{2}(\eta + h))$: *the phase of a fermion determinant in odd dimensions is the eta invariant*.

Worked example 10.52: The eta invariant on a circle, computed

Take $\mathcal{D}_a = -i\frac{\mathrm{d}}{\mathrm{d}\theta} + a$ on periodic functions on S^1, i.e. the Dirac operator of a charged particle on a circle in the constant gauge field a. Eigenfunctions $e^{in\theta}$ give eigenvalues $\lambda_n = n + a$, $n \in \mathbb{Z}$. Let $0 < a < 1$. The positive eigenvalues are $n + a$ ($n \geq 0$), the negative ones are $-(m + 1 - a)$ ($m \geq 0$). With the Hurwitz zeta function $\zeta_H(s, a) = \sum_{n\geq 0}(n + a)^{-s}$:

$$\begin{aligned}\eta_{\mathcal{D}_a}(0) &\overset{(1)}{=} \zeta_H(0, a) - \zeta_H(0, 1 - a)\\ &\overset{(2)}{=} \left(\tfrac{1}{2} - a\right) - \left(\tfrac{1}{2} - (1 - a)\right)\\ &= 1 - 2a.\end{aligned}$$

(1) Split the sum defining η into the positive and the negative eigenvalues.
(2) $\zeta_H(0, b) = \frac{1}{2} - b$, proved just below.

Proof of $\zeta_H(0, b) = \frac{1}{2} - b$. This is the same heat-kernel argument as (10.182). The "heat trace" of the spectrum $\{n + b\}_{n\geq 0}$ is a geometric series:

$$\sum_{n\geq 0} e^{-t(n+b)} \overset{(1)}{=} \frac{e^{-tb}}{1 - e^{-t}} \overset{(2)}{=} \frac{1 - tb + O(t^2)}{t\left(1 - \frac{t}{2} + O(t^2)\right)} \overset{(3)}{=} \frac{1}{t} + \left(\frac{1}{2} - b\right) + O(t),$$

$$\zeta_H(0, b) \overset{(4)}{=} \text{the } t^0 \text{ coefficient} = \tfrac{1}{2} - b.$$

(1) Sum of a geometric series with ratio e^{-t}.
(2) Expand e^{-tb} and $1 - e^{-t} = t - \frac{t^2}{2} + \cdots$.
(3) $\frac{1}{1-t/2} = 1 + \frac{t}{2} + \cdots$, and multiply out.
(4) $\Gamma(s)\zeta_H(s, b) = \int_0^\infty t^{s-1} \sum_n e^{-t(n+b)}\mathrm{d}t$; the $1/t$ term produces a pole at $s = 1$, the constant term c produces c/s near $s = 0$, and $\Gamma(s) \approx 1/s$, so $\zeta_H(0, b) = c$.

For the two endpoints: at $a = 0$ the spectrum $\mathbb{Z}$ is symmetric, $\eta = 0$, and $h = 1$ (the constant function). The combination $\xi(a) := \frac{1}{2}(\eta + h)$ is

$$\xi(a) = \tfrac{1}{2} - a \quad (0 < a < 1), \qquad \xi(0) = \tfrac{1}{2}, \qquad \xi(a + 1) = \xi(a) \text{ (same spectrum).}$$

So ξ decreases smoothly with slope -1 and jumps up by 1 at every integer, exactly when an eigenvalue $n + a$ passes from negative, through zero, to positive (Figure 10.40).

This one example contains the whole anatomy of the eta invariant:

1. its *variation* along a smooth family, away from crossings, is local, an integral of a curvature density (here $\mathrm{d}\xi/\mathrm{d}a = -1$);

2. it *jumps by integers* (by 2 for η, by 1 for ξ) when eigenvalues cross zero;

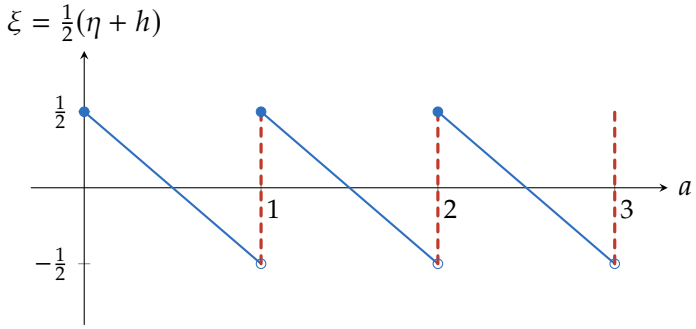


Figure 10.40: The reduced eta invariant $\xi = \frac{1}{2}(\eta + h)$ of $-i\,\mathrm{d}/\mathrm{d}\theta + a$ on a circle. Between integers it changes smoothly (slope -1): that part is *local*, the integral of the field strength. At each integer one eigenvalue crosses zero and ξ jumps by exactly 1 (red dashes): that part counts spectral flow. So $\exp(2\pi i\xi)$ is smooth, while $\exp(i\pi\xi)$ flips sign at every crossing.

3. going round a closed loop of operators, the smooth part changes by an amount that is minus the net spectral flow (here -1 for one circuit $a \to a+1$).

Worked example 10.53: The phase of a one-dimensional determinant

Put Example 10.52 into (10.217). For $0 < a < 1$, $\zeta_{|\mathcal{D}|}(0) = \zeta_H(0,a) + \zeta_H(0,1-a) = 0$ and $\eta = 1-2a$. For the modulus, Lerch's formula $\prod_{n\geq 0}(n+b) = \sqrt{2\pi}/\Gamma(b)$ (the zeta-regularised product) gives

$$\begin{aligned}\det \mathcal{D}_a &\overset{(1)}{=} \frac{\sqrt{2\pi}}{\Gamma(a)} \cdot \frac{\sqrt{2\pi}}{\Gamma(1-a)} \cdot e^{-\frac{i\pi}{2}(1-2a)} \\ &\overset{(2)}{=} 2\sin(\pi a)\, e^{i\pi(a-\frac{1}{2})} \\ &\overset{(3)}{=} 1 - e^{2\pi i a}.\end{aligned}$$

(1) The modulus is the product over positive eigenvalues $n+a$ times the product over |negative| eigenvalues $m+1-a$; the phase is (10.217) with $\zeta_{|\mathcal{D}|}(0) = 0$.
(2) The reflection formula $\Gamma(a)\Gamma(1-a) = \pi/\sin(\pi a)$.
(3) $2\sin(\pi a) = -i(e^{i\pi a} - e^{-i\pi a})$ and $e^{-i\pi/2} = -i$; multiply out.

Look at the two factors in line (2). The real factor $2\sin(\pi a)$, continued smoothly past $a = 1$, changes sign under $a \mapsto a+1$: this is the analogue of the Pfaffian's sign flip, a global anomaly. The phase factor $e^{i\pi(a-1/2)} = -i\exp\left(\frac{i}{2}\oint A\right)$ (with $\oint A = 2\pi a$) also flips sign, so the product $1 - e^{2\pi i a}$ is periodic. The factor $\exp(\frac{i}{2}\oint A)$ is a one-dimensional Chern–Simons term, a *local* functional of the gauge field. The zeta-function regularisation has silently added it, trading the global anomaly for a term that is not invariant under $a \mapsto -a$. This is exactly the trade-off of the "Try it" after Example 10.45, now produced by the eta invariant, and it is the one-dimensional version of the **parity anomaly** of odd-dimensional fermions.

Watch out: The sign convention in the phase

Choosing $-1 = e^{-i\pi}$ instead of $e^{+i\pi}$ in step (1) of (10.217) flips the sign of the exponent. The two choices differ by $e^{i\pi(\zeta-\eta)}$, a local counterterm times an integer power of -1. Different papers use different conventions; what is convention independent is the change of the phase around a closed loop, modulo the local part.

10.9.12 The adiabatic limit and Witten's formula

Now back to the eleven-manifold N. We want the change of the ten-dimensional chiral determinant as ρ^t runs from ρ to $f\cdot\rho$. The idea is that when t is a very slow variable, a spinor on N is approximately (a function of t)×(an eigenspinor of the ten-dimensional operator at time t), and the eleven-dimensional Dirac operator becomes ∂_t plus the ten-dimensional one, as in the SU(2) case, $\mathcal{D}_5=\partial_t+\not{D}_{A_t}$.

Working definition 10.14 (adiabatic limit)

Replace the family ρ^t on S^{10} by $\varepsilon\rho^t$ and let $\varepsilon\to 0$ (Nash p. 288). Equivalently, since $ds^2=dt^2+\varepsilon\rho^t dx^2=\varepsilon\,(dt^2/\varepsilon+\rho^t dx^2)$ and η does not see an overall scale, stretch the circle: $ds^2=dt^2/\varepsilon+\rho^t dx^2$ with $\varepsilon\to 0$ (Nash (10.63)). The circle has length $1/\sqrt{\varepsilon}\to\infty$, so the metric varies slowly along it. Then compute $\eta_{\mathcal{D}}(0)$ [the number of positive minus the number of negative eigenvalues of $\mathcal{D}$, made finite] and h [how many independent solutions $\psi\neq 0$ of $\mathcal{D}\psi=0$ there are, $\dim\operatorname{Ker}\mathcal{D}$] of the Dirac operator on N by expanding in powers of ε.

Key idea: Witten's formula, Nash (10.56)–(10.58)

Under the large diffeomorphism $\rho\mapsto f\cdot\rho$,

$$\det\bigl(\not{D}^*_\rho\not{D}_\rho\bigr)\longmapsto\det\bigl(\not{D}^*_{f\cdot\rho}\not{D}_{f\cdot\rho}\bigr)=C\,\det\bigl(\not{D}^*_\rho\not{D}_\rho\bigr),\qquad C=\exp\bigl[i\pi\,\eta_{\mathcal{D}}(0)\bigr],$$

with $\eta_{\mathcal{D}}$ the eta invariant of the Dirac operator on N (in the adiabatic limit). The path integral contains the square root, so the fermion effective action $S_\psi=\ln\sqrt{\det(\not{D}^*_\rho\not{D}_\rho)}$ changes by $\Delta S_\psi=\frac{i\pi}{2}\eta_{\mathcal{D}}(0)$, and

$$\begin{aligned}\text{no global anomaly}\ &\Longleftrightarrow\ \sqrt{C}=\exp\Bigl[\frac{i\pi}{2}\eta_{\mathcal{D}}(0)\Bigr]=1\\ &\Longleftrightarrow\ \eta_{\mathcal{D}}(0)=4k\ \Longleftrightarrow\ \Delta S_\psi=2\pi i k,\quad k\in\mathbb{Z}.\end{aligned}$$

Why is this plausible? The ten-dimensional chiral determinant, followed continuously along the loop ρ^t, has a modulus that returns to itself (it is diffeomorphism invariant) and a phase that may not. By (10.217), the phase of the eleven-dimensional determinant is $e^{\mp i\pi\eta/2}$, and in the adiabatic limit the eleven-dimensional determinant on N "is" the product over the loop of the ten-dimensional ones, with the end glued to the beginning by f. So the total phase accumulated going round the loop is $e^{\mp i\pi\eta/2}$. In Example 10.53 we saw exactly this in one dimension lower: the phase of the determinant along $a\in[0,1]$ changes by $e^{i\pi}$, which is $e^{-i\pi\Delta\eta/2}$ for the smooth change $\Delta\eta=-2$.

But $\eta(0)$ is not computable from the spectrum. Nobody can diagonalise the Dirac operator on an eleven-dimensional mapping torus. The way out is that eta invariants appear in index theorems for manifolds with boundary.

10.9.13 The Atiyah–Patodi–Singer theorem computes η

Recall from Chapter 4 the index theorem for a manifold with boundary. Let X be a compact 12-dimensional spin manifold with $\partial X=N$, with a metric that is a product $dt^2+\bar{\rho}$ near the boundary. Impose the Atiyah–Patodi–Singer (global, spectral) boundary condition, which projects out the non-negative modes of the boundary operator. Then the chiral Dirac operator

$\not{D}_X$ is Fredholm and (Nash (10.59))

$$\operatorname{index} \not{D}_X = \int_X \hat{A}(X) - \frac{h + \eta_{\mathcal{D}}(0)}{2}, \tag{10.218}$$

where $\mathcal{D}$ is the Dirac operator on N and $h = \dim \operatorname{Ker} \mathcal{D}$. The bulk term is the familiar local index density; the boundary correction $\xi = \frac{1}{2}(h + \eta)$ is exactly the combination whose jumps we watched in Figure 10.40. The formula is consistent: $\int_X \hat{A}$ changes smoothly under deformations of the metric, ξ changes smoothly except for integer jumps, and the index changes by the same integers.

Physical picture: Why a boundary term must be non-local

If X were closed, $\operatorname{index} \not{D}_X = \int_X \hat{A}$ would be an integer. With a boundary, $\int_X \hat{A}$ is some real number that depends on the metric near ∂X. The index is still an integer. Something must make up the fractional difference, and it cannot be a local boundary integral (local boundary integrals can be changed continuously while keeping the bulk fixed, and a continuous function cannot compensate the jump of an integer). The spectral asymmetry η of the boundary operator is the global object that does the job.

Using APS to compute the anomaly. Suppose N bounds a 12-manifold X. Then (10.218) gives $\frac{1}{2}\eta_{\mathcal{D}}(0) = \int_X \hat{A} - \operatorname{index} \not{D}_X - \frac{h}{2}$, and so (Nash (10.60))

$$\Delta S_\psi \overset{(1)}{=} \frac{i\pi}{2}\,\eta_{\mathcal{D}}(0) \overset{(2)}{=} i\pi\Big\{\int_X \hat{A} - \operatorname{index} \not{D}_X - \frac{h}{2}\Big\}. \tag{10.219}$$

(1) Witten's formula.
(2) Solve (10.218) for $\eta/2$ and multiply by $i\pi$.

In twelve dimensions the index of the Dirac operator is always *even*. The reason is the pseudo-reality we met for SU(2): in dimensions $8k + 4$ the chiral spinor representations are quaternionic, so the kernel and cokernel of $\not{D}_X$ carry an anti-linear J with $J^2 = -1$ and their complex dimensions are even (Kramers). Hence $i\pi \operatorname{index} \not{D}_X \in 2\pi i\mathbb{Z}$ and drops out of $e^{\Delta S}$. What remains is a *local* bulk integral (plus the harmless h): the global anomaly has been computed from characteristic classes on a twelve-manifold, one dimension above N, two above space-time.

Watch out: What if N bounds no X?

The argument needs $N = \partial X$. For the mapping tori of S^{10} this is fine (the relevant spin cobordism group vanishes, and in fact one can build X by filling in the fibre spheres). For a general space-time M this is a genuine hypothesis, and the modern theory of anomalies (Dai–Freed, Witten) treats the eta invariant itself as the fundamental object.

10.9.14 All the fields of the superstring

A superstring theory has more than spin-$\frac{1}{2}$ fields, and all of them must be included, both to cancel the local anomalies (Green–Schwarz) and to cancel the global one. Nash lists three kinds of chiral field in ten dimensions:

- spin-$\frac{1}{2}$ Weyl fermions, with operator the Dirac operator $\mathcal{D}$, number n_d (counted with sign: left minus right);

- spin-$\frac{3}{2}$ Rarita–Schwinger fields (gravitini), number n_{rs}. Their operator acts on vector-spinors ψ_μ. A vector-spinor contains a spin-$\frac{1}{2}$ part that is a gauge artefact (removed by the ghosts of local supersymmetry), so the physical contribution is $\eta_{rs} - \eta_{\mathcal{D}}$;
- self-dual antisymmetric tensor fields, number n_{sd}, with an operator of signature type; being "half" of an ordinary field (only the self-dual half propagates), they come with an extra factor $\frac{1}{2}$.

The total phase is (Nash (10.61))

$$\Delta S = \frac{\pi i}{2}\Big\{ n_d\,\eta_{\mathcal{D}}(0) + n_{rs}\big(\eta_{rs}(0) - \eta_{\mathcal{D}}(0)\big) - \frac{n_{sd}}{2}\,\eta_{sd}(0)\Big\}. \tag{10.220}$$

Worked example 10.54: Counting fields in the $E_8 \times E_8$ and $\mathrm{SO}(32)$ theories

The $N = 1$ supergravity multiplet coupled to super Yang–Mills in ten dimensions contains:

field	chirality	number
gravitino ψ_μ	left	1
dilatino λ	right	1
gauginos χ^a (adjoint of G)	left	$\dim G = 496$
self-dual tensor	–	0 (the 2-form B is not self-dual)

Hence $n_{rs} = 1$, $n_{sd} = 0$, and the net number of left-handed spin-$\frac{1}{2}$ fields is $n_d = 496 - 1 = 495$, which is Nash's value. Inserting into (10.220) gives the fermionic phase. Then one must add the change of the remaining term in the action, the Green–Schwarz counterterm $\int B \wedge X_8(F,R)$ that was introduced to cancel the local anomaly (the 12-form anomaly polynomial factorises as $X_4 \wedge X_8$ for exactly these two groups). With it, Witten found that the complete change is $\Delta S = 2\pi i k$: the theories have no global gravitational anomaly on S^{10}. The same calculation for type IIB supergravity, which has $n_{rs} = 2$, $n_d = -2$ and $n_{sd} = 1$, also gives $2\pi i k$.

10.9.15 Global anomalies are holonomies: Bismut and Freed

There is a geometrical statement behind (10.56). It completes the picture of the whole chapter: *local anomaly = curvature, global anomaly = holonomy* of a natural connection on the determinant line bundle.

The connection. The determinant line bundle $\det \operatorname{Index} \not{D}$ over Y carries the Quillen metric of Section 10.8.9 (zeta-regularised norm of the section $\det \not{D}$), and Bismut and Freed constructed a compatible unitary connection ∇^{BF}. Its connection 1-form, pulled back to the space of fields, is the variation of the effective action, $\mathrm{d} \ln \det$, i.e. the anomaly 1-form of §2. Two facts:

1. The *curvature* of ∇^{BF} is the degree-two component of the families index density $\int_M \hat{A}\,\mathrm{ch}$. It is the de Rham representative of $c_1(\det \operatorname{Index} \not{D})$: the local anomaly.
2. The *holonomy* of ∇^{BF} around a loop $\alpha : S^1 \to Y$ is (Nash (10.62)) $\exp\big[\int_{S^1} \alpha^* A\big]$, where A is the connection 1-form. When the curvature vanishes it depends only on the homotopy class of α: it is the global anomaly.

Key idea: Bismut–Freed holonomy formula, Nash (10.64)

Let $\alpha : S^1 \to Y$ be a loop, $N = \alpha^* Z$ the corresponding mapping torus, and give it the metric $ds^2 = dt^2/\varepsilon + \rho^t dx^2$ (Nash (10.63); "blowing up the metric" of the base circle). Then

$$\mathrm{hol}_\alpha\big(\nabla^{\mathrm{BF}}\big) = \exp\Big[\int_{S^1} \alpha^* A\Big] = (-1)^{\mathrm{index}\, \not{D}_\rho} \lim_{\varepsilon\to 0} \exp\big[-\pi i\,(h_\varepsilon + \eta_\varepsilon(0))\big],$$

with h_ε and η_ε those of the Dirac operator on N for the metric with parameter ε, and index $\not{D}_\rho$ the index of the chiral Dirac operator on the fibre. (Nash writes the exponent with $+\pi i$; the two signs differ by the convention of the pitfall above.) The holonomy is the global anomaly.

Physical picture: The Aharonov–Bohm effect of the space of fields

A charged particle going around a solenoid feels no magnetic field (zero curvature) along its path, yet its wavefunction returns with the phase $e^{ie\Phi}$ (holonomy). The effective action of the fermions is the "wavefunction" of the space of fields Y, and the determinant line bundle is the bundle it lives in. When the local anomaly cancels, the Bismut–Freed connection is flat, like the vector potential outside the solenoid, but its holonomy around a non-contractible loop in Y (a large gauge transformation or diffeomorphism) can still be non-trivial. That is a global anomaly. The "solenoid" is the non-trivial $\pi_1(Y)$ (Figure 10.41).

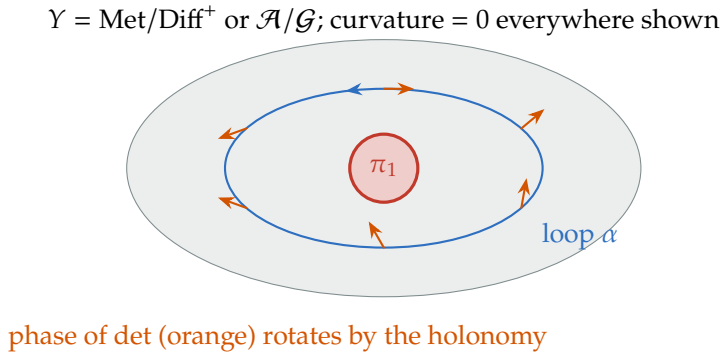


Figure 10.41: Local versus global anomaly as curvature versus holonomy. When the local anomaly cancels, the Bismut–Freed connection on the determinant line is flat on Y (grey), yet transporting the phase of the determinant (orange arrows) around a non-contractible loop α need not bring it back to itself, because the loop encircles a "hole" of Y (red), a non-trivial element of $\pi_1(Y)$. The mismatch is the global anomaly.

Torsion Chern classes are flat bundles with holonomy. This also explains the torsion classes of (10.211) and (10.213). A complex line bundle whose c_1 is torsion has a flat connection, and a flat line bundle is determined by its holonomy, a homomorphism $\pi_1(Y) \to \mathrm{U}(1)$. Such homomorphisms factor through $H_1(Y;\mathbb{Z})$.

Worked example 10.55: Flat line bundles when $\pi_1(Y) = \mathbb{Z}_N$

If $H_1(Y;\mathbb{Z}) = \mathbb{Z}_N$, a homomorphism $\chi : \mathbb{Z}_N \to \mathrm{U}(1)$ is fixed by $\chi(1) = \lambda$ with $\lambda^N = \chi(N\cdot 1) = \chi(0) = 1$. So the possible holonomies are the N-th roots of unity $\lambda = e^{2\pi i m/N}$, $m \in \mathbb{Z}_N$, and they match the elements of $T_1 = \mathbb{Z}_N \subseteq H^2(Y;\mathbb{Z})$ one for one.

- SU(2) on S^4: $N = 2$, the holonomy is ± 1; the real (Möbius) determinant bundle of

Figure 10.36 has holonomy -1 for one Weyl doublet, $(-1)^p$ for p doublets.

- Gravity on S^{10}: $N = 992$; a priori the holonomy of the determinant bundle could be any 992nd root of unity. Witten's computation shows that for the full superstring field content it is 1.

Beyond spheres and Dirac operators. Nash closes the discussion (p. 290) by noting that nothing was special about S^{10}: for any even-dimensional M and any first-order elliptic operator L on it, one builds the mapping torus N of a large diffeomorphism (and, if needed, a bounding X), blows up the metric of the circle, and finds that the global anomaly of L is the phase $\exp\left[\frac{\pi i}{2}\eta_{L_N}(0)\right]$, with L_N the operator on N induced by L.

Reading Nash (pp. 285–290, (10.50)–(10.64))

Nash's "991 exotic spheres in 11 dimensions" and "$\pi_0(\mathrm{Diff}^+(S^{10})) = \mathbb{Z}_{992}$" are the same fact (992 smooth structures, one standard). His (10.53) places the gravitational anomaly in $\mathbb{Z}_{992}$; (10.56) is Witten's adiabatic formula; (10.59) is the APS theorem (his Theorem 4.164 of Chapter 4), for a 12-manifold X with $\partial X = N$. In (10.61) the subtraction $\eta_{rs} - \eta_{\mathcal{D}}$ removes the gauge (ghost) spinor contained in ψ_μ, and $n_d = 495 = 496 - 1$ is gauginos minus dilatino. (10.62)–(10.64) are Bismut–Freed; the operator in the exponent is on the eleven-manifold N with the "blown-up" metric, and the sign in the exponent depends on orientation conventions.

10.9.16 Example: complex anomalies

Nash ends §5 with an example of a different kind of global effect. Anomalies of the global type force us to distinguish manifolds that are homeomorphic but not diffeomorphic (exotic spheres). One can go one step further and distinguish different *complex structures* on the same smooth manifold, as one must in string theory, where the worldsheet (and, in compactifications, space-time) is a complex manifold. The first examples where characteristic numbers tell complex structures apart occur in complex dimension five.

The manifold. Let (Nash (10.65))

$$M = \frac{\mathrm{U}(4)}{\mathrm{U}(2)\times\mathrm{U}(1)\times\mathrm{U}(1)} = \frac{G}{H}. \tag{10.221}$$

It is the space of flags $\{V_2 \subset V_3 \subset \mathbb{C}^4\}$ with $\dim V_k = k$ (equivalently, orthogonal decompositions $\mathbb{C}^4 = W_2 \oplus W_1 \oplus W_1'$ into subspaces of dimensions $2, 1, 1$). Its real dimension is

$$\dim_{\mathbb{R}} M = \dim \mathrm{U}(4) - \dim \mathrm{U}(2) - 2\dim \mathrm{U}(1) = 16 - 4 - 1 - 1 = 10,$$

so $\dim_{\mathbb{C}} M = 5$.

Complex structures from roots. The roots of U(4) are $e_i - e_j$ ($i \neq j$), where e_i picks out the i-th diagonal entry of the Cartan subalgebra; the root space of $e_i - e_j$ is spanned by the matrix unit E_{ij}. The roots of H are $\pm(e_1 - e_2)$ (the U(2) block on indices $1, 2$). The tangent space of M at the base point is spanned by the remaining root spaces, the $12 - 2 = 10$ roots $e_i - e_j$ with i, j in different blocks $\{1,2\}, \{3\}, \{4\}$. As in the treatment of G/T in Chapter 6, an invariant complex structure is fixed by declaring which half of these roots are "holomorphic" ($\bar\partial$ kills them): a choice of positive roots, i.e. of a Weyl chamber, or a lexicographic ordering. For a flag manifold

this amounts to *ordering the three blocks*: a root $e_i - e_j$ is positive if the block of i comes before the block of j. There are $3! = 6$ orderings. Reversing the order gives the complex conjugate structure, and swapping the two U(1) blocks is a symmetry of M, so up to equivalence there are two cases:

$$M_1 : \{1,2\} < \{3\} < \{4\} \quad \text{(U(2) block at the end)},$$
$$M_2 : \{3\} < \{1,2\} < \{4\} \quad \text{(U(2) block in the middle)}.$$

M_1 and M_2 are the same smooth manifold (hence certainly homeomorphic), with two different complex structures.

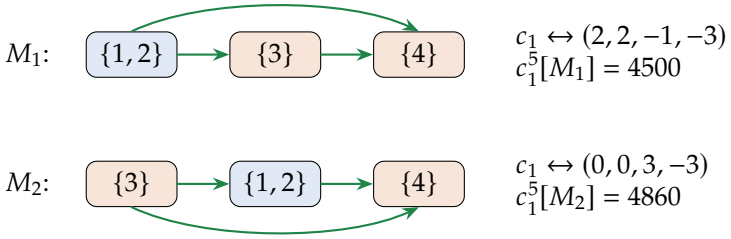


Figure 10.42: Two invariant complex structures on $\mathrm{U}(4)/(\mathrm{U}(2)\times\mathrm{U}(1)\times\mathrm{U}(1))$. An arrow from block B to block B' means the roots $e_i - e_j$ with $i \in B$, $j \in B'$ are holomorphic (positive); there are $2+2+1=5$ of them in each case, one for each complex dimension. Moving the U(2) block (blue) from the end to the middle changes which roots are holomorphic, and changes the Chern number c_1^5.

Computing c_1^5 by hand. Nash quotes (10.66), $c_1^5(M_1) = 4500$ and $c_1^5(M_2) = 4860$, citing his own paper. We can check it with a tool every physicist knows: counting states. Let $L = K^{-1}$ be the anticanonical line bundle (its c_1 is $c_1(M)$). Three facts:

1. (Riemann–Roch, large m) For an ample line bundle, $\dim H^0(M, L^m) = \int_M \mathrm{ch}(L^m)\,\mathrm{td}(M) = \frac{m^5}{5!}\,c_1(L)^5[M] + O(m^4)$, the higher cohomology vanishing (Kodaira).
2. (Borel–Weil) $H^0(G/H, L_\lambda)$ is the irreducible representation of U(4) with highest weight λ, for the positive system defined by the complex structure. For $L = K^{-1}$ the weight is the sum of the positive complementary roots.
3. (Weyl) Its dimension is $\prod_{\alpha>0} \frac{\langle \lambda+\rho,\alpha\rangle}{\langle \rho,\alpha\rangle}$, with ρ half the sum of all positive roots of U(4).

For $\lambda \to m\lambda$ with m large, the factor for the root $\alpha = e_1 - e_2$ of H is 1 (since $\langle \lambda, e_1 - e_2\rangle = 0$), and each of the five complementary roots contributes $m\langle\lambda,\alpha\rangle/\langle\rho,\alpha\rangle$. Comparing with fact 1:

$$c_1^5[M] = 5! \prod_{\alpha\in\Delta^+\setminus\Delta_H} \frac{\langle\lambda,\alpha\rangle}{\langle\rho,\alpha\rangle}. \tag{10.222}$$

Worked example 10.56: Warm-up: $\mathbb{CP}^1$ and $\mathbb{CP}^2$

$\mathbb{CP}^1 = \mathrm{U}(2)/(\mathrm{U}(1)\times\mathrm{U}(1))$: one complementary positive root $e_1 - e_2$, $\lambda = e_1 - e_2 \leftrightarrow (1,-1)$, $\rho = (\frac{1}{2}, -\frac{1}{2})$, so $c_1[\mathbb{CP}^1] = 1! \cdot \frac{2}{1} = 2$, the Euler characteristic of the sphere. ✓
$\mathbb{CP}^2 = \mathrm{U}(3)/(\mathrm{U}(1)\times\mathrm{U}(2))$, blocks $\{1\} < \{2,3\}$: complementary positive roots $e_1 - e_2$, $e_1 - e_3$; $\lambda = 2e_1 - e_2 - e_3 \leftrightarrow (2,-1,-1)$; $\rho = (1,0,-1)$. Then $\langle\lambda, e_1 - e_2\rangle = 3$, $\langle\rho, e_1 - e_2\rangle = 1$,

$\langle\lambda, e_1 - e_3\rangle = 3$, $\langle\rho, e_1 - e_3\rangle = 2$, and

$$c_1^2[\mathbb{CP}^2] = 2! \cdot \frac{3 \cdot 3}{1 \cdot 2} = 9,$$

the known value ($c_1 = 3H$, $H^2 = 1$). ✓

Worked example 10.57: $c_1^5(M_1) = 4500$ and $c_1^5(M_2) = 4860$

M_1: order $\{1,2\} < \{3\} < \{4\}$. The positive system of U(4) is the standard one, $e_i - e_j$ with $i < j$, so $\rho = (\frac{3}{2}, \frac{1}{2}, -\frac{1}{2}, -\frac{3}{2})$. The complementary positive roots are $e_1 - e_3, e_1 - e_4, e_2 - e_3, e_2 - e_4, e_3 - e_4$, whose sum is $\lambda = (2, 2, -1, -3)$. Tabulate:

α	$e_1 - e_3$	$e_1 - e_4$	$e_2 - e_3$	$e_2 - e_4$	$e_3 - e_4$
$\langle\lambda, \alpha\rangle$	3	5	3	5	2
$\langle\rho, \alpha\rangle$	2	3	1	2	1

$$c_1^5[M_1] \overset{(1)}{=} 120 \cdot \frac{3 \cdot 5 \cdot 3 \cdot 5 \cdot 2}{2 \cdot 3 \cdot 1 \cdot 2 \cdot 1} = 120 \cdot \frac{450}{12} = 4500.$$

(1) Formula (10.222) with $5! = 120$ and the tabulated pairings (the pairing is $\langle \sum_i c_i e_i, e_j - e_k\rangle = c_j - c_k$).

M_2: order $\{3\} < \{1,2\} < \{4\}$. Now the positive system orders the indices as $3, 1, 2, 4$, so ρ assigns $\frac{3}{2}, \frac{1}{2}, -\frac{1}{2}, -\frac{3}{2}$ to e_3, e_1, e_2, e_4. The complementary positive roots are $e_3 - e_1, e_3 - e_2, e_3 - e_4, e_1 - e_4, e_2 - e_4$, with sum $\lambda = 3e_3 - 3e_4 \leftrightarrow (0, 0, 3, -3)$.

α	$e_3 - e_1$	$e_3 - e_2$	$e_3 - e_4$	$e_1 - e_4$	$e_2 - e_4$
$\langle\lambda, \alpha\rangle$	3	3	6	3	3
$\langle\rho, \alpha\rangle$	1	2	3	2	1

$$c_1^5[M_2] \overset{(1)}{=} 120 \cdot \frac{3 \cdot 3 \cdot 6 \cdot 3 \cdot 3}{1 \cdot 2 \cdot 3 \cdot 2 \cdot 1} = 120 \cdot \frac{486}{12} = 4860.$$

(1) Formula (10.222) with $5! = 120$ and the tabulated pairings (the pairing is $\langle \sum_i c_i e_i, e_j - e_k\rangle = c_j - c_k$).

Both agree with Nash's (10.66), and they differ by $4860 - 4500 = 360$. (The same numbers come out of the Atiyah–Bott localisation formula, summing c_1^5/e over the 12 torus-fixed points of M.)

The physics: an action that sees the complex structure. Nash's toy model has action

$$S = \frac{i}{(2\pi)^5} \int_M (\operatorname{tr} F)^5,$$

with F the curvature of the (Hermitian) tangent bundle of the Kähler manifold M. Up to the conventional factor, $\frac{i}{2\pi} \operatorname{tr} F$ is the Chern–Weil representative of c_1, so S is (a constant times) the Chern number $c_1^5[M]$ of the complex structure in use. It is diffeomorphism invariant for a fixed complex structure: it is a topological number. Now let g and $\tilde{g}$ be the Kähler metrics adapted to the two complex structures (both invariant under U(4)). They are related by a smooth path g^t of metrics ($g^0 = g$, $g^1 = \tilde{g}$), but not by a holomorphic change of coordinates. Along the path

the complex structure jumps from one type to the other, and the action does not return to its value: it changes by

$$\Delta S = c_1^5[M_2] - c_1^5[M_1] = 360.$$

The example is not meant as a realistic physical model; it shows how an action built from characteristic classes of the complex structure can fail to be invariant under a deformation that is perfectly smooth from the real point of view, a potential "complex anomaly".

Recap: Global gravitational and complex anomalies

- Large diffeomorphisms: $\pi_0(\mathrm{Diff}^+(M)) \neq 0$. On T^2 this is the modular group $\mathrm{SL}(2,\mathbb{Z})$; on spheres it is the group of exotic spheres, $\pi_0(\mathrm{Diff}^+(S^{n-1})) \cong \Theta_n$ (clutching and Cerf).
- In the ten-dimensional superstring, $\pi_{10}(G) = 0$ kills global gauge anomalies; $\pi_0(\mathrm{Diff}^+(S^{10})) = \mathbb{Z}_{992}$ leaves a possible torsion class $c_1(\det \mathrm{Index}\, \not{D}_{\mathrm{grav}}) \in \mathbb{Z}_{992}$.
- Witten: use the interpolating metrics ρ^t to build the mapping torus N; in the adiabatic limit the phase change is $\exp(\frac{i\pi}{2}\eta_{\mathcal{D}}(0))$; no anomaly iff $\eta = 4k$.
- The eta invariant is the phase of an odd-dimensional determinant; on S^1, $\eta(-i\partial_\theta + a) = 1 - 2a$, and $\frac{1}{2}(\eta + h)$ jumps by the spectral flow.
- APS turns η into a local integral $\int_X \hat{A}$ on a bounding 12-manifold, the index being even; with all fields ($n_d = 495$, $n_{rs} = 1$) and the Green–Schwarz term, $\Delta S = 2\pi i k$.
- Bismut–Freed: local anomaly = curvature, global anomaly = holonomy of the determinant line bundle; torsion c_1 = flat bundle with root-of-unity holonomy.
- Complex anomalies: the flag manifold $\mathrm{U}(4)/(\mathrm{U}(2)\times\mathrm{U}(1)^2)$ has two complex structures with $c_1^5 = 4500$ and 4860; an action $\propto \int (\mathrm{tr}\, F)^5$ changes by 360.

10.10 Anomalies from a Hamiltonian perspective

Everything so far in this chapter has been Euclidean. We integrated over fermions in a path integral, obtained a determinant, and asked whether the determinant is gauge invariant. A physicist who learned quantum field theory from canonical quantisation will ask a fair question: *where is the anomaly in the Hamiltonian?* In canonical quantisation there is no determinant. There is a Hilbert space of states, a Hamiltonian, and a set of operators that generate gauge transformations. If gauge invariance is anomalous, something must go wrong with those operators.

The answer, which this section builds up step by step, is:

1. The one-particle Dirac Hamiltonians H_A form a family over the space of spatial gauge fields. Going round a loop of gauge fields, eigenvalues of H_A can cross zero: **spectral flow**.
2. The vacuum is the "filled Dirac sea" of negative-energy states. As A varies, the vacuum states form a *line bundle*, and this bundle has a curvature, exactly like the Berry phase of a spin in a slowly rotating magnetic field.
3. Because of the zero crossings, the Fock spaces are only well defined up to a phase. The gauge group then acts on states only *projectively*: $U(g)U(h) = \lambda\, U(gh)$ with a phase λ.

4. The infinitesimal version of that phase is an extra term in the commutator of gauge generators, the **Schwinger term**. With it, Gauss's law $G(\xi)|\text{phys}\rangle = 0$ has no solutions.

The topology that controls all this is the topology of spaces of Fredholm operators, and the bookkeeping rule that makes the Euclidean and the Hamiltonian pictures agree is: *only $n + p$ matters*, where n is the dimension of the manifold and p the number of parameters of the family.

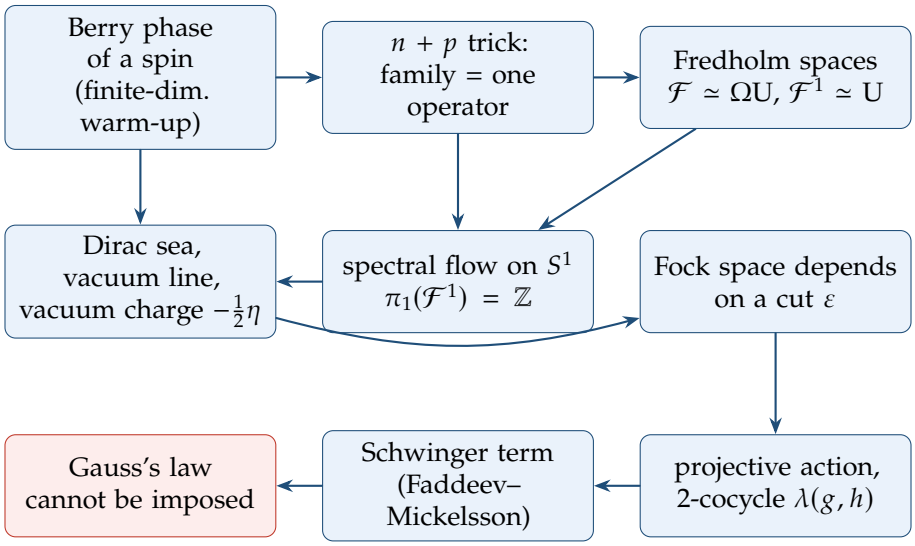


Figure 10.43: The logic of this section. Read along the arrows: a finite-dimensional warm-up (top left), the topology of families (top row), the concrete two-dimensional example (middle), and the second-quantised consequences (bottom row), ending at the physical statement in red: an anomalous gauge symmetry cannot be imposed as a constraint on states.

Reading Nash (p. 291, the opening of §6)

Nash starts §6 with the remark that "anomalies should not disappear in an alternative formulation, though they may manifest themselves in slightly different ways", and that "the central idea is still to use families of Dirac operators". The difference is which operators: in the Euclidean picture the chiral Dirac operator of the whole n-dimensional spacetime; in the Hamiltonian picture the self-adjoint Dirac Hamiltonian of $(n-1)$-dimensional space. The warm-up below (Berry's phase) is not in Nash, but it is the finite-dimensional model of every step in his §6.

10.10.1 Warm-up: the Berry phase of a spin in a magnetic field

Before any field theory, consider the simplest quantum system whose Hamiltonian depends on parameters: a spin-$\frac{1}{2}$ in a magnetic field. Everything we meet later (a family of Hamiltonians, a bundle of ground states, a curvature, a topological obstruction to choosing phases, and a degeneracy that is the source of it all) is already here in a 2×2 matrix.

The family. Let the field be $\vec{B} = B\,\vec{n}$ with $B > 0$ fixed and the unit vector

$$\vec{n} = (\sin\theta\cos\phi,\ \sin\theta\sin\phi,\ \cos\theta)$$

free. The parameter space is the sphere S^2 of field directions. Take

$$H(\vec{n}) = -B\,\vec{n}\cdot\vec{\sigma} = -B\begin{pmatrix} \cos\theta & \sin\theta\, e^{-i\phi} \\ \sin\theta\, e^{i\phi} & -\cos\theta \end{pmatrix},$$

with the Pauli matrices $\sigma_1 = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, $\sigma_2 = \left(\begin{smallmatrix} 0 & -i \\ i & 0 \end{smallmatrix}\right)$, $\sigma_3 = \left(\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right)$. Since $(\vec{n}\cdot\vec{\sigma})^2 = \mathbb{1}$, the eigenvalues are $\mp B$. The ground state (energy $-B$) is spin up along $\vec{n}$:

$$|\psi_N(\vec{n})\rangle = \begin{pmatrix} \cos\frac{\theta}{2} \\ e^{i\phi}\sin\frac{\theta}{2} \end{pmatrix}. \tag{10.223}$$

Check it: the first component of $(\vec{n}\cdot\vec{\sigma})|\psi_N\rangle$ is $\cos\theta\cos\frac{\theta}{2} + \sin\theta\sin\frac{\theta}{2} = \cos(\theta - \frac{\theta}{2}) = \cos\frac{\theta}{2}$, and the second is $e^{i\phi}(\sin\theta\cos\frac{\theta}{2} - \cos\theta\sin\frac{\theta}{2}) = e^{i\phi}\sin\frac{\theta}{2}$. So $\vec{n}\cdot\vec{\sigma}|\psi_N\rangle = |\psi_N\rangle$ and $H|\psi_N\rangle = -B|\psi_N\rangle$.

The vector (10.223) is smooth everywhere except at the south pole $\theta = \pi$, where it becomes $(0, e^{i\phi})$, which depends on the direction ϕ from which we approach the pole. Multiplying by the phase $e^{-i\phi}$ gives

$$|\psi_S(\vec{n})\rangle = e^{-i\phi}|\psi_N(\vec{n})\rangle = \begin{pmatrix} e^{-i\phi}\cos\frac{\theta}{2} \\ \sin\frac{\theta}{2} \end{pmatrix},$$

which is smooth at the south pole but not at the north pole. We will see that no choice of phase is smooth everywhere.

The adiabatic theorem produces a phase. Now let the direction move slowly along a closed loop $\vec{n}(t)$, $0 \le t \le T$, with $\vec{n}(T) = \vec{n}(0)$. If the motion is slow compared with the gap $2B$, the state stays in the instantaneous ground state up to a phase: $|\Psi(t)\rangle = c(t)\,e^{-i\int_0^t E\,\mathrm{d}t'}|\psi(\vec{n}(t))\rangle$ with $E = -B$ and $|\psi\rangle$ any smooth choice of phase along the loop. Put this into the Schrödinger equation ($\hbar = 1$):

$$\begin{aligned} i\partial_t|\Psi\rangle = H|\Psi\rangle &\overset{(1)}{=} i\dot{c}|\psi\rangle + Ec|\psi\rangle + ic\,\partial_t|\psi\rangle = Ec|\psi\rangle \quad \text{(times the common factor } e^{-i\int E}\text{)} \\ &\implies \dot{c} \overset{(2)}{=} -c\,\langle\psi|\partial_t\psi\rangle \\ &\implies c(T) \overset{(3)}{=} \exp\Big(-\oint\langle\psi|\mathrm{d}\psi\rangle\Big) \overset{(4)}{=} \exp\Big(i\oint\mathcal{A}\Big), \qquad \mathcal{A} \coloneqq i\langle\psi|\mathrm{d}\psi\rangle. \end{aligned} \tag{10.224}$$

(1) Differentiate the product; $H|\psi\rangle = E|\psi\rangle$ on the right.

(2) Take the inner product with $\langle\psi|$ and use $\langle\psi|\psi\rangle = 1$. (We have dropped the components of $\partial_t|\psi\rangle$ along the excited state: that is the adiabatic approximation.)

(3) Integrate the linear equation, with $c(0) = 1$; the time integral of $\langle\psi|\partial_t\psi\rangle\mathrm{d}t$ is the line integral of the 1-form $\langle\psi|\mathrm{d}\psi\rangle$ on parameter space.

(4) Definition of $\mathcal{A}$. Differentiating $\langle\psi|\psi\rangle = 1$ gives $\langle\psi|\mathrm{d}\psi\rangle + \overline{\langle\psi|\mathrm{d}\psi\rangle} = 0$, so $\langle\psi|\mathrm{d}\psi\rangle$ is imaginary and $\mathcal{A}$ is real.

The extra phase $\gamma = \oint\mathcal{A}$ is the **Berry phase**. It depends only on the path in parameter space, not on how fast it is traversed. Under a change of phase $|\psi\rangle \to e^{i\chi}|\psi\rangle$ the 1-form shifts by $\mathcal{A} \to \mathcal{A} - \mathrm{d}\chi$: it is a U(1) gauge field on parameter space, the **Berry connection**.

Worked example 10.58: The Berry phase of spin-$\frac{1}{2}$ is minus half the solid angle

In the north gauge (10.223):

$$\begin{aligned} \langle\psi_N|\mathrm{d}\psi_N\rangle &\overset{(1)}{=} \cos\tfrac{\theta}{2}\,\mathrm{d}\big(\cos\tfrac{\theta}{2}\big) + \sin\tfrac{\theta}{2}\,e^{-i\phi}\,\mathrm{d}\big(e^{i\phi}\sin\tfrac{\theta}{2}\big) \\ &\overset{(2)}{=} -\tfrac{1}{2}\cos\tfrac{\theta}{2}\sin\tfrac{\theta}{2}\,\mathrm{d}\theta + \tfrac{1}{2}\sin\tfrac{\theta}{2}\cos\tfrac{\theta}{2}\,\mathrm{d}\theta + i\sin^2\tfrac{\theta}{2}\,\mathrm{d}\phi \\ &= i\sin^2\tfrac{\theta}{2}\,\mathrm{d}\phi, \\ \mathcal{A}_N &\overset{(3)}{=} -\sin^2\tfrac{\theta}{2}\,\mathrm{d}\phi \overset{(4)}{=} -\tfrac{1}{2}(1-\cos\theta)\,\mathrm{d}\phi. \end{aligned}$$

(1) The components of $|\psi_N\rangle$ are real and $e^{i\phi}\sin\frac{\theta}{2}$; the bra conjugates the second.

(2) Product rule: $\mathrm{d}\cos\frac{\theta}{2} = -\frac{1}{2}\sin\frac{\theta}{2}\mathrm{d}\theta$, $\mathrm{d}(e^{i\phi}\sin\frac{\theta}{2}) = e^{i\phi}(\frac{1}{2}\cos\frac{\theta}{2}\mathrm{d}\theta + i\sin\frac{\theta}{2}\mathrm{d}\phi)$.
(3) $\mathcal{A} = i\langle\psi|\mathrm{d}\psi\rangle$ and $i\cdot i = -1$.
(4) Half-angle formula $\sin^2\frac{\theta}{2} = \frac{1}{2}(1-\cos\theta)$.

For a loop at fixed latitude θ_0, traversed once in the direction of increasing ϕ,

$$\gamma = \oint \mathcal{A}_N = -\tfrac{1}{2}(1-\cos\theta_0)\cdot 2\pi = -\tfrac{1}{2}\,\Omega(\theta_0), \qquad \Omega(\theta_0) = 2\pi(1-\cos\theta_0),$$

where Ω is the solid angle of the cap enclosed by the loop. The curvature is

$$\mathcal{F} = \mathrm{d}\mathcal{A}_N = -\tfrac{1}{2}\sin\theta\,\mathrm{d}\theta\wedge\mathrm{d}\phi,$$

minus one half of the area form of the unit sphere: the field of a magnetic monopole of "charge $-\frac{1}{2}$" sitting at $\vec{B}=0$. In the south gauge, $\mathcal{A}_S = i\langle\psi_S|\mathrm{d}\psi_S\rangle = \mathcal{A}_N + \mathrm{d}\phi = \frac{1}{2}(1+\cos\theta)\mathrm{d}\phi$, which is regular at $\theta=\pi$ (where $1+\cos\theta=0$). The two gauges differ by the transition function $e^{-i\phi}$ on the equator. The total flux is

$$\int_{S^2}\mathcal{F} \overset{(1)}{=} \int_{\text{north}}\mathrm{d}\mathcal{A}_N + \int_{\text{south}}\mathrm{d}\mathcal{A}_S \overset{(2)}{=} \oint_{\text{eq}}\mathcal{A}_N - \oint_{\text{eq}}\mathcal{A}_S$$
$$\overset{(3)}{=} -\oint_{\text{eq}}\mathrm{d}\phi = -2\pi, \qquad \text{so} \qquad \frac{1}{2\pi}\int_{S^2}\mathcal{F} = -1.$$

(1) Cut the sphere along the equator and use, on each half, the gauge that is regular there ($\mathcal{F}$ is gauge invariant).
(2) Stokes' theorem on each hemisphere; the equator is the boundary of the northern cap with the orientation of increasing ϕ, and of the southern cap with the opposite orientation.
(3) $\mathcal{A}_N - \mathcal{A}_S = -\mathrm{d}\phi$.

The integer -1 is the first Chern number of the **ground-state line bundle** $L \to S^2$ whose fibre over $\vec{n}$ is the line spanned by $|\psi(\vec{n})\rangle$. (With the conventions of Chapter T2, the connection 1-form of L is the imaginary $\omega = \langle\psi|\mathrm{d}\psi\rangle = -i\mathcal{A}$, and $c_1 = \frac{i}{2\pi}\int\mathrm{d}\omega = \frac{1}{2\pi}\int\mathcal{F}$.) The line spanned by $|\psi(\vec{n})\rangle$ in $\mathbb{C}^2$ is a point of $\mathbb{CP}^1$, so L is the tautological line bundle over $\mathbb{CP}^1 \cong S^2$, whose Chern number is indeed -1. The excited states form the complementary bundle with Chern number $+1$; the two together make the trivial bundle $S^2\times\mathbb{C}^2$, and $-1+1=0$.

Try it: Higher spin

For spin j in the state $|j,m\rangle$ along $\vec{n}$ (energy $-Bm$ in the field $H=-B\,\vec{n}\cdot\vec{J}$), show that the Berry phase is $\gamma = -m\,\Omega$ and the Chern number of that bundle is $-2m$. *Answer:* rotate $|j,m\rangle_{\hat{z}}$ by $e^{-i\phi J_3}e^{-i\theta J_2}e^{i\phi J_3}$; the same computation gives $\mathcal{A}_N = -m(1-\cos\theta)\mathrm{d}\phi$, so $\gamma = -m\Omega$ and $\frac{1}{2\pi}\int\mathcal{F} = -m\cdot 2 = -2m$. The Chern numbers over all $2j+1$ states add up to $-2\sum_m m = 0$.

What the warm-up teaches. Line up the ingredients with the anomaly problem:

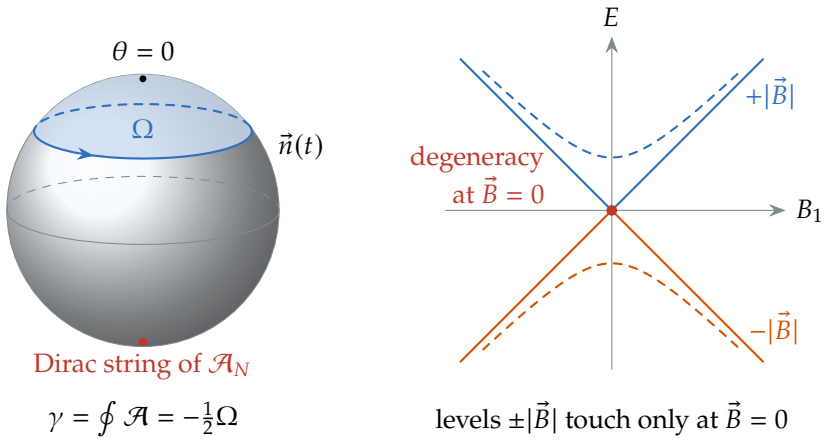


Figure 10.44: Left: the parameter space of the spin is the sphere of field directions. A slow loop $\vec{n}(t)$ encloses a cap of solid angle Ω and the ground state returns multiplied by $e^{-i\Omega/2}$. The north-gauge phase is singular at the south pole (red): that point is the end of a Dirac string, and it cannot be removed, because the Berry flux through the sphere is -2π. Right: a slice of the spectrum as a function of the field. The two levels meet only at $\vec{B} = 0$ (a "diabolical point"); a slice that misses $\vec{B} = 0$ (dashed) sees them repel. This degeneracy is the monopole whose flux the sphere measures.

Dictionary: Berry phase ↔ Hamiltonian anomaly

spin in a field	chiral fermions in a gauge field
parameter space S^2 of directions	space of gauge fields $\mathcal{A}$, or $\mathcal{A}/\mathcal{G}$
Hamiltonian $H(\vec{n})$	one-particle Dirac Hamiltonian H_A
ground state $\lvert\psi(\vec{n})\rangle$	Dirac sea (Fock vacuum) $\lvert\Omega(A)\rangle$
ground-state line bundle L	vacuum line bundle = a determinant line bundle
Berry curvature $\mathcal{F}$	local anomaly, Schwinger term
$c_1(L) = -1 \neq 0$: no global phase choice	no gauge-invariant choice of vacuum phase
degeneracy at $\vec{B} = 0$	zero modes of H_A (spectral flow)

The ground-state bundle is only defined on the space of *gapped* Hamiltonians. If we allowed $\vec{B} = 0$ the parameter space would be all of $\mathbb{R}^3$, which is contractible, and all topology would disappear. The topology is created by removing the degeneracy. The same is true for Dirac operators: all the topology of the anomaly comes from how zero modes sit in the space of gauge fields. To handle that in infinite dimensions we need the right spaces of operators: the Fredholm operators.

10.10.2 A family is a single operator in more dimensions

Let $\not{D}_t^{(n)}$ be a family of Dirac operators on an n-dimensional spin manifold M, depending on p real parameters $t = (t_1, \ldots, t_p)$ [Nash p. 292, (10.67)]:

$$\not{D}^{(n)}_{(t_1,\ldots,t_p)}.$$

Typically t are coordinates on a p-sphere S^p inside $\mathcal{A}/\mathcal{G}$ (as in Chapter 10, §2, where the local anomaly was detected on a 2-sphere of gauge orbits). Out of this family we can build one Dirac

operator on the $(n+p)$-dimensional manifold $M \times$ (parameter space) [Nash (10.68)]:

$$\not{D}^{(n+p)} = \not{D}^{(n)}_{(t_1,\dots,t_p)} + \sum_{i=1}^{p} \gamma_i \frac{\partial}{\partial t_i}, \tag{10.225}$$

where $\gamma_1, \dots, \gamma_p$ are the extra gamma matrices needed in $n+p$ dimensions. The parameters have become coordinates.

Why this is the right operator. Square it. If the new gamma matrices anticommute with the old ones and with each other ($\gamma_i\gamma_j + \gamma_j\gamma_i = -2\delta_{ij}$ in the convention where $(\gamma_i\partial_i)^2 = -\partial_i^2$, and γ_i anticommutes with the γ's inside $\not{D}^{(n)}$):

$$\begin{aligned}\big(\not{D}^{(n+p)}\big)^2 &\overset{(1)}{=} \big(\not{D}^{(n)}_t\big)^2 + \sum_i \big(\not{D}^{(n)}_t \gamma_i\partial_i + \gamma_i\partial_i \not{D}^{(n)}_t\big) + \sum_{i,j} \gamma_i\gamma_j\partial_i\partial_j \\ &\overset{(2)}{=} \big(\not{D}^{(n)}_t\big)^2 - \sum_i \partial_i^2 + \sum_i \gamma_i \big(\partial_i \not{D}^{(n)}_t\big).\end{aligned}$$

(1) Expand the square of a sum; γ_i are constant matrices.

(2) In the cross term, $\gamma_i\partial_i(\not{D}\,\psi) = \gamma_i(\partial_i\not{D})\psi + \gamma_i\not{D}\partial_i\psi$ and $\not{D}\gamma_i = -\gamma_i\not{D}$, so only the derivative of the family survives; in the last term only $i = j$ survives the symmetric sum, giving $\gamma_i^2\partial_i^2 = -\partial_i^2$.

So the $(n+p)$-dimensional Laplacian-type operator is "the square of the family, plus the Laplacian in the parameter directions, plus a term that knows how fast the family changes". When the family changes slowly, the last term is small, and the low-lying spectrum of $\not{D}^{(n+p)}$ is governed by the low-lying spectrum of $\not{D}^{(n)}_t$ as t moves: this is the same adiabatic idea as in the Berry phase and in Witten's global gravitational anomaly (Section 10.9.12).

Worked example 10.59: From a circle to a torus: $n = 1, p = 1$

Take $M = S^1$ with coordinate θ, and the family $D_t = -i\frac{\mathrm{d}}{\mathrm{d}\theta} + kt$ ($k \in \mathbb{Z}$, $0 \le t \le 1$), which will be Nash's example in Section 10.10.7. In two dimensions use $\gamma^\theta = \sigma_1$, $\gamma^t = \sigma_2$ and chirality σ_3. The two-dimensional Dirac operator on the (t,θ) cylinder is

$$\begin{aligned}\mathcal{D} &\overset{(1)}{=} \sigma_1\big(-i\partial_\theta + kt\big) + \sigma_2\big(-i\partial_t\big) \\ &\overset{(2)}{=} \begin{pmatrix} 0 & D_t - \partial_t \\ D_t + \partial_t & 0 \end{pmatrix}.\end{aligned}$$

(1) The Dirac operator $\gamma^\mu(-i\partial_\mu + a_\mu)$ with the gauge field $a = kt\,\mathrm{d}\theta$ (so $a_\theta = kt$, $a_t = 0$).

(2) $\sigma_1(\cdot)$ puts D_t in both off-diagonal slots; $\sigma_2(-i\partial_t) = \left(\begin{smallmatrix} 0 & -\partial_t \\ \partial_t & 0 \end{smallmatrix}\right)$.

The lower-left block $D_A := \partial_t + D_t$ maps positive-chirality spinors to negative-chirality ones: it is Nash's operator (10.85), the chiral Dirac operator on the torus for the field strength $F = \mathrm{d}a = k\,\mathrm{d}t \wedge \mathrm{d}\theta$. This is (10.225) with $n = p = 1$ and the single extra gamma matrix acting as "∂_t in the chiral block". So a one-parameter family of 1-dimensional operators *is* a single 2-dimensional operator.

The Hamiltonian version. Nash adds a second reading of the same trick. In a Hamiltonian treatment spacetime is $\mathbb{R}_{\text{time}} \times \Sigma^{n-1}$, and the object that matters is the Dirac Hamiltonian on the $(n-1)$-dimensional space Σ. A p-parameter family of n-dimensional Euclidean operators

becomes a $(p+1)$-parameter family $\not{D}_t^{(n-1)}$ of $(n-1)$-dimensional Hamiltonians, the extra parameter being time. The totals are equal:

$$n + p = (n-1) + (p+1).$$

Nash's claim is that the anomaly depends only on this total. We now make the claim precise with Fredholm operators.

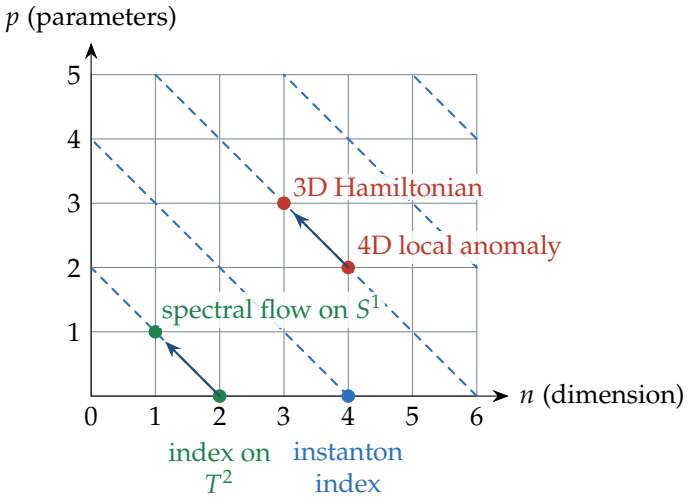


Figure 10.45: Bookkeeping of families. A point (n, p) is a p-parameter family of n-dimensional Dirac operators. The anomaly group depends only on the diagonal $n + p$ it lies on; it is $\mathbb{Z}$ on the dashed diagonals $n + p$ even and 0 in between. The arrows are the step from the Euclidean picture to the Hamiltonian one, $(n, p) \mapsto (n-1, p+1)$: the 4D local anomaly on a 2-sphere of gauge orbits becomes a 3-sphere family of 3D Hamiltonians, and the index on the torus becomes spectral flow on the circle.

10.10.3 Fredholm operators and the index map

Nash now recalls that elliptic operators on compact manifolds are *Fredholm* (Chapter 2) and that the space of Fredholm operators is related to K-theory (Chapter 3). We rebuild both facts here in the form we need.

Working definition 10.15 (Fredholm operator and its index)

A bounded operator [$\|T\psi\| \le C\,\|\psi\|$ for a fixed C. The last sentence shows how a Dirac operator is made so] $T : H \to H'$ between Hilbert spaces is **Fredholm** if $\dim \ker T < \infty$ and $\dim \operatorname{Coker} T < \infty$, where $\operatorname{Coker} T = H'/\operatorname{Im} T$ [H' with the range of T divided out, so its dimension counts the independent directions that T cannot reach] (the range is then automatically closed, and $\operatorname{Coker} T \cong \ker T^*$). Its **index** is

$$\operatorname{index} T = \dim \ker T - \dim \ker T^*.$$

How to test it: count the solutions of $T\psi = 0$ and of $T^*\chi = 0$. *Key property:* the index does not change under continuous deformations of T through Fredholm operators, nor under adding a compact operator [a norm limit of finite-rank operators, i.e. up to an arbitrarily small error it acts on only finitely many independent directions]. We write $\mathcal{F}(H)$ for the space of Fredholm operators $H \to H$ with the norm topology [two operators are close when $\|T - T'\|$ is small]. An unbounded elliptic operator such as a Dirac operator is made bounded without changing its kernel or the signs of its eigenvalues by $D \mapsto D(1 + D^*D)^{-1/2}$.

Worked example 10.60: Three Fredholm operators with their indices

(a) Finite dimensions. Every linear map $T : \mathbb{C}^a \to \mathbb{C}^b$ is Fredholm, and

$$\operatorname{index} T \overset{(1)}{=} \dim \operatorname{Ker} T - \dim \operatorname{Ker} T^* \overset{(2)}{=} (a - \operatorname{rank} T) - (b - \operatorname{rank} T) = a - b.$$

(1) Definition.
(2) Rank–nullity for T and for $T^* : \mathbb{C}^b \to \mathbb{C}^a$, which has the same rank.

The index does not depend on T at all, only on the dimensions: the simplest instance of "the index is a topological invariant". In infinite dimensions $a - b$ becomes $\infty - \infty$, and the index is the finite number that survives.
(b) The shift. On $\ell^2 = \{(c_0, c_1, c_2, \dots)\}$ with basis $e_0, e_1, \dots$, let $S e_j = e_{j+1}$. As a matrix,

$$S = \begin{pmatrix} 0 & 0 & 0 & \cdots \\ 1 & 0 & 0 & \cdots \\ 0 & 1 & 0 & \cdots \\ \vdots & & \ddots & \ddots \end{pmatrix}, \qquad S^*S = \mathbb{1}, \qquad SS^* = \mathbb{1} - |e_0\rangle\langle e_0|.$$

S is injective, so $\operatorname{Ker} S = 0$; its range misses e_0, and $\operatorname{Ker} S^* = \mathbb{C}e_0$. So index $S = 0 - 1 = -1$, index $S^* = +1$, and index $S^k = -k$. Every integer occurs as an index. Note that S is "almost invertible": S^* is an inverse up to the rank-one (compact) operator $|e_0\rangle\langle e_0|$. That is the general picture: T is Fredholm exactly when it is invertible modulo compact operators.
(c) A Dirac operator. $D_a = -i\frac{\mathrm{d}}{\mathrm{d}\theta} + a$ on S^1 has eigenvalues $n + a$, $n \in \mathbb{Z}$, on $e^{in\theta}$. It is self-adjoint, so $\operatorname{Ker} D_a = \operatorname{Ker} D_a^*$ and index $D_a = 0$; the kernel is $\mathbb{C}e^{-ia\theta}$ when $a \in \mathbb{Z}$ and zero otherwise. The index is 0 although the kernel jumps: kernel and cokernel jump together.

Families and the index bundle. Now let X be a compact space and [Nash (10.70)]

$$\alpha : X \to \mathcal{F}(H), \qquad x \mapsto \alpha(x),$$

a continuous map. This is a *family* α_X of Fredholm operators parametrised by X. For each x we have two finite-dimensional spaces $\operatorname{Ker} \alpha(x)$ and $\operatorname{Ker} \alpha(x)^*$. If their dimensions do not jump, they are the fibres of two vector bundles $\operatorname{Ker} \alpha$ and $\operatorname{Ker} \alpha^*$ over X, and [Nash (10.71)]

$$\operatorname{Index} \alpha_X := [\operatorname{Ker} \alpha] - [\operatorname{Ker} \alpha^*] \in K(X)$$

is a formal difference of bundles, an element of the K-theory of X. This is the index bundle of Chapter 10, §2, for an abstract family.

When the dimensions do jump, the formula needs a small repair, and the repair is worth seeing because it is the same idea as the "cut at ε" that will appear in the Fock-space construction.

Worked example 10.61: Repairing a jumping kernel

Take $X = [-1, 1]$ and the family of 1×1 matrices $T(x) = x : \mathbb{C} \to \mathbb{C}$. At $x \neq 0$ kernel and cokernel are zero; at $x = 0$ both are $\mathbb{C}$. Naively $[\operatorname{Ker}] - [\operatorname{Ker}^*]$ is "$0 - 0$" away from 0 and "$\mathbb{C} - \mathbb{C}$" at 0: the difference is the same but neither term is a bundle. Repair: add a fixed extra dimension to the source,

$$\tilde{T}(x) : \mathbb{C} \oplus \mathbb{C} \to \mathbb{C}, \qquad (v, w) \mapsto xv + w.$$

Now $\tilde{T}(x)$ is surjective for every x (take w), so $\operatorname{Ker}\tilde{T}^* = 0$, and

$$\operatorname{Ker}\tilde{T}(x) = \{(v, -xv)\} \cong \mathbb{C}$$

has constant dimension 1: a genuine (trivial) line bundle. The index bundle is $[\operatorname{Ker}\tilde{T}] - [\mathbb{C}_{\text{extra}}] = 1 - 1 = 0$, as it should be. In general one picks a finite-dimensional space W so that $\alpha(x) \oplus$ (inclusion of W) is onto for all x (possible because X is compact); the kernels then form a bundle, and $\operatorname{Index}\alpha_X = [\text{kernel bundle}] - [W]$.

Worked example 10.62: The Berry bundle is an index bundle

Over $X = S^2$ let $\alpha(\vec{n}) : \mathbb{C}^2 \to \mathbb{C}$ be $v \mapsto \langle\psi_\perp(\vec{n})|v\rangle$, the component along the excited state of Example 10.58 (well defined up to a phase, which does not change kernels). Each $\alpha(\vec{n})$ is onto, so $\operatorname{Ker}\alpha^* = 0$, and $\operatorname{Ker}\alpha(\vec{n}) = \mathbb{C}|\psi(\vec{n})\rangle$ is the ground-state line. Hence

$$\operatorname{Index}\alpha_{S^2} = [L] \in K(S^2), \qquad \text{rank } 1, \quad c_1 = -1.$$

The ordinary index (the rank, $2 - 1 = 1$) sees nothing. The first Chern class sees the Berry monopole. Anomalies live in exactly this extra information.

The Atiyah–Jänich theorem. The index bundle depends only on the homotopy class of α, and every element of $K(X)$ arises. That is [Nash (10.69), (10.72)]:

$$\operatorname{Index} : [X, \mathcal{F}(H)] \xrightarrow{\cong} K(X), \qquad [\alpha] \mapsto [\operatorname{Ker}\alpha] - [\operatorname{Ker}\alpha^*] = \operatorname{Index}\alpha_X. \tag{10.226}$$

Here $[X, Y]$ means homotopy classes of maps. Nash's name for this: $\mathcal{F}(H)$ is a **classifying space** for K-theory.

Working definition 10.16 (classifying space)

A space B *classifies* a kind of object $\mathcal{E}$ over spaces X if the objects over X [here: K-theory classes $[V] - [W]$ over X, with $B = \mathcal{F}(H)$], up to isomorphism, correspond to homotopy classes of maps $X \to B$ [maps counted as the same when one can be deformed continuously into the other]. *How to use it:* to classify objects over a sphere S^p it is enough to know the homotopy group $\pi_p(B)$ [maps $S^p \to B$ sending a chosen point of S^p to a chosen point of B, up to continuous deformation] (for connected B; in general, based maps).

Nash's footnote 6 recalls the model case from Chapter 1: $[X, \mathrm{Gr}(k, \infty)] \cong \mathrm{Vect}_k(X)$. The Grassmannian of k-planes in $\mathbb{C}^\infty$ classifies rank-k bundles. The map sends x to the fibre E_x, viewed as a k-plane in a big $\mathbb{C}^N$. For $k = 1$, $\mathrm{Gr}(1, \infty) = \mathbb{CP}^\infty$, and line bundles over S^2 are classified by $\pi_2(\mathbb{CP}^\infty) = \mathbb{Z}$, the Chern number. The Berry bundle is the map $\vec{n} \mapsto [\psi(\vec{n})] \in \mathbb{CP}^1 \subset \mathbb{CP}^\infty$, the identification $S^2 \cong \mathbb{CP}^1$, a map of degree one (the sign -1 of c_1 is the convention that the tautological bundle has $c_1 = -1$).

The homotopy type of $\mathcal{F}(H)$. Three facts pin down $\mathcal{F}(H)$ [Nash (10.73)]:

1. $\pi_0(\mathcal{F}) = \mathbb{Z}$, via the index: two Fredholm operators can be joined by a path of Fredholm operators if and only if they have the same index. (One direction is the stability of the index. The other uses the next fact.)

2. **Kuiper's theorem**: the unitary group $\mathrm{U}(H)$ of an infinite-dimensional Hilbert space is contractible. This is very unlike $\mathrm{U}(N)$, and it is why "every component looks the same":

each component of $\mathcal{F}$ is homotopy equivalent to $BU = \lim_N \mathrm{Gr}(N, 2N)$, the classifying space of stable bundles. Altogether $\mathcal{F} \simeq \mathbb{Z} \times BU$.

3. **Bott periodicity**: $\Omega U(\infty) \simeq \mathbb{Z} \times BU$, where $U(\infty) = \bigcup_N U(N)$ and Ω means based loops.

Combining the last two gives Nash's (10.73):

$$\mathcal{F}(H) \simeq \Omega U(\infty). \tag{10.227}$$

Worked example 10.63: A loop of unitaries gives a Fredholm operator: Toeplitz

The equivalence (10.227) can be made explicit. Let $H = L^2(S^1)$ with basis $z^j = e^{ij\theta}$, $j \in \mathbb{Z}$, and let $H_+ = \overline{\mathrm{span}}\{z^j : j \geq 0\}$ with orthogonal projection P_+. A loop $\gamma : S^1 \to U(1)$ acts on H by multiplication. Its **Toeplitz operator** is $T_\gamma = P_+ \gamma P_+ : H_+ \to H_+$. For $\gamma(z) = z$:

$$T_z(z^j) \overset{(1)}{=} P_+(z^{j+1}) = z^{j+1} \quad (j \geq 0),$$
$$\mathrm{index}\, T_z \overset{(2)}{=} \mathrm{index}\, S = -1 = -\,\mathrm{wind}(\gamma).$$

(1) Multiplying by z raises the power; the result is still in H_+.
(2) In the basis $z^0, z^1, \ldots$ this is exactly the shift S of Example 10.60(b); the winding number of $z \mapsto z$ is 1.

In general $\mathrm{index}\, T_\gamma = -\,\mathrm{wind}(\det \gamma)$ for loops in $U(N)$ (with $H_+ \otimes \mathbb{C}^N$). So the map "loop $\mapsto$ Toeplitz operator" takes the winding number, which labels $\pi_0(\Omega U) = \pi_1(U) = \mathbb{Z}$, to minus the index, which labels $\pi_0(\mathcal{F}) = \mathbb{Z}$.

10.10.4 Self-adjoint Fredholm operators and spectral flow

In odd dimensions there is no chirality, and the Dirac operator is self-adjoint. So we need the space $\mathcal{F}^1(H)$ of *self-adjoint* Fredholm operators. Its index is always zero ($\mathrm{Ker}\, T = \mathrm{Ker}\, T^*$), so the index cannot be the invariant. The invariant is subtler.

A self-adjoint Fredholm operator has real spectrum. Near zero there are only finitely many eigenvalues, each of finite multiplicity (otherwise the kernel of a small perturbation would be infinite-dimensional). Far from zero the spectrum may accumulate (for bounded operators) or run off to infinity (for the unbounded originals). The **essential spectrum** is where it accumulates. It splits $\mathcal{F}^1$ into three pieces [Nash p. 293]:

- $\mathcal{F}^1_+$: essential spectrum only positive; finitely many negative eigenvalues. Example: the Laplacian $-\mathrm{d}^2/\mathrm{d}\theta^2$ on S^1 (after bounding), or $\mathbb{1}$.
- $\mathcal{F}^1_-$: essential spectrum only negative; example: $-\mathbb{1}$.
- $\mathcal{F}^1_*$: essential spectrum on both sides; infinitely many eigenvalues of each sign. Example: every Dirac operator $-i\mathrm{d}/\mathrm{d}\theta + a$ on S^1, and every Dirac operator in odd dimensions.

Why $\mathcal{F}^1_\pm$ are contractible. If $T \in \mathcal{F}^1_+$, the straight-line path $T_s = (1-s)T + s\mathbb{1}$, $0 \leq s \leq 1$, stays inside $\mathcal{F}^1_+$: each T_s is self-adjoint, and its essential spectrum is $(1-s)\,\mathrm{ess}(T) + s$, still positive and away from 0. So the whole of $\mathcal{F}^1_+$ contracts to the point $\mathbb{1}$. The same with $-\mathbb{1}$ for $\mathcal{F}^1_-$. Physically: an operator that is bounded below cannot have spectral flow, because there is nowhere for eigenvalues to flow *from*. Nash stresses this at the end of his example (p. 296): "if the Dirac operator had a lowest bound no flow could take place".

The interesting piece. All the topology is in $\mathcal{F}^1_*$. Atiyah and Singer showed [Nash (10.74)–(10.75)]

$$\mathcal{F}^1 \simeq \mathcal{F}^1_* \simeq \Omega\mathcal{F}(H) \simeq \Omega(\Omega\mathrm{U}(\infty)) \simeq \mathrm{U}(\infty), \tag{10.228}$$

the last step being Bott periodicity again, $\Omega^2\mathrm{U} \simeq \mathrm{U}$.

The equivalence $\mathcal{F}^1_* \simeq \mathrm{U}(\infty)$ has a transparent picture. Bound the operator so its spectrum lies in $[-1,1]$, then push the essential spectrum to the two ends ±1 (this is a homotopy). Now map

$$T \longmapsto U_T = -\exp(i\pi T).$$

U_T is unitary. On the essential spectrum ±1 it equals $-e^{\pm i\pi} = 1$, so $U_T - \mathbb{1}$ is compact: U_T lies in the stable unitary group, which is a model of $\mathrm{U}(\infty)$. An eigenvalue λ of T becomes the point $-e^{i\pi\lambda}$ on the unit circle; $\lambda = 0$ goes to -1. *An eigenvalue crossing zero becomes an eigenvalue of U_T crossing -1* (Figure 10.46).

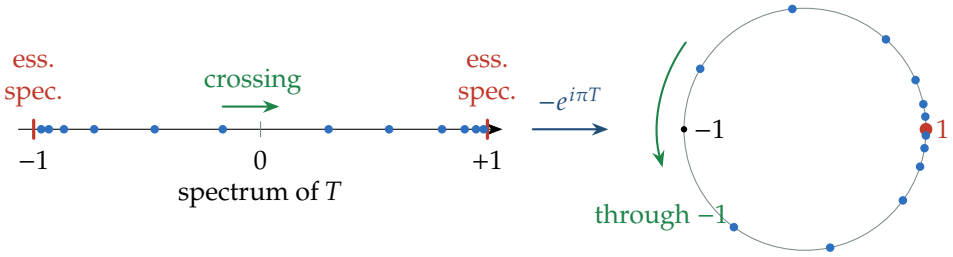


Figure 10.46: The map $T \mapsto -e^{i\pi T}$ from self-adjoint Fredholm operators (spectrum in $[-1,1]$, essential spectrum at the red ends) to unitaries differing from $\mathbb{1}$ by a compact operator. The essential spectrum goes to 1; zero goes to -1. A zero crossing of T (green) becomes a passage of an eigenvalue of U_T through -1, so spectral flow round a loop of T's is a winding number of unitaries.

Worked example 10.64: Spectral flow is a winding number: a telescoping sum

Take the loop $D_t = -i\frac{\mathrm{d}}{\mathrm{d}\theta} + t$, $0 \le t \le 1$ (the case $k = 1$ of Example 10.59). The eigenvalues are $\lambda_n(t) = n + t$. Bound them with $f(\lambda) = \lambda/\sqrt{1+\lambda^2} \in (-1,1)$. Then $U_t = -\exp(i\pi f(D_t))$ has eigenvalues $-e^{i\pi f(n+t)}$, and the change of the total phase of its (regularised) determinant over the loop is

$$\Delta \arg\det U_t \overset{(1)}{=} \pi \sum_{n\in\mathbb{Z}} \big[f(\lambda_n(1)) - f(\lambda_n(0))\big] \overset{(2)}{=} \pi \sum_{n\in\mathbb{Z}} \big[f(n+1) - f(n)\big]$$
$$\overset{(3)}{=} \pi\Big[\lim_{N\to\infty} f(N+1) - \lim_{N\to\infty} f(-N)\Big] \overset{(4)}{=} \pi\,[1-(-1)] = 2\pi.$$

(1) The phase of $\det U$ is the sum of the phases of the eigenvalues; follow each eigenvalue continuously in t.
(2) $\lambda_n(1) = n+1$ and $\lambda_n(0) = n$.
(3) The sum telescopes (sum from $-N$ to N and let $N \to \infty$).
(4) $f(\lambda) \to \pm1$ as $\lambda \to \pm\infty$.

So the loop of unitaries winds once: the element of $\pi_1(\mathrm{U}(\infty)) = \mathbb{Z}$ is 1. The spectrum of D_1 is the same set as that of D_0, but every eigenvalue has moved up one slot, and one of them crossed zero. The telescoping sum only "sees" the two ends of the spectrum. That is why spectral flow needs a spectrum unbounded in both directions: if the spectrum had a bottom, the second limit would be a fixed finite eigenvalue and the flow would be zero.

Homotopy groups. With $\mathcal{F} \simeq \Omega\mathrm{U}$ and $\mathcal{F}^1 \simeq \mathrm{U}$ we can read off the homotopy groups from those of the stable unitary group, $\pi_k(\mathrm{U}(\infty)) = \mathbb{Z}$ for k odd and 0 for k even (Bott; for small

k: $\pi_1(\mathrm{U}(1)) = \mathbb{Z}$ is the winding number, $\pi_3(\mathrm{SU}(2)) = \pi_3(S^3) = \mathbb{Z}$ is the instanton number, $\pi_2(\mathrm{U}(N)) = 0$). Hence [Nash (10.76)]:

$$\pi_k(\mathcal{F}(H)) \overset{(1)}{=} \pi_k(\Omega\mathrm{U}) \overset{(2)}{=} \pi_{k+1}(\mathrm{U}) = \begin{cases} \mathbb{Z} & k \text{ even} \\ 0 & k \text{ odd} \end{cases} \tag{10.229}$$

$$\pi_k(\mathcal{F}^1(H)) \overset{(3)}{=} \pi_k(\mathrm{U}) = \begin{cases} 0 & k \text{ even} \\ \mathbb{Z} & k \text{ odd} \end{cases} \tag{10.230}$$

(1) The homotopy equivalence (10.227).
(2) A map $S^k \to \Omega Y$ is the same as a map $S^k \times S^1 \to Y$ sending the based circles to the base point, i.e. a map $S^{k+1} \to Y$: $\pi_k(\Omega Y) = \pi_{k+1}(Y)$.
(3) The homotopy equivalence (10.228).

k	0	1	2	3	4	5
$\pi_k(\mathrm{U}(\infty))$	0	$\mathbb{Z}$	0	$\mathbb{Z}$	0	$\mathbb{Z}$
$\pi_k(\mathcal{F})$	$\mathbb{Z}$ (index)	0	$\mathbb{Z}$ (c_1 of index bundle)	0	$\mathbb{Z}$	0
$\pi_k(\mathcal{F}^1)$	0	$\mathbb{Z}$ (spectral flow)	0	$\mathbb{Z}$	0	$\mathbb{Z}$

The first entries have physical names. $\pi_0(\mathcal{F}) = \mathbb{Z}$ is the ordinary index (zero modes of an instanton). $\pi_1(\mathcal{F}^1) = \mathbb{Z}$ is spectral flow round a loop. $\pi_2(\mathcal{F}) = \mathbb{Z}$ is the first Chern class of the index bundle over a 2-sphere, which is how the local anomaly was detected in Chapter 10, §2, and which (Example 10.62) is the Berry monopole.

Watch out: Berry bundles and $\pi_2(\mathcal{F}^1) = 0$

The Berry example is a 2-sphere family of self-adjoint 2×2 matrices with a nontrivial Chern number, yet $\pi_2(\mathcal{F}^1) = 0$. There is no contradiction. The Berry bundle is defined over the space of *invertible* (gapped) self-adjoint operators, and records how the negative-energy subspace twists. Inside all of $\mathcal{F}^1$ the same sphere can be shrunk by passing through operators with zero modes, as the sphere $|\vec{B}| = 1$ shrinks to $\vec{B} = 0$ through the degeneracy. In infinite dimensions the gapped self-adjoint operators in $\mathcal{F}^1_*$ form a space homotopy equivalent to $\mathcal{F} \simeq \mathbb{Z} \times B\mathrm{U}$ (the "restricted Grassmannian" of negative-energy subspaces). Its π_2 is $\mathbb{Z}$: this is where the vacuum line bundles of Section 10.10.12 live.

10.10.5 Only $n + p$ matters

Back to anomalies. Restrict the parameter space to a sphere, $X = S^p$. Then [Nash (10.77)]

$$\not{D}^{(n)}_{(t_1,\dots,t_p)} \in \begin{cases} \mathcal{F}(H) & n \text{ even (the chiral operator } E^+ \to E^-), \\ \mathcal{F}^1(H) & n \text{ odd (the self-adjoint Dirac operator),} \end{cases}$$

and the topology of the anomaly is in the index bundle over S^p, i.e. in the homotopy class of the map $S^p \to \mathcal{F}$ or $S^p \to \mathcal{F}^1$ [Nash (10.78)]:

$$\pi_p(\mathcal{F}(H)) \quad (n \text{ even}), \qquad \pi_p(\mathcal{F}^1(H)) \quad (n \text{ odd}).$$

Now use the table.

- $n + p$ even. Then n, p are both even or both odd. If both even: $\pi_p(\mathcal{F}) = \mathbb{Z}$. If both odd: $\pi_p(\mathcal{F}^1) = \mathbb{Z}$. Either way $\mathbb{Z}$.

- $n+p$ odd. Then one is even and one odd. If n even, p odd: $\pi_p(\mathcal{F}) = 0$. If n odd, p even: $\pi_p(\mathcal{F}^1) = 0$. Either way 0.

A single formula covers all cases [Nash (10.79)]:

$$\text{anomaly group} \overset{(1)}{=} \begin{cases} \pi_p(\Omega \mathrm{U}) = \pi_{p+1}(\mathrm{U}) & n \text{ even} \\ \pi_p(\mathrm{U}) & n \text{ odd} \end{cases} \overset{(2)}{=} \pi_{n+p+1}(\mathrm{U}(\infty)) \overset{(3)}{=} \pi_{n+p}(\Omega \mathrm{U}(\infty)). \quad (10.231)$$

(1) (10.227) and (10.228).
(2) Bott periodicity makes $\pi_k(\mathrm{U})$ depend only on k mod 2. For n even $p+1 \equiv n+p+1$; for n odd $p \equiv n+p+1 \pmod 2$.
(3) $\pi_k(\Omega Y) = \pi_{k+1}(Y)$.

It is $\mathbb{Z}$ for $n+p$ even and 0 for $n+p$ odd. The Euclidean (n,p) and the Hamiltonian $(n-1,p+1)$ descriptions of the same anomaly lie on the same diagonal of Figure 10.45 and so have the same anomaly group.

Worked example 10.65: Checking the diagonal on three cases

(i) Instantons, $(n,p) = (4,0)$. A single chiral Dirac operator on S^4: $\pi_0(\mathcal{F}) = \mathbb{Z}$, the index = instanton number (Chapter 8). *(ii) The 4D local anomaly,* $(4,2)$*, and its Hamiltonian form* $(3,3)$. In §2 the anomaly of SU(N) on S^4 was the Chern class of the determinant bundle over a 2-sphere in $\mathcal{A}/\mathcal{G}$: $\pi_2(\mathcal{F}) = \mathbb{Z}$. In the Hamiltonian picture space is S^3, the Dirac Hamiltonian is self-adjoint, and the family is over a 3-sphere: $\pi_3(\mathcal{F}^1) = \mathbb{Z}$. Same diagonal $n+p=6$. *(iii) The torus and the circle,* $(2,0)$ *and* $(1,1)$. The chiral operator D_A on T^2 has an index ($\pi_0(\mathcal{F}) = \mathbb{Z}$). Its Hamiltonian version is the loop D_t of self-adjoint operators on S^1, whose invariant is spectral flow ($\pi_1(\mathcal{F}^1) = \mathbb{Z}$). This pair is Nash's worked example, done in full in Section 10.10.7, where we will see that the index on T^2 and the spectral flow on S^1 are the same integer k.

Physical picture: Why the parity of $n+p$

Each step up in dimension adds one gamma matrix; each extra parameter, through (10.225), also adds one. Chirality (and hence the distinction between $\mathcal{F}$ and $\mathcal{F}^1$) depends on whether the total number of gamma matrices is even or odd. So a parameter and a dimension are interchangeable, and only their total counts. The same count explains the familiar descent: a $2m$-dimensional anomaly comes from a $(2m+2)$-form ch_{m+1}, i.e. from "two more dimensions".

10.10.6 The real case and Witten's SU(2) anomaly again

Global anomalies can involve *real* operators. In Section 10.9.5 we saw that for an SU(2) doublet the pseudo-real structure makes the Weyl operator equivalent to a real skew-adjoint operator, whose kernel dimension is invariant only mod 2. So we should redo the Fredholm analysis on a real Hilbert space H_r [Nash p. 294].

Replacing H by H_r replaces U(∞) by the infinite orthogonal group O(∞), and complex K-theory K by real K-theory KO:

- $\mathcal{F}(H_r)$ classifies KO: $[X, \mathcal{F}(H_r)] \cong KO(X)$, by the real version of the index map (10.226).
- The space $\mathcal{F}^1_{\text{skew}}(H_r)$ of real skew-adjoint Fredholm operators is homotopy equivalent to O(∞) (Atiyah and Singer).

The global anomaly is captured by

$$\pi_0\big(\mathcal{F}^1_{\text{skew}}(H_r)\big) = \pi_0(\mathrm{O}(\infty)) = \mathbb{Z}_2,$$

in agreement with Witten's $\mathbb{Z}_2$ of Section 10.9. The two components of $\mathrm{O}(N)$ are det $= \pm 1$, and the corresponding invariant of a real skew operator is the parity of dim Ker.

Worked example 10.66: Why only the parity of the kernel survives

A real skew-symmetric matrix K ($K^T = -K$) has eigenvalues in pairs $\pm i\mu$ ($\mu > 0$) plus zeros. For $K = \left(\begin{smallmatrix} 0 & \mu \\ -\mu & 0 \end{smallmatrix}\right)$, the characteristic polynomial is $\lambda^2 + \mu^2$, roots $\pm i\mu$. Deforming K can move a pair $\pm i\mu$ to 0 (as $\mu \to 0$, two zero modes appear) or split two zero modes into a pair, but can never create or destroy a *single* zero mode. So

$$\dim \mathrm{Ker}\, K \bmod 2$$

is a deformation invariant: the **mod-2 index**. An odd-size real skew matrix always has a kernel ($\det K = \det K^T = (-1)^N \det K$ forces $\det K = 0$ for N odd). In Witten's argument, the five-dimensional operator built from a loop of four-dimensional SU(2) operators by (10.225) is of this kind, and its mod-2 index is 1 for the non-trivial loop: that is the sign flip of the Pfaffian.

k mod 8	0	1	2	3	4	5	6	7
$\pi_k(\mathrm{O}(\infty))$	$\mathbb{Z}_2$	$\mathbb{Z}_2$	0	$\mathbb{Z}$	0	0	0	$\mathbb{Z}$

Real Bott periodicity has period 8. The $\mathbb{Z}_2$'s in this table are the source of every mod-2 global anomaly; the $\mathbb{Z}_2$ at $k = 0$ is the one Nash uses.

Reading Nash (pp. 291–295, (10.67)–(10.79))

(10.67)–(10.68) are the $n + p$ trick of Section 10.10.2. (10.69) and (10.72) are the Atiyah–Jänich theorem; Nash writes the index map as $[\alpha] \mapsto [\mathrm{Ker}\,\alpha] - [\mathrm{Ker}\,\alpha^*]$, which is literally correct only when the kernel dimensions are constant; Example 10.61 shows the general repair. (10.73) is $\mathcal{F} \simeq \Omega\mathrm{U}(\infty)$: Kuiper plus Bott. In the text before (10.74) Nash's $\mathcal{F}^1_+$ is the part with "only a finite number of negative eigenvalues" and $\mathcal{F}^1_-$ the part with finitely many positive ones; both are contractible, so the topology is carried by $\mathcal{F}^1_*$. (10.74)–(10.75) are Atiyah–Singer plus Bott. (10.76) is our table. (10.77)–(10.79) is the parity count of Section 10.10.5; note that Nash's (10.79) is stated with $\Omega\mathrm{U}(\infty)$, i.e. it is $\pi_{n+p}(\mathcal{F})$. On p. 294–295 he treats the real case with $\mathrm{O}(\infty)$ and $\mathcal{F}^1_{\text{skew}}(H_r)$.

10.10.7 Example: spectral flow and an anomaly in two dimensions

We now do Nash's first example completely. It is the smallest system in which every idea of the section can be seen with explicit functions: an abelian U(1) gauge theory on the two-dimensional torus $T^2 = S^1 \times S^1$. We look at it twice. First as a Euclidean space-time, where the relevant object is the *index* of one chiral Dirac operator. Then as "space = circle, time = the other circle", where the relevant object is the *spectral flow* of a loop of self-adjoint operators. The diagonal of Figure 10.45 predicts that the two numbers agree, since $(n, p) = (2, 0)$ and $(1, 1)$ both have $n + p = 2$. We shall check that they do, and see *why* they do.

The Euclidean side. On the even-dimensional torus we have chirality, so we have a chiral Dirac operator $\not{D}_A$ coupled to a U(1) connection A. In two dimensions the local anomaly

computed in §2 is a two-form, $\propto \mathrm{d}(\text{gauge parameter}\cdot A)$, and on a single line bundle it integrates to something with no topology; what is left is the fact that the torus carries *non-trivial* U(1) bundles [Nash p. 295]. These are the two-dimensional abelian "instantons". Line bundles over T^2 are classified by one integer, the first Chern number

$$k = c_1(L)[T^2] = \frac{1}{2\pi}\int_{T^2} F \in \mathbb{Z},$$

(here in the physicists' convention with a real field strength F; in the anti-Hermitian convention $F_{\text{math}} = -iF$ and $c_1 = \frac{i}{2\pi}F_{\text{math}}$), and the index theorem of Chapter 4 gives

$$\operatorname{index} \not{D}_A = \int_{T^2} \operatorname{ch}(F)\,\hat{A}(T^2) = \int_{T^2} \frac{F}{2\pi} = k,$$

because the flat torus has $\hat{A} = 1$, and only the degree-two part $\operatorname{ch}_1 = F/2\pi$ of $\operatorname{ch}(F)$ can be integrated over a surface. In the language of Fredholm families this is the case where the parameter space X is a point: the index map $[X, \mathcal{F}(H)] \to K(X)$ becomes $\pi_0(\mathcal{F}) \to K(\text{point}) = \mathbb{Z}$, which is just "a Fredholm operator $\mapsto$ its index".

Worked example 10.67: The line bundle of charge k on the torus, explicitly

Use coordinates (t, θ) with $t \in [0,1]$ and $\theta \in [0, 2\pi]$. A section of a line bundle over T^2 is a function $\psi(t,\theta)$, 2π-periodic in θ, that returns to itself after $t \mapsto t+1$ only *up to a transition function*. Take

$$\psi(t+1,\theta) = e^{-ik\theta}\,\psi(t,\theta), \qquad a = kt\,\mathrm{d}\theta, \tag{10.232}$$

where a is the gauge field, entering the covariant derivative as $\mathrm{d} + ia$ (physicists' convention, charge 1). Check that the pair is consistent and compute the Chern number:

$$a(t+1) \overset{(1)}{=} a(t) + k\,\mathrm{d}\theta \overset{(2)}{=} a(t) + \mathrm{d}(k\theta),$$
$$F \overset{(3)}{=} \mathrm{d}a = k\,\mathrm{d}t \wedge \mathrm{d}\theta,$$
$$\frac{1}{2\pi}\int_{T^2} F \overset{(4)}{=} \frac{k}{2\pi}\int_0^1 \mathrm{d}t \int_0^{2\pi} \mathrm{d}\theta = k.$$

(1) Substitute $t+1$ in $a = kt\,\mathrm{d}\theta$.

(2) So the two ends of the t-interval differ by the gauge transformation with parameter $\chi = k\theta$. Under $\psi \mapsto e^{-i\chi}\psi$, $a \mapsto a + \mathrm{d}\chi$ the covariant derivative is covariant: $(\mathrm{d} + i(a + \mathrm{d}\chi))(e^{-i\chi}\psi) = e^{-i\chi}(\mathrm{d} + ia)\psi$. This is exactly the gluing (10.232). The function $e^{-ik\theta}$ is single valued on the circle only because $k \in \mathbb{Z}$: that is why the Chern number is an integer.

(3) $\mathrm{d}(kt\,\mathrm{d}\theta) = k\,\mathrm{d}t \wedge \mathrm{d}\theta$; the field is uniform.

(4) Orient the torus by $\mathrm{d}t \wedge \mathrm{d}\theta$.

The bundle is trivial exactly when $k = 0$. For $k \neq 0$ it has no nowhere-vanishing section: a section would have to wind k times round the θ-circle between $t = 0$ and $t = 1$, which forces a zero.

The Hamiltonian side. Now we switch viewpoint [Nash p. 295]. Declare the θ-circle to be *space* and t to be a parameter. On the spatial circle the Dirac operator is self-adjoint (there is no chirality in one dimension), and with the gauge field $a = kt\,\mathrm{d}\theta$ it is simply [Nash (10.80)]

$$\not{D}^1_t = -i\frac{\mathrm{d}}{\mathrm{d}\theta} + kt, \qquad 0 \le t \le 1, \quad k \in \mathbb{Z}. \tag{10.233}$$

This is the Dirac Hamiltonian of a charged particle of one chirality (a right mover) on a ring threaded by the magnetic flux $\oint a = 2\pi kt$. As t goes from 0 to 1 the flux grows by $2\pi k$, i.e. by k flux quanta. Since the other circle of the torus is no longer Euclidean time but a parameter, we are looking at a one-parameter family of one-dimensional operators: $n = 1$, $p = 1$. By (10.230) the anomaly is measured by an element of $\pi_1(\mathcal{F}^1(H)) = \mathbb{Z}$, and the integer is computed by the spectral flow technique of Atiyah and Lusztig (and of Atiyah–Patodi–Singer).

The spectrum. The functions $e^{in\theta}$, $n \in \mathbb{Z}$, are eigenfunctions:

$$\not{D}_t^1 \, e^{in\theta} = \big(-i \cdot in + kt\big) e^{in\theta} = (n + kt)\, e^{in\theta}.$$

So the eigenvalues are [Nash (10.81)]

$$\lambda_n(t) = n + kt, \qquad n = \dots, -1, 0, 1, \dots \tag{10.234}$$

At the two ends of the family [Nash (10.82)],

$$\operatorname{spec} \not{D}_0^1 = \{n : n \in \mathbb{Z}\}, \qquad \operatorname{spec} \not{D}_1^1 = \{n + k : n \in \mathbb{Z}\} = \mathbb{Z}.$$

As sets the two spectra coincide. But if we follow each eigenvalue continuously, we find [Nash (10.83)]

$$\lambda_n(1) = \lambda_n(0) + k : \tag{10.235}$$

every eigenvalue has moved up k places. In particular the k eigenvalues $\lambda_{-1}, \dots, \lambda_{-k}$, which were negative at $t = 0$, are positive (or zero) at $t = 1$. They have crossed zero.

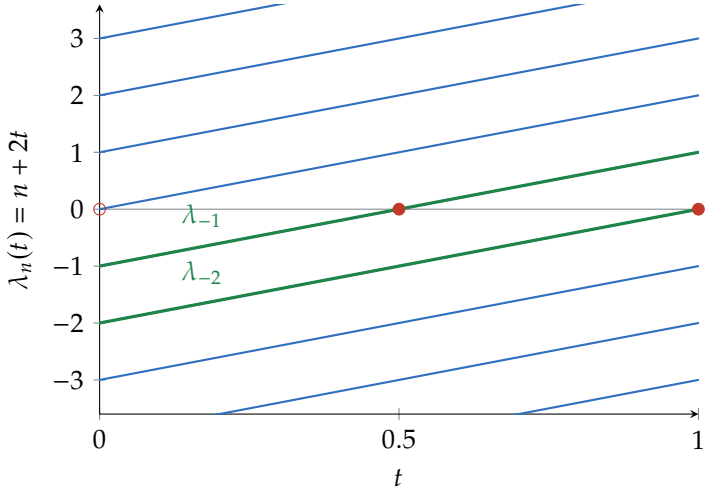


Figure 10.47: Spectral flow of $\not{D}_t^1 = -i\,\mathrm{d}/\mathrm{d}\theta + kt$ for $k = 2$. Each eigenvalue is a straight line of slope k. The set of eigenvalues is the same at $t = 0$ and $t = 1$ (all integers), but each line has climbed two places. Two lines (green) cross zero during the circuit, at the red dots $t = \frac{1}{2}$ and $t = 1$. The line through the open dot at $t = 0$ is the one that the next circuit counts at $t = 1$. The spectral flow is $2 = k$. The spectrum is unbounded in both directions, which is what makes a net upward drift possible.

Working definition 10.17 (spectral flow)

Let D_t, $0 \le t \le 1$, be a continuous family of self-adjoint Fredholm operators [real eigenvalues, with only finitely many independent zero modes and no pile-up of eigenvalues near 0, so the crossings of zero can be counted] whose end points have the same spectrum.

Follow the eigenvalues continuously in t. The **spectral flow** is

$$\mathrm{sf}(D_t) = \#\{\text{crossings from } <0 \text{ to } >0\} - \#\{\text{crossings from } >0 \text{ to } <0\}.$$

How to compute it: plot the eigenvalues against t (a "spectral-flow diagram" like Figure 10.47) and count signed crossings of the axis $\lambda = 0$. *Convention at the end points:* if some eigenvalue is exactly 0 at $t = 0$ (as here), replace D_t by $D_t - \delta$ with a small $\delta > 0$ that is not an eigenvalue; this does not change the answer for a closed loop and makes the count unambiguous. For Nash's family, with δ small, the crossings are at $t = (\delta - n)/k$ for $n = 0, -1, \dots, -(k-1)$: exactly k of them in $[0, 1]$.

The count is a (signed) integer, and in this example it is k [Nash p. 296]. Nash stresses two general points. First, in more complicated families both kinds of crossings happen, and only the net number is an invariant: a pair of crossings in opposite directions can be created or destroyed by a deformation of the family, a single crossing cannot. Second, *the spectrum must be unbounded in both directions*. The next example shows why.

Worked example 10.68: A loop of finite matrices has no spectral flow

Let D_t be a loop of Hermitian $N \times N$ matrices, $D_1 = D_0$. Let $N_-(t)$ be the number of negative eigenvalues. It changes only at crossings: it drops by one each time an eigenvalue goes from negative to positive and rises by one for each crossing the other way. So

$$\begin{aligned}\mathrm{sf}(D_t) &\overset{(1)}{=} \big[\text{number of } (- \to +)\big] - \big[\text{number of } (+ \to -)\big] \\ &\overset{(2)}{=} N_-(0) - N_-(1) \overset{(3)}{=} 0.\end{aligned}$$

(1) Definition of spectral flow.
(2) Each $(- \to +)$ crossing lowers N_- by one, each $(+ \to -)$ raises it by one; add up the changes along the path.
(3) $D_1 = D_0$, so they have the same number of negative eigenvalues.

A concrete loop with crossings: $D_t = \left(\begin{smallmatrix} \cos 2\pi t & 0 \\ 0 & -\cos 2\pi t \end{smallmatrix}\right)$. Its eigenvalues $\pm\cos 2\pi t$ cross zero at $t = \frac{1}{4}$ and $t = \frac{3}{4}$, once in each direction for each line: net flow 0. In infinite dimensions the argument fails because $N_- = \infty$ and "$\infty - \infty$" can be any integer. Nash's family is the simplest case: every eigenvalue climbs by k, and there is no bottom rung to push the spectrum against. For an operator bounded below (in $\mathcal{F}_+^1$) the argument above applies with N_- finite, so the spectral flow vanishes; this is the physical content of the contractibility of $\mathcal{F}_\pm^1$ in Section 10.10.4.

The family is a loop up to a gauge transformation. The family (10.233) is not periodic in t: $\not{D}_1^1 \neq \not{D}_0^1$ as operators. But the two end points are related by conjugation [Nash (10.84)]:

$$\not{D}_1^1 = e^{-ik\theta}\, \not{D}_0^1\, e^{ik\theta}. \tag{10.236}$$

Let us verify it on a test function $f(\theta)$:

$$\begin{aligned}e^{-ik\theta}\not{D}_0^1\big(e^{ik\theta}f\big) &\overset{(1)}{=} e^{-ik\theta}(-i)\frac{\mathrm{d}}{\mathrm{d}\theta}\big(e^{ik\theta}f\big) \\ &\overset{(2)}{=} e^{-ik\theta}(-i)\big(ik\,e^{ik\theta}f + e^{ik\theta}f'\big) \\ &\overset{(3)}{=} -if' + kf \overset{(4)}{=} \not{D}_1^1 f.\end{aligned}$$

(1) $\not{D}^1_0 = -i\,\mathrm{d}/\mathrm{d}\theta$.
(2) Product rule.
(3) Cancel $e^{-ik\theta}e^{ik\theta} = 1$; $(-i)(ik) = k$.
(4) $\not{D}^1_1 = -i\,\mathrm{d}/\mathrm{d}\theta + k$ by (10.233) at $t = 1$.

This is the familiar gauge covariance of the Dirac operator, $\not{D}_{A_g} = g^{-1}\not{D}_A g$ of §2, for the gauge transformation $g = e^{ik\theta}$, which is a map $S^1 \to \mathrm{U}(1)$ of winding number k: a *large* gauge transformation. So the path $t \mapsto \not{D}^1_t$ in the space of operators becomes a *closed loop* once we identify gauge equivalent operators, i.e. a loop in $\mathcal{A}/\mathcal{G}$ for the circle, and it is this loop whose class in $\pi_1(\mathcal{F}^1(H)) = \mathbb{Z}$ Nash computes.

If we prefer an honest loop inside $\mathcal{F}^1(H)$ itself, Kuiper's theorem (Section 10.10.3) supplies one. The unitary group of H is contractible, so there is a path U_s in $\mathrm{U}(H)$ from $U_0 = \mathbb{1}$ to $U_1 = e^{ik\theta}$ (it cannot consist of multiplication operators, since $e^{ik\theta}$ has winding number k, but it exists among all unitaries). Close the family by returning from $\not{D}^1_1$ to $\not{D}^1_0$ along $s \mapsto U_s^{-1}\cdots$, precisely $s \mapsto U_s \not{D}^1_1 U_s^{-1}$ run from $s = 0$ to $s = 1$: at $s = 0$ it is $\not{D}^1_1$ and at $s = 1$ it is $e^{ik\theta}\not{D}^1_1 e^{-ik\theta} = \not{D}^1_0$. Every operator on the return path is unitarily equivalent to $\not{D}^1_1$, so its spectrum does not move at all, and the return path contributes no spectral flow. The closed loop therefore has spectral flow k, and represents k times the generator of $\pi_1(\mathcal{F}^1) = \mathbb{Z}$.

Physical picture: What the gauge transformation does to the eigenvectors

Under conjugation by $e^{ik\theta}$ the eigenvector $e^{in\theta}$ of $\not{D}^1_0$ (eigenvalue n) becomes the eigenvector $e^{-ik\theta}e^{in\theta} = e^{i(n-k)\theta}$ of $\not{D}^1_1$, with eigenvalue $(n-k)+k = n$. So the gauge transformation matches the two ends of the family eigenvalue by eigenvalue, but it relabels the eigenvectors, shifting the label by k. The spectral flow is this relabelling made visible: following the eigenvector $e^{in\theta}$ continuously in t takes eigenvalue n to $n+k$, while the gauge identification says "eigenvalue $n+k$ at $t = 1$ is the same state as eigenvalue $n+k$ at $t = 0$". The mismatch is k rungs of an infinite ladder.

10.10.8 The index on the torus, computed

Nash ends his example with the remark that the integer k "is also the index of the chiral Dirac operator $\not{D}_A$ on the torus. To check this explicitly one would have to calculate the null spaces of $\not{D}_A$ and $\not{D}^*_A$ on the torus using" [Nash (10.85)]

$$\not{D}_A = \not{D}^1_t + \frac{\partial}{\partial t} = \frac{\partial}{\partial t} - i\frac{\partial}{\partial \theta} + kt. \tag{10.237}$$

This is the $n+p$ operator (10.225) for our family, as we saw in Example 10.59. It acts on sections of the charge-k line bundle of Example 10.67, so its domain is functions obeying the twisted periodicity (10.232). We now do the calculation.

Worked example 10.69: The zero modes of $\not{D}_A$ are theta functions

Step 1: reduce to ordinary differential equations. Expand in Fourier modes of the spatial circle, $\psi(t,\theta) = \sum_{n\in\mathbb{Z}} c_n(t)e^{in\theta}$. Then

$$\not{D}_A\psi = 0 \overset{(1)}{=} \sum_n \big(\dot{c}_n + (n+kt)c_n\big)e^{in\theta} = 0$$

$$\iff \dot{c}_n \overset{(2)}{=} -\lambda_n(t)\,c_n$$

$$\Longleftrightarrow\quad c_n(t) \overset{(3)}{=} a_n \exp\!\Big(-\frac{(n+kt)^2}{2k}\Big) \qquad (k\neq 0).$$

(1) Apply (10.237) term by term; $e^{in\theta}$ is an eigenfunction of $-i\partial_\theta + kt$ with eigenvalue $\lambda_n(t) = n + kt$.
(2) The exponentials are independent, so each coefficient vanishes. Each Fourier mode evolves in t with its own instantaneous eigenvalue.
(3) Integrate: $-\int \lambda_n \mathrm{d}t = -(nt + \frac{k}{2}t^2) = -\frac{(n+kt)^2}{2k} + \frac{n^2}{2k}$; absorb the constant $e^{n^2/2k}$ into the constant a_n.

Step 2: impose the twisted periodicity. The condition $\psi(t+1,\theta) = e^{-ik\theta}\psi(t,\theta)$ reads

$$\sum_n c_n(t+1)e^{in\theta} \overset{(1)}{=} \sum_m c_m(t)e^{i(m-k)\theta} \overset{(2)}{=} \sum_n c_{n+k}(t)e^{in\theta},$$

$$\text{so}\quad c_n(t+1) = c_{n+k}(t) \quad\Longleftrightarrow\quad a_n e^{-(n+k+kt)^2/2k} \overset{(3)}{=} a_{n+k}e^{-(n+k+kt)^2/2k}.$$

(1) Multiply the Fourier series by $e^{-ik\theta}$.
(2) Rename $m = n + k$.
(3) Insert the solution of Step 1: in $c_n(t+1)$ the exponent contains $n+k(t+1) = (n+k)+kt$, the same as in $c_{n+k}(t)$.

So $a_{n+k} = a_n$: the coefficients depend only on n mod k. For $k > 0$ there are exactly k free constants $a_0, \dots, a_{k-1}$, giving k solutions

$$\psi_r(t,\theta) = \sum_{j\in\mathbb{Z}} \exp\!\Big(-\frac{(r+jk+kt)^2}{2k}\Big)\, e^{i(r+jk)\theta}, \qquad r = 0,1,\dots,k-1. \tag{10.238}$$

Step 3: they are normalisable. Using $|e^{i(r+jk)\theta}| = 1$ and the orthogonality of different Fourier modes,

$$\|\psi_r\|^2 \overset{(1)}{=} 2\pi \sum_{j\in\mathbb{Z}} \int_0^1 \exp\!\Big(-\frac{(r+jk+kt)^2}{k}\Big)\mathrm{d}t \overset{(2)}{=} 2\pi \int_{-\infty}^{\infty} \exp\!\Big(-\frac{(r+ks)^2}{k}\Big)\mathrm{d}s \overset{(3)}{=} 2\pi\sqrt{\frac{\pi}{k}}.$$

(1) Integrate $|\psi_r|^2$ over $\theta \in [0, 2\pi]$: cross terms between different j integrate to zero.
(2) Put $s = t + j$: the sum over j of integrals over unit intervals is one integral over the whole line.
(3) Gaussian integral: substitute $u = r + ks$, $\int e^{-u^2/k}\mathrm{d}u/k = \sqrt{\pi k}/k$.

Step 4: the adjoint. $\not{D}_A^* = -\partial_t - i\partial_\theta + kt$ (the t-derivative changes sign, $\not{D}_t^1$ is self-adjoint). The same steps give $\dot{c}_n = +\lambda_n c_n$, i.e. $c_n \propto \exp(+(n+kt)^2/2k)$, which grows like a Gaussian in the wrong direction: for $k > 0$ the sum does not converge and there is no normalisable solution. Hence

$$\operatorname{index}\not{D}_A = \dim\operatorname{Ker}\not{D}_A - \dim\operatorname{Ker}\not{D}_A^* = k - 0 = k \qquad (k > 0).$$

For $k < 0$ the roles swap: $\operatorname{Ker}\not{D}_A = 0$ and $\dim\operatorname{Ker}\not{D}_A^* = |k|$, so again $\operatorname{index}\not{D}_A = k$. For $k = 0$ the bundle is trivial, both kernels are the constants, and the index is 0. In every case the index is the Chern number, as the index theorem said, and it equals the spectral flow of the Hamiltonian family.

Why they are called theta functions. Write $w = \theta + it$. Expanding the square in (10.238),

$$\psi_r = e^{-kt^2/2} \sum_{n \equiv r \ (k)} e^{-n^2/2k}\, e^{in(\theta+it)} = e^{-kt^2/2} \sum_{n \equiv r \ (k)} e^{-n^2/2k}\, e^{inw}.$$

Apart from the Gaussian prefactor, this is a holomorphic function of w, a **Jacobi theta function** with characteristic r/k. That is no accident: in the complex coordinate w the chiral Dirac operator on a surface is a $\bar\partial$-operator, $\partial_t - i\partial_\theta = 2i\,\partial_{\bar w}$ (check: $\partial_{\bar w} = \frac{1}{2}(\partial_\theta + i\partial_t)$), so its zero modes are holomorphic sections of the line bundle. The prefactor $e^{-kt^2/2}$ accounts for the term kt in (10.237): it is the change of trivialisation from the unitary gauge $a = kt\,\mathrm{d}\theta$ to a holomorphic one. The Riemann–Roch count of holomorphic sections of a degree k line bundle on a torus, $\dim H^0 = k$ (Chapter 5), is our count of k theta functions.

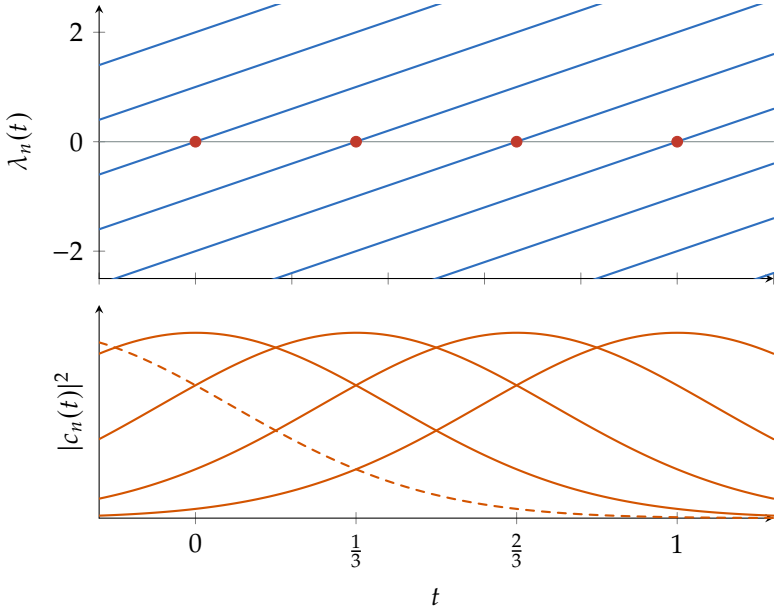


Figure 10.48: Index = spectral flow, seen in the zero modes ($k = 3$). Top: the spectral-flow diagram of $\not{D}^1_t$; eigenvalues cross zero at $t = 0, \frac{1}{3}, \frac{2}{3}, 1, \ldots$ (red). Bottom: the Fourier coefficients $|c_n(t)|^2 = e^{-(n+kt)^2/k}$ of the zero modes of $\not{D}_A$ on the torus. Each is a Gaussian centred exactly at the time when its eigenvalue $\lambda_n(t)$ crosses zero. One period of t contains k crossings, and the twisted periodicity glues the Gaussians into k independent theta functions.

10.10.9 Why the index equals the spectral flow

The picture in Figure 10.48 is general. Let D_t be a family of self-adjoint operators on a compact $(n-1)$-manifold M, depending on $t \in \mathbb{R}$ and approaching fixed invertible operators $D_\pm$ as $t \to \pm\infty$. On the cylinder $\mathbb{R} \times M$ consider

$$\mathcal{D} = \frac{\partial}{\partial t} + D_t.$$

Key idea: Index = spectral flow (Atiyah–Patodi–Singer)

index $\mathcal{D}$ = sf(D_t): the index of the n-dimensional operator $\partial_t + D_t$ is the net number of eigenvalues of the $(n-1)$-dimensional family that cross zero from below. For a loop (or a loop up to gauge transformation, as in Nash's example) the same holds on $S^1 \times M$, with the twisted boundary condition.

Why it is true (the adiabatic argument). Suppose first that D_t changes slowly and that its eigenvectors $\phi_n(t)$ (eigenvalues $\lambda_n(t)$) change slowly too. Look for zero modes of the form $\psi = c(t)\phi_n(t)$, neglecting $\dot\phi_n$ (the adiabatic approximation, as in the Berry-phase calculation (10.224)):

$$\mathcal{D}\psi = 0 \overset{(1)}{=} \dot c\,\phi_n + c\,\lambda_n(t)\phi_n \approx 0 \quad\Longrightarrow\quad c(t) \overset{(2)}{=} c(0)\exp\Big(-\int_0^t \lambda_n(t')\,\mathrm{d}t'\Big).$$

(1) $\partial_t(c\phi_n) = \dot c\phi_n + c\dot\phi_n$, and we drop $c\dot\phi_n$; $D_t\phi_n = \lambda_n\phi_n$.
(2) Solve the linear equation $\dot c = -\lambda_n c$.

When is c square integrable on $\mathbb{R}$? The exponent is $-\int\lambda_n$. As $t\to+\infty$ it must go to $-\infty$, so we need $\lambda_n > 0$ at late times; as $t\to-\infty$ we need $\int_0^t\lambda_n\to+\infty$ for $t\to-\infty$, which requires $\lambda_n<0$ at early times. So:

- a zero mode of $\mathcal{D}$ for every eigenvalue that goes from negative to positive;
- a zero mode of $\mathcal{D}^* = -\partial_t + D_t$ (where $c = e^{+\int\lambda_n}$) for every eigenvalue that goes from positive to negative;
- nothing for eigenvalues that do not change sign (the Gaussian grows at one end).

Subtracting, $\operatorname{index}\mathcal{D} = \#(-\to+) - \#(+\to-) = \operatorname{sf}(D_t)$. The simplest case $\lambda(t) = kt$ gives $c = e^{-kt^2/2}$: exactly the Gaussians of Example 10.69. The general statement (no adiabatic assumption) follows because both sides are deformation invariant: we may always deform the family to a slowly varying one without changing either the index or the spectral flow.

Watch out: Which way is "positive"?

The sign of the answer depends on writing $\partial_t + D_t$ rather than $\partial_t - D_t$, and on the orientation $\mathrm{d}t\wedge\mathrm{d}\theta$ of the torus. Flip either and the index becomes $-\operatorname{sf}$. What is convention independent is the equality of two integers, up to one overall sign fixed once and for all. With our choices, Nash's example gives index $=$ sf $= c_1 = k$.

10.10.10 The Dirac sea: spectral flow creates charge

So far the fermions were first quantised: $\not{D}^1_t$ is a one-particle Hamiltonian. To see the anomaly as a physicist does, we second quantise, and the spectral flow turns into a *non-conservation of charge*: the Hamiltonian form of the Adler–Bell–Jackiw anomaly.

The Dirac sea. Let H be the one-particle Hilbert space with a self-adjoint Hamiltonian h whose spectrum has no zero. Split $H = H^+\oplus H^-$ into the positive and negative spectral subspaces. Dirac's prescription for the ground state of many non-interacting fermions is: *fill every negative-energy level, leave every positive-energy level empty*. For a finite-dimensional H with H^- spanned by orthonormal eigenvectors $\phi_1,\dots,\phi_m$, the many-body ground state is the Slater determinant

$$|\Omega\rangle = \phi_1\wedge\phi_2\wedge\cdots\wedge\phi_m \in \Lambda^m(H).$$

Changing the basis of H^- by a unitary matrix u multiplies $|\Omega\rangle$ by $\det u$; so the *line* through $|\Omega\rangle$ is canonical, it is the line $\Lambda^m(H^-) = \det H^-$, but the vector is not. This is the first appearance of a determinant line in the Hamiltonian picture, and it is the many-body version of the Berry line of Example 10.58. (For a single spin, $H=\mathbb{C}^2$ and $m=1$: the "sea" is the ground state $|\psi(\vec n)\rangle$ itself.)

Adiabatic evolution with spectral flow. Now let the one-particle Hamiltonian be Nash's $\not{D}_t^1$ with t slowly increasing from 0 to 1. Start in the Dirac sea of $\not{D}_0^1$: modes with $\lambda_n(0) = n < 0$ filled. (Resolve the zero mode at $t = 0$ with the δ-shift of the working definition: the mode $n = 0$ is empty.) By the adiabatic theorem each one-particle level is carried along with its occupation number. At $t = 1$:

- the levels $n \le -k-1$ have $\lambda_n(1) = n + k \le -1$: still negative, still filled;
- the k levels $n = -k, \dots, -1$ have $\lambda_n(1) = n + k \in \{0, \dots, k-1\}$: now non-negative, but *still filled*;
- the levels $n \ge 0$ are positive and empty.

The Dirac sea of $\not{D}_1^1$ would have every negative level filled and every non-negative level empty. So the evolved state is the new vacuum *plus* k *particles* in the lowest k positive levels. The field has created k units of charge out of the vacuum (Figure 10.49).

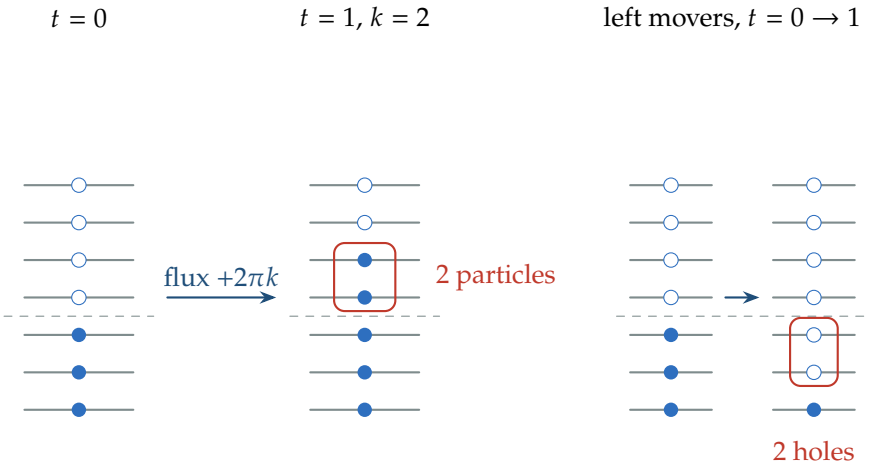


Figure 10.49: The Dirac sea under spectral flow. Levels are drawn by energy (the dashed line is zero energy); filled dots are occupied. Right movers (left two columns): every level climbs $k = 2$ rungs and keeps its occupation, so two occupied levels end up above zero, i.e. two particles are created. Left movers (right panel): the levels move down, two empty levels end up below zero, i.e. two holes (antiparticles) are created. The total charge $N_R + N_L$ is conserved, the chiral charge $N_R - N_L$ changes by $2k$.

Worked example 10.70: The anomaly equation of QED$_2$ from counting levels

Take a massless Dirac fermion on a spatial circle of length 2π: a right mover with $h_R = -i\partial_\theta + a$ and a left mover with $h_L = +i\partial_\theta - a$, where $a(t)$ is the (spatially constant) gauge potential, so the electric field is $E = \dot{a}$ (up to a convention-dependent sign that we fix this way). Let a grow from 0 to k. Count the charges:

$$\Delta N_R \overset{(1)}{=} \mathrm{sf}(h_R) = +k, \qquad \Delta N_L \overset{(2)}{=} \mathrm{sf}(h_L) = -k,$$
$$\Delta Q = \Delta N_R + \Delta N_L = 0, \qquad \Delta Q_5 = \Delta N_R - \Delta N_L = 2k,$$
$$2k \overset{(3)}{=} 2 \cdot \frac{1}{2\pi} \int \mathrm{d}t \oint \mathrm{d}\theta\, \dot{a} \overset{(4)}{=} \frac{1}{\pi} \int \mathrm{d}t \oint \mathrm{d}\theta\, E.$$

(1) h_R is Nash's family, with eigenvalues $n + a$ rising by k; by the adiabatic argument each $(- \to +)$ crossing puts one particle above the new sea.

(2) The eigenvalues of h_L are $-n - a$: they *fall* by k. Each $(+ \to -)$ crossing counts -1 in

sf and takes an empty level below zero, creating a hole, which is charge -1 relative to the new sea. So $\Delta N_L = \mathrm{sf}(h_L) = -k$.

(3) $\oint \mathrm{d}\theta \int \dot{a}\,\mathrm{d}t = 2\pi\,\Delta a = 2\pi k$.

(4) Definition $E = \dot{a}$.

So the vector charge is conserved and the axial charge obeys $\Delta Q_5 = \frac{1}{\pi}\int E$, the integrated form of the local law $\partial_\mu j_5^\mu = \frac{e}{\pi}E = \frac{e}{2\pi}\epsilon^{\mu\nu}F_{\mu\nu}$ (with $e = 1$), which the QFT notes derive from the vacuum polarisation diagram [QFT notes §22.2, (22.2.27)]. No Feynman diagram was needed: the anomaly *is* the spectral flow. If only the right mover were present and coupled to a (a chiral gauge theory), the gauge charge itself would not be conserved: that is a gauge anomaly, and it is the situation that the rest of this section analyses.

The vacuum charge and the eta invariant. How can a charge "appear" in an adiabatic process where every level keeps its occupation? Because the charge of an infinite Dirac sea is itself an infinite quantity that must be regularised, and the regularised value changes continuously. The charge operator, ordered symmetrically between particles and holes, is

$$Q = \tfrac{1}{2}\sum_n \left(a_n^\dagger a_n - a_n a_n^\dagger\right),$$

where a_n annihilates a fermion in the level λ_n. On the Dirac sea each filled level contributes $+\frac{1}{2}$ and each empty level $-\frac{1}{2}$:

$$Q_{\mathrm{vac}} \overset{(1)}{=} \tfrac{1}{2}\sum_{\lambda_n<0} 1 - \tfrac{1}{2}\sum_{\lambda_n>0} 1 \overset{(2)}{=} -\tfrac{1}{2}\sum_n \mathrm{sgn}(\lambda_n) \overset{(3)}{=} -\tfrac{1}{2}\,\eta_h(0). \tag{10.239}$$

(1) For a filled level $a^\dagger a = 1$, $aa^\dagger = 0$; for an empty level the reverse.

(2) Collect the two sums (no zero modes).

(3) Regularise the sum of signs by the eta function, exactly as in Section 10.9.11: $\sum \mathrm{sgn}\,\lambda_n |\lambda_n|^{-s}$ at $s = 0$.

For the right mover $h_R = -i\partial_\theta + a$ with $0 < a < 1$, Example 10.52 gave $\eta = 1 - 2a$, so

$$Q_{\mathrm{vac}}(a) = a - \tfrac{1}{2} \qquad (0 < a < 1).$$

Follow the adiabatically evolved state from $a = 0^+$ to $a = 1^-$: its charge rises smoothly from $-\frac{1}{2}$ to $+\frac{1}{2}$. At $a = 1$ an eigenvalue crosses zero, the ground state of $h_R(1)$ is again a sea of charge $-\frac{1}{2}$, and the evolved state, still of charge $+\frac{1}{2}$, is that new vacuum plus one particle (Figure 10.50). The smooth part of Q is *local*, $\mathrm{d}Q_{\mathrm{vac}}/\mathrm{d}a = 1 = \frac{1}{2\pi}\oint \mathrm{d}\theta$, the density of the anomaly; the integer jumps are the spectral flow.

Watch out: Charge is created, but nothing is non-unitary

Nothing in the evolution violates quantum mechanics. The many-body state evolves unitarily; only the *labelling* of states as "sea" or "particle" is tied to the instantaneous Hamiltonian, and after a spectral flow the old sea is not the new sea. In a vector-like theory (both chiralities with the same charge) the creation of right-moving particles is compensated by the creation of left-moving holes. In a chiral theory it is not, and the gauge charge is not conserved. Then Gauss's law, which says that the charge in a region equals the electric flux out of it, cannot hold as an operator statement. The rest of this section turns that sentence into mathematics.

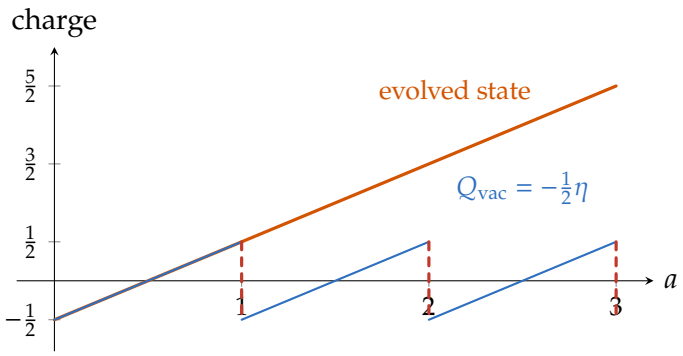


Figure 10.50: Charge of the Dirac sea of $h_R = -i\,\mathrm{d}/\mathrm{d}\theta + a$ as the flux a grows. Blue: the regularised vacuum charge $-\frac{1}{2}\eta(a)$, a sawtooth that drops back by one unit (red) each time an eigenvalue crosses zero. Orange: the charge of the state obtained by adiabatic evolution from the vacuum at $a = 0$; it rises steadily, one unit per flux quantum. The gap between the orange line and the blue sawtooth is the number of particles created: the spectral flow.

Reading Nash (pp. 295–296, (10.80)–(10.85))

Nash's example is "Spectral flow and an anomaly in two dimensions". (10.80) is our (10.233); (10.81)–(10.83) are (10.234)–(10.235); his remark that "k of the negative eigenvalues of $\not{D}_0^1$ have flowed up and become positive eigenvalues of $\not{D}_1^1$" is Figure 10.47. (10.84) is the gauge covariance (10.236); his sentence "if we quotient by this gauge transformation the family descends to a periodic family" is our loop up to gauge, closed in $\mathcal{F}^1$ by Kuiper's theorem. For (10.85) Nash only says that one "would have to calculate the null spaces"; we did so in Example 10.69. His closing warning, "if the Dirac operator had a lowest bound no flow could take place", is Example 10.68. The Dirac-sea interpretation, the QED_2 anomaly and the vacuum charge are not in Nash; they are the physics that makes his statement "anomalies should not disappear in an alternative formulation" concrete.

10.10.11 Example: Fock space and Gauss's law. The set-up

The two-dimensional example gave a *one*-parameter family and the group $\pi_1(\mathcal{F}^1) = \mathbb{Z}$. Nash's second example [Nash p. 297] is the four-dimensional non-abelian anomaly of Chapter 10's §2, seen from both sides. The question it answers is the one we started this section with: in canonical quantisation, what exactly goes wrong?

Two ways of seeing the same integer. Take a gauge group $K = \mathrm{SU}(N)$ with $N \geq 3$ and chiral fermions in the fundamental representation.

- **Euclidean.** Spacetime is S^4 ($n = 4$). The anomaly of §2 was the first Chern class of the determinant line bundle, evaluated on a 2-sphere in $\mathcal{A}/\mathcal{G}$. So we need a family of chiral Dirac operators $\not{D}_y^{(4)}$ with $y \in S^2$ ($p = 2$). The family is a map $S^2 \to \mathcal{F}$, and its class lives in $\pi_2(\mathcal{F}) = \mathbb{Z}$.
- **Hamiltonian.** Space is S^3 and time is extra ($n - 1 = 3$). The operators are the self-adjoint Dirac Hamiltonians $\not{D}_A^{(3)}$ on S^3, so a family is a map into $\mathcal{F}^1$, and to see an integer we need $\pi_p(\mathcal{F}^1) = \mathbb{Z}$, i.e. p odd. Nash takes $p = 3$: a family over $S^3 \subset \mathcal{A}_3/\mathcal{G}_3$.

The bookkeeping rule of Section 10.10.5 says these must agree: $n + p = 4 + 2 = 6$ in the Euclidean picture and $(n - 1) + p = 3 + 3 = 6$ in the Hamiltonian one. Both are measured by

$\pi_6(\Omega U(\infty)) = \pi_7(U(\infty)) = \mathbb{Z}$. Nash's summary is that "the cohomology generators shift from even to odd dimensions and the 2-sphere in $\mathcal{A}/\mathcal{G}$ becomes a 3-sphere".

Why a 2-sphere becomes a 3-sphere. The gauge-theory side of the same statement can be computed directly. Let $\mathcal{G}_m = \mathrm{Map}_*(S^m, K)$ be the based gauge group on S^m (gauge transformations equal to $\mathbb{1}$ at a base point, so that $\mathcal{G}_m$ acts freely on the connections $\mathcal{A}_m$). Then

$$\begin{aligned} \pi_2(\mathcal{A}_4/\mathcal{G}_4) &\overset{(1)}{=} \pi_1(\mathcal{G}_4) \overset{(2)}{=} \pi_1\big(\mathrm{Map}_*(S^4, K)\big) \overset{(3)}{=} \pi_5(K) \overset{(4)}{=} \mathbb{Z}, \\ \pi_3(\mathcal{A}_3/\mathcal{G}_3) &\overset{(5)}{=} \pi_2(\mathcal{G}_3) \overset{(6)}{=} \pi_2\big(\mathrm{Map}_*(S^3, K)\big) \overset{(7)}{=} \pi_5(K) = \mathbb{Z}. \end{aligned} \tag{10.240}$$

(1) The fibration $\mathcal{G} \to \mathcal{A} \to \mathcal{A}/\mathcal{G}$ has contractible total space ($\mathcal{A}$ is an affine space), so its long exact homotopy sequence collapses to $\pi_k(\mathcal{A}/\mathcal{G}) \cong \pi_{k-1}(\mathcal{G})$.
(2) Definition of $\mathcal{G}_4$.
(3) A based loop of based maps $S^4 \to K$ is a based map $S^1 \wedge S^4 = S^5 \to K$: $\pi_k(\mathrm{Map}_*(S^m, K)) = \pi_{k+m}(K)$.
(4) Bott: $\pi_5(\mathrm{SU}(N)) = \mathbb{Z}$ for $N \geq 3$ (the stable range, since $\pi_5(\mathrm{U}(\infty)) = \mathbb{Z}$).
(5) As in step 1, now on S^3.
(6) Definition of $\mathcal{G}_3$.
(7) As in step 3, with $k = 2$, $m = 3$: $2 + 3 = 5$.

So the same group $\pi_5(\mathrm{SU}(N))$, whose generator is behind the non-abelian anomaly in four dimensions, appears once as a 2-sphere in the Euclidean orbit space and once as a 3-sphere in the Hamiltonian one. The dimension of space went down by one, the dimension of the sphere went up by one, and the sum stayed fixed. This is the $n + p$ rule seen from the gauge group rather than from the operators. (For $K = \mathrm{SU}(2)$, $\pi_5(\mathrm{SU}(2)) = \mathbb{Z}_2$: no integer, so no perturbative anomaly, in line with Section 10.9.)

Physical picture: What "anomaly" will mean in the Hamiltonian picture

In the Euclidean picture the anomaly was a *line bundle* over $\mathcal{A}/\mathcal{G}$ with a non-zero Chern class in H^2. In the Hamiltonian picture Nash will find a *projective* bundle of Fock spaces over $\mathcal{A}/\mathcal{G}$ whose obstruction lives in $H^3(\mathcal{A}/\mathcal{G}; \mathbb{Z}) = \mathbb{Z}$ (one degree higher, matching the 3-sphere of (10.240)). Seen from the gauge group, it is a 2-cocycle on $\mathcal{G}$: the gauge group acts on states only up to phases. Infinitesimally, it is a Schwinger term. Physically, Gauss's law cannot be imposed.

10.10.12 The fermionic Fock space

In canonical quantisation of a gauge theory in temporal gauge $A_0 = 0$, the bosonic degree of freedom is the spatial connection $A \in \mathcal{A}$ on space (here S^3). A state is a wave functional $\Psi[A]$, just as a state of a particle is a wave function $\psi(x)$. With fermions present, $\Psi[A]$ is not a number: for each A it is a vector in the space of states of the fermions moving in the background A. That space is the fermionic Fock space $\mathfrak{F}(A)$ built on the one-particle Dirac Hamiltonian $\not{D}_A$. So we need Fock spaces first. We build them in finite dimensions, where every step is a matrix computation, and then see what changes when the one-particle space is infinite-dimensional.

Exterior algebras are many-fermion spaces. For a finite-dimensional one-particle space V with orthonormal basis $e_1, \dots, e_d$, the state with fermions in the levels $i_1 < \cdots < i_k$ is the antisymmetrised product $e_{i_1} \wedge \cdots \wedge e_{i_k}$. The Pauli principle is $e_i \wedge e_i = 0$. So the space of

k-fermion states is $\Lambda^k(V)$, of dimension $\binom{d}{k}$, and the space of all states is the exterior algebra

$$\Lambda^*(V) = \bigoplus_{k=0}^{d} \Lambda^k(V), \qquad \dim \Lambda^*(V) = 2^d.$$

Creation and annihilation operators are wedging and contracting: $a^\dagger(v)\omega = v \wedge \omega$ and $a(v) = \iota_v$, the contraction with $\langle v, \cdot \rangle$. They obey the canonical anticommutation relations $\{a(u), a^\dagger(v)\} = \langle u|v\rangle$, $\{a(u), a(v)\} = 0$. (Check on $V = \mathbb{C}$: $a^\dagger$ maps $1 \mapsto e$ and $e \mapsto 0$, a maps $e \mapsto 1$ and $1 \mapsto 0$, so $aa^\dagger + a^\dagger a = \mathbb{1}$: the 2×2 matrices $\left(\begin{smallmatrix}0&0\\1&0\end{smallmatrix}\right)$ and $\left(\begin{smallmatrix}0&1\\0&0\end{smallmatrix}\right)$ in the basis $(1, e)$.)

Particles and antiparticles. Now let $V = H$ carry a one-particle Hamiltonian h with no zero eigenvalue and split $H = H^+ \oplus H^-$ into its positive and negative spectral subspaces [Nash (10.86)]. Dirac's ground state fills H^-. Relative to it, an excitation is either a fermion added to a level of H^+ (a particle) or a fermion removed from a level of H^- (a hole, i.e. an antiparticle). Holes are described by the *complex conjugate* space $\bar{H}^-$: if the level $w \in H^-$ has the wave function $w(x)$, the hole has $\overline{w(x)}$, the opposite charge and the energy $-\lambda > 0$. The particle–antiparticle Fock space is Nash's (10.87):

Working definition 10.18 (Fermionic Fock space of a polarised Hilbert space)

Given $H = H^+ \oplus H^-$ (a **polarisation** [the split of the one-particle states into the positive-energy levels H^+ and negative-energy levels H^- of h]),

$$\mathfrak{F} = \Lambda^*\big(H^+ \oplus \bar{H}^-\big) = \bigoplus_{p,q\geq 0} \Lambda^p(H^+) \otimes \Lambda^q(\bar{H}^-). \tag{10.241}$$

The summand $\Lambda^p(H^+) \otimes \Lambda^q(\bar{H}^-)$ holds the states with p particles and q antiparticles. Its charge is $Q = p - q$, its energy is $\sum_{\text{particles}} \lambda + \sum_{\text{holes}} |\lambda|$, and the vacuum $|0\rangle = 1 \otimes 1 \in \Lambda^0 \otimes \Lambda^0$ is the unique state of lowest energy. To compute with it: pick an eigenbasis of h; a basis of $\mathfrak{F}$ is labelled by a finite set of occupied positive levels and a finite set of empty negative levels.

The second equality in (10.241) is the exterior algebra of a direct sum, $\Lambda^k(U \oplus W) = \bigoplus_{p+q=k} \Lambda^p(U) \otimes \Lambda^q(W)$: a basis k-vector of $U \oplus W$ is a wedge of p basis vectors of U and q of W, and can be reordered (up to sign) with the U factors first.

Worked example 10.71: A three-level Dirac sea

Let $H = \mathbb{C}^3$ with $h = \mathrm{diag}(-2, -1, +1)$ and eigenbasis e_1, e_2, e_3. Then $H^- = \mathrm{span}(e_1, e_2)$ and $H^+ = \mathrm{span}(e_3)$. The Fock space (10.241) decomposes by (p, q) with $p \in \{0, 1\}$, $q \in \{0, 1, 2\}$:

(p,q)	basis	dim	$Q = p - q$	energy	in $\Lambda^*(H)$
$(0,0)$	1	1	0	0	$e_1 \wedge e_2$
$(0,1)$	$\bar{e}_1,\ \bar{e}_2$	2	-1	2, 1	$e_2,\ -e_1$
$(0,2)$	$\bar{e}_1 \wedge \bar{e}_2$	1	-2	3	1
$(1,0)$	e_3	1	$+1$	1	$e_3 \wedge e_1 \wedge e_2$
$(1,1)$	$e_3 \otimes \bar{e}_1,\ e_3 \otimes \bar{e}_2$	2	0	3, 2	$e_3 \wedge e_2,\ -e_3 \wedge e_1$
$(1,2)$	$e_3 \otimes \bar{e}_1 \wedge \bar{e}_2$	1	-1	4	e_3

The total dimension is $1 + 2 + 1 + 1 + 2 + 1 = 8 = 2^3 = \dim \Lambda^*(\mathbb{C}^3)$. The last column shows the same state written as "which levels are occupied" in the exterior algebra of the whole of H, obtained by starting from the sea $e_1 \wedge e_2$ and removing the holes (contracting) and

adding the particles (wedging). Grouping by charge: $Q = -2, -1, 0, 1$ has dimensions $1, 3, 3, 1$, the dimensions of $\Lambda^0, \Lambda^1, \Lambda^2, \Lambda^3$ of H, as it must be, since "charge Q relative to a sea of 2" means "$2 + Q$ fermions".

The dependence on the sea is a line. The last column of the table used the sea vector $\Omega = e_1 \wedge e_2$. Had we used another orthonormal basis of H^-, related by $u \in \mathrm{U}(2)$, the sea vector would be $(\det u)\,\Omega$. So what is canonical is a map that needs one vector of the line $\det H^- = \Lambda^{\mathrm{top}}(H^-)$ as input:

$$\Lambda^p(H^+) \otimes \Lambda^q(\bar{H}^-) \otimes \det H^- \overset{(1)}{\cong} \Lambda^p(H^+) \otimes \Lambda^{m-q}(H^-) \xrightarrow{\wedge} \Lambda^{p+m-q}(H),$$

$$\text{so} \quad \mathfrak{F} \otimes \det H^- \overset{(2)}{\cong} \Lambda^*(H) \qquad (m = \dim H^-). \tag{10.242}$$

(1) Contracting a top form of H^- with q hole vectors $\bar{w}_1, \dots, \bar{w}_q$ (i.e. with $\iota_{w_1} \cdots \iota_{w_q}$) leaves an $(m-q)$-vector in H^-: "the sea with q levels emptied". This is an isomorphism because $\bar{H}^- \cong (H^-)^*$ via the inner product, and $\Lambda^q(V^*) \otimes \det V \cong \Lambda^{m-q}(V)$ is the Hodge-type duality of an m-dimensional space.

(2) Wedging with the particles and summing over p, q gives every degree $k = p+m-q$ of $\Lambda^*(H)$ exactly once, with matching dimensions $\sum_{p-q=k-m} \binom{\dim H^+}{p}\binom{m}{q} = \binom{\dim H}{k}$ (Vandermonde's identity).

So in finite dimensions $\mathfrak{F} \cong \Lambda^*(H) \otimes (\det H^-)^{-1}$: the particle–antiparticle Fock space is the "which levels are filled" space, tensored with a line. The line is the space of possible vacuum vectors. Projectively, $\mathfrak{F}$ and $\Lambda^*(H)$ are the same, and the same for every polarisation.

What changes in infinite dimensions. For a Dirac Hamiltonian H^- is infinite-dimensional (the spectrum is unbounded below), and "$e_1 \wedge e_2 \wedge \cdots$" with infinitely many factors is not a vector of any Hilbert space; nor is $\det H^-$ a well-defined line. So $\Lambda^*(H)$ is meaningless, and (10.241) is the only definition available. It depends on the polarisation: two splittings $H_1^+ \oplus H_1^-$ and $H_2^+ \oplus H_2^-$ give Fock spaces that can be compared only if the two negative subspaces are "close" (we make this precise in Section 10.10.18). When they differ by a finite-dimensional space, the comparison costs exactly the finite-dimensional line of the last paragraph. This is the whole mechanism of Nash's argument.

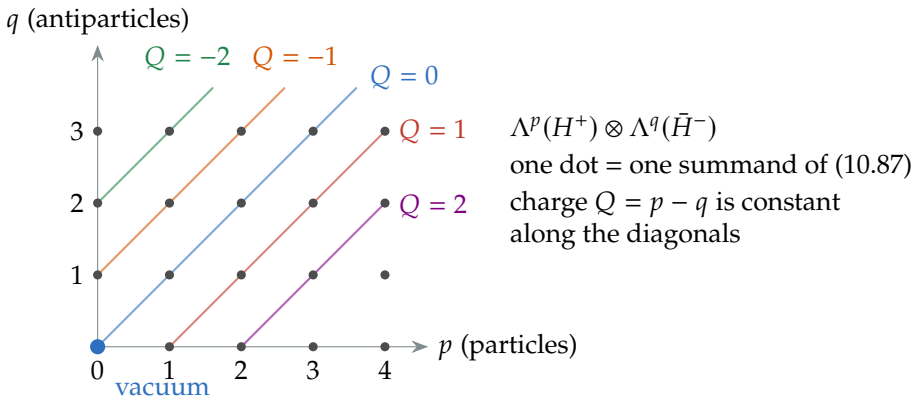


Figure 10.51: The Fock space (10.241) as a grid. Each dot (p, q) is the summand with p particles and q antiparticles; the vacuum is the blue dot at the origin. States of fixed charge lie on the coloured diagonals, and each diagonal is infinite: charge 0 contains the vacuum, all particle–hole pairs, and so on. Changing the polarisation by a k-dimensional space relabels the charges: every diagonal moves by k (Example 10.73).

Fock spaces over the space of connections. Put back the gauge field. For each spatial connection A, the one-particle Hamiltonian is the self-adjoint Dirac operator $\not{D}_A$ on S^3. Suppose for the moment that $\ker \not{D}_A = 0$. Then (10.241) with $H^{\pm} = H^{\pm}(A)$ gives a Fock space $\mathfrak{F}(A)$, and as A varies they form a bundle over $\mathcal{A}$. A gauge transformation $g \in \mathcal{G}$ maps A to $g \cdot A$ and conjugates the Dirac operator, $\not{D}_{g\cdot A} = g\, \not{D}_A\, g^{-1}$, so it maps eigenvectors of $\not{D}_A$ to eigenvectors of $\not{D}_{g\cdot A}$ with the same eigenvalues, and therefore maps $H^{\pm}(A) \to H^{\pm}(g \cdot A)$ and $\mathfrak{F}(A) \to \mathfrak{F}(g \cdot A)$. So we can hope to quotient [Nash (10.88)]:

$$\begin{array}{ccc} \mathfrak{F} & & \mathfrak{F}/\mathcal{G} \\ \downarrow & & \downarrow \\ \mathcal{A} & & \mathcal{A}/\mathcal{G} \end{array} \tag{10.243}$$

The physical states should be the square-integrable sections of $\mathfrak{F}/\mathcal{G} \to \mathcal{A}/\mathcal{G}$.

Dictionary: Why "sections of $\mathfrak{F}/\mathcal{G}$" is Gauss's law

physics	geometry
wave functional $\Psi[A]$ with values in fermion states	section of $\mathfrak{F} \to \mathcal{A}$
electric field $E = -i\,\delta/\delta A$	derivative along $\mathcal{A}$
Gauss's law $(D \cdot E - \rho)\Psi = 0$	Ψ constant along gauge orbits: $\Psi[g \cdot A] = U(g)\Psi[A]$
gauge-invariant wave functionals	sections of $\mathfrak{F}/\mathcal{G} \to \mathcal{A}/\mathcal{G}$

The middle row is the familiar statement that Gauss's law generates gauge transformations: the generator $G(\xi) = \int \operatorname{tr}\, \xi\,(D_i E_i - \rho)$ moves A by $D\xi$ (through $E = -i\delta/\delta A$) and rotates the fermions by ξ (through the charge density ρ). Requiring $G(\xi)\Psi = 0$ for all ξ says that Ψ is invariant under the combined transformation, i.e. that Ψ descends to the quotient. For this to make sense $\mathcal{G}$ must *act* on the Fock bundle: $U(g)U(h) = U(gh)$. Nash's point is that it does not.

10.10.13 The flaw: zero modes, and cutting the spectrum at ε

The construction assumed $\ker \not{D}_A = 0$. As A varies we cannot avoid connections where an eigenvalue passes through zero: that is exactly the spectral flow of Section 10.10.7, and it happens along a codimension-one wall in $\mathcal{A}$. On the wall the split $H = H^+ \oplus H^-$ is undefined, and across it $H^+(A)$ jumps by the crossing eigenvector. So the Fock spaces $\mathfrak{F}(A)$ do not form a smooth bundle over the whole of $\mathcal{A}$ [Nash p. 298].

Worked example 10.72: Where the polarisation jumps

Take Nash's circle family $\not{D}^1_t = -i\,\mathrm{d}/\mathrm{d}\theta + t$ with $k = 1$, so $\lambda_n(t) = n + t$. For $0 < t < 1$ the negative levels are $n \le -1$ and $H^+(t) = \overline{\operatorname{span}}\{e^{in\theta} : n \ge 0\}$. At $t = 1$ the level $n = -1$ reaches zero; for $1 < t < 2$ it is positive and $H^+(t) = \overline{\operatorname{span}}\{e^{in\theta} : n \ge -1\}$. So

$$H^+(1.5) = H^+(0.5) \oplus \mathbb{C}\, e^{-i\theta}.$$

The subspace jumps by one dimension at the wall $t = 1$, and nothing smooth connects the two Fock spaces through it. Everywhere else the subspace is constant (in this family the eigenvectors do not even move).

Nash's cure: move the cut. Nothing forces us to split the spectrum at zero. Choose a real number ε that is not an eigenvalue of $\not{D}_A$ and define [Nash (10.89)]

$$H_\varepsilon^+ = \text{span of eigenvectors with } \lambda > \varepsilon, \qquad H_\varepsilon^- = \text{span of eigenvectors with } \lambda < \varepsilon,$$

and the Fock space [Nash (10.90)]

$$\mathfrak{F}_\varepsilon(A) = \Lambda^*\big(H_\varepsilon^+ \oplus \bar{H}_\varepsilon^-\big). \tag{10.244}$$

The key observation is that this is smooth on a whole open set. Let

$$U_\varepsilon = \{A \in \mathcal{A} : \ \varepsilon \notin \operatorname{spec} \not{D}_A\}.$$

Eigenvalues depend continuously on A, so if ε is not an eigenvalue of $\not{D}_A$ it is not one for nearby A either: U_ε is open. On U_ε the spectral projection

$$P_\varepsilon^+(A) = \frac{1}{2\pi i} \oint_\Gamma \frac{\mathrm{d}z}{z - \not{D}_A}$$

(where Γ encircles the part of the spectrum above ε, crossing the real axis at ε and at $+\infty$) depends smoothly on A, because the contour stays away from the spectrum. So $H_\varepsilon^\pm(A)$ and $\mathfrak{F}_\varepsilon(A)$ form smooth bundles over U_ε. As ε ranges over $\mathbb{R}$ the sets U_ε cover $\mathcal{A}$: every A has a discrete spectrum, so some ε misses it.

Two cuts. On the overlap $U_\varepsilon \cap U_{\varepsilon'}$ we have two decompositions [Nash (10.89)], $H = H_\varepsilon^+ \oplus H_\varepsilon^- = H_{\varepsilon'}^+ \oplus H_{\varepsilon'}^-$, and two Fock spaces. Take $\varepsilon < \varepsilon'$ and let

$$J_{\varepsilon\varepsilon'}(A) = \text{span of eigenvectors of } \not{D}_A \text{ with } \varepsilon < \lambda < \varepsilon'.$$

It is *finite-dimensional*: $\not{D}_A$ is elliptic on a compact manifold, so it has only finitely many eigenvalues in a bounded interval (Chapter 2). Then:

$$H_\varepsilon^+ \overset{(1)}{=} H_{\varepsilon'}^+ \oplus J_{\varepsilon\varepsilon'}, \qquad H_{\varepsilon'}^- \overset{(2)}{=} H_\varepsilon^- \oplus J_{\varepsilon\varepsilon'},$$

$$\Longrightarrow \ H \overset{(3)}{=} H_{\varepsilon'}^+ \oplus J_{\varepsilon\varepsilon'} \oplus H_\varepsilon^-, \tag{10.245}$$

$$\mathfrak{F}_\varepsilon(A) \overset{(4)}{=} \Lambda^*\big(H_{\varepsilon'}^+ \oplus J_{\varepsilon\varepsilon'} \oplus \bar{H}_\varepsilon^-\big) \overset{(5)}{=} \Lambda^*(J_{\varepsilon\varepsilon'}) \otimes \Lambda^*\big(H_{\varepsilon'}^+ \oplus \bar{H}_\varepsilon^-\big), \tag{10.246}$$

$$\mathfrak{F}_{\varepsilon'}(A) \overset{(6)}{=} \Lambda^*\big(H_{\varepsilon'}^+ \oplus \bar{J}_{\varepsilon\varepsilon'} \oplus \bar{H}_\varepsilon^-\big) \overset{(7)}{=} \Lambda^*(\bar{J}_{\varepsilon\varepsilon'}) \otimes \Lambda^*\big(H_{\varepsilon'}^+ \oplus \bar{H}_\varepsilon^-\big). \tag{10.247}$$

(1) An eigenvalue above ε is either above ε' or in $(\varepsilon, \varepsilon')$; Nash's (10.91).
(2) An eigenvalue below ε' is either below ε or in $(\varepsilon, \varepsilon')$.
(3) Combine; the three pieces are mutually orthogonal (eigenspaces of a self-adjoint operator).
(4) Definition (10.244) with step 1.
(5) $\Lambda^*(U \oplus W) \cong \Lambda^*(U) \otimes \Lambda^*(W)$; Nash's (10.92).
(6) Definition with ε': now the J-levels are *below* the cut, so they are part of the sea, and their excitations are holes, described by $\bar{J}$. This is Nash's (10.93).
(7) As in step 5.

So the two Fock spaces share the big factor $\Lambda^*(H_{\varepsilon'}^+ \oplus \bar{H}_\varepsilon^-)$, and differ only in a finite factor: $\Lambda^*(J)$ versus $\Lambda^*(\bar{J})$ [Nash p. 299].

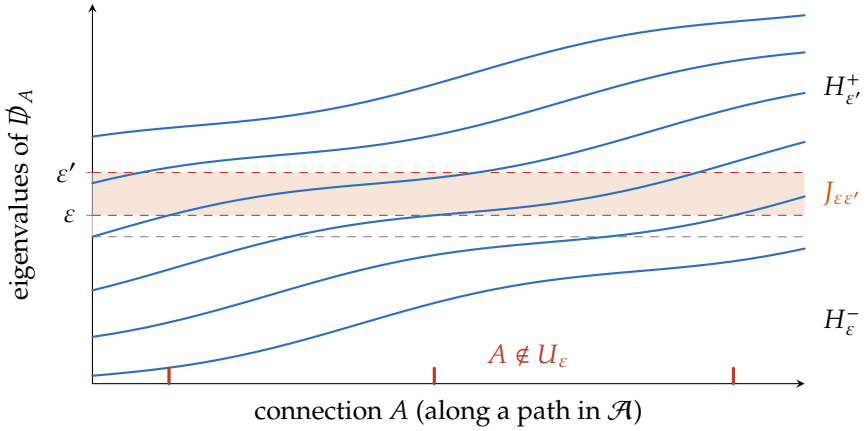


Figure 10.52: Eigenvalues of $\not{D}_A$ (blue) along a path of connections. A cut at ε (dashed red) defines a Fock space $\mathfrak{F}_\varepsilon(A)$ that is smooth as long as no eigenvalue touches the cut; the red ticks on the axis mark the connections where a line crosses ε, which are exactly the points outside U_ε. Between two cuts $\varepsilon < \varepsilon'$ lies the orange window: its eigenvectors span the finite-dimensional space $J_{\varepsilon\varepsilon'}(A)$. The number of lines in the window changes as lines enter and leave, but only where a line crosses one of the two cuts.

Watch out: The cut at ε is not "the vacuum".

For $\varepsilon \neq 0$ the vacuum $|0\rangle_\varepsilon$ of $\mathfrak{F}_\varepsilon$ is not the lowest energy state: if there are levels in $(0, \varepsilon)$ they are empty in $|0\rangle_\varepsilon$ but would be filled in the true ground state, or vice versa for $\varepsilon < 0$. That is harmless. We are not trying to identify the ground state; we are trying to build a smooth bundle of *state spaces*. The state space $\mathfrak{F}_\varepsilon$ contains the true ground state as one of its vectors (a state with finitely many particles or holes in the window).

10.10.14 Changing the cut costs a line: projective Fock spaces

We must compare $\Lambda^*(J)$ with $\Lambda^*(\bar{J})$, where $n = \dim J$. They are the Fock spaces of the *same* n levels, described once as "empty unless occupied" (particles above the cut ε) and once as "filled unless emptied" (holes below the cut ε'). A physical configuration of the n levels, "the levels in the set S are filled and the rest are empty", is described in the first language by the p particles $e_S = e_{s_1} \wedge \cdots \wedge e_{s_p} \in \Lambda^p(J)$ ($|S| = p$), and in the second by the $n - p$ holes $\bar{e}_{S^c} \in \Lambda^{n-p}(\bar{J})$ in the complementary levels. So the dictionary should send $\Lambda^p(J) \to \Lambda^{n-p}(\bar{J})$, and $e_S \mapsto (\text{number}) \cdot \bar{e}_{S^c}$. The only question is which number.

Working definition 10.19 (The isomorphism I^λ (Nash's (10.94)))

Let $e_1, \ldots, e_n$ be an orthonormal eigenbasis of $J = J_{\varepsilon\varepsilon'}(A)$ [the span of the eigenvectors of $\not{D}_A$ with eigenvalues between the two cuts ε and ε'] and choose a non-zero element of the line $\Lambda^n(J)$ [the multiples of $e_1 \wedge \cdots \wedge e_n$, the state with all n window levels filled], $\lambda = c\, e_1 \wedge \cdots \wedge e_n$ with $c \in \mathbb{C}^\times$. For each subset S define the sign $\sigma(S)$ by $e_{S^c} \wedge e_S = \sigma(S)\, e_1 \wedge \cdots \wedge e_n$, and set

$$I^\lambda_p : \Lambda^p(J) \to \Lambda^{n-p}(\bar{J}), \qquad I^\lambda_p(e_S) = c\, \sigma(S)\, \bar{e}_{S^c}, \tag{10.248}$$

extended linearly. Reading $\bar{e}_j$ as e_j, this is exactly Nash's condition $I^\lambda_p(\omega) \wedge \omega = \lambda$ on basis vectors $\omega = e_S$. It is an isomorphism (it maps a basis to a basis up to non-zero factors). Tensoring with the identity on the common factor [the levels outside the window, which both Fock spaces describe in the same way] gives $I^\lambda \otimes \mathbb{1} : \mathfrak{F}_\varepsilon(A) \to \mathfrak{F}_{\varepsilon'}(A)$.

Worked example 10.73: One level crosses the window

Let $n = \dim J = 1$, $J = \mathbb{C}e_1$. Then $\Lambda^*(J) = \mathrm{span}(1, e_1)$ and $\Lambda^*(\bar{J}) = \mathrm{span}(1, \bar{e}_1)$, and (10.248) gives

$$I^\lambda(1) = c\,\bar{e}_1, \qquad I^\lambda(e_1) = c \cdot 1.$$

In words: "the level is empty" (the ε-vacuum) is "one hole in the level" in the ε' description, and "one particle in the level" is the ε'-vacuum. Charges, counted relative to each vacuum, compare as

$$Q_{\varepsilon'}\big(I^\lambda(e_S \otimes \phi)\big) \overset{(1)}{=} Q(\phi) - (n-p) \overset{(2)}{=} \big(Q(\phi) + p\big) - n \overset{(3)}{=} Q_\varepsilon(e_S \otimes \phi) - \dim J.$$

(1) The ε'-state has the charge $Q(\phi)$ of the common factor $\phi \in \Lambda^*(H^+_{\varepsilon'} \oplus \bar{H}^-_{\varepsilon'})$ plus $n - p$ holes of charge -1.
(2) Rearrange.
(3) The ε-state has the same ϕ plus p particles.

So moving the cut up by a window containing $\dim J$ levels shifts every charge by $-\dim J$ (all the diagonals of Figure 10.51 move). This is the Fock-space version of the vacuum-charge sawtooth of Figure 10.50: the *label* "charge relative to the vacuum" depends on the cut, the physics does not.

Worked example 10.74: Two levels in the window: signs

For $n = 2$, $\lambda = c\, e_1 \wedge e_2$. The signs are $\sigma(\varnothing) = 1$ ($e_1 \wedge e_2 \wedge 1$), $\sigma(\{1\}) = -1$ ($e_2 \wedge e_1 = -e_1 \wedge e_2$), $\sigma(\{2\}) = +1$ ($e_1 \wedge e_2$), $\sigma(\{1,2\}) = 1$. Hence

$$I^\lambda(1) = c\,\bar{e}_1 \wedge \bar{e}_2, \quad I^\lambda(e_1) = -c\,\bar{e}_2, \quad I^\lambda(e_2) = c\,\bar{e}_1, \quad I^\lambda(e_1 \wedge e_2) = c.$$

As a matrix from the basis $(1, e_1, e_2, e_1 \wedge e_2)$ to $(\bar{e}_1 \wedge \bar{e}_2, \bar{e}_2, \bar{e}_1, 1)$ this is $c\ \mathrm{diag}(1, -1, 1, 1)$. Now change the orthonormal basis of J by $u \in \mathrm{U}(2)$. The wedge $e_1 \wedge e_2$ becomes $(\det u)\, e_1 \wedge e_2$. If we keep the *same* λ, then c becomes $c/\det u$ and the map I^λ is unchanged (it depends on λ, not on the basis). If instead we keep the same c, I changes by the phase $\det u$. Either way: the isomorphism is fixed by a vector of the line $\Lambda^2(J)$, and different choices differ by a non-zero complex number.

The ambiguity is a line; projectively it disappears. The last two examples show that I^λ depends on λ only through an overall factor: $I^{z\lambda} = z\, I^\lambda$ for $z \in \mathbb{C}^\times$. There is no canonical choice of λ in the line $\Lambda^n(J)$. But the induced map of *projective* spaces (rays),

$$\mathbb{P}\mathfrak{F}_\varepsilon(A) \xrightarrow{[I^\lambda]} \mathbb{P}\mathfrak{F}_{\varepsilon'}(A),$$

is the same for every λ [Nash p. 299]. Quantum states are rays anyway: $|\psi\rangle$ and $z|\psi\rangle$ are the same physical state. So we drop the label ε and write $\mathfrak{FP}(A)$ for the projective Fock space, canonically defined for every A, zero modes or not. The projective Fock spaces form a bundle over all of $\mathcal{A}$ [Nash (10.95)]:

$$\mathfrak{FP} = \{\mathfrak{FP}(A) : A \in \mathcal{A}\}, \tag{10.249}$$

glued on each overlap $U_\varepsilon \cap U_{\varepsilon'}$ by $[I^\lambda]$. Consistency on triple overlaps $\varepsilon < \varepsilon' < \varepsilon''$ holds because $J_{\varepsilon\varepsilon''} = J_{\varepsilon\varepsilon'} \oplus J_{\varepsilon'\varepsilon''}$, so $\Lambda^{\mathrm{top}} J_{\varepsilon\varepsilon''} = \Lambda^{\mathrm{top}} J_{\varepsilon\varepsilon'} \otimes \Lambda^{\mathrm{top}} J_{\varepsilon'\varepsilon''}$ and the composite of two I's is an I up to a scalar. And now gauge transformations cause no trouble: g maps each $J_{\varepsilon\varepsilon'}(A)$ to $J_{\varepsilon\varepsilon'}(g \cdot A)$ (same eigenvalues), hence commutes with the gluing up to scalars, so $\mathcal{G}$ acts smoothly on $\mathfrak{FP}$ and the quotient [Nash (10.96)]

$$\mathfrak{FP}/\mathcal{G} \longrightarrow \mathcal{A}/\mathcal{G} \tag{10.250}$$

is a perfectly good bundle of projective spaces.

Key idea: Where the anomaly now sits

The line bundles $\det J_{\varepsilon\varepsilon'} \to U_\varepsilon \cap U_{\varepsilon'}$ are the Hamiltonian determinant lines. Choosing the λ's is choosing sections of them. Over $\mathcal{A}$ (contractible) these choices can be made consistently, and $\mathfrak{F}\mathfrak{P}$ lifts to an honest vector bundle $\mathfrak{F}$ of Fock spaces. The question is whether this can be done *compatibly with* $\mathcal{G}$, i.e. whether the projective bundle $\mathfrak{F}\mathfrak{P}/\mathcal{G}$ over $\mathcal{A}/\mathcal{G}$ is the projectivisation of a vector bundle. The obstruction is a class in $H^3(\mathcal{A}/\mathcal{G};\mathbb{Z})$; for chiral fermions on S^3 it is the generator of $H^3(\mathcal{A}_3/\mathcal{G}_3;\mathbb{Z}) = \mathbb{Z}$ dual to the 3-sphere of (10.240). Seen from $\mathcal{G}$, it is a 2-cocycle.

Why projective bundles are classified by H^3. A reminder of the one-degree-up pattern (Chapter 1). Complex line bundles, i.e. U(1)-bundles, are classified by $H^2(X;\mathbb{Z})$ through c_1. The reason is that U(1) is a circle, whose classifying space $\mathbb{CP}^\infty$ has only $\pi_2 = \mathbb{Z}$. A bundle of projective Hilbert spaces has structure group $\mathrm{PU}(H) = \mathrm{U}(H)/\mathrm{U}(1)$. By Kuiper's theorem $\mathrm{U}(H)$ is contractible, so $\mathrm{PU}(H)$ has the homotopy of $B\mathrm{U}(1) = \mathbb{CP}^\infty$: only $\pi_2 = \mathbb{Z}$. Its classifying space $B\mathrm{PU}(H)$ therefore has only $\pi_3 = \mathbb{Z}$, and projective bundles over X are classified by $H^3(X;\mathbb{Z})$, the **Dixmier–Douady class**. The class vanishes exactly when the projective bundle comes from a vector bundle. For Nash's bundle, $X = \mathcal{A}/\mathcal{G}$, and the class is the anomaly.

From $H^3(\mathcal{A}/\mathcal{G})$ to $H^2(\mathcal{G})$: transgression. Nash phrases the obstruction as "a cocycle in $H^2(\mathcal{G})$ which corresponds under transgression to an element of $H^3(\mathcal{A}/\mathcal{G})$". The geometric content is the same as step 1 of (10.240). Take a 3-sphere Σ in $\mathcal{A}/\mathcal{G}$. Cut it open at one point. What remains is a 3-ball, and a 3-ball lifts to $\mathcal{A}$ (it is contractible, so the bundle $\mathcal{A} \to \mathcal{A}/\mathcal{G}$ is trivial over it). The lifted ball $B^3 \subset \mathcal{A}$ has its boundary 2-sphere inside a single gauge orbit, because the whole boundary was glued to one point of $\mathcal{A}/\mathcal{G}$. A gauge orbit is a copy of $\mathcal{G}$, so ∂B^3 is a 2-sphere in $\mathcal{G}$ (Figure 10.53). Integrating a closed 3-form Θ on $\mathcal{A}/\mathcal{G}$ over Σ and using Stokes' theorem on the ball turns the answer into an integral of a 2-form over a 2-sphere in the group:

$$\int_\Sigma \Theta \overset{(1)}{=} \int_{B^3} \pi^*\Theta \overset{(2)}{=} \int_{B^3} \mathrm{d}\beta \overset{(3)}{=} \int_{\partial B^3} \beta \overset{(4)}{=} \int_{S^2\subset\mathcal{G}} \beta\big|_{\text{orbit}}. \tag{10.251}$$

(1) $\pi : \mathcal{A} \to \mathcal{A}/\mathcal{G}$ maps B^3 onto Σ, one-to-one except on the boundary, which is collapsed to a point (measure zero).
(2) $\mathcal{A}$ is contractible, so the closed form $\pi^*\Theta$ is exact: $\pi^*\Theta = \mathrm{d}\beta$ (Poincaré lemma, Chapter T2).
(3) Stokes' theorem.
(4) ∂B^3 lies in one orbit $\cong \mathcal{G}$; restricted to it, β is closed ($\mathrm{d}\beta = \pi^*\Theta$ vanishes on vectors tangent to an orbit, since π kills them), so it defines a class in $H^2(\mathcal{G})$.

This is **transgression**: a degree-3 class downstairs becomes a degree-2 class on the fibre. In the Euclidean picture of §2 the same move turned $c_1 \in H^2(\mathcal{A}/\mathcal{G})$ into a class in $H^1(\mathcal{G})$, whose Lie-algebra version was the consistent anomaly, a 1-cocycle. Here the Lie-algebra version of the $H^2(\mathcal{G})$ class will be a 2-cocycle, the Schwinger term.

10.10.15 Symmetries that act only up to phases

The projective Fock bundle $\mathfrak{F}\mathfrak{P}/\mathcal{G}$ of (10.250) is a perfectly good bundle, and $\mathcal{G}$ acts on it. But a quantum state is a *vector*, not a ray, as soon as we want to add two states or ask whether a state is invariant. So the question Nash asks next is: when we lift the action of $\mathcal{G}$ from the rays back to the vectors, what goes wrong? The answer is a phase, and the phase is the anomaly. Before we meet it on Fock space we meet it where every step is a 2×2 matrix.

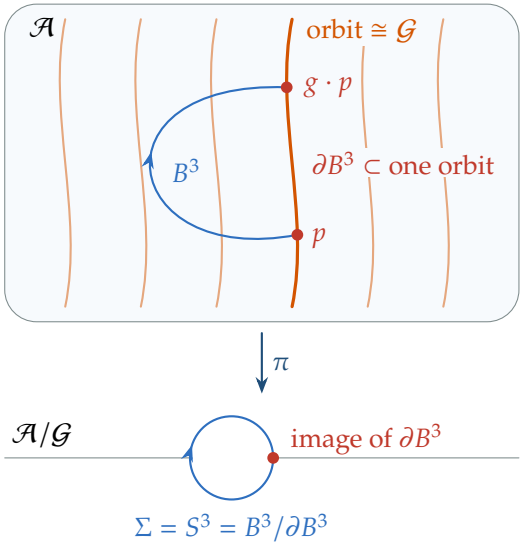


Figure 10.53: Transgression, drawn with every dimension lowered by two. The 3-sphere Σ in the orbit space (bottom) becomes a loop, and its lift, the 3-ball B^3 in the contractible space $\mathcal{A}$ (blue, top), becomes a path. The path starts at p and ends at $g \cdot p$ on the same gauge orbit (orange): its whole boundary (red; really a 2-sphere) lies in one orbit and is collapsed to one point by π. So a 3-sphere in $\mathcal{A}/\mathcal{G}$ is the same as a 2-sphere in $\mathcal{G}$, and $H^3(\mathcal{A}/\mathcal{G})$ corresponds to $H^2(\mathcal{G})$ by Stokes' theorem (10.251).

Why only rays are physical. Two unit vectors ψ and $e^{i\theta}\psi$ give the same probabilities $\left|\langle\phi|\psi\rangle\right|^2$ for every ϕ. So the physical content of a state is its ray $[\psi] = \{e^{i\theta}\psi\}$. A symmetry is a map on rays that preserves the transition probabilities. Wigner's theorem says that every such map comes from a unitary (or antiunitary) operator U on vectors, and that U is unique *up to a phase*. So if a group G acts by symmetries, each $g \in G$ gives an operator $U(g)$ that we may multiply by any phase we like. The group law $g \cdot h = gh$ then only requires that $U(g)U(h)$ and $U(gh)$ give the same map on rays:

Working definition 10.20 (Projective representation and its multiplier)

A **projective representation** of a group G on a Hilbert space is an assignment $g \mapsto U(g)$ of unitary operators such that

$$U(g)\,U(h) = \omega(g,h)\,U(gh), \qquad |\omega(g,h)| = 1. \tag{10.252}$$

The function $\omega : G \times G \to \mathrm{U}(1)$ is the **multiplier** (or **2-cocycle** [it obeys $\omega(g,h)\,\omega(gh,k) = \omega(h,k)\,\omega(g,hk)$, because $(U(g)U(h))U(k) = U(g)(U(h)U(k))$]). *How to compute it:* choose one operator $U(g)$ for every g, multiply two of them, and read off the phase by which the product differs from the chosen $U(gh)$. *How to test whether it is removable* [can be made 1 by changing each $U(g)$ by a phase, $U(g) \to \beta(g)U(g)$]: for every pair of commuting elements $gh = hg$, compute the **commutator phase** $\omega(g,h)/\omega(h,g)$. If one of them is not 1, no choice of phases turns U into an honest representation [one with $U(g)U(h) = U(gh)$ exactly].

The cocycle condition. The multiplier is not an arbitrary function. Operator products are associative, so the two ways of evaluating $U(g)U(h)U(k)$ must agree:

$$\big(U(g)U(h)\big)U(k) \overset{(1)}{=} \omega(g,h)\,U(gh)\,U(k) \overset{(2)}{=} \omega(g,h)\,\omega(gh,k)\,U(ghk),$$

$$U(g)\big(U(h)U(k)\big) \overset{(3)}{=} \omega(h,k)\,U(g)\,U(hk) \overset{(4)}{=} \omega(h,k)\,\omega(g,hk)\,U(ghk).$$

(1) (10.252) for the pair (g,h); the phase is a number and moves to the front.
(2) (10.252) for the pair (gh,k).
(3) (10.252) for the pair (h,k).
(4) (10.252) for the pair (g,hk).

Since $U(ghk)$ is invertible, the two phases are equal:

$$\omega(g,h)\,\omega(gh,k) = \omega(h,k)\,\omega(g,hk) \qquad \text{for all } g,h,k \in G. \tag{10.253}$$

This is the **2-cocycle condition**. It is the multiplicative version of the condition $\delta\omega = 0$ in group cohomology, and it plays for projective representations the role that $\mathrm{d}\omega = 0$ plays for forms.

Changing the phases: coboundaries. We were free to choose the phases. Replace $U(g)$ by $U'(g) = \beta(g)U(g)$ with any function $\beta : G \to \mathrm{U}(1)$. Then

$$U'(g)U'(h) \overset{(1)}{=} \beta(g)\beta(h)\,U(g)U(h) \overset{(2)}{=} \beta(g)\beta(h)\,\omega(g,h)\,U(gh) \overset{(3)}{=} \frac{\beta(g)\beta(h)}{\beta(gh)}\,\omega(g,h)\,U'(gh). \tag{10.254}$$

(1) Definition of U'; the numbers β move to the front.
(2) (10.252).
(3) $U(gh) = \beta(gh)^{-1}U'(gh)$.

So the new multiplier is $\omega' = \omega \cdot \beta(g)\beta(h)/\beta(gh)$. A multiplier of the form $\beta(g)\beta(h)/\beta(gh)$ is a **coboundary**: it can be removed completely by rephasing. The classes of cocycles modulo coboundaries form the group $H^2(G;\mathrm{U}(1))$, and a projective representation can be made into an honest one exactly when its class is zero. The commutator phase is a good detector because it is unchanged by rephasing: if $gh = hg$, then $\beta(g)\beta(h)/\beta(gh) = \beta(h)\beta(g)/\beta(hg)$, so the factor cancels in $\omega'(g,h)/\omega'(h,g)$.

Worked example 10.75: The Pauli matrices represent $\mathbb{Z}_2 \times \mathbb{Z}_2$ projectively

Let $G = \mathbb{Z}_2 \times \mathbb{Z}_2 = \{e,a,b,ab\}$, with $a^2 = b^2 = e$ and $ab = ba$. It is abelian, so every honest irreducible representation is one-dimensional. Choose

$$U(e) = \mathbb{1}, \quad U(a) = \sigma_x, \quad U(b) = \sigma_z, \quad U(ab) = \sigma_y,$$

with $\sigma_x = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, $\sigma_y = \left(\begin{smallmatrix} 0 & -i \\ i & 0 \end{smallmatrix}\right)$, $\sigma_z = \left(\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right)$. Multiply the matrices out:

$$U(a)U(b) = \sigma_x\sigma_z = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \overset{(1)}{=} -i\,\sigma_y = -i\,U(ab),$$

$$U(b)U(a) = \sigma_z\sigma_x = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \overset{(2)}{=} +i\,\sigma_y = +i\,U(ab).$$

(1) $-i\sigma_y = -i\left(\begin{smallmatrix} 0 & -i \\ i & 0 \end{smallmatrix}\right) = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$.
(2) $i\sigma_y = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$.

Doing the same for all pairs (using $\sigma_x\sigma_y = i\sigma_z$, $\sigma_y\sigma_z = i\sigma_x$, $\sigma_z\sigma_x = i\sigma_y$ and $\sigma_j^2 = \mathbb{1}$) gives the whole multiplier:

$\omega(g,h)$	$h=e$	a	b	ab
$g=e$	1	1	1	1
a	1	1	$-i$	i
b	1	i	1	$-i$
ab	1	$-i$	i	1

For instance $\omega(a,ab)$: $U(a)U(ab)=\sigma_x\sigma_y=i\sigma_z=iU(b)$ and $a\cdot ab=b$, so $\omega(a,ab)=i$. The commutator phase of the commuting pair (a,b) is

$$\frac{\omega(a,b)}{\omega(b,a)}=\frac{-i}{i}=-1\neq 1.$$

So the class is non-trivial: σ_x and σ_z anticommute, and no choice of phases can make them commute. The "honest" group that does act is generated by σ_x,σ_z and the phases: it is a non-abelian extension of the abelian $\mathbb{Z}_2\times\mathbb{Z}_2$ (the Pauli group).

Worked example 10.76: No invariant vector: a toy Gauss law

Suppose we demand a state invariant under the whole group, even allowing each element to act with its own phase: $\sigma_x\psi=\lambda\psi$ and $\sigma_z\psi=\mu\psi$. Then

$$\lambda\mu\,\psi\overset{(1)}{=}\sigma_x\sigma_z\psi\overset{(2)}{=}-\sigma_z\sigma_x\psi\overset{(3)}{=}-\mu\lambda\,\psi,$$

(1) Apply σ_z, then σ_x, using the two eigenvalue equations.
(2) $\sigma_x\sigma_z=-i\sigma_y=-\sigma_z\sigma_x$ from Example 10.75.
(3) Apply σ_x, then σ_z.

so $2\lambda\mu\psi=0$, and since $|\lambda|=|\mu|=1$, $\psi=0$. A non-trivial multiplier forbids invariant states. This two-line argument is, in miniature, why Gauss's law fails in an anomalous gauge theory (Section 10.10.20).

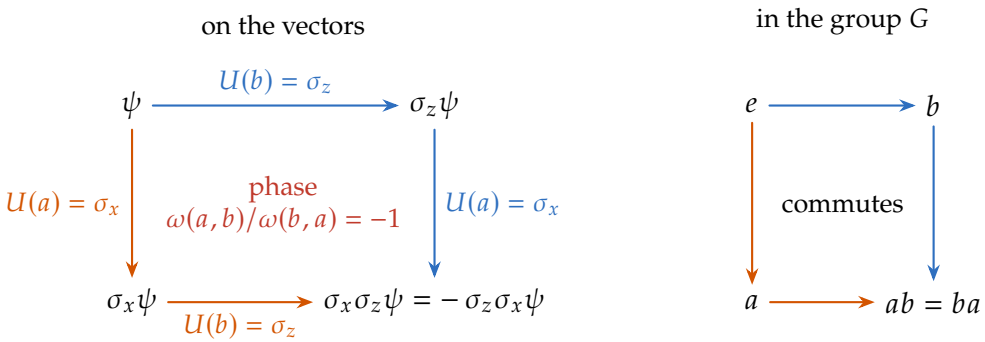


Figure 10.54: A projective representation, drawn as a square. In the group (right) going "b then a" or "a then b" lands on the same element. On the vectors (left) the two routes land on vectors that differ by the commutator phase, here -1. The phase does not depend on how the individual operators were phased, so it is an invariant of the projective representation; it is the simplest picture of a 2-cocycle.

Worked example 10.77: Spin $\frac{1}{2}$ is a projective representation of SO(3)

The rotation by angle φ about the unit vector $\vec{n}$ acts on a spin-$\frac{1}{2}$ state by $U(R_{\vec{n}}(\varphi)) = \exp(-\frac{i}{2}\varphi\,\vec{n}\cdot\vec{\sigma})$. Take $\vec{n} = \hat{z}$, so the operator is $\mathrm{diag}(e^{-i\varphi/2}, e^{i\varphi/2})$. The rotation by π composed with itself is the rotation by 2π, which is the identity of SO(3). But

$$U(R_z(\pi))\,U(R_z(\pi)) \overset{(1)}{=} \mathrm{diag}(e^{-i\pi/2}, e^{i\pi/2})^2 = \mathrm{diag}(e^{-i\pi}, e^{i\pi}) \overset{(2)}{=} -\mathbb{1} \overset{(3)}{=} -U(R_z(2\pi) = e),$$

(1) The formula for U at $\varphi = \pi$.
(2) $e^{\pm i\pi} = -1$.
(3) We chose $U(e) = \mathbb{1}$.

so $\omega(R_z(\pi), R_z(\pi)) = -1$. This multiplier takes only the values ± 1 and cannot be removed by any continuous choice of phases: its class is the generator of $H^2(\mathrm{SO}(3); \mathrm{U}(1)) = \mathbb{Z}_2$, which is the statement that the double cover $\mathrm{SU}(2) \to \mathrm{SO}(3)$ is non-trivial (Chapter T4). The honest group acting on spinors is the extension SU(2).

The extension that acts honestly. Every projective representation is an honest representation of a bigger group. On the set $\widehat{G} = \mathrm{U}(1) \times G$ define the product

$$(\lambda, g)\cdot(\mu, h) = \big(\lambda\mu\,\omega(g,h),\ gh\big). \tag{10.255}$$

It is associative exactly because of the cocycle condition:

$$\begin{aligned}\big((\lambda,g)(\mu,h)\big)(\nu,k) &\overset{(1)}{=} \big(\lambda\mu\nu\,\omega(g,h)\,\omega(gh,k),\ ghk\big)\\ &\overset{(2)}{=} \big(\lambda\mu\nu\,\omega(h,k)\,\omega(g,hk),\ ghk\big) \overset{(3)}{=} (\lambda,g)\big((\mu,h)(\nu,k)\big).\end{aligned}$$

(1) Apply (10.255) twice, left pair first.
(2) The cocycle condition (10.253).
(3) Apply (10.255) twice, right pair first.

And $(\lambda, g) \mapsto \lambda U(g)$ is an honest representation, because $\lambda U(g)\,\mu U(h) = \lambda\mu\,\omega(g,h)U(gh)$, which is the operator assigned to the product (10.255). The subgroup $\{(\lambda, e)\} \cong \mathrm{U}(1)$ commutes with everything (it is **central**), and dividing by it gives back G: $1 \to \mathrm{U}(1) \to \widehat{G} \to G \to 1$. For Example 10.77 the relevant part of $\widehat{G}$ is SU(2); for Example 10.75 it is the Pauli group.

10.10.16 Infinitesimal version: Lie-algebra cocycles

For a Lie group we can phase each one-parameter subgroup so that $U(e^{tX}) = e^{t\rho(X)}$ with $\rho(X)$ anti-Hermitian (a generator). What does the multiplier do to the commutation relations of the generators? Take X, Y in the Lie algebra and small s, t. In the group, $e^{sX}e^{tY}e^{-sX}e^{-tY} = e^{st[X,Y]+\cdots}$, where the dots are of third order. On the vectors the same product equals a phase times $U(e^{st[X,Y]})$. Write that phase as $1 + st\,c(X,Y) + \ldots$ with $c(X,Y) \in i\mathbb{R}$ and expand both sides to order st:

$$\begin{aligned}e^{s\rho(X)}e^{t\rho(Y)}e^{-s\rho(X)}e^{-t\rho(Y)} &\overset{(1)}{=} \mathbb{1} + st\,[\rho(X),\rho(Y)] + \ldots,\\ \big(1 + st\,c(X,Y)\big)\big(\mathbb{1} + st\,\rho([X,Y])\big) &\overset{(2)}{=} \mathbb{1} + st\,\big(\rho([X,Y]) + c(X,Y)\big) + \ldots\end{aligned}$$

(1) Expand each exponential to first order; the terms of order s^2 and t^2 cancel in pairs, and the surviving cross terms are $st(\rho(X)\rho(Y) - \rho(Y)\rho(X))$.

(2) Multiply out and keep order st.

Comparing,

$$[\rho(X),\rho(Y)] = \rho([X,Y]) + c(X,Y)\,\mathbb{1}. \tag{10.256}$$

The number $c(X,Y)$ is a **central term**. It is antisymmetric (swap X and Y in (10.256)) and bilinear. The Jacobi identity for the commutators of operators forces a cocycle condition:

$$0 \overset{(1)}{=} \sum_{\text{cyc}} \big[[\rho(X),\rho(Y)],\rho(Z)\big] \overset{(2)}{=} \sum_{\text{cyc}} \big[\rho([X,Y]) + c(X,Y),\ \rho(Z)\big]$$
$$\overset{(3)}{=} \sum_{\text{cyc}} \Big(\rho\big([[X,Y],Z]\big) + c([X,Y],Z)\Big) \overset{(4)}{=} \sum_{\text{cyc}} c([X,Y],Z),$$

(1) Jacobi identity for operators; "cyc" sums over the three cyclic permutations of (X,Y,Z).
(2) (10.256).
(3) The number $c(X,Y)$ commutes with $\rho(Z)$; then (10.256) again.
(4) Jacobi identity in the Lie algebra, $\sum_{\text{cyc}}[[X,Y],Z] = 0$, and linearity of ρ.

So $c([X,Y],Z) + c([Y,Z],X) + c([Z,X],Y) = 0$: a **Lie-algebra 2-cocycle**. Redefining the generators by constants, $\rho'(X) = \rho(X) + \alpha(X)$ with α linear, changes $c'(X,Y) = c(X,Y) - \alpha([X,Y])$: the commutator of the ρ' is that of the ρ, while $\rho([X,Y]) = \rho'([X,Y]) - \alpha([X,Y])$. So c is trivial (a coboundary) when it has the form $\alpha([X,Y])$.

Worked example 10.78: The Heisenberg group: canonical commutators are a cocycle

Let $G = \mathbb{R}^2$, the abelian group of phase-space translations (a,b), acting on wave functions $\psi(x)$ on the line by $W(a,b) = \exp\big(i(b\hat{x} - a\hat{p})\big)$ (units $\hbar = 1$, $[\hat{x},\hat{p}] = i$). Separately, $W(a,0) = e^{-ia\hat{p}}$ translates, $(W(a,0)\psi)(x) = \psi(x-a)$, and $W(0,b) = e^{ib\hat{x}}$ boosts the momentum. We need one standard identity.

Recall (the Baker–Campbell–Hausdorff formula when the commutator is central). If $[P,Q]$ commutes with P and Q, then $e^P e^Q = e^{P+Q+\frac{1}{2}[P,Q]}$. Proof: $F(t) = e^{tP}e^{tQ}$ obeys $F' = (P + e^{tP}Qe^{-tP})F = (P + Q + t[P,Q])F$, because the series $e^{tP}Qe^{-tP} = Q + t[P,Q] + \frac{t^2}{2}[P,[P,Q]] + \dots$ stops after two terms. Since $P+Q$ commutes with $[P,Q]$, the solution with $F(0) = \mathbb{1}$ is $F(t) = e^{t(P+Q)+\frac{t^2}{2}[P,Q]}$; differentiate to check. Set $t = 1$.

Now take $P = i(b\hat{x} - a\hat{p})$, $Q = i(b'\hat{x} - a'\hat{p})$:

$$[P,Q] \overset{(1)}{=} -\big[b\hat{x} - a\hat{p},\ b'\hat{x} - a'\hat{p}\big] \overset{(2)}{=} -\big(-ba'[\hat{x},\hat{p}] - ab'[\hat{p},\hat{x}]\big) \overset{(3)}{=} -i\,(ab' - ba'),$$
$$W(a,b)\,W(a',b') \overset{(4)}{=} e^{-\frac{i}{2}(ab'-ba')}\,W(a+a',\,b+b').$$

(1) $i \cdot i = -1$.
(2) Bilinearity; $[\hat{x},\hat{x}] = [\hat{p},\hat{p}] = 0$.
(3) $[\hat{x},\hat{p}] = i$ and $[\hat{p},\hat{x}] = -i$, so the bracket is $-ba'i + ab'i$.
(4) The BCH formula above; $P + Q = i((b+b')\hat{x} - (a+a')\hat{p})$.

So $\omega\big((a,b),(a',b')\big) = e^{-\frac{i}{2}(ab'-ba')}$. The group $\mathbb{R}^2$ is abelian, but the commutator phase of a translation and a boost is

$$\frac{\omega((a,0),(0,b))}{\omega((0,b),(a,0))} = \frac{e^{-\frac{i}{2}ab}}{e^{+\frac{i}{2}ab}} = e^{-iab} \neq 1.$$

Check it directly on a wave function: $W(a,0)W(0,b)\psi(x) = e^{ib(x-a)}\psi(x-a)$ while $W(0,b)W(a,0)\psi(x) = e^{ibx}\psi(x-a)$; the ratio is e^{-iab}. At the Lie algebra level the generators are $\rho(e_1) = -i\hat{p}$ and $\rho(e_2) = i\hat{x}$, and $[\rho(e_1),\rho(e_2)] = [\hat{p},\hat{x}] = -i$, while $[e_1,e_2] = 0$.

So (10.256) holds with $c(e_1, e_2) = -i$. Because the Lie algebra is abelian, $\alpha([e_1, e_2]) = 0$ for every α, and the central term cannot be removed. The canonical commutation relation *is* a non-trivial Lie-algebra 2-cocycle; the honest group is the Heisenberg group. The toy Gauss law fails here too: $\hat{x}\psi = 0$ and $\hat{p}\psi = 0$ would give $i\psi = [\hat{x}, \hat{p}]\psi = 0$.

Dictionary: Group cocycles and Lie-algebra cocycles

group level	**Lie-algebra level**
multiplier $\omega(g, h) \in \mathrm{U}(1)$	central term $c(X, Y) \in i\mathbb{R}$
$\omega(g, h)\omega(gh, k) = \omega(h, k)\omega(g, hk)$	$\sum_{\mathrm{cyc}} c([X, Y], Z) = 0$
rephasing $\beta(g)$	shift $\alpha(X)$
coboundary $\beta(g)\beta(h)/\beta(gh)$	coboundary $-\alpha([X, Y])$
central extension $\widehat{G}$	extended algebra $\mathfrak{g} \oplus i\mathbb{R}$
commutator phase of commuting g, h	$c(X, Y)$ for commuting X, Y

10.10.17 Gauge transformations on the Fock bundle: Nash's (10.97)

Now the real thing. Over each connection A on space we have a Fock space $\mathfrak{F}(A)$, defined up to the choice of a line (Section 10.10.14), and the projective spaces $\mathfrak{FP}(A)$ fit into a bundle on which $\mathcal{G}$ acts. A gauge transformation g maps the Dirac Hamiltonian at A to the Dirac Hamiltonian at $g \cdot A$ (they are unitarily equivalent by (10.236) and its non-abelian analogue), so it maps rays in $\mathfrak{FP}(A)$ to rays in $\mathfrak{FP}(g \cdot A)$. Lift this to vectors: for every g and every A choose a unitary

$$U_A(g) : \mathfrak{F}(A) \longrightarrow \mathfrak{F}(g \cdot A)$$

covering the action on rays. As for Wigner's theorem, each $U_A(g)$ is fixed up to a phase, but now the phase may depend on A.

The multiplier is a function of A. Compose two lifts. $U_A(h)$ goes from $\mathfrak{F}(A)$ to $\mathfrak{F}(h \cdot A)$, and $U_{h\cdot A}(g)$ continues to $\mathfrak{F}(gh \cdot A)$. The composite covers the same map of rays as $U_A(gh)$, so

$$U_{h\cdot A}(g)\, U_A(h) = \omega(g, h; A)\, U_A(gh), \qquad \omega(g, h; \cdot) \in \mathrm{Map}(\mathcal{A}, \mathrm{U}(1)). \tag{10.257}$$

This is Nash's (10.97), $g \cdot (h \cdot f_A) = \lambda(A)(gh) \cdot f_A$, with $\lambda(A) = \omega(\cdot, \cdot; A)$. The new feature compared with (10.252) is that the phase is a *function on the space of connections*.

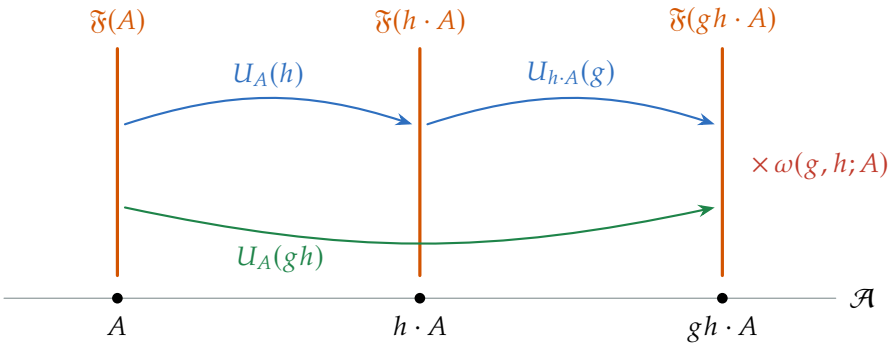


Figure 10.55: Nash's (10.97). The lifted gauge transformations carry the Fock space over A to the Fock spaces over the gauge-equivalent connections (orange fibres). Going in two steps (blue) or in one step (green) lands on the same ray; the vectors differ by a phase $\omega(g, h; A)$, which is a function of the starting point A.

The twisted cocycle condition. Associativity again, but now we must keep track of where each operator acts. Evaluate $U_{hk\cdot A}(g)\,U_{k\cdot A}(h)\,U_A(k)$, a map from $\mathfrak{F}(A)$ to $\mathfrak{F}(ghk\cdot A)$, in the two possible ways:

$$\big(U_{hk\cdot A}(g)\,U_{k\cdot A}(h)\big)U_A(k) \overset{(1)}{=} \omega(g,h;k\cdot A)\,U_{k\cdot A}(gh)\,U_A(k) \overset{(2)}{=} \omega(g,h;k\cdot A)\,\omega(gh,k;A)\,U_A(ghk),$$
$$U_{hk\cdot A}(g)\big(U_{k\cdot A}(h)\,U_A(k)\big) \overset{(3)}{=} \omega(h,k;A)\,U_{hk\cdot A}(g)\,U_A(hk) \overset{(4)}{=} \omega(h,k;A)\,\omega(g,hk;A)\,U_A(ghk).$$

(1) (10.257) with base point $k\cdot A$ (the pair g,h acts on the fibre over $k\cdot A$).
(2) (10.257) for the pair (gh,k) at base point A.
(3) (10.257) for the pair (h,k) at base point A.
(4) (10.257) for the pair (g,hk) at base point A.

Hence

$$\omega(g,h;k\cdot A)\,\omega(gh,k;A) = \omega(h,k;A)\,\omega(g,hk;A). \tag{10.258}$$

If ω does not depend on A this is (10.253). In general the first factor is evaluated at a moved point: $\mathcal{G}$ acts on the functions $\mathrm{Map}(\mathcal{A},\mathrm{U}(1))$ by moving their argument, and (10.258) is the cocycle condition of group cohomology *with coefficients in the $\mathcal{G}$-module* $\mathrm{Map}(\mathcal{A},\mathrm{U}(1))$. Rephasing $U'_A(g)=\beta(g;A)U_A(g)$ gives, exactly as in (10.254) but keeping track of base points,

$$U'_{h\cdot A}(g)\,U'_A(h) \overset{(1)}{=} \beta(g;h\cdot A)\,\beta(h;A)\,U_{h\cdot A}(g)\,U_A(h) \overset{(2)}{=} \frac{\beta(g;h\cdot A)\,\beta(h;A)}{\beta(gh;A)}\,\omega(g,h;A)\,U'_A(gh),$$

(1) Definition of U'; the lift of g acts at the point $h\cdot A$.
(2) (10.257), then $U_A(gh)=\beta(gh;A)^{-1}U'_A(gh)$.

so the removable multipliers are $\beta(g;h\cdot A)\beta(h;A)/\beta(gh;A)$ with β an arbitrary function of g and A. This is a much larger freedom than before, which is why a multiplier that looks non-trivial can sometimes be removed by an A-dependent rephasing. We will see an instance in Example 10.83.

The extension $\widehat{\mathcal{G}}$. The operators that act on sections Ψ of the Fock bundle ($\Psi(A)\in\mathfrak{F}(A)$) are the lifts $(\widehat{U}(g)\Psi)(g\cdot A)=U_A(g)\Psi(A)$ and the multiplications $(M_f\Psi)(A)=f(A)\Psi(A)$ by functions $f\in\mathrm{Map}(\mathcal{A},\mathrm{U}(1))$. A one-line computation shows that they do not commute:

$$(\widehat{U}(g)M_f\Psi)(g\cdot A) \overset{(1)}{=} U_A(g)\,f(A)\Psi(A) \overset{(2)}{=} f(A)\,(\widehat{U}(g)\Psi)(g\cdot A) \overset{(3)}{=} (M_{f\circ g^{-1}}\widehat{U}(g)\Psi)(g\cdot A),$$

(1) Definitions of $\widehat{U}(g)$ and M_f.
(2) $f(A)$ is a number; definition of $\widehat{U}(g)$.
(3) Write $f(A)=(f\circ g^{-1})(g\cdot A)$, where $(f\circ g^{-1})(B)=f(g^{-1}\cdot B)$.

so $\widehat{U}(g)M_f = M_{f\circ g^{-1}}\widehat{U}(g)$. The group generated by the $\widehat{U}(g)$ and the M_f is the extension

$$1\longrightarrow \mathrm{Map}(\mathcal{A},\mathrm{U}(1))\longrightarrow \widehat{\mathcal{G}}\longrightarrow \mathcal{G}\longrightarrow 1,$$

and the subgroup $\mathrm{Map}(\mathcal{A},\mathrm{U}(1))$ is *not* central: $\widehat{U}(g)$ moves the argument of f. This is the non-central extension of the gauge group met in Chapter 6, and it is what Nash means by "an extension of the group $\mathcal{G}$ by the group of projective multipliers $\lambda(A)$, which clearly belong to $\mathrm{Map}(\mathcal{A},S^1)$". When all the multipliers can be chosen constant, the extension is central and we are back in the situation of (10.255).

The infinitesimal version: Gauss generators and their algebra. Differentiate along one-parameter subgroups $g = e^{tX}$, $X \in \mathcal{LG}$ (a Lie-algebra-valued function on space). The lift becomes a first-order operator on sections,

$$\widehat{G}(X) = L_X + \widehat{\rho}_A(X), \tag{10.259}$$

where L_X differentiates a section along the infinitesimal gauge direction and $\widehat{\rho}_A(X)$ is the fermionic generator acting inside each fibre. We use the convention that e^{tX} moves A by

$$\delta_X A = \mathrm{d}X + [A, X] = D_A X, \tag{10.260}$$

the infinitesimal form of $A \mapsto g^{-1}Ag + g^{-1}\mathrm{d}g$ with $g = e^{tX}$ (expand to first order: $g^{-1}\mathrm{d}g \approx t\,\mathrm{d}X$ and $g^{-1}Ag \approx (\mathbb{1} - tX)A(\mathbb{1} + tX) \approx A + t[A, X]$). This is the right action $A \mapsto A^g$; in the left-action notation $g \cdot A$ of (10.257) it is the action of g^{-1}. We use it because with this sign the L_X are an honest representation of the Lie algebra, as we now check. On a function f of A, $(L_X f)(A) = \frac{\mathrm{d}}{\mathrm{d}t} f(A + t\,\delta_X A)|_{t=0}$.

The L_X represent the Lie algebra. We check $[L_X, L_Y] = L_{[X,Y]}$ on a function f. Write $\mathrm{D}f_A(\cdot)$ for the derivative of f at A. Then $L_Y f(A) = \mathrm{D}f_A(\delta_Y A)$, and $\delta_Y A = \mathrm{d}Y + [A, Y]$ itself depends on A, with derivative $[\delta_X A, Y]$ in the direction $\delta_X A$. So

$$\begin{aligned}
[L_X, L_Y]f &\overset{(1)}{=} \mathrm{D}f_A\big([\delta_X A, Y] - [\delta_Y A, X]\big) \\
&\overset{(2)}{=} \mathrm{D}f_A\big([\mathrm{d}X, Y] + [[A, X], Y] - [\mathrm{d}Y, X] - [[A, Y], X]\big) \\
&\overset{(3)}{=} \mathrm{D}f_A\big(\mathrm{d}[X, Y] + [A, [X, Y]]\big) \overset{(4)}{=} L_{[X,Y]}f.
\end{aligned}$$

(1) Product rule: $L_X L_Y f$ is the second derivative $\mathrm{D}^2 f_A(\delta_X A, \delta_Y A)$, which is symmetric and cancels in the commutator, plus $\mathrm{D}f_A$ of the derivative of $\delta_Y A$ along $\delta_X A$.
(2) (10.260) for $\delta_X A$ and $\delta_Y A$.
(3) $\mathrm{d}[X, Y] = [\mathrm{d}X, Y] + [X, \mathrm{d}Y] = [\mathrm{d}X, Y] - [\mathrm{d}Y, X]$; and $[[A, X], Y] - [[A, Y], X] = [[A, X], Y] + [[Y, A], X] = -[[X, Y], A] = [A, [X, Y]]$ by the Jacobi identity.
(4) (10.260) for $[X, Y]$.

Commutators with multiplication operators. For a function f of A (acting by multiplication) and any section Ψ, the product rule $L_Z(f\Psi) = (L_Z f)\Psi + f L_Z \Psi$ gives

$$[f, \widehat{G}(Z)] = -(L_Z f), \tag{10.261}$$

since $\widehat{\rho}_A(Z)$ acts fibrewise and commutes with multiplication by a number.

Working definition 10.21 (Schwinger term)

If the lifts [the operators $\widehat{U}(g)$ carrying the Fock space at A to the one at $g \cdot A$] are projective with an A-dependent multiplier [$\widehat{U}(g)\widehat{U}(h)$ differs from $\widehat{U}(gh)$ by a phase that depends on A], the generators [the infinitesimal lifts $\widehat{G}(X) = L_X + \widehat{\rho}_A(X)$, for $g = e^{tX}$] obey

$$[\widehat{G}(X), \widehat{G}(Y)] = \widehat{G}([X, Y]) + c(A; X, Y), \tag{10.262}$$

where $c(A; X, Y)$ is a function on $\mathcal{A}$ (acting by multiplication), bilinear and antisymmetric in X, Y. It is the **Schwinger term**. *How to compute it:* calculate the commutator of the regularised charge densities [the fermion bilinears with their infinite Dirac-sea value subtracted, i.e. normal ordered] of the fermions in the background A, and subtract $\widehat{G}([X, Y])$. *How to test it:* it is removable [can be made zero by shifting $\widehat{G}(X) \to \widehat{G}(X) + \alpha(X)$] iff $c = -(\delta\alpha)$ for some function $\alpha(A; X)$ linear in X, in the sense of (10.264) below.

The Jacobi identity now gives a cocycle condition that involves L_X:

$$0 \overset{(1)}{=} \sum_{\text{cyc}} \big[[\widehat{G}(X),\widehat{G}(Y)],\widehat{G}(Z)\big] \overset{(2)}{=} \sum_{\text{cyc}} \Big(\big[\widehat{G}([X,Y]),\widehat{G}(Z)\big] + \big[c(X,Y),\widehat{G}(Z)\big]\Big)$$
$$\overset{(3)}{=} \sum_{\text{cyc}} \Big(\widehat{G}\big([[X,Y],Z]\big) + c([X,Y],Z) - L_Z c(X,Y)\Big) \overset{(4)}{=} \sum_{\text{cyc}} \Big(c([X,Y],Z) - L_Z c(X,Y)\Big).$$

(1) Jacobi identity for operators.
(2) (10.262) for the inner commutator.
(3) (10.262) for the first term; (10.261) with $f = c(X,Y)$ for the second.
(4) Jacobi identity in $\mathcal{LG}$ and linearity of $\widehat{G}$.

Writing out the cyclic sum and using $c(Z,X) = -c(X,Z)$,

$$(\delta c)(X,Y,Z) \coloneqq L_X c(Y,Z) - L_Y c(X,Z) + L_Z c(X,Y) - c([X,Y],Z) + c([X,Z],Y) - c([Y,Z],X) = 0. \tag{10.263}$$

This is the Lie-algebra version of (10.258): the Chevalley–Eilenberg coboundary with coefficients in functions on $\mathcal{A}$. The L terms are new compared with Section 10.10.16; they are there because $\mathcal{G}$ acts on the coefficients.

Coboundaries. Shift the generators by functions, $\widehat{G}'(X) = \widehat{G}(X) + \alpha(X)$ with $\alpha(X) = \alpha(A;X)$ linear in X:

$$\begin{aligned}[\widehat{G}'(X),\widehat{G}'(Y)] &\overset{(1)}{=} [\widehat{G}(X),\widehat{G}(Y)] + [\widehat{G}(X),\alpha(Y)] + [\alpha(X),\widehat{G}(Y)] \\ &\overset{(2)}{=} \widehat{G}([X,Y]) + c(X,Y) + L_X\alpha(Y) - L_Y\alpha(X) \\ &\overset{(3)}{=} \widehat{G}'([X,Y]) + \underbrace{c(X,Y) + L_X\alpha(Y) - L_Y\alpha(X) - \alpha([X,Y])}_{c'(X,Y)}.\end{aligned} \tag{10.264}$$

(1) Bilinearity; $[\alpha(X),\alpha(Y)] = 0$ since both are multiplication operators.
(2) (10.262); (10.261) in the form $[\widehat{G}(X), f] = L_X f$.
(3) $\widehat{G}([X,Y]) = \widehat{G}'([X,Y]) - \alpha([X,Y])$.

So c can be removed iff $c = -(\delta\alpha)$ with $(\delta\alpha)(X,Y) = L_X\alpha(Y) - L_Y\alpha(X) - \alpha([X,Y])$. The anomaly is the class of c in this cohomology. Nash's obstruction, the class in $H^2(\mathcal{G})$ which transgresses to $H^3(\mathcal{A}/\mathcal{G};\mathbb{Z})$ ((10.251)), is its global form.

10.10.18 Schwinger terms from the anticommutation relations

Where does a Schwinger term come from in practice? From second quantisation of one-particle operators relative to a Dirac sea. The calculation is short and completely explicit, and it shows in one line why the answer is zero in finite dimensions "up to a coboundary", and non-zero in infinite dimensions.

Second quantisation of a one-particle operator. Let H have orthonormal basis e_i and let $a_i, a_i^\dagger$ obey the canonical anticommutation relations (CAR) $\{a_i, a_j^\dagger\} = \delta_{ij}$, $\{a_i, a_j\} = \{a_i^\dagger, a_j^\dagger\} = 0$. For a one-particle operator X with matrix $X_{ij} = \langle e_i | X e_j\rangle$ put

$$\widehat{X} = \sum_{i,j} X_{ij}\, a_i^\dagger a_j. \tag{10.265}$$

(For $X = \mathbb{1}$ this is the number operator; for $X = f(x)$ a multiplication operator it is the smeared density $\int f\,\psi^\dagger\psi$.) First the basic commutator:

$$\begin{aligned}[a_i^\dagger a_j,\, a_k^\dagger a_l] &\overset{(1)}{=} a_i^\dagger(\delta_{jk} - a_k^\dagger a_j)a_l - a_k^\dagger(\delta_{li} - a_i^\dagger a_l)a_j \\ &\overset{(2)}{=} \delta_{jk}\, a_i^\dagger a_l - \delta_{il}\, a_k^\dagger a_j - a_i^\dagger a_k^\dagger a_j a_l + a_k^\dagger a_i^\dagger a_l a_j \\ &\overset{(3)}{=} \delta_{jk}\, a_i^\dagger a_l - \delta_{il}\, a_k^\dagger a_j. \end{aligned} \tag{10.266}$$

(1) Write the commutator as $a_i^\dagger a_j a_k^\dagger a_l - a_k^\dagger a_l a_i^\dagger a_j$ and use the CAR $a_j a_k^\dagger = \delta_{jk} - a_k^\dagger a_j$ and $a_l a_i^\dagger = \delta_{li} - a_i^\dagger a_l$.
(2) Multiply out.
(3) $a_k^\dagger a_i^\dagger = -a_i^\dagger a_k^\dagger$ and $a_l a_j = -a_j a_l$, so the last term equals $+a_i^\dagger a_k^\dagger a_j a_l$ and cancels the one before it.

Multiplying by $X_{ij}Y_{kl}$ and summing:

$$[\widehat{X}, \widehat{Y}] \overset{(1)}{=} \sum_{i,l}\Big(\sum_j X_{ij}Y_{jl}\Big)a_i^\dagger a_l - \sum_{k,j}\Big(\sum_i Y_{ki}X_{ij}\Big)a_k^\dagger a_j \overset{(2)}{=} \widehat{[X,Y]}.$$

(1) (10.266); the δ's set $k = j$ in the first term and $l = i$ in the second.
(2) The brackets are $(XY)_{il}$ and $(YX)_{kj}$; rename the indices and use (10.265).

So second quantisation is an honest Lie-algebra homomorphism: no central term. In finite dimensions nothing more needs to be said.

Normal ordering relative to a Dirac sea. Now let $H = H^+ \oplus H^-$ be polarised by a one-particle Hamiltonian, and let the e_i be eigenvectors, with occupation numbers $n_i = 1$ for $e_i \in H^-$ (filled) and $n_i = 0$ for $e_i \in H^+$ (empty) in the Dirac sea $|\Omega\rangle$. Then $\langle\Omega|a_i^\dagger a_j|\Omega\rangle = \delta_{ij}n_i$. If H is infinite-dimensional, $\widehat{X}$ does not make sense on the sea: for $X = \mathbb{1}$, $\widehat{X}|\Omega\rangle = (\text{number of filled levels})|\Omega\rangle = \infty$. The cure is to subtract the vacuum value, i.e. to **normal order**:

$$:a_i^\dagger a_j: = a_i^\dagger a_j - \delta_{ij}n_i, \qquad :\widehat{X}: = \sum_{i,j} X_{ij}\,:a_i^\dagger a_j:. \tag{10.267}$$

In finite dimensions this is $\widehat{X} - \mathrm{tr}(P_- X)$, with P_- the projector onto H^-. Substitute $a^\dagger a = :a^\dagger a: + \delta n$ into (10.266):

$$\begin{aligned}[:a_i^\dagger a_j:, :a_k^\dagger a_l:] &\overset{(1)}{=} \delta_{jk}(:a_i^\dagger a_l: + \delta_{il}n_i) - \delta_{il}(:a_k^\dagger a_j: + \delta_{kj}n_k) \\ &\overset{(2)}{=} \delta_{jk}\,:a_i^\dagger a_l: - \delta_{il}\,:a_k^\dagger a_j: + \delta_{jk}\delta_{il}\,(n_i - n_j). \end{aligned} \tag{10.268}$$

(1) The numbers $\delta_{ij}n_i$ commute with everything, so the commutator of normal-ordered bilinears equals that of the bare ones, (10.266); then re-express the result in normal-ordered form.
(2) In the last term $k = j$ is enforced by δ_{kj}, so $n_k = n_j$.

The last term is new: a number. Multiplying by $X_{ij}Y_{kl}$ and summing, the operator parts give $:\widehat{[X,Y]}:$ as before, and

$$[:\widehat{X}:, :\widehat{Y}:] = :\widehat{[X,Y]}: + c(X,Y), \qquad c(X,Y) = \sum_{i,j} X_{ij}Y_{ji}\,(n_i - n_j). \tag{10.269}$$

Only pairs with one level filled and one empty contribute. With the block notation $X_{+-} = P_+ X P_-$ (from H^- to H^+) and $X_{-+} = P_- X P_+$,

$$c(X,Y) \overset{(1)}{=} \sum_{i\in -,\, j\in +} X_{ij}Y_{ji} - \sum_{i\in +,\, j\in -} X_{ij}Y_{ji} \overset{(2)}{=} \operatorname{tr}\big(X_{-+}Y_{+-}\big) - \operatorname{tr}\big(X_{+-}Y_{-+}\big) \overset{(3)}{=} \tfrac{1}{4}\operatorname{tr}\big(\epsilon\,[\epsilon,X]\,[\epsilon,Y]\big). \tag{10.270}$$

(1) $n_i - n_j = +1$ if i is filled and j empty, -1 in the reverse case, 0 otherwise.
(2) Matrix multiplication of the blocks.
(3) With $\epsilon = P_+ - P_-$ (the sign of the Hamiltonian), in block form $[\epsilon, X] = \left(\begin{smallmatrix} 0 & 2X_{+-} \\ -2X_{-+} & 0 \end{smallmatrix}\right)$, so $\epsilon[\epsilon,X][\epsilon,Y] = \left(\begin{smallmatrix} -4X_{+-}Y_{-+} & 0 \\ 0 & 4X_{-+}Y_{+-} \end{smallmatrix}\right)$; take $\frac{1}{4}$ of the trace.

Key idea: Why the Schwinger term is zero in finite dimensions and not in infinite ones

In finite dimensions, $c(X,Y) = \operatorname{tr}(P_-[X,Y])$ (the $--$ block of $[X,Y]$ contributes $\operatorname{tr}(X_{--}Y_{--} - Y_{--}X_{--}) = 0$). That is the coboundary $-\alpha([X,Y])$ of $\alpha(X) = -\operatorname{tr}(P_-X)$, and indeed $:\widehat{X}: = \widehat{X} - \operatorname{tr}(P_-X)$ is just a constant shift of an honest homomorphism. In infinite dimensions $\operatorname{tr}(P_-X)$ is infinite, so this α does not exist, but (10.270) is still finite whenever the off-diagonal blocks X_{+-}, X_{-+} are Hilbert–Schmidt (the product of two Hilbert–Schmidt operators is trace class). The Schwinger term is the finite remainder of a difference of two infinite traces. This form of the cocycle is due to Lundberg.

Worked example 10.79: The U(1) current algebra on a circle

Take one right-moving Weyl fermion on the circle: $H = L^2(S^1)$, $h = -i\,\mathrm{d}/\mathrm{d}\theta$, eigenvectors $e_k = e^{ik\theta}/\sqrt{2\pi}$ with eigenvalue $k \in \mathbb{Z}$. Fill $k \le -1$ and leave $k \ge 0$ empty (the zero mode is empty, as in Section 10.10.10): $n_k = 1$ for $k \le -1$. The density operators are the second-quantised multiplication operators. Multiplication by $e^{im\theta}$ has matrix

$$\big(e^{im\theta}\big)_{jk} = \langle e_j | e^{im\theta} e_k \rangle = \delta_{j,k+m},$$

since it moves e_k to e_{k+m}. Define the Fourier modes of the charge density by $\rho_m = :\widehat{e^{-im\theta}}: = \sum_k :a^\dagger_{k-m}a_k:$. Multiplication operators commute, $[e^{-im\theta}, e^{-in\theta}] = 0$, so the naive answer is $[\rho_m,\rho_n] = 0$. The correct answer is (10.269):

$$[\rho_m,\rho_n] \overset{(1)}{=} \sum_{i,j} \delta_{i,j-m}\,\delta_{j,i-n}\,(n_i - n_j) \overset{(2)}{=} \delta_{m+n,0} \sum_j \big(n_{j-m} - n_j\big) \overset{(3)}{=} m\,\delta_{m+n,0}.$$

(1) (10.269) with $X_{ij} = \delta_{i,j-m}$ and $Y_{ji} = \delta_{j,i-n}$; the operator part vanishes because $[X,Y] = 0$.
(2) The first δ sets $i = j - m$; then the second demands $j = j - m - n$.
(3) Count, for $m > 0$: $n_{j-m} - n_j$ is $+1$ exactly when $j - m \le -1$ and $j \ge 0$, i.e. for $j = 0, 1, \ldots, m-1$: m values, and 0 otherwise. For $m < 0$ the same count gives $-|m| = m$. For $m = 0$ it is 0.

So $[\rho_m,\rho_{-m}] = m$: for $m > 0$, $\rho_m/\sqrt{m}$ and $\rho_{-m}/\sqrt{m}$ obey the commutation relation of an annihilation and a creation operator. Indeed $\rho_m|\Omega\rangle = 0$ for $m > 0$ (the term $a^\dagger_{k-m}a_k$ needs k filled and $k-m$ empty, impossible), and $\|\rho_{-m}|\Omega\rangle\|^2 = \langle\Omega|\rho_m\rho_{-m}|\Omega\rangle = \langle\Omega|[\rho_m,\rho_{-m}]|\Omega\rangle = m$: the state $\rho_{-m}|\Omega\rangle$ is a superposition of the m particle–hole pairs that straddle the Fermi level (Figure 10.56). This is the starting point of bosonisation.

Position space. For real functions $f = \sum_m f_m e^{im\theta}$ and g put $\rho(f) = :\widehat{f}: = \sum_m f_m \rho_{-m}$. Then

$$[\rho(f), \rho(g)] \overset{(1)}{=} \sum_{m,n} f_m g_n \, [\rho_{-m}, \rho_{-n}] \overset{(2)}{=} -\sum_m m \, f_m g_{-m} \overset{(3)}{=} \frac{1}{2\pi i} \int_0^{2\pi} f \, g' \, \mathrm{d}\theta,$$

(1) Bilinearity.

(2) The result above with $(m, n) \to (-m, -n)$: $[\rho_{-m}, \rho_{-n}] = -m \, \delta_{m+n,0}$.

(3) $\int_0^{2\pi} f g' \mathrm{d}\theta = \sum_{m,n} f_m g_n (in) \, 2\pi \delta_{m+n,0} = -2\pi i \sum_m m f_m g_{-m}$.

which is the smeared form of $[\rho(\theta), \rho(\theta')] = \frac{1}{2\pi i} \, \delta'(\theta - \theta')$ (derivative with respect to θ; check by inserting into $\iint f(\theta) g(\theta') [\rho(\theta), \rho(\theta')]$ and integrating by parts). This is the classic Schwinger term of a chiral fermion in $1 + 1$ dimensions.

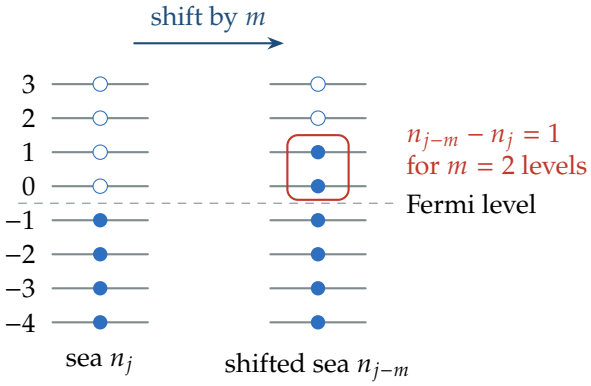


Figure 10.56: Where the m in $[\rho_m, \rho_{-m}] = m$ comes from. Left: the Dirac sea, levels $j \le -1$ filled. Right: the same occupation pattern shifted up by $m = 2$. Each of the two infinite sums $\sum_j n_j$ and $\sum_j n_{j-m}$ diverges, but their difference is finite: exactly the m levels just above the Fermi level (red box). The "trace of a commutator of two multiplication operators" is $\infty - \infty = m$.

Watch out: Chiral versus vector Schwinger terms

A left mover, $h_L = +i \, \mathrm{d}/\mathrm{d}\theta$, has energy $-k$ on e_k, so its sea fills $k \ge 1$. The same count as in step (3) gives $\sum_j (n_{j-m} - n_j) = -m$: the left-moving densities obey $[\rho^L_m, \rho^L_n] = -m \, \delta_{m+n,0}$. For a Dirac fermion the vector density $\rho = \rho^R + \rho^L$ therefore has *no* Schwinger term, while the axial density $\rho_5 = \rho^R - \rho^L$ has $[\rho_5(f), \rho(g)] = \frac{2}{2\pi i} \int f g' \, \mathrm{d}\theta$. The central term is a property of *chiral* charges. In a vector-like gauge theory the gauge generators are built from ρ, and the gauge algebra closes; in a chiral theory they are built from ρ^R alone, and it does not. The same bookkeeping produced $\Delta Q = 0$, $\Delta Q_5 = 2k$ in Example 10.70.

Worked example 10.80: Non-abelian currents in $1 + 1$: the Kac–Moody algebra

Give the right mover N colour components, $H = L^2(S^1) \otimes \mathbb{C}^N$, and let X, Y be $\mathfrak{u}(N)$-valued functions on the circle, $X = \sum_m X_m e^{im\theta}$ with X_m an $N \times N$ matrix. The sea is the same in each colour, so (10.269) gives the U(1) answer with a colour trace. For $X = X_m e^{im\theta}$ and $Y = Y_n e^{in\theta}$ the count of the previous example gives $c(X, Y) = -m \, \delta_{m+n,0} \, \mathrm{tr}(X_m Y_n)$, and summing over modes exactly as in the position-space step,

$$c(X, Y) = -\sum_m m \, \mathrm{tr}(X_m Y_{-m}) = \frac{1}{2\pi i} \int_0^{2\pi} \mathrm{tr}\big(X \, \partial_\theta Y\big) \mathrm{d}\theta. \tag{10.271}$$

This does not depend on any background connection: it is a genuine central term, and (10.269) with it is the level-one Kac–Moody (affine) algebra. At the group level it is the central extension of the loop group $\mathrm{Map}(S^1, \mathrm{U}(N))$ constructed in Chapter 6. The Lie-algebra cocycle condition of Section 10.10.16 can be checked by hand: $\oint \mathrm{tr}([X,Y]\,Z')$ summed cyclically vanishes, because $\mathrm{tr}([X,Y]Z') + \mathrm{tr}([Y,Z]X') + \mathrm{tr}([Z,X]Y') = \partial_\theta \,\mathrm{tr}(XYZ - YXZ)$ (expand both sides with the product rule and cyclicity of the trace), and a derivative integrates to zero on the circle.

When does normal ordering work? The Hilbert–Schmidt test. The state

$$:\widehat{X}:|\Omega\rangle = \sum_{i\in+,\, j\in-} X_{ij}\, a_i^\dagger a_j |\Omega\rangle$$

is a superposition of particle–hole pairs. (On the sea, $a_i^\dagger a_j|\Omega\rangle = 0$ if j is empty, since a_j finds nothing to remove, and also if $i \neq j$ is filled, since $a_i^\dagger$ is blocked by the Pauli principle; for $i = j$ filled it gives $|\Omega\rangle$, which the subtraction $\delta_{ij} n_i$ cancels. Only i empty and j filled survive.) Its squared norm is

$$\left\| :\widehat{X}:|\Omega\rangle \right\|^2 = \sum_{i\in+,\, j\in-} \left|X_{ij}\right|^2 = \|X_{+-}\|_{\mathrm{HS}}^2 \,. \tag{10.272}$$

(Distinct pairs (i,j) give orthogonal states of norm one.) So $:\widehat{X}:$ makes sense on the Fock space built on this sea exactly when X_{+-} is **Hilbert–Schmidt**, i.e. when the sum of the squares of its matrix elements converges. The same is true at the group level (the **Shale–Stinespring theorem**): a unitary U on H is implemented by a unitary on Fock space iff its off-diagonal blocks are Hilbert–Schmidt, and then the implementer is unique up to a phase. The implementable unitaries form the **restricted unitary group** $\mathrm{U}_{\mathrm{res}}(H)$, and the phases of the implementers give a *central* extension of it, whose Lie-algebra cocycle is (10.270).

Worked example 10.81: One space dimension passes, three fail

$d = 1$. Take $h = -i\,\mathrm{d}/\mathrm{d}\theta$ and X multiplication by a smooth $f = \sum_m f_m e^{im\theta}$. Then $X_{ij} = f_{i-j}$, and

$$\|X_{+-}\|_{\mathrm{HS}}^2 \overset{(1)}{=} \sum_{i\geq 0}\sum_{j\leq -1} \left|f_{i-j}\right|^2 \overset{(2)}{=} \sum_{m\geq 1} m\ |f_m|^2 < \infty.$$

(1) (10.272) with $H^+ = \{i \geq 0\}$ and $H^- = \{j \leq -1\}$.

(2) Group the terms by $m = i - j \geq 1$; the number of pairs with a given m is m (namely $j = -1, \ldots, -m$). Smooth f has rapidly decreasing f_m.

So every smooth gauge transformation on the circle is implementable, the extension is central, and the Schwinger term is the constant (10.271).

$d = 3$. Take a Weyl fermion on the 3-torus, $h = \vec{\sigma}\cdot\vec{p}$ with $\vec{p} \in \mathbb{Z}^3$. For each $\vec{p} \neq 0$ there is one positive-energy spinor $u_+(\vec{p})$ (spin along $\hat{\vec{p}}$) and one negative-energy spinor $u_-(\vec{p})$ (spin against $\hat{\vec{p}}$). Multiplication by $e^{i\vec{m}\cdot\vec{x}}$ maps momentum $\vec{p}$ to $\vec{p} + \vec{m}$; its $+-$ block has matrix elements $\langle u_+(\vec{p}+\vec{m})|u_-(\vec{p})\rangle$. We need the overlap of two spin states. Let $|\vec{n},+\rangle$ be the spin-$\frac{1}{2}$ state pointing along the unit vector $\vec{n}$ (the state $|\psi_N(\vec{n})\rangle$ of the Berry-phase example, Example 10.58). Its projector is $|\vec{n},+\rangle\langle\vec{n},+| = \frac{1}{2}(\mathbb{1} + \vec{n}\cdot\vec{\sigma})$: both sides are Hermitian with trace 1, and $\frac{1}{2}(\mathbb{1} + \vec{n}\cdot\vec{\sigma})$ has eigenvalue 1 on $|\vec{n},+\rangle$ (because $(\vec{n}\cdot\vec{\sigma})|\vec{n},+\rangle = |\vec{n},+\rangle$) and 0 on the orthogonal state. Then

$$\left|\langle\vec{n}_1,+|\vec{n}_2,+\rangle\right|^2 \overset{(1)}{=} \mathrm{tr}\big(|\vec{n}_1,+\rangle\langle\vec{n}_1,+|\,|\vec{n}_2,+\rangle\langle\vec{n}_2,+|\big)$$

$$\overset{(2)}{=} \tfrac{1}{4}\operatorname{tr}\big((\mathbb{1}+\vec{n}_1\cdot\vec{\sigma})(\mathbb{1}+\vec{n}_2\cdot\vec{\sigma})\big) \overset{(3)}{=} \tfrac{1}{2}\big(1+\vec{n}_1\cdot\vec{n}_2\big).$$

(1) $\operatorname{tr}(|a\rangle\langle a|b\rangle\langle b|) = \langle a|b\rangle\langle b|a\rangle = |\langle a|b\rangle|^2$.
(2) The projector formula.
(3) $\operatorname{tr}\mathbb{1} = 2$, $\operatorname{tr}\sigma_i = 0$, $\operatorname{tr}(\sigma_i\sigma_j) = 2\delta_{ij}$.

The negative-energy state at $\vec{p}$ points against $\hat{\vec{p}}$: $u_-(\vec{p}) = |-\hat{\vec{p}},+\rangle$. So

$$\|X_{+-}\|_{\mathrm{HS}}^2 \overset{(1)}{=} \sum_{\vec{p}} \tfrac{1}{2}\big(1-\hat{\vec{p}}\cdot\widehat{\vec{p}+\vec{m}}\big) \overset{(2)}{=} \sum_{\vec{p}}\Big(\frac{|\vec{m}_\perp|^2}{4|\vec{p}|^2} + O(|\vec{p}|^{-3})\Big) \overset{(3)}{=} \infty.$$

(1) The overlap formula with $\vec{n}_1 = \widehat{\vec{p}+\vec{m}}$ and $\vec{n}_2 = -\hat{\vec{p}}$.
(2) For large $|\vec{p}|$ the angle ϑ between $\vec{p}$ and $\vec{p}+\vec{m}$ is $|\vec{m}_\perp|/|\vec{p}|$ to leading order, where $\vec{m}_\perp$ is the part of $\vec{m}$ perpendicular to $\vec{p}$; and $1-\cos\vartheta \approx \vartheta^2/2$.
(3) The number of lattice points with $|\vec{p}| \approx P$ grows like $4\pi P^2$, so the sum behaves like $\int^\Lambda 4\pi P^2\,\mathrm{d}P/P^2 \propto \Lambda$: linearly divergent.

In one dimension only finitely many levels near the Fermi point are disturbed by f (Figure 10.56). In three dimensions a gauge transformation tilts the spin of *every* level by a small angle, and the small tilts add up to an infinite number of pairs. The sum $\sum|\vec{p}|^{-2}$ diverges while $\sum|\vec{p}|^{-4}$ converges, so X_{+-} is in the Schatten class $\mathcal{L}_4$ but not in $\mathcal{L}_2$.

Key idea: The dimension of space decides the type of extension

In $1+1$ dimensions the gauge group on space lies inside $\mathrm{U}_{\mathrm{res}}$ of a fixed polarisation: the Fock space can be built once and for all, the extension is *central*, and the Schwinger term is the constant (10.271). In $3+1$ dimensions it does not: there is no single Fock space on which all gauge transformations act. One must use a separate Fock space $\mathfrak{F}(A)$ for each background connection, built on the polarisation of $\not{D}_A$ itself (Section 10.10.12). Then the phases of the implementers depend on A, the extension is by $\mathrm{Map}(\mathcal{A},\mathrm{U}(1))$ and is non-central, as in (10.257), and the Schwinger term is a non-constant function on $\mathcal{A}$. This is precisely the situation Nash describes, and the next subsection computes that function.

10.10.19 The Faddeev–Mickelsson cocycle

We now want the function $c(A;X,Y)$ for chiral fermions on S^3. A direct Hamiltonian computation is possible (it is Mickelsson's: renormalise the charge densities separately on each $\mathfrak{F}(A)$ and compute their commutators), but it is long. Faddeev found the answer faster, by noticing that the Schwinger term is the next rung of the same *descent ladder* that produced the Euclidean anomaly in Chapter 10. We follow that route and do every step.

Ghosts: a bookkeeping device for cochains. A cochain of degree g on the gauge Lie algebra is a function $c(A;X_1,\dots,X_g)$, multilinear and antisymmetric in the X's. A compact way to handle antisymmetric multilinear functions is to replace the X's by a single anticommuting variable, the **ghost** $v = v^a T_a$, whose components $v^a(x)$ are Grassmann numbers. A ghost-number-g expression, a polynomial of degree g in v and its derivatives, determines a g-cochain: to evaluate it on $(X_1,\dots,X_g)$ replace each product $v^{a_1}\cdots v^{a_g}$ of ghost components by the antisymmetrised product $\sum_\pi \operatorname{sgn}(\pi) X^{a_1}_{\pi(1)}\cdots X^{a_g}_{\pi(g)}$. For $g=2$: $v^a v^b \mapsto X^a Y^b - Y^a X^b$. For example (up to an

overall sign that depends on how ghosts are ordered against $\mathrm{d}x^\mu$),

$$\operatorname{tr}(A\,\mathrm{d}v\,\mathrm{d}v) \;\longmapsto\; \operatorname{tr}\big(A\,(\mathrm{d}X\,\mathrm{d}Y - \mathrm{d}Y\,\mathrm{d}X)\big), \tag{10.273}$$

where $\mathrm{d}X\,\mathrm{d}Y$ means the wedge product of the matrix-valued 1-forms $\mathrm{d}X$ and $\mathrm{d}Y$ (the order of the matrices matters). Under this dictionary the BRST operator s becomes (up to sign) the coboundary δ of (10.263), and $s^2 = 0$ becomes $\delta^2 = 0$.

Conventions. All objects are graded by (form degree + ghost number) mod 2: A and v are odd, $\mathrm{d}A$ and $\mathrm{d}v$ are even. The operators d and s are odd derivations, $\mathrm{d}(xy) = (\mathrm{d}x)y + (-1)^{|x|}x\,\mathrm{d}y$ and the same for s, with $\mathrm{d}^2 = s^2 = 0$ and $s\mathrm{d} + \mathrm{d}s = 0$. The trace is graded-cyclic: $\operatorname{tr}(xy) = (-1)^{|x||y|}\operatorname{tr}(yx)$. The action of s is fixed by the **Russian formula** $(\mathrm{d}+s)(A+v) + (A+v)^2 = F$; sorting it by (form, ghost) degree gives $\mathrm{d}A + A^2 = F$ at $(2,0)$, and

$$\begin{aligned} (1,1):&\quad sA + \mathrm{d}v + Av + vA = 0 \implies sA = -\mathrm{d}v - Av - vA,\\ (0,2):&\quad sv + v^2 = 0 \implies sv = -v^2. \end{aligned} \tag{10.274}$$

(With v replaced by tX, $sA = -D_A v$ is minus the infinitesimal gauge transformation (10.260); the sign is the standard one for BRST.) From these, using $s\mathrm{d} = -\mathrm{d}s$:

$$\begin{aligned} s(\mathrm{d}A) &\overset{(1)}{=} -\mathrm{d}(sA) \overset{(2)}{=} \mathrm{d}(\mathrm{d}v + Av + vA) \overset{(3)}{=} \mathrm{d}A\,v - A\,\mathrm{d}v + \mathrm{d}v\,A - v\,\mathrm{d}A,\\ s(\mathrm{d}v) &\overset{(4)}{=} -\mathrm{d}(sv) \overset{(5)}{=} \mathrm{d}(v^2) \overset{(6)}{=} \mathrm{d}v\,v - v\,\mathrm{d}v. \end{aligned}$$

(1) $s\mathrm{d} = -\mathrm{d}s$.
(2) (10.274).
(3) $\mathrm{d}^2 v = 0$; $\mathrm{d}(Av) = \mathrm{d}A\,v - A\,\mathrm{d}v$ and $\mathrm{d}(vA) = \mathrm{d}v\,A - v\,\mathrm{d}A$ (graded Leibniz, A and v odd).
(4) $s\mathrm{d} = -\mathrm{d}s$.
(5) (10.274).
(6) Graded Leibniz with v odd.

(Check: $s^2 v = -s(v^2) = -(sv)v + v(sv) = v^3 - v^3 = 0$.)

The rung we need. In Chapter 10 (§2) the consistent non-abelian anomaly in four dimensions came out as a multiple of

$$\omega_4^1 = \operatorname{tr}\big(v\,\mathrm{d}\Phi\big), \qquad \Phi = A\,\mathrm{d}A + \tfrac{1}{2}A^3, \tag{10.275}$$

with normalisation $\frac{1}{24\pi^2}$ (up to convention-dependent factors of i), obtained by descent from $\operatorname{tr} F^3$ in six dimensions. The descent equations continue one more step: there is a 3-form ω_3^2 of ghost number 2 with

$$s\,\omega_4^1 + \mathrm{d}\,\omega_3^2 = 0. \tag{10.276}$$

We find it. Three short computations.

Step 1: move the derivative off the ghost.

$$\mathrm{d}\,\operatorname{tr}(v\Phi) \overset{(1)}{=} \operatorname{tr}(\mathrm{d}v\,\Phi) - \operatorname{tr}(v\,\mathrm{d}\Phi) \quad\Longrightarrow\quad \omega_4^1 \overset{(2)}{=} \operatorname{tr}(\mathrm{d}v\,\Phi) - \mathrm{d}\,\operatorname{tr}(v\Phi).$$

(1) Graded Leibniz; v is odd, and d commutes with the trace.
(2) Rearrange; definition (10.275).

Step 2: the BRST variation of Φ. Using (10.274) and the formula for $s(\mathrm{d}A)$,

$$\begin{aligned} s(A\,\mathrm{d}A) &\overset{(1)}{=} (sA)\,\mathrm{d}A - A\,s(\mathrm{d}A) \\ &\overset{(2)}{=} \big(-\mathrm{d}v - Av - vA\big)\mathrm{d}A - A\big(\mathrm{d}A\,v - A\,\mathrm{d}v + \mathrm{d}v\,A - v\,\mathrm{d}A\big) \\ &\overset{(3)}{=} -\mathrm{d}v\,\mathrm{d}A - vA\,\mathrm{d}A - A\,\mathrm{d}A\,v + A^2\mathrm{d}v - A\,\mathrm{d}v\,A, \\ s(A^3) &\overset{(4)}{=} (sA)A^2 - A(sA)A + A^2(sA) \\ &\overset{(5)}{=} -\mathrm{d}v\,A^2 - AvA^2 - vA^3 + A\,\mathrm{d}v\,A + A^2vA + AvA^2 - A^2\mathrm{d}v - A^3v - A^2vA \\ &\overset{(6)}{=} -\mathrm{d}v\,A^2 - vA^3 + A\,\mathrm{d}v\,A - A^2\mathrm{d}v - A^3v. \end{aligned}$$

(1) Graded Leibniz; A is odd.
(2) Insert sA and $s(\mathrm{d}A)$.
(3) Multiply out; the terms $-Av\,\mathrm{d}A$ and $+Av\,\mathrm{d}A$ cancel.
(4) Graded Leibniz twice; each A passed by s gives a sign.
(5) Insert sA in each slot and multiply out.
(6) $-AvA^2$ cancels $+AvA^2$ and $+A^2vA$ cancels $-A^2vA$.

Adding $s(A\,\mathrm{d}A) + \frac{1}{2}s(A^3)$:

$$s\Phi = -\mathrm{d}v\,\mathrm{d}A - vA\,\mathrm{d}A - A\,\mathrm{d}A\,v + \tfrac{1}{2}A^2\mathrm{d}v - \tfrac{1}{2}A\,\mathrm{d}v\,A - \tfrac{1}{2}\mathrm{d}v\,A^2 - \tfrac{1}{2}vA^3 - \tfrac{1}{2}A^3v. \tag{10.277}$$

Step 3: vary $\operatorname{tr}(\mathrm{d}v\,\Phi)$. Since $\mathrm{d}v$ is even, $s\,\operatorname{tr}(\mathrm{d}v\,\Phi) = \operatorname{tr}\big(s(\mathrm{d}v)\Phi\big) + \operatorname{tr}\big(\mathrm{d}v\,s\Phi\big)$. The two pieces are

$$\begin{aligned} \operatorname{tr}\big(s(\mathrm{d}v)\Phi\big) &= \operatorname{tr}(\mathrm{d}v\,vA\,\mathrm{d}A) - \operatorname{tr}(v\,\mathrm{d}v\,A\,\mathrm{d}A) + \tfrac{1}{2}\operatorname{tr}(\mathrm{d}v\,vA^3) - \tfrac{1}{2}\operatorname{tr}(v\,\mathrm{d}v\,A^3), \\ \operatorname{tr}\big(\mathrm{d}v\,s\Phi\big) &= -\operatorname{tr}(\mathrm{d}v\,\mathrm{d}v\,\mathrm{d}A) - \operatorname{tr}(\mathrm{d}v\,vA\,\mathrm{d}A) - \operatorname{tr}(\mathrm{d}v\,A\,\mathrm{d}A\,v) + \tfrac{1}{2}\operatorname{tr}(\mathrm{d}v\,A^2\mathrm{d}v) \\ &\quad - \tfrac{1}{2}\operatorname{tr}(\mathrm{d}v\,A\,\mathrm{d}v\,A) - \tfrac{1}{2}\operatorname{tr}(\mathrm{d}v\,\mathrm{d}v\,A^2) - \tfrac{1}{2}\operatorname{tr}(\mathrm{d}v\,vA^3) - \tfrac{1}{2}\operatorname{tr}(\mathrm{d}v\,A^3v). \end{aligned}$$

Now add them and simplify term by term:

$$\begin{aligned} s\,\operatorname{tr}(\mathrm{d}v\,\Phi) &\overset{(1)}{=} -\operatorname{tr}(v\,\mathrm{d}v\,A\,\mathrm{d}A) - \tfrac{1}{2}\operatorname{tr}(v\,\mathrm{d}v\,A^3) - \operatorname{tr}(\mathrm{d}v\,\mathrm{d}v\,\mathrm{d}A) - \operatorname{tr}(\mathrm{d}v\,A\,\mathrm{d}A\,v) \\ &\qquad + \tfrac{1}{2}\operatorname{tr}(\mathrm{d}v\,A^2\mathrm{d}v) - \tfrac{1}{2}\operatorname{tr}(\mathrm{d}v\,A\,\mathrm{d}v\,A) - \tfrac{1}{2}\operatorname{tr}(\mathrm{d}v\,\mathrm{d}v\,A^2) - \tfrac{1}{2}\operatorname{tr}(\mathrm{d}v\,A^3v) \\ &\overset{(2)}{=} -\operatorname{tr}(\mathrm{d}v\,\mathrm{d}v\,\mathrm{d}A) + \tfrac{1}{2}\operatorname{tr}(\mathrm{d}v\,A^2\mathrm{d}v) - \tfrac{1}{2}\operatorname{tr}(\mathrm{d}v\,\mathrm{d}v\,A^2) - \tfrac{1}{2}\operatorname{tr}(\mathrm{d}v\,A\,\mathrm{d}v\,A) \\ &\overset{(3)}{=} -\operatorname{tr}(\mathrm{d}v\,\mathrm{d}v\,\mathrm{d}A) \overset{(4)}{=} -\operatorname{tr}(\mathrm{d}A\,\mathrm{d}v\,\mathrm{d}v). \end{aligned}$$

(1) The terms $\pm\operatorname{tr}(\mathrm{d}v\,vA\,\mathrm{d}A)$ cancel, and so do $\pm\frac{1}{2}\operatorname{tr}(\mathrm{d}v\,vA^3)$.
(2) Graded cyclicity: moving the odd v from the end to the front past the odd block $\mathrm{d}v\,A\,\mathrm{d}A$ (parity $0+1+0$) costs a sign, so $\operatorname{tr}(\mathrm{d}v\,A\,\mathrm{d}A\,v) = -\operatorname{tr}(v\,\mathrm{d}v\,A\,\mathrm{d}A)$, and the pair cancels; in the same way $\operatorname{tr}(\mathrm{d}v\,A^3v) = -\operatorname{tr}(v\,\mathrm{d}v\,A^3)$ cancels its partner.
(3) $\operatorname{tr}(\mathrm{d}v\,A^2\mathrm{d}v) = \operatorname{tr}(\mathrm{d}v\,\mathrm{d}v\,A^2)$ (move the even $\mathrm{d}v$ from the end to the front), so those two cancel. And $\operatorname{tr}(\mathrm{d}v\,A\,\mathrm{d}v\,A) = 0$: moving the last (odd) A to the front past the odd block $\mathrm{d}v\,A\,\mathrm{d}v$ gives $-\operatorname{tr}(A\,\mathrm{d}v\,A\,\mathrm{d}v)$, and moving the last (even) $\mathrm{d}v$ of that to the front gives back $-\operatorname{tr}(\mathrm{d}v\,A\,\mathrm{d}v\,A)$; a quantity equal to minus itself is zero.
(4) Move the even $\mathrm{d}A$ to the front.

And finally $\mathrm{d}\,\operatorname{tr}(A\,\mathrm{d}v\,\mathrm{d}v) = \operatorname{tr}(\mathrm{d}A\,\mathrm{d}v\,\mathrm{d}v) - \operatorname{tr}\big(A\,\mathrm{d}(\mathrm{d}v\,\mathrm{d}v)\big) = \operatorname{tr}(\mathrm{d}A\,\mathrm{d}v\,\mathrm{d}v)$, since $\mathrm{d}^2v = 0$. Putting

the steps together:

$$\begin{aligned} s\,\omega_4^1 &\overset{(1)}{=} s\ \mathrm{tr}(\mathrm{d}v\,\Phi) - s\,\mathrm{d}\ \mathrm{tr}(v\Phi) \overset{(2)}{=} -\mathrm{tr}(\mathrm{d}A\,\mathrm{d}v\,\mathrm{d}v) + \mathrm{d}\,s\ \mathrm{tr}(v\Phi) \\ &\overset{(3)}{=} -\mathrm{d}\Big(\mathrm{tr}(A\,\mathrm{d}v\,\mathrm{d}v) - s\ \mathrm{tr}(v\Phi)\Big). \end{aligned} \tag{10.278}$$

(1) Step 1.
(2) Step 3, and $s\mathrm{d} = -\mathrm{d}s$.
(3) $\mathrm{tr}(\mathrm{d}A\,\mathrm{d}v\,\mathrm{d}v) = \mathrm{d}\ \mathrm{tr}(A\,\mathrm{d}v\,\mathrm{d}v)$.

So (10.276) holds with $\omega_3^2 = \mathrm{tr}(A\,\mathrm{d}v\,\mathrm{d}v) - s\ \mathrm{tr}(v\Phi)$. The second term is s-exact: in cochain language it is the coboundary of the local 1-cochain $\alpha(X) \propto \int \mathrm{tr}(X\Phi)$, and by (10.264) it can be dropped. (We checked (10.278) independently by implementing the graded algebra on a computer.)

Why the next rung is the Schwinger term. Integrate (10.276) over a four-dimensional slab $M = [t_0, t_1] \times S^3$ of spacetime:

$$s\int_M \omega_4^1 \;=\; -\int_M \mathrm{d}\,\omega_3^2 \;=\; -\int_{\{t_1\}\times S^3} \omega_3^2 + \int_{\{t_0\}\times S^3} \omega_3^2.$$

On a closed four-manifold the right side is zero: that is the Wess–Zumino consistency condition, $\delta a = 0$ for the anomaly $a = \int \omega_4^1$ (Chapter 10). On a slab it fails by terms that live on the initial and final *spatial* slices, and depend only on the fields there. In the Hamiltonian language the slices carry the states, and the failure of the gauge transformations acting on them to compose properly is exactly what the Schwinger term measures. So the Schwinger term on S^3 is the integral of ω_3^2 over space, translated into a 2-cochain by (10.273). (Matching the overall constant requires the full Hamiltonian computation; we quote it.)

Key idea: The Faddeev–Mickelsson cocycle

For chiral fermions on S^3 coupled to a connection A, the Schwinger term of the Gauss generators is

$$c(A; X, Y) = N\int_{S^3} \mathrm{tr}\big(A\,(\mathrm{d}X\,\mathrm{d}Y - \mathrm{d}Y\,\mathrm{d}X)\big), \qquad N = \frac{1}{24\pi^2}\ \text{(up to a factor of } i \text{ and a sign)}, \tag{10.279}$$

the Faddeev–Mickelsson cocycle. It is linear in A, so it is a *non-constant* function on $\mathcal{A}$: the extension it defines is by $\mathrm{Map}(\mathcal{A}, \mathrm{U}(1))$, not by $\mathrm{U}(1)$, as Nash says.

In components. Write $\mathrm{d}X = \partial_j X\,\mathrm{d}x^j$. Then

$$\begin{aligned} \mathrm{d}X\,\mathrm{d}Y - \mathrm{d}Y\,\mathrm{d}X &\overset{(1)}{=} \big(\partial_j X\,\partial_k Y - \partial_j Y\,\partial_k X\big)\mathrm{d}x^j \wedge \mathrm{d}x^k \overset{(2)}{=} \big(\partial_j X\,\partial_k Y + \partial_k Y\,\partial_j X\big)\mathrm{d}x^j \wedge \mathrm{d}x^k, \\ c(A; X, Y) &\overset{(3)}{=} N\int_{S^3} \epsilon^{ijk}\ \mathrm{tr}\big(A_i\,\{\partial_j X, \partial_k Y\}\big)\,\mathrm{d}^3x. \end{aligned}$$

(1) Definition of the wedge product of matrix-valued 1-forms: the matrices keep their order, the $\mathrm{d}x$'s are wedged.
(2) In the second term swap the dummy indices $j \leftrightarrow k$ and use $\mathrm{d}x^k \wedge \mathrm{d}x^j = -\mathrm{d}x^j \wedge \mathrm{d}x^k$.
(3) $A = A_i\,\mathrm{d}x^i$ and $\mathrm{d}x^i \wedge \mathrm{d}x^j \wedge \mathrm{d}x^k = \epsilon^{ijk}\mathrm{d}^3x$.

So only the *anticommutator* of the Lie-algebra matrices enters. This is the key to which groups are anomalous.

Worked example 10.82: SU(2) has no Faddeev–Mickelsson term; SU(3) has one

Expand in a basis of anti-Hermitian generators T_a: $\operatorname{tr}(A_i\{\partial_j X, \partial_k Y\}) = A_i^c\,\partial_j X^a\,\partial_k Y^b\,\operatorname{tr}\big(T_c\{T_a, T_b\}\big)$.

- SU(2), $T_a = -\frac{i}{2}\sigma_a$: $\{T_a, T_b\} = -\frac{1}{4}\{\sigma_a, \sigma_b\} = -\frac{1}{2}\delta_{ab}\mathbb{1}$, using $\{\sigma_a, \sigma_b\} = 2\delta_{ab}\mathbb{1}$. Then $\operatorname{tr}(T_c\{T_a, T_b\}) = -\frac{1}{2}\delta_{ab}\operatorname{tr} T_c = 0$. So $c \equiv 0$. This matches $\pi_5(\mathrm{SU}(2)) = \mathbb{Z}_2$: there is no integer class, so no perturbative anomaly, only Witten's global one (Section 10.9).
- SU(3), $T_a = -\frac{i}{2}\lambda_a$ (Gell-Mann matrices, with $\{\lambda_a, \lambda_b\} = \frac{4}{3}\delta_{ab}\mathbb{1} + 2d_{abe}\lambda_e$ and $\operatorname{tr}(\lambda_a\lambda_b) = 2\delta_{ab}$):

$$\operatorname{tr}\big(T_c\{T_a, T_b\}\big) \overset{(1)}{=} \Big(-\tfrac{i}{2}\Big)^3 \operatorname{tr}\big(\lambda_c\{\lambda_a, \lambda_b\}\big) \overset{(2)}{=} \tfrac{i}{8}\big(\tfrac{4}{3}\delta_{ab}\operatorname{tr}\lambda_c + 2d_{abe}\operatorname{tr}(\lambda_c\lambda_e)\big) \overset{(3)}{=} \tfrac{i}{2}\,d_{abc}.$$

  (1) Definition of T_a.
  (2) $(-\frac{i}{2})^3 = \frac{i}{8}$; the anticommutator formula.
  (3) $\operatorname{tr}\lambda_c = 0$ and $\operatorname{tr}(\lambda_c\lambda_e) = 2\delta_{ce}$.

  So $c(A; X, Y) = \frac{iN}{2}\int \epsilon^{ijk} d_{abc}\,A_i^c\,\partial_j X^a\,\partial_k Y^b\,\mathrm{d}^3x$, which is not zero (take $d_{888} = -1/\sqrt{3}$).

The totally symmetric symbol $d_{abc} = \frac{1}{4}\operatorname{tr}(\lambda_c\{\lambda_a, \lambda_b\})$ is the same one that controls the Euclidean anomaly in Chapter 10. With several species of fermions in representations R, the cocycles add, and the total is proportional to $\sum_R \pm \operatorname{tr}_R\big(T_c\{T_a, T_b\}\big)$, the sign depending on chirality. It vanishes exactly when the Euclidean anomaly cancels: the Hamiltonian and Lagrangian cancellation conditions are the same, as they must be.

Checking the cocycle identity by hand. Nash's statement that the multipliers define an extension requires c to obey (10.263). For (10.279) this can be verified directly, and the verification shows what each term does. Put $B(\beta; Y, Z) = \operatorname{tr}\big(\beta\,(\mathrm{d}Y\,\mathrm{d}Z - \mathrm{d}Z\,\mathrm{d}Y)\big)$ for a matrix 1-form β, so $c(A; Y, Z) = N\int B(A; Y, Z)$. Since c is linear in A and $\delta_X A = \mathrm{d}X + [A, X]$, $L_X c(Y, Z) = N\int B(\mathrm{d}X + [A, X]; Y, Z)$. The integrand of $(\delta c)(X, Y, Z)/N$ splits into a part without A and a part linear in A.

The part without A. Put $P = \operatorname{tr}(\mathrm{d}X\,\mathrm{d}Y\,\mathrm{d}Z)$ and $Q = \operatorname{tr}(\mathrm{d}X\,\mathrm{d}Z\,\mathrm{d}Y)$. Cyclicity for 1-forms, $\operatorname{tr}(\alpha\,\beta\gamma) = \operatorname{tr}(\beta\gamma\,\alpha)$ (moving a 1-form past a 2-form costs no sign), gives $\operatorname{tr}(\mathrm{d}Y\,\mathrm{d}Z\,\mathrm{d}X) = \operatorname{tr}(\mathrm{d}Z\,\mathrm{d}X\,\mathrm{d}Y) = P$ and $\operatorname{tr}(\mathrm{d}Y\,\mathrm{d}X\,\mathrm{d}Z) = \operatorname{tr}(\mathrm{d}Z\,\mathrm{d}Y\,\mathrm{d}X) = Q$. Then

$$\begin{aligned} &B(\mathrm{d}X; Y, Z) - B(\mathrm{d}Y; X, Z) + B(\mathrm{d}Z; X, Y) \\ &\overset{(1)}{=} (P - Q) - (Q - P) + (P - Q) \overset{(2)}{=} 3(P - Q) \overset{(3)}{=} 3\,\mathrm{d}\Big(\operatorname{tr}(X\,\mathrm{d}Y\,\mathrm{d}Z) - \operatorname{tr}(X\,\mathrm{d}Z\,\mathrm{d}Y)\Big). \end{aligned}$$

(1) Definition of B and the cyclic identities just listed.
(2) Arithmetic.
(3) $\mathrm{d}\operatorname{tr}(X\,\mathrm{d}Y\,\mathrm{d}Z) = \operatorname{tr}(\mathrm{d}X\,\mathrm{d}Y\,\mathrm{d}Z)$ because $\mathrm{d}^2 = 0$, and similarly for Q.

It is exact, so it integrates to zero on the closed S^3 (Stokes).

The part linear in A. It collects the $[A, \cdot\,]$ pieces of the L terms and the three bracket terms $-c([X, Y], Z) + c([X, Z], Y) - c([Y, Z], X)$, in which $\mathrm{d}[X, Y] = \mathrm{d}X\,Y - Y\,\mathrm{d}X + X\,\mathrm{d}Y - \mathrm{d}Y\,X$. Every term contains one undifferentiated X, Y or Z. Look at the terms with an undifferentiated X (the other two groups are the same computation with the letters permuted). From $L_X c(Y, Z)$:

$$\operatorname{tr}\big([A, X](\mathrm{d}Y\mathrm{d}Z - \mathrm{d}Z\mathrm{d}Y)\big) = \operatorname{tr}(AX\mathrm{d}Y\mathrm{d}Z) - \operatorname{tr}(XA\mathrm{d}Y\mathrm{d}Z) - \operatorname{tr}(AX\mathrm{d}Z\mathrm{d}Y) + \operatorname{tr}(XA\mathrm{d}Z\mathrm{d}Y).$$

From $-c([X, Y], Z)$ and $+c([X, Z], Y)$, keeping only the pieces $X\,\mathrm{d}Y - \mathrm{d}Y\,X$ of $\mathrm{d}[X, Y]$ and

$X\,\mathrm{d}Z - \mathrm{d}Z\,X$ of $\mathrm{d}[X,Z]$:

$$\begin{aligned}&-\operatorname{tr}\big(A(X\mathrm{d}Y - \mathrm{d}YX)\mathrm{d}Z\big) + \operatorname{tr}\big(A\,\mathrm{d}Z(X\mathrm{d}Y - \mathrm{d}YX)\big)\\ &\qquad + \operatorname{tr}\big(A(X\mathrm{d}Z - \mathrm{d}ZX)\mathrm{d}Y\big) - \operatorname{tr}\big(A\,\mathrm{d}Y(X\mathrm{d}Z - \mathrm{d}ZX)\big)\\ &\overset{(1)}{=} -\operatorname{tr}(AX\mathrm{d}Y\mathrm{d}Z) + \operatorname{tr}(AX\mathrm{d}Z\mathrm{d}Y) - \operatorname{tr}(A\,\mathrm{d}Z\mathrm{d}YX) + \operatorname{tr}(A\,\mathrm{d}Y\mathrm{d}ZX).\end{aligned}$$

(1) Multiply out. The terms $+\operatorname{tr}(A\,\mathrm{d}YX\mathrm{d}Z)$ and $-\operatorname{tr}(A\,\mathrm{d}YX\mathrm{d}Z)$ cancel, and so do $\pm\operatorname{tr}(A\,\mathrm{d}ZX\mathrm{d}Y)$.

Adding the two displays, the $\operatorname{tr}(AX\cdots)$ terms cancel, leaving

$$-\operatorname{tr}(XA\,\mathrm{d}Y\mathrm{d}Z) + \operatorname{tr}(XA\,\mathrm{d}Z\mathrm{d}Y) - \operatorname{tr}(A\,\mathrm{d}Z\mathrm{d}YX) + \operatorname{tr}(A\,\mathrm{d}Y\mathrm{d}ZX) \overset{(1)}{=} 0,$$

(1) Cyclicity moves the 0-form X from the end to the front with no sign: $\operatorname{tr}(A\,\mathrm{d}Y\mathrm{d}ZX) = \operatorname{tr}(XA\,\mathrm{d}Y\mathrm{d}Z)$ and $\operatorname{tr}(A\,\mathrm{d}Z\mathrm{d}YX) = \operatorname{tr}(XA\,\mathrm{d}Z\mathrm{d}Y)$.

So $\delta c = 0$: (10.279) is a cocycle. (A numerical check with random $\mathfrak{su}(3)$-valued fields on a lattice confirms it to rounding error.)

Worked example 10.83: The abelian Faddeev–Mickelsson term is a coboundary

For $K = \mathrm{U}(1)$ everything commutes, and the 1-forms anticommute: $\mathrm{d}Y\,\mathrm{d}X = -\mathrm{d}X\,\mathrm{d}Y$. So (10.279) becomes

$$c(A;X,Y) \overset{(1)}{=} 2N\int A\,\mathrm{d}X\,\mathrm{d}Y \overset{(2)}{=} 2N\int X\,\mathrm{d}A\,\mathrm{d}Y \overset{(3)}{=} 2N\int X\,\mathrm{d}Y\,\mathrm{d}A.$$

(1) $\mathrm{d}X\mathrm{d}Y - \mathrm{d}Y\mathrm{d}X = 2\,\mathrm{d}X\mathrm{d}Y$.
(2) $\mathrm{d}(XA\,\mathrm{d}Y) = \mathrm{d}X\,A\,\mathrm{d}Y + X\,\mathrm{d}A\,\mathrm{d}Y = -A\,\mathrm{d}X\mathrm{d}Y + X\,\mathrm{d}A\,\mathrm{d}Y$; integrate over the closed S^3 (Stokes).
(3) A 2-form commutes with a 1-form.

It is not constant, but it is trivial. Try $\alpha(X) = \beta\int X\,A\,\mathrm{d}A$. Under $A \mapsto A + t\,\mathrm{d}X$ only the undifferentiated A changes (to first order $\mathrm{d}A$ is unchanged), so

$$(\delta\alpha)(X,Y) \overset{(1)}{=} L_X\alpha(Y) - L_Y\alpha(X) \overset{(2)}{=} \beta\int\big(Y\,\mathrm{d}X\,\mathrm{d}A - X\,\mathrm{d}Y\,\mathrm{d}A\big) \overset{(3)}{=} -2\beta\int X\,\mathrm{d}Y\,\mathrm{d}A.$$

(1) Definition of $\delta\alpha$; $[X,Y] = 0$.
(2) $L_X\alpha(Y) = \beta\int Y\,\mathrm{d}X\,\mathrm{d}A$, and similarly for $L_Y\alpha(X)$.
(3) $\mathrm{d}(YX\,\mathrm{d}A) = (\mathrm{d}Y\,X + Y\,\mathrm{d}X)\mathrm{d}A$ integrates to zero, so $\int Y\,\mathrm{d}X\,\mathrm{d}A = -\int X\,\mathrm{d}Y\,\mathrm{d}A$.

With $\beta = N$, $c + \delta\alpha = 0$, so by (10.264) the redefined generators $\widehat{G}(X) + N\int XA\,\mathrm{d}A$ close without a Schwinger term. The same happens in $1+1$ dimensions: for U(1) the central term $K\int X\,\mathrm{d}Y$ of Example 10.79 equals $-\delta\alpha$ for $\alpha(X) = \frac{K}{2}\int XA$ (the same two-line computation with $\mathrm{d}A$ removed). An A-dependent redefinition can remove what no constant can. In both abelian cases the topology agrees: $\pi_3(\mathrm{U}(1)) = \pi_5(\mathrm{U}(1)) = 0$, so $H^3(\mathcal{A}/\mathcal{G};\mathbb{Z})$ has no class to carry an obstruction. (For abelian theories the Euclidean anomaly is also invisible to the topology of $\mathcal{G}$; it lives in *local* cohomology, as Nash remarks in §2. On the Hamiltonian side the price of the redefinition is that the new generators contain a Chern–Simons-like functional of A, so they no longer generate the plain gauge transformations.)

space	Schwinger term	extension	topology	obstruction?
S^1, U(1)	$\frac{1}{2\pi i}\int X\,\mathrm{d}Y$ (const.)	central	$\pi_3(\mathrm{U}(1)) = 0$	no
S^1, SU(N)	$\frac{1}{2\pi i}\int \mathrm{tr}(X\,\mathrm{d}Y)$ (const.)	central	$\pi_3 = \mathbb{Z}$	yes
S^3, U(1)	$2N\int X\,\mathrm{d}Y\,\mathrm{d}A$	by Map($\mathcal{A}$, U(1))	$\pi_5(\mathrm{U}(1)) = 0$	no
S^3, SU(2)	0	–	$\pi_5 = \mathbb{Z}_2$	only global
S^3, SU($N \geq 3$)	(10.279), $\propto d_{abc}$	by Map($\mathcal{A}$, U(1))	$\pi_5 = \mathbb{Z}$	yes

The fourth column is (10.240) in general form: on S^d, $\pi_3(\mathcal{A}/\mathcal{G}) = \pi_2(\mathcal{G}) = \pi_{d+2}(K)$.

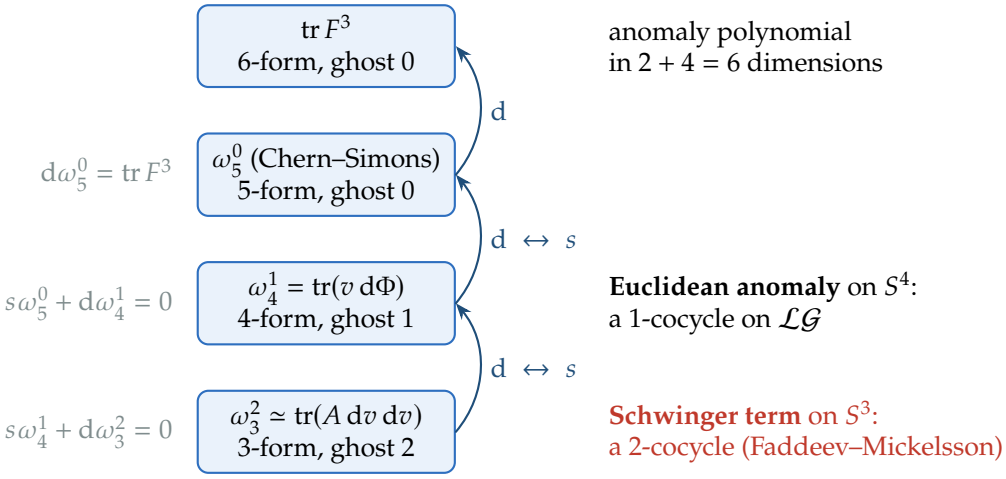


Figure 10.57: The descent ladder. Each rung loses one form degree and gains one ghost number; the equations on the left link neighbouring rungs. Read as cochains, the 4-form rung is the Euclidean (Lagrangian) anomaly in four dimensions, and the 3-form rung below it is the Hamiltonian anomaly, the Schwinger term, on three-dimensional space. This is the cohomological form of Nash's "$n + p$" bookkeeping: the degree of the anomaly polynomial is fixed, and it can be spent on spacetime or on parameters.

10.10.20 Gauss's law cannot be imposed

Nash ends the chapter with a physical consequence: "Gauss's law is violated for theories with this anomaly". We now make the statement precise.

Gauss's law as an operator statement. In temporal gauge $A_0 = 0$ the classical equations of a Yang–Mills field coupled to a charge density ρ^a include one without time derivatives, Gauss's law $(D_iE_i)^a = \rho^a$. Time evolution preserves it but cannot impose it, so in the quantum theory it becomes a condition on states. The electric field is the momentum conjugate to A: $E_i^a(x) = -i\,\delta/\delta A_i^a(x)$ acting on wave functionals $\Psi[A]$. The bosonic generator of gauge transformations is then

$$L_X \overset{(1)}{=} \int \mathrm{d}^3x\,(D_iX)^a\,\frac{\delta}{\delta A_i^a} \overset{(2)}{=} i\int \mathrm{d}^3x\,(D_iX)^a\,E_i^a \overset{(3)}{=} -i\int \mathrm{d}^3x\,X^a\,(D_iE_i)^a,$$

(1) The derivative along $\delta_X A = D_A X$, (10.260), written in components.

(2) $\delta/\delta A = iE$.

(3) Integrate by parts on the closed S^3. The derivative part is ordinary integration by parts; for the commutator part, with real structure constants $[T_b, T_c] = f_{bca}T_a$ totally antisymmetric, $f_{abc}A_i^bX^cE_i^a = -X^c\,f_{cba}A_i^bE_i^a$, which is $-X^c$ times the commutator part of $(D_iE_i)^c$.

so L_X is ($-i$ times) the smeared generator $\int X \cdot D \cdot E$. The full Gauss generator also rotates the fermions: $\widehat{G}(X) = L_X + :\widehat{\rho}_A(X)$: as in (10.259), the second term being the regularised fermion charge density smeared with X. **Gauss's law** is the requirement

$$\widehat{G}(X)\,\Psi = 0 \qquad \text{for all } X \in \mathcal{LG}. \tag{10.280}$$

Exponentiating, it says that Ψ is invariant under every gauge transformation in the identity component of $\mathcal{G}$: physical states are gauge-invariant wave functionals, i.e. they are really functions on $\mathcal{A}/\mathcal{G}$ (sections of $\mathfrak{F}/\mathcal{G}$, in Nash's language). This is how the configuration space $\mathcal{A}/\mathcal{G}$ of Chapter 10 enters the Hamiltonian picture.

The obstruction. Suppose (10.280) holds. Apply two generators:

$$0 \overset{(1)}{=} \widehat{G}(X)\widehat{G}(Y)\Psi - \widehat{G}(Y)\widehat{G}(X)\Psi \overset{(2)}{=} \widehat{G}([X,Y])\Psi + c(X,Y)\Psi \overset{(3)}{=} c(X,Y)\,\Psi.$$

(1) Each term is zero by (10.280).
(2) The commutator (10.262).
(3) (10.280) for the element $[X, Y]$.

Since $c(X, Y)$ acts by multiplication by the function $c(A; X, Y)$, this says $c(A; X, Y)\,\Psi[A] = 0$ for every A, X, Y. So $\Psi[A] = 0$ wherever $c(A; \cdot, \cdot) \neq 0$. For SU(3) and the cocycle of Example 10.82: choosing X and Y concentrated near a point x_0, with prescribed gradients there, shows that $c(A; \cdot, \cdot) = 0$ forces $d_{abc}A^c_i(x_0) = 0$ for all a, b, i. That forces $A_i(x_0) = 0$: if $x = A_i(x_0)$ (a traceless anti-Hermitian matrix) had $\mathrm{tr}(T_c\{T_a, x\}) = 0$ for all a, c, then the traceless part of $\{T_a, x\}$ would vanish for all a; taking $T_a \propto x$ gives $x^2 \propto \mathbb{1}$. The eigenvalues of the anti-Hermitian x are $i\mu_1, i\mu_2, i\mu_3$ with μ_k real, so $x^2 \propto \mathbb{1}$ means $|\mu_1| = |\mu_2| = |\mu_3|$; but three numbers $\pm\mu$ cannot sum to zero (as $\mathrm{tr}\,x = 0$ requires) unless $\mu = 0$. So Ψ vanishes at every $A \neq 0$, and by continuity everywhere. *There are no physical states.* This is the infinite-dimensional version of Example 10.76, and of the impossibility of $\hat{x}\psi = \hat{p}\psi = 0$ in Example 10.78.

The escape routes are exactly the two of Section 10.10.17. If c is a coboundary, $c = -\delta\alpha$, impose instead the modified law $(\widehat{G}(X) + \alpha(X))\Psi = 0$, whose generators close; this is what happens in Example 10.83. If c is not a coboundary, there is no consistent Gauss law and the gauge theory does not exist as a quantum theory, unless the fermion content is changed so that the cocycles of the different species cancel.

Nash's group-level phrasing. Nash says: "the non-triviality of the extension cocycle means that the identity of $\mathcal{G}$ is covered by a non-trivial element of $\mathrm{Map}(\mathcal{A}, S^1)$ ... the identity element acts non-trivially and will change the vacuum". Here is that sentence as a computation. Take two gauge transformations g, h that commute in $\mathcal{G}$. Then $gh = hg$, and (10.257) applied twice gives two lifts of the same element:

$$U_{h\cdot A}(g)\,U_A(h) \overset{(1)}{=} \omega(g,h;A)\,U_A(gh) \overset{(2)}{=} \frac{\omega(g,h;A)}{\omega(h,g;A)}\,U_{g\cdot A}(h)\,U_A(g).$$

(1) (10.257) for the pair (g, h).
(2) (10.257) for the pair (h, g) reads $U_{g\cdot A}(h)U_A(g) = \omega(h, g; A)U_A(hg)$, and $U_A(hg) = U_A(gh)$.

So the lifted group commutator $\widehat{U}(g)\widehat{U}(h)\widehat{U}(g)^{-1}\widehat{U}(h)^{-1}$, which lies over the *identity* of $\mathcal{G}$, acts on the fibre over $gh \cdot A$ by the phase $\omega(g, h; A)/\omega(h, g; A)$. If this is not identically 1, a wave functional invariant under all the $\widehat{U}$ would have to satisfy $\Psi = \frac{\omega(g,h;\cdot)}{\omega(h,g;\cdot)}\Psi$, hence vanish wherever the phase differs from 1. The identity of the gauge group, lifted to the Fock bundle, is not the

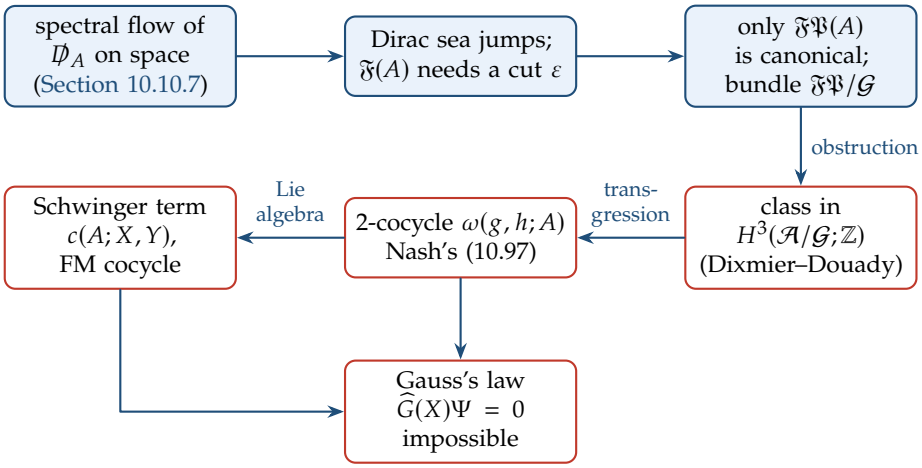


Figure 10.58: The logic of the Hamiltonian picture of anomalies, in the order Nash presents it. Blue boxes are constructions, red boxes are the anomaly in its successive disguises: a degree-3 class on the orbit space, a multiplier on the gauge group, a Schwinger term in the current algebra, and finally the failure of Gauss's law.

identity operator: it multiplies states by a non-constant phase. Infinitesimally, with $g = e^{sX}$ and $h = e^{tY}$, the phase is $1 + st\,c(A;X,Y) + \dots$ and we are back to the Schwinger term.

Reading Nash (pp. 297–300, (10.86)–(10.97))

Nash's example "Fock space and Gauss's law" starts with the counting $n+p = 4+2 = 3+3$ of (10.240). (10.86)–(10.88) are the polarisation, the Fock space (10.241) and the Fock bundles (10.243). His "flaw" (zero eigenvalues make the decomposition jump) and the cure by cuts at ε, ε', (10.89)–(10.93), are Section 10.10.13 and (10.245)–(10.247); the isomorphism I_p^λ of (10.94) is (10.248), worked out in Examples 10.73 and 10.74. The projective bundle (10.95)–(10.96) is (10.249)–(10.250). His (10.97), $g\cdot(h\cdot f_A) = \lambda(A)(gh)\cdot f_A$, is our (10.257) with $\lambda(A) = \omega(\cdot,\cdot;A)$. Nash states without computation that this gives the non-central extension of Chapter 6 and that it is "responsible for the Schwinger term" in the current algebra; we derived the twisted cocycle condition (10.258), its Lie-algebra form (10.263), computed Schwinger terms from the CAR (Section 10.10.18) and obtained the Faddeev–Mickelsson cocycle (10.279) by descent. His final paragraph on Gauss's law (p. 300) is Section 10.10.20. Nash refers to Pressley and Segal for the Fock-space constructions, and to Faddeev, Mickelsson and Segal for the cocycle.

Recap: Section 6: anomalies from a Hamiltonian perspective

- A Berry phase is the holonomy of a line bundle of ground states (Example 10.58); anomalies in the Hamiltonian picture are the same phenomenon for a Dirac sea.
- A p-parameter family of n-dimensional Dirac operators is one operator in $n+p$ dimensions (10.225); only $n+p$ matters, and the anomaly is measured by $\pi_{n+p}(\Omega\mathrm{U}(\infty))$.
- Fredholm operators classify K-theory, $[X,\mathcal{F}] = K(X)$ (10.226); self-adjoint ones are the odd partner, $\mathcal{F}^1 \simeq \mathrm{U}(\infty)$, with $\pi_1(\mathcal{F}^1) = \mathbb{Z}$ given by spectral flow; the real case gives $\mathbb{Z}_2$ and Witten's anomaly.
- On the circle, $-i\,\mathrm{d}/\mathrm{d}\theta + kt$ has spectral flow k, equal to the index on the torus (computed with theta functions) and to the number of particles created from the Dirac sea; the vacuum charge is $-\frac{1}{2}\eta$.

- Fock spaces $\Lambda^*(H^+ \oplus \bar{H}^-)$ depend on a polarisation; changing the cut changes them by a determinant line, so only the projective Fock bundle $\mathfrak{F}\mathfrak{P}/\mathcal{G}$ is canonical. Its obstruction to being a vector bundle is a class in $H^3(\mathcal{A}/\mathcal{G};\mathbb{Z})$.
- Lifting gauge transformations to vectors gives multipliers $\omega(g,h;A)$ obeying the twisted cocycle condition (10.258): an extension of $\mathcal{G}$ by $\mathrm{Map}(\mathcal{A},\mathrm{U}(1))$. Infinitesimally, a Schwinger term $c(A;X,Y)$ obeying (10.263).
- Schwinger terms are finite remainders of infinite traces, $c(X,Y) = \sum X_{ij}Y_{ji}(n_i - n_j)$. In $1+1$ they are constant (Kac–Moody); in $3+1$ they depend on A (Faddeev–Mickelsson, $\propto d_{abc}$), because gauge transformations fail the Hilbert–Schmidt test.
- A non-trivial cocycle forbids Gauss's law: $c(X,Y)\Psi = 0$ forces $\Psi = 0$. Anomaly cancellation is the same d_{abc} condition as in the Euclidean picture.

Chapter summary

This chapter asked one question in several guises: *can a symmetry of the classical theory survive quantisation?* For chiral fermions the quantum theory is built from a determinant (Euclidean picture) or from a Dirac sea (Hamiltonian picture), and both are defined only up to choices that the symmetry group may be unable to make consistently. An anomaly is the obstruction, and it is topological. The answers came in four layers.

1. *Local anomalies* (Nash §§1–3). The fermion integral gives $\sqrt{\det(\not{D}_A^* \not{D}_A)}$, a function of the gauge field that should be gauge invariant but, after regularisation, is not. Its variation along a gauge orbit is the anomalous Ward identity, computed by the triangle graph, by point splitting, or by Fujikawa's Jacobian with a heat-kernel regulator; its integral is the index of the Dirac operator. Topologically the determinant is a section of the determinant line bundle $\det \operatorname{Index} \not{D}$ over $\mathcal{A}/\mathcal{G}$, whose first Chern class is computed by the families index theorem; the anomaly is that class, or equivalently the 1-cocycle μ_1 on $\mathcal{G}$. Gravity is the same story with $\mathrm{Met}(M)/\mathrm{Diff}(M)$ in place of $\mathcal{A}/\mathcal{G}$ and $\hat{A}$ in place of ch, possible only in dimensions $4k-2$.

2. *The conformal anomaly and the critical dimension* (Nash §4). For the bosonic string the relevant group is the Weyl group $C^\infty_+(\Sigma)$ and the relevant family is the family of $\bar\partial$-operators over $\mathrm{Met}(\Sigma)/C^\infty_+$. The matter and ghost determinants together form a section of a determinant line whose first Chern class is $\frac{26-d}{24}\int x^2$: it vanishes exactly for $d = 26$. The same number comes out of the heat kernel, the central charge, Mumford's formula and the light-cone zero-point energy.

3. *Global anomalies* (Nash §5). When the symmetry group is disconnected, a determinant can change sign (or phase) between components while being invariant under infinitesimal transformations. For SU(2), $\pi_4(\mathrm{SU}(2)) = \mathbb{Z}_2$, a mod-2 index and a spectral flow make $\sqrt{\det}$ change sign, and an odd number of doublets is inconsistent. For gravity, large diffeomorphisms (linked to exotic spheres) are tested with the eta invariant of a mapping torus, computed by Atiyah–Patodi–Singer; Bismut and Freed identified global anomalies with holonomies of the determinant line.

4. *The Hamiltonian picture* (Nash §6). A p-parameter family in n dimensions is one operator in $n + p$ dimensions, and the anomaly is classified by $\pi_{n+p}(\Omega\mathrm{U}(\infty))$ through Fredholm operators. In canonical quantisation spectral flow moves levels across the Dirac sea, Fock spaces are defined only projectively, and the gauge group acts only up to phases depending on A: a 2-cocycle, infinitesimally the Schwinger term (constant Kac–Moody term in 1 + 1, Faddeev–Mickelsson term in 3 + 1). A non-trivial cocycle forbids Gauss's law.

The tables below list every concept of the chapter, what it means operationally (how to compute it or test for it), and where it is used again.

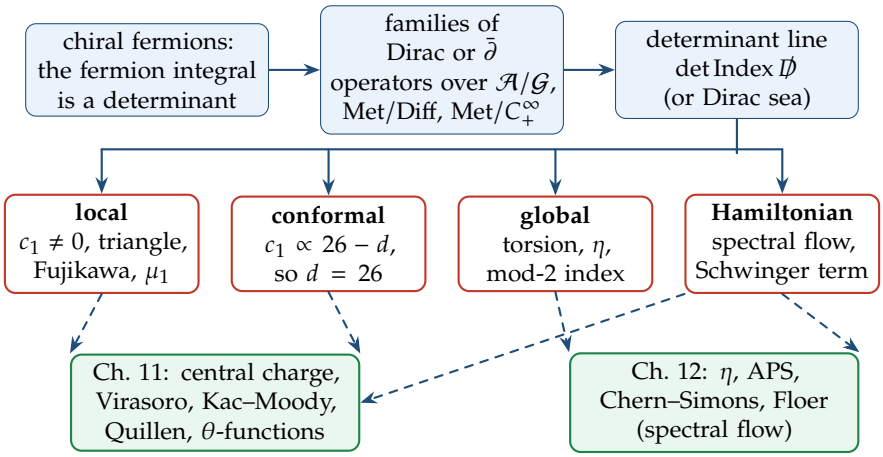


Figure 10.59: The chapter in one picture. Integrating out chiral fermions produces determinants of families of operators; these are sections of a determinant line bundle, and every kind of anomaly (red) is a topological property of that bundle or of its Hamiltonian counterpart. Dashed arrows point to the chapters where the same structures return.

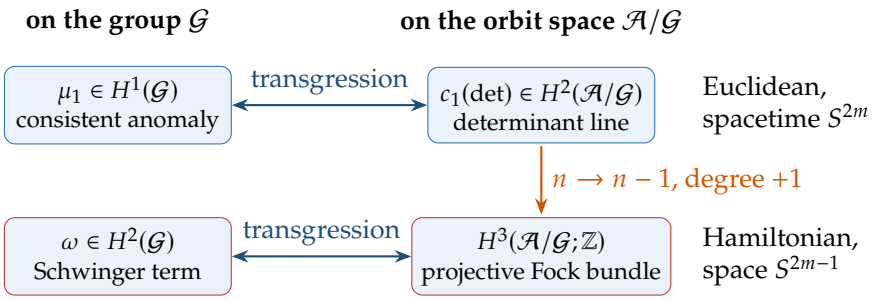


Figure 10.60: Where the anomaly lives, in the two pictures. Horizontally, transgression (Stokes' theorem on a lifted ball, Figure 10.53) trades a class on the orbit space for a class one degree lower on the group. Vertically, passing from spacetime to space lowers the dimension by one and raises the degree by one, keeping $n + p$ fixed.

Local anomalies in gauge theory (Nash §§1–2)

concept	operational meaning	used later
anomaly	a classical symmetry whose quantum (regularised) version cannot be kept; test: vary $\log Z$ along the symmetry and find a non-zero local term that no counterterm removes	throughout; Chapter 11, Chapter 12
chirality, γ_5, Weyl spinors	$\gamma_5 = (-i)^m \gamma_1 \cdots \gamma_{2m}$ splits spinors into ± 1 eigenspaces; explicit 2×2 and 4×4 matrices	Chapter 11 (chiral fields)
ABJ anomaly	$\partial_\mu j_5^\mu \propto \epsilon FF$; explains $\pi^0 \to 2\gamma$	Chapter 12 (instanton effects)
chiral Dirac operator	$\not{D}_A : \Gamma(E^+ \otimes F) \to \Gamma(E^- \otimes F)$, needs a spin structure ($w_1 = w_2 = 0$)	Chapter 12
index of $\not{D}_A$	$n_+ - n_- = \int \hat{A}\ \mathrm{ch}(F)$; on S^{2m} just $\int \mathrm{ch}_m(F) = k$	Chapter 12 (Donaldson)
Grassmann integral	$\int \mathrm{d}\bar\psi\, \mathrm{d}\psi\, e^{-\bar\psi M \psi} = \det M$; fermions give determinants	Chapter 11
gauge covariance	$\not{D}_{A^g} = g^{-1} \not{D}_A g$, so formally det is invariant; regularisation spoils it	
zeta-function determinant	$\det \Delta = e^{-\zeta'_\Delta(0)}$; not multiplicative	Chapter 11 (Quillen metric)
anomalous Ward identity	$\delta_f \log Z = \int f^a \langle (\nabla_\mu j_5^\mu)^a \rangle$, non-zero	
triangle graph, point splitting	one-loop computations of $\langle \partial j_5 \rangle$; the surface term of a linearly divergent integral	
Fujikawa Jacobian	$\lim_{M\to\infty} \mathrm{tr}\, \gamma_5 e^{\not{D}^2/M^2}$ = index density (heat kernel)	Chapter 12
integrated anomaly	$\int \langle \partial j_5 \rangle = 2\,$index: the anomaly counts zero modes	
consistent vs covariant anomaly	Wess–Zumino consistent ($\delta a = 0$) vs gauge covariant; differ by a local polynomial	
BRST and descent	$\mathrm{tr}\, F^{m+1} \to \omega^0_{2m+1} \to \omega^1_{2m} \to \omega^2_{2m-1}$ by d and s; the rung ω^1_{2m} is the anomaly	Chapter 12 (Chern–Simons)
anomaly cancellation	$\sum_R \pm \mathrm{tr}_R(T_a\{T_b, T_c\}) = 0$ (d_{abc} condition); holds in the Standard Model generation by generation	
T_g and μ_1	$\mu_1 = \mathrm{d}\log\det T_g$ on $\mathcal{G}$; closed; exact iff no anomaly; $[\mu_1] \in H^1(\mathcal{G})$	
$\pi_1(\mathcal{G}) = \pi_{n+1}(K)$	on S^4: $\pi_5(\mathrm{U}(N)) = 0, \mathbb{Z}_2, \mathbb{Z}$ for $N = 1, 2, \geq 3$	global anomalies (§5)
determinant line of a map	$\det O \in (\Lambda^{\max} \ker O)^* \otimes \Lambda^{\max} \ker O^*$; defined even when ker jumps	Chapter 11
det Index $\not{D}$	line bundle over $\mathcal{A}/\mathcal{G}$ whose sections are "determinants"; anomaly-free iff trivial	Chapter 11, Chapter 12
families index theorem	$\mathrm{ch}(\mathrm{Index}\, \not{D}) = \int_M \mathrm{ch}(E)\hat{A}$; gives $c_1(\det)$	Chapter 11
local cohomology	anomalies modulo local counterterms; sees the abelian anomaly that $H^*(\mathcal{G})$ misses	

Gravitational anomalies (Nash §3)

concept	operational meaning	used later
$\not{D}_\rho$ on $\mathrm{Met}(M)$	Dirac operator of the metric ρ; $\not{D}_{h\cdot\rho} = h^{-1}\not{D}_\rho h$ for $h \in \mathrm{Diff}_+(M)$	§5 (global)
$\mathrm{Met}(M)/\mathrm{Diff}_0(M)$	the gravitational orbit space; Diff_0 fixes a frame at a point so the action is free	Chapter 11 (moduli of surfaces)
gravitational anomaly	$\mathrm{ch}(\det \mathrm{Index}) = \int_M \hat{A}$; degree 2 only if $n = 4k-2$	
spin-$\frac{3}{2}$, self-dual forms	further anomaly polynomials (Alvarez-Gaumé–Witten); needed for cancellation	§5 (superstring)
2D gravitational anomaly	$\propto c_L - c_R$	Chapter 11
Lorentz anomaly	trade Diff for local frame rotations $\mathcal{G}_L$ in $\mathrm{Aut}(P) = \mathcal{G}_L \rtimes \mathrm{Diff}_0$	
Green–Schwarz mechanism	for SO(32) and $E_8 \times E_8$ the 12-form factorises and is cancelled by a 2-form field	§5

The critical dimension of strings (Nash §4)

concept	operational meaning	used later
Polyakov action, Weyl symmetry	$S = \frac{1}{2}\int \sqrt{g}\, g^{ab}\partial_a X \cdot \partial_b X$ is invariant under $g \to e^{2\sigma} g$ in 2D	Chapter 11
conformal (Weyl) anomaly	$\delta_\sigma \log \det{}'\Delta_g \neq 0$; for d bosons $\delta \log Z = \frac{d}{24\pi}\int \sqrt{g}\,\sigma R$	Chapter 11 (central charge)
heat kernel, $\zeta(0)$	$\mathrm{tr}\, e^{-t\Delta} \sim \frac{A}{4\pi t} + \frac{\chi}{6} + \ldots$; $\zeta_\Delta(0) = \frac{\chi}{6} - \dim\ker$ ($= -\frac{2}{3}$ on S^2)	Chapter 11, Chapter 12
Liouville action	integrate the anomaly: $\log Z[e^{2\sigma}g] - \log Z[g] \propto \int(\lvert\nabla\sigma\rvert^2 + R\sigma)$	Chapter 11
ghost determinant	$\det(\bar\partial^*\bar\partial)$ on K^* from gauge fixing; contributes -26	Chapter 11 (bc system)
bc central charge	$c = -2(6\lambda^2 - 6\lambda + 1)$; $\lambda = 2$ gives -26	Chapter 11
virtual bundle $K^* - \frac{d}{2}I$	raising a determinant to a power = tensoring the family with a virtual bundle	
Nash's families computation	$c_1(\det \mathrm{Index}\,\bar\partial_Y) = \int \mathrm{ch}(K^* - \frac{d}{2})\,\mathrm{td} = \frac{26-d}{24}\int x^2$	
Mumford's formula	$\lambda_j = \lambda_1^{6j^2-6j+1}$ on moduli space; same polynomial as c	Chapter 11
superstring $d = 10$	add fermions and superghosts: $\frac{3}{2}d - 15 = 0$	Chapter 11
light-cone route	$\frac{d-2}{2}\sum n = -\frac{d-2}{24}$ with $\zeta(-1) = -\frac{1}{12}$; massless vector needs $d = 26$	Chapter 11
Quillen metric	zeta-regularised norm on the determinant line; curvature = families index form	Chapter 11
Belavin–Knizhnik, Manin	in $d = 26$ the measure is $\lvert f\rvert^2$, f holomorphic; f via theta functions; torus: $\lvert\eta(\tau)\rvert^{-48}$	Chapter 11 (modular invariance)

Global anomalies (Nash §5)

concept	operational meaning	used later
global anomaly	Z changes under a gauge transformation not connected to $\mathbb{1}$; invisible to derivatives and de Rham cohomology	Chapter 12
$\pi_0(\mathcal{G}) = \pi_{2m}(K)$	components of the gauge group on S^{2m}; $\pi_4(\mathrm{SU}(2)) = \mathbb{Z}_2$ (suspended Hopf map, framed circles)	
toy global anomaly	charge-$\frac{1}{2}$ fermion on a circle: $Z = 2\cos\frac{a}{2}$ changes sign under $a \to a + 2\pi$	
pseudo-reality, Pfaffian	$J = i\sigma_2 K$ pairs eigenvalues (Kramers); $\sqrt{\det}$ is a Pfaffian, with a sign	
mod-2 index	dim ker mod2 of a real skew operator; deformation invariant	Chapter 12
spectral flow	net number of eigenvalues crossing 0 along a path; for Witten's loop it is odd	Chapter 12 (Floer)
Witten's SU(2) anomaly	odd number of doublets (more generally $\sum 2T(R)$ odd) is inconsistent	
torsion, UCT, Möbius bundle	$H^2(\mathcal{A}/\mathcal{G};\mathbb{Z}) = \mathbb{Z}_2$ is pure torsion; the real determinant line is a Möbius bundle	
large diffeomorphisms	$\pi_0(\mathrm{Diff}_+(S^n))$ = exotic $(n+1)$-spheres (Kervaire–Milnor); $\mathbb{Z}_{992}$ for $n = 10$	
mapping torus	glue $M \times [0,1]$ by f; the path of metrics becomes one metric on an $(n+1)$-manifold	Chapter 12
eta invariant	$\eta(0) = \sum \operatorname{sgn}\lambda\lvert\lambda\rvert^{-s}\big\rvert_{s=0}$; phase of a determinant $e^{i\pi\eta/2}$	Chapter 12 (Chern–Simons phase)
APS theorem	index $= \int_X \hat{A} - \frac{h+\eta}{2}$ on a manifold with boundary	Chapter 12
adiabatic limit, Witten's formula	scale the fibre metric; $\Delta S = \frac{i\pi}{2}\eta$ of the mapping torus	
superstring global check	$n_d = 495$, $n_{rs} = 1$, $n_{sd} = 0$: $\Delta S = 2\pi i k$, no anomaly	
Bismut–Freed	global anomaly = holonomy of the natural connection on det; computed by $\lim e^{i\pi(h+\eta)}$	Chapter 12
complex anomalies	distinct complex structures on $\mathrm{U}(4)/(\mathrm{U}(2)\times\mathrm{U}(1)^2)$ with $c_1^5 = 4500, 4860$; $\Delta S = 360$	

The Hamiltonian picture (Nash §6)

concept	operational meaning	used later
Berry phase	$\gamma = \oint i\langle\psi\|\mathrm{d}\psi\rangle$; for spin $\frac{1}{2}$, $-\frac{1}{2}\times$solid angle; holonomy of the ground-state line	Chapter 12
$n+p$ rule	a p-parameter family in n dimensions = one operator $\not{D}_t^{(n)} + \sum \gamma_i \partial_{t_i}$ in $n+p$	
Fredholm operators $\mathcal{F}(H)$	finite kernel and cokernel; $[X, \mathcal{F}] = K(X)$ (Atiyah–Jänich); $\mathcal{F} \simeq \Omega\mathrm{U}(\infty)$	Chapter 12
self-adjoint Fredholm $\mathcal{F}^1$	$\mathcal{F}^1_* \simeq \mathrm{U}(\infty)$; $\pi_k(\mathcal{F}) = \mathbb{Z}, 0$ and $\pi_k(\mathcal{F}^1) = 0, \mathbb{Z}$ (k even, odd)	Chapter 12
real case	$\mathcal{F}^1_{\mathrm{skew}}(H_r) \simeq \mathrm{O}(\infty)$, $\pi_0 = \mathbb{Z}_2$: Witten's anomaly again	
spectral flow on the circle	$-i\partial_\theta + kt$: eigenvalues $n + kt$, flow k = index on T^2 (theta functions)	Chapter 11
Dirac sea, vacuum charge	fill negative levels; $Q_{\mathrm{vac}} = -\frac{1}{2}\eta$; spectral flow creates k particles	Chapter 11
Fock space	$\Lambda^*(H^+ \oplus \bar{H}^-)$ for a polarisation; changing the cut costs $\Lambda^{\mathrm{top}} J_{\varepsilon\varepsilon'}$	Chapter 11
projective Fock bundle	$\mathfrak{F}\mathfrak{P}/\mathcal{G} \to \mathcal{A}/\mathcal{G}$; obstruction to lifting = Dixmier–Douady class in H^3	
transgression	$H^3(\mathcal{A}/\mathcal{G}) \to H^2(\mathcal{G})$ by Stokes on a lifted ball	
projective representation	$U(g)U(h) = \omega(g,h)U(gh)$; test: commutator phase $\neq 1$ (Pauli, spin $\frac{1}{2}$, Heisenberg)	Chapter 11
Nash's (10.97)	A-dependent multipliers; twisted cocycle; extension of $\mathcal{G}$ by $\mathrm{Map}(\mathcal{A}, \mathrm{U}(1))$	Chapter 11
Schwinger term	$[\widehat{G}(X), \widehat{G}(Y)] = \widehat{G}([X,Y]) + c(A;X,Y)$; $c = \sum X_{ij}Y_{ji}(n_i - n_j)$	Chapter 11 (Virasoro)
U(1) current algebra, Kac–Moody	$[\rho_m, \rho_n] = m\delta_{m+n,0}$; $c = \frac{1}{2\pi i}\oint \mathrm{tr}(X\,\mathrm{d}Y)$	Chapter 11 (WZW)
Shale–Stinespring, HS test	implementable iff off-diagonal blocks Hilbert–Schmidt; passes in 1 + 1, fails in 3 + 1	Chapter 11
Faddeev–Mickelsson cocycle	$N\int \mathrm{tr}(A(\mathrm{d}X\mathrm{d}Y - \mathrm{d}Y\mathrm{d}X))$; next rung of descent; $\propto d_{abc}$; zero for SU(2), trivial for U(1)	
Gauss's law obstruction	$\widehat{G}(X)\Psi = 0$ for all X forces $c\,\Psi = 0$, so $\Psi = 0$ when c is non-trivial	

Key idea: The one sentence to remember from Chapter 10

Integrating out chiral fermions turns a gauge (or diffeomorphism, or Weyl) symmetry into a question about the determinant line bundle of a family of elliptic operators over the orbit space; the anomaly is its topology, computed by the families index theorem (local anomalies, the critical dimension 26), by eta invariants and mod-2 indices (global anomalies), or, in the Hamiltonian picture, by spectral flow and the Schwinger 2-cocycle, and it must cancel for the quantum theory to exist.

Chapter 11

Conformal Quantum Field Theories

Where this chapter is going

Most of this book has been about topology leaking into quantum field theory: winding numbers, indices, anomalies. This chapter is about a *symmetry* that is so strong in two dimensions that it almost solves the theory by itself. The symmetry is **conformal invariance**: invariance under maps that preserve angles but may stretch lengths by a position-dependent amount. In d dimensions such maps form a finite-dimensional group. In two dimensions every holomorphic function is such a map, so the symmetry is infinite dimensional. Its Lie algebra is (a central extension of) the algebra of vector fields on the circle, the **Virasoro algebra** we met in Chapter 6. The single number c in its central term, the **central charge**, turns out to control everything: which operators can exist, what their scaling dimensions are, and, via statistical mechanics, the critical exponents of real magnets.

The logic runs as follows. First (§1) we ask why a physicist should care: critical points of statistical systems and short-distance limits of quantum field theories have no length scale, and scale invariance comes with conformal invariance. We meet Wilson's operator product expansion (OPE), the tool that turns "no length scale" into algebra. Next (§2) we go to two dimensions and use complex coordinates. The central objects are **primary fields**, which transform like tensors of weight $(h, \bar{h})$, and the stress tensor $T(z)$. Conformal Ward identities turn into contour integrals, and the Ward identities become OPEs such as $T(w)\phi(z) \sim h\phi/(w-z)^2 + \partial\phi/(w-z)$. The OPE of T with itself contains c. Then (§3) we expand $T(z)$ in modes L_n and find that they obey the Virasoro algebra. States of the theory are built by acting with L_{-n} on **highest-weight states** $|h\rangle = \phi(0)|0\rangle$. Demanding positive norms (the Kac determinant) forces c and h onto a discrete list when $c < 1$: the **unitary minimal models**, among them the critical Ising model. Later sections of Nash's chapter (§4–§5, the next part of these notes) put the theory on a torus, where modular invariance ties the representations together, and then discuss fusion rules and the axiomatic picture that feeds into knot theory in Chapter 12. This part covers §1–§3. Figure 11.1 shows how the pieces connect.

One physical thread runs through everything. In a theory without a length scale, the only thing a two-point function can be is a power law, $\langle A(x)A(0)\rangle \propto 1/|x|^{2\Delta}$. The exponent Δ (the **scaling dimension**) is the basic datum. The whole chapter is a machine for computing the allowed values of Δ and the rules for multiplying operators. In two dimensions the machine is powerful enough that for $c < 1$ it produces the complete list.

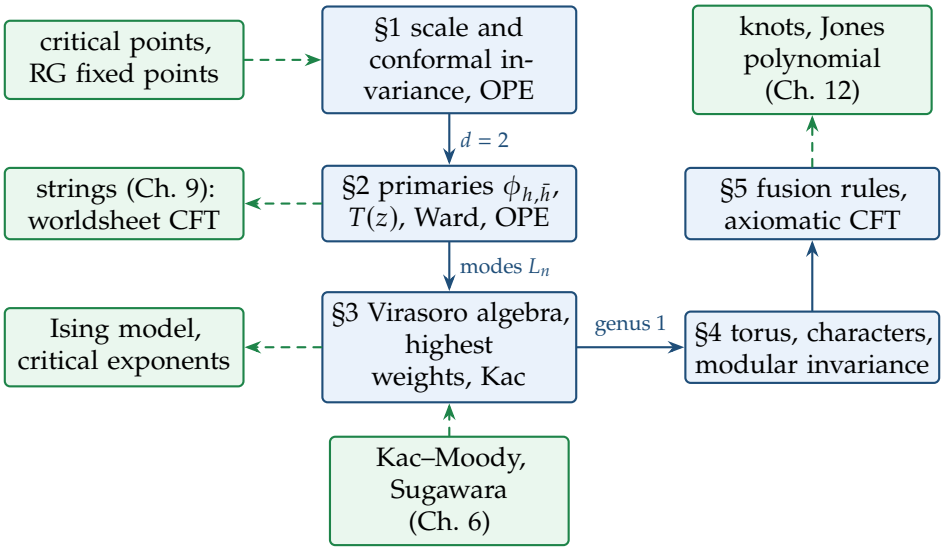


Figure 11.1: The logic of Chapter XI. Blue boxes are Nash's sections; green boxes are the physics they come from or feed. The left column (§1–§3) is this part: from "no length scale" to the Virasoro algebra and its unitary representations. The right column (§4–§5) puts the theory on a torus and on general Riemann surfaces.

Dictionary: the words of this chapter

maths	physics
conformal map	angle-preserving change of coordinates; local rescaling
holomorphic function $z \mapsto f(z)$	2D conformal transformation
weight $(h, \bar{h})$	scaling dimension $h + \bar{h}$, spin $h - \bar{h}$
Virasoro algebra, central charge c	algebra of modes of T; c counts degrees of freedom and measures the Casimir energy
highest-weight (Verma) module	conformal family: a primary and its descendants
unitary representation	theory with positive-norm Hilbert space
Kac determinant	Gram determinant of descendant states

11.1 Conformal invariance and quantum field theory

Nash opens with a remark that sounds negative: a conformally invariant universe would have no masses, so conformal invariance is not a symmetry of the real world. Why then study it? Because there are two regimes where the masses genuinely drop out. One is a statistical system at a second-order phase transition. The other is any quantum field theory probed at distances much shorter than all Compton wavelengths. In both regimes the theory has no length scale, and the absence of a length scale is an extraordinarily strong constraint. This section explains what "no length scale" means operationally, why it tends to upgrade itself to conformal invariance, and what tool (Wilson's operator product expansion) turns it into computable algebra. We end with the conformal group in d dimensions and see why $d = 2$ is special.

11.1.1 Scale transformations and scaling dimensions

Start with dimensional analysis. In units $\hbar = c = 1$ every quantity has a dimension that is a power of length. A mass m has dimension (length)$^{-1}$. A field has a dimension fixed by requiring that the action be dimensionless.

Worked example 11.1: Canonical dimension of a free scalar

The Euclidean action of a free massless scalar in d dimensions is $S = \frac{1}{2}\int \mathrm{d}^d x\, (\partial_\mu \varphi)^2$. We write $[X]$ for the mass dimension of X (so $[x] = -1$, $[\partial] = 1$, $[m] = 1$). Then

$$0 = [S] \overset{(1)}{=} [\mathrm{d}^d x] + 2[\partial] + 2[\varphi] \overset{(2)}{=} -d + 2 + 2[\varphi],$$

(1) The action is a pure number, and dimensions add under multiplication.
(2) $[\mathrm{d}^d x] = -d$ and $[\partial_\mu] = +1$.

So $[\varphi] = (d-2)/2$: dimension 1 in four dimensions, $1/2$ in three, and 0 in two. This is the **canonical dimension** d^c_φ. In two dimensions a free scalar is dimensionless, a fact Nash uses in §2.

A **scale transformation** (or **dilation**) is $x \mapsto \lambda x$ with a constant $\lambda > 0$. In a theory without dimensionful parameters, rescaling lengths must map the theory to itself, provided each field is rescaled by the appropriate power of λ.

Working definition 11.1 (scaling dimension)

A local field $A(x)$ has **scaling dimension** d_A (also written Δ_A) if the unitary operator $U(\lambda)$ implementing dilations [the operator on the space of states that carries out $x \mapsto \lambda x$] acts as

$$U(\lambda)\, A(x)\, U(\lambda)^{-1} = \lambda^{d_A}\, A(\lambda x). \tag{11.1}$$

Operational test: compute the two-point function; if the theory is scale invariant it is a pure power, $\langle A(x)A(0)\rangle = C/|x|^{2d_A}$, and the exponent is $2d_A$. The difference $\gamma_A = d_A - d^c_A$ between the true and the canonical dimension [the value read off from the action by dimensional analysis, $(d-2)/2$ for a scalar] is the **anomalous dimension** [the shift produced by interactions. It is zero in a free theory].

Why does the two-point function have to be a power? Suppose the vacuum is invariant, $U(\lambda)|0\rangle = |0\rangle$. Then

$$\langle A(x)A(0)\rangle \overset{(1)}{=} \langle 0|U(\lambda)^{-1}\,U(\lambda)A(x)U(\lambda)^{-1}\,U(\lambda)A(0)U(\lambda)^{-1}\,U(\lambda)|0\rangle \overset{(2)}{=} \lambda^{2d_A}\,\langle 0|A(\lambda x)A(0)|0\rangle = \lambda^{2d_A}\langle A(\lambda x)A(0)\rangle. \tag{11.2}$$

(1) Insert $1 = U^{-1}U$ three times; the outer factors act on the invariant vacuum and drop out.
(2) Apply (11.1) to each field; note $\lambda \cdot 0 = 0$.

If also rotations are a symmetry, $\langle A(x)A(0)\rangle = G(|x|)$ with $G(r) = \lambda^{2d_A} G(\lambda r)$ for every λ. Choosing $\lambda = 1/r$ gives $G(r) = G(1)\, r^{-2d_A}$: a pure power law. There is no room for an exponential $e^{-r/\xi}$, because ξ would be a length.

Nash's footnote describes $U(\lambda)$ as $\exp[\ln\lambda\,(x\,\partial/\partial x)]$. Here is what that operator does to a

function, which is the reason for the formula. Let $t = \ln\lambda$ and act on a monomial x^n:

$$\exp\!\left[t\,x\frac{\mathrm{d}}{\mathrm{d}x}\right]x^n \overset{(1)}{=} \sum_{k=0}^{\infty}\frac{t^k}{k!}\left(x\frac{\mathrm{d}}{\mathrm{d}x}\right)^k x^n \overset{(2)}{=} \sum_{k=0}^{\infty}\frac{(nt)^k}{k!}\,x^n$$
$$= e^{nt}x^n \overset{(3)}{=} (\lambda x)^n .$$

(1) Definition of the exponential of an operator as its power series.
(2) $x\,\mathrm{d}/\mathrm{d}x$ is the "degree counter": $x\,\mathrm{d}(x^n)/\mathrm{d}x = nx^n$, so x^n is an eigenfunction with eigenvalue n.
(3) $e^{nt} = \lambda^n$.

By linearity, $\exp[\ln\lambda\; x\,\mathrm{d}/\mathrm{d}x]\,f(x) = f(\lambda x)$ for any power series f. So the **dilation generator** is $D = x^\mu\partial_\mu$ (on functions), just as $P_\mu = \partial_\mu$ generates translations, $e^{a\cdot\partial}f(x) = f(x+a)$.

Worked example 11.2: Masses break scale invariance

The Euclidean propagator of a free scalar of mass m in three dimensions is the Yukawa potential $G_m(r) = e^{-mr}/(4\pi r)$. Rescale the separation:

$$G_m(\lambda r) = \frac{e^{-m\lambda r}}{4\pi\lambda r} \overset{(1)}{=} \lambda^{-1}\,\frac{e^{-(\lambda m)r}}{4\pi r} = \lambda^{-1}G_{\lambda m}(r).$$

(1) Regroup $m\lambda r$ as $(\lambda m)r$.

Scale invariance would require $G_m(\lambda r) = \lambda^{-2d_\varphi}G_m(r)$ with the *same* m. This holds for all λ only if $\lambda m = m$, that is $m = 0$. Then $G_0(r) = 1/(4\pi r)$, a power law with $2d_\varphi = 1$, matching the canonical dimension $1/2$ found above. A mass is a length scale (the Compton wavelength $1/m$), and a length scale is exactly what a scale-invariant theory cannot have.

11.1.2 Critical phenomena: where the length scale disappears

Take a ferromagnet, idealised as the **Ising model**: spins $\sigma_i = \pm 1$ on a square lattice with energy $E = -J\sum_{\langle ij\rangle}\sigma_i\sigma_j$. At high temperature the spins are disordered; at low temperature they align. In two dimensions there is a sharp transition at a critical temperature T_c (Onsager: $\sinh(2J/k_BT_c) = 1$). Measure the spin–spin correlation function at large separation r. Away from T_c it decays exponentially,

$$\langle\sigma(r)\sigma(0)\rangle - \langle\sigma\rangle^2 \;\sim\; \frac{e^{-r/\xi}}{r^{\,d-2+\eta}}, \qquad \xi \sim |T - T_c|^{-\nu}. \tag{11.3}$$

The length ξ is the **correlation length**: the size of a typical region of aligned spins. The exponents ν and η are **critical exponents**. As $T \to T_c$, $\xi \to \infty$. At T_c the only lengths left are the lattice spacing a (far too small to matter at large r) and r itself, so the correlator becomes a pure power.

Worked example 11.3: Reading off scaling dimensions from critical exponents

Onsager's solution of the 2D Ising model gives $\eta = 1/4$ and $\nu = 1$. At $T = T_c$, (11.3) with $\xi = \infty$ and $d = 2$ reads $\langle\sigma(r)\sigma(0)\rangle \propto r^{-1/4}$. Compare with the scale-invariant form $r^{-2\Delta_\sigma}$:

$$2\Delta_\sigma \overset{(1)}{=} d - 2 + \eta \overset{(2)}{=} \tfrac{1}{4} \quad\Longrightarrow\quad \Delta_\sigma = \tfrac{1}{8}.$$

(1) Match the power of r in (11.3) at $\xi = \infty$ with (11.2).
(2) $d = 2$, $\eta = 1/4$.

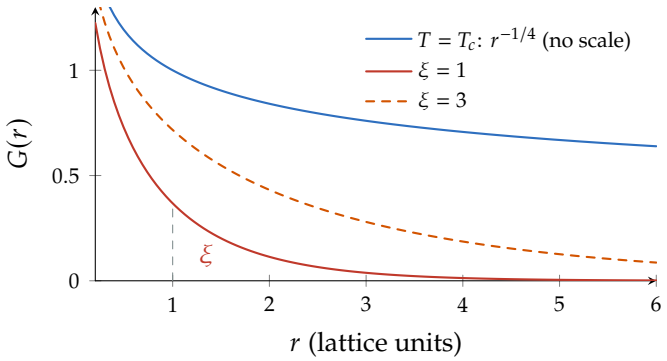


Figure 11.2: Spin–spin correlation of the 2D Ising model (schematic, using the exact exponent $\eta = 1/4$). Away from T_c (red, orange) the correlation dies off after a distance of order ξ. At T_c (blue) $\xi = \infty$ and the decay is a pure power: zooming in or out of the blue curve gives the same shape up to an overall factor. That self-similarity *is* scale invariance.

Now the energy operator ε (the local version of $\sigma_i\sigma_j$). Moving off criticality adds a term $t\int \mathrm{d}^2x\, \varepsilon(x)$ to the action, with $t \propto T - T_c$. The action is dimensionless, so $[t] = d - \Delta_\varepsilon$ in mass units. The only length that can be built from t is $\xi \propto t^{-1/(d-\Delta_\varepsilon)}$. Comparing with $\xi \propto t^{-\nu}$:

$$\nu \overset{(1)}{=} \frac{1}{d-\Delta_\varepsilon} \quad\Longrightarrow\quad \Delta_\varepsilon \overset{(2)}{=} d - \frac{1}{\nu} = 2 - 1 = 1.$$

(1) Dimensional analysis: $t^{1/(d-\Delta_\varepsilon)}$ has dimension of mass, so its inverse is a length.
(2) Solve for Δ_ε and insert $d = 2$, $\nu = 1$.

We will see in §3 that these two numbers, $\frac{1}{8}$ and 1, reappear as $2h$ with $h = \frac{1}{16}$ and $h = \frac{1}{2}$: the weights of the unitary minimal model with $c = \frac{1}{2}$. The critical exponents of a real magnet come out of the representation theory of the Virasoro algebra.

Nash adds a stronger statement: in two dimensions the critical theory is not just scale invariant but *conformally* invariant. Conformal transformations are those that look locally like a rotation times a rescaling, with the rescaling factor allowed to vary from point to point (the precise definition is in Section 11.1.5).

Physical picture: why local physics upgrades scale to conformal

A critical system has no preferred length anywhere. If the physics is also *local* (built from short-range couplings), then different regions do not know about each other's choice of units. So one should be able to rescale by different amounts in different places, as long as the map is locally a rescaling and a rotation and does not shear. Such maps are precisely the conformal maps. Polchinski turned this into a theorem for 2D unitary theories with a discrete spectrum of dimensions: scale plus rotation invariance imply conformal invariance. In higher dimensions it is believed but not proved in general.

11.1.3 Short distances, renormalisation and fixed points

The second regime is the short-distance limit of a massive quantum field theory. Nash states that the product $\varphi(x)\varphi(0)$ becomes independent of the mass as $x \to 0$. Here is what this means in the simplest case.

Worked example 11.4: The mass drops out at short distance

Expand the 3D Yukawa propagator for $mr \ll 1$:

$$G_m(r) = \frac{e^{-mr}}{4\pi r} \overset{(1)}{=} \frac{1}{4\pi r}\left(1 - mr + \tfrac{1}{2}m^2r^2 - \cdots\right)$$
$$= \underbrace{\frac{1}{4\pi r}}_{\text{massless, singular}} - \frac{m}{4\pi} + \frac{m^2 r}{8\pi} - \cdots .$$

(1) Taylor series of e^{-mr}.

The singular part, which dominates as $r \to 0$, is exactly the massless propagator. The mass only appears in terms that are finite or vanishing at $r = 0$. In four dimensions the same thing happens: $G_m(r) = mK_1(mr)/(4\pi^2 r) = 1/(4\pi^2 r^2) + \frac{m^2}{8\pi^2}\ln(mr) + \cdots$, using $K_1(u) = 1/u + \frac{u}{2}\ln(u/2) + \cdots$ for small u. The leading singularity is mass independent. Physically, at distances far below the Compton wavelength a particle does not know it is massive.

For an interacting theory the story is subtler, because the coupling constant itself changes with the scale at which we look: this is the **renormalisation group** (RG). Let us recall just enough of it (see [QFT notes §15.3–15.4] for a full treatment). Renormalised correlation functions depend on an arbitrary renormalisation scale M. Physics cannot depend on M, so changing M must be compensated by changing the renormalised coupling g and the normalisation of the fields. For the n-point function $G^{(n)}$ of a field φ this gives the **Callan–Symanzik equation**

$$\left[M\frac{\partial}{\partial M} + \beta(g)\frac{\partial}{\partial g} + n\,\gamma(g)\right]G^{(n)}(x_1, \ldots, x_n; M, g) = 0, \tag{11.4}$$

where $\beta(g) = M\,\mathrm{d}g/\mathrm{d}M$ says how the coupling runs and $\gamma(g)$ is the anomalous dimension of φ.

Working definition 11.2 (RG fixed point)

A **fixed point** is a value g^* with $\beta(g^*) = 0$: there the coupling stops running. *Operational test:* compute $\beta(g)$ (in perturbation theory, as a power series) and find its zeros. At a fixed point the theory has no intrinsic scale [no length or mass is singled out, so rescaling $x \mapsto \lambda x$ maps the theory to itself].

Let us see that a fixed point produces power laws with shifted exponents. Take the two-point function in a massless theory at $g = g^*$. By rotation invariance and dimensional analysis it has the form $G^{(2)}(x; M) = |x|^{-2d^c}F(M|x|)$, with d^c the canonical dimension and F a dimensionless function of the only dimensionless combination $M|x|$. Then

$$0 \overset{(1)}{=} \left[M\frac{\partial}{\partial M} + 2\gamma^*\right]|x|^{-2d^c}F(M|x|)$$
$$\overset{(2)}{=} |x|^{-2d^c}\left[uF'(u) + 2\gamma^* F(u)\right]_{u=M|x|},$$

(1) The Callan–Symanzik equation (11.4) with $n = 2$; the β term vanishes at the fixed point, and $\gamma^* = \gamma(g^*)$.

(2) Chain rule: $M\,\partial_M F(M|x|) = M|x|\,F'(M|x|)$.

so $uF' = -2\gamma^* F$, whose solution is $F(u) = C\,u^{-2\gamma^*}$. Hence

$$G^{(2)}(x) = C\,M^{-2\gamma^*}\,|x|^{-2(d^c+\gamma^*)}. \tag{11.5}$$

This is a pure power law with scaling dimension $d_\varphi = d^c + \gamma^*$. The anomalous dimension is literally an anomalous shift of the exponent. This is the content of Nash's footnote: "when the scale dimension differs from the canonical dimension the difference is the anomalous dimension γ_A".

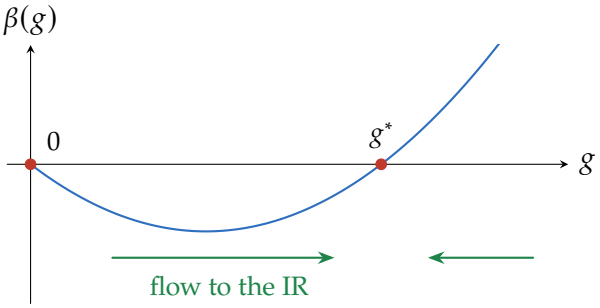


Figure 11.3: A β-function with two zeros. At each zero the coupling stops running and the theory is scale invariant. Since $\beta = M\,\mathrm{d}g/\mathrm{d}M$, lowering M moves g against the sign of β. With $\beta < 0$ between the zeros and $\beta > 0$ beyond g^*, the coupling approaches g^* from both sides at long distances (green arrows): g^* is an infrared-stable fixed point. This is the shape of $\beta(g) = -\varepsilon g + b g^2$ for φ^4 theory in $4-\varepsilon$ dimensions, whose Wilson–Fisher point describes critical magnets. The free point $g = 0$ is the ultraviolet end of the flow.

Watch out: a zero of β is necessary, not sufficient

Nash says a fixed point is a "necessary, but insufficient" condition for conformal invariance. Necessary: without $\beta(g^*) = 0$ the coupling runs and there is a scale (for QCD, Λ_{QCD}). Insufficient: a fixed point gives scale invariance, and scale invariance is weaker than conformal invariance. There are (rare, somewhat pathological) scale-invariant but not conformal theories, for example the theory of elasticity in two dimensions, which is not reflection positive. In practice, unitary, local fixed points are conformal.

11.1.4 Wilson's operator product expansion

Products of quantum fields at the same point are singular. The product $\varphi(x)\varphi(0)$ blows up as $x \to 0$, as the propagators above show. Wilson's idea was to organise the singularity. Seen from far away, the pair of operators at x and 0 looks like a single local disturbance at 0. Any local disturbance can be expanded in a basis of local operators $\mathcal{O}_n(0)$ at that point. The coefficients carry all the x dependence.

Working definition 11.3 (operator product expansion (OPE))

For two local fields A, B there are local operators $\mathcal{O}_n$ [fields sitting at the single point 0, such as $\varphi^2(0)$ or $\partial\varphi(0)$] and c-number functions $C_n(x)$ [ordinary numerical functions of x, not operators] such that, inside any correlation function with other fields placed far from 0,

$$A(x)B(0) \;\longrightarrow\; \sum_n C_n(x)\,\mathcal{O}_n(0) \qquad (x \to 0). \tag{11.6}$$

Operational recipe: for free fields, apply Wick's theorem [the product equals its normal-ordered form plus every way of replacing pairs of fields by their propagator] to $A(x)B(0)$, keep the contractions as c-number coefficients, and Taylor-expand the leftover normal-ordered operators [the uncontracted fields, with annihilation parts to the right, so their vacuum expectation value is zero] about 0. The C_n are singular at $x = 0$; ordering the $\mathcal{O}_n$

by dimension [in a scale-invariant theory $C_n(x)$ scales as $|x|^{d_n - d_A - d_B}$, so a smaller scaling dimension d_n means a stronger singularity] orders the terms from most to least singular.

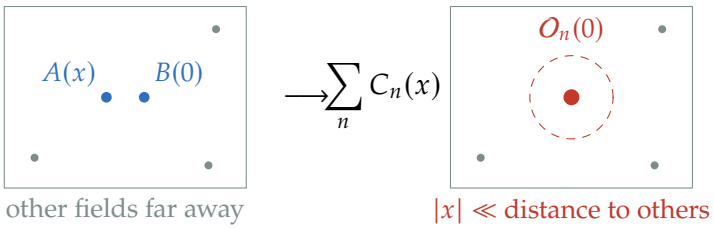


Figure 11.4: The OPE. Two operators close together (left), seen from far away, are indistinguishable from a combination of single local operators at one point (right). The coefficients $C_n(x)$ record how the pair's separation enters. The expansion is valid as long as no other field comes inside the dashed circle.

Worked example 11.5: OPE of a free massless scalar in four dimensions

Let φ be a free massless scalar with $\langle \varphi(x)\varphi(0)\rangle = 1/(4\pi^2 x^2)$. Wick's theorem writes the product as the contraction plus the normal-ordered product,

$$\varphi(x)\varphi(0) \overset{(1)}{=} \langle \varphi(x)\varphi(0)\rangle\, \mathbb{1} + :\varphi(x)\varphi(0):$$
$$\overset{(2)}{=} \frac{1}{4\pi^2 x^2}\,\mathbb{1} + :\varphi^2:(0) + x^\mu :\partial_\mu\varphi\;\varphi:(0) + \tfrac{1}{2}x^\mu x^\nu :\partial_\mu\partial_\nu\varphi\;\varphi:(0) + \cdots$$

(1) Wick's theorem for two free fields: one contraction or none.
(2) The normal-ordered product has no singularity, so we may Taylor-expand $\varphi(x)$ about 0 inside it.

So $\mathcal{O}_0 = \mathbb{1}$ with $C_0 = 1/(4\pi^2 x^2)$, $\mathcal{O}_1 = :\varphi^2:$ with $C_1 = 1$, $\mathcal{O}_{2,\mu} = :\partial_\mu\varphi\;\varphi:$ with $C_2^\mu = x^\mu$, and so on. Let us check the dimension bookkeeping (derived in general just below). With $d_\varphi = 1$: the identity has dimension 0 and $C_0 \sim |x|^{-2}$; $:\varphi^2:$ has dimension 2 and $C_1 \sim |x|^0$; $:\partial\varphi\;\varphi:$ has dimension 3 and $C_2 \sim |x|^{+1}$. In each case (power of $1/|x|$) $= d_\varphi + d_\varphi - d_{\mathcal{O}_n}$.

Now let us derive that bookkeeping rule in a scale-invariant theory. Nash defines the *strength* S_{C_n} of the singularity of C_n by

$$C_n(\lambda x) = \frac{C_n(x)}{\lambda^{S_{C_n}}}, \tag{11.7}$$

so that $C_n(x) \propto |x|^{-S_{C_n}}$. Conjugate the OPE (11.6) by $U(\lambda)$:

$$\lambda^{d_A+d_B}\, A(\lambda x)B(0) \overset{(1)}{=} U(\lambda)\, A(x)B(0)\, U(\lambda)^{-1} \overset{(2)}{=} \sum_n C_n(x)\, U(\lambda)\mathcal{O}_n(0)U(\lambda)^{-1}$$
$$\overset{(3)}{=} \sum_n C_n(x)\, \lambda^{d_{\mathcal{O}_n}}\mathcal{O}_n(0),$$
$$\lambda^{d_A+d_B}\, A(\lambda x)B(0) \overset{(4)}{=} \lambda^{d_A+d_B}\sum_n C_n(\lambda x)\,\mathcal{O}_n(0) \overset{(5)}{=} \sum_n \lambda^{d_A+d_B-S_{C_n}} C_n(x)\,\mathcal{O}_n(0).$$

(1) (11.1) for A and for B (at the origin), with UBU^{-1} and UAU^{-1} multiplied.
(2) Apply the OPE to $A(x)B(0)$; the C_n are numbers and commute with U.
(3) (11.1) for each $\mathcal{O}_n$ at the origin.

(4) Apply the OPE directly to $A(\lambda x)B(0)$.
(5) Definition (11.7).

The $\mathcal{O}_n$ are independent, so the coefficients must agree term by term:

$$d_A + d_B = d_{\mathcal{O}_n} + S_{C_n} \qquad \text{for each } n, \qquad \text{i.e.} \qquad C_n(x) \propto |x|^{d_{\mathcal{O}_n} - d_A - d_B}. \tag{11.8}$$

The most singular term comes from the operator of lowest dimension (usually the identity), and operators of higher dimension come with less singular or vanishing coefficients.

Watch out: a sign in Nash's (11.3)

Nash prints (11.3) as $d_A + d_B = d_{\mathcal{O}_n} - S_{C_n}$. With his own definition (11.2) of S_{C_n} the sign should be +, as derived above. The free-field example confirms it: for $\varphi\varphi \to \mathbb{1}$ in four dimensions, $d_A + d_B = 2$, $d_{\mathbb{1}} = 0$ and $C_0 \propto |x|^{-2}$ has $S = 2$; so $2 = 0 + 2$, whereas the printed sign would give $2 = 0 - 2$.

Nash remarks that the OPE "plays a prominent part" in two-dimensional conformal field theory. That is an understatement. In two dimensions the OPE of the stress tensor with any field *is* the conformal Ward identity (Section 11.3), and the OPE of the stress tensor with itself *is* the Virasoro algebra (Section 11.4).

11.1.5 The conformal group in d dimensions

We have been using the phrase "conformal transformation". Let us define it and find all of them.

Working definition 11.4 (conformal transformation)

A smooth invertible map $x \mapsto x'(x)$ of (a region of) $\mathbb{R}^d$ with the flat metric $\delta_{\mu\nu}$ is **conformal** if it rescales the metric by a position-dependent factor [a short step $\mathrm{d}x$ at x is mapped to a step $\Omega(x)$ times as long, the same factor in every direction],

$$\delta_{\rho\sigma}\,\frac{\partial x'^{\rho}}{\partial x^{\mu}}\,\frac{\partial x'^{\sigma}}{\partial x^{\nu}} = \Omega(x)^2\,\delta_{\mu\nu}. \tag{11.9}$$

Operational test: compute the Jacobian matrix $J = \partial x'/\partial x$ and check that $J^T J$ is a multiple of the identity at every point. Equivalently, $J(x)$ is a rotation (or reflection) times a scalar, so the map preserves angles between curves but not lengths.

Infinitesimally, write $x'^{\mu} = x^{\mu} + \epsilon^{\mu}(x)$ with ϵ small. Then $J = \mathbb{1} + \partial\epsilon$ and, to first order,

$$\delta_{\rho\sigma}\big(\delta^{\rho}_{\mu} + \partial_{\mu}\epsilon^{\rho}\big)\big(\delta^{\sigma}_{\nu} + \partial_{\nu}\epsilon^{\sigma}\big) \overset{(1)}{=} \delta_{\mu\nu} + \partial_{\mu}\epsilon_{\nu} + \partial_{\nu}\epsilon_{\mu} + O(\epsilon^2) \overset{(2)}{=} (1 + 2\sigma(x))\,\delta_{\mu\nu},$$

(1) Multiply out; indices are raised and lowered with δ, so $\epsilon_{\nu} = \epsilon^{\nu}$.
(2) Write $\Omega^2 = 1 + 2\sigma$ with σ small.

Taking the trace of $\partial_{\mu}\epsilon_{\nu} + \partial_{\nu}\epsilon_{\mu} = 2\sigma\delta_{\mu\nu}$ fixes $2\sigma = \frac{2}{d}\,\partial\cdot\epsilon$. So an infinitesimal conformal transformation satisfies the **conformal Killing equation**

$$\partial_{\mu}\epsilon_{\nu} + \partial_{\nu}\epsilon_{\mu} = \frac{2}{d}\,(\partial\cdot\epsilon)\,\delta_{\mu\nu}. \tag{11.10}$$

Solving the conformal Killing equation for $d > 2$. We show that ϵ is at most quadratic in x. Contract (11.10) with ∂^{ν}:

$$\Box\epsilon_{\mu} + \partial_{\mu}(\partial\cdot\epsilon) = \frac{2}{d}\,\partial_{\mu}(\partial\cdot\epsilon) \quad\Longrightarrow\quad \Box\epsilon_{\mu} = \Big(\frac{2}{d} - 1\Big)\partial_{\mu}(\partial\cdot\epsilon). \tag{11.11}$$

Now apply $\Box$ to (11.10) and use (11.11):

$$\frac{2}{d}\,\delta_{\mu\nu}\Box(\partial\cdot\epsilon) \overset{(1)}{=} \partial_\mu\Box\epsilon_\nu + \partial_\nu\Box\epsilon_\mu \overset{(2)}{=} 2\Big(\frac{2}{d}-1\Big)\partial_\mu\partial_\nu(\partial\cdot\epsilon).$$

(1) $\Box$ commutes with ∂_μ.
(2) Substitute (11.11); the two terms are equal because partial derivatives commute.

Taking the trace ($\mu=\nu$, summed with $\delta_{\mu\mu}=d$) of the two ends of this chain gives

$$2\,\Box(\partial\cdot\epsilon) \overset{(1)}{=} 2\Big(\frac{2}{d}-1\Big)\Box(\partial\cdot\epsilon) \quad\Longrightarrow\quad \frac{2(d-1)}{d}\,\Box(\partial\cdot\epsilon) \overset{(2)}{=} 0.$$

(1) Left: $\frac{2}{d}\delta_{\mu\mu}=2$. Right: $\partial_\mu\partial_\mu=\Box$.
(2) Move everything to one side: $2-2(\frac{2}{d}-1)=4-\frac{4}{d}=\frac{4(d-1)}{d}$; divide by 2.

So for $d\geq 2$, $\Box(\partial\cdot\epsilon)=0$. Then the left end of the chain vanishes, and the right end gives

$$(2-d)\,\partial_\mu\partial_\nu(\partial\cdot\epsilon)=0. \tag{11.12}$$

For $d>2$ this says $\partial\cdot\epsilon$ is at most linear in x. Next, the identity

$$\begin{aligned}2\partial_\mu\partial_\nu\epsilon_\rho &\overset{(1)}{=} \partial_\mu(\partial_\nu\epsilon_\rho+\partial_\rho\epsilon_\nu)+\partial_\nu(\partial_\mu\epsilon_\rho+\partial_\rho\epsilon_\mu)-\partial_\rho(\partial_\mu\epsilon_\nu+\partial_\nu\epsilon_\mu)\\ &\overset{(2)}{=} \frac{2}{d}\big(\delta_{\nu\rho}\partial_\mu+\delta_{\mu\rho}\partial_\nu-\delta_{\mu\nu}\partial_\rho\big)(\partial\cdot\epsilon)\end{aligned}$$

(1) Expand the right side: six terms, four cancel in pairs because partial derivatives commute, leaving $2\partial_\mu\partial_\nu\epsilon_\rho$.
(2) Replace each bracket using (11.10).

shows that the second derivatives of ϵ are constants (since $\partial\cdot\epsilon$ is linear). So ϵ is a polynomial of degree at most 2. Substituting the general quadratic into (11.10) and sorting the solutions gives exactly four types:

$$\begin{aligned}&\text{translation:} && \epsilon^\mu = a^\mu,\\ &\text{rotation:} && \epsilon^\mu = \omega^\mu{}_\nu x^\nu, \quad \omega_{\mu\nu}=-\omega_{\nu\mu},\\ &\text{dilation:} && \epsilon^\mu = \lambda\, x^\mu,\\ &\text{special conformal (SCT):} && \epsilon^\mu = 2(b\cdot x)x^\mu - b^\mu x^2.\end{aligned} \tag{11.13}$$

Worked example 11.6: The special conformal transformation satisfies (11.10)

For $\epsilon_\nu = 2(b\cdot x)x_\nu - b_\nu x^2$,

$$\begin{aligned}\partial_\mu\epsilon_\nu &\overset{(1)}{=} 2b_\mu x_\nu + 2(b\cdot x)\delta_{\mu\nu} - 2b_\nu x_\mu,\\ \partial_\mu\epsilon_\nu+\partial_\nu\epsilon_\mu &\overset{(2)}{=} 4(b\cdot x)\,\delta_{\mu\nu},\\ \partial\cdot\epsilon &\overset{(3)}{=} 2(b\cdot x)+2d\,(b\cdot x)-2(b\cdot x)=2d\,(b\cdot x).\end{aligned}$$

(1) Product rule; $\partial_\mu(b\cdot x)=b_\mu$, $\partial_\mu x_\nu=\delta_{\mu\nu}$, $\partial_\mu x^2=2x_\mu$.
(2) Symmetrise: the terms $2b_\mu x_\nu - 2b_\nu x_\mu$ are antisymmetric and cancel.
(3) Set $\mu=\nu$ in the first line and sum: $\delta_{\mu\mu}=d$.

So the right side of (11.10) is $\frac{2}{d}\cdot 2d(b\cdot x)\delta_{\mu\nu}=4(b\cdot x)\delta_{\mu\nu}$, equal to the left side. The local scale factor is $\sigma=\frac{1}{d}\partial\cdot\epsilon=2b\cdot x$: it varies from point to point, which is what distinguishes an SCT from a dilation.

Count the parameters: d translations, $d(d-1)/2$ rotations, one dilation and d special conformal transformations, in total

$$d + \frac{d(d-1)}{2} + 1 + d = \frac{(d+1)(d+2)}{2} = \dim \mathrm{SO}(d+1,1). \tag{11.14}$$

The conformal group of Euclidean $\mathbb{R}^d$ (with a point at infinity added) is indeed $\mathrm{SO}(d+1,1)$; in Minkowski signature it is $\mathrm{SO}(d,2)$. For $d=4$ this is the 15-dimensional group $\mathrm{SO}(4,2)$ of Maxwell's equations and of massless QCD at the classical level.

The finite SCT is best understood through the **inversion** $I : x^\mu \mapsto x^\mu/x^2$. An SCT is "invert, translate by $-b$, invert again":

$$x'^\mu = \frac{x^\mu - b^\mu x^2}{1 - 2\,b\cdot x + b^2 x^2}. \tag{11.15}$$

Expanding to first order in b gives $x'^\mu \approx x^\mu - b^\mu x^2 + 2(b\cdot x)x^\mu$, matching (11.13).

Worked example 11.7: Inversion is conformal

Let $x'^\mu = x^\mu/x^2$. Its Jacobian is

$$\frac{\partial x'^\mu}{\partial x^\nu} \overset{(1)}{=} \frac{\delta^\mu_\nu}{x^2} - \frac{2x^\mu x_\nu}{x^4} = \frac{1}{x^2}\,R^\mu{}_\nu, \qquad R^\mu{}_\nu := \delta^\mu_\nu - 2\frac{x^\mu x_\nu}{x^2}.$$

(1) Quotient rule, with $\partial_\nu (x^2)^{-1} = -2x_\nu/x^4$.

The matrix R is the reflection in the hyperplane orthogonal to x: it sends $x \mapsto -x$ and fixes every vector orthogonal to x. So $R^T R = \mathbb{1}$, and $J^T J = x^{-4}\mathbb{1}$: the inversion is conformal with $\Omega = 1/x^2$. (It reverses orientation, so it is not in the identity component; but composing two inversions with a translation in between gives the orientation-preserving SCT.)

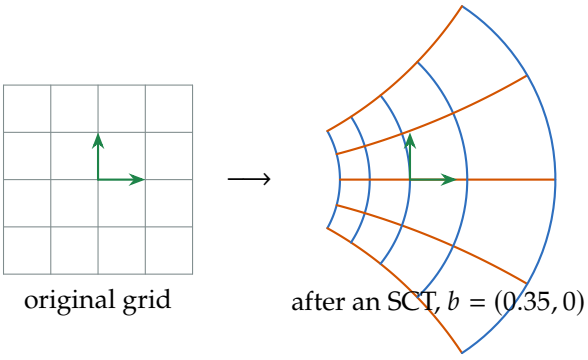


Figure 11.5: A special conformal transformation (11.15) in two dimensions. Straight grid lines become circular arcs, and the cells grow towards the right. But every crossing is still a right angle, and small cells stay square. Locally the map is a rotation times a rescaling; the rescaling factor $\Omega(x)$ changes from point to point.

Why scale invariance plus a traceless stress tensor gives conformal invariance. Here is the field-theory version of the upgrade from scale to conformal invariance. If we change coordinates by $\epsilon^\mu(x)$ depending on position, a translation-invariant action changes by an amount proportional to $\partial\epsilon$. The coefficient defines the (symmetric) stress tensor: $\delta S = -\int \mathrm{d}^d x\, T^{\mu\nu}\partial_\mu \epsilon_\nu$ (the overall sign and normalisation are conventions; see Section 11.3.10). For a conformal ϵ,

$$\delta S \overset{(1)}{=} -\frac{1}{2}\int \mathrm{d}^d x\, T^{\mu\nu}(\partial_\mu\epsilon_\nu + \partial_\nu\epsilon_\mu) \overset{(2)}{=} -\frac{1}{d}\int \mathrm{d}^d x\, T^\mu{}_\mu\,(\partial\cdot\epsilon). \tag{11.16}$$

(1) $T^{\mu\nu}$ is symmetric, so only the symmetric part of $\partial_\mu \epsilon_\nu$ contributes.
(2) The conformal Killing equation (11.10).

So a theory whose stress tensor is traceless, $T^\mu{}_\mu = 0$, is invariant under *all* conformal transformations, not just under dilations (for which $\partial \cdot \epsilon = d\lambda$ is constant). Scale invariance only requires $T^\mu{}_\mu$ to be a total divergence; tracelessness is the extra local condition.

Worked example 11.8: The 2D free boson has a traceless stress tensor

For $S = \frac{g}{2}\int \mathrm{d}^d x\,(\partial\varphi)^2$ the canonical symmetric stress tensor is $T_{\mu\nu} = g\big(\partial_\mu\varphi\,\partial_\nu\varphi - \frac{1}{2}\delta_{\mu\nu}(\partial\varphi)^2\big)$. Its trace is

$$T^\mu{}_\mu \overset{(1)}{=} g\left((\partial\varphi)^2 - \frac{d}{2}(\partial\varphi)^2\right) = g\,\frac{2-d}{2}\,(\partial\varphi)^2.$$

(1) $\delta_{\mu\mu} = d$.

In $d = 2$ this vanishes identically: the free massless boson in two dimensions is conformally invariant with no further work. In $d = 4$ the canonical tensor is not traceless; one must add the "improvement term" $-\frac{g}{6}(\partial_\mu\partial_\nu - \delta_{\mu\nu}\Box)\varphi^2$, which does not change the conserved charges but removes the trace on shell.

11.1.6 Why two dimensions is different

Go back to (11.12). For $d = 2$ the factor $(2-d)$ vanishes and the equation imposes no constraint at all. Write the conformal Killing equation out in components for $d = 2$:

$$\begin{aligned} \mu = \nu = 1: &\quad 2\partial_1\epsilon_1 = \partial_1\epsilon_1 + \partial_2\epsilon_2 \implies \partial_1\epsilon_1 = \partial_2\epsilon_2, \\ \mu = 1, \nu = 2: &\quad \partial_1\epsilon_2 + \partial_2\epsilon_1 = 0. \end{aligned}$$

These are the **Cauchy–Riemann equations** for the complex function $\epsilon(z) = \epsilon_1 + i\epsilon_2$ of $z = x_1 + ix_2$: the real part $u = \epsilon_1$ and imaginary part $v = \epsilon_2$ obey $u_x = v_y$, $u_y = -v_x$. So in two dimensions *every holomorphic function* $\epsilon(z)$ gives an infinitesimal conformal transformation. Holomorphic functions are specified by infinitely many Taylor or Laurent coefficients, so the local conformal algebra is infinite dimensional. Only a six-dimensional part of it, corresponding to $\mathrm{SO}(3,1) \cong \mathrm{SL}(2,\mathbb{C})/\mathbb{Z}_2$ by (11.14) with $d = 2$, consists of transformations that are defined and invertible on the whole plane plus the point at infinity. We will see this explicitly in Section 11.3.3.

Key idea: the leverage of two dimensions

In $d > 2$ conformal symmetry is a finite-dimensional group: it fixes two- and three-point functions but little more. In $d = 2$ it is infinite dimensional: there are infinitely many conserved charges, one for each holomorphic vector field. That is why "in two dimensions the situation is somewhat better" (Nash), and why exactly solvable interacting conformal theories exist there.

Nash lists where two-dimensional conformal theories occur: the worldsheet theory of string theory (Chapter 9), where conformal invariance is a gauge symmetry and c must cancel against the ghost contribution; the critical points of two-dimensional statistical systems; and the quantum field theory approach to knot invariants of Chapter 12. The original reference for everything in this chapter is Belavin, Polyakov and Zamolodchikov (BPZ, 1984).

Reading Nash (pp. 301–302, (11.1)–(11.3))

§1 is two pages of motivation. Nash mentions critical phenomena, the mass independence of $\varphi(x)\varphi(0)$ at short distance, the Callan–Symanzik equation and the need for a zero of $\beta(g)$, and states Wilson's OPE (11.1) with the scaling rule (11.2)–(11.3). His footnote 1 defines the scale dimension via $U(\lambda)A(x)U^{-1}(\lambda) = \lambda^{d_A}A(\lambda x)$ with $U(\lambda) = \exp[\ln\lambda\,(x\,\partial/\partial x)]$, and the anomalous dimension $\gamma_A = d_A - d_A^c$ (he writes $d_A^c - d_A$, a sign convention; with the Callan–Symanzik form (11.5) the natural sign is ours). We have added the derivation of power laws from scale invariance and from the Callan–Symanzik equation, the free-field OPE, the corrected sign in (11.3), and the whole of the conformal group in d dimensions, which Nash does not discuss but which explains his remark that the conformal group is infinite dimensional only in two dimensions.

Recap

- A mass is a length scale; theories without length scales have power-law correlators $\langle AA\rangle \propto |x|^{-2d_A}$.
- Critical points ($\xi = \infty$) and RG fixed points ($\beta(g^*) = 0$) are scale invariant; the scaling dimension is canonical plus anomalous, $d_\varphi = d^c + \gamma^*$.
- The OPE $A(x)B(0) = \sum_n C_n(x)\mathcal{O}_n(0)$ has $C_n \propto |x|^{d_{\mathcal{O}_n} - d_A - d_B}$.
- Conformal maps rescale the metric locally; in $d > 2$ they form $\mathrm{SO}(d+1,1)$ (translations, rotations, dilation, SCTs), in $d = 2$ they are all holomorphic maps.
- A traceless symmetric stress tensor means invariance under all conformal maps. In 2D the free boson qualifies directly.
- 2D Ising: $\Delta_\sigma = 1/8$, $\Delta_\varepsilon = 1$ from $\eta = 1/4$, $\nu = 1$; these will be explained by the Virasoro algebra.

11.2 Weyl rescalings versus conformal transformations

In Section 11.1.5 a conformal transformation was a change of coordinates $x \mapsto x'$ that rescales the flat metric by a factor $\Omega(x)^2$. Many books, and many conversations, use a second operation with almost the same formula: multiply the metric by $\Omega(x)^2$ and do nothing else. That second operation is a *Weyl rescaling*. The two are easy to confuse, because both put a factor Ω^2 in front of $g_{\mu\nu}$. They are not the same. A conformal transformation moves points around inside one fixed geometry. A Weyl rescaling leaves every point where it is and changes the geometry itself: a flat plane can become a sphere, and flat $\mathbb{R}^d$ can become a cylinder.

This section answers four questions. What exactly is the relation between the two operations? How does a primary field transform, and why does the exponent $-\Delta$ appear? How can we read off Δ without moving any point? And what does demanding invariance under *local* Weyl rescalings buy us? The last answer is concrete: it forces a free scalar in curved space to couple to the curvature as $\frac{1}{2}\xi R\varphi^2$ with $\xi = \frac{1}{6}$ in four dimensions, and it explains the "improvement term" of Example 11.8.

Conventions for this section. We work in Euclidean signature on a d-dimensional manifold M with metric $g_{\mu\nu}$, $\sqrt{g} := \sqrt{\det g_{\mu\nu}}$, and ∇_μ the covariant derivative of g (a reminder of Christoffel symbols and curvature is in Chapter T2). For a function f we write $(\nabla f)^2 := g^{\mu\nu}\partial_\mu f\,\partial_\nu f$ and

$\Box f := g^{\mu\nu}\nabla_\mu\nabla_\nu f = \frac{1}{\sqrt{g}}\partial_\mu\big(\sqrt{g}\,g^{\mu\nu}\partial_\nu f\big)$. The curvature convention is

$$R^\rho{}_{\sigma\mu\nu} = \partial_\mu\Gamma^\rho_{\nu\sigma} - \partial_\nu\Gamma^\rho_{\mu\sigma} + \Gamma^\rho_{\mu\lambda}\Gamma^\lambda_{\nu\sigma} - \Gamma^\rho_{\nu\lambda}\Gamma^\lambda_{\mu\sigma}, \qquad R_{\sigma\nu} = R^\mu{}_{\sigma\mu\nu}, \qquad R = g^{\sigma\nu}R_{\sigma\nu}, \tag{11.17}$$

for which a round sphere S^d of radius a has $R = d(d-1)/a^2 > 0$ (we check $d = 2$ below). Rescaling factors are written $\Omega = e^\sigma$, so $\sigma = \ln\Omega$. For small σ this is the σ of Section 11.1.5, where $\Omega^2 = 1 + 2\sigma$.

11.2.1 Two operations: moving the points, or changing the ruler

A conformal transformation moves points. Recall (11.9): a map $x \mapsto x' = f(x)$ of flat $\mathbb{R}^d$ is conformal if its Jacobian matrix $J = \partial x'/\partial x$ obeys $J^T J = \Omega(x)^2\,\mathbb{1}$. Read it as a statement about lengths. A short step $\mathrm{d}x$ at x is carried by the map to the step $\mathrm{d}x' = J\,\mathrm{d}x$ at x', and

$$|\mathrm{d}x'|^2 \overset{(1)}{=} \mathrm{d}x^T J^T J\,\mathrm{d}x \overset{(2)}{=} \Omega(x)^2\;|\mathrm{d}x|^2\,.$$

(1) $|\mathrm{d}x'|^2 = (\mathrm{d}x')^T\mathrm{d}x'$ and $\mathrm{d}x' = J\,\mathrm{d}x$.
(2) The conformal condition $J^T J = \Omega^2\mathbb{1}$.

Every short step near x is stretched by the same factor $\Omega(x)$, whatever its direction. So angles survive, small squares go to small squares, and only their size changes, by an amount that varies from point to point. The points themselves have moved: the point that was at x is now at x'.

A Weyl rescaling changes the ruler. Now keep every point fixed and replace the metric by

$$g_{\mu\nu}(x) \;\longmapsto\; \bar{g}_{\mu\nu}(x) = \Omega(x)^2\,g_{\mu\nu}(x) = e^{2\sigma(x)}g_{\mu\nu}(x). \tag{11.18}$$

Nothing moves. What changes is the reading of the ruler: a coordinate step $\mathrm{d}x$ at x used to have length $\sqrt{g_{\mu\nu}\mathrm{d}x^\mu\mathrm{d}x^\nu}$ and now has length $\Omega(x)$ times that. If Ω is constant, this is a change of units, like measuring in centimetres instead of metres. If Ω varies, it is a different change of units at every point, and in general it produces a different geometry. Figure 11.6 shows the two operations side by side.

When is the second operation secretly the first one? Suppose that after the Weyl rescaling we are allowed to relabel the points, by some smooth invertible map f (a **diffeomorphism**). If some relabelling makes the new rulers read exactly like the old ones, then nobody can tell that anything happened to the geometry, and the whole effect was just a change of coordinates. That is precisely the conformal case.

Working definition 11.5 (Weyl rescaling, and when it is conformal)

A **Weyl rescaling** is the replacement $g_{\mu\nu}(x) \mapsto \Omega(x)^2 g_{\mu\nu}(x)$ with $\Omega > 0$ [every point stays where it is, and a short step $\mathrm{d}x$ at x that used to measure ℓ now measures $\Omega(x)\,\ell$], made together with $\phi(x) \mapsto \Omega(x)^{-\Delta}\phi(x)$ for each field of scaling dimension Δ [a field that carries Δ powers of inverse length is re-expressed in the new local unit of length]. It is **conformal** when $\Omega^2 g$ is **diffeomorphic** to g [some smooth relabelling of the points, $x \mapsto x' = f(x)$, makes the new rulers read exactly like the old ones], that is, when

$$(f^*g)_{\mu\nu}(x) := g_{\rho\sigma}\big(f(x)\big)\,\frac{\partial f^\rho}{\partial x^\mu}\frac{\partial f^\sigma}{\partial x^\nu} = \Omega(x)^2\,g_{\mu\nu}(x) \tag{11.19}$$

[moving the points with f and then measuring with the old ruler gives the same lengths as measuring the unmoved points with the new ruler]. The map f is then a **conformal**

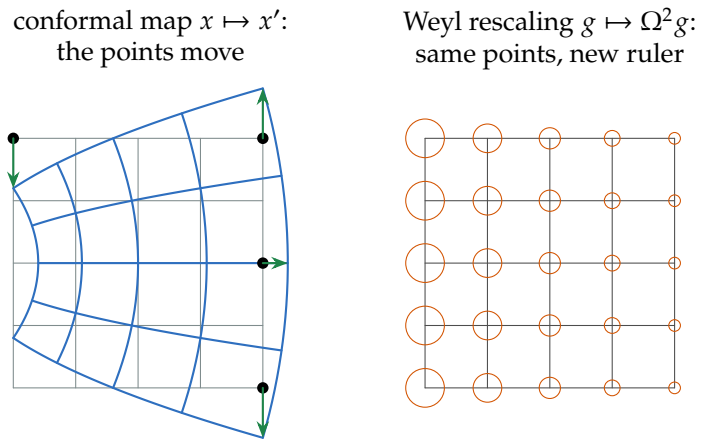


Figure 11.6: Left: the conformal map $z \mapsto z + 0.2\,z^2$ of the plane (grey grid before, blue grid after). The green arrows show four points being carried to new positions. The blue cells are still little squares with right-angled corners, but they are bigger on the right than on the left. Right: a Weyl rescaling with $\Omega = e^{0.6x}$. The grid does not move at all. Each orange circle is the set of points at one fixed *new* distance from the grid point at its centre. On the right, where Ω is large, the new ruler reads every step as long, so the circle of fixed new radius is small.

transformation of (M, g) [for flat $g = \delta$ this is exactly (11.9)]. *To test a given* Ω*:* compute the curvature of $\Omega^2 g$. Curvature cannot be changed by relabelling points, so if $\Omega^2 g$ and g have different curvature the rescaling is not conformal. For flat g the rescaling is conformal (at least locally) exactly when $\Omega^2 g$ is flat as well [its Riemann tensor is zero everywhere]. In $d = 2$ this is the single condition $\Box \ln \Omega = 0$ (see (11.34)).

Farnsworth, Luty and Prilepina put it in one sentence: "Conformal transformations are special Weyl transformations such that the transformed metric is diffeomorphic to the original metric." Most Weyl rescalings are not of this kind. A general $\Omega(x)$ is an arbitrary positive function, while in $d > 2$ the conformal Ω's form the finite-dimensional family found in Section 11.1.5. Figure 11.7 draws this as a picture of the space of all metrics.

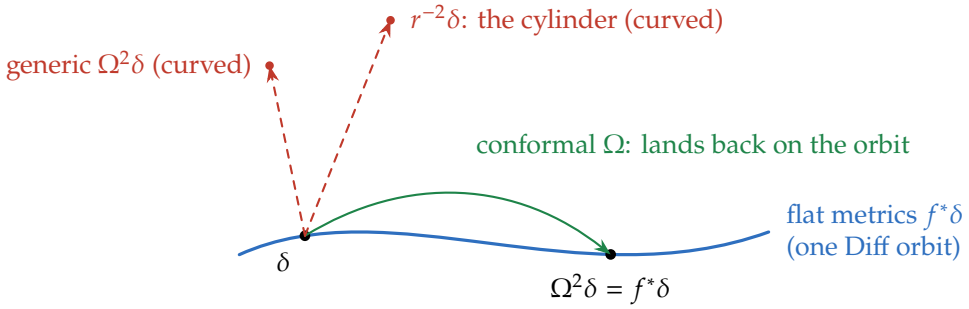


Figure 11.7: The space of all metrics on $\mathbb{R}^d$, drawn as a plane. The blue curve is the set of metrics obtained from the flat metric δ by relabelling points, $f^*\delta$: they all describe the same flat geometry. A Weyl rescaling moves us from δ to $\Omega^2\delta$. For a conformal Ω (green) we land back on the blue curve, so the rescaling can be undone by a diffeomorphism. For every other Ω (red) we land on a genuinely curved metric, for example the cylinder of Example 11.13, and no relabelling brings us back.

Why Weyl invariance implies conformal invariance. Let $S[g, \phi]$ be an action built covariantly from a metric and a scalar field, so that relabelling points does not change it: $S[f^*g, f^*\phi] =$

$S[g,\phi]$, where $(f^*\phi)(x) = \phi(f(x))$. Suppose it is also **Weyl invariant**: $S[\Omega^2 g, \Omega^{-\Delta}\phi] = S[g,\phi]$ for every positive function Ω. Take a conformal map f of flat space, with $f^*\delta = \Omega^2\delta$, and define the transformed field ϕ' by

$$\phi'\big(f(x)\big) = \Omega(x)^{-\Delta}\phi(x). \tag{11.20}$$

Then

$$S[\delta,\phi'] \overset{(1)}{=} S[f^*\delta,\ f^*\phi'] \overset{(2)}{=} S[\Omega^2\delta,\ \Omega^{-\Delta}\phi] \overset{(3)}{=} S[\delta,\phi]. \tag{11.21}$$

(1) Relabelling the points by f does not change a covariant action.
(2) $f^*\delta = \Omega^2\delta$ because f is conformal, and $(f^*\phi')(x) = \phi'(f(x)) = \Omega(x)^{-\Delta}\phi(x)$ by (11.20).
(3) Weyl invariance, used only for this special Ω.

So the flat-space action is unchanged when ϕ is replaced by ϕ': the theory is conformally invariant in flat space. The chain also shows what a conformal transformation *is* on fields: a Weyl rescaling (step 3, read backwards) followed by the relabelling f that undoes it (step 1). Steps 1 and 2 hold for any covariant action. So flat-space conformal invariance is exactly Weyl invariance tested only on the special Ω's of conformal maps, evaluated at $g = \delta$. That is why the converse can fail: a theory may pass the test for these few Ω's and fail it for the others. We return to this in Section 11.2.5.

11.2.2 Primary fields in d dimensions

In two dimensions a primary field was defined in Section 11.3.5 by its response to holomorphic maps. In d dimensions the definition has the same spirit. Under a conformal map the Jacobian is, by the operational test of (11.9),

$$\frac{\partial x'^{\mu}}{\partial x^{\nu}} = \Omega(x)\, R^{\mu}{}_{\nu}(x), \qquad R^T R = \mathbb{1}, \tag{11.22}$$

a rescaling by $\Omega(x)$ times a rotation (or reflection) $R(x)$. A primary field is one that responds to each of these two pieces in the simplest possible way: it is rescaled by a power of Ω and its indices are rotated by R.

Working definition 11.6 (primary tensor of rank (n,m) and dimension Δ)

A field $T^{\mu_1\cdots\mu_n}{}_{\nu_1\cdots\nu_m}$ with n upper and m lower indices is a **primary** of **dimension** Δ if under every conformal map $x \mapsto x'$ [a relabelling of points that stretches every short step near x by the same factor $\Omega(x)$] it transforms as

$$T'^{\mu_1\cdots\mu_n}{}_{\nu_1\cdots\nu_m}(x') = \left|\frac{\partial x'}{\partial x}\right|^{-(\Delta+n-m)/d} \frac{\partial x'^{\mu_1}}{\partial x^{\alpha_1}}\cdots\frac{\partial x'^{\mu_n}}{\partial x^{\alpha_n}}\,\frac{\partial x^{\beta_1}}{\partial x'^{\nu_1}}\cdots\frac{\partial x^{\beta_m}}{\partial x'^{\nu_m}}\, T^{\alpha_1\cdots\alpha_n}{}_{\beta_1\cdots\beta_m}(x) \tag{11.23}$$

[the usual tensor rule for a change of coordinates, times an extra power of the Jacobian determinant]. Here $|\partial x'/\partial x|$ is the absolute value of the Jacobian determinant [the factor by which the map multiplies small volumes]. *To test it:* check that the correlators of T at the new points equal the correlators of the right-hand side, for translations, rotations, the dilation and the inversion.

The power of the determinant looks odd until we evaluate it. The determinant of (11.22) is

$$\left|\frac{\partial x'}{\partial x}\right| \overset{(1)}{=} |\det(\Omega R)| \overset{(2)}{=} \Omega^d\ |\det R| \overset{(3)}{=} \Omega^d.$$

(1) The Jacobian is ΩR by (11.22).

(2) A $d \times d$ matrix multiplied by the number Ω has its determinant multiplied by Ω^d.
(3) $R^T R = \mathbb{1}$ gives $(\det R)^2 = 1$.

So the prefactor in (11.23) is $\Omega^{-(\Delta+n-m)}$. Each upper index brings one factor of the Jacobian, ΩR, and each lower index one factor of the inverse Jacobian, $\Omega^{-1} R^T$ (indices are raised and lowered with δ, and $R^{-1} = R^T$). In total

$$T'(x') \overset{(1)}{=} \Omega^{-(\Delta+n-m)}\, \Omega^n\, \Omega^{-m}\, (R \cdots R)(R^T \cdots R^T)\, T(x)$$
$$\overset{(2)}{=} \Omega(x)^{-\Delta}\, (R \cdots R)(R^T \cdots R^T)\, T(x), \tag{11.24}$$

(1) The determinant gives $\Omega^{-(\Delta+n-m)}$, the n Jacobians give Ω^n and the m inverse Jacobians give Ω^{-m}.
(2) $-(\Delta + n - m) + n - m = -\Delta$.

with one rotation matrix acting on each index. This is the meaning of the definition: *every primary, whatever its indices, is multiplied by* $\Omega^{-\Delta}$ *and rotated.* The index-dependent part of the exponent, $n - m$, is there only to cancel the powers of Ω hidden in the Jacobians.

The scalar case and the sign of the exponent. For a scalar, $n = m = 0$, and (11.24) reads

$$\phi'(x') = \left|\frac{\partial x'}{\partial x}\right|^{-\Delta/d} \phi(x) = \Omega(x)^{-\Delta}\, \phi(x). \tag{11.25}$$

Our convention is therefore: Ω is the factor by which the map *stretches* lengths at x ($\Omega > 1$ where it expands), and a primary of positive dimension is *weakened* by $\Omega^{-\Delta}$ where the map stretches. Let us check that this sign gives the right two-point function. Conformal invariance means that the new field ϕ' has the same correlators as ϕ, so $\langle \phi(x_1')\phi(x_2')\rangle = \langle \phi'(x_1')\phi'(x_2')\rangle = \Omega(x_1)^{-\Delta}\Omega(x_2)^{-\Delta}\langle \phi(x_1)\phi(x_2)\rangle$. Try $\langle \phi(x_1)\phi(x_2)\rangle = C/|x_1 - x_2|^{2\Delta}$.

- *Dilation* $x' = \lambda x$, $\Omega = \lambda$. The left side is $C/(\lambda\,|x_1 - x_2|)^{2\Delta}$, and the right side is $\lambda^{-\Delta}\lambda^{-\Delta} C/|x_1 - x_2|^{2\Delta}$. They agree. This is also (11.2).

- *Inversion* $x' = x/x^2$, $\Omega = 1/x^2$ (Example 11.7). First

$$\left|x_1' - x_2'\right|^2 \overset{(1)}{=} \frac{1}{x_1^2} - \frac{2\,x_1 \cdot x_2}{x_1^2 x_2^2} + \frac{1}{x_2^2} \overset{(2)}{=} \frac{|x_1 - x_2|^2}{x_1^2\, x_2^2}.$$

  (1) Expand $\left|x_1/x_1^2 - x_2/x_2^2\right|^2$ and use $x_i \cdot x_i = x_i^2$.
  (2) Put everything over the common denominator $x_1^2 x_2^2$: the numerator is $x_2^2 - 2x_1 \cdot x_2 + x_1^2 = |x_1 - x_2|^2$.

  So the left side is $C\,(x_1^2 x_2^2)^{\Delta}/|x_1 - x_2|^{2\Delta}$, and the right side is $(x_1^2)^{\Delta}(x_2^2)^{\Delta}\, C/|x_1 - x_2|^{2\Delta}$. They agree.

With the opposite sign, $\Omega^{+\Delta}$, both checks would fail. In $d = 2$ a holomorphic map $z' = f(z)$ has $\Omega = |f'(z)|$, so (11.25) is $\phi'(z') = |f'|^{-\Delta}\,\phi(z)$. This is (11.60) with $h = \bar{h} = \Delta/2$, since $(f')^{-h}(\bar{f}')^{-\bar{h}} = |f'|^{-2h}$. For $h \neq \bar{h}$ the leftover phase $(f'/|f'|)^{-(h-\bar{h})}$ is the rotation R acting on a field of spin $h - \bar{h}$.

The same law seen as a Weyl rescaling. By (11.21), a conformal transformation is a Weyl rescaling followed by the relabelling f. The relabelling is the ordinary tensor rule, which supplies the Jacobian factors in (11.23). Everything else must come from the Weyl step. So under a pure Weyl rescaling, with no point moving, a primary of rank (n, m) transforms as

$$T \longmapsto \Omega^{-w}\, T, \qquad w = \Delta + n - m \quad \text{(the \textbf{Weyl weight})}. \tag{11.26}$$

For a scalar the Weyl weight is the dimension. For a vector potential A_μ in four dimensions, $\Delta = 1$ and $(n, m) = (0, 1)$, so $w = 0$: A_μ does not change at all under a Weyl rescaling, which we will use in Example 11.12.

11.2.3 The scaling dimension from a constant Weyl rescaling

We now find Δ for the free scalar without moving a single point. Couple the massless scalar to a metric in the simplest covariant way,

$$S_{\text{kin}}[g, \varphi] = \frac{1}{2} \int \mathrm{d}^d x\, \sqrt{g}\, g^{\mu\nu}\, \partial_\mu \varphi\, \partial_\nu \varphi, \tag{11.27}$$

which is the action of Section 11.1.1 when $g = \delta$. Rescale the metric by a constant, $g \mapsto \lambda^2 g$, and the field by $\varphi \mapsto \lambda^{-\Delta}\varphi$. The coordinates x, and so $\mathrm{d}^d x$, do not change. The three metric-dependent pieces change as follows:

$$\sqrt{\det(\lambda^2 g)} \overset{(1)}{=} \lambda^d \sqrt{g}, \qquad (\lambda^2 g)^{\mu\nu} \overset{(2)}{=} \lambda^{-2} g^{\mu\nu}, \qquad \partial_\mu(\lambda^{-\Delta}\varphi) \overset{(3)}{=} \lambda^{-\Delta}\partial_\mu \varphi.$$

(1) Multiplying a $d \times d$ matrix by λ^2 multiplies its determinant by λ^{2d}.
(2) The inverse of $\lambda^2 g$ is $\lambda^{-2} g^{-1}$.
(3) λ is constant, so it passes through the derivative.

Therefore

$$S_{\text{kin}}[\lambda^2 g, \lambda^{-\Delta}\varphi] \overset{(1)}{=} \frac{1}{2} \int \mathrm{d}^d x\, \lambda^d \sqrt{g}\, \lambda^{-2} g^{\mu\nu}\, \lambda^{-2\Delta} \partial_\mu \varphi\, \partial_\nu \varphi \overset{(2)}{=} \lambda^{d-2-2\Delta}\, S_{\text{kin}}[g, \varphi]. \tag{11.28}$$

(1) Substitute the three rules above. $\mathrm{d}^d x$ is unchanged because no point moves.
(2) Collect the powers of λ: $d - 2 - 2\Delta$.

Invariance for every λ requires $d - 2 - 2\Delta = 0$:

$$\boxed{\Delta = \frac{d-2}{2}} \tag{11.29}$$

This is the canonical dimension of Section 11.1.1: 1 in $d = 4$, $\frac{1}{2}$ in $d = 3$, 0 in $d = 2$.

The same number comes out of the dilation $x \mapsto \lambda x$ with $\varphi'(x') = \lambda^{-\Delta}\varphi(x)$, in flat space. There the metric stays δ and instead $\mathrm{d}^d x' = \lambda^d \mathrm{d}^d x$ and $\partial'_\mu = \lambda^{-1}\partial_\mu$, and the action again scales as $\lambda^{d-2-2\Delta}$. The two calculations agree for a reason: pulling the flat metric back by the dilation $f(x) = \lambda x$ gives $f^*\delta = \lambda^2 \delta$, so a *constant* Weyl rescaling of flat space is always conformal, undone by a dilation. A constant λ cannot detect the difference between Weyl and conformal. The difference appears only when Ω varies from point to point.

Worked example 11.9: The field equation alone also fixes Δ

A symmetry must map solutions of the field equation $\partial^2\Phi = 0$ (flat space) to solutions. The infinitesimal dilation and special conformal transformation of a primary scalar are $\delta_D\Phi = (x\cdot\partial + \Delta)\Phi$ and $\delta_{K_\mu}\Phi = (K_\mu + 2\Delta x_\mu)\Phi$ with $K_\mu = 2x_\mu\, x\cdot\partial - x^2\partial_\mu$. (These are $\delta\Phi(x) := \Phi(x) - \Phi'(x)$, computed from (11.25) for $x' = x + \epsilon$ with $\Omega = 1 + \sigma$, where $\sigma = \lambda$ for the dilation and $\sigma = 2b\cdot x$ for the SCT, see Example 11.6.) We need three commutators, each checked by acting on a test function u:

$$
\begin{aligned}
[\partial^2, x_\mu]u &\overset{(1)}{=} 2\partial_\mu u,\\
[\partial^2, x\cdot\partial]u &\overset{(2)}{=} 2\partial^2 u,\\
[\partial^2, K_\mu]u &\overset{(3)}{=} 4x_\mu\partial^2 u + 2(2-d)\,\partial_\mu u.
\end{aligned}
$$

(1) $\partial^2(x_\mu u) = x_\mu\partial^2 u + 2\partial_\mu u$, since $\partial_\nu x_\mu = \delta_{\mu\nu}$ and $\partial^2 x_\mu = 0$.
(2) $\partial^2(x^\nu\partial_\nu u) = x^\nu\partial_\nu\partial^2 u + 2\partial^\nu\partial_\nu u$, by the same rule applied to each x^ν.
(3) Write $K_\mu = 2x_\mu x^\nu\partial_\nu - x^2\partial_\mu$. Using (1) twice, $[\partial^2, 2x_\mu x^\nu\partial_\nu]u = 4\partial_\mu u + 4x^\nu\partial_\nu\partial_\mu u + 4x_\mu\partial^2 u$. Also $[\partial^2, x^2]v = 2d\,v + 4x\cdot\partial v$, so $[\partial^2, x^2\partial_\mu]u = 2d\,\partial_\mu u + 4x^\nu\partial_\nu\partial_\mu u$. Subtract: the $x\cdot\partial\,\partial_\mu u$ terms cancel.

Now apply ∂^2 to the transformed field of a solution ($\partial^2\Phi = 0$):

$$
\begin{aligned}
\partial^2(\delta_D\Phi) &\overset{(1)}{=} (x\cdot\partial + 2 + \Delta)\,\partial^2\Phi = 0,\\
\partial^2(\delta_{K_\mu}\Phi) &\overset{(2)}{=} (K_\mu + 4x_\mu + 2\Delta x_\mu)\,\partial^2\Phi + \big[2(2-d) + 4\Delta\big]\partial_\mu\Phi \overset{(3)}{=} \big[2(2-d) + 4\Delta\big]\partial_\mu\Phi.
\end{aligned}
$$

(1) Move ∂^2 to the right with commutator (2).
(2) Move ∂^2 to the right with commutators (3) and (1).
(3) $\partial^2\Phi = 0$.

The dilation maps solutions to solutions for every Δ. The special conformal transformation does so only if $2(2-d) + 4\Delta = 0$, that is $\Delta = (d-2)/2$. So the field equation knows the dimension through the SCT, just as the action knows it through the (constant) rescaling.

11.2.4 Local Weyl invariance fixes the curvature coupling

With $\Delta = (d-2)/2$ the action (11.27) is invariant under *constant* rescalings. Is it invariant when $\sigma(x) = \ln\Omega(x)$ varies? The derivative of $e^{-\Delta\sigma}\varphi$ now hits σ too, and the answer is no. The cure is to add the only other term with two derivatives and two powers of φ that a metric allows,

$$
S[g,\varphi] = \frac{1}{2}\int \mathrm{d}^d x\,\sqrt{g}\,\Big(g^{\mu\nu}\partial_\mu\varphi\,\partial_\nu\varphi + \xi\,R\,\varphi^2\Big), \tag{11.30}
$$

where R is the scalar curvature (two derivatives of the metric) and ξ is a number to be found. In flat space $R = 0$ and the new term is invisible. To use it we first need to know how R responds to a Weyl rescaling.

Step 1: the Laplacian of a Weyl-rescaled metric. For $\bar g = e^{2\sigma}g$ and any function u,

$$
\begin{aligned}
\bar\Box u &\overset{(1)}{=} \frac{1}{\sqrt{\bar g}}\partial_\mu\big(\sqrt{\bar g}\,\bar g^{\mu\nu}\partial_\nu u\big) \overset{(2)}{=} \frac{e^{-d\sigma}}{\sqrt{g}}\,\partial_\mu\big(e^{(d-2)\sigma}\sqrt{g}\,g^{\mu\nu}\partial_\nu u\big)\\
&\overset{(3)}{=} e^{-2\sigma}\Big[\Box u + (d-2)\,g^{\mu\nu}\partial_\mu\sigma\,\partial_\nu u\Big].
\end{aligned} \tag{11.31}
$$

(1) The Laplacian on functions, written with $\sqrt{g}$ (conventions above).
(2) $\sqrt{\bar{g}} = e^{d\sigma}\sqrt{g}$ (the determinant of $e^{2\sigma}g$ is $e^{2d\sigma}\det g$) and $\bar{g}^{\mu\nu} = e^{-2\sigma}g^{\mu\nu}$.
(3) Product rule on $e^{(d-2)\sigma}$, whose derivative is $(d-2)e^{(d-2)\sigma}\partial_\mu\sigma$.

Step 2: the curvature under an infinitesimal rescaling. Let $\delta g_{\mu\nu} = 2\sigma g_{\mu\nu}$ with σ small, so $\delta g^{\mu\nu} = -2\sigma g^{\mu\nu}$. The change of the Christoffel symbols is a tensor (a difference of two connections), and we can compute it at a point in coordinates where $\Gamma = 0$, where $\nabla = \partial$:

$$\delta\Gamma^\lambda_{\mu\nu} \overset{(1)}{=} \tfrac{1}{2}g^{\lambda\kappa}\big(\nabla_\mu\delta g_{\nu\kappa} + \nabla_\nu\delta g_{\mu\kappa} - \nabla_\kappa\delta g_{\mu\nu}\big) \overset{(2)}{=} \delta^\lambda_\nu\partial_\mu\sigma + \delta^\lambda_\mu\partial_\nu\sigma - g_{\mu\nu}\nabla^\lambda\sigma.$$

(1) Vary $\Gamma^\lambda_{\mu\nu} = \frac{1}{2}g^{\lambda\kappa}(\partial_\mu g_{\nu\kappa} + \partial_\nu g_{\mu\kappa} - \partial_\kappa g_{\mu\nu})$ at a point where $\Gamma = 0$. The term with $\delta g^{\lambda\kappa}$ multiplies Γ and drops out, and there $\partial = \nabla$.
(2) $\nabla_\mu\delta g_{\nu\kappa} = 2(\partial_\mu\sigma)g_{\nu\kappa}$ because $\nabla g = 0$. Then contract with $\frac{1}{2}g^{\lambda\kappa}$.

Two contractions of this are needed:

$$g^{\mu\nu}\delta\Gamma^\lambda_{\mu\nu} \overset{(1)}{=} \nabla^\lambda\sigma + \nabla^\lambda\sigma - d\,\nabla^\lambda\sigma = (2-d)\nabla^\lambda\sigma, \qquad \delta\Gamma^\lambda_{\lambda\nu} \overset{(2)}{=} \partial_\nu\sigma + d\,\partial_\nu\sigma - \partial_\nu\sigma = d\,\partial_\nu\sigma.$$

(1) $g^{\mu\nu}g_{\mu\nu} = d$.
(2) $\delta^\lambda_\lambda = d$.

In coordinates with $\Gamma = 0$ at the point, (11.17) gives $R_{\mu\nu} = \partial_\lambda\Gamma^\lambda_{\mu\nu} - \partial_\nu\Gamma^\lambda_{\lambda\mu} + (\Gamma\Gamma)$, and the $\Gamma\Gamma$ terms have zero variation there. So $\delta R_{\mu\nu} = \nabla_\lambda\delta\Gamma^\lambda_{\mu\nu} - \nabla_\nu\delta\Gamma^\lambda_{\lambda\mu}$ (the **Palatini identity**). It is an equation between tensors, so it holds in every coordinate system. Now

$$\begin{aligned}\delta R &\overset{(1)}{=} \delta g^{\mu\nu}\,R_{\mu\nu} + g^{\mu\nu}\,\delta R_{\mu\nu} \overset{(2)}{=} -2\sigma R + \nabla_\lambda\big(g^{\mu\nu}\delta\Gamma^\lambda_{\mu\nu}\big) - \nabla^\mu\big(\delta\Gamma^\lambda_{\lambda\mu}\big)\\ &\overset{(3)}{=} -2\sigma R + (2-d)\Box\sigma - d\,\Box\sigma = -2\sigma R - 2(d-1)\,\Box\sigma. \end{aligned} \tag{11.32}$$

(1) $R = g^{\mu\nu}R_{\mu\nu}$ and the product rule.
(2) $\delta g^{\mu\nu} = -2\sigma g^{\mu\nu}$, and the Palatini identity, with $g^{\mu\nu}$ moved through ∇ (allowed since $\nabla g = 0$).
(3) The two contractions above.

Step 3: the finite rescaling. Build the finite rescaling $\bar{g} = e^{2\sigma}g$ out of many small ones along the path $g_t = e^{2t\sigma}g$, $0 \le t \le 1$. Going from t to $t + \mathrm{d}t$ is the small rescaling $g_{t+\mathrm{d}t} = e^{2\sigma\,\mathrm{d}t}g_t$ of the metric g_t. So (11.32), applied at the metric g_t with small parameter $\sigma\,\mathrm{d}t$, gives the rate of change of $R_t := R[g_t]$. Write $R_t = e^{-2t\sigma}u_t$. Then

$$\begin{aligned}&\frac{\mathrm{d}R_t}{\mathrm{d}t} \overset{(1)}{=} -2\sigma R_t - 2(d-1)\,\Box_t\sigma \overset{(2)}{=} -2\sigma R_t - 2(d-1)\,e^{-2t\sigma}\big[\Box\sigma + (d-2)\,t\,(\nabla\sigma)^2\big],\\ &\frac{\mathrm{d}R_t}{\mathrm{d}t} \overset{(3)}{=} -2\sigma R_t + e^{-2t\sigma}\frac{\mathrm{d}u_t}{\mathrm{d}t} \quad\Longrightarrow\quad \frac{\mathrm{d}u_t}{\mathrm{d}t} \overset{(4)}{=} -2(d-1)\big[\Box\sigma + (d-2)\,t\,(\nabla\sigma)^2\big].\end{aligned}$$

(1) (11.32) at the metric g_t, with $\Box_t$ its Laplacian.
(2) (11.31) with $\sigma \to t\sigma$ and $u = \sigma$. Here $\Box$ and $(\nabla\sigma)^2$ belong to the original g.
(3) Product rule on $R_t = e^{-2t\sigma}u_t$ (σ does not depend on t).
(4) Compare the two lines: the terms $-2\sigma R_t$ cancel, then divide by $e^{-2t\sigma}$.

The right side of the last equation is a polynomial in t, so we integrate from $u_0 = R$ to $t = 1$ and get $u_1 = R - 2(d-1)\Box\sigma - (d-1)(d-2)(\nabla\sigma)^2$. Hence

$$\boxed{\bar{R} = e^{-2\sigma}\Big[R - 2(d-1)\,\Box\sigma - (d-1)(d-2)\,(\nabla\sigma)^2\Big] \quad \text{for } \bar{g} = e^{2\sigma}g} \tag{11.33}$$

This is the standard formula (for example Wald, *General Relativity*, Appendix D). It is exact, not just first order in σ.

Worked example 11.10: Check in $d = 2$: a flat plane becomes a round sphere

In $d = 2$ the formula (11.33) loses its last term:

$$\bar{R} = e^{-2\sigma}\big(R - 2\,\Box\sigma\big) \qquad (d = 2). \tag{11.34}$$

Take flat $g = \delta$ ($R = 0$) and $e^{\sigma} = 2/(1 + r^2)$, with $r^2 = x_1^2 + x_2^2$. The metric $\bar{g} = 4(\mathrm{d}x_1^2 + \mathrm{d}x_2^2)/(1 + r^2)^2$ is the unit sphere written in stereographic coordinates, so we expect $\bar{R} = 2$. With $\sigma = \ln 2 - \ln(1 + r^2)$ and the flat Laplacian in polar coordinates, $\Box u = \frac{1}{r}\frac{\mathrm{d}}{\mathrm{d}r}(r\frac{\mathrm{d}u}{\mathrm{d}r})$ for $u(r)$:

$$\Box \ln(1 + r^2) \overset{(1)}{=} \frac{1}{r}\frac{\mathrm{d}}{\mathrm{d}r}\Big(\frac{2r^2}{1 + r^2}\Big) \overset{(2)}{=} \frac{1}{r}\cdot\frac{4r(1 + r^2) - 2r^2\cdot 2r}{(1 + r^2)^2} = \frac{4}{(1 + r^2)^2},$$
$$\bar{R} \overset{(3)}{=} -2e^{-2\sigma}\Box\sigma \overset{(4)}{=} 2\cdot\frac{(1 + r^2)^2}{4}\cdot\frac{4}{(1 + r^2)^2} = 2.$$

(1) $\frac{\mathrm{d}}{\mathrm{d}r}\ln(1 + r^2) = \frac{2r}{1+r^2}$, then multiply by r.
(2) Quotient rule. The numerator simplifies to $4r$.
(3) (11.34) with $R = 0$.
(4) $\Box\sigma = -\Box\ln(1 + r^2) = -4/(1 + r^2)^2$ and $e^{-2\sigma} = (1 + r^2)^2/4$.

The unit sphere has $\bar{R} = 2 = d(d - 1)/a^2$ with $d = 2$, $a = 1$, which confirms both the formula and the sign convention. The integral also checks Gauss–Bonnet (Chapter T2): $\int\sqrt{\bar{g}}\,\bar{R} = 2\cdot 4\pi = 4\pi\chi(S^2)$. And (11.34) gives the test promised in Definition 11.5: in $d = 2$, $e^{2\sigma}\delta$ is flat exactly when $\Box\sigma = 0$. For a holomorphic map f, $\sigma = \ln|f'| = \mathrm{Re}\ln f'$ is the real part of a holomorphic function, hence harmonic, as it must be.

Route 1: fix Δ first, then find ξ. Take $\Delta = (d - 2)/2$ from (11.29), and apply the finite rescaling $\bar{g} = e^{2\sigma}g$, $\bar{\varphi} = e^{-\Delta\sigma}\varphi$ to (11.30). The derivative of the rescaled field is

$$\partial_\mu\bar{\varphi} \overset{(1)}{=} e^{-\Delta\sigma}\partial_\mu\varphi - \Delta\, e^{-\Delta\sigma}\varphi\,\partial_\mu\sigma = e^{-\Delta\sigma}\big(\partial_\mu\varphi - \Delta\,\varphi\,\partial_\mu\sigma\big).$$

(1) Product rule. The factor $e^{-\Delta\sigma}$ now depends on x.

The kinetic density becomes

$$\sqrt{\bar{g}}\,\bar{g}^{\mu\nu}\partial_\mu\bar{\varphi}\,\partial_\nu\bar{\varphi} \overset{(1)}{=} e^{d\sigma}\sqrt{g}\;e^{-2\sigma}g^{\mu\nu}\;e^{-2\Delta\sigma}(\partial_\mu\varphi - \Delta\varphi\partial_\mu\sigma)(\partial_\nu\varphi - \Delta\varphi\partial_\nu\sigma)$$
$$\overset{(2)}{=} \sqrt{g}\,\Big[(\nabla\varphi)^2 - 2\Delta\,\varphi\,g^{\mu\nu}\partial_\mu\varphi\,\partial_\nu\sigma + \Delta^2\varphi^2(\nabla\sigma)^2\Big]. \tag{11.35}$$

(1) $\sqrt{\bar{g}} = e^{d\sigma}\sqrt{g}$, $\bar{g}^{\mu\nu} = e^{-2\sigma}g^{\mu\nu}$, and the derivative just computed.
(2) The exponents add to $(d - 2 - 2\Delta)\sigma = 0$. Then multiply out the bracket, using the symmetry of $g^{\mu\nu}$ for the two cross terms.

The middle term is a total derivative in disguise. Since $2\varphi\,\partial_\mu\varphi = \partial_\mu(\varphi^2)$,

$$\int \mathrm{d}^d x\,\sqrt{g}\,\big(-2\Delta\varphi\,g^{\mu\nu}\partial_\mu\varphi\,\partial_\nu\sigma\big) \overset{(1)}{=} -\Delta\int \mathrm{d}^d x\,\sqrt{g}\,g^{\mu\nu}\partial_\mu(\varphi^2)\,\partial_\nu\sigma \overset{(2)}{=} +\Delta\int \mathrm{d}^d x\,\sqrt{g}\,\varphi^2\,\Box\sigma. \tag{11.36}$$

(1) $2\varphi\partial_\mu\varphi = \partial_\mu(\varphi^2)$.

(2) Integrate by parts: $\partial_\mu(\varphi^2)\sqrt{g}g^{\mu\nu}\partial_\nu\sigma = \partial_\mu\big(\varphi^2\sqrt{g}g^{\mu\nu}\partial_\nu\sigma\big) - \varphi^2\,\partial_\mu(\sqrt{g}g^{\mu\nu}\partial_\nu\sigma)$. The first term integrates to zero (M closed, or fields decaying at infinity), and the second is $-\varphi^2\sqrt{g}\,\Box\sigma$.

So the kinetic term changes by

$$S_{\text{kin}}[\bar g,\bar\varphi] - S_{\text{kin}}[g,\varphi] = \frac{1}{2}\int \mathrm{d}^d x\,\sqrt{g}\,\Big[\Delta\,\varphi^2\,\Box\sigma + \Delta^2\varphi^2(\nabla\sigma)^2\Big]. \tag{11.37}$$

Now the curvature term, $S_\xi = \frac{\xi}{2}\int\sqrt{g}\,R\varphi^2$:

$$\begin{aligned} S_\xi[\bar g,\bar\varphi] &\overset{(1)}{=} \frac{\xi}{2}\int \mathrm{d}^d x\,e^{d\sigma}\sqrt{g}\;e^{-2\sigma}\big[R - 2(d-1)\Box\sigma - (d-1)(d-2)(\nabla\sigma)^2\big]\;e^{-2\Delta\sigma}\varphi^2 \\ &\overset{(2)}{=} S_\xi[g,\varphi] + \frac{1}{2}\int \mathrm{d}^d x\,\sqrt{g}\,\Big[-2\xi(d-1)\,\varphi^2\,\Box\sigma - \xi(d-1)(d-2)\,\varphi^2(\nabla\sigma)^2\Big]. \end{aligned} \tag{11.38}$$

(1) $\sqrt{\bar g} = e^{d\sigma}\sqrt{g}$, the formula (11.33) for $\bar R$, and $\bar\varphi^2 = e^{-2\Delta\sigma}\varphi^2$.
(2) The exponents again add to $(d - 2 - 2\Delta)\sigma = 0$. The R term reproduces $S_\xi[g,\varphi]$.

Adding (11.37) and (11.38), the total change of (11.30) is

$$S[\bar g,\bar\varphi] - S[g,\varphi] = \frac{1}{2}\int \mathrm{d}^d x\,\sqrt{g}\,\varphi^2\Big\{\big[\Delta - 2\xi(d-1)\big]\Box\sigma + \big[\Delta^2 - \xi(d-1)(d-2)\big](\nabla\sigma)^2\Big\}. \tag{11.39}$$

The two brackets must vanish separately. (Replace σ by $\epsilon\sigma$: the $\Box\sigma$ term is proportional to ϵ and the $(\nabla\sigma)^2$ term to ϵ^2, and the sum must vanish for every ϵ.) The $\Box\sigma$ bracket fixes ξ:

$$\boxed{\xi = \frac{\Delta}{2(d-1)} = \frac{d-2}{4(d-1)}} \tag{11.40}$$

Nothing is left to choose. Yet the $(\nabla\sigma)^2$ bracket must still vanish, and it does:

$$\Delta^2 - \xi(d-1)(d-2) \overset{(1)}{=} \Delta^2 - \frac{\Delta}{2(d-1)}(d-1)(d-2) \overset{(2)}{=} \Delta\Big(\Delta - \frac{d-2}{2}\Big) \overset{(3)}{=} 0. \tag{11.41}$$

(1) Insert $\xi = \Delta/(2(d-1))$ from (11.40).
(2) Cancel $(d-1)$ and take out the common factor Δ.
(3) $\Delta = (d-2)/2$ by (11.29).

So the action (11.30) with $\xi = (d-2)/(4(d-1))$ is exactly invariant under every finite local Weyl rescaling.

Why did the second condition come for free? Because Weyl rescalings compose: doing $e^{2\sigma_1}$ and then $e^{2\sigma_2}$ is doing $e^{2(\sigma_1+\sigma_2)}$. The $\Box\sigma$ bracket is all that survives to first order in σ, and first-order invariance at *every* metric already implies finite invariance. Indeed $F(t) := S[e^{2t\sigma}g, e^{-t\Delta\sigma}\varphi]$ has $\mathrm{d}F/\mathrm{d}t$ equal to a first-order variation of S about the configuration $(e^{2t\sigma}g, e^{-t\Delta\sigma}\varphi)$, which is zero. So $F(1) = F(0)$, and the $(\nabla\sigma)^2$ terms had to cancel.

Route 2: fix Δ and ξ at the same time. We can also leave Δ unknown and demand invariance under an infinitesimal rescaling $\delta g_{\mu\nu} = 2\sigma g_{\mu\nu}$, $\delta\varphi = -\Delta\sigma\varphi$ with an arbitrary small $\sigma(x)$. Then $\delta\sqrt{g} = d\,\sigma\sqrt{g}$ and $\delta g^{\mu\nu} = -2\sigma g^{\mu\nu}$. For the kinetic term,

$$\begin{aligned} \delta S_{\text{kin}} &\overset{(1)}{=} \frac{1}{2}\int \mathrm{d}^d x\,\sqrt{g}\,\Big[(d-2)\,\sigma\,(\nabla\varphi)^2 - 2\Delta\,g^{\mu\nu}\partial_\mu\varphi\,\partial_\nu(\sigma\varphi)\Big] \\ &\overset{(2)}{=} \int \mathrm{d}^d x\,\sqrt{g}\,\sigma\Big[\frac{d-2}{2}(\nabla\varphi)^2 + \Delta\,\varphi\,\Box\varphi\Big]. \end{aligned} \tag{11.42}$$

(1) The variations of $\sqrt{g}$ and $g^{\mu\nu}$ give $(d-2)\sigma$. The variation of $\partial_\mu\varphi\,\partial_\nu\varphi$ gives twice $\partial_\mu\varphi\,\partial_\nu(\delta\varphi)$, with $\delta\varphi = -\Delta\sigma\varphi$.
(2) Integrate by parts in the second term, moving ∂_ν off $\sigma\varphi$: $-\int\sqrt{g}\,g^{\mu\nu}\partial_\mu\varphi\,\partial_\nu(\sigma\varphi) = +\int\sqrt{g}\,\sigma\varphi\,\Box\varphi$.

For the curvature term we use (11.32):

$$\begin{aligned}\delta S_\xi &\overset{(1)}{=} \frac{\xi}{2}\int \mathrm{d}^d x\,\sqrt{g}\,\Big[(d-2\Delta)\,\sigma R\varphi^2 + \varphi^2\big(-2\sigma R - 2(d-1)\Box\sigma\big)\Big]\\ &\overset{(2)}{=} \frac{\xi}{2}\int \mathrm{d}^d x\,\sqrt{g}\,\sigma\Big[(d-2-2\Delta)R\varphi^2 - 2(d-1)\,\Box(\varphi^2)\Big]\\ &\overset{(3)}{=} \frac{\xi}{2}\int \mathrm{d}^d x\,\sqrt{g}\,\sigma\Big[(d-2-2\Delta)R\varphi^2 - 4(d-1)(\nabla\varphi)^2 - 4(d-1)\,\varphi\,\Box\varphi\Big].\end{aligned} \tag{11.43}$$

(1) $\delta\sqrt{g} = d\sigma\sqrt{g}$, $\delta(\varphi^2) = -2\Delta\sigma\varphi^2$, and δR from (11.32).
(2) Collect the $\sigma R\varphi^2$ terms. Integrate the $\Box\sigma$ term by parts twice, $\int\sqrt{g}\,\varphi^2\Box\sigma = \int\sqrt{g}\,\sigma\,\Box(\varphi^2)$.
(3) $\Box(\varphi^2) = 2(\nabla\varphi)^2 + 2\varphi\Box\varphi$.

Adding the two,

$$\delta S = \int \mathrm{d}^d x\,\sqrt{g}\,\sigma\Big[\underbrace{\Big(\tfrac{d-2}{2} - 2\xi(d-1)\Big)}_{A}(\nabla\varphi)^2 + \underbrace{\big(\Delta - 2\xi(d-1)\big)}_{B}\,\varphi\,\Box\varphi + \underbrace{\tfrac{\xi}{2}\big(d-2-2\Delta\big)}_{C}\,R\varphi^2\Big]. \tag{11.44}$$

Since $\sigma(x)$ is arbitrary, the bracket must vanish at every point, for every φ and every metric. The three structures can be varied independently at a point, so $A = B = C = 0$. Then $A = 0$ gives $\xi = (d-2)/(4(d-1))$, $B = 0$ gives $\Delta = 2\xi(d-1) = (d-2)/2$, and $C = 0$ is then automatic. Both numbers come out of one requirement.

The two routes are the same calculation sorted in two ways. Route 1 keeps σ undifferentiated only in the overall factor $e^{(d-2-2\Delta)\sigma}$ and sorts the rest by derivatives of σ. Route 2 moves every derivative off σ and sorts by what multiplies σ. To first order in σ, with Δ left free, Route 1 gives

$$\begin{aligned}\delta S &\overset{(1)}{=} \frac{1}{2}\int\sqrt{g}\,\Big\{(d-2-2\Delta)\,\sigma\big[(\nabla\varphi)^2 + \xi R\varphi^2\big] + \big[\Delta - 2\xi(d-1)\big]\varphi^2\,\Box\sigma\Big\}\\ &\overset{(2)}{=} \int\sqrt{g}\,\sigma\Big\{\tfrac{d-2-2\Delta}{2}(\nabla\varphi)^2 + C\,R\varphi^2 + B\big[(\nabla\varphi)^2 + \varphi\,\Box\varphi\big]\Big\}\\ &\overset{(3)}{=} \int\sqrt{g}\,\sigma\big[A(\nabla\varphi)^2 + B\,\varphi\Box\varphi + C\,R\varphi^2\big].\end{aligned}$$

(1) Expand $e^{(d-2-2\Delta)\sigma} \approx 1 + (d-2-2\Delta)\sigma$ in the first lines of (11.35) and (11.38) (before the exponent was set to zero), use (11.36), and drop the $(\nabla\sigma)^2$ terms, which are second order.
(2) Move $\Box$ from σ to φ^2 and use $\frac{1}{2}\Box(\varphi^2) = (\nabla\varphi)^2 + \varphi\Box\varphi$.
(3) $\frac{d-2-2\Delta}{2} + \Delta - 2\xi(d-1) = \frac{d-2}{2} - 2\xi(d-1) = A$.

This is exactly (11.44).

The conformally coupled scalar. In $d = 4$ the result is

$$\Delta = 1, \qquad \xi = \frac{4-2}{4\cdot 3} = \frac{1}{6}. \tag{11.45}$$

A scalar with the action (11.30) and $\xi = (d-2)/(4(d-1))$ is called a **conformally coupled scalar**: it is coupled to the curvature in exactly the way that makes it invariant under local

Weyl rescalings with $\varphi \mapsto \Omega^{-(d-2)/2}\varphi$. (With $\xi = 0$ it is called **minimally coupled**.) Its field equation is $E \coloneqq \frac{1}{\sqrt{g}}\frac{\delta S}{\delta \varphi} = -\Box\varphi + \xi R\varphi = 0$ (vary (11.30) and integrate by parts once). The operator $-\Box + \frac{d-2}{4(d-1)}R$ is the **conformal Laplacian**, and Weyl invariance of the action makes it transform neatly. Vary φ by η on both sides of $S[\bar{g},\bar{\varphi}] = S[g,\varphi]$ (with $\bar{\eta} = \Omega^{-\Delta}\eta$):

$$\int \mathrm{d}^d x\,\sqrt{g}\,E\,\eta \overset{(1)}{=} \int \mathrm{d}^d x\,\sqrt{\bar{g}}\,\bar{E}\,\bar{\eta} \overset{(2)}{=} \int \mathrm{d}^d x\,\sqrt{g}\,\Omega^{d-\Delta}\,\bar{E}\,\eta \quad\Longrightarrow\quad \bar{E} \overset{(3)}{=} \Omega^{-(d+2)/2}E.$$

(1) The first-order change of each side, with $\bar{E}$ the field-equation expression built from $\bar{g}$ and $\bar{\varphi}$.
(2) $\sqrt{\bar{g}} = \Omega^d\sqrt{g}$ and $\bar{\eta} = \Omega^{-\Delta}\eta$.
(3) η is arbitrary, and $d - \Delta = (d+2)/2$.

So φ solves the field equation in the metric g exactly when $\Omega^{-(d-2)/2}\varphi$ solves it in $\Omega^2 g$. For example, a spatially flat expanding universe has a metric of the form $a(\eta)^2\times$(flat), so a massless conformally coupled field in it obeys the flat-space wave equation after the rescaling (Parker's classic observation that such a field is not produced by the expansion).

d	2	3	4	6	10	$d \to \infty$
$\Delta = \frac{d-2}{2}$	0	$\frac{1}{2}$	1	2	4	∞
$\xi = \frac{d-2}{4(d-1)}$	0	$\frac{1}{8}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{2}{9}$	$\to \frac{1}{4}$

11.2.5 Why it matters

Weyl implies conformal, and the converse needs more. Chain (11.21) showed that Weyl invariance in curved space implies conformal invariance in flat space. It is instructive to see how little the flat-space test actually probes, on the free scalar. Infinitesimally, a conformal map $x' = x + \epsilon(x)$ has $\Omega = 1 + \bar{\sigma}$ with $\bar{\sigma} = \frac{1}{d}\partial\cdot\epsilon$ (Section 11.1.5). By steps 1 and 2 of (11.21), the change of the flat action under the conformal map equals the change under the Weyl rescaling $\bar{\sigma}$ at $g = \delta$. So by (11.44) with $R = 0$ and $\Delta = (d-2)/2$,

$$\begin{aligned}\delta_{\mathrm{conf}}S &\overset{(1)}{=} \int \mathrm{d}^d x\,\bar{\sigma}\Big[\big(\tfrac{d-2}{2} - 2\xi(d-1)\big)(\partial\varphi)^2 + \big(\tfrac{d-2}{2} - 2\xi(d-1)\big)\varphi\,\partial^2\varphi\Big] \\ &\overset{(2)}{=} \Big(\tfrac{d-2}{4} - \xi(d-1)\Big)\int \mathrm{d}^d x\,\bar{\sigma}\,\partial^2(\varphi^2) \overset{(3)}{=} \Big(\tfrac{d-2}{4} - \xi(d-1)\Big)\int \mathrm{d}^d x\,(\partial^2\bar{\sigma})\,\varphi^2 \overset{(4)}{=} 0. \end{aligned} \tag{11.46}$$

(1) (11.44) at $g = \delta$ with $\Delta = \frac{d-2}{2}$, where the R term is absent.
(2) $(\partial\varphi)^2 + \varphi\partial^2\varphi = \frac{1}{2}\partial^2(\varphi^2)$.
(3) Integrate by parts twice.
(4) $\partial\cdot\epsilon$ is at most linear in x for $d > 2$ (this was shown in Section 11.1.5), and harmonic in $d = 2$ (Cauchy–Riemann), so $\partial^2\bar{\sigma} = 0$ in every case.

The result is zero *for every* ξ. The conformal $\bar{\sigma}$'s have $\partial^2\bar{\sigma} = 0$, so they never feel the $\Box\sigma$ term that fixed ξ. The flat-space theory is conformal with or without the curvature coupling. Weyl invariance is a strictly stronger demand: it needs the extra ingredient $\xi R\varphi^2$, which happens to exist.

For some theories no such ingredient exists. Karananas and Monin, and Farnsworth, Luty and Prilepina, discuss the higher-derivative free scalars $S = \frac{1}{2}\int \varphi\,\Box^k\varphi$ (up to sign), with $\Delta_\varphi = (d-2k)/2$. In flat space they are conformally invariant, but when d is even and $d < 2k$ there is no Weyl-invariant curved-space version. The simplest case with the full conformal group is $k = 3$ in $d = 4$, $\mathcal{L} = \frac{1}{2}(\partial_\mu\Box\varphi)^2$: the curvature terms needed for Weyl invariance come

with coefficients that diverge at $d = 4$. All these examples have $\Delta_\varphi < 0$, so they are not unitary (their two-point functions grow with distance).

Key idea: when a conformal theory can be coupled Weyl-invariantly (Farnsworth–Luty–Prilepina)

Let a quantum field theory in flat space be conformally invariant, and suppose that (i) it can be coupled to a background metric so that relabelling points is a symmetry, (ii) its conformal charges are integrals of local currents, (iii) it is unitary, and (iv) $d \le 10$. Then it can be placed on any curved background so that $T^\mu{}_\mu = 0$ as an operator, up to c-number curvature terms (the trace anomaly below). Equivalently it is Weyl invariant, with scalar primaries transforming as $\mathcal{O} \mapsto \Omega^{-\Delta}\mathcal{O}$ (proved for operators of low dimension and spin).

Their argument has three steps, and each is something we have already met.

1. *Flat space: improve.* Condition (ii) forces the trace of the stress tensor to be a double derivative, $T^\mu{}_\mu = \partial_\mu\partial_\nu L^{\mu\nu}$, of a local operator $L^{\mu\nu}$ of dimension $d-2$. Adding curvature couplings $\xi R L + \xi' R_{\mu\nu}L^{\mu\nu}$ to the curved-space action (invisible in flat space) changes $T_{\mu\nu}$ by second derivatives of L and makes $T^\mu{}_\mu = 0$ in flat space. For the free scalar, $L^{\mu\nu} \propto \delta^{\mu\nu}\varphi^2$, and this is our $\xi R\varphi^2$.

2. *Curved space: what could be left.* The leftover trace vanishes in flat space, so it must contain at least one power of curvature. Curvature has dimension 2 and $T^\mu{}_\mu$ has dimension d, so the first candidate is $T^\mu{}_\mu = RX$ with X a scalar of dimension $d-2$. Unitarity requires scalar primaries to have $\Delta \ge (d-2)/2$. The next candidates, (curvature)$^2 \times Y$, need $\Delta_Y = d-4 \ge (d-2)/2$, that is $d \ge 6$. So for $d < 6$ only RX is possible.

3. *Weyl rescalings commute.* If $T^\mu{}_\mu = RX$, the theory is invariant under $\delta g_{\mu\nu} = 2\sigma g_{\mu\nu}$ combined with a shift of the source ρ_X that couples to X, $\delta\rho_X = \sigma R$. Two such transformations must commute on g (they do) and on ρ_X. Using (11.32),

$$
\begin{aligned}
[\delta_{\sigma_1}, \delta_{\sigma_2}]\rho_X &\overset{(1)}{=} \sigma_2\,\delta_{\sigma_1}R - \sigma_1\,\delta_{\sigma_2}R \overset{(2)}{=} \sigma_2\big(-2\sigma_1 R - 2(d-1)\Box\sigma_1\big) - \sigma_1\big(-2\sigma_2 R - 2(d-1)\Box\sigma_2\big) \\
&= 2(d-1)\big(\sigma_1\Box\sigma_2 - \sigma_2\Box\sigma_1\big).
\end{aligned}
$$

   (1) $\delta_{\sigma_1}(\sigma_2 R) = \sigma_2\,\delta_{\sigma_1}R$, since σ_2 is a fixed function. Subtract the same with $1 \leftrightarrow 2$.
   (2) (11.32).

   With $\sigma_1 = 1$ this is $2(d-1)\Box\sigma_2$, an arbitrary function. So the theory would be unchanged by an arbitrary shift of ρ_X alone, which means X has vanishing correlators: $X = 0$, and $T^\mu{}_\mu = 0$ after all. Higher d needs more bookkeeping of the same kind.

Every possible obstruction needs an operator of a special dimension ($d-2$, $d-4$, and so on). The non-unitary $\Box^k$ theories escape through exactly this door.

The improved stress tensor: the flat-space shadow of Weyl invariance. Define the stress tensor by varying the metric,

$$
T_{\mu\nu} := \frac{2}{\sqrt{g}}\frac{\delta S}{\delta g^{\mu\nu}}. \tag{11.47}
$$

For S_{kin} at $g = \delta$ this gives $\partial_\mu\varphi\partial_\nu\varphi - \frac{1}{2}\delta_{\mu\nu}(\partial\varphi)^2$, the canonical tensor of Example 11.8 with $g = 1$. Under an infinitesimal Weyl rescaling, $\delta g^{\mu\nu} = -2\sigma g^{\mu\nu}$ and $\delta\varphi = -\Delta\sigma\varphi$, so

$$
\delta S \overset{(1)}{=} \int \mathrm{d}^d x\,\frac{\delta S}{\delta g^{\mu\nu}}\,\delta g^{\mu\nu} + \int \mathrm{d}^d x\,\frac{\delta S}{\delta\varphi}\,\delta\varphi \overset{(2)}{=} -\int \mathrm{d}^d x\,\sqrt{g}\,\sigma\left(T^\mu{}_\mu + \Delta\,\varphi\,E\right). \tag{11.48}
$$

(1) Chain rule for the change of $S[g,\varphi]$.
(2) $\frac{\delta S}{\delta g^{\mu\nu}} = \frac{\sqrt{g}}{2} T_{\mu\nu}$ and $g^{\mu\nu}T_{\mu\nu} = T^{\mu}{}_{\mu}$. Also $\frac{\delta S}{\delta \varphi} = \sqrt{g}\,E$.

If S is Weyl invariant, this vanishes for every σ, so $T^{\mu}{}_{\mu} = -\Delta\varphi E$, which is zero on solutions ($E = 0$), in every background and in particular in flat space. This is the statement of Section 11.1.5 that tracelessness gives conformal invariance, now traced back to its origin.

For the scalar the ξ term contributes to $T_{\mu\nu}$ even in flat space, where the term itself vanishes. To see it we need δR for a general $\delta g^{\mu\nu}$. Write $h_{\mu\nu} = \delta g_{\mu\nu}$, $h = g^{\mu\nu}h_{\mu\nu}$, and repeat the Palatini computation:

$$g^{\mu\nu}\delta R_{\mu\nu} \overset{(1)}{=} \nabla_\lambda\big(g^{\mu\nu}\delta\Gamma^{\lambda}_{\mu\nu} - g^{\lambda\nu}\delta\Gamma^{\mu}_{\mu\nu}\big) \overset{(2)}{=} \nabla_\lambda\big(\nabla_\mu h^{\mu\lambda} - \tfrac{1}{2}\nabla^\lambda h - \tfrac{1}{2}\nabla^\lambda h\big) \overset{(3)}{=} g_{\mu\nu}\Box\,\delta g^{\mu\nu} - \nabla_\mu\nabla_\nu\,\delta g^{\mu\nu}.$$

(1) The Palatini identity, contracted with $g^{\mu\nu}$.
(2) From $\delta\Gamma^{\lambda}_{\mu\nu} = \frac{1}{2}g^{\lambda\kappa}(\nabla_\mu h_{\nu\kappa} + \nabla_\nu h_{\mu\kappa} - \nabla_\kappa h_{\mu\nu})$: the first contraction is $\nabla_\mu h^{\mu\lambda} - \frac{1}{2}\nabla^\lambda h$ and $\delta\Gamma^{\mu}_{\mu\nu} = \frac{1}{2}\nabla_\nu h$.
(3) $h^{\mu\nu} = -\delta g^{\mu\nu}$ (vary $g^{\mu\lambda}g_{\lambda\nu} = \delta^{\mu}_{\nu}$) and $h = -g_{\mu\nu}\delta g^{\mu\nu}$.

So $\delta R = R_{\mu\nu}\delta g^{\mu\nu} + g_{\mu\nu}\Box\delta g^{\mu\nu} - \nabla_\mu\nabla_\nu\delta g^{\mu\nu}$. (Check: with $\delta g^{\mu\nu} = -2\sigma g^{\mu\nu}$ it gives $-2\sigma R - 2d\,\Box\sigma + 2\Box\sigma$, which is (11.32).) In flat space only the derivative terms survive, and

$$\delta S_\xi\big|_{g=\delta} \overset{(1)}{=} \frac{\xi}{2}\int \mathrm{d}^d x\,\varphi^2\big(\delta_{\mu\nu}\partial^2 - \partial_\mu\partial_\nu\big)\delta g^{\mu\nu} \overset{(2)}{=} \frac{\xi}{2}\int \mathrm{d}^d x\,\delta g^{\mu\nu}\big(\delta_{\mu\nu}\partial^2 - \partial_\mu\partial_\nu\big)\varphi^2,$$

(1) $R = R_{\mu\nu} = 0$ in flat space, so neither $\delta\sqrt{g}\,R$ nor $R_{\mu\nu}\delta g^{\mu\nu}$ contributes.
(2) Integrate by parts twice.

so by (11.47) the flat-space stress tensor is

$$\Theta_{\mu\nu} = \partial_\mu\varphi\,\partial_\nu\varphi - \tfrac{1}{2}\delta_{\mu\nu}(\partial\varphi)^2 - \xi\big(\partial_\mu\partial_\nu - \delta_{\mu\nu}\partial^2\big)\varphi^2. \tag{11.49}$$

The last term is the **improvement term**. With $\xi = \frac{1}{6}$ in $d = 4$ it is exactly the term quoted in Example 11.8, and for general d it is the improvement found from the virial current in Chapter T5 (section "Fixed points are conformal field theories"). Its trace vanishes on shell:

$$\begin{aligned}\Theta^{\mu}{}_{\mu} &\overset{(1)}{=} \Big(1 - \frac{d}{2}\Big)(\partial\varphi)^2 + \xi(d-1)\,\partial^2(\varphi^2)\\ &\overset{(2)}{=} \frac{2-d}{2}(\partial\varphi)^2 + 2\xi(d-1)\big[(\partial\varphi)^2 + \varphi\,\partial^2\varphi\big]\\ &\overset{(3)}{=} \frac{d-2}{2}\,\varphi\,\partial^2\varphi.\end{aligned}$$

(1) $\delta^{\mu}{}_{\mu} = d$, so the improvement contributes $-\xi(1-d)\partial^2\varphi^2$.
(2) $\partial^2(\varphi^2) = 2(\partial\varphi)^2 + 2\varphi\partial^2\varphi$.
(3) $2\xi(d-1) = \frac{d-2}{2}$ by (11.40), so the $(\partial\varphi)^2$ terms cancel.

This is $-\Delta\varphi E$ with $E = -\partial^2\varphi$, as (11.48) predicted, and it is zero on shell. The improvement term is conserved identically and does not change the charges (Chapter T5), so flat-space physics cannot see ξ except through the trace. In short: the curvature coupling that Weyl invariance demands in curved space leaves behind, in flat space, exactly the improvement that makes $T_{\mu\nu}$ traceless.

The trace anomaly. Everything so far is classical. In the quantum theory the path integral must be regularised, and every regulator introduces a length measured with the metric g. A

Weyl rescaling rescales that length, so Weyl invariance can fail by a local c-number built from curvature. In $d = 2$ the result is $\langle T^{\mu}{}_{\mu} \rangle \propto c\, R$ with c the central charge. This **trace anomaly** (or Weyl anomaly) is computed for free bosons from the heat kernel in Chapter 10 (section "The Weyl anomaly of d free bosons"). It is the "c-number curvature terms" allowed in the Farnsworth–Luty–Prilepina statement above.

11.2.6 Examples

Worked example 11.11: The two-dimensional free boson: $\Delta = 0$ and $\xi = 0$

In $d = 2$ the formulae give $\Delta = (2-2)/2 = 0$ and $\xi = 0/(4 \cdot 1) = 0$. The reason is visible directly. The combination $\sqrt{g}\, g^{\mu\nu}$ is Weyl invariant in two dimensions:

$$\sqrt{\bar{g}}\, \bar{g}^{\mu\nu} \overset{(1)}{=} e^{2\sigma} \sqrt{g}\, e^{-2\sigma} g^{\mu\nu} = \sqrt{g}\, g^{\mu\nu}.$$

(1) $\sqrt{\bar{g}} = e^{d\sigma}\sqrt{g}$ with $d = 2$, and $\bar{g}^{\mu\nu} = e^{-2\sigma} g^{\mu\nu}$.

So $S_{\text{kin}} = \frac{1}{2}\int \sqrt{g}\, g^{\mu\nu}\partial_\mu\varphi\partial_\nu\varphi$ is unchanged with φ left alone ($\Delta = 0$). No derivative ever hits σ, so no $\Box\sigma$ appears, and no curvature term is needed. In fact a curvature term would *spoil* the invariance: by (11.34),

$$\begin{aligned} \frac{\xi}{2}\int \mathrm{d}^2x\, \sqrt{\bar{g}}\, \bar{R}\, \varphi^2 &\overset{(1)}{=} \frac{\xi}{2}\int \mathrm{d}^2x\, e^{2\sigma}\sqrt{g}\, e^{-2\sigma}(R - 2\Box\sigma)\, \varphi^2 \\ &= \frac{\xi}{2}\int \mathrm{d}^2x\, \sqrt{g}\, R\, \varphi^2 - \xi \int \mathrm{d}^2x\, \sqrt{g}\, \varphi^2\, \Box\sigma, \end{aligned}$$

(1) $\sqrt{\bar{g}} = e^{2\sigma}\sqrt{g}$, (11.34), and $\bar{\varphi} = \varphi$.

and the last term is not zero for general φ unless $\xi = 0$. (Without φ^2, the integral $\int \sqrt{g}\, R = 4\pi\chi$ is Weyl invariant, because $\int \sqrt{g}\, \Box\sigma = 0$ on a closed surface: Gauss–Bonnet again.)

Two remarks connect this with earlier sections. First, $\Delta = 0$ does not make φ a primary: its two-point function is a logarithm (see the discussion after (11.69)). Weyl weight zero is a statement about the action. Second, $\partial_\mu\varphi$ is a rank $(0,1)$ field with $\Delta = 1$, so by (11.26) its Weyl weight is $1 + 0 - 1 = 0$, which agrees with φ being unchanged. Quantum mechanically this theory has the trace anomaly with $c = 1$.

Worked example 11.12: Maxwell theory: Weyl invariant in $d = 4$ only

The curved-space Maxwell action is

$$S = \frac{1}{4}\int \mathrm{d}^d x\, \sqrt{g}\, g^{\mu\alpha} g^{\nu\beta} F_{\mu\nu} F_{\alpha\beta}, \qquad F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu.$$

The field strength needs no metric: the Christoffel symbols in $\nabla_\mu A_\nu - \nabla_\nu A_\mu$ cancel because they are symmetric in $\mu\nu$. Rescale $g \mapsto e^{2\sigma} g$ and leave A_μ alone:

$$\sqrt{\bar{g}}\, \bar{g}^{\mu\alpha}\bar{g}^{\nu\beta} F_{\mu\nu}F_{\alpha\beta} \overset{(1)}{=} e^{d\sigma} e^{-2\sigma} e^{-2\sigma} \sqrt{g}\, g^{\mu\alpha} g^{\nu\beta} F_{\mu\nu}F_{\alpha\beta} = e^{(d-4)\sigma} \sqrt{g}\, g^{\mu\alpha} g^{\nu\beta} F_{\mu\nu}F_{\alpha\beta}.$$

(1) $\sqrt{\bar{g}} = e^{d\sigma}\sqrt{g}$ and two inverse metrics, each $e^{-2\sigma} g^{-1}$. F is unchanged because A is.

In $d = 4$ the factor is 1 for every $\sigma(x)$: Maxwell theory is exactly Weyl invariant, with A_μ of Weyl weight 0. That is the value predicted by (11.26) for $\Delta_A = 1$, $(n, m) = (0, 1)$. No curvature term is needed, and none would be allowed, since a term like $R A_\mu A^\mu$ is not gauge invariant.

In $d \neq 4$ a *constant* rescaling can still be compensated, by $A \mapsto \lambda^{-w} A$ with $w = \frac{d-4}{2}$ (then $\Delta_A = w + 1 = \frac{d-2}{2}$). For a local σ, however,

$$\partial_\mu\big(e^{-w\sigma}A_\nu\big) - \partial_\nu\big(e^{-w\sigma}A_\mu\big) \overset{(1)}{=} e^{-w\sigma}\Big[F_{\mu\nu} - w\big(\partial_\mu\sigma\, A_\nu - \partial_\nu\sigma\, A_\mu\big)\Big].$$

(1) Product rule on each term.

The extra piece contains A_μ itself, not just F, so it changes under a gauge transformation $A \mapsto A + \partial\chi$. Any curvature term that could cancel it would also contain undifferentiated A and break gauge invariance. So Maxwell theory in $d \neq 4$ has no Weyl-invariant form. By the contrapositive of (11.21) this is consistent with what Chapter T5 finds in flat space: for $d \neq 4$ free Maxwell theory is scale invariant but not conformally invariant. Farnsworth, Luty and Prilepina note the general version: free p-form gauge fields with $d \neq 2(p+1)$ are scale invariant without being conformal. The flat-space trace, $T^\mu{}_\mu = (1 - \frac{d}{4})F^2$, vanishes precisely in $d = 4$.

Worked example 11.13: A Weyl rescaling that is not conformal: the plane becomes a cylinder

In polar coordinates on $\mathbb{R}^d$, $\delta = \mathrm{d}r^2 + r^2\mathrm{d}\Omega_{d-1}^2$, where $\mathrm{d}\Omega_{d-1}^2$ is the metric of the unit sphere S^{d-1}. With $\tau = \ln r$, $\mathrm{d}r = r\,\mathrm{d}\tau$, so

$$\delta = r^2\big(\mathrm{d}\tau^2 + \mathrm{d}\Omega_{d-1}^2\big) \qquad \Longrightarrow \qquad r^{-2}\,\delta = \mathrm{d}\tau^2 + \mathrm{d}\Omega_{d-1}^2.$$

The Weyl rescaling with $\Omega = 1/r$, that is $\sigma = -\ln r$, turns flat space (minus the origin) into the cylinder $\mathbb{R} \times S^{d-1}$. Every point stays where it is. The sphere of radius r becomes a sphere of radius 1 at height $\tau = \ln r$ (Figure 11.8). Is this rescaling conformal? Compute the curvature with (11.33), $R = 0$. We need, for $r \neq 0$,

$$\partial_\mu\sigma \overset{(1)}{=} -\frac{x_\mu}{r^2}, \qquad (\nabla\sigma)^2 = \frac{1}{r^2}, \qquad \Box\sigma \overset{(2)}{=} -\frac{1}{r^{d-1}}\frac{\mathrm{d}}{\mathrm{d}r}\Big(r^{d-1}\cdot\frac{1}{r}\Big) = -\frac{d-2}{r^2}.$$

(1) $\partial_\mu r = x_\mu/r$.

(2) The radial part of the flat Laplacian in d dimensions is $r^{1-d}\frac{\mathrm{d}}{\mathrm{d}r}(r^{d-1}\frac{\mathrm{d}}{\mathrm{d}r})$, and $\frac{\mathrm{d}}{\mathrm{d}r}\ln r = \frac{1}{r}$.

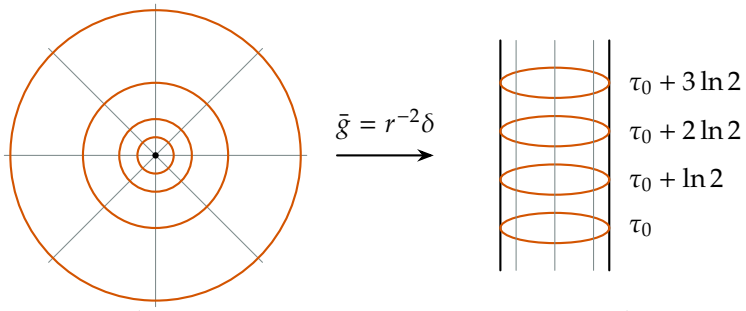
Then

$$\bar{R} \overset{(1)}{=} r^2\Big[-2(d-1)\Big(-\frac{d-2}{r^2}\Big) - (d-1)(d-2)\frac{1}{r^2}\Big] \overset{(2)}{=} (d-1)(d-2).$$

(1) (11.33) with $R = 0$ and $e^{-2\sigma} = r^2$.

(2) $2(d-1)(d-2) - (d-1)(d-2) = (d-1)(d-2)$.

This is the scalar curvature of the unit sphere S^{d-1}, as it should be for $\mathbb{R} \times S^{d-1}$. For $d > 2$ it is not zero, so the cylinder is curved and the rescaling is *not* conformal: no relabelling of points turns it back into flat space. For $d = 2$, $\bar{R} = 0$: the cylinder $\mathbb{R} \times S^1$ is flat, and the rescaling is undone by the conformal map $w = \ln z$, the cylinder map used in Example 11.34 and Section 11.4.3. For $d > 2$ this Weyl rescaling is what replaces that map in radial quantisation: a Weyl-invariant theory on flat space is the same as the theory on the cylinder, and dilations $r \mapsto \lambda r$ become time translations $\tau \mapsto \tau + \ln\lambda$.



flat $\mathbb{R}^d$ (a slice): spheres of radius $r_0, 2r_0, 4r_0, 8r_0$

cylinder $\mathbb{R} \times S^{d-1}$: unit spheres at height $\tau = \ln r$

Figure 11.8: The Weyl rescaling $\bar{g} = r^{-2}\delta$ of Example 11.13. On the left, concentric spheres in flat space, each twice the radius of the one inside it, and a few radial lines. The new ruler shrinks every length at distance r by the factor $1/r$, so all the spheres become unit spheres, stacked at equal spacing $\ln 2$ along the cylinder on the right (heights $\tau = \ln r$, with $\tau_0 = \ln r_0$), and the radial lines become straight lines along the cylinder. No point has moved. For $d > 2$ the cylinder is curved, so this rescaling is not a conformal map.

Try it: the conformal coupling in other dimensions

Show that in $d = 3$ the conformally coupled scalar has $\Delta = \frac{1}{2}$ and $\xi = \frac{1}{8}$, and in $d = 6$ it has $\Delta = 2$ and $\xi = \frac{1}{5}$. Then check (11.41) directly in $d = 3$. *Answer:* $\xi = \frac{d-2}{4(d-1)}$ gives $\frac{1}{8}$ and $\frac{4}{20} = \frac{1}{5}$. In $d = 3$, $\Delta^2 - \xi(d-1)(d-2) = \frac{1}{4} - \frac{1}{8} \cdot 2 \cdot 1 = 0$.

Watch out: Weyl is not conformal, and other traps

- *Different operations.* A conformal transformation moves points and keeps the geometry. A Weyl rescaling keeps the points and changes the geometry, usually from flat to curved (Example 11.13). Only the rescalings that a relabelling can undo are conformal (Definition 11.5). Writing "conformal transformation $g \mapsto \Omega^2 g$" for a general Ω is common and is exactly this confusion.
- *Weyl implies conformal, not conversely.* Chain (11.21) goes one way. The free scalar is conformal in flat space for any ξ (11.46), but Weyl invariant only for $\xi = \frac{d-2}{4(d-1)}$, and the non-unitary $\Box^k$ theories are conformal without any Weyl-invariant version.
- *The sign of R.* With our convention (11.17) spheres have $R > 0$, the Euclidean action is $\frac{1}{2}\int \sqrt{g}\,[(\nabla\varphi)^2 + \xi R\varphi^2]$, and the conformal operator is $-\Box + \xi R$ with $\xi = +\frac{1}{6}$ in $d = 4$. Books that define the Ricci tensor with the opposite sign (so spheres have $R < 0$) write $-\xi R\varphi^2$ for the same physics. Check any source on the round sphere before copying a sign. The derivation never used the signature: in Lorentzian signature the Weyl-invariant combination is still $g^{\mu\nu}\partial_\mu\varphi\,\partial_\nu\varphi + \xi R\varphi^2$, with R computed from (11.17) and $\sqrt{|g|}$ in place of $\sqrt{g}$, whatever overall sign the action carries.
- *How Ω is written.* We write $g \mapsto \Omega^2 g = e^{2\sigma} g$. Some papers write $g \mapsto \Omega\, g$ (Karananas–Monin), or give only the infinitesimal form $\delta g_{\mu\nu} = 2\sigma g_{\mu\nu}$ (Farnsworth–Luty–Prilepina). The exponent in $\phi \mapsto \Omega^{-\Delta}\phi$ changes accordingly, so convert before comparing.
- *Weyl weight is not dimension for tensors.* Under a pure Weyl rescaling a primary of rank (n, m) scales as $\Omega^{-(\Delta+n-m)}$ (11.26), not $\Omega^{-\Delta}$. The Maxwell field has $\Delta = 1$ but does not change at all.

Recap

- A conformal map $x \mapsto x'$ moves points and stretches every short step near x by $\Omega(x)$. A Weyl rescaling $g \mapsto \Omega^2 g$ moves nothing and changes the ruler.
- Conformal transformations are the Weyl rescalings that a diffeomorphism undoes, $f^* g = \Omega^2 g$. In flat space: $\Omega^2 \delta$ must be flat (in $d = 2$, $\Box \ln \Omega = 0$).
- A primary of rank (n, m) and dimension Δ transforms by (11.23). Net effect: multiplied by $\Omega^{-\Delta}$ and rotated. Scalar: $\phi'(x') = \Omega^{-\Delta}\phi(x)$, checked on the dilation and the inversion. Weyl weight $\Delta + n - m$.
- A constant rescaling of $\frac{1}{2}\int \sqrt{g}(\nabla\varphi)^2$ gives $\lambda^{d-2-2\Delta}$, so $\Delta = \frac{d-2}{2}$.
- $\bar{R} = e^{-2\sigma}[R - 2(d-1)\Box\sigma - (d-1)(d-2)(\nabla\sigma)^2]$. Local Weyl invariance of $\frac{1}{2}\int \sqrt{g}[(\nabla\varphi)^2 + \xi R\varphi^2]$: the $\Box\sigma$ terms fix $\xi = \frac{d-2}{4(d-1)}$ ($\frac{1}{6}$ in $d = 4$), and the $(\nabla\sigma)^2$ terms then cancel automatically. Doing it infinitesimally with Δ unknown fixes both numbers at once.
- Weyl invariance gives $T^\mu{}_\mu = 0$ on shell. In flat space the ξ term survives as the improvement term $-\xi(\partial_\mu\partial_\nu - \delta_{\mu\nu}\partial^2)\varphi^2$. Quantum mechanically the trace anomaly $\propto cR$ appears (Chapter 10).
- Farnsworth–Luty–Prilepina: local, unitary conformal theories with local conformal currents in $d \leq 10$ are Weyl invariant. Counterexamples are non-unitary ($\Box^k$ with d even, $d < 2k$).
- Examples: 2D boson ($\Delta = \xi = 0$, $\sqrt{g}g^{\mu\nu}$ invariant), Maxwell (Weyl invariant only in $d = 4$), plane to cylinder (Weyl but not conformal for $d > 2$).

11.3 Conformal field theories in two dimensions

We now know that in two dimensions the conformal transformations are the holomorphic maps. This section turns that fact into a working toolkit. The questions are: how do fields transform under a holomorphic change of coordinates? What does conformal symmetry force the correlation functions to be? And what plays the role of the conserved current, the object whose charges generate the symmetry? The answers are, in order: *primary fields* transform like tensors with two weights $(h, \bar{h})$; two-point functions are fixed up to a constant; and the generator is the holomorphic component $T(z)$ of the stress tensor, whose operator product with any field encodes the whole conformal Ward identity. Along the way we meet the two simplest conformal field theories, the free Majorana fermion and the free boson, and use them to check every formula.

Nash works on the complex plane $\mathbb{C}$ or on the Riemann sphere $\mathbb{C} \cup \{\infty\}$. General Riemann surfaces appear only in §5 of his chapter (the next part of these notes). We do the same.

11.3.1 Complex coordinates on the plane

Two-dimensional Euclidean space has coordinates (x_1, x_2) and metric $\mathrm{d}s^2 = \mathrm{d}x_1^2 + \mathrm{d}x_2^2$. Trade them for

$$z = x_1 + ix_2, \qquad \bar{z} = x_1 - ix_2. \tag{11.50}$$

For a real point, $\bar{z}$ is the complex conjugate of z. But in most calculations we treat z and $\bar{z}$ as *independent* variables, as if x_1, x_2 were complex. This is the trick that makes two-dimensional

conformal field theory work: a function of (x_1, x_2) becomes a function of $(z, \bar{z})$, and holomorphic means "depends on z only".

The derivatives dual to $z, \bar{z}$ are

$$\partial \equiv \partial_z = \tfrac{1}{2}(\partial_1 - i\partial_2), \qquad \bar{\partial} \equiv \partial_{\bar{z}} = \tfrac{1}{2}(\partial_1 + i\partial_2). \tag{11.51}$$

Check: $\partial z = \frac{1}{2}(1 - i \cdot i) = 1$ and $\partial \bar{z} = \frac{1}{2}(1 + i \cdot i) = 0$, and similarly $\bar{\partial}\bar{z} = 1$, $\bar{\partial} z = 0$. The Laplacian factorises,

$$4\partial\bar{\partial} \overset{(1)}{=} (\partial_1 - i\partial_2)(\partial_1 + i\partial_2) \overset{(2)}{=} \partial_1^2 + \partial_2^2 = \nabla^2. \tag{11.52}$$

(1) Definition (11.51); the two factors of $\frac{1}{2}$ give $\frac{1}{4}$, cancelled by the 4.
(2) Multiply out: the cross terms $i\partial_1\partial_2 - i\partial_2\partial_1$ cancel because partial derivatives commute.

The metric and the measure. Substituting $\mathrm{d}x_1 = \frac{1}{2}(\mathrm{d}z + \mathrm{d}\bar{z})$ and $\mathrm{d}x_2 = \frac{1}{2i}(\mathrm{d}z - \mathrm{d}\bar{z})$,

$$\mathrm{d}s^2 \overset{(1)}{=} \tfrac{1}{4}(\mathrm{d}z + \mathrm{d}\bar{z})^2 - \tfrac{1}{4}(\mathrm{d}z - \mathrm{d}\bar{z})^2 \overset{(2)}{=} \mathrm{d}z\,\mathrm{d}\bar{z}.$$

(1) $(\frac{1}{2i})^2 = -\frac{1}{4}$.
(2) The squares $\mathrm{d}z^2$ and $\mathrm{d}\bar{z}^2$ cancel; the cross terms add to $\frac{1}{4} \cdot 4\,\mathrm{d}z\,\mathrm{d}\bar{z}$.

So the metric components are $g_{z\bar{z}} = g_{\bar{z}z} = \frac{1}{2}$ and $g_{zz} = g_{\bar{z}\bar{z}} = 0$ (the symmetric tensor $\mathrm{d}z\,\mathrm{d}\bar{z}$ is $\frac{1}{2}(\mathrm{d}z \otimes \mathrm{d}\bar{z} + \mathrm{d}\bar{z} \otimes \mathrm{d}z)$). The inverse metric has $g^{z\bar{z}} = 2$. Indices are lowered with g: for a vector v, $v_z = g_{z\bar{z}} v^{\bar{z}} = \frac{1}{2} v^{\bar{z}}$.

For the area element, use forms (Chapter T1):

$$\mathrm{d}z \wedge \mathrm{d}\bar{z} \overset{(1)}{=} (\mathrm{d}x_1 + i\,\mathrm{d}x_2) \wedge (\mathrm{d}x_1 - i\,\mathrm{d}x_2) \overset{(2)}{=} -2i\,\mathrm{d}x_1 \wedge \mathrm{d}x_2. \tag{11.53}$$

(1) Definition (11.50).
(2) $\mathrm{d}x_1 \wedge \mathrm{d}x_1 = 0$; the cross terms are $-i\,\mathrm{d}x_1 \wedge \mathrm{d}x_2 + i\,\mathrm{d}x_2 \wedge \mathrm{d}x_1 = -2i\,\mathrm{d}x_1 \wedge \mathrm{d}x_2$.

In CFT texts "$\mathrm{d}z\,\mathrm{d}\bar{z}$" or "$\mathrm{d}^2 z$" usually means the positive measure $|\mathrm{d}z \wedge \mathrm{d}\bar{z}| = 2\,\mathrm{d}^2 x$, where $\mathrm{d}^2 x = \mathrm{d}x_1 \mathrm{d}x_2$. Conventions differ by this factor 2 from book to book. To stay safe we write all actions and integrals with $\mathrm{d}^2 x$.

Tensors in complex components. A tensor $T_{\mu\nu}$ has complex components obtained by the chain rule, $T_{zz} = \frac{\partial x^\mu}{\partial z}\frac{\partial x^\nu}{\partial z} T_{\mu\nu}$, with $\partial x_1/\partial z = \frac{1}{2}$ and $\partial x_2/\partial z = \frac{1}{2i}$. We compute these explicitly for the stress tensor in Section 11.3.10.

The key distributional identity. The function $1/z$ is holomorphic away from 0, so $\bar{\partial}(1/z) = 0$ for $z \neq 0$. At $z = 0$ it has a pole, and $\bar{\partial}(1/z)$ is a delta function:

$$\bar{\partial}\,\frac{1}{z} = \pi\,\delta^2(x), \qquad \partial\bar{\partial}\,\ln|z|^2 = \pi\,\delta^2(x). \tag{11.54}$$

Worked example 11.14: Why $\bar{\partial}(1/z) = \pi\delta^2(x)$: Stokes in the plane

A distribution is defined by what it does to a smooth test function $g(x)$ that vanishes outside a large disc D. The derivative of a distribution is defined by moving the derivative onto g with a minus sign. We also need Stokes' theorem (Chapter T2) in complex form: for

any function F, $\mathrm{d}(F\,\mathrm{d}z) = \bar\partial F\,\mathrm{d}\bar z \wedge \mathrm{d}z = 2i\,\bar\partial F\,\mathrm{d}^2x$ by (11.53), so $\int_R \mathrm{d}^2x\,\bar\partial F = \frac{1}{2i}\oint_{\partial R} F\,\mathrm{d}z$. Let D_ϵ be the disc of radius ϵ about 0. Then

$$\begin{aligned}\left\langle \bar\partial\frac{1}{z},\, g\right\rangle &\overset{(1)}{=} -\lim_{\epsilon\to0}\int_{D\setminus D_\epsilon} \mathrm{d}^2x\;\frac{\bar\partial g}{z} \overset{(2)}{=} -\lim_{\epsilon\to0}\int_{D\setminus D_\epsilon} \mathrm{d}^2x\;\bar\partial\Big(\frac{g}{z}\Big)\\ &\overset{(3)}{=} -\lim_{\epsilon\to0}\frac{1}{2i}\oint_{\partial(D\setminus D_\epsilon)}\frac{g}{z}\,\mathrm{d}z \overset{(4)}{=} \lim_{\epsilon\to0}\frac{1}{2i}\oint_{|z|=\epsilon}\frac{g}{z}\,\mathrm{d}z\\ &\overset{(5)}{=} \frac{1}{2i}\,2\pi i\,g(0) = \pi\,g(0).\end{aligned}$$

(1) Definition of the distributional derivative; $1/z$ is integrable in 2D, so removing a small disc changes nothing in the limit.
(2) Product rule: $\bar\partial(g/z) = (\bar\partial g)/z + g\,\bar\partial(1/z)$, and $\bar\partial(1/z) = 0$ away from the origin.
(3) Stokes' theorem in the complex form just stated.
(4) $g = 0$ on the outer circle. The inner circle is part of the boundary traversed clockwise; writing it counter-clockwise flips the sign.
(5) On $|z| = \epsilon$, $z = \epsilon e^{i\theta}$ and $\mathrm{d}z/z = i\,\mathrm{d}\theta$; as $\epsilon \to 0$, $g \to g(0)$ and $\oint i\,\mathrm{d}\theta = 2\pi i$.

So $\bar\partial(1/z) = \pi\delta^2(x)$. The second form in (11.54) follows from $\partial \ln|z|^2 = \partial(\ln z + \ln\bar z) = 1/z$. As a cross-check, $\nabla^2 \ln r = 2\pi\delta^2(x)$ is Gauss's law for a line charge in 2D, and $4\partial\bar\partial \ln|z|^2 = \nabla^2(2\ln r) = 4\pi\delta^2(x)$, consistent with (11.52).

11.3.2 Conformal maps are holomorphic maps

We saw in Section 11.1.6 that infinitesimal conformal maps are holomorphic. The finite statement is just as direct. Let $w = f(z)$ be holomorphic, $f = u + iv$. The Cauchy–Riemann equations $u_1 = v_2$, $u_2 = -v_1$ make the Jacobian matrix

$$J = \begin{pmatrix} u_1 & u_2 \\ v_1 & v_2 \end{pmatrix} = \begin{pmatrix} a & -b \\ b & a \end{pmatrix} = |f'(z)| \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \qquad f'(z) = a + ib = |f'|e^{i\theta}. \tag{11.55}$$

It is a rotation by $\arg f'$ times a stretch by $|f'|$. So $J^TJ = |f'|^2\mathbb{1}$, and in the language of (11.9), $\Omega = |f'(z)|$. Equivalently $\mathrm{d}w\,\mathrm{d}\bar w = |f'(z)|^2\mathrm{d}z\,\mathrm{d}\bar z$. Holomorphic maps preserve orientation; antiholomorphic maps $\bar f(\bar z)$ are conformal but reverse orientation.

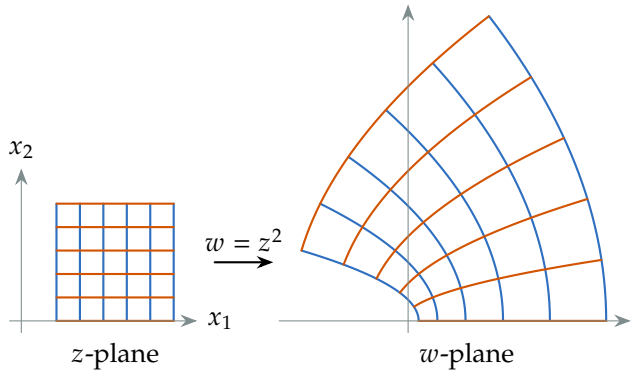


Figure 11.9: The holomorphic map $w = z^2$ applied to a rectangular grid. Vertical lines (blue) and horizontal lines (orange) become two families of parabolas. The images are curved and stretched, but they still cross at right angles: the Jacobian (11.55) is a rotation times a scale at every point. Where $|f'| = 2|z|$ is larger (far from the origin) the cells are bigger.

Worked example 11.15: Two maps we will use constantly

(a) *Rotation and dilation*: $f(z) = \lambda z$ with $\lambda = \rho e^{i\theta}$ complex. Then $f' = \lambda$, so the map rotates by θ and scales by ρ. In real coordinates this is the 2×2 matrix $\rho\left(\begin{smallmatrix}\cos\theta & -\sin\theta \\ \sin\theta & \cos\theta\end{smallmatrix}\right)$: one complex parameter packages a dilation and a rotation.
(b) *The exponential map*: $z = e^w$ with $w = \tau + i\sigma$. Then $|z| = e^\tau$ and $\arg z = \sigma$. The line $\tau =$ const is a circle of radius e^τ; the line $\sigma =$ const is a ray. Since $\sigma \sim \sigma + 2\pi$ gives the same z, the w-plane is really a cylinder of circumference 2π, with τ running along it. The far past $\tau \to -\infty$ is the origin $z = 0$, the far future $\tau \to +\infty$ is $z = \infty$. Here $f'(w) = e^w = z$, so a unit cell on the cylinder is stretched by $|z|$: circles far from the origin are bigger. This map is the basis of radial quantisation in Section 11.4.3.

11.3.3 The local conformal algebra: the Witt algebra

An infinitesimal holomorphic map is $z \mapsto z + \epsilon(z)$ with ϵ holomorphic in some region. Expand in a Laurent series, $\epsilon(z) = -\sum_n \epsilon_n z^{n+1}$. The map acts on functions through the vector field $\epsilon(z)\partial_z$, and it is natural to use the basis

$$\ell_n = -z^{n+1}\partial_z, \qquad \bar{\ell}_n = -\bar{z}^{n+1}\partial_{\bar{z}}, \qquad n \in \mathbb{Z}. \tag{11.56}$$

(The minus sign and the power $n+1$ are chosen so that the commutator comes out clean and so that ℓ_n lowers the power of z by n.)

Worked example 11.16: The Witt algebra commutator

Act on a test function $g(z)$:

$$\begin{aligned}[\ell_m, \ell_n]g &\overset{(1)}{=} z^{m+1}\partial\big(z^{n+1}\partial g\big) - z^{n+1}\partial\big(z^{m+1}\partial g\big) \\ &\overset{(2)}{=} z^{m+1}\big((n+1)z^n\partial g + z^{n+1}\partial^2 g\big) - z^{n+1}\big((m+1)z^m\partial g + z^{m+1}\partial^2 g\big) \\ &\overset{(3)}{=} (n-m)\, z^{m+n+1}\partial g \overset{(4)}{=} (m-n)\,\ell_{m+n}\, g.\end{aligned}$$

(1) $\ell_m\ell_n g = (-z^{m+1}\partial)(-z^{n+1}\partial g)$: the two minus signs cancel.
(2) Product rule.
(3) The second-derivative terms cancel; collect the first-derivative terms, $(n+1)-(m+1) = n-m$.
(4) $\ell_{m+n} = -z^{m+n+1}\partial$, so $z^{m+n+1}\partial = -\ell_{m+n}$.

So

$$[\ell_m, \ell_n] = (m-n)\,\ell_{m+n}, \qquad [\bar{\ell}_m, \bar{\ell}_n] = (m-n)\,\bar{\ell}_{m+n}, \qquad [\ell_m, \bar{\ell}_n] = 0. \tag{11.57}$$

This is the **Witt algebra** (two commuting copies). It is the Virasoro algebra (11.102) with $c = 0$. The central term will appear only in the quantum theory.

Which generators are global? The vector field $\ell_n = -z^{n+1}\partial_z$ is regular at $z = 0$ only if $n \geq -1$. To test $z = \infty$, use the coordinate $u = 1/z$ near infinity. Then $\partial_z = \frac{\mathrm{d}u}{\mathrm{d}z}\partial_u = -u^2\partial_u$, so

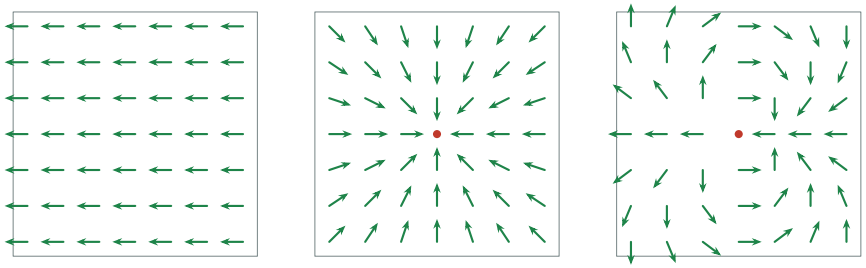
$$\ell_n = -z^{n+1}\partial_z \overset{(1)}{=} -u^{-n-1}(-u^2\partial_u) = u^{1-n}\partial_u,$$

(1) Substitute $z = 1/u$ and $\partial_z = -u^2\partial_u$.

which is regular at $u = 0$ only if $n \leq 1$. So on the whole Riemann sphere only

$$\ell_{-1} = -\partial_z, \qquad \ell_0 = -z\partial_z, \qquad \ell_1 = -z^2\partial_z \tag{11.58}$$

(and their barred partners) are globally defined. Their meaning: ℓ_{-1} generates translations, $\ell_0 + \bar{\ell}_0$ dilations, $i(\ell_0 - \bar{\ell}_0)$ rotations, and ℓ_1 special conformal transformations. That is $3 + 3 = 6$ real generators, matching $\dim \mathrm{SO}(3,1) = 6$ from (11.14). The other ℓ_n are perfectly good *local* symmetries (they are conformal Killing vectors on any annulus), and in quantum field theory, where the Ward identities are local, they constrain the theory just as much.



$\ell_{-1} = -\partial_z$: translation $\quad$ $\ell_0 = -z\partial_z$: dilation $\quad$ $\ell_1 = -z^2\partial_z$: SCT

Figure 11.10: The three global conformal vector fields on the plane, drawn as the flow $\dot{z} = \epsilon(z) = -z^{n+1}$ (arrows show direction only). Left: a uniform shift. Middle: everything flows radially into the fixed point 0 (red dot); multiplying by i would give rotation about 0 instead. Right: the field $-z^2$ has a double zero at the origin; its flow lines are circles through the origin, the images of straight lines under the inversion $z \mapsto 1/z$. Only these three (and i times them) extend smoothly to $z = \infty$.

Watch out: vector fields versus coordinate changes

The flow of $\ell_n = -z^{n+1}\partial_z$ moves points by $\delta z = -\epsilon_n z^{n+1}$ when we write the transformation as $\exp(\epsilon_n \ell_n)$. Some authors put the minus sign elsewhere. Nothing physical depends on it; what matters is that the commutator (11.57) has the sign $(m - n)$, which with our conventions it does.

11.3.4 The global conformal group: Möbius transformations

Exponentiating the three global generators gives the **Möbius transformations**

$$z \mapsto f(z) = \frac{az + b}{cz + d}, \qquad a, b, c, d \in \mathbb{C}, \quad ad - bc = 1. \tag{11.59}$$

They are the only one-to-one holomorphic maps of the Riemann sphere to itself. (A holomorphic map of the sphere is a rational function; to be one-to-one it must have exactly one pole and one zero, so it is of degree one.) Each one is conformal everywhere: $f'(z) = (ad - bc)/(cz + d)^2 = 1/(cz + d)^2$, which never vanishes.

The nicest fact about them is that composition is matrix multiplication. Represent f by the matrix $M_f = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}(2, \mathbb{C})$ and write z as the ratio $z = z_1/z_2$ of a column vector $(z_1, z_2)^T$.

Worked example 11.17: Composition is matrix multiplication

Let $f \leftrightarrow \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ and $g \leftrightarrow \left(\begin{smallmatrix} a' & b' \\ c' & d' \end{smallmatrix}\right)$. Then

$$f(g(z)) \overset{(1)}{=} \frac{a\,\frac{a'z+b'}{c'z+d'} + b}{c\,\frac{a'z+b'}{c'z+d'} + d} \overset{(2)}{=} \frac{a(a'z+b') + b(c'z+d')}{c(a'z+b') + d(c'z+d')}$$
$$\overset{(3)}{=} \frac{(aa' + bc')z + (ab' + bd')}{(ca' + dc')z + (cb' + dd')},$$

(1) Substitute $g(z)$ into f.
(2) Multiply numerator and denominator by $c'z + d'$.

(3) Collect powers of z.

and the coefficients are exactly the entries of the matrix product $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)\left(\begin{smallmatrix} a' & b' \\ c' & d' \end{smallmatrix}\right)$. The matrices M and $-M$ give the same map, since the signs cancel in the ratio. So the Möbius group is

$$\mathrm{SL}(2,\mathbb{C})/\{\pm\mathbb{1}\} \;\cong\; \mathrm{SO}^{+}(3,1),$$

of real dimension 6 (4 complex entries minus 1 complex condition). Its elements are generated by translations $\left(\begin{smallmatrix} 1 & b \\ 0 & 1 \end{smallmatrix}\right)$, rotation-dilations $\left(\begin{smallmatrix} \lambda^{1/2} & 0 \\ 0 & \lambda^{-1/2} \end{smallmatrix}\right)$ (giving $z \mapsto \lambda z$), and the inversion $\left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ (giving $z \mapsto -1/z$). The SCT is $\left(\begin{smallmatrix} 1 & 0 \\ -\beta & 1 \end{smallmatrix}\right)$: $z \mapsto z/(1-\beta z)$, which is the complex form of (11.15) with $\beta = b_1 - ib_2$.

The infinitesimal generators as 2×2 matrices. Near the identity, $M = \mathbb{1} + \epsilon X$ with $\operatorname{tr} X = 0$. The map is $z \mapsto z + \epsilon(X_{12} + (X_{11} - X_{22})z - X_{21}z^2)$ to first order. So

$$\ell_{-1} \leftrightarrow \begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix}, \quad \ell_0 \leftrightarrow \begin{pmatrix} -\frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}, \quad \ell_1 \leftrightarrow \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}.$$

(For example ℓ_1 moves z by $-\epsilon z^2$, which is $-X_{21}z^2$ with $X_{21} = 1$.) One checks directly that these matrices obey $[\ell_m, \ell_n] = (m-n)\ell_{m+n}$ with a minus sign overall: the map from vector fields to matrices is an anti-homomorphism, the usual price for letting matrices act on coordinates from the left. For instance $[X_{\ell_1}, X_{\ell_{-1}}] = \left(\begin{smallmatrix} 0 & 0 \\ 1 & 0 \end{smallmatrix}\right)\left(\begin{smallmatrix} 0 & -1 \\ 0 & 0 \end{smallmatrix}\right) - \left(\begin{smallmatrix} 0 & -1 \\ 0 & 0 \end{smallmatrix}\right)\left(\begin{smallmatrix} 0 & 0 \\ 1 & 0 \end{smallmatrix}\right) = \left(\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right) = -2X_{\ell_0}$, while (11.57) gives $[\ell_1, \ell_{-1}] = 2\ell_0$. This is the algebra $\mathfrak{sl}(2,\mathbb{C})$, which returns in Section 11.4.13.

Worked example 11.18: Three points fix a Möbius map

Given three distinct points z_1, z_2, z_3, the map

$$f(z) = \frac{(z-z_1)(z_2-z_3)}{(z-z_3)(z_2-z_1)}$$

is Möbius (after normalising $ad - bc = 1$) and sends $z_1 \mapsto 0$, $z_2 \mapsto 1$, $z_3 \mapsto \infty$: read off the zero, check $f(z_2) = 1$, read off the pole. Since six real parameters can be used to fix three complex points, a Möbius map is completely determined by where it sends three points. With four points $z_1, \dots, z_4$ one invariant survives: the image of z_4,

$$\eta = f(z_4) = \frac{z_{41}z_{23}}{z_{43}z_{21}}, \qquad z_{ij} = z_i - z_j,$$

the **cross-ratio**. This is why conformal symmetry fixes two- and three-point functions completely but leaves four-point functions depending on one variable.

11.3.5 Primary fields

In a conformal field theory, how does a field respond to a holomorphic change of coordinates $z \mapsto z' = f(z)$? The simplest possibility is that it behaves like a tensor with h lower z indices and $\bar{h}$ lower $\bar{z}$ indices: like $\phi_{zz\cdots z\bar{z}\cdots\bar{z}}(\mathrm{d}z)^h(\mathrm{d}\bar{z})^{\bar{h}}$, with the number of indices allowed to be any real number.

Working definition 11.7 (primary field of weight $(h,\bar{h})$)

A field $\phi(z,\bar{z})$ is **primary** with **conformal weights** $(h,\bar{h})$ if under every conformal map $z \mapsto z'(z)$ [a holomorphic change of coordinate, so z' depends on z but not on $\bar{z}$] the

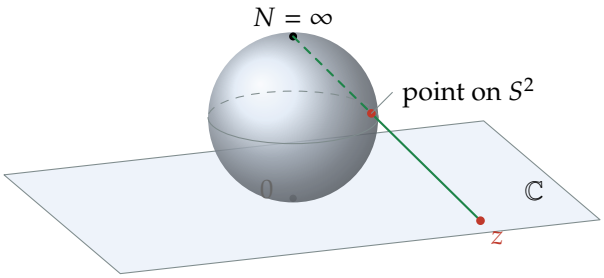


Figure 11.11: The Riemann sphere $\mathbb{C} \cup \{\infty\}$ by stereographic projection: the line from the north pole N through a point of the sphere meets the plane at z. As z runs off to infinity in any direction, the point on the sphere approaches N. On the sphere ∞ is an ordinary point, and a vector field is "globally defined" if it is smooth at N too; that is the test $u = 1/z$ applied to ℓ_n above.

transformed field ϕ' is related to ϕ by

$$\phi(z,\bar z) = \left(\frac{\mathrm{d}z'}{\mathrm{d}z}\right)^{h}\left(\frac{\mathrm{d}\bar z'}{\mathrm{d}\bar z}\right)^{\bar h}\phi'(z',\bar z'), \qquad \text{equivalently} \qquad \phi(z,\bar z)\,\mathrm{d}z^{h}\mathrm{d}\bar z^{\bar h} \text{ is invariant.} \tag{11.60}$$

Here $h, \bar h$ are real, and $\bar h$ is *not* the complex conjugate of h; they are independent labels. *Operational test:* compute the change of ϕ under an infinitesimal $z \mapsto z + \epsilon f(z)$; a primary gives $\delta\phi = \epsilon\big(hf'\phi + f\partial\phi\big)$ plus the barred analogue, for *every* holomorphic f (see (11.63)). If the formula holds only for the Möbius f (degree ≤ 2) [$f = 1, z, z^2$, which generate translations, rotations with dilations, and special conformal maps], the field is called **quasi-primary**.

The two weights package the two quantum numbers that a physicist knows.

Scaling dimension. Take $z' = \lambda z$ with $\lambda > 0$ real. Then $\mathrm{d}z'/\mathrm{d}z = \mathrm{d}\bar z'/\mathrm{d}\bar z = \lambda$ and (11.60) reads $\phi(z,\bar z) = \lambda^{h+\bar h}\phi'(\lambda z, \lambda\bar z)$. Comparing with (11.1), the scaling dimension is

$$\Delta = h + \bar h. \tag{11.61}$$

Spin. Take a rotation $z' = e^{i\theta}z$, $\bar z' = e^{-i\theta}\bar z$. Then $\phi(z,\bar z) = e^{i\theta(h-\bar h)}\phi'(z',\bar z')$. A field that picks up $e^{is\theta}$ under rotation by θ has spin s, so

$$s = h - \bar h. \tag{11.62}$$

Examples: a scalar has $h = \bar h$; a holomorphic current $J(z)$ has $(1,0)$, spin 1; the stress tensor $T(z)$ has $(2,0)$, spin 2 (the two indices of T_{zz}); a chiral fermion has $(\frac12, 0)$, spin $\frac12$.

Nash says that $h - \bar h$ is "therefore restricted to integer or half-integer values". The reason is the rotation by 2π: $e^{2\pi i s}$ must be 1 (bosons) or, allowing a sign, -1 (fermions, which are defined only up to sign). So $s \in \frac12\mathbb{Z}$. We see the same condition from the two-point function below.

Infinitesimal form. Let $z' = z + \epsilon f(z)$, so $\mathrm{d}z'/\mathrm{d}z = 1 + \epsilon f'(z)$. We define the variation of a field as the difference between the old and the new field at the *same* coordinate value, $\delta\phi(z) \coloneqq \phi(z) - \phi'(z)$. Working to first order in ϵ and showing only the holomorphic dependence:

$$\begin{aligned}\phi'(z) &\overset{(1)}{=} \phi'(z') - \epsilon f(z)\,\partial\phi'(z') \overset{(2)}{=} (1+\epsilon f')^{-h}\phi(z) - \epsilon f\,\partial\phi(z)\\ &\overset{(3)}{=} \phi(z) - \epsilon\big(hf'(z)\phi(z) + f(z)\partial\phi(z)\big),\\ \delta\phi(z,\bar z) &\overset{(4)}{=} \epsilon\big(hf'(z)\phi + f(z)\partial\phi\big) + \bar\epsilon\big(\bar h\bar f'(\bar z)\phi + \bar f(\bar z)\bar\partial\phi\big).\end{aligned} \tag{11.63}$$

(1) Taylor expand: $z = z' - \epsilon f(z)$, so $\phi'(z) = \phi'(z' - \epsilon f)$.
(2) The primary law (11.60) gives $\phi'(z') = (1 + \epsilon f')^{-h}\phi(z)$; in the $O(\epsilon)$ term we may replace ϕ' by ϕ.
(3) $(1 + \epsilon f')^{-h} = 1 - h\epsilon f' + O(\epsilon^2)$.
(4) Subtract from $\phi(z)$, and add the identical antiholomorphic calculation (independent parameter $\bar{\epsilon}$ and function $\bar{f}$).

This is Nash's (11.20). The term $f\partial\phi$ is the "transport" of the field, present for any scalar; the term $hf'\phi$ is the response of the h lower indices to the local rotation-and-scaling f'.

Watch out: which $\delta\phi$?

Nash defines $\delta\phi(x)$ as $\phi'(x') - \phi(x)$. With the primary law (11.60) that combination equals $-\epsilon h f'\phi$, without the transport term, and with the opposite sign to his (11.20). What he actually uses in (11.20) and the Ward identity is our $\delta\phi = \phi(z) - \phi'(z)$ (the "active" change of the field at a fixed point). The overall sign is a convention; what matters is that the same $\delta\phi$ appears on both sides of the Ward identity. We check the sign on the free boson in Example 11.26.

11.3.6 Conformal symmetry fixes the two-point function

Suppose ϕ_1, ϕ_2 are primary (quasi-primary is enough) with weights $(h_1, \bar{h}_1)$ and $(h_2, \bar{h}_2)$, and the vacuum is invariant under the global conformal group. Invariance means that correlators of the transformed fields equal correlators of the old fields, so from (11.60)

$$\langle \phi_1(z_1,\bar{z}_1)\phi_2(z_2,\bar{z}_2)\rangle = \prod_{i=1}^{2} f'(z_i)^{h_i}\,\bar{f}'(\bar{z}_i)^{\bar{h}_i}\,\langle \phi_1(f(z_1),\bar{f}(\bar{z}_1))\,\phi_2(f(z_2),\bar{f}(\bar{z}_2))\rangle \tag{11.64}$$

for every Möbius map f. Call the left side $G(z_1, z_2)$ (suppressing $\bar{z}$'s). We use the three kinds of Möbius map one at a time.

Worked example 11.19: Deriving the two-point function

Translations $f(z) = z + a$: $f' = 1$, so $G(z_1, z_2) = G(z_1 + a, z_2 + a)$ and G depends only on $z_{12} = z_1 - z_2$ (and $\bar{z}_{12}$).
Rotation-dilations $f(z) = \lambda z$: $f' = \lambda$, so $G(z_{12}) = \lambda^{h_1+h_2}\bar{\lambda}^{\bar{h}_1+\bar{h}_2}G(\lambda z_{12})$. Choosing $\lambda = 1/z_{12}$ gives

$$G = \frac{C_{12}}{z_{12}^{\,h_1+h_2}\,\bar{z}_{12}^{\,\bar{h}_1+\bar{h}_2}}, \qquad C_{12} = G(1).$$

Inversion $f(z) = -1/z$: $f'(z) = 1/z^2$ and $f(z_1) - f(z_2) = -\frac{1}{z_1} + \frac{1}{z_2} = \frac{z_{12}}{z_1 z_2}$. Look at the holomorphic factors of the right side of (11.64):

$$z_1^{-2h_1} z_2^{-2h_2}\,\frac{C_{12}}{(z_{12}/z_1 z_2)^{h_1+h_2}} \overset{(1)}{=} C_{12}\,\frac{z_1^{h_2-h_1}\,z_2^{h_1-h_2}}{z_{12}^{\,h_1+h_2}}.$$

(1) Collect the powers: $z_1^{-2h_1+h_1+h_2} = z_1^{h_2-h_1}$, similarly for z_2.

This must equal $C_{12}/z_{12}^{h_1+h_2}$ for all z_1, z_2. That is possible only if $h_1 = h_2$ (and likewise $\bar{h}_1 = \bar{h}_2$), or else $C_{12} = 0$. Therefore

$$\langle \phi_1(z_1,\bar{z}_1)\phi_2(z_2,\bar{z}_2)\rangle = \begin{cases} \dfrac{C_{12}}{z_{12}^{2h}\,\bar{z}_{12}^{2\bar{h}}} & h_1 = h_2 = h,\ \bar{h}_1 = \bar{h}_2 = \bar{h}, \\ 0 & \text{otherwise.} \end{cases} \tag{11.65}$$

In polar form $z_{12} = re^{i\theta}$, (11.65) is $C\, r^{-2(h+\bar{h})} e^{-2i\theta(h-\bar{h})}$. The modulus is the power law $r^{-2\Delta}$ of Section 11.1.1. The phase must return to itself (up to the fermionic sign) when $\theta \to \theta + 2\pi$, so $2(h-\bar{h}) \in \mathbb{Z}$: the spin is integer or half-integer, as claimed.

Watch out: the exponent in Nash's (11.9)

Nash prints (11.9) as $\langle \phi(z,\bar{z})\phi(0,0)\rangle = C/z^h\bar{z}^{\bar{h}}$. The exponents should be $2h$ and $2\bar{h}$, as derived above. The check is the free fermion (11.6): $\langle \psi(z)\psi(0)\rangle = C/z$ with $h = \frac{1}{2}$, so the exponent is $1 = 2h$, not h.

Try it: the three-point function

Repeat the argument for three quasi-primaries. Show that translations and rotation-dilations allow $\prod_{i<j} z_{ij}^{-a_{ij}}$ with $a_{12} + a_{13} = 2h_1$ etc. (exponents chosen so each field carries total weight h_i after inversion), and deduce

$$\langle \phi_1\phi_2\phi_3\rangle = \frac{C_{123}}{z_{12}^{h_1+h_2-h_3}\, z_{23}^{h_2+h_3-h_1}\, z_{13}^{h_1+h_3-h_2}} \times (\text{same with bars}).$$

Answer sketch: under inversion each $z_{ij} \mapsto z_{ij}/(z_i z_j)$, so the right side picks up $z_1^{a_{12}+a_{13}}\cdots$, which cancels the $z_1^{-2h_1}$ from $f'(z_1)^{h_1}$ exactly when $a_{12} + a_{13} = 2h_1$. Three linear equations for three unknowns give the exponents shown. The number C_{123} is a genuine dynamical datum of the theory: the OPE coefficient.

11.3.7 Example: the free Majorana fermion

The simplest interacting-looking conformal field theory is actually free: a two-dimensional Majorana (real) fermion. Let us build it from the Dirac equation, with an explicit representation.

Gamma matrices in two Euclidean dimensions. Take $\gamma^1 = \sigma^1$, $\gamma^2 = \sigma^2$ (the Pauli matrices), so $\{\gamma^\mu, \gamma^\nu\} = 2\delta^{\mu\nu}$ (Chapter T4). The Dirac operator is

$$\gamma^\mu\partial_\mu = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\partial_1 + \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}\partial_2 \overset{(1)}{=} \begin{pmatrix} 0 & \partial_1 - i\partial_2 \\ \partial_1 + i\partial_2 & 0 \end{pmatrix} \overset{(2)}{=} 2\begin{pmatrix} 0 & \partial \\ \bar{\partial} & 0 \end{pmatrix}. \tag{11.66}$$

(1) Add the matrices entry by entry.
(2) Definition (11.51) of $\partial, \bar{\partial}$.

Writing the two-component spinor as $\Psi = (\psi, \bar{\psi})^T$, the massless Dirac equation $\gamma^\mu\partial_\mu\Psi = 0$ splits into $2\partial\bar{\psi} = 0$ (first row) and $2\bar{\partial}\psi = 0$ (second row). The upper component is holomorphic, the lower one antiholomorphic. The chirality matrix $\gamma^3 = -i\gamma^1\gamma^2 = -i(i\sigma^3) = \sigma^3$ is diagonal in this basis, so "holomorphic" and "chirality +1" are the same statement. A **Majorana** fermion has real Grassmann components; for it the action is a sum of two decoupled chiral pieces. We look at the holomorphic piece.

The action and its conformal invariance. Nash's (11.5) is $S = \int \mathrm{d}z\mathrm{d}\bar{z}\, \psi\bar{\partial}\psi$ with an unspecified normalisation. We take

$$S_\psi = \frac{1}{2\pi}\int \mathrm{d}^2x\ \psi\,\bar{\partial}\psi. \tag{11.67}$$

Since ψ is Grassmann, $\psi\bar{\partial}\psi$ is not a total derivative (for a commuting field it would be $\frac{1}{2}\bar{\partial}(\psi^2)$). Let us check that S_ψ is invariant under any holomorphic map $z' = f(z)$ if ψ is primary of weight $(\frac{1}{2}, 0)$, that is $\psi'(z') = f'(z)^{-1/2}\psi(z)$:

$$\int \mathrm{d}^2x'\ \psi'\bar{\partial}'\psi' \overset{(1)}{=} \int |f'|^2\mathrm{d}^2x\ f'^{-1/2}\psi\ \bar{f}'^{-1}\bar{\partial}\big(f'^{-1/2}\psi\big)$$
$$\overset{(2)}{=} \int |f'|^2\mathrm{d}^2x\ f'^{-1/2}f'^{-1/2}\bar{f}'^{-1}\ \psi\bar{\partial}\psi \overset{(3)}{=} \int \mathrm{d}^2x\ \psi\,\bar{\partial}\psi.$$

(1) $\mathrm{d}^2x' = |f'|^2\mathrm{d}^2x$ (the Jacobian determinant of (11.55)); $\partial_{\bar{z}'} = (\mathrm{d}\bar{z}'/\mathrm{d}\bar{z})^{-1}\partial_{\bar{z}} = \bar{f}'^{-1}\bar{\partial}$; the transformation of ψ.
(2) f' is holomorphic, so $\bar{\partial}f'^{-1/2} = 0$ and it passes through the derivative.
(3) $|f'|^2 = f'\bar{f}'$ cancels all the factors.

The weight $\frac{1}{2}$ is forced: one $\bar{f}'$ from the measure is eaten by $\bar{\partial}$, and the remaining f' must be shared equally between two ψ's.

Worked example 11.20: The fermion propagator $\langle\psi(z)\psi(w)\rangle = 1/(z-w)$

For a Gaussian Grassmann integral with $S = \frac{1}{2}\int \psi K\psi$ and antisymmetric kernel K, the propagator is K^{-1}. Here (11.67) is $\frac{1}{2}\int \psi\,(\frac{1}{\pi}\bar{\partial})\psi$, so $K = \frac{1}{\pi}\bar{\partial}$ (antisymmetric, since integration by parts flips the sign of $\bar{\partial}$). The propagator $G(z,w) = \langle\psi(z)\psi(w)\rangle$ therefore solves

$$\frac{1}{\pi}\,\bar{\partial}_z G(z,w) = \delta^2(z-w) \quad\Longrightarrow\quad G(z,w) \overset{(1)}{=} \frac{1}{z-w},$$

(1) By (11.54), $\bar{\partial}\frac{1}{z-w} = \pi\delta^2(z-w)$; the solution is unique among functions that vanish at infinity (a bounded entire function is constant, by Liouville).

This is Nash's (11.6) with $C = 1$. Compare with (11.65): $\bar{h} = 0$ and $2h = 1$, so ψ is primary of weight $(\frac{1}{2}, 0)$, exactly the weight that made the action invariant. Consistency check: $1/(z-w)$ is antisymmetric under $z \leftrightarrow w$, as it must be for anticommuting fields (a commuting field would need a symmetric propagator). The half-integer spin shows up in the transformation law: under a rotation $f(z) = e^{i\theta}z$ we have $\psi' = e^{-i\theta/2}\psi$, and at $\theta = 2\pi$ this is $-\psi$. A fermion changes sign under a full turn.

11.3.8 Example: the free boson

Now a real scalar field $\varphi(z,\bar{z})$. Nash's (11.7) is $S = \frac{1}{4\pi}\int \mathrm{d}z\mathrm{d}\bar{z}\,\partial\varphi\bar{\partial}\varphi$. We fix the normalisation as

$$S_\varphi = \frac{1}{4\pi}\int \mathrm{d}^2x\ (\partial_\mu\varphi)^2 = \frac{1}{\pi}\int \mathrm{d}^2x\ \partial\varphi\,\bar{\partial}\varphi, \tag{11.68}$$

using $(\partial_\mu\varphi)^2 = 4\partial\varphi\bar{\partial}\varphi$ (from $\partial_1 = \partial + \bar{\partial}$, $\partial_2 = i(\partial - \bar{\partial})$). This normalisation makes the stress tensor exactly Nash's $T = -(\partial\varphi)^2$ in (11.28); see Example 11.31. The action is conformally invariant with φ of weight $(0,0)$: $\partial\varphi\,\bar{\partial}\varphi\,\mathrm{d}^2x$ picks up $f'^{-1}\bar{f}'^{-1}|f'|^2 = 1$. (In $d = 2$ a scalar is dimensionless: see Section 11.1.1.)

Worked example 11.21: The boson propagator is a logarithm

Integrate by parts: $S_\varphi = \frac{1}{2}\int \mathrm{d}^2x\,\varphi K\varphi$ with $K = -\frac{1}{2\pi}\nabla^2$. The propagator $G = \langle\varphi(x)\varphi(y)\rangle$ solves $KG = \delta^2(x-y)$:

$$-\frac{1}{2\pi}\nabla^2 G = \delta^2(x-y) \quad\Longrightarrow\quad G \overset{(1)}{=} -\ln|x-y| + \text{const} \overset{(2)}{=} -\tfrac{1}{2}\ln|z-w|^2 + \text{const}.$$

(1) Gauss's law in two dimensions: $\nabla^2 \ln r = 2\pi\delta^2(x)$.
(2) $|x-y|^2 = |z-w|^2 = (z-w)(\bar z - \bar w)$.

So

$$\langle\varphi(z,\bar z)\varphi(w,\bar w)\rangle = -\tfrac{1}{2}\ln(z-w) - \tfrac{1}{2}\ln(\bar z - \bar w) + \text{const}, \tag{11.69}$$

which is Nash's (11.8), $C\ln(z\bar z)$, with $C = -\frac{1}{2}$. The additive constant is an infrared ambiguity: the zero mode $\varphi \to \varphi+$const costs no action, so the massless propagator in 2D is defined only up to a constant (it needs an IR cutoff L: $-\ln(|x-y|/L)$).

A logarithm is not a power: $\ln(\lambda z) = \ln z + \ln\lambda$ shifts instead of scaling. So φ is *not* a primary field. As Nash says, this could have been anticipated: φ is dimensionless, so a primary φ would have $h = \bar h = 0$ and a constant two-point function, but the Laplacian needs a nonconstant Green function. The primary fields of the free boson are built from φ in two ways: by differentiating, and by exponentiating.

Worked example 11.22: $\partial\varphi$ is primary of weight $(1,0)$

Differentiate (11.69):

$$\langle\partial\varphi(z)\,\partial\varphi(w)\rangle \overset{(1)}{=} \partial_z\partial_w\left(-\tfrac{1}{2}\ln(z-w)\right) \overset{(2)}{=} \partial_z\left(\frac{1}{2(z-w)}\right) = -\frac{1}{2(z-w)^2}. \tag{11.70}$$

(1) The antiholomorphic and constant terms are killed by ∂_z.
(2) $\partial_w \ln(z-w) = -1/(z-w)$.

By (11.65) this is the two-point function of a field with $2h = 2$, $\bar h = 0$: $\partial\varphi$ has weight $(1,0)$. It is holomorphic by the equation of motion $\partial\bar\partial\varphi = 0$. It is the conserved U(1) current of the shift symmetry $\varphi \to \varphi + a$. The normalised current $J(z) := i\sqrt{2}\,\partial\varphi(z)$ has $\langle J(z)J(w)\rangle = 1/(z-w)^2$.

11.3.9 Free-field toolkit: normal ordering and Wick's theorem

For the free boson and fermion all correlation functions follow from the propagator by Wick's theorem. We will use it constantly, so let us set it up.

Working definition 11.8 (normal ordering and contraction)

For free fields, the **normal-ordered product** $:A_1\cdots A_n:$ of fields is the product with all self-contractions [propagators joining two fields inside the same product] removed; its vacuum expectation value is zero, and it is regular when two of the points coincide. The **contraction** of two fields is their propagator, e.g. the contraction of $\varphi(z)$ with $\varphi(w)$ is $\langle\varphi(z)\varphi(w)\rangle$. *Wick's theorem:* a product of normal-ordered clusters [factors such as $:\varphi(z)\varphi(z):$, each normal-ordered by itself] equals the sum over all ways of contracting pairs of fields from *different* clusters, each term being the product of the contractions times the normal-ordered product of what is left. Fermion contractions carry the sign of the

permutation needed to bring the pair together.

For example, with two clusters of two fields each,

$$:AB:(z)\,:CD:(w) = :ABCD: + \langle AC\rangle :BD: + \langle AD\rangle :BC: + \langle BC\rangle :AD: + \langle BD\rangle :AC: + \langle AC\rangle\langle BD\rangle + \langle AD\rangle\langle BC\rangle,$$

with no terms containing $\langle AB\rangle$ or $\langle CD\rangle$. We use this formula to compute $T(z)T(w)$ in Section 11.3.13.

Worked example 11.23: Vertex operators $e^{i\alpha\varphi}$ are primary with $h = \alpha^2/4$

Let $V_\alpha = :e^{i\alpha\varphi}:$. For a Gaussian variable X with $\langle X\rangle = 0$, $\langle e^X\rangle = e^{\frac{1}{2}\langle X^2\rangle}$ (complete the square). Normal ordering removes the self-contractions $\langle\varphi(z)^2\rangle$, which are infinite. So

$$\langle V_\alpha(z)V_{-\alpha}(w)\rangle \overset{(1)}{=} \exp\Big(\tfrac{1}{2}\big\langle(i\alpha\varphi(z) - i\alpha\varphi(w))^2\big\rangle\Big)\Big|_{\text{no self-contractions}} \overset{(2)}{=} \exp\big(\alpha^2\langle\varphi(z)\varphi(w)\rangle\big) \overset{(3)}{=} \exp\big(-\tfrac{\alpha^2}{2}\ln|z-w|^2\big) = \frac{1}{(z-w)^{\alpha^2/2}(\bar z-\bar w)^{\alpha^2/2}}.$$

(1) $X = i\alpha\varphi(z) - i\alpha\varphi(w)$ is Gaussian with mean zero.
(2) Expand the square, $-\alpha^2(\varphi_z^2 - 2\varphi_z\varphi_w + \varphi_w^2)$; drop the self-contractions φ_z^2, φ_w^2; the cross term gives $\frac{1}{2}\cdot 2\alpha^2\langle\varphi_z\varphi_w\rangle$.
(3) The propagator (11.69), with the constant set to zero.

Compare with (11.65): $2h = 2\bar h = \alpha^2/2$, so

$$V_\alpha = :e^{i\alpha\varphi}: \text{ has weights } (h,\bar h) = \Big(\frac{\alpha^2}{4}, \frac{\alpha^2}{4}\Big). \tag{11.71}$$

Also the charges must balance: $\langle V_\alpha V_\beta\rangle = 0$ unless $\alpha + \beta = 0$, because the shift symmetry $\varphi \to \varphi + a$ multiplies $V_\alpha V_\beta$ by $e^{i(\alpha+\beta)a}$. So a single free boson contains a continuous family of primaries with arbitrary $h \geq 0$: a theory with $c = 1$ and no bound on the number of primaries, unlike the minimal models of Section 11.4.12. We confirm $h = \alpha^2/4$ from the OPE with T in Example 11.28.

Physical picture: Coulomb gas

$\langle\prod_i V_{\alpha_i}(z_i)\rangle = \prod_{i<j}|z_i - z_j|^{\alpha_i\alpha_j}$ (for $\sum\alpha_i = 0$) is the Boltzmann weight $e^{-\beta E}$ of point charges α_i in two dimensions, where the Coulomb energy is logarithmic. The free boson is a two-dimensional electrostatics problem, and the charge neutrality condition is Gauss's law on the compact sphere.

Reading Nash (pp. 302–303, (11.4)–(11.9))

Nash defines primary fields by (11.4), identifies $h+\bar h$ as the scale dimension and $h-\bar h$ as the spin, and gives two examples. The Majorana fermion (11.5)–(11.6) has $\langle\psi(z)\psi(0)\rangle = C/z$ and is primary of weight $(\frac{1}{2},0)$; we derived $C = 1$ from the normalisation (11.67) and checked conformal invariance of the action. The boson (11.7)–(11.8) has a logarithmic propagator and is not primary; we found its primaries $\partial\varphi$ and $:e^{i\alpha\varphi}:$. Nash's (11.9) says scale invariance "alone" constrains the two-point function; in fact scale plus rotation invariance give $C/z^{h_1+h_2}\bar z^{\bar h_1+\bar h_2}$, and one needs the special conformal transformation (inversion) to force $h_1 = h_2$. His exponents $h, \bar h$ should read $2h, 2\bar h$.

11.3.10 The stress tensor in complex coordinates

Every continuous symmetry has a conserved current, and the charges built from the current generate the symmetry on the fields. For translations the current is the stress (energy–momentum) tensor $T_{\mu\nu}$. Conformal transformations are position-dependent translations, $x^\mu \mapsto x^\mu + \epsilon^\mu(x)$, so we expect the stress tensor to generate them too. This subsection sets up $T_{\mu\nu}$ with a definite normalisation, turns it into complex components, and shows that in a conformal theory it collapses to one holomorphic function $T(z)$ and one antiholomorphic function $\bar{T}(\bar{z})$.

The stress tensor as the response to a local translation. Take any field theory that is invariant under translations. Now let the translation depend on position, $\epsilon^\mu = \epsilon^\mu(x)$, with ϵ small and vanishing outside a bounded region. Each field changes by its "active" variation $\delta\phi = \phi(x) - \phi'(x)$ as in Section 11.3.5. For a scalar this is $\delta\phi = \epsilon^\mu \partial_\mu \phi$. If ϵ were constant the action would not change. So for x-dependent ϵ the change of the action can only involve derivatives of ϵ, and to first order it is linear in them.

Working definition 11.9 (stress tensor)

The **stress tensor** $T^{\mu\nu}$ is the coefficient of $\partial_\mu \epsilon_\nu$ in the change of the action under a position-dependent translation:

$$\delta S = \int \mathrm{d}^2x \; T^{\mu\nu}(x)\, \partial_\mu \epsilon_\nu(x) \qquad (\delta\phi = \epsilon^\mu \partial_\mu \phi + \dots). \tag{11.72}$$

Operational recipe: vary the action with $\delta\phi = \epsilon^\mu(x)\partial_\mu\phi$ (plus the spin terms for fields with spin [extra pieces that rotate the components of a vector or spinor field. A scalar has none]), integrate by parts until no derivative acts on anything except ϵ, and read off the coefficient. For theories with a Lorentz-invariant action it can be chosen symmetric, $T^{\mu\nu} = T^{\nu\mu}$. Setting ϵ constant inside a region shows that $\partial_\mu T^{\mu\nu} = 0$ on shell [when the fields obey their equations of motion].

Worked example 11.24: The stress tensor of the free boson

Take $S = \frac{g}{2}\int \mathrm{d}^2x\, \partial_\nu\varphi\, \partial_\nu\varphi$ with $g = \frac{1}{2\pi}$, which is (11.68). With $\delta\varphi = \epsilon^\mu\partial_\mu\varphi$,

$$\begin{aligned}
\delta S &\overset{(1)}{=} g \int \mathrm{d}^2x \; \partial_\nu\varphi\, \partial_\nu\big(\epsilon^\mu \partial_\mu\varphi\big) \\
&\overset{(2)}{=} g \int \mathrm{d}^2x \; \Big[(\partial_\nu\epsilon^\mu)\,\partial_\nu\varphi\,\partial_\mu\varphi + \epsilon^\mu\, \tfrac{1}{2}\partial_\mu\big(\partial_\nu\varphi\,\partial_\nu\varphi\big)\Big] \\
&\overset{(3)}{=} g \int \mathrm{d}^2x \; (\partial_\nu\epsilon^\mu)\Big[\partial_\nu\varphi\,\partial_\mu\varphi - \tfrac{1}{2}\delta_{\mu\nu}(\partial\varphi)^2\Big].
\end{aligned}$$

(1) Vary the two factors; they give equal contributions, cancelling the $\frac{1}{2}$.

(2) Product rule; in the second term $\partial_\nu\varphi\,\partial_\nu\partial_\mu\varphi = \frac{1}{2}\partial_\mu(\partial\varphi)^2$.

(3) Integrate the second term by parts (ϵ vanishes far away): $\int \epsilon^\mu \partial_\mu X = -\int(\partial_\mu\epsilon^\mu)X = -\int(\partial_\nu\epsilon^\mu)\delta_{\mu\nu}X$.

So $T_{\mu\nu} = g\big(\partial_\mu\varphi\,\partial_\nu\varphi - \frac{1}{2}\delta_{\mu\nu}(\partial\varphi)^2\big)$, the tensor of Example 11.8, symmetric and traceless in two dimensions.

Watch out: the sign of δS

In (11.16) we wrote $\delta S = -\int T^{\mu\nu}\partial_\mu \epsilon_\nu$. That was for the passive shift $\phi'(x) = \phi(x-\epsilon)$. Here we use the active variation $\delta\phi = \phi(x) - \phi'(x)$, which flips the sign of ϵ. Both give the same $T^{\mu\nu}$. The conclusion of (11.16), that a traceless T means invariance under all conformal maps, is insensitive to the sign.

Tracelessness. In a conformal theory $\delta S = 0$ for every conformal ϵ. By (11.16) this means $\int T^\mu{}_\mu\, \partial\cdot\epsilon = 0$ for all conformal ϵ. In two dimensions $\partial\cdot\epsilon$ for a holomorphic ϵ is $2\,\mathrm{Re}\,\epsilon'(z)$, which can be any harmonic function; that is enough freedom to conclude that $T^\mu{}_\mu$ vanishes (up to contact terms at the positions of other operators). So

$$T_{11} + T_{22} = 0, \tag{11.73}$$

which is Nash's (11.11). A symmetric traceless 2×2 tensor has only two independent components, for example $T_{11} = -T_{22}$ and T_{12}.

Complex components. As in Section 11.3.1, T_{ab} with $a, b \in \{z, \bar z\}$ is obtained by the chain rule with $\partial x_1/\partial z = \frac12$, $\partial x_2/\partial z = \frac{1}{2i} = -\frac{i}{2}$, $\partial x_1/\partial\bar z = \frac12$, $\partial x_2/\partial\bar z = \frac{i}{2}$:

$$\begin{aligned} T_{zz} &\overset{(1)}{=} \left(\tfrac12\right)^2 T_{11} + 2\cdot\tfrac12\cdot\left(-\tfrac{i}{2}\right)T_{12} + \left(-\tfrac{i}{2}\right)^2 T_{22} = \tfrac14\big(T_{11} - T_{22} - 2iT_{12}\big), \\ T_{\bar z\bar z} &\overset{(2)}{=} \tfrac14\big(T_{11} - T_{22} + 2iT_{12}\big), \\ T_{z\bar z} = T_{\bar z z} &\overset{(3)}{=} \tfrac12\cdot\tfrac12\, T_{11} + \tfrac12\cdot\tfrac{i}{2}\, T_{12} - \tfrac{i}{2}\cdot\tfrac12\, T_{21} - \tfrac{i}{2}\cdot\tfrac{i}{2}\, T_{22} = \tfrac14\big(T_{11} + T_{22}\big) = 0. \end{aligned} \tag{11.74}$$

(1) $T_{zz} = \frac{\partial x^\mu}{\partial z}\frac{\partial x^\nu}{\partial z} T_{\mu\nu}$; the two cross terms T_{12} and T_{21} are equal.
(2) Same computation with $\partial x_2/\partial\bar z = +\frac{i}{2}$ (complex conjugate coefficients).
(3) One index z, one index $\bar z$; the T_{12} terms cancel, $-\frac{i}{2}\cdot\frac{i}{2} = +\frac14$, and we use (11.73).

Watch out: a factor $\frac14$ in Nash's (11.12)

Nash writes $T_{zz} = T_{11} - T_{22} - 2iT_{12}$ and $T_{z\bar z} = T_{11} + T_{22}$, without the $\frac14$. This only rescales T, and Nash never uses the absolute normalisation of T_{zz}: he fixes the normalisation of $T(z)$ later, through the OPE (11.25). But if you check his formulas against a Lagrangian, you need the $\frac14$.

Conservation makes T_{zz} holomorphic. Raise the index with $g^{z\bar z} = g^{\bar z z} = 2$ and $g^{zz} = g^{\bar z\bar z} = 0$ (Section 11.3.1). The z component of the conservation law $\partial^\mu T_{\mu\nu} = 0$ reads

$$0 = g^{\mu\rho}\partial_\rho T_{\mu z} \overset{(1)}{=} g^{z\bar z}\partial_{\bar z} T_{zz} + g^{\bar z z}\partial_z T_{\bar z z} \overset{(2)}{=} 2\,\bar\partial T_{zz}, \tag{11.75}$$

(1) Only the off-diagonal inverse metric components are nonzero.
(2) $T_{\bar z z} = 0$ by tracelessness (11.74).

and similarly the $\bar z$ component gives $\partial T_{\bar z\bar z} = 0$. This is Nash's (11.13). So T_{zz} depends on z only and $T_{\bar z\bar z}$ on $\bar z$ only (away from other operator insertions).

Working definition 11.10 (the holomorphic stress tensor)

Define

$$T(z) := -2\pi\, T_{zz}(z), \qquad \bar{T}(\bar{z}) := -2\pi\, T_{\bar{z}\bar{z}}(\bar{z}). \tag{11.76}$$

Nash's (11.14) sets $T = T_{zz}$; the factor -2π is the standard choice that makes the Ward identity (11.83) [a contour integral of $f(w)T(w)$ around the fields gives their change under $z \mapsto z + \epsilon f(z)$] and the OPE (11.87) come out without stray constants. It is what Nash's formulas implicitly assume: with it, the free boson has $T = -(\partial\varphi)^2$, exactly his (11.28). *Operational test:* $T(z)$ is correctly normalised if its OPE with any primary field [a field for which $\phi\, \mathrm{d}z^h \mathrm{d}\bar{z}^{\bar{h}}$ is invariant] has the leading term $h\phi/(w-z)^2$ with the weight h read off from the two-point function.

Worked example 11.25: $T(z)$ for the free boson and the free fermion

Boson. From Example 11.24,

$$T_{zz} \overset{(1)}{=} \tfrac{g}{4}\left[(\partial_1\varphi)^2 - (\partial_2\varphi)^2 - 2i\,\partial_1\varphi\,\partial_2\varphi\right] \overset{(2)}{=} \tfrac{g}{4}\left(\partial_1\varphi - i\partial_2\varphi\right)^2 \overset{(3)}{=} g\,(\partial\varphi)^2,$$

(1) (11.74); the $\delta_{\mu\nu}$ term drops out of $T_{11} - T_{22}$ and does not contribute to T_{12}.
(2) The bracket is a perfect square.
(3) $\partial_1 - i\partial_2 = 2\partial$ by (11.51).

With $g = \frac{1}{2\pi}$ we get $T(z) = -2\pi\cdot\frac{1}{2\pi}(\partial\varphi)^2 = -(\partial\varphi)^2$, and in the quantum theory $T = -{:}\partial\varphi\,\partial\varphi{:}$. The classical equation of motion $\partial\bar{\partial}\varphi = 0$ shows directly that $\bar{\partial}T = -2\,\partial\varphi\,\partial\bar{\partial}\varphi = 0$.

Fermion. For $S_\psi = \frac{1}{2\pi}\int \mathrm{d}^2x\,\psi\bar{\partial}\psi$ it is quickest to vary with a purely holomorphic-index translation $\epsilon^z(z,\bar{z})$, $\epsilon^{\bar{z}} = 0$, under which the weight-$\frac{1}{2}$ field changes by $\delta\psi = \epsilon^z\partial\psi + \frac{1}{2}(\partial\epsilon^z)\psi$. The general formula (11.72) with a symmetric traceless T becomes, for such ϵ (lower the index: $\epsilon_{\bar{z}} = g_{\bar{z}z}\epsilon^z = \frac{1}{2}\epsilon^z$, and $T^{\bar{z}\bar{z}} = 4T_{zz}$),

$$\delta S = \int \mathrm{d}^2x\; T^{\bar{z}\bar{z}}\,\partial_{\bar{z}}\epsilon_{\bar{z}} = 2\int \mathrm{d}^2x\; T_{zz}\,\bar{\partial}\epsilon^z = -\frac{1}{\pi}\int \mathrm{d}^2x\; T(z)\,\bar{\partial}\epsilon^z. \tag{11.77}$$

Now compute δS_ψ directly:

$$\begin{aligned}
\delta S_\psi &\overset{(1)}{=} \frac{1}{\pi}\int \mathrm{d}^2x\;\delta\psi\,\bar{\partial}\psi \overset{(2)}{=} \frac{1}{\pi}\int \mathrm{d}^2x\,\Big(\epsilon\,\partial\psi + \tfrac{1}{2}(\partial\epsilon)\psi\Big)\bar{\partial}\psi \\
&\overset{(3)}{=} \frac{1}{\pi}\int \mathrm{d}^2x\;\epsilon\Big(\tfrac{1}{2}\partial\psi\,\bar{\partial}\psi - \tfrac{1}{2}\psi\,\partial\bar{\partial}\psi\Big) \overset{(4)}{=} \frac{1}{\pi}\int \mathrm{d}^2x\;\epsilon\Big(-\tfrac{1}{2}\bar{\partial}(\psi\,\partial\psi)\Big) \\
&\overset{(5)}{=} \frac{1}{2\pi}\int \mathrm{d}^2x\;(\bar{\partial}\epsilon)\,\psi\,\partial\psi.
\end{aligned}$$

(1) $\delta(\psi\bar{\partial}\psi) = \delta\psi\,\bar{\partial}\psi + \psi\,\bar{\partial}\delta\psi$; integrating the second term by parts and anticommuting the two Grassmann factors gives $+\delta\psi\,\bar{\partial}\psi$ again.
(2) Insert $\delta\psi$ (we write ϵ for ϵ^z).
(3) Integrate the $\partial\epsilon$ term by parts: $\frac{1}{2}(\partial\epsilon)\psi\bar{\partial}\psi \to -\frac{1}{2}\epsilon(\partial\psi\,\bar{\partial}\psi + \psi\,\partial\bar{\partial}\psi)$; combine with the first term.
(4) $\bar{\partial}(\psi\partial\psi) = \bar{\partial}\psi\,\partial\psi + \psi\,\partial\bar{\partial}\psi = -\partial\psi\,\bar{\partial}\psi + \psi\,\partial\bar{\partial}\psi$ (Grassmann sign), so the bracket is $-\frac{1}{2}\bar{\partial}(\psi\partial\psi)$.
(5) Integrate by parts once more.

Comparing with (11.77): $-\frac{1}{\pi}T = \frac{1}{2\pi}\psi\partial\psi$, so

$$T_\psi(z) = -\tfrac{1}{2}\,{:}\psi\,\partial\psi{:}(z), \qquad T_\varphi(z) = -{:}\partial\varphi\,\partial\varphi{:}(z). \tag{11.78}$$

(The same shortcut, applied to the boson with $\delta\varphi = \epsilon^z\partial\varphi$, gives $\delta S_\varphi = \frac{1}{\pi}\int(\bar\partial\epsilon)(\partial\varphi)^2$, hence again $T = -(\partial\varphi)^2$.)

11.3.11 The conformal Ward identity

We now derive the statement that T generates conformal transformations. The tool is the simplest fact about path integrals: renaming the integration variable does not change the integral.

Ward identities from a change of variables. Let $X = \phi_1(x_1)\cdots\phi_n(x_n)$, with path-integral average

$$\langle X\rangle = \frac{1}{Z}\int \mathcal{D}\phi\, X\, e^{-S[\phi]}.$$

Substitute $\phi = \tilde\phi + \delta\tilde\phi$, where δ is a transformation with a position-dependent parameter, and assume the measure is invariant ($\mathcal{D}\phi = \mathcal{D}\tilde\phi$; for conformal transformations this is where the central charge will eventually hide, but the Ward identity for T is unaffected). Then

$$\int \mathcal{D}\phi\; X[\phi]\, e^{-S[\phi]} \overset{(1)}{=} \int \mathcal{D}\tilde\phi\; X[\tilde\phi+\delta\tilde\phi]\, e^{-S[\tilde\phi+\delta\tilde\phi]} \overset{(2)}{=} \int \mathcal{D}\tilde\phi\; \big(X+\delta X\big)\big(1-\delta S\big)e^{-S}.$$

(1) Change of integration variable; the Jacobian is 1 by assumption.
(2) Expand to first order in the transformation.

The left side is the same as the zeroth-order term on the right, so the first-order terms must cancel:

$$\langle\delta X\rangle = \langle\delta S\; X\rangle, \qquad \delta X = \sum_{i=1}^{n}\phi_1(x_1)\cdots\delta\phi_i(x_i)\cdots\phi_n(x_n). \tag{11.79}$$

With (11.72) this is

$$\sum_i\langle\phi_1(x_1)\cdots\delta\phi_i(x_i)\cdots\phi_n(x_n)\rangle = \int \mathrm{d}^2x\;\partial_\mu\epsilon_\nu(x)\,\langle T^{\mu\nu}(x)\,\phi_1(x_1)\cdots\phi_n(x_n)\rangle. \tag{11.80}$$

Cutting off the transformation: Nash's (11.16). Now choose the transformation to be a genuine conformal map inside a region R that contains all the points x_i, and to be switched off outside: $\epsilon^\nu(x) = \epsilon f^\nu(x)\,\theta_R(x)$, where θ_R is 1 on R and 0 outside. Then $\partial_\mu\epsilon_\nu = \epsilon\,\theta_R\,\partial_\mu f_\nu + \epsilon f_\nu\,\partial_\mu\theta_R$, and $\partial_\mu\theta_R = -n_\mu\,\delta_{\partial R}$ is a delta function on the boundary with the outward normal n_μ. So (11.80) splits into a bulk and a boundary term:

$$\sum_i\langle\cdots\delta\phi_i\cdots\rangle = \epsilon\int_R \mathrm{d}^2x\;\partial^\mu f^\nu\,\langle T_{\mu\nu}\,X\rangle - \epsilon\oint_{\partial R}\mathrm{d}l\;n^\mu f^\nu\,\langle T_{\mu\nu}\,X\rangle. \tag{11.81}$$

Nash writes the boundary term as $\int_{\partial R}\mathrm{d}l_\mu\,\epsilon^{\mu\nu}\langle T_{\nu\sigma}X\rangle f^\sigma$ with the permutation symbol: for a boundary traversed counterclockwise, the outward normal times $\mathrm{d}l$ is the tangent rotated by $90°$, $n_\mu\mathrm{d}l = \epsilon_{\mu\nu}\mathrm{d}x^\nu$. Up to the sign, which depends on orientation conventions, this is (11.16).

If f is conformal, the bulk term vanishes: T is symmetric, so only the symmetric part of $\partial^\mu f^\nu$ contributes; by the conformal Killing equation (11.10) that part is $\frac{1}{2}(\partial\cdot f)\delta^{\mu\nu}$; and T is traceless. Only the boundary term survives. Nash says exactly this after (11.17).

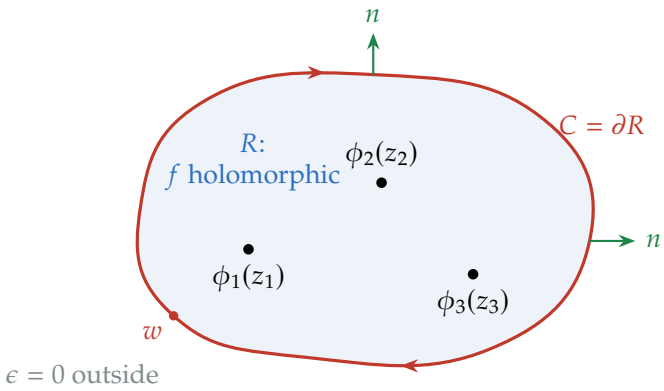


Figure 11.12: The geometry of the Ward identity. Inside R we perform the conformal transformation $z \mapsto z + \epsilon f(z)$; outside R we do nothing. The change of the fields at the insertion points (black) is balanced by the "flux" of the stress tensor through the edge C where the transformation is switched off. Since $T(w)$ is holomorphic away from the insertions, the contour C can be moved freely as long as it does not cross an insertion.

The boundary term in complex form: Nash's (11.17)–(11.18). Rather than convert $n^\mu f^\nu T_{\mu\nu}$ by brute force, go back to (11.77), which (together with its antiholomorphic partner) is (11.72) for a symmetric traceless T: $\delta S = -\frac{1}{\pi}\int \mathrm{d}^2x \left(T\,\bar\partial\epsilon^z + \bar T\,\partial\epsilon^{\bar z}\right)$. Take $\epsilon^z = \epsilon f(z)\theta_R$ with f holomorphic on and near R. We need one formula: for a function g that is smooth near ∂R,

$$\int \mathrm{d}^2x\; g\,\bar\partial\theta_R = -\frac{1}{2i}\oint_{\partial R} g\,\mathrm{d}z, \qquad \int \mathrm{d}^2x\; g\,\partial\theta_R = +\frac{1}{2i}\oint_{\partial R} g\,\mathrm{d}\bar z. \tag{11.82}$$

This follows by moving the derivative onto g (minus sign) and applying the complex Stokes formula $\int_R \mathrm{d}^2x\,\bar\partial F = \frac{1}{2i}\oint_{\partial R} F\mathrm{d}z$ of Example 11.14; for the second, $\mathrm{d}(F\mathrm{d}\bar z) = \partial F\,\mathrm{d}z\wedge\mathrm{d}\bar z = -2i\,\partial F\,\mathrm{d}^2x$ gives $\int_R \mathrm{d}^2x\,\partial F = -\frac{1}{2i}\oint F\mathrm{d}\bar z$. So, for the holomorphic half,

$$\delta S \overset{(1)}{=} -\frac{\epsilon}{\pi}\int \mathrm{d}^2x\; T(w)\,f(w)\,\bar\partial\theta_R \overset{(2)}{=} -\frac{\epsilon}{\pi}\left(-\frac{1}{2i}\right)\oint_{\partial R}\mathrm{d}w\; f(w)\,T(w) = \frac{\epsilon}{2\pi i}\oint_C \mathrm{d}w\; f(w)\,T(w),$$

(1) $\bar\partial(f\theta_R) = f\,\bar\partial\theta_R$ because f is holomorphic.
(2) (11.82) with $g = fT$; T is smooth near the boundary because no operators sit there.

and the antiholomorphic half gives $-\frac{\epsilon}{2\pi i}\oint_C \mathrm{d}\bar w\,\bar f(\bar w)\bar T(\bar w)$, with $C = \partial R$ counterclockwise. Inserting into (11.79):

$$\boxed{\sum_i \langle\cdots\delta\phi_i(z_i,\bar z_i)\cdots\rangle = \frac{\epsilon}{2\pi i}\oint_C \mathrm{d}w\, f(w)\,\langle T(w)\,X\rangle - \frac{\epsilon}{2\pi i}\oint_C \mathrm{d}\bar w\,\bar f(\bar w)\,\langle\bar T(\bar w)\,X\rangle} \tag{11.83}$$

This is Nash's (11.18).

Watch out: the orientation of the $\mathrm{d}\bar w$ integral

Nash writes both terms of (11.18) with a plus sign. The two versions agree: for an antiholomorphic integrand the counterclockwise integral $\oint \mathrm{d}\bar w/(\bar w - \bar z)$ equals $-2\pi i$ (take the complex conjugate of $\oint \mathrm{d}w/(w-z) = 2\pi i$). So "$-\oint_{\text{ccw}}\mathrm{d}\bar w$" is the same as "$+\oint$ with the residue in $\bar w$ counted positively", which is what Nash means. In practice: *evaluate the $\bar w$ contour by taking the residue in $\bar w$ with a plus sign*, exactly as for w.

From correlators to an operator statement: Nash's (11.19). The integrand $\langle T(w)X\rangle$ is holomorphic in w except at the insertion points z_i. By Cauchy's theorem the contour C around all the points can be shrunk to a sum of small circles C_i, one around each z_i (Figure 11.13). The variation on the left is also a sum over i. Since the points are arbitrary, the equality must hold term by term: the small circle around z_i gives $\delta\phi_i$. Dropping the other fields, which are spectators, we get an *operator* identity for the holomorphic part $\delta^h\phi$ of the variation:

$$\delta^h\phi(z,\bar z) = \frac{\epsilon}{2\pi i}\oint_z \mathrm{d}w\; f(w)\,T(w)\,\phi(z,\bar z), \tag{11.84}$$

the contour now being a small circle around z. This is Nash's (11.19). It is the precise meaning of "T generates conformal transformations": the conserved charge $Q_f = \frac{1}{2\pi i}\oint f(w)T(w)\mathrm{d}w$ acts on a field by circling it.

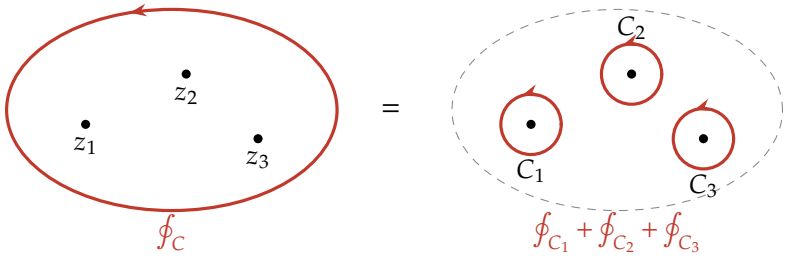


Figure 11.13: Contour deformation. The integrand $f(w)\langle T(w)X\rangle$ is holomorphic except at the insertion points. The big contour (left) can be shrunk until it hugs each point separately (right); in between, the integrand has no singularities, so nothing changes. Each small circle C_i computes the variation of the one field it encloses, and this is how the global identity (11.83) becomes the local operator statement (11.84).

11.3.12 The operator product of T with a primary field

We now have two expressions for the change of a primary field under $z \mapsto z + \epsilon f(z)$. One is the transformation law (11.63), $\delta^h\phi = \epsilon(hf'\phi + f\partial\phi)$ (Nash's (11.20)). The other is the contour integral (11.84). Equating them (Nash's (11.21)):

$$h\,f'(z)\,\phi(z,\bar z) + f(z)\,\partial\phi(z,\bar z) = \frac{1}{2\pi i}\oint_z \mathrm{d}w\; f(w)\,T(w)\,\phi(z,\bar z) \qquad \text{for every holomorphic } f. \tag{11.85}$$

Since f is arbitrary, this determines the singular part of the product $T(w)\phi(z)$. Nash's way of extracting it is to expand everything in powers of $w - z$.

The expansions (Nash's (11.22)). Expand f about z and the product $T(w)\phi(z)$ in a Laurent series about $w = z$, with operator-valued coefficients:

$$f(w) = \sum_{k\ge 0} \hat f_k(z)\,(w-z)^k, \qquad \hat f_k(z) = \frac{f^{(k)}(z)}{k!},$$

$$T(w)\phi(z,\bar z) = \sum_{n\in\mathbb{Z}} \frac{T_n(\phi)(z,\bar z)}{(w-z)^{n+2}} = \cdots + \frac{T_2(\phi)}{(w-z)^4} + \frac{T_1(\phi)}{(w-z)^3} + \frac{T_0(\phi)}{(w-z)^2} + \frac{T_{-1}(\phi)}{w-z} + T_{-2}(\phi) + \cdots$$

We label the coefficient of $(w-z)^{-n-2}$ by $T_n(\phi)$; this is the labelling of the Virasoro modes in §3, where $T_n(\phi)$ becomes $L_n\phi$.

Watch out: Nash's labels in (11.22)

Nash's (11.22) writes the coefficients of $(w-z)^{-2}$ and $(w-z)^{-1}$ as $T_{-2}(\phi)$ and $T_{-1}(\phi)$, and in (11.24) concludes $T_{-1}(\phi)=\partial\phi$, $T_{-2}(\phi)=h\phi$, $T_{-3}=T_{-4}=\cdots=0$. That is a different (and not quite consistent) labelling: in his (11.23) the index runs so that T_{-k-1} multiplies $\hat{f}_k$. We keep the §3 labelling throughout, in which the same results read $T_{-1}(\phi)=\partial\phi$, $T_0(\phi)=h\phi$ and $T_n(\phi)=0$ for $n\geq 1$. The content is identical.

Matching coefficients (Nash's (11.23)–(11.24)). Insert the expansions into the right side of (11.85). The residue theorem, $\frac{1}{2\pi i}\oint_z \mathrm{d}w\,(w-z)^m=\delta_{m,-1}$, picks out the terms where the total power of $(w-z)$ is -1:

$$\frac{1}{2\pi i}\oint_z \mathrm{d}w\, f(w)T(w)\phi(z) \overset{(1)}{=} \sum_{k\geq 0}\sum_{n} \hat{f}_k(z)\,T_n(\phi)\,\frac{1}{2\pi i}\oint_z \mathrm{d}w\,(w-z)^{k-n-2}$$
$$\overset{(2)}{=} \sum_{k\geq 0}\hat{f}_k(z)\,T_{k-1}(\phi)=\hat{f}_0\,T_{-1}(\phi)+\hat{f}_1\,T_0(\phi)+\hat{f}_2\,T_1(\phi)+\hat{f}_3\,T_2(\phi)+\cdots$$

(1) Multiply the two series and integrate term by term.
(2) The integral is 1 exactly when $k-n-2=-1$, i.e. $n=k-1$.

The left side of (11.85) is $\hat{f}_0\,\partial\phi+h\,\hat{f}_1\,\phi$, since $\hat{f}_0=f(z)$ and $\hat{f}_1=f'(z)$. At a fixed point z the numbers $\hat{f}_0,\hat{f}_1,\hat{f}_2,\ldots$ can be chosen independently (take $f(w)=(w-z)^k$ for one k at a time). Comparing coefficients:

$$T_{-1}(\phi)=\partial\phi,\qquad T_0(\phi)=h\,\phi,\qquad T_1(\phi)=T_2(\phi)=\cdots=0. \tag{11.86}$$

Nash does the matching at $z=0$, where $\hat{f}_k(0)=f_k$ are the Taylor coefficients of f; the result is the same at any z.

Key idea: the $T\phi$ OPE is the conformal Ward identity

For a primary field ϕ of weights $(h,\bar{h})$,

$$\begin{aligned} T(w)\,\phi(z,\bar{z}) &= \frac{h\,\phi(z,\bar{z})}{(w-z)^2}+\frac{\partial\phi(z,\bar{z})}{w-z}+\text{regular},\\ \bar{T}(\bar{w})\,\phi(z,\bar{z}) &= \frac{\bar{h}\,\phi(z,\bar{z})}{(\bar{w}-\bar{z})^2}+\frac{\bar{\partial}\phi(z,\bar{z})}{\bar{w}-\bar{z}}+\text{regular}. \end{aligned} \tag{11.87}$$

These are Nash's (11.25)–(11.27). The double pole carries the weight, the simple pole carries the derivative (the translation), and nothing more singular appears. Conversely, if a field has this OPE with T, then it is primary. The OPE is shorthand: the symbol $\sim$ ("equal up to regular terms") is often used, and all OPEs are understood inside correlation functions with other fields placed away from z.

The regular term $R(w-z)$ of Nash's (11.25)–(11.26) contains $T_{-2}(\phi)$, $T_{-3}(\phi)$, and so on. These are new fields: the descendants of §3. They are not fixed by the Ward identity, which only involves the singular part.

Worked example 11.26: Checking the Ward normalisation on the free boson

The boson φ has weight 0 but is not primary, so the check is to compute $T(w)\varphi(z)$ with $T=-{:}\partial\varphi\,\partial\varphi{:}$ and compare with $\delta\varphi=\epsilon f\partial\varphi$. Wick's theorem (Section 11.3.9) needs the

contraction

$$\langle \partial\varphi(w)\,\varphi(z)\rangle = \partial_w\big(-\tfrac{1}{2}\ln(w-z)\big) = -\frac{1}{2(w-z)}.$$

There is one field $\varphi(z)$ to contract, with either of the two $\partial\varphi(w)$:

$$T(w)\varphi(z) \overset{(1)}{=} -2\,\langle\partial\varphi(w)\varphi(z)\rangle\,\partial\varphi(w) + :\cdots: \overset{(2)}{=} \frac{\partial\varphi(w)}{w-z} + \text{reg.} \overset{(3)}{=} \frac{\partial\varphi(z)}{w-z} + \text{reg.}$$

(1) Two equivalent contractions; the leftover normal-ordered term is regular.
(2) Insert the contraction: $-2\cdot(-\frac{1}{2}) = 1$.
(3) Taylor expand $\partial\varphi(w) = \partial\varphi(z) + (w-z)\partial^2\varphi(z) + \cdots$; only the first term is singular.

There is no double pole, consistent with $h = 0$. Then (11.84) gives $\delta\varphi = \frac{\epsilon}{2\pi i}\oint \mathrm{d}w\, f(w)\frac{\partial\varphi(z)}{w-z} = \epsilon f(z)\partial\varphi(z)$, exactly the change of a scalar. The factor -2π in (11.76) is doing its job: with Nash's literal $T = T_{zz}$ we would have found $\delta\varphi = -\frac{1}{2\pi}\,\epsilon f\partial\varphi$.
Why is φ not primary, then? Because φ is not a well-defined local operator in the first place (its two-point function has the infrared constant), and the higher Ward identities with other fields detect this. The fields built from it, $\partial\varphi$ and $:e^{i\alpha\varphi}:$, are primary, as we now check.

Worked example 11.27: $T(w)\,\partial\varphi(z)$: a weight-one primary

Now the contraction is $\langle\partial\varphi(w)\partial\varphi(z)\rangle = -\frac{1}{2(w-z)^2}$, from (11.70).

$$T(w)\,\partial\varphi(z) \overset{(1)}{=} -2\,\langle\partial\varphi(w)\partial\varphi(z)\rangle\,\partial\varphi(w) + \text{reg.} \overset{(2)}{=} \frac{\partial\varphi(w)}{(w-z)^2} + \text{reg.}$$
$$\overset{(3)}{=} \frac{\partial\varphi(z)}{(w-z)^2} + \frac{\partial^2\varphi(z)}{w-z} + \text{reg.}$$

(1) Two ways to contract, as before.
(2) $-2\cdot(-\frac{1}{2}) = 1$.
(3) Taylor expand $\partial\varphi(w)$ to first order; the second-order term is regular.

This is (11.87) with $h = 1$ and $\phi = \partial\varphi$, confirming the weight $(1,0)$ found from the two-point function in Example 11.22.

Worked example 11.28: $T(w)\,V_\alpha(z)$: the vertex operator has $h = \alpha^2/4$

Let $V_\alpha = :e^{i\alpha\varphi}:$. Contract $\partial\varphi(w)$ with $\varphi(z)$ in the exponent: each contraction brings down a factor $i\alpha\langle\partial\varphi(w)\varphi(z)\rangle = -\frac{i\alpha}{2(w-z)}$ (think of $e^{i\alpha\varphi}$ as its power series; contracting one of the n factors of φ in $\varphi^n/n!$ leaves $\varphi^{n-1}/(n-1)!$, which resums to the exponential). T contains two $\partial\varphi$'s, so there are single and double contractions:

$$T(w)V_\alpha(z) \overset{(1)}{=} -\Big(-\frac{i\alpha}{2(w-z)}\Big)^2 V_\alpha(z) - 2\Big(-\frac{i\alpha}{2(w-z)}\Big):\partial\varphi(w)V_\alpha(z):$$
$$\overset{(2)}{=} \frac{\alpha^2/4}{(w-z)^2}\,V_\alpha(z) + \frac{i\alpha\,:\partial\varphi\,V_\alpha:(z)}{w-z} + \text{reg.} \overset{(3)}{=} \frac{\alpha^2/4}{(w-z)^2}\,V_\alpha(z) + \frac{\partial V_\alpha(z)}{w-z} + \text{reg.}$$

(1) Double contraction (both $\partial\varphi(w)$'s with the exponent) and single contraction (one of the two).
(2) $-(-i\alpha/2)^2 = -(-\alpha^2/4) = \alpha^2/4$ and $-2(-i\alpha/2) = i\alpha$; replace $\partial\varphi(w)$ by $\partial\varphi(z)$ in the singular term.
(3) $\partial V_\alpha = :i\alpha\,\partial\varphi\,e^{i\alpha\varphi}:$ by the chain rule.

So V_α is primary with $h = \alpha^2/4$ (and $\bar{h} = \alpha^2/4$ from $\bar{T}$), agreeing with the two-point function (11.71). Two independent methods, one answer: the normalisation of T is right.

Worked example 11.29: $T(w)\,\psi(z)$: the fermion has $h = \frac{1}{2}$

With $T = -\frac{1}{2}{:}\psi\partial\psi{:}$ and $\langle\psi(w)\psi(z)\rangle = \frac{1}{w-z}$, the two contractions of $\psi(z)$ with the factors of $T(w)$ are $\langle\psi(w)\psi(z)\rangle = \frac{1}{w-z}$ and $\langle\partial\psi(w)\psi(z)\rangle = -\frac{1}{(w-z)^2}$. Bringing $\psi(z)$ next to the field it contracts with costs a sign when it passes an odd number of fermions:

$$T(w)\psi(z) \overset{(1)}{=} -\tfrac{1}{2}\Big[-\langle\psi(w)\psi(z)\rangle\,\partial\psi(w) + \psi(w)\,\langle\partial\psi(w)\psi(z)\rangle\Big]$$
$$\overset{(2)}{=} \frac{\partial\psi(w)}{2(w-z)} + \frac{\psi(w)}{2(w-z)^2} \overset{(3)}{=} \frac{\frac{1}{2}\psi(z)}{(w-z)^2} + \frac{\partial\psi(z)}{w-z} + \text{reg.}$$

(1) In $\psi(w)\partial\psi(w)\psi(z)$, contracting $\psi(w)$ with $\psi(z)$ requires moving $\psi(z)$ past $\partial\psi(w)$: one sign. Contracting $\partial\psi(w)$ with $\psi(z)$ needs no move.
(2) Insert the contractions: $-\frac{1}{2}(-1) = \frac{1}{2}$ and $-\frac{1}{2}\cdot(-1) = \frac{1}{2}$.
(3) Expand $\psi(w) = \psi(z) + (w-z)\partial\psi(z) + \cdots$ in the double pole: it contributes $\frac{1}{2}\psi/(w-z)^2 + \frac{1}{2}\partial\psi/(w-z)$; add the other $\frac{1}{2}\partial\psi/(w-z)$.

So ψ is primary with $h = \frac{1}{2}$, as the propagator said (Example 11.20).

The Ward identity for correlation functions. Combining (11.83) with the OPE gives the most useful form of the Ward identity. The function $w \mapsto \langle T(w)\phi_1(z_1)\cdots\phi_n(z_n)\rangle$ is meromorphic, with poles only at the z_i given by the OPE, and (as we show in Section 11.4.13) it decays like w^{-4} at infinity. A meromorphic function on the sphere is fixed by its poles, so

$$\langle T(w)\,\phi_1(z_1)\cdots\phi_n(z_n)\rangle = \sum_{i=1}^{n}\Big(\frac{h_i}{(w-z_i)^2} + \frac{1}{w-z_i}\,\frac{\partial}{\partial z_i}\Big)\langle\phi_1(z_1)\cdots\phi_n(z_n)\rangle. \tag{11.88}$$

Worked example 11.30: The Ward identity on a two-point function

Take $G = \langle\phi(z_1)\phi(z_2)\rangle = z_{12}^{-2h}$ (suppressing the $\bar{z}$ dependence). Then $\partial_1 G = -2hG/z_{12}$ and $\partial_2 G = +2hG/z_{12}$. Write $a = w - z_1$, $b = w - z_2$, so $b - a = z_{12}$.

$$\langle T(w)\phi(z_1)\phi(z_2)\rangle \overset{(1)}{=} hG\Big[\frac{1}{a^2} + \frac{1}{b^2} - \frac{2}{z_{12}\,a} + \frac{2}{z_{12}\,b}\Big] \overset{(2)}{=} hG\Big[\frac{1}{a^2} + \frac{1}{b^2} - \frac{2}{ab}\Big]$$
$$\overset{(3)}{=} hG\Big(\frac{b-a}{ab}\Big)^2 = \frac{h\,z_{12}^{2-2h}}{(w-z_1)^2(w-z_2)^2}.$$

(1) (11.88) with the derivatives just computed.
(2) $\frac{1}{b} - \frac{1}{a} = \frac{a-b}{ab} = -\frac{z_{12}}{ab}$, so the last two terms are $\frac{2}{z_{12}}\cdot\big(-\frac{z_{12}}{ab}\big) = -\frac{2}{ab}$.
(3) A perfect square.

Three checks: the result has double poles at z_1 and z_2 with residue structure dictated by the OPE; it falls off like w^{-4} as $w \to \infty$ (the w^{-2} and w^{-3} terms have cancelled, which is precisely the statement that G is invariant under L_{-1}, L_0, L_1); and it vanishes if $h = 0$. For the free fermion ($h = \frac{1}{2}$, $G = 1/z_{12}$) one can verify it independently by Wick's theorem.

11.3.13 The operator product $T(w)T(z)$ and the central charge

The stress tensor is itself a field of the theory. Its holomorphic part $T(z)$ has, by the scaling argument, weight $(2,0)$: it is a component of a tensor with two lower z indices. Is it primary? Let us compute its OPE with itself in the free boson.

Worked example 11.31: The TT OPE of the free boson: $c = 1$

Let $T = -{:}\partial\varphi\,\partial\varphi{:}$ and write $u = w - z$ and $\langle\partial\varphi(w)\partial\varphi(z)\rangle = -\frac{1}{2u^2}$. Apply the four-field Wick formula of Section 11.3.9 with $A = B = \partial\varphi(w)$ and $C = D = \partial\varphi(z)$:

$$
\begin{aligned}
T(w)T(z) &\overset{(1)}{=} (-1)^2\Big[\,2\Big(-\frac{1}{2u^2}\Big)^2 + 4\Big(-\frac{1}{2u^2}\Big){:}\partial\varphi(w)\partial\varphi(z){:} + {:}\partial\varphi(w)^2\partial\varphi(z)^2{:}\Big] \\
&\overset{(2)}{=} \frac{1}{2u^4} - \frac{2}{u^2}\,{:}\partial\varphi(w)\partial\varphi(z){:} + \text{reg.} \\
&\overset{(3)}{=} \frac{1}{2u^4} - \frac{2}{u^2}\,{:}\partial\varphi\,\partial\varphi{:}(z) - \frac{2}{u}\,{:}\partial^2\varphi\,\partial\varphi{:}(z) + \text{reg.} \\
&\overset{(4)}{=} \frac{1/2}{(w-z)^4} + \frac{2\,T(z)}{(w-z)^2} + \frac{\partial T(z)}{w-z} + \text{reg.}
\end{aligned}
$$

(1) Two full contractions ($AC\,BD$ and $AD\,BC$) and four single contractions (AC, AD, BC, BD), all equal because $A = B$ and $C = D$.

(2) $2 \cdot \frac{1}{4u^4} = \frac{1}{2u^4}$ and $4 \cdot (-\frac{1}{2}) = -2$.

(3) Taylor expand $\partial\varphi(w) = \partial\varphi(z) + u\,\partial^2\varphi(z) + \frac{1}{2}u^2\partial^3\varphi(z) + \cdots$ inside the normal ordering; terms of order u^0 and higher are regular.

(4) ${-}{:}\partial\varphi\partial\varphi{:} = T$, and $\partial T = -2{:}\partial^2\varphi\,\partial\varphi{:}$.

This is Nash's (11.28), with the roles of w and z as he has them. The first term is a c-number times the identity operator I: it comes from the double contraction and is there in any theory with a T.

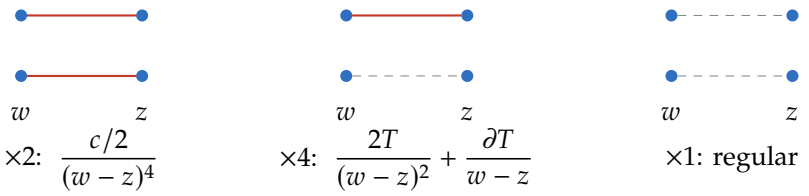


Figure 11.14: The Wick contractions in $T(w)T(z)$ for a free field. Each T is a pair of fields (blue dots); a red line is a propagator. Two propagators (left) leave no operator: a pure number, the central term. One propagator (middle) leaves one field at w and one at z, whose Taylor expansion gives the T and ∂T terms. No propagator (right) is the regular normal-ordered product. The central charge c is literally the sum of "vacuum bubbles" between the two T's, so for N independent free bosons it is N.

Working definition 11.11 (central charge)

In any conformal field theory the TT OPE has the form (Nash's (11.29))

$$
T(w)\,T(z) = \frac{c/2}{(w-z)^4} + \frac{2\,T(z)}{(w-z)^2} + \frac{\partial T(z)}{w-z} + \text{regular}. \tag{11.89}
$$

Here "regular" [terms that stay finite as $w \to z$]. The number c is the **central charge** [one number fixed by the theory, $c = 1$ for a free boson and $c = \frac{1}{2}$ for a free fermion]. *Operational*

recipe: compute the two-point function of T; by (11.89) $\langle T(w)T(z)\rangle = \frac{c/2}{(w-z)^4}$, so c is twice the coefficient of $(w-z)^{-4}$.

Why this form? The terms $2T/(w-z)^2 + \partial T/(w-z)$ are forced: they say that T has weight 2 and that T generates translations of itself, exactly as for any field in (11.87). Scaling allows one more singular term, a field of weight 0 at $(w-z)^{-4}$; the only weight-0 field in a unitary theory is the identity. A term at $(w-z)^{-3}$ would need a weight-1 field, and it is forbidden by the symmetry of the OPE under $w \leftrightarrow z$ (a $(w-z)^{-3}$ term is odd). So the only free datum is c.

Worked example 11.32: The free fermion has $c = \frac{1}{2}$

Nash's footnote 2 invites us to check $c = \frac{1}{2}$. Let $T = -\frac{1}{2}{:}\psi\partial\psi{:}$ and $u = w - z$. Name the four fields $A = \psi(w)$, $B = \partial\psi(w)$, $C = \psi(z)$, $D = \partial\psi(z)$. The contractions are

$$\langle AC\rangle = \frac{1}{u}, \qquad \langle AD\rangle = \partial_z \frac{1}{w-z} = \frac{1}{u^2},$$
$$\langle BC\rangle = \partial_w \frac{1}{w-z} = -\frac{1}{u^2}, \qquad \langle BD\rangle = \partial_w\partial_z \frac{1}{w-z} = -\frac{2}{u^3}.$$

Full contractions. To contract A with C and B with D in $ABCD$ we must swap B and C: one sign. To contract A with D and B with C we move D past C and B: two signs.

$$\tfrac{1}{4}\big[-\langle AC\rangle\langle BD\rangle + \langle AD\rangle\langle BC\rangle\big] \overset{(1)}{=} \tfrac{1}{4}\Big[\frac{2}{u^4} - \frac{1}{u^4}\Big] = \frac{1/4}{u^4}.$$

(1) $-\frac{1}{u}\cdot(-\frac{2}{u^3}) = \frac{2}{u^4}$ and $\frac{1}{u^2}\cdot(-\frac{1}{u^2}) = -\frac{1}{u^4}$; the prefactor is $(-\frac{1}{2})^2 = \frac{1}{4}$.

Comparing with $\frac{c/2}{u^4}$ gives $c = \frac{1}{2}$.

Single contractions. With the same sign rule, and writing ψ, ψ', ψ'' for $\psi(z)$ and its derivatives,

$$\begin{aligned}\tfrac{1}{4}\big[&-\langle AC\rangle{:}BD{:} + \langle AD\rangle{:}BC{:} + \langle BC\rangle{:}AD{:} - \langle BD\rangle{:}AC{:}\big]\\ &\overset{(1)}{=} \tfrac{1}{4}\Big[\frac{{:}\psi'\psi{:}}{u^2} + \frac{{:}\psi''\psi{:}}{u} - \frac{{:}\psi\psi'{:}}{u^2} + \frac{2{:}\psi'\psi{:}}{u^2} + \frac{{:}\psi''\psi{:}}{u}\Big] + \text{reg.}\\ &\overset{(2)}{=} \tfrac{1}{4}\Big[-\frac{4\,{:}\psi\psi'{:}}{u^2} - \frac{2\,{:}\psi\psi''{:}}{u}\Big] \overset{(3)}{=} \frac{2T(z)}{u^2} + \frac{\partial T(z)}{u}.\end{aligned}$$

(1) Expand the fields at w about z: $A = \psi + u\psi' + \cdots$, $B = \psi' + u\psi'' + \cdots$. The AC term is regular (it has only $1/u$ times ${:}\psi'\psi'{:} = 0$). The AD term gives $\frac{1}{u^2}{:}(\psi' + u\psi'')\psi{:}$. The BC term gives $-\frac{1}{u^2}{:}(\psi + u\psi')\psi'{:}$, whose $1/u$ part vanishes since ${:}\psi'\psi'{:} = 0$. The BD term gives $\frac{2}{u^3}{:}(\psi + u\psi' + \frac{1}{2}u^2\psi'')\psi{:}$, and ${:}\psi\psi{:} = 0$.

(2) Collect, using ${:}\psi'\psi{:} = -{:}\psi\psi'{:}$ and ${:}\psi''\psi{:} = -{:}\psi\psi''{:}$.

(3) $T = -\frac{1}{2}{:}\psi\psi'{:}$, so $2T = -{:}\psi\psi'{:}$; and $\partial T = -\frac{1}{2}({:}\psi'\psi'{:} + {:}\psi\psi''{:}) = -\frac{1}{2}{:}\psi\psi''{:}$.

So the fermion satisfies (11.89) with $c = \frac{1}{2}$. As Nash remarks, the value of c distinguishes theories: a single boson has $c = 1$, a single Majorana fermion $c = \frac{1}{2}$, and N decoupled copies of anything have N times the central charge (the T's add and the cross terms do not contract).

Physical picture: what c measures

The coefficient c counts degrees of freedom: a real boson counts 1, a Majorana fermion $\frac{1}{2}$ (half of a Dirac fermion, which is equivalent to a boson in two dimensions and has $c = 1$).

It also measures how the theory responds to the geometry. On a curved surface the trace of T no longer vanishes but becomes $\langle T^{\mu}{}_{\mu}\rangle = -\frac{c}{12}R$ (the **trace anomaly**, in Polchinski's conventions), and on a cylinder of circumference L the ground-state energy is shifted by $-\pi c/(6L)$, the Casimir energy we compute in Example 11.34. In string theory (Chapter 9) the requirement that the total c of matter plus ghosts vanish gives the critical dimension 26.

11.3.14 How T transforms: the Schwarzian derivative

The central term means T is not quite a primary field. Let us see exactly how it transforms.

Infinitesimal transformation (Nash's (11.30)). Apply the operator Ward identity (11.84) to $\phi = T$ and use (11.89):

$$\begin{aligned}\delta T(z) &\overset{(1)}{=} \frac{\epsilon}{2\pi i}\oint_z dw\, f(w)\Big[\frac{c/2}{(w-z)^4}+\frac{2T(z)}{(w-z)^2}+\frac{\partial T(z)}{w-z}\Big]\\ &\overset{(2)}{=} \epsilon\Big[\frac{c}{2}\frac{f'''(z)}{3!}+2f'(z)\,T(z)+f(z)\,\partial T(z)\Big] = \epsilon\Big[2f'T+f\,\partial T+\frac{c}{12}f'''\Big].\end{aligned} \tag{11.90}$$

(1) Operator Ward identity with the TT OPE.

(2) Cauchy's formula $\frac{1}{2\pi i}\oint_z \frac{f(w)}{(w-z)^{k+1}}dw = \frac{f^{(k)}(z)}{k!}$ for $k = 3, 1, 0$.

Without the last term this is the law (11.63) for a primary of weight 2. The extra term is a c-number, independent of the fields. It vanishes if and only if $f''' = 0$, that is, for $f = a + bz + cz^2$: the three global generators $\ell_{-1}, \ell_0, \ell_1$. So T is a **quasi-primary** field of weight 2: primary under Möbius maps only.

Finite transformation (Nash's (11.31)). The finite version is

$$T(z) = \Big(\frac{dz'}{dz}\Big)^2 T'(z') + \frac{c}{12}\{z', z\}, \qquad \{z', z\} := \frac{z'''}{z'} - \frac{3}{2}\Big(\frac{z''}{z'}\Big)^2, \tag{11.91}$$

where primes on z' mean derivatives with respect to z. The combination $\{z', z\}$ is the **Schwarzian derivative**.

Worked example 11.33: The finite law reproduces the infinitesimal one

Take $z' = z + \epsilon f(z)$. Then $dz'/dz = 1 + \epsilon f'$, $z'' = \epsilon f''$, $z''' = \epsilon f'''$, and

$$\{z', z\} = \frac{\epsilon f'''}{1+\epsilon f'} - \frac{3}{2}\Big(\frac{\epsilon f''}{1+\epsilon f'}\Big)^2 = \epsilon f''' + O(\epsilon^2).$$

Solve (11.91) for the new field and evaluate it at the old coordinate value:

$$\begin{aligned}T'(z) &\overset{(1)}{=} T'(z') - \epsilon f\,\partial T'(z') \overset{(2)}{=} (1+\epsilon f')^{-2}\Big[T(z) - \frac{c}{12}\epsilon f'''\Big] - \epsilon f\,\partial T(z)\\ &\overset{(3)}{=} T(z) - \epsilon\Big[2f'T + f\,\partial T + \frac{c}{12}f'''\Big] + O(\epsilon^2).\end{aligned}$$

(1) Taylor expand, since $z = z' - \epsilon f$.

(2) Solve (11.91) for $T'(z')$; in the $O(\epsilon)$ term replace T' by T.

(3) $(1+\epsilon f')^{-2} = 1 - 2\epsilon f' + \cdots$.

So $\delta T = T(z) - T'(z) = \epsilon[2f'T + f\partial T + \frac{c}{12}f''']$, which is (11.90).

Why the Schwarzian? The composition law. Transformation laws must compose: doing $z \to u \to w$ in two steps must give the same answer as doing $z \to w$ at once. For the tensorial part this is the chain rule, $(\mathrm{d}w/\mathrm{d}z)^2 = (\mathrm{d}w/\mathrm{d}u)^2(\mathrm{d}u/\mathrm{d}z)^2$. The extra term must therefore satisfy

$$\{w,z\} = \left(\frac{\mathrm{d}u}{\mathrm{d}z}\right)^2 \{w,u\} + \{u,z\}. \tag{11.92}$$

Let us verify that the Schwarzian does this. Write $S[w] = \{w,z\}$. The quickest route is the identity $\{w,z\} = \big(\ln w'\big)'' - \frac{1}{2}\big((\ln w')'\big)^2$, which follows from $(\ln w')' = w''/w'$ and $(\ln w')'' = w'''/w' - (w''/w')^2$. For $w(u(z))$ we have $\ln w'(z) = \ln w_u(u) + \ln u'(z)$, where $w_u = \mathrm{d}w/\mathrm{d}u$. Let $A = (\ln w_u)_u$ (a function of u) and $B = (\ln u')'$. Then

$$\begin{aligned}
(\ln w')' &\overset{(1)}{=} u'A + B, \qquad (\ln w')'' \overset{(2)}{=} u'^2 A_u + u''A + B', \\
\{w,z\} &\overset{(3)}{=} u'^2 A_u + u''A + B' - \tfrac{1}{2}\big(u'^2A^2 + 2u'AB + B^2\big) \\
&\overset{(4)}{=} u'^2\big(A_u - \tfrac{1}{2}A^2\big) + \big(B' - \tfrac{1}{2}B^2\big) + A\big(u'' - u'B\big) \overset{(5)}{=} u'^2\{w,u\} + \{u,z\}.
\end{aligned}$$

(1) Chain rule: $\frac{\mathrm{d}}{\mathrm{d}z}\ln w_u(u(z)) = u'A$.
(2) Differentiate again; $\frac{\mathrm{d}}{\mathrm{d}z}A = u'A_u$.
(3) Insert into the identity for the Schwarzian.
(4) Regroup.
(5) $u'B = u' \cdot u''/u' = u''$, so the last bracket vanishes; the other two brackets are the Schwarzians by the same identity.

So the Schwarzian is exactly the extra term that composes correctly. (In the language of Chapter 6, it is a one-cocycle on the group of diffeomorphisms; its infinitesimal version f''' is the Gelfand–Fuks cocycle that defines the Virasoro central extension.)

The Schwarzian vanishes exactly on Möbius maps (Nash's (11.32)). *Möbius* $\Rightarrow \{f,z\} = 0$. For $f = \frac{az+b}{cz+d}$ with $ad - bc = 1$:

$$f' = \frac{1}{(cz+d)^2}, \quad f'' = \frac{-2c}{(cz+d)^3}, \quad f''' = \frac{6c^2}{(cz+d)^4}, \qquad \{f,z\} \overset{(1)}{=} \frac{6c^2}{(cz+d)^2} - \frac{3}{2}\cdot\frac{4c^2}{(cz+d)^2} = 0.$$

(1) $f'''/f' = 6c^2/(cz+d)^2$ and $(f''/f')^2 = 4c^2/(cz+d)^2$.

$\{f,z\} = 0 \Rightarrow$ *Möbius.* This is Nash's chain (11.32), with the steps filled in. Let $g = f'$ (assume g is not constant; the constant case $f = az + b$ is Möbius with $c = 0$). Then

$$\begin{aligned}
0 = \{f,z\} &\overset{(1)}{=} \frac{g''}{g} - \frac{3}{2}\frac{g'^2}{g^2} \quad\Longrightarrow\quad \frac{g''}{g'} \overset{(2)}{=} \frac{3}{2}\frac{g'}{g} \quad\Longrightarrow\quad \ln g' \overset{(3)}{=} \tfrac{3}{2}\ln g + A \\
&\Longrightarrow\quad g^{-3/2}\,g' \overset{(4)}{=} e^A \quad\Longrightarrow\quad -2g^{-1/2} \overset{(5)}{=} e^A z + B' \quad\Longrightarrow\quad g \overset{(6)}{=} \frac{4}{(e^A z + B')^2}, \\
f &\overset{(7)}{=} -\frac{4e^{-A}}{e^A z + B'} + D \overset{(8)}{=} \frac{De^A z + (DB' - 4e^{-A})}{e^A z + B'}.
\end{aligned}$$

(1) $f'''/f' = g''/g$ and $f''/f' = g'/g$.
(2) Multiply by g/g'.
(3) Both sides are logarithmic derivatives; integrate, with constant A.
(4) Exponentiate.
(5) $\frac{\mathrm{d}}{\mathrm{d}z}(-2g^{-1/2}) = g^{-3/2}g'$; integrate, constant B'.
(6) Solve for g.

(7) Integrate g, constant D.
(8) Common denominator.

This is $(az+b)/(cz+d)$ with $a = De^{A}$, $b = DB' - 4e^{-A}$, $c = e^{A}$, $d = B'$, and $ad - bc = De^{A}B' - (DB' - 4e^{-A})e^{A} = 4$. Dividing a, b, c, d by 2 normalises $ad - bc = 1$. The three integration constants A, B', D of the third-order equation are the three complex parameters of $\mathrm{SL}(2,\mathbb{C})/\{\pm 1\}$.

Watch out: the constant in Nash's (11.32)

Nash's last line is $f = C\,\frac{az+b}{cz+d}$ with "C an undetermined constant". The undetermined constant actually enters additively (our D), and it is absorbed into a and b. A multiplicative constant is also harmless, since $z \mapsto Cz$ is itself Möbius. Either way the conclusion is the same: the solutions of $\{f, z\} = 0$ are exactly the Möbius maps.

Worked example 11.34: From the plane to the cylinder: the Casimir energy

Use the map of Example 11.15(b), $z = e^{w}$ with $w = \tau + i\sigma$ on a cylinder of circumference 2π. We want the stress tensor $T_{\mathrm{cyl}}(w)$ in the w coordinate. Apply (11.91) with the "old" coordinate w and the "new" one z:

$$T_{\mathrm{cyl}}(w) \overset{(1)}{=} \left(\frac{\mathrm{d}z}{\mathrm{d}w}\right)^{2} T(z) + \frac{c}{12}\{z, w\} \overset{(2)}{=} z^{2}\,T(z) + \frac{c}{12}\left(1 - \frac{3}{2}\right) = z^{2}\,T(z) - \frac{c}{24}.$$

(1) (11.91), with T_{cyl} the field in the w frame and T the field in the z frame.
(2) $\mathrm{d}z/\mathrm{d}w = z'' = z''' = e^{w} = z$, so $z'''/z' = 1$ and $(z''/z')^{2} = 1$.

On the plane the vacuum has $\langle T(z)\rangle = 0$ (by rotation and translation invariance: $\langle T(z)\rangle$ would have to be a constant of weight 2). On the cylinder, therefore, $\langle T_{\mathrm{cyl}}\rangle = -c/24$. The energy density on the cylinder is proportional to $T_{\mathrm{cyl}} + \bar{T}_{\mathrm{cyl}}$, and we will see in Section 11.4.3 that the Hamiltonian is $H = L_0 + \bar{L}_0 - \frac{c}{12}$. Restoring a circumference L multiplies energies by $2\pi/L$: the vacuum energy is $E_0 = -\frac{\pi c}{6L}$. This is the two-dimensional Casimir effect, and measuring E_0 in a numerical simulation of a lattice model is a standard way to determine c.

Reading Nash (pp. 303–307, (11.10)–(11.32))

Nash's aim here is the OPE of T with a primary field. He introduces the stress tensor, uses tracelessness and conservation to reduce it to $T(z)$ and $\bar{T}(\bar{z})$ ((11.10)–(11.14); his (11.12) lacks a factor $\frac{1}{4}$). He quotes the Ward identity (11.16) with a bulk and a boundary term, specialises to conformal f so only the boundary term survives, and obtains the contour form (11.18) and the operator equation (11.19). We derived all of these from a change of variables in the path integral and fixed the normalisation $T = -2\pi T_{zz}$, which Nash leaves implicit. Equating (11.19) with the transformation law (11.20) of a primary gives (11.21); matching the Laurent expansions (11.22)–(11.23) gives (11.24) and hence the OPEs (11.25)–(11.27). Nash then states the free-boson TT OPE (11.28) ("it is easy to calculate"; we computed it, and the fermion's too), the general form (11.29) with the central charge (footnote 2: $c = 1$ for the boson, $c = \frac{1}{2}$ for the fermion), the transformation (11.30) of T, the finite law (11.31) with the Schwarzian derivative, and the proof (11.32) that the Schwarzian vanishes only for Möbius maps. We added the stress tensors of the free fields, the Ward identity for correlators and its check on a two-point function, the explicit OPEs of T with $\partial\varphi$, $:e^{i\alpha\varphi}:$ and ψ, the composition law of the Schwarzian and the cylinder map.

Recap: Section 11.3

- In complex coordinates $z = x_1 + ix_2$, conformal maps are holomorphic maps. Locally they form the Witt algebra $[\ell_m, \ell_n] = (m-n)\ell_{m+n}$; globally on the sphere only $\ell_{-1}, \ell_0, \ell_1$ survive, generating the Möbius group $\mathrm{SL}(2,\mathbb{C})/\{\pm 1\}$.
- A primary field of weights $(h, \bar{h})$ transforms as $\phi\, \mathrm{d}z^h \mathrm{d}\bar{z}^{\bar{h}}$; $\Delta = h + \bar{h}$, $s = h - \bar{h}$. Its two-point function is $C/z_{12}^{2h}\bar{z}_{12}^{2\bar{h}}$.
- Free fermion: $\langle\psi\psi\rangle = 1/z$, $h = \frac{1}{2}$. Free boson: $\langle\varphi\varphi\rangle = -\frac{1}{2}\ln|z|^2$, not primary; $\partial\varphi$ has $h = 1$, $:e^{i\alpha\varphi}:$ has $h = \bar{h} = \alpha^2/4$.
- $T(z) = -2\pi T_{zz}$ is holomorphic; it generates conformal maps by $\delta\phi = \frac{\epsilon}{2\pi i}\oint f(w)T(w)\phi(z)\mathrm{d}w$.
- Primary $\Leftrightarrow T(w)\phi(z) \sim \frac{h\phi}{(w-z)^2} + \frac{\partial\phi}{w-z}$.
- $T(w)T(z) \sim \frac{c/2}{(w-z)^4} + \frac{2T}{(w-z)^2} + \frac{\partial T}{w-z}$; $c = 1$ (boson), $\frac{1}{2}$ (fermion).
- T transforms with the Schwarzian: $T = (z'_z)^2 T' + \frac{c}{12}\{z', z\}$, which vanishes only for Möbius maps; on the cylinder $\langle T_{\mathrm{cyl}}\rangle = -c/24$.

11.4 Relation to the Virasoro algebra

In Section 11.3 everything was phrased with correlation functions and contour integrals. A physicist also wants the Hamiltonian picture: a Hilbert space of states, operators acting on it, and the commutation relations of the symmetry generators. Two questions organise this section.

(a) The conserved charges built from $T(z)$ generate conformal maps. What algebra do they obey? The classical answer is the Witt algebra (11.57). The quantum answer is the **Virasoro algebra**, and the extra term in it is exactly the central charge c of the TT OPE.

(b) How are the states of a conformal field theory organised? The answer is that every primary field ϕ produces a state $|h\rangle$, the Virasoro generators with negative index build a tower of states on top of it, and demanding that all these states have positive norm puts strong restrictions on c and h.

The second question is the representation theory of the Virasoro algebra, which we met from the group side in Chapter 6. Here we see where the representations come from physically.

11.4.1 Two copies of $\mathrm{Diff}(S^1)$

Nash begins §3 by recalling that the two-dimensional conformal group "consists of the group of holomorphic and anti-holomorphic transformations" and that it can be realised as $\mathrm{Diff}(S^1) \times \mathrm{Diff}(S^1)$. Let us see what this means concretely, and where the statement needs a caveat.

Vector fields on the circle. A diffeomorphism of the circle close to the identity is $\theta \mapsto \theta + \epsilon v(\theta)$ with v a smooth real 2π-periodic function. Its generator is the vector field $v(\theta)\partial_\theta$. Expand v in

a Fourier series. A convenient complex basis of vector fields on the circle is

$$\ell_n\big|_{S^1} = i\,e^{in\theta}\,\partial_\theta, \qquad n \in \mathbb{Z}. \tag{11.93}$$

The name is deliberate: it is the restriction of the Witt generator $\ell_n = -z^{n+1}\partial_z$ of (11.56) to the unit circle $z = e^{i\theta}$.

$$\ell_n = -z^{n+1}\partial_z \overset{(1)}{=} -z^n\,(z\,\partial_z) \overset{(2)}{=} -e^{in\theta}\big(-i\,\partial_\theta\big) = i\,e^{in\theta}\partial_\theta.$$

(1) Split off one factor of z.
(2) On functions of $z = e^{i\theta}$ we have $\partial_\theta = \frac{\mathrm{d}z}{\mathrm{d}\theta}\partial_z = iz\,\partial_z$, so $z\partial_z = -i\partial_\theta$.

So the commutator $[\ell_m, \ell_n] = (m-n)\ell_{m+n}$ holds on the circle as well (you can check it directly with ∂_θ, exactly as in Example 11.16). The *real* vector fields are the real combinations

$$-\partial_\theta = i\ell_0, \qquad \cos n\theta\,\partial_\theta = -\tfrac{i}{2}\big(\ell_n + \ell_{-n}\big), \qquad \sin n\theta\,\partial_\theta = -\tfrac{1}{2}\big(\ell_n - \ell_{-n}\big).$$

So the Lie algebra $\mathrm{Vect}(S^1)$ of $\mathrm{Diff}(S^1)$ is a real form of the Witt algebra: the Witt algebra is its complexification. Figure 11.15 shows three of these vector fields.

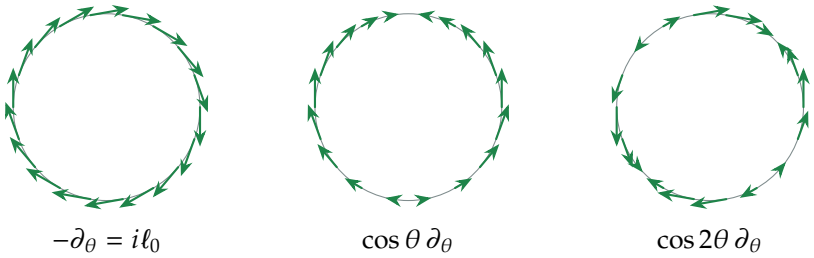


Figure 11.15: Three real vector fields on the circle. Left: rigid rotation (all arrows equal). Middle: $\cos\theta\,\partial_\theta$ moves points counterclockwise on the right half of the circle and clockwise on the left half, so the flow runs from the repelling zero at $\theta = 3\pi/2$ (bottom) to the attracting zero at $\theta = \pi/2$ (top). Right: $\cos 2\theta\,\partial_\theta$ has four zeros; it is a genuinely "local" deformation of the circle. The first two (together with $\sin\theta\,\partial_\theta$) generate the Möbius maps that preserve the unit disc; the $n \geq 2$ fields are the infinitely many extra directions of $\mathrm{Diff}(S^1)$.

Why boundary values? A holomorphic map defined on a neighbourhood of the unit circle (an annulus) is determined by its values on the circle, by Cauchy's formula. Restricting the holomorphic vector field $\epsilon(z)\partial_z$ to the circle therefore loses nothing. The antiholomorphic vector fields $\bar\epsilon(\bar z)\partial_{\bar z}$ give a second, independent copy. This is the sense of Nash's "$\mathrm{Diff}(S^1) \times \mathrm{Diff}(S^1)$: one S^1 corresponds to the boundary values of the holomorphic transformations and the other to the boundary values of the anti-holomorphic transformations."

Watch out: what "$\mathrm{Diff}(S^1) \times \mathrm{Diff}(S^1)$" really means

The statement is true at the level of Lie algebras after complexification:

$$\big(\text{local conformal algebra}\big) = \mathrm{Witt} \oplus \overline{\mathrm{Witt}} = \mathrm{Vect}(S^1)_{\mathbb{C}} \oplus \mathrm{Vect}(S^1)_{\mathbb{C}}.$$

It is *not* literally true at the level of groups, for two reasons. First, the group $\mathrm{Diff}(S^1)$ has no complexification: there is no Lie group whose Lie algebra is $\mathrm{Vect}(S^1)_{\mathbb{C}}$ (see Chapter 6). Second, in Euclidean signature z and $\bar z$ are complex conjugates, so a single real conformal map is one holomorphic function, not a pair. The two copies become genuinely independent only after continuing to Minkowski signature, where $z, \bar z \to x^0 \pm x^1$ are the two real light-cone coordinates and each light-cone circle carries its own $\mathrm{Diff}(S^1)$. For the quantum

theory, which only uses the Lie algebra and its representations, the statement is exactly right.

11.4.2 The modes of $T(z)$

The stress tensor $T(z)$ is holomorphic away from operator insertions. Put the only insertions at 0 and ∞ (we will see in a moment that this is the natural setting). Then T is holomorphic on the punctured plane and has a Laurent expansion. Since T has weight 2, it is natural to shift the powers by 2 so that the label n counts how much the mode changes the scaling dimension:

$$T(z) = \sum_{n\in\mathbb{Z}} L_n\, z^{-n-2}, \qquad L_n = \oint_0 \frac{\mathrm{d}w}{2\pi i}\, w^{n+1}\, T(w). \tag{11.94}$$

This is Nash's (11.33). The inversion formula follows from the residue theorem:

$$\oint_0 \frac{\mathrm{d}w}{2\pi i}\, w^{n+1} T(w) \overset{(1)}{=} \sum_{m\in\mathbb{Z}} L_m \oint_0 \frac{\mathrm{d}w}{2\pi i}\, w^{n+1-m-2} \overset{(2)}{=} \sum_{m\in\mathbb{Z}} L_m\, \delta_{n-m-1,-1} = L_n .$$

(1) Insert the Laurent series and integrate term by term (the series converges uniformly on the contour).

(2) $\oint_0 \frac{\mathrm{d}w}{2\pi i} w^k = \delta_{k,-1}$: only $k=-1$ has a residue.

The same formulas hold for $\bar T$ (Nash's (11.35)):

$$\bar T(\bar z) = \sum_{n\in\mathbb{Z}} \bar L_n\, \bar z^{-n-2}, \qquad \bar L_n = \oint_0 \frac{\mathrm{d}\bar w}{2\pi i}\, \bar w^{n+1}\, \bar T(\bar w), \tag{11.95}$$

with the $\bar w$ residue counted positively, as in the pitfall after (11.83).

The modes are the conformal charges. The operator Ward identity (11.84) says that the transformation $z \mapsto z + \epsilon f(z)$ is generated by $Q_f = \oint \frac{\mathrm{d}w}{2\pi i} f(w)T(w)$. For $f(w) = w^{n+1}$, the transformation $z \mapsto z + \epsilon z^{n+1}$ is the flow of $-\ell_n$, and the charge is exactly L_n. So

$$Q_f = \sum_n f_n\, L_{n-1} \quad \text{for} \quad f(w) = \sum_n f_n w^n, \tag{11.96}$$

and each mode L_n is the quantum generator of one of the Witt vector fields. (The sign relating L_n to ℓ_n is a convention about active versus passive maps and never matters below; what matters is that the L_n realise the algebra of the ℓ_n, up to a central term.)

Dictionary: which mode does what

mode	$f(w)$	transformation
L_{-1}	1	translation $z \mapsto z+\epsilon$
L_0	w	dilation and rotation $z \mapsto (1+\epsilon)z$
L_1	w^2	special conformal $z \mapsto z + \epsilon z^2$
L_n, $\lvert n\rvert \geq 2$	w^{n+1}	local conformal map, not globally defined

11.4.3 Radial quantisation

To talk about commutators of the L_n we need operators acting on a Hilbert space, which means we need a notion of time. Nash states the answer in two sentences: the Hamiltonian is the generator of dilations, $L_0 + \bar L_0$, and time ordering becomes radial ordering. Let us see why.

The plane is a cylinder. Recall the map $z = e^w$, $w = \tau + i\sigma$, of Example 11.15(b). A field theory on a spatial circle of circumference 2π evolving in Euclidean time τ lives on the cylinder $\mathbb{R} \times S^1$. The exponential map sends it to the punctured plane:

- a circle of constant time τ goes to the circle $|z| = e^\tau$;
- the infinite past $\tau \to -\infty$ goes to the origin $z = 0$;
- the infinite future $\tau \to +\infty$ goes to the point $z = \infty$;
- time translation $\tau \mapsto \tau + a$ goes to the dilation $z \mapsto e^a z$;
- space translation $\sigma \mapsto \sigma + b$ goes to the rotation $z \mapsto e^{ib} z$.

Since the map is conformal, a conformal field theory on the cylinder is the same thing as a conformal field theory on the plane.

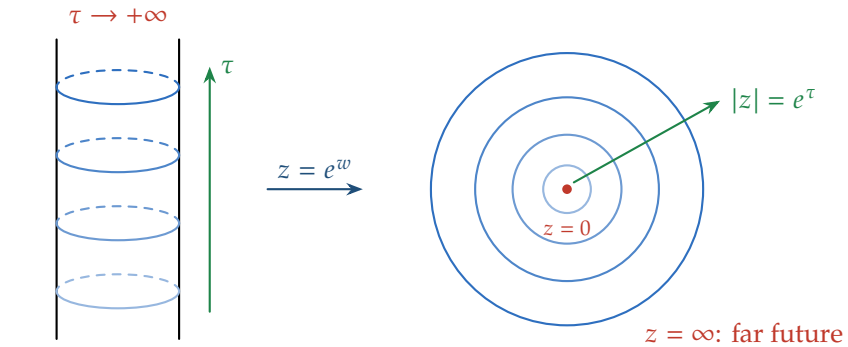


Figure 11.16: Radial quantisation. Circles of constant time on the cylinder (left) become concentric circles on the plane (right), with later times at larger radius. The infinite past shrinks to the single point $z = 0$, so an initial state is prepared by whatever operator we put at the origin: this is the state–operator correspondence. Time evolution is dilation, so the Hamiltonian is the dilation generator $L_0 + \bar{L}_0$.

The Hamiltonian. The Hamiltonian generates translations in τ, which on the plane are dilations $z \mapsto (1+\epsilon)z$, i.e. $f(z) = z$ in (11.96), and the same for $\bar{z}$. So on the plane H is represented by $L_0 + \bar{L}_0$. On the cylinder itself we can be more precise, using the transformation law of T with its Schwarzian. From Example 11.34, $T_{\text{cyl}}(w) = z^2 T(z) - \frac{c}{24}$. The charge generating $w \mapsto w + \epsilon$ on a circle of constant τ, where $\mathrm{d}w = i\,\mathrm{d}\sigma$, is

$$\oint \frac{\mathrm{d}w}{2\pi i}\, T_{\text{cyl}}(w) \overset{(1)}{=} \frac{1}{2\pi}\int_0^{2\pi} \mathrm{d}\sigma \left(z^2 \sum_n L_n z^{-n-2} - \frac{c}{24} \right) \overset{(2)}{=} \frac{1}{2\pi}\int_0^{2\pi} \mathrm{d}\sigma \left(\sum_n L_n\, e^{-n\tau} e^{-in\sigma} - \frac{c}{24} \right)$$
$$\overset{(3)}{=} L_0 - \frac{c}{24}. \tag{11.97}$$

(1) $\mathrm{d}w/(2\pi i) = \mathrm{d}\sigma/(2\pi)$ at fixed τ; insert the Schwarzian law and the mode expansion.
(2) $z = e^{\tau + i\sigma}$, so $z^{-n} = e^{-n\tau} e^{-in\sigma}$.
(3) $\int_0^{2\pi} e^{-in\sigma}\mathrm{d}\sigma = 2\pi\delta_{n,0}$.

Adding the antiholomorphic half,

$$H_{\text{cyl}} = L_0 + \bar{L}_0 - \frac{c}{12}, \qquad P_{\text{cyl}} = L_0 - \bar{L}_0. \tag{11.98}$$

The momentum P generates σ-translations, i.e. rotations of the plane, whose charge is $L_0 - \bar{L}_0$: the rotation $z \mapsto e^{i\epsilon} z$ has $f(z) = iz$ and $\bar{f}(\bar{z}) = -i\bar{z}$, so its charge is $i(L_0 - \bar{L}_0)$. An eigenstate of L_0 and $\bar{L}_0$ with eigenvalues $h, \bar{h}$ has energy $h + \bar{h} - \frac{c}{12}$ and momentum $h - \bar{h}$: the scaling dimension is the energy and the spin is the momentum. Why these, rather than mass and spin, are the right labels of a state is the subject of Section 11.4.6.

Time ordering becomes radial ordering. In the operator formalism, a Euclidean correlation function is a vacuum expectation value of a *time-ordered* product: operators at later times stand to the left. On the plane "later" means "further from the origin". So

$$\langle \phi_1(z_1) \cdots \phi_n(z_n) \rangle = \langle 0 | \, R\big(\phi_1(z_1) \cdots \phi_n(z_n)\big) | 0 \rangle, \qquad R\big(A(z)B(w)\big) = \begin{cases} A(z)B(w), & |z| > |w|, \\ \pm B(w)A(z), & |w| > |z|, \end{cases} \tag{11.99}$$

with the minus sign when both operators are fermionic. This is Nash's remark that "the points $z_1, \ldots, z_n$ in a correlation function are ordered by their differences $|z_i - z_j|$"; more precisely they are ordered by their distances $|z_i|$ from the origin, where the in-state sits. All the OPEs of Section 11.3 are statements about radially ordered products.

Commutators are contour integrals. Here is the one computational trick of this section. Let $A(w)$ be holomorphic and let $B(z)$ be any field. Consider the charge $A_C = \oint_C \frac{\mathrm{d}w}{2\pi i} A(w)$ on a circle C centred at 0. Then

$$\begin{aligned} \big[A_C, B(z)\big] &\overset{(1)}{=} \oint_{|w|>|z|} \frac{\mathrm{d}w}{2\pi i} A(w)B(z) - \oint_{|w|<|z|} \frac{\mathrm{d}w}{2\pi i} B(z)A(w) \\ &\overset{(2)}{=} \oint_{|w|>|z|} \frac{\mathrm{d}w}{2\pi i} R\big(A(w)B(z)\big) - \oint_{|w|<|z|} \frac{\mathrm{d}w}{2\pi i} R\big(A(w)B(z)\big) \\ &\overset{(3)}{=} \oint_z \frac{\mathrm{d}w}{2\pi i} A(w)B(z). \end{aligned} \tag{11.100}$$

(1) As operators, $A_C B(z)$ only makes sense (inside a correlator) when the charge acts at a later time than B, i.e. $|w| > |z|$, and $B(z)A_C$ when $|w| < |z|$. The charge is conserved, so the radius of the contour does not matter as long as it does not cross z.

(2) In each region the operator product is the radially ordered one.

(3) The radially ordered product is a single holomorphic function of w away from z (this is the content of the OPE). The difference of the two circles is a small circle around z (Figure 11.17).

So the commutator of a charge with a field is computed by the singular part of their OPE. For a double contour integral we apply the trick twice: first around z for fixed z, then around 0.

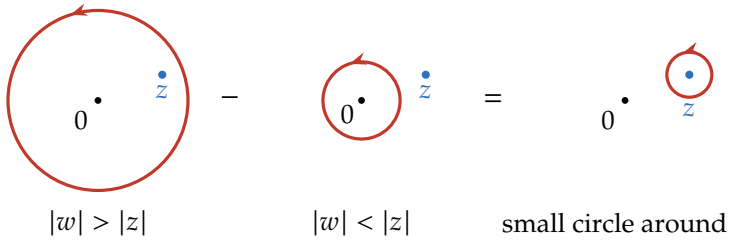


Figure 11.17: Why a commutator is a contour integral. The charge $\oint A$ acting after $B(z)$ is a circle outside z; acting before $B(z)$ it is a circle inside z. Their difference encloses only the point z, so only the singular terms of the OPE $A(w)B(z)$ contribute.

Worked example 11.35: The commutator of L_n with a primary field

Take $A = w^{n+1}T(w)$ and $B = \phi(z,\bar z)$ primary of weight $(h,\bar h)$.

$$[L_n,\phi(z,\bar z)] \overset{(1)}{=} \oint_z \frac{dw}{2\pi i}\, w^{n+1}\Big(\frac{h\,\phi(z,\bar z)}{(w-z)^2}+\frac{\partial\phi(z,\bar z)}{w-z}\Big)$$
$$\overset{(2)}{=} h(n+1)z^n\,\phi(z,\bar z)+z^{n+1}\,\partial\phi(z,\bar z). \tag{11.101}$$

(1) (11.100) with the $T\phi$ OPE (11.87); the regular terms have no residue.

(2) Cauchy: $\oint_z \frac{dw}{2\pi i}\frac{g(w)}{(w-z)^2} = g'(z)$ and $\oint_z \frac{dw}{2\pi i}\frac{g(w)}{w-z} = g(z)$, with $g = w^{n+1}$.

Check with $n=-1$: $[L_{-1},\phi]=\partial\phi$, so L_{-1} generates translations. With $n=0$: $[L_0,\phi(z)] = (h+z\partial)\phi$; at $z=0$ this is $h\phi(0)$, so L_0 measures the weight of a field sitting at the origin.

11.4.4 The Virasoro algebra from the TT OPE

Nash says: "If we combine this with the operator product expansion 11.29 we can verify that these generators have the commutation relations (11.34)." Here is the verification in full. It is the most important calculation of the chapter.

$$[L_m,L_n] \overset{(1)}{=} \oint_0 \frac{dz}{2\pi i}\, z^{n+1}\oint_z \frac{dw}{2\pi i}\, w^{m+1}\,T(w)T(z)$$
$$\overset{(2)}{=} \oint_0 \frac{dz}{2\pi i}\, z^{n+1}\oint_z \frac{dw}{2\pi i}\, w^{m+1}\Big[\frac{c/2}{(w-z)^4}+\frac{2T(z)}{(w-z)^2}+\frac{\partial T(z)}{w-z}\Big]$$
$$\overset{(3)}{=} \oint_0 \frac{dz}{2\pi i}\, z^{n+1}\Big[\frac{c}{2}\cdot\frac{(m+1)m(m-1)}{3!}z^{m-2}+2(m+1)z^m\,T(z)+z^{m+1}\,\partial T(z)\Big]$$
$$\overset{(4)}{=} \frac{c}{12}(m+1)m(m-1)\,\delta_{m+n,0}+2(m+1)\,L_{m+n}-(m+n+2)\,L_{m+n}$$
$$\overset{(5)}{=} (m-n)\,L_{m+n}+\frac{c}{12}\,m(m^2-1)\,\delta_{m+n,0}. \tag{11.102}$$

(1) Write $L_m = \oint \frac{dw}{2\pi i} w^{m+1}T(w)$ and $L_n = \oint \frac{dz}{2\pi i} z^{n+1}T(z)$. For fixed z, the commutator $[L_m,T(z)]$ is the w contour around z by (11.100); then integrate over z around 0.

(2) Insert the TT OPE (11.89); regular terms do not contribute to the w integral.

(3) Cauchy's formula $\oint_z \frac{dw}{2\pi i}\frac{g(w)}{(w-z)^{k+1}} = \frac{g^{(k)}(z)}{k!}$ with $g(w)=w^{m+1}$: $g'''=(m+1)m(m-1)w^{m-2}$, $g'=(m+1)w^m$, $g=w^{m+1}$.

(4) First term: $\oint \frac{dz}{2\pi i} z^{m+n-1} = \delta_{m+n,0}$. Second: $\oint \frac{dz}{2\pi i} z^{m+n+1}T(z) = L_{m+n}$ by (11.94). Third: integrate by parts on the closed contour, $\oint z^{m+n+2}\partial T = -(m+n+2)\oint z^{m+n+1}T$.

(5) $2(m+1)-(m+n+2) = m-n$, and $(m+1)m(m-1) = m(m^2-1)$.

Key idea: the Virasoro algebra (Nash's (11.34)–(11.35))

The modes of the stress tensor of any two-dimensional conformal field theory obey

$$[L_m,L_n] = (m-n)L_{m+n}+\frac{c}{12}m(m^2-1)\delta_{m+n,0},$$
$$[\bar L_m,\bar L_n] = (m-n)\bar L_{m+n}+\frac{c}{12}m(m^2-1)\delta_{m+n,0},$$

and $[L_m,\bar L_n]=0$. Each set spans a copy of the **Virasoro algebra** *Vir*: the Witt algebra plus one extra generator, the central element c, which commutes with everything (so in an

irreducible representation it is a number). The $\bar{L}_n$ commute with the L_n because $T(w)\bar{T}(\bar{z})$ has no singular terms: T is holomorphic and $\bar{T}$ antiholomorphic, so their product is regular.

The central term vanishes for $m \in \{-1, 0, 1\}$. So L_{-1}, L_0, L_1 close on themselves without any central term:

$$[L_0, L_{\pm 1}] = \mp L_{\pm 1}, \qquad [L_1, L_{-1}] = 2L_0. \tag{11.103}$$

This is $\mathfrak{sl}(2, \mathbb{C})$, the Lie algebra of the Möbius group, exactly the global part of the Witt algebra. The central charge is invisible to the global conformal group; it only shows up in the local transformations.

Physical picture: where the central term comes from

Classically the ℓ_n are vector fields and satisfy the Witt algebra exactly. The quantum L_n are quadratic in the fields ($T = -{:}\partial\varphi\partial\varphi{:}$), and normal ordering them requires subtracting an infinite constant. The subtraction cannot be made compatible with all the commutators at once; what is left over is the c-number $\frac{c}{12}m(m^2-1)$. In the contour calculation it is the $(w-z)^{-4}$ term, i.e. the double contraction of Figure 11.14. It is the same mechanism as an anomaly (Chapter 10): a classical symmetry algebra that the quantum theory realises only projectively.

Worked example 11.36: Which central terms are genuine?

Suppose we only knew that the L_n satisfy the Witt algebra up to a c-number, $[L_m, L_n] = (m-n)L_{m+n} + C(m)\delta_{m+n,0}$, with $C(-m) = -C(m)$ by antisymmetry. What can C be? The Jacobi identity $[[L_k, L_n], L_m] + [[L_n, L_m], L_k] + [[L_m, L_k], L_n] = 0$ with $k + n + m = 0$ gives, after the operator terms cancel (they obey the Witt Jacobi identity),

$$\begin{aligned} 0 &\overset{(1)}{=} (k-n)\,C(k+n) + (n-m)\,C(n+m) + (m-k)\,C(m+k) \\ &\overset{(2)}{=} -(k-n)\,C(m) - (n-m)\,C(k) - (m-k)\,C(n). \end{aligned}$$

(1) $[[L_k, L_n], L_m] = (k-n)[L_{k+n}, L_m]$, whose c-number part is $(k-n)C(k+n)\delta_{k+n+m,0}$; similarly for the other two.

(2) $k + n = -m$ and C is odd, so $C(k+n) = -C(m)$, and cyclically.

Take $n = 1$ and $k = -m-1$:

$$\begin{aligned} 0 &\overset{(1)}{=} (m+2)\,C(m) + (m-1)\,C(-m-1) - (2m+1)\,C(1) \\ &\overset{(2)}{=} (m+2)\,C(m) - (m-1)\,C(m+1) - (2m+1)\,C(1), \end{aligned}$$

(1) Substitute into the coefficients: $-(k-n) = m+2$, $-(n-m) = m-1$ (multiplying $C(k) = C(-m-1)$) and $-(m-k) = -(2m+1)$ (multiplying $C(n) = C(1)$).

(2) $C(-m-1) = -C(m+1)$.

that is, $(m-1)\,C(m+1) = (m+2)\,C(m) - (2m+1)\,C(1)$. For $m \geq 2$ this is a recursion that fixes every $C(m)$ once $C(1)$ and $C(2)$ are given. So the solutions form a two-dimensional space. Two solutions are $C(m) = m$ and $C(m) = m^3$:

$$\begin{aligned} C = m: &\quad (m-1)(m+1) = m^2 - 1, \qquad (m+2)m - (2m+1) = m^2 - 1 \ \checkmark \\ C = m^3: &\quad (m-1)(m+1)^3 = m^4 + 2m^3 - 2m - 1, \\ &\quad (m+2)m^3 - (2m+1) = m^4 + 2m^3 - 2m - 1 \ \checkmark \end{aligned}$$

Hence $C(m) = \alpha m^3 + \beta m$. The linear part is not genuine: shifting $L_0 \to L_0 + a$ changes $[L_m, L_{-m}] = 2mL_0 + C(m)$ into $2m(L_0 + a) + C(m) - 2ma$, i.e. $C(m) \to C(m) - 2am$. The standard choice fixes the shift by requiring $C(1) = 0$, so that $L_{\pm1}, L_0$ close into $\mathfrak{sl}(2)$; then $C(m) = \alpha(m^3 - m)$, and $\alpha = c/12$ is the convention. So the Virasoro algebra is the *unique* nontrivial central extension of the Witt algebra, and the one number c is all the freedom there is.

Watch out: the L_0 ambiguity and the Casimir energy

The freedom $L_0 \to L_0 + a$ is not just algebra. On the cylinder the natural generator is $L_0 - \frac{c}{24}$ (11.97), and in that basis the central term reads $\frac{c}{12}m^3$ with no linear piece. The plane generator L_0, with $L_0|0\rangle = 0$, is the one for which $\mathfrak{sl}(2)$ closes. The difference, $-\frac{c}{24}$, is the Casimir energy of Example 11.34. Different books use different L_0's; always check whether the vacuum has $L_0 = 0$ or $L_0 = -c/24$.

11.4.5 The free boson in modes: $c = 1$ by hand

The central charge of the free boson was found in Example 11.31 from the OPE. Now we see the same number in the Hilbert space, by building the L_n out of harmonic oscillators. This is the most concrete representation of the Virasoro algebra there is.

The current and its modes. With our normalisation $\langle \partial\varphi(w)\partial\varphi(z)\rangle = -\frac{1}{2(w-z)^2}$ ((11.70)), it is convenient to define the U(1) current

$$j(z) := i\sqrt{2}\,\partial\varphi(z), \qquad j(w)j(z) = \frac{1}{(w-z)^2} + \text{reg.}, \qquad T(z) = -{:}\partial\varphi\,\partial\varphi{:} = \tfrac{1}{2}{:}j\,j{:}(z). \tag{11.104}$$

(Check: $(i\sqrt{2})^2 \cdot (-\frac{1}{2}) = 1$, and $-(j/i\sqrt{2})^2 = \frac{1}{2}j^2$.) The current is a primary of weight 1 (Example 11.27), so its modes are labelled with a shift of 1:

$$j(z) = \sum_{n\in\mathbb{Z}} a_n\, z^{-n-1}, \qquad a_n = \oint_0 \frac{\mathrm{d}z}{2\pi i}\, z^n\, j(z). \tag{11.105}$$

Worked example 11.37: The oscillator algebra from the jj OPE

$$[a_m, a_n] \overset{(1)}{=} \oint_0 \frac{\mathrm{d}z}{2\pi i} z^n \oint_z \frac{\mathrm{d}w}{2\pi i} w^m \frac{1}{(w-z)^2} \overset{(2)}{=} \oint_0 \frac{\mathrm{d}z}{2\pi i} z^n\, m\, z^{m-1} \overset{(3)}{=} m\,\delta_{m+n,0}.$$

(1) The double-contour trick (11.100) with the OPE (11.104).
(2) Cauchy: the residue of $w^m/(w-z)^2$ at $w = z$ is $\frac{\mathrm{d}}{\mathrm{d}w}w^m = mw^{m-1}$.
(3) $\oint z^{m+n-1}\frac{\mathrm{d}z}{2\pi i} = \delta_{m+n,0}$.

So a_0 commutes with everything (it is the U(1) charge, the momentum of the zero mode), and for $n > 0$ the pairs (a_n, a_{-n}) are harmonic oscillators: $b_n = a_n/\sqrt{n}$ and $b_n^\dagger = a_{-n}/\sqrt{n}$ satisfy $[b_n, b_n^\dagger] = 1$. With $a_n^\dagger = a_{-n}$ (the factor i in j is what makes j Hermitian in radial quantisation), the Fock space built by the a_{-n} on a state annihilated by all $a_{n>0}$ has a positive inner product.

The Virasoro generators as oscillator bilinears. Insert (11.105) into $T = \frac{1}{2}{:}jj{:}$ and collect the coefficient of z^{-n-2}:

$$L_n = \frac{1}{2}\sum_{k\in\mathbb{Z}} {:}a_{n-k}\,a_k{:}, \qquad L_0 = \frac{1}{2}a_0^2 + \sum_{k\geq 1} a_{-k}\,a_k. \tag{11.106}$$

Normal ordering puts the annihilators $a_{k>0}$ to the right. It only matters for L_0: for $n \neq 0$ the two factors a_{n-k} and a_k commute (since $[a_{n-k}, a_k] = (n-k)\delta_{n,0} = 0$). In L_0 the unordered sum would be $\frac{1}{2}a_0^2 + \frac{1}{2}\sum_{k\geq 1}(a_{-k}a_k + a_k a_{-k}) = \frac{1}{2}a_0^2 + \sum_{k\geq 1} a_{-k}a_k + \frac{1}{2}\sum_{k\geq 1} k$, and the divergent constant $\frac{1}{2}\sum k$ is precisely the zero-point energy that normal ordering throws away. (Regularised as $\zeta(-1) = -\frac{1}{12}$, it gives $-\frac{1}{24}$: the Casimir energy $-c/24$ with $c = 1$ again.)

Worked example 11.38: The action of L_n on the oscillators

Since j is primary of weight 1, (11.101) gives $[L_m, j(z)] = (m+1)z^m j(z) + z^{m+1}\partial j(z)$. Taking the z^n moment:

$$[L_m, a_n] \overset{(1)}{=} \oint_0 \frac{\mathrm{d}z}{2\pi i}\, z^n\big((m+1)z^m\, j + z^{m+1}\partial j\big) \overset{(2)}{=} (m+1)\,a_{m+n} - (m+n+1)\,a_{m+n} = -n\,a_{m+n}.$$

(1) $a_n = \oint z^n j$, and L_m commutes with the c-number z^n.

(2) First term: definition of a_{m+n}. Second: integrate by parts, $\oint z^{m+n+1}\partial j = -(m+n+1)\oint z^{m+n} j$.

In particular $[L_0, a_{-n}] = n\,a_{-n}$: each creation operator a_{-n} raises the L_0 eigenvalue by n. One can check the same relation directly from (11.106) and $[a_m, a_n] = m\delta_{m+n,0}$.

Worked example 11.39: The central charge from a norm: $c = 1$

The vacuum $|0\rangle$ satisfies $a_n|0\rangle = 0$ for $n \geq 0$. The Virasoro algebra predicts, for $m \geq 2$,

$$\langle 0|L_m L_{-m}|0\rangle \overset{(1)}{=} \langle 0|[L_m, L_{-m}]|0\rangle \overset{(2)}{=} 2m\langle 0|L_0|0\rangle + \frac{c}{12}m(m^2-1) \overset{(3)}{=} \frac{c}{12}m(m^2-1).$$

(1) $L_m|0\rangle = 0$ for $m \geq -1$ (Section 11.4.7 below; for the boson it is visible from (11.106), since every term of L_m contains an annihilator or a_0).

(2) The Virasoro algebra (11.102).

(3) $L_0|0\rangle = 0$.

Now compute the left side with oscillators. In $L_{-m}|0\rangle = \frac{1}{2}\sum_k {:}a_{-m-k}a_k{:}|0\rangle$ only the terms with both factors creation operators survive: $k = -1, \ldots, -(m-1)$. Relabelling,

$$L_{-m}|0\rangle = \frac{1}{2}\sum_{k=1}^{m-1} a_{-k}\,a_{-(m-k)}|0\rangle.$$

For example $L_{-2}|0\rangle = \frac{1}{2}a_{-1}^2|0\rangle$ and $L_{-3}|0\rangle = a_{-1}a_{-2}|0\rangle$. The norm is

$$\langle 0|L_m L_{-m}|0\rangle \overset{(1)}{=} \frac{1}{4}\sum_{k,l=1}^{m-1}\langle 0|a_{m-l}\,a_l\,a_{-k}\,a_{-(m-k)}|0\rangle \overset{(2)}{=} \frac{1}{4}\sum_{k,l=1}^{m-1} l(m-l)\big(\delta_{l,k} + \delta_{l,m-k}\big)$$
$$\overset{(3)}{=} \frac{1}{2}\sum_{k=1}^{m-1} k(m-k) \overset{(4)}{=} \frac{1}{2}\Big(m\cdot\frac{m(m-1)}{2} - \frac{(m-1)m(2m-1)}{6}\Big) \overset{(5)}{=} \frac{m(m^2-1)}{12}.$$

(1) $L_m = L_{-m}^\dagger$ (Section 11.4.7), so the bra is $\frac{1}{2}\sum_l \langle 0|a_{m-l}a_l$.

(2) Move the annihilators to the right. Each must be absorbed by a creation operator with the opposite index, and $[a_p, a_{-p}] = p$: a_l pairs with a_{-k} ($l = k$) or with $a_{-(m-k)}$ ($l = m-k$), and a_{m-l} with the other.
(3) Both deltas give $k(m-k)$.
(4) $\sum_{k=1}^{m-1} k = \frac{m(m-1)}{2}$ and $\sum_{k=1}^{m-1} k^2 = \frac{(m-1)m(2m-1)}{6}$.
(5) $m(m-1)\left[\frac{m}{2} - \frac{2m-1}{6}\right] = m(m-1)\frac{m+1}{6}$, times $\frac{1}{2}$.

Comparing: $\frac{c}{12}m(m^2-1) = \frac{m(m^2-1)}{12}$ for every m, so $c = 1$. For $m = 2$: $\|L_{-2}|0\rangle\|^2 = \frac{1}{4}\left\|a_{-1}^2|0\rangle\right\|^2 = \frac{1}{4}\cdot 2 = \frac{1}{2} = \frac{c}{2}$.

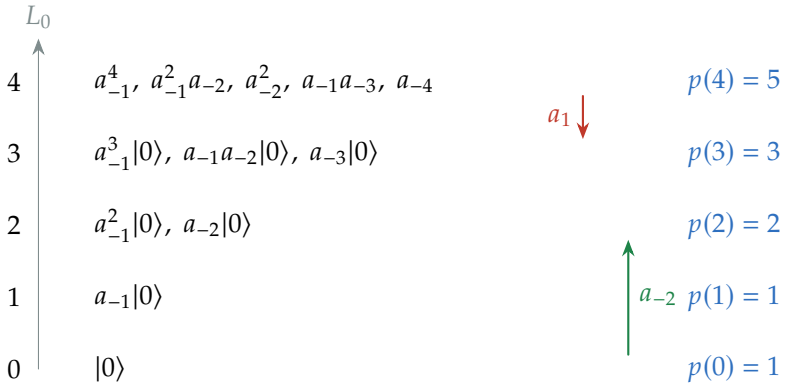


Figure 11.18: The Fock space of one free boson (zero U(1) charge), graded by L_0. A state at level N is a product of creation operators whose indices add up to N, so the number of states is the number $p(N)$ of partitions of N: $1, 1, 2, 3, 5, 7, 11, \ldots$. Creation operators a_{-n} move up by n, annihilators a_n down by n. The same counting will reappear for the Virasoro descendants of a generic highest-weight state.

Try it: the vertex-operator states

The state $|\alpha\rangle = V_\alpha(0)|0\rangle$ of Example 11.28 satisfies $a_n|\alpha\rangle = 0$ for $n > 0$ and $a_0|\alpha\rangle = \frac{\alpha}{\sqrt{2}}|\alpha\rangle$. Use (11.106) to find $L_0|\alpha\rangle$ and compare with the weight $h = \alpha^2/4$.
Answer: $L_0|\alpha\rangle = \frac{1}{2}a_0^2|\alpha\rangle = \frac{1}{2}\cdot\frac{\alpha^2}{2}|\alpha\rangle = \frac{\alpha^2}{4}|\alpha\rangle$. (The a_0 eigenvalue comes from the simple pole of $j(w)V_\alpha(z)$: one contraction gives $i\sqrt{2}\cdot i\alpha\cdot(-\frac{1}{2(w-z)}) = \frac{\alpha/\sqrt{2}}{w-z}$.)

11.4.6 How states are labelled: mass and spin versus dimension and spin

Radial quantisation gave us a Hilbert space and a Hamiltonian. Before we build states in it, we should ask what the *labels* of a state are. In ordinary quantum field theory we never hesitate: a one-particle state is labelled by its mass and its spin. In a conformal field theory these labels are replaced by the scaling dimension Δ and the spin s. This subsection explains why the old labels fail and why the new ones work.

Ordinary QFT: Wigner's classification. In a Poincaré-invariant theory (signature $(+,-,-,-)$ for this paragraph) the symmetry generators are the momenta P_μ and the Lorentz generators $J_{\mu\nu}$. A one-particle state is, by definition, a state in an **irreducible representation** of the Poincaré group: a space of states that the group maps into itself and that has no smaller invariant subspace. By Schur's lemma, an operator that commutes with every generator acts as a number

on an irreducible representation. Such operators are the **Casimirs**. The Poincaré group has two:

$$P^2 = P_\mu P^\mu, \qquad W^2 = W_\mu W^\mu, \quad W_\mu \coloneqq \tfrac{1}{2}\,\epsilon_{\mu\nu\rho\sigma} J^{\nu\rho} P^\sigma, \tag{11.107}$$

where W_μ is the **Pauli–Lubanski vector**. Their values label the representation. For a massive particle go to the rest frame, $p^\mu = (m,0,0,0)$, and write $J_i = \frac{1}{2}\epsilon_{ijk}J^{jk}$ for the angular momentum:

$$\begin{aligned} W_0 &\overset{(1)}{=} \tfrac{1}{2}\,\epsilon_{0\nu\rho0} J^{\nu\rho} m = 0, \\ W_i &\overset{(2)}{=} \tfrac{1}{2}\,\epsilon_{ijk0} J^{jk} m \overset{(3)}{=} -\tfrac{1}{2}\,\epsilon_{0ijk} J^{jk} m \overset{(4)}{=} -m\,J_i, \\ W^2 &\overset{(5)}{=} W_0^2 - \textstyle\sum_i W_i^2 = -m^2\,\vec{J}^{\,2} \overset{(6)}{=} -m^2\, s(s+1). \end{aligned}$$

(1) In the rest frame only $P^0 = m$ is nonzero, so $\sigma = 0$; then $\epsilon_{0\nu\rho0}$ has a repeated index and vanishes.
(2) Again only $\sigma = 0$ survives, and ν, ρ must be spatial.
(3) Moving the index 0 from the last slot to the first passes it through three others: an odd permutation, so a minus sign.
(4) $\epsilon_{0ijk} = \epsilon_{ijk}$ and the definition of J_i.
(5) Mostly-minus signature.
(6) On a spin-s multiplet $\vec{J}^{\,2} = s(s+1)$.

So a particle is labelled by (m^2, s), and its states $|p,\sigma\rangle$ run over the mass shell $p^2 = m^2$ and the $2s+1$ spin projections σ. (For $m = 0$ the rest frame does not exist and s is replaced by the helicity.) This is **Wigner's classification**.

Why mass fails in a CFT. A conformal theory contains the dilatation. Write the generators as Hermitian differential operators on functions,

$$P_\mu = -i\partial_\mu, \qquad D = -i\,x^\nu\partial_\nu, \qquad K_\mu = -i\big(2x_\mu\, x^\nu\partial_\nu - x^2\partial_\mu\big), \tag{11.108}$$

which are $-i$ times the real generators of Section 11.1.1 and (11.13). The key commutator is

$$\begin{aligned} [D, P_\mu] f &\overset{(1)}{=} (-i)^2\big(x^\nu\partial_\nu\partial_\mu f - \partial_\mu(x^\nu\partial_\nu f)\big) \overset{(2)}{=} -\big(x^\nu\partial_\nu\partial_\mu f - \delta^\nu_\mu\partial_\nu f - x^\nu\partial_\mu\partial_\nu f\big) \\ &\overset{(3)}{=} \partial_\mu f \overset{(4)}{=} i\,P_\mu f. \end{aligned} \tag{11.109}$$

(1) Definitions (11.108); $(-i)(-i) = -1$.
(2) Product rule: $\partial_\mu(x^\nu\partial_\nu f) = \delta^\nu_\mu\partial_\nu f + x^\nu\partial_\mu\partial_\nu f$.
(3) Partial derivatives commute, so the first and last terms cancel.
(4) $\partial_\mu = iP_\mu$.

In the same way (the coefficient of ∂ in K_μ has degree 2 in x instead of 0) one finds

$$[D, P_\mu] = iP_\mu, \qquad [D, K_\mu] = -iK_\mu, \qquad [K_\mu, P_\nu] = 2i\big(\eta_{\mu\nu}D - M_{\mu\nu}\big), \tag{11.110}$$

with $M_{\mu\nu} = i(x_\mu\partial_\nu - x_\nu\partial_\mu)$ the rotation generators. Now suppose $|p\rangle$ were a state of definite momentum, $P_\mu|p\rangle = p_\mu|p\rangle$ with $p^2 = m^2$. Conjugate P_μ by a finite dilatation:

$$\begin{aligned} e^{i\alpha D} P_\mu e^{-i\alpha D} &\overset{(1)}{=} P_\mu + [i\alpha D, P_\mu] + \tfrac{1}{2!}\big[i\alpha D, [i\alpha D, P_\mu]\big] + \cdots \\ &\overset{(2)}{=} P_\mu\Big(1 - \alpha + \tfrac{\alpha^2}{2!} - \cdots\Big) = e^{-\alpha} P_\mu, \end{aligned}$$

(1) The expansion $e^X Y e^{-X} = Y + [X,Y] + \frac{1}{2!}[X,[X,Y]] + \cdots$.

(2) $[i\alpha D, P_\mu] = i\alpha \cdot iP_\mu = -\alpha P_\mu$ by (11.109), and each further commutator gives another factor $-\alpha$.

and therefore

$$P_\mu\big(e^{-i\alpha D}|p\rangle\big) \overset{(1)}{=} e^{-i\alpha D}\big(e^{i\alpha D}P_\mu e^{-i\alpha D}\big)|p\rangle \overset{(2)}{=} e^{-\alpha}p_\mu\,\big(e^{-i\alpha D}|p\rangle\big).$$

(1) Insert $e^{-i\alpha D}e^{i\alpha D} = \mathbb{1}$ to the left of P_μ.
(2) The conjugation just computed, then $P_\mu|p\rangle = p_\mu|p\rangle$.

The dilated state has momentum $e^{-\alpha}p_\mu$ and mass squared $e^{-2\alpha}m^2$. So if one value $m^2 > 0$ occurs in the spectrum of P^2, every positive value occurs: the spectrum of P^2 is either $\{0\}$ or a continuum, and there is no discrete mass to label anything. Moreover D does not commute with P^2 (it rescales it), so P^2 is not a Casimir of the conformal group and cannot label its irreducible representations at all. And states of definite P_μ are plane waves, which are not normalisable. In an ordinary theory we cure this with wave packets around a sharp mass shell; in a CFT there is no mass shell to build them around.

Physical picture: no particles, only a continuum

The Fourier transform of a two-point function of dimension Δ in d dimensions is $\langle \mathcal{O}\mathcal{O}\rangle(p) \propto (p^2)^{\Delta-d/2}$. An ordinary particle shows up as a pole $1/(p^2-m^2)$. A power $(p^2)^{\Delta-d/2}$ has no pole, only a branch cut starting at $p^2 = 0$: a continuum of "masses". The one exception is $\Delta = \frac{d-2}{2}$, where the power is $(p^2)^{-1}$: a massless pole, i.e. a free massless field such as our free boson.

The CFT answer: diagonalise the dilatation. Instead of P_μ we diagonalise D. Radial quantisation makes this natural: on the cylinder $\mathbb{R}\times S^{d-1}$ the dilatation *is* the time translation, so its eigenstates are energy eigenstates. Space is now the compact sphere S^{d-1}, so the spectrum is discrete and the eigenstates are normalisable, just as a particle in a box has discrete, normalisable energy levels while a free particle on a line does not. It is convenient to use the generator with real eigenvalues in radial quantisation,

$$\mathcal{D} := -iD, \qquad [\mathcal{D}, P_\mu] = P_\mu, \qquad [\mathcal{D}, K_\mu] = -K_\mu, \qquad \mathcal{D}|\Delta\rangle = \Delta|\Delta\rangle, \tag{11.111}$$

where the commutators follow from (11.110): $[-iD, P_\mu] = -i\cdot iP_\mu = P_\mu$. (In radial quantisation Hermitian conjugation includes the inversion of (11.118), and it is $\mathcal{D}$, not D, that is Hermitian there.) Now the ladder:

$$\begin{aligned}\mathcal{D}\big(P_\mu|\Delta\rangle\big) &\overset{(1)}{=} \big(P_\mu\mathcal{D} + [\mathcal{D}, P_\mu]\big)|\Delta\rangle \overset{(2)}{=} (\Delta+1)\,P_\mu|\Delta\rangle,\\ \mathcal{D}\big(K_\mu|\Delta\rangle\big) &\overset{(3)}{=} \big(K_\mu\mathcal{D} + [\mathcal{D}, K_\mu]\big)|\Delta\rangle \overset{(4)}{=} (\Delta-1)\,K_\mu|\Delta\rangle.\end{aligned} \tag{11.112}$$

(1) Reorder: $\mathcal{D}P_\mu = P_\mu\mathcal{D} + [\mathcal{D}, P_\mu]$.
(2) $\mathcal{D}|\Delta\rangle = \Delta|\Delta\rangle$ and $[\mathcal{D}, P_\mu] = P_\mu$ from (11.111).
(3) The same reordering for K_μ.
(4) $[\mathcal{D}, K_\mu] = -K_\mu$.

So P_μ raises the dimension by one and K_μ lowers it by one, exactly like $a^\dagger$ and a for the harmonic oscillator. Physically, $P_\mu|\Delta\rangle$ is the state of the field $\partial_\mu\mathcal{O}$, and a derivative adds one unit of inverse length. If dimensions are bounded below (unitarity demands it, Section 11.4.11), repeated lowering must stop: there is a state with

$$K_\mu|\Delta\rangle = 0 \quad \text{for all } \mu. \tag{11.113}$$

This lowest state is the state of a **primary** operator, and the states $P_{\mu_1}\cdots P_{\mu_n}|\Delta\rangle$ above it are its **descendants**. The representation is labelled by the dimension Δ of the bottom state and by its spin under the rotations $M_{\mu\nu}$, which commute with $\mathcal{D}$.

The same structure in two dimensions. In $d=2$ the dictionary is

$$\mathcal{D} = L_0 + \bar{L}_0, \qquad \text{rotation} = L_0 - \bar{L}_0, \qquad P \leftrightarrow L_{-1},\ \bar{L}_{-1}, \qquad K \leftrightarrow L_1,\ \bar{L}_1. \tag{11.114}$$

Indeed $\ell_{-1} = -\partial_z$ is a translation and $\ell_1 = -z^2\partial_z$ is a special conformal transformation (Section 11.3), and the Virasoro algebra (11.102) gives $[L_0, L_{\mp 1}] = \pm L_{\mp 1}$, the two-dimensional form of (11.111). A state with L_0 and $\bar{L}_0$ eigenvalues $(h, \bar{h})$ is labelled by

$$\Delta = h + \bar{h} \quad \text{(scaling dimension)}, \qquad s = h - \bar{h} \quad \text{(spin)}, \tag{11.115}$$

and by (11.98) its energy on the cylinder is $\Delta - \frac{c}{12}$ and its momentum is s. What is new in two dimensions is that every L_{-n} with $n \geq 1$ raises the weight, not just L_{-1}. For a state $|h\rangle$ with $L_0|h\rangle = h|h\rangle$,

$$L_0\big(L_{-n}|h\rangle\big) \overset{(1)}{=} \big(L_{-n}L_0 + [L_0, L_{-n}]\big)|h\rangle \overset{(2)}{=} \big(h\,L_{-n} + n\,L_{-n}\big)|h\rangle = (h+n)\,L_{-n}|h\rangle. \tag{11.116}$$

(1) Reorder.
(2) (11.102) with $(m,n) \to (0,-n)$: $[L_0, L_{-n}] = (0-(-n))L_{-n} + \frac{c}{12}\cdot 0 = nL_{-n}$ (the central term has a factor $m = 0$).

So L_{-n} raises h by n and, by the same computation, L_n lowers it by n. The two-dimensional primary is annihilated by all the lowering operators,

$$L_n|h\rangle = 0 \quad (n > 0),$$

the infinite-dimensional version of (11.113). This is precisely the highest-weight condition of the next subsection. A state annihilated only by L_1 (the true K) is called **quasi-primary**: it is primary for the global group but not for the whole Virasoro algebra. T itself is the standard example: its state $L_{-2}|0\rangle$ obeys $L_1L_{-2}|0\rangle = [L_1, L_{-2}]|0\rangle = 3L_{-1}|0\rangle = 0$, but $L_2L_{-2}|0\rangle = [L_2, L_{-2}]|0\rangle = (4L_0 + \frac{c}{2})|0\rangle = \frac{c}{2}|0\rangle \neq 0$.

Worked example 11.40: The quadratic Casimir: labels are still Casimir values

The labels Δ, s are not a break with Wigner's idea; they are Casimir values of the global conformal group. For the global algebra $\{L_{-1}, L_0, L_1\}$, with $[L_1, L_{-1}] = 2L_0$, take $C := L_0^2 - \frac{1}{2}(L_1L_{-1} + L_{-1}L_1)$. It commutes with L_1:

$$\begin{aligned}[C, L_1] &\overset{(1)}{=} L_0[L_0, L_1] + [L_0, L_1]L_0 - \tfrac{1}{2}\big(L_1[L_{-1}, L_1] + [L_{-1}, L_1]L_1\big) \\ &\overset{(2)}{=} -(L_0L_1 + L_1L_0) - \tfrac{1}{2}\big(-2L_1L_0 - 2L_0L_1\big) = 0.\end{aligned}$$

(1) $[AB, C] = A[B, C] + [A, C]B$, and $[L_1, L_1] = 0$ removes the other two terms.
(2) $[L_0, L_1] = -L_1$ and $[L_{-1}, L_1] = -2L_0$.

The same computation with L_{-1} (all signs flipped) and the obvious $[C, L_0] = 0$ (each term of C has total weight zero) show that C commutes with all three generators, so it is constant on an irreducible representation. Evaluate it on the primary:

$$C|h\rangle \overset{(1)}{=} \Big(h^2 - \tfrac{1}{2}L_1L_{-1}\Big)|h\rangle \overset{(2)}{=} \Big(h^2 - \tfrac{1}{2}\big(L_{-1}L_1 + 2L_0\big)\Big)|h\rangle \overset{(3)}{=} h(h-1)|h\rangle.$$

(1) $L_0^2|h\rangle = h^2|h\rangle$, and $L_{-1}L_1|h\rangle = 0$ because $L_1|h\rangle = 0$.

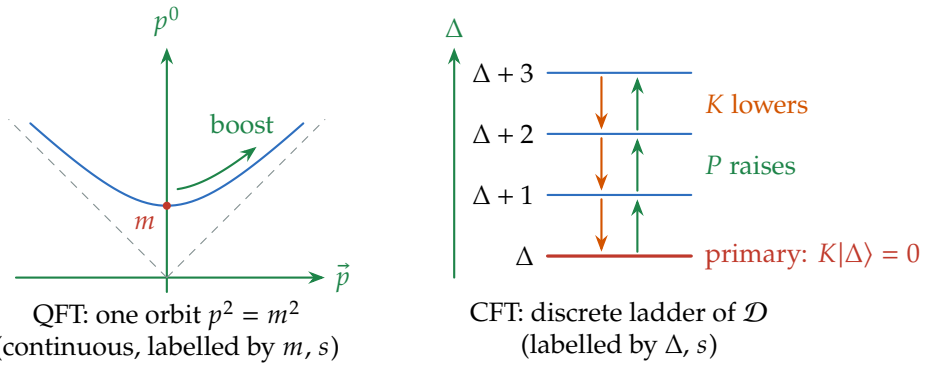
(2) $L_1 L_{-1} = L_{-1}L_1 + [L_1, L_{-1}] = L_{-1}L_1 + 2L_0$.
(3) $L_1|h\rangle = 0$ again, and $L_0|h\rangle = h|h\rangle$.

Adding the antiholomorphic copy, the Casimir of the full global group is

$$2\big(h(h-1) + \bar{h}(\bar{h}-1)\big) \overset{(1)}{=} (\Delta^2 + s^2) - 2\Delta = \Delta(\Delta-2) + s^2.$$

(1) $h = \frac{1}{2}(\Delta + s)$, $\bar{h} = \frac{1}{2}(\Delta - s)$, so $2(h^2 + \bar{h}^2) = \Delta^2 + s^2$ and $2(h + \bar{h}) = 2\Delta$.

This is the $d = 2$ case of the general $\mathrm{SO}(d+1,1)$ Casimir $\Delta(\Delta - d) + \ell(\ell + d - 2)$ with $\ell = |s|$. So a CFT representation is labelled by Casimir eigenvalues just as a particle is. The difference is which Casimirs: those of the conformal group, whose eigenvalues are functions of (Δ, s), rather than P^2 and W^2.



QFT: one orbit $p^2 = m^2$ (continuous, labelled by m, s)

CFT: discrete ladder of $\mathcal{D}$ (labelled by Δ, s)

Figure 11.19: Two ways of labelling states. Left: in a Poincaré-invariant theory a particle's states fill one mass hyperboloid $p^2 = m^2$; boosts move along it and the Casimirs $P^2 = m^2$, W^2 label the whole orbit. A dilatation would move the hyperboloid itself, so in a CFT no single m survives. Right: in a CFT we diagonalise the dilatation instead. Its spectrum is discrete; P_μ climbs the ladder by one unit, K_μ descends, and the bottom rung, killed by K_μ, is the primary.

	ordinary QFT	CFT (in $d = 2$ where specified)
symmetry group	Poincaré: P_μ, $J_{\mu\nu}$	conformal $\mathrm{SO}(d+1,1)$: P_μ, $M_{\mu\nu}$, D, K_μ; in $d = 2$ two copies of Virasoro
labels	Casimirs $P^2 = m^2$ and $W^2 = -m^2 s(s+1)$ (or helicity)	eigenvalues of $\mathcal{D}$ and rotation: $\Delta = h + \bar{h}$, $s = h - \bar{h}$ (Casimir $\Delta(\Delta - d) + \ell(\ell + d - 2)$)
spectrum of the label	m discrete, p_μ continuous	Δ discrete; P^2 continuous with no gap
basic object	a particle: the orbit $\lvert p, \sigma\rangle$ on the mass shell	a primary $\lvert h, \bar{h}\rangle = \phi(0)\lvert 0\rangle$ with K_μ (in 2D all $L_{n>0}$) annihilating it, plus its tower of descendants
moving within the representation	boosts and translations, at fixed m	P_μ raises Δ by 1, K_μ lowers it by 1; in 2D L_{-n} raises h by n
generator of time evolution	$H = P^0$ on $\mathbb{R}^{d-1}$	$\mathcal{D}$ on the sphere S^{d-1}; in 2D $H_{\mathrm{cyl}} = L_0 + \bar{L}_0 - \frac{c}{12}$
states are	asymptotic particles	local operators at the origin (state–operator correspondence)

Table 11.1: Labelling states in ordinary QFT and in CFT.

11.4.7 States from fields: highest-weight states

The vacuum. The vacuum $|0\rangle$ is the state on the cylinder at $\tau \to -\infty$ with nothing inserted, i.e. nothing at $z = 0$. So correlation functions with $|0\rangle$ as the in-state must be regular at the origin. Apply this to T itself:

$$T(z)|0\rangle = \sum_{n\in\mathbb{Z}} L_n|0\rangle\, z^{-n-2}$$

must be regular as $z \to 0$. The powers z^{-n-2} with $n \geq -1$ are singular, so

$$L_n|0\rangle = 0 \quad \text{for } n \geq -1, \qquad \text{in particular} \quad L_{-1}|0\rangle = L_0|0\rangle = L_1|0\rangle = 0. \tag{11.117}$$

The vacuum is invariant under the global conformal group. (Nash's (11.46) states the last part; we come back to it in Section 11.4.13.) The first nontrivial state is $L_{-2}|0\rangle = T(0)|0\rangle$, the state made by the stress tensor itself.

Hermitian conjugation. On the cylinder, Hermitian conjugation of a Euclidean field theory includes time reversal $\tau \to -\tau$. On the plane, $\tau \to -\tau$ with σ fixed is $z = e^{\tau+i\sigma} \mapsto e^{-\tau+i\sigma} = 1/\bar{z}$. For a primary of weights $(h, \bar{h})$ with real correlators, the conjugate is defined by

$$\phi(z,\bar{z})^\dagger = \bar{z}^{-2h}\, z^{-2\bar{h}}\, \phi\big(1/\bar{z}, 1/z\big). \tag{11.118}$$

The prefactor is the conformal factor of the map $z \mapsto 1/\bar{z}$, as in the primary law (11.60). Apply it to T, which has $(h, \bar{h}) = (2, 0)$:

$$\sum_n L_n^\dagger\, \bar{z}^{-n-2} \overset{(1)}{=} T(z)^\dagger \overset{(2)}{=} \bar{z}^{-4}\, T(1/\bar{z}) \overset{(3)}{=} \bar{z}^{-4} \sum_m L_m\, \bar{z}^{m+2} = \sum_m L_m\, \bar{z}^{m-2}.$$

(1) Conjugate the mode expansion; $(z^{-n-2})^* = \bar{z}^{-n-2}$.
(2) (11.118) for T.
(3) Mode expansion at the point $1/\bar{z}$.

Matching powers of $\bar{z}$, $-n-2 = m-2$, i.e. $m = -n$:

$$L_n^\dagger = L_{-n}, \qquad \bar{L}_n^\dagger = \bar{L}_{-n}. \tag{11.119}$$

This is what makes L_0 Hermitian (so its eigenvalues, the energies, are real) and what turns lowering operators into the adjoints of raising operators. It is the input for all the norm computations below. In particular the Virasoro algebra with c real is compatible with (11.119): taking the adjoint of (11.102) gives $[L_{-n}, L_{-m}] = (m-n)L_{-m-n} + \frac{c}{12}m(m^2-1)\delta_{m+n,0}$, which is (11.102) again with $(m, n) \to (-n, -m)$.

The state of a primary field (Nash's (11.36)–(11.38)). The field at the origin prepares a state at $\tau = -\infty$. Define, as Nash does,

$$|h\rangle := \phi(0)|0\rangle, \qquad |h,\bar{h}\rangle := \phi(0,0)|0\rangle. \tag{11.120}$$

Its properties follow from (11.101) and (11.117):

$$\begin{aligned} L_n|h\rangle &\overset{(1)}{=} [L_n, \phi(0)]|0\rangle + \phi(0)L_n|0\rangle \overset{(2)}{=} \big(h(n+1)z^n\phi(z) + z^{n+1}\partial\phi(z)\big)\big|_{z=0}|0\rangle \\ &\overset{(3)}{=} \begin{cases} h|h\rangle, & n = 0, \\ 0, & n \geq 1, \\ \partial\phi(0)|0\rangle, & n = -1. \end{cases} \end{aligned} \tag{11.121}$$

(1) Write $L_n\phi = [L_n, \phi] + \phi L_n$.
(2) (11.101) at $z = 0$; the second term vanishes for $n \geq -1$ by (11.117).
(3) For $n \geq 1$ both z^n and z^{n+1} vanish at $z = 0$; for $n = 0$ only $h\phi(0)$ survives; for $n = -1$ only $\partial\phi(0)$ survives.

So $L_0|h\rangle = h|h\rangle$ and $L_n|h\rangle = 0$ for $n \geq 1$: Nash's (11.37), which is exactly the defining property (6.57) of the highest-weight vector of Chapter 6. With $\bar{T}$ as well, $|h, \bar{h}\rangle$ is simultaneously highest weight for *Vir* and $\overline{Vir}$ (Nash's (11.38)), and the two towers built on it give a representation of $Vir \oplus \overline{Vir}$. Nash asks for a "single valued" primary: under a rotation by 2π a field picks up $e^{2\pi i(h-\bar{h})}$, so single-valuedness means integer spin $h - \bar{h}$ (fermions such as ψ, with spin $\frac{1}{2}$, need a sign and a choice of boundary condition).

Working definition 11.12 (highest-weight state)

A state $|h\rangle$ is a **highest-weight state** of weight h for a representation of *Vir* with central charge c [a space of states on which the L_n act with $[L_m, L_n] = (m-n)L_{m+n} + \frac{c}{12}m(m^2-1)\delta_{m+n,0}$] if

$$L_0|h\rangle = h|h\rangle, \qquad L_n|h\rangle = 0 \quad (n \geq 1).$$

Operational test: a state is highest weight if and only if it is annihilated by L_1 and L_2 (every L_n with $n \geq 3$ is a multiple of a commutator of these: $L_3 = -[L_1, L_2]$, $L_4 = -\frac{1}{2}[L_1, L_3]$, and so on by (11.102)). *State–operator form:* highest-weight states are exactly the states $\phi(0)|0\rangle$ of primary fields [fields for which $\phi\, dz^h d\bar{z}^{\bar{h}}$ is invariant under conformal maps].

Watch out: highest or lowest?

$|h\rangle$ has the *lowest* L_0 eigenvalue in its tower, since $[L_0, L_{-n}] = nL_{-n}$ makes every L_{-n} raise the energy. Physicists still call it "highest weight" because the roles of raising and lowering are reversed compared with, say, SU(2): the annihilation conditions $L_{n>0}|h\rangle = 0$ look like those of a highest-weight vector. Nash's footnote in Chapter 6 makes the same remark.

Why $L_n|h\rangle = 0$ is forced (the Chapter 6 argument). There is an independent argument, which Nash uses in Chapter 6. From the algebra, $[L_0, L_n] = -nL_n$, so

$$L_0\big(L_n|h\rangle\big) \overset{(1)}{=} \big(L_nL_0 + [L_0, L_n]\big)|h\rangle \overset{(2)}{=} (h-n)\,L_n|h\rangle.$$

(1) Reorder.
(2) $L_0|h\rangle = h|h\rangle$ and $[L_0, L_n] = (0-n)L_n$ from (11.102).

If $L_n|h\rangle$ were nonzero for some $n \geq 1$ it would have energy $h - n < h$. In a physical theory the energy $L_0 + \bar{L}_0$ is bounded below, and a state of lowest energy in its tower must therefore be annihilated by all $L_{n\geq 1}$. Both routes lead to the same place: the field-theory route shows that primary fields produce highest-weight states; the algebraic route shows that boundedness from below demands them.

Inner products and the out-state. The bra corresponding to $|h\rangle$ is, by (11.118),

$$\langle h| = \big(\phi(0,0)|0\rangle\big)^\dagger \overset{(1)}{=} \lim_{z\to 0}\langle 0|\,\bar{z}^{-2h}z^{-2\bar{h}}\,\phi\big(1/\bar{z}, 1/z\big) \overset{(2)}{=} \lim_{w\to\infty} w^{2h}\bar{w}^{2\bar{h}}\,\langle 0|\,\phi(w,\bar{w}).$$

(1) (11.118); the adjoint reverses the order, so $\langle 0|$ stands on the left.
(2) Set $w = 1/\bar{z}$, so $\bar{w} = 1/z$, $\bar{z}^{-2h} = w^{2h}$ and $z^{-2\bar{h}} = \bar{w}^{2\bar{h}}$.

In words: the out-state is made by putting ϕ at infinity and multiplying by the conformal factor that compensates the decay of the two-point function. Concretely,

$$\langle h|h\rangle \overset{(1)}{=} \lim_{w\to\infty} w^{2h}\bar{w}^{2\bar{h}}\,\langle\phi(w,\bar{w})\,\phi(0,0)\rangle \overset{(2)}{=} \lim_{w\to\infty} w^{2h}\bar{w}^{2\bar{h}}\,\frac{C}{w^{2h}\bar{w}^{2\bar{h}}} = C.$$

(1) The out-state is prepared by the operator at $z=\infty$, rescaled by its conformal factor.
(2) The two-point function (11.65).

We normalise primaries so that $C=1$, i.e. $\langle h|h\rangle = 1$. Positivity of the two-point coefficient is the first unitarity condition: it is the norm of a state.

11.4.8 Descendants and conformal families

Descendant states (Nash's (11.39)). The highest-weight conditions say that $L_{n\geq 1}$ kill $|h\rangle$ and L_0 multiplies it by h. The only operators that produce new states are the L_{-n} with $n\geq 1$ (Nash notes that "the condition $n<0$ follows from 11.37"). Acting repeatedly gives the **descendant** (or **secondary**) states

$$L_{-n_1}L_{-n_2}\cdots L_{-n_k}|h\rangle, \qquad n_1\geq n_2\geq\cdots\geq n_k\geq 1. \tag{11.122}$$

Nash writes the same states as $L_{-1}^{\alpha_1}L_{-2}^{\alpha_2}\cdots L_{-s}^{\alpha_s}|h\rangle$. The ordering does not matter for which states we get: any product of L_{-n}'s can be reordered into a fixed order using (11.102), at the cost of terms with fewer L's, which are again of the same form. So either ordering gives a spanning set.

The level. Each L_{-n} raises L_0 by n:

$$\begin{aligned} L_0\,L_{-n_1}\cdots L_{-n_k}|h\rangle &\overset{(1)}{=} \Big(\sum_{j=1}^{k} L_{-n_1}\cdots[L_0,L_{-n_j}]\cdots L_{-n_k} + L_{-n_1}\cdots L_{-n_k}L_0\Big)|h\rangle \\ &\overset{(2)}{=} \Big(\sum_j n_j + h\Big)\,L_{-n_1}\cdots L_{-n_k}|h\rangle. \end{aligned} \tag{11.123}$$

(1) Move L_0 to the right through each factor, picking up one commutator each time.
(2) $[L_0,L_{-n}]=nL_{-n}$ and $L_0|h\rangle = h|h\rangle$.

In Nash's notation (11.44), L_0 has eigenvalue $h+\sum_i i\alpha_i$ on $L_{-1}^{\alpha_1}\cdots L_{-s}^{\alpha_s}|h\rangle$. The integer $N=\sum_j n_j=\sum_i i\alpha_i$ is the **level** of the descendant.

Different levels are orthogonal. Nash: "Since the Virasoro algebra representation is unitary, the eigenvectors of L_0 are mutually orthogonal." The fact needed is only that L_0 is Hermitian, which is (11.119). For $|a\rangle, |b\rangle$ at levels $N_a\neq N_b$:

$$(h+N_a)\langle b|a\rangle \overset{(1)}{=} \langle b|L_0|a\rangle \overset{(2)}{=} \big(\langle a|L_0^\dagger|b\rangle\big)^* \overset{(3)}{=} \big(\langle a|L_0|b\rangle\big)^* \overset{(4)}{=} (h+N_b)\langle b|a\rangle,$$

(1) $|a\rangle$ is an L_0 eigenvector.
(2) Definition of the adjoint.
(3) $L_0^\dagger = L_0$.
(4) $|b\rangle$ is an eigenvector with real eigenvalue $h+N_b$.

so $(N_a-N_b)\langle b|a\rangle = 0$ and $\langle b|a\rangle = 0$. The space of descendants is therefore an orthogonal direct sum of finite-dimensional *levels*. This is the "grading" of Nash's footnote 3: $[\phi] = \bigoplus_{N\geq 0}[\phi]^N$.

How many states per level? At level N the states (11.122) correspond to the ways of writing $N = n_1 + \cdots + n_k$ with $n_1 \geq \cdots \geq n_k \geq 1$: the partitions of N. So at most $p(N)$ independent states:

N	states	$p(N)$
0	$\|h\rangle$	1
1	$L_{-1}\|h\rangle$	1
2	$L_{-2}\|h\rangle$, $L_{-1}^2\|h\rangle$	2
3	$L_{-3}\|h\rangle$, $L_{-2}L_{-1}\|h\rangle$, $L_{-1}^3\|h\rangle$	3
4	L_{-4}, $L_{-3}L_{-1}$, L_{-2}^2, $L_{-2}L_{-1}^2$, L_{-1}^4 (on $\|h\rangle$)	5

The generating function is $\sum_N p(N)q^N = \prod_{n\geq 1}(1-q^n)^{-1}$, the same as for the boson Fock space of Figure 11.18: each L_{-n} is like an independent oscillator of frequency n. Whether these $p(N)$ states are really linearly independent, and whether they have positive norm, depends on (c, h). That question occupies Sections 11.4.9 and 11.4.10.

Descendant fields (Nash's (11.40)–(11.42)). Every state is made by a field. The field that makes $L_{-n}|h\rangle$ is found by moving the contour of L_{-n} from around 0 to around the insertion point z of ϕ. Define, for any field ϕ at z,

$$\big(L_n\phi\big)(z) \equiv L_n(z)\phi(z) := \oint_z \frac{\mathrm{d}w}{2\pi i}\,(w-z)^{n+1}\,T(w)\,\phi(z), \tag{11.124}$$

Nash's (11.41). At $z = 0$ this is the L_n of (11.94) acting on $\phi(0)$, $L_n(0) = L_n$, and applying it to $|0\rangle$ gives the state: $(L_{-n}\phi)(0)|0\rangle = L_{-n}|h\rangle$. Comparing (11.124) with the Laurent expansion of Section 11.3.12, $(L_n\phi)(z)$ is just the coefficient $T_n(\phi)$ of $(w-z)^{-n-2}$ in $T(w)\phi(z)$. So the full OPE of T with ϕ is

$$T(w)\,\phi(z) = \sum_{n\in\mathbb{Z}} \frac{(L_n\phi)(z)}{(w-z)^{n+2}} = \frac{h\,\phi(z)}{(w-z)^2} + \frac{(L_{-1}\phi)(z)}{w-z} + (L_{-2}\phi)(z) + (w-z)(L_{-3}\phi)(z) + \cdots \tag{11.125}$$

for a primary (where $L_{n\geq 1}\phi = 0$). This is Nash's (11.45), and its first two terms are (11.26). The whole OPE of T with ϕ, including the regular part, is a list of the descendants of ϕ.

Watch out: the summation in Nash's (11.45)

Nash prints (11.45) as $\sum_{n=0}^{\infty}(w-z)^{-n-2}L_n(w)\phi(z)$. Read literally (sum over $n \geq 0$) that would contain only $L_0, L_1, L_2, \ldots$, i.e. only $h\phi/(w-z)^2$ for a primary, and lose the derivative term. The correct range is $n \in \mathbb{Z}$ (equivalently $n \leq 0$ for a primary, i.e. $\sum_{k\geq 0}(w-z)^{k-2}L_{-k}\phi$), with the argument z, not w, in L_n, as in (11.125).

Worked example 11.41: The first descendants, as fields

(a) L_{-1} is the derivative. From (11.124) with $n = -1$ and the OPE,

$$(L_{-1}\phi)(z) \overset{(1)}{=} \oint_z \frac{\mathrm{d}w}{2\pi i}\,T(w)\phi(z) \overset{(2)}{=} \oint_z \frac{\mathrm{d}w}{2\pi i}\left(\frac{h\phi}{(w-z)^2} + \frac{\partial\phi}{w-z}\right) = \partial\phi(z).$$

(1) $(w-z)^{n+1} = 1$ for $n = -1$.
(2) Only the simple pole has a residue.

So the derivative of a primary is its level-1 descendant. It is not itself primary (Nash: "the derivative of a primary field is not primary"): by (11.102), $L_1L_{-1}|h\rangle = [L_1, L_{-1}]|h\rangle = 2L_0|h\rangle = 2h|h\rangle \neq 0$ unless $h = 0$.
(b) The descendants of the identity. The identity field I has $h = 0$, and $T(w)I = T(w)$ has no

singular terms. Taylor expanding $T(w) = \sum_{k\ge 0} \frac{(w-z)^k}{k!}\partial^k T(z)$,

$$(L_{-n}I)(z) \overset{(1)}{=} \oint_z \frac{\mathrm{d}w}{2\pi i}\,(w-z)^{-n+1}\,T(w) \overset{(2)}{=} \sum_{k\ge 0} \frac{\partial^k T(z)}{k!} \oint_z \frac{\mathrm{d}w}{2\pi i}\,(w-z)^{k-n+1}$$
$$\overset{(3)}{=} \frac{\partial^{n-2}T(z)}{(n-2)!} \quad (n \ge 2),$$

(1) (11.124) with $\phi = I$.
(2) Taylor series.
(3) The integral is 1 when $k - n + 1 = -1$, i.e. $k = n - 2$.

and $(L_{-1}I) = 0$, $L_0 I = 0$. So $T = L_{-2}I$ is a descendant of the identity, and so are all its derivatives: Nash's remark that "$T(z)$ is a secondary field belonging to $[I]$". In the state language this is $L_{-2}|0\rangle = T(0)|0\rangle$.

(c) A level-2 descendant of the boson. In the free boson take $\phi = j = i\sqrt{2}\partial\varphi$ ($h = 1$). Its state is $a_{-1}|0\rangle$ (since $j(0)|0\rangle = \sum_n a_n z^{-n-1}|0\rangle|_{z=0}$ keeps only a_{-1}). Then

$$L_{-1}a_{-1}|0\rangle \overset{(1)}{=} [L_{-1}, a_{-1}]|0\rangle + a_{-1}L_{-1}|0\rangle \overset{(2)}{=} a_{-2}|0\rangle,$$
$$L_{-2}a_{-1}|0\rangle \overset{(3)}{=} [L_{-2}, a_{-1}]|0\rangle + a_{-1}L_{-2}|0\rangle \overset{(4)}{=} a_{-3}|0\rangle + \tfrac{1}{2}a_{-1}^3|0\rangle.$$

(1) Commute L_{-1} to the right.
(2) $[L_m, a_n] = -n a_{m+n}$ gives $[L_{-1}, a_{-1}] = a_{-2}$; and $L_{-1}|0\rangle = 0$.
(3) Same move.
(4) $[L_{-2}, a_{-1}] = a_{-3}$ and $L_{-2}|0\rangle = \frac{1}{2}a_{-1}^2|0\rangle$ (Example 11.39).

The first is the state of ∂j, as it must be. The Virasoro descendants of j are particular combinations of Fock states.

Descendant fields in general (Nash's (11.42)–(11.43)). Iterating (11.124), the field of the state $L_{-n_1}\cdots L_{-n_k}|h\rangle$ is $(L_{-n_1}\cdots L_{-n_k}\phi)(z)$, defined by nested contours, the innermost (L_{-n_k}) closest to z. Including the antiholomorphic generators,

$$\big(\bar L_{-\bar n_1}\cdots \bar L_{-\bar n_l}\, L_{-n_1}\cdots L_{-n_k}\phi\big)(z,\bar z) \quad\longleftrightarrow\quad \bar L_{-\bar n_1}\cdots \bar L_{-\bar n_l}\, L_{-n_1}\cdots L_{-n_k}|h,\bar h\rangle,$$

of weights $(h + \sum n_i,\ \bar h + \sum \bar n_j)$. The set of all of these, the primary together with all its descendants, is the **conformal family** $[\phi]$.

Working definition 11.13 (conformal family)

The **conformal family** $[\phi]$ of a primary field ϕ of weights $(h,\bar h)$ is the set of fields $(\bar L_{-\bar n_1}\cdots L_{-n_1}\cdots\phi)$; equivalently the representation of $Vir \oplus \overline{Vir}$ spanned by $|h,\bar h\rangle$ and its descendants [the states made from $|h,\bar h\rangle$ by any string of L_{-n} and $\bar L_{-\bar n}$ with $n,\bar n \ge 1$]. *Operational use:* the Ward identities [they trade each L_{-n} acting on a field for derivatives in the positions of the other fields] express every correlation function of descendants as a differential operator acting on the correlation function of the primaries, so a CFT is specified by its primaries and their correlators.

Operator products inside families. Nash remarks that the product of fields in a conformal family is again expressible through fields in the family, using $T(w)T(z)$ as the example. The precise statement, which is what the TT example shows, is that *fields of the identity family* $[I]$ (that

is, T, its derivatives and its normal-ordered products) act on any family $[\phi]$ without leaving it: the OPE (11.125) of T with ϕ involves only descendants of ϕ, and the TT OPE (11.89) involves only I, T and ∂T, all in $[I]$.

Watch out: two fields from the same family

Nash's sentence "if we take any two fields A and B in the same conformal family $[\phi]$ then the product $A(w)B(z)$ is expressible in a like manner over the fields in $[\phi]$" is true for $[\phi] = [I]$ but not in general. In the Ising model the spin field σ (weight $\frac{1}{16}$) has $\sigma \times \sigma = [I] + [\varepsilon]$: its OPE with itself contains the identity family and the energy family, not the family $[\sigma]$. In general the families appearing in $[\phi_i] \times [\phi_j]$ are governed by the **fusion rules**, which are the subject of Nash's §5 (the next part of these notes). What is always true is that $[I] \times [\phi] = [\phi]$.

Reading Nash (pp. 307–310, (11.33)–(11.45))

Nash defines the modes L_n (11.33), states that the TT OPE implies the Virasoro algebra (11.34) (we did the contour computation, showed via the Jacobi identity that c is the only possible central term, and verified $c = 1$ with oscillators), and adds the second copy (11.35). He identifies the Hamiltonian with $L_0 + \bar{L}_0$ and replaces time ordering by radial ordering in one sentence; we set up radial quantisation, the contour commutator trick, Hermitian conjugation $L_n^\dagger = L_{-n}$ and the cylinder Hamiltonian $L_0 + \bar{L}_0 - c/12$. His (11.36)–(11.38) define $|h\rangle = \phi(0)|0\rangle$ and assert (11.37); we derived it from the $T\phi$ OPE and the regularity of the vacuum, and also from boundedness of the energy as in his Chapter 6. (11.39)–(11.44) introduce descendant states and fields, the operators $L_n(z)$, the conformal family and the level; (11.45) is the full $T\phi$ OPE, whose printed summation range should be $n \in \mathbb{Z}$.

11.4.9 Verma modules: from one state to a representation

Nash's sentence is short: the tower of states $L_{-1}^{\alpha_1} \cdots L_{-s}^{\alpha_s}|h\rangle$ "comprise[s] an irreducible representation of the Virasoro algebra". Two pages later he adds that "there exist non-trivial null states which have zero norm" and that these "must be projected out in order to obtain an irreducible representation". These two statements pull in different directions. The first says the tower is irreducible. The second says that sometimes it is not, and something has to be removed. This subsection makes the construction precise. It is the most important piece of algebra in the chapter, so we go slowly and do one thing at a time.

A change of viewpoint: from a vector space to a module. In quantum mechanics we are used to starting from a Hilbert space $\mathcal{H}$ and then asking which operators act on it. For SU(2), for instance, we are handed the space $\mathbb{C}^{2j+1}$ and the spin matrices on it. The highest-weight construction runs the other way. We start with the *algebra* and one vector $|h\rangle$, and we *build* the space: it is whatever the algebra can produce from $|h\rangle$. What comes out is not a bare vector space. It is a vector space together with a rule saying how every element of the algebra acts on it. Mathematicians call such a thing a module.

Working definition 11.14 (module = representation space)

Let $\mathfrak{g}$ be a Lie algebra [in this chapter the Virasoro algebra, spanned by the L_n and c with their commutators]. A **$\mathfrak{g}$-module** is a complex vector space V together with a linear map $X \mapsto \rho(X) \in \mathrm{End}(V)$ [a linear operator from V to itself, a matrix once a basis is chosen]

such that

$$\rho\big([X,Y]\big) = \rho(X)\rho(Y) - \rho(Y)\rho(X) \qquad \text{for all } X, Y \in \mathfrak{g}.$$

Operational test: write down the action of each generator on a basis, and check that the commutators of the resulting matrices reproduce the brackets of $\mathfrak{g}$. A **submodule** is a subspace $W \subset V$ with $\rho(X)W \subset W$ for all X [acting with any generator on a vector of W gives a vector still in W]. A module is **irreducible** if its only submodules are $\{0\}$ and V.

Dictionary: the same object, two names

mathematician	physicist
$\mathfrak{g}$-module V	representation space; space of states on which $\mathfrak{g}$ acts
submodule	invariant subspace
irreducible module	irreducible representation (irrep)
module homomorphism	intertwiner; linear map commuting with $\mathfrak{g}$
cyclic vector	a state from which all others are reached by the generators

The point to keep in mind is that *a module need not come with an inner product*. The Virasoro relations, $L_0|h\rangle = h|h\rangle$ and $L_n|h\rangle = 0$ for $n \geq 1$, say nothing about lengths of vectors. The inner product is an extra structure that we add later, in Step 4 below, and it is at that step that the physics of positive norms (and hence of Hilbert spaces) enters. Keeping the two layers apart is what makes null states easy to understand.

Key idea: the pipeline

$$\underbrace{|h\rangle}_{\substack{\text{highest-weight}\\ \text{state}}} \xrightarrow{\substack{\text{generate}\\ \text{descendants}}} \underbrace{V(c,h) = U(vir_-)|h\rangle}_{\text{Verma module}} \xrightarrow{\substack{\text{quotient by the}\\ \text{null submodule } N}} \underbrace{L(c,h) = V(c,h)/N}_{\text{irreducible module}} .$$

The Verma module contains everything the Virasoro algebra can generate from $|h\rangle$, subject only to the highest-weight conditions. At special values of (c,h) it contains a *singular vector*, a descendant that is itself highest weight. That vector generates a whole submodule N of zero-norm states. We remove the entire submodule, not just the one vector, and what is left is the irreducible representation $L(c,h)$ that a CFT actually uses. For generic (c,h) there is nothing to remove and $V(c,h) = L(c,h)$.

Step 1: the Verma module

Split the Virasoro generators into three groups according to the sign of the index:

$$vir_- = \text{span}\{L_{-1}, L_{-2}, L_{-3}, \dots\}, \qquad vir_0 = \text{span}\{L_0, c\}, \qquad vir_+ = \text{span}\{L_1, L_2, L_3, \dots\}. \tag{11.126}$$

Each piece is a Lie subalgebra (for instance $[L_{-m}, L_{-n}] = (n-m)L_{-m-n}$ has negative index again, and the central term never appears because $-m-n \neq 0$). On a highest-weight state, vir_+ acts as zero and vir_0 acts by the numbers h and c. Only vir_- produces anything new.

The enveloping algebra, concretely. $U(vir_-)$, the **universal enveloping algebra** of vir_-, is a precise name for "all polynomials in the L_{-n}". Its elements are linear combinations of products

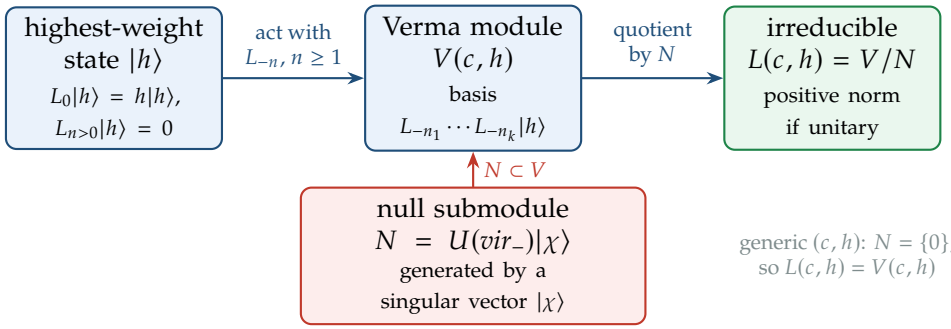


Figure 11.20: The construction of an irreducible highest-weight representation of the Virasoro algebra, in three stages. Read left to right. The only choice we make is (c,h). The Verma module is then completely determined, and the null submodule N (red) is whatever part of it happens to have zero norm. Removing N gives the representation that appears in a conformal field theory.

$L_{-n_1}L_{-n_2}\cdots L_{-n_k}$, and the only rule for manipulating them is that we may swap two neighbours at the cost of their commutator, $L_{-m}L_{-n} = L_{-n}L_{-m} + (n-m)L_{-m-n}$. Using this rule repeatedly, any product can be rewritten as a combination of *ordered* products with $n_1 \geq n_2 \geq \cdots \geq n_k$. For example

$$L_{-1}L_{-2} \overset{(1)}{=} L_{-2}L_{-1} + [L_{-1}, L_{-2}] \overset{(2)}{=} L_{-2}L_{-1} + L_{-3}.$$

(1) Swap the two factors and add the commutator.
(2) $[L_{-1}, L_{-2}] = (-1+2)L_{-3}$ by (11.102).

The **Poincaré–Birkhoff–Witt theorem** says that the ordered products are linearly independent in $U(vir_-)$: there are no relations among them other than those that follow from the commutators. So they form a basis.

Working definition 11.15 (Verma module)

Fix numbers c and h. The **Verma module** $V(c,h)$ is the vector space with basis

$$|h\rangle, \qquad L_{-n_1}L_{-n_2}\cdots L_{-n_k}|h\rangle \quad (n_1 \geq n_2 \geq \cdots \geq n_k \geq 1),$$

on which the Virasoro algebra acts as follows: L_{-n} $(n \geq 1)$ acts by multiplying on the left and then reordering [using $L_{-m}L_{-n} = L_{-n}L_{-m} + (n-m)L_{-m-n}$ until the indices decrease]; L_0 and $L_{n\geq 1}$ act by commuting them to the right with (11.102) until they hit $|h\rangle$, where $L_0|h\rangle = h|h\rangle$ and $L_{n\geq 1}|h\rangle = 0$; the central element acts as the number c. In one line, $V(c,h) = U(vir_-)|h\rangle$. *Operational test:* the dimension of the level-N subspace V_N [spanned by the basis vectors with $n_1 + \cdots + n_k = N$, all with L_0 eigenvalue $h + N$] is the number of partitions $p(N)$ [$p(1) = 1$, $p(2) = 2$, $p(3) = 3$, $p(4) = 5$], since the basis vectors at level N are labelled by partitions $N = n_1 + \cdots + n_k$.
Formal definition, for reference: $V(c,h) = U(Vir) \otimes_{U(vir_0 \oplus vir_+)} \mathbb{C}_{c,h}$ [all products of generators applied to one vector, where any L_0, $L_{n\geq 1}$ or c reaching that vector is replaced by its number], where $\mathbb{C}_{c,h}$ is the one-dimensional module on which vir_+ acts as 0, L_0 as h and the central element as c.

Worked example 11.42: The Verma module at levels 0, 1, 2, written out

Let us exhibit the representation. At levels $0, 1, 2$ the bases are

$$V_0 : |h\rangle, \qquad V_1 : L_{-1}|h\rangle, \qquad V_2 : L_{-2}|h\rangle,\ L_{-1}^2|h\rangle.$$

L_{-1} and L_{-2} just stack on the left. The interesting part is the action of the lowering operators

L_1 and L_2, which we compute by commuting them to the right. First L_1 on level 1:

$$L_1L_{-1}|h\rangle \overset{(1)}{=} [L_1, L_{-1}]|h\rangle + L_{-1}L_1|h\rangle \overset{(2)}{=} 2L_0|h\rangle = 2h|h\rangle.$$

(1) Commute L_1 to the right.
(2) $[L_1, L_{-1}] = 2L_0$ and $L_1|h\rangle = 0$.

Next L_1 on the two level-2 states:

$$L_1L_{-2}|h\rangle \overset{(1)}{=} [L_1, L_{-2}]|h\rangle \overset{(2)}{=} 3L_{-1}|h\rangle,$$
$$L_1L_{-1}^2|h\rangle \overset{(3)}{=} [L_1, L_{-1}]L_{-1}|h\rangle + L_{-1}L_1L_{-1}|h\rangle \overset{(4)}{=} 2L_0L_{-1}|h\rangle + 2h\,L_{-1}|h\rangle \overset{(5)}{=} (4h+2)\,L_{-1}|h\rangle.$$

(1) $L_1|h\rangle = 0$, so only the commutator survives.
(2) $[L_1, L_{-2}] = (1+2)L_{-1}$.
(3) Commute L_1 past the first L_{-1}.
(4) $[L_1, L_{-1}] = 2L_0$, and the previous computation for $L_1L_{-1}|h\rangle$.
(5) $L_0L_{-1}|h\rangle = (h+1)L_{-1}|h\rangle$ by (11.123).

And L_2 on the level-2 states:

$$L_2L_{-2}|h\rangle \overset{(1)}{=} [L_2, L_{-2}]|h\rangle \overset{(2)}{=} \left(4L_0 + \frac{c}{12}\cdot 2\cdot 3\right)|h\rangle = \left(4h + \frac{c}{2}\right)|h\rangle,$$
$$L_2L_{-1}^2|h\rangle \overset{(3)}{=} [L_2, L_{-1}]L_{-1}|h\rangle + L_{-1}[L_2, L_{-1}]|h\rangle \overset{(4)}{=} 3L_1L_{-1}|h\rangle + 3L_{-1}L_1|h\rangle \overset{(5)}{=} 6h|h\rangle.$$

(1) $L_2|h\rangle = 0$.
(2) $[L_2, L_{-2}] = 4L_0 + \frac{c}{12}\,2(2^2-1)$.
(3) Commute L_2 through both factors; the term $L_{-1}^2L_2|h\rangle$ vanishes.
(4) $[L_2, L_{-1}] = 3L_1$.
(5) $L_1L_{-1}|h\rangle = 2h|h\rangle$ and $L_1|h\rangle = 0$.

In matrix form, with the ordered bases above,

$$L_1\big|_{V_1\to V_0} = \begin{pmatrix} 2h \end{pmatrix}, \qquad L_1\big|_{V_2\to V_1} = \begin{pmatrix} 3 & 4h+2 \end{pmatrix}, \qquad L_2\big|_{V_2\to V_0} = \begin{pmatrix} 4h+\frac{c}{2} & 6h \end{pmatrix},$$

while $L_{-1}\big|_{V_1\to V_2} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ and $L_{-2}\big|_{V_0\to V_2} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$. As a check of the representation property, compare $[L_1, L_{-1}]$ with $2L_0$ on V_1:

$$\big(L_1L_{-1} - L_{-1}L_1\big)L_{-1}|h\rangle \overset{(1)}{=} L_1L_{-1}^2|h\rangle - L_{-1}\,(2h|h\rangle) \overset{(2)}{=} (4h+2-2h)\,L_{-1}|h\rangle \overset{(3)}{=} 2L_0\,L_{-1}|h\rangle.$$

(1) Definitions of the two products.
(2) The matrices above.
(3) $2L_0 = 2(h+1)$ on V_1.

Why "universal". Suppose W is *any* representation of *Vir* with central charge c containing a highest-weight vector $|h\rangle_W$. Then there is a map

$$\pi : V(c,h) \to W, \qquad L_{-n_1}\cdots L_{-n_k}|h\rangle \mapsto L_{-n_1}\cdots L_{-n_k}|h\rangle_W.$$

It is well defined because the basis of V has no relations to respect, and it commutes with the Virasoro action because both sides are computed by the same commutators. Its image is the submodule of W generated by $|h\rangle_W$. So every highest-weight representation is a quotient of

the Verma module: $V(c,h)$ is the biggest possible one, and all others are obtained from it by setting some combinations to zero. The remaining steps find out which combinations.

Step 2: singular vectors

Working definition 11.16 (singular vector)

A nonzero vector $|\chi\rangle \in V(c,h)$ at level $N \geq 1$ [a combination of basis vectors $L_{-n_1}\cdots L_{-n_k}|h\rangle$ with $n_1+\cdots+n_k = N$] is a **singular vector** (or null vector [once the inner product is defined, it is orthogonal to every vector of $V(c,h)$, itself included]) if it is itself highest weight:

$$L_n|\chi\rangle = 0 \quad (n \geq 1), \qquad L_0|\chi\rangle = (h+N)|\chi\rangle.$$

Operational test: it is enough to check $L_1|\chi\rangle = 0$ and $L_2|\chi\rangle = 0$, because L_1 and L_2 generate all of vir_+ [every L_n with $n \geq 3$ is a repeated commutator of L_1 and L_2] (see the working definition of a highest-weight state in Section 11.4.7). The L_0 condition is automatic for a vector of definite level.

Level 1. The only candidate is $|\chi\rangle = L_{-1}|h\rangle$. It is annihilated by L_2 automatically ($L_2L_{-1}|h\rangle = 3L_1|h\rangle = 0$), and $L_1L_{-1}|h\rangle = 2h|h\rangle$. So $L_{-1}|h\rangle$ is singular exactly when $h = 0$. This is the vacuum: we will see in Section 11.4.13 that $L_{-1}|0\rangle$ is the null vector Nash computes in (11.47).

Level 2. The general level-2 vector is, up to normalisation,

$$|\chi\rangle = \big(L_{-2} + \alpha\, L_{-1}^2\big)|h\rangle. \tag{11.127}$$

(The other possibility, a pure $L_{-1}^2|h\rangle$, is singular only if $h = 0$ and then it is a descendant of the level-1 singular vector.) Impose the two conditions, using the matrices of Example 11.42:

$$L_1|\chi\rangle \overset{(1)}{=} L_1L_{-2}|h\rangle + \alpha\, L_1L_{-1}^2|h\rangle \overset{(2)}{=} \big(3+\alpha(4h+2)\big)L_{-1}|h\rangle, \tag{11.128}$$

$$L_2|\chi\rangle \overset{(3)}{=} L_2L_{-2}|h\rangle + \alpha\, L_2L_{-1}^2|h\rangle \overset{(4)}{=} \Big(4h + \frac{c}{2} + 6\alpha h\Big)|h\rangle. \tag{11.129}$$

(1) Linearity.
(2) $L_1L_{-2}|h\rangle = 3L_{-1}|h\rangle$ and $L_1L_{-1}^2|h\rangle = (4h+2)L_{-1}|h\rangle$.
(3) Linearity.
(4) $L_2L_{-2}|h\rangle = (4h+\frac{c}{2})|h\rangle$ and $L_2L_{-1}^2|h\rangle = 6h|h\rangle$.

Both must vanish. The first fixes α, the second then fixes c in terms of h:

$$3+\alpha(4h+2) = 0 \overset{(1)}{\Rightarrow} \alpha = -\frac{3}{2(2h+1)}, \tag{11.130}$$

$$4h + \frac{c}{2} + 6\alpha h = 0 \overset{(2)}{\Rightarrow} \frac{c}{2} = -4h + \frac{9h}{2h+1} \overset{(3)}{\Rightarrow} c = \frac{2h(5-8h)}{2h+1}. \tag{11.131}$$

(1) Solve the linear equation ($h \neq -\frac{1}{2}$).
(2) Insert α: $6\alpha h = -\frac{9h}{2h+1}$.
(3) Multiply by 2 and use the common denominator $2h+1$: $c = \frac{2\left(-4h(2h+1)+9h\right)}{2h+1} = \frac{2(5h-8h^2)}{2h+1}$.

Equivalently, clearing the denominator,

$$16h^2 + 2(c-5)h + c = 0. \tag{11.132}$$

So a level-2 singular vector exists precisely on this curve in the (c,h) plane. For each c there are two roots h; we will meet them again as $h_{1,2}(c)$ and $h_{2,1}(c)$ in the Kac determinant.

Worked example 11.43: The two Ising weights lie on the level-2 curve

Put $c = \frac{1}{2}$ into (11.132):

$$16h^2 - 9h + \tfrac{1}{2} = 0 \overset{(1)}{=} h = \frac{9 \pm \sqrt{81 - 32}}{32} \overset{(2)}{=} h = \frac{9 \pm 7}{32} = \tfrac{1}{2} \text{ or } \tfrac{1}{16}.$$

(1) Quadratic formula for $16h^2 - 9h + \frac{1}{2} = 0$; the discriminant is $81 - 4 \cdot 16 \cdot \frac{1}{2} = 49$.
(2) $\sqrt{49} = 7$.

These are exactly the weights of the energy field ε ($h = \frac{1}{2}$) and the spin field σ ($h = \frac{1}{16}$) of the critical Ising model, whose scaling dimensions $2h = 1$ and $2h = \frac{1}{8}$ we extracted from the critical exponents in Example 11.3. The corresponding singular vectors are, from (11.130),

$$|\chi_\varepsilon\rangle = \left(L_{-2} - \tfrac{3}{4}L_{-1}^2\right)|\tfrac{1}{2}\rangle, \qquad |\chi_\sigma\rangle = \left(L_{-2} - \tfrac{4}{3}L_{-1}^2\right)|\tfrac{1}{16}\rangle.$$

A singular vector generates a whole submodule. A singular vector behaves in every respect like a new highest-weight state of weight $h + N$. So we can run Step 1 again starting from it: the negative modes acting on $|\chi\rangle$ produce

$$N_\chi := U(vir_-)|\chi\rangle = \text{span}\{|\chi\rangle,\ L_{-1}|\chi\rangle,\ L_{-2}|\chi\rangle,\ L_{-1}^2|\chi\rangle,\ L_{-3}|\chi\rangle, \dots\} \subset V(c,h). \tag{11.133}$$

This is a submodule: it is closed under all of *Vir*, not only under vir_-. For a positive mode acting on a typical element,

$$\begin{aligned} L_n\, L_{-k_1} \cdots L_{-k_j}|\chi\rangle &\overset{(1)}{=} \left(\sum_{i=1}^{j} L_{-k_1} \cdots [L_n, L_{-k_i}] \cdots L_{-k_j} + L_{-k_1} \cdots L_{-k_j}\, L_n\right)|\chi\rangle \\ &\overset{(2)}{=} \sum_{i=1}^{j} L_{-k_1} \cdots \left((n+k_i)L_{n-k_i} + \tfrac{c}{12}n(n^2-1)\delta_{n,k_i}\right) \cdots L_{-k_j}|\chi\rangle \\ &\overset{(3)}{=} (\text{an element of } U(vir_-))\,|\chi\rangle \ \in N_\chi. \end{aligned}$$

(1) Move L_n to the right, one commutator per factor.
(2) The Virasoro algebra, and $L_n|\chi\rangle = 0$.
(3) Each term again contains fewer lowering operators and one mode L_{n-k_i}. If $n - k_i < 0$ it is a lowering operator; if $n - k_i \geq 0$ we repeat the procedure, which terminates because each round reduces the number of factors. Whenever $L_{m\geq 1}$ reaches $|\chi\rangle$ it gives zero, L_0 gives the number $h + N$, and the central element gives c.

For example, at the first level of N_χ:

$$L_1 L_{-1}|\chi\rangle = 2L_0|\chi\rangle = 2(h+2)|\chi\rangle, \qquad L_2 L_{-1}|\chi\rangle = 3L_1|\chi\rangle = 0.$$

So a single singular vector is the seed of a whole copy of a highest-weight module, sitting inside $V(c,h)$ and starting at level N.

Step 3: the quotient, and why it is still a module

If $|\chi\rangle$ is a null state in the physical theory, it is not an independent state. Worse, since the algebra acts on it, all of $L_{-1}|\chi\rangle, L_{-2}|\chi\rangle, \dots$ are equally redundant. The whole submodule has to go. The mathematical tool for "setting a subspace to zero" is the quotient.

Working definition 11.17 (quotient module)

Let $N \subset V$ be a submodule [for example N_χ, spanned by all $L_{-k_1}\cdots L_{-k_j}|\chi\rangle$, which no L_n can leave]. Declare $v \sim v'$ if $v - v' \in N$. The equivalence classes $[v] = v + N$ [$[v]$ is the set of all vectors that differ from v by an element of N] form the vector space V/N, with $[v] + [w] = [v + w]$ and $\lambda[v] = [\lambda v]$. The algebra acts by

$$L_n[v] := [L_n v]. \tag{11.134}$$

Operational use: to compute in V/N, compute in V and then drop everything in N. In the example, $|\chi\rangle \sim 0$, $L_{-1}|\chi\rangle \sim 0$, $L_{-2}|\chi\rangle \sim 0$, and so on; in particular $L_{-2}|h\rangle \sim -\alpha\, L_{-1}^2|h\rangle$.

The definition (11.134) hides a question: the class $[v]$ has many representatives, so is $[L_n v]$ independent of which one we pick? It is, precisely because N is a submodule. Let $v \sim v'$. Then

$$L_n v - L_n v' \overset{(1)}{=} L_n(v - v') \overset{(2)}{\in} L_n N \overset{(3)}{\subset} N, \qquad \text{so} \qquad [L_n v] \overset{(4)}{=} [L_n v'].$$

(1) L_n is linear.
(2) $v - v' \in N$ by definition of $v \sim v'$.
(3) N is a submodule, i.e. closed under every L_n (and under c).
(4) Two vectors whose difference lies in N define the same class.

The commutation relations survive too, since they hold in V before we take classes:

$$L_m L_n[v] - L_n L_m[v] \overset{(1)}{=} [L_m L_n v - L_n L_m v] \overset{(2)}{=} \Big[(m-n)L_{m+n}v + \tfrac{c}{12}m(m^2-1)\delta_{m+n,0}\, v\Big]$$
$$\overset{(3)}{=} (m-n)L_{m+n}[v] + \tfrac{c}{12}m(m^2-1)\delta_{m+n,0}\,[v].$$

(1) Apply (11.134) twice, and linearity of $[\,\cdot\,]$.
(2) The Virasoro algebra holds on V.
(3) (11.134) again.

So V/N is a Virasoro module. Had we quotiented by the single vector $|\chi\rangle$ (the one-dimensional subspace $\mathbb{C}|\chi\rangle$), step (3) would fail: $L_{-1}|\chi\rangle \notin \mathbb{C}|\chi\rangle$, so L_{-1} would send the class $[0] = [|\chi\rangle]$ to two different classes, $[0]$ and $[L_{-1}|\chi\rangle] \neq [0]$. The action would not be defined. That is why we quotient by the submodule *generated* by the singular vector.

Step 4: where the inner product enters

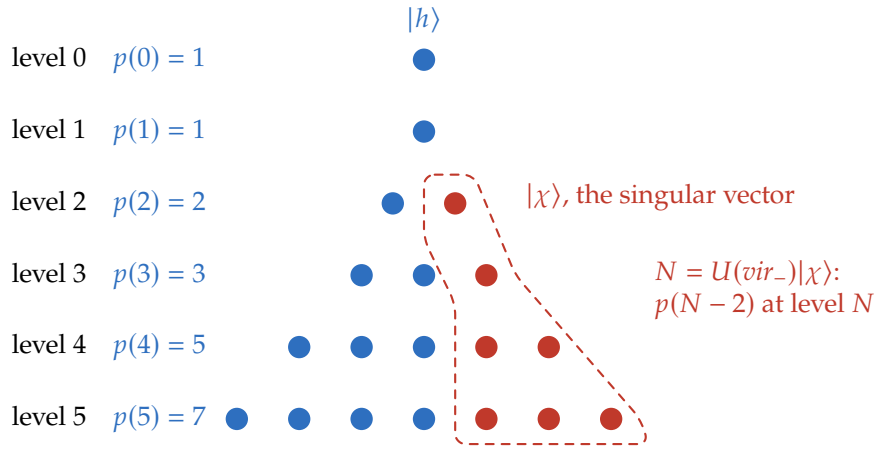
So far there has been no inner product anywhere. Now we add it. The physics input is the Hermiticity property $L_n^\dagger = L_{-n}$ that we derived from radial quantisation (11.119), and the normalisation $\langle h|h\rangle = 1$.

Working definition 11.18 (Shapovalov form)

The **Shapovalov form** (or Kac form) on $V(c,h)$ is the Hermitian form [a pairing $\langle u|v\rangle$, linear in v, with $\langle v|u\rangle = \overline{\langle u|v\rangle}$, but not assumed positive] defined on basis vectors by

$$\big\langle L_{-m_1}\cdots L_{-m_j}h \,\big|\, L_{-n_1}\cdots L_{-n_k}h\big\rangle := \langle h|\, L_{m_j}\cdots L_{m_1}\, L_{-n_1}\cdots L_{-n_k}|h\rangle,$$

where the right side is evaluated by commuting the raising operators L_m [the L_m with $m \geq 1$. Each lowers the level by m and kills $|h\rangle$] to the right until the result is a multiple of $|h\rangle$, and taking that multiple (with $\langle h|h\rangle = 1$). It is the unique Hermitian form with $\langle h|h\rangle = 1$ for which $L_n^\dagger = L_{-n}$ [an L_{-n} on the ket side can be moved to the bra side, where



$$\dim V_N = p(N) = 1, 1, 2, 3, 5, 7, \quad \dim N_N = 0, 0, 1, 1, 2, 3, \quad \dim L_N = 1, 1, 1, 2, 3, 4$$

Figure 11.21: The weight diagram of a Verma module $V(c,h)$, level by level, for (c,h) on the level-2 curve (11.132) but otherwise generic. Each dot is a basis state; level N has $p(N)$ of them. The red dots span the null submodule N generated by the level-2 singular vector $|\chi\rangle$: it looks like a second Verma module whose top sits at level 2, with $p(N-2)$ states at level N. The irreducible module $L(c,h) = V/N$ keeps only the blue dots. (At the minimal-model points, such as $c = \frac{1}{2}$, more singular vectors appear and N is larger; see Example 11.45.)

it becomes L_n]. *Operational use:* the matrices of this form, level by level, are the Gram matrices of Section 11.4.10.

The Shapovalov form is defined by algebra alone. Nothing guarantees that it is positive, or even nondegenerate. Two facts about it are the heart of the matter.

Fact 1: singular vectors have zero norm, and are orthogonal to everything. Let $|\chi\rangle$ be singular at level $N \geq 1$ and let $w = L_{-m_1} \cdots L_{-m_j}|h\rangle$ be any basis vector. If $j \geq 1$,

$$\langle w|\chi\rangle \overset{(1)}{=} \langle h| L_{m_j} \cdots L_{m_2} \left(L_{m_1}|\chi\rangle\right) \overset{(2)}{=} 0.$$

(1) Definition of the form; the adjoint of $L_{-m_1} \cdots L_{-m_j}$ is $L_{m_j} \cdots L_{m_1}$, and L_{m_1} acts first.
(2) $m_1 \geq 1$ and $|\chi\rangle$ is annihilated by every positive mode.

If $j = 0$, $w = |h\rangle$ is at level $0 \neq N$, and states at different levels are orthogonal (Section 11.4.8). So $\langle w|\chi\rangle = 0$ for every w, and in particular $\langle\chi|\chi\rangle = 0$. The same holds for every vector of the submodule N_χ: for $v = L_{-I}|\chi\rangle \in N_\chi$, $\langle w|v\rangle = \langle L_I w|\chi\rangle$ with L_I a product of raising operators, and $L_I w$ is again a combination of basis vectors, so the previous computation applies. A null submodule is a subspace of vectors orthogonal to the whole of V.

Fact 2: the vectors orthogonal to everything form the largest proper submodule. Define the **radical**

$$\mathrm{Rad}\, V := \{v \in V : \langle w|v\rangle = 0 \text{ for all } w \in V\}.$$

It is a submodule: for $v \in \mathrm{Rad}\, V$ and any w,

$$\langle w|L_n v\rangle \overset{(1)}{=} \langle L_{-n} w|v\rangle \overset{(2)}{=} 0.$$

(1) $L_n^\dagger = L_{-n}$.

(2) $L_{-n}w$ is another vector of V, and v is orthogonal to all of them.

Conversely, every proper submodule $M \subsetneq V$ lies inside the radical. The reason is a short chain. First, M is spanned by L_0-eigenvectors (since L_0 acts diagonalisably with finite-dimensional eigenspaces, and M is L_0-invariant), so it splits into levels, $M = \bigoplus_N M_N$. Second, $M_0 = 0$: if M contained $|h\rangle$ it would contain everything, $M = V$. Now take $v \in M_N$ with $N \geq 1$, and a basis vector $w = L_{-I}|h\rangle$ at level N (vectors at other levels are orthogonal to v anyway). Then

$$\langle w|v\rangle \overset{(1)}{=} \langle h|\left(L_I^{\dagger\dagger}\, v\right) \overset{(2)}{=} \langle h|\,(\text{an element of } M_0) \overset{(3)}{=} 0.$$

(1) Definition of the form: $\langle w|v\rangle = \langle h|(L_{-I})^\dagger v$, where $(L_{-I})^\dagger$ is a product of raising operators lowering the level by N.
(2) M is a submodule, so applying modes to $v \in M$ stays in M, and the result is at level 0.
(3) $M_0 = 0$.

So $\operatorname{Rad} V$ contains every proper submodule, and it is itself a proper submodule ($|h\rangle \notin \operatorname{Rad} V$ since $\langle h|h\rangle = 1$). It is the *maximal* proper submodule, and we take $N := \operatorname{Rad} V$.

Key idea: the irreducible module and its inner product

Let $N = \operatorname{Rad} V(c,h)$ be the null submodule. Then

(i) $L(c,h) = V(c,h)/N$ is irreducible: a submodule of the quotient pulls back to a submodule of V containing N, which by maximality is N or V.

(ii) The Shapovalov form descends to $L(c,h)$, by $\langle [w]|[v]\rangle := \langle w|v\rangle$. This is well defined because changing v by an element $n \in N$ changes $\langle w|v\rangle$ by $\langle w|n\rangle = 0$.

(iii) On $L(c,h)$ the form is *nondegenerate*: a class orthogonal to all classes has a representative in $\operatorname{Rad} V = N$, i.e. it is the class $[0]$.

Why unitarity needs the quotient. In a Hilbert space, $\|v\|^2 = 0$ forces $v = 0$. A physical state space therefore cannot contain a nonzero vector of zero norm. If the CFT contains a primary of weight h at a point (c,h) where $V(c,h)$ has singular vectors, the space of states built on ϕ cannot be $V(c,h)$: it must be a quotient in which all of N is zero. This is what Nash means by "projected out". In the field language, the field that would create $|\chi\rangle$ vanishes identically, for example $\left(L_{-2} - \frac{3}{2(2h+1)}L_{-1}^2\right)\phi = 0$ as an operator equation. Then:

- if the descended form on $L(c,h)$ is positive definite, its completion is a Hilbert space and we have a unitary representation;
- if the form on $L(c,h)$ still has a negative-norm vector, no quotient can cure it (the form is already nondegenerate), and the pair (c,h) is excluded from any unitary theory. This is the second half of Nash's remark: "states of negative squared norm can also occur and are disallowed."

The generic case. For generic (c,h) there is no singular vector at any level, the radical is $\{0\}$, and $V(c,h) = L(c,h)$ is already irreducible. We will see in Section 11.4.10 exactly which (c,h) are non-generic: they lie on countably many curves $h = h_{r,s}(c)$ in the (c,h) plane, the zeros of the Kac determinant. Nash's statement that the tower "comprise[s] an irreducible representation" is exactly right off these curves.

	Verma module $V(c,h)$	null submodule N	irreducible module $L(c,h)$
what it is	all formal descendants $L_{-n_1}\cdots L_{-n_k}\|h\rangle$	descendants of the singular vector(s); $=\operatorname{Rad} V$	V/N: Verma module with the null states set to zero
relations imposed	only $L_0\|h\rangle = h\|h\rangle$, $L_{n>0}\|h\rangle = 0$	—	in addition $\|\chi\rangle = 0$ and all its descendants
states at level N	$p(N)$	$\sum$ over singular vectors, e.g. $p(N-2)$ for one at level 2	$p(N) - \dim N_N$
Shapovalov form	may be degenerate or indefinite	identically zero (orthogonal to all of V)	nondegenerate; positive iff unitary
irreducible?	only for generic (c,h)	is itself a highest-weight module	always
role in physics	bookkeeping device, universal object	redundant, unphysical states	the space of states of one conformal family

Table 11.2: The three objects of the construction, side by side.

Worked example 11.44: The vacuum module: quotient by $L_{-1}|0\rangle$

Take $h = 0$. The level-1 computation above shows that $|\chi\rangle = L_{-1}|0\rangle$ is singular. It generates $N_\chi = U(vir_-)L_{-1}|0\rangle$, which has $p(N-1)$ states at level N. For generic c this is the whole radical (it matches the factor $h^{p(N-1)}$ in the Kac determinant, Section 11.4.10), so

$$\dim L(c,0)_N = p(N) - p(N-1):$$

N	0	1	2	3	4	5	6	7
$p(N)$	1	1	2	3	5	7	11	15
$p(N-1)$	0	1	1	2	3	5	7	11
$\dim L(c,0)_N$	1	0	1	1	2	2	4	4

A basis of the quotient is given by the ordered products with no L_{-1} at all, $L_{-n_1}\cdots L_{-n_k}|0\rangle$ with every $n_i \geq 2$: at level 4 these are $L_{-4}|0\rangle$ and $L_{-2}^2|0\rangle$. In field language they are $\frac{1}{2}\partial^2 T$ and $:TT:$, the level-4 fields of the identity family $[I]$ (Example 11.41(b)). The generating function of the dimensions is $\prod_{n\geq 2}(1-q^n)^{-1}$: the factor $(1-q)^{-1}$ that L_{-1} would have contributed is gone.

Worked example 11.45: The free fermion realises the Ising null vector

The free Majorana fermion has $c = \frac{1}{2}$ and contains the primary ψ of weight $\frac{1}{2}$ (Example 11.29). By Example 11.43, the Verma module $V(\frac{1}{2},\frac{1}{2})$ has the singular vector $|\chi_\varepsilon\rangle = (L_{-2} - \frac{3}{4}L_{-1}^2)|\frac{1}{2}\rangle$. In a genuine Hilbert space this vector must be *zero*. Let us check that the fermion Fock space does this automatically.

Modes. Since ψ has weight $\frac{1}{2}$, expand $\psi(z) = \sum_r \psi_r z^{-r-1/2}$, $\psi_r = \oint_0 \frac{dz}{2\pi i} z^{r-1/2}\psi(z)$. Regularity of $\psi(z)|0\rangle$ at $z=0$ requires $\psi_r|0\rangle = 0$ for $r > -\frac{1}{2}$, and single-valuedness around the origin (the Neveu–Schwarz sector, which contains the vacuum) makes $r \in \mathbb{Z} + \frac{1}{2}$. The contour trick (11.100), with an anticommutator because both operators are fermionic, and the OPE $\psi(w)\psi(z) \sim \frac{1}{w-z}$ give

$$\{\psi_r,\psi_s\} \overset{(1)}{=} \oint_0 \frac{dz}{2\pi i} z^{s-1/2} \oint_z \frac{dw}{2\pi i} \frac{w^{r-1/2}}{w-z} \overset{(2)}{=} \oint_0 \frac{dz}{2\pi i} z^{r+s-1} \overset{(3)}{=} \delta_{r+s,0}.$$

(1) Double-contour trick; the radial ordering of two fermions carries the sign that turns the commutator into an anticommutator.

(2) Residue of $w^{r-1/2}/(w-z)$ at $w=z$.
(3) Only z^{-1} has a residue.

In particular $\psi_{-1/2}^2 = \frac{1}{2}\{\psi_{-1/2},\psi_{-1/2}\} = \frac{1}{2}\delta_{-1,0} = 0$. The state of ψ is $|\tfrac{1}{2}\rangle = \psi(0)|0\rangle = \psi_{-1/2}|0\rangle$.

Action of L_m on the modes. Since ψ is primary of weight $\frac{1}{2}$, (11.101) gives $[L_m,\psi(z)] = \frac{1}{2}(m+1)z^m\psi + z^{m+1}\partial\psi$. Taking the $z^{r-1/2}$ moment exactly as in Example 11.38,

$$[L_m,\psi_r] \overset{(1)}{=} \tfrac{1}{2}(m+1)\,\psi_{m+r} - \left(m+r+\tfrac{1}{2}\right)\psi_{m+r} = -\left(r+\frac{m}{2}\right)\psi_{m+r}.$$

(1) The first term is the definition of ψ_{m+r}; the second is integration by parts, $\oint z^{m+r+1/2}\partial\psi = -(m+r+\frac{1}{2})\oint z^{m+r-1/2}\psi$.

The descendants. Using $L_{-1}|0\rangle = 0$:

$$L_{-1}|\tfrac{1}{2}\rangle \overset{(1)}{=} [L_{-1},\psi_{-1/2}]|0\rangle \overset{(2)}{=} \psi_{-3/2}|0\rangle,$$

$$L_{-1}^2|\tfrac{1}{2}\rangle \overset{(3)}{=} [L_{-1},\psi_{-3/2}]|0\rangle \overset{(4)}{=} 2\,\psi_{-5/2}|0\rangle,$$

$$L_{-2}|\tfrac{1}{2}\rangle \overset{(5)}{=} [L_{-2},\psi_{-1/2}]|0\rangle + \psi_{-1/2}L_{-2}|0\rangle \overset{(6)}{=} \tfrac{3}{2}\,\psi_{-5/2}|0\rangle + \beta\,\psi_{-1/2}\psi_{-3/2}\psi_{-1/2}|0\rangle \overset{(7)}{=} \tfrac{3}{2}\,\psi_{-5/2}|0\rangle.$$

(1) Commute L_{-1} to the right; $L_{-1}|0\rangle = 0$.
(2) $-(r+\frac{m}{2}) = -(-\frac{1}{2}-\frac{1}{2}) = 1$ for $m=-1$, $r=-\frac{1}{2}$.
(3) The same step once more.
(4) $-(-\frac{3}{2}-\frac{1}{2}) = 2$.
(5) Commute L_{-2} to the right; now $L_{-2}|0\rangle \neq 0$.
(6) $-(-\frac{1}{2}-1) = \frac{3}{2}$. For the second term: $L_{-2}|0\rangle = T(0)|0\rangle$ is bilinear in ψ, so it is a state of weight 2 with an even number of fermion modes; the only one is $\psi_{-3/2}\psi_{-1/2}|0\rangle$ (two distinct modes with $r_1+r_2=2$), so $L_{-2}|0\rangle = \beta\,\psi_{-3/2}\psi_{-1/2}|0\rangle$ for some number β.
(7) $\psi_{-1/2}\psi_{-3/2}\psi_{-1/2} = -\psi_{-3/2}\psi_{-1/2}^2 = 0$.

Hence

$$|\chi_\varepsilon\rangle = \left(L_{-2} - \tfrac{3}{4}L_{-1}^2\right)|\tfrac{1}{2}\rangle = \left(\tfrac{3}{2} - \tfrac{3}{4}\cdot 2\right)\psi_{-5/2}|0\rangle = 0.$$

The null vector is not merely orthogonal to everything; in the fermion Fock space it simply *is* zero. So the space of states of the free fermion built on ψ is not the Verma module $V(\frac{1}{2},\frac{1}{2})$ but a quotient of it. Counting confirms this. States of odd fermion number built on $|0\rangle$ at weight $\frac{1}{2}+N$ are

N	0	1	2	3	4
states	$\psi_{-1/2}$	$\psi_{-3/2}$	$\psi_{-5/2}$	$\psi_{-7/2}$	$\psi_{-9/2}$, $\psi_{-5/2}\psi_{-3/2}\psi_{-1/2}$
number	1	1	1	1	2
$p(N)$	1	1	2	3	5

(all acting on $|0\rangle$). The level-2 deficit of one state is $|\chi_\varepsilon\rangle$. At level 3 the deficit is two: $L_{-1}|\chi_\varepsilon\rangle$ and a second singular vector at level 3, which exists because $\frac{1}{2}$ is also a root of the level-3 factor of the Kac determinant (Section 11.4.10). This is the typical situation in a minimal model: the null submodule is generated by two singular vectors, and $L(c,h)$ is much smaller than $V(c,h)$. (As a check on β: $\|L_{-2}|0\rangle\|^2 = \frac{c}{2} = \frac{1}{4}$ from Example 11.39, and $\left\|\psi_{-3/2}\psi_{-1/2}|0\rangle\right\|^2 = 1$ with $\psi_r^\dagger = \psi_{-r}$, so $|\beta| = \frac{1}{2}$, consistent with $c=\frac{1}{2}$.)

Watch out: "irreducible" in Nash's sentence

Nash says the tower $L_{-1}^{\alpha_1}\cdots L_{-s}^{\alpha_s}|h\rangle$ comprises an irreducible representation. As a statement about *the space spanned by these vectors inside a Hilbert space* it is correct: in a unitary theory null vectors are automatically zero, so the span is $L(c,h)$, which is irreducible. As a statement about the Verma module, where the tower is taken as a formal basis, it is correct only off the Kac curves. The two readings differ exactly by the null submodule.

11.4.10 The Gram matrix and the Kac determinant

We now make Step 4 computable. The question is: for which (c,h) does the Shapovalov form have zero or negative directions? Since different levels are orthogonal, the form is block diagonal, one finite block per level. So the question reduces to linear algebra on finite matrices.

Working definition 11.19 (Gram matrix)

The **Gram matrix** at level N is the $p(N)\times p(N)$ matrix of inner products of the basis vectors of V_N,

$$M^{(N)}_{IJ}(c,h) = \big\langle L_{-I}h \,\big|\, L_{-J}h\big\rangle, \qquad L_{-I} = L_{-i_1}\cdots L_{-i_k}, \quad i_1\geq\cdots\geq i_k, \quad \textstyle\sum i = N.$$

Operational use: (i) $\det M^{(N)} = 0$ iff $V(c,h)$ has a null vector [a nonzero vector of V_N orthogonal to every vector of V_N] at level N; (ii) $M^{(N)}$ has a negative eigenvalue iff there is a negative-norm state [a vector v with $\langle v|v\rangle<0$] at level N; (iii) the representation can be unitary [every state has $\langle v|v\rangle\geq 0$, as a physical space of states requires] only if $M^{(N)}$ is positive semi-definite for every N.

The entries are polynomials in c and h, computed by commuting raising operators to the right. Level 1 is a single number:

$$M^{(1)} = \langle h|L_1L_{-1}|h\rangle \overset{(1)}{=} \langle h|[L_1,L_{-1}]|h\rangle \overset{(2)}{=} 2h. \tag{11.135}$$

(1) $L_1|h\rangle = 0$.
(2) $[L_1,L_{-1}] = 2L_0$ and $\langle h|h\rangle = 1$.

At level 2 the entries are read off from Example 11.42: each is $\langle h|$ applied to one of the computed vectors.

$$\langle h|L_2L_{-2}|h\rangle \overset{(1)}{=} 4h+\tfrac{c}{2}, \qquad \langle h|L_2L_{-1}^2|h\rangle \overset{(2)}{=} 6h,$$
$$\langle h|L_1^2L_{-2}|h\rangle \overset{(3)}{=} \langle h|L_1\,3L_{-1}|h\rangle = 6h, \qquad \langle h|L_1^2L_{-1}^2|h\rangle \overset{(4)}{=} (4h+2)\langle h|L_1L_{-1}|h\rangle = 4h(2h+1).$$

(1) $L_2L_{-2}|h\rangle = (4h+\frac{c}{2})|h\rangle$.
(2) $L_2L_{-1}^2|h\rangle = 6h|h\rangle$.
(3) $L_1L_{-2}|h\rangle = 3L_{-1}|h\rangle$, then (11.135).
(4) $L_1L_{-1}^2|h\rangle = (4h+2)L_{-1}|h\rangle$, then (11.135).

(The two off-diagonal entries agree, as they must for a Hermitian form with real c,h.) So, in the basis $(L_{-2}|h\rangle,\ L_{-1}^2|h\rangle)$,

$$M^{(2)} = \begin{pmatrix} 4h+\frac{c}{2} & 6h \\ 6h & 4h(2h+1)\end{pmatrix}, \qquad \det M^{(2)} = 2h\big(16h^2+2(c-5)h+c\big). \tag{11.136}$$

The determinant:

$$\det M^{(2)} \overset{(1)}{=} \Big(4h+\frac{c}{2}\Big)4h(2h+1)-36h^2 \overset{(2)}{=} 2h\Big[(8h+c)(2h+1)-18h\Big]$$
$$\overset{(3)}{=} 2h\big(16h^2+2ch+c-10h\big).$$

(1) $ad-bc$.
(2) Factor out $2h$: $(4h+\frac{c}{2})\cdot 4h(2h+1) = 2h(8h+c)(2h+1)$ and $36h^2 = 2h\cdot 18h$.
(3) Expand $(8h+c)(2h+1) = 16h^2+8h+2ch+c$ and subtract $18h$.

Look at the two factors. The factor $2h$ is $M^{(1)}$: it vanishes when $L_{-1}|h\rangle$ is null, and then its descendant $L_{-1}^2|h\rangle$ is null too. The second factor is *exactly* the level-2 curve (11.132) that we found by demanding $L_1|\chi\rangle = L_2|\chi\rangle = 0$. The determinant knows about singular vectors.

Level 3. The basis is $(L_{-3}|h\rangle,\ L_{-2}L_{-1}|h\rangle,\ L_{-1}^3|h\rangle)$. One entry as a sample; the others are computed the same way.

$$\langle h|L_1L_2\,L_{-2}L_{-1}|h\rangle \overset{(1)}{=} \langle h|L_1\Big([L_2,L_{-2}]L_{-1}+L_{-2}[L_2,L_{-1}]\Big)|h\rangle \overset{(2)}{=} \langle h|L_1\Big(\big(4L_0+\tfrac{c}{2}\big)L_{-1}+3L_{-2}L_1\Big)|h\rangle$$
$$\overset{(3)}{=} \Big(4(h+1)+\frac{c}{2}\Big)\langle h|L_1L_{-1}|h\rangle \overset{(4)}{=} h\,(8h+8+c).$$

(1) The adjoint of $L_{-2}L_{-1}$ is L_1L_2; commute L_2 to the right ($L_2|h\rangle = 0$).
(2) $[L_2,L_{-2}] = 4L_0+\frac{c}{2}$ and $[L_2,L_{-1}] = 3L_1$.
(3) $L_0L_{-1}|h\rangle = (h+1)L_{-1}|h\rangle$ and $L_1|h\rangle = 0$.
(4) (11.135): $\big(4h+4+\frac{c}{2}\big)2h$.

The full matrix and its determinant are

$$M^{(3)} = \begin{pmatrix} 6h+2c & 10h & 24h \\ 10h & h(8h+8+c) & 12h(3h+1) \\ 24h & 12h(3h+1) & 24h(h+1)(2h+1) \end{pmatrix}, \tag{11.137}$$

$$\det M^{(3)} = 48\,h^2\,\big(16h^2+2(c-5)h+c\big)\big(3h^2+(c-7)h+c+2\big). \tag{11.138}$$

(We checked these entries, and the determinants at levels 2, 3 and 4, by implementing the commutation rules on a computer; doing one or two entries by hand, as above, is the best way to trust them.) Read the factors:

- h^2: at $h=0$ the level-1 null vector $L_{-1}|h\rangle$ has $p(3-1)=2$ descendants at level 3 ($L_{-2}L_{-1}|h\rangle$ and $L_{-1}^3|h\rangle$), so the factor h appears twice.
- $16h^2+2(c-5)h+c$: the level-2 singular vector has $p(1)=1$ descendant at level 3, $L_{-1}|\chi\rangle$, so this factor appears once.
- $3h^2+(c-7)h+c+2$: new. It vanishes where a singular vector first appears at level 3.

At $c=\frac{1}{2}$ the new factor is $3h^2-\frac{13}{2}h+\frac{5}{2} = 3(h-\frac{1}{2})(h-\frac{5}{3})$. Its root $h=\frac{1}{2}$ is the second singular vector of the Ising energy field that we found by counting fermion states in Example 11.45.

The pattern: the Kac determinant. The structure we just saw continues at every level. A singular vector first appearing at level K contributes a factor that shows up at every level $N\geq K$, raised to the power equal to the number of its descendants at level N, which is $p(N-K)$. Kac found that the new singular vectors appear at levels $K=rs$, labelled by pairs of positive integers (r,s).

Key idea: the Kac determinant formula

$$\det M^{(N)}(c,h) = \alpha_N \prod_{\substack{r,s\geq 1\\ rs\leq N}} \big(h - h_{r,s}(c)\big)^{p(N-rs)}, \tag{11.139}$$

where $\alpha_N > 0$ is a constant independent of c and h, and the zeros are

$$h_{r,s}(c) = \frac{(r\,t-s)^2-(t-1)^2}{4t}, \qquad c = 13 - 6\Big(t+\frac{1}{t}\Big). \tag{11.140}$$

For each c the parameter t solves a quadratic; the two roots t and $1/t$ exchange $h_{r,s} \leftrightarrow h_{s,r}$, so the *set* of zeros depends only on c. For $c < 1$ it is convenient to write $t = \frac{m+1}{m}$ with $m > 0$ real; then

$$c = 1 - \frac{6}{m(m+1)}, \qquad h_{r,s} = \frac{\big((m+1)r - ms\big)^2 - 1}{4m(m+1)}. \tag{11.141}$$

What it means: $V(c,h)$ has a singular vector at level rs if and only if $h = h_{r,s}(c)$ for some $r, s \geq 1$.

Let us check that (11.141) follows from (11.140). With $t = \frac{m+1}{m}$,

$$c \overset{(1)}{=} 13 - 6\Big(\frac{m+1}{m} + \frac{m}{m+1}\Big) \overset{(2)}{=} 13 - 6\,\frac{2m^2+2m+1}{m(m+1)} \overset{(3)}{=} 13 - 12 - \frac{6}{m(m+1)},$$

$$h_{r,s} \overset{(4)}{=} \frac{\big(\frac{(m+1)r}{m} - s\big)^2 - \frac{1}{m^2}}{4\frac{m+1}{m}} \overset{(5)}{=} \frac{\big((m+1)r-ms\big)^2-1}{m^2}\cdot\frac{m}{4(m+1)} = \frac{\big((m+1)r-ms\big)^2-1}{4m(m+1)}.$$

(1) Insert t.
(2) Common denominator: $(m+1)^2 + m^2 = 2m^2 + 2m + 1$.
(3) $6(2m^2+2m)/(m(m+1)) = 12$.
(4) Insert t; $t - 1 = \frac{1}{m}$.
(5) Multiply numerator and denominator by m^2.

And a check against the level-2 factor: with (11.140), $h_{1,2} = \frac{3-2t}{4t}$ and $h_{2,1} = \frac{3t-2}{4}$, so

$$h_{1,2} + h_{2,1} \overset{(1)}{=} \frac{3t^2-4t+3}{4t} \overset{(2)}{=} \frac{5-c}{8}, \qquad h_{1,2}\,h_{2,1} \overset{(3)}{=} \frac{-6t^2+13t-6}{16t} \overset{(4)}{=} \frac{c}{16},$$

(1) Common denominator $4t$: $(3-2t) + t(3t-2)$.
(2) $5 - c = 5 - 13 + 6t + 6/t = (6t^2 - 8t + 6)/t$, and half of that over 4 is the left side.
(3) $(3-2t)(3t-2) = 9t - 6 - 6t^2 + 4t$.
(4) $c = (13t - 6t^2 - 6)/t$.

These are the sum and product of the roots of $16h^2 + 2(c-5)h + c$, so $16h^2 + 2(c-5)h + c = 16(h - h_{1,2})(h - h_{2,1})$, and (11.136) is (11.139) with $\alpha_2 = 32$ and $h_{1,1} = 0$. The level-4 determinant, computed the same way, contains the new factors $c + 8h - 1 = 8(h - h_{2,2})$ (indeed $h_{2,2} = \frac{1-c}{8}$) and $h - h_{1,4}$, $h - h_{4,1}$, together with $h^3(h-h_{1,2})^2(h-h_{2,1})^2(h-h_{1,3})(h-h_{3,1})$: the powers are $p(3) = 3$, $p(2) = 2$ and $p(1) = 1$, exactly as (11.139) predicts.

Why the formula has this shape (a sketch). Two facts make (11.139) plausible without the full proof. *Degree.* For large h, the diagonal entry for $L_{-I}|h\rangle$ with k factors grows like h^k and dominates its row, so $\det M^{(N)}$ is a polynomial in h of degree $\sum_{I\vdash N} k(I)$ (the total number of parts in all partitions of N). One checks that this equals $\sum_{rs\leq N} p(N-rs)$, the degree of the right-hand side. *Zeros.* Each zero $h_{r,s}$ comes with a singular vector at level rs, whose

descendants account for the multiplicity $p(N - rs)$. The location of the zeros is the hard part. Kac obtained it by exhibiting the singular vectors in free-field (Coulomb gas) realisations, which produce the square-completed form $(rt - s)^2$; Feigin and Fuchs proved the structure of the resulting submodules.

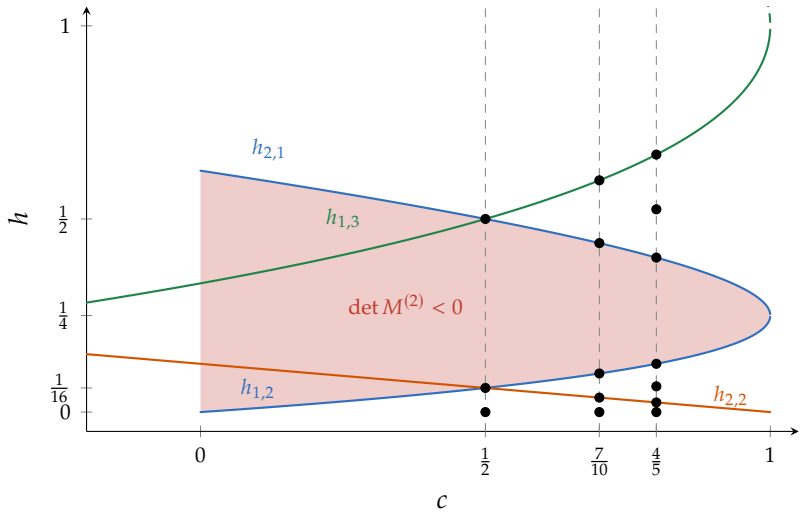


Figure 11.22: Zeros of the Kac determinant in the region $0 \le c < 1$: the curves $h_{1,2}$, $h_{2,1}$ (blue, level 2), $h_{1,3}$ (green) and $h_{3,1}$ (green dashed, level 3; it lies above the top of the plot except very near $c = 1$) and $h_{2,2}$ (orange, level 4). As $c \to 1$ every curve tends to $h_{r,s} \to (r-s)^2/4$: the blue curves meet at $h = \frac{1}{4}$, the green ones at $h = 1$, and $h_{2,2}$ (like every $h_{r,r}$) runs into $h = 0$. The shaded region between $h_{1,2}$ and $h_{2,1}$ is where $\det M^{(2)} < 0$, so there is a negative-norm state at level 2 and the representation cannot be unitary. The black dots on the dashed lines $c = \frac{1}{2}, \frac{7}{10}, \frac{4}{5}$ are the unitary minimal models of Section 11.4.12. Each one lies on two Kac curves at once, because $h_{r,s} = h_{m-r,\,m+1-s}$ (only a few of these curves are drawn), and none lies inside the shaded region.

Try it: the Kac zeros at $c = 1$ and the free boson

At $c = 1$, show that $t = 1$ and $h_{r,s} = \frac{(r-s)^2}{4}$. Compare with the weights $h = \alpha^2/4$ of the vertex operators $:e^{i\alpha\varphi}:$ of Example 11.23: which vertex operators sit at the degenerate values? *Answer:* $c = 13 - 6(t + 1/t) = 1$ gives $t + 1/t = 2$, so $t = 1$, and then $h_{r,s} = \frac{(r-s)^2-0}{4}$. The degenerate weights are $h = n^2/4$, $n \in \mathbb{Z}$, which are the vertex operators with $\alpha = \pm n$ (and $\partial\varphi$, with $h = 1$, is at $n = 2$). For these special charges the $c = 1$ Verma modules have singular vectors, and the boson Fock space realises the quotient.

11.4.11 Unitarity: which (c, h) can live in a Hilbert space?

The Kac determinant tells us where the Shapovalov form is *degenerate*. A physicist wants to know something slightly different: where is it *positive*? Nash puts it in one sentence (p. 310): "the need to construct the representation on a Hilbert space of positive norm plays an important part in finding those values of c and h which give rise to unitary irreducible representations of *Vir*." In Chapter 6 he quotes the answer, due to Friedan, Qiu and Shenker, as his (6.58). In this subsection we derive as much of that answer as can be done by hand. We also draw a picture of the rest, computed from the Gram matrices themselves.

Here is the question in operational form. A conformal family $[\phi]$ of a unitary CFT gives an irreducible module $L(c, h)$ whose inner product is positive definite. By Section 11.4.10, the

inner product is block diagonal by level. So the requirement is

$$\text{unitary} \iff M^{(N)}(c,h) \text{ is positive semi-definite for every level } N, \tag{11.142}$$

with the zero eigenvalues (the null states) then quotiented away, as in Step 4 of Section 11.4.9. A single negative eigenvalue at a single level is enough to rule out the pair (c,h). Such a state cannot be quotiented away, because it is not orthogonal to everything.

Physical picture: why negative norms are fatal

In quantum mechanics $\langle\psi|\psi\rangle$ is the total probability. If a state has $\langle\psi|\psi\rangle < 0$, then the Born rule assigns negative probabilities. Gauge theories do meet negative-norm states (the timelike photon in covariant gauge), but there a constraint such as the Gupta–Bleuler condition removes them. Here nothing removes them: the Virasoro generators *are* the symmetry, and the negative-norm state is produced from a physical state by applying the symmetry. A theory that contains $|h\rangle$ also contains $L_{-2}|h\rangle - L_{-1}^2|h\rangle$.

We now go through the (c,h) plane region by region. The result is summarised in Figure 11.23 and Figure 11.24.

The first two conditions: $h \geq 0$ and $c \geq 0$

The simplest descendants are the single-mode states $L_{-n}|h\rangle$, with $n \geq 1$. Their norms are

$$\begin{aligned} \|L_{-n}|h\rangle\|^2 &\overset{(1)}{=} \langle h|L_n L_{-n}|h\rangle \overset{(2)}{=} \langle h|\,[L_n, L_{-n}]\,|h\rangle \\ &\overset{(3)}{=} \langle h|\Big(2n\,L_0 + \frac{c}{12}\,n(n^2-1)\Big)|h\rangle \overset{(4)}{=} 2n\,h + \frac{c}{12}\,n(n^2-1). \end{aligned} \tag{11.143}$$

(1) The adjoint of L_{-n} is L_n, by (11.119).
(2) $L_n|h\rangle = 0$ for $n \geq 1$ (the highest-weight condition (11.121)), so $L_nL_{-n}|h\rangle = [L_n, L_{-n}]|h\rangle$.
(3) The Virasoro algebra (11.102) with $m = n$ and $n \to -n$: $(n-(-n))L_0 + \frac{c}{12}n(n^2-1)$.
(4) $L_0|h\rangle = h|h\rangle$ and $\langle h|h\rangle = 1$.

Two consequences follow at once.

$h \geq 0$. Put $n = 1$. The central term vanishes, and $\|L_{-1}|h\rangle\|^2 = 2h$. So a unitary highest-weight module has $h \geq 0$. Moreover, if $h = 0$ then $L_{-1}|h\rangle$ has zero norm. In a Hilbert space a zero-norm vector is zero, so $L_{-1}|h\rangle = 0$. In field language this says $\partial\phi(0)|0\rangle = 0$: the field does not depend on z. A holomorphic weight-zero primary in a unitary theory is therefore a constant, i.e. a multiple of the identity. In a unitary CFT the vacuum is the *only* state with $h = \bar{h} = 0$ (assuming a unique vacuum).

$c \geq 0$. Now let n grow at fixed h. The term $\frac{c}{12}n^3$ eventually dominates $2nh$. If $c < 0$ then $\|L_{-n}|h\rangle\|^2 < 0$ for all large n. So a unitary theory has $c \geq 0$. The cleanest case is the vacuum itself, $h = 0$, at $n = 2$:

$$\|L_{-2}|0\rangle\|^2 = \frac{c}{2}. \tag{11.144}$$

In field language $L_{-2}|0\rangle = T(0)|0\rangle$, and (11.144) is the statement that the two-point function of the stress tensor, $\langle T(z)T(0)\rangle = \frac{c/2}{z^4}$, has a positive coefficient. (We derive that two-point function in Section 11.4.13.) The central charge measures how strongly the stress tensor fluctuates, and positivity of that fluctuation is positivity of c.

Worked example 11.46: Negative central charge: the ghosts of string theory

The Faddeev–Popov ghosts of the bosonic string (Chapter 9) form a CFT with $c = -26$. Its vacuum module already contains a negative-norm state:

$$\|L_{-2}|0\rangle\|^2 \overset{(1)}{=} \frac{c}{2} = -13 < 0, \qquad \|L_{-3}|0\rangle\|^2 \overset{(2)}{=} \frac{c}{12}\cdot 3\cdot 8 = 2c = -52 < 0.$$

(1) (11.144) with $c = -26$.
(2) (11.143) with $h = 0$, $n = 3$: $n(n^2-1) = 24$.

This does not make string theory inconsistent. The ghosts are not physical particles. The matter CFT has $c = +26$, and the total central charge $26 - 26 = 0$ is what the BRST construction needs. The physical states are BRST cohomology classes, and on them the inner product is positive. The lesson is that negative central charges occur in useful non-unitary CFTs. They just never describe a space of physical states on their own.

$c = 0$ allows only the trivial theory. At $c = 0$ the single-mode norms $2nh$ are all positive for $h > 0$. We need a two-dimensional subspace to see the problem. Fix $N \geq 1$ and consider the two states $L_{-2N}|h\rangle$ and $L_{-N}^2|h\rangle$ at level $2N$. Write $c_N := \frac{c}{12}N(N^2-1)$, so that $[L_N, L_{-N}] = 2NL_0 + c_N$. The entries of their Gram matrix are

$$\begin{aligned}
\langle h|L_{2N}L_{-2N}|h\rangle &\overset{(1)}{=} 4Nh + c_{2N},\\
L_{2N}L_{-N}^2|h\rangle &\overset{(2)}{=} [L_{2N}, L_{-N}]\,L_{-N}|h\rangle \overset{(3)}{=} 3N\,L_NL_{-N}|h\rangle \overset{(4)}{=} 3N\,(2Nh + c_N)|h\rangle,\\
L_NL_{-N}^2|h\rangle &\overset{(5)}{=} \big([L_N, L_{-N}]L_{-N} + L_{-N}[L_N, L_{-N}]\big)|h\rangle \overset{(6)}{=} \big(2N(h+N) + c_N + 2Nh + c_N\big)L_{-N}|h\rangle,
\end{aligned}$$

(1) (11.143) with $n = 2N$.
(2) $L_{2N}|h\rangle = 0$, so only the commutator with the first L_{-N} survives, and then L_{2N} hits $L_{-N}|h\rangle$, whose level N is too low for L_{2N}: $L_{2N}L_{-N}|h\rangle = [L_{2N}, L_{-N}]|h\rangle = 3NL_N|h\rangle = 0$.
(3) $[L_{2N}, L_{-N}] = (2N+N)L_N$; no central term since $2N - N \neq 0$.
(4) $L_NL_{-N}|h\rangle = [L_N, L_{-N}]|h\rangle = (2Nh + c_N)|h\rangle$.
(5) Commute L_N through both factors; $L_N|h\rangle = 0$.
(6) L_0 acting on $L_{-N}|h\rangle$ gives $h + N$, and acting on $|h\rangle$ gives h.

So $\big\|L_{-N}^2|h\rangle\big\|^2 = (4Nh + 2N^2 + 2c_N)(2Nh + c_N)$. Set $c = 0$, so that all the c_N vanish. The Gram matrix in the basis $(L_{-2N}|h\rangle,\ L_{-N}^2|h\rangle)$ and its determinant are

$$\begin{aligned}
\det\begin{pmatrix} 4Nh & 6N^2h \\ 6N^2h & 2Nh(4Nh + 2N^2)\end{pmatrix} &\overset{(1)}{=} 8N^2h^2(4Nh + 2N^2) - 36N^4h^2\\
&\overset{(2)}{=} 32N^3h^3 - 20N^4h^2 \overset{(3)}{=} 4N^3h^2\,(8h - 5N).
\end{aligned}$$

(1) $ad - b^2$.
(2) Expand: $8N^2h^2 \cdot 4Nh = 32N^3h^3$ and $8N^2h^2 \cdot 2N^2 - 36N^4h^2 = -20N^4h^2$.
(3) Factor out $4N^3h^2$.

For any $h > 0$, choose an integer $N > \frac{8h}{5}$. The determinant is then negative, so this 2×2 block has one negative eigenvalue. *At $c = 0$ every $h > 0$ is excluded.* What remains is $h = 0$. But at $c = 0$, $\|L_{-2}|0\rangle\|^2 = 0$ as well, so $T(0)|0\rangle = 0$: the stress tensor vanishes. The only unitary $c = 0$ theory is the trivial one, with one state. (For $N = 1$ the determinant $4h^2(8h-5)$ is the level-2 Kac determinant at $c = 0$, $2h\big(16h^2 - 10h\big)$, as it should be.)

Watch out: non-unitary is not the same as unphysical

The $c = 0$ result does not say that $c = 0$ CFTs are useless. Two-dimensional percolation and self-avoiding polymers are critical, conformally invariant and have $c = 0$. Their CFTs are non-unitary: the transfer matrix of these lattice models is not a positive operator. Likewise the Yang–Lee edge singularity is described by a CFT with $c = -\frac{22}{5}$ and a primary of weight $h = -\frac{1}{5}$. Unitarity is the right requirement for a quantum field theory whose states are physical particles. It is a genuine extra assumption for statistical models.

$c \geq 1$: no further obstruction

For $c \geq 1$ we show that $h \geq 0$ is the only condition. The argument uses the Kac formula (11.140) and a continuity principle.

Key idea: continuity principle

The entries of $M^{(N)}(c, h)$ are polynomials in (c, h), so its eigenvalues depend continuously on (c, h). An eigenvalue can change sign only by passing through zero, and there $\det M^{(N)} = 0$. Hence *on a connected region of the (c, h) plane that contains no zero of the Kac determinant, the number of negative eigenvalues of $M^{(N)}$ is constant.*

So we need two facts: that there are no zeros in the region $c > 1$, $h > 0$, and that somewhere in that region there is no negative eigenvalue.

Where the zeros are. First rewrite $h_{r,s}$ from (11.140) so that its dependence on t is visible:

$$h_{r,s} \overset{(1)}{=} \frac{r^2t^2 - 2rs\,t + s^2 - t^2 + 2t - 1}{4t} \overset{(2)}{=} \frac{1}{4}\Big[(r^2-1)\,t + \frac{s^2-1}{t} - 2(rs-1)\Big]. \tag{11.145}$$

(1) Expand $(rt - s)^2 - (t - 1)^2$.
(2) Group by powers of t: $(r^2 - 1)t^2 - 2(rs - 1)t + (s^2 - 1)$, then divide by $4t$.

Next, which t belongs to which c? From $c = 13 - 6(t + 1/t)$ we get $t + \frac{1}{t} = \frac{13-c}{6}$. Three cases:

- $c < 1$: $t + \frac{1}{t} > 2$, so t is real and positive (and $t \neq 1$).
- $1 < c < 25$: $-2 < t + \frac{1}{t} < 2$. A real t has $|t + 1/t| \geq 2$, so t is complex. Since t and $1/t$ are the two roots of $t^2 - \frac{13-c}{6}t + 1 = 0$ with real coefficients, they are complex conjugates, and $t\bar{t} = t \cdot \frac{1}{t} = 1$. So $t = e^{i\theta}$ with $0 < \theta < \pi$.
- $c \geq 25$: $t + \frac{1}{t} \leq -2$, so t is real and negative.

Now look at $h_{r,s}$ in each case with $c > 1$.

Case $1 < c < 25$, $t = e^{i\theta}$. By (11.145),

$$\operatorname{Im} h_{r,s} \overset{(1)}{=} \frac{1}{4}\big[(r^2-1)\sin\theta - (s^2-1)\sin\theta\big] = \frac{(r^2-s^2)\sin\theta}{4},$$
$$h_{r,r} \overset{(2)}{=} \frac{1}{4}\big[(r^2-1)(e^{i\theta}+e^{-i\theta}) - 2(r^2-1)\big] \overset{(3)}{=} -\frac{(r^2-1)(1-\cos\theta)}{2} \leq 0.$$

(1) $\operatorname{Im} e^{\pm i\theta} = \pm\sin\theta$; the constant term is real.
(2) Put $s = r$ in (11.145).
(3) $e^{i\theta} + e^{-i\theta} = 2\cos\theta$, so the bracket is $2(r^2 - 1)(\cos\theta - 1)$.

For $r \neq s$ the imaginary part is nonzero ($\sin\theta > 0$), so $h_{r,s}$ is not real. For $r = s$ it is real but ≤ 0. Either way, no zero lies at real $h > 0$.

Case $c \geq 25$, $t < 0$. All three terms in (11.145) are ≤ 0: $(r^2-1)t \leq 0$, $(s^2-1)/t \leq 0$ and $-2(rs-1) \leq 0$. So $h_{r,s} \leq 0$.

Hence for $c > 1$ the Kac determinant has no zero anywhere in $h > 0$. The region $\{c > 1,\ h > 0\}$ is connected.

A point with no negative eigenvalue. Take h very large at fixed c. As we noted after (11.139), the diagonal entry of $M^{(N)}$ belonging to $L_{-i_1}\cdots L_{-i_k}|h\rangle$ grows like h^k with a positive coefficient, while the off-diagonal entries in its row grow more slowly. A symmetric matrix dominated by a positive diagonal is positive definite. So for $h \gg 1$ there is no negative eigenvalue. The continuity principle then gives:

Key idea: the region $c \geq 1$

For $c > 1$ and $h > 0$, every Gram matrix $M^{(N)}(c,h)$ is positive definite. The Verma module $V(c,h)$ is irreducible and its completion is a Hilbert space: every such (c,h) gives a unitary representation of *Vir*. At $c = 1$ the zeros $h_{r,s} = \frac{(r-s)^2}{4}$ are real and ≥ 0. By continuity from $c > 1$, the Gram matrices at $c = 1$ are positive *semi*-definite, and after quotienting the null states the representation is unitary for all $h \geq 0$. This is the first line of Nash's (6.58): "$c \geq 1$, $h \geq 0$".

The $c = 1$ free boson of Section 11.4.5 realises this explicitly: its Fock space is manifestly positive (the a_{-n} are ordinary creation operators) and it contains primaries of every weight $h = \frac{\alpha^2}{4} \geq 0$, the vertex operators of Example 11.23.

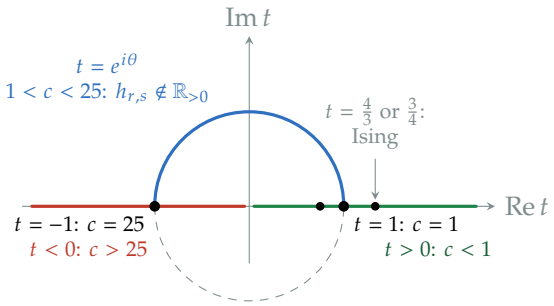


Figure 11.23: The parameter t of the Kac formula, $c = 13 - 6(t + 1/t)$. Values of c below 1 put t on the positive real axis (green), and there the zeros $h_{r,s}$ are real and many of them positive. For $1 < c < 25$, t lies on the unit circle (blue), and the zeros are either non-real or ≤ 0. For $c > 25$, t is negative (red) and all zeros are ≤ 0. The replacement $t \to 1/t$ (reflection in the unit circle) gives the same c and swaps $h_{r,s} \leftrightarrow h_{s,r}$. Only on the green half-line can the Kac determinant vanish at positive h, so only there can unitarity fail beyond $h \geq 0$.

$0 < c < 1$: the sieve

Below $c = 1$ everything changes. Now $t > 0$ is real and the curves $h_{r,s}(c)$ of Figure 11.22 run through the physical region $h > 0$. Crossing a curve can flip the sign of an eigenvalue. We look first at level 2, where everything can be done by hand.

Level 2. By (11.136) and (11.139), $\det M^{(2)} = 32\,h\,(h - h_{1,2})(h - h_{2,1})$. For $0 < c < 1$ both roots are real and positive, with $h_{1,2} < h_{2,1}$. If $h_{1,2} < h < h_{2,1}$, one factor is negative and the

others positive, so $\det M^{(2)} < 0$. A real symmetric 2×2 matrix with negative determinant has eigenvalues $\lambda_1\lambda_2 < 0$, i.e. one of each sign. So the entire band between the two level-2 curves is excluded. (It is the shaded region of Figure 11.22.)

Worked example 11.47: An explicit negative-norm state at $c = \frac{1}{2}$, $h = \frac{1}{4}$

At $c = \frac{1}{2}$ we have $h_{1,2} = \frac{1}{16}$ and $h_{2,1} = \frac{1}{2}$ (Example 11.43). The value $h = \frac{1}{4}$ lies between them. Take the simplest combination $|\psi\rangle = (L_{-2} - L_{-1}^2)|h\rangle$. With the level-2 entries $M_{11} = 4h + \frac{c}{2}$, $M_{12} = 6h$, $M_{22} = 4h(2h+1)$ of (11.136),

$$\langle\psi|\psi\rangle \overset{(1)}{=} M_{11} - 2M_{12} + M_{22} \overset{(2)}{=} 4h + \frac{c}{2} - 12h + 8h^2 + 4h = 8h^2 - 4h + \frac{c}{2}$$
$$\overset{(3)}{=} 8\cdot\frac{1}{16} - 4\cdot\frac{1}{4} + \frac{1}{4} = \frac{1}{2} - 1 + \frac{1}{4} = -\frac{1}{4}.$$

(1) Bilinearity: $\langle\psi|\psi\rangle = \langle L_{-2}h|L_{-2}h\rangle - 2\langle L_{-2}h|L_{-1}^2h\rangle + \langle L_{-1}^2h|L_{-1}^2h\rangle$ (the entries are real).
(2) Insert the entries.
(3) $h = \frac{1}{4}$, $c = \frac{1}{2}$.

So a CFT with $c = \frac{1}{2}$ cannot contain a primary of weight $\frac{1}{4}$. The same formula shows how narrow the window is: $8h^2 - 4h + \frac{1}{4} < 0$ exactly for $\frac{2-\sqrt{2}}{8} < h < \frac{2+\sqrt{2}}{8}$, i.e. $0.073 < h < 0.427$. This vector alone does not see the whole band $(\frac{1}{16}, \frac{1}{2})$. The eigenvector of the negative eigenvalue does. At $h = \frac{1}{4}$ the matrix is $\left(\begin{smallmatrix} 5/4 & 3/2 \\ 3/2 & 3/2 \end{smallmatrix}\right)$, with $\det = \frac{15}{8} - \frac{9}{4} = -\frac{3}{8}$ and eigenvalues $\frac{11\pm\sqrt{145}}{8} \approx 2.88,\ -0.13$.

Level 3 and beyond. At level 3 the new factor $3h^2 + (c-7)h + c + 2 = 3(h - h_{1,3})(h - h_{3,1})$ of (11.138) is negative between $h_{1,3}$ and $h_{3,1}$. Where the level-2 factor is positive, $\det M^{(3)} < 0$ there. The determinant of a 3×3 matrix is the product of three eigenvalues, so an odd number of them are negative, and at least one is. A second band is excluded. At level 4 the factors $h - h_{2,2}$, $h - h_{1,4}$ and $h - h_{4,1}$ bring more bands, and so on. Each level carves the (c, h) region further, like a sieve.

To see what survives, we computed the Gram matrices exactly up to level 12 ($p(12) = 77$ states), using the same commutation rules we used by hand for levels 2 and 3. For each point of a fine grid we recorded the first level with a negative eigenvalue. The result is Figure 11.24.

c	h	first negative level	comment
0.5	0.25	2	inside the level-2 band (Example 11.47)
0.7	0.2	2	inside the level-2 band
0.9	0.3	2	the band is narrow near $c = 1$, but 0.3 is in it
0.6	0.02	4	below $h_{1,2}$; caught by the $h_{2,2}$ factor
0.1	0.01	4	likewise
0.8	0.1	6	between the $c = \frac{4}{5}$ Kac values $\frac{1}{15}$ and $\frac{1}{8}$
0.6	0.5	6	just above $h_{2,1}(0.6) \approx 0.43$
0.95	0.05	none up to 9	not a minimal model: excluded at a higher level
0.7	0.1	none	$= (c_4, h_{1,2})$, a minimal model (tricritical Ising)
0.5	0.0625	none	$= (c_3, h_{1,2})$, the Ising spin field

Table 11.3: Sample points from the computation behind Figure 11.24. The first two columns are the point, the third is the lowest level with a negative eigenvalue of $M^{(N)}$. Points near $c = 1$ survive to high level: the sieve's holes get finer as $c \to 1$.

The picture suggests, and Friedan, Qiu and Shenker proved, the following.

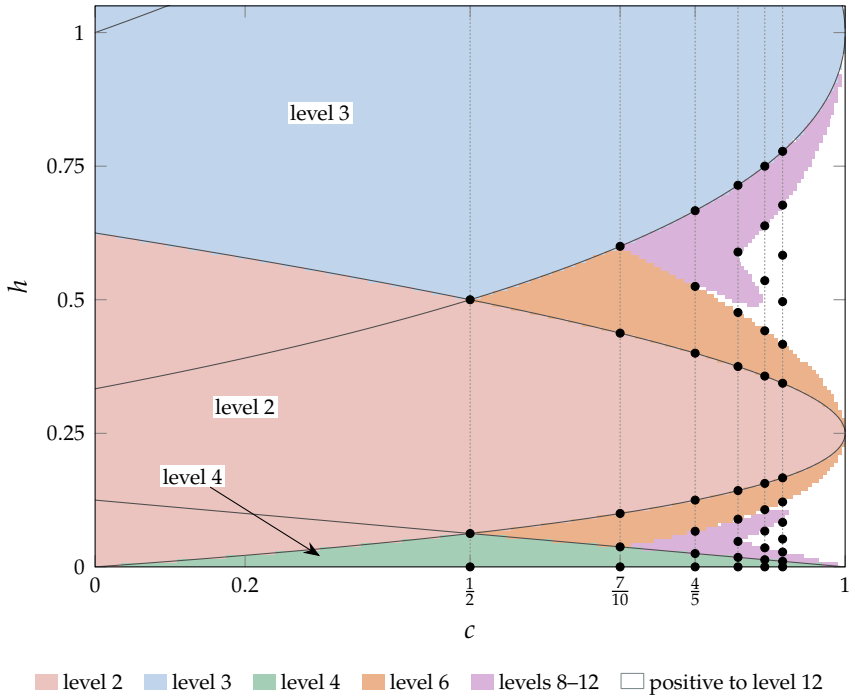


Figure 11.24: The unitarity sieve for $0 < c < 1$, computed from the exact Gram matrices $M^{(N)}(c, h)$ for $N \leq 12$. The colour of a point is the first level at which a negative-norm state appears. The thin black curves are the Kac curves $h_{r,s}(c)$ with $rs \leq 4$. The boundaries of the level-2, 3 and 4 regions run exactly along them, as the continuity principle demands; the higher-level regions are bounded in the same way by curves with $rs = 6, 8, 9, 12$, not drawn. Level 2 removes the band between $h_{1,2}$ and $h_{2,1}$, level 3 everything above $h_{2,1}$ in view, level 4 a thin wedge at small h, and higher levels bite ever closer to $c = 1$. The white region shrinks as the level grows. The black dots are the unitary minimal models $c = \frac{1}{2}, \frac{7}{10}, \frac{4}{5}, \frac{6}{7}, \frac{25}{28}, \frac{11}{12}$ (dotted lines) with $h = h_{r,s}$: they sit exactly on crossings of the curves, and they survive every level. The white area near $c = 1$ that is still uncoloured at level 12 is removed at higher levels, except for these dots.

Key idea: the Friedan–Qiu–Shenker theorem; Nash's (6.58)

For $0 \le c < 1$, the Virasoro module $L(c,h)$ is unitary *only* if

$$c = c_m := 1 - \frac{6}{m(m+1)}, \qquad h = h_{r,s} = \frac{\big((m+1)r - ms\big)^2 - 1}{4m(m+1)}, \tag{11.146}$$
$$m = 2,3,4,\dots, \qquad 1 \le r \le m-1, \quad 1 \le s \le m.$$

Together with the region $c \ge 1$, $h \ge 0$ this is the complete list of unitary highest-weight representations of *Vir*.

Some remarks on why this is plausible.

- *Why the special lines $c = c_m$.* The values c_m are exactly those for which $t = \frac{m+1}{m}$ is rational with numerator and denominator differing by 1. Then $h_{r,s} = h_{r+m,s+m+1}$, and the curves $h_{r,s}$ cross each other at the points (11.146). At a crossing the determinant vanishes to higher order, and sign changes coming from different curves can cancel. At a generic point between curves, FQS show that at some level an odd number of factors in (11.139) is negative.
- *Why they accumulate at $c = 1$.* As $m \to \infty$, $c_m \to 1$ and the table (11.146) fills in more and more of the half-line $h \ge 0$, in the limit giving the $c = 1$ values $h = \frac{(r-s)^2}{4}$ and everything in between.
- *Why $m = 2$ is trivial.* $c_2 = 0$ and the only entry is $h_{1,1} = h_{1,2} = 0$: the trivial theory, as we showed by hand above.

Existence: the coset construction. The FQS theorem is a list of *necessary* conditions. That each point on the list really carries a unitary representation was shown by Goddard, Kent and Olive (Nash's Chapter 6 cites them). They build *Vir* with $c = c_m$ inside a manifestly unitary theory, namely a product of SU(2) Kac–Moody representations (these appear in Section 11.4.14). By the Sugawara construction, the level-k SU(2) current algebra has a stress tensor with $c(k) = \frac{3k}{k+2}$. The difference of the stress tensors of $\mathrm{SU}(2)_k \times \mathrm{SU}(2)_1$ and of the diagonal $\mathrm{SU}(2)_{k+1}$ is again a Virasoro generator, commuting with the diagonal currents, with central charge

$$c \overset{(1)}{=} \frac{3k}{k+2} + \frac{3\cdot 1}{1+2} - \frac{3(k+1)}{k+3} \overset{(2)}{=} 1 + \frac{3\big[k(k+3) - (k+1)(k+2)\big]}{(k+2)(k+3)} \overset{(3)}{=} 1 - \frac{6}{(k+2)(k+3)}.$$

(1) The coset (difference) of Sugawara central charges, with $c(1) = 1$.
(2) Common denominator $(k+2)(k+3)$ for the first and third terms.
(3) $k^2 + 3k - (k^2 + 3k + 2) = -2$.

With $m = k+2$ this is c_m. The space of states of the coset theory is a subspace of a positive Hilbert space, so the representations are unitary, and one checks that all the weights $h_{r,s}$ of (11.146) occur. For $k = 1$ ($m = 3$) we get $c = \frac{1}{2}$, the Ising model.

Reading Nash (p. 310 and Chapter 6, pp. 186–187, (6.57)–(6.58))

In Chapter 6 Nash defines a highest-weight representation by (6.57), $L_0|h\rangle = h|h\rangle$, $L_n|h\rangle = 0$ for $n \ge 1$ (the same as his (11.37)), and quotes the unitary values (6.58):

$$c \ge 1,\ h \ge 0, \qquad \text{or} \qquad c = 1 - \frac{6}{(m+1)(m+2)}, \quad h = \frac{\big((m+2)p - (m+1)q\big)^2 - 1}{4(m+1)(m+2)},$$

with $m = 1, 2, \ldots$, $p = 1, \ldots, m$, $q = 1, \ldots, p$. His m is our $m - 1$: put $m_{\text{Nash}} = m - 1$, $p = r$, $q = s$ and one gets (11.146). His ranges $1 \le q \le p \le m_{\text{Nash}}$ pick one representative from each pair identified by the symmetry (11.147) below, so they list each distinct weight once. His $m = 1$ is our $m = 2$: $c = 0$, $h = 0$, the trivial theory. On p. 310 he adds that negative-norm states "are disallowed" and that positivity selects the allowed (c, h); in this subsection we derived $h \ge 0$, $c \ge 0$, the triviality of $c = 0$, the absence of obstructions for $c \ge 1$, and the level-2 and level-3 bands, and we displayed the rest of the sieve by computation.

11.4.12 The unitary minimal models

The discrete series has a striking property. At $c = c_m$ only finitely many values of h are allowed. So a unitary CFT with $c < 1$ can contain only finitely many conformal families. Such a theory is called a **minimal model**, and it is labelled by m. These are the simplest interacting quantum field theories that exist, and they are solvable: every correlation function can be computed exactly. Nash's §4 (the next part of these notes) takes the $m = 3$ model as its central example.

Working definition 11.20 (Kac table and minimal model $\mathcal{M}_m$)

Fix an integer $m \ge 3$. The **Kac table** is the rectangle of weights $h_{r,s}$ in (11.146) [$h_{r,s} = \frac{((m+1)r - ms)^2 - 1}{4m(m+1)}$], with $1 \le r \le m-1$ and $1 \le s \le m$. The **unitary minimal model** $\mathcal{M}_m$ is the CFT with $c = c_m$ [$c_m = 1 - \frac{6}{m(m+1)}$, so $c_3 = \frac{1}{2}$] whose primary fields are $\phi_{r,s}$ of weights $(h_{r,s}, h_{r,s})$, one for each distinct entry of the table. (We describe the "diagonal" model, with $\bar{h} = h$; other combinations are possible, see the pitfall below.) *How to use it:* to find the spectrum, evaluate $h_{r,s}$ on the rectangle and identify equal entries using the symmetry (11.147) [$h_{m-r,\,m+1-s} = h_{r,s}$, so each weight appears exactly twice].

The symmetry of the table. The table is symmetric under a half-turn:

$$h_{m-r,\,m+1-s} \overset{(1)}{=} \frac{\big((m+1)(m-r) - m(m+1-s)\big)^2 - 1}{4m(m+1)} \overset{(2)}{=} \frac{\big(-(m+1)r + ms\big)^2 - 1}{4m(m+1)} \overset{(3)}{=} h_{r,s}. \tag{11.147}$$

(1) (11.146) with $(r, s) \to (m - r, m + 1 - s)$.
(2) $(m+1)m - (m+1)r - m(m+1) + ms = -(m+1)r + ms$.
(3) The square does not see the overall sign.

No entry is fixed by the half-turn: that would need $r = \frac{m}{2}$ and $s = \frac{m+1}{2}$, but m and $m+1$ are not both even. So the $(m-1)m$ entries pair up, and $\mathcal{M}_m$ has

$$\#\{\text{primaries of } \mathcal{M}_m\} = \frac{m(m-1)}{2}. \tag{11.148}$$

The identity is always $\phi_{1,1} = \phi_{m-1,m}$, with $h_{1,1} = \frac{(m+1-m)^2 - 1}{4m(m+1)} = 0$.

Worked example 11.48: The Ising model, $m = 3$

For $m = 3$, $c_3 = 1 - \frac{6}{12} = \frac{1}{2}$ and $h_{r,s} = \frac{(4r-3s)^2 - 1}{48}$. The first row of the table is

$$h_{1,1} \overset{(1)}{=} \frac{(4-3)^2 - 1}{48} = 0, \quad h_{1,2} \overset{(1)}{=} \frac{(4-6)^2 - 1}{48} = \frac{3}{48} = \frac{1}{16}, \quad h_{1,3} \overset{(1)}{=} \frac{(4-9)^2 - 1}{48} = \frac{24}{48} = \frac{1}{2},$$

(1) (11.146) with $m = 3$, $r = 1$.

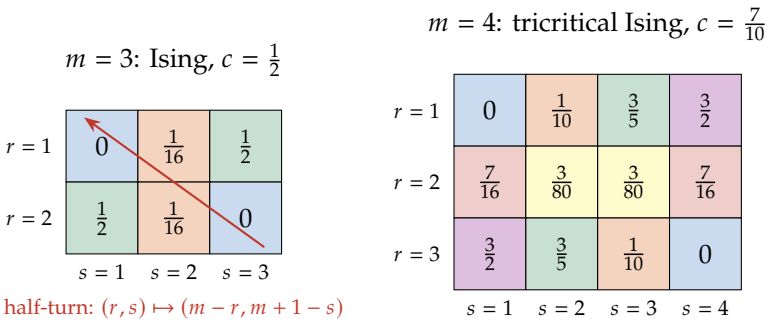


Figure 11.25: Kac tables of the first two unitary minimal models. Rows are r, columns are s. Equal colours mark entries identified by the half-turn symmetry (11.147), which rotates the rectangle about its centre (red arrow). The Ising table has $3 = \frac{3\cdot 2}{2}$ distinct weights $0, \frac{1}{16}, \frac{1}{2}$; the tricritical Ising table has $6 = \frac{4\cdot 3}{2}$ weights $0, \frac{3}{80}, \frac{1}{10}, \frac{7}{16}, \frac{3}{5}, \frac{3}{2}$.

and the second row is the first reversed, by (11.147). So the Ising CFT has three primaries:

$$\mathbb{1}\ (h=0), \qquad \sigma\ (h=\tfrac{1}{16}), \qquad \varepsilon\ (h=\tfrac{1}{2}).$$

The field σ is the continuum limit of the lattice spin, and ε is the energy density $\sigma_i\sigma_j$ on a bond. We already know one realisation: the free Majorana fermion of Section 11.3.7 has $c=\frac{1}{2}$, and $\varepsilon = i\psi\bar\psi$ has weights $(\frac{1}{2},\frac{1}{2})$. The spin field σ is not a polynomial in ψ: it is a twist field around which ψ changes sign.

Counting the states. Computing the rank of $M^{(N)}$ at the three points $(\frac{1}{2},0)$, $(\frac{1}{2},\frac{1}{16})$, $(\frac{1}{2},\frac{1}{2})$ gives the number of states of the irreducible module at each level:

level N	0	1	2	3	4	5	6	7	8	9
Verma $p(N)$	1	1	2	3	5	7	11	15	22	30
$L(\frac{1}{2},0)$	1	0	1	1	2	2	3	3	5	5
$L(\frac{1}{2},\frac{1}{16})$	1	1	1	2	2	3	4	5	6	8
$L(\frac{1}{2},\frac{1}{2})$	1	1	1	1	2	2	3	4	5	6

The irreducible modules are far smaller than the Verma module: at level 9 the vacuum module keeps 5 of 30 states. The free fermion confirms the first and third rows. Its Fock space is built by ψ_{-k}, $k\in\{\frac{1}{2},\frac{3}{2},\frac{5}{2},\dots\}$, each used at most once (Pauli). An even number of fermions gives integer level: at level 4 there are $\psi_{-7/2}\psi_{-1/2}|0\rangle$ and $\psi_{-5/2}\psi_{-3/2}|0\rangle$, i.e. 2 states, as in the row $L(\frac{1}{2},0)$. An odd number of fermions gives the module of ε, starting with $\psi_{-1/2}|0\rangle$ at weight $\frac{1}{2}$. (The chiral half of $\varepsilon = i\psi\bar\psi$ is ψ, of weight $\frac{1}{2}$.)

Worked example 11.49: Critical exponents of the Ising model from the Kac table

In Example 11.3 we went from Onsager's exponents to scaling dimensions. Now we run the argument backwards and *predict* the exponents from the Kac table. Recall from Section 11.1.2 that η is defined by the power law of the spin-spin correlation at T_c, $\langle\sigma(x)\sigma(0)\rangle \sim |x|^{-(d-2+\eta)}$, and ν by $\xi\sim|T-T_c|^{-\nu}$. A spinless primary of weights (h,h) has scale dimension $\Delta=2h$ and two-point function $|x|^{-2\Delta}$ (by the two-point function of Section 11.3.6). In $d=2$:

$$\eta \overset{(1)}{=} 2\Delta_\sigma \overset{(2)}{=} 2\cdot 2\cdot\frac{1}{16} = \frac{1}{4},$$
$$\nu \overset{(3)}{=} \frac{1}{2-\Delta_\varepsilon} \overset{(4)}{=} \frac{1}{2-1} = 1,$$

$$\beta \overset{(5)}{=} \frac{\nu\,(d-2+\eta)}{2} = \frac{1\cdot\frac{1}{4}}{2} = \frac{1}{8}.$$

(1) Compare $|x|^{-2\Delta_\sigma}$ with $|x|^{-(d-2+\eta)}$ at $d=2$.
(2) $\Delta_\sigma = h + \bar{h} = 2h_{1,2}$.
(3) Moving off T_c adds $\delta S = g\int \mathrm{d}^2x\,\varepsilon(x)$ with $g \propto T - T_c$. Since d^2x has dimension -2 and ε has Δ_ε, the coupling g has mass dimension $2-\Delta_\varepsilon$. The only length is then $\xi \sim g^{-1/(2-\Delta_\varepsilon)}$.
(4) $\Delta_\varepsilon = 2h_{1,3} = 1$.
(5) The scaling relation between the magnetisation exponent β and ν, η (from $\langle\sigma\rangle \sim \xi^{-\Delta_\sigma}$, so $\beta = \nu\Delta_\sigma$).

These are exactly Onsager's exact values for the square-lattice Ising model: $\eta = \frac{1}{4}$, $\nu = 1$, $\beta = \frac{1}{8}$. A lattice computation that took Onsager a tour de force comes out of (11.146) in three lines. The same recipe for the tricritical Ising model ($m = 4$) uses the leading energy field $\varepsilon = \phi_{1,2}$, $h = \frac{1}{10}$, so $\Delta_\varepsilon = \frac{1}{5}$ and $\nu = \frac{1}{2-1/5} = \frac{5}{9}$, the known tricritical value.

m	c_m	primaries	statistical model at criticality and weights h
2	0	1	trivial theory
3	$\frac{1}{2}$	3	Ising: $0,\ \frac{1}{16},\ \frac{1}{2}$
4	$\frac{7}{10}$	6	tricritical Ising: $0,\ \frac{3}{80},\ \frac{1}{10},\ \frac{7}{16},\ \frac{3}{5},\ \frac{3}{2}$
5	$\frac{4}{5}$	10	tetracritical Ising: $0, \frac{1}{40}, \frac{1}{15}, \frac{1}{8}, \frac{2}{5}, \frac{21}{40}, \frac{2}{3}, \frac{7}{5}, \frac{13}{8}, 3$; the 3-state Potts model uses the subset $0, \frac{1}{15}, \frac{2}{5}, \frac{2}{3}, \frac{7}{5}, 3$
6	$\frac{6}{7}$	15	tricritical 3-state Potts model (a subset, as for $m=5$)
∞	1	∞	the limit: $c=1$ theories such as the free boson

Table 11.4: The first unitary minimal models. The number of primaries is $\frac{m(m-1)}{2}$ from (11.148).

Watch out: diagonal and non-diagonal models, and non-unitary ones

Two cautions. First, the Kac table fixes which *holomorphic* weights may occur. A full CFT must also say which pairs $(h, \bar{h})$ occur and how often. The diagonal choice $\bar{h} = h$ is always consistent. Sometimes there are others; the 3-state Potts model at $c = \frac{4}{5}$ is one. Which combinations are allowed is decided by modular invariance on the torus, Nash's §4 (next part). Second, the FQS list assumes unitarity. Dropping it gives the non-unitary minimal models $\mathcal{M}(p,p')$ with coprime p, p' and $c = 1 - \frac{6(p-p')^2}{pp'}$. The Yang–Lee model $\mathcal{M}(5,2)$ has $c = -\frac{22}{5}$ and a primary with $h = -\frac{1}{5} < 0$, which our condition $h \geq 0$ forbids in a unitary theory.

Why "minimal": null vectors force the operator algebra to close

A CFT has an operator product expansion. If $\phi_{r,s}$ and $\phi_{r',s'}$ are in $\mathcal{M}_m$, their product must again be a sum of fields in $\mathcal{M}_m$. There are only finitely many to choose from, so the product must close on a finite set. What enforces this? The null vectors. In $L(c,h)$ a singular vector is *zero*, so the corresponding descendant field is the zero operator. Setting it to zero inside a correlation function gives a differential equation. We show how this works for the simplest case, a field with a level-2 null vector. (The full theory of fusion rules is in Nash's §5, the next part.)

Let $\phi := \phi_{1,2}$ with weight $h_\phi = h_{1,2}$. By (11.127) and (11.130) its level-2 null vector is $(L_{-2} + \alpha L_{-1}^2)|h_\phi\rangle$ with $\alpha = -\frac{3}{2(2h_\phi+1)}$. In the t parametrisation $h_{1,2} = \frac{3-2t}{4t}$, so $2h_\phi + 1 = \frac{3}{2t}$ and

$\alpha = -t$. Therefore the field equation is

$$\big(L_{-2}\phi\big)(z) = t\,\partial^2\phi(z), \tag{11.149}$$

using $L_{-1}\phi = \partial\phi$. Now sandwich $\phi(z)$ between two highest-weight states $|h\rangle = \phi_h(0)|0\rangle$ and $\langle h'|$ and define $G(z) \coloneqq \langle h'|\phi(z)|h\rangle$. This is a three-point function with the third field at infinity.

The z-dependence of G. By (11.101) with $m = 0$, $[L_0, \phi(z)] = (z\partial + h_\phi)\phi(z)$. Take the matrix element:

$$(h' - h)\,G(z) \overset{(1)}{=} \langle h'|\big(L_0\phi(z) - \phi(z)L_0\big)|h\rangle \overset{(2)}{=} (z\partial_z + h_\phi)\,G(z),$$

(1) $\langle h'|L_0 = h'\langle h'|$ (L_0 is Hermitian) and $L_0|h\rangle = h|h\rangle$.
(2) The commutator formula.

so $zG' = \kappa\,G$ with $\kappa \coloneqq h' - h - h_\phi$, and $G(z) = C\,z^\kappa$. The constant C is the OPE coefficient: $C \neq 0$ exactly when $[\phi_{h'}]$ appears in the product $\phi \times \phi_h$.

The descendant $L_{-2}\phi$ inside G. By Nash's (11.41) ((11.124)),

$$(L_{-2}\phi)(z) = \oint_z \frac{\mathrm{d}w}{2\pi i}(w - z)^{-1}T(w)\phi(z).$$

$$\begin{aligned}\langle h'|(L_{-2}\phi)(z)|h\rangle &\overset{(1)}{=} \oint_{|w|>|z|}\frac{\mathrm{d}w}{2\pi i}\frac{\langle h'|T(w)\phi(z)|h\rangle}{w-z} - \oint_{|w|<|z|}\frac{\mathrm{d}w}{2\pi i}\frac{\langle h'|\phi(z)T(w)|h\rangle}{w-z}\\ &\overset{(2)}{=} \sum_{k\geq 0} z^k\,\langle h'|L_{-k-2}\,\phi(z)|h\rangle + \sum_{k\geq 0} z^{-k-1}\langle h'|\phi(z)\,L_{k-1}|h\rangle\\ &\overset{(3)}{=} z^{-1}\langle h'|\phi(z)L_{-1}|h\rangle + h\,z^{-2}\,G(z) \overset{(4)}{=} -z^{-1}\,\partial_z G(z) + h\,z^{-2}\,G(z).\end{aligned}$$

(1) A small circle around z is the difference of a circle $|w| > |z|$ and a circle $|w| < |z|$ (both around 0), and radial ordering (11.99) puts $T(w)$ to the left when $|w| > |z|$ (Section 11.4.3).
(2) For $|w| > |z|$, $\frac{1}{w-z} = \sum_{k\geq 0} z^k w^{-k-1}$, and $\oint \frac{\mathrm{d}w}{2\pi i} w^{-k-1}T(w) = L_{-k-2}$ by the mode formula (11.94). For $|w| < |z|$, $\frac{1}{w-z} = -\sum_{k\geq 0} w^k z^{-k-1}$ and $\oint \frac{\mathrm{d}w}{2\pi i} w^k T(w) = L_{k-1}$; the two minus signs cancel.
(3) $\langle h'|L_{-n} = \big(L_n|h'\rangle\big)^\dagger = 0$ for $n \geq 1$, so the first sum vanishes. In the second, $L_{k-1}|h\rangle = 0$ for $k \geq 2$, and $L_0|h\rangle = h|h\rangle$ for $k = 1$.
(4) $\phi(z)L_{-1} = L_{-1}\phi(z) - [L_{-1}, \phi(z)] = L_{-1}\phi(z) - \partial\phi(z)$, and $\langle h'|L_{-1} = 0$.

The equation and its solutions. Insert this and $\langle h'|\partial^2\phi(z)|h\rangle = G''$ into (11.149):

$$0 \overset{(1)}{=} \Big(\frac{h}{z^2} - \frac{1}{z}\,\partial_z - t\,\partial_z^2\Big)C\,z^\kappa \overset{(2)}{=} \big(h - \kappa - t\,\kappa(\kappa-1)\big)\,C\,z^{\kappa-2}. \tag{11.150}$$

(1) The null-vector equation between $\langle h'|$ and $|h\rangle$.
(2) $\partial z^\kappa = \kappa z^{\kappa-1}$ and $\partial^2 z^\kappa = \kappa(\kappa-1)z^{\kappa-2}$.

So either $C = 0$, or κ solves the quadratic $t\kappa^2 + (1-t)\kappa - h = 0$. For a given h there are exactly two allowed values $h' = h + h_\phi + \kappa_\pm$. The product of $\phi_{1,2}$ with any primary contains *at most two* conformal families. With $h = h_{r,s}$ one checks (by inserting (11.140); we verified it by computer algebra) that the two roots are $h' = h_{r,s-1}$ and $h' = h_{r,s+1}$:

$$\phi_{1,2} \times \phi_{r,s} = \phi_{r,s-1} + \phi_{r,s+1}. \tag{11.151}$$

Worked example 11.50: Ising fusion rules from the null vector

In the Ising model $t = \frac{m+1}{m} = \frac{4}{3}$ and $\sigma = \phi_{1,2}$ with $h_\phi = \frac{1}{16}$. (Check: $\alpha = -\frac{3}{2(2/16+1)} = -\frac{3}{9/4} = -\frac{4}{3}$.)
$\sigma \times \sigma$. Put $h = \frac{1}{16}$:

$$\tfrac{4}{3}\kappa^2 - \tfrac{1}{3}\kappa - \tfrac{1}{16} = 0 \overset{(1)}{=} 64\kappa^2 - 16\kappa - 3 = 0 \overset{(2)}{=} \kappa = \frac{16 \pm \sqrt{256+768}}{128} = \frac{16 \pm 32}{128} = \frac{3}{8} \text{ or } -\frac{1}{8},$$

(1) Multiply by 48.
(2) Quadratic formula; $\sqrt{1024} = 32$.

and $h' = \frac{1}{16} + \frac{1}{16} + \kappa = \frac{1}{8} + \kappa \in \{\frac{1}{2}, 0\}$. So $\sigma \times \sigma = \mathbb{1} + \varepsilon$: two spins fuse into the identity (they can annihilate) or into the energy density (a bond). This is (11.151) with $(r, s) = (1, 2)$: $\phi_{1,1} + \phi_{1,3}$.
$\sigma \times \varepsilon$. Put $h = \frac{1}{2}$:

$$\tfrac{4}{3}\kappa^2 - \tfrac{1}{3}\kappa - \tfrac{1}{2} = 0 \overset{(1)}{=} 8\kappa^2 - 2\kappa - 3 = 0 \overset{(2)}{=} \kappa = \frac{2 \pm 10}{16} = \frac{3}{4} \text{ or } -\frac{1}{2},$$

(1) Multiply by 6.
(2) Quadratic formula; $\sqrt{4+96} = 10$.

so $h' = \frac{9}{16} + \kappa \in \{\frac{21}{16}, \frac{1}{16}\}$. The value $\frac{21}{16} = h_{1,4}$ is outside the Kac table. It is eliminated by the second null vector: ε is $\phi_{2,1}$ as well as $\phi_{1,3}$, so it has its own level-2 null vector, with $\alpha = -\frac{3}{2(2\cdot\frac{1}{2}+1)} = -\frac{3}{4}$, and the same computation with the roles of the two fields exchanged ($\varepsilon(z)$ between $\langle h'|$ and $|\sigma\rangle$) gives $\frac{3}{4}\kappa^2 + \frac{1}{4}\kappa - \frac{1}{16} = 0$, i.e. $\kappa \in \{\frac{1}{6}, -\frac{1}{2}\}$ and $h' = \frac{1}{16} + \frac{1}{2} + \kappa \in \{\frac{35}{48}, \frac{1}{16}\}$. The coupling $\langle \phi_{h'} \sigma \varepsilon \rangle$ must satisfy both equations, so $h' = \frac{1}{16}$: $\sigma \times \varepsilon = \sigma$. Truncation of (11.151) to the Kac table is exactly what makes the model minimal.

Try it: $\varepsilon \times \varepsilon$

Use the level-2 null vector of ε ($\alpha = -\frac{3}{4}$) with $h = \frac{1}{2}$ to find the families in $\varepsilon \times \varepsilon$.
Answer: $\frac{3}{4}\kappa^2 + \frac{1}{4}\kappa - \frac{1}{2} = 0$, i.e. $3\kappa^2 + \kappa - 2 = 0$, so $\kappa \in \{\frac{2}{3}, -1\}$ and $h' = 1 + \kappa \in \{\frac{5}{3}, 0\}$. The value $\frac{5}{3} = h_{3,1}$ lies outside the table (it is removed by the level-3 null vector of $\varepsilon = \phi_{1,3}$), leaving $\varepsilon \times \varepsilon = \mathbb{1}$. In the free fermion this is $\psi(z)\psi(w) \sim \frac{1}{z-w}$: two fermions fuse only into the identity family.

Physical picture: the whole story in one sentence

A null vector is a relation among the states of a family. As a field equation it becomes a differential equation for correlators, and differential equations have only finitely many solutions (here, two exponents $\kappa_\pm$). This finiteness, imposed at every null vector, is what cuts the operator algebra down to the $\frac{m(m-1)}{2}$ families of the Kac table.

11.4.13 The global subalgebra $\mathfrak{sl}(2, \mathbb{C})$ and the vacuum

The Virasoro algebra has infinitely many generators. Which of them are symmetries in the everyday sense, i.e. leave the vacuum unchanged? Nash answers in two lines, (11.46)–(11.47): exactly the three generators L_{-1}, L_0, L_1. In this subsection we explain why, what these three generators do, and what they buy us: the two-point functions, the $\langle TT \rangle$ correlator, and the behaviour of T at infinity that we used in Section 11.3.12.

The subalgebra. For $m, n \in \{-1, 0, 1\}$ the central term $\frac{c}{12} m(m^2-1)\delta_{m+n,0}$ of (11.102) vanishes, because $m(m^2-1) = (m-1)m(m+1)$ is zero for $m = 0, \pm 1$. So L_{-1}, L_0, L_1 close among themselves, as in (11.103):

$$[L_0, L_{-1}] = L_{-1}, \qquad [L_0, L_1] = -L_1, \qquad [L_1, L_{-1}] = 2L_0.$$

We have met this algebra in three guises already. It is worth putting them side by side, since each is a different *representation* of the same three commutators.

Dictionary: three faces of $\mathfrak{sl}(2,\mathbb{C})$

vector fields on the sphere	$\ell_{-1} = -\partial_z$ (translation), $\ell_0 = -z\partial_z$ (dilation and rotation), $\ell_1 = -z^2\partial_z$ (special conformal), from Section 11.3.3
2×2 matrices	the traceless matrices generating $\mathrm{SL}(2,\mathbb{C})$ acting by $z \mapsto \frac{az+b}{cz+d}$, Section 11.3.4
operators on states	L_{-1}, L_0, L_1 acting on the space of states; L_{-1} is the momentum P, L_0 the dilation (energy on the cylinder), L_1 the special conformal generator K

The angular momentum algebra is a useful comparison. Put $J_3 = L_0$, $J_+ = L_{-1}$, $J_- = L_1$. Then $[J_3, J_\pm] = \pm J_\pm$, and $[J_-, J_+] = 2J_3$. This is $[J_+, J_-] = 2J_3$ with the opposite sign: it is the algebra of $\mathrm{SU}(1,1) \cong \mathrm{SL}(2,\mathbb{R})$, the non-compact real form, not of $\mathrm{SU}(2)$. That sign has a visible consequence in the next example.

Worked example 11.51: The $\mathfrak{sl}(2)$ tower on a primary

Let $|h\rangle$ be a highest-weight state and $e_k := L_{-1}^k |h\rangle$, $k = 0, 1, 2, \dots$ These are the states $\partial^k \phi(0)|0\rangle$. First,

$$L_1 e_k \overset{(1)}{=} \sum_{j=0}^{k-1} L_{-1}^j \, [L_1, L_{-1}] \, L_{-1}^{k-1-j} |h\rangle \overset{(2)}{=} \sum_{j=0}^{k-1} L_{-1}^j \, 2L_0 \, L_{-1}^{k-1-j} |h\rangle$$

$$\overset{(3)}{=} \sum_{j=0}^{k-1} 2\,(h + k - 1 - j)\, e_{k-1} \overset{(4)}{=} k\,(2h + k - 1)\, e_{k-1}.$$

(1) Move L_1 to the right through the k factors, one commutator at a time; at the end $L_1|h\rangle = 0$.

(2) $[L_1, L_{-1}] = 2L_0$.

(3) L_0 acts on $L_{-1}^{k-1-j}|h\rangle$ with eigenvalue $h + k - 1 - j$; the remaining L_{-1}^j makes the total e_{k-1}.

(4) $\sum_{j=0}^{k-1}(h + k - 1 - j) = kh + \frac{k(k-1)}{2}$; doubling gives $k(2h + k - 1)$.

In the basis $(e_0, e_1, e_2, \dots)$ the three generators are the infinite matrices

$$L_0 = \begin{pmatrix} h & & & \\ & h+1 & & \\ & & h+2 & \\ & & & \ddots \end{pmatrix}, \quad L_{-1} = \begin{pmatrix} 0 & & & \\ 1 & 0 & & \\ & 1 & 0 & \\ & & \ddots & \ddots \end{pmatrix}, \quad L_1 = \begin{pmatrix} 0 & 2h & & \\ & 0 & 2(2h+1) & \\ & & 0 & \ddots \\ & & & \ddots \end{pmatrix}.$$

The norms follow by induction:

$$\|e_k\|^2 \overset{(1)}{=} \langle h | L_1^k L_{-1}^k | h \rangle \overset{(2)}{=} k(2h + k - 1)\, \|e_{k-1}\|^2 \overset{(3)}{=} k!\,(2h)(2h+1)\cdots(2h+k-1).$$

(1) $(L_{-1}^k)^\dagger = L_1^k$.

(2) $\|e_k\|^2 = \langle e_{k-1}|L_1 e_k\rangle$ and the formula for $L_1 e_k$.
(3) Iterate down to $\|e_0\|^2 = 1$.

For $h > 0$ every norm is positive and the tower never ends: an infinite-dimensional unitary representation, as it must be for a non-compact group. (For SU(2) the analogous coefficient is $k(2j - k + 1)$, with a minus sign, and it vanishes at $k = 2j + 1$: the tower ends and the representation is finite, with $2j + 1$ states.) For $h = 0$ the tower stops immediately, $\|e_1\|^2 = 2h = 0$: this is the vacuum, next.

The vacuum is an $\mathfrak{sl}(2)$ singlet. Regularity of $T(z)|0\rangle$ at $z = 0$ gave $L_n|0\rangle = 0$ for all $n \geq -1$ in (11.117). Nash obtains the L_{-1} part differently, and his argument is instructive because it uses only the Hilbert-space structure. Start from (11.46),

$$L_0|0\rangle = L_1|0\rangle = 0, \tag{11.152}$$

which says that the vacuum has zero energy and is a highest-weight state with $h = 0$. Then (11.47):

$$\|L_{-1}|0\rangle\|^2 \overset{(1)}{=} \langle 0|L_1 L_{-1}|0\rangle \overset{(2)}{=} \langle 0|[L_1, L_{-1}]|0\rangle \overset{(3)}{=} 2\langle 0|L_0|0\rangle \overset{(4)}{=} 0. \tag{11.153}$$

(1) $L_{-1}^\dagger = L_1$, (11.119).
(2) $L_1|0\rangle = 0$, so $L_1 L_{-1}|0\rangle = [L_1, L_{-1}]|0\rangle$.
(3) $[L_1, L_{-1}] = 2L_0$.
(4) $L_0|0\rangle = 0$.

In a Hilbert space a vector of zero norm is zero, so $L_{-1}|0\rangle = 0$. The vacuum is annihilated by the whole of $\mathfrak{sl}(2,\mathbb{C})$ and is invariant under the Möbius group. Taking adjoints of $L_n|0\rangle = 0$ ($n \geq -1$) gives $\langle 0|L_n = 0$ for $n \leq 1$. So:

$$L_n|0\rangle = 0 \ \ (n \geq -1), \qquad \langle 0|L_n = 0 \ \ (n \leq 1), \qquad \text{both for } n \in \{-1, 0, 1\}. \tag{11.154}$$

The in-vacuum and the out-vacuum are invariant under exactly the same three generators. Figure 11.26 shows the bookkeeping.

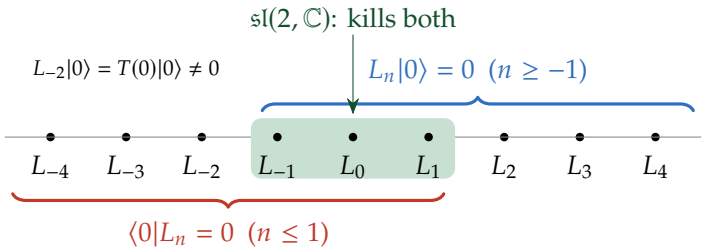


Figure 11.26: Which Virasoro modes annihilate the vacuum. The blue brace marks the modes that kill the in-vacuum $|0\rangle$ (regularity of $T(z)|0\rangle$ at $z = 0$); the red brace marks those that kill the out-vacuum $\langle 0|$ (regularity at $z = \infty$). Only the overlap $\{L_{-1}, L_0, L_1\}$ (green) kills both, so only the global conformal transformations are symmetries of vacuum correlators. The modes $L_{n\leq -2}$ create states, starting with the stress-tensor state $L_{-2}|0\rangle$.

The global Ward identities. Take a correlator of primaries $G = \langle 0|\phi_1(z_1)\cdots\phi_k(z_k)|0\rangle$ (suppressing $\bar z$). For $n \in \{-1, 0, 1\}$,

$$0 \overset{(1)}{=} \langle 0|L_n\,\phi_1\cdots\phi_k|0\rangle - \langle 0|\phi_1\cdots\phi_k\,L_n|0\rangle \overset{(2)}{=} \sum_{i=1}^{k}\langle 0|\phi_1\cdots[L_n,\phi_i(z_i)]\cdots\phi_k|0\rangle$$
$$\overset{(3)}{=} \sum_{i=1}^{k}\Big(z_i^{n+1}\partial_{z_i} + (n+1)h_i\,z_i^n\Big)G.$$

(1) Both terms vanish by (11.154).
(2) The commutator of L_n with a product is the sum of commutators with each factor (Leibniz rule).
(3) (11.101): $[L_n,\phi(z)] = \big(z^{n+1}\partial + (n+1)hz^n\big)\phi(z)$ for a primary; this operator does not act on the other fields.

Written out, these are three first-order differential equations:

$$\sum_i \partial_i G = 0, \qquad \sum_i \big(z_i\partial_i + h_i\big)G = 0, \qquad \sum_i \big(z_i^2\partial_i + 2h_i z_i\big)G = 0. \tag{11.155}$$

They are the statements that G is invariant under translations, under dilations (with the weights h_i as scaling dimensions) and under special conformal transformations. The first two hold in any scale-invariant theory. The third is the new constraint of conformal symmetry.

Worked example 11.52: The two-point function from the global Ward identities

Take $k = 2$. The first equation says G depends only on $z_{12} = z_1 - z_2$. The second then says $(z_1 - z_2)G'(z_{12}) + (h_1 + h_2)G = 0$, so $G = C\,z_{12}^{-(h_1+h_2)}$. Insert this into the third:

$$\big(z_1^2\partial_1 + z_2^2\partial_2 + 2h_1z_1 + 2h_2z_2\big)G \overset{(1)}{=} \Big(-(h_1+h_2)\frac{z_1^2 - z_2^2}{z_{12}} + 2h_1z_1 + 2h_2z_2\Big)G$$
$$\overset{(2)}{=} \big(-(h_1+h_2)(z_1+z_2) + 2h_1z_1 + 2h_2z_2\big)G$$
$$\overset{(3)}{=} (h_1 - h_2)(z_1 - z_2)\,G.$$

(1) $\partial_1 G = -(h_1+h_2)G/z_{12}$ and $\partial_2 G = +(h_1+h_2)G/z_{12}$.
(2) $z_1^2 - z_2^2 = (z_1+z_2)z_{12}$.
(3) Collect: the coefficient of z_1 is $h_1 - h_2$, that of z_2 is $h_2 - h_1$.

This must vanish. So either $C = 0$ or $h_1 = h_2$:

$$\langle\phi_1(z_1)\phi_2(z_2)\rangle = \frac{C\,\delta_{h_1h_2}}{z_{12}^{2h_1}}.$$

Scale invariance alone gave only $z_{12}^{-(h_1+h_2)}$, as in Nash's (11.9); the special conformal generator adds that *fields of different weights are uncorrelated*. This is the field-theory version of the orthogonality of L_0 eigenspaces. The same three equations fix the three-point function up to one constant, $\langle\phi_1\phi_2\phi_3\rangle = C_{123}\,z_{12}^{h_3-h_1-h_2}z_{23}^{h_1-h_2-h_3}z_{13}^{h_2-h_1-h_3}$; the next part of these notes develops it.

Try it: the dilation equation for the three-point function

Check that the three-point function just quoted satisfies $\sum_i (z_i\partial_i + h_i)G = 0$.
Answer: G is homogeneous in the z_i of degree $(h_3-h_1-h_2)+(h_1-h_2-h_3)+(h_2-h_1-h_3) = -(h_1+h_2+h_3)$. Euler's theorem for homogeneous functions gives $\sum_i z_i\partial_i G = -(h_1+h_2+h_3)G$, which cancels $\sum_i h_i G$.

Worked example 11.53: The TT two-point function from the modes

The stress tensor is not primary, so the example above does not apply directly. The modes do the job. For $|z| > |w|$,

$$\begin{aligned}\langle 0|T(z)T(w)|0\rangle &\overset{(1)}{=} \sum_{m,n\in\mathbb{Z}} z^{-m-2}w^{-n-2}\langle 0|L_mL_n|0\rangle \overset{(2)}{=} \sum_{m\geq 2} z^{-m-2}w^{m-2}\langle 0|L_mL_{-m}|0\rangle \\ &\overset{(3)}{=} \frac{c}{12}\sum_{m\geq 2} m(m^2-1)\,z^{-m-2}w^{m-2} \overset{(4)}{=} \frac{c}{2}z^{-4}\sum_{j\geq 0}\binom{j+3}{3}\left(\frac{w}{z}\right)^j \\ &\overset{(5)}{=} \frac{c}{2}z^{-4}\left(1-\frac{w}{z}\right)^{-4} = \frac{c/2}{(z-w)^4}.\end{aligned}$$

(1) The mode expansion (11.94), with $T(z)$ to the left because $|z| > |w|$.
(2) $L_n|0\rangle = 0$ unless $n \leq -2$, and $\langle 0|L_m = 0$ unless $m \geq 2$; also $\langle 0|L_mL_n|0\rangle = 0$ unless $m+n = 0$ (the state $L_mL_n|0\rangle$ has L_0 eigenvalue $-(m+n)$, and it must be 0 to overlap $\langle 0|$).
(3) $\langle 0|L_mL_{-m}|0\rangle = \langle 0|[L_m, L_{-m}]|0\rangle = \frac{c}{12}m(m^2-1)$ as in (11.143) with $h = 0$.
(4) Put $m = j+2$: $\frac{c}{12}(j+3)(j+2)(j+1) = \frac{c}{2}\binom{j+3}{3}$, and $z^{-m-2}w^{m-2} = z^{-4}(w/z)^j$.
(5) The binomial series $\sum_{j\geq 0}\binom{j+3}{3}x^j = (1-x)^{-4}$ (differentiate the geometric series three times and divide by 3!).

Also $\langle 0|T(z)|0\rangle = \sum_n z^{-n-2}\langle 0|L_n|0\rangle = 0$, since $\langle 0|L_n|0\rangle$ vanishes for $n \leq 1$ by (11.154) and for $n \geq -1$ by (11.117). On the plane the vacuum energy density is zero; on the cylinder it is $-\frac{c}{24}$ (Section 11.3.14). The coefficient $\frac{c}{2}$ is the one in the TT OPE (11.89), as it must be: the leading term of the OPE is the two-point function.

The stress tensor at infinity. In Section 11.3.12 we used that correlators of $T(z)$ decay like z^{-4} as $z \to \infty$. Here is why. The point $z = \infty$ of the sphere is an ordinary point, with local coordinate $w = -1/z$. The map $z \mapsto w = -1/z$ is Möbius, so its Schwarzian derivative vanishes (Nash's (11.32), proved in Section 11.3.14), and T transforms as an honest weight-2 tensor:

$$\tilde{T}(w) \overset{(1)}{=} \left(\frac{\mathrm{d}z}{\mathrm{d}w}\right)^2 T(z) + \frac{c}{12}\{z,w\} \overset{(2)}{=} \left(\frac{1}{w^2}\right)^2 T(z) \overset{(3)}{=} z^4\,T(z).$$

(1) The finite transformation law (11.91) of T under the change of coordinate $z = z(w) = -1/w$.
(2) $\frac{\mathrm{d}z}{\mathrm{d}w} = \frac{1}{w^2}$, and the Schwarzian of a Möbius map is zero.
(3) $w^{-4} = z^4$.

The field $\tilde{T}(w)$ is the stress tensor in the coordinate chart around ∞. If nothing is inserted at ∞ (the out-state is the vacuum), then $\tilde{T}(w)$ must be regular at $w = 0$. That is, $z^4T(z)$ stays bounded as $z \to \infty$, so $T(z) = O(z^{-4})$. In modes: $z^4T(z) = \sum_n L_n z^{2-n}$ is bounded at large z inside $\langle 0|\cdots$ precisely when $\langle 0|L_n = 0$ for $n \leq 1$, which is (11.154) again. (The Ward identity check in Example 11.30 showed the same w^{-4} falloff directly.)

The true symmetry group. Nash concludes (p. 311) that "the true invariance of the theory is that of the Möbius group $\mathrm{SL}(2,\mathbb{C})/\{\mp I\}$." Three facts combine here.

1. *Only the Möbius maps are global.* A conformal (holomorphic) map of the whole Riemann sphere to itself which is invertible is a rational function of degree one, i.e. Möbius (Section 11.3.4). The other ℓ_n, such as $\ell_2 = -z^3\partial_z$, generate vector fields with poles at 0 or ∞; they do not exponentiate to maps of the sphere.

2. *Only the Möbius generators fix the vacuum*, by (11.154). The remaining L_n are still useful: they give the local Ward identities and organise the states into representations. But they do not leave $|0\rangle$ alone.

3. *Under Möbius maps T is a true tensor.* The Schwarzian term in (11.91) vanishes exactly for Möbius maps (Nash's (11.32), Section 11.3.14), so under the global group T transforms like a primary of weight $(2,0)$, even though it is not primary under the full local algebra. The central charge is invisible to the global group.

Watch out: "infinite-dimensional symmetry" does not mean infinitely many conserved charges of the vacuum

It is tempting to read the Virasoro algebra as an infinite set of symmetries under which everything is invariant. It is not. The L_n with $|n| \geq 2$ are generated by $\oint \epsilon(z)T(z)$ with $\epsilon(z) = z^{n+1}$ singular somewhere on the sphere. They are symmetries of the *local* structure: they produce Ward identities that relate correlators with and without T insertions. The vacuum is not invariant under them; in fact they create the whole vacuum module $L_{-2}|0\rangle, L_{-3}|0\rangle, L_{-2}^2|0\rangle, \dots$ out of it.

11.4.14 Other sources of conformal field theories

Every CFT carries a representation of $Vir \oplus \overline{Vir}$, so any construction of such representations is a potential source of CFTs. Nash (p. 311) names two: representations of Kac–Moody algebras (studied as CFTs by Gepner and Witten), and superconformal algebras (Friedan, Martinec and Shenker). We describe both concretely. The mechanism in each case is the same: a bigger chiral algebra, containing $T(z)$, whose representations are easier to build and control.

Current algebras and the Sugawara construction

Currents. Suppose a CFT has a continuous internal symmetry with a compact Lie group G. Noether's theorem gives conserved currents. In two dimensions, as for T_{ab}, conservation and tracelessness split a current into a holomorphic part $J^a(z)$ and an antiholomorphic part. The J^a, $a = 1, \dots, \dim G$, are primaries of weight $(1,0)$. Their OPE is fixed by symmetry up to one number k:

$$J^a(z)J^b(w) = \frac{k\,\delta^{ab}}{(z-w)^2} + \frac{i f^{abc} J^c(w)}{z-w} + \text{reg.} \tag{11.156}$$

Here f^{abc} are the structure constants of $\mathfrak{g}$, $[T^a, T^b] = i f^{abc} T^c$, normalised by $f^{acd} f^{bcd} = 2h^\vee \delta^{ab}$, where $h^\vee$ is the dual Coxeter number ($h^\vee = N$ for $\mathrm{SU}(N)$, and $h^\vee = 0$ for $\mathrm{U}(1)$). The simple pole says that the charges rotate the currents into each other as the adjoint representation; the double pole is the Schwinger term, the analogue of the central term in TT. The integer k is the **level**. (Why an integer: see the WZW remark below.)

Worked example 11.54: The affine Kac–Moody algebra from the JJ OPE

Expand $J^a(z) = \sum_n J^a_n z^{-n-1}$, so $J^a_n = \oint_0 \frac{\mathrm{d}z}{2\pi i} z^n J^a(z)$. As for the oscillators in Example 11.37,

$$[J^a_m, J^b_n] \overset{(1)}{=} \oint_0 \frac{\mathrm{d}z}{2\pi i} z^n \oint_z \frac{\mathrm{d}w}{2\pi i} w^m \Big(\frac{k\delta^{ab}}{(w-z)^2} + \frac{i f^{abc} J^c(z)}{w-z} \Big)$$
$$\overset{(2)}{=} \oint_0 \frac{\mathrm{d}z}{2\pi i} \Big(k\delta^{ab}\, m\, z^{m+n-1} + i f^{abc} z^{m+n} J^c(z) \Big) \overset{(3)}{=} i f^{abc} J^c_{m+n} + k\, m\, \delta^{ab} \delta_{m+n,0}.$$

(1) The contour commutator (11.100) and the OPE (11.156) with the roles $w \leftrightarrow z$ (the OPE is symmetric after relabelling $a \leftrightarrow b$, c).
(2) Residues at $w = z$: of $w^m/(w-z)^2$ it is $m z^{m-1}$, of $w^m/(w-z)$ it is z^m.
(3) $\oint z^{m+n-1} = \delta_{m+n,0}$ and $\oint z^{m+n} J^c = J^c_{m+n}$.

This is the **affine Kac–Moody algebra** $\hat{\mathfrak{g}}_k$, the central extension of the loop algebra $\mathfrak{g} \otimes \mathbb{C}[z, z^{-1}]$ met in Chapter 6. The zero modes J^a_0 obey $[J^a_0, J^b_0] = i f^{abc} J^c_0$: the ordinary Lie algebra, the global charges. For $G = \mathrm{U}(1)$ we recover the oscillators $[a_m, a_n] = m\delta_{m+n,0}$ of (11.104)–(11.105) with $k = 1$.

The Sugawara construction. Given the currents, there is a stress tensor built out of them:

Key idea: Sugawara stress tensor

$$T(z) = \frac{1}{2(k+h^\vee)} \sum_a :J^a J^a:(z), \qquad c = \frac{k\ \dim G}{k+h^\vee}. \tag{11.157}$$

With this T each J^a is primary of weight 1, and a primary transforming in the representation R of G has $h = \frac{C_2(R)}{k+h^\vee}$, with C_2 the quadratic Casimir.

The shift $k \to k + h^\vee$ is a quantum effect. Classically one would guess the prefactor $\frac{1}{2k}$, as for the free boson. But in computing the OPE $T(z)J^b(w)$ one must contract J^b with each J^a inside $:J^aJ^a:$. The double pole of (11.156) gives k and the simple pole, used twice, gives $f^{acd}f^{bcd} = 2h^\vee\delta^{ab}$. Requiring the coefficient of $J^b(w)/(z-w)^2$ to be 1 forces the prefactor $\frac{1}{2(k+h^\vee)}$. The central charge then follows from the fourth-order pole of TT.

For U(1) ($\dim G = 1$, $h^\vee = 0$, $k = 1$): $T = \frac{1}{2}:jj:$ and $c = 1$, which is (11.104). For SU(2) at level k: $\dim G = 3$, $h^\vee = 2$, so

$$c\big(\mathrm{SU}(2)_k\big) = \frac{3k}{k+2}, \qquad h_j = \frac{j(j+1)}{k+2}, \quad j = 0, \tfrac{1}{2}, \dots, \tfrac{k}{2}. \tag{11.158}$$

These are the central charges used in the coset construction of Section 11.4.11. At $k = 1$ we get $c = 1$, and indeed:

Worked example 11.55: $\mathrm{SU}(2)_1$ out of one free boson

With the free boson of Section 11.4.5, put

$$J^3 = j = i\sqrt{2}\,\partial\varphi, \qquad J^\pm = :e^{\pm 2i\varphi}:, \qquad J^1 = \tfrac{1}{\sqrt{2}}(J^+ + J^-), \quad J^2 = \tfrac{1}{i\sqrt{2}}(J^+ - J^-).$$

The vertex operators $e^{\pm 2i\varphi}$ have weight $\frac{\alpha^2}{4} = 1$ (Example 11.23), exactly right for currents. Their OPE follows from the chiral propagator $\langle\varphi(z)\varphi(w)\rangle = -\frac{1}{2}\ln(z-w)$:

$$:e^{2i\varphi(z)}::e^{-2i\varphi(w)}: \overset{(1)}{=} e^{(2i)(-2i)\langle\varphi(z)\varphi(w)\rangle}\, :e^{2i\varphi(z)-2i\varphi(w)}: \overset{(2)}{=} (z-w)^{-2}\, :e^{2i(\varphi(z)-\varphi(w))}:$$

$$\overset{(3)}{=} \frac{1}{(z-w)^2} + \frac{2i\,\partial\varphi(w)}{z-w} + \text{reg.} = \frac{1}{(z-w)^2} + \frac{\sqrt{2}\,J^3(w)}{z-w} + \text{reg.}$$

(1) Wick's theorem for exponentials: $:e^A::e^B: = e^{\langle AB\rangle}:e^{A+B}:$ for A, B linear in the free field (Section 11.3.9).
(2) $4\cdot\left(-\frac{1}{2}\ln(z-w)\right) = -2\ln(z-w)$.
(3) Taylor expand $\varphi(z)-\varphi(w) = (z-w)\partial\varphi(w)+\dots$ in the exponent. Finally $2i\partial\varphi = \sqrt{2}\,j$.

From this one gets $J^1J^1 \sim \frac{1}{2}(J^+J^- + J^-J^+) \sim \frac{1}{(z-w)^2}$, the double pole with $k=1$, and simple poles reproducing $f^{abc} = \sqrt{2}\,\epsilon^{abc}$ (for which $f^{acd}f^{bcd} = 4\delta^{ab}$, i.e. $h^\vee = 2$). The Sugawara tensor $\frac{1}{6}\sum_a :J^aJ^a:$ coincides with the boson's $\frac{1}{2}:jj:$, with $c=1$. The spin-$\frac{1}{2}$ primary has $h_{1/2} = \frac{3/4}{3} = \frac{1}{4}$, the weight of $:e^{\pm i\varphi}:$. One free boson at this special radius has an enhanced SU(2) symmetry.

WZW models. The CFTs whose chiral algebra is $\hat{\mathfrak{g}}_k$ are the Wess–Zumino–Witten models. Their field is a map $g:\Sigma\to G$ and their action is a sigma-model term plus k times the Wess–Zumino term $\frac{1}{12\pi}\int_B \operatorname{tr}(g^{-1}\mathrm{d}g)^3$, an integral over a 3-manifold B with boundary Σ. Different choices of B change this term by 2π times an integer (the integral of the closed 3-form over a closed 3-manifold, a degree in $\pi_3(G)=\mathbb{Z}$), so e^{iS} is well defined only if k is an integer. This is the topological origin of the quantised level, and of Nash's "level k representations". Gepner and Witten determined the spectrum of these models: the primaries are the integrable highest-weight representations of $\hat{\mathfrak{g}}_k$, finitely many for each k; for $\mathrm{SU}(2)_k$ they are the spins $j \le \frac{k}{2}$ of (11.158). Like the minimal models, they are rational CFTs, and their fusion rules are the subject of Nash's §5.

Superconformal field theories

The second source is supersymmetry. Add to $T(z)$ a fermionic current $G(z)$ of weight $\frac{3}{2}$, the supercurrent. The $\mathcal{N}=1$ **superconformal algebra** is defined by the OPEs

$$T(z)G(w) = \frac{\frac{3}{2}G(w)}{(z-w)^2} + \frac{\partial G(w)}{z-w}, \qquad G(z)G(w) = \frac{2c/3}{(z-w)^3} + \frac{2T(w)}{z-w}, \tag{11.159}$$

plus regular terms. The first says G is primary of weight $\frac{3}{2}$; the second says that G is a "square root" of T. Because G is a fermion it may be periodic or antiperiodic around the origin: $G(z) = \sum_r G_r z^{-r-3/2}$ with $r\in\mathbb{Z}+\frac{1}{2}$ (Neveu–Schwarz) or $r\in\mathbb{Z}$ (Ramond). The same contour computation as for the Virasoro algebra, with anticommutators for G, gives

$$\{G_r, G_s\} = 2L_{r+s} + \frac{c}{3}\Big(r^2-\frac{1}{4}\Big)\delta_{r+s,0}, \qquad [L_m, G_r] = \Big(\frac{m}{2}-r\Big)G_{m+r}.$$

The anticommutator $\{G_{-1/2}, G_{1/2}\} = 2L_0$ is the two-dimensional version of "$\{Q,Q^\dagger\}=2H$".

Worked example 11.56: The free superconformal multiplet: $c=\frac{3}{2}$

Take one free boson and one free Majorana fermion, with $T = -(\partial\varphi)^2 - \frac{1}{2}\psi\partial\psi$ and $c = 1+\frac{1}{2} = \frac{3}{2}$. Try $G = a\,\psi\,\partial\varphi$. The contractions are $\psi(z)\psi(w)\sim\frac{1}{z-w}$ and $\partial\varphi(z)\partial\varphi(w)\sim -\frac{1}{2(z-w)^2}$.

$$G(z)G(w) \overset{(1)}{=} a^2\Big[\frac{1}{z-w}\cdot\frac{-1}{2(z-w)^2} + \frac{:\partial\varphi(z)\partial\varphi(w):}{z-w} - \frac{:\psi(z)\psi(w):}{2(z-w)^2}\Big] + \dots$$

$$\overset{(2)}{=} a^2\Big[-\frac{1}{2(z-w)^3}+\frac{(\partial\varphi)^2(w)}{z-w}-\frac{:\partial\psi\,\psi:(w)}{2(z-w)}\Big]+\text{reg.}$$
$$\overset{(3)}{=} -\frac{a^2}{2}\Big[\frac{1}{(z-w)^3}+\frac{2\big(-(\partial\varphi)^2-\frac{1}{2}\psi\partial\psi\big)(w)}{z-w}\Big]+\text{reg.}$$

(1) Wick's theorem: both pairs contracted, the fermions only, or the bosons only ($\partial\varphi(z)$ commutes past $\psi(w)$, so there is no sign).
(2) Taylor expand: $:\partial\varphi(z)\partial\varphi(w):\,=(\partial\varphi)^2(w)+O(z-w)$ and $:\psi(z)\psi(w):\,=(z-w):\partial\psi\,\psi:(w)+O((z-w)^2)$, since $\psi(w)^2=0$.
(3) Factor out $-\frac{a^2}{2}$, and use $:\partial\psi\,\psi:\,=-:\psi\partial\psi:$ (fermions anticommute).

Compare with (11.159): we need $-\frac{a^2}{2}=1$, and then the triple pole is $1=\frac{2c}{3}$ with $c=\frac{3}{2}$, and the simple pole is $2T$. Both work with $a=i\sqrt{2}$: $G=i\sqrt{2}\,\psi\,\partial\varphi$. The boson and fermion are superpartners.

The superconformal algebra has its own Kac determinant and its own unitary discrete series (Friedan, Qiu and Shenker again),

$$c=\frac{3}{2}\Big(1-\frac{8}{m(m+2)}\Big),\qquad m=3,4,\ldots$$

The first member, $m=3$, has $c=\frac{3}{2}\cdot\frac{7}{15}=\frac{7}{10}$: the tricritical Ising model of Section 11.4.12 is supersymmetric. In string theory the superconformal algebra is the gauge symmetry of the worldsheet of the superstring: ten free multiplets give $c=10\cdot\frac{3}{2}=15$, which cancels the -15 of the superconformal ghosts (Chapter 9).

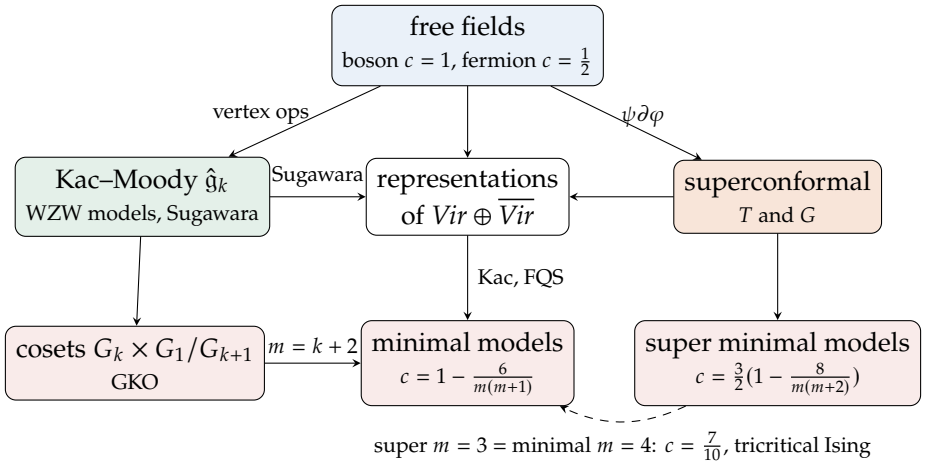


Figure 11.27: Where conformal field theories come from. Everything flows into representations of $Vir\oplus\overline{Vir}$ (centre). Free fields give the simplest ones; currents built from them (vertex operators) give Kac–Moody algebras, whose Sugawara stress tensors give Virasoro representations; differences of Sugawara tensors (cosets) realise the whole unitary minimal series; supercurrents give the superconformal series, whose first member coincides with the $m=4$ minimal model.

Reading Nash (pp. 310–311, (11.46)–(11.47))

Nash identifies the Möbius subalgebra $\mathfrak{sl}(2,\mathbb{C})=\{L_{-1},L_0,L_1\}$, states that the vacuum is annihilated by L_0 and L_1 (11.46), and shows $\|L_{-1}|0\rangle\|^2=0$ (11.47); we reproduced that chain, noted that it forces $L_{-1}|0\rangle=0$ only because we are in a Hilbert space, and gave the independent argument from regularity. He concludes that the true invariance is

$\mathrm{SL}(2,\mathbb{C})/\{\mp I\}$, that these are the only invertible conformal maps of the Riemann sphere, and that T is a true tensor under them because the Schwarzian vanishes. We added the $\mathfrak{sl}(2)$ tower, the global Ward identities and the two-point function, $\langle TT\rangle = \frac{c/2}{(z-w)^4}$ from modes, and the z^{-4} falloff of T. His closing paragraph of §3 lists Kac–Moody algebras (Gepner–Witten) and superconformal symmetry (Friedan–Martinec–Shenker) as further sources of CFTs; we derived the affine algebra from the JJ OPE, stated the Sugawara construction with the $\mathrm{SU}(2)_1$ example, and built the free $c = \frac{3}{2}$ superconformal multiplet. His §4 then turns to statistical mechanics, which is the next part of these notes.

Recap: Section 11.4

- The local conformal algebra is two copies of the Witt algebra (the complexified $\mathrm{Vect}(S^1)$); quantum mechanically each becomes the Virasoro algebra $[L_m, L_n] = (m-n)L_{m+n} + \frac{c}{12}m(m^2-1)\delta_{m+n,0}$, with $L_n = \oint \frac{\mathrm{d}z}{2\pi i} z^{n+1}T(z)$. The central term follows from the $\frac{c/2}{(z-w)^4}$ term of the TT OPE; for the free boson $c = 1$.
- Radial quantisation: time is $\ln|z|$, the Hamiltonian is $L_0 + \bar{L}_0$ (minus $\frac{c}{12}$ on the cylinder), $L_n^\dagger = L_{-n}$, and states correspond to fields: $|h\rangle = \phi(0)|0\rangle$ with $L_0|h\rangle = h|h\rangle$, $L_{n>0}|h\rangle = 0$.
- Descendants $L_{-n_1}\cdots L_{-n_k}|h\rangle$ form the Verma module $V(c,h)$, graded by level with $p(N)$ states. Singular vectors generate the null submodule N, and the physical representation is $L(c,h) = V/N$. For generic (c,h), $V = L$.
- The Gram matrices and the Kac determinant $\det M^{(N)} \propto \prod_{rs\le N}(h-h_{r,s})^{p(N-rs)}$ locate the singular vectors.
- Unitarity requires $h \ge 0$ and $c \ge 0$, allows everything with $c \ge 1$, $h \ge 0$, and for $c < 1$ only the discrete series $c = 1 - \frac{6}{m(m+1)}$, $h = h_{r,s}$ (Nash's (6.58), Friedan–Qiu–Shenker); the coset construction realises each point.
- The minimal models have $\frac{m(m-1)}{2}$ primaries; $m = 3$ is the Ising model with $h = 0, \frac{1}{16}, \frac{1}{2}$, giving $\eta = \frac{1}{4}$, $\nu = 1$. Null vectors become differential equations and truncate the fusion rules, e.g. $\sigma \times \sigma = \mathbb{1} + \varepsilon$.
- Only L_{-1}, L_0, L_1 annihilate both $|0\rangle$ and $\langle 0|$: the Möbius group is the true symmetry of the vacuum, and gives the global Ward identities.
- Further CFTs come from Kac–Moody algebras (Sugawara, $c = \frac{k\dim G}{k+h^\vee}$, WZW models) and superconformal algebras (the $c = \frac{3}{2}$ free multiplet, the super minimal series).

11.5 Critical phenomena and the two-dimensional Ising model

The first three sections of this chapter built an abstract machine. A two-dimensional conformal field theory has a holomorphic energy–momentum tensor $T(z)$, whose modes L_n obey the Virasoro algebra with central charge c. Its primary fields ϕ carry weights $(h, \bar{h})$, and every primary generates a tower of descendants. That is a lot of structure. A physicist is entitled to ask two blunt questions.

- *Where do such theories occur in nature?*

- *What do the numbers c and h measure?*

Nash's answer, in his §4, is statistical mechanics. A magnet at its Curie point, a liquid–gas mixture at its critical point, or a binary alloy at its ordering temperature has fluctuations on all length scales. Its long-distance physics is scale invariant, and in two dimensions it is conformally invariant. The conformal weights of the primary fields then *are* the measured critical exponents. The simplest example, the two-dimensional Ising model, turns out to be the CFT with $c = \frac{1}{2}$ and exactly three primary fields.

This section supplies the statistical-mechanics background Nash assumes: the lattice model, its phase transition, the critical exponents, and the dictionary between exponents and scaling dimensions. The next section (Section 11.6) identifies the Ising model with the $c = \frac{1}{2}$ representation theory. The two after that put the theory on a torus and check modular invariance, which is Nash's actual topic.

Watch out: Page numbers

Throughout this part, [Nash p. n] is the *printed* page number of Nash. The PDF page is printed +10. Our Nash range is printed pp. 311–321.

11.5.1 Where Section 3 leaves us: the Möbius group

Nash opens our page range by tidying up one point. The Virasoro algebra has infinitely many generators L_n, $n \in \mathbb{Z}$. Yet only three of them, L_{-1}, L_0, L_1, exponentiate to honest, globally defined, invertible transformations of the Riemann sphere $\hat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$. The others are only infinitesimal symmetries: they are holomorphic near a point, but they cannot be extended to invertible maps of the whole sphere.

Recall that L_n is represented on functions by the vector field

$$\ell_n = -z^{n+1}\frac{\partial}{\partial z}. \tag{11.160}$$

At $z = 0$ this is regular only for $n \geq -1$. To look at $z = \infty$, use the coordinate $w = -1/z$. Then $\partial_z = w^2\partial_w$ and $z^{n+1} = (-1)^{n+1}w^{-n-1}$, so

$$\begin{aligned} \ell_n &\overset{(1)}{=} -z^{n+1}\,\frac{\mathrm{d}w}{\mathrm{d}z}\,\frac{\partial}{\partial w} \\ &\overset{(2)}{=} -(-1)^{n+1}w^{-n-1}\,w^2\,\frac{\partial}{\partial w} \\ &= (-1)^n\,w^{1-n}\,\frac{\partial}{\partial w}. \end{aligned}$$

(1) Chain rule: a vector field $f(z)\partial_z$ becomes $f(z)(\mathrm{d}w/\mathrm{d}z)\partial_w$ in the new coordinate.
(2) $w = -1/z$ gives $\mathrm{d}w/\mathrm{d}z = 1/z^2 = w^2$ and $z = -1/w$.

This is regular at $w = 0$ (that is, at $z = \infty$) only for $n \leq 1$. So the vector fields regular on the *whole* sphere are exactly $\ell_{-1}, \ell_0, \ell_1$:

$$\underbrace{\ell_{-1} = -\partial_z}_{\text{translation}}, \qquad \underbrace{\ell_0 = -z\partial_z}_{\text{dilation, rotation}}, \qquad \underbrace{\ell_1 = -z^2\partial_z}_{\text{special conformal}}.$$

They generate the **Möbius group**.

Working definition 11.21 (Möbius transformation)

A Möbius transformation is a map of the Riemann sphere [the plane $\mathbb{C}$ with one point ∞ added, so that $m_M(-d/c)=\infty$ and $m_M(\infty)=a/c$] of the form

$$z \longmapsto m_M(z)=\frac{az+b}{cz+d}, \qquad M=\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}(2,\mathbb{C}), \quad ad-bc=1.$$

How to use it: to compose two Möbius maps, multiply their matrices. The matrices M and $-M$ give the same map, so the group of maps is $\mathrm{SL}(2,\mathbb{C})/\{\pm I\}=\mathrm{PSL}(2,\mathbb{C})$. It has three complex (six real) parameters, one for each of ℓ_{-1},ℓ_0,ℓ_1 [they integrate to the translations $z+b$, the rescalings with rotations λz, and the special conformal maps $z/(cz+1)$].

The composition rule is the reason for writing the map with a matrix. Let $M=\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ and $N=\left(\begin{smallmatrix} \alpha & \beta \\ \gamma & \delta \end{smallmatrix}\right)$. Then

$$\begin{aligned} m_M\big(m_N(z)\big) &\overset{(1)}{=} \frac{a\,\frac{\alpha z+\beta}{\gamma z+\delta}+b}{c\,\frac{\alpha z+\beta}{\gamma z+\delta}+d} \\ &\overset{(2)}{=} \frac{a(\alpha z+\beta)+b(\gamma z+\delta)}{c(\alpha z+\beta)+d(\gamma z+\delta)} \\ &\overset{(3)}{=} \frac{(a\alpha+b\gamma)z+(a\beta+b\delta)}{(c\alpha+d\gamma)z+(c\beta+d\delta)} = m_{MN}(z). \end{aligned}$$

(1) Substitute $m_N(z)$ for z in m_M.
(2) Multiply numerator and denominator by $\gamma z+\delta$.
(3) Collect powers of z; the coefficients are the entries of the matrix product MN.

Physical picture: Why only Möbius maps are global

A holomorphic bijection f of the sphere onto itself is a meromorphic function that takes the value ∞ exactly once, with multiplicity one. A meromorphic function on the sphere is a rational function $P(z)/Q(z)$, and it takes each value as many times as its degree. Degree one means $f=(az+b)/(cz+d)$. The infinitely many other "conformal transformations" generated by L_n, $|n|\geq 2$, are only local: $z\mapsto z+\varepsilon z^{n+1}$ is invertible near $z=0$ but not on all of $\hat{\mathbb{C}}$.

Nash's second remark concerns the energy–momentum tensor. Under a finite conformal map $z\mapsto z'$, [Nash p. 307, (11.31)] gives

$$T(z)=\left(\frac{\mathrm{d}z'}{\mathrm{d}z}\right)^2 T'(z')+\frac{c}{12}\{z',z\}, \qquad \{f,z\}=\frac{f'''}{f'}-\frac{3}{2}\left(\frac{f''}{f'}\right)^2. \tag{11.161}$$

The **Schwarzian derivative** $\{f,z\}$ is the obstruction to T being a tensor of weight 2. For a Möbius map it vanishes. Nash proves the converse; here is the direct check. Take $f(z)=(az+b)/(cz+d)$ with $ad-bc=1$:

$$\begin{aligned} f'(z) &\overset{(1)}{=} \frac{a(cz+d)-c(az+b)}{(cz+d)^2}=\frac{1}{(cz+d)^2}, \\ f''(z) &\overset{(2)}{=} \frac{-2c}{(cz+d)^3}, \qquad f'''(z)=\frac{6c^2}{(cz+d)^4}, \\ \{f,z\} &\overset{(3)}{=} \frac{6c^2}{(cz+d)^2}-\frac{3}{2}\left(\frac{-2c}{cz+d}\right)^2=\frac{6c^2-6c^2}{(cz+d)^2}=0. \end{aligned}$$

(1) Quotient rule, then $ad - bc = 1$.
(2) Differentiate $(cz + d)^{-2}$ twice more.
(3) $f'''/f' = 6c^2/(cz + d)^2$ and $f''/f' = -2c/(cz + d)$.

So *with respect to Möbius maps, $T(z)$ is a true tensor of weight* $(2, 0)$. This is the global statement behind the fact, proved in §3, that the vacuum is annihilated by L_{-1}, L_0, L_1 [Nash p. 310, (11.46)–(11.47)]: the vacuum is Möbius invariant, although it is not invariant under the whole Virasoro algebra. We will use Möbius invariance repeatedly in Section 11.9 to fix two- and three-point functions.

Watch out: Two different "conformal groups"

In the physics literature "the conformal group in two dimensions" means either the global group $\mathrm{PSL}(2, \mathbb{C})$ (finite dimensional, acts on the whole sphere) or the local algebra $\mathrm{Vir} \oplus \overline{\mathrm{Vir}}$ (infinite dimensional, acts only near a point, and centrally extended). The correlation functions on the sphere are exactly invariant under the first. They obey Ward identities, not invariance, under the second.

Where conformal field theories come from

Nash then observes that a CFT is, at heart, a collection of representations of $\mathrm{Vir} \oplus \overline{\mathrm{Vir}}$. Anything that produces Virasoro representations can therefore produce CFTs. He lists three sources, and his §4 adds a fourth. We collect them in Figure 11.28.

- **Kac–Moody (affine Lie) algebras.** In chapter 6 we met the loop group LG and its central extension. A representation of its Lie algebra at level k yields a Virasoro representation through the **Sugawara construction** $T(z) = \frac{1}{2k + c(G)} \sum_a : J^a(z) J^a(z) :$ in Nash's normalisation, with central charge $c = 2k \dim G/(2k + c(G))$ [Nash p. 187, (6.60)]. In the common normalisation where the level is written k and the dual Coxeter number $h^\vee$, this reads $c = k \dim G/(k + h^\vee)$; for SU(2), $c = 3k/(k + 2)$. The CFTs built this way are the **Wess–Zumino–Witten (WZW) models**; Gepner and Witten studied them as string backgrounds. They reappear in chapter 12, where their conformal blocks become the state spaces of Chern–Simons theory.

- **Superconformal symmetry.** Add a fermionic current $G(z)$ of weight $\frac{3}{2}$ to $T(z)$. The resulting super-Virasoro algebra gives two-dimensional superconformal field theories (Friedan, Martinec and Shenker), the world-sheet theories of the superstring. The simplest example is a free boson plus a free Majorana fermion, with $c = 1 + \frac{1}{2} = \frac{3}{2}$.

- **Strings.** The bosonic string is a CFT with $c = 26$ on every Riemann surface (chapter 9, and Section 11.11 below).

- **Critical statistical mechanics.** The subject of §4. Nash promises "an example of a conformal field theory of genus one". He means that the natural observable, the partition function with periodic boundary conditions, lives on a torus.

11.5.2 The Ising model on a square lattice

The Ising model is a cartoon of a uniaxial magnet. Put a spin $s_i \in \{+1, -1\}$ at every site i of a square lattice with N sites. Each spin likes to be aligned with its four nearest neighbours. The

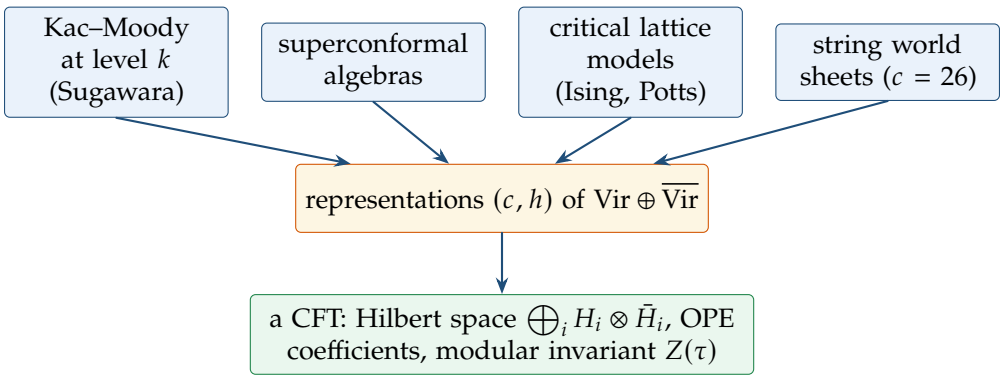


Figure 11.28: Four routes to a conformal field theory. Each source on top produces representations of two commuting Virasoro algebras; a CFT is a consistent way of assembling them. What to see: the Virasoro representation theory of §3 is the common currency, and statistical mechanics is only one of its customers.

energy of a configuration $\{s\}$ is

$$\mathcal{H}[s] = -J \sum_{\langle ij \rangle} s_i s_j \; - \; B \sum_i s_i, \tag{11.162}$$

where $\langle ij \rangle$ runs over nearest-neighbour pairs (bonds), $J > 0$ is the ferromagnetic coupling and B an external magnetic field. At temperature T the probability of a configuration is $e^{-\beta \mathcal{H}}/Z$ with $\beta = 1/k_B T$, and

$$Z(\beta, B) = \sum_{\{s\}} e^{-\beta \mathcal{H}[s]} \tag{11.163}$$

is the **partition function**. It is a finite sum of 2^N terms. All thermodynamics follows from it: the free energy per site is $f = -(k_B T/N) \ln Z$, the magnetisation per site is $M = \langle s_i \rangle = -\partial f / \partial B$, and so on. We write $K = \beta J$ for the dimensionless coupling.

Physical picture: Two competing tendencies

At low temperature (K large) energy wins. Almost all spins point the same way, and $M \neq 0$ even at $B = 0$: the system has spontaneously chosen a direction. At high temperature (K small) entropy wins. Spins fluctuate nearly independently and $M = 0$. Somewhere in between, at a critical temperature T_c, the order disappears. In two dimensions this happens continuously: $M \to 0$ as $T \to T_c^-$ with no latent heat. This is a **second-order** (continuous) phase transition.

Periodic boundary conditions put the model on a torus

A finite lattice has edges, and edge spins have fewer neighbours. The standard cure is to identify the right edge with the left one and the top edge with the bottom one. Then every site has four neighbours. But a rectangle with opposite sides identified is a *torus* (Figure 11.29). This innocent choice is the reason Nash's section on statistical mechanics is about modular invariance: the Ising partition function with periodic boundary conditions is a partition function on a torus.

High- and low-temperature expansions, and where T_c is

We can locate T_c exactly, without Onsager's full solution, by a beautiful argument of Kramers and Wannier. It also introduces the key idea that configurations of the model are *loops*, a picture that the continuum theory inherits.

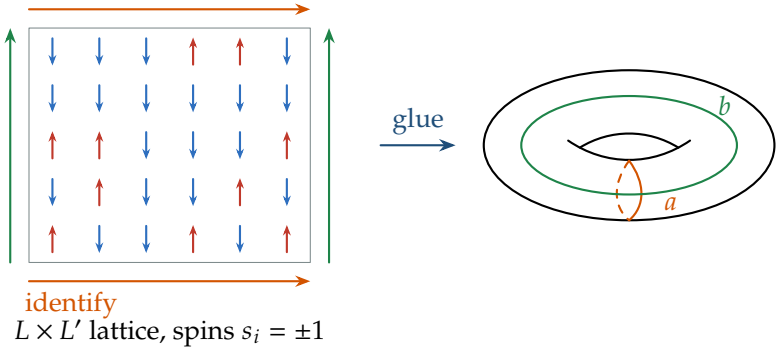


Figure 11.29: Periodic boundary conditions. Identifying the opposite edges of the lattice (orange with orange, green with green) makes every site equivalent and turns the rectangle into a torus with two independent cycles a and b. What to see: the lattice partition function is secretly a partition function on a torus, and its continuum limit will depend on the torus's shape through a single complex number τ (Section 11.7).

High-temperature expansion. Set $B = 0$ and $t_K = \tanh K$. Because $s_i s_j = \pm 1$, the Boltzmann weight of one bond is

$$\begin{aligned} e^{K s_i s_j} &\overset{(1)}{=} \cosh K + s_i s_j \sinh K \\ &= \cosh K \left(1 + s_i s_j \, t_K\right). \end{aligned}$$

(1) For $x = \pm 1$, $e^{Kx} = \cosh K + x \sinh K$: check $x = 1$ and $x = -1$ separately.

Now multiply over all $2N$ bonds (a periodic square lattice with N sites has $2N$ bonds) and sum over spins:

$$\begin{aligned} Z &\overset{(1)}{=} (\cosh K)^{2N} \sum_{\{s\}} \prod_{\langle ij \rangle} (1 + s_i s_j \, t_K) \\ &\overset{(2)}{=} (\cosh K)^{2N} \sum_{G \subseteq \text{bonds}} t_K^{|G|} \sum_{\{s\}} \prod_{\langle ij \rangle \in G} s_i s_j \\ &\overset{(3)}{=} 2^N (\cosh K)^{2N} \sum_{G \text{ closed}} t_K^{|G|}. \end{aligned} \tag{11.164}$$

(1) Pull the factor $\cosh K$ out of every bond.

(2) Expand the product: each bond contributes either 1 or $t_K s_i s_j$. A term is labelled by the set G of bonds that contributed $t_K s_i s_j$; $|G|$ is the number of such bonds.

(3) In $\prod_G s_i s_j$ the spin s_i appears once for every bond of G touching i. Since $\sum_{s=\pm 1} s = 0$ and $\sum_{s=\pm 1} s^2 = 2$, the spin sum is 2^N if every site touches an even number of bonds of G, and zero otherwise. Such G are unions of closed loops: "closed graphs".

Low-temperature expansion. Start from the all-up configuration, with weight e^{2NK}. Any other configuration is described by its domain walls: the set of bonds with $s_i \neq s_j$. Draw each such wall as the segment of the *dual* lattice (the lattice of plaquette centres) that crosses the broken bond. The walls form closed loops on the dual lattice, because every plaquette has an even number of broken bonds around it. Each broken bond changes $K s_i s_j$ from K to $-K$, a factor e^{-2K}. The configuration and its global flip have the same walls, so

$$Z = 2\, e^{2NK} \sum_{G^* \text{ closed on dual lattice}} \left(e^{-2K}\right)^{|G^*|}. \tag{11.165}$$

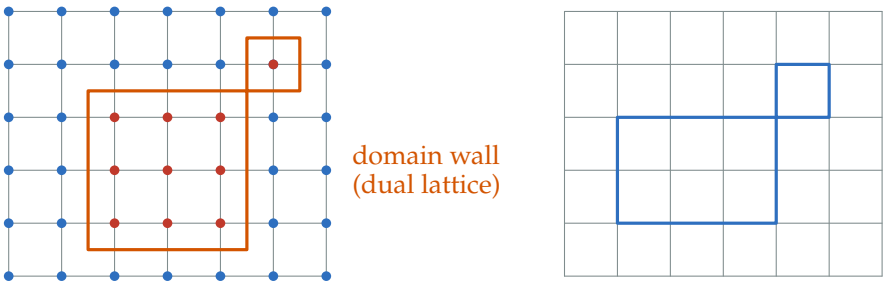


Figure 11.30: The two loop expansions of the Ising partition function. Left: at low temperature, a configuration is a sea of up spins (blue) with islands of down spins (red); the islands are bounded by closed domain walls on the dual lattice, each unit of wall costing e^{-2K}. Right: at high temperature the expansion (11.164) is a sum over closed graphs on the original lattice, each bond costing $\tanh K$. What to see: the two sums are the *same* sum over closed graphs, so the model at coupling K and at the dual coupling K^* are the same model.

The square lattice is its own dual. So (11.164) and (11.165) are the same loop sum if the weights per unit length agree. Define the dual coupling K^* by

$$e^{-2K^*} = \tanh K. \tag{11.166}$$

Then the high-temperature model at K and the low-temperature model at K^* have proportional partition functions (up to analytic prefactors and boundary effects from non-contractible loops on the torus). Small K maps to large K^*: the map exchanges the ordered and disordered phases. If there is a single critical point, it must be at the fixed point $K^* = K$.

Worked example 11.57: The critical coupling of the square-lattice Ising model

Solve $e^{-2K_c} = \tanh K_c$:

$$\begin{aligned} e^{-2K_c} = \frac{e^{K_c} - e^{-K_c}}{e^{K_c} + e^{-K_c}} &\overset{(1)}{=} e^{-K_c} + e^{-3K_c} = e^{K_c} - e^{-K_c} \\ &\overset{(2)}{=} 1 + e^{-2K_c} = e^{2K_c} - 1 \\ &\overset{(3)}{=} \sinh 2K_c = 1. \end{aligned}$$

(1) Cross-multiply by $e^{K_c} + e^{-K_c}$.
(2) Multiply by e^{K_c}.
(3) Rearrange to $e^{2K_c} - e^{-2K_c} = 2$.

So $K_c = \frac{1}{2}\ln(1+\sqrt{2}) \approx 0.4407$, that is $k_B T_c/J = 2/\ln(1+\sqrt{2}) \approx 2.269$. The general duality relation can be written symmetrically as $\sinh 2K \sinh 2K^* = 1$: from (11.166), $\sinh 2K^* = 2\tanh K^*/(1-\tanh^2 K^*)$ with $\tanh K^* = (1-e^{-2K^*})/(1+e^{-2K^*})$, and a two-line computation gives $1/\sinh 2K$.

11.5.3 Critical exponents

Near T_c thermodynamic quantities have power-law singularities. Let

$$t = \frac{T - T_c}{T_c}$$

be the reduced temperature. The two-point function $G(r) = \langle s(x)s(0)\rangle - \langle s\rangle^2$ at separation $r = |x|$ decays exponentially, $G \sim e^{-r/\xi}$, where ξ is the **correlation length**. As $t \to 0$ the

correlation length diverges. Six exponents summarise the singular behaviour [QFT notes §16.2.1, Table 16.2.1]:

exponent	definition	quantity	2D Ising (exact)
ν	$\xi \sim \lvert t\rvert^{-\nu}$	correlation length	1
η	$G(r) \sim r^{-(d-2+\eta)}$ at $t=0$	two-point function	$\frac{1}{4}$
α	$C \sim \lvert t\rvert^{-\alpha}$	specific heat	0 (log)
β	$M \sim (-t)^{\beta}$, $B=0$	spontaneous magnetisation	$\frac{1}{8}$
γ	$\chi \sim \lvert t\rvert^{-\gamma}$	susceptibility	$\frac{7}{4}$
δ	$M \sim B^{1/\delta}$ at $t=0$	critical isotherm	15

Here d is the space dimension ($d=2$ for us), C the specific heat and $\chi = \partial M/\partial B$ the susceptibility. The last column comes from Onsager's exact solution (1944), Yang's magnetisation (1952) and later work. The exponents are **universal**: any two-dimensional system with a two-fold degenerate ordered phase and short-range interactions (triangular lattice, anisotropic couplings, lattice gases, some adsorbed monolayers) has the same six numbers, although T_c changes. Universality is the first hint that the exponents are properties of a continuum field theory, not of the lattice.

Worked example 11.58: Reading off $\beta = \frac{1}{8}$ from Yang's formula

For $T < T_c$, Yang showed $M = \left(1 - \sinh^{-4} 2K\right)^{1/8}$. Near K_c, write $K = K_c + \delta K$ with $\delta K > 0$ small:

$$\sinh 2K \overset{(1)}{=} \sinh 2K_c + 2\cosh(2K_c)\,\delta K + O(\delta K^2) \overset{(2)}{=} 1 + 2\sqrt{2}\,\delta K + O(\delta K^2),$$

$$1 - \sinh^{-4} 2K \overset{(3)}{=} 1 - (1 - 8\sqrt{2}\,\delta K) + O(\delta K^2) = 8\sqrt{2}\,\delta K + O(\delta K^2),$$

$$M \overset{(4)}{=} (8\sqrt{2})^{1/8}\,(\delta K)^{1/8}\big(1 + O(\delta K)\big) \;\propto\; (-t)^{1/8}.$$

(1) Taylor expansion of $\sinh 2K$ about K_c.
(2) $\sinh 2K_c = 1$, so $\cosh 2K_c = \sqrt{1+1} = \sqrt{2}$.
(3) $(1+x)^{-4} \approx 1 - 4x$ with $x = 2\sqrt{2}\,\delta K$.
(4) $K = J/k_B T$, so $\delta K \approx -K_c\, t$ is proportional to $-t$.

The magnetisation vanishes with the strikingly small exponent $\frac{1}{8}$: it drops almost vertically at T_c (Figure 11.31). Mean-field theory would give $\frac{1}{2}$.

Worked example 11.59: $\nu = 1$ from Kramers–Wannier duality

For $T > T_c$ Onsager's solution gives the correlation length along a lattice axis as $\xi^{-1} = 2(K^* - K)$, with K^* the dual coupling (11.166). Near K_c the dual map is $K^* \approx K_c - (K - K_c)$ (it is an involution with fixed point K_c, so its derivative there is -1). Hence

$$\xi^{-1} \overset{(1)}{=} 2\big(K_c - (K - K_c) - K\big) = 4(K_c - K)$$
$$\overset{(2)}{=} 4K_c\, t + O(t^2),$$

(1) Insert the linearised dual map.
(2) $K = K_c/(1+t) \approx K_c(1-t)$.

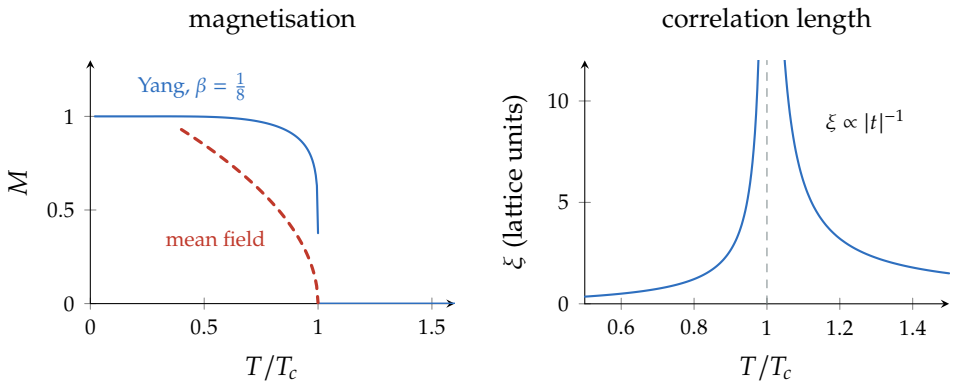


Figure 11.31: Left: the exact spontaneous magnetisation of the square-lattice Ising model (blue) against the mean-field square-root law (red, schematic). The exponent $\frac{1}{8}$ makes M fall off a cliff at T_c. Right: the correlation length $\xi^{-1} \propto |K - K^*|$ diverges linearly on both sides ($\nu = 1$). What to see: at $T = T_c$ there is no length scale left in the problem, which is the precondition for scale and conformal invariance.

so $\xi \propto t^{-1}$ and $\nu = 1$. The derivative of the involution is -1 because differentiating $(K^*)^* = K$ at the fixed point gives $(\mathrm{d}K^*/\mathrm{d}K)^2 = 1$, and the map is decreasing.

Scaling and the renormalisation group

Why power laws? The modern answer is the renormalisation group (RG). Coarse grain the lattice by a factor b (replace blocks of $b \times b$ spins by one effective spin). The couplings change, $(t, B) \mapsto (t', B')$. At the critical point $(0, 0)$ the model looks the same at every scale: it is a **fixed point** of the RG. Near the fixed point the flow is linear,

$$t' = b^{\,y_t}\, t, \qquad B' = b^{\,y_h}\, B, \tag{11.167}$$

with **RG eigenvalues** $y_t, y_h > 0$ (both perturbations grow under coarse graining: they are **relevant**). Lengths shrink by b, so $\xi(t', B') = \xi(t, B)/b$. The free energy per unit area, f, has dimension (length)$^{-d}$, so its singular part obeys

$$f(t, B) = b^{-d}\, f\left(b^{y_t} t,\; b^{y_h} B\right). \tag{11.168}$$

Every exponent follows from (11.168) by choosing b cleverly. We do that below, after seeing where y_t and y_h come from in a field theory.

11.5.4 Why the critical point is conformally invariant

Nash asserts that "at a second order phase transition the conformal invariance of the theory means…". Let us see where the conformal invariance comes from. At T_c the correlation length is infinite, so the continuum limit is a Euclidean field theory with no length scale. It is therefore invariant under translations and rotations (inherited from the lattice at long distance) and under dilations $x \mapsto \lambda x$ (no scale). The claim is that it is then invariant under the larger group of local conformal maps.

The mechanism is the energy–momentum tensor. For a local field theory the response of the action to any infinitesimal change of coordinates $x^\mu \mapsto x^\mu + \varepsilon^\mu(x)$ is

$$\delta S = -\frac{1}{2\pi} \int \mathrm{d}^2x\; T^{\mu\nu}\, \partial_\mu \varepsilon_\nu \tag{11.169}$$

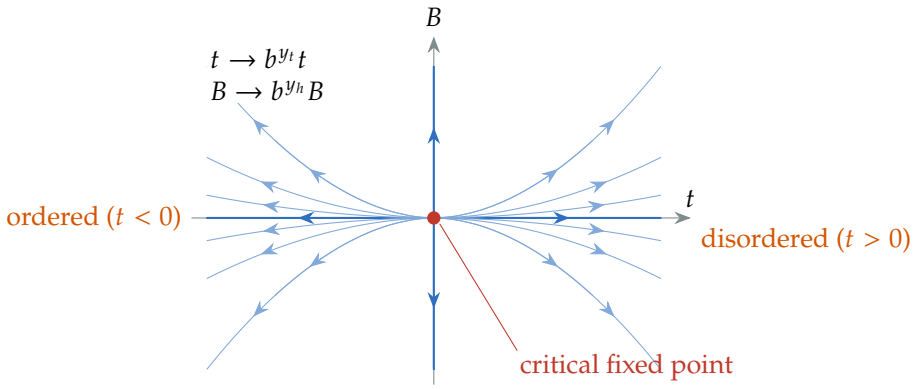


Figure 11.32: Renormalisation-group flow near the Ising critical point. Coarse graining pushes every point away from the origin along the relevant directions t (temperature) and B (magnetic field); since $y_h > y_t$ the B direction grows faster. What to see: the critical theory is the fixed point itself, and the two RG eigenvalues y_t, y_h encode all six critical exponents.

(the normalisation $1/2\pi$ is the one standard in two-dimensional CFT; Nash's (11.16) is the same statement for correlation functions). Rotation invariance makes $T^{\mu\nu}$ symmetric. Scale invariance forces $T^{\mu}{}_{\mu} = 0$: for $\varepsilon^\mu = \lambda x^\mu$ we have $\partial_\mu \varepsilon_\nu = \lambda \delta_{\mu\nu}$ and $\delta S = -(\lambda/2\pi)\int T^{\mu}{}_{\mu}$, which must vanish for every region, so (for a theory without scale) the trace vanishes as an operator, up to contact terms. This is Nash's (11.11), $T_{11} + T_{22} = 0$.

Now take a general ε and split $\partial_\mu \varepsilon_\nu$ into its symmetric traceless part, its trace, and its antisymmetric part:

$$\begin{aligned} \delta S &\overset{(1)}{=} -\frac{1}{2\pi}\int T^{\mu\nu}\,\tfrac{1}{2}\big(\partial_\mu\varepsilon_\nu + \partial_\nu\varepsilon_\mu\big) \\ &\overset{(2)}{=} -\frac{1}{2\pi}\int T^{\mu\nu}\,\tfrac{1}{2}\Big(\partial_\mu\varepsilon_\nu + \partial_\nu\varepsilon_\mu - \delta_{\mu\nu}\,\partial\cdot\varepsilon\Big) \;-\; \frac{1}{4\pi}\int T^{\mu}{}_{\mu}\,\partial\cdot\varepsilon \\ &\overset{(3)}{=} -\frac{1}{2\pi}\int T^{\mu\nu}\,\tfrac{1}{2}\Big(\partial_\mu\varepsilon_\nu + \partial_\nu\varepsilon_\mu - \delta_{\mu\nu}\,\partial\cdot\varepsilon\Big). \end{aligned}$$

(1) $T^{\mu\nu}$ is symmetric, so only the symmetric part of $\partial_\mu\varepsilon_\nu$ survives.
(2) Add and subtract the trace part $\frac{1}{2}\delta_{\mu\nu}\partial\cdot\varepsilon$.
(3) $T^{\mu}{}_{\mu} = 0$.

So the action is invariant under every ε that satisfies the **conformal Killing equation**

$$\partial_\mu\varepsilon_\nu + \partial_\nu\varepsilon_\mu = \delta_{\mu\nu}\,\partial\cdot\varepsilon. \tag{11.170}$$

In two dimensions, with $\varepsilon = \varepsilon^1 + i\varepsilon^2$ and $z = x^1 + ix^2$, the $(\mu\nu) = (11)$ and (12) components of (11.170) read $\partial_1\varepsilon^1 = \partial_2\varepsilon^2$ and $\partial_1\varepsilon^2 = -\partial_2\varepsilon^1$. These are the Cauchy–Riemann equations: $\varepsilon(z)$ is *any* holomorphic function. That is the infinite-dimensional conformal symmetry of §1–§3. In $d > 2$ the same equation allows only a finite-dimensional family (translations, rotations, dilations and special conformal maps).

Watch out: Scale invariance does not always imply conformal invariance

The argument needs $T^{\mu}{}_{\mu} = 0$ *as an operator*. Scale invariance alone only gives $T^{\mu}{}_{\mu} = \partial_\mu K^\mu$ for some local current K^μ (the "virial current"). In two dimensions, for unitary theories with a discrete spectrum, Zamolodchikov and Polchinski showed that K^μ can always be removed by improving $T^{\mu\nu}$, so scale plus rotation invariance gives conformal invariance. Critical lattice models satisfy these hypotheses; this is the quiet assumption behind Nash's

sentence.

11.5.5 Critical exponents are conformal weights

In the continuum theory at T_c the lattice spin becomes a local field $\sigma(x)$ and the local energy density $s_i s_j - \langle s_i s_j \rangle$ becomes a field $\varepsilon(x)$. Both are scaling fields: under $x \mapsto \lambda x$ they pick up factors $\lambda^{-\Delta}$. In a CFT they are primary fields with weights $(h, \bar{h})$, scaling dimension $\Delta = h + \bar{h}$ and spin $h - \bar{h} = 0$ (they are scalars). By [Nash p. 303, (11.9)], their two-point functions at T_c are

$$\langle \phi(z,\bar{z})\,\phi(0,0)\rangle = \frac{C}{z^{2h}\bar{z}^{2\bar{h}}} = \frac{C}{|z|^{2\Delta}} \qquad (h = \bar{h}). \tag{11.171}$$

(Nash's (11.9) writes $z^h \bar{z}^{\bar{h}}$; the exponent $2h$ is forced by the L_0 Ward identity, see Section 11.9.)

Away from T_c the critical theory is perturbed. Changing the temperature couples to the energy density and changing B couples to the spin:

$$S = S_{\mathrm{CFT}} + g_t \int \mathrm{d}^2x\ \varepsilon(x) + g_h \int \mathrm{d}^2x\ \sigma(x), \qquad g_t \propto t,\ \ g_h \propto B. \tag{11.172}$$

Each term in the action must be dimensionless. Under $x \mapsto bx$, the measure d^2x scales as b^d and the field as $b^{-\Delta}$, so the coupling must scale as $b^{\,d-\Delta}$. Comparing with (11.167):

$$y_t = d - \Delta_\varepsilon, \qquad y_h = d - \Delta_\sigma. \tag{11.173}$$

A perturbation is relevant exactly when $\Delta < d$.

Key idea: The dictionary between exponents and dimensions

All six critical exponents follow from the two scaling dimensions $\Delta_\sigma = 2h_\sigma$ and $\Delta_\varepsilon = 2h_\varepsilon$:

$$\nu = \frac{1}{d-\Delta_\varepsilon}, \quad \eta = 2\Delta_\sigma + 2 - d, \quad \alpha = 2 - \frac{d}{d-\Delta_\varepsilon}, \quad \beta = \frac{\Delta_\sigma}{d-\Delta_\varepsilon}, \quad \gamma = \frac{d-2\Delta_\sigma}{d-\Delta_\varepsilon}, \quad \delta = \frac{d-\Delta_\sigma}{\Delta_\sigma}.$$

For the Ising CFT, $h_\sigma = \frac{1}{16}$ and $h_\varepsilon = \frac{1}{2}$ (Section 11.6), so $\Delta_\sigma = \frac{1}{8}$, $\Delta_\varepsilon = 1$, and with $d = 2$ these give exactly the Onsager–Yang values $\nu = 1$, $\eta = \frac{1}{4}$, $\alpha = 0$, $\beta = \frac{1}{8}$, $\gamma = \frac{7}{4}$, $\delta = 15$.

We now derive each formula. Throughout, d is general; set $d = 2$ at the end.

ν. Iterate the RG until t' is of order one: choose $b = |t|^{-1/y_t}$. At that scale the correlation length is some number ξ_0 of order one, so

$$\xi(t) \overset{(1)}{=} b\,\xi(b^{y_t}t) \overset{(2)}{=} |t|^{-1/y_t}\,\xi(\pm 1) \overset{(3)}{=} \xi_0\,|t|^{-1/(d-\Delta_\varepsilon)}.$$

(1) Lengths measured in coarse-grained units are b times smaller.
(2) Choose $b = |t|^{-1/y_t}$, so the rescaled temperature is ± 1.
(3) $y_t = d - \Delta_\varepsilon$ from (11.173).

Hence $\nu = 1/(d - \Delta_\varepsilon)$. For Ising: $\nu = 1/(2-1) = 1$.

η. At T_c, $G(r) = \langle \sigma(r)\sigma(0)\rangle \propto r^{-2\Delta_\sigma}$ by (11.171). The definition $G \sim r^{-(d-2+\eta)}$ gives $d - 2 + \eta = 2\Delta_\sigma$, so $\eta = 2\Delta_\sigma + 2 - d$. For Ising: $\eta = 2 \cdot \frac{1}{8} = \frac{1}{4}$.

α. Put $B = 0$ and choose $b = |t|^{-1/y_t}$ in (11.168):

$$f(t,0) \overset{(1)}{=} |t|^{d/y_t}\,f(\pm 1, 0), \qquad C \propto \frac{\partial^2 f}{\partial t^2} \overset{(2)}{=} |t|^{d/y_t - 2},$$

(1) $b^{-d} = |t|^{d/y_t}$.
(2) The specific heat is the second temperature derivative of the free energy (up to a smooth factor T).

so $\alpha = 2 - d/y_t = 2 - d\nu$ (the **hyperscaling** relation). For Ising $\alpha = 2 - 2 = 0$. An exponent zero here means a logarithm: $f \propto t^2 \ln|t|$, which is Onsager's famous logarithmic specific heat.

β. The magnetisation is $M = -\partial f/\partial B$ at $B = 0$. Differentiating (11.168) with respect to B brings down a factor b^{y_h}:

$$M(t) \overset{(1)}{=} b^{\,y_h - d}\, M(b^{y_t} t, 0) \overset{(2)}{=} |t|^{(d-y_h)/y_t}\, M(-1, 0),$$

(1) Chain rule on $f(b^{y_t}t, b^{y_h}B)$.
(2) Choose $b = |t|^{-1/y_t}$, $t < 0$.

so $\beta = (d - y_h)/y_t = \Delta_\sigma/(d - \Delta_\varepsilon)$. For Ising $\beta = \frac{1}{8}/1 = \frac{1}{8}$, as Yang found.

γ. Differentiate twice with respect to B: $\chi(t) = b^{2y_h - d}\chi(b^{y_t}t, 0)$, so $\gamma = (2y_h - d)/y_t = (d - 2\Delta_\sigma)/(d - \Delta_\varepsilon)$. For Ising $\gamma = (2 - \frac{1}{4})/1 = \frac{7}{4}$.

δ. Now put $t = 0$ and choose $b = B^{-1/y_h}$: $M(0, B) = b^{y_h - d}M(0, 1) = B^{(d-y_h)/y_h}M(0, 1)$, so $1/\delta = (d - y_h)/y_h$ and $\delta = y_h/(d - y_h) = (d - \Delta_\sigma)/\Delta_\sigma$. For Ising $\delta = (2 - \frac{1}{8})/\frac{1}{8} = 15$.

Worked example 11.60: Checking the scaling relations for the Ising model

Because six exponents come from two numbers, there must be four relations among them. With the Ising values:

$$\begin{aligned} \alpha + 2\beta + \gamma &= 0 + \tfrac{1}{4} + \tfrac{7}{4} = 2 && \text{(Rushbrooke)},\\ \gamma &= \beta(\delta - 1) = \tfrac{1}{8}\cdot 14 = \tfrac{7}{4} && \text{(Widom)},\\ \gamma &= \nu(2 - \eta) = 1 \cdot \tfrac{7}{4} = \tfrac{7}{4} && \text{(Fisher)},\\ d\nu &= 2 - \alpha = 2 && \text{(Josephson)}. \end{aligned}$$

All four hold. They hold for *any* values of $\Delta_\sigma, \Delta_\varepsilon$: for instance Fisher's relation is $\nu(2 - \eta) = \frac{1}{d-\Delta_\varepsilon}(d - 2\Delta_\sigma) = \gamma$. For comparison, mean-field theory ($\beta = \frac{1}{2}, \gamma = 1, \delta = 3, \nu = \frac{1}{2}, \eta = 0, \alpha = 0$) violates Josephson's relation in $d = 2$ ($d\nu = 1 \neq 2$): mean-field theory is wrong in two dimensions.

Dictionary: Statistical mechanics ↔ conformal field theory

lattice at $T = T_c$, $B = 0$	CFT (fixed point of the RG)
lattice spin s_i	primary field σ, weights $(\frac{1}{16}, \frac{1}{16})$
energy density $s_i s_j - \langle s_i s_j\rangle$	primary field ε, weights $(\frac{1}{2}, \frac{1}{2})$
constant function 1	identity $\mathbf{1}$, weights $(0, 0)$
temperature deviation t	coupling of $\int \varepsilon$, RG eigenvalue $2 - \Delta_\varepsilon = 1$
magnetic field B	coupling of $\int \sigma$, RG eigenvalue $2 - \Delta_\sigma = \frac{15}{8}$
critical exponents	functions of h_σ, h_ε (key idea above)
universality class	the CFT (its c and its operator content)
periodic $L \times L'$ lattice	torus with modulus τ (Section 11.7)

Reading Nash (p. 311, top of §4)

Nash spends two lines on the physics we have just laid out: "In statistical mechanics the imposition of periodic boundary conditions leads to partition functions Z defined on a

torus; at a second order phase transition the conformal invariance of the theory means that Z should be invariant under the modular transformations". The periodic boundary conditions are Figure 11.29; the second-order transition is the Kramers–Wannier point $\sinh 2K_c = 1$; its conformal invariance is the argument of Section 11.5.4. The link between critical exponents and conformal weights is not stated by Nash at all, but it is the reason the weights $(\frac{1}{16}, \frac{1}{16})$ and $(\frac{1}{2}, \frac{1}{2})$ of his next paragraph are physically meaningful. The preceding paragraph (end of his §3) is our Section 11.5.1.

Recap

- Only L_{-1}, L_0, L_1 generate global maps of the sphere: the Möbius group $\mathrm{SL}(2,\mathbb{C})/\{\pm I\}$. The Schwarzian vanishes on it, so $T(z)$ is a genuine weight-2 tensor under Möbius maps.
- CFTs come from Kac–Moody algebras (Sugawara), superconformal algebras, strings and critical statistical mechanics.
- The 2D Ising model has a continuous transition at $\sinh 2K_c = 1$, located by Kramers–Wannier duality between two loop expansions.
- At T_c, $\xi = \infty$; scale, rotation and locality give $T^{\mu}{}_{\mu} = 0$, hence invariance under all holomorphic $\varepsilon(z)$.
- Perturbing couplings scale with RG eigenvalues $y = d - \Delta$; all six exponents are functions of $\Delta_\sigma, \Delta_\varepsilon$. The Ising values $\Delta_\sigma = \frac{1}{8}, \Delta_\varepsilon = 1$ reproduce Onsager's exact exponents.

11.6 The Ising model as the $c = \frac{1}{2}$ conformal field theory

The last section ended with a promise. If the critical Ising model is a CFT, and if its spin and energy fields have weights $h_\sigma = \frac{1}{16}$ and $h_\varepsilon = \frac{1}{2}$, then all six Onsager exponents follow. Nash states the needed facts in one sentence:

> "if we choose the two dimensional Ising model it turns out that the conformal field theory has $c = 1/2$ and contains only three primary fields whose conformal weights are $(0,0)$, $(1/16, 1/16)$ and $(1/2, 1/2)$."

His footnote adds that this can be checked from the Virasoro representation formula (6.58) with $m = 2$. This section does that check, and more. We want to see *why* $c = \frac{1}{2}$ forces exactly these three weights, what the states in each representation look like, and how the whole theory can be built from one free fermion. The question we answer is:

Given only the Virasoro algebra with $c = \frac{1}{2}$ and unitarity, which weights h are allowed, and what do their representations contain?

11.6.1 From one state to a whole representation: Verma modules

We recall the algebra of §3 and of Chapter 6. The Virasoro algebra is spanned by L_n, $n \in \mathbb{Z}$, and a central element acting as the number c:

$$[L_m, L_n] = (m-n)L_{m+n} + \frac{c}{12}\, m(m^2-1)\,\delta_{m+n,0}. \tag{11.174}$$

L_0 measures energy (in radial quantisation, the dilation generator). Since $[L_0, L_{-n}] = nL_{-n}$, the operator L_{-n} raises the L_0 eigenvalue by n and L_n lowers it by n. So $L_{n>0}$ are *lowering* (annihilation-like) operators and $L_{n<0}$ are *raising* (creation-like) operators, exactly as a and $a^\dagger$ for the harmonic oscillator.

A vector space, or a module?

A physicist starts with a Hilbert space of states and asks how symmetry operators act on it. Here we reverse the logic. We start with a single state $|h\rangle$ and *build* the space by acting on it with the algebra. What comes out is not just a vector space: it is a vector space together with a rule for how every L_n acts on every vector. The algebraists' name for that pair is a **module**.

Working definition 11.22 (module = representation space)

A **module** over the Virasoro algebra is a complex vector space V together with linear maps $L_n : V \to V$ (one for each n) that satisfy the commutation relations (11.174) with c acting as a number [the central element acts as c times the identity on all of V]. "Module" and "representation space" mean the same thing. *How to specify one:* give a basis and the matrix of every L_n in it. A **submodule** is a subspace $N \subseteq V$ that every L_n maps into itself. A module is **irreducible** if its only submodules are $\{0\}$ and V [starting from any nonzero vector, repeated action of the L_n and linear combinations reach all of V].

Notice what this definition does *not* contain: an inner product. The Hilbert-space structure is extra data, and it enters at a separate step below (the Shapovalov form). Keeping the two apart is what makes null states understandable. The whole construction is a three-stage pipeline (Figure 11.33).

Key idea: The highest-weight pipeline

$$|h\rangle \xrightarrow{\text{descendants}} V(c,h) = U(\text{vir}_-)|h\rangle \xrightarrow{\text{quotient by } N} L(c,h) = V(c,h)/N.$$

$V(c,h)$ is the **Verma module**: everything the algebra can produce from $|h\rangle$, with no relations beyond the highest-weight conditions. $L(c,h)$ is the **irreducible** module: the actual space of states of a conformal family in a CFT. We quotient by the *entire submodule* generated by the singular vectors, never by a single vector.

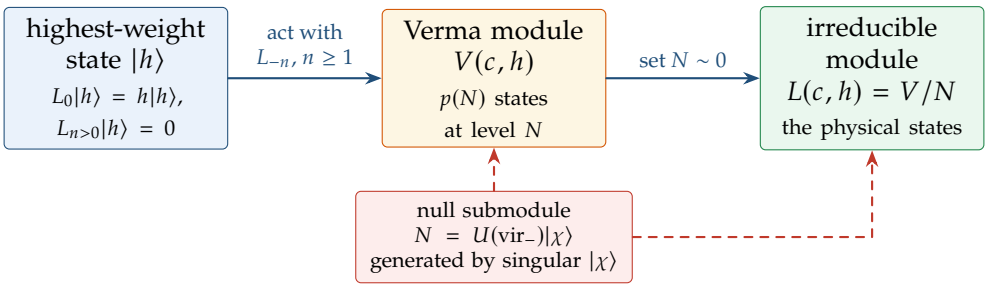


Figure 11.33: The highest-weight pipeline. One state generates the Verma module; the singular vectors inside it generate a submodule N (red), and dividing it out gives the irreducible module. What to see: N is removed as a whole; for generic (c,h) it is $\{0\}$ and the second arrow does nothing.

Step 1: the highest-weight state and its descendants

A **highest-weight state** $|h\rangle$ obeys

$$L_0|h\rangle = h|h\rangle, \qquad L_n|h\rangle = 0 \quad (n \geq 1) \tag{11.175}$$

[Nash p. 186, (6.57)]. It is the lowest-energy state of its family: the lowering operators have nowhere to go. In a CFT it is the state created by a primary field at the origin, $|h\rangle = \phi(0)|0\rangle$. The raising operators produce the **descendants** $L_{-n_1}L_{-n_2}\cdots L_{-n_k}|h\rangle$, with $n_1 \geq n_2 \geq \cdots \geq n_k \geq 1$. Such a state has L_0 eigenvalue $h + N$, where $N = n_1 + \cdots + n_k$ is its **level**. We may order the L_{-n} decreasingly because any other order can be rewritten with the commutators (11.174). The number of such monomials at level N is $p(N)$, the number of partitions of N:

level N	basis of $V(c,h)$ at level N	$p(N)$
0	$\|h\rangle$	1
1	$L_{-1}\|h\rangle$	1
2	$L_{-2}\|h\rangle,\ L_{-1}^2\|h\rangle$	2
3	$L_{-3}\|h\rangle,\ L_{-2}L_{-1}\|h\rangle,\ L_{-1}^3\|h\rangle$	3
4	$L_{-4},\ L_{-3}L_{-1},\ L_{-2}^2,\ L_{-2}L_{-1}^2,\ L_{-1}^4$ (on $\|h\rangle$)	5

In the Verma module these monomials are declared linearly independent. The span is

$$V(c,h) = U(\mathrm{vir}_-)|h\rangle = \mathrm{span}\{L_{-n_1}\cdots L_{-n_k}|h\rangle\}, \tag{11.176}$$

where $U(\mathrm{vir}_-)$, the **universal enveloping algebra** of the raising operators, simply means "all polynomials in the L_{-n}, $n \geq 1$, with products reordered by the commutators".

Worked example 11.61: The Verma module as explicit matrices, levels 0–2

A module is specified by how the operators act. On the basis above we compute the action of L_{-1}, L_1, L_{-2}, L_2 between levels 0, 1, 2. The raising operators are trivial: $L_{-1}|h\rangle$ is basis vector 1 of level 1, $L_{-1}(L_{-1}|h\rangle) = L_{-1}^2|h\rangle$ is basis vector 2 of level 2, and $L_{-2}|h\rangle$ is basis vector 1 of level 2. The lowering operators need the algebra. The three building blocks are

$$L_1L_{-1}|h\rangle \overset{(1)}{=} [L_1, L_{-1}]|h\rangle = 2L_0|h\rangle = 2h|h\rangle,$$
$$L_2L_{-2}|h\rangle \overset{(2)}{=} \big(4L_0 + \tfrac{c}{12}\cdot 2\cdot 3\big)|h\rangle = \big(4h + \tfrac{c}{2}\big)|h\rangle,$$
$$L_1L_{-2}|h\rangle \overset{(3)}{=} 3L_{-1}|h\rangle,$$
$$L_1L_{-1}^2|h\rangle \overset{(4)}{=} \big(2L_0 + L_{-1}L_1\big)L_{-1}|h\rangle \overset{(5)}{=} \big(2(h+1) + 2h\big)L_{-1}|h\rangle = (4h+2)L_{-1}|h\rangle,$$
$$L_2L_{-1}^2|h\rangle \overset{(6)}{=} \big([L_2, L_{-1}]L_{-1} + L_{-1}L_2L_{-1}\big)|h\rangle \overset{(7)}{=} 3L_1L_{-1}|h\rangle = 6h|h\rangle.$$

(1) $L_1|h\rangle = 0$; (11.174) with $m = 1, n = -1$ gives $[L_1, L_{-1}] = 2L_0$ (the central term has $m(m^2-1) = 0$).
(2) $[L_2, L_{-2}] = 4L_0 + \frac{c}{12}2(4-1)$, and $L_2|h\rangle = 0$.
(3) $[L_1, L_{-2}] = 3L_{-1}$, and $L_1|h\rangle = 0$.
(4) $L_1L_{-1} = [L_1, L_{-1}] + L_{-1}L_1 = 2L_0 + L_{-1}L_1$.
(5) $L_{-1}|h\rangle$ has $L_0 = h + 1$, and $L_1L_{-1}|h\rangle = 2h|h\rangle$ by the first line.
(6) Commute L_2 once past L_{-1}.
(7) $[L_2, L_{-1}] = 3L_1$; the second term dies because $L_2L_{-1}|h\rangle = 3L_1|h\rangle + L_{-1}L_2|h\rangle = 0$. Then the first line.

So, as matrices between the levels (columns labelled by the basis of the source level, rows by the basis of the target level),

$$L_1\big|_{1\to 0} = \begin{pmatrix} 2h \end{pmatrix}, \qquad L_1\big|_{2\to 1} = \begin{pmatrix} 3 & 4h+2 \end{pmatrix}, \qquad L_2\big|_{2\to 0} = \begin{pmatrix} 4h+\frac{c}{2} & 6h \end{pmatrix}, \qquad L_{-1}\big|_{1\to 2} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{11.177}$$

This is the Verma module "as a representation": a graded vector space with explicit matrices. Every higher level works the same way, with bigger matrices.

Step 2: where the inner product comes in (the Shapovalov form)

So far there is no notion of length. A quantum theory needs one. Physical input fixes it: $T(z)$ is Hermitian in radial quantisation, which gives $L_n^\dagger = L_{-n}$; and we normalise $\langle h|h\rangle = 1$. These two rules determine the inner product of any two descendants. To compute $\langle h|L_{n_k}\cdots L_{n_1}L_{-m_1}\cdots L_{-m_j}|h\rangle$, move the lowering operators to the right with (11.174) until they hit $|h\rangle$ and die. The resulting Hermitian form on $V(c,h)$ is the **Shapovalov form** (in physics usually just "the Kac form"). Different levels are orthogonal, because L_0 is Hermitian and they have different L_0 eigenvalues. Within level N the form is the $p(N)\times p(N)$ **Gram matrix** $G^{(N)}$.

Working definition 11.23 (Gram matrix and unitarity test)

List the $p(N)$ basis descendants [the vectors $L_{-n_1}\cdots L_{-n_k}|h\rangle$ with $n_1 \geq \cdots \geq n_k \geq 1$ and $n_1+\cdots+n_k = N$] at level N as $|v_1\rangle,\ldots,|v_{p(N)}\rangle$ and compute $G^{(N)}_{ab} = \langle v_a|v_b\rangle$ as just described. *Test:* the representation can be **unitary** (live in a Hilbert space) only if every $G^{(N)}$ is positive semi-definite [all eigenvalues ≥ 0, so no combination of the $|v_a\rangle$ has negative norm squared]. A zero eigenvalue signals a **null vector**: a nonzero element of $V(c,h)$ orthogonal to everything. A negative eigenvalue forbids the pair (c,h) outright.

Worked example 11.62: The Gram matrix at levels 1 and 2

Level 1. One state, $L_{-1}|h\rangle$, and $\langle h|L_1L_{-1}|h\rangle = 2h$ from the first building block. So unitarity needs $h \geq 0$, and $h = 0$ gives a null vector $L_{-1}|0\rangle$. For the vacuum this is welcome: L_{-1} generates translations, and the vacuum is translation invariant.

Level 2. Take $|v_1\rangle = L_{-2}|h\rangle$, $|v_2\rangle = L_{-1}^2|h\rangle$.

$$\begin{aligned} G_{11} &= \langle h|L_2L_{-2}|h\rangle = 4h+\tfrac{c}{2}, \\ G_{12} &\overset{(1)}{=} \langle h|L_2L_{-1}^2|h\rangle \overset{(2)}{=} 6h, \\ G_{22} &\overset{(3)}{=} \langle h|L_1L_1L_{-1}^2|h\rangle \overset{(4)}{=} (4h+2)\,\langle h|L_1L_{-1}|h\rangle = (4h+2)\,2h. \end{aligned}$$

(1) $\langle v_1| = (L_{-2}|h\rangle)^\dagger = \langle h|L_2$.
(2) The last building block. ($G_{21} = G_{12}^* = 6h$ as well.)
(3) $(L_{-1}^2|h\rangle)^\dagger = \langle h|L_1^2$.
(4) The fourth building block, then level 1.

Hence

$$G^{(2)} = \begin{pmatrix} 4h+\frac{c}{2} & 6h \\ 6h & 4h(2h+1) \end{pmatrix}, \qquad \det G^{(2)} = 2h\big(16h^2+2(c-5)h+c\big). \tag{11.178}$$

The determinant: $(4h+\frac{c}{2})(8h^2+4h) - 36h^2 = 32h^3+(4c-20)h^2+2ch$, which factorises as shown.

The determinant vanishes at $h = 0$ (the level-1 null vector, now seen again as $L_{-1}^2|0\rangle$) and at the two roots

$$h_\pm(c) = \frac{(5-c) \pm \sqrt{(1-c)(25-c)}}{16}. \tag{11.179}$$

These are the first two nontrivial **Kac curves**, usually written $h_{2,1}$ and $h_{1,2}$ (Figure 11.34).

Worked example 11.63: The Ising weights from one quadratic equation

Put $c = \frac{1}{2}$ in (11.179):

$$h_\pm\left(\tfrac{1}{2}\right) \overset{(1)}{=} \frac{\frac{9}{2} \pm \sqrt{\frac{1}{2}\cdot\frac{49}{2}}}{16} = \frac{\frac{9}{2} \pm \frac{7}{2}}{16} \overset{(2)}{=} \left\{\tfrac{1}{2},\ \tfrac{1}{16}\right\}.$$

(1) $5 - \frac{1}{2} = \frac{9}{2}$, $1 - \frac{1}{2} = \frac{1}{2}$ and $25 - \frac{1}{2} = \frac{49}{2}$, so the square root is $\frac{7}{2}$.
(2) $8/16 = \frac{1}{2}$ and $1/16$.

So at $c = \frac{1}{2}$ the only weights with a null vector at level ≤ 2 are exactly Nash's 0, $\frac{1}{16}$ and $\frac{1}{2}$. These are the dimensions of the identity, the spin field and the energy field.

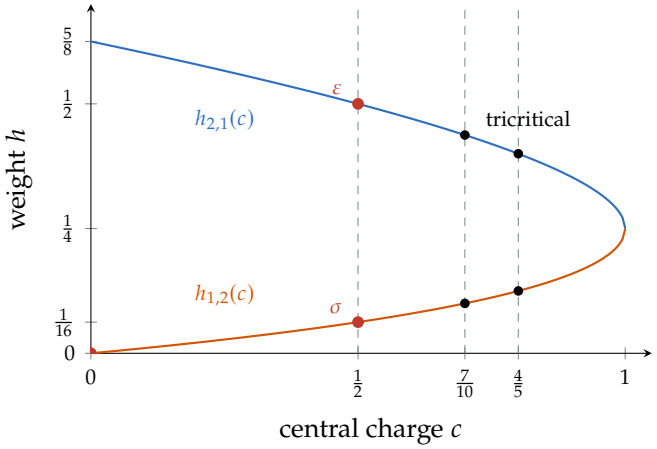


Figure 11.34: The two level-2 Kac curves (11.179). A highest-weight representation with (c, h) on one of the curves has a null vector at level 2. The dashed verticals are the first unitary minimal models $c = \frac{1}{2}$ (Ising, red dots: $h = \frac{1}{2}, \frac{1}{16}$), $c = \frac{7}{10}$ (tricritical Ising) and $c = \frac{4}{5}$ (tetracritical Ising / three-state Potts). What to see: the two curves meet at $c = 1$, $h = \frac{1}{4}$; below $c = 1$ they cut out a discrete list of weights at each allowed c.

Step 3: a singular vector appears

A zero of the Gram determinant means some level-2 combination is special. We now find it directly, without the determinant. Call a descendant $|\chi\rangle$ **singular** if it is itself a highest-weight state: $L_n|\chi\rangle = 0$ for all $n \ge 1$. At level 2 try

$$|\chi\rangle = \left(L_{-2} + \alpha L_{-1}^2\right)|h\rangle. \tag{11.180}$$

It is enough to impose $L_1|\chi\rangle = 0$ and $L_2|\chi\rangle = 0$: every L_n with $n \ge 3$ is a commutator of these two ($L_3 = [L_2, L_1]$, $L_4 = \frac{1}{2}[L_3, L_1]$, and so on), and a commutator of two operators that kill $|\chi\rangle$

also kills it. Using the building blocks,

$$L_1|\chi\rangle \overset{(1)}{=} L_1L_{-2}|h\rangle + \alpha L_1L_{-1}^2|h\rangle \overset{(2)}{=} \big(3 + \alpha(4h+2)\big)L_{-1}|h\rangle,$$
$$L_2|\chi\rangle \overset{(3)}{=} L_2L_{-2}|h\rangle + \alpha L_2L_{-1}^2|h\rangle \overset{(4)}{=} \big(4h + \tfrac{c}{2} + 6h\alpha\big)|h\rangle.$$

(1) Linearity.
(2) Third and fourth building blocks: $3L_{-1}|h\rangle$ and $(4h+2)L_{-1}|h\rangle$.
(3) Linearity.
(4) Second and fifth building blocks: $(4h + \frac{c}{2})|h\rangle$ and $6h|h\rangle$.

Both brackets must vanish. The first fixes α, the second then fixes a relation between h and c:

$$\alpha = -\frac{3}{2(2h+1)},$$
$$c \overset{(1)}{=} -2(4h + 6h\alpha) \overset{(2)}{=} -8h + \frac{18h}{2h+1} = \frac{2h(5-8h)}{2h+1}. \tag{11.181}$$

(1) Solve $4h + \frac{c}{2} + 6h\alpha = 0$ for c.
(2) Insert α: $-12h\alpha = 18h/(2h+1)$; then $-8h(2h+1) + 18h = 2h(5-8h)$.

Solving (11.181) for h gives back $16h^2 + 2(c-5)h + c = 0$, the quadratic factor of $\det G^{(2)}$: the singular vector exists exactly on the Kac curves. At $h = \frac{1}{16}$ we get $\alpha = -\frac{4}{3}$ and $c = \frac{1}{8}\cdot\frac{9}{2}\cdot\frac{8}{9} = \frac{1}{2}$; at $h = \frac{1}{2}$ we get $\alpha = -\frac{3}{4}$ and $c = 1\cdot 1/2 = \frac{1}{2}$. So the Ising model has the two singular vectors

$$|\chi_\sigma\rangle = \Big(L_{-2} - \tfrac{4}{3}L_{-1}^2\Big)|\sigma\rangle, \qquad |\chi_\varepsilon\rangle = \Big(L_{-2} - \tfrac{3}{4}L_{-1}^2\Big)|\varepsilon\rangle. \tag{11.182}$$

In Section 11.9.7 these two equations, set to zero, become differential equations for correlation functions, and they produce the fusion rule $\sigma \times \sigma = \mathbf{1} + \varepsilon$.

Worked example 11.64: A singular vector is a null vector

Why does "singular" (killed by lowering operators) imply "null" (zero norm, orthogonal to everything)? Take any level-2 state $|v\rangle = X|h\rangle$ with X a combination of L_{-2} and L_{-1}^2. Then $X^\dagger$ is a combination of L_2 and L_1^2, and

$$\langle v|\chi\rangle \overset{(1)}{=} \langle h|X^\dagger|\chi\rangle \overset{(2)}{=} \langle h|\big(a\,L_2 + b\,L_1L_1\big)|\chi\rangle \overset{(3)}{=} 0.$$

(1) $\langle v| = (X|h\rangle)^\dagger = \langle h|X^\dagger$.
(2) Write $X^\dagger = aL_2 + bL_1^2$ for some numbers a, b.
(3) $L_2|\chi\rangle = 0$ and $L_1|\chi\rangle = 0$.

States at other levels are orthogonal to $|\chi\rangle$ automatically. So $|\chi\rangle$ is orthogonal to the whole Verma module, itself included: $\langle\chi|\chi\rangle = 0$. As a numerical check at $h = \frac{1}{16}$, $c = \frac{1}{2}$, (11.178) is $G^{(2)} = \left(\begin{smallmatrix} 1/2 & 3/8 \\ 3/8 & 9/32 \end{smallmatrix}\right)$, and with the coefficient vector $v = (1, -\frac{4}{3})$ of $|\chi_\sigma\rangle$,

$$\langle\chi_\sigma|\chi_\sigma\rangle = v^{\mathsf{T}}G^{(2)}v \overset{(1)}{=} \tfrac{1}{2} - 2\cdot\tfrac{4}{3}\cdot\tfrac{3}{8} + \tfrac{16}{9}\cdot\tfrac{9}{32} = \tfrac{1}{2} - 1 + \tfrac{1}{2} = 0.$$

(1) $v^{\mathsf{T}}Gv = G_{11} + 2v_2G_{12} + v_2^2G_{22}$ with $G_{11} = \frac{1}{4} + \frac{1}{4}$, $G_{12} = \frac{6}{16}$, $G_{22} = \frac{4}{16}\cdot\frac{9}{8}$.

The other eigenvalue of $G^{(2)}$ is its trace, $\frac{25}{32} > 0$.

Step 4: one singular vector generates a whole submodule

$|\chi\rangle$ is a highest-weight state of weight $h+2$. So the raising operators act on it just as they act on $|h\rangle$, and generate a copy of a Verma module sitting inside $V(c,h)$:

$$N_\chi = U(\mathrm{vir}_-)|\chi\rangle = \mathrm{span}\{|\chi\rangle,\ L_{-1}|\chi\rangle,\ L_{-2}|\chi\rangle,\ L_{-1}^2|\chi\rangle,\ L_{-3}|\chi\rangle,\ \dots\}. \tag{11.183}$$

It has $p(N-2)$ states at level N: one at levels 2 and 3, two at level 4, three at level 5, and so on. For instance, in the level-3 basis of the table,

$$L_{-1}|\chi\rangle \overset{(1)}{=} L_{-1}L_{-2}|h\rangle + \alpha L_{-1}^3|h\rangle \overset{(2)}{=} L_{-3}|h\rangle + L_{-2}L_{-1}|h\rangle + \alpha L_{-1}^3|h\rangle,$$

(1) Apply L_{-1} to (11.180).
(2) $L_{-1}L_{-2} = L_{-2}L_{-1} + [L_{-1}, L_{-2}]$ and $[L_{-1}, L_{-2}] = (-1+2)L_{-3} = L_{-3}$.

a definite nonzero vector with coordinates $(1, 1, \alpha)$.

Two facts make N_χ the right thing to remove.

- **It is a submodule.** Raising operators keep us in N_χ by construction. A lowering operator L_n acting on $L_{-m_1}\cdots L_{-m_k}|\chi\rangle$ can be commuted to the right; every commutator produces another L, and when the lowering operators reach $|\chi\rangle$ they kill it. What survives is again of the form (raising operators)$|\chi\rangle$.
- **It is entirely null.** Any element of N_χ is $Y|\chi\rangle$ with $Y \in U(\mathrm{vir}_-)$. For any $|v\rangle \in V(c,h)$, $\langle v|Y|\chi\rangle = \langle Y^\dagger v|\chi\rangle$, and $Y^\dagger v$ is some vector of $V(c,h)$ (at level 2, if the level is to match). We showed $|\chi\rangle$ is orthogonal to all of those, so $\langle v|Y\chi\rangle = 0$.

Watch out: Why not throw away only $|\chi\rangle$?

Suppose we divided by the one-dimensional subspace $\mathbb{C}|\chi\rangle$ only. In the quotient $|\chi\rangle \sim 0$, so linearity demands $L_{-1}|\chi\rangle \sim L_{-1}\,0 = 0$. But $L_{-1}|\chi\rangle$ is a nonzero level-3 vector not in $\mathbb{C}|\chi\rangle$, so in that quotient it is *not* zero. The action of L_{-1} would give two different answers on the same equivalence class: it is not well defined. The subspace we divide by must be closed under the algebra. The smallest such subspace containing $|\chi\rangle$ is exactly N_χ.

Step 5: the quotient, and why it is still a module

Let N be the maximal proper submodule of $V(c,h)$: the sum of the submodules generated by all its singular vectors (there can be more than one, as the Ising σ example below shows). Define

$$L(c,h) = V(c,h)/N, \qquad v \sim v' \iff v - v' \in N. \tag{11.184}$$

An element of $L(c,h)$ is an equivalence class $[v] = v + N$. In particular $|\chi\rangle \sim 0$, $L_{-1}|\chi\rangle \sim 0$, $L_{-2}|\chi\rangle \sim 0$, and so on: the whole null tower is set to zero. Define the action on classes by $L_n[v] := [L_n v]$. This is well defined, because if $v \sim v'$ then

$$L_n v - L_n v' \overset{(1)}{=} L_n(v - v') \overset{(2)}{\in} L_n N \overset{(3)}{\subseteq} N \quad\Longrightarrow\quad [L_n v] = [L_n v'].$$

(1) L_n is linear.
(2) $v - v' \in N$ by the definition of $\sim$.
(3) N is a submodule: every L_n maps it into itself.

The commutation relations survive because they hold on representatives: $[L_m, L_n][v] = [(L_m L_n - L_n L_m)v] = [((m-n)L_{m+n} + \frac{c}{12}m(m^2-1)\delta_{m+n,0})v]$. So $L(c,h)$ is a Virasoro module. The inner product also descends, because N is orthogonal to everything: for $n, n' \in N$,

$$\langle v + n \,|\, v' + n' \rangle \overset{(1)}{=} \langle v|v'\rangle + \langle v|n'\rangle + \langle n|v'\rangle + \langle n|n'\rangle \overset{(2)}{=} \langle v|v'\rangle.$$

(1) Sesquilinearity.
(2) Every element of N is orthogonal to all of $V(c,h)$.

On the quotient the form has no null vectors left. If in addition it has no negative directions, $L(c,h)$ is a genuine (pre-)Hilbert space: this is the physical conformal family $[\phi]$ of Nash's §5.

Physical picture: Why unitarity needs the null states removed

In a Hilbert space the only vector of zero length is the zero vector: $\langle\psi|\psi\rangle = 0 \Rightarrow |\psi\rangle = 0$. A null state in $V(c,h)$ has zero norm and zero overlap with every state, so no measurement can ever detect it: every amplitude involving it vanishes. Keeping it would make the inner product degenerate; removing it costs nothing physical. The quotient is the precise way of saying "these states are zero". For the correlation functions this becomes a strong statement: inserting the field of a null state gives zero, which yields differential equations (Section 11.9.7).

Key idea: Generic weights have no null states

If (c,h) lies on none of the Kac curves $h = h_{r,s}(c)$, every Gram determinant is nonzero, $N = \{0\}$, and $V(c,h) = L(c,h)$ is already irreducible. Null states are the exception, and the exceptions are exactly where the interesting (minimal, rational) CFTs live.

	what it is	dimension at level N
Verma module $V(c,h)$	all formally generated descendants $L_{-n_1}\cdots L_{-n_k}\|h\rangle$	$p(N)$
null submodule N	descendants of the singular vectors; zero norm, orthogonal to everything	$p(N-2)$ for one level-2 singular vector (more if there are more)
irreducible $L(c,h)$	V/N: the physical states of a conformal family	$p(N) - \dim N_N$; equals $p(N)$ for generic (c,h)

Worked example 11.65: The weight diagram of the Ising spin module

Take $c = \frac{1}{2}$, $h = \frac{1}{16}$. In the Kac table of Figure 11.36, σ has two labels, $\phi_{1,2}$ and $\phi_{2,2}$, and the null vector of $\phi_{r,s}$ sits at level rs. So there are two singular vectors, at levels 2 and 4. Counting states level by level:

level N	0	1	2	3	4
Verma V: $p(N)$	1	1	2	3	5
generated by $\|\chi_\sigma\rangle$ (level 2): $p(N-2)$	0	0	1	1	2
generated by the level-4 singular vector: $p(N-4)$	0	0	0	0	1
irreducible L	1	1	1	2	2

The last row is confirmed independently by the free-fermion construction of Section 11.6.3, which gives $1, 1, 1, 2, 2$ states for the Ramond ground state tower. (The two null submodules first overlap at higher level, which the inclusion–exclusion character formula of Section 11.7 accounts for.) Figure 11.35 draws the same count.

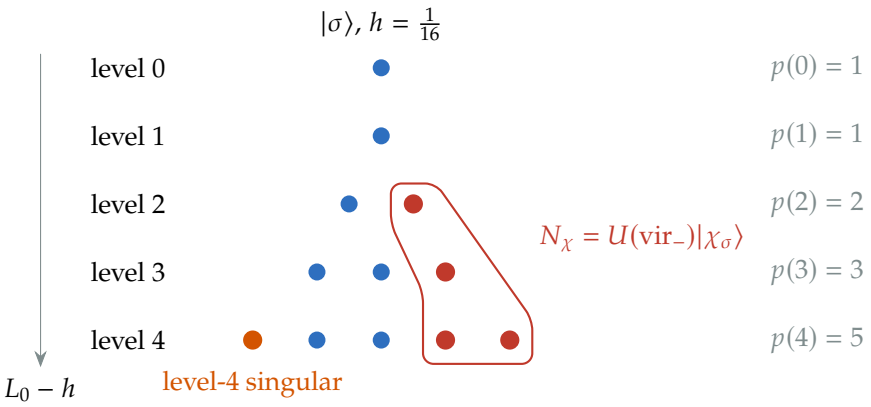


Figure 11.35: The weight diagram of the Verma module $V(\frac{1}{2}, \frac{1}{16})$ by level: $1, 1, 2, 3, 5$ states at levels 0–4 (dots). The red dots and the red outline are the submodule generated by the level-2 singular vector $|\chi_\sigma\rangle$ ($1, 1, 2$ states at levels $2, 3, 4$); the orange dot is a second singular vector at level 4. What to see: removing all coloured dots leaves $1, 1, 1, 2, 2$ blue dots, the states of the irreducible module $L(\frac{1}{2}, \frac{1}{16})$, i.e. of the physical spin field's family.

11.6.2 The Kac formula and the unitary series

Level 2 was a lucky shortcut. The general answer is Kac's determinant formula: $\det G^{(N)}$ is, up to a positive constant, a product of factors $(h - h_{r,s}(c))^{p(N-rs)}$ over positive integers r, s with $rs \le N$. For $c < 1$ Friedan, Qiu and Shenker showed that positivity at *all* levels leaves only a discrete set of (c, h). This is Nash's (6.58):

Key idea: The unitary discrete series [Nash p. 186, (6.58)]

For $c < 1$ a unitary highest-weight representation exists only if

$$c = 1 - \frac{6}{(m+1)(m+2)}, \qquad h = \frac{\big((m+2)p - (m+1)q\big)^2 - 1}{4(m+1)(m+2)}, \tag{11.185}$$

with $m = 1, 2, 3, \ldots$, $p = 1, \ldots, m$, $q = 1, \ldots, p$. For each m the list of h is finite. The corresponding CFTs are the **unitary minimal models**.

Physical picture: Why a discrete list

Each level adds more Kac curves $h_{r,s}(c)$. Between the curves the Gram determinant changes sign. Below $c = 1$ the curves accumulate, and at a generic point some level already has a negative-norm state. Only at the special points where many curves cross simultaneously do all the determinants stay non-negative. These are the points (11.185). Above $c = 1$ the curves have complex h and no longer cut the positive quadrant, so every $h \ge 0$ is allowed.

Worked example 11.66: Nash's footnote 4: $m = 2$ is the Ising model

For $m = 2$: $c = 1 - \frac{6}{3\cdot 4} = \frac{1}{2}$, and $4(m+1)(m+2) = 48$, so $h = \big((4p-3q)^2 - 1\big)/48$ with $1 \le q \le p \le 2$:

$$\begin{aligned}(p,q) = (1,1): &\quad h \overset{(1)}{=} \tfrac{(4-3)^2-1}{48} = 0,\\ (p,q) = (2,1): &\quad h = \tfrac{(8-3)^2-1}{48} = \tfrac{24}{48} = \tfrac{1}{2},\\ (p,q) = (2,2): &\quad h = \tfrac{(8-6)^2-1}{48} = \tfrac{3}{48} = \tfrac{1}{16}.\end{aligned}$$

(1) Substitute into (11.185); the other two lines are the same substitution.

Exactly three weights: $0, \frac{1}{2}, \frac{1}{16}$. This is Nash's claim.

Worked example 11.67: The next model: $m = 3$, the tricritical Ising model

For $m = 3$: $c = 1 - \frac{6}{20} = \frac{7}{10}$, and $h = \big((5p-4q)^2 - 1\big)/80$:

(p,q)	$(1,1)$	$(2,1)$	$(2,2)$	$(3,1)$	$(3,2)$	$(3,3)$
$(5p-4q)^2-1$	0	35	3	120	48	8
h	0	$\frac{7}{16}$	$\frac{3}{80}$	$\frac{3}{2}$	$\frac{3}{5}$	$\frac{1}{10}$

Six primaries. The level-2 ones, $h_{\pm}(\frac{7}{10}) = \frac{7}{16}, \frac{1}{10}$, lie on the curves of Figure 11.34. This theory describes the Ising model with vacancies at its tricritical point, and it is also the simplest superconformal minimal model. In general the m-th model has $m(m+1)/2$ primaries.

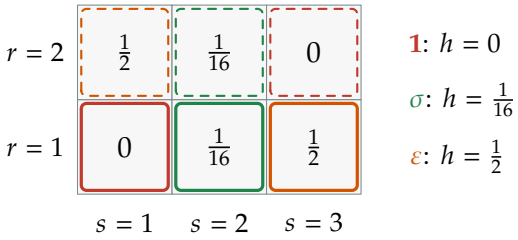


Figure 11.36: The Kac table of the Ising model in the modern labelling $h_{r,s} = \big((4r-3s)^2-1\big)/48$, $1 \le r \le 2$, $1 \le s \le 3$. The table is symmetric under $(r,s) \mapsto (3-r, 4-s)$ (a rotation by 180°), so each weight appears twice; boxes of the same colour are the same field. What to see: six entries, three distinct primaries. Nash's labels (p,q) with $q \le p$ pick one representative of each pair.

Watch out: Two labellings of the Kac table

Nash writes the weights with (p,q), $1 \le q \le p \le m$. Most CFT texts write $h_{r,s} = \frac{(r(m+2)-s(m+1))^2-1}{4(m+1)(m+2)}$ with $1 \le r \le m$, $1 \le s \le m+1$, and identify $h_{r,s} = h_{m+1-r,\,m+2-s}$. Both give the same list. In the modern labels, $\sigma = \phi_{1,2} = \phi_{2,2}$ and $\varepsilon = \phi_{2,1} = \phi_{1,3}$. The null vector of $\phi_{r,s}$ appears at level rs, which is why σ and ε both have level-2 null vectors.

Matching the lattice

Which lattice quantities are these three fields? The lattice model has a global $\mathbb{Z}_2$ symmetry $s_i \mapsto -s_i$. The identity $\mathbf{1}$ and the energy density ε are even; the spin σ is odd. A scalar primary

of weight (h, h) has dimension $\Delta = 2h$, so

$$\Delta_{\mathbf{1}} = 0, \qquad \Delta_\sigma = \tfrac{1}{8}, \qquad \Delta_\varepsilon = 1,$$

which are exactly the values that reproduced Onsager's exponents in Section 11.5.5. The most relevant $\mathbb{Z}_2$-odd operator is the spin, and the most relevant nontrivial $\mathbb{Z}_2$-even operator is the energy. Nothing else is needed: the critical Ising model *is* the $m = 2$ minimal model, with the diagonal combination $(h, \bar{h}) = (h, h)$ of left and right weights.

Watch out: There is a fourth lattice field with $h = \frac{1}{16}$

Kramers–Wannier duality maps σ to the **disorder operator** μ, which creates a domain-wall line ending at a point. It has the same weights $(\frac{1}{16}, \frac{1}{16})$ but is not local with respect to σ: taking one around the other gives a sign. It does not appear as a separate primary in the partition function below, which counts only fields local with respect to the spin. Nash's list of three is the list for the spin (order) description.

11.6.3 The Ising model is a free Majorana fermion

Abstract representation theory tells us which weights occur. A concrete model tells us what the states *are*. Onsager's solution, as reformulated by Kaufman and by Schultz, Mattis and Lieb, shows that the transfer matrix of the Ising model is the exponential of a quadratic form in fermions. At the critical point these fermions become massless. So the critical Ising model is the theory of one free Majorana fermion, the example of [Nash p. 303, (11.5)–(11.6)].

In part 11a we met this fermion. Its holomorphic component $\psi(z)$ has weight $(\frac{1}{2}, 0)$ and two-point function $\langle \psi(z)\psi(w)\rangle = 1/(z-w)$. Its energy–momentum tensor $T = -\frac{1}{2}\psi\partial\psi$ has central charge $c = \frac{1}{2}$. The antiholomorphic partner $\bar{\psi}(\bar{z})$ has weight $(0, \frac{1}{2})$. The product

$$\varepsilon(z, \bar{z}) = i\,\psi(z)\bar{\psi}(\bar{z}) \tag{11.186}$$

has weight $(\frac{1}{2}, \frac{1}{2})$: it is the energy field. (It is the mass term $im\psi\bar{\psi}$ of the fermion action. Moving off T_c gives the fermion a mass proportional to t, which is (11.172) with $g_t \propto t$.)

Modes and the two sectors

On the plane expand

$$\psi(z) = \sum_r \psi_r\, z^{-r-1/2}, \qquad \{\psi_r, \psi_s\} = \delta_{r+s,0}. \tag{11.187}$$

The exponent $-r - \frac{1}{2}$ is the usual $-r - h$ with $h = \frac{1}{2}$, so that ψ_r lowers L_0 by r. There are two consistent choices of the mode labels:

- **Neveu–Schwarz (NS)** sector: $r \in \mathbb{Z} + \frac{1}{2}$. Then $\psi(z)$ is single valued on the plane.
- **Ramond (R)** sector: $r \in \mathbb{Z}$. Then $\psi(z)$ changes sign when z goes once around the origin: $z^{-r-1/2}$ picks up $e^{-2\pi i(r+1/2)} = -1$.

The origin of the plane is the far past of a cylinder. Map $z = e^w$, $w = \tau_E + i\theta$, with $\theta \sim \theta + 2\pi$. A weight-$\frac{1}{2}$ field transforms with $(\mathrm{d}z/\mathrm{d}w)^{1/2} = z^{1/2}$:

$$\psi_{\mathrm{cyl}}(w) \overset{(1)}{=} \left(\frac{\mathrm{d}z}{\mathrm{d}w}\right)^{1/2} \psi(z) \overset{(2)}{=} \sum_r \psi_r\, z^{-r} \overset{(3)}{=} \sum_r \psi_r\, e^{-r\tau_E}\, e^{-ir\theta}.$$

(1) Primary-field transformation law with $h = \frac{1}{2}$.
(2) $z^{1/2} \cdot z^{-r-1/2} = z^{-r}$.
(3) $z = e^{\tau_E + i\theta}$.

So on the cylinder the roles swap: NS fermions ($r \in \mathbb{Z} + \frac{1}{2}$) are *antiperiodic* in θ, and R fermions ($r \in \mathbb{Z}$) are *periodic*. These are the two spin structures on a circle (Chapter T4). Figure 11.37 shows both pictures.

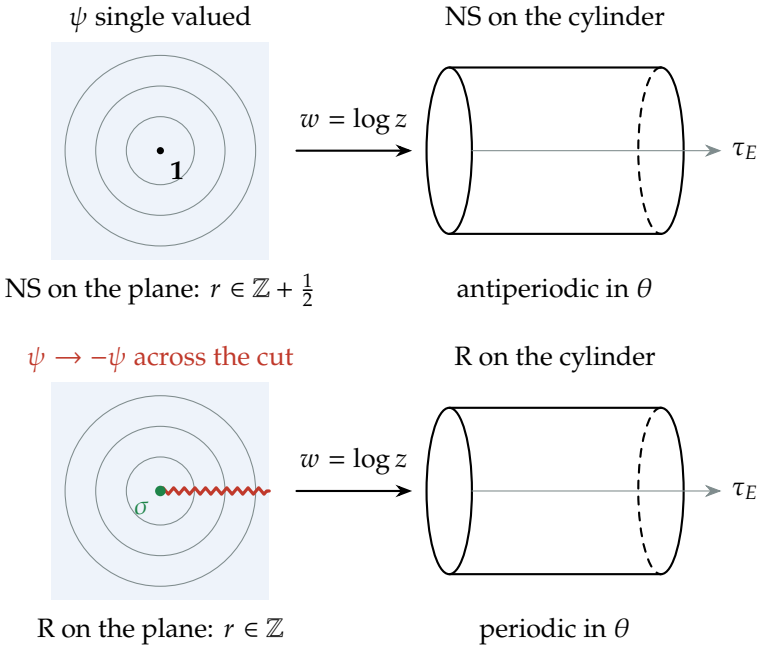


Figure 11.37: The two sectors of the Majorana fermion, one per row. On the plane (left column) the NS fermion is single valued, and the state at the origin is the vacuum; the R fermion changes sign around the origin, so a branch cut (zigzag) must end on an operator there, the spin field σ. Under $z = e^w$ (right column) the plane becomes a cylinder and the boundary conditions around the circle swap: NS is antiperiodic, R is periodic. What to see: inserting σ is the same as changing the spin structure around the insertion point.

The NS sector contains 1 and ε

The NS vacuum $|0\rangle$ obeys $\psi_r|0\rangle = 0$ for $r > 0$. States are created by ψ_{-r}, $r = \frac{1}{2}, \frac{3}{2}, \ldots$. Each can be used at most once, because $\psi_{-r}^2 = \frac{1}{2}\{\psi_{-r}, \psi_{-r}\} = 0$. A state $\psi_{-r_1} \cdots \psi_{-r_k}|0\rangle$ with $r_1 > \cdots > r_k > 0$ has $L_0 = r_1 + \cdots + r_k$. The fermion number $(-1)^F$ splits the NS space into an even part (containing $|0\rangle$) and an odd part (containing $\psi_{-1/2}|0\rangle$, with $L_0 = \frac{1}{2}$). Since T is bilinear in ψ, the Virasoro generators preserve $(-1)^F$, so each part is a Virasoro representation:

L_0	even states (**1** module)	odd states (ε module)
0	$\|0\rangle$	
$\frac{1}{2}$		$\psi_{-1/2}\|0\rangle$
$\frac{3}{2}$		$\psi_{-3/2}\|0\rangle$
2	$\psi_{-3/2}\psi_{-1/2}\|0\rangle$	
$\frac{5}{2}$		$\psi_{-5/2}\|0\rangle$
3	$\psi_{-5/2}\psi_{-1/2}\|0\rangle$	
$\frac{7}{2}$		$\psi_{-7/2}\|0\rangle$
4	$\psi_{-7/2}\psi_{-1/2}\|0\rangle,\ \psi_{-5/2}\psi_{-3/2}\|0\rangle$	
$\frac{9}{2}$		$\psi_{-9/2}\|0\rangle,\ \psi_{-5/2}\psi_{-3/2}\psi_{-1/2}\|0\rangle$

Compare the Virasoro count. In the vacuum module, level 1 is empty ($L_{-1}|0\rangle = 0$), level 2 has only $L_{-2}|0\rangle$, level 3 only $L_{-3}|0\rangle$, level 4 has $L_{-4}|0\rangle$ and $L_{-2}^2|0\rangle$. That is $1, 0, 1, 1, 2$: the even column. In the ε module, levels $0, 1, 2, 3, 4$ above $h = \frac{1}{2}$ contain $1, 1, 1, 1, 2$ states. A generic Verma module would have $1, 1, 2, 3, 5$; the null vector $(L_{-2} - \frac{3}{4}L_{-1}^2)|\varepsilon\rangle$ of (11.182) removes one state at level 2, and a second null vector at level 3 (the $rs = 1 \cdot 3$ one of $\phi_{1,3}$) removes more. The fermion does the bookkeeping for us.

The R sector contains σ

In the Ramond sector there is a zero mode ψ_0 with $\psi_0^2 = \frac{1}{2}$. The ground state $|\sigma\rangle$ obeys $\psi_n|\sigma\rangle = 0$ for $n > 0$. Its weight is the conformal weight of the operator that sits at the end of the branch cut in Figure 11.37. We can compute it exactly.

Worked example 11.68: $h_\sigma = \frac{1}{16}$ from the Ramond propagator

Step 1: the propagator in the R ground state. Only $\langle\psi_n\psi_{-n}\rangle$ with $n \geq 0$ survive, with $\langle\psi_n\psi_{-n}\rangle = \{\psi_n, \psi_{-n}\} = 1$ for $n > 0$ and $\langle\psi_0\psi_0\rangle = \frac{1}{2}$. So, for $|w| < |z|$,

$$\begin{aligned}\langle\sigma|\psi(z)\psi(w)|\sigma\rangle &\overset{(1)}{=} \sum_{n\geq 0}\langle\psi_n\psi_{-n}\rangle\, z^{-n-1/2}w^{n-1/2} \\ &\overset{(2)}{=} (zw)^{-1/2}\Big(\tfrac{1}{2} + \sum_{n\geq 1}\big(\tfrac{w}{z}\big)^n\Big) \overset{(3)}{=} (zw)^{-1/2}\,\frac{z+w}{2(z-w)} \\ &= \frac{\sqrt{z/w} + \sqrt{w/z}}{2\,(z-w)}.\end{aligned}$$

(1) Insert the mode expansion (11.187) with $r \in \mathbb{Z}$; $\psi_n|\sigma\rangle = 0$ for $n > 0$ kills all other terms.
(2) Pull out $(zw)^{-1/2}$; the $n = 0$ term gives $\frac{1}{2}$.
(3) Geometric series: $\frac{1}{2} + \frac{w/z}{1-w/z} = \frac{1}{2} + \frac{w}{z-w} = \frac{z+w}{2(z-w)}$.

(The same computation in the NS sector, $r \in \mathbb{Z} + \frac{1}{2}$, gives exactly $1/(z-w)$.)

Step 2: expand near $z = w$. Put $z = w(1+u)$:

$$\frac{\sqrt{z/w} + \sqrt{w/z}}{2} \overset{(1)}{=} \frac{2+u}{2\sqrt{1+u}} \overset{(2)}{=} \big(1 + \tfrac{u}{2}\big)\big(1 - \tfrac{u}{2} + \tfrac{3u^2}{8} + O(u^3)\big) = 1 + \frac{u^2}{8} + O(u^3).$$

(1) $\sqrt{1+u} + 1/\sqrt{1+u} = (2+u)/\sqrt{1+u}$.

(2) Binomial series $(1+u)^{-1/2} = 1 - \frac{u}{2} + \frac{3}{8}u^2 - \cdots$.

With $z - w = wu$ the propagator is $\frac{1}{z-w} + \frac{z-w}{8w^2} + O\big((z-w)^2\big)$.

Step 3: the expectation value of T. Normal ordering subtracts the singular part $1/(z-w)$, so

$$\langle\sigma|{:}\psi(w)\partial\psi(w){:}|\sigma\rangle \overset{(1)}{=} \lim_{z\to w} \partial_w\Big(\frac{z-w}{8w^2}\Big) = -\frac{1}{8w^2},$$
$$\langle\sigma|T(w)|\sigma\rangle \overset{(2)}{=} -\tfrac{1}{2}\Big(-\frac{1}{8w^2}\Big) = \frac{1}{16\,w^2}.$$

(1) $\langle\psi(z)\partial_w\psi(w)\rangle = \partial_w$ of the propagator; the regular part's derivative at $z = w$ is $-1/(8w^2)$ (the other terms vanish as $z \to w$).

(2) $T = -\frac{1}{2}{:}\psi\partial\psi{:}$.

Step 4: read off h. For any highest-weight state, $\langle h|T(w)|h\rangle = \sum_n w^{-n-2}\langle h|L_n|h\rangle = h/w^2$, because only L_0 has a diagonal matrix element. Comparing: $h_\sigma = \frac{1}{16}$.

Physical picture: Zero-point energies

The same number comes out of the Casimir energy. Each fermion mode of frequency ω contributes $-\omega/2$ to the ground-state energy. On the unit circle, NS modes have $\omega = r \in \mathbb{N} - \frac{1}{2}$ and R modes have $\omega = n \in \mathbb{N}$. With zeta-function regularisation, $\sum_{n\geq 1} n = -\frac{1}{12}$ and $\sum_{r\in\mathbb{N}-1/2} r = \frac{1}{24}$. So $E_{\mathrm{NS}} = -\frac{1}{48}$ and $E_{\mathrm{R}} = +\frac{1}{24}$. The first is $-c/24$ for $c = \frac{1}{2}$ (the vacuum Casimir energy); the difference is $E_{\mathrm{R}} - E_{\mathrm{NS}} = \frac{1}{24} + \frac{1}{48} = \frac{1}{16} = h_\sigma$.

Above $|\sigma\rangle$ the R states are $\psi_{-n_1}\cdots\psi_{-n_k}|\sigma\rangle$ with $n_1 > \cdots > n_k \geq 1$. Their counting at levels $0, 1, 2, 3, 4$ is $1, 1, 1, 2, 2$ (for level 3: ψ_{-3} and $\psi_{-2}\psi_{-1}$). The zero mode ψ_0 maps the R ground state to a state of the same weight; for a single Majorana fermion the two can be identified up to the choice of $(-1)^F$, a subtlety that shows up as the factor $\sqrt{2}$ in (11.204) below.

Dictionary: Ising model ↔ free Majorana fermion

identity $\mathbf{1}$, $h = 0$	NS states with $(-1)^F = +1$
energy ε, $h = \frac{1}{2}$	NS states with $(-1)^F = -1$; $\varepsilon = i\psi\bar{\psi}$
spin σ, $h = \frac{1}{16}$	R ground state; end point of a branch cut
$T \neq T_c$	mass $m \propto t$ for the fermion
$c = \frac{1}{2}$	one real fermion (half a complex one)
periodic/antiperiodic on the torus	the four spin structures (Section 11.7)

Reading Nash (p. 311–312, (11.49)–(11.50) and footnote 4)

Nash asserts $c = \frac{1}{2}$ and the three weights and cites (6.58). We have checked (6.58) at $m = 2$, and derived the same weights independently from the level-2 Gram determinant (11.178) and from the fermion (11.187). Nash's weights are the pairs $(h, \bar{h})$; the Ising model uses only the diagonal pairs (h, h), which will be the content of his formula (11.50).

Recap

- Unitarity is positivity of the Gram matrices. At level 2, $\det G^{(2)} = 2h(16h^2 + 2(c-5)h + c)$.

- At $c = \frac{1}{2}$ its roots are $h = 0, \frac{1}{16}, \frac{1}{2}$; the null vectors are $(L_{-2} - \frac{4}{3}L_{-1}^2)|\sigma\rangle$ and $(L_{-2} - \frac{3}{4}L_{-1}^2)|\varepsilon\rangle$.
- The Kac formula (6.58) at $m = 2$ gives the same list; $m = 3$ is the tricritical Ising model with six primaries.
- The Ising model is a free Majorana fermion. The NS sector splits into the **1** and ε modules by $(-1)^F$; the R sector is the σ module, with $h_\sigma = \frac{1}{16}$ computed from the Ramond propagator.

11.7 Partition functions on the torus

We now have a CFT, the $c = \frac{1}{2}$ theory with three primaries, and a lattice model that it describes. Nash's §4 asks a global question about it. Put the critical Ising model on a periodic lattice. Its partition function is a single number for each lattice shape. What does that number depend on, and what symmetry must it have?

The answer has three parts, and this section derives each one.

1. In the continuum limit, a periodic lattice is a flat torus, and a CFT only sees its *shape*: one complex number τ with $\operatorname{Im}\tau > 0$, the **modular parameter**.
2. Different τ can describe the same torus. They are related by $\tau \mapsto (a\tau + b)/(c\tau + d)$ with integer entries and $ad - bc = 1$: Nash's (11.48). So $Z(\tau)$ must be invariant under this **modular group**.
3. Computing Z in the operator language gives a trace over the Hilbert space, which organises itself into a sum of products of Virasoro characters: Nash's (11.49). For Ising it is (11.50).

11.7.1 A torus is a lattice in the complex plane

Take two complex numbers ω_1, ω_2 that are not real multiples of each other. They generate a lattice

$$\Lambda = \{\, m\omega_1 + n\omega_2 : m, n \in \mathbb{Z} \,\} \subset \mathbb{C}, \tag{11.188}$$

and the quotient $\mathbb{C}/\Lambda$ (identify $z \sim z + \omega_1 \sim z + \omega_2$) is a torus. It inherits the flat metric $|\mathrm{d}z|^2$ and, more importantly for us, the complex coordinate z up to translations: it is a Riemann surface of genus one. The fundamental parallelogram with corners $0, \omega_1, \omega_2, \omega_1 + \omega_2$ is one copy of the torus (Figure 11.38).

The Ising lattice of Figure 11.29 with L columns and L' rows, in units of the lattice spacing, is $\omega_1 = L$, $\omega_2 = iL'$. If we glue the top edge to the bottom edge with a shift by s columns (a "helical" boundary condition), $\omega_2 = s + iL'$.

A CFT cannot tell apart two tori related by a conformal map. The map $z \mapsto z/\omega_1$ is conformal: a rotation and a dilation. It takes Λ to the lattice generated by 1 and

$$\tau = \frac{\omega_2}{\omega_1}. \tag{11.189}$$

We order the generators so that $\operatorname{Im}\tau > 0$ (swap them if necessary). So the conformal class of the torus is fixed by the single number τ in the upper half plane $\mathbb{H}$.

Working definition 11.24 (modular parameter)

For a torus $\mathbb{C}/(\mathbb{Z}\omega_1 + \mathbb{Z}\omega_2)$ [the plane with z, $z+\omega_1$ and $z+\omega_2$ counted as the same point], the **modular parameter** is $\tau = \omega_2/\omega_1$ with the labelling chosen so that $\operatorname{Im}\tau > 0$. *How to read it:* rescale and rotate so that one period is 1; the other period is τ. For an $L \times L'$ rectangular lattice $\tau = iL'/L$; a twist by s sites [top edge glued to bottom edge after a shift by s columns] adds s/L to $\operatorname{Re}\tau$.

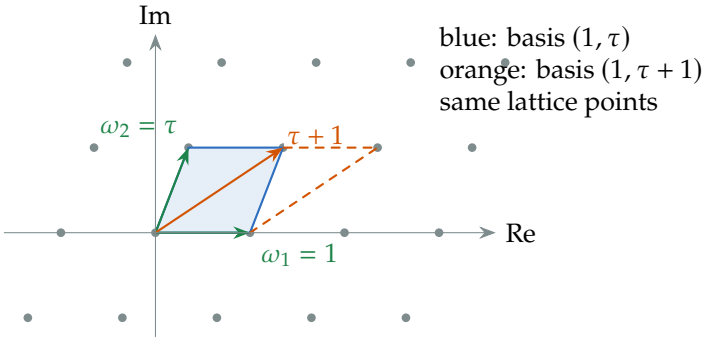


Figure 11.38: A torus is the plane modulo a lattice. The blue parallelogram, spanned by 1 and τ, is one fundamental domain. The orange one, spanned by 1 and $\tau+1$, contains the same area and tiles the plane with the *same* lattice points. What to see: τ and $\tau+1$ describe the same torus; this is the generator T of the modular group.

11.7.2 The modular group

The lattice Λ does not determine its generators. Any other pair

$$\begin{pmatrix}\omega_2'\\ \omega_1'\end{pmatrix} = \begin{pmatrix}a & b\\ c & d\end{pmatrix}\begin{pmatrix}\omega_2\\ \omega_1\end{pmatrix}, \qquad a,b,c,d \in \mathbb{Z}, \tag{11.190}$$

generates a sublattice of Λ. It generates all of Λ exactly when the matrix can be inverted *with integer entries*, which means $ad - bc = \pm 1$ (the inverse matrix has $1/(ad-bc)$ in front). Then the new modular parameter is

$$\tau' \overset{(1)}{=} \frac{\omega_2'}{\omega_1'} \overset{(2)}{=} \frac{a\omega_2 + b\omega_1}{c\omega_2 + d\omega_1} \overset{(3)}{=} \frac{a\tau + b}{c\tau + d}. \tag{11.191}$$

(1) Definition (11.189) in the new basis.
(2) The basis change (11.190).
(3) Divide numerator and denominator by ω_1.

This is Nash's (11.48). The sign of the determinant is fixed by the upper half plane:

$$\operatorname{Im}\frac{a\tau+b}{c\tau+d} \overset{(1)}{=} \operatorname{Im}\frac{(a\tau+b)(c\bar\tau+d)}{|c\tau+d|^2} \overset{(2)}{=} \frac{\operatorname{Im}\left(ad\,\tau + bc\,\bar\tau\right)}{|c\tau+d|^2} \overset{(3)}{=} \frac{(ad-bc)\operatorname{Im}\tau}{|c\tau+d|^2}.$$

(1) Multiply top and bottom by the complex conjugate of the denominator.
(2) The numerator is $ac|\tau|^2 + ad\,\tau + bc\,\bar\tau + bd$; the first and last terms are real.
(3) $\operatorname{Im}\bar\tau = -\operatorname{Im}\tau$.

So $\operatorname{Im}\tau' > 0$ requires $ad - bc = +1$: the basis change must preserve orientation. The matrices form $\mathrm{SL}(2,\mathbb{Z})$. Finally M and $-M$ give the same τ', so the group that acts on $\mathbb{H}$ is

$$\mathrm{PSL}(2,\mathbb{Z}) = \mathrm{SL}(2,\mathbb{Z})/\{\pm I\}, \tag{11.192}$$

which is Nash's $\mathrm{SL}(2,\mathbb{Z})/\{\mp I\}$. Composition of the maps is matrix multiplication, exactly as for Möbius maps (Section 11.5.1): $\mathrm{SL}(2,\mathbb{Z})$ is the subgroup of $\mathrm{SL}(2,\mathbb{C})$ with integer entries.

Working definition 11.25 (the generators S and T)

$$T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} : \ \tau \mapsto \tau + 1, \qquad S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} : \ \tau \mapsto -\frac{1}{\tau}.$$

Every element of $\mathrm{PSL}(2,\mathbb{Z})$ [a change of period basis, acting as $\tau \mapsto \frac{a\tau+b}{c\tau+d}$ with integers a,b,c,d and $ad-bc=1$] is a product of S's and T's. *Consequence (Nash's (11.51)):* to show that a function $Z(\tau)$ is modular invariant [unchanged under every such change of basis, so it depends only on the torus and not on the chosen periods] it suffices to check $Z(\tau+1) = Z(\tau)$ and $Z(-1/\tau) = Z(\tau)$.

Geometrically (Figure 11.39): T keeps the cycle a (along ω_1) and replaces the cycle b by $b + a$. This is a **Dehn twist**: cut the torus along a, rotate one edge by a full turn, and glue back. S exchanges the two cycles (with an orientation flip), which for a rectangular torus $\tau = i\beta$ exchanges space and time: $-1/(i\beta) = i/\beta$.

Worked example 11.69: The relations $S^2 = (ST)^3 = 1$

Multiply the matrices:

$$S^2 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} = -I,$$

$$ST \overset{(1)}{=} \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & 1 \end{pmatrix},$$

$$(ST)^2 = \begin{pmatrix} 0 & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 0 & -1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} -1 & -1 \\ 1 & 0 \end{pmatrix}, \qquad (ST)^3 = \begin{pmatrix} -1 & -1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 0 & -1 \\ 1 & 1 \end{pmatrix} = -I.$$

(1) ST means "first T, then S" as maps: $\tau \mapsto \tau+1 \mapsto -1/(\tau+1)$, which is indeed $(0\cdot\tau - 1)/(\tau+1)$.

In $\mathrm{PSL}(2,\mathbb{Z})$, $-I$ is the identity, so $S^2 = (ST)^3 = 1$. These are the only relations: $\mathrm{PSL}(2,\mathbb{Z})$ is the free product $\mathbb{Z}_2 * \mathbb{Z}_3$. Check on a point: ST sends $\tau = e^{2\pi i/3} \mapsto -1/(e^{2\pi i/3}+1) = -1/e^{i\pi/3} = e^{2\pi i/3}$, a fixed point of order 3; S fixes $\tau = i$.

Worked example 11.70: Writing a modular transformation as a word in S and T

Take $M = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 1 \end{smallmatrix}\right)$, $\tau \mapsto (2\tau+1)/(\tau+1)$. The algorithm is Euclid's: use $T^{\pm 1}$ to reduce $|a|$ against $|c|$, and S to swap them.

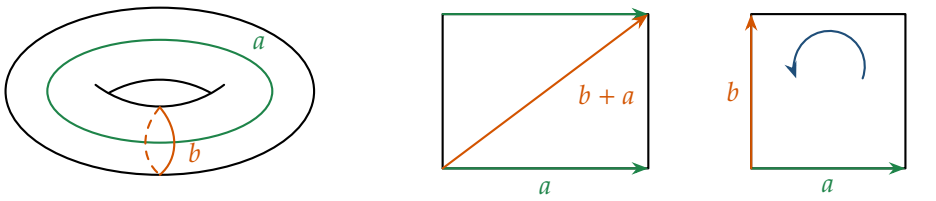
$$T^{-1}M \overset{(1)}{=} \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix},$$

$$S\,T^{-1}M \overset{(2)}{=} \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} -1 & -1 \\ 1 & 0 \end{pmatrix} = (ST)^2.$$

(1) Subtract row 2 from row 1: $a \to a - c$.

(2) Swap the rows (with a sign).

So $M = T\,S^{-1}(ST)^2 = T\,S^{-1}STST = T^2ST$ (using $S^{-1} = S$ in PSL and cancelling). Check on the map: $\tau \overset{T}{\mapsto} \tau+1 \overset{S}{\mapsto} -\frac{1}{\tau+1} \overset{T^2}{\mapsto} 2 - \frac{1}{\tau+1} = \frac{2\tau+1}{\tau+1}$. ✓



cycles a (ω_1) and b (ω_2) T: $b \mapsto b + a$ (Dehn twist) S: $(a, b) \mapsto (b, -a)$

Figure 11.39: The two generators of the modular group acting on the cycles of the torus. T cuts along a and re-glues with a full twist, so the new b cycle is the old $b + a$; S rotates the fundamental square by 90°, exchanging the cycles. What to see: neither changes the torus itself, only our choice of cycles. For a rectangular torus, S exchanges the "space" and "time" directions.

The fundamental domain

Every $\tau \in \mathbb{H}$ can be moved by $\mathrm{PSL}(2,\mathbb{Z})$ into

$$\mathcal{F} = \left\{\tau \in \mathbb{H} : \ |\mathrm{Re}\,\tau| \le \tfrac{1}{2},\ |\tau| \ge 1\right\}, \tag{11.193}$$

and two interior points of $\mathcal{F}$ are never related. The reason is a one-line calculation: $\mathrm{Im}(M\tau) = \mathrm{Im}\,\tau/|c\tau + d|^2$, and only finitely many pairs (c, d) make $|c\tau + d| < 1$. So in any orbit there is a point of maximal $\mathrm{Im}\,\tau$. Translate it by T^n into the strip $|\mathrm{Re}\,\tau| \le \frac{1}{2}$. If it had $|\tau| < 1$, then S would raise its imaginary part to $\mathrm{Im}\,\tau/|\tau|^2$, contradicting maximality. So $\mathcal{F}$ is the **moduli space of tori**: each conformal class of torus appears once. It is the genus-one case of the moduli spaces $\mathcal{M}_p$ that come back in Section 11.11.

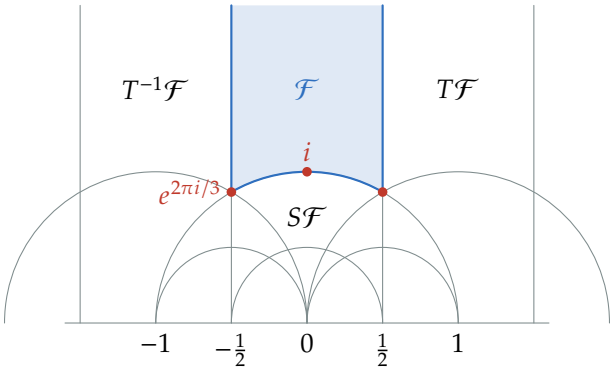


Figure 11.40: The fundamental domain $\mathcal{F}$ of $\mathrm{PSL}(2,\mathbb{Z})$ (blue) and some of its images: $T^{\pm 1}\mathcal{F}$ beside it and $S\mathcal{F}$ inside the unit circle. The vertical edges are glued by T, the arc to itself by S ($\tau \mapsto -\bar{\tau}$ on it). The red points are the orbifold points $\tau = i$ (fixed by S) and $\tau = e^{2\pi i/3}$ (fixed by ST). What to see: $\mathcal{F}$ is the space of shapes of tori; a modular invariant $Z(\tau)$ is really a function on $\mathcal{F}$.

Watch out: The modular group is not a symmetry of the lattice model

For a finite lattice, $\tau = iL'/L$ and $\tau \mapsto -1/\tau$ just means "turn the page by 90°", a trivial symmetry. The nontrivial content of modular invariance appears only when we write Z in the operator language, where the two cycles play very different roles (space and Euclidean time). Then $Z(-1/\tau) = Z(\tau)$ says that the spectrum of the Hamiltonian on a circle of length L determines, and is determined by, the spectrum on a circle of length L'. That is a strong constraint on the spectrum.

11.7.3 The Hamiltonian on a circle

To write Z as a trace we need the Hamiltonian of the CFT on a spatial circle of length L. Take the infinite cylinder with coordinate $w = t + ix$, $x \sim x + L$, $t \in \mathbb{R}$ Euclidean time, and map it to the plane by

$$z = e^{2\pi w/L}. \tag{11.194}$$

The far past $t \to -\infty$ is the origin $z = 0$, where the in-state sits in radial quantisation (§3). Circles of constant t are circles $|z| = e^{2\pi t/L}$. So time translation $t \mapsto t + \delta t$ is the dilation $z \mapsto e^{2\pi\delta t/L} z$, generated by $L_0 + \bar{L}_0$; translation in x is the rotation $z \mapsto e^{2\pi i \delta x/L} z$, generated by $L_0 - \bar{L}_0$.

The constant term comes from the Schwarzian. Apply the transformation law (11.161) with the cylinder coordinate w in the role of z and the plane coordinate z in the role of z'. With $k = 2\pi/L$ we have $z' = kz$, $z'' = k^2 z$, $z''' = k^3 z$ (primes are $\mathrm{d}/\mathrm{d}w$), so

$$\begin{aligned} T_{\mathrm{cyl}}(w) &\overset{(1)}{=} \left(\frac{\mathrm{d}z}{\mathrm{d}w}\right)^2 T(z) + \frac{c}{12}\{z, w\} \overset{(2)}{=} k^2 z^2 T(z) + \frac{c}{12}\left(k^2 - \tfrac{3}{2}k^2\right) \\ &\overset{(3)}{=} \left(\frac{2\pi}{L}\right)^2 \Big(\sum_{n\in\mathbb{Z}} L_n z^{-n} - \frac{c}{24}\Big). \end{aligned} \tag{11.195}$$

(1) Transformation law of T, [Nash p. 307, (11.31)].
(2) $\{z, w\} = z'''/z' - \frac{3}{2}(z''/z')^2 = k^2 - \frac{3}{2}k^2 = -\frac{1}{2}k^2$.
(3) Mode expansion $T(z) = \sum_n L_n z^{-n-2}$, so $z^2 T(z) = \sum_n L_n z^{-n}$.

The Hamiltonian is $H = \frac{1}{2\pi}\int_0^L \mathrm{d}x\, \big(T_{\mathrm{cyl}} + \bar{T}_{\mathrm{cyl}}\big)$ at fixed t (this normalisation is the one for which H generates unit time translations, as the next line confirms), and the momentum is $P = \frac{1}{2\pi}\int_0^L \mathrm{d}x\, (T_{\mathrm{cyl}} - \bar{T}_{\mathrm{cyl}})$ up to a factor i. Since $z^{-n} = e^{-2\pi n(t+ix)/L}$, the x-integral kills every $n \neq 0$:

$$\begin{aligned} H &\overset{(1)}{=} \frac{1}{2\pi} L \left(\frac{2\pi}{L}\right)^2 \Big(L_0 - \frac{c}{24} + \bar{L}_0 - \frac{c}{24}\Big) = \frac{2\pi}{L}\Big(L_0 + \bar{L}_0 - \frac{c}{12}\Big), \\ P &= \frac{2\pi}{L}\big(L_0 - \bar{L}_0\big). \end{aligned} \tag{11.196}$$

(1) $\int_0^L \mathrm{d}x\, e^{-2\pi i n x/L} = L\,\delta_{n,0}$; the same for the antiholomorphic half with $\bar{c} = c$.

Worked example 11.71: The Casimir energy measures c

The ground state is the vacuum, $L_0 = \bar{L}_0 = 0$. Its energy on a circle of length L is

$$E_0(L) = -\frac{2\pi}{L}\,\frac{c}{12} = -\frac{\pi c}{6L}. \tag{11.197}$$

This is a Casimir energy: it exists only because the circle is finite, and it is negative. On a lattice (or for the quantum Ising chain, whose transfer matrix is the lattice Hamiltonian) one measures $E_0(L) = e_\infty L - \pi c\, v/(6L) + \dots$, where e_∞ is a non-universal energy density and v the "speed of light" of the chain. Fitting the $1/L$ term gives c. Numerically this gives $c = 0.500$ for the Ising chain, $c = 0.700$ for the tricritical Ising chain and $c = 0.800$ for the three-state Potts chain (Blöte, Cardy, Nightingale; Affleck, 1986). The excited states give the weights: $E_i - E_0 = \frac{2\pi v}{L}(h_i + \bar{h}_i + N + \bar{N})$, a tower above each primary. So the entire operator content of §5 is visible in the spectrum of a finite chain. For the free Majorana fermion on a circle of length $L = 2\pi$, the zero-point sums of the intuition box in Section 11.6.3 gave $E_{\mathrm{NS}} = -\frac{1}{48}$ per chirality, i.e. total $-\frac{1}{24} = -c/12$ with $c = \frac{1}{2}$: the same answer.

11.7.4 The torus partition function as a trace

Now close the cylinder into a torus. Propagate for a Euclidean time t, translate in space by x, and identify the final circle with the initial one. The result is the torus with periods $\omega_1 = L$ (space) and $\omega_2 = x + it$, i.e.

$$\tau = \frac{x + it}{L}. \tag{11.198}$$

In quantum mechanics, "evolve and identify" is a trace. So the path integral on the torus is

$$\begin{aligned} Z(\tau) &\overset{(1)}{=} \operatorname{Tr}\, e^{-tH}\, e^{ixP} \\ &\overset{(2)}{=} \operatorname{Tr}\, \exp\!\left[-\frac{2\pi t}{L}\left(L_0 + \bar{L}_0 - \frac{c}{12}\right) + \frac{2\pi i x}{L}\left(L_0 - \bar{L}_0\right)\right] \\ &\overset{(3)}{=} \operatorname{Tr}\, \exp\!\left[2\pi i\tau\left(L_0 - \frac{c}{24}\right) - 2\pi i\bar{\tau}\left(\bar{L}_0 - \frac{c}{24}\right)\right] \\ &\overset{(4)}{=} \operatorname{Tr}\, q^{\,L_0 - c/24}\, \bar{q}^{\,\bar{L}_0 - c/24}, \qquad q = e^{2\pi i\tau}. \end{aligned} \tag{11.199}$$

(1) Periodic Euclidean time gives a trace (as $\operatorname{Tr} e^{-\beta H}$ in statistical mechanics); the spatial shift inserts the translation operator e^{ixP} before the trace closes.
(2) Insert (11.196).
(3) Collect the coefficient of L_0: $-2\pi t/L + 2\pi i x/L = 2\pi i(x + it)/L = 2\pi i\tau$; the coefficient of $\bar{L}_0$ is its complex conjugate $-2\pi i\bar{\tau}$. The constants: $-\frac{2\pi t}{L}\cdot(-\frac{c}{12}) = 2\pi \operatorname{Im}\tau\, \frac{c}{12}$, and $2\pi i\tau(-\frac{c}{24}) - 2\pi i\bar{\tau}(-\frac{c}{24}) = 2\pi \operatorname{Im}\tau\, \frac{c}{12}$ too.
(4) Write the exponentials as powers of q and $\bar{q} = e^{-2\pi i\bar{\tau}}$.

Here $\bar{q}$ is the complex conjugate of q, so $|q| = e^{-2\pi \operatorname{Im}\tau} < 1$ and the trace converges. The $-c/24$ shifts are the Casimir energy: they are exactly what makes Z modular invariant, as we will see.

Watch out: Lattice versus continuum partition function

The lattice Z at T_c is $e^{-f_\infty A}\, Z_{\mathrm{CFT}}(\tau)\,(1 + \dots)$, where $A = LL'$ is the area, f_∞ a non-universal bulk free energy, and the dots are corrections that vanish as the lattice spacing goes to zero. Only the universal factor $Z_{\mathrm{CFT}}(\tau)$ is modular invariant and depends only on τ. It is what Nash calls Z.

Characters

The Hilbert space of a CFT is a sum of products of irreducible representations of the two Virasoro algebras [Nash p. 311]:

$$\mathcal{H} = \bigoplus_{h,\bar{h}} m_{h,\bar{h}}\, L(c,h) \otimes \overline{L(c,\bar{h})}. \tag{11.200}$$

The trace (11.199) factorises on each summand, because L_0 acts only on the first factor and $\bar{L}_0$ only on the second. Each factor gives a **character**.

Working definition 11.26 (Virasoro character)

The character of the irreducible module $L(c,h)$ [the Verma module $V(c,h)$ with all its null vectors removed] is

$$\chi_h(\tau) = \operatorname{Tr}_{L(c,h)}\, q^{\,L_0 - c/24} = q^{\,h - c/24} \sum_{N \geq 0} d_N\, q^N,$$

where d_N is the number of independent states at level N [states with L_0 eigenvalue $h+N$]. *How to compute it:* count the states of the irreducible module level by level (the weight diagram of Figure 11.35) and put them into a power series.

Then

$$Z(\tau) \overset{(1)}{=} \sum_{h,\bar{h}} m_{h,\bar{h}} \, \mathrm{Tr}_{L(c,h)} \, q^{L_0 - c/24} \, \mathrm{Tr}_{\overline{L(c,\bar{h})}} \, \bar{q}^{\bar{L}_0 - c/24} \overset{(2)}{=} \sum_{h,\bar{h}} m_{h,\bar{h}} \, \chi_h(\tau) \, \bar{\chi}_{\bar{h}}(\bar{\tau}), \tag{11.201}$$

(1) The trace over a direct sum is the sum of traces; the trace over a tensor product of an operator $A \otimes B$ is $\mathrm{Tr}\, A \cdot \mathrm{Tr}\, B$.
(2) Definition of the character; the barred one is the complex conjugate function of τ.

which is Nash's (11.49). The multiplicities $m_{h,\bar{h}}$ are non-negative integers (they count copies of a representation), and $m_{0,0} = 1$ if the vacuum is unique.

Worked example 11.72: The Verma character and the partition generating function

For generic (c,h), $L(c,h) = V(c,h)$ has $p(N)$ states at level N, so

$$\chi_h^{\text{Verma}}(\tau) = q^{h-c/24} \sum_{N \geq 0} p(N) \, q^N \overset{(1)}{=} q^{h-c/24} \prod_{n=1}^{\infty} \frac{1}{1-q^n} \overset{(2)}{=} \frac{q^{\,h-(c-1)/24}}{\eta(\tau)}. \tag{11.202}$$

(1) Expand each factor: $\frac{1}{1-q^n} = \sum_{k_n \geq 0} q^{n k_n}$. A term of the product is a choice of $k_1, k_2, \ldots$, i.e. "use the part n exactly k_n times", and contributes $q^{\sum n k_n}$. So the coefficient of q^N counts partitions of N. (Equivalently: the state $L_{-1}^{k_1} L_{-2}^{k_2} \cdots |h\rangle$ is the partition with k_n parts equal to n.)
(2) Dedekind's function $\eta(\tau) = q^{1/24} \prod_n (1-q^n)$, Nash's (11.52), studied in the next section.

For the vacuum at generic c the only null state is $L_{-1}|0\rangle$ and its descendants, a Verma module starting at level 1, so $\chi_0 = q^{-c/24}(1-q)\prod_n (1-q^n)^{-1}$: the factor $(1-q)$ removes $p(N-1)$ states at level N.

Worked example 11.73: Spin must be an integer: T-invariance

Under $\tau \mapsto \tau + 1$, $q = e^{2\pi i \tau}$ is unchanged, but $q^{h-c/24}$ is multiplied by a phase:

$$\chi_h(\tau+1) \overset{(1)}{=} e^{2\pi i (h-c/24)} \, q^{h-c/24} \sum_N d_N q^N = e^{2\pi i (h-c/24)} \chi_h(\tau),$$

$$Z(\tau+1) \overset{(2)}{=} \sum_{h,\bar{h}} m_{h,\bar{h}} \, e^{2\pi i (h-c/24)} e^{-2\pi i (\bar{h}-c/24)} \chi_h(\tau) \bar{\chi}_{\bar{h}}(\bar{\tau}) = \sum_{h,\bar{h}} m_{h,\bar{h}} \, e^{2\pi i (h-\bar{h})} \chi_h \bar{\chi}_{\bar{h}}.$$

(1) $e^{2\pi i(\tau+1)a} = e^{2\pi i a} e^{2\pi i \tau a}$, and $e^{2\pi i N} = 1$ for integer levels N.
(2) Apply the first line to both factors; the antiholomorphic one gets the complex-conjugate phase.

So $Z(\tau+1) = Z(\tau)$ iff $h - \bar{h} \in \mathbb{Z}$ whenever $m_{h,\bar{h}} \neq 0$. The number $h - \bar{h}$ is the spin of the field (its eigenvalue under rotations, generated by $L_0 - \bar{L}_0$). T-invariance says: *every local field has integer spin*. It is the operator form of the statement that the Dehn twist of Figure 11.39 does nothing.

11.7.5 The Ising partition function

For the Ising model Nash writes

$$Z_{\text{Ising}} = \chi_0\bar{\chi}_0 + \chi_{1/2}\bar{\chi}_{1/2} + \chi_{1/16}\bar{\chi}_{1/16} = |\chi_0|^2 + |\chi_{1/2}|^2 + |\chi_{1/16}|^2, \tag{11.203}$$

his (11.50): $m_{h,\bar{h}} = \delta_{h,\bar{h}}$ for $h \in \{0, \frac{1}{2}, \frac{1}{16}\}$. Each primary appears once, paired with itself, so all three fields are spinless: $\mathbf{1} = (0,0)$, $\varepsilon = (\frac{1}{2}, \frac{1}{2})$, $\sigma = (\frac{1}{16}, \frac{1}{16})$. This is the **diagonal modular invariant**. To check its modular invariance we need the three characters explicitly. The free fermion gives them.

Worked example 11.74: The Ising characters from the Majorana fermion

NS sector. The NS Fock space is built by the modes ψ_{-r}, $r = \frac{1}{2}, \frac{3}{2}, \ldots$, each either empty or occupied. The trace factorises mode by mode:

$$\mathrm{Tr}_{\text{NS}}\, q^{L_0 - c/24} \overset{(1)}{=} q^{-1/48} \prod_{r \in \mathbb{N} - \frac{1}{2}} (1 + q^r), \qquad \mathrm{Tr}_{\text{NS}}\, (-1)^F q^{L_0 - c/24} \overset{(2)}{=} q^{-1/48} \prod_{r \in \mathbb{N} - \frac{1}{2}} (1 - q^r).$$

(1) $c/24 = 1/48$; an occupied mode ψ_{-r} adds r to L_0, so each mode contributes $1 + q^r$ (empty + occupied).

(2) An occupied mode also flips the sign of $(-1)^F$.

The projector onto even (odd) fermion number is $\frac{1}{2}(1 \pm (-1)^F)$, and by Section 11.6.3 the even part is the $\mathbf{1}$ module and the odd part the ε module. Hence

$$\chi_0 \pm \chi_{1/2} = q^{-1/48} \prod_{n=1}^{\infty} (1 \pm q^{n-1/2}), \qquad \chi_{1/16} = q^{1/24} \prod_{n=1}^{\infty} (1 + q^n). \tag{11.204}$$

R sector. The modes ψ_{-n}, $n \geq 1$, give $\prod_n (1 + q^n)$, and the ground state has $L_0 = \frac{1}{16}$, so the prefactor is $q^{1/16 - 1/48} = q^{1/24}$.

Watch out: The $\sqrt{2}$ of a single Majorana zero mode

The Ramond zero mode obeys $\psi_0^2 = \frac{1}{2}$. A Clifford algebra with one generator has no irreducible representation of dimension "1 with a fermion number": the smallest module with a $(-1)^F$ anticommuting with ψ_0 is two dimensional. For *two* Majorana fermions (one Dirac fermion) the two zero modes act on $\mathbb{C}^2$ like $\sigma_1/\sqrt{2}$ and $\sigma_2/\sqrt{2}$ (Chapter T4), and the R ground state is doubly degenerate. For one Majorana fermion the "degeneracy" is formally $\sqrt{2}$, which is why the full Ramond trace is $\sqrt{2}\,\chi_{1/16}$. The Ising model has a single spin field σ, and $\chi_{1/16}$ itself is the character that appears in Z.

Worked example 11.75: Checking the characters against the Verma-module count

Expand (11.204) (a short computer or hand expansion) and strip the prefactors $q^{h-1/48}$:

level N	0	1	2	3	4	5	6	7	8
χ_0: d_N	1	0	1	1	2	2	3	3	5
$\chi_{1/2}$: d_N	1	1	1	1	2	2	3	4	5
$\chi_{1/16}$: d_N	1	1	1	2	2	3	4	5	6
Verma: $p(N)$	1	1	2	3	5	7	11	15	22

The Verma-module picture predicts these numbers. The null vectors of $\phi_{r,s}$ sit at levels rs and $(2-r+1)(3-s+1)$, i.e. at levels $1, 6$ for $\mathbf{1} = \phi_{1,1}$, levels $2, 3$ for $\varepsilon = \phi_{2,1}$ and levels $2, 4$ for $\sigma = \phi_{1,2}$. Removing their submodules,

$$d_N^{(\mathbf{1})} = p(N) - p(N-1) - p(N-6) + \dots, \qquad d_6^{(\mathbf{1})} \overset{(1)}{=} 11 - 7 - 1 = 3,$$
$$d_N^{(\varepsilon)} = p(N) - p(N-2) - p(N-3) + \dots, \qquad d_4^{(\varepsilon)} \overset{(2)}{=} 5 - 2 - 1 = 2,$$
$$d_N^{(\sigma)} = p(N) - p(N-2) - p(N-4) + \dots, \qquad d_6^{(\sigma)} \overset{(3)}{=} 11 - 5 - 2 = 4,$$

(1) $p(6) = 11$, $p(5) = 7$, $p(0) = 1$.
(2) $p(4) = 5$, $p(2) = 2$, $p(1) = 1$.
(3) $p(6) = 11$, $p(4) = 5$, $p(2) = 2$.

all matching the fermion table. The dots are the inclusion–exclusion corrections from the overlap of the two submodules; they start at level 11 for $\mathbf{1}$, 7 for ε and 10 for σ. The full answer is the **Rocha-Caridi formula** $\chi_{r,s} = \eta(\tau)^{-1} \sum_{k\in\mathbb{Z}} \left(q^{(24k+4r-3s)^2/48} - q^{(24k+4r+3s)^2/48} \right)$ for the Ising model. So the free fermion and the abstract representation theory agree to every order we care to check.

Spin structures: why the fermion knows about modular invariance

The fermion picture explains *why* this particular combination of characters is modular invariant. A Majorana fermion on a torus needs a spin structure: a choice of periodic (P) or antiperiodic (A) boundary condition around each of the two cycles, four choices in all. Label them by (space, time). The traces we just computed are path integrals with definite spin structures:

- antiperiodic in space is the NS sector (on the cylinder, NS fermions are antiperiodic, Section 11.6.3);
- a plain trace gives *antiperiodic* time, because fermion path integrals with a trace have an extra sign; inserting $(-1)^F$ makes time periodic.

So, for the chiral fermion, $\mathrm{Tr}_{\mathrm{NS}}\, q^{L_0 - c/24} \leftrightarrow (A, A)$, $\mathrm{Tr}_{\mathrm{NS}} (-1)^F q^{L_0 - c/24} \leftrightarrow (A, P)$, $\mathrm{Tr}_{\mathrm{R}}\, q^{L_0 - c/24} \leftrightarrow (P, A)$, and $\mathrm{Tr}_{\mathrm{R}} (-1)^F q^{L_0 - c/24} \leftrightarrow (P, P)$, which vanishes because of the zero mode. A modular transformation relabels the cycles, so it permutes the spin structures (Figure 11.41): S swaps $(A, P) \leftrightarrow (P, A)$, and T swaps $(A, A) \leftrightarrow (A, P)$. An honest theory of the lattice spins must *sum over all spin structures*, and a sum over all of them is automatically modular invariant. Summing |fermion partition function| over the four spin structures gives

$$\tfrac{1}{2}\left(\left|\chi_0 + \chi_{1/2}\right|^2 + \left|\chi_0 - \chi_{1/2}\right|^2 + \left|\sqrt{2}\,\chi_{1/16}\right|^2 + 0 \right) \overset{(1)}{=} |\chi_0|^2 + |\chi_{1/2}|^2 + |\chi_{1/16}|^2 = Z_{\mathrm{Ising}}. \tag{11.205}$$

(1) $|a+b|^2 + |a-b|^2 = 2|a|^2 + 2|b|^2$, and $|\sqrt{2}\chi_{1/16}|^2 = 2|\chi_{1/16}|^2$.

This is the physical reason behind Nash's "routine" check. The factor $\frac{1}{2}$ is the projector $\frac{1}{2}(1 + (-1)^F)$ that keeps only states local with respect to the spin. We verify the modular invariance explicitly, with theta functions, in the next section.

Reading Nash (pp. 311–312, (11.48)–(11.51))

Nash's §4 is a single page: periodic boundary conditions give a torus; conformal invariance gives invariance under (11.48); the Hilbert space $\bigoplus_i H_i \otimes \bar{H}_i$ gives the character sum (11.49); the Ising model gives (11.50); and invariance under the generators S, T suffices

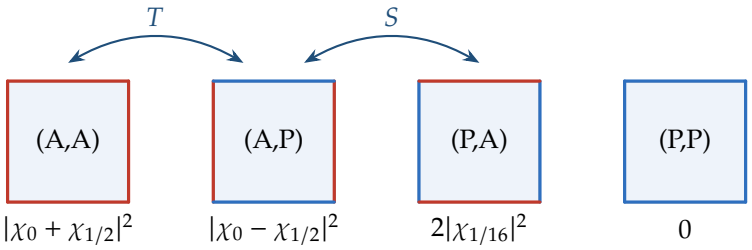


Figure 11.41: The four spin structures on a torus and the (left×right) fermion partition functions they give, in terms of Ising characters. The modular group permutes them: T exchanges (A,A) and (A,P), S exchanges (A,P) and (P,A), and (P,P) is invariant (its partition function vanishes by the zero mode). What to see: summing over all spin structures is modular invariant by construction, and the sum is $2Z_{\text{Ising}}$.

(11.51). We have derived each step: τ as the shape of the lattice (Section 11.7.1), (11.48) as a change of lattice basis (11.191), (11.49) from the cylinder Hamiltonian (11.199)–(11.201), and the three Ising characters from the free fermion (11.204). Note that Nash's $\chi_h(\tau)$ is our χ_h including the factor $q^{-c/24}$; without that factor Z would not be modular invariant.

Recap

- A torus is $\mathbb{C}/(\mathbb{Z}+\mathbb{Z}\tau)$; its conformal class is $\tau \in \mathbb{H}$ modulo $\mathrm{PSL}(2,\mathbb{Z})$, generated by $T:\tau\mapsto\tau+1$ and $S:\tau\mapsto -1/\tau$, with $S^2=(ST)^3=1$. The fundamental domain $\mathcal{F}$ is the moduli space of tori.
- On a circle of length L: $H=\frac{2\pi}{L}(L_0+\bar L_0-\frac{c}{12})$; the $-\frac{c}{12}$ is the Casimir energy from the Schwarzian, and it lets one measure c on a lattice.
- $Z(\tau)=\mathrm{Tr}\, q^{L_0-c/24}\bar q^{\bar L_0-c/24}=\sum m_{h,\bar h}\chi_h\bar\chi_{\bar h}$. T-invariance forces integer spins.
- The Ising characters are fermion traces (11.204); they agree with the Verma-module counting; Z_{Ising} is the sum over the four spin structures, hence modular invariant.

11.8 Modular forms and the modular invariance of Z

We have reduced Nash's claim to two equations, $Z(\tau+1)=Z(\tau)$ and $Z(-1/\tau)=Z(\tau)$. The first was easy: it only asks that every field have integer spin (Section 11.7.4). The second is hard. Under $\tau\mapsto -1/\tau$ the variable $q=e^{2\pi i\tau}$ goes to $\tilde q=e^{-2\pi i/\tau}$, and a power series in q has no obvious relation to a power series in $\tilde q$. If $\tau=it$ is purely imaginary, $q=e^{-2\pi t}$ and $\tilde q=e^{-2\pi/t}$: a long thin torus becomes a short fat one. Physically, S exchanges "a long time on a short circle" with "a short time on a long circle".

So the question of this section is:

Which functions of q have a known behaviour under $\tau\mapsto -1/\tau$, and how do we write the characters in terms of them?

The answer is the classical theory of modular forms, which Nash sketches in (11.52)–(11.59). We need three tools: the Dedekind η function, the Jacobi θ functions, and one piece of Fourier analysis, the Poisson summation formula. With these we prove the invariance of Z_{Ising} (Nash

calls this "a routine matter"), compute the free boson partition function, and see how modular invariance classifies theories.

11.8.1 The Dedekind eta function

Working definition 11.27 (Dedekind η function [Nash p. 312, (11.52)])

For τ in the upper half plane $\mathbb{H}$ and $q = e^{2\pi i\tau}$,

$$\eta(\tau) = e^{\pi i\tau/12}\prod_{n=1}^{\infty}(1-e^{2\pi in\tau}) = q^{1/24}\prod_{n=1}^{\infty}(1-q^n). \tag{11.206}$$

How to use it: $1/\eta(\tau)$ is "$q^{-1/24}$ times the generating function of partitions" [$\prod_n(1-q^n)^{-1} = \sum_N p(N)\,q^N = 1+q+2q^2+3q^3+5q^4+\cdots$]. It is the character of one free boson's oscillators [the states $a_{-n_1}\cdots a_{-n_k}|0\rangle$, of which there are $p(N)$ with $n_1+\cdots+n_k = N$]. Since $|q|<1$ the product converges, and η has no zeros in $\mathbb{H}$.

We already met $1/\eta$ in (11.202): the Verma module has $p(N)$ states at level N, and $\sum_N p(N)q^N = \prod_n(1-q^n)^{-1}$. The same function counts the states of a free boson. Its oscillators a_{-n}, $n\geq 1$, each raise L_0 by n, and each can be applied any number of times. A state $a_{-1}^{k_1}a_{-2}^{k_2}\cdots|0\rangle$ has level $\sum nk_n$, so

$$\mathrm{Tr}_{\text{oscillators}}\; q^{L_0-1/24} \overset{(1)}{=} q^{-1/24}\prod_{n=1}^{\infty}\Big(\sum_{k_n=0}^{\infty}q^{nk_n}\Big) \overset{(2)}{=} q^{-1/24}\prod_{n=1}^{\infty}\frac{1}{1-q^n} \overset{(3)}{=} \frac{1}{\eta(\tau)}. \tag{11.207}$$

(1) $c=1$ for one boson, so the shift is $q^{-1/24}$. The oscillators for different n are independent, so the trace is a product over n of traces over the occupation number k_n.
(2) Geometric series.
(3) Definition (11.206).

Worked example 11.76: The first terms of η, and Euler's pentagonal numbers

Multiply out the first few factors:

$$\begin{aligned}&(1-q)(1-q^2)(1-q^3)(1-q^4)(1-q^5)(1-q^6)(1-q^7)\cdots\\ &\quad = 1-q-q^2+q^5+q^7-q^{12}-q^{15}+\cdots.\end{aligned}$$

Almost everything cancels. Euler found the pattern: the surviving exponents are the **pentagonal numbers** $n(3n-1)/2$ for $n\in\mathbb{Z}$ $(0,1,2,5,7,12,15,\ldots)$, with sign $(-1)^n$. Including the prefactor,

$$\eta(\tau) \overset{(1)}{=} \sum_{n\in\mathbb{Z}}(-1)^n\,q^{\,n(3n-1)/2+1/24} \overset{(2)}{=} \sum_{n\in\mathbb{Z}}(-1)^n\,q^{\,(6n-1)^2/24}. \tag{11.208}$$

(1) Euler's pentagonal number theorem, times $q^{1/24}$.
(2) $(6n-1)^2/24 = (36n^2-12n+1)/24 = n(3n-1)/2+1/24$.

Check: $n=0,1,-1,2,-2$ give $(6n-1)^2 = 1,25,49,121,169$, that is $q^{1/24}$ times q^0,q^1,q^2,q^5,q^7 with signs $+,-,-,+,+$. The form (11.208), a Gaussian sum over a lattice, is what makes an S-transformation law possible: Gaussians are the functions that the Fourier transform maps to Gaussians.

The T law

This one is immediate:

$$\eta(\tau+1) \overset{(1)}{=} e^{\pi i(\tau+1)/12}\prod_n\left(1-e^{2\pi i n\tau}e^{2\pi i n}\right) \overset{(2)}{=} e^{\pi i/12}\,e^{\pi i\tau/12}\prod_n\left(1-e^{2\pi i n\tau}\right) = e^{\pi i/12}\,\eta(\tau). \quad (11.209)$$

(1) Substitute $\tau+1$ in (11.206).
(2) $e^{2\pi i n}=1$ for integer n; only the prefactor sees the shift.

This is Nash's (11.58). So η is *not* invariant: it picks up a 24th root of unity. The S law, Nash's (11.59),

$$\eta(-1/\tau) = (-i\tau)^{1/2}\,\eta(\tau), \quad (11.210)$$

needs "some additional work". We do that work in Section 11.8.3 after building the θ functions. Here $(-i\tau)^{1/2}$ is the branch that is positive for $\tau = it$, $t>0$; this is well defined because $-i\tau$ lies in the right half plane.

Watch out: Why η and not $\prod(1-q^n)$

The prefactor $q^{1/24}$ looks like decoration. It is not. Without it, $\prod(1-q^n)$ has no simple S law at all. With it, η transforms by the simple factor (11.210). This is the same $-c/24$ that the Casimir energy (11.197) put into the characters, for $c=1$. The zero-point energy of the cylinder is exactly what the torus needs to be modular invariant.

11.8.2 Modular forms

Nash now digresses to define modular forms. The idea is simple: we want functions on $\mathbb{H}$ that are not invariant under $\mathrm{PSL}(2,\mathbb{Z})$, but change by a *known, simple* factor.

Working definition 11.28 (modular form of weight k, Nash's convention [Nash p. 312, (11.54)])

A holomorphic function f on $\mathbb{H}$ is a **modular form of weight k** (or "dimension $-2k$") if the object

$$F(\tau) = f(\tau)\,(d\tau)^k$$

is invariant under every $m\in\mathrm{PSL}(2,\mathbb{Z})$ [every map $m(\tau)=\frac{a\tau+b}{c\tau+d}$ with integers a,b,c,d and $ad-bc=1$]. Nash calls F an **automorphic form** of weight k. *How to test it:* check $f(m(\tau))=(c\tau+d)^{2k}f(\tau)$ [since $dm(\tau)=d\tau/(c\tau+d)^2$, the factor f must make up $(c\tau+d)^{2k}$] for the two generators S and T (and require f to stay bounded as $\mathrm{Im}\,\tau\to\infty$).

The transformation law comes from how $d\tau$ transforms. For $m(\tau)=(a\tau+b)/(c\tau+d)$ with $ad-bc=1$, the derivative $m'(\tau)=(c\tau+d)^{-2}$ (we computed exactly this for Möbius maps in Section 11.5.1). Invariance of F means

$$\begin{aligned} F(m(\tau)) = F(\tau) &\overset{(1)}{=} f(m(\tau))\,\big(m'(\tau)\big)^k(d\tau)^k = f(\tau)(d\tau)^k \\ &\overset{(2)}{=} f(m(\tau)) = (c\tau+d)^{2k}f(\tau). \end{aligned} \quad (11.211)$$

(1) Pull back: $d(m(\tau)) = m'(\tau)\,d\tau$.
(2) $m'(\tau)^k=(c\tau+d)^{-2k}$; cancel $(d\tau)^k$.

This is Nash's (11.55).

Watch out: Nash's weight is half the usual weight

In almost every modern text, a modular form of weight w obeys $f(m(\tau)) = (c\tau + d)^w f(\tau)$. Nash's "weight k" is the modern weight $w = 2k$: he counts powers of $\mathrm{d}\tau$, which carries $(c\tau + d)^{-2}$. So Nash's "η^{24} is a modular form of weight 6" is the modern "$\Delta = \eta^{24}$ is the cusp form of weight 12". Both say (11.213) below.

Why is this consistent, that is, why does it suffice to check S and T? Write $j(M,\tau) = c\tau + d$ for $M = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$. Then j is a **cocycle**:

$$j(MN,\tau) \overset{(1)}{=} \frac{\mathrm{d}\tau}{\mathrm{d}(MN\tau)}\Big|^{1/2} \overset{(2)}{=} \left(\frac{\mathrm{d}\tau}{\mathrm{d}(N\tau)} \cdot \frac{\mathrm{d}(N\tau)}{\mathrm{d}(MN\tau)}\right)^{1/2} \overset{(3)}{=} j(N,\tau)\, j(M,N\tau), \tag{11.212}$$

(1) $\mathrm{d}(M\tau)/\mathrm{d}\tau = j(M,\tau)^{-2}$, so $j = (\mathrm{d}\tau/\mathrm{d}(M\tau))^{1/2}$ (up to a sign, which one checks directly by multiplying the matrices).
(2) Chain rule for the composite map $\tau \mapsto N\tau \mapsto M(N\tau)$.
(3) Each factor is the corresponding j^2; take square roots.

So if $f(M\tau) = j(M,\tau)^{2k} f(\tau)$ and $f(N\tau) = j(N,\tau)^{2k} f(\tau)$, then $f(MN\tau) = j(M,N\tau)^{2k} f(N\tau) = j(M,N\tau)^{2k} j(N,\tau)^{2k} f(\tau) = j(MN,\tau)^{2k} f(\tau)$. The law for generators implies the law for every word in them.

Worked example 11.77: η^{24} is a modular form: Nash's (11.56)–(11.57)

Raise the two laws (11.209) and (11.210) to the 24th power:

$$\eta^{24}(\tau+1) \overset{(1)}{=} \left(e^{\pi i/12}\right)^{24} \eta^{24}(\tau) = e^{2\pi i} \eta^{24}(\tau) = \eta^{24}(\tau),$$
$$\eta^{24}(-1/\tau) \overset{(2)}{=} (-i\tau)^{12}\, \eta^{24}(\tau) \overset{(3)}{=} \tau^{12}\, \eta^{24}(\tau).$$

(1) The T law.
(2) The S law; $\left((-i\tau)^{1/2}\right)^{24} = (-i\tau)^{12}$, with no branch ambiguity left since the exponent is an integer.
(3) $(-i)^{12} = \left((-i)^4\right)^3 = 1$.

These are Nash's (11.57). T has $c = 0, d = 1$, so $(c\tau + d)^{12} = 1$; S has $c = 1, d = 0$, so $(c\tau + d)^{12} = \tau^{12}$. By the cocycle argument,

$$\eta^{24}\big(m(\tau)\big) = (c\tau + d)^{12}\, \eta^{24}(\tau) \qquad \text{for all } m \in \mathrm{PSL}(2,\mathbb{Z}), \tag{11.213}$$

which is Nash's (11.56): η^{24} has Nash-weight 6 ($12 = 2 \cdot 6$). Its expansion begins

$$\eta^{24}(\tau) \overset{(1)}{=} q\,(1-q)^{24}(1-q^2)^{24}\cdots \overset{(2)}{=} q\Big(1 - 24q + \big[\tbinom{24}{2} - 24\big]q^2 + \dots\Big)$$
$$= q - 24q^2 + 252q^3 - \cdots.$$

(1) $(q^{1/24})^{24} = q$.
(2) From $(1-q)^{24}$: $1 - 24q + \binom{24}{2}q^2$; from $(1-q^2)^{24}$: $-24q^2$. $\binom{24}{2} = 276$ and $276 - 24 = 252$.

The coefficients $1, -24, 252, -1472, \dots$ are Ramanujan's $\tau(n)$.

Physical picture: Why 24 is the bosonic string's number

$1/\eta^{24}$ is the partition function of 24 free bosons' oscillators, $q^{-1}(1 + 24q + 324q^2 + \dots)$. These are the 24 transverse oscillators of the bosonic string in 26 dimensions (Chapter 9). The

leading q^{-1} is the tachyon, and $-c/24 = -24/24 = -1$ is the famous intercept. The fact that η^{24} is a modular form, while η alone is not, is one way of seeing why the light-cone string needs exactly 24 transverse directions for its one-loop amplitude to be modular invariant.

11.8.3 Poisson summation and the theta functions

Everything in the S law comes from one fact of Fourier analysis.

Key idea: Poisson summation

For a smooth, rapidly decaying $f:\mathbb{R}\to\mathbb{C}$ with Fourier transform $\hat f(k)=\int_{\mathbb{R}} f(x)e^{-2\pi i k x}\,dx$,

$$\sum_{n\in\mathbb{Z}} f(n) = \sum_{k\in\mathbb{Z}} \hat f(k). \tag{11.214}$$

A sum over a lattice equals the sum of the Fourier transform over the dual lattice.

Why it is true: periodise f. The function $P(x)=\sum_n f(x+n)$ has period 1, so it has a Fourier series $P(x)=\sum_k c_k e^{2\pi i k x}$. Its coefficients are

$$c_k \overset{(1)}{=} \int_0^1 P(x)e^{-2\pi i k x}dx \overset{(2)}{=} \sum_{n\in\mathbb{Z}}\int_0^1 f(x+n)e^{-2\pi i k(x+n)}dx \overset{(3)}{=} \int_{\mathbb{R}} f(y)e^{-2\pi i k y}dy = \hat f(k).$$

(1) Fourier coefficient of a period-1 function.
(2) Insert P; $e^{-2\pi i k n}=1$, so we may add n to x in the exponent.
(3) Substitute $y=x+n$; the intervals $[n,n+1]$ tile $\mathbb{R}$.

Now set $x=0$: $\sum_n f(n)=P(0)=\sum_k c_k=\sum_k \hat f(k)$.

The functions we need are Gaussians. For $\operatorname{Im}\tau>0$ and $f(x)=e^{i\pi\tau x^2}$,

$$\hat f(k) \overset{(1)}{=} \int_{\mathbb{R}} e^{i\pi\tau x^2-2\pi i k x}dx \overset{(2)}{=} e^{-i\pi k^2/\tau}\int_{\mathbb{R}} e^{i\pi\tau(x-k/\tau)^2}dx \overset{(3)}{=} (-i\tau)^{-1/2}\,e^{-i\pi k^2/\tau}. \tag{11.215}$$

(1) Definition of $\hat f$.
(2) Complete the square: $\tau x^2-2kx=\tau(x-k/\tau)^2-k^2/\tau$.
(3) Shift the contour back to the real axis (the integrand decays in the strip in between), then $\int e^{-\pi a y^2}dy=a^{-1/2}$ with $a=-i\tau$, $\operatorname{Re} a=\operatorname{Im}\tau>0$.

A narrow Gaussian has a wide Fourier transform and vice versa (Figure 11.42). This is the uncertainty principle, and it is exactly the exchange of a thin torus with a fat one.

Working definition 11.29 (Jacobi theta functions)

With $q=e^{2\pi i\tau}$ (so $q^{n^2/2}=e^{i\pi\tau n^2}$),

$$\theta_3(\tau)=\sum_{n\in\mathbb{Z}} q^{n^2/2},\qquad \theta_4(\tau)=\sum_{n\in\mathbb{Z}}(-1)^n q^{n^2/2},\qquad \theta_2(\tau)=\sum_{n\in\mathbb{Z}} q^{(n+1/2)^2/2}. \tag{11.216}$$

How to think of them: each is a Gaussian summed over a one-dimensional lattice ($\mathbb{Z}$ or $\mathbb{Z}+\tfrac12$), possibly with a sign [$e^{i\pi\tau n^2}$ is a Gaussian in n, decaying because $\operatorname{Im}\tau>0$]. They are partition functions of a particle on a circle [the state of momentum n has energy proportional to n^2 and contributes $q^{n^2/2}$ to the trace], with momentum quantised in $\mathbb{Z}$ or in $\mathbb{Z}+\tfrac12$.

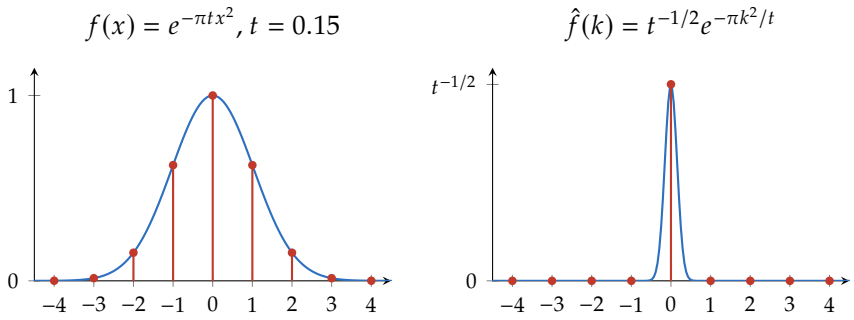


Figure 11.42: Poisson summation for a Gaussian, $\tau = it$. Left: a wide Gaussian sampled at the integers; many samples contribute, so $\theta_3(it)$ is large and slowly converging. Right: its Fourier transform is a narrow, tall Gaussian; only the $k = 0$ sample matters, giving $t^{-1/2}$. The two red sums are equal. What to see: $\theta_3(it) = t^{-1/2}\theta_3(i/t)$ trades a slowly converging sum for a rapidly converging one, which is the whole content of the S transformation.

S and T laws of the theta functions

Worked example 11.78: The S laws by Poisson summation

θ_3. Apply (11.214) to $f(x) = e^{i\pi\tau x^2}$:

$$\theta_3(\tau) \overset{(1)}{=} \sum_n e^{i\pi\tau n^2} \overset{(2)}{=} (-i\tau)^{-1/2} \sum_k e^{-i\pi k^2/\tau} \overset{(3)}{=} (-i\tau)^{-1/2}\, \theta_3(-1/\tau).$$

(1) Definition (11.216).
(2) Poisson summation with (11.215).
(3) $e^{-i\pi k^2/\tau} = e^{i\pi(-1/\tau)k^2}$ is the θ_3 summand at $-1/\tau$.

$\theta_4 \to \theta_2$. Now $f(x) = e^{i\pi\tau x^2} e^{i\pi x}$, so that $f(n)$ carries $(-1)^n$. The extra phase shifts the Fourier variable by $\frac{1}{2}$: $\hat{f}(k) = (-i\tau)^{-1/2} e^{-i\pi(k-1/2)^2/\tau}$ (repeat (11.215) with $k \to k - \frac{1}{2}$). Hence

$$\theta_4(\tau) \overset{(1)}{=} \sum_n e^{i\pi\tau n^2} e^{i\pi n} \overset{(2)}{=} (-i\tau)^{-1/2} \sum_k e^{-i\pi(k-1/2)^2/\tau} \overset{(3)}{=} (-i\tau)^{-1/2}\, \theta_2(-1/\tau).$$

(1) $(-1)^n = e^{i\pi n}$.
(2) Poisson summation.
(3) $k - \frac{1}{2}$ runs over $\mathbb{Z} + \frac{1}{2}$ as k runs over $\mathbb{Z}$.

Replacing τ by $-1/\tau$ in the last line, and using $(i/\tau)^{-1/2} = (-i\tau)^{1/2}$, gives the partner law. Altogether:

$$\theta_3(-\tfrac{1}{\tau}) = (-i\tau)^{1/2}\theta_3(\tau), \quad \theta_2(-\tfrac{1}{\tau}) = (-i\tau)^{1/2}\theta_4(\tau), \quad \theta_4(-\tfrac{1}{\tau}) = (-i\tau)^{1/2}\theta_2(\tau). \tag{11.217}$$

The T laws are one-liners. Since $n^2 \equiv n \pmod 2$, $e^{i\pi n^2} = (-1)^n$; and $(n + \frac{1}{2})^2 = n(n+1) + \frac{1}{4}$ with $n(n+1)$ even. So

$$\theta_3(\tau + 1) = \theta_4(\tau), \qquad \theta_4(\tau + 1) = \theta_3(\tau), \qquad \theta_2(\tau + 1) = e^{i\pi/4}\, \theta_2(\tau). \tag{11.218}$$

S swaps $\theta_2 \leftrightarrow \theta_4$ and T swaps $\theta_3 \leftrightarrow \theta_4$. Compare Figure 11.41: this is the same permutation as the spin structures $(A, P) \leftrightarrow (P, A)$ and $(A, A) \leftrightarrow (A, P)$. That is no accident, as we now see.

Product formulas: the Jacobi triple product

To connect the θ's with the fermion characters we need them as products. The bridge is Jacobi's identity, valid for $|q|<1$ and $y\neq 0$:

$$\sum_{n\in\mathbb{Z}} q^{n^2/2}y^n = \prod_{n=1}^{\infty}(1-q^n)\big(1+yq^{n-1/2}\big)\big(1+y^{-1}q^{n-1/2}\big). \tag{11.219}$$

Physical picture: The triple product is bosonisation

Read the right side as a trace over the NS Fock space of a complex (Dirac) fermion: modes $\psi^\dagger_{-r}$ (charge $+1$) and ψ_{-r} (charge -1), $r\in\mathbb{N}-\frac12$, each empty or occupied, with y counting charge. That gives $\prod(1+yq^r)(1+y^{-1}q^r)$. Read the left side as a compact boson: a sector of charge n has lowest weight $n^2/2$ (the lowest state of charge n fills the n lowest levels $\frac12+\frac32+\cdots+\frac{2n-1}{2}=\frac{n^2}{2}$), and on top of it sits a tower of neutral excitations counted by $\prod(1-q^n)^{-1}$. Multiplying both sides by $\prod(1-q^n)^{-1}$, the identity says: *the Dirac fermion Fock space equals* $\bigoplus_n$(*charge* n) $\otimes$ (*boson Fock space*). This is the boson–fermion correspondence of [Nash p. 303] in its sharpest form.

Worked example 11.79: Checking the triple product at charge 0 and 1

Expand the right side of (11.219), divided by $\prod(1-q^n)$, and collect the coefficient of y^0 up to q^2. The neutral states are: nothing (1); $\psi^\dagger_{-1/2}\psi_{-1/2}$ (q); $\psi^\dagger_{-3/2}\psi_{-1/2}$ and $\psi^\dagger_{-1/2}\psi_{-3/2}$ $(2q^2)$. So the fermion count is $1+q+2q^2+\cdots$. The boson count is $\prod(1-q^n)^{-1}=1+q+2q^2+\cdots$: the same. At y^1: the lowest state is $\psi^\dagger_{-1/2}$ with weight $\frac12=1^2/2$, then $\psi^\dagger_{-3/2}$ $(q^{3/2})$, then $\psi^\dagger_{-5/2}$ and $\psi^\dagger_{-1/2}\psi^\dagger_{-3/2}\psi_{-1/2}$ $(2q^{5/2})$: so $q^{1/2}(1+q+2q^2+\dots)$, again matching.

Setting $y=1$, $y=-1$ and $y=q^{1/2}$ in (11.219) gives

$$\theta_3(\tau)\overset{(1)}{=}\prod_{n\geq1}(1-q^n)(1+q^{n-1/2})^2,\qquad \theta_4(\tau)\overset{(2)}{=}\prod_{n\geq1}(1-q^n)(1-q^{n-1/2})^2,$$
$$\theta_2(\tau)\overset{(3)}{=}2q^{1/8}\prod_{n\geq1}(1-q^n)(1+q^n)^2. \tag{11.220}$$

(1) $y=1$: the left side is θ_3.
(2) $y=-1$: the left side is θ_4.
(3) $y=q^{1/2}$: the left side is $\sum q^{(n^2+n)/2}=q^{-1/8}\sum q^{(n+1/2)^2/2}=q^{-1/8}\theta_2$. On the right, $(1+q^n)(1+q^{n-1})$; the $n=1$ factor $1+q^0=2$ gives the 2, and the rest reassembles to $\prod_n(1+q^n)^2$.

Worked example 11.80: The S law of η: Nash's "additional work"

First, a product identity:

$$\theta_2\theta_3\theta_4\overset{(1)}{=}2q^{1/8}\prod_n(1-q^n)^3\Big[(1+q^n)(1+q^{n-1/2})(1-q^{n-1/2})\Big]^2$$
$$\overset{(2)}{=}2q^{1/8}\prod_n(1-q^n)^3\Big[(1+q^n)(1-q^{2n-1})\Big]^2\overset{(3)}{=}2q^{1/8}\prod_n(1-q^n)^3\overset{(4)}{=}2\eta^3. \tag{11.221}$$

(1) Multiply the three products (11.220).
(2) $(1+x)(1-x)=1-x^2$ with $x=q^{n-1/2}$.
(3) Euler: $\prod_n(1+q^n)=\prod_n\frac{1-q^{2n}}{1-q^n}=\frac{\prod_{\text{even }m}(1-q^m)}{\prod_{\text{all }m}(1-q^m)}=\prod_{\text{odd }m}\frac{1}{1-q^m}$, so the bracket is 1.

(4) $\eta^3 = q^{1/8}\prod(1-q^n)^3$.

Now apply S to both sides, using (11.217):

$$2\eta^3(-1/\tau) \overset{(1)}{=} \theta_2(-\tfrac{1}{\tau})\theta_3(-\tfrac{1}{\tau})\theta_4(-\tfrac{1}{\tau}) \overset{(2)}{=} (-i\tau)^{3/2}\,\theta_4\theta_3\theta_2(\tau) \overset{(3)}{=} (-i\tau)^{3/2}\,2\eta^3(\tau).$$

(1) (11.221) at $-1/\tau$.
(2) Each θ gives a factor $(-i\tau)^{1/2}$; θ_2 and θ_4 swap.
(3) (11.221) at τ.

So $\eta(-1/\tau) = \omega\,(-i\tau)^{1/2}\eta(\tau)$ with $\omega^3 = 1$. Both sides are holomorphic and nonzero on $\mathbb{H}$, so ω is a continuous function with values in $\{1, e^{\pm 2\pi i/3}\}$: it is constant. At $\tau = i$, $-1/\tau = i$ and $(-i\cdot i)^{1/2} = 1$, so $\eta(i) = \omega\,\eta(i)$ and $\omega = 1$. This proves (11.210), Nash's (11.59). A numerical check at $\tau = 0.23 + 0.91i$ agrees to 30 digits.

11.8.4 The Ising model is modular invariant

Compare the fermion characters (11.204) with the products (11.220):

$$\frac{\theta_3}{\eta} \overset{(1)}{=} \frac{\prod(1-q^n)(1+q^{n-1/2})^2}{q^{1/24}\prod(1-q^n)} = \Big(q^{-1/48}\prod(1+q^{n-1/2})\Big)^2 \overset{(2)}{=} (\chi_0+\chi_{1/2})^2,$$
$$\frac{\theta_4}{\eta} \overset{(3)}{=} (\chi_0-\chi_{1/2})^2, \qquad \frac{\theta_2}{2\eta} \overset{(4)}{=} \frac{q^{1/8}\prod(1+q^n)^2}{q^{1/24}} = \Big(q^{1/24}\prod(1+q^n)\Big)^2 = \chi_{1/16}^2. \tag{11.222}$$

(1) Insert the product formulas; the $\prod(1-q^n)$ cancel, and $q^{-1/24} = (q^{-1/48})^2$.
(2) The NS character (11.204) with the + sign.
(3) The same with − signs.
(4) $q^{1/8-1/24} = q^{1/12} = (q^{1/24})^2$; then the R character.

So the three Ising characters are square roots of theta functions over η:

$$\chi_0 = \tfrac{1}{2}\Big(\sqrt{\tfrac{\theta_3}{\eta}} + \sqrt{\tfrac{\theta_4}{\eta}}\Big), \qquad \chi_{1/2} = \tfrac{1}{2}\Big(\sqrt{\tfrac{\theta_3}{\eta}} - \sqrt{\tfrac{\theta_4}{\eta}}\Big), \qquad \chi_{1/16} = \sqrt{\tfrac{\theta_2}{2\eta}}. \tag{11.223}$$

(The square roots are the branches that are positive for $\tau = it$.)

The partition function in closed form

$$\begin{aligned} Z_{\text{Ising}} &\overset{(1)}{=} \tfrac{1}{2}\big|\chi_0+\chi_{1/2}\big|^2 + \tfrac{1}{2}\big|\chi_0-\chi_{1/2}\big|^2 + |\chi_{1/16}|^2 \\ &\overset{(2)}{=} \frac{1}{2}\Big|\frac{\theta_3}{\eta}\Big| + \frac{1}{2}\Big|\frac{\theta_4}{\eta}\Big| + \frac{1}{2}\Big|\frac{\theta_2}{\eta}\Big| = \frac{|\theta_2|+|\theta_3|+|\theta_4|}{2\,|\eta|}. \end{aligned} \tag{11.224}$$

(1) Parallelogram law $|a|^2+|b|^2 = \tfrac{1}{2}|a+b|^2 + \tfrac{1}{2}|a-b|^2$.
(2) (11.222): $|\chi_0 \pm \chi_{1/2}|^2 = |\theta_{3,4}/\eta|$ and $|\chi_{1/16}|^2 = |\theta_2|/(2|\eta|)$.

Worked example 11.81: The routine verification: $Z(-1/\tau) = Z(\tau+1) = Z(\tau)$

S. From (11.210) and (11.217), every $|\theta_i|$ and $|\eta|$ gets the same factor $|-i\tau|^{1/2} = |\tau|^{1/2}$:

$$Z(-\tfrac{1}{\tau}) \overset{(1)}{=} \frac{|\tau|^{1/2}\big(|\theta_4|+|\theta_3|+|\theta_2|\big)}{2|\tau|^{1/2}|\eta|} \overset{(2)}{=} Z(\tau).$$

(1) S laws; θ_2 and θ_4 are exchanged.
(2) The factor cancels, and the sum is symmetric in θ_2, θ_4.

T. From (11.209) and (11.218), $|\eta|$ and $|\theta_2|$ are unchanged (phases), and $\theta_3 \leftrightarrow \theta_4$. The sum is symmetric, so $Z(\tau+1) = Z(\tau)$. That is all: the Ising partition function is modular invariant. Numerically, at $\tau = 0.23 + 0.91i$ all three of $Z(\tau)$, $Z(-1/\tau)$, $Z(\tau+1)$ equal $1.894632739812\ldots$.

Nash's formula (11.53)

Nash writes the squared moduli with η alone, at the arguments $\tau/2$, τ and 2τ. To get there, express the half-integer products through η:

$$\eta(\tau/2) \overset{(1)}{=} q^{1/48}\prod_{m\geq 1}(1-q^{m/2}) \overset{(2)}{=} q^{1/48}\prod_{n}(1-q^{n})\prod_{n}(1-q^{n-1/2}) \overset{(3)}{=} q^{1/48-1/24}\,\eta(\tau)\prod_{n}(1-q^{n-1/2}),$$
$$\eta(2\tau) \overset{(4)}{=} q^{1/12}\prod_{n}(1-q^{2n}) \overset{(5)}{=} q^{1/12}\prod_{n}(1-q^{n})(1+q^{n}) = q^{1/24}\,\eta(\tau)\prod_{n}(1+q^{n}).$$

(1) (11.206) at $\tau/2$: $q \to q^{1/2}$.
(2) Split m into even ($m = 2n$) and odd ($m = 2n-1$).
(3) $\prod(1-q^n) = q^{-1/24}\eta(\tau)$.
(4) (11.206) at 2τ: $q \to q^2$.
(5) $1 - q^{2n} = (1-q^n)(1+q^n)$.

So $\prod(1-q^{n-1/2}) = q^{1/48}\eta(\tau/2)/\eta(\tau)$ and $\prod(1+q^n) = q^{-1/24}\eta(2\tau)/\eta(\tau)$. With Euler's identity, $\prod(1+q^{n-1/2}) = \prod(1-q^{2n-1})/\prod(1-q^{n-1/2})$, and $\prod(1-q^{2n-1}) = \prod(1-q^n)/\prod(1-q^{2n}) = q^{1/24}\eta(\tau)/\eta(2\tau)$. Inserting into (11.204) and (11.224):

$$|\chi_0|^2 + |\chi_{1/2}|^2 = \frac{1}{2}\left\{\left|\frac{\eta^2(\tau/2)}{\eta^2(\tau)}\right| + \left|\frac{\eta^4(\tau)}{\eta^2(2\tau)\,\eta^2(\tau/2)}\right|\right\}, \qquad |\chi_{1/16}|^2 = \left|\frac{\eta^2(2\tau)}{\eta^2(\tau)}\right|. \tag{11.225}$$

The first is Nash's (11.53) exactly: the two terms are $|\theta_4/\eta|$ and $|\theta_3/\eta|$.

Watch out: A stray factor of 2 in Nash's (11.53)

Nash writes $|\chi_{1/16}|^2 = 2\,|\eta^2(2\tau)/\eta^2(\tau)|$. The factor 2 is a misprint. From (11.204), $\chi_{1/16} = q^{1/24}(1+q+\ldots)$, so $|\chi_{1/16}|^2 \to |q|^{1/12}$ as $q \to 0$, while $|\eta^2(2\tau)/\eta^2(\tau)| = |q|^{2/12-1/12}(1+\ldots) = |q|^{1/12}(1+\ldots)$ with coefficient 1. A numerical check confirms it: at $\tau = 0.23 + 0.91i$ both sides of our (11.225) equal $0.6214738\ldots$. With Nash's factor 2 the partition function would contain the spin field twice and would *fail* the S test.

The modular group acts on the characters: the S and T matrices

The S law does more than prove invariance. It says that the *vector* of characters transforms linearly. Using (11.217), $\sqrt{\theta_3/\eta} \mapsto \sqrt{\theta_3/\eta}$, $\sqrt{\theta_4/\eta} \mapsto \sqrt{\theta_2/\eta} = \sqrt{2}\,\chi_{1/16}$ and $\sqrt{\theta_2/2\eta} \mapsto \sqrt{\theta_4/2\eta}$ (the factors $(-i\tau)^{1/2}$ cancel between numerator and denominator). So, in the basis $(\chi_0, \chi_{1/2}, \chi_{1/16})$,

$$\chi_0(-\tfrac{1}{\tau}) \overset{(1)}{=} \tfrac{1}{2}\left(\sqrt{\tfrac{\theta_3}{\eta}} + \sqrt{2}\,\chi_{1/16}\right) \overset{(2)}{=} \tfrac{1}{2}\chi_0 + \tfrac{1}{2}\chi_{1/2} + \tfrac{1}{\sqrt{2}}\chi_{1/16},$$
$$\chi_{1/2}(-\tfrac{1}{\tau}) = \tfrac{1}{2}\chi_0 + \tfrac{1}{2}\chi_{1/2} - \tfrac{1}{\sqrt{2}}\chi_{1/16}, \qquad \chi_{1/16}(-\tfrac{1}{\tau}) \overset{(3)}{=} \tfrac{1}{\sqrt{2}}(\chi_0 - \chi_{1/2}). \tag{11.226}$$

(1) (11.223) and the S images just listed.
(2) $\sqrt{\theta_3/\eta} = \chi_0 + \chi_{1/2}$.
(3) $\sqrt{\theta_4/2\eta} = (\chi_0 - \chi_{1/2})/\sqrt{2}$.

That is, $\chi_i(-1/\tau) = \sum_j S_{ij}\chi_j(\tau)$ and $\chi_i(\tau+1) = \sum_j T_{ij}\chi_j(\tau)$ with

$$S = \frac{1}{2}\begin{pmatrix} 1 & 1 & \sqrt{2} \\ 1 & 1 & -\sqrt{2} \\ \sqrt{2} & -\sqrt{2} & 0 \end{pmatrix}, \qquad T = \begin{pmatrix} e^{-2\pi i/48} & 0 & 0 \\ 0 & e^{2\pi i\,23/48} & 0 \\ 0 & 0 & e^{2\pi i/24} \end{pmatrix}. \tag{11.227}$$

The entries of T are $e^{2\pi i(h-c/24)}$ with $h = 0, \frac{1}{2}, \frac{1}{16}$ and $c/24 = \frac{1}{48}$.

Worked example 11.82: The characters carry a representation of the modular group

S is real and symmetric, and $S^2 = 1$:

$$(S^2)_{11} = \tfrac{1}{4}(1+1+2) = 1, \qquad (S^2)_{12} = \tfrac{1}{4}(1+1-2) = 0, \qquad (S^2)_{13} = \tfrac{1}{4}(\sqrt{2}-\sqrt{2}+0) = 0,$$
$$(S^2)_{33} = \tfrac{1}{4}(2+2+0) = 1,$$

and the others by symmetry. So S is also orthogonal, i.e. unitary. Multiplying the matrices (numerically is quickest) gives $(ST)^3 = 1$ exactly. So S and T satisfy the defining relations of PSL$(2,\mathbb{Z})$ (Section 11.7.2): *the characters of a rational CFT carry a finite-dimensional unitary representation of the modular group.* Modular invariance of a combination $Z = \sum_{ij} M_{ij}\chi_i\bar{\chi}_j$ is then the matrix statement

$$S^{\mathsf{T}} M\, \bar{S} = M, \qquad T^{\mathsf{T}} M\, \bar{T} = M, \tag{11.228}$$

i.e. M commutes with S and T (both are unitary and symmetric). For Ising, $M = I$ works because S is unitary and T is diagonal. This matrix S is the protagonist of Verlinde's formula in Section 11.10.5.

11.8.5 The free boson on a circle

The brief promised the free boson's partition function, and it is the best illustration of modular invariance: we can compute it both in the operator language and in the path-integral language, and modular invariance is the statement that the two agree.

Conventions

We take the action $S = \frac{1}{4\pi}\int \mathrm{d}^2x\,(\partial_\mu\varphi)^2$, which gives $\langle\varphi(x)\varphi(0)\rangle = -\ln|x| = -\frac{1}{2}\ln|z|^2$, the stress tensor $T = -(\partial\varphi)^2$ with $c = 1$, and vertex operators $:e^{ip\varphi}:$ of weight $(p^2/4, p^2/4)$ as in Nash's §5. Now make the target a circle: $\varphi \sim \varphi + 2\pi R$.

On a spatial circle of length 2π two kinds of zero modes appear.

- **Momentum.** The wave function $e^{ip\varphi_0}$ of the centre of mass must be single valued under $\varphi_0 \to \varphi_0 + 2\pi R$, so $p = n/R$, $n \in \mathbb{Z}$.
- **Winding.** The field may wind around the target: $\varphi(x+2\pi) = \varphi(x) + 2\pi R w$, $w \in \mathbb{Z}$. This is a topological sector that a non-compact boson does not have.

The zero-mode energy is kinetic plus gradient energy. With canonical momentum $\Pi = \frac{1}{2\pi}\partial_t\varphi$ and total momentum $\int \Pi = n/R$, and $\partial_x\varphi = Rw$ for the winding,

$$H_0 \overset{(1)}{=} \int_0^{2\pi} \mathrm{d}x \left[\pi\,\Pi^2 + \frac{1}{4\pi}(\partial_x\varphi)^2\right] \overset{(2)}{=} 2\pi\left[\pi\frac{(n/R)^2}{4\pi^2} + \frac{R^2w^2}{4\pi}\right] = \frac{1}{2}\Big(\frac{n^2}{R^2} + R^2w^2\Big).$$

(1) The Hamiltonian density of $\mathcal{L} = \frac{1}{4\pi}(\dot\varphi^2 - \varphi'^2)$ is $\pi\Pi^2 + \frac{1}{4\pi}\varphi'^2$.
(2) Constant zero modes: $\Pi = (n/R)/2\pi$, $\varphi' = Rw$; the integral gives a factor 2π.

Since $H = L_0 + \bar L_0 - \frac{1}{12}$ on this circle, the zero modes contribute $h + \bar h = \frac{1}{2}(n^2/R^2 + R^2w^2)$, and the momentum $P = L_0 - \bar L_0$ is nw (the product of the two quantum numbers). Both are reproduced by

$$h = \tfrac{1}{4}p_L^2, \quad \bar h = \tfrac{1}{4}p_R^2, \qquad p_L = \frac{n}{R} + wR, \quad p_R = \frac{n}{R} - wR. \tag{11.229}$$

Indeed $\frac{1}{4}(p_L^2 + p_R^2) = \frac{1}{2}(n^2/R^2 + w^2R^2)$ and $\frac{1}{4}(p_L^2 - p_R^2) = \frac{1}{4}\cdot 4nw = nw$.

The partition function, Hamiltonian form

Each sector (n, w) carries a Fock space of left and right oscillators, contributing $1/\eta$ and $1/\bar\eta$ by (11.207). Hence

$$Z(\tau; R) = \frac{1}{|\eta(\tau)|^2}\sum_{n,w\in\mathbb{Z}} q^{\,p_L^2/4}\,\bar q^{\,p_R^2/4} = \frac{1}{|\eta(\tau)|^2}\sum_{n,w\in\mathbb{Z}} \exp\Big[-\pi\tau_2\Big(\frac{n^2}{R^2} + w^2R^2\Big) + 2\pi i\tau_1 nw\Big], \tag{11.230}$$

with $\tau = \tau_1 + i\tau_2$. (The second form: $q^a\bar q^b = e^{2\pi i\tau_1(a-b)}e^{-2\pi\tau_2(a+b)}$ with $a + b = \frac{1}{2}(n^2/R^2 + w^2R^2)$, $a - b = nw$.)

T-invariance is immediate: $\tau_1 \to \tau_1 + 1$ multiplies each term by $e^{2\pi i nw} = 1$, and $|\eta|$ is invariant. The spins $h - \bar h = nw$ are integers, as they must be.

The partition function, path-integral form

For S we Poisson-resum over the momentum n. At fixed w, the n-sum is a Gaussian sum:

$$\sum_{n\in\mathbb{Z}} e^{-\pi\tau_2 n^2/R^2 + 2\pi i\tau_1 wn} \overset{(1)}{=} \sum_{m\in\mathbb{Z}}\int_{\mathbb{R}} \mathrm{d}x\; e^{-\pi\tau_2 x^2/R^2 + 2\pi i x(\tau_1 w - m)} \overset{(2)}{=} \frac{R}{\sqrt{\tau_2}}\sum_{m\in\mathbb{Z}} e^{-\pi R^2(m-\tau_1 w)^2/\tau_2}. \tag{11.231}$$

(1) Poisson summation (11.214); we relabelled $k \to m$.
(2) The Gaussian integral $\int e^{-\pi a x^2 + 2\pi i b x}\mathrm{d}x = a^{-1/2}e^{-\pi b^2/a}$ with $a = \tau_2/R^2$ and $b = \tau_1 w - m$.

Multiply by the winding factor $e^{-\pi\tau_2 w^2R^2}$ and combine the exponents: $(m - \tau_1 w)^2 + \tau_2^2 w^2 = |m - w\tau|^2$. So

$$Z(\tau; R) = \frac{R}{\sqrt{\tau_2}\,|\eta(\tau)|^2}\sum_{m,w\in\mathbb{Z}} \exp\Big[-\frac{\pi R^2}{\tau_2}\,|m - w\tau|^2\Big]. \tag{11.232}$$

This form has a direct meaning. On the torus $\mathbb{C}/(\mathbb{Z} + \mathbb{Z}\tau)$, a classical solution can wind m times around the target as we go around one cycle and w times around the other: $\varphi = 2\pi R(\dots)$ linear in the coordinates. Its action is $\pi R^2|m - w\tau|^2/\tau_2$. The factor $R/(\sqrt{\tau_2}|\eta|^2)$ is the fluctuation determinant and the zero-mode volume. So (11.232) is "sum over classical windings × one-loop determinant", the path integral done exactly.

Worked example 11.83: Modular invariance of the compact boson

Under a general $m(\tau) = (a\tau + b)/(c\tau + d)$:

$$\frac{|m - w\,\tau'|^2}{\tau_2'} \overset{(1)}{=} \frac{|m(c\tau + d) - w(a\tau + b)|^2}{|c\tau + d|^2} \cdot \frac{|c\tau + d|^2}{\tau_2} \overset{(2)}{=} \frac{\big|(md - wb) - (wa - mc)\tau\big|^2}{\tau_2}.$$

(1) $m - w\tau' = \big(m(c\tau + d) - w(a\tau + b)\big)/(c\tau + d)$, and $\tau_2' = \tau_2/|c\tau + d|^2$ (Section 11.7.2).
(2) Collect the constant and the τ terms.

The map $(m, w) \mapsto (md - wb,\ wa - mc)$ is given by an integer matrix of determinant $ad - bc = 1$, so it is a bijection of $\mathbb{Z}^2$: the double sum is just reordered. Meanwhile the prefactor $\sqrt{\tau_2}|\eta|^2$ is invariant: under S, $\sqrt{\tau_2} \to \sqrt{\tau_2}/|\tau|$ and $|\eta|^2 \to |\tau||\eta|^2$ by (11.210); under T both are unchanged. So $Z(\tau; R)$ is modular invariant *for every radius* R. (Numerically, at $R = 1.37$ and $\tau = 0.23 + 0.91i$: $Z = 2.3375179\ldots$ for τ, $-1/\tau$ and $\tau + 1$.)

Worked example 11.84: The non-compact boson

As $R \to \infty$ only $w = 0$ survives in (11.230), and the momentum sum becomes an integral, $\sum_n \to R \int \mathrm{d}p$ with $p = n/R$:

$$Z(\tau; R) \overset{(1)}{=} \frac{R}{|\eta|^2} \int_{\mathbb{R}} \mathrm{d}p\ q^{p^2/4} \bar{q}^{p^2/4} \overset{(2)}{=} \frac{R}{|\eta|^2} \int_{\mathbb{R}} \mathrm{d}p\ e^{-\pi\tau_2 p^2} \overset{(3)}{=} \frac{R}{\sqrt{\tau_2}\,|\eta(\tau)|^2}.$$

(1) Riemann sum with spacing $1/R$.
(2) $q^{p^2/4}\bar{q}^{p^2/4} = e^{2\pi i(\tau - \bar{\tau})p^2/4} = e^{-\pi\tau_2 p^2}$.
(3) Gaussian integral.

The factor R is the volume of the target, an overall normalisation. The function $1/(\sqrt{\tau_2}|\eta|^2)$ is modular invariant, which is the $m = w = 0$ term of (11.232). This is the torus partition function of the free field of [Nash p. 303, (11.7)–(11.8)].

T-duality and the moduli space of $c = 1$ theories

Look at (11.229) again. Exchanging $n \leftrightarrow w$ and $R \leftrightarrow 1/R$ sends $p_L \to p_L$ and $p_R \to -p_R$. Since only p_R^2 enters, the spectrum and hence Z are unchanged:

$$Z(\tau; R) = Z(\tau; 1/R). \tag{11.233}$$

This is **T-duality**: a boson on a circle of radius R is indistinguishable from one on radius $1/R$, with momentum and winding exchanged. (Numerically: $Z(\tau; 1.37) = Z(\tau; 1/1.37)$ to 30 digits.) In the path-integral form it is far from obvious; in the Hamiltonian form it is a relabelling. Figure 11.43 shows the lattice of (p_L, p_R).

So the compact bosons form a one-parameter family of CFTs, all with $c = 1$, labelled by $R \in [1, \infty)$ (the half-line, after T-duality). This is our first example of a **moduli space of conformal field theories**. Some points on it are special (Figure 11.44).

Worked example 11.85: Two special radii

$R = 1$, the self-dual point. Look for states with $(h, \bar{h}) = (1, 0)$: holomorphic currents. $h = p_L^2/4 = 1$ and $p_R = 0$ need $n/R = wR$ and $n/R + wR = \pm 2$. At $R = 1$ this is $n = w = \pm 1$. So two extra currents $J^\pm = :e^{\pm 2i\varphi_L}:$ appear, alongside $J^3 = i\partial\varphi$. Together they form SU(2) at level 1, whose Sugawara central charge $c = 3k/(k + 2) = 1$ matches (Section 11.5.1).

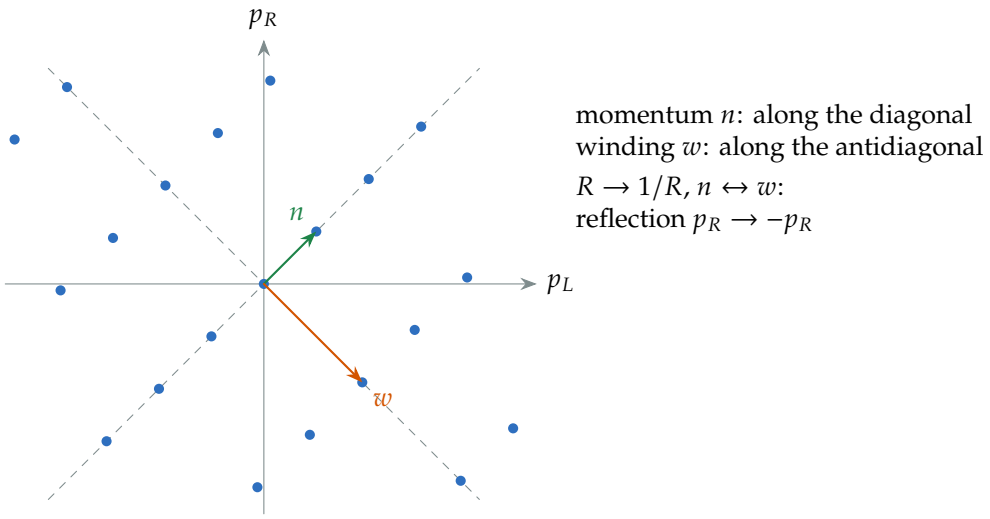


Figure 11.43: The charge lattice of the compact boson at $R = 1.37$: points $(p_L, p_R) = (n/R + wR,\ n/R - wR)$. Momentum steps (green) have length $\sqrt{2}/R$ along the diagonal, winding steps (orange) length $\sqrt{2}R$ along the antidiagonal. What to see: the lattice for radius $1/R$ is the mirror image in the p_L axis, and the spectrum depends only on p_L^2, p_R^2, so the two theories are identical. The lattice is even ($\frac{1}{4}(p_L^2 - p_R^2) = nw \in \mathbb{Z}$) and self-dual, which is the lattice version of T- and S-invariance.

T-duality is part of this enhanced SU(2) symmetry.

$R = \sqrt{2}$, the Dirac fermion. A complex fermion is two Majorana fermions with a common spin structure. Summing over the four spin structures as in (11.205), but now with $c = 1$,

$$Z_{\text{Dirac}} = \frac{|\theta_2|^2 + |\theta_3|^2 + |\theta_4|^2}{2|\eta|^2}. \tag{11.234}$$

This is modular invariant by exactly the argument for Ising (now with $|\theta_i|^2$ and $|\eta|^2$). Numerically it equals $Z(\tau;\sqrt{2})$ (and $Z(\tau;1/\sqrt{2})$): at $\tau = 0.1 + 1.3i$ both are $2.4914614645\ldots$. The two theories are the same: this is bosonisation on the torus, including all spin structures. The triple product (11.219) is the charge-resolved version of this equality.

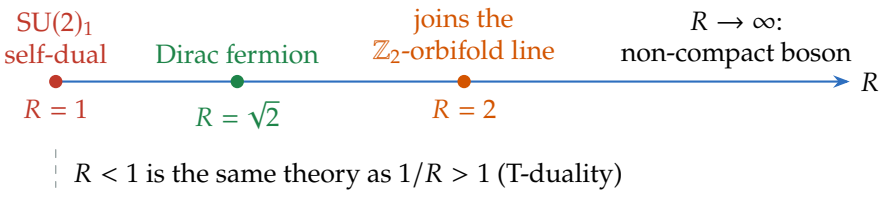


Figure 11.44: The moduli space of compact $c = 1$ bosons is the half-line $R \geq 1$. Every point is a modular-invariant CFT; the partition function varies continuously with R. Marked: the self-dual radius with enhanced $SU(2)_1$ symmetry, the Dirac point, and the point $R = 2$ where this line meets a second $c = 1$ line, the bosons on the orbifold $S^1/\mathbb{Z}_2$ (the circle at $R = 2$ equals the orbifold at $R = 1$). Two decoupled Ising models are not on this line: Z_{Ising}^2 is the orbifold at the Dirac radius $R = \sqrt{2}$. What to see: modular invariance alone does not fix a CFT; there can be continuous families.

Watch out: Two Majoranas are not always a Dirac fermion

The product Z_{Ising}^2 is modular invariant, but it is *not* Z_{Dirac}. In Z_{Ising}^2 the two Majorana fermions sum over their spin structures independently; in Z_{Dirac} they share one. At $\tau = 0.1 + 1.3i$, $Z_{\text{Ising}}^2 = 3.657\ldots$ while $Z_{\text{Dirac}} = 2.491\ldots$. Different sums over spin structures give different (but each consistent) theories. This is the simplest example of the freedom that the classification problem below has to control.

11.8.6 What modular invariance buys: Cardy's constraints

Nash remarks that modular invariance is "a considerable restriction on Z" and cites Cardy. Here are the two most famous consequences.

The high-temperature limit is controlled by the vacuum

Take a rectangular torus, $\tau = it$, and let $t \to 0$: the time circle is much shorter than the space circle, a high temperature. Directly, every state contributes and the sum is hard. After S, $\tilde\tau = i/t$ and $\tilde q = e^{-2\pi/t} \to 0$, so the sum is dominated by its lowest term, the vacuum:

$$Z(it) \overset{(1)}{=} Z(i/t) \overset{(2)}{=} \tilde q^{-c/24}\,\bar{\tilde q}^{-c/24}\big(1 + O(\tilde q^{h_{\min}})\big) \overset{(3)}{=} e^{\pi c/(6t)}\big(1 + O(e^{-2\pi h_{\min}/t})\big). \tag{11.235}$$

(1) S-invariance.
(2) In $\operatorname{Tr} \tilde q^{L_0 - c/24}\bar{\tilde q}^{\bar L_0 - c/24}$ the lowest state is the vacuum ($L_0 = \bar L_0 = 0$, unique); every other state is suppressed by at least $|\tilde q|^{h_{\min}}$, with $h_{\min}$ the lowest nonzero weight.
(3) $\tilde q^{-c/24} = e^{2\pi c/(24t)}$, squared.

For Ising, $c = \frac{1}{2}$ and $h_{\min} = \frac{1}{16}$:

t	$Z_{\text{Ising}}(it)$ (exact, from (11.224))	$e^{\pi c/6t}$	ratio
0.2	3.7754	3.7025	1.020
0.1	13.7135	13.7082	1.0004
0.05	187.9147	187.9146	1.0000002

Put units back: space circle L, inverse temperature $\beta = tL$ (with the velocity set to 1). Then $\ln Z = \pi c L/(6\beta)$, i.e. the free energy per unit length is

$$f(T) = -\frac{\pi c}{6}\,T^2, \qquad C(T) = \frac{\pi c}{3}\,T. \tag{11.236}$$

The low-temperature specific heat of a critical chain measures c. This is the finite-temperature twin of the Casimir energy (11.197), and modular invariance is what makes them twins: S swaps the circle of length L with the circle of length β.

Cardy's formula for the density of states

Write $Z(it) = \int \rho(\Delta) e^{-2\pi t(\Delta - c/12)}\mathrm{d}\Delta$, with $\Delta = L_0 + \bar L_0$ and ρ the density of states. For small t the integral is dominated by large Δ, and a saddle point must reproduce (11.235). Assume $\rho(\Delta) \approx e^{S(\Delta)}$ and maximise the exponent $S(\Delta) - 2\pi t\Delta$; then $\ln Z \approx S - 2\pi t\Delta$ at the saddle, a Legendre transform. Inverting it:

$$S(\Delta) \overset{(1)}{=} \min_t\Big(\frac{\pi c}{6t} + 2\pi t\Delta\Big) \overset{(2)}{=} \frac{\pi c}{6t_*} + 2\pi t_*\Delta\Big|_{t_* = \sqrt{c/12\Delta}} \overset{(3)}{=} 2\pi\sqrt{\frac{c\,\Delta}{3}}.$$

(1) Inverse Legendre transform of $\ln Z(t) = \pi c/6t$.
(2) Stationarity: $-\pi c/(6t^2) + 2\pi\Delta = 0$.
(3) At the stationary point both terms equal $\pi\sqrt{c\Delta/3}$.

This is **Cardy's formula**: $\rho(\Delta) \sim \exp\!\left(2\pi\sqrt{c\Delta/3}\right)$. The growth of the number of high-energy states is fixed by the central charge alone. (For the non-chiral string, the same argument gives the Hagedorn growth of the spectrum; for black holes it counts microstates in Strominger–Vafa.) The lesson is Nash's: a single symmetry of Z ties the ultraviolet (many states) to the infrared (the vacuum energy).

11.8.7 Classifying modular invariants: the ADE pattern

Modular invariance can be used as a classification tool. Fix a chiral algebra and its characters χ_i with S and T matrices. Look for all matrices M with

$$M_{ij} \in \mathbb{Z}_{\geq 0}, \qquad M_{00} = 1, \qquad [M, S] = [M, T] = 0. \tag{11.237}$$

Each solution is a candidate partition function $\sum M_{ij}\chi_i\bar{\chi}_j$ of a consistent theory. For Ising, the only solution is $M = I$: the lattice model is the *only* CFT built from $c = \frac{1}{2}$ Virasoro modules. (Try it: T-invariance forbids off-diagonal terms because the three T-phases differ, so M is diagonal; then $S^{\mathsf{T}}MS = M$ with $M = \mathrm{diag}(1, a, b)$ forces $a = b = 1$ from the first row.)

The richest test case is SU(2) at level k (the WZW models of Figure 11.28). There are $k+1$ primaries, labelled by $a = 2j \in \{0, 1, \ldots, k\}$, with

$$h_a = \frac{a(a+2)}{4(k+2)}, \qquad c = \frac{3k}{k+2}, \qquad S_{ab} = \sqrt{\frac{2}{k+2}}\,\sin\frac{\pi(a+1)(b+1)}{k+2}. \tag{11.238}$$

Cappelli, Itzykson and Zuber found *all* solutions of (11.237) (1987). Writing χ_λ with $\lambda = a+1 \in \{1, \ldots, k+1\}$ (the dimension of the SU(2) representation), the list is:

name	level	partition function
A_{k+1}	all k	$\sum_{\lambda=1}^{k+1} \lvert\chi_\lambda\rvert^2$
$D_{2\rho+2}$	$k = 4\rho$	$\sum_{\lambda\ \text{odd},\ \lambda<2\rho+1} \lvert\chi_\lambda + \chi_{k+2-\lambda}\rvert^2 + 2\lvert\chi_{2\rho+1}\rvert^2$
$D_{2\rho+1}$	$k = 4\rho - 2$	$\sum_{\lambda\ \text{odd}} \lvert\chi_\lambda\rvert^2 + \sum_{\lambda\ \text{even}} \chi_\lambda\bar{\chi}_{k+2-\lambda}$
E_6	$k = 10$	$\lvert\chi_1 + \chi_7\rvert^2 + \lvert\chi_4 + \chi_8\rvert^2 + \lvert\chi_5 + \chi_{11}\rvert^2$
E_7	$k = 16$	$\lvert\chi_1 + \chi_{17}\rvert^2 + \lvert\chi_5 + \chi_{13}\rvert^2 + \lvert\chi_7 + \chi_{11}\rvert^2 + \lvert\chi_9\rvert^2 + \left[(\chi_3 + \chi_{15})\bar{\chi}_9 + \text{c.c.}\right]$
E_8	$k = 28$	$\lvert\chi_1 + \chi_{11} + \chi_{19} + \chi_{29}\rvert^2 + \lvert\chi_7 + \chi_{13} + \chi_{17} + \chi_{23}\rvert^2$

(We checked each of D_4, D_5, E_6, E_7, E_8 numerically against (11.228).)

Worked example 11.86: Why the labels are Dynkin diagrams

Take E_6 at $k = 10$. The labels λ that appear on the diagonal of M (with multiplicity) are $1, 4, 5, 7, 8, 11$. These are exactly the **exponents** of the Lie algebra E_6, and $k + 2 = 12$ is its **Coxeter number**. The same holds for every entry:

invariant	$k+2$	diagonal labels in M	Lie algebra: Coxeter number, exponents
A_{k+1}	$k+2$	$1, 2, \ldots, k+1$	A_{k+1}: $h = k+2$, $1, \ldots, k+1$
D_4 ($k=4$)	6	$1, 3, 3, 5$	D_4: $h = 6$, $1, 3, 3, 5$
E_6 ($k=10$)	12	$1, 4, 5, 7, 8, 11$	E_6: $h = 12$, $1, 4, 5, 7, 8, 11$
E_8 ($k=28$)	30	$1, 7, 11, 13, 17, 19, 23, 29$	E_8: $h = 30$, same list

For D_4: $M = |\chi_1 + \chi_5|^2 + 2|\chi_3|^2$ has diagonal entries $M_{11} = M_{55} = 1$ and $M_{33} = 2$, i.e. the labels $1, 5, 3, 3$. Nobody has a fully satisfying explanation of why modular invariance "knows" the simply laced Dynkin diagrams; it is one of the correspondences Nash calls "intriguing". The same ADE list classifies finite subgroups of SU(2), simple singularities $\mathbb{C}^2/\Gamma$, and quivers of finite type.

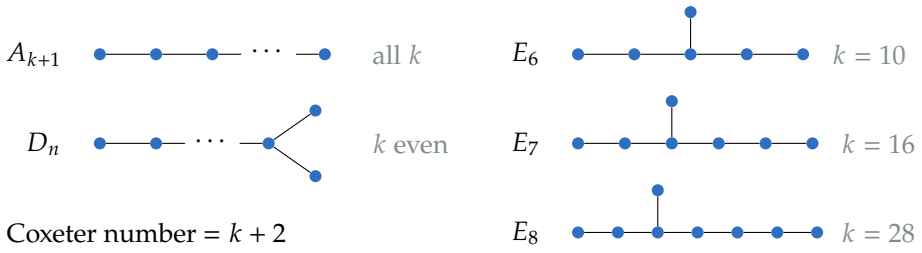


Figure 11.45: The simply laced Dynkin diagrams and the levels at which the corresponding $SU(2)_k$ modular invariants occur. What to see: the classification of modular-invariant partition functions, a purely analytic problem about q-series, reproduces the ADE list; the Virasoro minimal models are then classified by *pairs* (A_m, G) with Coxeter numbers $m+1$ and $m+2$, e.g. Ising = (A_2, A_3) and three-state Potts = (A_4, D_4).

Why pairs of Lie algebras, as Nash says? The minimal models can be built as cosets $SU(2)_k \times SU(2)_1/SU(2)_{k+1}$, so their invariants are products of an $SU(2)_{m-1}$ invariant and an $SU(2)_m$ invariant, one of each type. The Coxeter numbers of the pair are $m+1$ and $m+2$, the two integers in $c = 1 - 6/((m+1)(m+2))$. The Ising model, $m = 2$, is (A_2, A_3) and has only the diagonal invariant: that is the statement "$M = I$ is the only solution" above. The first model with a non-diagonal invariant is $m = 5$, $c = \frac{4}{5}$, where (A_4, D_4) is the three-state Potts model: a genuinely different lattice model from the tetracritical Ising model (A_4, A_5) at the same c.

Reading Nash (pp. 312–313, (11.52)–(11.59))

Nash defines η (11.52), relates the Ising $|\chi|^2$ to η at $\tau/2, \tau, 2\tau$ (11.53), defines modular forms of weight k via $f(\tau)\mathrm{d}\tau^k$ (11.54)–(11.55), states η^{24}'s law (11.56)–(11.57) and η's laws (11.58)–(11.59), and leaves the invariance of Z as "a routine matter". We derived (11.53) from the fermion products (correcting the factor 2 in its second line), proved (11.59) through $2\eta^3 = \theta_2\theta_3\theta_4$ and Poisson summation, and did the routine matter in (11.224). Note Nash's "weight 6" for η^{24} is weight 12 in modern language. His last paragraph (Cardy; Cappelli–Itzykson–Zuber and the A, D, E series) is our Sections 11.8.6 and 11.8.7.

Recap

- $\eta = q^{1/24}\prod(1-q^n)$: $\eta(\tau+1) = e^{i\pi/12}\eta$, $\eta(-1/\tau) = (-i\tau)^{1/2}\eta$. Modular forms change by $(c\tau+d)^{2k}$ (Nash's weight k); η^{24} has Nash-weight 6.
- Poisson summation turns a Gaussian lattice sum at τ into one at $-1/\tau$: $\theta_3 \to \theta_3$,

$\theta_2 \leftrightarrow \theta_4$. The triple product (bosonisation) gives θ's as products and $2\eta^3 = \theta_2\theta_3\theta_4$.

- $Z_{\text{Ising}} = (|\theta_2| + |\theta_3| + |\theta_4|)/2|\eta|$ is manifestly S- and T-invariant. The characters carry a unitary representation (S, T) of $\mathrm{PSL}(2,\mathbb{Z})$.
- Compact boson: $Z = |\eta|^{-2} \sum q^{p_L^2/4} \bar{q}^{p_R^2/4}$; Poisson resummation gives the winding (path-integral) form, which is manifestly invariant. $Z(R) = Z(1/R)$; a line of $c = 1$ theories.
- Cardy: $Z \approx e^{\pi c/6t}$ at high temperature; $\rho \sim e^{2\pi\sqrt{c\Delta/3}}$. CIZ: modular invariants of $\mathrm{SU}(2)_k$ follow the ADE classification.

11.9 Operator products and correlation functions

The torus taught us which representations a CFT contains. It did not tell us how the fields *interact*: what happens when two of them come close, and what the correlation functions are. Nash's §5 turns to this question. Its message is that conformal symmetry is so strong that a whole CFT is fixed by a small, finite amount of data:

the list of primary fields with their weights $(h_i, \bar{h}_i)$, the central charge c, and one number C_{ij}^{k} for each triple of primaries.

Everything else (every correlation function of every field, at any number of points) is then determined by symmetry and by a consistency condition, crossing symmetry. This section shows how, step by step:

1. a conformal family is a primary field plus its descendants, and correlators of descendants reduce to correlators of primaries;
2. two- and three-point functions of primaries are fixed up to constants;
3. the four-point function is fixed up to a function of one complex variable, the cross ratio;
4. the operator product expansion (OPE) of two primaries is fixed by the C_{ij}^{k};
5. inserting the OPE into the four-point function splits it into *conformal blocks*;
6. for the Ising model we compute everything explicitly: the blocks of $\langle\sigma\sigma\sigma\sigma\rangle$, the fusion $\sigma \times \sigma = \mathbf{1} + \varepsilon$, and the constant $C_{\sigma\sigma\varepsilon} = \tfrac{1}{2}$.

Figure 11.46 shows how the pieces feed into each other.

Conventions for this section

We use the holomorphic language of §§2–3. A primary field $\phi(z, \bar{z})$ of weights $(h, \bar{h})$ has the operator product with the energy–momentum tensor

$$T(z)\,\phi(w,\bar{w}) = \frac{h\,\phi(w,\bar{w})}{(z-w)^2} + \frac{\partial_w \phi(w,\bar{w})}{z-w} + \text{regular}, \tag{11.239}$$

Nash's (11.26). The Virasoro modes are $L_n = \oint \frac{\mathrm{d}z}{2\pi i}\, z^{n+1} T(z)$, and (11.239) is equivalent to the commutator

$$[L_n, \phi(z)] = z^n \big(z\partial_z + (n+1)h\big)\phi(z), \tag{11.240}$$

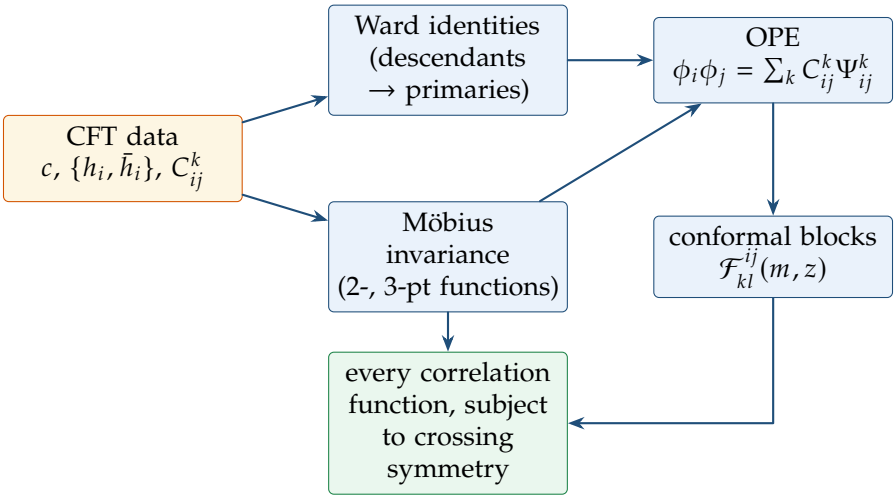


Figure 11.46: The logic of this section. What to see: the only inputs are the spectrum and the three-point constants; symmetry does the rest, and crossing symmetry is the one remaining condition that the data must satisfy.

where we suppress $\bar{z}$. The in-state of a primary is $|h\rangle = \phi(0)|0\rangle$ and its out-state is $\langle h| = \lim_{w\to\infty} w^{2h}\bar{w}^{2\bar{h}}\langle 0|\phi(w,\bar{w})$. Throughout, a formula written for the holomorphic side has an identical antiholomorphic partner, and we write only the first.

11.9.1 Conformal families and descendant fields

Nash starts with a theory that has finitely many primary fields $\phi_1, \dots, \phi_F$. Each generates a **conformal family** $[\phi_i]$: the primary together with everything the Virasoro raising operators make from it. A theory with finitely many families is a **rational conformal field theory** (RCFT). The Ising model is rational with $F = 3$; the non-compact free boson is not (it has a continuum of primaries V_p). The Hilbert space is Nash's

$$\mathcal{H} = \bigoplus_{i=1}^{F} H_i \otimes \bar{H}_i,$$

where $H_i = L(c, h_i)$ is the irreducible Virasoro module of Section 11.6.1 and $\bar{H}_i$ its antiholomorphic partner. (This is (11.200) with the diagonal multiplicities $m_{h,\bar{h}} = \delta_{h\bar{h}}$.)

Multi-index notation

A descendant state is a string of raising operators acting on $|h_i\rangle$. Nash packs the string into a multi-index. For $I = (i_1, \dots, i_s)$ with $i_1 \geq \cdots \geq i_s \geq 1$ he writes

$$L^I = L_{-i_1}\cdots L_{-i_s}, \qquad |I| = i_1 + \cdots + i_s, \qquad \bar{L}^J = \bar{L}_{-j_1}\cdots\bar{L}_{-j_t}, \tag{11.241}$$

Nash's (11.60). Then $L^I\bar{L}^J|h_i\rangle$ has weights $(h_i + |I|, \bar{h}_i + |J|)$. The multi-indices at level N are the partitions of N, so the table in Section 11.6.1 is also the list of multi-indices: $I = (1)$; $I = (2), (1,1)$; $I = (3), (2,1), (1,1,1)$; and so on.

From descendant states to descendant fields

In radial quantisation every state is created by a field at the origin (the state–field correspondence). So each descendant state must be created by a **descendant field**. Nash denotes it

$$\Phi_i^{I,J}(z,\bar z) = L^I(z)\,\bar L^J(\bar z)\,\phi_i(z,\bar z), \tag{11.242}$$

his (11.61). What does $L_{-n}(z)$ acting on a field at z mean? It is the same contour integral as L_{-n}, but centred on z instead of on the origin.

Working definition 11.30 (descendant field)

For a field $A(w)$ and $n \geq 1$,

$$(L_{-n}A)(w) := \oint_w \frac{\mathrm{d}z}{2\pi i}\,(z-w)^{1-n}\,T(z)A(w),$$

where the small contour encircles w and no other insertion. *How to compute it:* write the OPE $T(z)A(w)$ [for a primary A of weight h it begins $\frac{hA(w)}{(z-w)^2} + \frac{\partial A(w)}{z-w}$, followed by nonsingular terms] as a Laurent series in $z-w$; then $(L_{-n}A)(w)$ is the coefficient of $(z-w)^{n-2}$. Iterating gives $L^I(w)\phi(w)$ for any multi-index.

At $w=0$ the definition is literally L_{-n} acting on the state:

$$(L_{-n}A)(0)|0\rangle = \oint \frac{\mathrm{d}z}{2\pi i}\,z^{1-n}T(z)A(0)|0\rangle = L_{-n}A(0)|0\rangle,$$

since $z^{1-n} = z^{(-n)+1}$ is the weight in the definition of L_{-n}.

Worked example 11.87: The first descendants of a primary and of the identity

Read off coefficients from (11.239):

$$(L_{-1}\phi)(w) \overset{(1)}{=} \oint_w \frac{\mathrm{d}z}{2\pi i}\Big(\frac{h\phi(w)}{(z-w)^2} + \frac{\partial\phi(w)}{z-w} + \dots\Big) \overset{(2)}{=} \partial\phi(w),$$
$$(L_0\phi)(w) \overset{(3)}{=} \oint_w \frac{\mathrm{d}z}{2\pi i}\,(z-w)\,T(z)\phi(w) = h\,\phi(w),$$
$$(L_{-2}\phi)(w) \overset{(4)}{=} \oint_w \frac{\mathrm{d}z}{2\pi i}\,\frac{T(z)\phi(w)}{z-w} \overset{(5)}{=} :T\phi:(w).$$

(1) The definition with $n=1$, so $(z-w)^{1-n}=1$.
(2) Residue theorem: only the simple pole contributes.
(3) $n=0$: now the double pole becomes a simple pole with residue $h\phi$.
(4) $n=2$.
(5) The integrand's residue is the $(z-w)^0$ term of the OPE, which is by definition the normal-ordered product.

So L_{-1} is differentiation: the family $[\phi]$ contains all derivatives $\partial^k\phi$. For the identity field $\mathbf{1}$ the OPE $T(z)\mathbf{1} = T(z)$ is regular, so $L_{-1}\mathbf{1}=0$ (the vacuum is translation invariant) and $(L_{-2}\mathbf{1})(w) = T(w)$. *The energy–momentum tensor is a descendant of the identity*, at level 2. This is why T is not a primary field (its transformation law has the Schwarzian term) and yet belongs to the theory with no new label: it lives in the family $[\mathbf{1}]$.

Figure 11.47 draws a conformal family as a tower of fields, next to the contour that defines a descendant.

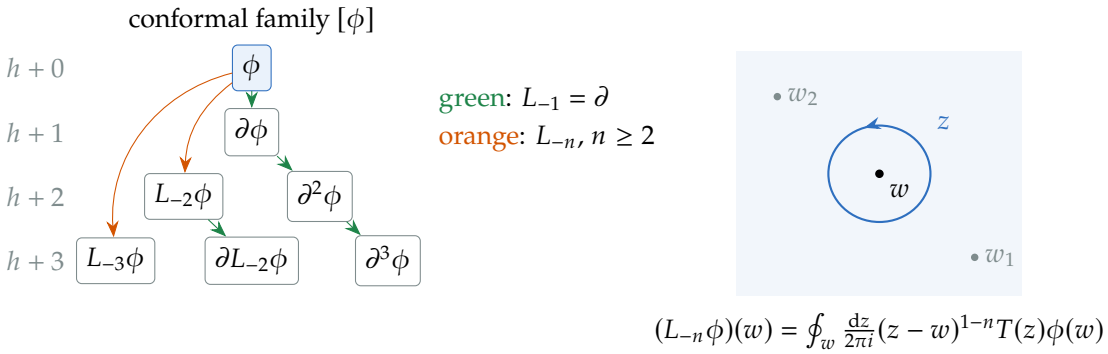


Figure 11.47: Left: the conformal family of a primary ϕ up to level 3, as fields (compare the states of Figure 11.35). Each level is a weight $h+N$; the green arrows are derivatives. Right: the contour that defines a descendant field encloses only w. What to see: a descendant is the primary "dressed" by a loop of T; the other insertions w_1, w_2 lie outside the loop, which is what lets us move the loop onto them.

11.9.2 Correlators of descendants reduce to primaries

Nash's most general correlation function (11.62) is

$$\left\langle \Phi_{i_1}^{I_1,J_1}(z_1,\bar z_1)\cdots\Phi_{i_n}^{I_n,J_n}(z_n,\bar z_n)\right\rangle, \tag{11.243}$$

and he asserts that the Ward identities express it through the correlators $\langle\phi_{i_1}\cdots\phi_{i_n}\rangle$ of primaries alone. Here is the mechanism, in one line and then in detail: *a descendant is a contour of T around a primary; deform the contour until it wraps the other insertions, and there the $T\phi$ OPE turns it into derivatives.*

Key idea: Descendant Ward identity

For primaries ϕ_i of weights h_i at points w_i, and any field ϕ at w (primary or not),

$$\left\langle (L_{-n}\phi)(w)\,\phi_1(w_1)\cdots\phi_N(w_N)\right\rangle = \mathcal{L}_{-n}\left\langle \phi(w)\,\phi_1(w_1)\cdots\phi_N(w_N)\right\rangle, \tag{11.244}$$

with the differential operator

$$\mathcal{L}_{-n} = \sum_{i=1}^{N}\left[\frac{(n-1)h_i}{(w_i-w)^n} - \frac{1}{(w_i-w)^{n-1}}\frac{\partial}{\partial w_i}\right], \qquad n\geq 1. \tag{11.245}$$

Let $X=\phi_1(w_1)\cdots\phi_N(w_N)$. As a function of z, the correlator $\langle T(z)\phi(w)X\rangle$ is meromorphic on the sphere, with poles only at w and the w_i, and it falls off like z^{-4} at infinity (because T is regular at $z=\infty$ in the coordinate $u=1/z$, where $T(z)=u^4T'(u)$). Then

$$\begin{aligned}\left\langle (L_{-n}\phi)(w)X\right\rangle &\overset{(1)}{=} \oint_w \frac{dz}{2\pi i}(z-w)^{1-n}\left\langle T(z)\phi(w)X\right\rangle \\ &\overset{(2)}{=} -\sum_{i=1}^{N}\oint_{w_i}\frac{dz}{2\pi i}(z-w)^{1-n}\left\langle T(z)\phi(w)X\right\rangle \\ &\overset{(3)}{=} -\sum_{i=1}^{N}\oint_{w_i}\frac{dz}{2\pi i}(z-w)^{1-n}\left[\frac{h_i}{(z-w_i)^2}+\frac{1}{z-w_i}\frac{\partial}{\partial w_i}\right]\left\langle\phi(w)X\right\rangle\end{aligned}$$

$$\overset{(4)}{=} -\sum_{i=1}^{N}\Big[(1-n)(w_i-w)^{-n}h_i + (w_i-w)^{1-n}\frac{\partial}{\partial w_i}\Big]\big\langle\phi(w)X\big\rangle \overset{(5)}{=} \mathcal{L}_{-n}\big\langle\phi(w)X\big\rangle. \tag{11.246}$$

(1) The definition of the descendant field.
(2) On the sphere the sum of all residues of a meromorphic 1-form is zero (Stokes' theorem on the sphere minus small discs). The 1-form $(z-w)^{1-n}\langle T\phi X\rangle\,\mathrm{d}z$ decays like $z^{-3-n}\mathrm{d}z$ at infinity, so there is no residue there for $n \geq 1$. Hence the contour around w equals minus the sum of contours around the w_i (Figure 11.48).
(3) Near w_i only the singular part of the OPE $T(z)\phi_i(w_i)$, (11.239), has a residue; the other fields are spectators.
(4) Residues: $\oint_{w_i}\frac{f(z)}{(z-w_i)^2} = f'(w_i)$ and $\oint_{w_i}\frac{f(z)}{z-w_i} = f(w_i)$, with $f(z) = (z-w)^{1-n}$.
(5) Rearranged, this is (11.245).

Since $\mathcal{L}_{-n}$ only involves the positions and weights of the *other* fields, we can iterate: a field $L_{-n_1}L_{-n_2}\phi$ gives $\mathcal{L}_{-n_1}\mathcal{L}_{-n_2}$ acting on $\langle\phi X\rangle$ (with care about the order: the inner operator is applied first, and the outer one also acts on the w-dependence through the $(w_i - w)$ factors). Descendants of the *other* fields are handled by the same argument, one at a time. That is Nash's claim: all of (11.243) follows from the primary correlators.

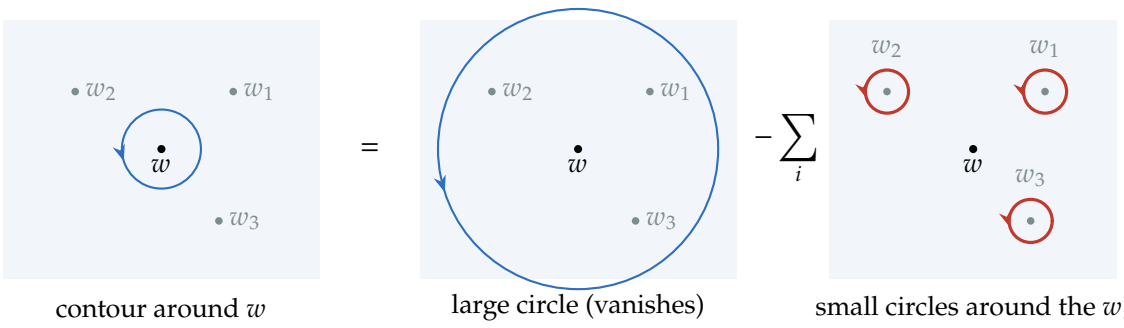


Figure 11.48: Contour deformation behind the Ward identity (11.246). The small loop around w is blown up into a large circle (which contributes nothing because $T \sim z^{-4}$ at infinity) minus loops around every other insertion (red). What to see: the action of T on one field is traded for its action on all the others, where the OPE (11.239) turns it into h_i's and derivatives.

Worked example 11.88: Two checks of the Ward operator

$n = 1$. (11.245) gives $\mathcal{L}_{-1} = -\sum_i \partial_{w_i}$. On the other hand $L_{-1}\phi = \partial_w\phi$. So the identity says $\partial_w\langle\phi X\rangle = -\sum_i \partial_{w_i}\langle\phi X\rangle$, i.e.

$$\Big(\partial_w + \sum_i \partial_{w_i}\Big)\langle\phi(w)X\rangle = 0,$$

which is translation invariance. Good.
$n = 2$ **on the identity.** Take $\phi = \mathbf{1}$, so $(L_{-2}\mathbf{1})(w) = T(w)$ and $\langle\mathbf{1}X\rangle = \langle X\rangle$:

$$\begin{aligned}\langle T(w)\,\phi_1(w_1)\cdots\phi_N(w_N)\rangle &\overset{(1)}{=} \sum_i\Big[\frac{h_i}{(w_i-w)^2} - \frac{1}{w_i-w}\partial_{w_i}\Big]\langle X\rangle \\ &\overset{(2)}{=} \sum_i\Big[\frac{h_i}{(w-w_i)^2} + \frac{1}{w-w_i}\partial_{w_i}\Big]\langle X\rangle.\end{aligned}$$

(1) (11.244) with $n = 2$.

(2) $(w_i - w)^2 = (w - w_i)^2$ and $-1/(w_i - w) = 1/(w - w_i)$.

This is the conformal Ward identity of §2 (Nash's (11.17)–(11.19) in correlator form): the singular behaviour of T near each insertion is exactly (11.239), and nothing else.

Watch out: Descendants of the identity are not zero

Since $\langle \mathbf{1} \rangle = 1$ and $\mathcal{L}_{-n}$ acts on the other fields, correlators like $\langle T(w)T(w') \rangle$ are not given by (11.244) with primaries only: T is not primary, and its OPE with itself has the extra $\frac{c/2}{(z-w)^4}$ term. So for descendants of $\mathbf{1}$ one must keep the central term; for example $\langle T(w)T(w') \rangle = \frac{c/2}{(w-w')^4}$. The reduction still works, but c now enters. This is how the central charge gets into every correlator.

11.9.3 Two- and three-point functions

Among all Ward identities, three are special. They come from the global conformal transformations, the Möbius maps generated by L_{-1}, L_0, L_1 (Section 11.5.1). Their statement is that correlators of primaries are *exactly* invariant, with no contour needed.

Working definition 11.31 (global Ward identities)

For primaries ϕ_i of weights h_i [fields for which $\phi_i \, \mathrm{d}z^{h_i} \mathrm{d}\bar{z}^{\bar{h}_i}$ is invariant under conformal maps] at points z_i, the function $G = \langle \phi_1(z_1) \cdots \phi_n(z_n) \rangle$ obeys

$$\sum_i \partial_i G = 0, \qquad \sum_i \big(z_i \partial_i + h_i\big) G = 0, \qquad \sum_i \big(z_i^2 \partial_i + 2h_i z_i\big) G = 0, \tag{11.247}$$

with $\partial_i = \partial/\partial z_i$ (and the same with bars). *How to use it:* three linear PDEs in n variables; they cut the number of free variables from n to $n - 3$ [for $n \geq 3$ a Möbius map can move z_1, z_2, z_3 to $0, 1, \infty$, so only the positions of the others remain free].

Where do they come from? The previous example gives $\langle T(z)X \rangle = \sum_i [h_i (z - z_i)^{-2} + (z - z_i)^{-1} \partial_i] \langle X \rangle$. Expand at large z:

$$\langle T(z)X \rangle \overset{(1)}{=} \frac{1}{z} \sum_i \partial_i \langle X \rangle + \frac{1}{z^2} \sum_i \big(z_i \partial_i + h_i\big) \langle X \rangle + \frac{1}{z^3} \sum_i \big(z_i^2 \partial_i + 2h_i z_i\big) \langle X \rangle + O(z^{-4}).$$

(1) $\frac{1}{z - z_i} = \frac{1}{z} + \frac{z_i}{z^2} + \frac{z_i^2}{z^3} + \cdots$ and $\frac{1}{(z-z_i)^2} = \frac{1}{z^2} + \frac{2z_i}{z^3} + \cdots$; collect powers.

Since $\langle T(z)X \rangle = O(z^{-4})$, the coefficients of z^{-1}, z^{-2}, z^{-3} vanish: that is (11.247). They are the statements that L_{-1}, L_0, L_1 annihilate both $|0\rangle$ and $\langle 0|$.

Worked example 11.89: The two-point function, Nash's (11.63) first line

Let $G(z_1, z_2) = \langle \phi_i(z_1) \phi_j(z_2) \rangle$. The first identity makes G a function of $z_{12} = z_1 - z_2$ only. The second:

$$0 \overset{(1)}{=} \big(z_1 \partial_1 + z_2 \partial_2 + h_i + h_j\big) G(z_{12}) \overset{(2)}{=} \big(z_{12} G'(z_{12}) + (h_i + h_j) G\big) \quad \Longrightarrow \quad G \overset{(3)}{=} \frac{C_{ij}}{z_{12}^{\,h_i + h_j}}.$$

(1) The L_0 identity.
(2) Chain rule: $\partial_1 G = G'$, $\partial_2 G = -G'$.
(3) The solution of $xG' = -aG$ is $G = Cx^{-a}$.

The third identity now gives a condition on the weights:

$$\big(z_1^2\partial_1 + z_2^2\partial_2 + 2h_i z_1 + 2h_j z_2\big)G \overset{(1)}{=} \Big[-(h_i+h_j)\frac{z_1^2 - z_2^2}{z_{12}} + 2h_i z_1 + 2h_j z_2\Big]G$$
$$\overset{(2)}{=} \big[-(h_i+h_j)(z_1+z_2) + 2h_i z_1 + 2h_j z_2\big]G = (h_i - h_j)\, z_{12}\, G.$$

(1) $\partial_1 G = -(h_i+h_j)G/z_{12}$ and $\partial_2 G = +(h_i+h_j)G/z_{12}$.
(2) $z_1^2 - z_2^2 = (z_1+z_2)z_{12}$.

So $C_{ij} = 0$ unless $h_i = h_j$ (and $\bar h_i = \bar h_j$). Including the antiholomorphic side,

$$\langle \phi_i(z_1,\bar z_1)\phi_j(z_2,\bar z_2)\rangle = \frac{C_{ij}}{z_{12}^{\,h_i+h_j}\,\bar z_{12}^{\,\bar h_i+\bar h_j}} = \frac{C_{ij}}{z_{12}^{\,2h_i}\,\bar z_{12}^{\,2\bar h_i}},$$

which is Nash's formula. One usually chooses a basis of primaries with $C_{ij} = \delta_{ij}$.

Worked example 11.90: The three-point function, Nash's (11.63) second line

Try $G = C\, z_{12}^a z_{13}^b z_{23}^d$; it already obeys the L_{-1} identity. With $\partial_1 = \frac{a}{z_{12}} + \frac{b}{z_{13}}$, $\partial_2 = -\frac{a}{z_{12}} + \frac{d}{z_{23}}$, $\partial_3 = -\frac{b}{z_{13}} - \frac{d}{z_{23}}$ (acting on $\log G$):

$$\frac{1}{G}\sum_i z_i^2\partial_i G \overset{(1)}{=} a\frac{z_1^2 - z_2^2}{z_{12}} + b\frac{z_1^2 - z_3^2}{z_{13}} + d\frac{z_2^2 - z_3^2}{z_{23}} \overset{(2)}{=} a(z_1+z_2) + b(z_1+z_3) + d(z_2+z_3).$$

(1) Collect the terms with the same z_{ij}.
(2) Difference of squares, as before.

Adding $2h_1z_1 + 2h_2z_2 + 2h_3z_3$ and demanding that the coefficient of each z_i vanish:

$$a+b = -2h_1, \qquad a+d = -2h_2, \qquad b+d = -2h_3$$
$$\Longrightarrow \quad a = -(h_1+h_2-h_3),\ \ b = -(h_1+h_3-h_2),\ \ d = -(h_2+h_3-h_1).$$

(The L_0 identity, $a+b+d = -(h_1+h_2+h_3)$, is then automatic.) So

$$\langle \phi_i(z_1)\phi_j(z_2)\phi_k(z_3)\rangle = \frac{C_{ijk}}{z_{12}^{\,h_i+h_j-h_k}\, z_{13}^{\,h_i+h_k-h_j}\, z_{23}^{\,h_j+h_k-h_i}} \times (\text{same with bars}), \tag{11.248}$$

Nash's (11.63). *One number C_{ijk} per triple* is all the freedom.

Worked example 11.91: Checking (11.248) on vertex operators

For the free boson, Wick's theorem (the Coulomb-gas formula of §2) gives, for $\alpha_1+\alpha_2+\alpha_3 = 0$, $\langle V_{\alpha_1}(z_1)V_{\alpha_2}(z_2)V_{\alpha_3}(z_3)\rangle = \prod_{i<j}|z_{ij}|^{\alpha_i\alpha_j}$. With $h_i = \alpha_i^2/4$, the predicted exponent of z_{12} is

$$-(h_1+h_2-h_3) \overset{(1)}{=} -\tfrac14\big(\alpha_1^2+\alpha_2^2-(\alpha_1+\alpha_2)^2\big) = \tfrac12\alpha_1\alpha_2,$$

(1) $\alpha_3 = -(\alpha_1+\alpha_2)$.

and $|z_{12}|^{\alpha_1\alpha_2} = z_{12}^{\alpha_1\alpha_2/2}\bar z_{12}^{\alpha_1\alpha_2/2}$ has exactly this exponent. Similarly for z_{13}, z_{23}. Here $C_{123} = 1$ if the charges add to zero and 0 otherwise.

Physical picture: Why three points and not four

A Möbius map has three complex parameters and can send any three distinct points to any other three. So there is nothing Möbius invariant to build from three points: the three-point function has no room for an arbitrary function. With four points there is exactly one invariant left over, the cross ratio. With n points there are $n-3$.

11.9.4 Four points and the anharmonic ratio

Working definition 11.32 (cross ratio)

For four distinct points of the Riemann sphere [$\mathbb{C}$ with the one point ∞ added], the **anharmonic (cross) ratio** is

$$z_{12,34} = \frac{(z_1 - z_2)(z_3 - z_4)}{(z_1 - z_3)(z_2 - z_4)}, \tag{11.249}$$

Nash's (11.64), and similarly $\bar{z}_{12,34}$. *Test:* it is unchanged by every Möbius map [$z \mapsto \frac{az+b}{cz+d}$ with $ad - bc = 1$, applied to all four points at once], and it takes every value in $\mathbb{C} \setminus \{0, 1\}$ as the points vary.

Worked example 11.92: The cross ratio is Möbius invariant

It suffices to check the generators: translations $z \mapsto z + b$ and scalings $z \mapsto az$ obviously leave (11.249) alone. For the inversion $z \mapsto -1/z$,

$$z_i - z_j \;\mapsto\; -\frac{1}{z_i} + \frac{1}{z_j} \overset{(1)}{=} \frac{z_i - z_j}{z_i z_j}, \qquad \text{so} \qquad z_{12,34} \;\mapsto\; \frac{\frac{z_{12}}{z_1 z_2}\,\frac{z_{34}}{z_3 z_4}}{\frac{z_{13}}{z_1 z_3}\,\frac{z_{24}}{z_2 z_4}} \overset{(2)}{=} z_{12,34}.$$

(1) Common denominator.

(2) Numerator and denominator both carry $1/(z_1 z_2 z_3 z_4)$.

Every Möbius map is a composition of these (Section 11.5.1).

Now use the invariance to move three of the points to $\infty, 1, 0$. The Möbius map $f(\zeta) = \frac{(\zeta - z_4)(z_2 - z_1)}{(\zeta - z_1)(z_2 - z_4)}$ sends $z_1 \mapsto \infty$, $z_2 \mapsto 1$, $z_4 \mapsto 0$, and $z_3 \mapsto z_{12,34}$. (Check: $f(z_3) = \frac{z_{34} z_{21}}{z_{31} z_{24}} = \frac{z_{12} z_{34}}{z_{13} z_{24}}$.) This is why Nash places the four fields at $(\infty, 1, z, 0)$: the one remaining position z *is* the cross ratio. Figure 11.49 shows it.

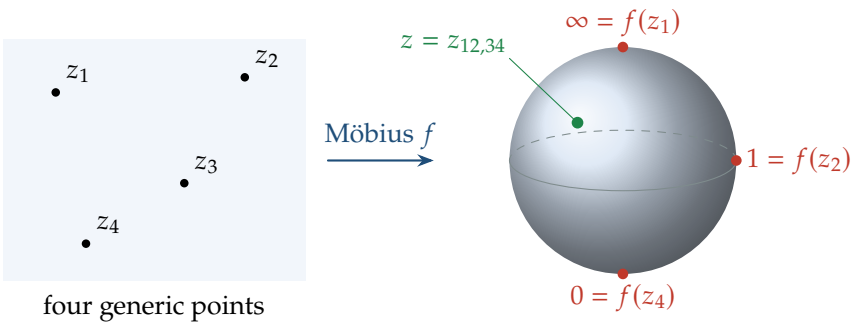


Figure 11.49: Using the three complex parameters of PSL(2, $\mathbb{C}$) to fix three of four points at $\infty, 1, 0$. What to see: the one point that remains free (green) is exactly the cross ratio, so a four-point function is a function of one complex variable z and its conjugate.

Key idea: The four-point function

Global conformal invariance fixes the four-point function of primaries up to one function of the cross ratio:

$$\langle\phi_1(z_1)\phi_2(z_2)\phi_3(z_3)\phi_4(z_4)\rangle = \prod_{i<j} z_{ij}^{h/3-h_i-h_j}\,\bar{z}_{ij}^{\bar{h}/3-\bar{h}_i-\bar{h}_j}\,Y\big(z_{12,34},\bar{z}_{12,34}\big), \tag{11.250}$$

with $h=\sum_i h_i$. Nash's arbitrary function F is Y. It is not determined by symmetry alone.

Why this prefactor? Any product $\prod z_{ij}^{\gamma_{ij}}$ with $\sum_{j\neq i}\gamma_{ij}=-2h_i$ for each i solves (11.247) (the same computation as for three points, with each z_i's coefficient giving one condition). The choice $\gamma_{ij}=h/3-h_i-h_j$ is symmetric and works: $\sum_{j\neq i}\gamma_{ij}=3\cdot\frac{h}{3}-3h_i-(h-h_i)=-2h_i$. Any other choice differs by a power of the cross ratio, which can be absorbed into Y. So "$n\geq 4$ needs further information": symmetry has used up its three parameters, and what remains is dynamics.

Worked example 11.93: A four-point function with no free function at all

For vertex operators with charges $(\alpha,-\alpha,\alpha,-\alpha)$ at $(\infty,1,z,0)$, the Coulomb-gas formula gives, after the factor $|z_1|^{4h_1}$ that defines the out-state has removed the dependence on z_1,

$$G(z,\bar{z})=\big\langle V_\alpha(\infty)V_{-\alpha}(1)V_\alpha(z,\bar{z})V_{-\alpha}(0)\big\rangle \overset{(1)}{=} |1-z|^{-\alpha^2}\,|1-0|^{\alpha^2}\,|z-0|^{-\alpha^2}=\big|z(1-z)\big|^{-\alpha^2}.$$

(1) Only the pairs not involving $z_1=\infty$ survive: $\alpha_2\alpha_3=-\alpha^2$, $\alpha_2\alpha_4=\alpha^2$, $\alpha_3\alpha_4=-\alpha^2$.

It factorises as $|z^{-\alpha^2/2}(1-z)^{-\alpha^2/2}|^2$: a single holomorphic function times its conjugate. As $z\to 0$ it behaves as $z^{-\alpha^2/2}=z^{h_p-h_3-h_4}$ with $h_p=0$: the two vertex operators fuse into the identity (and its descendants, from expanding $(1-z)^{-\alpha^2/2}$). The symmetry $G(z)=G(1-z)$ is the crossing symmetry discussed next.

Crossing symmetry and associativity

The four-point function of four *identical* fields must not care how we label the points. Permuting (z_1,z_2,z_3,z_4) changes the cross ratio by one of six maps,

$$z\;\mapsto\; z,\quad 1-z,\quad \frac{1}{z},\quad \frac{1}{1-z},\quad \frac{z-1}{z},\quad \frac{z}{z-1}, \tag{11.251}$$

the **anharmonic group** ($\cong S_3$). For example, swapping z_2 and z_4 sends $\frac{z_{12}z_{34}}{z_{13}z_{24}}\mapsto\frac{z_{14}z_{32}}{z_{13}z_{42}}$, and $1-z_{12,34}=\frac{z_{13}z_{24}-z_{12}z_{34}}{z_{13}z_{24}}=\frac{z_{14}z_{23}}{z_{13}z_{24}}$ by the identity $z_{13}z_{24}=z_{12}z_{34}+z_{14}z_{23}$. So $z\mapsto 1-z$. Invariance of the correlator under these maps (with the prefactors adjusted) is **crossing symmetry**.

Nash, following Belavin, Polyakov and Zamolodchikov, notes that crossing symmetry is equivalent to *associativity of the operator algebra*. Here is why. We can evaluate $\langle\phi_1\phi_2\phi_3\phi_4\rangle$ by first fusing ϕ_3 with ϕ_4 (as $z\to 0$) and then pairing the result with $\phi_1\phi_2$; or by first fusing ϕ_2 with ϕ_3 (as $z\to 1$). These are the two bracketings $\phi_1\phi_2(\phi_3\phi_4)$ and $\phi_1(\phi_2\phi_3)\phi_4$ of an operator product. If the OPE is associative the two must agree wherever both converge; and this agreement, written out, is the crossing relation between the expansions around $z=0$ and $z=1$ (Figure 11.50).

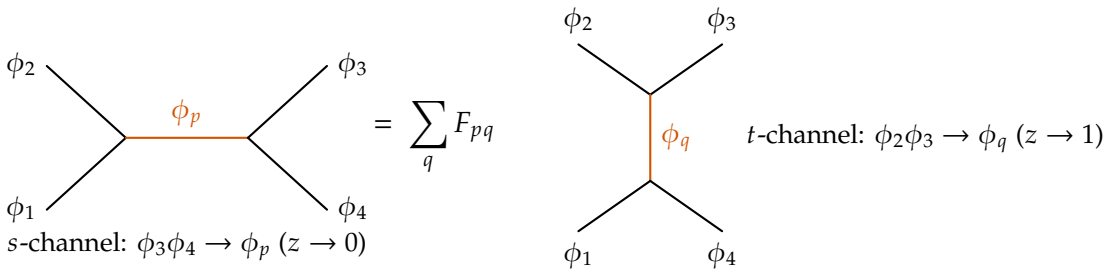


Figure 11.50: Crossing symmetry as associativity. The same four-point function can be expanded by fusing $\phi_3\phi_4$ first (left) or $\phi_2\phi_3$ first (right). Each tree is a sum over intermediate families (orange), and the two sums are related by a constant matrix F_{pq}. What to see: the diagram is the CFT version of $(ab)c = a(bc)$; demanding it is a strong constraint on the C_{ij}^{k}.

11.9.5 The operator product expansion of two primaries

To use associativity we need the OPE itself. Nash writes the general OPE of two primaries as

$$\phi_i(z,\bar z)\,\phi_j(0,0) = \sum_k \sum_{L,\bar L} \frac{C_{ij}^{k,L,\bar L}}{z^{\,h_i+h_j-h_k-|L|}\,\bar z^{\,\bar h_i+\bar h_j-\bar h_k-|\bar L|}}\,\Phi_k^{L,\bar L}(0,0), \tag{11.252}$$

his (11.65). The powers of z are forced by scaling: the left side has total weight $h_i + h_j$, a descendant $\Phi_k^{L,\bar L}$ has weight $h_k + |L|$, and z^a has weight $-a$. So $a = h_k + |L| - h_i - h_j$, the exponent shown.

The key claim is Nash's (11.66): the coefficient factorises,

$$C_{ij}^{k,L,\bar L} = C_{ij}^{k}\,\beta_{ij}^{kL}\,\bar\beta_{ij}^{k\bar L}, \tag{11.253}$$

where C_{ij}^{k} is the coefficient of the primary itself ($\beta^{k,\emptyset} = 1$) and the β's are *determined by conformal symmetry*, i.e. by c, h_i, h_j, h_k alone. With the β's Nash packages the family's contribution into

$$\Psi_{ij}^{k}(z,\bar z;0,0) = \sum_{L,\bar L} \frac{\beta_{ij}^{kL}\,\bar\beta_{ij}^{k\bar L}}{z^{\,h_i+h_j-h_k-|L|}\,\bar z^{\,\bar h_i+\bar h_j-\bar h_k-|\bar L|}}\,\Phi_k^{L,\bar L}(0,0),$$
$$\phi_i(z,\bar z)\phi_j(0,0) = \sum_k C_{ij}^{k}\,\Psi_{ij}^{k}(z,\bar z;0,0), \tag{11.254}$$

his (11.67)–(11.68). Ψ_{ij}^{k} is "the contribution of the family $[\phi_k]$ to the product $\phi_i\phi_j$".

Why the β's are fixed: the level-one coefficient

Apply the OPE to the vacuum: $\phi_j(0)|0\rangle = |h_j\rangle$, so (holomorphic part)

$$\phi_i(z)|h_j\rangle = \sum_k C_{ij}^{k}\,z^{\,h_k-h_i-h_j}\Big(|h_k\rangle + z\,\beta^{(1)}L_{-1}|h_k\rangle + z^2\big(\beta^{(2)}L_{-2} + \beta^{(1,1)}L_{-1}^2\big)|h_k\rangle + \cdots\Big). \tag{11.255}$$

Now act with L_1 on both sides and compare the coefficients of $z^{h_k-h_i-h_j+1}|h_k\rangle$.

$$\text{left: } L_1\phi_i(z)|h_j\rangle \overset{(1)}{=} [L_1,\phi_i(z)]|h_j\rangle \overset{(2)}{=} \big(z^2\partial_z + 2h_i z\big)\phi_i(z)|h_j\rangle$$
$$\overset{(3)}{=} \sum_k C_{ij}^k \big(h_k - h_i - h_j + 2h_i\big) z^{h_k-h_i-h_j+1}|h_k\rangle + \cdots,$$

$$\text{right: } \sum_k C_{ij}^k z^{h_k-h_i-h_j}\, z\,\beta^{(1)} L_1 L_{-1}|h_k\rangle \overset{(4)}{=} \sum_k C_{ij}^k\, 2h_k\,\beta^{(1)}\, z^{h_k-h_i-h_j+1}|h_k\rangle.$$

(1) $L_1|h_j\rangle = 0$, since $|h_j\rangle$ is a highest-weight state.
(2) (11.240) with $n = 1$.
(3) Insert (11.255) and differentiate the leading term: $(z^2\partial_z + 2h_i z)z^a = (a + 2h_i)z^{a+1}$.
(4) $L_1|h_k\rangle = 0$ kills the leading term, and $L_1L_{-1}|h_k\rangle = 2h_k|h_k\rangle$.

Hence

$$\beta_{ij}^{k,(1)} = \frac{h_k + h_i - h_j}{2h_k}. \tag{11.256}$$

A pure number built from weights. At level 2 the same trick with L_1^2 and L_2 gives two equations for $(\beta^{(2)}, \beta^{(1,1)})$, whose matrix is the Gram matrix $G^{(2)}$ of (11.178).

Worked example 11.94: Level 2, and what a null vector does to the OPE

Pair both sides of (11.255) with the level-2 bras $\langle h_k|L_2$ and $\langle h_k|L_1^2$. The right side gives $G^{(2)}\beta$, with $\beta = (\beta^{(2)}, \beta^{(1,1)})$. The left side is computed with (11.240): in general $\langle h_k|L_n\phi_i(z)|h_j\rangle = \langle h_k|[L_n,\phi_i(z)]|h_j\rangle = (h_k + nh_i - h_j)\,C_{ij}^k\, z^{h_k-h_i-h_j+n}$, because $z^n(z\partial_z + (n+1)h_i)z^a = (a + (n+1)h_i)z^{a+n}$ with $a = h_k - h_i - h_j$. Applying this twice for L_1^2:

$$G^{(2)}\begin{pmatrix}\beta^{(2)}\\ \beta^{(1,1)}\end{pmatrix} = \begin{pmatrix} h_k + 2h_i - h_j \\ (h_k + h_i - h_j)(h_k + h_i - h_j + 1)\end{pmatrix}.$$

Generic h_k: $G^{(2)}$ is invertible and β is unique. *Ising*, $\sigma\times\sigma \to \varepsilon$: $h_i = h_j = \frac{1}{16}$, $h_k = \frac{1}{2}$, $c = \frac{1}{2}$. Then

$$G^{(2)} \overset{(1)}{=} \begin{pmatrix}\frac{9}{4} & 3\\ 3 & 4\end{pmatrix}, \qquad \text{right side} \overset{(2)}{=} \begin{pmatrix}\frac{9}{16}\\ \frac{3}{4}\end{pmatrix} = \frac{3}{16}\begin{pmatrix}3\\4\end{pmatrix}.$$

(1) (11.178) at $h = \frac{1}{2}$, $c = \frac{1}{2}$: $4h + \frac{c}{2} = \frac{9}{4}$, $6h = 3$, $4h(2h+1) = 4$. Its determinant is $9 - 9 = 0$: the null vector $|\chi_\varepsilon\rangle$ of (11.182).
(2) $\frac{1}{2} + \frac{1}{8} - \frac{1}{16} = \frac{9}{16}$ and $\frac{1}{2}\cdot\frac{3}{2} = \frac{3}{4}$.

The matrix is singular, but the right side lies in its range (both rows of $G^{(2)}$ are multiples of $(3,4)$, and so is the right side). So a solution exists, for instance $\beta = (\frac{1}{4}, 0)$, and it is unique up to adding the null vector $(1, -\frac{3}{4})$, which is zero in the irreducible module anyway. This is not an accident: the OPE $\sigma\times\sigma$ is allowed to contain ε precisely because this consistency condition holds. If it failed, $C_{\sigma\sigma}^{\varepsilon}$ would have to vanish. *Null vectors restrict which families may appear in an OPE*: this is the seed of the fusion rules.

Watch out: The OPE is not a Taylor series

(11.252) is an exact operator statement inside correlators, convergent for $|z|$ smaller than the distance to the next insertion. It is not an asymptotic expansion: in a CFT the OPE converges. The non-integer powers $z^{h_k-h_i-h_j}$ make the correlator multivalued as a function of z for fixed k, even though the full correlator (summed over holomorphic and antiholomorphic

parts) is single valued.

11.9.6 Conformal blocks

Insert the OPE into the four-point function, with the points at Nash's positions $(\infty, 1, z, 0)$:

$$\begin{aligned}\left\langle \phi_i(\infty)\phi_j(1)\phi_k(z,\bar z)\phi_l(0)\right\rangle &\overset{(1)}{=} \sum_m C_{kl}^m \left\langle \phi_i(\infty)\phi_j(1)\,\Psi_{kl}^m(z,\bar z;0,0)\right\rangle \\ &\overset{(2)}{=} \sum_m C_{kl}^m \sum_{L,\bar L} \frac{\beta_{kl}^{mL}\bar\beta_{kl}^{m\bar L}}{z^{h_k+h_l-h_m-|L|}\bar z^{h_k+h_l-h_m-|\bar L|}} \left\langle \phi_i(\infty)\phi_j(1)\Phi_m^{L,\bar L}(0,0)\right\rangle \\ &\overset{(3)}{=} \sum_m C_{kl}^m\, C_{ijm}\, \mathcal{F}_{kl}^{ij}(m,z)\,\bar{\mathcal{F}}_{kl}^{ij}(m,\bar z). \end{aligned} \tag{11.257}$$

(1) The OPE (11.254) of $\phi_k(z)\phi_l(0)$, valid for $|z|<1$.
(2) The definition of Ψ.
(3) Each three-point function of descendants is, by the Ward identities of Section 11.9.2, a number times C_{ijm}; the sum factorises into a holomorphic sum times an antiholomorphic one.

The functions

$$\begin{aligned}\mathcal{F}_{kl}^{ij}(m,z) &= \sum_L \frac{\beta_{kl}^{mL}}{z^{\,h_k+h_l-h_m-|L|}} \frac{\left\langle \phi_i(\infty)\phi_j(1)L^L\phi_m(0)\right\rangle}{C_{ijm}} \\ &= z^{\,h_m-h_k-h_l}\left(1+\frac{(h_m+h_k-h_l)(h_m+h_j-h_i)}{2h_m}\,z+O(z^2)\right)\end{aligned} \tag{11.258}$$

are the **conformal blocks**: Nash's (11.69). (Nash normalises with $\sqrt{C_{ij}}$ factors and writes $C_{kl}^m C_{ij}^m$; with $C_{ij}=\delta_{ij}$ these are the same.) Each block depends only on the six numbers $c, h_i, h_j, h_k, h_l, h_m$, and is a power series in z times a leading power.

Worked example 11.95: The level-one coefficient of a block

The coefficient of z in (11.258) is $\beta^{(1)}\times\frac{\langle h_i|\phi_j(1)L_{-1}|h_m\rangle}{C_{ijm}}$. The first factor is (11.256) with $(i,j,k)\to(k,l,m)$. For the second, move L_{-1} to the left:

$$\begin{aligned}\langle h_i|\phi_j(1)L_{-1}|h_m\rangle &\overset{(1)}{=} \langle h_i|\big(L_{-1}\phi_j(1)-[L_{-1},\phi_j(1)]\big)|h_m\rangle \overset{(2)}{=} -\partial_w\langle h_i|\phi_j(w)|h_m\rangle\big|_{w=1} \\ &\overset{(3)}{=} -\partial_w\big(C_{ijm}\,w^{h_i-h_j-h_m}\big)\big|_{w=1} = (h_m+h_j-h_i)\,C_{ijm}.\end{aligned}$$

(1) Commutator.
(2) $\langle h_i|L_{-1}=(L_1|h_i\rangle)^\dagger=0$, and $[L_{-1},\phi_j(w)]=\partial_w\phi_j(w)$.
(3) $\langle h_i|\phi_j(w)|h_m\rangle$ is the three-point function with points at $\infty, w, 0$, which by (11.248) is $C_{ijm}w^{h_i-h_j-h_m}$.

Multiplying by $\beta^{(1)}=(h_m+h_k-h_l)/2h_m$ gives (11.258). For the identity family ($h_m=0$) this coefficient must be dropped: $L_{-1}|0\rangle=0$, and the block begins $z^{-h_k-h_l}(1+\frac{2h_kh_j}{c}z^2+\cdots)$ when $h_k=h_l$, $h_i=h_j$, from the level-2 state $L_{-2}|0\rangle$ with norm $c/2$.

More points; blocks as holomorphic objects

For $n \geq 4$ points the same procedure (fuse pairs repeatedly) writes the correlator as

$$\langle \phi_{i_1}(z_1,\bar z_1)\cdots\phi_{i_n}(z_n,\bar z_n)\rangle = \sum_{I,J} \mathcal{F}^I(z_1,\dots,z_{n-3})\, h_{IJ}\, \bar{\mathcal{F}}^J(\bar z_1,\dots,\bar z_{n-3}), \tag{11.259}$$

Nash's (11.70). The indices I, J label the choices of intermediate families in a chosen tree of fusions, and h_{IJ} is a Hermitian matrix made of the C_{ij}^k. After three points are fixed by Möbius invariance, the blocks depend on $n-3$ coordinates: the coordinates on the moduli space $\mathcal{M}_{0,n}$ of the n-punctured sphere.

The correlator is single valued on the configuration space (it is a physical quantity). The blocks are *not*: $\mathcal{F}(m,z) \sim z^{h_m - h_k - h_l}$ picks up a phase $e^{2\pi i(h_m-h_k-h_l)}$ when z circles 0. Circling $z=1$ mixes the blocks among themselves, by the crossing matrix of Figure 11.50. So the blocks form a vector-valued multivalued holomorphic function whose **monodromy** is a representation of the fundamental group of $\mathcal{M}_{0,n}$, and (11.259) is single valued because h_{IJ} is an invariant Hermitian form for that representation. Equivalently: the blocks are holomorphic sections of a flat vector bundle over $\mathcal{M}_{0,n}$, with the h_{IJ} giving a flat Hermitian metric. Nash will return to this picture for surfaces of any genus in Section 11.10.

11.9.7 The Ising four-spin function

Everything so far is general. We now compute one four-point function exactly, from nothing but the null vector (11.182) of the spin field. This is the classic BPZ computation, and it produces the fusion rule $\sigma \times \sigma = \mathbf{1} + \varepsilon$ and the constant $C_{\sigma\sigma\varepsilon}$.

Step 1: the null vector becomes a differential equation

In the irreducible module $|\chi_\sigma\rangle = (L_{-2} - \frac{4}{3}L_{-1}^2)|\sigma\rangle = 0$. By the state–field correspondence, the field $(L_{-2}\sigma)(w) - \frac{4}{3}\partial_w^2\sigma(w)$ is zero, so its correlators vanish. The descendant Ward identity (11.244) converts this into a PDE for $G = \langle\sigma(w)\sigma(w_1)\sigma(w_2)\sigma(w_3)\rangle$:

$$\Big[\tfrac{4}{3}\partial_w^2 - \sum_{i=1}^{3}\Big(\frac{h_\sigma}{(w_i-w)^2} - \frac{1}{w_i-w}\partial_{w_i}\Big)\Big]G = 0. \tag{11.260}$$

For a general level-2 null vector $(L_{-2} - aL_{-1}^2)\phi$ with $a = \frac{3}{2(2h+1)}$ the same holds with $\frac{4}{3} \to a$.

Step 2: reduce to an ODE in the cross ratio

Put $w = x$, $w_1 = 0$, $w_2 = 1$, $w_3 \to \infty$, and write $f(x) = \lim w_3^{2h_3} G$. We need $\partial_{w_1}G$ and $\partial_{w_2}G$ in terms of $\partial_x G$. For large w_3, $G \propto w_3^{-2h_3}$, so $\partial_{w_3}G \to 0$ and $w_3\partial_{w_3}G \to -2h_3 G$. The first two global Ward identities (11.247) then read

$$\partial_{w_1}f + \partial_{w_2}f \overset{(1)}{=} -\partial_x f,$$
$$0\cdot\partial_{w_1}f + 1\cdot\partial_{w_2}f \overset{(2)}{=} -x\partial_x f - (h + h_1 + h_2 - h_3)f,$$

(1) L_{-1} identity, with $\partial_{w_3} \to 0$.
(2) L_0 identity at $w_1 = 0$, $w_2 = 1$, with $w_3\partial_{w_3} \to -2h_3$; h is the weight of the field at x.

so $\partial_{w_2} f = -xf' - \kappa f$ and $\partial_{w_1} f = (x-1)f' + \kappa f$, with $\kappa = h + h_1 + h_2 - h_3$. Insert into the $\mathcal{L}_{-2}$ operator (the $i = 3$ term vanishes as $w_3 \to \infty$):

$$\begin{aligned}\mathcal{L}_{-2} f &\overset{(1)}{=} \frac{h_1}{x^2} f + \frac{1}{x}\partial_{w_1} f + \frac{h_2}{(1-x)^2} f - \frac{1}{1-x}\partial_{w_2} f \\ &\overset{(2)}{=} \left[\frac{h_1}{x^2} + \frac{h_2}{(1-x)^2} + \frac{\kappa}{x(1-x)}\right] f + \left[\frac{x-1}{x} + \frac{x}{1-x}\right] f' \\ &\overset{(3)}{=} \left[\frac{h_1}{x^2} + \frac{h_2}{(1-x)^2} + \frac{\kappa}{x(1-x)}\right] f + \frac{2x-1}{x(1-x)}\, f'.\end{aligned}$$

(1) (11.245) with $n = 2$: $w_1 - w = -x$, $w_2 - w = 1 - x$.
(2) Insert the two derivatives; $\frac{\kappa}{x} + \frac{\kappa}{1-x} = \frac{\kappa}{x(1-x)}$.
(3) $\frac{(x-1)(1-x)+x^2}{x(1-x)} = \frac{2x-1}{x(1-x)}$.

So the null-vector equation $a f'' = \mathcal{L}_{-2} f$ becomes the ordinary differential equation

$$a\, f''(x) + \frac{1-2x}{x(1-x)}\, f'(x) - \left[\frac{h_1}{x^2} + \frac{h_2}{(1-x)^2} + \frac{h + h_1 + h_2 - h_3}{x(1-x)}\right] f(x) = 0. \tag{11.261}$$

For four spins: $a = \frac{4}{3}$ and $h = h_1 = h_2 = h_3 = \frac{1}{16}$, so the last numerator is $\frac{1}{8}$.

Step 3: the indicial equation gives the fusion rule

Near $x = 0$ try $f \sim x^\lambda$. The most singular terms are of order $x^{\lambda-2}$:

$$\begin{aligned}a\lambda(\lambda-1) + \lambda - h_1 = 0 \quad &\xrightarrow{a=\frac{4}{3},\ h_1=\frac{1}{16}} \quad \tfrac{4}{3}\lambda^2 - \tfrac{1}{3}\lambda - \tfrac{1}{16} = 0 \\ &\overset{(1)}{=} 64\lambda^2 - 16\lambda - 3 = 0 \quad \Longrightarrow \quad \lambda \overset{(2)}{=} \tfrac{16 \pm 32}{128} = \left\{-\tfrac{1}{8},\ \tfrac{3}{8}\right\}.\end{aligned} \tag{11.262}$$

(1) Multiply by 48.
(2) Quadratic formula: $\sqrt{256 + 768} = 32$.

(The coefficient of f' is $\frac{1-2x}{x(1-x)} \approx \frac{1}{x}$ near 0, which gives the $+\lambda$.) Compare with the block (11.258), which starts as $x^{h_m - h_\sigma - h_\sigma} = x^{h_m - 1/8}$:

$$h_m - \tfrac{1}{8} = -\tfrac{1}{8} \;\Rightarrow\; h_m = 0, \qquad h_m - \tfrac{1}{8} = \tfrac{3}{8} \;\Rightarrow\; h_m = \tfrac{1}{2}.$$

Only the identity and the energy field can appear in $\sigma \times \sigma$. *A second-order ODE has exactly two solutions, and they are the two families*: $\sigma \times \sigma = \mathbf{1} + \varepsilon$. Notice that σ itself cannot appear (it would need $\lambda = -\frac{1}{16}$), which is the $\mathbb{Z}_2$ symmetry $\sigma \to -\sigma$ derived from scratch.

Step 4: solve the ODE

Remove the leading power by $f = (x(1-x))^{-1/8} u(x)$. The resulting equation is solved by $u = \sqrt{1 \pm \sqrt{1-x}}$ (one checks this by direct substitution; we verified it symbolically). The two solutions with the normalisation $x^{h_m - 1/8}(1 + O(x))$ are

$$\mathcal{F}_{\mathbf{1}}(x) = \frac{\sqrt{\frac{1}{2}\left(1 + \sqrt{1-x}\right)}}{(x(1-x))^{1/8}}, \qquad \mathcal{F}_{\varepsilon}(x) = \frac{\sqrt{2\left(1 - \sqrt{1-x}\right)}}{(x(1-x))^{1/8}}. \tag{11.263}$$

Worked example 11.96: The blocks have the predicted expansions

With $\sqrt{1-x} = 1 - \frac{x}{2} - \frac{x^2}{8} - \cdots$:

$$x^{1/8}\mathcal{F}_{\mathbf{1}} \overset{(1)}{=} (1-x)^{-1/8}\Big(1 - \tfrac{x}{4} - \tfrac{x^2}{16} + \cdots\Big)^{1/2} \overset{(2)}{=} 1 + \tfrac{x^2}{64} + O(x^3),$$
$$x^{-3/8}\mathcal{F}_{\varepsilon} \overset{(3)}{=} (1-x)^{-1/8}\Big(1 + \tfrac{x}{4} + \cdots\Big)^{1/2} \overset{(4)}{=} 1 + \tfrac{x}{4} + \tfrac{9x^2}{64} + O(x^3).$$

(1) $\frac{1}{2}(1+\sqrt{1-x}) = 1 - \frac{x}{4} - \frac{x^2}{16} - \cdots$.
(2) Multiply out the series; the $O(x)$ terms cancel.
(3) $2(1-\sqrt{1-x}) = x(1 + \frac{x}{4} + \cdots)$, and $x^{1/2}\cdot x^{-1/8} = x^{3/8}$.
(4) Multiply out.

Check against the general theory. For the identity block the $O(x)$ term is absent ($L_{-1}|0\rangle = 0$) and the x^2 coefficient is $2h_\sigma^2/c = 2\cdot\frac{1}{256}\cdot 2 = \frac{1}{64}$. For the ε block (11.258) predicts $\frac{(h_\varepsilon + h_\sigma - h_\sigma)^2}{2h_\varepsilon} = \frac{1/4}{1} = \frac{1}{4}$. Both agree.

Step 5: crossing symmetry fixes $C_{\sigma\sigma\varepsilon}$

The full correlator must be a single-valued combination

$$G(x,\bar{x}) = |\mathcal{F}_{\mathbf{1}}(x)|^2 + C_{\sigma\sigma\varepsilon}^2\,|\mathcal{F}_{\varepsilon}(x)|^2.$$

(No cross terms: the left and right families must match, since σ has $h = \bar{h}$ and the Ising model is diagonal.) Around $x = 0$ each term is single valued. Around $x = 1$ the blocks behave like $\sqrt{1 \pm \sqrt{1-x}}$, which mix when $\sqrt{1-x}$ changes sign. To find the mixing, parametrise $x = \sin^2 2\varphi$, so $\sqrt{1-x} = \cos 2\varphi$ and $\sqrt{x} = \sin 2\varphi$. Then (dropping the common prefactor $(x(1-x))^{-1/8}$, which is symmetric under $x \leftrightarrow 1-x$)

$$\sqrt{\tfrac{1}{2}(1+\cos 2\varphi)} = \cos\varphi, \qquad \sqrt{2(1-\cos 2\varphi)} = 2\sin\varphi,$$
$$\sqrt{\tfrac{1}{2}(1+\sin 2\varphi)} \overset{(1)}{=} \tfrac{1}{\sqrt{2}}(\cos\varphi + \sin\varphi), \qquad \sqrt{2(1-\sin 2\varphi)} \overset{(2)}{=} \sqrt{2}(\cos\varphi - \sin\varphi).$$

(1) $1 + \sin 2\varphi = (\cos\varphi + \sin\varphi)^2$.
(2) $1 - \sin 2\varphi = (\cos\varphi - \sin\varphi)^2$.

The first row is the blocks at x; the second row is the blocks at $1-x$. Solving the linear system,

$$\mathcal{F}_{\mathbf{1}}(x) = \tfrac{1}{\sqrt{2}}\mathcal{F}_{\mathbf{1}}(1-x) + \tfrac{1}{2\sqrt{2}}\mathcal{F}_{\varepsilon}(1-x), \qquad \mathcal{F}_{\varepsilon}(x) = \sqrt{2}\,\mathcal{F}_{\mathbf{1}}(1-x) - \tfrac{1}{\sqrt{2}}\mathcal{F}_{\varepsilon}(1-x). \tag{11.264}$$

Now rescale the ε block by the constant we are looking for: with $\tilde{\mathcal{F}}_\varepsilon = \frac{1}{2}\mathcal{F}_\varepsilon$ the relations become

$$\begin{pmatrix}\mathcal{F}_{\mathbf{1}}\\ \tilde{\mathcal{F}}_\varepsilon\end{pmatrix}(x) = \frac{1}{\sqrt{2}}\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix}\begin{pmatrix}\mathcal{F}_{\mathbf{1}}\\ \tilde{\mathcal{F}}_\varepsilon\end{pmatrix}(1-x),$$

an *orthogonal* matrix. So $|\mathcal{F}_{\mathbf{1}}|^2 + |\tilde{\mathcal{F}}_\varepsilon|^2$ is crossing symmetric, and any other ratio is not. Hence

$$C_{\sigma\sigma\varepsilon}^2 = \tfrac{1}{4}, \qquad C_{\sigma\sigma\varepsilon} = \tfrac{1}{2}. \tag{11.265}$$

The matrix $\frac{1}{\sqrt{2}}\left(\begin{smallmatrix}1 & 1\\ 1 & -1\end{smallmatrix}\right)$ is the **fusion** (or crossing) **matrix** F_{pq} of Figure 11.50 for four spins.

Worked example 11.97: The four-spin function in closed form

Add the two terms:

$$G(x,\bar{x}) \overset{(1)}{=} \frac{\left|1+\sqrt{1-x}\right|+\left|1-\sqrt{1-x}\right|}{2\,|x(1-x)|^{1/4}} \overset{(2)}{=} \frac{1}{|x(1-x)|^{1/4}}\sqrt{\frac{1+|x|+|1-x|}{2}}.$$

(1) $|\mathcal{F}_1|^2 = \frac{1}{2}|1+s|\,|x(1-x)|^{-1/4}$ and $\frac{1}{4}|\mathcal{F}_\varepsilon|^2 = \frac{1}{2}|1-s|\,|x(1-x)|^{-1/4}$, with $s=\sqrt{1-x}$.
(2) Square: $(|1+s|+|1-s|)^2 = |1+s|^2+|1-s|^2+2|1-s^2| = 2+2|s|^2+2|x|$, and $|s|^2=|1-x|$.

The last form is manifestly symmetric under $x \leftrightarrow 1-x$, and it is single valued (only absolute values appear). Under $x \to 1/x$, which swaps $\sigma(0)$ with $\sigma(\infty)$, it obeys $G(1/x) = |x|^{1/4}G(x)$, the correct covariance. Numerically, at $x = 0.3+0.4i$ both $G(x)$ and $G(1-x)$ equal $1.34766\ldots$. On the real segment $0<x<1$ the square root is 1 and $G = (x(1-x))^{-1/4}$ (Figure 11.51).

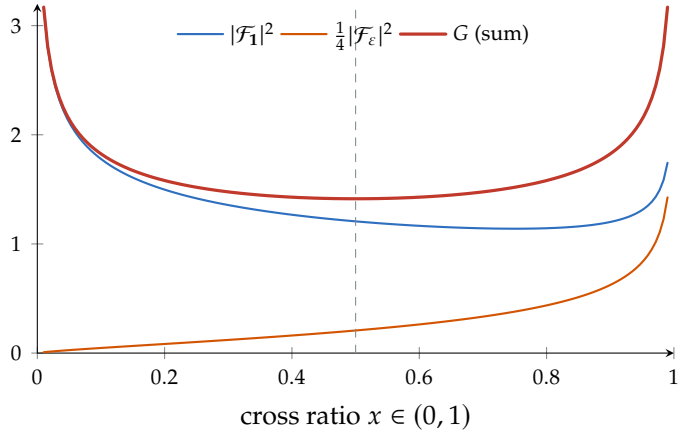


Figure 11.51: The Ising four-spin function on the real segment $0<x<1$. The identity block (blue) dominates near $x=0$, where $\sigma(x)$ approaches $\sigma(0)$ and the leading singularity $x^{-1/4} = |x|^{-2\Delta_\sigma}$ is the two-point function. The ε block (orange) grows towards $x=1$. What to see: neither block is symmetric about $x=\frac{1}{2}$, but their sum with the weight $C^2_{\sigma\sigma\varepsilon} = \frac{1}{4}$ (red) is: crossing symmetry fixes the OPE coefficient.

Physical picture: What the lattice sees

G is the correlation of four spins at criticality, in units where the two-point function is $|z|^{-1/4}$. The exponent $\frac{1}{4} = 2\Delta_\sigma$ is η of Section 11.5.3. The constant $C_{\sigma\sigma\varepsilon} = \frac{1}{2}$ is a universal amplitude: it says how strongly two nearby spins look like an energy density. On the lattice it is measured by comparing $\langle s_1 s_2\, E_3\rangle$ with $\langle s_1 s_2\rangle\langle E_3 E_4\rangle^{1/2}$. It is a prediction of the CFT with no free parameters.

Try it: The other Ising fusions from indicial equations

Use (11.262) in the general form $a\lambda(\lambda-1)+\lambda-h_1=0$, with $h_m = \lambda+h+h_1$. (a) For the σ null vector ($a=\frac{4}{3}$, $h=\frac{1}{16}$) fusing with ε ($h_1=\frac{1}{2}$), show $\lambda \in \{-\frac{1}{2},\frac{3}{4}\}$, hence $h_m \in \{\frac{1}{16},\frac{21}{16}\}$. (b) For the ε null vector ($a=\frac{3}{4}$, $h=\frac{1}{2}$) fusing with ε, show $h_m \in \{0,\frac{5}{3}\}$. *Answer:* $\frac{21}{16}$ and $\frac{5}{3}$ are not Ising weights, so their coefficients vanish, and $\sigma\times\varepsilon=\sigma$, $\varepsilon\times\varepsilon=\mathbf{1}$. (In (a), $8\lambda^2-2\lambda-3=0$; in (b), $3\lambda^2+\lambda-2=0$.)

Reading Nash (pp. 313–315, (11.60)–(11.70))

Nash introduces the families $[\phi_i]$, the notation L^I, $\Phi_i^{I,J}$ (11.60)–(11.61), the general correlator (11.62), and states that the Ward identities reduce it to primaries; we derived the reducing operators (11.245). His (11.63) is our (11.248) with its two-point companion; his (11.64) is the cross ratio. (11.65)–(11.68) are the OPE and its factorisation, for which we computed the level-one β (11.256) and the level-two system. (11.69) defines the blocks with the points at $(\infty, 1, z, 0)$, and (11.70) is the n-point form. The Ising computation of Section 11.9.7 is not in Nash; it is the standard illustration of all of these formulas at once, and it shows concretely what "crossing symmetry implies associativity" means.

Recap

- A descendant field is a loop of T around a primary; $L_{-1} = \partial$, $L_{-2}\mathbf{1} = T$. Contour deformation turns descendant insertions into the operators $\mathcal{L}_{-n}$, so all correlators reduce to those of primaries.
- Möbius invariance fixes $\langle\phi_i\phi_j\rangle = C_{ij}\delta_{h_ih_j}z_{12}^{-2h_i}\cdots$ and $\langle\phi_i\phi_j\phi_k\rangle$ up to C_{ijk}; the four-point function up to a function of the cross ratio.
- The OPE coefficients factorise as $C_{ij}^k\beta\bar{\beta}$, with $\beta^{(1)} = (h_k + h_i - h_j)/2h_k$; the four-point function is a sum over families of blocks $\mathcal{F}(m,z)\bar{\mathcal{F}}(m,\bar{z})$.
- Blocks are multivalued; the correlator is a monodromy-invariant Hermitian combination. Crossing symmetry = associativity of the OPE.
- Ising: the null vector gives a second-order ODE; its two exponents give $\sigma\times\sigma = \mathbf{1} + \varepsilon$; crossing fixes $C_{\sigma\sigma\varepsilon} = \frac{1}{2}$ and $G = |x(1-x)|^{-1/4}\sqrt{(1+|x|+|1-x|)/2}$.

11.10 Fusion rules and the Verlinde formula

The operator product expansion of Section 11.9 is a rich object: it contains the full set of constants C_{ij}^k and all the descendant coefficients β. Very often we do not need all of that. We only want to know *which* conformal families can appear when two fields collide, and how many independent ways there are for this to happen. That coarse information is a table of non-negative integers N_{ij}^k, the **fusion rules**. This section answers three questions about them.

1. *What are they, and what algebra do they obey?* They define a commutative, associative product $\phi_i \times \phi_j = \sum_k N_{ij}^k\phi_k$ on the set of families. Associativity is crossing symmetry again, now at the level of counting.
2. *How do they generalise to curved worldsheets?* On a surface of genus p with n punctures the conformal blocks form a vector space $V(\Sigma_{p,n})$. The fusion number N_{ij}^k is its dimension for the three-punctured sphere, and by cutting a surface into three-punctured spheres every $\dim V(\Sigma_{p,n})$ is computed from the N_{ij}^k alone.
3. *How do we compute them?* Verlinde's formula: the modular transformation $S : \tau \mapsto -1/\tau$ of the torus characters diagonalises the fusion rules, so the N_{ij}^k are read off from the matrix S of Section 11.8.4.

At the end we meet the constraints that Riemann surfaces of genus $p > 0$ impose on the singularities of fields: the Weierstrass gap theorem.

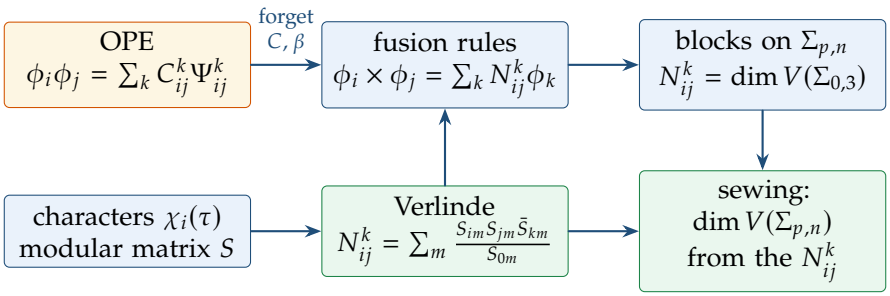


Figure 11.52: The logic of this section. What to see: the fusion rules sit in the middle. They come from the OPE by forgetting the continuous data, they are computed from the torus (the matrix S), and they determine the dimensions of the spaces of conformal blocks on every surface.

11.10.1 Which families appear in an operator product

Take two primaries ϕ_i and ϕ_j. Their OPE (11.255) is a sum over families $[\phi_k]$, each with coefficient C_{ij}^k. Nash's question (p. 316) is: *which k actually occur?* The answer is read off from three-point functions. Pair the OPE with ϕ_k at infinity:

$$\big\langle \phi_k(\infty)\,\phi_i(z)\,\phi_j(0)\big\rangle \overset{(1)}{=} \sum_l C_{ij}^l\, z^{h_l-h_i-h_j}\bar{z}^{\bar{h}_l-\bar{h}_i-\bar{h}_j}\big\langle \phi_k(\infty)\,\big(\phi_l(0)+\text{descendants}\big)\big\rangle$$
$$\overset{(2)}{=} C_{ij}^{k'}\,C_{k'k}\; z^{h_{k'}-h_i-h_j}\bar{z}^{\bar{h}_{k'}-\bar{h}_i-\bar{h}_j},$$

(1) Insert the OPE (11.255) of $\phi_i(z)\phi_j(0)$.

(2) The two-point function $\langle\phi_k\phi_l\rangle$ vanishes unless ϕ_l is the conjugate $\phi_{k'}$ of ϕ_k (the field with $C_{k'k}\neq 0$, Section 11.9.3); descendants of ϕ_l have zero two-point function with the primary ϕ_k (they are orthogonal to it, being at a different L_0 level).

So $[\phi_{k'}]$ occurs in $\phi_i\phi_j$ exactly when the three-point function $\langle\phi_k\phi_i\phi_j\rangle$ is non-zero. Nash's phrase is that a non-vanishing three-point function is "a fusing together of three representations of the conformal group".

Working definition 11.33 (fusion rules)

For each ordered pair of primaries, the **fusion coefficient** N_{ij}^k is the number of independent ways in which the family $[\phi_k]$ [ϕ_k together with all its descendants $L^I\bar{L}^J\phi_k$] appears in the operator product $\phi_i\phi_j$. Equivalently, it is the number of linearly independent chiral three-point couplings [holomorphic halves of three-point functions, depending on z_1, z_2, z_3 only] $\langle\phi_{k'}\phi_i\phi_j\rangle$ allowed by the chiral algebra [the holomorphic symmetry generators, whose modes act on each family] (the Virasoro algebra, or a larger algebra such as a Kac–Moody algebra). Verlinde packages them into the formal product (Nash (11.71))

$$\phi_i \times \phi_j = \sum_k N_{ij}^k\,\phi_k. \tag{11.266}$$

Operational test: N_{ij}^k is the number of conformal blocks [independent holomorphic functions into which a correlator splits, one for each family that can be exchanged] with $[\phi_k]$ in the intermediate channel. For a Virasoro minimal model it is 0 or 1 and can be read off from the null-vector differential equations [setting a null vector to zero inside a correlator gives a differential equation in the positions], as in (11.262).

Watch out: The fusion product is not a product of fields, nor a tensor product

$\phi_i \times \phi_j$ is a *formal* product. It forgets the positions z, the constants C_{ij}^k and the descendant coefficients: only the integers survive. It is also not the tensor product of Virasoro representations. If L_n acts on H_i and on H_j with central charge c, then $L_n \otimes 1 + 1 \otimes L_n$ acts on $H_i \otimes H_j$ with central charge $2c$, so $H_i \otimes H_j$ is not even a module of the same algebra. Fusion is the right replacement: it builds a module with the *same* c out of two modules, using the geometry of two points on the sphere.

The fusion algebra

The integers N_{ij}^k obey four rules. Each has a one-line reason.

1. **Commutativity:** $N_{ij}^k = N_{ji}^k$. The OPE $\phi_i(z)\phi_j(0)$ and $\phi_j(0)\phi_i(z)$ are the same product of commuting (bosonic, Euclidean) fields; the families that appear cannot depend on the order.
2. **Identity:** $N_{0j}^k = \delta_j^k$. The identity field $\phi_0 = \mathbf{1}$ multiplies nothing: $\mathbf{1} \times \phi_j = \phi_j$.
3. **Conjugation:** $N_{ij}^0 = \delta_{j,i'}$, where i' labels the conjugate family, the unique one with $\langle \phi_i \phi_{i'} \rangle \neq 0$. The identity appears in $\phi_i \phi_j$ only if the two-point function is non-zero.
4. **Associativity:** $\sum_m N_{ij}^m N_{mk}^l = \sum_m N_{jk}^m N_{im}^l$, i.e. $(\phi_i \times \phi_j) \times \phi_k = \phi_i \times (\phi_j \times \phi_k)$.

The last rule is crossing symmetry. Consider the four-point function $\langle \phi_{l'}(\infty)\phi_i(1)\phi_j(z)\phi_k(0)\rangle$. In the channel $z \to 0$ we first fuse $\phi_j\phi_k \to \phi_m$, then $\phi_i\phi_m \to \phi_l$: the number of blocks is $\sum_m N_{jk}^m N_{im}^l$. In the channel $z \to 1$ we first fuse $\phi_i\phi_j \to \phi_m$, then $\phi_m\phi_k \to \phi_l$: the number of blocks is $\sum_m N_{ij}^m N_{mk}^l$. The crossing matrix of Figure 11.50 is an invertible change of basis between the two sets of blocks, so the two counts must agree.

Working definition 11.34 (fusion matrices)

For each primary ϕ_i define the $F \times F$ matrix [F is the number of primary families] N_i with entries $(N_i)_j{}^k = N_{ij}^k$. It is "multiplication by ϕ_i", written as a matrix in the basis $\phi_0, \ldots, \phi_{F-1}$ of the fusion algebra [formal sums of the ϕ_k with integer coefficients, multiplied by (11.266)]. The four rules become

$$N_0 = \mathbb{1}, \qquad N_i N_j = N_j N_i, \qquad N_i N_j = \sum_k N_{ij}^k N_k. \tag{11.267}$$

So the matrices N_i *represent the fusion algebra*: this is the statement of Nash (p. 317) that "the N_{ij}^k form a representation of the fusion algebra 11.71".

Let us check the third relation in (11.267) from the associativity rule, entry by entry:

$$(N_i N_j)_a{}^b \overset{(1)}{=} \sum_m (N_i)_a{}^m (N_j)_m{}^b \overset{(2)}{=} \sum_m N_{ia}^m N_{jm}^b \overset{(3)}{=} \sum_m N_{ji}^m N_{ma}^b \overset{(4)}{=} \sum_k N_{ij}^k (N_k)_a{}^b.$$

(1) Matrix multiplication.
(2) The definition $(N_i)_a{}^m = N_{ia}^m$.
(3) Associativity in the form $\phi_j \times (\phi_i \times \phi_a) = (\phi_j \times \phi_i) \times \phi_a$, with commutativity used to write $N_{jm}^b = N_{mj}^b$ and so on.

(4) Rename $m \to k$, use $N_{ji}^{k} = N_{ij}^{k}$, and $(N_k)_a{}^b = N_{ka}^{b} = N_{ak}^{b}$.

In words: multiplying by ϕ_i and then by ϕ_j is the same as multiplying by the product $\phi_i \times \phi_j$.

Worked example 11.98: The Ising fusion algebra and its matrices

The Ising model has $F = 3$ families, ordered $(\mathbf{1}, \varepsilon, \sigma)$. From Section 11.9.7 and its exercise,

$$\sigma \times \sigma = \mathbf{1} + \varepsilon, \qquad \sigma \times \varepsilon = \sigma, \qquad \varepsilon \times \varepsilon = \mathbf{1}, \tag{11.268}$$

plus the identity rules. Every field is self-conjugate. The fusion matrices (row j, column k) are

$$N_{\mathbf{1}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad N_{\varepsilon} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad N_{\sigma} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}.$$

For instance the third row of N_σ says $\sigma \times \sigma = \mathbf{1} + \varepsilon$. Now check that they represent (11.268):

$$N_\sigma N_\sigma \overset{(1)}{=} \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}\begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \overset{(2)}{=} \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix} \overset{(3)}{=} N_{\mathbf{1}} + N_{\varepsilon},$$

$$N_\sigma N_\varepsilon \overset{(4)}{=} \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \overset{(5)}{=} N_\sigma, \qquad N_\varepsilon N_\varepsilon \overset{(6)}{=} N_{\mathbf{1}}.$$

(1) The matrix of "multiply by σ", applied twice.
(2) Row times column: e.g. the $(3,3)$ entry is $1 \cdot 1 + 1 \cdot 1 + 0 = 2$.
(3) Add the matrices $N_{\mathbf{1}}$ and N_ε above.
(4) Multiply by ε first, then by σ.
(5) Right-multiplying by the permutation N_ε swaps the first two columns, which are equal.
(6) N_ε is a transposition, so it squares to the identity.

The entry 2 in N_σ^2 is the statement that $\sigma \times \sigma \times \sigma$ contains σ twice: there are *two* conformal blocks for $\langle\sigma\sigma\sigma\sigma\rangle$, as we found in (11.263). The algebra is also $\mathbb{Z}_2$-graded: give $\mathbf{1}$, ε charge 0 and σ charge 1; charges add under fusion. This is the spin-flip symmetry $\sigma \to -\sigma$ of the lattice.

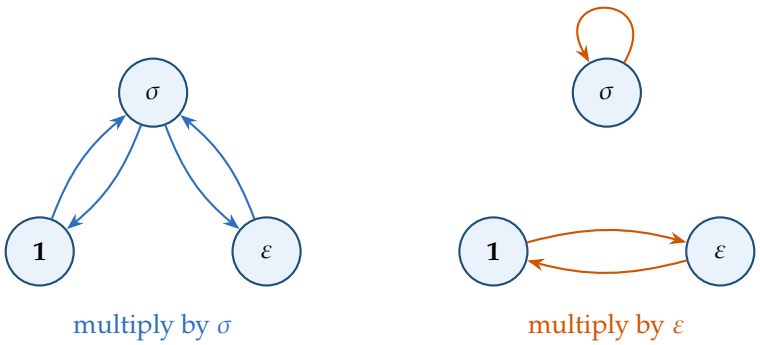


Figure 11.53: Fusion graphs of the Ising model: an arrow $a \to b$ for every occurrence of ϕ_b in $\phi_i \times \phi_a$. The graphs are the fusion matrices N_σ and N_ε drawn as adjacency matrices. What to see: the σ-graph is the Dynkin diagram A_3 (a path $\mathbf{1} - \sigma - \varepsilon$); the number of paths of length n from $\mathbf{1}$ to ϕ_b counts the blocks of $\langle\sigma \cdots \sigma\, \phi_b\rangle$ with n spins.

Worked example 11.99: Associativity checked by counting blocks

Count the blocks of $\langle\sigma\sigma\sigma\sigma\rangle$ in two ways, with $i=j=k=l=\sigma$:

$$\sum_m N_{\sigma\sigma}^{m} N_{m\sigma}^{\sigma} \overset{(1)}{=} N_{\sigma\sigma}^{\mathbf{1}} N_{\mathbf{1}\sigma}^{\sigma} + N_{\sigma\sigma}^{\varepsilon} N_{\varepsilon\sigma}^{\sigma} + N_{\sigma\sigma}^{\sigma} N_{\sigma\sigma}^{\sigma} \overset{(2)}{=} 1\cdot 1 + 1\cdot 1 + 0\cdot 0 = 2,$$

(1) Sum over the intermediate family $m \in \{\mathbf{1}, \varepsilon, \sigma\}$.
(2) The Ising fusion rules (11.268); in particular $\sigma \notin \sigma\times\sigma$.

and $\sum_m N_{\sigma\sigma}^{m} N_{\sigma m}^{\sigma}$ gives the same 2 by commutativity. Now a mixed case, $\langle \varepsilon\,\sigma\,\sigma\,\varepsilon\rangle$ (the family l at infinity is $\varepsilon' = \varepsilon$): in the s-channel we fuse $\sigma\times\varepsilon = \sigma$ first, and $\sigma\times\sigma \ni \varepsilon$ once: one block. In the t-channel we fuse $\varepsilon\times\sigma = \sigma$, then $\sigma\times\sigma \ni \varepsilon$ once: one block. They agree, as they must.

11.10.2 Vertex operators: fusion is momentum conservation

Nash's simple example (p. 316) is the free scalar field. We recall from §2 (Nash's (11.8)) that φ itself is not primary: its two-point function is a logarithm, $\langle\varphi(z,\bar z)\varphi(0)\rangle = -\frac{1}{2}\ln|z|^2$ in the normalisation of Section 11.8.5, and a logarithm is not a power. Its exponentials are primary.

Working definition 11.35 (vertex operator)

For $p\in\mathbb{R}$ the **vertex operator** is

$$V_p(z,\bar z) = :\exp\big(ip\varphi(z,\bar z)\big): = \sum_{n\ge 0}\frac{(ip)^n}{n!}:\varphi^n:(z,\bar z),$$

with all contractions inside the exponential removed. It is a primary field of weights $(h,\bar h) = (p^2/4, p^2/4)$ [read off from $\langle V_p(z,\bar z)V_{-p}(0)\rangle = |z|^{-p^2} = z^{-2h}\bar z^{-2\bar h}$]. *Operational meaning:* it creates a state of momentum p in the φ direction, i.e. it shifts the zero mode [the constant part of φ, whose conjugate momentum is $\hat p$]: $V_p|0\rangle \propto |p\rangle$ with $\hat p|p\rangle = p|p\rangle$.

To compute the OPE we need one identity for free fields: if A and B are linear in φ, then

$$:e^{A}::e^{B}: = :e^{A+B}:\, e^{\langle AB\rangle}. \tag{11.269}$$

It follows from Wick's theorem (the free-field toolkit of §2): each term of the product with n contractions between A's and B's gives $\langle AB\rangle^n/n!$ times the normal-ordered rest, and the sum over n is the exponential. (It is the field-theory version of the Baker–Campbell–Hausdorff formula $e^Ae^B = e^{A+B}e^{[A,B]/2}$ for operators with a central commutator.) Now take $A = ip\varphi(z,\bar z)$ and $B = iq\varphi(0)$:

$$\begin{aligned} V_p(z,\bar z)V_q(0) &\overset{(1)}{=} :e^{ip\varphi(z,\bar z)+iq\varphi(0)}:\, e^{(ip)(iq)\langle\varphi(z,\bar z)\varphi(0)\rangle} \\ &\overset{(2)}{=} :e^{ip\varphi(z,\bar z)+iq\varphi(0)}:\, e^{\frac{pq}{2}\ln|z|^2} \\ &\overset{(3)}{=} |z|^{pq}\Big(:e^{i(p+q)\varphi(0)}: + ip\,z\,:\partial\varphi\, e^{i(p+q)\varphi}:(0) + \cdots\Big) \\ &\overset{(4)}{=} z^{pq/2}\,\bar z^{pq/2}\,V_{p+q}(0) + \cdots . \end{aligned} \tag{11.270}$$

(1) The identity (11.269).
(2) $(ip)(iq) = -pq$ and $\langle\varphi\varphi\rangle = -\frac{1}{2}\ln|z|^2$, so the exponent is $+\frac{pq}{2}\ln|z|^2$.

(3) $e^{\frac{pq}{2}\ln|z|^2} = |z|^{pq}$; Taylor expand $\varphi(z,\bar z) = \varphi(0) + z\partial\varphi(0) + \bar z\bar\partial\varphi(0) + \cdots$ inside the normal ordering.

(4) The leading term is the vertex operator of momentum $p+q$; the dots are its descendants, with extra powers of z and $\bar z$.

Check the exponent against the general OPE (11.255), which says the leading power is $z^{h_{p+q}-h_p-h_q}$:

$$h_{p+q} - h_p - h_q = \frac{(p+q)^2 - p^2 - q^2}{4} = \frac{pq}{2}.$$

It agrees. So the only family in $V_p \times V_q$ is $[V_{p+q}]$ (Nash's (11.73)):

$$\phi_p \times \phi_q = \phi_{p+q}, \qquad N_{pq}^r = \delta_{r,p+q}. \tag{11.271}$$

Watch out: The exponent in Nash's (11.72)

Nash writes the prefactor as $|z|^{pq/2} = z^{pq/4}\bar z^{pq/4}$. With the weights $(p^2/4, p^2/4)$ that he states, the exponent must be $h_{p+q} - h_p - h_q = pq/2$, i.e. the prefactor is $|z|^{pq}$ as in (11.270). The two are consistent only if the propagator were $-\frac{1}{4}\ln|z|^2$, which would make the weight $p^2/8$. The fusion rule (11.271) is unaffected.

Worked example 11.100: Momentum conservation from the zero-mode integral

Why must the charges in a correlator of vertex operators add to zero? Split $\varphi = \varphi_0 + \tilde\varphi$ into the constant zero mode and the rest. In the path integral the zero mode appears only through the vertex operators, so

$$\Big\langle \prod_{a=1}^{n} V_{p_a}(z_a) \Big\rangle \overset{(1)}{=} \int \mathrm{d}\varphi_0\, e^{i(p_1+\cdots+p_n)\varphi_0} \Big\langle \prod_a :e^{ip_a\tilde\varphi(z_a)}: \Big\rangle_{\tilde\varphi}$$
$$\overset{(2)}{=} 2\pi\,\delta\Big(\sum_a p_a\Big) \prod_{a<b} |z_{ab}|^{p_a p_b}.$$

(1) The action does not depend on φ_0 (it contains only derivatives), and $V_p = e^{ip\varphi_0}{:}e^{ip\tilde\varphi}{:}$.

(2) $\int \mathrm{d}\varphi_0\, e^{ik\varphi_0} = 2\pi\delta(k)$; the remaining Gaussian integral is (11.269) applied repeatedly.

The δ-function is translation invariance of the target: shifting $\varphi \to \varphi + a$ is a symmetry with conserved charge $\hat p$. For the three-point function this says $\langle V_{-r}V_pV_q\rangle \neq 0$ only for $r = p+q$: that is $N_{pq}^r = \delta_{r,p+q}$ again. For the compact boson of Section 11.8.5, $p = n/R$ is quantised and the δ-function becomes a Kronecker delta.

Nash's extension (p. 316) is to d scalar fields X^μ, $\mu = 1,\dots,d$, the string embedding coordinates of Chapter 9. The vertex operator $V_p = {:}e^{ip_\mu X^\mu}{:}$ has weight $p^2/4$ with $p^2 = p_\mu p^\mu$, the OPE prefactor is $|z|^{p\cdot q}$, and the fusion rule $V_p \times V_q = V_{p+q}$ is the conservation of d-momentum at a vertex of a string diagram. For the closed bosonic string V_p with $h = \bar h = 1$, i.e. $p^2 = 4$ in these units, is the tachyon vertex operator.

Physical picture: Group-like fusion

When every $\phi_i \times \phi_j$ contains a single family, the fusion algebra is the group ring of an abelian group: the momentum lattice for free bosons, $\mathbb{Z}_2$ for the spin-flip sector of Ising's $\{\mathbf{1}, \varepsilon\}$, $\mathbb{Z}_N$ for the parafermion-like vertex operators of a boson at a rational radius. Families with several outputs, like $\sigma \times \sigma = \mathbf{1} + \varepsilon$, are *non-abelian*; they are what makes rational CFT (and the anyons of Chapter 12) interesting.

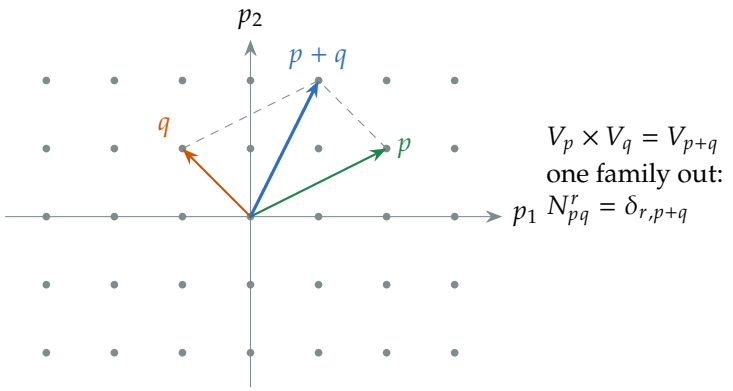


Figure 11.54: Fusion of vertex operators for two compact bosons: the charges live on a lattice (grey), and fusion is vector addition. What to see: the fusion algebra of a free boson is the group algebra of the charge lattice. A group-like fusion rule has exactly one output, so every block space is one- or zero-dimensional.

11.10.3 Conformal blocks on a surface of genus p

In string theory the worldsheet can be any Riemann surface. So we need to say what a conformal block is on a surface Σ of genus p. Nash's answer (p. 317) has three steps.

1. All the local formulas of Section 11.9 (OPEs, Ward identities, null-vector equations) hold near any point of Σ, in a local holomorphic coordinate z. They are statements about neighbourhoods, and every neighbourhood of a Riemann surface looks like a disc in $\mathbb{C}$.

2. A **conformal block** on Σ with n punctures $P_1, \dots, P_n$ carrying primaries $\phi_{i_1}, \dots, \phi_{i_n}$ is a (multivalued) function of the complex structure of Σ and the positions of the punctures, satisfying all the local Ward identities and glued consistently. Write $\Sigma_{p,n}$ for a genus-p surface with n distinct punctures.

3. The blocks form a vector space $V(\Sigma_{p,n})$. Its dimension counts the independent chiral "channels".

The simplest case is the sphere with three punctures. Its blocks are the chiral three-point couplings, one for each occurrence of ϕ_k in $\phi_i \times \phi_j$, so (Nash (11.74))

$$N_{ij}^{k} = \dim V\big(\Sigma_{0,3}\big) \quad \text{with punctures labelled } i,\ j,\ k'. \tag{11.272}$$

Working definition 11.36 (moduli space $\mathcal{M}_{p,n}$)

$\mathcal{M}_{p,n}$ is the set of all complex structures [choices of which local coordinates z count as holomorphic, which fixes angles but not lengths] on a genus-p surface with n marked points, where two are identified if a biholomorphic map [an invertible holomorphic map with holomorphic inverse] carries one to the other (and the marked points to the marked points). A point of $\mathcal{M}_{p,n}$ is "a shape" of the punctured surface [for $p = 1$, $n = 0$ it is a modular parameter $\tau \in \mathbb{H}$, with τ and $\frac{a\tau+b}{c\tau+d}$ counted as the same point]. *Operational meaning:* its complex coordinates are the parameters a conformally invariant quantity can depend on after all the Möbius, translation and rescaling freedom has been used up.

Nash's footnote 5 gives $\dim_{\mathbb{C}} \mathcal{M}_{p,n} = 3p-3+n$. Here is where it comes from. An infinitesimal change of complex structure that fixes the punctures is a class in $H^1(\Sigma, T\Sigma(-P_1 - \cdots - P_n))$

(vector fields vanishing at the punctures), and the holomorphic sections H^0 of the same bundle are the infinitesimal automorphisms, which do not change the shape. So

$$\begin{aligned}\dim \mathcal{M}_{p,n} &\overset{(1)}{=} h^1\big(T\Sigma(-\textstyle\sum P_a)\big) - h^0\big(T\Sigma(-\textstyle\sum P_a)\big)\\ &\overset{(2)}{=} -\operatorname{index}\bar\partial_{T\Sigma(-\sum P_a)}\\ &\overset{(3)}{=} -\Big(\deg T\Sigma(-\textstyle\sum P_a) + 1 - p\Big)\\ &\overset{(4)}{=} -\big((2-2p-n)+1-p\big) = 3p-3+n. \end{aligned} \tag{11.273}$$

(1) Deformations minus automorphisms (the stacky dimension; it is the honest dimension whenever the surface has no continuous automorphisms, i.e. whenever $3p-3+n \geq 0$ and $(p,n) \neq (1,0)$).
(2) The index of the $\bar\partial$ operator on sections of a holomorphic line bundle is $h^0 - h^1$ (Chapter 4).
(3) Riemann–Roch: $\operatorname{index}\bar\partial_L = \deg L + 1 - p$ (Chapter 5).
(4) $\deg T\Sigma = \chi(\Sigma) = 2-2p$ (a vector field has $2-2p$ zeros counted with sign), and each forced zero at a puncture lowers the degree by one.

Worked example 11.101: The first few moduli spaces

- $\mathcal{M}_{0,3}$: $3\cdot 0 - 3 + 3 = 0$. A point. Any three points of the sphere can be moved to $(0,1,\infty)$ by a Möbius map, so there is no shape to vary. This is why three-point functions contain only constants (Section 11.9.3).
- $\mathcal{M}_{0,4}$: dimension 1. The shape of four points is the cross ratio z of (11.249), with $z \neq 0,1,\infty$. So $\mathcal{M}_{0,4} = \mathbb{CP}^1 \setminus \{0,1,\infty\}$, and four-point blocks are functions of z.
- $\mathcal{M}_{1,1}$: dimension 1. A torus with one marked point is fixed by τ modulo $\mathrm{SL}(2,\mathbb{Z})$: $\mathcal{M}_{1,1}$ is the fundamental domain of Figure 11.40. (Without the marked point, translations are automorphisms and (11.273) gives 0, the "1 − 1" of one modulus minus one automorphism.)
- $\mathcal{M}_{2,0}$: dimension 3. A genus-2 surface is a double cover of the sphere branched at 6 points, $y^2 = \prod_{a=1}^{6}(x-e_a)$; three of the e_a can be fixed by Möbius, leaving $6-3=3$ parameters. This matches $3p-3$.

In general $\dim \mathcal{M}_p = 3p-3$ for $p \geq 2$ is Nash's result (5.99) from Chapter 5.

As the surface changes shape, the space of blocks moves with it: $V(\Sigma_{p,n})$ varies holomorphically with the point of $\mathcal{M}_{p,n}$, and its dimension does not jump (it is a count of channels, an integer that cannot change continuously). So the spaces fit together into a holomorphic vector bundle

$$\mathcal{V}_{p,n} \longrightarrow \mathcal{M}_{p,n}, \qquad \text{fibre over } [\Sigma] \;=\; V(\Sigma_{p,n}),$$

and a conformal block is a holomorphic section of it. For $p=0$ this is the remark of Section 11.9.6: the blocks are holomorphic on $\mathcal{M}_{0,n}$, and their monodromy is the flat structure of the bundle. In general $\mathcal{V}_{p,n}$ carries a *projectively* flat connection (flat up to a scalar curvature proportional to c), which we meet again in the Friedan–Shenker picture of Section 11.11.

11.10.4 Sewing: every block space from three-punctured spheres

The final sentence of Nash's p. 317 contains a powerful idea: "any punctured Riemann surface $\Sigma_{p,n}$ can be assembled by joining together $\Sigma_{0,3}$'s". Physically, it is the statement that every

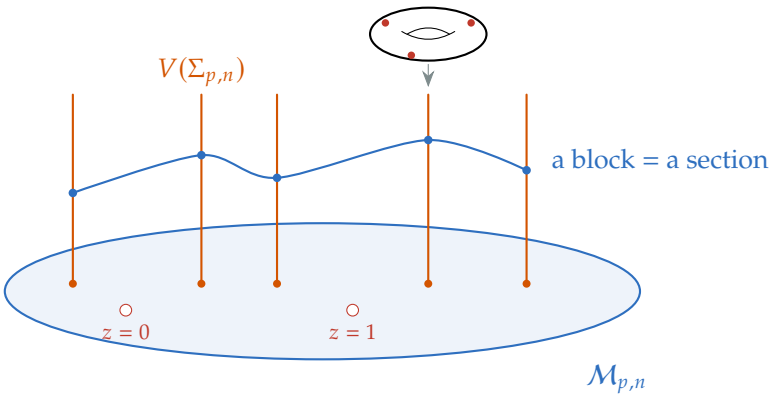


Figure 11.55: Conformal blocks as a vector bundle over moduli space. Each point of the base $\mathcal{M}_{p,n}$ is a shape of the punctured surface (one is drawn above its fibre); the fibre is the space of blocks $V(\Sigma_{p,n})$ for that shape. What to see: a block is a holomorphic section, and going round a hole of the base (for $\mathcal{M}_{0,4}$ the removed points $z = 0, 1$) returns it to a different vector of the same fibre: monodromy.

string diagram is built from cubic vertices and propagators.

Key idea: Pants decomposition

Cut a surface $\Sigma_{p,n}$ along $3p - 3 + n$ disjoint simple closed curves. The pieces are $2p - 2 + n$ **pairs of pants**: spheres with three holes (or punctures). Every $\Sigma_{p,n}$ with $2p - 2 + n > 0$ can be cut this way.

The counting is Euler characteristic bookkeeping. A pair of pants has $\chi = 2 - 3 = -1$; gluing along circles ($\chi = 0$) adds Euler characteristics; so the number of pants is $-\chi(\Sigma_{p,n}) = 2p - 2 + n$. Each pant has three boundary circles, giving $3(2p - 2 + n)$ circles; n of them are the punctures and the rest are glued in pairs, so the number of cuts is $\frac{1}{2}\big(3(2p - 2 + n) - n\big) = 3p - 3 + n$, the dimension of $\mathcal{M}_{p,n}$. This is no coincidence: each cut has a length and a twist, which together form one complex modulus (the plumbing parameter of Section 11.11).

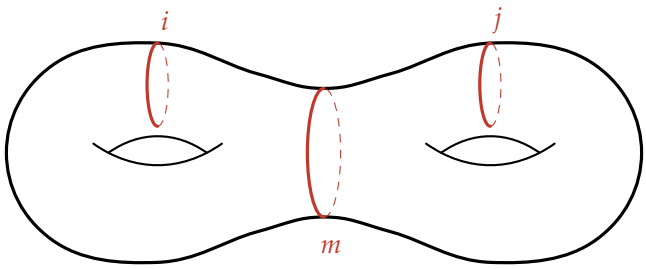
Key idea: The sewing rule for dimensions

Cut Σ along a circle C into Σ_1 and Σ_2 (or, if C does not separate, into one surface Σ' with two new punctures). Then

$$\dim V(\Sigma) = \sum_m \dim V(\Sigma_1;\, m)\, \dim V(\Sigma_2;\, m') \qquad \Big(\text{or} \sum_m \dim V(\Sigma';\, m, m')\Big), \tag{11.274}$$

where m runs over the primaries and the new punctures carry m and its conjugate m'.

Why is this true? Near the circle C the surface looks like a long thin cylinder. Every state crossing the cylinder decomposes into families: by (11.200) the Hilbert space on a circle is $\bigoplus_m H_m \otimes \bar{H}_m$. The chiral blocks sort themselves according to which H_m flows through C: this is the insertion of a complete set of states, the "$\sum_m$", and it is the same mechanism that expanded the four-point function into blocks in (11.257). For each m the blocks factorise as (blocks of Σ_1 with m at the new puncture) $\otimes$ (blocks of Σ_2 with m'), and dimensions multiply.



$$\dim V(\Sigma_{2,0}) = \sum_m \Big(\sum_i N_{im}^{\,i}\Big)\Big(\sum_j N_{jm'}^{\,j}\Big) = \sum_m \operatorname{tr} N_m \ \operatorname{tr} N_{m'}$$

Figure 11.56: A pants decomposition of a closed genus-2 surface: $3p-3=3$ cuts (red) and $2p-2=2$ pairs of pants. The waist cut m separates two one-holed tori; the cut i (resp. j) on each handle turns a one-holed torus into a pair of pants whose three boundaries carry i, i' and m. What to see: summing over the labels of the cuts, with one fusion number per pair of pants, gives the dimension of the space of genus-2 blocks.

Worked example 11.102: Dimensions from the Ising fusion rules

(a) Four spins on the sphere. Cut $\Sigma_{0,4}$ into two pants joined along one circle with label m:

$$\dim V(\Sigma_{0,4};\sigma,\sigma,\sigma,\sigma) \overset{(1)}{=} \sum_m N_{\sigma\sigma}^{m} N_{m\sigma}^{\sigma} \overset{(2)}{=} 2,$$

(1) (11.274) with (11.272) for each pant (all fields are self-conjugate).
(2) The count of the associativity example.

the two blocks $\mathcal{F}_{\mathbf{1}}$, $\mathcal{F}_{\varepsilon}$ of (11.263).

(b) The torus without punctures. Cut along one non-separating circle; the result is a sphere with two punctures labelled m and m':

$$\dim V(\Sigma_{1,0}) \overset{(1)}{=} \sum_m N_{mm'}^{\mathbf{1}} \overset{(2)}{=} \sum_m 1 = F = 3.$$

(1) (11.274); a sphere with two punctures is a sphere with three punctures, one of them the identity.
(2) The conjugation rule: $N_{mm'}^{\mathbf{1}} = 1$ for every m.

These three blocks are the three characters $\chi_0, \chi_{1/2}, \chi_{1/16}$: the torus blocks are the characters, and the torus partition function $\sum_m |\chi_m|^2$ is the Hermitian combination (11.259) at genus one.

(c) The torus with one puncture ϕ_j. Cut the handle: a sphere with punctures m, m', j, so $\dim V(\Sigma_{1,1};j) = \sum_m N_{jm}^{m} = \operatorname{tr} N_j$. For Ising: $\operatorname{tr} N_{\mathbf{1}} = 3$, $\operatorname{tr} N_{\varepsilon} = 1$, $\operatorname{tr} N_{\sigma} = 0$ (read off the diagonals of the matrices above). So $\langle\sigma\rangle_{\text{torus}} = 0$, the $\mathbb{Z}_2$ symmetry again, and $\langle\varepsilon\rangle_{\text{torus}}$ has exactly one block (with σ circulating round the handle).

(d) Genus two. Use Figure 11.56:

$$\dim V(\Sigma_{2,0}) \overset{(1)}{=} \sum_m \operatorname{tr} N_m \ \operatorname{tr} N_{m'} \overset{(2)}{=} 3^2 + 1^2 + 0^2 = 10.$$

(1) Two one-punctured tori glued along the waist, each contributing (c).
(2) The traces from (c), with $m = \mathbf{1}, \varepsilon, \sigma$.

The answer must not depend on how we cut. For $\Sigma_{0,4}$ the s- and t-channel decompositions

give the two sides of the associativity rule, so this independence *is* associativity. A theorem of Moore and Seiberg says that for genus zero and genus one the relations between different pants decompositions are generated by a handful of moves (the fusion move on $\Sigma_{0,4}$, the S move on $\Sigma_{1,1}$, and braiding), so consistency on those surfaces is enough for every surface.

Working definition 11.37 (the handle operator)

Attaching one handle to a surface is a sum over the family circulating through it [cutting the handle leaves two new punctures labelled i and i', and one sums over i]. As a matrix acting on the label of a puncture, it is

$$\mathcal{H} = \sum_i N_i N_{i'}, \qquad \dim V(\Sigma_{p,1}; j) = (\mathcal{H}^p)_j{}^{\mathbf{1}}, \qquad \dim V(\Sigma_{p,0}) = (\mathcal{H}^p)_{\mathbf{1}}{}^{\mathbf{1}}. \tag{11.275}$$

with N_i the fusion matrices [$(N_i)_j{}^k = N_{ij}^k$] and $V(\Sigma_{p,1}; j)$ [the space of conformal blocks on a genus-p surface with one puncture carrying ϕ_j]. *Operational meaning:* to add a handle next to a puncture labelled a, fuse a with a pair i, i' in all possible ways.

For the Ising model the handle operator can be computed in closed form:

$$\mathcal{H} \overset{(1)}{=} N_{\mathbf{1}}^2 + N_\varepsilon^2 + N_\sigma^2 \overset{(2)}{=} \mathbb{1} + \mathbb{1} + (\mathbb{1} + N_\varepsilon) = 3\,\mathbb{1} + N_\varepsilon,$$
$$\mathcal{H}^2 \overset{(3)}{=} 9\,\mathbb{1} + 6N_\varepsilon + N_\varepsilon^2 \overset{(4)}{=} 10\,\mathbb{1} + 6N_\varepsilon, \qquad \mathcal{H}^3 \overset{(5)}{=} 36\,\mathbb{1} + 28N_\varepsilon.$$

(1) All Ising fields are self-conjugate, $i' = i$.
(2) $N_{\mathbf{1}} = \mathbb{1}$, $N_\varepsilon^2 = \mathbb{1}$ and $N_\sigma^2 = \mathbb{1} + N_\varepsilon$ from Example 11.98.
(3) Square: $N_{\mathbf{1}}$ commutes with everything.
(4) $N_\varepsilon^2 = \mathbb{1}$.
(5) $(10\,\mathbb{1} + 6N_\varepsilon)(3\,\mathbb{1} + N_\varepsilon) = 30\,\mathbb{1} + 10N_\varepsilon + 18N_\varepsilon + 6\,\mathbb{1}$.

Since $(N_\varepsilon)_{\mathbf{1}}{}^{\mathbf{1}} = 0$, the numbers of blocks on closed surfaces of genus $1, 2, 3$ are $3, 10, 36$. Genus 2 agrees with part (d) of the example, obtained by cutting differently.

11.10.5 The Verlinde formula

The fusion rules so far came from hard work: null vectors, differential equations and indicial exponents. Verlinde (1988) found that they are hidden in something much simpler, the behaviour of the torus characters under $\tau \mapsto -1/\tau$. This is Nash's p. 317: "the matrix representing the action of S on the characters can be used to diagonalise the fusion algebra".

Key idea: The Verlinde formula

Let $\chi_i(-1/\tau) = \sum_j S_{ij}\chi_j(\tau)$ define the modular S matrix of a rational CFT, with 0 the label of the identity. Then S is symmetric and unitary, the fusion matrices are simultaneously diagonalised by S,

$$N_i = S\,\Lambda_i\,S^{-1}, \qquad \Lambda_i = \operatorname{diag}\!\left(\frac{S_{im}}{S_{0m}}\right)_m, \tag{11.276}$$

and therefore

$$N_{ij}^k = \sum_m \frac{S_{im}\,S_{jm}\,\bar{S}_{km}}{S_{0m}}. \tag{11.277}$$

We explain the formula in three steps: why the N_i can be diagonalised at all, why S is the diagonalising matrix, and why the eigenvalues are ratios of entries of S.

Step 1: one basis that diagonalises all the N_i at once

By (11.267) the N_i commute. They are also normal: $N_i^{\mathsf{T}} = N_{i'}$ (because $N_{ij}^{k} = N_{ik'}^{j'}$, a relabelling of the three punctures of $\Sigma_{0,3}$), and $N_{i'}$ is again one of the commuting family, so $N_i N_i^{\mathsf{T}} = N_i N_{i'} = N_{i'} N_i = N_i^{\mathsf{T}} N_i$. A family of commuting normal matrices can be **diagonalised simultaneously**: there is *one* unitary change of basis U such that $U^{-1} N_i U$ is diagonal for *every* i.

Watch out: What "commuting matrices share eigenvectors" does and does not say

It does *not* say the matrices are diagonal. In the basis of primary fields the N_i are full matrices of non-negative integers. What commutation (plus normality) buys is the *existence* of some other basis, the common eigenvectors $\psi^{(m)}$, in which all of them become diagonal together. Normality matters. Without it, commuting matrices are only guaranteed to share *one* eigenvector, and may not be diagonalisable at all. For example, $\left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right)$ commutes with $\mathbb{1}$, and both have the eigenvector $(1, 0)$, but the first has no basis of eigenvectors. So "simultaneously diagonalisable" is a statement about a basis that exists, not about the matrices as written.

Let $\psi^{(m)}$, $m = 0, \ldots, F-1$, be the common eigenvectors and $\lambda_i^{(m)}$ the eigenvalue of N_i on $\psi^{(m)}$. Applying (11.267) to $\psi^{(m)}$ gives

$$\lambda_i^{(m)} \lambda_j^{(m)} = \sum_k N_{ij}^{k} \lambda_k^{(m)}. \tag{11.278}$$

So for each m the numbers $\phi_i \mapsto \lambda_i^{(m)}$ are a one-dimensional representation of the fusion algebra: a way of replacing each field by a number that respects the fusion rules. There are F of them.

Step 2: why S diagonalises (Verlinde's loop operators)

Think of the torus as a cylinder, of circumference 2π in the space direction, whose ends have been glued after a Euclidean time $2\pi \operatorname{Im}\tau$. Call the space circle the a-cycle and the time circle the b-cycle (Figure 11.57). The torus blocks are the characters: in $\chi_j(\tau) = \operatorname{Tr}_{H_j} q^{L_0 - c/24}$ the family j propagates around the b-cycle.

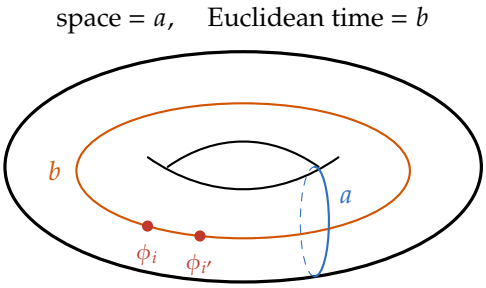
Now define an operation on blocks: create the pair ϕ_i, $\phi_{i'}$ out of the vacuum at some point (allowed, since $\mathbf{1} \in \phi_i \times \phi_{i'}$), carry ϕ_i once around a closed curve C, and let the pair annihilate back into the vacuum. The result is again a torus block, so this defines a linear map $\phi_i(C)$ on the F-dimensional space $V(\Sigma_{1,0})$.

- $C = a$ (around space, at fixed time). The whole operation happens on one time slice, so it is an operator on each H_j. The loop can be slid past any insertion of T (the stress tensor is single valued), so the operator commutes with all L_n. By Schur's lemma it is a number on the irreducible module H_j: $\phi_i(a)\chi_j = \lambda_i^{(j)} \chi_j$. *Diagonal.*

- $C = b$ (around time). On every time slice between creation and annihilation, the state is ϕ_i fused with the family j that was circulating: it lies in $\bigoplus_k N_{ij}^{k} H_k$. When the pair annihilates, family k is circulating instead of j: $\phi_i(b)\chi_j = \sum_k N_{ij}^{k} \chi_k$. *The fusion matrix.*

- The modular transformation $S : \tau \mapsto -1/\tau$ exchanges the two cycles. In the basis of characters it is the matrix S, so $\phi_i(b) = S\, \phi_i(a)\, S^{-1}$.

Putting the three together: $N_i = S\Lambda_i S^{-1}$, with Λ_i diagonal. The columns of S are the common eigenvectors $\psi^{(m)}$ of Step 1.



$$\phi_i(a)\,\chi_j = \frac{S_{ij}}{S_{0j}}\,\chi_j$$

$$\phi_i(b)\,\chi_j = \textstyle\sum_k N_{ij}^k\,\chi_k$$

$$S:\ a \leftrightarrow b$$

Figure 11.57: Verlinde's loop operators on the torus. A pair ϕ_i, $\phi_{i'}$ is created, ϕ_i travels round a cycle, and the pair annihilates. What to see: round the space cycle a the operation acts inside one module and is a number; round the time cycle b it changes the circulating family by fusion. The modular transformation S swaps the cycles, so S turns the diagonal matrices into the fusion matrices.

Step 3: the eigenvalues

The identity family is special: $\phi_i \times \mathbf{1} = \phi_i$. Use it to find the eigenvalues. Take the $(0, m)$ entry of $N_i S = S\Lambda_i$:

$$(N_i S)_{0m} \overset{(1)}{=} \sum_k (N_i)_0{}^k\, S_{km} \overset{(2)}{=} \sum_k N_{i0}^k\, S_{km} \overset{(3)}{=} S_{im},$$

$$(S\Lambda_i)_{0m} \overset{(4)}{=} S_{0m}\,\lambda_i^{(m)}.$$

(1) Matrix multiplication.
(2) The definition of the fusion matrix, row 0.
(3) The identity rule $N_{i0}^k = N_{0i}^k = \delta_i^k$.
(4) Λ_i is diagonal with entries $\lambda_i^{(m)}$.

Therefore $\lambda_i^{(m)} = S_{im}/S_{0m}$. (For a unitary theory $S_{0m} > 0$ for all m, so the division is allowed; this positivity is the statement that the vacuum character dominates as $\tau \to 0$, Cardy's argument of Section 11.8.6.) Now write out $N_i = S\Lambda_i S^{-1}$ entry by entry:

$$N_{ij}^k \overset{(1)}{=} \sum_m S_{jm}\,\lambda_i^{(m)}\,(S^{-1})_{mk} \overset{(2)}{=} \sum_m S_{jm}\,\frac{S_{im}}{S_{0m}}\,\bar{S}_{km} = \sum_m \frac{S_{im}S_{jm}\bar{S}_{km}}{S_{0m}},$$

(1) Matrix product, with $(N_i)_j{}^k = N_{ij}^k$.
(2) Step 3 for λ; S unitary and symmetric, so $(S^{-1})_{mk} = \bar{S}_{km}$.

which is (11.277). (Had the orientation conventions given $S^{-1}\Lambda_i S$ instead, the result would be the complex conjugate of (11.277); since N_{ij}^k is real, nothing changes.)

Watch out: What was proved, and when

Step 2 is a physical argument: it assumes that the loop operators are well defined and behave as described, which Verlinde conjectured. The formula was derived from the axioms of rational CFT by Moore and Seiberg (1988–89) and proved rigorously for rational vertex operator algebras by Huang (2005). In every example we can check, it works; the

next examples check it.

Worked example 11.103: Verlinde for Ising: $\sigma \times \sigma = \mathbf{1} + \varepsilon$ from the torus

The Ising S matrix of (11.227), in the order $(\mathbf{1}, \varepsilon, \sigma)$, is real, with $S_{0m} = (\frac{1}{2}, \frac{1}{2}, \frac{\sqrt{2}}{2})$. Compute $N_{\sigma\sigma}^k$ for each k:

$$N_{\sigma\sigma}^{\mathbf{1}} \overset{(1)}{=} \frac{S_{\sigma\mathbf{1}}S_{\sigma\mathbf{1}}S_{\mathbf{11}}}{S_{\mathbf{11}}} + \frac{S_{\sigma\varepsilon}S_{\sigma\varepsilon}S_{\mathbf{1}\varepsilon}}{S_{\mathbf{1}\varepsilon}} + \frac{S_{\sigma\sigma}S_{\sigma\sigma}S_{\mathbf{1}\sigma}}{S_{\mathbf{1}\sigma}} \overset{(2)}{=} \tfrac{1}{2} + \tfrac{1}{2} + 0 = 1,$$

$$N_{\sigma\sigma}^{\varepsilon} \overset{(3)}{=} \frac{(\frac{\sqrt{2}}{2})^2 \cdot \frac{1}{2}}{\frac{1}{2}} + \frac{(-\frac{\sqrt{2}}{2})^2 \cdot \frac{1}{2}}{\frac{1}{2}} + 0 = 1,$$

$$N_{\sigma\sigma}^{\sigma} \overset{(4)}{=} \frac{(\frac{\sqrt{2}}{2})^2 \cdot \frac{\sqrt{2}}{2}}{\frac{1}{2}} + \frac{(-\frac{\sqrt{2}}{2})^2 \cdot (-\frac{\sqrt{2}}{2})}{\frac{1}{2}} + 0 = \frac{\sqrt{2}}{2} - \frac{\sqrt{2}}{2} = 0.$$

(1) (11.277) with $i = j = \sigma$, $k = \mathbf{1}$, summing over $m = \mathbf{1}, \varepsilon, \sigma$.

(2) $S_{\sigma\mathbf{1}} = \frac{\sqrt{2}}{2}$, $S_{\sigma\varepsilon} = -\frac{\sqrt{2}}{2}$, $S_{\sigma\sigma} = 0$; the first two terms are $\frac{1}{2}$ each.

(3) Now $\bar{S}_{km} = S_{\varepsilon m} = (\frac{1}{2}, \frac{1}{2}, -\frac{\sqrt{2}}{2})$; the third term has the factor $S_{\sigma\sigma}^2 = 0$.

(4) $S_{\sigma m} = (\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}, 0)$; the two surviving terms cancel.

So $\sigma \times \sigma = \mathbf{1} + \varepsilon$: the result of the four-spin differential equation of Section 11.9.7, now obtained from nothing but the modular properties of the characters. The eigenvalues of N_σ are $S_{\sigma m}/S_{0m} = (\sqrt{2}, -\sqrt{2}, 0)$; indeed $\det(N_\sigma - \lambda) = -\lambda^3 + 2\lambda$ for the matrix of Example 11.98.

Working definition 11.38 (quantum dimension)

The **quantum dimension** of ϕ_i is $d_i = S_{i0}/S_{00}$ [S is the matrix with $\chi_i(-1/\tau) = \sum_j S_{ij}\chi_j(\tau)$. For Ising $d_\varepsilon = 1$ and $d_\sigma = \sqrt{2}$], the eigenvalue of N_i on the column $m = 0$ [the eigenvector $(S_{k0})_k$, all of whose entries are positive]. By (11.278) $d_i d_j = \sum_k N_{ij}^k d_k$, and since N_i has non-negative entries, d_i is its largest eigenvalue (Perron–Frobenius [if a matrix with non-negative entries has an eigenvector with all entries positive, that eigenvalue is the largest]). *Operational meaning:* the number of blocks with n copies of ϕ_i grows like d_i^n.

For Ising $d_{\mathbf{1}} = d_\varepsilon = 1$ and $d_\sigma = \sqrt{2}$; check $d_\sigma^2 = 2 = d_{\mathbf{1}} + d_\varepsilon$. A system of $2n$ spin fields has 2^{n-1} blocks, i.e. $(\sqrt{2})^{2n}/2$: this is the degeneracy of $2n$ Majorana zero modes (the non-abelian anyons of the $p + ip$ superconductor, whose edge theory is the chiral Ising model).

The dimension of block spaces from S

Combine the handle operator (11.275) with the diagonal form (11.276):

$$\dim V(\Sigma_{p,0}) \overset{(1)}{=} (\mathcal{H}^p)_0{}^0 \overset{(2)}{=} \Big(S\Big(\sum_i \Lambda_i \Lambda_{i'}\Big)^p S^{-1}\Big)_0^{\ 0}$$

$$\overset{(3)}{=} \sum_m S_{0m}\Big(\sum_i \frac{|S_{im}|^2}{S_{0m}^2}\Big)^p \bar{S}_{0m} \overset{(4)}{=} \sum_m S_{0m}^{2-2p}. \tag{11.279}$$

(1) The handle operator applied p times to the identity label.

(2) Each $N_i = S\Lambda_i S^{-1}$; the $S^{-1}S$ in between cancel.

(3) The diagonal entries: $\lambda_{i'}^{(m)} = S_{i'm}/S_{0m} = \bar{S}_{im}/S_{0m}$ (charge conjugation is complex conjugation of S).

(4) $\sum_i |S_{im}|^2 = 1$ by unitarity, and S_{0m} is real positive.

This is the **Verlinde dimension formula**. For Ising: $\sum_m S_{0m}^{2-2p} = 2 \cdot 2^{2p-2} + 2^{p-1}$, which is $1, 3, 10, 36$ for $p = 0, 1, 2, 3$, exactly the numbers found by sewing.

Worked example 11.104: $\mathbb{Z}_N$ fusion: Verlinde is Fourier analysis

Take $F = N$ fields ϕ_a, $a \in \mathbb{Z}_N$, with group-like fusion $\phi_a \times \phi_b = \phi_{a+b \bmod N}$. (A compact boson at radius $R^2 = 2N$ in suitable units, or SU(N) at level 1, realises it.) The S matrix is the discrete Fourier transform, $S_{ab} = N^{-1/2} e^{2\pi i ab/N}$. Verlinde gives

$$N_{ab}^c \overset{(1)}{=} \sum_{m=0}^{N-1} \frac{N^{-1/2} e^{2\pi i am/N}\, N^{-1/2} e^{2\pi i bm/N}\, N^{-1/2} e^{-2\pi i cm/N}}{N^{-1/2}}$$
$$\overset{(2)}{=} \frac{1}{N} \sum_{m=0}^{N-1} e^{2\pi i (a+b-c)m/N} \overset{(3)}{=} \delta_{c,\, a+b \bmod N}.$$

(1) (11.277) with $S_{0m} = N^{-1/2}$.
(2) Collect the phases.
(3) A geometric sum of N-th roots of unity is N if the exponent is a multiple of N and 0 otherwise.

The Verlinde formula is the statement that the characters of the group $\mathbb{Z}_N$ (the columns of S) turn convolution into multiplication. For a non-abelian fusion algebra the same thing happens, with the columns of S playing the role of a character table. The block dimensions are $\sum_m S_{0m}^{2-2p} = N \cdot N^{p-1} = N^p$ at genus p.

Worked example 11.105: $\mathrm{SU}(2)_k$: truncated Clebsch–Gordan

For SU(2) at level k the primaries are the spins $j = 0, \frac{1}{2}, \ldots, \frac{k}{2}$ and S is given in (11.238). Evaluating (11.277) with trigonometric sums gives

$$j_1 \times j_2 = \sum_{j=|j_1-j_2|}^{\min(j_1+j_2,\, k-j_1-j_2)} j \qquad \text{(step 1)}. \tag{11.280}$$

This is the Clebsch–Gordan rule of angular momentum, cut off at the top by $j \le k - j_1 - j_2$. Three cases (all checked numerically from S):

- $k = 1$: $\frac{1}{2} \times \frac{1}{2} = 0$. The spin 1 of ordinary angular momentum is removed. The algebra is $\mathbb{Z}_2$, and the block dimensions are 2^p.
- $k = 2$: $\frac{1}{2} \times \frac{1}{2} = 0 + 1$, $\frac{1}{2} \times 1 = \frac{1}{2}$, $1 \times 1 = 0$. This is the Ising fusion algebra with $\sigma \leftrightarrow \frac{1}{2}$, $\varepsilon \leftrightarrow 1$. In the order $(0, 1, \frac{1}{2})$ the S matrix of (11.238) with $k = 2$ is *identical* to the Ising S of (11.227). The theories are different ($c = \frac{3}{2}$ versus $\frac{1}{2}$, and $h_{1/2} = \frac{3}{16}$ versus $h_\sigma = \frac{1}{16}$): fusion rules and S do not determine the theory; the T matrix distinguishes them. ($\mathrm{SU}(2)_2$ is three Majorana fermions, $3 \times \frac{1}{2} = \frac{3}{2}$.)
- $k = 3$: $1 \times 1 = 0 + 1$. The pair $\{0, 1\}$ closes under fusion; with $d_1 = S_{10}/S_{00} = \sin\frac{3\pi}{5} / \sin\frac{\pi}{5} = \frac{1+\sqrt{5}}{2}$, the golden ratio, and $d_1^2 = 1 + d_1$. This is the "Fibonacci" fusion rule, whose block spaces have Fibonacci-number dimensions.

The level-k dimension formula is $\dim V(\Sigma_{p,0}) = \left(\frac{k+2}{2}\right)^{p-1} \sum_{a=1}^{k+1} \left(\sin\frac{\pi a}{k+2}\right)^{2-2p}$, the famous Verlinde formula for the dimension of the space of non-abelian theta functions (holomorphic

sections of the k-th power of a line bundle over the moduli space of flat SU(2) connections on Σ). For $k = 2$ it gives $1, 3, 10, 36, \ldots$, the Ising numbers again.

11.10.6 Kac–Moody theories and the dimensions (11.75)

Nash recalls (p. 317) that CFTs can be built from representations of the loop group $\widetilde{LG}$, i.e. of the Kac–Moody algebra $\hat{\mathfrak{g}}_k$ (the WZW models of Figure 11.28). A primary is then labelled by an irreducible representation R of G (with a level-dependent upper bound), and a primary field $\phi_R(z)$ is a vector of fields transforming in R:

$$[J_0^a, \phi_R(z)] = -\, t_R^a\, \phi_R(z),$$

where t_R^a are the representation matrices of the generators and J_0^a is the zero mode of the current, the global G charge. For the next chapter Nash needs only spheres, so he writes $V(S^2_{R_1,\ldots,R_n})$ for the space of blocks on the sphere with punctures carrying $R_1, \ldots, R_n$.

Key idea: Nash (11.75)

$$\dim V(S^2_R) = \begin{cases} 1, & R \text{ trivial}, \\ 0, & \text{otherwise}, \end{cases} \qquad \dim V(S^2_{R_i,R_j}) = g_{ij} \equiv N_{ij}^0 = \begin{cases} 1, & R_i = R_j^*, \\ 0, & \text{otherwise}, \end{cases} \tag{11.281}$$

where R^* is the **dual** (complex conjugate) representation.

Both statements follow from the global G symmetry and Schur's lemma. The vacuum is G-invariant, $J_0^a|0\rangle = 0$ and $\langle 0|J_0^a = 0$. So for one puncture

$$0 \overset{(1)}{=} \langle 0|\,[J_0^a, \phi_R(z)]\,|0\rangle \overset{(2)}{=} -\, t_R^a\, \langle 0|\phi_R(z)|0\rangle,$$

(1) J_0^a annihilates the vacuum on both sides.
(2) The transformation law of ϕ_R.

The one-point function is a vector $v \in R$ killed by every generator: a G-invariant vector. In an irreducible R the invariant vectors form a subrepresentation, which is either 0 or all of R; it is all of R only if G acts trivially. Hence $\dim V(S^2_R) = 1$ for R trivial and 0 otherwise. For two punctures the same argument shows that $\langle \phi_{R_i}\phi_{R_j}\rangle$ is an invariant vector in $R_i \otimes R_j$, i.e. an invariant bilinear pairing, i.e. a G-map $R_j \to R_i^*$. By Schur's lemma such a map is zero unless $R_j \cong R_i^*$, and then it is unique up to a constant. That is g_{ij}.

Worked example 11.106: Duals for SU(2) and SU(3)

SU(2). Every representation is self-dual. For the doublet, the invariant pairing is the antisymmetric tensor $\epsilon = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$: for $U \in \mathrm{SU}(2)$,

$$U^{\mathsf{T}} \epsilon\, U \overset{(1)}{=} (\det U)\, \epsilon \overset{(2)}{=} \epsilon,$$

(1) For any 2×2 matrix, $U^{\mathsf{T}}\epsilon U = (\det U)\epsilon$ (multiply out with $U = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$: the $(1,2)$ entry is $ad - bc$).
(2) $\det U = 1$.

So $\langle \phi_{1/2}\phi_{1/2}\rangle \propto \epsilon_{\alpha\beta}$, the spin singlet $\frac{1}{\sqrt{2}}(|\uparrow\downarrow\rangle - |\downarrow\uparrow\rangle)$, and $g_{\frac{1}{2}\frac{1}{2}} = 1$.

SU(3). The triplet $\mathbf{3}$ is not self-dual. Take $U = \mathrm{diag}(e^{i\theta}, e^{i\theta}, e^{-2i\theta})$. Its character in $\mathbf{3}$ is $\mathrm{tr}\, U = 2e^{i\theta} + e^{-2i\theta}$ and in $\mathbf{3}^*$ it is $\mathrm{tr}\, \bar{U} = 2e^{-i\theta} + e^{2i\theta}$; these differ for generic θ, so $\mathbf{3} \not\cong \mathbf{3}^* = \bar{\mathbf{3}}$.

Hence $g_{\mathbf{33}} = 0$ and $g_{\mathbf{3\bar{3}}} = 1$: a quark two-point function needs an antiquark. At level 1 the three primaries $\mathbf{1}, \mathbf{3}, \bar{\mathbf{3}}$ have $\mathbb{Z}_3$ fusion, $\mathbf{3} \times \mathbf{3} = \bar{\mathbf{3}}$, and their S matrix is the $N = 3$ Fourier matrix of the previous example.

With three punctures the same argument gives, at large level, the number of invariants in $R_1 \otimes R_2 \otimes R_3$: the Clebsch–Gordan multiplicity of R_3^* in $R_1 \otimes R_2$. At finite level there is an additional constraint from a null vector of the affine algebra (for $\mathrm{SU}(2)_k$ it is $(J_{-1}^{+})^{k+1-2j}|j\rangle = 0$), and it produces the truncation in (11.280): $j_1 + j_2 + j_3 \le k$. The metric g_{ij} lowers indices, $N_{ijk} = \sum_l N_{ij}^{l} g_{lk}$, and the lowered symbol N_{ijk} is totally symmetric in its three labels: it is the dimension of the space of blocks on a sphere with three punctures, and the punctures can be listed in any order. These spaces $V(S^2_{R_1,\dots,R_n})$ are exactly the Hilbert spaces of Chern–Simons theory on a sphere with n marked points that Chapter 12 uses to build knot invariants.

$\mathrm{SU}(2)_3$: $j_1 \times j_2$	0	$\frac{1}{2}$	1	$\frac{3}{2}$
0	0	$\frac{1}{2}$	1	$\frac{3}{2}$
$\frac{1}{2}$	$\frac{1}{2}$	$0+1$	$\frac{1}{2}+\frac{3}{2}$	1
1	1	$\frac{1}{2}+\frac{3}{2}$	$0+1$	$\frac{1}{2}$
$\frac{3}{2}$	$\frac{3}{2}$	1	$\frac{1}{2}$	0

Every entry follows from (11.280) with $k = 3$: for example 1×1 would contain $0, 1, 2$ by angular momentum, but $2 > k - 1 - 1 = 1$ is cut. In the table the field $\frac{3}{2}$ acts as a simple current: $\frac{3}{2} \times j = \frac{3}{2} - j$.

11.10.7 Higher genus: the Weierstrass gap theorem

Nash ends his discussion of blocks with a single sentence: “the Weierstrass gap theorem is a constraint on the nature of the meromorphic singularities when $p > 0$. This is also a constraint on the properties of some of the correlation functions”. Let us unpack both halves.

The theorem

On the Riemann sphere, for any point P and any $n \ge 1$ there is a meromorphic function with a pole of order exactly n at P and no other poles: $1/(z - P)^n$. On a surface of higher genus this fails.

Key idea: Weierstrass gap theorem

Let Σ have genus $p \ge 1$ and let $P \in \Sigma$. Among the integers $n = 1, 2, 3, \dots$ there are exactly p values, all in the range $1 \le n \le 2p - 1$, for which *no* meromorphic function on Σ has a pole of order exactly n at P and is holomorphic elsewhere. These are the **gaps** at P. For generic P the gaps are $\{1, 2, \dots, p\}$; the finitely many points with a different gap sequence are the **Weierstrass points**.

The proof is Riemann–Roch. Let $\ell(nP)$ be the dimension of the space of meromorphic functions with at most an n-fold pole at P, and K the canonical bundle (holomorphic 1-forms,

degree $2p-2$). Then

$$\ell(nP) \overset{(1)}{=} n+1-p+\ell(K-nP),$$
$$\ell(nP) \overset{(2)}{=} n+1-p \qquad \text{for } n \geq 2p-1,$$

(1) Riemann–Roch for the divisor nP (Chapter 5): $\ell(D)-\ell(K-D)=\deg D+1-p$.
(2) If $n>2p-2$ then $K-nP$ has negative degree, and a line bundle of negative degree has no non-zero holomorphic sections.

Now count. $\ell(0)=1$ (the constants) and $\ell((2p-1)P)=p$. Going from $n-1$ to n, ℓ grows by 0 or 1 (a new function is fixed, up to lower-order ones, by the coefficient of its leading pole). In the $2p-1$ steps from $n=0$ to $n=2p-1$ it grows by $p-1$ in total. So it stays constant at exactly $(2p-1)-(p-1)=p$ steps: those n are the gaps. For $n \geq 2p$ it grows at every step: no more gaps.

Worked example 11.107: Gaps on the torus and on genus two

Torus ($p=1$). The only gap is $n=1$: there is no meromorphic function on a torus with a single simple pole. Reason: a meromorphic function f on the torus gives a meromorphic 1-form $f\,\mathrm{d}z$, and the sum of the residues of a 1-form on a compact surface is zero (Stokes' theorem on the surface minus small discs, Chapter T2). A single simple pole would have non-zero residue. At $n=2$ there is the Weierstrass function $\wp(z)=z^{-2}+O(z^2)$, and $\wp'$ has a triple pole; all $n \geq 2$ occur. Every point of a torus looks the same (translations), so there are no Weierstrass points.
Genus two. Write Σ as $y^2=\prod_{a=1}^{6}(x-e_a)$, a double cover of the x-sphere branched over the six points e_a. At a branch point $P_a=(e_a,0)$ the local coordinate is y, and $x-e_a \sim y^2$. The function $1/(x-e_a)$ has a double pole at P_a and nothing else. So $n=2$ is not a gap at P_a, and the gaps there are $\{1,3\}$. At a generic point the gaps are $\{1,2\}$. The six branch points are the Weierstrass points of a genus-2 surface.

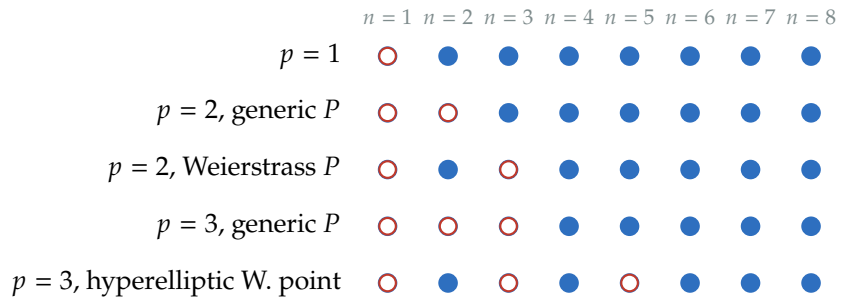


Figure 11.58: Gap sequences. A filled dot at n means a meromorphic function with a pole of order exactly n at P (and no other poles) exists; an open red circle is a gap. What to see: there are always exactly p gaps, all below $2p$; on the sphere ($p=0$) the row would be all filled; the pattern can change only at special points, the Weierstrass points.

Why a field theorist cares

A correlation function on Σ in which a chiral field $A(z)$ of weight h appears, with all other insertions at the single point P, is, as a function of z, a meromorphic section of K^h with poles only at P. The local OPE of A with the fields at P specifies the singular terms. On the plane every choice of singular terms is realised by some global function. On Σ, Riemann–Roch decides which are, and the answer depends on h and on p.

Worked example 11.108: Charge conservation on any surface

Let $J(z)$ be a U(1) current ($h = 1$) and put one charged field $\phi_q(P)$ on Σ, so $J(z)\phi_q(P) \sim \frac{q}{z-P}\phi_q(P)$. Then $\langle J(z)\phi_q(P)\rangle$ is a meromorphic 1-form with at most a simple pole at P. Riemann–Roch for $K + nP$:

$$\ell(K+nP) \overset{(1)}{=} (2p-2+n) + 1 - p + \ell(-nP) \overset{(2)}{=} p - 1 + n \quad (n \geq 1),$$

(1) Riemann–Roch with $\deg(K + nP) = 2p - 2 + n$ and $K - (K + nP) = -nP$.

(2) $\ell(-nP) = 0$: a holomorphic function vanishing at P is zero.

At $n = 1$ this is $p = \ell(K)$: there is no 1-form with a genuine simple pole at one point. So the residue $q\langle\phi_q\rangle$ must vanish: a single charged field has zero expectation value, on every surface. This "gap" at $n = 1$ for 1-forms is the residue theorem, and physically it is charge conservation. For the stress tensor ($h = 2$) the same count gives $\ell(2K + nP) = 3p - 3 + n$ for $p \geq 2$: no gaps, so T can have singularities of every order, and the $3p - 3$ holomorphic quadratic differentials that exist without any pole are the cotangent vectors to $\mathcal{M}_p$ that we meet in Section 11.11.

These constraints are the higher-genus version of the Ward identities. In the operator formalism developed by Alvarez-Gaumé, Gomez, Moore, Reina and Vafa, a genus-p surface with a disc removed around P is represented by a state $\langle\Sigma, P|$ in the Hilbert space of the boundary circle, and that state must be annihilated by exactly those modes of each chiral field that extend holomorphically over $\Sigma \setminus P$. Which modes extend is the Riemann–Roch count above, with its gaps. Vafa used this circle of ideas, together with consistency of sewing at genus two, to show that in a rational CFT the phases $e^{2\pi i(h_i - c/24)}$ are roots of unity: the central charge and the conformal weights are rational numbers.

Reading Nash (pp. 316–318, (11.71)–(11.75))

Nash defines the fusion rules through Verlinde's formal product (11.71), gives the vertex-operator example (11.72)–(11.73) (our (11.270), with the corrected exponent), and states $N_{ij}^k = \dim V(\Sigma_{0,3})$ (11.74). His footnote 5, $\dim \mathcal{M}_{p,n} = 3p - 3 + n$, is our (11.273). The paragraph on Verlinde's work (S diagonalises the fusion algebra; N_{ij}^k in terms of S; sewing gives $\dim V(\Sigma_{p,n})$) is expanded in Sections 11.10.4 and 11.10.5. (11.75) is (11.281), with the Schur's-lemma derivation. The gap theorem is only mentioned by Nash; the proof and the charge-conservation example are ours.

Recap

- N_{ij}^k counts the occurrences of $[\phi_k]$ in $\phi_i\phi_j$. The fusion algebra is commutative and associative with unit $\mathbf{1}$; associativity is crossing symmetry. The matrices $(N_i)_j{}^k$ represent it.
- Vertex operators: $V_p \times V_q = V_{p+q}$, momentum conservation; the OPE prefactor is $|z|^{pq}$.
- Blocks on $\Sigma_{p,n}$ form a vector bundle over $\mathcal{M}_{p,n}$, $\dim \mathcal{M}_{p,n} = 3p - 3 + n$; $N_{ij}^k = \dim V(\Sigma_{0,3})$; sewing along pants decompositions computes every $\dim V(\Sigma_{p,n})$.
- Verlinde: $N_i = S\Lambda_i S^{-1}$, $N_{ij}^k = \sum_m S_{im}S_{jm}\bar{S}_{km}/S_{0m}$, $\dim V(\Sigma_{p,0}) = \sum_m S_{0m}^{2-2p}$. Ising: $\sigma \times \sigma = \mathbf{1} + \varepsilon$ and $1, 3, 10, 36$ blocks at genus $0, 1, 2, 3$.

- Kac–Moody theories: $\dim V(S^2_R) = \delta_{R,\mathrm{triv}}$, $\dim V(S^2_{R_i,R_j}) = \delta_{R_i,R_j^*}$ (Schur's lemma); $\mathrm{SU}(2)_k$ fusion is Clebsch–Gordan truncated at $j_1 + j_2 + j_3 \le k$.
- Weierstrass: at each point of a genus-p surface exactly p pole orders are impossible; for currents the gap at $n = 1$ is charge conservation.

11.11 Axiomatic approaches I: conformal field theory as geometry on moduli space

So far we have met a conformal field theory in three guises. On the sphere it is a list of primary fields with weights $(h_i, \bar{h}_i)$ and operator product coefficients C_{ij}^k that obey crossing symmetry (Section 11.9). On the torus it is a partition function $Z(\tau)$, a sum of products of characters that is invariant under the modular group (Sections 11.7 and 11.8). On a surface of genus p with n punctures it is a finite-dimensional space of conformal blocks $V(\Sigma_{p,n})$ that varies holomorphically with the shape of the surface (Section 11.10). These are three views of one object. We now ask the question that ties them together:

What is a conformal field theory, if we are not allowed to start from a Lagrangian?

Every answer so far has had the same shape. A CFT *assigns numbers (or operators) to Riemann surfaces*, and the assignment is compatible with cutting a surface into pieces and gluing it back. Two ways of making this precise are mentioned by Nash.

1. *Friedan and Shenker* (this section) look at *all closed surfaces at once*. The partition functions of all genera, and of disconnected surfaces, are one object living on one big space, the **universal moduli space** $\mathcal{R}$. The axioms are statements about the analytic geometry of $\mathcal{R}$: which line bundle the partition function is a section of, and how it behaves where surfaces degenerate.

2. *Segal* (Section 11.12) looks at *surfaces with boundary*. A surface with incoming and outgoing boundary circles is a "spacetime history" of closed strings, and the CFT assigns to it an operator between Hilbert spaces. The axioms say that gluing surfaces corresponds to composing operators.

The first is natural for string theory, whose vacuum amplitude is an integral over moduli spaces. The second is the ancestor of the axioms for topological field theory in Chapter 12.

11.11.1 The string vacuum amplitude as a sum over moduli spaces

The motivating example is the closed bosonic string (Chapter 9). Its vacuum amplitude is Polyakov's path integral over worldsheet metrics g and embeddings $X : \Sigma \to \mathbb{R}^{26}$, summed over the topology of Σ:

$$Z = \sum_{p=0}^{\infty} g_s^{2p-2} \int \frac{\mathcal{D}g\,\mathcal{D}X}{\mathrm{Vol}(\mathrm{Diff} \ltimes \mathrm{Weyl})}\, e^{-S[g,X]}. \tag{11.282}$$

Here p is the genus, g_s is the string coupling (each handle costs a factor g_s^2, and the Euler characteristic is $\chi = 2-2p$), and we divide by the volume of the gauge group of diffeomorphisms

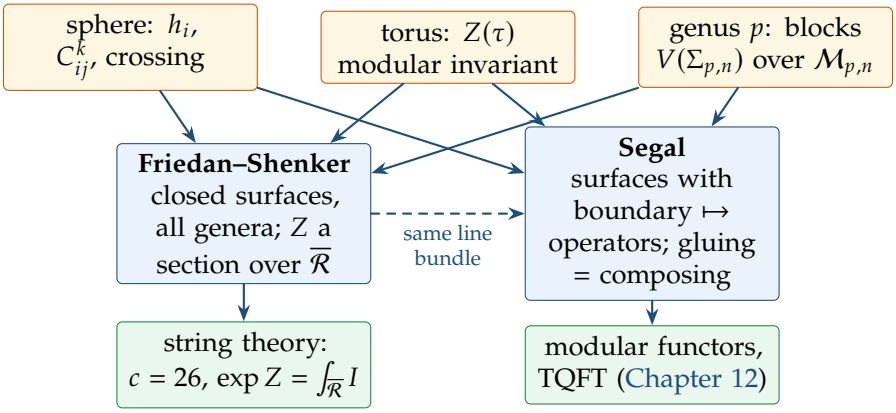


Figure 11.59: The logic of the two axiomatic approaches. What to see: the three concrete pictures of a CFT (top row) are all special cases of "amplitudes assigned to Riemann surfaces, compatible with gluing". Friedan and Shenker package the closed surfaces into one space; Segal uses surfaces with boundary. Both find that the amplitudes are not functions but sections of a line bundle fixed by c.

and Weyl rescalings $g \mapsto e^{2\omega} g$. In chapter 9 we saw that after gauge fixing the infinite-dimensional integral over metrics collapses to a finite-dimensional integral. What survives is exactly the set of metrics that cannot be reached from each other by a diffeomorphism and a Weyl rescaling. A metric up to Weyl rescaling is a complex structure. So what survives is the set of complex structures up to diffeomorphism.

Working definition 11.39 (Moduli space $\mathcal{M}_p$)

$\mathcal{M}_p$ is the set of Riemann surfaces of genus p [surfaces with p handles, with holomorphic coordinates z chosen patch by patch], where two surfaces are the same if a holomorphic bijection maps one onto the other. *Operationally:* $\mathcal{M}_p$ is the list of parameters in the shape of a genus-p surface that no change of coordinates can remove.

- $p = 0$: there is one Riemann sphere; $\mathcal{M}_0$ is a point.
- $p = 1$: a torus is $\mathbb{C}/(\mathbb{Z} + \tau\mathbb{Z})$ with τ in the upper half plane $\mathbb{H}$, and τ, τ' give the same torus iff they differ by $\mathrm{PSL}(2,\mathbb{Z})$ [$\tau' = \frac{a\tau+b}{c\tau+d}$ with integers a, b, c, d and $ad - bc = 1$] (Section 11.7.2). So $\mathcal{M}_1 = \mathbb{H}/\mathrm{PSL}(2,\mathbb{Z})$, one complex parameter, pictured as the fundamental domain (11.193) [$|\operatorname{Re}\tau| \le \frac{1}{2}$, $|\tau| \ge 1$, which holds one τ for each torus, up to edge identifications].
- $p \ge 2$: $\mathcal{M}_p$ has complex dimension $3p - 3$.

Where does $3p - 3$ come from? An infinitesimal change of complex structure is a **Beltrami differential** $\mu = \mu_{\bar z}{}^{z}\, \mathrm{d}\bar z \otimes \partial_z$: it deforms the holomorphic coordinate by $\mathrm{d}z \mapsto \mathrm{d}z + \mu\, \mathrm{d}\bar z$. Changes that come from an infinitesimal diffeomorphism $z \mapsto z + v(z,\bar z)$ have $\mu = \partial_{\bar z} v$, and they do not change the surface. So the tangent space to $\mathcal{M}_p$ is "Beltrami differentials modulo $\bar\partial$ of vector fields", which is the Dolbeault cohomology group $H^1(\Sigma, K^{-1})$ (K is the canonical line bundle, whose sections are 1-forms $f(z)\mathrm{d}z$; K^{-1} is the tangent bundle). Its dimension is computed by Riemann–Roch (Chapter 5): for a line bundle L of degree d on a genus-p surface, $\dim H^0(L) - \dim H^1(L) = d + 1 - p$. Serre duality says $H^1(L) \cong H^0(K \otimes L^{-1})^*$. Then

$$\dim_{\mathbb{C}} T_{[\Sigma]}\mathcal{M}_p \overset{(1)}{=} \dim H^1(\Sigma, K^{-1}) \overset{(2)}{=} \dim H^0(\Sigma, K^2)$$
$$\overset{(3)}{=} \deg(K^2) + 1 - p + \dim H^0(\Sigma, K^{-1}) \overset{(4)}{=} (4p - 4) + 1 - p + 0 = 3p - 3. \qquad (11.283)$$

(1) Tangent vectors to $\mathcal{M}_p$ are Beltrami differentials modulo $\bar\partial v$.
(2) Serre duality with $L = K^{-1}$: $K \otimes L^{-1} = K^2$. Elements of $H^0(K^2)$ are **holomorphic quadratic differentials** $\varphi(z)\,\mathrm{d}z^2$.
(3) Riemann–Roch for $L = K^2$, with $H^1(K^2) \cong H^0(K^{-1})^*$ by Serre duality again.
(4) $\deg K = 2p-2$ (a 1-form has $2p-2$ zeros), so $\deg K^2 = 4p-4$; K^{-1} has negative degree for $p \geq 2$, so it has no holomorphic sections: there are no holomorphic vector fields.

Two facts from this chain are used below. First, the cotangent space to $\mathcal{M}_p$ is the space of holomorphic quadratic differentials, $T^*_{[\Sigma]}\mathcal{M}_p \cong H^0(\Sigma, K^2)$. The stress tensor $T(z)\mathrm{d}z^2$ is a quadratic differential, so it is naturally a cotangent vector to moduli space. Second, for $p = 0, 1$ the last term is not zero: the sphere has three holomorphic vector fields (the Möbius generators L_{-1}, L_0, L_1) and the torus has one (translation), and the formula $\dim H^0(K^2) - \dim H^0(K^{-1}) = 3p-3$ gives $0-3=-3$ and $1-1=0$. The negative count is the number of conformal Killing vectors that must be fixed by fixing the positions of punctures, which is why a sphere needs three punctures before it has any moduli ($\dim \mathcal{M}_{0,n} = n-3$, (11.273)).

After gauge fixing, (11.282) becomes Nash's formula

$$Z = \sum_{p=0}^{\infty} \int_{\mathcal{M}_p} I_p, \tag{11.284}$$

[Nash p. 318, (11.76)], where I_p is a density on $\mathcal{M}_p$ (a top-degree form of real degree $2(3p-3)$), built from the partition function of the matter CFT on Σ, the Faddeev–Popov (ghost) determinant, and the power g_s^{2p-2}, which Nash absorbs into I_p. Each term is one *connected* closed surface. The job of a CFT, from this point of view, is to supply the integrand.

Worked example 11.109: The one-loop integrand of the bosonic string

At $p = 1$ everything is explicit. The integrand is, up to a τ-independent constant (see Polchinski, *String Theory*, §7.3),

$$I_1 = \frac{\mathrm{d}^2\tau}{\tau_2^2}\left(\tau_2^{-12}\,|\eta(\tau)|^{-48}\right), \qquad \tau = \tau_1 + i\tau_2, \quad \mathrm{d}^2\tau = \mathrm{d}\tau_1\,\mathrm{d}\tau_2. \tag{11.285}$$

The factor in brackets is the partition function of the 24 transverse bosons $X^2, \dots, X^{25}$; the ghosts cancel the oscillators of the other two, and the leftover powers of τ_2 combine into the measure $\mathrm{d}^2\tau/\tau_2^2$. We check the two ingredients.
One non-compact boson. Its states are a momentum $k \in \mathbb{R}$ times the oscillator Fock space, with $L_0 = \bar L_0 = k^2/4 + (\text{levels})$ (the normalisation of Section 11.8.5, in which $e^{ik\phi}$ has weight $k^2/4$). The trace (11.199), per unit length of the target line, is

$$Z_{\mathbb{R}}(\tau) \overset{(1)}{=} \int_{-\infty}^{\infty} \frac{\mathrm{d}k}{2\pi}\,(q\bar q)^{k^2/4}\,\frac{1}{|\eta(\tau)|^2} \overset{(2)}{=} \int_{-\infty}^{\infty} \frac{\mathrm{d}k}{2\pi}\,e^{-\pi\tau_2 k^2}\,\frac{1}{|\eta(\tau)|^2} \overset{(3)}{=} \frac{1}{2\pi\sqrt{\tau_2}}\,\frac{1}{|\eta(\tau)|^2}.$$

(1) A continuum of momentum states replaces the sum over (n, w) of the compact boson; the oscillators give $\left|q^{-1/24}\prod_n (1-q^n)^{-1}\right|^2 = |\eta|^{-2}$ as in (11.207).
(2) $q\bar q = \left|e^{2\pi i\tau}\right|^2 = e^{-4\pi\tau_2}$.
(3) Gaussian integral $\int \mathrm{d}k\, e^{-ak^2} = \sqrt{\pi/a}$ with $a = \pi\tau_2$.

Twenty-four copies give $\tau_2^{-12}\,|\eta|^{-48}$, up to constants.
Modular invariance. Under $\tau \mapsto \tau' = (a\tau+b)/(c\tau+d)$ with $ad-bc=1$:

$$\tau_2' \overset{(1)}{=} \mathrm{Im}\,\frac{(a\tau+b)(c\bar\tau+d)}{|c\tau+d|^2} \overset{(2)}{=} \frac{(ad-bc)\,\tau_2}{|c\tau+d|^2} = \frac{\tau_2}{|c\tau+d|^2},$$

$$\mathrm{d}^2\tau' \overset{(3)}{=} \left|\frac{\mathrm{d}\tau'}{\mathrm{d}\tau}\right|^2 \mathrm{d}^2\tau \overset{(4)}{=} \frac{\mathrm{d}^2\tau}{|c\tau+d|^4},$$
$$|\eta(\tau')|^{48} \overset{(5)}{=} |c\tau+d|^{24}\,|\eta(\tau)|^{48}.$$

(1) Multiply numerator and denominator by $c\bar\tau + d$.
(2) $\mathrm{Im}\big[(a\tau+b)(c\bar\tau+d)\big] = \mathrm{Im}\big[ac\,|\tau|^2 + ad\tau + bc\bar\tau + bd\big] = (ad-bc)\tau_2$.
(3) A holomorphic map $\tau \mapsto \tau'$ scales areas by $|\mathrm{d}\tau'/\mathrm{d}\tau|^2$ (its real Jacobian).
(4) $\mathrm{d}\tau'/\mathrm{d}\tau = \big(a(c\tau+d) - c(a\tau+b)\big)/(c\tau+d)^2 = 1/(c\tau+d)^2$.
(5) η^{24} is a modular form of weight 12, (11.213).

So $\mathrm{d}^2\tau/\tau_2^2$ is invariant (it is the hyperbolic area element of $\mathbb{H}$), and $\tau_2^{12}\,|\eta|^{48} \mapsto |c\tau+d|^{-24}\,\tau_2^{12}\,|c\tau+d|^{24}\,|\eta|^{48}$ is invariant too. Hence I_1 is a well-defined density on $\mathcal{M}_1$, and $\int_{\mathcal{M}_1} I_1$ may be computed over the fundamental domain. Modular invariance is not a luxury: without it the integral over $\mathcal{M}_1$ would not even be defined.

Reading the spectrum. Expand in q:

$$|\eta|^{-48} = (q\bar q)^{-1}\big(1 + 24q + 324q^2 + \dots\big)\big(1 + 24\bar q + 324\bar q^2 + \dots\big). \tag{11.286}$$

A term $q^{m-1}\bar q^{n-1}$ is a string state with left and right oscillator levels m, n. In the part of the fundamental domain with $\tau_2 > 1$ the τ_1 integral runs over a full period:

$$\int_{-1/2}^{1/2} \mathrm{d}\tau_1\; q^{m-1}\bar q^{n-1} \overset{(1)}{=} e^{-2\pi\tau_2(m+n-2)} \int_{-1/2}^{1/2} \mathrm{d}\tau_1\, e^{2\pi i\tau_1(m-n)} \overset{(2)}{=} e^{-4\pi\tau_2(m-1)}\,\delta_{mn}.$$

(1) $q^a\bar q^b = e^{2\pi i\tau_1(a-b)}e^{-2\pi\tau_2(a+b)}$.
(2) The integral of $e^{2\pi ik\tau_1}$ over one period is δ_{k0}.

Only states with equal left and right levels survive. This is **level matching**: $L_0 = \bar L_0$, the statement that a closed string has no preferred point. The surviving terms are the tachyon ($m = n = 0$, growing like $e^{4\pi\tau_2}$), then $24\times 24 = 576$ massless states ($m = n = 1$: the graviton, the antisymmetric tensor and the dilaton, in light-cone gauge), then massive levels. The tachyon makes the integral diverge at $\tau_2 \to \infty$. We will see below that $\tau_2 \to \infty$ is a *degenerate* torus, a point we must add to $\mathcal{M}_1$ to compactify it.

11.11.2 Disconnected surfaces and the universal moduli space

The sum (11.284) contains only connected surfaces. In quantum field theory the sum of *all* vacuum diagrams, connected or not, is the exponential of the sum of the connected ones. This is the **linked-cluster theorem**, and the $1/n!$ in the exponential series is the symmetry factor for permuting n identical connected pieces.

Worked example 11.110: Linked clusters in a zero-dimensional field theory

Take one real variable x with "action" $\frac12 x^2 - Jx$. The generating function is

$$\mathcal{Z}(J) \overset{(1)}{=} \frac{\int_{-\infty}^{\infty}\mathrm{d}x\, e^{-x^2/2+Jx}}{\int_{-\infty}^{\infty}\mathrm{d}x\, e^{-x^2/2}} \overset{(2)}{=} e^{J^2/2} \overset{(3)}{=} \sum_{n=0}^{\infty}\frac{1}{n!}\Big(\frac{J^2}{2}\Big)^n.$$

(1) Normalised Gaussian integral with a source J.
(2) Complete the square: $-\frac12 x^2 + Jx = -\frac12(x-J)^2 + \frac12 J^2$, and shift $x \mapsto x + J$.

(3) Exponential series.

The only connected diagram is one propagator joining two sources, with value $W = J^2/2$ (the $\frac{1}{2}$ is its symmetry factor: swapping its two ends). A diagram with n disjoint propagators is W^n, divided by $n!$ because the n identical pieces can be permuted. As a check, the coefficient of J^{2n} is $1/(2^n n!)$, and $\langle x^{2n}\rangle = (2n)!/(2^n n!) = (2n-1)!!$, which is the number of ways of pairing $2n$ points: Wick's theorem. So $\mathcal{Z} = e^W$: *all diagrams* = exp(connected diagrams), with $1/n!$ as the symmetry factor of n identical components.

The string version of this is Nash's

$$\exp[Z] \overset{(1)}{=} \exp\Big[\sum_{p=0}^{\infty}\int_{\mathcal{M}_p} I_p\Big] \overset{(2)}{=} \prod_{p=0}^{\infty}\exp\Big[\int_{\mathcal{M}_p} I_p\Big] \overset{(3)}{=} \prod_{p=0}^{\infty}\sum_{n=0}^{\infty}\frac{1}{n!}\Big\{\int_{\mathcal{M}_p} I_p\Big\}^n, \tag{11.287}$$

[Nash p. 318, (11.77)].

(1) Insert (11.284).
(2) $e^{a+b} = e^a e^b$, applied to every term (formally, for infinitely many terms).
(3) Exponential series for each factor.

The term $\frac{1}{n!}\{\int_{\mathcal{M}_p} I_p\}^n$ is the contribution of n disjoint surfaces, all of genus p. Choosing one term from each factor of the product gives a surface with n_0 spheres, n_1 tori, n_2 genus-two components, and so on. So $\exp Z$ is the sum over *all* closed surfaces, connected or not.

Now comes the geometric idea. Can we write each such term as an integral over one space, the space of all possibly disconnected surfaces? A set of n surfaces of genus p is an *unordered* list of n points of $\mathcal{M}_p$. The space of such lists is the **symmetric product**

$$\mathrm{Sym}^n(\mathcal{M}_p) = \underbrace{\mathcal{M}_p \times \cdots \times \mathcal{M}_p}_{n} \big/ S_n, \tag{11.288}$$

where the permutation group S_n permutes the factors. (By convention Sym^0 is one point: the empty surface.) The integrand on a disconnected surface is the product of the integrands of its components: amplitudes of independent pieces multiply. Then

$$\int_{\mathrm{Sym}^n(\mathcal{M}_p)} I_p \cdots I_p \overset{(1)}{=} \frac{1}{n!}\int_{\mathcal{M}_p\times\cdots\times\mathcal{M}_p} I_p(\Sigma_1)\cdots I_p(\Sigma_n) \overset{(2)}{=} \frac{1}{n!}\Big\{\int_{\mathcal{M}_p} I_p\Big\}^n. \tag{11.289}$$

(1) The map $\mathcal{M}_p^n \to \mathrm{Sym}^n \mathcal{M}_p$ is $n!$-to-one away from the diagonals (where two surfaces coincide), and the diagonals have measure zero. So the integral upstairs counts every point $n!$ times.
(2) Fubini: the integrand is a product of functions of separate variables.

So the $1/n!$ of the linked-cluster theorem is the $1/n!$ of the symmetric product. Friedan and Shenker collect everything into the **universal moduli space**

$$\mathcal{R} = \prod_{p=0}^{\infty}\Big\{\bigcup_{n=0}^{\infty}\mathrm{Sym}^n(\mathcal{M}_p)\Big\}, \tag{11.290}$$

[Nash p. 319, (11.78)]. A point of $\mathcal{R}$ is a choice, for each p, of a finite unordered set of genus-p surfaces (almost all of them empty): that is, one closed Riemann surface with any number of components. With I the product integrand, the full vacuum amplitude is one integral:

$$\int_{\mathcal{R}} I \overset{(1)}{=} \prod_{p=0}^{\infty}\sum_{n=0}^{\infty}\int_{\mathrm{Sym}^n(\mathcal{M}_p)} I_p\cdots I_p \overset{(2)}{=} \prod_{p=0}^{\infty}\sum_{n=0}^{\infty}\frac{1}{n!}\Big\{\int_{\mathcal{M}_p} I_p\Big\}^n \overset{(3)}{=} \exp[Z]. \tag{11.291}$$

(1) An integral over a product of spaces of a product integrand is the product of integrals; an integral over a disjoint union is the sum.
(2) (11.289).
(3) (11.287).

This is Nash's (11.79), $\exp[Z] = \int_{\mathcal{R}} I$. Its content is a change of viewpoint: the basic object is not the sequence of numbers $\int_{\mathcal{M}_p} I_p$ but a single integrand I on the single space $\mathcal{R}$, which is *multiplicative*: on a disconnected surface $\Sigma = \Sigma_1 \sqcup \Sigma_2$ it satisfies $I(\Sigma) = I(\Sigma_1)I(\Sigma_2)$.

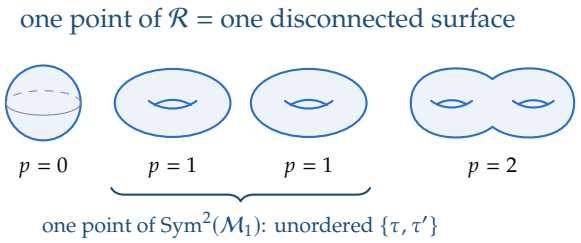


Figure 11.60: A point of the universal moduli space $\mathcal{R}$. What to see: it is a closed surface with several components, here one sphere, two tori and one genus-two surface. The two tori form an unordered pair, a point of $\mathrm{Sym}^2(\mathcal{M}_1)$; the $\frac{1}{2!}$ in $\exp Z$ is the statement that swapping them gives the same point. The integrand I is the product of the integrands of the components.

Watch out: Formal products

The product over p in (11.290) is infinite, and the series in p does not converge (see Section 11.11.6). Equations (11.287)–(11.291) are identities of formal power series in the coupling g_s, hidden inside I_p. They organise the perturbation series; they do not by themselves define a number.

11.11.3 Nodes, stable curves and the compactified moduli space

The moduli spaces $\mathcal{M}_p$ are not compact. The fundamental domain of $\mathcal{M}_1$ runs off to $\tau_2 \to \infty$, and we saw that the string integrand misbehaves exactly there. What does a surface look like near such an "end" of moduli space? The torus with $\tau = i\tau_2$ is a cylinder of circumference 1 and length τ_2 with its ends glued. As $\tau_2 \to \infty$ the cylinder becomes infinitely long and thin *relative to its length*; conformally, a long thin tube is the same as a short tube whose waist shrinks to a point. In the limit a circle on the surface has been pinched to a point. The result is a surface with a **node**.

Working definition 11.40 (Node; plumbing)

A **node** is a point near which the surface looks like two discs glued at their centres: locally $\{(x, y) \in \mathbb{C}^2 : xy = 0\}$, the union of the x-axis and the y-axis. Its smoothing is the **plumbing** $xy = t$ with a small complex number $t \neq 0$: remove the disc $|x| < |t|$ from one side and $|y| < |t|$ from the other, and identify the annuli $|t| < |x| < 1$ and $|t| < |y| < 1$ by $y = t/x$. The **plumbing parameter** t is a local complex coordinate on moduli space transverse to the degenerate surfaces [those with a node. For the torus $t = q = e^{2\pi i\tau}$, so $t \to 0$ is $\tau_2 \to \infty$], and $t = 0$ is the node.

The neck created by plumbing is the annulus $|t| < |x| < 1$. In the coordinate $w = \ln x = u + i\theta$ it is a cylinder of circumference 2π and length $\ell = \ln(1/|t|)$: small t means a long neck. A state propagating down the neck evolves by the cylinder Hamiltonian (11.196) with $L = 2\pi$, $H = L_0 + \bar{L}_0 - \frac{c}{12}$, and the phase of t twists one end relative to the other by the angle $\theta_0 = \arg t$,

which is the momentum operator $P = L_0 - \bar{L}_0$. So the plumbing operator is

$$\begin{aligned} e^{-\ell H}\, e^{i\theta_0 P} &\overset{(1)}{=} \exp\Big[\ln|t|\,\Big(L_0 + \bar{L}_0 - \frac{c}{12}\Big) + i\arg t\,\big(L_0 - \bar{L}_0\big)\Big] \\ &\overset{(2)}{=} \exp\Big[(\ln|t| + i\arg t)\Big(L_0 - \frac{c}{24}\Big) + (\ln|t| - i\arg t)\Big(\bar{L}_0 - \frac{c}{24}\Big)\Big] \\ &\overset{(3)}{=} t^{\,L_0 - c/24}\;\bar{t}^{\,\bar{L}_0 - c/24}. \end{aligned} \tag{11.292}$$

(1) $\ell = -\ln|t|$ and $\theta_0 = \arg t$; the operators commute, so the exponentials combine.
(2) Split $\frac{c}{12} = \frac{c}{24} + \frac{c}{24}$ and regroup the L_0 and $\bar{L}_0$ terms.
(3) $\ln|t| + i\arg t = \ln t$ and $\ln|t| - i\arg t = \ln\bar{t}$.

This is the same operator as $q^{L_0 - c/24}\bar{q}^{\bar{L}_0 - c/24}$ in the torus trace (11.199), and that is no accident.

Worked example 11.111: The torus is a plumbed sphere; q is the plumbing parameter

Take the Riemann sphere with the two points 0 and ∞, with local coordinates $x = z$ near 0 and $y = 1/z$ near ∞. Plumb them with parameter $t = q$: the relation $xy = q$ identifies z (near 0) with the point $1/y = z/q$ (near ∞). The result is the annulus $|q| < |z| < 1$ with its two boundary circles identified by $z \sim z/q$, i.e. the quotient $\mathbb{C}^\times/(z \sim qz)$. Writing $z = e^{2\pi i u}$,

$$\mathbb{C}^\times/(z \sim qz) \overset{(1)}{=} \mathbb{C}\Big/\Big(u \sim u + 1,\; u \sim u + \tfrac{\ln q}{2\pi i}\Big) \overset{(2)}{=} \mathbb{C}/(\mathbb{Z} + \tau\mathbb{Z}), \qquad q = e^{2\pi i\tau}.$$

(1) $z = e^{2\pi i u}$ maps $\mathbb{C}$ onto $\mathbb{C}^\times$ with $u \sim u + 1$; multiplying z by q shifts u by $\ln q/(2\pi i)$.
(2) Definition of τ by $q = e^{2\pi i\tau}$.

So the torus with modulus τ is the sphere with 0 and ∞ plumbed together with parameter $q = e^{2\pi i\tau}$. The end of $\mathcal{M}_1$, $\tau_2 \to \infty$, is $q \to 0$: the torus becomes a sphere with 0 and ∞ glued to each other, a *nodal* torus. Its partition function is dominated by the lowest state through the node. For the Ising model the characters of Section 11.7.5 give

$$\begin{aligned} (q\bar{q})^{1/48} Z_{\text{Ising}} &\overset{(1)}{=} \big|1 + q^2 + \ldots\big|^2 + \big|q^{1/2}(1 + q + \ldots)\big|^2 + \big|q^{1/16}(1 + q + \ldots)\big|^2 \\ &\overset{(2)}{=} 1 + (q\bar{q})^{1/16}\big(1 + q + \bar{q} + \ldots\big) + (q\bar{q})^{1/2}\big(1 + q + \bar{q} + \ldots\big) + q^2 + \bar{q}^2 + \ldots \end{aligned}$$

(1) $\chi_h = q^{h - 1/48}(1 + \ldots)$ with the expansions $q^{1/48}\chi_0 = 1 + q^2 + q^3 + 2q^4 + \ldots$, $q^{1/48}\chi_{1/2} = q^{1/2}(1 + q + q^2 + q^3 + 2q^4 + \ldots)$, $q^{1/48}\chi_{1/16} = q^{1/16}(1 + q + q^2 + 2q^3 + \ldots)$ (Section 11.7.5).
(2) Multiply out and order by powers of $|q|$.

As $q \to 0$: the prefactor $(q\bar{q})^{-c/24}$ is the Casimir energy of the long neck; the coefficient 1 of the leading term is the partition function of the nodal surface (a sphere, normalised to 1); and the corrections $q^h\bar{q}^{\bar{h}}$ list the states that flow through the node: σ at $|q|^{1/8}$, then its descendants, ε at $|q|$, the stress tensor $T = L_{-2}\mathbf{1}$ at q^2.

We can now say precisely what Nash means by the **factorisation condition**. Suppose the pinched circle separates Σ into two pieces Σ_1 (genus p_1) and Σ_2 (genus $p_2 = p - p_1$). Cut along the neck. Each piece, with a disc removed, defines a state in the Hilbert space of the circle: $|\Sigma_2\rangle$, and dually $\langle\Sigma_1|$ (this is Segal's point of view, Section 11.12; on the sphere it is the state–operator correspondence). Inserting a complete orthonormal set of $L_0, \bar{L}_0$ eigenstates $|k\rangle$ between them,

$$\begin{aligned} Z(\Sigma_t) &\overset{(1)}{=} \langle\Sigma_1|\, t^{L_0 - c/24}\,\bar{t}^{\bar{L}_0 - c/24}\,|\Sigma_2\rangle \overset{(2)}{=} |t|^{-c/12}\sum_k t^{h_k}\bar{t}^{\bar{h}_k}\,\langle\Sigma_1|k\rangle\langle k|\Sigma_2\rangle \\ &\overset{(3)}{=} |t|^{-c/12}\Big(Z(\Sigma_1)\,Z(\Sigma_2) + O\big(|t|^{2h_{\min}}\big)\Big). \end{aligned} \tag{11.293}$$

(1) The plumbed surface is Σ_1 and Σ_2 joined by the neck operator (11.292).
(2) Insert $\sum_k |k\rangle\langle k| = \mathbb{1}$; $L_0|k\rangle = h_k|k\rangle$.
(3) For a unitary theory all weights are ≥ 0 and only the vacuum has $h = \bar{h} = 0$, so it dominates as $t \to 0$. The vacuum is the state made by an empty disc, so $\langle \Sigma_1 | 0 \rangle$ is the partition function of Σ_1 with the hole filled in: the closed surface Σ_1. Here $h_{\min}$ is the smallest non-zero value of $(h + \bar{h})/2$.

In words: *the partition function of a surface with a node is the product of the partition functions of the surfaces obtained by pinching*, once the universal factor $|t|^{-c/12}$ (which depends only on c and on the choice of metric near the neck) is stripped off. If the pinched circle does not separate Σ, the same argument gives a trace instead of a matrix element, and the leading term is the partition function of the genus $p-1$ surface obtained by cutting and filling in the two holes; the torus example above is the case $p = 1$. Notice what the singular terms record: the powers t^{h_k} are the conformal weights. The singular behaviour at the node is the spectrum of the theory.

Worked example 11.112: The four-punctured sphere and its three boundary points

Fix four points on the sphere at Nash's positions $(\infty, 1, z, 0)$. The only modulus is the anharmonic ratio z, so $\mathcal{M}_{0,4} = \mathbb{CP}^1 \setminus \{0, 1, \infty\}$. The three missing points are where two punctures collide: $z \to 0$ (the fields at z and 0), $z \to 1$ (at z and 1), $z \to \infty$ (at z and ∞). Colliding punctures is not allowed on a smooth surface. The stable-curve picture replaces the collision by a new sphere that "bubbles off": the two colliding points sit on a small sphere attached by a node to the original one, which keeps the other two points. Each of the two spheres now carries three special points (two punctures and the node). The four-point function near $z = 0$ is, by the block expansion (11.257) with (11.258), for scalar fields,

$$\langle \phi_i(\infty)\phi_j(1)\phi_k(z,\bar{z})\phi_l(0)\rangle \overset{(1)}{=} \sum_m C_{kl}^m C_{ijm}\, \mathcal{F}_{kl}^{ij}(m,z)\, \bar{\mathcal{F}}_{kl}^{ij}(m,\bar{z})$$
$$\overset{(2)}{=} \sum_m C_{kl}^m C_{ijm}\; |z|^{2(h_m - h_k - h_l)} \big(1 + O(z,\bar{z})\big).$$

(1) The OPE of $\phi_k(z)\phi_l(0)$, summed over families m.
(2) Each block starts as $z^{h_m - h_k - h_l}$ times a power series in z.

Two features are Nash's two remarks about the compactification divisor. First, the blocks are singular there, with exponents fixed by the weights h_m of the intermediate families: the boundary is "a locus of singularities". Second, the coefficient of each singular term is the product $C_{kl}^m \cdot C_{ijm}$ of the two three-point functions of the two component spheres of the nodal curve: this is factorisation again. The plumbing parameter at $z = 0$ is z itself. Adding the three points gives the compact space $\overline{\mathcal{M}}_{0,4} = \mathbb{CP}^1$.

Working definition 11.41 (Stable curve; $\overline{\mathcal{M}}_p$)

A **stable curve** of genus p is a connected compact complex curve whose only singularities are nodes, such that every component [a piece left over when every node is pulled apart into two separate points] that is a sphere contains at least three *special points* (nodes or marked points), and every component that is a torus contains at least one. *Operational test:* for each component, remove its special points; the result must have negative Euler characteristic, $2 - 2g_i - n_i < 0$ [g_i is the genus of the component and n_i its number of special points]. Equivalently it has only finitely many automorphisms [holomorphic bijections

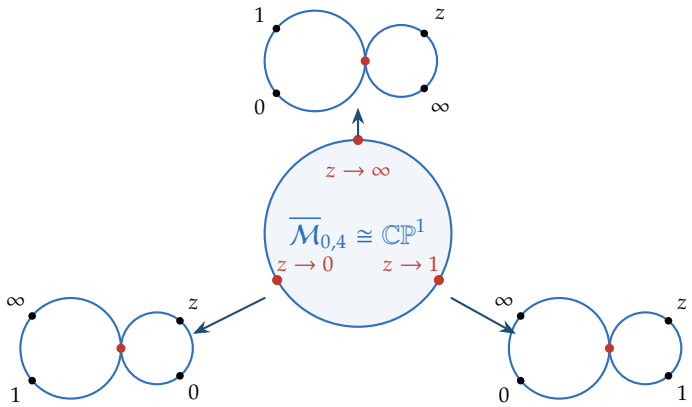


Figure 11.61: The compactified moduli space of four points on a sphere. What to see: the smooth surfaces fill $\mathbb{CP}^1$ minus three points (red). At each red point two punctures would collide; instead a small sphere bubbles off, carrying the two colliding punctures and attached at a node (red dot) to a sphere with the other two. Every component has three special points, so the nodal curves are *stable*. The three points are the three OPE channels (kl), (jk), (ik) of the four-point function, and crossing symmetry relates their expansions.

of the component onto itself that fix each special point], so no continuous symmetries (unfixed conformal Killing vectors) are left over. The **genus** is the genus of any smoothing [the surface obtained by replacing each node $xy = 0$ with a plumbing $xy = t$, $t \neq 0$]. Deligne and Mumford showed that stable curves of genus $p \geq 2$ form a compact space $\overline{\mathcal{M}}_p \supset \mathcal{M}_p$ of the same dimension $3p - 3$, the **moduli space of stable curves**.

The stability condition is exactly the one that makes the plumbing coordinates work: a sphere with fewer than three special points could be rescaled or rotated for free, and it would contribute a spurious continuous symmetry. The added set $\overline{\mathcal{M}}_p \setminus \mathcal{M}_p$ is where at least one plumbing parameter vanishes, one complex equation $t = 0$, so it has complex codimension one. A codimension-one subvariety is a **divisor**; this one is the **compactification divisor** D (often written $\partial\mathcal{M}_p$). It has components labelled by the topology of the pinched circle:

- Δ_0: a non-separating circle is pinched; the generic point is an irreducible curve with one node, whose normalisation has genus $p - 1$ with two points identified;
- Δ_i, $1 \leq i \leq \lfloor p/2 \rfloor$: a separating circle is pinched; the generic point is a genus-i and a genus-$(p - i)$ surface meeting at one node.

Conformal blocks and partition functions are generically singular along D, with singularities $t^{h_k - c/24}$ controlled by the spectrum, as in (11.293). So D is the locus of singularities of the theory, and the singularities are not a nuisance: they encode the spectrum, and their coefficients encode factorisation.

Worked example 11.113: Compactifying $\mathcal{M}_1$: the cusp and the discriminant

The modular function $j(\tau) = q^{-1} + 744 + 196884\,q + \ldots$ is invariant under $\mathrm{PSL}(2,\mathbb{Z})$ and takes every complex value exactly once on the fundamental domain, so $j : \mathcal{M}_1 \to \mathbb{C}$ is a bijection. The missing point $j = \infty$ is $q = 0$, the nodal torus of the plumbing example. So $\overline{\mathcal{M}}_1 = \mathcal{M}_1 \cup \{q = 0\} \cong \mathbb{CP}^1$: we add one point, the *cusp* $\tau = i\infty$ of the fundamental domain. In algebraic terms a torus is a cubic curve $y^2 = 4x^3 - g_2 x - g_3$; it is smooth exactly when

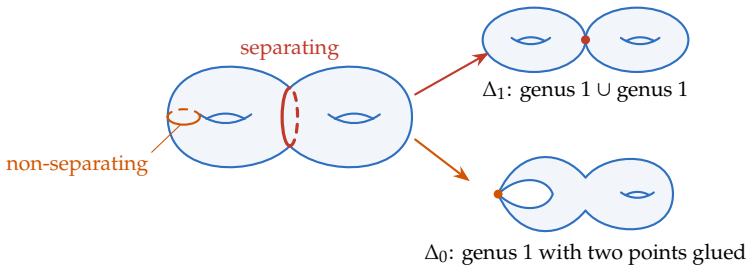


Figure 11.62: The two ways a genus-two surface can degenerate. What to see: pinching the separating circle (red) gives two tori meeting at a node, a point of the boundary component Δ_1; pinching a non-separating circle (orange) gives a torus with two of its points glued, a point of Δ_0. In each case one plumbing parameter t has gone to zero, so the boundary has complex codimension one in $\overline{\mathcal{M}}_2$.

the discriminant $g_2^3 - 27g_3^2$ is non-zero, and that discriminant is $(2\pi)^{12}\eta(\tau)^{24}$. Since

$$\eta(\tau)^{24} = q\prod_{n\geq 1}(1-q^n)^{24}$$

has a simple zero at $q = 0$ and no other zeros, the function η^{24} "detects" the compactification divisor. The string integrand (11.285) contains $|\eta|^{-48}$: it has a pole on the divisor, and (11.286) shows that the pole is the tachyon, the state of negative weight flowing through the node. In a unitary CFT all weights are non-negative, and the only singularity at the node is the universal factor $|q|^{-c/12}$.

When nodes are allowed everywhere, the universal moduli space is replaced by its **stable** version $\overline{\mathcal{R}}$, built from the $\overline{\mathcal{M}}_p$ exactly as in (11.290), and the vacuum amplitude becomes

$$\exp[Z] = \int_{\overline{\mathcal{R}}} I, \tag{11.294}$$

[Nash p. 319, (11.80)]. The gain is conceptual. A nodal surface is on the boundary between connected and disconnected surfaces: pinching a separating circle of a connected genus-p surface produces, after the node is resolved, a disconnected pair. So $\overline{\mathcal{R}}$ glues the pieces $\mathrm{Sym}^n\,\mathcal{M}_p$ of $\mathcal{R}$ together along their boundaries, and the factorisation condition (11.293) says that the integrand I is *continuous* across the gluing (after stripping the universal factor). Multiplicativity on disconnected surfaces and factorisation at nodes become one condition on one integrand on one space.

11.11.4 The stress tensor as a connection on moduli space

So far we have treated the partition function as a *function* on moduli space. For $c \neq 0$ it is not one. To define $Z(\Sigma)$ we must pick a metric in the conformal class of Σ, and the conformal anomaly (Chapter 10; chapter 9 for the string) says that a Weyl rescaling $g \mapsto e^{2\omega}g$ multiplies Z by $\exp\left(\frac{c}{24\pi}S_L[\omega, g]\right)$, where $S_L = \int \mathrm{d}^2x\sqrt{g}\,(g^{ab}\partial_a\omega\partial_b\omega + R\,\omega)$ is the Liouville action. There is no canonical metric on a general Riemann surface that varies holomorphically with moduli. So Z is defined only after a choice, and changing the choice multiplies it by a non-zero factor. An object "defined up to a non-zero factor that depends on local choices" is a *section of a line bundle*. To compare its values at nearby points of moduli space we need a *connection*. The stress tensor supplies it.

Step 1: T is the response to a change of moduli. By definition T_{ab} is the response of the theory to a change of metric. A change of complex structure by a Beltrami differential μ ($\mathrm{d}z \mapsto \mathrm{d}z + \mu\,\mathrm{d}\bar{z}$) is a change of the metric component $g_{z\bar{z}}$, and it couples to $T = T_{zz}$:

$$\delta_\mu \ln Z = \frac{1}{2\pi}\int_\Sigma \mathrm{d}^2 z\,\big(\mu\,\langle T(z)\rangle + \bar{\mu}\,\langle \bar{T}(\bar{z})\rangle\big), \qquad \mathrm{d}^2 z = 2\,\mathrm{d}x\,\mathrm{d}y, \tag{11.295}$$

with the normalisation in which $T(z)\phi(0) \sim h\phi(0)/z^2$.

Worked example 11.114: The torus: $\partial_\tau \ln Z$ is the zero mode of T

On the torus the trace formula (11.199) makes this a one-line computation, treating τ and $\bar{\tau}$ as independent:

$$\partial_\tau \ln Z \overset{(1)}{=} \frac{1}{Z}\,\mathrm{Tr}\Big[2\pi i\Big(L_0 - \frac{c}{24}\Big) q^{L_0 - c/24}\bar{q}^{\bar{L}_0 - c/24}\Big] \overset{(2)}{=} 2\pi i\,\Big\langle L_0 - \frac{c}{24}\Big\rangle_\tau \overset{(3)}{=} \frac{i}{2\pi}\int_0^1 \mathrm{d}x\,\langle T_{\mathrm{cyl}}\rangle_\tau .$$

(1) $\partial_\tau q^a = 2\pi i\, a\, q^a$ with $q = e^{2\pi i \tau}$.
(2) Thermal expectation value on the torus.
(3) (11.195) with $L = 1$: $\int_0^1 \mathrm{d}x\, T_{\mathrm{cyl}} = (2\pi)^2(L_0 - \frac{c}{24})$, all modes $n \neq 0$ integrating to zero.

The change $\tau \mapsto \tau + \delta\tau$ is produced by the constant Beltrami differential $\mu \propto \delta\tau/\tau_2$, and the right-hand side is the integral of $\langle T\rangle$ against it, as (11.295) says. For the Ising model at large τ_2, $\langle L_0 - \frac{c}{24}\rangle \to -\frac{1}{48}$: the Casimir energy.

Step 2: the pairing with Beltrami differentials. Tangent vectors to $\mathcal{M}_p$ are Beltrami differentials modulo $\bar{\partial} v$ (Section 11.11.1). If $\varphi = \varphi(z)\mathrm{d}z^2$ is a *holomorphic quadratic differential*, then $\int \mu\varphi$ depends only on the class of μ:

$$\int_\Sigma \mathrm{d}^2 z\,(\mu + \partial_{\bar{z}} v)\,\varphi \overset{(1)}{=} \int_\Sigma \mathrm{d}^2 z\,\mu\varphi - \int_\Sigma \mathrm{d}^2 z\, v\,\partial_{\bar{z}}\varphi \overset{(2)}{=} \int_\Sigma \mathrm{d}^2 z\,\mu\,\varphi. \tag{11.296}$$

(1) Integrate by parts. $v\varphi\,\mathrm{d}z$ is a globally defined 1-form ($v\partial_z$ is a vector field, $\varphi\,\mathrm{d}z^2$ a quadratic differential), so by Stokes the total derivative integrates to zero on the closed surface Σ.
(2) φ is holomorphic: $\partial_{\bar{z}}\varphi = 0$.

This is the Serre duality pairing between $H^1(\Sigma, K^{-1}) = T\mathcal{M}_p$ and $H^0(\Sigma, K^2) = T^*\mathcal{M}_p$. Each holomorphic quadratic differential is a cotangent vector, i.e. the value at $[\Sigma]$ of a $(1,0)$-form on $\mathcal{M}_p$. So if $\langle T(z)\rangle \mathrm{d}z^2$ were a holomorphic quadratic differential, (11.295) would say that $\partial \ln Z$ is a well-defined 1-form, and Z would be a function.

Step 3: it is not, because of the Schwarzian. Away from insertions $\langle T(z)\rangle$ is holomorphic, but by Nash's (11.81),

$$T'(z')(\mathrm{d}z')^2 = T(z)\,\mathrm{d}z^2 + \frac{c}{12}\{z, z'\}(\mathrm{d}z')^2, \tag{11.297}$$

(our (11.161) with the primes exchanged), it changes by an inhomogeneous term under a change of chart. Now look at the *difference* of two stress tensors T_1, T_2 obeying the same law (for instance $\langle T\rangle$ at two different points of moduli space, written in the same chart, or two different reference choices):

$$\begin{aligned}\big(T_1'(z') - T_2'(z')\big)(\mathrm{d}z')^2 &\overset{(1)}{=} \Big(T_1\mathrm{d}z^2 + \tfrac{c}{12}\{z,z'\}(\mathrm{d}z')^2\Big) - \Big(T_2\mathrm{d}z^2 + \tfrac{c}{12}\{z,z'\}(\mathrm{d}z')^2\Big)\\ &\overset{(2)}{=} \big(T_1(z) - T_2(z)\big)\mathrm{d}z^2.\end{aligned} \tag{11.298}$$

(1) (11.297) for each of T_1 and T_2.
(2) The Schwarzian terms are the same (they depend only on the change of chart and on c), so they cancel.

So the difference is a genuine quadratic differential, a tensor. This is word for word the situation of gauge potentials. With Nash's convention (Chapter T3) $A \mapsto g^{-1}Ag + g^{-1}\mathrm{d}g$, so for two connections

$$A_1' - A_2' \overset{(1)}{=} \left(g^{-1}A_1 g + g^{-1}\mathrm{d}g\right) - \left(g^{-1}A_2 g + g^{-1}\mathrm{d}g\right) \overset{(2)}{=} g^{-1}(A_1 - A_2)\, g,$$

(1) Gauge transformation of each potential.
(2) The inhomogeneous terms $g^{-1}\mathrm{d}g$ cancel.

and $A_1 - A_2$ is a tensor (a 1-form with values in the adjoint bundle), while each A_i on its own is only a connection. By Serre duality (11.296), a difference $T_1 - T_2$ is a 1-form on $\overline{\mathcal{M}}_p$; therefore $\langle T\rangle \mathrm{d}z^2$ itself is a *connection* 1-form on $\overline{\mathcal{M}}_p$, and Z is a section of the line bundle on which it acts. This is Nash's statement on p. 320.

Dictionary: Stress tensor ↔ gauge potential

gauge theory	CFT over moduli space
potential A	$\langle T(z)\rangle\, \mathrm{d}z^2$
gauge transformation g	change of holomorphic chart $z \mapsto z'$
inhomogeneous term $g^{-1}\mathrm{d}g$	$\frac{c}{12}\{z, z'\}(\mathrm{d}z')^2$
$A_1 - A_2$ is a tensor	$T_1 - T_2$ is a quadratic differential
tensor ↔ 1-form	Serre duality: $H^0(K^2) = T^*\mathcal{M}_p$
section of a bundle	partition function Z
curvature	a $(1,1)$-form on $\overline{\mathcal{M}}_p$ proportional to c

The object that transforms like $\frac{c}{12}\{z, z'\}$ has a name.

Working definition 11.42 (Projective connection)

A **projective connection** on a Riemann surface is a choice, in every holomorphic chart z_α, of a holomorphic function $S_\alpha(z_\alpha)$, such that on overlaps

$$S_\beta(z_\beta)\,(\mathrm{d}z_\beta)^2 = S_\alpha(z_\alpha)\,(\mathrm{d}z_\alpha)^2 + \{z_\alpha, z_\beta\}(\mathrm{d}z_\beta)^2.$$

Here $\{z_\alpha, z_\beta\}$ is the Schwarzian derivative [$\frac{z_\alpha'''}{z_\alpha'} - \frac{3}{2}\left(\frac{z_\alpha''}{z_\alpha'}\right)^2$, with primes meaning $\mathrm{d}/\mathrm{d}z_\beta$. It vanishes for Möbius changes of chart]. *Operationally:* it transforms like $\frac{12}{c}T$. The difference of two projective connections is a holomorphic quadratic differential [$q(z)\,\mathrm{d}z^2$ with q holomorphic, changing between charts with no extra term], so on a surface of genus $p \geq 2$ the projective connections form an affine space [fix one of them. Every other one is that one plus a quadratic differential] of complex dimension $\dim H^0(K^2) = 3p - 3$. By (11.297), $\frac{12}{c}\langle T\rangle$ is a projective connection.

For the definition to be consistent, the rule must compose correctly: going from chart z to z' and then to z'' must agree with going directly. The law (11.297) applied twice gives

$$T''(\mathrm{d}z'')^2 \overset{(1)}{=} T'(\mathrm{d}z')^2 + \tfrac{c}{12}\{z', z''\}(\mathrm{d}z'')^2 \overset{(2)}{=} T\mathrm{d}z^2 + \tfrac{c}{12}\Big[\{z, z'\}(\mathrm{d}z')^2 + \{z', z''\}(\mathrm{d}z'')^2\Big],$$

(1) (11.297) for the change $z' \to z''$.

(2) (11.297) for the change $z \to z'$.

while the direct change gives $T\mathrm{d}z^2 + \frac{c}{12}\{z,z''\}(\mathrm{d}z'')^2$. So consistency is the **cocycle identity** of the Schwarzian,

$$\{z,z''\} = \{z,z'\}\Big(\frac{\mathrm{d}z'}{\mathrm{d}z''}\Big)^2 + \{z',z''\}, \qquad \text{i.e.} \qquad \{f\circ g,x\} = \big(\{f,y\}\circ g\big)\,g'(x)^2 + \{g,x\}, \tag{11.299}$$

which is true: it follows from the chain rule by differentiating $f(g(x))$ three times (we checked it symbolically for several f, g). It is the analogue of $(gh)^{-1}\mathrm{d}(gh) = h^{-1}(g^{-1}\mathrm{d}g)h + h^{-1}\mathrm{d}h$, the identity that makes gauge transformations compose. A special case: if g is Möbius, $\{g,x\} = 0$, and the Schwarzian is invariant under Möbius changes of the inner variable.

Worked example 11.115: The same vacuum, two values of $\langle T\rangle$

In the plane the vacuum has $\langle T(z)\rangle = 0$ (rotation and scale invariance forbid anything else for a field of weight 2). Map to the cylinder of circumference L by $z = e^{2\pi w/L}$, as in (11.194). With $k = 2\pi/L$,

$$\langle T_{\mathrm{cyl}}(w)\rangle \overset{(1)}{=} \Big(\frac{\mathrm{d}z}{\mathrm{d}w}\Big)^2\langle T(z)\rangle + \frac{c}{12}\{z,w\} \overset{(2)}{=} 0 + \frac{c}{12}\Big(k^2 - \frac{3}{2}k^2\Big) = -\frac{c}{24}\Big(\frac{2\pi}{L}\Big)^2.$$

(1) (11.161) with the cylinder coordinate w as the new chart.
(2) $z' = kz$, $z'' = k^2 z$, $z''' = k^3 z$ (primes are $\mathrm{d}/\mathrm{d}w$), so $\{z,w\} = k^2 - \frac{3}{2}k^2$.

The state is the same; the number $\langle T\rangle$ is not. It depends on the chart, exactly as a gauge potential depends on the gauge. The physical content is in differences: the energy density $-\frac{\pi c}{6L^2}$ on the cylinder, i.e. the Casimir energy (11.197), is $\frac{1}{2\pi}$ times the constant $\langle T_{\mathrm{cyl}}\rangle + \langle \bar{T}_{\mathrm{cyl}}\rangle$ relative to the plane, and it is measurable.

Worked example 11.116: Canonical projective structures by uniformisation

Is there a natural reference projective connection? Choose charts in which all transition functions are Möbius maps. Then $S_\alpha = 0$ in every chart is consistent, because $\{z_\alpha, z_\beta\} = 0$ for Möbius transitions. Such charts exist on every surface:

- $p = 0$: the charts z and $w = 1/z$ of $\mathbb{CP}^1$; the transition $1/z$ is Möbius.
- $p = 1$: $\mathbb{C}/(\mathbb{Z}+\tau\mathbb{Z})$ with the flat coordinate z; transitions are translations $z \mapsto z+m+n\tau$.
- $p \geq 2$: the uniformisation theorem writes $\Sigma = \mathbb{H}/\Gamma$ with $\Gamma \subset \mathrm{PSL}(2,\mathbb{R})$; the coordinate of $\mathbb{H}$ is a chart with Möbius transitions.

In such a chart $\frac{12}{c}\langle T\rangle$ is a genuine holomorphic quadratic differential, i.e. a cotangent vector to $\mathcal{M}_p$, and (11.295) turns it into the 1-form $\partial \ln Z$ *in the trivialisation defined by the chart*. On the torus this is what we did in the previous example: $\partial_\tau \ln Z$ in the flat chart. Change the family of charts, and $\partial \ln Z$ shifts by a 1-form: a gauge transformation.

The connection is not flat. Its curvature is a $(1,1)$-form on $\overline{\mathcal{M}}_p$ proportional to c, so that Z can be made a *function* only when the relevant power of the line bundle is trivial. Nash phrases this by saying that T is a projective connection and gives rise to a **projective line bundle** E_c over $\mathcal{R}$: "projective" because its transition functions are fixed only up to the c-dependent anomaly factors. Since T is obtained by varying moduli, Z can be viewed as obtained by integrating this variation over $\overline{\mathcal{R}}$: $\ln Z$ is a "potential" whose derivative is the connection 1-form $\langle T\rangle$.

11.11.5 Which line bundle: E_c, holomorphic factorisation and $c = 26$

Friedan and Shenker identify the bundle and state the structure of Z:

Key idea: Friedan–Shenker

The anomaly line bundle is a power of a determinant line bundle, $E_c = (\det)^{c/2}$, and the partition function is the norm squared of a holomorphic section,

$$Z = h(\Psi, \Psi), \qquad \Psi \text{ a holomorphic section of } \mathcal{V} = E_c \otimes W,$$

where W is a vector bundle over $\overline{\mathcal{R}}$ with a projectively flat hermitian metric h. The fibre of W is the space of conformal blocks of the theory (Section 11.10.3), and its projectively flat connection is the one mentioned there.

Here is what the ingredients are, and how to see them in examples.

Determinant lines and modular forms. On a surface of genus p the holomorphic 1-forms form a p-dimensional space $H^0(\Sigma, K)$. Its top exterior power is a complex line, and as Σ varies these lines form the **Hodge line bundle** $\lambda = \det H^0(\Sigma, K)$ over $\mathcal{M}_p$; similarly $\lambda_n = \det H^0(\Sigma, K^n)$. (It is also the determinant line bundle of the $\bar\partial$ operator, the bundle of Chapter 4.) At genus one we can see sections of λ explicitly. The torus $\mathbb{C}/(\mathbb{Z} + \tau\mathbb{Z})$ has the 1-form $\mathrm{d}z$. Under $\tau \mapsto \tau' = (a\tau + b)/(c\tau + d)$, the same torus is $\mathbb{C}/(\mathbb{Z} + \tau'\mathbb{Z})$ with coordinate $z' = z/(c\tau + d)$:

$$\mathbb{Z} + \tau'\mathbb{Z} \overset{(1)}{=} \frac{1}{c\tau + d}\big((c\tau + d)\mathbb{Z} + (a\tau + b)\mathbb{Z}\big) \overset{(2)}{=} \frac{1}{c\tau + d}\big(\mathbb{Z} + \tau\mathbb{Z}\big), \qquad \text{so} \qquad \mathrm{d}z' = \frac{\mathrm{d}z}{c\tau + d}.$$

(1) Put τ' over the common denominator.
(2) $(c\tau + d)m + (a\tau + b)n = (cm + an)\tau + (dm + bn)$, and $\left(\begin{smallmatrix} d & b \\ c & a \end{smallmatrix}\right)$ is invertible over $\mathbb{Z}$, so these run over all of $\mathbb{Z} + \tau\mathbb{Z}$.

So the generator $\mathrm{d}z$ of the fibre of λ rescales by $(c\tau + d)^{-1}$. A section $f(\tau)(\mathrm{d}z)^k$ of λ^k is well defined iff $f(\tau') = (c\tau + d)^k f(\tau)$: *sections of λ^k over $\mathcal{M}_1$ are modular forms of weight k* (modern convention). For example η^{24} is a section of λ^{12}. Since $\mathrm{d}\tau' = \mathrm{d}\tau/(c\tau + d)^2$, Nash's invariant $f(\tau)\mathrm{d}\tau^k$ of (11.54) needs f of modern weight $2k$: that is why η^{24} has "weight 6" in Nash's convention (the pitfall in Section 11.8.2). Nash's modular forms of weight k are sections of $K_{\mathcal{M}_1}^k$, the k-th power of the canonical bundle of $\mathcal{M}_1$.

Worked example 11.117: Holomorphic factorisation of the one-loop string

The integrand (11.285) is, up to a constant,

$$I_1 = \left|\frac{\mathrm{d}\tau}{\eta(\tau)^{24}}\right|^2 \tau_2^{-14}.$$

The first factor is $|F|^2$ with $F = \mathrm{d}\tau/\eta^{24}$ holomorphic (away from the cusp). The second is a positive "metric". Under a modular transformation

$$F(\tau') \overset{(1)}{=} \frac{(c\tau + d)^{-2}\mathrm{d}\tau}{(c\tau + d)^{12}\eta(\tau)^{24}} = (c\tau + d)^{-14}F(\tau),$$
$$\tau_2'^{-14} \overset{(2)}{=} |c\tau + d|^{28}\,\tau_2^{-14},$$

(1) $\mathrm{d}\tau' = (c\tau + d)^{-2}\mathrm{d}\tau$ and η^{24} has weight 12.
(2) $\tau_2' = \tau_2/|c\tau + d|^2$.

so the holomorphic section F and the hermitian metric τ_2^{-14} transform oppositely and $I_1 = h(F,F)$ is invariant. This is the simplest instance of "Z is the norm squared of a holomorphic section". At higher genus the same structure is the theorem of Belavin and Knizhnik: the bosonic string measure on $\mathcal{M}_p$ is $|F_p|^2 (\det \operatorname{Im}\Omega)^{-13}$, with Ω the period matrix and F_p a holomorphic, nowhere-vanishing section of $K_{\mathcal{M}_p} \otimes \lambda^{-13}$.

Why $c = 26$. Why should $K_{\mathcal{M}_p} \otimes \lambda^{-13}$ have a nowhere-vanishing holomorphic section? Two facts.

$$K_{\mathcal{M}_p} = \det T^*\mathcal{M}_p \overset{(1)}{=} \det H^0(\Sigma, K^2) = \lambda_2 \overset{(2)}{=} \lambda^{6\cdot 2^2 - 6\cdot 2 + 1} = \lambda^{13}. \tag{11.300}$$

(1) (11.283): the cotangent space is the space of quadratic differentials, so the top power of $T^*\mathcal{M}_p$ is λ_2.

(2) Mumford's isomorphism $\lambda_n \cong \lambda^{6n^2-6n+1}$ on $\mathcal{M}_p$, $p \geq 2$ (a consequence of Grothendieck–Riemann–Roch applied to the family of curves; Chapter 4 for the index-theoretic viewpoint).

So $K_{\mathcal{M}_p} \otimes \lambda^{-13}$ is trivial, and a trivial bundle has a nowhere-vanishing section. Now count powers. The integrand of a string theory must be a density on $\mathcal{M}_p$, i.e. a section of $K_{\mathcal{M}_p} \otimes \bar{K}_{\mathcal{M}_p}$: a $(3p-3, 3p-3)$-form, which is what one integrates over a complex $(3p-3)$-dimensional space. The ghosts supply the $K_{\mathcal{M}_p}$ (their zero modes are the quadratic differentials). The matter CFT of central charge c contributes a section of the anomaly bundle, a fixed power $\lambda^{-c/2}$. The two combine into something with no leftover anomaly exactly when $c/2 = 13$, i.e. $c = 26$. At $c = 26$, E_c is (up to duality) $\lambda^{13} \cong K_{\mathcal{M}_p}$, the determinant of the cotangent bundle: the partition function has become a density that can be integrated. This is the geometric meaning of the critical dimension.

Watch out: The exponent in $E_c = (\det T^*\mathcal{R})^{c/2}$

Nash writes $E_c = (\det T^*\mathcal{R})^{c/2}$. Read literally, with $\det T^*\mathcal{R}$ the canonical bundle K, this would give K^{13} at $c = 26$, which is not what the argument needs. The consistent reading is that "det" denotes the determinant (Hodge) line bundle λ of the $\bar\partial$ operator, so $E_c = \lambda^{c/2}$ (or its dual, depending on conventions). By Mumford's isomorphism (11.300) this is the same as $E_c = K^{c/26}$, and at $c = 26$ it is the canonical bundle $\det T^*\mathcal{R}$ itself. Keep the formula $E_c \cong \lambda^{c/2}$, and $E_{26} \cong \det T^*\mathcal{R}$.

Worked example 11.118: The Ising model: W, h, and why "projectively" flat

At genus one the blocks of the Ising model are the three characters, and the vector $\chi = (\chi_0, \chi_{1/2}, \chi_{1/16})$ is a holomorphic section of a rank-3 bundle over $\mathcal{M}_1$: modular transformations act on it by the constant matrices S and T of (11.227). With the hermitian metric $h_{ij} = \delta_{ij}$,

$$Z_{\text{Ising}} = h(\chi,\chi) = |\chi_0|^2 + |\chi_{1/2}|^2 + |\chi_{1/16}|^2,$$

Nash's (11.50). Constant transition matrices mean a flat connection, and S, T unitary means h is parallel. Where is the "projective"? The phases of T are $e^{2\pi i(h_i - c/24)}$: the part $e^{2\pi i h_i}$ belongs to the blocks, and the part $e^{-2\pi i c/24}$ belongs to the anomaly bundle E_c (it is the phase of η^{-c} under $\tau \mapsto \tau + 1$, by (11.209)). Strip off E_c by using $T_{\text{top}} = \operatorname{diag}(e^{2\pi i h_i}) = e^{2\pi i c/24}\, T$ instead. Then

$$(ST_{\text{top}})^3 \overset{(1)}{=} e^{2\pi i c/8}\, (ST)^3 \overset{(2)}{=} e^{2\pi i c/8}\, S^2 \overset{(3)}{=} e^{2\pi i/16}\, \mathbb{1}.$$

(1) $T_{\text{top}} = e^{2\pi i c/24} T$ and scalars commute; three factors give $e^{2\pi i c/8}$.

(2) The characters carry a genuine representation of $\mathrm{SL}(2,\mathbb{Z})$, so $(ST)^3 = S^2$ (checked numerically with (11.227)).
(3) For Ising $S^2 = \mathbb{1}$ (every field is self-conjugate) and $c = \frac{1}{2}$.

The relation $(ST)^3 = S^2$ of the modular group now holds only up to the phase $e^{2\pi i c/8}$: after removing E_c, the blocks W carry a *projective* representation, and the connection on W is only projectively flat. The obstruction is a function of c alone. This phase reappears as the framing anomaly of Chern–Simons theory in Chapter 12.

11.11.6 Infinite genus and non-perturbative effects

Nash closes the Friedan–Shenker discussion with a speculation: if $\overline{\mathcal{R}}$ is extended to a space $\overline{\mathcal{R}}_\infty$ that includes surfaces of infinite genus, then non-perturbative effects might be incorporated, by studying $p \to \infty$ and expanding in inverse powers of p. Why would one want this? Because the sum over genera is a power series in g_s^2, and it does not converge. For closed strings the genus-p term grows like $(2p)!$ (Gross and Periwal; Shenker). A series whose terms grow factorially is asymptotic, and its ambiguity is non-perturbative in the coupling.

Worked example 11.119: How small is the ambiguity of a factorially divergent series?

Take $a_n = n!\, g^n$ with $0 < g \ll 1$. The terms decrease while $a_{n+1}/a_n = (n+1)g < 1$, i.e. up to $n_* \approx 1/g$, and increase afterwards. The best one can do is to stop at the smallest term, and the error is of the size of that term:

$$a_{n_*} \overset{(1)}{=} n_*!\, g^{n_*} \overset{(2)}{=} \sqrt{2\pi n_*}\left(\frac{n_*}{e}\right)^{n_*} n_*^{-n_*} \overset{(3)}{=} \sqrt{\frac{2\pi}{g}}\, e^{-1/g}.$$

(1) Stop at $n_* = 1/g$.
(2) Stirling's formula $n! \approx \sqrt{2\pi n}\,(n/e)^n$, and $g = 1/n_*$.
(3) $(n_*/e)^{n_*} n_*^{-n_*} = e^{-n_*} = e^{-1/g}$.

The error $e^{-1/g}$ vanishes faster than any power of g: no finite order of perturbation theory can see it. For strings $a_p \sim (2p)!\, g_s^{2p}$ gives, by the same argument with $n = 2p$, an ambiguity of order e^{-C/g_s}. Such effects are the ones that a large-p analysis of $\overline{\mathcal{R}}_\infty$ was meant to capture; in modern string theory they are produced by D-branes. In the quantum-mechanical double well the analogous effects are the instantons of Chapter 8, with action $\propto 1/g$.

Reading Nash (pp. 318–320, (11.76)–(11.81))

Nash introduces the axiomatic approaches of Friedan–Shenker and Segal, and Friedan–Shenker's *gauge system*: a CFT as analytic geometry on the universal moduli space $\mathcal{R}$, with string theory as the case $c = 26$. (11.76) is our (11.284); (11.77) is (11.287), the exponential of connected diagrams; (11.78) and (11.79) are (11.290) and (11.291), with the $1/n!$ explained by (11.289). The nodes, the factorisation condition, the stable compactification $\overline{\mathcal{M}}_p$, the compactification divisor and (11.80) are in Section 11.11.3, where the plumbing formula (11.292) and the factorisation formula (11.293) are ours. (11.81) is (11.297); the tensor property of differences is (11.298), with the Serre-duality step spelled out in (11.296). The projective connection, E_c, Z as a norm squared of a section of $E_c \otimes W$, and the remark on $\overline{\mathcal{R}}_\infty$ are in Sections 11.11.5 and 11.11.6; see the pitfall there on the exponent of E_c. Nash's footnote 6 says the compactification is the natural one; it is the Deligne–Mumford

compactification by stable curves.

Recap

- The string vacuum amplitude is $Z = \sum_p \int_{\mathcal{M}_p} I_p$, with $\dim_{\mathbb{C}} \mathcal{M}_p = 3p - 3$ (Riemann–Roch for K^2). At genus one the integrand $\tau_2^{-14} \left| \mathrm{d}\tau / \eta^{24} \right|^2$ is modular invariant, and its q-expansion lists the string spectrum with level matching.
- All vacuum diagrams: $\exp Z = \int_{\mathcal{R}} I$ with $\mathcal{R} = \prod_p \bigcup_n \mathrm{Sym}^n \mathcal{M}_p$; the $1/n!$ of the exponential is the $1/n!$ of the symmetric product.
- Ends of moduli space are nodes, reached by plumbing $xy = t$. The neck propagates by $t^{L_0 - c/24} \bar{t}^{\bar{L}_0 - c/24}$; so partition functions factorise at nodes, and the singular exponents are the conformal weights. Stable curves compactify $\mathcal{M}_p$ to $\overline{\mathcal{M}}_p$; the boundary is a divisor, and $\exp Z = \int_{\overline{\mathcal{R}}} I$.
- $\langle T \rangle \mathrm{d}z^2$ transforms with a Schwarzian, so it is a projective connection and, via Serre duality, a connection 1-form on $\overline{\mathcal{M}}_p$; differences are tensors, as for gauge potentials.
- $Z = h(\Psi, \Psi)$ for a holomorphic section Ψ of $E_c \otimes W$; $E_c \cong \lambda^{c/2}$; with Mumford's $K_{\mathcal{M}_p} \cong \lambda^{13}$ the string integrand is a density exactly at $c = 26$. After removing E_c the blocks carry a projective representation, with phase $e^{2\pi i c/8}$.

11.12 Axiomatic approaches II: Segal's definition

Friedan and Shenker look at closed surfaces. But a physicist's first instinct about a quantum theory is Hamiltonian: there is a space of states on a slice of space, and time evolution is an operator. Segal's definition of a conformal field theory takes this instinct seriously for *every* Riemann surface, not only for the cylinder. The question this section answers is:

If a closed string can split and join, what is the "time-evolution operator" of an arbitrary worldsheet, and what rules must these operators obey?

The answer: a surface Σ whose boundary consists of q incoming and r outgoing circles defines an operator $U_\Sigma : H^{\otimes q} \to H^{\otimes r}$, and gluing surfaces along circles composes the operators. That is all. Everything we have met (the Hamiltonian, the vacuum, the OPE, the torus partition function, modular invariance, the conformal blocks) is a special case, and the rules of the game are those of a *functor* between two categories. We build up to this in steps, starting with ordinary quantum mechanics.

11.12.1 Warm-up: quantum mechanics as a rule for intervals

In Euclidean quantum mechanics, a time interval of length t is turned into the operator $U_t = e^{-tH}$ on the Hilbert space H. Three properties make this assignment work.

1. *Gluing is composition.* Putting an interval of length s after one of length t gives an interval of length $s + t$, and $U_s U_t = U_{s+t}$.
2. *Closing is a trace.* Gluing the two ends of an interval of length β gives a circle, and the circle is assigned the number $\mathrm{Tr}\, U_\beta = \mathrm{Tr}\, e^{-\beta H}$: the thermal partition function.

3. *Kernels are path integrals.* In the position basis U_t has a kernel $K_t(x_0, x_t) = \langle x_t | e^{-tH} | x_0 \rangle$, and Feynman's formula writes it as an integral over paths $x(s)$, $0 \le s \le t$, with fixed endpoints: $K_t(x_0, x_t) = \int_{[x_0, x_t]} e^{-S[x]} \mathcal{D}x$.

Worked example 11.120: Gluing two free-particle intervals

For a free particle ($m = \hbar = 1$, $H = -\frac{1}{2}\partial_x^2$) the Euclidean kernel is the heat kernel $K_t(x, y) = (2\pi t)^{-1/2} e^{-(x-y)^2/2t}$. Gluing means integrating over the position y at the junction:

$$\begin{aligned}\int_{-\infty}^{\infty} \mathrm{d}y\, K_s(x,y) K_t(y,z) &\overset{(1)}{=} \frac{1}{2\pi\sqrt{st}} \int \mathrm{d}y\, \exp\Big[-\frac{(x-y)^2}{2s} - \frac{(y-z)^2}{2t}\Big] \\ &\overset{(2)}{=} \frac{1}{2\pi\sqrt{st}}\, e^{-\frac{(x-z)^2}{2(s+t)}} \int \mathrm{d}y\, \exp\Big[-\frac{s+t}{2st}\big(y - y_*\big)^2\Big] \\ &\overset{(3)}{=} \frac{1}{2\pi\sqrt{st}}\, e^{-\frac{(x-z)^2}{2(s+t)}} \sqrt{\frac{2\pi st}{s+t}} = K_{s+t}(x,z).\end{aligned}$$

(1) Insert the kernels.
(2) Complete the square in y, with $y_* = (tx + sz)/(s+t)$; the y-independent remainder is $-(x-z)^2/2(s+t)$.
(3) Gaussian integral $\int \mathrm{d}y\, e^{-a(y-y_*)^2} = \sqrt{\pi/a}$ with $a = (s+t)/2st$.

The composition law $U_s U_t = U_{s+t}$ holds kernel by kernel: gluing intervals is integrating over the field at the junction.

Worked example 11.121: Closing the interval: the oscillator on a circle

For the harmonic oscillator $H = \frac{1}{2}(p^2 + x^2)$ the kernel is Mehler's

$$K_t(x,y) = \frac{1}{\sqrt{2\pi \sinh t}} \exp\Big[-\frac{(x^2+y^2)\cosh t - 2xy}{2\sinh t}\Big].$$

Gluing the ends means setting $y = x$ and integrating:

$$\begin{aligned}\operatorname{Tr} e^{-tH} &\overset{(1)}{=} \int \mathrm{d}x\, K_t(x,x) \overset{(2)}{=} \frac{1}{\sqrt{2\pi\sinh t}} \int \mathrm{d}x\, \exp\Big[-x^2\,\frac{\cosh t - 1}{\sinh t}\Big] \\ &\overset{(3)}{=} \frac{1}{\sqrt{2\pi\sinh t}} \sqrt{\frac{\pi \sinh t}{\cosh t - 1}} \overset{(4)}{=} \frac{1}{2\sinh(t/2)}.\end{aligned} \tag{11.301}$$

(1) The trace of an operator with kernel K is $\int K(x,x)\,\mathrm{d}x$.
(2) At $y = x$ the exponent is $-(2x^2\cosh t - 2x^2)/(2\sinh t)$.
(3) Gaussian integral with $a = (\cosh t - 1)/\sinh t$.
(4) $\cosh t - 1 = 2\sinh^2(t/2)$.

Compare the sum over the spectrum $E_n = n + \frac{1}{2}$: $\sum_n e^{-t(n+1/2)} = e^{-t/2}/(1 - e^{-t}) = 1/(2\sinh(t/2))$. The same number, computed once as a path integral on a circle and once as a trace over states. We will use (11.301) mode by mode for the free boson below.

11.12.2 Surfaces with boundary and their operators

A closed string at an instant is a circle. So in string theory the analogue of an endpoint of an interval is a boundary circle, and the analogue of an interval is a Riemann surface Σ with

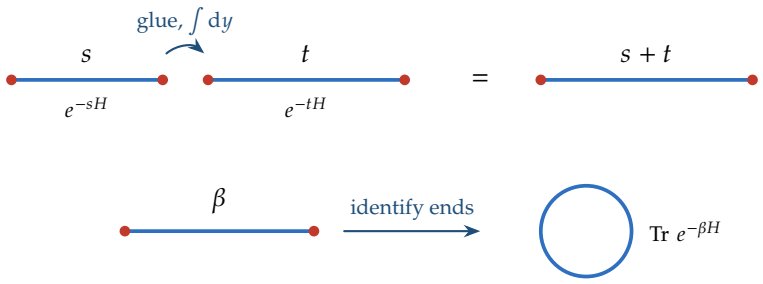


Figure 11.63: Quantum mechanics as a rule for intervals. What to see: an interval of length t is assigned the operator e^{-tH}; gluing endpoints (integrating over the position at the junction) composes operators; gluing the two ends of one interval into a circle takes the trace. Segal's axioms are the same three rules with intervals replaced by Riemann surfaces and endpoints replaced by boundary circles.

boundary $\partial\Sigma = S^1 \sqcup \cdots \sqcup S^1$, n circles in all. Each boundary circle comes with a *parametrisation*, an identification with the standard circle $\{|z| = 1\}$ (this is what lets us glue two circles in a definite way), and with a label: *incoming* or *outgoing*, fixed by whether the orientation induced from Σ agrees or disagrees with the standard one.

Physical picture: Punctures versus boundary circles

Nash remarks that a surface with n boundary circles is, up to homotopy, a surface with n punctures. Topologically: Σ minus an open disc deformation-retracts onto Σ minus a point, since the punctured disc $\{0 < |z| \le 1\}$ retracts onto its boundary circle, so the two have the same topology. Physically: a puncture at which a local operator ϕ sits, together with a local coordinate z vanishing there, is the same thing as a boundary circle $\{|z| = 1\}$ carrying the state $|\phi\rangle = \phi(0)|0\rangle$. This is the state–operator correspondence of radial quantisation (Nash §3, earlier in this chapter). A local coordinate at a puncture is exactly the data needed to parametrise the small circle around it.

From the string point of view q of the circles are strings coming in and $r = n - q$ are strings going out, and the history in between should be described by an operator

$$U_{\Sigma,\partial\Sigma} : \underbrace{H \otimes \cdots \otimes H}_{q \text{ factors}} \longrightarrow \underbrace{H \otimes \cdots \otimes H}_{r \text{ factors}}, \tag{11.302}$$

[Nash p. 320, (11.82)]. Here H is the Hilbert space of one closed string (the space of states of the CFT on one circle). For a CFT the operator depends only on the conformal structure of Σ and on the boundary parametrisations, not on a metric; but because of the anomaly, only up to a scalar factor, as we will see in Section 11.12.5.

11.12.3 The Hilbert space of a string, and the kernel of U

What is H, concretely? Nash follows the path-integral picture. Take the simplest surface with boundary, $\Sigma_{p,2}$: genus p with two boundary circles, one incoming and one outgoing. A moment in the history of the string is a loop, so the configuration space is a space of loops, and a history is a path of loops,

$$\alpha : [0, t] \longrightarrow L\Sigma_{p,2} = \mathrm{Map}(S^1, \Sigma_{p,2}), \tag{11.303}$$

with $\alpha(0) = \alpha_0$ and $\alpha(t) = \alpha_t$ the two boundary circles and t the duration of the evolution, [Nash p. 320, (11.83)]. The quantum states are wave functions on the configuration space,

$$H = L^2(L\Sigma_{p,2}), \tag{11.304}$$

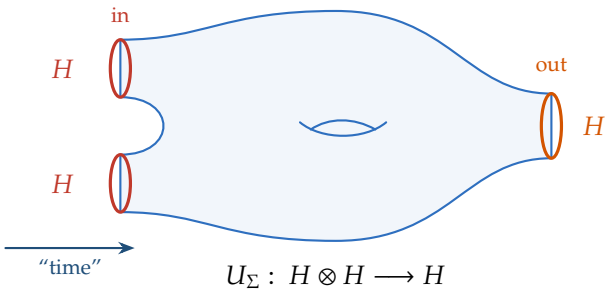


Figure 11.64: A Segal morphism. What to see: a genus-one surface with two incoming boundary circles (red) and one outgoing circle (orange). Two closed strings come in, join, and one leaves, with a loop (the handle) in between. The CFT assigns to it a linear map $H \otimes H \to H$. With no handle this surface is the "pair of pants", and its map is the operator product expansion.

(11.84). If S is a conformally invariant action (for example the string action), the propagation kernel from the loop α_0 to the loop α_t is the path integral over all histories with those ends,

$$K(\alpha_0, \alpha_t) = \int_{[\alpha_0, \alpha_t]} e^{-S(\alpha)}\, \mathcal{D}\alpha, \tag{11.305}$$

(11.85), and the operator acts on wave functions by

$$\big(U_{\Sigma_{p,2}}\psi\big)(\alpha_t) = \int K(\alpha_0, \alpha_t)\, \psi(\alpha_0)\, \mathcal{D}\alpha_0, \tag{11.306}$$

(11.86). These are the formulas of the quantum-mechanics warm-up, with a point x replaced by a loop α: (11.305) is Feynman's kernel and (11.306) is "multiply by the kernel and integrate".

Watch out: What the loops are, and what L^2 means

Two cautions. (i) In a sigma model with target X (spacetime, for the string) the field on a circle is a loop *in* X, $x : S^1 \to X$, and the wave functions live on the loop space LX; the surface $\Sigma_{p,2}$ is what the moving loop sweeps out. Nash's $L\Sigma_{p,2}$ should be read in this spirit: an element records how a circle sits in the history. For a general CFT there is no target at all, and "the field configuration on a circle" is replaced by an abstract state. (ii) There is no Lebesgue measure on an infinite-dimensional loop space, so neither L^2 nor $\mathcal{D}\alpha$ is literally defined. Segal's axioms are designed to avoid both: they postulate the space H and the operators U_Σ directly, and (11.304)–(11.306) become a heuristic for constructing examples. For free fields the heuristic can be made exact, because every integral is Gaussian, as the next example shows.

Worked example 11.122: The disc amplitude is the vacuum wave functional

Take a free boson ϕ with action $S = \frac{1}{2}\int \mathrm{d}^2x\, |\nabla\phi|^2$, and let the unit disc D have its boundary circle outgoing. Expand the boundary value in Fourier modes,

$$\phi(\sigma) = \frac{1}{\sqrt{2\pi}} \sum_{n\in\mathbb{Z}} \phi_n\, e^{in\sigma}, \qquad \phi_{-n} = \overline{\phi_n}.$$

The disc defines a state: a wave functional of the boundary loop, $\Psi_D[\phi] = \int_{\phi|_{\partial D}=\phi} e^{-S}$. The integral is Gaussian, so up to a constant it is $e^{-S_{\mathrm{cl}}}$, where S_{cl} is the action of the classical solution with the given boundary value. That solution is harmonic, $\phi_{\mathrm{cl}} = \frac{1}{\sqrt{2\pi}} \sum_n \phi_n r^{|n|} e^{in\sigma}$,

so

$$S_{\mathrm{cl}} \overset{(1)}{=} \frac{1}{2}\int_D \big(\nabla\cdot(\phi_{\mathrm{cl}}\nabla\phi_{\mathrm{cl}}) - \phi_{\mathrm{cl}}\nabla^2\phi_{\mathrm{cl}}\big) \overset{(2)}{=} \frac{1}{2}\oint_{\partial D}\phi_{\mathrm{cl}}\,\partial_r\phi_{\mathrm{cl}}\,\mathrm{d}\sigma$$
$$\overset{(3)}{=} \frac{1}{2}\int_0^{2\pi}\mathrm{d}\sigma\,\frac{1}{2\pi}\sum_{m,n}\phi_m\phi_n\,|n|\;e^{i(m+n)\sigma} \overset{(4)}{=} \frac{1}{2}\sum_{n\in\mathbb{Z}}|n|\,\big|\phi_n\big|^2. \tag{11.307}$$

(1) $\big|\nabla\phi\big|^2 = \nabla\cdot(\phi\nabla\phi) - \phi\nabla^2\phi$.
(2) Stokes (the divergence theorem) on the disc, and $\nabla^2\phi_{\mathrm{cl}} = 0$; the outward normal is ∂_r.
(3) At $r = 1$, $\partial_r r^{|n|} = |n|$.
(4) $\int_0^{2\pi} e^{i(m+n)\sigma}\mathrm{d}\sigma = 2\pi\delta_{m,-n}$ and $\phi_{-n} = \overline{\phi_n}$.

So $\Psi_D[\phi] \propto \exp\big(-\sum_{n\geq 1} n\,\big|\phi_n\big|^2\big)$. Compare the Hamiltonian on the circle: $H = \frac{1}{2}\int_0^{2\pi}\mathrm{d}\sigma\,\big(\Pi^2 + (\partial_\sigma\phi)^2\big) = \frac{1}{2}\sum_n\big(|\Pi_n|^2 + n^2\,\big|\phi_n\big|^2\big)$. Each mode $n \neq 0$ is a harmonic oscillator of frequency $|n|$, whose ground state is $\exp(-\frac{1}{2}\,|n|\,\big|\phi_n\big|^2)$ per real degree of freedom; the pair $\pm n$ gives $\exp(-n\,\big|\phi_n\big|^2)$. This is Ψ_D. *The disc is the vacuum*: an empty disc creates $|0\rangle$, the state–operator correspondence for the identity operator. The ground energy is the sum of zero-point energies,

$$E_0 = \sum_{n\neq 0}\frac{|n|}{2} = \sum_{n\geq 1} n \;\to\; \zeta(-1) = -\frac{1}{12} = -\frac{c}{12}, \qquad c = 1,$$

with the zeta-function value $\zeta(-1) = -\frac{1}{12}$ (the regularisation that respects the conformal symmetry). This is the Casimir energy (11.197) for $L = 2\pi$.

11.12.4 Gluing the ends: the trace condition and modular invariance

Join the two boundary circles of $\Sigma_{p,2}$ to each other. The result is a closed surface $\tilde{\Sigma}_{p,2}$ of genus $p+1$. Exactly as closing an interval gave a trace, Segal requires

$$\mathrm{tr}\big(U_{\Sigma_{p,2}}\big) = U_{\tilde{\Sigma}_{p,2}}, \tag{11.308}$$

[Nash p. 321, (11.87)], where the operator of a closed surface is a map $\mathbb{C}\to\mathbb{C}$, i.e. a number: its partition function. For this to make sense $U_{\Sigma_{p,2}}$ must be trace class, which it is whenever the surface has non-zero area (the factor q^{L_0} below damps high-energy states).

The case $p = 0$ is the one we know. $\Sigma_{0,2}$ is a cylinder, or conformally an annulus $A_q = \{|q| \leq |z| \leq 1\}$, with $|z| = 1$ incoming and $|z| = |q|$ outgoing, the outgoing circle parametrised by $\sigma \mapsto q e^{i\sigma}$. Evolution from the outer to the inner circle is the plumbing operator (11.292):

$$U_{A_q} = q^{\,L_0 - c/24}\,\bar{q}^{\,\bar{L}_0 - c/24}. \tag{11.309}$$

Gluing the two circles of A_q (with the rotation by $\arg q$ built into the parametrisation) gives the torus with modulus τ, $q = e^{2\pi i\tau}$, exactly as in the plumbing example of Section 11.11.3. So (11.308) is

$$Z(\tau) = \mathrm{tr}\;U_{A_q} = \mathrm{Tr}\;q^{L_0 - c/24}\bar{q}^{\bar{L}_0 - c/24},$$

the trace formula (11.199). The trace condition *is* the statement "periodic Euclidean time gives a trace".

Worked example 11.123: The free-boson annulus, mode by mode

For the free boson and $\tau = i\tau_2$ (so $q = e^{-t}$ with $t = 2\pi\tau_2$ the length of a cylinder of circumference 2π), the annulus operator is e^{-tH}, and H is a sum of independent oscillators of frequency n, two for each $n \geq 1$ (the modes $\pm n$), plus the zero mode. The trace factorises into Mehler traces (11.301):

$$\mathrm{Tr}_{\mathrm{osc}}\, e^{-tH} \overset{(1)}{=} \prod_{n\geq 1}\Big(\frac{1}{2\sinh(nt/2)}\Big)^2 \overset{(2)}{=} \prod_{n\geq 1}\frac{e^{-nt}}{(1-e^{-nt})^2} \overset{(3)}{=} e^{-t\sum_{n\geq 1} n}\prod_{n\geq 1}\frac{1}{(1-q^n)^2}$$
$$\overset{(4)}{=} q^{-1/12}\prod_{n\geq 1}\frac{1}{(1-q^n)^2} \overset{(5)}{=} \frac{1}{|\eta(i\tau_2)|^2}.$$

(1) Independent modes: the trace of a tensor product is the product of traces; each oscillator of frequency n evolves for time t, i.e. (11.301) with $t \to nt$.
(2) $2\sinh(x/2) = e^{x/2}(1-e^{-x})$, squared.
(3) Collect the prefactors; $e^{-nt} = q^n$.
(4) $\sum_{n\geq 1} n = \zeta(-1) = -\frac{1}{12}$, so $e^{t/12} = q^{-1/12}$.
(5) For real q, $|\eta|^2 = \eta^2 = q^{1/12}\prod(1-q^n)^2$.

The zero mode adds the factor $(2\pi\sqrt{\tau_2})^{-1}$ of the example in Section 11.11.1. So Segal's trace condition, evaluated by gluing Mehler kernels, reproduces the partition function of the free boson.

Modular invariance now comes out as a *consistency condition*. The axioms say that $U_{\tilde{\Sigma}}$ depends only on the closed surface, not on how it was cut open. A torus can be cut along either of its two cycles. Cutting along the a-cycle gives the annulus A_q, $q = e^{2\pi i\tau}$; cutting along the b-cycle gives an annulus $A_{\tilde{q}}$ with $\tilde{q} = e^{-2\pi i/\tau}$, because the roles of space and time are exchanged and the modulus becomes $-1/\tau$. The two traces must agree:

$$\mathrm{Tr}\; q^{L_0 - c/24}\bar{q}^{\bar{L}_0 - c/24} = \mathrm{Tr}\; \tilde{q}^{\,L_0 - c/24}\bar{\tilde{q}}^{\,\bar{L}_0 - c/24}, \qquad \text{i.e.} \qquad Z(\tau) = Z(-1/\tau). \tag{11.310}$$

Similarly, cutting along the a-cycle but choosing a different parametrisation (a Dehn twist) replaces τ by $\tau + 1$. So Segal's axioms imply Nash's (11.51). The same mechanism at higher genus is the requirement that $U_{\tilde{\Sigma}}$ be invariant under the mapping class group, the large diffeomorphisms of $\tilde{\Sigma}$.

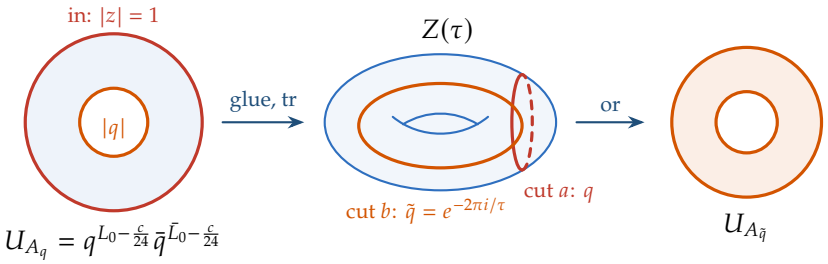


Figure 11.65: Why Segal's axioms imply modular invariance. What to see: the torus (middle) can be cut open along the a-cycle (red), giving the annulus A_q (left), or along the b-cycle (orange), giving the annulus $A_{\tilde{q}}$ with $\tilde{q} = e^{-2\pi i/\tau}$ (right). The trace condition applied to the two annuli must give the same number, since it is the same torus: that is $Z(\tau) = Z(-1/\tau)$.

11.12.5 Sewing surfaces and the multiplier λ

The trace condition glues two circles of *one* surface. The other basic operation glues two *different* surfaces. If $\Sigma_{p,2}$ and $\Sigma'_{p,2}$ each have one incoming and one outgoing circle, sew the outgoing circle of Σ' to the incoming circle of Σ (using the parametrisations to match points). The result is a surface $\Sigma_{p,2} \cup \Sigma'_{p,2}$ with again two boundary circles, and Segal requires

$$U_{\Sigma_{p,2}} \circ U_{\Sigma'_{p,2}} = \lambda\, U_{\Sigma_{p,2} \cup \Sigma'_{p,2}}, \tag{11.311}$$

[Nash p. 321, (11.88)], with λ a non-zero scalar. Why not simply $\lambda = 1$? In quantum mechanics $U_s U_t = U_{s+t}$ exactly. Here the answer is the central charge.

Worked example 11.124: Round annuli compose exactly

The round annuli form a semigroup: sewing $A_{q'}$ (outer circle $|z| = 1$, inner $|z| = |q'|$) to A_q by rescaling the second one by q' gives the annulus $\{|qq'| \le |z| \le 1\} = A_{qq'}$. And

$$U_{A_q}\, U_{A_{q'}} \overset{(1)}{=} q^{L_0 - c/24} \bar{q}^{\bar{L}_0 - c/24}\, q'^{L_0 - c/24} \bar{q}'^{\bar{L}_0 - c/24} \overset{(2)}{=} (qq')^{L_0 - c/24} (\overline{qq'})^{\bar{L}_0 - c/24} = U_{A_{qq'}}.$$

(1) (11.309) for each annulus.
(2) L_0 and $\bar{L}_0$ commute with each other and with themselves, so the powers add.

Here $\lambda = 1$, because the normalisation $-c/24$ was chosen consistently (it is the Casimir energy of the flat cylinder metric, which is additive in the length).

For general surfaces there is no such preferred normalisation. Deform the boundary parametrisations of an annulus by diffeomorphisms of the circle. The "annuli with arbitrary boundary parametrisation" form a semigroup $\mathcal{A}$ which contains, as its boundary, the group $\mathrm{Diff}(S^1)$ (the infinitely thin annuli), and $\mathcal{A}$ is a kind of complexification of $\mathrm{Diff}(S^1)$. Segal's axioms give a map $A \mapsto U_A$, and (11.311) says it is a representation up to scalars: $U_A U_{A'} = \lambda(A, A') U_{AA'}$. Infinitesimally, $\mathrm{Diff}(S^1)$ is generated by the vector fields $\ell_n = -z^{n+1}\partial_z$ (Section 11.5.1), represented by the L_n, and the projective factor becomes the central term of the Virasoro algebra (11.174). Can a clever choice of phases remove it? Try shifting L_0 by a constant a, which is all the freedom one has (the central term only appears in $[L_m, L_{-m}]$):

$$[L_m, L_{-m}] \overset{(1)}{=} 2mL_0 + \frac{c}{12}(m^3 - m) \overset{(2)}{=} 2m\,(L_0 + a) + \Big(\frac{c}{12}(m^3 - m) - 2ma\Big). \tag{11.312}$$

(1) The Virasoro algebra (11.174) with $n = -m$.
(2) Add and subtract $2ma$.

In terms of $L'_0 = L_0 + a$ the central term is $\frac{c}{12}m^3 - (\frac{c}{12} + 2a)m$. The part linear in m can be removed by $a = -c/24$ (and this is exactly the shift to $L_0 - c/24$ that appears on the cylinder). The cubic part $\frac{c}{12}m^3$ cannot be removed by any choice. So for $c \neq 0$ the representation of $\mathrm{Diff}(S^1)$, and hence of the annulus semigroup, is genuinely projective, and $\lambda \neq 1$ in general. For $m = 0, \pm 1$ the central term $m^3 - m$ vanishes: the Möbius subgroup is represented honestly, as we knew.

Key idea: The multiplier is the anomaly

The scalar λ in (11.311) is the finite, global form of the central charge. In Segal's precise formulation each surface carries a complex line Det_Σ (built from determinant lines, raised to a power fixed by c), sewing multiplies these lines, and U_Σ is an element of $\mathrm{Hom}(H^{\otimes q}, H^{\otimes r}) \otimes \mathrm{Det}_\Sigma$. Choosing a trivialisation of each line produces the numbers λ. As Σ varies over moduli space the lines Det_Σ form a line bundle: the anomaly bundle E_c of Friedan and Shenker (Section 11.11.5). The two axiom systems meet here.

11.12.6 The category of circles and surfaces

All the rules so far are instances of one structure. Segal organises them as follows [Nash p. 321].

Working definition 11.43 (The Segal category C)

- *Objects:* finite disjoint unions of parametrised circles [each circle comes with a chosen identification with the standard circle $|z| = 1$], $S^1 \sqcup \cdots \sqcup S^1$, including the empty union $\varnothing$.
- *Morphisms* from X to Y: Riemann surfaces Σ whose boundary is X (incoming) together with Y (outgoing), with the parametrisations, up to biholomorphisms that respect them [invertible holomorphic maps between surfaces that match the boundary parametrisations point by point].
- *Composition:* if $\Sigma' : X \to Y$ and $\Sigma : Y \to Z$, then $\Sigma \circ \Sigma' : X \to Z$ is the surface obtained by sewing Σ to Σ' along their common boundary Y [each outgoing circle of Σ' is glued to the matching incoming circle of Σ, point by point through the parametrisations].
- *Disjoint union* $\sqcup$ of objects and of morphisms makes C a *monoidal* category [circles and surfaces can be placed side by side, and adding $\varnothing$ changes nothing], with unit $\varnothing$.

A **conformal field theory** in Segal's sense is then a (projective, monoidal) **functor** U from C to a category $\mathcal{H}$ of topological vector spaces:

- to k circles it assigns $U(S^1 \sqcup \cdots \sqcup S^1) = H^{\otimes k}$, and to $\varnothing$ the ground field $\mathbb{C}$;
- to a morphism $\Sigma : X \to Y$ it assigns the operator $U_\Sigma : U(X) \to U(Y)$ of (11.302);
- it respects composition up to the multiplier, $U_{\Sigma\circ\Sigma'} = \lambda^{-1} U_\Sigma U_{\Sigma'}$, as in (11.311);
- it respects disjoint union, $U_{\Sigma\sqcup\Sigma'} = U_\Sigma \otimes U_{\Sigma'}$ (independent systems have tensor-product state spaces and amplitudes);
- gluing an outgoing circle of Σ to an incoming circle of the same surface is a partial trace, as in (11.308).

Nash calls such a functor a **modular functor**. (In later usage that name is reserved for the chiral structure of Section 11.12.7; the full theory is "a CFT in Segal's sense".) A closed surface is a morphism $\varnothing \to \varnothing$, so U_Σ is a linear map $\mathbb{C} \to \mathbb{C}$: a number, the partition function. The multiplicativity of Friedan and Shenker on disconnected surfaces is the disjoint-union axiom.

Worked example 11.125: The Ising pair of pants

Take the pants of Figure 11.66 with round boundary circles in standard position (the three-punctured sphere with punctures at $0, 1, \infty$ and the unit-disc coordinates there). Its operator $U_P : H \otimes H \to H$ is known once we know its matrix elements between primaries, because the descendants are fixed by the Virasoro Ward identities (Section 11.9.2). With the unit-normalised primaries $\mathbf{1}, \sigma, \varepsilon$ of the Ising model,

$$\langle \varepsilon | U_P | \sigma \otimes \sigma \rangle \overset{(1)}{=} \langle \varepsilon(\infty)\sigma(1)\sigma(0)\rangle \overset{(2)}{=} C_{\sigma\sigma\varepsilon} = \tfrac{1}{2},$$
$$\langle \mathbf{1} | U_P | \sigma \otimes \sigma \rangle \overset{(3)}{=} C_{\sigma\sigma\mathbf{1}} = 1, \qquad \langle \sigma | U_P | \sigma \otimes \sigma \rangle \overset{(4)}{=} 0.$$

(1) Matrix elements of the pants between primary states are three-point functions with

surface (morphism)	operator U_Σ	physics
disc, boundary outgoing: $\varnothing \to S^1$	$\mathbb{C} \to H$, $1 \mapsto \lvert 0\rangle$	the vacuum (example of Section 11.12.3)
disc with ϕ at the centre	$1 \mapsto \lvert\phi\rangle = \phi(0)\lvert 0\rangle$	state–operator correspondence
disc, boundary incoming: $S^1 \to \varnothing$	$H \to \mathbb{C}$, $\psi \mapsto \langle 0\rvert\psi\rangle$	the vacuum bra
annulus A_q: $S^1 \to S^1$	$q^{L_0 - c/24}\bar{q}^{\bar{L}_0 - c/24}$	Hamiltonian and momentum; plumbing
pair of pants: $S^1 \sqcup S^1 \to S^1$	$H \otimes H \to H$	the OPE; matrix elements C_{ijk}
sphere with four holes	$H^{\otimes 3} \to H$	four-point function; its two cuttings give crossing
torus: $\varnothing \to \varnothing$	tr U_{A_q}	$Z(\tau)$, modular invariant
closed genus p: $\varnothing \to \varnothing$	a number	$Z(\Sigma)$, integrand I_p of (11.284)

Table 11.5: The Segal functor on the basic surfaces. What to see: every object of the earlier sections (vacuum, states, evolution, OPE, correlation functions, partition functions) is the operator of some surface.

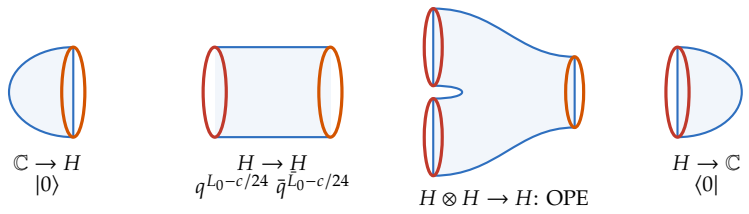


Figure 11.66: The building blocks of the Segal category. What to see: the disc with outgoing boundary (orange) creates the vacuum, the cylinder evolves, the pair of pants multiplies two states (the OPE), and the cap with incoming boundary (red) projects on the vacuum. Every Riemann surface with boundary can be sewn from these pieces (a pants decomposition plus discs), so the whole functor is determined by H, the annulus operators and the pants operator, subject to consistency.

the punctures at $\infty, 1, 0$ (the out-state at ∞).

(2) (11.248) at these points is the constant C_{ijk}; $C^2_{\sigma\sigma\varepsilon} = \frac{1}{4}$ from the exact four-point function, (11.265), with the sign convention $C > 0$.

(3) The two-point normalisation $\langle\sigma\sigma\rangle = 1$.

(4) The $\mathbb{Z}_2$ symmetry $\sigma \to -\sigma$: an odd number of spins has zero correlator; equivalently $N^{\sigma}_{\sigma\sigma} = 0$.

The non-zero pattern of the pants matrix elements is exactly the fusion rule $\sigma \times \sigma = \mathbf{1} + \varepsilon$ of (11.268). So the fusion rules are the "support" of the pants operator.

Worked example 11.126: Associativity of sewing is crossing symmetry

A sphere with four holes can be sewn from two pairs of pants in two ways: glue along a circle separating holes $\{3,4\}$ from $\{1,2\}$, or along a circle separating $\{2,3\}$ from $\{1,4\}$ (Figure 11.67). The functor axioms say both give the same operator. Evaluating each one by inserting a complete set of states in the sewn circle,

$$\sum_{m} C^{m}_{kl}\, C_{ijm}\, \left|\mathcal{F}^{ij}_{kl}(m,z)\right|^2 = \sum_{m'} C^{m'}_{jk}\, C_{il\,m'}\, \left|\mathcal{F}^{il}_{jk}(m',1-z)\right|^2,$$

(for scalar fields) which is the crossing relation of Section 11.9.4, checked for the Ising spin four-point function in (11.264). *Functoriality contains the bootstrap.* Moore and Seiberg showed that, conversely, consistency of sewing on the four-holed sphere and on the one-holed torus is enough to guarantee consistency on every surface.

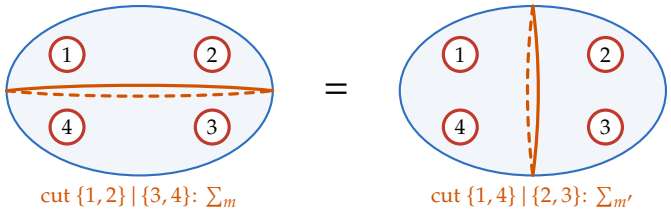


Figure 11.67: The four-holed sphere sewn from two pairs of pants in two ways. What to see: the orange circle is where the two pants are sewn, and where a complete set of intermediate states m (left) or m' (right) is inserted. Since both give the same surface, the two sums are equal: crossing symmetry is associativity of sewing.

Watch out: No identity morphism

A category needs identity morphisms, and $\mathcal{C}$ has none: the identity would be an annulus of zero width, which is not a Riemann surface. The annuli A_q with $q \to 1$ approximate it, and $U_{A_q} \to \mathbb{1}$ strongly (on each state, $q^{L_0} \to 1$). Strictly $\mathcal{C}$ is a "category without units", or one adds the thin annuli by hand. This is a technical point; it does not affect anything above.

11.12.7 Modular functors: where blocks and fusion rules live

Nash ends with the remark that in Segal's framework concrete notions such as conformal blocks and fusion rules can be identified. They live in the *chiral* half of the theory. The full amplitudes U_Σ are real-analytic in the moduli; but, as the block decomposition (11.257) showed, they are sums of products of holomorphic and antiholomorphic pieces. The holomorphic pieces, for a surface whose boundary circles are *labelled* by primaries, span finite-dimensional vector spaces.

The rules they obey define a (chiral) modular functor.

Working definition 11.44 (Modular functor (chiral data of a rational CFT))

Fix the finite set of primaries $\{0 = \mathbf{1}, 1, \dots, F-1\}$ with the conjugation $i \mapsto i^*$ [ϕ_{i^*} is the unique primary with $\langle \phi_i \phi_{i^*} \rangle \neq 0$]. A **modular functor** assigns to each surface Σ with boundary circles labelled $i_1, \dots, i_n$ a finite-dimensional vector space $V(\Sigma; i_1, \dots, i_n)$ (the conformal blocks), such that:

1. *Disjoint union:* $V(\Sigma \sqcup \Sigma') = V(\Sigma) \otimes V(\Sigma')$.
2. *Gluing:* if Σ is cut along a circle into Σ_{cut} with two new boundary circles, $V(\Sigma) \cong \bigoplus_m V(\Sigma_{\text{cut}}; m, m^*)$ ((11.274)).
3. *Normalisation:* $\dim V(S^2; i) = \delta_{i0}$ and $\dim V(S^2; i, j) = \delta_{ij^*}$: Nash's (11.75), (11.281).
4. *Holomorphy:* as Σ varies, the spaces form a holomorphic vector bundle over moduli space with a projectively flat connection [a rule for carrying blocks along paths in moduli space. Around a small closed loop a block comes back multiplied by a number]; its monodromy [what carrying the blocks once around a non-contractible loop does to them] gives a projective action of the mapping class group [diffeomorphisms of Σ modulo those that can be deformed to the identity, $\mathrm{SL}(2,\mathbb{Z})$ for the torus] on each $V(\Sigma)$.

The fusion numbers are $N_{ij}^k = \dim V(S^2; i, j, k^*)$, (11.272).

Worked example 11.127: The torus has one block per primary

Cut the torus along a non-separating circle. What is left is an annulus, i.e. a sphere with two holes, with labels m and m^*:

$$\dim V(T^2) \overset{(1)}{=} \sum_{m=0}^{F-1} \dim V(S^2; m, m^{**}) \overset{(2)}{=} \sum_{m=0}^{F-1} \delta_{m,m} = F.$$

(1) The gluing axiom, with the new circles labelled m and m^*.
(2) The normalisation axiom: $\dim V(S^2; i, j) = \delta_{ij^*}$ with $i = m$, $j = m^*$, and $m^{**} = m$.

So $V(T^2)$ has dimension equal to the number of primaries, and a basis is given by the characters $\chi_m(\tau)$: for the Ising model, $F = 3$ and $V(T^2) = \mathrm{span}(\chi_0, \chi_{1/2}, \chi_{1/16})$. The mapping class group of the torus is $\mathrm{SL}(2,\mathbb{Z})$, and it acts on $V(T^2)$ by the matrices S and T of (11.227), projectively after the c-dependent phase is removed (the example at the end of Section 11.11.5). The same gluing rule with Verlinde's formula gave $\dim V(\Sigma_2) = 10$ for the Ising model at genus two, (11.279).

The full CFT is recovered from the modular functor by "gluing the two halves": the Hilbert space is $H = \bigoplus_i H_i \otimes \bar{H}_i$ (11.200), and each amplitude is a hermitian pairing of a holomorphic block with an antiholomorphic one, $U_\Sigma = \sum h^{I\bar{J}} \mathcal{F}_I \bar{\mathcal{F}}_{\bar{J}}$, as $Z_{\text{Ising}} = \sum_i |\chi_i|^2$ and (11.257). This is the same statement as Friedan–Shenker's $Z = h(\Psi, \Psi)$.

Forget the complex structure altogether and keep only the topology. What remains is a rule that assigns finite-dimensional vector spaces to topological surfaces and linear maps to three-dimensional cobordisms between them: a **topological quantum field theory**. Atiyah's axioms for it (Chapter 12) are Segal's with "Riemann surface" replaced by "manifold" and with no anomaly multiplier, up to framing. Witten showed that Chern–Simons theory with

gauge group G at level k is such a theory, and that its state space on a surface Σ is the space of conformal blocks $V(\Sigma)$ of the G Kac–Moody (WZW) model at level k: the modular functor of this chapter is the Hilbert space of the next.

	quantum mechanics	Segal CFT / modular functor	Atiyah TQFT (Chapter 12)
space	a point	a circle (parametrised)	closed surface Σ
spacetime	interval $[0,t]$	Riemann surface with boundary	3-manifold with $\partial M = \Sigma$
state space	H (∞-dim.)	$H = \bigoplus_i H_i \otimes \bar{H}_i$ (CFT); $V(\Sigma)$ finite-dim. (chiral)	$V(\Sigma)$, finite-dim.
amplitude	e^{-tH}	U_Σ; blocks $\mathcal{F} \in V(\Sigma)$	vector $Z(M) \in V(\Sigma)$
gluing	$U_s U_t = U_{s+t}$	sewing, (11.311); trace, (11.308)	gluing 3-manifolds
anomaly	none	multiplier λ, central charge c	framing phase $e^{2\pi i c/24}$
closed case	Tr $e^{-\beta H}$	$Z(\Sigma)$, e.g. $Z(\tau)$	$Z(M)$, knot invariants

Table 11.6: Three "functorial" quantum theories. What to see: the same three rules (spacetimes give operators, gluing composes, closing traces) organise quantum mechanics, conformal field theory and topological field theory; only the geometry of "space" and "spacetime" changes.

Reading Nash (pp. 320–321, (11.82)–(11.88))

Segal's approach: a surface whose boundary is n circles (homotopically, n punctures), q incoming and r outgoing, gives $U_{\Sigma,\partial\Sigma} : H^{\otimes q} \to H^{\otimes r}$, (11.82) = (11.302). For $\Sigma_{p,2}$ Nash describes H as L^2 of the loop space, with the history α (11.83), H (11.84), the kernel (11.85) and the action of U (11.86): our (11.303)–(11.306), with the caveats of the pitfall in Section 11.12.3 and the exactly solvable free-boson examples added. The trace condition (11.87) is (11.308); for $p = 0$ it gives the torus partition function and modular invariance, which we derived from cutting the torus two ways. The sewing condition (11.88) is (11.311), with λ explained by the Virasoro central term. The category C, the functor to topological vector spaces, and the identification of blocks and fusion rules are in Sections 11.12.6 and 11.12.7.

Recap

- Segal's CFT is quantum mechanics with intervals replaced by Riemann surfaces: a surface with q incoming and r outgoing parametrised circles gives $U_\Sigma : H^{\otimes q} \to H^{\otimes r}$.
- Heuristically $H = L^2(\text{loops})$ and U has a path-integral kernel; for free fields this is exact: the disc gives the vacuum functional $e^{-\sum n|\phi_n|^2}$, and gluing Mehler kernels gives $|\eta|^{-2}$.
- Closing a surface is a trace (11.87); for the annulus $U = q^{L_0 - c/24}\bar{q}^{\bar{L}_0 - c/24}$ this is $Z(\tau)$, and cutting the torus two ways gives $Z(\tau) = Z(-1/\tau)$.
- Sewing composes operators up to a scalar λ (11.88); λ is the global form of the Virasoro central term, whose m^3 part cannot be removed.

- The whole structure is a projective monoidal functor from the category of circles and surfaces; disc = vacuum, pants = OPE, associativity of sewing = crossing.
- The chiral data form a modular functor: finite-dimensional block spaces $V(\Sigma)$ with gluing and normalisation axioms (11.75); $\dim V(T^2) = F$, and the mapping class group acts by S, T. Its topological shadow is the TQFT of Chapter 12.

11.13 The c-theorem: the flow runs downhill

The renormalisation group was built in the toolkit chapter Chapter T5. Recall its picture in three sentences. Lowering the cutoff moves a theory through the space of all couplings; the fixed points of this flow have no length scale, and in two dimensions (for unitary theories with a discrete spectrum) they are conformal field theories (Section T5.5.4). Near a fixed point the flow leaves along relevant directions and comes in along irrelevant ones (Figure T5.11), and a quantum field theory is a trajectory from one fixed point to another (Section T5.6). This section adds the one global fact about the flow that is special to conformal field theory: it runs downhill.

Integrating out throws information away (Section T5.1.1), so one expects a long-distance theory to have "fewer degrees of freedom" than the short-distance theory it came from. For a two-dimensional CFT there is a natural count of degrees of freedom: the central charge. A free boson has $c = 1$ and a free Majorana fermion $c = \frac{1}{2}$ (Section 11.3.13); n decoupled copies of anything have n times the central charge; and c fixes the Casimir energy $-\pi c/6L$ of a system on a circle of length L and the growth of the number of high-energy states (Cardy's formula, $\ln \rho(E) \simeq 2\pi\sqrt{cLE/6\pi}$ at large E, Section 11.8.6). Zamolodchikov (1986) turned the expectation into a theorem: there is a function on theory space that equals c at every fixed point and decreases along every RG flow. Sources: Nakayama (arXiv:1302.0884, §5.1) for the construction, Komargodski (arXiv:1112.4538, eq. (1.8)) for the sum rule, and Komargodski–Schwimmer (arXiv:1107.3987) for the four-dimensional analogue.

What we take from Chapter T5. In complex coordinates $z = x^1 + ix^2$ put $T = -2\pi T_{zz}$ and $\Theta = -2\pi T_{z\bar{z}} = -\frac{\pi}{2} T^{\mu}{}_{\mu}$ (Eq. (T5.94)); at a fixed point $\Theta = 0$ and $T = T(z)$ is the holomorphic stress tensor of Eq. (11.76). Conservation reads $\bar{\partial} T + \partial \Theta = 0$. Rotation invariance fixes the two-point functions up to three functions of $r^2 = z\bar{z}$,

$$\langle T(z,\bar{z})T(0)\rangle = \frac{F}{z^4}, \qquad \langle T(z,\bar{z})\Theta(0)\rangle = \frac{G}{z^3\bar{z}}, \qquad \langle \Theta(z,\bar{z})\Theta(0)\rangle = \frac{H}{z^2\bar{z}^2}$$

(Eq. (T5.96)), and conservation, inserted in these correlators, gives (Eqs. (T5.97) and (T5.98)), with dots meaning $\mathrm{d}/\mathrm{d}\ln r^2$,

$$\dot{F} = -\dot{G} + 3G, \qquad \dot{G} - G = -\dot{H} + 2H.$$

At a scale-invariant point F, G, H are constants, these force $G = H = 0$, and reflection positivity turns $\langle \Theta\Theta \rangle = 0$ into $\Theta = 0$ (Section T5.5.4). Away from a fixed point the same two equations contain more.

11.13.1 Zamolodchikov's C-function

Keep this notation, now in a theory that is *not* at a fixed point, so that F, G, H depend on $r^2 = z\bar{z}$. Dots denote $\mathrm{d}/\mathrm{d}\ln r^2$. The conservation equations (T5.97) and (T5.98) are two first-order

equations for three functions. Look for a combination whose derivative contains no G and no $\dot{H}$.

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\ln r^2}\big(2F + aG + bH\big) &\overset{(1)}{=} 2(-\dot{G} + 3G) + a\dot{G} + b\dot{H} = (a-2)\dot{G} + 6G + b\dot{H} \\ &\overset{(2)}{=} (a-2)(G - \dot{H} + 2H) + 6G + b\dot{H} \\ &= (a+4)\,G + (b - a + 2)\,\dot{H} + 2(a-2)\,H. \end{aligned} \tag{11.313}$$

(1) (T5.97): $\dot{F} = -\dot{G} + 3G$.
(2) (T5.98): $\dot{G} = G - \dot{H} + 2H$.

The G term vanishes for $a = -4$, and then the $\dot{H}$ term vanishes for $b = a - 2 = -6$. The last coefficient is $2(-6) = -12$.

Key idea: Zamolodchikov's c-theorem

Define

$$C := 2F - 4G - 6H. \tag{11.314}$$

Then

$$\frac{\mathrm{d}C}{\mathrm{d}\ln r^2} = -12\,H = -12\,r^4\langle\Theta(x)\Theta(0)\rangle \;\le\; 0. \tag{11.315}$$

At a fixed point $C = c$, the central charge. Along an RG flow from a CFT with central charge c_{UV} (reached as $r \to 0$) to one with c_{IR} (reached as $r \to \infty$), C decreases monotonically, so

$$c_{\mathrm{UV}} \;\ge\; c_{\mathrm{IR}},$$

with equality only if $\Theta = 0$ throughout, that is, if there is no flow.

Three ingredients make this work, and each is something we have already proved.

- *The sign.* $H = r^4\langle\Theta\Theta\rangle \ge 0$ by reflection positivity, (T5.99). This is where unitarity enters.
- *The fixed-point value.* At a fixed point $\Theta = 0$, so $G = H = 0$, and $\langle T(z)T(0)\rangle = \frac{c/2}{z^4}$ (Eq. (11.89)) gives $F = \frac{c}{2}$ and $C = 2F = c$.
- *A function on theory space.* C is dimensionless, so it can depend on r only through dimensionless combinations of r and the couplings: $C = C(\hat{g}_i(r))$, the couplings measured in units of the distance r at which the correlators are evaluated. (The stress tensor is conserved, so it needs no renormalisation and has no anomalous dimension; Nakayama eqs. (5.8)–(5.10) show that the Callan–Symanzik equation then removes any other r dependence.) So (11.315) says C decreases along the flow in theory space.

Physical picture: why "fewer degrees of freedom"

$C(r)$ measures the strength of the TT correlation at separation r, that is, how many independent fluctuating fields carry energy and momentum at that scale. A field of mass m contributes at $r \ll 1/m$ and stops contributing at $r \gg 1/m$, where its correlations die exponentially. So $C(r)$ counts the modes that are still "light" at scale r, and that number can only go down as r grows. The next example shows this happening explicitly.

Trace and β-function. Let the theory be a fixed point perturbed by one operator, $S = S_{\mathrm{CFT}} + g\int \mathrm{d}^2x\,\Phi(x)$, with Φ of dimension $\Delta = 2 - y$, so g has mass dimension y. The trace

of the stress tensor is the response of the action to a uniform stretch of all lengths (this is the two-dimensional case of Eq. (T5.91), repeated here with the normalisation of this chapter). With the stress tensor defined by the response to the metric, $\delta S = \frac{1}{2}\int \sqrt{g}\, T_{\mu\nu}\delta g^{\mu\nu}$ (for the free boson this reproduces Example 11.24), a stretch $g_{\mu\nu} \to (1+2\sigma)g_{\mu\nu}$ gives $\delta S = -\sigma \int T^{\mu}{}_{\mu}$. The CFT part does not respond (in flat space), while in the perturbation $\sqrt{g} \to (1+2\sigma)\sqrt{g}$ and the primary $\Phi \to (1-\Delta\sigma)\Phi$:

$$-\sigma \int T^{\mu}{}_{\mu} \overset{(1)}{=} g\,(2-\Delta)\,\sigma \int \mathrm{d}^2x\,\Phi \quad\Longrightarrow\quad T^{\mu}{}_{\mu} \overset{(2)}{=} -y\,g\,\Phi = -\beta(g)\,\Phi, \qquad \Theta \overset{(3)}{=} \frac{\pi}{2}\,\beta(g)\,\Phi.$$

(1) Collect the factors of $1+\sigma$ from the measure and from Φ.
(2) $2-\Delta = y$; and $y\,g$ is the classical β-function (T5.6) of g ($\beta = \mathrm{d}\hat{g}/\mathrm{d}t$ with t the logarithm of the length scale). In the quantum theory the full β appears; this is the operator form of the Callan–Symanzik equation.
(3) $\Theta = -\frac{\pi}{2}T^{\mu}{}_{\mu}$.

Check on the massive free boson $S = \frac{1}{4\pi}\int[(\partial\varphi)^2 + m^2\varphi^2]$: here $\Phi = \varphi^2$, $g = m^2/4\pi$, $y = 2$, so $T^{\mu}{}_{\mu} = -m^2\varphi^2/2\pi$. Directly, $T_{\mu\nu} = \frac{1}{2\pi}\big[\partial_\mu\varphi\partial_\nu\varphi - \frac{1}{2}\delta_{\mu\nu}((\partial\varphi)^2 + m^2\varphi^2)\big]$ has trace $\frac{1}{2\pi}[(\partial\varphi)^2 - (\partial\varphi)^2 - m^2\varphi^2] = -m^2\varphi^2/2\pi$. They agree.

So $H = \frac{\pi^2}{4}\beta^2 r^4\langle\Phi(x)\Phi(0)\rangle$. Write $r^4\langle\Phi\Phi\rangle =: \mathsf{G}(\hat{g})$, the (dimensionless) **Zamolodchikov metric**, normalised so that $\mathsf{G} = 1$ at the UV fixed point when $\langle\Phi(x)\Phi(0)\rangle = |x|^{-2\Delta}$ and g is measured in units of r. With $t = \ln r$ (so $\mathrm{d}\ln r^2 = 2\,\mathrm{d}t$), (11.315) becomes

$$\frac{\mathrm{d}C}{\mathrm{d}t} = -24H = -6\pi^2\,\mathsf{G}(\hat{g})\,\beta(\hat{g})^2. \tag{11.316}$$

Along the flow $\mathrm{d}t = \mathrm{d}\hat{g}/\beta$, so

$$c_{\mathrm{UV}} - c_{\mathrm{IR}} = 6\pi^2 \int_{\hat{g}_{\mathrm{UV}}}^{\hat{g}_{\mathrm{IR}}} \mathsf{G}(\hat{g})\,\beta(\hat{g})\,\mathrm{d}\hat{g}. \tag{11.317}$$

11.13.2 The sum rule, and a free massive boson

Integrating (11.315) over all scales gives a formula for the total drop that needs no knowledge of the couplings:

$$\begin{aligned} c_{\mathrm{UV}} - c_{\mathrm{IR}} &\overset{(1)}{=} 12\int_0^\infty H\;\mathrm{d}\ln r^2 \overset{(2)}{=} 24\int_0^\infty \frac{\mathrm{d}r}{r}\,r^4\,\frac{\pi^2}{4}\,\langle T^{\mu}{}_{\mu}(x)T^{\mu}{}_{\mu}(0)\rangle \\ &\overset{(3)}{=} \frac{6\pi^2}{2\pi}\int \mathrm{d}^2x\;|x|^2\langle T^{\mu}{}_{\mu}(x)T^{\mu}{}_{\mu}(0)\rangle = 3\pi\int \mathrm{d}^2x\;|x|^2\,\langle T^{\mu}{}_{\mu}(x)T^{\mu}{}_{\mu}(0)\rangle. \end{aligned} \tag{11.318}$$

(1) Integrate (11.315) from $r = 0$ ($C = c_{\mathrm{UV}}$) to $r = \infty$ ($C = c_{\mathrm{IR}}$).
(2) $\mathrm{d}\ln r^2 = 2\mathrm{d}r/r$ and $\Theta = -\frac{\pi}{2}T^{\mu}{}_{\mu}$.
(3) $\mathrm{d}^2x = 2\pi r\,\mathrm{d}r$ for a rotation-invariant integrand, so $\int r^3\mathrm{d}r\,f = \frac{1}{2\pi}\int \mathrm{d}^2x\,|x|^2 f$.

This is Cardy's sum rule; Komargodski derives it (his (1.8)) from the dilaton effective action, a different route to the same formula.

Worked example 11.128: A massive boson flows from $c = 1$ to $c = 0$

Take $S = \frac{1}{4\pi}\int[(\partial\varphi)^2 + m^2\varphi^2]$. At short distances it is the free boson ($c = 1$); at long distances nothing is left ($c = 0$). So the sum rule must give exactly 1. We found $T^{\mu}{}_{\mu} = -\frac{m^2}{2\pi}\varphi^2$. The

kinetic operator is $\frac{1}{2\pi}(k^2+m^2)$ in momentum space, so the propagator is $\tilde G(k) = 2\pi/(k^2+m^2)$ (in position space $K_0(m|x|)$, which tends to the massless $-\ln|x|$ of Eq. (11.69) as $m \to 0$, up to a constant). By Wick's theorem $\langle\varphi^2(x)\varphi^2(0)\rangle_{\text{conn}} = 2\,G(x)^2$, so

$$\mathcal{F}(x) := \langle T^{\mu}{}_{\mu}(x) T^{\mu}{}_{\mu}(0)\rangle = \frac{m^4}{4\pi^2}\cdot 2G(x)^2.$$

It is easiest in momentum space, since $\int \mathrm{d}^2x\,|x|^2\mathcal{F}(x) = -\nabla_p^2\tilde{\mathcal{F}}(p)\big|_{p=0}$ (differentiate $\tilde{\mathcal{F}}(p) = \int \mathrm{d}^2x\, e^{-ip\cdot x}\mathcal{F}(x)$ twice). The product G^2 becomes a convolution:

$$\tilde{\mathcal{F}}(p) \overset{(1)}{=} \frac{m^4}{2\pi^2}\int \frac{\mathrm{d}^2k}{(2\pi)^2}\,\frac{2\pi}{k^2+m^2}\,\frac{2\pi}{(p-k)^2+m^2} \overset{(2)}{=} \frac{m^4}{2\pi^2}\int_0^1 \mathrm{d}x \int \mathrm{d}^2\ell\,\frac{1}{\big(\ell^2+m^2+x(1-x)p^2\big)^2}$$
$$\overset{(3)}{=} \frac{m^4}{2\pi}\int_0^1 \frac{\mathrm{d}x}{m^2+x(1-x)p^2} \overset{(4)}{=} \frac{m^2}{2\pi} - \frac{p^2}{2\pi}\int_0^1 x(1-x)\,\mathrm{d}x + O(p^4) = \frac{m^2}{2\pi} - \frac{p^2}{12\pi} + O(p^4).$$

(1) The Fourier transform of a product is the convolution of the transforms (with $\int \mathrm{d}^2k/(2\pi)^2$).

(2) Feynman's parameter formula $\frac{1}{AB} = \int_0^1 \frac{\mathrm{d}x}{[xA+(1-x)B]^2}$ with $A = (p-k)^2+m^2$, $B = k^2+m^2$. Then $xA+(1-x)B = (k-xp)^2 + x(1-x)p^2 + m^2$; shift $\ell = k - xp$.

(3) $\int \mathrm{d}^2\ell\,(\ell^2+D)^{-2} = \pi\int_0^\infty \mathrm{d}(\ell^2)\,(\ell^2+D)^{-2} = \pi/D$.

(4) Expand $1/(m^2+ap^2) = m^{-2} - ap^2m^{-4} + \cdots$; $\int_0^1 x(1-x)\mathrm{d}x = \frac{1}{6}$.

In two dimensions $\nabla_p^2 p^2 = 4$, so $\int \mathrm{d}^2x\,|x|^2\mathcal{F} = -(-\frac{4}{12\pi}) = \frac{1}{3\pi}$, and the sum rule (11.318) gives

$$c_{\text{UV}} - c_{\text{IR}} = 3\pi\cdot\frac{1}{3\pi} = 1.$$

Exactly one degree of freedom has been lost, as it must.

We can even write down C at every scale. With $\Theta = \frac{m^2}{4}\varphi^2$, $H = r^4\langle\Theta\Theta\rangle = \frac{m^4r^4}{8}K_0(mr)^2$, and integrating (11.315) from 0 to r,

$$C(r) = 1 - 24\int_0^r \frac{\mathrm{d}r'}{r'}H(r') = 1 - 3\int_0^{mr} u^3K_0(u)^2\,\mathrm{d}u. \tag{11.319}$$

The integral $\int_0^\infty u^3K_0(u)^2\mathrm{d}u = \frac{1}{3}$ (the position-space version of the calculation above), so $C(\infty) = 0$. Figure 11.68 plots it.

11.13.3 Short flows: conformal perturbation theory

Now a flow between two non-trivial CFTs. Perturb a CFT by a primary Φ that is only *slightly* relevant, $\Delta = 2-y$ with $0 < y \ll 1$, normalised by $\langle\Phi(x)\Phi(0)\rangle = |x|^{-2\Delta}$, and with OPE

$$\Phi(x)\Phi(0) = \frac{1}{|x|^{2\Delta}} + \frac{C}{|x|^{\Delta}}\Phi(0) + \cdots \tag{11.320}$$

(the OPE rule Eq. (11.8) fixes the power $|x|^{\Delta-2\Delta}$). The number $C = C_{\Phi\Phi}{}^{\Phi}$ is an OPE coefficient of the CFT.

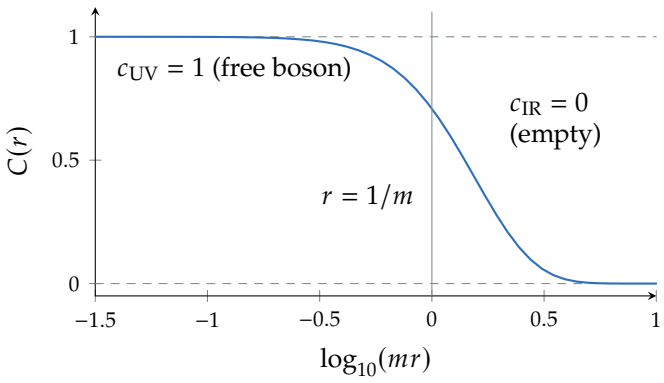


Figure 11.68: Zamolodchikov's C-function for a free boson of mass m, (11.319), computed numerically. What to see: at distances well below the Compton wavelength $1/m$ the boson is effectively massless and $C \approx 1$; near $r \sim 1/m$ it drops; beyond a few Compton wavelengths the field no longer fluctuates and $C \to 0$. The curve is monotonic, as (11.315) requires.

Worked example 11.129: The β-function to second order

Regularise with a short-distance cutoff a: in the expansion of $e^{-g\int\Phi}$, no two operators may come closer than a. Increase the cutoff to $a(1+\delta t)$. In the second-order term $\frac{g^2}{2}\iint\Phi\Phi$ we lose the pairs with separation between a and $a(1+\delta t)$. By the OPE, their contribution is

$$\frac{g^2}{2}\int \mathrm{d}^2x \int_{a<|y|<a(1+\delta t)} \mathrm{d}^2y\ \Phi(x+y)\Phi(x) \overset{(1)}{=} \frac{g^2}{2}\, C\, a^{-\Delta}\,(2\pi a\cdot a\,\delta t)\int \mathrm{d}^2x\,\Phi(x)+\cdots$$
$$= \pi C g^2 a^{2-\Delta}\,\delta t\int \mathrm{d}^2x\,\Phi(x)+\cdots .$$

(1) Use (11.320) (the identity term only shifts the free energy); the thin annulus has area $2\pi a\cdot a\delta t$ and on it $|y|^{-\Delta}=a^{-\Delta}$.

To keep all correlation functions unchanged we must put this contribution back into the first-order term, $-g\int\Phi$: the coupling becomes $g' = g - \pi C g^2 a^{2-\Delta}\delta t$. In the dimensionless coupling $\hat g = g\,a^y$ (measured in units of the cutoff), which also picks up the canonical factor $(1+y\,\delta t)$,

$$\beta(\hat g) = \frac{\mathrm{d}\hat g}{\mathrm{d}t} = y\,\hat g - \pi C\,\hat g^2 + O(\hat g^3). \tag{11.321}$$

For $C > 0$ there is a fixed point at $\hat g^* = y/(\pi C)$, of order y, so it is within reach of perturbation theory.

At the new fixed point the slope is $\beta'(\hat g^*) = y - 2\pi C\hat g^* = -y$: the operator that was relevant with eigenvalue $+y$ at the UV point is irrelevant with eigenvalue $-y$ at the IR point, so its dimension there is $\Delta_{\mathrm{IR}} = 2 + y + O(y^2)$.

Worked example 11.130: How much c is lost

Insert (11.321) in (11.317) with $\mathsf{G} = 1 + O(\hat g)$:

$$c_{\mathrm{UV}} - c_{\mathrm{IR}} \overset{(1)}{=} 6\pi^2\int_0^{\hat g^*}\left(y\,\hat g - \pi C\hat g^2\right)\mathrm{d}\hat g \overset{(2)}{=} 6\pi^2\left(\frac{y\,\hat g^{*2}}{2} - \frac{\pi C\,\hat g^{*3}}{3}\right) \overset{(3)}{=} 6\pi^2\left(\frac{y^3}{2\pi^2C^2} - \frac{y^3}{3\pi^2C^2}\right) = \frac{y^3}{C^2}.$$

(1) (11.317); corrections from $\mathsf{G}-1$ and from the $\hat g^3$ term of β are of relative order y.
(2) Integrate the polynomial.

(3) $\hat{g}^* = y/\pi C$.

So $c_{\mathrm{IR}} = c_{\mathrm{UV}} - y^3/C^2 + O(y^4)$: positive, as the theorem demands, and small, of third order in the distance y from marginality.

11.13.4 Flows between minimal models

The unitary minimal models $\mathcal{M}_m$ (Section 11.4.12) have $c_m = 1 - \frac{6}{m(m+1)}$, increasing towards 1 as $m \to \infty$. Their primaries $\phi_{r,s}$ have weights given by the Kac formula Eq. (11.146). Take $\phi_{1,3}$:

$$h_{1,3} \overset{(1)}{=} \frac{\big((m+1)-3m\big)^2 - 1}{4m(m+1)} = \frac{(2m-1)^2-1}{4m(m+1)} \overset{(2)}{=} \frac{m-1}{m+1}, \qquad \Delta_{1,3} = 2h_{1,3} = 2 - \frac{4}{m+1}.$$

(1) Eq. (11.146) with $r = 1$, $s = 3$.

(2) $(2m-1)^2 - 1 = (2m-2)(2m) = 4m(m-1)$.

So $\phi_{1,3}$ is relevant with $y = 4/(m+1)$, small when m is large: exactly the situation of Section 11.13.3. (For the Ising model, $m = 3$, $\phi_{1,3}$ is the energy operator ε with $\Delta = 1$; the perturbation is not small, and the flow goes to a massive theory, $c = 0 = c_2$.)

Worked example 11.131: $\mathcal{M}_m \to \mathcal{M}_{m-1}$

Zamolodchikov (1987) computed $C_{(1,3)(1,3)}{}^{(1,3)} = \frac{4}{\sqrt{3}}\big(1 + O(1/m)\big)$ from the Coulomb-gas construction. Then Example 11.130 predicts

$$c_{\mathrm{UV}} - c_{\mathrm{IR}} \overset{(1)}{=} \frac{y^3}{C^2} \overset{(2)}{=} \frac{64}{(m+1)^3} \cdot \frac{3}{16} = \frac{12}{(m+1)^3} \approx \frac{12}{m^3}.$$

(1) Example 11.130.

(2) $y = 4/(m+1)$, $C^2 = 16/3$.

Compare with the drop from $\mathcal{M}_m$ to the next model down:

$$c_m - c_{m-1} \overset{(1)}{=} \frac{6}{(m-1)m} - \frac{6}{m(m+1)} \overset{(2)}{=} \frac{6\big((m+1)-(m-1)\big)}{(m-1)m(m+1)} = \frac{12}{m(m^2-1)} \approx \frac{12}{m^3}.$$

(1) $c_m = 1 - 6/(m(m+1))$ and $c_{m-1} = 1 - 6/((m-1)m)$.

(2) Common denominator.

They agree at leading order in $1/m$. The dimension check agrees too. The IR fixed point should have an irrelevant operator with $\Delta_{\mathrm{IR}} = 2 + y = 2 + \frac{4}{m+1}$. In $\mathcal{M}_{m-1}$ (replace m by $m-1$ in the Kac formula),

$$h_{3,1} = \frac{(3m-(m-1))^2 - 1}{4(m-1)m} = \frac{(2m+1)^2-1}{4m(m-1)} = \frac{m+1}{m-1}, \qquad \Delta_{3,1} = 2 + \frac{4}{m-1},$$

which matches at leading order. So the perturbation $\phi_{1,3}$ of $\mathcal{M}_m$ flows to $\mathcal{M}_{m-1}$, arriving along its irrelevant operator $\phi_{3,1}$. Numerically (with $y = 4/(m+1)$):

m	c_m	$c_m - c_{m-1}$ (exact)	y^3/C^2	ratio
5	0.8000	0.1000	0.0556	0.56
10	0.9455	0.01212	0.00902	0.74
20	0.9857	0.001504	0.001296	0.86
50	0.99765	9.60×10^{-5}	9.05×10^{-5}	0.94

The ratio tends to 1 with $O(1/m)$ corrections, as expected from the neglected terms. The exact statement (confirmed by the exact S-matrix and numerical methods) is that the flow ends precisely at $\mathcal{M}_{m-1}$.

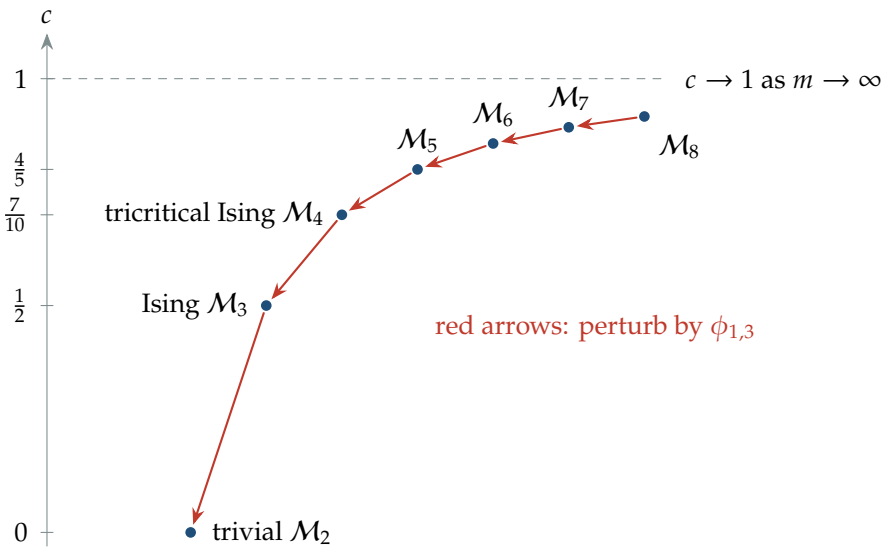


Figure 11.69: The staircase of flows between unitary minimal models. Each dot is $\mathcal{M}_m$ at height $c_m = 1 - 6/(m(m+1))$; each red arrow is the flow generated by the relevant operator $\phi_{1,3}$, which lands on $\mathcal{M}_{m-1}$. The last step, Ising $\to$ trivial, is the thermal perturbation that makes the Ising model massive. What to see: every arrow goes down in c, and the steps shrink like $12/m^3$ as the models accumulate below $c = 1$.

11.13.5 The RG as a gradient flow; the Morse picture

With several couplings $\hat{g}^i$ and $\Theta = \frac{\pi}{2}\beta^i\Phi_i$, the same steps give $H = \frac{\pi^2}{4}\beta^i \mathsf{G}_{ij}\beta^j$ with the **Zamolodchikov metric**

$$\mathsf{G}_{ij}(\hat{g}) = r^4\langle\Phi_i(x)\Phi_j(0)\rangle\big|_{|x|=r}, \qquad \frac{\mathrm{d}C}{\mathrm{d}t} = -6\pi^2\,\beta^i\mathsf{G}_{ij}\beta^j. \tag{11.322}$$

The metric is positive definite: for any real vector v, $v^i\mathsf{G}_{ij}v^j = r^4\langle(v\cdot\Phi)(x)\,(v\cdot\Phi)(0)\rangle \geq 0$ by reflection positivity, with equality only for $v = 0$ when the Φ_i are independent. In conformal perturbation theory one checks more: to the order computed, $C(\hat{g}) = c_{\mathrm{UV}} - 6\pi^2\big(\frac{y}{2}\hat{g}^2 - \frac{\pi C}{3}\hat{g}^3\big)$ satisfies

$$\frac{\partial C}{\partial \hat{g}^i} = -6\pi^2\,\mathsf{G}_{ij}\,\beta^j, \qquad \text{that is} \qquad \beta^i = -\mathsf{G}^{ij}\,\partial_j\Big(\frac{C}{6\pi^2}\Big). \tag{11.323}$$

(For one coupling: $\partial_{\hat{g}}C = -6\pi^2(y\hat{g} - \pi C\hat{g}^2) = -6\pi^2\beta$.) This is the **gradient-flow** property: the RG flows down the gradient of $C/6\pi^2$ with respect to the metric G, which is exactly the gradient flow of Morse theory, [Nash Ch. VII], $\dot{x}^i = -g^{ij}\partial_j f$ (Chapter 7). In general two-dimensional theories (11.323) holds up to extra terms that vanish at fixed points (Nakayama eq. (5.11)); the monotonicity (11.322) holds exactly.

Key idea: the Morse index of C counts relevant operators

Fixed points are critical points of C ($\beta = 0 \Leftrightarrow \partial C = 0$ when G is invertible). Differentiating (11.323) at a fixed point, where $\beta = 0$,

$$\partial_i \partial_j C = -6\pi^2\, \mathsf{G}_{ik}\, \partial_j \beta^k = -6\pi^2\, (\mathsf{G}M)_{ij},$$

with M the stability matrix of Section T5.3.1. So $M = -\mathsf{G}^{-1}\mathrm{Hess}(C)/6\pi^2$. Since G is positive, M is similar to $-\mathsf{G}^{-1/2}\mathrm{Hess}(C)\,\mathsf{G}^{-1/2}/6\pi^2$, which is congruent to $-\mathrm{Hess}(C)$; by Sylvester's law of inertia (Chapter 7) the number of positive eigenvalues y_a of M equals the number of negative eigenvalues of the Hessian:

$$\text{Morse index of } C \text{ at a CFT} = \text{number of relevant operators.}$$

UV fixed points are maxima or saddles of C; an IR-stable fixed point (no relevant operators) is a local minimum.

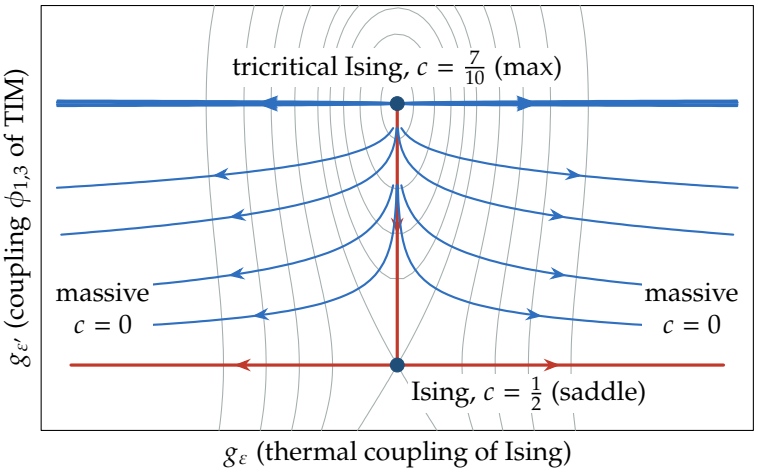


Figure 11.70: The RG as gradient flow on a two-parameter slice of theory space (a schematic height function $C = h(g_{\varepsilon'})e^{-g_\varepsilon^2}$ with the correct values at the fixed points; flow lines computed as its gradient flow). Grey: level sets of C. Blue: flows leaving the tricritical Ising point, a maximum of C with two relevant $\mathbb{Z}_2$-even directions. Thick red: the special flows along $\phi_{1,3}$ from tricritical to Ising, and from the Ising saddle (one relevant direction, ε) down to the two massive phases. What to see: every flow crosses the level sets downwards, fixed points are the critical points of C, and the number of downhill directions at each is its number of relevant operators, its Morse index.

Watch out: what the gradient picture does and does not give

Two consequences are solid: there are no limit cycles (a function that strictly decreases along the flow cannot return to its starting value), and there is no "chaotic" RG in unitary 2D theories. The analogy with Chapter 7 should not be pushed further without care: theory space is infinite-dimensional, not compact, and G can degenerate, so there is no established set of "Morse inequalities" relating the numbers of CFTs of each index. The flows between fixed points are the analogue of the flow lines between critical points that build the Morse–Witten complex.

11.13.6 Four dimensions: the a-theorem

In four dimensions the two-point function of T does not produce a monotonic function; the right quantity turned out to be a coefficient of the trace anomaly. On a curved background a 4D CFT has

$$\langle T^{\mu}{}_{\mu} \rangle = a\, E_4 - c\, W_{\mu\nu\rho\sigma} W^{\mu\nu\rho\sigma}$$

(the convention of Komargodski–Schwimmer; signs vary in the literature), where W is the Weyl tensor and E_4 the Euler density, whose integral is $32\pi^2\chi(M)$ by Gauss–Bonnet–Chern (the four-dimensional version of the Gauss–Bonnet theorem of Chapter T2). Cardy conjectured in 1988 that a, not c, decreases.

Key idea: the a-theorem (Komargodski and Schwimmer, 2011)

For any unitary RG flow in four dimensions between a UV CFT and an IR CFT, $a_{\rm UV} > a_{\rm IR}$. The proof couples the theory to a background "dilaton" $\tau(x)$ by replacing every mass scale $M \to M e^{-\tau}$. Anomaly matching forces the low-energy effective action of τ to contain the Wess–Zumino term $2(a_{\rm UV} - a_{\rm IR}) \int \mathrm{d}^4x\, [2(\partial\tau)^2 \Box\tau - (\partial\tau)^4]$. Its coefficient appears in the forward $2 \to 2$ "dilaton scattering" amplitude as $A(s) \propto (a_{\rm UV} - a_{\rm IR}) s^2$, and a dispersion relation gives

$$a_{\rm UV} - a_{\rm IR} = \frac{1}{4\pi} \int_{s>0} \frac{\mathrm{Im}\, A(s)}{s^3}\, \mathrm{d}s > 0,$$

positive by the optical theorem (Komargodski, eq. (1.15)). The 2D analogue of this argument, with a two-dilaton term $\frac{c_{\rm UV}-c_{\rm IR}}{24\pi} \int (\partial\tau)^2$, reproduces the sum rule (11.318).

Worked example 11.132: Counting with free fields: QCD and pions

For free fields, in units of $1/(90(8\pi)^2)$, a real scalar has $a = 1$, a Weyl fermion $a = \frac{11}{2}$ and a gauge field $a = 62$ (Komargodski–Schwimmer, app. A). QCD with N_c colours and N_f massless quarks is free in the UV (asymptotic freedom): $N_c^2 - 1$ gluons and $2N_cN_f$ Weyl fermions. In the IR, chiral symmetry breaking leaves $N_f^2 - 1$ free massless pions. So

$$a_{\rm UV} \overset{(1)}{=} 62(N_c^2 - 1) + \tfrac{11}{2} \cdot 2N_cN_f = 62(N_c^2 - 1) + 11N_cN_f, \qquad a_{\rm IR} \overset{(2)}{=} N_f^2 - 1.$$

(1) Each Dirac quark of each colour is two Weyl fermions.
(2) Each pion is a real scalar.

For $N_c = 3$, $N_f = 2$: $a_{\rm UV} = 496 + 66 = 562$ and $a_{\rm IR} = 3$. The theorem holds with room to spare. It would fail only for $N_f \gtrsim 44$ at $N_c = 3$, far beyond the $N_f < 16.5$ required for asymptotic freedom. In three dimensions the analogous monotonic quantity is the free energy on the three-sphere (F-theorem, Casini–Huerta).

Recap

- From the conservation equations, $C = 2F - 4G - 6H$ obeys $\mathrm{d}C/\mathrm{d}\ln r^2 = -12r^4\langle\Theta\Theta\rangle \le 0$; $C = c$ at fixed points, so $c_{\rm UV} \ge c_{\rm IR}$.
- $\Theta = \frac{\pi}{2}\beta^i\Phi_i$ gives $\mathrm{d}C/\mathrm{d}t = -6\pi^2\beta\mathsf{G}\beta$ with the positive Zamolodchikov metric G.
- Sum rule $c_{\rm UV} - c_{\rm IR} = 3\pi \int \mathrm{d}^2x\, x^2 \langle T^{\mu}{}_{\mu} T^{\nu}{}_{\nu} \rangle$; a massive boson gives exactly 1.
- Near-marginal flows: $\beta = y\hat{g} - \pi C\hat{g}^2$, $\Delta c = y^3/C^2$; for $\phi_{1,3}$ in $\mathcal{M}_m$ this is $12/m^3$, the drop to $\mathcal{M}_{m-1}$.

- The RG is (to the order checked) the gradient flow of C; the Morse index of C at a CFT is its number of relevant operators.
- In 4D the a-anomaly decreases (Komargodski–Schwimmer).

Chapter summary

This chapter asked what a quantum field theory looks like when it has no length scale at all, and found that in two dimensions the answer is astonishingly rigid. The story in five steps:

1. *Why conformal* (Nash §1). At a second-order phase transition the correlation length diverges, so no length scale is left; scale invariance plus rotations and locality (a traceless stress tensor) give conformal invariance. Short-distance limits of massive theories, and RG fixed points, are conformal. The operator product expansion $A(x)B(0) = \sum_n C_n(x)O_n(0)$ is the organising tool, and scale invariance fixes the powers in $C_n(x)$ by the scaling dimensions.

2. *Two dimensions* (Nash §2). Conformal maps of the plane are holomorphic maps, so the local conformal algebra is infinite dimensional (the Witt algebra), while the global group is the Möbius group $\mathrm{SL}(2,\mathbb{C})/\{\pm\mathbb{1}\}$. Primary fields transform with weights $(h,\bar{h})$. The stress tensor splits into $T(z)$ and $\bar{T}(\bar{z})$; the Ward identity turns into the OPE $T(w)\phi(z) \sim h\phi/(w-z)^2 + \partial\phi/(w-z)$, and $T(w)T(z)$ contains the central charge c, which makes T transform with a Schwarzian.

3. *Virasoro* (Nash §3). The modes L_n of T obey the Virasoro algebra. Radial quantisation turns primaries into highest-weight states; descendants fill a Verma module; null vectors are removed; unitarity restricts (c,h), and for $c<1$ only the discrete series (Kac table) survives.

4. *Statistical mechanics* (Nash §4, Sections 11.5 to 11.8). The critical Ising model is the $c=\frac{1}{2}$ minimal model with primaries of weights $0, \frac{1}{16}, \frac{1}{2}$, and its critical exponents are conformal weights. On the torus the partition function is a sum of products of characters, and modular invariance under $\tau \mapsto (a\tau+b)/(c\tau+d)$ is a strong constraint, verified through the theta and eta functions and classified (ADE).

5. *Operator products, fusion, axioms* (Nash §5, Sections 11.9 to 11.12). Correlators reduce to primaries; two- and three-point functions are fixed up to constants; four-point functions are sums of conformal blocks; the fusion rules N_{ij}^{k} count blocks and are diagonalised by the modular S matrix (Verlinde). The axiomatic approaches of Friedan–Shenker and Segal repackage all of this as amplitudes on Riemann surfaces compatible with gluing, which leads directly to the topological field theories of Chapter 12.

Formula sheet

Key idea: The formulas to remember

primary: $$\phi'(z',\bar{z}') = \left(\frac{\mathrm{d}z'}{\mathrm{d}z}\right)^{-h}\left(\frac{\mathrm{d}\bar{z}'}{\mathrm{d}\bar{z}}\right)^{-\bar{h}}\phi(z,\bar{z})$$

$T\phi$ OPE: $$T(w)\phi(z) \sim \frac{h\,\phi(z)}{(w-z)^2} + \frac{\partial\phi(z)}{w-z}$$

TT OPE: $$T(w)T(z) \sim \frac{c/2}{(w-z)^4} + \frac{2T(z)}{(w-z)^2} + \frac{\partial T(z)}{w-z}$$

Schwarzian: $$T(z) = \left(\tfrac{\mathrm{d}z'}{\mathrm{d}z}\right)^2 T'(z') + \tfrac{c}{12}\{z',z\}$$

Virasoro: $$[L_m, L_n] = (m-n)L_{m+n} + \tfrac{c}{12}m(m^2-1)\delta_{m+n,0}$$

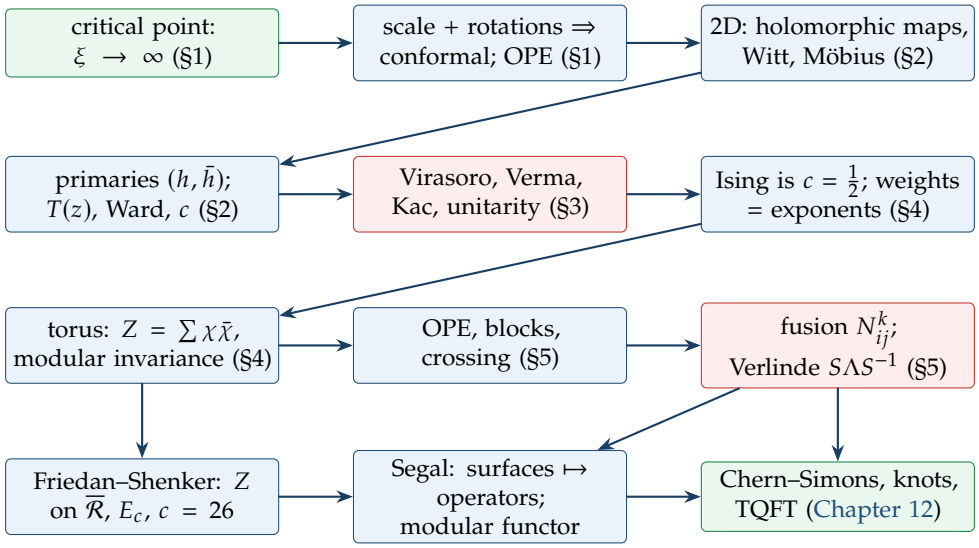


Figure 11.71: The logic of Chapter 11. What to see: the physics of critical points (green, top left) forces conformal invariance; in two dimensions this gives the Virasoro algebra and its representations (red, the algebraic core); the torus and the sphere give two sets of consistency conditions (modular invariance, crossing), which meet in the fusion rules and the Verlinde formula (red); the axiomatic approaches package everything as amplitudes on Riemann surfaces, which is the starting point of topological field theory (green, bottom right).

minimal (6.58):	$c = 1 - \frac{6}{(m+1)(m+2)}, \quad h = \frac{((m+2)p-(m+1)q)^2-1}{4(m+1)(m+2)}$, Ising: $m = 2$
cylinder:	$H = \frac{2\pi}{L}\left(L_0 + \bar{L}_0 - \frac{c}{12}\right), \quad E_0 = -\frac{\pi c}{6L}$
torus:	$Z(\tau) = \mathrm{Tr}\; q^{L_0 - c/24}\bar{q}^{\bar{L}_0 - c/24} = \sum M_{h\bar{h}}\chi_h(\tau)\overline{\chi_{\bar{h}}(\tau)}$
eta:	$\eta(\tau+1) = e^{i\pi/12}\eta(\tau), \quad \eta(-1/\tau) = \sqrt{-i\tau}\,\eta(\tau)$
Ising:	$Z = \frac{\lvert\theta_2\rvert + \lvert\theta_3\rvert + \lvert\theta_4\rvert}{2\lvert\eta\rvert}, \quad \sigma \times \sigma = \mathbf{1} + \varepsilon$
2-, 3-point:	$\langle\phi_i\phi_j\rangle = \frac{C_{ij}}{z_{12}^{2h}\bar{z}_{12}^{2\bar{h}}}, \quad \langle\phi_1\phi_2\phi_3\rangle = \frac{C_{123}}{z_{12}^{h_1+h_2-h_3} z_{23}^{h_2+h_3-h_1} z_{13}^{h_1+h_3-h_2}} \times \text{c.c.}$
Verlinde:	$N_{ij}^k = \sum_m \frac{S_{im}S_{jm}\bar{S}_{km}}{S_{0m}}, \quad \dim V(\Sigma_p) = \sum_m S_{0m}^{2-2p}$
moduli:	$\dim_{\mathbb{C}} \mathcal{M}_{p,n} = 3p - 3 + n, \quad T^*\mathcal{M}_p = H^0(K^2)$
plumbing:	$xy = t \;\Rightarrow$ neck operator $t^{L_0 - c/24}\bar{t}^{\bar{L}_0 - c/24}$
Segal:	$U_\Sigma : H^{\otimes q} \to H^{\otimes r}, \quad U_\Sigma U_{\Sigma'} = \lambda\, U_{\Sigma\cup\Sigma'}, \quad \mathrm{tr}\; U_{\Sigma_{p,2}} = U_{\bar{\Sigma}_{p,2}}$

Concepts, what they mean operationally, and where they are used

concept	operational meaning	where it is used
Nash §1: conformal invariance and QFT		
scaling dimension d_A	$A(\lambda x) \mapsto \lambda^{-d_A} A$; read off from $\langle AA \rangle \propto \lvert x \rvert^{-2d_A}$	critical exponents (§4); OPE powers
anomalous dimension	$\gamma_A = d_A - d_A^{\mathrm{c}}$, from the Callan–Symanzik equation at a fixed point	RG language of §4
RG fixed point	$\beta(g_*) = 0$; necessary, not sufficient, for conformal invariance	critical points; string backgrounds (Chapter 9)
Wilson OPE	$A(x)B(0) = \sum_n C_n(x) O_n(0)$, $C_n \propto \lvert x \rvert^{d_{O_n} - d_A - d_B}$	§5 OPE, fusion rules
conformal group in d dim.	solutions of the conformal Killing equation, $\mathrm{SO}(d+1,1)$, dim. $\frac{(d+1)(d+2)}{2}$	contrast with $d = 2$
Nash §2: conformal field theories in two dimensions		
conformal = holomorphic	f with $\partial_{\bar z} f = 0$ preserves angles, scales lengths by $\lvert f' \rvert$	everything in 2D
Witt algebra	$\ell_n = -z^{n+1}\partial_z$, $[\ell_m, \ell_n] = (m-n)\ell_{m+n}$	Virasoro; $\mathrm{Diff}(S^1)$ (Chapter 6)
Möbius group	$z \mapsto \frac{az+b}{cz+d}$, generated by $\ell_{-1}, \ell_0, \ell_1$; kills $\{f, z\}$	fixing three points; 2-, 3-point functions
primary field $(h, \bar h)$	transforms by $(f')^h (\bar f')^{\bar h}$; OPE with T stops at $(w-z)^{-2}$	all of §3–§5
free fermion, free boson	$c = \frac{1}{2}$ and $c = 1$; ψ primary $(\frac{1}{2}, 0)$, $\partial\phi$ primary $(1,0)$, $e^{ip\phi}$ primary $(\frac{p^2}{4}, \frac{p^2}{4})$	Ising (§4); vertex-operator fusion (§5); strings
stress tensor $T(z)$	$T = T_{zz}$, holomorphic; generates conformal maps via $\oint \varepsilon(z) T(z)$	Ward identities, L_n, moduli connection
central charge c	coefficient of $(w-z)^{-4}$ in TT; Casimir energy $-\pi c/6L$	Virasoro, modular T-phase, $c = 26$, anomaly
Schwarzian $\{f, z\}$	$f'''/f' - \frac{3}{2}(f''/f')^2$; zero iff f Möbius	cylinder shift $-c/24$; projective connection

Table 11.7: Chapter 11 concepts, part 1 (Nash §1–§2).

concept	operational meaning	where it is used
Nash §3: relation to the Virasoro algebra		
Virasoro algebra	modes $L_n = \oint \frac{dz}{2\pi i} z^{n+1} T(z)$; central term $\frac{c}{12} m(m^2-1)$	representations (Chapter 6); Segal's λ
radial quantisation	time $= \ln\lvert z\rvert$, $H = L_0 + \bar{L}_0$; state $\lvert h\rangle = \phi(0)\lvert 0\rangle$	state–operator; Segal's discs
highest-weight state	$L_0\lvert h\rangle = h\lvert h\rangle$, $L_{n>0}\lvert h\rangle = 0$	Verma modules, characters
Verma module $V(c,h)$	span of $L_{-n_1}\cdots L_{-n_k}\lvert h\rangle$; $p(N)$ states at level N	characters $q^{h-c/24}/\prod(1-q^n)$
null vector / singular vector	descendant annihilated by L_1, L_2; zero norm; generates a submodule N	irreducible $L = V/N$; BPZ equations
Gram matrix, Kac determinant	matrix of inner products at level N; zeros at $h = h_{r,s}(c)$	unitarity; Ising weights
unitary minimal models	$c = 1 - \frac{6}{(m+1)(m+2)}$ (Nash's (6.58)), weights $h_{p,q}$	Ising ($m=2$), tricritical Ising ($m=3$), Potts
$\mathfrak{sl}(2,\mathbb{C})$ vacuum	L_{-1}, L_0, L_1 annihilate $\lvert 0\rangle$	Möbius covariance of correlators
Kac–Moody, Sugawara	currents J^a at level k; $T \propto \sum :J^a J^a:$; WZW	$SU(2)_k$ fusion; Chern–Simons (Chapter 12)
Nash §4: statistical mechanics (Sections 11.5 to 11.8)		
Ising model	$H = -J\sum \sigma_i\sigma_j$ on $\mathbb{Z}^2$; T_c from Kramers–Wannier	the prototype critical CFT
critical exponents	$\alpha, \beta, \gamma, \delta, \nu, \eta$; $\nu = 1/(2-2h_\varepsilon)$, $\eta = 4h_\sigma$	universality classes
Ising = minimal model	$c = \frac{1}{2}$; $\mathbf{1}, \sigma, \varepsilon$ with $h = 0, \frac{1}{16}, \frac{1}{2}$	fusion, Verlinde, blocks (§5)
torus, modulus τ	$\mathbb{C}/(\mathbb{Z}+\tau\mathbb{Z})$; $\tau \sim \frac{a\tau+b}{c\tau+d}$; fundamental domain	$\mathcal{M}_1$; one-loop strings
torus partition function	$Z = \mathrm{Tr}\, q^{L_0-c/24}\bar{q}^{\bar{L}_0-c/24} = \sum M_{h\bar{h}}\chi_h\bar{\chi}_{\bar{h}}$	modular invariance; Cardy; Segal's trace
character χ_h	$\mathrm{Tr}_{H_h}\, q^{L_0-c/24}$; a vector in the block space $V(T^2)$	Verlinde's S; modular functor
Dedekind η, theta functions	$\eta = q^{1/24}\prod(1-q^n)$; θ_i via Poisson summation	all explicit modular checks
modular form of weight k	$f(\frac{a\tau+b}{c\tau+d}) = (c\tau+d)^k f(\tau)$; a section of λ^k	η^{24}, discriminant; string measure
compact boson, T-duality	$Z(R) = \lvert\eta\rvert^{-2}\sum q^{p_L^2/4}\bar{q}^{p_R^2/4}$; $Z(R) = Z(2/R)$	strings on circles; $c=1$ moduli
ADE classification	modular invariants of $SU(2)_k \leftrightarrow$ ADE diagrams	classification of CFTs

Table 11.8: Chapter 11 concepts, part 2 (Nash §3–§4).

concept	operational meaning	where it is used
Nash §5: operator products and fusion rules (Sections 11.9 and 11.10)		
conformal family $[\phi]$	ϕ and all $L^I\bar{L}^J\phi$; closes under OPE with T	rational CFT
Ward reduction	$\langle (L_{-n}\phi)\cdots\rangle = \mathcal{L}_{-n}\langle\phi\cdots\rangle$, a differential operator	descendants from primaries
2-, 3-, 4-point functions	fixed by Möbius up to C_{ij}, C_{ijk}, and a function of the cross-ratio z	OPE coefficients; blocks
OPE of primaries	$\phi_i\phi_j = \sum_k C_{ij}^k \Psi_{ij}^k$, descendant coefficients β fixed by Virasoro	crossing, bootstrap
conformal block $\mathcal{F}(m,z)$	contribution of one family m to a four-point function; $z^{h_m-h_k-h_l}(1+\dots)$	monodromy; braid group (Chapter 12)
crossing symmetry	the s- and t-channel block expansions agree; associativity of the OPE	bootstrap; Moore–Seiberg
BPZ equation	a null vector gives a differential equation for correlators	exact Ising $\langle\sigma\sigma\sigma\sigma\rangle$
fusion rules N_{ij}^k	multiplicity of $[\phi_k]$ in $\phi_i\phi_j$; $\dim V(\Sigma_{0,3})$	Verlinde; knot invariants
block spaces $V(\Sigma_{p,n})$	finite-dim. spaces over $\mathcal{M}_{p,n}$, $\dim \mathcal{M}_{p,n} = 3p-3+n$; sewing rule	Hilbert spaces of Chern–Simons (Chapter 12)
Verlinde formula	S diagonalises fusion: $N_i = S\Lambda_i S^{-1}$	dimension formulas; quantum dimensions
dual rep., (11.75)	$\dim V(S_R^2) = \delta_{R,0}$, $\dim V(S_{R_i,R_j}^2) = \delta_{R_i,R_j^*}$	normalisation axiom of TQFT
Weierstrass gaps	exactly p impossible pole orders at a point of Σ_p	higher-genus Ward identities
Nash §5: axiomatics (Sections 11.11 and 11.12)		
genus sum, $\mathcal{R}$	$\exp Z = \int_{\mathcal{R}} I$, $\mathcal{R} = \prod_p \bigcup_n \mathrm{Sym}^n \mathcal{M}_p$	string perturbation theory
node, plumbing, stable curve	$xy = t$; neck operator $t^{L_0-c/24}\bar{t}^{\bar{L}_0-c/24}$; $\overline{\mathcal{M}}_p$	factorisation; degenerations
projective connection	transforms like $\frac{12}{c}T$; differences are quadratic differentials	T as a connection on E_c
anomaly bundle E_c	$E_c \cong \lambda^{c/2}$; $Z = h(\Psi,\Psi)$, $\Psi \in \Gamma(E_c \otimes W)$	$c = 26$ via Mumford; framing
Segal functor	surfaces with boundary $\mapsto$ operators; sewing = composition up to λ	Atiyah's TQFT axioms (Chapter 12)
modular functor	block spaces with gluing and normalisation axioms; projective mapping-class action	Chern–Simons, Jones polynomial (Chapter 12)

Table 11.9: Chapter 11 concepts, part 3 (Nash §5).

Which tool answers which question

- *What are the critical exponents of a 2D transition?* Identify the CFT by its central charge (Casimir energy of a finite chain) and its primaries; then $\nu = 1/(2 - 2h_\varepsilon)$, $\eta = 4h_\sigma$, and the others by scaling relations.
- *Which operators can a CFT contain?* Unitarity and the Kac determinant for $c < 1$; modular invariance of the torus partition function for the combinations $(h, \bar{h})$ and their multiplicities.
- *What is a correlation function?* Möbius invariance for two and three points; conformal blocks and crossing for four; null vectors for exact differential equations.
- *Which fields appear in a product?* The fusion rules, computed from the modular S matrix by Verlinde's formula, or from the indicial exponents of a BPZ equation.
- *How many independent amplitudes on a surface of genus p?* Sew three-punctured spheres: $\dim V(\Sigma_p) = \sum_m S_{0m}^{2-2p}$.
- *What is a CFT without a Lagrangian?* A section of a line bundle over the (stable, universal) moduli space (Friedan–Shenker), or a functor from surfaces to operators (Segal). Forgetting the complex structure gives a TQFT, the subject of Chapter 12.

Chapter 12

Topological Quantum Field Theories

Where this chapter is going

Every quantum field theory we have met so far needs a metric. The Yang–Mills action contains $F \wedge \star F$, and the Hodge star needs a metric. The string and conformal field theories of Chapters 9 and 11 need a little less: only the conformal class of a metric on a Riemann surface. This last chapter asks what happens if we throw away even that. Can we build a quantum field theory whose only input is the bare smooth manifold? Such a theory is a **topological quantum field theory** (TQFT). Its partition function is a number attached to a manifold that does not change when we deform the metric, so it is a *topological invariant*. Its Hilbert spaces are finite-dimensional, and every state has zero energy.

This sounds like a theory in which nothing happens. The surprise of the chapter is the opposite. TQFTs turn out to be machines for producing the deepest invariants known in low-dimensional topology, and the machines are built from ingredients we already have: Morse theory and its supersymmetric quantum mechanics (Chapter 7), instantons and their moduli spaces (Chapter 8), spectral flow and the index theorem (Chapters 4 and 10), and conformal field theory (Chapter 11).

The chapter has three strands.

- **Floer theory (§2).** Witten's Morse complex of Chapter 7 is rebuilt on the infinite-dimensional space $\mathcal{A}/\mathcal{G}$ of gauge orbits of SU(2) connections on a 3-manifold. The Morse function is the Chern–Simons functional. Its critical points are flat connections and its gradient flow lines are instantons on $M \times \mathbb{R}$. The result is Floer homology, a new invariant of homology 3-spheres, graded by $\mathbb{Z}_8$.
- **Donaldson theory (§3).** Moduli spaces of instantons on a 4-manifold carry cohomology classes. Integrating them gives Donaldson's polynomial invariants, which can tell apart 4-manifolds that are homeomorphic but not diffeomorphic. Witten showed that they are correlation functions of a topological field theory, and cutting a 4-manifold along a 3-manifold expresses them through Floer groups.
- **Chern–Simons theory and knots (§§4–7).** The Chern–Simons functional, used as an action, defines a TQFT in three dimensions. Expectation values of Wilson loops in it are knot invariants: the linking number in the abelian theory, the Jones polynomial in the SU(2) theory. Surgery and the modular data of a conformal field theory make the computation explicit.

This part of the notes covers §§1–3. Figure 12.1 shows how the ideas feed into each other.

One idea runs through the whole chapter, and it is worth stating now. In Chapter 7 the cohomology of a finite-dimensional manifold X was computed as the space of *zero-energy*

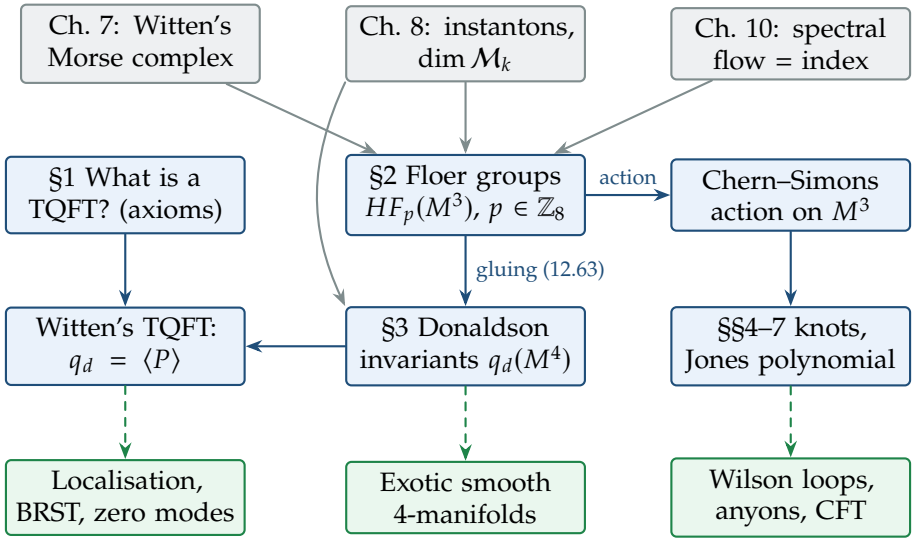


Figure 12.1: The logic of Chapter XII. Grey boxes are earlier chapters that feed in, blue boxes are this chapter, green boxes are the physics or mathematics that comes out. Read the middle column top to bottom: Morse theory on the space of gauge orbits gives Floer homology, and Floer homology is the "space of states" that glues Donaldson's 4-manifold invariants together. This part covers §§1–3.

states of a supersymmetric quantum mechanics on X. Classical vacua sat at the critical points of a function f, and instantons (steepest-descent paths) tunnelling between them decided which classical vacua survived as true quantum ground states. Floer theory is the same story with X replaced by $\mathcal{A}/\mathcal{G}$. Donaldson theory is its four-dimensional "relativistic" completion. Chern–Simons theory takes the Morse function itself and uses it as an action. In each case, a topological invariant is a *count* (of vacua, instantons or flat connections) done in a way that no deformation of the metric can change.

12.1 Introduction: quantum field theory without a metric

What is the least amount of structure on which one can build a quantum field theory? In the previous chapter the data were a Riemann surface Σ and its complex structure. Once those are fixed, the Hilbert space of the conformal field theory can be constructed. Nash's opening move is to discard the complex structure too, and with it every trace of a metric. A quantum field theory built from Σ alone (or from a manifold M of any dimension alone) is a topological quantum field theory.

This section explains what such a theory is, why its dynamics is "trivial" and yet its content is not, and how one axiomatises it. Nash gives the physical description in one paragraph (printed p. 322) and mentions at the end of the chapter that "an axiomatic approach, somewhat analogous to that introduced for conformal field theories, can be given, cf. Atiyah". We give those axioms here, with pictures and small examples, because they are the cleanest way to see what the later sections are computing.

12.1.1 No metric, no dynamics

Start with ordinary quantum mechanics or field theory on a spacetime M with metric $g_{\mu\nu}$. Time evolution is generated by the Hamiltonian H, and H is the integral over space of the energy

density T_{00}, a component of the energy–momentum tensor

$$T_{\mu\nu} = \frac{2}{\sqrt{g}} \frac{\delta S}{\delta g^{\mu\nu}}. \tag{12.1}$$

(We use the Euclidean sign conventions; the overall normalisation will not matter.) Now suppose the action S does not depend on the metric at all. Then $\delta S/\delta g^{\mu\nu} = 0$, so $T_{\mu\nu} = 0$, and

$$H \overset{(1)}{=} \int_{\text{space}} T_{00}\, \mathrm{d}^{d-1}x \overset{(2)}{=} 0. \tag{12.2}$$

(1) The Hamiltonian is the charge of time translations, and the conserved current of translations is the energy–momentum tensor.
(2) A metric-independent action has vanishing energy–momentum tensor, by (12.1).

So every state has energy zero. There is nothing like a particle moving from one place to another, because "distance" and "speed" do not exist without a metric. There are no forces. This is what Nash means by "no conventional dynamics".

A second consequence is that the Hilbert space is usually *finite-dimensional*. Here is the heuristic. In an ordinary field theory the Hilbert space is infinite-dimensional because there are infinitely many excitations (particles with any momentum). Each excitation costs energy. If all states have zero energy there is nowhere for the infinitely many excitations to live; what is left is a finite set of "vacua", typically labelled by topological data (flat connections, homology classes, conformal blocks). We will see this explicitly in the examples below.

Physical picture: Where the non-triviality lives: tunnelling between vacua

If $H = 0$, what can such a theory possibly compute? The answer is that it still has *several* vacua, and the quantum amplitudes between them are non-trivial numbers that depend only on the topology. In the Morse-theory language of Chapter 7: perturbation theory around each critical point of a function f produces one candidate ground state, but instantons tunnelling between critical points can pair some of them up and lift them. The surviving ground states count the cohomology. The information is in *which* vacua are connected by tunnelling, and with what sign. That is Nash's sentence "the non-triviality of the theory is reflected in the existence of tunnelling between vacua".

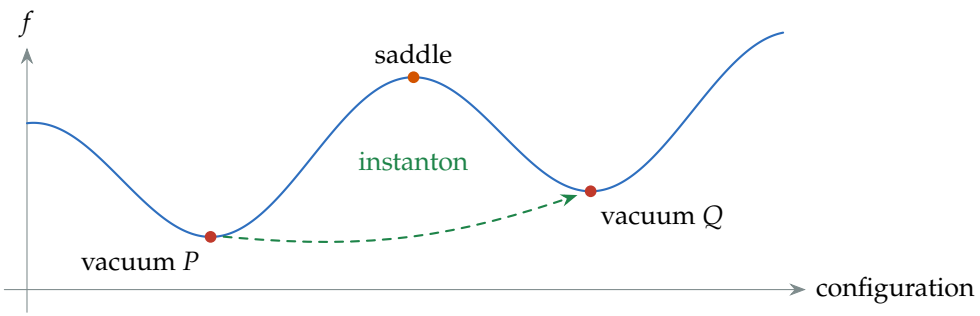


Figure 12.2: Two classical vacua (red) separated by a barrier. With $H = 0$ nothing moves classically, but the Euclidean path integral contains instantons (green) that tunnel from one vacuum to another. A topological theory is controlled entirely by the pattern and signs of such tunnelling events.

12.1.2 Two ways to be topological

There are two ways to make the physics independent of the metric, and this chapter uses both.

Working definition 12.1 (Schwarz-type and Witten-type TQFTs)

A quantum field theory is **topological** if its correlation functions (including the partition function) do not change when the metric is varied [the answers depend only on the shape of the manifold up to smooth deformation, not on any lengths or angles]. *Test:* vary $g_{\mu\nu}$ inside a correlator and show the answer is zero.

- **Schwarz type**: the action contains no metric at all, so $T_{\mu\nu} = 0$ identically [$T_{\mu\nu} = \delta S/\delta g^{\mu\nu}$, the response of the action to a change of metric]. Example: the Chern–Simons action $\int \mathrm{tr}(A \wedge \mathrm{d}A + \frac{2}{3} A \wedge A \wedge A)$ on a 3-manifold (Section 12.5).
- **Witten type** or **cohomological**: the action does contain a metric, but there is a fermionic symmetry Q [a symmetry that turns bosonic fields into fermionic ones and back, like a BRST operator] with $Q^2 = 0$ (on gauge-invariant quantities) such that both the action and $T_{\mu\nu}$ are Q-exact [equal to Q acting on something. The vacuum expectation value of any such term is zero]: $S = \{Q, V\}$, $T_{\mu\nu} = \{Q, V_{\mu\nu}\}$. Example: Witten's theory of Donaldson invariants (Section 12.7).

In a Witten-type theory, H is not zero as an operator, but it is Q-exact, and the physical states are Q-cohomology classes. On them H acts as zero. The supersymmetric quantum mechanics of Chapter 7 is the prototype: $H = \{Q, Q^\dagger\}$, and the ground states represent the cohomology.

12.1.3 Atiyah's axioms: a TQFT is a machine on cobordisms

The cleanest way to say what a TQFT *is* goes back to Atiyah (1988), following Segal's axioms for conformal field theory. The idea is to write down the structure that a path integral would have, if it existed, and take that structure as the definition.

Think of a d-dimensional Euclidean spacetime M whose boundary is a disjoint union of $(d-1)$-dimensional "space slices". In a path integral:

- a closed space slice Σ has a space of states $Z(\Sigma)$ (wave functionals of the fields on Σ);
- a spacetime M with boundary $\partial M = \Sigma$ produces a state $Z(M) \in Z(\Sigma)$: the path integral over fields on M with prescribed boundary values on Σ, viewed as a function of those boundary values;
- a spacetime M running from an "in" slice Σ_0 to an "out" slice Σ_1 gives a linear map $Z(M) : Z(\Sigma_0) \to Z(\Sigma_1)$, the Euclidean evolution operator.

Such an M is called a **cobordism** from Σ_0 to Σ_1. Orientation matters: $\bar{\Sigma}$ denotes Σ with the opposite orientation.

Working definition 12.2 (Atiyah's axioms for a d-dimensional TQFT)

A TQFT Z assigns

- to every closed oriented $(d-1)$-manifold Σ [a space slice, compact with no boundary], a finite-dimensional complex vector space $Z(\Sigma)$;
- to every compact oriented d-manifold M with $\partial M = \Sigma$ a vector $Z(M) \in Z(\Sigma)$ [the state on Σ prepared by the path integral over M],

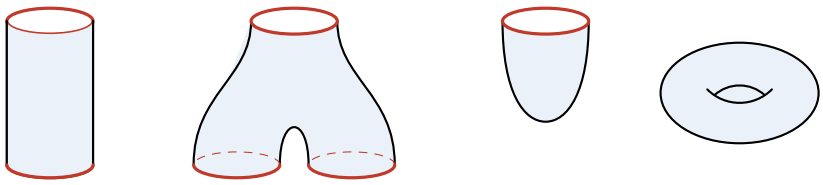


Figure 12.3: Cobordisms in two dimensions and what a TQFT assigns to them. Time runs upward: the bottom circles are "in", the top circles are "out". Red circles are boundaries, each carrying the vector space $V = Z(S^1)$. The cylinder is the identity map, whatever its length. The disc with its circle on top is the unit $1 \in V$; turned upside down it is the trace $\theta : V \to \mathbb{C}$. A closed surface has no boundary, so it is a map $\mathbb{C} \to \mathbb{C}$, that is, a number.

such that:

(A1) *Naturality.* A diffeomorphism $\Sigma \to \Sigma'$ induces an isomorphism $Z(\Sigma) \to Z(\Sigma')$, compatible with the vectors $Z(M)$.

(A2) *Orientation reversal.* $Z(\bar{\Sigma}) = Z(\Sigma)^*$ (the dual space).

(A3) *Multiplicativity.* $Z(\Sigma_1 \sqcup \Sigma_2) = Z(\Sigma_1) \otimes Z(\Sigma_2)$, and $Z(\varnothing) = \mathbb{C}$.

(A4) *Gluing.* If M_1 has boundary $\bar{\Sigma}_0 \sqcup \Sigma_1$ and M_2 has boundary $\bar{\Sigma}_1 \sqcup \Sigma_2$, then $Z(M_1 \cup_{\Sigma_1} M_2) = Z(M_2) \circ Z(M_1)$, where by (A2), (A3) we read $Z(M_i)$ as linear maps $Z(\Sigma_{i-1}) \to Z(\Sigma_i)$.

(A5) *Identity.* $Z(\Sigma \times [0,1]) = \mathrm{id}_{Z(\Sigma)}$.

Operational reading: to compute Z of anything, cut it into simple pieces, compute each piece as a vector or matrix, and multiply the matrices. *Formally*, a TQFT is a symmetric monoidal functor from the cobordism category [objects are closed $(d-1)$-manifolds, arrows are the d-manifolds joining them, and "monoidal" means $\sqcup$ goes to $\otimes$] to the category of vector spaces.

Let us see what these axioms buy us.

Closed manifolds give numbers. A closed d-manifold M has $\partial M = \varnothing$, so $Z(M) \in Z(\varnothing) = \mathbb{C}$. This number is the partition function. Axiom (A1) says it depends only on the diffeomorphism class of M: it is an invariant.

The Hamiltonian vanishes. The cylinder $\Sigma \times [0,T]$ is the Euclidean evolution operator e^{-TH}. Axiom (A5) says it is the identity, for every T (a cylinder of length T is diffeomorphic to one of length 1). So

$$\begin{aligned} -H &\overset{(1)}{=} \frac{\mathrm{d}}{\mathrm{d}T} e^{-TH}\Big|_{T=0} \\ &\overset{(2)}{=} \frac{\mathrm{d}}{\mathrm{d}T} \mathrm{id} \\ &= 0. \end{aligned}$$

(1) Differentiate the evolution operator at $T = 0$.
(2) $Z(\Sigma \times [0,T]) = \mathrm{id}$ for all T, by (A5) and (A1).

This is the axiomatic form of (12.2).

The dimension of the space of states is a partition function. Glue the two ends of the cylinder $\Sigma \times [0,1]$ together. The result is $\Sigma \times S^1$. Gluing an in-boundary to an out-boundary is taking

a trace (it is the "sum over intermediate states" of the path integral with periodic boundary conditions). So

$$\begin{aligned} Z(\Sigma \times S^1) &\overset{(1)}{=} \mathrm{Tr}_{Z(\Sigma)}\big(Z(\Sigma \times [0,1])\big) \\ &\overset{(2)}{=} \mathrm{Tr}_{Z(\Sigma)}\,\mathrm{id} \\ &= \dim Z(\Sigma). \end{aligned} \tag{12.3}$$

(1) Gluing the out-end to the in-end contracts the $Z(\Sigma)$ and $Z(\Sigma)^*$ indices of the cylinder, which is a trace.
(2) Axiom (A5).

Physically, (12.3) is the statement $\mathrm{Tr}\, e^{-\beta H} = \mathrm{Tr}\,\mathrm{id} = \dim \mathcal{H}$ when $H = 0$: the thermal partition function counts ground states. The same formula will turn up in the knot sections of the chapter, and it is the reason finite-dimensionality is forced: $\dim Z(\Sigma)$ is a (finite) number.

Worked example 12.1: The simplest TQFT: $d = 1$

In one dimension the closed 0-manifolds are finite sets of oriented points. Let $Z(+) = V$, a vector space of dimension n, so $Z(-) = V^*$ by (A2). Pick a basis e_i of V and the dual basis e^i of V^*.
An interval bent into a "cap" has boundary $\{+\} \sqcup \{-\}$ as an in-boundary; it is a map $\mathrm{ev} : V \otimes V^* \to \mathbb{C}$, the evaluation $v \otimes \alpha \mapsto \alpha(v)$. Bent the other way (a "cup") it is a map $\mathrm{coev} : \mathbb{C} \to V \otimes V^*$, $1 \mapsto \sum_i e_i \otimes e^i$. The circle is a cup followed by a cap. So

$$\begin{aligned} Z(S^1) &\overset{(1)}{=} \mathrm{ev} \circ \mathrm{coev}(1) \\ &\overset{(2)}{=} \textstyle\sum_{i=1}^{n} e^i(e_i) \\ &\overset{(3)}{=} n. \end{aligned}$$

(1) Cut the circle into two arcs and use the gluing axiom (A4).
(2) Definitions of coev and ev.
(3) $e^i(e_j) = \delta^i_j$.

This is (12.3) with $\Sigma =$ point. Physically it is quantum mechanics of a system with an n-dimensional Hilbert space and $H = 0$: the partition function on a circle of any circumference β is $\mathrm{Tr}\, e^{-\beta \cdot 0} = n$.

Worked example 12.2: Two dimensions: a commutative Frobenius algebra

In two dimensions the only closed connected 1-manifold is the circle, so the theory is determined by the single space $V = Z(S^1)$ and by what the basic cobordisms of Figure 12.3 do. The pair of pants (two circles in, one out) is a product $m : V \otimes V \to V$. The disc read as "one circle in, nothing out" is a linear functional $\theta : V \to \mathbb{C}$ (the "trace"). The disc read the other way is a unit $1 \in V$. Because pants can be deformed into each other in the ways drawn by topologists, m is associative and commutative, and $\langle a, b\rangle = \theta(ab)$ is a non-degenerate pairing. Such a structure is a **commutative Frobenius algebra**.
Take the simplest one: $V = \mathbb{C}^n$ with basis $e_1, \dots, e_n$ of idempotents, $e_i e_j = \delta_{ij} e_i$, unit $1 = \sum_i e_i$, and $\theta(e_i) = \theta_i \neq 0$. The pairing is $\langle e_i, e_j\rangle = \delta_{ij}\theta_i$, so the copairing (the cylinder bent into a U with two out-circles) is $\sum_i e_i \otimes e_i/\theta_i$. A handle is "copairing then multiply":

$$h \overset{(1)}{=} m\Big(\textstyle\sum_i \theta_i^{-1}\, e_i \otimes e_i\Big)$$

$$\overset{(2)}{=} \textstyle\sum_i \theta_i^{-1}\, e_i .$$

(1) A handle is a tube that splits and rejoins: copairing followed by the product.
(2) $e_i e_i = e_i$.

A closed genus-g surface is a disc (the unit 1), then g handles, then a cap (θ):

$$\begin{aligned} Z(\Sigma_g) &\overset{(1)}{=} \theta\big(h^g \cdot 1\big) \\ &\overset{(2)}{=} \theta\Big(\textstyle\sum_i \theta_i^{-g} e_i\Big) \\ &= \textstyle\sum_{i=1}^{n} \theta_i^{1-g}. \end{aligned}$$

(1) Cut Σ_g into a disc, g one-holed tori, and a disc, and glue with (A4).
(2) Idempotents multiply diagonally, so $h^g = \sum_i \theta_i^{-g} e_i$, and $h^g \cdot 1 = h^g$.

Check: for the torus $g = 1$ we get $\sum_i 1 = n = \dim V$, exactly as (12.3) demands for $\Sigma \times S^1 = S^1 \times S^1$.

Worked example 12.3: Gauge theory with a finite gauge group counts flat connections

Here is a physical 2d TQFT. Let G be a finite group and do "lattice gauge theory" on a surface with gauge group G. Since G is discrete, there is no curvature to integrate: every connection is flat, and a flat connection is determined up to gauge by its holonomies, a homomorphism $\rho : \pi_1(\Sigma) \to G$ up to conjugation. The path integral, with the natural normalisation (divide by the volume $|G|$ of the gauge group at a base point), is

$$Z(\Sigma_g) \;=\; \frac{\#\operatorname{Hom}\big(\pi_1(\Sigma_g), G\big)}{|G|}. \tag{12.4}$$

One can show it is the Frobenius algebra of class functions on G, with $\theta_R = (\dim R/|G|)^2$ for the irreducible representations R, so that $Z(\Sigma_g) = \sum_R (\dim R/|G|)^{2-2g}$. Let us test this on $G = S_3$ (order 6; irreducible representations of dimensions $1, 1, 2$; three conjugacy classes).
Sphere. $\pi_1(S^2) = 1$, so only the trivial homomorphism exists:

$$Z(S^2) \overset{(1)}{=} \tfrac{1}{6} \overset{(2)}{=} \tfrac{1^2+1^2+2^2}{6^2} = \tfrac{6}{36}.$$

(1) Formula (12.4) with one homomorphism.
(2) Frobenius formula at $g = 0$; the numbers agree because $\sum_R (\dim R)^2 = |G|$.

Torus. $\pi_1(T^2) = \mathbb{Z}^2$, so a homomorphism is a commuting pair (a, b). For each a the number of b commuting with it is the order of the centraliser of a: 6 for the identity, 2 for each of the 3 transpositions, 3 for each of the 2 three-cycles. So

$$Z(T^2) \overset{(1)}{=} \tfrac{1}{6}\,(6 + 3\cdot 2 + 2\cdot 3) = \tfrac{18}{6} = 3 \overset{(2)}{=} 1 + 1 + 1.$$

(1) Count commuting pairs using centralisers.
(2) Frobenius formula at $g = 1$: one term per irreducible representation.

And $3 = \dim Z(S^1)$, the number of class functions, as (12.3) requires.

Physical picture: The same set will come back

The partition function of a gauge theory whose only field configurations are flat connections *counts* $\mathrm{Hom}(\pi_1, G)/G$. Keep this in mind: in §2 the critical points of the Chern–Simons functional on a 3-manifold are exactly the flat SU(2) connections, $\mathrm{Hom}(\pi_1(M), \mathrm{SU}(2))/\mathrm{Ad}\,\mathrm{SU}(2)$, and Casson's invariant is a *signed* count of them.

Watch out: Axioms are not a construction

Atiyah's axioms tell us what structure a TQFT has. They do not tell us how to find one in dimension 3 or 4. The theories of this chapter are constructed from physics: an action, a path integral, and a lot of analysis (Floer, Donaldson) to make the resulting counts rigorous. In dimension 4, Floer homology is only a partial fit to the axioms: it is defined only for homology 3-spheres and is $\mathbb{Z}_8$-graded rather than a plain vector space.

12.1.4 Where the first TQFT comes from

Topological field theories are not restricted to two dimensions. In three dimensions (Chern–Simons theory) they compute knot invariants; in four dimensions (Witten's theory) they compute Donaldson invariants. Nash's plan is to find the first one by *generalising a quantum-mechanical construction we already understand*: the supersymmetric quantum mechanics whose ground states are the de Rham cohomology of a manifold. Replacing the manifold by the infinite-dimensional space of gauge orbits, and the Morse function by the Chern–Simons functional, leads to Floer homology, and at the very end, to the Chern–Simons TQFT. This is the subject of the next four sections of these notes.

Dictionary: TQFT words

Mathematics	Physics
closed $(d-1)$-manifold Σ	a space slice
$Z(\Sigma)$	Hilbert space of states on Σ
cobordism $M : \Sigma_0 \to \Sigma_1$	Euclidean spacetime history
$Z(M) : Z(\Sigma_0) \to Z(\Sigma_1)$	path integral with fixed boundary values; evolution operator
$Z(\Sigma \times I) = \mathrm{id}$	$H = 0$
$Z(M)$, M closed	partition function; topological invariant
$Z(\Sigma \times S^1) = \dim Z(\Sigma)$	$\mathrm{Tr}\, e^{-\beta H}$ counts ground states
gluing axiom	inserting a complete set of states

Reading Nash (p. 322, §1)

Nash's §1 is one paragraph. He says: a CFT needs Σ plus a complex structure; drop the complex structure and the metric and you get a TQFT; without a metric there are no distances, forces or conventional dynamics; the Hamiltonian has only zero eigenstates and the Hilbert space is usually finite-dimensional; the theory is non-trivial because of tunnelling between vacua; there is an axiomatic approach (Atiyah) analogous to the one for CFTs; and the first example will arise from a field-theory version of the Morse theory of his Chapter 7. We have expanded each clause: $H = 0$ from $T_{\mu\nu} = 0$ or from the cylinder axiom; finite-dimensionality from $Z(\Sigma \times S^1) = \dim Z(\Sigma)$; and explicit TQFTs in dimensions 1 and 2.

Recap

- A TQFT is a quantum field theory whose correlation functions do not depend on the metric. Closed manifolds get numbers (invariants), space slices get finite-dimensional state spaces.
- Schwarz type: metric-free action (Chern–Simons). Witten type: action and $T_{\mu\nu}$ are Q-exact (Donaldson–Witten).
- Atiyah's axioms: cut, compute pieces, compose. Consequences: $H = 0$ and $Z(\Sigma\times S^1) = \dim Z(\Sigma)$.
- 2d TQFTs are commutative Frobenius algebras; $Z(\Sigma_g) = \sum_i \theta_i^{1-g}$. Finite-group gauge theory counts flat connections.
- The content of a TQFT is in tunnelling between its vacua.

12.2 Floer's programme and the Chern–Simons function

In Chapter 7 a function f on a finite-dimensional manifold X was enough to rebuild the cohomology of X: critical points of f became basis vectors of a complex, and gradient flow lines (instantons of a supersymmetric quantum mechanics) became the matrix elements of its coboundary. Floer asked: what if X is the space of all gauge fields on a 3-manifold, and f is the most natural gauge-theory function there is? This section sets up the stage: which space, which bundle, which gauge group and which function. The key discovery is that the natural function, the Chern–Simons functional, is not quite gauge invariant: it shifts by *integers*. That integer shift will eventually produce the $\mathbb{Z}_8$ grading of Floer homology.

Nash's §2 is long. We split it into four sections of these notes: this one (the function), Section 12.3 (critical points, Hessian, relative index), Section 12.4 (gradient flow and instantons, the number 8) and Section 12.5 (Floer groups, Casson's invariant, Chern–Simons theory and Floer's Hamiltonian).

12.2.1 Reminder: Witten's Morse complex

Here is what we need from Chapter 7, in one place. Let X be a compact Riemannian manifold of dimension n and $f : X \to \mathbb{R}$ a Morse function (all critical points non-degenerate). At a critical point x the Hessian $\partial^2 f/\partial x^i \partial x^j$ is a symmetric matrix; its number of negative eigenvalues is the **Morse index** λ_x.

The supersymmetric quantum mechanics on X has Hilbert space

$$\mathcal{H} \;=\; \bigoplus_{p\geq 0} \Omega^p(X), \tag{12.5}$$

the space of all differential forms (Nash's (12.1)). A p-form is a state with p fermions, so even p are bosonic and odd p are fermionic. The supercharge is the exterior derivative d, and the Hamiltonian is the Hodge Laplacian $\mathrm{d}\mathrm{d}^* + \mathrm{d}^*\mathrm{d}$, whose zero-energy states (harmonic forms) represent $H^*_{\mathrm{dR}}(X)$.

Witten deformed d to $\mathrm{d}_t = e^{-tf}\,\mathrm{d}\,e^{tf}$. For large t the low-energy states localise at the critical points of f: one approximate ground state per critical point, a p-form if the index is p. So the

approximate ground states span

$$C^p \;=\; \bigoplus_{\text{crit. pts } x,\ \lambda_x = p} \mathbb{R}\cdot x. \tag{12.6}$$

Tunnelling (instantons) between critical points lifts some of them. The exact statement is that there is a **coboundary**

$$\cdots \xrightarrow{\delta} C^p \xrightarrow{\delta} C^{p+1} \xrightarrow{\delta} \cdots, \qquad \delta x \;=\; \sum_{\lambda_y = \lambda_x + 1} n(x,y)\, y, \tag{12.7}$$

where $n(x,y)$ is the number of gradient flow lines running between y and x, counted with signs from orientations. Then $\delta^2 = 0$ and $\operatorname{Ker}\delta/\operatorname{Im}\delta \cong H^*_{\mathrm{dR}}(X)$. This is Nash's (12.12).

Worked example 12.4: The Morse complex of a circle with four critical points

Let $X = S^1$ with angle ϑ, and $f = \cos 2\vartheta$. Then $f' = -2\sin 2\vartheta$ vanishes at $\vartheta = 0, \pi/2, \pi, 3\pi/2$, and $f'' = -4\cos 2\vartheta$ is -4 at $0, \pi$ (maxima, index 1) and $+4$ at $\pi/2, 3\pi/2$ (minima, index 0). Call the minima m_1, m_2 and the maxima M_0, M_π. So $C^0 = \mathbb{R}m_1 \oplus \mathbb{R}m_2$ and $C^1 = \mathbb{R}M_0 \oplus \mathbb{R}M_\pi$.

Each minimum is joined to each maximum by exactly one flow line (an arc of the circle). Orient S^1 anticlockwise and give a flow line from a minimum up to a maximum the sign $+1$ if it runs anticlockwise, -1 if clockwise. From m_1 ($\vartheta = \pi/2$) the arc to M_π is anticlockwise and the arc to M_0 is clockwise. From m_2 ($\vartheta = 3\pi/2$) the arc to M_0 ($= 2\pi$) is anticlockwise and the arc to M_π clockwise. Hence

$$\delta m_1 = M_\pi - M_0, \qquad\qquad \delta m_2 = M_0 - M_\pi.$$

The kernel of δ on C^0 is spanned by $m_1 + m_2$, and the image is spanned by $M_0 - M_\pi$. So

$$H^0 \overset{(1)}{=} \operatorname{Ker}\delta|_{C^0} = \mathbb{R}(m_1 + m_2) \cong \mathbb{R},$$
$$H^1 \overset{(2)}{=} C^1/\operatorname{Im}\delta = \frac{\mathbb{R}M_0 \oplus \mathbb{R}M_\pi}{\mathbb{R}(M_0 - M_\pi)} \cong \mathbb{R}.$$

(1) There is no C^{-1}, so H^0 is the kernel alone.
(2) There is no C^2, so everything in C^1 is closed.

These are the Betti numbers $b_0 = b_1 = 1$ of the circle, although f has four critical points. The two "extra" vacua have been paired off by tunnelling.

Floer's idea is to run exactly this machine with X infinite-dimensional:

	Chapter 7	Floer
space X	compact manifold M	gauge orbits $\mathcal{A}/\mathcal{G}$ over a 3-manifold
function f	any Morse function	the Chern–Simons functional
critical points	$\mathrm{d}f = 0$	flat connections
flow lines	$\dot{x} = -\nabla f$	instantons on $M \times \mathbb{R}$
index	finite, λ_x	infinite; only differences defined, mod 8
output	$H^*_{\mathrm{dR}}(M)$	Floer groups $HF_p(M)$, $p \in \mathbb{Z}_8$

With an arbitrary f on an arbitrary $\mathcal{A}/\mathcal{G}$ this would be hopeless. Floer made very specific choices, which we now describe.

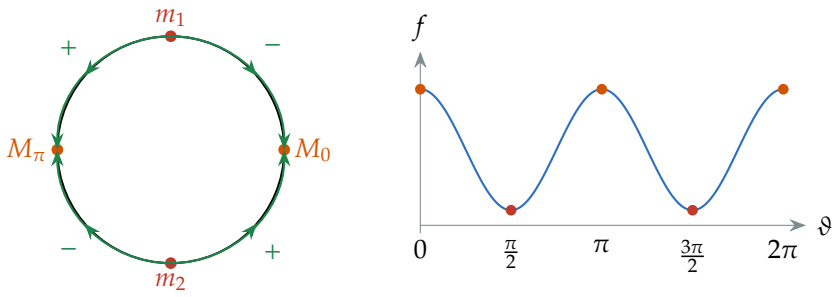


Figure 12.4: Left: the circle with $f = \cos 2\vartheta$, two minima (red) and two maxima (orange). Green arrows are the four flow lines from minima up to maxima, with the orientation signs used in the coboundary. Right: the graph of f. Each minimum sees the two maxima with opposite signs, so $\delta(m_1 + m_2) = 0$ and one combination of maxima is exact.

12.2.2 The stage: SU(2) connections on a homology 3-sphere

Floer's choice (Nash's (12.2)) is

$$\mathcal{A} \;=\; \{\text{SU(2) connections on a closed orientable 3-manifold } M\}. \tag{12.8}$$

Throughout the Floer part of this chapter M is three-dimensional. (In the Donaldson sections, Section 12.6 onwards, M will be four-dimensional; we will warn again.)

Connections as matrices of 1-forms

Recall the conventions (see Chapter T3). Over a region where the bundle is trivial, an SU(2) connection is a 2×2 matrix of 1-forms

$$A = A^a_\mu \, T_a \, \mathrm{d}x^\mu, \qquad T_a = -\tfrac{i}{2}\sigma_a, \tag{12.9}$$

with the Pauli matrices

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

The T_a are anti-Hermitian and traceless, so they span $\mathfrak{su}(2)$. From $\sigma_a\sigma_b = \delta_{ab} + i\varepsilon_{abc}\sigma_c$ we get

$$[T_a, T_b] \overset{(1)}{=} -\tfrac{1}{4}[\sigma_a, \sigma_b] \overset{(2)}{=} -\tfrac{1}{4} \cdot 2i\varepsilon_{abc}\sigma_c \overset{(3)}{=} \varepsilon_{abc} T_c,$$
$$\operatorname{tr}(T_a T_b) \overset{(4)}{=} -\tfrac{1}{4}\operatorname{tr}(\sigma_a\sigma_b) = -\tfrac{1}{2}\delta_{ab}.$$

(1) $T_aT_b = (-\frac{i}{2})^2\sigma_a\sigma_b = -\frac{1}{4}\sigma_a\sigma_b$.
(2) The symmetric part δ_{ab} cancels in the commutator.
(3) $-\frac{1}{4}\cdot 2i\,\varepsilon_{abc}\sigma_c = \varepsilon_{abc}\big(-\frac{i}{2}\sigma_c\big)$, and $-\frac{i}{2}\sigma_c = T_c$.
(4) $\operatorname{tr}\sigma_a\sigma_b = 2\delta_{ab}$.

So $-\operatorname{tr}(XY)$ is a positive-definite inner product on $\mathfrak{su}(2)$: for $X = X^aT_a$, $-\operatorname{tr}(X^2) = \frac{1}{2}\sum_a (X^a)^2$. The curvature is $F = \mathrm{d}A + A \wedge A$, and a gauge transformation $g : M \to \mathrm{SU}(2)$ acts by

$$A \mapsto A^g = g^{-1}Ag + g^{-1}\mathrm{d}g, \qquad F \mapsto F^g = g^{-1}Fg. \tag{12.10}$$

Dictionary: Gauge-field conventions

Nash and these notes use anti-Hermitian A. A physicist writes $A_{\text{math}} = -i\, g_{\text{YM}}\, A^a_{\text{phys}}\, \frac{\sigma_a}{2}$ with real A^a_{phys} and coupling g_{YM}; then $F_{\text{math}} = -i\, g_{\text{YM}}\, F^a_{\text{phys}}\, \frac{\sigma_a}{2}$ and $-\operatorname{tr}(F_{\text{math}} \wedge \star F_{\text{math}}) = \frac{g_{\text{YM}}^2}{2} F^a_{\text{phys}} \wedge \star F^a_{\text{phys}}$, a positive energy density. For SU(2) the physicist's A^a and our A^a of (12.9) are related by $A^a = g_{\text{YM}} A^a_{\text{phys}}$.

Homology 3-spheres

Nash adds an assumption in a footnote: $H_1(M;\mathbb{Z}) = 0$.

Working definition 12.3 (homology 3-sphere)

A closed oriented 3-manifold M is a **homology 3-sphere** if it has the same integer homology as S^3 [$H_k(M;\mathbb{Z})$ records the closed k-dimensional cycles in M that are not boundaries, counted with integer coefficients]:

$$H_0(M;\mathbb{Z}) = \mathbb{Z}, \quad H_1(M;\mathbb{Z}) = 0, \quad H_2(M;\mathbb{Z}) = 0, \quad H_3(M;\mathbb{Z}) = \mathbb{Z}.$$

Test: compute $H_1(M;\mathbb{Z})$, which is the abelianisation $\pi_1(M)/[\pi_1,\pi_1]$ of the fundamental group [the loop group $\pi_1(M)$ with every commutator $aba^{-1}b^{-1}$ set equal to 1]. If it vanishes, M is a homology sphere; the other groups follow automatically.

Why do the others follow? $H_0 = H_3 = \mathbb{Z}$ for any closed connected oriented 3-manifold. For H_2 we use Poincaré duality and the universal coefficient theorem:

$$\begin{aligned} H_2(M;\mathbb{Z}) &\overset{(1)}{=} H^1(M;\mathbb{Z}) \\ &\overset{(2)}{=} \operatorname{Hom}\big(H_1(M;\mathbb{Z}),\mathbb{Z}\big) \\ &\overset{(3)}{=} 0. \end{aligned}$$

(1) Poincaré duality in dimension 3: $H_k \cong H^{3-k}$.
(2) H^1 is always torsion-free and equals the homomorphisms from H_1 to $\mathbb{Z}$ (a class in H^1 is "what you get when you integrate a closed 1-form around 1-cycles").
(3) $H_1 = 0$.

In words: a homology sphere has no non-contractible surfaces and no *abelian* loops, but it may have a highly non-abelian fundamental group.

Worked example 12.5: Homology spheres that are not spheres

- S^3 itself, with $\pi_1 = 1$.
- The **Poincaré homology sphere** $\Sigma(2,3,5)$. It can be built by gluing opposite faces of a solid dodecahedron with a $\pi/5$ twist, or as the quotient $\mathrm{SU}(2)/I^*$, where $I^* \subset \mathrm{SU}(2)$ is the binary icosahedral group of order 120. Then $\pi_1 = I^*$, which has the presentation $\langle x, y \mid (xy)^2 = x^3 = y^5 \rangle$. This group is *perfect*: every element is a product of commutators, so $H_1 = I^*/[I^*,I^*] = 0$ even though π_1 has 120 elements.
- The **Brieskorn spheres** $\Sigma(p,q,r) = \{z_1^p + z_2^q + z_3^r = 0\} \cap S^5 \subset \mathbb{C}^3$ for pairwise coprime $p,q,r \geq 2$. The case $(2,3,5)$ is the Poincaré sphere. Their fundamental groups are $\langle x,y,z \mid x^p = y^q = z^r = xyz \rangle$ (the element xyz is central).

Let us check $H_1(\Sigma(2,3,5)) = 0$ from the presentation. To abelianise, let x, y commute and write the group additively. The relations $(xy)^2 = x^3 = y^5$ become two linear equations for the two generators x, y:

$$\begin{aligned} 2x + 2y = 3x \ \text{ and } \ 3x = 5y &\overset{(1)}{=} x = 2y \ \text{ and } \ 3x = 5y \\ &\overset{(2)}{=} 6y = 5y \\ &\overset{(3)}{=} y = 0, \ \ x = 0. \end{aligned}$$

(1) Rearranging the first equation, $x = 2y$.
(2) Substitute $x = 2y$ into the second.
(3) Subtract; then $x = 2y = 0$.

So both generators die: $H_1 = 0$. The same computation in matrix form: the relation matrix has rows $(1, -2)$ and $(3, -5)$, with determinant $-5 + 6 = 1$, and an integer matrix of determinant ± 1 presents the trivial group. For a general Brieskorn sphere, the analogous determinant is ± 1 precisely because p, q, r are pairwise coprime.

Watch out: Presentations and abelianisation

Abelianising is a mechanical recipe: write each relation additively, put the coefficients into an integer matrix, and read off $\mathbb{Z}^{\#\text{gens}}$ modulo the row lattice. The order of the result (when finite) is the gcd of the maximal minors. A small slip in a presentation changes the answer completely: $\langle a, b \mid a^2 = b^3 = (ab)^5 \rangle$, which looks similar, has abelianisation $\mathbb{Z}_{19}$, not 0. Always use a presentation you have checked against a known invariant.

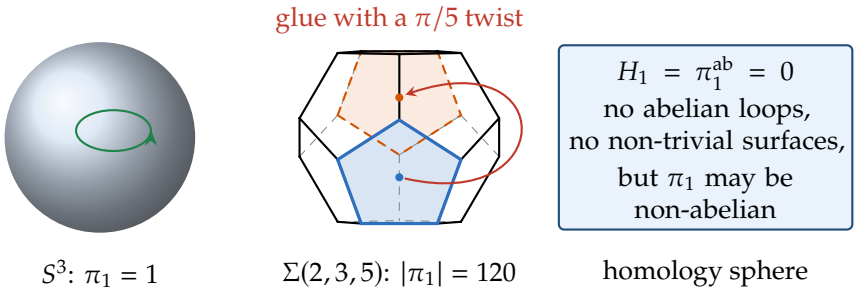


Figure 12.5: Two homology 3-spheres. Left: S^3, where every loop contracts. Middle: the Poincaré sphere, obtained from a dodecahedron by gluing each face to the opposite one after a twist by $\pi/5$; the red arrow shows one such pair, the front face (blue) and the back face (dashed orange). Its fundamental group has 120 elements but no abelian quotient. Floer theory detects exactly this kind of non-abelian topology, through flat SU(2) connections.

Why impose $H_1 = 0$? Nash says: "to avoid difficulties with reducible connections". We will see the reason in Section 12.3.2: on a homology sphere, the only flat connection with a large stabiliser is the trivial one.

12.2.3 Every SU(2) bundle over a 3-manifold is trivial

Before writing down f, Nash digresses to show that there is only one SU(2) bundle over M, the product $M \times \text{SU}(2)$. This matters because then a connection is a single global matrix of 1-forms, and gauge transformations are simply maps $M \to \text{SU}(2)$.

Picture 1: $\mathrm{SU}(2)$ is the 3-sphere of unit quaternions

Every $g \in \mathrm{SU}(2)$ has the form

$$g = \begin{pmatrix} \alpha & -\bar{\beta} \\ \beta & \bar{\alpha} \end{pmatrix}, \qquad |\alpha|^2 + |\beta|^2 = 1, \tag{12.11}$$

since the columns must be orthonormal and $\det g = 1$. Writing $\alpha = x^4 + ix^3$, $\beta = x^2 + ix^1$ (one convenient choice), $g = x^4\,\mathbf{1} + i(x^1\sigma_1 + x^2\sigma_2 + x^3\sigma_3)$ and the condition is $\sum_\mu (x^\mu)^2 = 1$. So $\mathrm{SU}(2)$ is the unit sphere $S^3 \subset \mathbb{R}^4$. The matrices $1, i\sigma_1, i\sigma_2, i\sigma_3$ multiply like the quaternions $1, \mathbf{k}, \mathbf{j}, \mathbf{i}$ (up to ordering), so $\mathrm{SU}(2)$ is the group of unit quaternions.

Picture 2: the classifying space $B\mathrm{SU}(2) = \mathbb{H}P^\infty$

Principal G-bundles over M are classified by homotopy classes of maps $[M, BG]$ into the classifying space (Chapter 1). In Chapter 7 we saw $B\mathrm{U}(1) = \mathbb{C}P^\infty$: $\mathrm{U}(1)$ acts freely on the unit sphere $S^\infty \subset \mathbb{C}^\infty$ by scalar multiplication, S^∞ is contractible, and the quotient is $\mathbb{C}P^\infty$. Replace $\mathbb{C}$ by the quaternions $\mathbb{H}$: the unit quaternions $\mathrm{SU}(2)$ act freely on the unit sphere $S^\infty \subset \mathbb{H}^\infty$ by right multiplication, and the quotient is quaternionic projective space. Hence

$$B\mathrm{SU}(2) = \mathbb{H}P^\infty.$$

$\mathbb{H}P^n$ has a cell decomposition with one cell in each dimension $0, 4, 8, \dots, 4n$ (the quaternionic analogue of $\mathbb{C}P^n = \mathbb{C}P^{n-1} \cup e^{2n}$: a point of $\mathbb{H}P^n$ not in $\mathbb{H}P^{n-1}$ is $[q_0 : \cdots : q_{n-1} : 1]$, a copy of $\mathbb{H}^n = \mathbb{R}^{4n}$). Since all cells are in dimensions divisible by 4, no boundary map can connect two of them, and each contributes one Betti number:

$$\begin{aligned} P_t(B\mathrm{SU}(2)) &\overset{(1)}{=} \sum_k b_k\, t^k \\ &\overset{(2)}{=} 1 + t^4 + t^8 + \cdots \\ &\overset{(3)}{=} \frac{1}{1 - t^4}. \end{aligned} \tag{12.12}$$

(1) Definition of the Poincaré polynomial (series).
(2) One cell, hence one Betti number, in each dimension $4j$; the cellular boundary maps vanish because adjacent cells differ in dimension by 4, not 1.
(3) Geometric series in t^4.

This is Nash's (12.3).

Picture 3: cellular approximation

The **n-skeleton** of a cell complex is the union of its cells of dimension at most n. The **cellular approximation theorem** (Whitehead) says that any map from an n-dimensional cell complex into a cell complex Y is homotopic to one landing in the n-skeleton of Y. Apply it to M (dimension 3) and $Y = \mathbb{H}P^\infty$: the 3-skeleton of $\mathbb{H}P^\infty$ is just the 0-cell, a point. So every map $M \to B\mathrm{SU}(2)$ is homotopic to a constant map, and the constant map classifies the trivial bundle. Hence every $\mathrm{SU}(2)$ bundle over M is isomorphic to $M \times \mathrm{SU}(2)$.

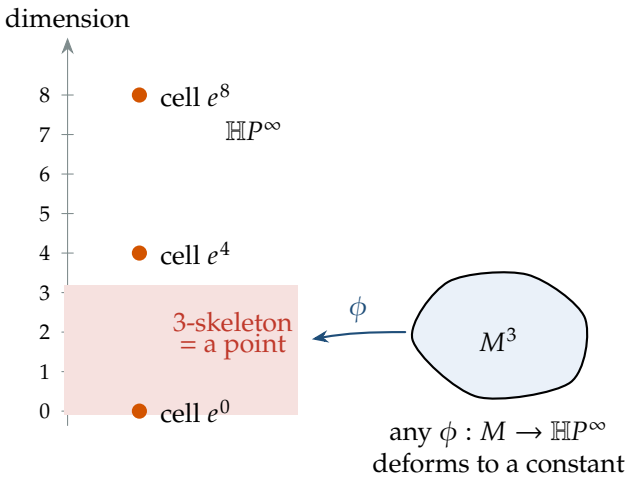


Figure 12.6: Why SU(2) bundles over a 3-manifold are trivial. The classifying space $\mathbb{H}P^\infty$ has cells only in dimensions $0, 4, 8, \dots$. A map from a 3-dimensional M can be pushed down into the 3-skeleton (shaded), which is a single point. The constant map classifies the trivial bundle.

Physical picture: The same fact, by hand

Here is a more hands-on version. Build M out of cells of dimension $0, 1, 2, 3$. Over each cell the bundle is trivial (cells are contractible). Try to extend a global trivialisation cell by cell. On a k-cell D^k we already have a trivialisation over its boundary S^{k-1}, and on the cell itself another one; they differ by a map $S^{k-1} \to \mathrm{SU}(2) = S^3$. We can adjust the trivialisation on D^k to match if that map extends over D^k, i.e. if it is null-homotopic. For $k \le 3$ the sphere S^{k-1} has dimension at most 2, and every map $S^j \to S^3$ with $j \le 2$ is null-homotopic ($\pi_j(S^3) = 0$ for $j < 3$). So the extension never fails. The first possible obstruction lives on 4-cells and is measured by $\pi_3(S^3) = \mathbb{Z}$: that is the instanton number, which is exactly why SU(2) bundles over 4-manifolds are labelled by an integer k.

Consequences

Because the bundle is trivial:

- a connection is a global $\mathfrak{su}(2)$-valued 1-form A on M, and $\mathcal{A}$ is an affine space: any two connections differ by an $\mathfrak{su}(2)$-valued 1-form. Its tangent space at any point is

$$T_A\mathcal{A} = \Omega^1(M) \otimes \mathfrak{su}(2), \tag{12.13}$$

  which Nash writes as $\Omega^1(M) \times \mathfrak{su}(2)$;

- the gauge group is the group of all smooth maps,

$$\mathcal{G} = \mathrm{Map}(M, \mathrm{SU}(2)), \tag{12.14}$$

  with pointwise multiplication, acting by (12.10). Nash calls this the "untwisted" form: over a non-trivial bundle, gauge transformations would be sections of a twisted bundle of groups.

12.2.4 The Chern–Simons function

Now the function. Floer's choice is the integral of the Chern–Simons secondary characteristic class of Chapter 8 (Nash's (12.4)):

$$\boxed{f : \mathcal{A} \to \mathbb{R}, \qquad f(A) = -\frac{1}{8\pi^2}\int_M \operatorname{tr}\Big(A\wedge \mathrm{d}A + \tfrac{2}{3}A\wedge A\wedge A\Big).} \tag{12.15}$$

We will also write $\mathrm{CS}(A) = \operatorname{tr}(A\wedge \mathrm{d}A + \frac{2}{3}A^3)$ for the 3-form in the integrand, so $f(A) = -\frac{1}{8\pi^2}\int_M \mathrm{CS}(A)$. (Wedge products of matrix-valued forms are understood; $A^3 = A\wedge A\wedge A$.)

Why this function: it is a "potential" for $\operatorname{tr} F\wedge F$

Where does (12.15) come from? From four dimensions. The instanton-number density $\operatorname{tr}(F\wedge F)$ is closed, and locally exact:

$$\begin{aligned}
\mathrm{d}\,\mathrm{CS}(A) &\overset{(1)}{=} \operatorname{tr}(\mathrm{d}A\wedge \mathrm{d}A) + \tfrac{2}{3}\operatorname{tr}\,\mathrm{d}(A\wedge A\wedge A)\\
&\overset{(2)}{=} \operatorname{tr}(\mathrm{d}A\wedge \mathrm{d}A) + 2\operatorname{tr}(A\wedge A\wedge \mathrm{d}A)\\
&\overset{(3)}{=} \operatorname{tr}(\mathrm{d}A\wedge \mathrm{d}A) + \operatorname{tr}(\mathrm{d}A\wedge A^2) + \operatorname{tr}(A^2\wedge \mathrm{d}A) + \operatorname{tr}(A^4)\\
&\overset{(4)}{=} \operatorname{tr}(F\wedge F).
\end{aligned} \tag{12.16}$$

(1) $\mathrm{d}\operatorname{tr}(A\wedge \mathrm{d}A) = \operatorname{tr}(\mathrm{d}A\wedge \mathrm{d}A) - \operatorname{tr}(A\wedge \mathrm{d}\mathrm{d}A)$ and $\mathrm{d}\mathrm{d} = 0$.

(2) $\mathrm{d}(A^3) = \mathrm{d}A\,A^2 - A\,\mathrm{d}A\,A + A^2\,\mathrm{d}A$; under the trace, moving a p-form past a q-form costs $(-1)^{pq}$, so $\operatorname{tr}(\mathrm{d}A\,A^2) = \operatorname{tr}(A^2\mathrm{d}A)$ and $\operatorname{tr}(A\,\mathrm{d}A\,A) = -\operatorname{tr}(A^2\mathrm{d}A)$; the three terms add to $3\operatorname{tr}(A^2\mathrm{d}A)$.

(3) Split $2\operatorname{tr}(A^2\mathrm{d}A)$ as $\operatorname{tr}(\mathrm{d}A\,A^2) + \operatorname{tr}(A^2\mathrm{d}A)$ and add $\operatorname{tr}(A^4)$, which is zero: $\operatorname{tr}(A\cdot A^3) = (-1)^{1\cdot 3}\operatorname{tr}(A^3\cdot A) = -\operatorname{tr}(A^4)$.

(4) Expand $F\wedge F = (\mathrm{d}A + A^2)\wedge(\mathrm{d}A + A^2)$.

So on a 4-manifold W with boundary $\partial W = M$, Stokes gives $\int_W \operatorname{tr}(F\wedge F) = \int_M \mathrm{CS}(A)$. With the normalisation of the instanton number $k = \frac{1}{8\pi^2}\int \operatorname{tr}(F\wedge F)$ (conventions of Chapter 8; see §12.4.2 for the sign), $-f(A)$ is "the instanton number of W read off from its boundary". That is why the $\frac{1}{8\pi^2}$ is there.

Worked example 12.6: The abelian Chern–Simons functional is magnetic helicity

For a U(1) gauge field write $A = -i\,a$ with a real 1-form $a = a_i\mathrm{d}x^i$ on $\mathbb{R}^3$ (or T^3). Then $A\wedge A = 0$ and

$$\begin{aligned}
\mathrm{CS}(A) &\overset{(1)}{=} (-i)^2\, a\wedge \mathrm{d}a\\
&\overset{(2)}{=} -\,\vec{a}\cdot(\nabla\times\vec{a})\,\mathrm{d}^3x = -\,\vec{a}\cdot\vec{B}\,\mathrm{d}^3x.
\end{aligned}$$

(1) The cubic term vanishes for an abelian field, and there is no trace to take.

(2) In $\mathbb{R}^3$, $a\wedge \mathrm{d}a = \epsilon^{ijk}a_i\partial_j a_k\,\mathrm{d}^3x$, which is $\vec{a}\cdot(\nabla\times\vec{a})\,\mathrm{d}^3x$ (Dray's dictionary between forms and vector calculus, Chapter T1).

The integral $\int \vec{a}\cdot\vec{B}\,\mathrm{d}^3x$ is the **magnetic helicity** of plasma physics. For two thin closed flux tubes with fluxes Φ_1, Φ_2 it equals $2\,\Phi_1\Phi_2\,\mathrm{Lk}(C_1, C_2)$, twice the product of fluxes times the Gauss linking number of their core curves. So the Chern–Simons functional measures *how field lines are linked*. This is the seed of the knot theory in the later sections of the chapter.

Note that it involves no metric: $a \wedge \mathrm{d}a$ is a 3-form, which can be integrated without a volume element.

12.2.5 The Chern–Simons function is gauge invariant only modulo integers

Would f descend to a function on $\mathcal{A}/\mathcal{G}$? Only if $f(A^g) = f(A)$ for every gauge transformation. The truth (Nash's (12.5)) is

$$\boxed{f(A^g) = f(A) + n, \qquad n = \deg g \in \mathbb{Z}.} \tag{12.17}$$

Nash quotes this from Nash and Sen. We now derive it.

Step 1: how CS changes

Write $\theta = g^{-1}\mathrm{d}g$ (the pull-back of the Maurer–Cartan form) and $\eta = \mathrm{d}g\, g^{-1}$. Two identities do all the work:

$$\mathrm{d}\eta \overset{(1)}{=} -\mathrm{d}g \wedge \mathrm{d}(g^{-1}) \overset{(2)}{=} \mathrm{d}g\, g^{-1} \wedge \mathrm{d}g\, g^{-1} = \eta \wedge \eta, \tag{12.18}$$
$$\operatorname{tr}(\eta^3) \overset{(3)}{=} \operatorname{tr}\big(g\,\theta^3 g^{-1}\big) = \operatorname{tr}(\theta^3).$$

(1) $\mathrm{d}(\mathrm{d}g\, g^{-1}) = \mathrm{d}\mathrm{d}g\, g^{-1} - \mathrm{d}g \wedge \mathrm{d}g^{-1}$, and $\mathrm{d}\mathrm{d}g = 0$.
(2) Differentiate $g\, g^{-1} = 1$: $\mathrm{d}(g^{-1}) = -g^{-1}\mathrm{d}g\, g^{-1}$.
(3) $\eta = g\,\theta\, g^{-1}$, and the trace is conjugation invariant.

Now $A^g = g^{-1}(A + \eta)g$. Put $\tilde{A} = A + \eta$, so $A^g = g^{-1}\tilde{A}g$. Differentiating,

$$\begin{aligned}\mathrm{d}A^g &\overset{(1)}{=} \mathrm{d}(g^{-1})\,\tilde{A}\,g + g^{-1}\mathrm{d}\tilde{A}\,g - g^{-1}\tilde{A} \wedge \mathrm{d}g \\ &\overset{(2)}{=} g^{-1}\big(\mathrm{d}\tilde{A} - \eta \wedge \tilde{A} - \tilde{A} \wedge \eta\big)g.\end{aligned}$$

(1) Leibniz rule; the minus sign appears because d passes the 1-form $\tilde{A}$ before hitting g.
(2) $\mathrm{d}(g^{-1}) = -g^{-1}\mathrm{d}g\, g^{-1} = -g^{-1}\eta$ and $\mathrm{d}g = \eta\, g$.

Therefore, using the cyclicity of the trace (for 1-forms $\operatorname{tr}(\alpha\beta\gamma) = \operatorname{tr}(\gamma\alpha\beta)$),

$$\begin{aligned}\mathrm{CS}(A^g) &\overset{(1)}{=} \operatorname{tr}\big(\tilde{A}\,\mathrm{d}\tilde{A} - \tilde{A}\eta\tilde{A} - \tilde{A}^2\eta\big) + \tfrac{2}{3}\operatorname{tr}\tilde{A}^3 \\ &\overset{(2)}{=} \mathrm{CS}(\tilde{A}) - 2\operatorname{tr}(\tilde{A}^2\eta).\end{aligned}$$

(1) The conjugation by g drops out of every trace.
(2) $\operatorname{tr}(\tilde{A}\eta\tilde{A}) = \operatorname{tr}(\tilde{A}^2\eta)$ by cyclicity.

Now expand $\tilde{A} = A + \eta$ in both terms. Using $\mathrm{d}\eta = \eta^2$ and cyclicity repeatedly:

$$\begin{aligned}\mathrm{CS}(\tilde{A}) &= \mathrm{CS}(A) + \operatorname{tr}(A\,\mathrm{d}\eta) + \operatorname{tr}(\eta\,\mathrm{d}A) + \operatorname{tr}(\eta\,\mathrm{d}\eta) + 2\operatorname{tr}(A^2\eta) + 2\operatorname{tr}(A\eta^2) + \tfrac{2}{3}\operatorname{tr}\eta^3, \\ -2\operatorname{tr}(\tilde{A}^2\eta) &= -2\operatorname{tr}(A^2\eta) - 4\operatorname{tr}(A\eta^2) - 2\operatorname{tr}\eta^3.\end{aligned}$$

(In the first line the cube $\operatorname{tr}(A+\eta)^3$ gives $\operatorname{tr}A^3 + 3\operatorname{tr}A^2\eta + 3\operatorname{tr}A\eta^2 + \operatorname{tr}\eta^3$, multiplied by $\frac{2}{3}$. In the second, $\operatorname{tr}(A+\eta)^2\eta = \operatorname{tr}A^2\eta + \operatorname{tr}A\eta^2 + \operatorname{tr}\eta A\eta + \operatorname{tr}\eta^3$ and $\operatorname{tr}\eta A\eta = \operatorname{tr}A\eta^2$.) Adding, and using $\operatorname{tr}(A\,\mathrm{d}\eta) = \operatorname{tr}(A\eta^2)$ and $\operatorname{tr}(\eta\,\mathrm{d}\eta) = \operatorname{tr}\eta^3$,

$$\begin{aligned}\mathrm{CS}(A^g) &\overset{(1)}{=} \mathrm{CS}(A) + \operatorname{tr}(\eta\,\mathrm{d}A) - \operatorname{tr}(A\eta^2) - \tfrac{1}{3}\operatorname{tr}\eta^3 \\ &\overset{(2)}{=} \mathrm{CS}(A) + \mathrm{d}\,\operatorname{tr}(A \wedge \eta) - \tfrac{1}{3}\operatorname{tr}(\theta^3).\end{aligned} \tag{12.19}$$

(1) Collect: $\operatorname{tr} A\eta^2$ appears with coefficient $1+2-4=-1$, $\operatorname{tr} A^2\eta$ with $2-2=0$, $\operatorname{tr}\eta^3$ with $1+\frac{2}{3}-2=-\frac{1}{3}$.
(2) $\mathrm{d}\operatorname{tr}(A\eta)=\operatorname{tr}(\mathrm{d}A\,\eta)-\operatorname{tr}(A\,\mathrm{d}\eta)=\operatorname{tr}(\eta\,\mathrm{d}A)-\operatorname{tr}(A\eta^2)$, and $\operatorname{tr}\eta^3=\operatorname{tr}\theta^3$ by (12.18).

Step 2: integrate

On a closed M the exact term integrates to zero (Stokes, $\partial M=\varnothing$):

$$\begin{aligned} f(A^g) &\overset{(1)}{=} -\frac{1}{8\pi^2}\int_M \mathrm{CS}(A)+\frac{1}{8\pi^2}\cdot\frac{1}{3}\int_M \operatorname{tr}(\theta^3) \\ &\overset{(2)}{=} f(A)+\frac{1}{24\pi^2}\int_M \operatorname{tr}\big((g^{-1}\mathrm{d}g)^3\big). \end{aligned} \tag{12.20}$$

(1) Integrate (12.19) and multiply by $-\frac{1}{8\pi^2}$; the total derivative drops out.
(2) Definition (12.15) of f.

The last term depends only on the map $g: M\to \mathrm{SU}(2)=S^3$, and it is its **degree**.

Step 3: the integral is the degree

The 3-form $\frac{1}{24\pi^2}\operatorname{tr}(\theta^3)$ is the pull-back by g of a 3-form ω on $\mathrm{SU}(2)$, namely $\omega=\frac{1}{24\pi^2}\operatorname{tr}((h^{-1}\mathrm{d}h)^3)$ with h the coordinate on the group. For any map g between closed oriented 3-manifolds, $\int_M g^*\omega=\deg(g)\int_{S^3}\omega$ (the degree counts preimages with orientation signs; Chapter 1). So we only need $\int_{\mathrm{SU}(2)}\omega=1$.

Worked example 12.7: The identity map has winding number one

Take $M=S^3=\mathrm{SU}(2)$ and $g=\mathrm{id}$, i.e. $g=x^4\mathbf{1}+ix^a\sigma_a$. At the identity element ($x^4=1$, $x^a=0$) the tangent directions are $\mathrm{d}x^a$ and $\theta=g^{-1}\mathrm{d}g=i\sigma_a\,\mathrm{d}x^a$. Then

$$\begin{aligned} \theta^3 &\overset{(1)}{=} i^3\,\sigma_a\sigma_b\sigma_c\,\mathrm{d}x^a\wedge\mathrm{d}x^b\wedge\mathrm{d}x^c \\ &\overset{(2)}{=} i^3\sum_{a,b,c}\varepsilon_{abc}\,\sigma_a\sigma_b\sigma_c\,\mathrm{d}x^1\wedge\mathrm{d}x^2\wedge\mathrm{d}x^3 \\ &\overset{(3)}{=} i^3\cdot 6i\cdot\mathbf{1}\,\mathrm{d}x^1\wedge\mathrm{d}x^2\wedge\mathrm{d}x^3=6\cdot\mathbf{1}\,\mathrm{d}x^{123}, \\ \operatorname{tr}\theta^3 &= 12\,\mathrm{d}x^{123}. \end{aligned}$$

(1) Expand the matrix product of three matrix-valued 1-forms.
(2) $\mathrm{d}x^a\wedge\mathrm{d}x^b\wedge\mathrm{d}x^c=\varepsilon_{abc}\,\mathrm{d}x^{123}$.
(3) For a,b,c a permutation of $1,2,3$, $\sigma_a\sigma_b\sigma_c=\varepsilon_{abc}\,\sigma_1\sigma_2\sigma_3=\varepsilon_{abc}\,i\mathbf{1}$; so the sum is $\sum\varepsilon_{abc}^2\,i=6i$.

The form $\operatorname{tr}\theta^3$ is invariant under left translation ($(hg)^{-1}\mathrm{d}(hg)=g^{-1}\mathrm{d}g$), so it equals 12 times the round volume form everywhere on the unit S^3, oriented so that $\mathrm{d}x^{123}>0$ at $\mathbf{1}$ (this is the boundary orientation of the unit ball in $\mathbb{H}=\mathbb{R}^4$ with ordered basis $1,i,j,k$). The volume of the unit S^3 is $2\pi^2$. Hence

$$\frac{1}{24\pi^2}\int_{S^3}\operatorname{tr}\theta^3\overset{(1)}{=}\frac{12\cdot 2\pi^2}{24\pi^2}=1.$$

(1) $\operatorname{tr}\theta^3=12\,\mathrm{vol}$ and $\mathrm{vol}(S^3)=2\pi^2$.

Putting the three steps together proves (12.17), with $n = \deg g$. For a general closed M, $[M, S^3] \cong \mathbb{Z}$ via the degree (Hopf's theorem), so $\mathcal{G} = \mathrm{Map}(M, \mathrm{SU}(2))$ has one connected component for each integer: gauge transformations are "large" or "small" according to their degree, and only small ones leave f unchanged.

Worked example 12.8: A gauge transformation of degree one on $\mathbb{R}^3 \cup \{\infty\}$

On $\mathbb{R}^3$ with $r = |\vec{x}|$, set

$$g(\vec{x}) = \frac{(r^2 - 1)\mathbf{1} + 2i\,\vec{x}\cdot\vec{\sigma}}{r^2 + 1}.$$

It is a unit quaternion: $\big((r^2-1)^2 + 4r^2\big)/(r^2+1)^2 = 1$. It tends to $\mathbf{1}$ as $r \to \infty$, so it extends to $S^3 = \mathbb{R}^3 \cup \{\infty\}$, and $g(\vec{0}) = -\mathbf{1}$. Every $h \in \mathrm{SU}(2)$ other than $\mathbf{1}$ is hit exactly once (this is inverse stereographic projection), so $\deg g = \pm 1$, with the sign fixed by orientations. Physicists know this as the **hedgehog**: the direction $\hat{x}\cdot\vec{\sigma}$ in the Lie algebra points radially outward. Applying it to $A = 0$ gives the pure-gauge field $g^{-1}\mathrm{d}g$, whose Chern–Simons number is $f = \pm 1$ although its curvature vanishes. In the temporal-gauge quantisation of Yang–Mills theory the classical vacua $A = g^{-1}\mathrm{d}g$ are labelled by this integer, and the instanton tunnels between neighbouring ones (Chapter 1).

Step 4: descend to $\mathcal{A}/\mathcal{G}$ with values in a circle

Since f changes by integers, compose it with $\mathbb{R} \to \mathbb{R}/\mathbb{Z}$:

$$\bar{f} : \mathcal{A}/\mathcal{G} \longrightarrow \mathbb{R}/\mathbb{Z}, \qquad \bar{f}([A]) = f(A) \bmod 1.$$

This is well defined. Its derivative $\mathrm{d}\bar{f}$ is an honest 1-form on $\mathcal{A}/\mathcal{G}$, so critical points and gradient flow make sense. Think of the angle ϑ on a circle: not a function, but $\mathrm{d}\vartheta$ is a perfectly good 1-form.

Worked example 12.9: Climbing from one flat connection to its gauge copy

Let g have degree 1 on M and $\theta = g^{-1}\mathrm{d}g$. Both $A = 0$ and $A = \theta$ are flat (the second is a gauge transform of the first). Walk along the straight line $A_s = s\,\theta$, $0 \le s \le 1$, in $\mathcal{A}$. Then

$$\mathrm{d}A_s \overset{(1)}{=} s\,\mathrm{d}\theta \overset{(2)}{=} -s\,\theta^2, \qquad F_s = \mathrm{d}A_s + A_s^2 = (s^2 - s)\,\theta^2,$$

$$\mathrm{CS}(A_s) \overset{(3)}{=} \mathrm{tr}\big(s\theta \wedge (-s\theta^2)\big) + \tfrac{2}{3}s^3\,\mathrm{tr}\,\theta^3 = \big(-s^2 + \tfrac{2}{3}s^3\big)\,\mathrm{tr}\,\theta^3,$$

$$f(A_s) \overset{(4)}{=} -\frac{1}{8\pi^2}\big(-s^2 + \tfrac{2}{3}s^3\big)\cdot 24\pi^2 = 3s^2 - 2s^3.$$

(1) θ does not depend on s.
(2) Maurer–Cartan: $\mathrm{d}(g^{-1}\mathrm{d}g) = -g^{-1}\mathrm{d}g \wedge g^{-1}\mathrm{d}g$ (differentiate g^{-1} as in (12.18)).
(3) Substitute into $\mathrm{CS} = \mathrm{tr}(A\mathrm{d}A + \frac{2}{3}A^3)$.
(4) $\int_M \mathrm{tr}\,\theta^3 = 24\pi^2 \deg g = 24\pi^2$.

So f rises smoothly from 0 to 1. The endpoints are critical ($\frac{\mathrm{d}}{\mathrm{d}s}(3s^2 - 2s^3) = 6s(1-s)$ vanishes at $s = 0, 1$), matching the fact that they are flat, and in between the connection is not flat ($F_s \neq 0$ for $0 < s < 1$). In $\mathcal{A}/\mathcal{G}$ the two endpoints are the same point, so this path is a closed loop, and going once around it raises f by 1.

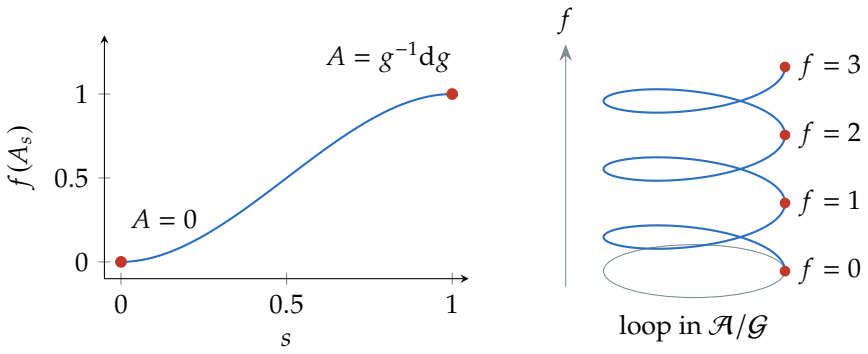


Figure 12.7: Left: the Chern–Simons function along the straight path $A_s = s\, g^{-1}\mathrm{d}g$ with $\deg g = 1$ rises from 0 to 1; both ends are flat connections (red), hence critical points. Right: the same fact globally. The loop in $\mathcal{A}/\mathcal{G}$ lifts to a spiral staircase in $\mathcal{A}$; each turn climbs by one unit of f. So f is well defined only as an angle, f mod 1.

Watch out: A function with integer ambiguity is not a sickness

Physicists meet this all the time. The action of a charged particle in a monopole field, the θ-term, and the Wess–Zumino term are all defined only modulo a constant. What must be single valued is e^{iS}, or here the 1-form $\mathrm{d}f$. For Floer theory $\mathrm{d}f$ is all we need; for the Chern–Simons *quantum* theory we will need $e^{2\pi i k f}$ to be single valued, which forces $k \in \mathbb{Z}$ (Section 12.5.4).

Reading Nash (pp. 322–324, (12.1)–(12.5))

Nash recalls the supersymmetric Morse complex of his Chapter 7 (our (12.5), (12.7)); states Floer's choice (12.2) with the footnote defining homology 3-spheres; shows that all SU(2) bundles over M^3 are trivial via $B\mathrm{SU}(2) = \mathbb{H}P^\infty$, the Poincaré series (12.3) and cellular approximation; notes $\mathcal{G} = \mathrm{Map}(M, \mathrm{SU}(2))$ and $T_A\mathcal{A} = \Omega^1(M) \times \mathfrak{su}(2)$; defines f in (12.4); and quotes (12.5), $f(A_g) = f(A) + n$. We derived (12.5) in full, identified n as the degree of g and computed it for the identity map. Note that Nash writes A_g for our A^g.

Recap

- Floer's plan: Witten's Morse complex with $X = \mathcal{A}/\mathcal{G}$ and f = Chern–Simons.
- M is a homology 3-sphere: $H_1(M;\mathbb{Z}) = 0$, hence $H_2 = 0$; π_1 can still be large (Poincaré sphere).
- SU(2) bundles over 3-manifolds are trivial (cells of $\mathbb{H}P^\infty$ start in dimension 4). So A is a global matrix of 1-forms, $\mathcal{G} = \mathrm{Map}(M, \mathrm{SU}(2))$, $T_A\mathcal{A} = \Omega^1 \otimes \mathfrak{su}(2)$.
- $f(A) = -\frac{1}{8\pi^2}\int \mathrm{tr}(A\mathrm{d}A + \frac{2}{3}A^3)$ satisfies $\mathrm{d}\,\mathrm{CS} = \mathrm{tr}\, F \wedge F$ and $f(A^g) = f(A) + \deg g$, with $\deg g = \frac{1}{24\pi^2}\int \mathrm{tr}(g^{-1}\mathrm{d}g)^3$.
- So f is an $\mathbb{R}/\mathbb{Z}$-valued function on $\mathcal{A}/\mathcal{G}$; $\mathrm{d}f$ is a genuine 1-form.

12.3 Critical points, the Hessian and the relative Morse index

We now have a function f on $\mathcal{A}$, defined on $\mathcal{A}/\mathcal{G}$ up to integers. A Morse theorist asks three questions about any function. Where are its critical points? What does the Hessian

look like at each one? How many negative eigenvalues does the Hessian have? For the Chern–Simons function the first two questions have beautiful answers: the critical points are the flat connections, and the Hessian is the "curl" operator $\star \mathrm{d}_A$. The third question has a shocking answer: *infinitely many*. Floer's way out, the relative Morse index defined by spectral flow, is the subject of the second half of the section.

12.3.1 The first variation: $\mathrm{d}f = -F/4\pi^2$

To find critical points we differentiate f along a straight line in $\mathcal{A}$ (Nash's (12.6)–(12.7)). Since $\mathcal{A}$ is an affine space, every tangent vector $a \in T_A\mathcal{A} = \Omega^1(M) \otimes \mathfrak{su}(2)$ gives a line

$$A_t = A + t\,a, \qquad t \in \mathbb{R}, \tag{12.21}$$

and $\mathrm{d}f(A)$ applied to a is $\frac{\mathrm{d}}{\mathrm{d}t} f(A_t)|_{t=0}$. The condition for a critical point is

$$\mathrm{d}f(A) = 0, \quad \text{i.e.} \quad \frac{\mathrm{d}}{\mathrm{d}t} f(A + ta)\Big|_{t=0} = 0 \ \text{ for every } a. \tag{12.22}$$

Here d is the exterior derivative *on the infinite-dimensional space* $\mathcal{A}$, not on M. We shall be careful to say which one we mean.

Two rules of the matrix-valued form calculus will be used over and over. For a p-form α and a q-form β with values in matrices,

$$\operatorname{tr}(\alpha \wedge \beta) = (-1)^{pq} \operatorname{tr}(\beta \wedge \alpha), \qquad \mathrm{d} \operatorname{tr}(\alpha \wedge \beta) = \operatorname{tr}(\mathrm{d}\alpha \wedge \beta) + (-1)^p \operatorname{tr}(\alpha \wedge \mathrm{d}\beta). \tag{12.23}$$

The first holds because the trace is cyclic on the matrix indices and reordering the forms costs $(-1)^{pq}$. The second is the Leibniz rule.

Expand the integrand of $f(A_t)$ in powers of t. The quadratic piece is

$$\operatorname{tr}\big(A_t \wedge \mathrm{d}A_t\big) = \operatorname{tr}(A\,\mathrm{d}A) + t\big[\operatorname{tr}(a\,\mathrm{d}A) + \operatorname{tr}(A\,\mathrm{d}a)\big] + t^2 \operatorname{tr}(a\,\mathrm{d}a),$$

and the cubic piece is

$$\begin{aligned} \tfrac{2}{3}\operatorname{tr}(A_t^3) = \tfrac{2}{3}\operatorname{tr}(A^3) &+ \tfrac{2}{3}t\big[\operatorname{tr}(aAA) + \operatorname{tr}(AaA) + \operatorname{tr}(AAa)\big] \\ &+ \tfrac{2}{3}t^2\big[\operatorname{tr}(aaA) + \operatorname{tr}(aAa) + \operatorname{tr}(Aaa)\big] + \tfrac{2}{3}t^3 \operatorname{tr}(a^3). \end{aligned}$$

(All products are wedge products.) This is Nash's middle line of (12.7). Now simplify the linear terms:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\operatorname{CS}(A_t)\Big|_{t=0} &\overset{(1)}{=} \operatorname{tr}(a\,\mathrm{d}A) + \operatorname{tr}(A\,\mathrm{d}a) + 2\operatorname{tr}(a\,A^2) \\ &\overset{(2)}{=} \operatorname{tr}(a\,\mathrm{d}A) + \operatorname{tr}(\mathrm{d}A\,a) - \mathrm{d}\operatorname{tr}(A\,a) + 2\operatorname{tr}(a\,A^2) \\ &\overset{(3)}{=} 2\operatorname{tr}\big(a \wedge (\mathrm{d}A + A \wedge A)\big) - \mathrm{d}\operatorname{tr}(A \wedge a) \\ &\overset{(4)}{=} 2\operatorname{tr}(F \wedge a) - \mathrm{d}\operatorname{tr}(A \wedge a). \end{aligned} \tag{12.24}$$

(1) The three cubic terms are equal: $\operatorname{tr}(AaA) = \operatorname{tr}(a\,A^2)$ and $\operatorname{tr}(A^2a) = \operatorname{tr}(a\,A^2)$ by the graded cyclicity (12.23) (moving a 1-form past a 2-form costs $(-1)^2 = +1$), so $\frac{2}{3}\cdot 3 = 2$.
(2) Leibniz (12.23): $\mathrm{d}\operatorname{tr}(A\,a) = \operatorname{tr}(\mathrm{d}A\,a) - \operatorname{tr}(A\,\mathrm{d}a)$.
(3) $\operatorname{tr}(\mathrm{d}A \wedge a) = \operatorname{tr}(a \wedge \mathrm{d}A)$ (2-form past 1-form, sign $+1$).
(4) $F = \mathrm{d}A + A \wedge A$ and $\operatorname{tr}(a \wedge F) = \operatorname{tr}(F \wedge a)$.

Integrate over the closed manifold M; the exact term drops out by Stokes, and multiplying by $-1/8\pi^2$,

$$f(A+ta) = f(A) - \frac{t}{4\pi^2}\int_M \operatorname{tr}(F\wedge a) + O(t^2). \tag{12.25}$$

This is the last line of Nash's (12.7). Reading off the linear term gives Nash's (12.8):

$$\boxed{\mathrm{d}f(A) = -\frac{F(A)}{4\pi^2}} \qquad \text{meaning} \qquad \mathrm{d}f_a(A) = -\frac{1}{4\pi^2}\int_M \operatorname{tr}\big(F(A)\wedge a\big). \tag{12.26}$$

Working definition 12.4 (the curvature as a 1-form on $\mathcal{A}$ (Nash's footnote 3))

On M^3 the curvature $F(A)$ is an $\mathfrak{su}(2)$-valued 2-form. It becomes a 1-form on the space $\mathcal{A}$ [the space of all SU(2) connections A on M, infinite-dimensional] by the rule

$$a \longmapsto F_a(A) := \int_M \operatorname{tr}\big(F(A)\wedge a\big), \qquad a \in T_A\mathcal{A}.$$

Here $a \in T_A\mathcal{A}$ is a tangent vector [a small change $A \to A + ta$ of the connection, itself an $\mathfrak{su}(2)$-valued 1-form on M]. *Why this works:* a 2-form wedged with a 1-form on a 3-manifold is a 3-form, which can be integrated; the result is linear in a. So a field strength on space is the same thing as a covector on configuration space. The finite-dimensional analogue is $\sum_i F_i\,\mathrm{d}x^i$, with the index i replaced by the pair (point of M, Lie algebra direction, form index).

Physical picture: Electric and magnetic fields

In a Hamiltonian formulation of Yang–Mills theory on space M, the configuration variable is $A_i(\vec{x})$. The magnetic field is $B^i = \frac{1}{2}\varepsilon^{ijk}F_{jk}$, which is $(\star F)^i$. Equation (12.26) says the gradient of the Chern–Simons function is the magnetic field:

$$\frac{\delta f}{\delta A_i^a(\vec{x})} \;\propto\; B_a^i(\vec{x}).$$

The critical points are the *magnetic vacua*, configurations with no magnetic field anywhere. This will be exactly the potential energy of Floer's Hamiltonian in Section 12.5.5.

12.3.2 Critical points are flat connections

From (12.26), $\mathrm{d}f(A) = 0$ for all a precisely when $F(A) = 0$. (Take $a = \star F$: then $\int \operatorname{tr}(F\wedge\star F) = -\|F\|^2$, which vanishes only if $F = 0$.) So:

Key idea: critical points of Chern–Simons

The critical points of the Chern–Simons function are the **flat connections**, $F(A) = 0$. Modulo gauge, they are in one-to-one correspondence with representations of the fundamental group,

$$\{\text{flat } A\}/\mathcal{G} \;\longleftrightarrow\; \operatorname{Hom}\big(\pi_1(M), \mathrm{SU}(2)\big)\,/\,\mathrm{Ad}\,\mathrm{SU}(2). \tag{12.27}$$

Nash states (12.27) as (12.9)–(12.10). Here is where it comes from.

Holonomy of a flat connection

Pick a base point $x_0 \in M$. For a loop γ based at x_0, the **holonomy** is the path-ordered exponential (Chapter T3)

$$\mathrm{hol}_A(\gamma) = \mathcal{P}\exp\Big(-\oint_\gamma A\Big) \in \mathrm{SU}(2),$$

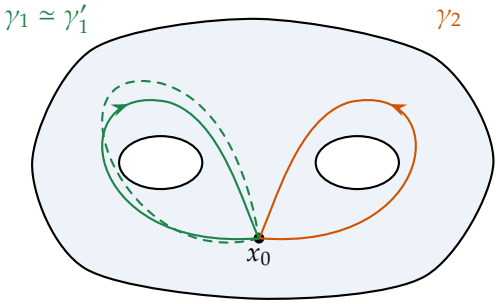
the matrix by which parallel transport around γ rotates the fibre over x_0. (The sign in the exponent is convention; nothing below depends on it.) For a general connection, deforming γ changes its holonomy, and the rate of change is the curvature: by the non-abelian Stokes theorem, the holonomy around a small loop bounding an area element $\Delta S^{\mu\nu}$ is $1 - F_{\mu\nu}\Delta S^{\mu\nu} + \cdots$. If $F = 0$ the holonomy does not change under deformation, so it depends only on the homotopy class $[\gamma] \in \pi_1(M, x_0)$. Composing loops composes parallel transports, so

$$\rho_A : \pi_1(M, x_0) \to \mathrm{SU}(2), \qquad \rho_A([\gamma_1][\gamma_2]) = \rho_A([\gamma_1])\,\rho_A([\gamma_2]) \tag{12.28}$$

is a homomorphism (up to an ordering convention).

A gauge transformation g changes parallel transport by $g(x_0)^{-1}$ at the start and $g(x_0)$ at the end, so it conjugates the whole representation: $\rho_{A^g} = g(x_0)^{-1}\rho_A\, g(x_0)$. That is the quotient by $\mathrm{Ad}\,\mathrm{SU}(2)$ in (12.27). Conversely, given ρ, build the flat bundle $\tilde{M} \times_\rho \mathrm{SU}(2)$ from the universal cover $\tilde{M}$: the product connection on $\tilde{M} \times \mathrm{SU}(2)$ descends, and its holonomy is ρ. So the correspondence is one-to-one.

$$F = 0\text{: } \mathrm{hol}(\gamma_1) = \mathrm{hol}(\gamma_1') = \rho(a), \quad \mathrm{hol}(\gamma_2) = \rho(b)$$



flat connection $\iff$ $\rho : \pi_1(M, x_0) \to \mathrm{SU}(2)$, up to conjugation

Figure 12.8: A flat connection assigns an SU(2) matrix to every loop through the base point x_0. Because $F = 0$, sliding a loop (green solid to green dashed) does not change its holonomy, so only the homotopy class matters. Different loops that cannot be deformed into each other (green and orange) can have different, even non-commuting, holonomies. (The surface drawn is a 2d cartoon of a 3-manifold with non-trivial π_1.)

Worked example 12.10: Flat connections on S^3 and on the 3-torus

S^3. $\pi_1(S^3) = 1$, so $\mathrm{Hom}(1, \mathrm{SU}(2))$ is a single point: the only flat connection up to gauge is $A = 0$. (The pure-gauge fields $g^{-1}\mathrm{d}g$ of Example 12.9 are all gauge equivalent to 0; they are the same point of $\mathcal{A}/\mathcal{G}$, with values of f differing by integers.) So f has exactly one critical point on $\mathcal{A}/\mathcal{G}$ over S^3.

T^3 (not a homology sphere, but instructive). $\pi_1(T^3) = \mathbb{Z}^3$, generated by three commuting loops. A homomorphism sends them to three *commuting* matrices $U_1, U_2, U_3 \in \mathrm{SU}(2)$.

Commuting elements of SU(2) lie in a common maximal torus, so after conjugation

$$U_j = \begin{pmatrix} e^{i\phi_j} & 0 \\ 0 & e^{-i\phi_j} \end{pmatrix}, \qquad \phi_j \in [0, 2\pi),$$

and the remaining freedom is the Weyl reflection $\phi_j \to -\phi_j$ (conjugation by $\left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$). A representing flat connection on $T^3 = \mathbb{R}^3/(2\pi\mathbb{Z})^3$ is the constant one

$$A = -\frac{i}{2\pi}\,\sigma_3 \sum_j \phi_j \, \mathrm{d}x^j,$$

whose holonomy round the j-th circle is $\exp(-2\pi A_j) = \exp(i\phi_j\sigma_3) = U_j$. It is constant, so $\mathrm{d}A = 0$, and diagonal, so $A \wedge A = 0$; hence $F = 0$. The space of flat connections is

$$\mathrm{Hom}(\mathbb{Z}^3, \mathrm{SU}(2))/\mathrm{Ad}\,\mathrm{SU}(2) \;=\; T^3/\mathbb{Z}_2,$$

a 3-dimensional orbifold. The critical points of f are *not isolated*, and every one of them is reducible (it commutes with the U(1) of diagonal matrices). This is exactly the kind of degenerate situation Floer's homology-sphere condition is designed to avoid.

Reducible and irreducible flat connections

A connection is **reducible** if some gauge transformations other than the centre $\pm\mathbf{1}$ leave it fixed. For a flat connection, $A^g = A$ means $g(x_0)$ commutes with every holonomy, i.e. with the whole image of ρ. So:

Working definition 12.5 (irreducible flat connection)

A flat SU(2) connection with holonomy ρ [the map $\rho : \pi_1(M) \to \mathrm{SU}(2)$ sending each loop to the parallel-transport matrix round it] is **irreducible** if the only matrices commuting with all of $\rho(\pi_1)$ are $\pm\mathbf{1}$. *Test:* the image of ρ is non-abelian. (If the image is abelian it lies in a maximal torus $\mathrm{U}(1) \subset \mathrm{SU}(2)$ [after conjugation, the diagonal matrices $\mathrm{diag}(e^{i\phi}, e^{-i\phi})$], and that whole U(1) commutes with it; if it is non-abelian, two non-commuting elements have only $\pm\mathbf{1}$ as common centraliser [the matrices that commute with both].)

Now Nash's argument. Suppose ρ is reducible. Then $\rho(\pi_1)$ lies in a $\mathrm{U}(1) \subset \mathrm{SU}(2)$, which is abelian. A homomorphism into an abelian group kills all commutators, so it factors through the abelianisation:

$$\pi_1(M) \;\longrightarrow\; \pi_1(M)/[\pi_1, \pi_1] = H_1(M; \mathbb{Z}) \;\longrightarrow\; \mathrm{U}(1).$$

On a homology sphere $H_1 = 0$, so ρ is trivial. Hence:

Key idea: the only reducible is the trivial connection

On a homology 3-sphere, every non-trivial flat SU(2) connection is irreducible, and the trivial connection $\theta = 0$ is the unique reducible one.

Watch out: Nash's "all irreducible"

Nash writes that the flat connections "are also all irreducible". That is true of all *non-trivial* ones. The trivial connection is always there and is maximally reducible (its stabiliser is all of SU(2)). In Floer's construction it is treated separately: the Floer complex is built from

the irreducible flat connections, and the trivial connection enters only as a reference point and as a possible end of flow lines, handled by extra analysis. On a homology sphere at least it is isolated, since $H^1(M;\mathfrak{su}(2)) = H^1(M)\otimes\mathfrak{su}(2) = 0$.

Worked example 12.11: The Poincaré sphere has two irreducible flat connections

$\pi_1(\Sigma(2,3,5)) = I^*$, the binary icosahedral group, which is itself a subgroup of SU(2). The inclusion $I^* \hookrightarrow \mathrm{SU}(2)$ is therefore a homomorphism, and it is irreducible because I^* is non-abelian. Composing with the outer automorphism of I^* (which exchanges the two classes of elements of order 5, i.e. swaps $e^{2\pi i/5}$ with $e^{4\pi i/5}$ in the eigenvalues) gives a second, inequivalent one: an element of order 10 has trace $2\cos(\pi/5) = \frac{1+\sqrt{5}}{2}$ in one and $2\cos(3\pi/5) = \frac{1-\sqrt{5}}{2}$ in the other, and conjugation cannot change a trace. A theorem of Fintushel and Stern shows these are all. So

$$\mathrm{Hom}(I^*, \mathrm{SU}(2))/\mathrm{Ad}\,\mathrm{SU}(2) = \{\theta,\ \rho_1,\ \rho_2\},$$

the trivial connection plus two irreducible ones. The function $\bar{f}$ on $\mathcal{A}/\mathcal{G}$ has exactly three critical points, and two of them are the building blocks of the Floer complex.

12.3.3 The Hessian is the curl operator $\star \mathrm{d}_A$

At a critical point, the second-order term of $f(A + ta)$ defines the Hessian. We expand to order t^2 (Nash's (12.11)); the coefficient of $t^2/2$ is the quadratic form $\mathrm{Hess}_A(a,a)$.

First we need the covariant derivative on $\mathfrak{su}(2)$-valued 1-forms. For a p-form ω with values in the adjoint representation,

$$\mathrm{d}_A\omega = \mathrm{d}\omega + A\wedge\omega - (-1)^p\,\omega\wedge A, \qquad \text{so} \qquad \mathrm{d}_A a = \mathrm{d}a + A\wedge a + a\wedge A \ \text{ for a 1-form } a. \tag{12.29}$$

(For $p = 1$ the "commutator" is an anticommutator of matrix-valued 1-forms, because the forms themselves anticommute.)

$$\begin{aligned}
\mathrm{Hess}_A(a,a) &\overset{(1)}{=} 2\times\Big(-\frac{1}{8\pi^2}\Big)\int_M \Big\{\mathrm{tr}(a\wedge\mathrm{d}a) + \tfrac{2}{3}\big[\mathrm{tr}(aaA) + \mathrm{tr}(aAa) + \mathrm{tr}(Aaa)\big]\Big\} \\
&\overset{(2)}{=} -\frac{1}{4\pi^2}\int_M \mathrm{tr}\big\{a\wedge(\mathrm{d}a + 2\,a\wedge A)\big\} \\
&\overset{(3)}{=} -\frac{1}{4\pi^2}\int_M \mathrm{tr}\big\{a\wedge\mathrm{d}a + a\wedge A\wedge a + a\wedge a\wedge A\big\} \\
&\overset{(4)}{=} -\frac{1}{4\pi^2}\int_M \mathrm{tr}\big(a\wedge\mathrm{d}_A a\big) \\
&\overset{(5)}{=} \frac{1}{4\pi^2}\,\big\langle a, \star\mathrm{d}_A a\big\rangle.
\end{aligned} \tag{12.30}$$

(1) The t^2 terms of $f(A_t)$ collected above, multiplied by 2 because the Hessian is the coefficient of $t^2/2$.

(2) The three cubic terms are all equal to $\mathrm{tr}(a\wedge a\wedge A)$ by graded cyclicity (moving a 1-form past a 2-form is free), so $\frac{2}{3}\cdot 3 = 2$.

(3) Split $2\,\mathrm{tr}(aaA)$ as $\mathrm{tr}(aAa) + \mathrm{tr}(aaA)$, again by cyclicity.

(4) $a\wedge\mathrm{d}_A a = a\wedge(\mathrm{d}a + Aa + aA)$ by (12.29).

(5) The L^2 inner product on $\mathfrak{su}(2)$-valued forms is $\langle\alpha,\beta\rangle = -\int_M \mathrm{tr}(\alpha\wedge\star\beta)$, and $\star\star = 1$ on forms on a Riemannian 3-manifold, so $\langle a, \star\mathrm{d}_A a\rangle = -\int \mathrm{tr}(a\wedge\mathrm{d}_A a)$.

Nash's (12.11) has the same content; his first line keeps the $\frac{2}{3}$ and the three cubic terms visible, and the simplification in his second line uses exactly step (2).

Note that we did not use flatness to get (12.30): it is the second-order term at any A. At a critical point it is the honest Hessian. Using the metric to turn the quadratic form into an operator,

$$\mathrm{Hess}_A(a,b) = \langle a, H(A)b\rangle, \qquad \boxed{4\pi^2\, H(A) = \star \mathrm{d}_A} \qquad \text{acting on } \Omega^1(M)\otimes \mathfrak{su}(2). \tag{12.31}$$

Working definition 12.6 (the Hessian of Chern–Simons)

$H(A)$ takes an $\mathfrak{su}(2)$-valued 1-form a, differentiates it covariantly [$\mathrm{d}_A a = \mathrm{d}a + A\wedge a + a\wedge A$] to a 2-form, and turns the 2-form back into a 1-form with the Hodge star [in orthonormal coordinates on the 3-manifold, $\star$ sends $\mathrm{d}x^1\wedge\mathrm{d}x^2$ to $\mathrm{d}x^3$, and cyclically]. *In vector-calculus language*, for $A=0$ and one colour component, it is the curl: $\star\mathrm{d}a \leftrightarrow \nabla\times\vec{a}$ (Dray's dictionary, Chapter T1). It is symmetric: $\langle a, \star\mathrm{d}_A b\rangle = \langle \star\mathrm{d}_A a, b\rangle$ (integrate by parts on the closed M).

Why the Hessian is a disaster for ordinary Morse theory

The Morse index is the number of negative eigenvalues of the Hessian. The operator $\star\mathrm{d}_A$ is *first order*. Like the Dirac operator, and unlike the Laplacian, a first-order self-adjoint operator typically has spectrum unbounded in *both* directions. So the "Morse index" of every critical point is infinite. The following example shows this in the simplest case.

Worked example 12.12: The spectrum of curl on the 3-torus

Take $M = T^3 = \mathbb{R}^3/(2\pi\mathbb{Z})^3$, $A=0$, and one Lie-algebra component, so $H \propto \star\mathrm{d}$ acting on ordinary real 1-forms, i.e. curl acting on vector fields $\vec{a}(\vec{x})$. Fourier-expand: $\vec{a} = \vec{\varepsilon}\, e^{i\vec{k}\cdot\vec{x}}$ with $\vec{k}\in\mathbb{Z}^3$. Then

$$\nabla\times\left(\vec{\varepsilon}\, e^{i\vec{k}\cdot\vec{x}}\right) \overset{(1)}{=} i\,\vec{k}\times\vec{\varepsilon}\, e^{i\vec{k}\cdot\vec{x}}.$$

(1) $\partial_j e^{i\vec{k}\cdot\vec{x}} = ik_j e^{i\vec{k}\cdot\vec{x}}$ and $\vec{\varepsilon}$ is constant.

So on each Fourier mode curl is the 3×3 matrix $C_{\vec{k}}\,\vec{\varepsilon} = i\vec{k}\times\vec{\varepsilon}$. For $\vec{k} = (0,0,\kappa)$, $\kappa>0$,

$$C_{\vec{k}} = i\kappa\begin{pmatrix} 0 & -1 & 0\\ 1 & 0 & 0\\ 0 & 0 & 0\end{pmatrix}, \qquad \text{eigenvalues } +\kappa,\ -\kappa,\ 0,$$

with eigenvectors $\vec{\varepsilon}_\pm = (1,\pm i,0)/\sqrt{2}$ (the two circular polarisations, i.e. the two helicities) and $\vec{\varepsilon}_0 = (0,0,1)$ (the longitudinal direction, $\vec{a} = \nabla(\text{something})$). Check one:

$$C_{\vec{k}}\,\vec{\varepsilon}_+ \overset{(1)}{=} \frac{i\kappa}{\sqrt{2}}(-i,\,1,\,0) \overset{(2)}{=} \frac{\kappa}{\sqrt{2}}(1,\,i,\,0) = +\kappa\,\vec{\varepsilon}_+.$$

(1) The matrix sends $(\varepsilon_1,\varepsilon_2,\varepsilon_3)$ to $i\kappa(-\varepsilon_2,\varepsilon_1,0)$, and $\varepsilon_2 = i$, $\varepsilon_1 = 1$.
(2) $i\cdot(-i) = 1$ and $i\cdot 1 = i$.

The same computation gives $C_{\vec{k}}\vec{\varepsilon}_- = -\kappa\,\vec{\varepsilon}_-$ and $C_{\vec{k}}\vec{\varepsilon}_0 = 0$. By rotation invariance the same holds for every $\vec{k}$ with $\kappa = |\vec{k}|$. The full spectrum is

$$\mathrm{Spec}(\star\mathrm{d}) = \{+|\vec{k}|,\,-|\vec{k}| : \vec{k}\in\mathbb{Z}^3\setminus 0\}\ \cup\ \{0 \text{ (longitudinal and constant modes)}\}.$$

There are infinitely many positive *and* infinitely many negative eigenvalues: $|\vec{k}|$ grows without bound in both helicities. The Morse index at $A=0$ is $+\infty$, and so is the "co-index".

Worked example 12.13: Curl on the round 3-sphere

On the unit $S^3 = \mathrm{SU}(2)$ use the left-invariant coframe. From the Maurer–Cartan equation $\mathrm{d}\theta = -\theta \wedge \theta$ for $\theta = g^{-1}\mathrm{d}g = \theta^a T_a$ we get

$$\mathrm{d}\theta^c\, T_c \overset{(1)}{=} -\theta^a \wedge \theta^b\, T_a T_b \overset{(2)}{=} -\tfrac{1}{2}\theta^a \wedge \theta^b\, [T_a, T_b] \overset{(3)}{=} -\tfrac{1}{2}\varepsilon_{abc}\, \theta^a \wedge \theta^b\, T_c,$$

(1) Expand $\theta \wedge \theta$ in components.
(2) $\theta^a \wedge \theta^b$ is antisymmetric in a, b, so only the antisymmetric part of $T_a T_b$ survives.
(3) $[T_a, T_b] = \varepsilon_{abc} T_c$.

so $\mathrm{d}\theta^1 = -\theta^2 \wedge \theta^3$ and cyclically. At the identity $g = \mathbf{1} + ix^a\sigma_a$, so $\theta = i\sigma_a \mathrm{d}x^a = -2T_a\,\mathrm{d}x^a$, i.e. $\theta^a = -2\,\mathrm{d}x^a$. Hence $e^a := -\tfrac{1}{2}\theta^a$ is an orthonormal coframe for the round metric (it is at the identity, and everything is left-invariant), and

$$\star\mathrm{d}e^1 \overset{(1)}{=} \star\big(\tfrac{1}{2}\theta^2 \wedge \theta^3\big) \overset{(2)}{=} \star\big(2\,e^2 \wedge e^3\big) \overset{(3)}{=} 2\,e^1.$$

(1) $\mathrm{d}e^1 = -\tfrac{1}{2}\mathrm{d}\theta^1 = \tfrac{1}{2}\theta^2 \wedge \theta^3$.
(2) $\theta^a = -2e^a$, so $\theta^2 \wedge \theta^3 = 4e^2 \wedge e^3$.
(3) $\star(e^2 \wedge e^3) = e^1$ for an oriented orthonormal coframe.

So the three left-invariant 1-forms are eigenvectors of curl with eigenvalue $+2$. The same computation with right-invariant forms gives -2. In general the eigenvalues of $\star\mathrm{d}$ on co-closed 1-forms on the unit S^3 are $\pm(n+1)$, $n = 1, 2, 3, \ldots$, each with multiplicity $n(n+2)$: a spectrum symmetric about zero and unbounded both ways. (The $n = 1$ case is what we just computed: ± 2, multiplicity 3.)

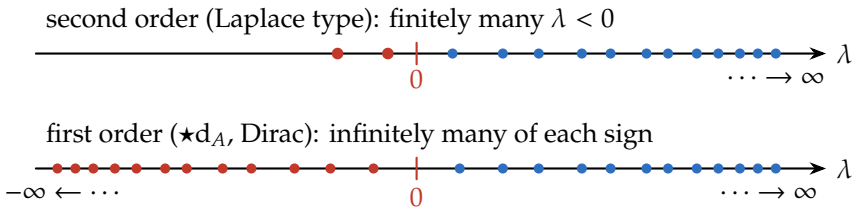


Figure 12.9: Why the Chern–Simons Hessian has infinite Morse index. Top: a Laplace-type operator (Hessian of the energy of paths, or of the Yang–Mills action) is bounded below, so only finitely many eigenvalues (red) are negative. Bottom: the first-order operator $\star\mathrm{d}_A$ has eigenvalues accumulating at both $+\infty$ and $-\infty$, like the curl on T^3 or S^3. Counting the red dots gives ∞.

Physical picture: Why geodesics and Yang–Mills were fine

In Chapter 7 the infinite-dimensional Morse functions were the energy $E(\gamma) = \frac{1}{2}\int |\dot\gamma|^2 \mathrm{d}t$ of paths, and the Yang–Mills action. Their Hessians are $-\frac{\mathrm{d}^2}{\mathrm{d}t^2} + (\text{curvature})$ and $\mathrm{d}_A^*\mathrm{d}_A + \cdots$: second order, like $p^2 + V$ in quantum mechanics. Such an operator has a lowest eigenvalue; only finitely many eigenvalues lie below zero, and the Morse index is finite. The Chern–Simons Hessian is first order, like $\vec\sigma \cdot \vec p$ for a Weyl fermion: its spectrum is a "Dirac sea", filled with infinitely many negative states.

12.3.4 The relative Morse index

How does one do Morse theory when every index is infinite? Floer's observation is that the complex of Chapter 7 never uses the index itself. Look at the coboundary (12.7) (Nash's (12.12)):

$$\cdots \xrightarrow{\delta} C^{p} \xrightarrow{\delta} C^{p+1} \xrightarrow{\delta} \cdots$$

It only needs to know *which pairs* of critical points have indices differing by one. That is, it only uses *differences* of indices. And while $\infty - \infty$ is meaningless as arithmetic, it can be made meaningful as a count of eigenvalues that *cross zero*.

Finite dimensions: index difference = eigenvalue crossings

Let X be finite-dimensional, P, Q critical points of f, and $A(t)$ a path from P to Q. Extend the Hessian to a family of symmetric matrices $H(t)$ along the path (any interpolation will do). At $t = 0$ it has λ_P negative eigenvalues; at $t = 1$ it has λ_Q. Eigenvalues move continuously, so the only way the count can change is by eigenvalues passing through zero. Hence

$$\lambda_P - \lambda_Q \;=\; \#\{\text{eigenvalues crossing } - \to + \} - \#\{\text{eigenvalues crossing } + \to - \} \;=:\; \mathrm{SF}\big(H(t)\big). \tag{12.32}$$

The right side is the **spectral flow** of the family. In finite dimensions both sides make sense and agree. In infinite dimensions only the right side makes sense, and it is still a finite integer, because only finitely many eigenvalues can cross zero along a compact path (the spectrum near zero is discrete for an elliptic operator).

Working definition 12.7 (spectral flow and the relative Morse index)

Let $H(t)$, $0 \le t \le 1$ (or $-\infty < t < \infty$), be a continuous family of self-adjoint operators with discrete spectrum and invertible ends [H has no zero eigenvalue at the two endpoints of the path]. The **spectral flow** σ is the net number of eigenvalues that cross zero from negative to positive as t increases (Chapter 10, where it was the mechanism of the anomaly). *To compute:* draw the eigenvalues as curves $\lambda_j(t)$ and count the signed zero crossings.
For Chern–Simons, choose a steepest-descent path $A(t)$ from A_P to A_Q (Nash's (12.13)) [$A(t)$ runs downhill in f as fast as possible, with grad f the vector dual to $\mathrm{d}f = -F/4\pi^2$ under the L^2 metric],

$$\frac{\mathrm{d}A(t)}{\mathrm{d}t} = -\operatorname{grad} f\big(A(t)\big), \qquad A(-\infty) = A_P,\; A(+\infty) = A_Q, \tag{12.33}$$

consider the family $4\pi^2 H(A(t)) = \star \mathrm{d}_{A(t)}$, and define the **relative Morse index** (Nash's (12.14))

$$\lambda_{A_P, A_Q} := \sigma(A_P, A_Q) = \mathrm{SF}\big(\star \mathrm{d}_{A(t)}\big). \tag{12.34}$$

In finite dimensions this is $\lambda_P - \lambda_Q$.

Nash describes it as "the net number of negative eigenvalues of $H(A_P)$ which flow into positive eigenvalues of $H(A_Q)$", and calls it a "renormalised" measure of the difference of Morse indices. The renormalisation is literal: we subtract two infinite numbers by pairing their eigenvalues along a path, exactly as one renormalises the charge of the Dirac sea by counting only level crossings.

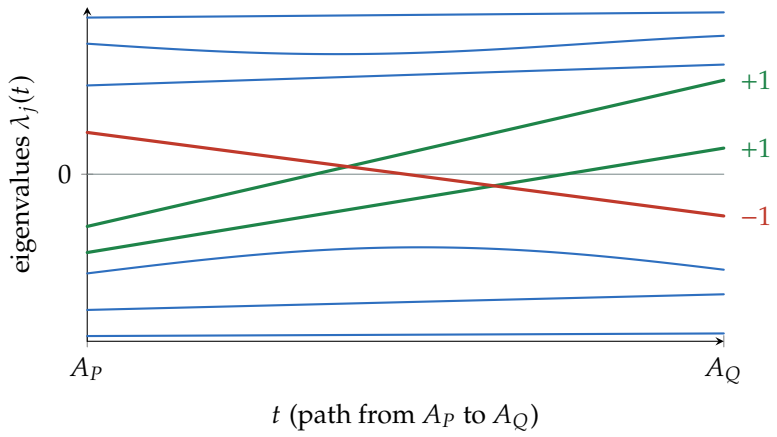


Figure 12.10: Spectral flow. The eigenvalues of $\star \mathrm{d}_{A(t)}$ along a path from A_P to A_Q extend infinitely far up and down (only a window is drawn). Blue curves never cross zero and do not count. The two green curves cross upward (+1 each), the red curve crosses downward (−1). The spectral flow is $1 + 1 - 1 = 1$: the relative Morse index $\lambda_{A_P,A_Q} = 1$, although both Morse indices are infinite.

Worked example 12.14: Spectral flow of the Dirac operator on a circle

The simplest infinite-dimensional case. On S^1 with angle ϑ, consider the family of self-adjoint operators

$$D_s = -i\frac{\mathrm{d}}{\mathrm{d}\vartheta} + s, \qquad s \in \mathbb{R},$$

(the Dirac operator coupled to a constant U(1) gauge field s). The eigenfunctions are $e^{in\vartheta}$, $n \in \mathbb{Z}$, with eigenvalues

$$D_s\, e^{in\vartheta} \overset{(1)}{=} (n+s)\, e^{in\vartheta}.$$

(1) $-i\,\partial_\vartheta e^{in\vartheta} = n\, e^{in\vartheta}$.

The spectrum $\{n + s\}$ is unbounded both ways: each D_s has infinitely many negative eigenvalues. Now move s from $-\frac{1}{2}$ to $\frac{1}{2}$. Every eigenvalue moves up by one unit; exactly one of them ($n = 0$) crosses zero, upward. So

$$\mathrm{SF}\big(D_s,\ s \in [-\tfrac{1}{2}, \tfrac{1}{2}]\big) = +1.$$

Also note that $D_{s+1} = e^{-i\vartheta} D_s\, e^{i\vartheta}$: shifting s by 1 is a (large) gauge transformation, so $s \in [0, 1]$ is a *closed loop* in the space of gauge orbits, and the spectral flow round it is 1. This is the toy model for the statement below that the spectral flow round a loop in $\mathcal{A}/\mathcal{G}$ is not zero but 8.

Spectral flow is an index

Why is spectral flow an *integer invariant*, independent of the details of the path? Because it equals the index of an operator in one dimension more. This is the bridge Nash uses to compute it.

Worked example 12.15: $\operatorname{index}(\frac{\mathrm{d}}{\mathrm{d}t}+\lambda(t))$ counts one crossing

Take one eigenvalue: a smooth function $\lambda(t)$ on $\mathbb{R}$ with limits $\lambda_\pm = \lambda(\pm\infty) \neq 0$, and the operator $D = \frac{\mathrm{d}}{\mathrm{d}t} + \lambda(t)$ on $L^2(\mathbb{R})$. Its kernel:

$$Du = 0 \overset{(1)}{=} u(t) = u(0)\exp\Big(-\int_0^t \lambda(s)\,\mathrm{d}s\Big) \sim \begin{cases} e^{-\lambda_+ t}, & t \to +\infty, \\ e^{-\lambda_- t} = e^{+\lambda_-|t|}, & t \to -\infty. \end{cases}$$

(1) Separate variables in $u' = -\lambda u$; at large $|t|$, $\int_0^t \lambda \approx \lambda_\pm t$.

So $u \in L^2$ iff $\lambda_+ > 0$ and $\lambda_- < 0$. The adjoint $D^* = -\frac{\mathrm{d}}{\mathrm{d}t} + \lambda(t)$ has solutions $\exp(+\int\lambda)$, in L^2 iff $\lambda_+ < 0$ and $\lambda_- > 0$. Therefore

$$\operatorname{index} D = \dim\operatorname{Ker} D - \dim\operatorname{Ker} D^* = \begin{cases} +1 & \lambda : - \longrightarrow + \ \text{(upward crossing)}, \\ -1 & \lambda : + \longrightarrow - \ \text{(downward crossing)}, \\ 0 & \text{no net crossing.} \end{cases}$$

For a family $H(t)$ with many eigenvalues, decompose (approximately) into eigenvalue curves and add: $\operatorname{index}\big(\frac{\mathrm{d}}{\mathrm{d}t} + H(t)\big) = \mathrm{SF}(H(t))$. (This is the Atiyah–Patodi–Singer theorem; the diagonal case above is its whole mechanism.)

So: the relative Morse index is the index of the four-dimensional operator $\frac{\mathrm{d}}{\mathrm{d}t} + \star\mathrm{d}_{A(t)}$ on $M \times \mathbb{R}$. In the next section we shall see that this operator is precisely the deformation operator of instantons, and its index is computed by the Atiyah–Singer formula of Chapter 8.

12.3.5 Path dependence and the fundamental group of $\mathcal{A}/\mathcal{G}$

Is $\sigma(A_P, A_Q)$ independent of the chosen path? Take two paths $A(t)$ and $A'(t)$ from A_P to A_Q. Running along A and back along A' is a loop ℓ, and

$$\mathrm{SF}(A) - \mathrm{SF}(A') = \mathrm{SF}(\ell).$$

Spectral flow round a loop is the index of $\frac{\mathrm{d}}{\mathrm{d}t} + \star\mathrm{d}_{A(t)}$ on $M \times S^1$ (the t-direction closed up into a circle), an Atiyah–Singer index. An index is a deformation invariant, so $\mathrm{SF}(\ell)$ depends only on the homotopy class of ℓ. Two possibilities:

- In $\mathcal{A}$, which is an affine space (contractible), every loop is contractible, so $\mathrm{SF}(\ell) = 0$ and $\sigma(A_P, A_Q)$ is well defined for points of $\mathcal{A}$.
- But the critical points are really points of $\mathcal{A}/\mathcal{G}$, and paths there can wind around non-contractible loops.

Key idea: Nash's footnote 4: $\pi_1(\mathcal{A}/\mathcal{G}) = \mathbb{Z}$

For SU(2) on a closed 3-manifold,

$$\pi_1(\mathcal{A}/\mathcal{G}) \;\cong\; \pi_0(\mathcal{G}) \;\cong\; [M, S^3] \;\cong\; \mathbb{Z}, \tag{12.35}$$

the integer being the degree of a gauge transformation.

Here is the argument, with the steps Nash compresses.

1. $\mathcal{G}$ acts on $\mathcal{A}$ and (after removing the few points with large stabilisers, or using the based gauge group of maps with $g(x_0) = \mathbf{1}$) freely, so $\mathcal{G} \to \mathcal{A} \to \mathcal{A}/\mathcal{G}$ is a fibration: a principal $\mathcal{G}$-bundle.

2. A fibration has a long exact sequence of homotopy groups; its end is

$$\pi_1(\mathcal{A}) \to \pi_1(\mathcal{A}/\mathcal{G}) \xrightarrow{\partial} \pi_0(\mathcal{G}) \to \pi_0(\mathcal{A}).$$

   The map ∂ takes a loop in $\mathcal{A}/\mathcal{G}$, lifts it to a path in $\mathcal{A}$ from A to some A^g, and returns the component of g.

3. $\mathcal{A}$ is contractible (it is an affine space: shrink everything to a chosen A_0 along straight lines). So $\pi_1(\mathcal{A}) = \pi_0(\mathcal{A}) = 0$ and ∂ is an isomorphism.

4. $\pi_0(\mathcal{G}) = \pi_0(\mathrm{Map}(M, \mathrm{SU}(2)))$ is the set of homotopy classes of maps $M^3 \to S^3$. By Hopf's theorem these are classified by the degree, an integer.

The generating loop is precisely the one in Example 12.9: the straight path from A to A^g with $\deg g = 1$, closed up in $\mathcal{A}/\mathcal{G}$.

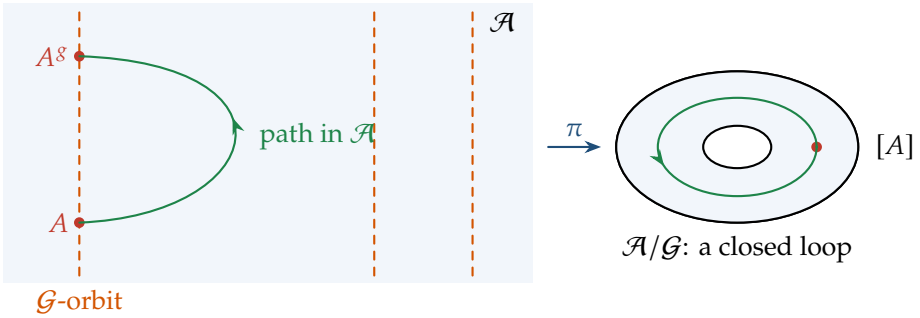


Figure 12.11: Why $\pi_1(\mathcal{A}/\mathcal{G}) = \pi_0(\mathcal{G})$. Left: the contractible space $\mathcal{A}$, with gauge orbits drawn as dashed fibres. A path from A to its gauge copy A^g projects to a closed loop in $\mathcal{A}/\mathcal{G}$ (right). The loop is non-contractible exactly when g lies in a non-trivial component of $\mathcal{G}$, i.e. has non-zero degree. The "hole" is a cartoon of that non-contractibility.

Consequently $\sigma(A_P, A_Q)$ on $\mathcal{A}/\mathcal{G}$ is defined only modulo the spectral flow round the generating loop: if that number is N, then σ is well defined in $\mathbb{Z}/N\mathbb{Z}$. Computing N is an index computation on $M \times S^1$, and it is done in the next section. The answer will be $N = 8$.

Watch out: Chern–Simons values and the grading

Going once round the generating loop changes f by 1 (Example 12.9) and the relative index by $N = 8$. So the pair (f mod 1, σ mod 8) is well defined, while $f \in \mathbb{R}$ and $\sigma \in \mathbb{Z}$ are defined only on $\mathcal{A}$. In more refined versions of Floer theory one keeps track of both at once, using $\mathcal{A}/\mathcal{G}_0$ with $\mathcal{G}_0$ the degree-zero gauge transformations; then the grading is a genuine integer.

Reading Nash (pp. 324–327, (12.6)–(12.14))

Nash expands $f(A + ta)$ to first order to get $\mathrm{d}f = -F/4\pi^2$ (12.7)–(12.8) and explains in footnote 3 how F is a 1-form on $\mathcal{A}$. He identifies the critical points with flat connections, i.e. with $\mathrm{Hom}(\pi_1(M), \mathrm{SU}(2))$ modulo conjugation (12.9)–(12.10), and argues irreducibility from $H_1 = 0$. He expands to second order to get the Hessian $4\pi^2 H(A) = \star \mathrm{d}_A$ (12.11), observes that it is first order with infinite Morse index, and contrasts it with the geodesic

and Yang–Mills cases. He recalls the coboundary (12.12) and defines the relative Morse index by the spectral flow along a steepest-descent path (12.13)–(12.14), discussing path dependence and $\pi_1(\mathcal{A}/\mathcal{G}) = \mathbb{Z}$ in footnote 4. We added the graded-trace details, the holonomy construction, the T^3 and Poincaré-sphere examples, the explicit curl spectra, the circle example of spectral flow, and the one-dimensional computation showing that spectral flow is an index.

Recap

- $\mathrm{d}f(A) = -F(A)/4\pi^2$: the gradient of Chern–Simons is the magnetic field. Critical points = flat connections = $\mathrm{Hom}(\pi_1, \mathrm{SU}(2))/\mathrm{Ad}$.
- On a homology sphere every non-trivial flat connection is irreducible; the trivial one is the unique reducible.
- Hessian: $4\pi^2 H(A) = \star \mathrm{d}_A$, a first-order "curl" operator with spectrum unbounded in both directions; Morse indices are infinite.
- Relative index λ_{A_P, A_Q} = spectral flow of $\star \mathrm{d}_{A(t)}$ = $\mathrm{index}\big(\frac{\mathrm{d}}{\mathrm{d}t} + \star \mathrm{d}_{A(t)}\big)$; finite.
- $\pi_1(\mathcal{A}/\mathcal{G}) = \mathbb{Z}$, so the relative index is defined modulo the spectral flow round the generating loop.

12.4 Gradient flow lines are instantons

In Witten's Morse complex the coboundary counts gradient flow lines between critical points. So to build Floer's complex we need to understand the gradient flow of the Chern–Simons function. The remarkable fact, and the reason Floer theory is a gauge theory rather than a piece of abstract analysis, is that a gradient flow line of Chern–Simons on a 3-manifold M *is* an instanton on the 4-manifold $M \times \mathbb{R}$. Once we see this, the missing number from the last section, the spectral flow round a loop, is an instanton-moduli dimension, and we know how to compute those (Chapter 8). The answer is 8.

12.4.1 A metric on $\mathcal{A}$ and the gradient of f

A gradient needs a metric. The flow (Nash's (12.15)) is

$$\frac{\mathrm{d}A(t)}{\mathrm{d}t} = -\operatorname{grad} f\big(A(t)\big), \tag{12.36}$$

and "grad" converts the 1-form $\mathrm{d}f$ into a vector using an inner product on $T_A\mathcal{A}$. The tangent vectors are $\mathfrak{su}(2)$-valued 1-forms on M, so the natural choice (Nash's (12.16)) is the L^2 inner product built from the metric on M and the trace:

$$\langle a, b\rangle = -\int_M \operatorname{tr}(a \wedge \star b), \qquad a, b \in T_A\mathcal{A} \cong \Omega^1(M) \otimes \mathfrak{su}(2). \tag{12.37}$$

Let us check that it is positive. Write $a = a^i_\mu\, T_i\, \mathrm{d}x^\mu$ with $\operatorname{tr}(T_iT_j) = -\frac{1}{2}\delta_{ij}$ (from (12.9)). Then

$$\begin{aligned} -\operatorname{tr}(a \wedge \star a) &\overset{(1)}{=} -\operatorname{tr}(T_iT_j)\, a^i_\mu a^j_\nu\, \mathrm{d}x^\mu \wedge \star \mathrm{d}x^\nu \\ &\overset{(2)}{=} \tfrac{1}{2}\,\delta_{ij}\, a^i_\mu a^j_\nu\, g^{\mu\nu}\,\mathrm{vol} \end{aligned}$$

$$= \tfrac{1}{2}\sum_i |a^i|^2 \,\mathrm{vol} \;\geq 0.$$

(1) Pull the Lie-algebra generators out of the wedge product.
(2) $\mathrm{tr}(T_iT_j) = -\tfrac{1}{2}\delta_{ij}$, and $\mathrm{d}x^\mu \wedge \star \mathrm{d}x^\nu = g^{\mu\nu}\,\mathrm{vol}$ (the defining property of $\star$, Chapter T1).

So $\langle a, a\rangle = \tfrac{1}{2}\int_M \sum_i |a^i|^2 \mathrm{vol}$, the sum of the squared lengths of the colour components: the obvious "kinetic" metric.

The gradient is defined by $\langle \mathrm{grad}\, f, a\rangle = \mathrm{d}f_a$ for every a (Nash's (12.17)). Using (12.26):

$$\begin{aligned} \mathrm{d}f_a &\overset{(1)}{=} -\frac{1}{4\pi^2}\,F_a(A) \\ &\overset{(2)}{=} -\frac{1}{4\pi^2}\int_M \mathrm{tr}(F\wedge a) \\ &\overset{(3)}{=} -\frac{1}{4\pi^2}\int_M \mathrm{tr}\big(a\wedge\star(\star F)\big) \\ &\overset{(4)}{=} \left\langle a, \frac{\star F}{4\pi^2}\right\rangle \;=\; \left\langle \frac{\star F}{4\pi^2}, a\right\rangle. \end{aligned} \tag{12.38}$$

(1) Equation (12.26), written with Nash's notation $F_a(A)$ of footnote 3.
(2) Definition of $F_a(A)$.
(3) $\mathrm{tr}(F\wedge a) = \mathrm{tr}(a\wedge F)$ (2-form past 1-form), and $F = \star\star F$ because $\star\star = 1$ on 2-forms on a Riemannian 3-manifold. Here $\star F$ is an $\mathfrak{su}(2)$-valued 1-form.
(4) Definition (12.37) of the inner product, with $b = \star F/4\pi^2$; it is symmetric.

This is Nash's (12.18), and it gives (12.19):

$$\boxed{\mathrm{grad}\, f(A) = \frac{\star F(A)}{4\pi^2}}. \tag{12.39}$$

The gradient flow is $\mathrm{d}A/\mathrm{d}t = -\star F/4\pi^2$. Rescale time, $t \to 4\pi^2 t$, to absorb the constant; the flow equation (Nash's (12.20)) is

$$\boxed{\frac{\mathrm{d}A}{\mathrm{d}t} = -\star F(A)}. \tag{12.40}$$

In components on M with $A = A_i\,\mathrm{d}x^i$ and $B^i = (\star F)^i = \tfrac{1}{2}\varepsilon^{ijk}F_{jk}$ (the non-abelian magnetic field), (12.40) reads $\dot{A}_i = -B_i$: *the connection flows against its own magnetic field.*

How does f change along the flow? It decreases, at a rate set by the magnetic energy:

$$\frac{\mathrm{d}}{\mathrm{d}t}f(A(t)) \overset{(1)}{=} \mathrm{d}f_{\dot{A}} \overset{(2)}{=} \left\langle \frac{\star F}{4\pi^2}, \dot{A}\right\rangle \overset{(3)}{=} -\frac{1}{4\pi^2}\,\|\star F\|^2 \;\leq 0. \tag{12.41}$$

(1) Chain rule: the derivative of f along a curve is $\mathrm{d}f$ applied to the velocity.
(2) Equation (12.38) with $a = \dot{A}$.
(3) The flow (12.40): $\dot{A} = -\star F$.

So f is strictly decreasing except at flat connections, where the flow stops.

Worked example 12.16: The abelian flow on the 3-torus is not an initial-value problem

Take one abelian component on T^3, so the flow is linear: $\dot{\vec{a}} = -\nabla\times\vec{a}$. Expand in the curl eigenmodes found in Example 12.12: a mode of helicity $\pm$ and wavenumber $\kappa = |\vec{k}|$ has

$\nabla \times \vec{a} = \pm\kappa\, \vec{a}$, so

$$\dot{\vec{a}}_{\vec{k},\pm} = \mp\kappa\, \vec{a}_{\vec{k},\pm} \overset{(1)}{=} \vec{a}_{\vec{k},\pm}(t) = e^{\mp\kappa t}\, \vec{a}_{\vec{k},\pm}(0).$$

(1) Solve the scalar linear ODE for each mode separately.

Positive-helicity modes decay forward in time; negative-helicity modes *grow* like $e^{\kappa t}$, faster and faster as $\kappa \to \infty$. So for generic initial data the forward flow does not exist even for a short time, and neither does the backward flow. Contrast the heat equation $\dot{u} = \Delta u$, whose modes all decay forward. The Chern–Simons flow is not a parabolic evolution; it behaves like the Cauchy–Riemann equations $\partial_t u + i\partial_x u = 0$, where one cannot choose initial data freely. The right way to pose it is as a *boundary value problem* in t, with conditions at $t = \pm\infty$. The only finite-energy solutions linking two flat connections are those in which each growing mode is present only for $t < 0$ and each decaying mode only for $t > 0$, glued smoothly. This is exactly the setting of elliptic equations in four dimensions, and it leads to instantons.

12.4.2 The flow equation is the anti-self-duality equation on $M \times \mathbb{R}$

Now think of the path $A(t)$ as a single object on the 4-manifold $M \times \mathbb{R}$, with coordinates (x^1, x^2, x^3, t) and $x^4 \equiv t$. Define a connection $\mathbf{A}$ on $M \times \mathbb{R}$ by

$$\mathbf{A} = A(t), \qquad \text{i.e.} \qquad \mathbf{A}_i(x,t) = A_i(x,t), \quad \mathbf{A}_4 = 0. \tag{12.42}$$

Physicists call $\mathbf{A}_4 = 0$ **temporal gauge**: covariant differentiation in the t-direction is just $\partial/\partial t$. (Any connection on $M \times \mathbb{R}$ can be put in this gauge by solving an ODE in t for the gauge transformation.) Give $M \times \mathbb{R}$ the product metric $g_M + \mathrm{d}t^2$ and the orientation $\mathrm{d}t \wedge \mathrm{vol}_M$, i.e. the ordered coframe $(e^4, e^1, e^2, e^3) = (\mathrm{d}t, e^1, e^2, e^3)$.

Step 1: the four-dimensional curvature

Write $\mathbf{d} = \mathrm{d}_M + \mathrm{d}t \wedge \partial_t$ for the exterior derivative on $M \times \mathbb{R}$, with d_M the derivative along M. Then

$$\begin{aligned} \mathbf{F} &\overset{(1)}{=} \mathbf{dA} + \mathbf{A} \wedge \mathbf{A} \\ &\overset{(2)}{=} \mathrm{d}_M A + \mathrm{d}t \wedge \partial_t A + A \wedge A \\ &\overset{(3)}{=} F(t) + \mathrm{d}t \wedge \dot{A}(t), \end{aligned} \tag{12.43}$$

(1) Definition of curvature on $M \times \mathbb{R}$.
(2) $\mathbf{A} = A(t)$ has no $\mathrm{d}t$ component, so the only extra term in $\mathbf{dA}$ is the t-derivative.
(3) $F(t) = \mathrm{d}_M A + A \wedge A$ is the curvature of $A(t)$ on the slice $M \times \{t\}$.

In components, $\mathbf{F}_{ij} = F_{ij}$ and $\mathbf{F}_{4i} = \partial_t A_i$. This is Nash's remark that "the curvature components $\mathbf{F}_{4i}$ reduce to $\mathbf{F}_{4i} \equiv \mathrm{d}A/\mathrm{d}t$". Physically $\mathbf{F}_{4i}$ is the (Euclidean) electric field E_i, so (12.43) is the familiar $\mathbf{F} = \vec{E} \wedge \mathrm{d}t + \vec{B}$ split, with $\vec{E} = \dot{\vec{A}}$ in temporal gauge.

Step 2: the Hodge star on $M \times \mathbb{R}$

For a 1-form α on M (independent of how it varies in t), the four-dimensional star of the two kinds of 2-forms is

$$\star_4(\mathrm{d}t \wedge \alpha) = \star_3 \alpha, \qquad \star_4(\star_3 \alpha) = \mathrm{d}t \wedge \alpha. \tag{12.44}$$

Check on the coframe, with $\alpha = e^1$: $\star_3 e^1 = e^2 \wedge e^3$. The defining property $\omega \wedge \star\omega = |\omega|^2 \,\mathrm{vol}_4$ with $\mathrm{vol}_4 = e^4 \wedge e^1 \wedge e^2 \wedge e^3$ gives

$$(e^4 \wedge e^1) \wedge \star_4(e^4 \wedge e^1) = \mathrm{vol}_4 \overset{(1)}{=} \star_4(e^4 \wedge e^1) = e^2 \wedge e^3 = \star_3 e^1,$$
$$(e^2 \wedge e^3) \wedge \star_4(e^2 \wedge e^3) = \mathrm{vol}_4 \overset{(2)}{=} \star_4(e^2 \wedge e^3) = e^4 \wedge e^1.$$

(1) $e^4 \wedge e^1 \wedge e^2 \wedge e^3 = \mathrm{vol}_4$ exactly.
(2) $e^2 \wedge e^3 \wedge e^4 \wedge e^1 = e^4 \wedge e^1 \wedge e^2 \wedge e^3$, since moving the 2-form $e^4 \wedge e^1$ past the 2-form $e^2 \wedge e^3$ costs $(-1)^{2\cdot 2} = +1$.

The other components follow by cyclic permutation of $1, 2, 3$.

Step 3: compare

Apply (12.44) to (12.43), with $F = \star_3(\star_3 F)$:

$$\star_4 \mathbf{F} = \mathrm{d}t \wedge \star_3 F + \star_3 \dot{A}.$$

So the anti-self-duality equation $\mathbf{F} = -\star_4 \mathbf{F}$ reads

$$\begin{aligned} & F + \mathrm{d}t \wedge \dot{A} \overset{(1)}{=} -\mathrm{d}t \wedge \star_3 F - \star_3 \dot{A} \\ \iff \quad & \dot{A} \overset{(2)}{=} -\star_3 F \quad \text{and} \quad F = -\star_3 \dot{A} \\ \iff \quad & \dot{A} \overset{(3)}{=} -\star_3 F. \end{aligned} \tag{12.45}$$

(1) Substitute (12.43) and the formula for $\star_4 \mathbf{F}$.
(2) Match the parts containing $\mathrm{d}t$ and the parts tangent to M separately; a 2-form on $M \times \mathbb{R}$ splits uniquely this way.
(3) The second equation is the first with $\star_3$ applied to both sides ($\star_3 \star_3 = 1$): the two halves are the same equation.

Key idea: flow lines are instantons

With the product metric and the orientation $\mathrm{d}t \wedge \mathrm{vol}_M$, the Chern–Simons gradient flow $\dot{A} = -\star F$ on M is *exactly* the anti-self-duality equation

$$\mathbf{F} = -\star \mathbf{F} \quad \text{on } M \times \mathbb{R} \tag{12.46}$$

for the connection $\mathbf{A} = A(t)$ in temporal gauge. With the opposite orientation (or the flow run backwards) one gets the self-duality equation; this is Nash's (12.21), $\mathbf{F} = \mp \star \mathbf{F}$.

12.4.3 Boundary conditions and the energy identity

$M \times \mathbb{R}$ is not compact, so (12.46) needs boundary conditions at the two ends. The natural ones (Nash's (12.22)) are that the flow starts and ends at critical points:

$$A(-\infty) = A_P, \qquad A(+\infty) = A_Q, \qquad A_P, A_Q \text{ flat}. \tag{12.47}$$

Nash remarks that this is consistent with the usual requirement of finite Yang–Mills action: finite action forces $\mathbf{F} \to 0$ at the ends, i.e. the slices become flat. Let us compute the action and see that it is not only finite but *determined* by the endpoints.

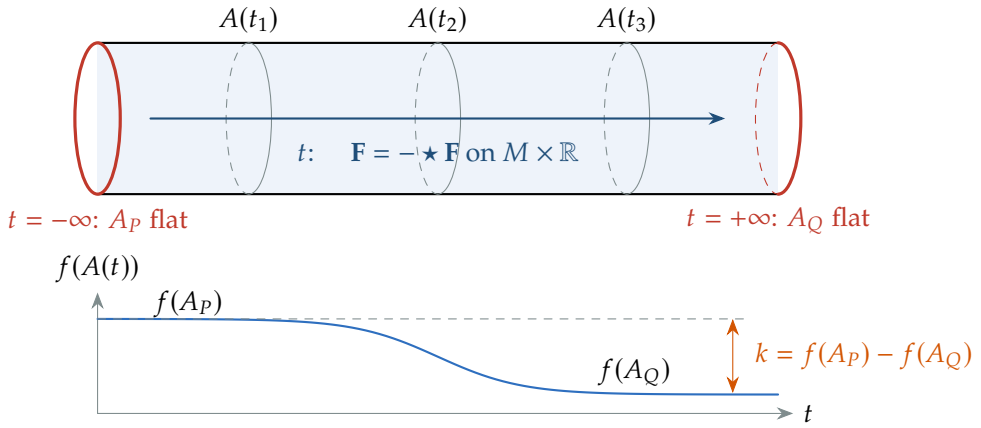


Figure 12.12: Top: a flow line of the Chern–Simons function is a 1-parameter family of connections $A(t)$ on the slices $M \times \{t\}$. Stacked together they form a connection on the infinite cylinder $M \times \mathbb{R}$, and the flow equation is the anti-self-duality equation there. The ends are flat connections (red). Bottom: along the flow, f decreases monotonically; the total drop is the instanton number k, a sum of the energy on all the slices.

$$\begin{aligned} S_{\mathrm{YM}} &:= -\int_{M\times\mathbb{R}} \operatorname{tr}(\mathbf{F}\wedge\star\mathbf{F}) \overset{(1)}{=} \int_{M\times\mathbb{R}} \operatorname{tr}(\mathbf{F}\wedge\mathbf{F}) \\ &\overset{(2)}{=} \int_{M\times\mathbb{R}} \mathbf{d}\,\mathrm{CS}(\mathbf{A}) \\ &\overset{(3)}{=} \int_M \mathrm{CS}(A_Q) - \int_M \mathrm{CS}(A_P) \\ &\overset{(4)}{=} 8\pi^2\big(f(A_P) - f(A_Q)\big). \end{aligned} \tag{12.48}$$

(1) Anti-self-duality: $\star\mathbf{F} = -\mathbf{F}$.
(2) The transgression formula (12.16), $\operatorname{tr}(F\wedge F) = \mathrm{d}\,\mathrm{CS}(A)$, which holds in any dimension.
(3) Stokes on $M \times [-T, T]$, $T \to \infty$. With the orientation $\mathrm{d}t \wedge \mathrm{vol}_M$ the outward normal at $t = +T$ is $+\partial_t$ and the boundary orientation is $+\mathrm{vol}_M$; at $t = -T$ it is $-\mathrm{vol}_M$. The 3-form $\mathrm{CS}(\mathbf{A})$ restricted to a slice is $\mathrm{CS}(A(t))$ since $\mathbf{A}$ has no $\mathrm{d}t$ component.
(4) Definition (12.15): $\int_M \mathrm{CS} = -8\pi^2 f$.

With the instanton number normalised as $k = \frac{1}{8\pi^2}\int \operatorname{tr}(\mathbf{F}\wedge\mathbf{F})$,

$$\boxed{\,k = f(A_P) - f(A_Q) \geq 0, \qquad S_{\mathrm{YM}} = 8\pi^2 k\,}. \tag{12.49}$$

The inequality is (12.41): f only goes down along a flow. The action of a flow line is its height drop.

Physical picture: Instantons as tunnelling, again

In Chapter 7 instantons were steepest-descent paths between critical points, and the Euclidean action of an instanton equalled the height difference $f(P) - f(Q)$: that was the Bogomolny trick $\frac{1}{2}\int |\dot{x} \pm \nabla f|^2 \geq \mp\int \dot{x}\cdot\nabla f$. Equation (12.48) is literally the same computation one dimension up. The famous Yang–Mills bound $S_{\mathrm{YM}} \geq 8\pi^2|k|$ is the Bogomolny bound of the supersymmetric quantum mechanics on $\mathcal{A}/\mathcal{G}$ whose "potential" is the Chern–Simons function.

Worked example 12.17: The BPST instanton is a flow line on $S^3 \times \mathbb{R}$

Remove the origin from $\mathbb{R}^4$ and use polar coordinates $r > 0$ and a point of the unit S^3. Put $\tau = \ln r$. Then

$$\mathrm{d}r^2 + r^2\, g_{S^3} \overset{(1)}{=} e^{2\tau}\mathrm{d}\tau^2 + e^{2\tau} g_{S^3} = e^{2\tau}\big(\mathrm{d}\tau^2 + g_{S^3}\big).$$

(1) $r = e^\tau$, so $\mathrm{d}r = e^\tau \mathrm{d}\tau$.

So $\mathbb{R}^4 \setminus \{0\}$ is *conformal* to the cylinder $S^3 \times \mathbb{R}$ with the product metric. The Hodge star on 2-forms in four dimensions is conformally invariant (rescaling $g \to \Omega^2 g$ multiplies $\omega \wedge \star\omega$ and vol by the same Ω^0 factor on 2-forms), so an anti-self-dual connection on $\mathbb{R}^4$ is an anti-self-dual connection on $S^3 \times \mathbb{R}$, i.e. a Chern–Simons flow line on S^3 with time $\tau = \ln r$. The orientation $\mathrm{d}\tau \wedge \mathrm{vol}_{S^3}$ is the standard one of $\mathbb{R}^4$ (radial direction first, then the boundary orientation of the sphere).

Take the BPST instanton of instanton number $k = 1$ centred at the origin (Chapter 8). In regular gauge it is smooth at $r = 0$, where $A \to 0$; as $r \to \infty$ it approaches a pure gauge $h^{-1}\mathrm{d}h$ with $h : S^3 \to \mathrm{SU}(2)$ of degree ± 1. So as a flow line on $S^3 \times \mathbb{R}$:

- $\tau \to -\infty$: $A \to A_P = 0$, with $f(A_P) = 0$;
- $\tau \to +\infty$: $A \to A_Q = h^{-1}\mathrm{d}h$, and (12.49) forces $f(A_Q) = f(A_P) - k = -1$.

Both ends are the *same* point of $\mathcal{A}/\mathcal{G}$, the trivial connection θ. The flow line descends the spiral staircase of Figure 12.7 by one full turn: it is a closed loop in $\mathcal{A}/\mathcal{G}$ representing the generator of $\pi_1(\mathcal{A}/\mathcal{G})$. Its scale parameter ρ is the time-translation of the flow line ($\rho \to e^c\rho$ is $\tau \to \tau + c$), and its centre gives four more parameters, so these flow lines form the 5-dimensional moduli space of $k = 1$ instantons on $\mathbb{R}^4$. The number $5 = 8 - 3$ is the index $8k$ computed below, minus $3 = \dim \mathrm{SU}(2)$ because both ends sit at the reducible connection θ whose stabiliser is all of SU(2).

12.4.4 Linearising the flow

To count flow lines we need their deformation theory: how many independent directions can one move a flow line and stay a flow line? Perturb a solution $A(t)$ to $A(t) + \varepsilon a(t)$ and keep first order in ε (Nash's (12.23)):

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\big(A + \varepsilon a\big) &\overset{(1)}{=} -\star F(A + \varepsilon a) \\ \frac{\mathrm{d}A}{\mathrm{d}t} + \varepsilon\frac{\mathrm{d}a}{\mathrm{d}t} &\overset{(2)}{=} -\star F(A) - \varepsilon\ \star \mathrm{d}_{A(t)}a + O(\varepsilon^2) \\ \frac{\mathrm{d}a}{\mathrm{d}t} &\overset{(3)}{=} -\star\, \mathrm{d}_{A(t)}a. \end{aligned} \tag{12.50}$$

(1) The perturbed path must satisfy the flow equation.
(2) $F(A + \varepsilon a) = F(A) + \varepsilon\, \mathrm{d}_A a + \varepsilon^2 a \wedge a$: the linear term is $\mathrm{d}a + A \wedge a + a \wedge A = \mathrm{d}_A a$ by (12.29).
(3) Subtract the unperturbed equation and divide by ε.

So a tangent vector to the space of flow lines is a solution of (Nash's (12.24))

$$\Big(\frac{\mathrm{d}}{\mathrm{d}t} + \star\mathrm{d}_{A(t)}\Big)a(t) = 0. \tag{12.51}$$

This is the same operator whose index is the spectral flow (Example 12.15): $\frac{\mathrm{d}}{\mathrm{d}t} + 4\pi^2 H(A(t))$. So

$$\dim\{\text{flow lines near } A(t)\} = \mathrm{index}\Big(\frac{\mathrm{d}}{\mathrm{d}t} + \star\mathrm{d}_{A(t)}\Big) = \sigma(A_P, A_Q), \tag{12.52}$$

exactly as in finite dimensions, where the space of flow lines from P to Q has dimension $\lambda_P - \lambda_Q$ (Figure 12.13). The relative Morse index is the dimension of the space of instantons connecting the two flat connections (when the relevant cokernel vanishes, which holds for generic data).

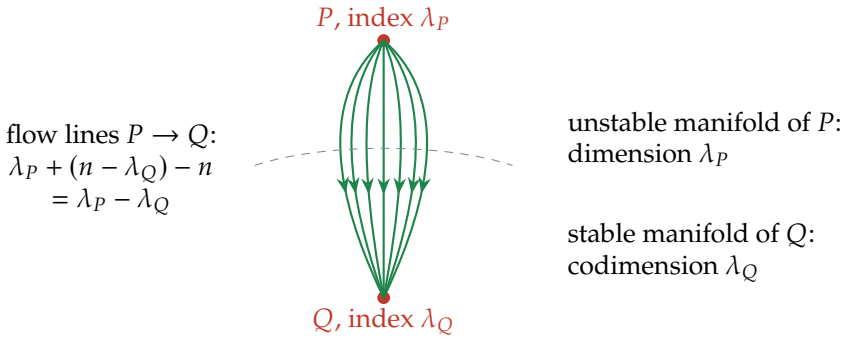


Figure 12.13: Finite-dimensional picture of (12.52). Flow lines from P leave along the λ_P-dimensional unstable manifold of P and must arrive in the stable manifold of Q, which has codimension λ_Q. Generic intersection has dimension $\lambda_P - \lambda_Q$. In Floer theory both numbers are infinite, but their difference is the index of $\frac{\mathrm{d}}{\mathrm{d}t} + \star \mathrm{d}_{A(t)}$.

The operator is the instanton deformation operator

Nash says that "after taking proper account of gauge equivalent a's", (12.51) is the deformation problem for instantons. Here is the proper account. A general perturbation of the 4d connection is $\mathbf{a} = a_0\,\mathrm{d}t + a$, with $a_0(x,t) \in \Omega^0(M) \otimes \mathfrak{su}(2)$ and $a(t) \in \Omega^1(M) \otimes \mathfrak{su}(2)$. The instanton deformation complex of Chapter 8 (Nash's (8.104), with self-dual and anti-self-dual exchanged because we linearise $\mathbf{F}^+ = 0$) asks for

$$\mathbf{d}_{\mathbf{A}}^+ \mathbf{a} = 0 \quad \text{(linearised ASD)}, \qquad \mathbf{d}_{\mathbf{A}}^* \mathbf{a} = 0 \quad \text{(gauge fixing)}.$$

Using $\mathbf{d}_{\mathbf{A}}\mathbf{a} = \mathrm{d}_A a + \mathrm{d}t \wedge (\dot{a} - \mathrm{d}_A a_0)$ and (12.44), the self-dual part of a 2-form $\beta + \mathrm{d}t \wedge \gamma$ is determined by $\gamma + \star_3 \beta$. So the two equations become

$$\begin{aligned} \dot{a} - \mathrm{d}_A a_0 + \star \mathrm{d}_A a &= 0, \\ -\dot{a}_0 + \mathrm{d}_A^* a &= 0, \end{aligned}$$

i.e.

$$\left(\frac{\mathrm{d}}{\mathrm{d}t} + L_{A(t)}\right)\begin{pmatrix} a_0 \\ a \end{pmatrix} = 0, \qquad L_A = \begin{pmatrix} 0 & -\mathrm{d}_A^* \\ -\mathrm{d}_A & \star \mathrm{d}_A \end{pmatrix} \text{ on } (\Omega^0 \oplus \Omega^1)(M) \otimes \mathfrak{su}(2). \tag{12.53}$$

L_A is self-adjoint (the off-diagonal entries are adjoints of each other). It is the **extended Hessian**: the curl operator $\star \mathrm{d}_A$ of (12.31) bordered by the gauge directions $\mathrm{d}_A a_0$. At an irreducible, non-degenerate flat connection L_A is invertible, and on the gauge-fixed slice $\mathrm{d}_A^* a = 0$ it reduces to $\star \mathrm{d}_A$. This is the precise sense of Nash's statement that $\frac{\mathrm{d}}{\mathrm{d}t} + \star \mathrm{d}_{A(t)}$ is the instanton deformation operator.

Worked example 12.18: The extended Hessian at $A = 0$ on the 3-torus

Return to Example 12.12. On a Fourier mode $e^{i\vec{k}\cdot\vec{x}}$ the bordered operator acts on $(a_0, \vec{a})$, with $\mathrm{d}_A = \nabla$ on functions, $\mathrm{d}_A^* = -\nabla\cdot$ on 1-forms, and $\star\mathrm{d} = \nabla\times$:

$$L_{\vec{k}} = \begin{pmatrix} 0 & i\vec{k}^{\mathsf{T}} \\ -i\vec{k} & i\vec{k}\times \end{pmatrix} \quad \text{on} \quad \mathbb{C} \oplus \mathbb{C}^3.$$

On the transverse helicity modes it gives $\pm|\vec{k}|$ as before. The longitudinal mode $\vec{a} = \hat{\vec{k}}$ and the scalar a_0 now mix: on the pair (a_0, $\hat{\vec{k}}$-component) the matrix is $\left(\begin{smallmatrix} 0 & i|\vec{k}| \\ -i|\vec{k}| & 0 \end{smallmatrix}\right)$, with eigenvalues $\pm|\vec{k}|$. So the zero eigenvalues that came from the gauge directions have become $\pm|\vec{k}|$: gauge fixing removes the spurious kernel. What remains at $\vec{k} = 0$ is a 4-dimensional kernel (constant a_0 and constant $\vec{a}$), reflecting that on T^3 the flat connection 0 is reducible ($H^0 \neq 0$) and degenerate ($H^1 \neq 0$). On a homology sphere $H^1 = 0$, and away from the trivial connection $H^0 = 0$ too.

12.4.5 The spectral flow round a loop is $8k$

We can now compute the missing number. Close up the t-direction: a loop in $\mathcal{A}/\mathcal{G}$ is a path $A(t)$, $0 \le t \le 1$, with $A(1) = A(0)^g$. Glue the ends of $M \times [0,1]$ using g. The result is an SU(2) bundle P over $M \times S^1$, and $\mathbf{A}$ is a connection on it. The bundle is non-trivial when $\deg g \neq 0$: g is the clutching function, and its degree is the instanton number,

$$k = \frac{1}{8\pi^2}\int_{M\times S^1} \operatorname{tr}(\mathbf{F}\wedge\mathbf{F}) = f(A(0)) - f(A(1)) = f(A(0)) - f(A(0)^g) = -\deg g,$$

by the same Stokes computation as (12.48) and the gauge shift (12.17). (The sign just reflects our orientation choice; what matters is $|k| = |\deg g|$.)

The spectral flow round the loop is the index of $\frac{\mathrm{d}}{\mathrm{d}t} + \star\mathrm{d}_{A(t)}$ on the closed 4-manifold $M \times S^1$, which is the dimension of the moduli space of instantons there. From Chapter 8 (Nash's (8.115) and (8.124)):

$$\operatorname{index}\Big(\frac{\mathrm{d}}{\mathrm{d}t} + \star\mathrm{d}_{A(t)}\Big) = \dim \mathcal{M}_k = p_1(\mathrm{ad}_{\mathbb{C}}P) - \frac{\dim G}{2}\big(\chi(M\times S^1) - \tau(M\times S^1)\big). \tag{12.54}$$

This is Nash's (12.25). We need three numbers.

(i) The Pontryagin number. For $G = \mathrm{SU}(2)$, in Nash's normalisation (Chapter 8, (8.127)–(8.131)),

$$p_1(\mathrm{ad}_{\mathbb{C}}P) = 8k, \qquad \dim G = 3. \tag{12.55}$$

(ii) The signature. The signature τ of a 4-manifold is the signature of the intersection form on H_2. Use the Künneth formula:

$$\begin{aligned} H_2(M\times S^1) &\overset{(1)}{=} \big(H_2(M)\otimes H_0(S^1)\big) \oplus \big(H_1(M)\otimes H_1(S^1)\big) \\ &\overset{(2)}{=} 0 \oplus 0 = 0 \quad\Longrightarrow\quad \tau(M\times S^1) = 0. \end{aligned} \tag{12.56}$$

(1) Künneth: $H_n(X\times Y) = \bigoplus_{p+q=n} H_p(X)\otimes H_q(Y)$ (plus torsion terms, which vanish here); $H_2(S^1) = 0$ kills the third term.
(2) M is a homology sphere: $H_1(M) = H_2(M) = 0$.

(Nash writes this as $H_2(M\times S^1) = H_2(M) = 0$, which is the same thing once one notes $H_1(M) = 0$; this is Nash's (12.26).) With no 2-cycles there is no intersection form at all, so its signature is 0.

(iii) The Euler characteristic. The Euler characteristic is multiplicative (Nash's (12.27)):

$$\chi(M\times S^1) \overset{(1)}{=} \chi(M)\,\chi(S^1) \overset{(2)}{=} (1-0+0-1)(1-1) = 0. \tag{12.57}$$

(1) Künneth again: the Poincaré polynomial of a product is the product of the Poincaré polynomials, and $\chi = P_{-1}$.

(2) Betti numbers of a homology 3-sphere are $1,0,0,1$; of the circle, $1,1$. (Either factor alone would do: every closed odd-dimensional manifold has $\chi = 0$ by Poincaré duality, $b_j = b_{n-j}$.)

Putting (i)–(iii) into (12.54):

$$\operatorname{index}\Big(\frac{\mathrm{d}}{\mathrm{d}t} + \star \mathrm{d}_{A(t)}\Big) \overset{(1)}{=} 8k - \frac{3}{2}(0-0) = 8k. \tag{12.58}$$

(1) Equations (12.55), (12.56) and (12.57).

This is Nash's (12.28). Back in three dimensions it says (Nash's (12.29)):

Key idea: the magic number 8

The spectral flow of $\star \mathrm{d}_{A(t)}$ round a loop in $\mathcal{A}/\mathcal{G}$ of degree k is $8k$. The generating loop ($k = 1$) has spectral flow 8. Hence the relative Morse index of the Chern–Simons function is well defined only modulo 8:

$$\lambda_{A_P, A_Q} \in \mathbb{Z}/8\mathbb{Z}.$$

Watch out: Normalisations of p_1 and the sign of k

Nash's (12.25) uses his normalisation $p_1(\mathrm{ad}_{\mathbb{C}}P) = 8k$ from Chapter 8; in the most common modern convention one writes $\dim \mathcal{M}_k = 8c_2(E) - 3(1 - b_1 + b_2^+)$ for an SU(2) bundle E with $c_2(E) = k$. On $M \times S^1$ with M a homology sphere, $b_1 = 1$ and $b_2^+ = 0$, so this also gives $8k - 3(1 - 1 + 0) = 8k$. The formulae agree; only the bookkeeping of p_1 differs (see the corresponding warning in Chapter 8). The sign of k is fixed by the orientation; with ours, instantons have $k \geq 0$ and flow lines go downhill in f.

Worked example 12.19: Betti numbers of $\Sigma(2,3,5) \times S^1$

Poincaré polynomials multiply:

$$P_t(\Sigma \times S^1) \overset{(1)}{=} (1 + t^3)(1 + t) = 1 + t + t^3 + t^4.$$

(1) $P_t(\Sigma) = 1 + t^3$ for any homology 3-sphere, $P_t(S^1) = 1 + t$.

So $b_0 = b_1 = b_3 = b_4 = 1$ and $b_2 = 0$. Then $\chi = 1 - 1 + 0 - 1 + 1 = 0$ and $\tau = 0$ (no H_2). Compare S^4 ($1 + t^4$, $\chi = 2$) where the same formula gives $\dim \mathcal{M}_k = 8k - \frac{3}{2} \cdot 2 = 8k - 3$: the -3 that appeared at the end of Example 12.17. On $M \times S^1$ there is no such correction; all of $8k$ is spectral flow.

Physical picture: Where 8 comes from

Why eight and not one? The index of the instanton deformation operator counts real parameters of instantons, and for SU(2) each unit of instanton number carries eight: on $\mathbb{R}^4$ they are the four coordinates of the instanton's centre, its size, and three SU(2) orientation angles (the global gauge rotation). The three orientation angles are "lost" on S^4 and $\mathbb{R}^4$ because the endpoint is reducible, leaving $8k - 3$; on $M \times S^1$ nothing is lost and all $8k$ are counted. So one trip round $\mathcal{A}/\mathcal{G}$ passes eight eigenvalues of the curl operator through zero.

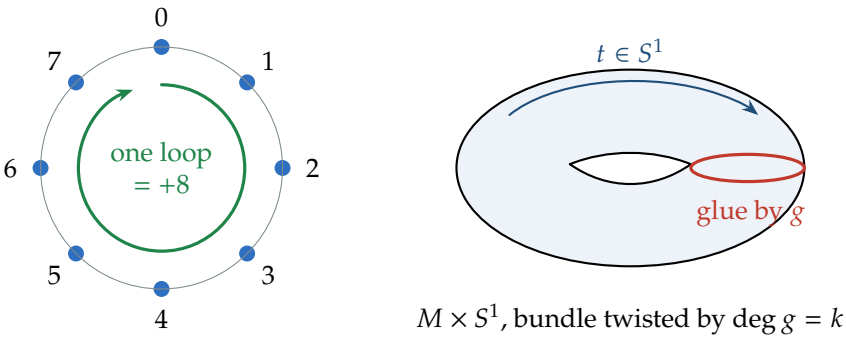


Figure 12.14: Left: because the spectral flow round the generating loop of $\pi_1(\mathcal{A}/\mathcal{G}) = \mathbb{Z}$ is 8, relative Morse indices live on a clock with eight hours. Right: the loop closes $M \times [0,1]$ into $M \times S^1$, glued by the gauge transformation g; the resulting bundle has instanton number $k = \pm \deg g$, and the index theorem on this closed 4-manifold gives $8k$.

Reading Nash (pp. 327–330, (12.15)–(12.29))

Nash introduces the metric (12.16) on $T_A\mathcal{A}$, computes grad $f = \star F/4\pi^2$ (12.17)–(12.19) and the flow $\dot{A} = -\star F$ (12.20). He extends $A(t)$ to $M \times \mathbb{R}$ with $\mathbf{A}_4 = 0$ and product metric and states that the flow is the (anti-)self-duality equation (12.21), with boundary conditions (12.22). He linearises (12.23)–(12.24) and computes the index via Chapter 8's (8.115), (8.124) (12.25), using $p_1 = 8k$, $\tau = 0$ (12.26), $\chi = 0$ (12.27), to get $8k$ (12.28), the spectral flow round a loop (12.29). We added: positivity of the metric; the monotonicity of f; the ill-posedness of the flow as an initial value problem; the full Hodge-star verification of anti-self-duality with our orientation; the energy identity $S_{\mathrm{YM}} = 8\pi^2(f(A_P) - f(A_Q))$; the BPST instanton as a flow line; the gauge-fixed extended Hessian; and the Künneth computations.

Recap

- With the L^2 metric, grad $f = \star F/4\pi^2$; the flow (rescaled) is $\dot{A} = -\star F$, and f decreases along it.
- In temporal gauge on $M \times \mathbb{R}$, $\mathbf{F} = F + \mathrm{d}t \wedge \dot{A}$, and the flow equation is exactly $\mathbf{F} = -\star\mathbf{F}$: flow lines are instantons with flat limits A_P, A_Q.
- Energy identity: $S_{\mathrm{YM}} = 8\pi^2 k$, $k = f(A_P) - f(A_Q)$.
- The linearised flow is $\left(\frac{\mathrm{d}}{\mathrm{d}t} + \star \mathrm{d}_A\right) a = 0$, the gauge-fixed instanton deformation operator; its index is the dimension of the space of flow lines and equals the relative index.
- On $M \times S^1$: $\tau = \chi = 0$, index $= 8k$. The relative Morse index is defined mod 8.

12.5 Floer homology, Casson's invariant and the Chern–Simons theory

We now have every ingredient of a Morse complex on $\mathcal{A}/\mathcal{G}$: critical points (irreducible flat connections), relative indices (spectral flow, defined mod 8), and flow lines (instantons on $M \times \mathbb{R}$). This section assembles them. We build the complex and its homology, the Floer groups $HF_p(M)$; we see why they are invariants of the 3-manifold; we compute them in the simplest non-trivial case; and we take their Euler characteristic, which turns out to be twice a classical invariant of Casson. Then we look at the Chern–Simons function from two more angles that

Nash raises at the end of his §2: as the *action* of a quantum field theory in three dimensions, and as the potential in a supersymmetric Hamiltonian whose ground states are the Floer groups.

12.5.1 The Floer complex

The chain groups

Let M be a homology 3-sphere, and suppose (after a small perturbation if necessary) that every irreducible flat connection is non-degenerate, i.e. the curl operator $\star \mathrm{d}_\alpha$ has no kernel on the gauge slice. Write

$$\mathcal{R}^*(M) = \{\alpha_1, \dots, \alpha_N\} \subset \mathrm{Hom}(\pi_1(M), \mathrm{SU}(2))/\mathrm{Ad}\,\mathrm{SU}(2)$$

for the irreducible ones. The trivial connection θ is the only reducible one (Section 12.3); it is *left out* of the complex, because its stabiliser SU(2) makes the flow lines into and out of it behave differently. It is still useful as a reference point for the grading.

Each α_i gets a degree in $\mathbb{Z}_8$: its relative Morse index with respect to θ,

$$\mu(\alpha) := \sigma(\alpha, \theta) \bmod 8, \tag{12.59}$$

the spectral flow of $\star \mathrm{d}_{A(t)}$ along any path from α to θ (Section 12.3.4). Different paths change it by multiples of 8 (Section 12.4.5), so $\mu(\alpha) \in \mathbb{Z}_8$ is well defined. (Conventions differ by a constant shift, usually $3 = \dim \mathrm{SU}(2)$, coming from the stabiliser of θ. Only differences matter for the complex.) The **Floer chain groups** are the free abelian groups

$$C_p := \bigoplus_{\alpha \in \mathcal{R}^*(M),\ \mu(\alpha) = p} \mathbb{Z} \cdot \alpha, \qquad p \in \mathbb{Z}_8. \tag{12.60}$$

This is (12.6) with "critical point of index p" replaced by "irreducible flat connection of relative index p mod 8".

The boundary operator

For $\alpha, \beta \in \mathcal{R}^*(M)$ let $\mathcal{M}(\alpha, \beta)$ be the space of flow lines from α to β modulo gauge, i.e. instantons on $M \times \mathbb{R}$ with limits α at $t = -\infty$ and β at $t = +\infty$ (Section 12.4). By (12.52) its dimension is the relative index $\sigma(\alpha, \beta)$. Time translation $t \mapsto t + c$ acts on it freely (a non-constant flow line is not translation invariant), so

$$\dim\big(\mathcal{M}(\alpha, \beta)/\mathbb{R}\big) = \sigma(\alpha, \beta) - 1.$$

When $\mu(\alpha) - \mu(\beta) = 1$, the quotient is a finite set of points: a finite number of *unparametrised* instantons. Each carries a sign ± 1 (the orientation of the moduli space at that point; physically, the sign of a fermion determinant). Their signed count is

$$n(\alpha, \beta) := \#\big(\mathcal{M}(\alpha, \beta)/\mathbb{R}\big) \in \mathbb{Z}, \tag{12.61}$$

and the **Floer boundary operator** is

$$\partial \alpha := \sum_{\mu(\beta) = \mu(\alpha) - 1} n(\alpha, \beta)\, \beta. \tag{12.62}$$

It lowers the degree by one. Unlike (12.7), which counted flow lines *up* and raised the degree (a coboundary), this counts flow lines *down* and lowers it (a boundary). The two are transposes of each other; Nash, following Floer, uses the boundary version.

Because degrees live in $\mathbb{Z}_8$, the complex closes up on itself (Nash's (12.30)):

$$\cdots \xrightarrow{\partial} C_7 \xrightarrow{\partial} C_6 \xrightarrow{\partial} \cdots \xrightarrow{\partial} C_1 \xrightarrow{\partial} C_0 \xrightarrow{\partial} C_7 \xrightarrow{\partial} \cdots \tag{12.63}$$

Watch out: The complex is cyclic, not finite

Nash draws (12.30) as $0 \to C_7 \to C_6 \to \cdots \to C_0 \to 0$, with zeros at both ends. Read literally, that would mean no flow lines from degree 0 to degree 7. But $7 \equiv 0 - 1 \pmod 8$, so there is in general a map $\partial : C_0 \to C_7$, and the complex is a *cycle* of eight groups, as in (12.63). What Nash means by "length 8" is that there are only eight distinct groups, however many critical points and however infinite the index.

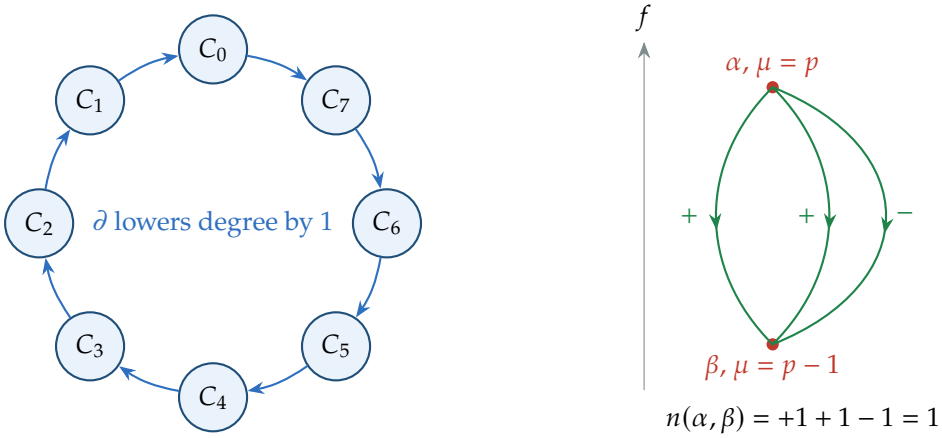


Figure 12.15: Left: the Floer complex has one group for each residue mod 8, and the boundary operator goes once round the clock. Right: the matrix element $n(\alpha, \beta)$ is the signed number of instantons on $M \times \mathbb{R}$ (flow lines modulo time translation) from α down to β, when their relative index is 1.

Why $\partial^2 = 0$: broken flow lines

Nash says that $\partial^2 = 0$ "can be done (formally) with exactly the same instanton tunnelling method" as in Chapter 7. Here is that method, as a Stokes-type argument. Take α and γ with $\mu(\alpha) - \mu(\gamma) = 2$. The coefficient of γ in $\partial^2\alpha$ is

$$\begin{aligned}(\partial^2\alpha)_\gamma &\overset{(1)}{=} \sum_\beta n(\alpha,\beta)\, n(\beta,\gamma) \\ &\overset{(2)}{=} \#\{\text{broken flow lines } \alpha \to \beta \to \gamma\} \\ &\overset{(3)}{=} \#\,\partial\big(\overline{\mathcal{M}(\alpha,\gamma)/\mathbb{R}}\big) \\ &\overset{(4)}{=} 0. \end{aligned} \tag{12.64}$$

(1) Apply (12.62) twice; β runs over generators of degree $\mu(\alpha) - 1$.

(2) A pair (flow line $\alpha \to \beta$, flow line $\beta \to \gamma$) is a **broken flow line**; its sign is the product of the two signs.

(3) $\mathcal{M}(\alpha,\gamma)/\mathbb{R}$ has dimension $2 - 1 = 1$. The *gluing theorem* says it can be compactified to a compact 1-manifold whose boundary points are exactly the broken flow lines, with the product orientation.

(4) A compact 1-manifold is a finite union of closed intervals and circles; each interval has two ends with opposite orientations, so the signed count of boundary points is zero.

Step (4) is the one-dimensional case of "$\int_M \mathrm{d}\alpha = \int_{\partial M} \alpha$": the fundamental theorem of calculus, $\int_a^b h' = h(b) - h(a)$, applied to the constant function $h = 1$ on each interval.

Watch out: Why no bubbles in the 1-dimensional moduli space

In four dimensions a sequence of instantons can concentrate curvature at a point and "bubble off" an instanton on S^4 (Section 12.6.5). Could this add extra ends in step (3)? A bubble carries away instanton number 1, hence (by Section 12.4.5) index 8. What is left is a flow line of index $2 - 8 = -6 < 0$, and for generic data there are none. So bubbling cannot happen in the moduli spaces that define ∂ and ∂^2. This is the same number 8 that made the grading cyclic.

Worked example 12.20: A dented sphere: tunnelling cancels a pair

Let $X = S^2$ be dented like a kidney bean standing on end, with height function f. It has two maxima a_1, a_2 (index 2), a saddle s between them (index 1) and one minimum m (index 0). The chain groups are $C_2 = \mathbb{Z}a_1 \oplus \mathbb{Z}a_2$, $C_1 = \mathbb{Z}s$, $C_0 = \mathbb{Z}m$.
From each maximum exactly one flow line runs down to the saddle, and from the saddle two flow lines run down to the minimum, one on each side. Orienting the unstable manifolds gives

$$\partial a_1 = s, \qquad \partial a_2 = -s, \qquad \partial s = m - m = 0, \qquad \partial m = 0.$$

Then

$$H_2 \overset{(1)}{=} \operatorname{Ker} \partial|_{C_2} = \mathbb{Z}(a_1 + a_2) \cong \mathbb{Z},$$
$$H_1 \overset{(2)}{=} \operatorname{Ker} \partial|_{C_1} / \operatorname{Im} \partial|_{C_2} = \mathbb{Z}s / \mathbb{Z}s = 0,$$
$$H_0 \overset{(3)}{=} C_0 / \operatorname{Im} \partial|_{C_1} = \mathbb{Z}m \cong \mathbb{Z}.$$

(1) There is no C_3, so H_2 is the kernel; $\partial(xa_1 + ya_2) = (x - y)s$ vanishes iff $x = y$.
(2) $\partial s = 0$, and $\operatorname{Im} \partial|_{C_2} = \mathbb{Z}s$.
(3) $\partial s = 0$, so nothing in C_0 is a boundary.

These are the Betti numbers of S^2. The pair (a_2, s) has been "cancelled by tunnelling". Now check $\partial^2 a_1 = 0$ geometrically: the flow lines from a_1 to m form a 1-parameter family (they sweep out the cap of a_1). At the two ends of the family the flow line breaks at s, once on each side, giving the two broken lines $a_1 \to s \to m$ with signs $+1$ and -1. They cancel, as (12.64) says they must.

12.5.2 The Floer groups

With $\partial^2 = 0$ we can take homology (Nash's (12.31)):

$$\boxed{HF_p(M) = \frac{\operatorname{Ker}\big(\partial : C_p \to C_{p-1}\big)}{\operatorname{Im}\big(\partial : C_{p+1} \to C_p\big)}, \qquad p \in \mathbb{Z}_8.} \tag{12.65}$$

(Nash labels the map $C_{p+1} \to C_p$ as ∂_p, so his formula reads $\operatorname{Ker} \partial_{p-1} / \operatorname{Im} \partial_p$; it is the same thing.) These are the **Floer homology groups** of M.

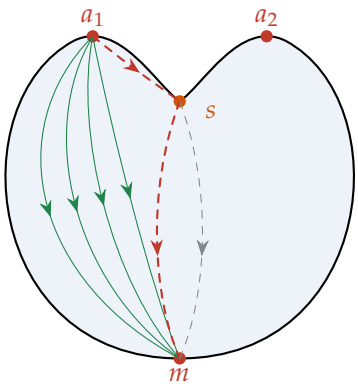


Figure 12.16: The dented sphere of Example 12.20, seen from the front. Flow lines from the maximum a_1 to the minimum m come in a 1-parameter family. As the parameter runs to its ends, the flow line is pulled onto the saddle s and breaks into two pieces. The two broken lines (front and back) have opposite signs, which is why $\partial^2 a_1 = 0$. The same picture, with instantons on $M \times \mathbb{R}$ in place of flow lines, proves $\partial^2 = 0$ in Floer's complex.

Working definition 12.8 (Floer homology of a homology 3-sphere)

To compute $HF_*(M)$:

1. List the irreducible flat SU(2) connections, i.e. the irreducible representations $\pi_1(M) \to \mathrm{SU}(2)$ up to conjugation. Perturb if any is degenerate [its Hessian $\star \mathrm{d}_A$ has a nonzero zero mode a that is not a gauge direction].
2. Give each one a degree $\mu \in \mathbb{Z}_8$ by the spectral flow of the curl operator $\star \mathrm{d}_{A(t)}$ from it to the trivial connection.
3. For each pair with degrees differing by one, count the instantons on $M \times \mathbb{R}$ [anti-self-dual connections, the same as steepest-descent paths $A(t)$ of the Chern–Simons function] running between them, modulo time translation, with signs. This is the matrix of ∂.
4. Take kernel modulo image in each of the eight degrees.

Formally $HF_*(M)$ is the Morse homology [chains are the critical points, and ∂ counts the flow lines between them] of the Chern–Simons function on the irreducible part of $\mathcal{A}/\mathcal{G}$ [gauge classes of irreducible connections].

Why it is an invariant

The construction used a metric on M (to define $\star$, hence the curl operator and the flow), the product metric on $M \times \mathbb{R}$ (to define instantons), and the L^2 metric (12.37) on $\mathcal{A}$ (to define the gradient). Nash stresses that $HF_*(M)$ does not depend on these choices. The mechanism is again the one we know from Chapter 7. Take two metrics g_0, g_1 on M and a path g_s between them. Put the metric $g_{s(t)}$ on the slice $M \times \{t\}$, with $s(t) = 0$ for $t \ll 0$ and $s(t) = 1$ for $t \gg 0$. Counting instantons on this cylinder from a generator of the g_0-complex to one of the same degree of the g_1-complex defines a map

$$\Phi : C_*(M; g_0) \to C_*(M; g_1), \qquad \Phi \partial_0 = \partial_1 \Phi,$$

the second equation by another broken-flow-line argument. Running the path backwards gives Ψ in the other direction, and a homotopy of homotopies shows that $\Psi\Phi$ and $\Phi\Psi$ induce the

identity on homology. So Φ induces an isomorphism $HF_*(M; g_0) \cong HF_*(M; g_1)$.

Physical picture: Invariance is the cylinder axiom

In the language of Section 12.1.3, the Floer groups are the state space $Z(M)$ attached to the 3-manifold M, and the instanton counts on the cylinder $M \times \mathbb{R}$ with varying metric are the operator $Z(M \times I)$. Atiyah's cylinder axiom $Z(M \times I) = \mathrm{id}$ is precisely the statement that changing the metric does not change the state space. Floer's invariance proof is that axiom, verified by analysis.

Key idea: Floer homology

For a homology 3-sphere M, the Floer groups $HF_p(M)$, $p \in \mathbb{Z}_8$, are topological invariants of M. They are built from the infinite-dimensional space $\mathcal{A}/\mathcal{G}$, yet there are only eight of them and each is finitely generated. They are the Morse homology of the Chern–Simons function, and physically the zero-energy states of a supersymmetric gauge theory on $M \times \mathbb{R}$ (Section 12.5.5).

Worked example 12.21: The 3-sphere and the Poincaré sphere

$M = S^3$. Here $\pi_1 = 1$, so the only flat connection is θ, which is excluded. Every chain group is zero:

$$HF_p(S^3) = 0 \quad \text{for all } p.$$

This vanishing is used in Section 12.8.4 to show that Donaldson invariants of connected sums vanish.

$M = \Sigma(2,3,5)$. By Example 12.11 there are exactly two irreducible flat connections, ρ_1 and ρ_2. Fintushel and Stern computed their degrees: with the standard orientation and grading convention, $\mu(\rho_1) = 1$ and $\mu(\rho_2) = 5$. Then

$$\partial\rho_1 \in C_0 \overset{(1)}{=} 0, \qquad \partial\rho_2 \in C_4 \overset{(1)}{=} 0,$$

(1) C_0 and C_4 are zero: no generator has degree 0 or 4.

So the boundary operator vanishes for degree reasons, without counting a single instanton, and

$$HF_*(\Sigma(2,3,5)) = \big(0,\ \mathbb{Z},\ 0,\ 0,\ 0,\ \mathbb{Z},\ 0,\ 0\big) \qquad (p = 0, 1, \dots, 7).$$

A non-zero Floer group proves at once that $\Sigma(2,3,5)$ is not S^3, which is no surprise since $\pi_1 \neq 1$. More interesting is the general pattern found by Fintushel and Stern: for every Brieskorn sphere $\Sigma(p,q,r)$ all irreducible flat connections have *odd* degree, so $\partial = 0$ always and HF_* is free on the flat connections.

Worked example 12.22: Degree reasons in general

The argument above is worth isolating. If every generator of a $\mathbb{Z}_8$-graded complex has degree of the same parity, then

$$\partial(C_{\mathrm{odd}}) \subset C_{\mathrm{even}} \overset{(1)}{=} 0 \quad \Longrightarrow \quad \partial = 0 \quad \Longrightarrow \quad HF_p = C_p.$$

(1) ∂ changes the degree by one, hence changes the parity, and the even chain groups are zero.

So $\operatorname{rank} HF_* = \#\mathcal{R}^*(M)$: every flat connection survives as a quantum ground state. Tunnelling can only lift pairs of vacua whose degrees differ by one; if all vacua are "fermionic",

none can be lifted. This is the Floer analogue of the statement that a Morse function with critical points only of even index (like the height on $\mathbb{CP}^n$) is *perfect*.

12.5.3 The Euler characteristic and Casson's invariant

The Euler characteristic of a $\mathbb{Z}_8$-graded complex

A $\mathbb{Z}$-graded complex has an Euler characteristic $\sum(-1)^p \operatorname{rank} H_p$. With only a $\mathbb{Z}_8$ grading, is $(-1)^p$ still meaningful? Yes, because 8 is even: p and $p+8$ have the same parity. So define

$$\chi_F(M) := \sum_{p=0}^{7} (-1)^p \operatorname{rank} HF_p(M). \tag{12.66}$$

As in finite dimensions it can be computed from the chains, with no knowledge of ∂:

$$\begin{aligned}\sum_p (-1)^p \operatorname{rank} HF_p &\overset{(1)}{=} \sum_p (-1)^p \left(\operatorname{rank} Z_p - \operatorname{rank} B_p\right) \\ &\overset{(2)}{=} \sum_p (-1)^p \left(\operatorname{rank} C_p - \operatorname{rank} B_{p-1} - \operatorname{rank} B_p\right) \\ &\overset{(3)}{=} \sum_p (-1)^p \operatorname{rank} C_p \;=\; \sum_{\alpha \in \mathcal{R}^*(M)} (-1)^{\mu(\alpha)}.\end{aligned} \tag{12.67}$$

(1) $Z_p = \operatorname{Ker}(\partial|_{C_p})$, $B_p = \operatorname{Im}(\partial|_{C_{p+1}}) \subset C_p$, and $HF_p = Z_p/B_p$.
(2) Rank–nullity for $\partial : C_p \to C_{p-1}$: $\operatorname{rank} C_p = \operatorname{rank} Z_p + \operatorname{rank} B_{p-1}$.
(3) In the sum over the cycle $p \in \mathbb{Z}_8$, the term B_{p-1} appears with sign $(-1)^p$ and again (as $B_{p'}$ with $p' = p-1$) with sign $-(-1)^{p'} = (-1)^p \ldots$ with opposite total sign; the B's cancel in pairs. This uses only that $(-1)^p$ is well defined mod 8.

So $\chi_F(M)$ is a signed count of irreducible flat connections, the sign of each being the parity of its spectral flow. For $\Sigma(2,3,5)$, $\chi_F = (-1)^1 + (-1)^5 = -2$.

The finite-dimensional model: Poincaré–Hopf

Nash explains (12.67) through a classical theorem. Let X be a compact manifold and ω a 1-form (equivalently, via a metric, a vector field v) with isolated non-degenerate zeros. At a zero x, write $\omega = \omega_i \, \mathrm{d}x^i$ and form the matrix $J_{ij} = \partial_j \omega_i(x)$. The **index** of the zero is $\operatorname{sgn} \det J$.

Key idea: Poincaré–Hopf

The sum of the indices of the zeros of any 1-form (or vector field) with isolated non-degenerate zeros on a compact manifold X is the Euler characteristic:

$$\sum_{\omega(x)=0} \operatorname{sgn} \det\big(\partial_j \omega_i(x)\big) = \chi(X).$$

For $\omega = \mathrm{d}f$, J is the Hessian, $\operatorname{sgn}\det J = (-1)^{\lambda_x}$, and the theorem is $\sum_x (-1)^{\lambda_x} = \chi(X)$: the Euler characteristic of the Morse complex.

Why it is true (sketch). The sum of indices does not change when ω is deformed: zeros can move, and can be created or annihilated only in pairs of opposite index (a fold $\omega = x^2 - \varepsilon$ has

two zeros with slopes $\pm 2\sqrt{\varepsilon}$ that merge and vanish as $\varepsilon \to 0^-$). So the sum is an invariant of X, and computing it for one Morse function identifies it with $\sum (-1)^p b_p$ by the Morse complex.

Worked example 12.23: Poincaré–Hopf on the sphere and the torus

Sphere. The height $f = z$ on the unit S^2 has a minimum at the south pole and a maximum at the north pole. Near either pole use (x, y) as coordinates, $f = \pm\sqrt{1 - x^2 - y^2} \approx \pm(1 - \tfrac{1}{2}(x^2 + y^2))$, so the Hessian is $\mp \mathbf{1}_2$ and

$$\sum (-1)^{\lambda_x} \overset{(1)}{=} \det(-\mathbf{1}_2)^{\mathrm{sgn}} + \det(+\mathbf{1}_2)^{\mathrm{sgn}} = (+1) + (+1) = 2 = \chi(S^2).$$

(1) $\operatorname{sgn}\det(-\mathbf{1}_2) = (-1)^2 = +1$ at the maximum (index 2) and $\operatorname{sgn}\det \mathbf{1}_2 = +1$ at the minimum (index 0).

Torus. On the square torus $T^2 = \mathbb{R}^2/(2\pi\mathbb{Z})^2$ take $f = \cos x + \cos y$. Then $\omega = \mathrm{d}f = -\sin x\,\mathrm{d}x - \sin y\,\mathrm{d}y$ vanishes at the four points with $x, y \in \{0, \pi\}$, and $J = \operatorname{diag}(-\cos x, -\cos y)$:

(x, y)	$(0,0)$	(π,π)	$(0,\pi)$	$(\pi,0)$
J	$\operatorname{diag}(-1,-1)$	$\operatorname{diag}(1,1)$	$\operatorname{diag}(-1,1)$	$\operatorname{diag}(1,-1)$
type	max	min	saddle	saddle
$\operatorname{sgn}\det J$	$+1$	$+1$	-1	-1

The sum is $1 + 1 - 1 - 1 = 0 = \chi(T^2)$. Figure 12.17 shows the vector field.

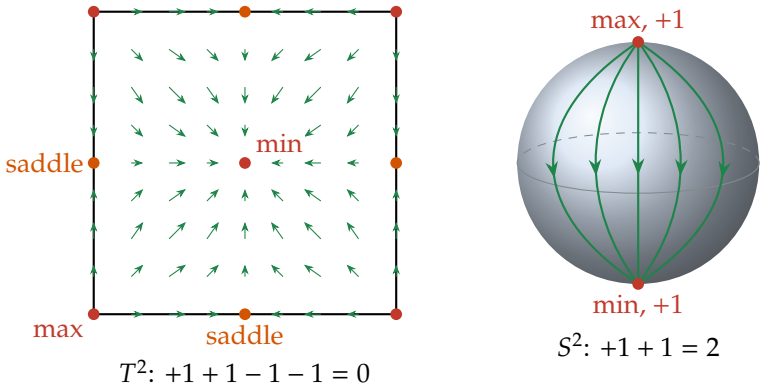


Figure 12.17: Poincaré–Hopf. Left: the downhill vector field $-\nabla f$ for $f = \cos x + \cos y$ on the square torus (opposite edges identified, so the four corners are one point). Its four zeros have indices $+1$ (maximum and minimum) and -1 (the two saddles), summing to $\chi(T^2) = 0$. Right: on the sphere the height function has only a maximum and a minimum, each of index $+1$: $\chi(S^2) = 2$. In Floer theory the vector field is $-\star F$ on $\mathcal{A}/\mathcal{G}$, its zeros are the flat connections, and the index of each is the parity of its spectral flow.

The infinite-dimensional version

Now carry this over (Nash's (12.33)). By (12.26) the curvature F is, up to the factor $-1/4\pi^2$, the 1-form $\mathrm{d}f$ on $\mathcal{A}$. It is gauge covariant, $F(A^g) = g^{-1}F(A)g$, and the pairing $\int \operatorname{tr}(F \wedge a)$ is gauge invariant, so it descends to a 1-form on $\mathcal{A}/\mathcal{G}$ (on the part where $\mathcal{G}$ acts freely). Its zeros are the flat connections modulo gauge,

$$\operatorname{Hom}\big(\pi_1(M), \mathrm{SU}(2)\big) \,/\, \operatorname{Ad}\mathrm{SU}(2),$$

and the "index" of each irreducible zero is sgn det of the Hessian, which in infinite dimensions has to be defined as $(-1)^{\text{spectral flow}}$, i.e. $(-1)^{\mu(\alpha)}$. The signed count of zeros is therefore exactly (12.67). Taubes made this precise: he perturbed F by a small Fredholm perturbation (if needed) to make all zeros non-degenerate, and used spectral flow to decide whether a zero's sign changes as the perturbation or the metric is varied. The result is Taubes' theorem (Nash's (12.32)):

$$\boxed{\chi_F(M) = 2\lambda(M)}, \tag{12.68}$$

where $\lambda(M)$ is the **Casson invariant**. In particular $\chi_F(M)$ is always even.

Casson's invariant

Casson defined $\lambda(M)$ in the mid-1980s by a finite-dimensional intersection count, long before the gauge theory reading was available. Nash calls the definition complicated; its idea is simple and pretty.

1. *Split M.* Every closed 3-manifold is a union of two handlebodies H_1, H_2 of some genus g (solid "pretzels") glued along their common boundary surface Σ_g: a **Heegaard splitting** $M = H_1 \cup_\Sigma H_2$.

2. *Represent.* Let $\mathcal{R}(X) = \mathrm{Hom}(\pi_1(X), \mathrm{SU}(2))/\text{conj}$. Van Kampen's theorem says $\pi_1(M)$ is generated by $\pi_1(H_1)$ and $\pi_1(H_2)$ with the relations coming from $\pi_1(\Sigma)$. So a representation of $\pi_1(M)$ is a pair of representations of the handlebodies that agree on Σ:

$$\mathcal{R}(M) = \mathcal{R}(H_1) \cap \mathcal{R}(H_2) \;\subset\; \mathcal{R}(\Sigma).$$

3. *Count dimensions.* $\pi_1(\Sigma_g)$ has $2g$ generators and one relation; $\pi_1(H_i)$ is free on g generators. For irreducible representations,

$$\dim \mathcal{R}^*(\Sigma_g) \overset{(1)}{=} 3 \cdot 2g - 3 - 3 = 6g - 6,$$
$$\dim \mathcal{R}^*(H_i) \overset{(2)}{=} 3g - 3.$$

   (1) Each generator gives an element of SU(2) (3 parameters); the relation is an SU(2)-valued equation (3 conditions); conjugation removes 3 more.
   (2) g free elements of SU(2), minus conjugation.

   So each $\mathcal{R}^*(H_i)$ is half-dimensional in $\mathcal{R}^*(\Sigma)$, and two half-dimensional submanifolds meet generically in isolated points. (In fact $\mathcal{R}^*(\Sigma)$ is symplectic, with form $\int_\Sigma \mathrm{tr}(a \wedge b)$, and the $\mathcal{R}^*(H_i)$ are Lagrangian.)

4. *Intersect.* Casson's invariant is

$$\lambda(M) = \tfrac{1}{2}\,\#\big(\mathcal{R}^*(H_1) \cap \mathcal{R}^*(H_2)\big), \tag{12.69}$$

   the signed intersection number, with orientations from a fixed orientation of M. The factor $\frac{1}{2}$ is part of the definition; Casson proved the result is an integer independent of the splitting.

The intersection points are exactly the irreducible flat connections on M, the same points that generate the Floer complex. So (12.68) is the statement that two ways of signing them (by intersection orientation, and by the parity of spectral flow) agree up to the overall factor.

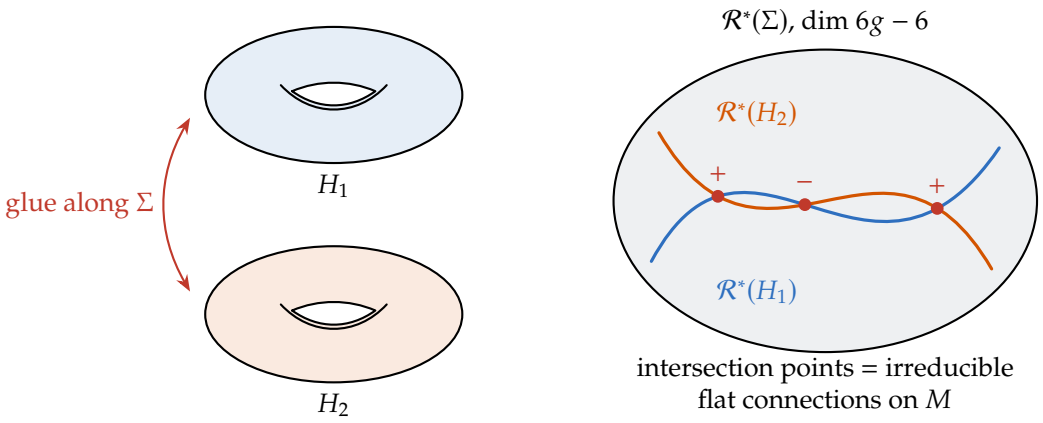


Figure 12.18: Casson's construction. Left: a Heegaard splitting of genus 1 (two solid tori glued along a torus; genus 1 gives only S^3, $S^1\times S^2$ and lens spaces, so real examples need higher genus). Right: in the space of irreducible representations of the surface group, the representations that extend over H_1 and those that extend over H_2 are two half-dimensional submanifolds. They meet at the irreducible representations of $\pi_1(M)$. Half the signed count is $\lambda(M)$; here it would be $\frac{1}{2}(+1-1+1)$, so the three points drawn are only a cartoon: a real example has an even signed count.

Worked example 12.24: Properties and values of Casson's invariant

- $\lambda(S^3)=0$: there are no irreducible representations of the trivial group.
- $\lambda(-M)=-\lambda(M)$: reversing the orientation of M reverses all intersection signs. The Floer version: reversing orientation replaces f by $-f$, the curl operator by its negative, and the degree μ by $-\mu$ up to a shift of 3, which flips every parity.
- For the Poincaré sphere, oriented as the link of the singularity $z_1^2+z_2^3+z_3^5=0$, $\lambda(\Sigma(2,3,5))=-1$, consistent with $\chi_F=-2$ in Example 12.21 and (12.68).
- $\lambda(M)$ mod 2 is the Rokhlin invariant of M (the signature of any spin 4-manifold bounded by M, divided by 8, mod 2). For $\Sigma(2,3,5)$ it bounds the E_8 plumbing, of signature ∓ 8, so the Rokhlin invariant is 1, matching $\lambda=-1$ mod 2.

Casson's *surgery formula* makes λ computable. If M_n is obtained from S^3 by $1/n$ surgery on a knot K (cut out a tubular neighbourhood of K and glue it back with a twist), then

$$\lambda(M_{n+1})-\lambda(M_n)=\tfrac{1}{2}\Delta_K''(1),$$

where Δ_K is the symmetrised Alexander polynomial of K (see the knot sections of this chapter). For the trefoil, $\Delta(t)=t-1+t^{-1}$:

$$\Delta''(t)\overset{(1)}{=}\frac{\mathrm{d}}{\mathrm{d}t}\bigl(1-t^{-2}\bigr)=2t^{-3},\qquad \tfrac{1}{2}\Delta''(1)\overset{(2)}{=}1.$$

(1) $\Delta'(t)=1-t^{-2}$, then differentiate again.
(2) Evaluate at $t=1$.

Since $M_0=S^3$ has $\lambda=0$, the manifolds $M_{\pm1}$ obtained by ±1 surgery on the trefoil have $\lambda=\pm1$. One of them is the Poincaré sphere (which one, and with which orientation, is a matter of conventions for the handedness of the trefoil).

12.5.4 The Chern–Simons function as an action

Nash ends his §2 by turning the Morse function into an action. This is the seed of the second half of the chapter (Witten's knot invariants), so we set it up carefully here.

A metric-free action

Define (Nash's (12.34))

$$S(A) = -\frac{1}{8\pi^2}\int_M \operatorname{tr}\Big(A\wedge \mathrm{d}A + \tfrac{2}{3}A\wedge A\wedge A\Big), \tag{12.70}$$

which is our $f(A)$ of (12.15), now for any closed oriented 3-manifold M and any compact gauge group. Look at what went into it: A, the wedge product, d, a trace, and the integral of a 3-form over an oriented 3-manifold. *No metric*, no Hodge star, no $\sqrt{g}$. So S is automatically invariant under all orientation-preserving diffeomorphisms of M, and the stress tensor $T_{\mu\nu} = \delta S/\delta g^{\mu\nu}$ is identically zero. This is the Schwarz type of TQFT of Section 12.1.2.

The classical equations of motion are the critical points of S, already found in (12.26) (Nash's (12.35)):

$$F(A) = 0.$$

Unlike in Yang–Mills theory, the classical solutions are not just the trivial connection: on a 3-manifold with $\pi_1(M) \neq 1$ there are flat connections with non-trivial holonomy (Example 12.11), and these are the classical vacua.

Quantisation of the level

S is gauge invariant only modulo integers (12.17): $S(A^g) = S(A) + \deg g$. Classically this is harmless, because the equations of motion only see δS. Quantum mechanically, the path integral in $2+1$ Lorentzian dimensions weights each configuration with $\exp(iS)$ (Nash's (12.36)), and $\exp(iS)$ changes by $e^{i\deg g}$ under a large gauge transformation. So $\exp(iS)$ is not a function on $\mathcal{A}/\mathcal{G}$. The cure (Nash's (12.37)) is to multiply the action by $2\pi k$ with k an integer:

$$\begin{aligned}\exp\big(2\pi i k\, S(A)\big) &\overset{(1)}{=} \exp\Big[2\pi i k\cdot\Big(-\frac{1}{8\pi^2}\Big)\int_M \operatorname{tr}\Big(A\wedge \mathrm{d}A + \tfrac{2}{3}A^3\Big)\Big]\\ &\overset{(2)}{=} \exp\Big[-\frac{ik}{4\pi}\int_M \operatorname{tr}\Big(A\wedge \mathrm{d}A + \tfrac{2}{3}A\wedge A\wedge A\Big)\Big],\end{aligned} \tag{12.71}$$

(1) Definition (12.70).
(2) $2\pi\cdot\frac{1}{8\pi^2} = \frac{1}{4\pi}$.

and under a gauge transformation

$$\exp\big(2\pi i k\, S(A^g)\big) \overset{(1)}{=} \exp\big(2\pi i k\, S(A)\big)\,\exp\big(2\pi i k\deg g\big) \overset{(2)}{=} \exp\big(2\pi i k\, S(A)\big).$$

(1) The shift formula (12.17).
(2) $k\deg g \in \mathbb{Z}$, and $e^{2\pi i\cdot\text{integer}} = 1$.

The integer k is the **level** of the Chern–Simons theory. With $k\notin\mathbb{Z}$ the weight is multivalued on $\mathcal{A}/\mathcal{G}$; with $k\in\mathbb{Z}$ the theory makes sense. This is the same logic as Dirac's quantisation condition: $e^{ie\oint A}$ must be single-valued, so $eg\in 2\pi\mathbb{Z}$.

Watch out: Lorentzian or Euclidean?

Nash puts the theory in $2+1$ Lorentzian dimensions to get e^{iS}. For a topological term the distinction is immaterial: S contains no metric, so nothing changes under Wick rotation, and the weight stays e^{iS} (a phase) even on a Euclidean 3-manifold. Topological terms such as Chern–Simons and the θ-term are "imaginary" in Euclidean signature. That is why their path integrals are oscillatory, not damped.

Worked example 12.25: Level quantisation seen on a single gauge transformation

Take $M = S^3$, $A = 0$ and g of degree 1 (Example 12.7). Then $A^g = g^{-1}\mathrm{d}g$, and

$$S(g^{-1}\mathrm{d}g) \overset{(1)}{=} S(0) + 1 = 1, \qquad e^{iS(g^{-1}\mathrm{d}g)} \overset{(2)}{=} e^{i} \neq e^{iS(0)} = 1,$$

(1) The shift formula with $\deg g = 1$ and $S(0) = 0$.
(2) The unnormalised weight differs on two gauge-equivalent configurations.

while $e^{2\pi i k S(g^{-1}\mathrm{d}g)} = e^{2\pi i k} = 1 = e^{2\pi i k S(0)}$ for $k \in \mathbb{Z}$. So the weight of the pure-gauge configuration $g^{-1}\mathrm{d}g$, which is physically the same as $A = 0$, agrees with it only at integer level.

Worked example 12.26: Abelian Chern–Simons on a torus has a k-dimensional Hilbert space

Here is the promise of Section 12.1, a finite-dimensional Hilbert space, made explicit. Take $G = \mathrm{U}(1)$, write $A = ia$ with a a real 1-form (the physicists' potential), so $\mathrm{tr}(A \wedge \mathrm{d}A) = -a \wedge \mathrm{d}a$ and the cubic term vanishes. The level-k action (12.71) is $\frac{k}{4\pi}\int a \wedge \mathrm{d}a$. Put it on $\Sigma \times \mathbb{R}$ with $\Sigma = T^2$, coordinates $x^1, x^2 \in [0, 2\pi)$ and time t, in the gauge $a_0 = 0$.
The equation of motion from varying a_0 (Gauss's law) is that a is flat on Σ. Flat connections on T^2, modulo small gauge transformations, are constant: $a = \frac{1}{2\pi}(\theta_1\,\mathrm{d}x^1 + \theta_2\,\mathrm{d}x^2)$, with $\theta_j = \oint_{\gamma_j} a$ the holonomy round the jth cycle. Large gauge transformations $e^{i(n_1x^1+n_2x^2)}$ shift $\theta_j \to \theta_j + 2\pi n_j$, so the θ_j are angles. Substituting into the action,

$$\begin{aligned}\frac{k}{4\pi}\int a \wedge \mathrm{d}a &\overset{(1)}{=} \frac{k}{4\pi}\int a \wedge \mathrm{d}t \wedge \dot{a} \\ &\overset{(2)}{=} -\frac{k}{4\pi}\int \mathrm{d}t \wedge \frac{(\theta_1\dot{\theta}_2 - \theta_2\dot{\theta}_1)}{4\pi^2}\,\mathrm{d}x^1 \wedge \mathrm{d}x^2 \\ &\overset{(3)}{=} -\frac{k}{4\pi}\int \mathrm{d}t\,(\theta_1\dot{\theta}_2 - \theta_2\dot{\theta}_1) \overset{(4)}{=} -\frac{k}{2\pi}\int \mathrm{d}t\,\theta_1\dot{\theta}_2 + (\text{boundary term}).\end{aligned}$$

(1) The θ_j depend only on t, so $\mathrm{d}a = \mathrm{d}t \wedge \dot{a}$.
(2) Move $\mathrm{d}t$ to the front past the 1-form a (one sign), and $a \wedge \dot{a} = \frac{1}{4\pi^2}(\theta_1\dot{\theta}_2 - \theta_2\dot{\theta}_1)\mathrm{d}x^1 \wedge \mathrm{d}x^2$.
(3) $\int_{T^2} \mathrm{d}x^1 \wedge \mathrm{d}x^2 = 4\pi^2$ with orientation $\mathrm{d}t \wedge \mathrm{d}x^1 \wedge \mathrm{d}x^2$.
(4) $\theta_2\dot{\theta}_1 = \frac{\mathrm{d}}{\mathrm{d}t}(\theta_1\theta_2) - \theta_1\dot{\theta}_2$.

The Lagrangian $L = -\frac{k}{2\pi}\theta_1\dot{\theta}_2$ has the form $p\dot{q}$ with $q = \theta_2$, $p = -\frac{k}{2\pi}\theta_1$. It is first order in time: the two holonomies are canonically conjugate, and the phase space is the torus of flat connections itself, of finite area. Canonical quantisation $[q, p] = i$ gives

$$[\theta_1, \theta_2] = \frac{2\pi i}{k}$$

(the overall sign depends on the orientation and does not matter). Only the periodic operators $W_j = e^{i\theta_j}$, the Wilson loops, are well defined. Using $e^X e^Y = e^Y e^X e^{[X,Y]}$ when

$[X, Y]$ is a number,

$$W_1 W_2 \overset{(1)}{=} W_2 W_1 \, e^{[i\theta_1,\, i\theta_2]} \overset{(2)}{=} W_2 W_1 \, e^{-2\pi i/k} \;=:\; q\, W_2 W_1.$$

(1) Baker–Campbell–Hausdorff with a central commutator.
(2) $[i\theta_1, i\theta_2] = -[\theta_1, \theta_2] = -2\pi i/k$.

Its irreducible representation is the *clock and shift* pair on $\mathbb{C}^k$ with basis $e_0, \dots, e_{k-1}$:

$$W_1 e_j = q^j e_j, \qquad W_2 e_j = e_{j+1} \text{ (indices mod } k).$$

Check: $W_1 W_2 e_j = q^{j+1} e_{j+1}$ and $q W_2 W_1 e_j = q \cdot q^j e_{j+1}$, equal. No smaller representation exists: if $W_1 v = \mu v$, then $W_1(W_2 v) = q\mu\, W_2 v$, so $\mu, q\mu, q^2\mu, \dots$ are all eigenvalues, and q is a primitive kth root of unity, so there are k distinct ones. For $k = 2$: $q = -1$, $W_1 = \sigma_3$, $W_2 = \sigma_1$, and indeed $\sigma_3\sigma_1 = -\sigma_1\sigma_3$. So

$$\dim Z(T^2) = k, \qquad Z(T^3) = Z(T^2 \times S^1) = \dim Z(T^2) = k$$

by the trace axiom of Section 12.1.3. The Hilbert space is finite because the classical phase space (flat connections on T^2) is compact, with symplectic area proportional to k: one quantum state per unit of area $2\pi\hbar$.

12.5.5 Floer's supersymmetric Hamiltonian

In Chapter 7, Morse theory was the large-t limit of a supersymmetric quantum mechanics with Hamiltonian H_t, and the cohomology was its space of exact ground states. Nash's last point in §2 is that Floer homology has the same interpretation, with the Chern–Simons function as the "superpotential" of a quantum mechanics on $\mathcal{A}$, i.e. a quantum field theory on M.

Finite dimensions: the deformed Laplacian

Let $X = \mathbb{R}^n$ with the flat metric (the general case adds curvature terms that do not change the structure). Introduce fermion operators $a^{i*} = \mathrm{d}x^i \wedge$ (creates a form index) and $a^i = \iota_{\partial_i}$ (removes one), with

$$\{a^i, a^{j*}\} = \delta^{ij}, \qquad \{a^i, a^j\} = \{a^{i*}, a^{j*}\} = 0.$$

Then $\mathrm{d} = a^{i*}\partial_i$ and $\mathrm{d}^* = -a^i \partial_i$ on forms. Witten's deformation is $\mathrm{d}_t = e^{-tf}\,\mathrm{d}\, e^{tf}$ and $\mathrm{d}_t^* = e^{tf}\mathrm{d}^* e^{-tf}$, i.e.

$$\mathrm{d}_t = a^{i*} D_i, \quad D_i = \partial_i + t f_i; \qquad \mathrm{d}_t^* = -a^j \bar{D}_j, \quad \bar{D}_j = \partial_j - t f_j,$$

with $f_i = \partial_i f$, $f_{ij} = \partial_i \partial_j f$. The Hamiltonian is

$$\begin{aligned} H_t = \{\mathrm{d}_t, \mathrm{d}_t^*\} &\overset{(1)}{=} -\big(a^{i*} a^j D_i \bar{D}_j + a^j a^{i*} \bar{D}_j D_i\big) \\ &\overset{(2)}{=} -a^{i*} a^j \big[D_i, \bar{D}_j\big] - \bar{D}_i D_i \\ &\overset{(3)}{=} 2t\, f_{ij}\, a^{i*} a^j - \partial_i \partial_i - t\, f_{ii} + t^2 f_i f_i \\ &\overset{(4)}{=} -\partial_i \partial_i + t^2\, |\nabla f|^2 + t\, f_{ij}\, [a^{i*}, a^j]. \end{aligned} \tag{12.72}$$

(1) Expand the anticommutator; the a's commute with the differential operators $D, \bar{D}$.
(2) Write $a^j a^{i*} = \delta^{ij} - a^{i*} a^j$; the $a^{i*} a^j$ terms combine into a commutator of D's.

(3) $[D_i, \bar{D}_j] = [\partial_i, -tf_j] + [tf_i, \partial_j] = -2tf_{ij}$, and $\bar{D}_i D_i = \partial_i\partial_i + tf_{ii} - t^2 f_i f_i$ (the cross terms $\pm tf_i\partial_i$ cancel).
(4) $f_{ij}[a^{i*}, a^j] = f_{ij}(2a^{i*}a^j - \delta^{ij}) = 2f_{ij}a^{i*}a^j - f_{ii}$.

This is the second line of Nash's (12.38): Laplacian $+ t^2\langle \operatorname{grad} f, \operatorname{grad} f\rangle + t$ (Hessian)$\times$ (fermion bilinear).

Watch out: The first line of Nash's (12.38)

Nash writes $H_t = \mathrm{d}_t\mathrm{d}_t^* + \mathrm{d}_t^*\mathrm{d}_t = e^{-ft}(\mathrm{dd}^* + \mathrm{d}^*\mathrm{d})e^{ft}$. The first equality is the definition; the second is not right. Since $\mathrm{d}_t^* = e^{+tf}\,\mathrm{d}^* e^{-tf}$ has the opposite exponent, $\mathrm{d}_t\mathrm{d}_t^* = e^{-tf}\,\mathrm{d}\,e^{2tf}\,\mathrm{d}^* e^{-tf}$, which is not a conjugate of dd^*. Indeed, if H_t were conjugate to H_0 it would have the same spectrum for all t, and Witten's localisation would never happen. What *is* conjugate is the complex: $\mathrm{d}_t = e^{-tf}\,\mathrm{d}\,e^{tf}$, which is why $\operatorname{Ker}\mathrm{d}_t/\operatorname{Im}\mathrm{d}_t$ (hence the number of zero modes of H_t) is independent of t. The correct explicit form is the second line, (12.72).

Worked example 12.27: The harmonic oscillator is the simplest Witten Laplacian

Take $n = 1$ and $f = \frac{1}{2}x^2$. The form degrees are 0 and 1, and $[a^*, a] = a^*a - aa^*$ equals -1 on 0-forms (a kills them, $aa^* = 1$) and $+1$ on 1-forms (a^* kills them, $a^*a = 1$). So

$$H_t\big|_{\Omega^0} \overset{(1)}{=} -\partial_x^2 + t^2x^2 - t, \qquad H_t\big|_{\Omega^1} \overset{(1)}{=} -\partial_x^2 + t^2x^2 + t.$$

(1) (12.72) with $f_x = x$, $f_{xx} = 1$.

The oscillator $-\partial_x^2 + t^2x^2$ has levels $t(2m+1)$, $m = 0, 1, \dots$. So the spectrum on 0-forms is $\{0, 2t, 4t, \dots\}$ with ground state $e^{-tx^2/2} = e^{-tf}$, and on 1-forms it is $\{2t, 4t, \dots\}$. There is exactly one zero-energy state, in degree 0 = the Morse index of the minimum. Every excited level is paired between degrees 0 and 1 by the supercharge: this is the boson–fermion degeneracy that will make fermion and boson determinants cancel in Section 12.7.4.

Infinite dimensions: the Floer Hamiltonian

On $\mathcal{A}$ the coordinates are the field values $A_i^a(x)$, one for each point $x \in M$, each direction i and each colour a. The exterior derivative on $\mathcal{A}$ is a functional derivative (Nash's (12.39)):

$$\mathrm{d} = \int_M \operatorname{tr}\Big(a \wedge \frac{\delta}{\delta A}\Big), \tag{12.73}$$

where $a = a_i^a(x)\,T_a\,\mathrm{d}x^i$ is the fermionic analogue of a^{i*}: it multiplies a form on $\mathcal{A}$ by the 1-form $\delta A_i^a(x)$ (that is what the tangent vector a in (12.21) becomes when promoted to an operator). Its adjoint a^* removes such a factor. The operators satisfy canonical anticommutation relations, schematically $\{a_i^a(x), a_j^{b*}(y)\} = \delta^{ab}\delta_{ij}\delta^3(x-y)$.

Now substitute the Chern–Simons function into (12.72), term by term.

- *Laplacian.* $\mathcal{A}$ is an affine space with the flat metric (12.37), so $-\partial_i\partial_i$ becomes $-\int_M \operatorname{tr}(\delta/\delta A)^2$, the functional Laplacian.
- *Potential.* By (12.39),

$$\langle \operatorname{grad} f, \operatorname{grad} f\rangle \overset{(1)}{=} -\frac{1}{16\pi^4}\int_M \operatorname{tr}\big(\star F \wedge \star\star F\big) \overset{(2)}{=} -\frac{1}{16\pi^4}\int_M \operatorname{tr}\big(F \wedge \star F\big),$$

(1) The metric (12.37) with $a = b = \star F/4\pi^2$.

(2) $\star\star = 1$ on a Riemannian 3-manifold, and $\operatorname{tr}(\star F \wedge F) = \operatorname{tr}(F \wedge \star F)$ (1-form past 2-form).

which is the magnetic energy $\frac{1}{2}\int \sum_a |\vec{B}^a|^2$ up to normalisation.

- *Fermion term.* The Hessian is $4\pi^2 H(A) = \star \mathrm{d}_A$ (12.31), so $f_{ij}[a^{i*}, a^j]$ becomes a curl sandwiched between a and a^*: $\int_M \operatorname{tr}(a \wedge \mathrm{d}_A a^*)$, a Dirac-like (first-order) coupling of the fermions to the gauge field.

Absorbing the constants into t and the normalisation of a, one gets Nash's (12.40):

$$\mathcal{H}_t = \int_M \Big\{ -\operatorname{tr}\Big(\frac{\delta}{\delta A}\Big)^2 - 2t\ \operatorname{tr}\big(a \wedge \mathrm{d}_A a^*\big) - t^2\ \operatorname{tr}\big(F \wedge \star F\big) \Big\}. \tag{12.74}$$

Physical picture: It is a Yang–Mills Hamiltonian

In temporal gauge the canonical momentum of A_i is the electric field, $E_i = -i\,\delta/\delta A_i$, so $-\operatorname{tr}(\delta/\delta A)^2$ is the electric energy $\frac{1}{2}\vec{E}^2$ (with our sign of the trace), and $-\operatorname{tr}(F \wedge \star F)$ is the magnetic energy $\frac{1}{2}\vec{B}^2$. So (12.74) is

$$\mathcal{H}_t \sim \int_M \big(\vec{E}^2 + t^2 \vec{B}^2\big) + \text{fermions},$$

the Hamiltonian of Yang–Mills theory on M with coupling $\sim t^{-1/2}$, coupled to adjoint fermions. The states are functionals $\Psi[A]$ which must be gauge invariant (Gauss's law), i.e. functions on $\mathcal{A}/\mathcal{G}$. Large t is weak coupling, where the low-lying states localise on the classical vacua $F = 0$: the flat connections.

The fermions a, a^* are 1-forms on M, i.e. they carry a vector index, and yet they anticommute. Nash notes this "unconventional property of possessing integral spin". In a relativistic theory spin and statistics are tied together, but here there is no Lorentz symmetry forcing the connection: these are exactly the *twisted* fermions of topological field theory, which look like ghosts. The Fock space is graded by fermion number, and the grading that survives in infinite dimensions is the spectral-flow one (Nash's (12.41)):

$$\mathcal{H}_t = \bigoplus_{p=0}^{7} \Delta_p, \tag{12.75}$$

where Δ_p is the restriction of $\mathcal{H}_t$ to states of degree p mod 8. Hodge theory ("harmonic representatives") then suggests

$$HF_p(M) \ \cong\ \operatorname{Ker} \Delta_p = \{\text{ground states of degree } p\}.$$

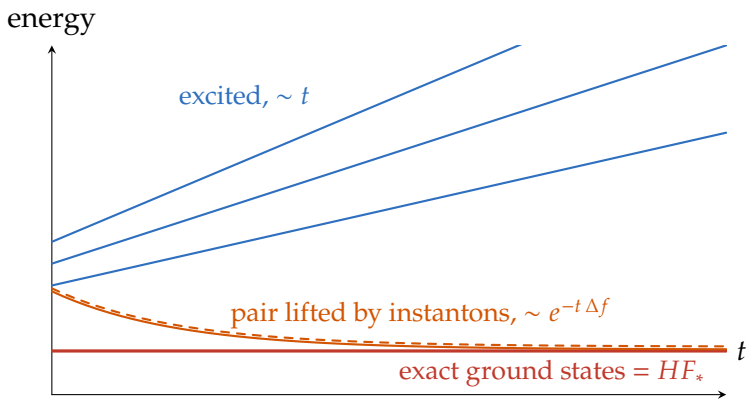


Figure 12.19: Schematic spectrum of Floer's Hamiltonian $\mathcal{H}_t$ as t grows. The excited levels rise like t (oscillator frequencies around each vacuum). One approximate ground state sits at each flat connection; pairs of them in adjacent degrees joined by an instanton are split away from zero by an energy of order $e^{-t\,\Delta f}$, where $\Delta f = S_{\mathrm{YM}}/8\pi^2$ is the instanton action. What stays exactly at zero is Floer homology.

Dictionary: Floer theory as supersymmetric quantum mechanics

Floer theory	Supersymmetric quantum mechanics / gauge theory
$\mathcal{A}/\mathcal{G}$	configuration space (gauge-invariant wavefunctionals)
Chern–Simons f	superpotential
flat connection α	classical vacuum
relative index $\mu(\alpha)$ mod 8	fermion number of the vacuum
instanton on $M \times \mathbb{R}$	tunnelling path
$n(\alpha,\beta)$	tunnelling amplitude, signed
∂	supercharge Q at large t
$HF_p(M)$	exact zero-energy states of degree p
$\chi_F(M) = 2\lambda(M)$	Witten index $\mathrm{Tr}(-1)^F e^{-\beta\mathcal{H}}$
metric independence	$\mathcal{H}_t = \{Q, Q^*\}$ is Q-exact

The Witten-index line deserves a remark. $\mathrm{Tr}(-1)^F e^{-\beta\mathcal{H}_t}$ receives no contribution from the paired excited states (a boson and a fermion with the same energy cancel), so it equals the number of bosonic minus fermionic ground states, $\chi_F(M)$, for all β and t. Casson's invariant is (half) the Witten index of Floer's quantum mechanics. Nash's closing remark, that this non-relativistic system has a relativistic four-dimensional generalisation (Witten), is the subject of the next two sections: the partition function of that 4d theory on a closed 4-manifold computes Donaldson's invariants.

Reading Nash (pp. 330–332, (12.30)–(12.41))

From the relative index mod 8 Nash gets a complex with eight groups (12.30) and defines $HF_p(M)$ (12.31); he states that $\partial^2 = 0$ follows by the instanton-tunnelling argument of his Chapter 7 and that the Floer groups are metric independent, and quotes Casson's invariant and Taubes' theorem $\chi_F = 2\lambda$ (12.32), explained via zeros of the 1-form F on $\mathcal{A}/\mathcal{G}$ (12.33). He then introduces the Chern–Simons action (12.34), its field equation $F = 0$ (12.35), the weight e^{iS} (12.36) and the level (12.37); finally he writes Floer's supersymmetric Hamiltonian (12.38)–(12.41). We have filled in: the grading convention, the explicit boundary operator,

the broken-flow-line proof of $\partial^2 = 0$ and the absence of bubbles, the invariance argument, HF of S^3 and $\Sigma(2,3,5)$, the rank-nullity computation of χ_F, Poincaré–Hopf with examples, Casson's Heegaard construction and surgery formula, the level-quantisation check, the k-dimensional Hilbert space of abelian Chern–Simons on T^2, and the derivation of the Witten Laplacian. Note that Nash's complex (12.30) is really cyclic, and that the first line of his (12.38) should be replaced by the second.

Recap

- Floer chains: irreducible flat connections, graded mod 8 by spectral flow from θ. Boundary: signed count of instantons on $M \times \mathbb{R}$ between generators of adjacent degree, modulo translation.
- $\partial^2 = 0$ because broken flow lines are the ends of 1-dimensional moduli spaces; no bubbling since a bubble costs index 8.
- $HF_p(M)$, $p \in \mathbb{Z}_8$, is a topological invariant; $HF(S^3) = 0$; $HF(\Sigma(2,3,5)) = \mathbb{Z}$ in degrees 1 and 5.
- $\chi_F(M) = \sum(-1)^{\mu(\alpha)} = 2\lambda(M)$ (Taubes), with λ Casson's intersection count on a Heegaard splitting.
- The Chern–Simons functional as an action is metric-free; its weight $e^{2\pi i k S}$ is gauge invariant for integer level k. Abelian Chern–Simons on T^2 has a k-dimensional Hilbert space.
- Floer homology is the space of ground states of $\mathcal{H}_t = -\int \mathrm{tr}(\delta/\delta A)^2 - 2t \int \mathrm{tr}(a \wedge \mathrm{d}_A a^*) - t^2 \int \mathrm{tr}(F \wedge \star F)$, a Yang–Mills Hamiltonian with twisted fermions.

12.6 Donaldson's polynomial invariants

From here on M is a four-dimensional manifold. In the Floer sections it was three-dimensional; the switch matters for every Hodge star below.

Here is the question this section answers. Suppose two closed 4-manifolds are *homeomorphic*: there is a continuous bijection between them with a continuous inverse. Are they *diffeomorphic*, that is, can the homeomorphism be chosen smooth? In every dimension except four the answer is understood and is "usually yes, and when no, for reasons one can list". In dimension four it is wildly false, and the first tool that could see the difference was built out of instantons.

A physicist should care for two reasons. First, the tool is a gauge theory: the invariants are integrals over the moduli space of Yang–Mills instantons, the same objects that dominate the Euclidean path integral of QCD. Second, Witten showed that the tool *is* a quantum field theory: Donaldson's invariants are correlation functions of a twisted supersymmetric Yang–Mills theory. That is the prototype of a Witten-type TQFT (Section 12.1.2).

Physical picture: Why instantons can see smooth structure

The anti-self-duality equation $F = -\star F$ needs a metric, and a metric needs a smooth structure (it is a smooth tensor field). So the space $\mathcal{M}_k$ of solutions knows about the smooth structure. The homeomorphism type does not enter the equation at all. If we can extract from $\mathcal{M}_k$ numbers that do not depend on the metric we chose, those numbers are

invariants of the *smooth* manifold, and they have a chance of distinguishing two smooth structures on the same topological manifold.

We first recall what Nash did in Chapter 1 with the first moduli space $\mathcal{M}_1$ (Chapter 1). There M had a positive definite intersection form. The moduli space $\mathcal{M}_1$ of SU(2) instantons of charge 1 is then a 5-manifold. Near its "end", where instantons shrink to zero size, it looks like $M \times (0, \epsilon)$: an infinitely concentrated instanton is labelled by the point where it sits. It has finitely many singular points, cones on $\mathbb{CP}^2$, one for each reducible connection. So $\mathcal{M}_1$ is a cobordism between M and a disjoint union of copies of $\mathbb{CP}^2$, and that forces the intersection form to be diagonal. Donaldson's later insight, which is what Nash's §3 describes, is that *all* the moduli spaces $\mathcal{M}_k$, $k = 1, 2, \ldots$, contain information, and that one can package it into polynomials on the second homology.

12.6.1 The stage: simply connected four-manifolds

Nash's setting is: M a smooth, simply connected, oriented, closed 4-manifold with a Riemannian metric g. Let us collect what we need to know about such an M.

Homology

Simply connected means $\pi_1(M) = 1$, so $H_1(M;\mathbb{Z}) = 0$ (it is the abelianisation of π_1). By Poincaré duality $H_3(M;\mathbb{Z}) \cong H^1(M;\mathbb{Z}) = 0$ as well. The only interesting group is $H_2(M;\mathbb{Z})$. It has no torsion: by the universal coefficient theorem its torsion equals that of H_1, which is zero. So

$$H_2(M;\mathbb{Z}) \cong \mathbb{Z}^{b_2}, \qquad b_0 = b_4 = 1, \quad b_1 = b_3 = 0.$$

Elements of $H_2(M;\mathbb{Z})$ can be pictured as closed oriented surfaces $\Sigma \subset M$ (every class is represented by an embedded surface in a 4-manifold).

The intersection form

Two surfaces in a 4-manifold generically meet in isolated points: two 2-dimensional objects in 4 dimensions have $2 + 2 - 4 = 0$ dimensional intersection, just as two lines in a plane meet in points.

Working definition 12.9 (intersection form)

For $\Sigma_1, \Sigma_2 \in H_2(M;\mathbb{Z})$, the **intersection form** $Q(\Sigma_1, \Sigma_2) = \Sigma_1 \cdot \Sigma_2 \in \mathbb{Z}$ is computed by either of:

- *counting:* move Σ_1, Σ_2 to meet transversally [at each meeting point the two tangent planes together span the 4-dimensional tangent space of M] and count intersection points, each with sign $+1$ if the orientation of Σ_1 followed by that of Σ_2 is the orientation of M, and -1 otherwise;
- *integrating:* with α_i closed 2-forms Poincaré dual to Σ_i [$\int_{\Sigma_i} \beta = \int_M \alpha_i \wedge \beta$ for every closed 2-form β] (Section 12.6.3), $Q(\Sigma_1, \Sigma_2) = \int_M \alpha_1 \wedge \alpha_2$.

Q is symmetric (2-forms commute), bilinear, and *unimodular*: its matrix in a basis of $H_2(M;\mathbb{Z})$ has determinant ± 1 (this is Poincaré duality over $\mathbb{Z}$). Formally: the cup product [the integer-coefficient version of the wedge product of closed forms] pairing $H^2(M;\mathbb{Z}) \times H^2(M;\mathbb{Z}) \to H^4(M;\mathbb{Z}) = \mathbb{Z}$.

Diagonalise Q over $\mathbb{R}$. The number of positive eigenvalues is b_2^+, the number of negative ones is b_2^-, and

$$b_2 = b_2^+ + b_2^-, \qquad \tau(M) = b_2^+ - b_2^- \ \text{(the signature)}, \qquad \chi(M) = \sum_j (-1)^j b_j = 2 + b_2.$$

Q is **even** if $Q(\Sigma, \Sigma)$ is even for every Σ, and **odd** otherwise. Freedman's theorem (Chapter 1) says that a simply connected closed topological 4-manifold is determined up to homeomorphism by Q (plus one $\mathbb{Z}_2$ invariant when Q is odd), and that every unimodular form occurs. So Q is the *whole* topology. Anything the instantons see beyond Q is smooth structure.

$b_2^\pm$ from the Hodge star

The numbers $b_2^\pm$ have a direct meaning in terms of forms, which is how they enter the instanton equations. On a Riemannian 4-manifold the Hodge star maps 2-forms to 2-forms and squares to one:

$$\star\star\,\omega \overset{(1)}{=} (-1)^{p(n-p)}\,\omega \overset{(2)}{=} (-1)^{2\cdot 2}\,\omega = \omega.$$

(1) The general rule for a p-form in n Riemannian dimensions (Chapter T1).
(2) $p = 2$, $n = 4$.

So $\star$ has eigenvalues ± 1 on 2-forms, and splits them into **self-dual** ($\star\omega = \omega$) and **anti-self-dual** ($\star\omega = -\omega$) parts: $\Lambda^2 = \Lambda^2_+ \oplus \Lambda^2_-$, each of rank 3. The star commutes with the Laplacian, so it preserves the harmonic 2-forms, $\mathcal{H}^2 = \mathcal{H}^2_+ \oplus \mathcal{H}^2_-$. For a harmonic self-dual form ω,

$$Q(\omega, \omega) = \int_M \omega \wedge \omega \overset{(1)}{=} \int_M \omega \wedge \star\omega \overset{(2)}{=} \|\omega\|^2 > 0, \tag{12.76}$$

(1) $\omega = \star\omega$.
(2) $\omega \wedge \star\omega = |\omega|^2\,\mathrm{vol}$ defines the pointwise norm; $\|\cdot\|^2$ is its integral.

and for an anti-self-dual one the same chain gives $-\|\omega\|^2 < 0$. Since $\mathcal{H}^2 \cong H^2_{\mathrm{dR}}(M)$ (Hodge's theorem), Q is positive on $\mathcal{H}^2_+$ and negative on $\mathcal{H}^2_-$, and these are Q-orthogonal. Hence

$$b_2^+ = \dim \mathcal{H}^2_+(g), \qquad b_2^- = \dim \mathcal{H}^2_-(g).$$

The *numbers* do not depend on g, but the *subspaces* do: as the metric varies, the plane $\mathcal{H}^2_+(g)$ moves around inside $H^2_{\mathrm{dR}}(M) \cong \mathbb{R}^{b_2}$. This will matter when we discuss reducible connections.

Worked example 12.28: $S^2 \times S^2$: the hyperbolic form

Take two round unit spheres with coordinates (θ_1, ϕ_1) and (θ_2, ϕ_2) and the product metric. An oriented orthonormal coframe is

$$e^1 = \mathrm{d}\theta_1, \quad e^2 = \sin\theta_1\,\mathrm{d}\phi_1, \quad e^3 = \mathrm{d}\theta_2, \quad e^4 = \sin\theta_2\,\mathrm{d}\phi_2, \qquad \mathrm{vol} = e^1 \wedge e^2 \wedge e^3 \wedge e^4.$$

The area forms are $\omega_1 = e^1 \wedge e^2$ and $\omega_2 = e^3 \wedge e^4$, each with total area 4π on its own factor. Their stars:

$$\star\omega_1 \overset{(1)}{=} e^3 \wedge e^4 = \omega_2, \qquad \star\omega_2 \overset{(2)}{=} e^1 \wedge e^2 = \omega_1.$$

(1) $e^{12} \wedge e^{34} = \mathrm{vol}$, so $\star e^{12} = e^{34}$.
(2) $e^{34} \wedge e^{12} = e^{12} \wedge e^{34} = \mathrm{vol}$ (2-forms commute).

Both forms are closed and co-closed, hence harmonic, and

$$\omega_1 + \omega_2 \text{ is self-dual}, \qquad \omega_1 - \omega_2 \text{ is anti-self-dual}.$$

So $b_2^+ = b_2^- = 1$. Now the intersection form. Let $S_1 = S^2 \times \{p\}$ and $S_2 = \{p\} \times S^2$. The form $\omega_2/4\pi$ is Poincaré dual to S_1: for any closed 2-form $\alpha = a_1\omega_1/4\pi + a_2\omega_2/4\pi$ (plus exact),

$$\int_M \alpha \wedge \frac{\omega_2}{4\pi} \overset{(1)}{=} \frac{a_1}{16\pi^2} \int_M \omega_1 \wedge \omega_2 \overset{(2)}{=} a_1 \overset{(3)}{=} \int_{S_1} \alpha.$$

(1) $\omega_2 \wedge \omega_2 = 0$ (it is a 4-form built only from the second factor's 2-dimensional coframe).
(2) $\int_M \omega_1 \wedge \omega_2 = (4\pi)(4\pi)$ by Fubini.
(3) On S_1 only ω_1 survives the restriction, and it integrates to 4π.

Similarly $\omega_1/4\pi$ is dual to S_2. Therefore

$$Q = \begin{pmatrix} S_1 \cdot S_1 & S_1 \cdot S_2 \\ S_2 \cdot S_1 & S_2 \cdot S_2 \end{pmatrix} = \frac{1}{16\pi^2} \begin{pmatrix} \int \omega_2^2 & \int \omega_2\omega_1 \\ \int \omega_1\omega_2 & \int \omega_1^2 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} =: H.$$

The counting picture agrees: $S^2 \times \{p\}$ and $S^2 \times \{q\}$ are disjoint for $p \neq q$, so $S_1 \cdot S_1 = 0$; and S_1 meets S_2 in the single point (p, p). The eigenvalues of H are ± 1, with eigenvectors $S_1 \pm S_2$, matching the harmonic forms $\omega_1 \pm \omega_2$. H is even: $(xS_1 + yS_2)^2 = 2xy$.

Worked example 12.29: $\mathbb{CP}^2$: the form $(+1)$

$H_2(\mathbb{CP}^2;\mathbb{Z}) = \mathbb{Z}$, generated by a complex projective line $L = \{z_2 = 0\} \cong \mathbb{CP}^1$. Two distinct lines in the projective plane meet in exactly one point, and complex submanifolds always meet positively (the complex orientation of L_1 followed by that of L_2 at a transverse intersection is the complex orientation of $\mathbb{CP}^2$). So $Q = (L \cdot L) = (+1)$: $b_2^+ = 1$, $b_2^- = 0$. The self-dual harmonic form is the Fubini–Study Kähler form (Kähler forms are self-dual; we check this on $\mathbb{C}^2$ in Example 12.39). Reversing the orientation gives $\overline{\mathbb{CP}}^2$ with $Q = (-1)$.

Connected sums

Nash's footnote 6 defines the **connected sum** $X\#Y$ of two n-manifolds: remove a small open n-ball from each, and glue the two resulting boundary spheres S^{n-1} together (with an orientation-reversing identification, so that the orientations of X and Y extend over the sum). The removed balls carry no second homology, so

$$H_2(X\#Y) = H_2(X) \oplus H_2(Y), \qquad Q_{X\#Y} = Q_X \oplus Q_Y, \qquad b_2^\pm(X\#Y) = b_2^\pm(X) + b_2^\pm(Y).$$

(Surfaces in X can be pushed off the ball, so they never meet surfaces in Y.)

12.6.2 Instantons and the dimension of $\mathcal{M}_k$

The equation and the conventions

Let $P \to M$ be a principal G-bundle with a connection A, and use the metric to define $\star$. Nash's instantons are now the **anti-self-dual** (ASD) connections (Nash's (12.42)):

$$F = -\star F. \tag{12.77}$$

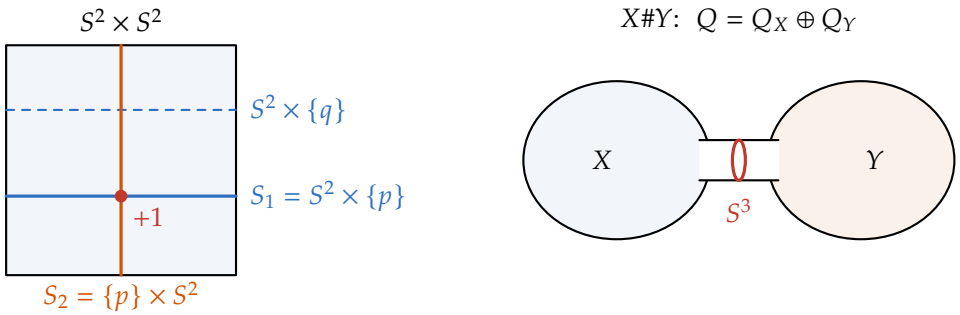


Figure 12.20: Left: in $S^2 \times S^2$ (drawn schematically as a square, one factor horizontal and one vertical) the sphere S_1 can be pushed to the disjoint parallel copy $S^2 \times \{q\}$, so $S_1 \cdot S_1 = 0$; it meets S_2 once, positively. Right: a connected sum is two manifolds joined by a thin tube whose cross-section is a 3-sphere. Surfaces live on one side or the other, so the intersection form is block diagonal.

M	b_2	b_2^+	b_2^-	τ	χ	Q
S^4	0	0	0	0	2	(empty)
$\mathbb{CP}^2$	1	1	0	1	3	(1)
$\overline{\mathbb{CP}}^2$	1	0	1	−1	3	(−1)
$S^2 \times S^2$	2	1	1	0	4	H (even)
$\mathbb{CP}^2\#\overline{\mathbb{CP}}^2$	2	1	1	0	4	(1) ⊕ (−1) (odd)
K3 surface	22	3	19	−16	24	$3H \oplus 2(-E_8)$ (even)
elliptic surface $E(3)$	34	5	29	−24	36	$5(1) \oplus 29(-1)$ (odd)
$5\mathbb{CP}^2\#29\overline{\mathbb{CP}}^2$	34	5	29	−24	36	$5(1) \oplus 29(-1)$ (odd)

Table 12.1: Simply connected 4-manifolds used in this section. E_8 is the positive definite even form of rank 8 (Chapter 1). The last two rows have the same Q, so by Freedman they are homeomorphic; we will see in Section 12.6.6 that Donaldson's invariants tell them apart.

This is the same equation as the flow equation on $M^3 \times \mathbb{R}$ (Eq. (12.46)), now on a closed 4-manifold. Our conventions are those of the Floer sections: A is anti-Hermitian, $F = \mathrm{d}A + A \wedge A$, and the instanton number is

$$k = \frac{1}{8\pi^2} \int_M \operatorname{tr}(F \wedge F). \tag{12.78}$$

For an ASD connection $k \geq 0$:

$$8\pi^2 k = \int_M \operatorname{tr}(F \wedge F) \overset{(1)}{=} -\int_M \operatorname{tr}(F \wedge \star F) \overset{(2)}{=} \tfrac{1}{2} \int_M \textstyle\sum_a |F^a|^2 \,\mathrm{vol} \;\geq 0,$$

(1) $F = -\star F$.

(2) Expand $F = F^a T_a$ with $\operatorname{tr}(T_a T_b) = -\frac{1}{2}\delta_{ab}$ (Eq. (12.9)); the same computation as the positivity of the metric (12.37).

with equality only for $F = 0$. In the physicists' dictionary ($A_{\text{math}} = -igA^a_{\text{phys}} \frac{\sigma^a}{2}$) this is the familiar statement that an anti-instanton or instanton (depending on one's orientation convention) saturates the bound $S_{\text{YM}} \geq 8\pi^2 |k|/g^2$.

The **moduli space** is

$$\mathcal{M}_k = \{A \text{ on } P_k \;:\; F_A = -\star F_A\} / \mathcal{G},$$

with P_k the bundle of instanton number k and $\mathcal{G}$ its gauge group. For $G = \mathrm{SU}(2)$ the bundle is classified by k alone (over a 4-dimensional complex, the only obstruction to triviality of an SU(2) bundle lives in $H^4(M; \pi_3(\mathrm{SU}(2))) = H^4(M; \mathbb{Z}) = \mathbb{Z}$, compare Section 12.2.3 where H^4 vanished for a 3-manifold).

The deformation complex

The dimension of $\mathcal{M}_k$ is the number of independent ways to deform an instanton and stay an instanton, not counting gauge directions. Perturb $A \to A + a$ with $a \in \Omega^1(\mathrm{ad}P)$. Then $F_{A+a} = F_A + \mathrm{d}_A a + a \wedge a$, so the linearised equation says the self-dual part of $\mathrm{d}_A a$ vanishes. Gauge directions are $a = \mathrm{d}_A \xi$ for $\xi \in \Omega^0(\mathrm{ad}P)$; we remove them by asking a to be L^2-orthogonal to them, $\mathrm{d}_A^* a = 0$. So the tangent space is the kernel of

$$D_A = \mathrm{d}_A^* \oplus \mathrm{d}_A^+ : \Omega^1(\mathrm{ad}P) \longrightarrow \Omega^0(\mathrm{ad}P) \oplus \Omega^2_+(\mathrm{ad}P), \qquad \mathrm{d}_A^+ a = \tfrac{1}{2}\big(\mathrm{d}_A a + \star \mathrm{d}_A a\big). \tag{12.79}$$

Equivalently, it is the middle cohomology H^1_A of the **deformation complex**

$$0 \longrightarrow \Omega^0(\mathrm{ad}P) \xrightarrow{\mathrm{d}_A} \Omega^1(\mathrm{ad}P) \xrightarrow{\mathrm{d}_A^+} \Omega^2_+(\mathrm{ad}P) \longrightarrow 0. \tag{12.80}$$

It is a complex because $\mathrm{d}_A^+ \mathrm{d}_A \xi = [F_A^+, \xi] = 0$ when A is ASD. This is Nash's complex (8.104) with self-dual and anti-self-dual exchanged: Chapter 8 studied self-dual instantons, whose linearisation lands in Ω^2_-. The three cohomology groups mean:

$$H^0_A = \{\xi : \mathrm{d}_A \xi = 0\} = \text{stabiliser of } A, \qquad H^1_A = T_{[A]}\mathcal{M}_k, \qquad H^2_A = \text{obstructions}.$$

If A is irreducible, $H^0_A = 0$; for generic metrics (and $b_2^+ > 0$), $H^2_A = 0$. Then $\dim \mathcal{M}_k = \dim H^1_A =$ − index of the complex, and the index theorem computes it.

Worked example 12.30: The symbol of D_A is a quaternion

Look at the simplest case: flat $\mathbb{R}^4$, one abelian component, orientation $\mathrm{d}x^1 \wedge \mathrm{d}x^2 \wedge \mathrm{d}x^3 \wedge \mathrm{d}x^4$. A basis of self-dual 2-forms is

$$\eta^1 = \mathrm{d}x^{12} + \mathrm{d}x^{34}, \quad \eta^2 = \mathrm{d}x^{13} - \mathrm{d}x^{24}, \quad \eta^3 = \mathrm{d}x^{14} + \mathrm{d}x^{23},$$

(check one: $\mathrm{d}x^{13} \wedge \mathrm{d}x^{24} = -\mathrm{vol}$, so $\star \mathrm{d}x^{13} = -\mathrm{d}x^{24}$). With $f_{\mu\nu} = \partial_\mu a_\nu - \partial_\nu a_\mu$, the four equations $Da = 0$ are

$$\begin{aligned} -\mathrm{d}^* a &= \partial_1 a_1 + \partial_2 a_2 + \partial_3 a_3 + \partial_4 a_4 = 0, \\ f_{12} + f_{34} &= \partial_1 a_2 - \partial_2 a_1 + \partial_3 a_4 - \partial_4 a_3 = 0, \\ f_{13} - f_{24} &= \partial_1 a_3 - \partial_3 a_1 - \partial_2 a_4 + \partial_4 a_2 = 0, \\ f_{14} + f_{23} &= \partial_1 a_4 - \partial_4 a_1 + \partial_2 a_3 - \partial_3 a_2 = 0. \end{aligned}$$

Four equations for four unknowns: the system is square, which is why an index can be finite. Replace ∂_μ by $i\xi_\mu$ (and drop the i) to get the **symbol**, the matrix acting on (a_1, a_2, a_3, a_4):

$$\sigma(\xi) = \begin{pmatrix} \xi_1 & \xi_2 & \xi_3 & \xi_4 \\ -\xi_2 & \xi_1 & -\xi_4 & \xi_3 \\ -\xi_3 & \xi_4 & \xi_1 & -\xi_2 \\ -\xi_4 & -\xi_3 & \xi_2 & \xi_1 \end{pmatrix}.$$

Multiply it by its transpose. The diagonal entries are $\xi_1^2 + \xi_2^2 + \xi_3^2 + \xi_4^2$; the off-diagonal ones vanish, e.g.

$$\begin{aligned} (\sigma\sigma^{\mathsf{T}})_{12} &\overset{(1)}{=} \xi_1(-\xi_2) + \xi_2\xi_1 + \xi_3(-\xi_4) + \xi_4\xi_3 = 0, \\ (\sigma\sigma^{\mathsf{T}})_{34} &\overset{(2)}{=} (-\xi_3)(-\xi_4) + \xi_4(-\xi_3) + \xi_1\xi_2 + (-\xi_2)\xi_1 = 0. \end{aligned}$$

(1) Row 1 dotted with row 2.
(2) Row 3 dotted with row 4; the other four pairs work the same way.

So $\sigma(\xi)\sigma(\xi)^{\mathsf{T}} = |\xi|^2\, \mathbb{1}_4$: $\sigma(\xi)$ is $|\xi|$ times an orthogonal matrix, invertible for every $\xi \neq 0$. This is the definition of an **elliptic** operator (Chapter 2). The matrix is right multiplication by the quaternion $\xi_1 - \xi_2\mathbf{i} - \xi_3\mathbf{j} - \xi_4\mathbf{k}$ in the basis $1, \mathbf{i}, \mathbf{j}, \mathbf{k}$ of $\mathbb{H} = \mathbb{R}^4$ (a unit quaternion times $|\xi|$): D is a "quaternionic Cauchy–Riemann operator", a Dirac operator in disguise. Its ellipticity is why $\mathcal{M}_k$ is finite-dimensional.

The dimension formula

Chapter 8 computed the index of the complex for *self-dual* instantons (Nash's (8.124)): $\dim \mathcal{M}_k = p_1(\mathrm{ad}_{\mathbb{C}}P) - \frac{\dim G}{2}(\chi - \tau)$. Our ASD connections on M are self-dual connections on $\overline{M}$, the same manifold with the opposite orientation. Reversing orientation leaves χ alone, flips τ, and swaps $b_2^+ \leftrightarrow b_2^-$. The instanton number is, by construction, the positive number k in both descriptions. So

$$\begin{aligned} \dim \mathcal{M}_k &\overset{(1)}{=} p_1(\mathrm{ad}_{\mathbb{C}}P) - \frac{\dim G}{2}\big(\chi(\overline{M}) - \tau(\overline{M})\big) \\ &\overset{(2)}{=} p_1(\mathrm{ad}_{\mathbb{C}}P) - \frac{\dim G}{2}\big(\chi(M) + \tau(M)\big) \\ &\overset{(3)}{=} p_1(\mathrm{ad}_{\mathbb{C}}P) - \frac{\dim G}{2}\big(2 + b_2^+ + b_2^- + b_2^+ - b_2^-\big) \\ &= p_1(\mathrm{ad}_{\mathbb{C}}P) - \dim G\,(1 + b_2^+). \end{aligned} \tag{12.81}$$

(1) Nash's (8.124) applied on $\overline{M}$, where A is self-dual.
(2) $\chi(\overline{M}) = \chi(M)$ and $\tau(\overline{M}) = -\tau(M)$.
(3) $\chi = 2 + b_2$ for a simply connected closed 4-manifold, and $\tau = b_2^+ - b_2^-$.

This is Nash's (12.43). The combination $1 + b_2^+$ has a direct meaning: it is $b_0 - b_1 + b_2^+$, the Euler characteristic of the *untwisted* complex $\Omega^0 \to \Omega^1 \to \Omega^2_+$ (the cohomologies are the constants, the harmonic 1-forms and the self-dual harmonic 2-forms). The index theorem says: twisted index = (curvature term) − dim $G\times$ (untwisted index).

Specialising to SU(2)

Nash next writes $p_1(\mathrm{ad}_{\mathbb{C}}P) = -2c_2(\mathrm{ad}_{\mathbb{C}}P)$ (his (12.44)), and for $G = \mathrm{SU}(2)$, $c_2(\mathrm{ad}_{\mathbb{C}}P) = -4k$. Let us check the factor 4 by splitting principle. Let E be the rank-2 complex vector bundle of P (fundamental representation), with formal Chern roots $\pm x$ (they sum to zero because SU(2) has $c_1 = 0$). The complexified adjoint bundle is $\mathrm{ad}_{\mathbb{C}}P = \mathrm{Sym}^2 E$, the spin-1 representation, whose weights are the sums of pairs of weights of E: $2x, 0, -2x$. Then

$$\begin{aligned} c_2(\mathrm{ad}_{\mathbb{C}}P) &\overset{(1)}{=} (2x)(0) + (2x)(-2x) + (0)(-2x) \\ &= -4x^2 \overset{(2)}{=} 4\,c_2(E). \end{aligned} \tag{12.82}$$

(1) c_2 is the second elementary symmetric function of the Chern roots.
(2) $c_2(E) = (x)(-x) = -x^2$.

With Nash's orientation convention (Chapter 8's (8.127)) $c_2(E) = -k$, so $c_2(\mathrm{ad}_{\mathbb{C}}P) = -4k$ and $p_1(\mathrm{ad}_{\mathbb{C}}P) = 8k$. Also $\dim \mathrm{SU}(2) = 3$. Hence Nash's (12.45):

$$\boxed{\dim \mathcal{M}_k = 8k - 3\,(1 + b_2^+)} \tag{12.83}$$

Watch out: Normalisations of p_1 and c_2

The standard definition is $p_1(V) = -c_2(V \otimes \mathbb{C})$ for a real bundle V, not $-2c_2$; and in the most common convention the instanton number of an ASD SU(2) connection is $c_2(E) = +k$. With those conventions one writes

$$\dim \mathcal{M}_k = -2p_1(\mathrm{ad}P) - 3(1 - b_1 + b_2^+) = 8c_2(E) - 3(1 + b_2^+).$$

Nash's letters differ (see the matching warning in Chapter 8), but the formula a reader should remember is convention-free: *eight real parameters per unit of instanton number, minus three for each of* $1 + b_2^+$.

Parity and Nash's condition that b_2^+ be odd

$8k$ is even, so $\dim \mathcal{M}_k$ is even exactly when $3(1 + b_2^+)$ is even, i.e. when b_2^+ is odd. Nash insists on this because the invariants will be integrals of products of *2-forms* over $\mathcal{M}_k$, which can only fill up an even-dimensional space (Section 12.6.4). For complex algebraic surfaces, the most important examples, the condition holds automatically: $b_2^+ = 1 + 2p_g$ is always odd (Example 12.39).

Worked example 12.31: Counting dimensions

Using (12.83) and Table 12.1:

- S^4: $b_2^+ = 0$, so $\dim \mathcal{M}_k = 8k - 3$; for $k = 1$ it is 5, the centre (4) and scale (1) of the BPST instanton (Chapter 8). It is odd, consistent with $b_2^+ = 0$ being even: S^4 has no Donaldson polynomials (it has no H_2 anyway).

- $\mathbb{CP}^2$, $S^2 \times S^2$: $b_2^+ = 1$, $\dim \mathcal{M}_k = 8k - 6$: $2, 10, 18, \ldots$
- K3: $b_2^+ = 3$, $\dim \mathcal{M}_k = 8k - 12$. For $k = 1$ this is -4: generically there are no charge-1 instantons at all. For $k = 2$ it is 4, for $k = 3$ it is 12.
- $E(3)$: $b_2^+ = 5$, $\dim \mathcal{M}_k = 8k - 18$: $6, 14, 22, \ldots$ for $k = 3, 4, 5, \ldots$
- $b_2^+ = 7$: $\dim \mathcal{M}_k = 8k - 24$, which is *zero* for $k = 3$. Then $\mathcal{M}_3$ is a finite set of points and the invariant is a signed count (Nash's (12.64); Section 12.8.5).

The dimension grows by 8 with each unit of k, the same 8 that made the Floer grading periodic (Section 12.4.5).

Reducible connections and the role of b_2^+

Irreducibility is what made $H^0_A = 0$. When can an SU(2) instanton be reducible? A reducible SU(2) connection has stabiliser containing a U(1), so the bundle splits, $E = L \oplus L^{-1}$, and $A = \operatorname{diag}(\alpha, -\alpha)$ with α a U(1) connection on the line bundle L. Let $e = c_1(L) \in H^2(M;\mathbb{Z})$. Then

$$\begin{aligned} k = \frac{1}{8\pi^2}\int_M \operatorname{tr}(F \wedge F) &\overset{(1)}{=} \frac{1}{8\pi^2}\int_M 2\,F_\alpha \wedge F_\alpha \\ &\overset{(2)}{=} \frac{1}{8\pi^2}\int_M 2(-2\pi i)^2\, c \wedge c \\ &\overset{(3)}{=} -Q(e,e). \end{aligned} \tag{12.84}$$

(1) $F = \operatorname{diag}(F_\alpha, -F_\alpha)$ and $F_\alpha \wedge F_\alpha$ appears twice in the trace.
(2) $c = \frac{i}{2\pi}F_\alpha$ is the Chern–Weil representative of $e = c_1(L)$, so $F_\alpha = -2\pi i\, c$.
(3) $(-2\pi i)^2 = -4\pi^2$, and $\int_M c \wedge c = Q(e,e)$.

F is ASD if and only if F_α is, and a closed ASD 2-form is automatically co-closed ($\mathrm{d} \star F_\alpha = -\mathrm{d}F_\alpha = 0$), hence harmonic. So:

Key idea: reducibles and b_2^+

Reducible ASD SU(2) connections of charge k correspond to classes $e \in H^2(M;\mathbb{Z})$ with $Q(e,e) = -k$ whose harmonic representative is anti-self-dual, i.e. e lies in $\mathcal{H}^2_-(g)$, i.e. e is Q-orthogonal to the b_2^+-dimensional subspace $\mathcal{H}^2_+(g)$. That is b_2^+ conditions on the metric. For a generic metric there are none when $b_2^+ \geq 1$; along generic *paths* of metrics there are none when $b_2^+ \geq 2$.

Watch out: The case $b_2^+ = 1$

When $b_2^+ = 1$ the metrics with reducibles form walls of codimension 1. Two metrics on opposite sides of a wall cannot be joined without passing through a reducible, and the invariants can jump there (**wall-crossing**). The invariants of $\mathbb{CP}^2$ and $S^2 \times S^2$ are therefore defined only relative to a chamber. Nash's statement that the invariants are metric-independent is correct for $b_2^+ > 1$ ($b_2^+ \geq 3$, given oddness). When $b_2^+ = 0$ reducibles are unavoidable; that is precisely the situation of Donaldson's diagonalisation theorem in Chapter 1, where the reducibles were the cone points of $\mathcal{M}_1$.

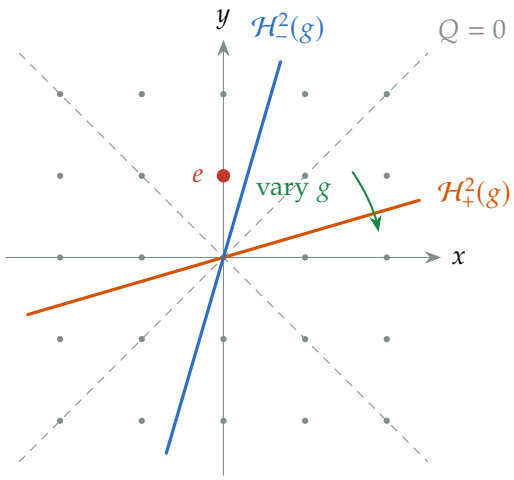


Figure 12.21: $H^2(M;\mathbb{R})$ for $M = \mathbb{CP}^2\#\overline{\mathbb{CP}}^2$, with $Q = x^2 - y^2$. The dashed light cone is $Q = 0$; lattice points are integral classes. The metric determines a positive line $\mathcal{H}^2_+(g)$ (orange) and its Q-orthogonal negative line $\mathcal{H}^2_-(g)$ (blue); they rotate like a Lorentz boost as g varies. A reducible instanton with $k = 1$ appears exactly when $\mathcal{H}^2_-(g)$ passes through a class e with $e^2 = -1$ (red). With $b_2^+ = 1$ this happens on a *wall* of codimension one in the space of metrics, so a path of metrics can be forced to cross it.

Reading Nash (pp. 332–333, (12.42)–(12.45))

Nash recalls that Chapter 1 used $\mathcal{M}_1$ for smoothability questions, and that Donaldson used all $\mathcal{M}_k$ to build invariants of simply connected 4-manifolds, also reachable by Witten's relativistic version of Floer's quantum mechanics. He takes M smooth, simply connected, oriented and closed with a metric, writes the ASD equation (12.42), quotes $\dim \mathcal{M}_k = p_1(\mathrm{ad}_{\mathbb{C}}P) - \dim G(1 + b_2^+)$ (12.43) "in the notation of chapter 8", rewrites it with $p_1 = -2c_2$ (12.44), remarks that it is even when b_2^+ is odd, and for SU(2) uses $c_2(\mathrm{ad}_{\mathbb{C}}P) = -4k$ to get $8k - 3(1 + b_2^+)$ (12.45). We have supplied: the homology of simply connected 4-manifolds, the intersection form and $b_2^\pm$ via $\star$, the examples $S^2 \times S^2$ and $\mathbb{CP}^2$, the deformation complex and its elliptic symbol, the orientation-reversal derivation of (12.43) from Chapter 8's (8.124), the splitting-principle check of the factor 4, and the role of b_2^+ in excluding reducibles.

12.6.3 From surfaces in M to 2-forms on $\mathcal{M}_k$

We now know that $\mathcal{M}_k$ is a manifold of dimension $8k - 3(1 + b_2^+)$. The next question is: *what number can we extract from it?* A manifold by itself gives only crude numbers (its dimension, its Euler characteristic), and $\mathcal{M}_k$ is not even compact. Donaldson's idea is to integrate differential forms over $\mathcal{M}_k$. To integrate, we need top-degree forms on $\mathcal{M}_k$, and to build those we need a supply of lower-degree forms that are natural, i.e. that come from the geometry of M and not from arbitrary choices.

The supply comes from the surfaces in M. Nash states the result as a map (his (12.47))

$$m : H_2(M) \longrightarrow H^2(\mathcal{M}_k), \tag{12.85}$$

which turns a surface $\Sigma \subset M$ into a 2-form $m(\Sigma)$ on the moduli space. (Most of the literature writes μ for this map; we use Nash's m and μ interchangeably.) Building m takes three ingredients: Poincaré duality, which turns surfaces into forms on M; the universal bundle, which turns forms on M into forms on $M \times \mathcal{M}_k$; and integration over M, which removes the M-dependence. We take them in turn.

Ingredient 1: Poincaré duality

A surface Σ is a place to integrate 2-forms. A 2-form is a thing to integrate. Poincaré duality says that, on a closed oriented manifold, these two kinds of object carry the same information: integrating over Σ is the same as wedging with a fixed form and integrating over all of M. Nash gives the definition in his footnote 5.

Working definition 12.10 (Poincaré dual of a submanifold)

Let M be closed and oriented of dimension n, and let $i : X \hookrightarrow M$ be a closed oriented submanifold of dimension r. The **Poincaré dual** of X is the class $X^* = [\omega] \in H^{n-r}_{\mathrm{dR}}(M)$ [a closed $(n-r)$-form ω, with two such forms counted as the same when they differ by an exact form] such that

$$\int_X i^*\alpha = \int_M \alpha \wedge \omega \qquad \text{for every closed } r\text{-form } \alpha. \tag{12.86}$$

Here $i^*\alpha$ is the pull-back [α restricted to X, fed only vectors tangent to X]. *How to find it:* take a thin tubular neighbourhood of X [the points of M within a small distance of X, shaped like X times a small normal disc], and let ω be a closed $(n-r)$-form supported in it whose integral over every small normal disc (oriented so that "X followed by the normal disc" is the orientation of M) equals 1. Then ω represents X^*. Nash writes a^* for the dual of a class $a \in H_2(M)$.

The degrees are forced. The left side of (12.86) integrates an r-form over an r-manifold. On the right, $\alpha \wedge \omega$ must be an n-form, so ω has degree $n-r$. For a surface in a 4-manifold, $n - r = 4 - 2 = 2$: *the Poincaré dual of a surface is again a 2-form*. This coincidence is special to the middle dimension, and it is why H_2 of a 4-manifold carries a symmetric form.

Why does the bump form work? Near X, split coordinates into r along X and $n-r$ normal to it. A closed form α is nearly constant across a thin normal disc. So integrating $\alpha \wedge \omega$ over the tube is integrating α along X and ω across the disc, and the second factor is 1. Figure 12.22 shows this for a circle on the torus.

Worked example 12.32: Poincaré duals of circles on the torus

Let $T^2 = \mathbb{R}^2/\mathbb{Z}^2$ with coordinates $x, y \in [0,1)$ and orientation $\mathrm{d}x \wedge \mathrm{d}y$. Let $a = \{y = y_0\}$, oriented by ∂_x, and $b = \{x = x_0\}$, oriented by ∂_y. Every closed 1-form is cohomologous to $\alpha = p\,\mathrm{d}x + q\,\mathrm{d}y$ with constants p, q. We look for $\omega_a = c\,\mathrm{d}y$:

$$\int_{T^2} \alpha \wedge \omega_a \overset{(1)}{=} \int_{T^2} (p\,\mathrm{d}x + q\,\mathrm{d}y) \wedge c\,\mathrm{d}y \overset{(2)}{=} pc \int_{T^2} \mathrm{d}x \wedge \mathrm{d}y = pc,$$

(1) Insert the two forms.
(2) $\mathrm{d}y \wedge \mathrm{d}y = 0$; the torus has unit area.

and this must equal $\int_a \alpha = p$. So $c = 1$ and $a^* = [\mathrm{d}y]$. The bump version is $\omega_a = \rho(y - y_0)\,\mathrm{d}y$ with $\int \rho = 1$ and ρ supported in $|y - y_0| < \epsilon$: it is cohomologous to $\mathrm{d}y$ because the difference $(\rho - 1)\mathrm{d}y = \mathrm{d}\big(\int^y (\rho - 1)\big)$ is exact (the integral of $\rho - 1$ over a full period vanishes, so the primitive is periodic).
For b, try $\omega_b = c\,\mathrm{d}x$:

$$\int_{T^2} \alpha \wedge \omega_b \overset{(1)}{=} qc \int_{T^2} \mathrm{d}y \wedge \mathrm{d}x \overset{(2)}{=} -qc,$$

(1) Only the $q\,\mathrm{d}y$ term survives against $\mathrm{d}x$.

(2) $\mathrm{d}y \wedge \mathrm{d}x = -\mathrm{d}x \wedge \mathrm{d}y$.

and this must equal $\int_b \alpha = q$. So $b^* = [-\mathrm{d}x]$. Now intersect:

$$\int_{T^2} a^* \wedge b^* \overset{(1)}{=} \int_{T^2} \mathrm{d}y \wedge (-\mathrm{d}x) \overset{(2)}{=} +1,$$

(1) The two duals just found.
(2) $-\mathrm{d}y \wedge \mathrm{d}x = \mathrm{d}x \wedge \mathrm{d}y$.

The geometric count agrees: a and b meet once, at (x_0, y_0), and there "∂_x followed by ∂_y" is the orientation of T^2, so the sign is $+1$. The minus sign in b^* is not a nuisance; it is what makes wedge product and signed counting agree.

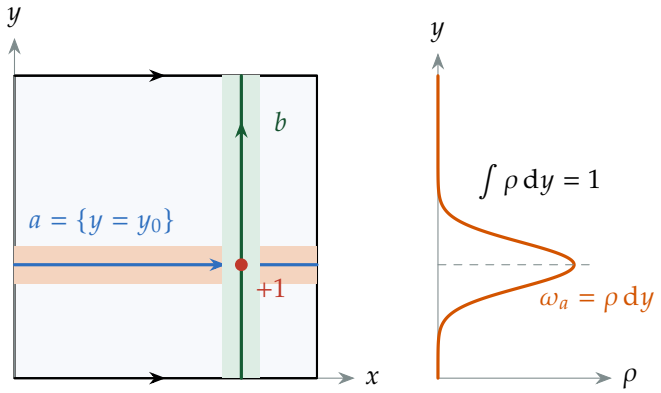


Figure 12.22: The Poincaré dual of a curve is a form concentrated on a thin tube around it. The circle a (blue) on the torus (a square with opposite sides identified) has dual $\omega_a = \rho(y)\,\mathrm{d}y$, a bump across the orange strip with unit total weight (profile on the right). Wedging with ω_a and integrating over the torus is therefore the same as integrating along a. The product of the two bump forms of a and b is concentrated on the small square where the strips cross, and integrates to the intersection sign $+1$.

Key idea: Wedge product of duals counts intersections

For closed oriented submanifolds $X, Y \subset M$ of complementary dimension meeting transversally,

$$\int_M X^* \wedge Y^* = \#(X \cap Y) \quad \text{(counted with signs)}.$$

Why: the product of the bump forms of X and Y is supported near the intersection points; near each one the two tubes cross like the two strips in Figure 12.22, and the integral is ± 1 depending on orientations. For surfaces in a 4-manifold this is the statement, used in Section 12.6.1, that $Q(\Sigma_1, \Sigma_2) = \int_M \Sigma_1^* \wedge \Sigma_2^*$.

Worked example 12.33: The line in $\mathbb{CP}^2$ and the Fubini–Study form

On $\mathbb{CP}^2$ with affine coordinates (z_1, z_2) (the chart $[1 : z_1 : z_2]$, which misses only a line at infinity, a set of measure zero), consider

$$\omega = \frac{i}{2\pi}\,\partial\bar\partial \log(1 + |z|^2) = \frac{i}{2\pi} \sum_{j,l} g_{j\bar l}\,\mathrm{d}z_j \wedge \mathrm{d}\bar z_l, \qquad g_{j\bar l} = \frac{(1 + |z|^2)\delta_{jl} - \bar z_j z_l}{(1 + |z|^2)^2}.$$

This is the (normalised) Fubini–Study Kähler form. It is closed because $\mathrm{d} = \partial + \bar\partial$ and $\partial^2 = \bar\partial^2 = 0$, $\partial\bar\partial = -\bar\partial\partial$.
Its integral over a line. Restrict to the line $L = \{z_2 = 0\}$, with coordinate $z = z_1 = re^{i\theta}$. There $\omega|_L = \frac{i}{2\pi}\,\mathrm{d}z \wedge \mathrm{d}\bar z/(1+r^2)^2$, and

$$\int_L \omega \overset{(1)}{=} \frac{1}{2\pi}\int_{\mathbb{R}^2} \frac{2\,\mathrm{d}x\wedge\mathrm{d}y}{(1+r^2)^2} \overset{(2)}{=} \frac{1}{\pi}\cdot 2\pi\int_0^\infty \frac{r\,\mathrm{d}r}{(1+r^2)^2} \overset{(3)}{=} 2\cdot\frac{1}{2} = 1.$$

(1) $i\,\mathrm{d}z\wedge\mathrm{d}\bar z = i(\mathrm{d}x + i\mathrm{d}y)\wedge(\mathrm{d}x - i\mathrm{d}y) = 2\,\mathrm{d}x\wedge\mathrm{d}y$.
(2) Polar coordinates.
(3) $\int_0^\infty r(1+r^2)^{-2}\mathrm{d}r = \left[-\frac{1}{2}(1+r^2)^{-1}\right]_0^\infty = \frac{1}{2}$.

So ω integrates to 1 on the generator L of $H_2(\mathbb{CP}^2) = \mathbb{Z}$. Since $H^2_{\mathrm{dR}}(\mathbb{CP}^2) = \mathbb{R}$, this says $[\omega] = L^*$ (both pair to 1 with L, and the pairing is non-degenerate).
Its square. First the determinant of g: the matrix $(1+|z|^2)\mathbb{1} - \bar z z^{\mathsf T}$ has eigenvalue $1+|z|^2$ on vectors orthogonal to $\bar z$ and eigenvalue $(1+|z|^2) - |z|^2 = 1$ on $\bar z$ itself, so $\det g = (1+|z|^2)(1+|z|^2)^{-4} = (1+|z|^2)^{-3}$. Then

$$\int_{\mathbb{CP}^2}\omega\wedge\omega \overset{(1)}{=} \left(\frac{i}{2\pi}\right)^2\int 2\det g\;\mathrm{d}z_1\wedge\mathrm{d}\bar z_1\wedge\mathrm{d}z_2\wedge\mathrm{d}\bar z_2$$
$$\overset{(2)}{=} \frac{1}{4\pi^2}\int_{\mathbb{R}^4}\frac{2\cdot 4\,\mathrm{d}^4x}{(1+r^2)^3} \overset{(3)}{=} \frac{2}{\pi^2}\cdot 2\pi^2\int_0^\infty\frac{r^3\,\mathrm{d}r}{(1+r^2)^3} \overset{(4)}{=} 4\cdot\frac{1}{4} = 1.$$

(1) Expanding the wedge square, only terms with $\{j,k\} = \{l,m\} = \{1,2\}$ survive; reordering the four 1-forms gives $2(g_{1\bar1}g_{2\bar2} - g_{1\bar2}g_{2\bar1}) = 2\det g$ times $\mathrm{d}z_1\mathrm{d}\bar z_1\mathrm{d}z_2\mathrm{d}\bar z_2$.
(2) $(i\,\mathrm{d}z_1\wedge\mathrm{d}\bar z_1)\wedge(i\,\mathrm{d}z_2\wedge\mathrm{d}\bar z_2) = 4\,\mathrm{d}^4x$ with the complex orientation, and $\det g = (1+r^2)^{-3}$.
(3) The integrand is radial; the unit 3-sphere has volume $2\pi^2$.
(4) With $u = r^2$: $\int_0^\infty r^3(1+r^2)^{-3}\mathrm{d}r = \frac{1}{2}\int_0^\infty u(1+u)^{-3}\mathrm{d}u = \frac{1}{2}\cdot\frac{1}{2}$.

So $\int L^*\wedge L^* = 1 = L\cdot L$, matching the geometric count in Example 12.29: two lines meet in one point. We will use this computation twice more: as the model of a "positive" class in Section 12.6.6, and as the simplest instance of the integrals that define q_d.

Ingredient 2: the space of gauge orbits as a classifying space

Nash writes "let $E_{\mathcal G}$ be the universal $\mathcal G$-bundle over $B\mathcal G$" and, a line later, "$\mathcal M_k \subset B\mathcal G$". Both statements hide a construction, which we now supply.

Classifying spaces, in one paragraph. For a group H, a **universal bundle** is a principal H-bundle $EH \to BH$ whose total space EH is contractible. BH is the **classifying space**. Its defining property: every principal H-bundle over any (reasonable) space Y is the pullback of EH along a map $Y \to BH$, unique up to homotopy. So characteristic classes of H-bundles are pulled back from $H^*(BH)$. We met two examples in Section 12.2.3: $S^\infty \to \mathbb{CP}^\infty$ for U(1), and $S^\infty \to \mathbb{H}P^\infty$ for SU(2).

The gauge group version. Let $\mathcal A_k$ be the space of all connections on P_k, and $\mathcal A_k^*$ the irreducible ones. Then:

1. $\mathcal A_k$ is contractible: it is an affine space, and $A_t = (1-t)A_0 + tA_1$ joins any two connections.

2. $\mathcal A_k^*$ is still (weakly) contractible: the reducibles form a subset of infinite codimension, and removing a subset of infinite codimension from an infinite-dimensional contractible space does not create any homotopy (just as $\mathbb{R}^N$ minus a point has $\pi_j = 0$ for $j < N-1$,

and $N = \infty$).

3. The gauge group acts freely on $\mathcal{A}_k^*$ *except* for the central elements $\pm\mathbb{1}$, which fix every connection (for constant $g = \pm\mathbb{1}$, $g^{-1}Ag + g^{-1}\mathrm{d}g = A$). So $\bar{\mathcal{G}} = \mathcal{G}/\{\pm\mathbb{1}\}$ acts freely.

A free action of $\bar{\mathcal{G}}$ on a contractible space is exactly the data of a universal bundle. Hence

$$\mathcal{A}_k^* \longrightarrow \mathcal{B}_k^* := \mathcal{A}_k^*/\mathcal{G} \quad \text{is the universal } \bar{\mathcal{G}}\text{-bundle, and} \quad \mathcal{B}_k^* \simeq B\bar{\mathcal{G}}. \tag{12.87}$$

Nash's "$B\mathcal{G}$" is this $\mathcal{B}_k^*$, and his (12.49), $\mathcal{M}_k \subset B\mathcal{G}$, is the obvious inclusion of the irreducible instantons into all irreducible gauge orbits.

Ingredient 3: the universal bundle over $M \times \mathcal{B}^*$

A point of $\mathcal{B}_k^*$ is a gauge orbit $[A]$. Each connection A lives on the same bundle P over M. So we have a *family* of bundles-with-connection parametrised by $\mathcal{B}_k^*$. Families of bundles over M parametrised by Y are bundles over $M \times Y$. Here is the bundle.

Working definition 12.11 (the universal bundle over $M \times \mathcal{B}^*$)

The gauge group acts on $\mathcal{A}_k^*$ (by gauge transformations) and on the adjoint bundle $\mathrm{ad}P$ [a copy of $\mathfrak{su}(2)$ over each point x of M, where the Lie-algebra-valued fields live] (a gauge transformation is a bundle automorphism, so it acts fibrewise on $\mathrm{ad}P$ by Ad_g). Let it act on both at once and form the quotient [a pair $(A,(x,v))$ and its gauge transform $(A^g,(x,g^{-1}vg))$ become the same point]

$$\mathbb{E} := \big(\mathcal{A}_k^* \times \mathrm{ad}P\big)/\mathcal{G} \longrightarrow \mathcal{B}_k^* \times M, \qquad [A,\,(x,v)] \mapsto ([A],\,x). \tag{12.88}$$

It is a real rank-3 vector bundle (an SO(3)-bundle). Its fibre over $([A],x)$ is "the Lie algebra at x, as seen by the connection A". Its restriction to $\{[A]\} \times M$ is $\mathrm{ad}P$ itself.

Why $\mathrm{ad}P$ *and not* P*?* The central element $-\mathbb{1}$ acts trivially on $\mathcal{A}_k^*$ but as -1 on the fibres of P (or of the rank-2 bundle E). The quotient $(\mathcal{A}^* \times E)/\mathcal{G}$ would then not be a vector bundle: the fibre over $([A],x)$ would be $E_x/\{\pm1\}$. On the adjoint bundle $-\mathbb{1}$ acts as $\mathrm{Ad}_{-\mathbb{1}} = +1$, so the problem disappears. The price is that $\mathbb{E}$ is an SO(3)-bundle, and its natural characteristic class is $p_1(\mathbb{E})$. On any subspace where the SU(2) lift exists, $p_1(\mathrm{ad}E) = -4c_2(E)$ (a real form of (12.82)). So Nash's "$c_2(E_{\mathcal{G}})$" is to be read as

$$c_2 := -\tfrac{1}{4}\,p_1(\mathbb{E}) \;\in H^4(\mathcal{B}_k^* \times M;\mathbb{Q}). \tag{12.89}$$

The $\frac{1}{4}$ is why careful authors say the invariants are rational numbers which turn out to be integers (or integers up to a known power of 2).

The map m

We have a 4-class $c_2(\mathbb{E})$ on the ∞-dimensional space $\mathcal{B}_k^* \times M$. Split it by bidegree, (p,q) meaning p directions along $\mathcal{B}_k^*$ and q along M. By the Künneth theorem,

$$H^4(\mathcal{B}_k^*\times M) = \bigoplus_{p+q=4} H^p(\mathcal{B}_k^*)\otimes H^q(M) = H^4(\mathcal{B}^*)\otimes H^0(M) \oplus H^2(\mathcal{B}^*)\otimes H^2(M) \oplus H^0(\mathcal{B}^*)\otimes H^4(M), \tag{12.90}$$

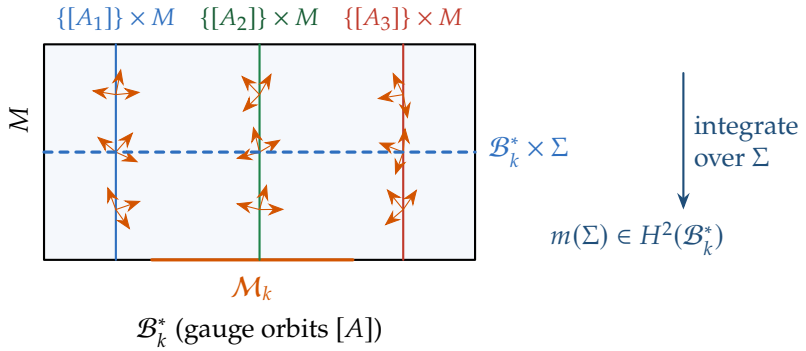


Figure 12.23: The universal bundle. The rectangle is $\mathcal{B}^*_k \times M$; each vertical slice $\{[A]\} \times M$ is a copy of M carrying the bundle $\mathrm{ad}P$ (orange triads) with the connection A. As $[A]$ moves horizontally the internal frames twist, and $c_2(\mathbb{E})$ measures that twisting. Restricting to $\mathcal{B}^*_k \times \Sigma$ (dashed) and integrating out the two Σ-directions leaves a 2-form on $\mathcal{B}^*_k$, and hence on $\mathcal{M}_k$ (orange segment).

where we used $H^1(M) = H^3(M) = 0$ for simply connected M (and the odd cohomology of $\mathcal{B}^*$ plays no role below). Nash's construction (12.48)–(12.51) extracts the middle piece. In a chain:

$$\begin{aligned} m(a) &\overset{(1)}{=} \int_M c_2(\mathbb{E}) \wedge a^* \\ &\overset{(2)}{=} \int_M c_2(\mathbb{E})^{(2,2)} \wedge a^* \\ &\overset{(3)}{=} \int_\Sigma c_2(\mathbb{E})^{(2,2)}\Big|_{\mathcal{B}^*\times\Sigma} \in H^2(\mathcal{B}^*_k), \qquad a = [\Sigma]. \end{aligned} \tag{12.91}$$

(1) Nash's definition (12.51): wedge with the Poincaré dual a^* (a form on M, pulled back to $\mathcal{B}^* \times M$), which gives a 6-form (12.48), then integrate over the four M-directions (12.50). Integration over the fibre M of the projection $\mathcal{B}^* \times M \to \mathcal{B}^*$ lowers the degree by 4: $6 - 4 = 2$.

(2) Only components with exactly two M-directions survive: a^* already supplies two, and the integral needs four in total.

(3) The defining property (12.86) of the Poincaré dual, applied slice by slice in $\mathcal{B}^*$.

Finally restrict from $\mathcal{B}^*_k$ to the submanifold $\mathcal{M}_k$ (Nash's (12.49)): $m(a)|_{\mathcal{M}_k} \in H^2(\mathcal{M}_k)$. The map is linear in a, because $a \mapsto a^*$ is linear and integration is linear.

Physical picture: What $m(\Sigma)$ measures

A 2-form on $\mathcal{M}_k$ assigns a number to every small 2-dimensional family of instantons, or, after integrating, to every closed surface S of instantons in $\mathcal{M}_k$. Take such a family $S \subset \mathcal{M}_k$. It gives a bundle $\mathbb{E}|_{S\times\Sigma}$ over the 4-manifold $S \times \Sigma$, and $\langle m(\Sigma), S\rangle = \int_{S\times\Sigma} c_2(\mathbb{E})$ is its instanton number. So $m(\Sigma)$ asks: *as the instanton moves around the family S, how much do its internal frames over Σ twist?*

A formula for the representative

It is worth having an explicit form, because the same expression will reappear in Witten's field theory as an observable. The bundle $\mathbb{E}$ carries a natural connection, whose curvature $\mathbb{F}$ at a point $([A], x)$ has three pieces (Atiyah–Singer, Donaldson):

- bidegree $(0,2)$ (two M-directions): $\mathbb{F}^{(0,2)} = F_A(x)$, the curvature of A itself;

- bidegree $(1,1)$: for a tangent vector $a \in T_{[A]}\mathcal{B}^*$, represented by a 1-form $a \in \Omega^1(\mathrm{ad}P)$ in Coulomb gauge $\mathrm{d}_A^* a = 0$, and a tangent vector v to M, $\mathbb{F}^{(1,1)}(a, v) = a(v)$: *the mixed curvature is the tangent vector itself*;
- bidegree $(2,0)$: for two tangent vectors a, b, $\mathbb{F}^{(2,0)}(a,b) = \Phi(a,b)$, where $\Phi(a,b) \in \Omega^0(\mathrm{ad}P)$ solves the Laplace equation $\mathrm{d}_A^* \mathrm{d}_A \Phi = -2 \star [a \wedge \star b]$ (up to a normalisation constant depending on conventions). It exists because A is irreducible, so $\mathrm{d}_A^* \mathrm{d}_A$ is invertible.

(The $(2,0)$ piece is the curvature of the Coulomb-gauge "connection" on $\mathcal{A}^* \to \mathcal{B}^*$: two horizontal directions a, b fail to commute by an infinitesimal gauge transformation, and $\Phi(a,b)$ is its generator.) Then Chern–Weil (Chapter 8) gives $c_2(\mathbb{E}) = \frac{1}{8\pi^2}\operatorname{tr}(\mathbb{F} \wedge \mathbb{F})$ in our conventions, and its $(2,2)$ part is

$$\begin{aligned} \operatorname{tr}(\mathbb{F} \wedge \mathbb{F})^{(2,2)} &\overset{(1)}{=} \operatorname{tr}\big(\mathbb{F}^{(0,2)}\mathbb{F}^{(2,0)} + \mathbb{F}^{(2,0)}\mathbb{F}^{(0,2)} + \mathbb{F}^{(1,1)} \wedge \mathbb{F}^{(1,1)}\big) \\ &\overset{(2)}{=} 2\operatorname{tr}\big(F_A\,\Phi\big) + \operatorname{tr}\big(\mathbb{F}^{(1,1)} \wedge \mathbb{F}^{(1,1)}\big). \end{aligned} \tag{12.92}$$

(1) Expand $\mathbb{F} = \mathbb{F}^{(0,2)} + \mathbb{F}^{(1,1)} + \mathbb{F}^{(2,0)}$ and keep the products whose bidegrees add to $(2,2)$.
(2) The trace is cyclic, and the $(2,0)$ piece Φ is a 0-form on M, so it commutes past F_A under the trace.

Evaluated on two tangent vectors a, b of $\mathcal{M}_k$ and integrated over Σ, this gives

$$m(\Sigma)(a,b) = \frac{1}{4\pi^2}\int_\Sigma \operatorname{tr}\big(\Phi(a,b)\,F_A + a \wedge b\big) \qquad \text{(up to an overall sign fixed by orientations).} \tag{12.93}$$

Keep an eye on the two terms, "ΦF" and "$a \wedge b$". In Witten's theory they will be ϕF and $\frac{1}{2}\psi \wedge \psi$ (Section 12.7.5).

Worked example 12.34: A toy universal bundle: flat $\mathrm{U}(1)$ connections on T^2

Everything above can be seen in a finite-dimensional abelian model, one dimension lower. Let $M = T^2 = \mathbb{R}^2/\mathbb{Z}^2$ and consider flat U(1) connections on the trivial line bundle,

$$A_\theta = i(\theta_1\,\mathrm{d}x + \theta_2\,\mathrm{d}y), \qquad \theta \in \mathbb{R}^2.$$

(Anti-Hermitian, so U(1) connections are i times real 1-forms.) A gauge transformation $g = e^{2\pi i(mx+ny)}$, $m, n \in \mathbb{Z}$, is well defined on T^2 and sends

$$A_\theta \mapsto g^{-1}A_\theta g + g^{-1}\mathrm{d}g \overset{(1)}{=} A_\theta + 2\pi i(m\,\mathrm{d}x + n\,\mathrm{d}y) \overset{(2)}{=} A_{\theta + 2\pi(m,n)}.$$

(1) U(1) is abelian, so $g^{-1}Ag = A$; and $g^{-1}\mathrm{d}g = 2\pi i(m\,\mathrm{d}x + n\,\mathrm{d}y)$.
(2) Read off the new θ.

So the moduli space of flat connections is the *dual torus* $\hat{T} = \mathbb{R}^2/(2\pi\mathbb{Z})^2$, with coordinates θ_1, θ_2. The universal line bundle over $\hat{T} \times T^2$ (the **Poincaré line bundle**) carries the "universal connection" $\mathbb{A} = i(\theta_1 \mathrm{d}x + \theta_2 \mathrm{d}y)$, viewed as a 1-form on the product (it is defined on $\mathbb{R}^2 \times T^2$ and glued by the gauge transformations above). Its curvature is

$$\mathbb{F} = \mathrm{d}\mathbb{A} \overset{(1)}{=} i\big(\mathrm{d}\theta_1 \wedge \mathrm{d}x + \mathrm{d}\theta_2 \wedge \mathrm{d}y\big),$$

(1) d on the product acts on the θ's too; $\mathrm{d}(\mathrm{d}x) = 0$.

This is purely of mixed bidegree $(1,1)$: $F_A = 0$ since the connections are flat, and $\Phi = 0$ since the group is abelian. The mixed term is the tangent vector: moving θ_1 by $\delta\theta_1$ changes A by $i\,\delta\theta_1\,\mathrm{d}x$, which is what $\mathbb{F}$ pairs with ∂_{θ_1}. The first Chern class is

$c_1 = \frac{i}{2\pi}\mathbb{F} = -\frac{1}{2\pi}(\mathrm{d}\theta_1 \wedge \mathrm{d}x + \mathrm{d}\theta_2 \wedge \mathrm{d}y)$. Now slant with the circle $a = \{y = y_0\}$, which has dual $a^* = \mathrm{d}y$ (Example 12.32). Integrating over T^2 (the x, y directions, placed on the right):

$$m(a) = \int_{T^2} c_1 \wedge a^* \overset{(1)}{=} -\frac{1}{2\pi}\int_{T^2} \mathrm{d}\theta_1 \wedge \mathrm{d}x \wedge \mathrm{d}y \overset{(2)}{=} -\frac{\mathrm{d}\theta_1}{2\pi}.$$

(1) The $\mathrm{d}\theta_2 \wedge \mathrm{d}y$ term dies against $\mathrm{d}y$.
(2) $\int_{T^2} \mathrm{d}x \wedge \mathrm{d}y = 1$, with the $\hat{T}$-form $\mathrm{d}\theta_1$ pulled out to the left.

So a 1-cycle of M produced a 1-form on moduli space, whose integral round the θ_1-circle of $\hat{T}$ is -1. Likewise $b = \{x = x_0\}$, with $b^* = -\mathrm{d}x$, gives $m(b) = -\frac{1}{2\pi}\int \mathrm{d}\theta_2 \wedge \mathrm{d}y \wedge (-\mathrm{d}x) = -\mathrm{d}\theta_2/2\pi$ (since $\mathrm{d}y \wedge \mathrm{d}x = -\mathrm{d}x \wedge \mathrm{d}y$). The "toy invariant" is then

$$\int_{\hat{T}} m(a) \wedge m(b) = \frac{1}{4\pi^2}\int_{\hat{T}} \mathrm{d}\theta_1 \wedge \mathrm{d}\theta_2 = 1,$$

a number built from cycles in M and the geometry of the moduli space. In Donaldson theory the cycles are surfaces, the forms have degree 2, and the integral is over $\mathcal{M}_k$; the logic is identical.

Try it: Degrees in the toy model

In the toy model M had dimension 2 and the curvature class had degree 2. Show that a j-cycle of M gives a $(2-j)$-form on $\hat{T}$, and check that the 0-cycle "a point p" gives $m(p) = c_1|_{\hat{T}\times\{p\}} = 0$. *Answer:* $c_1 \wedge$ (dual of a j-cycle) has degree $2 + (2-j)$; integrating over the 2-dimensional M leaves $2 - j$. For a point the dual is the volume form of T^2, which kills both terms of c_1 (each contains a T^2-direction), so $m(p) = 0$: the Poincaré bundle is trivial on each slice $\hat{T} \times \{p\}$. In Donaldson theory the analogous point class is *not* zero; it is the degree-4 class of Section 12.7.5.

12.6.4 The polynomials q_d

Now we can integrate. Choose d surfaces $a_1, \dots, a_d \in H_2(M;\mathbb{Z})$. Each gives a 2-form $m(a_i)$ on $\mathcal{M}_k$. Their wedge product is a $2d$-form. It can be integrated over $\mathcal{M}_k$ exactly when it has top degree:

$$2d = \dim \mathcal{M}_k \qquad \Longleftrightarrow \qquad d = \tfrac{1}{2}\dim \mathcal{M}_k = 4k - \tfrac{3}{2}(1 + b_2^+). \tag{12.94}$$

This is why Nash insisted that $\dim \mathcal{M}_k$ be even: products of 2-forms have even degree, so on an odd-dimensional moduli space no product of m's has top degree and every integral of the kind below is zero.

Working definition 12.12 (Donaldson's polynomial invariant)

For $d = \frac{1}{2}\dim \mathcal{M}_k$ (Nash's (12.52)),

$$q_d(M)(a_1, \dots, a_d) = \int_{\overline{\mathcal{M}}_k} m(a_1) \wedge \cdots \wedge m(a_d) \ \in \mathbb{Z}, \tag{12.95}$$

where $\overline{\mathcal{M}}_k$ is a compactification of $\mathcal{M}_k$ [$\mathcal{M}_k$ with the limits added in which instantons shrink to points] (Section 12.6.5) and the metric is generic [chosen so that $\mathcal{M}_k$ is a smooth manifold of the expected dimension]. *How to compute, geometrically:* represent each $m(a_i)$ by a codimension-2 submanifold $V_{a_i} \subset \mathcal{M}_k$ (its Poincaré dual in $\mathcal{M}_k$), put them in general position [moved slightly so that they meet transversally, in finitely many isolated points],

and count the points of $V_{a_1} \cap \cdots \cap V_{a_d}$ with signs. Formally, q_d is a symmetric d-linear form on $H_2(M;\mathbb{Z})$, i.e. an element of $\mathrm{Sym}^d(H^2(M))$, or equivalently a homogeneous polynomial of degree d on $H_2(M)$ (Nash's (12.46)).

Let us unpack each property.

Multilinear: each $m(a_i)$ is linear in a_i and the wedge product is linear in each factor.

Symmetric: 2-forms commute. For two of them,

$$m(a_1) \wedge m(a_2) \overset{(1)}{=} (-1)^{2\cdot 2}\, m(a_2) \wedge m(a_1) = m(a_2) \wedge m(a_1),$$

(1) $\alpha \wedge \beta = (-1)^{pq}\beta \wedge \alpha$ for a p-form and a q-form (Chapter T1).

so swapping any two arguments leaves the integrand, hence q_d, unchanged.

A polynomial: a symmetric multilinear form is determined by its values on the diagonal, $q_d(a) \coloneqq q_d(a,\dots,a)$, which is a homogeneous polynomial of degree d in the coordinates of a with respect to a basis of $H_2(M;\mathbb{Z}) \cong \mathbb{Z}^{b_2}$. The multilinear form is recovered by **polarisation**. For $d = 2$:

$$q_2(a+b) \overset{(1)}{=} q_2(a,a) + q_2(a,b) + q_2(b,a) + q_2(b,b) \overset{(2)}{=} q_2(a) + 2q_2(a,b) + q_2(b),$$

(1) Bilinearity.
(2) Symmetry.

so $q_2(a,b) = \frac{1}{2}\big(q_2(a+b) - q_2(a) - q_2(b)\big)$. The same trick, with more terms, works for every d. This is the same relation as between a quadratic form and its bilinear form: q_2 "looks like" an intersection form, and q_d is a degree-d generalisation.

Integer valued: the geometric definition as a signed point count makes this plausible. Behind it lie two technical facts. (i) $\mathcal{M}_k$ is orientable, and an orientation is fixed by choosing an orientation of $H^2_+(M)$ (a **homology orientation**); reversing it flips the sign of q_d. This is why Nash later says the invariants are defined "up to a sign". (ii) The classes $m(a)$ are integral after the correct normalisation of (12.89); with other normalisations one gets rational numbers with bounded denominators.

Worked example 12.35: Toy polynomials from simple moduli spaces

Suppose, purely to see how a polynomial arises, that the moduli space were a compact complex manifold on which the m-classes are combinations of a few standard 2-forms.

1. *Moduli* $= \mathbb{CP}^2$, $m(a) = \lambda(a)\,h$ with $h = L^*$ the hyperplane class and $\lambda : H_2(M) \to \mathbb{Z}$ linear. Then, by Example 12.33,

$$q_2(a) = \int_{\mathbb{CP}^2} m(a)^2 \overset{(1)}{=} \lambda(a)^2 \int_{\mathbb{CP}^2} h \wedge h \overset{(2)}{=} \lambda(a)^2.$$

   (1) Linearity: $m(a)^2 = \lambda(a)^2 h^2$.
   (2) $\int h^2 = 1$: two lines meet in one point.

   A perfect square of a linear form.

2. *Moduli* $= \mathbb{CP}^1 \times \mathbb{CP}^1$, $m(a) = \lambda_1(a)h_1 + \lambda_2(a)h_2$, with h_i the area form of the i-th factor, normalised to $\int_{\mathbb{CP}^1} h_i = 1$. Then

$$q_2(a) \overset{(1)}{=} \int \big(\lambda_1^2 h_1^2 + 2\lambda_1\lambda_2\, h_1 \wedge h_2 + \lambda_2^2 h_2^2\big) \overset{(2)}{=} 2\lambda_1(a)\lambda_2(a).$$

(1) Expand the square; the 2-forms commute.
(2) $h_i \wedge h_i = 0$ (a 4-form built from a 2-dimensional factor), and $\int h_1 \wedge h_2 = 1$ by Fubini, exactly as in Example 12.28.

The polynomial is now *indefinite*, like the hyperbolic form H.

The geometry of the moduli space, through the intersection numbers of its 2-forms, determines the shape of the polynomial. Real Donaldson polynomials are built in this way, but from moduli spaces that are (usually) far from explicit.

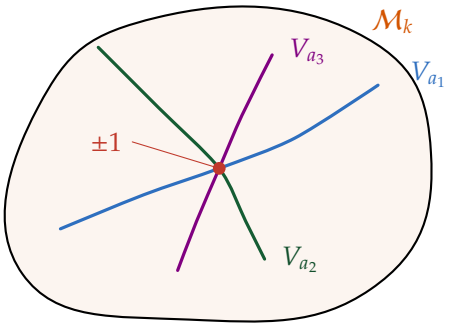


Figure 12.24: The geometric meaning of q_d. Each surface $a_i \subset M$ defines a codimension-2 "divisor" V_{a_i} in the moduli space (drawn as curves in a schematic picture). In a $2d$-dimensional $\mathcal{M}_k$, d divisors in general position meet in finitely many points, and q_d counts them with signs. The same picture with $\mathcal{M}_k = \mathbb{CP}^n$ and V's hyperplanes gives $\int h^n = 1$.

The divisors V_Σ

Donaldson constructs V_Σ concretely. Restrict each instanton A to a small tubular neighbourhood $\nu(\Sigma)$ of the surface. The restricted connections define a family of Dirac-type operators on Σ (a $\bar\partial$-operator coupled to $A|_\Sigma$), and the determinant of such a family is a complex line bundle $\mathcal{L}_\Sigma$ over the space of connections on $\nu(\Sigma)$, with $c_1(\mathcal{L}_\Sigma) = m(\Sigma)$ up to a factor. A generic section of $\mathcal{L}_\Sigma$ vanishes on a real codimension-2 set: this is V_Σ, the instantons for which the coupled Dirac operator on Σ has a zero mode. The key feature is *locality*: V_Σ is defined by what the instanton does near Σ. That locality is what makes the cutting-and-gluing arguments of Section 12.8 work, and it is the geometric counterpart of the fact that $m(\Sigma)$ in (12.93) is an integral over Σ.

12.6.5 Compactness: bubbles and the compactification $\overline{\mathcal{M}}_k$

Nash's (12.52) integrates over "a compactification $\overline{\mathcal{M}}_k$". Integrals of top-degree forms over non-compact manifolds are not topological (think of $\int_{\mathbb{R}} \mathrm{d}f = f(\infty) - f(-\infty)$). So we need to know how $\mathcal{M}_k$ fails to be compact, and what to add.

How a sequence of instantons can escape

In a sequence of connections in $\mathcal{M}_k$ the total action is fixed, $\int \frac{1}{2}|F|^2 = 8\pi^2 k$. The only way a sequence can fail to converge (after gauge transformations) is by **bubbling**: the action density piles up at finitely many points, each carrying a quantised amount $8\pi^2$ (a whole unit of instanton number), while away from these points the connections converge to an instanton of lower charge. This is **Uhlenbeck's compactness theorem**. It is the non-abelian cousin of a familiar phenomenon: a family of Gaussians of fixed area converging to a delta function.

Worked example 12.36: A BPST instanton shrinking to a point

The BPST instanton of Chapter 8 with centre 0 and scale ρ has action density

$$e_\rho(x) = \tfrac{1}{2}\sum_a |F^a|^2 = \frac{48\,\rho^4}{(|x|^2+\rho^2)^4} \qquad (x \in \mathbb{R}^4).$$

Its total is independent of ρ:

$$\int_{\mathbb{R}^4} e_\rho \,\mathrm{d}^4x \overset{(1)}{=} 48\rho^4 \cdot 2\pi^2 \int_0^\infty \frac{r^3\,\mathrm{d}r}{(r^2+\rho^2)^4} \overset{(2)}{=} 48\rho^4 \cdot \pi^2 \int_0^\infty \frac{u\,\mathrm{d}u}{(u+\rho^2)^4}$$
$$\overset{(3)}{=} 48\rho^4\pi^2 \cdot \frac{1}{6\rho^4} = 8\pi^2.$$

(1) Radial integrand; $\mathrm{vol}(S^3) = 2\pi^2$.
(2) $u = r^2$, $r^3\mathrm{d}r = \frac{1}{2}u\,\mathrm{d}u$.
(3) $\int_0^\infty u(u+c)^{-4}\mathrm{d}u = c^{-2}B(2,2) = c^{-2}/6$ with $c = \rho^2$ (substitute $u = cs$ and use $B(2,2) = \Gamma(2)^2/\Gamma(4) = 1/6$).

So e_ρ always carries exactly $8\pi^2 = 8\pi^2 \cdot 1$. The peak value is $e_\rho(0) = 48/\rho^4$, and at a fixed $x \neq 0$, $e_\rho(x) \approx 48\rho^4/|x|^8 \to 0$ as $\rho \to 0$. Hence, as distributions,

$$e_\rho \xrightarrow{\rho\to 0} 8\pi^2\,\delta^4(x).$$

The connection itself converges, away from 0, to a flat (pure gauge) connection: all the curvature has gone into the "bubble" at the origin. On a compact M the same happens in a small coordinate ball: a charge-k instanton can converge to a charge-$(k-1)$ instanton plus a delta function of weight $8\pi^2$ at a point $x \in M$.

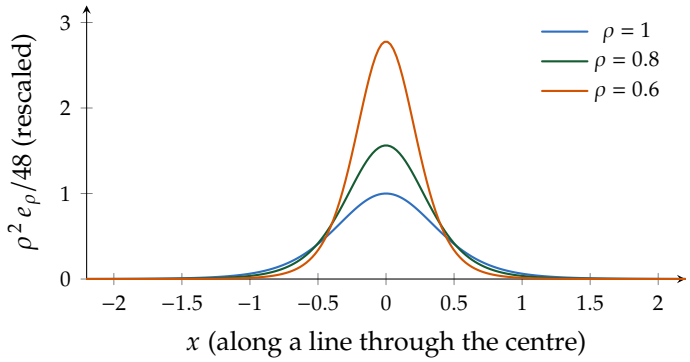


Figure 12.25: Bubbling. The action density of a BPST instanton of scale ρ, along a line through its centre (plotted as $\rho^6/(x^2+\rho^2)^4$, i.e. e_ρ times $\rho^2/48$, to keep the curves on one scale). As ρ decreases the profile narrows and its height grows like ρ^{-2} in this normalisation (ρ^{-4} unnormalised), while the four-dimensional integral stays $8\pi^2$. In the limit the density is $8\pi^2\delta^4$: a unit of instanton number concentrated at a point.

The Uhlenbeck compactification

The limits just described are "ideal instantons": a pair $(A', \{x_1, \dots, x_l\})$ with $A' \in \mathcal{M}_{k-l}$ and l points of M (possibly repeated), unordered. Adding them gives

$$\overline{\mathcal{M}}_k = \mathcal{M}_k \cup \mathcal{M}_{k-1} \times M \cup \mathcal{M}_{k-2} \times \mathrm{Sym}^2 M \cup \cdots \cup \mathcal{M}_0 \times \mathrm{Sym}^k M, \tag{12.96}$$

where $\mathrm{Sym}^l M = M^l/S_l$ is the space of unordered l-tuples. This space is compact. Its lower strata are small. Count dimensions using (12.83):

$$\dim\bigl(\mathcal{M}_{k-l}\times \mathrm{Sym}^l M\bigr) \overset{(1)}{=} 8(k-l) - 3(1+b_2^+) + 4l \\ \overset{(2)}{=} \dim \mathcal{M}_k - 4l. \tag{12.97}$$

(1) Each lost unit of charge removes 8 parameters from the moduli space and adds 4 (the position of the bubble in the 4-manifold).
(2) $\dim \mathcal{M}_k = 8k - 3(1+b_2^+)$.

So every lower stratum has codimension at least 4. (Where did the other four parameters go? A bubble on S^4 has 8 parameters: 4 for its centre, 1 for its scale and 3 for its SU(2) orientation relative to the background. In the limit only the centre survives; scale and orientation collapse.)

Key idea: Why codimension 4 is enough

A top-degree form integrated over $\overline{\mathcal{M}}_k$ does not see subsets of codimension ≥ 2 (they have measure zero), and more importantly the divisors V_{a_i} can be chosen to avoid the lower strata in the intersection count as long as those strata have codimension large enough. Donaldson shows that for k large compared with b_2^+ (the **stable range**, roughly $4k > 3(1+b_2^+)$+const) the cut-down moduli spaces $V_{a_1}\cap\cdots\cap V_{a_j}$ meet the ends in codimension at least two, and q_d is well defined and metric independent. Outside the stable range one must work harder.

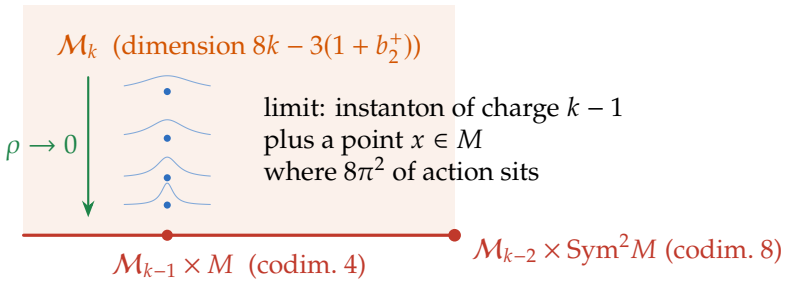


Figure 12.26: The Uhlenbeck compactification, schematically. A sequence in the top stratum $\mathcal{M}_k$ (blue dots, each with its narrowing density profile) runs to the edge as one bubble shrinks. The limit is a point of the stratum $\mathcal{M}_{k-1}\times M$ (red edge), four dimensions smaller. Two simultaneous bubbles land in a stratum of codimension 8 (red corner). The ends are thin, so a count of intersection points in the interior is not disturbed by them.

Worked example 12.37: $\mathcal{M}_1(S^4)$: the five-ball and its boundary

On S^4 with the round metric every charge-1 instanton is a BPST instanton, labelled by a centre $x_0\in\mathbb{R}^4\cup\{\infty\}=S^4$ and a scale $\rho>0$; conformal symmetry lets us trade (x_0,ρ) for a point of the open unit ball $B^5\subset\mathbb{R}^5$. The centre of the ball is the most spread out, SO(5)-symmetric instanton; points near the boundary sphere are instantons concentrated near the corresponding point of S^4. So

$$\mathcal{M}_1(S^4) = B^5, \qquad \overline{\mathcal{M}}_1(S^4) = \bar{B}^5 = B^5\cup(\mathcal{M}_0\times S^4),$$

with $\mathcal{M}_0=\{\theta\}$ the trivial connection. The boundary stratum is S^4 itself, of dimension $4 = 5-1$: here the codimension is only 1, not 4, because $\mathcal{M}_0$ is the reducible trivial connection, whose stabiliser SU(2) has absorbed the three orientation parameters. This is the "collar" $M\times(0,\epsilon)$ of Chapter 1: a manifold bounding its own moduli space. Donaldson's

diagonalisation proof used precisely this kind of boundary, while the polynomial invariants are built in the opposite regime, where the ends are thin.

Watch out: Where $\mathcal{M}_k$ sits versus where we integrate

The forms $m(a)$ are defined on $\mathcal{B}_k^*$, which contains only irreducible connections, and on $\mathcal{M}_k$. Integrating over $\overline{\mathcal{M}}_k$ requires extending them over the ideal points, where $\mathcal{B}_k^*$ gives no guidance. Donaldson's solution is the geometric one of Figure 12.24: define q_d as an intersection number of divisors V_{a_i} chosen, using their locality, to stay away from the bubbling points. So "$\int_{\overline{\mathcal{M}}_k}$" in (12.52) is best read as a count of points in the interior $\mathcal{M}_k$.

Reading Nash (pp. 333–334, (12.46)–(12.52), footnote 5)

Nash states that $q_d(M)$ is a symmetric integer polynomial of degree d on $H_2(M;\mathbb{Z})$ (12.46), built from a map $m : H_2(M) \to H^2(\mathcal{M}_k)$ (12.47). To construct m he takes the universal $\mathcal{G}$-bundle $E_{\mathcal{G}}$ over $B\mathcal{G}$, forms $c_2(E_{\mathcal{G}}) \wedge a^* \in H^6(M \times B\mathcal{G})$ (12.48), uses $\mathcal{M}_k \subset B\mathcal{G}$ (12.49), and integrates over M (12.50)–(12.51). Then q_d is the integral of $m(a_1) \wedge \cdots \wedge m(a_d)$ over a compactification $\overline{\mathcal{M}}_k$ with $d = \dim \mathcal{M}_k/2$ (12.52). Footnote 5 defines the Poincaré dual. We have supplied: Poincaré duals as bump forms with the torus and $\mathbb{CP}^2$ computed; $\mathcal{B}^* = \mathcal{A}^*/\mathcal{G}$ as a classifying space and the explicit universal bundle (with the $\pm\mathbb{1}$ issue that forces ad$P$ and p_1); the Künneth bookkeeping behind (12.50); the explicit curvature representative of $m(\Sigma)$; the Poincaré line bundle toy model; symmetry, polarisation and toy polynomials; the divisor picture; and Uhlenbeck's compactification with the BPST bubbling computation. Nash's "$E_{\mathcal{G}}$ over $B\mathcal{G}$" should be read as the bundle $\mathbb{E}$ over $\mathcal{B}^* \times M$ of (12.88): the universal bundle of the gauge group lives over $\mathcal{B}^*$, but c_2 needs the M-directions too.

12.6.6 Two theorems: algebraic surfaces and connected sums

We have defined the numbers $q_d(M)(a_1, \ldots, a_d)$. Nash is candid about them: they "are, a priori, not very easy to calculate since they require detailed knowledge of the instanton moduli space". For a general 4-manifold nobody can write down $\mathcal{M}_k$. But two *general* statements can be made without knowing $\mathcal{M}_k$ in detail, and they point in opposite directions:

- on a *complex algebraic surface*, the q_d are *non-zero* once d is large (Nash's (12.53));
- on a *connected sum* $M_1 \# M_2$ with $b_2^+(M_i) > 0$ on both sides, all the q_d *vanish* (Nash's (12.54)–(12.55)).

A manifold of the first kind and one of the second kind can have the same intersection form. Freedman then says they are homeomorphic, and Donaldson's invariants say they are not diffeomorphic. That is the promised payoff. We take the two theorems in turn and then put them together.

Complex surfaces in one page

A **complex surface** is a 4-manifold covered by charts with complex coordinates (z_1, z_2), the transition functions being holomorphic. It is **algebraic** if it can be embedded in some $\mathbb{CP}^N$ as the common zero set of homogeneous polynomials (Chapter 5). The examples to keep in mind are the smooth surfaces of degree d in $\mathbb{CP}^3$,

$$X_d = \{[z_0 : z_1 : z_2 : z_3] \in \mathbb{CP}^3 : \; z_0^d + z_1^d + z_2^d + z_3^d = 0\}.$$

They are all simply connected (Lefschetz), and their topology is fixed by d. We quote the standard formulas (they follow from the adjunction formula and the Hirzebruch signature theorem of Chapter 4):

$$\chi(X_d) = d\,(d^2 - 4d + 6), \qquad \tau(X_d) = -\tfrac{1}{3}\,d\,(d^2 - 4), \qquad p_g(X_d) = \binom{d-1}{3}. \tag{12.98}$$

Here p_g, the **geometric genus**, is the number of independent holomorphic 2-forms $f(z_1, z_2)\,\mathrm{d}z_1 \wedge \mathrm{d}z_2$ on the surface.

Worked example 12.38: The hypersurfaces X_d: b_2^+ is always odd

For simply connected M, $b_2 = \chi - 2$ and $b_2^\pm = \tfrac{1}{2}(b_2 \pm \tau)$. For the quartic $d = 4$:

$$b_2^+(X_4) \overset{(1)}{=} \tfrac{1}{2}\big(\chi - 2 + \tau\big) \overset{(2)}{=} \tfrac{1}{2}\big(4 \cdot 6 - 2 - \tfrac{1}{3} \cdot 4 \cdot 12\big) = \tfrac{1}{2}(24 - 2 - 16) = 3.$$

(1) $b_2 = \chi - 2$ and $b_2^+ = \tfrac{1}{2}(b_2 + \tau)$.

(2) (12.98) with $d = 4$: $\chi = 4(16 - 16 + 6) = 24$ and $\tau = -\tfrac{4}{3}(16 - 4) = -16$.

This is the K3 surface of Table 12.1. Doing the same for the first few degrees:

d	χ	τ	b_2^+	b_2^-	p_g	surface
1	3	1	1	0	0	$\mathbb{CP}^2$
2	4	0	1	1	0	quadric $\cong S^2 \times S^2$
3	9	−5	1	6	0	cubic $\cong \mathbb{CP}^2 \# 6\overline{\mathbb{CP}}^2$
4	24	−16	3	19	1	K3
5	55	−35	9	44	4	quintic
6	108	−64	21	85	10	sextic

In every row $b_2^+ = 1 + 2p_g$, which is odd. The cubic is the surface containing the famous 27 lines of Chapter 5. We now explain why $b_2^+ = 1 + 2p_g$ holds for every Kähler surface.

The Kähler form is self-dual

An algebraic surface inherits a metric from the Fubini–Study metric of $\mathbb{CP}^N$, and with it a **Kähler form** ω: a closed real 2-form which, in suitable local holomorphic coordinates at any given point, looks like the flat one on $\mathbb{C}^2$. So everything about how ω sits inside Λ^2 can be read off in flat $\mathbb{C}^2$.

Worked example 12.39: Self-dual and anti-self-dual forms on $\mathbb{C}^2$

Write $z_1 = x^1 + ix^2$, $z_2 = x^3 + ix^4$, and orient $\mathbb{C}^2 = \mathbb{R}^4$ by $\mathrm{vol} = \mathrm{d}x^{1234}$ (the complex orientation). Recall the self-dual basis η^1, η^2, η^3 of Example 12.30. First the Kähler form:

$$\omega := \frac{i}{2}\big(\mathrm{d}z_1 \wedge \mathrm{d}\bar{z}_1 + \mathrm{d}z_2 \wedge \mathrm{d}\bar{z}_2\big) \overset{(1)}{=} \frac{i}{2}\big(-2i\,\mathrm{d}x^{12} - 2i\,\mathrm{d}x^{34}\big) = \mathrm{d}x^{12} + \mathrm{d}x^{34} = \eta^1.$$

(1) $\mathrm{d}z \wedge \mathrm{d}\bar{z} = (\mathrm{d}x + i\mathrm{d}y) \wedge (\mathrm{d}x - i\mathrm{d}y) = -i\,\mathrm{d}x \wedge \mathrm{d}y + i\,\mathrm{d}y \wedge \mathrm{d}x = -2i\,\mathrm{d}x \wedge \mathrm{d}y$.

Next the holomorphic 2-form $\Omega = \mathrm{d}z_1 \wedge \mathrm{d}z_2$:

$$\begin{aligned}\Omega &\overset{(1)}{=} (\mathrm{d}x^1 + i\mathrm{d}x^2) \wedge (\mathrm{d}x^3 + i\mathrm{d}x^4) \\ &\overset{(2)}{=} \big(\mathrm{d}x^{13} - \mathrm{d}x^{24}\big) + i\big(\mathrm{d}x^{14} + \mathrm{d}x^{23}\big) = \eta^2 + i\,\eta^3.\end{aligned}$$

(1) Definitions of z_1, z_2.

(2) Multiply out: $i \cdot i\, dx^2 \wedge dx^4 = -dx^{24}$, and the two cross terms are $i\, dx^{14}$ and $i\, dx^{23}$.

So the three self-dual forms are exactly ω, $\operatorname{Re}\Omega$ and $\operatorname{Im}\Omega$. The anti-self-dual forms are the other three, and they are all of type $(1,1)$ and orthogonal to ω:

$$\tfrac{i}{2}\big(dz_1 \wedge d\bar{z}_1 - dz_2 \wedge d\bar{z}_2\big) \overset{(1)}{=} dx^{12} - dx^{34},$$
$$dz_1 \wedge d\bar{z}_2 \overset{(2)}{=} \big(dx^{13} + dx^{24}\big) - i\big(dx^{14} - dx^{23}\big).$$

(1) The same computation as for ω, with a relative minus sign.
(2) $(dx^1 + i dx^2) \wedge (dx^3 - i dx^4) = dx^{13} - i dx^{14} + i dx^{23} + dx^{24}$.

Check one anti-self-duality directly: $dx^{13} \wedge dx^{24} = -\text{vol}$ gives $\star dx^{13} = -dx^{24}$, so $\star(dx^{13} + dx^{24}) = -(dx^{13} + dx^{24})$. In one line:

$$\Lambda^2_+ \otimes \mathbb{C} = \mathbb{C}\,\omega \;\oplus\; \Lambda^{2,0} \oplus \Lambda^{0,2}, \qquad \Lambda^2_- \otimes \mathbb{C} = \Lambda^{1,1}_0 \;\text{ (the } (1,1)\text{-forms orthogonal to } \omega).$$

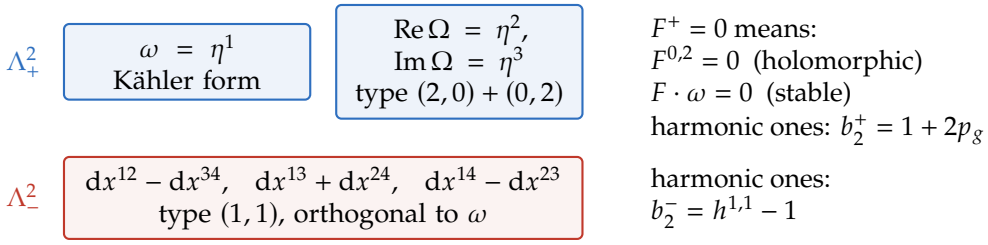


Figure 12.27: How the complex structure of $\mathbb{C}^2$ splits the six 2-forms. The three self-dual forms (blue) are the Kähler form and the real and imaginary parts of $dz_1 \wedge dz_2$; the three anti-self-dual forms (red) are the $(1,1)$-forms perpendicular to ω. Read off: an ASD curvature has no $(0,2)$ part and no ω-component, and the self-dual harmonic forms are ω plus the holomorphic 2-forms.

Two consequences follow at once from Example 12.39.

First, $b_2^+ = 1 + 2p_g$. The self-dual harmonic forms on a compact Kähler surface are spanned by the (harmonic) Kähler form and by the real and imaginary parts of the holomorphic 2-forms, which are automatically harmonic. Counting real dimensions,

$$b_2^+ \overset{(1)}{=} \dim_{\mathbb{R}} \mathcal{H}^2_+ \overset{(2)}{=} 1 + \dim_{\mathbb{R}} \big(\mathcal{H}^{2,0} \oplus \mathcal{H}^{0,2}\big)_{\mathbb{R}} \overset{(3)}{=} 1 + 2p_g. \tag{12.99}$$

(1) b_2^+ is the dimension of the self-dual harmonic 2-forms (Section 12.6.1).
(2) The pointwise splitting of Λ^2_+ in Example 12.39 passes to harmonic forms: the ω-direction contributes only multiples of ω itself, and the other two directions contribute the real forms $\operatorname{Re}\alpha$, $\operatorname{Im}\alpha$ for holomorphic 2-forms α.
(3) The holomorphic 2-forms form a complex vector space of dimension p_g, which is $2p_g$ real dimensions.

So b_2^+ is odd, and Nash's parity requirement is met by every algebraic surface.

Second, ASD connections are holomorphic bundles. Split the curvature by type, $F = F^{2,0} + F^{1,1} + F^{0,2}$. For an SU(2) connection $F^{2,0} = -(F^{0,2})^\dagger$, so the reality of F ties the two. Then

$$F = -\star F \quad\Longleftrightarrow\quad F^{0,2} = 0 \;\text{ and }\; \omega \cdot F^{1,1} = 0, \tag{12.100}$$

because F^+ has exactly the components along ω, $\Lambda^{2,0}$ and $\Lambda^{0,2}$. Now $F^{0,2}$ is the square of the $(0,1)$-part of the covariant derivative: $F^{0,2} = (\bar{\partial}_A)^2$. An operator $\bar{\partial}_A$ with $\bar{\partial}_A^2 = 0$ is precisely what it takes to define holomorphic sections, so it makes E a *holomorphic* vector bundle. The second condition in (12.100), the "Hermitian–Yang–Mills" condition, turns out to be equivalent

to an algebraic condition called **stability**. This is the theorem of Donaldson (for surfaces) and Uhlenbeck–Yau (in general), sometimes called the Kobayashi–Hitchin correspondence:

$$\mathcal{M}_k(M, g_{\text{Kähler}}) \;\cong\; \big\{\text{stable holomorphic rank-2 bundles } E \to M \\ \text{with } c_1(E) = 0,\; c_2(E) = k\big\}\big/\text{iso}.$$

The right side is a problem in algebraic geometry, and algebraic geometers know a great deal about it. For example, it is a complex (quasi-projective) variety, and its complex dimension can be computed by the Riemann–Roch theorem.

Worked example 12.40: Riemann–Roch reproduces $\dim \mathcal{M}_k$

Deformations of a holomorphic bundle E are measured by the sheaf cohomology of the trace-free endomorphisms $\operatorname{End}_0 E$ (Chapter 3): $H^1(\operatorname{End}_0 E)$ is the tangent space and $H^2(\operatorname{End}_0 E)$ the obstructions, while $H^0(\operatorname{End}_0 E) = 0$ for a stable bundle. Riemann–Roch for a rank-3 bundle $V = \operatorname{End}_0 E$ with $c_1(V) = 0$ on a surface gives $\chi(V) = 3\,\chi(\mathcal{O}_M) - c_2(V)$, and $c_2(\operatorname{End}_0 E) = 4c_2(E)$ by (12.82). Hence

$$\begin{aligned}\dim_{\mathbb{R}} \mathcal{M}_k &\overset{(1)}{=} 2\big(h^1(\operatorname{End}_0 E) - h^2(\operatorname{End}_0 E)\big) \overset{(2)}{=} -2\,\chi(\operatorname{End}_0 E)\\ &\overset{(3)}{=} 2\big(4k - 3\,\chi(\mathcal{O}_M)\big) \overset{(4)}{=} 8k - 6\,(1 + p_g)\\ &\overset{(5)}{=} 8k - 3\,(1 + b_2^+).\end{aligned}$$

(1) The expected complex dimension is $h^1 - h^2$; real dimension is twice that.
(2) $\chi = h^0 - h^1 + h^2$ and $h^0 = 0$.
(3) Riemann–Roch as quoted, with $c_2(E) = k$.
(4) $\chi(\mathcal{O}_M) = 1 - q + p_g$ with $q = \dim H^1(\mathcal{O}_M) = \tfrac{1}{2}b_1 = 0$ for simply connected M.
(5) $b_2^+ = 1 + 2p_g$ from (12.99).

This is Nash's (12.45), obtained this time by algebraic geometry rather than by the Atiyah–Singer index theorem. The agreement is no accident: Riemann–Roch is the index theorem for the Dolbeault complex (Chapter 4).

Why the invariants are non-zero: positivity

Here is the mechanism behind Nash's (12.53). It rests on one elementary fact.

Key idea: Complex things intersect positively

If two complex submanifolds of a complex manifold meet transversally, every intersection point counts $+1$, never -1. If V is a compact complex submanifold of complex dimension n in a Kähler manifold, then $\int_V \omega^n > 0$. So a count of intersections of complex objects cannot cancel.

Worked example 12.41: Complex linear maps preserve orientation

The sign of a transverse intersection point is the sign of a real determinant: the determinant of the real matrix whose columns are the oriented tangent vectors of the two submanifolds. For complex submanifolds these columns come in complex pairs, and the matrix is the real form of a complex matrix. Take first a 1×1 complex matrix, multiplication by $c = a + ib$ on

$\mathbb{C} = \mathbb{R}^2$:

$$\det_{\mathbb{R}}(c) \overset{(1)}{=} \det \begin{pmatrix} a & -b \\ b & a \end{pmatrix} = a^2 + b^2 \overset{(2)}{=} |c|^2 \ \geq 0.$$

(1) $(a+ib)(x+iy) = (ax-by)+i(bx+ay)$, so in the basis $(1, i)$ the real matrix is as shown.
(2) Definition of the modulus.

For an $n \times n$ complex matrix C, bring it to upper triangular form by a complex change of basis (which changes $\det_{\mathbb{R}}$ by $|\det_{\mathbb{C}}|^2$ of the change, a positive number). The real form of a triangular matrix is block triangular with 2×2 diagonal blocks of the kind above, so

$$\det_{\mathbb{R}} C \overset{(1)}{=} \prod_{j=1}^{n} \det_{\mathbb{R}}(\lambda_j) \overset{(2)}{=} \prod_{j=1}^{n} |\lambda_j|^2 = |\det_{\mathbb{C}} C|^2 \ \geq 0.$$

(1) The determinant of a block triangular matrix is the product of the determinants of the diagonal blocks; λ_j are the diagonal entries (eigenvalues).
(2) The 1×1 case, applied to each block.

For the two coordinate lines $\{z_2 = 0\}$ and $\{z_1 = 0\}$ in $\mathbb{C}^2$ the tangent vectors $(\partial_{x^1}, \partial_{x^2})$ followed by $(\partial_{x^3}, \partial_{x^4})$ give the identity matrix: sign $+1$. Any other transverse pair of complex lines is obtained by a complex linear map, so has sign $+1$ as well. This is why $L \cdot L = +1$ in $\mathbb{CP}^2$ (Example 12.29), and why a degree-n curve meets a line in exactly n points counted positively.

With this fact the non-vanishing argument runs as follows.

1. Choose the Kähler metric. By the Kobayashi–Hitchin correspondence, $\mathcal{M}_k$ is a complex variety.
2. Take $a = [H]$, the class of a hyperplane section $H = M \cap \mathbb{CP}^{N-1}$ (a complex curve in M). Donaldson shows that the divisor $V_H \subset \mathcal{M}_k$ of Figure 12.24 can be chosen to be a *complex* hypersurface, and that for k large $m(H)$ is represented by a positive $(1,1)$-form on $\mathcal{M}_k$, playing the role of a Kähler form on the moduli space.
3. For k large the moduli space is non-empty and has the expected dimension $2d$ (Gieseker, Taubes).
4. Then $q_d(H, \dots, H) = \int_{\mathcal{M}_k} m(H)^d$ is, up to a positive factor $d!$, the volume of $\mathcal{M}_k$ in this Kähler-like form: a sum of $+1$'s, one for each point of $V_{H_1} \cap \cdots \cap V_{H_d}$. It is strictly positive.

Key idea: Donaldson's non-vanishing theorem (Nash's (12.53))

If M is a simply connected complex algebraic surface, there is a d_0 such that

$$q_d(M) \neq 0 \qquad \text{for all } d \geq d_0,$$

and in fact $q_d(M)(H, \dots, H) > 0$ for the hyperplane class H. The mechanism is positivity: intersections of complex divisors in a complex moduli space all count $+1$.

Physical picture: A toy version of positivity

On $\mathbb{CP}^n$ with hyperplane class h, $\int_{\mathbb{CP}^n} h^n = 1$: n hyperplanes meet in one point, and it counts $+1$. On any complex projective variety $V \subset \mathbb{CP}^N$ of dimension n, $\int_V h^n = \deg V$, the

number of points in which V meets a generic linear subspace of complementary dimension, all counted +1. The Donaldson invariant of an algebraic surface is essentially the "degree" of its instanton moduli space.

The wedge computation in Example 12.39 also shows the volume form of the moduli-space picture in miniature. On $\mathbb{C}^2$,

$$\tfrac{1}{2}\,\omega\wedge\omega \overset{(1)}{=} \tfrac{1}{2}\big(\mathrm{d}x^{12}+\mathrm{d}x^{34}\big)\wedge\big(\mathrm{d}x^{12}+\mathrm{d}x^{34}\big) \overset{(2)}{=} \tfrac{1}{2}\cdot 2\,\mathrm{d}x^{12}\wedge\mathrm{d}x^{34} = \mathrm{vol}\ > 0,$$

(1) $\omega = \mathrm{d}x^{12}+\mathrm{d}x^{34}$.
(2) $\mathrm{d}x^{12}\wedge\mathrm{d}x^{12}=0$, and the two cross terms are equal because 2-forms commute.

so $\omega^n/n!$ is the Riemannian volume form on a Kähler manifold of complex dimension n (here $n=2$), and its integral over any complex submanifold is that submanifold's volume: positive.

Connected sums: the vanishing theorem

Now the opposite statement. Let $M = M_1\#M_2$ (Nash's (12.54); the connected sum was described in Figure 12.20). The two pieces are joined by a tube whose cross-section is S^3. Since we may choose any metric, choose one in which the tube is a very long cylinder $S^3\times[-T,T]$. What do instantons look like as $T\to\infty$?

Step 1: the energy splits. An instanton has finite action $8\pi^2 k$. On a very long cylinder most of the slices $S^3\times\{t\}$ must carry almost no curvature, otherwise the action would grow with T. So in the middle of the neck the connection is almost flat. The only flat SU(2) connection on S^3 is the trivial one ($\pi_1(S^3)=1$; see Example 12.10). The instanton therefore splits into an instanton A_1 of charge k_1 concentrated in M_1 and an instanton A_2 of charge k_2 concentrated in M_2, with $k_1+k_2=k$ and both asymptotically trivial on the neck.

Step 2: gluing needs a parameter. Conversely, given A_1 and A_2 we can glue them back into an approximate instanton on M and then correct it to an exact one (Taubes' gluing). But to glue we must identify the fibre of P_1 over the neck with the fibre of P_2. Both connections are trivial there, so any constant SU(2) matrix u will do. Replacing u by $-u$ changes nothing, since $-\mathbb{1}$ is central and acts trivially on connections. So the gluing data is a point of

$$\mathrm{SU}(2)/\{\pm\mathbb{1}\} = \mathrm{SO}(3), \qquad \dim \mathrm{SO}(3) = 3.$$

Physically, u is the relative colour orientation of the two instantons. The same three parameters appeared as the "orientations" of a BPST instanton in Example 12.37.

Step 3: the dimensions add up, with three to spare. For irreducible A_1, A_2 we get, near the long-neck limit,

$$\mathcal{M}_k(M_1\#M_2) \;\approx\; \mathcal{M}_{k_1}(M_1)\times\mathcal{M}_{k_2}(M_2)\times\mathrm{SO}(3). \tag{12.101}$$

Check the dimension with (12.83):

$$\begin{aligned}\dim\mathcal{M}_{k_1}(M_1)+\dim\mathcal{M}_{k_2}(M_2)+3 &\overset{(1)}{=} \big(8k_1-3-3b_2^+(M_1)\big)+\big(8k_2-3-3b_2^+(M_2)\big)+3\\ &\overset{(2)}{=} 8k-3-3\big(b_2^+(M_1)+b_2^+(M_2)\big)\\ &\overset{(3)}{=} 8k-3\big(1+b_2^+(M)\big) \overset{(4)}{=} \dim\mathcal{M}_k(M).\end{aligned}$$

(1) (12.83) on each piece.
(2) $k_1+k_2=k$, and $-3-3+3=-3$.
(3) b_2^+ is additive under connected sum (Section 12.6.1).

(4) (12.83) on M.

The count closes. The picture (12.101) is therefore consistent, and it is the whole moduli space near the long-neck limit.

Step 4: cutting down. By multilinearity it is enough to evaluate q_d on classes each of which lies on one side, say $a_1, \dots, a_r \in H_2(M_1)$ and $a_{r+1}, \dots, a_d \in H_2(M_2)$. The divisors V_{a_i} are *local* (Section 12.6.4): V_{a_i} for $a_i \subset M_1$ only looks at the connection near a_i, i.e. at A_1. So in (12.101) the cut-down moduli space is

$$\Big(\mathcal{M}_{k_1}(M_1) \cap V_{a_1} \cap \cdots \cap V_{a_r}\Big) \times \Big(\mathcal{M}_{k_2}(M_2) \cap V_{a_{r+1}} \cap \cdots \cap V_{a_d}\Big) \times \mathrm{SO}(3).$$

Write $D_i = \dim \mathcal{M}_{k_i}(M_i)$. The three factors have dimensions $D_1 - 2r$, $D_2 - 2(d-r)$ and 3, and they must add up to 0 for a point count:

$$(D_1 - 2r) + \big(D_2 - 2(d-r)\big) \overset{(1)}{=} (D_1 + D_2) - 2d \overset{(2)}{=} (2d-3) - 2d = -3.$$

(1) Collect terms.
(2) Step 3: $D_1 + D_2 + 3 = \dim \mathcal{M}_k(M) = 2d$.

Two integers summing to -3 cannot both be ≥ 0. So one of the two cut-down factors has negative expected dimension, which means it is empty for generic choices. The cut-down moduli space of M is empty, and the count is zero. Put differently: the three SO(3) directions can never be cut down by divisors coming from either side, because every V_{a_i} is pulled back from one of the factors, and a top-degree form pulled back from a space of smaller dimension vanishes.

Where $b_2^+ > 0$ is used. Step 3 assumed A_1 and A_2 irreducible. The dangerous case is $k_2 = 0$: then A_2 is the trivial connection, whose stabiliser is all of SO(3), so the gluing parameter is used up by the stabiliser and no spare dimensions remain. However, the trivial connection on M_2 has obstruction space $H^2_A = \mathcal{H}^2_+(M_2) \otimes \mathfrak{su}(2)$, of dimension $3b_2^+(M_2)$. When $b_2^+(M_2) > 0$ these obstructions cut the glued family down by $3b_2^+(M_2)$ dimensions, below what is needed, and again nothing survives. When $b_2^+(M_2) = 0$, as for $M_2 = \overline{\mathbb{CP}}^2$, there are no obstructions and the argument fails, correctly so, as the next pitfall explains. The same caution applies to reducible A_1 or A_2 with $k_i > 0$, which are excluded by $b_2^+(M_i) > 0$ and a generic metric (Section 12.6.2).

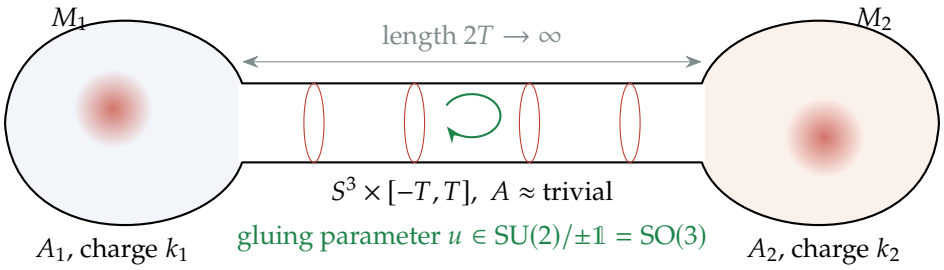


Figure 12.28: Stretching the neck of a connected sum. The instanton's action cannot live on the long neck, where the connection becomes flat, hence trivial (S^3 is simply connected). What remains are an instanton A_1 in M_1 and A_2 in M_2, glued with a relative colour rotation u. These three extra parameters can never be cut down by surfaces from either side, so every Donaldson invariant vanishes.

Key idea: Donaldson's vanishing theorem (Nash's (12.55))

If $M = M_1 \# M_2$ with $b_2^+(M_1) > 0$ and $b_2^+(M_2) > 0$, then

$$q_d(M) = 0 \qquad \text{for all } d.$$

The mechanism: near a long neck the moduli space is (moduli of M_1) $\times$ (moduli of M_2) $\times$SO(3), and the three gluing directions cannot be cut down by any product of m-classes.

Watch out: The condition $b_2^+ > 0$ on *both* sides is essential

Blowing up a point of a complex surface X gives $X\#\overline{\mathbb{CP}}^2$, which is again an algebraic surface. By the non-vanishing theorem its invariants are non-zero, although it is a connected sum. There is no contradiction: $b_2^+(\overline{\mathbb{CP}}^2) = 0$. Indeed the invariants of $X\#\overline{\mathbb{CP}}^2$ are computed from those of X by a "blow-up formula". The exceptional sphere e with $e \cdot e = -1$ supports reducible instantons on the $\overline{\mathbb{CP}}^2$ side (Figure 12.21), and these contribute.

Worked example 12.42: The dimension bookkeeping on a concrete sum

Take $M = \mathbb{CP}^2\#(4\mathbb{CP}^2\#29\overline{\mathbb{CP}}^2)$, so $M_1 = \mathbb{CP}^2$ with $b_2^+ = 1$ and $M_2 = 4\mathbb{CP}^2\#29\overline{\mathbb{CP}}^2$ with $b_2^+ = 4$. Then $b_2^+(M) = 5$ and $\dim \mathcal{M}_k(M) = 8k - 18$. For $k = 3$: $\dim \mathcal{M}_3(M) = 6$ and $d = 3$. The possible splittings $k = k_1 + k_2$ with both pieces irreducible include $(k_1, k_2) = (1, 2)$:

$$\dim \mathcal{M}_1(\mathbb{CP}^2) + \dim \mathcal{M}_2(M_2) + 3 \overset{(1)}{=} (8 - 6) + (16 - 15) + 3 = 2 + 1 + 3 = 6.$$

(1) $\dim \mathcal{M}_k = 8k - 3(1 + b_2^+)$ with $b_2^+ = 1$ and $b_2^+ = 4$ respectively.

Three divisors must cut the six dimensions down to points. With r of them from $\mathbb{CP}^2$ and $3 - r$ from M_2, the first factor has dimension $2 - 2r$ and the second $1 - 2(3 - r) = 2r - 5$; their sum is -3 for every r. For $r = 0$: $(2, -5)$; $r = 1$: $(0, -3)$; $r = 2$: $(-2, -1)$; $r = 3$: $(-4, 1)$. Each row has a negative entry, so the cut-down space is empty. (Note that the moduli space of M_2 has odd dimension here: nothing forbids that for the pieces, only for M.)

Putting the two theorems together: exotic smooth structures

Now the punchline, which Nash states at the end of §3 ("they have the potential to distinguish homeomorphic manifolds which have distinct diffeomorphic structures"). We need a manifold of each kind with the same intersection form.

Worked example 12.43: The elliptic surface $E(3)$ and $5\mathbb{CP}^2\#29\overline{\mathbb{CP}}^2$

The **elliptic surfaces** $E(n)$ are simply connected algebraic surfaces fibred over $\mathbb{CP}^1$ with generic fibre a torus; $E(1) = \mathbb{CP}^2\#9\overline{\mathbb{CP}}^2$ and $E(2) = \text{K3}$. Their Chern numbers are $c_1^2 = 0$ and $c_2 = \chi = 12n$. The signature theorem for surfaces gives $\tau = \frac{1}{3}(c_1^2 - 2c_2)$. So

$$\begin{aligned}
\tau(E(n)) &\overset{(1)}{=} \tfrac{1}{3}(0 - 24n) = -8n, \\
b_2^+(E(n)) &\overset{(2)}{=} \tfrac{1}{2}\big(\chi - 2 + \tau\big) = \tfrac{1}{2}(12n - 2 - 8n) = 2n - 1, \\
b_2^-(E(n)) &\overset{(3)}{=} \tfrac{1}{2}\big(\chi - 2 - \tau\big) = \tfrac{1}{2}(12n - 2 + 8n) = 10n - 1.
\end{aligned}$$

(1) Hirzebruch's signature theorem written in Chern numbers.
(2) $b_2 = \chi - 2$ and $b_2^+ = \frac{1}{2}(b_2 + \tau)$.
(3) $b_2^- = \frac{1}{2}(b_2 - \tau)$.

For $n = 3$: $b_2^+ = 5$, $b_2^- = 29$, and the form is odd (the $E(n)$ with n odd are not spin). An odd indefinite unimodular form is determined by its rank and signature and is diagonal, so

$$Q_{E(3)} \cong 5\,(+1) \oplus 29\,(-1) = Q_{5\mathbb{CP}^2\#29\overline{\mathbb{CP}}^2}.$$

Now compare:

- $E(3)$ is algebraic: $q_d(E(3)) \neq 0$ for large d (12.53).
- $X = 5\mathbb{CP}^2\#29\overline{\mathbb{CP}}^2 = \mathbb{CP}^2\#(4\mathbb{CP}^2\#29\overline{\mathbb{CP}}^2)$ is a connected sum of two pieces with $b_2^+ = 1$ and $b_2^+ = 4$, both positive: $q_d(X) = 0$ for all d (12.55).

Freedman's theorem says $E(3)$ and X are homeomorphic (same odd form, and both have the same $\mathbb{Z}_2$ invariant, since both are smooth). If they were diffeomorphic, the diffeomorphism would carry $q_d(E(3))$ to $\pm q_d(X) = 0$. It cannot. So $E(3)$ and X are homeomorphic but *not* diffeomorphic: the topological manifold underlying X carries at least two smooth structures.

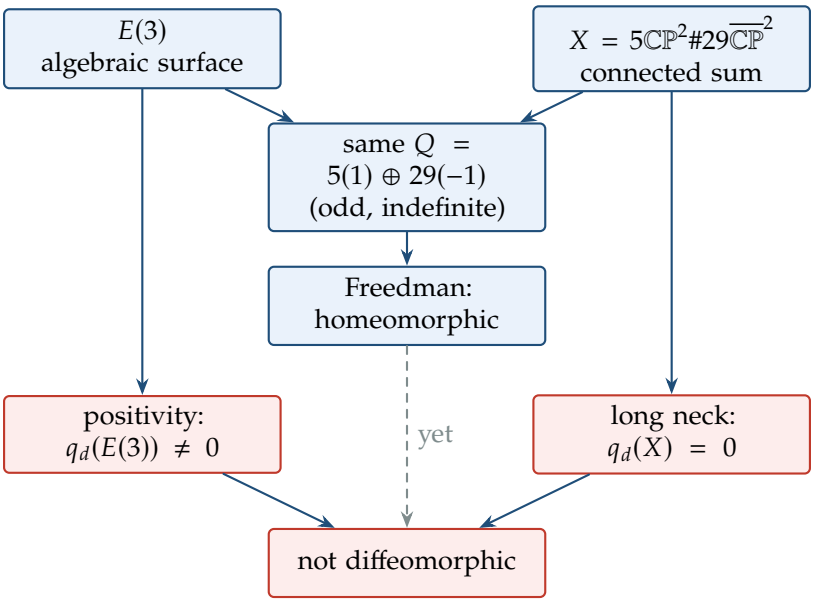


Figure 12.29: The logic of an exotic pair. The top half is topology: equal intersection forms force a homeomorphism. The bottom half is gauge theory: the Donaldson invariants, which are diffeomorphism invariants, are non-zero on one side and zero on the other.

Physical picture: What the invariants "see"

The vanishing theorem says that q_d detects whether a manifold can be pulled apart along an S^3 into two pieces that each carry a self-dual harmonic form. The non-vanishing theorem says that complex surfaces cannot be pulled apart in this way, although their intersection forms would allow it. The invariants measure a kind of smooth "irreducibility" that the intersection form, and hence the topology, cannot see. The same argument gives many more exotic pairs, e.g. $E(n)$ against $(2n-1)\mathbb{CP}^2\#(10n-1)\overline{\mathbb{CP}}^2$ for every odd $n \geq 3$; Nash

cites Ebeling for further examples.

Try it: Why not the K3 surface?

K3 has $Q = 3H \oplus 2(-E_8)$, even. Is there a connected sum of standard pieces with the same form, to play the role of X?
Answer: the candidate would be $3(S^2 \times S^2)\#2$ (a manifold with form $-E_8$), but by Donaldson's diagonalisation theorem (Chapter 1) there is no smooth closed simply connected manifold with definite form $-E_8$. So K3 has no "obvious" connected-sum rival, and the exotic pair argument is run instead with $E(n)$, n odd, where the diagonal rival $(2n-1)\mathbb{CP}^2\#(10n-1)\overline{\mathbb{CP}}^2$ exists.

Reading Nash (pp. 334–335, (12.53)–(12.55), footnote 6)

Nash remarks that the q_d are hard to compute because they need detailed knowledge of $\mathcal{M}_k$; that for a complex algebraic surface "a positivity argument shows" $q_d \neq 0$ for $d \geq d_0$ (12.53); and that if $M = M_1\#M_2$ (12.54) with $b_2^+(M_i) > 0$ for both, then $q_d(M) = 0$ for all d (12.55). Footnote 6 defines the connected sum (remove a disc from each manifold and sew along the new boundaries). He gives no argument for either statement. We have supplied: the hypersurfaces X_d and their Betti numbers; the self-duality of the Kähler form and the splitting of Λ^2 on $\mathbb{C}^2$; $b_2^+ = 1 + 2p_g$; the holomorphic reading of the ASD equation and the Riemann–Roch check of (12.45); positivity of complex intersections; the neck-stretching argument with its SO(3) gluing parameter and dimension count; the role of $b_2^+ > 0$ and the blow-up caveat; and the exotic pair $E(3)$ versus $5\mathbb{CP}^2\#29\overline{\mathbb{CP}}^2$ that realises Nash's closing remark on p. 339.

Recap

- The stage is a smooth, simply connected, closed, oriented 4-manifold. Its topology is the unimodular intersection form Q on $H_2(M;\mathbb{Z})$; $b_2^\pm$ are the dimensions of self-dual and anti-self-dual harmonic 2-forms.
- For SU(2), $\dim \mathcal{M}_k = 8k - 3(1 + b_2^+)$: eight parameters per unit charge, minus three for each of $1 + b_2^+$. It is even exactly when b_2^+ is odd. Generic metrics avoid reducible instantons when $b_2^+ \geq 1$, and paths of metrics avoid them when $b_2^+ \geq 2$.
- Surfaces $\Sigma \subset M$ give 2-forms $m(\Sigma)$ on $\mathcal{M}_k$ via Poincaré duality, the universal bundle and integration over M. Geometrically, $m(\Sigma)$ is dual to a divisor V_Σ of instantons with a zero mode on Σ.
- $q_d(a_1, \dots, a_d) = \int_{\mathcal{M}_k} m(a_1) \wedge \cdots \wedge m(a_d)$ with $2d = \dim \mathcal{M}_k$, a symmetric integer polynomial on $H_2(M)$. It counts the points of $V_{a_1} \cap \cdots \cap V_{a_d}$. The integral makes sense because bubbling happens in high codimension.
- Algebraic surfaces: positivity of complex intersections makes $q_d \neq 0$ for large d. Connected sums with $b_2^+ > 0$ on both sides: the SO(3) gluing parameter makes all $q_d = 0$.
- Together with Freedman's theorem this gives homeomorphic but not diffeomorphic pairs, e.g. $E(3)$ and $5\mathbb{CP}^2\#29\overline{\mathbb{CP}}^2$.

12.7 Witten's topological field theory

Donaldson's invariants are integrals over moduli spaces of instantons. Instantons are what dominate a Euclidean path integral at weak coupling. So it is natural to ask: *is there a quantum field theory whose correlation functions are exactly Donaldson's integrals?* Atiyah posed the question in 1987, with a hint: Floer's supersymmetric quantum mechanics on $\mathcal{A}/\mathcal{G}$ (Section 12.5.5) should be the Hamiltonian version of a four-dimensional theory on $M^3 \times \mathbb{R}$. Witten answered it in 1988. His theory is Nash's "relativistic generalisation" of Floer's Hamiltonian, and it is the prototype of a Witten-type (cohomological) TQFT (Section 12.1.2).

The logic of this section is the one of Chapter 7, where Witten's supersymmetric quantum mechanics computed Morse homology. Four facts do all the work:

1. there is a fermionic symmetry Q with $Q^2 = 0$ (up to gauge);
2. the action and the energy-momentum tensor are Q-exact;
3. hence the correlation functions of Q-invariant observables do not depend on the metric or on the coupling;
4. hence they can be computed at infinitesimal coupling, where the path integral collapses onto the classical minima: the instantons.

We first meet the fields, then establish 1–4, then do the whole computation in a zero-dimensional toy model, and finally identify the observables that produce q_d.

Watch out: Conventions in this section

Nash's (12.56) is written in the physicists' way: the trace is positive-definite on the fields and factors of i sit in front of the fermion terms. Reading it with our anti-Hermitian fields, replace tr by $-\operatorname{tr}$ (i.e. use the positive inner product $\langle X, Y\rangle = -\operatorname{tr} XY$ of (12.37)). None of the arguments below depends on these normalisations: they only use which terms are present and how Q acts. Instantons are anti-self-dual, $F^+ = 0$, as in the rest of this part.

12.7.1 The fields and the action

The field content

All fields take values in the Lie algebra $\mathfrak{su}(2)$, i.e. they are sections of $\mathrm{ad}P$. There are three bosons and three fermions:

field	type	form degree	statistics	U	geometric role
A_μ	gauge field	1	boson	0	point of $\mathcal{A}$
ψ_μ	fermion	1	fermion	+1	tangent vector to $\mathcal{A}$
ϕ	scalar	0	boson	+2	curvature of $\mathcal{A} \to \mathcal{A}/\mathcal{G}$
λ	scalar	0	boson	−2	partner of ϕ
η	scalar	0	fermion	−1	gauge directions
$\chi_{\mu\nu}$	self-dual 2-form	2, $\chi = \star\chi$	fermion	−1	the equation $F^+ = 0$

ϕ and λ are "spinless" (Nash's word): they are scalars, not spinors, even though λ comes from a gaugino of the supersymmetric theory we start from. The fermions $(\eta, \psi_\mu, \chi_{\mu\nu})$ are a 0-form, a 1-form and a self-dual 2-form. Nash's key observation is that this is exactly the

anti-instanton deformation complex (8.104), which in our ASD conventions is (12.80):

$$\underbrace{\Omega^0(\mathrm{ad}P)}_{\eta} \xrightarrow{\mathrm{d}_A} \underbrace{\Omega^1(\mathrm{ad}P)}_{\psi_\mu} \xrightarrow{\mathrm{d}_A^+} \underbrace{\Omega^2_+(\mathrm{ad}P)}_{\chi_{\mu\nu}}.$$

So the fermions are a copy of the three spaces that control the moduli space: ψ lives where the deformations live, χ where the equation lives, and η where the gauge transformations live. This is not a coincidence; it is how the theory was designed.

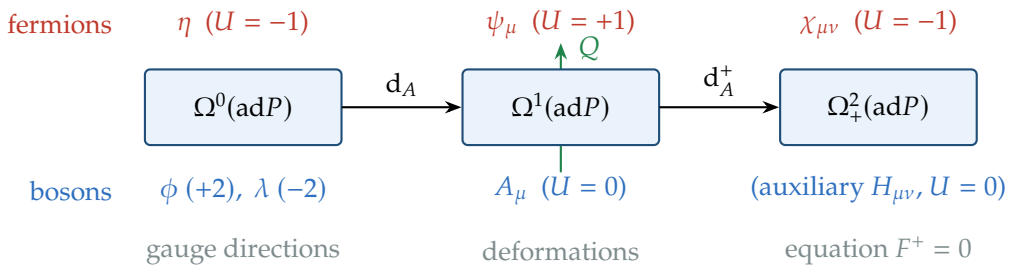


Figure 12.30: The fields of Witten's theory sit on top of the instanton deformation complex. Each fermion (red) lives in one of the three spaces; Q maps the gauge field A_μ to its fermionic partner ψ_μ, a tangent vector to $\mathcal{A}$. The zero modes of ψ are the tangent vectors to $\mathcal{M}_k$; the would-be zero modes of η and χ are H^0_A and H^2_A, which vanish for irreducible, unobstructed instantons.

Worked example 12.44: Where the fields come from: twisting $N=2$ super Yang–Mills

Witten obtained the theory by "twisting" the $N=2$ supersymmetric Yang–Mills theory in four Euclidean dimensions. The rotation group is $\mathrm{SO}(4)=\big(\mathrm{SU}(2)_L\times\mathrm{SU}(2)_R\big)/\mathbb{Z}_2$, and representations are labelled by two spins (j_L,j_R): a vector is $(\frac12,\frac12)$ and the two kinds of 2-forms $\Lambda^2_\pm$ are $(1,0)$ and $(0,1)$. The $N=2$ theory also has an internal symmetry $\mathrm{SU}(2)_I$ rotating its two supercharges, and a $\mathrm{U}(1)_R$ symmetry. Its fields, with their (j_L,j_R,j_I):

$$A_\mu:(\tfrac12,\tfrac12,0),\qquad \lambda^A_\alpha:(\tfrac12,0,\tfrac12),\qquad \bar\lambda_{A\dot\alpha}:(0,\tfrac12,\tfrac12),\qquad B,\bar B:(0,0,0),$$

with B a complex scalar of $\mathrm{U}(1)_R$ charge ±2. The **twist** declares the new rotation group to be $\mathrm{SU}(2)_L\times\mathrm{SU}(2)'_R$ with $\mathrm{SU}(2)'_R$ the diagonal subgroup of $\mathrm{SU}(2)_R\times\mathrm{SU}(2)_I$. So we simply combine j_R and j_I with the Clebsch–Gordan rule $\frac12\otimes\frac12=0\oplus1$, $0\otimes\frac12=\frac12$:

$$\begin{aligned}\lambda:(\tfrac12,\,0\otimes\tfrac12)&\overset{(1)}{=}(\tfrac12,\tfrac12)\;\longrightarrow\;\psi_\mu,\\ \bar\lambda:(0,\,\tfrac12\otimes\tfrac12)&\overset{(2)}{=}(0,0)\oplus(0,1)\;\longrightarrow\;\eta\oplus\chi_{\mu\nu},\\ B,\bar B:(0,0\otimes0)&=(0,0)\;\longrightarrow\;\phi,\lambda.\end{aligned}$$

(1) Spin 0 times spin $\frac12$ is spin $\frac12$; $(\frac12,\frac12)$ is the vector representation.
(2) $\frac12\otimes\frac12=0\oplus1$: a scalar and a 2-form of one chirality (which we call self-dual).

Counting components per colour: λ and $\bar\lambda$ have $2\times2+2\times2=8$ real fermionic components after the Euclidean reality conditions, and ψ,η,χ have $4+1+3=8$. The $\mathrm{U}(1)_R$ charge becomes Nash's U-number. The supercharges transform like the gauginos, so one of them, the $(0,0)$ piece of $\bar Q_{A\dot\alpha}$, becomes a *scalar* after the twist. That scalar supercharge is Q. Because it is a scalar, it survives on any curved M: no covariantly constant spinors are needed.

The action

Nash's (12.56) is

$$S = \int_M d^4x\sqrt{g}\ \mathrm{tr}\Big\{\tfrac{1}{4}F_{\mu\nu}F^{\mu\nu} + \tfrac{1}{4}F_{\mu\nu}{}^*F^{\mu\nu} + \tfrac{1}{2}\phi D_\mu D^\mu\lambda + iD_\mu\psi_\nu\chi^{\mu\nu} - i\eta D_\mu\psi^\mu \\ - \tfrac{i}{8}\phi[\chi_{\mu\nu},\chi^{\mu\nu}] - \tfrac{i}{2}\lambda[\psi_\mu,\psi^\mu] - \tfrac{i}{2}\phi[\eta,\eta] - \tfrac{1}{8}[\phi,\lambda]^2\Big\}, \tag{12.102}$$

with ${}^*F_{\mu\nu} = \frac{1}{2}\varepsilon_{\mu\nu\rho\sigma}F^{\rho\sigma}$. Let us read it term by term.

The gauge terms. The first two terms together are the square of the self-dual part of F. Split $F = F^+ + F^-$ with ${}^*F^\pm = \pm F^\pm$, and write $X\cdot Y = X_{\mu\nu}Y^{\mu\nu}$ (summed, with the trace understood). Self-dual and anti-self-dual 2-forms are orthogonal: $F^+\cdot F^- = {}^*F^+\cdot F^- = F^+\cdot{}^*F^- = -F^+\cdot F^-$, so $F^+\cdot F^- = 0$ (we used that $\star$ is symmetric, $X\cdot{}^*Y = {}^*X\cdot Y$). Then

$$\begin{aligned}\tfrac{1}{4}F\cdot F + \tfrac{1}{4}F\cdot{}^*F &\overset{(1)}{=} \tfrac{1}{4}\big(F^+\cdot F^+ + F^-\cdot F^-\big) + \tfrac{1}{4}\big(F^+ + F^-\big)\cdot\big(F^+ - F^-\big)\\ &\overset{(2)}{=} \tfrac{1}{4}\big(F^+\cdot F^+ + F^-\cdot F^-\big) + \tfrac{1}{4}\big(F^+\cdot F^+ - F^-\cdot F^-\big)\\ &= \tfrac{1}{2}\,F^+\cdot F^+.\end{aligned} \tag{12.103}$$

(1) Expand $F\cdot F$ using $F^+\cdot F^- = 0$, and use ${}^*F = {}^*F^+ + {}^*F^- = F^+ - F^-$.
(2) Multiply out; the cross terms vanish by orthogonality.

So the bosonic gauge action is $\frac{1}{2}\int|F^+|^2 \geq 0$, and its minima are exactly the configurations with $F^+ = 0$: the anti-self-dual instantons, Nash's (12.57), $F_{\mu\nu} = -{}^*F_{\mu\nu}$. Equivalently: the second term is topological, $\int\sqrt{g}\ \mathrm{tr}\,F_{\mu\nu}{}^*F^{\mu\nu}\,d^4x = 2\int\mathrm{tr}\,F\wedge F$, which is a fixed multiple of the instanton number, so within a fixed topological sector the gauge part of S is the ordinary Yang–Mills action plus a constant. Yang–Mills with a θ-term tuned so that the action vanishes on instantons.

The fermion kinetic terms. $iD_\mu\psi_\nu\chi^{\mu\nu}$ pairs χ with the self-dual part of $D\psi$ (only that part survives contraction with a self-dual χ), i.e. with $\mathrm{d}_A^+\psi$. $-i\eta D_\mu\psi^\mu$ pairs η with $\mathrm{d}_A^*\psi$. Together they are $\langle(\eta,\chi), D_A\psi\rangle$ with $D_A = \mathrm{d}_A^*\oplus\mathrm{d}_A^+$, the operator (12.79) whose kernel is $T\mathcal{M}_k$.

The scalar terms. $\frac{1}{2}\phi D_\mu D^\mu\lambda$ is a kinetic term for the pair (ϕ,λ) built from the covariant Laplacian $\mathrm{d}_A^*\mathrm{d}_A$ on Ω^0. The Yukawa-like terms couple ϕ to $\chi\chi$ and $\eta\eta$, and λ to $\psi\psi$. The last term is a potential $\propto|[\phi,\lambda]|^2$.

The U-number

Assign $U = (2,-2,-1,0,1,-1)$ to $(\phi,\lambda,\eta,A_\mu,\psi_\mu,\chi_{\mu\nu})$ (derivatives carry $U=0$). Every term of (12.102) has total $U = 0$:

term	$\phi DD\lambda$	$D\psi\,\chi$	$\eta D\psi$	$\phi\chi\chi$	$\lambda\psi\psi$	$\phi\eta\eta$	$[\phi,\lambda]^2$
U	$2-2$	$1-1$	$-1+1$	$2-1-1$	$-2+1+1$	$2-1-1$	0

So U is a classical symmetry, a "ghost number". Quantum mechanically it is anomalous, exactly like the axial charge of QCD in an instanton background (Chapter 10): the path-integral measure is not U-invariant when there are fermion zero modes. We compute the anomaly in Section 12.7.4: it is $\dim\mathcal{M}_k = 8k - 3(1+b_2^+)$. For fixed M this depends on k only through $8k$, so, as Nash says, U is not conserved but is conserved *mod* 8. This is the same 8 as the $\mathbb{Z}_8$ grading of Floer homology (Section 12.4.5), and for the same reason.

12.7.2 The BRST symmetry

The action is invariant under a fermionic symmetry. Following Witten, with a Grassmann parameter ε of $U = -1$,

$$\begin{aligned} \delta A_\mu &= i\varepsilon\psi_\mu, & \delta\phi &= 0, & \delta\lambda &= 2i\varepsilon\eta, \\ \delta\eta &= \tfrac{1}{2}\varepsilon[\phi,\lambda], & \delta\psi_\mu &= -\varepsilon D_\mu\phi, & \delta\chi_{\mu\nu} &= \varepsilon\big(F_{\mu\nu} + {}^*F_{\mu\nu}\big). \end{aligned} \tag{12.104}$$

Write $\delta X = \varepsilon\, QX$; then Q is an odd derivation that raises U by one ($A \to \psi$: $0 \to 1$; $\psi \to D\phi$: $1 \to 2$; $\lambda \to \eta$: $-2 \to -1$; $\eta \to [\phi,\lambda]$: $-1 \to 0$; $\chi \to F^+$: $-1 \to 0$). Physicists call such a symmetry a BRST symmetry, after the symmetry of gauge-fixed Yang–Mills theory, because it too is fermionic and squares to zero on the gauge-invariant quantities.

Worked example 12.45: Q^2 is a gauge transformation generated by ϕ

Recall that an infinitesimal gauge transformation with parameter $\xi \in \Omega^0(\mathrm{ad}P)$ acts by $\delta_\xi A = -D\xi$ and $\delta_\xi X = [\xi, X]$ on adjoint fields (Chapter T3). Apply Q twice, using $QA = i\psi$, $Q\psi = -D\phi$, $Q\phi = 0$, $Q\lambda = 2i\eta$, $Q\eta = \frac{1}{2}[\phi,\lambda]$:

$$\begin{aligned} Q^2 A_\mu &\overset{(1)}{=} Q(i\psi_\mu) = -iD_\mu\phi \overset{(2)}{=} \delta_{i\phi}A_\mu, \\ Q^2\psi_\mu &\overset{(3)}{=} -Q(D_\mu\phi) = -[QA_\mu,\phi] = -i[\psi_\mu,\phi] = [i\phi,\psi_\mu] \overset{(4)}{=} \delta_{i\phi}\psi_\mu, \\ Q^2\lambda &\overset{(5)}{=} Q(2i\eta) = 2i\cdot\tfrac{1}{2}[\phi,\lambda] = [i\phi,\lambda] = \delta_{i\phi}\lambda, \\ Q^2\phi &= 0 = [i\phi,\phi] = \delta_{i\phi}\phi. \end{aligned}$$

(1) $QA = i\psi$ and $Q\psi = -D\phi$.
(2) Gauge transformation with $\xi = i\phi$: $\delta_\xi A = -D(i\phi)$.
(3) $D_\mu\phi = \partial_\mu\phi + [A_\mu,\phi]$ and $Q\phi = 0$, so only A is varied.
(4) $\delta_\xi\psi = [\xi,\psi]$ for an adjoint field.
(5) $Q\lambda = 2i\eta$, $Q\eta = \frac{1}{2}[\phi,\lambda]$.

So $Q^2 = \delta_{\text{gauge}}(i\phi)$. On gauge-invariant functionals $Q^2 = 0$, and the Q-cohomology ("Q-closed modulo Q-exact") is well defined. The field ϕ is thus the generator of the gauge transformation that Q^2 produces. In the mathematical language, Q is the differential of the Cartan model of $\mathcal{G}$-equivariant cohomology of $\mathcal{A}$, with $\psi \leftrightarrow$ "$\mathrm{d}A$" (1-forms on $\mathcal{A}$) and $\phi \leftrightarrow$ the curvature of $\mathcal{A} \to \mathcal{A}/\mathcal{G}$; its cohomology is the de Rham cohomology of $\mathcal{A}/\mathcal{G}$. This is the precise sense in which the field theory computes the topology of the space of gauge orbits.

Dictionary: Witten's supersymmetric quantum mechanics versus Witten's TQFT

Chapter 7: SUSY QM on a manifold X	Donaldson–Witten theory on M^4
coordinates x^i	gauge field $A_\mu(x)$
fermions $\psi^i \leftrightarrow \mathrm{d}x^i$	$\psi_\mu \leftrightarrow$ "$\mathrm{d}A_\mu$"
supercharge $Q = \mathrm{d} + t\,\mathrm{d}f\wedge$	BRST charge Q
fermion number	U-number
Morse critical points	instantons $F^+ = 0$
wave functions = forms on X	observables = forms on $\mathcal{A}/\mathcal{G}$
$\ker\Delta \cong H^*(X)$	Q-cohomology $\cong H^*(\mathcal{A}/\mathcal{G})$

12.7.3 Why the answers do not depend on the metric or the coupling

The Ward identity

Everything rests on one identity. If the measure $\mathcal{DF}$ (all fields) and the action are Q-invariant, then for every functional X,

$$\langle QX \rangle = \int \mathcal{DF}\, e^{-S}\, QX = 0. \tag{12.105}$$

The reason is that Q acts on the fields as a vector field, and the integral of a derivative is zero. (We check this explicitly in the toy model of Section 12.7.4.) Physically: Q annihilates the vacuum, so a Q-exact operator has zero expectation value, as in any theory with unbroken supersymmetry.

S and $T_{\mu\nu}$ are Q-exact

Witten showed that, up to the topological term $\int \operatorname{tr} F \wedge F$,

$$S = \{Q, V\}, \qquad T_{\mu\nu} \coloneqq \frac{\delta S}{\delta g^{\mu\nu}} = \{Q, V_{\mu\nu}\}, \qquad V_{\mu\nu} = \frac{\delta V}{\delta g^{\mu\nu}}, \tag{12.106}$$

for a functional V of $U = -1$. (Here $\{Q, V\}$ means QV; the anticommutator notation recalls that in the operator formalism Q acts on operators by graded commutator.) The topological term does not depend on the metric at all. The toy model in Section 12.7.4 shows how such a V looks: schematically $V = \int \operatorname{tr}\big(\chi \cdot (F^{+} - \tfrac{1}{4}H) + \dots\big)$ with an auxiliary field $H_{\mu\nu}$, the dots being terms built from λ, η, ψ that produce the scalar part of S.

Metric independence

Let P be a Q-invariant observable ($QP = 0$) that does not depend on the metric. Vary the metric, $g \to g + \delta g$. Then

$$\begin{aligned}
\delta_g \langle P \rangle &\overset{(1)}{=} \int \mathcal{DF}\, \big(-\delta_g S\big)\, e^{-S}\, P \\
&\overset{(2)}{=} -\int \mathcal{DF}\, e^{-S}\, \{Q, \delta_g V\}\, P \\
&\overset{(3)}{=} -\int \mathcal{DF}\, e^{-S}\, Q\big(\delta_g V\, P\big) \\
&\overset{(4)}{=} -\big\langle Q(\delta_g V\, P) \big\rangle \overset{(5)}{=} 0.
\end{aligned} \tag{12.107}$$

(1) Only the action depends on g; P does not, and (for the natural measure) neither does $\mathcal{DF}$.
(2) $\delta_g S = \int \delta g^{\mu\nu} T_{\mu\nu} = \{Q, \delta_g V\}$ by (12.106).
(3) Q is a derivation and $QP = 0$, so $(Q\delta_g V)P = Q(\delta_g V\, P)$.
(4) Definition of $\langle\cdot\rangle$.
(5) The Ward identity (12.105).

So $\langle P \rangle$ is a smooth invariant: it depends on the smooth structure (through the existence of metrics and of the action) but not on the metric. In the Hamiltonian picture, the Hamiltonian $H = \int T_{00}$ is Q-exact, so it vanishes on Q-cohomology: every physical state has zero energy, as promised in Section 12.1.1.

Coupling independence

Put a coupling constant in front: $S_e = \frac{1}{e^2}\{Q, V\}$. The same chain with $\frac{\mathrm{d}}{\mathrm{d}(1/e^2)}$ in place of δ_g:

$$\frac{\mathrm{d}}{\mathrm{d}(1/e^2)}\langle P\rangle_e \overset{(1)}{=} -\big\langle \{Q,V\}\,P\big\rangle_e \overset{(2)}{=} -\big\langle Q(VP)\big\rangle_e \overset{(3)}{=} 0.$$

(1) Differentiate e^{-S_e}.
(2) $QP = 0$.
(3) The Ward identity.

So we may take $e \to 0$. There the path integral is dominated by the minima of S, and the semiclassical (Gaussian) approximation around them is *exact*. This is the step Nash describes as "we can evaluate them in a small coupling limit".

Physical picture: Why a theory with $\hbar$-independent answers is "topological"

In an ordinary QFT, the loop expansion in e^2 (or $\hbar$) generates an infinite series of corrections. Here all corrections beyond one loop cancel between bosons and fermions, because Q pairs them. The one-loop answer is exact, and it is a count: a sum over classical solutions of determinant ratios that are ± 1.

12.7.4 Localisation: the path integral becomes an integral over $\mathcal{M}_k$

Before the infinite-dimensional case, here is the entire mechanism in a model with one bosonic variable. It is the zero-dimensional ancestor of both Witten's theory and the Morse theory of Chapter 7.

Worked example 12.46: A zero-dimensional TQFT counts zeros with signs

Let $s : \mathbb{R} \to \mathbb{R}$ be a smooth function with $|s(x)| \to \infty$ as $|x| \to \infty$. The "fields" are two bosons x, H and two Grassmann variables ψ, χ, with $U(x) = U(H) = 0$, $U(\psi) = 1$, $U(\chi) = -1$. The BRST operator is

$$Qx = \psi, \quad Q\psi = 0, \quad Q\chi = H, \quad QH = 0; \qquad \text{i.e. } Q = \psi\frac{\partial}{\partial x} + H\frac{\partial}{\partial \chi},$$

and plainly $Q^2 = 0$. Take $V = \chi\big(\frac{1}{2}H - i\,s(x)\big)$ and $S_t = t\,QV$:

$$\begin{aligned} QV &\overset{(1)}{=} (Q\chi)\big(\tfrac{1}{2}H - is\big) - \chi\,Q\big(\tfrac{1}{2}H - is\big) \\ &\overset{(2)}{=} \tfrac{1}{2}H^2 - iHs(x) + i\,\chi\,s'(x)\,\psi. \end{aligned}$$

(1) Q is an odd derivation: it picks up a sign when it passes the odd variable χ.
(2) $Q\chi = H$, and $Q\,s(x) = s'(x)\,Qx = s'(x)\psi$, $QH = 0$.

Define $Z_t = i\int \frac{\mathrm{d}x\,\mathrm{d}H}{2\pi}\,\mathrm{d}\psi\,\mathrm{d}\chi\; e^{-S_t}$ (the factor i fixes the overall normalisation). Do the integrals one at a time. The Grassmann integral, with $\int \mathrm{d}\psi\,\psi = \int \mathrm{d}\chi\,\chi = 1$:

$$\int \mathrm{d}\psi\,\mathrm{d}\chi\; e^{-it\chi s'\psi} \overset{(1)}{=} \int \mathrm{d}\psi\,\mathrm{d}\chi\,\big(1 - it\,s'\,\chi\psi\big) \overset{(2)}{=} -it\,s'(x).$$

(1) The exponential series stops after the linear term, because $(\chi\psi)^2 = 0$.
(2) $\int \mathrm{d}\chi\,\chi\psi = \psi$ and then $\int \mathrm{d}\psi\,\psi = 1$; the constant term integrates to zero.

The H integral is a Gaussian:

$$\int \frac{\mathrm{d}H}{2\pi}\, e^{-t(\frac{1}{2}H^2 - iHs)} \overset{(1)}{=} \int \frac{\mathrm{d}H}{2\pi}\, e^{-\frac{t}{2}(H-is)^2}\, e^{-\frac{t}{2}s^2} \overset{(2)}{=} \frac{1}{\sqrt{2\pi t}}\, e^{-\frac{t}{2}s(x)^2}.$$

(1) Complete the square: $\frac{1}{2}H^2 - iHs = \frac{1}{2}(H-is)^2 + \frac{1}{2}s^2$.

(2) $\int e^{-tu^2/2}\mathrm{d}u = \sqrt{2\pi/t}$ (shifting the contour by is is harmless).

Putting the pieces together, with $\Phi(u) = \int_{-\infty}^{u} e^{-v^2/2}\mathrm{d}v/\sqrt{2\pi}$ the normal distribution function:

$$\begin{aligned} Z_t &\overset{(1)}{=} \int \mathrm{d}x\, \sqrt{\frac{t}{2\pi}}\, s'(x)\, e^{-\frac{t}{2}s(x)^2} \\ &\overset{(2)}{=} \int \mathrm{d}x\, \frac{\mathrm{d}}{\mathrm{d}x}\Phi\big(\sqrt{t}\, s(x)\big) \\ &\overset{(3)}{=} \Phi\big(\sqrt{t}\, s(+\infty)\big) - \Phi\big(\sqrt{t}\, s(-\infty)\big). \end{aligned} \tag{12.108}$$

(1) $i \cdot (-it\, s') \cdot (2\pi t)^{-1/2} = \sqrt{t/2\pi}\, s'$.

(2) Chain rule: $\frac{\mathrm{d}}{\mathrm{d}x}\Phi(\sqrt{t}\, s) = \sqrt{t}\, s'\, e^{-ts^2/2}/\sqrt{2\pi}$.

(3) Fundamental theorem of calculus.

Since $\Phi(+\infty) = 1$ and $\Phi(-\infty) = 0$:

$$Z_t = \begin{cases} +1, & s(-\infty) = -\infty,\ s(+\infty) = +\infty, \\ -1, & s(-\infty) = +\infty,\ s(+\infty) = -\infty, \\ 0, & s(\pm\infty) \text{ of the same sign.} \end{cases}$$

Three observations, each a piece of Nash's p. 336.

1. *Z_t is independent of t.* This is the coupling independence ($t = 1/e^2$). Directly from the Ward identity: $\frac{\mathrm{d}Z_t}{\mathrm{d}t} = -i\int (QV)e^{-tQV} = -i\int Q\big(Ve^{-tQV}\big) = 0$, because $Q = \psi\partial_x + H\partial_\chi$ is a sum of derivatives: $\int \mathrm{d}x\, \partial_x(\cdots) = 0$ by the Gaussian decay, and $\int \mathrm{d}\chi\, \partial_\chi(\cdots) = 0$ because a χ-derivative contains no χ.

2. *Z_t counts zeros of s with signs.* For s with only simple zeros x_j, $Z = \sum_j \operatorname{sgn} s'(x_j)$: going from $-\infty$ to $+\infty$, s crosses zero upwards once more than downwards. For $s = x^3 - x$: zeros at $-1, 0, 1$ with $s' = 2, -1, 2$, so $Z = 1 - 1 + 1 = 1$. For $s = x^2 - 1$: $s'(\mp 1) = \mp 2$, so $Z = -1 + 1 = 0$, in agreement with (12.108).

3. *As $t \to \infty$ the integrand localises*, and each zero contributes a ratio of determinants equal to ± 1. Near x_j, $s \approx s'_j\, y$ with $y = x - x_j$. The bosonic integrals (x and H) give $\int \mathrm{d}y\, (2\pi t)^{-1/2} e^{-ts_j'^2 y^2/2} = 1/(t|s'_j|)$: one over the bosonic "determinant". The fermionic integral gives $i \cdot (-its'_j) = ts'_j$: the fermionic determinant. The ratio is $s'_j/|s'_j| = \pm 1$.

Worked example 12.47: Zero modes must be soaked up: observables are forms

Now take two bosons $(x, y) \in \mathbb{R}^2$ but still one equation, $s(x, y) = y$. The zero set $\{y = 0\}$ is a line: a "moduli space" of dimension 1. There are fermions $\psi^x = Qx$, $\psi^y = Qy$ and one χ,

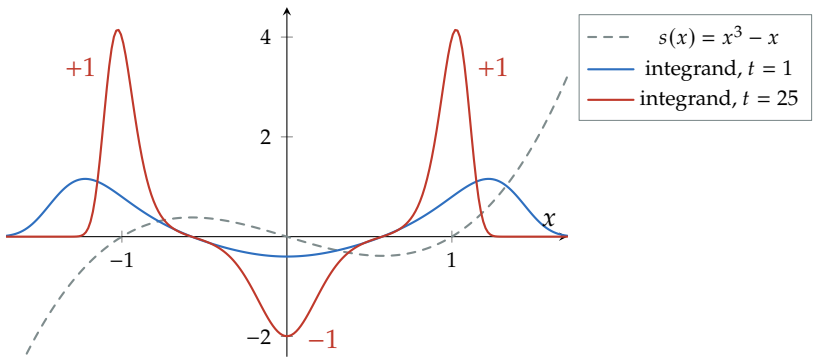


Figure 12.31: Localisation in the toy model. The integrand $\sqrt{t/2\pi}\, s'(x) e^{-t s(x)^2/2}$ has total integral 1 for every t. At $t = 1$ (blue) it is spread out; at $t = 25$ (red) it has collapsed onto narrow peaks at the zeros of s (dashed), each of area $\operatorname{sgn} s'(x_j)$: $+1 - 1 + 1 = 1$. In Witten's theory the zeros of s become the instantons and the peaks become the contributions of $\mathcal{M}_k$.

so the action $S_t = t\,Q\big(\chi(\tfrac{1}{2}H - iy)\big) = t\big(\tfrac{1}{2}H^2 - iHy + i\chi\psi^y\big)$ contains no ψ^x at all. Then

$$Z = \int \mathrm{d}\psi^x \,(\text{something independent of } \psi^x) = 0.$$

The unpaired ψ^x is a *fermion zero mode*, pointing along the moduli space. To get a non-zero answer we must insert an observable that contains ψ^x, for instance $\mathcal{O} = f(x)\,\psi^x$. It is Q-closed: $Q\mathcal{O} = f'(x)\psi^x\psi^x = 0$. Its expectation value is

$$\begin{aligned}\langle f(x)\psi^x\rangle &\overset{(1)}{=} \int \mathrm{d}x\, f(x) \int \mathrm{d}\psi^x\, \psi^x \;\times\; i\int \frac{\mathrm{d}y\,\mathrm{d}H}{2\pi}\mathrm{d}\psi^y\mathrm{d}\chi\; e^{-t(\frac{1}{2}H^2 - iHy + i\chi\psi^y)} \\ &\overset{(2)}{=} \int \mathrm{d}x\, f(x)\cdot 1\cdot 1 = \int_{\{y=0\}} f(x)\,\mathrm{d}x.\end{aligned}$$

(1) The integrand factorises into an (x, ψ^x) part and a (y, H, ψ^y, χ) part.

(2) $\int \mathrm{d}\psi^x\psi^x = 1$, and the second factor is the one-dimensional model (12.108) with $s(y) = y$, which equals 1.

The observable $f(x)\psi^x$ is really the 1-form $f(x)\,\mathrm{d}x$ ($\psi^x \leftrightarrow \mathrm{d}x$), and the answer is its integral over the zero set. The U-numbers match: the measure $\mathrm{d}\psi^x\mathrm{d}\psi^y\mathrm{d}\chi$ has $U = -1 - 1 + 1 = -1$ (a Grassmann measure $\mathrm{d}\psi$ carries the opposite charge of ψ), which is minus the dimension of the zero set, so only an observable of $U = +1$, a 1-form, can have a non-zero expectation value.

Dictionary: From the toy model to Donaldson–Witten theory

toy model	Witten's theory on M^4
x	A_μ (modulo gauge: ϕ, λ, η do the dividing)
$\psi = Qx$	ψ_μ, with $QA_\mu = i\psi_\mu$
the equation $s(x)$	F^+, a section of Ω^2_+ over $\mathcal{A}/\mathcal{G}$
χ, H	$\chi_{\mu\nu}$ and an auxiliary self-dual $H_{\mu\nu}$
$\frac{t}{2}s^2$ after integrating out H	$\frac{1}{2}\lvert F^+\rvert^2$, i.e. $\frac{1}{4}F\cdot F + \frac{1}{4}F\cdot {}^*F$
$i\chi s'\psi$	$iD_\mu\psi_\nu\chi^{\mu\nu}$, i.e. $\chi\, \mathrm{d}_A^+\psi$
zeros of s	instantons, $\mathcal{M}_k$
$\operatorname{sgn} s'(x_j)$	orientation sign of an instanton $((-1)^{n_i})$
$t \to \infty$	weak coupling $e \to 0$
$f(x)\psi^x$	$P(\mathcal{F})$, a form on $\mathcal{M}_k$

The real computation, step by step

Now repeat the three observations with Nash's (12.58),

$$\langle P\rangle = \int \mathcal{D}\mathcal{F}\; e^{-S}\, P(\mathcal{F}), \tag{12.109}$$

where $\mathcal{F}$ denotes all the fields and P is a polynomial in them.

Step 1: the minima. At $e \to 0$ we need S minimal. The bosonic part of S is a sum of squares: $\frac{1}{2}|F^+|^2$, the kinetic term of (ϕ, λ) and $|[\phi, \lambda]|^2$ (with Euclidean reality conditions making them positive). They vanish exactly when

$$F^+ = 0, \qquad D_\mu\phi = 0, \qquad D_\mu\lambda = 0.$$

These are also the fixed points of Q ($\delta\chi = 0$ and $\delta\psi = 0$ in (12.104)): the path integral localises on Q-fixed points, exactly as the toy model localised on $s = 0$. A covariantly constant ϕ generates a gauge transformation fixing A ($\delta_\phi A = -D\phi = 0$), so it lies in H^0_A, the Lie algebra of the stabiliser. For irreducible A this is zero. Hence Nash's remark "we also need ϕ and λ to vanish for irreducible connections": the minima are $\phi = \lambda = 0$ and $A \in \mathcal{M}_k$.

Step 2: the non-zero modes cancel. Expand all fields to quadratic order around a minimum. The bosonic fluctuation a of A (in a gauge $\mathrm{d}_A^* a = 0$) has quadratic action $\frac{1}{2}\langle a, D_A^* D_A a\rangle$, the expansion of $\frac{1}{2}|F^+|^2$ plus gauge fixing. The fermions ψ and (η, χ) have quadratic action $\langle(\eta, \chi), D_A\psi\rangle$. The key fact is that $D_A^* D_A$ and $D_A D_A^*$ have the same non-zero eigenvalues:

$$D_A^* D_A\, v = \mu v,\ \mu \neq 0 \overset{(1)}{=} D_A D_A^*\,(D_A v) = D_A(D_A^* D_A v) = \mu\,(D_A v), \qquad D_A v \overset{(2)}{=\!\!\!\neq} 0.$$

(1) Apply D_A to the eigenvalue equation.
(2) If $D_A v = 0$ then $\mu v = D_A^* D_A v = 0$, contradicting $\mu \neq 0$, $v \neq 0$.

(and symmetrically in the other direction, so the non-zero spectra coincide with multiplicities, as in the proof of the McKean–Singer formula, Chapter 4). Therefore

$$\frac{\text{fermions}}{\text{bosons}} \overset{(1)}{=} \frac{\det'(D_A)}{\det'(D_A^* D_A)^{1/2}} \overset{(2)}{=} \frac{\pm\prod_j \sqrt{\mu_j}}{\prod_j \sqrt{\mu_j}} = \pm 1.$$

(1) A Grassmann Gaussian gives a determinant; a bosonic one gives an inverse square root of a determinant. Primes: zero modes omitted.

(2) The singular values of D_A are the square roots $\sqrt{\mu_j}$ of the common non-zero eigenvalues; the sign depends on orientation conventions, exactly the $\operatorname{sgn} s'$ of the toy model.

The scalars (ϕ, λ) and the remaining fermions cancel in the same way. This is Nash's "Boson–Fermion ratio of determinants, a ratio which supersymmetry constrains to be ∓ 1 since Bosonic and Fermionic eigenvalues are equal in pairs".

Step 3: the zero modes remain. The zero modes of the bosonic fluctuation a are $\ker D_A = H^1_A = T_{[A]}\mathcal{M}_k$: moving along them costs no action, so they are not Gaussian integrals but an honest integral over $\mathcal{M}_k$. The fermionic zero modes are: for ψ, again $\ker D_A$ ($n = \dim \mathcal{M}_k$ of them); for η and χ, $\ker D_A^* = H^0_A \oplus H^2_A = 0$ (none, for irreducible unobstructed instantons). The n Grassmann zero modes $\psi_0^1, \dots, \psi_0^n$, like ψ^x in Example 12.47, must be soaked up by P. Expanding $P(\mathcal{F})$ and keeping the term with exactly n factors of ψ_0 gives

$$\langle P \rangle = \int_{\mathcal{M}_k} P_n, \tag{12.110}$$

Nash's (12.59), where P_n is an n-form on $\mathcal{M}_k$ ($\psi_0^i \leftrightarrow \mathrm{d}x^i$ on $\mathcal{M}_k$).

Step 4: the selection rule. Each zero mode ψ_0^i has $U = 1$, so each measure factor $\mathrm{d}\psi_0^i$ has $U = -1$:

$$\begin{aligned} U(\text{measure}) &\overset{(1)}{=} -\#(\psi \text{ zero modes}) + \#(\eta \text{ zero modes}) + \#(\chi \text{ zero modes}) \\ &\overset{(2)}{=} -\dim H^1_A + \dim H^0_A + \dim H^2_A \overset{(3)}{=} -\dim \mathcal{M}_k = -\big(8k - 3(1 + b_2^+)\big). \end{aligned} \tag{12.111}$$

(1) ψ has $U = +1$; η, χ have $U = -1$; the non-zero modes come in pairs of opposite U and contribute nothing net.
(2) The zero modes are the cohomology groups of the deformation complex (Figure 12.30).
(3) $H^0_A = H^2_A = 0$, and the index theorem (12.83). (In general the alternating sum is the index, which is still the expected dimension.)

So $\langle P \rangle \neq 0$ requires $U(P) = \dim \mathcal{M}_k$. Since the dimension changes by 8 when k changes by 1, the net U-violation is fixed mod 8 on a given M. With the right choice of P, (12.110) is literally Donaldson's integral (12.95). Which P? That is the subject of the next subsection.

12.7.5 Observables and the descent equations

We need Q-invariant, gauge-invariant, metric-independent observables with the right U-number. Start from the only gauge-invariant polynomial in ϕ of lowest degree:

$$W_0 = \tfrac{1}{2} \operatorname{tr} \phi^2, \qquad U(W_0) = 4.$$

It is Q-invariant because $Q\phi = 0$. It is a function on M, so it depends on the point x where it is inserted. But its derivative is Q-exact. With $W_1 \coloneqq \operatorname{tr}(\phi\, \psi)$, a 1-form on M of $U = 3$:

$$\mathrm{d}W_0 \overset{(1)}{=} \operatorname{tr}(\phi\, \mathrm{d}\phi) \overset{(2)}{=} \operatorname{tr}(\phi\, D\phi) \overset{(3)}{=} -\operatorname{tr}(\phi\, Q\psi) \overset{(4)}{=} -Q \operatorname{tr}(\phi\, \psi) = -QW_1.$$

(1) $\mathrm{d}\frac{1}{2} \operatorname{tr} \phi^2 = \frac{1}{2} \operatorname{tr}(\mathrm{d}\phi\, \phi + \phi\, \mathrm{d}\phi)$, and the trace is cyclic.
(2) $D\phi = \mathrm{d}\phi + [A, \phi]$ and $\operatorname{tr}(\phi[A, \phi]) = 0$ (cyclicity).
(3) $Q\psi = -D\phi$.
(4) $Q\phi = 0$, so Q passes ϕ and acts only on ψ.

The pattern continues. Define

$$W_2 = \operatorname{tr}\big(\tfrac{1}{2}\psi \wedge \psi + i\phi F\big), \qquad W_3 = i \operatorname{tr}(\psi \wedge F), \qquad W_4 = -\tfrac{1}{2} \operatorname{tr}(F \wedge F), \tag{12.112}$$

with $U(W_k) = 4 - k$. They obey Witten's **descent equations**

$$QW_0 = 0, \qquad \mathrm{d}W_k = -QW_{k+1} \ \ (k = 0,1,2,3), \qquad \mathrm{d}W_4 = 0. \tag{12.113}$$

Let us verify the next step, $\mathrm{d}W_1 = -QW_2$. Here Q commutes with $\mathrm{d}x^\mu$ and anticommutes with the Grassmann-odd components ψ_μ. First,

$$\begin{aligned} Q\ \mathrm{tr}(\tfrac{1}{2}\psi\wedge\psi) &\overset{(1)}{=} \tfrac{1}{2}\,\mathrm{tr}\big((Q\psi_\mu)\psi_\nu - \psi_\mu(Q\psi_\nu)\big)\mathrm{d}x^\mu\wedge\mathrm{d}x^\nu \\ &\overset{(2)}{=} -\tfrac{1}{2}\,\mathrm{tr}\big(D_\mu\phi\,\psi_\nu - \psi_\mu D_\nu\phi\big)\mathrm{d}x^{\mu\nu} \overset{(3)}{=} -\,\mathrm{tr}(D\phi\wedge\psi), \end{aligned}$$

(1) Q is an odd derivation acting on the Grassmann components.
(2) $Q\psi_\mu = -D_\mu\phi$.
(3) $\mathrm{tr}(\psi_\mu D_\nu\phi)\mathrm{d}x^{\mu\nu} = \mathrm{tr}(D_\nu\phi\,\psi_\mu)\mathrm{d}x^{\mu\nu} = -\,\mathrm{tr}(D\phi\wedge\psi)$ (cyclicity, $D\phi$ Grassmann-even, then $\mathrm{d}x^{\mu\nu} = -\mathrm{d}x^{\nu\mu}$).

and $Q\ \mathrm{tr}(\phi F) = \mathrm{tr}(\phi\,QF) = \mathrm{tr}(\phi\,D(QA)) = i\,\mathrm{tr}(\phi\,D\psi)$, using $\delta F = D\,\delta A$. So

$$-QW_2 \overset{(1)}{=} \mathrm{tr}(D\phi\wedge\psi) - i\cdot i\ \mathrm{tr}(\phi\,D\psi) = \mathrm{tr}(D\phi\wedge\psi + \phi\,D\psi) \overset{(2)}{=} \mathrm{d}\ \mathrm{tr}(\phi\,\psi) = \mathrm{d}W_1.$$

(1) The two computations just made.
(2) Leibniz rule; the $[A,\cdot]$ terms cancel inside the trace.

This fixes the coefficient i in W_2.

Try it: The last two descent equations

Check $\mathrm{d}W_2 = -QW_3$ and $\mathrm{d}W_3 = -QW_4$, using $DF = 0$ (Bianchi), $QF = iD\psi$ and $\mathrm{tr}(\psi\wedge D\psi) = -\,\mathrm{tr}(D\psi\wedge\psi)$.
Answer: $\mathrm{d}W_2 = \mathrm{tr}(D\psi\wedge\psi) + i\,\mathrm{tr}(D\phi\wedge F)$ and $QW_3 = i\,\mathrm{tr}(Q\psi\wedge F) - i\,\mathrm{tr}(\psi\wedge QF) = -i\,\mathrm{tr}(D\phi\wedge F) - \mathrm{tr}(D\psi\wedge\psi)$. For the last one, $\mathrm{d}W_3 = i\,\mathrm{tr}(D\psi\wedge F)$ and $QW_4 = -\,\mathrm{tr}(QF\wedge F) = -i\,\mathrm{tr}(D\psi\wedge F)$.

Key idea: Observables from cycles

For a k-dimensional closed submanifold (cycle) $\gamma\subset M$ define $I(\gamma) = \int_\gamma W_k$. Then:

$$Q\,I(\gamma) \overset{(1)}{=} \int_\gamma QW_k \overset{(2)}{=} -\int_\gamma \mathrm{d}W_{k-1} \overset{(3)}{=} -\int_{\partial\gamma} W_{k-1} \overset{(4)}{=} 0,$$

(1) Q acts on fields, not on the point of integration.
(2) Descent equation (12.113).
(3) Stokes' theorem (Chapter T2).
(4) γ is closed: $\partial\gamma = \varnothing$.

so $I(\gamma)$ is Q-invariant. If $\gamma = \partial\beta$ is a boundary, the same steps in reverse give $I(\gamma) = \int_\beta \mathrm{d}W_k = -Q\int_\beta W_{k+1}$, which is Q-exact and so has zero expectation value by (12.105). Hence $\langle I(\gamma_1)\cdots I(\gamma_r)\rangle$ depends only on the *homology classes* of the γ_i: the field theory turns homology of M into observables.

Matching with Donaldson's classes

For a surface Σ, $I(\Sigma) = \int_\Sigma \mathrm{tr}(\tfrac{1}{2}\psi\wedge\psi + i\phi F)$ has $U = 2$. In the localised theory, ψ is replaced by its zero modes, which are tangent vectors a, b to $\mathcal{M}_k$, and ϕ by its value forced by the λ

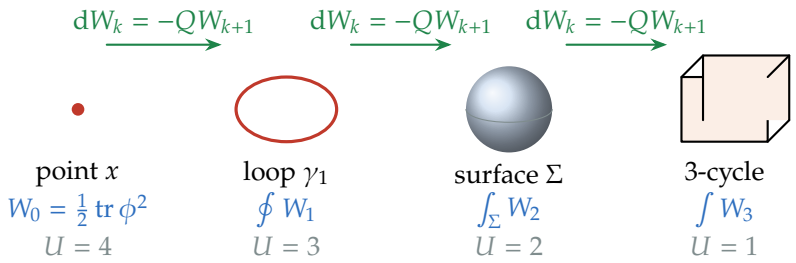


Figure 12.32: The descent ladder. Each W_k is a k-form on M of U-number $4-k$, and its exterior derivative is Q of the next one. Integrated over a k-cycle it gives a Q-invariant observable that depends only on the homology class. For simply connected M only points and surfaces matter ($b_1 = b_3 = 0$): they give the classes of degree 4 and 2 on $\mathcal{M}_k$.

equation of motion, $\mathrm{d}_A^* \mathrm{d}_A \phi \propto [\psi, \star\psi]$, which is the $\Phi(a,b)$ of Section 12.6.3. So

$$I(\Sigma) \longrightarrow \int_\Sigma \operatorname{tr}\big(a\wedge b + \Phi(a,b)F_A\big) \;\propto\; m(\Sigma)(a,b),$$

exactly the two terms of (12.93): the field theory's observable *is* Donaldson's 2-form. Therefore, when $2d = \dim \mathcal{M}_k$,

$$\big\langle I(\Sigma_1)\cdots I(\Sigma_d)\big\rangle \overset{(1)}{=} \int_{\mathcal{M}_k} m(\Sigma_1)\wedge\cdots\wedge m(\Sigma_d) \overset{(2)}{=} q_d(M)(\Sigma_1,\ldots,\Sigma_d), \tag{12.114}$$

(1) Localisation (12.110) with $P = \prod I(\Sigma_i)$, whose U-number $2d$ equals $\dim \mathcal{M}_k$ as the selection rule demands.
(2) Definition (12.95).

up to normalisation. Similarly $W_0(x)$ becomes a 4-form ν on $\mathcal{M}_k$ (the "point class", independent of x by descent), and the general Donaldson invariant is $\langle I(\Sigma_1)\cdots I(\Sigma_r)\, W_0^s\rangle$ with $2r + 4s = \dim \mathcal{M}_k$. Nash's q_d are the $s = 0$ ones.

Watch out: What the field theory does and does not prove

Witten's derivation is a physicist's argument: it assumes the path integral exists, is Q-invariant and localises. It explains *why* Donaldson's integrals are metric-independent and *what* they are (correlation functions), and it reproduces their structural properties. It does not by itself handle compactness of $\mathcal{M}_k$ or reducible connections; those are the hard analysis of Section 12.6.5. Its real power appeared later, when Seiberg and Witten used the physical *strong-coupling* description of the same theory to compute the invariants.

Reading Nash (pp. 335–336, (12.56)–(12.59))

Nash says Witten obtained the $q_d(M)$ as correlation functions of a BRST-supersymmetric TQFT, the relativistic generalisation of Floer's theory, and describes only its principal features. He writes the action (12.56), lists the fields $(\phi, \lambda, \eta, \psi_\mu, \chi_{\mu\nu})$, notes that ϕ, λ are spinless and that (η, ψ, χ) are a 0-form, a 1-form and a self-dual 2-form matching the anti-instanton deformation complex (8.104), and assigns $U = (2, -2, -1, 0, 1, -1)$, conserved only mod 8. Metric independence follows from $S = \{Q, V\}$, $T = \{Q, V'\}$; coupling independence lets one take weak coupling, where the minima are instantons (12.57) with $\phi, \lambda = 0$ for irreducibles. For $\langle P\rangle = \int \mathcal{D}\mathcal{F}\, \exp(-S)P$ (12.58) the zero modes are tangent to $\mathcal{M}_k$, the non-zero modes give a boson–fermion ratio ∓ 1, and $\langle P\rangle = \int_{\mathcal{M}_k} P_n$ (12.59), non-zero only when $U(P) = \dim \mathcal{M}_k$. We have supplied: the twisting origin of the fields; the F^+

form of the gauge action; the U-check of every term; the BRST transformations and $Q^2 =$ gauge; the metric- and coupling-independence chains; the complete zero-dimensional toy model, with and without zero modes; the spectral pairing behind the determinant ratio; the U-anomaly as an index; and the descent equations that produce the observables and match them to Donaldson's $m(\Sigma)$.

Recap

- Witten's theory has fields $A, \psi, \phi, \lambda, \eta, \chi$; the fermions (η, ψ, χ) fill the deformation complex $\Omega^0 \to \Omega^1 \to \Omega^2_+$. It is the twist of $N = 2$ super Yang–Mills, which makes one supercharge Q a scalar.
- The gauge action is $\frac{1}{2}|F^+|^2$, minimised by instantons. U is a classical symmetry, broken by $\dim \mathcal{M}_k$ units by zero modes, hence conserved mod 8.
- Q^2 is a gauge transformation generated by ϕ; S and $T_{\mu\nu}$ are Q-exact. The Ward identity $\langle QX \rangle = 0$ then gives independence of the metric and of the coupling.
- At weak coupling the path integral localises on Q-fixed points: $F^+ = 0$, $\phi = \lambda = 0$. Non-zero modes cancel to ± 1 because D^*D and DD^* share their non-zero spectrum; zero modes turn the integrand into a form on $\mathcal{M}_k$: $\langle P \rangle = \int_{\mathcal{M}_k} P_n$.
- The toy model $Z = \int \sqrt{t/2\pi}\, s' e^{-ts^2/2} \mathrm{d}x = \sum \operatorname{sgn} s'(x_j)$ contains the whole mechanism.
- The descent equations $\mathrm{d}W_k = -QW_{k+1}$ turn k-cycles into observables; $\int_\Sigma W_2$ is Donaldson's $m(\Sigma)$, and $\langle \prod I(\Sigma_i) \rangle = q_d$.

12.8 Cutting along homology spheres: Floer meets Donaldson

We now have two theories that look as if they were made for each other. Floer's theory (Section 12.5) attaches to a closed oriented homology 3-sphere N a $\mathbb{Z}_8$-graded group $HF_*(N)$, built from flat connections and the instantons on the cylinder $N \times \mathbb{R}$. Donaldson's theory (Section 12.6) attaches to a closed oriented 4-manifold M a list of numbers $q_d(M)$, built from the instantons on M. Atiyah's axioms (Section 12.1.3) tell us what a TQFT in $3+1$ dimensions should do: to a closed 3-manifold it assigns a vector space $Z(N)$, to a 4-manifold with boundary N it assigns a vector in $Z(N)$, and when a closed 4-manifold is cut into two pieces its number is the pairing of the two vectors.

So the question of this section is:

If we cut M along a homology 3-sphere N, can we compute $q_d(M)$ from the two halves, with the Floer group $HF_*(N)$ playing the role of $Z(N)$?

The answer is yes, and it is Nash's formula (12.63),

$$[C^+] \bullet [C^-] = q_d(M).$$

To understand it we need four things, which are the four subsections below: how to cut a 4-manifold along a homology sphere and what this does to the intersection form; what an instanton on a manifold with boundary is, and why its "boundary value" is a flat connection; why the set of such instantons defines a *Floer homology class*; and why pairing two such classes reproduces the count of instantons on the closed manifold. After that (Section 12.8.5 onwards)

we look at the simplest case, $d = 0$, read the formula in both directions, and set up the relative invariants.

Throughout, N is a closed oriented homology 3-sphere ($H_1(N;\mathbb{Z}) = 0$), the gauge group is SU(2), and instantons are anti-self-dual, $F^+ = 0$. We use the conventions of Section 12.4: on a cylinder $N \times \mathbb{R}$ with orientation $\mathrm{d}t \wedge \mathrm{vol}_N$ the ASD equation is the downward gradient flow $\dot{A} = -\star F_A$ of the Chern–Simons function f, and a flow line from α (at $t = -\infty$) to β (at $t = +\infty$) has charge $k = f(\alpha) - f(\beta) \geq 0$ (12.49).

12.8.1 Splitting a 4-manifold along a homology sphere

The operation $M = M^+ \cup_h M^-$

Take two compact oriented 4-manifolds M^+ and M^- with boundaries

$$\partial M^+ = N, \qquad \partial M^- = \overline{N},$$

where $\overline{N}$ is N with the opposite orientation. Glue them along the common boundary with an orientation-reversing identification. The result is a closed oriented 4-manifold, which Nash writes (12.60)

$$M = M^+ \cup_h M^-. \tag{12.115}$$

The subscript h is a reminder that the gluing surface is a *homology* sphere, not an ordinary one. Nash calls the two copies of the boundary N^+ and N^-. They are the same 3-manifold N; what differs is the orientation each inherits from its side (Figure 12.33).

Watch out: Why the two boundaries have opposite orientations

The rule for the orientation of a boundary is *outward normal first* (Chapter T2): an ordered basis (v_1, v_2, v_3) of $T\partial X$ is positive if $(n_{\text{out}}, v_1, v_2, v_3)$ is positive in X. Near the gluing surface, write $M \supset N \times (-\varepsilon, \varepsilon)$ with coordinate t and orientation $\mathrm{d}t \wedge \mathrm{vol}_N$, with M^+ at $t < 0$ and M^- at $t > 0$. The outward normal of M^+ is $+\partial_t$, so ∂M^+ carries $+\mathrm{vol}_N$: it is N. The outward normal of M^- is $-\partial_t$, so ∂M^- carries $-\mathrm{vol}_N$: it is $\overline{N}$. This is the same reason that the two ends of an interval $[a, b]$ come with opposite signs in $\int_a^b g'\,\mathrm{d}t = g(b) - g(a)$.

The simplest example is the **connected sum** (Section 12.6.6). Remove a small ball from each of two closed manifolds M_1, M_2 and glue along the two boundary 3-spheres:

$$M_1 \# M_2 = (M_1 \setminus B^4) \cup_{S^3} (M_2 \setminus B^4).$$

Here $N = S^3$, which is certainly a homology sphere. The operation $\cup_h$ generalises this by allowing any homology sphere, such as the Poincaré sphere $\Sigma(2,3,5)$ or the Brieskorn spheres $\Sigma(2,3,6n-1)$.

What the splitting does to H_2 and to Q

Nash says the decomposition exists when the intersection form splits, $Q(M) = Q(M^+) \oplus Q(M^-)$. Let us first see the easy direction: if M is split along a homology sphere, then Q splits.

The tool is the **Mayer–Vietoris sequence**. It is the homology version of the inclusion–exclusion principle: if $X = U \cup V$, a cycle in X can be cut into a piece in U and a piece in V whose boundaries agree on $U \cap V$. The precise statement is that the sequence

$$\cdots \to H_p(U \cap V) \xrightarrow{(i_*, -j_*)} H_p(U) \oplus H_p(V) \to H_p(X) \xrightarrow{\partial} H_{p-1}(U \cap V) \to \cdots \tag{12.116}$$

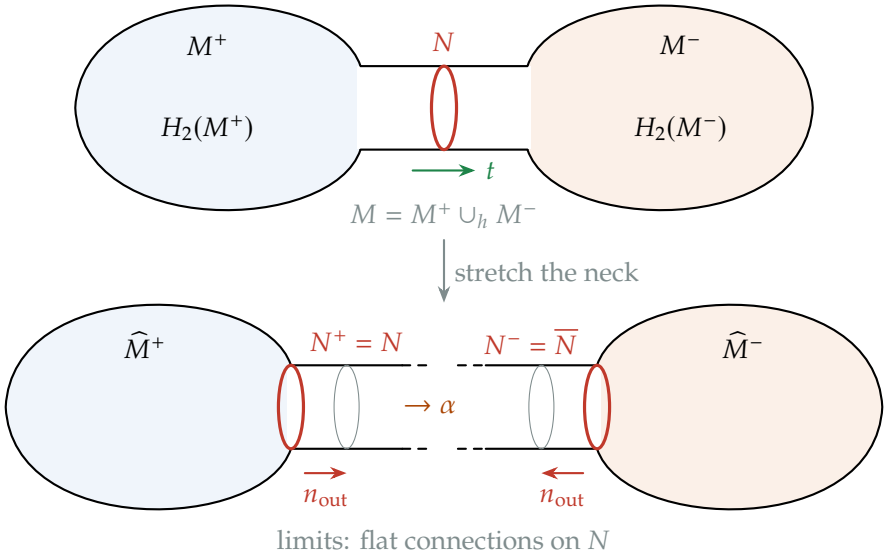


Figure 12.33: Cutting a closed 4-manifold along a homology sphere N. Top: M is the union of M^+ and M^- along N. Bottom: stretching the neck to infinite length turns each half into a manifold with a cylindrical end. See the outward normals point in opposite t-directions, which is why $\partial M^+ = N$ but $\partial M^- = \overline{N}$. On each end an instanton tends to a flat connection α on N.

is exact: at each spot the image of the incoming map equals the kernel of the outgoing map. The middle arrow adds a cycle in U and a cycle in V; the map ∂ takes a cycle in X, cuts it along $U \cap V$, and returns the boundary of the U-piece. We take $U \simeq M^+$, $V \simeq M^-$ (thickened slightly) and $U \cap V \simeq N$.

Two facts about N are needed:

$$H_1(N) \overset{(1)}{=} 0, \qquad H_2(N) \overset{(2)}{=} H^1(N) \overset{(3)}{=} \operatorname{Hom}\big(H_1(N), \mathbb{Z}\big) = 0. \tag{12.117}$$

(1) N is a homology sphere.
(2) Poincaré duality on the closed oriented 3-manifold N: $H_p \cong H^{3-p}$.
(3) H^1 is the dual of H_1 (a 1-cocycle is determined by its values on loops; the universal coefficient theorem), and $H_1 = 0$.

This is Nash's footnote 1 of Section 12.2 again, now doing real work. Put $p = 2$ in (12.116):

$$\underbrace{H_2(N)}_{=0} \longrightarrow H_2(M^+) \oplus H_2(M^-) \longrightarrow H_2(M) \longrightarrow \underbrace{H_1(N)}_{=0}. \tag{12.118}$$

Exactness at $H_2(M^+) \oplus H_2(M^-)$ says the middle map has kernel equal to the image of 0, so it is injective. Exactness at $H_2(M)$ says its image is the kernel of the map to 0, i.e. everything, so it is surjective. Hence

$$\boxed{H_2(M) \cong H_2(M^+) \oplus H_2(M^-)} \tag{12.119}$$

which is the decomposition Nash uses in (12.67).

The intersection form splits too. A class $a^+ \in H_2(M^+)$ is represented by a surface inside M^+, and $a^- \in H_2(M^-)$ by a surface inside M^-. These surfaces are disjoint, so

$$\begin{aligned} Q(a^+ + a^-,\, b^+ + b^-) &\overset{(1)}{=} Q(a^+, b^+) + Q(a^+, b^-) + Q(a^-, b^+) + Q(a^-, b^-) \\ &\overset{(2)}{=} Q^+(a^+, b^+) + Q^-(a^-, b^-), \end{aligned} \tag{12.120}$$

(1) Q is bilinear.
(2) Surfaces on opposite sides do not meet, so their intersection number is zero; surfaces on the same side meet only inside that side, so Q restricted to $H_2(M^\pm)$ is the intersection form $Q^\pm$ of $M^\pm$.

In matrix language: in a basis adapted to (12.119), the matrix of Q is block diagonal, $Q = \begin{pmatrix} Q^+ & 0 \\ 0 & Q^- \end{pmatrix}$. Each block is itself unimodular ($\det Q^\pm = \pm 1$), because $\det Q = \det Q^+ \det Q^- = \pm 1$ with integer factors. Consequently b_2^+ and b_2^- add:

$$b_2^\pm(M) = b_2^\pm(M^+) + b_2^\pm(M^-). \tag{12.121}$$

The converse: Freedman and Taylor

Nash quotes the converse: *if $Q(M) = Q_1 \oplus Q_2$ is any splitting of the intersection form of a closed simply connected smooth 4-manifold, then $M = M^+ \cup_h M^-$ along some homology sphere, with $Q(M^\pm) = Q_{1,2}$* (Freedman and Taylor, 1977). The point to appreciate is the contrast with connected sums. Every splitting of Q is realised by cutting along a homology sphere, but *not* by cutting along an honest S^3. We saw why in Example 12.43: $E(3)$ has $Q = 5(1) \oplus 29(-1)$, which splits into pieces with $b_2^+ > 0$, yet $E(3)$ is not a connected sum of two such pieces, because it has non-zero Donaldson invariants. The homology spheres are the extra room that smooth 4-manifolds need.

Worked example 12.48: Splitting the elliptic surfaces: nucleus and Milnor fibre

The elliptic surfaces $E(n)$ (with $E(1) = \mathbb{CP}^2 \# 9\overline{\mathbb{CP}}^2$ and $E(2) = \mathrm{K3}$) contain a torus fibre F and a sphere section S with

$$F \cdot F = 0, \qquad F \cdot S = 1, \qquad S \cdot S = -n.$$

A neighbourhood of $F \cup S$ is a 4-manifold with boundary, the **nucleus** $N(n)$. Its boundary is the Brieskorn sphere $\Sigma(2,3,6n-1)$, and its complement is the **Milnor fibre** $M(2,3,6n-1)$ of the singularity $z_1^2 + z_2^3 + z_3^{6n-1} = 0$. So

$$E(n) = N(n) \cup_{\Sigma(2,3,6n-1)} M(2,3,6n-1)$$

is a splitting of exactly Nash's kind. Let us check that the forms fit.

The nucleus. In the basis (F, S) the form is $Q_{N(n)} = \begin{pmatrix} 0 & 1 \\ 1 & -n \end{pmatrix}$, with determinant -1: unimodular, as it must be for a manifold bounded by a homology sphere. For $n = 2$ change basis to $e_1 = F$, $e_2 = S + F$:

$$e_2 \cdot e_2 \overset{(1)}{=} S \cdot S + 2\,S \cdot F + F \cdot F \overset{(2)}{=} -2 + 2 + 0 = 0,$$
$$e_1 \cdot e_2 \overset{(1)}{=} F \cdot S + F \cdot F \overset{(2)}{=} 1,$$

(1) Bilinearity and symmetry of the intersection form.
(2) The values $F^2 = 0$, $F \cdot S = 1$, $S^2 = -2$.

so $Q_{N(2)} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = H$, the hyperbolic form of $S^2 \times S^2$ (Example 12.28), with $b_2^+ = b_2^- = 1$. For $n = 1$ the same change of basis gives $e_2 \cdot e_2 = -1 + 2 = 1$ and $S \cdot e_2 = -1 + 1 = 0$, so in the basis $(S, S + F)$ one finds $Q_{N(1)} = \mathrm{diag}(-1, 1)$.

The Milnor fibre. Its rank is the Milnor number $\mu = (p-1)(q-1)(r-1)$. For $(2,3,11)$ this is $1 \cdot 2 \cdot 10 = 20$, and for $(2,3,5)$ it is $1 \cdot 2 \cdot 4 = 8$.

The bookkeeping for K3. $b_2(\mathrm{K3}) = 22 = 2+20$, and the signature is $-16 = 0+\tau(M(2,3,11))$, so the Milnor fibre has $b_2^+ = 2$, $b_2^- = 18$, and (being even) $Q_{M(2,3,11)} = 2H \oplus 2(-E_8)$. Altogether

$$Q(\mathrm{K3}) = \underbrace{H}_{N(2)} \oplus \underbrace{2H \oplus 2(-E_8)}_{M(2,3,11)} = 3H \oplus 2(-E_8),$$

in agreement with Table 12.1, and by (12.121) $b_2^+ = 1+2 = 3$. *Both sides have* $b_2^+ > 0$: this is a splitting of the type Nash needs, along $N = \Sigma(2,3,11)$.
The bookkeeping for $E(1)$. Here $(1) \oplus 9(-1) = \mathrm{diag}(-1,1) \oplus (-E_8)$ over the integers (both are odd, indefinite, unimodular, of rank 10 and signature -8), with the $-E_8$ carried by $M(2,3,5)$, the E_8 plumbing bounded by the Poincaré sphere. The Milnor-fibre side now has $b_2^+ = 0$, so this splitting does *not* satisfy Nash's hypothesis.

12.8.2 Instantons on a manifold with a cylindrical end

Nash says that on $M^\pm$ "boundary conditions are required since the $M^\pm$ are not closed. The boundary conditions are just the specifying of a connection on the boundaries." Let us make this precise, because the precise version is what makes everything work.

Stretch the neck

The Donaldson invariants do not depend on the metric (Section 12.7.3 and Section 12.8.8). So we may choose a metric on M that is a product $\mathrm{d}t^2 + g_N$ on a long neck $N \times [-T, T]$, and let $T \to \infty$. In the limit each half acquires an infinite cylindrical end (bottom of Figure 12.33):

$$\widehat{M}^+ = M^+ \cup_N \big(N \times [0,\infty)\big), \qquad \widehat{M}^- = \big(N \times (-\infty, 0]\big) \cup_N M^-.$$

On the cylindrical part of $\widehat{M}^+$ everything is as in Section 12.4.2: in temporal gauge an ASD connection is a path $A(t)$ of connections on N with

$$\frac{\mathrm{d}A}{\mathrm{d}t} = -\star_N F_{A(t)} = -4\pi^2 \operatorname{grad} f\big(A(t)\big) \tag{12.122}$$

by (12.45) and (12.39). So an instanton on $\widehat{M}^+$ is "an instanton on M^+ with a Chern–Simons gradient flow line attached".

Finite action forces a flat limit

Finite Yang–Mills action on the end means $\int_0^\infty \|\dot{A}\|^2 \,\mathrm{d}t < \infty$. Along a gradient flow the action on $[t_1, t_2]$ is the drop of f (the same chain as (12.48)), so $f(A(t))$ decreases and converges. Then $\|\dot{A}\|^2 = \|\star F\|^2 \to 0$ along a sequence of times, and with the compactness of the space of flat connections one shows

$$\lim_{t\to\infty} A(t) = \alpha, \qquad \alpha \text{ a flat connection on } N. \tag{12.123}$$

This is the four-dimensional statement "finite action forces the field strength to vanish at infinity", which for $\mathbb{R}^4$ gave the pure-gauge boundary condition of the BPST instanton (Chapter 8). Here the "sphere at infinity" is N, and "pure gauge" is replaced by "flat", because N has non-trivial flat connections.

So Nash's "boundary condition" is: *choose the limiting flat connection*. For each $\alpha \in \mathcal{R}^*(N)$ (the irreducible flat connections of (12.27), perturbed if necessary so that each is non-degenerate) define

$$\mathcal{M}(M^+;\alpha) := \big\{A \text{ ASD on } \widehat{M}^+,\ \ \|F_A\| < \infty,\ \ A(t) \to \alpha \text{ as } t \to \infty\big\}/\mathcal{G}, \tag{12.124}$$

and similarly $\mathcal{M}(M^-;\alpha)$ with $A(t) \to \alpha$ as $t \to -\infty$ on $\widehat{M}^-$. These are Nash's solution sets $C^{\pm}$: more precisely, C^+ is the collection of all the $\mathcal{M}(M^+;\alpha)$ as α varies.

Working definition 12.13 (The moduli space of a manifold with a cylindrical end)

To describe $\mathcal{M}(M^+;\alpha)$:

1. Attach $N \times [0,\infty)$ to M^+ and use a product metric there [$g_N + \mathrm{d}t^2$, the same metric on every slice $N \times \{t\}$].
2. Solve $F_A^+ = 0$ [the self-dual part of the curvature vanishes, so A is an instanton] with finite action. On the end, the solution is a Chern–Simons flow line (12.122).
3. Record its limit $\alpha = \lim_{t\to\infty} A(t)$, a flat connection on N. Group the solutions by their limit.
4. The dimension of $\mathcal{M}(M^+;\alpha)$ is the index of the linearised operator [independent solutions of the linearised instanton equation, minus independent obstructions] (Section 12.4.4) with decay at the end; it depends on α through its Floer degree [the $\mathbb{Z}_8$ grading $\mu(\alpha)$ fixed by spectral flow], as in (12.127) below.

The charge becomes fractional

On a closed manifold the instanton number k is an integer. On $\widehat{M}^+$ it is not, and the Chern–Simons function tells us exactly what it is. Recall the local identity $\operatorname{tr}(F\wedge F) = \mathrm{d}\operatorname{tr}\big(A\wedge\mathrm{d}A + \frac{2}{3}A\wedge A\wedge A\big)$ ((12.16)), and $f(A) = -\frac{1}{8\pi^2}\int_N \operatorname{tr}\big(A\mathrm{d}A + \frac{2}{3}A^3\big)$. Suppose first that the bundle over $\widehat{M}^+$ is trivialised, so that A is a global 1-form. Truncating the end at $t = T$,

$$\begin{aligned}\frac{1}{8\pi^2}\int_{M^+\cup N\times[0,T]} \operatorname{tr}(F\wedge F) &\overset{(1)}{=} \frac{1}{8\pi^2}\int_{\partial(M^+\cup N\times[0,T])} \operatorname{tr}\big(A\mathrm{d}A + \tfrac{2}{3}A^3\big)\\ &\overset{(2)}{=} \frac{1}{8\pi^2}\int_{N\times\{T\}} \operatorname{tr}\big(A\mathrm{d}A + \tfrac{2}{3}A^3\big) \overset{(3)}{=} -f\big(A(T)\big) \xrightarrow{T\to\infty} -f(\alpha).\end{aligned} \tag{12.125}$$

(1) Stokes' theorem applied to the transgression formula.
(2) The only boundary is $N \times \{T\}$, with the orientation $+N$ (outward normal $+\partial_t$, as in the pitfall above).
(3) Definition of f.

In general the bundle over $\widehat{M}^+$ need not be trivial, and trivialising it near the end changes f by an integer ((12.17)). So the charge on $\widehat{M}^+$ satisfies

$$k^+ \equiv -f(\alpha) \pmod 1, \qquad \text{and likewise} \qquad k^- \equiv +f(\alpha) \pmod 1, \tag{12.126}$$

the sign flipping because $\partial\widehat{M}^- = \overline{N}$. When we glue, the total charge $k = k^+ + k^-$ is an integer again, as it must be.

Worked example 12.49: Fractional instantons ending on the Poincaré sphere

The Poincaré sphere $\Sigma(2,3,5)$ has two irreducible flat connections, ρ_1 and ρ_2 (Example 12.11). Their Chern–Simons values were computed by Fintushel and Stern; in one common normalisation

$$f(\rho_1) \equiv \tfrac{1}{120}, \qquad f(\rho_2) \equiv \tfrac{49}{120} \pmod 1.$$

(The denominators are $120 = |\pi_1(\Sigma(2,3,5))|$, the order of the binary icosahedral group, which is no accident: a flat connection is a representation of π_1, and $120\,f$ is an integer invariant of it.) So by (12.126) an instanton on a manifold with a $\Sigma(2,3,5)$-end limiting to ρ_1 carries charge $k^+ \in -\frac{1}{120} + \mathbb{Z}$, e.g. $k^+ = \frac{119}{120}$. Gluing it to an instanton on the other side with $k^- \in \frac{1}{120} + \mathbb{Z}$ gives an integer total. The Chern–Simons function on the gluing surface tells the two halves how to share the charge.

Dimensions: the Floer grading reappears

What is $\dim \mathcal{M}(M^+;\alpha)$, and how does it depend on α? The cleanest way to see it is to glue a flow line onto the end. Suppose $u \in \mathcal{M}(M^+;\alpha)$ and γ is a flow line on $N \times \mathbb{R}$ from α to β. Placing γ far down the end of $\widehat{M}^+$, after u has settled down to α, gives an approximate instanton limiting to β, which can be corrected to an exact one (this is the gluing construction of Section 12.5.1). So families glue, and dimensions add (the index is additive under gluing along a non-degenerate irreducible α):

$$\dim \mathcal{M}(M^+;\beta) = \dim \mathcal{M}(M^+;\alpha) + \sigma(\alpha,\beta), \tag{12.127}$$

where $\sigma(\alpha,\beta) = \dim \mathcal{M}(\alpha,\beta)$ is the relative Morse index ((12.52)), i.e. the spectral flow of the curl operator $\star \mathrm{d}_A$. Since $\sigma(\alpha,\beta) \equiv \mu(\alpha) - \mu(\beta) \pmod 8$ (12.59), we get

$$\dim \mathcal{M}(M^+;\alpha) + \mu(\alpha) \equiv \text{a constant} \pmod 8,$$

the constant depending only on M^+. *The limiting flat connections of the zero-dimensional moduli spaces all have the same Floer degree mod 8.* This is why the class we are about to build is homogeneous.

On the closed manifold, the same additivity gives the dimension of the glued moduli space. The general formula for gluing along a flat connection α is

$$\dim \mathcal{M}_k(M) = \dim \mathcal{M}(M^+;\alpha) + \dim \mathcal{M}(M^-;\alpha) + \dim H^0_\alpha + \dim H^1_\alpha, \tag{12.128}$$

where $H^0_\alpha = \operatorname{Ker} \mathrm{d}_\alpha$ on $\Omega^0(\mathfrak{su}(2))$ is the Lie algebra of the stabiliser of α, and $H^1_\alpha = \operatorname{Ker}(\star \mathrm{d}_\alpha)$ on $\operatorname{Ker} \mathrm{d}^*_\alpha$ is the space of flat deformations. The two corrections count the gluing parameters: a rotation of one half relative to the other by an element of the stabiliser, and a flat deformation of the neck.

Worked example 12.50: The two extreme cases: irreducible α and S^3

Irreducible, non-degenerate α. Irreducible means the stabiliser is the centre $\{\pm\mathbb{1}\}$, which is discrete, so $H^0_\alpha = 0$. Non-degenerate means the Hessian $\star \mathrm{d}_\alpha$ has no kernel on $\operatorname{Ker} \mathrm{d}^*_\alpha$, so $H^1_\alpha = 0$ (Section 12.3.3). Then (12.128) reads

$$\dim \mathcal{M}_k(M) = \dim \mathcal{M}(M^+;\alpha) + \dim \mathcal{M}(M^-;\alpha)\,:$$

no gluing parameters. The two halves lock together rigidly.
The connected sum, $N = S^3$. The only flat connection is the trivial one, θ. Its stabiliser is all of SU(2), so $H^0_\theta = \mathfrak{su}(2)$ has dimension 3; and $H^1_\theta = H^1(S^3) \otimes \mathfrak{su}(2) = 0$. Writing

$D_i = \dim \mathcal{M}_{k_i}(M_i)$ for the closed pieces,

$$\dim \mathcal{M}_k(M_1 \# M_2) \overset{(1)}{=} D_1 + D_2 + \dim H^0_\theta + \dim H^1_\theta$$
$$\overset{(2)}{=} D_1 + D_2 + 3.$$

(1) (12.128), with $\mathcal{M}(M_i \setminus B^4; \theta) \cong \mathcal{M}_{k_i}(M_i)$ (an instanton on a punctured manifold with finite action extends over the puncture, by removable singularities).
(2) $\dim \mathfrak{su}(2) = 3$ and $H^1(S^3) = 0$.

This is exactly the count of (12.101): the "+3" there is the SO(3) of gluing parameters, and it is the reason the invariants of a connected sum vanish. In Floer's language the same fact appears as $HF_*(S^3) = 0$, because the reducible θ is not a generator of the Floer complex (Example 12.21).

12.8.3 The Floer classes $[C^\pm]$

Nash writes (12.62): "The boundary conditions allow us to construct two Floer homology classes $[C^\mp] \in HF(N^\mp)$." Here is the construction, in the simplest case where no surfaces are used to cut down ($d = 0$). The general case is in Section 12.8.6.

Counting the rigid instantons on $\widehat{M}^+$

For each generator α of the Floer complex (12.60) of N, consider the moduli spaces $\mathcal{M}(M^+; \alpha)$ that are zero-dimensional. For generic metrics they are finite sets of points, each with a sign ± 1 from an orientation (the same signs as in (12.61)). Let $\#\mathcal{M}(M^+; \alpha)^0$ be the signed count, taken to be 0 when the moduli space is not zero-dimensional. Define the Floer chain

$$\psi^+ := \sum_{\alpha \in \mathcal{R}^*(N)} \#\mathcal{M}(M^+; \alpha)^0 \, \alpha \ \in C_*(N). \tag{12.129}$$

By the dimension formula (12.127), all the α that appear have the same degree, so $\psi^+ \in C_p(N)$ for one $p \in \mathbb{Z}_8$.

Key idea: The rigid instantons form a cycle

$\partial\psi^+ = 0$. So ψ^+ defines a Floer homology class

$$[C^+] := [\psi^+] \in HF_p(N).$$

It does not depend on the metric on M^+ (or on the other choices), because changing them changes ψ^+ by a boundary.

Why $\partial\psi^+ = 0$. This is the same mechanism as $\partial^2 = 0$ in Section 12.5.1: *a compact 1-manifold has an even number of ends, which cancel in pairs when counted with orientation*. Fix a generator β one degree below the α's, and look at the one-dimensional moduli space $\mathcal{M}(M^+; \beta)^1$. It is a 1-manifold, not compact, and its non-compactness is completely described: a sequence of instantons in it can only break by sending energy down the end. The limit is a *broken configuration*: a rigid instanton $u \in \mathcal{M}(M^+; \alpha)^0$ followed by a rigid flow line $\gamma \in \widehat{\mathcal{M}}(\alpha, \beta)$ on $N \times \mathbb{R}$ (Figure 12.34). (Bubbling is excluded in such low dimensions; the dimension count is the one of Section 12.6.5.) Conversely every broken configuration is the end of exactly one

component, by gluing. So

$$\begin{aligned} 0 &\overset{(1)}{=} \#\big(\text{oriented ends of } \mathcal{M}(M^+;\beta)^1\big) \\ &\overset{(2)}{=} \sum_{\alpha} \#\mathcal{M}(M^+;\alpha)^0 \cdot n(\alpha,\beta) \\ &\overset{(3)}{=} \big(\text{coefficient of } \beta \text{ in } \partial\psi^+\big). \end{aligned} \tag{12.130}$$

(1) A compact oriented 1-manifold with boundary is a disjoint union of closed intervals and circles; each interval has one end counted +1 and one counted −1.

(2) The ends are the broken configurations (u,γ), and the sign of an end is the product of the signs of u and γ; for fixed α there are $\#\mathcal{M}(M^+;\alpha)^0$ choices of u and $n(\alpha,\beta)$ (signed) choices of γ.

(3) By (12.62), $\partial\psi^+ = \sum_\alpha \psi^+_\alpha\,\partial\alpha = \sum_\beta\big(\sum_\alpha \psi^+_\alpha n(\alpha,\beta)\big)\beta$.

Every coefficient vanishes, so $\partial\psi^+ = 0$.

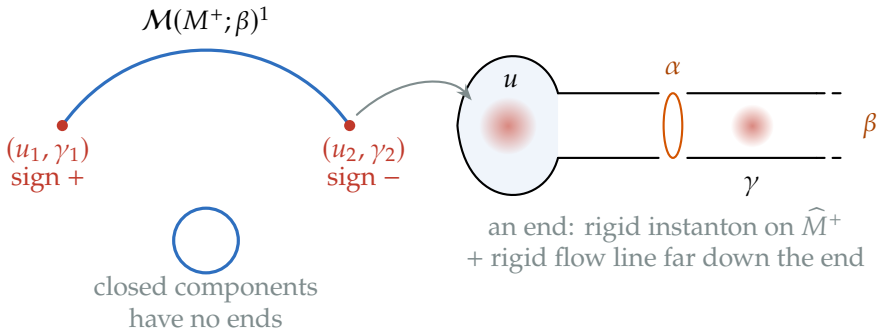


Figure 12.34: Why ψ^+ is a Floer cycle. The one-dimensional moduli space $\mathcal{M}(M^+;\beta)^1$ is a union of arcs and circles. Each end of an arc is a broken configuration: a rigid instanton u on $\widehat{M}^+$ limiting to α, followed by a rigid flow line γ from α to β that has slid infinitely far down the end. The ends of each arc cancel, and the total count of ends is the coefficient of β in $\partial\psi^+$.

Worked example 12.51: A cycle condition in numbers

Suppose N had three generators, α_1,α_2 in degree p and β in degree $p-1$, with instanton counts $n(\alpha_1,\beta)=1$ and $n(\alpha_2,\beta)=2$. The Floer complex in these degrees is

$$C_p = \mathbb{Z}\alpha_1 \oplus \mathbb{Z}\alpha_2 \xrightarrow{\partial=(1\ \ 2)} C_{p-1} = \mathbb{Z}\beta.$$

Then $\operatorname{Ker}\partial$ is spanned by $2\alpha_1-\alpha_2$. The rigid-instanton chain $\psi^+ = a\,\alpha_1 + b\,\alpha_2$ of *any* manifold M^+ bounded by N (with these degrees) is forced by (12.130) to satisfy $a+2b=0$: the instanton counts on the two halves are not independent, they are constrained by the counts on the neck. For instance $\psi^+ = 2\alpha_1-\alpha_2$ is allowed, and $\psi^+=\alpha_1$ is not.

Independence of choices

Change the metric on M^+ along a path g_s, $s\in[0,1]$. The parametrised moduli space $\bigcup_s \mathcal{M}_{g_s}(M^+;\alpha)\times\{s\}$ is one dimension bigger. Its zero-dimensional part consists of isolated instantons that occur at special values of s with limit α' of degree $p+1$; count them to get a chain $\chi\in C_{p+1}(N)$. Counting the ends of the one-dimensional parametrised moduli spaces as above gives

$$\psi^+_{g_1} - \psi^+_{g_0} = \pm\,\partial\chi. \tag{12.131}$$

The ends are now of three kinds: the endpoints at $s = 0$ and $s = 1$ (contributing $\psi^+_{g_1} - \psi^+_{g_0}$) and the broken configurations at intermediate s (contributing $\partial\chi$). So the class $[\psi^+]$ is unchanged. This is the Floer-theoretic version of the metric independence of q_d, and it is the same "count the ends" argument once more.

The other half: a cocycle

On $\widehat{M}^-$ the end points the other way: the connection comes *from* α as $t \to -\infty$. A broken configuration now consists of a flow line γ from β to α *followed* by a rigid instanton on $\widehat{M}^-$ limiting to α. Define

$$\phi^- := \sum_\alpha \#\mathcal{M}(M^-;\alpha)^0 \; \alpha^*,$$

a linear function on $C_*(N)$ ($\alpha^*(\beta) = \delta_{\alpha\beta}$). The count of ends of $\mathcal{M}(M^-;\beta)^1$ gives $\sum_\alpha n(\beta,\alpha)\,\phi^-_\alpha = 0$ for every β, i.e.

$$\phi^- \circ \partial = 0 : \tag{12.132}$$

ϕ^- is a *cocycle*, and defines a class in Floer *cohomology* $HF^*(N)$. Reversing the orientation of N turns the flow of f into the flow of $-f$ and exchanges "down" with "up", so $HF^*(N)$ is the same thing as the Floer homology of $\overline{N}$ up to a shift of the grading (Example 12.24). This is Nash's remark that $N^\pm$ "inherit opposite orientations from M and so the $HF(N^\mp)$ are dual to one another":

$$[C^+] \in HF_*(N), \qquad [C^-] \in HF_*(\overline{N}) \cong HF^*(N) \;\text{ (up to a grading shift).} \tag{12.133}$$

12.8.4 The pairing formula

We can now read Nash's (12.61) and (12.63).

$C^+ \cap C^- = \mathcal{M}_k$

An instanton on the closed manifold M with a very long neck looks, on each half, like an instanton on $\widehat{M}^\pm$, and on the neck like a nearly flat connection $\approx \alpha$. Conversely, if $u^+ \in \mathcal{M}(M^+;\alpha)$ and $u^- \in \mathcal{M}(M^-;\alpha)$ have the *same* limit α, they can be glued across the neck into an instanton on M. So, for a long enough neck,

$$\mathcal{M}_k(M) \;\cong\; \bigcup_\alpha \mathcal{M}(M^+;\alpha) \times \mathcal{M}(M^-;\alpha) \qquad (\text{with } k^+ + k^- = k). \tag{12.134}$$

This is what Nash's (12.61), $C^+ \cap C^- = \mathcal{M}_k$, means: restrict each instanton on M^+ to the neck, and each on M^-; the instantons on M are the pairs whose restrictions *match*. If we think of $C^\pm$ as subsets of the space $\mathcal{B}^*(N)$ of connections on N modulo gauge (the boundary values), then $\mathcal{M}_k$ is literally their intersection. By Example 12.50 there are no extra gluing parameters, because α is irreducible.

The count

Take the case $\dim \mathcal{M}_k(M) = 0$. By (12.128) with $H^0_\alpha = H^1_\alpha = 0$, the only contributions come from α with $\dim \mathcal{M}(M^+;\alpha) + \dim \mathcal{M}(M^-;\alpha) = 0$. Generically a moduli space of negative dimension

is empty, so both must be zero-dimensional. Then

$$\begin{aligned} q_0(M) &\overset{(1)}{=} \#\mathcal{M}_k(M) \\ &\overset{(2)}{=} \sum_\alpha \#\mathcal{M}(M^+;\alpha)^0 \cdot \#\mathcal{M}(M^-;\alpha)^0 \\ &\overset{(3)}{=} \sum_\alpha \phi^-_\alpha \psi^+_\alpha \overset{(4)}{=} \phi^-(\psi^+) \overset{(5)}{=} \langle [C^-], [C^+] \rangle. \end{aligned} \tag{12.135}$$

(1) For $\dim \mathcal{M}_k = 0$ the invariant is the signed count of instantons (Section 12.8.5).
(2) The bijection (12.134), and the sign of a glued instanton is the product of the signs of its halves (orientations multiply under gluing).
(3) Definitions of ψ^+ (12.129) and ϕ^-.
(4) $\alpha^*(\beta) = \delta_{\alpha\beta}$.
(5) ψ^+ is a cycle and ϕ^- a cocycle, and the evaluation of a cocycle on a cycle depends only on their classes (next paragraph).

This is Nash's (12.63), $[C^+]\bullet[C^-] = q_d(M)$, with the pairing $\bullet$ being the evaluation of cohomology on homology, i.e. the Poincaré-duality pairing between $HF_*(N)$ and $HF_*(\overline{N})$.

Why the pairing is well defined on classes. If we change ψ^+ by a boundary and ϕ^- by a coboundary $\eta \circ \partial$,

$$\begin{aligned} (\phi^- + \eta\circ\partial)(\psi^+ + \partial\chi) &\overset{(1)}{=} \phi^-(\psi^+) + \phi^-(\partial\chi) + \eta(\partial\psi^+) + \eta(\partial\partial\chi) \\ &\overset{(2)}{=} \phi^-(\psi^+). \end{aligned}$$

(1) Linearity.
(2) $\phi^- \circ \partial = 0$ (12.132), $\partial\psi^+ = 0$ (12.130), and $\partial^2 = 0$.

This is the same algebra that made the Morse-theoretic pairing between cycles and cocycles well defined in Chapter 7.

Physical picture: The pairing as a transition amplitude

In the Hamiltonian picture of Section 12.5.5, the Floer groups are the zero-energy states of a supersymmetric theory on N. Then $[C^+]$ is the state that the path integral on M^+ prepares on its boundary, $|M^+\rangle \in \mathcal{H}_N$, and $[C^-]$ is the dual state $\langle M^-|$ prepared by the other half. The number of the closed manifold is the overlap,

$$q(M) = \langle M^-|M^+\rangle,$$

exactly as a Euclidean path integral over a spacetime cut along a time slice is $\langle \text{out}|(\text{evolution})|\text{in}\rangle$. With a long neck the evolution operator e^{-TH} projects onto $H = 0$, which is why only Floer homology (the zero-energy states) survives.

Worked example 12.52: Only one degree can contribute

Take $N = \Sigma(2,3,5)$ with $HF_* = \mathbb{Z}\rho_1 \oplus \mathbb{Z}\rho_2$ in degrees 1 and 5 (Example 12.21). Suppose $M = M^+ \cup_{\Sigma(2,3,5)} M^-$ and the rigid instantons on $\widehat{M}^+$ limit to degree-1 connections, with $\psi^+ = a\,\rho_1$ for an integer a. For the rigid instantons of $\widehat{M}^-$ to pair non-trivially with ψ^+ they must also limit to ρ_1: $\phi^- = c\,\rho_1^* + (\text{terms in } \rho_2^*)$. Then

$$q_0(M) \overset{(1)}{=} \phi^-(\psi^+) \overset{(2)}{=} c\,a\,\rho_1^*(\rho_1) = ac.$$

(1) (12.135).
(2) $\rho_2^*(\rho_1) = 0$.

The whole global count has become a product of two local counts, one on each side. If M^+ had only produced instantons ending on ρ_2, the invariant would vanish. (This example only illustrates the bookkeeping; we do not claim a specific M realising given a and c.)

The rest of Nash's discussion is about what this formula means: the simplest case $d = 0$, the two directions in which it can be read, and the generalisation to $d > 0$ with surfaces on both sides. We take these in turn.

12.8.5 The simplest case: $d = 0$ and a signed count of instantons

When is the moduli space a finite set?

The invariant q_d is a polynomial of degree $d = \frac{1}{2}\dim \mathcal{M}_k$ (12.94). The lowest possible degree is $d = 0$: a polynomial of degree zero is a constant, so $q_0(M)$ is a single integer. Nash's (12.64) is the condition for this to happen:

$$\dim \mathcal{M}_k = 8k - 3(1 + b_2^+) = 0. \tag{12.136}$$

Which manifolds allow it? Solve for b_2^+:

$$8k = 3(1 + b_2^+) \overset{(1)}{\Longrightarrow} 8 \mid 3(1 + b_2^+) \overset{(2)}{\Longrightarrow} 8 \mid (1 + b_2^+)$$
$$\overset{(3)}{\Longrightarrow} b_2^+ = 8j - 1, \qquad k = \tfrac{3(1+b_2^+)}{8} = 3j, \qquad j = 1, 2, 3, \ldots$$

(1) $8k$ is divisible by 8.
(2) $\gcd(3, 8) = 1$, so 8 divides the other factor.
(3) Write $1 + b_2^+ = 8j$ and solve for k.

So $b_2^+ \in \{7, 15, 23, \ldots\}$ with charges $k \in \{3, 6, 9, \ldots\}$. As a check, b_2^+ is odd, as it must be for $\dim \mathcal{M}_k$ to be even (Section 12.6.4).

Worked example 12.53: Which surfaces have a q_0?

For a simply connected complex surface $b_2^+ = 2p_g + 1$, where p_g is the number of independent holomorphic 2-forms (Section 12.6.6). The elliptic surfaces $E(n)$ have $p_g = n - 1$, hence

$$b_2^+\big(E(n)\big) = 2(n - 1) + 1 = 2n - 1,$$

which gives $1, 3, 5$ for $E(1) = \mathbb{CP}^2\#9\overline{\mathbb{CP}}^2$, $E(2) = \mathrm{K3}$ and $E(3)$, in agreement with Table 12.1 and Example 12.43. The condition $b_2^+ = 8j - 1$ becomes $2n - 1 = 8j - 1$, i.e. $n = 4j$. So:

manifold	b_2^+	k with $\dim \mathcal{M}_k = 0$	next dimensions
$E(4)$	7	3	$\dim \mathcal{M}_4 = 8$, $\dim \mathcal{M}_5 = 16$
$E(8)$	15	6	$\dim \mathcal{M}_7 = 8, \ldots$
$E(12)$	23	9	$\dim \mathcal{M}_{10} = 8, \ldots$

For $E(4)$ at $k = 3$ the "polynomial" $q_0(E(4))$ is a number: the signed number of charge-3 instantons for a generic metric. By the positivity theorem of Section 12.6.6 (for large enough k) the invariants of algebraic surfaces are non-zero; the number at $k = 3$ is not guaranteed by that theorem, which is why Nash writes "$q_d \neq 0$ for $d \geq d_0$".

The count, and the partition function

When $\dim \mathcal{M}_k = 0$ (and the metric is generic, and there are no reducibles), $\mathcal{M}_k$ is a finite set of points $\{A_1, \dots, A_r\}$; it is compact because bubbling would need 8 dimensions to spare (Section 12.6.5). The invariant is

$$q_0(M) = \sum_{i=1}^{r} \varepsilon_i, \qquad \varepsilon_i = (-1)^{n_i} \in \{\pm 1\}, \tag{12.137}$$

where ε_i is the orientation of $\mathcal{M}_k$ at A_i. Orienting a point means choosing a sign. Here it is the sign of the determinant of the linearised operator $D_{A_i} = \mathrm{d}^*_{A_i} \oplus \mathrm{d}^+_{A_i}$ of (12.79), measured against a reference, and the reference is fixed once and for all by orienting the space $H^2_+(M)$ of self-dual harmonic forms (a **homology orientation**).

From the field-theory side this is the partition function. In Witten's theory a correlation function $\langle P \rangle$ is non-zero only when $U(P) = \dim \mathcal{M}_k$ (12.111). For $\dim \mathcal{M}_k = 0$ the observable is $P = 1$, and $\langle 1 \rangle$ is the partition function Z. Localisation (12.110) sums over the classical minima, and each isolated minimum contributes the ratio of the fermion determinant to the square root of the boson determinant, which is ± 1 because the non-zero eigenvalues pair up. This gives Nash's (12.65):

$$Z = \sum_i (-1)^{n_i}, \qquad n_i \in \{0, 1\}. \tag{12.138}$$

This is the four-dimensional version of the toy model of Example 12.46, $Z = \sum_j \operatorname{sgn} s'(x_j)$ (12.108): there the sign was the sign of the "fermion determinant" $s'(x_j)$, here it is the sign of $\det D_{A_i}$.

Watch out: Why the answer is a signed count, not a count

The unsigned number r of instantons does depend on the metric: as the metric moves, pairs of instantons can be born or annihilate, exactly as a pair of zeros of $s(x)$ in the toy model can appear when a local minimum of s dips below the axis. At such a birth the two new instantons have opposite signs. So r jumps by 2 while $\sum \varepsilon_i$ does not change. Only the signed count is an invariant.

Middle-dimensional intersections

Nash's formula (12.63) at $d = 0$ says that the signed sum (12.137) is also the *intersection pairing* of the two Floer classes, and he concludes that the Floer groups behave like the "middle-dimensional" homology of $\mathcal{A}/\mathcal{G}$. Here is a finite-dimensional model of what this means.

On a closed oriented manifold X of dimension $2m$, two cycles of the middle dimension m generically meet in isolated points. Each point gets a sign: $+1$ if a positive basis of the first cycle's tangent space followed by a positive basis of the second's is a positive basis of TX, and -1 otherwise. The signed count is the **intersection number** $a \cdot b$, which depends only on the homology classes. The intersection form of a 4-manifold (Section 12.6.1) is the case $m = 2$.

Worked example 12.54: Curves on the torus

Take $X = T^2 = \mathbb{R}^2/\mathbb{Z}^2$, $m = 1$. A closed curve of homology class (p, q) winds p times in the x-direction and q times in the y-direction; the straight line of slope q/p represents it. Two

straight curves of classes (p,q) and (r,s) meet in $|ps-qr|$ points, all with the same sign, so

$$(p,q)\cdot(r,s) \overset{(1)}{=} \det\begin{pmatrix} p & r \\ q & s \end{pmatrix} = ps - qr.$$

(1) At each crossing the sign compares the ordered pair of tangent vectors $\big((p,q),(r,s)\big)$ with the standard orientation (∂_x,∂_y), which is the sign of this determinant; the number of crossings is $|\det|$ because the linear map with these columns has degree det on the torus.

For $a=(1,0)$ and $b=(0,1)$ this is 1. Now wiggle a into the curve a' of Figure 12.35: it now crosses b three times, with signs $+,-,+$. The count of points went from 1 to 3, but

$$a'\cdot b = (+1)+(-1)+(+1) = 1 = a\cdot b.$$

The wiggle created a cancelling pair, exactly as in the pitfall above.

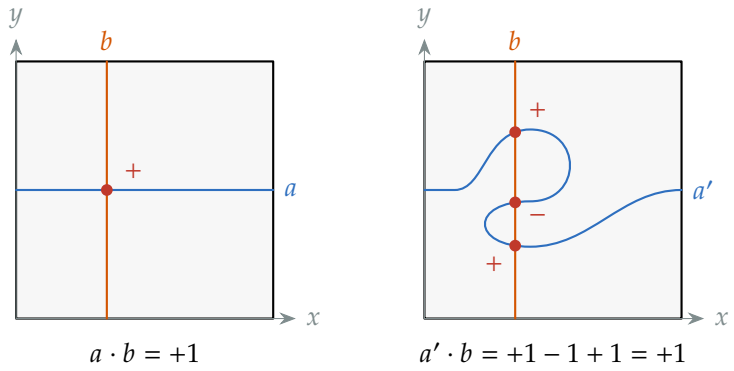


Figure 12.35: The finite-dimensional model of (12.63) at $d=0$. On the torus (opposite sides identified), two middle-dimensional cycles meet in isolated points with signs. Deforming a into a' creates two new crossings of opposite sign; the signed count is unchanged. In gauge theory the cycles are the boundary values $C^{\pm}$ of instantons on the two halves, the ambient space is $\mathcal{A}/\mathcal{G}$ of the gluing surface, and the crossings are the instantons on the closed manifold.

In the gauge-theory problem, C^+ and C^- are sets of boundary values on N, i.e. subsets of $\mathcal{B}^*(N)=\mathcal{A}^*/\mathcal{G}$ of the gluing surface, and $\mathcal{M}_k = C^+\cap C^-$ (12.134). Both $C^{\pm}$ are infinite-dimensional *and* of infinite codimension: roughly, C^+ contains the connections on N that extend as instantons over M^+, which is "half" of all connections, in the sense that the Hessian of f has infinitely many positive and infinitely many negative directions and C^+ is tangent to one half. Ordinary homology cannot measure such "semi-infinite" cycles. Floer homology can: its grading is a *relative* index (Section 12.3.4), exactly what is needed to say which semi-infinite cycles can intersect in points. That is the sense in which $HF_*(N)$ is the middle-dimensional homology of $\mathcal{A}/\mathcal{G}$.

Physical picture: The Atiyah–Floer picture

There is a finite-dimensional version of this which one can actually draw. Casson's construction (Section 12.5.3) splits N along a surface Σ_g into two handlebodies. The flat SU(2) connections on Σ_g form a space $\mathcal{R}(\Sigma_g)$ of dimension $6g-6$, and those extending over each handlebody form two subspaces of dimension $3g-3$: *exactly half*. Their intersection points are the flat connections on N, and half their signed count is the Casson invariant (12.69). The **Atiyah–Floer conjecture** says that the Floer homology of N is the

Floer homology of this pair of half-dimensional (Lagrangian) submanifolds. The same "half-dimensional cycles and their intersections" structure appears one dimension up, in (12.63).

Reading (12.63) in both directions

Nash stresses that $[C^+] \bullet [C^-] = q_d(M)$ carries information both ways.

From Donaldson to Floer. Suppose $q_d(M) \neq 0$ for some d, and $M = M^+ \cup_h M^-$ with $b_2^+(M^\pm) > 0$. Then

$$0 \neq q_d(M) \overset{(1)}{=} \langle [C^-], [C^+] \rangle \quad \Longrightarrow \quad [C^+] \neq 0 \quad \Longrightarrow \quad HF_*(N) \neq 0.$$

(1) Nash's (12.63), in the form (12.135) (and its generalisation (12.142) below for $d > 0$).

So every homology sphere along which one can cut an algebraic surface (into pieces with $b_2^+ > 0$) has non-zero Floer homology. The hypothesis $b_2^+(M^\pm) > 0$ is there for the same reason as on closed manifolds: it keeps reducible instantons out of the moduli spaces of the pieces, so that the relative invariants are defined.

Worked example 12.55: K3 and the Brieskorn sphere $\Sigma(2,3,11)$

By Example 12.48, K3 $= N(2) \cup_{\Sigma(2,3,11)} M(2,3,11)$ with $b_2^+ = 1$ and 2 on the two sides. The Donaldson invariants of K3 are non-zero (it is an algebraic surface; the first one, at $k = 2$, has $\dim \mathcal{M}_2 = 4$ and $d = 2$). So the argument above predicts $HF_*(\Sigma(2,3,11)) \neq 0$. Check it independently with Casson's invariant. For Brieskorn spheres Fintushel and Stern showed $\lambda(\Sigma(p,q,r)) = \frac{1}{8}\tau\big(M(p,q,r)\big)$, the signature of the Milnor fibre over 8. Then

$$\chi_F\big(\Sigma(2,3,11)\big) \overset{(1)}{=} 2\lambda\big(\Sigma(2,3,11)\big) \overset{(2)}{=} 2 \cdot \tfrac{1}{8}\tau\big(M(2,3,11)\big) \overset{(3)}{=} 2 \cdot \tfrac{-16}{8} = -4 \neq 0.$$

(1) Taubes' theorem (12.68).
(2) The Fintushel–Stern formula quoted above.
(3) $\tau(M(2,3,11)) = -16$ from Example 12.48.

A non-zero Euler characteristic forces a non-zero group. (The same formula for $\Sigma(2,3,5)$ gives $\chi_F = 2 \cdot (-8)/8 = -2$, matching Example 12.21.) In fact all four irreducible flat connections of $\Sigma(2,3,11)$ have odd degree, so by Example 12.22 $HF_*(\Sigma(2,3,11))$ is free of rank 4.

From Floer to Donaldson. Suppose instead that the gluing surface is a genuine 3-sphere, so $M^+ \cup_h M^- = \widehat{M}^+ \# \widehat{M}^-$ with $\widehat{M}^\pm$ the closed manifolds obtained by capping off with balls. Then

$$q_d(M) \overset{(1)}{=} \langle [C^-], [C^+] \rangle \overset{(2)}{=} \langle 0, 0 \rangle = 0.$$

(1) Nash's (12.63).
(2) $HF_*(S^3) = 0$ (Example 12.21): there are no irreducible flat connections on S^3, so every Floer class is zero.

This recovers the connected-sum vanishing theorem (12.55) of Section 12.6.6. The two proofs are the same proof: in (12.101) the obstruction was the SO(3) of gluing parameters of the reducible θ; here it is the absence of θ from the Floer complex (Example 12.50).

12.8.6 Relative invariants and the evaluation rule

For $d > 0$ we must cut the moduli spaces down by the divisors V_a, $a \in H_2(M)$ (Section 12.6.4), and it matters on which side of N each surface lies.

Polynomials on $H_2(M^+)$ with values in $HF_*(N)$

The key property is locality: V_a is defined by what the connection does near a surface representing a (Section 12.6.3). If $a \in H_2(M^+)$, pick the surface inside M^+; then V_a only constrains the M^+-half of a glued instanton. So for $a_1, \dots, a_i \in H_2(M^+)$ define, generalising (12.129),

$$q^i(M^+)(a_1, \dots, a_i) := \sum_{\alpha} \#\big(\mathcal{M}(M^+;\alpha) \cap V_{a_1} \cap \cdots \cap V_{a_i}\big)\, \alpha, \tag{12.139}$$

where only the α with $\dim \mathcal{M}(M^+;\alpha) = 2i$ contribute (each divisor cuts down 2 dimensions, so the intersection is a finite set of signed points). The same "count the ends" argument as in (12.130), now applied to the one-dimensional cut-down moduli spaces, shows that (12.139) is a cycle. Its class

$$q^i(M^+)(a_1, \dots, a_i) \in HF_*(N)$$

is symmetric and multilinear in the a's. These are Nash's **relative Donaldson invariants**. They are polynomials in the variables (12.68)

$$\underbrace{H_2(M^+) \times \cdots \times H_2(M^+)}_{i \text{ factors}},$$

but their values are *not* numbers: they are Floer homology classes of the boundary. This is the precise meaning of Nash's statement that "the relative Donaldson invariants for a non-closed manifold are not integer valued, but rather take values in the Floer groups of its boundary". The $d = 0$ class $[C^+]$ of Section 12.8.3 is $q^0(M^+)$. Similarly $q^j(M^-)$ takes values in the dual group $HF^*(N)$.

Symmetric powers split: (12.66)–(12.67)

A symmetric multilinear function of d arguments in $H = H_2(M)$ is the same as a linear function on the **symmetric power** $\mathrm{Sym}^d H$. Concretely: if $x_1, \dots, x_n$ are coordinates on H (a basis of the dual), Sym^d is the space of homogeneous polynomials of degree d in the x's, with basis the monomials $x_1^{e_1} \cdots x_n^{e_n}$, $e_1 + \cdots + e_n = d$. Counting monomials ("d stars and $n-1$ bars"):

$$\dim \mathrm{Sym}^d(\mathbb{R}^n) = \binom{n+d-1}{d}. \tag{12.140}$$

Nash applies the splitting (12.119) $H = V \oplus W$, with $V = H_2(M^+)$ and $W = H_2(M^-)$, to the d-fold product (12.66). Every monomial in the coordinates x on V and y on W factors uniquely as (a monomial of degree i in x)×(a monomial of degree $d-i$ in y). So (Nash's (12.67))

$$\mathrm{Sym}^d(V \oplus W) = \bigoplus_{i=0}^{d} \mathrm{Sym}^i(V) \otimes \mathrm{Sym}^{d-i}(W). \tag{12.141}$$

Nash writes the sum as $\sum_i \mathrm{Sym}^i \mathrm{Sym}^{d-i}$; the product of the two factors is a tensor product.

Worked example 12.56: Checking the splitting by counting

Take $V = \mathbb{R}^2$ with coordinates x_1, x_2, $W = \mathbb{R}$ with coordinate y, and $d = 2$. The left side of (12.141) has the basis $x_1^2,\ x_1x_2,\ x_2^2,\ x_1y,\ x_2y,\ y^2$, six monomials, in agreement with (12.140): $\binom{3+2-1}{2} = \binom{4}{2} = 6$. Sorting them by their degree i in the x's:

$$\begin{aligned}\dim \mathrm{Sym}^2(V\oplus W) &\overset{(1)}{=} \dim\big(\mathrm{Sym}^2V\otimes\mathrm{Sym}^0W\big) + \dim\big(\mathrm{Sym}^1V\otimes\mathrm{Sym}^1W\big)\\ &\quad + \dim\big(\mathrm{Sym}^0V\otimes\mathrm{Sym}^2W\big)\\ &\overset{(2)}{=} 3\cdot 1 + 2\cdot 1 + 1\cdot 1 = 6.\end{aligned}$$

(1) (12.141); dim of a direct sum adds, of a tensor product multiplies.
(2) $\{x_1^2, x_1x_2, x_2^2\}$, $\{x_1y, x_2y\}$, $\{y^2\}$, with (12.140) for each factor.

In general the count is the identity $\binom{m+n+d-1}{d} = \sum_i \binom{m+i-1}{i}\binom{n+d-i-1}{d-i}$ (a form of Vandermonde's identity), which is (12.141) with $\dim V = m$, $\dim W = n$.

The evaluation rule (12.69)

Now evaluate $q_d(M)$ on a class $x = x^+ + x^-$ with $x^\pm \in H_2(M^\pm)$. As a homogeneous polynomial, $q_d(M)(x)$ means the symmetric multilinear form with all d arguments equal to x. Then

$$\begin{aligned}q_d(M)(x^+ + x^-) &\overset{(1)}{=} \sum_{S\subset\{1,\dots,d\}} q_d(M)\big(y_1^S,\dots,y_d^S\big), \qquad y_j^S = \begin{cases} x^+, & j\in S\\ x^-, & j\notin S\end{cases}\\ &\overset{(2)}{=} \sum_{i=0}^{d}\binom{d}{i}\, q_d(M)(\underbrace{x^+,\dots,x^+}_{i},\underbrace{x^-,\dots,x^-}_{d-i})\\ &\overset{(3)}{=} \sum_{i=0}^{d}\binom{d}{i}\Big\langle q^{d-i}(M^-)(x^-,\dots,x^-),\ q^i(M^+)(x^+,\dots,x^+)\Big\rangle. \end{aligned}\tag{12.142}$$

(1) Multilinearity: expand each of the d slots into its x^+ and x^- parts, giving 2^d terms, one for each subset S of slots filled with x^+.
(2) Symmetry: the term depends only on $i = |S|$, and there are $\binom{d}{i}$ subsets of size i.
(3) The gluing argument of Section 12.8.4, with the divisors for x^+ inside M^+ and those for x^- inside M^- (locality); the cut-down moduli spaces glue exactly as the zero-dimensional ones did in (12.135).

This is Nash's (12.69), $q_d(M) = \sum_{i=0}^{d} q_d^i(M^+)\bullet q_d^{d-i}(M^-)$: each term of the splitting (12.141) of the symmetric power is computed as a Floer pairing of a relative invariant of one side with a relative invariant of the other.

Watch out: Binomial coefficients and degrees

Nash's (12.69) has no binomial coefficients. That is a matter of normalisation: if the relative invariants are regarded as elements of $\mathrm{Sym}^i(V)^*\otimes HF$ (as polynomials rather than multilinear forms), the $\binom{d}{i}$ is absorbed. Also, not every term can be non-zero: $q^i(M^+)$ lands in one degree of $HF_*(N)$ and $q^{d-i}(M^-)$ in one degree of $HF^*(N)$, determined by the dimension formula (12.127), and the pairing vanishes unless these degrees match.

Worked example 12.57: A degree-2 invariant cut in half

Suppose $H_2(M^+) = \mathbb{Z}x$ and $H_2(M^-) = \mathbb{Z}y$ (one generator each) and $d = 2$. For a general class $ax + by$,

$$q_2(M)(ax+by) \overset{(1)}{=} a^2 \left\langle q^0(M^-), q^2(M^+)(x,x)\right\rangle + 2ab \left\langle q^1(M^-)(y), q^1(M^+)(x)\right\rangle + b^2 \left\langle q^2(M^-)(y,y), q^0(M^+)\right\rangle.$$

(1) (12.142) with $x^+ = ax$, $x^- = by$, $d = 2$, and multilinearity to pull out the powers of a and b.

The three coefficients of the quadratic form $q_2(M)$ are three separate Floer pairings. The a^2 coefficient uses only moduli spaces on M^+ cut down by two copies of V_x, glued to rigid instantons on M^-; the ab coefficient cuts down once on each side; and so on. If the Floer degrees of $q^1(M^\pm)$ do not match, the mixed coefficient is forced to be zero, and $q_2(M)$ has no ab term: the cutting has told us something about the shape of the polynomial.

12.8.7 The Floer–Donaldson theory as a TQFT

We can now put the chapter's opening axioms (Section 12.1.3) next to what Floer and Donaldson built. The comparison is the real content of the phrase "Donaldson theory is a topological quantum field theory".

Dictionary: Atiyah's axioms and the Floer–Donaldson theory

Axiom / TQFT datum	Floer–Donaldson realisation
space of states $Z(N)$ of a closed 3-manifold	$HF_*(N;\mathbb{Q})$, N a homology 3-sphere
$Z(\overline{N}) = Z(N)^*$	$HF_*(\overline{N}) \cong HF^*(N)$ (grading shifted) (12.133)
vector $Z(X) \in Z(\partial X)$	relative invariants $q^i(X) \in HF_*(\partial X)$ (12.139)
number $Z(M)$ of a closed 4-manifold	$q_d(M)$ (or its generating function)
gluing $Z(M^+ \cup M^-) = \langle Z(M^-), Z(M^+)\rangle$	Nash's (12.63), (12.135), (12.142)
cobordism $W : N_1 \to N_2$ gives $Z(W) : Z(N_1) \to Z(N_2)$	count rigid instantons on W with two cylindrical ends
$Z(N \times [0,1]) = \mathrm{id}$	only constant flow lines are rigid (Example 12.58)
$Z(N \times S^1) = \dim Z(N)$	$\chi_F(N) = 2\lambda(N)$ (Example 12.59)

Cobordisms act on Floer groups

Let W be a compact oriented 4-manifold with $\partial W = \overline{N_1} \sqcup N_2$: a cobordism from N_1 to N_2. Attach an incoming end $N_1 \times (-\infty, 0]$ and an outgoing end $N_2 \times [0, \infty)$. For flat connections α_1 on N_1 and α_2 on N_2 let $\mathcal{M}(W; \alpha_1, \alpha_2)$ be the finite-action instantons with these limits, and define

$$Z(W)\alpha_1 := \sum_{\alpha_2} \#\mathcal{M}(W;\alpha_1,\alpha_2)^0 \, \alpha_2. \tag{12.143}$$

Counting the ends of the one-dimensional moduli spaces now gives broken configurations of two kinds, a flow line on the incoming end or one on the outgoing end, and the count says $Z(W)\partial = \partial Z(W)$: $Z(W)$ is a chain map and so acts on HF_*. Gluing two cobordisms along N_2 gives, by the same argument as (12.134), $Z(W_2 \cup_{N_2} W_1) = Z(W_2) \circ Z(W_1)$ on homology (Figure 12.36). A manifold M^+ with $\partial M^+ = N$ is a cobordism from the empty set to N, and $Z(M^+)$ applied to $1 \in Z(\varnothing) = \mathbb{Q}$ is the class $[C^+]$.

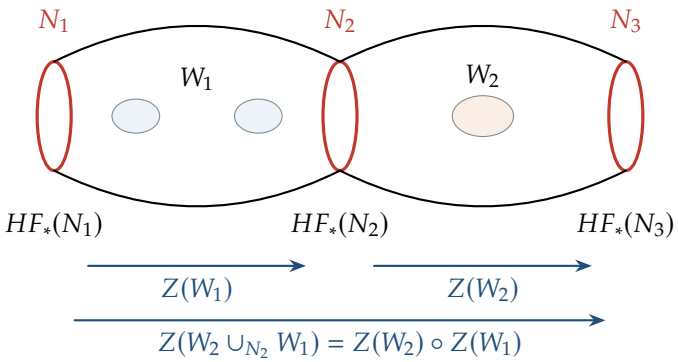


Figure 12.36: Cobordisms act on Floer groups. A 4-manifold W_1 from N_1 to N_2 induces a map $HF_*(N_1) \to HF_*(N_2)$ by counting rigid instantons with prescribed limits on its two ends; gluing cobordisms composes the maps. This is Atiyah's gluing axiom realised by instantons, and (12.63) is the special case in which the two outer boundaries are empty.

Worked example 12.58: The cylinder acts as the identity

Take $W = N \times [0,1]$ with its ends attached: this is the infinite cylinder $N \times \mathbb{R}$. The instantons on it with limits α and β are the Chern–Simons flow lines from α to β, a space of dimension $\sigma(\alpha,\beta)$ (12.52). Time translation $t \mapsto t + c$ maps flow lines to flow lines. On a non-constant flow line it acts freely, so a non-constant flow line comes in a family of dimension ≥ 1 and is never rigid. The only rigid instantons are the constant ones, $A(t) \equiv \alpha$, which exist only for $\beta = \alpha$ and count +1. Hence

$$Z(N \times \mathbb{R})\,\alpha \overset{(1)}{=} \sum_{\beta} \#\mathcal{M}(\alpha,\beta)^0\,\beta \overset{(2)}{=} \alpha.$$

(1) Definition (12.143).
(2) Only $\beta = \alpha$ contributes, with the constant flow line counted once.

So $Z(N \times \mathbb{R}) = \mathrm{id}$, which is axiom (A5) of Section 12.1.3. The non-rigid flow lines are not wasted: after dividing by translations they are what the boundary operator ∂ counts.

Worked example 12.59: The partition function on $N \times S^1$ is Casson's invariant

In a TQFT, closing the cylinder into a circle takes a trace, $Z(N \times S^1) = \mathrm{Tr}_{Z(N)}\,\mathrm{id} = \dim Z(N)$ (12.3). When the space of states is graded into bosons and fermions, as the Floer groups are (even and odd degrees, following the Morse-theory dictionary of Section 12.2.1), the path integral on a circle with periodic boundary conditions for the fermions computes the *supertrace* $\mathrm{Tr}(-1)^F$, the Witten index. So Witten's theory predicts

$$Z(N \times S^1) \overset{(1)}{=} \mathrm{Tr}_{HF_*(N)}(-1)^F \overset{(2)}{=} \sum_{p=0}^{7} (-1)^p \,\mathrm{rank}\, HF_p(N) \overset{(3)}{=} \chi_F(N) \overset{(4)}{=} 2\lambda(N).$$

(1) The trace axiom, with the fermion-parity sign that periodic fermions produce.
(2) $(-1)^F = (-1)^p$ on HF_p.
(3) Definition of the Floer Euler characteristic (Section 12.5.3).
(4) Taubes' theorem (12.68).

For the Poincaré sphere this is $2 \cdot (-1) = -2$, the two fermionic ground states ρ_1, ρ_2 of Example 12.21. This is a physicist's statement: $N \times S^1$ has $b_2^+ = 0$ (its H_2 is $H_2(N) \oplus H_1(N) \otimes$

$H_1(S^1) = 0$), so it has no Donaldson polynomials in the sense of Section 12.6. The field theory nevertheless assigns it a number, and the number it assigns is Casson's invariant, a result of Taubes that the TQFT picture makes natural.

Watch out: A partial TQFT

The Floer–Donaldson theory satisfies Atiyah's axioms only partially. The 3-manifolds are restricted (homology spheres, so that the only reducible flat connection is θ and it can be discarded); the 4-manifolds need $b_2^+ > 0$ on each piece (to avoid reducible instantons) and $b_2^+ > 1$ for metric independence of the closed invariants; the grading is only mod 8; and the vector "$Z(M)$" of a closed manifold is really a whole polynomial (or a generating function of all the q_d) rather than one number. What survives is the essential structure: local pieces, states on the cuts, and gluing by pairing.

12.8.8 What kind of invariants are the q_d?

Nash closes the section with a list of properties. Each deserves an explanation.

There are infinitely many, with relations

For each k with $8k - 3(1 + b_2^+) \geq 0$ we get one polynomial q_d, with $d = 4k - \frac{3}{2}(1 + b_2^+)$ (12.94); as k grows, d grows by 4 each time. So there are infinitely many q_d for each M. Nash remarks that there may be relations among them. There are, and for a large class of manifolds (those of **simple type**) they are strong enough to determine every q_d from finitely many numbers. The result, due to Kronheimer and Mrowka, packages all the invariants into a generating function of the form

$$\mathbb{D}_M(x) = e^{Q(x)/2} \sum_r c_r \, e^{K_r \cdot x},$$

with finitely many *basic classes* $K_r \in H^2(M;\mathbb{Z})$ and rational coefficients c_r. For K3 there is a single basic class $K = 0$, and the formula says that every q_d(K3) is, up to normalisation, a power of the intersection form. In particular the first one, at $d = 2$, is proportional to $Q(x) = x \cdot x$ itself (Donaldson's computation). The basic classes turned out to be the Seiberg–Witten classes, which is where the physics of this chapter eventually led.

They do not depend on the metric

Nash: "the Donaldson invariants are independent of the metric required to define the instantons". The physics argument was Q-exactness (Section 12.7.3). The mathematical argument is Stokes' theorem on a cobordism of moduli spaces, and it is worth writing out. Join two generic metrics by a path g_t, $t \in [0,1]$, and form the parametrised moduli space

$$\mathcal{W} = \bigcup_{t \in [0,1]} \overline{\mathcal{M}}_k(g_t) \times \{t\},$$

of dimension $2d + 1$. For a generic path it is a manifold with boundary

$$\partial \mathcal{W} = \overline{\mathcal{M}}_k(g_1) \sqcup \overline{\overline{\mathcal{M}}_k(g_0)}$$

(the bar on the second copy meaning reversed orientation), provided no reducible connection is met along the path (Figure 12.37). The classes $m(a)$ are defined on the whole space of

connections, so they are closed 2-forms $\mu(a)$ on $\mathcal{W}$ restricting to $m(a)$ on each end. Then

$$\begin{aligned} q_d(g_1) - q_d(g_0) &\overset{(1)}{=} \int_{\overline{\mathcal{M}}_k(g_1)} \mu(a_1) \wedge \cdots \wedge \mu(a_d) - \int_{\overline{\mathcal{M}}_k(g_0)} \mu(a_1) \wedge \cdots \wedge \mu(a_d) \\ &\overset{(2)}{=} \int_{\partial \mathcal{W}} \mu(a_1) \wedge \cdots \wedge \mu(a_d) \\ &\overset{(3)}{=} \int_{\mathcal{W}} \mathrm{d}\big(\mu(a_1) \wedge \cdots \wedge \mu(a_d)\big) \overset{(4)}{=} 0. \end{aligned} \tag{12.144}$$

(1) Definition (12.95) for each metric.
(2) $\partial\mathcal{W}$ is the g_1 end minus the g_0 end (outward normal first: at $t = 1$ it is $+\partial_t$, at $t = 0$ it is $-\partial_t$).
(3) Stokes' theorem on the $(2d + 1)$-dimensional $\mathcal{W}$.
(4) Each $\mu(a_i)$ is closed, so their wedge product is closed.

This is the most basic use of Stokes' theorem: an exact form integrates to zero over something with no boundary, and a difference of two boundary integrals is an integral of d of something closed. (The compactification adds lower strata of codimension at least 2, which do not affect Stokes' theorem; this is part of the analysis of Section 12.6.5.)

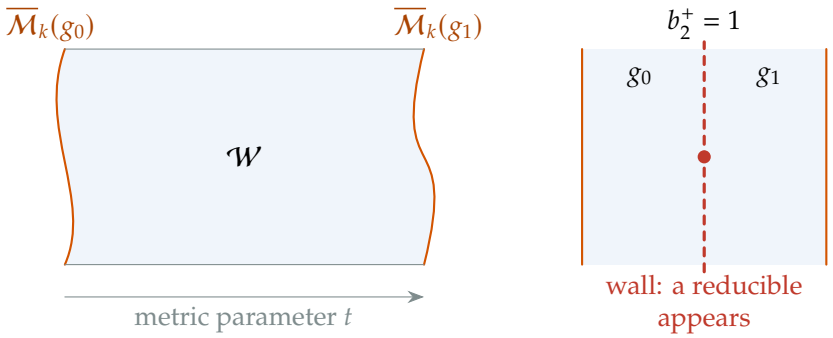


Figure 12.37: Metric independence by Stokes' theorem. Left: as the metric moves from g_0 to g_1 the moduli spaces sweep out a manifold $\mathcal{W}$ whose boundary is the two ends, so the two integrals agree. Right: reducible connections occur in codimension b_2^+ in the space of metrics; when $b_2^+ = 1$ a path of metrics can be forced across such a wall, $\mathcal{W}$ acquires an extra boundary there, and the invariant jumps (compare Figure 12.21).

They are invariant under orientation-preserving diffeomorphisms, up to sign

Let $\varphi : M \to M'$ be an orientation-preserving diffeomorphism and g' a metric on M'. Pull back the metric to $g = \varphi^* g'$. Then φ is an isometry $(M, g) \to (M', g')$, and pulling back connections $A' \mapsto \varphi^* A'$ maps g'-instantons to g-instantons (the ASD condition only uses the metric and the orientation, and both are preserved). So $\varphi^* : \mathcal{M}_k(M', g') \to \mathcal{M}_k(M, g)$ is a diffeomorphism, and it carries the divisor of $\varphi_* a$ to the divisor of a. Therefore

$$\begin{aligned} q_d(M')(\varphi_* a_1, \ldots, \varphi_* a_d) &\overset{(1)}{=} \int_{\mathcal{M}_k(M', g')} m(\varphi_* a_1) \wedge \cdots \wedge m(\varphi_* a_d) \\ &\overset{(2)}{=} \pm \int_{\mathcal{M}_k(M, g)} m(a_1) \wedge \cdots \wedge m(a_d) \overset{(3)}{=} \pm q_d(M)(a_1, \ldots, a_d). \end{aligned}$$

(1) Definition (12.95) on M'.

(2) Change of variables under the diffeomorphism φ^* of moduli spaces, which pulls $m(\varphi_* a)$ back to $m(a)$; the sign is $+1$ if φ^* preserves the orientations of the moduli spaces and -1 otherwise.
(3) Metric independence on M.

The orientation of $\mathcal{M}_k$ is fixed by the homology orientation (an orientation of H^2_+). So the sign is $+1$ or -1 according as φ preserves or reverses the chosen orientation of $H^2_+(M)$: it is $\det(\varphi^*|_{H^2_+})$.

Worked example 12.60: Complex conjugation on $\mathbb{CP}^2$

Let $c : \mathbb{CP}^2 \to \mathbb{CP}^2$ be complex conjugation of the homogeneous coordinates. In local coordinates $(z_1, z_2) = (x_1 + iy_1, x_2 + iy_2)$ it is $(x_1, y_1, x_2, y_2) \mapsto (x_1, -y_1, x_2, -y_2)$, with Jacobian determinant $(+1)(-1)(+1)(-1) = +1$: orientation-preserving. It maps the line $L = \mathbb{CP}^1$ to itself reversing its orientation, so $c_*[L] = -[L]$; and it is antiholomorphic, so it sends the Kähler form ω, which spans $H^2_+(\mathbb{CP}^2)$, to $-\omega$. The sign in the rule above is therefore -1, and the rule predicts

$$q_d(c_* a) = -q_d(a).$$

Check it against homogeneity. With $a = [L]$, $c_* a = -a$ and

$$q_d(c_* a) \overset{(1)}{=} q_d(-a) \overset{(2)}{=} (-1)^d q_d(a) \overset{(3)}{=} -q_d(a).$$

(1) $c_*[L] = -[L]$.
(2) q_d is homogeneous of degree d.
(3) For $\mathbb{CP}^2$, $b_2^+ = 1$ and $2d = 8k - 6$, so $d = 4k - 3$ is odd.

The two computations agree: the sign from the homology orientation is exactly what the parity of d requires. (Conjugation is an isometry of the Fubini–Study metric, so the chamber subtleties of $b_2^+ = 1$ do not enter.)

No relation under orientation reversal

Reversing the orientation of M gives $\overline{M}$ with

$$Q(\overline{M}) = -Q(M), \qquad b_2^+(\overline{M}) = b_2^-(M), \qquad b_2^-(\overline{M}) = b_2^+(M),$$

and the Hodge star changes sign, so anti-self-dual on $\overline{M}$ means self-dual on M. The instantons of $\overline{M}$ are the anti-instantons of M: a different equation with differently sized moduli spaces, $\dim \mathcal{M}_k(\overline{M}) = 8k - 3(1 + b_2^-(M))$. There is no reason for the two sets of invariants to be related, and in general they are not.

Worked example 12.61: K3 and $\mathbb{CP}^2$ with both orientations

For K3, $(b_2^+, b_2^-) = (3, 19)$. Then

$$\dim \mathcal{M}_k(\mathrm{K3}) = 8k - 12, \qquad \dim \mathcal{M}_k(\overline{\mathrm{K3}}) = 8k - 3(1 + 19) = 8k - 60.$$

The first invariant of K3 appears at $k = 2$, that of $\overline{\mathrm{K3}}$ at $k = 8$: they are unrelated polynomials. For $\mathbb{CP}^2$ the contrast is even starker. $\overline{\mathbb{CP}}^2$ has $b_2^+ = 0$, so $\dim \mathcal{M}_k = 8k - 3$ is odd and $\overline{\mathbb{CP}}^2$ has no Donaldson polynomials at all, while $\mathbb{CP}^2$ ($b_2^+ = 1$) has them in every dimension $8k - 6$.

Differential, not topological, invariants

The last point is the reason anyone cares. The ASD equation involves derivatives, so it needs a *smooth* structure; a mere homeomorphism cannot transport instantons from one manifold to another. So the q_d are invariants of the smooth manifold, and two manifolds that are homeomorphic (same Q, by Freedman's theorem, in the simply connected case) may have different q_d. When that happens they are not diffeomorphic. We saw one such pair in Example 12.43: $E(3)$ and $5\mathbb{CP}^2\#29\overline{\mathbb{CP}}^2$. Nash cites Ebeling, who used the q_d to show that certain pairs of complex surfaces (complete intersections in projective space) are homeomorphic but not diffeomorphic. In each case the mechanism is the contrast of Section 12.6.6: one member has non-zero invariants (positivity), the other has them all equal to zero (a splitting along S^3, or along a homology sphere with trivial Floer groups), or both are non-zero but different polynomials.

Reading Nash (pp. 336–339, (12.60)–(12.69))

Nash says: cut $M = M^+ \cup_h M^-$ along homology spheres (12.60), which is possible when Q splits with $b_2^+(M^\pm) > 0$ (Freedman–Taylor). With boundary conditions given by connections on $N^\pm$, the instanton sets $C^\pm$ on the two halves satisfy $C^+ \cap C^- = \mathcal{M}_k$ (12.61) and define Floer classes $[C^\mp] \in HF(N^\mp)$ (12.62); since $N^\pm$ have opposite orientations the groups are dual, and Donaldson shows that $[C^+] \bullet [C^-] = q_d(M)$ (12.63). For $d = 0$ one needs $8k - 3(1 + b_2^+) = 0$ (12.64); $\mathcal{M}_k$ is a finite set, q_0 a signed count, and in the field theory it is the partition function $Z = \sum_i (-1)^{n_i}$ (12.65). The Floer groups are like the middle-dimensional homology of $\mathcal{A}/\mathcal{G}$. Read one way, (12.63) shows that the gluing spheres of algebraic surfaces have non-trivial HF; read the other way, gluing along S^3 forces $q_d = 0$, recovering (12.55). Relative invariants take values in Floer groups; splitting the product (12.66) and the symmetric powers (12.67) leads to polynomials $q_d^i(M^+)$ in the variables (12.68) and the evaluation rule (12.69). Finally the q_d are infinite in number, metric independent, invariant up to sign under orientation-preserving diffeomorphisms, unrelated under orientation reversal, and differential rather than topological (Ebeling). We have supplied: the boundary-orientation rule; the Mayer–Vietoris proof that H_2 and Q split; the nucleus/Milnor-fibre splitting of $E(n)$ and K3 with explicit forms; the cylindrical-end moduli spaces and the flat limit; the fractional charge $k^\pm \equiv \mp f(\alpha)$; the gluing dimension formula with the H^0, H^1 corrections and its S^3 case; the construction of $[C^+]$ as a cycle by counting ends; the cocycle $[C^-]$; the derivation of the pairing and its well-definedness; the $b_2^+ = 8j - 1$ analysis of (12.64); the torus model of middle-dimensional intersections; the Casson check for $\Sigma(2,3,11)$; the binomial form of (12.69); the dictionary with Atiyah's axioms, the cylinder, and $Z(N \times S^1) = 2\lambda$; and the Stokes proof of metric independence, the diffeomorphism sign with the $\mathbb{CP}^2$ check, and the orientation-reversal examples.

Recap

- Cutting M along a homology sphere N splits H_2 and Q (Mayer–Vietoris, using $H_1(N) = H_2(N) = 0$). Every splitting of Q is realised this way (Freedman–Taylor), though not always along S^3.
- On a half with a cylindrical end an instanton is an instanton on $M^\pm$ with a Chern–Simons flow line attached; finite action makes it converge to a flat connection α, and its charge is $\mp f(\alpha)$ mod 1.
- Counting rigid instantons on $\widehat{M}^+$ by their limits gives a Floer cycle ψ^+ (the ends of 1-dimensional moduli spaces cancel), hence a class $[C^+] \in HF_*(N)$; the other half

gives a cocycle, i.e. a class of $\overline{N}$.

- Gluing along irreducible α has no gluing parameters, so $q_d(M) = \langle [C^-], [C^+] \rangle$. Along S^3 the only flat connection is the reducible θ: $HF(S^3) = 0$, and $q_d = 0$.
- For $d = 0$ ($b_2^+ = 8j - 1$, $k = 3j$) the invariant is the signed count of instantons, the partition function $Z = \sum (-1)^{n_i}$ of Witten's theory, and an intersection number of "middle-dimensional" cycles.
- For $d > 0$ the relative invariants $q^i(M^+)$ are HF-valued polynomials; $\mathrm{Sym}^d(V \oplus W) = \bigoplus \mathrm{Sym}^i V \otimes \mathrm{Sym}^{d-i} W$ and $q_d(x^+ + x^-) = \sum \binom{d}{i} \langle q^{d-i}(M^-), q^i(M^+) \rangle$.
- This is Atiyah's axiom system in $3 + 1$ dimensions: $Z(N) = HF_*(N)$, the cylinder is the identity, $Z(N \times S^1) = \chi_F = 2\lambda$.
- The q_d are metric independent (Stokes on a cobordism of moduli spaces, for $b_2^+ > 1$), diffeomorphism invariant up to the sign $\det \varphi^*|_{H^2_+}$, unrelated under orientation reversal, and sensitive to the smooth structure.

The Floer groups entered this story as the space of states of a three-dimensional slice. The same question can be asked of the three-dimensional Chern–Simons theory whose action is the function f of Section 12.2: quantise it, and ask what its states on a surface and its path integral over a 3-manifold compute. The answer involves knots, and that is where the chapter goes next.

12.9 Knots, links and when two of them are the same

Here is the question that drives the next two sections. A Wilson loop is the trace of the holonomy of a gauge field around a closed curve C in space-time. In an ordinary gauge theory its expectation value depends on the length of C, the area it spans, the metric, and so on. In a *topological* field theory in three dimensions, such as the Chern–Simons theory of Chapter 12, nothing depends on the metric. So $\langle W(C) \rangle$ can only depend on how C sits inside the three-manifold, up to smooth deformation. It depends only on the *knot type* of C. If we have several loops, it depends on how they are tangled with one another: on the *link type*.

So the correlation functions of a topological gauge theory in three dimensions are *knot invariants*. To recognise what the path integral computes, we first need to know what knots are, when two of them count as "the same", and which invariants mathematicians already had. That is the content of Nash's §4. He gives it in three pages. We take it slowly, because everything in the Chern–Simons calculation of Sections 12.12 and 12.13 is checked against these invariants.

The physicist has met knot theory before, perhaps without the name. Gauss invented the linking number while doing magnetostatics (Section 12.10). Kelvin's "vortex atom" theory of 1867 proposed that atoms are knotted vortex tubes in the ether, and it was Tait's attempt to tabulate atoms that produced the first knot tables. A charged particle going round a solenoid feels the Aharonov–Bohm phase: this is the simplest knot invariant, a winding number, measured by a phase.

12.9.1 What a knot is

Working definition 12.14 (knot, link)

A **knot** is a smooth closed curve in $\mathbb{R}^3$ that does not cross itself. Concretely, it is a smooth map

$$\gamma : [0, 2\pi] \to \mathbb{R}^3, \qquad \gamma(0) = \gamma(2\pi),$$

with $\gamma'(t) \neq 0$ everywhere [the point never stops moving, so the curve has no cusps] and $\gamma(s) \neq \gamma(t)$ for $s \neq t$ in $[0, 2\pi)$. We usually identify the knot with its image $K = \gamma([0, 2\pi])$. A **link** with p **components** is p such curves that are disjoint from one another. A knot is a link with one component.
Operational test: to specify a knot, write down a periodic parametrisation and check that it never revisits a point.
Formal version: a knot is an embedding $S^1 \hookrightarrow \mathbb{R}^3$ (or into $S^3 = \mathbb{R}^3 \cup \{\infty\}$).

Nash works in $\mathbb{R}^3$ or its one-point compactification S^3. For knots the two are interchangeable: a knot is compact, so it misses a neighbourhood of ∞, and every deformation in S^3 can be pushed off the point ∞. The reason to prefer S^3 is that it is a *closed* three-manifold, which is where Chern–Simons theory is defined without boundary terms.

Worked example 12.62: Three explicit knots

1. The **unknot**: $\gamma(t) = (\cos t, \sin t, 0)$, the round circle.

2. The **trefoil** (Nash's "clover leaf"):

$$\gamma(t) = \big(\sin t + 2\sin 2t,\ \cos t - 2\cos 2t,\ \sin 3t\big). \tag{12.145}$$

   Its projection to the xy-plane is the three-lobed curve of Figure 12.38. The height $z = \sin 3t$ decides, at each of the three points where the projection crosses itself, which strand is on top.

3. The **torus knots**. Draw a doughnut of core radius 2 and tube radius 1. The curve that winds p times around the z-axis while winding q times around the tube,

$$\gamma_{p,q}(t) = \big((2 + \cos qt)\cos pt,\ (2 + \cos qt)\sin pt,\ -\sin qt\big), \tag{12.146}$$

   is the torus knot $T(p, q)$ when $\gcd(p, q) = 1$. $T(2, 3)$ is the trefoil again. When $\gcd(p, q) = d > 1$ the curve closes up after $1/d$ of the way round and the formula gives a link $T(p, q)$ with d components (shift t by $2\pi j/q$ for the j-th one).

Why is $\gamma_{p,q}$ injective when $\gcd(p, q) = 1$? Suppose $\gamma_{p,q}(s) = \gamma_{p,q}(t)$. The point lies on the torus, and $(pt \bmod 2\pi,\ qt \bmod 2\pi)$ are its two angles on the torus. So $p(s - t) \in 2\pi\mathbb{Z}$ and $q(s - t) \in 2\pi\mathbb{Z}$. With $\gcd(p, q) = 1$ there are integers a, b with $ap + bq = 1$, so $s - t = a\,p(s - t) + b\,q(s - t) \in 2\pi\mathbb{Z}$, and $s = t$ on the circle.

The first thing Nash stresses is this.

Key idea: The knottedness is in the embedding

As topological spaces all knots are the same: each is homeomorphic to S^1 (use the parameter t). What distinguishes a trefoil from a circle is not the curve but *how it sits in the surrounding*

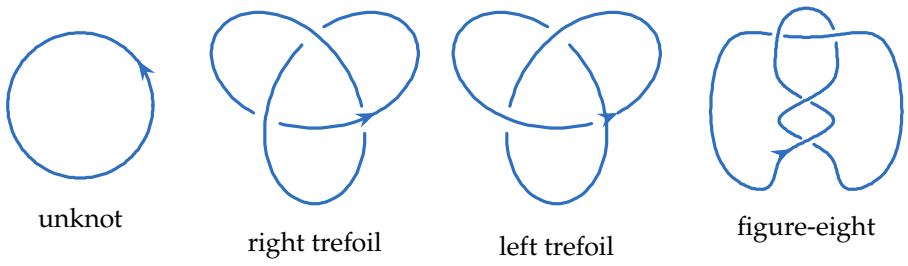


Figure 12.38: The first few knots. Each is a copy of the circle; what differs is how it sits in space. At a crossing the gap marks the strand that passes underneath. The two trefoils are mirror images: every crossing has been switched. The figure-eight knot has four crossings.

space. The information lives in the *complement* $\mathbb{R}^3 \setminus K$, not in K.

This is exactly the situation of a physicist who measures a curve from outside. You cannot walk along a knot and discover it is knotted; every point looks like a line. You discover it by walking *around* it: measuring holonomies, phases, fluxes. That is why the invariants below are built from the complement (the knot group), from loops around the knot (Wilson loops) and from fields sourced by currents on it (the Gauss integral).

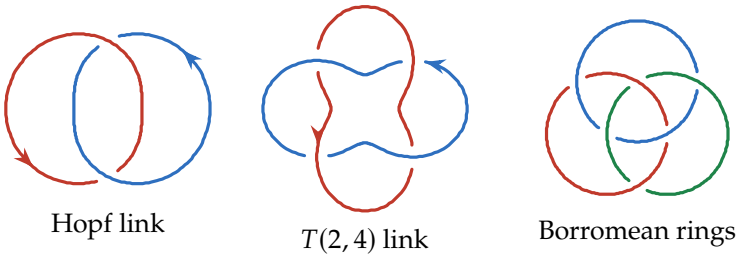


Figure 12.39: Three links. The Hopf link is two circles through each other once. The $T(2,4)$ link winds twice. In the Borromean rings no two rings are linked (remove any one and the other two fall apart), yet the three together cannot be separated.

12.9.2 Why three dimensions: all knots untie in four

Why is knot theory a subject about $\mathbb{R}^3$? Look at the other dimensions.

In the plane. A closed curve in $\mathbb{R}^2$ without self-crossings is, by the Jordan–Schoenflies theorem, always deformable to a round circle. There is no room to tangle.

In four dimensions. In $\mathbb{R}^3$ a knot cannot be untied because one strand would have to pass through another. Think of a crossing: strand a passes over strand b. To change it into an under-crossing we must move a down, and in $\mathbb{R}^3$ it has to go through b. In $\mathbb{R}^4 = \{(x, y, z, w)\}$ we can do this instead:

1. lift a short piece of a near the crossing into $w = 1$ (the rest of the knot stays at $w = 0$);
2. move that piece down in z, past the height of b; this does not hit b, because b is at $w = 0$ and the moving piece is at $w = 1$;
3. lower the piece back to $w = 0$.

The net effect is a crossing change, obtained without any collision. It is the same trick as in one lower dimension. Two strands crossing in the plane must meet; in space one simply goes over the other.

Any knot diagram can be turned into an unknot by changing some of its crossings. (Walk along the knot from a starting point and make each crossing an over-crossing the first time you reach it. The result, a "descending" diagram, lies like a spiral staircase that can be lifted off from the top.) So every knot in $\mathbb{R}^3$ becomes the unknot inside $\mathbb{R}^4$.

Nash quotes the general statement: an m-sphere embedded in $\mathbb{R}^n$ can always be unknotted if

$$n > \tfrac{3}{2}(m+1). \tag{12.147}$$

For $m = 1$ this is $n > 3$, so knotted circles need exactly $n = 3$. There is a cruder count that shows where such bounds come from. Deforming an m-dimensional sphere sweeps out an $(m+1)$-dimensional track. The track generically meets the m-dimensional sphere itself in a set of dimension $(m+1) + m - n$. If this is negative, collisions can be avoided, so all knots untie once $n \geq 2m+2$. The true bound (12.147) is sharper.

m	crude count $n \geq 2m+2$	bound (12.147)	knots exist in
1	4	$n \geq 4$	$\mathbb{R}^3$
2	6	$n \geq 5$	$\mathbb{R}^4$ ("2-knots")
3	8	$n \geq 7$	$\mathbb{R}^5, \mathbb{R}^6$

Nash also notes that other spaces can be knotted in $\mathbb{R}^3$: for example, a knotted torus is the boundary of a thickened knot. We will not need this.

12.9.3 When are two knots the same? Equivalence and isotopy

We need a precise notion of "the same knot". Nash gives two, and the difference between them is the difference between a knot and its mirror image.

Working definition 12.15 (equivalent knots, Nash (12.70))

Knots K_1 and K_2 are **equivalent** if there is a homeomorphism $\alpha : \mathbb{R}^3 \to \mathbb{R}^3$ (a continuous bijection with continuous inverse) that carries one onto the other:

$$\alpha K_1 = K_2. \tag{12.148}$$

Operational test: find a continuous invertible change of coordinates of all of space that moves one curve onto the other. The equivalence classes [each class collects all the knots equivalent to one given knot] are the **knot types**.

Examples of such α: translations, rotations, rescalings $x \mapsto \lambda x$ ($\lambda > 0$), any smooth invertible "stretching" of space, and also reflections, such as

$$\rho(x, y, z) = (x, y, -z). \tag{12.149}$$

The **mirror image** of a knot K is $\bar{K} = \rho K$. Applying ρ to the trefoil (12.145) replaces $\sin 3t$ by $-\sin 3t$. The projection is unchanged, but at every crossing the strand that was on top is now below. That is the left trefoil of Figure 12.38. By definition, K and $\bar{K}$ are always equivalent.

This is too coarse for a physicist who cares about handedness, like a chemist distinguishing a molecule from its mirror image. A reflection is not a motion you can perform on a piece of string. What you can do continuously is captured by the second definition.

Working definition 12.16 (isotopic knots, Nash (12.71))

K_1 and K_2 have the same **isotopy type** if there is a family of homeomorphisms

$$\{\alpha_t\}, \qquad t \in [0,1], \tag{12.150}$$

of $\mathbb{R}^3$, with $\alpha_t(x)$ continuous in t and x together, $\alpha_0 = \mathrm{id}$ and $\alpha_1 K_1 = K_2$. *Operational meaning:* a film, starting from "do nothing", in which all of space flows continuously and at the end K_1 sits where K_2 is. Moving a piece of string about on a table, without cutting it, is an isotopy.

Isotopic knots are equivalent: just take $\alpha = \alpha_1$. The converse fails, and the reason is orientation.

Key idea: Isotopies preserve orientation

Every α_t of an isotopy preserves orientation. For smooth α_t this is a one-line continuity argument:

$$\begin{aligned} \det D\alpha_1 &\overset{(1)}{=} \det D\alpha_0 \cdot \frac{\det D\alpha_1}{\det D\alpha_0} \\ &\overset{(2)}{=} (+1) \cdot (\text{a number that never passes through } 0) \\ &\overset{(3)}{=} \text{positive.} \end{aligned}$$

(1) Trivial rewriting, with $D\alpha_t$ the Jacobian matrix of α_t at a fixed point x.

(2) $\alpha_0 = \mathrm{id}$ has Jacobian $\mathbb{1}$, with determinant 1. Each α_t is invertible, so $\det D\alpha_t \neq 0$ for all t.

(3) A continuous real function of t that starts at $+1$ and is never zero stays positive (intermediate value theorem).

The reflection ρ of (12.149) has Jacobian $\mathrm{diag}(1,1,-1)$, with determinant -1. So ρ is never the end of an isotopy. Conversely, one can show that every orientation-preserving homeomorphism of $\mathbb{R}^3$ is the end of some isotopy. So:

$$\text{isotopic} \iff \text{equivalent through an orientation-preserving homeomorphism.}$$

A knot isotopic to its own mirror image is **amphichiral**. The figure-eight knot is amphichiral (try it with a piece of string, or see Example 12.79). The trefoil is not: the right and left trefoils are equivalent but not isotopic. Nash states this; the proof needs an invariant that can tell the two apart, and the Alexander polynomial and the knot group cannot (Section 12.11.2). The Jones polynomial can (Section 12.11.5).

Watch out: Isotopy is not homotopy (Nash's footnote 7)

The definition of isotopy looks like the definition of a homotopy. The crucial difference is that in an isotopy *every* α_t is invertible, so the curve $\alpha_t K_1$ is an embedded knot at every moment. A homotopy of the curve only requires a continuous family of maps $S^1 \to \mathbb{R}^3$, and these may pass through self-intersections. Under homotopy all knots are trivial: at a self-intersection a strand passes through itself, which changes a crossing (Figure 12.40), and we saw above that crossing changes can untie anything. Indeed $\pi_1(\mathbb{R}^3) = 0$, so every loop is homotopic to a constant. Homotopy forgets exactly the information knot theory is about.

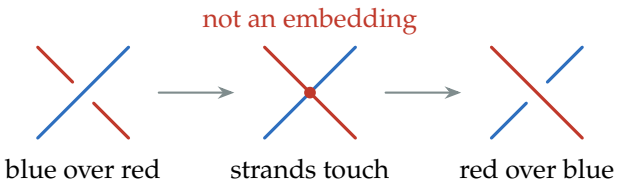


Figure 12.40: A crossing change. A homotopy may pass one strand through another (middle frame); an isotopy may not, because every intermediate curve must be embedded. In $\mathbb{R}^4$ the middle frame can be avoided by lifting the blue strand into the fourth dimension.

Tame and wild knots. Nash separates off a pathology. A knot is **tame** if it is equivalent to a **polygonal** knot, one made of finitely many straight segments. Otherwise it is **wild**. The standard wild knot ties infinitely many smaller and smaller trefoils into a string, with sizes 2^{-n}, so that they accumulate at a point (Figure 12.41). Near the accumulation point the curve is continuous but has no tangent line, and no finite polygon captures it. Every *smooth* knot is tame: approximate it by a fine inscribed polygon; when the polygon is fine enough, straightening each small arc into a chord is an isotopy. From now on, like Nash, we consider only smooth, and hence tame, knots.

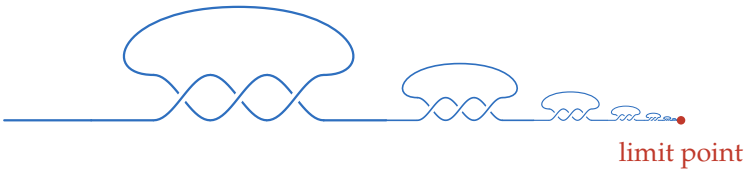


Figure 12.41: A wild knot: infinitely many ever smaller trefoils tied end to end in one string, accumulating at a point (sizes halving each time). The curve is continuous, but at the red point it has no tangent and cannot be replaced by a finite polygon. Smooth knots never do this.

Orientations. An **oriented** knot comes with a direction of travel (the direction of increasing t). Several invariants below (crossing signs, linking numbers, the Jones polynomial of a link) need orientations. Reversing the orientation of a *knot* does not change any of the polynomials we meet. For a link, reversing one component can change them. A knot isotopic to itself with the orientation reversed is **invertible**; all knots with fewer than eight crossings are, and the first non-invertible one has eight crossings.

12.9.4 Knot diagrams and the Reidemeister moves

Nobody computes with parametrisations like (12.145). One computes with pictures. Nash mentions that knot calculations use projections to two dimensions, and that one must then prove the answer does not depend on the projection. Here is the machinery that makes this precise.

Working definition 12.17 (knot diagram)

Project the knot to a plane, for instance by $(x, y, z) \mapsto (x, y)$. The projection is **regular** if it has only finitely many self-intersection points, each a transverse crossing [the two strands cross at a nonzero angle, they do not just touch] of exactly two strands. A **knot diagram** is a regular projection together with the information, at each crossing, of which strand is on top: we break the lower strand.

Operational fact: almost every direction of projection is regular. If a projection has a triple point or a tangency, tilt the plane slightly.

Different diagrams can represent the same knot. How do we know when?

Key idea: Reidemeister's theorem

Two diagrams represent isotopic knots if and only if one can be turned into the other by a finite sequence of

- planar isotopies (sliding the picture about in the plane without creating or destroying crossings), and
- the three **Reidemeister moves** R1, R2, R3 of Figure 12.42.

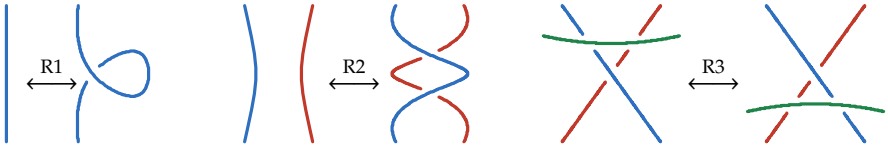


Figure 12.42: The Reidemeister moves. R1 adds or removes a kink. R2 slides one strand entirely over another, creating or removing two crossings. R3 slides a strand (green) over a crossing of the other two. Only the part of the diagram inside a small disc changes; outside it everything is fixed.

Why these three moves? Here is the physicist's argument. An isotopy is a film of knots K_t. Look at the film projected to the plane. At most moments the projection is regular, and it changes only by planar isotopy. The diagram changes type only at isolated moments when the projection is *not* regular. A generic one-parameter family meets only the simplest (codimension-one) failures of regularity, and there are exactly three:

1. a strand is viewed edge-on and its projection has a cusp: a small kink is born or dies (R1);
2. two strands become tangent in the projection and then pass through each other there, one staying on top: two crossings are born or die (R2);
3. three strands pass through a common point of the projection: a triple point, across which a strand slides over a crossing (R3).

Worse coincidences (a tangency at a triple point, say) need two parameters to arrange, and a generic film avoids them, just as a generic path in the plane avoids a given point.

Key idea: How to build a knot invariant

Define a quantity from a diagram. Check that it is unchanged by R1, R2 and R3. Then it is an invariant of the isotopy type of the knot. Every invariant in Sections 12.10 and 12.11 is checked this way. The Chern–Simons definition in Section 12.12 is different: it is manifestly three-dimensional and needs no diagram at all.

Crossing signs and the writhe. Give the knot an orientation. At each crossing the two strands then have directions, and there are two kinds of crossing (Figure 12.43). The **sign** of a crossing is

$$\varepsilon = \left(\hat{t}_{\text{over}} \times \hat{t}_{\text{under}}\right) \cdot \hat{z} \quad \text{(normalised to } \pm 1\text{)}, \tag{12.151}$$

where $\hat{t}$ are the unit tangents, $\hat{z}$ points towards the viewer, and the diagram lies in the xy-plane. Equivalently (right-hand rule): point the thumb of the right hand along the over-strand; if the fingers curl in the direction of the under-strand, the crossing is positive.

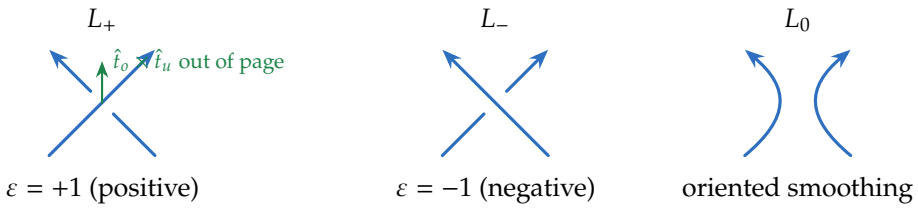


Figure 12.43: The two kinds of oriented crossing and the oriented smoothing. Both strands point upwards. In L_+ the over-strand runs south-west to north-east; in L_- it runs south-east to north-west. L_0 reconnects the four ends without a crossing, respecting the arrows. These are the three local pictures in every skein relation of Section 12.11.

Check (12.151) on L_+ of Figure 12.43. The over-strand has direction $\hat{t}_o \propto (1,1,0)$ and the under-strand $\hat{t}_u \propto (-1,1,0)$. Then $(\hat{t}_o \times \hat{t}_u)_z \propto 1\cdot 1 - 1\cdot(-1) = 2 > 0$, so $\varepsilon = +1$.

Working definition 12.18 (writhe)

The **writhe** of an oriented diagram D is the sum of all its crossing signs [$\varepsilon = +1$ for an L_+ crossing, -1 for an L_- crossing], $w(D) = \sum_{\text{crossings}} \varepsilon$. *Test for invariance:* an R2 move [sliding one strand over another] creates two crossings of opposite sign; R3 [sliding a strand across a crossing] moves crossings about without changing their signs; R1 [adding or removing a kink] creates a single crossing of sign ± 1. So w is invariant under R2 and R3 but *not* under R1: it is a property of the diagram, not of the knot.

For the trefoil drawn in Figure 12.38 with the orientation of increasing t, all three crossings are positive for the right trefoil and negative for the left one: $w = +3$ and $w = -3$. (Mirroring switches every over/under, and switching the over/under at a crossing flips its sign.) The figure-eight diagram has two positive and two negative crossings: $w = 0$. The failure of R1-invariance is not a nuisance to be discarded. We will meet it again as the *framing* of a Wilson loop (Section 12.12.8): w is the self-linking number of the knot thickened into a ribbon lying flat in the page.

12.9.5 A first invariant: tricolouring

Before the heavy machinery, a toy invariant that already proves the trefoil is knotted.

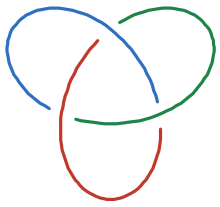
Working definition 12.19 (tricolouring)

An **arc** of a diagram is a piece of the curve running from one under-crossing to the next. A **tricolouring** assigns one of three colours to every arc such that at each crossing the three arcs meeting there (the over-arc and the two halves of the under-strand) are either *all the same colour* or *all different*. Let $\tau(D)$ be the number of tricolourings.

Worked example 12.63: Tricolourings of the unknot and the trefoil

The round circle has one arc and no crossings, so any of the 3 colours works: $\tau = 3$. These three constant colourings exist for every diagram; call a colouring *non-trivial* if it uses more than one colour.

The trefoil diagram has three arcs a, b, c and three crossings, and each crossing involves all three arcs. So the rule says: a, b, c are all equal or all different. All equal: 3 ways. All different: $3! = 6$ ways. So $\tau(\text{trefoil}) = 9$ (Figure 12.44).



at each crossing: over-arc and both under-arcs are all equal or all different.

Here: red, blue, green meet at every crossing, so this is a valid colouring.

Figure 12.44: A non-trivial tricolouring of the trefoil. The three arcs meet at each of the three crossings, and they carry three different colours. The round circle has no non-trivial tricolouring, so the trefoil is not the unknot.

Why is τ an invariant? Check the moves.

- *R1.* The kink has one crossing whose over-arc and one under-arc are the same arc. So all three arcs must be equal: the kink carries the colour of the strand, with nothing new.
- *R2.* Two strands, one over the other. If the lower strand's outer ends have the same colour as the upper strand, the middle piece must share it. If they have colour x and the upper strand has $y \neq x$, the middle piece must be the third colour z at both crossings, consistently. Either way colourings before and after correspond one to one.
- *R3.* A case check (there are only a few colour patterns for the three incoming strands); in each case the outgoing colours can be matched, one to one.

So τ is a knot invariant, and $\tau(\text{trefoil}) = 9 \neq 3 = \tau(\text{unknot})$: *the trefoil is really knotted*. For the figure-eight knot the same count gives $\tau = 3$, so tricolouring cannot distinguish the figure-eight from the unknot. We need stronger invariants.

Colourings as linear algebra. Label the colours by $\mathbb{Z}_3 = \{0, 1, 2\}$. For three elements $x, y, z \in \mathbb{Z}_3$, "all equal or all different" is the same as $x + y + z \equiv 0 \pmod 3$ (check: $0+1+2 = 3$ and $3x \equiv 0$; any other pattern, like $0, 0, 1$, gives a non-zero sum). Since $3 \equiv 0$, this is also $2x_{\text{over}} \equiv x_{\text{under},1} + x_{\text{under},2} \pmod 3$. So tricolourings are the solutions of a linear system mod 3, with one equation per crossing. For the trefoil,

$$\begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} \equiv 0 \pmod 3.$$

Mod 3 every entry of this matrix is $\equiv -1 \equiv 2$ times the all-ones matrix, whose rank is 1. So the solution space is 2-dimensional over $\mathbb{Z}_3$, with $3^2 = 9$ elements, as counted. Replacing 3 by another prime p gives **Fox p-colourings**. They are controlled by one integer, the *determinant* of the knot, which we will meet again as $|\Delta(-1)|$ in Section 12.11.2.

12.9.6 The knot group

Nash says the classification of knots proceeds by constructing invariants, and the first one he names is the fundamental group of the complement. It is the most direct way to "walk around" the knot.

Reminder: the fundamental group. For a space X and a base point x_0, $\pi_1(X, x_0)$ is the set of loops starting and ending at x_0, where two loops are identified if one can be continuously deformed into the other (keeping x_0 fixed). Loops are multiplied by doing one after the other; the inverse is the loop run backwards; the identity is the constant loop. See Chapter 1.

Working definition 12.20 (knot group, Nash (12.72))

The **knot group** of K is

$$\pi_1(\mathbb{R}^3 \setminus K). \tag{12.152}$$

Operational meaning: the group of ways of travelling around the knot and coming home, up to deformation. Equivalent knots have homeomorphic complements [the map α carries $\mathbb{R}^3 \setminus K_1$ onto $\mathbb{R}^3 \setminus K_2$], hence isomorphic groups.

Worked example 12.64: The unknot: the Aharonov–Bohm group

Let K be the unit circle in the xy-plane. A loop in the complement can wind any integer number of times through the disc bounded by K, and the winding number is the only thing that survives deformation. So $\pi_1(\mathbb{R}^3 \setminus K) \cong \mathbb{Z}$, generated by a small loop μ that links K once (a **meridian**).

Physically: run a steady current I around K. For a loop ℓ in the complement, Ampère's law gives

$$\oint_\ell \vec{B} \cdot \mathrm{d}\vec{l} = \mu_0 I \, n(\ell),$$

with $n(\ell)$ the number of times ℓ threads K. The integer $n(\ell)$ is precisely the element of $\pi_1 \cong \mathbb{Z}$ represented by ℓ. A charged particle carried around ℓ picks up an Aharonov–Bohm phase that depends only on $n(\ell)$. This is the abelian version of the Wilson loops of Section 12.12.

For a knotted K the group is non-abelian, and a direct recipe computes a presentation of it from any diagram.

Working definition 12.21 (Wirtinger presentation)

Take an oriented diagram with arcs $x_1, \dots, x_n$ (and n crossings; for a knot diagram each crossing ends one arc and starts another). Put the base point high above the page, towards the viewer.

- *Generators.* To each arc x_i associate the loop that comes down from the base point, passes once under x_i from right to left (with respect to the arc's direction), and goes back up. We call this loop x_i too.
- *Relations.* At a crossing where the over-arc is x_k and the under-strand goes from x_i into x_j, there is one relation:

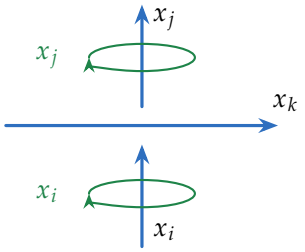
$$x_j = x_k^{\varepsilon} \, x_i \, x_k^{-\varepsilon}, \tag{12.153}$$

  where $\varepsilon = \pm 1$ is the crossing sign.

Then $\pi_1(\mathbb{R}^3 \setminus K)$ is generated by $x_1, \dots, x_n$ subject to these n relations [every element is a product of the $x_i^{\pm 1}$, and two products are equal exactly when the relations force it], and any one of the relations follows from the others.

Why the relation holds (Figure 12.45): a loop passing under the arc x_i just before the crossing can be slid along the under-strand, underneath the over-arc x_k, to a loop under x_j. Sliding

it under x_k drags it around x_k, and that is conjugation by x_k. The sign ε records which way round x_k goes. (With other conventions for the generators, the exponent is $-\varepsilon$; this changes nothing below.)



Slide the loop around x_i up the under-strand. To get past the over-arc it must go round x_k once: $x_j = x_k^{\pm 1} x_i\, x_k^{\mp 1}$.

Figure 12.45: The Wirtinger relation at one crossing. The green loops encircle the arcs just below and just above the over-arc x_k. They differ by conjugation by the loop around x_k.

Worked example 12.65: The trefoil group and a representation by matrices

Label the three arcs of the trefoil x, y, z so that at the three crossings the over-arcs are x, y, z in turn. With all crossings of the same sign, the Wirtinger relations are, in a suitable labelling,

$$z = x\,y\,x^{-1}, \qquad x = y\,z\,y^{-1}, \qquad y = z\,x\,z^{-1}.$$

Eliminate z using the first relation and substitute into the second:

$$\begin{aligned} x &\overset{(1)}{=} y\,z\,y^{-1} \\ &\overset{(2)}{=} y\,(xyx^{-1})\,y^{-1}, \end{aligned}$$

(1) The second Wirtinger relation.
(2) Substitute $z = xyx^{-1}$ from the first relation.

so $x\,y\,x = y\,x\,y$ after multiplying on the right by $y\,x$. (The third relation then follows automatically, as promised.) Hence

$$\pi_1(\mathbb{R}^3 \setminus \text{trefoil}) = \langle x, y \mid xyx = yxy \rangle. \tag{12.154}$$

This is the braid group on three strands. With $a = xyx$ and $b = xy$ one finds $b^3 = xyxyxy = (xyx)(yxy) = (xyx)(xyx) = a^2$, which gives the other common form $\langle a, b \mid a^2 = b^3 \rangle$, the torus-knot presentation for $T(2,3)$.

Is this group really different from $\mathbb{Z}$? We prove it is non-abelian by exhibiting a representation with non-commuting matrices. Send x and y to the transpositions (12) and (23) of three objects, written as permutation matrices:

$$x \mapsto P_x = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad y \mapsto P_y = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

Check the relation:

$$\begin{aligned} P_x P_y P_x &\overset{(1)}{=} \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} P_x \overset{(2)}{=} \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \\ P_y P_x P_y &\overset{(3)}{=} \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} P_y \overset{(4)}{=} \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}. \end{aligned}$$

(1) Row i of P_xP_y is row i of P_x times P_y: rows $(0,1,0)$, $(1,0,0)$, $(0,0,1)$ of P_x pick out rows $2,1,3$ of P_y.
(2) Multiplying by P_x on the right swaps the first two columns.
(3) Similarly, P_yP_x has rows $1,3,2$ of P_x.
(4) Multiplying by P_y on the right swaps the last two columns.

Both products equal the matrix of (13), so $xyx = yxy$ holds in the image and we have a genuine representation of the knot group. But $P_xP_y \neq P_yP_x$ (the first is the cycle (123), the second (132)). The image is the whole permutation group S_3, which is non-abelian. Therefore the trefoil group is not $\mathbb{Z}$, and the trefoil is not the unknot, a second proof.
The two proofs are the same proof. A homomorphism $\pi_1 \to S_3$ sending every Wirtinger generator to a transposition is a choice of one of the three transpositions (12), (23), (13) for each arc, subject to (12.153). Conjugating one transposition by another gives the third one, unless they are equal. That is exactly the tricolouring rule, with colours = transpositions.

Physical picture: Representations of the knot group are flat connections

A homomorphism $\rho : \pi_1(\mathbb{R}^3 \setminus K) \to G$ assigns a group element to each loop in the complement, compatibly with composition. That is exactly the holonomy of a *flat* G-connection on the complement: the Wilson loop $P \exp \oint A$ of a connection with $F = 0$ depends only on the homotopy class of the loop (Chapter T3). So "counting representations of the knot group" means "counting flat gauge fields on the knot complement". Flat connections are exactly the classical solutions of Chern–Simons theory (Section 12.13.3); this is the first hint of why a gauge theory computes knot invariants.

What the knot group cannot see. All Wirtinger generators of a knot are conjugate to each other (at each crossing x_j is conjugate to x_i, and the diagram is connected). Abelianising (setting all commutators to 1) therefore sends them all to one generator t:

$$H_1(\mathbb{R}^3 \setminus K) = \pi_1/[\pi_1, \pi_1] \cong \mathbb{Z} = \{t^n\}$$

for every knot. The abelian part sees nothing; all the information is in the non-abelian structure. The Alexander polynomial of Section 12.11.2 extracts some of that information and packages it as a Laurent polynomial in this t. Also, the complements of K and of its mirror image $\bar{K}$ are homeomorphic (by ρ), so the knot group can never detect chirality.

Reading Nash (pp. 339–340, (12.70)–(12.72))

Nash defines knots as embeddings of S^1 in $\mathbb{R}^3$ or S^3, notes that they are all homeomorphic to S^1, and quotes the unknotting criterion $n > 3(m+1)/2$ (his reference is Kyle; the smooth statement is due to Haefliger). He defines links with p components, equivalence by a homeomorphism $\alpha K_1 = K_2$ (12.70) and isotopy through a family α_t with $\alpha_0 = \mathrm{id}$ (12.71). His footnote 7 is our pitfall on homotopy versus isotopy. His "isotopic knots are equivalent, but the converse is false, because the α_t all preserve orientation" is our determinant argument. "Tame" and "wild" are as above, and his first invariant is the knot group (12.72). Diagrams, Reidemeister moves, crossing signs, the writhe, tricolourings and the Wirtinger presentation are not in Nash; they are the standard tools that make his statements checkable, and we use them throughout.

Recap

- A knot is an embedded circle in $\mathbb{R}^3$ or S^3; the knottedness is in the complement, not in the circle.
- All knots untie in $\mathbb{R}^4$: an extra dimension lets strands pass each other. Knotted m-spheres need $n \le \tfrac{3}{2}(m+1)$.
- Equivalence uses any homeomorphism of space; isotopy uses a continuous film starting at the identity, which preserves orientation. A knot and its mirror image are equivalent, but need not be isotopic (trefoil).
- Homotopy allows self-crossing and trivialises every knot. We consider smooth, hence tame, knots.
- Diagrams represent the same knot iff they are related by R1, R2, R3. An invariant is anything unchanged by the three moves. The writhe is invariant under R2, R3 but not R1.
- Tricolourings and the knot group $\langle x, y \mid xyx = yxy \rangle$ (with its S_3 representation) both prove the trefoil is knotted. Representations of the knot group are flat connections on the complement.

12.10 The linking number: Gauss, Ampère and a degree

The simplest question about two disjoint closed curves C_1 and C_2 is: *how many times does one wind around the other?* The answer is an integer, the **linking number** $L(C_1, C_2)$. It is the oldest knot invariant. Gauss wrote down an integral formula for it in 1833, and it is exactly the quantity that the abelian Chern–Simons path integral will produce in Section 12.12 (Nash's (12.89)–(12.90)). So before the polynomials of Section 12.11, we take the linking number apart completely. We shall meet it in four guises:

1. a *count of crossings* in a diagram;
2. a *flux*: the current through one loop measured by Ampère's law around the other;
3. the *Gauss double integral*, which is Biot–Savart in disguise;
4. the *degree* of a map from a torus to a sphere.

That the four agree is the whole content of this section. Each guise explains a different feature. The crossing count makes it computable. The flux makes it obviously an integer. The Gauss integral makes it obviously independent of any projection. The degree makes it obviously unchanged by deformations.

Throughout, C_1 and C_2 are oriented, disjoint, smooth closed curves in $\mathbb{R}^3$, parametrised as $x(s)$ and $y(t)$ with $s, t \in [0, 2\pi]$. We write $\dot{x} = \mathrm{d}x/\mathrm{d}s$ and $\dot{y} = \mathrm{d}y/\mathrm{d}t$.

12.10.1 Counting crossings

Working definition 12.22 (linking number from a diagram)

Draw a diagram of the two-component link $C_1 \cup C_2$. Look only at the crossings *between* C_1 and C_2 (ignore the self-crossings of each component). Each has a sign $\varepsilon = \pm 1$ given by (12.151) [$+1$ for an L_+ crossing, -1 for an L_- crossing]. Then

$$L(C_1, C_2) = \frac{1}{2} \sum_{\substack{\text{crossings of} \\ C_1 \text{ with } C_2}} \varepsilon = \sum_{\substack{\text{crossings where} \\ C_1 \text{ is over } C_2}} \varepsilon. \tag{12.155}$$

Why are the two sums in (12.155) equal? Walk along C_2 in the diagram. Each time C_2 passes under C_1 and then later back out, the total number of crossings between the two is even, and the "C_1 over" and "C_2 over" signed counts are equal. This is not obvious from the diagram alone; it follows from the degree picture of Section 12.10.4, where the two counts are the degree of the same map computed at two antipodal points of the sphere. We use the symmetric form $\frac{1}{2}\sum$ when we want to check invariance, and the one-sided form when we want to compute quickly.

Invariance under the Reidemeister moves.

- R1 creates or removes a self-crossing of one component. It does not enter the sum.
- R2 between C_1 and C_2 creates two inter-component crossings. The two strands are the same two directed strands at both, and one of them is over at both. Reversing which of the two crossings we look at reverses the relative direction of the under-strand, so their signs are opposite. The sum is unchanged. (R2 between a component and itself does not enter.)
- R3 moves one strand across a crossing of two others. The three crossings survive with the same over/under information and the same directions, so the same signs.

So L is a link invariant. Because every crossing has sign ± 1 and the one-sided count is a sum of such signs, $L \in \mathbb{Z}$.

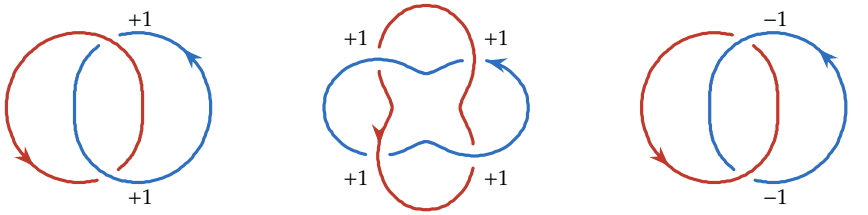


Figure 12.46: Counting crossings. Only crossings between the blue and the red component count, each with its sign. The Hopf link has two, both $+1$, so $L = \frac{1}{2}(1+1) = 1$. The $T(2,4)$ link has four positive crossings: $L = 2$. The mirror image switches every crossing and flips every sign.

Worked example 12.66: Linking numbers of simple links

1. *The unlink* (two circles far apart) has no crossings between the components: $L = 0$.
2. *The Hopf link* of Figure 12.46 (left). Two inter-component crossings, both positive:

$$L = \tfrac{1}{2}(+1 + 1) = 1.$$

The one-sided count agrees: blue passes over red exactly once (at the bottom), with sign +1.

3. *The torus links* $T(2, 2n)$, two parallel strands twisted n full times around each other. The diagram has $2n$ crossings, all between the two components, all of the same sign: $L = \pm n$.

4. *The Whitehead link*. One round circle, and one component that clasps it: it goes under, over, then back under and over in the reverse direction. There are four inter-component crossings with signs $+1, +1, -1, -1$, so $L = 0$. Yet the two components cannot be pulled apart. The linking number does not see this.

5. *The Borromean rings* (Figure 12.39). Each pair of rings has two crossings with opposite signs, so every pairwise linking number is 0, even though the three rings are linked.

Three properties follow at once from (12.155).

$$\begin{aligned} L(C_2, C_1) &= L(C_1, C_2) && \text{(the symmetric sum does not care which is first),} \\ L(-C_1, C_2) &= -L(C_1, C_2) && \text{(reversing one strand flips the sign of every crossing it is in),} \\ L(\rho C_1, \rho C_2) &= -L(C_1, C_2) && \text{(the mirror (12.149) switches every crossing).} \end{aligned}$$

The last line means that the linking number, unlike the knot group, *does* detect chirality of links: the Hopf link with $L = +1$ is not isotopic to its mirror image with $L = -1$ (as oriented links).

12.10.2 Ampère's law: why the linking number is an integer

Now forget diagrams and think about magnetostatics. Let a steady unit current $I = 1$ flow along C_2. We use units with $\mu_0 = 1$. The magnetic field is given by the Biot–Savart law,

$$\vec{B}(x) = \frac{1}{4\pi} \oint_{C_2} \frac{d\vec{y} \times (x - y)}{|x - y|^3}. \tag{12.156}$$

Ampère's law says that the circulation of $\vec{B}$ around any closed loop equals the current threading the loop:

$$\oint_{C_1} \vec{B} \cdot d\vec{x} = (\text{current through a surface bounded by } C_1). \tag{12.157}$$

Let us see why the right side is an integer, and which integer it is.

Working definition 12.23 (linking number as a flux)

Choose any oriented surface S with boundary $\partial S = C_1$ (orient S so that C_1 runs anticlockwise when S is seen from the tip of its normal $\hat{n}$). Count the points where C_2 pierces S, with sign $+1$ if C_2 crosses in the direction of $\hat{n}$ and -1 otherwise. The signed count is $L(C_1, C_2)$.

Such a surface always exists. If C_1 is a round circle, S is the flat disc. For any knot, Seifert's algorithm (Section 12.11.2) constructs an orientable surface with boundary C_1, a **Seifert surface**. The surface may be pierced by C_2 many times, but the signed count is independent of the choice of S. Two choices S and S' together bound a region of space (glue S to S' with the orientation of S' reversed); C_2 enters that region as often as it leaves, so the counts through S and S' agree.

Now apply Stokes' theorem. Because the current is a sum of delta functions along C_2,

$$\oint_{C_1} \vec{B} \cdot \mathrm{d}\vec{x} \overset{(1)}{=} \int_S (\nabla \times \vec{B}) \cdot \hat{n}\, \mathrm{d}A \overset{(2)}{=} \int_S \vec{j} \cdot \hat{n}\, \mathrm{d}A \overset{(3)}{=} \sum_{\text{piercings } p} \varepsilon_p \overset{(4)}{=} L(C_1, C_2). \tag{12.158}$$

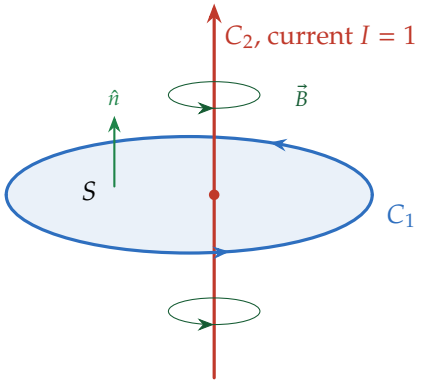
(1) Stokes' theorem for the surface S with $\partial S = C_1$ (Chapter T2). $\vec{B}$ is smooth on S except where C_2 pierces it, where it has an integrable $1/\rho$ singularity; cutting out tiny discs around the piercings and letting them shrink gives the same result.

(2) The magnetostatic Maxwell equation $\nabla \times \vec{B} = \vec{j}$ (with $\mu_0 = 1$), which Biot–Savart solves.

(3) $\vec{j}$ is concentrated on C_2, a unit current along the curve. Each piercing contributes ± 1 according to the direction of crossing: the flux of a unit current through a surface it crosses once.

(4) The flux description of the linking number.

The left side is a continuous function of the curves (as long as they do not touch). The right side is an integer. So the left side cannot change under deformation: it is a topological invariant, and an integer. This is Nash's remark "from Ampère's law, where precisely this integral occurs, it follows that $L(C_l, C_m)$ is an integer".



one piercing, along $\hat{n}$:

$$\oint_{C_1} \vec{B} \cdot \mathrm{d}\vec{x} = \int_S \vec{j} \cdot \hat{n}\, \mathrm{d}A = +1$$

Figure 12.47: The linking number as a flux. The disc S spans C_1; its normal $\hat{n}$ is fixed by the right-hand rule from the arrow on C_1. The current on C_2 pierces S once along $\hat{n}$, so Ampère's law gives circulation $+1$. Deform the curves and the piercing may move, but it cannot disappear without C_2 crossing C_1.

12.10.3 The Gauss integral

Substituting Biot–Savart into Ampère turns the linking number into a double integral over the two curves, with no surface and no diagram in sight:

$$L(C_1, C_2) \overset{(1)}{=} \oint_{C_1} \vec{B} \cdot \mathrm{d}\vec{x} \overset{(2)}{=} \frac{1}{4\pi} \oint_{C_1} \oint_{C_2} \frac{\mathrm{d}\vec{x} \cdot \left(\mathrm{d}\vec{y} \times (x - y)\right)}{|x - y|^3} \overset{(3)}{=} \frac{1}{4\pi} \oint_{C_1} \oint_{C_2} \frac{(x - y) \cdot (\mathrm{d}\vec{x} \times \mathrm{d}\vec{y})}{|x - y|^3}$$

$$\overset{(4)}{=} \frac{\varepsilon_{ijk}}{4\pi} \oint_{C_1} \mathrm{d}x^i \oint_{C_2} \mathrm{d}y^j \, \frac{(x-y)^k}{|x-y|^3}. \tag{12.159}$$

(1) Ampère's law (12.158).
(2) Insert the Biot–Savart field (12.156) of a unit current on C_2.
(3) Cyclic symmetry of the triple product, $\vec{a}\cdot(\vec{b}\times\vec{c}) = \vec{c}\cdot(\vec{a}\times\vec{b})$, with $\vec{a} = \mathrm{d}\vec{x}$, $\vec{b} = \mathrm{d}\vec{y}$, $\vec{c} = x - y$.
(4) In components, $\vec{c}\cdot(\vec{a}\times\vec{b}) = \varepsilon_{ijk}a^i b^j c^k$.

The last line is exactly Nash's (12.90). In parametrised form, which is what one feeds to a computer,

$$L(C_1, C_2) = \frac{1}{4\pi} \int_0^{2\pi} \mathrm{d}s \int_0^{2\pi} \mathrm{d}t \, \frac{\big(x(s) - y(t)\big)\cdot\big(\dot{x}(s)\times\dot{y}(t)\big)}{|x(s) - y(t)|^3}. \tag{12.160}$$

Worked example 12.67: The Gauss integral for a circle and a line

Take C_1 the unit circle in the xy-plane, anticlockwise seen from above, and C_2 the z-axis traversed upwards. The z-axis is not a closed curve in $\mathbb{R}^3$, but it is one in $S^3 = \mathbb{R}^3 \cup \{\infty\}$: it closes up through the point at infinity. (Equivalently, think of a huge circle, most of which is far away and contributes nothing.) Parametrise

$$x(s) = (\cos s, \sin s, 0), \qquad y(u) = (0, 0, u), \quad u \in \mathbb{R}.$$

Then $\dot{x} = (-\sin s, \cos s, 0)$, $\dot{y} = (0, 0, 1)$ and $x - y = (\cos s, \sin s, -u)$. Compute the integrand piece by piece:

$$\begin{aligned} \dot{x}\times\dot{y} &\overset{(1)}{=} (\cos s\cdot 1 - 0,\ 0 - (-\sin s)\cdot 1,\ 0) = (\cos s, \sin s, 0), \\ (x-y)\cdot(\dot{x}\times\dot{y}) &\overset{(2)}{=} \cos^2 s + \sin^2 s + 0 = 1, \\ |x-y|^3 &\overset{(3)}{=} (1+u^2)^{3/2}. \end{aligned}$$

(1) $(\vec{a}\times\vec{b})_x = a_y b_z - a_z b_y$, and cyclically, with $\vec{a} = \dot{x}$, $\vec{b} = \dot{y}$.
(2) Dot product with $x - y = (\cos s, \sin s, -u)$.
(3) $|x-y|^2 = \cos^2 s + \sin^2 s + u^2$.

So

$$\begin{aligned} L(C_1, C_2) &\overset{(1)}{=} \frac{1}{4\pi} \int_0^{2\pi} \mathrm{d}s \int_{-\infty}^{\infty} \frac{\mathrm{d}u}{(1+u^2)^{3/2}} \\ &\overset{(2)}{=} \frac{1}{4\pi}\cdot 2\pi \cdot \left[\frac{u}{\sqrt{1+u^2}}\right]_{-\infty}^{\infty} \\ &\overset{(3)}{=} \frac{1}{4\pi}\cdot 2\pi\cdot 2 = 1. \end{aligned}$$

(1) The parametrised Gauss formula (12.160) with the integrand just computed.
(2) The s-integral is trivial. For the u-integral, $\frac{\mathrm{d}}{\mathrm{d}u}\frac{u}{\sqrt{1+u^2}} = \frac{(1+u^2)-u^2}{(1+u^2)^{3/2}}$.
(3) $u/\sqrt{1+u^2} \to \pm 1$ as $u \to \pm\infty$.

The z-axis goes up through the disc spanned by the circle, which is the direction of the circle's right-hand normal: one positive piercing, and the integral agrees. Physically, $2\pi\cdot\frac{1}{2\pi\rho}\big|_{\rho=1} = 1$ is the circulation of the field $B_\phi = 1/(2\pi\rho)$ of an infinite wire.

Worked example 12.68: Two coplanar circles

Let C_1 and C_2 be any two disjoint closed curves lying in the plane $z = 0$. Then $\dot{x}$ and $\dot{y}$ lie in the plane, so $\dot{x} \times \dot{y}$ points along $\hat{z}$. But $x - y$ also lies in the plane, so $(x - y) \cdot (\dot{x} \times \dot{y}) = 0$ at every point. The integrand vanishes identically: $L = 0$. Two curves in a plane are never linked, as the crossing count (no crossings at all) also says.

Watch out: The integrand is not a total derivative

The Gauss integrand is not zero and not constant for a general link; only its integral is topological. For the Hopf link of Figure 12.46 the integrand is large where the curves come close and small elsewhere. Deform the curves and the integrand changes everywhere, but the integral stays at 1. A numerical evaluation of (12.160) on the torus-link parametrisations (12.146) with 400×400 points gives 1.0001 for the Hopf link $T(2, 2)$ and 2.0002 for $T(2, 4)$.

12.10.4 The Gauss integral as a degree

Why is the Gauss integral unchanged by deformations, directly and without Ampère? Because it is the *degree* of a map, and degrees are integers that cannot jump.

Working definition 12.24 (the Gauss map)

The **Gauss map** of the pair (C_1, C_2) assigns to each pair of parameters the unit vector pointing from the point on C_1 to the point on C_2:

$$\Psi : T^2 \to S^2, \qquad \Psi(s, t) = \frac{y(t) - x(s)}{|y(t) - x(s)|}. \tag{12.161}$$

Here $T^2 = S^1 \times S^1$ is the torus of pairs (s, t), oriented by $\mathrm{d}s \wedge \mathrm{d}t$ [s counts as the first direction and t as the second, which fixes the sign of the degree]. It is well defined because the curves never meet.

The **degree** of a map between closed oriented surfaces of the same dimension counts, with signs, how many times it covers the target (Chapter 1). Operationally there are two ways to compute it:

- *by integration:* $\deg \Psi = \frac{1}{4\pi} \int_{T^2} \Psi^* \omega$, where ω is the area form of the unit sphere, with total area 4π;
- *by counting:* pick a point $\hat{n} \in S^2$ that is a regular value; then $\deg \Psi = \sum_{\Psi(p)=\hat{n}} \pm 1$, with $+1$ where Ψ preserves orientation near p and -1 where it reverses it.

Step 1: the integral is the Gauss integral. The area form of the unit sphere at the point $u \in S^2$, evaluated on two tangent vectors $\vec{a}, \vec{b}$, is the signed area $u \cdot (\vec{a} \times \vec{b})$. Put $\vec{r} = x - y$, so $\Psi = -\vec{r}/|\vec{r}|$. Then

$$\partial_s \Psi \overset{(1)}{=} -\frac{\dot{x}}{|\vec{r}|} + (\dots)\,\vec{r}, \qquad \partial_t \Psi \overset{(2)}{=} \frac{\dot{y}}{|\vec{r}|} + (\dots)\,\vec{r},$$

$$\Psi^* \omega(\partial_s, \partial_t) \overset{(3)}{=} \Psi \cdot (\partial_s \Psi \times \partial_t \Psi)$$

$$\overset{(4)}{=} \left(-\frac{\vec{r}}{|\vec{r}|}\right) \cdot \left(-\frac{\dot{x}}{|\vec{r}|} \times \frac{\dot{y}}{|\vec{r}|}\right) = \frac{\vec{r} \cdot (\dot{x} \times \dot{y})}{|\vec{r}|^3}.$$

(1) $\partial_s \vec{r} = \dot{x}$ and $\partial_s (1/|\vec{r}|) \propto \vec{r}$; only the first term will matter.

(2) $\partial_t \vec{r} = -\dot{y}$, so $\partial_t \Psi = +\dot{y}/|\vec{r}| + (\ldots)\vec{r}$.
(3) Definition of the area form, and of pullback: $\Psi^*\omega$ evaluated on ∂_s, ∂_t is ω evaluated on their images (Chapter T1).
(4) Any term proportional to $\vec{r}$ inside the cross product is killed by the dot product with $\vec{r}$ (a triple product with a repeated vector vanishes).

So $\frac{1}{4\pi}\int_{T^2} \Psi^*\omega$ is exactly the Gauss integral (12.160):

$$L(C_1, C_2) = \deg \Psi. \tag{12.162}$$

Step 2: why the degree is invariant. Deform the curves. Ψ changes continuously, so $\frac{1}{4\pi}\int \Psi^*\omega$ changes continuously. But it is always an integer (the counting description). A continuous integer-valued function is constant. The only way to change the degree is to make Ψ undefined, which happens exactly when $y(t) = x(s)$ for some s, t: the curves pass through each other.

Step 3: counting gives the crossing formula. Choose $\hat{n} = -\hat{z}$, pointing away from a viewer who looks down on the xy-plane. Then $\Psi(s,t) = -\hat{z}$ means that $y(t)$ lies directly *below* $x(s)$: in the diagram, C_1 crosses over C_2 there. So the preimages of $-\hat{z}$ are exactly the crossings where C_1 is over C_2, and the local sign of Ψ at such a point is the crossing sign ε. (Check it on the Hopf link: one such crossing, sign +1, and $L = +1$.) If instead we choose $\hat{n} = +\hat{z}$, the preimages are the crossings where C_2 is over C_1. Both counts compute the same degree, which proves that the two sums in (12.155) agree.

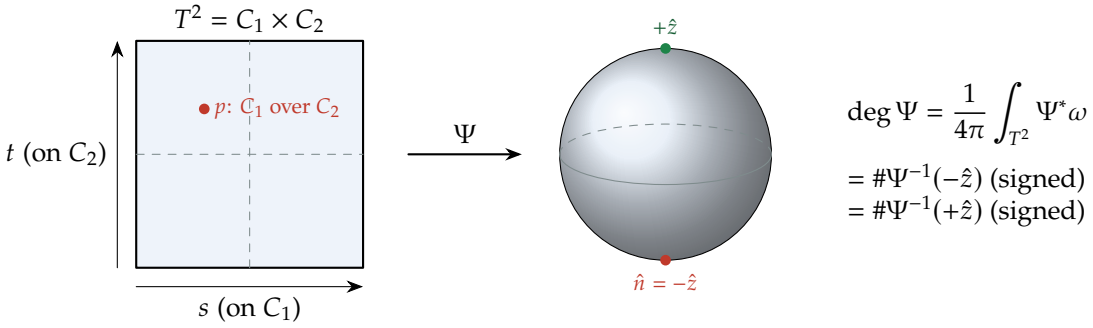


Figure 12.48: The Gauss map sends a pair of points, one on each curve, to the direction from the first to the second. The torus of pairs wraps around the sphere L times. Pairs mapped to the south pole are the crossings where C_1 lies over C_2; pairs mapped to the north pole are those where C_2 lies over C_1.

Worked example 12.69: The degree for the circle and the line

Return to Example 12.67. There $\Psi(s,u) = \big(-\cos s, -\sin s, u\big)/\sqrt{1+u^2}$. As u runs from $-\infty$ to ∞ the third component runs from -1 to $+1$: from the south pole to the north pole. As s goes once round, Ψ goes once round the z-axis. So Ψ covers every point of the sphere (apart from the poles, which are limits) exactly once, and $|\deg \Psi| = 1$. The sign is +1 by Step 1, because the integrand is positive.

12.10.5 Homology and physics: what the linking number measures

Two more ways to see the same integer, both important later.

Homology of the complement. In Example 12.64 we saw that the first homology of the complement of a knot is $H_1(\mathbb{R}^3 \setminus C_2) \cong \mathbb{Z}$, generated by a small meridian loop μ_2 around C_2.

The curve C_1 lies in this complement, so it defines a class

$$[C_1] = L(C_1, C_2)\,[\mu_2] \in H_1(\mathbb{R}^3 \setminus C_2) \cong \mathbb{Z}. \tag{12.163}$$

The integer is the linking number: the homomorphism $H_1(\mathbb{R}^3 \setminus C_2) \to \mathbb{Z}$ is "circulation of the field of a unit current on C_2", which is 1 on μ_2. In gauge-theory words: a particle of charge q transported around C_1 in the field of a unit magnetic flux tube along C_2 picks up the Aharonov–Bohm phase

$$\exp\Big(iq \oint_{C_1} \vec{A} \cdot \mathrm{d}\vec{x}\Big) = \exp\big(iq\, L(C_1, C_2)\big).$$

The Wilson loop of a $U(1)$ field with a flux tube measures the linking number. The abelian Chern–Simons theory of Section 12.12 is a theory in which each Wilson loop *is* such a flux tube for the others.

Magnetic helicity. The abelian Chern–Simons action is, up to a constant, the integral $\int A \wedge \mathrm{d}A$. For a real $U(1)$ field in $\mathbb{R}^3$ in the vector-calculus dictionary of Chapter T1, $A \leftrightarrow \vec{A}$, $\mathrm{d}A \leftrightarrow \vec{B} = \nabla \times \vec{A}$, and

$$\int_{\mathbb{R}^3} A \wedge \mathrm{d}A = \int_{\mathbb{R}^3} \vec{A} \cdot \vec{B}\, \mathrm{d}^3x,$$

which is the **magnetic helicity** of plasma physics (Woltjer, Moffatt). It measures linking of field lines.

Worked example 12.70: Helicity of two linked flux tubes

Let the magnetic field be concentrated in two thin, untwisted flux tubes along closed curves C_1, C_2 carrying fluxes Φ_1 and Φ_2. So $\vec{B} = \Phi_1 \vec{\delta}_{C_1} + \Phi_2 \vec{\delta}_{C_2}$, where $\int \vec{v} \cdot \vec{\delta}_C\, \mathrm{d}^3x = \oint_C \vec{v} \cdot \mathrm{d}\vec{x}$. Write $\vec{A} = \Phi_1 \vec{A}_1 + \Phi_2 \vec{A}_2$ with $\nabla \times \vec{A}_i$ the field of a unit tube along C_i. Then

$$\begin{aligned} \int \vec{A} \cdot \vec{B}\, \mathrm{d}^3x &\overset{(1)}{=} \sum_{i,j} \Phi_i \Phi_j \oint_{C_i} \vec{A}_j \cdot \mathrm{d}\vec{x} \\ &\overset{(2)}{=} \sum_{i \neq j} \Phi_i \Phi_j\, L(C_i, C_j) + (\text{self terms}) \\ &\overset{(3)}{=} 2\,\Phi_1 \Phi_2\, L(C_1, C_2) + (\text{self terms}). \end{aligned}$$

(1) Insert $\vec{B}$ as delta functions and expand $\vec{A}$.

(2) By Stokes, $\oint_{C_i} \vec{A}_j \cdot \mathrm{d}\vec{x}$ is the flux of the unit tube j through a surface spanning C_i, which is the signed number of piercings: $L(C_i, C_j)$ (the flux description, with the roles of $\vec{B}$ and $\vec{j}$ exchanged).

(3) Symmetry $L(C_1, C_2) = L(C_2, C_1)$.

The "self terms" $i = j$ measure how each tube is twisted and coiled around itself. They are exactly the self-linking numbers that force the framing of Section 12.12.8. Nash's formula (12.92), $\exp\big(\frac{i\pi}{k} \sum_{l,m} n_l n_m L(C_l, C_m)\big)$, is this helicity with fluxes $\Phi_l \propto n_l$ put into the exponent.

12.10.6 Self-linking: the first sign of framing

What if $C_1 = C_2 = C$? The Gauss formula (12.160) with both parameters on the same curve,

$$\mathrm{Wr}(C) = \frac{1}{4\pi} \oint_C \oint_C \frac{(x-y)\cdot(\mathrm{d}\vec{x}\times\mathrm{d}\vec{y})}{|x-y|^3}, \tag{12.164}$$

is called the **writhe** of the space curve C. (Its relative, the writhe $w(D)$ of a diagram, was defined in Section 12.9.4.) Near the diagonal $s = t$ both numerator and denominator vanish. Expand $x = x(s)$, $y = x(t)$ with $h = t - s$ small:

$$\begin{aligned} x - y &\overset{(1)}{=} -\dot{x}\,h - \tfrac{1}{2}\ddot{x}\,h^2 + O(h^3), \\ \dot{x}(s)\times\dot{x}(t) &\overset{(2)}{=} \dot{x}\times\ddot{x}\,h + O(h^2), \\ (x-y)\cdot\big(\dot{x}(s)\times\dot{x}(t)\big) &\overset{(3)}{=} -\tfrac{1}{2}\,\ddot{x}\cdot(\dot{x}\times\ddot{x})\,h^3 + O(h^4) = O(h^4), \\ |x-y|^3 &\overset{(4)}{=} |\dot{x}|^3\,|h|^3 + O(h^4). \end{aligned}$$

(1) Taylor expansion of $x(t) = x(s+h)$ about s.
(2) $\dot{x}(s+h) = \dot{x} + \ddot{x}\,h + \ldots$, and $\dot{x}\times\dot{x} = 0$.
(3) The term $-\dot{x}h\cdot(\dot{x}\times\ddot{x})h$ vanishes (repeated vector); the term $-\frac{1}{2}\ddot{x}h^2\cdot(\dot{x}\times\ddot{x})h$ vanishes for the same reason. What survives is of order h^4 (it involves the third derivative, i.e. the torsion of the curve).
(4) $|x-y| = |\dot{x}|\,|h| + O(h^2)$.

So the integrand is bounded near the diagonal and $\mathrm{Wr}(C)$ is a finite number. But it is *not* an integer and *not* an invariant. The degree argument fails, because the Gauss map $\Psi(s,t) = (x(t)-x(s))/|x(t)-x(s)|$ is undefined on the whole diagonal $s = t$: the torus now has a circle cut out of it, and the counting argument has a boundary to leak through. Deform a curve smoothly and Wr changes smoothly. A flat curve in a plane has $\mathrm{Wr} = 0$ (by Example 12.68); lift part of it into a helix and Wr takes any real value you like.

In quantum field theory the trouble is worse: the self-term involves the propagator $\langle A_i(x)A_j(y)\rangle$ at coincident points, which is singular. Nash states the difficulty this way: $L(C_l, C_l)$ has singularities at $x = y$. The cure, framing, keeps the answer topological at the price of one extra integer per component. We return to it in Section 12.12.8.

Try it: Orientation and mirror

Compute L for the Hopf link with the red component reversed, and for the $T(2,6)$ link. *Answer:* reversing one component flips all crossing signs: $L = -1$. $T(2,6)$ has six positive inter-component crossings: $L = 3$.

Reading Nash (pp. 344–345, (12.89)–(12.90))

Nash meets the linking number as the output of the abelian Chern–Simons path integral (12.89), written with $\varepsilon_{ijk}\oint \mathrm{d}x^i \oint \mathrm{d}y^j\,(x-y)^k/|x-y|^3$, and identifies it as Gauss's linking number $L(C_l, C_m)$ (12.90), "one of the oldest invariants in knot theory". He notes that it arose from Ampère's law and that this makes it an integer. Our chain (12.159) is exactly his (12.90); the diagram count, the degree of the Gauss map and the helicity are not in Nash, but they explain why the integral is an invariant and why a gauge theory produces it. His remark that $L(C_l, C_l)$ is singular is our Section 12.10.6.

Recap

- $L(C_1, C_2)$ is half the signed number of crossings between the two components in any diagram; it is an integer and a link invariant.
- Ampère: $L = \oint_{C_1} \vec{B} \cdot \mathrm{d}\vec{x}$ for the field of a unit current on C_2, equal to the signed number of times C_2 pierces a surface spanning C_1.
- Gauss: $L = \frac{1}{4\pi} \oint \oint (x - y) \cdot (\mathrm{d}x \times \mathrm{d}y)/|x - y|^3$, which is Biot–Savart inserted into Ampère (Nash (12.90)).
- The Gauss integral is the degree of the map $T^2 \to S^2$ giving the direction between the two points; this is why it is invariant, and counting preimages of a pole recovers the crossing count.
- L is symmetric, flips sign under reversing one component or under mirroring, and misses the Whitehead link and Borromean rings.
- The same integral for one curve, the writhe, is finite but not invariant. This is the origin of framing.

12.11 Knot polynomials: Alexander, Jones and their skein relations

The tricolouring number and the knot group proved that the trefoil is knotted. Neither can tell the right trefoil from the left one, and the knot group is hard to compare between two knots: deciding whether two presentations give the same group is, in general, hopeless. We want invariants that are *numbers we can compute and compare*. The answer found by Alexander in 1928, and in a much stronger form by Jones in 1984, is to attach to every knot a polynomial.

This section answers three questions.

1. Where does the variable t come from, and what does the Alexander polynomial measure? (The abelian shadow of the knot group, seen by an infinite "crystal" built from the knot complement.)

2. How do we compute these polynomials in practice? (By cutting the problem into smaller problems with *skein relations*, or by a state sum like a partition function in statistical mechanics.)

3. Why is the Jones polynomial the one a physicist should care about? (Because it sees chirality, and because Section 12.12 and Chapter 12 show it is a Wilson-loop expectation value in Chern–Simons theory.)

Nash states the results in two pages, [Nash pp. 340–342, (12.73)–(12.83)], and says the properties "can be easily checked". We check all of them.

12.11.1 Laurent polynomials, and what "defined up to $\pm t^n$" means

Working definition 12.25 (Laurent polynomial)

A **Laurent polynomial** (Nash: an *L*-**polynomial**) in t is a finite sum $\sum_{k=m}^{M} a_k t^k$ in which negative powers are allowed. We write $\mathbb{Z}[t, t^{-1}]$ for the Laurent polynomials with integer coefficients. When half-integer powers $t^{1/2}$ occur (for links with an even number of components) we work in $\mathbb{Z}[t^{1/2}, t^{-1/2}]$.

Examples: $t^2 - t + 1$, $t^{-1} - 1 + t$ and $t + t^3 - t^4$ are Laurent polynomials; $t^{-1/2} + t^{1/2}$ is one in $t^{1/2}$.

The **units** of $\mathbb{Z}[t, t^{-1}]$, the elements with an inverse inside $\mathbb{Z}[t, t^{-1}]$, are exactly $\pm t^n$: indeed $(\pm t^n)(\pm t^{-n}) = 1$, and a product of two Laurent polynomials with more than one term has more than one term. The Alexander polynomial will turn out to be defined only up to multiplication by a unit. This is the same kind of ambiguity as an overall phase of a wavefunction, and it is removed by a normalisation convention.

Working definition 12.26 (span and normalised form)

The **span** of a non-zero Laurent polynomial is (largest power) − (smallest power). It is unchanged by multiplying by $\pm t^n$, so it is a well-defined invariant of anything defined up to units. The **normalised form** is the unique representative with no negative powers and a positive constant term.

For $t^{-1} - 1 + t$: span $= 1 - (-1) = 2$, normalised form $t^2 - t + 1$.

12.11.2 The Alexander polynomial

Where the variable t comes from. At the end of Section 12.9.6 we saw that abelianising the knot group always gives $\mathbb{Z} = \{t^n\}$: the only abelian information is how many times a loop winds around the knot. Here t stands for "one meridian", one loop around the knot. The Alexander polynomial is the first-order non-abelian information, written in the language of this t.

Physical picture: The Alexander polynomial is a Bloch determinant

Cut the knot complement open along a surface bounded by the knot. We get a "unit cell" with two faces, a top copy and a bottom copy of the surface. Stack infinitely many unit cells, gluing the top of each to the bottom of the next. The result is the **infinite cyclic cover**: a one-dimensional "crystal" whose lattice translation is t (going once around the knot moves us up one cell). Loops in this crystal (its first homology) form a module over the translations, just as the states of an electron in a crystal form representations of the lattice translations. In Bloch's language $t = e^{ik}$ is the phase acquired per cell. The Alexander polynomial is the determinant of the matrix that presents this module, the analogue of $\det H(k)$ for a tight-binding Hamiltonian. Its zeros are the "Bloch phases" at which the crystal has an extra loop. The two recipes below compute this determinant: the Fox calculus from the knot group, and the Seifert matrix from the surface.

Recipe 1: from the knot group (Fox calculus)

The Fox derivative is a way of differentiating words in a group. It is the discrete analogue of the product rule.

Working definition 12.27 (Fox derivative)

For words in generators $x_1, \dots, x_n$ [products such as $x_1 x_2^{-1} x_1$ of generators and their inverses], the **Fox derivative** $\partial/\partial x_j$ [its values are integer combinations of words, such as $1 - x_1 x_2$] is fixed by three rules:

$$\frac{\partial x_i}{\partial x_j} = \delta_{ij}, \qquad \frac{\partial (uv)}{\partial x_j} = \frac{\partial u}{\partial x_j} + u\,\frac{\partial v}{\partial x_j}, \qquad \frac{\partial x_j^{-1}}{\partial x_j} = -x_j^{-1}.$$

(The third follows from the first two applied to $x_j x_j^{-1} = 1$: $0 = 1 + x_j\,\partial x_j^{-1}/\partial x_j$.) In words: scan the word from left to right; every time you meet x_j, write down the prefix before it with a plus sign; every time you meet x_j^{-1}, write the prefix *including* x_j^{-1} with a minus sign. *Recipe for* Δ. Take a Wirtinger presentation [one generator per arc of the diagram, one relation per crossing] with n generators and $n-1$ relations $r_i = 1$ (one relation can always be dropped). Form the $(n-1)\times n$ **Alexander matrix** $\partial r_i/\partial x_j$ and then set every generator equal to t. Delete any one column. The determinant of the remaining square matrix is $\Delta(t)$, up to a unit $\pm t^n$.

Worked example 12.71: Alexander polynomial of the trefoil from its group

From (12.154) the trefoil group has generators x, y (both Wirtinger generators, both mapping to t) and one relation $xyx = yxy$, i.e. $r = xyx\,y^{-1}x^{-1}y^{-1} = 1$. Differentiate with respect to x. The letter x appears in position 1 (prefix: empty word 1) and position 3 (prefix xy); x^{-1} appears in position 5 (prefix including it: $xyxy^{-1}x^{-1}$):

$$\begin{aligned}\frac{\partial r}{\partial x} &\overset{(1)}{=} 1 + xy - xyxy^{-1}x^{-1}\\ &\overset{(2)}{=} 1 + t^2 - t^{1+1+1-1-1}\\ &= t^2 - t + 1.\end{aligned}$$

(1) The scanning rule.
(2) Abelianise: every generator becomes t, and exponents add.

Now with respect to y: y in position 2 (prefix x), y^{-1} in positions 4 and 6 (prefixes $xyxy^{-1}$ and $xyxy^{-1}x^{-1}y^{-1}$):

$$\begin{aligned}\frac{\partial r}{\partial y} &\overset{(1)}{=} x - xyxy^{-1} - xyxy^{-1}x^{-1}y^{-1}\\ &\overset{(2)}{=} t - t^2 - t^0\\ &= -(t^2 - t + 1).\end{aligned}$$

(1) The scanning rule again.
(2) Abelianise; the last word is r itself, which becomes $t^0 = 1$.

The Alexander matrix is the 1×2 matrix $\big(t^2 - t + 1,\ -(t^2 - t + 1)\big)$. Delete either column: $\Delta(t) = t^2 - t + 1$, up to sign. This is Nash's (12.74).

Consistency check. There is a "fundamental formula" of the Fox calculus, $\sum_j (\partial r/\partial x_j)(x_j - 1) = r - 1$, which is the discrete version of $\sum_j \mathrm{d}x_j\,\partial_j f = \mathrm{d}f$. Abelianised, with $r = 1$, it

says $(t-1)\sum_j(\partial r/\partial x_j) = 0$: *the columns of the Alexander matrix add to zero.* Ours do: $(t^2-t+1)-(t^2-t+1)=0$. This is why deleting different columns gives the same answer.

Recipe 2: from a surface (the Seifert matrix)

The second recipe is geometric and makes the properties of Δ transparent.

Working definition 12.28 (Seifert surface and Seifert matrix)

A **Seifert surface** for an oriented knot K is a compact, connected, oriented surface S in $\mathbb{R}^3$ with boundary $\partial S = K$. If S has genus g [S is a disc with g handles attached], its first homology has a basis of $2g$ loops [every closed loop on S is, up to deformation on S, an integer combination of these $2g$] $a_1, \dots, a_{2g}$ on S. Let a_j^+ be the loop a_j pushed slightly off S in the direction of the positive normal. The **Seifert matrix** is the $2g \times 2g$ integer matrix

$$V_{ij} = L\big(a_i, a_j^+\big),$$

with L the linking number of Section 12.10. Then

$$\Delta_K(t) = \det\big(V - tV^T\big). \tag{12.165}$$

How to find S (Seifert's algorithm). In an oriented diagram, replace every crossing by its oriented smoothing L_0 [cut both strands and reconnect the four ends without a crossing, following the arrows] (Figure 12.43). The result is a set of disjoint oriented circles, the **Seifert circles**. Fill each with a disc (stacking nested discs at different heights). At each old crossing attach a small band with a half-twist joining the two discs there. The result is an oriented surface bounded by K.

A useful fact, which we will use: for any two loops a, b on the oriented surface,

$$L(a, b^+) - L(a^+, b) = a \cdot b, \tag{12.166}$$

the signed number of intersection points of a and b on S. The reason: pushing b up versus pushing a up differs, near each intersection point, by moving one strand from one side of the other to the other side, which changes the linking number by ± 1. Since $L(a^+, b) = L(a, b^-) = L(b^-, a)$ and pushing b down is the same as pushing a up, $L(a^+, b) = V_{ji}$. So $V - V^T = J$, the matrix of intersection numbers of the basis loops.

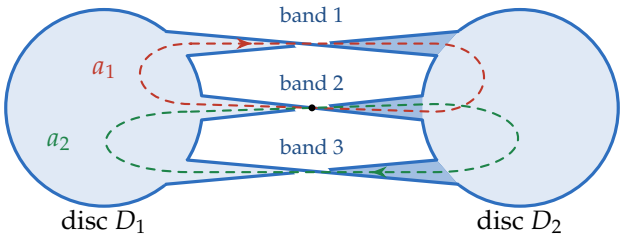


Figure 12.49: A Seifert surface for the trefoil, from Seifert's algorithm applied to the closed braid $\widehat{\sigma_1^3}$: two discs (one for each Seifert circle) joined by three bands, each with a half-twist (light shading: the front face of the band, dark: its back face, which the twist brings into view). The boundary of the surface is the trefoil. The loop a_1 (red) runs through bands 1 and 2, the loop a_2 (green) through bands 2 and 3; they meet once, at the dot in band 2. These two loops are the basis used for the Seifert matrix in Example 12.72.

Worked example 12.72: Seifert matrices of the trefoil and the figure-eight

Trefoil. Draw the trefoil as the closure of the two-strand braid σ_1^3 (Figure 12.51). The oriented smoothings give two Seifert circles, so the surface is two discs joined by three half-twisted bands (Figure 12.49). Its Euler characteristic is $\chi = (\text{discs}) - (\text{bands}) = 2 - 3 = -1$, and a surface with one boundary circle has $\chi = 1 - 2g$, so $g = 1$. A basis of loops: a_1 runs through bands 1 and 2, a_2 through bands 2 and 3.

The entries come from two facts. A loop running once through a band with a (right-handed) half-twist links its own push-off by $-\frac{1}{2}$ for that band; a_1 uses two bands, so $V_{11} = -1$, and likewise $V_{22} = -1$. The loops a_1, a_2 cross once (in band 2), so by (12.166) $V_{12} - V_{21} = \pm 1$; pushing a_2 up lifts it over a_1, which links once, while pushing a_1 up does not. So

$$V = \begin{pmatrix} -1 & 1 \\ 0 & -1 \end{pmatrix}, \qquad V - V^T = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

Then

$$\begin{aligned} \Delta(t) &\overset{(1)}{=} \det \begin{pmatrix} -1+t & 1 \\ -t & -1+t \end{pmatrix} \\ &\overset{(2)}{=} (t-1)^2 + t \\ &= t^2 - t + 1. \end{aligned}$$

(1) $V - tV^T$ entry by entry: the $(1,2)$ entry is $1 - t \cdot 0$, the $(2,1)$ entry is $0 - t \cdot 1$.
(2) 2×2 determinant: $(t-1)^2 - (1)(-t)$.

The same answer as the Fox calculus, as it must be.

Figure-eight. Seifert's algorithm on the standard diagram gives a genus-one surface with, in a suitable basis, $V = \begin{pmatrix} 1 & 1 \\ 0 & -1 \end{pmatrix}$ (one loop twists one way, the other the opposite way, which is the amphichirality of 4_1 in miniature). Then

$$\begin{aligned} \Delta_{4_1}(t) &\overset{(1)}{=} \det \begin{pmatrix} 1-t & 1 \\ -t & -1+t \end{pmatrix} \overset{(2)}{=} -(1-t)^2 + t \\ &= -(t^2 - 3t + 1) \;\doteq\; t^2 - 3t + 1, \end{aligned}$$

(1) $V - tV^T$ entry by entry.
(2) $(1-t)(t-1) - (1)(-t)$.

where $\doteq$ means "equal up to a unit". So $\Delta_{4_1} \neq \Delta_{3_1}$: the figure-eight knot is not a trefoil.

The properties Nash lists, (12.73). Nash says Δ is only determined up to $\mp t^n$, that $\Delta(1) = \mp 1$, that $\Delta(1/t) = t^n \Delta(t)$ for some even n, and that the coefficients are integers. All four follow from (12.165).

Integer coefficients. V is an integer matrix, so $\det(V - tV^T)$ is a polynomial in t with integer coefficients.

Only up to units. Different Seifert surfaces and bases give different matrices V. A change of basis $V \mapsto PVP^T$ with $\det P = \pm 1$ changes the determinant by $(\det P)^2 = 1$. Adding a tube to the surface (the other move relating two Seifert surfaces) enlarges V by a block and multiplies the determinant by $\pm t$. Hence Δ is well defined up to $\pm t^n$.

Value at $t = 1$.

$$\begin{aligned}\Delta(1) &\overset{(1)}{=} \det(V - V^T)\\ &\overset{(2)}{=} \det J\\ &\overset{(3)}{=} \det\begin{pmatrix} 0 & 1\\ -1 & 0\end{pmatrix}^{g} = 1.\end{aligned}$$

(1) Set $t = 1$ in (12.165).
(2) By (12.166), $V - V^T$ is the intersection matrix J of the loops on S.
(3) On a genus-g surface one can choose loops $a_1, b_1, \ldots, a_g, b_g$ with a_i meeting only b_i, once: J is block diagonal with g blocks $\left(\begin{smallmatrix} 0 & 1\\ -1 & 0\end{smallmatrix}\right)$, each of determinant 1.

After multiplying by an arbitrary unit $\pm t^n$ we get Nash's $\Delta(1) = \mp 1$.

Symmetry.

$$\begin{aligned}\Delta(t) &\overset{(1)}{=} \det\big((V - tV^T)^T\big) \overset{(2)}{=} \det\big(V^T - tV\big)\\ &\overset{(3)}{=} \det\big(-t\,(V - t^{-1}V^T)\big) \overset{(4)}{=} (-t)^{2g}\det\big(V - t^{-1}V^T\big)\\ &\overset{(5)}{=} t^{2g}\,\Delta(1/t).\end{aligned} \tag{12.167}$$

(1) A matrix and its transpose have the same determinant.
(2) $(V - tV^T)^T = V^T - tV$.
(3) Factor out $-t$: $-t(V - t^{-1}V^T) = -tV + V^T$.
(4) The matrix is $2g \times 2g$, so pulling a scalar out of the determinant gives its $2g$-th power.
(5) Definition (12.165) at t^{-1}, and $2g$ is even.

So $\Delta(1/t) = t^{-2g}\Delta(t)$: Nash's (12.73) with the even integer $n = -2g$. The coefficients read the same forwards and backwards (for $t^2 - t + 1$: $1, -1, 1$; for $t^2 - 3t + 1$: $1, -3, 1$). The chain also shows that the degree of Δ is at most $2g$: *the Alexander polynomial bounds the genus of the knot from below*. For the trefoil, degree 2 means no Seifert surface of genus 0 (a disc) exists, which is a third proof that it is knotted.

Normalisations, (12.74)–(12.75). Since Δ is defined up to $\pm t^n$, these are all the same invariant of the trefoil:

$$t^2 - t + 1, \qquad \frac{1}{t^2} - \frac{1}{t} + 1 = t^{-2}(t^2 - t + 1), \qquad t - 1 + \frac{1}{t} = t^{-1}(t^2 - t + 1).$$

The first is the normalised form (no negative powers, positive constant term). Its degree is 2, which equals the span. The last is the **symmetric** or **Conway** normalisation Δ^{Sym}: it is invariant under $t \leftrightarrow t^{-1}$ and has $\Delta^{\mathrm{Sym}}(1) = +1$. By (12.167) every knot has one: multiply by t^{-g} and fix the sign so that $\Delta(1) = 1$.

Watch out: Two misprints in Nash (12.75)

Nash prints the symmetric form of the trefoil polynomial as "$t - 1 - \frac{1}{t}$", and says it satisfies "$\Delta(0) = 1$". Both are misprints. Multiplying $t^2 - t + 1$ by t^{-1} gives $t - 1 + \frac{1}{t}$ (a plus sign); and a Laurent polynomial with a t^{-1} term is not even defined at $t = 0$. The property meant is $\Delta^{\mathrm{Sym}}(1) = 1$: indeed $1 - 1 + 1 = 1$.

Mirror images. Under the reflection ρ of (12.149), linking numbers change sign and the positive side of the surface is exchanged with the negative side, so the Seifert matrix becomes $\bar{V} = -V^T$. Then

$$\Delta_{\bar{K}}(t) \overset{(1)}{=} \det\!\left(-V^T + tV\right) \overset{(2)}{=} \det\!\left((-V + tV^T)^T\right) \overset{(3)}{=} (-1)^{2g} \det\!\left(V - tV^T\right) = \Delta_K(t).$$

(1) Formula (12.165) with $\bar{V} = -V^T$, whose transpose is $-V$.
(2) Write the matrix as a transpose.
(3) Transpose-invariance of the determinant, and pull out -1 from a $2g \times 2g$ matrix.

So, as Nash says, the right and left trefoils have the same Alexander polynomial. The Alexander polynomial cannot see chirality, just like the knot group it came from.

The determinant and colourings. The integer $|\Delta(-1)|$ is called the **determinant** of the knot. For the trefoil $|1 + 1 + 1| = 3$; for the figure-eight $|1 + 3 + 1| = 5$. A knot has a non-trivial Fox p-colouring (for p an odd prime) exactly when p divides its determinant. So the trefoil is tricolourable and the figure-eight is not (it is 5-colourable instead), in agreement with Section 12.9.5. The toy invariant of the last section is the Alexander polynomial evaluated at $t = -1$ and reduced mod 3.

12.11.3 Skein relations: computing by cutting and reconnecting

Fox calculus and Seifert matrices work, but for a big diagram they are laborious. Conway (1969) found that the Alexander polynomial obeys a local *recursion* which Nash quotes. It is the idea that later made the Jones polynomial computable, and it is the relation that Chern–Simons theory must reproduce.

Working definition 12.29 (skein triple, Nash (12.79))

Three oriented links L_+, L_-, L_0 form a **skein triple** if their diagrams are identical except inside one small disc, where they look like the three pictures of Figure 12.43: a positive crossing, a negative crossing, and the oriented smoothing [the two strands cut and reconnected without a crossing, following the arrows]. If $L_\pm$ have n crossings, L_0 has $n - 1$.

The number of components changes in a definite way: if the two strands at the crossing belong to the same component of $L_\pm$, smoothing cuts that component into two; if they belong to different components, smoothing joins them into one. So L_0 has one component more or one fewer than $L_\pm$.

Key idea: The Alexander–Conway skein relation, Nash (12.80)

With Δ in its symmetric (Conway) normalisation,

$$\Delta_{L_+}(t) - \Delta_{L_-}(t) = \left(t^{1/2} - t^{-1/2}\right) \Delta_{L_0}(t), \qquad \Delta_{\text{unknot}} = 1. \tag{12.168}$$

It is convenient to write $z = t^{1/2} - t^{-1/2}$; then Δ becomes a polynomial in z, the **Conway polynomial** $\nabla(z)$.

Why these two rules determine Δ for every link. Pick a crossing. Either L_+ or L_- is the link we want; the relation expresses it through the other one (same number of crossings, one crossing changed) and L_0 (one crossing fewer). Recall from Section 12.9.2 that changing suitable

crossings turns any diagram into a "descending" diagram of an unlink. So by induction (on the number of crossings, and for a fixed number, on the number of crossing changes needed), everything reduces to the values on unlinks. This is Nash's remark that "iteration of this relation expresses any $\Delta_L(t)$ in terms of the Alexander polynomial of a finite number of unlinked, unknotted circles".



Figure 12.50: The skein triple at a kink. Both L_+ and L_- are unknots (an R1 move removes the kink), while the oriented smoothing L_0 splits off a small circle, giving the two-component unlink. Feeding this triple into a skein relation computes the value of the invariant on unlinks.

Worked example 12.73: The unlink, the Hopf link and the trefoil by skein

The two-component unlink. Apply the skein relation to the kink in Figure 12.50. L_+ and L_- are both unknots (an R1 move removes the kink), and L_0 is two separate circles, the unlink U_2:

$$\begin{aligned} z\,\Delta_{U_2} &\overset{(1)}{=} \Delta_{L_+} - \Delta_{L_-} \\ &\overset{(2)}{=} 1 - 1 = 0. \end{aligned}$$

(1) Skein relation (12.168).
(2) Both are the unknot.

So $\Delta_{U_2} = 0$, and the same argument with extra circles gives $\Delta_{U_p} = 0$ for all $p \geq 2$. (Geometrically: two disjoint circles bound a disconnected surface; tubing it together produces a Seifert matrix with a zero row.)

The positive Hopf link H_+ is the closure of σ_1^2. At either crossing, $L_+ = H_+$, switching gives $L_- = U_2$, and smoothing gives $L_0 =$ the closure of σ_1, an unknot:

$$\begin{aligned} \Delta_{H_+} &\overset{(1)}{=} \Delta_{U_2} + z\,\Delta_{\text{unknot}} \\ &\overset{(2)}{=} 0 + z = t^{1/2} - t^{-1/2}. \end{aligned}$$

(1) Skein relation solved for Δ_{L_+}.
(2) The two values just found.

The trefoil is the closure of σ_1^3. At one crossing, $L_+ = 3_1$, switching gives the closure of $\sigma_1^2\sigma_1^{-1} = \sigma_1$, an unknot, and smoothing gives the closure of σ_1^2, the Hopf link:

$$\begin{aligned} \Delta_{3_1} &\overset{(1)}{=} \Delta_{\text{unknot}} + z\,\Delta_{H_+} \\ &\overset{(2)}{=} 1 + z^2 \\ &\overset{(3)}{=} 1 + t - 2 + t^{-1} = t - 1 + t^{-1}. \end{aligned}$$

(1) Skein relation.
(2) $\Delta_{H_+} = z$ from above.

(3) $z^2 = (t^{1/2} - t^{-1/2})^2 = t - 2 + t^{-1}$.

This is Δ^{Sym} of the trefoil once more, now in half a dozen lines. The Conway polynomial is $\nabla_{3_1} = 1 + z^2$.

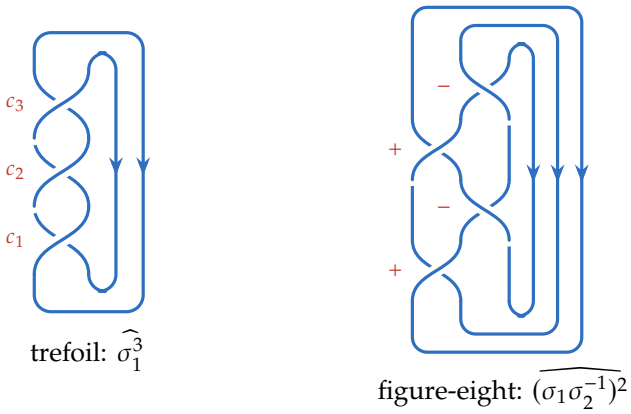


Figure 12.51: Knots as closed braids. Left: the right trefoil as the closure of σ_1^3, three positive crossings c_1, c_2, c_3 between two strands, the ends joined round the back. Right: the figure-eight knot as the closure of $\sigma_1\sigma_2^{-1}\sigma_1\sigma_2^{-1}$ on three strands; the crossing signs alternate $+, -, +, -$ from the bottom up, so the writhe is 0.

Worked example 12.74: The figure-eight knot by skein

Write 4_1 as the closure of the three-strand braid $\sigma_1\sigma_2^{-1}\sigma_1\sigma_2^{-1}$ (Figure 12.51). Apply the skein relation at the first (positive) crossing (Figure 12.52).

- $L_+ = 4_1$.
- L_- is the closure of $\sigma_1^{-1}\sigma_2^{-1}\sigma_1\sigma_2^{-1}$. It is an unknot (its Jones polynomial, computed below by the same method, is 1, and one can untangle it by hand).
- L_0 is the closure of $\sigma_2^{-1}\sigma_1\sigma_2^{-1}$. By cyclically moving the last letter to the front (conjugation, which does not change the closure) it is the closure of $\sigma_2^{-2}\sigma_1$. The strand that passes through σ_1 only once can be pulled straight (a Markov move, the braid version of R1), leaving the closure of σ^{-2} on two strands: the *negative* Hopf link H_-, with $\Delta_{H_-} = -z$ (switch both crossings of H_+: the skein computation above with every sign reversed).

Then

$$\begin{aligned}\Delta_{4_1} &\overset{(1)}{=} \Delta_{\text{unknot}} + z\,\Delta_{H_-}\\ &\overset{(2)}{=} 1 - z^2\\ &\overset{(3)}{=} 1 - t + 2 - t^{-1} = -t + 3 - t^{-1},\end{aligned}$$

(1) Skein relation, solved for Δ_{L_+}.
(2) $\Delta_{H_-} = -z$.
(3) $z^2 = t - 2 + t^{-1}$.

in agreement with the Seifert matrix (up to the unit $-t^{-1}$).

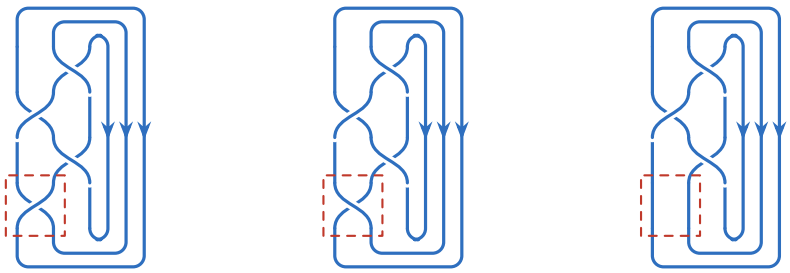


Figure 12.52: The skein triple used for the figure-eight knot, at the lowest crossing (dashed box). Switching it gives an unknot L_-; smoothing it gives the closure of $\sigma_2^{-1}\sigma_1\sigma_2^{-1}$, which is the negative Hopf link.

Watch out: Sign conventions in skein relations

Nash writes (12.80) as $\Delta_{L_+} - \Delta_{L_-} + (t^{1/2} - t^{-1/2})\Delta_{L_0} = 0$, which is (12.168) with $z \to -z$, i.e. $t^{1/2} \to -t^{1/2}$ (or $L_+ \leftrightarrow L_-$). Nash's pictures of $L_\pm$ carry no arrows, so his convention cannot be read off. For *knots* nothing depends on it: $\nabla_K(z)$ contains only even powers of z (a knot has an odd number, one, of components, and each skein step changes the number of components by one and the power of z by one). For links with an even number of components the overall sign of Δ depends on the convention. We use (12.168) with the crossing signs of Figure 12.43 throughout.

12.11.4 The Kauffman bracket: a knot polynomial as a partition function

The Jones polynomial was discovered inside von Neumann algebras, but the quickest honest route to it is Kauffman's (1987). It is a *state sum*, built exactly like the partition function of a lattice model in statistical mechanics: sum over configurations of local choices, each weighted by a Boltzmann-like factor. The "spins" are the ways of resolving each crossing.

Working definition 12.30 (Kauffman bracket)

The **Kauffman bracket** assigns to every unoriented link diagram D [no arrows on the strands] a Laurent polynomial $\langle D \rangle$ in a variable A, fixed by three rules:

$$\begin{aligned} \langle \bigcirc \rangle &= 1, \\ \langle D \sqcup \bigcirc \rangle &= d\,\langle D \rangle, \qquad d := -A^2 - A^{-2}, \\ \left\langle \text{crossing} \right\rangle &= A \left\langle \text{A-smoothing} \right\rangle + A^{-1} \left\langle \text{B-smoothing} \right\rangle. \end{aligned} \tag{12.169}$$

The **A-smoothing** at a crossing joins the two regions that the over-strand sweeps out when rotated anticlockwise until it meets the under-strand; the B-smoothing joins the other two (Figure 12.53). A disjoint circle [an extra circle drawn apart from D, crossing nothing] multiplies by d.

Closed formula (state sum). A **state** s is a choice of A or B at every crossing. Smoothing every crossing according to s gives a collection of $|s|$ disjoint circles. With $a(s)$ and $b(s)$ the numbers of A's and B's,

$$\langle D \rangle = \sum_{\text{states } s} A^{a(s)-b(s)}\, d^{\,|s|-1}. \tag{12.170}$$

For a crossing drawn with the over-strand from south-west to north-east, rotating it

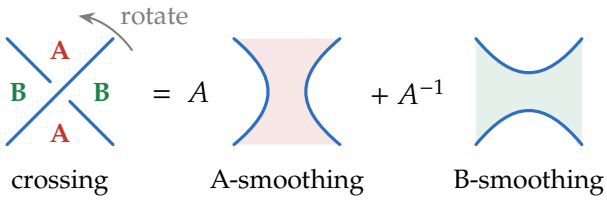


Figure 12.53: The Kauffman bracket rule. Rotate the over-strand anticlockwise: the two regions it sweeps are labelled A, the other two B. The A-smoothing opens a channel joining the two A regions (here: the vertical smoothing), the B-smoothing joins the two B regions (the horizontal one). The weights are A and A^{-1}.

anticlockwise sweeps the regions above and below the crossing. The A-smoothing therefore opens a channel between the top and bottom regions and gives the vertical smoothing ") (". If both strands point upwards this is a *positive* crossing L_+, and its A-smoothing is the oriented smoothing L_0. For a negative crossing the A-smoothing is the horizontal one. We use this constantly.

Dictionary: State sum ↔ partition function

crossing	lattice site
choice A/B at a crossing	spin up/down
weight $A^{\pm 1}$ per crossing	local Boltzmann weight
factor d per closed loop	fugacity of a loop (loop models)
$\langle D \rangle$	partition function $Z = \sum_s e^{-\beta E(s)}$

Kauffman showed that at special values of A the bracket of a planar graph is the partition function of the Potts model on it. The algebra that multiplies the smoothings (below) is the Temperley–Lieb algebra, first found in the transfer matrices of that model.

Invariance under R2 and R3. Take two strands forming an R2 configuration and expand both crossings. Of the four states, one gives the two strands passing straight by each other (") (", with weight $A \cdot A^{-1} = 1$. The other three give the turned-back picture "≍": two of them (AA and BB) directly, with weights A^2 and A^{-2}, and the remaining mixed state with an extra little circle in the middle, weight $A^{-1} A\, d = d$. Therefore

$$\begin{aligned} \big\langle \text{R2 tangle} \big\rangle &\overset{(1)}{=} \big\langle\,)\,(\,\big\rangle + \big(A^2 + A^{-2} + d\big)\big\langle \asymp \big\rangle \\ &\overset{(2)}{=} \big\langle\,)\,(\,\big\rangle. \end{aligned} \tag{12.171}$$

(1) The four states just listed; the small circle is removed by the rule $\langle D \sqcup \bigcirc \rangle = d\langle D \rangle$.

(2) $d = -A^2 - A^{-2}$: *this is why d has the value it has*. The loop fugacity is chosen precisely to cancel the unwanted term.

R3 then follows from R2. Expand the crossing of the two lower strands in both sides of the R3 move. The A-terms on the two sides are equal diagrams outright, and the B-terms are related by two R2 moves, which we just showed are harmless. So $\langle D \rangle$ is unchanged by R2 and R3.

R1 and the writhe. A kink is not harmless. For the positive kink of Figure 12.50 (left), the A-smoothing is the oriented one and splits off a small circle, while the B-smoothing just

straightens the strand:

$$\begin{aligned}\big\langle \text{positive kink} \big\rangle &\overset{(1)}{=} A\, d \,\big\langle \,\big|\, \big\rangle + A^{-1} \big\langle \,\big|\, \big\rangle \\ &\overset{(2)}{=} \big(-A^3 - A^{-1} + A^{-1}\big)\big\langle \,\big|\, \big\rangle = -A^3 \,\big\langle \,\big|\, \big\rangle . \end{aligned} \tag{12.172}$$

(1) Bracket rule at the kink's crossing; the A-state has an extra circle.
(2) $Ad = -A^3 - A^{-1}$.

A negative kink gives $-A^{-3}$ in the same way (A and B exchange roles). So each R1 move multiplies $\langle D \rangle$ by $-A^{\pm 3}$ and changes the writhe $w(D)$ of Section 12.9.4 by ± 1. The two defects cancel in

$$f_L(A) \coloneqq \big(-A^3\big)^{-w(D)} \langle D \rangle , \tag{12.173}$$

for any oriented diagram D of L. f_L is unchanged by R1, R2 and R3, so it is an invariant of oriented links.

Physical picture: The bracket is an invariant of ribbons

$\langle D \rangle$ on its own is invariant under R2 and R3 only ("regular isotopy"). Think of the knot as a flat ribbon lying in the page. R2 and R3 moves slide the ribbon about; an R1 move puts a full twist into it. So $\langle D \rangle$ is an invariant of *framed* knots, and $-A^3$ is the price of one unit of framing. We shall meet exactly this structure in Chern–Simons theory: Wilson loops need a framing, and changing it multiplies $\langle W \rangle$ by a phase (Section 12.12.8). The writhe correction (12.173) is the knot theorist's way of choosing the canonical framing.

Worked example 12.75: The Hopf link

H_+ is the closure of σ_1^2, two positive crossings. Both crossings are positive, so A-smoothing = oriented smoothing = two vertical strands "$\|$", and B-smoothing = cup-cap "$\asymp$". The four states:

state	loops $\lvert s \rvert$	$a - b$	$A^{a-b} d^{\lvert s \rvert - 1}$
AA	2	2	$A^2 d$
AB	1	0	1
BA	1	0	1
BB	2	−2	$A^{-2} d$

(AA: two parallel strands closed up, two circles. One B: the cup-cap joins the two strands into one circle. BB: the cap of one box and the cup of the next close off a small circle, and the rest forms a second one.)

$$\begin{aligned}\langle H_+ \rangle &\overset{(1)}{=} A^2 d + 2 + A^{-2} d \\ &\overset{(2)}{=} -A^4 - 1 + 2 - 1 - A^{-4} \\ &= -A^4 - A^{-4} .\end{aligned}$$

(1) Sum of the table.
(2) $A^{\pm 2} d = -A^{\pm 4} - 1$.

With $w(H_+) = +2$, $f_{H_+} = (-A^3)^{-2}(-A^4 - A^{-4}) = -A^{-2} - A^{-10}$.

Worked example 12.76: The trefoil: the full state sum

Take the right trefoil as the closure of σ_1^3, with crossings c_1, c_2, c_3 (Figure 12.51). All three are positive, so again A = "$\|$" and B = "$\asymp$" in each box. There are $2^3 = 8$ states, drawn in Figure 12.54. Counting loops:

- AAA: three identity boxes; the closure gives 2 circles.
- One B (three states): the single cup-cap connects the two strands: 1 circle.
- Two B's (three states): between two cup-caps there is either a small circle trapped (if they are adjacent) or, going around the closure, a second circle (if not). Either way 2 circles.
- BBB: two small circles between the three cup-caps and one big one: 3 circles.

So

$$\begin{aligned}\langle 3_1\rangle &\overset{(1)}{=} A^3 d + 3A + 3A^{-1} d + A^{-3} d^2 \\ &\overset{(2)}{=} \left(-A^5 - A\right) + 3A + \left(-3A - 3A^{-3}\right) + \left(A + 2A^{-3} + A^{-7}\right) \\ &\overset{(3)}{=} -A^5 - A^{-3} + A^{-7}. \end{aligned} \tag{12.174}$$

(1) State sum (12.170): the weight A^{a-b} is A^3, A, A^{-1}, A^{-3} for 0, 1, 2, 3 B's, and the loop counts are 2, 1, 2, 3.
(2) Expand: $A^3 d = -A^5 - A$; $3A^{-1} d = -3A - 3A^{-3}$; $A^{-3} d^2 = A^{-3}(A^4 + 2 + A^{-4})$.
(3) Collect: the A terms cancel, $-1 + 3 - 3 + 1 = 0$; the A^{-3} terms give $-3 + 2 = -1$.

The writhe is $w = +3$, so

$$\begin{aligned} f_{3_1}(A) &\overset{(1)}{=} (-A^3)^{-3}\left(-A^5 - A^{-3} + A^{-7}\right) \\ &\overset{(2)}{=} -A^{-9}\left(-A^5 - A^{-3} + A^{-7}\right) \\ &= A^{-4} + A^{-12} - A^{-16}. \end{aligned} \tag{12.175}$$

(1) Definition (12.173).
(2) $(-A^3)^{-3} = -A^{-9}$.

12.11.5 The Jones polynomial

Working definition 12.31 (Jones polynomial)

The **Jones polynomial** of an oriented link L is

$$V_L(t) = f_L(A)\Big|_{A=t^{-1/4}} = \left[\left(-A^3\right)^{-w(D)}\langle D\rangle\right]_{A=t^{-1/4}}. \tag{12.176}$$

Recipe: draw a diagram, compute the bracket by the state sum, multiply by $(-A^3)^{-w}$ [w is the writhe, the sum of the crossing signs. This factor cancels the change of $\langle D\rangle$ under R1], substitute $A = t^{-1/4}$. For a knot only integer powers of t survive; for links with an even number of components, half-integer powers appear.

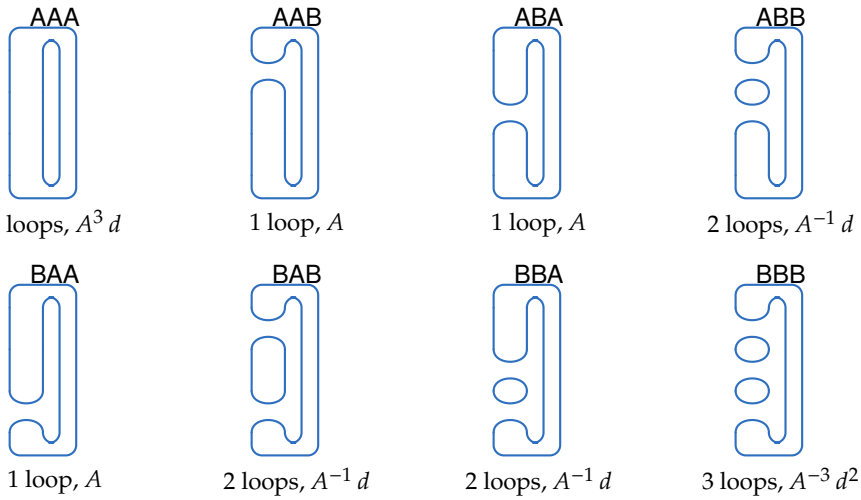


Figure 12.54: The eight states of the trefoil diagram $\widehat{\sigma_1^3}$. Each crossing is replaced by its A-smoothing (two vertical strands) or its B-smoothing (a cup and a cap); the letters list the choices from the bottom crossing c_1 to the top crossing c_3. Under each state: the number $|s|$ of loops and the weight $A^{a-b}d^{|s|-1}$. Summing the weights gives (12.174).

Worked example 12.77: The trefoil and its mirror image, Nash (12.76)

From (12.175), with $A^{-4} = t$:

$$V_{3_1}(t) = t + t^3 - t^4. \tag{12.177}$$

This is Nash's V_L for the clover leaf; our right trefoil (positive crossings) is his.
For the mirror image, every crossing is switched. Switching a crossing exchanges its A- and B-smoothings (rotate the over-strand the other way), so the state sum (12.170) is the same with $A \leftrightarrow A^{-1}$; and d is symmetric under this. The writhe changes sign. Hence

$$f_{\bar{L}}(A) \overset{(1)}{=} \left(-A^3\right)^{+w(D)} \langle D \rangle \big|_{A \to A^{-1}} \overset{(2)}{=} f_L(A^{-1}),$$

(1) Bracket with $A \to A^{-1}$ and writhe $-w$.
(2) $(-A^3)^w = \left(-(A^{-1})^3\right)^{-w}$, so this is (12.173) evaluated at A^{-1}.

and $A \to A^{-1}$ means $t \to t^{-1}$:

$$V_{\bar{L}}(t) = V_L(1/t), \qquad V_{\bar{3}_1}(t) = \frac{1}{t} + \frac{1}{t^3} - \frac{1}{t^4}.$$

This is the second half of Nash's (12.76) and the third property in his (12.77).

Key idea: The Jones polynomial detects chirality

$t + t^3 - t^4 \neq t^{-1} + t^{-3} - t^{-4}$. So the right and left trefoils are *not isotopic*, although they are equivalent (via the reflection ρ) and have the same knot group and the same Alexander polynomial. This settles the claim of Section 12.9.3. A knot with $V_K(t) \neq V_K(1/t)$ can never be amphichiral.

The Jones skein relation, derived. Nash quotes (12.81). We derive it from the bracket. At a crossing, let L_∞ denote the *unoriented* smoothing ("$\asymp$" when the strands point up). From the

A/B rule of the previous subsection:

$$\langle L_+\rangle = A\langle L_0\rangle + A^{-1}\langle L_\infty\rangle,$$
$$\langle L_-\rangle = A^{-1}\langle L_0\rangle + A\langle L_\infty\rangle.$$

Multiply the first by A, the second by A^{-1} and subtract: L_∞ drops out and

$$A\langle L_+\rangle - A^{-1}\langle L_-\rangle = (A^2 - A^{-2})\langle L_0\rangle. \tag{12.178}$$

Now convert to f. Let w_0 be the writhe of L_0; since the other crossings are the same, $w(L_\pm) = w_0 \pm 1$. Using $\langle D\rangle = (-A^3)^{w(D)} f$:

$$\begin{aligned} A(-A^3)^{w_0+1} f_{L_+} - A^{-1}(-A^3)^{w_0-1} f_{L_-} &\overset{(1)}{=} (A^2 - A^{-2})(-A^3)^{w_0} f_{L_0} \\ \Longrightarrow \quad -A^4 f_{L_+} + A^{-4} f_{L_-} &\overset{(2)}{=} (A^2 - A^{-2}) f_{L_0} \\ \Longrightarrow \quad t^{-1} V_{L_+} - t\, V_{L_-} &\overset{(3)}{=} \left(t^{1/2} - t^{-1/2}\right) V_{L_0}. \end{aligned} \tag{12.179}$$

(1) Substitute $\langle D\rangle = (-A^3)^w f$ into (12.178).
(2) Divide by $(-A^3)^{w_0}$: $A(-A^3) = -A^4$ and $A^{-1}(-A^3)^{-1} = -A^{-4}$.
(3) Put $A = t^{-1/4}$, so $A^4 = t^{-1}$, $A^{-4} = t$, $A^{\pm 2} = t^{\mp 1/2}$, and multiply by -1.

Together with $V_{\text{unknot}} = 1$ this determines V on all links, by the same induction as for Δ.

Watch out: Nash's (12.81) is the mirror convention

Nash prints $tV_{L_+} - t^{-1}V_{L_-} + (t^{1/2} - t^{-1/2})V_{L_0} = 0$. This is (12.179) with $t \to t^{-1}$ (equivalently with $L_+ \leftrightarrow L_-$). With his relation read literally in our crossing convention, the trefoil with positive crossings would get $t^{-1} + t^{-3} - t^{-4}$. Since his pictures of $L_\pm$ have no arrows, the two statements are consistent once L_+ and L_- are exchanged. What is convention-independent: the trefoil and its mirror get $t + t^3 - t^4$ and $t^{-1} + t^{-3} - t^{-4}$, one each.

Unlinks, Nash (12.83). Apply (12.179) to the kink of Figure 12.50: $L_\pm$ are unknots and $L_0 = U_2$:

$$\begin{aligned} \left(t^{1/2} - t^{-1/2}\right) V_{U_2} &\overset{(1)}{=} t^{-1}\cdot 1 - t\cdot 1 \\ \Longrightarrow \quad V_{U_2} &\overset{(2)}{=} -\frac{t - t^{-1}}{t^{1/2} - t^{-1/2}} \overset{(3)}{=} -\left(t^{1/2} + t^{-1/2}\right). \end{aligned}$$

(1) Skein relation with $V_{\text{unknot}} = 1$.
(2) Divide.
(3) $t - t^{-1} = (t^{1/2} - t^{-1/2})(t^{1/2} + t^{-1/2})$.

The same argument with $p-1$ extra circles present (they are spectators) shows that each additional split unknotted component multiplies V by this factor:

$$V_{U_p}(t) = \left\{-\frac{t - t^{-1}}{t^{1/2} - t^{-1/2}}\right\}^{p-1} = \left(-t^{1/2} - t^{-1/2}\right)^{p-1}, \tag{12.180}$$

which is Nash's (12.83). It is also immediate from the bracket: p disjoint circles have $\langle U_p\rangle = d^{p-1}$ and $w = 0$, and $d = -A^2 - A^{-2} = -t^{-1/2} - t^{1/2}$.

Worked example 12.78: Hopf link and trefoil by the Jones skein relation

For H_+ at one crossing, $L_+ = H_+$, $L_- = U_2$, $L_0 =$ unknot:

$$\begin{aligned} t^{-1}V_{H_+} &\overset{(1)}{=} t\,V_{U_2} + \left(t^{1/2} - t^{-1/2}\right) \\ &\overset{(2)}{=} -t^{3/2} - t^{1/2} + t^{1/2} - t^{-1/2} = -t^{3/2} - t^{-1/2}, \end{aligned}$$

(1) Skein relation (12.179) solved for $t^{-1}V_{L_+}$.
(2) $V_{U_2} = -(t^{1/2} + t^{-1/2})$.

so $V_{H_+} = -t^{1/2} - t^{5/2}$. (Check with the bracket: $f_{H_+} = -A^{-2} - A^{-10}$ and $A^{-2} = t^{1/2}$, $A^{-10} = t^{5/2}$. It agrees.)
For the trefoil, $L_+ = 3_1$, $L_- =$ unknot, $L_0 = H_+$:

$$\begin{aligned} t^{-1}V_{3_1} &\overset{(1)}{=} t + \left(t^{1/2} - t^{-1/2}\right)\left(-t^{1/2} - t^{5/2}\right) \\ &\overset{(2)}{=} t - t - t^3 + 1 + t^2 \\ \Longrightarrow \quad V_{3_1} &= t + t^3 - t^4. \end{aligned}$$

(1) Skein relation with $V_{L_-} = 1$ and $V_{L_0} = V_{H_+}$.
(2) Multiply out: $t^{1/2}\cdot(-t^{1/2}) = -t$, $t^{1/2}\cdot(-t^{5/2}) = -t^3$, $-t^{-1/2}\cdot(-t^{1/2}) = 1$, $-t^{-1/2}\cdot(-t^{5/2}) = t^2$.

The state sum and the skein relation agree, as they must.

Worked example 12.79: The figure-eight knot is amphichiral

Use the skein triple of Figure 12.52: $L_+ = 4_1$, $L_- =$ unknot, $L_0 = H_-$. The negative Hopf link is the mirror of H_+, so $V_{H_-}(t) = V_{H_+}(1/t) = -t^{-1/2} - t^{-5/2}$. Then

$$\begin{aligned} t^{-1}V_{4_1} &\overset{(1)}{=} t + \left(t^{1/2} - t^{-1/2}\right)\left(-t^{-1/2} - t^{-5/2}\right) \\ &\overset{(2)}{=} t - 1 - t^{-2} + t^{-1} + t^{-3} \\ \Longrightarrow \quad V_{4_1}(t) &\overset{(3)}{=} t^2 - t + 1 - t^{-1} + t^{-2}. \end{aligned}$$

(1) Skein relation (12.179).
(2) Multiply out the four products.
(3) Multiply by t.

The answer is symmetric under $t \leftrightarrow t^{-1}$, consistent with the figure-eight being isotopic to its mirror image (it is: a string model can be rearranged into its mirror in a few moves). A direct state sum over the $2^4 = 16$ states of the braid-closure diagram gives $\langle 4_1\rangle = A^8 - A^4 + 1 - A^{-4} + A^{-8}$ with $w = 0$, the same polynomial. The same skein method applied to L_- itself returns $V_{L_-} = 1$, as claimed in Example 12.74.

The special values, Nash (12.77) and (12.78). Nash says these "can be easily checked for the clover leaf". Let us check them for the trefoil, and then see why they hold for every link.

For the trefoil, $V = t + t^3 - t^4$:

- $V(1) = 1 + 1 - 1 = 1 = (-2)^0$.
- $V(-1) = -1 - 1 - 1 = -3$, and $\Delta^{\mathrm{Sym}}(-1) = -1 - 1 - 1 = -3$.

- With $\omega = e^{2\pi i/3}$, $\omega^3 = 1$ and $\omega^4 = \omega$: $V(\omega) = \omega + 1 - \omega = 1$.
- $V'(t) = 1 + 3t^2 - 4t^3$, so $V'(1) = 1 + 3 - 4 = 0$.

Why $V_L(1) = (-2)^{p-1}$. At $t = 1$ the skein relation (12.179) becomes $V_{L_+}(1) - V_{L_-}(1) = 0$: crossing changes are free. Undo crossings until L becomes the unlink U_p with the same number of components. Then $V_L(1) = V_{U_p}(1) = (-1-1)^{p-1} = (-2)^{p-1}$ by (12.180).

Why $V_K(\omega) = 1$ for knots. Put $t = \omega$, $t^{1/2} = e^{i\pi/3}$. Then $t^{1/2} - t^{-1/2} = 2i\sin\frac{\pi}{3} = i\sqrt{3}$, $t^{-1} - t = -2i\sin\frac{2\pi}{3} = -i\sqrt{3}$, and $V_{U_2}(\omega) = -2\cos\frac{\pi}{3} = -1$. Claim: $V_L(\omega) = (-1)^{p-1}$ for every p-component link. It holds for unlinks. If it holds for L_0 (with $p \pm 1$ components) and for one of $L_\pm$, it holds for the other: if, say, $V_{L_-}(\omega) = (-1)^{p-1}$, then

$$\omega^{-1}V_{L_+}(\omega) \overset{(1)}{=} \omega\,(-1)^{p-1} + i\sqrt{3}\,(-1)^p \overset{(2)}{=} (-1)^{p-1}\big(\omega - i\sqrt{3}\big) \overset{(3)}{=} (-1)^{p-1}\,\omega^{-1},$$

(1) Skein relation, with the hypothesis for L_- and for L_0 (which has $p \pm 1$ components, so its sign is $(-1)^p$).
(2) Factor out $(-1)^{p-1}$.
(3) $\omega - i\sqrt{3} = -\frac{1}{2} + \frac{\sqrt{3}}{2}i - i\sqrt{3} = -\frac{1}{2} - \frac{\sqrt{3}}{2}i = \omega^{-1}$.

so $V_{L_+}(\omega) = (-1)^{p-1}$. By induction it holds for all links, and for knots ($p = 1$) it is Nash's $V_K(e^{2\pi i/3}) = 1$.

Why $V_K'(1) = 0$ for knots. Differentiate (12.179) with respect to t and set $t = 1$:

$$-V_{L_+}(1) + V'_{L_+}(1) - V_{L_-}(1) - V'_{L_-}(1) \overset{(1)}{=} \big(\tfrac{1}{2} + \tfrac{1}{2}\big)V_{L_0}(1)$$
$$\Longrightarrow \quad -1 + V'_{K_+}(1) - 1 - V'_{K_-}(1) \overset{(2)}{=} -2$$
$$\Longrightarrow \quad V'_{K_+}(1) = V'_{K_-}(1).$$

(1) Product rule on the left; on the right the derivative of $t^{1/2} - t^{-1/2}$ at 1 is $\frac{1}{2} + \frac{1}{2}$, and the term with V'_{L_0} is multiplied by $t^{1/2} - t^{-1/2} = 0$ at $t = 1$.
(2) For knots $K_\pm$, $V(1) = 1$; and L_0 then has two components, so $V_{L_0}(1) = -2$.

Crossing changes do not change $V'(1)$; unknotting, $V'_K(1) = V'_{\text{unknot}}(1) = 0$.

Why $V_L(-1) = \Delta_L(-1)$. At $t = -1$ both skein relations become the same recursion. With $t^{1/2} = i$, (12.179) reads $-V_{L_+} + V_{L_-} = 2i\,V_{L_0}$, i.e. $V_{L_+} - V_{L_-} = -2i\,V_{L_0}$, and Nash's form of the Alexander relation (12.80) at $t^{1/2} = i$ reads $\Delta_{L_+} - \Delta_{L_-} = -2i\,\Delta_{L_0}$. The initial values also agree: $V = \Delta = 1$ on the unknot, and on split links $V_{U_p}(-1) = (-i - i^{-1})^{p-1} = 0 = \Delta_{U_p}$. Two functions obeying the same recursion with the same initial values are equal. For knots the choice of square root is irrelevant, as in the trefoil check above.

Where the Jones polynomial came from: a representation of the braid group. Nash mentions that the Jones polynomial "originated in the theory of finite dimensional von Neumann algebras". Here is the part of that story we can make completely explicit, in the spirit of exhibiting representations.

Every link is the closure of a braid (Alexander's theorem), as in Figure 12.51. On two strands, the bracket rule says that a positive crossing is a linear combination of two pictures: the identity $\mathbb{1}$ (two vertical strands) and the cup-cap e:

$$\sigma = A\,\mathbb{1} + A^{-1}e, \qquad \sigma^{-1} = A^{-1}\mathbb{1} + A\,e.$$

Stacking pictures is multiplication. Stacking e on e traps one small circle: $e^2 = d\,e$. This two-dimensional algebra with basis $\{\mathbb{1}, e\}$ is the **Temperley–Lieb algebra** TL_2. Check that σ^{-1} really is the inverse:

$$\begin{aligned}\sigma\sigma^{-1} &\overset{(1)}{=} \mathbb{1} + A^2 e + A^{-2} e + e^2 \\ &\overset{(2)}{=} \mathbb{1} + \bigl(A^2 + A^{-2} + d\bigr)e \overset{(3)}{=} \mathbb{1}.\end{aligned}$$

(1) Multiply out $(A\mathbb{1} + A^{-1}e)(A^{-1}\mathbb{1} + Ae)$.
(2) $e^2 = d\,e$.
(3) $d = -A^2 - A^{-2}$.

This is the R2 computation (12.171) as algebra. Closing a picture up is a linear functional, the **Markov trace**: closing $\mathbb{1}$ gives two circles, weight d; closing e gives one circle, weight 1. So $\mathrm{tr}(\mathbb{1}) = d$, $\mathrm{tr}(e) = 1$, and $\langle\hat{\beta}\rangle = \mathrm{tr}(\beta)$.

Worked example 12.80: All $(2, n)$ torus links from a 2×2 matrix

Left multiplication by σ on TL_2, in the basis $(\mathbb{1}, e)$:

$$\sigma \cdot \mathbb{1} = A\,\mathbb{1} + A^{-1}e, \qquad \sigma \cdot e = Ae + A^{-1}d\,e = -A^{-3}e, \qquad [\sigma] = \begin{pmatrix} A & 0 \\ A^{-1} & -A^{-3} \end{pmatrix}.$$

(We used $A + A^{-1}d = A - A - A^{-3} = -A^{-3}$.) The eigenvalues are A and $-A^{-3}$. The eigenvector for $-A^{-3}$ is e, and the one for A is $\mathbb{1} - e/d$ (check: $\sigma(\mathbb{1} - e/d) = A\mathbb{1} + A^{-1}e + A^{-3}e/d = A\mathbb{1} + \bigl(A^{-1}d + A^{-3}\bigr)e/d = A\mathbb{1} - Ae/d$). Writing $\mathbb{1} = (\mathbb{1} - e/d) + e/d$, the braid σ^n acts on each piece by its eigenvalue, and closing up:

$$\begin{aligned}\bigl\langle T(2,n)\bigr\rangle &\overset{(1)}{=} \mathrm{tr}\bigl(\sigma^n\bigr) \overset{(2)}{=} A^n\,\mathrm{tr}\bigl(\mathbb{1} - e/d\bigr) + (-A^{-3})^n\,\mathrm{tr}(e)/d \\ &\overset{(3)}{=} A^n\Bigl(d - \frac{1}{d}\Bigr) + \frac{(-A^{-3})^n}{d}. \qquad (12.181)\end{aligned}$$

(1) The closure of σ^n is the torus link $T(2, n)$, and closing is the Markov trace.
(2) $\sigma^n\mathbb{1} = A^n(\mathbb{1} - e/d) + (-A^{-3})^n e/d$ by the eigen-decomposition.
(3) $\mathrm{tr}\,\mathbb{1} = d$, $\mathrm{tr}\,e = 1$.

Checks: $n = 0$ gives d, the two-component unlink. $n = 1$ gives $Ad - (A + A^{-3})/d = Ad + A^{-1} = -A^3$, a kinked unknot, as in (12.172). $n = 3$ gives $A^3 d - (A^3 + A^{-9})/d$; since $A^6 + A^{-6} = (A^2 + A^{-2})(A^4 - 1 + A^{-4})$, $(A^3 + A^{-9})/d = -A + A^{-3} - A^{-7}$, and the total is $-A^5 - A - (-A + A^{-3} - A^{-7}) = -A^5 - A^{-3} + A^{-7}$, which is (12.174). The trefoil state sum of eight terms has been replaced by diagonalising a 2×2 matrix.

Physical picture: Transfer matrices

This is the transfer-matrix method of statistical mechanics. The braid direction is "time"; σ is the transfer matrix of one time step; the closure is the trace that imposes periodic boundary conditions; the bracket is $\mathrm{tr}(T^n)$, dominated by eigenvalues. Jones's original construction is the same with TL_n on n strands, realised as operators on a Hilbert space (a von Neumann algebra), and with the trace normalised by $\mathrm{tr}(\mathbb{1}) = 1$. In Chapter 12 these Hilbert spaces reappear as the state spaces of Chern–Simons theory on a punctured sphere, with σ the operator that exchanges two punctures.

12.11.6 HOMFLY, and what polynomials cannot see

Both skein relations have the same shape: $x\,P_{L_+} + y\,P_{L_-} + z\,P_{L_0} = 0$ with some coefficients. Nash mentions that Freyd, Yetter, Hoste, Lickorish, Millett and Ocneanu ("HOMFLY") showed that one can keep the coefficients *free*: there is a unique invariant, homogeneous of degree zero in three variables x, y, z, satisfying this with $P_{\text{unknot}} = 1$. Dividing out the homogeneity leaves two variables. In the common normalisation:

$$a\,P_{L_+} - a^{-1}P_{L_-} = z\,P_{L_0}, \qquad P_{\text{unknot}} = 1. \tag{12.182}$$

Comparing with (12.168) and (12.179):

$$\Delta_L(t) = P_L\big(a = 1,\ z = t^{1/2} - t^{-1/2}\big), \qquad V_L(t) = P_L\big(a = t^{-1},\ z = t^{1/2} - t^{-1/2}\big). \tag{12.183}$$

Worked example 12.81: HOMFLY polynomial of the trefoil

Unlink. From the kink: $aP_{\text{unknot}} - a^{-1}P_{\text{unknot}} = zP_{U_2}$, so $P_{U_2} = (a - a^{-1})/z$.
Hopf link. $L_+ = H_+$, $L_- = U_2$, $L_0 =$ unknot:

$$\begin{aligned} aP_{H_+} &\overset{(1)}{=} a^{-1}P_{U_2} + z \\ \Longrightarrow \quad P_{H_+} &\overset{(2)}{=} a^{-2}\frac{a - a^{-1}}{z} + a^{-1}z = \big(a^{-1} - a^{-3}\big)z^{-1} + a^{-1}z. \end{aligned}$$

(1) Skein relation (12.182).
(2) Divide by a and insert P_{U_2}.

Trefoil. $L_+ = 3_1$, $L_- =$ unknot, $L_0 = H_+$:

$$\begin{aligned} aP_{3_1} &\overset{(1)}{=} a^{-1} + z\,P_{H_+} \overset{(2)}{=} a^{-1} + a^{-1} - a^{-3} + a^{-1}z^2 \\ \Longrightarrow \quad P_{3_1} &= 2a^{-2} - a^{-4} + a^{-2}z^2. \end{aligned}$$

(1) Skein relation.
(2) Insert P_{H_+} and multiply by z.

Now specialise. $a = 1$: $2 - 1 + z^2 = 1 + z^2$, the Conway polynomial of Example 12.73. $a = t^{-1}$, $z^2 = t - 2 + t^{-1}$: $2t^2 - t^4 + t^2(t - 2 + t^{-1}) = t + t^3 - t^4$, the Jones polynomial. One calculation, both of Nash's polynomials.

What polynomials cannot see. Nash ends §4 with a warning: "for the Jones and Homfly polynomials there exist distinct links on which these polynomials have the same values" (Birman; Kanenobu). Examples:

- *Mutants.* Cut a diagram along a circle meeting it in four points, rotate the inside by 180° and reglue. The result (a **mutant**) has the same Alexander, Jones and HOMFLY polynomials, because every skein computation can be done outside the circle first. The Kinoshita–Terasaka and Conway knots (11 crossings) are mutants, and are distinct.
- *Kanenobu's knots.* An infinite family of distinct knots with the same HOMFLY polynomial.
- Whether $V_K(t) = 1$ forces K to be the unknot is still an open question.

This is why a more general machine is desirable, and Chern–Simons theory provides one: each gauge group G and representation R gives an invariant, and the Jones polynomial is only the SU(2), spin-$\tfrac{1}{2}$ member of the family.

	unknot	H_+	3_1 (right)	$\bar{3}_1$ (left)	4_1
components p	1	2	1	1	1
tricolourings τ	3	9	9	9	3
Δ^{Sym}	1	$t^{1/2}-t^{-1/2}$	$t-1+t^{-1}$	$t-1+t^{-1}$	$-t+3-t^{-1}$
$\lvert\Delta(-1)\rvert$	1	2	3	3	5
$V(t)$	1	$-t^{1/2}-t^{5/2}$	$t+t^3-t^4$	$t^{-1}+t^{-3}-t^{-4}$	$t^2-t+1-t^{-1}+t^{-2}$
writhe of diagram used	0	+2	+3	−3	0

Invariants of the smallest knots and links. Only the Jones row separates the two trefoils. (For H_+, $\tau = 9$: the two circles may be coloured independently.)

Reading Nash (pp. 340–342, (12.72)–(12.83))

Nash lists, without derivation: the Alexander polynomial as an integer Laurent polynomial defined up to $\mp t^n$, with $\Delta(1) = \mp 1$ and $\Delta(1/t) = t^n\Delta(t)$, n even (12.73); the trefoil value (12.74) and its other normalisations (12.75), including Δ^{Sym}; the equality of Δ for mirror images; the Jones polynomial and its trefoil values (12.76); the special values (12.77)–(12.78); HOMFLY; the skein triple (12.79); the Alexander skein relation (12.80); the Jones skein relation (12.81)–(12.82) and the unlink formula (12.83). We have derived each of these: the properties of Δ from the Seifert matrix (12.167), the Jones polynomial and its skein relation from the Kauffman bracket (12.179), and the special values by induction on skein relations. Watch Nash's misprint in (12.75) and his sign conventions in (12.80)–(12.81), discussed in the pitfalls above. The Fox calculus, Seifert matrices, the bracket, the Temperley–Lieb matrix and the HOMFLY trefoil computation are not in Nash.

Recap

- The Alexander polynomial $\Delta(t)$ is a Laurent polynomial defined up to $\pm t^n$. Compute it by Fox calculus on the knot group or as $\det(V - tV^T)$ from a Seifert matrix. Trefoil: $t^2 - t + 1$; figure-eight: $t^2 - 3t + 1$.
- From the Seifert matrix: $\Delta(1) = \pm 1$, $\Delta(1/t) = t^{-2g}\Delta(t)$, $\deg\Delta \le 2g$, and Δ is blind to mirror images.
- Skein relations reduce any link to unlinks: $\Delta_+ - \Delta_- = (t^{1/2}-t^{-1/2})\Delta_0$ and $t^{-1}V_+ - tV_- = (t^{1/2}-t^{-1/2})V_0$, with $V_{U_p} = (-t^{1/2}-t^{-1/2})^{p-1}$.
- The Kauffman bracket is a state sum with loop fugacity $d = -A^2 - A^{-2}$, invariant under R2, R3; R1 costs $-A^{\pm 3}$, the framing factor. $V_L(t) = (-A^3)^{-w}\langle D\rangle$ at $A = t^{-1/4}$.
- Right trefoil: $\langle 3_1\rangle = -A^5 - A^{-3} + A^{-7}$, $V = t + t^3 - t^4$; mirror: $V(1/t)$. Jones detects chirality.
- $V(1) = (-2)^{p-1}$, $V(-1) = \Delta(-1)$, $V_K(e^{2\pi i/3}) = 1$, $V_K'(1) = 0$, all by skein induction.
- Temperley–Lieb: $\sigma = A\mathbb{1} + A^{-1}e$, $e^2 = de$; the bracket is a transfer-matrix trace. HOMFLY unifies Δ and V; no polynomial is complete.

12.12 Wilson loops and abelian Chern–Simons theory

We now have a stock of knot invariants: the linking number, the Alexander polynomial and the Jones polynomial. Every one of them was defined from a *picture*, a projection of the knot onto a plane, and every one had to be checked against the three Reidemeister moves. Nash's §5 turns

to Witten's discovery that the Jones polynomial is also a *correlation function* of a quantum field theory in three dimensions, the Chern–Simons theory. This section does the first, fully explicit, part of that story: the abelian theory, where the path integral is Gaussian and we can evaluate it exactly. The answer is the linking number, and along the way we shall meet the one subtlety, framing, that survives in the non-abelian case.

12.12.1 Witten's programme

Nash lists the reasons why a field-theoretic definition is valuable [Nash pp. 342–343]. Let us spell them out.

1. *An intrinsically three-dimensional definition.* A knot lives in $\mathbb{R}^3$ (or S^3), not in a plane. A definition from a diagram depends a priori on the direction of projection, and different knots can even have the same projection if we forget the over/under information. Every diagrammatic invariant therefore needs a proof that it does not depend on the projection: the Reidemeister check of Section 12.9.4. A path integral over gauge fields on the three-manifold, with the knot entering as a Wilson loop, uses no projection at all. If the integral makes sense and is diffeomorphism invariant, then it is automatically an isotopy invariant.

2. *Generalisation.* The same definition works in any closed oriented three-manifold M, not only S^3, and for any compact gauge group G and any choice of representations on the components. The Jones polynomial becomes one member ($G = \mathrm{SU}(2)$, spin $\frac{1}{2}$) of a large family.

3. *The price.* Evaluating the non-abelian path integral is not a Gaussian integral. Witten did it by cutting M into pieces and using the Hilbert spaces of the theory on surfaces, which turn out to be the spaces of conformal blocks of a two-dimensional *conformal field theory*. As Nash remarks, the machinery of conformal field theory is an indispensable part of the construction. That calculation is the business of the next section of this chapter (see Chapter 12).

Key idea: The strategy for proving that $\langle W \rangle$ is the Jones polynomial

By Section 12.11.5, the Jones polynomial is the *unique* function of oriented links that obeys the skein relation (12.179) and takes the value 1 on the unknot (and then $V_{U_p} = (-t^{1/2} - t^{-1/2})^{p-1}$ on unlinks). So to identify a Chern–Simons correlation function with $V_L(t)$ one must show two things: (i) it obeys the skein relation, for a suitable identification of t with the level k; (ii) it takes the right values on unlinked, unknotted circles. Nothing else needs checking.

In the present section we test the idea where everything can be computed: the gauge group U(1). We shall see that the abelian correlation function obeys a skein relation of its own, a degenerate one, and that comparing it with the Jones relation shows exactly what non-abelian structure has to add.

12.12.2 Conventions: two ways of writing a gauge field

As in the rest of these notes, a connection is a Lie-algebra-valued one-form $A = A_\mu \, \mathrm{d}x^\mu$ whose values are *anti-Hermitian* matrices, with curvature $F = \mathrm{d}A + A \wedge A$ and gauge transformation $A \mapsto g^{-1}Ag + g^{-1}\mathrm{d}g$. Physicists usually write a Hermitian field with a coupling g and generators T^a (Hermitian, $[T^a, T^b] = if^{abc}T^c$).

Dictionary: Mathematicians' and physicists' gauge fields

mathematicians (Nash)	physicists
A_μ anti-Hermitian	$A_\mu^{\text{phys}} = A_\mu^a T^a$ Hermitian
A_μ	$-i\,g\,A_\mu^a T^a$
$D_\mu = \partial_\mu + A_\mu$	$D_\mu = \partial_\mu - igA_\mu^a T^a$
$F = \mathrm{d}A + A \wedge A$	$F = -\frac{i}{2}\,g\,F_{\mu\nu}^a T^a\,\mathrm{d}x^\mu \wedge \mathrm{d}x^\nu$
U(1): $A = i\,a$, a real	a_μ = the electromagnetic potential (with $e = 1$)
$\operatorname{tr} P \exp \oint_C A$	$\operatorname{tr} P \exp\bigl(-ig \oint_C A_\mu^a T^a \mathrm{d}x^\mu\bigr)$

For U(1) the Lie algebra is $i\mathbb{R}$, so an anti-Hermitian A is i times a real one-form a. Nash, in the abelian formulas (12.87)–(12.88), writes the *real* field and calls it A again; the Wilson line appears as $\exp(in \oint A)$ with an explicit i. We follow him: in Sections 12.12.6 to 12.12.8 A denotes a real one-form, the ordinary electromagnetic vector potential.

12.12.3 Wilson loops: the holonomy made into a number

Parallel transport as a product of small steps. Take a closed curve C, parametrised by $x(s)$, $0 \le s \le 1$, with $x(0) = x(1) = x_0$. Chop it into N short steps $x_j = x(j/N)$ with displacements $\Delta x_j = x_{j+1} - x_j$. Along each step, attach the matrix

$$V_j = \exp\bigl(A_\mu(x_j)\,\Delta x_j^\mu\bigr) \approx \mathbb{1} + A_\mu(x_j)\,\Delta x_j^\mu, \tag{12.184}$$

which lies in the gauge group G because $A_\mu(x_j)$ lies in its Lie algebra. (In lattice gauge theory V_j is the "link variable".) The **holonomy** of A around C is the ordered product

$$\mathrm{Hol}_C(A) = P \exp \oint_C A \coloneqq \lim_{N\to\infty} V_0 V_1 \cdots V_{N-1}. \tag{12.185}$$

The symbol P (path ordering) records that the factors are *not* combined into $\exp(\sum_j A(x_j)\Delta x_j)$: the matrices $A(x_j)$ need not commute, and the order in which they are multiplied is the order of the points along the curve.

Working definition 12.32 (Wilson loop, Nash (12.84))

For a closed curve C, a connection A and an irreducible representation R of G [a rule turning each $g \in G$ into an $n \times n$ matrix, with no proper subspace of $\mathbb{C}^n$ left invariant, such as spin j for SU(2)], the **Wilson loop** is

$$W(R, C) = \operatorname{tr}_R\, P \exp \oint_C A, \tag{12.186}$$

the trace, in the representation R, of the holonomy. *To compute it:* solve the parallel-transport equation (or multiply the link variables) around the loop to get a matrix in G, represent the matrix in R, and take the trace.

Why the trace: gauge invariance. Under $A \mapsto A^g = g^{-1}Ag + g^{-1}\mathrm{d}g$ each link variable transforms as

$$\begin{aligned} g(x_j)^{-1}\,V_j\,g(x_{j+1}) &\overset{(1)}{=} g^{-1}\bigl(\mathbb{1} + A_\mu \Delta x_j^\mu\bigr)\bigl(g + \partial_\mu g\,\Delta x_j^\mu\bigr) + O(\Delta x^2) \\ &\overset{(2)}{=} \mathbb{1} + \bigl(g^{-1}A_\mu g + g^{-1}\partial_\mu g\bigr)\Delta x_j^\mu + O(\Delta x^2) \end{aligned}$$

$$\overset{(3)}{=} V_j^{\,g} + O(\Delta x^2). \tag{12.187}$$

(1) Expand V_j to first order and Taylor expand $g(x_{j+1}) = g(x_j) + \partial_\mu g\, \Delta x_j^\mu$; all functions at x_j.
(2) Multiply out and keep first-order terms; $g^{-1}g = \mathbb{1}$.
(3) The bracket is A_μ^g, so this is the link variable of the transformed connection.

So $V_j^g = g_j^{-1} V_j g_{j+1}$ with $g_j = g(x_j)$. In the ordered product the inner factors cancel in pairs,

$$V_0^g V_1^g \cdots V_{N-1}^g = g_0^{-1} V_0\, g_1 g_1^{-1}\, V_1\, g_2 \cdots g_{N-1}^{-1} V_{N-1} g_N = g_0^{-1} (V_0 \cdots V_{N-1}) g_0,$$

using $g_N = g(x_0) = g_0$ on a closed loop. The holonomy is conjugated by the gauge transformation at the base point, and its trace is invariant: $W(R, C)$ is a gauge-invariant function of A. It also does not depend on the base point, since moving x_0 along C cyclically permutes the factors.

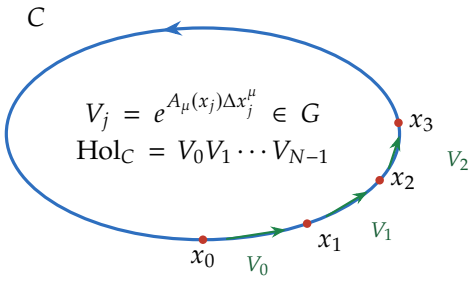


Figure 12.55: The holonomy as an ordered product. The loop C is cut into short steps; the step from x_j to x_{j+1} carries the group element $V_j \approx \mathbb{1} + A_\mu(x_j)\Delta x_j^\mu$ (the link variable of lattice gauge theory). The holonomy is the product in the order of the steps, starting at the base point x_0; a gauge transformation replaces it by $g(x_0)^{-1}\,\mathrm{Hol}_C\; g(x_0)$, so its trace, the Wilson loop, is gauge invariant.

Watch out: Ordering conventions

With Nash's sign, $\exp(+\oint A)$, the earlier points along the curve stand to the *left*, as in (12.185); this is what made (12.187) work. Many physics texts write $P\exp(-\oint A)$ with later points to the left, like the time-ordered exponential $T\exp(-i\int H\,\mathrm{d}t)$ of the Dirac picture. The two holonomies are inverse to each other, and for a unitary matrix $\mathrm{tr}\,U^{-1} = \overline{\mathrm{tr}\,U}$. For SU(2), whose representations are all equivalent to their complex conjugates, the traces agree.

Worked example 12.82: U(1): the Aharonov–Bohm phase

For $G = \mathrm{U}(1)$, write $A = i\,a$ with a real. All the $V_j = e^{ia(x_j)\cdot\Delta x_j}$ commute, so path ordering is empty:

$$W(C) \overset{(1)}{=} \exp\Big(i \oint_C a_\mu\, \mathrm{d}x^\mu\Big) \overset{(2)}{=} \exp\Big(i \int_S (\nabla \times \vec{a}) \cdot \hat{n}\, \mathrm{d}S\Big) \overset{(3)}{=} e^{i\Phi_S},$$

(1) Commuting factors: the product of exponentials is the exponential of the sum, which becomes the line integral.
(2) Stokes' theorem, for any surface S bounded by C.
(3) $\nabla \times \vec{a} = \vec{b}$ is the magnetic field, and its flux through S is Φ_S.

This is the phase an electron of unit charge picks up when carried around C: the Aharonov–Bohm effect. It is non-trivial even where $\vec{b} = 0$ on C itself. A Wilson loop measures the flux

of the gauge field that it encloses.

Worked example 12.83: SU(2): path ordering matters

Take $G = \mathrm{SU}(2)$, the fundamental representation, and the constant connection $A = i\alpha\,\sigma_1\,\mathrm{d}x + i\beta\,\sigma_2\,\mathrm{d}y$ in the plane (σ_i the Pauli matrices). Let C be the boundary of the unit square, traversed $(0,0)\to(1,0)\to(1,1)\to(0,1)\to(0,0)$. On each edge A is constant, so the link variables along the four edges are $e^{i\alpha\sigma_1}$, $e^{i\beta\sigma_2}$, $e^{-i\alpha\sigma_1}$, $e^{-i\beta\sigma_2}$ (the last two edges run in the $-x$ and $-y$ directions). If the matrices commuted the product would be $\mathbb{1}$ and $W = 2$. They do not. First conjugate σ_2 by $e^{i\alpha\sigma_1} = \cos\alpha + i\sigma_1\sin\alpha$ (write $c = \cos\alpha$, $s = \sin\alpha$):

$$\begin{aligned} e^{i\alpha\sigma_1}\sigma_2 e^{-i\alpha\sigma_1} &\overset{(1)}{=} c^2\sigma_2 + isc\,(\sigma_1\sigma_2 - \sigma_2\sigma_1) + s^2\sigma_1\sigma_2\sigma_1 \\ &\overset{(2)}{=} c^2\sigma_2 + isc\,(2i\sigma_3) - s^2\sigma_2 \\ &\overset{(3)}{=} \cos 2\alpha\;\sigma_2 - \sin 2\alpha\;\sigma_3 \equiv m. \end{aligned}$$

(1) Multiply out $(c + is\sigma_1)\sigma_2(c - is\sigma_1)$.
(2) $[\sigma_1,\sigma_2] = 2i\sigma_3$ and $\sigma_1\sigma_2\sigma_1 = -\sigma_2$ (anticommute, then $\sigma_1^2 = \mathbb{1}$).
(3) Double-angle formulas.

m is again a unit Pauli vector, so $m^2 = \mathbb{1}$ and $e^{i\beta m} = \cos\beta + i\sin\beta\, m$. Now

$$\begin{aligned} W &\overset{(1)}{=} \mathrm{tr}\big(e^{i\alpha\sigma_1}e^{i\beta\sigma_2}e^{-i\alpha\sigma_1}e^{-i\beta\sigma_2}\big) \overset{(2)}{=} \mathrm{tr}\big(e^{i\beta m}\,e^{-i\beta\sigma_2}\big) \\ &\overset{(3)}{=} \mathrm{tr}\big[(\cos\beta + i\sin\beta\, m)(\cos\beta - i\sin\beta\,\sigma_2)\big] \\ &\overset{(4)}{=} 2\cos^2\beta + \sin^2\beta\;\mathrm{tr}(m\sigma_2) \overset{(5)}{=} 2\cos^2\beta + 2\sin^2\beta\cos 2\alpha. \end{aligned}$$

(1) The ordered product of the four link variables.
(2) $e^{i\alpha\sigma_1}e^{i\beta\sigma_2}e^{-i\alpha\sigma_1} = \exp(i\beta\, e^{i\alpha\sigma_1}\sigma_2 e^{-i\alpha\sigma_1})$.
(3) Expand both exponentials.
(4) $\mathrm{tr}\,\mathbb{1} = 2$; the terms linear in Pauli matrices are traceless.
(5) $\mathrm{tr}(m\sigma_2) = \cos 2\alpha\;\mathrm{tr}\,\sigma_2^2 = 2\cos 2\alpha$.

For small α,β: $W \approx 2 - 4\alpha^2\beta^2$. Compare the curvature: $F = \mathrm{d}A + A\wedge A = [A_x, A_y]\,\mathrm{d}x\wedge\mathrm{d}y = (i\alpha)(i\beta)[\sigma_1,\sigma_2]\,\mathrm{d}x\wedge\mathrm{d}y = -2i\alpha\beta\,\sigma_3\,\mathrm{d}x\wedge\mathrm{d}y$. A small loop has holonomy $\approx \exp(F_{xy}\times\text{area})$, whose trace is $2\cos(2\alpha\beta) \approx 2 - 4\alpha^2\beta^2$. It agrees: the deviation of a Wilson loop from $\dim R$ measures the non-abelian flux inside it, and that flux here comes *entirely* from the commutator term $A\wedge A$, since $\mathrm{d}A = 0$.

Wilson loops in other representations. Every element of SU(2) is conjugate to $\mathrm{diag}(e^{i\theta}, e^{-i\theta})$. In the spin-$j$ representation (dimension $2j+1$) this element acts on the basis $|j,m\rangle$ by $e^{2im\theta}$, so

$$W(j,C) = \sum_{m=-j}^{j} e^{2im\theta} = \frac{\sin\big((2j+1)\theta\big)}{\sin\theta}, \tag{12.188}$$

the **character** of spin j (geometric series). For $j = \frac{1}{2}$: $2\cos\theta$. For $j = 1$: $1 + 2\cos 2\theta$. A Wilson loop in representation R is the character χ_R evaluated on the holonomy.

Trace normalisations. Nash warns, before (12.87), that different representations have "traces" whose normalisations differ by rational factors. Concretely: for generators T^a of a simple Lie

algebra, $\mathrm{tr}_R(T^aT^b) = x_R\,\delta^{ab}$ with a number x_R (the Dynkin index) that depends on R. For SU(2) with $T^a = J^a$ in spin j,

$$\mathrm{tr}_j\big(J_3^2\big) = \sum_{m=-j}^{j} m^2 = \frac{j(j+1)(2j+1)}{3}: \qquad x_{1/2} = \tfrac{1}{2}, \quad x_1 = 2, \quad x_{3/2} = 5.$$

The ratios are rational. The Chern–Simons action is defined with the trace in one fixed representation (for SU(N), the fundamental), and this normalisation fixes what "level k" means; the Wilson loops can then carry any R. For U(1) with charge n the analogue of x_R is n^2, and that n^2 is exactly what will appear in the exponent of our answer.

12.12.4 The correlation functions and why they are topological

Let L be a link with components $C_1, \dots, C_p$, and assign an irreducible representation R_l to each C_l. The object of interest is the normalised correlation function, Nash's (12.85),

$$\big\langle W(R_1, C_1)\cdots W(R_p, C_p)\big\rangle = \frac{1}{Z(M)}\int \mathcal{D}A\; W(R_1, C_1)\cdots W(R_p, C_p) \\ \times \exp\Big[-\frac{ik}{4\pi}\int_M \mathrm{tr}\big(A\wedge \mathrm{d}A + \tfrac{2}{3}A\wedge A\wedge A\big)\Big], \tag{12.189}$$

with $Z(M)$ the same integral without the Wilson loops. The integral is over all connections on M modulo gauge transformations.

Key idea: No metric anywhere

Look at what enters (12.189).

- The Chern–Simons three-form $\mathrm{tr}(A\,\mathrm{d}A + \frac{2}{3}A^3)$ is built from wedge products and d alone. It is integrated over the oriented M without a volume form. There is no metric and no Hodge star.
- The Wilson loops are integrals of one-forms along curves: again no metric.

Compare Yang–Mills theory, $\int \mathrm{tr}\,F\wedge\star F$, where the Hodge star needs a metric; there a Wilson loop has an "area law" $\langle W\rangle \sim e^{-\sigma\cdot\mathrm{Area}}$, a geometric answer. In Chern–Simons theory the correlation function can only depend on the topology of the pair (M, L). A diffeomorphism of M carrying L to L' can be used to change variables in the integral, so $\langle W(L)\rangle = \langle W(L')\rangle$: the correlator is an isotopy invariant, with no Reidemeister moves to check.

Two caveats make this a "chance", in Nash's word, rather than a theorem. First, to define $\mathcal{D}A$ modulo gauge transformations one must fix a gauge, and gauge fixing does use a metric. The metric dependence must cancel; in the abelian case we shall watch it cancel. Second, the self-linking of a component will turn out to need one extra piece of topological data, a framing (Section 12.12.8). Also, of course, a topological invariant may be trivial; the computation below shows that it is not.

12.12.5 Gaussian integrals with sources: the finite-dimensional rehearsal

The abelian theory is a Gaussian integral with a linear source. Here is everything we need, in finite dimensions first.

One variable. For $\lambda > 0$ the Gaussian integral and its oscillatory cousin, the Fresnel integral, are

$$\int_{-\infty}^{\infty} e^{-\lambda x^2/2}\mathrm{d}x = \sqrt{\frac{2\pi}{\lambda}}, \qquad \int_{-\infty}^{\infty} e^{\mp i\lambda x^2/2}\mathrm{d}x = \sqrt{\frac{2\pi}{\lambda}}\, e^{\mp i\pi/4}. \tag{12.190}$$

(The second follows from the first by rotating the contour through $\mp 45°$, i.e. by setting $\lambda \to \pm i\lambda$ and taking the principal root.) The phase $e^{\mp i\pi/4}$ depends only on the *sign* of the quadratic form. Now add a source J and complete the square:

$$\frac{\int e^{-i\lambda x^2/2 + iJx}\,\mathrm{d}x}{\int e^{-i\lambda x^2/2}\,\mathrm{d}x} \overset{(1)}{=} \frac{\int e^{-i\lambda (x - J/\lambda)^2/2}\,\mathrm{d}x}{\int e^{-i\lambda x^2/2}\,\mathrm{d}x}\, e^{iJ^2/2\lambda} \overset{(2)}{=} e^{iJ^2/2\lambda}. \tag{12.191}$$

(1) $-\frac{\lambda}{2}x^2 + Jx = -\frac{\lambda}{2}\left(x - \frac{J}{\lambda}\right)^2 + \frac{J^2}{2\lambda}$.
(2) Shift the integration variable $x \to x + J/\lambda$; the two integrals cancel.

Two lessons. The normalised answer is $e^{i\Phi_{\mathrm{cl}}}$, the phase evaluated at the stationary point $x_{\mathrm{cl}} = J/\lambda$; and all the awkward prefactors (the $\sqrt{2\pi/\lambda}$ and the Fresnel phase) cancel against the normalisation.

Many variables. For a real symmetric invertible $n \times n$ matrix K and a source vector J, the same completion of the square with $x_{\mathrm{cl}} = K^{-1}J$ gives

$$\left\langle e^{iJ^T x} \right\rangle := \frac{\int \mathrm{d}^n x\; e^{-\frac{i}{2}x^T K x + iJ^T x}}{\int \mathrm{d}^n x\; e^{-\frac{i}{2}x^T K x}} = \exp\!\left(\frac{i}{2}\, J^T K^{-1} J\right). \tag{12.192}$$

Differentiating twice with respect to J at $J = 0$ gives the **propagator** $\langle x_i x_j \rangle = -i(K^{-1})_{ij}$, and (12.192) can be rewritten as $\langle e^X \rangle = e^{\langle X^2 \rangle/2}$ for the linear function $X = iJ^T x$: for a Gaussian distribution the exponential of a linear function is determined by the two-point function. *The whole abelian Chern–Simons calculation is this formula with K^{-1} replaced by a Green's function.*

Zero modes and sources that do not see them. Gauge symmetry makes K degenerate: the action does not change along gauge directions. Here is the finite-dimensional model. Let

$$K = \begin{pmatrix} \lambda & 0 \\ 0 & 0 \end{pmatrix}, \qquad J = \begin{pmatrix} j \\ 0 \end{pmatrix}.$$

The direction x_2 is a "gauge direction": neither the action nor the source depends on it. Then

$$\frac{\int \mathrm{d}x_1 \mathrm{d}x_2\; e^{-i\lambda x_1^2/2 + ijx_1}}{\int \mathrm{d}x_1 \mathrm{d}x_2\; e^{-i\lambda x_1^2/2}} \overset{(1)}{=} \frac{\left(\int \mathrm{d}x_2\right) \int \mathrm{d}x_1\, e^{-i\lambda x_1^2/2 + ijx_1}}{\left(\int \mathrm{d}x_2\right) \int \mathrm{d}x_1\, e^{-i\lambda x_1^2/2}} \overset{(2)}{=} e^{ij^2/2\lambda}.$$

(1) The integrand does not depend on x_2, so the integral factorises; $\int \mathrm{d}x_2$ is infinite, the "volume of the gauge group".
(2) The infinite volumes cancel in the ratio, and the x_1 integral is (12.191).

So the rule is: *restrict to a slice transverse to the gauge directions and invert K there* ("gauge fixing"). This is legitimate because the source has no component along the zero mode. If it had one, the integral would be ill-defined, which is the finite-dimensional shadow of the fact that a gauge-*non*-invariant observable has no expectation value. The choice of slice does not matter: any other slice $x_2 = f(x_1)$ gives the same ratio. In field theory the slice is a gauge condition such as $\nabla \cdot A = 0$, and the Jacobian that appears for a curved slice is the Faddeev–Popov determinant; for the abelian theory it does not depend on A and cancels between numerator and denominator. We shall need it in earnest in Section 12.13.

12.12.6 Abelian Chern–Simons theory: action and Wilson loops

The action, Nash (12.86) and (12.88). For $G = \mathrm{U}(1)$ the cubic term vanishes ($A \wedge A = 0$ for a one-form with values in a commutative algebra), so the action is quadratic. With the real field $A = A_i \,\mathrm{d}x^i$ on $\mathbb{R}^3$ we have

$$\begin{aligned} A \wedge \mathrm{d}A &\overset{(1)}{=} A_i \,\mathrm{d}x^i \wedge \partial_j A_k \,\mathrm{d}x^j \wedge \mathrm{d}x^k \\ &\overset{(2)}{=} \varepsilon^{ijk} A_i \,\partial_j A_k \,\mathrm{d}^3x \overset{(3)}{=} \vec{A} \cdot (\nabla \times \vec{A}) \,\mathrm{d}^3x. \end{aligned} \tag{12.193}$$

(1) $\mathrm{d}A = \partial_j A_k \,\mathrm{d}x^j \wedge \mathrm{d}x^k$.
(2) $\mathrm{d}x^i \wedge \mathrm{d}x^j \wedge \mathrm{d}x^k = \varepsilon^{ijk} \,\mathrm{d}x^1 \wedge \mathrm{d}x^2 \wedge \mathrm{d}x^3$ for the standard orientation.
(3) $(\nabla \times \vec{A})^i = \varepsilon^{ijk} \partial_j A_k$.

So the abelian Chern–Simons action is the **magnetic helicity** $\int \vec{A} \cdot \vec{B} \,\mathrm{d}^3x$ of the field $\vec{B} = \nabla \times \vec{A}$, which we met as a linking number in Example 12.70.

Gauge invariance. Under $A \mapsto A + \mathrm{d}\lambda$,

$$\begin{aligned} \int_M (A + \mathrm{d}\lambda) \wedge \mathrm{d}(A + \mathrm{d}\lambda) &\overset{(1)}{=} \int_M A \wedge \mathrm{d}A + \int_M \mathrm{d}\lambda \wedge \mathrm{d}A \\ &\overset{(2)}{=} \int_M A \wedge \mathrm{d}A + \int_M \mathrm{d}(\lambda \,\mathrm{d}A) \overset{(3)}{=} \int_M A \wedge \mathrm{d}A, \end{aligned}$$

(1) $\mathrm{d}\mathrm{d}\lambda = 0$.
(2) $\mathrm{d}(\lambda \,\mathrm{d}A) = \mathrm{d}\lambda \wedge \mathrm{d}A + \lambda \,\mathrm{d}\mathrm{d}A$ and $\mathrm{d}\mathrm{d}A = 0$.
(3) Stokes' theorem on the closed manifold M ($\partial M = \varnothing$).

The representations and the Wilson loops, Nash (12.87). The irreducible representations of $\mathrm{U}(1) = \{e^{i\theta}\}$ are one-dimensional, $e^{i\theta} \mapsto e^{in\theta}$ with $n \in \mathbb{Z}$ (the charge). The integer condition is needed for the map to be well defined on the circle: θ and $\theta + 2\pi$ are the same group element. Path ordering is empty and the trace of a 1×1 matrix is its entry, so

$$W(n, C) = \exp\Bigl(in \oint_C A \Bigr), \qquad n \in \mathbb{Z}. \tag{12.194}$$

The same integrality shows up as gauge invariance. A gauge transformation $g = e^{i\lambda}$ need only be single valued as a group element, so going once round C the function λ may jump by $2\pi w$ with $w \in \mathbb{Z}$ (for example $\lambda = \varphi$, the azimuthal angle round a line). Then $\oint_C \mathrm{d}\lambda = 2\pi w$ and $W(n, C)$ is multiplied by $e^{2\pi i n w} = 1$ precisely because $n \in \mathbb{Z}$.

Wilson loops are currents. The line integral along C can be written as a volume integral with a delta-function current:

$$\oint_C A = \int \mathrm{d}^3x \; J_C^i(x) A_i(x), \qquad J_C^i(x) := \oint_C \mathrm{d}y^i \,\delta^3(x - y). \tag{12.195}$$

J_C is the current density of a unit current flowing along the wire C. It is conserved: for any test function f,

$$\int \mathrm{d}^3x \; f \,\partial_i J_C^i \overset{(1)}{=} -\int \mathrm{d}^3x \; J_C^i \,\partial_i f \overset{(2)}{=} -\oint_C \mathrm{d}y^i \,\partial_i f(y) \overset{(3)}{=} 0.$$

(1) Integrate by parts.
(2) Insert (12.195) and do the x integral with the delta function.
(3) The integral of a gradient round a closed curve vanishes.

So $\nabla \cdot \vec{J}_C = 0$: a closed loop of current has no sources or sinks. This is the field-theory form of gauge invariance of the Wilson loop, and it is exactly the condition, from the finite-dimensional rehearsal, that the source not couple to the zero modes $A = \mathrm{d}\lambda$.

The integral to be done. For a link with charges $n_1, \dots, n_p$ put $\vec{J} = \sum_l n_l \vec{J}_{C_l}$. The abelian correlation function (12.88) is

$$\big\langle W(n_1, C_1) \cdots W(n_p, C_p) \big\rangle = \frac{1}{Z} \int \mathcal{D}A \; \exp\Big[-\frac{ik}{4\pi} \int \vec{A} \cdot (\nabla \times \vec{A}) \, \mathrm{d}^3x + i \int \vec{J} \cdot \vec{A} \, \mathrm{d}^3x \Big]. \tag{12.196}$$

It has exactly the form (12.192), with the index i of x_i replaced by the pair (vector index, point of space), and with

$$K = \frac{k}{2\pi} \nabla \times \qquad \Big(\text{so that } \tfrac{1}{2} A \cdot KA = \tfrac{k}{4\pi} A \cdot \nabla \times A \Big). \tag{12.197}$$

Following Nash we take $M = S^3$, which for a link is the same as $\mathbb{R}^3$: S^3 is $\mathbb{R}^3$ plus one point at infinity, a link can always be isotoped to miss that point, and fields that fall off at infinity see no difference. On S^3 there are also no non-trivial flat U(1) connections ($H^1(S^3) = 0$), so the only zero modes of K are the gauge directions.

12.12.7 Doing the integral: the propagator of Chern–Simons theory

By (12.192) the answer is $\exp\big(\frac{i}{2} \iint J^i(x) \, G_{ij}(x, y) \, J^j(y)\big)$, where $G = K^{-1}$ on the gauge-fixed slice. We find G in Fourier space, where K becomes a 3×3 matrix we can simply look at.

The curl as a matrix. With $f(x) = \int \frac{\mathrm{d}^3p}{(2\pi)^3} e^{ip\cdot x} \tilde{f}(p)$ the derivative ∂_j becomes multiplication by ip_j, so $(\nabla \times \vec{A})_i = \varepsilon_{ijk} \partial_j A_k$ becomes $P_{ik}(p) \tilde{A}_k$ with

$$P_{ik}(p) = i \, \varepsilon_{ijk} \, p_j, \qquad \tilde{K}(p) = \frac{k}{2\pi} P(p). \tag{12.198}$$

Take $\vec{p} = p \, \hat{z}$. Then only $j = 3$ contributes: $P_{12} = i\varepsilon_{132} p = -ip$, $P_{21} = i\varepsilon_{231} p = ip$, and every other entry vanishes:

$$P(p\hat{z}) = p \begin{pmatrix} 0 & -i & 0 \\ i & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = p \, \big(\sigma_2 \oplus 0 \big). \tag{12.199}$$

This is a Pauli matrix. Its eigenvectors and eigenvalues:

eigenvector	eigenvalue of P	meaning
$\hat{z}$ (along $\vec{p}$)	0	longitudinal, $\tilde{A} \propto \vec{p}$, i.e. $A = \nabla\lambda$: pure gauge
$(1, i, 0)/\sqrt{2}$	$+p$	circular polarisation, helicity $+1$
$(1, -i, 0)/\sqrt{2}$	$-p$	circular polarisation, helicity -1

(Check: $\sigma_2 (1, i)^T = (-i \cdot i, \; i \cdot 1)^T = (1, i)^T$.) Three facts can be read off. The zero mode is exactly the gauge direction, as promised. The operator is first order, so its eigenvalues grow like p

rather than p^2. And the eigenvalues come with *both* signs, so the Gaussian is not damped in any direction: the Chern–Simons "kinetic term" is a Fresnel integral whose phases $e^{\mp i\pi/4}$, one per mode, depend on the spectral asymmetry of the curl. That phase is the origin of the η-invariant in Section 12.13. For the normalised correlator it cancels.

Inverting on the transverse modes. The square of P is the projector onto the transverse plane, times p^2:

$$\begin{aligned}(P^2)_{il} &\overset{(1)}{=} i^2\,\varepsilon_{ijm}p_j\,\varepsilon_{mnl}p_n \overset{(2)}{=} -\big(\delta_{in}\delta_{jl}-\delta_{il}\delta_{jn}\big)p_jp_n\\ &\overset{(3)}{=} p^2\delta_{il}-p_ip_l = p^2\,(\Pi_T)_{il}.\end{aligned} \tag{12.200}$$

(1) Matrix product of (12.198) with itself.
(2) $\varepsilon_{ijm}\varepsilon_{mnl}=\varepsilon_{mij}\varepsilon_{mnl}=\delta_{in}\delta_{jl}-\delta_{il}\delta_{jn}$.
(3) Contract; $\Pi_T=\mathbb{1}-\hat p\hat p^{\,T}$ projects out the longitudinal direction.

(In (12.199): $\sigma_2^2=\mathbb{1}$ on the first two components, 0 on the third.) Hence P/p^2 inverts P on transverse vectors, and the propagator in momentum space, in the gauge $\nabla\cdot\vec A=0$ (i.e. $\vec p\cdot\tilde A=0$), is

$$\tilde G_{il}(p)=\frac{2\pi}{k}\,\frac{P_{il}(p)}{p^2}=\frac{2\pi}{k}\,\frac{i\,\varepsilon_{ijl}\,p_j}{p^2},\qquad \tilde K\tilde G=\Pi_T. \tag{12.201}$$

Back to position space. We need one Fourier transform, that of $1/p^2$. Since $-\nabla^2\frac{1}{4\pi r}=\delta^3(x)$ and $-\nabla^2\to p^2$,

$$\int\frac{\mathrm{d}^3p}{(2\pi)^3}\,\frac{e^{ip\cdot r}}{p^2}=\frac{1}{4\pi\,|r|}. \tag{12.202}$$

Then

$$\begin{aligned}G_{il}(x,y) &\overset{(1)}{=} \frac{2\pi}{k}\int\frac{\mathrm{d}^3p}{(2\pi)^3}\,e^{ip\cdot(x-y)}\,\frac{i\,\varepsilon_{ijl}\,p_j}{p^2}\\ &\overset{(2)}{=} \frac{2\pi}{k}\,\varepsilon_{ijl}\,\frac{\partial}{\partial x^j}\,\frac{1}{4\pi\,|x-y|}\\ &\overset{(3)}{=} -\frac{1}{2k}\,\varepsilon_{ijl}\,\frac{(x-y)_j}{|x-y|^3}\overset{(4)}{=}\frac{1}{2k}\,\varepsilon_{ilj}\,\frac{(x-y)^j}{|x-y|^3}.\end{aligned} \tag{12.203}$$

(1) Inverse Fourier transform of (12.201).
(2) $ip_j\,e^{ip\cdot(x-y)}=\partial_{x^j}e^{ip\cdot(x-y)}$; pull the derivative out and use (12.202).
(3) $\partial_j\,|r|^{-1}=-r_j/|r|^3$, and $\frac{2\pi}{4\pi k}=\frac{1}{2k}$.
(4) $\varepsilon_{ijl}=-\varepsilon_{ilj}$; indices are Euclidean, so up and down are the same.

This is the Chern–Simons propagator, $\langle A_i(x)A_l(y)\rangle=-iG_{il}(x,y)$. It is the Biot–Savart kernel.

The same thing, the physicist's way: Ampère's law. There is a quicker route which shows what is going on. The stationary point of the exponent in (12.196) satisfies

$$\begin{aligned}0 &\overset{(1)}{=} \frac{\delta}{\delta A_i(x)}\Big[-\frac{k}{4\pi}\int\vec A\cdot(\nabla\times\vec A)+\int\vec J\cdot\vec A\Big]\\ &\overset{(2)}{=} -\frac{k}{2\pi}\,(\nabla\times\vec A)_i(x)+J_i(x)\quad\Longrightarrow\quad \nabla\times\vec A_{\mathrm{cl}}=\frac{2\pi}{k}\,\vec J.\end{aligned} \tag{12.204}$$

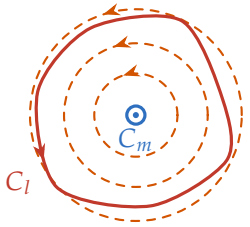
(1) Stationary phase: the classical solution dominates a Gaussian exactly, as in (12.191).

(2) $\int \delta\vec{A}\cdot\nabla\times\vec{A} = \int \vec{A}\cdot\nabla\times\delta\vec{A}$ (integrate by parts; the curl is symmetric), so the variation of the quadratic term is twice $\int \delta\vec{A}\cdot\nabla\times\vec{A}$.

This is Ampère's law $\nabla\times\vec{B} = \vec{J}$ for the field $\vec{B} = (k/2\pi)\vec{A}_{\mathrm{cl}}$. In the gauge $\nabla\cdot\vec{A} = 0$ its solution is the Biot–Savart field (12.156):

$$\vec{A}_{\mathrm{cl}} = \frac{2\pi}{k}\sum_m n_m\,\vec{B}_{C_m}, \qquad \vec{B}_{C_m}(x) = \frac{1}{4\pi}\oint_{C_m}\frac{\mathrm{d}\vec{y}\times(x-y)}{|x-y|^3}. \tag{12.205}$$

In words: *the Chern–Simons field produced by a Wilson loop is the magnetic field of a wire carrying the current n along the loop*, scaled by $2\pi/k$. Equivalently, the Chern–Simons equation of motion $F = \mathrm{d}A = (2\pi/k)\star J$ attaches a flux $2\pi n/k$ to each charge n. This flux attachment is what makes Chern–Simons theory the effective theory of the fractional quantum Hall effect.



field lines of $\vec{A}_{\mathrm{cl}} = \frac{2\pi}{k}\,n_m\vec{B}_{C_m}$

$$\oint_{C_l}\vec{A}_{\mathrm{cl}}\cdot\mathrm{d}\vec{x} = \frac{2\pi}{k}\,n_m\,L(C_l,C_m)$$

Figure 12.56: Flux attachment. Seen end on, a Wilson loop C_m of charge n_m acts as a wire carrying current n_m; the stationary Chern–Simons field $\vec{A}_{\mathrm{cl}}$ is its Biot–Savart field scaled by $2\pi/k$. A second Wilson loop C_l measures its circulation, which by Ampère's law is $2\pi n_m/k$ times the number of times C_l winds around C_m.

The result, Nash (12.89)–(12.92). Evaluate the exponent at the stationary point:

$$\begin{aligned}\Phi_{\mathrm{cl}} &\overset{(1)}{=} -\frac{k}{4\pi}\int\vec{A}_{\mathrm{cl}}\cdot\frac{2\pi}{k}\vec{J} + \int\vec{J}\cdot\vec{A}_{\mathrm{cl}} \overset{(2)}{=} \frac{1}{2}\int\vec{J}\cdot\vec{A}_{\mathrm{cl}}\,\mathrm{d}^3x\\ &\overset{(3)}{=} \frac{\pi}{k}\sum_{l,m} n_l n_m\oint_{C_l}\vec{B}_{C_m}\cdot\mathrm{d}\vec{x} \overset{(4)}{=} \frac{\pi}{k}\sum_{l,m} n_l n_m\,L(C_l,C_m).\end{aligned} \tag{12.206}$$

(1) Use the equation of motion (12.204) to replace $\nabla\times\vec{A}_{\mathrm{cl}}$.
(2) $-\frac{1}{2}+1 = \frac{1}{2}$.
(3) Insert $\vec{J} = \sum_l n_l\vec{J}_{C_l}$ and (12.205), and do the volume integral with the delta functions (12.195).
(4) Ampère's law in the form (12.158): the circulation round C_l of the field of a unit current on C_m is $L(C_l,C_m)$.

Writing out step 3 with the Biot–Savart integral instead of step 4 gives

$$\begin{aligned}\big\langle W(n_1,C_1)\cdots W(n_p,C_p)\big\rangle &\overset{(1)}{=} e^{i\Phi_{\mathrm{cl}}} \overset{(2)}{=} \exp\Big[\frac{i}{4k}\,\varepsilon_{ijk}\sum_{l,m=1}^{p} n_l n_m\oint_{C_l}\mathrm{d}x^i\oint_{C_m}\mathrm{d}y^j\,\frac{(x-y)^k}{|x-y|^3}\Big]\\ &\overset{(3)}{=} \exp\Big[\frac{i\pi}{k}\sum_{l,m} n_l n_m\,L(C_l,C_m)\Big].\end{aligned} \tag{12.207}$$

(1) The Gaussian formula (12.192): the normalised integral is the phase at the stationary point. The fluctuation determinant is independent of J and cancels against Z.
(2) $\frac{\pi}{k}\cdot\frac{1}{4\pi} = \frac{1}{4k}$, and $\oint\vec{B}_{C_m}\cdot\mathrm{d}\vec{x} = \frac{1}{4\pi}\oint\oint\mathrm{d}\vec{x}\cdot\big(\mathrm{d}\vec{y}\times(x-y)\big)/|x-y|^3$ with $\mathrm{d}\vec{x}\cdot(\mathrm{d}\vec{y}\times\vec{r}) = \varepsilon_{ijk}\mathrm{d}x^i\mathrm{d}y^j r^k$. This is Nash's (12.89), and the same as $\frac{i}{2}\iint JGJ$ with the propagator (12.203).

(3) Gauss's formula (12.159), which is Nash's (12.90).

This is Nash's (12.92). The Gaussian path integral of abelian Chern–Simons theory computes the Gauss linking numbers of the Wilson loops: *a topological quantum field theory produces a knot invariant*, exactly as advertised.

Watch out: Two things that look metric dependent but are not

(1) The propagator (12.203) was derived in the gauge $\nabla \cdot \vec{A} = 0$, which uses the flat metric. In another gauge G_{il} changes by terms of the form $\partial_i \Lambda_l(x, y)$ or $\partial_{y^l} \Lambda_i$, and these integrate to zero against the conserved current J (the chain after (12.195)). (2) On S^3 with its round metric the Green's function is a different function, but the Gauss integral is a degree (Section 12.10.4) and any kernel inverting $\nabla\times$ gives the same integer. Only the diagonal terms $l = m$ are delicate; they are the subject of Section 12.12.8.

Watch out: The sign of k

Nash's (12.86) carries tr and an anti-Hermitian A, his (12.88) carries the real A. With $A_{\text{anti-Herm}} = iA_{\text{real}}$, $\operatorname{tr}(A \wedge \mathrm{d}A) \to -A \wedge \mathrm{d}A$, which flips the sign of k. We used (12.88) as printed; with the opposite sign every k in (12.207) becomes $-k$, i.e. the answer is complex conjugated, which is the same as replacing the link by its mirror image. Nash's (12.89) agrees with our sign.

Worked example 12.84: The Hopf link and the unlink

For two components the double sum in (12.207) has two off-diagonal terms, $(l, m) = (1, 2)$ and $(2, 1)$, equal by symmetry of L:

$$\big\langle W(n_1, C_1) W(n_2, C_2) \big\rangle = \exp\Big[\frac{2\pi i}{k}\, n_1 n_2\, L(C_1, C_2)\Big] \times (\text{self-linking factors}).$$

For the Hopf link, $L = \pm 1$ and the cross term is $e^{\pm 2\pi i n_1 n_2/k}$. For the two-component unlink it is 1. For the Whitehead link it is also 1 (Section 12.10: $L = 0$), although the Whitehead link is not split: the abelian theory is as blind as the linking number itself.

Physical picture: Anyons

In $2 + 1$ dimensions a Wilson line in space-time is the world-line of a charged particle, and the Chern–Simons term attaches flux $2\pi n/k$ to charge n. Take two particles of unit charge and let one circle the other: the world-lines form a Hopf link, and the amplitude picks up $e^{2\pi i/k}$, an Aharonov–Bohm phase from the attached flux. Half a circuit, an exchange, gives $e^{i\pi/k}$: the particles are **anyons** with statistics angle $\theta = \pi/k$, neither bosons ($\theta = 0$) nor fermions ($\theta = \pi$). In the Laughlin state of the fractional quantum Hall effect at filling $\nu = 1/k$ the quasiparticles are exactly such anyons, and the low-energy theory is abelian Chern–Simons at level k.

12.12.8 Self-linking and framing

The trouble with $l = m$. The diagonal terms of (12.207) ask for $L(C_l, C_l)$, the Gauss integral of a curve with itself. In Section 12.10.6 we saw that this integral converges (the integrand is bounded on the diagonal) and defines the writhe $\operatorname{Wr}(C)$; but it is not an integer and not an isotopy invariant, because the degree argument breaks down on the diagonal $x = y$. In quantum field theory the problem is sharper: the term involves the propagator $\langle A_i(x) A_j(y) \rangle$ at

$x = y$, the product of two field operators at the same point, which has no unique meaning. Any definition requires a regularisation, and the answer depends on it. This is Nash's remark that $L(C_l, C_l)$ "has singularities due to points where $x - y = 0$".

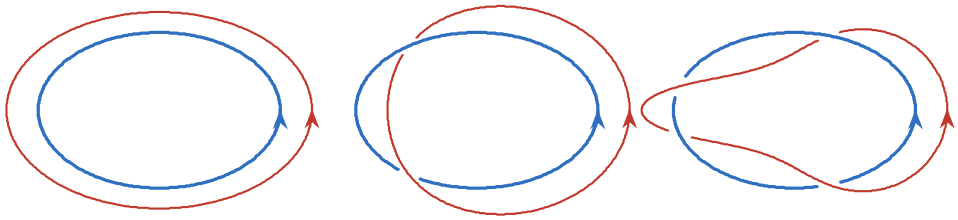
Point splitting. The standard cure for a coincident-point singularity is point splitting: evaluate one of the two fields at a nearby point. For a Wilson loop this means replacing one copy of C by a nearby parallel curve C'. Then $L(C, C')$ is a linking number of two *disjoint* curves, an honest integer and an isotopy invariant of the pair. But there are many choices of C'.

Working definition 12.33 (Framing, Nash (12.91))

A **framing** of a knot C is a choice of a nowhere-vanishing normal vector field $\vec{v}$ along C [at each point of C, a nonzero vector perpendicular to the tangent, varying smoothly], or equivalently a thin ribbon with C as one edge. The other edge, $C' = C + \varepsilon\vec{v}$, is the **push-off**. The **self-linking number** of the framed knot is defined to be

$$L(C, C) := L(C, C'). \tag{12.208}$$

To compute it: draw C and C' and count half the signed crossings between them (Section 12.10.1).



$t = 0$: $L(C, C') = 0$ $t = 1$: $L(C, C') = 1$ $t = 2$: $L(C, C') = 2$

Figure 12.57: Framings of the unknot. The blue curve is C, the red curve its push-off $C' = C + \varepsilon\vec{v}$, drawn in perspective. With the flat framing ($t = 0$) C' is a parallel circle and $L(C, C') = 0$. Each full twist of $\vec{v}$ about the tangent makes C' wind once more around C: two more crossings of the same sign in the picture, one more unit of linking.

How much freedom. Twisting the ribbon once more around C (a full 360° turn of $\vec{v}$ about the tangent) makes C' wind once more around C, so $L(C, C')$ changes by ± 1 (Figure 12.57). Every integer occurs. So a framing is, up to deformation, exactly an integer's worth of extra data per component. "A little thought", as Nash says, shows that this is the only ambiguity: two normal fields along C can be deformed into each other if and only if they give the same $L(C, C')$.

Where the non-integer writhe went. The Călugăreanu–White formula says that for a ribbon,

$$L(C, C') = \mathrm{Wr}(C) + \mathrm{Tw}(C, \vec{v}), \tag{12.209}$$

where $\mathrm{Wr}(C)$ is the writhe integral (12.164) and $\mathrm{Tw} = \frac{1}{2\pi}\oint (\hat{t} \times \vec{v}) \cdot \frac{\mathrm{d}\vec{v}}{\mathrm{d}s}\,\mathrm{d}s$ (for unit $\vec{v} \perp \hat{t}$) is the total rotation of the ribbon about its core, in units of full turns. Neither term on the right is an integer or an invariant; their sum is. As $\varepsilon \to 0$ the Gauss integral $L(C, C'_\varepsilon)$ does *not* tend to $\mathrm{Wr}(C)$: the integrand develops a spike near the diagonal whose integral is exactly Tw. This is the "singularity" at $x = y$: a delta-function-like contribution whose weight depends on how the limit is taken. DNA biophysicists use (12.209) daily, since a closed DNA loop is a ribbon: twisting it (Tw) makes it supercoil (Wr).

Worked example 12.85: Two framings of a round circle

Let C be the unit circle in the xy-plane. It lies in a plane, so $\mathrm{Wr}(C) = 0$ (Example 12.68). (a) The radial framing $\vec{v} = \hat{r}$ keeps the ribbon flat in the plane: C' is a larger concentric circle, disjoint from the disc bounded by C, so $L(C, C') = 0 = \mathrm{Tw}$. (b) The framing $\vec{v} = \cos(ts)\,\hat{r} - \sin(ts)\,\hat{z}$, with s the angle along C, rotates t times round the tangent: C' pierces the flat disc spanning C once for each full turn, always in the same direction, so $L(C, C') = t = \mathrm{Tw}$. Figure 12.57 shows $t = 0, 1, 2$.

The blackboard framing. Given a diagram D of C, there is a natural framing: take $\vec{v}$ lying in the plane of the paper (the blackboard), so that C' is drawn parallel to C everywhere in the picture. Away from crossings C and C' do not cross. At each crossing of D the two strands and their two parallel copies make four crossings (Figure 12.58), all with the same sign ε as the original one; exactly two of them are between C and C'. Hence

$$L(C, C'_{\mathrm{bb}}) \overset{(1)}{=} \frac{1}{2} \sum_{\substack{\text{crossings of} \\ C \text{ with } C'}} \varepsilon \overset{(2)}{=} \frac{1}{2} \sum_{\text{crossings } c \text{ of } D} 2\,\varepsilon_c \overset{(3)}{=} w(D). \tag{12.210}$$

(1) The crossing formula (12.155) for the linking number.
(2) Each crossing c of D produces two crossings between C and C', both of sign ε_c.
(3) Definition of the writhe of the diagram.

The blackboard self-linking number is the writhe of the diagram. This explains why the writhe appeared in the Kauffman bracket: an R1 move changes $w(D)$ by ± 1, i.e. it changes the blackboard framing by one twist (flatten a kink in a ribbon and you get a twisted ribbon), and the bracket is an invariant of framed knots.

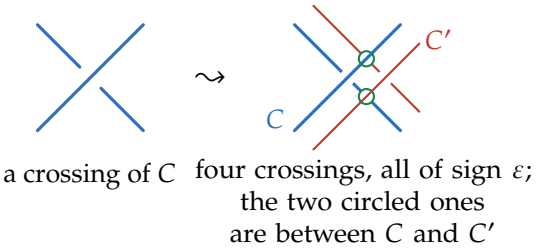


Figure 12.58: The blackboard framing. The push-off C' (red) is drawn parallel to C (blue) in the plane of the diagram. Each crossing of the diagram becomes four crossings of the same sign; two of them are between C and C', so each crossing contributes $\frac{1}{2} \cdot 2\varepsilon = \varepsilon$ to $L(C, C')$, and $L(C, C'_{\mathrm{bb}}) = w(D)$.

The canonical framing in S^3. For a knot in S^3 (or any homology sphere) there is a preferred framing: push C off along a Seifert surface S (Section 12.11.2). The push-off C' lies on a parallel copy of S, which C does not meet, so the signed count of piercings is 0 and $L(C, C') = 0$. This **Seifert** (or zero) framing is why the self-linking ambiguity is harmless for knots in S^3: one always knows how to convert to the canonical choice. In a general three-manifold no such canonical choice need exist, and one must carry the framing along as data.

The correlation function in terms of framed links, Nash (12.92)–(12.94). With each component framed, all the $L(C_l, C_m)$ in (12.207) are integers, and

$$\big\langle W(n_1, C_1) \cdots W(n_p, C_p) \big\rangle = \exp\Big[\frac{i\pi}{k} \sum_{l,m} n_l n_m\, L(C_l, C_m)\Big] \tag{12.211}$$

is an invariant of *framed* links. Change the framing of C_l by t extra twists. Then $L(C_l, C_l) \to L(C_l, C_l) + t$, the diagonal term $l = m$ of the double sum changes by $n_l^2 t$, and nothing else changes:

$$\begin{aligned}\big\langle W(n_1, C_1) \cdots W(n_p, C_p) \big\rangle &\overset{(1)}{=} \exp\Big[\frac{i\pi}{k}\Big(\sum_{l,m} n_l n_m L(C_l, C_m) + n_l^2 t\Big)\Big] \\ &\overset{(2)}{=} \exp\Big[\frac{i\pi}{k} n_l^2 t\Big] \exp\Big[\frac{i\pi}{k} \sum_{l,m} n_l n_m L(C_l, C_m)\Big].\end{aligned} \tag{12.212}$$

(1) Only the (l, l) entry of the matrix of linking numbers changes.
(2) Split the exponential. This is Nash's (12.93).

So under a change of framing,

$$\langle W \rangle \longmapsto \exp\Big[\frac{2\pi i\, t\, n_l^2}{2k}\Big] \langle W \rangle, \tag{12.213}$$

Nash's (12.94): each twist costs a phase, not a new unknown, and one can always translate between framings.

Key idea: Framing phases are conformal weights

Write the phase per twist in (12.213) as $e^{2\pi i h_n}$ with $h_n = n^2/2k$. In the two-dimensional conformal field theory of a compact boson at level k (the one Witten's construction associates with U(1) Chern–Simons), the primary field of charge n has conformal weight exactly $h_n = n^2/2k$. A full twist of a ribbon is a 2π rotation of the particle whose world-line it is, and $e^{2\pi i h}$ is the phase of a spin-h object under a 2π rotation. In the SU(2) theory the same argument gives $e^{2\pi i h_j}$ with $h_j = j(j+1)/(k+2)$, and for the Kauffman bracket the corresponding factor is $-A^3$. This is the first of many places where Nash's remark about conformal field theory being indispensable comes true.

Worked example 12.86: The framed unknot and the framed trefoil

A single unknot of charge n with framing f (self-linking f) has $\langle W \rangle = e^{i\pi n^2 f/k}$. With the canonical framing $f = 0$ it is 1.
Draw the right trefoil in the standard way; $w(D) = +3$, so in the blackboard framing $L(C, C'_{\mathrm{bb}}) = 3$ and $\langle W(n, 3_1) \rangle_{\mathrm{bb}} = e^{3i\pi n^2/k}$. Converting to the canonical framing with $t = -3$ gives $\langle W(n, 3_1) \rangle_0 = 1$, the same as the unknot. *For a single knot in S^3 with canonical framing the abelian theory always gives* 1: it sees linking between components, but nothing of the knotting of a single component.

12.12.9 The abelian skein relation, and what is missing

Let every component carry charge $n = 1$ and use the blackboard framing of a diagram D of L. The double sum in (12.211) then collects every crossing of D exactly once with its sign:

$$\begin{aligned}\sum_{l,m} L(C_l, C_m) &\overset{(1)}{=} \sum_l w(D_l) + 2 \sum_{l<m} L(C_l, C_m) \\ &\overset{(2)}{=} \sum_l \sum_{c \in D_l} \varepsilon_c + \sum_{\substack{c \text{ between} \\ \text{different components}}} \varepsilon_c \overset{(3)}{=} w(D).\end{aligned} \tag{12.214}$$

(1) Diagonal terms by (12.210); off-diagonal terms in pairs by symmetry.
(2) The writhe of the sub-diagram D_l counts the self-crossings of C_l; twice the linking number counts the crossings between two components.
(3) Together these are all crossings of D.

So, with $q := e^{i\pi/k}$,

$$\langle W(L)\rangle_{\mathrm{bb}} = q^{\,w(D)}. \tag{12.215}$$

Now take a skein triple. Switching one crossing from L_+ to L_- lowers w by 2:

$$q^{-1}\langle W(L_+)\rangle - q\,\langle W(L_-)\rangle = 0. \tag{12.216}$$

Compare the Jones relation (12.179), $t^{-1}V_{L_+} - tV_{L_-} = (t^{1/2} - t^{-1/2})V_{L_0}$. The left-hand sides have the same shape; the abelian theory has *no* L_0 *term*. Also, for split unlinks the abelian theory gives $\langle W(U_p)\rangle = 1$, where Jones has $(-t^{1/2} - t^{-1/2})^{p-1}$: each extra circle is worth 1 rather than the "quantum dimension" $-(t^{1/2} + t^{-1/2})$.

Key idea: Where the L_0 term comes from

In the non-abelian theory the propagator carries group indices, $\langle A_i^a(x)A_j^b(y)\rangle \propto \delta^{ab}$, and each propagator joining two strands inserts $\sum_a T^a \otimes T^a$ acting on the two strands' representation spaces. For SU(2) in the spin-$\frac{1}{2}$ representation, $T^a = \sigma^a/2$, this operator can be written with the *swap* $\mathsf{P}|u\rangle \otimes |v\rangle = |v\rangle \otimes |u\rangle$:

$$\sum_{a=1}^{3} \sigma^a \otimes \sigma^a = 2\mathsf{P} - \mathbb{1}, \qquad \sum_a T^a \otimes T^a = \tfrac{1}{2}\big(\mathsf{P} - \tfrac{1}{2}\mathbb{1}\big). \tag{12.217}$$

(Check on $|\uparrow\uparrow\rangle$: $\sigma^3 \otimes \sigma^3$ gives $+1$, $\sigma^1 \otimes \sigma^1$ and $\sigma^2 \otimes \sigma^2$ give $|\downarrow\downarrow\rangle$ and $i^2|\downarrow\downarrow\rangle$, which cancel; so the sum is $1 = 2 \cdot 1 - 1$. On the singlet $\mathsf{P} = -1$ and the sum is $\vec{\sigma}_1 \cdot \vec{\sigma}_2 = -3 = 2(-1) - 1$.) The identity $\mathbb{1}$ acts like the abelian theory: it lets the two strands go straight through. The swap P *reconnects* them, and reconnecting the four ends at a crossing without the crossing is exactly the smoothing L_0. So the extra term in the Jones skein relation is the footprint of non-commutativity. The honest derivation of the Jones relation, with the exact coefficients as functions of k, uses the conformal-field-theory machinery of Chapter 12; the identity (12.217) is its leading-order shadow.

Try it: Charges mod k

Show that in (12.211) replacing n_1 by $n_1 + 2k$ changes nothing, for any framed link. (Cross terms change by $e^{2\pi i \cdot 2n_m L} = 1$; the diagonal term changes by $e^{i\pi(4kn_1+4k^2)L_{11}/k} = 1$.) So the abelian theory at level k has only finitely many distinct Wilson lines, a first glimpse of the finiteness of the state spaces of a topological field theory.

Reading Nash (pp. 342–345, (12.84)–(12.94))

Nash introduces Witten's programme (an intrinsically three-dimensional definition of the Jones polynomial, needing conformal field theory, and to be verified through the skein relation and the unlink values), defines the Wilson line (12.84) and the normalised correlation functions (12.85), and observes that they need no non-topological data. He then specialises to U(1) (12.86)–(12.88), and states that completing the square on S^3 gives (12.89), i.e. the Gauss linking numbers (12.90), integers by Ampère's law. The self-linking singularity is cured by framing (12.91); the correlator is (12.92), and a change of framing by

t twists multiplies it by $e^{i\pi n_l^2 t/k}$ (12.93)–(12.94). We have supplied: the construction of the holonomy and its gauge invariance, the Gaussian integral with sources and zero modes, the propagator (in Fourier space, where the curl is a Pauli matrix, and via Ampère's law), the blackboard and Seifert framings, the Călugăreanu–White formula, the abelian skein relation and its comparison with Jones, and the physics of flux attachment and anyons.

Recap

- The Wilson loop $W(R,C) = \mathrm{tr}_R\, P \exp \oint_C A$ is the trace of the holonomy, an ordered product of link variables; gauge transformations conjugate the holonomy, so the trace is invariant. For U(1) it is the Aharonov–Bohm phase $e^{in\oint A}$, $n \in \mathbb{Z}$.
- Chern–Simons correlators contain no metric; they are candidate isotopy invariants. Gauge fixing uses a metric, but the answer does not depend on it.
- Abelian theory: Gaussian integral with the conserved source $J = \sum n_l J_{C_l}$. The kernel is $\frac{k}{2\pi}\nabla\times$; in Fourier space the curl is $p\,\sigma_2$ with eigenvalues 0 (gauge), $\pm p$ (helicities).
- The classical field of a Wilson loop is its Biot–Savart field times $2\pi/k$ (flux attachment). The correlator is $\exp\left[\frac{i\pi}{k}\sum_{l,m} n_l n_m L(C_l, C_m)\right]$.
- Self-linking needs a framing: $L(C,C) := L(C,C')$, with $L = \mathrm{Wr} + \mathrm{Tw}$. Blackboard framing gives the writhe $w(D)$; Seifert framing gives 0. A change by t twists multiplies $\langle W\rangle$ by $e^{i\pi n^2 t/k} = e^{2\pi i h_n t}$, $h_n = n^2/2k$.
- With unit charges, $\langle W\rangle_{\mathrm{bb}} = q^{w(D)}$ and $q^{-1}W_+ - qW_- = 0$: the Jones relation without the L_0 term. The L_0 term comes from $\sum_a T^a \otimes T^a = \frac{1}{2}(\mathsf{P} - \frac{1}{2})$.

12.13 Chern–Simons theory at weak coupling: flat connections and torsion

The abelian theory was Gaussian, so it could be done exactly. For a non-abelian group the cubic term $\frac{2}{3}A \wedge A \wedge A$ spoils this. Nash's second piece of evidence for the topological nature of the theory is the *weak-coupling*, or large-k, limit: expand the path integral around its stationary points, do the resulting Gaussian integral, and look at what comes out. The stationary points are the flat connections, and the Gaussian integral produces a ratio of determinants of Laplacians that turns out to be a classical topological invariant, the Ray–Singer analytic torsion. This section fills in every step of that argument [Nash pp. 346–350, (12.95)–(12.113)].

The ingredients are standard tools of quantum field theory, each of which we first rehearse in finite dimensions:

1. the method of stationary phase (Section 12.13.2);
2. the expansion of the action to second order about a flat connection (Section 12.13.3);
3. gauge fixing with a Lagrange multiplier and Faddeev–Popov ghosts (Section 12.13.4);
4. Gaussian integrals as determinants, regularised with the zeta function, and the phase given by the eta invariant (Section 12.13.6);
5. the recognition of the answer as the analytic torsion (Section 12.13.7).

Figure 12.59 shows how these steps feed into one another.

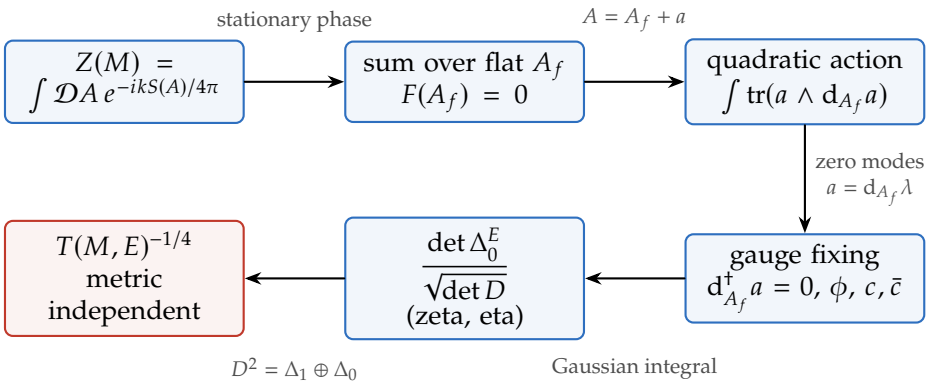


Figure 12.59: The logic of the weak-coupling computation. Read it clockwise from the top left: stationary phase localises the path integral on flat connections; the action is expanded to second order about each; the gauge zero modes are removed with a multiplier and ghosts; the resulting Gaussian is a ratio of determinants; and because D squares to the Laplacians the ratio is a power of the Ray–Singer torsion, from which the metric has disappeared.

12.13.1 The coupling constant is $1/k$

In physics notation the gauge coupling g sits in the commutator term of the curvature. With Hermitian generators $[T^a, T^b] = i f^{abc} T^c$ and the dictionary of Section 12.12.2, Nash's (12.95) reads

$$F(A) = \tfrac{1}{2} F^a_{\mu\nu} T^a \, \mathrm{d}x^\mu \wedge \mathrm{d}x^\nu, \qquad F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu + g\, f^{abc} A^b_\mu A^c_\nu. \tag{12.218}$$

In the language of forms: if $F = \mathrm{d}A + g\, A \wedge A$ then the Chern–Simons form that goes with it is $\mathrm{tr}(A \wedge \mathrm{d}A + \frac{2}{3} g\, A \wedge A \wedge A)$. Rescale $A = \tilde{A}/g$:

$$\begin{aligned} \mathrm{tr}\Big(A \wedge \mathrm{d}A + \tfrac{2}{3} g\, A \wedge A \wedge A\Big) &\overset{(1)}{=} \mathrm{tr}\Big(\frac{\tilde{A} \wedge \mathrm{d}\tilde{A}}{g^2} + \frac{2}{3}\, g\, \frac{\tilde{A} \wedge \tilde{A} \wedge \tilde{A}}{g^3}\Big) \\ &\overset{(2)}{=} \frac{1}{g^2}\, \mathrm{tr}\Big(\tilde{A} \wedge \mathrm{d}\tilde{A} + \tfrac{2}{3} \tilde{A} \wedge \tilde{A} \wedge \tilde{A}\Big). \end{aligned} \tag{12.219}$$

(1) Substitute $A = \tilde{A}/g$: the quadratic term gets g^{-2}, the cubic term $g \cdot g^{-3}$.
(2) Both terms now carry the same factor g^{-2}.

(Nash writes the quadratic term of (12.95) as $i g\, f^{abc} A^b_\mu A^c_\nu$. The factor of i depends on whether the generators are taken Hermitian or anti-Hermitian and on how A is normalised; with our Hermitian T^a and $[T^a, T^b] = i f^{abc} T^c$ the component form above has no i. What matters here is only that g multiplies the part of F quadratic in A.) So the rescaled field has the standard curvature $\tilde{F} = \mathrm{d}\tilde{A} + \tilde{A} \wedge \tilde{A}$ and the action becomes Nash's (12.96),

$$-\frac{ik}{4\pi g^2} \int_M \mathrm{tr}\Big(\tilde{A} \wedge \mathrm{d}\tilde{A} + \tfrac{2}{3} \tilde{A} \wedge \tilde{A} \wedge \tilde{A}\Big).$$

Only the combination k/g^2 occurs. Small coupling $g \to 0$ is therefore the same as large level $k \to \infty$, and from now on we set $g = 1$ and let k be large. The analogy is exact: in $\int \mathcal{D}A\, e^{iS/\hbar}$ the classical limit is $\hbar \to 0$, and here the exponent is $ik\, S_{\mathrm{CS}}/4\pi$, so

$$\hbar_{\mathrm{eff}} = \frac{4\pi}{k}.$$

(There is no freedom to vary k continuously: invariance under large gauge transformations forces $k \in \mathbb{Z}$. But large integers are allowed, and the large-k expansion is the semiclassical expansion.)

12.13.2 Stationary phase: the finite-dimensional model

Working definition 12.34 (Stationary phase)

For a smooth real function f on $\mathbb{R}^n$ with isolated non-degenerate critical points x_c (where $\nabla f = 0$ and the Hessian $H_c = \partial^2 f(x_c)$ is invertible),

$$\int \mathrm{d}^n x \; g(x)\, e^{ikf(x)} \xrightarrow{k\to\infty} \sum_{x_c} g(x_c)\, e^{ikf(x_c)} \left(\frac{2\pi}{k}\right)^{n/2} \frac{e^{i\pi\, \mathrm{sgn}(H_c)/4}}{|\det H_c|^{1/2}}, \tag{12.220}$$

where $\mathrm{sgn}(H_c)$ (the **signature**) is the number of positive minus the number of negative eigenvalues of H_c. *Recipe:* find the critical points; at each, evaluate the phase; expand f to second order and do the Fresnel integral (12.190) [$\int \mathrm{d}x\, e^{ik\lambda x^2/2} = (2\pi/k\,|\lambda|)^{1/2} e^{i\pi\,\mathrm{sgn}(\lambda)/4}$ for real $\lambda \neq 0$], one eigenvalue at a time.

Why: away from critical points e^{ikf} oscillates faster and faster as k grows and the integral cancels itself out; near x_c, $f \approx f(x_c) + \frac{1}{2}(x-x_c)^T H_c (x-x_c)$, and the Gaussian width is $\sim k^{-1/2}$. This is exactly how the classical path is singled out in Feynman's path integral: the paths near the stationary one add up in phase, and all the others cancel.

Worked example 12.87: The Bessel function

Take $f(x) = \cos x$ on the circle $0 \le x < 2\pi$. The exact integral is a Bessel function, $\int_0^{2\pi} e^{ik\cos x}\mathrm{d}x = 2\pi J_0(k)$. Critical points: $x = 0$ ($f = 1$, $f'' = -1$) and $x = \pi$ ($f = -1$, $f'' = +1$). Then

$$\int_0^{2\pi} e^{ik\cos x}\mathrm{d}x \overset{(1)}{=} \sqrt{\frac{2\pi}{k}}\left(e^{ik}e^{-i\pi/4} + e^{-ik}e^{+i\pi/4}\right)$$
$$\overset{(2)}{=} 2\sqrt{\frac{2\pi}{k}}\cos\left(k - \frac{\pi}{4}\right).$$

(1) Formula (12.220): $n = 1$, $|f''| = 1$ at both points, signatures -1 and $+1$.
(2) $e^{i\theta} + e^{-i\theta} = 2\cos\theta$ with $\theta = k - \pi/4$.

This is the textbook asymptotic form $J_0(k) \approx \sqrt{2/\pi k}\,\cos(k - \pi/4)$: multiply by 2π and compare. Two things to notice for later. The *phases* $e^{\pm i\pi/4}$ come from the signs of the Hessian eigenvalues, and they matter: they produce the $-\pi/4$ shift. And the *power* $k^{-1/2}$ counts the dimension of the integral.

12.13.3 Flat connections and the expansion of the action

The partition function and its stationary points, Nash (12.97)–(12.98). The partition function is

$$Z(M) = \int \mathcal{D}A \; \exp\left[-\frac{ik}{4\pi}\, S(A)\right], \qquad S(A) = \int_M \mathrm{tr}\left(A\wedge \mathrm{d}A + \tfrac{2}{3} A\wedge A\wedge A\right). \tag{12.221}$$

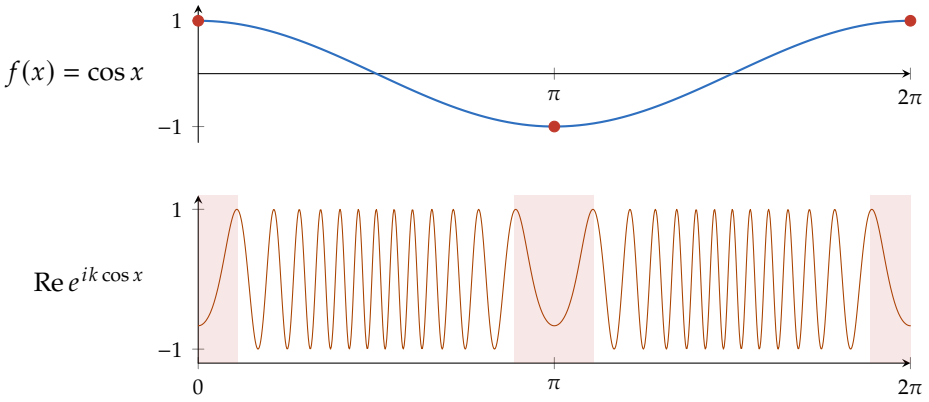


Figure 12.60: Stationary phase for $\int_0^{2\pi} e^{ik\cos x}\,dx$ at $k = 40$. Top: the phase function $f = \cos x$ with its critical points (red). Bottom: the real part of the integrand. Away from the critical points it oscillates rapidly and its positive and negative lobes cancel; only the shaded windows of width $\sim k^{-1/2}$ around $x = 0$ and $x = \pi$ contribute. In the path integral the same mechanism picks out the flat connections.

To find its stationary points, vary $A \to A + \delta A$. We need two facts about traces of matrix-valued forms: for a p-form α and a q-form β, $\operatorname{tr}(\alpha \wedge \beta) = (-1)^{pq} \operatorname{tr}(\beta \wedge \alpha)$ (cyclicity of the trace combined with the sign for moving forms past each other), and $\mathrm{d}\operatorname{tr}(\alpha \wedge \beta) = \operatorname{tr}(\mathrm{d}\alpha \wedge \beta) + (-1)^p \operatorname{tr}(\alpha \wedge \mathrm{d}\beta)$. Then

$$\begin{aligned} \delta S &\overset{(1)}{=} \int_M \operatorname{tr}\Big(\delta A \wedge \mathrm{d}A + A \wedge \mathrm{d}\delta A + \tfrac{2}{3}\big(\delta A \wedge A \wedge A + A \wedge \delta A \wedge A + A \wedge A \wedge \delta A\big)\Big) \\ &\overset{(2)}{=} \int_M \operatorname{tr}\big(2\,\delta A \wedge \mathrm{d}A + 2\,\delta A \wedge A \wedge A\big) \overset{(3)}{=} 2\int_M \operatorname{tr}\big(\delta A \wedge F\big). \end{aligned} \tag{12.222}$$

(1) First-order terms of $S(A + \delta A)$.

(2) $\int \operatorname{tr}(A \wedge \mathrm{d}\delta A) = \int \operatorname{tr}(\mathrm{d}A \wedge \delta A) - \int \mathrm{d}\operatorname{tr}(A \wedge \delta A) = \int \operatorname{tr}(\delta A \wedge \mathrm{d}A)$, using Stokes on closed M and then moving the 1-form δA past the 2-form $\mathrm{d}A$ (sign $(-1)^2 = +1$). The three cubic terms are equal: moving a 1-form past a 2-form inside the trace costs no sign.

(3) $F = \mathrm{d}A + A \wedge A$.

So $\delta S = 0$ for all δA exactly when

$$F(A) = 0\,: \tag{12.223}$$

the classical solutions of Chern–Simons theory are the **flat connections**. By stationary phase, for large k the partition function should become a sum over flat connections, each weighted by $e^{-ikS(A_f)/4\pi}$ and by a Gaussian fluctuation integral.

Working definition 12.35 (Flat connections as representations of π_1)

A flat connection has holonomy that depends only on the homotopy class of the loop [two loops at the base point count as the same if one can be deformed into the other]: by the non-abelian Stokes theorem the holonomy round a small contractible loop is $\mathbb{1} + F \cdot (\text{area}) + \cdots = \mathbb{1}$. So a flat connection defines a homomorphism $\rho : \pi_1(M) \to G$ [going round one loop and then another gives the product of the two holonomies], well defined up to conjugation [$\rho \to h\rho h^{-1}$ for one fixed $h \in G$] (the choice of base point and of gauge), and conversely every such homomorphism comes from a flat connection. *To list the flat connections, list the homomorphisms $\pi_1(M) \to G$ up to conjugation.*

Worked example 12.88: Flat connections on some three-manifolds

- $M = S^3$: $\pi_1 = 1$, so the only flat connection is the trivial one, $A_f = 0$ (up to gauge).
- $M = S^1 \times S^2$, $G = \mathrm{U}(1)$: $\pi_1 = \mathbb{Z}$, so the flat connections are labelled by one holonomy $e^{i\theta}$ round the S^1, a whole circle's worth. They are not isolated; this is the situation with zero modes that Nash excludes at the end.
- $M = L(p, q)$, a lens space (S^3 divided by the $\mathbb{Z}_p$ action $(z_1, z_2) \mapsto (e^{2\pi i/p} z_1, e^{2\pi i q/p} z_2)$), $G = \mathrm{U}(1)$: $\pi_1 = \mathbb{Z}_p$, so there are p flat connections, with holonomies $e^{2\pi i m/p}$, $m = 0, \dots, p-1$. They are isolated points.
- M = the Poincaré homology sphere, $G = \mathrm{SU}(2)$: π_1 is the binary icosahedral group of order 120, and apart from the trivial one there are exactly two irreducible flat SU(2) connections.

The expansion about a flat connection, Nash (12.99). Put $A = A_f + a$. We need the covariant derivative on Lie-algebra-valued forms, $\mathrm{d}_{A_f} a = \mathrm{d}a + A_f \wedge a + a \wedge A_f$ for a 1-form a (the graded commutator $[A_f, a]$). Collect powers of a:

$$\begin{aligned} S(A_f + a) &\overset{(1)}{=} S(A_f) + 2\int_M \mathrm{tr}\big(a \wedge F(A_f)\big) + \int_M \mathrm{tr}\big(a \wedge \mathrm{d}a + 2\,A_f \wedge a \wedge a\big) + \tfrac{2}{3}\int_M \mathrm{tr}(a \wedge a \wedge a) \\ &\overset{(2)}{=} S(A_f) + 2\int_M \mathrm{tr}\big(F(A_f) \wedge a\big) + \int_M \mathrm{tr}\big(a \wedge \mathrm{d}_{A_f} a\big) + \tfrac{2}{3}\int_M \mathrm{tr}(a \wedge a \wedge a) \\ &\overset{(3)}{=} S(A_f) + \int_M \mathrm{tr}\big(a \wedge \mathrm{d}_{A_f} a\big) + O(a^3). \end{aligned} \tag{12.224}$$

(1) The term linear in a is (12.222) with $\delta A = a$ (the two linear pieces $\mathrm{tr}(A_f \wedge \mathrm{d}a) + \mathrm{tr}(a \wedge \mathrm{d}A_f)$ of $\mathrm{tr}(A \wedge \mathrm{d}A)$ become $2\,\mathrm{tr}(a \wedge \mathrm{d}A_f)$ after the same integration by parts). The quadratic piece of $\mathrm{tr}(A \wedge \mathrm{d}A)$ is simply $\mathrm{tr}(a \wedge \mathrm{d}a)$. From the cubic term come the three orderings $A_f aa$, $aA_f a$, aaA_f, which are equal by cyclicity (moving a 1-form past a 2-form inside the trace costs no sign), so $\frac{2}{3} \cdot 3 = 2$ copies of $\mathrm{tr}(A_f \wedge a \wedge a)$.

(2) $\mathrm{tr}(a \wedge \mathrm{d}_{A_f} a) = \mathrm{tr}(a \wedge \mathrm{d}a) + \mathrm{tr}(a \wedge A_f \wedge a) + \mathrm{tr}(a \wedge a \wedge A_f) = \mathrm{tr}(a \wedge \mathrm{d}a) + 2\,\mathrm{tr}(A_f \wedge a \wedge a)$, again by cyclicity; and $\mathrm{tr}(a \wedge F) = \mathrm{tr}(F \wedge a)$.

(3) $F(A_f) = 0$.

The expansion *terminates* at third order: the action is a cubic polynomial in A. Nash writes the first two lines as (12.99).

The quadratic form $\int \mathrm{tr}(a \wedge \mathrm{d}_{A_f} a)$ is the non-abelian version of the helicity $\int \vec{a} \cdot \nabla \times \vec{a}$ of (12.193). It is invariant under infinitesimal gauge transformations $a \to a + \mathrm{d}_{A_f}\lambda$ precisely because A_f is flat: $\mathrm{d}_{A_f}\mathrm{d}_{A_f}\lambda = [F(A_f), \lambda] = 0$. So once more the kernel has zero modes along the gauge directions, and we must fix a gauge.

12.13.4 Gauge fixing: Lagrange multipliers and ghosts

Faddeev–Popov in two dimensions. The whole procedure can be seen in an ordinary double integral with a rotation symmetry. Let $I = \int \mathrm{d}^2x\, e^{-V(r)}$ with $r = |x|$. The "gauge group" is the rotation group SO(2); the "gauge orbits" are circles; a "gauge condition" is a curve $G(x) = 0$ meeting each orbit, say $G(x) = x_2$ (the x_1-axis). Insert 1 in the form

$$1 = \frac{1}{2}\int_0^{2\pi} \mathrm{d}\theta\; \delta\big(G(R_\theta x)\big)\left|\frac{\partial G(R_\theta x)}{\partial \theta}\right|, \qquad x \neq 0, \tag{12.225}$$

where R_θ is rotation by θ. This is the formula $\int \mathrm{d}\theta\, \delta(h(\theta))\,|h'(\theta)| = \#\{\text{zeros of } h\}$ applied to $h(\theta) = G(R_\theta x) = (R_\theta x)_2$, which vanishes at the two angles that rotate x onto the positive and the negative x_1-axis; hence the $\frac{1}{2}$. Then

$$I \overset{(1)}{=} \frac{1}{2}\int_0^{2\pi} \mathrm{d}\theta \int \mathrm{d}^2x\ \delta(x_2)\ |x_1|\ e^{-V(r)}$$
$$\overset{(2)}{=} \frac{1}{2}\cdot 2\pi \int_{-\infty}^{\infty} \mathrm{d}x_1\ |x_1|\ e^{-V(|x_1|)} \overset{(3)}{=} 2\pi \int_0^{\infty} r\, e^{-V(r)}\, \mathrm{d}r. \tag{12.226}$$

(1) Insert (12.225), change variables $x \to R_{-\theta}x$ (the integrand and d^2x are rotation invariant), and evaluate the Jacobian on the slice: $\partial_\theta(R_\theta x)_2 = x_1$ at $x_2 = 0$ (the derivative of a rotation is $\partial_\theta R_\theta x = J\, R_\theta x$ with $J = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$, whose second component is the first component of $R_\theta x$). The $\frac{1}{2}$ is the one in (12.225): each circle meets the x_1-axis *twice*, at $\pm r$.
(2) The θ integral gives the "volume of the gauge group", 2π; the delta function sets $x_2 = 0$.
(3) The integrand is even in x_1.

This is polar coordinates. The **Faddeev–Popov determinant** is the Jacobian $|x_1| = r$, the size of the orbit through the slice; the factor 2π is the gauge-group volume that we divide out in field theory; and the double intersection is a toy **Gribov ambiguity**. Both the δ-function and the determinant are now written as integrals, so that everything becomes an exponent.

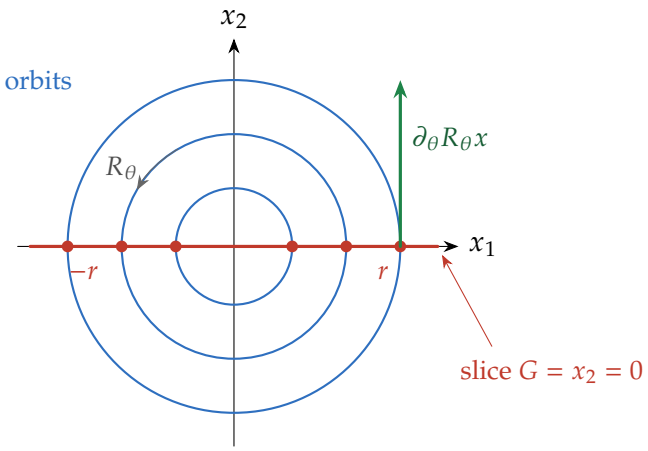


Figure 12.61: Faddeev–Popov in the plane. The "gauge orbits" of rotations are the blue circles; the gauge slice $x_2 = 0$ (red) meets each orbit twice, at $\pm r$ (a toy Gribov copy). The Faddeev–Popov determinant is the speed with which the orbit crosses the slice, the green vector $\partial_\theta(R_\theta x)_2 = x_1$ of length r: integrating over the slice with weight r and multiplying by the orbit "volume" 2π is just polar coordinates.

The delta function as a Lagrange multiplier. $\delta(G) = \int \frac{\mathrm{d}\phi}{2\pi}\, e^{i\phi G}$. Enforcing the gauge condition is the same as adding $\phi\, G$ to the action and integrating over the new field ϕ.

The determinant as a Grassmann integral. Grassmann variables $\theta_1, \bar\theta_1, \ldots$ anticommute ($\theta_i\theta_j = -\theta_j\theta_i$, so $\theta_i^2 = 0$). A function of one Grassmann variable is $a + b\theta$, and integration is defined by $\int \mathrm{d}\theta\, 1 = 0$, $\int \mathrm{d}\theta\, \theta = 1$; with pairs we use the convention $\int \mathrm{d}\bar\theta\, \mathrm{d}\theta\ \theta\bar\theta = 1$. Then for a 1×1 "matrix" b:

$$\int \mathrm{d}\bar\theta\, \mathrm{d}\theta\ e^{-\bar\theta b\theta} = \int \mathrm{d}\bar\theta\, \mathrm{d}\theta\ (1 - \bar\theta b\theta) = \int \mathrm{d}\bar\theta\, \mathrm{d}\theta\ (1 + b\,\theta\bar\theta) = b.$$

The series stops because $\theta^2 = 0$. For an $n \times n$ matrix B the same computation gives $\det B$, *not* $(\det B)^{-1/2}$ as for a real bosonic Gaussian. Here is $n = 2$ in full. Only the second-order term $\frac{1}{2}(\bar\theta B\theta)^2$ contains all four variables:

$$\begin{aligned}\tfrac{1}{2}(\bar\theta B\theta)^2 &\overset{(1)}{=} B_{11}B_{22}\,\bar\theta_1\theta_1\bar\theta_2\theta_2 + B_{12}B_{21}\,\bar\theta_1\theta_2\bar\theta_2\theta_1 \\ &\overset{(2)}{=} \big(-B_{11}B_{22} + B_{12}B_{21}\big)\,\bar\theta_1\theta_1\theta_2\bar\theta_2 \overset{(3)}{=} \det B\;\theta_1\bar\theta_1\,\theta_2\bar\theta_2 .\end{aligned}$$

(1) In $\sum B_{ij}B_{kl}\bar\theta_i\theta_j\bar\theta_k\theta_l$ we need $\{i,k\} = \{j,l\} = \{1,2\}$; the terms with $i = 2, k = 1$ repeat those with $i = 1, k = 2$ (the bilinears $\bar\theta_i\theta_j$ are even and commute), cancelling the $\frac{1}{2}$.
(2) $\bar\theta_2\theta_2 = -\theta_2\bar\theta_2$; and in the second term move θ_1 two places left, past $\bar\theta_2$ and θ_2 (sign $(+1)$).
(3) $\bar\theta_1\theta_1 = -\theta_1\bar\theta_1$.

Integrating with $\int \mathrm{d}\bar\theta_1\mathrm{d}\theta_1\mathrm{d}\bar\theta_2\mathrm{d}\theta_2\;\theta_1\bar\theta_1\theta_2\bar\theta_2 = 1$ gives $\det B$.

Try it: Bosons against fermions

For a complex variable $z = x + iy$ and $b > 0$ show that $\int \frac{\mathrm{d}x\,\mathrm{d}y}{\pi}\, e^{-b\bar z z} = \frac{1}{b}$, and hence that an n-component complex Gaussian gives $(\det B)^{-1}$. (Polar coordinates: $\int_0^\infty 2r\, e^{-br^2}\mathrm{d}r = 1/b$.) So a pair of commuting fields $\bar z, z$ gives $\det^{-1}$, a pair of anticommuting fields $\bar\theta, \theta$ gives $\det^{+1}$, and one real field gives $\det^{-1/2}$.

So a determinant in the numerator is a Gaussian integral over anticommuting **ghost** fields $c, \bar c$. The ghosts are the Faddeev–Popov Jacobian in disguise; they are not physical particles.

The gauge-fixed partition function, Nash (12.100)–(12.101). Pick a metric on M and impose the covariant Lorenz gauge

$$\mathrm{d}^\dagger_{A_f} a = 0, \tag{12.227}$$

where $\mathrm{d}^\dagger_{A_f}$ is the adjoint of d_{A_f} with respect to the metric ($\mathrm{d}^\dagger = -\star\,\mathrm{d}\star$ on 1-forms in three dimensions; Nash writes $\mathrm{d}^*_{A_f}$). For $A_f = 0$ on $\mathbb{R}^3$ this is $\nabla\cdot\vec a = 0$, the gauge used in Section 12.12.7. The infinitesimal gauge transformation $a \to a + \mathrm{d}_{A_f}\lambda$ changes the gauge function by $\mathrm{d}^\dagger_{A_f}\mathrm{d}_{A_f}\lambda$, so the Faddeev–Popov operator is $\mathrm{d}^\dagger_{A_f}\mathrm{d}_{A_f} = \Delta_0$, the covariant Laplacian on Lie-algebra-valued functions. Assembling the pieces,

$$\begin{aligned}Z(M) = \sum_{A_f} e^{-ikS(A_f)/4\pi} \int \mathcal{D}a\,\mathcal{D}\phi\,\mathcal{D}c\,\mathcal{D}\bar c \\ \times \exp\Big[-\frac{ik}{4\pi}\int_M \Big\{\mathrm{tr}\big(a\wedge \mathrm{d}_{A_f}a\big) + 2\,\mathrm{tr}\big(\phi\wedge\star\mathrm{d}^\dagger_{A_f}a\big) + \mathrm{tr}\big(\mathrm{d}_{A_f}\bar c\wedge\star\mathrm{d}_{A_f}c\big)\Big\}\Big].\end{aligned} \tag{12.228}$$

Here ϕ is the Lagrange multiplier, a Lie-algebra-valued function (Nash uses the equivalent 3-form $\star\phi$), and $c, \bar c$ are Lie-algebra-valued Grassmann functions. We have scaled the gauge-fixing and ghost terms by the same factor $ik/4\pi$ for convenience; this only changes the overall constant.

Sections of the flat bundle. The fields $a, \phi, c, \bar c$ are matrix-valued forms. At a flat connection A_f they are best viewed as forms with values in the vector bundle E (the adjoint bundle, with fibre the Lie algebra) on which A_f acts: Nash's

$$\Omega^p(M, E) = \Gamma\big(M, E\otimes\Lambda^p T^*M\big).$$

The operator d_{A_f} maps $\Omega^p(M,E) \to \Omega^{p+1}(M,E)$ and, because $F(A_f)=0$, it squares to zero. So $\Omega^0(M,E) \to \Omega^1(M,E) \to \Omega^2(M,E) \to \Omega^3(M,E)$ is a complex, the de Rham complex twisted by the flat bundle E, with cohomology $H^p(M;E)$. For $A_f = 0$ it is the ordinary de Rham complex tensored with the Lie algebra.

12.13.5 The operator D and its square

Nash packages the bosonic fields into $v = (a, \phi)$, a 1-form and a function (or 3-form), and writes the quadratic part of the bosonic action as $\langle v, Dv\rangle$ for a single operator D. The inner product on E-valued forms is

$$\langle \alpha, \beta\rangle = -\int_M \mathrm{tr}(\alpha \wedge \star\beta), \tag{12.229}$$

which is positive definite because $-\mathrm{tr}(X^2) = \mathrm{tr}(X^\dagger X) \geq 0$ for an anti-Hermitian matrix X.

Working definition 12.36 (The operator D, Nash (12.102)–(12.103))

On $\Omega^1(M,E) \oplus \Omega^0(M,E)$ define

$$D = \begin{pmatrix} \star\mathrm{d}_{A_f} & \mathrm{d}_{A_f} \\ \mathrm{d}^\dagger_{A_f} & 0 \end{pmatrix}, \qquad D\begin{pmatrix} a \\ \phi \end{pmatrix} = \begin{pmatrix} \star\mathrm{d}_{A_f} a + \mathrm{d}_{A_f}\phi \\ \mathrm{d}^\dagger_{A_f} a \end{pmatrix}. \tag{12.230}$$

Here $\mathrm{d}^\dagger_{A_f}$ is the adjoint of d_{A_f} [taken with the inner product (12.229). Setting $\mathrm{d}^\dagger_{A_f} a = 0$ is the covariant Lorenz gauge]. It is first order and self-adjoint: $\star\mathrm{d}_{A_f}$ is self-adjoint on 1-forms in three dimensions (this is the curl), and the off-diagonal entries are adjoints of each other. Replacing the function ϕ by the 3-form $\star\phi$ gives Nash's version on $\Omega^1 \oplus \Omega^3$. D is (a twisted form of) the **odd signature operator** [its eta invariant is what fixes the phase of the one-loop answer below].

The quadratic form:

$$\begin{aligned} \langle v, Dv\rangle &\overset{(1)}{=} \langle a, \star\mathrm{d}_{A_f} a\rangle + \langle a, \mathrm{d}_{A_f}\phi\rangle + \langle \phi, \mathrm{d}^\dagger_{A_f} a\rangle \\ &\overset{(2)}{=} -\int_M \mathrm{tr}\big(a \wedge \star\star\, \mathrm{d}_{A_f} a\big) + 2\langle \phi, \mathrm{d}^\dagger_{A_f} a\rangle \\ &\overset{(3)}{=} -\int_M \mathrm{tr}\big(a \wedge \mathrm{d}_{A_f} a\big) - 2\int_M \mathrm{tr}\big(\phi \wedge \star\mathrm{d}^\dagger_{A_f} a\big). \end{aligned} \tag{12.231}$$

(1) Expand with (12.230).
(2) Definition (12.229) for the first term; the second equals the third by the definition of the adjoint.
(3) $\star\star = 1$ on every form in a three-dimensional Riemannian manifold, and (12.229) again.

This is (minus) the bosonic exponent of (12.228), Nash's (12.104). Likewise the ghost term is $-\langle \bar c, \mathrm{d}^\dagger_{A_f}\mathrm{d}_{A_f} c\rangle$.

Watch out: Signs in D

Nash writes both off-diagonal entries of D as "$\mathrm{d}_{A_f}\star$". Read literally (the composite of $\star$ and d), the two entries are *minus* each other's adjoints: for $A_f = 0$ on $\mathbb{R}^3$ they are $\nabla\phi$ and $\nabla\cdot\vec a$, and $\int \vec a\cdot\nabla\phi = -\int \phi\,\nabla\cdot\vec a$. Then $\langle v, Dv\rangle$ would lose its gauge-fixing term altogether. The intended operator is the self-adjoint one (12.230), which uses $\mathrm{d}^\dagger = -\star\,\mathrm{d}\star$ in one entry; this is what makes the cross term come out doubled, as in his (12.104).

Watch out: A missing star in the ghost term

Nash's (12.101) and (12.104) write the ghost term as $\mathrm{tr}(\mathrm{d}_{A_f}\bar{c}\wedge\mathrm{d}_{A_f}c)$. That is a 2-form, which cannot be integrated over a three-manifold. The intended term is $\mathrm{tr}(\mathrm{d}_{A_f}\bar{c}\wedge\star\mathrm{d}_{A_f}c)$, the covariant version of $\int\nabla\bar{c}\cdot\nabla c\,\mathrm{d}^3x$, as in (12.228). With the star it equals $-\langle\bar{c},\mathrm{d}^\dagger_{A_f}\mathrm{d}_{A_f}c\rangle$ after one integration by parts, which is what (12.104)–(12.105) use.

Worked example 12.89: D as a 4×4 matrix

Take $A_f=0$ and one Lie-algebra direction, on $\mathbb{R}^3$ (or a large torus). In Fourier space $\mathrm{d}\phi\to ip\,\tilde{\phi}$, $\mathrm{d}^\dagger a=-\nabla\cdot\vec{a}\to -ip\cdot\tilde{a}$, and $\star\mathrm{d}a\to P(p)\tilde{a}$ with the curl matrix (12.198). For $\vec{p}=p\,\hat{z}$, in the basis (a_1,a_2,a_3,ϕ),

$$D(p\hat{z})=p\begin{pmatrix}0&-i&0&0\\ i&0&0&0\\ 0&0&0&i\\ 0&0&-i&0\end{pmatrix}=p\,\big(\sigma_2\oplus(-\sigma_2)\big). \tag{12.232}$$

Compare the curl alone, (12.199), which had a zero eigenvalue on the longitudinal mode a_3. In D the longitudinal mode is paired with the Lagrange multiplier ϕ by the block $-p\sigma_2$, and the zero eigenvalue is gone: *the gauge-fixed operator is invertible*. Its eigenvalues are $+p,-p$ (transverse helicities) and $+p,-p$ (the longitudinal–multiplier pair). And

$$D(p\hat{z})^2=p^2\,\mathbb{1}_4\,:$$

D^2 is the Laplacian p^2 on each of the four components, three from the 1-form and one from the function.

Try it: Helicity eigenvectors

Find the eigenvectors of the transverse block $p\sigma_2$ of (12.232). (Answer: $(1,\pm i)/\sqrt{2}$ with eigenvalues $\pm p$; as 3-vectors $(1,\pm i,0)/\sqrt{2}$, the circular polarisations of a photon moving along $\hat{z}$.) Check also $D(p\hat{z})^2=p^2\mathbb{1}_4$ from $\sigma_2^2=\mathbb{1}$.

D^2 in general, Nash (12.111).

$$\begin{aligned}D^2&\overset{(1)}{=}\begin{pmatrix}\star\mathrm{d}_{A_f}\star\mathrm{d}_{A_f}+\mathrm{d}_{A_f}\mathrm{d}^\dagger_{A_f}&\star\mathrm{d}_{A_f}\mathrm{d}_{A_f}\\ \mathrm{d}^\dagger_{A_f}\star\mathrm{d}_{A_f}&\mathrm{d}^\dagger_{A_f}\mathrm{d}_{A_f}\end{pmatrix}\\ &\overset{(2)}{=}\begin{pmatrix}\mathrm{d}^\dagger_{A_f}\mathrm{d}_{A_f}+\mathrm{d}_{A_f}\mathrm{d}^\dagger_{A_f}&0\\ 0&\mathrm{d}^\dagger_{A_f}\mathrm{d}_{A_f}\end{pmatrix}=\begin{pmatrix}\Delta_1^E&0\\ 0&\Delta_0^E\end{pmatrix}.\end{aligned} \tag{12.233}$$

(1) Multiply the block matrices (12.230).

(2) On 2-forms in three dimensions $\mathrm{d}^\dagger=\star\mathrm{d}\star$, so $\star\mathrm{d}_{A_f}\star\mathrm{d}_{A_f}=\mathrm{d}^\dagger_{A_f}\mathrm{d}_{A_f}$ on 1-forms. The off-diagonal blocks contain $\mathrm{d}_{A_f}\mathrm{d}_{A_f}=[F(A_f),\cdot\,]=0$ *because A_f is flat* (for the lower one, $\mathrm{d}^\dagger_{A_f}\star\mathrm{d}_{A_f}=\pm\star\mathrm{d}_{A_f}\mathrm{d}_{A_f}$).

So D is a square root of the Laplacian on $\Omega^1\oplus\Omega^0$, a Dirac-type operator. Nash writes Δ_3^E in place of Δ_0^E, which is the same operator after the Hodge star identifies 3-forms with functions. Taking determinants,

$$(\det D)^2=\det\Delta_1^E\cdot\det\Delta_0^E. \tag{12.234}$$

12.13.6 Gaussian integrals as determinants: zeta and eta

The formal answer, Nash (12.105)–(12.106). With $\mu = ik/4\pi$ the fluctuation integral is

$$\int \mathcal{D}v\,\mathcal{D}c\,\mathcal{D}\bar{c}\;\exp\!\big[\mu\big(\langle v, Dv\rangle + \langle \bar{c}, \mathrm{d}_{A_f}^{\dagger}\mathrm{d}_{A_f} c\rangle\big)\big].$$

The bosonic integral is a (Fresnel) Gaussian and gives $\det(\mu D)^{-1/2}$; the ghost integral gives $\det(\mu\Delta_0^E)$ with no square root, because c and $\bar{c}$ are independent (as for complex variables). Formally, therefore,

$$\int \mathcal{D}v\,\mathcal{D}c\,\mathcal{D}\bar{c}\;e^{\mu(\dots)} = C\,\frac{\det\!\big(\mathrm{d}_{A_f}^{\dagger}\mathrm{d}_{A_f}\big)}{\sqrt{\det D}}, \tag{12.235}$$

where the constant C collects the μ-dependence and the phase. Both determinants are infinite products and need to be defined.

Working definition 12.37 (Zeta-function determinant)

For a positive elliptic operator A [an operator such as Δ_0^E on the compact M, whose eigenvalues are discrete and grow without bound, so their product diverges] with eigenvalues $\lambda_j > 0$, define

$$\zeta_A(s) = \sum_j \lambda_j^{-s}, \qquad \det A := e^{-\zeta_A'(0)}. \tag{12.236}$$

The sum converges for $\operatorname{Re} s$ large and is continued analytically [there is exactly one function, analytic except at isolated poles, that equals the sum wherever the sum converges] to $s = 0$. *Why this definition:* for a finite matrix, $-\zeta'(0) = \sum_j \ln\lambda_j = \ln\prod_j \lambda_j$, since $\frac{\mathrm{d}}{\mathrm{d}s}\lambda^{-s} = -\lambda^{-s}\ln\lambda$.

The scaling rule Nash quotes follows at once:

$$\begin{aligned}\zeta_{\mu A}(s) = \mu^{-s}\zeta_A(s) \;&\Longrightarrow\; \zeta_{\mu A}'(0) \overset{(1)}{=} -\ln\mu\;\zeta_A(0) + \zeta_A'(0)\\ &\Longrightarrow\; \det(\mu A) \overset{(2)}{=} \mu^{\zeta_A(0)}\det A.\end{aligned} \tag{12.237}$$

(1) Product rule at $s = 0$, with $\frac{\mathrm{d}}{\mathrm{d}s}\mu^{-s} = -\mu^{-s}\ln\mu$.
(2) Exponentiate $-\zeta_{\mu A}'(0)$.

For an $n\times n$ matrix, $\zeta(0) = n$ and this is $\det(\mu A) = \mu^n \det A$. For an operator $\zeta(0)$ is a finite number, the "regularised dimension". For positive μ this is harmless and $\mu^{\zeta(0)}$ goes into C. For $\mu = ik/4\pi$ it is not: $i^{\zeta(0)}$ has no meaning until we say how the phases of the individual eigenvalues are to be combined.

Worked example 12.90: A determinant on a circle with a flat bundle

Let $M = S^1$ of length 2π and E the flat U(1) bundle with holonomy $e^{2\pi i\alpha}$, $0 < \alpha < 1$. Its sections are functions with $\psi(x + 2\pi) = e^{2\pi i\alpha}\psi(x)$; a basis is $e^{i(n+\alpha)x}$, $n \in \mathbb{Z}$, and the Laplacian $\Delta_0^E = -\mathrm{d}^2/\mathrm{d}x^2$ has eigenvalues $(n+\alpha)^2$. Using the Hurwitz zeta function $\zeta_H(s,a) = \sum_{m\ge 0}(m+a)^{-s}$, with the classical values $\zeta_H(0,a) = \frac{1}{2} - a$ and $\partial_s\zeta_H(0,a) = \ln\Gamma(a) - \frac{1}{2}\ln 2\pi$:

$$\zeta_\Delta(s) \overset{(1)}{=} \sum_{n\ge 0}(n+\alpha)^{-2s} + \sum_{m\ge 0}(m+1-\alpha)^{-2s} = \zeta_H(2s,\alpha) + \zeta_H(2s,1-\alpha),$$

$$\zeta'_\Delta(0) \overset{(2)}{=} 2\big[\ln\Gamma(\alpha)+\ln\Gamma(1-\alpha)-\ln 2\pi\big] \overset{(3)}{=} 2\ln\frac{\pi}{2\pi\sin\pi\alpha} = -2\ln\big(2\sin\pi\alpha\big),$$
$$\det\Delta_0^E \overset{(4)}{=} \big(2\sin\pi\alpha\big)^2 = \big|1-e^{2\pi i\alpha}\big|^2.$$

(1) Split $n\geq 0$ and $n=-m-1<0$; then $|n+\alpha| = m+1-\alpha$.
(2) Differentiate: the chain rule gives the factor 2; use the value of $\partial_s\zeta_H(0,a)$.
(3) The reflection formula $\Gamma(\alpha)\Gamma(1-\alpha)=\pi/\sin\pi\alpha$.
(4) $\det = e^{-\zeta'(0)}$; and $\big|1-e^{i\theta}\big|^2 = 2-2\cos\theta = 4\sin^2(\theta/2)$.

Three lessons. The regularised infinite product is a finite, explicit function of the holonomy. As $\alpha\to 0$ it vanishes: the constant function becomes a zero mode, which is the zero-mode problem Nash mentions at the end. And $\zeta_\Delta(0) = (\frac12-\alpha)+(\frac12-(1-\alpha)) = 0$, so by (12.237) the determinant does not change if we rescale the circle (all eigenvalues scale by the same factor): it does not depend on the metric.

The phase: the eta invariant. For a real Gaussian with eigenvalue λ, (12.190) gives a phase $e^{i\pi\,\mathrm{sgn}(\lambda)/4}$ (with $e^{+i\lambda x^2/2}$ in the exponent). For infinitely many eigenvalues of both signs the total phase is $\exp\big(\frac{i\pi}{4}\sum_j \mathrm{sgn}\,\lambda_j\big)$, a divergent sum which is regularised in the same spirit as the determinant:

$$\eta_D(s) = \sum_j \mathrm{sgn}(\lambda_j)\,|\lambda_j|^{-s}, \qquad \text{phase} = \exp\Big(\frac{i\pi}{4}\,\eta_D(0)\Big) \quad \text{for } e^{+i\langle v,Dv\rangle/2}. \tag{12.238}$$

$\eta_D(0)$, the **eta invariant** of Atiyah, Patodi and Singer, measures the *spectral asymmetry* of D: how many more positive than negative eigenvalues it has. For the matrix (12.232) the spectrum is symmetric and $\eta=0$; on a curved M, or with a non-trivial flat bundle, it is not.

Worked example 12.91: An eta invariant on the circle

The first-order operator $-i\,\mathrm{d}/\mathrm{d}x$ on the same twisted circle has eigenvalues $n+\alpha$, $n\in\mathbb{Z}$: positive for $n\geq 0$, negative for $n<0$. Then

$$\eta(s) \overset{(1)}{=} \sum_{n\geq 0}(n+\alpha)^{-s} - \sum_{m\geq 0}(m+1-\alpha)^{-s} = \zeta_H(s,\alpha)-\zeta_H(s,1-\alpha),$$
$$\eta(0) \overset{(2)}{=} \big(\tfrac12-\alpha\big)-\big(\tfrac12-(1-\alpha)\big) = 1-2\alpha.$$

(1) Positive and negative eigenvalues separately, as in Example 12.90.
(2) $\zeta_H(0,a)=\frac12-a$.

At $\alpha=\frac12$ the spectrum $\{\pm\frac12,\pm\frac32,\dots\}$ is symmetric and $\eta=0$. As $\alpha\to 0^+$, $\eta\to 1$: the eigenvalue α is about to cross zero, and η jumps by 2 when it does. So the phase of a Chern–Simons one-loop integral depends on the flat connection through such spectral asymmetries.

Nash (following Witten) absorbs the μ-dependence and this phase into the constant C, noting that the phase is of the form $\exp[i\pi\eta(0)/2]$. The precise coefficient depends on conventions (the sign of the exponent, whether η is taken for D or for the operator on all odd forms); the structure is what matters: a modulus given by determinants, and a phase given by the eta invariant. Witten showed that the eta-invariant phase, combined with a gravitational Chern–Simons term needed to cancel its metric dependence, is what shifts k to $k+2$ in the exact SU(2) answers. The shift is derived from the eta invariant in Section T5.7.2.

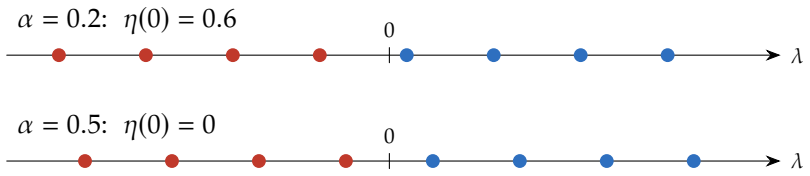


Figure 12.62: Spectral asymmetry. The eigenvalues $n+\alpha$ of $-i\,\mathrm{d}/\mathrm{d}x$ on a circle with holonomy $e^{2\pi i\alpha}$: positive ones blue, negative ones red. For $\alpha = 0.2$ the whole lattice is shifted to the right, so there are "more" positive eigenvalues and $\eta(0) = 1-2\alpha = 0.6$. For $\alpha = \frac{1}{2}$ the spectrum is symmetric about 0 and $\eta(0) = 0$. As $\alpha \to 0$ the eigenvalue α approaches zero, which is where the determinant $\left|1-e^{2\pi i\alpha}\right|^2$ vanishes.

The result, Nash (12.107).

$$Z(M) \approx \sum_{A_f} C\, e^{-ikS(A_f)/4\pi}\, \frac{\det\!\left(\mathrm{d}^\dagger_{A_f}\mathrm{d}_{A_f}\right)}{\sqrt{\det D}}. \tag{12.239}$$

This is the one-loop (first semiclassical) approximation. Now we look at the ratio of determinants.

12.13.7 The Ray–Singer analytic torsion

Working definition 12.38 (Analytic torsion, Nash (12.108))

For a closed oriented Riemannian n-manifold M and a flat bundle E [here the adjoint bundle, with the flat connection A_f] with $H^p(M;E) = 0$ for all p [every d_{A_f}-closed E-valued form is d_{A_f}-exact, so no Δ^E_p has a zero mode], let $\Delta^E_p = \mathrm{d}^\dagger_{A_f}\mathrm{d}_{A_f} + \mathrm{d}_{A_f}\mathrm{d}^\dagger_{A_f}$ on $\Omega^p(M,E)$. The **analytic torsion** $T(M,E)$ is defined by

$$\ln T(M,E) = \sum_{p=0}^{n} (-1)^p\, p\, \ln\det\Delta^E_p. \tag{12.240}$$

(Ray and Singer include a factor $\frac{1}{2}$; Nash omits it, and we follow Nash.) *Recipe:* compute the zeta-determinant of the Laplacian in each degree and combine them with the weights $(-1)^p p$.

Three dimensions, Nash (12.109)–(12.110). For $n = 3$ the weights are $0, -1, +2, -3$, so

$$T(M,E) = \left(\det\Delta^E_1\right)^{-1}\left(\det\Delta^E_2\right)^{2}\left(\det\Delta^E_3\right)^{-3}. \tag{12.241}$$

Now use Hodge duality. The star $\star : \Omega^p \to \Omega^{n-p}$ is an isometry and commutes with the Laplacian, $\star\Delta_p = \Delta_{n-p}\star$; hence if $\Delta_p\omega = \lambda\omega$ then $\Delta_{n-p}(\star\omega) = \lambda\,\star\,\omega$, the spectra agree, and

$$\det\Delta^E_p = \det\Delta^E_{n-p}. \tag{12.242}$$

With $n = 3$: $\det\Delta_2 = \det\Delta_1$ and $\det\Delta_3 = \det\Delta_0$. Therefore

$$T(M,E) \overset{(1)}{=} \left(\det\Delta^E_1\right)^{-1}\left(\det\Delta^E_1\right)^{2}\left(\det\Delta^E_0\right)^{-3} \overset{(2)}{=} \left(\det\Delta^E_0\right)^{-3}\det\Delta^E_1, \tag{12.243}$$

(1) Formula (12.241) with (12.242).
(2) Combine the powers of $\det\Delta_1$: $-1+2 = 1$.

which is Nash's (12.110).

The Chern–Simons ratio is a power of the torsion, Nash (12.112).

$$\frac{\det\left(\mathrm{d}_{A_f}^{\dagger}\mathrm{d}_{A_f}\right)}{\sqrt{\det D}} \overset{(1)}{=} \frac{\det\Delta_0^E}{\left(\det\Delta_1^E\,\det\Delta_0^E\right)^{1/4}} \overset{(2)}{=} \frac{\left(\det\Delta_0^E\right)^{3/4}}{\left(\det\Delta_1^E\right)^{1/4}} \overset{(3)}{=} T(M,E)^{-1/4}. \tag{12.244}$$

(1) The ghost operator is $\mathrm{d}_{A_f}^{\dagger}\mathrm{d}_{A_f} = \Delta_0^E$ on functions (there $\mathrm{d}^{\dagger}$ of a function is zero); and $\sqrt{\det D} = \left((\det D)^2\right)^{1/4}$ with (12.234), up to the phase already put into C.

(2) $\det\Delta_0 \big/ (\det\Delta_0)^{1/4} = (\det\Delta_0)^{3/4}$.

(3) Raise (12.243) to the power $-\frac{1}{4}$: $T^{-1/4} = (\det\Delta_0)^{3/4}(\det\Delta_1)^{-1/4}$.

Hence Nash's (12.113):

$$Z(M) \approx \sum_E C\, e^{-ikS(E)/4\pi}\, T(M,E)^{-1/4}, \tag{12.245}$$

the sum running over flat bundles E (flat connections up to gauge equivalence), $S(E)$ being the Chern–Simons invariant of the flat connection.

Key idea: Why this is evidence of topological invariance

Each factor in (12.245) is topological. $S(A_f)$ is the Chern–Simons invariant of a flat connection: it is unchanged by small deformations (its variation is $2\int \mathrm{tr}(\delta A \wedge F) = 0$ at a flat connection) and depends only on the flat bundle. And the analytic torsion, although built from the metric-dependent Laplacians, is *independent of the metric*: this is the theorem of Ray and Singer (for acyclic E). Schwarz observed that exactly this combination appears in the quadratic Chern–Simons integral, and established the relation explicitly in the abelian case, where the action is exactly quadratic and the one-loop formula is exact. The metric we introduced to fix the gauge has dropped out, as a topological theory requires. (The remaining metric dependence sits in the eta-invariant phase inside C, and is cancelled as described above.)

Worked example 12.92: Torsion of a circle

For $M = S^1$ ($n = 1$) the formula (12.240) has one term, $\ln T = -\ln\det\Delta_1^E$. On a circle the 1-forms $f\,\mathrm{d}x$ have the same spectrum as the functions f, so by Example 12.90,

$$T(S^1,E) = \left|1 - e^{2\pi i\alpha}\right|^{-2}.$$

It depends only on the holonomy, not on the length of the circle (the metric), as promised. Up to the power (fixed by Ray and Singer's factor $\frac{1}{2}$ and by orientation conventions) this is $|1-u|^{\pm 1}$ for holonomy u, which is the combinatorial (Reidemeister) torsion of the circle computed from its cell structure: one 0-cell e_0 and one 1-cell e_1, with twisted boundary map $\partial e_1 = (u-1)e_0$. Ray and Singer conjectured, and Cheeger and Müller proved, that analytic torsion equals Reidemeister torsion in general.

Physical picture: The Alexander polynomial is a torsion too

Apply the same combinatorial recipe to the complement of a knot in S^3, with the flat U(1) bundle whose holonomy round a meridian is t. The cell structure comes from a Wirtinger presentation, the boundary maps are the Fox derivatives of Example 12.71, and the torsion is (Milnor)

$$\tau\left(S^3 \setminus K; t\right) = \frac{\Delta_K(t)}{t-1}.$$

So the determinants of the Chern–Simons one-loop integral and the Alexander polynomial of Section 12.11.2 are the same kind of object, a regularised determinant of a twisted boundary (or Laplace) operator, and the Alexander polynomial is the torsion of the knot complement with a flat U(1) bundle. The circle example above is the case of the unknot, whose complement is homotopic to a circle: $\tau = 1/(t-1)$, $\Delta = 1$.

Zero modes. Nash ends with a caveat. By Hodge theory, the kernel of Δ_p^E is isomorphic to the twisted cohomology $H^p(M;E)$. If some $H^p(M;E) \neq 0$ then Δ_p^E has zero eigenvalues and $\det \Delta_p^E = 0$, and the formulas above break down (as in Example 12.90 at $\alpha \to 0$). Physically: $H^0(M;E) \neq 0$ means the flat connection has a non-trivial stabiliser (constant gauge transformations that leave it fixed), and $H^1(M;E) \neq 0$ means it can be deformed through flat connections (a moduli space, as for $S^1 \times S^2$ in Example 12.88). Both happen for the trivial connection on S^3, where $H^0 = H^3 = \mathfrak{g}$. Nash restricts to acyclic E; the general case requires integrating over the moduli space and dividing by the stabiliser volume. Each zero mode also changes the power of k in the stationary phase formula, one factor $k^{-1/2}$ per missing Gaussian direction (Example 12.87).

Worked example 12.93: The trivial connection on S^3 and the exact answer

For $G = \mathrm{SU}(2)$ on S^3 the only flat connection is trivial, and its stabiliser is all of SU(2), three dimensions of constant gauge transformations. The zero modes of Δ_0 are the three constant Lie-algebra-valued functions, which are gauge directions we may not integrate over as if they were fluctuations. Treating them correctly leaves $Z(S^3) \propto k^{-3/2}$ for large k, one $k^{-1/2}$ per stabiliser dimension. Witten's exact answer, derived with the conformal field theory methods of Chapter 11 as applied to Chern–Simons theory later in this chapter, is

$$Z(S^3) = \sqrt{\frac{2}{k+2}}\,\sin\frac{\pi}{k+2} \;\approx\; \sqrt{2}\,\pi\,k^{-3/2} \qquad (k \to \infty),$$

which indeed decays like $k^{-3/2}$. The weak-coupling analysis correctly predicts the leading behaviour of the exact result, and the shift $k \to k+2$ is the eta-invariant effect mentioned above.

Reading Nash (pp. 346–350, (12.95)–(12.113))

Nash introduces the coupling (12.95), rescales it into the level (12.96) so that small coupling means large k, and argues by stationary phase that $Z(M)$ (12.97) is dominated by flat connections (12.98). He expands the action about a flat connection (12.99), fixes the gauge $\mathrm{d}_{A_f}^* a = 0$ (12.100) with a multiplier and ghosts (12.101), packages the bosons into the operator D (12.102)–(12.104), rewrites Z (12.105), evaluates the Gaussian as a ratio of zeta-function determinants with the μ-dependence and an eta-invariant phase absorbed into C (12.106)–(12.107), and identifies the ratio as $T(M,E)^{-1/4}$ via the Ray–Singer torsion (12.108)–(12.113), metric independent by Schwarz. He assumes all $H^p(M;E)$ vanish. We have added: the stationary-phase formula and the Bessel example; the proof that $\delta S = 2\int \mathrm{tr}(\delta A \wedge F)$ and the exact expansion; flat connections as representations of π_1 with examples; Faddeev–Popov in two dimensions; Grassmann integration; the explicit 4×4 matrix of D; the computation of D^2 and where flatness enters; zeta and eta functions on a twisted circle; the torsion of a circle and its relation to the Alexander polynomial; the zero-mode count and the comparison with $Z(S^3)$. Note the sign issue in Nash's D (pitfall above).

Recap

- The coupling appears only as k/g^2; weak coupling is large k, with $\hbar_{\text{eff}} = 4\pi/k$.
- $\delta S_{\text{CS}} = 2\int \operatorname{tr}(\delta A \wedge F)$: the stationary points are flat connections, i.e. homomorphisms $\pi_1(M) \to G$. $S(A_f + a) = S(A_f) + \int \operatorname{tr}(a \wedge \mathrm{d}_{A_f} a) + \frac{2}{3}\int \operatorname{tr} a^3$.
- Gauge fixing $\mathrm{d}^\dagger_{A_f} a = 0$: a Lagrange multiplier ϕ and ghosts $c, \bar{c}$ (a Grassmann Gaussian gives det, not $\det^{-1/2}$). Ghost operator Δ_0^E.
- $D = \left(\begin{smallmatrix} \star\mathrm{d} & \mathrm{d} \\ \mathrm{d}^\dagger & 0 \end{smallmatrix}\right)$ is self-adjoint, invertible after gauge fixing, and $D^2 = \Delta_1^E \oplus \Delta_0^E$ because $F(A_f) = 0$.
- Determinants by zeta functions, $\det A = e^{-\zeta'(0)}$, $\det(\mu A) = \mu^{\zeta(0)} \det A$; phases by the eta invariant. Twisted circle: $\det \Delta = \left|1 - e^{2\pi i \alpha}\right|^2$, $\eta = 1 - 2\alpha$.
- $\det \Delta_0 / \sqrt{\det D} = T(M,E)^{-1/4}$ with $T = (\det \Delta_0)^{-3} \det \Delta_1$ in three dimensions, so $Z(M) \approx \sum_E C\, e^{-ikS(E)/4\pi} T(M,E)^{-1/4}$: metric independent, as a topological theory must be.

12.14 Chern–Simons theory and the Jones polynomial

In the previous sections we met two faces of Chern–Simons theory. With the abelian group U(1) the path integral is Gaussian, and a pair of Wilson loops gives back the Gauss linking number. With a non-abelian group and a weak coupling, the path integral localises on flat connections and produces the Ray–Singer torsion. Neither computation tells us what the *exact* non-abelian answer is. For $G = \mathrm{SU}(2)$ we would like to know

$$\left\langle W(R_1, C_1) \cdots W(R_p, C_p) \right\rangle = ?$$

at finite level k, for an arbitrary link $C_1 \cup \cdots \cup C_p$ in S^3. The integrand is not Gaussian, so we cannot just complete the square.

Witten's way out is a piece of "cut and paste" physics. He never evaluates the path integral directly. Instead he

(i) cuts the 3-manifold along surfaces Σ;

(ii) quantises the theory on $\Sigma \times \mathbb{R}$, obtaining a Hilbert space $\mathcal{H}_\Sigma$ that turns out to be *finite-dimensional*;

(iii) recognises $\mathcal{H}_\Sigma$ as the space of conformal blocks of a two-dimensional conformal field theory (the WZW model of Chapter 11), which is already fully understood;

(iv) glues the pieces back together, using only linear algebra in these small vector spaces.

The payoff is spectacular. Step (iv), applied to a 2-dimensional Hilbert space, produces the skein relation of the Jones polynomial with no knot diagrams in sight. Section 12.15 pushes the same idea further: cutting out solid tori and regluing them (surgery) computes invariants of 3-manifolds themselves. Figure 12.63 shows the logic of this part.

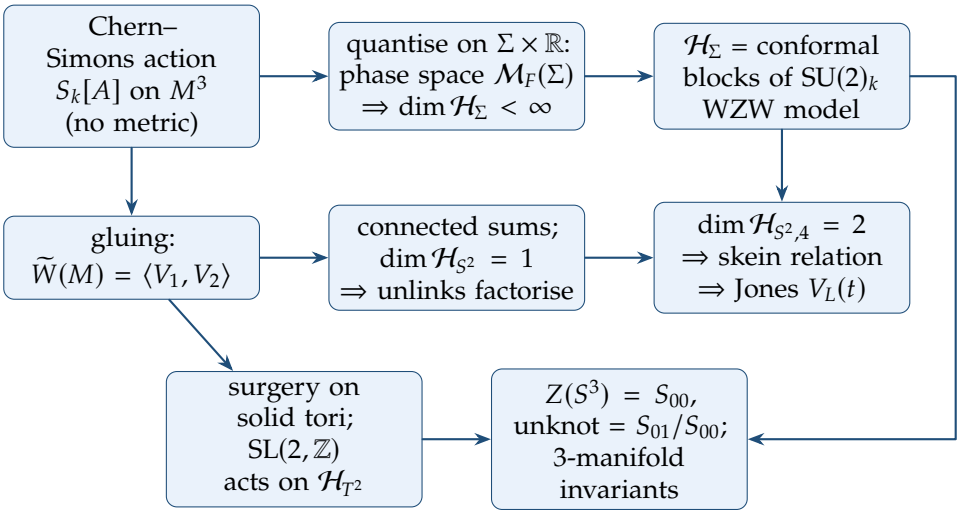


Figure 12.63: The logic of Nash's §6–§7. Read the top row first: the quantisation of Chern–Simons theory on a surface gives finite-dimensional Hilbert spaces, identified with conformal blocks. The middle row uses only gluing and dimension counting to get the skein relation. The bottom row replaces spheres by tori and gets 3-manifold invariants.

12.14.1 Conventions for this part

We keep the conventions of this chapter. The gauge field A is a $\mathfrak{su}(2)$-valued 1-form with anti-Hermitian values, $F = \mathrm{d}A + A \wedge A$, and gauge transformations act as $A \mapsto g^{-1}Ag + g^{-1}\mathrm{d}g$. The Chern–Simons function is

$$f(A) \;=\; -\frac{1}{8\pi^2}\int_M \mathrm{tr}\Big(A \wedge \mathrm{d}A + \tfrac{2}{3}A \wedge A \wedge A\Big), \tag{12.246}$$

which jumps by the integer $\deg g$ under a large gauge transformation. The Chern–Simons **action at level** k is

$$S_k[A] \;=\; 2\pi k\, f(A) \;=\; -\frac{k}{4\pi}\int_M \mathrm{tr}\Big(A \wedge \mathrm{d}A + \tfrac{2}{3}A \wedge A \wedge A\Big), \qquad k \in \mathbb{Z}, \tag{12.247}$$

and the path integral weight is $e^{iS_k[A]}$, which is gauge invariant because $e^{2\pi i k \deg g} = 1$.

Dictionary: anti-Hermitian versus Hermitian gauge fields

Physicists write $A = -iA^a_{\text{phys}}T^a$ with Hermitian generators $T^a = \sigma^a/2$, so $\mathrm{tr}(T^aT^b) = \frac{1}{2}\delta^{ab}$. Then $\mathrm{tr}(A \wedge \mathrm{d}A) = -\frac{1}{2}A^a_{\text{phys}} \wedge \mathrm{d}A^a_{\text{phys}}$, and (12.247) becomes the familiar

$$S_k = \frac{k}{4\pi}\int \mathrm{tr}_{\text{phys}}\Big(A \wedge \mathrm{d}A - \tfrac{2i}{3}A \wedge A \wedge A\Big) = \frac{k}{8\pi}\int \Big(A^a \wedge \mathrm{d}A^a + \tfrac{1}{3}\varepsilon^{abc}A^a \wedge A^b \wedge A^c\Big).$$

For U(1) we simply write $S_k = \frac{k}{4\pi}\int A \wedge \mathrm{d}A$ with a real 1-form A, as in the abelian calculation of the previous section.

Two numbers will appear everywhere:

$$q \;:=\; \exp\Big(\frac{2\pi i}{k+2}\Big), \qquad [n]_q \;:=\; \frac{q^{n/2} - q^{-n/2}}{q^{1/2} - q^{-1/2}} \;=\; \frac{\sin\big(n\pi/(k+2)\big)}{\sin\big(\pi/(k+2)\big)}. \tag{12.248}$$

The **quantum integer** $[n]_q$ tends to the ordinary integer n as $k \to \infty$ (so $q \to 1$). For example $[2]_q = q^{1/2} + q^{-1/2} = 2\cos\big(\pi/(k+2)\big)$ and $[3]_q = q + 1 + q^{-1}$. The shift $k \to k+2$ is the famous quantum correction: 2 is the dual Coxeter number of SU(2) (for SU(N) it is N).

12.14.2 Why a surface, and why punctures

Suppose M is a 3-manifold with knots inside it, and pick a surface $\Sigma \subset M$. Near Σ the manifold looks like a slab $\Sigma \times (-\varepsilon, \varepsilon)$. Call the second factor "time". Then the knots are *worldlines*: each time a knot crosses Σ it leaves a mark, a point of Σ where a charged particle sits at that instant (Figure 12.64).

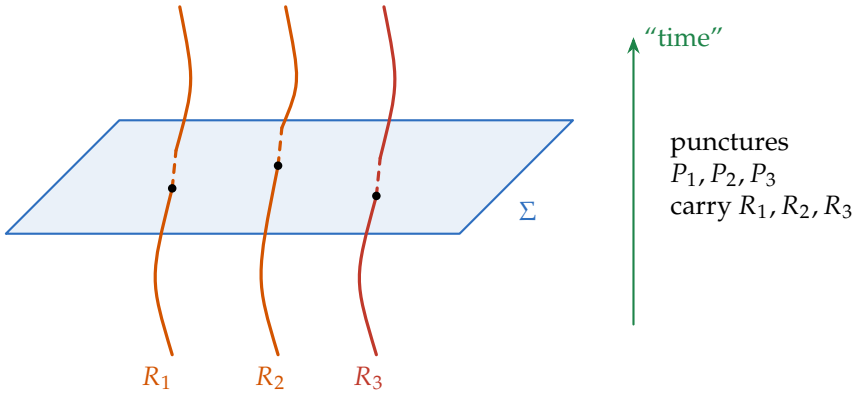


Figure 12.64: A surface Σ inside a 3-manifold, pierced by three Wilson lines. Near Σ the 3-manifold is $\Sigma \times$ time and the Wilson lines are worldlines of static charges. What we should quantise is Σ *together with* its marked points, each labelled by the representation of the line that passes through it.

So the natural objects are Riemann surfaces with marked points, each marked point labelled by a representation of G. These are exactly the data on which a conformal field theory is defined: a CFT assigns a space of conformal blocks $V(\Sigma_{p,n})$ to a genus-p surface with n punctures carrying primary fields (Chapter 11). This coincidence of data is the first hint that the two theories are secretly one.

Physical picture: Anyons

In $2+1$ dimensions a Wilson line is the worldline of a heavy charged particle. The Chern–Simons term attaches magnetic flux to charge, so moving one particle around another produces an Aharonov–Bohm-like phase, or, in the non-abelian case, a *matrix*. Particles whose exchange is a matrix rather than ± 1 are called non-abelian anyons. Knot invariants are the bookkeeping of these exchange matrices along the worldlines.

12.14.3 Canonical quantisation on $\Sigma \times \mathbb{R}$

We now quantise the theory on $M = \Sigma \times \mathbb{R}$ in the ordinary Hamiltonian way. Split the gauge field into a time component and a spatial part,

$$A = A_0\,\mathrm{d}t + \tilde{A}, \qquad \tilde{A} = A_1 \mathrm{d}x^1 + A_2 \mathrm{d}x^2,$$

and split the exterior derivative the same way, $\mathrm{d} = \mathrm{d}t\,\partial_t + \tilde{\mathrm{d}}$, where $\tilde{\mathrm{d}}$ differentiates along Σ only. Write $\tilde{F} = \tilde{\mathrm{d}}\tilde{A} + \tilde{A} \wedge \tilde{A}$ for the curvature of $\tilde{A}$ on the slice.

The action in $2+1$ form

Only the terms containing exactly one $\mathrm{d}t$ survive in a 3-form on $\Sigma \times \mathbb{R}$. Collecting them:

$$\operatorname{tr}(A \wedge \mathrm{d}A)\big|_{\mathrm{d}t} \overset{(1)}{=} \operatorname{tr}\big(A_0\,\mathrm{d}t \wedge \tilde{\mathrm{d}}\tilde{A} + \tilde{A} \wedge \tilde{\mathrm{d}}A_0 \wedge \mathrm{d}t + \tilde{A} \wedge \mathrm{d}t \wedge \partial_t \tilde{A}\big)$$

$$\overset{(2)}{=} \mathrm{d}t \wedge \mathrm{tr}\big(A_0\,\tilde{\mathrm{d}}\tilde{A} + \tilde{A}\wedge\tilde{\mathrm{d}}A_0 - \tilde{A}\wedge\partial_t\tilde{A}\big)$$

$$\overset{(3)}{=} \mathrm{d}t \wedge \mathrm{tr}\big(2A_0\,\tilde{\mathrm{d}}\tilde{A} - \tilde{A}\wedge\partial_t\tilde{A}\big) - \mathrm{d}t\wedge\tilde{\mathrm{d}}\ \mathrm{tr}(\tilde{A}A_0), \tag{12.249}$$

$$\tfrac{2}{3}\,\mathrm{tr}(A\wedge A\wedge A)\big|_{\mathrm{d}t} \overset{(4)}{=} \tfrac{2}{3}\cdot 3\,\mathrm{d}t\wedge\mathrm{tr}\big(A_0\,\tilde{A}\wedge\tilde{A}\big) = 2\,\mathrm{d}t\wedge\mathrm{tr}\big(A_0\,\tilde{A}\wedge\tilde{A}\big). \tag{12.250}$$

(1) $\mathrm{d}A = \tilde{\mathrm{d}}A_0\wedge\mathrm{d}t + \mathrm{d}t\wedge\partial_t\tilde{A} + \tilde{\mathrm{d}}\tilde{A}$; multiply by $A = A_0\mathrm{d}t + \tilde{A}$ and keep the terms with a single $\mathrm{d}t$.
(2) Move $\mathrm{d}t$ to the front: past a 2-form it costs nothing, past the 1-form $\tilde{A}$ it costs a sign.
(3) Leibniz: $\tilde{\mathrm{d}}\ \mathrm{tr}(\tilde{A}A_0) = \mathrm{tr}(\tilde{\mathrm{d}}\tilde{A}\,A_0) - \mathrm{tr}(\tilde{A}\wedge\tilde{\mathrm{d}}A_0)$, so the middle term equals $\mathrm{tr}(A_0\tilde{\mathrm{d}}\tilde{A})$ minus an exact form.
(4) The cube $(A_0\mathrm{d}t + \tilde{A})^3$ has three terms with one $\mathrm{d}t$; after moving $\mathrm{d}t$ to the front and using graded cyclicity of the trace, all three equal $\mathrm{d}t\wedge\mathrm{tr}(A_0\tilde{A}\wedge\tilde{A})$.

Adding the two pieces and dropping the exact term (it integrates to zero on a closed Σ), we find

$$S_k[A] \overset{(1)}{=} -\frac{k}{4\pi}\int \mathrm{d}t\int_\Sigma \mathrm{tr}\big(2A_0(\tilde{\mathrm{d}}\tilde{A} + \tilde{A}\wedge\tilde{A}) - \tilde{A}\wedge\partial_t\tilde{A}\big)$$

$$\overset{(2)}{=} \underbrace{\frac{k}{4\pi}\int \mathrm{d}t\int_\Sigma \mathrm{tr}\big(\tilde{A}\wedge\partial_t\tilde{A}\big)}_{\text{“}p\,\dot{q}\text{”}} \underbrace{-\frac{k}{2\pi}\int \mathrm{d}t\int_\Sigma \mathrm{tr}\big(A_0\,\tilde{F}\big)}_{\text{constraint}}. \tag{12.251}$$

(1) Insert (12.249) and (12.250) into (12.247), with orientation $\mathrm{d}t\wedge\mathrm{vol}_\Sigma$.
(2) Recognise $\tilde{\mathrm{d}}\tilde{A} + \tilde{A}\wedge\tilde{A} = \tilde{F}$.

Read (12.251) like a mechanics Lagrangian. There is a term of the form $p\,\dot{q}$, first order in time. There is a term linear in A_0, which has no time derivative at all: A_0 is a **Lagrange multiplier**. There is *no* term of the form $-H(p,q)$. So

Key idea: The Chern–Simons phase space

(a) Varying A_0 imposes the **Gauss law** $\tilde{F} = 0$: the spatial connection must be flat.

(b) The constraint generates gauge transformations on Σ, so physical configurations are flat connections modulo gauge,

$$\mathcal{M}_F(\Sigma) = \{\tilde{A}\text{ on }\Sigma:\ \tilde{F} = 0\}/\mathcal{G}.$$

(c) The Hamiltonian vanishes, $\mathcal{H} = 0$: nothing evolves, as befits a topological theory.

Choosing the gauge $A_0 = 0$ (Nash's choice) removes the multiplier; the Gauss law is then imposed on states, and we are left with the space $\mathcal{M}_F(\Sigma)$ of flat connections modulo gauge as phase space.

Worked example 12.94: Abelian Chern–Simons: the commutator of A_1 and A_2

Take $G = \mathrm{U}(1)$ and $S_k = \frac{k}{4\pi}\int A\wedge\mathrm{d}A$ with a real 1-form A. In components $A\wedge\mathrm{d}A = \varepsilon^{\mu\nu\rho}A_\mu\partial_\nu A_\rho\,\mathrm{d}^3x$, with $\varepsilon^{012} = 1$. The terms with a time derivative are

$$\frac{k}{4\pi}\varepsilon^{i0j}A_i\partial_0A_j \overset{(1)}{=} -\frac{k}{4\pi}(A_1\dot{A}_2 - A_2\dot{A}_1)$$

$$\overset{(2)}{=} \frac{k}{2\pi}A_2\dot{A}_1 - \frac{k}{4\pi}\partial_t(A_1A_2).$$

(1) $\varepsilon^{i0j} = -\varepsilon^{0ij}$, and ε^{0ij} is $+1$ for $(i,j) = (1,2)$.

(2) $-A_1\dot{A}_2 = A_2\dot{A}_1 - \partial_t(A_1A_2)$.

Dropping the total time derivative, the Lagrangian is $p\dot{q}$ with $q = A_1(x)$ and $p = \frac{k}{2\pi}A_2(x)$ at each point of Σ. Canonical quantisation $[q,p] = i$ then gives

$$\big[A_1(x), A_2(y)\big] = \frac{2\pi i}{k}\,\delta^2(x-y).$$

The two spatial components of the gauge field are conjugate to each other, like x and p of a particle. The effective Planck constant is $2\pi/k$: large k is the classical limit. This is the same algebra as the guiding-centre coordinates of an electron in the lowest Landau level, which is why Chern–Simons theory describes the quantum Hall effect.

For a non-abelian group the same computation, with tr in the dictionary normalisation, gives $[A_1^a(x), A_2^b(y)] = \frac{4\pi i}{k}\delta^{ab}\delta^2(x-y)$ (the extra factor 2 comes from $\mathrm{tr}(T^aT^b) = \frac{1}{2}\delta^{ab}$). On the phase space $\mathcal{M}_F(\Sigma)$ the corresponding **symplectic form** is

$$\omega(a,b) \;=\; -\frac{k}{2\pi}\int_\Sigma \mathrm{tr}(a\wedge b), \qquad a,b \in T_{[\tilde{A}]}\mathcal{M}_F \subset \Omega^1(\Sigma;\mathfrak{su}(2)), \tag{12.252}$$

the minus sign making it positive for anti-Hermitian fields. It is non-degenerate on $\mathcal{M}_F$ because the directions it would miss (the gauge directions) have been divided out; this is the Atiyah–Bott construction of the symplectic structure on the moduli space of flat connections.

Counting states: phase-space volume

The semiclassical rule of thumb is

$$\dim\mathcal{H} \;\approx\; \frac{\text{phase-space volume}}{(2\pi\hbar)^d} \;=\; \int_{\mathcal{M}_F}\frac{\omega^d}{d!\,(2\pi)^d}, \qquad 2d = \dim\mathcal{M}_F. \tag{12.253}$$

Since $\omega \propto k$ and $\mathcal{M}_F$ is *compact* (it is built from holonomies in the compact group G), the right-hand side is finite and grows like k^d. This is the reason the Hilbert spaces are finite-dimensional, in sharp contrast with ordinary field theories.

Worked example 12.95: U(1) Chern–Simons on a torus has k states

Let $\Sigma = T^2$ with periodic coordinates $x, y \in [0,1)$. A flat U(1) connection is gauge equivalent to a constant one, $\tilde{A} = \theta_1\mathrm{d}x + \theta_2\mathrm{d}y$, and large gauge transformations $e^{2\pi i(mx+ny)}$ shift $\theta_1 \to \theta_1 + 2\pi m$, $\theta_2 \to \theta_2 + 2\pi n$. So $\mathcal{M}_F$ is the torus of holonomies $(e^{i\theta_1}, e^{i\theta_2})$. The Lagrangian of the previous example restricted to these modes is

$$L \overset{(1)}{=} \frac{k}{2\pi}\int_{T^2} A_2\dot{A}_1\,\mathrm{d}x\,\mathrm{d}y = \frac{k}{2\pi}\,\theta_2\,\dot{\theta}_1,$$

$$\text{volume} \overset{(2)}{=} \int_0^{2\pi}\mathrm{d}\theta_1\int_0^{2\pi}\mathrm{d}p_1 = 2\pi\cdot\frac{k}{2\pi}\cdot 2\pi = 2\pi k,$$

$$\dim\mathcal{H}_{T^2} \overset{(3)}{=} \frac{2\pi k}{2\pi} = k.$$

(1) The fields are constant and the torus has unit area.
(2) The momentum is $p_1 = \frac{k}{2\pi}\theta_2$, which ranges over an interval of length $\frac{k}{2\pi}\cdot 2\pi = k$.
(3) One state per cell of area $2\pi\hbar = 2\pi$.

Here the semiclassical count is exact. The k wavefunctions are theta functions: in the representation where θ_1 is diagonal, periodicity of θ_1 quantises p_1 to integers, periodicity

of θ_2 (i.e. of p_1 modulo k) then leaves the k values $p_1 \in \mathbb{Z}_k$.

Worked example 12.96: SU(2) on a torus: the pillowcase and $k+1$ states

Flat SU(2) connections on T^2 are determined by two holonomies U_1, U_2 around the two cycles. Because $\pi_1(T^2) = \mathbb{Z}^2$ is abelian, they commute: $U_1 U_2 U_1^{-1} U_2^{-1} = 1$ (the holonomy round the boundary of the square). Commuting unitaries can be diagonalised together:

$$U_1 = \begin{pmatrix} e^{i\theta_1} & 0 \\ 0 & e^{-i\theta_1} \end{pmatrix}, \qquad U_2 = \begin{pmatrix} e^{i\theta_2} & 0 \\ 0 & e^{-i\theta_2} \end{pmatrix}.$$

The only remaining freedom is the Weyl element $\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$, which swaps the diagonal entries and sends $(\theta_1, \theta_2) \to (-\theta_1, -\theta_2)$. So $\mathcal{M}_F(T^2) = T^2/\mathbb{Z}_2$, a sphere with four corners, the "pillowcase" of Figure 12.65. The SU(2) theory restricted to the diagonal is a U(1) theory at level $2k$ (the trace of σ_3^2 is 2), which has $2k$ states labelled by $n \in \mathbb{Z}_{2k}$. The Weyl group identifies $n \sim -n$ and keeps only even combinations:

$$\dim \mathcal{H}_{T^2} \overset{(1)}{=} \underbrace{2}_{n=0,\,k} + \underbrace{\tfrac{2k-2}{2}}_{\text{pairs } \{n,-n\}} \overset{(2)}{=} k+1.$$

(1) $n = 0$ and $n = k$ are their own images under $n \to -n$ modulo $2k$; the other $2k-2$ values pair up.
(2) Arithmetic.

A more careful treatment (which includes the one-loop shift $k \to k+2$ and uses Weyl-*odd* theta functions at level $2(k+2)$) gives the same count, $(2(k+2)-2)/2 = k+1$, and identifies the wavefunctions with the numerators of the $SU(2)_k$ characters. Either way the answer is $k+1$, the number of integrable representations of SU(2) at level k (Section 12.14.6). This is our first glimpse of the dictionary with conformal field theory.

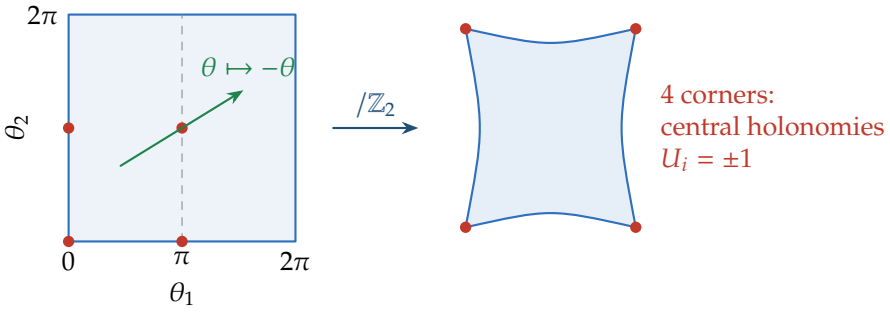


Figure 12.65: The moduli space of flat SU(2) connections on a torus. Left: the torus of angles (θ_1, θ_2) of the two commuting holonomies. The Weyl group acts by $\theta \mapsto -\theta$, whose fixed points are the four red dots. Right: the quotient is a "pillowcase", a sphere with four conical corners. Quantising it at level k gives $k+1$ states.

Worked example 12.97: The sphere: exactly one state

On $\Sigma = S^2$ every loop is contractible, $\pi_1(S^2) = 0$, so every flat connection has trivial holonomy and is gauge equivalent to $\tilde{A} = 0$. Hence $\mathcal{M}_F(S^2)$ is a single point. Quantising a

point gives a one-dimensional Hilbert space:

$$\dim \mathcal{H}_{S^2} = 1.$$

This humble fact, which Nash quotes from the CFT side as (11.75), is the engine behind the factorisation of connected sums in Section 12.14.9.

12.14.4 Kähler quantisation, line bundles and Wilson lines

Nash says that $\mathcal{M}_F$ is "a compact Kähler manifold" and that the Hilbert spaces are "the sections of the kth power of the determinant line bundle". Let us unpack both statements.

Geometric quantisation in one paragraph

To quantise a phase space (P, ω) that is not a cotangent bundle, one needs three things.

(a) A **prequantum line bundle** $L \to P$ with a connection whose curvature is $-i\omega$. Such an L exists only if $\omega/2\pi$ has integer periods: this is the Dirac quantisation condition in disguise, and for Chern–Simons theory it is the quantisation of the level k.

(b) A **polarisation**, i.e. a choice of "half of the coordinates" on which wavefunctions depend. On a Kähler manifold the natural choice is: wavefunctions are *holomorphic*.

(c) Then $\mathcal{H} = H^0(P, L)$, the holomorphic sections of L. For compact P this is finite-dimensional, and for large k its dimension is the volume (12.253) (Riemann–Roch).

For $P = \mathcal{M}_F(\Sigma)$ the Kähler structure comes from a complex structure on Σ: the Hodge star on 1-forms of a Riemann surface squares to -1, so $a \mapsto \star a$ is a complex structure J on $T\mathcal{M}_F$, compatible with ω. The line bundle is $\mathcal{L}^k$, where $\mathcal{L}$ is Quillen's **determinant line bundle** of the family of operators $\bar\partial_{\tilde A}$; its curvature is the basic symplectic form. So

$$\mathcal{H}_\Sigma \;=\; H^0\big(\mathcal{M}_F(\Sigma),\, \mathcal{L}^k\big). \tag{12.254}$$

Notice what has happened. To define $\mathcal{H}_\Sigma$ we chose a complex structure on Σ, although the theory is topological. The resolution is that the spaces for different complex structures form a vector bundle over the moduli space of Riemann surfaces with a projectively flat connection (Hitchin's connection, the Knizhnik–Zamolodchikov connection in CFT language), so they are canonically identified up to a phase. That phase is the **framing anomaly** that will reappear in Section 12.15.7.

Worked example 12.98: Quantising the sphere gives spin $n/2$ (Borel–Weil)

The simplest compact Kähler phase space is $P = \mathbb{CP}^1 = S^2$ with ω equal to $n/2$ times the area form of the unit sphere, so the total area is $2\pi n$. The prequantum line bundle is $\mathcal{O}(n)$, whose holomorphic sections in the affine coordinate z are polynomials of degree at most n:

$$\mathcal{H} = \operatorname{span}\{1, z, z^2, \dots, z^n\}, \qquad \dim \mathcal{H} = n + 1 = \frac{2\pi n}{2\pi} + 1.$$

(The $+1$ is the Riemann–Roch correction to the volume count.) SU(2) acts on these polynomials by the differential operators

$$J_+ = z^2\partial_z - nz, \qquad J_- = -\partial_z, \qquad J_3 = z\partial_z - \tfrac{n}{2}.$$

Check the algebra on a test function $f(z)$:

$$[J_3, J_+]f \overset{(1)}{=} [z\partial_z, z^2\partial_z]f - n[z\partial_z, z]f = z^2 f' - nzf = J_+ f,$$
$$[J_+, J_-]f \overset{(2)}{=} -[z^2\partial_z, \partial_z]f + n[z, \partial_z]f = 2zf' - nf = 2J_3 f.$$

(1) $z\partial_z(z^2 f') - z^2\partial_z(zf') = z^2 f'$ and $z\partial_z(zf) - z\,(zf)'$ reduces to zf.
(2) $z^2 f'' - (z^2 f')' = -2zf'$ and $zf' - (zf)' = -f$.

On the basis, $J_3 z^m = (m - \frac{n}{2})z^m$, so the weights run from $-\frac{n}{2}$ to $+\frac{n}{2}$: this is the spin-$\frac{n}{2}$ representation. The lesson: *quantising a sphere of area $2\pi n$ produces the irreducible representation of dimension $n+1$*. Such a sphere is a coadjoint orbit of SU(2).

This example explains how Wilson lines enter the Hamiltonian picture. A Wilson line $W(R, C) = \mathrm{tr}_R\, P \exp \oint_C A$ can be rewritten as a path integral over an auxiliary "spin" living on the curve, whose phase space is the coadjoint orbit that quantises to R (Kirillov's orbit method). When C pierces Σ at a point P, the spin at P couples to the gauge field, and the Gauss law becomes $\tilde{F} = (\text{charge})\,\delta_P$. Classically this means the holonomy around P is no longer trivial but is fixed to lie in a definite conjugacy class, determined by the representation. Quantum mechanically, the physical Hilbert space $\mathcal{H}_{\Sigma; R_1, \dots, R_n}$ is built from the flat connections on $\Sigma \setminus \{P_1, \dots, P_n\}$ with these prescribed holonomies, tensored with the spins. It is again finite-dimensional.

Physical picture: Gauss law = gauge invariance of the total state

For a large level k the prescribed holonomies are close to 1, and the only remnant of the Gauss law is that the total state of all the spins must be *gauge invariant*: $\mathcal{H}_{S^2; R_1, \dots, R_n} \approx \mathrm{Inv}(R_1 \otimes \cdots \otimes R_n)$, the space of singlets. This is Nash's remark that "the physical Hilbert space must be SU(2)-invariant". At finite k some singlets are lost (Section 12.14.6).

12.14.5 The Hilbert spaces are conformal blocks

The WZW model of Chapter 11 built from the loop group LG at level k has, on each punctured Riemann surface, a finite-dimensional space of conformal blocks $V(\Sigma_{p,n})$, and these form a vector bundle over the moduli space of punctured Riemann surfaces. Witten's claim, which Nash states, is

$$\mathcal{H}_{\Sigma_{p,n}; R_1, \dots, R_n} \;\cong\; V\big(\Sigma_{p,n}; R_1, \dots, R_n\big). \tag{12.255}$$

Why the Gauss law is the WZW Ward identity

Here is the mechanism, at the physicist level (normalisations suppressed). Give Σ a complex coordinate z and use the holomorphic polarisation: wavefunctions are functionals $\Psi[A_{\bar z}]$ of the $(0,1)$ part of the connection. The canonical commutator of Example 12.94 makes A_z act as a derivative,

$$A_z \;\longrightarrow\; \frac{2\pi}{k}\,\frac{\delta}{\delta A_{\bar z}}.$$

The Gauss law $F_{z\bar z} = \partial_z A_{\bar z} - \partial_{\bar z} A_z + [A_z, A_{\bar z}] = 0$ then becomes a first-order functional differential equation,

$$\Big(\frac{k}{2\pi}\,\partial_z A_{\bar z} - \partial_{\bar z}\frac{\delta}{\delta A_{\bar z}} - \Big[A_{\bar z}, \frac{\delta}{\delta A_{\bar z}}\Big]\Big)\Psi[A_{\bar z}] = 0. \tag{12.256}$$

Now look at the WZW model with a background field $A_{\bar z}$ coupled to its holomorphic current $J(z)$: its generating functional $Z[A_{\bar z}] = \langle \exp \int \mathrm{tr}(J A_{\bar z}) \rangle$ obeys exactly this equation, which is

the Ward identity expressing $\bar\partial J + [A_{\bar z}, J] = \frac{k}{2\pi}\partial A_{\bar z}$, the anomalous conservation of the current. So the Chern–Simons physical states *are* WZW generating functionals, i.e. conformal blocks. With punctures the same argument produces the blocks with primary fields inserted at the punctures.

Watch out: Chiral, not full, CFT

The conformal blocks of a CFT are *chiral* (holomorphic) objects. The full correlation functions of the 2d CFT are sesquilinear combinations $\sum h_{IJ}\mathcal{F}_I\bar{\mathcal{F}}_J$ of blocks. Chern–Simons theory sees only one chirality: $\mathcal{H}_\Sigma$ is the space of blocks itself, not the space of full correlators.

12.14.6 The data of SU(2) at level k

From now on $G = \mathrm{SU}(2)$. Everything we need from the CFT side fits in one box.

Working definition 12.39 ($\mathrm{SU}(2)_k$ modular data)

- *Labels.* The integrable representations [the only spins that give consistent primary fields at level k] of the level-k current algebra are the spins $j = 0, \frac12, 1, \dots, \frac{k}{2}$. There are $k+1$ of them. Nash labels them by $i = 2j \in \{0, 1, \dots, k\}$, so $i = 0$ is the trivial representation and $i = 1$ the defining (2-dimensional) one.
- *Conformal weights* [$\phi_j(z)$ picks up a factor λ^{-h_j} under $z \to \lambda z$] of the primary fields, and the central charge c:
$$h_j = \frac{j(j+1)}{k+2} = \frac{i(i+2)}{4(k+2)}, \qquad c = \frac{3k}{k+2}.$$
(Sugawara: $h = C_2(j)/(k+h^\vee)$ with $C_2(j) = j(j+1)$, $h^\vee = 2$.)
- *Fusion rules* [which spins j can appear when ϕ_{j_1} and ϕ_{j_2} are brought together]. $j_1 \times j_2 = \sum_j j$ where j runs in integer steps from $|j_1 - j_2|$ to $\min(j_1 + j_2,\ k - j_1 - j_2)$. The lower limit is the Clebsch–Gordan rule; the upper limit is the level-k truncation.
- *Modular S-matrix* [the matrix, on the blocks of the torus, of the swap of its two circles] (Nash (12.157); derived in Section 12.15.6):
$$S_{ij} = \sqrt{\frac{2}{k+2}}\,\sin\frac{(i+1)(j+1)\pi}{k+2}, \qquad i, j = 0, \dots, k.$$
- *Dimensions.* $\dim \mathcal{H}_\Sigma$ for any punctured surface is computed by cutting Σ into pairs of pants [spheres with three holes] and summing over the labels of the cuts, with $N_{j_1 j_2}^{\ j} \in \{0, 1\}$ from the fusion rules at each pair of pants. Equivalently, Verlinde's formula:
$$\dim \mathcal{H}_{\Sigma_g; j_1, \dots, j_n} = \sum_{m=0}^{k} S_{0m}^{2-2g-n} \prod_{r=1}^{n} S_{i_r m}. \tag{12.257}$$

Worked example 12.99: Four spin-$\frac{1}{2}$ punctures on a sphere

This is the Hilbert space behind the skein relation. First the classical count (large k): we need the number of singlets in $R^{\otimes 4}$ with $R = \frac{1}{2}$.

$$\begin{aligned}\tfrac{1}{2}\otimes\tfrac{1}{2}\otimes\tfrac{1}{2}\otimes\tfrac{1}{2} &\overset{(1)}{=} (0\oplus 1)\otimes(0\oplus 1)\\ &\overset{(2)}{=} (0\otimes 0)\oplus(0\otimes 1)\oplus(1\otimes 0)\oplus(1\otimes 1)\\ &\overset{(3)}{=} 0\oplus 1\oplus 1\oplus(0\oplus 1\oplus 2).\end{aligned}$$

(1) Pair the factors: $\frac{1}{2}\otimes\frac{1}{2} = 0\oplus 1$ (singlet plus triplet); dimension check $2\cdot 2 = 1+3$.
(2) Distribute.
(3) $1\otimes 1 = 0\oplus 1\oplus 2$; dimension check $16 = 1+3+3+(1+3+5)$.

The spin 0 appears twice, so $\dim \mathrm{Inv}(R^{\otimes 4}) = 2$: this is Nash's "the trivial representation occurs precisely twice". A basis is labelled by the spin $j \in \{0,1\}$ of the pair (P_1, P_2) (Figure 12.66).

Now at level k. The intermediate spin $j = 1$ is allowed only if $1 \le k/2$, i.e. $k \ge 2$; the fusion $\frac{1}{2}\times\frac{1}{2}$ at $k = 1$ has upper limit $\min(1, 1-1) = 0$. So

$$\dim \mathcal{H}_{S^2;\frac{1}{2},\frac{1}{2},\frac{1}{2},\frac{1}{2}} = \begin{cases} 2, & k \ge 2,\\ 1, & k = 1.\end{cases}$$

Verlinde's formula agrees: for $k = 2$, with $S_{0m} = \frac{1}{2}(1,\sqrt{2},1)$ and $S_{1m} = \frac{1}{2}(\sqrt{2},0,-\sqrt{2})$,

$$\sum_m S_{0m}^{-2} S_{1m}^{4} \overset{(1)}{=} 4\cdot\tfrac{1}{4} + 2\cdot 0 + 4\cdot\tfrac{1}{4} = 2.$$

(1) $S_{00}^{-2} = 4$, $S_{10}^{4} = (\sqrt{2}/2)^4 = \frac{1}{4}$; the middle term vanishes; the last term equals the first.

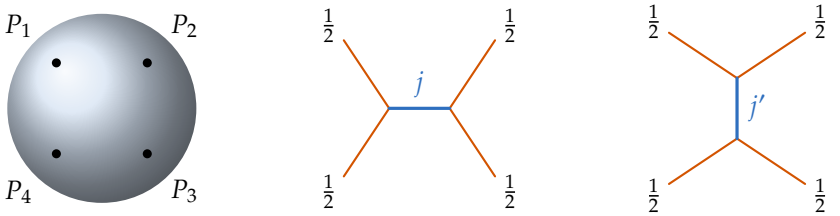


Figure 12.66: The two-dimensional Hilbert space of a sphere with four spin-$\frac{1}{2}$ punctures. A basis is obtained by choosing which pair of punctures to fuse first and labelling the intermediate line by its spin, 0 or 1. Two such choices give two bases, related by a 2×2 "fusion matrix" F that we write down in Section 12.14.12.

Worked example 12.100: Higher genus: Verlinde numbers

With no punctures, (12.257) reads $\dim \mathcal{H}_{\Sigma_g} = \sum_m S_{0m}^{2-2g}$.

- $g = 0$: $\sum_m S_{0m}^2 = 1$, because the rows of the orthogonal matrix S are unit vectors. This is Example 12.97 again.

	$n=0$	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$
$k=1$	1	0	1	0	1	0	1
$k=2$	1	0	1	0	2	0	4
$k=3$	1	0	1	0	2	0	5
$k\to\infty$: $\dim \mathrm{Inv}(R^{\otimes n})$	1	0	1	0	2	0	5

Table 12.2: $\dim \mathcal{H}_{S^2}$ with n punctures, each carrying spin $\frac{1}{2}$, from Verlinde's formula (12.257). Odd n gives zero (no singlet in an odd number of spin-$\frac{1}{2}$'s). At large k the numbers are the Catalan numbers $1,1,2,5,\dots$ counting singlets; small k truncates them. The $n=0$ entry is Example 12.97; $n=2$ gives Nash's $\dim V(S^2_{R_i,R_j}) = g_{ij}$; $n=4$ gives the skein relation.

- $g=1$: $\sum_m S_{0m}^0 = k+1$, matching the pillowcase count of Example 12.96.
- $g=2$, $k=1$: $S_{0m} = \frac{1}{\sqrt{2}}(1,1)$, so $\dim = 2\cdot(\frac{1}{\sqrt{2}})^{-2} = 4 = 2^g$. In general $k=1$ gives 2^g.
- $g=2$, $k=2$: $S_{0m} = (\frac{1}{2}, \frac{1}{\sqrt{2}}, \frac{1}{2})$, so $\dim = (\frac{1}{2})^{-2} + (\frac{1}{\sqrt{2}})^{-2} + (\frac{1}{2})^{-2} = 4+2+4 = 10$.

For large k and $g \geq 2$ the growth is $\dim \sim k^{3g-3}$, in agreement with the semiclassical count (12.253), since $\dim_{\mathbb{R}} \mathcal{M}_F(\Sigma_g) = 6g-6$.

Reading Nash (pp. 350–351, §6 opening)

Nash compresses this whole section into one paragraph: choose $A_0 = 0$ on $\Sigma\times\mathbb{R}$; the phase space is the moduli space $\mathcal{M}_F$ of flat connections, which is compact Kähler; its quantisation gives finite-dimensional $\mathcal{H}_\Sigma$ or $\mathcal{H}_{\Sigma_{p,n}}$, spaces of sections of the kth power of the determinant line bundle; these coincide with the conformal blocks $V(\Sigma)$ of chapter 11 (see Witten, Hitchin). The dimension facts he later uses are his (11.75): $\dim V(S^2_{R_i}) = \delta_{i0}$ and $\dim V(S^2_{R_i,R_j}) = g_{ij}$, with $g_{ij}=1$ if $R_i = R_j^*$ and 0 otherwise. For SU(2) every representation is self-conjugate, so $g_{ij} = \delta_{ij}$. Both follow from (12.257) and the orthogonality of S.

Recap

- Chern–Simons theory on $\Sigma\times\mathbb{R}$ is first order in time, has $\mathcal{H}=0$, and its Gauss law forces flat spatial connections.
- Phase space $= \mathcal{M}_F(\Sigma)$, compact, symplectic form $\propto k$, so $\dim\mathcal{H}_\Sigma < \infty$: U(1) on T^2 gives k, SU(2) on T^2 gives $k+1$, S^2 gives 1.
- Wilson lines piercing Σ are punctures carrying representations; at large k the Hilbert space is the space of singlets.
- $\mathcal{H}_\Sigma$ = space of conformal blocks of the $\mathrm{SU}(2)_k$ WZW model, because the Gauss law is the WZW Ward identity.
- Four spin-$\frac{1}{2}$ punctures on S^2: $\dim = 2$ for $k \geq 2$.

12.14.7 Notation for Wilson-line correlators

Before we cut anything we need a compact way to write the quantities we are after. Recall that a **Wilson line** along an oriented closed curve C, in a representation R of G, is

$$W(R,C) \;=\; \operatorname{tr}_R\, P\exp\oint_C A,$$

the trace, in the representation R, of the holonomy of A around C (Chapter T3). A link $L = C_1 \cup \cdots \cup C_p$ with a representation R_a attached to each component C_a gives a product of p Wilson lines. Its path-integral average is

$$\Big\langle W(R_1,C_1)\cdots W(R_p,C_p)\Big\rangle \;=\; \frac{\widetilde{W}\big(M,\{R\}\big)}{Z(M)}, \qquad \{R\} = R_1,\dots,R_p, \tag{12.258}$$

where the numerator and denominator are the *unnormalised* path integrals

$$\widetilde{W}\big(M,\{R\}\big) \coloneqq \int_{\mathcal{A}/\mathcal{G}} \mathcal{D}A\; e^{iS_k[A]} \prod_{a=1}^{p} W(R_a,C_a), \tag{12.259}$$

$$Z(M) \coloneqq \int_{\mathcal{A}/\mathcal{G}} \mathcal{D}A\; e^{iS_k[A]}. \tag{12.260}$$

The symbol $\{R\}$ records the whole "representation content" of the link: which curves, and which representation on each. When we want to be explicit we write the representations out, for instance

$$\widetilde{W}(M,R_1,R_2) \qquad (p=2), \tag{12.261}$$

and when there are no Wilson lines at all ($p=0$) we write

$$\widetilde{W}(M) \;=\; Z(M), \tag{12.262}$$

because with no insertions (12.259) is just the partition function. Finally the **normalised correlation function** is

$$W\big(M,\{R\}\big) \;\coloneqq\; \frac{\widetilde{W}\big(M,\{R\}\big)}{Z(M)}. \tag{12.263}$$

Working definition 12.40 (unnormalised and normalised correlators)

- $\widetilde{W}(M,\{R\})$: the path integral over connections on M of e^{iS_k} times the product of Wilson lines. It depends on M, on the link (up to isotopy), on the representations and on a framing of each component [a thin ribbon with the curve as one edge, which fixes its self-linking number] (Section 12.14.16).
- $W(M,\{R\}) = \widetilde{W}/Z(M)$: the same, divided by the value with no Wilson lines. This is the "expectation value" a physicist would quote. For $M = S^3$ it will be a Laurent polynomial in $q^{1/4}$ [with $q = e^{2\pi i/(k+2)}$].
- Operationally we never do the integrals. We compute both $\widetilde{W}$ and Z by cutting M into pieces and gluing vectors in the finite-dimensional spaces $\mathcal{H}_\Sigma$ [the space of states on a cut surface Σ, the space of conformal blocks].

From now on $G = \mathrm{SU}(2)$ and, unless stated otherwise, every Wilson line carries the defining 2-dimensional representation, spin $j = \frac{1}{2}$, which Nash calls R (and labels $i = 1$).

12.14.8 A path integral on a manifold with boundary is a vector

The whole method rests on one principle, which every physicist already knows from quantum mechanics.

The quantum-mechanical prototype

The propagator of a particle, $K(x_f,t_f;x_i,t_i)=\langle x_f|e^{-iH(t_f-t_i)}|x_i\rangle$, is a path integral over paths from x_i to x_f. If we stop the paths at an intermediate time t and record where they are, $y=x(t)$, the integral splits:

$$\begin{aligned} K(x_f,t_f;x_i,t_i) &\overset{(1)}{=} \int_{x(t_i)=x_i}^{x(t_f)=x_f} \mathcal{D}x\; e^{iS[x]} \\ &\overset{(2)}{=} \int \mathrm{d}y \int_{x(t)=y}^{x(t_f)=x_f} \mathcal{D}x_2\, e^{iS[x_2]} \int_{x(t_i)=x_i}^{x(t)=y} \mathcal{D}x_1\, e^{iS[x_1]} \\ &\overset{(3)}{=} \int \mathrm{d}y\; K(x_f,t_f;y,t)\,K(y,t;x_i,t_i). \end{aligned} \tag{12.264}$$

(1) Definition of the propagator as a sum over histories.

(2) Every path is a path x_1 up to time t followed by a path x_2 after it, joined at $y=x(t)$; the action is an integral of a local Lagrangian, so $S[x]=S[x_1]+S[x_2]$ and e^{iS} factorises; finally we integrate over the joining point y.

(3) Each inner integral is again a propagator.

In words: the path integral over the *first half* of spacetime, with the fields on the cut held fixed, is a *wavefunction* of those fields, $\psi_1(y)=K(y,t;x_i,t_i)$. The second half gives another function $\psi_2(y)=K(x_f,t_f;y,t)$. The full answer is the pairing $\int \mathrm{d}y\,\psi_2(y)\psi_1(y)$.

The same statement for a 3-manifold

Now let M be a closed 3-manifold and $\Sigma\subset M$ a closed surface that separates it into two pieces, $M=\widetilde{M}_1\cup_\Sigma\widetilde{M}_2$. The Chern–Simons action is the integral of a local 3-form, so it is additive, $S_k[A]=S_k[A|_{\widetilde{M}_1}]+S_k[A|_{\widetilde{M}_2}]$ (the boundary terms cancel between the two sides). Repeat (12.264) with y replaced by the connection a on Σ:

$$\begin{aligned} Z(M) &\overset{(1)}{=} \int \mathcal{D}a \underbrace{\int_{A|_\Sigma=a} \mathcal{D}A_2\; e^{iS_k[A_2]}}_{\Psi_2[a]} \underbrace{\int_{A|_\Sigma=a} \mathcal{D}A_1\; e^{iS_k[A_1]}}_{\Psi_1[a]} \\ &\overset{(2)}{=} \big\langle \Psi_1,\Psi_2\big\rangle. \end{aligned} \tag{12.265}$$

(1) Split every connection on M into its pieces on $\widetilde{M}_1$ and $\widetilde{M}_2$, agreeing on Σ; use additivity of the action; integrate over the common boundary value a.

(2) The integral over a is the natural pairing of a wavefunction on the boundary of $\widetilde{M}_1$ with one on the boundary of $\widetilde{M}_2$.

Gauge invariance (the Gauss law of Section 12.14.3) forces each Ψ_i to lie in the physical Hilbert space, so

$$\Psi_1=V_1\in\mathcal{H}_\Sigma,\qquad \Psi_2=V_2\in\mathcal{H}_{\bar\Sigma}.$$

If Wilson lines run inside $\widetilde{M}_1$ or $\widetilde{M}_2$ they are simply included in the corresponding integral, and V_1, V_2 depend on them. If they cross Σ, the surface acquires punctures and the vectors live in $\mathcal{H}_{\Sigma;R_1,\dots,R_n}$.

Watch out: What is "held fixed" on the boundary

The Chern–Simons action is first order in derivatives, like $p\dot{q}$. In quantum mechanics with $L = p\dot{q} - H$ one cannot fix both q and p at the final time; one fixes q (or p, or $q + ip$). Likewise on Σ we cannot fix both components of a. In the holomorphic polarisation of Section 12.14.5 we fix $a_{\bar{z}}$, and $\Psi[a_{\bar{z}}]$ is a conformal block. Nash's "heuristically a wave function" is this.

Why the other side gives the dual space

The two pieces induce *opposite* orientations on the common surface. If we orient Σ as the boundary of $\widetilde{M}_1$ (outward normal first), then as the boundary of $\widetilde{M}_2$ it is $\bar{\Sigma}$, the same surface with the opposite orientation. The Chern–Simons action is odd under orientation reversal, because it is the integral of a 3-form with no metric in it:

$$\begin{aligned} S_k^{-M}[A] &\overset{(1)}{=} -\frac{k}{4\pi}\int_{-M}\operatorname{tr}\Big(A\wedge \mathrm{d}A + \tfrac{2}{3}A^3\Big) \\ &\overset{(2)}{=} +\frac{k}{4\pi}\int_{M}\operatorname{tr}\Big(A\wedge \mathrm{d}A + \tfrac{2}{3}A^3\Big) = -S_k^{M}[A], \end{aligned} \tag{12.266}$$

(1) Definition (12.247) on the reversed manifold $-M$.
(2) Integrating a top form over $-M$ gives minus the integral over M.

So reversing orientation replaces e^{iS} by $e^{-iS} = \overline{e^{iS}}$, and wavefunctions by their complex conjugates. The Hilbert space of $\bar{\Sigma}$ is the complex conjugate of that of Σ, and with the inner product this is the dual space: $\mathcal{H}_{\bar{\Sigma}} \cong \mathcal{H}_\Sigma^*$. That is why the pairing in (12.265) is the dual pairing of a vector with a covector. It is the axiom "orientation reversal $\mapsto$ dual space" of Atiyah's TQFT axioms, met at the start of this chapter.

Key idea: Gluing

If $\partial\widetilde{M}_1 = \Sigma$ and $\partial\widetilde{M}_2 = \bar{\Sigma}$, the path integrals over the pieces are vectors $V_1 \in \mathcal{H}_\Sigma$ and $V_2 \in \mathcal{H}_\Sigma^*$, and the path integral over the glued manifold is their pairing:

$$\widetilde{W}\big(M, \{R\}\big) = \big\langle V_1, V_2 \big\rangle. \tag{12.267}$$

This is Nash's (12.121). He compares it to the factorisation of string amplitudes on intermediate states (Chapter 9): cutting a world sheet along a circle and inserting a complete set of states.

Worked example 12.101: Two consistency checks: the cylinder and the circle

(a) Take $\widetilde{M} = \Sigma \times [0,1]$. It has two boundary components, $\bar{\Sigma}\times\{0\}$ and $\Sigma\times\{1\}$, so its path integral is a vector in $\mathcal{H}_\Sigma^* \otimes \mathcal{H}_\Sigma = \operatorname{End}(\mathcal{H}_\Sigma)$, i.e. an operator. Since the Hamiltonian vanishes, this operator is $e^{-i\mathcal{H}\cdot 1} = \mathrm{id}$. Gluing a cylinder onto anything therefore changes nothing, as it must: $\Sigma\times[0,1]\cup_\Sigma \widetilde{M}_1 \cong \widetilde{M}_1$.
(b) Glue the two ends of the cylinder to each other. The result is $\Sigma \times S^1$, and gluing the two ends of an operator is taking its trace:

$$Z(\Sigma\times S^1) \overset{(1)}{=} \sum_a \big\langle e_a,\ \mathrm{id}\, e_a\big\rangle \overset{(2)}{=} \dim \mathcal{H}_\Sigma.$$

(1) The identity operator written in a basis e_a of $\mathcal{H}_\Sigma$ with dual basis e^a; gluing the outgoing

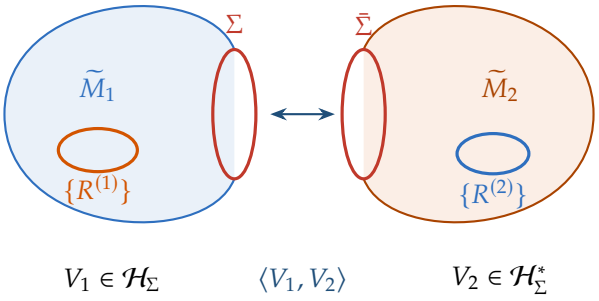


Figure 12.67: Cutting a closed 3-manifold along a surface Σ. The path integral over the left piece, with all Wilson lines inside it, is a vector V_1 in the Hilbert space of Σ; the right piece sees Σ with the opposite orientation and gives a covector V_2. Gluing is pairing. When no Wilson line crosses Σ, the surface has no punctures.

end to the incoming end contracts the upper index with the lower one.

(2) $\mathrm{tr}\ \mathrm{id} = \dim$.

This is Nash's (12.151), which we meet again in Section 12.15.5. With $\Sigma = S^2$ it gives $Z(S^2 \times S^1) = 1$.

12.14.9 Connected sums factorise

What a connected sum is

Given two closed oriented 3-manifolds M_1 and M_2, remove a small open 3-ball from each. What is left, $\widetilde{M}_1$ and $\widetilde{M}_2$, has a 2-sphere as boundary. Glue the two boundary spheres together by an orientation-reversing identification. The result is the **connected sum**

$$M \;=\; M_1 \# M_2 . \tag{12.268}$$

Up to diffeomorphism it does not depend on where the balls were removed. Two simple facts to keep in mind: $M \# S^3 \cong M$ (removing a ball from S^3 leaves a ball, and gluing a ball back into the hole of $\widetilde{M}$ gives M again), and $\pi_1(M_1 \# M_2) = \pi_1(M_1) * \pi_1(M_2)$, the free product, by van Kampen's theorem, because the joining sphere is simply connected.

Conversely, given M and an embedded 2-sphere that separates it, we can *decompose* M as a connected sum by cutting along the sphere and capping each half with a ball. For the first half of the argument we insist that *no knot crosses the joining sphere*. The link then splits into two sets, $\{R^{(1)}\}$ inside M_1 and $\{R^{(2)}\}$ inside M_2 (Figure 12.68).

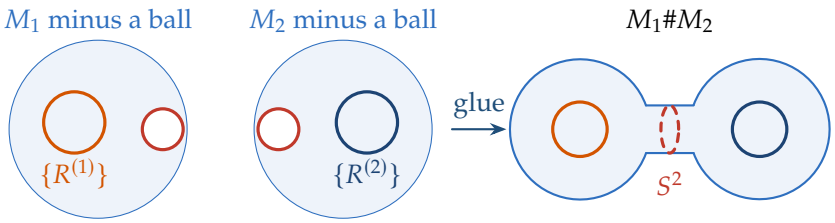


Figure 12.68: The connected sum. Remove a ball (red boundary sphere) from each manifold and glue along the spheres. The knots stay on their own sides: no Wilson line crosses the joining S^2, so the joining sphere carries no punctures and its Hilbert space is one-dimensional.

The factorisation argument

Our goal is Nash's (12.120):

$$Z(S^3)\,\widetilde{W}\big(M,\{R\}\big) \;=\; \widetilde{W}\big(M_1,\{R^{(1)}\}\big)\,\widetilde{W}\big(M_2,\{R^{(2)}\}\big). \tag{12.269}$$

We need four vectors in $\mathcal{H}_{S^2}$ or its dual.

- $V_1 \in \mathcal{H}_{S^2}$: the path integral over $\widetilde{M}_1$ with the Wilson lines $\{R^{(1)}\}$ inside.
- $V_2 \in \mathcal{H}^*_{S^2}$: the path integral over $\widetilde{M}_2$ with $\{R^{(2)}\}$ inside.
- $U_1 \in \mathcal{H}_{S^2}$ and $U_2 \in \mathcal{H}^*_{S^2}$: the path integrals over two empty 3-balls B_1, B_2, with the boundary orientations of $\widetilde{M}_1$ and $\widetilde{M}_2$ respectively.

Now glue these four pieces in the four possible ways. Each gluing is an instance of (12.267):

$$\begin{aligned}
\widetilde{M}_1 \cup \widetilde{M}_2 = M &\;\Longrightarrow\; \widetilde{W}(M,\{R\}) = \langle V_1, V_2\rangle,\\
B_1 \cup B_2 = S^3 &\;\Longrightarrow\; Z(S^3) = \langle U_1, U_2\rangle,\\
\widetilde{M}_1 \cup B_2 = M_1 &\;\Longrightarrow\; \widetilde{W}(M_1,\{R^{(1)}\}) = \langle V_1, U_2\rangle,\\
B_1 \cup \widetilde{M}_2 = M_2 &\;\Longrightarrow\; \widetilde{W}(M_2,\{R^{(2)}\}) = \langle U_1, V_2\rangle.
\end{aligned} \tag{12.270}$$

The second line uses that two balls glued along their boundary make a 3-sphere, the 3-dimensional version of two hemispheres making S^2 (Figure 12.69). The last two lines say that capping the hole of $\widetilde{M}_i$ with a ball gives M_i back.

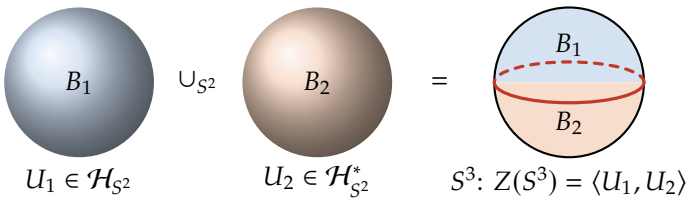


Figure 12.69: Two 3-balls glued along their boundary 2-spheres form S^3. The picture shows the analogy one dimension down: two discs (hemispheres) glued along their boundary circle (red equator) form S^2. Each empty ball defines a vector in the one-dimensional space $\mathcal{H}_{S^2}$ or its dual.

The whole argument now rests on $\dim \mathcal{H}_{S^2} = 1$ (Example 12.97; Nash's (11.75) with no marked points, which corresponds to the trivial representation). In a one-dimensional space every vector is a multiple of any nonzero one. Since $Z(S^3) = \langle U_1, U_2\rangle \neq 0$ (we will find $Z(S^3) = S_{00} > 0$), both U_1 and U_2 are nonzero, so we may write $V_1 = aU_1$ and $V_2 = bU_2$ for some numbers a, b. Then

$$\begin{aligned}
\langle U_1, U_2\rangle\langle V_1, V_2\rangle &\overset{(1)}{=} \langle U_1, U_2\rangle\, ab\, \langle U_1, U_2\rangle\\
&\overset{(2)}{=} \big(a\langle U_1, U_2\rangle\big)\big(b\langle U_1, U_2\rangle\big)\\
&\overset{(3)}{=} \langle V_1, U_2\rangle\, \langle U_1, V_2\rangle.
\end{aligned} \tag{12.271}$$

(1) Substitute $V_1 = aU_1$, $V_2 = bU_2$ and use bilinearity of the pairing.
(2) Regroup the numbers.
(3) Again by bilinearity, $\langle V_1, U_2\rangle = a\langle U_1, U_2\rangle$ and $\langle U_1, V_2\rangle = b\langle U_1, U_2\rangle$.

Reading each factor off (12.270), the left side is $Z(S^3)\widetilde{W}(M,\{R\})$ and the right side is

$$\widetilde{W}\big(M_1,\{R^{(1)}\}\big)\,\widetilde{W}\big(M_2,\{R^{(2)}\}\big) \tag{12.272}$$

so we have shown

$$Z(S^3)\,\widetilde{W}\big(M,\{R\}\big) = \widetilde{W}\big(M_1,\{R^{(1)}\}\big)\,\widetilde{W}\big(M_2,\{R^{(2)}\}\big), \tag{12.273}$$

which after dividing by $Z(S^3)^2$ is the symmetric form

$$\frac{\widetilde{W}\big(M,\{R\}\big)}{Z(S^3)} = \frac{\widetilde{W}\big(M_1,\{R^{(1)}\}\big)}{Z(S^3)}\,\frac{\widetilde{W}\big(M_2,\{R^{(2)}\}\big)}{Z(S^3)}. \tag{12.274}$$

Physical picture: Why a one-dimensional space forces factorisation

A one-dimensional Hilbert space means the joining sphere can carry only one state: "nothing" passes through it. Whatever happens in $\widetilde{M}_1$ can influence $\widetilde{M}_2$ only through that single state, so the two halves are statistically independent, up to the overall normalisation $Z(S^3)$ that measures the "amplitude of the empty sphere". It is the topological version of cluster decomposition.

Reading Nash (pp. 351–352, (12.119)–(12.126))

Nash writes the key identity as $\langle U_1,U_2\rangle\langle V_1,V_2\rangle = \langle V_1,U_1\rangle\langle U_2,V_2\rangle$ and then says "$\langle V_1,U_1\rangle$ corresponds to $\widetilde{M}_1$ 'evolving' into M_1". Read literally this pairs two vectors of the same space. The intended meaning is: pair V_1 with the *ball that caps off* $\widetilde{M}_1$. With our labels, where U_2 is the capping ball for $\widetilde{M}_1$ (it carries the orientation of $\widetilde{M}_2$'s side), this is (12.271). The content is unchanged.

Worked example 12.102: A finite-group check of the factorisation

The factorisation (12.273) holds in *any* TQFT with $\dim\mathcal{H}_{S^2}=1$, so we can test it on a theory where we can do the path integral exactly. Take gauge theory with a *finite* gauge group Γ (the untwisted Dijkgraaf–Witten theory). A connection is then flat automatically and is determined by its holonomies, and the path integral is a finite sum over homomorphisms $\pi_1(M)\to\Gamma$ divided by the size of the gauge group:

$$Z(M) = \frac{\big|\mathrm{Hom}(\pi_1(M),\Gamma)\big|}{|\Gamma|}.$$

Then $Z(S^3)=1/|\Gamma|$ (only the trivial homomorphism from $\pi_1=1$), and for a connected sum

$$\begin{aligned} Z(M_1\#M_2) &\overset{(1)}{=} \frac{\big|\mathrm{Hom}(\pi_1(M_1)*\pi_1(M_2),\Gamma)\big|}{|\Gamma|} \\ &\overset{(2)}{=} \frac{\big|\mathrm{Hom}(\pi_1(M_1),\Gamma)\big|\,\big|\mathrm{Hom}(\pi_1(M_2),\Gamma)\big|}{|\Gamma|} \\ &\overset{(3)}{=} \frac{Z(M_1)\,Z(M_2)}{Z(S^3)}. \end{aligned}$$

(1) Van Kampen: π_1 of a connected sum is the free product.
(2) A homomorphism out of a free product is the same as an independent choice of homomorphism out of each factor.

(3) Multiply and divide by $|\Gamma|$ and use $Z(S^3) = 1/|\Gamma|$.

This is exactly (12.273) with no Wilson lines. For a concrete number take $\Gamma = \mathbb{Z}_2$ and $M_1 = M_2 = \mathbb{RP}^3$ ($\pi_1 = \mathbb{Z}_2$): $Z(\mathbb{RP}^3) = 2/2 = 1$ and $Z(\mathbb{RP}^3\#\mathbb{RP}^3) = |\mathrm{Hom}(\mathbb{Z}_2 * \mathbb{Z}_2, \mathbb{Z}_2)|/2 = 4/2 = 2 = 1 \cdot 1/(1/2)$.

12.14.10 Iterating: unlinks

Nothing stops us from decomposing again. If

$$M = M_1 \# M_2 \# \cdots \# M_p \tag{12.275}$$

with no knot crossing any of the $p-1$ joining spheres, then applying (12.274) once splits off M_1, and applying it to the rest repeatedly gives, by induction on p,

$$\frac{\widetilde{W}(M, \{R\})}{Z(S^3)} = \frac{\widetilde{W}(M_1, \{R^{(1)}\})}{Z(S^3)} \cdots \frac{\widetilde{W}(M_p, \{R^{(p)}\})}{Z(S^3)}. \tag{12.276}$$

The case that matters most is $M = S^3$ and $\{R\}$ a collection of p **unlinked, unknotted circles** (the p-component **unlink**). The circles can be separated by $p-1$ disjoint 2-spheres, one enclosing each circle (Figure 12.70). Cutting along these spheres and capping writes $S^3 = S^3\# \cdots \# S^3$ with exactly one unknot in each copy. Every factor on the right of (12.276) is then the same number, and since $M_i = M = S^3$ we can replace $Z(S^3)$ by the normalisation of each factor:

$$\frac{\widetilde{W}(S^3, \{R\})}{Z(S^3)} = \frac{\widetilde{W}(S^3, R_1)}{Z(S^3)} \cdots \frac{\widetilde{W}(S^3, R_p)}{Z(S^3)}, \qquad \text{i.e.} \qquad W(S^3, \{R\}) = W(S^3, R_1) \cdots W(S^3, R_p). \tag{12.277}$$

The normalised expectation value of an unlink is the product of the expectation values of its components. With all R_a equal to the fundamental representation, and writing $W(\bigcirc) := W(S^3, R)$ for one unknot,

$$W(S^3, \text{unlink of } p) \;=\; W(\bigcirc)^p. \tag{12.278}$$

We still do not know the number $W(\bigcirc)$. The skein relation will give it in Section 12.14.14, and surgery will give it again, independently, in Section 12.15.6.

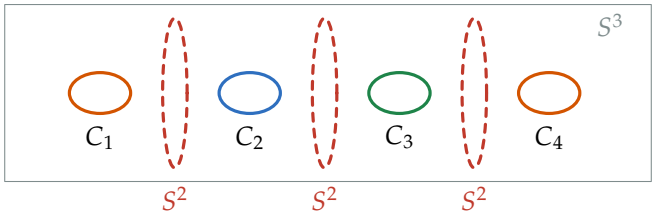


Figure 12.70: A 4-component unlink in S^3. Three 2-spheres (red, dashed) that meet no circle cut S^3 into four pieces, each containing one unknot. By (12.276) the normalised expectation value is the product of the four single-circle values.

Worked example 12.103: Connected sums of 3-manifolds: Z is multiplicative up to $Z(S^3)$

With no Wilson lines, (12.273) reads $Z(M_1\#M_2) = Z(M_1)Z(M_2)/Z(S^3)$. Two quick consequences.

(a) $M_2 = S^3$: $Z(M_1\#S^3) = Z(M_1)Z(S^3)/Z(S^3) = Z(M_1)$, consistent with $M_1\#S^3 \cong M_1$.

(b) $M_1 = M_2 = S^2 \times S^1$, for which $Z = \dim \mathcal{H}_{S^2} = 1$ (Example 12.101):

$$Z\big(\#^g (S^2 \times S^1)\big) \overset{(1)}{=} \frac{Z(S^2 \times S^1)^g}{Z(S^3)^{g-1}} \overset{(2)}{=} Z(S^3)^{1-g} = S_{00}^{1-g}.$$

(1) Apply (12.273) $g-1$ times.
(2) $Z(S^2 \times S^1) = 1$, and $Z(S^3) = S_{00}$, which we derive in Section 12.15.6.

Compare the Verlinde formula $\dim \mathcal{H}_{\Sigma_g} = \sum_m S_{0m}^{2-2g}$ of Example 12.100: the *dominant* term $m = 0$ has the same shape. This is not an accident; both are sums over the labels running round cycles, and here only the trivial label contributes.

Worked example 12.104: Connected sum of knots: the Jones polynomial is multiplicative

Now let the joining sphere be pierced by a knot *twice*. Given knots K_1 and K_2 in S^3, their **connected sum** $K_1 \# K_2$ is made by cutting each open at one point and splicing the loose ends together (Figure 12.71). A 2-sphere separates the two "halves" and meets the knot in two points, carrying R and (because the orientation of the curve runs in on one side and out on the other) R^*, which is R again for SU(2).
The Hilbert space of a sphere with two marked points has dimension $\dim V(S^2_{R_i,R_j}) = g_{ij}$ (Nash's (11.75)), which is 1 for $R_i = R_j$. So the whole argument above goes through verbatim, with U_1, U_2 now the path integrals over balls containing a single *unknotted* arc. Gluing two such balls gives an unknot in S^3, so $\langle U_1, U_2 \rangle = \widetilde{W}(S^3, \bigcirc)$ replaces $Z(S^3)$:

$$\widetilde{W}(S^3, \bigcirc)\, \widetilde{W}(S^3, K_1 \# K_2) \overset{(1)}{=} \widetilde{W}(S^3, K_1)\, \widetilde{W}(S^3, K_2),$$
$$\frac{W(S^3, K_1 \# K_2)}{W(\bigcirc)} \overset{(2)}{=} \frac{W(S^3, K_1)}{W(\bigcirc)} \frac{W(S^3, K_2)}{W(\bigcirc)}.$$

(1) The four-vector identity (12.271) in the one-dimensional space $\mathcal{H}_{S^2;R,R}$.
(2) Divide by $Z(S^3)^2 W(\bigcirc)^2$ and use $\widetilde{W} = Z(S^3)\, W$ on S^3.

The ratio $W(S^3, K)/W(\bigcirc)$ will turn out to be the Jones polynomial $V_K(t)$ (Section 12.14.14). So we have derived, from $\dim \mathcal{H}_{S^2;R,R} = 1$ alone, the known property

$$V_{K_1 \# K_2}(t) = V_{K_1}(t)\, V_{K_2}(t).$$

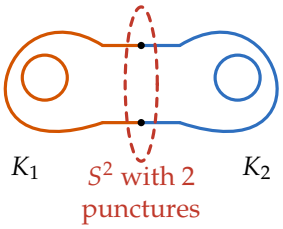
For instance, with the right-handed trefoil $V = t + t^3 - t^4$ and its mirror image $V = t^{-1} + t^{-3} - t^{-4}$ (Example 12.109), the **granny knot** (trefoil # trefoil) and the **square knot** (trefoil # mirror trefoil) have

$$V_{\text{granny}} = (t + t^3 - t^4)^2 = t^2 + 2t^4 - 2t^5 + t^6 - 2t^7 + t^8,$$
$$V_{\text{square}} = (t + t^3 - t^4)(t^{-1} + t^{-3} - t^{-4}) = -t^{-3} + t^{-2} - t^{-1} + 3 - t + t^2 - t^3.$$

The Jones polynomial tells these two apart, while the Alexander polynomial, $(t^2 - t + 1)^2$ for both, does not.

Try it: Three punctures

Why can a sphere never be pierced by a single Wilson line of a closed link (or by three spin-$\frac{1}{2}$ lines)? What does $\dim \mathcal{H}$ say?
Answer. A closed curve crosses a separating sphere an even number of times (it must come



$$\dim \mathcal{H}_{S^2;R,R} = 1 \Rightarrow V_{K_1\#K_2} = V_{K_1} V_{K_2}$$

Figure 12.71: The connected sum of two knots. A sphere meeting the knot in two points separates the two summands. The small circles stand for whatever knotting each summand contains. Because the twice-punctured sphere has a one-dimensional Hilbert space, the same argument that factorised unlinks factorises the Jones polynomial.

back). Algebraically: $\dim \mathcal{H}_{S^2;\frac{1}{2}} = \delta_{\frac{1}{2},0} = 0$ and $\dim \mathcal{H}_{S^2;\frac{1}{2},\frac{1}{2},\frac{1}{2}} = 0$ (Table 12.2), so such configurations would have vanishing amplitude anyway. There is no singlet in an odd tensor power of spin $\frac{1}{2}$.

Recap

- $\widetilde{W}(M,\{R\})$ is the unnormalised path integral with Wilson lines, $Z(M) = \widetilde{W}(M)$, and $W = \widetilde{W}/Z$.
- Path integral over a manifold with boundary Σ = vector in $\mathcal{H}_\Sigma$; the other side gives the dual; gluing = pairing (Nash's (12.121)). Cylinder = identity, $\Sigma \times S^1$ = trace.
- $\dim \mathcal{H}_{S^2} = 1$ forces $Z(S^3)\widetilde{W}(M_1\#M_2) = \widetilde{W}(M_1)\widetilde{W}(M_2)$.
- Unlink of p circles: $W = W(\bigcirc)^p$.
- $\dim \mathcal{H}_{S^2;R,R} = 1$ gives $V_{K_1\#K_2} = V_{K_1} V_{K_2}$.

12.14.11 Knots through the joining sphere

So far no knot was allowed to cross the joining sphere, and the result was factorisation. Now we do the opposite: we let the link cross the joining sphere, and extract the *skein relation*. Recall its shape (Nash's (12.118), in the standard normalisation of the Jones polynomial):

$$t^{-1} V_{L_+}(t) - t\, V_{L_-}(t) \;=\; \left(t^{1/2} - t^{-1/2}\right) V_{L_0}(t), \tag{12.279}$$

where L_+, L_-, L_0 are three links that are identical except inside a small ball, where they look like a positive crossing, a negative crossing and no crossing (Figure 12.72). Nash's version has L_+ and L_- interchanged; we come back to this in Section 12.14.13. His summary of the quantum field theory content is striking: the skein relation "amounts to the statement that any three vectors in a two-dimensional vector space are linearly dependent". Let us see exactly how.

The set-up

Take a link L in M and choose a crossing of some diagram of it. Surround the crossing by a small ball $\widetilde{M}_2$; call the rest $\widetilde{M}_1$. Now $M = \widetilde{M}_1 \cup_{S^2} \widetilde{M}_2$, but the joining sphere is pierced at *four* points, where the two strands enter and leave the ball. Each puncture carries the fundamental representation R (as a point of S^2 it carries R or R^* according to whether the strand enters or

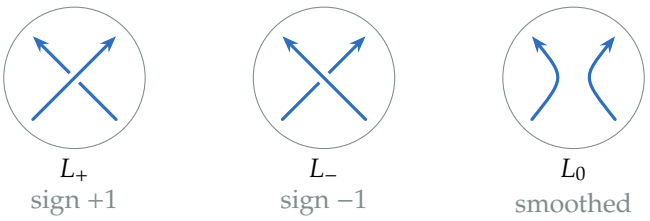


Figure 12.72: The skein triple. Both strands point upwards. In L_+ the strand going from lower left to upper right passes over; in L_- it passes under; in L_0 the crossing is replaced by two parallel arcs, respecting the orientations. Outside the grey ball the three links agree. Each picture is the inside of the ball $\widetilde{M}_2$, whose boundary sphere is pierced at four points.

leaves; for SU(2) these agree). By the gluing rule,

$$\widetilde{W}\big(M,\{R\}\big) = \big\langle V_1, V_2\big\rangle, \qquad V_1 \in \mathcal{H}_{S^2;R,R,R,R}, \quad V_2 \in \mathcal{H}^*_{S^2;R,R,R,R}. \tag{12.280}$$

Reading Nash (p. 353, footnote 8)

Why four punctures rather than two? In a skein relation the two strands in the ball may belong to two different components of the link, or to the same one. To be completely general we allow each of the two strands to belong to its own Wilson loop. Each loop must enter and leave the ball, so there are two entry points and two exit points on the sphere: a 4-punctured S^2.

The Hilbert space of this 4-punctured sphere is the one we computed in Example 12.99:

$$\mathcal{H}_{S^2;R,R,R,R} \equiv \mathcal{H}_{S^2,4}, \qquad \dim \mathcal{H}_{S^2,4} = 2 \quad (k \geq 2),$$

because the trivial representation occurs exactly twice in $R \otimes R \otimes R \otimes R$ (the physical Hilbert space must be SU(2)-invariant).

Three links, three vectors

Do the decomposition three times, for the three links L_+, L_-, L_0 that differ only inside the ball. The outside $\widetilde{M}_1$, with everything in it, is the same each time, so it gives the *same* vector V_1. Only the inside changes, giving three vectors $V_2^+, V_2^-, V_2^0 \in \mathcal{H}^*_{S^2,4}$. Nash's (12.131) is

$$\widetilde{W}\big(M,\{R^+\}\big) = \langle V_1, V_2^+\rangle, \quad \widetilde{W}\big(M,\{R^-\}\big) = \langle V_1, V_2^-\rangle, \quad \widetilde{W}\big(M,\{R^0\}\big) = \langle V_1, V_2^0\rangle, \tag{12.281}$$

where $\{R^\pm\}, \{R^0\}$ denote the Wilson-line content of $L_\pm, L_0$. Three vectors in a two-dimensional space are linearly dependent: there are numbers $\alpha^+, \alpha^-, \alpha^0$, not all zero, with

$$\begin{aligned} &\alpha^+ V_2^+ + \alpha^- V_2^- + \alpha^0 V_2^0 = 0 \\ \Longrightarrow \quad &\alpha^+ \widetilde{W}\big(M,\{R^+\}\big) + \alpha^- \widetilde{W}\big(M,\{R^-\}\big) + \alpha^0 \widetilde{W}\big(M,\{R^0\}\big) = 0, \end{aligned} \tag{12.282}$$

the second line following from the first by pairing with V_1. This is already a skein relation. Two questions remain: what are the α's, and why are they the *same* for every link (that is, why do they not depend on the vectors V_2, which depend on how the ball sits in the link)? Both are answered by the braiding matrix.

Worked example 12.105: Three vectors in $\mathbb{C}^2$: the coefficients are determinants

Here is an explicit formula for the dependence. For $u, v, w \in \mathbb{C}^2$ write $[u, v] := u_1 v_2 - u_2 v_1$. Then

$$[v, w]\, u + [w, u]\, v + [u, v]\, w = 0. \tag{12.283}$$

To see it, look at the i-th component ($i = 1, 2$):

$$[v, w]u_i + [w, u]v_i + [u, v]w_i \overset{(1)}{=} \det \begin{pmatrix} u_i & v_i & w_i \\ u_1 & v_1 & w_1 \\ u_2 & v_2 & w_2 \end{pmatrix} \overset{(2)}{=} 0.$$

(1) Expand the 3×3 determinant along its first row; the three cofactors are $[v, w]$, $-[u, w] = [w, u]$ and $[u, v]$.

(2) The first row equals row $i + 1$, and a determinant with two equal rows vanishes.

For example, $u = (1, 0)$, $v = (1, 1)$, $w = (2, 3)$ give $[v, w] = 1$, $[w, u] = 3$, $[u, v] = 1$, and indeed $(1, 0) + 3(1, 1) - \ldots$ wait: $1 \cdot (1, 0) + 3 \cdot (1, 1) + 1 \cdot (2, 3)$ is not zero. The signs matter: $[w, u] = w_1 u_2 - w_2 u_1 = 0 - 3 = -3$, so the correct combination is $(1, 0) - 3(1, 1) + (2, 3) = (0, 0)$. The lesson for the skein relation is that the α's are not arbitrary: they are built from the vectors themselves, and we must find a reason for them to be universal.

12.14.12 The braiding matrix

Diffeomorphisms of the punctured sphere act on the Hilbert space

Look again at the three pictures in Figure 12.72. Each is a ball containing two arcs, attached to the same four points on the boundary. Take L_0 and grab the two upper endpoints on the boundary sphere. Rotate them around each other by half a turn, anticlockwise as seen from above, dragging the arcs along. The two parallel arcs acquire a crossing: we get L_+. Another half turn in the same direction gives two full twists; a half turn in the opposite direction gives L_-. So

$$L_+ = \sigma(L_0), \qquad L_- = \sigma^{-1}(L_0),$$

where σ is the **half-twist** diffeomorphism of the ball that exchanges two of the punctures on the boundary (Figure 12.73).

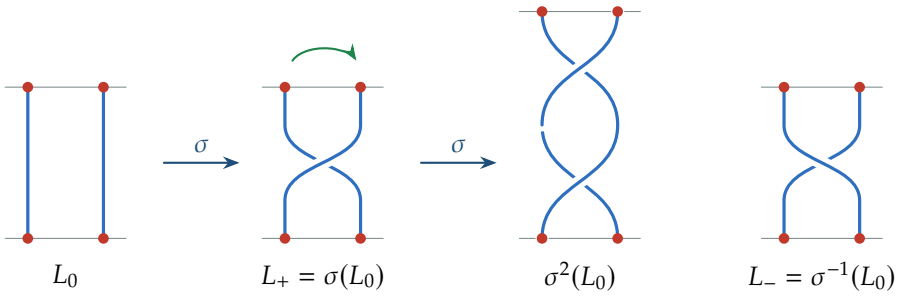


Figure 12.73: The half-twist. Draw the ball as a cylinder with the four punctures (red) on the top and bottom discs. Rotating the two top punctures around each other by half a turn turns the parallel arcs L_0 into L_+; doing it again gives a full twist $\sigma^2(L_0)$; rotating the other way gives L_-. Because these are diffeomorphisms of the punctured sphere, they act on $\mathcal{H}_{S^2,4}$ as 2×2 matrices.

A diffeomorphism ϕ of a punctured surface maps the Hilbert space to itself: it is a symmetry of the classical phase space (it maps flat connections to flat connections, respecting the punctures and their labels), so after quantisation it is a linear operator $\rho(\phi)$ on $\mathcal{H}_\Sigma$. Diffeomorphisms

isotopic to the identity act trivially (the theory is topological), so only the **mapping class group** (diffeomorphisms modulo isotopy) acts. Applied to the half-twist: the vector of the ball containing $\sigma(L_0)$ is the operator σ applied to the vector of the ball containing L_0,

$$V_2^+ = \sigma\, V_2^0, \qquad V_2^- = \sigma^{-1} V_2^0. \tag{12.284}$$

(Strictly, σ acts on the dual space $\mathcal{H}^*_{S^2,4}$ in which the V_2's live; we use the same letter.) Nash writes this as

$$V_2^0 = B V_2^+, \qquad V_2^- = B V_2^0 = B^2 V_2^+, \tag{12.285}$$

with his **braiding matrix** $B = \sigma^{-1}$ (his (12.133)–(12.134)). It is the same statement read in the other direction.

The braid group

The half-twists of adjacent punctures generate the **braid group**. For n strands it has generators $\sigma_1, \dots, \sigma_{n-1}$ (σ_i crosses strand i over strand $i+1$) and relations

$$\sigma_i \sigma_{i+1} \sigma_i = \sigma_{i+1} \sigma_i \sigma_{i+1}, \qquad \sigma_i \sigma_j = \sigma_j \sigma_i \quad (|i-j| \geq 2). \tag{12.286}$$

The first is the **braid relation** (Figure 12.74), the third Reidemeister move in disguise. Unlike the permutation group, there is *no* relation $\sigma_i^2 = 1$: twisting twice is not the same as doing nothing.

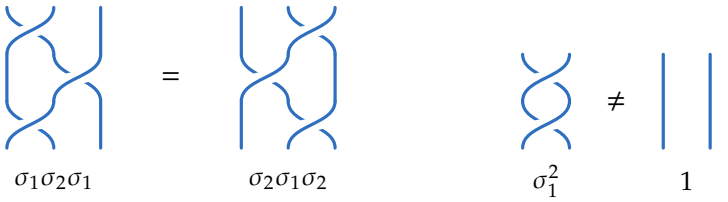


Figure 12.74: Left: the braid relation $\sigma_1\sigma_2\sigma_1 = \sigma_2\sigma_1\sigma_2$ (read from bottom to top). The two braids are isotopic: slide the middle crossing across. Right: a full twist of two strands is *not* isotopic to the trivial braid with the ends fixed, so $\sigma^2 \neq 1$ in the braid group, and its matrix need not be the identity.

The eigenvalues come from the operator product expansion

Choose the basis of $\mathcal{H}_{S^2,4}$ adapted to the pair of punctures that the half-twist exchanges: basis vector e_j means "the two exchanged strands fuse into total spin j", with $j \in \{0, 1\}$ (the s-channel basis of Figure 12.66). Exchanging two particles does not change their total spin, so σ is *diagonal* in this basis, $\sigma e_j = \lambda_j e_j$. To find λ_j we use the conformal blocks. Near each other, two primary fields of spin $\frac{1}{2}$ and weight $h := h_{1/2}$ have the operator product expansion

$$\phi_{1/2}(z_1)\,\phi_{1/2}(z_2) \;\sim\; \sum_{j=0,1} (z_1 - z_2)^{h_j - 2h} \left[\phi_j(z_2) + \dots\right],$$

so the block e_j behaves like $(z_1 - z_2)^{h_j - 2h}$. Half an anticlockwise turn of z_1 around z_2 sends $z_1 - z_2 \mapsto e^{i\pi}(z_1 - z_2)$, and multiplies the block by $e^{i\pi(h_j - 2h)}$. The exchange also swaps the two spin-$\frac{1}{2}$ labels, which costs the Clebsch–Gordan symmetry sign ε_j: the singlet is antisymmetric ($\varepsilon_0 = -1$), the triplet symmetric ($\varepsilon_1 = +1$). With $h_j = j(j+1)/(k+2)$ and $q = e^{2\pi i/(k+2)}$, so that

$e^{i\pi x/(k+2)} = q^{x/2}$:

$$\lambda_0 \overset{(1)}{=} \varepsilon_0\, e^{i\pi(h_0 - 2h_{1/2})} \overset{(2)}{=} -\exp\!\left(i\pi\left(0 - \frac{2\cdot\frac{3}{4}}{k+2}\right)\right) \overset{(3)}{=} -q^{-3/4},$$
$$\lambda_1 \overset{(4)}{=} \varepsilon_1\, e^{i\pi(h_1 - 2h_{1/2})} \overset{(5)}{=} \exp\!\left(i\pi\,\frac{2-\frac{3}{2}}{k+2}\right) = q^{1/4}. \tag{12.287}$$

(1) Phase of the block under the half-exchange, times the symmetry sign.
(2) $h_0 = 0$ and $h_{1/2} = \frac{1}{2}\cdot\frac{3}{2}/(k+2) = \frac{3}{4}/(k+2)$; $\varepsilon_0 = -1$.
(3) $e^{-i\pi\,(3/2)/(k+2)} = q^{-3/4}$.
(4) The same rule for the triplet.
(5) $h_1 = 2/(k+2)$, $\varepsilon_1 = +1$; then $e^{i\pi(1/2)/(k+2)} = q^{1/4}$.

Which sense of rotation is the "positive" crossing is a convention; we fix it by the framing check in Example 12.112, which requires this choice. So in the s-channel basis

$$\sigma \;=\; \begin{pmatrix} -q^{-3/4} & 0 \\ 0 & q^{1/4} \end{pmatrix}, \qquad \det\sigma = -q^{-1/2}, \qquad \operatorname{tr}\sigma = q^{1/4} - q^{-3/4}. \tag{12.288}$$

Physical picture: The classical limit is the permutation of two spins

As $k \to \infty$, $q \to 1$ and $\sigma \to \operatorname{diag}(-1,+1)$. On $\mathbb{C}^2 \otimes \mathbb{C}^2$ the operator that is -1 on the singlet and $+1$ on the triplet is the *swap* $P(u\otimes v) = v\otimes u$. So at infinite level a crossing just exchanges the two spins, $\sigma^2 = 1$, and the theory cannot tell over from under. The quantum deformation $q \neq 1$ splits the two eigenvalues apart by the phases $q^{-3/4}$ and $q^{1/4}$; that splitting is what makes over and under different, and it vanishes like $1/k$.

The second generator and the fusion matrix

With four punctures there are two independent half-twists: σ_1 exchanging punctures 1 and 2, and σ_2 exchanging 2 and 3. In the basis adapted to $(1,2)$, σ_1 is diagonal; σ_2 is diagonal in the basis adapted to $(2,3)$. The change of basis between the two "channels" is the **fusion matrix** F (Moore–Seiberg). For four spin-$\frac{1}{2}$ punctures it is

$$F = \frac{1}{[2]}\begin{pmatrix} 1 & \sqrt{[3]} \\ \sqrt{[3]} & -1 \end{pmatrix}, \qquad F^2 = \mathrm{id}, \qquad \det F = -1, \tag{12.289}$$

the quantum version of the $6j$ symbols of the rotation group, and $\sigma_2 = F\sigma_1 F$. The check $F^2 = \mathrm{id}$ uses $[2]^2 = [3] + 1$ (true for all q: $(q^{1/2} + q^{-1/2})^2 = q + 2 + q^{-1}$).

Worked example 12.106: The braid relation holds

We verify $\sigma_1\sigma_2\sigma_1 = \sigma_2\sigma_1\sigma_2$ with $D := \sigma_1 = \operatorname{diag}(a,b)$, $a = -q^{-3/4}$, $b = q^{1/4}$ and $\sigma_2 = FDF$. First reduce the braid relation to a statement about the single matrix $M := DF$:

$$\sigma_1\sigma_2\sigma_1 = \sigma_2\sigma_1\sigma_2 \overset{(1)}{=} \iff DFDFD = FDFDFDF$$
$$\overset{(2)}{=} \iff (DF)^3F = F(DF)^3$$
$$\overset{(3)}{=} \impliedby (DF)^3 = c\,\mathrm{id} \;\text{ for some number } c.$$

(1) Substitute $\sigma_2 = FDF$; on the right, $FDF\cdot D\cdot FDF$.
(2) Multiply the left side on the right by $F^2 = \mathrm{id}$: $DFDFD = DFDFDF\cdot F = (DF)^3F$; the

right side is $F(DF)^3$.

(3) A multiple of the identity commutes with F.

A 2×2 matrix with eigenvalues μ and $\mu\omega$, $\omega = e^{\pm 2\pi i/3}$, cubes to $\mu^3\mathrm{id}$. Its trace and determinant satisfy $(\operatorname{tr} M)^2 = \mu^2(1+\omega)^2 = \mu^2\omega = \det M$, using $1+\omega+\omega^2 = 0$. So it is enough to check $(\operatorname{tr} M)^2 = \det M$:

$$(\operatorname{tr} DF)^2 \overset{(4)}{=} \left(\frac{a-b}{[2]}\right)^2 \overset{(5)}{=} \left(\frac{-q^{-1/4}(q^{-1/2}+q^{1/2})}{[2]}\right)^2 = q^{-1/2},$$
$$\det(DF) \overset{(6)}{=} ab\,\det F = -ab \overset{(7)}{=} q^{-1/2}.$$

(4) The diagonal of DF is $(aF_{11}, bF_{22}) = (a,-b)/[2]$.

(5) $a-b = -q^{-3/4}-q^{1/4} = -q^{-1/4}(q^{-1/2}+q^{1/2}) = -q^{-1/4}[2]$.

(6) $\det(DF) = \det D\,\det F$ and $\det F = -1$.

(7) $ab = -q^{-3/4}q^{1/4} = -q^{-1/2}$.

They agree, so the braid relation holds for every k. Notice that it is the *specific* eigenvalues (12.287) that make this work: the ratio $b/a = -q$ and the entries of F conspire.

Worked example 12.107: Level $k=2$: Hadamard and phase gates

At $k=2$, $q = e^{2\pi i/4} = i$, $[2] = \sqrt{2}$, $[3] = 1$. Then $a = -q^{-3/4} = -e^{-3\pi i/8} = e^{i\pi/8}\cdot i$ and $b = e^{i\pi/8}$, so

$$\sigma_1 = e^{i\pi/8}\begin{pmatrix} i & 0\\ 0 & 1\end{pmatrix}, \qquad F = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1\\ 1 & -1\end{pmatrix}, \qquad \sigma_2 = F\sigma_1 F = \frac{e^{i\pi/8}}{2}\begin{pmatrix} 1+i & i-1\\ i-1 & 1+i\end{pmatrix}.$$

F is the Hadamard matrix, and σ_1 is (up to a phase) the "phase gate" of quantum computing. The check of the previous example is $(\operatorname{tr} DF)^2 = \big((i-1)/\sqrt{2}\big)^2 = -i = i\cdot(-1) = \det(DF)$ (with the common phase $e^{i\pi/8}$ stripped off). These are the braiding matrices of the Ising anyons that appear in the $\nu = 5/2$ quantum Hall state: particles whose exchange acts by a matrix on a degenerate ground state.

12.14.13 The skein relation, derived

Cayley–Hamilton does the work

Every 2×2 matrix satisfies its own characteristic equation, $\sigma^2 - (\operatorname{tr}\sigma)\,\sigma + (\det\sigma)\,\mathrm{id} = 0$. Multiply by σ^{-1} and apply the result to V_2^0:

$$\begin{aligned} 0 &\overset{(1)}{=} \sigma V_2^0 - (\operatorname{tr}\sigma)V_2^0 + (\det\sigma)\,\sigma^{-1}V_2^0\\ &\overset{(2)}{=} V_2^+ - (q^{1/4}-q^{-3/4})V_2^0 - q^{-1/2}V_2^-\\ &\overset{(3)}{=} q^{1/4}V_2^+ - q^{-1/4}V_2^- - (q^{1/2}-q^{-1/2})V_2^0. \end{aligned} \tag{12.290}$$

(1) Cayley–Hamilton divided by σ, applied to V_2^0.

(2) (12.284) and the trace and determinant (12.288).

(3) Multiply by $q^{1/4}$: $q^{1/4}(q^{1/4}-q^{-3/4}) = q^{1/2}-q^{-1/2}$ and $q^{1/4}q^{-1/2} = q^{-1/4}$.

Pair with V_1:

$$q^{1/4}\,\widetilde{W}_{\mathrm{bb}}(L_+) - q^{-1/4}\,\widetilde{W}_{\mathrm{bb}}(L_-) = \big(q^{1/2}-q^{-1/2}\big)\widetilde{W}_{\mathrm{bb}}(L_0). \tag{12.291}$$

The subscript "bb" is explained in Section 12.14.16: it means the links carry the "blackboard" framing of the diagram.

Key idea: Why the coefficients are universal

The coefficients in (12.290) are the trace and determinant of σ, i.e. symmetric functions of its eigenvalues. They depend only on the representation labels on the four punctures and on k, *not* on the vectors V_2^0, V_1, and so not on the rest of the link. This is what turns "three vectors in $\mathbb{C}^2$ are dependent" into a *universal* skein relation. It also means the relation holds in *any* 3-manifold M, not only S^3.

From blackboard framing to Nash's coefficients

In L_+ and L_- the local crossing contributes ± 1 to the writhe of the diagram, relative to L_0. The framing of a Wilson line (below) multiplies $\widetilde{W}$ by $\theta := e^{2\pi i h_{1/2}} = q^{3/4}$ per unit of framing, and the blackboard framing of a diagram D has total framing equal to its writhe $w(D)$. Define the **writhe-normalised** quantity

$$X(D) := \theta^{-w(D)}\,\widetilde{W}_{\mathrm{bb}}(D),$$

which is independent of the diagram. Then $w(L_\pm) = w(L_0) \pm 1$ and (12.291) becomes

$$\begin{aligned} 0 &\overset{(1)}{=} q^{1/4}\theta^{w_0+1}X(L_+) - q^{-1/4}\theta^{w_0-1}X(L_-) - (q^{1/2}-q^{-1/2})\theta^{w_0}X(L_0) \\ &\overset{(2)}{=} q\,X(L_+) - q^{-1}X(L_-) - (q^{1/2}-q^{-1/2})X(L_0). \end{aligned} \tag{12.292}$$

(1) Substitute $\widetilde{W}_{\mathrm{bb}}(D) = \theta^{w(D)}X(D)$, with $w_0 = w(L_0)$.
(2) Divide by θ^{w_0} and use $q^{1/4}\theta = q^{1/4}q^{3/4} = q$.

Compare Nash's coefficients (12.135),

$$\begin{aligned} \alpha^+ &= -\exp\!\left[\frac{2\pi i}{k+2}\right] = -q, \qquad \alpha^- = \exp\!\left[-\frac{2\pi i}{k+2}\right] = q^{-1}, \\ \alpha^0 &= \exp\!\left[\frac{\pi i}{k+2}\right] - \exp\!\left[-\frac{\pi i}{k+2}\right] = q^{1/2}-q^{-1/2}. \end{aligned} \tag{12.293}$$

Then $\alpha^+X(L_+) + \alpha^-X(L_-) + \alpha^0X(L_0) = 0$ is exactly -1 times (12.292). The "inessential overall constant" Nash mentions is this -1, and the "change of framing" in his footnote 9 is our passage from $\widetilde{W}_{\mathrm{bb}}$ to X. Now introduce Nash's variable

$$t := \exp\!\left[-\frac{2\pi i}{k+2}\right] = q^{-1}, \qquad t^{1/2} := -q^{-1/2} = -e^{-\pi i/(k+2)}. \tag{12.294}$$

With this square root, $t^{1/2} - t^{-1/2} = -q^{-1/2} + q^{1/2} = \alpha^0$, and (12.292) reads

$$\boxed{\,t^{-1}X(L_+) - t\,X(L_-) = \left(t^{1/2}-t^{-1/2}\right)X(L_0).\,} \tag{12.295}$$

This is the Jones skein relation (12.279), now for the path integral.

Watch out: Nash's (12.137): which crossing is which, and which square root

Nash writes the final relation as $t\widetilde{W}(M,\{R^+\}) - t^{-1}\widetilde{W}(M,\{R^-\}) + (t^{1/2}-t^{-1/2})\widetilde{W}(M,\{R^0\}) = 0$, with t multiplying L_+. Substituting his own (12.293) and (12.294) gives $-t^{-1}X_+ + tX_- + \alpha^0X_0 = 0$, i.e. t multiplies L_-. The two versions differ by the exchange $L_+ \leftrightarrow L_-$, equivalently by replacing every link by its mirror image, or t by t^{-1}. His (12.118) has the same interchange

relative to the standard (12.279). Since the unknot is its own mirror image, none of this affects the unknot and unlink values below; it does decide whether the right-handed trefoil gets $t + t^3 - t^4$ or $t^{-1} + t^{-3} - t^{-4}$. Separately, $t^{1/2}$ must be the square root $-e^{-\pi i/(k+2)}$, *not* the principal root $e^{-\pi i/(k+2)}$, for $\alpha^0 = t^{1/2} - t^{-1/2}$ to hold and for the unknot value (12.138) to agree with the surgery value (12.158).

The same relation as a Kauffman bracket

There is a second way to read σ. Let U be the operator that projects the two exchanged strands onto their singlet and multiplies by [2]: in the s-channel basis, $U = \mathrm{diag}([2], 0)$. Pictorially U is a "cap" followed by a "cup": the two strands annihilate into the vacuum and are recreated, and $U^2 = [2]\,U$ because the closed loop in the middle is an unknot, worth [2]. Then

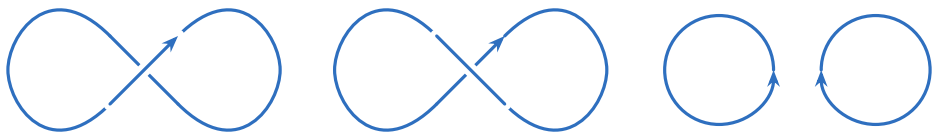
$$\sigma \overset{(1)}{=} q^{1/4}\,\mathrm{id} - q^{-1/4}\,U, \tag{12.296}$$

(1) Check on the basis: on the triplet $U = 0$ and the right side is $q^{1/4} = \lambda_1$; on the singlet it is $q^{1/4} - q^{-1/4}(q^{1/2} + q^{-1/2}) = -q^{-3/4} = \lambda_0$.

In words: *a crossing equals $q^{1/4}$ times the vertical smoothing minus $q^{-1/4}$ times the horizontal smoothing.* With $A := q^{1/4}$ and $U_K := -U$ this is Kauffman's bracket relation $\langle\sigma\rangle = A\langle\,)(\,\rangle + A^{-1}\langle\asymp\rangle$ with loop value $\delta = -A^2 - A^{-2} = -[2]$, met in the discussion of knot polynomials earlier in this chapter. The standard dictionary $t = A^{-4}$ then gives $t = q^{-1}$, in agreement with (12.294). So Chern–Simons theory at level k *is* the Kauffman bracket evaluated at $A = e^{\pi i/(2(k+2))}$.

12.14.14 The unknot and the Jones polynomial

We now compute $W(S^3, \{R\})$ when $\{R\}$ is a single unknotted circle, following Nash (12.138). Draw the unknot with one crossing, a single kink (Figure 12.75). Then L_+ is the unknot, L_- is again the unknot (a kink the other way), and L_0 is two unlinked circles.



L_+: unknot L_-: unknot L_0: two circles

Figure 12.75: The skein triple for the unknot. A figure-of-eight curve with one crossing is an unknot whichever strand is on top (Reidemeister I untwists it). Smoothing the crossing according to the orientation splits it into two disjoint circles. So $L_+ = L_- = \bigcirc$ and $L_0 = \bigcirc\,\bigcirc$.

Divide the skein relation by $Z(S^3)$, so that it holds for the normalised correlators $W = \widetilde{W}/Z$, and use the unlink rule (12.278): $W(L_0) = W(\bigcirc)^2$. In Nash's form (12.137) (where $L_\pm$ are interchanged, which does not matter here because $L_+ = L_-$):

$$\begin{aligned} 0 &\overset{(1)}{=} t\,W(\bigcirc) - t^{-1}W(\bigcirc) + \big(t^{1/2} - t^{-1/2}\big)W(\bigcirc)^2 \\ &\overset{(2)}{=} W(\bigcirc)\Big[(t - t^{-1}) + \big(t^{1/2} - t^{-1/2}\big)W(\bigcirc)\Big], \\ W(\bigcirc) &\overset{(3)}{=} -\frac{t - t^{-1}}{t^{1/2} - t^{-1/2}} \overset{(4)}{=} -\big(t^{1/2} + t^{-1/2}\big) \overset{(5)}{=} q^{1/2} + q^{-1/2} = [2]_q. \end{aligned} \tag{12.297}$$

(1) The skein relation with $L_+ = L_- = \bigcirc$, $L_0 = \bigcirc\,\bigcirc$, and (12.278) with $p = 2$.

(2) Factor out $W(\bigcirc)$.
(3) $W(\bigcirc) \neq 0$ (a vanishing unknot would make every link vanish), so the bracket vanishes.
(4) $t - t^{-1} = (t^{1/2} - t^{-1/2})(t^{1/2} + t^{-1/2})$.
(5) $t^{1/2} = -q^{-1/2}$ from (12.294), so $-(t^{1/2} + t^{-1/2}) = q^{-1/2} + q^{1/2}$.

So the unknot has

$$W(\bigcirc) = [2]_q = 2\cos\frac{\pi}{k+2} \xrightarrow{k\to\infty} 2.$$

At infinite level this is the dimension of the representation: a Wilson loop in a weakly fluctuating field is $\mathrm{tr}_R(\mathrm{id}) = \dim R = 2$. The **quantum dimension** $[2]_q$ is its finite-k deformation.

For a p-component unlink, (12.278) gives Nash's (12.139):

$$W\big(S^3, \text{unlink of } p\big) = \left\{-\frac{t - t^{-1}}{t^{1/2} - t^{-1/2}}\right\}^p = [2]_q^p. \tag{12.298}$$

Dividing by one unknot factor defines the knot polynomial:

$$V_L(t) \;:=\; -\frac{t^{1/2} - t^{-1/2}}{t - t^{-1}}\, W(S^3, \{R\}) \;=\; \frac{W(S^3, \{R\})}{W(\bigcirc)}. \tag{12.299}$$

(For links with framing, W here means the writhe-normalised $X/Z(S^3)$.)

Key idea: Witten's theorem (Nash's (12.140))

$V_L(t)$ satisfies the Jones skein relation (it is W divided by a constant, and the skein relation is linear), and it takes the value $V_{\text{unlink of } p} = [2]^{p-1} = \big(-(t^{1/2} + t^{-1/2})\big)^{p-1}$ on unlinks, which is Nash's (12.83). Since the skein relation and the unlink values determine the Jones polynomial uniquely (any link can be reduced to unlinks by crossing changes and smoothings), $V_L(t)$ *is* the Jones polynomial, evaluated at $t = e^{-2\pi i/(k+2)}$. Knowing it at infinitely many roots of unity, we may "analytically continue" to all t: a Laurent polynomial is determined by its values at infinitely many points.

Worked example 12.108: The Hopf link by the skein relation

The positive Hopf link H_+ is two circles linked once, with both crossings positive. Change one crossing: the circles come apart, $L_- =$ unlink of 2. Smooth it instead: the two circles merge into one unknot, $L_0 = \bigcirc$. With $V(\bigcirc) = 1$ and $V(\bigcirc\,\bigcirc) = \delta := -(t^{1/2} + t^{-1/2})$:

$$\begin{aligned} t^{-1}V(H_+) &\overset{(1)}{=} t\,\delta + (t^{1/2} - t^{-1/2})\cdot 1 \\ &\overset{(2)}{=} -t^{3/2} - t^{1/2} + t^{1/2} - t^{-1/2} = -t^{3/2} - t^{-1/2}, \\ V(H_+) &\overset{(3)}{=} -t^{5/2} - t^{1/2}. \end{aligned}$$

(1) The skein relation (12.295) for V, solved for the L_+ term.
(2) Expand $t\delta = -t^{3/2} - t^{1/2}$.
(3) Multiply by t.

In terms of q (with $t^{1/2} = -q^{-1/2}$): $V(H_+) = q^{-1/2} + q^{-5/2}$. We will reproduce this from the modular S-matrix in Example 12.127.

Worked example 12.109: The right-handed trefoil by the skein relation

The right-handed trefoil T is the closure of the braid σ_1^3 (Figure 12.76); all three crossings are positive. Pick one. Changing it gives $\sigma_1^2\sigma_1^{-1} = \sigma_1$, whose closure is an unknot; smoothing it gives the closure of σ_1^2, the positive Hopf link. So $L_+ = T$, $L_- = \bigcirc$, $L_0 = H_+$:

$$\begin{aligned} t^{-1}V(T) &\overset{(1)}{=} t\cdot 1 + (t^{1/2} - t^{-1/2})\bigl(-t^{5/2} - t^{1/2}\bigr) \\ &\overset{(2)}{=} t - t^3 - t + t^2 + 1 = 1 + t^2 - t^3, \\ V(T) &\overset{(3)}{=} t + t^3 - t^4. \end{aligned}$$

(1) Skein relation with $V(\bigcirc) = 1$ and the Hopf value from Example 12.108.
(2) $(t^{1/2} - t^{-1/2})(-t^{5/2} - t^{1/2}) = -t^3 - t + t^2 + 1$.
(3) Multiply by t.

This is Nash's (12.76). The mirror image (left-handed trefoil) has all crossings negative and gives $V(t^{-1}) = t^{-1} + t^{-3} - t^{-4}$: the Jones polynomial detects chirality.

Try it: The figure-eight knot

In the standard 4-crossing diagram of the figure-eight knot 4_1, pick a positive crossing. Changing it unknots the knot; smoothing it gives the *negative* Hopf link, with $V(H_-) = -t^{-1/2} - t^{-5/2}$ (the mirror of Example 12.108). Find V_{4_1}.
Answer. $t^{-1}V = t + (t^{1/2} - t^{-1/2})(-t^{-1/2} - t^{-5/2}) = t - 1 + t^{-1} - t^{-2} + t^{-3}$, so $V_{4_1} = t^2 - t + 1 - t^{-1} + t^{-2}$. It is symmetric under $t \to t^{-1}$, as it must be: the figure-eight knot is isotopic to its mirror image.

12.14.15 A second route: eigenvalues and closed braids

The skein relation reduces everything to unlinks, but there is a more direct computation for links that are closures of 2-strand braids, which uses the eigenvalues themselves. Close up the braid σ^n by joining the top ends to the bottom ends (Figure 12.76). For $n = 0$ we get two unlinked circles, for $n = \pm 1$ an unknot with a kink, for $n = 2$ the Hopf link, $n = 3$ the trefoil, $n = 5$ the cinquefoil (the $(2, 5)$ torus knot). In general the closure of σ^n is the $(2, n)$ **torus link** $T(2, n)$, a knot for n odd and a two-component link for n even.

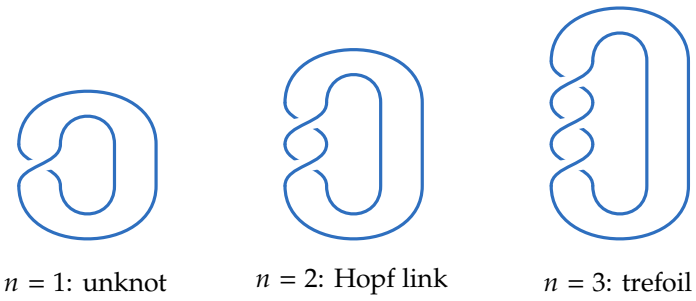


Figure 12.76: Closures of the 2-strand braids σ^n (crossings stacked vertically, closing strands on the right). $n = 1$ is an unknot with one kink, $n = 2$ the Hopf link, $n = 3$ the right-handed trefoil. The link lies inside a solid torus (a thickened circle around which both strands wind), and the braid acts diagonally on the fusion channels of the two strands.

The quantum trace

The closed braid lies inside a solid torus, the neighbourhood of the circle around which it winds. Think of the two strands as a single composite line. In the fusion channel j the pair behaves as one Wilson line of spin j running round the solid torus, and the braid σ^n merely multiplies that channel by λ_j^n. Summing over channels and closing up,

$$W_{\mathrm{bb}}\big(\widehat{\sigma^n}\big) \;=\; \sum_{j=0,1} \lambda_j^{\,n}\, W\big(\bigcirc_j\big) \;=\; \lambda_0^{\,n} + [3]\,\lambda_1^{\,n}, \tag{12.300}$$

where $W(\bigcirc_j)$ is the unknot carrying spin j. We will show in Section 12.15.6 that $W(\bigcirc_j) = S_{0,2j}/S_{00} = [2j+1]$, so $W(\bigcirc_0) = 1$ and $W(\bigcirc_1) = [3]$. This formula is a **quantum trace**: an ordinary trace $\sum_j (\dim j)\lambda_j^n$ with the dimensions replaced by quantum dimensions. At $q = 1$ it is the trace of P^n on $\mathbb{C}^2 \otimes \mathbb{C}^2$, which is 4 for n even and 2 for n odd.

Worked example 12.110: Checks: $n = 0$ and $n = 2$

For $n = 0$ the closure is two unlinked unknots:

$$W_{\mathrm{bb}}(\widehat{\sigma^0}) \overset{(1)}{=} 1 + [3] \overset{(2)}{=} [2]^2 = W(\bigcirc)^2.$$

(1) (12.300) with $n = 0$.
(2) $[2]^2 = q + 2 + q^{-1} = [3] + 1$.

In agreement with the unlink rule. For $n = 2$ (the Hopf link; the two crossings are between different components, so each component has zero self-writhe and the blackboard framing is the zero framing):

$$W_{\mathrm{bb}}(\widehat{\sigma^2}) \overset{(3)}{=} q^{-3/2} + [3]\,q^{1/2} \overset{(4)}{=} q^{-3/2} + q^{3/2} + q^{1/2} + q^{-1/2} \overset{(5)}{=} [4].$$

(3) $\lambda_0^2 = q^{-3/2}$, $\lambda_1^2 = q^{1/2}$.
(4) $[3]q^{1/2} = (q + 1 + q^{-1})q^{1/2}$.
(5) $[4] = (q^2 - q^{-2})/(q^{1/2} - q^{-1/2}) = q^{3/2} + q^{1/2} + q^{-1/2} + q^{-3/2}$.

So the zero-framed Hopf link has $W = [4]$; we will meet this again as S_{11}/S_{00}.

Worked example 12.111: All $(2, n)$ torus knots at once

Let n be odd. The diagram of $\widehat{\sigma^n}$ has writhe n, so the writhe-normalised value is $\theta^{-n} W_{\mathrm{bb}}$ with $\theta = q^{3/4}$. Write $x := q^{-1} = t$. Then

$$\begin{aligned} V_{T(2,n)} &\overset{(1)}{=} \frac{\theta^{-n}\big(\lambda_0^n + [3]\lambda_1^n\big)}{[2]} \overset{(2)}{=} \frac{-q^{-3n/2} + [3]\,q^{-n/2}}{[2]} \\ &\overset{(3)}{=} \frac{x^{1/2}\big[\,x^{n/2}(x^{-1} + 1 + x) - x^{3n/2}\big]}{1 + x} \overset{(4)}{=} x^{\frac{n-1}{2}}\,\frac{1 + x + x^2 - x^{n+1}}{1 + x}. \end{aligned}$$

(1) $V = W/W(\bigcirc)$ with $W(\bigcirc) = [2]$, after writhe normalisation.
(2) $\theta^{-n}\lambda_0^n = q^{-3n/4}(-1)^n q^{-3n/4} = -q^{-3n/2}$ for n odd; $\theta^{-n}\lambda_1^n = q^{-3n/4}q^{n/4} = q^{-n/2}$.
(3) Substitute $q = x^{-1}$; $[3] = x^{-1} + 1 + x$ and $[2] = x^{-1/2}(1 + x)$, so dividing by $[2]$ brings a factor $x^{1/2}$.
(4) Collect powers: $x^{1/2}x^{n/2}x^{-1} = x^{(n-1)/2}$.

Since $1-x^{n+1}=(1+x)(1-x+x^2-\cdots-x^n)$ for n odd, the fraction is a polynomial:

$$\frac{1+x+x^2-x^{n+1}}{1+x}=\frac{(1-x^{n+1})+x(1+x)}{1+x}=1+x^2-x^3+x^4-\cdots-x^n.$$

Hence

$$V_{T(2,3)}=t+t^3-t^4,\qquad V_{T(2,5)}=t^2+t^4-t^5+t^6-t^7,\qquad V_{T(2,n)}=t^{\frac{n-1}{2}}\left(1+t^2-t^3+\cdots-t^n\right).$$

The trefoil agrees with the skein computation of Example 12.109: two independent routes through the same Hilbert space.

12.14.16 Framing

Why Wilson lines need a framing

In the abelian theory we saw that the self-linking of a Wilson loop is ill-defined: the Gauss integral $\oint\oint$ over the same curve twice is singular at coincident points, and has to be regularised by pushing one copy off the curve along a normal vector field. The choice of push-off is a **framing**: a ribbon whose core is the knot. Different framings differ by an integer, the number of full twists of the ribbon around the core (Figure 12.77). The same singularity occurs in the non-abelian theory, so $\widetilde{W}(M,\{R\})$ is really an invariant of *framed* links.

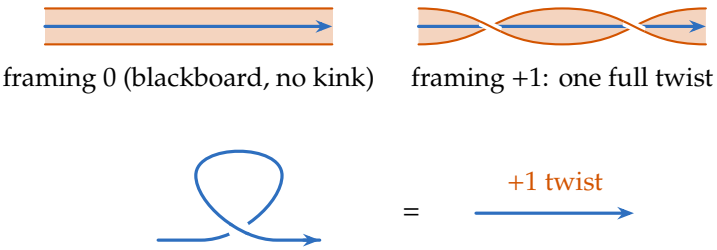


Figure 12.77: Framings. A framed knot is a thin ribbon; framings are counted by full twists. Drawing a knot on paper gives it the "blackboard" framing (the ribbon lies flat on the page); a kink in the drawing then contributes one full twist. This is why Reidemeister move I changes the blackboard framing, and why the writhe correction θ^{-w} is needed.

How the correlator changes: Nash's Dehn twist argument

Nash's argument is short and beautiful. Cut M along a surface Σ that the Wilson line C crosses at a single point P, so that $M=M_1\cup_\Sigma M_2$. Before regluing, apply to ∂M_1 a **Dehn twist** by 2π around P: rotate a small disc around P by a full turn, interpolating to the identity on an annulus around it. The knot is unchanged as a curve, but its ribbon has been twisted once around the core: the framing has changed by 1. Doing it τ times changes the framing by τ.

On the Hilbert space this twist acts as a linear map on $\mathcal{H}_{\Sigma;\ldots,R,\ldots}$. In conformal field theory a 2π rotation around the insertion point of a primary field of weight h_R is generated by $L_0-\bar{L}_0$ and, on the chiral blocks, multiplies by $e^{2\pi i h_R}$: the primary field transforms as $\phi(e^{2\pi i}z)=e^{2\pi i h_R}\phi(z)$ under the rotation, a consequence of ϕ having spin h_R. So

$$\widetilde{W}\big(M,\{R\}\big)\;\longmapsto\;e^{2\pi i\tau h_R}\,\widetilde{W}\big(M,\{R\}\big)\qquad\text{(framing of }C\text{ changed by }\tau).\tag{12.301}$$

For the fundamental representation,

$$\theta := e^{2\pi i h_{1/2}} = \exp\!\Big(2\pi i\,\frac{3/4}{k+2}\Big) = q^{3/4}.$$

In the classical limit $\theta \to 1$: framing dependence is a quantum effect, of order $1/k$, exactly like the self-linking ambiguity in the abelian theory.

Worked example 12.112: The kink: eigenvalues and conformal weights agree

Here is a strong consistency check between two pieces of CFT data that we used independently. The closure of σ^1 is an unknot with a single positive kink, i.e. an unknot with framing +1. By the framing rule its value should be $\theta\, W(\bigcirc) = q^{3/4}[2]$. By the quantum trace (12.300),

$$\begin{aligned} W_{\mathrm{bb}}(\widehat{\sigma}) &\overset{(1)}{=} \lambda_0 + [3]\lambda_1 \overset{(2)}{=} -q^{-3/4} + q^{1/4}(q+1+q^{-1}) \\ &\overset{(3)}{=} q^{5/4} + q^{1/4} \overset{(4)}{=} q^{3/4}\big(q^{1/2}+q^{-1/2}\big) = \theta\,[2]. \end{aligned}$$

(1) (12.300) with $n = 1$.
(2) The eigenvalues (12.287) and $[3] = q + 1 + q^{-1}$.
(3) The terms $-q^{-3/4}$ and $q^{1/4}q^{-1}$ cancel.
(4) Factor out $q^{3/4}$.

The braiding eigenvalues (fusion channels) and the twist (conformal weight of a single line) are consistent. Had we chosen the opposite sense of rotation for the positive crossing, we would have found $\theta^{-1}[2]$: this is the check that fixed the convention in (12.287). It is an instance of the general **ribbon** identity $\sum_j \lambda_j\,[2j+1] = \theta_{1/2}\,[2]$.

Worked example 12.113: The unknot again, in blackboard framing

In (12.297) we treated L_+ and L_- as the same unknot. In blackboard framing they are not the same framed knot: $L_\pm$ is the unknot with framing ± 1, worth $\theta^{\pm 1}W(\bigcirc)$. Nash's footnote 9 warns that the calculation "involves changing the framing". Let us redo it with (12.291), which is valid for blackboard framings:

$$\begin{aligned} 0 &\overset{(1)}{=} q^{1/4}\theta\, W(\bigcirc) - q^{-1/4}\theta^{-1}W(\bigcirc) - (q^{1/2}-q^{-1/2})W(\bigcirc)^2 \\ &\overset{(2)}{=} W(\bigcirc)\big[(q-q^{-1}) - (q^{1/2}-q^{-1/2})W(\bigcirc)\big], \\ W(\bigcirc) &\overset{(3)}{=} \frac{q-q^{-1}}{q^{1/2}-q^{-1/2}} = q^{1/2}+q^{-1/2} = [2]. \end{aligned}$$

(1) (12.291) divided by $Z(S^3)$, with the framed values of $L_\pm$ and the unlink rule for L_0 (two zero-framed circles).
(2) $q^{1/4}\theta = q$ and $q^{-1/4}\theta^{-1} = q^{-1}$.
(3) Divide by $W(\bigcirc) \neq 0$.

The same answer: the framing bookkeeping and the writhe normalisation are equivalent.

Dictionary: framed invariants and the Jones polynomial

Chern–Simons, $\mathrm{SU}(2)_k$, spin $\frac{1}{2}$	Knot theory
$q = e^{2\pi i/(k+2)}$, $t = q^{-1}$	Jones variable t; Kauffman $A = q^{1/4}$
$W(\bigcirc) = [2]_q$	loop value $-(t^{1/2} + t^{-1/2}) = -A^2 - A^{-2}$ (up to sign)
twist $\theta = e^{2\pi i h_{1/2}} = q^{3/4}$	$-A^3$ up to sign: Reidemeister I factor
$\widetilde{W}_{\mathrm{bb}}(D)/Z(S^3)$	framed invariant (Kauffman bracket)
$\theta^{-w(D)}\widetilde{W}_{\mathrm{bb}}(D)/(Z(S^3)[2])$	$V_L(t)$
zero-framed W of a link	$V_L(t)\cdot[2]\cdot\theta^{2\sum_{a<b}\mathrm{lk}(C_a,C_b)}$

Reading Nash (pp. 353–355, (12.130)–(12.140) and the framing paragraph)

Nash's route: four punctures, three configurations (12.130), three pairings (12.131), linear dependence (12.132); the braiding matrix B with $V_2^0 = BV_2^+$, $V_2^- = B^2V_2^+$ (12.133)–(12.134), with the α's (12.135) taken from Moore–Seiberg; $t = e^{-2\pi i/(k+2)}$ (12.136); the skein relation for $\widetilde{W}$ (12.137); the unknot (12.138), unlinks (12.139), and the identification with the Jones polynomial (12.140). The framing paragraph at the end of §6 says that a $2\pi\tau$ Dehn twist around P changes the framing by τ and multiplies $\widetilde{W}$ by $e^{2\pi i\tau h_R}$. We have supplied: the eigenvalues from the OPE, explicit σ and F matrices, the braid-relation check, Cayley–Hamilton as the source of the α's, the framing correction that turns our coefficients into Nash's, the Kauffman-bracket form, and the Hopf link, trefoil, figure-eight and $(2,n)$ torus knot computations.

Recap

- A crossing sits in a ball whose boundary is a 4-punctured sphere, $\dim \mathcal{H}_{S^2,4} = 2$; three vectors in $\mathbb{C}^2$ are dependent.
- The half-twist σ acts on $\mathcal{H}_{S^2,4}$ with eigenvalues $-q^{-3/4}$ (singlet) and $q^{1/4}$ (triplet), read off from the OPE; Cayley–Hamilton gives $q^{1/4}\widetilde{W}_{\mathrm{bb}}(L_+) - q^{-1/4}\widetilde{W}_{\mathrm{bb}}(L_-) = (q^{1/2} - q^{-1/2})\widetilde{W}_{\mathrm{bb}}(L_0)$.
- Writhe normalisation ($\theta = q^{3/4}$ per unit framing) turns this into the Jones skein relation with $t = q^{-1}$, $t^{1/2} = -q^{-1/2}$.
- $W(\bigcirc) = [2]$, unlinks $[2]^p$; $V_L = W/W(\bigcirc)$ is the Jones polynomial: Hopf $-t^{1/2} - t^{5/2}$, trefoil $t + t^3 - t^4$, 4_1: $t^2 - t + 1 - t^{-1} + t^{-2}$.
- Closed 2-braids: $W_{\mathrm{bb}}(\widehat{\sigma^n}) = \lambda_0^n + [3]\lambda_1^n$.
- Changing the framing by τ multiplies $\widetilde{W}$ by $e^{2\pi i\tau h_R}$ (Dehn twist around the puncture).

12.15 Surgery and the Jones polynomial

Jones defined his polynomial for links in $\mathbb{R}^3$ or S^3. The knot diagram on a sheet of paper is a picture in $\mathbb{R}^2$, and the whole combinatorial machinery of Reidemeister moves lives there. Quantum field theory does not care about the paper. The Chern–Simons path integral

$$\widetilde{W}\big(M,\{R\}\big) = \int_{\mathcal{A}/\mathcal{G}} \mathcal{D}A\; e^{iS_k[A]} \prod_a W(R_a, C_a)$$

makes sense for a link in *any* closed oriented 3-manifold M. So we immediately get a generalisation of the Jones polynomial to links in M, and, taking the empty link, a number $Z(M)$ attached to M itself. The question of this section is: *how do we compute these numbers?*

The gluing rules of Section 12.14.8 reduce the question to two things: a supply of simple pieces whose vectors we know, and a way of assembling an arbitrary M from them. The pieces will be *solid tori*; the assembly instruction is **surgery**. Topologists proved long ago that every closed oriented 3-manifold can be obtained from S^3 by removing a few solid tori and gluing them back differently. Witten used this in two ways, both of which Nash describes: to define and compute the invariants on any M, and to recompute the Jones polynomial in S^3 itself, this time from the modular S-matrix rather than from the braiding.

In this section G can be any compact simple Lie group; we return to SU(2) when we need explicit numbers.

12.15.1 Solid tori, meridians and longitudes

A **solid torus** is $D^2 \times S^1$: a disc swept round a circle, the shape of a doughnut or an inner tube including the rubber. Its boundary is the ordinary torus $T^2 = S^1 \times S^1$. On the boundary there are two distinguished kinds of closed curve (Figure 12.78).

Working definition 12.41 (meridian and longitude)

Let $V = D^2 \times S^1$, with $D^2 = \{|z| \le 1\}$ and $S^1 = \{e^{i\phi}\}$.

- A **meridian** is the curve $m = \partial D^2 \times \{\text{pt}\}$. Test: it bounds a disc *inside* V, so it is contractible in V [it can be shrunk to a point without leaving V]. Up to orientation and isotopy it is the unique such curve.
- A **longitude** is a curve $l = \{\text{pt}\} \times S^1$ on ∂V. Test: it meets the meridian exactly once, and it is homotopic in V to the core circle $\{0\} \times S^1$ [it can be deformed into the core circle without leaving V], which generates $\pi_1(V) = \mathbb{Z}$.

The meridian is intrinsic to V. The longitude is *not* unique: $l + n\,m$ (go once along, and n times around) is an equally good longitude for every $n \in \mathbb{Z}$. Choosing a longitude is exactly the same as choosing a **framing** of the core circle: the longitude is the edge of the ribbon.

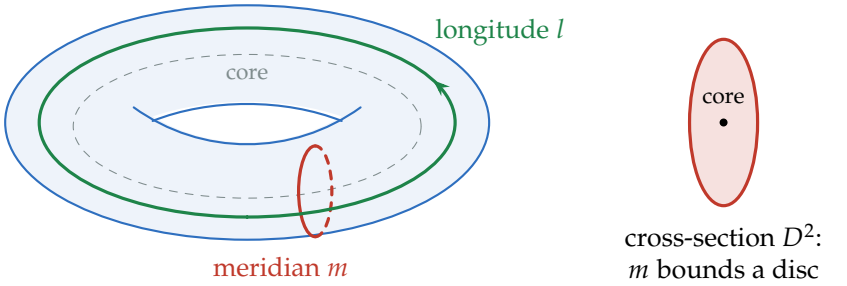


Figure 12.78: A solid torus $V = D^2 \times S^1$. The red meridian m goes once around the tube and bounds a disc inside V (right). The green longitude l goes once along the tube, parallel to the dashed core circle. In V the meridian is contractible and the longitude generates $\pi_1(V) = \mathbb{Z}$. Any closed curve on ∂V is homologous to $a\,m + b\,l$ for integers a, b.

Diffeomorphisms of the torus are $\mathrm{SL}(2,\mathbb{Z})$ matrices

Every closed curve on T^2 is homologous to $a\,m + b\,l$ with $a, b \in \mathbb{Z}$: $H_1(T^2;\mathbb{Z}) = \mathbb{Z}^2$ with basis (m, l). A diffeomorphism K of T^2 maps curves to curves, hence acts on H_1 by an integer 2×2 matrix. It has an integer inverse and it preserves orientation, so its determinant is $+1$. Conversely, every such matrix is realised by a linear map of the square $\mathbb{R}^2/\mathbb{Z}^2$, and two diffeomorphisms that induce the same matrix are isotopic. So

$$\{\text{orientation-preserving diffeos of } T^2\}/\text{isotopy} \;=\; \mathrm{SL}(2,\mathbb{Z}), \tag{12.302}$$

where

$$\mathrm{SL}(2,\mathbb{Z}) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} : \; a,b,c,d \in \mathbb{Z},\; ad - bc = 1 \right\}.$$

This is the **mapping class group** of the torus. It is the same $\mathrm{SL}(2,\mathbb{Z})$ that acts on the modular parameter τ of a torus in Chapter 5 and Chapter 11, via $\tau \mapsto (a\tau + b)/(c\tau + d)$; the matrix $-\mathbb{1}$ acts trivially on τ, which is why Nash writes $\mathrm{SL}(2,\mathbb{Z})/\{\pm\mathbb{1}\}$.

Working definition 12.42 (the generators S and T)

Acting on column vectors of coefficients in the basis (m, l):

$$S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} : \; m \mapsto l,\; l \mapsto -m; \qquad T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} : \; m \mapsto m,\; l \mapsto l + m.$$

S is a quarter turn of the square: it swaps the two cycles (up to a sign that keeps the orientation). T is a **Dehn twist** along the meridian: cut the torus along m, twist one side by a full turn, reglue. Curves crossing m pick up an extra copy of m. On τ they act as $S : \tau \mapsto -1/\tau$ and $T : \tau \mapsto \tau + 1$. They generate $\mathrm{SL}(2,\mathbb{Z})$ [every matrix in $\mathrm{SL}(2,\mathbb{Z})$ is a product of S, T and their inverses] and obey

$$S^2 = (ST)^3 = -\mathbb{1}, \qquad S^4 = \mathbb{1}. \tag{12.303}$$

Let us check the relations by multiplying out; this is the first place where the framing anomaly of Section 12.15.7 will later hide.

$$S^2 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \overset{(1)}{=} \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix},$$

$$ST = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \overset{(2)}{=} \begin{pmatrix} 0 & -1 \\ 1 & 1 \end{pmatrix}, \qquad (ST)^2 \overset{(3)}{=} \begin{pmatrix} -1 & -1 \\ 1 & 0 \end{pmatrix},$$

$$(ST)^3 \overset{(4)}{=} \begin{pmatrix} -1 & -1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 0 & -1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}.$$

(1) Row times column: the $(1,1)$ entry is $0 \cdot 0 + (-1) \cdot 1 = -1$, and so on.
(2) $(1,1)$: 0; $(1,2)$: $0 \cdot 1 + (-1) \cdot 1 = -1$; $(2,1)$: 1; $(2,2)$: $1 + 0 \cdot \ldots = 1$.
(3) Square the previous matrix: $(1,1) = 0 \cdot 0 + (-1) \cdot 1 = -1$, $(1,2) = 0 \cdot (-1) + (-1) \cdot 1 = -1$, $(2,1) = 0 + 1 = 1$, $(2,2) = -1 + 1 = 0$.
(4) Multiply once more.

An element of order six (ST has $(ST)^6 = \mathbb{1}$) is a rotation of the hexagonal lattice; S, of order four, is a rotation of the square lattice. These are the two special tori with extra symmetry, $\tau = i$ and $\tau = e^{2\pi i/3}$ (Figure 12.79).

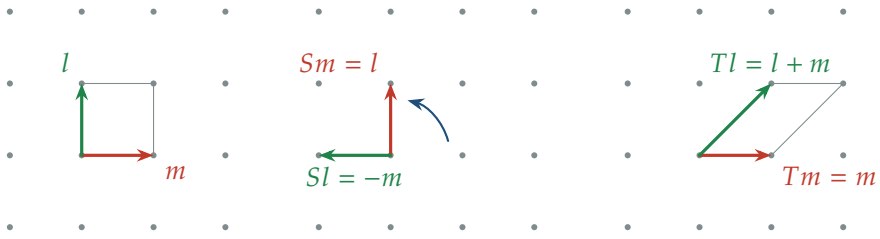


Figure 12.79: How S and T act on the lattice $H_1(T^2) = \mathbb{Z}^2$. The torus is the plane modulo this lattice; a closed curve is a lattice vector. S rotates the basis by a quarter turn, exchanging the roles of meridian and longitude. T shears it, adding one meridian to the longitude: this is the Dehn twist, and it is also a change of framing of the core by one unit. Both preserve the area of the fundamental cell (determinant 1).

Physical picture: A Dehn twist is a change of framing

Take the longitude l of a solid torus and apply T^n: it becomes $l + nm$, a curve that goes once along the tube while winding n times around it. As an edge of a ribbon around the core, this ribbon has n full twists. So *T acting on the boundary of a solid torus is the change of framing of its core circle by one unit.* This is the same Dehn twist that Nash used at the end of §6 to change the framing of a Wilson line (Section 12.14.16); now it acts on a whole solid torus.

12.15.2 Gluing two solid tori

The simplest closed 3-manifolds are made of two solid tori V_1, V_2 glued along their boundaries by a diffeomorphism $K : \partial V_2 \to \partial V_1$. The result depends only on the class of K in $\mathrm{SL}(2,\mathbb{Z})$, and in fact only on *where the meridian m_2 goes*: once the meridian disc of V_2 is glued in, what remains of V_2 is a 3-ball, and there is only one way to fill a spherical hole.

Working definition 12.43 (genus-one gluings)

Glue V_2 to V_1 so that $m_2 \mapsto q\, m_1 + p\, l_1$ with coprime integers p, q [no common factor other than ± 1, which is what lets $q\, m_1 + p\, l_1$ be drawn as one curve that does not cross itself]. Then $M = V_1 \cup_K V_2$ is

- $S^2 \times S^1$ if $m_2 \mapsto m_1$ ($p = 0$);
- S^3 if $m_2 \mapsto l_1$ ($p = 1$, $q = 0$), i.e. $K = S$;
- the **lens space** $L(p, q)$ in general.

Test: $\pi_1(M) = \mathbb{Z}_p$. Indeed $\pi_1(V_1) = \mathbb{Z}$ is generated by l_1 (the meridian m_1 is contractible in V_1), and gluing in the meridian disc of V_2 kills the class of $q\, m_1 + p\, l_1 = p\, l_1$ [that loop now bounds a disc, so it becomes the identity in $\pi_1(M)$]. So $\pi_1(M) = \mathbb{Z}/p\mathbb{Z}$.

S^3 is two solid tori

Nash's description of S^3 as two solid tori "linked together, as if belonging to a chain" becomes very concrete in complex coordinates. Write

$$S^3 = \{(z_1, z_2) \in \mathbb{C}^2 : \ |z_1|^2 + |z_2|^2 = 1\}$$

and split it according to which coordinate is larger:

$$V_1 = \{|z_1| \le |z_2|\}, \qquad V_2 = \{|z_1| \ge |z_2|\}, \qquad V_1 \cap V_2 = \{|z_1| = |z_2| = \tfrac{1}{\sqrt{2}}\} = T^2.$$

Each piece is a solid torus. In V_1 we have $|z_2|^2 \ge \frac{1}{2}$, so $z_2 \neq 0$ and $\arg z_2 \in S^1$ is a good circle coordinate, while $w = z_1/|z_2|$ ranges over the disc $|w| \le 1$:

$$V_1 \overset{(1)}{=} \left\{\left(w\, r,\ \sqrt{1 - r^2\,|w|^2}\ e^{i\phi_2}\right)\right\} \ \cong\ D^2_w \times S^1_{\phi_2},$$

(1) Here $r = |z_2|$ is fixed by $|z_1|^2 + |z_2|^2 = 1$ and $z_1 = wr$: $r^2(1 + |w|^2) = 1$. Every point of V_1 is given uniquely by (w, ϕ_2) with $|w| \le 1$.

The same with $1 \leftrightarrow 2$ for V_2. Now read off the curves on the common boundary torus $\{(e^{i\phi_1}, e^{i\phi_2})/\sqrt{2}\}$:

- the meridian of V_1 bounds the disc D^2_w, so it is the circle in which ϕ_1 varies: $m_1 = \{\phi_2 = \text{const}\}$;
- the meridian of V_2 is the circle in which ϕ_2 varies: $m_2 = \{\phi_1 = \text{const}\} = l_1$.

So the gluing identifies the meridian of each torus with the longitude of the other: $S^3 = V_1 \cup_S V_2$. The swap $(z_1, z_2) \mapsto (z_2, z_1)$ is an isometry of S^3 exchanging V_1 and V_2. Stereographic projection to $\mathbb{R}^3$ turns it into Nash's "inversion which interchanges the interior of T with its exterior": a round torus with radii (a, b) goes to one with radii (b, a). The two core circles $\{z_1 = 0\}$ and $\{z_2 = 0\}$ are the cores of the two tori, and they form a Hopf link (Figure 12.80).

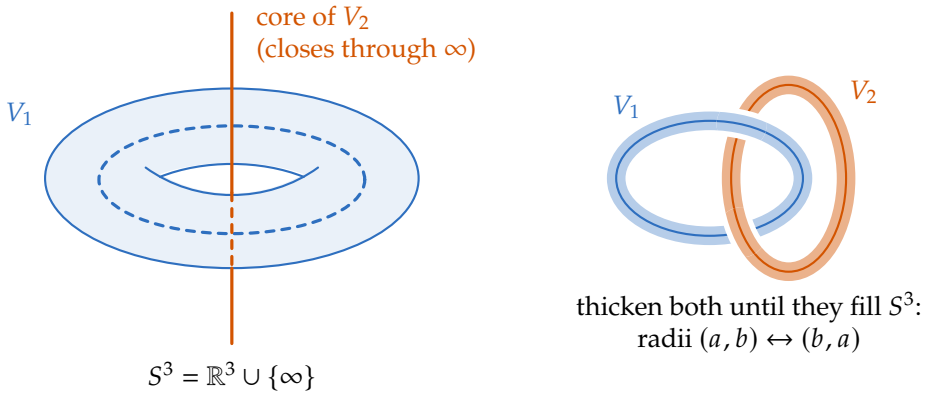


Figure 12.80: S^3 as two solid tori. Left: in $\mathbb{R}^3 \cup \{\infty\}$ the complement of a standard solid torus V_1 is another solid torus V_2 whose core (orange) runs through the hole and closes up through the point at infinity. Right: Nash's "chain" picture. The cores of V_1 and V_2 form a Hopf link; the meridian of each torus is the longitude of the other, so the gluing map is S.

$S^2 \times S^1$ is two solid tori

Take two discs D and D' and glue them along their boundary circles by the identity: the result is a sphere S^2 (two hemispheres). Multiply everything by a circle:

$$(D \times S^1) \cup_{\mathrm{id}} (D' \times S^1) = (D \cup D') \times S^1 = S^2 \times S^1.$$

Here the gluing map is the identity on $\partial D \times S^1$: it sends the meridian $\partial D'$ to the meridian ∂D (Figure 12.81). The cores $\{0\} \times S^1$ and $\{0'\} \times S^1$ are the circles $\{N\} \times S^1$ and $\{S\} \times S^1$ through the north and south poles; they are parallel and unlinked.

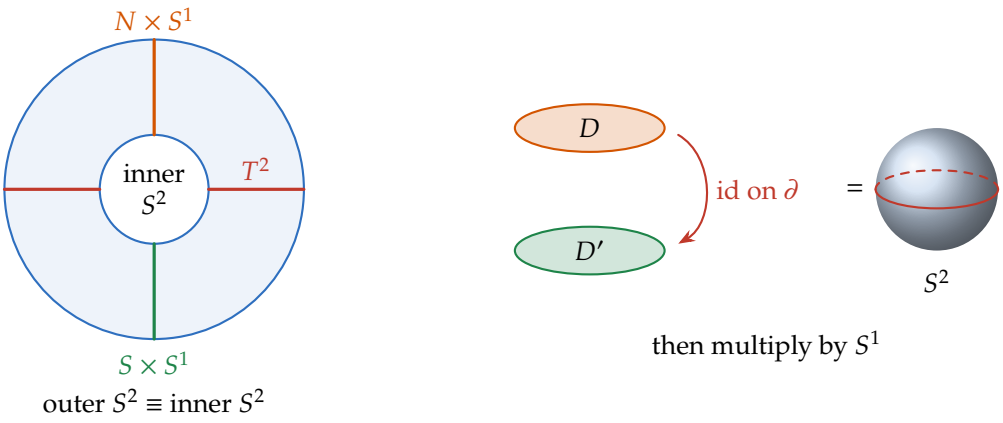


Figure 12.81: $S^2 \times S^1$ as two solid tori. Left: draw $S^2 \times S^1$ as a thick spherical shell whose inner and outer spheres are identified (the radial direction is the S^1). The red torus T^2 = equator $\times S^1$, which this slice cuts in two horizontal segments, splits it into $D \times S^1$ (containing the orange circle $N \times S^1$) and $D' \times S^1$ (containing the green circle $S \times S^1$). Right: $D \cup D'$ glued by the identity on their rims is a sphere. The gluing map on T^2 is the identity, so meridian goes to meridian.

Worked example 12.114: The fundamental groups tell the three apart

Use the test in the working definition.

- $S^2 \times S^1$: $m_2 \mapsto m_1$, which is already trivial in V_1, so nothing new is killed: $\pi_1 = \mathbb{Z}$, generated by the S^1 factor. Correct, since $\pi_1(S^2) = 0$.
- S^3: $m_2 \mapsto l_1$, the generator; $\pi_1 = \mathbb{Z}/1 = 0$.
- $L(2,1)$: $m_2 \mapsto m_1 + 2l_1$, so $\pi_1 = \mathbb{Z}_2$. This lens space is $\mathbb{RP}^3 = \mathrm{SO}(3)$: the famous fact that a 2π rotation is not contractible in SO(3) but a 4π rotation is.

More generally $L(p,1)$ is $S^3/\mathbb{Z}_p$, with $\mathbb{Z}_p$ acting by $(z_1, z_2) \mapsto (e^{2\pi i/p} z_1, e^{2\pi i/p} z_2)$, a free action; its π_1 is $\mathbb{Z}_p$ because S^3 is simply connected.

12.15.3 Dehn surgery

Now take *any* closed 3-manifold M and a knot C inside it. Nash's recipe (Figure 12.82):

(1) Thicken C into a tubular neighbourhood $M_2 \cong D^2 \times S^1$, a solid torus with core C.

(2) Cut it out. What is left, $M_1 = M \setminus \operatorname{int} M_2$, has boundary $\partial M_1 = \partial M_2 \cong T^2$.

(3) Glue M_2 back in, but first act on its boundary with a diffeomorphism K of T^2.

The result is a new closed 3-manifold,

$$M^K \;=\; M_1 \cup_{T^2} M_2^K. \tag{12.304}$$

This is Nash's (12.141). If K extends to a diffeomorphism of the solid torus (for instance if $K = T^n$, which only twists the tube), then $M^K \cong M$: nothing has changed. The interesting surgeries move the meridian.

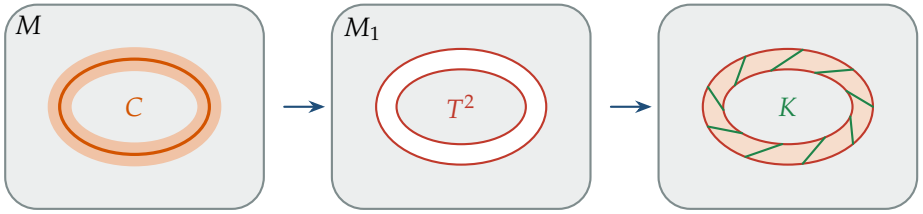


Figure 12.82: Surgery along a knot C (Nash's (12.141)). Remove a solid-torus neighbourhood of C, leaving M_1 with a torus boundary; glue the solid torus back after applying a diffeomorphism K of T^2 (the green "twisted" identification). The new manifold M^K depends only on where K sends the meridian of M_2.

Working definition 12.44 (Dehn surgery on a knot in S^3)

Let $C \subset S^3$ be a knot, μ the meridian of its neighbourhood [a small circle round C that bounds a disc pierced once by C] and λ the **Seifert longitude**: the longitude that has linking number 0 with C (the "zero framing"). For coprime integers p, q, the p/q **surgery** on C glues the solid torus back so that its meridian goes to $p\,\mu + q\,\lambda$. When $q = 1$ this is **integer surgery** with framing $n = p$: the meridian of the new solid torus goes to the framing curve $\lambda + n\mu$ of C. So an integer surgery is specified by a **framed knot**, and a surgery on a link by a **framed link**.

Worked example 12.115: Surgeries on the unknot

For the unknot $U \subset S^3$ the complement of the neighbourhood is the *other* solid torus V_1 of Figure 12.80, and its meridian is $m_1 = \lambda$ (the Seifert longitude of U bounds a disc in the complement) while its longitude is $l_1 = \mu$. The new meridian goes to

$$p\,\mu + q\,\lambda \overset{(1)}{=} q\,m_1 + p\,l_1,$$

(1) Substitute $\mu = l_1$, $\lambda = m_1$.

so, by the genus-one working definition, p/q surgery on the unknot is $L(p, q)$. Special cases:

- 0-surgery ($p = 0$, $q = 1$): $S^2 \times S^1$;
- ± 1-surgery: $L(\pm 1, 1) = S^3$; more generally $1/n$ surgery gives S^3 for every n;
- n-surgery: $L(n, 1)$, e.g. $\mathbb{RP}^3$ for $n = 2$.

The case 0 reversed is Nash's calculation: *starting* from $S^2 \times S^1$ and doing surgery with $K = S$ on the circle $\{N\} \times S^1$ produces S^3.

Worked example 12.116: The Poincaré sphere from the trefoil

Integer surgery with framing -1 on the left-handed trefoil produces the Poincaré homology sphere $\Sigma(2, 3, 5)$, the manifold met in the discussion of Floer homology earlier in this chapter. Why it has trivial homology is a one-line calculation: for integer surgery on a knot with framing n, $H_1(M) = \mathbb{Z}/n\mathbb{Z}$, because the only relation imposed on the meridian class μ

(which generates H_1 of the knot complement) is

$$0 \overset{(1)}{=} [\lambda + n\mu] \overset{(2)}{=} n[\mu] \quad \text{in } H_1,$$

(1) The new meridian disc kills the framing curve.
(2) The Seifert longitude bounds a surface in the complement, so $[\lambda] = 0$.

With $n = \pm 1$ we get $H_1 = 0$: *every* ± 1 surgery on *any* knot is a homology sphere. The trefoil is special in that π_1 is the binary icosahedral group of order 120, not trivial. We shall compute its Chern–Simons invariant in Example 12.133.

Every 3-manifold is surgery on a link

Nash quotes the key fact from surgery theory: "after enough surgery any manifold M can be transformed into S^3".

Key idea: Lickorish–Wallace theorem

Every closed, connected, oriented 3-manifold is obtained from S^3 by integer surgery on a framed link $L \subset S^3$. One may even take every component unknotted with framing ± 1.

Why it is plausible. Every closed oriented 3-manifold has a **Heegaard splitting**: it is two handlebodies of some genus g glued along their common boundary surface Σ_g by a diffeomorphism ϕ (the genus-one case is the lens spaces above). S^3 itself has a splitting of every genus. So M and S^3 differ only in the gluing map, ϕ_M versus ϕ_{S^3}. Lickorish proved that every diffeomorphism of Σ_g is, up to isotopy, a product of Dehn twists along a finite list of curves. A Dehn twist on the gluing surface can be undone by a ± 1 surgery on a copy of the twisting curve pushed slightly into the 3-manifold (cutting out a neighbourhood of the curve and regluing with $T^{\pm 1}$ relative to its surface framing is exactly a twist of the two sides against each other). So M is S^3 with finitely many ± 1 surgeries, one per Dehn twist.

When do two surgeries give the same manifold? Kirby moves

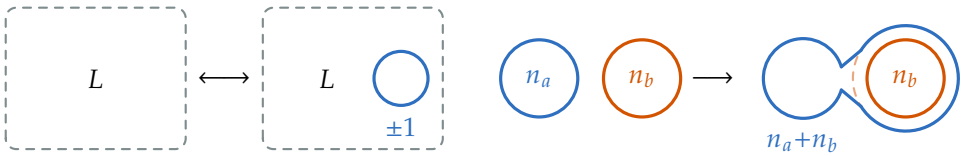
Different framed links can give the same 3-manifold. Kirby found the complete list of moves relating them (Figure 12.83).

Working definition 12.45 (Kirby moves)

Two framed links in S^3 give homeomorphic 3-manifolds if and only if they are related by a sequence of isotopies and the moves

- **K1 (blow-up/down):** add or delete an unknot with framing ± 1 [its push-off links it $+1$ or -1 times] that is unlinked from everything else;
- **K2 (handle slide):** replace a component C_a by the band sum [join the two curves by a thin band and replace the band's two short ends by its two long edges] $C_a \# _b C_b'$ of C_a with a parallel copy C_b' of another component C_b, the copy being pushed off along the framing of C_b. Framings and linking numbers change according to $n_a \mapsto n_a + n_b \pm 2\,\mathrm{lk}(C_a, C_b)$.

A **3-manifold invariant** computed from a framed link must be unchanged by both moves.



K1: add an isolated ±1 unknot K2: slide C_a over C_b

Figure 12.83: The two Kirby moves. K1: an isolated ±1-framed unknot can be added or removed (surgery on it is S^3, and taking a connected sum with S^3 changes nothing). K2: a component C_a can be band-summed with a parallel push-off (dashed) of another component C_b; the new blue component follows C_a, then runs around the push-off of C_b. For unlinked components the framing becomes $n_a + n_b$. Any 3-manifold invariant computed from a framed link must be unchanged by these moves.

Worked example 12.117: The slam-dunk: a Hopf link with framings $(n, 0)$ is S^3

Take a Hopf link $C_1 \cup C_2$ with framings n on C_1 and 0 on C_2. Surgery on C_2 alone (with framing 0) gives $S^2 \times S^1$ (Example 12.115), and in it C_1 becomes the circle $\{\text{pt}\} \times S^1$, a core of one of the two solid tori. Surgery on a core of a genus-one splitting just changes the gluing map: the new meridian goes to $\lambda_1 + n\mu_1$, which in the $S^2 \times S^1$ picture is the curve $m + n\,l$ on the splitting torus with respect to the other torus … and a curve that meets the meridian of the other torus exactly once always gives S^3 (genus one, $p = 1$). So the answer is S^3 for every n. As a Kirby calculus exercise: slide C_1 over C_2 repeatedly to reduce n to 0 or ±1, then use K1. We will check that the Chern–Simons invariant knows this (Example 12.131).

12.15.4 The surgery formula

Now bring in the quantum theory. Surgery is cutting and regluing, and in Section 12.14.8 we learned what the path integral does under cutting and regluing. Cut M along the torus $T^2 = \partial M_2$. The path integral over M_1, with its Wilson lines, is a vector V_1 in (the dual of) $\mathcal{H}_{T^2}$; the path integral over the solid torus M_2, with its Wilson lines, is a vector V_2; and

$$\widetilde{W}\big(M, \{R\}\big) = \langle V_1, V_2 \rangle.$$

When we reglue after applying K to ∂M_2, the vector V_2 is moved by the operator that represents K on $\mathcal{H}_{T^2}$ (Section 12.14.12 explained why diffeomorphisms of a surface act on its Hilbert space). We call this operator K too. Hence Nash's (12.142):

$$\widetilde{W}\big(M^K, \{R\}\big) = \big\langle V_1,\, K\,V_2 \big\rangle. \tag{12.305}$$

Everything now depends on knowing $\mathcal{H}_{T^2}$, a basis of it, and the matrices of K.

A basis of $\mathcal{H}_{T^2}$: Wilson lines along the core

From Example 12.96, $\dim \mathcal{H}_{T^2}$ equals the number of integrable representations of the loop group at level k: for SU(2) it is $k + 1$. Nash writes $\dim \mathcal{H}_{T^2} = n + 1$ for a general group, $n + 1$ being the number of primary fields of the WZW model. There is a natural basis:

Working definition 12.46 (the basis V_2^i)

Let $V_2^i \in \mathcal{H}_{T^2}$ be the vector given by the path integral over the solid torus M_2 with one Wilson line in the representation R_i running once along its core [the central circle $\{0\} \times S^1$]

(parallel to the boundary):

$$V_2^i = \text{(solid torus with core Wilson line } R_i\text{)}, \qquad i = 0, 1, \ldots, n.$$

R_0 is the trivial representation. For it the Wilson line is $\mathrm{tr}_{R_0} P \exp \oint A = 1$: the line is absent. So V_2^0 is the empty solid torus.

Why are these a basis? The quick argument is by counting and by the CFT dictionary. In the holomorphic picture, the path integral on the solid torus with boundary modulus τ gives a conformal block on T^2; with a line R_i along the core it gives the **character**

$$\chi_i(\tau) = \mathrm{Tr}_{\mathcal{H}_i}\, e^{2\pi i\tau(L_0 - c/24)} = e^{2\pi i\tau(h_i - c/24)}\big(\dim R_i + O(e^{2\pi i\tau})\big) \tag{12.306}$$

of the WZW representation $\mathcal{H}_i$ with highest weight R_i (Chapter 11). Characters with different h_i have different leading powers of $e^{2\pi i\tau}$, so they are linearly independent, and there are $n+1$ of them. Physically: in the Hamiltonian picture on the solid torus, the circle direction is time, the disc is space, and a line along the core is a static charge at the centre of the disc; the trace over states with that charge is the character.

Physical picture: What a vector in $\mathcal{H}_{T^2}$ "is"

A state on the boundary torus of a solid torus is the same as "what is running along the core". The vacuum V_2^0 is nothing; V_2^i is a charge R_i. The meridian of the solid torus surrounds the core, so an observer who measures the holonomy around the meridian sees the charge R_i: in the V_2^i basis, Wilson loops along the meridian are *diagonal*. Wilson loops along the longitude instead *add* charge to the core, i.e. act by fusion.

Any operator on $\mathcal{H}_{T^2}$ has a matrix in this basis. Following Nash (12.143), write

$$K\, V_2^i = K^i{}_j\, V_2^j \qquad \text{(sum over } j\text{)}. \tag{12.307}$$

Three consequences

(a) No Wilson lines. If M contains no Wilson lines at all, then $\widetilde{W}(M, \{R\}) = Z(M)$ and V_2 is the empty solid torus, $V_2 = V_2^0$ (Nash's (12.144)–(12.145)):

$$Z(M) = \langle V_1, V_2^0 \rangle. \tag{12.308}$$

Do the surgery:

$$Z(M^K) \overset{(1)}{=} \langle V_1,\, KV_2^0 \rangle \overset{(2)}{=} \langle V_1,\, K^0{}_j V_2^j \rangle \overset{(3)}{=} K^0{}_j \langle V_1, V_2^j \rangle \tag{12.309}$$

$$\overset{(4)}{=} K^0{}_j\, \widetilde{W}\big(M, R_j\big). \tag{12.310}$$

(1) The general surgery formula (12.305) with $V_2 = V_2^0$.
(2) Expand KV_2^0 in the basis, (12.307).
(3) The pairing is bilinear.
(4) $\langle V_1, V_2^j \rangle$ glues back the *original* solid torus, now containing a Wilson line R_j along its core, i.e. along the knot C (Nash's (12.147)).

Read (12.310) (Nash's (12.148)) carefully. The left side is a partition function without Wilson lines on the *new* manifold M^K. The right side is a sum of correlation functions with one Wilson line on the *old* manifold M. Surgery trades topology for Wilson lines.

(b) A Wilson line parallel to the core. If the solid torus M_2 already contains an unknotted Wilson line R_i running parallel to its boundary, then $V_2 = V_2^i$ and (Nash's (12.149))

$$\widetilde{W}(M^K, R_i) \overset{(1)}{=} \langle V_1,\, K V_2^i \rangle \overset{(2)}{=} K^i{}_j \langle V_1, V_2^j \rangle \overset{(3)}{=} K^i{}_j\, \widetilde{W}(M, R_j). \tag{12.311}$$

(1) (12.305) with $V_2 = V_2^i$.
(2) (12.307) and bilinearity.
(3) As in the previous chain.

(c) Lines in M_1 too. Nothing changes if M_1 contains Wilson lines $\{R\}$: they are part of V_1, and $\widetilde{W}(M^K, \{R\}, R_i) = K^i{}_j \widetilde{W}(M, \{R\}, R_j)$.

The matrices of S and T

To use these formulas we need $K^i{}_j$ for $K = S$ and $K = T$. The CFT supplies them, because $\mathrm{SL}(2,\mathbb{Z})$ acts on the characters (12.306) by changing the modular parameter.

Working definition 12.47 (S and T on $\mathcal{H}_{T^2}$)

$$T:\ \chi_i(\tau+1) = e^{2\pi i(h_i - c/24)}\chi_i(\tau), \qquad S:\ \chi_i(-1/\tau) = \sum_j S_{ij}\,\chi_j(\tau).$$

Here $\chi_i(\tau)$ [the vector V_2^i written as a function of the modular parameter τ of the boundary torus] is the character, h_i the conformal weight and c the central charge. So in the basis V_2^i, $T = \mathrm{diag}\big(e^{2\pi i(h_i - c/24)}\big)$ and S is the modular S-matrix. For $\mathrm{SU}(2)_k$, $h_i = \frac{i(i+2)}{4(k+2)}$, $c = \frac{3k}{k+2}$ and

$$S_{ij} = \sqrt{\frac{2}{k+2}}\,\sin\frac{(i+1)(j+1)\pi}{k+2} \qquad \text{(Nash's (12.157)).}$$

S is real, symmetric and orthogonal, $S^2 = \mathbb{1}$, and $(ST)^3 = \mathbb{1}$.

The T rule is immediate from (12.306): every term in χ_i is $e^{2\pi i\tau(h_i - c/24 + N)}$ with integer N, and shifting $\tau \to \tau + 1$ multiplies it by $e^{2\pi i(h_i - c/24)}e^{2\pi i N}$. The S rule is Poisson resummation. Here it is in the simplest case.

Worked example 12.118: The S-matrix of $\mathrm{SU}(2)_1$ by Poisson resummation

At level 1 there are two representations, $j = 0$ and $j = \frac{1}{2}$, with $h = 0$ and $h = \frac{1}{4}$, and $c = 1$. The WZW model is a free boson at the self-dual radius, and its characters are theta functions over the eigenvalues of J_3 (integers for $j = 0$, half-integers for $j = \frac{1}{2}$):

$$\chi_0(\tau) = \frac{1}{\eta(\tau)}\sum_{n\in\mathbb{Z}} e^{2\pi i\tau n^2}, \qquad \chi_{1/2}(\tau) = \frac{1}{\eta(\tau)}\sum_{n\in\mathbb{Z}} e^{2\pi i\tau(n+\frac{1}{2})^2}.$$

Take $\tau = it$ with $t > 0$, so $-1/\tau = i/t$. The Poisson summation formula $\sum_n f(n) = \sum_m \hat{f}(m)$, with $\hat{f}(m) = \int f(x)e^{-2\pi i m x}\mathrm{d}x$, applied to the Gaussian $f(x) = e^{-2\pi t x^2}$ gives

$$\sum_n e^{-2\pi t n^2} \overset{(1)}{=} \frac{1}{\sqrt{2t}}\sum_m e^{-\pi m^2/(2t)} \overset{(2)}{=} \frac{1}{\sqrt{2t}}\Big[\sum_p e^{-2\pi p^2/t} + \sum_p e^{-2\pi(p+\frac{1}{2})^2/t}\Big].$$

(1) The Fourier transform of $e^{-2\pi t x^2}$ is $(2t)^{-1/2}e^{-\pi m^2/(2t)}$ (complete the square).

(2) Split m into even $m = 2p$ and odd $m = 2p+1$: $\pi m^2/(2t) = 2\pi p^2/t$ or $2\pi(p+\frac{1}{2})^2/t$.

The bracket is $\eta(i/t)\big[\chi_0(i/t)+\chi_{1/2}(i/t)\big]$, and the Dedekind function obeys $\eta(i/t) = \sqrt{t}\,\eta(it)$. Dividing by $\eta(it)$:

$$\chi_0(it) \overset{(3)}{=} \frac{\sqrt{t}}{\sqrt{2t}}\big[\chi_0(i/t)+\chi_{1/2}(i/t)\big] = \frac{1}{\sqrt{2}}\big[\chi_0(-1/\tau)+\chi_{1/2}(-1/\tau)\big].$$

(3) Combine the previous chain with $\eta(i/t) = \sqrt{t}\,\eta(it)$.

The same computation with $n+\frac{1}{2}$ gives $\chi_{1/2} = \frac{1}{\sqrt{2}}[\chi_0 - \chi_{1/2}]$ (the half-shift produces the sign $e^{-2\pi i m/2} = (-1)^m$). Since $S^2 = \mathbb{1}$ we can invert, and

$$S = \frac{1}{\sqrt{2}}\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix} = \sqrt{\tfrac{2}{3}}\begin{pmatrix}\sin\frac{\pi}{3} & \sin\frac{2\pi}{3}\\ \sin\frac{2\pi}{3} & \sin\frac{4\pi}{3}\end{pmatrix},$$

in agreement with the general formula at $k = 1$.

Worked example 12.119: $(ST)^3 = \mathbb{1}$ at level 1, and where the $c/24$ matters

At $k = 1$, $T = \mathrm{diag}\big(e^{-2\pi i/24}, e^{2\pi i(1/4-1/24)}\big) = e^{-i\pi/12}\,\mathrm{diag}(1,i)$. Put $N = \begin{pmatrix}1 & i\\ 1 & -i\end{pmatrix}$, so that $ST = \frac{e^{-i\pi/12}}{\sqrt{2}}N$.

$$N^2 \overset{(1)}{=} \begin{pmatrix}1+i & 1+i\\ 1-i & -1+i\end{pmatrix}, \qquad N^3 \overset{(2)}{=} \begin{pmatrix}2+2i & 0\\ 0 & 2+2i\end{pmatrix},$$

$$(ST)^3 \overset{(3)}{=} \frac{e^{-i\pi/4}}{2\sqrt{2}}(2+2i)\,\mathbb{1} \overset{(4)}{=} e^{-i\pi/4}\,e^{i\pi/4}\,\mathbb{1} = \mathbb{1}.$$

(1) Row 1: $(1\cdot 1 + i\cdot 1,\ 1\cdot i + i\cdot(-i)) = (1+i, 1+i)$; row 2: $(1-i,\ i-1)$.

(2) Row 1 times N: $(1+i)+(1+i) = 2+2i$ and $(1+i)i-(1+i)i = 0$; row 2: $(1-i)+(-1+i) = 0$ and $(1-i)i+(-1+i)(-i) = 2+2i$.

(3) Cube the prefactor: $(e^{-i\pi/12})^3/(\sqrt{2})^3$.

(4) $2+2i = 2\sqrt{2}\,e^{i\pi/4}$.

Now drop the $-c/24$, i.e. use $T' = \mathrm{diag}(e^{2\pi i h_i}) = \mathrm{diag}(1,i)$, which is what the framing rule (12.301) alone would suggest. The same computation gives $(ST')^3 = e^{i\pi/4}\mathbb{1} = e^{2\pi i c/8}\mathbb{1}$. So T' and S satisfy the $\mathrm{SL}(2,\mathbb{Z})$ relations only *up to a phase*: the action on $\mathcal{H}_{T^2}$ is a **projective** representation. The phase $e^{2\pi i c/8}$ is the germ of the framing anomaly of Section 12.15.7. (Numerically, $(ST)^3 = \mathbb{1}$ and $(ST')^3 = e^{2\pi i c/8}\mathbb{1}$ hold at every level; we checked $k = 1,2,3,4,10$.)

Worked example 12.120: Level 2 in full

At $k = 2$: labels $i = 0,1,2$ (spins $0, \frac{1}{2}, 1$), $h = (0, \frac{3}{16}, \frac{1}{2})$, $c = \frac{3}{2}$, and

$$S = \frac{1}{2}\begin{pmatrix}1 & \sqrt{2} & 1\\ \sqrt{2} & 0 & -\sqrt{2}\\ 1 & -\sqrt{2} & 1\end{pmatrix}, \qquad T = e^{-i\pi/8}\,\mathrm{diag}\big(1,\ e^{3\pi i/8},\ -1\big).$$

Check $S^2 = \mathbb{1}$ on one entry: $(S^2)_{00} = \frac{1}{4}(1+2+1) = 1$, and $(S^2)_{01} = \frac{1}{4}(\sqrt{2}+0-\sqrt{2}) = 0$. The rows of S are orthonormal. This is the modular data of the Ising model tensored with a

free fermion (Example 12.107): the zero in the middle, $S_{11} = 0$, says that the spin-$\frac{1}{2}$ Hopf link vanishes at $k = 2$, since $[4]_q = \sin(4\pi/4)/\sin(\pi/4) = 0$.

Reading Nash (pp. 355–357, (12.141)–(12.149))

Nash's §7 begins by observing that nothing in QFT restricts us to S^3, and that surgery theory lets us reach every M from S^3. He defines the surgery $M^K = M_1 \cup_{T^2} M_2^K$ (12.141), notes that K acts linearly on $\mathcal{H}_{T^2}$, so $\widetilde{W}(M^K, \{R\}) = \langle V_1, KV_2\rangle$ (12.142), introduces the basis $V_2^0, \dots, V_2^n$ of $\mathcal{H}_{T^2}$ labelled by primaries (Wilson lines along the core; V_2^0 trivial because the Wilson line collapses to the identity) and the matrix $K^i{}_j$ (12.143), and derives $Z(M^K) = K^0{}_j \widetilde{W}(M, R_j)$ (12.144)–(12.148) and $\widetilde{W}(M^K, R_i) = K^i{}_j \widetilde{W}(M, R_j)$ (12.149). Nash writes K_i^j with the indices placed differently; only the order "row = old label, column = new label" matters. We have supplied: the meridian/longitude, $\mathrm{SL}(2,\mathbb{Z})$ and its relations, the three genus-one gluings with the explicit $S^3 \subset \mathbb{C}^2$ decomposition, Dehn surgery and framings, Lickorish–Wallace and Kirby moves, the character argument for the basis, and explicit S, T matrices with the Poisson derivation and the $(ST)^3$ check.

Recap

- Curves on $\partial(D^2 \times S^1)$: the meridian bounds a disc, the longitude is defined up to framing; diffeos of T^2 form $\mathrm{SL}(2,\mathbb{Z})$, generated by S (swap) and T (Dehn twist = unit change of framing).
- Two solid tori glued with $m_2 \mapsto q m_1 + p l_1$ give $L(p, q)$; S^3 uses S, $S^2 \times S^1$ uses the identity.
- Dehn surgery: remove a solid torus around a knot, reglue with K. Every 3-manifold is integer surgery on a framed link in S^3; Kirby moves K1, K2 relate different presentations.
- Quantum surgery formula: $\widetilde{W}(M^K, \{R\}) = \langle V_1, KV_2\rangle$; basis V_2^i = core Wilson line R_i; $Z(M^K) = K^0{}_j \widetilde{W}(M, R_j)$.
- $T = \mathrm{diag}\, e^{2\pi i(h_i - c/24)}$, S = modular S-matrix; without the $c/24$ the representation is only projective.

12.15.5 The Hamiltonian formalism: partition functions are traces

To carry out a complete surgery calculation we must pick a definite M and a definite K. Before doing so Nash inserts "a useful piece of Hamiltonian formalism". It answers the question: *what is the path integral on a 3-manifold that is a surface times a circle?* The answer is short, and we will use it at every step that follows: it is a trace, and because the Hamiltonian vanishes, it counts states.

Periodic time gives a trace

Start with ordinary quantum mechanics of one particle, position x, Hamiltonian H. Feynman's formula writes the propagator as a path integral over paths with fixed ends:

$$\langle x_f|\, e^{-iHt}\, |x_i\rangle = \int_{x(0)=x_i}^{x(t)=x_f} \mathcal{D}x\; e^{iS[x]}.$$

Now make time periodic. Set $x_f = x_i$ and integrate over the common endpoint. The paths are then loops, i.e. maps of the circle $S^1 = [0,t]/(0 \sim t)$ into configuration space, and

$$\int_{\text{loops}} \mathcal{D}x\; e^{iS[x]} \overset{(1)}{=} \int \mathrm{d}x\; \langle x|\, e^{-iHt}\, |x\rangle \overset{(2)}{=} \sum_n \int \mathrm{d}x\; \langle x|e^{-iHt}|n\rangle\langle n|x\rangle$$
$$\overset{(3)}{=} \sum_n \langle n|e^{-iHt}|n\rangle \overset{(4)}{=} \operatorname{Tr}\, e^{-iHt}. \tag{12.312}$$

(1) Split the loop integral into "paths from x back to x" followed by an integral over the base point x; the inner integral is the propagator with $x_f = x_i = x$.
(2) Insert a complete set of energy eigenstates $\sum_n |n\rangle\langle n| = \mathbb{1}$.
(3) Move $\langle n|x\rangle$ to the front and use $\int \mathrm{d}x\, |x\rangle\langle x| = \mathbb{1}$.
(4) The sum of diagonal matrix elements in any basis is the trace.

The same argument works for any field theory: the space is Σ, the fields on Σ play the role of x, and the path integral over fields on $\Sigma \times S^1$ that are periodic in time is the trace of the evolution operator over $\mathcal{H}_\Sigma$. Nash writes $\operatorname{tr} \exp(i\mathcal{H}t)$; the sign in the exponent depends on which way round the circle we call "forward", and will not matter.

Worked example 12.121: The harmonic oscillator: a trace weights each state

Rotate to imaginary time, $t = -i\beta$. For the oscillator $H = \omega(a^\dagger a + \frac{1}{2})$,

$$\operatorname{Tr}\, e^{-\beta H} \overset{(1)}{=} \sum_{n=0}^{\infty} e^{-\beta\omega(n+\frac{1}{2})} \overset{(2)}{=} \frac{e^{-\beta\omega/2}}{1-e^{-\beta\omega}} \overset{(3)}{=} \frac{1}{2\sinh(\beta\omega/2)}.$$

(1) The eigenvalues of H are $\omega(n+\frac{1}{2})$, $n = 0, 1, 2, \ldots$
(2) Geometric series with ratio $e^{-\beta\omega} < 1$.
(3) Multiply numerator and denominator by $e^{\beta\omega/2}$.

This is the partition function of statistical mechanics. Each state is counted with the weight $e^{-\beta E_n}$. If *every* energy were zero, each state would be counted with weight 1 and the trace would be the number of states. That is exactly the situation in a topological theory. (With fermions and periodic boundary conditions the same trace becomes the supertrace, the Witten index of Chapter 7.)

Chern–Simons theory: $Z(\Sigma \times S^1) = \dim \mathcal{H}_\Sigma$

In Section 12.14.3 we split the Chern–Simons action on $\Sigma \times \mathbb{R}$ into a "$p\dot{q}$" term plus A_0 times the Gauss-law constraint $\tilde{F} = 0$. There is no other term. So the canonical Hamiltonian $\mathcal{H} = p\dot{q} - L$ is A_0 times the constraint, and it vanishes on the physical Hilbert space: $\mathcal{H} = 0$. (Physically: no metric, so no energy density to write down; the stress tensor $T_{\mu\nu} = \delta S/\delta g^{\mu\nu}$ vanishes identically.) Hence Nash's (12.150)–(12.151):

$$Z(\Sigma \times S^1) \overset{(1)}{=} \operatorname{Tr}_{\mathcal{H}_\Sigma} \exp(i\mathcal{H}t) \overset{(2)}{=} \operatorname{Tr}_{\mathcal{H}_\Sigma}(\mathbb{1}) \overset{(3)}{=} \dim \mathcal{H}_\Sigma. \tag{12.313}$$

(1) Periodic time gives a trace, (12.312), with space Σ and time circle of length t.
(2) $\mathcal{H} = 0$, so $\exp(i\mathcal{H}t) = \mathbb{1}$ for every t. In particular Z does not depend on the length of the circle, as a topological answer must not.
(3) The trace of the identity on a finite-dimensional space is its dimension.

We met this already as the "glue the cylinder to itself" check of Example 12.101; the Hamiltonian argument says the same thing in the language of operators.

Now let Σ carry marked points $P_1, \dots, P_n$ labelled by representations $R_1, \dots, R_n$: Wilson lines piercing Σ. As Σ evolves round the time circle, each P_a traces out the circle $\{P_a\} \times S^1$. These circles are parallel copies of the time direction. They are unknotted and unlinked from one another, each running once round the S^1 factor (Figure 12.84). The trace is now over the Hilbert space of the punctured surface, and Nash's (12.152) follows by the same chain:

$$\widetilde{W}\big(\Sigma \times S^1, \{R\}\big) = \dim \mathcal{H}_{\Sigma,\{R\}}. \tag{12.314}$$

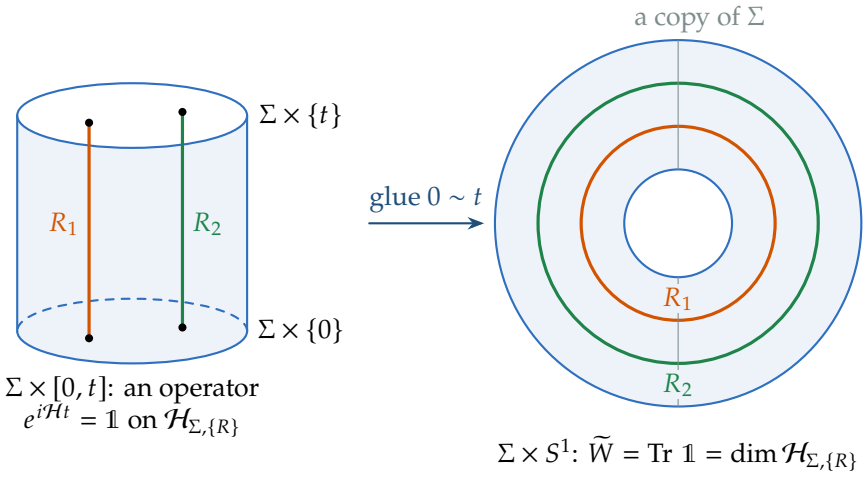


Figure 12.84: Nash's (12.150)–(12.152). Left: the path integral on $\Sigma \times [0, t]$ is the evolution operator on $\mathcal{H}_{\Sigma,\{R\}}$, which is the identity because $\mathcal{H} = 0$. Right: identifying the two ends turns time into a circle (drawn as the angular direction of the ring; each radial slice is a copy of Σ) and the operator into its trace. The marked points trace out parallel, unknotted, unlinked circles $\{P_a\} \times S^1$.

Worked example 12.122: The whole Hilbert-space dictionary of $S^2 \times S^1$

Take $\Sigma = S^2$ with 0, 1, 2 or 3 marked points. The dimensions come from the fusion rules of $\mathrm{SU}(2)_k$ (Section 12.14.6), and (12.314) turns them into correlation functions of parallel circles $\{P_a\} \times S^1$ in $S^2 \times S^1$. Label representations by Nash's $i = 2j$.

$$Z(S^2 \times S^1) \overset{(1)}{=} \dim \mathcal{H}_{S^2} = 1,$$

$$\widetilde{W}(S^2 \times S^1; R_i) \overset{(2)}{=} \dim \mathcal{H}_{S^2;R_i} = \delta_{i0},$$

$$\widetilde{W}(S^2 \times S^1; R_i, R_j) \overset{(3)}{=} \dim \mathcal{H}_{S^2;R_i,R_j} = \delta_{ij},$$

$$\widetilde{W}(S^2 \times S^1; R_i, R_j, R_m) \overset{(4)}{=} \dim \mathcal{H}_{S^2;R_i,R_j,R_m} = N_{ijm}.$$

(1) The sphere with no punctures has a one-dimensional Hilbert space (Example 12.97).
(2) A single charge on a closed sphere violates Gauss's law unless it is neutral: the only invariant in R_i is in R_0. This is Nash's (11.75), $\dim V(S^2_{R_i}) = \delta_{i0}$.
(3) Two charges: $\mathrm{Inv}(R_i \otimes R_j)$ is one-dimensional if $R_j = R_i^*$ and zero otherwise; for SU(2), $R_i^* = R_i$. This is Nash's $g_{ij} = \delta_{ij}$.

(4) Three charges: the number of ways $R_i \otimes R_j$ contains R_m at level k, the **fusion coefficient** $N_{ijm} \in \{0, 1\}$.

At $k = 2$ (spins $0, \frac{1}{2}, 1$) the non-zero fusion coefficients are those with $\frac{1}{2} \times \frac{1}{2} \ni 0, 1$, $\frac{1}{2} \times 1 \ni \frac{1}{2}$ and $1 \times 1 \ni 0$. The level-2 truncation removes $1 \times 1 \to 1$ (allowed classically), so *three parallel spin-1 circles in* $S^2 \times S^1$ *have* $\widetilde{W} = 0$ *at* $k = 2$ *but* 1 *for* $k \geq 3$. A topological amplitude sees the level.

Worked example 12.123: The three-torus and other products

With $\Sigma = T^2$, (12.313) and the pillowcase count of Example 12.96 give

$$Z(T^3) = Z(T^2 \times S^1) = \dim \mathcal{H}_{T^2} = k + 1 \qquad (\mathrm{SU}(2)_k),$$

and $Z(T^3) = k$ for $\mathrm{U}(1)_k$ (Example 12.95). For a surface of genus g, $Z(\Sigma_g \times S^1) = \sum_m S_{0m}^{2-2g}$, the Verlinde numbers of Example 12.100: $Z(\Sigma_2 \times S^1) = 10$ at $k = 2$. Note the pattern: Z of a product with a circle is always a non-negative *integer*, while $Z(S^3)$ will turn out to be an irrational number. The integrality is the fingerprint of a trace over a genuine vector space.

Twisted time: mapping tori

The space-time need not be a product globally. Let Σ evolve for one unit of time and then, before closing time up, apply a diffeomorphism K of Σ: the point $(x, 0)$ is glued to $(Kx, 1)$. The result is the **mapping torus**

$$\Sigma \times_K S^1 = \Sigma \times [0, 1] \big/ \big((x, 0) \sim (Kx, 1)\big),$$

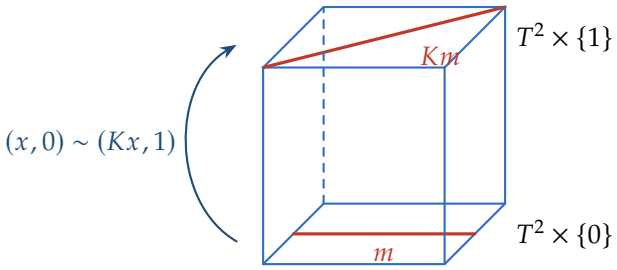
a bundle over the circle with fibre Σ and monodromy K (Figure 12.85). The path integral over $\Sigma \times [0, 1]$ is the identity operator, the twisted gluing inserts the operator $\rho(K)$ that represents K on $\mathcal{H}_\Sigma$ (Section 12.14.12), and gluing the ends takes the trace. So (Nash's (12.153))

$$Z(\Sigma \times_K S^1) \overset{(1)}{=} \mathrm{Tr}\big(\rho(K) \exp(i\mathcal{H}t)\big) \overset{(2)}{=} \mathrm{Tr}\ \rho(K). \tag{12.315}$$

(1) Evolve with $e^{i\mathcal{H}t}$, then apply $\rho(K)$, then trace: the same chain as (12.312) with $\rho(K)$ inserted before the ends are identified.

(2) $\mathcal{H} = 0$.

With $K = \mathrm{id}$ we recover $\Sigma \times S^1$ and $\mathrm{Tr}\ \mathbb{1} = \dim$.



$K = \mathbb{1}$: T^3, $\quad Z = k + 1$
$K = -\mathbb{1}$: $Z = \mathrm{Tr}\ C = k + 1$
$K = S$: $Z = \mathrm{Tr}\ S \in \mathbb{Z}$
general: $Z = \mathrm{Tr}\ \rho(K)$

Figure 12.85: A mapping torus of T^2. Take the slab $T^2 \times [0, 1]$ and glue the bottom to the top after applying K. Here K is the Dehn twist T: the horizontal meridian m on the bottom is glued to the sheared curve Km on the top. The partition function is the trace of the operator that represents K on $\mathcal{H}_{T^2}$ (Nash's (12.153)).

Worked example 12.124: Mapping tori of the torus

On $\mathcal{H}_{T^2}$ we know the operators of S and T in the basis V^i of core Wilson lines (Section 12.15.4).

- $K = -\mathbb{1}$, the involution $(x, y) \mapsto (-x, -y)$ of the square torus. It is S^2 in $\mathrm{SL}(2, \mathbb{Z})$, and it reverses the direction of the core, so it maps a Wilson line R_i to one in the conjugate representation: $\rho(-\mathbb{1}) = C$, the **charge conjugation** matrix $C_{ij} = \delta_{i,j^*}$. For SU(2) every representation is self-conjugate, $C = \mathbb{1}$, and $Z = \mathrm{Tr}\, C = k + 1$.
- $K = S$. Since $S^2 = \mathbb{1}$ on $\mathcal{H}_{T^2}$, the eigenvalues of S are ± 1, and $\mathrm{Tr}\, S = n_+ - n_-$ is an integer. At $k = 2$:

$$\mathrm{Tr}\, S \overset{(1)}{=} S_{00} + S_{11} + S_{22} \overset{(2)}{=} \tfrac{1}{2} + 0 + \tfrac{1}{2} = 1.$$

  (1) The trace is the sum of diagonal entries.
  (2) The level-2 matrix of Example 12.120.

  So S has eigenvalues $(+1, +1, -1)$ at $k = 2$. At $k = 1$, $\mathrm{Tr}\, S = \frac{1}{\sqrt{2}}(1 - 1) = 0$; at $k = 3$ the diagonal entries are $\sqrt{\frac{2}{5}} \sin \frac{\pi}{5}\, (1, 1, -1, -1)$ and again $\mathrm{Tr}\, S = 0$.

For $K = T$ the trace $\sum_i e^{2\pi i (h_i - c/24)}$ is a sum of phases whose overall phase depends on the conventions discussed in Example 12.119: it is only defined up to the framing anomaly. The traces of $\pm\mathbb{1}$ and S are free of this ambiguity.

Watch out: Which Hilbert space do the marked points live in?

In (12.314) the Wilson lines must run *parallel to the time circle*, once round. A Wilson line that winds around the S^1 twice, or that is knotted inside $\Sigma \times S^1$, is not a set of marked points on a single copy of Σ, and the formula does not apply. Likewise, a Wilson loop lying inside one copy $\Sigma \times \{t_0\}$ is not a marked point at all: it is an *operator* acting on $\mathcal{H}_\Sigma$ (for example, a loop around a cycle of T^2; see Example 12.128).

12.15.6 From $S^2 \times S^1$ to S^3: the unknot from the S-matrix

We now have everything needed for Nash's showcase calculation. We take

$$M = S^2 \times S^1, \qquad K = S, \qquad M^K = S^3,$$

and use the surgery formulas (12.310) and (12.311). The inputs on $S^2 \times S^1$ are the counting results of Example 12.122; the output will be the partition function of S^3 and the value of the unknot, obtained this time with no skein relation and no braiding.

The surgery

In Section 12.15.2 we saw both manifolds as two solid tori glued together:

$$S^2 \times S^1 = (D \times S^1) \cup_{\mathrm{id}} (D' \times S^1), \qquad S^3 = (D \times S^1) \cup_S (D' \times S^1).$$

The first piece and the second piece are the same in both; only the gluing differs, by S. So: start from $S^2 \times S^1$, cut out the solid torus $M_2 = D \times S^1$ (a neighbourhood of the circle $C = \{N\} \times S^1$ through the north pole), apply S to its boundary, and glue it back. The result is S^3. This is exactly Nash's description "cutting $S^2 \times S^1$ into two solid tori and then acting with the inversion

on one of them before gluing the pieces back together". In the language of Example 12.115, it is the inverse of 0-surgery on the unknot.

Follow the other core circle, $\{S\} \times S^1$ through the south pole. It lies in $M_1 = D' \times S^1$ and is untouched by the surgery. In S^3 it is the core of one of the two Heegaard solid tori: an *unknot*. So if we place a Wilson line R_i on $\{S\} \times S^1$ before the surgery, we obtain an unknotted Wilson line R_i in S^3 after it (Figure 12.86).

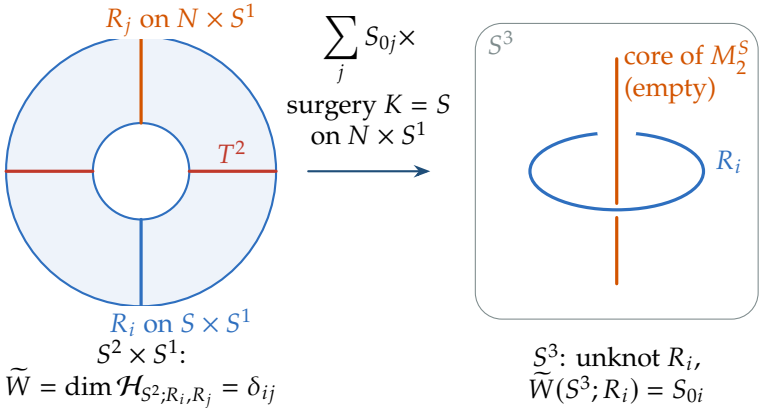


Figure 12.86: Nash's surgery from $S^2 \times S^1$ to S^3. Left: $S^2 \times S^1$ as a spherical shell with inner and outer spheres identified; the two radial segments are the circles through the poles. The red torus (the equator times S^1, seen in this slice as two horizontal segments) cuts it into two solid tori, the upper and lower halves. The surgery formula expands the reglued solid torus as a sum over Wilson lines R_j on its core (orange) with weights S_{0j}. Right: after the surgery the blue circle, which never moved, is an unknot in S^3; the orange circle is the core of the reglued torus, which carries no Wilson line. It runs through the blue unknot and closes through the point at infinity.

The calculation

Apply (12.310) with no Wilson lines, and the variant (c) of Section 12.15.4 with a line R_i in M_1. The matrix of $K = S$ in the basis V_2^j is the modular S-matrix, $K^0_{\ j} = S_{0j}$. This is Nash's (12.154):

$$Z(S^3) \overset{(1)}{=} S_{0j}\, \widetilde{W}\big(S^2 \times S^1, R_j\big) \overset{(2)}{=} S_{0j}\, \delta_{j0} = S_{00}, \tag{12.316}$$

$$\widetilde{W}(S^3, R_i) \overset{(3)}{=} S_{0j}\, \widetilde{W}\big(S^2 \times S^1, R_i, R_j\big) \overset{(4)}{=} S_{0j}\, \delta_{ij} = S_{0i}, \tag{12.317}$$

$$W(S^3, R_i) \overset{(5)}{=} \frac{\widetilde{W}(S^3, R_i)}{Z(S^3)} \overset{(6)}{=} \frac{S_{0j} \dim \mathcal{H}_{S^2;R_i,R_j}}{S_{0j} \dim \mathcal{H}_{S^2;R_j}} \overset{(7)}{=} \frac{S_{0i}}{S_{00}}. \tag{12.318}$$

(1) Nash's (12.148) with $K = S$, $M = S^2 \times S^1$, $M^K = S^3$; the sum over j puts a line R_j on the core $N \times S^1$ of the removed solid torus.
(2) The single circle $N \times S^1$ is a marked point on S^2 carried round the time circle: $\widetilde{W} = \dim \mathcal{H}_{S^2;R_j} = \delta_{j0}$ by (12.314) and (11.75).
(3) The same with the extra line R_i on $S \times S^1$, which lies in M_1 and is carried along unchanged (Nash's (12.149) in the form (c)).
(4) Two marked points: $\dim \mathcal{H}_{S^2;R_i,R_j} = g_{ij} = \delta_{ij}$ (Nash's (12.155)).
(5) Definition of the normalised correlation function.
(6) Nash's way of writing the two previous results, with the dimensions of (12.151)–(12.152) kept visible.
(7) The dimensions are δ_{ij} and δ_{j0}; the sums over j collapse. This is Nash's (12.156).

Three remarks.

- $Z(S^3)$ *is not* 1. The simplest closed 3-manifold has partition function S_{00}, which equals $\sqrt{2/(k+2)}\,\sin(\pi/(k+2))$: a small irrational number. This is why Nash always divides by $Z(M)$.
- *No knot diagram was used.* The unknot value came from the gluing of two solid tori and from counting states on a sphere.
- *A direct route.* Since $S^3 = V_1 \cup_S V_2$ with V^i the solid torus containing a core line R_i, the same numbers are $Z(S^3) = \langle V^0, SV^0\rangle = S_{00}$ and $\widetilde{W}(S^3, R_i) = \langle V^i, SV^0\rangle = S_{i0}$, using $\langle V^i, V^j\rangle = \widetilde{W}(S^2\times S^1; R_i, R_j) = \delta_{ij}$. Nash's surgery calculation is this statement read in a different order.

Where the sines come from

Nash quotes the matrix elements of S from Gepner and Witten (his (12.157)):

$$S_{ij} = \sqrt{\frac{2}{k+2}}\,\sin\!\Big(\frac{(i+1)(j+1)\pi}{k+2}\Big), \qquad i,j = 0,1,\dots,k. \tag{12.319}$$

We verified this at $k=1$ by Poisson resummation in Example 12.118. For general k the derivation has the same shape, and it is worth seeing because it explains *why* a sine appears: the sine is the Weyl character formula of SU(2).

Step 1: the Weyl character formula. The character of the spin-j representation of SU(2) on the rotation $e^{i\theta J_3}$ is

$$\chi_j(\theta) \overset{(1)}{=} \sum_{m=-j}^{j} e^{im\theta} \overset{(2)}{=} e^{-ij\theta}\,\frac{e^{i(2j+1)\theta}-1}{e^{i\theta}-1} \overset{(3)}{=} \frac{e^{i(j+\frac12)\theta}-e^{-i(j+\frac12)\theta}}{e^{i\theta/2}-e^{-i\theta/2}} \overset{(4)}{=} \frac{\sin\big((2j+1)\theta/2\big)}{\sin(\theta/2)}. \tag{12.320}$$

(1) J_3 has eigenvalues $m = -j, -j+1, \dots, j$ on the spin-j multiplet.
(2) Factor out the first term and sum the geometric series of $2j+1$ terms with ratio $e^{i\theta}$.
(3) Multiply numerator and denominator by $e^{-i\theta/2}$.
(4) $e^{ix} - e^{-ix} = 2i\sin x$ in numerator and denominator.

The numerator is an *antisymmetrised* exponential (odd under $\theta\to-\theta$, the Weyl reflection), and the denominator is the same expression for $j=0$ shifted by $\frac12$: the **Weyl vector** $\rho = \frac12$.

Step 2: the affine version. For the level-k current algebra the sum over m becomes a sum over a lattice that also contains the momentum modes of the currents. The result (the Weyl–Kac character formula) replaces each exponential by a theta function

$$\Theta_{m,K}(\tau,z) := \sum_{n\in\mathbb{Z}+\frac{m}{2K}} e^{2\pi iK\tau n^2}\,e^{2\pi iKnz}, \qquad m\in\mathbb{Z}/2K\mathbb{Z},$$

and reads, with $K = k+2$ and i Nash's label,

$$\chi_i(\tau,z) = \frac{\Theta_{i+1,k+2}(\tau,z) - \Theta_{-i-1,k+2}(\tau,z)}{\Theta_{1,2}(\tau,z) - \Theta_{-1,2}(\tau,z)}. \tag{12.321}$$

Here z is the chemical potential for J_3, which we need to keep the antisymmetric numerator from vanishing identically. We can check this formula against what we know. The leading term of $\Theta_{m,K}$ is $n = m/2K$, with $e^{2\pi i\tau\, m^2/4K}$. So the leading power of $e^{2\pi i\tau}$ in χ_i is

$$\frac{(i+1)^2}{4(k+2)} - \frac18 \overset{(1)}{=} \frac{2(i+1)^2-(k+2)}{8(k+2)} \overset{(2)}{=} \frac{i(i+2)}{4(k+2)} - \frac{k}{8(k+2)} \overset{(3)}{=} h_i - \frac{c}{24},$$

(1) Common denominator $8(k+2)$.
(2) $2(i+1)^2 - 2 = 2i(i+2)$; split off $-k/(8(k+2))$.
(3) $h_i = i(i+2)/4(k+2)$ and $c/24 = 3k/(24(k+2)) = k/(8(k+2))$.

exactly as (12.306) demands; and the coefficient of that leading power is $(y^{(i+1)/2} - y^{-(i+1)/2})/(y^{1/2} - y^{-1/2})$ with $y = e^{2\pi i z}$, the Weyl character (12.320) of spin $i/2$ with $e^{i\theta} = y$: the ground states form one copy of the representation R_i.

Step 3: modular transformation of one theta function. Poisson resummation, exactly as in Example 12.118 but with the extra linear term $2\pi i K n z$ (completing the square produces a Gaussian in z), gives

$$\Theta_{m,K}\Big(-\frac{1}{\tau}, \frac{z}{\tau}\Big) = A_K(\tau, z) \sum_{m'=0}^{2K-1} e^{-i\pi m m'/K}\, \Theta_{m',K}(\tau, z), \qquad A_K = \sqrt{\frac{-i\tau}{2K}}\, e^{i\pi K z^2/(2\tau)}.$$

The phases $e^{-i\pi m m'/K}$ are a *discrete Fourier transform* on $\mathbb{Z}/2K\mathbb{Z}$: Poisson resummation exchanges a lattice with its dual, and the finite group of cosets with its group of characters.

Step 4: antisymmetrise. Write $N_m := \Theta_{m,K} - \Theta_{-m,K}$ for the numerator.

$$N_m\Big(-\frac{1}{\tau}, \frac{z}{\tau}\Big) \overset{(1)}{=} A_K \sum_{m'=0}^{2K-1} \big(e^{-i\pi m m'/K} - e^{+i\pi m m'/K}\big)\Theta_{m',K} \overset{(2)}{=} -2i\, A_K \sum_{m'=0}^{2K-1} \sin\frac{\pi m m'}{K}\, \Theta_{m',K}$$
$$\overset{(3)}{=} -2i\, A_K \sum_{m'=1}^{K-1} \sin\frac{\pi m m'}{K}\, \big(\Theta_{m',K} - \Theta_{-m',K}\big) = -2i\, A_K \sum_{m'=1}^{K-1} \sin\frac{\pi m m'}{K}\, N_{m'}. \quad (12.322)$$

(1) Step 3 for $\Theta_{m,K}$ and for $\Theta_{-m,K}$ (replace m by $-m$).
(2) $e^{-ix} - e^{ix} = -2i \sin x$.
(3) Pair the term m' with the term $2K - m' \equiv -m'$: since $\sin(\pi m(2K - m')/K) = -\sin(\pi m m'/K)$ for integer m, the pair is $\sin(\pi m m'/K)(\Theta_{m'} - \Theta_{-m'})$. The terms $m' = 0$ and $m' = K$ have $\sin = 0$.

The denominator of (12.321) is the case $K = 2$, $m = 1$, for which the sum has the single term $m' = 1$ with $\sin(\pi/2) = 1$: it transforms into $-2iA_2$ times itself. Dividing, and writing $m = i + 1$, $m' = j + 1$:

$$\chi_i\Big(-\frac{1}{\tau}, \frac{z}{\tau}\Big) \overset{(4)}{=} \frac{A_{k+2}}{A_2} \sum_{j=0}^{k} \sin\frac{\pi(i+1)(j+1)}{k+2}\, \chi_j(\tau, z) \overset{(5)}{=} e^{i\pi k z^2/(2\tau)} \sum_{j=0}^{k} S_{ij}\, \chi_j(\tau, z).$$

(4) (12.322) for numerator ($K = k + 2$) and denominator ($K = 2$); the factors $-2i$ cancel.
(5) $A_{k+2}/A_2 = \sqrt{2/(k+2)}\, e^{i\pi(k+2-2)z^2/(2\tau)}$. The z-dependent phase disappears at $z = 0$ and is the standard anomaly factor of current-algebra characters; the number in front of the sine is Nash's normalisation.

This is (12.319). (We also checked it numerically, by summing the theta series at a random τ, for $k = 1, 2, 3, 5$.)

Worked example 12.125: S is orthogonal: the discrete sine transform

Set $K = k + 2$ and $a = i + 1$, $b = j + 1 \in \{1, \dots, K - 1\}$. Then

$$\sum_{m=1}^{K-1} \sin\frac{\pi a m}{K} \sin\frac{\pi b m}{K} \overset{(1)}{=} \frac{1}{2} \sum_{m=0}^{2K-1} \sin\frac{\pi a m}{K} \sin\frac{\pi b m}{K}$$

$$\overset{(2)}{=} \frac{1}{4}\sum_{m=0}^{2K-1}\Big[\cos\frac{\pi(a-b)m}{K} - \cos\frac{\pi(a+b)m}{K}\Big] \overset{(3)}{=} \frac{1}{4}\big[2K\,\delta_{ab} - 0\big] = \frac{K}{2}\,\delta_{ab}.$$

(1) The summand is unchanged under $m \to 2K - m$ (both sines flip sign) and vanishes at $m = 0$ and $m = K$; so the sum over a full period $0 \le m < 2K$ is twice the sum over $1 \le m \le K-1$.
(2) $\sin x \sin y = \frac{1}{2}[\cos(x-y) - \cos(x+y)]$.
(3) $\sum_{m=0}^{2K-1} e^{i\pi nm/K}$ is a sum of all powers of the $2K$-th root of unity $e^{i\pi n/K}$: it is $2K$ if $n \equiv 0 \pmod{2K}$ and 0 otherwise. Now $|a-b| < K$ vanishes only for $a = b$, and $2 \le a+b \le 2K-2$ is never $\equiv 0$.

Multiplying by the prefactor $(\sqrt{2/K})^2 = 2/K$ gives $\sum_m S_{im}S_{jm} = \delta_{ij}$: S is orthogonal. It is also symmetric, so $S^2 = SS^{\mathsf{T}} = \mathbb{1}$, the relation we used for the mapping torus of S. Taking $i = j = 0$, $\sum_m S_{0m}^2 = 1$: this was the genus-zero Verlinde count $\dim \mathcal{H}_{S^2} = 1$ of Example 12.100.

The unknot, once more

For the Jones polynomial the Wilson line carries the defining representation, $i = 1$ in Nash's labels. Then (12.318) and (12.319) give Nash's (12.158):

$$W(S^3, R_1) \overset{(1)}{=} \frac{S_{01}}{S_{00}} \overset{(2)}{=} \frac{\sin\big(2\pi/(k+2)\big)}{\sin\big(\pi/(k+2)\big)} \overset{(3)}{=} q^{1/2} + q^{-1/2} \overset{(4)}{=} -\frac{t - t^{-1}}{t^{1/2} - t^{-1/2}}. \tag{12.323}$$

(1) (12.318) with $i = 1$.
(2) (12.319): the common prefactor $\sqrt{2/(k+2)}$ cancels.
(3) $\sin 2x/\sin x = 2\cos x$ with $x = \pi/(k+2)$, and $2\cos x = q^{1/2} + q^{-1/2}$ for $q = e^{2\pi i/(k+2)}$.
(4) With $t = q^{-1}$ and $t^{1/2} = -q^{-1/2}$ (our convention (12.294)): $-(t - t^{-1})/(t^{1/2} - t^{-1/2}) = -(t^{1/2} + t^{-1/2}) = q^{-1/2} + q^{1/2}$.

This is the unknot value $[2]_q$ that the skein relation gave in (12.297): the braiding route and the surgery route agree.

Watch out: The minus sign in Nash's (12.158)

If one reads $t^{1/2}$ in (12.158) as $e^{-i\pi/(k+2)}$, the obvious square root of Nash's $t = e^{-2\pi i/(k+2)}$, then $(t - t^{-1})/(t^{1/2} - t^{-1/2}) = t^{1/2} + t^{-1/2} = 2\cos(\pi/(k+2)) > 0$ and the second line of (12.158) is *minus* the first. The two lines agree only with the other branch, $t^{1/2} = -e^{-i\pi/(k+2)} = -q^{-1/2}$, which is the branch that makes the Chern–Simons skein relation coincide with Jones's (Section 12.14.13). The sign of $t^{1/2}$ is not cosmetic: it is the sign that makes a circle count $+2$ (not -2) at $k = \infty$.

Worked example 12.126: Coloured unknots: quantum dimensions

For a Wilson line in the spin-j representation ($i = 2j$), (12.318) gives

$$W(S^3, R_i) \overset{(1)}{=} \frac{\sin\big((i+1)\pi/(k+2)\big)}{\sin\big(\pi/(k+2)\big)} \overset{(2)}{=} [i+1]_q \overset{(3)}{=} \chi_{i/2}\Big(\frac{2\pi}{k+2}\Big).$$

(1) (12.319) with $j = 0$ in both numerator and denominator.
(2) The quantum integer (12.248) with $n = i + 1$.
(3) The Weyl character (12.320) at the angle $\theta = 2\pi/(k+2)$.

This is the **quantum dimension** d_i of R_i, the number we assumed in the quantum trace (12.300): $d_0 = 1$, $d_1 = [2]$, $d_2 = [3]$. As $k \to \infty$, $d_i \to i+1 = \dim R_i$. At finite level (Figure 12.87):

- $d_i = d_{k-i}$, because $\sin((k+2-(i+1))\pi/(k+2)) = \sin((i+1)\pi/(k+2))$;
- $d_k = 1$: the top representation $j = k/2$ is a **simple current**, whose fusion with anything gives a single representation;
- the **total quantum dimension** is fixed by orthogonality:

$$\mathcal{D}^2 := \sum_{i=0}^{k} d_i^2 \overset{(4)}{=} \frac{1}{S_{00}^2} \sum_i S_{0i}^2 \overset{(5)}{=} \frac{1}{S_{00}^2} \overset{(6)}{=} \frac{k+2}{2\sin^2\bigl(\pi/(k+2)\bigr)}.$$

(4) $d_i = S_{0i}/S_{00}$.
(5) Row 0 of the orthogonal matrix S is a unit vector (Example 12.125).
(6) (12.319) with $i = j = 0$.

So $Z(S^3) = S_{00} = 1/\mathcal{D}$. At $k = 1$: $d = (1,1)$, $\mathcal{D}^2 = 2$; at $k = 2$: $d = (1,\sqrt{2},1)$, $\mathcal{D}^2 = 4$; at $k = 3$: $d = (1,\phi,\phi,1)$ with $\phi = 2\cos(\pi/5) = \frac{1+\sqrt{5}}{2}$ the golden ratio, $\mathcal{D}^2 = 2+2\phi^2 = 5+\sqrt{5} \approx 7.236$.

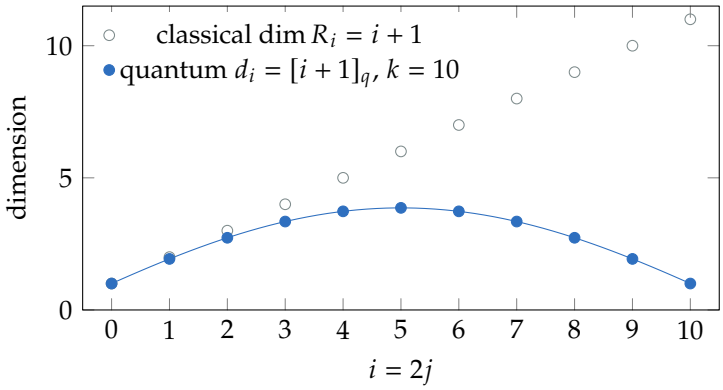


Figure 12.87: Quantum dimensions $d_i = S_{0i}/S_{00}$ at level $k = 10$ (filled) against the classical dimensions $i+1$ (open). For small spins they agree closely. Then they bend over, reach a maximum $1/\sin(\pi/12) \approx 3.86$ in the middle, and come back to $d_{10} = 1$. The quantum dimensions lie on a sine curve, because they are the Weyl character evaluated at the angle $2\pi/(k+2)$.

Physical picture: Why the character is evaluated at a special angle

The ratio $S_{ij}/S_{0j} = \chi_{i/2}\bigl(2\pi(j+1)/(k+2)\bigr)$ is a Wilson loop in the representation R_i, evaluated on a *fixed* holonomy $\exp\bigl(2\pi i\,\frac{j+1}{k+2}\,\sigma_3\bigr)$. Classically, on a solid torus containing a charge R_j on its core, the flat connection on the complement has meridian holonomy $\exp(2\pi i\,\frac{j}{k}\sigma_3)$: a charge j is a magnetic flux $\propto j/k$ in Chern–Simons theory, which ties charge to flux. The quantum corrections shift $j \to j+1$ (the Weyl vector) and $k \to k+2$ (the dual Coxeter number). The unknot, $j = 0$, sees the smallest possible flux $2\pi/(k+2)$.

The Hopf link from the S-matrix

The same splitting $S^3 = V_1 \cup_S V_2$ computes the first non-trivial link. Put a Wilson line R_i on the core of V_1 and R_j on the core of V_2. The two cores $\{z_1 = 0\}$ and $\{z_2 = 0\}$ form a Hopf link (Figure 12.80), and the gluing rule gives it at once.

Worked example 12.127: The coloured Hopf link

$$\widetilde{W}\big(S^3;\mathrm{Hopf}_{ij}\big) \overset{(1)}{=} \big\langle V^i,\, S\,V^j\big\rangle \overset{(2)}{=} \sum_m S_{jm}\big\langle V^i, V^m\big\rangle \overset{(3)}{=} \sum_m S_{jm}\delta_{im} = S_{ij},$$

$$W\big(S^3;\mathrm{Hopf}_{ij}\big) \overset{(4)}{=} \frac{S_{ij}}{S_{00}} \overset{(5)}{=} \frac{\sin\big((i+1)(j+1)\pi/(k+2)\big)}{\sin\big(\pi/(k+2)\big)} = \big[(i+1)(j+1)\big]_q.$$

(1) The gluing formula (12.305): $V_1 = V^i$ is the solid torus with core line R_i, $V_2 = V^j$ likewise, glued by S.
(2) Expand SV^j in the basis, (12.307).
(3) $\langle V^i, V^m\rangle$ glues two solid tori by the identity: two parallel lines in $S^2 \times S^1$, i.e. δ_{im} by Example 12.122.
(4) Divide by $Z(S^3) = S_{00}$.
(5) (12.319); the prefactors cancel.

Each core is pushed off along the longitude of its own solid torus, which is the meridian of the other; that curve bounds a meridian disc disjoint from the first core, so the framing is the zero framing. Now compare with the braiding calculations of Section 12.14.15:

- spin $\frac{1}{2}$ on both, $i = j = 1$: $W = [4]_q$, the value of the zero-framed Hopf link found from the eigenvalues in Example 12.110;
- converting to the Jones polynomial with the dictionary of Section 12.14.16 (divide by one unknot [2], and by $\theta^{2\,\mathrm{lk}} = q^{3/2}$ because the zero framing is not the blackboard framing of the standard 2-crossing diagram, which has writhe 2):

$$V(H_+) \overset{(6)}{=} q^{-3/2}\,\frac{[4]}{[2]} \overset{(7)}{=} q^{-3/2}\big(q + q^{-1}\big) = q^{-1/2} + q^{-5/2},$$

  (6) Normalisation of Section 12.14.16.
  (7) $[4]/[2] = (q^2 - q^{-2})/(q - q^{-1}) = q + q^{-1}$.

  exactly the skein answer of Example 12.108;
- as $k \to \infty$, $W \to (i+1)(j+1) = \dim R_i \dim R_j$: at infinite level linked and unlinked circles cannot be told apart, as for the abelian theory at zero coupling.

At $k = 2$ the spin-$\frac{1}{2}$ Hopf link vanishes, $[4]_{q=i} = \sin\pi/\sin(\pi/4) = 0$, the zero in the middle of the level-2 S-matrix of Example 12.120.

The Verlinde formula from surgery

The S-matrix also produces the fusion rules, i.e. the dimensions of all the Hilbert spaces we have been using. The argument uses Wilson loops as *operators* on $\mathcal{H}_{T^2}$.

Worked example 12.128: Fusion coefficients from S

Let $W_i(\gamma)$ be the operator on $\mathcal{H}_{T^2}$ obtained by inserting a Wilson loop R_i along a curve γ on a copy of the boundary torus, pushed slightly into the solid torus. Act on the basis V^j (a solid torus with core line R_j).

- *Longitude l.* The loop runs parallel to the core. Push it onto the core: the two parallel lines R_i and R_j fuse into $R_i \otimes R_j = \bigoplus_m N_{ij}^{\ m} R_m$, so $W_i(l)\, V^j = \sum_m N_{ij}^{\ m} V^m$.
- *Meridian m.* The loop encircles the core once. A small ball around the loop meets the core in a line segment, so its boundary is a sphere with two punctures R_j, R_j, whose Hilbert space is one-dimensional. The loop can therefore only multiply by a number: $W_i(m)\, V^j = c_{ij} V^j$. To find c_{ij}, glue in the complementary solid torus with S: the core closes into an unknot R_j in S^3 and the meridian loop becomes a Hopf-linked unknot R_i. So $S_{ij} = c_{ij} S_{0j}$, i.e. $c_{ij} = S_{ij}/S_{0j}$.

S exchanges meridian and longitude, so $W_i(l) = S\, W_i(m)\, S^{-1}$. In components, with $S^{-1} = S$ real symmetric:

$$N_{ij}^{\ m} \overset{(1)}{=} \big(S\, W_i(m)\, S\big)_{jm} \overset{(2)}{=} \sum_l S_{jl}\, \frac{S_{il}}{S_{0l}}\, S_{lm}. \tag{12.324}$$

(1) The matrix of $W_i(l)$ in the basis V^j is $N_{ij}^{\ m}$; conjugate the diagonal matrix $W_i(m)$ by S.

(2) $W_i(m) = \mathrm{diag}_l\big(S_{il}/S_{0l}\big)$.

This is the **Verlinde formula**: the S-matrix diagonalises the fusion rules. Check at $k = 2$, with the matrix of Example 12.120 (rows $S_{1l} = \frac{\sqrt{2}}{2}(1, 0, -1)$, $S_{0l} = \frac{1}{2}(1, \sqrt{2}, 1)$, $S_{2l} = \frac{1}{2}(1, -\sqrt{2}, 1)$):

$$N_{11}^{\ 0} \overset{(3)}{=} \sum_l S_{1l}^2\, \frac{S_{0l}}{S_{0l}} = \tfrac{1}{2} + 0 + \tfrac{1}{2} = 1,$$

$$N_{11}^{\ 1} \overset{(4)}{=} \sum_l S_{1l}^2\, \frac{S_{1l}}{S_{0l}} = \tfrac{1}{2}\cdot\sqrt{2} + 0 + \tfrac{1}{2}\cdot(-\sqrt{2}) = 0,$$

$$N_{11}^{\ 2} \overset{(5)}{=} \sum_l S_{1l}^2\, \frac{S_{2l}}{S_{0l}} = \tfrac{1}{2}\cdot 1 + 0 + \tfrac{1}{2}\cdot 1 = 1.$$

(3) (12.324) with $i = j = 1$, $m = 0$; $S_{l0} = S_{0l}$.

(4) $S_{1l}/S_{0l} = (\sqrt{2},\, 0,\, -\sqrt{2})$; the middle term has $S_{11}^2 = 0$.

(5) $S_{2l}/S_{0l} = (1, -1, 1)$.

So $\frac{1}{2} \times \frac{1}{2} = 0 + 1$, and spin $\frac{1}{2}$ is not in $\frac{1}{2} \times \frac{1}{2}$ (it could not be: $\frac{1}{2} \otimes \frac{1}{2}$ contains only integer spins). Gluing pairs of pants with these coefficients gives the dimension formula (12.257) that we used from the start.

Dictionary: the S-matrix, four ways

Viewpoint	What S_{ij} is
2d CFT	modular transformation of characters, $\chi_i(-1/\tau) = \sum_j S_{ij}\chi_j(\tau)$
3d topology	gluing map of the genus-one splitting of S^3 (meridian $\leftrightarrow$ longitude)
knots	unnormalised Hopf link coloured (i, j); S_{0i} is the unknot, $S_{00} = Z(S^3)$
algebra	diagonalises fusion: $N_{ij}^{\ m} = \sum_l S_{jl}S_{il}S_{lm}/S_{0l}$

Reading Nash (pp. 357–360, (12.150)–(12.158))

Nash introduces the Hamiltonian formalism in one paragraph: $Z(\Sigma \times S^1) = \operatorname{tr}\exp(i\mathcal{H}t)$ (12.150), $\mathcal{H} = 0$ so $Z = \dim \mathcal{H}_\Sigma$ (12.151), marked points give unlinked unknotted circles with $\widetilde{W} = \dim \mathcal{H}_{\Sigma,\{R\}}$ (12.152), and a twisted product $\Sigma \times_K S^1$ has $Z = \operatorname{tr}(K)$ (12.153). He then chooses $M = S^2 \times S^1$, describes the inversion of $\mathbb{R}^3$ that swaps the inside and outside of a solid torus, identifies it with $S \in \mathrm{SL}(2,\mathbb{Z})/\{\pm\mathbb{1}\}$ from chapter 5, and obtains (12.154)–(12.156). The S-matrix (12.157) is quoted from Gepner and Witten, and (12.158) is the unknot, "in agreement with 12.138". We have added: the derivation of the trace formula; the $S^2 \times S^1$ dictionary including fusion and level truncation; mapping tori of T^2 with explicit traces; the derivation of (12.157) from the Weyl–Kac formula; orthogonality; quantum dimensions; the Hopf link; the Verlinde formula; and the branch issue in (12.158).

12.15.7 Invariants of 3-manifolds: the Reshetikhin–Turaev formula

Nash's surgery calculation used one surgery, on one very simple circle. The Lickorish–Wallace theorem of Section 12.15.3 says that *every* closed oriented 3-manifold is integer surgery on some framed link in S^3. So the question is now: *given a framed link $L \subset S^3$, what is the partition function of the manifold M_L that surgery on L produces?* If we can answer it, we have a formula for $Z(M)$ for every M, in terms of coloured link invariants in S^3, which the skein relation and the braiding already compute.

The answer was found by Witten on the physics side and by Reshetikhin and Turaev on the mathematics side (who built it rigorously from the quantum group $U_q(\mathfrak{sl}_2)$ at $q = e^{2\pi i/(k+2)}$). The result is often called the **Witten–Reshetikhin–Turaev** (WRT) invariant. We derive it in three steps: a single integer surgery is always the matrix S; many surgeries multiply; and a phase must be fixed.

Step 1: an integer surgery is always S

Let $C \subset M$ be a knot with a framing n. Its solid-torus neighbourhood M_2 has a meridian $m = \mu$ and, because of the framing, a preferred longitude $l = \lambda + n\mu$ (the edge of the ribbon; see Section 12.15.1). Integer surgery with framing n glues the new solid torus so that its meridian goes to exactly this curve l. In the basis (m, l) of ∂M_2 that is the map

$$m \mapsto l, \qquad \text{i.e.} \qquad K = S \quad (\text{possibly followed by } T^a).$$

Why "possibly followed by T^a"? Only the image of the meridian matters (Section 12.15.2), and ST^a also sends $m \mapsto l$, because T fixes m. On the Hilbert space T^a acts on the empty solid torus by the phase $e^{-2\pi i ac/24}$, since V^0 has $h_0 = 0$. So the choice of K is unique up to such phases, and we postpone them to Step 3.

Now apply Nash's surgery formula (12.310):

$$Z(M_C) \overset{(1)}{=} \langle V_1,\, S\,V^0 \rangle \overset{(2)}{=} \sum_j S_{0j} \langle V_1, V^j \rangle \overset{(3)}{=} \sum_j S_{0j}\; \widetilde{W}\big(M; C_j^{(n)}\big). \tag{12.325}$$

(1) (12.305) with $K = S$ and $V_2 = V^0$, the empty solid torus.
(2) Expand $SV^0 = \sum_j S^0{}_j V^j$ with $S^0{}_j = S_{0j}$ (Section 12.15.4).
(3) $\langle V_1, V^j \rangle$ glues back the *original* solid torus with a Wilson line R_j on its core. The core line of V^j is pushed off along the longitude l of M_2, so as a framed knot it is C with framing n; we write $C_j^{(n)}$.

So one integer surgery replaces the knot by a sum over all colours, weighted by the first row of S. This combination deserves a name.

Working definition 12.48 (the Kirby colour Ω)

The **Kirby colour** is the vector

$$\Omega := S\,V^0 = \sum_{j=0}^{k} S_{0j}\, V^j \;\in \mathcal{H}_{T^2}.$$

"Colouring a component C of a link by Ω" means: compute the coloured link invariant with C in each representation R_j, multiply by S_{0j}, and sum over j. Physically, Ω is the state on the boundary of a solid torus that is *empty* as seen from the other side of the gluing: it is the vacuum of the dual solid torus, rewritten as a superposition of all charges on the core. For $\mathrm{SU}(2)_k$, $S_{0j} = S_{00}\, d_j$ with $d_j = [j+1]_q$ the quantum dimension [$d_j = \sin\frac{(j+1)\pi}{k+2} / \sin\frac{\pi}{k+2}$, which tends to the ordinary dimension $j+1$ of R_j as $k \to \infty$] (Example 12.126), so each colour is weighted by its quantum dimension.

Worked example 12.129: The Kirby colour kills every charge that threads it

Take any link containing a strand of colour R_i, and add a small 0-framed unknot U encircling that strand once. Colour U by Ω. From Example 12.128, a meridian loop of colour R_j around a line of colour R_i multiplies the correlator by S_{ji}/S_{0i}. Summing with the Kirby weights:

$$\sum_j S_{0j}\,\frac{S_{ji}}{S_{0i}} \overset{(1)}{=} \frac{(S^2)_{0i}}{S_{0i}} \overset{(2)}{=} \frac{\delta_{0i}}{S_{0i}} = \frac{\delta_{i0}}{S_{00}}.$$

(1) Matrix multiplication: $\sum_j S_{0j} S_{ji} = (S^2)_{0i}$.
(2) $S^2 = \mathbb{1}$ for SU(2) (Example 12.125).

An Ω-coloured unknot acts as a *projector onto zero total charge* through its disc (Figure 12.88). The topology says the same thing. Surgery on a 0-framed unknot that bounds a disc meeting the strand once takes a connected sum with $S^2 \times S^1$, and the strand now pierces the sphere $S^2 \times \{\mathrm{pt}\}$ exactly once. A single charge on a closed sphere violates Gauss's law unless it is neutral, i.e. $\dim \mathcal{H}_{S^2;R_i} = \delta_{i0}$ (Nash's (11.75)). For $i = 0$ the factor $1/S_{00}$ is the connected-sum rule $Z(M\#N) = Z(M)Z(N)/Z(S^3)$ of Section 12.14.9, with $Z(S^2 \times S^1) = 1$.

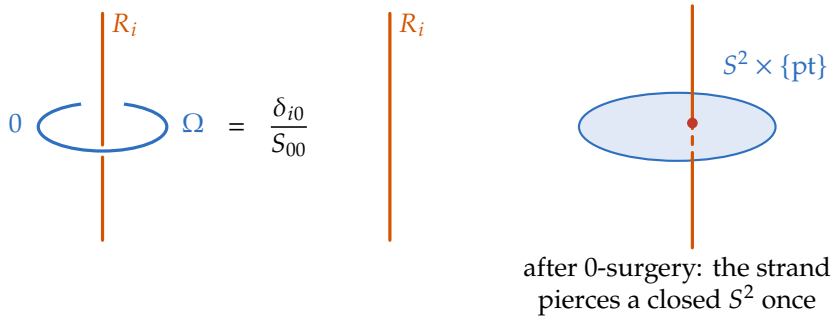


Figure 12.88: The killing property of the Kirby colour (Example 12.129). Left: a strand of colour R_i threading a 0-framed, Ω-coloured unknot equals δ_{i0}/S_{00} times the strand alone. Right: why. Surgery on that unknot creates an $S^2 \times S^1$ summand, and the strand pierces its sphere once (red dot). The sphere is drawn as the disc the unknot bounded, which the surgery caps off; it does not separate $S^2 \times S^1$, so a single piercing is possible. Gauss's law on a closed sphere allows only zero total charge.

Step 2: many components

Let $L = C_1 \cup \cdots \cup C_m$ be a framed link in S^3. The solid-torus neighbourhoods of the components are disjoint, so we can do the surgeries one at a time. After the first surgery, the other components are Wilson lines in the part M_1 that is carried along unchanged (variant (c) of Section 12.15.4). Repeating (12.325) m times,

$$Z_{\text{naive}}(M_L) = \sum_{j_1,\dots,j_m=0}^{k} \Big(\prod_{a=1}^{m} S_{0j_a}\Big)\, \widetilde{W}\big(S^3; L_{j_1\dots j_m}\big), \tag{12.326}$$

where $L_{j_1\dots j_m}$ is the link with component C_a in the representation R_{j_a} and with its framing n_a, and $\widetilde{W}$ is the *unnormalised* correlator (so the empty link gives $\widetilde{W}(S^3;\varnothing) = Z(S^3) = S_{00}$). In words: *colour every component by* Ω.

The framings and the linkings of L are collected in one matrix.

Working definition 12.49 (the linking matrix and its signature)

For a framed link $L = C_1 \cup \cdots \cup C_m$, the **linking matrix** is the symmetric integer matrix

$$B_{aa} = n_a \ (\text{framing of } C_a), \qquad B_{ab} = \text{lk}(C_a, C_b)\ (a \neq b).$$

Here n_a is the framing [the linking number of C_a with its push-off] of C_a. Its **signature** $\sigma(B)$ is the number of positive eigenvalues minus the number of negative ones. Two useful facts: $|\det B|$ is the order of $H_1(M_L)$ [the number of elements of the group $H_1(M_L;\mathbb{Z})$] when it is finite (and $\det B = 0$ exactly when $H_1(M_L)$ is infinite), and B is the intersection form of the 4-manifold W_L obtained by attaching a 2-handle [a copy of $D^2 \times D^2$, glued onto $\partial B^4 = S^3$ along a thickened C_a with framing n_a] to the 4-ball along each C_a, whose boundary is M_L.

For example, the Hopf link with framings $(n, 0)$ has $B = \left(\begin{smallmatrix} n & 1 \\ 1 & 0 \end{smallmatrix}\right)$, with $\det B = -1$, so $H_1 = 0$. Its eigenvalues are $\frac{1}{2}\big(n \pm \sqrt{n^2+4}\big)$, one positive and one negative, and $\sigma = 0$. A single unknot with framing p has $B = (p)$ and $\sigma = \operatorname{sgn} p$.

Step 3: the Kirby test and the framing anomaly

A formula for $Z(M)$ must give the same answer for two framed links related by Kirby moves (Figure 12.83). Test (12.326) on the simplest move, K1: add an isolated unknot $U_\pm$ with framing ± 1. Surgery on it gives back S^3, and a connected sum with S^3 does nothing, so Z should not change. What does (12.326) say? An isolated component factorises by the connected-sum rule of Section 12.14.9: $\widetilde{W}(L \sqcup U_j^{(\pm 1)}) = \widetilde{W}(L)\,\widetilde{W}(U_j^{(\pm 1)})/S_{00}$. The framing rule (12.301) gives $\widetilde{W}(U_j^{(\pm 1)}) = \theta_j^{\pm 1} S_{0j}$ with $\theta_j := e^{2\pi i h_j}$. So adding $U_\pm$ multiplies Z_{naive} by

$$\frac{1}{S_{00}} \sum_j S_{0j}^2\, \theta_j^{\pm 1}. \tag{12.327}$$

This is a **Gauss sum**, and we can evaluate it exactly.

Worked example 12.130: The Gauss sum is a phase times S_{00}

In Example 12.119 we found that S and $T' = \mathrm{diag}(\theta_j)$ obey $(ST')^3 = e^{2\pi i c/8}\,\mathbb{1}$ rather than $\mathbb{1}$. Rearranging,

$$ST'ST'ST' = e^{2\pi i c/8}\,\mathbb{1} \overset{(1)}{=} ST'S = e^{2\pi i c/8}\,T'^{-1}S^{-1}T'^{-1} \overset{(2)}{=} e^{2\pi i c/8}\,T'^{-1}ST'^{-1}.$$

(1) Multiply on the right by $T'^{-1}S^{-1}T'^{-1}$.
(2) $S^{-1} = S$ because $S^2 = \mathbb{1}$.

Take the $(0,0)$ entry of both sides:

$$\sum_j S_{0j}\,\theta_j\,S_{j0} \overset{(3)}{=} e^{2\pi i c/8}\,\theta_0^{-1} S_{00}\,\theta_0^{-1} \overset{(4)}{=} e^{2\pi i c/8}\,S_{00}. \tag{12.328}$$

(3) T' is diagonal, so $(ST'S)_{00} = \sum_j S_{0j}\theta_j S_{j0}$ and $(T'^{-1}ST'^{-1})_{00} = \theta_0^{-1}S_{00}\theta_0^{-1}$.
(4) $h_0 = 0$, so $\theta_0 = 1$.

The complex conjugate gives $\sum_j S_{0j}^2\theta_j^{-1} = e^{-2\pi i c/8}S_{00}$. Check at $k=1$, where $S_{0j}^2 = \frac{1}{2}$, $\theta = (1, i)$ and $c = 1$: $\sum_j S_{0j}^2\theta_j = \frac{1+i}{2} = \frac{e^{i\pi/4}}{\sqrt{2}} = e^{2\pi i/8}S_{00}$. Figure 12.89 shows the sum at $k=3$ as a chain of arrows in the complex plane.

So K1 fails, but only by a phase: adding $U_\pm$ multiplies Z_{naive} by $e^{\pm 2\pi i c/8}$. The fix is to find a quantity built from the framed link that also changes by ± 1 under K1 and does not change under K2. The signature does exactly this.

- *K1.* Adding $U_\pm$ adds a block (± 1) to the linking matrix, $B \mapsto B \oplus (\pm 1)$, so $\sigma \mapsto \sigma \pm 1$.
- *K2.* Sliding C_a over C_b replaces the homology class $[C_a]$ by $[C_a] + [C_b]$. With new basis vectors $f_a = e_a + e_b$, $f_b = e_b$,

$$B'_{aa} \overset{(1)}{=} (e_a + e_b)^{\mathsf{T}} B\,(e_a + e_b) = n_a + n_b + 2\,\mathrm{lk}(C_a, C_b),$$
$$B'_{ab} \overset{(2)}{=} (e_a + e_b)^{\mathsf{T}} B\,e_b = \mathrm{lk}(C_a, C_b) + n_b.$$

(1) The framing of a band sum is the self-intersection of the sum of the classes: expand the bilinear form.
(2) The pushed-off copy of C_b links C_b exactly n_b times; this is what the framing means.

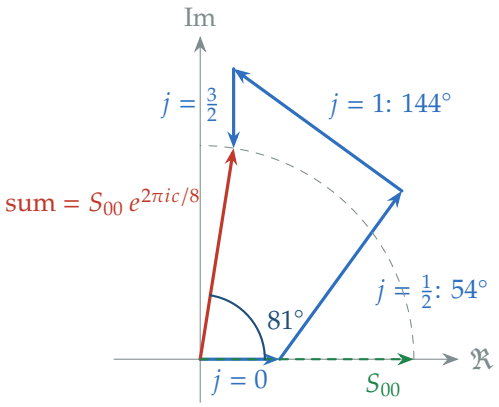


Figure 12.89: The Gauss sum $\sum_j S_{0j}^2 \theta_j$ at level $k = 3$, drawn as a chain of arrows: arrow j has length S_{0j}^2 and direction $2\pi h_j$ ($h_j = 0, \frac{3}{20}, \frac{8}{20}, \frac{15}{20}$). The red resultant has exactly the length $S_{00} \approx 0.372$ that an invariant would need (dashed circle), but it points at $81° = 2\pi c/8$ with $c = \frac{9}{5}$ instead of along the green real axis. Adding a +1-framed unknot therefore rotates the naive surgery sum by $e^{2\pi i c/8}$: this is the framing anomaly.

This is the rule quoted in Section 12.15.3. In matrix form $B' = P^{\mathsf{T}} B P$ with P an integer matrix of determinant 1. By Sylvester's law of inertia a congruence does not change the numbers of positive and negative eigenvalues: $\sigma(B') = \sigma(B)$.

Topologically: K2 is just an isotopy inside M_L (the strand of C_a is pushed across the meridian disc of the new solid torus glued in along C_b), so the path integral cannot see it, and the only thing that can go wrong is the phase. K1 changes the 4-manifold W_L by a connected sum with $\pm\mathbb{CP}^2$, which changes its signature by ± 1.

Key idea: The Witten–Reshetikhin–Turaev invariant

Let M_L be integer surgery on a framed link $L = C_1 \cup \cdots \cup C_m \subset S^3$ with linking matrix B. Then

$$Z(M_L) = e^{-2\pi i\, c\, \sigma(B)/8} \sum_{j_1,\dots,j_m=0}^{k} \Big(\prod_{a=1}^{m} S_{0 j_a}\Big)\, \widetilde{W}\big(S^3; L_{j_1 \dots j_m}\big), \qquad c = \frac{3k}{k+2}, \tag{12.329}$$

is independent of the choice of L: it is an invariant of the 3-manifold. It is normalised so that $Z(S^3) = S_{00}$ and $Z(S^2 \times S^1) = 1$.

Physical picture: The phase is a gravitational Chern–Simons counterterm

Witten found the same phase from the path integral. To define the Chern–Simons theory at the quantum level, one must choose a metric for the gauge fixing. The metric dependence cancels *except* for a phase $\exp\big(\frac{ic}{24} \cdot \frac{1}{4\pi} \int \operatorname{tr}(\omega\, \mathrm{d}\omega + \frac{2}{3}\omega^3)\big)$, with ω the Levi-Civita connection: the gravitational Chern–Simons term, with the coefficient $c/24$ that is the central charge of the boundary CFT. The phase is removed by choosing a framing of the tangent bundle of M (a **2-framing** in Atiyah's sense), and changing the framing by s units multiplies Z by $e^{2\pi i c s/24}$. Atiyah showed that every M has a *canonical* 2-framing. If $M = \partial W$, the gravitational Chern–Simons term of the canonical framing is fixed by the first Pontryagin number of W, and $\int_W p_1 = 3\sigma(W)$ (Hirzebruch). So the phase is $e^{2\pi i\,(c/24)\cdot 3\sigma(W)} = e^{2\pi i c\sigma/8}$, which is exactly the factor in (12.329). The $c/24$ of the Virasoro algebra, the $c/8$ of $(ST')^3$ and the signature of a 4-manifold are one and the same anomaly.

Worked example 12.131: Kirby checks of the formula

We check (12.329) on the surgeries of Section 12.15.3.

(i) The empty link and the 0-framed unknot. With no components the sum is $\widetilde{W}(S^3;\varnothing) = S_{00} = Z(S^3)$. With one 0-framed unknot ($\sigma = 0$),

$$Z \overset{(1)}{=} \sum_j S_{0j}\,\widetilde{W}(S^3;U_j^{(0)}) \overset{(2)}{=} \sum_j S_{0j}^2 \overset{(3)}{=} 1 = Z(S^2\times S^1).$$

(1) (12.329) with $m = 1$, $B = (0)$.
(2) The zero-framed unknot is S_{0j} (Section 12.15.6).
(3) Row 0 of the orthogonal matrix S is a unit vector.

(ii) The ±1-framed unknot. With $B = (\pm1)$, $\sigma = \pm1$:

$$Z \overset{(4)}{=} e^{\mp 2\pi i c/8}\sum_j S_{0j}^2\theta_j^{\pm1} \overset{(5)}{=} e^{\mp 2\pi i c/8}\,e^{\pm 2\pi i c/8}S_{00} = S_{00} = Z(S^3).$$

(4) (12.329); the framing ±1 contributes $\theta_j^{\pm1}$.
(5) The Gauss sum (12.328) and its conjugate.

(iii) Blowing down a linked $+1$ unknot. This is the general form of K1: a $+1$-framed unknot U that links one strand of colour R_i once. Kirby calculus says we may delete U provided we lower the framing of the strand by $\mathrm{lk}^2 = 1$. Colour U by Ω. The meridian loop R_j with framing $+1$ multiplies by $\theta_j S_{ji}/S_{0i}$, so the sum over j is

$$\sum_j S_{0j}\,\theta_j\,\frac{S_{ji}}{S_{0i}} \overset{(6)}{=} \frac{(ST'S)_{0i}}{S_{0i}} \overset{(7)}{=} \frac{e^{2\pi i c/8}\,\theta_0^{-1}S_{0i}\,\theta_i^{-1}}{S_{0i}} = e^{2\pi i c/8}\,\theta_i^{-1}.$$

(6) The sum is the $(0,i)$ entry of $ST'S$.
(7) $ST'S = e^{2\pi i c/8}T'^{-1}ST'^{-1}$ from Example 12.130.

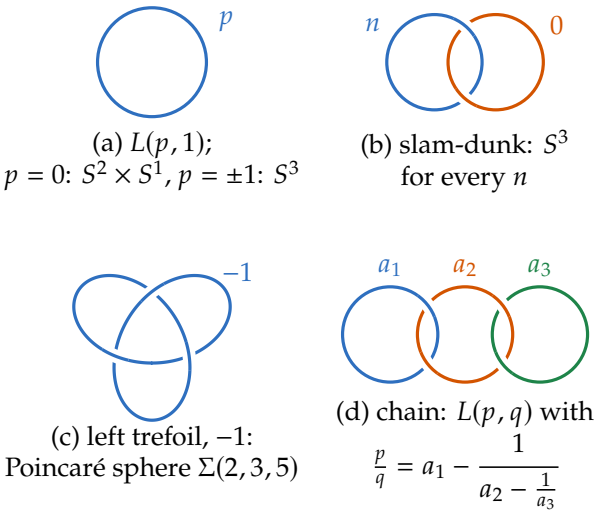
The factor θ_i^{-1} lowers the framing of the strand by one, exactly as Kirby's rule demands, and the phase $e^{2\pi i c/8}$ is cancelled by the change of signature: the linking matrix loses one positive eigenvalue. For instance the chain of two unknots with framings $(a,1)$ must give $L(a-1,1)$. We checked numerically at $k = 1,2,3$ and $a = 2,3,4$ that (12.329) gives identical values for both presentations.

(iv) The slam-dunk. For the Hopf link with framings $(n,0)$ ($\sigma = 0$), the zero-framed Hopf link is S_{ij} (Example 12.127), and the framing n on the first component gives θ_i^n:

$$Z \overset{(8)}{=} \sum_{i,j} S_{0i}S_{0j}\,S_{ij}\,\theta_i^n \overset{(9)}{=} \sum_i S_{0i}\,\theta_i^n\,(S^2)_{i0} \overset{(10)}{=} S_{00}\,\theta_0^n = S_{00}.$$

(8) (12.329) with two components.
(9) Do the sum over j first: $\sum_j S_{ij}S_{j0} = (S^2)_{i0}$.
(10) $(S^2)_{i0} = \delta_{i0}$ and $\theta_0 = 1$.

This is $Z(S^3)$ for every n, as the slam-dunk of Example 12.117 says it must be. The 0-framed component acts as the projector of Example 12.129 on the other one.



(a) $L(p,1)$; $p=0$: $S^2\times S^1$, $p=\pm1$: S^3

(b) slam-dunk: S^3 for every n

(c) left trefoil, -1: Poincaré sphere $\Sigma(2,3,5)$

(d) chain: $L(p,q)$ with $\frac{p}{q}=a_1-\dfrac{1}{a_2-\frac{1}{a_3}}$

Figure 12.90: A gallery of surgery diagrams; the numbers are the framings. (a) One unknot gives a lens space. (b) A Hopf link with framings $(n,0)$ gives S^3 (the 0-framed circle "dunks" the other). (c) -1 surgery on the left-handed trefoil gives the Poincaré homology sphere. (d) A linear chain of unknots gives the lens space whose p/q is the continued fraction of the framings (up to orientation). Every closed oriented 3-manifold has such a picture, and (12.329) turns the picture into a number.

Worked example 12.132: Lens spaces $L(p,1)$

$L(p,1)$ is p-surgery on the unknot (Example 12.115). For $p>0$, $\sigma=1$ and

$$Z\big(L(p,1)\big)=e^{-2\pi i c/8}\sum_{j=0}^{k}S_{0j}^2\,\theta_j^{\,p}. \tag{12.330}$$

(For $p<0$ take the complex conjugate: $L(-p,1)$ is $L(p,1)$ with the orientation reversed.) In operator language this is $\langle V^0, ST'^{\,p}S\,V^0\rangle$ up to the phase: the manifold is two solid tori glued by ST^pS, in agreement with Section 12.15.2.

Level 1. Here $S_{0j}^2=\frac12$, $\theta=(1,i)$, $c=1$:

$$Z\big(L(p,1)\big)\overset{(1)}{=}e^{-i\pi/4}\,\tfrac12\big(1+i^{\,p}\big),$$

(1) (12.330) with $e^{-2\pi i/8}=e^{-i\pi/4}$ and $\theta_{1/2}^{\,p}=i^p$.

which gives $\frac{1}{\sqrt2}$, 0, $-\frac{i}{\sqrt2}$, $e^{-i\pi/4}$ for $p=1,2,3,4$. The value at $p=1$ is $Z(S^3)$, as it must be.

Level 2, $p=2$ ($\mathbb{RP}^3=\mathrm{SO}(3)$). Here $S_{0j}^2=(\frac14,\frac12,\frac14)$, $h=(0,\frac{3}{16},\frac12)$, $c=\frac32$:

$$Z(\mathbb{RP}^3)\overset{(2)}{=}e^{-3\pi i/8}\Big(\tfrac14+\tfrac12e^{3\pi i/4}+\tfrac14e^{2\pi i}\Big)\overset{(3)}{=}\tfrac12\big(e^{-3\pi i/8}+e^{3\pi i/8}\big)\overset{(4)}{=}\cos\frac{3\pi}{8}\approx0.383.$$

(2) $\theta_j^2=e^{4\pi i h_j}=(1,e^{3\pi i/4},e^{2\pi i})$ and $e^{-2\pi i c/8}=e^{-3\pi i/8}$.

(3) $\frac14+\frac14=\frac12$; distribute the prefactor.

(4) $e^{ix}+e^{-ix}=2\cos x$.

A table of (12.330) (computed numerically):

	$p=1$ (S^3)	$p=2$ ($\mathbb{RP}^3$)	$p=3$	$p=4$
$k=1$	0.7071	0	$-0.7071\,i$	$0.7071-0.7071\,i$
$k=2$	0.5	0.3827	$-0.3536+0.3536\,i$	$-0.2706-0.6533\,i$
$k=3$	0.3717	0	$0.5721+0.1859\,i$	$-0.4684+0.2387\,i$

Two features are worth noticing. First, $Z(\mathbb{RP}^3) = 0$ at odd level (we checked $k = 1, 3, 5, 7, 9$). In the Gauss sum the terms with j and $\frac{k}{2} - j$ cancel in pairs. Second, Z distinguishes the lens spaces, even though at level 1 it only takes the few values allowed by an abelian Gauss sum.

The Poincaré sphere

Here is a manifold that is not a lens space. The Poincaré homology sphere $\Sigma(2, 3, 5)$ has the homology of S^3 but a fundamental group of order 120 (Example 12.116). It is -1 surgery on the left-handed trefoil (Figure 12.90(c)).

We need the *coloured* trefoil. The left trefoil is the closure of the 2-braid σ^{-3}. Colour both strands (they are one knot) by spin j. The two strands can fuse into total spin $l \in \{0, 1, \dots, 2j\}$, with $l \le k - 2j$ at level k. In each channel the half-twist σ acts by the number

$$\lambda_l^{(j)} = (-1)^{2j-l}\, e^{i\pi(h_l - 2h_j)} = (-1)^{2j-l}\, q^{\,l(l+1)/2 - j(j+1)}, \tag{12.331}$$

which is (12.287) with the same two ingredients: the exchange phase read off from the operator product (Section 12.14.12), and the symmetry sign of channel l in $j \otimes j$ (the top channel $l = 2j$ is symmetric, and the sign alternates downwards). For $j = \frac{1}{2}$ it gives $\lambda_0 = -q^{-3/4}$ and $\lambda_1 = q^{1/4}$ as before. The quantum trace (12.300) then gives the blackboard-framed value (writhe -3), and we raise the framing from -3 to -1 with θ_j^2:

$$\widetilde{W}\big(S^3; \mathrm{LT}_j^{(-1)}\big) = S_{00}\, \theta_j^2 \sum_l [2l+1]\, \big(\lambda_l^{(j)}\big)^{-3}. \tag{12.332}$$

Worked example 12.133: $Z(\Sigma(2,3,5))$ at levels 1 and 2

With $B = (-1)$, $\sigma = -1$, and (12.329) reads $Z = e^{+2\pi i c/8} \sum_j S_{0j}\, \widetilde{W}(S^3; \mathrm{LT}_j^{(-1)})$.

Level 1 ($q = e^{2\pi i/3}$, $c = 1$). The colour $j = 0$ contributes $S_{00}^2 = \frac{1}{2}$. For $j = \frac{1}{2}$ only the channel $l = 0$ survives ($l \le k - 2j = 0$), with $\lambda_0 = -q^{-3/4}$ and $[1] = 1$:

$$\theta_{1/2}^2\, \lambda_0^{-3} \overset{(1)}{=} q^{3/2} \cdot \big(-q^{9/4}\big) \overset{(2)}{=} -q^{15/4} \overset{(3)}{=} -e^{2\pi i \cdot 5/4} = -i,$$

$$Z\big(\Sigma(2,3,5)\big) \overset{(4)}{=} e^{i\pi/4}\Big(\tfrac{1}{2} + \tfrac{1}{2}(-i)\Big) \overset{(5)}{=} e^{i\pi/4}\, \frac{e^{-i\pi/4}}{\sqrt{2}} = \frac{1}{\sqrt{2}}.$$

(1) $\theta_{1/2} = q^{3/4}$ and $(-q^{-3/4})^{-3} = -q^{9/4}$.
(2) Add the exponents.
(3) $q^{15/4} = e^{2\pi i \cdot 15/12}$.
(4) $S_{0,1/2} S_{00} = \frac{1}{2}$ and $e^{2\pi i c/8} = e^{i\pi/4}$.
(5) $1 - i = \sqrt{2}\, e^{-i\pi/4}$.

So $Z(\Sigma(2,3,5)) = Z(S^3)$ at level 1. That is expected: at $k = 1$ the theory is equivalent to an abelian one, which only sees H_1, and $H_1 = 0$ for both manifolds.

Level 2 ($q = i$, $c = \frac{3}{2}$, $S_{0j} = (\frac{1}{2}, \frac{\sqrt{2}}{2}, \frac{1}{2})$, $[3]_{q=i} = 1$). Colour by colour:

- $j = 0$: $S_{00} \cdot S_{00} = \frac{1}{4}$.
- $j = \frac{1}{2}$: channels $l = 0, 1$. Using $\theta^2(\lambda_0^{-3} + [3]\lambda_1^{-3}) = q^{3/2}(-q^{9/4} + q^{-3/4}) = e^{3\pi i/8} - e^{15\pi i/8} = \sqrt{2}\, e^{5\pi i/8}$, the contribution is $\frac{\sqrt{2}}{2} \cdot \frac{1}{2} \cdot \sqrt{2}\, e^{5\pi i/8} = \frac{1}{2} e^{5\pi i/8}$.

- $j = 1$: only $l = 0$ ($l \le k - 2j = 0$), $\lambda_0^{(1)} = q^{-2} = -1$, $\theta_1^2 = q^4 = 1$: contribution $\frac{1}{2} \cdot \frac{1}{2} \cdot (-1) = -\frac{1}{4}$.

The first and last cancel, and

$$Z\big(\Sigma(2,3,5)\big) \overset{(6)}{=} e^{3\pi i/8} \cdot \tfrac{1}{2} e^{5\pi i/8} \overset{(7)}{=} -\tfrac{1}{2}.$$

(6) $e^{2\pi i c/8} = e^{3\pi i/8}$ times the sum.
(7) $e^{i\pi} = -1$.

Now $Z(S^3) = +\frac{1}{2}$ at $k = 2$: the invariant distinguishes the Poincaré sphere from S^3 although their homology is the same. At higher levels the values (numerically) are

k	3	4	5	6
$Z(\Sigma(2,3,5))$	$-0.3008 + 0.9256\,i$	0.2887	$-0.8460 + 0.0448\,i$	$0.7325\,i$
$Z(S^3)$	0.3717	0.2887	0.2319	0.1913

+1 surgery on the right-handed trefoil gives $\Sigma(2,3,5)$ with the opposite orientation, and the complex-conjugate values. The invariant is *chiral*: it knows the orientation of M.

Physical picture: What $|Z|$ remembers: flat connections

Why is $|Z(\Sigma(2,3,5))|$ so much larger than $Z(S^3)$? At large k the path integral localises on flat connections, as in the weak-coupling analysis of the previous section (Nash's (12.107)–(12.113)). S^3 has only the trivial one. The Poincaré sphere has in addition two irreducible flat SU(2) connections, coming from the two irreducible 2-dimensional representations of its fundamental group (the binary icosahedral group). Their Chern–Simons invariants are $-\frac{1}{120}$ and $-\frac{49}{120}$ mod 1. Each irreducible connection contributes a term of size $O(1)$ (no zero modes) with phase $e^{2\pi i(k+2)\,\mathrm{CS}}$, while the trivial connection contributes only $O(k^{-3/2})$. So $|Z|$ oscillates at order 1, and its oscillation frequencies are the Chern–Simons invariants. The Floer and Casson story of the start of this chapter counts the same irreducible flat connections.

Worked example 12.134: Large k: $Z(S^3)$ and the zero modes of the trivial connection

The exact answer is $Z(S^3) = S_{00} = \sqrt{2/(k+2)}\,\sin(\pi/(k+2))$. For large k,

$$Z(S^3) \overset{(1)}{=} \sqrt{\frac{2}{k+2}}\left(\frac{\pi}{k+2} - \frac{\pi^3}{6(k+2)^3} + \dots\right) \overset{(2)}{=} \sqrt{2}\,\pi\,(k+2)^{-3/2}\Big(1 + O(k^{-2})\Big).$$

(1) Taylor expansion $\sin x = x - x^3/6 + \dots$ with $x = \pi/(k+2)$.
(2) Collect the powers of $k+2$.

Numerically, at $k = 10, 50, 100$ the exact values are 0.10566, 0.011841, 0.0043122, and $\sqrt{2}\pi(k+2)^{-3/2}$ gives 0.10688, 0.011848, 0.0043129; with k in place of $k+2$ the approximation is much worse (0.1405, 0.01257, 0.00444).
The power $k^{-3/2}$ is exactly what the weak-coupling formula predicts, once the assumption stated at the end of Nash's §5 is dropped. S^3 is simply connected, so the only flat connection is $A = 0$, and at $A = 0$ the cohomology of the flat bundle is

$$H^0(S^3; \mathfrak{su}(2)) = \mathfrak{su}(2) \text{ (constant gauge transformations)}, \qquad H^1(S^3; \mathfrak{su}(2)) = 0.$$

Nash's Gaussian formula assumed all H^p vanish. Here the ghost operator $\mathrm{d}^*\mathrm{d}$ on Ω^0 has

dim $H^0 = 3$ zero modes: the constant gauge transformations that leave $A = 0$ fixed, i.e. the stabiliser SU(2). They must be removed from the determinant and replaced by the volume of the stabiliser. A Gaussian mode of the action $\frac{k}{4\pi}\int \ldots$ contributes a factor $\propto k^{-1/2}$. After the regularised count the net power is $k^{-(\dim H^0 - \dim H^1)/2} = k^{-3/2}$, and the constant in front ($\sqrt{2}\pi$) comes from the volume of SU(2) and the Ray–Singer torsion of S^3. The shift $k \to k+2$ is the one-loop correction (the η-invariant phase of Nash's (12.106)), and the exact sin is what the one-loop answer cannot see. The shift is derived in Section T5.7.2.

Watch out: Normalisations of $Z(M)$ differ by phases

Different books multiply (12.329) by different phases and powers of S_{00}. For instance, some normalise $Z(S^3) = 1$ by dividing by S_{00} once per component and once overall, and some use a different framing of M and so a different power of $e^{2\pi i c/24}$. Two things are convention-independent: $|Z(M)|$, and ratios $Z(M)/Z(M')$ computed in one convention. Reversing the orientation of M always complex-conjugates Z.

Reading Nash (pp. 355–360: the invariant $Z(M)$)

Nash states the aim of §7 in one sentence: surgery lets Witten "generalise V_L to an arbitrary 3-manifold M" and gives "a new way to compute V_L for S^3". He works out the surgery on $S^2 \times S^1$ and does not write the general formula. The general formula (12.329) and its checks are therefore additions here. They use only Nash's (12.142)–(12.149), the S-matrix (12.157), and the framing rule at the end of §6. The Kirby-colour sum is the many-component version of Nash's (12.148), and the signature phase is the framing anomaly that Nash's footnote 9 anticipates.

12.15.8 Beyond SU(2): the HOMFLY polynomial

Everything in Section 12.14 was done for $G = \mathrm{SU}(2)$ and Wilson lines in the 2-dimensional representation. Nash ends his account with one sentence: for $G \neq \mathrm{SU}(2)$ we obtain a generalisation of the Jones polynomial, and for $G = \mathrm{SU}(N)$ with N allowed to vary, a *two-variable* polynomial, which is the HOMFLY polynomial of Freyd et al. once the Wilson lines carry the defining representation. The question of this subsection is: *what exactly changes when* SU(2) *becomes* SU(N), *and why does the answer depend on two variables?*

The strategy of Section 12.14.11 was a chain of four steps:

(a) count the dimension of the Hilbert space of a sphere with four punctures: it was 2;

(b) compute the two eigenvalues of the half-twist σ from the conformal weights;

(c) a 2×2 matrix satisfies a quadratic equation, and that equation is the skein relation;

(d) correct for framing with the twist phase $\theta = e^{2\pi i h}$.

We now run exactly these four steps for SU(N). Only the numbers change.

The data of $\mathrm{SU}(N)$ at level k

Working definition 12.50 ($\mathrm{SU}(N)_k$ data needed for knots)

- *Deformation parameter.* The dual Coxeter number [the integer $h^\vee$ that is added to k in the exact formulas, $h^\vee = 2$ for SU(2)] of $\mathrm{SU}(N)$ is $h^\vee = N$, and the quantum shift is $k \to k + N$. We set

$$q := \exp\Big(\frac{2\pi i}{k+N}\Big), \qquad [n]_q = \frac{q^{n/2} - q^{-n/2}}{q^{1/2} - q^{-1/2}}.$$

  For $N = 2$ this is the q of (12.248). The quantum integer $[n]_q$ [equals $\sin\frac{n\pi}{k+N} / \sin\frac{\pi}{k+N}$, and tends to n as $k \to \infty$].

- *Integrable representations.* A representation of $\mathrm{SU}(N)$ is labelled by a Young diagram [left-justified rows of boxes, each row no longer than the one above. One box is the defining representation] with at most $N-1$ rows. At level k it is integrable if its first row has at most k boxes.

- *Conformal weights* (Sugawara, as in Chapter 11): $h_\lambda = C_2(\lambda)/(k+N)$, with the quadratic Casimir C_2 [the number with $\sum_a T^a T^a = C_2(\lambda)\,\mathbb{1}$ in the representation λ] normalised so that $\mathrm{tr}(T^a T^b) = \frac{1}{2}\delta^{ab}$ in the fundamental representation.

- *Central charge*: $c = k(N^2-1)/(k+N)$.

We need the Casimir of three representations: the fundamental $V = \mathbb{C}^N$, and the two pieces of $V \otimes V$. Write $\mathrm{Sym}^2 V$ (symmetric tensors, Young diagram $\square\square$, dimension $N(N+1)/2$) and $\Lambda^2 V$ (antisymmetric tensors, Young diagram $\begin{smallmatrix}\square\\\square\end{smallmatrix}$, dimension $N(N-1)/2$). For $N = 2$ these are the triplet and the singlet.

Worked example 12.135: Casimirs from the Fierz identity

The N^2 matrices $\mathbb{1}/\sqrt{N}$ and $\sqrt{2}\,T^a$ ($a = 1, \dots, N^2-1$) are an orthonormal basis of all $N \times N$ matrices for the inner product $\langle X, Y\rangle = \mathrm{tr}(X^\dagger Y)$. Completeness of an orthonormal basis, $\sum_B B_{ij}\overline{B_{lk}} = \delta_{il}\delta_{jk}$, reads

$$\frac{1}{N}\,\delta_{ij}\delta_{kl} + 2\sum_a (T^a)_{ij}(T^a)_{kl} = \delta_{il}\delta_{kj}$$

(we used that T^a is Hermitian, so $\overline{(T^a)_{lk}} = (T^a)_{kl}$). As an operator on $V \otimes V$, the right side is the swap $P(u \otimes v) = v \otimes u$. So

$$\sum_a T^a \otimes T^a = \frac{1}{2}\Big(P - \frac{1}{N}\,\mathbb{1}\Big). \tag{12.333}$$

This is the $\mathrm{SU}(N)$ version of $\vec{S}_1 \cdot \vec{S}_2 = \frac{1}{2}(P - \frac{1}{2})$ for two spins $\frac{1}{2}$. Now compute Casimirs. On the fundamental,

$$C_2(V)\,\mathbb{1} \overset{(1)}{=} \sum_a (T^a T^a)_{ij} \overset{(2)}{=} \frac{1}{2}\Big(N\,\delta_{ij} - \frac{1}{N}\delta_{ij}\Big) = \frac{N^2-1}{2N}\,\delta_{ij}.$$

(1) Definition of the Casimir: $C_2 = \sum_a T^a T^a$, a multiple of $\mathbb{1}$ on an irreducible representation (Schur).

(2) Set $j = k$ in the index form of the Fierz identity and sum over j: $\delta_{il}\delta_{jj} \to N\delta_{il}$ and $\delta_{ij}\delta_{jl} \to \delta_{il}$, then divide by 2.

On $V \otimes V$ the generators are $T^a \otimes \mathbb{1} + \mathbb{1} \otimes T^a$, so

$$C_2(V \otimes V) \overset{(1)}{=} C_2(V) \otimes \mathbb{1} + \mathbb{1} \otimes C_2(V) + 2\sum_a T^a \otimes T^a \overset{(2)}{=} \frac{N^2-1}{N} + P - \frac{1}{N}.$$

(1) Square the sum of the two generators and collect the cross terms.

(2) $C_2(V) = \frac{N^2-1}{2N}$ on each factor, and (12.333).

The swap is $P = +1$ on symmetric and $P = -1$ on antisymmetric tensors. Hence

$$C_2(\mathrm{Sym}^2 V) = \frac{N^2-1}{N} + 1 - \frac{1}{N} = \frac{(N-1)(N+2)}{N},$$
$$C_2(\Lambda^2 V) = \frac{N^2-1}{N} - 1 - \frac{1}{N} = \frac{(N+1)(N-2)}{N}.$$

For $N = 2$: $C_2(V) = \frac{3}{4}$, $C_2(\mathrm{Sym}^2) = 2 = 1 \cdot 2$ and $C_2(\Lambda^2) = 0$, the spin values $j(j+1)$ for $j = \frac{1}{2}, 1, 0$. Dividing by $k + N$ gives the conformal weights h_V, h_S, h_A.

Step (a): the four-punctured sphere is still 2-dimensional

For SU(2) all four punctures of the sphere around a crossing carried the same representation, because the 2-dimensional representation is equivalent to its dual. For SU(N) with $N \geq 3$ it is not: the fundamental V and the antifundamental $\bar{V}$ are different. So we must look at orientations. Two strands of a crossing both enter the ball through the lower pair of points and leave through the upper pair (Figure 12.72). A strand leaving the ball is a puncture labelled V; a strand entering it is a puncture labelled $\bar{V}$ (or the other way round; only consistency matters). The Hilbert space is therefore $\mathcal{H}_{S^2;V,V,\bar{V},\bar{V}}$. At large k its dimension is the number of invariants:

$$\dim \mathrm{Inv}\big(V \otimes V \otimes \bar{V} \otimes \bar{V}\big) \overset{(1)}{=} \dim \mathrm{Hom}_G\big(V \otimes V,\, V \otimes V\big)$$
$$\overset{(2)}{=} \dim \mathrm{Hom}_G\big(S \oplus A,\, S \oplus A\big) \overset{(3)}{=} 1 + 1 = 2.$$

(1) $\bar{V} \otimes \bar{V} \cong (V \otimes V)^*$, and an invariant in $X \otimes Y^*$ is the same thing as a G-equivariant linear map $Y \to X$.

(2) $V \otimes V = S \oplus A$ with $S = \mathrm{Sym}^2 V$, $A = \Lambda^2 V$.

(3) Schur's lemma: an equivariant map sends $S \to S$ and $A \to A$ by multiples of the identity and cannot mix the inequivalent S and A. So there are exactly two independent maps, for example $\mathbb{1}$ and P, or the projectors $\frac{1}{2}(\mathbb{1} \pm P)$.

At finite level we must also keep only integrable intermediate states: $\Lambda^2 V$ (one column of two boxes) is integrable for $k \geq 1$, but $\mathrm{Sym}^2 V$ (one row of two boxes) only for $k \geq 2$. So for $k \geq 2$ the space is 2-dimensional, exactly as for SU(2), and the "three vectors in a 2-dimensional space" argument of (12.283) goes through unchanged. (At $k = 1$ it is 1-dimensional; see Example 12.138.)

Step (b): the braiding eigenvalues

The half-twist exchanging the two outgoing punctures acts diagonally in the basis labelled by the fusion channel (S or A). Its eigenvalue is the phase of the conformal block, times the

exchange sign $\varepsilon_S = +1$, $\varepsilon_A = -1$, as in (12.287):

$$\lambda_S \overset{(1)}{=} \varepsilon_S\, e^{i\pi(h_S - 2h_V)} \overset{(2)}{=} \exp\!\Big(\frac{i\pi}{k+N}\Big[\frac{(N-1)(N+2)}{N} - \frac{N^2-1}{N}\Big]\Big) \overset{(3)}{=} q^{\frac{N-1}{2N}},$$
$$\lambda_A \overset{(4)}{=} \varepsilon_A\, e^{i\pi(h_A - 2h_V)} \overset{(5)}{=} -\exp\!\Big(\frac{i\pi}{k+N}\Big[\frac{(N+1)(N-2)}{N} - \frac{N^2-1}{N}\Big]\Big) \overset{(6)}{=} -q^{-\frac{N+1}{2N}}. \tag{12.334}$$

(1) Half a turn of one field around the other in the symmetric channel (same rule as for SU(2)).
(2) The Casimirs of Example 12.135 divided by $k+N$.
(3) The bracket is $\frac{(N-1)(N+2-N-1)}{N} = \frac{N-1}{N}$, and $e^{i\pi x/(k+N)} = q^{x/2}$.
(4) The same rule in the antisymmetric channel, with the exchange sign -1.
(5) The Casimirs again.
(6) The bracket is $\frac{(N+1)(N-2-N+1)}{N} = -\frac{N+1}{N}$.

For $N = 2$ these are $\lambda_S = q^{1/4} = \lambda_1$ and $\lambda_A = -q^{-3/4} = \lambda_0$, the SU(2) values of (12.287): the triplet is the symmetric channel and the singlet the antisymmetric one. The framing phase of a single fundamental line is

$$\theta_V := e^{2\pi i h_V} = \exp\!\Big(2\pi i\,\frac{N^2-1}{2N(k+N)}\Big) = q^{\frac{N^2-1}{2N}}, \tag{12.335}$$

which is $q^{3/4}$ for $N = 2$, as in Section 12.14.16.

Step (c): the Hecke relation

On a 2-dimensional space with eigenvalues $\lambda_S \neq \lambda_A$ the matrix σ obeys its characteristic equation $(\sigma - \lambda_S)(\sigma - \lambda_A) = 0$. This is the **Hecke relation**. Rewrite it with σ^{-1}:

$$\begin{aligned} 0 &\overset{(1)}{=} \sigma - (\lambda_S + \lambda_A)\mathbb{1} + \lambda_S\lambda_A\,\sigma^{-1} \\ &\overset{(2)}{=} \sigma - q^{-\frac{1}{2N}}\big(q^{\frac12} - q^{-\frac12}\big)\mathbb{1} - q^{-\frac{1}{N}}\,\sigma^{-1}, \\ \text{so}\quad q^{\frac{1}{2N}}\,\sigma - q^{-\frac{1}{2N}}\,\sigma^{-1} &\overset{(3)}{=} \big(q^{\frac12} - q^{-\frac12}\big)\mathbb{1}. \end{aligned} \tag{12.336}$$

(1) Expand $(\sigma - \lambda_S)(\sigma - \lambda_A) = 0$ and multiply by σ^{-1}.
(2) $\lambda_S + \lambda_A = q^{\frac{N-1}{2N}} - q^{-\frac{N+1}{2N}} = q^{-\frac{1}{2N}}(q^{\frac12} - q^{-\frac12})$, and $\lambda_S\lambda_A = -q^{\frac{(N-1)-(N+1)}{2N}} = -q^{-\frac{1}{N}}$.
(3) Move the $\mathbb{1}$ term to the right and multiply by $q^{\frac{1}{2N}}$.

Apply this operator identity to the vector V_0 of the smoothed crossing, as in Section 12.14.13: $\sigma V_0 = V_+$ and $\sigma^{-1}V_0 = V_-$. Pairing with the vector V_1 of the outside of the ball gives, for blackboard-framed correlators,

$$q^{\frac{1}{2N}}\,W_{\mathrm{bb}}(L_+) - q^{-\frac{1}{2N}}\,W_{\mathrm{bb}}(L_-) = \big(q^{\frac12} - q^{-\frac12}\big)\,W_{\mathrm{bb}}(L_0).$$

Step (d): framing, and the result

The framing-independent invariant is $X(L) := \theta_V^{-w(L)} W_{\mathrm{bb}}(L)$, where w is the writhe of the diagram (Section 12.14.16). The three diagrams have writhes $w(L_\pm) = w(L_0) \pm 1$. Substitute $W_{\mathrm{bb}}(L) = \theta_V^{w(L)} X(L)$ and divide by $\theta_V^{w(L_0)}$:

$$\big(q^{\frac{1}{2N}}\theta_V\big)X(L_+) - \big(q^{\frac{1}{2N}}\theta_V\big)^{-1}X(L_-) \overset{(1)}{=} \big(q^{\frac12} - q^{-\frac12}\big)X(L_0),$$

$$\boxed{a\,X(L_+) - a^{-1}X(L_-) = z\,X(L_0), \qquad a = q^{N/2}, \quad z = q^{1/2} - q^{-1/2}.} \tag{12.337}$$

(1) The writhe bookkeeping just described.
The exponent is $\frac{1}{2N} + \frac{N^2-1}{2N} = \frac{N}{2}$, so $q^{\frac{1}{2N}}\theta_V = q^{N/2}$.

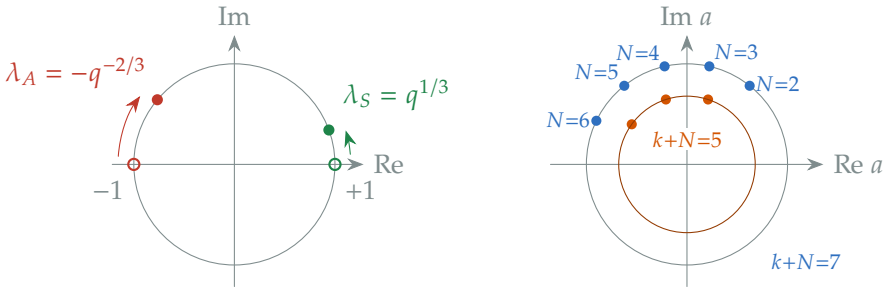
This is the **HOMFLY skein relation** (named after Hoste, Ocneanu, Millett, Freyd, Lickorish and Yetter; also called HOMFLY-PT, after Przytycki and Traczyk). Its two variables are

$$a = \exp\Big(\frac{i\pi N}{k+N}\Big), \qquad z = 2i\sin\Big(\frac{\pi}{k+N}\Big). \tag{12.338}$$

These are Nash's "variables determined by N and k".

Key idea: The HOMFLY polynomial from $\mathrm{SU}(N)$ Chern–Simons theory

Wilson lines in the defining representation of SU(N) at level k obey the skein relation (12.337). With the normalisation $P(\text{unknot}) = 1$, the ratio $P_L = X(L)/X(\bigcirc)$ is the two-variable HOMFLY polynomial $P_L(a,z)$ evaluated at (12.338). For $N = 2$ it is the Jones polynomial. The mechanism is the same as for SU(2): a 2-dimensional space of conformal blocks, and a quadratic relation for the braiding.



(a) SU(3), $k = 3$: $q = e^{2\pi i/6}$ (b) points $a = e^{i\pi N/(k+N)}$

Figure 12.91: (a) The two braiding eigenvalues of SU(3) at level 3 on the unit circle. Open circles are the classical values ±1 (the swap P); the arrows show how the quantum corrections $q^{(N-1)/2N}$ and $-q^{-(N+1)/2N}$ move them apart. (b) For fixed $k+N = n$ the variable $z = 2i\sin(\pi/n)$ is fixed, and the admissible $N = 2, \dots, n-1$ give $n-2$ points a on the upper unit semicircle (drawn on circles of different radii only to separate the two values of n). As n grows these sets become dense; this is why the SU(N) invariants for all N and k determine a genuinely two-variable polynomial.

Worked example 12.136: The unknot and the quantum dimension $[N]_q$

Take the one-crossing diagram of the unknot with a positive kink. Then L_+ is the kinked unknot, L_- the unknot with a negative kink, and L_0 is two disjoint circles. X ignores framing, so $X(L_\pm) = X(\bigcirc)$, and X of two unlinked circles is $X(\bigcirc)^2$ (connected sums, Section 12.14.9). The skein relation gives

$$\big(a - a^{-1}\big)X(\bigcirc) \overset{(1)}{=} z\,X(\bigcirc)^2 \quad\Longrightarrow\quad X(\bigcirc) \overset{(2)}{=} \frac{a-a^{-1}}{z} \overset{(3)}{=} \frac{q^{N/2}-q^{-N/2}}{q^{1/2}-q^{-1/2}} = [N]_q.$$

(1) (12.337) for this triple.
(2) Divide by $zX(\bigcirc) \neq 0$.
(3) Insert $a = q^{N/2}$, $z = q^{1/2} - q^{-1/2}$.

As $k \to \infty$, $[N]_q \to N = \dim V$: the classical Wilson loop of the trivial connection, $\mathrm{tr}_V \mathbb{1} = N$. The surgery side (Section 12.15.6) gives the same thing as S_{0V}/S_{00}, the quantum dimension of V. For $N = 2$ it is $[2] = W(\bigcirc)$ of (12.323).
Two more checks, which test step (b) against step (d). A single strand of V with a positive

kink closes to $\widehat{\sigma}$, and the quantum trace (12.300) with the SU(N) data is

$$W_{\mathrm{bb}}(\widehat{\sigma}) = \lambda_S\, d_S + \lambda_A\, d_A, \qquad d_S = \frac{[N][N+1]}{[2]}, \quad d_A = \frac{[N][N-1]}{[2]},$$

where d_S, d_A are the quantum dimensions of $\mathrm{Sym}^2 V$ and $\Lambda^2 V$ (they reduce to $N(N\pm1)/2$ at $q=1$). With n crossings, $W_{\mathrm{bb}}(\widehat{\sigma^n}) = \lambda_S^n d_S + \lambda_A^n d_A$. For $n=0$, $d_S + d_A = [N]([N+1]+[N-1])/[2] = [N]^2$, using $[N+1]+[N-1] = [2][N]$: two unlinked circles. For $n=1$ one finds $\theta_V[N]$, the kinked unknot. (Both are easy to check numerically for any N and k; the $N=2$ case is Example 12.112.)

Worked example 12.137: The Hopf link and the trefoil, for all N at once

Normalise $P_L = X(L)/X(\bigcirc)$, so $P(\bigcirc)=1$ and the unlink of two has $P = [N]_q = (a-a^{-1})/z$.
Hopf link. Take the positive Hopf link H_+ and change one crossing (as in Example 12.108): $L_+ = H_+$, $L_- =$ unlink, $L_0 = \bigcirc$.

$$a\,P(H_+) \overset{(1)}{=} a^{-1}\,\frac{a-a^{-1}}{z} + z \quad\Longrightarrow\quad P(H_+) \overset{(2)}{=} \frac{a^{-1}-a^{-3}}{z} + a^{-1}z.$$

(1) (12.337) solved for the L_+ term, with the unlink and unknot values.
(2) Divide by a.

Trefoil. The right-handed trefoil is the closure of σ^3. Change one crossing: $L_- = \widehat{\sigma}$ is an unknot, $L_0 = \widehat{\sigma^2} = H_+$.

$$P(T) \overset{(1)}{=} a^{-2}\cdot 1 + a^{-1}z\,P(H_+) \overset{(2)}{=} a^{-2} + a^{-1}z\Big(\frac{a^{-1}-a^{-3}}{z} + a^{-1}z\Big) \overset{(3)}{=} 2a^{-2} - a^{-4} + z^2a^{-2}.$$

(1) Skein relation divided by a.
(2) The Hopf result.
(3) Multiply out: $a^{-2} + a^{-2} - a^{-4} + a^{-2}z^2$.

Specialisations. Put $N=2$: $a = q = t^{-1}$ and, with the square root $t^{1/2} = -q^{-1/2}$ of (12.294), $z = q^{1/2} - q^{-1/2} = t^{1/2} - t^{-1/2}$ so that $z^2 = t - 2 + t^{-1}$. Then

$$P(T)\big|_{N=2} \overset{(1)}{=} 2t^2 - t^4 + t^2\big(t - 2 + t^{-1}\big) = t + t^3 - t^4,$$

(1) Substitute $a^{-2} = t^2$, $a^{-4} = t^4$ and z^2.

which is Nash's (12.76), the Jones polynomial of the trefoil (and agrees with Example 12.109). Put instead $a=1$: the skein relation becomes Conway's $X_+ - X_- = zX_0$, and with $z = t^{1/2} - t^{-1/2}$, $P(T)|_{a=1} = 1 + z^2 = t - 1 + t^{-1}$, the Alexander polynomial of the trefoil (Nash's (12.74), up to the factor t). Formally $a = q^{N/2} = 1$ is "$N=0$"; the unknot value $[0]_q = 0$ then vanishes, which is why only the normalised P survives this limit. (The supergroup U(1|1), whose defining representation has superdimension $1-1=0$, realises it.)

Worked example 12.138: Level 1: $\mathrm{SU}(N)_1$ is secretly abelian

At $k=1$ the symmetric channel is not integrable. The quantum dimension knows this: $q = e^{2\pi i/(N+1)}$ and $[N+1]_q = \sin\pi/\sin\frac{\pi}{N+1} = 0$, so $d_S = 0$. The 4-punctured sphere is 1-dimensional and $\sigma = \lambda_A$ is a number. Every fundamental line has $[N]_q = \sin\frac{N\pi}{N+1}/\sin\frac{\pi}{N+1} =$

1: all quantum dimensions are 1, the hallmark of an abelian theory. For a 2-strand closure,

$$X\left(\widehat{\sigma^n}\right) \overset{(1)}{=} \theta_V^{-n}\lambda_A^n\, d_A \overset{(2)}{=} \left(-q^{-\frac{N^2-1}{2N}-\frac{N+1}{2N}}\right)^n \overset{(3)}{=} \left(-q^{-\frac{N+1}{2}}\right)^n \overset{(4)}{=} 1.$$

(1) The quantum trace with only the antisymmetric channel.
(2) At this q, $[N] = 1$ and $[N-1] = [2]$ (because $\sin\frac{(N-1)\pi}{N+1} = \sin\frac{2\pi}{N+1}$), so $d_A = [N][N-1]/[2] = 1$; then insert (12.335) and (12.334).
(3) $\frac{N^2-1+N+1}{2N} = \frac{N+1}{2}$.
(4) $q^{(N+1)/2} = e^{i\pi} = -1$.

So in the writhe-normalised form the Hopf link and the trefoil are both trivial. The information sits in the framing: the zero-framed Hopf link is $\theta_V^2 X(H_+) = e^{-2\pi i/N}$ (for $N = 3$: $e^{-2\pi i/3}$). This is a pure phase, the exponential of a linking number, exactly the kind of answer the abelian theory of the previous section gave through Gauss's formula. The lines of $\mathrm{SU}(N)_1$ are abelian anyons with statistics $e^{-2\pi i/N}$.

Worked example 12.139: Why all N and k together determine a two-variable polynomial

One $\mathrm{SU}(N)$ theory at one level gives one number $P_L(a,z)$. Why does Nash say the answer is a polynomial in two variables? Write the HOMFLY polynomial as $P_L(a,z) = \sum_{m=-M}^{M} a^m\, p_m(z)$ with Laurent polynomials p_m. Fix $n = k+N$. Then $z_n = 2i\sin(\pi/n)$ is fixed, while $N = 2,\dots,n-1$ gives the $n-2$ distinct values $a = e^{i\pi N/n}$ (Figure 12.91(b)). The function $a \mapsto a^M P_L(a, z_n)$ is a polynomial of degree at most $2M$, so once $n-2 > 2M$ its values at those points fix all its coefficients (Vandermonde). Thus every $p_m(z_n)$ is known, for infinitely many distinct z_n; a Laurent polynomial is fixed by infinitely many values. Conversely, for a single N the values only see the one-variable specialisation $a = q^{N/2}$, $z = q^{1/2} - q^{-1/2}$. So a single group gives a one-variable invariant, and the family $\mathrm{SU}(N)$, $N = 2,3,\dots$, gives the two-variable one.

Watch out: Many conventions for HOMFLY

The HOMFLY polynomial appears in the literature as $P(l,m)$ with $lP_+ + l^{-1}P_- + mP_0 = 0$, as $P(v,z)$ with $v^{-1}P_+ - vP_- = zP_0$, and as $P(x,y,z)$ with $xP_+ + yP_- + zP_0 = 0$ (the original homogeneous form of Freyd et al., which Nash cites). They differ by $a \leftrightarrow a^{-1}$, by signs and by i's, and by the exchange $L_+ \leftrightarrow L_-$, which is reflection of the link. Always test a convention on the trefoil: with ours, the right-handed trefoil has only non-positive powers of a.

Other groups and other representations

The number of terms in the skein relation equals the dimension of the space of conformal blocks on the 4-punctured sphere, which at large k is the number of irreducible pieces of $R \otimes R$ (when each occurs once). A d-dimensional space makes σ obey a polynomial equation of degree d, so a relation among $\sigma^{-1}, \mathbb{1}, \sigma, \dots$ with $d+1$ terms.

group and representation R	$R \otimes R$	channels	polynomial
U(1), charge n	charge $2n$	1	$e^{i\cdot\text{linking}}$ (previous section)
SU(2), spin $\frac{1}{2}$	$0 \oplus 1$	2	Jones $V_L(t)$
SU(N), $V = \mathbb{C}^N$	$\mathrm{Sym}^2 V \oplus \Lambda^2 V$	2	HOMFLY $P_L(a,z)$
SO(N), $V = \mathbb{R}^N$	$\mathbf{1} \oplus \Lambda^2 V \oplus \mathrm{Sym}^2_0 V$	3	Kauffman $F_L(a,z)$
Sp($2n$), $V = \mathbb{C}^{2n}$	$\mathbf{1} \oplus \Lambda^2_0 V \oplus \mathrm{Sym}^2 V$	3	Kauffman (other branch)
SU(2), spin 1	$0 \oplus 1 \oplus 2$	3	coloured Jones / Kauffman
G_2, $\mathbf{7}$	$\mathbf{1} \oplus \mathbf{7} \oplus \mathbf{14} \oplus \mathbf{27}$	4	G_2 invariant (Kuperberg)

A new diagram appears when R is equivalent to its dual (real or pseudo-real). Then a strand may turn back, and the invariant $\mathbf{1} \subset R \otimes R$ is a "cup": two outgoing punctures joined to each other. The fourth diagram L_∞ of Figure 12.92 is allowed, and the Kauffman relation reads $L_+ - L_- = z(L_0 - L_\infty)$ (for framed links). For SU(N) with $N \geq 3$ the fundamental is complex, orientation forbids joining two outgoing punctures, and L_∞ is absent: two channels, three terms. For SU(2) the fundamental is pseudo-real, and indeed the Jones polynomial has an unoriented version, the Kauffman bracket of the previous section, in which L_∞ appears as the second smoothing. There, however, the space is still 2-dimensional, and L_∞ can be eliminated in favour of L_0 and the crossings.

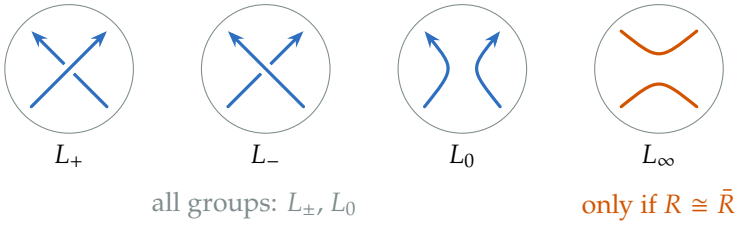


Figure 12.92: The diagrams that can appear in a skein relation. The first three respect the orientation of the strands, and are all that SU(N) with its complex defining representation allows. The fourth joins the two lower punctures to each other and the two upper ones to each other. It needs an invariant pairing $R \otimes R \to \mathbb{C}$, so it exists for SO($N$) and Sp($2n$) (and for SU(2)), and it gives the Kauffman polynomial.

Physical picture: Why the answer is a polynomial

Why are all these invariants *polynomials* (in $q^{\pm 1/2}$ or in a, z) and not arbitrary functions of k? Because the skein relation reduces any link, crossing by crossing, to unlinks, and each step multiplies by a Laurent monomial in the eigenvalues. The eigenvalues are powers of q because the conformal weights are rational multiples of $1/(k + h^\vee)$. The only non-polynomial ingredient, the unknot value $[N]_q$, is removed by the normalisation $P(\bigcirc) = 1$. Physics supplies the relation; the polynomial structure is combinatorics.

12.15.9 Other topological field theories

Nash closes the chapter by observing that Chern–Simons theory is far from the only topological field theory. He mentions two more: a topological sigma model in two dimensions (Witten, Floer, Gromov), whose ground states are "like Floer groups", and a Chern–Simons theory of gravity in $2+1$ dimensions with gauge group ISO$(2,1)$ (Witten), whose Hilbert spaces are infinite-dimensional and which has topology-changing amplitudes. Finally he points to the axiomatic approach of Atiyah. We take them in turn, each time with the same questions: *what are the fields, why is the theory topological, and what does it compute?*

The two-dimensional topological sigma model

The fields and the energy bound. A **sigma model** is a field theory whose field is a map $\phi : \Sigma \to X$ from a surface (the worldsheet) to a target manifold X, with action the Dirichlet energy $E(\phi) = \frac{1}{2}\int_\Sigma |\mathrm{d}\phi|^2 \, \mathrm{d}A$. (For $X = \mathbb{R}^n$ these are n free massless scalars; for a curved X it is the string worldsheet theory of Chapter 9.) Take X to be a **Kähler manifold**: it carries a metric g, a complex structure J (a linear map on each tangent space with $J^2 = -\mathbb{1}$, "multiplication by i"), and the 2-form $\omega(u, v) = g(Ju, v)$, which is closed.

The key identity is a Bogomolny-type completion of the square, exactly as for instantons (Chapter 8). Work first with $X = \mathbb{C}$, $\phi = u + iv$, and local coordinates (x, y) on Σ:

$$\begin{aligned} \tfrac{1}{2}|\mathrm{d}\phi|^2 &\overset{(1)}{=} \tfrac{1}{2}(u_x^2 + u_y^2 + v_x^2 + v_y^2) \\ &\overset{(2)}{=} \tfrac{1}{2}\big[(u_x - v_y)^2 + (u_y + v_x)^2\big] + (u_x v_y - u_y v_x) \\ &\overset{(3)}{=} 2\,|\partial_{\bar w}\phi|^2 + \frac{\phi^*(\mathrm{d}u \wedge \mathrm{d}v)}{\mathrm{d}x \wedge \mathrm{d}y}. \end{aligned} \tag{12.339}$$

(1) The Euclidean norm of the four partial derivatives.

(2) Add and subtract $u_x v_y - u_y v_x$: the squares produce $-2u_x v_y$ and $+2u_y v_x$, which the last bracket cancels.

(3) With $w = x + iy$, $\partial_{\bar w} = \frac{1}{2}(\partial_x + i\partial_y)$ gives $2\partial_{\bar w}\phi = (u_x - v_y) + i(u_y + v_x)$; and the Jacobian $u_x v_y - u_y v_x$ is the pull-back of the area form $\omega = \mathrm{d}u \wedge \mathrm{d}v$ divided by $\mathrm{d}x \wedge \mathrm{d}y$.

The same pointwise identity holds for any Kähler X (it is linear algebra in one tangent space):

$$E(\phi) = \int_\Sigma |\bar\partial_J \phi|^2 \, \mathrm{d}A \; + \; \int_\Sigma \phi^*\omega \;\geq\; \int_\Sigma \phi^*\omega. \tag{12.340}$$

Since $\mathrm{d}\omega = 0$, the last integral depends only on the homology class $\phi_*[\Sigma] \in H_2(X)$ (Stokes, Chapter T2: changing ϕ by a homotopy changes $\int \phi^*\omega$ by $\int \mathrm{d}(\dots) = 0$). It is a topological charge. The bound is saturated exactly by the **holomorphic maps**, $\bar\partial_J \phi = 0$: these are the **worldsheet instantons**.

Dictionary: Instantons in four dimensions and in two

Yang–Mills on M^4	sigma model on Σ^2
$\frac{1}{2}\int \lvert F\rvert^2 = \frac{1}{2}\int \lvert F \pm \star F\rvert^2 \mp \int \operatorname{tr} F \wedge F$	$E = \int \lvert\bar\partial\phi\rvert^2 + \int \phi^*\omega$
self-duality $F = \pm \star F$	Cauchy–Riemann $\bar\partial_J \phi = 0$
instanton number k	degree $d = \phi_*[\Sigma]$
weight $e^{-8\pi^2 k/g^2}$	weight e^{-td}, $t = \int_{\text{line}} \omega$
moduli space $\mathcal{M}_k$	moduli space of holomorphic curves
Donaldson invariants (§3)	Gromov–Witten invariants

Why it is topological. Witten's construction adds fermions and a BRST-type charge Q with $Q^2 = 0$ (up to symmetries), exactly as in the Donaldson theory of §3. The action becomes $S = \{Q, V\} + \int \phi^*\omega$ for some V. The first term has metric-independent correlators by the argument of §3 (its variation with respect to the worldsheet metric is Q-exact); the second is topological by (12.340). The coupling in front of $\{Q, V\}$ can be sent to infinity, and the path integral localises on the zeros of $\bar\partial_J \phi$: the holomorphic maps. Gromov's contribution was the analysis of these maps (the **pseudo-holomorphic curves**, allowed also for non-integrable J) and in particular their compactness: sequences can only degenerate by "bubbling" off spheres, the 2-dimensional analogue of Uhlenbeck's bubbling of instantons.

The observables are labelled by cohomology classes $\alpha \in H^*(X)$: pick a point $z \in \Sigma$ and a cycle Poincaré dual to α, and ask that $\phi(z)$ lie on that cycle. A correlation function counts the holomorphic maps satisfying all such conditions, each weighted by e^{-td}.

How many conditions? The dimension of the moduli space. The tangent space to the moduli space at a holomorphic ϕ is the kernel of the linearised operator, a $\bar\partial$ operator on sections of ϕ^*TX. Its index is given by Riemann–Roch (Chapter 4): for Σ of genus g and $\dim_{\mathbb{C}} X = n$,

$$\dim_{\mathbb{C}} \mathcal{M}_d = n(1-g) + \int_\Sigma \phi^* c_1(TX) \qquad \text{(when there is no cokernel).} \tag{12.341}$$

This is the "selection rule" (the ghost-number anomaly): a correlator can be non-zero only if the total complex codimension of the conditions equals $\dim_{\mathbb{C}} \mathcal{M}_d$.

Worked example 12.140: The three-point function of $\mathbb{CP}^1$ and quantum cohomology

Let $\Sigma = X = \mathbb{CP}^1$, the Riemann sphere, with $c_1(T\mathbb{CP}^1)$ equal to 2 on the fundamental class (Euler characteristic 2). A holomorphic map of degree d is a rational function $\phi(z) = f(z)/g(z)$ with $\deg f, \deg g \le d$ and no common root. Let $x \in H^2(\mathbb{CP}^1)$ be the class of a point; the condition "$\phi(z_i) = p_i$" has complex codimension 1.

$$\dim_{\mathbb{C}} \mathcal{M}_d \overset{(1)}{=} 1\cdot(1-0) + 2d = 2d+1.$$

(1) (12.341) with $n=1$, $g=0$ and $\int \phi^* c_1 = d\cdot 2$. Directly: f and g have $2(d+1)$ coefficients, minus one for the overall scale, giving $2d+1$.

Three point conditions need $2d+1=3$, so $d=1$. Degree-1 maps are the Möbius transformations $\phi(z) = (\alpha z+\beta)/(\gamma z+\delta)$, which form $\mathrm{PSL}(2,\mathbb{C})$, of complex dimension 3. Now require $\phi(0)=p_1$, $\phi(1)=p_2$, $\phi(\infty)=p_3$ for three distinct points. There is exactly one such Möbius map (it is fixed by the cross-ratio). Hence

$$\langle \mathcal{O}_x(0)\,\mathcal{O}_x(1)\,\mathcal{O}_x(\infty)\rangle = 1\cdot e^{-t},$$

a one-instanton effect: at $d=0$ (constant maps) three distinct points can never be hit, so classically this correlator vanishes. The correlators with the unit $1 \in H^0$ are $\langle \mathcal{O}_1\mathcal{O}_1\mathcal{O}_x\rangle = 1$ (at $d=0$; total codimension $1 = \dim \mathcal{M}_0$) and $\langle \mathcal{O}_1\mathcal{O}_x\mathcal{O}_x\rangle = 0$ (it would need $2d+1=2$). Define a product by $\langle x*y,\, z\rangle = \langle \mathcal{O}_x\mathcal{O}_y\mathcal{O}_z\rangle$, using the classical pairing $\langle 1,x\rangle = 1$, $\langle 1,1\rangle = \langle x,x\rangle = 0$. Then $\langle x*x, x\rangle = e^{-t}$ and $\langle x*x, 1\rangle = 0$ give

$$x * x = e^{-t}\,\mathbf{1},$$

the **quantum cohomology** ring of $\mathbb{CP}^1$. Classically $x\cup x = 0$ (two points on a sphere do not meet); the instanton deforms the ring, and the classical ring returns as $t\to\infty$ (a very large sphere, where instantons are infinitely suppressed).

Floer-like ground states. Now take the worldsheet to be a cylinder $S^1\times\mathbb{R}$, with $t\in S^1$ the space direction and $s\in\mathbb{R}$ the Euclidean time. A state is a functional on the space LX of loops $\gamma : S^1 \to X$. The structure is that of §2 of this chapter, with $\mathcal{A}/\mathcal{G}$ replaced by LX and the Chern–Simons function replaced by the **symplectic action** $\mathcal{A}(\gamma) = \int_D \bar\gamma^*\omega$, the ω-area of a disc $\bar\gamma$ bounded by γ (defined up to the periods of ω, just as the Chern–Simons function is defined up to integers). Let us compute its gradient in the simplest target, $X = \mathbb{C} = \mathbb{R}^2$ with $\omega = \mathrm{d}x\wedge\mathrm{d}y$ and $J(\xi_x,\xi_y) = (-\xi_y,\xi_x)$.

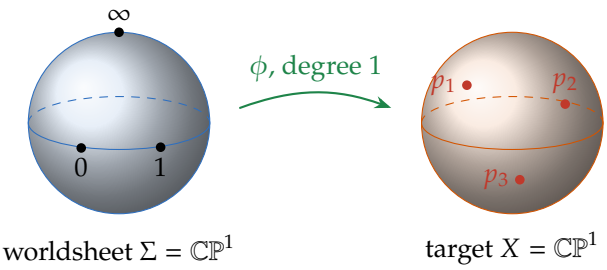


Figure 12.93: The three-point function of the $\mathbb{CP}^1$ sigma model. Constant maps cannot send three points of the worldsheet to three different points of the target. Degree-1 holomorphic maps (Möbius transformations) form a 3-dimensional family, and exactly one of them sends $0, 1, \infty$ to p_1, p_2, p_3. So the correlator is a pure one-instanton effect, e^{-t}.

$$\begin{aligned}
\mathcal{A}(\gamma) &\overset{(1)}{=} \frac{1}{2}\oint (x\,\mathrm{d}y - y\,\mathrm{d}x),\\
\delta\mathcal{A} &\overset{(2)}{=} \frac{1}{2}\oint \big(\xi_x\dot{y} + x\dot{\xi}_y - \xi_y\dot{x} - y\dot{\xi}_x\big)\mathrm{d}t \overset{(3)}{=} \oint \big(\xi_x\dot{y} - \xi_y\dot{x}\big)\mathrm{d}t = \oint \omega(\xi,\dot{\gamma})\,\mathrm{d}t,\\
\operatorname{grad}\mathcal{A} &\overset{(4)}{=} -J\dot{\gamma}.
\end{aligned} \tag{12.342}$$

(1) Green's theorem (Stokes in the plane): the area enclosed is $\frac{1}{2}\oint(x\,\mathrm{d}y - y\,\mathrm{d}x)$.
(2) Vary $\gamma \to \gamma + \xi$ to first order.
(3) Integrate by parts round the closed loop: $\oint x\dot{\xi}_y = -\oint \dot{x}\xi_y$ and $-\oint y\dot{\xi}_x = \oint \dot{y}\xi_x$.
(4) The gradient is defined by $\delta\mathcal{A} = \oint g(\xi, \operatorname{grad}\mathcal{A})$ with $g(a,b) = \omega(a,Jb) = a\cdot b$. We need $J\operatorname{grad} = \dot{\gamma}$, so $\operatorname{grad} = -J\dot{\gamma}$.

The gradient flow $\partial_s u = \operatorname{grad}\mathcal{A}$ for a family of loops $u(s,t)$ is therefore

$$\partial_s u + J\,\partial_t u = 0, \tag{12.343}$$

which for $u = x + iy$ and $w = s + it$ is $2\partial_{\bar{w}} u = 0$: the flow lines of the symplectic action are *holomorphic cylinders*. This is Floer's equation, the exact analogue of "gradient flow of the Chern–Simons function = instanton on $M\times\mathbb{R}$" in §2. The critical points are the loops with $\dot{\gamma} = 0$: constant loops (or, after adding a Hamiltonian term, the periodic orbits of a Hamiltonian flow).

Worked example 12.141: Infinite index in both directions, again

Expand a loop in $\mathbb{C}$ in Fourier modes, $\gamma(t) = \sum_n c_n e^{int}$. Then

$$\mathcal{A}(\gamma) \overset{(1)}{=} \frac{1}{2}\operatorname{Im}\oint \bar{\gamma}\,\mathrm{d}\gamma \overset{(2)}{=} \frac{1}{2}\operatorname{Im}\sum_{m,n}\bar{c}_m c_n\, in\int_0^{2\pi} e^{i(n-m)t}\mathrm{d}t \overset{(3)}{=} \pi\sum_n n\,|c_n|^2.$$

(1) With $\gamma = x + iy$: $\bar{\gamma}\,\mathrm{d}\gamma = (x\,\mathrm{d}x + y\,\mathrm{d}y) + i(x\,\mathrm{d}y - y\,\mathrm{d}x)$, and the real part integrates to zero round a loop.
(2) Insert the Fourier series, $\mathrm{d}\gamma = \sum_n in\,c_n e^{int}\mathrm{d}t$.
(3) Orthogonality: the integral is $2\pi\delta_{mn}$; and $\operatorname{Im}(2\pi in\,|c_n|^2) = 2\pi n\,|c_n|^2$.

So $\mathcal{A}$ is a quadratic form with infinitely many positive directions ($n > 0$, loops winding anticlockwise) and infinitely many negative ones ($n < 0$). Its Morse index is infinite, exactly like the Hessian $\star\mathrm{d}_A$ of the Chern–Simons function in §2, whose eigenvalues on S^3 or T^3 also run to $\pm\infty$. The cure is the same: only *relative* indices, computed by spectral flow,

are used, and the Floer groups are built from critical points and holomorphic cylinders between them.

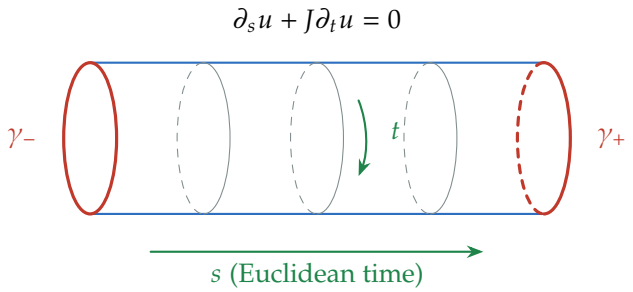


Figure 12.94: A Floer trajectory: a holomorphic cylinder $u(s,t)$ that flows from one critical loop γ_- at $s = -\infty$ to another γ_+ at $s = +\infty$. The grey circles are the intermediate loops $u(s,\cdot)$. This is the 2-dimensional counterpart of an instanton on $M \times \mathbb{R}$ joining two flat connections; counting such cylinders defines the boundary operator of the Floer complex, and the supersymmetric ground states of the sigma model on the circle are its homology.

Dictionary: Two Floer theories side by side

	instanton Floer theory (§2)	symplectic Floer theory
configuration space	$\mathcal{A}/\mathcal{G}$ over a 3-manifold M	loop space LX
Morse function	Chern–Simons $f(A)$, mod $\mathbb{Z}$	action $\mathcal{A}(\gamma)$, mod periods
critical points	flat connections	constant loops / periodic orbits
flow lines	instantons on $M \times \mathbb{R}$	holomorphic cylinders
Hessian	$\star \mathrm{d}_A$, index $\pm\infty$	$-J\,\mathrm{d}/\mathrm{d}t$, index $\pm\infty$
field theory	3d $\to$ 4d Donaldson TQFT	1d $\to$ 2d sigma model
ground states	$HF_*(M)$	$HF_*(X) \cong H_*(X)$

For a compact symplectic X (under technical conditions that exclude holomorphic spheres of negative index) Floer proved that the symplectic Floer groups are isomorphic to the ordinary homology of X. Since the chain complex is generated by the fixed points of a Hamiltonian symplectomorphism (the periodic orbits), this gives the **Arnold conjecture**: such a map has at least $\sum_i b_i(X)$ fixed points (when they are non-degenerate). This was Floer's original motivation, and it is the sense of Nash's phrase "quantum ground states that are like Floer groups".

Chern–Simons gravity in $2+1$ dimensions

General relativity in three spacetime dimensions is a remarkable test case: the vacuum Einstein equations say that spacetime is *flat*. (In three dimensions the Riemann tensor is determined by the Ricci tensor, and $R_{\mu\nu} = 0$ forces $R_{\mu\nu\rho\sigma} = 0$. The number of graviton polarisations in d dimensions is $d(d-3)/2$, which is 0 for $d = 3$.) So the theory has no local degrees of freedom; everything is global, just as in Chern–Simons theory. Witten showed that it *is* a Chern–Simons theory.

The gauge group. Use the first-order (frame) formalism: a **dreibein** $e^a = e^a_\mu \mathrm{d}x^\mu$ ($a = 0,1,2$, with $g_{\mu\nu} = \eta_{ab} e^a_\mu e^b_\nu$ and $\eta = \mathrm{diag}(-1,1,1)$) and a spin connection, which in three dimensions we dualise to a vector of 1-forms $\omega^a = \frac{1}{2}\varepsilon^{abc}\omega_{bc}$. Indices are raised with η, and $\varepsilon_{012} = 1$. The Poincaré group $\mathrm{ISO}(2,1)$ has generators J_a (Lorentz) and P_a (translations). Exhibit them

concretely as 4×4 matrices acting on $(x^0,x^1,x^2,1)$, the affine representation:

$$J_a=\begin{pmatrix} j_a & 0\\ 0 & 0\end{pmatrix},\quad (j_a)^b{}_c=\varepsilon^b{}_{ac},\qquad P_a=\begin{pmatrix} 0 & e_a\\ 0 & 0\end{pmatrix},$$

where e_a is the a-th unit column vector. A translation matrix has its only entries in the last column and a zero last row, so the product of two of them vanishes: $[P_a,P_b]=0$ at once. Multiplying out the other products gives

$$[J_a,J_b]=\varepsilon_{abc}J^c,\qquad [J_a,P_b]=\varepsilon_{abc}P^c,\qquad [P_a,P_b]=0. \tag{12.344}$$

(For $[J_a,P_b]$: the commutator has last column $j_a e_b$, whose c-th entry is $\varepsilon^c{}_{ab}$, so it is $\varepsilon^c{}_{ab}P_c=\varepsilon_{abc}P^c$.)

The invariant form. Chern–Simons theory needs an invariant, non-degenerate bilinear form on the Lie algebra; for SU(2) it was $-\operatorname{tr}$. For $\mathfrak{iso}(2,1)$ the Killing form $K(X,Y)=\operatorname{tr}(\operatorname{ad}_X\operatorname{ad}_Y)$ is *degenerate*: a direct computation gives $K(J_a,J_b)=4\eta_{ab}$ and $K(P_a,\cdot)=0$, because ad_{P_a} maps everything into the abelian ideal spanned by the P's. There is, however, another invariant form,

$$\langle J_a,P_b\rangle=\eta_{ab},\qquad \langle J_a,J_b\rangle=\langle P_a,P_b\rangle=0, \tag{12.345}$$

which is non-degenerate. Invariance, $\langle[X,Y],Z\rangle+\langle Y,[X,Z]\rangle=0$, is checked on basis elements; for instance with $X=J_a$, $Y=J_b$, $Z=P_c$ it reads $\varepsilon_{abd}\eta^{de}\eta_{ec}+\varepsilon_{acd}\eta^{de}\eta_{be}=\varepsilon_{abc}+\varepsilon_{acb}=0$.

The action is Einstein–Hilbert. Form the $\mathfrak{iso}(2,1)$ connection $A=e^aP_a+\omega^aJ_a$. With $[A\wedge A]:=A^i\wedge A^j[X_i,X_j]$ (so that $A\wedge A=\frac12[A\wedge A]$ in a matrix representation), the curvature is

$$F=\mathrm{d}A+\tfrac12[A\wedge A]\overset{(1)}{=}\big(\mathrm{d}e^a+\varepsilon^a{}_{bc}\,\omega^b\wedge e^c\big)P_a+\big(\mathrm{d}\omega^a+\tfrac12\varepsilon^a{}_{bc}\,\omega^b\wedge\omega^c\big)J_a=:T^aP_a+R^aJ_a.$$

(1) Using (12.344): the $\omega\omega$ terms give $\omega^b\wedge\omega^c\varepsilon_{bca}J^a$; the two mixed terms $\omega^b\wedge e^c[J_b,P_c]$ and $e^c\wedge\omega^b[P_c,J_b]$ are equal and give $2\,\omega^b\wedge e^c\varepsilon_{bca}P^a$; $[P,P]=0$.

Here T^a is the *torsion* and R^a the (dualised) *Riemann curvature*. The Chern–Simons 3-form built with (12.345) is

$$\begin{aligned}\operatorname{CS}(A):=\Big\langle A\wedge\mathrm{d}A+\tfrac13 A\wedge[A\wedge A]\Big\rangle&\overset{(1)}{=}\big(e_a\wedge\mathrm{d}\omega^a+\omega_a\wedge\mathrm{d}e^a\big)+\varepsilon_{abc}\,e^a\wedge\omega^b\wedge\omega^c\\&\overset{(2)}{=}2\,e_a\wedge\mathrm{d}\omega^a+\varepsilon_{abc}\,e^a\wedge\omega^b\wedge\omega^c+\mathrm{d}\big(e_a\wedge\omega^a\big)\\&\overset{(3)}{=}2\,e_a\wedge R^a+\mathrm{d}\big(e_a\wedge\omega^a\big).\end{aligned} \tag{12.346}$$

(1) $\langle A\wedge\mathrm{d}A\rangle$ pairs e with $\mathrm{d}\omega$ and ω with $\mathrm{d}e$ through (12.345). In $\langle A\wedge[A\wedge A]\rangle$, e^aP_a pairs with the J-part $\varepsilon_{bcd}\omega^b\wedge\omega^cJ^d$ of $[A\wedge A]$, giving $\varepsilon_{abc}e^a\wedge\omega^b\wedge\omega^c$, and ω^aJ_a pairs with the P-part, giving twice that; the total $3\varepsilon_{abc}e^a\omega^b\omega^c$ times $\frac13$.
(2) $\mathrm{d}(e_a\wedge\omega^a)=\mathrm{d}e_a\wedge\omega^a-e_a\wedge\mathrm{d}\omega^a$, and $\omega_a\wedge\mathrm{d}e^a=\mathrm{d}e^a\wedge\omega_a$ (a 1-form and a 2-form commute).
(3) $e_a\wedge R^a=e_a\wedge\mathrm{d}\omega^a+\frac12\varepsilon_{abc}e^a\wedge\omega^b\wedge\omega^c$.

So, up to a boundary term, $\frac12\int\operatorname{CS}(A)=\int e_a\wedge R^a$. In components, $e_a\wedge R^a=\frac12\varepsilon_{abc}e^a\wedge R^{bc}$ equals $\pm\frac12\sqrt{|g|}\,R\,\mathrm{d}^3x$ (the determinant of e comes from $e^a_\rho e^d_\mu e^e_\nu\varepsilon^{\rho\mu\nu}=\det(e)\,\varepsilon^{ade}$, and $\varepsilon_{abc}\varepsilon^{ade}$ contracts $R^{bc}{}_{de}$ to twice the Ricci scalar; the sign is fixed by orientation and signature). This is the Einstein–Hilbert Lagrangian. Hence

$$S_{\mathrm{EH}}=\frac{1}{8\pi G}\int_M e_a\wedge R^a=\frac{k}{4\pi}\int_M\operatorname{CS}(A),\qquad k=\frac{1}{4G}, \tag{12.347}$$

a Chern–Simons action for ISO(2,1) with the form (12.345) in place of the trace.

Field equations and gauge symmetry. The Chern–Simons equation of motion is $F = 0$, i.e. $T^a = 0$ and $R^a = 0$: the spin connection is torsion-free (so it is the Levi-Civita connection of e) and its curvature vanishes: spacetime is flat, which is the vacuum Einstein equation in three dimensions. The gauge transformations with parameter $u = \rho^a P_a + \tau^a J_a$ act as $\delta A = \mathrm{d}u + [A, u]$: the τ's are local Lorentz rotations, and the ρ's are "local translations" $\delta e^a = \mathrm{d}\rho^a + \varepsilon^a{}_{bc}\omega^b\rho^c$, $\delta\omega = 0$. These are the diffeomorphisms, on shell. Indeed, for a vector field ξ,

$$\mathcal{L}_\xi A \overset{(1)}{=} \iota_\xi \mathrm{d}A + \mathrm{d}(\iota_\xi A) \overset{(2)}{=} \iota_\xi F - [\iota_\xi A, A] + \mathrm{d}(\iota_\xi A) \overset{(3)}{=} \iota_\xi F + D(\iota_\xi A),$$

(1) Cartan's formula $\mathcal{L}_\xi = \iota_\xi \mathrm{d} + \mathrm{d}\iota_\xi$ on forms (Chapter T1).
(2) $\iota_\xi F = \iota_\xi \mathrm{d}A + \frac{1}{2}\iota_\xi[A \wedge A]$ and $\frac{1}{2}\iota_\xi[A \wedge A] = [\iota_\xi A, A]$.
(3) $D\lambda = \mathrm{d}\lambda + [A, \lambda]$ for a Lie-algebra-valued function λ, and $-[\lambda, A] = [A, \lambda]$.

On shell $F = 0$, so a diffeomorphism acts as a gauge transformation with parameter $u = \iota_\xi A = \xi^\mu e^a_\mu P_a + \xi^\mu \omega^a_\mu J_a$. The theory is *generally covariant*, yet its action contains no metric: the metric $g_{\mu\nu} = \eta_{ab} e^a_\mu e^b_\nu$ is one of the fields, not a background. This is the sense of Nash's "a gravity theory which is generally covariant but possesses no metric".

Worked example 12.142: A point particle is a Wilson line: the conical deficit

Put a static particle of mass m at the origin of space. The metric $\mathrm{d}s^2 = -\mathrm{d}t^2 + h_{ij}\mathrm{d}x^i\mathrm{d}x^j$ is ultrastatic, and then $R_{00} = 0$ and the scalar curvature is that of the spatial metric h. The 00 Einstein equation reads

$$G_{00} \overset{(1)}{=} R_{00} - \tfrac{1}{2}g_{00}R \overset{(2)}{=} \tfrac{1}{2}R^{(2)} \overset{(3)}{=} K \overset{(4)}{=} 8\pi G\, m\, \delta^2(x),$$

(1) Definition of the Einstein tensor.
(2) $R_{00} = 0$, $g_{00} = -1$ and $R = R^{(2)}$ for an ultrastatic metric.
(3) On a surface the scalar curvature is twice the Gauss curvature K.
(4) Einstein's equation $G_{\mu\nu} = 8\pi G T_{\mu\nu}$ with energy density $T_{00} = m\,\delta^2(x)$.

So space is flat except at the particle, where all the curvature sits: it is a *cone*. By Gauss–Bonnet (Chapter T2), the total curvature $\int K\,\mathrm{d}A = 8\pi G m$ equals the deficit angle of the cone. Cut a wedge of angle $8\pi G m$ out of the plane and glue the edges (Figure 12.95). In gauge-theory language, the holonomy of the flat connection A around a loop encircling the particle is a Lorentz rotation by $8\pi G m$, i.e. $\exp(8\pi G m\, J_0)$. A particle is a Wilson line of ISO(2, 1), labelled by a representation (mass and spin), just as knots were Wilson lines of SU(2).

Why the Hilbert spaces are infinite-dimensional. For compact G, the phase space of Chern–Simons theory on Σ is the moduli space $\mathcal{M}_F(\Sigma)$ of flat connections, which is compact, and a compact phase space of volume $\mathcal{V}$ carries about $\mathcal{V}/(2\pi\hbar)^n$ states (Example 12.95). For ISO(2, 1) the moduli space of flat connections is *non-compact*. On a closed surface of genus $g \geq 2$, the component that describes spacetimes $\Sigma \times \mathbb{R}$ is the cotangent bundle $T^*\mathcal{T}_g$ of the Teichmüller space $\mathcal{T}_g$ (the flat SO(2, 1) part ω gives the hyperbolic structure of the spatial slice, and the translation part e gives a cotangent vector). For the torus $\mathcal{T}_1$ is the upper half plane of modular parameters τ. The quantisation of a cotangent bundle T^*N is the space of wave functions $L^2(N)$, exactly as $T^*\mathbb{R}$ (phase space (x, p)) gives $L^2(\mathbb{R})$ in ordinary quantum mechanics. That space is infinite-dimensional. So in Nash's words, "the Hilbert spaces are no longer finite dimensional".

There is a second difference. For SU(2) the level k had to be an integer because $\pi_3(\mathrm{SU}(2)) = \mathbb{Z}$ allows large gauge transformations that shift the action by $2\pi k$ (Section 12.14.1). The group

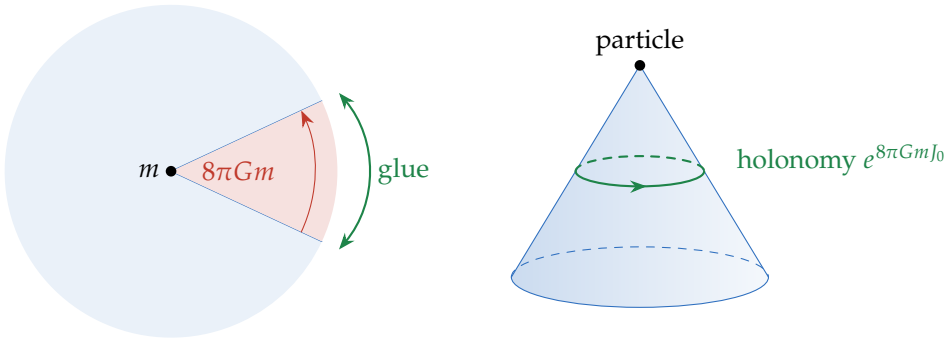


Figure 12.95: A massive particle in 2 + 1 gravity. Left: remove a wedge of angle $8\pi Gm$ from flat space and identify the two edges. Right: the result is a cone, flat everywhere except at the tip. A vector carried round the green loop comes back rotated by the deficit angle: the particle is a Wilson line of the flat ISO(2, 1) connection, and the rotation is its holonomy.

ISO(2, 1) retracts onto its maximal compact subgroup SO(2), which has $\pi_3 = 0$. There are no large gauge transformations of this kind, and the coupling $1/G$ is not quantised.

Topology change. In the Atiyah framework (below), a 3-manifold M with $\partial M = \bar{\Sigma}_1 \sqcup \Sigma_2$ gives a linear map $Z(M) : \mathcal{H}_{\Sigma_1} \to \mathcal{H}_{\Sigma_2}$. If Σ_1 and Σ_2 are different surfaces, this is an *amplitude for space to change its topology*. In Einstein's theory in metric form such processes are hard even to formulate, since a Lorentzian metric on a topology-changing cobordism must degenerate somewhere. In the Chern–Simons form the dreibein e is allowed to be degenerate (it is just a component of a connection), and Witten found that these amplitudes can be non-zero. Figure 12.96 shows the simplest picture.

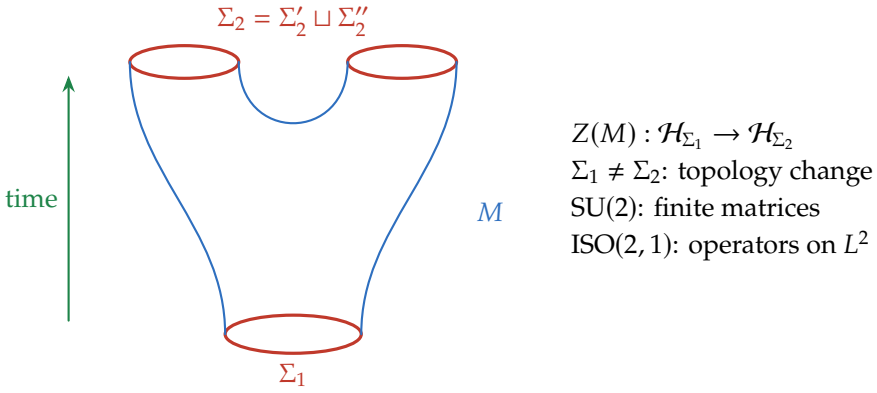


Figure 12.96: A cobordism M from one spatial slice Σ_1 to two. Each red circle stands for a whole closed surface (one dimension is suppressed). The path integral over M is a linear map between the Hilbert spaces of the initial and final slices. When the slices are topologically different, this is a topology-changing amplitude; in 2 + 1 gravity written as a Chern–Simons theory it can be computed and need not vanish.

Physical picture: Adding a cosmological constant

With a cosmological constant Λ the vacuum solutions have constant curvature instead of zero curvature, and the gauge group changes: it is SO(3, 1) for $\Lambda > 0$ (de Sitter) and $\mathrm{SO}(2,2) \cong (\mathrm{SL}(2,\mathbb{R}) \times \mathrm{SL}(2,\mathbb{R}))/\mathbb{Z}_2$ for $\Lambda < 0$ (anti-de Sitter). In the second case the action splits into two $\mathrm{SL}(2,\mathbb{R})$ Chern–Simons actions of opposite levels. The Poincaré group is the

contraction $\Lambda \to 0$ of both, in which the translations P_a stop failing to commute.

The common language: Atiyah's axioms

Nash's last sentence points to the axiomatic approach of Atiyah, analogous to that of Segal for conformal field theories. The axioms were stated at the start of this chapter: a TQFT assigns a vector space $Z(\Sigma)$ to each closed oriented surface and a vector $Z(M) \in Z(\partial M)$ to each oriented 3-manifold, compatibly with gluing. Looking back, every calculation of §6–§7 was an application of one of them. The table records where each axiom was used.

axiom	Chern–Simons realisation	used in
$\Sigma \mapsto Z(\Sigma)$, finite-dim.	$\mathcal{H}_\Sigma$ = conformal blocks of $\mathrm{SU}(2)_k$	Sections 12.14.3 and 12.14.5
$Z(\bar\Sigma) = Z(\Sigma)^*$	$S_k \to -S_k$ under orientation reversal	(12.266)
$Z(\Sigma_1 \sqcup \Sigma_2) = Z(\Sigma_1) \otimes Z(\Sigma_2)$	path integral factorises over pieces	Section 12.14.8
$Z(M) \in Z(\partial M)$; gluing = pairing	$\widetilde{W}(M) = \langle V_1, V_2 \rangle$	(12.267), connected sums
$Z(\Sigma \times I) = \mathbb{1}$	no Hamiltonian, $H = 0$	Example 12.101
$Z(\Sigma \times S^1) = \dim Z(\Sigma)$	$\mathrm{tr}\, e^{iHt} = \mathrm{tr}\, \mathbb{1}$	(12.313)
$\mathrm{Diff}(\Sigma)$ acts on $Z(\Sigma)$	braid group σ; $\mathrm{SL}(2,\mathbb{Z})$ via S, T	Sections 12.14.12 and 12.15.4
(extra structure) framings	twist $e^{2\pi i h}$; anomaly $e^{2\pi i c/24}$	Sections 12.14.16 and 12.15.7
(extra structure) marked points	Wilson lines $\leftrightarrow$ punctures with labels R	Example 12.99

Two lessons stand out. First, the representation of the mapping class group is only *projective*: this is the framing anomaly, which forced the signature phase in (12.329). Chern–Simons theory is a TQFT for 3-manifolds with a little extra structure (a 2-framing, or a choice of signature). Second, the axioms by themselves say nothing about which vector spaces occur; the physics (quantisation of flat connections, conformal field theory) is what supplies the actual $\mathcal{H}_\Sigma$, the matrices S and T, and the braiding. The gravity example shows that relaxing finite-dimensionality leads to theories that are topological but not of Atiyah's finite type.

Reading Nash (p. 360: other groups and other theories)

[Nash p. 360] "When $G \neq \mathrm{SU}(2)$ we obtain a generalisation of the original Jones polynomial. If $G = \mathrm{SU}(N)$ and N is allowed to vary we obtain a two variable polynomial whose variables are determined by N and k." Translation: the whole of §6 goes through with the $\mathrm{SU}(N)$ conformal weights; the Hecke relation (12.336) gives the HOMFLY skein relation (12.337) with $a = q^{N/2}$, $z = q^{1/2} - q^{-1/2}$ and $q = e^{2\pi i/(k+N)}$. "On adjusting the representations of the Wilson lines to be the defining representation of $\mathrm{SU}(N)$ we can recover the Homfly polynomial of Freyd et al." This is Example 12.137; other representations give more terms (Figure 12.92).

Nash then mentions "a topological sigma model (cf. Witten, Floer and Gromov) which, when quantised, exhibits quantum ground states that are like Floer groups": this is Eq. (12.343) and the dictionary of the two Floer theories. Next, "a Chern–Simons topological theory of gravity ... where the group $G = \mathrm{ISO}(2,1)$ ... the moduli space of flat $\mathrm{ISO}(2,1)$-connections is non-compact and the Hilbert spaces are no longer finite dimensional ... non-trivial amplitudes which change the topology of space": this is (12.347), the cotangent-bundle phase space and Figure 12.96. Finally "an axiomatic approach, somewhat analogous to

that introduced for conformal field theories, can be given, cf. Atiyah": the table above.

Recap: Section : surgery and the Jones polynomial

- A solid torus neighbourhood of a knot can be cut out and reglued by a diffeomorphism K of its boundary torus; this is surgery, and by Lickorish–Wallace every closed oriented 3-manifold arises by integer surgery on a link in S^3. Kirby moves say when two links give the same manifold.
- $\mathcal{H}_{T^2}$ has the basis $V^0, \dots, V^k$ (Wilson lines of spin $j = i/2$ on the core), and $\mathrm{SL}(2,\mathbb{Z})$ acts on it through S and T. Surgery changes correlators by $\widetilde{W}(M^K, R_i) = K^i{}_j \widetilde{W}(M, R_j)$ (Nash (12.149)).
- The Hamiltonian is zero, so $Z(\Sigma \times S^1) = \dim \mathcal{H}_\Sigma$ and a mapping torus gives $\operatorname{tr} K$.
- Gluing two solid tori by S turns $S^2 \times S^1$ into S^3. Hence $Z(S^3) = S_{00}$ and the unknot is $S_{0i}/S_{00} = [i+1]_q$; the Hopf link is S_{ij}/S_{00}. This reproduces the skein-relation answer from the modular S-matrix alone.
- For many surgeries, each component is coloured by the Kirby colour $\Omega = \sum_j S_{0j} V^j$; the framing anomaly adds the phase $e^{-2\pi i c\, \sigma(L)/8}$. The result, $Z(M)$, is the Witten–Reshetikhin–Turaev invariant (lens spaces, Poincaré sphere).
- For $\mathrm{SU}(N)$ the same four steps give the HOMFLY polynomial, with $a = q^{N/2}$, $z = q^{1/2} - q^{-1/2}$; groups with self-dual representations give the Kauffman polynomial.
- Other TQFTs: the topological sigma model (holomorphic maps, Gromov–Witten invariants, symplectic Floer homology), and $2+1$ gravity as $\mathrm{ISO}(2,1)$ Chern–Simons theory (no local degrees of freedom, infinite-dimensional Hilbert spaces, topology change). All of them fit Atiyah's axioms, possibly with extra structure.

Chapter summary

This chapter asked one question in three settings: *what can a quantum field theory compute if it never uses a metric?* The answer each time is a topological invariant of the space on which the theory lives, and the mechanism each time is the same pair of tools from the earlier chapters: the path integral localises on the solutions of a first-order equation (flat connections, instantons, holomorphic maps), and the answer is a count of those solutions with signs or phases, or a trace over a finite-dimensional Hilbert space.

1. *What a TQFT is* (Nash §1). Remove the metric from a field theory. Then the stress tensor vanishes (or is Q-exact), the Hamiltonian is zero, the states are all ground states, and the amplitudes depend only on topology. Atiyah's axioms turn this into linear algebra: a vector space for each closed surface, a vector for each 3-manifold bounding it, and gluing by pairing.

2. *Floer homology* (Nash §2). The Chern–Simons function $f(A)$ on the space of SU(2) connections on a homology 3-sphere is a Morse function whose critical points are flat connections and whose gradient lines are instantons on $M \times \mathbb{R}$. Witten's Morse theory of Chapter 7, with relative indices from spectral flow, gives the $\mathbb{Z}_8$-graded groups $HF_*(M)$; their Euler characteristic is twice Casson's invariant.

3. *Donaldson invariants* (Nash §3). Integrating products of 2-dimensional classes over the instanton moduli space $\mathcal{M}_k$ of a 4-manifold gives polynomials q_d on $H_2(M)$. Witten wrote them as correlators of a BRST-exact (cohomological) field theory. Cutting along a homology sphere expresses them as pairings in Floer homology.

4. *Knots* (Nash §4). Knots, links, isotopy, and the classical and modern polynomials (Alexander, Jones, HOMFLY), all characterised by skein relations and the unknot value.

5. *Chern–Simons theory and knots* (Nash §5). Wilson loops in the Chern–Simons theory are knot invariants. For U(1) the path integral is Gaussian and gives the Gauss linking number, with a framing ambiguity for self-linking. At weak coupling the non-abelian theory localises on flat connections, with a one-loop factor given by the Ray–Singer torsion and a phase from the η-invariant.

6. *Chern–Simons theory and the Jones polynomial* (Nash §6, Section 12.14). Quantising on $\Sigma \times \mathbb{R}$ gives finite-dimensional Hilbert spaces, identified with the conformal blocks of the $\mathrm{SU}(2)_k$ WZW model. Gluing and connected sums give unlinks; the 2-dimensional space of the four-punctured sphere forces a three-term linear relation, which is the Jones skein relation with $q = e^{2\pi i/(k+2)}$.

7. *Surgery* (Nash §7, Section 12.15). Cutting out solid tori and regluing them by $\mathrm{SL}(2,\mathbb{Z})$ builds every 3-manifold. The modular matrices S, T act on $\mathcal{H}_{T^2}$, giving $Z(S^3) = S_{00}$, the unknot as S_{01}/S_{00}, the Hopf link as S_{ij}/S_{00}, and the Witten–Reshetikhin–Turaev invariant of any 3-manifold. The same scheme with $\mathrm{SU}(N)$ gives HOMFLY. Other TQFTs: the topological sigma model and 2 + 1 gravity.

Figure 12.97 draws the logic. After it come a formula sheet, two examples that tie the chapter together, the concept tables, and a guide to which tool answers which question.

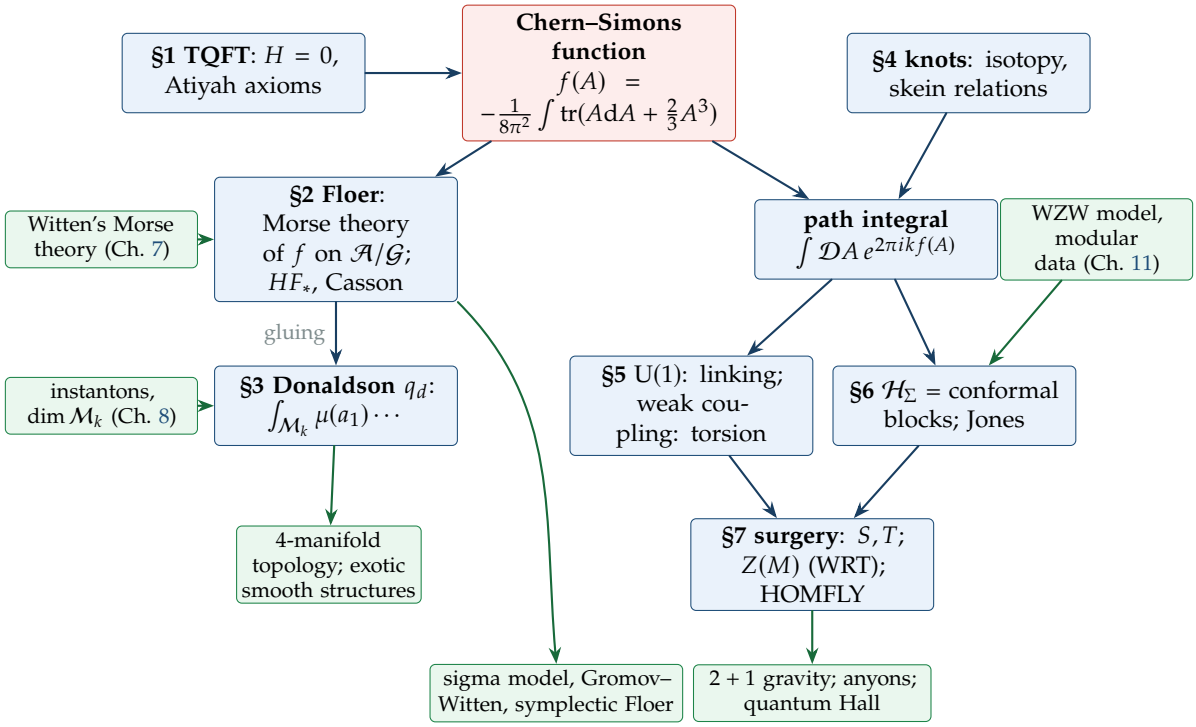


Figure 12.97: Chapter 12 in one picture. The red box, the Chern–Simons function, is used in two ways. On the left it is a Morse function on the space of connections (§2), whose gradient lines are the instantons behind Donaldson's invariants (§3). On the right it is an action, $2\pi k f$, and its path integral with Wilson loops computes knot polynomials (§4–§6) and 3-manifold invariants (§7). Green boxes: the earlier chapters that feed in, and the directions in which these ideas continue.

The formula sheet

The Chern–Simons function and action. With anti-Hermitian A,

$$f(A) = -\frac{1}{8\pi^2}\int_M \mathrm{tr}\Big(A\wedge \mathrm{d}A + \tfrac{2}{3}A\wedge A\wedge A\Big), \qquad f(A^g) = f(A) + \deg g, \qquad S_k = 2\pi k\, f(A),\ k\in\mathbb{Z}.$$

Its variation and flow (§2): $\mathrm{d}f(a) = -\frac{1}{4\pi^2}\int \mathrm{tr}(F\wedge a)$, so critical points are flat connections; $\mathrm{grad}\, f = \star F/4\pi^2$ and the flow $\dot A = -\star F$ is the anti-self-duality equation on $M\times\mathbb{R}$. The spectral flow round a loop in $\mathcal{A}/\mathcal{G}$ is $8k$, whence the $\mathbb{Z}_8$ grading of $HF_*(M)$, and $\chi(HF_*) = 2\lambda(M)$ (Taubes).

Instanton moduli and Donaldson (§3): for SU(2), $\dim \mathcal{M}_k = 8k - 3(1+b_2^+)$ and $q_d(a_1,\dots,a_d) = \int_{\mathcal{M}_k}\mu(a_1)\wedge\cdots\wedge\mu(a_d)$ with $\mu(a) = \int_a c_2(\mathbb{E})$ and $d = \frac{1}{2}\dim\mathcal{M}_k$.

Knot polynomials (§4): the Jones skein $t^{-1}V_+ - tV_- = (t^{1/2}-t^{-1/2})V_0$ with $V(\bigcirc) = 1$; right-handed trefoil $V = t + t^3 - t^4$; HOMFLY $aP_+ - a^{-1}P_- = zP_0$ (Section 12.15.8).

Abelian Chern–Simons (§5): for $S = \frac{k}{4\pi}\int A\wedge \mathrm{d}A$, $\langle\prod_l W(n_l, C_l)\rangle = \exp\big(\frac{i\pi}{k}\sum_{l,m} n_l n_m\, L(C_l, C_m)\big)$ (the overall sign of the exponent depends on the orientation convention), with the Gauss linking number $L(C_1,C_2) = \frac{1}{4\pi}\oint\oint \frac{(\vec x-\vec y)\cdot(\mathrm{d}\vec x\times\mathrm{d}\vec y)}{|\vec x-\vec y|^3}$ and the self-linking fixed by a framing.

Weak coupling (§5): $Z(M) \approx \sum_{\text{flat } A_f} C\, e^{iS_k(A_f)}\, T_{A_f}^{-1/4}$ for isolated, acyclic A_f (Nash (12.113)), with T the Ray–Singer torsion and a further phase fixed by the η-invariant.

Exact $\mathrm{SU}(2)_k$ *data* (§6–§7): $q = e^{2\pi i/(k+2)}$, $h_j = \frac{j(j+1)}{k+2}$, $c = \frac{3k}{k+2}$, $S_{ij} = \sqrt{\frac{2}{k+2}}\sin\frac{(i+1)(j+1)\pi}{k+2}$, $T_{ij} = \delta_{ij}e^{2\pi i(h_i - c/24)}$, braiding eigenvalues $\lambda_0 = -q^{-3/4}$, $\lambda_1 = q^{1/4}$, framing phase $\theta = q^{3/4}$.

Gluing and surgery (§6–§7):

$$\widetilde{W}(M) = \langle V_1, V_2 \rangle, \qquad Z(M_1 \# M_2) = \frac{Z(M_1) Z(M_2)}{Z(S^3)}, \qquad \widetilde{W}(M^K, R_i) = K^i{}_j \, \widetilde{W}(M, R_j),$$

$$Z(S^3) = S_{00}, \qquad W(\bigcirc_i) = \frac{S_{0i}}{S_{00}} = [i+1]_q, \qquad W(\mathrm{Hopf}_{ij}) = \frac{S_{ij}}{S_{00}}, \qquad Z(\Sigma \times S^1) = \dim \mathcal{H}_\Sigma.$$

The WRT invariant of the manifold M_L obtained by integer surgery on a framed link $L = L_1 \cup \cdots \cup L_n$ with linking matrix of signature $\sigma(L)$:

$$Z(M_L) = e^{-2\pi i c\, \sigma(L)/8} \sum_{j_1, \ldots, j_n} \prod_{r=1}^{n} S_{0 j_r} \, \widetilde{W}\big(S^3; L_{j_1 \ldots j_n}\big) \qquad \text{(Eq. (12.329))}.$$

Two examples that tie the chapter together

Worked example 12.143: One family of flat connections, seen by four invariants

Take the Poincaré homology sphere $\Sigma(2,3,5)$ (Example 12.116). Its fundamental group is the binary icosahedral group of order 120, and $\mathrm{Hom}(\pi_1, \mathrm{SU}(2))/\text{conjugation}$ consists of the trivial connection and two irreducible flat connections A_1, A_2. Every invariant of the chapter reads off the same list.

- *Floer* (§2): the chain complex has one generator for each irreducible flat connection. Their relative index is even (they sit in degrees 1 and 5 in the usual grading), so the boundary operator, which changes the degree by one, vanishes and HF_* has rank 2.
- *Casson* (§2): $\chi(HF_*) = (-1)^1 + (-1)^5 = -2$, and Taubes' $\chi(HF_*) = 2\lambda$ gives $\lambda(\Sigma(2,3,5)) = -1$.
- *Chern–Simons values*: $f(A_1) = -\frac{1}{120}$ and $f(A_2) = -\frac{49}{120}$ modulo 1. These are the phases $e^{2\pi i (k+2) f(A_i)}$ that make $|Z|$ oscillate at large k (Example 12.133).
- *WRT* (§7): surgery on the left-handed trefoil with framing -1 gives $Z(\Sigma(2,3,5))$ exactly, e.g. $Z = -\frac{1}{2}$ at $k = 2$, and the large-k asymptotics are a sum of one term per flat connection: the trivial one contributing $O(k^{-3/2})$, the irreducible ones $O(1)$.

The pattern (flat connections as critical points, weighted by $e^{2\pi i k f}$ and a one-loop determinant, or counted with signs) is the whole chapter in miniature. The sum over critical points is the stationary phase version of the path integral; the signed count is its supersymmetric version.

Worked example 12.144: The unknot three ways

The Wilson loop of an unknotted circle in the spin-$\frac{1}{2}$ representation of SU(2), computed by the three methods of the chapter.

$$\begin{aligned}
\text{classical } (k \to \infty): \quad & W(\bigcirc) \overset{(1)}{=} \mathrm{tr}_{\mathbb{C}^2} \mathbb{1} = 2, \\
\text{skein relation (§6)}: \quad & W(\bigcirc) \overset{(2)}{=} -\frac{t - t^{-1}}{t^{1/2} - t^{-1/2}} \overset{(3)}{=} q^{1/2} + q^{-1/2}, \\
\text{surgery and } S \text{ (§7)}: \quad & W(\bigcirc) \overset{(4)}{=} \frac{S_{01}}{S_{00}} \overset{(5)}{=} \frac{\sin\big(2\pi/(k+2)\big)}{\sin\big(\pi/(k+2)\big)} \overset{(6)}{=} 2\cos\frac{\pi}{k+2}.
\end{aligned}$$

(1) At $k \to \infty$ the path integral is dominated by the trivial flat connection $A = 0$, whose

holonomy is $\mathbb{1}$.

(2) The unknot value from the skein relation, Nash's (12.138) (and the second line of (12.158)).

(3) $t = q^{-1}$ and $t^{1/2} = -q^{-1/2}$: numerator $-(q^{-1} - q)$, denominator $-(q^{-1/2} - q^{1/2})$, and $(q - q^{-1})/(q^{1/2} - q^{-1/2}) = q^{1/2} + q^{-1/2}$.

(4) Nash's (12.156): gluing two solid tori by S.

(5) Nash's (12.157) with $i = 1$ and $i = 0$.

(6) $\sin 2x = 2 \sin x \cos x$.

All three agree, since $q^{1/2} + q^{-1/2} = 2 \cos \frac{\pi}{k+2}$, which tends to 2 as $k \to \infty$. The quantum correction is $W(\bigcirc) - 2 \approx -\pi^2/(k+2)^2$.

Concepts of the chapter

In the tables, "operational meaning" says how to compute or recognise the thing; "used in" says where it reappears within the chapter, and, since this is the last chapter of the book, where it leads beyond it.

Topological field theory in general (§1)

concept	operational meaning	used in
TQFT	correlators independent of the metric: check that $T_{\mu\nu} = \delta S/\delta g^{\mu\nu}$ vanishes or is $\{Q,\cdot\}$-exact	every section
Schwarz vs Witten type	metric absent from S (Chern–Simons) vs $S = \{Q, V\}$ (Donaldson, sigma model)	§3, §5, Section 12.15.9
Atiyah axioms	$\Sigma \mapsto Z(\Sigma)$, $M \mapsto Z(M) \in Z(\partial M)$, gluing = pairing, $Z(\Sigma \times S^1) = \dim Z(\Sigma)$	§6–§7 (table in Section 12.15.9)
$H = 0$	time evolution trivial; states are ground states; $Z(\Sigma \times S^1)$ counts them	(12.313)
2d TQFT	commutative Frobenius algebra; $Z(\Sigma_g) = \sum_i \theta_i^{1-g}$	Verlinde formula (12.257)

Floer theory and the Chern–Simons function (§2)

concept	operational meaning	used in
homology 3-sphere	closed 3-manifold with $H_1 = 0$ (so $H_2 = 0$); only reducible flat connection is trivial	§2, §3 gluing; $\Sigma(2,3,5)$
trivial SU(2) bundles over M^3	$B\mathrm{SU}(2) = \mathbb{H}P^\infty$ has cells only in dimensions $4m$	$\mathcal{G} = \mathrm{Map}(M, \mathrm{SU}(2))$
Chern–Simons function f	integrate the CS 3-form; $\mathbb{R}/\mathbb{Z}$-valued on $\mathcal{A}/\mathcal{G}$; jumps by $\deg g$	§5 action, §6–§7
flat connections	$F = 0$; $= \mathrm{Hom}(\pi_1 M, \mathrm{SU}(2))/\mathrm{conj}$ via holonomy	§5 stationary phase; Example 12.143
Hessian $\star \mathrm{d}_A$	first-order, spectrum unbounded both ways; infinite Morse index	relative index; Example 12.141
spectral flow	net number of eigenvalues crossing 0 along a path; = index of $\partial_t + \star \mathrm{d}_{A(t)}$	Floer grading; Chapter 10
gradient flow = instantons	$\dot{A} = -\star F$ on $M \times \mathbb{R}$ is $F = -\star_4 F$	§3 gluing; Floer equation (12.343)
Floer homology $HF_*(M)$	complex on irreducible flat connections, boundary counts instantons, $\mathbb{Z}_8$-graded	§3 pairing $[C^+]\cdot[C^-]$
Casson invariant λ	signed count of irreducible flat connections; $\chi(HF_*) = 2\lambda$	Example 12.143
Floer's Hamiltonian	Witten Laplacian of $\mathrm{d}_t = e^{-tf}\,\mathrm{d}\,e^{tf}$ on $\mathcal{A}/\mathcal{G}$; ground states = HF_*	sigma-model ground states

Donaldson invariants (§3)

concept	operational meaning	used in
intersection form, $b_2^{\pm}$	symmetric matrix of H_2 pairings; count its positive/negative eigenvalues	dim $\mathcal{M}_k$; signature $\sigma(L)$ in (12.329)
dim $\mathcal{M}_k$	index of the ASD deformation complex: $8k - 3(1 + b_2^+)$	selection rules, §3
μ-map	slant product of c_2 of the universal bundle with $a \in H_2(M)$	q_d
Donaldson polynomial q_d	integrate $\mu(a_1)\cdots\mu(a_d)$ over compactified $\mathcal{M}_k$	smooth structures; beyond: Seiberg–Witten
cohomological TQFT	fields $(A, \psi, \phi, \dots)$ with Q^2 = gauge; $S = \{Q, V\}$; localise at Q-fixed points	sigma model, Section 12.15.9
ghost number U	selection rule: $\sum U(\mathcal{O}_i) = \dim \mathcal{M}_k$	(12.341)
Floer gluing	$M = M^+ \cup_N M^-$; $q_d(M) = \langle [C^+], [C^-] \rangle$ in $HF_*(N)$	Atiyah axioms; $\widetilde{W} = \langle V_1, V_2 \rangle$

Knots and knot invariants (§4)

concept	operational meaning	used in
knot, link, isotopy	embedded circles; same if related by Reidemeister moves on a diagram	§5–§7
knot group	$\pi_1(S^3 \setminus K)$ from the Wirtinger presentation	Alexander; flat connections on the complement
linking number	half the signed count of crossings between two components; Gauss integral	§5 abelian CS; framing
framing, writhe	choice of a parallel push-off; writhe = signed crossing count of a diagram	Section 12.14.16, (12.337)
skein triple $L_\pm, L_0$	three links differing inside one ball at one crossing	Section 12.14.13
Alexander $\Delta(t)$	$\Delta_+ - \Delta_- = (t^{1/2} - t^{-1/2})\Delta_0$, or Seifert/Fox matrices	$a = 1$ of HOMFLY
Jones $V(t)$	skein relation plus $V(\bigcirc) = 1$; Kauffman bracket	§6 derivation; Example 12.144
HOMFLY $P(a, z)$	$aP_+ - a^{-1}P_- = zP_0$	SU(N) Chern–Simons, Section 12.15.8
braid, closure	strands going up; closing gives every link (Alexander)	quantum trace (12.300)

Chern–Simons theory and knots (§5)

concept	operational meaning	used in
Wilson loop $W(R, C)$	$\mathrm{tr}_R\ P \exp \oint_C A$: holonomy in representation R	every knot computation
level quantisation	e^{iS_k} gauge invariant iff $k \in \mathbb{Z}$ (large gauge transformations)	§6; contrast ISO(2, 1)
abelian CS	Gaussian integral with sources; propagator = Biot–Savart kernel; answer $e^{i\cdot\text{linking}}$	$\mathrm{SU}(N)_1$, Example 12.138
weak coupling	stationary phase at flat connections; gauge fixing with ghosts	large-k checks Example 12.134
ζ- and η-regularisation	$\det A = e^{-\zeta'_A(0)}$; phase of $\det^{-1/2}$ from $\eta(0)$	shift $k \to k+2$; framing
Ray–Singer torsion	a ratio of ζ-determinants of the Laplacians on forms of each degree (Nash (12.108)); metric-independent	one-loop $Z(M)$

Chern–Simons theory and the Jones polynomial (§6)

concept	operational meaning	used in
phase space $\mathcal{M}_F(\Sigma)$	flat connections on Σ mod gauge, symplectic form $\frac{k}{4\pi}\int \mathrm{tr}\,\delta A \wedge \delta A$	quantisation; gravity (non-compact)
$\mathcal{H}_\Sigma$	holomorphic sections of $\mathcal{L}^k$ over $\mathcal{M}_F$; = conformal blocks; dimension by Verlinde	§7 everywhere
Wilson lines = punctures	a line crossing Σ is a marked point with label R (Borel–Weil)	skein, surgery
$\dim \mathcal{H}_{S^2} = 1$	sphere with no punctures carries one state	connected sums, unlinks
$\dim \mathcal{H}_{S^2,\frac{1}{2}^4} = 2$	two fusion channels (spin 0, 1)	skein relation
braiding matrix	half-twist on the 4-punctured sphere; eigenvalues $-q^{-3/4}, q^{1/4}$	skein; $\mathrm{SU}(N)$ Hecke
fusion matrix F	change of channel basis (quantum $6j$)	braid relation check
quantum dimension	$[n]_q = \sin(n\pi/(k+2))/\sin(\pi/(k+2))$	unknot, Kirby colour
framing anomaly	twist of a line multiplies by $e^{2\pi i h_R}$	writhe normalisation; T-matrix

Surgery and 3-manifold invariants (§7)

concept	operational meaning	used in
solid torus, meridian, longitude	$D^2 \times S^1$; meridian bounds a disc, longitude runs along the core	surgery
mapping class group of T^2	$\mathrm{SL}(2,\mathbb{Z})$ acting on $H_1(T^2)$; generators S, T	$\mathcal{H}_{T^2}$ representation
Heegaard/genus-1 gluing	two solid tori glued by K: S^3 ($K = S$), $S^2 \times S^1$ ($K = \mathbb{1}$), lens spaces	$Z(S^3) = S_{00}$
Dehn surgery p/q	remove a knot neighbourhood, reglue so the meridian goes to $p\mu + q\lambda$	Lickorish–Wallace; Poincaré sphere
Kirby moves	blow-up/down (± 1 unknots) and handle slides relate links giving the same M	invariance test of $Z(M)$
surgery formula	$\widetilde{W}(M^K, R_i) = K^i{}_j \widetilde{W}(M, R_j)$	every closed M
mapping torus	$\Sigma \times_K S^1$; $Z = \mathrm{tr}\, K$	$Z(T^3) = k + 1$
modular S, T	S_{ij} of Nash (12.157); $T = e^{2\pi i(h - c/24)}$; $(ST)^3 = S^2$	unknot, Hopf, Verlinde, WRT
Kirby colour Ω	$\sum_j S_{0j} V^j$ on each surgery component	WRT formula
WRT invariant $Z(M)$	(12.329): coloured link invariant, Gauss-sum phase correction	lens spaces, $\Sigma(2,3,5)$
Hecke relation	$(\sigma - \lambda_S)(\sigma - \lambda_A) = 0$ gives HOMFLY with $a = q^{N/2}$	Section 12.15.8
topological sigma model	localises on holomorphic maps; counts them (Gromov–Witten)	quantum cohomology, Example 12.140
ISO(2,1) Chern–Simons	$A = eP + \omega J$; $S_{\mathrm{CS}} = S_{\mathrm{EH}}$; non-compact phase space	$2+1$ gravity

Which tool answers which question

question	tool
Are two homology 3-spheres different?	Floer groups HF_*, Casson λ, or $Z(M)$ at several k
Are two smooth 4-manifolds (homeomorphic) diffeomorphic?	Donaldson polynomials q_d
Are two knots isotopic?	Jones/HOMFLY/Kauffman polynomials; different values $\Rightarrow$ different knots
Is a knot equivalent to its mirror image?	$V(t)$ vs $V(1/t)$ (the trefoil is chiral)
Linking of two curves, without a diagram	Gauss integral, or abelian Chern–Simons
Exact non-abelian Wilson loops at level k	skein relation from the 4-punctured sphere; or S_{ij}/S_{00}
Partition function of a closed 3-manifold	surgery presentation plus (12.329)
Number of states on a surface	Verlinde formula (12.257) or $Z(\Sigma \times S^1)$
Large-k behaviour, flat connections, torsion	stationary phase (Nash §5), Example 12.134
Counts of holomorphic curves	topological sigma model, Example 12.140
Quantum gravity without local degrees of freedom	ISO(2, 1) Chern–Simons, (12.347)

Physical picture: Where this goes next

The ideas of this chapter did not stop in 1991. In four dimensions, Donaldson's invariants were largely replaced by the Seiberg–Witten invariants, which count solutions of an abelian monopole equation and come from the low-energy limit of the same supersymmetric theory. The Jones polynomial was "categorified" by Khovanov: it is the Euler characteristic of a homology theory, just as Casson's invariant is the Euler characteristic of Floer homology. The topological sigma model became the A-model of mirror symmetry. Abelian Chern–Simons theory at level k is the effective theory of the fractional quantum Hall effect, and the non-abelian theories describe anyons; $SU(2)_3$ (whose anyons have quantum dimension $[2]_q = 2\cos(\pi/5)$, the golden ratio) is universal for topological quantum computation. And $2+1$ gravity with negative cosmological constant, written as two $SL(2,\mathbb{R})$ Chern–Simons theories, is the simplest setting of the AdS/CFT correspondence. Each of these starts from a calculation in this chapter: a metric-independent action, a localisation, and a gluing rule.

Chapter 13

The Mass Gap

Where this chapter is going

The strong nuclear force has a range of about 1.4×10^{-15} m. The electromagnetic force has no range at all: the Coulomb field of a charge falls off as a power of the distance, never exponentially. Both forces come from gauge theories. Why is one short-ranged and the other long-ranged? The answer is a single property of the spectrum of the Hamiltonian, the **mass gap**: an energy $\Delta > 0$ below which there are no excitations of any kind. A theory with a gap has short-range correlations, with range $1/\Delta$. A theory without one has correlations that decay only as a power.

The chapter has three movements. First we say precisely what a gap is and why it forces correlations to decay exponentially (Sections 13.1 and 13.2). Then we collect the ways a gap can be produced. The elementary ones are a mass term, the Higgs mechanism and dimensional transmutation (Section 13.3). The deeper ones are topological: instantons that screen a photon (Section 13.4), a Chern–Simons term that gives a photon a mass with no Higgs field (Section 13.5), a θ-term that decides whether a spin chain is gapped (Section 13.6), and anomalies that *forbid* a trivial gap (Section 13.7), with supersymmetry and topological field theory closing the loop (Section 13.8). Last, we state the Yang–Mills mass gap problem carefully and separate what is known from what is not (Section 13.9). Figure 13.1 shows how the pieces fit, and which earlier chapters each piece leans on.

Dictionary: the words of this chapter

word	what it means physically
mass gap Δ	smallest energy of any excitation above the vacuum
gapped / gapless	$\Delta > 0$ / excitations of arbitrarily small energy
correlation length ξ	range of correlations, $\xi = 1/\Delta$ in a relativistic theory
spectral density $\rho(\mu^2)$	how strongly a field creates states of invariant mass μ
dimensional transmutation	a mass $\mu\, e^{-1/(\beta_0 g^2)}$ made from a dimensionless coupling
monopole-instanton	a tunnelling event in 2+1 dimensions that changes magnetic flux
topological mass	a photon mass from a Chern–Simons term
θ-term	$i\theta\times$(integer winding number) in the Euclidean action
't Hooft anomaly	a symmetry that cannot be coupled to background fields consistently
trivially gapped	gapped, one vacuum, no topological order

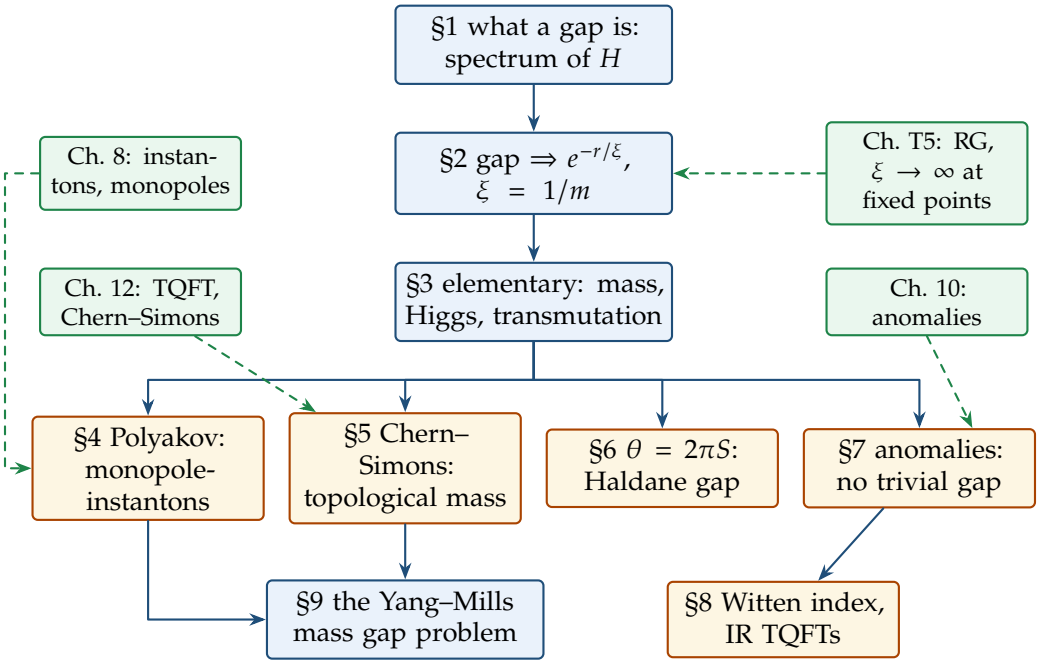


Figure 13.1: The logic of the chapter. The spine (blue) defines the gap and turns it into exponential decay of correlations. The orange boxes are the topological mechanisms, each a different way for topology to open (or to forbid) a gap. The green boxes are the earlier chapters that each piece uses. Everything feeds into the precise statement of the Yang–Mills problem at the bottom.

Reading Nash (no single chapter)

Nash never discusses the spectrum of a Hamiltonian, so there is no "mass gap" in his book. His tools, however, are everywhere in this chapter. Monopoles and their topology are his Chapter VIII, §4 (p. 242 onwards). The Chern–Simons form is his Chapter VIII, §2 (p. 223) and Chapter XII (abelian Chern–Simons theory on S^3, pp. 343–344, (12.86)–(12.89)). Supersymmetric quantum mechanics and Morse theory are Chapter VII, §4 (p. 207, (7.60)–(7.69)). Anomalies are his Chapter X. Here those tools are put to one physical use: deciding whether a theory has a gap.

Conventions. We use natural units $\hbar = c = 1$, except in Section 13.2.6, where $\hbar$ and c are restored to get numbers. Spacetime has dimension d, space has dimension $d-1$. Euclidean signature is the default. When a Minkowski metric is needed we use $\eta = \mathrm{diag}(+1,-1,\ldots,-1)$ and say so. Abelian gauge fields are written in the physicists' real convention, $F = \mathrm{d}A$, with covariant derivative $D_\mu = \partial_\mu - iA_\mu$ on a field of charge one. The dictionary to the anti-Hermitian convention of Chapter T3 is $A_{\text{math}} = -iA$, so $F_{\text{math}} = -iF$.

13.1 What a mass gap is

Here is a question that sounds as if it has an easy answer. *If the mass gap is the mass of the lightest particle, doesn't every theory with massive particles have a gap?* It does not. Quantum electrodynamics has a massive electron, and it has no gap. The reason is that the gap is not about the lightest *particle we care about*. It is about the lightest *excitation of any kind*: massive particles, massless particles, and continua of many-particle states. A single massless particle anywhere in the theory closes the gap. This section makes that precise.

13.1.1 The spectrum of the Hamiltonian

A quantum field theory is, among other things, a Hilbert space $\mathcal{H}$ with two commuting operators: the Hamiltonian H, which generates time translations, and the momentum $\vec{P}$, which generates space translations. Their joint eigenstates $|n\rangle$ obey

$$H|n\rangle = E_n|n\rangle, \qquad \vec{P}|n\rangle = \vec{p}_n|n\rangle.$$

The **vacuum** $|\Omega\rangle$ is the state of lowest energy E_0. It is translation invariant, $\vec{P}|\Omega\rangle = 0$. We may always shift H by a constant so that $E_0 = 0$, and we will often do so.

The set of all energies E_n is the **spectrum** of H. In a finite box the spectrum is a discrete list of numbers. In infinite volume, momenta are continuous and the spectrum contains whole intervals. Picture it as marks on an energy axis: a mark at E_0 for the vacuum, and then marks, or shaded bands, for everything else.

Two remarks keep us out of trouble later.

- In infinite volume a single particle of mass m has energies $\sqrt{m^2 + \vec{p}^2}$ for every $\vec{p}$, so the one-particle states fill the whole interval $[m, \infty)$ of the spectrum of H. To see a particle as an isolated line we look at states of *zero total momentum* (the rest frame), or equivalently at the invariant mass $M = \sqrt{H^2 - \vec{P}^2}$. The gap itself does not care about this distinction, because the bottom of the spectrum is the same either way.

- Positivity of energy, $E_n \geq E_0$, together with Lorentz invariance, forces every state to satisfy $(E_n - E_0)^2 \geq \vec{p}_n^{\,2}$. This is the **spectral condition**: every state lies inside the forward light cone in the $(E, \vec{p})$ plane.

Working definition 13.1 (mass gap)

A quantum system has a **mass gap** $\Delta > 0$ if, in infinite volume, the spectrum of its Hamiltonian H [the list of energies E for which $H|\psi\rangle = E|\psi\rangle$ has a solution] consists of the vacuum energy E_0 [the lowest energy, carried by one state or by finitely many states that no local measurement can tell apart energetically] and then nothing in the open interval $(E_0, E_0 + \Delta)$ [no state at all, whether one particle, many particles or a continuum, has energy strictly between E_0 and $E_0 + \Delta$]. The largest such Δ is the gap. In a Lorentz-invariant theory it equals the smallest invariant mass m of any state other than the vacuum [the mass of the lightest thing that can be created, massless particles included]. *To test for a gap:* put the system in a box of size L, discard the finitely many lowest levels whose splitting from E_0 shrinks exponentially in L, and check that the next level stays at least Δ above E_0 as $L \to \infty$. *Equivalent test in a relativistic theory:* every connected correlation function of local operators decays at least as fast as $e^{-\Delta r}$ at distance r (Section 13.2).

Formal version (as stated by Jaffe and Witten for the Clay problem, with $E_0 = 0$): the theory has a mass gap Δ if H has no spectrum in the interval $(0, \Delta)$; the supremum of such Δ is the mass m, and one requires $m < \infty$ (A. Jaffe and E. Witten, *Quantum Yang–Mills theory*, Clay Mathematics Institute problem description, §4).

13.1.2 The gap is the lightest excitation, not the lightest particle

In a Lorentz-invariant theory, take any state $|n\rangle$ other than the vacuum, with energy E_n and momentum $\vec{p}_n$. Its **invariant mass** M_n is defined by $M_n^2 = (E_n - E_0)^2 - \vec{p}_n^{\,2}$, which is non-negative

by the spectral condition. Let m be the smallest invariant mass that occurs. Then

$$E_n - E_0 \overset{(1)}{=} \sqrt{M_n^2 + \vec{p}_n^{\,2}} \overset{(2)}{\geq} M_n \overset{(3)}{\geq} m. \tag{13.1}$$

(1) Solve the definition of M_n for $E_n - E_0$, taking the positive root because $E_n \geq E_0$.
(2) $\vec{p}_n^{\,2} \geq 0$, and the square root is increasing.
(3) m is the smallest invariant mass in the spectrum, by definition.

The bound is reached: a state of invariant mass m at rest has $E_n - E_0 = m$ exactly. So in a Lorentz-invariant theory

$$\Delta \;=\; m \;=\; \min\{\text{invariant masses of all states other than } |\Omega\rangle\}.$$

The word "all" carries the whole weight. If *any* state has $M_n = 0$, or if invariant masses accumulate at zero, then $\Delta = 0$.

Key idea: gap versus particle mass

A theory is gapped when *every* excitation costs at least a finite energy. "The theory has a massive particle" is not enough. A single massless particle, or a continuum of states reaching down to zero energy, closes the gap, however heavy the other particles are.

Worked example 13.1: the free massive scalar is gapped

A free real scalar of mass m has Hamiltonian (normal ordered, so $E_0 = 0$)

$$H = \int \frac{\mathrm{d}^{d-1}p}{(2\pi)^{d-1}}\, \omega_{\vec{p}}\, a^\dagger_{\vec{p}} a_{\vec{p}}, \qquad \omega_{\vec{p}} = \sqrt{m^2 + \vec{p}^{\,2}},$$

where $a^\dagger_{\vec{p}}$ creates one particle of momentum $\vec{p}$. Every state is built from $n \geq 0$ particles, $a^\dagger_{\vec{p}_1} \cdots a^\dagger_{\vec{p}_n} |0\rangle$. For such a state with $n \geq 1$,

$$E - E_0 \overset{(1)}{=} \omega_{\vec{p}_1} + \cdots + \omega_{\vec{p}_n} \overset{(2)}{\geq} n\, m \overset{(3)}{\geq} m.$$

(1) $[H, a^\dagger_{\vec{p}}] = \omega_{\vec{p}} a^\dagger_{\vec{p}}$, so each creation operator raises the energy by $\omega_{\vec{p}}$, and $H|0\rangle = 0$.
(2) Each $\omega_{\vec{p}_i} = \sqrt{m^2 + \vec{p}_i^{\,2}} \geq m$.
(3) $n \geq 1$.

So $\Delta = m$. At zero total momentum the spectrum is: 0 (the vacuum), m (one particle at rest), and the continuum $[2m, \infty)$ (two particles with momenta $\pm\vec{k}$, energy $2\sqrt{m^2 + \vec{k}^2}$), then $[3m, \infty)$, and so on (Figure 13.2(a)).

Worked example 13.2: QED has a massive electron and no gap

QED contains the electron, $m_e = 0.511\,\mathrm{MeV}$, and the photon, which is massless. A state with two photons of momenta $\vec{k}$ and $-\vec{k}$ has total momentum zero and energy

$$E - E_0 \overset{(1)}{=} |\vec{k}| + |-\vec{k}| \overset{(2)}{=} 2|\vec{k}|.$$

(1) A photon of momentum $\vec{k}$ has energy $|\vec{k}|$, since its mass is zero, and the two photons are (to leading order) free.
(2) $|-\vec{k}| = |\vec{k}|$.

Given any proposed gap $\Delta > 0$, choose $|\vec{k}| < \Delta/2$. This state has energy below Δ. So no $\Delta > 0$ works: QED is gapless, although its only charged particle is massive (Figure 13.2(b)). Physically, the long range of the Coulomb force and the absence of a gap are the same fact.

Worked example 13.3: Goldstone bosons close the gap

When a continuous symmetry is spontaneously broken, there are gapless excitations, the **Goldstone bosons** (Goldstone's theorem). The picture is simple. The vacua form a continuous family (a circle of values of a phase, say). A long-wavelength wave in which the field slowly rotates from one vacuum to a neighbouring one costs almost no energy, and its energy goes to zero as the wavelength goes to infinity: $\omega = c\,|\vec{k}|$ for some speed c. Examples:

- phonons in a crystal (broken translations), $\omega = c_s|\vec{k}|$;
- spin waves in a Heisenberg antiferromagnet in three dimensions (broken spin rotations);
- the pions of QCD (broken chiral symmetry). Here the symmetry is also broken slightly *explicitly* by the quark masses, so the pions have a small mass, $m_\pi \approx 140\,\text{MeV}$, much less than the proton mass 938 MeV. In the limit of massless quarks QCD would be gapless.

In each case there are states of arbitrarily small energy, so $\Delta = 0$ (Figure 13.2(c)). We return to this in Section 13.3.2, where we see that gauging the broken symmetry removes the Goldstone boson and opens a gap.

Watch out: "massive theory"

Physicists say "a massive theory" to mean a theory in which *all* excitations are massive. Such a theory is gapped by (13.1). A theory in which *some* particles are massive, like QED or the Standard Model, need not be gapped. There are also gapless theories with no particles at all: a conformal field theory (Chapter 11) has a continuum of states reaching down to zero energy and no isolated one-particle lines.

13.1.3 Multi-particle continua and the energy–momentum plane

It helps to draw the spectrum in the plane of energy E (with $E_0 = 0$) against momentum $p = |\vec{p}|$. Each state is a point in the forward light cone $E \geq p$. A particle of mass m traces the hyperbola $E = \sqrt{m^2 + p^2}$.

Two particles of mass m with momenta $\vec{k}_1, \vec{k}_2$ have total momentum $\vec{p} = \vec{k}_1 + \vec{k}_2$ and energy $E = \omega_{\vec{k}_1} + \omega_{\vec{k}_2}$. The smallest energy at fixed total momentum is reached when the two particles move together, each carrying half the momentum:

$$E_{\min}(\vec{p}) \overset{(1)}{=} 2\,\omega_{\vec{p}/2} \overset{(2)}{=} 2\sqrt{m^2 + \tfrac{1}{4}\vec{p}^{\,2}} = \sqrt{(2m)^2 + \vec{p}^{\,2}}. \tag{13.2}$$

(1) $\omega_{\vec{k}}$ is a convex function of $\vec{k}$, so for fixed $\vec{k}_1 + \vec{k}_2 = \vec{p}$ the sum $\omega_{\vec{k}_1} + \omega_{\vec{k}_2}$ is smallest at $\vec{k}_1 = \vec{k}_2 = \vec{p}/2$.
(2) Definition of ω.

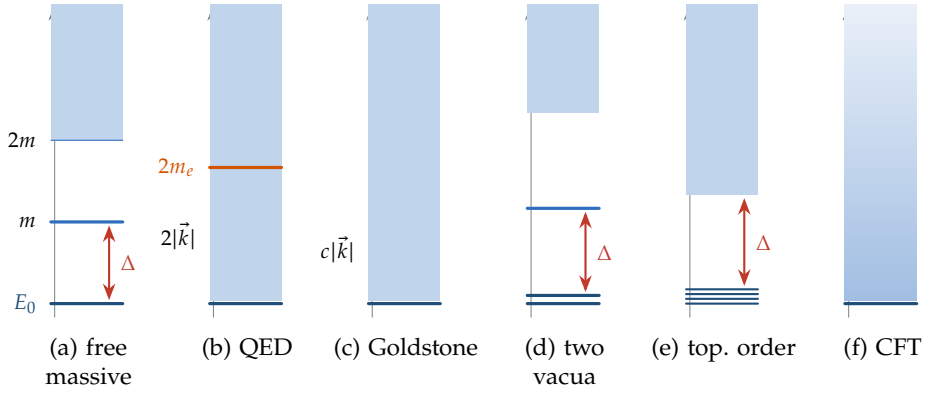


Figure 13.2: Spectra at zero total momentum, energy upwards. Dark lines are vacua, blue lines are one-particle states at rest, shaded bands are continua. (a) A free massive field: a gap $\Delta = m$, then the two-particle continuum at $2m$. (b) QED: the heavy electron pair threshold at $2m_e$ is irrelevant, because two back-to-back photons fill the spectrum down to zero. (c) Goldstone bosons do the same. (d) A spontaneously broken discrete symmetry: two vacua, then a gap. (e) Topological order on a torus: several vacua, split only by an amount exponentially small in the system size, then a gap. (f) A conformal field theory: a continuum from zero, with no particles.

So the two-particle states fill the region above the hyperbola of mass $2m$. They form a continuum: an interval of invariant masses $[2m, \infty)$. A **bound state** of two particles would show up as an extra hyperbola with mass between m and $2m$ (like the deuteron below the proton–neutron threshold). The gap is the lowest of all these curves at $p = 0$ (Figure 13.3).

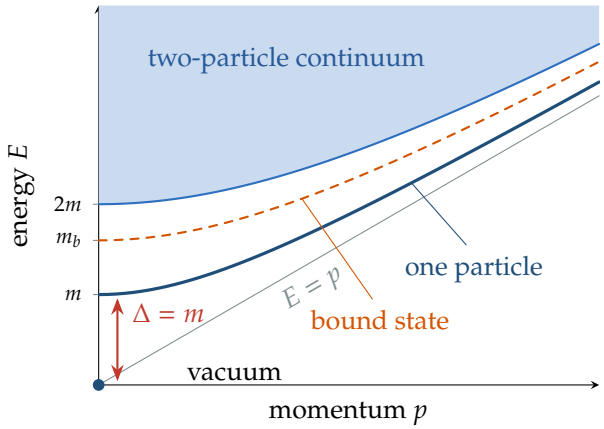


Figure 13.3: The energy–momentum plane of a gapped theory. Every state lies above the light line $E = p$. A particle of mass m is the hyperbola $E = \sqrt{m^2 + p^2}$ (dark blue). A bound state of mass $m_b < 2m$ is another hyperbola (dashed orange). Two-particle states fill the whole region above the hyperbola of mass $2m$ (shaded). The gap is the height of the lowest curve at $p = 0$. In a gapless theory some curve, or some shaded region, touches the origin.

Try it: the $3m$ threshold

In ϕ^4 theory the field ϕ is odd under $\phi \to -\phi$, so acting with ϕ on the vacuum creates only states with an odd number of particles. Where does the continuum seen by the operator ϕ start? *Answer:* one particle at m, then three particles at $3m$. The two-particle continuum at $2m$ is created by even operators such as ϕ^2.

13.1.4 Gapped phases with several vacua

A gap does not require a unique vacuum. The working definition allows finitely many vacua, and two important classes of gapped phases need this.

Broken discrete symmetry. Take a real scalar in the double-well potential $V(\phi) = \frac{\lambda}{4}(\phi^2 - v^2)^2$. It has two classical vacua $\phi = \pm v$, exchanged by $\phi \to -\phi$. In infinite volume the quantum theory also has two vacua $|\Omega_\pm\rangle$, and above each one every excitation is massive: small oscillations have mass $m^2 = V''(\pm v) = 2\lambda v^2$, and in $1+1$ dimensions there are also **kinks**, domain walls interpolating from $-v$ to $+v$, with a finite mass. The spectrum is two vacuum lines and then a gap (Figure 13.2(d)). Jaffe and Witten describe exactly this situation for the rigorously constructed two-dimensional ϕ^4 theory at strong coupling: two phases, each with a mass gap and a unique vacuum. Choosing one vacuum is choosing one "phase".

Topological order. Some gapped phases have several vacua that are not related by any symmetry. Their number depends on the topology of space: one vacuum on a sphere, four on a torus for the simplest example (the $\mathbb{Z}_2$ toric code of Section 13.8.2). No local operator distinguishes these vacua, and in a box of size L they are split only by an amount of order $e^{-L/\xi}$ (Figure 13.2(e)). The low-energy physics is a topological quantum field theory (Chapter 12).

Key idea: three kinds of gapped phase

- *Trivially gapped:* one vacuum on every spatial manifold, and no topological response. This is the phase of a free massive field.
- *Gapped with broken symmetry:* several vacua, permuted by a broken discrete symmetry.
- *Gapped with topological order:* vacua whose number depends on the topology of space, not related by symmetry.

In all three the spectrum above the vacua has a gap. Section 13.7 shows that anomalies can rule out the first kind while allowing the other two.

13.1.5 Finite volume

The definition insists on infinite volume, and for good reason. In a finite box every spectrum is discrete, so every finite system is "gapped" in a trivial sense. What matters is how the lowest gaps behave as the size L grows. There are three typical behaviours.

Worked example 13.4: a massless scalar on a circle

Put a free massless scalar on a circle of circumference L (one space dimension). Periodicity quantises the momenta, $p_n = 2\pi n/L$ with $n \in \mathbb{Z}$. The lowest excitation with nonzero momentum has energy

$$E_1 - E_0 \overset{(1)}{=} |p_{\pm 1}| \overset{(2)}{=} \frac{2\pi}{L} \xrightarrow{L\to\infty} 0.$$

(1) A massless particle has $E = |p|$.
(2) $n = \pm 1$.

In a finite box the gap is $2\pi/L > 0$, but it closes like $1/L$. This $1/L$ closing is the fingerprint of a gapless theory. For a massive scalar the same lowest state (one particle at rest, $p = 0$) has $E_1 - E_0 = m$ for every L, and the gap stays open.

Worked example 13.5: the transverse-field Ising chain shows all three behaviours

Take L spin-$\frac{1}{2}$ sites on a ring with

$$H = -J \sum_{j=1}^{L} \sigma_j^z \sigma_{j+1}^z - h \sum_{j=1}^{L} \sigma_j^x, \qquad J, h > 0,$$

where $\sigma_j^{x,z}$ are Pauli matrices on site j. This model can be solved exactly by turning spins into free fermions (the Jordan–Wigner transformation), and the single-fermion energies are

$$\varepsilon_k = 2\sqrt{J^2 + h^2 - 2Jh\cos k}, \qquad k \in \left\{\tfrac{2\pi n}{L}\right\}.$$

The smallest value is at $k = 0$:

$$\varepsilon_0 \overset{(1)}{=} 2\sqrt{J^2 + h^2 - 2Jh} \overset{(2)}{=} 2\sqrt{(J-h)^2} = 2|J - h|.$$

(1) $\cos 0 = 1$.
(2) Complete the square.

- $h > J$ *(disordered):* a unique vacuum, spins mostly along x, and a gap $2(h - J)$.
- $h < J$ *(ordered):* two nearly degenerate lowest states, close to $|{\uparrow\uparrow}\cdots\rangle \pm |{\downarrow\downarrow}\cdots\rangle$. To connect $|{\uparrow\uparrow}\cdots\rangle$ to $|{\downarrow\downarrow}\cdots\rangle$ the term $-h\sigma^x$ must flip all L spins, so their splitting appears first at order h^L in perturbation theory and is of order $J(h/J)^L = Je^{-L\ln(J/h)}$, exponentially small in L. Above this pair there is a gap $2(J - h)$.
- $h = J$ *(critical):* $\varepsilon_k = 4J|\sin(k/2)|$, so the lowest excitation with $k = 2\pi/L$ costs about $2\pi J/L$. The gap closes like $1/L$, and the long-distance theory is a conformal field theory (the critical Ising model of Chapter 11).

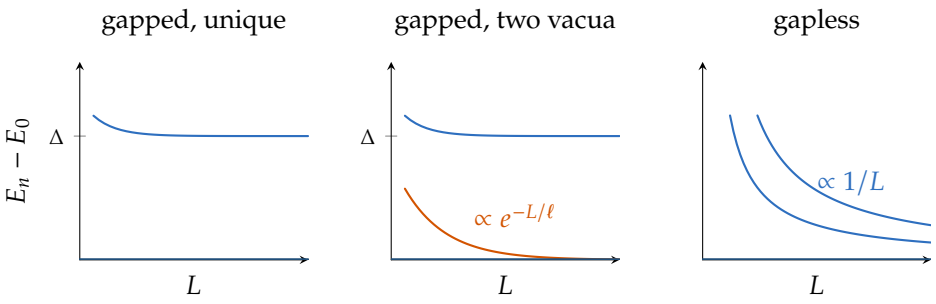


Figure 13.4: How the lowest energy differences behave as the box size L grows. Left: a gapped theory with one vacuum, whose first excitation tends to a finite Δ. Middle: a gapped theory with two vacua. The second level (orange) falls onto the vacuum exponentially fast, and the gap Δ sits above the pair. Right: a gapless theory, where every level falls like $1/L$ and nothing is left in infinite volume.

The exponential splitting in the ordered phase has a general origin. To go from one vacuum to the other, the whole system must tunnel. In Euclidean language the tunnelling event is a domain wall that sweeps across space once, an instanton (Chapter 8) whose action is the wall tension σ times the area of space, L^{d-1}. The splitting is therefore of order $e^{-\sigma L^{d-1}}$, which vanishes as $L \to \infty$. This is why the working definition discards "finitely many levels whose splitting shrinks exponentially". They become exactly degenerate vacua in infinite volume.

Watch out: a finite-volume gap proves nothing

Numerical studies always work in finite volume, and they always find a gap. The question is whether it survives $L \to \infty$. A lattice gauge theory calculation that reports a glueball mass is really reporting that the lowest excitation stays at a fixed value in physical units as both $L \to \infty$ and the lattice spacing $a \to 0$. Jaffe and Witten stress that a gap which is *uniform in the volume* is exactly what current mathematical methods cannot yet prove for Yang–Mills theory (Section 13.9).

theory	lightest excitation	gapped?
free scalar of mass m	particle of mass m	yes, $\Delta = m$
QED	two photons, $E = 2\lvert\vec{k}\rvert$	no
QCD with massless quarks	pions (Goldstone bosons)	no
real QCD	pions, $m_\pi \approx 140\,\text{MeV}$	yes, but only because $m_q \neq 0$
pure Yang–Mills, $d = 4$	glueball (conjectured)	believed yes, unproven
superconductor (Higgs phase)	massive photon	yes
critical Ising model	continuum from 0	no

Recap

- The mass gap is a property of the spectrum of H in infinite volume: nothing between the vacuum energy and $E_0 + \Delta$.
- In a Lorentz-invariant theory Δ is the smallest invariant mass of *any* state. One massless particle, or a continuum reaching zero, closes the gap. QED and theories with Goldstone bosons are gapless.
- Multi-particle states form continua starting at thresholds such as $2m$. Bound states sit below them.
- Gapped phases may have several vacua, from broken discrete symmetry or from topological order.
- In a box of size L, gapped theories keep a finite Δ (after exponentially split vacua are merged). Gapless theories have levels falling like $1/L$.

13.2 The gap makes correlations decay exponentially

We almost never diagonalise a Hamiltonian. What we can compute, on a lattice, in perturbation theory or in an experiment, are *correlation functions*: how much a measurement at one point is correlated with a measurement at another. So we need a bridge between the spectrum and correlations. The bridge is to insert a complete set of energy eigenstates between two field operators. The result is simple and exact: *a gap Δ makes every connected correlation function decay like $e^{-\Delta r}$*. In a relativistic theory the range of correlations is therefore $\xi = 1/\Delta$, which is the Compton wavelength of the lightest particle.

13.2.1 Imaginary time and the spectral sum

In real time, a state of energy E evolves by the phase e^{-iEt}. Phases oscillate and never decay. Replace t by $-i\tau$ (a **Wick rotation**). Then e^{-iEt} becomes $e^{-E\tau}$, and high energies are exponentially

damped. This is also the time variable of the Euclidean path integral, with weight e^{-S_E}. The **Euclidean Heisenberg field** at imaginary time τ and position $\vec{x}$ is

$$\phi(\tau,\vec{x}) = e^{H\tau}\, e^{-i\vec{P}\cdot\vec{x}}\, \phi(0)\, e^{i\vec{P}\cdot\vec{x}}\, e^{-H\tau}, \tag{13.3}$$

where $\phi(0)$ is a Hermitian field operator at the origin. (The sign in the spatial translation makes $\langle\Omega|\phi(0,\vec{x})|\vec{p}\rangle = e^{i\vec{p}\cdot\vec{x}}\langle\Omega|\phi(0)|\vec{p}\rangle$, as for a free field $\phi \ni a_{\vec{p}}e^{i\vec{p}\cdot\vec{x}}$.)

For $\tau > 0$, the **Euclidean two-point function** is $G(\tau,\vec{x}) = \langle\Omega|\phi(\tau,\vec{x})\phi(0,\vec{0})|\Omega\rangle$. Write $c_n = \langle\Omega|\phi(0)|n\rangle$ for the amplitude with which ϕ connects the vacuum to the eigenstate $|n\rangle$. Then

$$\begin{aligned} G(\tau,\vec{x}) &\overset{(1)}{=} \langle\Omega|e^{H\tau}e^{-i\vec{P}\cdot\vec{x}}\phi(0)e^{i\vec{P}\cdot\vec{x}}e^{-H\tau}\,\phi(0)|\Omega\rangle \\ &\overset{(2)}{=} \sum_n \langle\Omega|e^{H\tau}e^{-i\vec{P}\cdot\vec{x}}\phi(0)e^{i\vec{P}\cdot\vec{x}}e^{-H\tau}|n\rangle\langle n|\phi(0)|\Omega\rangle \\ &\overset{(3)}{=} \sum_n e^{E_0\tau}\, e^{-E_n\tau}\, e^{i\vec{p}_n\cdot\vec{x}}\, \langle\Omega|\phi(0)|n\rangle\langle n|\phi(0)|\Omega\rangle \\ &\overset{(4)}{=} \sum_n |c_n|^2\, e^{-(E_n-E_0)\tau + i\vec{p}_n\cdot\vec{x}}. \end{aligned} \tag{13.4}$$

(1) Definition (13.3), and $\phi(0,\vec{0}) = \phi(0)$.

(2) Insert the identity $\mathbb{1} = \sum_n |n\rangle\langle n|$, a sum over all joint eigenstates of H and $\vec{P}$ (it includes integrals over continuous momenta).

(3) Acting to the right, $e^{i\vec{P}\cdot\vec{x}}e^{-H\tau}|n\rangle = e^{i\vec{p}_n\cdot\vec{x}}e^{-E_n\tau}|n\rangle$. Acting to the left, $\langle\Omega|e^{H\tau}e^{-i\vec{P}\cdot\vec{x}} = e^{E_0\tau}\langle\Omega|$ because $\vec{P}|\Omega\rangle = 0$.

(4) $\phi(0)$ is Hermitian, so $\langle n|\phi(0)|\Omega\rangle = \overline{\langle\Omega|\phi(0)|n\rangle} = \overline{c_n}$.

For $\tau < 0$ the operators are ordered the other way (Euclidean time ordering), and the same steps give (13.4) with τ replaced by $|\tau|$.

The term $n = \Omega$ is the constant $|\langle\Omega|\phi|\Omega\rangle|^2 = \langle\phi\rangle^2$. It does not depend on the separation, and it carries no information about excitations. Subtracting it defines the **connected** two-point function

$$G_c(\tau,\vec{x}) = G(\tau,\vec{x}) - \langle\phi\rangle^2 = \sum_{n\neq\Omega} |c_n|^2\, e^{-(E_n-E_0)|\tau| + i\vec{p}_n\cdot\vec{x}}. \tag{13.5}$$

If there are several vacua, we work in one of them. In infinite volume a local operator cannot take one vacuum to another (they differ everywhere in space), so $c_n = 0$ for the other vacua and they do not appear in the sum.

Physical picture: oscillation in real time, decay in imaginary time

Every term in (13.5) is an exponential $e^{-(E_n-E_0)|\tau|}$. At large $|\tau|$ the terms with larger energy die first, and the sum is dominated by the *lowest* state that ϕ can create. Undo the Wick rotation and the same term is a phase $e^{-i(E_n-E_0)t}$: the gap is the lowest frequency at which the vacuum can ring when struck by ϕ.

13.2.2 From a gap to exponential decay

Now assume a gap, $E_n - E_0 \geq \Delta$ for all $n \neq \Omega$. Fix a small reference time $\tau_0 > 0$ (we need $\tau_0 > 0$ because at $\tau = 0$ the sum is $\langle\phi(0)^2\rangle$, which is ultraviolet divergent for a point-like field). For

$\tau \geq \tau_0$ at $\vec{x} = 0$,

$$G_c(\tau, \vec{0}) \overset{(1)}{=} \sum_{n \neq \Omega} |c_n|^2 \, e^{-(E_n - E_0)\tau_0} \, e^{-(E_n - E_0)(\tau - \tau_0)}$$
$$\overset{(2)}{\leq} e^{-\Delta(\tau - \tau_0)} \sum_{n \neq \Omega} |c_n|^2 \, e^{-(E_n - E_0)\tau_0}$$
$$\overset{(3)}{=} e^{-\Delta(\tau - \tau_0)} \, G_c(\tau_0, \vec{0}). \tag{13.6}$$

(1) (13.5) at $\vec{x} = 0$, with the exponent split as $\tau = \tau_0 + (\tau - \tau_0)$.
(2) Every term is non-negative, $\tau - \tau_0 \geq 0$, and $E_n - E_0 \geq \Delta$, so each factor $e^{-(E_n - E_0)(\tau - \tau_0)}$ is at most $e^{-\Delta(\tau - \tau_0)}$.
(3) The remaining sum is (13.5) at $\tau = \tau_0$, $\vec{x} = 0$.

At nonzero $\vec{x}$ the phases can only help:

$$|G_c(\tau, \vec{x})| \overset{(1)}{\leq} \sum_{n \neq \Omega} |c_n|^2 \, e^{-(E_n - E_0)|\tau|} \left| e^{i \vec{p}_n \cdot \vec{x}} \right| \overset{(2)}{=} G_c(\tau, \vec{0}).$$

(1) The triangle inequality, $|\sum a_n| \leq \sum |a_n|$.
(2) $|e^{i\theta}| = 1$, and what is left is (13.5) at $\vec{x} = 0$.

So in *imaginary time* the correlation decays at least as $e^{-\Delta|\tau|}$.

What about two points at the same time, separated in space? In a Lorentz-invariant theory, the Euclidean correlation functions are invariant under rotations of $\mathbb{R}^d$, which mix τ with $\vec{x}$. So $G_c(\tau, \vec{x}) = F\left(\sqrt{\tau^2 + \vec{x}^2}\right)$ for some function F of the Euclidean distance alone. Then

$$G_c(0, \vec{x}) \overset{(1)}{=} F(|\vec{x}|) \overset{(2)}{=} G_c(|\vec{x}|, \vec{0}) \overset{(3)}{\leq} e^{-\Delta(|\vec{x}| - \tau_0)} \, G_c(\tau_0, \vec{0}). \tag{13.7}$$

(1) Rotation invariance at $\tau = 0$.
(2) The same function F evaluated at a purely temporal separation of the same length.
(3) The bound (13.6) with $\tau = |\vec{x}|$.

Key idea: the gap gives clustering

If H has a gap Δ, then every connected two-point function of local operators decays at least as $e^{-\Delta r}$, in imaginary time and, in a Lorentz-invariant theory, also at spacelike separation r. This is the **clustering** property that Jaffe and Witten single out as "an important consequence of the existence of a mass gap": for any $C < \Delta$, $|\langle \Omega, O(\vec{x}) O(\vec{y}) \Omega \rangle| \leq e^{-C|\vec{x} - \vec{y}|}$ for large $|\vec{x} - \vec{y}|$, when $\langle O \rangle = 0$.

Watch out: without Lorentz invariance

Step (1) of (13.7) used rotation invariance of Euclidean spacetime. A lattice magnet has no such symmetry, so the argument above only gives decay in imaginary time. Decay in space is still true for lattice systems with local interactions and a gap, but it needs a separate proof (the Hastings–Koma theorem, which gives equal-time correlations bounded by $e^{-r/\xi}$ with ξ set by the gap and the speed at which information spreads).

Reading off masses. The same formula tells us how masses are measured in lattice simulations. Integrate over space to project onto zero total momentum:

$$C(\tau) = \int \mathrm{d}^{d-1}x \; G_c(\tau, \vec{x}) \overset{(1)}{=} \sum_{n \neq \Omega} |c_n|^2 \, (2\pi)^{d-1} \delta^{d-1}(\vec{p}_n) \, e^{-(E_n - E_0)|\tau|} \tag{13.8}$$

(1) $\int \mathrm{d}^{d-1}x\, e^{i\vec{p}\cdot\vec{x}} = (2\pi)^{d-1}\delta^{d-1}(\vec{p})$.

Only states at rest survive, and their energies are their masses. At large τ the lightest state created by ϕ dominates, $C(\tau) \approx A\, e^{-m\tau}$.

Worked example 13.6: the effective-mass plateau

Suppose ϕ creates two states at rest, of masses $m_1 < m_2$, with weights $A_1, A_2 > 0$, so $C(\tau) = A_1 e^{-m_1\tau} + A_2 e^{-m_2\tau}$. Define the **effective mass** $m_{\text{eff}}(\tau) = -\mathrm{d}\ln C/\mathrm{d}\tau$. Then

$$m_{\text{eff}}(\tau) \overset{(1)}{=} \frac{m_1 A_1 e^{-m_1\tau} + m_2 A_2 e^{-m_2\tau}}{A_1 e^{-m_1\tau} + A_2 e^{-m_2\tau}}$$
$$\overset{(2)}{=} \frac{m_1 + m_2\, r\, e^{-(m_2-m_1)\tau}}{1 + r\, e^{-(m_2-m_1)\tau}}, \qquad r = A_2/A_1,$$
$$\overset{(3)}{=} m_1 + (m_2 - m_1)\, r\, e^{-(m_2-m_1)\tau} + O\big(e^{-2(m_2-m_1)\tau}\big).$$

(1) $-\mathrm{d}\ln C/\mathrm{d}\tau = -C'/C$, and $-\frac{\mathrm{d}}{\mathrm{d}\tau}e^{-m\tau} = m\, e^{-m\tau}$.
(2) Divide numerator and denominator by $A_1 e^{-m_1\tau}$.
(3) Expand $1/(1+\epsilon) = 1 - \epsilon + O(\epsilon^2)$ with $\epsilon = re^{-(m_2-m_1)\tau}$, and collect terms.

The effective mass falls from above onto a *plateau* at m_1, and it approaches the plateau at a rate set by the splitting $m_2 - m_1$ (Figure 13.5). Lattice glueball masses are read off from exactly such plateaux, with ϕ a gauge-invariant operator like $\mathrm{tr}\, F_{ij}F_{ij}$ summed over a time slice.

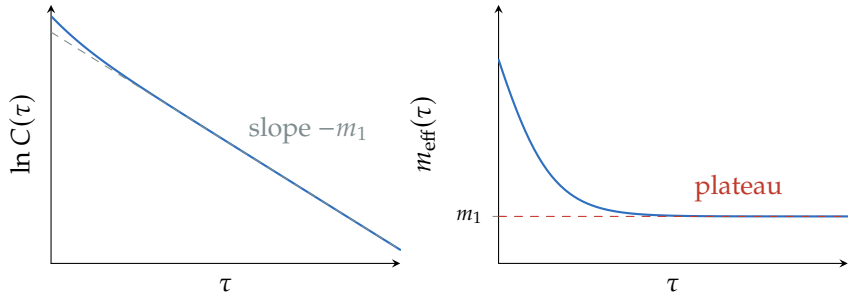


Figure 13.5: Reading a mass off a Euclidean correlator. Left: at zero momentum, $\ln C(\tau)$ becomes a straight line whose slope is minus the lightest mass. Right: the effective mass $-\mathrm{d}\ln C/\mathrm{d}\tau$ falls onto a plateau at m_1. The heavier state only bends the curve at small τ.

The converse. The implication also runs the other way in a relativistic theory. If every connected correlator of local operators decays at least as e^{-mr}, then the zero-momentum projection (13.8) decays at least as $e^{-m\tau}$ for every local operator. Since local operators acting on the vacuum create (a dense set of) all states, no state lies below m. So "gap" and "exponential clustering of all local correlators" are two ways of saying the same thing.

13.2.3 The Källén–Lehmann representation

Lorentz invariance lets us organise the spectral sum by invariant mass. Write $p = (p^0, \vec{p})$ for a d-momentum and $p^2 = (p^0)^2 - \vec{p}^{\,2}$ for its Minkowski square. Each state $|n\rangle$ has $p_n = (E_n - E_0, \vec{p}_n)$. Define the **spectral density** ρ by

$$\sum_{n\neq\Omega} |c_n|^2\, (2\pi)^d\, \delta^d(p - p_n) = 2\pi\, \theta(p^0)\, \rho(p^2). \tag{13.9}$$

The left side is a sum of positive delta functions, so $\rho \geq 0$. It is Lorentz invariant (each Lorentz transformation permutes the states $|n\rangle$ without changing $|c_n|$ for a scalar ϕ), so it can depend on p only through p^2 and the sign of p^0. The spectral condition says it vanishes unless $p^2 \geq 0$ and $p^0 > 0$.

Working definition 13.2 (spectral density)

The **spectral density** $\rho(\mu^2)$ of a field ϕ is the non-negative function defined by (13.9) [how strongly ϕ, acting on the vacuum, creates states of invariant mass μ]. A single particle of mass m shows up as a delta function $Z\,\delta(\mu^2 - m^2)$ [$Z = |c|^2$ is the probability-like weight with which ϕ creates that particle], and many-particle states show up as a continuous part starting at a threshold [the smallest total invariant mass of the particles, for example $2m$]. *To test for a gap:* the gap is the lowest μ at which ρ of some local operator is nonzero.

Now rewrite (13.5) with ρ. For $\tau > 0$,

$$
\begin{aligned}
G_c(\tau,\vec{x}) &\overset{(1)}{=} \int \frac{\mathrm{d}^d p}{(2\pi)^d} \sum_{n\neq\Omega} |c_n|^2 (2\pi)^d \delta^d(p - p_n)\, e^{-p^0\tau + i\vec{p}\cdot\vec{x}} \\
&\overset{(2)}{=} \int \frac{\mathrm{d}^d p}{(2\pi)^{d-1}}\, \theta(p^0)\rho(p^2)\, e^{-p^0\tau + i\vec{p}\cdot\vec{x}} \\
&\overset{(3)}{=} \int_0^\infty \mathrm{d}\mu^2\, \rho(\mu^2) \int \frac{\mathrm{d}^d p}{(2\pi)^{d-1}}\, \theta(p^0)\, \delta(p^2 - \mu^2)\, e^{-p^0\tau + i\vec{p}\cdot\vec{x}} \\
&\overset{(4)}{=} \int_0^\infty \mathrm{d}\mu^2\, \rho(\mu^2) \int \frac{\mathrm{d}^{d-1} p}{(2\pi)^{d-1}\, 2\omega_\mu(\vec{p})}\, e^{-\omega_\mu(\vec{p})\tau + i\vec{p}\cdot\vec{x}} \\
&\overset{(5)}{=} \int_0^\infty \mathrm{d}\mu^2\, \rho(\mu^2)\, G_\mu(\tau,\vec{x}).
\end{aligned}
\tag{13.10}
$$

(1) In each term of (13.5), write the exponential as $\int \mathrm{d}^d p\, \delta^d(p - p_n)\, e^{-p^0\tau + i\vec{p}\cdot\vec{x}}$, and move the sum inside the p integral.

(2) Definition (13.9); the factor 2π cancels one power of 2π.

(3) Insert $1 = \int_0^\infty \mathrm{d}\mu^2\, \delta(p^2 - \mu^2)$, allowed because $p^2 \geq 0$ wherever $\rho(p^2) \neq 0$, and then replace $\rho(p^2)$ by $\rho(\mu^2)$.

(4) Do the p^0 integral. With $\omega_\mu(\vec{p}) = \sqrt{\vec{p}^{\,2} + \mu^2}$, $\delta\big((p^0)^2 - \omega^2\big) = \big[\delta(p^0 - \omega) + \delta(p^0 + \omega)\big]/(2\omega)$, and $\theta(p^0)$ keeps the first term.

(5) The inner integral is the Euclidean propagator G_μ of a free field of mass μ, as we check next.

The check. The free Euclidean propagator is $G_\mu(\tau,\vec{x}) = \int \frac{\mathrm{d}^d k}{(2\pi)^d} \frac{e^{ik_0\tau + i\vec{k}\cdot\vec{x}}}{k_0^2 + \vec{k}^2 + \mu^2}$. Do the k_0 integral first, for $\tau > 0$, with $\omega = \sqrt{\vec{k}^2 + \mu^2}$:

$$
\int_{-\infty}^{\infty} \frac{\mathrm{d}k_0}{2\pi}\, \frac{e^{ik_0\tau}}{k_0^2 + \omega^2} \overset{(1)}{=} \frac{1}{2\pi}\, 2\pi i\, \mathrm{Res}_{k_0 = i\omega} \frac{e^{ik_0\tau}}{(k_0 - i\omega)(k_0 + i\omega)} \overset{(2)}{=} i\, \frac{e^{-\omega\tau}}{2i\omega} = \frac{e^{-\omega\tau}}{2\omega}.
$$

(1) For $\tau > 0$, $e^{ik_0\tau}$ decays in the upper half plane, so we close the contour there. The only pole inside is $k_0 = i\omega$.

(2) The residue is $e^{i(i\omega)\tau}/(2i\omega) = e^{-\omega\tau}/(2i\omega)$.

What remains is exactly the inner integral of step (4) with μ in place of the mass. For a free field of mass m, the only states created by ϕ are one-particle states $|\vec{q}\rangle$ with $c = 1$ and

$\sum_n \to \int \frac{\mathrm{d}^{d-1}q}{(2\pi)^{d-1}2\omega_q}$, and (13.9) gives

$$\int \frac{\mathrm{d}^{d-1}q}{(2\pi)^{d-1}2\omega_q}(2\pi)^d\delta(p^0-\omega_q)\delta^{d-1}(\vec{p}-\vec{q}) \overset{(1)}{=} \frac{2\pi\,\delta(p^0-\omega_p)}{2\omega_p} \overset{(2)}{=} 2\pi\,\theta(p^0)\,\delta(p^2-m^2),$$

(1) The $\vec{q}$ integral sets $\vec{q}=\vec{p}$.
(2) The identity used in step (4) above, read backwards.

so $\rho(\mu^2)=\delta(\mu^2-m^2)$ and (13.10) returns G_m, as it must.

Key idea: Källén–Lehmann

Any two-point function in a unitary Lorentz-invariant theory is a positive superposition of free propagators,

$$G_c(x)=\int_0^\infty \mathrm{d}\mu^2\,\rho(\mu^2)\,G_\mu(x), \qquad \rho\geq 0.$$

If $\rho(\mu^2)=0$ for $\mu<\Delta$, then G_c is a superposition of propagators each of which decays at least as $e^{-\Delta r}$ (next subsection). The gap is the bottom of the support of ρ.

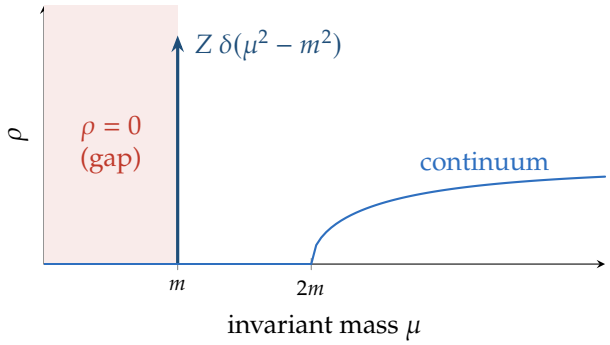


Figure 13.6: A typical spectral density in a gapped theory. Nothing below the lightest mass m (shaded). A delta function at $\mu=m$ for the one-particle state, with weight Z. A continuum from the two-particle threshold $2m$ upwards. In a gapless theory the continuum would reach down to $\mu=0$.

Try it: a gapless spectral density

Take $\rho(\mu^2)=(\mu^2)^a$ for $\mu>0$, with $a>-1$. Show, using the scaling $G_\mu(x)=\mu^{d-2}g(\mu|x|)$ of the free propagator (evident from (13.11) below), that $G_c(x)\propto|x|^{-(d+2a)}$. *Answer:* substitute $\mu=s/|x|$; then $\int \mathrm{d}\mu^2\,\mu^{2a}\mu^{d-2}g(\mu|x|)=|x|^{-(2a+d)}\int \mathrm{d}s^2\,s^{2a+d-2}g(s)$. A continuum reaching down to zero gives a power law, the hallmark of a conformal field theory (Chapter 11), where $2\Delta_\phi=d+2a$.

13.2.4 The free massive propagator: Bessel functions and Yukawa

We now compute $G_m(x)=\int \frac{\mathrm{d}^dk}{(2\pi)^d}\frac{e^{ik\cdot x}}{k^2+m^2}$ in d Euclidean dimensions, with $r=|x|$. The trick is to turn the denominator into a Gaussian.

$$G_m(x) \overset{(1)}{=} \int_0^\infty \mathrm{d}s \int \frac{\mathrm{d}^dk}{(2\pi)^d}\,e^{ik\cdot x-s(k^2+m^2)}$$
$$\overset{(2)}{=} \int_0^\infty \mathrm{d}s\; e^{-sm^2}\,(4\pi s)^{-d/2}\,e^{-r^2/(4s)}$$

$$\overset{(3)}{=} (4\pi)^{-d/2}\, m^{d-2} \int_0^\infty \mathrm{d}t\; t^{-d/2}\, e^{-t-(mr)^2/(4t)}$$
$$\overset{(4)}{=} (4\pi)^{-d/2}\, m^{d-2} \cdot 2\Big(\frac{mr}{2}\Big)^{1-d/2} K_{d/2-1}(mr)$$
$$\overset{(5)}{=} \frac{1}{(2\pi)^{d/2}} \Big(\frac{m}{r}\Big)^{d/2-1} K_{d/2-1}(mr). \tag{13.11}$$

(1) The **Schwinger parametrisation** $1/A = \int_0^\infty \mathrm{d}s\, e^{-sA}$ for $A > 0$, with $A = k^2 + m^2$.
(2) The k integral factorises into d Gaussian integrals. In each, complete the square, $ik_jx_j - sk_j^2 = -s(k_j - ix_j/2s)^2 - x_j^2/4s$, and use $\int \mathrm{d}k\, e^{-sk^2} = \sqrt{\pi/s}$, so $\int \frac{\mathrm{d}k_j}{2\pi} e^{ik_jx_j - sk_j^2} = (4\pi s)^{-1/2} e^{-x_j^2/4s}$. The product over j gives $(4\pi s)^{-d/2} e^{-r^2/4s}$.
(3) Change variable to $t = sm^2$: $\mathrm{d}s = \mathrm{d}t/m^2$, $s^{-d/2} = m^d t^{-d/2}$ and $r^2/4s = (mr)^2/4t$.
(4) The integral representation of the modified Bessel function,

$$K_\nu(z) = \tfrac{1}{2}(z/2)^\nu \int_0^\infty \mathrm{d}t\; t^{-\nu-1} e^{-t-z^2/4t},$$

with $\nu = d/2 - 1$ (so $t^{-\nu-1} = t^{-d/2}$) and $z = mr$. It says that $\int_0^\infty \mathrm{d}t\, t^{-d/2} e^{-t-z^2/4t}$ equals $2(z/2)^{-\nu} K_\nu(z)$.
(5) Collect constants: $2 \cdot 2^\nu (4\pi)^{-d/2} = 2^{d/2}\, 2^{-d}\, \pi^{-d/2} = (2\pi)^{-d/2}$, and $m^{d-2}(mr)^{-\nu} = m^\nu r^{-\nu}$.

Large distances. For $z \gg 1$ the Bessel integral is dominated by the minimum of the exponent $f(t) = t + z^2/4t$, a Laplace (saddle-point) approximation.

$$K_\nu(z) \overset{(1)}{=} \tfrac{1}{2}\big(\tfrac{z}{2}\big)^\nu \int_0^\infty \mathrm{d}t\; t^{-\nu-1} e^{-f(t)} \overset{(2)}{\approx} \tfrac{1}{2}\big(\tfrac{z}{2}\big)^\nu \big(\tfrac{z}{2}\big)^{-\nu-1} e^{-z} \sqrt{\frac{2\pi}{4/z}} \overset{(3)}{=} \sqrt{\frac{\pi}{2z}}\; e^{-z}.$$

(1) The integral representation, with $f(t) = t + z^2/4t$.
(2) $f'(t) = 1 - z^2/4t^2$ vanishes at $t_* = z/2$, where $f(t_*) = z$ and $f''(t_*) = z^2/(2t_*^3) = 4/z$. Laplace's method: $\int g(t) e^{-f(t)} \mathrm{d}t \approx g(t_*) e^{-f(t_*)} \sqrt{2\pi/f''(t_*)}$, with $g = t^{-\nu-1}$ slowly varying on the width $\sqrt{z/4}$ of the peak.
(3) $\frac{1}{2} \cdot \frac{2}{z} \cdot \sqrt{\pi z/2} = \sqrt{\pi/2z}$.

Inserting this into (13.11),

$$G_m(r) \overset{(1)}{\approx} \frac{1}{(2\pi)^{d/2}} \Big(\frac{m}{r}\Big)^{\frac{d-2}{2}} \sqrt{\frac{\pi}{2mr}}\; e^{-mr} \overset{(2)}{=} \frac{1}{2m} \Big(\frac{m}{2\pi r}\Big)^{\frac{d-1}{2}} e^{-mr}, \qquad mr \gg 1. \tag{13.12}$$

(1) The asymptotic form of K_ν.
(2) Collect powers: $m^{(d-2)/2-1/2} = m^{(d-3)/2} = m^{-1} m^{(d-1)/2}$, $r^{-(d-2)/2-1/2} = r^{-(d-1)/2}$, and $(2\pi)^{-d/2}(\pi/2)^{1/2} = \frac{1}{2}(2\pi)^{-(d-1)/2}$.

So the free propagator decays as e^{-mr} times a power $r^{-(d-1)/2}$. At short distances, $mr \ll 1$, the mass is invisible and G_m reduces to the massless power law $G_0 \propto r^{-(d-2)}$ (for $d > 2$). The crossover happens at $r \sim 1/m$ (Figure 13.7).

Worked example 13.7: three dimensions: the Yukawa form, directly

For $d = 3$, $\nu = \frac{1}{2}$ and $K_{1/2}(z) = \sqrt{\pi/2z}\, e^{-z}$ *exactly*, so (13.11) gives $G_m = e^{-mr}/(4\pi r)$ with no approximation. It is worth getting this directly, with spherical coordinates $\vec{k}$ at angle

arccos u to $\vec{x}$:

$$G_m(r) \overset{(1)}{=} \frac{1}{(2\pi)^2}\int_0^\infty \mathrm{d}k\,\frac{k^2}{k^2+m^2}\int_{-1}^{1}\mathrm{d}u\,e^{ikru} \overset{(2)}{=} \frac{1}{(2\pi)^2}\int_0^\infty \mathrm{d}k\,\frac{k^2}{k^2+m^2}\,\frac{2\sin kr}{kr}$$
$$\overset{(3)}{=} \frac{1}{4\pi^2 r}\,\mathrm{Im}\int_{-\infty}^{\infty}\mathrm{d}k\,\frac{k\,e^{ikr}}{k^2+m^2} \overset{(4)}{=} \frac{1}{4\pi^2 r}\,\mathrm{Im}\Big(2\pi i\cdot\frac{im\,e^{-mr}}{2im}\Big) = \frac{e^{-mr}}{4\pi r}.$$

(1) $\mathrm{d}^3k = 2\pi k^2\,\mathrm{d}k\,\mathrm{d}u$; the 2π from the azimuth cancels one of the three factors 2π.
(2) $\int_{-1}^{1} e^{iau}\mathrm{d}u = 2\sin a/a$.
(3) $k\sin kr/(k^2+m^2)$ is even in k, so $\int_0^\infty = \frac{1}{2}\int_{-\infty}^{\infty}$, and $\sin kr = \mathrm{Im}\,e^{ikr}$.
(4) Close the contour in the upper half plane ($r>0$). The pole at $k=im$ has residue $im\,e^{-mr}/(2im) = e^{-mr}/2$.

Worked example 13.8: two and four dimensions

$d=4$: $G_m = \frac{m}{4\pi^2 r}K_1(mr)$. For $mr \ll 1$, $K_1(z)\approx 1/z$ and $G_m \approx 1/(4\pi^2 r^2)$, the massless propagator. For $mr \gg 1$, (13.12) gives $G_m \approx \frac{\sqrt{m}}{2(2\pi r)^{3/2}}e^{-mr}$. $d=2$: $G_m = \frac{1}{2\pi}K_0(mr)$. For $mr \ll 1$, $K_0(z)\approx -\ln(z/2)-\gamma_E$ (γ_E is Euler's constant), a logarithm. The massless propagator in two dimensions is $-\frac{1}{2\pi}\ln r$, which does not even decay. This infrared disaster will force a mass to appear in Section 13.3.4.

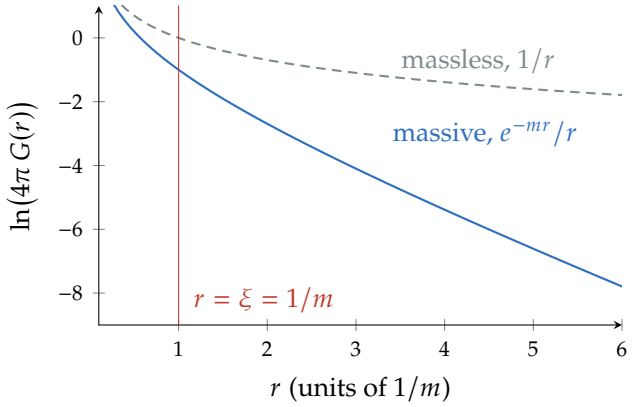


Figure 13.7: The three-dimensional propagator on a logarithmic scale. At distances below $\xi = 1/m$ the massive propagator (blue) follows the massless one (grey, dashed). Beyond ξ it falls exponentially, a straight line of slope $-m$ at large r, while the massless one decays only as a power.

Worked example 13.9: the Yukawa potential and the range of a force

Couple a scalar of mass m to static point sources through $\mathcal{L} \ni g\,\varrho(\vec{x})\phi$. The static field energy is $E[\phi] = \int \mathrm{d}^3x\big[\frac{1}{2}(\nabla\phi)^2 + \frac{1}{2}m^2\phi^2 - g\varrho\phi\big]$, and minimising it gives $(-\nabla^2+m^2)\phi = g\varrho$. For two sources $\varrho = \delta^3(\vec{x}-\vec{x}_1)+\delta^3(\vec{x}-\vec{x}_2)$ at distance r:

$$E_{\min} \overset{(1)}{=} -\tfrac{1}{2}\,g\int \mathrm{d}^3x\,\varrho\,\phi \overset{(2)}{=} -\tfrac{1}{2}\,g^2\int \mathrm{d}^3x\,\mathrm{d}^3y\;\varrho(\vec{x})\,G_m(\vec{x}-\vec{y})\,\varrho(\vec{y})$$
$$\overset{(3)}{=} (\text{self-energies}) \;-\; g^2\,G_m(r) \overset{(4)}{=} (\text{self-energies}) \;-\; \frac{g^2}{4\pi}\,\frac{e^{-mr}}{r}.$$

(1) Integrate $(\nabla\phi)^2$ by parts and use the field equation: the quadratic terms become $\frac{1}{2}g\int\varrho\phi$.

(2) The solution is $\phi = g\, G_m * \varrho$, since G_m is the Green function of $-\nabla^2 + m^2$ in three dimensions.
(3) Of the four terms, the two with $\vec{x}, \vec{y}$ at the same source do not depend on r. The two cross terms are equal.
(4) Example 13.7.

The interaction $V(r) = -\frac{g^2}{4\pi}\frac{e^{-mr}}{r}$ is attractive and has range $1/m$. For $m = 0$ it is Coulomb's (or Newton's) law. This was Yukawa's argument of 1935: the nuclear force has a range of about 1.4 fm, so it must be carried by a particle of mass about $\hbar c/1.4\,\text{fm} \approx 140\,\text{MeV}$. That particle is the pion.

13.2.5 The correlation length and the renormalisation group

The **correlation length** is the decay length of connected correlations,

$$\xi = -\lim_{r\to\infty} \frac{r}{\ln G_c(r)}.$$

By (13.12) and (13.10), in a relativistic theory $\xi = 1/m$, where m is the lightest mass that the operator can create (the power-law prefactor does not change the limit). Taking the lightest over all local operators, $\xi = 1/\Delta$.

Worked example 13.10: the Ising chain: transfer matrix, gap and ξ

The one-dimensional Ising chain at coupling K (weight $\prod_i e^{K s_i s_{i+1}}$, $s_i = \pm 1$) is a Euclidean "quantum mechanics" in which the lattice direction is imaginary time. Its **transfer matrix** is the 2×2 matrix of Boltzmann weights between neighbours,

$$T = \begin{pmatrix} e^{K} & e^{-K} \\ e^{-K} & e^{K} \end{pmatrix}, \qquad T\begin{pmatrix}1\\1\end{pmatrix} = 2\cosh K \begin{pmatrix}1\\1\end{pmatrix}, \quad T\begin{pmatrix}1\\-1\end{pmatrix} = 2\sinh K\begin{pmatrix}1\\-1\end{pmatrix}.$$

Write $T = \lambda_1 e^{-aH}$: H is a Hamiltonian with eigenvalues 0 and $\Delta = a^{-1}\ln(\lambda_1/\lambda_2)$, where a is the lattice spacing. The spin s is the matrix σ^z, which maps the eigenvector $(1,1)$ to $(1,-1)$. For an infinite chain,

$$\langle s_0 s_r\rangle \overset{(1)}{=} \frac{\langle 1|\sigma^z T^r \sigma^z|1\rangle}{\lambda_1^r} \overset{(2)}{=} |\langle 2|\sigma^z|1\rangle|^2 \Big(\frac{\lambda_2}{\lambda_1}\Big)^r \overset{(3)}{=} (\tanh K)^r \overset{(4)}{=} e^{-r\,a\Delta}.$$

(1) For a long chain, $\mathrm{Tr}(\cdots T^N)$ is dominated by the top eigenvector $|1\rangle = (1,1)/\sqrt{2}$, which plays the role of the vacuum.
(2) Insert $|2\rangle\langle 2|$ with $|2\rangle = (1,-1)/\sqrt{2}$, the only state that $\sigma^z|1\rangle$ has overlap with, since $\sigma^z|1\rangle = |2\rangle$.
(3) $|\langle 2|\sigma^z|1\rangle| = 1$ and $\lambda_2/\lambda_1 = \tanh K$.
(4) Definition of Δ.

This is (13.4) in miniature. The correlation length in lattice units is $\xi = 1/(a\Delta) = -1/\ln\tanh K$. At low temperature (large K), $\ln\tanh K \approx -2e^{-2K}$, so $\xi \approx \frac{1}{2}e^{2K}$, which is huge but finite for any $K < \infty$. Under the decimation step of Chapter T5, $\tanh K' = \tanh^2 K$, so $\xi(K') = -1/(2\ln\tanh K) = \xi(K)/2$: the correlation length measured in the new lattice units halves, as it must.

Key idea: gap, correlation length and the RG

In a lattice model the correlation length ξ is measured in lattice units, and the physical mass is $m = 1/(\xi a)$. Under coarse graining by a factor b, $\xi \mapsto \xi/b$ (Chapter T5). So a *fixed point* of the RG has $\xi = 0$ or $\xi = \infty$.

- Away from a critical point ξ is finite. Coarse graining drives $\xi \to 0$: the long-distance limit of a gapped theory has no local degrees of freedom left. It is trivial, or a topological field theory (Section 13.8.2).
- At a critical point $\xi = \infty$, and $m = 0$. Near it, $\xi \propto |t|^{-\nu}$, so the gap closes as $m \propto |t|^{\nu}$, where t is the distance from criticality.
- To define a continuum QFT with a finite mass m from a lattice, we must take $a \to 0$ at fixed m, so $\xi = 1/(ma) \to \infty$: the continuum limit is taken *at* a critical point, approached along a trajectory.

13.2.6 The Compton wavelength

Restore $\hbar$ and c. A mass m becomes an inverse length through

$$\xi = \frac{1}{m} \quad \longrightarrow \quad \lambda\!\!\!^{-}_C = \frac{\hbar}{mc},$$

the (reduced) **Compton wavelength**. So the correlation length of a gapped relativistic theory is the Compton wavelength of its lightest particle. This length has a second meaning: it is the length below which a one-particle description fails and quantum *field* theory becomes unavoidable.

Two wavelengths, two thresholds. Quantum mechanics becomes important when the **de Broglie wavelength** $\lambda_{\mathrm{dB}} = h/p$ of a particle is comparable to the size of the region it moves in. Quantum field theory becomes important when the particle is probed on distances comparable to $\lambda\!\!\!^{-}_C$. The argument is the uncertainty principle plus $E = mc^2$. Suppose a particle is localised within Δx. Then

$$\Delta p \overset{(1)}{\gtrsim} \frac{\hbar}{\Delta x}, \qquad \text{so} \qquad \Delta x < \lambda\!\!\!^{-}_C = \frac{\hbar}{mc} \overset{(2)}{\Longrightarrow} c\,\Delta p > mc^2. \tag{13.13}$$

(1) The Heisenberg uncertainty relation, up to a factor of order one.
(2) Substitute $\Delta x < \hbar/(mc)$ into $\Delta p \gtrsim \hbar/\Delta x$ and multiply by c.

A relativistic particle with momentum spread $\Delta p \gtrsim mc$ has an energy spread $c\Delta p \gtrsim mc^2$, comparable to the threshold $2mc^2$ for creating a particle–antiparticle pair. So an attempt to confine an electron to a region smaller than its Compton wavelength puts enough energy into the region to create electron–positron pairs. The "one electron" is no longer one electron, the particle number is not conserved, and the only consistent description is a field. Correlations of the field extend over $\lambda\!\!\!^{-}_C$ because virtual pairs live that long: $\Delta t \sim \hbar/(2mc^2)$, over which they travel $c\Delta t \sim \lambda\!\!\!^{-}_C/2$.

Worked example 13.11: numbers for the electron and the pion

Use $\hbar c = 197.327\,\mathrm{MeV\,fm}$ ($1\,\mathrm{fm} = 10^{-15}\,\mathrm{m}$).

$$\lambda\!\!\!^{-}_C(e) \overset{(1)}{=} \frac{\hbar c}{m_e c^2} = \frac{197.327\ \mathrm{MeV\,fm}}{0.511\ \mathrm{MeV}} \overset{(2)}{=} 386.2\ \mathrm{fm} = 3.86 \times 10^{-13}\ \mathrm{m},$$
$$\lambda\!\!\!^{-}_C(\pi^{\pm}) \overset{(1)}{=} \frac{\hbar c}{m_\pi c^2} = \frac{197.327\ \mathrm{MeV\,fm}}{139.57\ \mathrm{MeV}} \overset{(2)}{=} 1.414\ \mathrm{fm}.$$

(1) Multiply top and bottom of $\hbar/(mc)$ by c.
(2) Arithmetic. (The unreduced Compton wavelength is 2π times larger, $h/(m_e c) = 2.43 \times 10^{-12}$ m for the electron.)

The pion value is the range of the nuclear force (Example 13.9). Compare with atomic physics. The Bohr radius is $a_0 = \lambda\!\!\!^{-}_C(e)/\alpha = 386.2\,\text{fm} \times 137.04 = 5.29 \times 10^4\,\text{fm} = 0.529\,\text{Å}$, where $\alpha \approx 1/137$ is the fine-structure constant. Atoms are 137 times larger than the electron's Compton wavelength. That is why non-relativistic quantum mechanics describes them so well, and why the effects of the quantum electromagnetic field (the Lamb shift, $g-2$) are small corrections, suppressed by powers of α. By contrast, a 1 eV electron has $p = \sqrt{2m_e E} = 1.01\,\text{keV}/c$ and de Broglie wavelength $\lambda_{\text{dB}} = h/p = 1.23\,\text{nm}$: quantum mechanics matters at the nanometre scale, field theory at the scale of 10^{-13} m.

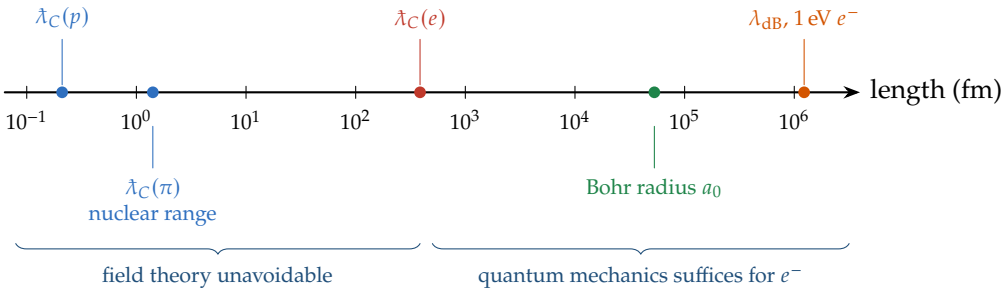


Figure 13.8: Length scales on a logarithmic axis. Each reduced Compton wavelength $\hbar/mc$ is the correlation length of the corresponding field and the range of the force it carries. For the electron, the Bohr radius sits a factor 137 above $\lambda\!\!\!^{-}_C(e)$, so atoms live in the quantum-mechanical regime. Nuclear physics lives at the pion's Compton wavelength, where pair creation of pions is always at hand.

Recap

- Inserting energy eigenstates gives the exact spectral sum $G_c = \sum_{n\neq\Omega} |c_n|^2 e^{-(E_n-E_0)|\tau|+i\vec{p}_n\cdot\vec{x}}$.
- A gap Δ bounds G_c by $e^{-\Delta|\tau|}$, and by Euclidean rotation invariance also by $e^{-\Delta r}$ at spacelike separation. Conversely, clustering of all local correlators implies a gap.
- Källén–Lehmann: $G_c = \int \mathrm{d}\mu^2 \rho(\mu^2) G_\mu$ with $\rho \geq 0$. The gap is the bottom of the support of ρ.
- The free propagator is $(2\pi)^{-d/2}(m/r)^{d/2-1}K_{d/2-1}(mr)$, decaying as $r^{-(d-1)/2}e^{-mr}$. In $d=3$ it is Yukawa's $e^{-mr}/4\pi r$.
- $\xi = 1/\Delta = \hbar/mc$: the correlation length is the Compton wavelength of the lightest particle. It is finite away from critical points and diverges at RG fixed points.
- Probing below $\lambda\!\!\!^{-}_C$ creates pairs. That is the scale where quantum field theory takes over from quantum mechanics.

13.3 Elementary ways to make a gap

A theory is gapless when it has massless excitations. Massless excitations are usually there for a reason: a symmetry forbids their mass. So opening a gap means getting around that protection. There are three elementary ways. Put in a mass by hand, where the symmetries allow it. Let a gauge field "eat" a would-be Goldstone boson (the Higgs mechanism). Or let quantum effects create a mass out of a theory that has none classically (dimensional transmutation). This section treats all three, and solves one theory of the third kind exactly at large N. The topological mechanisms of the later sections are best understood against this background.

Key idea: what keeps things massless

- *Gauge invariance* forbids a photon mass term $m^2 A_\mu A^\mu$, since it changes under $A \to A + \mathrm{d}\alpha$.
- *Chiral symmetry* forbids a fermion mass term $\bar\psi\psi$, since it couples left- and right-handed components that rotate differently.
- *Spontaneously broken continuous symmetry* gives Goldstone bosons, whose mass is forbidden by the shift $\theta \to \theta + c$ of the Goldstone field.

Every gap-generating mechanism in this chapter defeats one of these protections.

13.3.1 An explicit mass

The simplest way is to write a mass into the Lagrangian. For a real scalar in Minkowski signature, $\mathcal{L} = \frac{1}{2}\partial_\mu\phi\partial^\mu\phi - \frac{1}{2}m^2\phi^2$, the field equation and a plane wave give

$$(\partial_t^2 - \nabla^2 + m^2)\phi = 0, \quad \phi = e^{-i\omega t + i\vec{k}\cdot\vec{x}} \overset{(1)}{\Longrightarrow} -\omega^2 + \vec{k}^2 + m^2 = 0 \overset{(2)}{\Longrightarrow} \omega = \sqrt{\vec{k}^2 + m^2} \geq m.$$

(1) $\partial_t \to -i\omega$ and $\nabla \to i\vec{k}$ on the plane wave.

(2) Take the positive root. The minimum over $\vec{k}$ is at $\vec{k} = 0$.

This dispersion relation, quantised, gives the gapped spectrum of Example 13.1. A Dirac mass $m\bar\psi\psi$ does the same for fermions.

Worked example 13.12: the Proca field: a vector with an explicit mass

Take $\mathcal{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \frac{1}{2}m^2 A_\mu A^\mu$ in 3+1 dimensions. The field equation is $\partial_\mu F^{\mu\nu} + m^2 A^\nu = 0$. Then

$$0 \overset{(1)}{=} \partial_\nu\partial_\mu F^{\mu\nu} + m^2\partial_\nu A^\nu \overset{(2)}{=} m^2\,\partial_\nu A^\nu,$$
$$0 \overset{(3)}{=} \partial_\mu\partial^\mu A^\nu - \partial^\nu\partial_\mu A^\mu + m^2 A^\nu \overset{(4)}{=} (\Box + m^2)A^\nu.$$

(1) Take ∂_ν of the field equation.

(2) $\partial_\nu\partial_\mu$ is symmetric and $F^{\mu\nu}$ antisymmetric, so the first term vanishes. Hence $\partial_\nu A^\nu = 0$ for $m \neq 0$.

(3) Write out $F^{\mu\nu} = \partial^\mu A^\nu - \partial^\nu A^\mu$ in the field equation.

(4) Use $\partial_\mu A^\mu = 0$ from the first line.

Each component obeys the Klein–Gordon equation, and the constraint $\partial \cdot A = 0$ removes one of the four components, leaving three polarisations of mass m. The price is gauge invariance: $m^2 A_\mu A^\mu$ is not invariant under $A \to A + \partial\alpha$. For the gluons of Yang–Mills theory such a term would destroy the consistency of the quantum theory, so gauge bosons need a subtler mechanism.

Worked example 13.13: a small explicit breaking gives a small gap

Take a complex scalar with a U(1)-symmetric Mexican-hat potential, whose minima form the circle $|\phi| = v/\sqrt{2}$. Write $\phi = \frac{v}{\sqrt{2}} e^{i\theta}$ on the circle. The kinetic term becomes $|\partial\phi|^2 = \frac{v^2}{2}(\partial\theta)^2$, and θ is a massless Goldstone boson. Now tilt the hat with a small term $-\epsilon \operatorname{Re}\phi$, $\epsilon > 0$. On the circle of minima,

$$-\epsilon \operatorname{Re}\phi \overset{(1)}{=} -\frac{\epsilon v}{\sqrt{2}} \cos\theta \overset{(2)}{=} -\frac{\epsilon v}{\sqrt{2}} + \frac{\epsilon v}{2\sqrt{2}}\,\theta^2 + O(\theta^4).$$

(1) $\operatorname{Re} e^{i\theta} = \cos\theta$.
(2) $\cos\theta = 1 - \theta^2/2 + \ldots$

Compare with $\frac{v^2}{2}\left[(\partial\theta)^2 - m_\theta^2\theta^2\right]$ (the canonical form for the field $v\theta$). Matching the θ^2 terms, $\frac{v^2}{2}m_\theta^2 = \frac{\epsilon v}{2\sqrt{2}}$, so $m_\theta^2 = \epsilon/(\sqrt{2}\,v)$. The gap grows like $\sqrt{\epsilon}$. The same mechanism gives the pions their mass, $m_\pi^2 \propto m_u + m_d$ (the Gell-Mann–Oakes–Renner relation): the quark masses tilt the hat of broken chiral symmetry.

13.3.2 The Higgs mechanism

Gauge the U(1) of the previous example. The **abelian Higgs model** is

$$\mathcal{L} = -\tfrac{1}{4}F_{\mu\nu}F^{\mu\nu} + |D_\mu\phi|^2 - \lambda\left(|\phi|^2 - \tfrac{v^2}{2}\right)^2, \qquad D_\mu = \partial_\mu - ieA_\mu,$$

in Minkowski signature, with gauge coupling e written explicitly. It is invariant under $\phi \to e^{ie\alpha}\phi$, $A_\mu \to A_\mu + \partial_\mu\alpha$. Without the gauge field, the phase of ϕ would be a Goldstone boson (Figure 13.9). With it, the phase can be removed by a gauge transformation: choose α so that ϕ is real and positive everywhere (the **unitary gauge**), and write $\phi = (v + h)/\sqrt{2}$ with h a real field. Then

$$\begin{aligned} |D_\mu\phi|^2 &\overset{(1)}{=} \tfrac{1}{2}\left|\partial_\mu h - ieA_\mu(v+h)\right|^2 \overset{(2)}{=} \tfrac{1}{2}(\partial_\mu h)^2 + \tfrac{1}{2}e^2(v+h)^2 A_\mu A^\mu \\ &= \tfrac{1}{2}(\partial_\mu h)^2 + \tfrac{1}{2}\,e^2v^2\,A_\mu A^\mu + e^2vh\,A_\mu A^\mu + \tfrac{1}{2}e^2h^2 A_\mu A^\mu, \end{aligned} \tag{13.14}$$

$$\lambda\left(|\phi|^2 - \tfrac{v^2}{2}\right)^2 \overset{(3)}{=} \lambda\left(vh + \tfrac{1}{2}h^2\right)^2 = \lambda v^2h^2 + \lambda vh^3 + \tfrac{1}{4}\lambda h^4. \tag{13.15}$$

(1) Insert $\phi = (v + h)/\sqrt{2}$, which is real.
(2) $|a - ib|^2 = a^2 + b^2$ for real $a = \partial_\mu h$, $b = eA_\mu(v+h)$ (with the Lorentz contraction understood).
(3) $|\phi|^2 - v^2/2 = \frac{1}{2}(v+h)^2 - \frac{1}{2}v^2 = vh + \frac{1}{2}h^2$.

Read off the quadratic terms. $\frac{1}{2}e^2v^2A_\mu A^\mu$ is a Proca mass term with $m_A^2 = e^2v^2$. $\lambda v^2h^2 = \frac{1}{2}m_h^2h^2$ gives $m_h^2 = 2\lambda v^2$. So

$$m_A = ev, \qquad m_h = \sqrt{2\lambda}\,v. \tag{13.16}$$

Count degrees of freedom in 3+1 dimensions. Before: a massless photon (two polarisations) and a complex scalar (two real fields), four in all. After: a massive vector (three polarisations) and one real scalar h, again four. The Goldstone boson has become the longitudinal polarisation of the massive photon. Every excitation is massive: *the Higgs phase is gapped*, with $\Delta = \min(m_A, m_h)$.

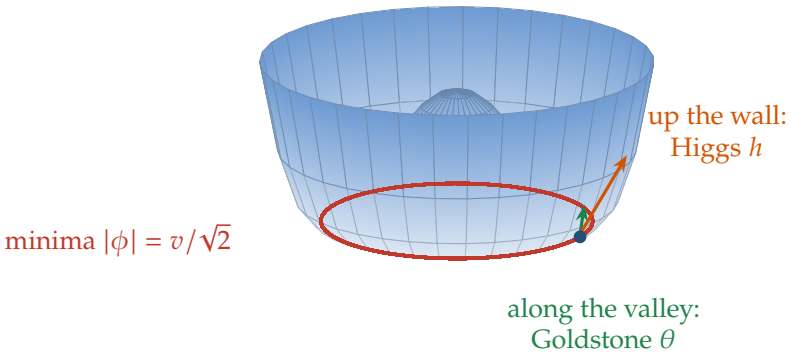


Figure 13.9: The Mexican-hat potential over the complex ϕ plane. The vacua form a circle (red). Moving up the wall costs energy, so the radial mode h is massive. Moving along the valley costs none, so the angular mode is a massless Goldstone boson. When the U(1) is gauged, motion along the valley is a gauge transformation, the Goldstone mode is absorbed by the photon, and the photon becomes massive.

Worked example 13.14: the Meissner effect: a gap seen as screening

A superconductor is a Higgs phase, with ϕ the Cooper-pair field. A static magnetic field inside obeys the Proca equation with $A_0 = 0$. In Coulomb gauge $\nabla \cdot \vec{A} = 0$ it reads $\nabla^2 \vec{A} = m_A^2 \vec{A}$, and taking the curl, $\nabla^2 \vec{B} = m_A^2 \vec{B}$. For a field $\vec{B} = B(x)\hat{\vec{z}}$ entering a superconductor filling $x > 0$,

$$B''(x) = m_A^2 B(x) \overset{(1)}{\Longrightarrow} B(x) = B_0\, e^{-m_A x} + B_1 e^{m_A x} \overset{(2)}{\Longrightarrow} B(x) = B_0\, e^{-x/\lambda_L}, \quad \lambda_L = \frac{1}{m_A}.$$

(1) General solution of a linear ODE with constant coefficients.
(2) B must stay bounded deep inside, so $B_1 = 0$.

The magnetic field is expelled, except for a skin of thickness equal to the Compton wavelength of the massive photon, the **London penetration depth**. For a typical superconductor $\lambda_L \approx 50\,\text{nm}$, so the photon mass inside is $\hbar c/\lambda_L \approx 197\,\text{MeV fm}/(5\times 10^7\,\text{fm}) \approx 4\,\text{eV}$. This is Section 13.2 in action: a gap makes the static response decay exponentially.

Watch out: the Standard Model is not gapped

In the electroweak theory the Higgs mechanism gives the W and Z bosons masses of 80.4 and 91.2 GeV, but it leaves one combination of gauge fields, the photon, massless. So the Standard Model, like QED, is gapless. Also, a "broken gauge symmetry" is a figure of speech: gauge transformations are redundancies of the description and cannot break. The gauge-invariant statement is that the Higgs phase is gapped and screens charges.

13.3.3 Dimensional transmutation

Pure Yang–Mills theory in four dimensions has one parameter, the coupling g, and g is dimensionless. Classically there is no mass anywhere: gluons are massless waves, and the theory is scale invariant. How can it have a gap? A mass needs a scale.

The scale comes from quantum mechanics. To define the quantum theory we must regularise it at some scale, and the coupling then depends on the scale μ at which it is measured

(Chapter T5). Write the running as

$$\mu \frac{\mathrm{d}g^2}{\mathrm{d}\mu} = -\beta_0\, g^4 + O(g^6), \tag{13.17}$$

where $\beta_0 > 0$ means **asymptotic freedom**: g shrinks at high energies and grows at low ones. For SU(N) Yang–Mills, $\beta_0 = \frac{11N}{24\pi^2}$. (The usual form $\mu\,\mathrm{d}g/\mathrm{d}\mu = -\frac{b_0 g^3}{16\pi^2}$ with $b_0 = \frac{11N}{3}$ gives this, since $\mu\,\mathrm{d}g^2/\mathrm{d}\mu = 2g\,\mu\,\mathrm{d}g/\mathrm{d}\mu$.) Solve the leading-order equation:

$$\mu \frac{\mathrm{d}}{\mathrm{d}\mu}\left(\frac{1}{g^2}\right) \overset{(1)}{=} -\frac{1}{g^4}\,\mu\frac{\mathrm{d}g^2}{\mathrm{d}\mu} \overset{(2)}{=} \beta_0 \quad \overset{(3)}{\Longrightarrow} \quad \frac{1}{g^2(\mu)} = \beta_0 \ln\frac{\mu}{\Lambda}. \tag{13.18}$$

(1) Chain rule, $\mathrm{d}(1/u) = -\mathrm{d}u/u^2$ with $u = g^2$.
(2) (13.17) at leading order.
(3) Integrate in $\ln\mu$. The integration constant is written as $-\beta_0 \ln\Lambda$, which defines the scale Λ.

Solving for Λ,

$$\Lambda = \mu\, e^{-1/(\beta_0 g^2(\mu))}. \tag{13.19}$$

The right side contains μ and $g(\mu)$, but the combination does not depend on μ:

$$\mu\frac{\mathrm{d}\Lambda}{\mathrm{d}\mu} \overset{(1)}{=} \Lambda + \mu\, e^{-1/(\beta_0 g^2)}\cdot\left(-\frac{1}{\beta_0}\right)\mu\frac{\mathrm{d}}{\mathrm{d}\mu}\frac{1}{g^2} \overset{(2)}{=} \Lambda - \frac{\beta_0}{\beta_0}\Lambda = 0.$$

(1) Product rule and chain rule on (13.19).
(2) $\mu\,\mathrm{d}(1/g^2)/\mathrm{d}\mu = \beta_0$ by (13.18).

Λ is a renormalisation-group invariant mass. It is the *only* mass in pure Yang–Mills theory, so by dimensional analysis every physical mass is a pure number times Λ:

$$m = c\,\Lambda = c\,\mu\, e^{-1/(\beta_0 g^2(\mu))}. \tag{13.20}$$

A dimensionless coupling has been traded for a mass. This is **dimensional transmutation**.

Worked example 13.15: the gap is invisible in perturbation theory

Let $f(x) = e^{-1/x}$ for $x > 0$, where $x = \beta_0 g^2$. Every derivative of f has the form $f^{(n)}(x) = P_n(1/x)\,e^{-1/x}$ with P_n a polynomial. (Induction: $\frac{\mathrm{d}}{\mathrm{d}x}[P(1/x)e^{-1/x}] = [-x^{-2}P'(1/x) + x^{-2}P(1/x)]e^{-1/x}$, again a polynomial in $1/x$ times $e^{-1/x}$.) With $y = 1/x$,

$$\lim_{x\to 0^+} f^{(n)}(x) \overset{(1)}{=} \lim_{y\to\infty} P_n(y)\,e^{-y} \overset{(2)}{=} 0.$$

(1) Substitute $y = 1/x \to \infty$.
(2) An exponential beats any polynomial.

So the Taylor series of $m(g)$ around $g = 0$ is identically zero, to every order. No finite order of perturbation theory in g can see the mass gap. This is why the gap of Yang–Mills theory is a genuinely non-perturbative problem (Figure 13.10).

Physical picture: instantons and powers of Λ

A Yang–Mills instanton (Chapter 8) has action $8\pi^2/g^2$, so it contributes with weight

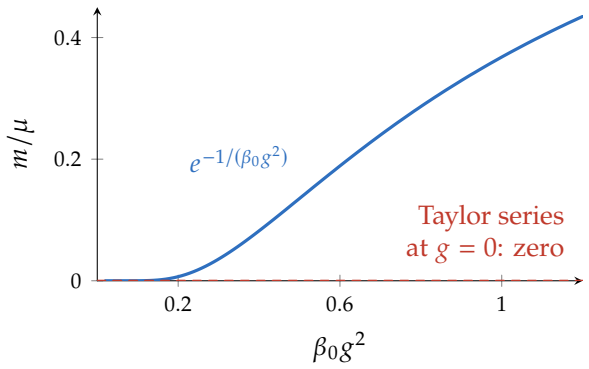


Figure 13.10: The dynamically generated mass as a function of the coupling. The curve leaves zero so flatly that all its derivatives vanish at $g = 0$: the Taylor series (red, dashed) is identically zero. A perturbative expansion in g therefore predicts massless gluons to all orders, although the true theory is (believed to be) gapped.

$e^{-8\pi^2/g^2(\mu)}$. With $1/g^2 = \frac{b_0}{8\pi^2}\ln(\mu/\Lambda)$,

$$e^{-8\pi^2/g^2(\mu)} = e^{-b_0\ln(\mu/\Lambda)} = \left(\frac{\Lambda}{\mu}\right)^{b_0}.$$

Instanton effects are powers of Λ, the same non-analytic scale. This is suggestive, but it does not mean instantons *cause* the gap. The next example shows a theory where they demonstrably do not, while Section 13.4 shows one where they do.

13.3.4 A solvable example: the $\mathbb{CP}^{N-1}$ model at large N

We want a theory that is classically massless, asymptotically free, and in which the gap can actually be computed. The two-dimensional $\mathbb{CP}^{N-1}$ model is such a theory, at large N.

The model. Take N complex fields $z_a(x)$, $a = 1, \dots, N$, on two-dimensional Euclidean space, subject to the constraint $\sum_a |z_a|^2 = N/g^2$ (a sphere S^{2N-1} in $\mathbb{C}^N$). Identify $z_a \sim e^{i\alpha(x)} z_a$ (a local U(1)). The field lives on $S^{2N-1}/\mathrm{U}(1) = \mathbb{CP}^{N-1}$, the space of complex lines in $\mathbb{C}^N$ (for $N = 2$, $\mathbb{CP}^1 = S^2$). The action is

$$S = \int \mathrm{d}^2x \sum_{a=1}^{N} |D_\mu z_a|^2, \qquad D_\mu = \partial_\mu - iA_\mu, \tag{13.21}$$

where A_μ is an auxiliary gauge field with no kinetic term. (Its field equation sets $A_\mu = \frac{i}{2}\frac{\bar{z}\partial_\mu z - \partial_\mu \bar{z}\, z}{\bar{z}z}$, which makes the action invariant under the local phase.) The coupling g is dimensionless in two dimensions, since $[z] = 0$.

Classically gapless. Pick a point on $\mathbb{CP}^{N-1}$, say $z = (\sqrt{N}/g, 0, \dots, 0)$. It breaks the global SU(N) symmetry, and the fluctuations along $\mathbb{CP}^{N-1}$ are $N-1$ massless complex Goldstone fields. But in two dimensions this picture cannot survive. The fluctuation of a massless field at a point is $\langle \delta z^2 \rangle \propto \int \frac{\mathrm{d}^2p}{(2\pi)^2}\frac{1}{p^2} = \frac{1}{2\pi}\int \frac{\mathrm{d}p}{p}$, which diverges logarithmically at small p. The field wanders all over the target space, and the would-be order is destroyed (the **Mermin–Wagner–Coleman theorem**: no spontaneous breaking of a continuous symmetry in two dimensions). Something must give. At large N we can see exactly what.

The Lagrange multiplier. Impose the constraint with a field $\lambda(x)$:

$$Z = \int \mathcal{D}z\,\mathcal{D}\bar{z}\,\mathcal{D}A\,\mathcal{D}\lambda\ \exp\Big[-\int \mathrm{d}^2x\Big(\sum_a |D_\mu z_a|^2 + i\lambda\Big(\sum_a |z_a|^2 - \frac{N}{g^2}\Big)\Big)\Big].$$

(The λ integral of $e^{-i\lambda X}$ is a delta function $\delta(X)$.) Now z appears only quadratically, $\int \bar{z}\,\mathcal{O}\,z$ with $\mathcal{O} = -D^2 + i\lambda$. For one complex field $\int \mathcal{D}z\mathcal{D}\bar{z}\,e^{-\int \bar{z}\mathcal{O}z} \propto (\det\mathcal{O})^{-1}$, the complex Gaussian integral. With N fields,

$$Z \overset{(1)}{=} \int \mathcal{D}A\,\mathcal{D}\lambda\,(\det\mathcal{O})^{-N}\,e^{\,i\frac{N}{g^2}\int\lambda} \overset{(2)}{=} \int \mathcal{D}A\,\mathcal{D}\lambda\ e^{-S_{\text{eff}}},$$
$$S_{\text{eff}} = N\Big[\operatorname{Tr}\ln\big(-D^2 + i\lambda\big) - \frac{i}{g^2}\int \mathrm{d}^2x\,\lambda\Big]. \tag{13.22}$$

(1) The N complex Gaussian integrals, one per z_a, each giving $(\det\mathcal{O})^{-1}$.
(2) $(\det\mathcal{O})^{-N} = e^{-N\ln\det\mathcal{O}}$ and $\ln\det = \operatorname{Tr}\ln$.

The whole effective action is proportional to N. As $N\to\infty$ the remaining integral over A and λ is dominated by its saddle point, exactly as $\int \mathrm{d}x\,e^{-Nf(x)}$ is dominated by the minimum of f. (The fluctuations around the saddle are suppressed by $1/N$.)

The gap equation. Look for a saddle with $A_\mu = 0$ and constant $i\lambda = m^2$. (The saddle lies at imaginary λ. The contour of the λ integral can be deformed to pass through it.) For constant fields Tr ln is the volume V times a momentum integral, so $S_{\text{eff}} = V\,\mathcal{V}(m^2)$ with

$$\frac{\mathcal{V}(m^2)}{N} = \int^{\Lambda} \frac{\mathrm{d}^2p}{(2\pi)^2}\,\ln(p^2+m^2) - \frac{m^2}{g^2}, \tag{13.23}$$

with an ultraviolet cutoff Λ on $|p|$. The saddle point condition is $\mathrm{d}\mathcal{V}/\mathrm{d}m^2 = 0$:

$$\frac{1}{g^2} \overset{(1)}{=} \int^{\Lambda} \frac{\mathrm{d}^2p}{(2\pi)^2}\,\frac{1}{p^2+m^2} \overset{(2)}{=} \frac{1}{2\pi}\int_0^{\Lambda} \frac{p\,\mathrm{d}p}{p^2+m^2} \overset{(3)}{=} \frac{1}{4\pi}\ln\Big(1+\frac{\Lambda^2}{m^2}\Big). \tag{13.24}$$

(1) Differentiate (13.23) under the integral sign.
(2) Polar coordinates, $\mathrm{d}^2p = 2\pi p\,\mathrm{d}p$.
(3) Substitute $u = p^2$: $\int_0^{\Lambda^2} \frac{\mathrm{d}u/2}{u+m^2} = \frac{1}{2}\ln\frac{\Lambda^2+m^2}{m^2}$.

This is the **gap equation**. It has a direct meaning. The constraint says $\langle\sum_a |z_a|^2\rangle = N/g^2$. At the saddle, each z_a is a free field of mass m, so $\langle|z_a|^2\rangle = G_m(0) = \int \frac{\mathrm{d}^2p}{(2\pi)^2}\frac{1}{p^2+m^2}$, and (13.24) is exactly the constraint. With $m = 0$ the right side would diverge at small p: the constraint can be satisfied *only* with a mass. Solving,

$$m \overset{(1)}{=} \frac{\Lambda}{\sqrt{e^{4\pi/g^2}-1}} \overset{(2)}{\approx} \Lambda\,e^{-2\pi/g^2}. \tag{13.25}$$

(1) Exponentiate (13.24): $1 + \Lambda^2/m^2 = e^{4\pi/g^2}$.
(2) For small g, $e^{4\pi/g^2} \gg 1$.

This is (13.20) with $c = 1$, $\mu = \Lambda$ and $\beta_0 = \frac{1}{2\pi}$. Indeed, holding m fixed while changing the cutoff requires $1/g^2(\Lambda) = \frac{1}{2\pi}\ln(\Lambda/m)$, which is (13.18) with $\beta_0 = 1/2\pi$: the $\mathbb{CP}^{N-1}$ model is asymptotically free.

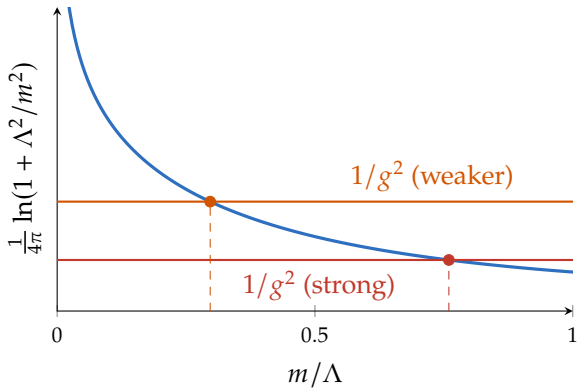


Figure 13.11: Graphical solution of the gap equation (13.24). The blue curve is the fluctuation $\langle|z|^2\rangle$ per field for fields of mass m, which grows without bound as $m \to 0$ because of the infrared logarithm. The horizontal lines are the value $1/g^2$ that the constraint demands. They always cross the curve, so a mass is always generated. Weaker coupling (higher line) gives an exponentially smaller mass.

What the gapped theory looks like. At the saddle, all N fields z_a have the same mass m: the SU(N) symmetry is unbroken, and there are no Goldstone bosons, as Mermin–Wagner demanded. One step beyond the saddle, the loop of z fields generates a kinetic term $\frac{N}{48\pi m^2}F_{\mu\nu}F_{\mu\nu}$ for the gauge field A_μ, so A becomes a genuine two-dimensional photon. In two dimensions the Coulomb potential between charges is linear (the field lines cannot spread), so the z quanta are *confined*. The physical particles are gauge-invariant $\bar{z}z$ bound states, with masses near $2m$. The theory is gapped.

Worked example 13.16: numbers, and why instantons are not the cause

For $g^2 = 2$, (13.25) gives $m = \Lambda/\sqrt{e^{2\pi}-1} = 0.043\,\Lambda$. For $g^2 = 1$ it gives $m = \Lambda/\sqrt{e^{4\pi}-1} = 0.0019\,\Lambda$. Halving g^2 cut the mass by a factor of more than twenty. The $\mathbb{CP}^{N-1}$ model has instantons, labelled by $\pi_2(\mathbb{CP}^{N-1}) = \mathbb{Z}$, with action $S_{\text{inst}} = 2\pi N/g^2$ in our normalisation. So

$$e^{-S_{\text{inst}}} \overset{(1)}{=} \left(e^{-2\pi/g^2}\right)^N \overset{(2)}{=} \left(\frac{m}{\Lambda}\right)^N.$$

(1) $S_{\text{inst}} = N \cdot 2\pi/g^2$.
(2) (13.25).

At large N instanton effects are suppressed by the N-th power of the tiny number m/Λ, while the gap is of order m itself. The gap of the $\mathbb{CP}^{N-1}$ model is *not* caused by instantons: it comes from the collective effect of the N fluctuating fields, the infrared logarithm.

13.3.5 A fermionic example: the Gross–Neveu model

The same logic works for fermions, and shows a gapped phase with two vacua. Take N massless two-component Dirac fermions ψ_i in two Euclidean dimensions, with the attractive four-fermion interaction

$$\mathcal{L}_E = \sum_i \bar{\psi}_i \partial\!\!\!/ \psi_i - \frac{g^2}{2N}\Big(\sum_i \bar{\psi}_i \psi_i\Big)^2.$$

Use the Euclidean gamma matrices $\gamma_1 = \sigma^1$, $\gamma_2 = \sigma^2$ (Hermitian, $\{\gamma_\mu, \gamma_\nu\} = 2\delta_{\mu\nu}$) and $\gamma_5 = \sigma^3$ (Chapter T4). The model has a discrete chiral symmetry $\psi \to \gamma_5\psi$, $\bar{\psi} \to -\bar{\psi}\gamma_5$, under which

$\bar\psi\psi \to -\bar\psi\psi$. It forbids a mass term.

Introduce an auxiliary field σ: $\mathcal{L}_E = \bar\psi_i(\not\partial+\sigma)\psi_i + \frac{N}{2g^2}\sigma^2$. (Completing the square, $\sigma = -\frac{g^2}{N}\bar\psi\psi$, returns the original Lagrangian.) The chiral symmetry now acts as $\sigma \to -\sigma$. Integrate out the fermions: the Grassmann Gaussian integral gives $\det(\not\partial + \sigma)^N$, with a *positive* power. For constant σ, in momentum space $\not\partial \to i\not p$, and the 2×2 determinant is $\det(i\not p + \sigma) = \sigma^2 + p^2$ (the eigenvalues of the Hermitian matrix $\not p$ are $\pm|p|$). The effective potential per unit volume is

$$\frac{\mathcal{V}(\sigma)}{N} = -\int^{\Lambda} \frac{\mathrm{d}^2p}{(2\pi)^2}\,\ln(p^2+\sigma^2) + \frac{\sigma^2}{2g^2}. \tag{13.26}$$

Its extrema:

$$\frac{1}{N}\frac{\mathrm{d}\mathcal{V}}{\mathrm{d}\sigma} \overset{(1)}{=} -\int^{\Lambda} \frac{\mathrm{d}^2p}{(2\pi)^2}\frac{2\sigma}{p^2+\sigma^2} + \frac{\sigma}{g^2} \overset{(2)}{=} \sigma\Big[\frac{1}{g^2} - \frac{1}{2\pi}\ln\Big(1+\frac{\Lambda^2}{\sigma^2}\Big)\Big]. \tag{13.27}$$

(1) Differentiate under the integral.
(2) The integral is twice the one in (13.24).

So either $\sigma = 0$ or $\frac{1}{g^2} = \frac{1}{2\pi}\ln(1+\Lambda^2/\sigma^2)$, that is

$$\sigma = \pm m, \qquad m = \frac{\Lambda}{\sqrt{e^{2\pi/g^2}-1}} \approx \Lambda\, e^{-\pi/g^2}.$$

Which is the minimum? The second derivative at $\sigma = 0$ contains $-2\int \frac{\mathrm{d}^2p}{(2\pi)^2}\frac{1}{p^2}$, which is $-\infty$ from the infrared. So $\sigma = 0$ is a maximum, and the minima are at $\sigma = \pm m$ (Figure 13.12). In either vacuum the fermions have mass m: the gap is generated dynamically, with $\beta_0 = 1/\pi$. The chiral $\mathbb{Z}_2$ is spontaneously broken, so there are *two* vacua, and in two dimensions there are also kinks interpolating between them. This is a gapped phase of the second kind in Section 13.1.4.

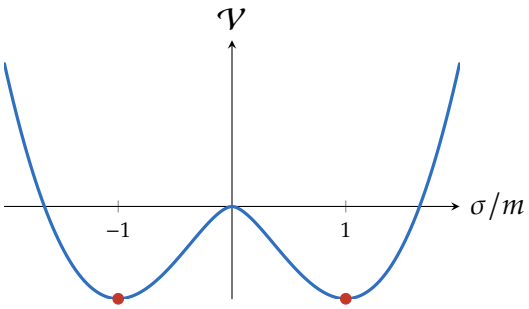


Figure 13.12: The large-N effective potential of the Gross–Neveu model, $\mathcal{V} \propto \sigma^2(\ln(\sigma^2/m^2) - 1)$ (from (13.26) after removing a constant). The symmetric point $\sigma = 0$ is a maximum, for any coupling. The two minima $\sigma = \pm m$ are the two vacua, related by the broken chiral symmetry, and in each the fermions have mass m.

Try it: the potential in closed form

Using $\int_0^{\Lambda^2}\ln(u+\sigma^2)\mathrm{d}u = (\Lambda^2+\sigma^2)\ln(\Lambda^2+\sigma^2) - \Lambda^2 - \sigma^2\ln\sigma^2$ and expanding for $\Lambda \gg \sigma$, show that $\mathcal{V}/N = \frac{\sigma^2}{4\pi}\big[\ln\frac{\sigma^2}{\Lambda^2} - 1\big] + \frac{\sigma^2}{2g^2} + \text{const} = \frac{\sigma^2}{4\pi}\big[\ln\frac{\sigma^2}{m^2} - 1\big] + \text{const}$, with $m = \Lambda e^{-\pi/g^2}$. *Answer:* the Λ dependence of the first form is absorbed exactly by $1/g^2 = \frac{1}{2\pi}\ln(\Lambda/m)^2$, which is the renormalisation of the coupling.

13.3.6 Three dimensions: a gap only beyond a critical coupling

In two dimensions the infrared logarithm forced a gap at any coupling. In three dimensions the massless fluctuation integral $\int \frac{\mathrm{d}^3p}{(2\pi)^3}\frac{1}{p^2}$ converges at small p, and the story changes. Repeat the $\mathbb{CP}^{N-1}$ (or O(N)) calculation in $d = 3$, where g^2 now has dimensions of length. The gap equation is

$$\frac{1}{g^2} \overset{(1)}{=} \int^{\Lambda} \frac{\mathrm{d}^3p}{(2\pi)^3}\frac{1}{p^2+m^2} \overset{(2)}{=} \frac{1}{2\pi^2}\int_0^{\Lambda} \mathrm{d}p\left(1-\frac{m^2}{p^2+m^2}\right) \overset{(3)}{=} \frac{1}{2\pi^2}\left(\Lambda - m\arctan\frac{\Lambda}{m}\right) \overset{(4)}{\approx} \frac{\Lambda}{2\pi^2} - \frac{m}{4\pi}. \tag{13.28}$$

(1) The same saddle-point condition as (13.24), in three dimensions.

(2) $\mathrm{d}^3p = 4\pi p^2\mathrm{d}p$, and $\frac{p^2}{p^2+m^2} = 1 - \frac{m^2}{p^2+m^2}$.

(3) $\int_0^{\Lambda} \frac{\mathrm{d}p}{p^2+m^2} = \frac{1}{m}\arctan\frac{\Lambda}{m}$.

(4) $\arctan(\Lambda/m) \to \pi/2$ for $\Lambda \gg m$.

Define the critical coupling by $1/g_c^2 = \Lambda/(2\pi^2)$. Then

$$m = 4\pi\left(\frac{1}{g_c^2} - \frac{1}{g^2}\right), \tag{13.29}$$

which is positive only for $g > g_c$. For $g < g_c$ the gap equation has no solution with $m > 0$: the constraint is satisfied by a condensate, the symmetry is broken, and the theory has Goldstone bosons. So there are two phases (Figure 13.13): a gapless ordered phase at weak coupling and a gapped symmetric phase at strong coupling, separated by a critical point at $g = g_c$. Near it, $m \propto (g - g_c)$, so the correlation length $\xi = 1/m \propto (g - g_c)^{-\nu}$ with $\nu = 1$. The critical point is an RG fixed point with $\xi = \infty$ (Chapter T5), and the large-N value $\nu = 1$ is the $N \to \infty$ limit of the O(N) Wilson–Fisher exponents.

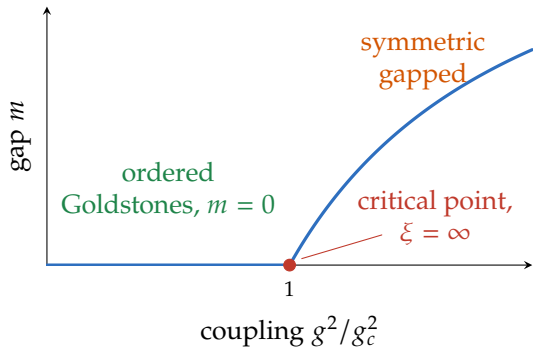


Figure 13.13: The large-N sigma model in three dimensions. Below the critical coupling the symmetry is broken and Goldstone bosons make the theory gapless. Above it the gap (13.29) rises linearly from zero. At the critical point the correlation length diverges: this is an RG fixed point, a conformal field theory.

Recap

- Masslessness is protected by gauge invariance, chiral symmetry or Goldstone's theorem. Opening a gap means evading the protection.
- Explicit masses work where symmetry allows. A small explicit breaking gives a

small gap (pions).

- The Higgs mechanism: the gauge field eats the Goldstone boson, with $m_A = ev$ and $m_h = \sqrt{2\lambda}v$. The Higgs phase is gapped and screens magnetic fields over $1/m_A$.
- Dimensional transmutation: $m = c\,\mu\, e^{-1/(\beta_0 g^2(\mu))}$, an RG invariant that vanishes to all orders in perturbation theory.
- At large N, the $\mathbb{CP}^{N-1}$ gap equation $1/g^2 = \int \frac{\mathrm{d}^2p}{(2\pi)^2}\frac{1}{p^2+m^2}$ forces $m = \Lambda e^{-2\pi/g^2}$ through the infrared logarithm. Gross–Neveu gives $m = \Lambda e^{-\pi/g^2}$ with two vacua. In three dimensions a gap appears only beyond a critical coupling, with $\nu = 1$.

Sources. The large-N solutions of the $\mathbb{CP}^{N-1}$ and Gross–Neveu models follow D. Tong, *Gauge Theory* (lecture notes, Cambridge), §7.3 and §7.4.1, which reproduce the original analyses of D'Adda, Lüscher and Di Vecchia and of Witten (1979) and of Gross and Neveu (1974).

13.4 A gap from instantons: Polyakov's compact QED in 2+1 dimensions

Maxwell theory in $2+1$ dimensions has one photon polarisation, and it is massless. So pure electromagnetism in the plane is gapless, and charges feel a potential that grows only logarithmically with distance. In 1977 Polyakov showed that this changes completely when the gauge group is the *compact* group U(1), for example when electromagnetism arises from SU(2) broken by a Higgs field. Then magnetic monopoles exist as finite-action configurations of Euclidean spacetime. They are instantons (Chapter 8). Summing over a dilute gas of them turns the massless photon into a massive one, with mass proportional to $e^{-S_0/2}$, where S_0 is the monopole action, and turns the logarithmic potential into a linear one. The gap and confinement both come from a topological charge: the degree of a map $S^2 \to S^2$.

This is also the prototype of what Jaffe and Witten call the need for a "non-classical change of variables or duality transformation" to see a mass gap. Here the duality is explicit, and the whole calculation can be done.

13.4.1 Maxwell theory in three dimensions and the dual photon

Work in three-dimensional Euclidean space with action

$$S = \frac{1}{4e^2}\int \mathrm{d}^3x\, F_{\mu\nu}F_{\mu\nu}. \tag{13.30}$$

The coupling is in front of the action, so a charge-one particle couples as $e^{i\oint A}$, and e^2 has dimensions of mass (since $[A]=1$ and $[\mathrm{d}^3x]=-3$). In three dimensions an antisymmetric $F_{\mu\nu}$ has three components, the same number as a vector. Package them as the **dual field strength**

$$B_\mu = \tfrac{1}{2}\epsilon_{\mu\nu\rho}F_{\nu\rho}, \qquad F_{\mu\nu}F_{\mu\nu} = 2B_\mu B_\mu.$$

The Bianchi identity $\mathrm{d}F = 0$ reads $\partial_\mu B_\mu = 0$: the lines of B never end. A magnetic monopole of charge q at a point X is exactly a place where they do end, with total flux $2\pi q$ (Dirac quantisation for charge-one matter):

$$\partial_\mu B_\mu = 2\pi\sum_i q_i\,\delta^3(x - X_i), \qquad q_i \in \mathbb{Z}. \tag{13.31}$$

Worked example 13.17: Maxwell theory is a free scalar in disguise

Without monopoles, integrate over unconstrained fields B_μ and impose $\partial \cdot B = 0$ with a Lagrange multiplier $\sigma(x)$, using $\int \mathcal{D}\sigma\, e^{\frac{i}{2\pi}\int \sigma\, \partial\cdot B} \propto \delta[\partial \cdot B]$:

$$Z = \int \mathcal{D}B\, \mathcal{D}\sigma\ \exp\Big[-\int \mathrm{d}^3x \Big(\frac{1}{2e^2} B_\mu B_\mu + \frac{i}{2\pi} B_\mu \partial_\mu \sigma\Big)\Big],$$

after integrating $\frac{i}{2\pi}\int \sigma\, \partial \cdot B$ by parts. Complete the square in B:

$$\frac{1}{2e^2} B^2 + \frac{i}{2\pi} B \cdot \partial\sigma \overset{(1)}{=} \frac{1}{2e^2}\Big(B + \frac{ie^2}{2\pi}\partial\sigma\Big)^2 + \frac{e^2}{8\pi^2}(\partial\sigma)^2.$$

(1) Expand the square: $\frac{1}{2e^2}\big(B^2 + \frac{ie^2}{\pi} B\cdot\partial\sigma - \frac{e^4}{4\pi^2}(\partial\sigma)^2\big) = \frac{B^2}{2e^2} + \frac{i}{2\pi} B\cdot\partial\sigma - \frac{e^2}{8\pi^2}(\partial\sigma)^2$, so adding back $\frac{e^2}{8\pi^2}(\partial\sigma)^2$ reproduces the left side.

The Gaussian integral over the shifted B is a constant, so

$$Z \propto \int \mathcal{D}\sigma\ \exp\Big[-\frac{e^2}{8\pi^2}\int \mathrm{d}^3x\, (\partial_\mu\sigma)^2\Big], \qquad B_\mu = -\frac{ie^2}{2\pi}\,\partial_\mu\sigma \text{ (on shell)}. \tag{13.32}$$

The field σ is the **dual photon**. It is a free massless scalar: one degree of freedom, as a $2+1$ dimensional photon should have. Why is σ periodic, $\sigma \sim \sigma + 2\pi$? Because with monopoles present, (13.31) turns the multiplier term into $e^{\frac{i}{2\pi}\int \sigma\,\partial\cdot B} = \prod_i e^{iq_i\sigma(X_i)}$, which is unchanged by $\sigma \to \sigma + 2\pi$ exactly when all $q_i \in \mathbb{Z}$.

The free action (13.32) is invariant under the shift $\sigma \to \sigma + c$. This is a genuine global symmetry, called $\mathrm{U}(1)_{\text{top}}$ ("topological" or magnetic), whose conserved current is $J_\mu = \frac{1}{2\pi} B_\mu$ and whose charge is the magnetic flux through space divided by 2π. In the Coulomb phase the vacuum picks a value of σ, so $\mathrm{U}(1)_{\text{top}}$ is spontaneously broken and the photon is its Goldstone boson. That is the symmetry reason the $2+1$ dimensional photon is massless. To give it a mass, something must break $\mathrm{U}(1)_{\text{top}}$ explicitly. The insertion $e^{iq\sigma(X)}$ does exactly that: it shifts by e^{iqc}. So *monopoles are the only thing that can gap the photon*, and the question is whether they occur in the theory.

13.4.2 Monopole-instantons

In pure Maxwell theory with gauge group $\mathbb{R}$ there are no monopoles. They appear when the U(1) is compact and the short-distance theory can smooth out the Dirac singularity. The cleanest example is the three-dimensional **Georgi–Glashow model**: SU(2) gauge theory with a real scalar Φ^a in the adjoint representation, whose potential forces $|\Phi| = v$ at low energies. The vacuum value $\Phi^a = v\,\delta^{a3}$ breaks SU(2) to the U(1) of rotations about the 3-axis. The charged gauge bosons $W^\pm$ get mass $m_W = ev$, and below m_W we are left with compact U(1) electromagnetism.

The topology. A configuration with finite Euclidean action must have $|\Phi| \to v$ on the sphere at infinity, S^2_∞. So $\Phi/|\Phi|$ restricted to S^2_∞ is a map $S^2_\infty \to S^2$ (the vacuum manifold $\mathrm{SU}(2)/\mathrm{U}(1) = S^2$). Such maps are classified by their degree, $\pi_2(S^2) = \mathbb{Z}$, the number of times the image wraps the target sphere. The degree equals the magnetic charge q of the unbroken U(1). The **'t Hooft–Polyakov monopole** is the smooth solution with $q = 1$: the hedgehog $\Phi^a \to v\, x^a/|x|$ (Figure 13.14).

In $3+1$ dimensions this configuration, extended in time, is a particle. In three Euclidean

dimensions the same configuration is localised at a *point* of spacetime: it is an event, an instanton. What does it do? Its flux $2\pi q$ leaves every sphere around it. A spatial slice below the event and a spatial slice above it therefore carry magnetic fluxes differing by $2\pi q$: the monopole-instanton is a tunnelling event that changes the magnetic flux through space by one quantum. Its action is

$$S_0 = \frac{4\pi m_W}{e^2}\,\varepsilon\!\left(\frac{m_H}{m_W}\right), \tag{13.33}$$

where m_H is the mass of the neutral Higgs particle and ε is a slowly varying function of order one ($\varepsilon = 1$ in the limit $m_H \to 0$, where the Bogomolny bound is saturated, rising to about 1.8 as $m_H \to \infty$). The theory is weakly coupled when $e^2 \ll m_W$, and then $S_0 \gg 1$.

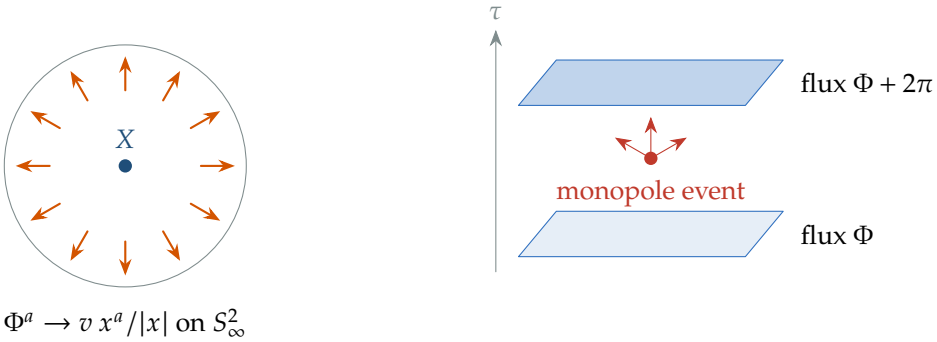


Figure 13.14: Left: the 't Hooft–Polyakov hedgehog. On the sphere at infinity the Higgs field points radially, a map $S^2 \to S^2$ of degree one. Right: in three-dimensional Euclidean spacetime the same configuration is an event (red). The magnetic flux through the spatial slice after the event exceeds the flux before it by one quantum 2π: the monopole is an instanton that tunnels between flux sectors.

Worked example 13.18: monopoles interact through a three-dimensional Coulomb law

Far from its core, a monopole of charge q at X has $B_\mu = \frac{q}{2}\frac{(x-X)_\mu}{|x-X|^3} = -\frac{q}{2}\partial_\mu \frac{1}{|x-X|}$, whose flux through any sphere around X is $\frac{q}{2}\cdot 4\pi = 2\pi q$, as required. For two monopoles the action $\frac{1}{2e^2}\int (B_1+B_2)^2$ contains the self-actions and a cross term:

$$S_{12} \overset{(1)}{=} \frac{1}{e^2}\int \mathrm{d}^3x\, B_1\cdot B_2 \overset{(2)}{=} \frac{q_1q_2}{4e^2}\int \mathrm{d}^3x\, \partial_\mu \frac{1}{|x-X_1|}\,\partial_\mu \frac{1}{|x-X_2|}$$
$$\overset{(3)}{=} -\frac{q_1q_2}{4e^2}\int \mathrm{d}^3x\, \frac{1}{|x-X_1|}\,\nabla^2 \frac{1}{|x-X_2|} \overset{(4)}{=} \frac{\pi\, q_1q_2}{e^2\,|X_1-X_2|}.$$

(1) The cross term of $\frac{1}{2e^2}(B_1+B_2)^2$ is $\frac{1}{e^2}B_1\cdot B_2$.
(2) The far-field form above, for both monopoles.
(3) Integrate by parts. The surface term vanishes because the integrand falls like $1/|x|^3$.
(4) $\nabla^2 \frac{1}{|x-X_2|} = -4\pi\delta^3(x-X_2)$.

Like charges repel and opposite charges attract, with a $1/r$ law: the monopole-instantons form a three-dimensional **Coulomb gas**. The same result comes from the dual photon. The action (13.32) is $\frac{1}{2}\kappa(\partial\sigma)^2$ with $\kappa = e^2/4\pi^2$, so the propagator is $\langle\sigma(x)\sigma(0)\rangle = \frac{1}{\kappa}\frac{1}{4\pi|x|} = \frac{\pi}{e^2|x|}$ (Example 13.7 with $m=0$). For a Gaussian field, $\langle e^{i\sum_i q_i\sigma(X_i)}\rangle = e^{-\frac{1}{2}\sum_{i,j} q_iq_j\langle\sigma(X_i)\sigma(X_j)\rangle}$, and the $i\neq j$ terms give $e^{-\sum_{i<j}\pi q_iq_j/(e^2|X_i-X_j|)} = e^{-\sum_{i<j}S_{ij}}$. Inserting $e^{iq\sigma(X)}$ *is* inserting a monopole.

13.4.3 The dilute gas and the sine-Gordon model

At weak coupling the monopoles are rare and far apart, so the path integral is a sum over configurations of N_+ monopoles and N_- anti-monopoles ($q = \pm 1$) at arbitrary positions. Each one carries the weight e^{-S_0} times a factor K from the Gaussian fluctuations around it (a one-loop determinant with the zero modes traded for the position integrals, exactly as for the instantons of Chapter 8). K has dimensions of (mass)^3, so that $K\,\mathrm{d}^3X$ is dimensionless. Write $\zeta = Ke^{-S_0}$ for the **fugacity**. The $1/N_\pm!$ removes the overcounting of identical objects. Then

$$
\begin{aligned}
Z &\overset{(1)}{=} \sum_{N_+,N_-} \frac{\zeta^{N_++N_-}}{N_+!\,N_-!} \int \prod_i \mathrm{d}^3X_i \; e^{-\sum_{i<j} \pi q_i q_j/(e^2|X_i-X_j|)} \\
&\overset{(2)}{=} \int \mathcal{D}\sigma \, e^{-\frac{e^2}{8\pi^2}\int(\partial\sigma)^2} \sum_{N_+,N_-} \frac{1}{N_+!\,N_-!} \Big(\zeta \int \mathrm{d}^3X \, e^{i\sigma(X)}\Big)^{N_+} \Big(\zeta \int \mathrm{d}^3X \, e^{-i\sigma(X)}\Big)^{N_-} \\
&\overset{(3)}{=} \int \mathcal{D}\sigma \; \exp\Big[-\int \mathrm{d}^3x \, \frac{e^2}{8\pi^2}(\partial\sigma)^2 + \zeta \int \mathrm{d}^3x \, \big(e^{i\sigma} + e^{-i\sigma}\big)\Big] \\
&\overset{(4)}{=} \int \mathcal{D}\sigma \; \exp\Big[-\int \mathrm{d}^3x \Big(\frac{e^2}{8\pi^2}(\partial\sigma)^2 - 2\zeta\cos\sigma\Big)\Big].
\end{aligned}
\tag{13.34}
$$

(1) The dilute monopole gas, with the Coulomb interaction of Example 13.18 (the monopole self-energies are part of S_0).
(2) Example 13.18: the Coulomb factor is the Gaussian average of $\prod_i e^{iq_i\sigma(X_i)}$ over the dual photon.
(3) $\sum_N \frac{x^N}{N!} = e^x$, once for N_+ and once for N_-.
(4) $e^{i\sigma} + e^{-i\sigma} = 2\cos\sigma$.

The Coulomb gas of monopole-instantons is equivalent to the three-dimensional **sine-Gordon model**. This is the duality Jaffe and Witten cite for the rigorous proof of Debye screening in a Coulomb gas.

The photon mass. The potential $-2\zeta\cos\sigma$ has minima at $\sigma \in 2\pi\mathbb{Z}$ (Figure 13.15). Expand about $\sigma = 0$:

$$
\frac{e^2}{8\pi^2}(\partial\sigma)^2 - 2\zeta\cos\sigma \overset{(1)}{=} -2\zeta + \frac{e^2}{8\pi^2}(\partial\sigma)^2 + \zeta\sigma^2 + O(\sigma^4) \overset{(2)}{=} -2\zeta + \frac{e^2}{8\pi^2}\Big[(\partial\sigma)^2 + \frac{8\pi^2\zeta}{e^2}\,\sigma^2\Big] + \dots
$$

(1) $\cos\sigma = 1 - \sigma^2/2 + O(\sigma^4)$.
(2) Factor out the coefficient of the kinetic term.

Comparing with $\frac{\kappa}{2}[(\partial\sigma)^2 + m^2\sigma^2]$, the dual photon has mass

$$
m_\gamma^2 = \frac{8\pi^2\zeta}{e^2} = \frac{8\pi^2 K}{e^2}\, e^{-S_0}, \qquad m_\gamma \propto e^{-S_0/2}. \tag{13.35}
$$

The photon has a mass, and pure compact U(1) gauge theory in $2+1$ dimensions is gapped. The mass is exponentially small at weak coupling, $m_\gamma \ll e^2 \ll m_W$, which is what justifies using the low-energy dual photon theory with the cosine as a small perturbation.

Physical picture: Debye screening by a plasma of instantons

In an ordinary plasma, mobile charges screen an electric field over the Debye length. Here the "plasma" lives in Euclidean spacetime, and its particles are monopole-instantons, with density $n = \langle N_+ + N_-\rangle/V = 2\zeta$ at leading order. They screen *magnetic* field. The Debye

mass of a Coulomb gas is (interaction strength)×(density). Our interaction is $\frac{\pi}{e^2 r} = \frac{4\pi^2}{e^2}\cdot\frac{1}{4\pi r}$, of strength $4\pi^2/e^2$, and indeed $m_\gamma^2 = \frac{4\pi^2}{e^2}\cdot 2\zeta$. The photon mass is the screening mass of the instanton plasma.

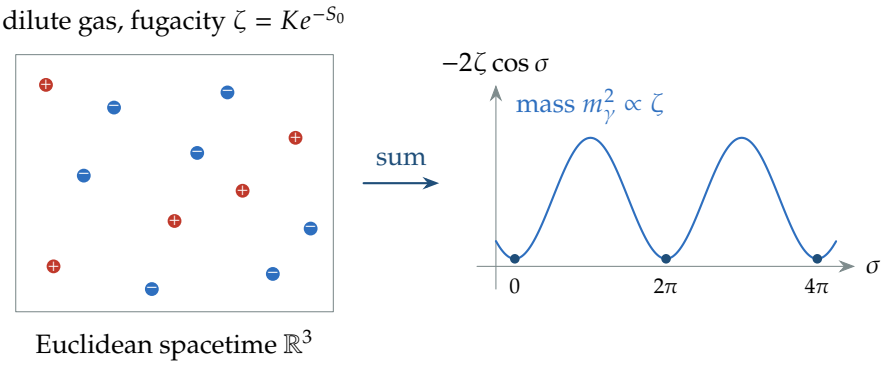


Figure 13.15: Left: a typical configuration in the path integral of compact QED in $2+1$ dimensions, a dilute gas of monopoles (+) and anti-monopoles (−) scattered through Euclidean spacetime, interacting by a $1/r$ Coulomb law. Right: summing over the gas produces a periodic potential for the dual photon. Its minima are the vacua $\sigma \in 2\pi\mathbb{Z}$, and its curvature there is the photon mass squared.

13.4.4 Confinement

A mass gap is not the same thing as confinement, but here both follow from the same potential. Rotate back to Minkowski signature, where the on-shell relation of (13.32) becomes real: $\frac{1}{2}\epsilon^{\mu\nu\rho}F_{\nu\rho} = \frac{e^2}{2\pi}\partial^\mu\sigma$. In components the electric field is $E_i = F_{0i}$, and $E_i = \frac{e^2}{2\pi}\epsilon_{ij}\partial_j\sigma$ (up to an overall sign convention). So an electric field is a *gradient* of σ at right angles to it.

Before the monopoles. The field of two opposite static charges is the two-dimensional Coulomb field: its lines spread out over the plane, and the energy grows as $\frac{e^2}{2\pi}\ln r$ (log confinement). In σ language this is a smooth gradient spread over all of space.

After the monopoles. A spread-out gradient of σ now costs potential energy $-2\zeta\cos\sigma$ everywhere it moves σ off a minimum. The cheaper option is to keep σ at a minimum almost everywhere and let it jump by 2π, from one minimum to the next, across a thin wall. Across such a **kink** the electric flux is

$$\int E_\perp \,\mathrm{d}\ell \overset{(1)}{=} \frac{e^2}{2\pi}\int \partial_\ell\sigma \,\mathrm{d}\ell \overset{(2)}{=} \frac{e^2}{2\pi}\cdot 2\pi = e^2,$$

(1) $E = \frac{e^2}{2\pi}\epsilon\cdot\nabla\sigma$: the flux through a line crossing the wall is the change of σ along the line.
(2) σ changes by 2π across the kink.

which is exactly the flux of one unit of charge (Gauss's law from (13.30) is $\nabla\cdot\vec{E} = e^2\rho$). So the electric flux between the charges is squeezed into a string of width $\sim 1/m_\gamma$: a **flux tube** (Figure 13.16). A string of fixed tension T (energy per unit length) stretched between the charges gives the linear potential $V(r) = T\,r$. Charges are confined.

Worked example 13.19: the string tension

The tension is the minimum energy per unit length of a kink in which σ goes from 0 to 2π along a coordinate y. With $A = e^2/8\pi^2$ and $U(\sigma) = 2\zeta(1-\cos\sigma) \geq 0$ (the potential shifted to vanish at the minima),

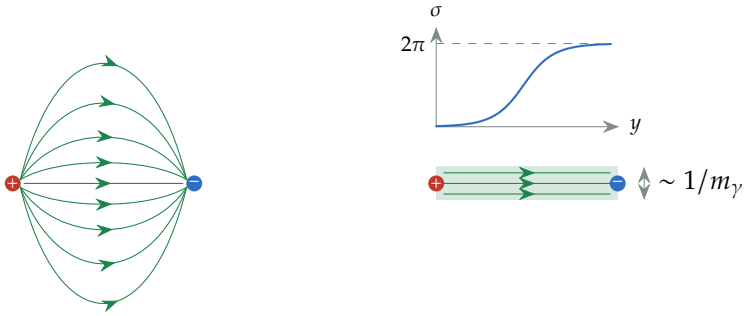
$$T \overset{(1)}{=} \int \mathrm{d}y \left[A\,\sigma'^2 + U(\sigma)\right] \overset{(2)}{=} \int \mathrm{d}y \left[\left(\sqrt{A}\,\sigma' - \sqrt{U}\right)^2 + 2\sqrt{AU}\,\sigma'\right] \overset{(3)}{\geq} 2\sqrt{A}\int_0^{2\pi} \sqrt{U(\sigma)}\,\mathrm{d}\sigma$$
$$\overset{(4)}{=} 2\sqrt{2A\zeta}\int_0^{2\pi} \sqrt{1-\cos\sigma}\,\mathrm{d}\sigma \overset{(5)}{=} 2\sqrt{2A\zeta}\cdot\sqrt{2}\int_0^{2\pi} \sin\frac{\sigma}{2}\,\mathrm{d}\sigma \overset{(6)}{=} 16\sqrt{A\zeta}.$$

(1) Energy per unit length of a static configuration depending on y.
(2) Complete the square: $(a-b)^2 + 2ab = a^2 + b^2$ with $a = \sqrt{A}\sigma'$, $b = \sqrt{U}$.
(3) Drop the non-negative square. The remaining term is $\int 2\sqrt{AU}\,\frac{\mathrm{d}\sigma}{\mathrm{d}y}\mathrm{d}y = \int_0^{2\pi} 2\sqrt{AU}\,\mathrm{d}\sigma$. Equality holds for the kink obeying $\sqrt{A}\sigma' = \sqrt{U}$ (the Bogomolny trick).
(4) $U = 2\zeta(1-\cos\sigma)$.
(5) $1-\cos\sigma = 2\sin^2(\sigma/2)$, and $\sin(\sigma/2) \geq 0$ on $[0, 2\pi]$.
(6) $\int_0^{2\pi} \sin(\sigma/2)\mathrm{d}\sigma = 4$, and $2\sqrt{2}\cdot\sqrt{2}\cdot 4 = 16$.

With $A = e^2/8\pi^2$ and $\zeta = e^2 m_\gamma^2/8\pi^2$ from (13.35),

$$T = 16\sqrt{\frac{e^2}{8\pi^2}\cdot\frac{e^2 m_\gamma^2}{8\pi^2}} = \frac{2\,e^2\,m_\gamma}{\pi^2}.$$

The string tension is proportional to the photon mass: both vanish when the monopoles are switched off. For a large Wilson loop of area $\mathcal{A}$, the same wall spans the loop in Euclidean spacetime, and $\langle W(C)\rangle \sim e^{-T\mathcal{A}}$, the **area law**.



no monopoles: $V \propto \ln r$ monopole plasma: $V = T\,r$

Figure 13.16: Left: without monopoles, the electric field lines of two opposite charges spread over the whole plane, and the energy grows only logarithmically. Right: in the monopole plasma the dual photon σ must sit at a minimum of its cosine potential almost everywhere. It jumps by 2π across a thin wall (inset), and the wall carries exactly one unit of electric flux. The flux is confined to a tube of width about $1/m_\gamma$, and the energy grows linearly with separation.

Try it: how small is the gap?

Take the BPS limit $\varepsilon = 1$ and $e^2/m_W = 0.5$. Compute S_0 and the suppression factor $e^{-S_0/2}$. *Answer:* $S_0 = 4\pi m_W/e^2 = 8\pi \approx 25.1$, so $e^{-S_0/2} = e^{-12.6} \approx 3.5\times 10^{-6}$. The photon mass, and the string tension with it, are some six orders of magnitude below the W mass: an

enormous hierarchy produced by a topological effect.

13.4.5 What was topological, and what is known rigorously

Key idea: Polyakov's mechanism in four lines

1. Non-compact $2+1$ dimensional QED is gapless: the photon is the Goldstone boson of the magnetic symmetry $\mathrm{U}(1)_{\mathrm{top}}$, $\sigma \to \sigma + c$.
2. Compactness of U(1) allows finite-action monopole-instantons, labelled by the degree $q \in \pi_2(S^2) = \mathbb{Z}$.
3. Each monopole inserts $e^{\pm i\sigma}$, which breaks $\mathrm{U}(1)_{\mathrm{top}}$ explicitly. The dilute gas generates $-2\zeta \cos\sigma$, and $m_\gamma^2 = 8\pi^2 K e^{-S_0}/e^2$.
4. The same cosine squeezes electric flux into tubes of tension $T = 2e^2 m_\gamma/\pi^2$: linear confinement.

Watch out: the mechanism is fragile

Three things can switch it off. (i) In $3+1$ dimensions monopoles are particles, not instantons, and compact QED_4 keeps a massless photon (a Coulomb phase) at weak coupling. (ii) Massless charged fermions give each monopole-instanton fermion zero modes (Chapter 4), and then $e^{i\sigma}$ alone has vanishing expectation value; QED_3 with enough massless fermions is believed to be gapless. (iii) At high temperature the Euclidean time direction becomes a small circle, the monopole gas turns into a gas of monopole–anti-monopole pairs, and confinement is lost. The topological charge is always there, but whether it opens a gap depends on dimension, matter content and temperature.

Dictionary: compact QED_3 and a superconductor

superconductor (Example 13.14)	compact QED_3
Cooper pairs condense	monopoles proliferate
photon gets a mass	dual photon gets a mass
magnetic flux expelled or squeezed into tubes	electric flux squeezed into tubes
magnetic charges would be confined	electric charges are confined

Compact QED_3 is a *dual superconductor*: the roles of electric and magnetic are exchanged. The 't Hooft–Mandelstam picture of confinement in four-dimensional Yang–Mills theory is the conjecture that the QCD vacuum is also a dual superconductor, with magnetic monopoles condensing. In four dimensions no one has made this into a controlled calculation for pure Yang–Mills theory. It has been made precise in supersymmetric theories (Seiberg and Witten, 1994), where monopoles become light and their condensation can be followed exactly.

Rigorous status. For the lattice version of compact U(1) gauge theory in three dimensions (with the Villain action), Göpfert and Mack proved in 1982 that static charges are confined and that there is a mass gap, for all values of the coupling. Polyakov's model is therefore one of the few gauge theories in which a gap generated by topology is established mathematically, though on the lattice rather than in the continuum.

Reading Nash (Ch. VIII, §§4–5, pp. 242–257)

Nash treats monopoles as static solutions of the Yang–Mills–Higgs equations in 3 + 1 dimensions: their topological charge, the Bogomolny equations, and their moduli spaces. Everything he says about a single monopole applies unchanged here, because a static monopole in $\mathbb{R}^3$ and a monopole-instanton in Euclidean $\mathbb{R}^3$ are the same field configuration. What changes is its interpretation: here the third coordinate is Euclidean time, the monopole is an event, and the physical question is what a gas of such events does to the spectrum.

Recap

- In 2 + 1 dimensions Maxwell theory is dual to a free periodic scalar, the dual photon, $S = \frac{e^2}{8\pi^2}\int(\partial\sigma)^2$.
- Compact U(1) has monopole-instantons with action $S_0 = 4\pi m_W \varepsilon/e^2$, which insert $e^{\pm i\sigma}$ and interact by a three-dimensional Coulomb law.
- The dilute gas sums to the sine-Gordon model, giving $m_\gamma^2 = 8\pi^2 K e^{-S_0}/e^2$: a gap from instantons, the Debye mass of a plasma of topological events.
- Electric flux is carried by kinks of σ with tension $T = 2e^2 m_\gamma/\pi^2$: linear confinement and an area law.

Sources. A. M. Polyakov, *Quark confinement and topology of gauge groups*, Nucl. Phys. B120 (1977) 429, reproduced in D. Tong, *Gauge Theory*, §8.1 and §8.3 (whose normalisation of the monopole charge, $\pm 4\pi$, differs from our ± 1, which is why his potential reads $\cos 2\sigma$). M. Göpfert and G. Mack, Commun. Math. Phys. 82 (1982) 545.

13.5 A gap from a topological term: the Chern–Simons photon

The Higgs mechanism gives a photon a mass by adding a charged scalar that condenses. Polyakov's mechanism needs a compact gauge group and a gas of instantons. In 2 + 1 dimensions there is a third way, which needs neither extra fields nor instantons. Add to the Maxwell action the Chern–Simons term, a term that knows nothing about the metric. The photon becomes massive, gauge invariance is untouched, and the long-distance theory is a topological quantum field theory (Chapter 12). This is the **topologically massive gauge theory** of Deser, Jackiw and Templeton (1982).

13.5.1 The Chern–Simons term

For an abelian gauge field on a three-manifold M, the Chern–Simons term is

$$S_{\mathrm{CS}} = \frac{k}{4\pi}\int_M A\wedge \mathrm{d}A = \frac{k}{4\pi}\int \mathrm{d}^3x\;\epsilon^{\mu\nu\rho}A_\mu\partial_\nu A_\rho, \tag{13.36}$$

with $\epsilon^{012} = 1$. It has three properties we need.

No metric. The integrand is a 3-form built from A and $\mathrm{d}A$ alone. Neither the wedge product nor d involves the metric (Chapter T1). So S_{CS} is unchanged by any change of the metric, and its stress tensor $T_{\mu\nu} \propto \delta S/\delta g^{\mu\nu}$ vanishes. A pure Chern–Simons theory has zero Hamiltonian: it has no energy to give to anything. That is what "topological" means here.

Gauge invariance up to a boundary. Under $A \to A + \mathrm{d}\alpha$,

$$\delta\big(A \wedge \mathrm{d}A\big) \overset{(1)}{=} \mathrm{d}\alpha \wedge \mathrm{d}A \overset{(2)}{=} \mathrm{d}(\alpha\, \mathrm{d}A),$$

(1) $\mathrm{d}(A + \mathrm{d}\alpha) = \mathrm{d}A$ since $\mathrm{d}^2 = 0$, so only the first factor changes.
(2) $\mathrm{d}(\alpha\, \mathrm{d}A) = \mathrm{d}\alpha \wedge \mathrm{d}A + \alpha\, \mathrm{d}^2 A$ and $\mathrm{d}^2 A = 0$.

So $\delta S_{\mathrm{CS}} = \frac{k}{4\pi}\int_{\partial M} \alpha\, \mathrm{d}A$ by Stokes, which vanishes when M has no boundary and α is single valued. For large gauge transformations on nontrivial manifolds, $e^{iS_{\mathrm{CS}}}$ is invariant only if the **level** k is an integer (Chapter 12). We write $\kappa = k/2\pi$, so the term is $\frac{\kappa}{2}\epsilon^{\mu\nu\rho}A_\mu \partial_\nu A_\rho$.

Odd under parity and time reversal. The ϵ tensor changes sign under a reflection of one coordinate. So a Chern–Simons term can only appear in a theory that breaks parity and time reversal. This is the price of the mechanism.

Working definition 13.3 (topological mass)

A $2+1$ dimensional gauge field has a **topological mass** when its action contains, besides the Maxwell term $-\frac{1}{4e^2}F^2$, a Chern–Simons term $\frac{k}{4\pi}\int A \wedge \mathrm{d}A$ [a term built from A and $\mathrm{d}A$ only, with no metric, so it contributes no energy of its own]. The photon then has mass $m = |k|e^2/2\pi$ [the only mass that can be made from e^2, which carries units of mass in $2+1$ dimensions, and the pure number k], while gauge invariance is kept [the action changes only by a boundary term under $A \to A + \mathrm{d}\alpha$] and no field condenses [there is no Higgs scalar and no vacuum expectation value]. *To test for it:* the photon propagator has a single pole at $p^2 = m^2$, the static field of a charge Q decays as e^{-mr}, and the charge carries magnetic flux $2\pi Q/k$.

13.5.2 The field equation and the mass

Use Minkowski signature $\eta = \mathrm{diag}(+1,-1,-1)$, so that $\epsilon_{012} = +1$ as well. The **Maxwell–Chern–Simons** Lagrangian is

$$\mathcal{L} = -\frac{1}{4e^2}F_{\mu\nu}F^{\mu\nu} + \frac{\kappa}{2}\,\epsilon^{\mu\nu\rho}A_\mu \partial_\nu A_\rho. \tag{13.37}$$

Vary A. For the Maxwell term, $\delta(-\frac{1}{4e^2}F^2) = -\frac{1}{e^2}F^{\mu\nu}\partial_\mu \delta A_\nu$, which after integration by parts is $\frac{1}{e^2}(\partial_\mu F^{\mu\nu})\delta A_\nu$. For the Chern–Simons term,

$$\delta\Big(\frac{\kappa}{2}\epsilon^{\mu\nu\rho}A_\mu\partial_\nu A_\rho\Big) \overset{(1)}{=} \frac{\kappa}{2}\epsilon^{\mu\nu\rho}\big(\delta A_\mu\, \partial_\nu A_\rho + A_\mu\, \partial_\nu \delta A_\rho\big) \overset{(2)}{=} \frac{\kappa}{2}\epsilon^{\mu\nu\rho}\delta A_\mu\, \partial_\nu A_\rho - \frac{\kappa}{2}\epsilon^{\mu\nu\rho}\partial_\nu A_\mu\, \delta A_\rho$$
$$\overset{(3)}{=} \kappa\, \epsilon^{\mu\nu\rho}\delta A_\mu\, \partial_\nu A_\rho \overset{(4)}{=} \kappa\, \tilde{F}^\mu\, \delta A_\mu, \qquad \tilde{F}^\mu := \tfrac{1}{2}\epsilon^{\mu\nu\rho}F_{\nu\rho}.$$

(1) Product rule.
(2) Integrate the second term by parts (dropping a total derivative).
(3) In the second term rename $\rho \to \mu \to \nu \to \rho$ cyclically, using $\epsilon^{\rho\mu\nu} = \epsilon^{\mu\nu\rho}$, and then swap $\mu \leftrightarrow \nu$ inside $\partial_\nu A_\mu$, which costs a sign that cancels the minus. Both terms become equal.
(4) $\epsilon^{\mu\nu\rho}\partial_\nu A_\rho = \frac{1}{2}\epsilon^{\mu\nu\rho}(\partial_\nu A_\rho - \partial_\rho A_\nu) = \frac{1}{2}\epsilon^{\mu\nu\rho}F_{\nu\rho}$.

The field equation is therefore

$$\frac{1}{e^2}\partial_\mu F^{\mu\nu} + \kappa\, \tilde{F}^\nu = 0. \tag{13.38}$$

The dual field $\tilde{F}^\mu$ is gauge invariant and obeys the Bianchi identity $\partial_\mu \tilde{F}^\mu = \frac{1}{2}\epsilon^{\mu\nu\rho}\partial_\mu F_{\nu\rho} = 0$. In $2+1$ dimensions $\tilde{F}$ contains all of F: $F_{\nu\rho} = \epsilon_{\nu\rho\sigma}\tilde{F}^\sigma$ (contract $\tilde{F}^\mu = \frac{1}{2}\epsilon^{\mu\nu\rho}F_{\nu\rho}$ with $\epsilon_{\alpha\beta\mu}$ and use

$\epsilon_{\alpha\beta\mu}\epsilon^{\mu\nu\rho} = \delta^\nu_\alpha\delta^\rho_\beta - \delta^\rho_\alpha\delta^\nu_\beta$). Write the equation in terms of $\tilde{F}$ alone. With $m \coloneqq \kappa e^2$,

$$\frac{1}{e^2}\partial_\mu F^{\mu\nu} \overset{(1)}{=} \frac{1}{e^2}\epsilon^{\mu\nu\sigma}\partial_\mu\tilde{F}_\sigma \overset{(2)}{=} -\frac{1}{e^2}\epsilon^{\nu\mu\sigma}\partial_\mu\tilde{F}_\sigma \quad \overset{(3)}{\Longrightarrow} \quad \epsilon^{\nu\mu\sigma}\partial_\mu\tilde{F}_\sigma = m\,\tilde{F}^\nu. \tag{13.39}$$

(1) $F^{\mu\nu} = \epsilon^{\mu\nu\sigma}\tilde{F}_\sigma$ (raise the indices of $F_{\nu\rho} = \epsilon_{\nu\rho\sigma}\tilde{F}^\sigma$).
(2) $\epsilon^{\mu\nu\sigma} = -\epsilon^{\nu\mu\sigma}$.
(3) Insert into (13.38) and multiply by $-e^2$.

This is a first-order equation. Apply it twice. Lower the indices of (13.39), $\epsilon_{\lambda\alpha\nu}\partial^\alpha\tilde{F}^\nu = m\tilde{F}_\lambda$, and act with $\epsilon_{\lambda\alpha\nu}\partial^\alpha$ on (13.39) itself:

$$\begin{aligned} m^2\tilde{F}_\lambda &\overset{(1)}{=} m\,\epsilon_{\lambda\alpha\nu}\partial^\alpha\tilde{F}^\nu \overset{(2)}{=} \epsilon_{\lambda\alpha\nu}\epsilon^{\nu\mu\sigma}\partial^\alpha\partial_\mu\tilde{F}_\sigma \overset{(3)}{=} \left(\delta^\mu_\lambda\delta^\sigma_\alpha - \delta^\sigma_\lambda\delta^\mu_\alpha\right)\partial^\alpha\partial_\mu\tilde{F}_\sigma \\ &\overset{(4)}{=} \partial_\lambda\left(\partial^\alpha\tilde{F}_\alpha\right) - \partial^\alpha\partial_\alpha\tilde{F}_\lambda \overset{(5)}{=} -\Box\tilde{F}_\lambda. \end{aligned} \tag{13.40}$$

(1) The lowered form of (13.39), multiplied by m.
(2) Replace $m\tilde{F}^\nu$ by the left side of (13.39).
(3) $\epsilon_{\lambda\alpha\nu}\epsilon^{\nu\mu\sigma} = \epsilon_{\nu\lambda\alpha}\epsilon^{\nu\mu\sigma} = \delta^\mu_\lambda\delta^\sigma_\alpha - \delta^\sigma_\lambda\delta^\mu_\alpha$ (cyclic reordering, and $\epsilon_{012}\epsilon^{012} = +1$ in this signature).
(4) Carry out the contractions.
(5) Bianchi identity $\partial^\alpha\tilde{F}_\alpha = 0$, and $\Box = \partial^\alpha\partial_\alpha$.

So $(\Box + m^2)\tilde{F} = 0$: every component of the field strength obeys the Klein–Gordon equation with

$$m_{\text{top}} = |\kappa|\,e^2 = \frac{|k|\,e^2}{2\pi}. \tag{13.41}$$

The dispersion relation is $\omega^2 = \vec{k}^2 + m^2_{\text{top}}$: a gap (Figure 13.17).

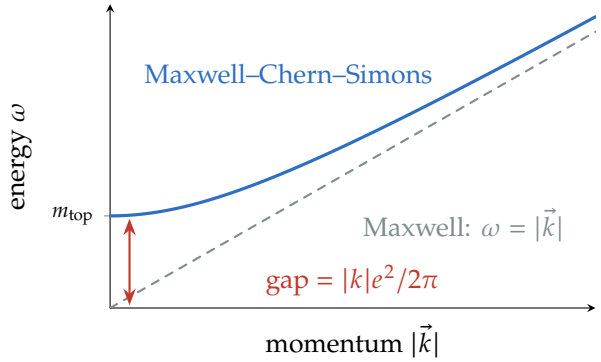


Figure 13.17: Dispersion relations of the 2 + 1 dimensional photon. Pure Maxwell theory (grey, dashed) is gapless, $\omega = |\vec{k}|$. Adding a Chern–Simons term at level k lifts the curve to $\omega = \sqrt{\vec{k}^2 + m^2_{\text{top}}}$ (blue), with a gap set by the level and the coupling, and no other field involved.

13.5.3 One polarisation with a definite spin

How many states does the massive photon have? A Proca photon in 2 + 1 dimensions (Example 13.12 with three spacetime dimensions) has two polarisations, spin +1 and spin −1. The Chern–Simons photon has only one.

Worked example 13.20: the rest frame

Look for a solution at rest, $\tilde{F}^\mu = f^\mu e^{-iEt}$ with constant f^μ and $E > 0$. The Bianchi identity gives $\partial_\mu \tilde{F}^\mu = \partial_0 \tilde{F}^0 = -iE f^0 e^{-iEt} = 0$, so $f^0 = 0$. The spatial components of (13.39) involve only the time derivative ($\mu = 0$):

$$m f^i \overset{(1)}{=} \epsilon^{i0j}(-iE) f_j \overset{(2)}{=} -iE\, \epsilon^{0ij} f^j \overset{(3)}{=} -iE\, \varepsilon_{ij} f^j,$$

(1) $\partial_0 \to -iE$ and the plane wave factor cancels.
(2) $\epsilon^{i0j} = -\epsilon^{0ij}$ and $f_j = -f^j$ (lowering a spatial index costs a sign). The two signs cancel.
(3) Write $\varepsilon_{ij} = \epsilon^{0ij}$, the two-dimensional antisymmetric symbol with $\varepsilon_{12} = 1$.

In matrix form this is

$$-iE \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} f^1 \\ f^2 \end{pmatrix} = m \begin{pmatrix} f^1 \\ f^2 \end{pmatrix}.$$

The matrix $\left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$ has eigenvalues $\pm i$, with eigenvectors $(1, \pm i)$. So $-iE \cdot (\pm i) = \pm E$ must equal m. Since $E > 0$, only *one* eigenvector works, $E = |m|$, with the sign $\pm = \operatorname{sgn} m = \operatorname{sgn} \kappa$. For $\kappa > 0$ the solution is $f \propto (1, i)$, so the real field is $(\cos Et,\ \sin Et)$: it rotates counterclockwise. For $\kappa < 0$ it rotates clockwise. The Chern–Simons photon is a single massive particle of spin $\operatorname{sgn} \kappa = \pm 1$ (Figure 13.18).

The counting matches: the Maxwell photon has one massless state. The Chern–Simons term does not add a state; it gives the existing one a mass and a handedness. The absent partner of opposite spin is the visible sign that parity is broken.

2 + 1 dimensional theory	states	mass	spin
Maxwell	1	0	none (parity even)
Proca, $\frac{1}{2}m^2A^2$	2	m	+1 and −1
Maxwell–Chern–Simons	1	$\lvert k\rvert e^2/2\pi$	$\operatorname{sgn} k$
pure Chern–Simons	0	—	—

Try it: the propagator pole

Add the gauge-fixing term $-\frac{1}{2\xi e^2}(\partial \cdot A)^2$ and invert the quadratic form in momentum space. The result (Dunne, *Aspects of Chern–Simons theory*, eq. (26)) is

$$\Delta_{\mu\nu}(p) = e^2 \frac{p^2 g_{\mu\nu} - p_\mu p_\nu - i\kappa e^2 \epsilon_{\mu\nu\rho} p^\rho}{p^2 (p^2 - \kappa^2 e^4)} + \xi e^2 \frac{p_\mu p_\nu}{(p^2)^2}.$$

Which poles are physical? *Answer:* the pole at $p^2 = \kappa^2 e^4$ is the massive photon. The $1/p^2$ poles multiply $p^2 g_{\mu\nu} - p_\mu p_\nu$ or $p_\mu p_\nu$, which vanish between conserved currents at $p^2 = 0$ (or are gauge artefacts), so there is no massless particle.

13.5.4 Charges carry flux, and their fields are screened

Couple a conserved current, $\mathcal{L} \ni -A_\mu j^\mu$. The field equation becomes $\frac{1}{e^2}\partial_\mu F^{\mu\nu} + \kappa \tilde{F}^\nu = j^\nu$. Its time component is Gauss's law. With $E^i = F^{i0}$ and $B = \tilde{F}^0 = F_{12}$ it reads

$$\frac{1}{e^2} \nabla \cdot \vec{E} + \kappa B = \rho. \tag{13.42}$$

Worked example 13.21: a static charge

Put a charge Q at the origin, $\rho = Q\delta^2(\vec{x})$. The fields are components of $\tilde{F}$, which obeys $(\Box + m^2)\tilde{F} =$ (local source terms), so away from the charge the static fields obey $(-\nabla^2 + m^2)\tilde{F} = 0$ and fall off like the two-dimensional Yukawa function $K_0(mr) \sim e^{-mr}/\sqrt{r}$ (Example 13.8). Integrate (13.42) over a disc of radius $R \gg 1/m$:

$$Q \overset{(1)}{=} \frac{1}{e^2} \oint_{|\vec{x}|=R} \vec{E} \cdot \hat{\vec{n}}\, \mathrm{d}\ell + \kappa \int_{|\vec{x}|<R} B\, \mathrm{d}^2x \overset{(2)}{\longrightarrow} \kappa\, \Phi \overset{(3)}{\Longrightarrow} \Phi = \frac{Q}{\kappa} = \frac{2\pi Q}{k}.$$

(1) Integrate (13.42), using the divergence theorem on the first term.

(2) As $R \to \infty$ the boundary term vanishes because $\vec{E}$ decays exponentially. Φ is the total magnetic flux.

(3) $\kappa = k/2\pi$.

The charge is surrounded by a cloud of magnetic flux $2\pi Q/k$, and its electric field is screened: nothing of the charge is felt beyond $1/m$ by local measurements. What remains at long distances is purely topological. A second charge carried around the first picks up an Aharonov–Bohm phase from the attached flux, and exchanging two unit charges gives the phase $e^{i\pi/k}$: the particles are **anyons** (Chapter 12).

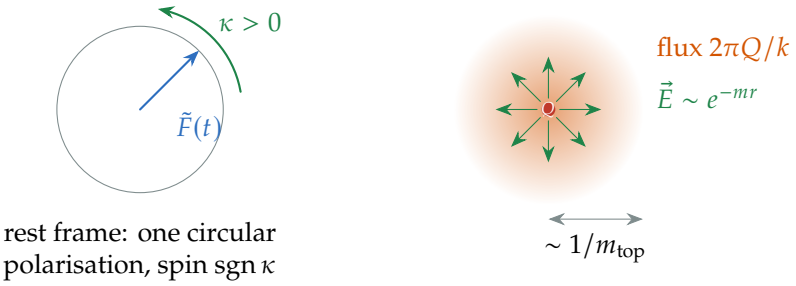


Figure 13.18: Left: at rest, the Chern–Simons photon's field strength rotates in one sense only, fixed by the sign of the level: a single state of spin ±1. Right: a static charge Q is dressed by a cloud of magnetic flux $2\pi Q/k$ (orange) of size $1/m_{\text{top}}$. Its electric field (green) is screened exponentially. At long distances only the attached flux remains, visible through Aharonov–Bohm phases.

13.5.5 The infrared: a topological field theory

Compare the two terms of (13.37) by dimension. In $2+1$ dimensions $[A] = 1$, so F^2 has dimension 4 and needs the coefficient $1/e^2$ of dimension -1, while $A\,\mathrm{d}A$ has dimension 3 and a dimensionless coefficient. At energies $E \ll m_{\text{top}} = \kappa e^2$, the Maxwell term is suppressed by E/m_{top} and drops out. What is left is pure Chern–Simons theory, $\mathrm{U}(1)_k$: a TQFT with no local degrees of freedom and zero Hamiltonian. So Maxwell–Chern–Simons theory is *gapped with topological order* (Section 13.1.4): its vacua are the states of the TQFT, and the massive photon sits a gap m_{top} above them.

Worked example 13.22: k vacua on a torus

Put $\mathrm{U}(1)_k$ Chern–Simons theory on a spatial torus with coordinates $x, y \in [0, 1)$. The low-energy configurations are flat connections, $A = a_1(t)\,\mathrm{d}x + a_2(t)\,\mathrm{d}y$, with holonomies e^{ia_1}, e^{ia_2} around the two cycles, and $a_j \sim a_j + 2\pi$ by large gauge transformations. Then

$$A \wedge \mathrm{d}A \overset{(1)}{=} (a_1\mathrm{d}x + a_2\mathrm{d}y) \wedge (\dot{a}_1\,\mathrm{d}t \wedge \mathrm{d}x + \dot{a}_2\,\mathrm{d}t \wedge \mathrm{d}y) \overset{(2)}{=} (a_2\dot{a}_1 - a_1\dot{a}_2)\,\mathrm{d}t \wedge \mathrm{d}x \wedge \mathrm{d}y,$$

$$S_{\mathrm{CS}} \overset{(3)}{=} \frac{k}{4\pi}\int \mathrm{d}t\,(a_2\dot{a}_1 - a_1\dot{a}_2) \overset{(4)}{=} \frac{k}{2\pi}\int \mathrm{d}t\,a_2\,\dot{a}_1.$$

(1) $\mathrm{d}A = \dot{a}_1\mathrm{d}t\wedge\mathrm{d}x + \dot{a}_2\mathrm{d}t\wedge\mathrm{d}y$.
(2) $\mathrm{d}x\wedge\mathrm{d}t\wedge\mathrm{d}y = -\mathrm{d}t\wedge\mathrm{d}x\wedge\mathrm{d}y$ and $\mathrm{d}y\wedge\mathrm{d}t\wedge\mathrm{d}x = +\mathrm{d}t\wedge\mathrm{d}x\wedge\mathrm{d}y$. Terms with $\mathrm{d}x\wedge\mathrm{d}x$ or $\mathrm{d}y\wedge\mathrm{d}y$ vanish.
(3) The x, y integrals give 1.
(4) Integrate $-a_1\dot{a}_2$ by parts to $+a_2\dot{a}_1$.

This is a first-order action, like $\int p\,\dot{q}$: a_1 is the coordinate and $p_1 = \frac{k}{2\pi}a_2$ its momentum. Quantising, $[a_1, a_2] = \frac{2\pi i}{k}$. The gauge-invariant operators are the holonomies $U = e^{ia_1}$ and $V = e^{ia_2}$, and $e^Xe^Y = e^Ye^Xe^{[X,Y]}$ for a central commutator gives

$$UV = e^{-2\pi i/k}\,VU.$$

The smallest matrices obeying this are $k\times k$: the clock and shift matrices $U = \mathrm{diag}(1,\omega,\omega^2,\ldots,\omega^{k-1})$ and V the cyclic shift, with $\omega = e^{-2\pi i/k}$. So the Hilbert space on the torus is k-dimensional: k degenerate vacua, whose number depends on the topology of space. On a sphere, with no noncontractible loops, there is one.

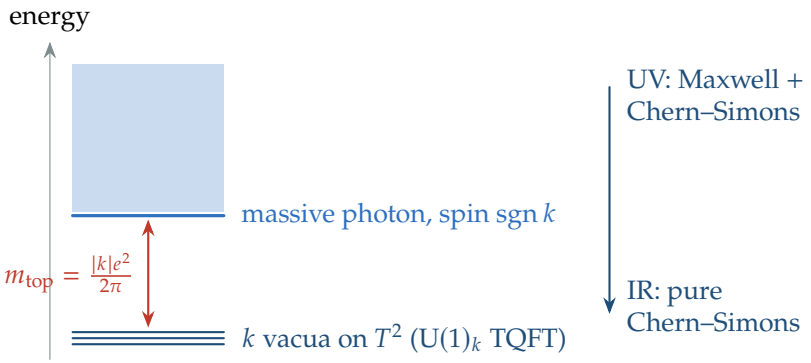


Figure 13.19: The spectrum of Maxwell–Chern–Simons theory on a spatial torus. The photon sits a gap m_{top} above a set of k exactly degenerate vacua. Below the gap only the topological field theory $\mathrm{U}(1)_k$ survives: the Maxwell term fades out along the RG flow (right), leaving a theory with no local excitations and a vacuum degeneracy that counts the topology of space.

Physical picture: where Chern–Simons terms come from

A Chern–Simons term is not an exotic addition. Integrate out a massive two-component Dirac fermion of mass M and charge one in $2+1$ dimensions: the one-loop vacuum polarisation contains a parity-odd piece, which at momenta $\ll |M|$ is a Chern–Simons term at level $\frac{1}{2}\,\mathrm{sgn}\,M$. (The half-integer is the **parity anomaly**, Chapter 10: a single massless Dirac fermion cannot be regularised without breaking parity.) In a quantum Hall system, integrating out the gapped electrons gives a Chern–Simons term whose level is the Hall conductance in units of e^2/h. Gapped matter leaves behind topological terms, and topological terms gap the gauge field.

Key idea: a gap from topology, not from condensation

In Maxwell–Chern–Simons theory the photon mass $|k|e^2/2\pi$ comes from a metric-independent term. No field condenses, no symmetry is spontaneously broken, and

gauge invariance is intact. The only inputs are the integer level k and the coupling e^2. The theory is gapped, parity violating, and at long distances a TQFT whose vacuum degeneracy on a torus is $|k|$.

Reading Nash (Ch. XII, pp. 343–344, (12.86)–(12.89))

Nash computes the correlation functions of Wilson lines in *pure* abelian Chern–Simons theory on S^3. The action is quadratic, the integral is Gaussian, and the answer is an exponential of Gauss linking numbers of the knots. There are no local degrees of freedom at all: only the topology of the knots matters. Adding the Maxwell term, as here, gives the theory exactly one local degree of freedom, a massive photon, and leaves Nash's topological answer as its long-distance limit. His Chern–Simons form (Ch. VIII, §2, p. 223) is the non-abelian version of $A \wedge \mathrm{d}A$.

Recap

- $S_{\mathrm{CS}} = \frac{k}{4\pi}\int A \wedge \mathrm{d}A$ is metric independent, gauge invariant up to boundary terms ($k \in \mathbb{Z}$), and odd under parity.
- Maxwell + Chern–Simons gives $\frac{1}{e^2}\partial_\mu F^{\mu\nu} + \kappa \tilde{F}^\nu = 0$, which squares to $(\Box + m^2)\tilde{F} = 0$ with $m = |k|e^2/2\pi$.
- There is one massive state, of spin $\operatorname{sgn} k$.
- A charge Q binds flux $2\pi Q/k$, and its electric field is screened: charges become anyons.
- Below the gap the theory is the TQFT $\mathrm{U}(1)_k$, with $|k|$ vacua on a torus: gapped with topological order.

Sources. S. Deser, R. Jackiw and S. Templeton, *Topologically massive gauge theories*, Ann. Phys. 140 (1982) 372; reviewed in G. Dunne, *Aspects of Chern–Simons theory* (hep-th/9902115), §2.2, whose equations (21)–(26) are those of this section. D. Tong, *Gauge Theory*, §8.4–8.5, for level quantisation and the fermion-induced term.

13.6 θ-terms and the Haldane gap

Take a chain of quantum spins of size S with antiferromagnetic nearest-neighbour coupling,

$$H = J\sum_j S_j \cdot S_{j+1}, \qquad J > 0. \tag{13.43}$$

Classically, and in spin-wave theory, all values of S behave alike: the spins want to alternate (the **Néel** pattern), and the excitations are gapless spin waves. In 1983 Haldane argued that this is wrong in a way that depends on S in a startling manner. For *integer* S the chain has a gap. For *half-integer* S it is gapless. The difference comes from a term in the effective action that is invisible in the classical equations of motion, a **θ-term**, with $\theta = 2\pi S$. This section builds the idea from the simplest example, derives the θ-term from the spin chain, and explains why $\theta = \pi$ is special.

13.6.1 A θ-term in quantum mechanics: the particle on a ring

A particle of moment of inertia I moves on a circle, angle $\varphi \sim \varphi + 2\pi$. Add a total-derivative term:

$$L = \frac{I}{2}\dot{\varphi}^2 + \frac{\theta}{2\pi}\dot{\varphi}.$$

The added term is $\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\theta}{2\pi}\varphi\right)$, so the Euler–Lagrange equation $I\ddot{\varphi} = 0$ does not see it. Classically θ is irrelevant. Quantum mechanically it is not:

$$p \overset{(1)}{=} \frac{\partial L}{\partial \dot{\varphi}} = I\dot{\varphi} + \frac{\theta}{2\pi}, \qquad H \overset{(2)}{=} p\dot{\varphi} - L = \frac{1}{2I}\left(p - \frac{\theta}{2\pi}\right)^2, \qquad E_n \overset{(3)}{=} \frac{1}{2I}\left(n - \frac{\theta}{2\pi}\right)^2, \quad n \in \mathbb{Z}.$$

(1) Canonical momentum.
(2) Legendre transform: substitute $\dot{\varphi} = (p - \theta/2\pi)/I$.
(3) Wavefunctions $e^{in\varphi}$ must be single valued on the circle, so $p = -i\partial_\varphi$ has integer eigenvalues n.

The spectrum depends on θ, periodically with period 2π (shifting $\theta \to \theta + 2\pi$ relabels $n \to n + 1$). Physically θ is a magnetic flux threading the ring, and this is the Aharonov–Bohm effect. In Euclidean time with period β, a path winds $w \in \mathbb{Z}$ times, and the θ-term contributes

$$-i\frac{\theta}{2\pi}\int_0^\beta \dot{\varphi}\,\mathrm{d}\tau = -i\theta\, w \qquad \Longrightarrow \qquad Z = \sum_{w\in\mathbb{Z}} e^{i\theta w}\, Z_w,$$

a phase that weights each topological sector (Figure 13.20).

Worked example 13.23: the gap closes at $\theta = \pi$

At $\theta = 0$ the ground state is $n = 0$ with $E_0 = 0$, and the first excited states $n = \pm 1$ have $E = 1/2I$: a gap $1/2I$. At $\theta = \pi$,

$$E_0 \overset{(1)}{=} \frac{1}{2I}\left(0 - \frac{1}{2}\right)^2 = \frac{1}{8I}, \qquad E_1 \overset{(2)}{=} \frac{1}{2I}\left(1 - \frac{1}{2}\right)^2 = \frac{1}{8I},$$

(1) $n = 0$, $\theta/2\pi = \frac{1}{2}$.
(2) $n = 1$.

so the two lowest levels are exactly degenerate. Time reversal sends $\theta \to -\theta$. It is a symmetry only at $\theta = 0$ and at $\theta = \pi$ (where $-\pi \equiv \pi \bmod 2\pi$), and at $\theta = \pi$ it exchanges the two ground states, $n = 0 \leftrightarrow n = 1$. In the path-integral language the sectors with even and odd winding enter with opposite signs, $e^{i\pi w} = (-1)^w$, and this destructive interference is what removes the unique ground state.

Working definition 13.4 (θ-term)

A **θ-term** is a contribution $-i\theta\, Q$ to the Euclidean action, where Q is an integer-valued topological charge [a winding number, unchanged by any small deformation of the field, like the number of times a path goes round a ring]. It is a total derivative, so it leaves the classical field equations unchanged [the Euler–Lagrange equations of L and of $L + \frac{\mathrm{d}}{\mathrm{d}t}f$ are the same], but it weights each topological sector by $e^{i\theta Q}$ [paths that wind differently interfere with different phases]. Physics depends on θ only modulo 2π. Time reversal (or parity) sends $\theta \to -\theta$, so it is a symmetry only at $\theta = 0$ and $\theta = \pi$. *To test for one:* find the integer Q of the field configurations (for example $\pi_1(S^1)$, $\pi_2(S^2)$, the instanton number of Chapter 8), and check whether the action contains Q with an imaginary coefficient.

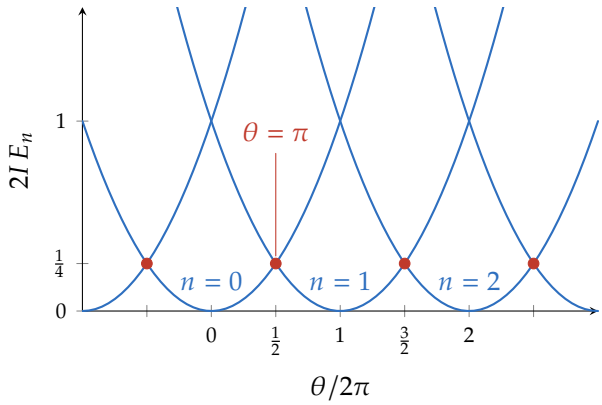


Figure 13.20: Energy levels of a particle on a ring as a function of the θ-angle. Each parabola is one momentum n. The spectrum is periodic in θ with period 2π. At $\theta = 0$ the ground state is unique and gapped. At $\theta = \pi$ (red) two parabolas cross and the ground state is doubly degenerate: a topological term, invisible classically, has closed the gap.

The examples we need are:

- four-dimensional Yang–Mills theory, $\frac{\theta}{8\pi^2}\int \operatorname{tr} F \wedge F$, with Q the instanton number (Chapter 8);
- the two-dimensional O(3) sigma model, a unit vector field $\vec{n}(x,\tau) \in S^2$, with

$$Q = \frac{1}{4\pi}\int \mathrm{d}x\,\mathrm{d}\tau\ \vec{n}\cdot\left(\partial_x\vec{n}\times\partial_\tau\vec{n}\right) \in \mathbb{Z}, \tag{13.44}$$

  the degree of the map from spacetime (compactified to S^2) to the target S^2. The integrand $\vec{n}\cdot(\partial_x\vec{n}\times\partial_\tau\vec{n})\,\mathrm{d}x\,\mathrm{d}\tau$ is the area element of the target sphere pulled back to spacetime, so $4\pi Q$ is the total area covered, counted with orientation.

13.6.2 A single spin: the Berry phase is a solid angle

A spin S can be described by the direction $\vec{n}$ in which it points. The **spin coherent state** $|\vec{n}\rangle$ is the state $|S,S\rangle$ (maximal S_z) rotated so that its axis points along $\vec{n}$, with $\langle\vec{n}|\vec{S}|\vec{n}\rangle = S\vec{n}$. When $\vec{n}(\tau)$ moves round a closed loop, the state acquires a **Berry phase**, and in the path integral for the spin this phase is the whole of the kinetic term. We compute it for a loop at fixed polar angle ϑ_0, with azimuth ϕ going once round.

Worked example 13.24: Berry phase of a spin

Take $|\vec{n}\rangle = e^{iS\phi}e^{-i\phi S_z}e^{-i\vartheta_0 S_y}|S,S\rangle$. (The prefactor $e^{iS\phi}$ makes the state smooth at the north pole.) The Berry phase is $\gamma = i\oint\langle\vec{n}|\partial_\tau\vec{n}\rangle\,\mathrm{d}\tau$:

$$\langle\vec{n}|\partial_\tau|\vec{n}\rangle \overset{(1)}{=} iS\dot\phi - i\dot\phi\,\langle S,S|e^{i\vartheta_0 S_y}S_z e^{-i\vartheta_0 S_y}|S,S\rangle \overset{(2)}{=} iS\dot\phi - i\dot\phi\,S\cos\vartheta_0,$$

$$\gamma \overset{(3)}{=} -S\oint(1-\cos\vartheta_0)\,\mathrm{d}\phi \overset{(4)}{=} -S\cdot 2\pi(1-\cos\vartheta_0) \overset{(5)}{=} -S\,\Omega.$$

(1) Differentiate the two ϕ-dependent factors. The operator $e^{i\phi S_z}\cdots e^{-i\phi S_z}$ from the conjugation leaves S_z unchanged.
(2) Rotating by ϑ_0 about the y axis turns S_z into $S_z\cos\vartheta_0$ + (a component of S_x), and $\langle S,S|S_x|S,S\rangle = 0$.
(3) Multiply by i and integrate.

(4) ϕ increases by 2π.
(5) $\Omega = 2\pi(1 - \cos\vartheta_0)$ is the solid angle of the cap enclosed by the loop.

The phase is S times the solid angle enclosed by the path of $\vec{n}$, as if the spin were a charge moving on the unit sphere in the field of a magnetic monopole of strength S at the centre.

The loop bounds two caps, of solid angles Ω and $4\pi - \Omega$ (with opposite orientation). The two choices give phases differing by $4\pi S$, so the path-integral weight $e^{iS\Omega}$ is unambiguous only if $e^{4\pi iS} = 1$, that is $2S \in \mathbb{Z}$. Spin quantisation is Dirac quantisation for the monopole of strength S (Figure 13.21).

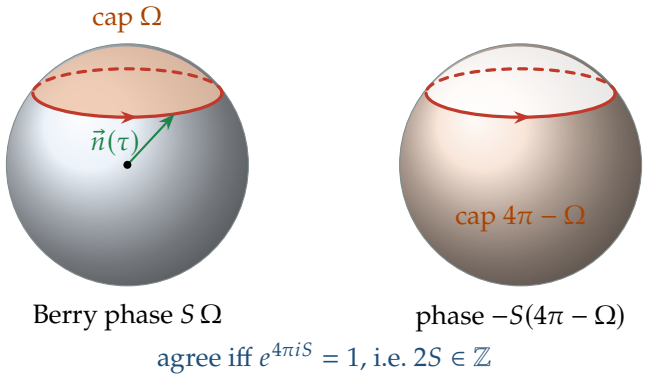


Figure 13.21: The Berry phase of a spin is S times the solid angle swept by its direction $\vec{n}$. The same loop (red) bounds the upper cap (left) and, with the opposite orientation, the rest of the sphere (right). The two answers differ by $4\pi S$, so the quantum amplitude is well defined only for integer or half-integer spin: Dirac's quantisation condition for a monopole of strength S.

We will also need how the solid angle changes when the loop is deformed. Move the loop $\vec{n}(\tau)$ to $\vec{n} + \delta\vec{n}$. The enclosed solid angle changes by the area of the thin strip between the old and the new loop. The strip has sides $\partial_\tau\vec{n}\,\mathrm{d}\tau$ and $\delta\vec{n}$, and its oriented area on the unit sphere (normal $\vec{n}$) is $\vec{n}\cdot(\delta\vec{n}\times\partial_\tau\vec{n})\,\mathrm{d}\tau$. So

$$\delta\Omega = \oint \mathrm{d}\tau\;\vec{n}\cdot\left(\delta\vec{n}\times\partial_\tau\vec{n}\right) \tag{13.45}$$

(with an orientation convention whose overall sign plays no role below).

13.6.3 The spin chain becomes the $\mathrm{O}(3)$ sigma model

For large S the spins of (13.43) stay close to a Néel pattern whose direction varies slowly. Let the lattice spacing be a and write

$$\vec{S}_j \approx S\,(-1)^j\,\vec{n}(x_j) + (\text{small uniform part}),$$

with $\vec{n}$ a slowly varying unit vector, the local Néel direction (Figure 13.22). The Euclidean action is the sum of the Berry phases of all spins and the energy. We compute three pieces.

(a) The staggered Berry phases give the θ-term. The spin at site j points along $(-1)^j\vec{n}_j$. The antipodal loop $-\vec{n}$ encloses the solid angle $-\Omega[\vec{n}]$ (mod 4π), so the total Berry phase is

$S\sum_j(-1)^j\Omega[\vec n_j]$. Pair the sites $(2k,2k+1)$:

$$S\sum_j(-1)^j\Omega[\vec n_j]\overset{(1)}{=}S\sum_k\big(\Omega[\vec n_{2k}]-\Omega[\vec n_{2k+1}]\big)\overset{(2)}{\approx}-S\sum_k a\,\partial_x\Omega\overset{(3)}{\approx}-\frac{S}{2}\int \mathrm{d}x\;\partial_x\Omega$$
$$\overset{(4)}{=}-\frac{S}{2}\int \mathrm{d}x\,\mathrm{d}\tau\;\vec n\cdot(\partial_x\vec n\times\partial_\tau\vec n)\overset{(5)}{=}-2\pi S\,Q. \tag{13.46}$$

(1) Group even and odd sites.
(2) $\vec n_{2k+1}\approx\vec n_{2k}+a\,\partial_x\vec n$, so the difference of solid angles is $a\,\partial_x\Omega$ to first order.
(3) The pairs are spaced by $2a$, so $\sum_k 2a\,f(x_{2k})\to\int \mathrm{d}x\,f$.
(4) (13.45) with $\delta\vec n=\partial_x\vec n\,\mathrm{d}x$.
(5) Definition (13.44) of Q.

So the chain produces a θ-term with

$$\boxed{\theta=2\pi S.} \tag{13.47}$$

(Pairing $(2k-1,2k)$ instead flips the sign, $\theta=-2\pi S$. Since $e^{\pm 2\pi i SQ}$ are equal for $2S\in\mathbb{Z}$, nothing depends on this choice.)

(b) Twisting the Néel direction costs gradient energy. For the Néel part, $\vec S_j\cdot\vec S_{j+1}=-S^2\vec n_j\cdot\vec n_{j+1}$. Expand to second order:

$$\vec n(x)\cdot\vec n(x+a)\overset{(1)}{=}\vec n\cdot\big(\vec n+a\vec n'+\tfrac{a^2}{2}\vec n''\big)\overset{(2)}{=}1+\tfrac{a^2}{2}\,\vec n\cdot\vec n''\overset{(3)}{=}1-\tfrac{a^2}{2}(\vec n')^2,$$

(1) Taylor expansion, $\vec n'=\partial_x\vec n$.
(2) $\vec n\cdot\vec n=1$ and, differentiating, $\vec n\cdot\vec n'=0$.
(3) Differentiating $\vec n\cdot\vec n'=0$ once more gives $\vec n\cdot\vec n''=-(\vec n')^2$.

so each bond costs $\frac{JS^2a^2}{2}(\partial_x\vec n)^2$ above the Néel energy, and the energy per unit length is $\frac{JS^2a}{2}(\partial_x\vec n)^2$.

(c) Canting and the time-derivative term. A spin configuration that moves in time must cant slightly: in each cell of two sites write $\vec S_{2k}\propto S(\vec n+a\vec l)$, $\vec S_{2k+1}\propto -S(\vec n-a\vec l)$, normalised, with $\vec l\perp\vec n$ a small uniform magnetisation. Canting costs energy:

$$\vec S_{2k}\cdot\vec S_{2k+1}\overset{(1)}{=}-S^2\,\frac{(\vec n+a\vec l)\cdot(\vec n-a\vec l)}{1+a^2\vec l^{\,2}}\overset{(2)}{=}-S^2\,\frac{1-a^2\vec l^{\,2}}{1+a^2\vec l^{\,2}}\overset{(3)}{\approx}-S^2\big(1-2a^2\vec l^{\,2}\big),$$

(1) Each vector $\vec n\pm a\vec l$ has length $\sqrt{1+a^2\vec l^{\,2}}$ because $\vec n\perp\vec l$, and we divide by it.
(2) $\vec n\cdot\vec l=0$.
(3) $(1-\epsilon)/(1+\epsilon)\approx 1-2\epsilon$.

so each bond costs $2JS^2a^2\vec l^{\,2}$. With two bonds per cell of length $2a$, the energy per unit length is $2JS^2a\,\vec l^{\,2}$. The canting also changes the Berry phases. The two spins of a cell contribute $S\,\Omega[\vec n+a\vec l]-S\,\Omega[\vec n-a\vec l]\approx S\int \mathrm{d}\tau\,\vec n\cdot(2a\vec l\times\partial_\tau\vec n)$ by (13.45), which per unit length is $S\,\vec l\cdot(\partial_\tau\vec n\times\vec n)$. The Euclidean Lagrangian density for $\vec l$ is therefore $2JS^2a\,\vec l^{\,2}-iS\,\vec l\cdot(\partial_\tau\vec n\times\vec n)$. The field $\vec l$ appears quadratically, so we integrate it out:

$$\int \mathrm{d}^2l\;e^{-\left[A\vec l^{\,2}-i\vec b\cdot\vec l\right]}\overset{(1)}{=}\frac{\pi}{A}\,e^{-\vec b^{\,2}/4A},\qquad \frac{\vec b^{\,2}}{4A}\overset{(2)}{=}\frac{S^2(\partial_\tau\vec n)^2}{8JS^2a}=\frac{(\partial_\tau\vec n)^2}{8Ja},$$

(1) Complete the square, $A\vec{l}^2 - i\vec{b}\cdot\vec{l} = A(\vec{l} - i\vec{b}/2A)^2 + \vec{b}^2/4A$, for the two components of $\vec{l}$ perpendicular to $\vec{n}$.
(2) $A = 2JS^2a$ and $\vec{b} = S\,\partial_\tau\vec{n}\times\vec{n}$, whose square is $S^2(\partial_\tau\vec{n})^2$ because $\partial_\tau\vec{n}\perp\vec{n}$.

The result. Adding (a), (b) and (c), the long-wavelength Euclidean action of the spin chain is

$$S_E = \frac{1}{2g}\int \mathrm{d}x\,\mathrm{d}\tau\left[v\,(\partial_x\vec{n})^2 + \frac{1}{v}(\partial_\tau\vec{n})^2\right] - i\theta\,Q, \qquad g = \frac{2}{S}, \quad v = 2JSa, \quad \theta = 2\pi S. \tag{13.48}$$

To read off g and v, match $\frac{v}{2g} = \frac{JS^2a}{2}$ and $\frac{1}{2gv} = \frac{1}{8Ja}$: multiplying gives $\frac{1}{4g^2} = \frac{S^2}{16}$, and dividing gives $v^2 = 4J^2S^2a^2$. The time direction can be rescaled by v to make the action Lorentz invariant, with v playing the role of the speed of light. This is the O(3) **nonlinear sigma model** with a θ-term. It is also the $\mathbb{CP}^1$ model of Section 13.3.4, since $\mathbb{CP}^1 = S^2$.

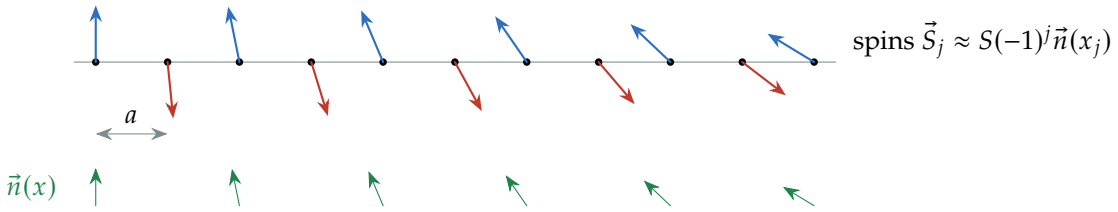


Figure 13.22: The Néel field of an antiferromagnetic chain. Neighbouring spins (blue and red) point in nearly opposite directions. Their common axis, the Néel direction $\vec{n}(x)$ (green, drawn below the chain), rotates slowly along the chain. Slow rotations in space cost gradient energy. Slow rotations in time force a small canting of each pair, and the Berry phases of the alternating spins leave behind the topological term $-2\pi iSQ$.

13.6.4 Integer versus half-integer spin

The weight of a topological sector is

$$e^{i\theta Q} = e^{2\pi iSQ} = \begin{cases} 1 & S\in\mathbb{Z}, \\ (-1)^Q & S\in\mathbb{Z}+\frac{1}{2}. \end{cases}$$

Integer spin: the O(3) model at $\theta = 0$ is gapped. With no phase, (13.48) is the plain two-dimensional O(3) sigma model. Like the $\mathbb{CP}^{N-1}$ model it is asymptotically free, with one-loop running $\mu\,\mathrm{d}g/\mathrm{d}\mu = -g^2/2\pi$ (for O(N) the coefficient is $(N-2)/2\pi$). By the argument of Section 13.3.3, a mass is generated.

Worked example 13.25: estimating the Haldane gap

The bare coupling $g = 2/S$ is defined at the lattice scale, momentum $\mu_0 \sim 1/a$, which corresponds to the energy $v/a = 2JS$. Integrating the running coupling,

$$\frac{1}{g(\mu)} \overset{(1)}{=} \frac{1}{g(\mu_0)} - \frac{1}{2\pi}\ln\frac{\mu_0}{\mu} \quad \overset{(2)}{\Longrightarrow} \quad \mu_* \overset{(3)}{=} \mu_0\,e^{-2\pi/g(\mu_0)} \overset{(4)}{=} \frac{1}{a}\,e^{-\pi S},$$

(1) $\frac{\mathrm{d}}{\mathrm{d}\ln\mu}\frac{1}{g} = -\frac{1}{g^2}\frac{\mathrm{d}g}{\mathrm{d}\ln\mu} = \frac{1}{2\pi}$, integrated from μ up to μ_0.
(2) The coupling becomes strong ($1/g\to 0$) at the scale μ_*.
(3) Solve $1/g(\mu_0) = \frac{1}{2\pi}\ln(\mu_0/\mu_*)$.
(4) $g(\mu_0) = 2/S$ and $\mu_0 \sim 1/a$.

The gap is $\Delta \sim v\mu_* \sim JS\, e^{-\pi S}$ and the correlation length is $\xi \sim a\, e^{\pi S}$. (A two-loop treatment refines the prefactor to $\Delta \sim JS^2 e^{-\pi S}$.) The gap is exponentially small for large spin, consistent with the gapless classical limit, but it is nonzero for every integer S. Numerically, the spin-1 chain has $\Delta = 0.4105\,J$ and $\xi \approx 6$ sites, and the spin-2 chain has $\Delta \approx 0.087\,J$ and $\xi \approx 50$ sites. The gap has been seen by neutron scattering in quasi-one-dimensional spin-1 compounds such as NENP.

Half-integer spin: $\theta = \pi$ and destructive interference. Now sectors of odd Q come with a minus sign. In the integer case the gap is produced by the disordering effect of all the topologically nontrivial fluctuations of $\vec{n}$. At $\theta = \pi$ those fluctuations cancel in pairs, just as the even and odd windings did for the particle on a ring (Example 13.23). The result, established by Affleck and Haldane (1987) using the non-abelian bosonisation of Chapter 11, is that the O(3) model at $\theta = \pi$ flows to a gapless conformal field theory, the $\mathrm{SU}(2)_1$ Wess–Zumino–Witten model. Three independent pieces of evidence support this.

- *Exact solution for $S = \frac{1}{2}$.* Bethe's ansatz solves the spin-$\frac{1}{2}$ chain exactly. Its elementary excitations, spinons, have dispersion $\varepsilon(k) = \frac{\pi J}{2}|\sin k|$, which vanishes at $k = 0, \pi$: gapless.
- *Lieb–Schultz–Mattis.* For every half-integer S, a rigorous theorem (Section 13.7.2) forbids a unique gapped ground state that preserves translation symmetry.
- *An anomaly at $\theta = \pi$.* The O(3) model at $\theta = \pi$ has a mixed 't Hooft anomaly between its SO(3) symmetry and time reversal (or translation, from the lattice point of view). It forbids a trivially gapped phase (Section 13.7).

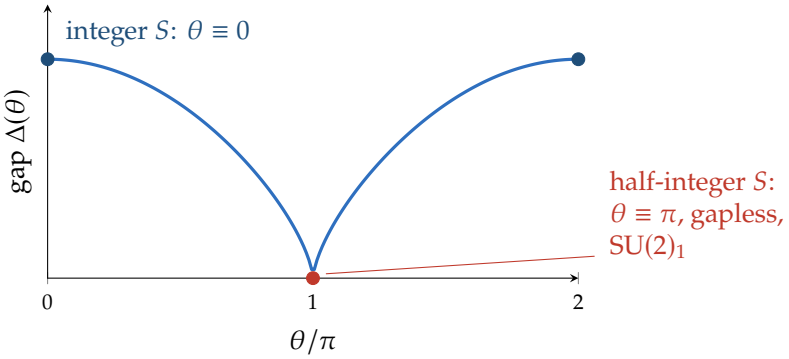


Figure 13.23: Schematic gap of the O(3) sigma model as a function of θ, at the coupling relevant for the Heisenberg chain. The gap is largest at $\theta = 0$ (integer spin, dark dots) and closes at $\theta = \pi$ (half-integer spin, red dot), where the theory is the gapless $\mathrm{SU}(2)_1$ conformal field theory. Only the two time-reversal-invariant values, 0 and π, can be realised by a Heisenberg chain.

The role of $\theta = \pi$. Two facts make $\theta = \pi$ special. First, it is one of only two values compatible with time reversal, so a time-reversal- invariant microscopic model (such as the Heisenberg chain) must land on $\theta = 0$ or $\theta = \pi$. Second, at $\theta = \pi$ the anomaly leaves only two options: the theory is gapless, or a symmetry breaks spontaneously and the ground state is degenerate. Both occur. The Heisenberg chain is gapless. Adding a next-nearest-neighbour coupling J_2 drives it, beyond $J_2/J \approx 0.241$, into a *dimerised* phase: gapped, with two ground states that break translation by one site. At $J_2 = J/2$ (the Majumdar–Ghosh point) the two dimer coverings are exact ground states (Figure 13.24). What cannot happen at $\theta = \pi$ is a unique gapped ground state.

Worked example 13.26: the AKLT chain: an exactly gapped spin-1 chain

Integer spin chains can be made exactly solvable. For spin 1, consider

$$H_{\mathrm{AKLT}} = J\sum_j \Big[\vec{S}_j\cdot\vec{S}_{j+1} + \tfrac{1}{3}(\vec{S}_j\cdot\vec{S}_{j+1})^2\Big].$$

For two spins 1 with total spin $s\in\{0,1,2\}$, $x=\vec{S}_j\cdot\vec{S}_{j+1}=\frac{1}{2}[s(s+1)-4]$ takes the values $-2,-1,1$. The projector onto $s=2$ is $P_2=\frac{(x+2)(x+1)}{(1+2)(1+1)}$ (it vanishes at $x=-2,-1$ and equals 1 at $x=1$). So

$$x+\tfrac{1}{3}x^2 \overset{(1)}{=} \tfrac{1}{3}\big(x^2+3x+2\big)-\tfrac{2}{3} \overset{(2)}{=} 2P_2-\tfrac{2}{3},$$

(1) Add and subtract $\frac{2}{3}$.
(2) $x^2+3x+2=(x+2)(x+1)=6P_2$.

and $H_{\mathrm{AKLT}}=J\sum_j(2P_2^{(j,j+1)}-\frac{2}{3})$: a sum of projectors, minus a constant. Now build a state as in Figure 13.24: write each spin 1 as two spin-$\frac{1}{2}$'s symmetrised, and join neighbouring halves into singlets. On any bond, the two spin-1's share one singlet, so their total spin is at most 1 and P_2 annihilates the state. Since every $P_2\geq 0$, this **valence-bond solid** is an exact ground state, with energy $-\frac{2}{3}J$ per bond. Affleck, Kennedy, Lieb and Tasaki proved that it is the unique ground state on a ring and that there is a gap above it: a rigorous Haldane gap. On an open chain the two unpaired spin-$\frac{1}{2}$'s at the ends give a fourfold ground-state degeneracy, the signature of the **Haldane phase** as a symmetry-protected topological phase: trivially gapped in the bulk, with a topological response at its edges.

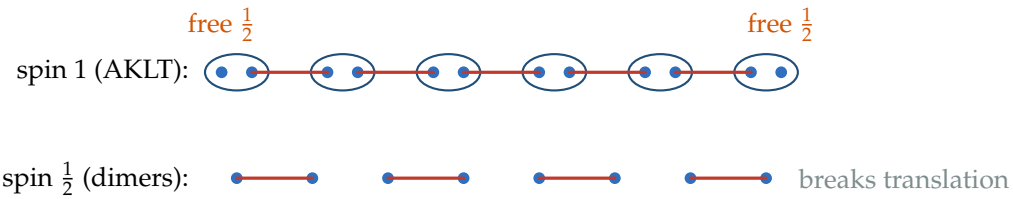


Figure 13.24: Top: the AKLT valence-bond solid. Each spin 1 (ellipse) is two spin-$\frac{1}{2}$'s, and every bond carries one singlet (red). The state is translation invariant, unique on a ring and gapped. On an open chain two spin-$\frac{1}{2}$'s are left free at the ends. Bottom: for spin $\frac{1}{2}$ each site has only one spin-$\frac{1}{2}$, so singlets can cover only every other bond. A gapped state must pick one of two dimer patterns, breaking translation. This is the lattice face of $\theta=2\pi S$.

Key idea: Haldane's conjecture

The Heisenberg antiferromagnetic chain of spin S is, at long distances, the O(3) sigma model with $g=2/S$ and $\theta=2\pi S$. For integer S ($\theta\equiv 0$) the model is gapped, with $\Delta\sim JS^2e^{-\pi S}$. For half-integer S ($\theta\equiv\pi$) the topological sectors interfere destructively and the model is gapless, flowing to the $\mathrm{SU}(2)_1$ conformal field theory. The gap is decided by a term that does not appear in any classical equation of motion.

Recap

- A θ-term $-i\theta Q$ weights topological sectors by $e^{i\theta Q}$. It is invisible classically, periodic in θ, and time-reversal invariant only at $\theta=0,\pi$.

- For the particle on a ring, $E_n = (n - \theta/2\pi)^2/2I$, and the ground state becomes degenerate at $\theta = \pi$.
- A spin's Berry phase is S times a solid angle. Staggering it along a Néel chain produces $\theta = 2\pi S$ for the O(3) model, with $g = 2/S$ and $v = 2JSa$.
- Integer S: gapped (Haldane gap $0.41J$ for $S = 1$, AKLT exactly). Half-integer S: gapless, or gapped only by breaking translation.

Sources. F. D. M. Haldane, the 1981 preprint *Ground state properties of antiferromagnetic chains with unrestricted spin* (arXiv:1612.00076), and Phys. Rev. Lett. 50 (1983) 1153; G. Sierra, *On the application of the non-linear sigma model to spin chains and spin ladders* (cond-mat/9610057), whose equations (6)–(8), (27) and (28) and Table 4 give the mapping $\theta = 2\pi S$, $g = 2/S$, $c = 2JS$ and the numerical gaps; I. Affleck, T. Kennedy, E. H. Lieb and H. Tasaki, Phys. Rev. Lett. 59 (1987) 799; I. Affleck and F. D. M. Haldane, Phys. Rev. B36 (1987) 5291.

13.7 Anomalies forbid a trivial gap

The previous sections were about producing gaps. This one is about the opposite: a topological obstruction that makes a gap *impossible*, unless something else happens. The obstruction is an 't Hooft anomaly. Its lattice version is the Lieb–Schultz–Mattis theorem, which is why half-integer spin chains cannot have a Haldane gap. The logic is short enough to state at once.

Key idea: anomaly matching and gaps

If a theory has an 't Hooft anomaly for a symmetry G, then its low-energy limit must reproduce the same anomaly. A *trivially* gapped phase (one vacuum, no topological order, G unbroken) cannot reproduce any anomaly. So the theory must be

1. gapless, or
2. gapped with G spontaneously broken (several vacua, or Goldstone bosons), or
3. gapped with topological order, a nontrivial TQFT in the infrared (Chapter 12) that carries the anomaly.

13.7.1 't Hooft anomalies

Take a theory with a global symmetry G (not gauged). We may always couple the symmetry currents to a *background* gauge field A, a fixed classical field that we are free to choose, and compute the partition function $Z[A]$. It is a generating function: derivatives with respect to A give correlation functions of the currents. If the symmetry were perfect, $Z[A]$ would be invariant under background gauge transformations $A \to A^\lambda$.

Working definition 13.5 ('t Hooft anomaly)

A global symmetry G has an **'t Hooft anomaly** if the partition function in a background gauge field A [a fixed, non-dynamical field coupled to the conserved currents of G, like an external electromagnetic field] transforms under background gauge transformations as $Z[A^\lambda] = e^{i\alpha[\lambda,A]}\, Z[A]$ with a phase α that cannot be removed by adding a local counterterm to the action [an integral of a polynomial in A and its derivatives, which one is always free to add because it does not change the physics at separated points]. The symmetry itself is exact [currents are conserved when $A = 0$, and states still fall into representations of G].

The anomaly is an obstruction to gauging G, not a breaking of G. *To test for one:* compute the variation of $Z[A]$ in the simplest background, for example from the triangle diagram of massless fermions (Chapter 10) or from spectral flow (Example 13.27).

Why the anomaly is the same at all energies. 't Hooft's argument: add to the theory a set of free massless "spectator" fermions, chosen to have exactly the opposite anomaly. The combined theory has no anomaly, so G can be gauged with a tiny coupling that does not disturb the dynamics. A consistent gauge theory must be consistent at every energy scale, so at low energies the original theory's infrared degrees of freedom, plus the spectators (which are free and do not change), must again cancel the anomaly. The infrared of the original theory therefore has exactly the anomaly of the ultraviolet. The anomaly is an invariant of the renormalisation-group flow.

Why a trivially gapped theory has none. This is where the gap enters. If the theory is gapped with a unique vacuum, every connected correlation function of currents decays exponentially (Section 13.2). Expanding $\ln Z[A]$ in powers of A, each coefficient is a connected correlator of currents, and exponentially decaying kernels can be expanded in derivatives. So

$$-\ln Z[A] = W[A] = \int \mathrm{d}^d x \, \left(c_0 + c_1 A^2 + c_2 (\partial A)^2 + \dots \right) + (\text{topological terms}), \tag{13.49}$$

a *local* functional of the background (up to topological terms such as Chern–Simons terms at integer levels, which are gauge invariant). But an anomaly is, by definition, a variation of $\ln Z$ that is *not* the variation of any local functional. If W is local, its variation is trivially the variation of a local functional. Contradiction: a trivially gapped theory has no 't Hooft anomaly. The contrapositive is the key idea above. In a gapless theory $W[A]$ is nonlocal (for example the $1/p^2$ pole of the two-dimensional anomaly, which is the massless boson of bosonisation). With spontaneous breaking, W depends on which vacuum is chosen and on the Goldstone fields. With topological order, the TQFT coupled to A carries the anomaly.

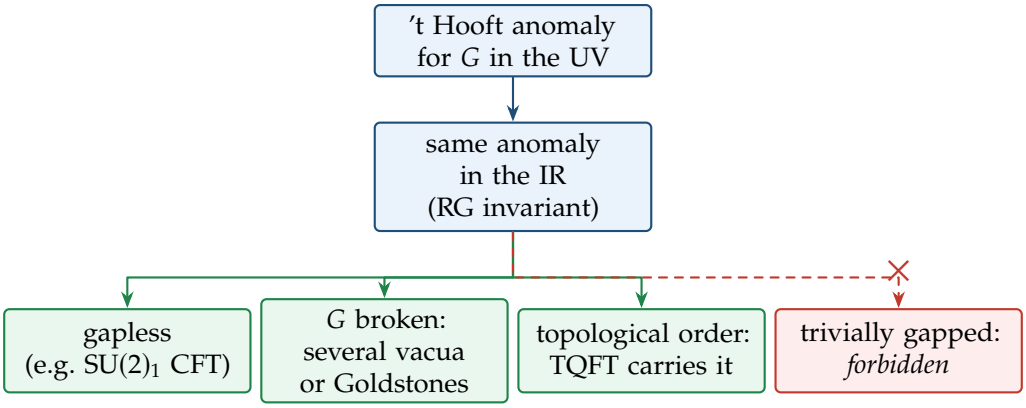


Figure 13.25: The logic of anomaly matching. An anomaly present at short distances is present at long distances. A trivially gapped infrared, whose response to background fields is local, cannot carry it (red). What remains is a gapless theory, a phase with spontaneously broken symmetry, or a topologically ordered phase whose TQFT carries the anomaly (green).

Worked example 13.27: spectral flow: the massless Dirac fermion in two dimensions

A massless Dirac fermion in 1 + 1 dimensions is a right-mover and a left-mover, with $E = +p$ and $E = -p$. It has two U(1) symmetries, the vector charge $Q_V = N_R + N_L$ and the

axial charge $Q_A = N_R - N_L$ ($N_{R,L}$ count particles minus holes). Put it on a ring of length L and thread a flux Φ of the $\mathrm{U}(1)_V$ background field through the ring. With $A_x = \Phi/L$, the covariant momentum is $p - A_x$, and the single-particle levels are

$$E_R(n) = \frac{2\pi}{L}\left(n - \frac{\Phi}{2\pi}\right), \qquad E_L(n) = -\frac{2\pi}{L}\left(n - \frac{\Phi}{2\pi}\right), \qquad n \in \mathbb{Z}.$$

As Φ goes from 0 to 2π, every right-moving level moves down by one spacing and every left-moving level moves up by one spacing (Figure 13.26). One right-moving level crosses $E = 0$ from above to below and becomes filled; one left-moving level crosses from below to above and becomes empty. So

$$\Delta N_R \overset{(1)}{=} +1, \qquad \Delta N_L \overset{(2)}{=} -1, \qquad \Delta Q_A \overset{(3)}{=} \Delta N_R - \Delta N_L = 2, \qquad \Delta Q_V = 0.$$

(1) The Dirac sea of right-movers fills all levels with $E < 0$; one more level has entered it.
(2) One left-moving level has left the sea: a hole.
(3) Definition of Q_A.

The Hamiltonian at $\Phi = 2\pi$ is gauge equivalent to the one at $\Phi = 0$, yet the ground state has become a state with $Q_A = 2$. This is the two-dimensional chiral anomaly, $\partial_\mu j^\mu_A = \frac{1}{\pi}E$ integrated over the process ($\int E\,\mathrm{d}t\,\mathrm{d}x = \Phi = 2\pi$). It is a mixed 't Hooft anomaly between $\mathrm{U}(1)_V$ and $\mathrm{U}(1)_A$. Now suppose the theory could be deformed, keeping both symmetries, into a trivially gapped one with a unique ground state of $Q_A = 0$. Threading the flux slowly would then return the ground state to itself, contradicting $\Delta Q_A = 2$. Indeed, the only mass term, $m\bar\psi\psi$, breaks $\mathrm{U}(1)_A$: to open a gap one must give up a symmetry. The same flux-threading logic, on a lattice, is the Lieb–Schultz–Mattis argument.

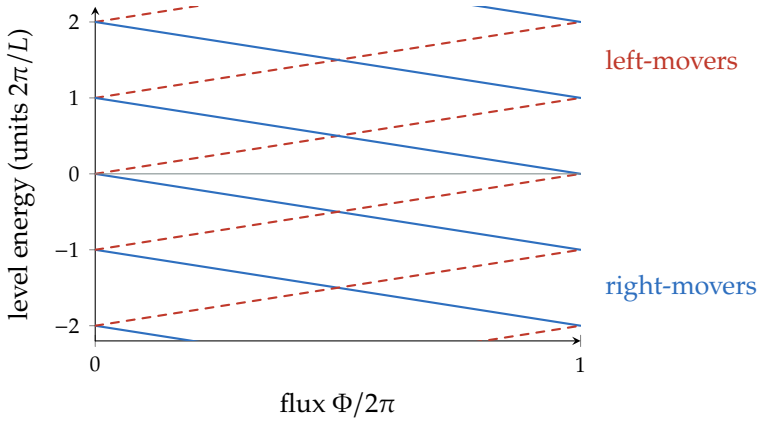


Figure 13.26: Spectral flow of a massless Dirac fermion on a ring as one flux quantum is threaded. Right-moving levels (blue) all slide down by one spacing and left-moving levels (red, dashed) slide up. At the end the spectrum is the same, but one right-mover has sunk into the Dirac sea and one left-mover has left it. The axial charge has changed by two: an anomaly, visible as a level crossing through zero energy.

13.7.2 The Lieb–Schultz–Mattis theorem

Key idea: Lieb–Schultz–Mattis, and Oshikawa's extension

A translation-invariant lattice system with a conserved U(1) charge (for spins, the total S^z), at a filling of ν particles per unit cell, cannot have a unique gapped ground state unless ν is an integer. For $\nu = p/q$ in lowest terms, a gapped phase has at least q degenerate ground states. For spin-S chains at zero magnetisation, $\nu = S$ (count $S + S_j^z$ as the number of bosons at site j), so half-integer spin chains cannot be trivially gapped.

The original argument (one dimension, spin $\frac{1}{2}$). Take the Heisenberg chain (13.43) with $S = \frac{1}{2}$ on a ring of L sites (L even), with ground state $|\Psi_0\rangle$ of energy E_0, total $S^z = 0$ and momentum P_0. Apply a slow twist,

$$U = \exp\Big(\frac{2\pi i}{L}\sum_{j=1}^{L} j\, S_j^z\Big),$$

which rotates spin j about the z axis by the angle $2\pi j/L$. Two facts follow.

(i) The twisted state has low energy. Since $e^{-i\alpha S^z} S^+ e^{i\alpha S^z} = e^{-i\alpha} S^+$, each bond term transforms as $U^\dagger S_j^+ S_{j+1}^- U = e^{2\pi i/L} S_j^+ S_{j+1}^-$ (the angles differ by $2\pi/L$ on every bond, including the bond from L to 1 where they differ by $2\pi/L - 2\pi$). Write $H = J\sum_j \big[\frac{1}{2}(S_j^+ S_{j+1}^- + S_j^- S_{j+1}^+) + S_j^z S_{j+1}^z\big]$ and $c = \langle\Psi_0|S_j^+ S_{j+1}^-|\Psi_0\rangle$, which is real and independent of j by translation and reflection symmetry. Then

$$\begin{aligned}\langle\Psi_0|U^\dagger H U|\Psi_0\rangle - E_0 &\overset{(1)}{=} \frac{J}{2}\sum_{j=1}^{L}\big[(e^{2\pi i/L}-1)+(e^{-2\pi i/L}-1)\big]c \overset{(2)}{=} JLc\Big(\cos\frac{2\pi}{L}-1\Big)\\ &\overset{(3)}{\approx} -\frac{2\pi^2 Jc}{L} = O\Big(\frac{1}{L}\Big).\end{aligned} \tag{13.50}$$

(1) The S^zS^z terms commute with U and cancel. The two hopping terms pick up $e^{\pm 2\pi i/L}$, and both have expectation value c.
(2) L equal terms, and $e^{ix} + e^{-ix} - 2 = 2(\cos x - 1)$.
(3) $\cos x - 1 \approx -x^2/2$ with $x = 2\pi/L$.

(ii) The twisted state is orthogonal to the ground state. Let T translate by one site, $T S_j^z T^{-1} = S_{j+1}^z$. Then

$$\begin{aligned}TUT^{-1} &\overset{(1)}{=} \exp\Big(\frac{2\pi i}{L}\sum_{j=1}^{L} j\, S_{j+1}^z\Big) \overset{(2)}{=} \exp\Big(\frac{2\pi i}{L}\sum_{k=1}^{L}(k-1)S_k^z + 2\pi i S_1^z\Big)\\ &\overset{(3)}{=} U\, e^{-\frac{2\pi i}{L}S_{\text{tot}}^z}\, e^{2\pi i S_1^z} \overset{(4)}{=} -U.\end{aligned} \tag{13.51}$$

(1) Conjugate each S_j^z.
(2) Relabel $k = j + 1$. The last term, $j = L$, is $L\, S_{L+1}^z = L\, S_1^z$ on the ring. The $k = 1$ term of $\sum_k (k-1)S_k^z$ is zero, so the sum equals $\sum_{k=1}^{L}(k-1)S_k^z + L\, S_1^z$, and $\frac{2\pi i}{L}\cdot L = 2\pi i$.
(3) The operators commute, so the exponential splits.
(4) $S_{\text{tot}}^z = 0$, and $e^{2\pi i S_1^z} = e^{\pm i\pi} = -1$ for spin $\frac{1}{2}$.

So $TU|\Psi_0\rangle = -UT|\Psi_0\rangle = e^{i(P_0+\pi)}U|\Psi_0\rangle$: the twisted state has momentum $P_0 + \pi$, and is therefore orthogonal to $|\Psi_0\rangle$.

Together, (i) and (ii) give a state orthogonal to the ground state with energy at most $O(1/L)$ above it. As $L \to \infty$, either there is a second ground state (degeneracy, as in the dimerised phase) or there are excitations of arbitrarily low energy (gaplessness). A unique gapped ground state is impossible. For integer spin, step (4) gives $e^{2\pi i S_1^z} = +1$: no momentum shift, and no conclusion. That is exactly the room the Haldane gap needs.

Oshikawa's flux-threading argument. In higher dimensions the energy estimate (i) fails (the twist costs energy proportional to the cross-section of the system). Oshikawa replaced it by a topological argument borrowed from Laughlin's explanation of the quantum Hall effect. Give the conserved particles a fictitious charge, make the system a ring in the x direction (length L, cross-section of C unit cells), and thread a flux Φ through the ring (Figure 13.27).

1. Increase Φ adiabatically from 0 to 2π. The Hamiltonian stays translation invariant throughout, so the momentum P_x of the evolving state is conserved.
2. At $\Phi = 2\pi$ the Hamiltonian is gauge equivalent to the original one: $H(2\pi) = UH(0)U^{-1}$ with $U = \exp\left(\frac{2\pi i}{L}\sum_{\vec{r}} x_{\vec{r}} n_{\vec{r}}\right)$, the large gauge transformation (the analogue of the twist above).
3. By the same computation as (13.51), undoing the gauge transformation shifts the momentum by $2\pi\nu C$.
4. If the system is gapped, adiabatic evolution takes a ground state to a ground state. If $\nu C \notin \mathbb{Z}$, the final ground state has a different momentum from the initial one, so it is a different ground state.

Repeating, for $\nu = p/q$ with C coprime to q, there are at least q degenerate ground states. Hastings (2004) made the argument rigorous in any dimension.

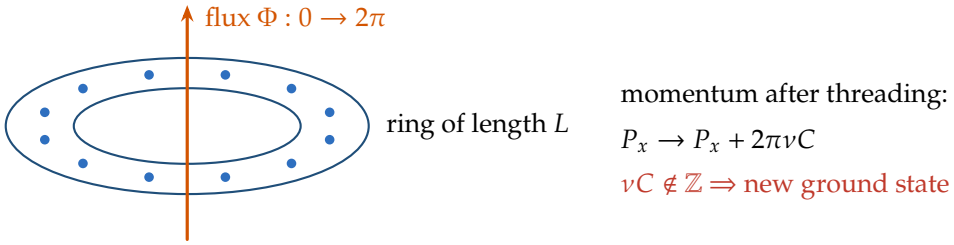


Figure 13.27: Oshikawa's argument. The lattice system (dots) is a ring threaded by a fictitious flux. Adding one flux quantum slowly returns the Hamiltonian to itself up to a gauge transformation, but shifts the crystal momentum of the state by $2\pi\nu C$. At fractional filling the state that comes back is orthogonal to the one that went in, so a gapped system must have several ground states.

Physical picture: LSM is an anomaly

In field-theory language, LSM says that lattice translation and the U(1) (or SO(3)) spin rotations have a mixed 't Hooft anomaly whenever there is a half-integer spin per unit cell. All three escape routes of Figure 13.25 are realised. The spin-$\frac{1}{2}$ Heisenberg chain is gapless (the $SU(2)_1$ CFT, which carries the anomaly). The J_1–J_2 chain dimerises (broken translation). In two dimensions a spin liquid with $\mathbb{Z}_2$ topological order can be gapped and symmetric: its TQFT carries the anomaly, through anyons that carry spin $\frac{1}{2}$ (Section 13.8.2). Oshikawa pointed out that such a state must have a topological ground-state degeneracy on the torus, consistent with the argument above.

Worked example 13.28: Yang–Mills at $\theta = \pi$

Four-dimensional SU(N) Yang–Mills theory has a θ-term $\frac{\theta}{8\pi^2}\int \operatorname{tr} F \wedge F$ (Definition 13.4), and is time-reversal invariant at $\theta = 0$ and $\theta = \pi$. It also has a "one-form" symmetry, the centre $\mathbb{Z}_N$ of SU(N), which acts on Wilson loops rather than on local operators. Gaiotto, Kapustin, Komargodski and Seiberg (2017) showed that at $\theta = \pi$ there is a mixed 't Hooft anomaly between time reversal and the $\mathbb{Z}_N$ centre symmetry. In their words, "at $\theta = \pi$ the vacuum cannot be a trivial non-degenerate gapped state", while at $\theta = 0$ it is believed to be. If the theory is gapped and confining at $\theta = \pi$ (as expected at large N), then time reversal must be spontaneously broken there, with two degenerate vacua and a domain wall between them that carries a Chern–Simons theory. The two-dimensional analogue is the $\mathbb{CP}^{N-1}$ model at $\theta = \pi$: for $N = 2$ it is gapless (Haldane's case), and for $N > 2$ time reversal is broken. So even for the theory of the Clay problem, anomalies tell us something rigorous about where the gap *can* be trivial and where it cannot.

Watch out: what anomalies do not tell you

Anomaly matching is a constraint, not a solution. It says that a trivially gapped phase is impossible, and it lists the alternatives, but it does not say which one is realised. Nor does the *absence* of an anomaly imply a gap: the three-dimensional O(N) model at weak coupling (Section 13.3.6) has no anomaly and is gapless because of spontaneous symmetry breaking.

Recap

- An 't Hooft anomaly is an obstruction to coupling a global symmetry to background gauge fields. It is invariant under the RG.
- A trivially gapped theory has a local response $W[A]$, hence no anomaly. So an anomalous theory is gapless, breaks the symmetry, or is topologically ordered.
- Spectral flow of a Dirac fermion under flux threading exhibits the two-dimensional anomaly: $\Delta Q_A = 2$.
- LSM: a twist U gives a state of energy $O(1/L)$ and momentum shifted by π for half-integer spin. Oshikawa's flux threading extends this to all dimensions: fractional filling forbids a unique gapped ground state.
- SU(N) Yang–Mills at $\theta = \pi$ cannot be trivially gapped.

Sources. M. Oshikawa, *Commensurability, excitation gap and topology in quantum many-particle systems on a periodic lattice* (cond-mat/9911137), eqs. (1)–(4); M. B. Hastings, *Lieb–Schultz–Mattis in higher dimensions* (cond-mat/0305505); E. H. Lieb, T. Schultz and D. J. Mattis, Ann. Phys. 16 (1961) 407; D. Gaiotto, A. Kapustin, Z. Komargodski and N. Seiberg, *Theta, time reversal, and temperature* (arXiv:1703.00501), §1; T. Sulejmanpasic and Y. Tanizaki (arXiv:1802.02153) and Z. Komargodski, A. Sharon, R. Thorngren and X. Zhou (arXiv:1705.04786) for the anomaly of the $\mathbb{CP}^1$ model at $\theta = \pi$.

13.8 Supersymmetry, the Witten index, and gapped phases as TQFTs

Two more ideas complete the picture. The Witten index is a topological count of vacua that cannot change under deformations: it detects, without solving the theory, whether zero-energy vacua exist. And every gapped phase, looked at below its gap, is a topological quantum field theory. Together they explain why the language of Nash's book (Morse theory, TQFT) is the natural language for gapped quantum systems.

13.8.1 Supersymmetric quantum mechanics and the Witten index

Take a particle on a line with a two-state internal space ("fermion number" 0 or 1) and a real function $W(x)$, the **superpotential**. Define

$$Q = \sigma^{-}(p + iW'), \qquad Q^{\dagger} = \sigma^{+}(p - iW'), \qquad \sigma^{+} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \ \sigma^{-} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix},$$

with $p = -i\,\mathrm{d}/\mathrm{d}x$. Then $Q^2 = 0$ (since $(\sigma^{-})^2 = 0$), and

$$\begin{aligned} H \coloneqq \tfrac{1}{2}\{Q, Q^{\dagger}\} &\overset{(1)}{=} \tfrac{1}{2}\big[\sigma^{-}\sigma^{+}(p + iW')(p - iW') + \sigma^{+}\sigma^{-}(p - iW')(p + iW')\big] \\ &\overset{(2)}{=} \tfrac{1}{2}\big(p^2 + W'^2 + W''\,\sigma^3\big). \end{aligned} \tag{13.52}$$

(1) Expand the anticommutator, $QQ^{\dagger} + Q^{\dagger}Q$.

(2) $(p \pm iW')(p \mp iW') = p^2 + W'^2 \mp i[p, W']$ and $[p, W'] = -iW''$, so the products are $p^2 + W'^2 \mp W''$. Also $\sigma^{-}\sigma^{+} = \mathrm{diag}(0, 1)$ and $\sigma^{+}\sigma^{-} = \mathrm{diag}(1, 0)$, which combine into $W''\sigma^3$.

Let $(-1)^F = -\sigma^3$ (the lower component is "bosonic"). It anticommutes with Q and $Q^{\dagger}$.

Worked example 13.29: nonzero energies come in boson–fermion pairs

Let $Q_1 = Q + Q^{\dagger}$. It is Hermitian, anticommutes with $(-1)^F$, and $Q_1^2 = QQ^{\dagger} + Q^{\dagger}Q = 2H$ (using $Q^2 = 0$). Let $|b\rangle$ be bosonic with $H|b\rangle = E|b\rangle$, $E > 0$, and set $|f\rangle = Q_1|b\rangle/\sqrt{2E}$. Then

$$\langle f|f\rangle \overset{(1)}{=} \frac{\langle b|Q_1^2|b\rangle}{2E} = 1, \qquad (-1)^F|f\rangle \overset{(2)}{=} -|f\rangle, \qquad H|f\rangle \overset{(3)}{=} E|f\rangle.$$

(1) Q_1 is Hermitian and $Q_1^2 = 2H$.

(2) $(-1)^F Q_1 = -Q_1(-1)^F$.

(3) $[H, Q_1] = \tfrac{1}{2}[Q_1^2, Q_1] = 0$.

So every positive energy is shared by as many bosons as fermions. Only the zero-energy states can be unpaired. Hence the **Witten index**

$$I = \mathrm{Tr}\,(-1)^F e^{-\beta H} = n_B(E = 0) - n_F(E = 0) \tag{13.53}$$

does not depend on β, and it cannot change under any deformation that keeps the large-x behaviour of W (states can only enter or leave $E = 0$ in boson–fermion pairs). If $I \neq 0$, there are exact zero-energy vacua and supersymmetry is unbroken.

Worked example 13.30: counting vacua with Morse theory

Zero-energy states solve $Q\psi = Q^{\dagger}\psi = 0$. For the upper component, $(p + iW')\psi_{\uparrow} = 0$ gives $\psi_{\uparrow} = e^{W}$. For the lower one, $(p - iW')\psi_{\downarrow} = 0$ gives $\psi_{\downarrow} = e^{-W}$. Only normalisable solutions count.

- $W = \frac{1}{4}x^4 - \frac{1}{2}a^2x^2$: e^{-W} is normalisable, e^{W} is not. Exactly one bosonic vacuum, $I = 1$.
- $W = \frac{1}{3}x^3 - a^2x$: neither e^{W} nor e^{-W} is normalisable (the cubic grows with opposite signs at $\pm\infty$). No vacuum, $I = 0$: supersymmetry is broken.

Now compare with a semiclassical count. Near each critical point x_c of W ($W'(x_c) = 0$), $H \approx \frac{1}{2}[p^2 + W''(x_c)^2(x - x_c)^2] + \frac{1}{2}W''(x_c)\sigma^3$: a harmonic oscillator of frequency $|W''(x_c)|$ with zero-point energy $\frac{1}{2}|W''|$, cancelled by $\frac{1}{2}W''\sigma^3$ in the component with $\sigma^3 = -\operatorname{sgn} W''$. So each critical point gives one perturbative vacuum, bosonic if $W'' > 0$ and fermionic if $W'' < 0$, and

$$I = \sum_{x_c} \operatorname{sgn} W''(x_c).$$

For the quartic, the critical points are 0 ($W'' = -a^2$) and $\pm a$ ($W'' = 2a^2$): $I = -1 + 1 + 1 = 1$. For the cubic, $\pm a$ with $W'' = \pm 2a$: $I = 1 - 1 = 0$. The semiclassical count gives the exact index, although in the cubic case the two perturbative vacua are lifted by instantons tunnelling between $\pm a$, to an energy of order $e^{-2[W(-a)-W(a)]}$. This is Witten's proof of the Morse inequalities in miniature (Chapter 7): on a manifold, $Q = \mathrm{d} + \mathrm{d}^*$ deformed by a Morse function, and the index is the Euler characteristic.

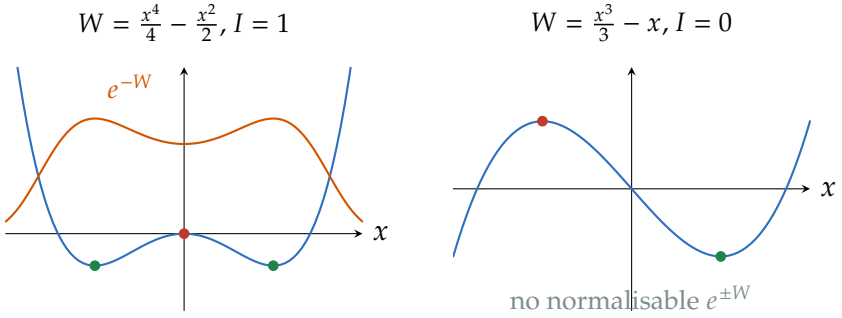


Figure 13.28: Supersymmetric quantum mechanics and Morse theory. The superpotential W is blue. Its minima (green) give bosonic perturbative vacua and its maxima (red) fermionic ones. Left: the signed count is 1, and the exact vacuum e^{-W} (orange) exists. Right: the count is 0, no exact vacuum exists, and the two perturbative vacua are lifted by tunnelling: supersymmetry is broken.

The index and the gap. In a supersymmetric field theory the Witten index, computed in a finite spatial box (where the spectrum is discrete), counts vacua. The four-dimensional $\mathcal{N} = 1$ supersymmetric SU(N) Yang–Mills theory has $\operatorname{Tr}(-1)^F = N$. This matches the expected physics: the theory confines, is gapped, and has N vacua, which spontaneously break a discrete $\mathbb{Z}_{2N}$ chiral symmetry to $\mathbb{Z}_2$ through a gaugino condensate. A nonzero index guarantees exact vacua. It does not by itself guarantee a gap. And if supersymmetry is spontaneously broken in a field theory, Goldstone's theorem gives a massless fermion (the goldstino) and the gap closes.

Reading Nash (Ch. VII, §4, p. 207, (7.60)–(7.69))

Nash introduces supersymmetry generators Q_1, Q_2 with $Q_1^2 = H + P$, $Q_2^2 = H - P$ (7.61), the operator $(-1)^F$ (7.63), and the realisation $Q_1 = \mathrm{d} + \mathrm{d}^*$, $Q_2 = i(\mathrm{d} - \mathrm{d}^*)$ on differential forms (7.68), with even forms bosonic and odd forms fermionic. Then $\operatorname{Tr}(-1)^F$ counts harmonic forms with signs, which is the Euler characteristic. Our one-dimensional example is the case $M = \mathbb{R}$ with a Morse function W, where the "forms" are the two components of ψ.

13.8.2 Gapped phases are topological field theories

Suppose a theory has a gap Δ. Probe it at energies far below Δ (or distances far beyond ξ). Only the vacua are left, and on the vacua the Hamiltonian is a constant, which we set to zero. What remains is: for every spatial manifold Σ, a finite-dimensional space $\mathcal{H}(\Sigma)$ of vacua, and for every spacetime, an amplitude that depends only on its topology (after removing non-universal local terms such as the vacuum energy density times the volume). These are the data of a **topological quantum field theory** in the sense of Atiyah (Chapter 12).

Key idea: the infrared of a gapped phase

The long-distance limit of a gapped phase is a TQFT. The long-distance limit of a gapless phase is a scale-invariant theory (a CFT, or free massless particles). Classifying gapped phases of matter is therefore largely classifying TQFTs, possibly with symmetry.

gapped phase	$\dim \mathcal{H}(\Sigma)$, genus g	example
trivially gapped	1	free massive field
broken $\mathbb{Z}_2$	2 (any Σ)	ordered Ising chain, Gross–Neveu
$\mathbb{Z}_2$ topological order	4^g	toric code, $\mathbb{Z}_2$ spin liquid
$\mathrm{U}(1)_k$ Chern–Simons	$\lvert k\rvert^g$	Maxwell–Chern–Simons, quantum Hall
SPT (e.g. Haldane phase)	1, but edge modes	AKLT chain

Worked example 13.31: the toric code: a gap and a topological degeneracy

Put a spin $\frac{1}{2}$ on every edge of a square lattice drawn on a closed surface Σ. For each vertex v let $A_v = \prod \sigma^x$ over the four edges touching v, and for each face p let $B_p = \prod \sigma^z$ over its four edges. Take

$$H = -\sum_v A_v - \sum_p B_p .$$

Every A_v commutes with every B_p: they share either no edge or exactly two, and $\sigma^x\sigma^z = -\sigma^z\sigma^x$ on each shared edge, so the two signs cancel. All terms can be diagonalised together, with eigenvalues ± 1. A ground state has all $A_v = B_p = +1$. An excitation flips at least one sign, and since the operators that create excitations (σ^z or σ^x on an edge) flip them in pairs, it costs at least 4: an exact gap. Now count the ground states. There are 2^E states for E edges. Each independent constraint $A_v = 1$ or $B_p = 1$ halves the count. The constraints are not all independent: $\prod_v A_v = 1$ and $\prod_p B_p = 1$ identically (each edge appears twice). With V vertices and F faces,

$$\dim \mathcal{H}_0 \overset{(1)}{=} 2^{E-(V-1)-(F-1)} \overset{(2)}{=} 2^{2-\chi(\Sigma)} \overset{(3)}{=} 2^{2g} = 4^g .$$

(1) $V-1$ independent vertex constraints and $F-1$ independent face constraints.
(2) Euler's formula $V - E + F = \chi(\Sigma)$ (Chapter 1).
(3) $\chi = 2 - 2g$ for a surface of genus g.

One ground state on the sphere, four on the torus. The four torus ground states are distinguished by the eigenvalues of $\prod \sigma^z$ along the two noncontractible loops, the elements of $H^1(T^2;\mathbb{Z}_2) = \mathbb{Z}_2^2$ (Figure 13.29). No local operator can tell them apart, which is why in a perturbed model their splitting is only of order $e^{-L/\xi}$.

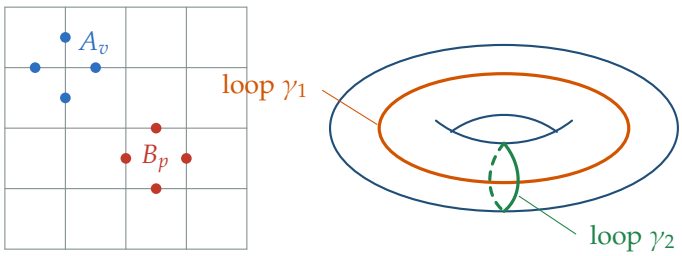


Figure 13.29: The toric code. Left: spins live on edges. The vertex term A_v (blue) acts with σ^x on the four edges at a vertex, and the face term B_p (red) with σ^z on the four edges of a face. Right: on a torus the ground states are labelled by the values of σ^z-strings around the two noncontractible loops. They are four vacua that no local measurement distinguishes: topological order, and a TQFT in the infrared.

Recap

- In supersymmetric theories, nonzero energies pair up, and the Witten index $\mathrm{Tr}(-1)^F = n_B - n_F$ at zero energy is a deformation invariant, computable semiclassically as a Morse count $\sum \operatorname{sgn} W''(x_c)$.
- A nonzero index guarantees exact vacua. Spontaneous supersymmetry breaking in field theory produces a massless goldstino.
- Below its gap, every gapped phase is a TQFT: finite vacuum spaces $\mathcal{H}(\Sigma)$ that depend only on topology. The toric code has $\dim \mathcal{H} = 4^g$, from Euler's formula.

13.9 The Yang–Mills mass gap problem

We can now state the most famous open problem about gaps, and see which parts of this chapter bear on it. Pure Yang–Mills theory in four dimensions is the theory of gluons without quarks. Classically it describes massless waves, like Maxwell theory. Quantum mechanically it is believed to have a mass gap, to confine, and to have as its lightest particles colourless bound states of gluons, the **glueballs**. Nobody has proved any of this. In 2000 the Clay Mathematics Institute made it one of its seven Millennium Prize Problems, with the problem description written by Jaffe and Witten.

13.9.1 The statement

Key idea: Yang–Mills existence and mass gap (Jaffe and Witten)

Prove that for any compact simple gauge group G, a non-trivial quantum Yang–Mills theory exists on $\mathbb{R}^4$ and has a mass gap $\Delta > 0$. Existence includes establishing axiomatic properties at least as strong as those cited in [the Wightman axioms and the Osterwalder–Schrader axioms].

Every phrase carries weight. We unpack them in turn.

"Compact simple gauge group G". For example $\mathrm{SU}(N)$ with $N \geq 2$, $\mathrm{SO}(N)$, $\mathrm{Sp}(N)$, or an exceptional group. The Lie algebra has no abelian factor, so U(1) (Maxwell theory, which is free

and gapless) is excluded. The classical theory is the Yang–Mills action $\frac{1}{4g^2}\int \operatorname{Tr} F \wedge \star F$ with $F = \mathrm{d}A + A \wedge A$ (Chapter T3).

"Exists". The theory must be a quantum field theory in the precise sense of the axioms. In Minkowski form (Wightman), there is a Hilbert space carrying a unitary representation of the Poincaré group, the energy is non-negative, there is a unique Poincaré-invariant vacuum Ω, and there are local field operators that commute at spacelike separation. Jaffe and Witten specify which fields: local operators "in correspondence with the gauge-invariant local polynomials in the curvature F and its covariant derivatives, such as $\operatorname{Tr} F_{ij}F_{kl}(x)$". Their correlation functions must agree at short distances with asymptotic freedom and perturbative renormalisation theory, including the existence of a stress tensor and an operator product expansion.

Working definition 13.6 (Euclidean QFT in the sense of Osterwalder and Schrader)

A **Euclidean quantum field theory** is a collection of correlation functions $S_n(x_1, \dots, x_n)$ of local fields on $\mathbb{R}^d$ [the Euclidean path-integral averages $\langle \phi(x_1) \cdots \phi(x_n) \rangle$, for example of $\operatorname{Tr} F^2$] that are invariant under Euclidean motions, symmetric (or antisymmetric for fermions), clustering [they factorise when two groups of points are pulled far apart, as in (13.7)], and **reflection positive** [reflecting in the plane $\tau = 0$ and complex conjugating gives a non-negative inner product, which is what later becomes the positivity of probabilities in the Hilbert space]. *Why it matters:* the Osterwalder–Schrader reconstruction theorem turns such correlation functions into a Wightman theory, with Hilbert space, Hamiltonian and vacuum. The spectral sum of Section 13.2.1 is exactly this reconstruction, read backwards.

"Non-trivial". Not a free theory. (A free theory of massless fields would exist, but would not be Yang–Mills theory.)

"Has a mass gap $\Delta > 0$". With $H\Omega = 0$, the Hamiltonian has no spectrum in the interval $(0, \Delta)$ (Definition 13.1). Jaffe and Witten note the consequence we derived in Section 13.2: exponential clustering, $|\langle \Omega, \mathcal{O}(\vec{x})\mathcal{O}(\vec{y})\Omega \rangle| \le e^{-C|\vec{x}-\vec{y}|}$ for $C < \Delta$. They also exclude one shortcut: "weak existence (compactness)" of a limit of approximations does not count, unless the properties of the limit (the gap and the axioms) are also established.

Watch out: two problems, not one

The problem has two halves, and both are open. *Existence*: no interacting quantum field theory in four dimensions has ever been constructed to satisfy the axioms. *Mass gap*: even granted existence, the gap must be proved. The gap is also not the same as confinement. A Higgs phase is gapped and not confining (Section 13.3.2). Proving confinement (an area law for Wilson loops) is listed by Jaffe and Witten as a natural *extension* of the problem, not as part of it.

13.9.2 Why it is hard

The gap is invisible classically and perturbatively. Classical Yang–Mills waves are massless. Perturbation theory in g describes massless gluons to every order. The expected gap is $\Delta = c\,\Lambda = c\,\mu\, e^{-1/(\beta_0 g^2(\mu))}$ with $\beta_0 = \frac{11N}{24\pi^2}$ for SU(N) (Section 13.3.3), and this function has a vanishing Taylor series at $g = 0$ (Example 13.15). Jaffe and Witten: "because four-dimensional gauge theory is a theory in which the mass gap is not classically visible, to demonstrate it

may require a non-classical change of variables or duality transformation". Polyakov's model (Section 13.4) is the example where such a change of variables works.

The continuum limit is at weak coupling. On a lattice with spacing a the theory is a well-defined finite-dimensional integral in finite volume. But the continuum limit $a \to 0$ at fixed physical Δ requires $\Delta a \to 0$, that is $\xi \to \infty$ in lattice units, which by (13.20) means $g(1/a) \to 0$ in a precise way.

Worked example 13.32: how the lattice coupling must be tuned

With $\mu = 1/a$ in (13.18), $1/g^2(1/a) = \beta_0 \ln \frac{1}{a\Lambda}$. Halving the lattice spacing changes the coupling by

$$\frac{1}{g^2(2/a)} - \frac{1}{g^2(1/a)} \overset{(1)}{=} \beta_0 \ln 2 \overset{(2)}{=} \frac{33}{24\pi^2} \times 0.693 = 0.097.$$

(1) Difference of (13.18) at $\mu = 2/a$ and $\mu = 1/a$.
(2) $\beta_0 = 11N/24\pi^2$ with $N = 3$.

Lattice practitioners use $\beta_{\text{lat}} = 6/g^2$ for SU(3), so halving a means raising β_{lat} by about 0.58 at one loop. (Simulations near $\beta_{\text{lat}} \approx 6$ find roughly this rate, with corrections from higher loops.) The dimensionless gap Δa must fall by a factor of two at each step, forever, while the physical Δ stays put. A proof must control an infinite sequence of such steps uniformly: this is an RG problem of exactly the type of Chapter T5, in the hardest possible regime.

The volume must go to infinity. Jaffe and Witten stress that even a complete construction on a four-torus would leave open "the existence of a mass gap that is uniform in the volume", and that "no present ideas point the direction" to it. Finite-volume gaps always exist (Section 13.1.5); the whole content is their persistence as $L \to \infty$.

Four dimensions is the borderline. Scalar ϕ^4 theory, the simplest interacting candidate, exists with a mass gap in two and three dimensions but is *trivial* (becomes free in the continuum limit) in five or more dimensions, and, as proved by Aizenman and Duminil-Copin in 2021 for the lattice regularisation, also in four. Asymptotic freedom is what makes Yang–Mills theory the natural candidate for a non-trivial four-dimensional theory. It is also what puts the interesting physics at strong coupling.

13.9.3 What is known and what is not

Numerical evidence. Monte Carlo simulations of lattice Yang–Mills theory, extrapolated to $a \to 0$ and $L \to \infty$, find a gapped spectrum of glueballs and an area law for Wilson loops. For SU(3) the lightest glueball has $J^{PC} = 0^{++}$ and mass about 3.5 times the square root of the string tension, that is about 1.5–1.7 GeV depending on how physical units are assigned. Heavier states include a 2^{++} near 2.4 GeV and a 0^{-+} near 2.6 GeV (Figure 13.30). The ratios approach finite limits as $N \to \infty$. These results are extremely convincing physically. They are not proofs, because they are finite-volume, finite-spacing numerics with statistical errors.

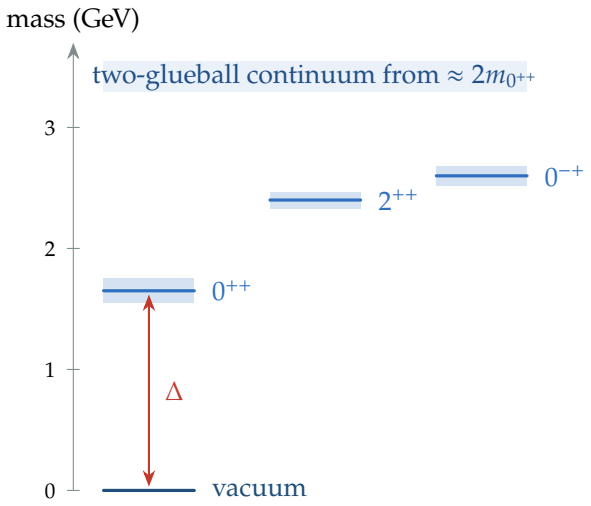


Figure 13.30: Schematic spectrum of pure SU(3) Yang–Mills theory from lattice simulations, at zero momentum. Bands indicate the rough spread between determinations. The lightest state, the scalar 0^{++} glueball, sets the gap. Compare the free massive field in Figure 13.2(a): the shape is the same, but here nothing in the Lagrangian carries a mass.

Worked example 13.33: what a 1.7 GeV gap means for correlations

If the lightest glueball has $m_{0^{++}}c^2 \approx 1.7\,\text{GeV}$, then by Section 13.2.6 the correlation length of pure Yang–Mills theory is

$$\xi \overset{(1)}{=} \frac{\hbar c}{m_{0^{++}}c^2} \overset{(2)}{=} \frac{197.3\ \text{MeV fm}}{1700\ \text{MeV}} \approx 0.12\ \text{fm}.$$

(1) $\xi = 1/\Delta$ with $\hbar$ and c restored.
(2) $\hbar c = 197.3\,\text{MeV fm}$.

So the connected correlator of $\operatorname{Tr} F^2$ at two points 1 fm apart is suppressed by roughly $e^{-1/0.12} = e^{-8.6} \approx 2 \times 10^{-4}$ relative to its value at the scale ξ, while in a gapless theory it would fall only as a power of the distance. A lattice with spacing a must resolve ξ, so $a \ll 0.12\,\text{fm}$, and must contain many correlation lengths, so $L \gg 0.12\,\text{fm}$: modern simulations use $a \approx 0.05$–$0.1\,\text{fm}$ and $L \approx 2$–$4\,\text{fm}$. Proving the Clay statement amounts to proving that this exponential suppression survives $a \to 0$ and $L \to \infty$.

	scalar ϕ^4	abelian gauge	Yang–Mills
$d = 2$	exists, gapped	Higgs: exists, gapped	solvable
$d = 3$	exists, gapped	compact U(1) lattice: gapped, confining	partial (lattice, probabilistic)
$d = 4$	trivial (lattice)	free photon (gapless)	**open**: existence and mass gap

Figure 13.31: A map of what is established, by spacetime dimension (rows) and type of theory (columns). Green cells are constructed with the stated properties, the red cell is proved trivial, yellow cells are partial. The Clay problem is the framed cell: four-dimensional Yang–Mills theory, the only asymptotically free candidate for a non-trivial four-dimensional quantum field theory.

Rigorous results. Here is what has been proved, as surveyed by Jaffe and Witten and in the mathematical literature since.

- *Lattice gauge theory at strong coupling.* For any compact gauge group, the Wilson lattice theory at large g has a mass gap and an area law (Osterwalder and Seiler, 1978). But the continuum limit lives at $g \to 0$, not at large g.
- *Lower dimensions and simpler theories.* ϕ^4 in two and three dimensions (Glimm, Jaffe and others) has a mass gap. The two-dimensional abelian Higgs model was constructed and shown to satisfy the Osterwalder–Schrader axioms (Brydges, Fröhlich and Seiler), with a mass gap proved on the lattice (Balaban, Brydges, Imbrie and Jaffe). Three-dimensional compact U(1) lattice gauge theory confines with a gap (Göpfert and Mack).
- *Balaban's programme.* A renormalisation-group analysis of lattice Yang–Mills theory on three- and four-dimensional tori, establishing ultraviolet stability uniformly in the lattice spacing. It is "an important step toward establishing the existence of the continuum limit on a compactified space-time", but it does not reach gauge invariant correlation functions, infinite volume, or the gap.
- *Two dimensions.* Two-dimensional Yang–Mills theory is exactly solvable, and recent probabilistic work has constructed Yang–Mills measures in two dimensions and made progress in three. The four-dimensional case remains out of reach.

Physics arguments. Each gives strong evidence for a gap, in a setting that differs from pure four-dimensional Yang–Mills theory in one controlled way.

- $2+1$ *dimensions:* Polyakov's compact QED (Section 13.4) and, for non-abelian theories with an adjoint Higgs field, the same monopole mechanism.
- *Two-dimensional analogues:* the $\mathbb{CP}^{N-1}$ model is asymptotically free, classically massless and provably gapped at large N (Section 13.3.4).
- *Supersymmetry:* in $\mathcal{N} = 2$ theories softly broken to $\mathcal{N} = 1$ (Seiberg and Witten), monopole condensation gives a gap and confinement exactly, the dual superconductor of Section 13.4.5; $\mathcal{N} = 1$ super Yang–Mills has N gapped vacua (Section 13.8.1).
- *Large N and holography:* gauge theories dual to gravity in higher dimensions have gapped glueball spectra computable from wave equations in the dual geometry.
- *Anomalies:* at $\theta = \pi$ the theory cannot be trivially gapped (Example 13.28), which constrains how the gap at $\theta = 0$ can evolve with θ.

statement	status	source of confidence
4D YM exists as an axiomatic QFT	open	asymptotic freedom, Balaban
4D YM has a mass gap	open	lattice numerics, analogues
4D YM confines (area law)	open	lattice numerics
lattice YM gapped at strong coupling	proved	cluster expansion
ϕ^4_2, ϕ^4_3 exist with a gap	proved	constructive QFT
compact $\mathrm{U}(1)_3$ (lattice) gapped, confining	proved	Göpfert–Mack
$\mathbb{CP}^{N-1}_2$ gapped at $N = \infty$	derived	saddle point, exact at $N = \infty$
$\mathrm{SU}(N)$ YM at $\theta = \pi$ not trivially gapped	derived	't Hooft anomaly

Physical picture: what the gap would buy

Jaffe and Witten note that clustering "may make it possible to apply mathematical results established on $\mathbb{R}^4$ to any 4-manifold". The reason is locality: in a gapped theory, what happens in one region is independent of what happens a few correlation lengths away, so a theory on a curved four-manifold can be pieced together from flat patches, with only topological information surviving at long distances. Witten's derivation of the Seiberg–Witten invariants from the topologically twisted $\mathcal{N} = 2$ theory (the setting of Donaldson theory in Chapter 12) used exactly this kind of argument, for the supersymmetric extension where the infrared is under control. A gapped theory is one whose long-distance limit is a TQFT (Section 13.8.2): the mass gap is the bridge between quantum field theory and the topology of Nash's book.

Key idea: the chapter in one paragraph

A mass gap is the absence of low-energy excitations of any kind, and it is equivalent to exponential clustering with $\xi = 1/\Delta = \hbar/mc$. Gaps are opened by explicit masses, by the Higgs mechanism and by dimensional transmutation, and, more subtly, by topology: monopole-instantons screen a photon, a Chern–Simons term gives it a mass, and a θ-term decides whether a spin chain is gapped. Topology can also forbid a trivial gap, through 't Hooft anomalies and the Lieb–Schultz–Mattis theorem. Whatever produces it, a gap leaves a TQFT at long distances. For four-dimensional Yang–Mills theory, none of these mechanisms has yet been turned into a proof.

Recap

- The Clay problem: construct 4D Yang–Mills theory for compact simple G satisfying the Wightman or Osterwalder–Schrader axioms, and prove $\operatorname{spec} H \cap (0, \Delta) = \varnothing$ for some $\Delta > 0$.
- The gap is invisible classically and to all orders of perturbation theory, the continuum limit sits at weak coupling, and the gap must be uniform in the volume.
- Lattice numerics give a 0^{++} glueball near 1.5–1.7 GeV. Rigorous results exist only at strong lattice coupling, in lower dimensions or for simpler theories.

Sources. A. Jaffe and E. Witten, *Quantum Yang–Mills theory*, official problem description of the Clay Mathematics Institute (open PDF at claymath.org), §§1–6, from which all quotations in this section are taken. M. Aizenman and H. Duminil-Copin, Ann. Math. 194 (2021) 163, for the triviality of lattice ϕ^4_4. C. Morningstar and M. Peardon, Phys. Rev. D60 (1999) 034509, for the glueball spectrum.

Chapter summary

concept	operational meaning	where
mass gap Δ	no state with energy in $(E_0, E_0 + \Delta)$ in infinite volume	Definition 13.1
invariant mass, spectral condition	$M^2 = E^2 - \vec{p}^2 \geq 0$, so $\Delta = \min M$	Section 13.1.2
continuum threshold	many-particle states start at $2m$ (or $3m$)	Section 13.1.3
degenerate vacua	broken discrete symmetry or topological order, split by $e^{-L/\xi}$	Sections 13.1.4 and 13.1.5
spectral sum	$G_c = \sum \lvert c_n\rvert^2 e^{-(E_n - E_0)\lvert\tau\rvert + i\vec{p}_n\cdot\vec{x}}$	Section 13.2.1
clustering	gap $\Leftrightarrow \lvert G_c\rvert \leq Ce^{-\Delta r}$	Section 13.2.2
effective mass	$-\mathrm{d}\ln C/\mathrm{d}\tau \to m$: how lattices measure gaps	Example 13.6
spectral density ρ	$G_c = \int \mathrm{d}\mu^2 \rho\, G_\mu$, $\rho \geq 0$, gap = bottom of support	Definition 13.2
massive propagator	$(2\pi)^{-d/2}(m/r)^{d/2-1}K_{d/2-1}(mr)$, Yukawa in $d = 3$	Section 13.2.4
correlation length	$\xi = 1/\Delta$, finite off criticality, ∞ at RG fixed points	Section 13.2.5, Chapter T5
Compton wavelength	$\hbar/mc$: range of force, onset of pair creation	Section 13.2.6
Higgs mechanism	$m_A = ev$, $m_h = \sqrt{2\lambda}v$, Meissner screening	Section 13.3.2
dimensional transmutation	$m = c\mu e^{-1/(\beta_0 g^2)}$, zero Taylor series	Section 13.3.3
gap equation	$1/g^2 = \int \frac{\mathrm{d}^d p}{(2\pi)^d}\frac{1}{p^2+m^2}$ ($\mathbb{CP}^{N-1}$, O(N)), GN analogue	Sections 13.3.4 to 13.3.6
dual photon, monopole-instanton	$\frac{e^2}{8\pi^2}(\partial\sigma)^2$, $e^{\pm i\sigma}$ insertions, $\pi_2(S^2)$	Section 13.4, Chapter 8
Polyakov mass, string tension	$m_\gamma^2 = 8\pi^2 K e^{-S_0}/e^2$, $T = 2e^2 m_\gamma/\pi^2$	Sections 13.4.3 and 13.4.4
topological mass	$m = \lvert k\rvert e^2/2\pi$ from $\frac{k}{4\pi}A\,\mathrm{d}A$, one state of spin sgn k	Definition 13.3, Chapter 12
θ-term	$-i\theta Q$, periodic, T-invariant only at $0, \pi$	Definition 13.4
Haldane gap	spin chain = O(3) model with $\theta = 2\pi S$, $g = 2/S$	Section 13.6
't Hooft anomaly	obstruction to gauging; forbids a trivial gap	Definition 13.5, Chapter 10
Lieb–Schultz–Mattis	fractional filling: gapless, degenerate, or topological	Section 13.7.2
Witten index	$\mathrm{Tr}(-1)^F$, Morse count of vacua	Section 13.8.1, Chapter 7
IR TQFT	gapped phase at long distances; toric code $\dim\mathcal{H} = 4^g$	Section 13.8.2, Chapter 12
Yang–Mills problem	existence + gap for compact simple G on $\mathbb{R}^4$	Section 13.9